PROCEEDINGS OF THE
INTERDISCIPLINARY DOCTORAL SCHOOL
2012-2013 ACADEMIC YEAR
FACULTY OF INFORMATION TECHNOLOGY
PÁZMÁNY PÉTER CATHOLIC UNIVERSITY
BUDAPEST
2013

Faculty of Information Technology
Pázmány Péter Catholic University

PhD PROCEEDINGS

PROCEEDINGS OF THE INTERDISCIPLINARY DOCTORAL SCHOOL 2012-2013 ACADEMIC YEAR FACULTY OF INFORMATION TECHNOLOGY PÁZMÁNY PÉTER CATHOLIC UNIVERSITY BUDAPEST

June, 2013

Pázmány University ePress
Budapest, 2013

Kiadja a Pázmány Egyetem eKiadó
2013
Budapest

Felelős kiadó
Ft. Dr. Szuromi Szabolcs Anzelm O. Praem.
a Pázmány Péter Katolikus Egyetem rektora

Készült a
TÁMOP-4.2.1.B-11/2/KMR-2011-0002 és a TÁMOP-4.2.2/B-10/1-2010-0014
projekt keretében, és az Új Széchenyi Terv támogatásával

A kiadvány megjelentetését az EMMI a
53724-2/2012/FOFEJL számú szerződés alapján támogatta.

Cover image by **András Laki**, A simple microfluidic technique has been developed to detect living parasites from veterinarian blood using a monolithic polydimethylsiloxane (PDMS) structure. Several intravenous parasitosis can be observed by this developed microcapillary system such as dirofilariasis or Lyme disease. Inside this microfluidic device a special flow-through separator structure has been implemented, which contains a cylindrical active zone, where the microfilariae or other few micron-size parasitic infections remain trapped. The center region is partially surrounded by rectangular cross-section shaped microcapillaries.
TOP: Geometric description of the nematode filter; BOTTOM: The manufactured filter during veterinarian test.

A borítón **Laki András** ábrája látható: Az általunk fejlesztett, monolitikus (polidimetilsziloxán (PDMS)-üvegtechnikával előállított) mikrofluidikai eszköz segítségével kimutathatjuk vérben élő paraziták jelenlétét állatorvosi mintákból. Készülékünkkel több parazitológiai fertőzés detektálható, mint például a Dirofilaria fajok és a Lyme-kór. Mikrofluidikai eszközünkben egy speciális, keresztülfolyásos szűrőt implementáltunk, melynek központi aktív régiójában maradnak vissza a kiszűrt, pár mikron nagyságú paraziták. A központi aktív régiót mikrokapilláris csatornák szegélyezik.
FELÜL: A szűrő geometriai leírása; ALUL: A legyártott szűrő állatorvosi használat során.

HU ISSN 1788-9197

Contents

Introduction

It is our pleasure to publish this Annual Proceedings again to demonstrate the genuine interdisciplinary research done at the Jedlik Laboratories by young talents working in the Interdisciplinary Doctoral School of the Faculty of Information Technology at Pázmány Péter Catholic University. The scientific results of our PhD students show the main recent research directions in which our faculty is engaged. Thanks are also due to the supervisors and consultants, as well as to the five collaborating National Research Laboratories of the Hungarian Academy of Sciences, the Semmelweis Medical School and the University of Pannonia. The collaborative work with the partner universities, especially, Katolieke Universiteit Leuven, Politecnico di Torino, Technische Universität München, University of California at Berkeley, University of Notre Dame, Universidad de Sevilla, Universita di Catania is gratefully acknowledged.

As an important development of this special collaboration, we were able to jointly accredit a new undergraduate curriculum on Molecular Bionics with the Semmelweis Medical School, the first of this kind in Europe.

We acknowledge the many sponsors of the research reported here. Namely,

- the Hungarian National Research Fund (OTKA),
- the Hungarian Academy of Sciences (MTA),
- the National Development Agency (NFÜ),
- the Gedeon Richter Co.,
- the Office of Naval Research (ONR) of the US,
- NVIDIA Ltd.,
- Eutecus Inc., Berkeley, CA,
- MorphoLogic Ltd., Budapest,
- Analogic Computers Ltd., Budapest,
- AnaFocus Ltd., Seville,

and some other companies and individuals.

Needless to say, the resources and support of the Pázmány Péter Catholic University is gratefully acknowledged.

Budapest, June 2013.

Tamás Roska	Gábor Prószéky	Péter Szolgay
Head of the Jedlik Laboratory	Chairman of the Board of the Doctoral School	Head of the Doctoral School

Content processing for light field displaying

Vamsi Kiran Adhikarla
(Supervisor: Péter Szolgay)
v.kiran@holografika.com

***Abstract* -- In this paper, we present a view synthesis method for generating multiview image sequences for 3DTV systems using a sparse set of views obtained from cameras in a multiview linear camera configuration. First, the input images are analyzed to extract information about sparse disparity and a mesh is constructed using a set of sparse disparities on all virtual views. Then, for each virtual view, pixels are interpolated inside the mesh by formulating and solving special warping functions and by fitting a uniform bicubic surface to the original data points from the input images. The method is fully automatic and can generate visually pleasing virtual views. Furthermore, we do not need any post processing operations like occlusion handling, hole filling or inpainting because of the warp driven approach. The method also supports view extrapolation in a limited range and can be implemented in real time, which is extremely needed in the present scenario.**

***Keywords* — HoloVizio, multiview video, lightfield displays, image warping, view synthesis**

I. Introduction

Stereoscopic 3D is a widely popular 3D technology for creating and enhancing the illusion of depth by presenting two perspectives of a scene separately to the left and right eye of the viewer. Very efficient and accurate methods are already available to create and handle such stereoscopic 3D data to ensure high quality end-user experience. However, in many cases two views are not sufficient to reproduce all natural 3D cues, and the user must necessarily wear glasses for 3D perception in stereoscopic 3D. Multiview autostereoscopic 3D display technology is designed to address these shortcomings of stereoscopic 3D. Autostereoscopic 3D is a glasses free technology and the display uses a separate lens arrangement for transmitting/ blocking light in specific directions. These displays can project multiple views and also accommodate motion parallax to allow more natural 3D depth cue.

The field of view (FOV) of a multiview autostereoscopic displays is very limited because of the smaller number of views (typically 5-9). On the other hand, the transition between the two successive views is not smooth when the user moves around in front of the display. LightField Displays (LFDs) address these shortcomings of multiview autostereoscopic displays. LFDs can provide very large FOVs with continuous and smooth transition between individual views and it is also possible to extend the motion parallax in vertical direction. Fig. 1 illustrates the principle of an LFD. An array of optical modules project light beams to hit a special holographic screen at various angles of incidence. The holographic screen then does the necessary optical transformation to distribute the light in various directions. The resulting 3D images are more natural since the light beams emitted correspond to the collection of light rays from each three dimensional coordinate in real world. HoloVizio, an LFD which is built on this principle has been proposed and developed by Holografika [5].

3D content creation today is dominated by stereo in all applications because it has less complexity, and is predicted to remain standard over many years [1]. Thus, it is needed to convert a limited number of views to a much larger number of views. LFDs support almost 20 times the interaxial distance of typical stereoscopic 3D content which makes content creation more tedious. Many ways to generate the required N views from M views ($M<N$) have been already proposed. These can be divided in to two main categories: depth based methods and warping based methods. Depth Image Based Rendering (DIBR) [3] is a very popular technique that falls under the first category and makes use of depth information in the scene to discriminate between different depth layers to generate virtual views.

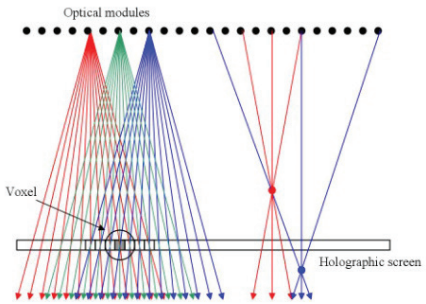


Fig. 1. Concept of HoloVizio LFD.

In many cases, the depth generation [4] process is ill-posed and this makes it necessary to have a pre-processing algorithm, to refine the initial depth map. Fully automatic depth generation with reliable accuracy and robustness remains an unsolved problem today. On the other hand, warping based methods [1] are simple to use and completely

V. K. Adhikarla, "Content processing for light field displaying,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 9-12.

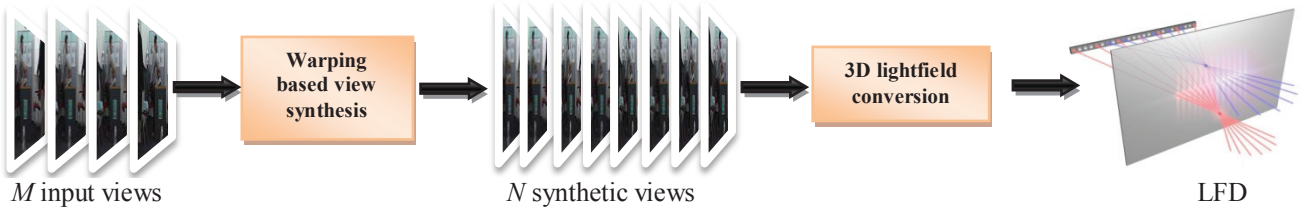


Fig. 2. Concept of multiview generation & visualization on a LFD.

automatic, providing high quality results. They work directly on the input images and do not rely on depth estimation, which reduces the amount of processing. Thus it is more convenient to investigate possibility of real time implementation of such algorithms. In this paper, we present a warping based approach to synthesize the virtual views for a LFD automatically.

The rest of this paper is organized as follows. In the next section we give an overview of the system concept. Details of the algorithm are described in section 3. Then, section 4 presents the results, and finally, section 5 concludes the paper.

II. System Concept

As already mentioned in section 1, LFDs support wide FOVs and thus, the larger the number of views, the better the quality of the displayed image. The system uses an LFD requiring N views, and these N views are generated from a smaller number M of input views (M < N). Also we assume that the input *M* views are rectified i.e., the epipolar lines of all views are horizontally aligned. Fig. 2 illustrates the concept of the system.

It is a tedious task to produce the required *N* views for an LFD using *N* real time cameras because of the physical size of the cameras. Also we assume that the input images are projected to a common image plane (rectified), which imposes constraints on alignment and synchronization of the cameras. To fill the large FOV of the LFD, along with interpolation, we also extrapolate the views to a limited extent. In contrast to the DIBR methods, extrapolation does not result in serious problems due to disocclusions and hence there is no need for any further post-processing steps. The algorithm is described in detail in the following section. The resulting *N* views are processed using a lightfield converter which encodes the views in to a suitable format as required for a LFD.

III. Algorithm Description

In order to transform one view into another, we need a non-linear transformation. Information on this non-linear transformation is stored in an image, which is normally referred to as disparity map. The disparity map carries information on how much each pixel is shifted (horizontally) from one view to another. In the present algorithm, we first estimate a sparse disparity set and then use it to generate virtual views.

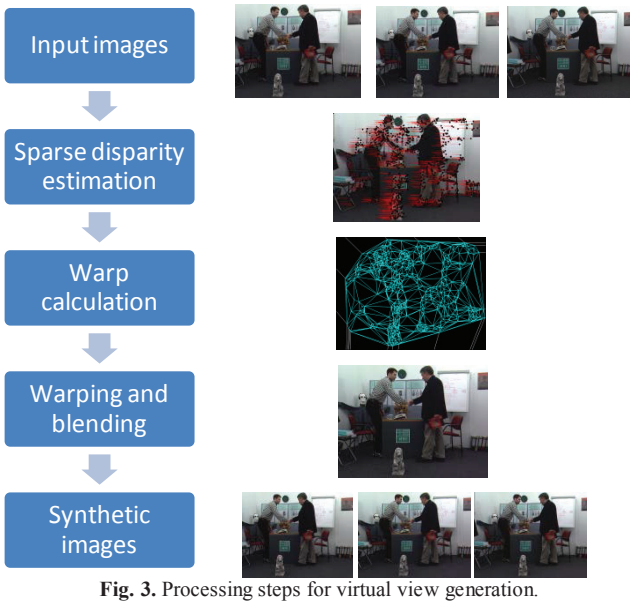


Fig. 3. Processing steps for virtual view generation.

Let us represent the set of input images as $\{I_1, I_2, I_3, \ldots I_m\}$. As shown in Fig. 3, the overall algorithm contains three steps.

A. Sparse disparity estimation

Feature extraction matching is applied to detect reliable and accurate disparities. In addition to that, the extracted features are matched across all *M* input images as shown in Fig. 4, to ensure the robustness of the extracted features.

B. Warp calculation

Warping distorts the input images and transforms them to a new perspective. Different regions of the image should be affected in a different manner and, in order to achieve this, we divide the image into various regions by incorporating a simple triangular mesh.

Fig. 4. Feature matching across all input images.

1. Feature relocation

To generate a specific intermediate view, first we relocate all the sparse features extracted in the first step to the new location on the intermediate image as shown in Fig. 5. The destination locations for each feature are calculated by properly weighing the available feature locations on the input images.

2. Segregating the warping zones

The next step is to isolate the regions on input images which should be affected by a single warping function. These regions are referred to as warping zones. We construct a triangular mesh on each intermediate image using the approach in [2]. The vertices of each triangle denote a zone border, represented by a six element vector, $t_k = \{x_{1k}, y_{1k}, x_{2k}, y_{2k}, x_{3k}, y_{3k}\}$, where (x_{1k}, y_{1k}), (x_{2k}, y_{2k}) & (x_{3k}, y_{3k}) are the coordinates of the vertices of a triangle t_k on a specific intermediate image. Thus for every intermediate image, we have a set of warping boundaries, denoted by a vector $\mathbf{T} = \{t_1, t_2, t_3, \ldots, t_p\}$ which contains the border information. Note that the number of zones may differ from one intermediate image to other. For a specific intermediate image, we fill the warping zones on it by considering immediate left and right images to it.

3. Defining the warp

Now we will solve a simple warp function for each warping zone for a specific intermediate view to identify the pixel locations on the sources images.

Consider an intermediate view; $I_{1.5}$ between the pair of images I_1 & I_2. For this view we have a set of warping zones in a vector **T**. Consider a single warping zone $t_k = \{x_{1k}, y_{1k}, x_{2k}, y_{2k}, x_{3k}, y_{3k}\}$. Let the coordinates of all the pixels inside this warping zone are represented as (x_1, y_1), (x_2, y_2), (x_3, y_3) ….. (x_l, y_l). Now we define two matrices *M1* and *M2* as follows:

$$M1 = \begin{bmatrix} x_{1k} & x_{2k} & x_{3k} \\ y_{1k} & y_{2k} & y_{3k} \\ 1 & 1 & 1 \end{bmatrix}, M2 = \begin{bmatrix} x_1 & x_2 & x_3 & x_l \\ y_1 & y_2 & y_3 & \ldots y_l \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad (1)$$

As we have a set of sparse disparities already calculated, we know the coordinates of the borders of this warping zone on the source left and right images.

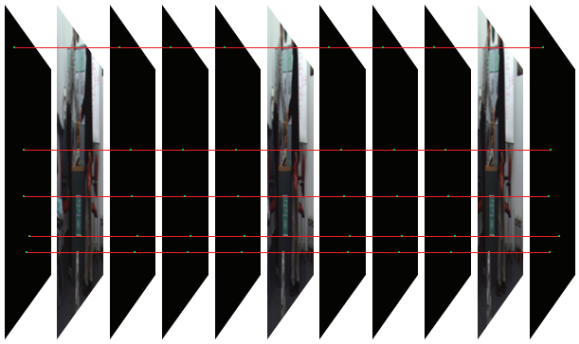


Fig. 5. Feature transformation to all the intermediate views.

These are represented by two separate matrices: *M1_L* & *M1_R*, respectively as below.

$$M1\_L = \begin{bmatrix} x_{1k\_l} & x_{2k\_l} & x_{3k\_l} \\ y_{1k\_l} & y_{2k\_l} & y_{3k\_l} \\ 1 & 1 & 1 \end{bmatrix} \quad (2)$$

$$M1\_R = \begin{bmatrix} x_{1k\_r} & x_{2k\_r} & x_{3k\_r} \\ y_{1k\_r} & y_{2k\_r} & y_{3k\_r} \\ 1 & 1 & 1 \end{bmatrix} \quad (3)$$

(x_{1k_l}, y_{1k_l}), (x_{2k_l}, y_{2k_l}) & (x_{3k_l}, y_{3k_l}) are the borders of the warping zone on the source left image (in this case I_1) and similarly (x_{1k_r}, y_{1k_r}), (x_{2k_r}, y_{2k_r}) & (x_{3k_r}, y_{3k_r}) are the borders of the warping zone on the source right image (in this case I_2). To identify the candidate pixels on the source left and right images, we solve the following equations

$$M2\_L = (M1\_L \times M1^{-1}) \times M2 \quad (4)$$

$$M2\_R = (M1\_R \times M1^{-1}) \times M2 \quad (5)$$

The matrices *M2_L* and *M2_R* are of same dimension as *M2* with the first two rows containing the *x* and *y* coordinates of the target pixels on the left and right images respectively.

C. Warping and blending

From the warp calculation stage, we have target pixel locations for each warping zone on an intermediate image. As the warp calculation process involves finding the matrix inverse, we may have the target pixel location as floating point values. The pixel values at these floating point locations are interpolated by fitting a bicubic surface to the data points on source left and right images.

The target pixels obtained from source left/right images are blended individually in to the warping zones. Let P_1 and P_2 be the target pixels on the source images I_k and I_{k+1} for a pixel P_{in} on an intermediate view at I_{k+frac}. Then P_{in} can be computed as:

$$P_{in} = (1 - frac) \times P_1 + frac \times P_2 \quad (5)$$

IV. Experimental Results

The performance of the algorithm is evaluated by considering different test image sequences obtained using the experimental settings defined in MPEG. As an input, we considered three equally spaced views and the generated views are Along with these input images, a set of three

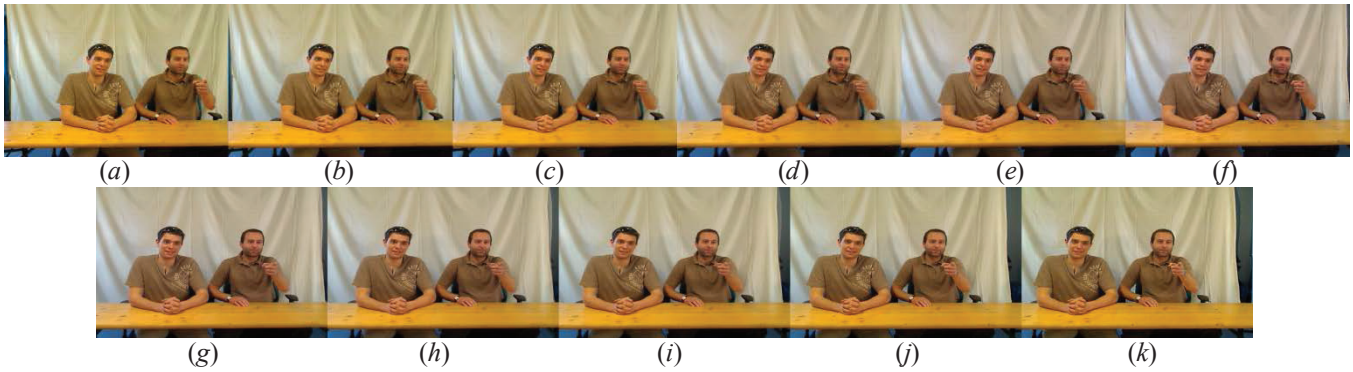


Fig. 6. Output Views of sequence Tele Conference.

Note: Images (*b*), (*f*) & (*j*) are the source input images. Images (*c*), (*d*), (*e*), (*g*), (*h*) & (*i*) are a set of interpolated images and images (*a*) and (*k*) are extrapolated.

views are locally captured using a three horizontally aligned camera rig and synthesized views corresponding to these images are also presented. Because of the spacing constraint, the generated views from a locally generated images are shown in Fig. 6.

A. Limitations and future work

In many cases, the artifacts resulting from the method are in the form of ghosting or blurring in the synthetic views as illustrated in Fig. 7. These are due to the lack of a sufficient number of correspondence points in these areas. Another reason for these artifacts is the inaccuracy when matching the sparse disparities. These limitations can be handled in a better way, by obtaining a different set of sparse disparities from a different algorithm (e.g. optical flow) and populating the existing disparities with the new disparities. However compared to DIBR, the algorithm is robust and allows us to generate images corresponding to very large FOV with limited artifacts without requiring any post processing.

V. Conclusions

We presented a reliable, fast and automatic method to create content for LFD. We followed a warping driven approach which relies on the sparse disparities and constructs a mesh. A set of appropriate warping functions are formulated and solved for each region inside the mesh The method can generate good quality intermediate views and also support extrapolation. With simple warping functions, the method greatly reduces the complexity in the multiview content creation process. As the warping based approaches are continuous, the method will not introduce any holes in the synthesized views thus further reducing the complexity associated with the post processing steps and thus the algorithm can be a potential alternative to DIBR.

VI. Acknowledgements

The research leading to these results has received funding from the DIVA Marie Curie Action of the People programme of the European Union's Seventh Framework Programme FP7/2007- 2013/ under REA grant agreement 290227.

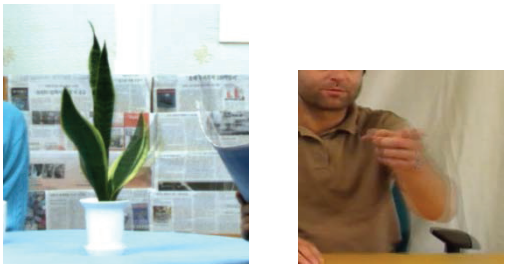


Fig. 7. Blurring artifacts.

VII. References

[1] M. Farre, O. Wang, M. Lang, N. Stefanoski, A. Hornung, and A. Smolic. Automatic content creation for multiview autostereoscopic displays using image domain warping. In *Proc. ICME*, 2011, pp.1-6.

[2] L.P Chew. Constrained Delaunay triangulations. In *Proceedings of the Third Annual Symposium on Computational Geometry*, 1987, pp. 215-222.

[3] A. Smolic, K. Muller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand. Intermediate view interpolation based on multiview video plus depth for advanced 3d video systems. In *Proc. ICIP 2008, IEEE International Conference on Image Processing*, pp. 2448–2451. IEEE, 2008.

[4] Wang, Daolei and Lim, Kah Bin. Obtaining depth map from segment-based stereo matching using graph cuts. *J. Vis. Comun. Image Represent,* 1047-3203, 2011, Vol. 22, pp.325-331.

[5] T. Balogh, P. T. Kovacs and Z. Megyesi. HoloVizio 3D display system. In *Proceedings of the First International Conference on Immersive Telecommunications*, ser. ImmersCom '07. ICST, Brussels, Belgium, 2007, pp. 19:1–19:5.

Examination of bacterial mutants in open and closed models

Dóra Bihary
(Supervisor: Dr. Sándor Pongor)
bihary.dora@itk.ppke.hu

***Abstract*—Bacteria use a mechanism called quorum sensing for inter- and intraspecies communication. This is a concentration based phenomena: bacteria emit chemical compounds and they respond to its above threshold concentration. In this paper I summarize results obtained with open and closed bacterial models. In these two cases we examined quorum sensing cheaters. These cheaters are mutant forms of the original wild type species that do not release as much chemical compounds - they do not take as big part in communication and cooperation - as wild type species do. This way they are not able to perform a swarming population on their own but in a population where wild type species can be found as well they can take part in the swarming of the other species and overgrow them because of their lower energy consumption.**

***Keywords*-quorum sensing; wild type; signal negative; signal blind; closed model; open model**

I. Introduction

Bacterial (or other similar) models can be classified in many ways [1]. We can classify models according to how they represent bacteria, space, medium or bacterial behavior. In the next paragraphs I will give a short overview of these classification methods.

In the simplest case bacteria can be represented as a continuous mass that grows and diffuses in space [2]. These models are described with reaction-diffusion equations. In these models the individual representation of bacteria disappears. This is the reason why we usually represent bacteria as agents - as interacting entities. These interactions can be potential-based or rule-based. Potentials, like Lennard-Jones potential are frequently used in such models [3], they make a potential field for each bacteria. The movement in the next step is based on the distance from the surrounding agents. It is violated to go too close to each other and we can define an optimal distance between agents where they prefer to be. Rule-based models give rules that a certain agent can follow during movement (e.g. try to avoid crowded places, try to suit your direction to your mate's direction; try to suit your speed to your mate's speed, try to move toward your mates etc. [4]). These rules are usually sequentially evaluated beginning with the most important ones - like "do not collide with others".

The space that surrounds bacteria sometimes has no importance and so it is not represented in some models. In more complex cases we describe the space by coordinates. This can happen in 1, 2 or 3 dimensions depending on the actual model. In both cases we can talk about open [4] or closed [5] systems. A closed model has fix or periodic boundary conditions at its margins. In our work we compared closed and open models so these will be discussed in detail later.

Medium in our case means the rest of the model that is not the object of our work (e.g. all but bacteria). This medium sometimes may not be represented (vacuum), we can describe it at the level of physical forces or chemical particles [6]. In bacterial models the most commonly used representation is to describe medium as continuous mass where the participating materials can diffuse in space and time.

A bacterial colony consists of bacteria that try to achieve a common goal via a more or less common behavioral pattern. This way we can talk about the coordination of agents. This coordination can happen on several levels, we can for example coordinate the speed or direction of the movement. We can widen this concept to the inner states of agents or even to the whole genome [7].

II. Biological background

The communication mechanism of bacteria is called quorum sensing, it is based on the emission of chemical compounds [8] [9]. Bacteria secret a basal amount of signal that, in an open environment, continuously diffuses away. If there is a sufficient number of bacteria present in a small environment, the concentration of the emitted signal can raise and eventually reach a threshold concentration. The bacteria sense this threshold concentration and change their metabolism: they increase the production of signals, and start the production of factors. When in turn the concentration of factors reaches a threshold in the environment, the bacteria increase their movement, food intake and their division rate. They enter an active state - they start swarming.

Signal molecules are usually called as communication materials since their function is to sign for each other that they are present on the surface. On the other hand factor molecules are usually called as cooperation materials, or public goods. This is because these factors are chemical compounds that are not needed in a basal state, the production of them is energy consuming, but they are sufficient for swarming - e.g. siderophores, surfactants.

Our model organism was *Pseudomonas aeruginosa* an opportunistic pathogen that can potentially cause death in patients of cystic fibrosis.

In the simulations we examined three kinds of bacteria: wild type (WT), signal negative (SN) and signal blind (SB) [10]. The form of a bacteria that can perform all the above

D. Bihary, "Examination of bacterial mutants in open and closed models,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 13-16.

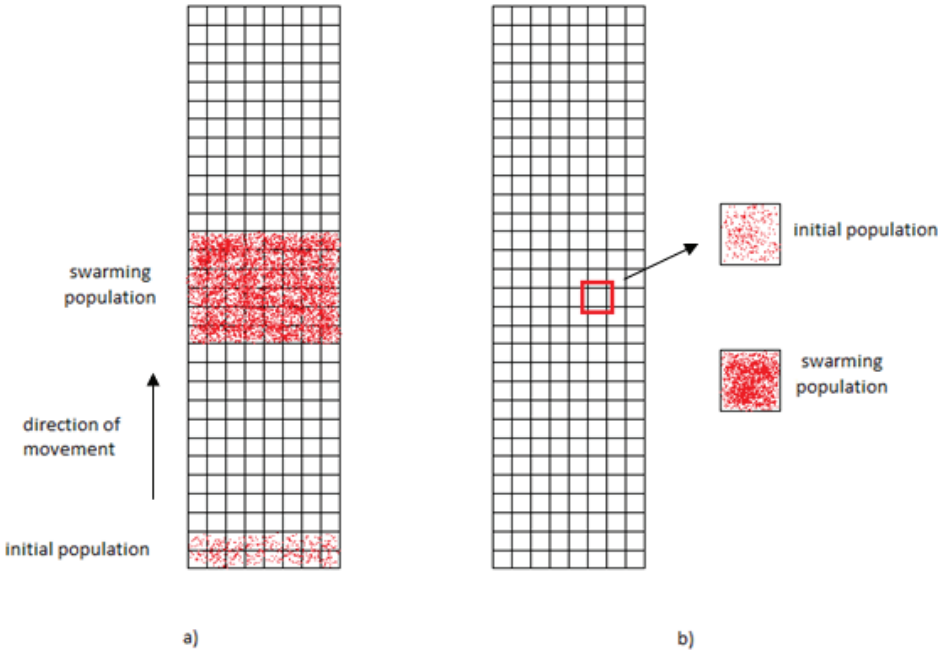


Fig. 1. In our open model at the beginning of the simulation bacteria are at the bottom of surface, during simulation they go upwards (a); the closed model can be imagined as one single cell of the open model (b).

mentioned aspects of quorum sensing is called WT. SN is a mutant type where species do not take part in the production of signal molecules however they can react for outer signal concentration (for example consumed by WT species) and as an effect they can switch to an active state and consume factor molecules. SB mutants consume signal in a basal amount, but do never react to signal concentration. They never consume factors, but in the presence of enough factors they can switch to a swarming state and enjoy the advantages of the swarming. These two mutant species are sometimes called as cheaters because they swarm together with other species without properly taking part in the creation of circumstances that make it possible to swarm.

III. Model environment

Based on the above mentioned categories we can classify our model as well. Our goal was to represent bacterial colonies in a way that best fits the bacteria observed in nature. In order to this our bacteria are individual agents, not a continuous mass, the interactions between bacteria are represented by a rule-based solution (we recently started simulations where bacteria interact with Lennard-Jones potentials). The medium in our model is a continuous mass where the participating materials (food, signal and factor) can diffuse in space and time. Based on the quorum sensing concept we can say that the behavior of bacteria is represented in our model in a way that the inner state of bacteria is locally synchronized. This means that at a given space all bacteria (belonging to the same species) will have the same state, therefore the same speed, same energy consumption, they will produce the same amount of signal and factor, etc.

The representation of space will be discussed in detail in the next section.

IV. Representation of space

Our goal was to compare open and closed systems which means that the space was represented in two different ways during the simulations.

A. Open model

A model is open if there is no limitation in cell growth, at least in one direction we do never let to reach bacteria the end of the space. This is not an uncommon representation in biological experiments bacteria are not always let to run as long as they can reach the sides of petri dishes either.

Our original model was an open model where the space was represented as an infinite long rectangle (see Fig. 1, (a)). This surface is represented as a cell structure. One certain cell can only carry a certain number of bacteria, and inside of it the food, signal and factor concentration values are constant. Between these cells the chemical compounds can diffuse. Bacteria are placed at the beginning of each simulation at the bottom of this surface (first 4 row) and during the simulation they move upwards (where new nutrient sources can be found). At the sides of this rectangular there is a periodic boundary which means that for example if one would ”step down” from the surface in the right direction, it will appear at the left side of the rectangle. Bacteria move randomly, their movement is

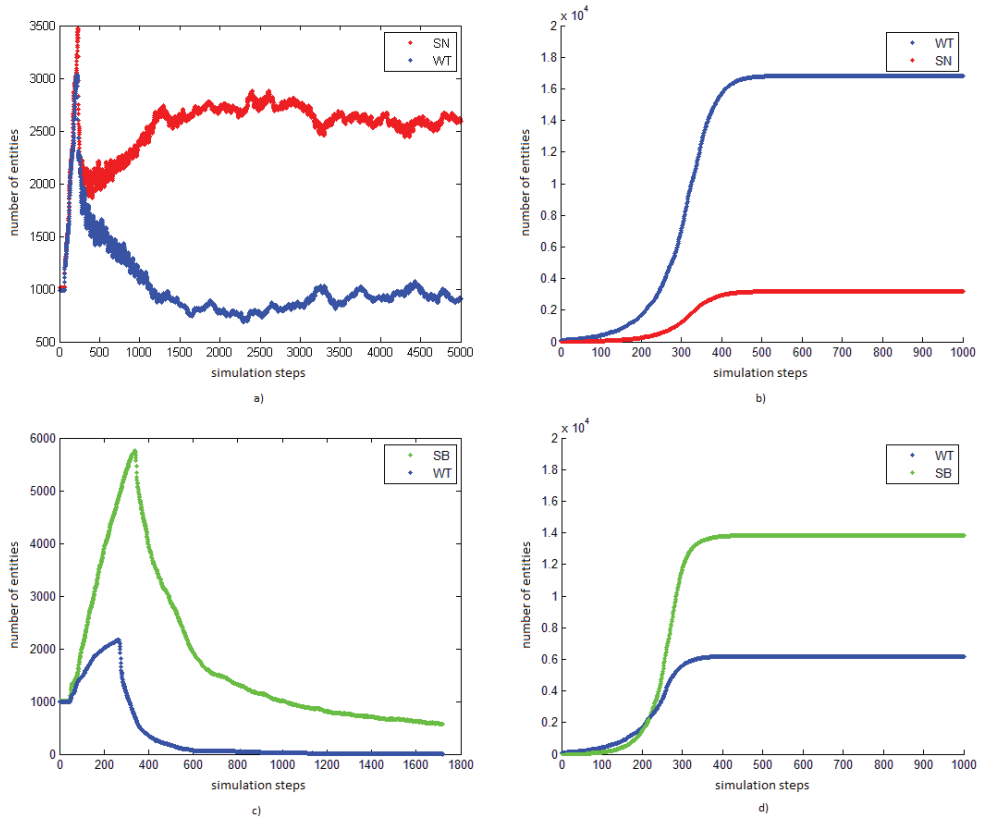


Fig. 2. Simulation results with wild type and signal negative bacteria in open (a) and closed (b) environment and wild type and signal blind bacteria in open (c) and closed (d) environment; x axes denote time step y axes the actual population sizes. Blue curve belongs to WT, red curve to SN and green curve to SB bacteria.

limited only by the food - those who would go down where no more food remained will die in time. In order to obtain bacteria close to each other like in a natural colony we introduced the so called border advancement that does not let bacteria discover a new cell as long as enough number of bacteria tried to move into it.

Bacteria do the following things in each simulation step: they produce signal and factor, according to the new concentrations they get a new state, they eat, divide and move.

B. Closed model

We call a model closed when we let the bacteria reach the ends of the surface in each direction during the simulation. In this case we always have to define what happens at the boundaries (boundary conditions).

We simplified the above described open model to be able to simulate closed systems. In the closed model we use the same principles but the lattice consists only of one cell (see Fig. 1, (b)). We implemented this cell with periodic boundary condition on all its sides. This means a well mixed consortium where there is no need for diffusion or movement - hence all parameters are the same in a certain cell. We introduced infinite amount of food, which in contrast with the open model means that hypothetically the population can grow infinitely. To avoid this we maximized the total number of bacteria in a way that the bigger the population, the smaller the reproduction rate will become. This results a population of fixed (e.g. 20 000) number of bacteria. If we give finite amount of food into a closed space the result is quite predictable - when the entire nutrient disappears from the system in a few steps all bacteria will die out.

Bacteria do the following things in each simulation step: they produce signal and factor, according to the new concentrations they get a new state, they eat and divide.

V. Results

We examined cases where the initial population consisted of WT and SN or WT and SB species. For both cases we compared the open and the closed model.

In a mixed consortium that consists of WT and SN bacteria, SN species can swarm together with the WT ones, but they have a small metabolic advantage against them hence they do not take part in signal consumption. The usual expectation is that SN species - due to their less energy consumption - overgrow WT ones but they do let them swarm with a small amount of species because the signal and factor consumed by WT bacteria is essential for SN species to remain in swarming state. As you can see in Figure 2 (a) and (b) in the open case this holds however in the closed system the SN mutants are not able to overgrow WT species. This happens because

species do not have to compete for nutrient in this model, which means that a mutant species must have more metabolic advantage instead of the WTs than in an open model in order to produce a bigger population. Sufficiently lowering the energy consumption of SN species in the swarming state they could overgrow WT species (data not shown).

SB mutants have bigger metabolic advantage against WT species than SN ones because although they do consume signals in a basal amount but they never take part in factor consumption, and it is more costly to emit factors than signals. This results in an open population where WT and SB species are present that SB species overgrow WT ones in a way that results the extinction of WT species. This is a greedy approach because hence SB mutants can not hold on the swarming state on their own the big population they could grow with the help of WT species will shrink without them. In closed systems SB mutants still overgrow WT ones however without a real competition for food both of the species will survive. In Figure 2 (c) and (d) you can see this phenomena for open and closed systems.

In our closed system the signal, factor and food rate is the same in the whole space. This means that after a certain number of steps - when bacteria could perform enough amount of signal and factor - all bacteria are in swarming state. This way the only parameter that defines the ratio between WT and mutant species at the end of the simulation is the energy consumption of mutant species in swarming state. As we can see the parameter value that barely lets mutants overgrow WT species is somewhere between the energy consumption values of SN and SB species that were sufficient in the open system.

VI. Future plan

Our future plan is to compare the results obtained with the simulation environment with real biological results. We will build our work on two group's experiments. An open space is evaluated in the work of Netotea et. al. [10] and a closed one in the work of Diggle et. al. [11] Based on these results we will validate our simulation model.

Acknowledgment

This project was developed within the PhD program of the Multidisciplinary Doctoral School, Faculty of Information Technology, Pázmány Péter Catholic University, Budapest. Thanks are due to my supervisor, Prof. Sándor Pongor and to Ádám Kerényi, who is the major developer of the agent-based simulation program.

References

[1] Leon Sterling Taveter and Kuldar. *The Art of Agent-Oriented Modeling*. The MIT Press, 2009.

[2] Irene Giardina. Collective behavior in animal groups: theoretical models and empirical studies. *HFSP journal*, 2(4):205–19, August 2008.

[3] A Mogilner, L Edelstein-Keshet, L Bent, and A Spiros. Mutual interactions, potentials, and individual distance in a social aggregation. *Journal of mathematical biology*, 47(4):353–89, October 2003.

[4] C W Reynolds. Flocks, herds and schools: a distributed behavioral model. In *Computer Graphics*, volume 21, pages 25–34. ACM, 1987.

[5] K Kawasaki, A Mochizuki, M Matsushita, T Umeda, and N Shigesada. Modeling spatio-temporal patterns generated by Bacillus subtilis. *J Theor Biol*, 188(2):177–185, 1997.

[6] Josephine R Chandler, Silja Heilmann, John E Mittler, and E Peter Greenberg. Acyl-homoserine lactone-dependent eavesdropping promotes competition in a laboratory co-culture model. *The ISME journal*, 6(12):2219–28, December 2012.

[7] Martin Thanbichler. Synchronization of chromosome dynamics and cell division in bacteria. *Cold Spring Harbor perspectives in biology*, 2(1):a000331, January 2010.

[8] M B Miller and B L Bassler. Quorum sensing in bacteria. *Annu Rev Microbiol*, 55:165–199, 2001.

[9] V Venturi, A Kerenyi, B Reiz, D Bihary, and S Pongor. Locality versus globality in bacterial signalling: can local communication stabilize bacterial communities? *Biology Direct*, 5:30, 2010.

[10] S Netotea, I Bertani, L Steindler, V Venturi, S Pongor, and A Kerenyi. A simple model for the early events of quorum sensing in Pseudomonas aeruginosa: modeling bacterial swarming as the movement of an "activation zone". *Biol Direct*, 4:6, 2008.

[11] S P Diggle, A S Griffin, G S Campbell, and S A West. Cooperation and conflict in quorum-sensing bacterial populations. *Nature*, 450(7168):411–414, 2007.

Myoelectric signal analysis using an embedded SoC

Bence József Borbély
(Supervisor: Dr. Péter Szolgay)
borbely.bence@itk.ppke.hu

***Abstract*—An implementation for the analysis of human myoelectric signals (MES) is presented. Offline recorded multi-channel signals of forearm muscles are processed with an embedded SoC having field programmable on-chip modules in order to classify different movement patterns to control human-assisting electromechanical systems with multiple degrees of freedom (e.g. a prosthetic hand). Benchmark results of an ANSI C implementation are shown to assess the raw performance of the used ARM cores of the SoC. Possible computational bottlenecks are located based on these results and suggestions on custom hardware implementations are made to fully utilize the flexibility and performance of the used hardware platform.**

***Keywords*—EMG processing; embedded system; field programmable**

I. Introduction

Electric signals measured at different skin surface locations carry important features of particular biological subsystems. Two prominent examples are current state of health monitoring by analysing the electrocardiogram (ECG) signal measured from the chest surface or the steady-state cognitive concentration on some event characterized by the electroencephalogram (EEG) recorded from the surface of the head. An other, very important type is the mioelectric signal (MES) which can be measured from the covering skin of muscles. The MES directly reflects the summed motor unit activity, thus it can be related to muscle contraction and exerted force (however these relationships are highly non-linear in most cases). The importance of this signal type is that it can be used to analyse movement patterns at the muscle activation level, or tell specific movement intents even in the case when the actuated end-effector is absent from the system - like in the case of hand amputations where the hand itself is missing but the muscles responsible for main wrist and finger movements are still present in the forearm.

This study focuses on the processing and classification of MES data to utilize the flexibility and performance of an embedded platform in a test environment for prosthesis control. As a prototype system a widely recognized pattern recognition scheme was implemented to process four to eight forearm MES channels using time-domain signal features and an LDA classifier.

II. Methods

A. The pattern recognition method

The idea behind the standard pattern recognition based myoelectric control is to measure signals from multiple channels during different predefined isometric contractions of muscles (different states) and store specific features of these recordings as separate state descriptors (offline, supervised learning). After this stage an online stream of data from the same recording sites can be obtained and classified to categorize the actual signals into one of the trained classes. MES data is non-stationary and stochastic in nature therefore most of the related analyses apply processing windows to extract descriptive features of the signal. In the current implementation a 150 ms long processing window was used because it has been shown that this length enables optimal performance for this type of classifiers [1].

The spatial selectivity (the number of separable movement classes) in the system is highly determined by the number of separate recording channels. Previous studies justified that in the case of lower arm recordings four channels of MES are suitable to classify online measured data into one of six separate classes with high efficiency [2]. Based on these results a four-channel system was implemented as the basis of the test environment, but for testing reasons it was extended to have five, six and eight virtual channels to estimate performance in more complex recording environments. Because we had only four real channel recordings, six possible output classes were used in every case.

In real prosthetic applications overall latency and response time are critical factors of device acceptance which are determined by the processing window length and the amount of processing window shift (or sampling delay) during operation. Among these two factors window shift value can be varied to obtain different temporal resolutions, resulting that shorter shifts yield better response times at the cost of computational overhead.

1) Signal features: To characterize signal windows and to reduce data dimension the standard four element time-domain feature (TDF) set was calculated for each data window (150 ms) and channel in the performed simulations. These features were the mean absolute value (MAV), number of zero crossings (NZC), number of slope sign changes (NSSC) and the waveform length (WL) as described in previous studies [3], [4]. It is important to note that these features give only estimations of specific signal properties (e.g. NZC $\sim$ frequency) but it has been shown that they provide as good basis as frequency-domain features for classification of stationary signals for less computational cost and induce lower latency in the system [2].

2) LDA classifier: To partition the feature space into six subspaces (or classes) for pattern classification, linear discriminant analysis (LDA) was applied as desribed in [5]. The reason for using LDA is that it can reduce feature space dimensionality taking the separate subspaces into account. More specifically it finds those projection vectors in the complete feature space (in this case with dimension of (4 TDF $\times$ # channels)) which best separate the individual classes when the dataset is projected. After the projection vectors are calculated

B. J. Borbély, "Myoelectric signal analysis using an embedded SoC,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 17-20.

(# projection vectors $\ll$ # feature space dimensions) data points from the complete feature space are projected to get a more separable set of target classes having lower dimension (# projection vectors).

During online operation the actual recorded data is first transformed into the feature space (by calculating its time-domain features) followed by the projection to the same vectors obtained with the LDA algorithm. The classification takes place when these projected values are compared to the stored projections of the target classes and class labels are assigned to the data based on its distance (e.g. Euclidean) from the stored class values.

B. Recorded and simulated data

Four channels of MES were recorded (F_s =1 kHz, resolution: 16 bit) from one subject during six different isometric muscle contraction classes following the method described in [6]. The recordings were performed independently from the processing system. The recording electrodes were placed on the forearm above the wrist flexors and extensors and on each side of the forearm, roughly at middle length. Separate data sets were recorded to train and test the classifier (with average length of 25 s for each class). Testing was performed using an appended array of test recordings in pseudo-random order as the input stream. For simulation reasons the measured MES data were extended to have five, six and eight virtual channels using a perturbed version of the original recordings.

C. Main algorithm

The main steps of the practical realization as described previously are shown in Algorithm 1. It is important to note that this implementation is used for offline testing with previously measured training and test data, not for online streaming and processing of the input signals. However, the algorithmic design allows the extension of the system to have real-time functionality with only minor modifications.

Algorithm 1 Offline EMG classification

1: **procedure** EMGCLASSMAIN
2: // Calculate and store the time-domain features of the training dataset
3: *PreprocessTrainingData*($N_{channels}, WinShift$)
4: // Calculate and store the LDA projection vectors which best separates the training dataset
5: *TrainLDAClassifier*($preprocessedData$)
6: // Assign class labels to the test signal windows based on the the separated training dataset
7: *ClassifyTestData*($inputData, LDAdata$)
8: **end procedure**

The three main parts of the system were developed to allow easy separation of the main processing steps. *PreprocessTrainingData*($N_{channels}, WinShift$) calculates and stores all time-domain features of the training data based on the number of channels and the amount of processing window shift, decreasing the dimensionality at the first place. The second function, *TrainLDAClassifier*($preprocessedData$) calculates and stores the LDA projection vectors, which best separates the training dataset in the time-domain feature space. After these vectors are calculated, the training dataset is projected to reduce its dimension and prepare it for classification.

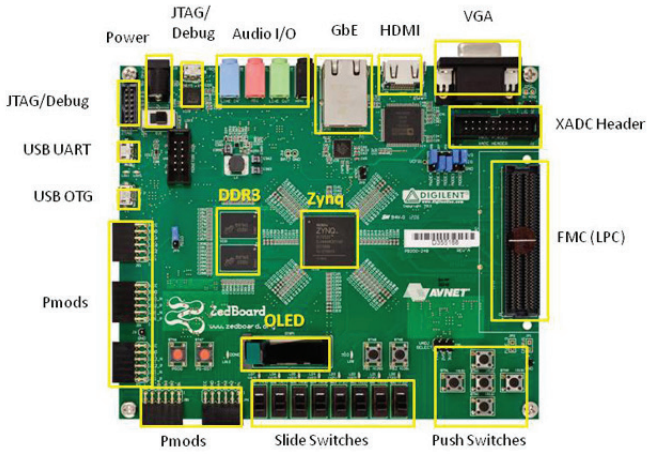


Fig. 1. The prototyping platform - the Digilent Zedboard

The last part, *ClassifyTestData*($inputData, LDAdata$) performs the classification of all input data window using LDA projection vectors calculated during classifier training and assigns the most probable class label to each of these windows.

III. Implementation

A. The Zynq-7000 platform

To implement the EMG processing system (Section II. A.) in hardware the Digilent Zedboard [7] was chosen (Figure 1), which is based on a Xilinx Zynq – 7020 SoC architecture [8]. The Zynq – 7000 family integrates the ARM Cortex-A9 dual core PS (Processing System) and the 28nm Xilinx Series-7 PL (Programmable Logic) fabric. The unique features of this system are the tight integration of the embedded microprocessor and the FPGA using standard AXI4 bus interfaces and the so-called processor centric approach, where the PS is initialized in the first step and the PL is configured in the second step during the startup sequence.

The processor system contains private L1 instruction and data caches for both cores that run at up to 667MHz clock frequency. In addition, 512KB L2 cache is shared between the cores. Time-critical data and code can be stored in the 256KB on chip SRAM memory, while larger data storage can be provided by using DDR2 or DDR3 memory components using the integrated external memory controller. The Processor System has wide variety of different I/O interfaces to connect the system to the outside world such as UART, I2C, SPI, USB Gigabit Ethernet to name a few.

The Programmable Logic, which is based on the Atrix family, is connected to the PS via several AXI4 interconnects; four 32bit wide interfaces are dedicated to low latency access to the registers of the peripherals implemented in the PL. Four 64bit wide high performance AXI4 buses are available for fast transfer of large amounts of data between the PL and the different memories. For tightly integrated coprocessors, which should share data with the software part running on the PS,

a specialized 64-bit wide coherent AXI4 bus connected to the snoop protocol of the L2 cache is also available.

B. ANSI C implementation

The algorithm outlined in Section II. C. was implemented in ANSI C on a laptop computer having an Intel Core i5-540M CPU running at 2.53 GHz. The extracted time-domain features were $MAV, NZC, NSSC$ and WL. Self-written implementations were used for all numerical calculation methods. The development system was running Ubuntu Linux 12.04 LTS operating system and the *gcc* compiler [9] was used to generate executables. To compile the sources to the ARM cores of the Zynq processor, gcc's cross compiler version (*arm-linux-gnueabi-gcc*) was used with the same options as the desktop version. For optimal performance the *-O3* compiler option was used in both situations.

1) Time-domain feature extraction: As the basis of pattern recognition the time-domain properties of the multi-channel MES signal are calculated by the *PreprocessTrainingData*($N_{channels}, WinShift$) function, as described in Algorithm 2.

Algorithm 2 Calculation of time-domain features

```
1: function PREPROCESSTRAININGDATA (N_channels, WinShift)
2:     for all sample windows determined by WinLength and WinShift do
3:         calculate and store MAV;
4:         calculate and store NZC;
5:         calculate and store NSSC;
6:         calculate and store WL;
7:     end for
8: end function
```

This function enables the characterization of the actual MES sample window with less data than the window has itself. The calculated signal features are:

- MAV: the mean of the summed absolute values within the processing window, giving an estimate of signal power
- NZC: the number of zero crossings having a magnitude difference between two consecutive samples exceeding a predefined threshold value within the processing window, giving a rough estimate of signal frequency
- $NSSC$: the number of slope sign changes having a magnitude difference between three consecutive samples exceeding a predefined threshold value within the processing window, giving an other measure of signal frequency
- WL: the total length of the waveform within the processing window, providing information on waveform complexity

2) Calculation of LDA projection vectors: Linear Discriminant Analysis is a statistical method for dimension reduction mostly used in classification problems. The actual algorithm involves covariance an inverse matrix calculations, followed by determining the eigenvalues and eigenvectors of a linear system, as described is Algorithm 3.

Algorithm 3 Calculation of LDA projection vectors

```
1: function TRAINLDACLASSIFIER(preprocessedData)
2:     calculate the means and sample covariances of all of the training data sets separately;
3:     calculate the between-class scatter matrix for all classes (S_b);
4:     calculate the within-class scatter matrix for all classes (S_w);
5:     estimate the eigenvalues of the linear system determined by (S_w^-1 S_b);
6:     calculate the exact eigenvectors from the estimated eigenvalues ⇒ LDA projection vectors
7: end function
```

The sample covariances are calculated as

$$\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T,$$

where $\mathbf{C}$ is the covariance matrix, $\mathbf{x_i}$-s are the observation vectors and $\bar{\mathbf{x}}$ is the mean of the observation vectors. Inverse matrix calculation was implemented using the Gauss-Jordan elimination method. Because of performance considerations, the eigenvalues of the linear system (calculated from the within-class and between-class scatter matrices) was only estimated by the QR Iteration algorithm applying 10 iterations only (Algorithm 4).

Algorithm 4 QR Iteration algorithm

```
1: procedure QRITERATION
2:     // for a given matrix A
3:     A_0 = A;
4:     // iterate finite number of steps
5:     for k = 1, 2, ..., 10 do
6:         // compute the QR decomposition of A_{k-1}
7:         Q_k R_k = A_{k-1};
8:         // calculate the new A_k and iterate
9:         A_k = R_k Q_k;
10:    end for
11: end procedure
```

The method results an almost upper-triangular matrix that's main diagonal contains the estimated eigenvalues. After sorting these values in a descending order they can be used to determine the exact eigenvectors of the system (which are the LDA projection vectors) using the Inverse Iteration algorithm (Algorithm 5).

3) Classification of the simulated test data: *ClassifyTestData*($inputData, LDAdata$) performs class label assignment based on time-domain feature extraction and the calculated LDA projection vectors to all of the simulated

Algorithm 5 Inverse Iteration algorithm

procedure INVERSEITERATION
  // for a given matrix A and and approximated eigenvalue μ, initialize $\mathbf{b_0}$ as a random vector
  $\mathbf{b_0} = rand(length(\mathbf{A}), 1)$;
  // iterate until $\mathbf{b_k}$ converges
  for $k = 1, 2, \ldots$ **do**
    // determine the actual normalization factor
    $C_k = \|(\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{b_k}\|$;
    // calculate the new $\mathbf{b_k}$ and iterate
    $\mathbf{b_{k+1}} = \frac{(\mathbf{A}-\mu\mathbf{I})^{-1}\mathbf{b_k}}{C_k}$;
  end for
end procedure

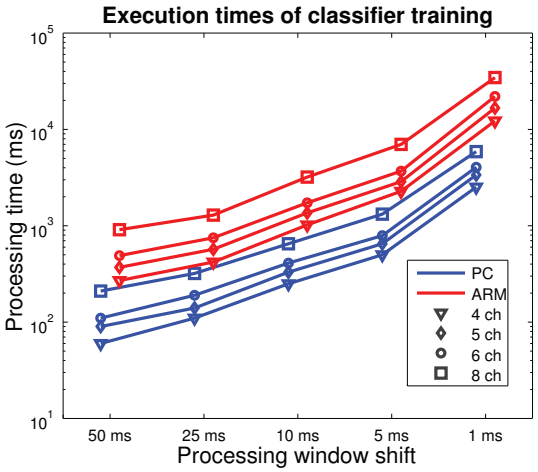


Fig. 2. Benchmark results - training

test data described in Section II. B. This classification process is performed offline for simulation reasons, but can be extended to have real-time functionality in the future.

IV. Results and Future Work

Benchmark tests with various channel numbers and processing window shift values were performed on the development PC described in Section III. 2. and on the Zedboard itself. 20 different test conditions were analysed using 4 different channel numbers (4, 5, 6, 8) and 5 different processing window shift values (50 ms, 25 ms, 10 ms, 5 ms, 1 ms) in all possible combinations. As it can be seen in Figure 2, processing times increased exponentially across window shift values and the number of analysed channels (note the logarithmic scale of the figure). In addition, approximately one order of magnitude difference can be seen between the PC and ARM running times. The results show that the ARM cores of the Zynq SoC are suitable for standard classifier training (having 50-25 ms shifts between processing windows up to 8 channels), but have performance drop-downs at more frequent analysis steps (e.g. training the classifier with 1 ms window shift yields more than 100 s average training time).

On the other hand, analysis of data window classification running times showed that however execution times still differ approximately with an order of magnitude, the raw performance of the ARM cores are suitable for real-time

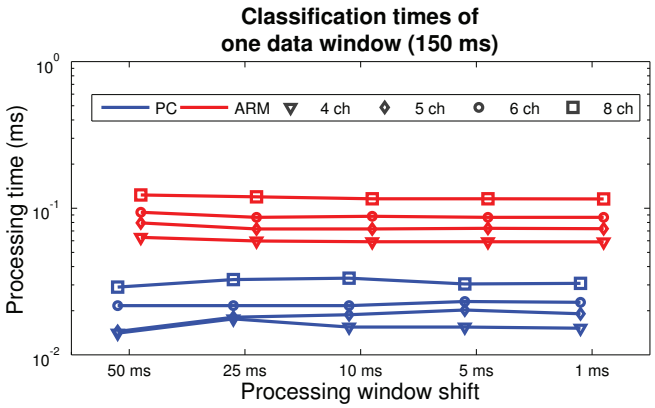


Fig. 3. Benchmark results - classification performance

data classification based on the applied labelling method (the average classification time is about 0.1 ms compared to the 1 ms inter-sample time of EMG recordings).

From the above data it looks like that the main bottleneck of the built-in ARM cores is the relatively low performance during classifier training, involving data- and computation-intensive operations. The solution to this problem can be to design custom hardware elements responsible for fast vector and matrix operations and implement them in the SoC's FPGA fabric. The final system that uses both the ARM cores and the custom hardware elements would allow real-time measurement and processing of biosignals in a high performance embedded environment with low power consumption.

Acknowledgment

This research project was supported by the OTKA Grant No. K84267. The support of the grants TÁMOP-4.2.1.B-11/2/KMR-2011-0002 and TÁMOP-4.2.2/B-10/1-2010-0014 is gratefully acknowledged.

References

[1] L. H. Smith, L. J. Hargrove, B. a. Lock, and T. a. Kuiken, "Determining the optimal window length for pattern recognition-based myoelectric control: balancing the competing effects of classification error and controller delay." *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 19, no. 2, pp. 186–92, Apr. 2011.

[2] L. J. Hargrove, K. Englehart, and B. Hudgins, "A comparison of surface and intramuscular myoelectric signal classification." *IEEE Trans. Biomed. Eng.*, vol. 54, no. 5, pp. 847–53, May 2007.

[3] B. Hudgins, P. Parker, and R. N. Scott, "A new strategy for multifunction myoelectric control." *IEEE Trans. Biomed. Eng.*, vol. 40, no. 1, pp. 82–94, Jan. 1993.

[4] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control." *IEEE Trans. Biomed. Eng.*, vol. 50, no. 7, pp. 848–54, Jul. 2003.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[6] B. Borbély, "Design and simulation of a processing system for myoelectric data," in Hungarian, Pázmány Péter Catholic University, 2012.

[7] "Digilent Zedboard (official webpage - 2013)." [Online]. Available: http://www.zedboard.org

[8] "Xilinx official webpage (2013)." [Online]. Available: http://www.xilinx.com

[9] "GCC, the GNU Compiler Collection." [Online]. Available: http://gcc.gnu.org

Subcellular localization of the components of the nitric oxide system in the hypothalamic paraventricular nucleus of mice

Erzsébet Farkas
(Supervisor: Csaba Fekete)
farkas.erzsebet@koki.mta.hu

Abstract—**Nitric oxide (NO) is a gaseous transmitter. In the hypothalamic paraventricular nucleus (PVN), it has been implicated in the regulation of energy homeostasis and neuroendocrine control. However, little information is known about the subcellular localization of the components of the NO system in the PVN and whether NO is utilized as an anterograde and/or retrograde messenger by the parvocellular neurons of this nucleus. Neuronal nitric oxide synthase (nNOS) is the enzyme responsible for the NO production of neurons, and the primary mechanism mediating the effects of NO on target neurons is the soluble guanylate cyclase-facilitated production of cGMP. Soluble guanylate cyclase is a heterodimer composed of α (α1 and α2) and β (β1 and β2) subunits, the most prevalent form being the α1/β1 heterodimer. Using antisera against nNOS and the soluble guanylate cyclase α1 and β1 subunits, immuno-electron microscopy was performed to determine the subcellular localization of these proteins in the parvocellular part of the PVN in mice. nNOS was abundantly present in neuronal perikarya and dendrites and also in axon varicosities. In perikarya and dendrites, nNOS-immunoreactivity was widely distributed in the cytoplasm, primarily associated with the endoplasmatic reticulum. nNOS-immunoreacivity was also found to be associated with the perikaryal plasma membrane in close proximity to both symmetric and asymmetric synapses, as well as within axon varicosities forming both symmetric and asymmetric synapses. The soluble guanylate cyclase α1 subunit was found in dendrites and axon varicosities and associated with both the pre- and postsynaptic densities of the synapses. The α1 subunit was associated with both symmetric and asymmetric types of synapses, whereas the β1 subunit was primarily observed in dendrites and frequently associated with the postsynaptic density of synapses. On rare occasions when the β1 subunit was observed in axon varicosities, the immunoreactive varicosities formed symmetric type synapses.**

In summary, these data indicate that nitric oxide may be utilized as both an anterograde and retrograde transmitter in the parvocellular part of the PVN.

Keywords-*nitric oxide; paraventricular nucleus; electron microscopy*

I. Introduction

Nitric oxide (NO) is a gaseous transmitter that has extremely short half-life in biological systems. It is synthesized by the nitric oxide synthase (NOS) enzymes from the amino acid L-arginine. There are three NOS isoforms: endothelial NOS (eNOS), inducible NOS (iNOS) and neuronal NOS (nNOS). All are present in the nervous system, however, nNOS is the principal isoform that is utilized by neurons. NO primarily exerts its effect through the soluble guanylyl cyclase enzyme (sGC). Soluble GC is a heterodimer molecule. Four sGC subunits, α1, α2, ß1 and ß2, have been identified. The most common form in mammalian tissues is the α1/ß1 heterodimer. The activity of sGC results in an increase of intracellular cyclic GMP (cGMP) levels. This way, NO causes accumulation of cGMP in cells leading to activation of multiple downstream targets, including kinases, ion channels and phosphodiesterases. NO can be utilized as both anterograde and retrograde transmitter.

The hypothalamic paraventricular nucleus (PVN) has been implicated in the regulation of energy homeostasis and the neuroendocrine systems. NO has been shown to play important role in the regulation of these functions including the regulation of food intake, TRH gene expression and CRH release, suggesting that NO may influence the parvocellular neurons of the PVN. However, little information is available about the localization and function of this transmitter system in the PVN.

Therefore, we performed ultrastructural studies to determine the precise localization of the components of the NO transmitter system in the parvocellular part of the PVN and to understand whether NO can be utilized as an anterograde and/or retrograde transmitter in this nucleus.

E. Farkas, "Subcellular localization of the components of the nitric oxide system in the hypothalamic paraventricular nucleus of mice," in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds. Faculty of Information Technology, Pázmány Péter Catholic University. Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 21-24.

II. Material and Methodes

The experiments were carried out on ten adult, male, CD1 mice, weighing 30–35 g, housed under standard environmental conditions (light between 06:00 and 18:00 h, temperature 22±1 ^{o}C, mouse chow and water *ad libitum*). All experimental protocols were reviewed and approved by the Animal Welfare Committee at the Institute of Experimental Medicine of the Hungarian Academy of Sciences.

Animals were deeply anesthetized with ketamine/xylazine (ketamine 50 mg/kg, xylazine 10 mg/kg body weight, ip). Five minutes later, the animals were perfused transcardially with 10 ml 0.01 M phosphate-buffered saline (PBS), pH 7.4, followed sequentially by 10 ml of 4% paraformaldehyde in Na-acetate buffer, pH 6.0, and then by 50 ml of 4% paraformaldehyde in Borax buffer, pH 8.5. The brains were rapidly removed and stored in 4% paraformaldehyde in 0.1 M phosphate buffer (PB), pH 7.4, for 24 h at 4 ºC.

Serial, 25µm thick coronal sections were cut on a Leica VT 1000S vibratome (Leica Microsystems, Wetzlar, Germany). The sections were treated with 0.5% H_2O_2 in PBS for 15 min. The sections were cryoprotected in 15% sucrose in PBS for 15 min at room temperature and in 30% sucrose in PBS overnight at 4^{o}C and then, quickly frozen over liquid nitrogen to improve antibody penetration into the tissue. To detect the nNOS-immunoreactivity, pretreated sections were then placed into rabbit anti-nNOS serum (1:200) for 4 days at 4 ^{o}C and after rinsing in PBS and 0.1% cold water fish gelatin/1% bovine serum albumin (BSA) in PBS, incubated in donkey anti-rabbit IgG conjugated with 0.8 nm colloidal gold (Electron Microscopy Sciences, Fort Washington, PA) diluted at 1:100 in PBS containing 0.1% cold water fish gelatin and 1% BSA. After rinsing in 0.2 M sodium citrate, pH 7.5, the gold particles were silver intensified with the Aurion R-Gent SE-LM Kit. (Amersham–Pharmacia Biotech UK, Buckinghamshire, UK). Sections were osmicated, and then treated with 2% uranyl acetate in 70% ethanol for 30 min. Following dehydration in an ascending series of ethanol and propylene oxide, the sections were flat embedded in Durcupan ACM epoxy resin (Fluka) on liquid release agent (Electron Microscopy Sciences)-coated slides, and polymerized at 56 oC for 2 days.

To detect the localization of the sGC subunits, sections were incubated in rabbit antiserum against sGC α1 or ß1 (1:4000) for 4 days at 4 ºC, followed by biotinylated anti-rabbit IgG (1:500) and avidin-biotin-peroxidase complex (ABC Elite 1:1000). Immunoreactivity was detected in 0.05% DAB/0.15%Ni-ammonium-sulfate/0.005% H_2O_2 in 0.05 M Tris buffer, pH 7.6. The immunoreaction product was silver-intensified by using the Gallyas method. After immunostaining, the sections were embedded in Durcupan ACM epoxy resin (Fluka) and then 60–70 nm thick utlrasections were cut with Leica ultracut UCT ultramicrotome (Leica Microsystems, Wetzlar, Germany). The ultrathin sections were mounted onto Formvar-coated single slot grids, contrasted with 2% lead citrate and examined with a Jeol-100 C transmission electron microscope.

Primary antisera	Dilution
rabbit antiserum against nNOS (rabbit polyclonal antibody, Zymed Laboratories, San Francisco, CA)	1:200
rabbit antiserum against soluble guanylyl cyclase α1 (rabbit polyclonal antibody; catalog number G4280, lot number 011K4888; Sigma)	1:4000
rabbit antiserum against soluble guanylyl cyclase ß1 (rabbit polyclonal antibody; catalog number 160897, lot number 134521,Cayman Chemical, Ann Arbor, MI)	1:4000

Specificity of antisera

The specificity of nNOS, sGC α1 and ß1 antisera was reported previously (Szabadits et al., J Neurosci, 2007).

III. Results

Neuronal NOS-immunoreactivity was abundantly present in neuronal perikarya and dendrites and also in axon varicosities in the parvocellular part of the PVN. In perikarya and dendrites, nNOS-immunoreactivity was widely distributed in the cytoplasm, primarily associated with the endoplasmatic reticulum. nNOS-immunoreacivity was also found to be associated with the perikaryal plasma membrane in close proximity to both symmetric and asymmetric synapses, as well as within axon varicosities forming both symmetric and asymmetric synapses. The soluble guanylate cyclase α1 subunit was found in dendrites and also in axon varicosities, and was closely located to both the pre- and postsynaptic sides of synapses in many instances. The α1 subunit was associated with both symmetric and asymmetric types of synapses. The soluble guanylate cyclase α1 subunit was found in dendrites and also in axon varicosities, and was closely located to both the pre- and postsynaptic sides of synapses in many instances. The α1 subunit was associated with both symmetric and asymmetric types of synapses.

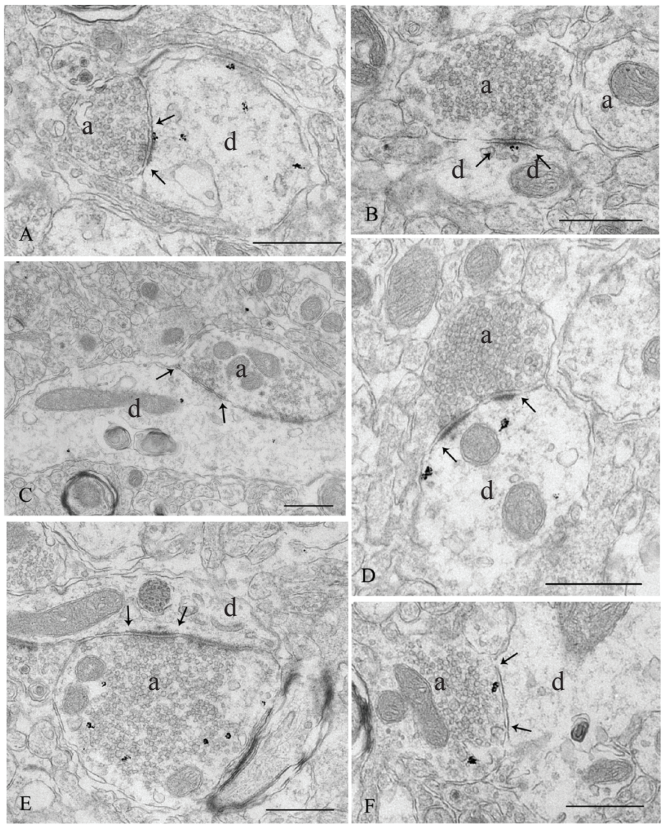


Figure 1. Electron micrographs illustrate the localization of nNOS-immunoreactivity in dendrites (A-D) axons (E, F) and a neuronal perikaryon (G) in the parvocellular part of the paraventricular nucleus in mice. The nNOS-immunoreactivity is labeled with highly electron dense gold–silver granules. nNOS-immunoreactivity can be observed in dendrites in the proximity of the postsynaptic density of both asymmetric (A, B, D) and symmetric (C) synapses. nNOS-immunoreactivity can also be observed in axon varicosities forming asymmetric (E) or symmetric synapses (F). A low-power magnification image shows a nNOS-IR perikaryon (G). In perikarya, nNOS immunoreactivity was widely distributed in the cytoplasm and primarily associated with the endoplasmatic reticulum. Arrows point to synapses. Scale bars=0.5μm in (A-G). a= axon; d= dendrite; Nu= nucleus

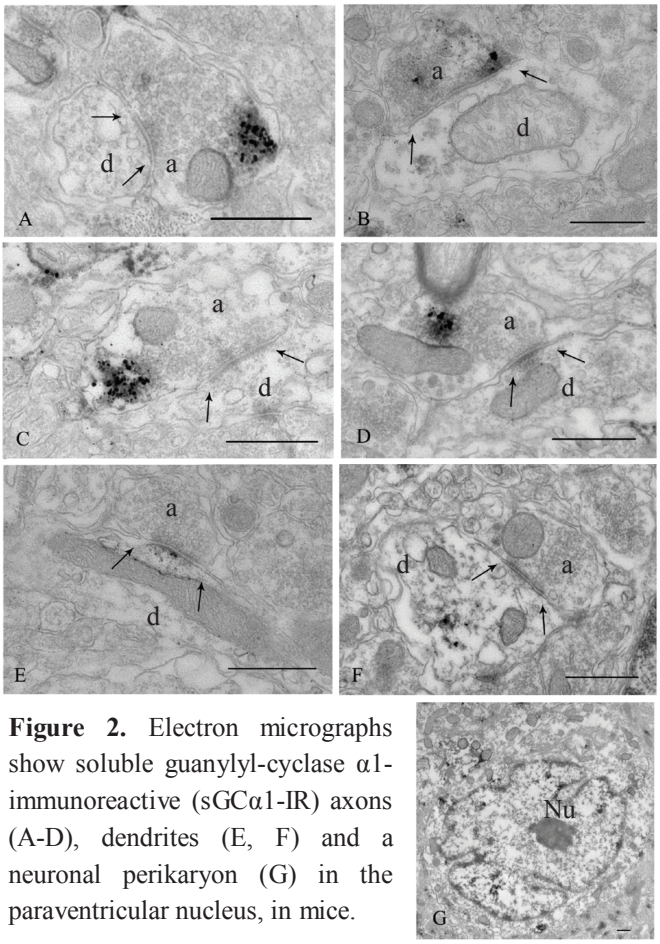


Figure 2. Electron micrographs show soluble guanylyl-cyclase α1-immunoreactive (sGCα1-IR) axons (A-D), dendrites (E, F) and a neuronal perikaryon (G) in the paraventricular nucleus, in mice.

The sGCα1-immunoreactivity is recognized by the presence of the electron dense silver granules. sGCα1-IR axon varicosities form both symmetric (A, D) and asymmetric type synapses (B, E). Similarly both asymmetric (C) and symmetric type synapses (F) are formed on the sGCα1-IR dendrites. Low-power magnification image shows sGCα1-IR perikaryon (G) sGCα1-IR was widely distributed in the cytoplasm. Arrows point to synapses. Scale bars=0,5μm in (A-G). a= axon; d= dendrite; Nu= nucleus

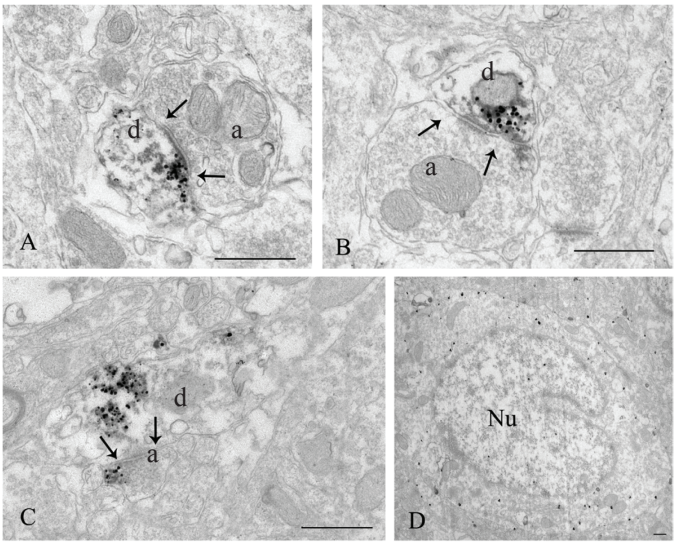


Figure 3. Electron micrographs illustrate (arrows) soluble guanylate-cyclase ß1-immunoreactive (sGCß1-IR) dendrites (A, B, C) axon (C) and neuronal perikaryon (D) in the parvocellular part of the paraventricular nucleus in mice. Soluble sGCß1-IR is recognized by the presence of electron dense silver granules. Asymmetric (A) and symmetric synapses (B, C) are observed on sGCß1-IR dendrites. sGCß1-immunoreactivity is typically associated with the postsynaptic membrane of synapses. A sGCß1-IR axon terminal forms a symmetric synapse with an IR dendrite (C). Low-power magnification image shows a sGCß1-IR perikaryon (G). Arrows indicate the synapses. Scale bars=0.5µm in (A-D). a= axon; d= dendrite; Nu= nucleus

IV. Conclusion

1) nNOS-immunoreactivity is present in both the pre- and postsynaptic sites of inhibitory symmetric and excitatory asymmetric synapses as well.

2) Similar to nNOS, soluble guanylyl-cyclase α1 is also present in both the pre- and postsynaptic elements of symmetric and asymmetric synapses.

3) Soluble guanylyl-cyclase ß1 is localized to dendrites and perikarya, and in several instances, present close to the postsynaptic side of both symmetric and asymmetric synapses.

4) Soluble guanylyl-cyclase ß1 subunit was only very rarely seen in axon varicosities.

5) Our data suggest that nitric oxide can be utilized as both an anterograde and retrograde transmitter in the parvocellular part of the PVN in mice.

References

[1] Esplugues, J. V. (2002). *NO as a signalling molecule in the nervous system.* British Journal of Pharmacology.

[2] Jin-Dong Ding, A. B. (2004). Distribution of Soluble Guanylyl Cyclase in the Rat Brain. *The journal of comparative neurology* , 437-448.

[3] Kadekaro, M. (2004). Nitric oxide modulation of the hypothalamo-neurohypophyseal system. *Brazilian Journal of Medicinal and Biological Research* , 441-450.

[4] Stojilkovic, Y. J. (2006). Molecular cloning and characterization of alpha1-soluble guanylyl cyclase gene promoter in rat pituitary cells. *Journal of Molecular Endocrinology* , 503-515.

Patterns of neuronal synchronous population activity in the human neocortex in vitro

Katharina Hofer
(Supervisor: István Ulbert)
katharina.hofer@gmx.net

Abstract— **In vitro human studies of cortical network activity employ tissue that was surgically removed from epileptic patients for treatment reasons if pharmacological treatment proved to be ineffective. Similar to interictal spikes seen on scalp EEG recordings, spontaneous population activity (SPA) was also observed in vitro in human neocortical slice preparations. In this study, for the first time, neocortical tissue originating from patients with brain tumor was incorporated as non-epileptic control to study SPA in vitro.**
We investigated the cellular and network properties of SPA in neocortical slices. Using a 24 channel laminar microelectrode, SPA could be observed in epileptic as well as in non-epileptic neocortical tissue at similar ratios. SPAs consist of a local field potential gradient (LFPg) transient, high frequency oscillations (HFOs) and increased cell firing (multi unit activity, MUA).
More experiments will further help to elucidate the subtle border between physiological (non-epileptic) and pathological (epileptic) neuronal population activity.

***Keywords*- neuronal population activity; laminar multielectrode; neocortical slice preparation; neuroscience; human;**

***Abbreviations*- synchronous population activity (SPA); local field potential gradient (LFPg); multi unit activity (MUA); high frequency oscillations (HFOs); artificial cerebrospinal fluid (ACSF); current source density (CSD); time frequency analysis (TFR)**

I. Introduction

Cortical neural network activity and it's pathology is extensively studied in vitro and in vivo in various model organisms such as rat, mouse or monkey [1-4]. Epilepsy, one of the most common neurological disorders, is thought to be related to hyperactivity of neuronal circuits. Pharmacological treatment is often effective, but significant numbers of patients resist pharmacotherapy. Surgical tissue removal in these patients offers a remarkable possibility to study living human tissue known to be intimately involved in the generation of this neurological disorder.

Similar to interictal spikes recorded on the scalp EEG, spontaneous population activity (SPA) could be observed in vitro in human epileptic neocortical (see results section and [5]) and hippocampal [6-9] slice preparations in a physiological perfusion solution. The interictal spikes recorded from the scalp of humans are considered to be signs of epileptic activity. Interictal discharges consist of high-amplitude, fast EEG transients, defined as spikes, usually followed by a slow wave that lasts several hundreds of milliseconds (for review see [10]. The waveform of interictal-like discharges in slice preparations of human hippocampal and neocortical tissue derived from epileptic patients shows certain similarities to in vivo recorded interictal spikes [6].

In this study, for the first time, healthy neocortical tissue originating from patients with brain tumor but without epilepsy was used as control to study SPA in vitro. The preliminary data suggest that SPA is not only generated in neocortical slices of epileptic patients but also in neocortical slices of patients with tumor but without epilepsy.

II. Methods

A. Tissue preparation

Postoperative neocortical human tissue (about 0.5 - 1.5cm^3) was obtained from epileptic or tumor patients during their brain surgery and was immediately transferred into ice cold, oxygenated cutting solution (256mM sucrose, 10mM D-glucose, 25mM $NaHCO_3$, 1mM KCl, 1mM $CaCl_2$, 10mM $MgCl_2$, phenol red, saturated with carbogen gas: 95% O_2, 5% CO_2). The pia mater and big blood vessels were removed and the tissue was cut into slices (500µm), perpendicular to the cortical layers.

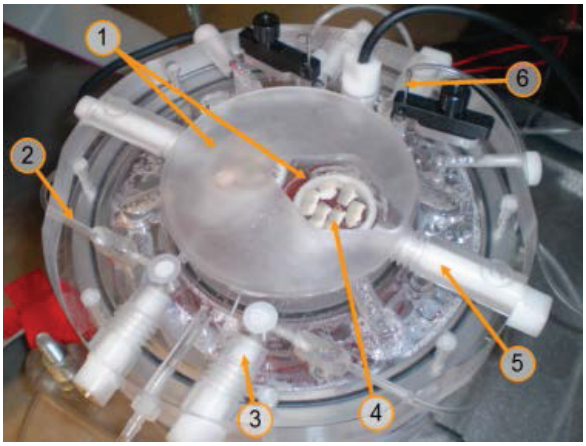


Figure 1. (1) superfusion chambers, (2) ACSF input, (3) reference electrode, (4) brain slices, (5) temperature controller, (6) ACSF output.

Supported by OKTA K81354, OKTA PD77864, ANR-TÉT Neurogen, ANR-TÉT Multisca, TÁMOP-4.2.1.B-11/2/KMR-2011-0002, Bolyai Research Fellowship (toLW), PhD student grant by the hungarian government (to KH)

K. Hofer, "Patterns of neuronal synchronous population activity in the human neocortex in vitro,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 25-28.

The slices were transferred into the interface chamber (Fig. 1), where one side of the tissue was exposed to the humidified carbogen gas, while the other side was superfused with artificial cerebrospinal fluid (ACSF; 124mM NaCl, 10mM D-glucose, 25mM $NaHCO_3$, 3.5mM KCl, 1mM $CaCl_2$, 1mM $MgCl_2$; saturated with 95% O_2, 5% CO_2) at 33°C. The slices were equilibrated for 1h before recording.

B. Recordings

For extracellular recordings (local field potential gradient, LFPg), the laminar electrode (Fig. 2) was placed onto the surface of the neocortical slice without applying pressure but assuring the contact between the electrode contacts and the tissue (Fig. 3). A custom made voltage gradient amplifier of pass-band 0.01Hz to 10kHz was used [11-14]. The data were recorded at 20kHz sampling rate.

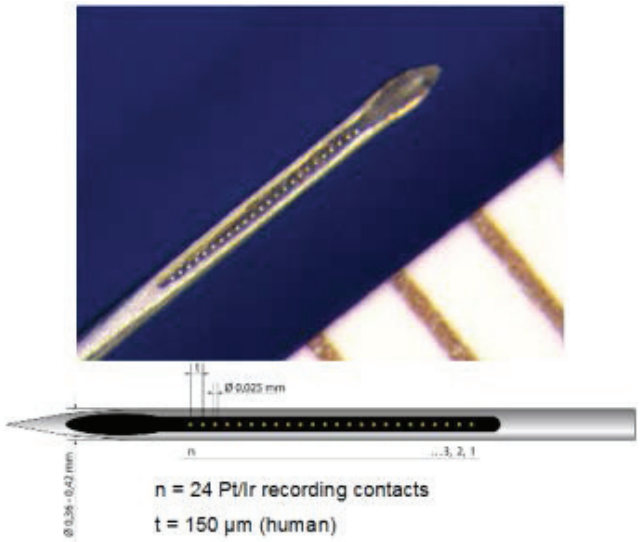


Figure 2. Microscopic photograph and scheme of the bottom view of the laminar multielectrode. As the contacts are arranged in a linear fashion with 150µm intercontact distance, they can cover all 6 neocortical layers.

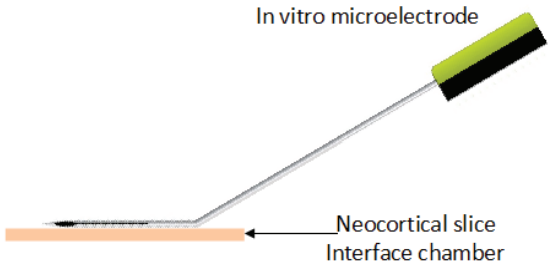


Figure 3. Electrode placement on the tissue.

C. Data analysis

Data were analyzed using the NeuroScan Edit 4.3 program (Compumedics NeuroScan, Charlotte, NC, USA) and routines written for Matlab (The MathWorks, Natick, MA, USA) [12-14]

Current source density (CSD), which estimates population trans-membrane currents was obtained from LFPg recordings by a Hamming-window spatial smoothing followed by applying one additional spatial derivation.

For the MUA analysis, a high pass filter (500Hz) was applied, neuronal firing was detected and clustered (to isolate the activity of different cells) using routines written for Matlab.

SPAs were detected on the optimal channel using an amplitude threshold of 3x the standard deviation after filtering from 3-30Hz. The detected SPAs were then averaged and further processed using the peak of the LFPg on the channel representing the cell layer for correlations with cell firing (MUA).

III. RESULTS

A. Presence of SPA in human neocortical slices in vitro

Spontaneous population activity was generated in neocortical slices of patients with epilepsy (in 48% of slices, n=45/94) or with tumor but without epilepsy (in 42% of slices, n=54/129; see Fig. 4). Fig. 5 shows that SPA was composed of a LFPg transient with superimposed high frequency oscillations (HFOs) and increased cell firing (MUA). CSD analysis indicates that SPA was locally generated. The time frequency analysis (TFR) also showed the emergence of high frequency oscillations of about 200Hz.

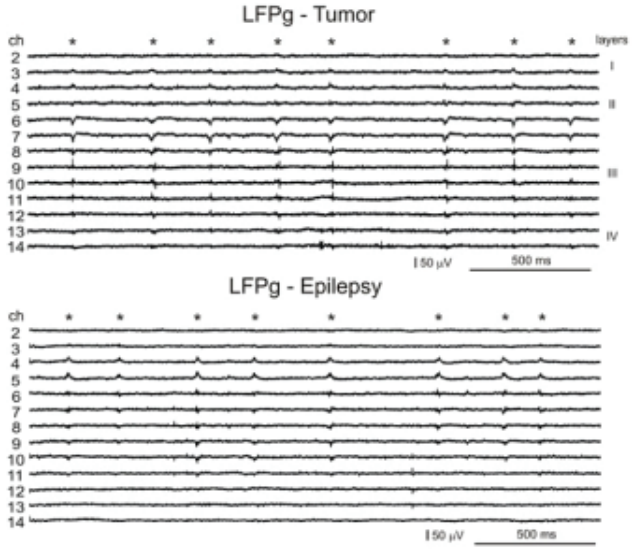


Figure 4. Sample recordings from neocortical slices from a tumor patient (upper panel) and an epileptic patient (lower panel) showing SPA in vitro. It is similar in tumor and in epilepsy patients.

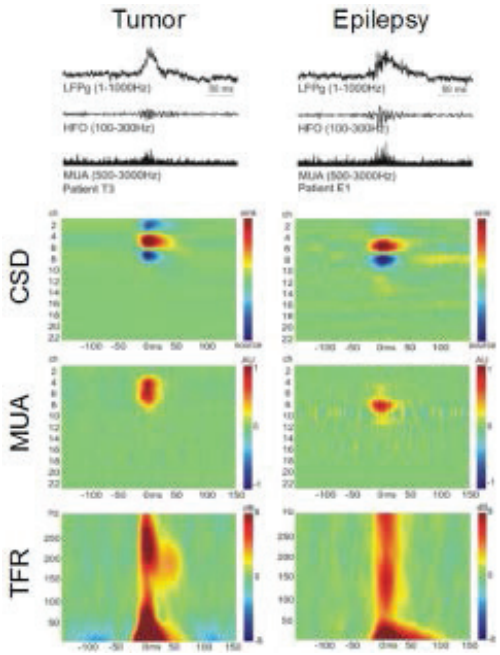


Figure 5. Top panel: samples for LFPg (1-1000Hz band pass), HFOs (100-300Hz band pass) and MUA (500-3000Hz band pass) for in vitro recordings from tumor (left) and epileptic (right) patients. Lower panels: CSD, cell activity (MUA) and TFR.

B. Different patterns of SPA

Different types of SPA were found in the human neocortex in vitro, based on their location (Fig. 6). For both epileptic and non-epileptic tissue, they were most abundant in the supragranular layers. The ratios of occurring types were similar in neocortical slices of epileptic and tumor patients (data not shown). In about 15% of the slices, multiple types of SPA was observed in parallel, in neocortical tissue of both tumor and epileptic patients (data not shown).

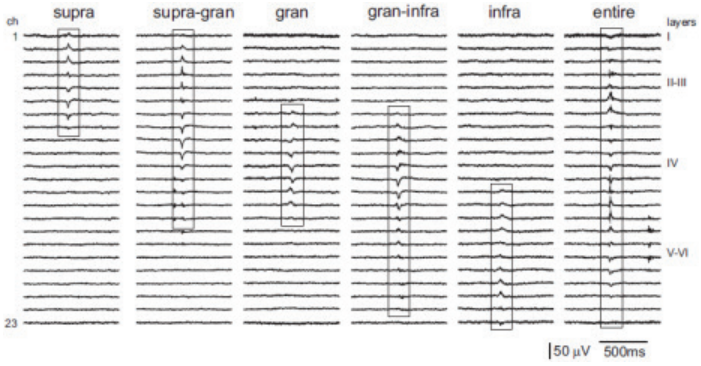


Figure 6. SPAs occurred in different locations within the neocortical layers.

C. Firing behavior of single neurons during SPA

Using cell clustering methods (Fig. 7), we could describe the firing properties of single neurons both in tumor (n=158 cells) and epileptic (n=173 cells) tissue.

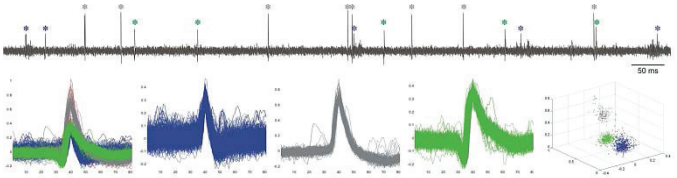


Figure 7. Illustration of cell clustering: Sample recording containing cell activity of multiple cells (top, scale bar: 50ms) and the APs of the isolated cells (bottom).

The peak firing of the cells was determined compared to the peak of the SPA. Different firing patterns could be distinguished in relation to the SPAs (Fig. 8). The ratio of cells showing increased firing during SPA was increased in epileptic tissue (67%) in comparison to non-epileptic tissue (45%). When comparing the neuronal firing activity to the phases of the SPA (Fig. 8), we found that cells from epileptic and non-epileptic tissue both show all differentiated firing patterns, but at slightly different ratios (Fig. 9). In general, cells in the epileptic neocortex seem to fire earlier than cells in non-epileptic tissue related to the SPA peak.

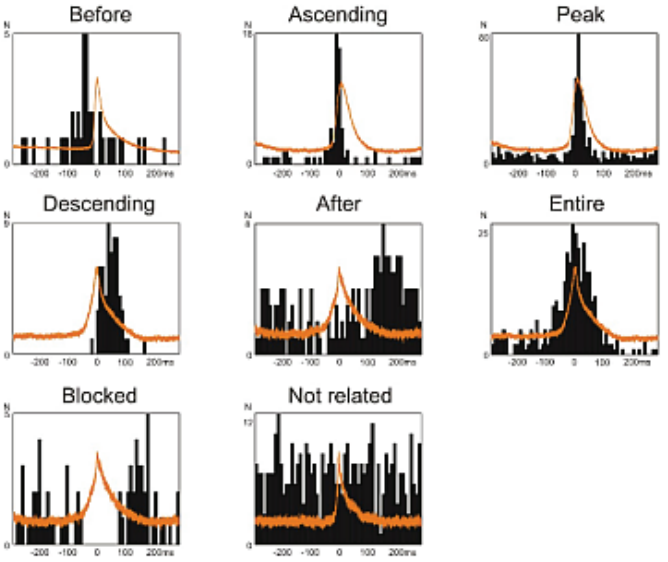


Figure 8. Firing properties of single neurons in relation to peaks of synchronous population activity.

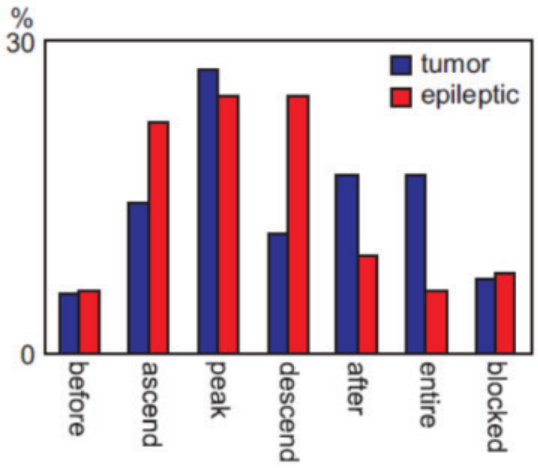


Figure 9. Ratios of the neurons firing related to SPA. Cells firing in a non-related manner (see Fig. 8: not related) were not included.

IV. Conclusion

Interictal spikes recorded on the scalp EEG, are associated with epileptic activity. SPA in vitro is similar to the interictal spikes in EEG recordings. However, we could show that SPA is generated in both epileptic and non-epileptic human neocortical tissue slices. Although it can cover any and all cortical layers, SPA occurs most often in the supragranular layers. In general, the cellular and network properties of SPAs showed only slight differences in tissue slices derived from epileptic and tumor patients. This indicates that in vitro occurring SPA cannot be directly related to epileptic processes.

V. Future plans

In the future, further experiments using pharmacological tools to affect SPA will be performed. In addition, intracellular recordings followed by cell filling will be implemented in addition to the laminar extracellular electrode.

Acknowledgment

The author wishes to acknowledge Dr. István Ulbert and Dr. Lucia Wittner for excellent supervision as well as Dr. Kinga Tóth, Dr. Dániel Fabó and Dr. György Karmos for collaboration and advice as well as the neurosurgeons Dr. Attila Bagó, Dr. Loránd Erőss and Dr. László Entz from the National Neuroscience Institute (OITI) for their collaboration.

References

[1] V. Bouilleret, F. Loup, T. Kiener, C. Marescaux and J.M. Fritschy, "Early loss of interneurons and delayed subunit-specific changes in GABA(A)-receptor expression in a mouse model of mesial temporal lobe epilepsy," Hippocampus 2000, 10: pp.305-24.

[2] Z. Maglóczky and T.F. Freund, "Selective neuronal death in the contralateral hippocampus following unilateral kainate injections into the CA3 subfield," Neuroscience 1993, 56: pp.317-35.

[3] J.O. McNamara, M.C. Byrne, R.M. Dasheiff and J.G. Fitz, "The kindling model of epilepsy: a review," Prog Neurobiol 1980, 15: pp.139-59.

[4] L. Turski, E.A. Cavalhiero, M. Sieklucka-Dziuba, C. Ikonomidou-Turski, S.J. Czuczwar and W.A. Turski, "Seizures produced by pilocarpine: Neuropathological sequelae and activity of glutamate decarboxylase in the rat forebrain," Brain Res. 1986, 398: pp.37-48.

[5] R. Kohling, A. Lucke, H. Straub, E.J. Speckmann, I. Tuxhorn, P. Wolf et al, "Spontaneous sharp waves in human neocortical slices excised from epileptic patients," Brain 1998, 121 (Pt 6): pp.1073-87.

[6] I. Cohen, V. Navarro, S. Clemenceau, M. Baulac and R. Miles, "On the origin of interictal activity in human temporal lobe epilepsy in vitro," Science 2002, 298: pp.1418-21.

[7] G. Huberfeld, L. Wittner, S. Clemenceau, M. Baulac, K. Kaila, R. Miles et al, "Perturbed chloride homeostasis and GABAergic signaling in human temporal lobe epilepsy," J Neurosci 2007, 27: pp.9866-73.

[8] L. Wittner, G. Huberfeld, S. Clemenceau, L. Erőss, E. Dezamis, L. Entz et al, "The epileptic human hippocampal cornu ammonis 2 region generates spontaneous interictal-like activity in vitro," Brain 2009, 132: pp.3032-46.

[9] C. Wozny, A. Knopp, T.N. Lehmann, U. Heinemann and J. Behr, "The subiculum: a potential site of ictogenesis in human temporal lobe epilepsy," Epilepsia 2005, 46 Suppl 5: pp.17-21.

[10] M. de Curtis and G. Avanzini, "Interictal spikes in focal epileptogenesis," Prog Neurobiol 2001, 63: pp.541-67.

[11] I. Ulbert, E. Halgren, G. Heit and G. Karmos, "Multiple microelectrode-recording system for human intracortical applications," J Neurosci Methods 2001, 106: pp.69-79.

[12] I. Ulbert, G. Heit, J. Madsen, G. Karmos and Halgren E, "Laminar analysis of human neocortical interictal spike generation and propagation: current source density and multiunit analysis in vivo," Epilepsia 2004, 45 Suppl 4: pp.48-56

[13] I. Ulbert, Z. Maglóczky, L. Erőss, S. Czirják, J. Vajda, L. Bognár, S. Tóth, Z. Szabó, P. Halász, D. Fabó, E. Halgren, T.F. Freund and G. Karmos, "In vivo laminar electrophysiology co-registered with histology in the hippocampus of patients with temporal lobe epilepsy," Exp Neurol 2004, 187: pp.310-318.

[14] D. Fabó, Z. Maglóczky, L. Wittner, A. Pék, L. Erőss, S. Czirják, J. Vajda, A. Sólyom, G. Rásonyi, A. Szűcs, A. Kelemen, V. Juhos, L. Grand, B. Dombovári, P. Halász, T.F. Freund, E. Halgren, G. Karmos, I. Ulbert, "Properties of in vivo interictal spike generation in the human subiculum," Brain 2008, 131: pp.485-99.

An extended spell checker for unknown words

Balázs Indig
(Supervisor: Dr. Gábor Proszéky)
indig.balazs@itk.ppke.hu

***Abstract*—Spell checking is considered a solved problem, but with the rapid development of the natural language processing the new results are slowly extending the means of spell checking towards grammar checking. In this article I review some of the spell checking error classes in a broader sense, the related problems, their state-of-the-art solutions and their different nature on different types of languages (English and Hungarian), arguing that these methods are insufficient for some language classes. Finally, I present my own method of batch spell checking in large volumes of coherent text.**

***Keywords*-spellchecking; context-sensitive; batch-correction**

I. Introduction

Tools called "spell checkers" are widely used in current word processing systems as an error correcting tool. By the rapid changing of the Internet and computers, the current spell checking is gaining an increasing importance in our lives by the growing capacity of computers, because of the increasing number of ways and volumes content created. Traditionally, spell checkers did subsequent word-by-word analysis, and then transferred to do the analysis while typing. This made it possible for spell checkers to have significance beyond word processors. Nowadays spell checkers can be found everywhere from web browsers to e-mail clients and people use them actively. As in the beginning, today as well the basic principle is the word-by-word analysis, thus the spell checking procedure is stuck at word level. Developers in the IT industry concentrate on these local tools, for example the increasingly better support of agglutinative languages and word compounding appeared approximately 5-6 years ago[1], and in the meantime dictionaries follow the changes of individual languages (by adding new words). Meanwhile, in the field of Natural Language Processing things are developing rapidly as well, but these novel approaches have rarely been applied in spell checking systems yet. A 10 million word English corpus has less than 100,000 different word forms, a corpus of the same size for Hungarian contains well over 800,000[2]. While an open class English word has about 46 different word forms, it has several hundred or thousand different productively suffixed forms in agglutinating languages[3]. The standard tools, which have been proven good in English cannot be applied without any modification. In the literature there exist a lot of separate algorithms that have proven good for partial problems in the English language. I am going to review these state-of-the-art methods and I am going to argue that they cannot be applied because of the nature of the Hungarian language. I will describe my paradigm of spell checking in detail.

All of the aforementioned methods have something in common. They are working with a larger volume of texts. I will set another constraint: I will suppose that all the texts which are examined are coherent. So I can rely on the text-level information, which lies in the text to be extracted, examined and used to improve spell checking performance.
I want to show that spelling errors can be widely different. One must classify these errors and make special sub-solutions for each class to locate and correct most of the errors found in current Hungarian texts with the lowest false positive rate as possible.

II. Types of spelling errors

The academic Hungarinan spelling rules are very complex. They involve semantic features like substance names, occupation names, etc. and the way one should imagine the word: e.g. "légikísérő" is written in one word because the word "kísérő" is in the air physically and not figuratively. The rough listing of the types of errors is as follows:

- in-word errors: One take a word, and modify it by edit distance (e.g. the so called Damerau-Levenshtein distance[4][5]), so the word does not become some other valid word. This is the oldest error observed and most of the errors in English can be corrected by searching the word no more than one distance from the erroneous form. The English language is so sparse that there are only a few candidates. In Hungarian this type of error has not been a problem for a long time. There are several models for this type of errors (e.g. the Noisy Channel Model[6]), but the rate of these errors is much lower then in English.
- real-word errors: One take a word, and modify it, so the modified word becomes a valid meaningful word that has nothing to do with its context. For example: "He had lots of *honey* (money), he wanted to buy a bigger house." These errors must be approached differently. If one knows that the writer has a specific mother tongue and English is his second language one can collect statistical information about the typical misspellings and use them to correct errors [7]. In this type one must distinguish between the words that changed their word species and those which did not. (e.g. money $\rightarrow$ honey, defuse $\rightarrow$ diffuse) In Hungarian there are more word species, so there are more errors of this type.
- word compounding errors: One take two words, and write them as one or take a compound word and write it in two words. The real problem is that the former can be detected and corrected at word level, but the latter cannot.

B. Indig, "An extended spell checker for unknown words,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 29-32.

The Hungarian Academy rules are so complex in this case that in Hungarian a lot of errors fall into this class.

- Out of Vocabulary (OOV) errors: The traditional spell checkers work with a list of words or the list of stems and the production rules (these two are together called lexicon), but there are open word-classes and the spell checker must distinguish between the unknown or OOV words and the misspelled ones. Not to mention the right and consistent use of these words. This can only be detected in a larger volume of coherent text.
- punctuation errors: The right punctuation in the text is not closely related to spell checking, but helps people and the programs to interpret the written text. And can be checked and corrected with the same tool-set as the aforementioned error classes.
- grammar errors: These kind of errors cannot be clearly separated from the cases mentioned above, so I list this class here.

A. How Hungarian and English differ

There are several tools that work language independently, but the most important resources are language dependent. With the help of the self-developed tools in the MTA-PPKE-NLPG research group I can split any raw text to sentences and tokenize it[8]. I can recognize named-entities for future use[9]. Then with the POS-tagger I can couple every word with a tag that reflects its distributional preferences and therefore can classify them into groups[10]. The number of the groups vary from language to language. For example, in English there are only 36 and in Hungarian there are more than 1000 word class tags[11][12]. This makes the task much harder for Hungarian, and the problem becomes even worse when one restricts the domain to clinical texts[13]. As Hungarian is a highly inflected language there are many word forms that belong to the same stem. And there are many homonyms as well, so all in all it is far less sparse than English. Therefore the error types mentioned above cannot be corrected by word-level easily. One can apply Machine Learning methods for extracting features from the context and make decisions, but the liberal word ordering of the Hungarian language makes this task ineffective.

III. Methods in the literature

The current state-of-the-art methods approaching different parts of the whole spell checking. I will list some techniques and argue that they cannot work in Hungarian.

- Take the function words and record their contextual features, because subsequent function words can identify what should come after them and that can be checked for validity[14]. This technique has been successfully applied for the German language on compound words and punctuations. In Hungarian the function words can be omitted and therefore this method cannot achieve much success.
- Make a confusion set of the common misspellings and their right forms[15]. This method can be successfully applied for accenting and word-sense disambiguation. But only on languages that are not inflected and have few word forms. In Hungarian the morphological production rules can be theoretically infinite, and the resources are not available. If the right resource existed, then still one would face the sparse data problem. This highlights other problems: for example, to use stop words or not, and when to use the real word form over the distributional tag. It is desired to automatically choose the right candidate suggestion, but the sufficient features cannot be retrieved from the text because of data sparsity. One way to help this is to rank the suggestions by weighing the edit distance[16].
- One can approach by defining a hash function that collide only on the misspelled and right spelled words and therefore one gets automatically the correct word form for the misspelled word[17][18]. This method can only work if one has a list of misspelled words and the correct forms to train the hash function to work as expected.

IV. My own method

Text corpora forms a consistent closely related text in one topic. That information can be used. I am trying to reduce the number of false positive results of traditional spell checking algorithms. At the same time I want to collect information of the new words and make their usage more consistent[1] by the interaction of the user. I also want to reduce the time consumed by the proofreading of the text by classifying the spelling errors by the stems and guessed production paradigms, so the user does not have to correct every occurrence of the same misspelling (or those which belong to the same stem) one-by-one[19]. This method would stay at word level, but will not be restricted to a fixed lexicon that is integrated into the spell checking programs. I use all of our tools in pipeline and make statistical inferences from the decorated text.[20]

A. Statistical methods on the decorated text

The text was split into sentences and tokens, then I added the POS-tag and lemma for every token with the information of the candidate lemma-tag couples. I also added the information, whether a token is recognised as a correct word form or not. Then I examined the following features of the tokens:

- the frequency of each word form
- the frequency of lemmas of the incorrect words
- the combination of the above

While examining word forms classified by their lemmas, one can find features that characterize the Hungarian morphological production system, which is hard-coded in the morphological analyser[21] for the fixed list of words. If one can find a sufficient number and quality of word forms one can construct an inflectional paradigm that makes a good point to examine the less frequent words against. If these words meet the expectations of their lemma's paradigm, then they

[1] as the program has no information about how the different forms of these words should be spelled

are considered good, otherwise they are considered misspelled and the user is asked to decide. The paradigm also helps to generate suggestions of the misspelled word. They come from the paradigm and it is not necessary for them to appear in the text. The possibility of automatically correcting these words becomes available. There is a threshold that must be set in order to distinguish between low frequency misspelled words and the ones that are too frequent to be misspelled. This threshold can be set safely between 3-5. As the non-systematic misspellings are so diverse that there cannot be such coincidence. The systematic misspellings are considered to be right as the program does not have any external information of the text. Just helps to increase the consistence of the text. The words that are above the threshold are considered "certainly good", the others need to be checked with the extended spell checker. From "certainly good", frequent word forms and their lemmas, the program generates the paradigms. With that, the program checks the other "possibly misspelled" words. The traditional spell checkers' engines can be extended to accept the new words and generate an inflectional paradigm to work with. This can save a lot of time and effort as generating the suggestions is not a trivial task. The classified word forms with their accompanying suggestions can be displayed to the user at once and he can accept or decline the suggestions for each occurrence by examining the context of the word without even proofreading the whole document, just looking at the critical parts of the text if it is necessary. To apply the changes at once the program must map the corrected text to the original one. This could be done for example by Dynamic Time Warping (DTW)[22]. By finding anchors in the text and make the two versions parallel. This could be very useful on environments with special formatted texts, where the formatting is destroyed during the preprocessing steps.

B. Adapting POS-tagger to the text with a posteriori information

The tokenized text is passed to the POS-tagger, to couple each word with its stem, tag and the possible other candidates. For the known words this task is easy. The morphology module can help the tagger, but when it comes to the new words, that are not known either by the morphology module or by the POS-tagger the number of candidates can grow from one up to ten. These candidates mostly differ in the lemmas of the words. The statistical module tries to guess the appropriate lemmas. But this module does not care for the words seen previously. Guessing is totally local to the word in the text. No context is taken into account, but the information is lying in the text. Therefore, after the preprocessing task my program selects the lemmas of the unknown words (choosing also from the candidates) in the text which are frequent enough to not being noise (see table II). I feed these selected lemmas to the POS-tagger. In another pass the POS-tagger selects the fed lemma from the candidates unconditionally if he can. This method can be repeated and all the repetitions improve the performance of the guesser for the current text to a level and decrease the number of the candidates which the POS-tagger chooses from. (There can still be more candidate tags for the same lemma.)

V. Results

The efficiency of the method was tested on two corpora (table I). One is a book (Orwell: 1984) full of theoretically good, but self-invented words. Some of these words are not known by the spell checker but those words are in control. The other is taken from the Internet, contains newspaper articles from a specific site. The size of the two corpora is almost identical. The language model is taken from Szeged corpus 2.0 [12]. The table shows two stages before and after the following heuristic filtering: I filtered out the tokens that were definitely some affix or were not containing four alphabetic letters beside each other (table I). With this filtering, I hope that the real words come into view. Later, I worked with these set of tokens.

TABLE I
The statistics of the used corpora

	1984		*Articles*	
Filtering:	*before*	*after*	*before*	*after*
Tokens:	99913	50586	74053	40716
Tokens (unique):	20393	18211	20916	18465
Not known by Humor:	301	283	1431	1224
Not known by Humor (unique):	181	168	1029	886

TABLE II
Example of word form frequencies

word form	**frequency**	**stem**
Obama	40	Obama
Obamaáról	1	Obamaá
Obamák	1	Obamá
Obama-kormány	1	Obama-kormány
Obamának	3	Obam
Obamának	3	Obamá
Obamára	1	Obamá
Obamáról	3	Obam
Obamáról	3	Obamá
Obamát	5	Obam
Obamát	5	Obamát
Obamával	1	Obamával

As seen in table III, there were many words that were found to be good and with the traditional spell checking methods would become false positives. There were word forms above the threshold and these were selected to be the base of the inflection paradigm for other flexed form of the same stem (see table IV). Finally, the remaining words were considered to be misspellings and suggestions were generated (see table V). In table V one can see the faults of the trivial suggestion generation algorithm. This can be vastly improved by using the engine of some traditional spell checker program.

VI. Conclusion

The described method can correct a wider class of the aforementioned misspellings than the traditional spell checkers. This initial phase of the research shows that with my new method the entire proofreading process becomes simpler and faster as the size of the text grows. The amount of text processed per unit of time clearly increases.

TABLE III
RESULTS

	1984	*Articles*
Stems altered:	34	65
Stems altered (unique):	19	48
Frequent stems:	14	55
Frequent word forms:	40	51
Inflection paradigms:	17	58
Suggestions (for new words):	3	8

TABLE IV
GOOD INFLECTION PARADIGMS

1984	
Stem	
beszélír	
Good form	**Rare form**
beszélírba	beszélírja
beszélírral	beszélírtól
beszélír	
beszélírt	

Articles	
Stem	
Obama	
Good form	**Rare form**
Obamának	Obamáék
Obamáról	Obamára
Obamát	Obamával
Obama	

TABLE V
SUGGESTIONS

Articles	
Misspelled word	**Suggestion**
BruxInfo	Bruxinfo
Gingrics	Gingrich
Mtelekom	MTelekom
Obamaáról	Obamáról
Osama	Obama
Sandber	Sandberg
stent	sztent
Unicredit	UniCredit

1984	
Misspelled word	**Suggestion**
aszondom	Aszondom
beszélírja	beszélírba
jógondoló	jógondol

VII. FUTURE WORK

The method is currently not able to make corrections automatically, but beside this the other paths of future research are:

- extending the spell checker program's lexicon efficiently
- building a misspelling dictionary
- making collaborated spell checking and correction easier with shared lexica
- rapid domain adaptation

These workflows are quite demanding today, with my proposed method it becomes much easier.

ACKNOWLEDGMENT

I would like to thank my Professor and Colleagues for their help.

REFERENCES

[1] N. László. (2005, Jul.) Hunspell, hungarian spell checker. [Online]. Available: http://sourceforge.net/projects/hunspell/

[2] C. Oravecz and P. Dienes, "Efficient stochastic part-of-speech tagging for hungarian," in *In Proc. of the Third LREC, Las Palmas, Espanha*, 2002, p. 710717.

[3] O. György and N. Attila, "Purepos – an open source morphological disambiguator," in *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*, 2012.

[4] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, no. 3, pp. 171–176, Mar. 1964. [Online]. Available: http://doi.acm.org/10.1145/363958.363994

[5] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady.*, vol. 10, no. 8, pp. 707–710, Feb. 1966.

[6] M. D. Kernighan, K. W. Church, and W. A. Gale, "A spelling correction program based on a noisy channel model," in *Proceedings of the 13th conference on Computational linguistics - Volume 2*, ser. COLING '90. Stroudsburg, PA, USA: Association for Computational Linguistics, 1990, pp. 205–210. [Online]. Available: http://dx.doi.org/10.3115/997939.997975

[7] A. Rozovskaya and D. Roth, "Generating confusion sets for context-sensitive error correction," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 961–970. [Online]. Available: http://dl.acm.org/citation.cfm?id=1870658.1870752

[8] B. Indig, "Puretoken: egy új tokenizáló eszköz." Szeged: Szegedi Egyetem, 01/2013 2013.

[9] R. Farkas, G. Szarvas, and R. Ormándi, "Improving a state-of-the-art named entity recognition system using the world wide web," in *Proceedings of the 7th industrial conference on Advances in data mining: theoretical aspects and applications*, ser. ICDM'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 163–172. [Online]. Available: http://dl.acm.org/citation.cfm?id=1770770.1770787

[10] A. Novák, G. Orosz, and B. Indig, "Javában taggelünk," Szegedi Egyetem. Szeged: Szegedi Egyetem, 12/2011 2011.

[11] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," *COMPUTATIONAL LINGUISTICS*, vol. 19, no. 2, pp. 313–330, 1993.

[12] D. Csendes, J. Csirik, and T. Gyimthy, "The szeged corpus: A pos tagged and syntactically annotated hungarian natural language corpus." in *TSD*, ser. Lecture Notes in Computer Science, P. Sojka, I. Kopecek, and K. Pala, Eds., vol. 3206. Springer, 2004, pp. 41–48. [Online]. Available: http://dblp.uni-trier.de/db/conf/tsd/tsd2004.html#CsendesCG04

[13] B. Stomach and V. Hit, "Novel applications of the stomach-hit algorithm," *Commun. ACM*, vol. 8, no. 13, pp. 1687–1693, Apr. 1987. [Online]. Available: http://doi.acm.org/14.1343/345538.356446

[14] R. Kese, F. Dudda, G. Heyer, and M. Kugler, "Extended spelling correction for german," in *Proceedings of the Third Conference on Applied Natural Language Processing*. Trento, Italy: Association for Computational Linguistics, March 1992, pp. 126–132. [Online]. Available: http://www.aclweb.org/anthology/A92-1017

[15] M. P. Jones and J. H. Martin, "Contextual spelling correction using latent semantic analysis," in *Proceedings of the fifth conference on Applied natural language processing*, ser. ANLC '97. Stroudsburg, PA, USA: Association for Computational Linguistics, 1997, pp. 166–173. [Online]. Available: http://dx.doi.org/10.3115/974557.974582

[16] M. A. Elmi and M. Evens, "Spelling correction using context," in *In Proceedings of COLING/ACL 98*. Morgan Kaufmann Publishers, 1998, pp. 360–364.

[17] M. Reynaert, "Text-Induced Spelling Correction," Ph.D. dissertation, Tilburg University, Tilburg, The Netherlands, 2005. [Online]. Available: http://ilk.uvt.nl/˜mre/TISC.PhD.MartinReynaert.pdf.gz

[18] ——, "Text induced spelling correction," in *Proceedings of the 20th international conference on Computational Linguistics*, ser. COLING '04. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. [Online]. Available: http://dx.doi.org/10.3115/1220355.1220475

[19] B. Indig and G. Prószéky, "Ismeretlen szavak helyes kezelése kötegelt helyesírás-ellenőrző programmal." Szeged: Szegedi Egyetem, 01/2013 2013.

[20] G. Prószéky and B. Kis, "A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ser. ACL '99. Stroudsburg, PA, USA: Association for Computational Linguistics, 1999, pp. 261–268. [Online]. Available: http://dx.doi.org/10.3115/1034678.1034723

[21] A. Novák and T. M. Pintér, "Milyen a még jobb humor?" Szegedi Egyetem. Szeged: Szegedi Egyetem, 12/2006 2006.

[22] R. Bellman and R. Kalaba, "On adaptive control processes," *Automatic Control, IRE Transactions on*, vol. 4, no. 2, pp. 1–9, Nov. 1959. [Online]. Available: http://dx.doi.org/10.1109/tac.1959.1104847

Metabolic changes during differentiation of neural stem cells

Attila Jády
(Supervisor: EmíliaMadarász)
jah@net-home.eu

***Abstract*—According to previous results [1], neural stem cells survive at much lower oxygen supply than neurons, both in vivo and in vitro. In order to understand the diverse O_2-demand, metabolic analyses were carried out on one-cell derived populations of neural stem cells representing progenitors of the neural plate /early neural tube (NE-4C) [2] and the adult neurogenic zones (HC_A and SVZ_M) [3]. Depending on origin and developmental stages, different stem cells displayed different responses to supplementing the "starvation" medium with single metabolites (glucose, lactate, β-OH-butyrate, amino acids). The data indicate that the basic metabolism shifts with the advancement of neural differentiation, and the metabolic profile reflects the origin and stage of differentiation of distinct neural stem/progenitor populations.**

***Keywords*: neural stem cell; metabolism; neuronal differentiation; oxygen consumption**

I. INTRODUCTION

Previous data indicated that neural stem cells and their differentiating progenies require significantly different environment for survival. Besides the needs for growth factors, adhesive surfaces and cell activation patterns, the changes in metabolism play important roles in decision on integration or decay of young neural cells in the course of development, regeneration and physiological neuron-replacement.

Under hypoxic ($[O_2] \leq 1$ (v/v)%) conditions, neural stem cells survive and proliferate but can not differentiate; under hypoxic conditions, committed neural precursors and maturing neurons die [1]. The composition of mitochondrial membranes is also changing during the formation of neurons as it was shown by the presence of TSPO 18 kDa (PBR; peripheral benzodiazepine receptor) in stem cells and early neuronal progenitors [4], but not in mature neurons.

Biochemical reasons underlying developmentally regulated changes of the metabolic machinery are not understood.

In order to explore biochemical processes behind differentiation-dependent metabolic changes, the O_2-consumption of one-cell derived neural stem cell populations representing progenitors of early brain vesicles (NE-4C) [2] and adult neurogenic zones (HC_A and SVZ_M) [3] were investigated. In vitro induced neural differentiation of these cells provided models to investigate some metabolic characteristics of developing neural cells at defined stages of differentiation.

II. CHARACTERIZATION OF NON-INDUCED AND PARTIALLY DIFFERENTIATED NEURAL STEM/PROGENITOR CELLS. SCHEDULE OF IN VITRO DIFFERENTIATION.

A. NE-4C, embryonic neural stem cells

The NE-4C neural stem cell linewas derived from the forebrain of a 9-day old, $p53^{-/-}$ mouse embryo [2].In non-induced state, the cell divide continuously and display epithel-like morphology. Treatment with 10^{-8}-10^{-6}M retinoic acid (RA) initiates neural differentiation of NE-4C cells resulting the formation of mature neurons approximately by the 7th , and astrocytes after the 14th days of induction.Neuron formation proceeds through well characterizedsteps [2, 5, 6, 7]: starts with aggregationofinduced cells (Day1-3)followed by migration out of the aggregates (Days 4-5) and formation of loose neuronal networks (Days 7-) on top of monolayer of substrate-attached non-neuronal cells. (Fig. 1, 2)

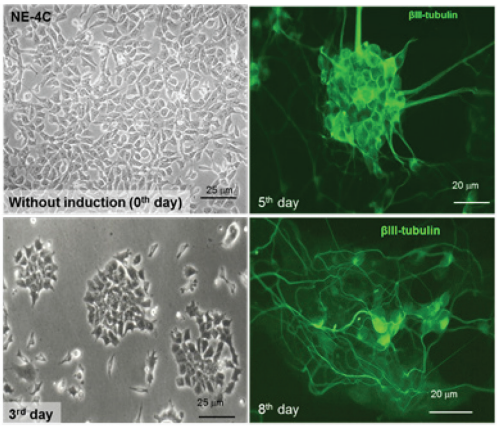


Figure 1. NE-4C cells, derived from the anterior brain vesicles of p53-deficient mouse embryo (E9), proliferate as non-differentiated epithel-like cells in maintaining cultures (RA0) , but give rise to neurons (5th and 8th days) if induced by all-trans retinoic acid (RA).

Tematic Grant of Richter Gedeon Pharmaceutical Inc.,
Grant No.: 4700148815

A. Jády, "Metabolic changes during differentiation of neural stem cells,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 33-36.

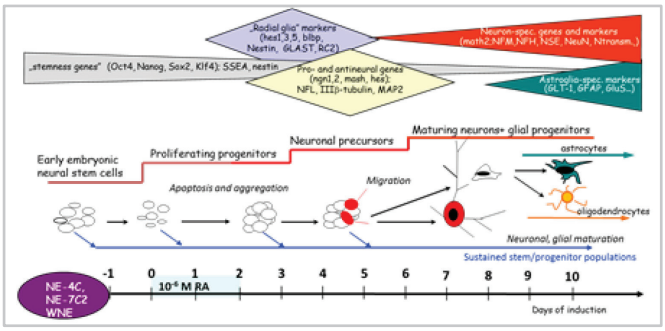


Figure 2. The differentiation proceeds through reproducible stages which had been morphologically, biochemically and physiologically characterized [5, 6, 7].

B. *SVZ_M and HC_A, adult neural stem cells*

The SVZ_M neural progenitor line was derived from the subventriculare zone and the HC_A line was cloned from the hippocampus of adult (P62) mice. The cells were grown on AK-c(RGDfC)-coated surfaces, in serum-free conditions with EGF (epidermal growth factor) supplementation [3]. These cells display characteristics of radial glial (RGl) cells and give rise to neurons upon EGF withdrawal. Neurons develop by the 5th day of induction. [3, 8] (Fig. 3, 4)

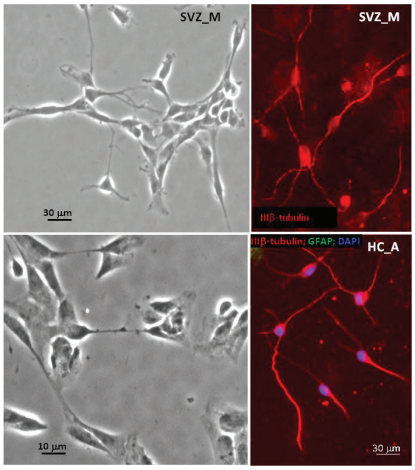


Figure 3. SVZ_M and HC_A cells, derived from the adult mouse subventriculare zone and hippocampus, respectively, proliferate as non-differentiated epithel-like cells in maintaining cultures, but give rise to neurons if induced by EGF withdrawal.

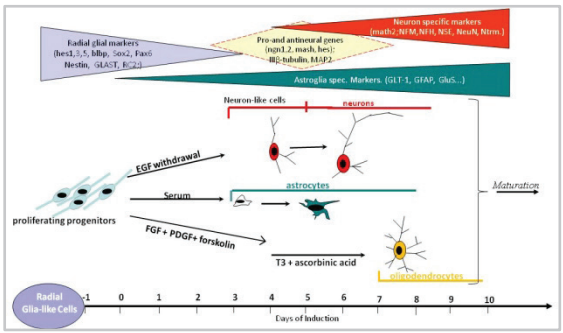


Figure 4. The differentiation proceeds through reproducible stages which had been morphologically, biochemically and physiologically characterized [3, 8].

III. EXPERIMENTAL SETUP

A. *Cell cultures*

The cells were maintained in the appropriate media (table 1 and 2) in water-saturated air atmosphere containing 5% CO_2, at 37 °C. The culture media were changed on every 2nd day. The cells were serially split using 0,05 (w/v) % tripsin with 1mM EDTA [Invitrogen (Gibco)].

Cells were seeded into 96-well Seahorse plates (1-3 x 10^4 cells/well) coated with appropriate adhesive peptides and were maintained as non-differentiated stem cells or were induced with appropriate treatment (table 1) for neural differentiation.

TABLE 1. CELL LINE DATA

Cell line name	**Embryonic cell line**	**Adult cell lines**	
	NE-4C	***HC_A***	***SVZ_M***
Derived from	9 days old mouse embryo forebrain	adult mouse hippocampus	adult mouse subventriculare zone
Medium	5% FCS MEM	high glc DMEM + F12 + B27 + EGF	
Induction medium	Serum-free (in the first 48 hours with 10^{-6} M retionic acid)	high glc DMEM + F12 + B27	
Plate coated with	poly-L-lysine	AK-cyclo[RGDfC]	

TABLE 2. TISSUE CULTURE MEDIA

5% FCS MEM medium	**Serum-free medium**	**High glc DMEM + F12 + B27 (+ EGF)**
•MEM – minimum essential medium [Sigma], •5% heat-inactivated FCS – foetal calf serum [PAA], •0.2 M L-glutamin [Sigma] •0.04 mg/ml Gentamicin [Chinoin]	•50% DMEM – Dulbecco's modified Eagle medium [Sigma], •50% F12 HAM [Sigma], •1% ITS – Insulin–Transferin–Selenium [Gibco], •0.2 M L-glutamin [Sigma,] •0.04 mg/ml Gentamicin [Chinoin]	•50% DMEM – Dulbecco's modified Eagle medium •50% F12 HAM •2% B27 (with retinal) •(40 ng/ml EGF)

The ACSF (artificial cerebrospinal fluid) solution contains 45 mMNaCl [Reanal], 3 mMKCl [Reanal], 2 mM $CaCl_2$ [Sigma], 1 mM $MgCl_2$ [Sigma], 10 mM HEPES [Sigma]. The pH was 7.2. For metabolic assays, the base media were supplemented with one of the following metabolites: 5 mM D-glucose (glc) [Reanal], 5 mM Na-lactate (lac), 5 mM D,L-β-hydroxi-butyrate (βOHB) or5 mM non-essential amino acid mixture (aa) [Gibco].

B. *Determinationof O_2 consumption with Seahorse Cell Metabolism Analyzer*

The Seahorse XF96 Extracellular Flux Analyzer (Seahorse Bioscience)was employed to determine the impact of various metabolic fuel substrates on the mitochondrial bioenergetic processes of non-induced and differentiating NE-4C, SVZ_M

and HC_A cells. Fluorimetric sensors enabled sensitive in situ measurement of O_2 consumption rate (OCR) and the rate of the extracellular pH drift (ECAR) in a 2.28 µl fluid volume above the cells. (Fig. 5)

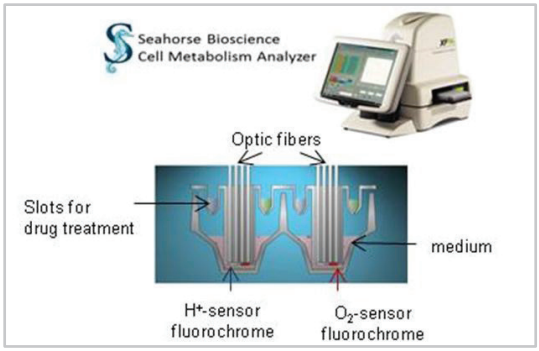


Figure 5. The measuring devices

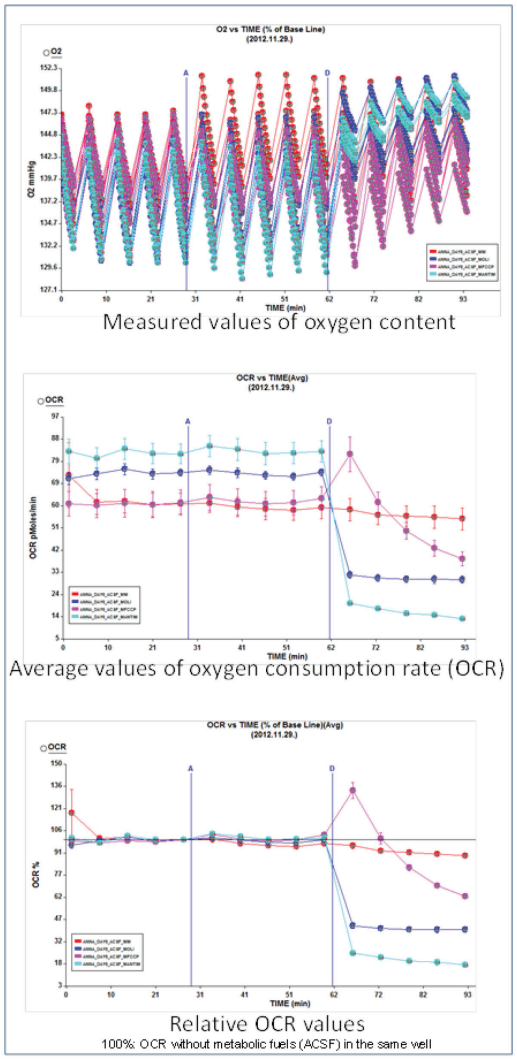


Figure 6. The measured data

In assay-mode, the device reduces the sensing volume to 2.28 µl fluid volume above the cells and produces a gas-tight measuring well. The oxygen content in the cell-covering fluid decreases with time in parallel with the O_2-consumption of cells. The device records the oxygen content in every 15 seconds for 3 minutes, then introduces atmospheric O_2 into the cell covering media by opening up and mixing the wells for 3 minutes. The oxygenation and assay steps alternate. Through ports in the assay-plate, solutions can be introduced to the assay space with a 5-min mixing period. Acidification (pH) of the extracellular medium was measured in parallel with oxygen content in each well. The data are plotted as OCR (oxygen consumption rate: pmole O_2 consumption/min) and ECAR (extracellular acidification rate mpH/min) as a function of time. For comparing reactions of defined wells to added material, OCR and ECAR values were related to those in non-treated control state and plotted as relative OCR and ECAR values. (Fig. 6)

The oxygen consumption rate (OCR) indicates the respiration activity, because the electron transport chain consumes oxygen.The extracellular acidification rate (ECAR)demonstrates mainly the intensity of glycolytic activity, because its lactate production acidifies the environment.

The metabolic state of cells was tested at the end of metabolic assays by monitoring mitochondrial responses to respiration blocking drugs.

Oligomycin blocks ATP synthase (Fig.7) resulting in reduced hydrogen ions consumption and accumulation of hydrogen ions in the intermembrane space of mitochondria. As a consequence, the electron transport chain will be blocked and the oxygen consumption decreased.

FCCP(fluoro3-carbonil cianide-methoxy-phenylhydrazone)opens free routes for hydrogen ionsthrough the inner mitochondrial membrane (Fig. 7)resulting in heavyincrease inthe oxygen consumption.

Cellsresponding accurately to the above drugs possessed functional mitochondria, thus were regarded “healthy”.

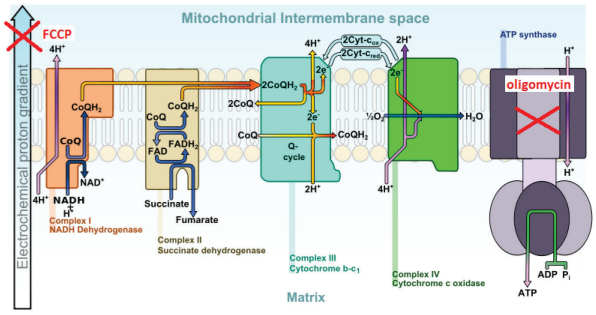


Figure 7. The effects of the used drugs

C. *The metabolic treatment*

For assaying metabolic characteristics, two different approacheswere used. As a common firststep, all cells were thoroughly washed with metabolite-free ACSF to remove metabolic components of the maintaining media. (Fig. 8)

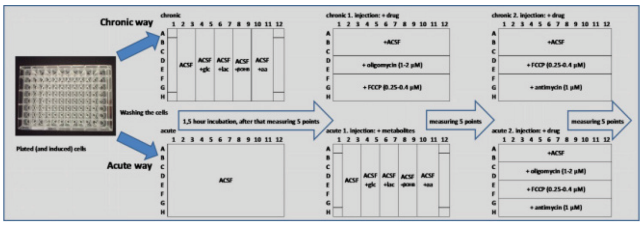


Figure 8. The flow chart of the metabolic treatments

In *chronic assays*, 180 µl ACSFsupplemented with the required metabolite was added to each well. According to the probed metabolites, there were 5 treatment-groups on each plate: without metabolite (starvation), with glucose, lactate, β-hydroxi-butyrate (keton-body) or aminoacid mixture.After 1.5 hour incubation, 5 OCR and ECAR data (60 – 60 datapoints) were recorded. After recording the metabolite effects, mitochondrial drugs were injected one after other, and 5 data points were measured from each treatment.

In *acute assays*, 180 µl ACSFwas added without any metabolite supplementation. (So the cells were starving.) After 1.5 hour incubation, 5 OCR and ECAR data (60 – 60 datapoints) were recorded as a baseline. A supplementary metabolite was then added through the injection port and further 5 data points were measured. At the end of metabolite testing, mitochondrial drugs were injected and 5 data points were measured.

IV. Brief Results

During differentiation,the oxygen consumption and metabolite requirements of developing neural cells change significantly. Results obtained on viability, O_2-consumption and extracellular acidification of distinct neural stem cell lines demonstrated that different neural stem/progenitor populations and also their differentiating progenies display specific, cell-type and developmental stage-dependent demandsfor survival.

Acknowledgment

I would like to thank Prof. LászlóTretter,TündeKovács, Susan Van-Wert andKatalinGaál for their theoretical and technical help.

I also want to thank the help and guidance to my supervisior, Prof. EmíliaMadarász.

This work has been supported by the Tematic Grant of Richter Gedeon Pharmaceutical Inc., Grant No.: 4700148815

References

[1] Anita Zádori, Viktor Antal Ágoston, Kornél Demeter, Nóra Hádinger, Linda Várady, Tímea Kőhídi, Anna Gőbl, Zoltán Nagy, Emília Madarász, “Survival and differentiation of neuroectodermal cells with stem cell properties at different oxygen levels”; Experimental Neurology, vol. 227, pp. 136–148,2011

[2] Katalin Schlett and Emília Madarász, “Retinoic Acid Induced Neural Differentiation in a Neuroectodermal Cell Line Immortalized by p53 Deficiency”; Journal of Neuroscience Research, vol. 47, pp. 405–415, 1997

[3] Károly Markó, Tímea Kőhídi, Nóra Hádinger,Márta Jelitai, Gábor Mező, Emília Madarász, “Isolation of radial glia-like neural stem cells from fetal and adult mouse forebrain via selective adhesion to a novel adhesive peptide-conjugate”; PLoS One (6) e28538, 2011

[4] Varga B, Markó K, Hádinger N, Jelitai M, Demeter K, Tihanyi K, Vas A, Madarász E.,“Translocator protein (TSPO 18kDa) is expressed by neural stem and neuronal precursor cells”; Neurosci.Letts., vol. 462, pp. 257-262, 2009

[5] Herberth B, Pataki A, Jelitai M, Schlett K, Deák F, Spät A, Madarász E.,“Changes of KCl sensitivity of proliferating neural progenitors during in vitro neurogenesis”; J Neurosci Res., vol. 67, pp. 574-582, 2002

[6] Jelitai M, Schlett K, Varju P, Eisel U, Madarász E., “Regulated appearance of NMDA receptor subunits and channel functions during in vitro neuronal differentiation”**;** J Neurobiol., vol. 51, pp. 54-65, 2002

[7] Jelitai M, Anderová M, Chvátal A, Madarász E., “Electrophysiological characterization of neural stem/progenitor cells during in vitro differentiation: study with an immortalized neuroectodermal cell line”; J Neurosci Res., vol. 85, pp. 1606-1617, 2007

[8] Madarasz E., “Diversity of neural stem/progenitor populations. In: Neural Stem Cells”; InTech Book Series; Ed. L.Bonfantini; in press, 2013

Evaluation of Speech Music Transitions in Radio Programs Based on Acoustic Features

Mátyás Jani
(Supervisor: György Takács, Gergely Lukács)
jani.matyas@itk.ppke.hu

***Abstract*—The final target of our research project is to create an automatic program editor for radio. There are several algorithms for music playlist generation, but no reference has been found for mixed speech and music playlists. As for a first step we studied the elements of existing radio programs. A simple subjective opinion test has been constructed to evaluate the ability of normal listeners to discriminate the well edited and the randomly selected speech and consecutive music pairs. Significant difference has been found in the opinions between the well harmonizing pairs and the pairs having dissimilar characteristic. We tried to predict the opinion values based on the basic acoustic features of the speech and the music signals. Some relations can be established based on statistical methods in between the acoustic features and the opinion values. We hope that by using content based features this prediction can be more accurate.**

***Keywords*-radio program; speech; music; speech-music transition; acoustic features;**

I. Introduction

The Internet makes it possible to create personalized audio programs for Internet radio or even audio social media [1]. It is important to provide the listeners an adequate program, but to have a smooth transition between two consecutive programs is also a relevant factor. At the conventional radio channels this is the task of the program editor. There are several existing algorithms for the play-list generation of the online music radios which take these into consideration [2] [3].

The novelty of the present work is that it examines the transitions between the speech and music parts, which have a fundamental relevance for automatically edited program streams that contain both speech and music. Some kind of personal association is used when the music editor selects the music for the speech. This association may based on the content of the pair but it can be affected by acoustic factors as well.

The goal of the research was to investigate the latter connection, so the prediction of the transition quality by matching the non-textual information in the speech and the music features. The hypothesis is that there exists an acoustic relationship, our tests were performed in the hope to discover this. The non-textual information contains the prosodical features, the emotion and also the changes which are done by the sound engineer.

II. Related Work

The playlist generation has gained significant research and industrial interest over the last ca. 5 years. A good entry point can be found e.g. in the dissertation [4] and tutorial [5] of Ben Fields (et. al.). In the playlist generation, beside other aspects (such as the selection of the pieces of music), also the matching of pieces following each other is an issue. Also, different methods and measures for the acoustic similarity of pieces of music were introduced [6] [7]. They aim at selecting pieces of music similar to those a user likes, rather than at a smooth acoustic transition between two pieces. In general, no work, either general or focusing on the match of consequent pairs, is known on mixed music and speech playlist generation.

Recommender systems are related in some respects to the playlist generation. However, recommender systems typically deliver an unordered set of "products" (e.g., pieces of music), without organising them sequentially. The very limited work dealing with sequential recommender systems [8] is concerned with general aspects of human reception, not specifically with pieces of audio material. In [9] a sequential recommender for video is introduced, however, the emphasis is put on other aspects, such as implicit feedback and dealing with the cold-start problem, rather than on our focus, smooth listening experience.

III. Method

A. Overview

Pairs of speech and music pieces from real radio recordings suitable for acoustic tests were selected. Our test set contained original pairs from radio programs presumably edited by professional editors and randomly selected pairs as well.

A listening test was constructed to collect subjective opinion ratings for the speech and music pairs. The test was conducted involving normal listeners and the result was collected.

A set of potentially relevant basic acoustic features was selected for the speech and the music parts. These features were extracted from the recordings used in the test.

The three types of the collected data – the original pairing matched by editor, the subjective ratings from the listeners and the automatically extracted acoustic features – were analysed. We were looking for the relationships, patterns in the data with the final goal of predicting the subjective ratings using acoustic features.

B. Test Set

The aim at the test set selection was to provide speech and music pairs together with a professional opinion both for the subjective opinion test and the acoustic feature extraction.

M. Jani, "Evaluation of speech music transitions in radio programs based on acoustic features,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 37-40.

Several radio stations were used as sources. We cut a total of 25 pairs of speech and consecutive music parts from the radio streams. We preserved the original pairing for having professionally edited positive samples. Other 25 pairs were made by combining the same speech and music parts using a random pairing that excluded the original pairs. With this two groups we had 50 transitions, 25 originally positive (matching) and 25 negative samples.

The test set was a mixture reflecting the variety of the different radio stations. There were male, female speakers in the speech parts, and some cuts had more than one speakers. There were 4 transitions where background music was present under the speech. The music parts consisted of mixed musical styles (classical, rock, pop). Some of them had only instrumental play, but others had singers in English, Swedish and some kind of Arabic language.

In order to avoid the impact of the content we used recordings from radio stations the language of which was not understood by the test subjects. If the listeners do not understand spoken sentences in the selected language then the content will not affect the acoustic based opinion about the transition between the speech and the music. The Swedish language was selected as it is not widely understood and spoken among the target audience of the test.

The level of the loudness for each music and the speech part was normalized to remove the effect of the different loudness in the different radio streams.

The length of the audio clip was selected, so that it is long enough for a well-established opinion, but short enough for an efficient test. The chosen length for each speech and each music part was approximately 10 seconds, so every transition was 20 ± 0.5 seconds long.

Different transition effects were used (i.e. crossfading) between the speech and music programs in the original radio streams. As this may influence our tests we discarded the crossfading parts in the recordings. By removing the crossfading we had speech and music parts with sharp beginnings and endings. To make the transitions sound more natural we applied 1 second fade in to both the speech and music parts and we applied 1 second fade out to the music part (to the end of the transition). The ending of the speech was not modified as sentence finishing parts were selected.

C. Subjective Opinion Test

The subjective measurement was realized in the form of an internet poll. Although this may not provide an equally quiet environment for the listeners and the test makers have less observation on the filling process, it is closer to a realistic radio or music listening environment for each listener. It may boost the willingness for participating in the test as well.

The poll consisted of two parts. The first part had several personal questions, the second contained the subjective opinion test with the radio recordings. The questions in the first part were about the music preference, favourite radio station, musical qualification and language knowledge of the listener. The language knowledge of the listeners was tested with short sound clips. The listener had to answer whether they recognized 5 different words in them. This method also had the advantage to present the sound intensity level for the listeners which was used later on, so they could adjust the volume meter on their speaker.

The second part of the poll contained the 50 different transitions. Only 5 transitions were displayed on each page, the listener could rate these transitions in one. The listener had to make a binary decision for each transition. They had to decide if it was cut from a real radio program or the speech and music part was selected randomly in their opinion. It was possible to have a break during the test and the listeners were asked to stop or pause the filling process when they were tired.

The test was time consuming and we expected that not all listeners would provide an answer for every transition. In order to keep the number of ratings for each transitions approximately equal the transitions having the least number of ratings but not yet rated by the actual listener were selected. The order of the presented transitions on one page was randomized to overcome the order effect.

The technical solution of the poll was a web page with HTML5 audio elements. We provided a fallback mode, where the listeners could download the audio files in mp3 format in order to support web browsers lacking the audio tag feature. This kind of fallback mode is quite uncomfortable but the target audience had much larger likelihood for having HTML5 audio compatible browser than the average internet users.

The target audience was the students of the Information Technology Faculty of the Pázmány Péter Catholic University. Totally 78 listeners participated in the test in one week, 27 of them rated all of the 50 transitions. Every transition received between 38 and 40 ratings, the total number of ratings was 1955. Those who checked that they understand at least 5 Swedish words in the first part of the test made 195 ratings in the second part. Interestingly the ratio of "good" answers for this latter group is only 1 percent higher than for the rest of the listeners.

D. Signal Processing Based Evaluation of Speech and Music Samples

The features were selected based on previous experiences. The number of features was kept low because the number of the transitions was also limited.

The chosen features for the speech were: (1) average of the fundamental frequency (sf_f0_avg), (2) standard deviation of the fundamental frequency (sf_f0_stddev), (3) speech tempo (sf_tempo), (4) spectral coefficients, (5) dynamic range (sf_dynrange). The following features were chosen for the music: (1) music tempo (mf_tempo) (2) dynamic range (mf_dynrange) (3) spectral coefficients.

The fundamental frequency extraction was done using the WaveSurfer program [10], using the ESPS method with default settings.

For the dynamic range we implemented the ITU-recommendation of loudness unit measure [11] and we used the loudness range measure [12] on top of it.

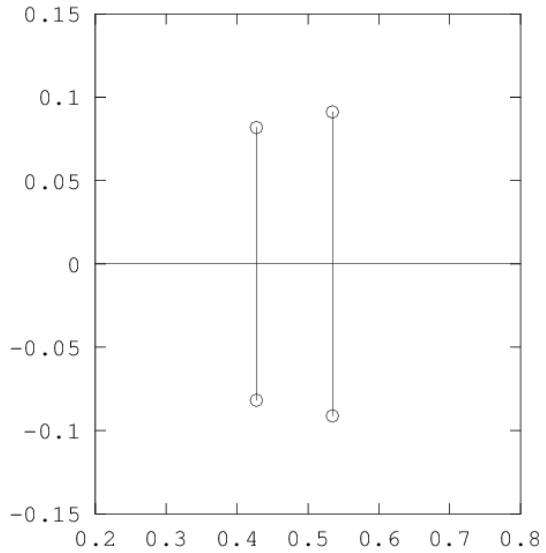


Fig. 1. Mean and standard deviation of the yes-ratio (right vertical segment is for true original matches, left is for randomly selected)

The music tempo was detected by listening, and the measure was the quarter beat per second. For longer music samples it will be possible to detect it automatically. The speech tempo was also detected manually. We used the vowel per second measure.

The spectral coefficients were extracted using OpenSMILE. Overlapping frames were used, the frame size was $0.025s$ with step size $0.01s$. We extracted 16 Mel spectral coefficients, calculated their average over the frames and fitted a second order polynomial (parabola) on the average spectral points of each speech and music part. The coefficients of the polynomials were used as separate speech and music features.

IV. Results

A. Data Distribution

The subjective ratings made by the listeners can be aggregated by counting the number of the "yes" votes for each transition (when they choose that the speech and music was taken from an actual radio program in their opinion) and the "no" votes (when they choose that it was randomly paired in their opinion). We use the yes-ratio to map the subjective ratings for each transition to the $[0, 1]$ interval. It is the ratio between the "yes" votes and all votes for a given transition:

$$\text{yes-ratio} = \frac{\text{\#yes votes}}{\text{\#votes}}$$

If every vote is a "yes" then the yes-ratio $= 1$ and if every vote is a "no" then the yes-ratio $= 0$.

In Figure 1. the mean and standard deviation of the yes-ratio is displayed for the original two group of the transitions (original matches, randomly paired - this is the original match property of the transition). There is a small difference, so originally matching transitions received more "yes" votes, and non-matching pairs received more "no" votes in general.

	match	random
> 0.5	14	8
≤ 0.5	11	17

	match	random
> 0.6	10	4
< 0.4	5	12

TABLE I
Confusion matrices

Fold	Cross-correlation
1	0.37
2	0.61
3	0.55
Average	0.51

TABLE II
Cross-correlations for the different cross-validation folds

The confusion matrices (see Table I.) also support this finding. The columns contain transitions distributed by the original match property. The rows contain transitions where the given restriction for the yes-ratio is valid. The left matrix shows all transitions, the right shows only those where the ratings form a higher majority of either "yes" or "no" votes by at least 20%.

B. Results of PCA

For each transition a vector was put together using the speech and music features described in Section III-D as components. The ratio of the music and speech tempo and the ratio of the dynamic ranges were also combined in the joint feature vector so it had 14 features in total. We used principal component analysis and regression to discover the important parameters on the normalized input data. We used only those transitions which had a yes-ratio > 0.6 or yes-ratio < 0.4. The cross-correlation between the true yes-ratio and the predicted was 0.85. By using only the first five principal components it reduced to 0.59. The most important parameters were the spectral coefficients in both speech and music, and the speech features seemed to be more dominant in prediction, their regression weights were bigger in absolute value.

This result is not too convincing as we used the same instances for training and testing. To avoid this we split the instances in three groups and made 3-fold cross-validation learning by training with two groups and testing with one in each possible setup. The reason for not using the generally favored 10-fold is that it would leave too few instances in one fold, and the calculation of the cross-correlation would be less meaningful. The cross-correlation values between the original yes-ratios and the predicted values by using all coefficients are concluded in Table II. for the three group.

C. Prediction Based on Data Mining Techniques

The set of the acoustic features directly extracted from the audio files was extended with additional, calculated attributes. We took into consideration that most data mining algorithms cannot compare several attributes directly or they require large amounts of data for that. Therefore, we calculated some comparison attributes that might make sense in the application domain. The following attributes turned out to be interesting:

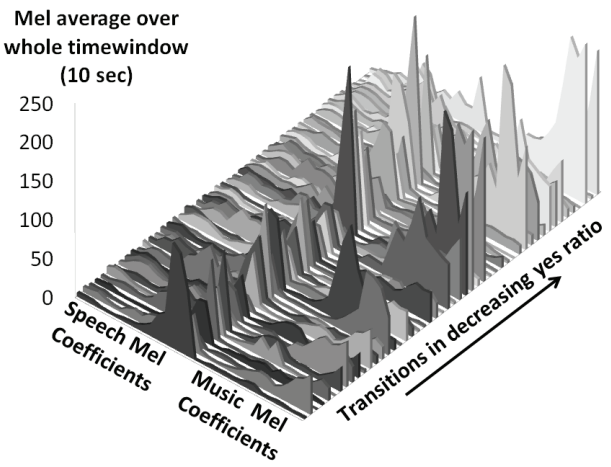


Fig. 2. The avg of the mel spectral coefficients in the speech-music pairs

- relative tempo difference: the difference of the normalized speech and music tempo, divided by the speech tempo (temponorm_diffrel)
- the largest spectral coefficient of the music (mf_mel_max)
- the largest spectral coefficient of the speech (sf_mel_max)
- the sum of the largest spectral coefficient of the parts (mel_maxsum = mf_mel_max + sf_mel_max)

The relationship between the largest spectral coefficients of the speech and music parts is supported by a visual investigation of the coefficients (see Figure 2).

Different data mining methods were used to discover patterns between the music and speech features, including the comparison attributes, on the one side and the class attribute match on the other, using the Weka toolkit [13]. A shortened summery of the methods and findings:

Interestingly, that some features of only one part influenced the match – without considering any features of the other part. Most notably the mf_mel_max (largest mel coefficient in the music part) allows a prediction rate of ca. 66%, with ca 7 times (!) lower false-negative than false-positive rate. I.e. high mf_mel_max values indicate a no-match fairly reliably, though low-values do not guarantee a match. The above performance was achieved with the Naive Bayes method using, because of the limited data amount, 3-fold cross validation. Other methods, such as Multilayer perceptron, Support Vector Machine, Ripper or C4.5 were also tested, but were found to perform less well than Naïve Bayes.

Speeches with a lower tempo (sf_tempo) fit in general better. Using this single feature, the Naïve Bayes method has a prediction performance of 62%.

Concerning the comparison attributes, the following patterns were found. The sum of the largest spectral coefficient values for speech and music (mel_maxsum) performs very similarly to mf_mel_max, both concerning the success rate (ca 66%) as well as the low number of false-negatives.

The best overall performance – 70% – was achieved using the attributes (1) speech tempo (sf_tempo), (2) music dynamic range (mf_dynragen), (3) the sum of the maximal spectral coefficients for the speech and music part (melmax_sum) and (4) the relative tempo difference (temponorm_diffrel) and the Naïve Bayes method.

V. Summary

We studied speech to music transitions for customized playlist generation. Our investigations revealed some specific relationships between the acoustic features and the quality of match. Further work is planned with a larger dataset to make the results more generalizable and precise. Also, semantical matching between the speech and the song texts will be studied.

Acknowledgment

The support of grants TÁMOP-4.2.1./B-11/2/KMR-2011-002 and TÁMOP-4.2.2./B-10/1-2010-0014 is acknowledged.

References

[1] Gergely Lukács, Beáta Pethesné Dávid, and Bea Madocsai, "Impact of personalized audio social media on social networks," in *XXXIII. Sunbelt Social Networks Conference of the International Network for Social Network Analysis Abstract Proceedings*, vol. To appear, Hamburg, Germany, 2013.

[2] Arthur Flexer, Dominik Schnitzer, Martin Gasser, and Gerhard Widmer, "Playlist generation using start and end songs," in *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, 2008, pp. 173–178.

[3] B. Logan, "Content-based playlist generation: Exploratory experiments," in *ISMIR 2002, 3rd International Conference on Music Information Retrieval, Paris, France, October 13-17, 2002, Proceedings*, 2002, pp. 295–296.

[4] Ben Fields, "Contextualize your listening: The playlist as recommendation engine," PhD, Goldsmiths, University of London, London, 2011.

[5] Ben Fields and Paul Lamere, "Finding a path through the jukebox – the playlist tutorial, ISMIR," Utrecht, 2010.

[6] J.-J. Aucouturier and F. Pachet, "Finding songs that sound the same," in *Proc.1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, Nov. 2002, pp. 1–8.

[7] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *Proc. of IEEE International Conference on Multimedia and Expo, ICME*. IEEE, 2001, pp. 745–748.

[8] Eliens A. and Wang Y., "EXPERT ADVICE AND REGRET FOR SERIAL RECOMMENDERS," in *Proc. EUROMEDIA 2007*. Delft, Netherlands: Eurosis, 2007, pp. 111–118.

[9] D. Zibriczky, B. Hidasi, Z. Petres, and D. Tikk, "Personalized recommendation of linear content on interactive TV platforms: beating the cold start and noisy implicit user feedback," in *Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP), Montreal*, 2012.

[10] K. Sjölander and J. Beskow, "WaveSurfer," http://www.speech.kth.se/wavesurfer/, [Online; accessed 14-February-2013].

[11] ITU, "Recommendation ITU-R BS.1770-3 algorithms to measure audio programme loudness and true-peak audio level," 2012.

[12] EBU, "Tech doc 3342 loudness range: A measure to supplement loudness normalisation in accordance with EBU r 128," 2011.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, p. 10–18, Nov. 2009.

[14] Mátyás Jani, György Takács, and Gergely Lukács, "Evaluation of speech music transitions in radio programs based on acoustic features," in *Proc. of 11th International Workshop on Content-Based Multimedia Indexing, CBMI*, vol. To appear, 2013.

Simulation-based Investigation of Temporal and Spatial Characteristics of Photodynamics in Two-Photon Microscope

Imre Benedek Juhász
(Supervisor: Dr. Árpád Csurgay)
juhim@digitus.itk.ppke.hu

***Abstract*—In this paper a computer program is presented that simulates the state transitions taking place in the sample investigated by a two-photon microscope. The simulator is based on a three level (ground state, excited state, triplet state) photodynamic model with four different photobleaching pathways. The sample is modeled by a homogenous fluorophore solution that is divided into cubic cells. The number of fluorophore molecules per state is calculated in each laser cycle, for every cell. Illumination is fixed to one focus and it is described as a Gaussian beam. The simulator enables the investigation of both temporal and spatial characteristics of photodynamics in case of different microscope operating parameters (e.g. laser power, pulse length, pulse repetition rate) and by this it can help in the optimization of these parameters in order to reach better image quality and reduced photobleaching.**

***Keywords*-two-photon microscope; numerical model; simulation; photodynamics; photobleaching;**

I. INTRODUCTION

The operation of the two-photon microscope is based on the phenomenon of two-photon absorption which means that two photons are absorbed simultaneously by one electron whilst the electron gets into an excited state. When the electron returns to the ground state, it emits a fluorescent photon which can be detected. In two-photon microscopes mode-locked lasers are used as light source. Laser pulses are focused on the sample in which around the focus the intensity of the illuminating light is high enough to evoke two-photon absorption. However as we are getting farther from the focus, the probability of two-photon absorption decreases in a large extent since its rate is proportional to the square of the irradiating light intensity. Thus the collected fluorescent photons can be considered to originate from a small volume around the focus. The illuminating laser beam scans across the sample while fluorescent photons are collected from each voxel. Finally a computer constructs an image based on the number of detected photons, where the intensity of the pixels refers to the number of fluorescent molecules in the corresponding volume element of the sample.

As it was mentioned, the probability of two-photon excitation is quadratically proportional to the light intensity, thus in case of a Gaussian beam, it changes in space. After a fluorophore molecule was excited, it can undergo different transitions: for instance it can return to the ground state while emitting a fluorescent photon or it can absorb further photons what might induce molecular transformations. These transformations can irreversibly inhibit the molecule to function as a fluorophore, thus it will not be capable of fluorescence any more. This process is called photobleaching and this is a significant limiting factor of image quality and examination duration.

The aim of this study is the numerical analysis of the aforementioned processes. This paper has two main parts: the first one presents the computational model; the second one demonstrates some simulation examples.

II. THE MODEL

The basic idea of the simulator came from [1] in which a simple three-level photodynamical model of two-photon microscope was described. This model has been completed by elements from [2], [3], [4] what led to a combined model that is delineated in the following section.

A. The Illuminating Laser Beam

1) Beam Profile

The illuminating laser beam is described as a Gaussian beam (as in [2]). The attenuation of light in the sample is neglected (as in [4]). Because of the cylindrical symmetry a point in space can be denoted with two cylindrical coordinates, namely the distance r from the beam axis, and the distance z measured from the focus along the beam axis. The light intensity in point (r,z) at time instance t is

$$I(r,z,t) = I_0(t)\left(\frac{W_{E0}}{W_E(z)}\right)^2 \exp\left(-\frac{2r^2}{W_E^2(z)}\right) \quad (1)$$

where

$$W_E(z) = W_{E0}\sqrt{1+\left(\frac{z}{z_0}\right)^2}\,, \quad (2)$$

$$z_{E0} = \frac{\pi W_{E0}^2}{\lambda_{exc}}, \quad (3)$$

I. B. Juhász, "Simulation-based investigation of temporal and spatial characteristics of photodynamics in two-photon microscope," in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds. Faculty of Information Technology, Pázmány Péter Catholic University. Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 41-45.

and

$$W_{E0} = \frac{\lambda_{exc}}{\pi\theta} \tag{4}$$

is the waist radius, λ_{exc} is the wavelength of excitation light, θ is the beam divergence [5]. E in subscript refers to excitation. I_0 can be determined from the laser power as:

$$I(t) = \frac{2P(t)}{\pi W_{E0}^2}. \tag{5}$$

2) Pulse Parameters

The simulator uses the following laser properties as input parameters:

- wavelength (λ_{exc}),
- laser power in time average (P_{avg}),
- pulse length given as full width at half maximum (T_{pulse})
- time period between consecutive laser pulses (T_{rep}).

There are two pulse shapes implemented in the simulator: the square pulse and the Gaussian pulse [1]; the latter one is defined as

$$P_G(t) = P_{G,max} \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{6}$$

B. The Sample

1) Photodynamic states of fluorophore molecules

The sample is modeled as a homogenous solution of a fluorophore. The simulated volume is divided into cubic cells with a properly chosen edge length Δx. In each cell the number of fluorophore molecules is stored state by state. The states and state transitions are modeled by a three-level photodynamical model taken from [1], [3], [4] with modifications, including four different photobleaching routes. The model contains the following states (Fig. 1.):

- ground state S_0,
- excited state S_1,
- triplet state T_1,
- four photobleached states B_1, B_2, B_3, B_4.

In the model the following transitions are possible:

- $S_0 \rightarrow S_1$ during two-photon absorption,
- $S_1 \rightarrow B_1$ during one-photon absorption,
- $S_1 \rightarrow B_2$ during two-photon absorption,
- $T_1 \rightarrow B_3$ during one-photon absorption,
- $T_1 \rightarrow B_4$ during two-photon absorption,
- $S_1 \rightarrow S_0$ non-radiative relaxation (internal conversion),
- $S_1 \rightarrow S_0$ radiative relaxation followed by fluorescent photon emission,
- $S_1 \rightarrow T_1$ intersystem crossing,
- $T_1 \rightarrow S_0$ non-radiative relaxation.

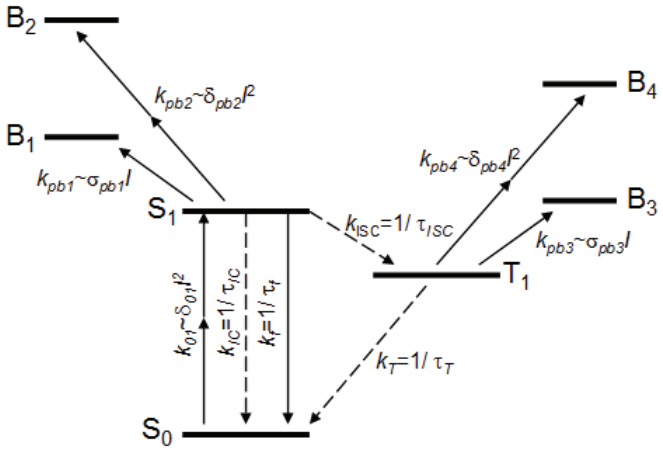


Figure 1. Simplified Jablonski diagram: photodynamic states and possible state transitions. S_0, S_1, T_1 denotes ground state, excited state and triplet state respectively; B_1, B_2, B_3, and B_4 denotes photobleached states. Solid lines refer to transitions with either photon absorption or emission; dashed lines denote transitions without participation of photon.

Transitions related to photon absorption are characterized by either one-photon or two-photon absorption cross sections (δ_{01}, σ_{pb1}, δ_{pb2}, σ_{pb3}, δ_{pb4} respectively) and depend on illumination intensity either linearly or quadratically (see (14) and (15)), while the rest of the transitions (hereunder relaxation transitions) are intensity-independent, and can be characterized by time constants τ_{IC}, τ_f, τ_{ISC}, τ_T respectively.

2) Transition Rates

a) Transitions during Laser Pulses

Let us assume that $T_{pulse} << \tau_{IC}, \tau_f, \tau_{ISC}, \tau_T$, thus during a laser pulse the relaxation transitions can be neglected, and we can assume that a fluorophore molecule is excited at most once per pulse [1], [2], [4]. During laser pulses the following transitions are simulated for each volume element (r,z) (from [3] with modifications):

$$\dot{S}_0(r,z,t) = -k_{01}(r,z,t)S_0(r,z,t) \tag{7}$$

$$\dot{S}_1(r,z,t) = k_{01}(r,z,t)S_0(r,z,t) - \left(k_{pb1}(r,z,t) + k_{pb2}(r,z,t)\right)S_1(r,z,t) \tag{8}$$

$$\dot{T}_1(t) = -\left(k_{pb3} + k_{pb4}\right)T_1(r,z,t) \tag{9}$$

$$\dot{B}_1(r,z,t) = k_{pb1}(r,z,t)S_1(r,z,t) \tag{10}$$

$$\dot{B}_2(r,z,t) = k_{pb2}(r,z,t)S_1(r,z,t) \tag{11}$$

$$\dot{B}_3(r,z,t) = k_{pb3}(r,z,t)T_1(r,z,t) \qquad (12)$$

$$\dot{B}_4(r,z,t) = k_{pb4}(r,z,t)T_1(r,z,t) \qquad (13)$$

Rate constant of a transition from the initial state i to final state f is

$$k_{if}(r,z,t) = \sigma_{if}\frac{\lambda_{exc}}{hc}I(r,z,t) \qquad (14)$$

if the transition is induced by one-photon absorption, and

$$k_{if}(r,z,t) = \frac{1}{2}\delta_{if}\left(\frac{\lambda_{exc}}{hc}\right)^2 I^2(r,z,t) \qquad (15)$$

if it is coupled to two-photon absorption, where σ_{if} and δ_{if} are the one-photon and two-photon absorption cross sections at wavelength λ_{exc} respectively, h is Planck's constant, c is velocity of light, whereas $I(r,z,t)$ is the illuminating light intensity at point (r,z) at time instance t [3]. Space and time dependence of the rate constants arises from the space and time dependence of the illumination intensity.

In a given (r,z) volume element every laser pulse is considered to be the same, defined by the pulse shape function, thus it is enough to calculate the rate constants once. By numerical integration of the above differential equations the transition probabilities in each volume element for a whole pulse can be obtained (cf. [2]), thus the transitions can be calculated in one step by a simple multiplication. When a molecule is excited from ground state to excited state, it is assumed to be instantly able to be photobleached, so transition probabilities for $S_0 \rightarrow B_1$ and $S_0 \rightarrow B_2$ transitions are also determined.

b) Transitions between Laser Pulses

Between laser pulses the following relaxation transitions take place (from [3] with modifications):

$$\dot{S}_0(t) = \left(k_{IC} + k_f\right)S_1(t) + k_T T_1(t), \qquad (16)$$

$$\dot{S}_1(t) = -\left(k_{IC} + k_f + k_{ISC}\right)S_1(t), \qquad (17)$$

$$\dot{T}_1(t) = k_{ISC}S_1(t) - k_T T_1(t) \qquad (18)$$

where the rate constants are $k_{IC} = 1/\tau_{IC}$, $k_f = 1/\tau_f$, $k_{ISC} = 1/\tau_{ISC}$, $k_T = 1/\tau_T$. By solving these differential equations the transition probabilities for one laser cycle can be obtained again.

3) Diffusion

Diffusion of fluorophore molecules is modelled in the following way: In each laser cycle the probability that a molecule moves to the neighbouring volume element along one of the coordinate axes during T_{rep} time is

$$p_{diff} = \frac{DT_{rep}}{(\Delta x)^2} \qquad (19)$$

where D is the diffusion coefficient and Δx is the size of the volume cells. It can be shown, that after large number of steps, the evolving distribution agrees with the distribution derived from Fick's law.

C. Photon Detection Probability

The probability that a fluorescent photon emitted from volume element at (r,z) is detected is

$$p_{det}(r,z) = \eta\Phi_0 Y_0(r,z), \qquad (20)$$

where η is the efficiency of the detector, Φ_0 is the fractional solid angle of observation, and $Y_0(r,z)$ is the observation beam profile that describes the space dependence of detection probability [2]. In case of confocal observation $Y_0(r,z)$ is a Gaussian beam profile (see (1)-(4)), and $\Phi_0 = \lambda_{emit}^2/(4\pi^2 n^2 W_{O0}^2)$, where λ_{emit} is the wavelength of emitted light, n is the refractive index of the sample, and W_{O0} is the waist radius of the Gaussian beam for observation. In case of full aperture detection $\Phi_0 = NA^2/(4n^2)$ and $Y_0(r,z)=1$ [2].

III. Simulations and Results

Example simulations were run to illustrate the capabilities of the model. In order to get realistic results, fluorophore parameters were set to be in the same order of magnitude that data found in [3], [4] for real fluorophores. Unless stated otherwise the following parameters were used: $\tau_f = 3\cdot10^{-9}$ s, $\tau_{ISC} = 10^{-6}$ s, $\tau_{IC} = \infty$ s i.e. internal conversion was disabled, $\tau_T = 2.5\cdot10^{-6}$ s, $D = 2\cdot10^{-10}$ m^2/s, $\delta_{01} = 2\cdot10^{-49}$ cm^{4}s, $\sigma_{pb1} = 5\cdot10^{-23}$ cm^2; $\delta_{pb2} = 0$ cm^{4}s, $\sigma_{pb3} = 0$ cm^2, $\delta_{pb4} = 0$ cm^{4}s, i.e. these three photobleaching pathways were disabled; Gaussian pulse shape with full width at half maximum of $T_{pulse} = 100$ fs was used, pulse repetition rate was set to 80 MHz ($T_{rep} = 1.25\cdot10^{-8}$s); beam divergence was 60°.

The power dependence of two-photon excitation probability ($S_0 \rightarrow S_1$ transition) in the focus was examined (Fig. 2.). As it can be seen, excitation saturates at about 40 mW laser power. Further increase of laser power does not enhance fluorescence in the focus, but it raises photobleaching and expands the excitation volume (Fig. 3.) thus worsens image quality.

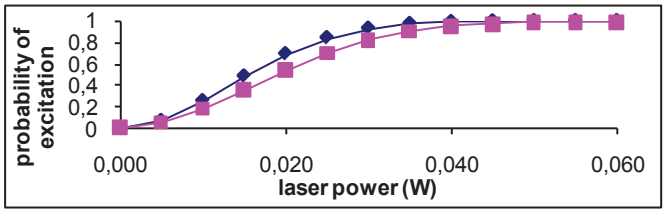


Figure 2. Laser power dependence of excitation probability in the focus. Dependence of two-photon excitation probability in the focus on laser power is plotted in case of square (blue) and Gaussian (purple) pulse shape.

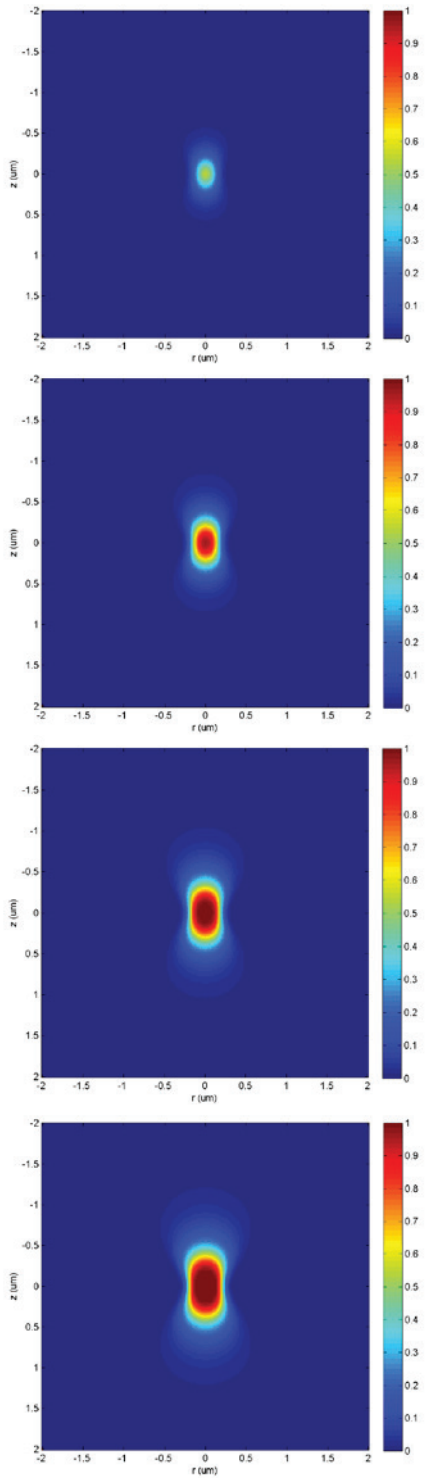


Figure 3. Local probability of two-photon excitation ($S_0 \rightarrow S_1$ transition) in case of different laser powers. Applied laser powers were 20 mW, 40 mW, 60 mW, 80 mW (from top to bottom). Note the expandation of excitation volume as laser power increases.

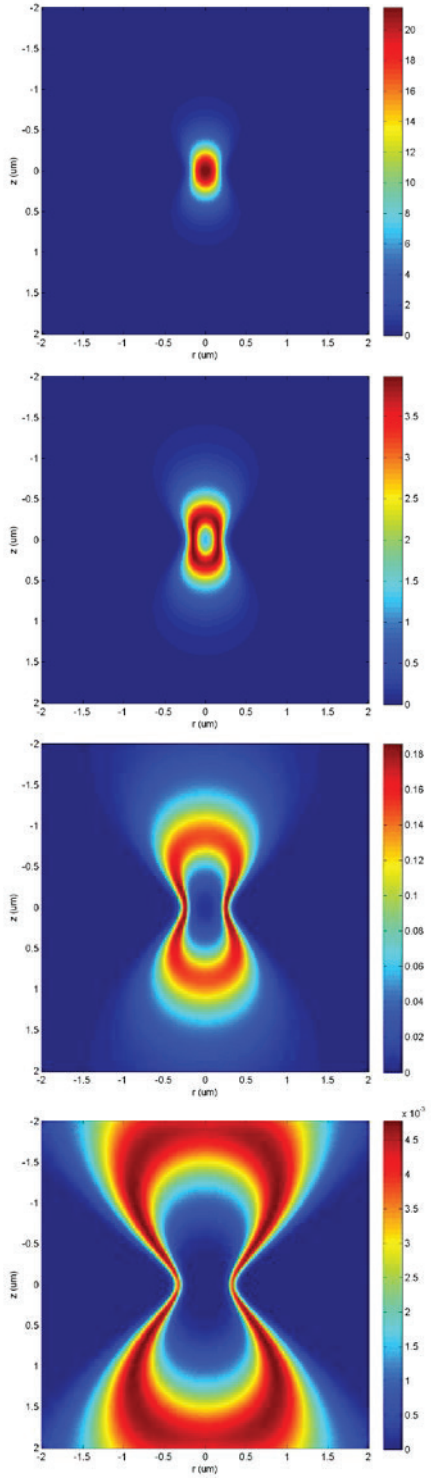


Figure 4. Cumulated number of emitted photons per volume element during 100 μs illmunition in case of different photobleaching probabilities. Absorption cross section σ_{pb1} characterizing $S_1 \rightarrow B_1$ transition was set to the following values: $5 \cdot 10^{-23}$ cm^2, $5 \cdot 10^{-21}$ cm^2, $5 \cdot 10^{-19}$ cm^2, $5 \cdot 10^{-17}$ cm^2 (from top to bottom). Note the 'hole' around at the focus due to high local photobleaching as well as the expansion of fluorescent volume, and decrease of the photon number in case of increasing absorption cross section σ_{pb1}.

In the second experiment, the photodynamic property of the fluorophore was changed, namely the absorption cross section σ_{pb1} characterizing $S_1 \rightarrow B_1$ photobleaching transition (Fig. 4.).

As the rate of photobleaching increases, the molecules around the focus are photobleached. As a result the number of emitted photons decreases, and their source (along an 8-like shape) is getting farther from the focus, therefore the resolution of the microscope (thus image quality) deteriorates.

IV. Conclusion

The above presented simulator is able to quantify the photodynamic processes including photobleaching in a two-photon microscope, therefore it might help to find better microscope operating parameters for higher image quality. Also in the reverse direction, the simulator might be useful when photodynamical parameters are searched for. In the first approach however, the lack of the state transition parameters can be a limiting factor.

V. References

[1] J. Mertz, "Molecular photodynamics involved in multi-photon excitation fluorescence microscopy," *Eur. Phys. J. -At. Mol. Opt. Plasma Phys.*, vol. 3, no. 1, pp. 53–66, 1998.

[2] Martin Kauert, Patrick C. Stoller, Martin Frenz, and Jaro Rička, "Absolute measurement of molecular two-photon absorption cross-sections using a fluorescence saturation technique," *Opt. Express*, vol. 14, no. 18, Sep. 2006.

[3] C. Eggeling, A. Volkmer, and C. A. M. Seidel, "Molecular Photobleaching Kinetics of Rhodamine 6G by One- and Two-Photon Induced Confocal Fluorescence Microscopy," *ChemPhysChem*, vol. 6, no. 5, pp. 791–804, May 2005.

[4] R. Niesner, W. Roth, and K.-H. Gericke, "Photophysical Aspects of Single-Molecule Detection by Two-Photon Excitation with Consideration of Sequential Pulsed Illumination," *ChemPhysChem*, vol. 5, no. 5, pp. 678–687, May 2004.

[5] Bahaa E. A. Saleh and Malvin Carl Teich, *Fundamentals of Photonics*. New York: John Wiley & Sons, Inc., 1991.

Integrated microcapillary system for microfluidic parasite analysis

András J. Laki
(Supervisors: Kristóf Iván Ph.D., Pierluigi Civera Ph.D., Dr. Éva Fok)
laki.andras@itk.ppke.hu

***Abstract*—We present the use of a simple microfluidic technique to detect living parasites from veterinarian blood using a monolithic polydimethylsiloxane (PDMS) structure. Several intravenous parasitosis can be observed by this developed microcapillary system such as dirofilariasis or Lyme disease. Inside this microfluidic device a special flow-through separator structure has been implemented, which contains a cylindrical Active Zone, where the microfilariae or other few micron-size parasitic infections remain trapped. The center region is partially surrounded by rectangular cross-section shaped microcapillaries. The developed test can be optimized for a specific nematode or parasite detection by changing the capillary width.**

I. Introduction

Present-day requirements of biomedical engineering comprehend real-time analysis, portable medical attendance and low-cost throughways tests. There is a strong desire to use fast-tests before an expensive laboratory examinations. Plenty of clinical fast-tests exist that are well-known and prevalent such as urinalysis, blood-sugar test, pregnancy test, etc. Results of medical examinations are instantaneous or near real-time, low-cost and approximately responsible. Generally clinical fast-tests, which insure examination from one well-defined measurement, require significant sample volume, thus more ampules of blood sample can be necessary from one patient to make a complex analysis. Due to the use of Micro Total Analysis Systems (μTAS), which is inspired to integrate different medical tests into one unique device, this required blood volume can be reduced to the volume of a droplet.

Due to the emergence of cardiopulmonary parasitoses a stand-alone, laboratory-independent microfluidic detector has been developed. A flow-through microfluidic filter has been designed applying a special microcapillary structure. The diameter of the microcapillaries has the same geometric parameters within one device. 48 different microfluidic channel system has been created with different microcapillary width in the range of $6.1\mu m$ up to $83.6\mu m$. The monolithic polydimethylsiloxane (PDMS) microfluidic structures have been developed to detect intravenous few micron-size parasitic infections form a blood sample. The fabrication of the constructed devices are based on soft-lithography techniques. Previously finite element calculations have been made by Comsol Multiphysics software to optimize the velocity and pressure profile of liquid flow solving the Navier-Stokes equations.

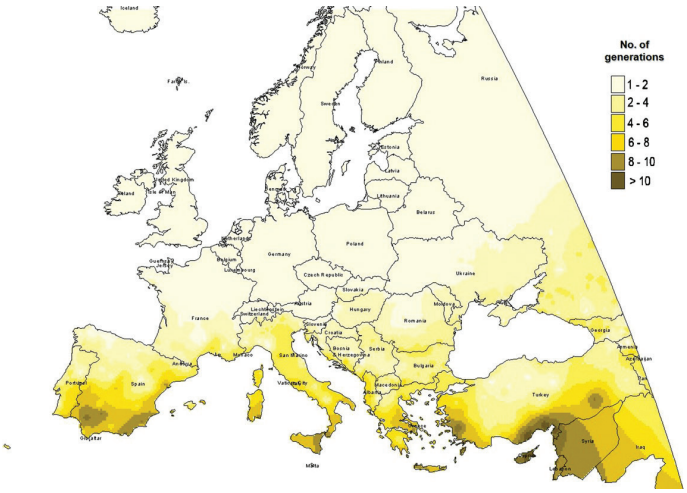


Fig. 1. Emergence of cardiopulmonary parasitoses in Europe. Yearly average predicted number of Dirofilaria generations obtained by Linear Kriging interpolation [1]

A. Emergence of cardiopulmonary parasitoses

Nematodes affecting cardiopulmonary system has recently become the focus of attention due to the increasing number of veterinarian infections cased by heartworm (*Angiostrongylus vasorum* and *Dirofilaria immitis*) and lungworm (*Aelurostrongylus abstrusus*, *Crenosoma vulpis* and *Eucoleus aerophilus*) parasites. The reasons for the apparent emergence of cardiopulmonary parasitoses in pets are unknown but several factors such as global warming, changes in vector seasonal population dynamics and movements in animal populations, may play a role in the recent rise in reports of infection in the various countries [2]. Each fifth pet is infected by parasitoses in the Mediterranean region, Southern and Eastern Europe, North Africa and Asia. The yearly average predicted number of Dirofilaria generations obtained is shown on Fig. 1. Also human parasitoses are well-known from several case histories [3].

This project takes aim at a realization of a laboratory-independent detection method for dirofilariasis. The nematodes of genus *Dirofilaria* belong to family *Onchocercidae* and subfamily *Dirofilariinae* of the order *Spirurida*. The genus *Dirofilaria* collect several species including *Dirofilaria immitis* and *Dirofilaria repens*. The species of Dirofilaria genus are slender, long filarial worms, which lead into oesophagus

A. J. Laki, "Integrated microcapillary system for microfluidic parasite analysis,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 47-50.

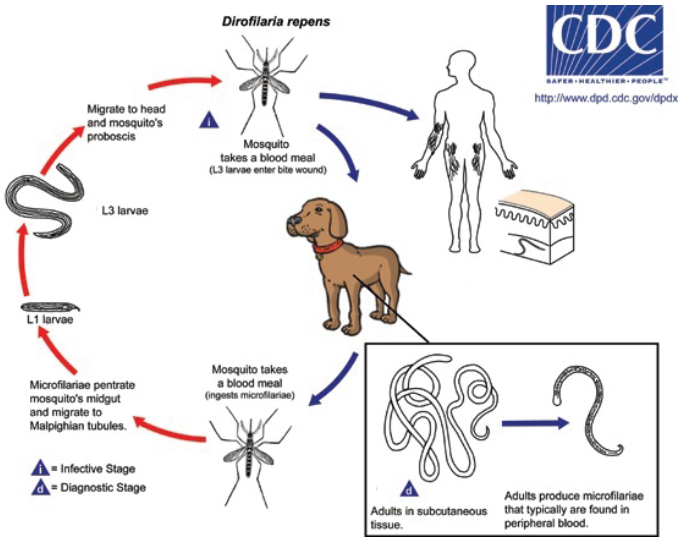


Fig. 2. The life cycle of Dirofilaria repens

differentiated into muscular and glandular regions without distinction. The length of adult females can elongate 250 to $310mm$ meanwhile the its diameter is around 1 to $1.3mm$. The adult males can reach 120 to $200mm$ length and 0.7 to $0.9mm$ width. These nematodes are ovoviviparous and the evolving unsheathed microfilariae live in bloodstream. The length of its larvae is 290 to $330\mu m$ and its width is $5-7\mu m$ [4].

The life cycle of species of Dirofilaria genus consists of five larval stages in vertebral host and an arthropod (mosquito) intermediate host and vector. The development period of the microfilariae mainly depends on the temperature inside the species of intermediate host (from 10 to 21 days at around 25°C). In infective stage the larvae (in stage L2) migrate to the Malpighian tubule lumen of the mosquito, while during subsequent nutrition of the intermediate host the larvae enter to subcutaneous connective tissue of definitive host. 48 hours after inoculation infective larvae (in stage L3) can be found in subcutaneous tissue, 70 days later nematodes (in stage L4) move in the muscle and subcutaneous tissues. 100 days later the larvae (in stage L5) enter to the thoracic and abdominal cavities. The gravid females of the D. immitis can be discovered in pulmonary arteries and in the right heart 180 days after the inoculation of the definitive host. The adult D. repens, which mainly aries from the D. immitis in length, remain in the subcutaneous tissues and its life cycle is represented on Fig. 2.

Several diagnostic methods have been developed to explore the existence of intravenous nematodes or to determine its volumetric population from serological samples. The following enumeration represents a scale of executive complexity in inverse proportion to currently used diagnostic methods [1]: fresh blood smear, modified Knott test, filter test, histochemical stain based test, Enzyme-linked Immunosorbent Assay (ELISA), Immunochromatographic tests, antibody tests and Polymerase Chain Reaction (PCR).

II. Types of microfluidic separator

Microfluidic separation techniques can be divided into two fields: active, which requires external forces or passive techniques. The active separation can be categorized into acoustophoresis [5]–[8], chemophoresis [9], electrophoresis [10]–[12], magnetophoresis [13], uses mechanical forces [14] and optophoresis [15].

Without external active forces the particle separation, which is mainly based on changing the geometry of the channels, is called as hydrophoresis, which is classified into subclasses: Batch Separation Procedures (BSP) such as Hydrodynamic Chromatography (HC) [16] and Continuous-Flow Separation Procedures (CFSP). The sample loading method is the main difference between these two passive procedures. The BSP works with quantized inlet volumes, meanwhile the CFSP loads sample continuously. The CFSP procedure is distinguished subclasses: using centrifugal extraction [17], Dean flow in cylindric channels [18], Determinisitc Cell Rolling (DCR) [19], Deterministic Lateral Displacement (DLD) [20], using the elasto-inertial effect [21], flow through separators [22], hydrocyclones [23], using membrane [24], Pinched-Flow Fraction (PFF) [25] or using the Zweifack-Fung effect [26].

III. Integrated microcapillary system for microfluidic parasite analysis

The developed flow-thought nematode filter (FTNF) is based on a common microfluidics-based particle separation technique, easy to implement in cheap disposable plastic chips, that we believe is well suited for the task of removing parasites from blood cells in order to aid detection. The mechanism of separation by FTNF is based on the interaction of nematodes suspended in whole blood with an ordered array of microcapillary system that the fluid is forced to flow through under low Reynolds number conditions, meanwhile the detectable larvae are trapped before.

A. Computational Fluid Dynamics (CFD) simulations

The designed flow-through microseparators had been geometrically optimize by the velocity and pressure profiles. A Computational Fluid Dynamics (CFD) solver was used to numerically calculate the Navier-Stokes equations (Eq. 1).

$$\rho\left(\frac{Dv_i}{Dt}+v\nabla v\right)=-\nabla p+\mu\Delta v+F_i \qquad (1)$$

where Dv_i/Dt in the unsteady, while $v\nabla v$ is the convective acceleration. These forces came form the property of volume, while the other part of the equation is the divergence stress and other body forces (F_i). The sum of the pressure gradient (∇p) and the viscosity ($\nu\Delta v$) is the divergence of stress.

1) Geometrical description of the capillary system: The geometry parameters of this microfluidic separator is shown on Fig. 3. Microfluidic channels are $20\mu m$ high and the width of the input and output channel is $400\mu m$. The radius of the central unit is $1.2mm$, the radius of the active zone (where parasites remain during the filtration) is $1mm$. The internal cylindric piers, which columns are holding the top of the active

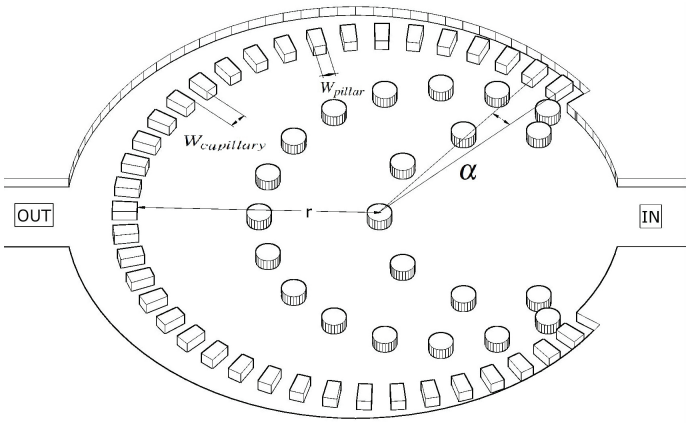


Fig. 3. Geometric description of the developed flow-thought nematode filter. The α angle is the structural repetition of the microcapillaries, r is the radius of the active zone, W_{pillar} is the width of the pillars, $W_{capillary}$ is the width of the capillary channel.

zone, are in a polar array with radius of $0.5mm$ and $0.7mm$ and its diameter is $100\mu m$. The active zone is surrounded by a layer of well-defined microcapillaries. The repetition angel between the microchannels is α. The following trigonometrical equations describe the connection between the α and the deterministic diameter.

$$sin\frac{\alpha}{2} = \frac{W_{pillar} + W_{capillary}}{2r} \tag{2}$$

$$W_{capillary} = 2rsin\frac{\alpha}{2} - W_{pillar} \tag{3}$$

$$\alpha = 2sin^{-1}\left[\frac{W_{pillar} + W_{capillary}}{2r}\right] \tag{4}$$

where the width of the pillars (W_{pillar}) is the same ($52.8\mu m$) in each device. The radius of the all central unit is $1.2mm$, while the radius (r) of the active zone is $1mm$. The internal cylindric piers, which columns are holding the top of the active zone, are in a polar array with radius of $0.5mm$ and $0.7mm$ and its diameter is $100\mu m$. The cross-section width of microcapillary channels ($W_{capillary}$) is identical within one device, but each has different width from $6.1\mu m$ up to $83.6\mu m$. The width of the microcapillary channels is correlated with α , which is described by Eq. 4. Those rigid particles, which have greater diameter than $W_{capillary}$, will be filtered out from the liquid flow. All amount of sample is pushed through this structure thus this structure could be a efficient method for parasite detection.

2) Velocity and pressure profile of the developed structure: The all domain of microfluidic channel is filled by a Non-newtonian fluid using the average parameters of the intravenous blood stream. The viscosity is $3.53 \cdot 10^{-3} Pa\ s$, the density is $1060kg/m3$. The velocity and the pressure profile were calculated at constant room temperate ($20°C$) and constant external pressure ($101kPa$).

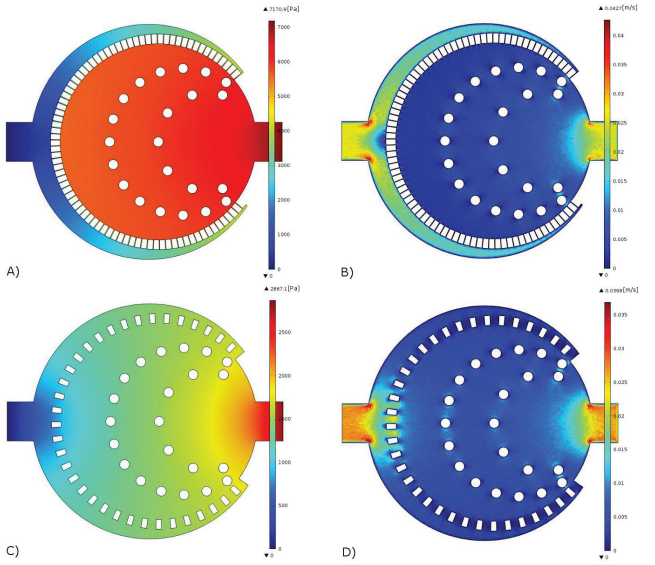


Fig. 4. Pressure and velocity profiles at same initial conditions (mean flow velocity on the inlet is $0.02m/s$, viscosity is $3.53 \cdot 10^{-3} Pa\ s$, density is $1060kg/m3$). A) Pressure profile inside the thinnest capillary system (α is $3.4°$), where pressure drop is $7190Pa$. B) Velocity profile (α is $3.4°$), where the maximum value is $0.042m/s$. C) Pressure profile inside the thickest capillary system (α is $8°$), where pressure drop is $2878Pa$. D) Velocity profile (α is $8°$), where the maximum value is $0.0389m/s$.

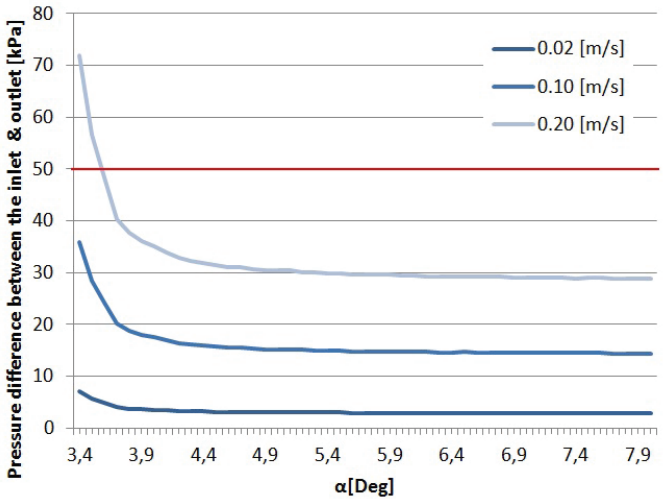


Fig. 5. Pressure difference between inlet and outlet in the function of α angle at fix inlet velocity ($0.02m/s$, $0.1m/s$, $0.2m/s$)

The pressure drop is a critical physical parameter of a filter structure. If the pressure is significant the trapped flexible particles can be squeezed through the microcapillary structure, while using and abnormal pressure the filter can be destroyed. The two boundary cases of the velocity and the pressure profile of the developed structure is shown on Fig. 4 at same inlet velocity ($0.02m/s$). The pressure difference between inlet and outlet in the function of α angle is represented on Fig. 5. The decreasing value of $W_{capillary}$ increases the pressure drop within the device.

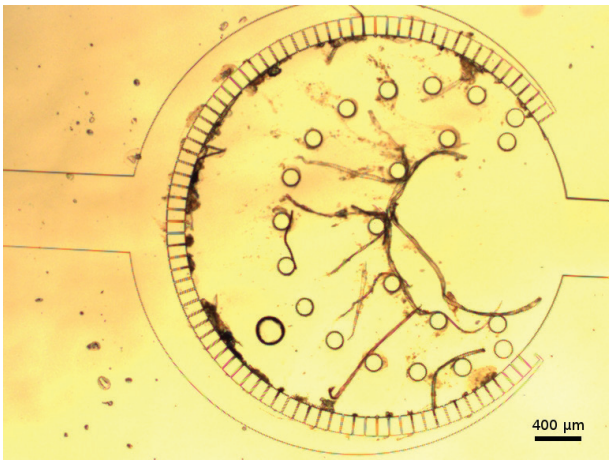


Fig. 6. A result of veterinarian measurement using the developed flow-thought nematode filter

IV. Conclusion

Pressure and velocity profile have been calculated to predict the pressure drop to secure the efficiency of the developed device. We have successfully shown how intravenous nematodes can be detected using the developed flow-thought nematode filter. 48 different microfluidic devices have been designed, fabricated and tested to uncover dirofilariasis from veterinarian blood samples.

V. Acknowledgment

I would like to thank Olga Jacsó for the biological samples and her kind help. I acknowledge Péter Fürjes and Zoltán Fekete for their kind help in the device fabrication. The support of grants TÁMOP-4.2.1.B-11/2/KMR-2011-0002 and TÁMOP-4.2.2/B-10/1-2010-0014 is gratefully acknowledged.

References

[1] C. Genchi, L. Venco, and M. Genchi, "Guideline for the laboratory diagnosis of canine and feline dirofilaria infections," *Mappe Parassitologiche*, pp. 139–144, Feb. 2007.

[2] D. Traversa, A. Di Cesare, and G. Conboy, "Canine and feline cardiopulmonary parasitic nematodes in europe: emerging and underestimated," *Parasites & Vectors*, vol. 3, no. 1, p. 62, Jul. 2010.

[3] L. Rinaldi, V. Musella, C. Genchi, and G. Cringoli, "Geographical Information Systems in health applications: experience on filariosis," *Mappe Parassitologiche*, pp. 19–38, Feb. 2007.

[4] G. Cancrini and S. Gabrielli, "Vectors of dirofilaria nematodes: biology, behaviour and host/parasite relationships," *Mappe Parassitologiche*, pp. 47–58, Feb. 2007.

[5] J. Shi, S. Yazdi, S.-C. S. Lin, X. Ding, I.-K. Chiang, K. Sharp, and T. J. Huang, "Three-dimensional continuous particle focusing in a microfluidic channel via standing surface acoustic waves (SSAW)," *Lab on a Chip*, vol. 11, no. 14, pp. 2319–2324, Jun. 2011.

[6] T. Laurell, F. Petersson, and A. Nilsson, "Chip integrated strategies for acoustic separation and manipulation of cells and particles," *Chem. Soc. Rev.*, vol. 36, no. 3, pp. 492–506, Feb. 2007.

[7] J. Nam, H. Lim, D. Kim, and S. Shin, "Separation of platelets from whole blood using standing surface acoustic waves in a microchannel," *Lab Chip*, vol. 11, no. 19, pp. 3361–3364, Sep. 2011.

[8] J. Shi, H. Huang, Z. Stratton, Y. Huang, and T. J. Huang, "Continuous particle separation in a microfluidic channel via standing surface acoustic waves (SSAW)," *Lab on a Chip*, vol. 9, no. 23, pp. 3354–3359, Dec. 2009.

[9] T. Kim, L.-J. Cheng, M.-T. Kao, E. F. Hasselbrink, L. Guo, and E. Meyhfer, "Biomolecular motor-driven molecular sorter," *Lab on a Chip*, vol. 9, no. 9, pp. 1282–1285, May 2009.

[10] M. J. Hilhorst, G. W. Somsen, and G. J. de Jong, "Capillary electrokinetic separation techniques for profiling of drugs and related products," *ELECTROPHORESIS*, vol. 22, no. 12, pp. 2542–2564, Jul. 2001.

[11] A. Valero, T. Braschler, A. Rauch, N. Demierre, Y. Barral, and P. Renaud, "Tracking and synchronization of the yeast cell cycle using dielectrophoretic opacity," *Lab Chip*, vol. 11, no. 10, pp. 1754–1760, May 2011.

[12] Y. Li, C. Dalton, H. J. Crabtree, G. Nilsson, and K. V. I. S. Kaler, "Continuous dielectrophoretic cell separation microfluidic device," *Lab on a Chip*, vol. 7, no. 2, pp. 239–248, Jan. 2007.

[13] A. I. Rodrguez-Villarreal, M. D. Tarn, L. A. Madden, J. B. Lutz, J. Greenman, J. Samitier, and N. Pamme, "Flow focussing of particles and cells based on their intrinsic properties using a simple diamagnetic repulsion setup," *Lab Chip*, vol. 11, no. 7, pp. 1240–1248, Apr. 2011.

[14] G. H. Kwon, Y. Y. Choi, J. Y. Park, D. H. Woo, K. B. Lee, J. H. Kim, and S.-H. Lee, "Electrically-driven hydrogel actuators in microfluidic channels: fabrication, characterization, and biological application," *Lab on a Chip*, vol. 10, no. 12, pp. 1604–1610, Jun. 2010.

[15] K. H. Lee, S. B. Kim, K. S. Lee, and H. J. Sung, "Enhancement by optical force of separation in pinched flow fractionation," *Lab Chip*, vol. 11, no. 2, pp. 354–357, Jan. 2011.

[16] A. W. Browne, L. Ramasamy, T. P. Cripe, and C. H. Ahn, "A lab-on-a-chip for rapid blood separation and quantification of hematocrit and serum analytes," *Lab on a Chip*, vol. 11, no. 14, pp. 2440–2446, Jun. 2011.

[17] S. Haeberle, T. Brenner, R. Zengerle, and J. Ducre, "Centrifugal extraction of plasma from whole blood on a rotating disk," *Lab on a Chip*, vol. 6, no. 6, pp. 776–781, May 2006.

[18] A. A. S. Bhagat, S. S. Kuntaegowdanahalli, and I. Papautsky, "Continuous particle separation in spiral microchannels using dean flows and differential migration," *Lab on a Chip*, vol. 8, no. 11, pp. 1906–1914, Nov. 2008.

[19] S. Choi, J. M. Karp, and R. Karnik, "Cell sorting by deterministic cell rolling," *Lab on a Chip*, vol. 12, no. 8, pp. 1427–1430, Apr. 2012.

[20] H. N. Joensson, M. Uhln, and H. A. Svahn, "Droplet size based separation by deterministic lateral displacementseparating droplets by cell-induced shrinking," *Lab on a Chip*, vol. 11, no. 7, pp. 1305–1310, Apr. 2011.

[21] J. Nam, H. Lim, D. Kim, H. Jung, and S. Shin, "Continuous separation of microparticles in a microfluidic channel via the elasto-inertial effect of non-newtonian fluid," *Lab on a Chip*, vol. 12, no. 7, pp. 1347–1354, Mar. 2012.

[22] J. S. Shim and C. H. Ahn, "An on-chip whole blood/plasma separator using hetero-packed beads at the inlet of a microchannel," *Lab on a Chip*, vol. 12, no. 5, pp. 863–866, Feb. 2012.

[23] P. Bhardwaj, P. Bagdi, and A. K. Sen, "Microfluidic device based on a micro-hydrocyclone for particle-liquid separation," *Lab on a Chip*, vol. 11, no. 23, pp. 4012–4021, Nov. 2011.

[24] H. Wei, B.-h. Chueh, H. Wu, E. W. Hall, C.-w. Li, R. Schirhagl, J.-M. Lin, and R. N. Zare, "Particle sorting using a porous membrane in a microfluidic device," *Lab on a Chip*, vol. 11, no. 2, pp. 238–245, Jan. 2011.

[25] A. A. S. Bhagat, H. W. Hou, L. D. Li, C. T. Lim, and J. Han, "Pinched flow coupled shear-modulated inertial microfluidics for high-throughput rare blood cell separation," *Lab on a Chip*, vol. 11, no. 11, pp. 1870–1878, Jun. 2011.

[26] O. Forouzan, J. M. Burns, J. L. Robichaux, W. L. Murfee, and S. S. Shevkoplyas, "Passive recruitment of circulating leukocytes into capillary sprouts from existing capillaries in a microfluidic system," *Lab on a Chip*, vol. 11, no. 11, pp. 1924–1932, Jun. 2011.

Flow-through functionalized PDMS microfluidic device for sandwich ELISA

Gábor Zsolt Nagy
(Supervisor: Kristóf Iván Ph.D.)
nagy.gabor.zsolt@itk.ppke.hu

***Abstract*—In this research I try to design a flow-through functionalized PDMS-glass microfluidic device, which is able to detect specified antigens from blood, wine or urine. The surface of PDMS channel is going to be modified for sandwich enzyme-linked immunosorbent assay (ELISA). For detection I plan to use a spectroscopical method. Three models have been designed and tested in a flow simulation program to choose an optimal geometry for fabrication.**

***Keywords*-PDMS; flow-through; ELISA; microfluidic channel**

I. INTRODUCTION

Infectious diseases - caused by viruses, bacteria, parasites or fungi – and different kinds allergies can be cured effectively in the 21st century [1]. We need to detect and diagnose the source of the cause as early and as properly as it is possible. The immunoassay tests - such as ELISA (Enzyme-linked immunosorbent assay) – are potential methods for the proper detection and screening. ELISA is widely used in medical diagnostics to detect proteins based on their binding to immobilized antibodies. For example in sandwich ELISA there are two kinds of antibodies: primary antibody - which is immobilized on the solid surface and binds the antigen from the sample - and the secondary antibody – which is labeled with an enzyme for detection. Through the binding of antigen-antibodies and the conversion of labeling enzyme we can measure the amount of the antigen in the sample quantitatively. The problem with ELISA, that this assay is slow (several hours), requires large volumes of sample and reagent and can be performed only in laboratories [2]. Point-Of-Care and Lab-On-A-Chip technologies are rapidly improving areas of biomedical and medical diagnostics. These microfluidics based technologies need small volumes of reagents and sample, and delivery of results with fast turnaround time [3]. It is a great challenge to transform immune assays from microplates into microfluidic devices reducing the costs and operation time.

II. TYPES OF ELISA SOLUTIONS IN MICROFLUIDIC CHANNELS

A. *Microbeads in microfluidic channel*

Microbeads are used for concentrating specific molecules and increasing the available surface for the immune complex. The microbead made of different materials: polystyrene, glass[10], magnetical material. Polystyrene microbeads are coated with capture antibodies through physisorption or chemisorption and introduced to the fabricated microchannel. A dam structure in the channel functions to stop the microbeads from escaping. At first, the sample with the antigens is introduced into the channel. The coated antibodies capture the antigens. Enzyme-conjugated secondary antibodies are then introduced and capture the antigens. Finally, substrates are applied and dye molecules produced by the enzymatic reaction [4],[6]. The microbeads can be placed in different parts of the microfluidic channel depending on where the dam structures are placed. A sample model (Fig. 1) shows the microbeads in the channel [8]. Magnetic microbeads can be manipulated by magnetic field with permanent magnets, electromagnets. Changing the magnetic field around the microfluidic channel, the beads can be repositioned in the channel from A point to B and this can help the connection of antigens and antibodies [9].

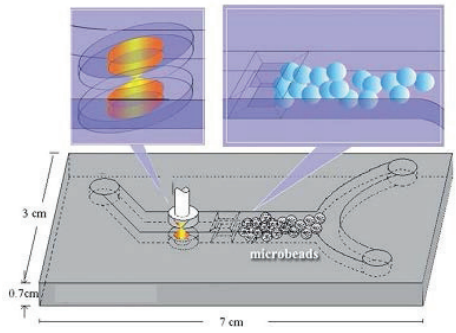


Fig. 1: Microbeads in microfluidic channel [8]

B. *Miniature microplate using LOM technology*

This device is a miniature 96 sample ELISA-lab-on-a-chip. This multilayered microfluidic channel constructed using Laminated Object Manufacturing (LOM). It is made from six acryl(poly(methyl methacrylate) core and five polycarbonate layers. It utilizes large surface area of nanotubes for increasing the surface for the immunocomplex. The antibodies are electrostatically adsorbed onto the carbon nanotubes. [5]

G. Zs. Nagy, "Flow-through functionalized PDMS microfluidic device for sandwich ELISA,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 51-54.

C. *Flow-through functionalized channel*

In a flow-through functionalized microfluidic device the channel wall (partly or full length) will be the surface for the immobilization of antibodies. PDMS (polydimethylsiloxane) is widely used as channel material for flow through functionalization. It is biocompatible and its surfaces can be modified for ELISA. There are three ways to immobilize antibodies to the PDMS wall:

- Passive adsorption: the PDMS wall binds protein passively so it can bind the capture antibodies without any modifications. [7]
- Site-selective binding to immobilized protein: by a chemical adsorption a selected protein bounds to the wall, and the capture antibody binds to this immobilized protein [7]
- Chemical adsorption: through oxidation processes aldehyde groups are produced (OP1), the PDMS wall is oxidized (OP2) and they bounds through Shiff bases (Fig. 2). The capture antibody binds covalent to the created layer [1].

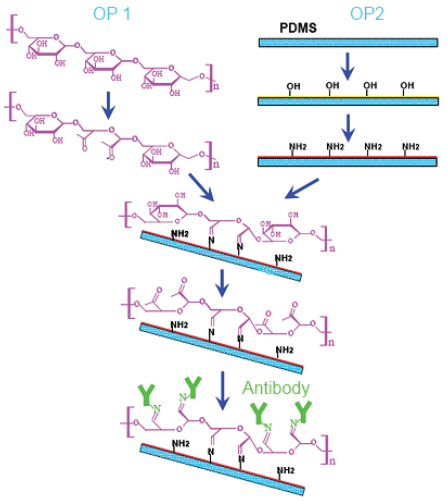


Fig. 2: Schematics of chemical adsorption flow-through modification of PDMS microchannel [1]

III. MODELS AND FINITE ELEMENT SIMULATIONS

The aim is to create flow-through functionalized PDMS-glass microfluidic device, which is used to detect specified antigens from blood, wine or urine. The first step is to design the geometry of the microfluidic channel. I created three models in AutoCAD software. All of the models have the same height (20 µm) and width (200 µm). The 1:10 ratio is an important factor for the later fabrication method. In the first model I created a simple channel without any additional structures in it. In the second and the third models the channel is divided into three parts, and in the middle area two kinds of objects were placed. I used round structures in the second one, squared structures in the third one.

After designing I tested the channels in Comsol flow simulation software. In the starting conditions the followings has been set: the number, quality and mass of testing particles; inlet pressure; the mean flow rate on the channel's inlet; testing fluid's quality, type of flow; time stepping settings. Running the simulation a velocity profile, pressure profile and a particle tracing profile in the microfluidic channel has been obtained. The particle tracing shows the particle's trajectories in the channel. This is an important information, because it makes calculable the particles (antibodies) binding probability to a given surface.

The three models have three different profiles in all three measured categories. In all models I used laminar flow.

In the first model (Fig. 3) the velocity is increasing continuously to the center of channel, where it reaches 0,6-0,7 m/s in the red area. In the particle tracing (Fig. 6) only 2-3 particles stucked to the wall, the others flowed through the channel. This caused by difference of velocities. The particles get out from the main flow towards the wall only with small probabilities.

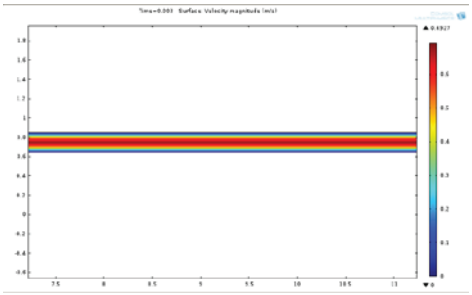


Fig. 3: Velocity profile in the channel without additional structures

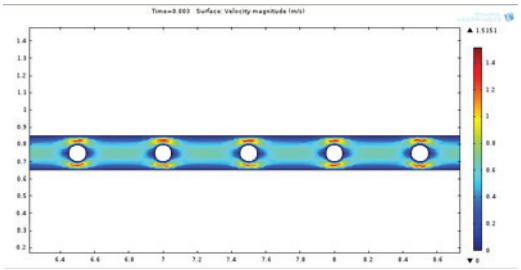


Figure 4: Velocity profile in the channel with round structures

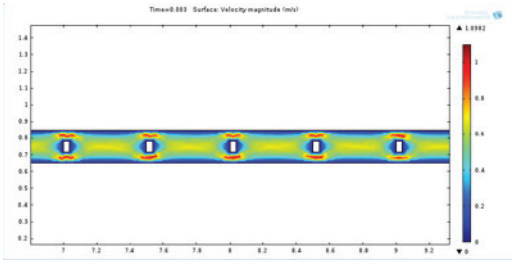


Fig. 5: Velocity profile in the channel with squared structures

In the second model the circles modify the velocity in the center flow (Fig. 4). Between the circles the velocity is mostly 0,6 m/s. The highest flow rate, 0,9-1 m/s, is between the channel wall and the circle. By the inserted circles the cross-section decreased and the flow rate increased at these points. The particles' trajectories also changed (Fig. 7). 4-6 particles stucked on the surface of the circle from the direction of inlet, and more particles reached the wall. Some of the particles reach a circle, but they bounce back into the wall.

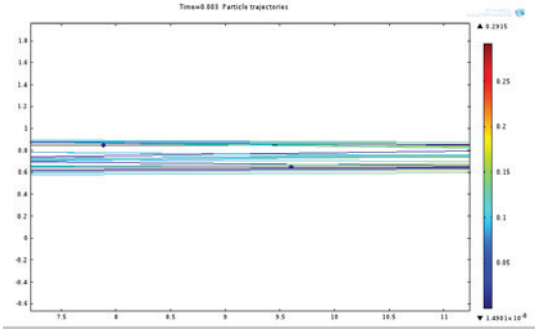


Fig. 6: Particle trajectories in the channel without additional structures

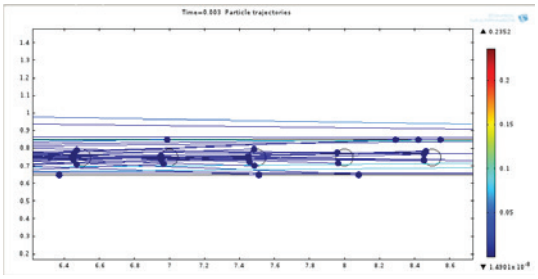


Fig. 7: Particle trajectories in the channel with round structures

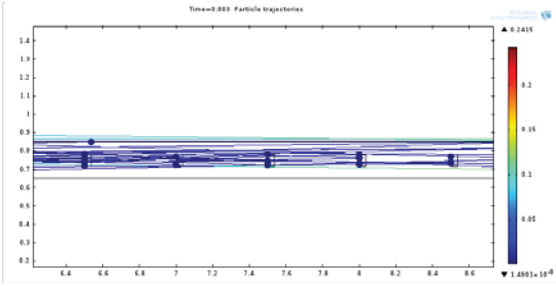


Fig. 8: Particle trajectories in the channel with squared structures

The third model contains squared structures, which modified the velocity different way than the circles (Fig. 5). The highest flow rate also between the structure and wall (0,9-1 m/s). But the flow rate in the center of the channel is higher (0,6-0,7 m/s). In the particle tracing (Fig. 8) 4-5 particles stuck on the surface of rectangles, from the direction of the inlet, but only a few particles reached the wall.

IV. PLANNED METHOD FOR SANDWICH ELISA IN PDMS MICROCHANNEL

In the microfluidic channel I plan to implement a standard sandwich ELISA method. The theoretical model (Fig. 9) shows the main components of the ELISA:

- primary capture antibody coated on the wall of PDMS channel
- antigen
- secondary antibody labelled by special enzyme
- enzyme substrate

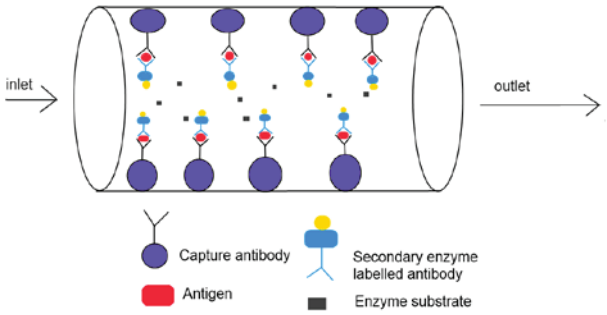


Fig. 9: Schematic of sandwich ELISA in PDMS microfluidic channel

During the implementation I would try two different ways: full flow-through and a half flow-through solution. The main difference between them, how long time the different components spend in the channel.

A. Flow-through solution

1) Coating the capture antibody

a) Coat the antibody by passive adsorption

b) Dilute the capture antibody in bicarbonate or carbonate until we reach a concentration of 8-10 µg/ml.

c) Pump diluted antibodies through the channel

d) Wash the channel through with washing buffer (PBS (Phosphate Buffered Saline)) two times to clean out the unnecessary coating solution residual.

2) Blocking

a) Mix blocking solution from BSA and PBS, additionally some kind of non-ionic detergent.

b) Pump the blocking solution through the channel to block the non-specific binding sites on the surface of PDMS

c) Wash the channel twice with washing buffer

3) Adding the sample

a) Flow through the sample

b) Wash the channel through with washing buffer twice

4) Adding the secondary antibodies

a) Flow through the enzyme labelled (streptavidin-HRP) secondary antibody

b) Wash the channel through with washing buffer twice

5) Adding enzyme substrate

a) Pump the enzyme substrate (TMB) into the channel

6) Adding STOP solution

a) Pump the STOP solution into the channel after 10-15 minutes adding TMB substrate.

7) Absorbance detection with photodiode

*B. **Half flow-through solution***

In the half flow-through solution I plan to change the length of some steps of the previously described method. In the coating step I pump the amount of diluted antibody, that fills the whole channel, then I stop pumping, close the inlet, outlet and put the device into a vacuum space. I keep the device in this vacuum to diffuse the water vapor out of PDMS and help the binding of antibodies on to the wall. In the 2.-5. steps I pump the current component into the channel in the amount that fills out the channel. I wait 15-30 minutes, then continue to the next step. Waiting time is for testing better bindings, more effective reactions.

CONCLUSIONS

The used geometry inside the microfluidic channel influences the particles trajectories. The particles reached and bound to wall surface with a lower probability in a simple channel, than in a channel using additional structures. According to the flow simulations I will use round structures in my device and implement the sandwich ELISA method modified for microfluidic channel.

ACKNOWLEDGMENT

I would like to acknowledge my supervisor, Kristóf Iván for his kind help and his knowledge on this multidisciplinary microfluidical field. Also the support TÁMOP-4.2.1./B-11/2/KMR-2011-002 and TÁMOP-4.2.2./B-10/1-2010-0014 is kindly acknowledged.

REFERENCES

[1] L. Yu, C. M. Li, Y. Liu, J. Gao, W. Wang, and Y. Gan, "Flow-through functionalized PDMS microfluidic channels with dextran derivative for ELISAs," *Lab Chip*, vol. 9, no. 9, pp. 1243–1247, May 2009.

[2] S. Sun, M. Yang, Y. Kostov, and A. Rasooly, "ELISA-LOC: lab-on-a-chip for enzyme-linked immunodetection," *Lab Chip*, vol. 10, no. 16, pp. 2093–2100, Jul. 2010.

[3] C. D. Chin, V. Linder, and S. K. Sia, "Commercialization of microfluidic point-of-care diagnostic devices," Lab Chip, vol. 12, no. 12, pp. 2118–2134, May 2012.

[4] T. Ohashi, K. Mawatari, K. Sato, M. Tokeshi, and T. Kitamori, "A micro-ELISA system for the rapid and sensitive measurement of total and specific immunoglobulin E and clinical application to allergy diagnosis," *Lab Chip*, vol. 9, no. 7, pp. 991–995, Apr. 2009.

[5] S. Sun, M. Yang, Y. Kostov, and A. Rasooly, "ELISA-LOC: lab-on-a-chip for enzyme-linked immunodetection," *Lab Chip*, vol. 10, no. 16, pp. 2093–2100, Jul. 2010.

[6] M. Ihara, A. Yoshikawa, Y. Wu, H. Takahashi, K. Mawatari, K. Shimura, K. Sato, T. Kitamori, and H. Ueda, "Micro OS-ELISA: Rapid noncompetitive detection of a small biomarker peptide by open-sandwich enzyme-linked immunosorbent assay (OS-ELISA) integrated into microfluidic device," *Lab Chip*, vol. 10, no. 1, pp. 92–100, Jan. 2010.

[7] E. Eteshola and D. Leckband, "Development and characterization of an ELISA assay in PDMS microfluidic channels," *Sensors and Actuators B: Chemical*, vol. 72, no. 2, pp. 129–133, Jan. 2001.

[8] Y. Dong, Y. Xu, Z. Liu, Y. Fu, T. Ohashi, Y. Tanaka, K. Mawatari, and T. Kitamori, "Rapid screening swine foot-and-mouth disease virus using micro-ELISA system," *Lab Chip*, vol. 11, no. 13, pp. 2153–2155, Jun. 2011.

[9] M. Herrmann, T. Veres, and M. Tabrizian, "Enzymatically-generated fluorescent detection in micro-channels with internal magnetic mixing for the development of parallel microfluidic ELISA," *Lab Chip*, vol. 6, no. 4, pp. 555–560, Mar. 2006.

[10] D. N. Kim, Y. Lee, and W.-G. Koh, "Fabrication of microfluidic devices incorporating bead-based reaction and microarray-based detection system for enzymatic assay," *Sensors and Actuators B: Chemical*, vol. 137, no. 1, pp. 305–312, Mar. 2009.

New insights in neuroscience with two-photon lasermicroscopy

Dénes Pálfi
(Supervisor: Dr. Balázs Rózsa)
denes.palfi@gmail.com

Abstract— **In this brief study I would like to show our recent in vitro experimental results. For the better clarity the study is separated to three projects which we have done in the recent past. First we developed some new chemical compounds which are better in any chemical, biological and pysical propeties compared to commercial available materials (1). These compounds were tested in biological experiments to verify our forecast. In the second project (2) we wished to mimick the interneuronal ripple oscillations which we found in fast-spiking, hippocampal parvalbumin-expressing interneurons (FS-PV INs) during spontaneous Sharp Wave activity. Sharp Wave-Ripple (SW-R) activity, which has been considered as an oscillatory network event, is associated with the reactivation of neuronal ensembles within specific circuits during memory formation. Using fast 3D two-photon trajectory scanning and a new caged glutamate compound with enhanced efficacy, we show that in the active state dendritic integration is supralinear and Ca2+ spikes are generated. The origin of interneuronal ripple oscillations were verified with pharmacology. The third interest area is the Silicon Carbide Quantum Dots for bioimaging (SiC QDs) (3). Luminescent nanocrystals or quantum dots give great potential for bio-analysis as well as optoelectronics. Two-photon excitation showed significant response from silicon carbide nanocrystals that were injected into hippocampal CA1 pyramidal cells. Further plans are to continue the development of GABA compounds and studying hippocampal network activity during SW-R.**

Keywords- **two-photon microscopy; glutamate uncaging; sharp wave activity; parvalbumin-expressing interneuron; dendrite; quantum dots**

I. NOVEL MORE EFFECTIVE GLUTEMATE AND GABA UNCAGING COMPOUNDS

Two-photon photochemical uncaging revolutionised many areas of cell biology and neurobiology because it could allow rapid photochemical release of neurotransmitters therefore being capable of mimicking fast synaptic quantal release[1]. In a combination with fast laser scanning methods neurotransmitters could be released in complex spatio-temporal patterns. Glutamate is one of the major transmitters in the nervous system and therefore several caged glutamate compounds have been developed[2,3]. They are generally used in neurophysiology experiments, investigating postsynaptic mechanisms, and fast dendritic signal integration[4,5]. However, only 2(*S*)-2-amino-5-(4-methoxy-7-nitro-2,3-dihydro-indol-1-yl)-5-oxo-pentanoic acid (MNI-Glu) has being used widely in two-photon experiments. This was due to the strict constraint for an appropriate caged compound coupled with the efficient uncaging with high chemical yield, fast two-photon induced release and low spontaneous hydrolysis rate[3].

In order to compensate for spontaneous hydrolysis of materials with fast photochemical reaction, first we wished to understand the relationship between spontaneous hydrolysis, two-photon photochemical action cross section, two-photon spectrum and chemical structure. Therefore, we have synthesized four chemical analogues of MNI-Glu (**Fig. 1**) and compared measured data with the predictions from the quantum chemical model. MNI-Glu trifluoroacetate (MNI-Glu•TFA) was prepared by a well-known method, but we isolated the TFA salt without chromatography with a better yield than previously reported methods. For the synthesis of 2(*S*)-2-amino-5-(4-methoxy-5,7-dinitro-2,3-dihydro-indol-1-yl)-5-oxo-pentanoic acid trifuoroacetate (DNI-Glu•TFA; **Fig. 1**) we have developed a new synthetic route, more efficient than the previously reported methods[6].

The reaction mechanism of the release of the glutamate from MNI-Glu•TFA and from DNI-Glu•TFA has been investigated by first principle molecular quantum mechanics using the DALTON program, the results are shown in **Fig. 2**. The mapped reaction mechanism is identical for MNI-Glu•TFA and DNI-Glu•TFA, but the energy values are somewhat smaller for DNI-Glu•TFA (values are shown in bracket), in general smaller by about 4–10 kJ mol^{-1}. Thus the photochemistry of DNI-Glu•TFA is observed to occur with a greater quantum yields as computed by G09.

Next, we compared the overall photochemical yield of MNI-Glu•TFA, DNI-Glu•TFA, MNI-Ulg•TFA and DNI-Ulg•TFA. The same laser energy which induced only a small somatic voltage and dendritic Ca^{2+} response (1.43±0.15 mV and 1.22±0.11 ΔF/F) in the presence of MNI-Glu•TFA, elicited approximately 10-fold higher responses in the presence of DNI-Glu•TFA (15.22±0.28 mV and 7.95±0.25 ΔF/F; **Fig. 3**). However, due to the nonlinear input-output property of dendritic and somatic membrane compartments the efficiency of the fast photochemical release could not be determined at a single laser intensity. Therefore we have performed a series of uncaging measurements at increasing laser intensities and plotted the responses as a function of the second order of the laser intensity and measured the increased photochemical yield as a relative *x*-axis shift of the responses in the presence of DNI-Glu•TFA and MNI-Glu•TFA by calculating distance between the two point sets using unconstrained nonlinear

D. Pálfi, "New insights in neuroscience with two-photon lasermicroscopy,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 55-58.

optimization **(Fig. 3c)**. The average distance between the two point sets revealed that release of glutamate with the same rate from MNI-Glu•TFA requires 7.17 ± 0.84-fold higher two-photon excitation when compared to DNI-Glu•TFA ($p < 0.00001$, n= 10), which corresponds to the value predicted by quantum chemical modeling (10-fold increase). Next, we repeated these experiments by comparing the efficiency of the photochemical release of the reversely coupled compounds (MNI-Ulg•TFA and DNI-Ulg•TFA) relative to the one of MNI-Glu•TFA. Again, in good agreement with the reduced photochemical release proposed by quantum chemical modeling, uncaging responses were reduced to 51.07 ± 6.76 % ($p < 0.001$, n = 4) and to 125,3 ± 4,1 % ($p = 0.003$, n = 3) in the presence of MNI-Ulg•TFA and DNI-Ulg•TFA as compared to MNI-Glu•TFA. These data show that the efficiency of photochemical release calculated from the quantum chemical model and from the uncaging measurements correlated well for all the four uncaging materials (R = 0.98).

II. Dendritic spikes and ripples in parvalbumin interneurons

Enhanced green fluorescent protein-expressing FS-PV Ins[7] in CA1 stratum pyramidale were identified with two-photon imaging and were filled with a fluorescent Ca^{2+} indicator (OGB-1 or Fluo-4) via a somatic recording pipette. All of the recorded neurons were characterized as typical fast-spiking interneurons. For Ca^{2+} imaging, a reference z-stack was taken with 3D acousto-optical imaging in order to select multiple long dendritic segments covering the majority of the dendritic arbor(**Fig. 4**). Then we performed simultaneous fast 3D trajectory scanning[8] along the dendritic segments during SPW-R events. Interestingly, the SPW-EPSP-associated dendritic Ca^{2+} signals invaded the majority of the distal apical but not the basal dendritic arbor (**Fig. 5**).

In order to find the minimum number of excitatory inputs that are needed to evoke the regenerative event in an all-or-none manner, we simulated excitatory inputs by using two-photon glutamate uncaging in short, temporally and spatially clustered patterns[9]. This method allowed the selective measurement of postsynaptic mechanisms and avoided any potential variability due to presynaptic mechanisms. To model the long-lasting and large-amplitude SPW-EPSPs (29.37±2.49 ms, 22.6±1.7 mV, n=12, see **Fig. 4**) from small unitary inputs (< 1 mV)[10], tens of unitary inputs must be activated in a short time window. With the application of DNI-Glu·TFA, it became possible to rapidly and repetitively activate the required high number of unitary inputs (up to 60) in a short time period (4.39±0.33 ms, n=35 cells) without inducing a detectable level of phototoxicity in the dendrites. In order to investigate the functional role of different ion channels that might satisfy criterion #5, we activated spatially and temporally clustered patterns of 43.8±2.9 inputs (above the second threshold, but below the threshold of somatic AP generation) at distal dendritic segments in control conditions and in the presence of various voltage- and ligand-gated ion channel blockers for comparison. Therefore, we imaged long dendritic segments and activated the inputs. These lateral dendritic regions of active propagation appeared in the spatial distribution of the Ca^{2+} amplitude as a plateau following a sharp drop in amplitude at the border of the input and lateral dendritic regions(**Fig. 6a-b**). The mean effect of the VGCC, voltage-gated Na^{+} channel, NMDA and AMPA receptor inhibitors was generally smaller in the input region as compared to the control values (**Fig. 6a-d**), but the combined application of AMPA and NMDA receptor blockers (AP5 and CNQX) reduced the Ca^{2+} responses to almost zero. Our data indicate that VGCCs are mainly responsible for the Ca^{2+} influx in the lateral dendritic region, and thus for the dendritic Ca^{2+} spike. Hippocampal interneurons express P/Q-, R-, L-, N-, and T-type VGCCs[11], but we found that the L-type VGCC blocker Nimodipine had the greatest effect on the Ca^{2+} responses of FS-PV INs, both in the present study and in our earlier data[12]. In line with other observations, we noted that IEM-1460 (a blocker of Ca^{2+}-permeable AMPA receptors) also had a large effect on the postsynaptic Ca^{2+}-influx[13]. Our experimental results suggest that dendritic voltage-gated Ca^{2+} and Na^{+} channels may be primarily responsible for the supralinear responses and the accompanying fast interneuronal ripple oscillations.

III. Silicon Carbide Quantum Dots for Bioimaging

For in vivo studies red or near-infrared region excitation and emission are desirable. Two-photon excitation of Silicon Carbide Quantum Dots (SiC QDs) may work efficiently for in vivo bioimaging because short near-infrared pulses can penetrate deeply in organic tissues (unlike photons in the ultraviolet or visible region), and our SiC QDs absorb single photons in the ultraviolet region which corresponds to absorption in the near-infrared region by two-photon processes. Indeed, we demonstrate here that our SiC QDs have a strong emission due to two-photon excitation. Fluorescence were detected from SiC QDs injected to neuron cells by a two-photon microscope. Two-photon imaging was performed using a Femto2D Two-Photon Laser Scanning Microscope (2PLSM; Femto2D, Femtonics Ltd.) with an ultrafast Ti:Sapphire laser (Mai Tai, Spectra Physics) tuned from 740 nm to 850 nm14. Optimal excitation wavelength for imaging in red channel (emission wavelengths between 600 and 700 nm) was around 830 nm. The excitation laser energy was ~60 mW before the objective. Whole cell patch-clamp recordings were made from CA1 pyramidal neurons in acute 300 µm mouse hippocampal slices. Cells were filled with intracellular solution (ICs) and SiC QDs via the patch-pipette (6-9 MΩ). Images were taken at ~40 µm under the slice surface. Cells were identified as CA1 pyramidal cells by video microscopy using oblique infrared illumination and by analyzing their response to somatic current injections, serving also as a measure of cell viability (Fig. 7C). The soma was clearly visible in the red channel, and even the apical dendrite near to the soma could be resolved (Fig. 7B). However the signal contrast was relatively weak in the green channel (425-525 nm, Fig.7A). which is a bit unexpected since the maximum emission of SiC QDs in water is in the green channel by single photon excitation. Possible, the altered ion and protein concentrations might generate a change in the emission spectra and fluorescent intensity as it also occurs in

case of other fluorescent dyes Maravall et al.15. Further investigations are needed to clear this issue. Z-stacks of red channel images were also taken confirming homogeneous loading of soma and proximal dendrites (Fig. 7D). As we expected from the bioinert property of SiC QDs, no sign of cellular damage of the pyramidal neurons was observable during the long (> 1 hour) measurement time. In addition, basic electrophysiological properties (action potential threshold, peak width and amplitude; membrane potential; firing rate) have not changed significantly during the measurement indicating the lack of cellular toxicity. This study implies that our SiC QDs may be good candidates as luminescent biomarkers for neuroscience.

REFERENCES

[1] Matsuzaki, M. et al. Dendritic spine geometry is critical for AMPA receptor expression in hippocampal CA1 pyramidal neurons. Nat Neurosci 4, 1086-1092, doi:10.1038/nn736nn736 [pii] (2001).

[2] Kantevari, S., Matsuzaki, M., Kanemoto, Y., Kasai, H. & Ellis-Davies, G. C. R. Two-color, two-photon uncaging of glutamate and GABA. Nature Methods 7, 123-125 (2010).

[3] Warther, D. et al. Two-photon uncaging: New prospects in neuroscience and cellular biology. Bioorganic and Medicinal Chemistry 18, 7753-7758 (2010).

[4] Katona, G. et al. Roller coaster scanning reveals spontaneous triggering of dendritic spikes in CA1 interneurons. Proceedings of the National Academy of Sciences of the United States of America 108, 2148-2153 (2011).

[5] Losonczy, A., Makara, J. K. & Magee, J. C. Compartmentalized dendritic plasticity and input feature storage in neurons. Nature 452, 436-441 (2008).

[6] Fedoryak, O. D., Sul, J. Y., Haydon, P. G. & Ellis-Davies, G. C. Synthesis of a caged glutamate for efficient one- and two-photon photorelease on living cells. Chem Commun (Camb), 3664-3666, doi:10.1039/b504922a (2005).

[7] Meyer, A. H., Katona, I., Blatow, M., Rozov, A. & Monyer, H. In vivo labeling of parvalbumin-positive interneurons and analysis of electrical coupling in identified neurons. J Neurosci 22, 7055-7064 (2002).

[8] Katona, G. et al. Fast two-photon in vivo imaging with three-dimensional random-access scanning in large tissue volumes. Nat Methods 9, 201-208 (2012).

[9] Losonczy, A. & Magee, J. C. Integrative properties of radial oblique dendrites in hippocampal CA1 pyramidal neurons. Neuron 50, 291-307 (2006).

[10] Ali, A. B., Deuchars, J., Pawelzik, H. & Thomson, A. M. CA1 pyramidal to basket and bistratified cell EPSPs: dual intracellular recordings in rat hippocampal slices. J Physiol 507 (Pt 1), 201-217 (1998).

[11] Vinet, J. & Sik, A. Expression pattern of voltage-dependent calcium channel subunits in hippocampal inhibitory neurons in mice. Neuroscience 143, 189-212 (2006).

[12] Chiovini, B. et al. Enhanced dendritic action potential backpropagation in parvalbumin-positive basket cells during sharp wave activity. Neurochem Res 35, 2086-2095 (2010).

[13] Topolnik, L. Dendritic calcium mechanisms and long-term potentiation in cortical inhibitory interneurons. Eur J Neurosci 35, 496-506 (2012).

[14] G. Katona, G. Szalay, P. Maák, A. Kaszás, M. Veress, D. Hillier, B. Chiovini, E.S. Vizi, B. Roska and B. Rózsa: Fast two-photon in vivo imaging with three-dimensional random-access scanning in large tissue volumes, Nat. Methods 9, 201 (2012).

[15] M. Maravall, Z.F. Mainen, B.L. Sabatini and K. Svoboda: Estimating intracellular calcium concentrations and buffering without wavelength ratioing, Biophys. J. 78, (2000)

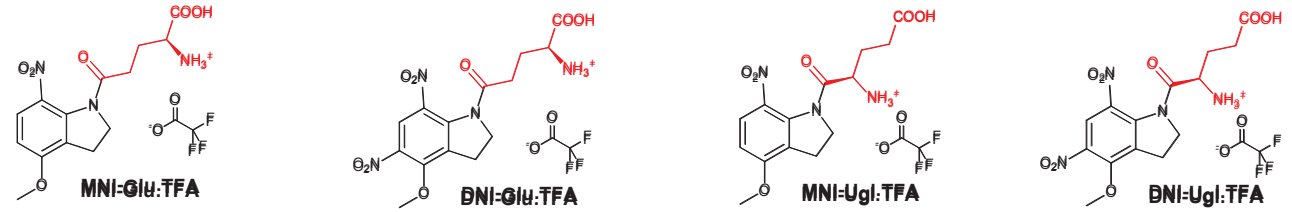


Figure 1: Chemical structure of the four syntetized cage compound.

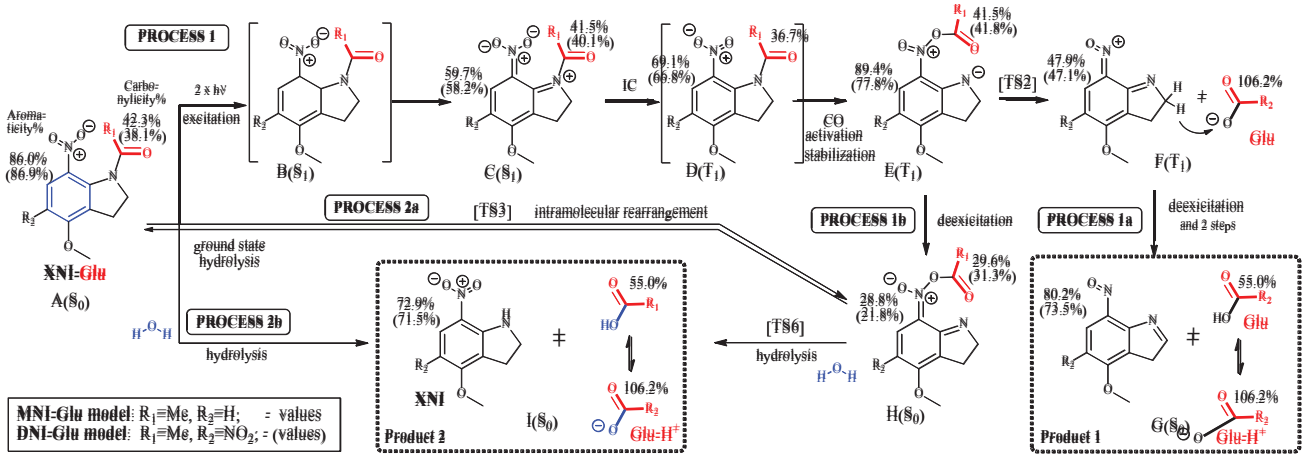


Figure 2: Quantum chemical model of compounds with high two-photon photochemical release.

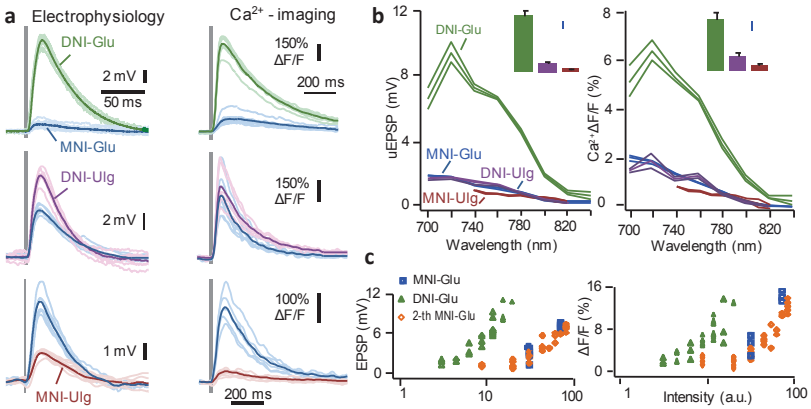


Figure 3: DNI-Glu•TFA in the presence of enzymatic compensation is a more effective caged glutamate compound as compared to MNI-Glu•TFA. a Two-photon uncaging evoked somatic EPSPs (uEPSPs) (left) and simultaneously measured Ca2+ transients (Right) in the presence of DNI-Glu•TFA, DNI-Ulg•TFA, MNI-Ulg•TFA as compared to the responses measured in the presence of MNI-Glu•TFA. Pairwise comparison was performed under the same conditions (using the same laser intensity, concentration and uncaging time in the same dendritic locations). Bold line shows the averages. Gray bars represent uncaging time. b Normalized amplitude of uEPSPs (left) and Ca2+ transients (right) evoked in the presence DNI-Glu•TFA, DNI-Ulg•TFA, MNI-Ulg•TFA, MNI-Ulg•TFA as a function of wavelength (mean ±s.e.m.; DNI-Glu•TFA, n=11; DNI-Ulg•TFA, n=4; MNI-Ulg•TFA, n=3, MNI-Glu•TFA, n = 6). (Inset) Normalized uEPSPs (left) and Ca2+ transients (right) in a (mean ± s.e.m). Responses were normalized to MNI-Glu•TFA responses. Blue bars show one. c Ca2+ transient and uEPSP amplitudes plotted as a function of the uncaging laser intensity in the presence of MNI-Glu•TFA (blue) following wash-in of DNI-Glu•TFA (green) and after recovery in MNI-Glu•TFA (orange).

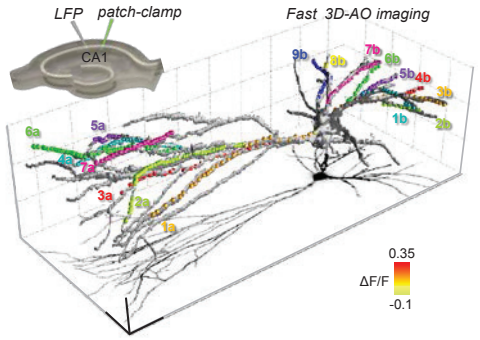


Figure 4: Full dendritic arborization of a FS-PV IN (top) imaged by 3D acousto-optical scanning in the hippocampal CA1 region. Colored spheres represent locations for the point-by-point trajectory scanning. Inset, experimental setup.

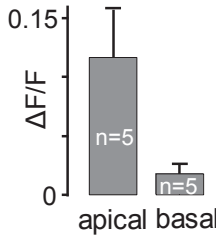


Figure 5: Average apical and basal dendritic Ca2+ signals during SPW-EPSPs (mean±s.e.m., n=5 cells).

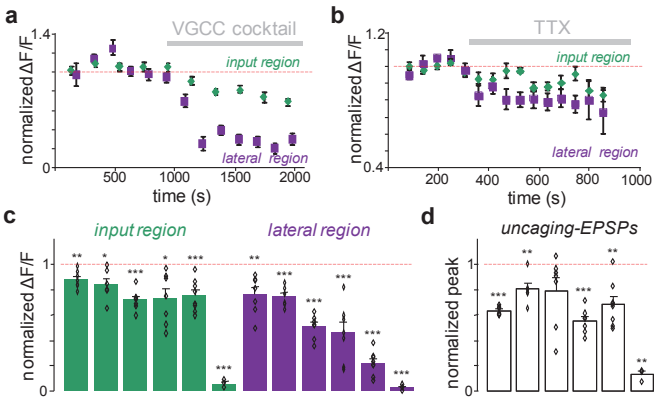


Figure 6: a Time course of the effect of VGCC cocktail on Ca2+ responses in the input (green) and lateral dendritic (magenta) regions. **b**, The same as a respectively, but for TTX. **c** Effect of different ion channel blockers on the peak amplitude of Ca2+ transients. **d**, The same as c, but for simultaneously recorded EPSPs.

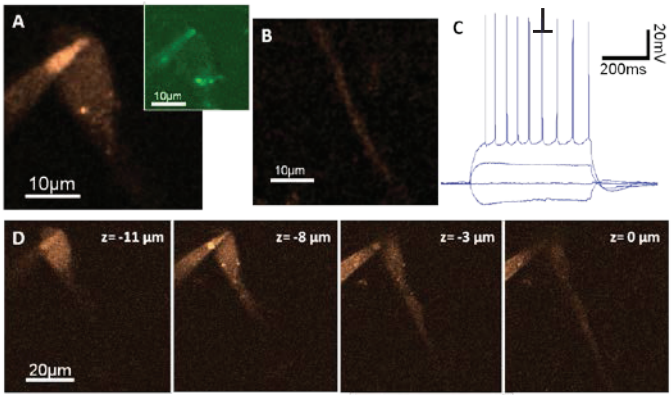


Figure 7: Two-photon imaging of a neuron labelled by SiC QDs. (**A**) Representative 2PLSM image of a CA1 hippocampal pyramidal neuron filled with SiC QDs. Red fluorescence signal (600-700 nm, red channel) was generated at 830 nm excitation. Inset, the same neuron, but fluorescent signal was collected from 425 nm to 525 nm (green channel). (**B**) Red channel image of the apical dendrite of the neuron. (**C**) Somatic membrane voltage responses of the neuron induced by somatic current injection show normal functioning of the neuron. (**D**) 2PLSM images collected at different depths using 800 nm excitation.

Effects of Fractalkine/CX3CR1 system on the development of obesity

Ágnes Polyák
(Supervisor: Dr. Krisztina Kovács)
polyak.agnes@itk.ppke.hu

***Abstract*—Obesity is characterized by, among others, chronic low-grade inflammation, immune cell infiltration into adipose tissue, proinflammatory cytokine production, insulin resistance. Fractalkine is a chemokine, which participates in attraction and adhesion of immune cells. The mechanisms by which monocytes traffic to adipose tissue are incompletely understood and a key question in the field. We investigated the effects of fractalkine receptor deficiency in the development of obesity. After 10 weeks of high fat diet or normal chow feeding glucose tolerance test was performed on fractalkine receptor deficient (CX3CR1 GFP/GFP) and control (CX3CR1 +/GFP) mice. Afterwards they were sacrificed, organs were harvested. Histological and real-time PCR analysis was used to reveal the differences between groups. We found that lack of fractalkine/fractalkine receptor system attenuated the symptoms of obesity: CX3CR1 GFP/GFP mice gained less weight, epididymal fat pads and average adipocyte size were smaller, produced less chemokines and proinflammatory cytokines and did not develop insulin resistance. Compared to others' our results suggest that lack of fractalkine signaling may slow down the development of obesity and attenuates comorbid immune-related alterations.**

***Keywords* - obesity, high fat diet, fractalkine, CX3CR1**

***Abbreviations* - CX3CL1/FKN-fractalkine, CX3CR1-fractalkine receptor, GFP-green fluorescent protein, ATM-adipose tissue macrophages, HFD-high fat diet, ND-normal diet, MCP-1-monocyte chemoattractant protein-1, IL-interleukin, TNFa- tumor necrosis factor alpha, GTT-glucose tolerance test.**

I. Introduction

High-fat diet (HFD)-induced obesity has emerged as a state of chronic low-grade inflammation characterized by a progressive infiltration of immune cells, particularly macrophages, into obese adipose tissue. Adipose tissue macrophages (ATM) present immense plasticity. In early obesity, M2 anti-inflammatory macrophages acquire an M1 pro-inflammatory phenotype. Pro-inflammatory cytokines including TNF-α, IL-6 and IL-1β produced by M1 ATM exacerbate local inflammation promoting insulin resistance, which consequently, can lead to type-2 diabetes mellitus. However, the triggers responsible for ATM recruitment and activation are not fully understood. Adipose tissue-derived chemokines are significant players in driving ATM recruitment during obesity. [1]

Fractalkine, a chemokine that signals through a single known receptor (CX3CR1), is expressed on numerous cells: on activated endothelial, smooth muscle cells, macrophages, and adipocytes [2]. It is synthesized as a trans-membrane protein with the CX3C chemokine domain displayed on an extended highly glycosylated, mucin-like stalk [3, 4] (Fig. 1). The transmembrane form of fractalkine is capable of mediating adhesion of cells expressing the G protein–coupled receptor CX3CR1 [2] and it mediates monocyte adhesion to human adipocytes [5]. A soluble form can be released from its membrane form by extracellular cleavage and then act as a classical chemoattractant for CX3CR1 expressing leukocytes. The expression of fractalkine has reportedly been enhanced by inflammatory stimuli, i.e., TNF-α, interferon (IFN)-γ and lipopolysaccharide [2].

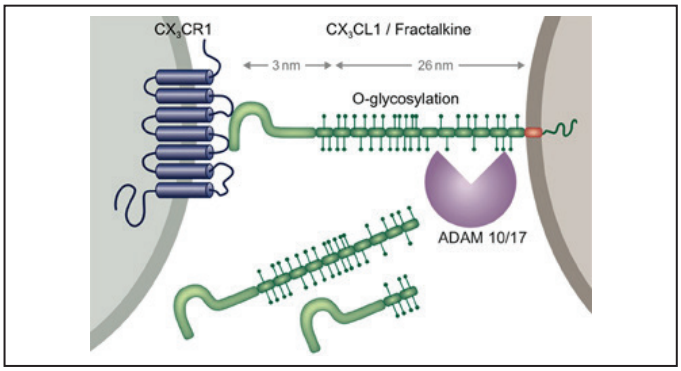


Figure 1. Transmembrane and soluble form of fractalkine, and its receptor: CX3CR1 [6].

II. Materials and Methods

A. Animals and diet

Experiments were performed in male CX3CR1 +/GFP, and CX3CR1 GFP/GFP mice. CX3CR1 +/GFP mice was used as control group as there were no significant difference between +/GFP and +/+ mice in previous experiments. CX3CR1/GFP mice were obtained from the European Mouse Mutant Archive (EMMA), backcrossed for more than 10 generations to C57Bl/6 [7]. In these mice, the cx3cr1 gene was replaced by a GFP reporter gene. At 35 days of age both CX3CR1 +/GFP (n=8) and CX3CR1 GFP/GFP (n=10) mice were randomly distributed into two equal groups. The first group, normal diet (ND), received standard chow (VRF1 (P), Special Diets Services (SDS), Witham, Essex, UK). The second group, high fat diet (HFD) was given standard chow mixed with lard (Spar Budget) (2:1) for 10 weeks. The mice were housed in groups of 4-5. Animals had free access to food and water and were maintained under

Á. Polyák, "Effects of Fractalkine/CX3CR1 system on the development of obesity,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 59-62.

controlled conditions: temperature, 21°C±1° C; humidity, 65%; light-dark cycle, 12-h light/12-h dark cycle, lights on at 07:00. All procedures were conducted in accordance with the guidelines set by the European Communities Council Directive (86/609 EEC) and approved by the Institutional Animal Care and Use Committee of the Institute of Experimental Medicine.

B. Experimental design

Mice were fed with ND or HFD for 10 weeks, body weight and food consumption was measured weekly. In the 10th week glucose tolerance (GTT) test was performed after overnight fasting. Two days after the GTT, mice were decapitated, trunk blood was collected, organs were harvested and stored at -70°C for RT-PCR, or fixed in 4% PFA for histology.

C. Glucose tolerance test

Mice were fasted overnight (15 h) and then injected intraperitoneally with 2 mg/g of body weight D-glucose (20% stock solution in saline). Blood glucose was measured from tail vein by DCont Personal Blood Glucose Meter (77 Elektronika Kft. Hungary) at 0 min (just before glucose injection) and at 15-, 30-, 60-, 90- and 120-min intervals after the glucose load.

D. Histology

Tissues were fixed by immersion in 4% paraformaldehyde in 0.1 M phosphate buffer, pH 7.4 (PB) for 3 days. Subsequently, they were stored in 1% paraformaldehyde in 0.1 M PB at 4°C. Tissues were paraffin-embedded, sectioned and stained with H&E stain. Slides were digitalized with Pannoramic Digital Slide Scanner (3DHISTECH Kft., Hungary) and analyzed with ImageJ software (NIH, USA).

E. Quantitative real-time PCR

Total RNA was isolated from epididymal white adipose tissue (EWAT) samples with QIAGEN RNeasyMiniKit (Qiagen, Valencia, CA, USA) according the manufacturer's instruction. To eliminate genomic DNA contamination DNase I treatment was used (100 ml Rnase free DNase I (1 uDNase I, Fermentas) solution was added). Sample quality control and the quantitative analysis were carried out by NanoDrop (Thermo Scientific). Amplification was not detected in the RT-minus controls. cDNA synthesis was performed with the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Foster City, CA, USA). The designed primers (Invitrogen) were used in the real-time PCR reaction with Power SYBR Green PCR master mix (Applied Biosystems, Foster City, CA, USA) on ABI StepOne instrument. The gene expression was analyzed by ABI StepOne2.0program. The amplicon was tested by Melt Curve Analysis on ABI StepOne instrument. Experiments were normalized to Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) expression.

F. Primer design

Primers used for the comparative CT (threshold cycle) experiments were designed by the Primer Express 3.0 program. Primer sequences were the following:

GAPDH: f: TGA CGT GCC GCC TGG AGA AA
r: AGT GTA GCC CAA GAT GCC CTT CAG
IL1a: f: CCA TAA CCC ATG ATC TGG AAG AG
r: GCT TCA TCA GTT TGT ATC TCA AAT CAC
IL1b: f: CTC GTG GTG TCG GAC CCA TAT GA
r: TGA GGC CCA AGG CCA CAG GT
IL6: f: TCC GGA GAG GAG ACT TCA CA
r: TGC AAG TGC ATC ATC GTT GT
TNFa: f: CAG CCG ATG GGT TGT ACC TT
r: GGC AGC CTT GTG CCT TGA
MCP-1: f: CCAGCACCAGCACCAGCCAA
r: TGGATGCTCCAGCCGGCAAC
FKN: f: CCG CGT TCT TCC ATT TGT GT
r: GGT CAT CTT GTC GCA CAT GATT
GFP: f: GGA CGA CGG CAA CTA CAA GA
r: AAG TCG ATG CCC TTC AGC TC

G. Statistical analysis

The results are shown as means ± SEM. Statistical analysis was performed by factorial ANOVA with Newman–Keuls post-hoc test in Statistica 11 (StatSoft Inc.); $p < 0.05$ was considered significant.

III. RESULTS

A. Body weight change, adipose tissue weight and food consumption

HFD feeding leads to excessive body weight gain in control (+/GFP) mice. CX3CR1 deficient mice gained lower body weight in HFD fed group at the 8th to 10th weeks (Fig. 2A), although there were no difference in total food consumption (Fig. 2B). Epididymal adipose tissue (EWAT) fat pads were significantly larger in HFD groups, but in GFP/GFP mice they were smaller compared to +/GFP mice (Fig. 2C).

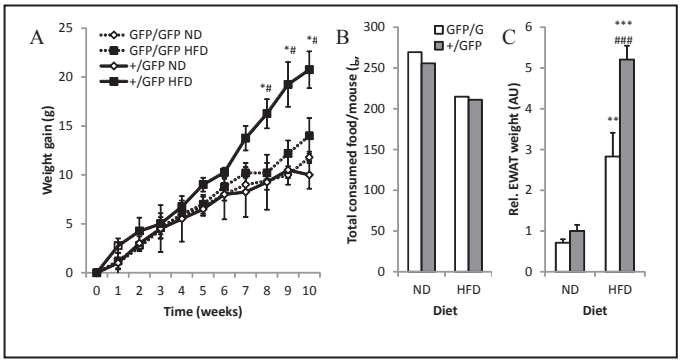


Figure 2. (A) Weight gain during 10 weeks of HFD. (B) Total consumed food during the diet. (C) Relative EWAT weight at the end of the experiment. * p<0.05, ** p<0.01, ***<0.0001 vs. ND, # p<0.05, ###<0.001 vs. GFP/GFP.

B. Glucose tolerance test

In GTT blood glucose level increases after intraperitoneal glucose load, which is followed by plasma insulin release. In normal diet fed mice insulin decreased the elevated blood glucose level, and after 120 min it returned to normal level. HFD induced glucose intolerance in control group, the blood

glucose level remained high. Lack of CX3CR1 prevented the development of glucose intolerance (Fig. 3).

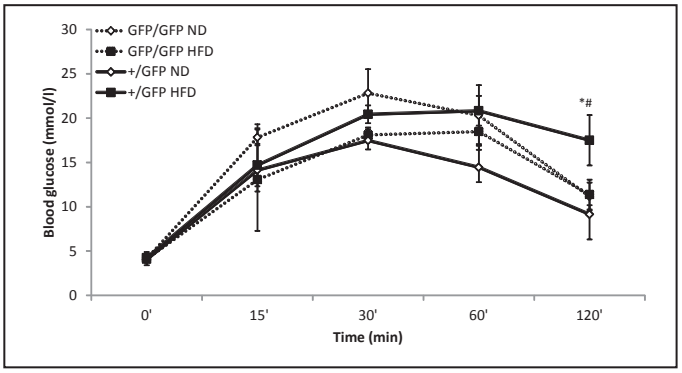


Figure 3. Results of Glucose Tolerance Test. * p<0.05 vs. ND, # p<0.05 vs. GFP/GFP.

C. Histology

HFD feeding resulted in 3.4 fold elevation in adipose cell size in +/GFP mice, and 2.38 fold elevation in GFP/GFP mice vs. ND mice (Fig. 4). CX3CR1 deficiency resulted in smaller cell size expansion (p<0.001) in HFD fed mice.

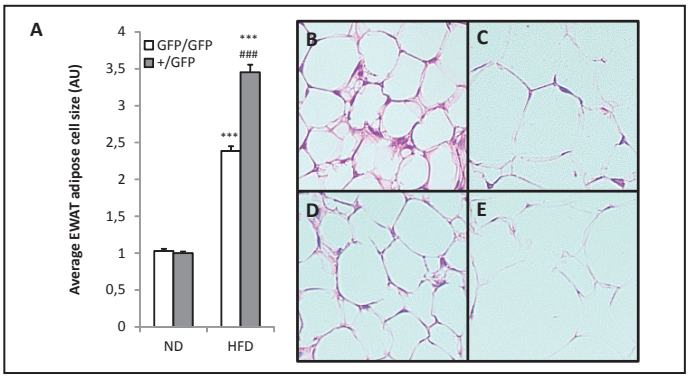


Figure 4. (A) Average adipose cell size in EWAT, arbitrary unit. *** p<0.0001 vs. ND, ### p<0.001 vs. GFP/GFP. Adipose cells in EWAT H&E staining 20x magnification, (B) CX3CR1 GFP/GFP + ND, (C) CX3CR1 GFP/GFP + HFD, (D) CX3CR1 +/GFP + ND, (D) CX3CR1 +/GFP + HFD.

D. Quantitative real-time PCR

The levels of chemokines participating in monocyte recruitment and the level of GFP, which refers to the amount of CX3CR1 expressing macrophages, were elevated in HFD fed mice. CX3CR1 deficiency resulted in smaller level of monocyte chemoattractant protein 1 and fewer macrophages (GFP positive cells) (Fig. 5).

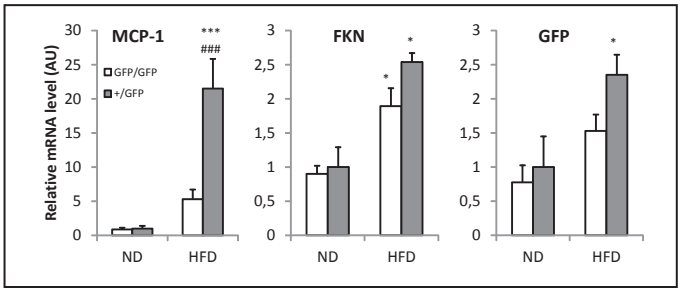


Figure 5. Chemokines (MCP-1, FKN) produced by epididymal adipose tissue. Rate of GFP+ monocites in EWAT. * p<0.05, ***<0.0001 vs. ND, ###<0.001 vs. GFP/GFP

High fat diet increased proinflammatory cytokine production in adipose tissue. Relative mRNA levels of IL1a, TNFa were significantly less elevated in CX3CR1 deficient mice (Fig. 6).

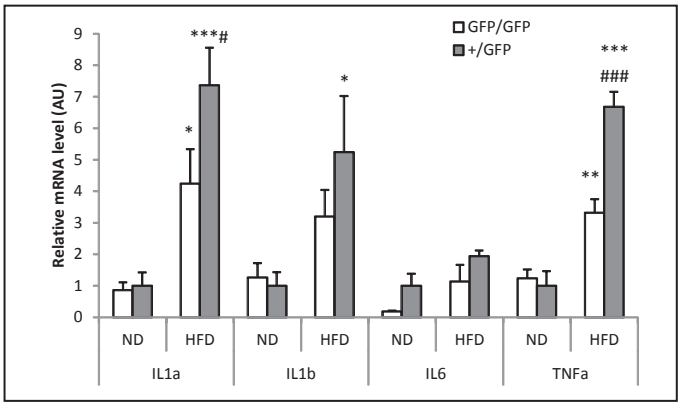


Figure 6. Proinflammatory cytokines produced by epididymal white adipose tissue. * p<0.05, ** p<0.01, ***<0.0001 vs. ND, # p<0.05, ###<0.001 vs. GFP/GFP.

IV. SUMMARY

We investigated the effect of CX3CL1/CX3CR1 system on the development of obesity. To induce obesity, we fed CX3CR1 +/GFP and CX3CR1 GFP/GFP mice with high fat diet for 10 weeks. HFD resulted in excessive body weight gain, impaired glucose tolerance, enlarged adipose cells, elevated chemoattractant levels, GFP positive cell infiltration into adipose tissue, and elevated proinflammatory cytokine production in control group (+/GFP). Lack of CX3CR1 attenuated the metabolic and immune symptoms of obesity. HFD fed CX3CR1 GFP/GFP mice did not gain more body weight than control, normal fed mice, although their epididymal adipose tissue, and adipocyte size was larger. They did not develop glucose intolerance and produced less chemokines and proinflammatory cytokines than CX3CR1 +/GFP mice.

In an experiment performed by Morris et al. [8], after 20 weeks of HFD, mice developed obesity induced insulin resistance and CX3CR1 deficiency did not alter the monocyte trafficking into adipose tissue.

These results suggest that lack of fractalkine may slow down the development of obesity.

REFERENCES

[1] Finucane, O.M., et al., Insights into the role of macrophage migration inhibitory factor in obesity and insulin resistance. Proc Nutr Soc, 2012. 71(4): p. 622-33.

[2] Cefalu, W.T., Fractalkine: a cellular link between adipose tissue inflammation and vascular pathologies. Diabetes, 2011. 60(5): p. 1380-2.

[3] Bazan, J.F., et al., A new class of membrane-bound chemokine with a CX3C motif. Nature, 1997. 385(6617): p. 640-4.

[4] Pan, Y., et al., Neurotactin, a membrane-anchored chemokine upregulated in brain inflammation. Nature, 1997. 387(6633): p. 611-7.

[5] Shah, R., et al., Fractalkine is a novel human adipochemokine associated with type 2 diabetes. Diabetes, 2011. 60(5): p. 1512-8.

[6] Wolf, Y., et al., Microglia, seen from the CX3CR1 angle. Front Cell Neurosci, 2013. 7: p. 26.

[7] Jung, S., et al., Analysis of fractalkine receptor CX(3)CR1 function by targeted deletion and green fluorescent protein reporter gene insertion. Molecular and Cellular Biology, 2000. 20(11): p. 4106-4114.

[8] Morris, D.L., et al., CX3CR1 deficiency does not influence trafficking of adipose tissue macrophages in mice with diet-induced obesity. Obesity (Silver Spring), 2012. 20(6): p. 1189-99.

Biomimetic Test bed Hand

Norbert Sárkány
(Supervisors: Dr. György Cserey, Dr. Péter Szolgay)
sarkany.norbert@itk.ppke.hu

***Abstract*—This paper presents a design of an anthropomorphic biomimetic test bed hand, focusing on the design of the fingers and its bio-inspired flexor-extensor like control. The kinematic description, the detailed explanation and presentation of the 3D CAD design are included. The description of the applied 3D touch and magnetic sensors are also detailed in the article. Functional simulation results and also the first experiments of the hardware prototype gave promising results and show that the approach can be an effective solution for the need of a hand test bed.**

***Keywords*-robotics; bionics; biomimetic; test bed; robotic hand**

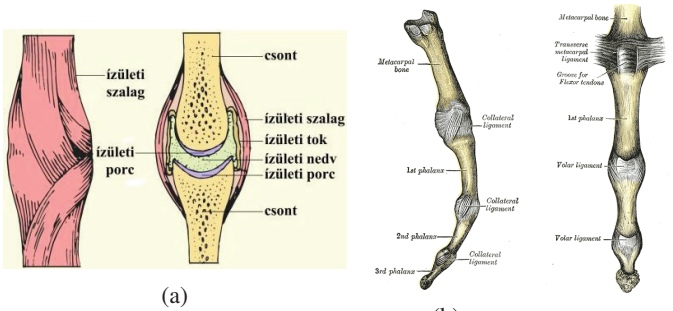
I. Introduction

In the last twenty years there was an extensive research about robotic hands, their goal was to design and develop an anthropomorphic dexterous hand [1], [2], [3], [4], [5], [6]. There are two designs, one with a local control where the actuators are in the hand [1], [3], [4], its reduces the amount of space which it requires, and the weight. The reduced weight is always an important aspect but in many cases this reduces the DOF. The second design is where the actuated structure and the actuator mechanism are separated and connected with artificial tendons, such a hand is capable to do manipulation tasks like a human hand can do, here every joint has an independent control, and there are no passively controlled joints. The commercially available prosthetics are similar to the first type but are limited in their movement capability and they have a lack of sensory information and less of control. In this paper a design of an anthropomorphic biomimetic hand test bed is discussed. Primarily the design of the finger and the concept of the actuation system is shown. The main goal of the research is to have a fully functional biomimetic hand test bed. Which can be used in robotic applications, research and it also give a basis of new prosthetics design too.

In Section. II the anatomical bases are presented, Section III shows the control and sensing structures in the human nevus system. Section IV discuss the finger kinematics and representation. Section V shows the biomimetic test bed setup for one finger, Section VI shows the concept of control and sensig in a finger. Finally, conclusions and future work are discussed in Section. VII.

II. Anatomical Bases

A finger is a limb of our body a tool of manipulation and sensing, it is found in the hand. Normally a human has five of them Thumb, Index, Middle, Ring and Little finger. The Thumb is structurally a little different from the other five, it has two phalanges and the other four have three. The components which constitute the fingers are the muscle, the ligaments, and the articulations. These mechanical structures make it possible to have 24 DOF and to achieve a smooth, compliant and accurate movement.



(a) (b) (c)

Fig. 1: General articulations structure (Fig. 1a), finger and its articulations(Fig. 1b , Fig. 1c)

A. Articulations

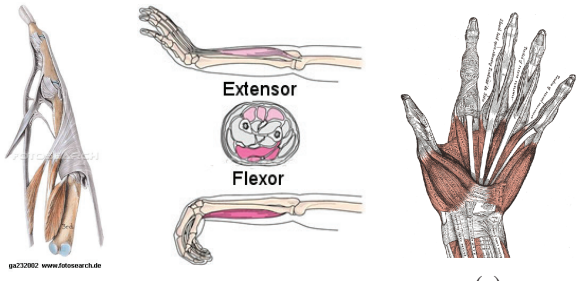
In our skeletal structure bones can connect in two ways, continues and interrupted. The continues connections (synarthrosis) also can be divided in to two groups transient and permanent. The first one is by evolving bones and the other at the bone connections. The interrupted connections are the joints (articulations) (Fig. 1). There is a more or less gap between the connecting bone parts. We talk about simple articulations when two bones meet. The components in the articulation can be permanent or collateral. In the simple articulations from the two bone parts one has a globular articulation head and the other has a concave score. The head of the articulation is not always globular, it can be cylindrical, elliptical. The shape of the articulation head is important because it determines the extent and quality of movement.

Figure 1 shows the structure and the parts of the articulations. In the articulation are the two connecting bone parts with articulation cartilage at the end, in the small gepp between the bones is the synovial fluid. The bones are actually held together by the articulation socket and fibers. The articulation system decided here is similar in all kind of articulation types in the hand. These articulation types are saddle joint, cylinder joint and ball joint.

B. Actuation anatomical components

The most common forces transmission agent in the human body is the tendon. The dense connective tissue is the most

N. Sárkány, "Biomimetic test bed hand,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 63-66.



(a) (b) (c)

Fig. 2: Human finger tendons articulations (Fig. 2a), Extensor-Flexor mechanism (Fig. 2b), Human hand muscle and tendon structure (Fig. 2c)

arranged form of connective tissues typical occurrence are tendons. Tendons are like strong braided ropes, they contains a lots of small fibers bundled to geather.

Muscles tissues are the source of force in the body due to this are we capable to move our limbs. There are three main type of muscle in the body: skeletal, smooth, cardiac. The process of movement in the hand has two parts an extensor and a flexor muscle. This way our body can move smooth and gently. Almost all degrees of freedom have its extensor flexor component.

III. Control and Sensing

The Central Nervous System (CNS) contains the spinal cord and several part of the brain. For the involuntary control of our body the main center is our spinal cord from skull down.

A. Somatic and autonomic motoric system

There are two types of control of our muscles a reflex action and a voluntary. The reflexes are not conscious they emerge from a low level of control from the spinal cord. These low level controls can be divided in two groups based on their source: muscle origin or external (skin) origin reflex.

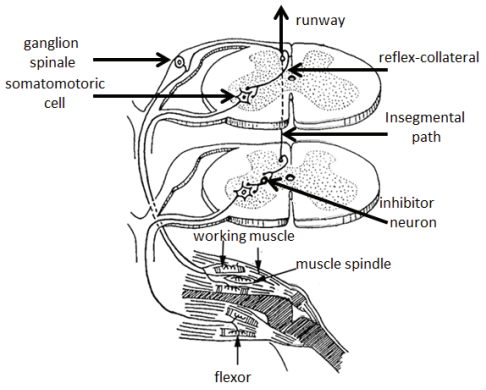


Fig. 3: Patela reflex

1) Proprioceptive (Myotacticus) reflex: The meaning of muscle origin reflex is that the reflex source is in the skeletal muscle, the receptor and the effector can be found in the same muscle. This kind of reflex is the myotacticus reflex. It is responsible for the control and to keep the length of the muscle and for its tension. For example the patela reflex (Fig. 3).

2) External (foreign) reflex: The meaning of the external origin is that the source of the reflex comes from the skin this implies that, the source of the stimulus the receptor is elsewhere then the effector. The main role for this reflex is to protect our system from external damage and is responsible for the sensing of heat, pain, and touch. The Sensory Epithelium is the part of the epithelium. Its role is to sense the stimulus of the outer world in our skin, it contains several kind of sensing receptor shown in figure 4. This receptors are the following: Mechanoreceptors which respond to mechanical stimuli (Ruffini's end organ (sustained pressure), Meissner's corpuscle (changes in texture, slow vibrations), Pacinian corpuscle (deep pressure, fast vibrations), Merkel's disc (sustained touch and pressure)), Thermoreceptores responsible for heat sensation, Nocioceptors responsible for sensing extreme pain and the Free nerve endings.

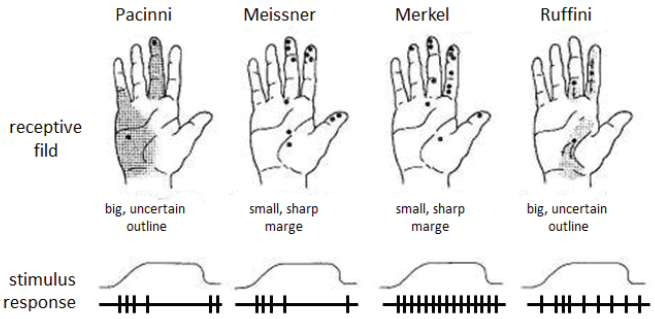


Fig. 4: The mechano receptors of the hand, its receptive fild, and the response to a stimulus.

B. Gamma -loop

CNS controls the muscles via nervous runways, it manipulates the reflex arcs. The nervus runaways afferent fibers activates the $A\gamma$ motoneurons. This neurons sense the passive elongation of the muscle. The elongation stimulate the specific receptors and they change theirs firing pattern. As a result the brain or a brain stem motoneuron activates and it activates the specific work muscle. So the gamma-loop (Fig. 5) is a system which is controlled by the CNS. Its basic functions are: posture, maintain stability, to change the intensity of the movement, exceptionally braking and acceleration and the execution of slow but sturdy movements.

IV. Finger Kinematics

First of all we must define the kinematic description of the system. A hand has five fingers the Thumb, the Index, the Middle, the Ring and the Little finger. The last four are similar in structure, kinematics and constraints. The thumb has two and the other four three links.

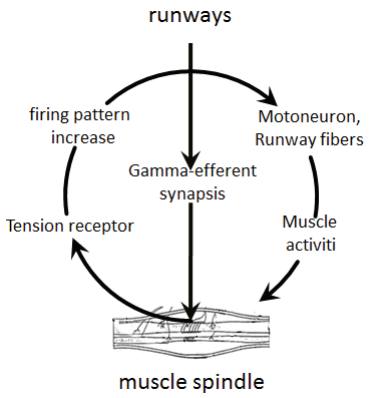
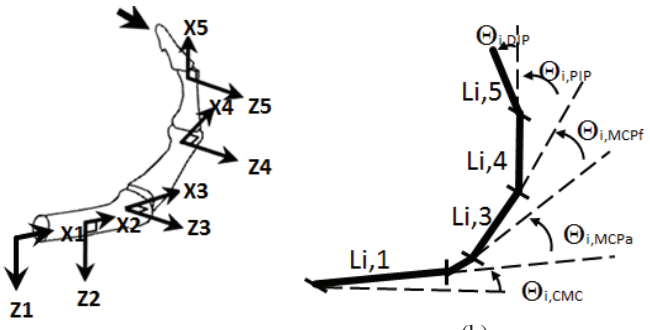


Fig. 5: Gamma -loop

A. Kinematic configuration

Figure 6 shows the kinematic configuration of the finger. Links and joints are defined as $L_{i,j}$ and $\Theta_{i,j}$, where i represents a finger (i = Index, Middle, Ring, Little) and j is the appropriate link or joint. The same kinematic configuration is used for the Index, Middle, Ring and Little finger. These configuration is define by five joints and four link, which are as follows metacarpal($L_{i,1}$), proximal($L_{i,3}$), middle($L_{i,4}$) and distal($L_{i,5}$). The joints are defined as carpometacarpal($\Theta_{i,CMC}$), metacarpo phalangeal is an articulation sellaris because of that it has 2 DOF's one for the adduction–abduction ($\Theta_{i,MCPa}$),and one for flexion–extension ($\Theta_{i,MCPf}$), proximal interrphalangial($\Theta_{i,PIP}$) and distal phalangeal ($\Theta_{i,DIP}$).



(a) (b)

Fig. 6: The kinematic representation of the human finger.

B. Forward Kinematics

Table I shows the Denavit-Hartenberg parameters of the finger. Joint are represent by $\Theta_{i,j}$, link are defined with $a_{i,j}$ this describes the length of a phalanges. The $d_{i,j}$parameter is 0 because the length of the phalanges are fixed, and $\alpha_{i,j}$ is the angle separation of Z_{i-1}, Z_i axis.

The forward kinematic of the finger is shown in equation IV-B, which is based on table I.

	$\Theta_{i,j}$	$d_{i,j}$	$a_{i,j}$	$\alpha_{i,j}$
1	$\Theta_{i,CMC}$	0	$L_{i,1}$	$\pi/2$
2	$\Theta_{i,MCPe}$	0	0	$-\pi/2$
3	$\Theta_{i,MCPf}$	0	$L_{i,3}$	0
4	$\Theta_{i,PIP}$	0	$L_{i,4}$	0
5	$\Theta_{i,DIP}$	0	$L_{i,5}$	0

TABLE I: The D-H parameters for a finger

$$O_i = {}^{-1}_{0}T_i(u_i) * {}^{0}_{1}T_i(\Theta_{i,CMC}) * {}^{1}_{2}T_i(\Theta_{i,MCPa}) * {}^{2}_{3}T_i(\Theta_{i,MCPf}) * {}^{3}_{4}T_i(\Theta_{i,PIP}) * {}^{4}_{5}T_i(\Theta_{i,DIP}) \quad (1)$$

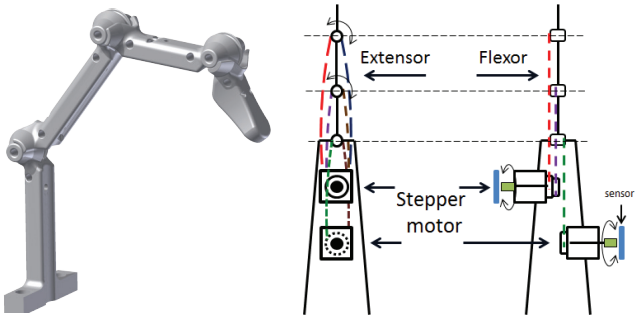
Where the phrases in equation IV-B are; O_i a matrix contains position and orientation of the $i-$ th fingertip with respect to the center of th wrist, u_i vector between the center of the wrist and the $i-$ th finger reference frame, ${}^{0}_{5}T_i(\Theta_{i,j})$ homogeneous matrix between $i-$ th finger reference frame and the finger tip. This matrix is the homogenous transformation matrix for a given phalanges.

$${}^{k-1}_{k}T_i(\Theta_{i,j}) = \begin{pmatrix} c(\Theta_{i,j}) & -c(\alpha_{i,j})s(\Theta_{i,j}) & s(\alpha_{i,j})s(\alpha_{i,j}) & a_{i,j}c(\Theta_{i,j}) \\ s(\Theta_{i,j}) & c(\alpha_{i,j})c(\Theta_{i,j}) & s(\alpha_{i,j})c(\alpha_{i,j}) & a_{i,j}s(\Theta_{i,j}) \\ 0 & s(\alpha_{i,j}) & c(\alpha_{i,j}) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

V. Biomimetic finger test bed

The Biomimetic finger is designed by reproducing, as close as possible to the size and kinematics of the human finger. Figure 7a shows the first prototype of the biomimetic hands finger

The setup of a finger actuation, sensing and control is shown in Figure 7b. A finger has two actuator, the $\Theta_{i,PIP}$,$\Theta_{i,DIP}$ joints have one common actuator but with various ratios to achieve the linkage between the links. The $\Theta_{i,MCP}$ has one actuator in this setup the joint has only 1 DOF. Every DOF has an extensor-flexor component, this method is accomplished by one actuator just like shown in Figure 7b.



(a) (b)

Fig. 7: The testbed concept of the biomimetic finger 7b, and the finger prototype 7a.

An actuator unit contains the stepper motor, sensors (po-

sition, torque, strain-gauges) and the extensor-flexor tendon pairs.

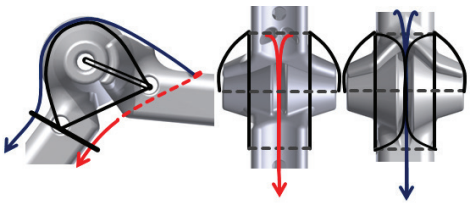


Fig. 8: The articulation of the biomimetic joint.

In figure 8 is the assemble of one biomimetic joint with the covering fibers and tendons for the corresponding link.

VI. Control & Sensing

The methods which have been mentioned in Section III provide the basis for the control and sensing methods which are be realized in the test bed.

A. Control

The concept of the control is "*divide et impera*", who wants to know everything, really knows nothing. This resource focus on the low level control of the test bed, to achieve a stable resting posture with compliance. The concept is shown in Figure 9, basically the complete fine motor control is separated from the abstract high control level. What does that mean? For example the manipulation task is to hold a bottle of water in the hand while rotating the wrist the direction of the gravity is changing at the level of joints and this changing force is not adjusted by the high level because this kind of information do not even get to it. That is the concept; we do not want to bother the high level with this kind of information.

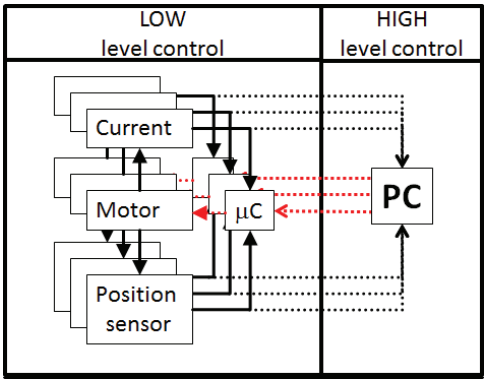


Fig. 9: The block diagram of the biomimetic joint control model

B. Sensing

Just like in the human body we want to have adequate telemetry. This can be achieved with a lot of different type of sensor. What can we sense? We can sense touch with a lot of types of sensors based on different technologies (e.g. capacitive, resistive, optical, etc.). We can sense torque a many different ways calculated from the position or measuring current. In this resource we use four kinds of sensors. 3D touch sensors in the phalanges, strain-gauges in the articulations to measure the strain in different directions, magnetic rotary sensors to have the accurate angle information of an articulation, basic current measuring to determ the torque between two links during a manipulation or a resting posture.

VII. Conclusion

The functional test showed promising results, but there is still room for improvement. First of all the investigation of the movement of the human hand to achieve "human-like" behavior, the muscle structure of the forearm to improve a better optimal strategy for the actuation system to reduce the amount of actuators. A second important objective that we are pursuing is to implement a global–control, based on the functional, biological structure of the cerebellum. This very challenging goal could ultimately lead to the development of a novel biomechatronic hand.

Acknowledgment

I would like to thank the multidisciplinary doctoral school at the Faculty of Information Technology of the Pázmány Péter Catholic University. The author is also grateful to the members of the Robotics lab for the discussions and their suggestions. Special thanks for project TÁMOP-4.2.1./B-11/2/KMR-2011-002, TÁMOP-4.2.2./B-10/1-2010-0014 for the suppor.

References

[1] H. Kawasaki, T. Komatsu, and K. Uchiyama, "Dexterous anthropomorphic robot hand with distributed tactile sensor: Gifu hand II," *Mechatronics, IEEE/ASME Transactions on*, vol. 7, no. 3, pp. 296–303, 2002.

[2] V. Weghe, M. Rogers, M. Weissert, and Y. Matsuoka, "The ACT hand: design of the skeletal structure," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 4, pp. 3375–3379, IEEE, 2004.

[3] J. Butterfass, M. Grebenstein, H. Liu, and G. Hirzinger, "DLR-Hand II: Next generation of a dextrous robot hand," in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, vol. 1, pp. 109–114, IEEE, 2006.

[4] M. Carrozza, B. Massa, S. Micera, R. Lazzarini, M. Zecca, and P. Dario, "The development of a novel prosthetic hand-ongoing research and preliminary results," *Mechatronics, IEEE/ASME Transactions on*, vol. 7, no. 2, pp. 108–114, 2002.

[5] S. Jacobsen, E. Iversen, D. Knutti, R. Johnson, and K. Biggers, "Design of the Utah/MIT dextrous hand," in *Robotics and Automation. Proceedings. 1986 IEEE International Conference on*, vol. 3, pp. 1520–1532, IEEE, 2002.

[6] C. Lovchik and M. Diftler, "The robonaut hand: A dexterous robot hand for space," in *Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on*, vol. 2, pp. 907–912, IEEE, 2002.

[7] R. Drake, W. Vogl, and A. Mitchell, *Gray's anatomy for students*. Elsevier/Churchill Livingstone Philadelphia, 2005.

[8] J. Doyle and M. Botte, *Surgical anatomy of the hand and upper extremity*. Lippincott Williams & Wilkins, 2003.

[9] J. Criag, "Introduction to robotics: mechanics and control," 2005.

[10] R. Tubiana, "Architecture and functions of the hand," *The hand*, vol. 1, 1981.

[11] G. Vasarhelyi, M. Adam, E. Vazsonyi, Z. Vizvary, A. Kis, I. Barsony, and C. Ducso, "Characterization of an integrable single-crystalline 3-d tactile sensor," *Sensors Journal, IEEE*, vol. 6, no. 4, pp. 928–934, 2006.

[12] G. Monkman, S. Hesse, and R. Steinmann, *Robot grippers*. John Wiley and Sons, 2007.

Kisspeptinimmunoreactivity in human gonadotropin-releasing hormone neurons

MátéSipos
(Supervisors: Erik Hrabovszky Dr., Krisztina Kovács Dr.)
sipos.mate@koki.mta.hu

I. Summary

Background: Central kisspeptin signaling to gonadotropin releasing hormone (GnRH) neurons plays a crucial role in mammalian reproduction and represents a potential therapeutic target for the treatment of human infertility. While our understanding of this communication mostly relies on experimental data from laboratory animals, recent immunohistochemical studies on post mortem hypothalami demonstrated that neuroanatomical characteristics of the hypothalamic kisspeptin system largely differ in humans and extensively studied model animals.

Aim: The goal of the present study was to address whether GnRH neurons in the human synthesize kisspeptin, in addition to receiving the previously identified kisspeptin-immunoreactive inputs.

Methods: Histological sections were prepared from autopsy samples of male (n=11; 21-78 yr) and female (n=6; 57-70 yr) human individuals. Immunofluorescence experiments, followed by confocal analysis, were carried out for the simultaneous detection of kisspeptin-54 and GnRH.

Results: A subpopulation of GnRH-immunoreactivesomata (9.5±4.3% in men and 7.9±2.7% in postmenopausal women) and their dendrites expressed kisspeptinimmunoreactivity; GnRH-immunoreactive axons, in contrast, contained kisspeptin rarely and kisspeptin signal was absent from GnRH-immunoreactive terminals in the infundibular stalk. Preabsorbtion of kisspeptin antibodies with a mixture of kisspeptin-related RF-amide peptides (neuropeptides VF, FF, AF and prolactin releasing peptide) did not eliminate kisspeptin signal in GnRH neurons, providing support for labeling specificity.

Conclusion: This study reveals an interesting colocalization phenomenon between kisspeptin and GnRH in 8-10% of human GnRH neurons. We propose that release of endogenous kisspeptin, which likely occurs from the somatodendritic compartment, participates in autocrine/paracrine regulatory mechanisms in the microenvironment of dual-labeled GnRH neurons.

II. Introduction

Kisspeptin (KP) signaling through its receptor (KISSR1; previously called GPR54) plays a pivotal role in human reproduction; inactivating mutations of the genes encoding for KP (1) or KISSR1 (2,3) cause hypogonadotropichypogonadism. KP acts as a potent stimulator of LH secretion in various species, including humans (4,5). Our understanding about the mechanisms of KP actions relies on information obtained primarily from studies of laboratory animals. In rodents, KP-induced LH secretion can be prevented by the gonadotropin-releasing hormone (GnRH) antagonist acyline (6), indicating that KP acts via increasing hypothalamic GnRH release. KP directly activates GnRH neurons which express KISS1R (7) and respond to KP with depolarization (8) and cFos expression (7). Immunohistochemical evidence from humans indicates that KP synthesizing hypothalamic neurons communicate with the GnRH neuronal system via forming axo-somatic, axo-dendritic and axo-axonal contacts (9,10). Additional central mechanisms whereby KP influences the reproductive axis remain poorly understood. Based on preliminary evidence that we obtained in the course of earlier immunofluorescence experiments (10-12), here we investigated the presence of KP immunoreactivity in human GnRH neurons that may underlie putative autocrine/paracrine regulatory mechanisms.

III. Materials and methods

Human subjects: Hypothalamic tissue samples from eleven men of variable ages (21-78 ys) and six postmenopausal women (57-70 ys) were obtained from autopsies at the Forensic Medicine Department of the University of Debrecen with the permission of the Regional Committee of Science and Research Ethics of the University of Debrecen (DEOEC RKEB/IKEB: 3183-2010). Selection criteria included sudden causes of death, lack of known neurological and endocrine disorders and post mortem delay below 36h.

Section preparation: The hypothalamic tissue blocks were immersion-fixed with 4% formaldehyde in 0.1M phosphate buffer saline (PBS; pH 7.4) for 7-14 days. The blocks were trimmed to include the optic chiasmarostrally, the mammillary bodies caudally and the anterior commissure dorsally (9,10). Two sagittal cuts were placed 2cm bilaterally from the midline. The blocks were bisected into right and left halves and then, infiltrated with 20%

M. Sipos, "Kisspeptinimmunoreactivity in human gonadotropin-releasing hormone neurons,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 67-70.

sucrose for 5 days. The right hemihypothalami were sectioned coronally at 30µm with a Leica SM 2000R freezing microtome (Leica Microsystems, NusslochGmbh, Germany Leica Microsystems).

Pretreatments: The tissues were permeabilized and endogenous peroxidase activity reduced using a mixture of 0.2% Triton X-100 and 0.5% H2O2 for 30 min. Antigen epitopes were unmasked using 0.1M citrate buffer (pH 6.0) at 80 ºC for 30 min (10). For immunofluorescence, the sections were additionally treated with Sudan black (9,10). Every 24th preoptic-hypothalamic section was used both in the dual-immunoperoxidase and in the dual-immunofluorescence experiments.

Simultaneous detection of KP and GnRH with two-color immunoperoxidasehistochemistry: Detection of KP: As in a series of recent immunohistochemical experiments on human hypothalami (9-12), the sections were incubated for 48 h at 4C in sheep polyclonal antibodies against human kisspeptin-54 (GQ2; 1:200,000)(4), followed by biotin-SP-antisheepIgG (Jackson ImmunoResearch Laboratories, West Grove, PA, USA; 1:500) and the ABC Elite reagent (Vector, Burlingame, CA; 1:1000) for 60 min each. The peroxidase reaction was visualized with nickel-intensified diaminobenzidine, and enhanced with silver-gold. Detection of GnRH: GnRH neurons were detected with a recently characterized guinea pig antiserum (#1018; 1:50,000) (10). The primary antibodies were reacted with biotin-SP-anti-guinea pig IgG (Jackson ImmunoResearch Laboratories; 1:500; 1h) and then, the ABC Elite reagent (Vector; 1:1000; 1h). The peroxidase reaction was visualized with brown diaminobenzidine.

Dual-label fluorescent immunohistochemistry: Another series of sections was used to study the colocalization of KP and GnRH. Incubation in a cocktail of primary antibodies (sheep anti-kisspeptin, 1:1000; guinea pig anti-GnRH, 1:3000) for 48h at 4ºC was followed by a cocktail of fluorochrom-conjugated secondary antibodies (anti-sheep-Cy3, 1:1000 and anti-guinea pig-FITC, 1:250; Jackson ImmunoResearch) for 6h.

Specificity control experiments: A series of sections was used in control experiments to minimize the possibility of non-specific KP labeling. Controls included omission of either the primary or the secondary antibodies from the labeling cocktails. Although in vitro cross-reactivity of the GQ2 antibodies with other KP-related human RF amide peptides has already been addressed and found to be below 0.01% (4), here we also tested with immunohistochemistry if KP immunoreactivity persists after a 12h preabsorbtion of the GQ2 primary antibodies with 10µg/ml of each of four different human RF amide peptides: neuropeptides VF, FF, AF and prolactin releasing peptide.

Section mounting and coverslipping: The immuno-peroxidase labeled sections were mounted from Elvanol and coverslipped with DPX mounting medium (FlukaChemie; Buchs, Switzerland). Immunofluorescent specimens were mounted on silanized slides and coverslipped with Mowiol.

Digital photography: Light and fluorescent microscopic images were scanned with an AxioCamMRc 5 digital camera mounted on a Zeiss AxioImager M1 microscope using the AxioVision 4.6 software (Carl Zeiss, Göttingen, Germany). Confocal images were prepared with a Radiance 2100 confocal microscope (Bio-Rad Laboratories, Hemel Hempstead, UK) using laser excitation lines 488 nm for FITC and 543 nm for Cy3 and dichroic/emission filters 560 nm/500–530 nm for FITC and 560–610 nm for Cy3. Emission cross-talk was avoided using the lambda strobing mode. Images were processed with the Adobe Photoshop CS software (Adobe Systems, San José, CA, USA).

IV. Results

Results of dual- immunoperoxidase labeling: In accordance with previous observations (9-12), many KP-R cell bodies were detectable in the Infundibular nucleus using the GQ2 primary antibodies and the black silver-gold-intensified nickel diaminobenzidinechromogen (Fig. A). Typical KP neurons were round, faintly stained and their dendritic arborization was visualized poorly (Figs. A2, A3), whereas some bipolar and intensively stained KP-IR cells that resembled GnRH neurons were also detectable (Fig. A1). These atypical KP cells occurred sporadically in preoptic/hypothalamic regions, without showing characteristic topography. GnRH-IR neurons were scattered throughout the preoptic area and the hypothalamus and occurred relatively often in the Inf where they intermingled with KP-IR cells (Fig. A). The absence of gray-black KP labeling in briownGnRH-IR neurons indicated that many GnRH cells are devoid of KP immunoreactivity (Fig. A) but did not exclude GnRH synthesis in fusiform KP neurons that were stained black (Fig. A1).

Results of colocalization experiments with fluorescent immunohistochemistry: In dual-immunofluorescent specimens, most KP-IR and GnRH-IR elements were distinct. Occassionally, the soma and proximal dendrites of GnRH neurons were KP-immunopositive (Figs. B, C). Dual-immunolabeled neurons occurred sporadically in the preoptic area (Fig. B) and in various hypothalamic nuclei, including the Inf (Fig. C). Their varicose axons were devoid of KP labeling (Fig. B1), with rare exceptions (Fig. B2). GnRH-IR and KP-IR axons in the infundibular stalk (Figs. E, E1) and their hypophysiotropic terminals were distinct. Quantitative analysis of 380 GnRH-IR perikarya identified 34 KP-IR cells (8.9±3.6%; mean±SEM). The extent of KP colocalization varied among individual samples without detectable effects of sex (9.5±4.3% in men and 7.9±2.7% in postmenopausal women; P=0.80 by ANOVA) and age (P=0.86, by multiple regression). Furthermore, KP colocalization in rostrally positioned GnRH neurons of the preoptic area and rostral hypothalamus (plates 20-23 of Mai (13) and in caudal

hypothalamic GnRH neurons (plates 24-28 of Mai (13)) did not differ (7.4±2.4% and 13.0±0.7%, respectively; P=0.12 by ANOVA).Omission of any one of the primary or secondary antibodies eliminated double-labeled neurons. Furthermore, presence of KP signal in GnRH neurons was unaffected by preabsorbtion of the primary antibody cocktail with four different RF amide peptides (Fig. 1D).

V. Discussion

Results of this dual-label immunohistochemical study provide novel evidence for KP immunoreactivity in a subset (9%) of GnRH neurons and their dendritic processes in the human hypothalamus.

Animal models of human reproduction have innate limitations due to many species differences in hypothalamic regulatory mechanisms, also related to central KP signaling to GnRH neurons. The majority of KP neurons in the ARC co-synthesize neurokinin B and dynorphinA, as reported first for the sheep (14) and later for the rodent (15). In contrast, we have recently shown that the colocalization between NKB and KP in the human Inf is only partial and its extent highly depend on age (12) and sex (10). In addition, we observed only very low degree of overlap between KP-IR and dynorphin A-IR neuronal elements (11). The partial colocalization of KP and GnRH we report in this study reflects an interesting further difference between the human and previously studied laboratory animals. While rat GnRH neurons do not express KP immunoreactivity (16), 90% of GnRH neurons in the ovine preoptic area contained KP immunoreactivity in a previous study (16). Later studies identified that this colocalization was due to non-specific antiserum binding (14) and results could not be confirmed with more specific KP antibodies (17). Here we recognize that antibody cross-reaction with unwanted targets always represents a pitfall in immunohistochemical studies. In the absence of knockout control approaches that would be readily available in mice, we cannot fully exclude that KP immunoreactivity in human GnRH neurons results from the presence of KP-like molecules and not KP per se. To minimize this possibility, we have replicated the colocalization studies using primary antibodies that had been preabsorbed with four different RF-amide peptides similar to KP at their C-terminal '-RF-amide' motif. The persistance of KP/GnRHcolocalization in this control study makes it unlikely that the GQ2 antibodies significantly cross-react with other RF-amide peptides, as concluded previously using radioimmunoassay (4). The functional significance of KP in human GnRH neurons requires clarification. It seems likely that KP-IR neuronal afferents on the cell bodies, dendrites and axons of human GnRH neurons (9,10) represent the primary route whereby central KP signaling influences reproduction in humans. In addition to this communication route conserved in animals and humans, the interesting colocalization phenomenon we report in this study suggests that endogenous KP may be synthesized and likely, released from about 9% of human GnRH neurons to participate in autocrine/paracrine regulatory mechanisms. GnRH neurons in different species are anatomically interconnected by classical synapses (18), cytoplasmic bridges (18) and dendro-dendritic appositions (19). Release of endogenous KP from activated human GnRH neurons may contribute to the chemical stimulation of neighboring GnRH neurons within the network via KISSR1. The somewhat unexpected absence of KP immunoreactivity in the majority of GnRH-IR axons and their terminals around the hypophysial portal vasculature of the postinfundibular eminence indicates that most of this putative communication may take place at the somato-dendritic compartment of the GnRH cell.

In summary, this study provided novel proof for KP immunoreactivity in a subset of human GnRH neurons and their dendrites, but typically, not in GnRH-IR axons and their terminals. The release of endogenous KP from activated GnRH neurons may underlie an important autocrine/paracrine signaling mechanism for the coordinated electric and secretory activity of the human GnRH neuronal network.

VI. Further plans

A) Mouse GnRH IL-1 receptor deficient line (exon 5 floxed) with CRE-Lox technology is currently being bred to investigate IL-1 effects on the reproductive axis. Cre-Lox recombination technique allows us to carry out a specific deletion of IL-1 receptor on GnRH neurons.

B) Future experiments include the investigation of KP receptor GPR54 distribution with a unique RNA probe in situ hybridization histochemistry technique. We are planning to explore interactions between KP and the HPA axis with morphological methods such as quantitative analysis of IHC labeled sections to reveal possible neuronal contacts between CRH and KP neurons. We are planning to measure stress effects in chronic and acute stress situations on the expression of KP and KP receptors.

C) endocrine disruptor compounds are natural or synthetic molecules with the ability to disrupt physiological endocrine functions. Our further plans include investigations of the effects of various endocrine disruptors on the reproductive axis of pre-pubertal and adult rodents. We are planning to monitor the effects of these EDCs on hormone (GnRH, LH, FSH, KP, ACTH) secretion and receptor (GnRH, KP, CRH, AVP) expression, that are involved in the regulation of the reproductive axis.

VII. References

[1] Opaloglu AK, Tello JA, Kotan LD, Ozbek MN, Yilmaz MB, Erdogan S, Gurbuz F, Temiz F, Millar RP, Yuksel B. **Inactivating KISS1 mutation and hypogonadotropichypogonadism. The New England journal of medicine.** 2012;366(7):629-635.

[2] de Roux N, Genin E, Carel JC, Matsuda F, Chaussain JL, Milgrom E. **Hypogonadotropichypogonadism due to loss of function of the KiSS1-derived peptide receptor GPR54.**ProcNatlAcadSci U S A. 2003;100(19):10972-10976.

[3] Seminara SB, Messager S, Chatzidaki EE, Thresher RR, Acierno JS, Jr., Shagoury JK, Bo-Abbas Y, Kuohung W, Schwinof KM, Hendrick AG,

Zahn D, Dixon J, Kaiser UB, Slaugenhaupt SA, Gusella JF, O'Rahilly S, Carlton MB, Crowley WF, Jr., Aparicio SA, Colledge WH. **The GPR54 gene as a regulator of puberty.** N Engl J Med. 2003;349(17):1614-1627.

[4] Dhillo WS, Chaudhri OB, Patterson M, Thompson EL, Murphy KG, Badman MK, McGowan BM, Amber V, Patel S, Ghatei MA, Bloom SR. **Kisspeptin-54 stimulates the hypothalamic-pituitary gonadal axis in human males.**J ClinEndocrinolMetab. 2005;90(12):6609-6615.

[5] Dhillo WS, Chaudhri OB, Thompson EL, Murphy KG, Patterson M, Ramachandran R, Nijher GK, Amber V, Kokkinos A, Donaldson M, Ghatei MA, Bloom SR. **Kisspeptin-54 stimulates gonadotropin release most potently during the preovulatory phase of the menstrual cycle in women.**J ClinEndocrinolMetab. 2007;92(10):3958-3966.

[6] Gottsch ML, Cunningham MJ, Smith JT, Popa SM, Acohido BV, Crowley WF, Seminara S, Clifton DK, Steiner RA. **A role for kisspeptins in the regulation of gonadotropin secretion in the mouse.**Endocrinology. 2004;145(9):4073-4077.

[7] Irwig MS, Fraley GS, Smith JT, Acohido BV, Popa SM, Cunningham MJ, Gottsch ML, Clifton DK, Steiner RA. **Kisspeptin activation of gonadotropin releasing hormone neurons and regulation of KiSS-1 mRNA in the male rat.**Neuroendocrinology. 2004;80(4):264-272.

[8] Han SK, Gottsch ML, Lee KJ, Popa SM, Smith JT, Jakawich SK, Clifton DK, Steiner RA, Herbison AE. **Activation of gonadotropin-releasing hormone neurons by kisspeptin as a neuroendocrine switch for the onset of puberty.**J Neurosci. 2005;25(49):11349-11356.

[9] Hrabovszky E, Ciofi P, Vida B, Horvath MC, Keller E, Caraty A, Bloom SR, Ghatei MA, Dhillo WS, Liposits Z, Kallo I. **The kisspeptin system of the human hypothalamus: sexual dimorphism and relationship with gonadotropin-releasing hormone and neurokinin B neurons.**Eur J Neurosci. 2010;31(11):1984-1998.

[10] Hrabovszky E, Molnar CS, Sipos M, Vida B, Ciofi P, Borsay BA, Sarkadi L, Herczeg L, Bloom SR, Ghatei MA, Dhillo WS, Kallo I, Liposits Z. **Sexual dimorphism of kisspeptin and neurokinin B immunoreactive neurons in the infundibular nucleus of aged men and women.**Frontiers in Endocrinology. 2011;2.

[11] Hrabovszky E, Sipos MT, Molnar CS, Ciofi P, Borsay BA, Gergely P, Herczeg L, Bloom SR, Ghatei MA, Dhillo WS, Liposits Z. **Low degree of overlap between kisspeptin, neurokinin B, and dynorphinimmunoreactivities in the infundibular nucleus of young male human subjects challenges the KNDy neuron concept.**Endocrinology. 2012;153(10):4978-4989.

[12] Molnar CS, Vida B, Sipos MT, Ciofi P, Borsay BA, Racz K, Herczeg L, Bloom SR, Ghatei MA, Dhillo WS, Liposits Z, Hrabovszky E. **Morphological evidence for enhanced kisspeptin and neurokinin B signaling in the infundibular nucleus of the aging man.**Endocrinology. 2012;153(11):5428-5439.

[13] Mai J, Assheuer J, Paxinos G, eds. **Atlas of the human brain.** San Diego: Academic Press; 1997.

[14] Goodman RL, Lehman MN, Smith JT, Coolen LM, de Oliveira CV, Jafarzadehshirazi MR, Pereira A, Iqbal J, Caraty A, Ciofi P, Clarke IJ. **Kisspeptin neurons in the arcuate nucleus of the ewe express both dynorphin A and neurokinin B.** Endocrinology. 2007;148(12):5752-5760.

[15] Navarro VM, Gottsch ML, Chavkin C, Okamura H, Clifton DK, Steiner RA.**Regulation of gonadotropin-releasing hormone secretion by kisspeptin/dynorphin/neurokinin B neurons in the arcuate nucleus of the mouse.**J Neurosci. 2009;29(38):11859-11866.

[16] Pompolo S, Pereira A, Estrada KM, Clarke IJ. **Colocalization of kisspeptin and gonadotropin-releasing hormone in the ovine brain.**Endocrinology. 2006;147(2):804-810.

[17] Smith JT, Coolen LM, Kriegsfeld LJ, Sari IP, Jaafarzadehshirazi MR, Maltby M, Bateman K, Goodman RL, Tilbrook AJ, Ubuka T, Bentley GE, Clarke IJ, Lehman MN. **Variation in kisspeptin and RFamide-related peptide (RFRP) expression and terminal connections to gonadotropin-releasing hormone neurons in the brain: a novel medium for seasonal breeding in the sheep.**Endocrinology. 2008;149(11):5770-5782.

[18] Witkin JW, O'Sullivan H, Silverman AJ. **Novel associations among gonadotropin-releasing hormone neurons.**Endocrinology. 1995;136(10):4323-4330.

[19] Campbell RE, Gaidamaka G, Han SK, Herbison AE. **Dendro-dendritic bundling and shared synapses between gonadotropin-releasing hormone neurons.**ProcNatlAcadSci U S A. 2009;106(26):10835-10840.

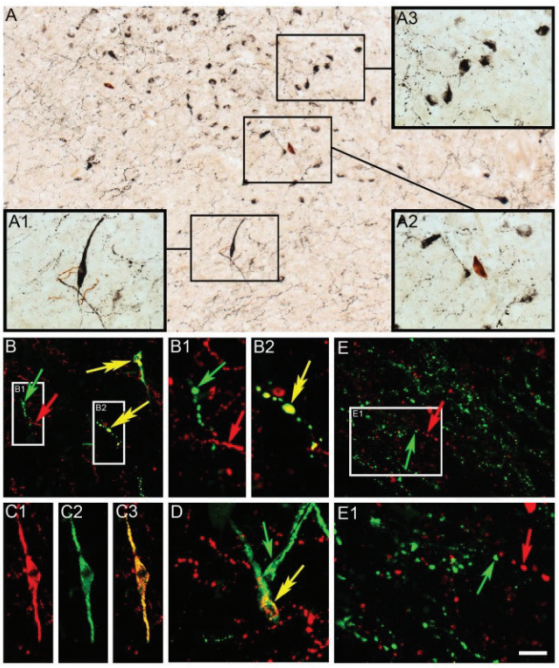


Fig.Demonstration of kisspeptinimmunoreactivity in a subset of gonadotropin-releasing hormone neurons. A representative image of the infundibular nucleus in **A** (62 year old male) illustrates kisspeptin (KP; black) and gonadotropin-releasing hormone (GnRH; brown) immunoreactivities. The high-power inset in **A1** shows an atypical fusiform KP-immunoreactive (IR) neuron which highly resembles GnRH neurons. Insets **A2** and **A3**, in turn, illustrate the typical KP neurons from the Inf; these cells are stained lightly, exhibit round shape and show no extensive labeling of their dendritic tree. While the dark KP labeling in **A1** may mask the GnRH signal of a putative GnRH/KP dual-phenotype neuron, the lack of any KP signal in other GnRH neurons (brown neuronal structures in **A2**) indicates that many GnRH cells are devoid of KP synthesis. Confocal images of dual-immunofluorescent sections reveal that KP (red) and GnRH (green) signals are usually distinct (red and green arrows), but occasionally overlap (yellow double-arrows). Sporadic GnRH-IR neurons co-labeled for KP occur in all regions where GnRH neurons are present, including the preoptic area (**B**; 21 year old male). High-power insets reveal that the varicose axons of GnRH neurons tend not to contain KP labeling (**B1**), except for very rare cases of GnRH/KP double-labeled axons (**B2**). Various hypothalamic regions including the infundibular nucleus contain scattered neurons containing KP (**C1**) as well as GnRH (**C2**) signals. (**C3** is merged from **C1** and **C2**; 64 year old male). KP immunoreactivity of a subset of GnRH neurons persists if the primary KP antibody is preabsorbed with 10µg/ml of each of four different RF-amide peptides: neuropeptides VF, FF, AF and prolactin releasing peptide (**D**; 70 year old female). Note that single-labeled (green arrow) and double-labeled (double-arrow) GnRH neurons coexist at the same hypothalamic site. While about 9% of the GnRH-IR cell bodies exhibit KP labeling, hypophyisotropicGnRH axon projections in the infundibular stalk tend to be devoid of any KP signal (**E**; 63 year old female). Inset **E1** shows at high magnification that GnRH-IR and KP-IR axons are distinct. Scale bar=38µm in **A**, 21µm in **A1-3**, 33µm in **B**, 13µm in **B1-2**, 25µm in **C1-3**, **D**, 50µm in **E** and 20µm in **E1**.

A computer-aided setup for studying relations between EMG prediction, signals and muscular activity

Ádám Vály
(Supervisors: József Laczkó, Péter Szolgay)
valy.adam@itk.ppke.hu

Abstract–**The objective of this paper is to lay the basics for a computer system designed to examine relations between actual and predicted EMG signals combined with a movement analysis instrument. Such system would enable researchers to better understand muscular-neural relationships and improve EMG prediction methods for use in prosthetics.**

Keywords–**EMG prediction, bionics, software development**

I. INTRODUCTION

Human arm movements and EMG are the centerpiece studies of biomechanics. The exact relations between muscular electric signals and associated kinematics are not yet fully understood. In most cases, a special setup is required to carry out experiments on preselected subjects, such as in [1], [2], [3], [4] and [5]. A new, general purpose hardware and software system would enable a greater variety of experiments to be conducted using the same architecture.

In this paper, a universal EMG-motion processing system is proposed. This device can serve multiple purposes:

- EMG research – given a 6 channel EMG recording instrument, one could measure the electric signals associated with basic arm movements. Simple shoulder, elbow, wrist and finger movement EMGs could be processed to better understand muscle control.

- Rehabilitation – subjects affected with the partial or complete absence of neural control of their limbs could use this system to train their muscles. Researchers and therapists can design better training excercises considering recorded and processed EMG signals.

- Prosthetic design – considering a patient with a need for a basic or complex arm prostethic, researchers could use this system to design a device that uses associated electric inputs from nerve signals from another location. These signals could then drive an instrument, such as the one detailed in [7].

The combined EMG processing, prediction and motion system realizes a complex experimental platform, where multiple paradigms could be realized.

II. RESEARCH

Nowadays, Brain Computer Interfaces are the main signal sources for prosthetic control, which use scalp EEG signals as information source and control [8]. There may be a need for applications that use nerves that were originally responsible for controlling muscles as their input. There is significant amount of anatomical data and research available that show which nerves innervate certain muscles, however, the dynamics of EMG signals are still unclear, due to the complexity of even simple movement patterns. These problems may be better understood, if we had information on basic signal-motion relationships, and how these complex signals are generated and what is the role of each motor neuron. It is interesting to note the robustness of movement execution, meaning that in the presence of signal-dependent motor noise, the success ratio of movement tasks do not degrade significantly, as detailed in [4].

There are several effective methods for predicting muscle activity (EMG), such as polynomial curve fitting, Bayesian density estimation and dynamic neural networks. Johnson and Fuglevand [1] found that the neural network method may be best suited for prosthetics use. Their article also gives details on how to program signal processing units and off-line calculation software, Matlab (*Mathworks,* Natick, MA, USA).

Another question is that what forces are associated with a reaching or a holding task, and how does the geometry of the limb affect kinematics. Articles [3] and [9] describe relevant experiments and results.

III. TEST PROCEDURE

For a convenient test environment, surface EMG would be used to measure electric signals. sEMG signals would be received from six channels similar to the method described in [5]. The sEMG electrodes are attached to the subject's muscles, depending on the test planned. For EMG measurements, the able-bodied participant is asked to do the movement previously specified in the test schedule. The first processing stage is realized in the 6-channel sEMG instrument, and then sent to the control PC via Bluetooth. The qualities of sEMG signals are detailed in [6].

Á. Vály, "A computer-aided setup for studying relations between EMG prediction, signals and muscular activity,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 71-74.

Reversing this, the same setup combined with the movement analysis device can be used to measure kinematics and test EMG prediction methods. If we have a pre-defined kinematic, previously predicted electric signals could be fed into the participant's limb, after which we can calculate the error from the new kinematic compared to the desired.

For therapic applications, one should be able to use the system without the help of a trained operator, meaning that the participant should be able to select the correct tests.

The program combining these devices should provide the testing environment, controls, indicators and the necessary procedures to document experiments, in MS Excel, for example.

IV. SOFTWARE DEVELOPMENT

The control software of the complex EMG instrument has to fulfill many functions in different roles. It would be responsible for configuring the associated hardware and variables for the selected task, giving feedback to the user via indicators and graphs, supervising measurement and control processes and if necessary, correcting errors during operation. Therefore, the system has to be very flexible but robust.

To create a program that matches these criteria, two software development systems can be used:

- National Instruments LabVIEW
- MATLAB Simulink

The first task in programming this software is getting to know the hardware that will be integrated into the system. LabVIEW and Simulink both offer ready to use serial communication tools, so the long process of creating a low-level code to control a serial port can be avoided.

Both environments are graphical based and provide a convenient way to develop, simulate and operate instruments. The graphical method allows other developers to easily contribute to the software, if necessary. It also allows program diagrams and flowcharts to be printed and presented.

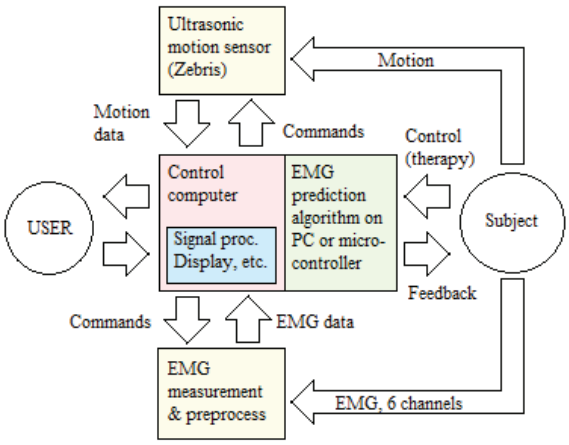


Figure 1.: the architecture of the proposed system.

Figure 1 depicts the architectural elements and relations for the planned software. The key role is played by the control software, which runs on the central PC and interfaces with the instruments via USB or other universal communication protocol, sending commands and receiving raw of preprocessed measurement data.

The EMG measurement instrument is instructed by the control computer to record and process EMG signals. It should be examined, wether the recorded and/or processed signals should be sent real-time to the control computer, or it would be better to do the transfer after all processes are completed. The received data may also contribute to the operation of other units, like in the case of a therapy.

The muscle activity prediction unit may be an application specific hardware or a software running on the control PC. The realization of this part depends on the ability of the programmer to utilize multi-core processors and parallelize control and prediction tasks. The prediction algorithm can be designed in a feedback manner, meaning that the same prediction algorithm may not be applicable to different participants (due to their physiological differences), and may need adjustment. The EMG signals recorded from the subject may affect the operation of the predictor. Testing the algorithm requires an apparatus which enables electric signals to be fed into the participant's limb.

The method for examining the correctness of the prediction process looks as the following:

- a desired kinematic is defined and recorded on the control computer
- the predictor calculates the signals
- the signals are fed into the subject
- participant makes the move
- the ultrasonic motion detector measures the movement made by the subject
- the movement data is sent to the control PC
- the error of the kinematic is calculated from the recorded movement and the desired kinematic
- the prediction algorithm is adjusted according to the error
- the experiment is repeated until the error is reduced below an acceptable level

With the process completed, the algorithm is adjusted to the subject, and produces the desired kinematic with a tolerable or with no error. The test data is correctly recorded in an appropriate format or in a printable record.

The Zebris ultrasonic motion sensor uses ultrasound to measure the position of microphones attached to the participant. The control computer instructs Zebris to begin

the test and send test data. Measurement data can be processed run-time or off-line according to the application. Statistical calculations can be made om the PC if a sufficiently large data set is available. This feature allows researchers to have a new perspective on the relations between EMG and kinematics.

Other features of the architectural elements can be exploited to create new and enhance existing features.

A possible operational flowchart of the control software can be depicted as follows:

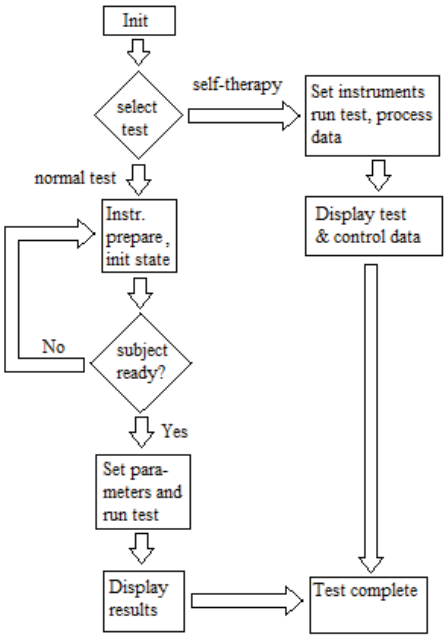


Figure 2.: test flowchart

In the first step, the program initializes, loads the control environment and requests a status report on the connected hardware. After that, the user is prompted to select a test type. There are two types of tests depending on the user of the system. Obviously, all tests require participants. In the case of normal tests, there is also an operator who is aware of the test conditions, schedule and tasks. The operator sets the software to perform the predefined test program, observes and if necessary, interacts with the control software. The operator is also responsible for communicating with the test participant.

In case of a therapy, the subject can control the system to prepare the hardware for a training excercise, or show test data, history and statistics. Here, the participant is in control and observing the system.

V. CONCLUSION

After the basic concepts of the test system have been specified and validated, sub-tasks can be designed and implemented. A modular programming paradigm allows for flexibility, robustness, good error handling and program expandability.

By combining an EMG measuring instrument, an EMG prediction system and an ultrasonic motion detector (such as the Zebris at PPKE ITK) with an advanced control software, a powerful, multi-purpose system can be created.

Graphical programming tools allow to create a flexible, user-friendly testing environment for analyzing kinematics, EMG and muscular activity. USB interface connects the specific hardware to the control PC. Measured and processed data may affect the working parameters of other elements.

The system can be used for both research and therapy, depending on the selected test and the needs of the participant and the operator.

Research into EMG characteristics, novel prostethic design and EMG prediction methods greatly help and define the specification of the system.

VI. ACKNOWLEDGMENTS

The author wishes to acknowledge the guidance of József Laczkó and Péter Szolgay. Regarding the Zebris Ultrasonic instument, the help of Zsolt Győrffy and Bence Borbély is much appreciated. This work is supported by the Péter Pázmány Catholic University, grants TÁMOP-4.2.1./B-11/2-KMR-2011-0002 and TÁMOP-4.2.2./B-10/1-2011-0014 by the European Union.

VII. REFERENCES

[1] L. A. Johnson, A. J. Fuglevand:
„Evaluation of probabilistic methods to predict muscle activity: implications for neuroprosthetics,”
Journal of Neural Engineering, 2009/6, 055008

[2] Au A T C, R. F. Kirsch:
„EMG-based prediction of shoulder and elbow kinematics in able-bodied and spinal cord injured individuals,”
IEEE Trans. on Rehabilitation Engineering, Vol. 8, Issue 4, December 2000.

[3] R. Tibold, J. Laczkó:
„The Effect of Load on Torques in Point-to-Point Arm Movements: A 3D Model,”
Journal of Motor Behavior, Vol 44, No. 5, 2012.

[4] S. M. Radhakrishnan, S. N. Baker and A. Jackson:
„Learning a Novel Myoelectric-Controlled Interface Task,”
Journal of Neurophysiology, 100:2397-2408, 2008.

[5] W. Wu, A. Shaikhouni, J.P. Donoghue and M.J. Black:
„Closed-Loop Neural Control of Cursor Motion using a Kalman Filter,”
Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco, CA, USA, September 1-5, 2004.

[6] G. Kamen, D. A. Gabriel:
„Essentials of Electromyography,”
Human Kinetics, 2010, ISBN(10): 0-7360-6712-4

[7] N.Sárkány:
„The Design of a Biomimetic Joint,”

Proceedings of the Multidisciplinary Doctoral School, 2011-2012 Academic Year, Péter Pázmány Catholic University, Faculty of Information Technology, ISSN 1788-9197

[8] C. Kuo, J. L. Knight, C. A. Dressel, A.W.L. Chiu: *„Non-Invasive BCI for the Decoding of Intended Arm Reaching Movement in Prosthetic Limb Control”,* American Journal for Biomedical Engineering, 2012, 2(4): 155-162

[9] R. Tibold, G. Fazekas and J. Laczko: *Three-Dimensional Model to Predict Muscle Forces and Their Relation to Motor Variances in Reaching Arm Movements,* Journal of Applied Biomechanics, 2011, 27, 362-374

More effective boilerplate removal: the GoldMiner algorithm

István Endrédy
(Supervisor: Dr. Gábor Prószéky)
endredy.istvan.gergely@itk.ppke.hu

Abstract— **The ever-increasing web is an important source from which large-scale corpora can be built efficiently. However, dynamically generated web pages often contain much irrelevant and duplicated text, which, due to overrepresenting repeated content, impairs the quality of the corpus. In this article, we present an automatic text extraction procedure that, enhancing a previously published boilerplate-removal algorithm, minimizes the occurrence of irrelevant duplicated content in corpora downloaded from the web more effectively than previous tools.**

***Keywords*- corpus building; web crawler; boilerplate removal**

I. THE TASK

When constructing corpora from web content, the extraction of relevant text from dynamically generated HTML pages is not a trivial task due to the great amount of irrelevant repeated text that needs to be identified and removed so that it does not compromise the quality of the corpus. This task, called boilerplate removal in the literature, consists of categorizing HTML content as valuable vs. irrelevant, filtering out menus, headers and footers, advertisements, and structure repeated on many pages.

In this paper, we present a boilerplate-removal algorithm that removes irrelevant content from crawled content more effectively than previous tools. The structure of our paper is as follows. First, we present two freely available tools that we used as baselines when evaluating the performance of our system. The second tool, jusText, is also used as part of our boilerplate-removal algorithm. This is followed by presentation of the enhanced system, called GoldMiner, and evaluation of the results.

A. The Body Text Extraction (BTE) algorithm

The basic insights underlying the BTE algorithm [1] are the following:

1. the relevant part of the HTML content is usually a contiguous stretch,

2. the density of HTML tags is lower in it, thain in boilerplate content.

Based on these two assumptions, the algorithm performs a search for the longest stretch of text in which the number of intervening tags is minimal. The idea is simple, but the result is often wrong with the algorithm failing to extract the most relevant part of the content in situations where, contrary to the tag density assumption, it contains a segment with a higher tag-to-text ratio, for example, if tables are included or advertisements interrupt the article. In this case, a significant part of the valuable content (or the whole) may be lost or replaced by entirely irrelevant content.

B. The JusText algorithm

The jusText algorithm [2] splits HTML content into paragraphs at block-level tags that are generally used to partition HTML content into logical units, such as <p>, <td>, <h1> etc. Using various features of these blocks of text such as the number of links, words and stopwords, the algorithm performs a rule-based classification of the blocks using various thresholds and a language-dependent list of function words tagging each unit ‘good’, ‘almost good’, ‘bad’ or ‘too short’. The latter tag applies to units too short to categorize reliably. After initial classification, ‘almost good’ and ‘too short’ units surrounded by ‘good’ ones are reclassified as ‘good’. The text to be extracted consists of all units classified as ‘good’ in the final classification. The algorithm performs quite well even for extreme pages.

However, inspection of the corpus generated by using the jusText algorithm to crawl news portals revealed that many expressions that should obviously be unique to a single piece of text were still very strongly over-represented. Examples in Table 2 are from a corpus crawled from Hungarian news portals applying jusText as a boilerplate-removal tool.

We found that the problem is caused primarily by jusText failing to eliminate leads of related and recommended articles and by index pages containing only article headlines and leads. Leads and headlines of the set of current articles advertised on every article body page during the limited time span of the crawl are thus heavily overrepresented in the corpus. This is illustrated on Table 2.

I. Endrédy, “More effective boilerplate removal: the GoldMiner algorithm,”
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 75-78.

TABLE 1. RESULTS ON 2000 PAGES FROM VARIOUS NEWS PORTALS

Algorithm		Sentences	Unique sentences	Characters	Characters in unique sentences	Rate of unique sentences %	Rate of chars. in unique sentences %
origo.hu	all plain text	264 423	63 594	16 218 753	7 048 011	24%	43%
	BTE	60 682	33 269	12 016 560	7 499 307	54%	62%
	JusText	58 670	30 168	8 425 059	4 901 528	51%	58%
	GoldMiner	22 475	21 242	3 076 288	3 051 376	**94%**	**99%**
nol.hu	all plain text	509 408	144 003	25 358 477	12 570 527	28%	49%
	BTE	154 547	107 573	24 292 755	13 544 130	69%	55%
	JusText	186 727	128 782	14 167 718	11 665 284	68%	82%
	GoldMiner	162 674	123 716	12 326 113	11 078 914	**76%**	**89%**
index.hu	all plain text	232 132	55 466	9 115 415	4 542 925	23%	49%
	BTE	51 713	26 176	5 756 176	4 061 697	50%	70%
	JusText	40 970	29 223	4 371 693	3 441 337	71%	78%
	GoldMiner	13 062	11 887	1 533 957	1 489 131	**91%**	**97%**

TABLE **2**. EXAMPLES OF PHRASES OVERREPRESENTED DUE TO INADEQUATE BOILERPLATE REMOVAL (PATTERN: N (+ N) + ADJ + N)

Phrase	Occurr.
Utasi Árpi-szerű mesemondó. 'Utasi Árpi-like storyteller.'	10,587
A cumisüveg potenciális veszélyforrás. 'The feeding-bottle is a potential source of hazard.'	1.578
Obama amerikai elnök, 'U.S. President Obama,'	292
etióp atléta: cseh jobbhátvéd 'Ethiopian athlete: Czech right-back'	39,328
Barack Obama amerikai elnök 'U.S. President Barack Obama'	2,372
George Bush amerikai elnök 'U.S. President George Bush'	1,626

II. THE GOLDMINER ALGORITHM

The problem can be solved more efficiently if we step up to a level higher than that of individual web pages. After our first attempts at defining a good general procedure for identifying unwanted parts of pages were less successful than expected, we decided to take an optimistic stance and look for what is good instead of what is bad.

We based our approach on the following observations:

- The relevant part of the HTML content is usually a contiguous stretch (see the BTE approach).
- Within a web domain/subdomain, the internal structure (the HTML code) of dynamically generated pages generally contains common patterns that can help us identify relevant content.

The algorithm takes a sample of the pages of the domain/subdomain and tries to locate the common patterns in the HTML code within the sample that identify the beginning and the end of valuable content. For example, news portals typically advertise recent and related articles by displaying their headline and lead next to the actual article. Although this usually seems to be relevant content to jusText, it is in fact just boilerplate content like menus or advertisements that has little or nothing to do with the actual article. Not filtering them out results in thousands of duplicates in the corpus, strongly overrepresenting this content.

The algorithm learns the HTML tags identifying the beginning and the end of the article for each web domain/subdomain, and only content within this stretch of the page is kept. In addition, since it may still be the case that the body of the article is interrupted with advertisements or other boilerplate content at several points, it is submitted for further processing to the jusText boilerplate removal algorithm. An advantage of this solution is that text from pages with no article content (thematic index pages, tag clouds, search page results,

etc.) will not be added to the corpus since the domain-specific HTML tag pattern is not present on them. The algorithm automatically discards the contents of these pages. On the other hand, all pages are, of course, still used as a source of URLs for the crawl.

A. A detailed description of the algorithm

The first phase of the crawl of a domain is taking a sample based on which the domain-specific HTML tag pattern is identified. The algorithm downloads a sample of some 100 pages, applying jusText categorization to each page, which breaks content into paragraphs and evaluates them. Repetitions of individual extracted paragraphs (identified as "good" by jusText) over different pages in the sample are identified by the GoldMiner algorithm, and these paragraphs are reclassified as bad. Unique paragraphs remain classified as "good". Next, it finds the nearest common parent HTML tag of the good paragraphs in the DOM hierarchy on each page. At the end of the learning phase, the most frequent common good parent tag is identified as the winner.

We do not usually get optimal results, however, if the closing tag pair of this parent tag is simply chosen as the tag marking the end of the article. The span enclosed by the parent tag pair may contain bad paragraphs, too. In this case, the algorithm would not find the optimal cut points. Therefore, it performs another search for the optimal starting and endpoint within the content of the previously selected tag, which may be a series of tags. With the selection of the cutting points, the learning phase for the domain is finished. As the URL domain is crawled afterwards, only the content between the domain-specific beginning endpoint tag patterns is passed to the jusText boilerplate removal algorithm. Of course, pages used during the learning phase are also handled this way.

During the learning phase, GoldMiner uses only pages where the length of the extracted paragraphs reaches a threshold. Without using a threshold, it failed to learn the optimal cutting points on some domains where thematic opening pages are more frequent than pages containing articles.

III. RESULTS

Table 1 shows the results of the algorithm compared with that of BTE and jusText on three Hungarian news portals: origo.hu, index.hu, nol.hu. The sample corpora quoted in Table 1 were generated crawling just the first 2,000 pages from the domains above. Using GoldMiner, the ratio of duplicates in the corpus was reduced considerably compared to what other algorithms produced.

The results clearly show that the algorithm effectively reduces unnecessary duplication in crawled corpora. Currently we have no estimate of how the different algorithms perform in terms of the amount/ratio of lost relevant content.

If we investigate the phrases from Table 2 on the corpora made by GoldMiner, we can observe the following results: the overrepresented sentences and phrases are eliminated, and the really frequented phrases are now well represented. Table 3 shows these results.

TABLE **3**. TOP LIST OF PHRASE FREQUENCIES WITH GOLDMINER, PATTERN: NOUN + ADJECTIVE + NOUN

Phrase	Occurr.
Obama amerikai elnök, 'U.S. President Obama, '	47
A cumisüveg potenciális veszélyforrás. 'The feeding-bottle is a potential source of hazard.'	1
'Utasi Árpi-szerű mesemondó. 'Utasi Árpi-like storyteller.'	1

And what had happened with Ethiopian athlete and Czech right-back? Table 4 shows its new results.

TABLE **4**. TOP LIST OF PHRASE FREQUENCIES WITH GOLDMINER, PATTERN: NOUN + NOUN + ADJECTIVE + NOUN

Phrase	Occurr.
Matolcsy György nemzetgazdasági miniszter 'Minister of National Economy György Matolcsy'	694
Barack Obama amerikai elnök 'U.S. President Barack Obama'	664
Sólyom László köztársasági elnök 'President László Sólyom'	367
Angela Merkel német kancellár 'German Chancellor Angela Merkel'	345
...	
etióp atléta: cseh jobbhátvéd 'Ethiopian athlete: Czech right-back'	1

"..." means many rows until we to Ethiopian line

It is very interesting that the patterns shown in table 4 gave a little "who is who" knowledge, a small wikipedia. We observed if a corpus is large enough, its sentences contains valuable world knowledge. We can extract this information with the help of simple phrase frequency. It could be a future project to build corpus driven world knowledge. It would be up to date (if the corpus is fresh), and it can say many things about persons, their relations, without any human intervention.

We were happy to see that unfrequented phrases were eliminated from the list (Table 4).

IV. CONCLUSION

In this paper, we presented a new algorithm that can eliminate boilerplate content from crawled web pages more efficiently than previous algorithms after identifying recurring HTML tag patterns in the dynamically generated web pages coming from a single domain.

ACKNOWLEDGMENT

I would like to say thank you to Attila Novák who gave me lots of good ideas and inspiration, and to Gábor Prószéky for the opportunity to take part in this project.

REFERENCES

[1] Aidan Finn, Nicholas Kushmerick, and Barry Smyth., Fact or fiction: Content classification for digital libraries. In DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries, 2001.

[2] Jan Pomikálek. Removing Boilerplate and Duplicate Content from Web Corpora [online]. Disertační práce, Masarykova univerzita, Fakulta informatiky, 2011.

Region-merging based on contour-structure of clusters in over-segmented image

Anna Horváth
(Supervisor: Dr. Kristóf Karacs)
horvath.anna@itk.ppke.hu

***Abstract*—There are different merging strategies usually based on similarity indexes or if possible semantic information to merge clusters in over-segmented images. This paper introduces a technique based on coherent region contours. Candidate points for the contours are detected and those along the same contour marked with identical ID. Clusters along each contour are merged respectively.**

***Keywords* oversegmentation; region-contour;**

I. INTRODUCTION

In image understanding segmentation is one of the most useful tools. On one hand background-foreground segmentation, or even distinguishing among foreground elements may often be necessary. As a result of any of the different techniques the output image is over-segmented, and requires further processing to enable us detecting and/or recognizing objects.

II. BACKGROUND

A. *Oversegmentation*

The input images are generated with mean shift(MS) algorithm. Means shift is an iterative, non-parametric kernel based method [1]. As a first step a feature space is defined for the image. The MS kernel is convolved with the feature space elements and locates dense regions (mean points) as a result of convergence. The output of means shift is usually an over-segmented image, as illustrated on Figure 1(with random colors) and Figure 2 (with mean colors).

Figure 1.: Example for over-segmented image as an output of mean-shift algorithm.

Figure 2.:Oversegmented image with mean colors

B. *T points*

The contour lines are built up from clusters. These clusters are candidates to be merged, as the contour indicates an image region which was sliced up. Figure 3. illustrates an image part where contour can be well detected for human eyes, to locate part of the wing of the airplane from the sky (background).

Figure 3.: Image part of over-segmented image to illustrate the combined contour of region by clusters.

The points which reside on the contours are usually have a T-shape where the letter's horizontal part represents the part of the cluster border in the joint contour. The vertical part indicates, that the other two clusters tumble into this contour in a perpendicular or similar way.

A. Horváth, "Region-merging based on contour-structure of clusters in over-segmented image,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 79-81.

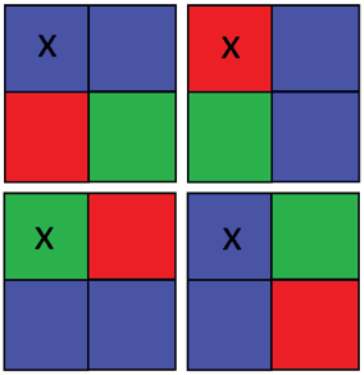


Figure 4.: Types of candidate pixels for contours: possible T points.

III. ALROGITHM

The input image to out algorithm is not only segmented, but also an indexed image based on the segent ID of the pixel.

Our algorithm consist of the following steps:

- Detect points in image where exactly three clusters meet.
    - See illustration on Figure 4. Pixels marked with x are selected as candidates. Currently only the upper-left corner candidates are selected, but this is to be extended to the other three variants. Let us call this candidate pixel window.
- Examine *NxN* size local neighborhood of each candidate found in the previous step (if their neighborhood is within image borders).
    - count frequency of each color in the neighborhood
    - check if the three most "popular" colors are the same as in the 2x2 elements in the candidate pixel window.
    - approximate the incidence of the line represented by the border between each two cluster-parts (falling in the neighborhood).
    - If any of the three approximated lines meets the third with and angle around 180 degrees (with a tolerance) then mark point as contour candidate.
- Index contour candidates with contour indexes, each contour with unique ID:
    - recursively find other candidates along the horizontal line of the T of the contour.
    - if found repeat, if not step to an unindexed contour-point.
- Merge clusters based on chains: current version only links contour points.

Parameters to be set: N as neighborhood size, and in our further work the angle of incidence will also be a parameter with a fuzzy output.

Figure 5.: Result image about the wing after processing.

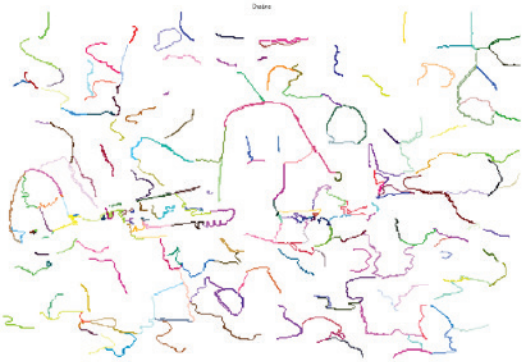


Figure 6.: Result total image processing.

IV. RESULTS AND DISCUSSION

Our algorithm can detect contours and mark them as coherent. Figure 5 and 6 illustrate outputs for the image in Figure 1 and for a part of the image. Still, the algorithm is not robust enough and candidates in the first step are not chosen precise enough. We investigate the upper left pixel for each 2x2 region instead of examining the middle element of a 3x3 region with more complex criteria of being a T-point. This issue is under labor right now.

Each T-point is now defined to belong to a single chain – this concept will be changed to multiple chains to enable the detection of long, coherent contours.

In Figure 7 one can see a typical case when out algorithm is useful. Based on color similarity no further merge is proposed. However, based on the detected lines our algorithm would propose a merge, see Figure 8 for the current output, and

Figure 9 for the expected results after the mentioned corrections. As a next step we would like to implement the vote on merge clusters part and test on large image set to examine robustness.

Figure 7: Input image and merged image based on similarity.

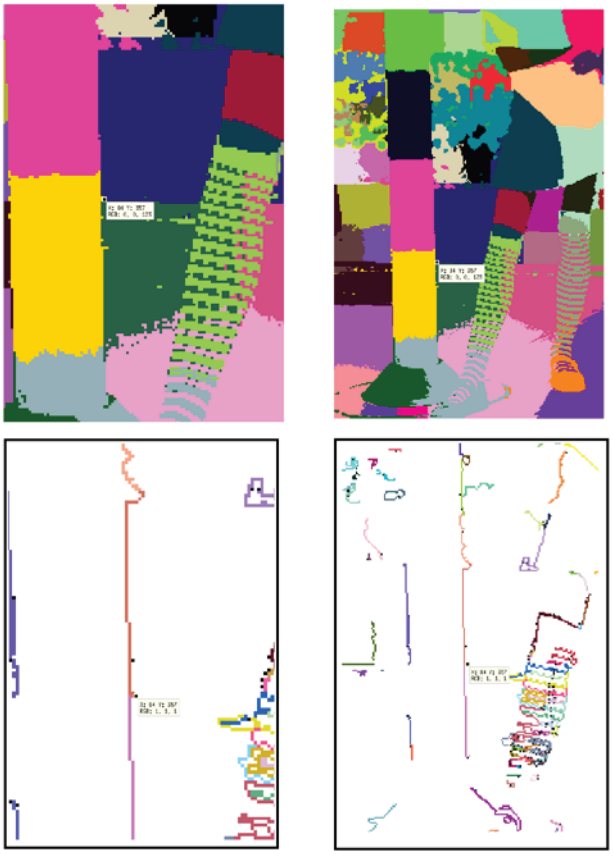


Figure 8: On this Figure we present the output of our current algorithm. The lines give already a fine estimation of the geometry of the pillar, and merge will be proposed along that as shown in Figure 9.

Figure 9: Expected output after corrections. Merge is proposed based on the T points and chains.

REFERENCES

[1] D. Comaniciu and P. Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5):603{619, 2002

High-Resolution, Multi-Channel, FPGA-Based Time-To-Digital Converter

Balázs Gy. Jákli
(Supervisor: György Cserey)
jakli.balazs@itk.ppke.hu

***Abstract*—In this paper we present a novel high-resolution multi-channel FPGA-based time-to-digital converter (TDC). We designed and implemented a complex electronic circuit on the FPGA, whose overall accuracy is several orders of magnitude greater than the accuracy of the FPGA used in digital mode. Our sensor device contains simple circuit elements that are cheap and easily accessible (Xilinx Spartan 3 and Spartan 6). Using our design, many channels (80-100 channels) can be implemented on a larger FPGA. The prototype of our TDC has been implemented and functionally verified by experiments and measurements. By a certified pulse generator 20 ps precision has been measured over the range of 3 ns. Using more precise clock signal this range may be extended. The achieved resolution is 10 ps. Its resolution, channel number and range can be configured dynamically, which makes it suitable for effective use in industrial purposes.**

I. Introduction

Integrated circuits which can measure time-intervals with high precision are called time-to-digital converters (TDC). The majority of time measurement tools on the market primarily use only a few channels (1-8 channels). A channel is characterized by two inputs, it measures the time delay between their signals. TDC has applications in various fields of industrial engineering. It is used for LIDAR [1], medical imaging applications [2] and time-of-flight mass spectrometry as well as it is used in logic analyzers, where extremely high resolution and accuracy are needed. Multi-channel TDC can be specifically apply to determine the 3D spatial localization of events with cm precision in a particle physics detector. It is hard and expensive to design a multi-channel TDC, but there are situations when they can be used effectively: for example, by using them it is possible to determine with cm precision the 3D spatial localization of events happening in a particle physics detector, moreover, they can be applied in multi-channel logical analyzer tools, where extremely high resolution and accuracy are needed.

Until the recent years, most TDC designs have been developed and implemented for specific applications, but the latest trends shows that configurable circuit implementations can be successful using field programmable gate arrays (FPGA)[6-13]. The advantages of the FPGA are lower development time and lower cost in small series.

Many approaches have been tried to use the opportunities of the FPGA to measure time [6-13]. One of the latest FPGA-based TDCs is documented in paper [3] that is close in precision to our solution, but they used the Virtex II Pro FPGA, which is more expensive, and it is not suitable for the realization of multi-channel TDCs. [4] and [5] are presenting multi-channel TDCs (about 96 channels), but their resolution has the order of magnitude of 1 ns.

The majority of TDCs on the market primarily use only a few channels (1-8 channels). A channel is characterized by two input signal; it measures the time delay between their signals. TDC has applications in various fields of industrial engineering. Multi-channel TDC can be specifically applied to determine the 3D spatial localization of events with cm precision in a particle physics detector. Until the recent years, most TDC designs have been developed and implemented for specific applications, but trends show that configurable circuit implementations can be successful using field programmable gate arrays (FPGA) [6] [7] [8]. The advantages of the FPGA are lower development time and lower cost in small series. Some of the latest FPGA-based TDCs documented in paper [7], [8] are close in precision to our solution, but they used the Virtex II Pro FPGA which is a more expensive and complex device. Another issue is that even though high-precision FPGA based implementations are achievable, the measurement range is usually relatively short. Therefore, our aim was to provide this property as well in our design.

II. Motivation

The measurement of time is essential in electrical measurements, practicing engineers use these techniques in many cases. An analog-digital conversion is performed by a TDC. The analog input value is a relative time difference of two events between each other, such as the delay between two signals. The output is a value in a discrete range, which is the approximated result of the converted input value by quantization. The unit of the conversion, the resolution of the TDC is the shortest time which makes changes, either increasing of decreasing of the output value. Currently, this magnitude is in pico-second range.

The number of TDSs have reached a marketable level by the very large scale integration of circuits, but the vast majority of commercially available devices have only a few channels (channels 1-8). A channel is characterized by two inputs, the relative temporal delay is measured between the inputs. These sensors are implemented on ASICs (Application Specific Integrated Circuits). Multi-channel TDCs are difficult and expensive, but there are tasks where they can be used effectively: for example, using TDCs in particle detectors, 3D spatial localization of events can be determined with precision

B. Gy. Jákli, "High-resolution, multi-channel, FPGA-based time-to-digital converter,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 83-86.

in the order of cm, and also their usage can be necessary in other particle physics experiments. Nowadays these particle detectors are very expensive and have only a relatively small accuracy. A cost-effective, high precision, multi-channel TDC would be suitable to fulfill a currently empty market segment. These TDCs would be used to complete measurements, which could not have been done by physicists previously, as well as in multi-channel logic analyzer devices, where high resolution and high accuracy is required.

III. Proposed Architecture

We designed and implemented an experimental prototype of our device. Besides the Xilinx Spartan 3 FPGA chip, we placed a dsPic to do the control and communication to the FPGA chip. This solution gives the opportunity to re-configure the channel number and the size of the actual measurement range of our system. The next version of the device used two Spartan 6 FPGAs.

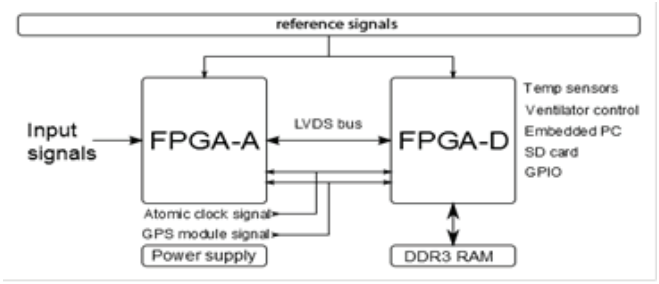


Fig. 1. System Overview: FPGA-A does the time-to-digital conversion. FPGA-D is the control FPGA responsible for the temperature-stability, control processes and measurements of the reference signals and their parameters.

The digital clock signal interferes with the measurement, so the most of the digital electronics are synthesised in a separate FPGA-D ('D' stands for 'digital'). The signal measurement has a dedicated FPGA (FPGA-A), which is used in an analog method, without clock signals. Clocking is applied only when the measured values are read. The digital FPGA controls the system functions, including the temperature stabilization and data logging to DDR3 RAM and flash memory. The system has an embedded computer for user interface, FPGA reprogramming and communication.

The system has a modular architecture. The different functions are implemented on different PCB boards. The processor board has two Xilinx Spartan 6 LX25 devices, DDR3 RAM memory, 8 SATA headers for impedance controlled signal input, and 10 temperature sensors for the temperature stabilization. The board has only 8 layers to decrease costs. After validation, the PCB needs another 4 ground planes to improve noise immunity.

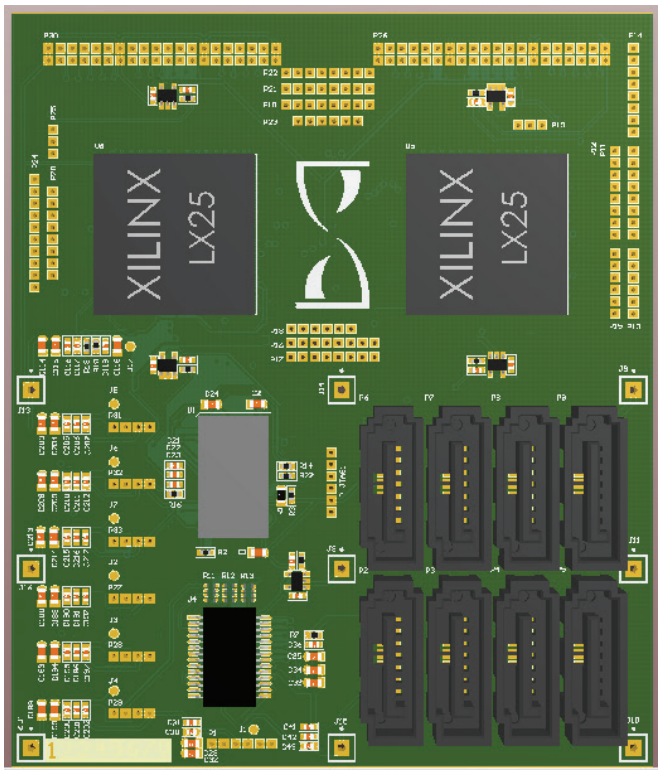


Fig. 2. FPGA measurement board

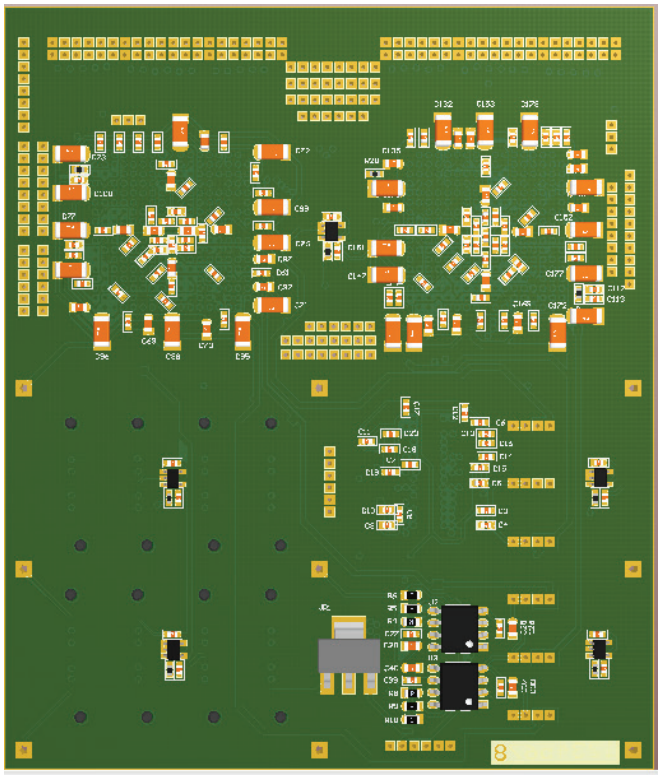


Fig. 3. FPGA measurement board bottom side

The system needs accurate reference clock signal for absolute time measurement. For the required accuracy, this needs

to be an atomic clock signal. Becouse atomic clocks are expensive, this low cost TDC instrument has two fallback subsystems. One is a GPS module, with a 1pps (pulse-per-second) signal output. This is very accurate, becouse GPS system is based on atomic clocks. The other subsystem is a matrix of oscillators. In the current device, we have a 8 * 8 oscillators, each with a 125 MHz output signal. The digital FPGA measures these reference signals, and merges them using stochastic methods. This system genenerates a 8 nanosecond timebase for absolute measurements. Becouse the oscillators has a significant jitter and their output slowly changes while the oscillator ages, this oscillator matrix has to be recalibrated sometimes using the GPS signal.

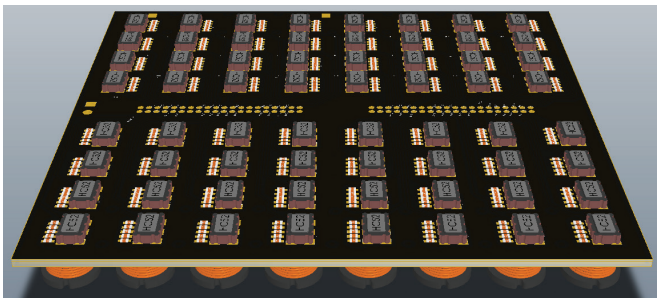


Fig. 4. Oscillator matrix

To achieve correct signal interference ratios, the PCB is build up from 4 layers, with one full ground and one full power plane.

IV. Experimental Results

We measured the pulse delay between two signals generated by a two channel function generator (Type: Agilent 81130A) A personal computer triggered the signals, then read out the output values of the TDC conversion via USB interface (see Fig. 6).

On the upper side of Fig. 7 we see the output of the TDC conversion depending on the delay set on the function generator. At the bottom side of Fig. 7 , we see the error between the theoretical and the measured delays. As we can see, the accuracy of the measurement is about 20 pico-secundums, which is the accuracy of the function generator. It is assumed that our method is capable of even more accurate results.

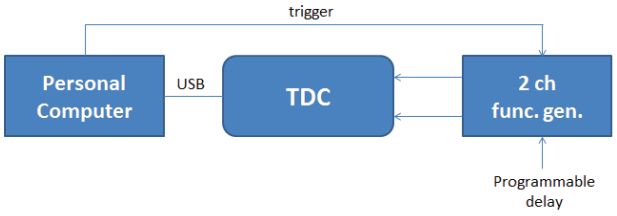


Fig. 5. We measured the pulse delay between two signals generated by a two channel function generator (Type: Agilent 81130A) A personal computer triggered the signals, then read out the output values of the TDC conversion via USB interface.

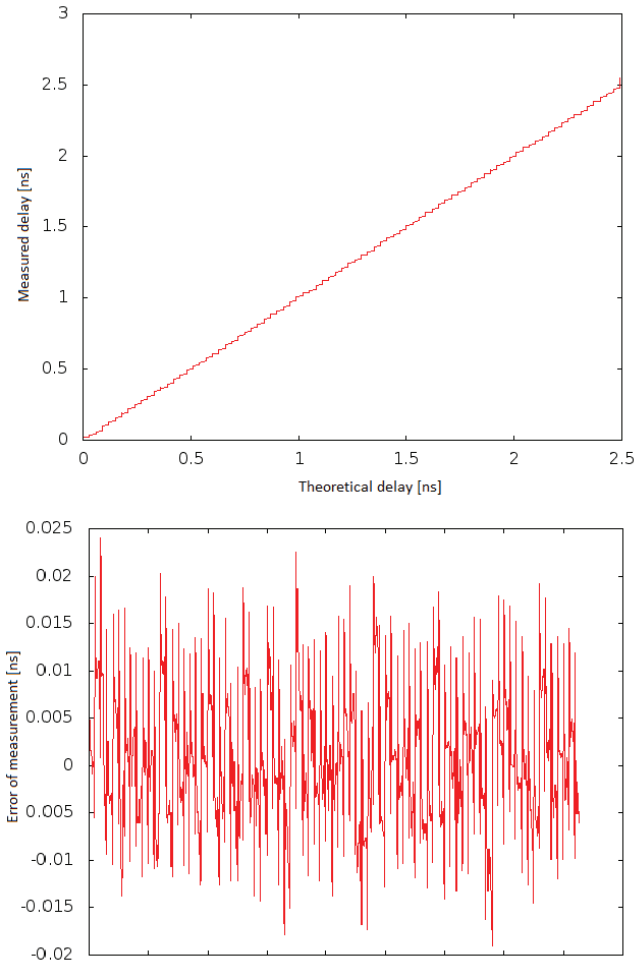


Fig. 6. On the upper side we see the output of the TDC conversion depending on the delay set on the function generator. At the bottom side, we see the error between the theoretical and the measured delays. As we can see, the accuracy of the measurement is about 20 pico-secundums, which is the accuracy of the function generator. It is assumed that our method is capable of even more accurate results.

The main components of the system can be seen in Fig.1. The experimental prototype of our device has been implemented (see Fig. 2.) and its functionality is confirmed by experiments and measurements (see Fig.3). We have captured the raw stochastic bitstream output of our TDC. In our first results we did not use any calibration on the bitstream, instead we have simply summed the bits as integers. This summation decreased the stochastic noise enough to allow us to see the working time measurements as seen in Fig.3. We have also tested the shifting of our measurements range and it worked exactly as precise as the clock source we have used as a reference source for shifting by whole clock cycles therefore extending the range to practically infinity (the precision is limited only by the frequency stability and the jitter of the clock source). Our prototype was applied on the Xilinx Spartan 6 FPGAs. The parameters of our innovative sensor can be configured flexibly.

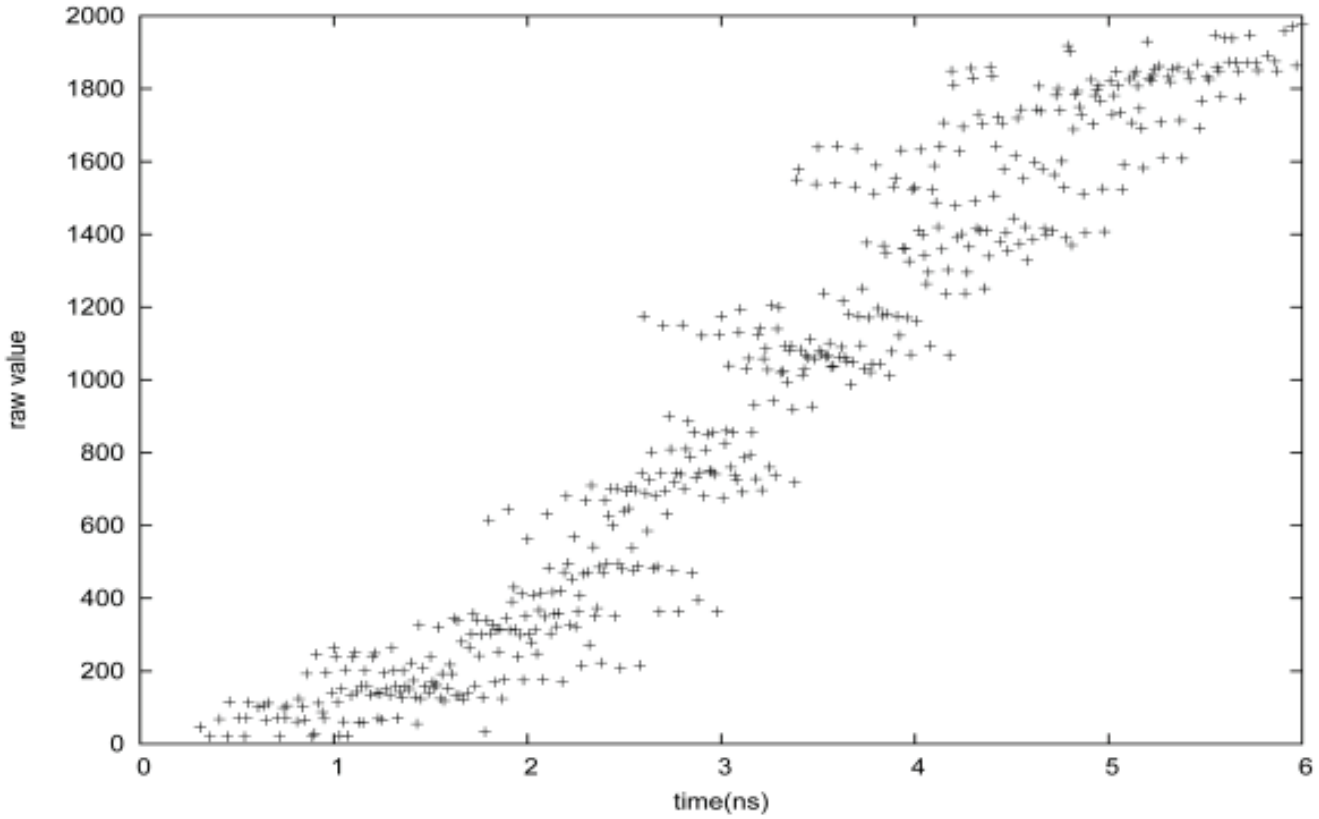


Fig. 7. Measured raw uncalibrated data from our TDC is depicted on the figure versus the input time difference of the impulses.

V. Conclusions and Future Work

The experimental prototype of our device has been implemented and its functionality is confirmed by experiments and measurements (see Fig. 6). We measured 20 ps accuracy rate by a certified pulse generator (20 ps, where the range is 3 ns, see Fig. 7). The range of the measurement may be extended by using a more accurate clock signal. Our prototype was applied on the smallest Xilinx Spartan 3 FPGA. Using a greater FPGA, up to 100 high-precision channels can be implemented. The parameters of our innovative sensor can be configured flexible. Using active cooling (Peltier-element) the range of the sensor system can be extended, while operation range becomes more stable. One of the fundamental innovations in our approach is that rather than only compare the two input signals in a channel, we have included the many calibrated reference signals as well in our comparison, and thus, we have been able to achieve high precision and low-noise in our measurements. After comparing the calibrated signals to the input signals, the resulting absolute values render a measurement range that is theoretically infinite and practically very large. In the event that we have multi-inputs, they are simultaneously compared to the reference signals. Another innovation is the design and measurement method of the internal structure of the FPGA, which provides a stochastic bit-stream based time-to-digital converter (TDC). We can reconfigure our system dynamically by changing the channel number or measurement accuracy of each channel even during operation. The expected achievable accuracy of the system is 1ps, and based on our measurements it is certainly less than 10ps.

References

[1] I. Nissinen and J. Kostamovaara, "On-chip voltage reference-based time-to-digital converter for pulsed time-of-flight laser radar measurements," *Instrumentation and Measurement, IEEE Transactions on*, vol. 58, no. 6, pp. 1938–1948, 2009.

[2] A. Yousif and J. Haslett, "A fine resolution tdc architecture for next generation pet imaging," *Nuclear Science, IEEE Transactions on*, vol. 54, no. 5, pp. 1574–1582, 2007.

[3] M. Daigneault and J. David, "A high-resolution time-to-digital converter on fpga using dynamic reconfiguration," *Instrumentation and Measurement, IEEE Transactions on*, vol. 60, no. 6, pp. 2070–2079, 2011.

[4] J. Wu, S. Hansen, and Z. Shi, "Adc and tdc implemented using fpga," in *Nuclear Science Symposium Conference Record, 2007. NSS'07. IEEE*, vol. 1. IEEE, 2007, pp. 281–286.

[5] V. Bocci, R. Nobrega *et al.*, "A multichannel tdc based on ring-delay time multiplexer: A prototype," in *Nuclear Science Symposium Conference Record, 2007. NSS'07. IEEE*, vol. 1. IEEE, 2007, pp. 720–724.

[6] M.-A. Daigneault and J. David, "A high-resolution time-to-digital converter on fpga using dynamic reconfiguration," *Instrumentation and Measurement, IEEE Transactions on*, vol. 60, no. 6, pp. 2070 –2079, june 2011.

[7] J. Wang, S. Liu, L. Zhao, X. Hu, and Q. An, "The 10-ps multitime measurements averaging tdc implemented in an fpga," *Nuclear Science, IEEE Transactions on*, vol. 58, no. 4, pp. 2011 –2018, aug. 2011.

[8] S. Qi, et al, "A fast improved fat tree encoder for wave union tdc in an fpga," in *arXiv preprint, arXiv:1303.6849*, 2013.

Attentional Modulation of Visual Cortical Responses to Sequential Stimuli – a Single-Trial Approach

Balázs Knakker
(Supervisor: Dr. Zoltán Vidnyánszky)
knakker.balazs@digitus.itk.ppke.hu

***Abstract*—Visual cortical processing for a stimulus that is part of a temporal sequence can be different from responses to stimuli viewed in isolation. Also, the influence on attention on responses for sequential stimuli is not well-known. To elucidate this, we presented sequences of simple and compound word and face stimuli, and cued participants to attend to either of the two categories, while EEG was measured. Analysis of average event related potentials and single-trial evoked components revealed robust attentional effects amidst stimulus sequences that were missing for the first stimulus. Sequential face stimuli induced gradual response habituation, which did not counteract but strengthened attentional effects. From single trial peak distributions we inferred that the jitter in the timing of responses might only have a secondary role in generating these patterns, as amplitude changes were observable for single trial responses as well. Within-subject analyses found that attentional effects were present individually in the majority of the subjects.**

***Keywords*-habituation, attention, word, face, event-related potentials, single-trial**

I. Introduction

In neuroscientific experiments on vision, subjects are usually seated in a dark room, asked not to move and to perform totally artificial tasks designed to disentangle delicate cortical processing steps based on electrophysiological signals and behavioral measures. In contrast, real-world vision is a continuous, active, purposeful and adaptive sampling of the environment. For example, during reading the visual system scans the text (more or less) word-by-word by means of eye movements. Even if we disregard eye movements, the situation is still far from most experimental settings, because 'stimuli' – bits of the text at distinct fixations – come at a relatively fast pace. So, instead of separate events, they are processed more as a continuous stream of visual information, where after each stimulus, the visual system can adjust its state to facilitate the perception and integration of the forthcoming ones.

To disentangle how cognitive-perceptual processes characterized under strict experimental control work under more natural circumstances, a research program needs to be pursued in which the variables conventionally excluded and controlled are re-introduced as variables of interest into the experimental paradigms. In the present study, we aimed at characterizing the effects of attention on the visual cortical processing of uninterrupted sequences of stimuli.

The other, just as important motivation of this study is of methodological nature. The Event Related Potentials (ERP) technique is a widely used classical method in neuroscientific research on vision. In most cases the method involves reducing the data to a set of subject-level averages, on which peaks are defined and identified. The latencies and amplitudes of these peaks are the subject of analysis. But this way, we lose information contained in intertrial variability in the EEG, and we also cannot tell how robust the effects are in each individual subject. To tackle this, several single-trial analysis approaches have been developed recently[1], [2], [3]. It is debated whether averaged ERPs arise from additive evoked components – basically the peaks being present in every single segment – or from simply the reorganization of ongoing EEG[4], but it seems reasonable to assume that this simple additive model of ERP generation holds true at least in part[5]. Thus, perhaps the most straightforward solution is to look for the single-trial peaks corresponding to those we know from averaged ERP responses[6] – this is the method that was tested and evaluated during this research. In this framework, the shape, latency and amplitude of averaged ERP components are expected to arise from the amplitude and latency distributions of the corresponding single-trial peaks. It is well known, for example, that a smaller average ERP peak amplitude can be caused by both an actual amplitude decrements in every single trial and/or the increased jitter of single-trial peak latencies[6]. So, compared to using amplitudes from a given latency in each trial (or using the whole segment point-by-point), we have the advantage that peak latency jitter, a physiologically relevant variable, is also taken into account.

The N1 componentis considered a prominent electrophysiological correlate of the higher-level structural encoding cascadethat operates in the inferior temporal cortex[7], and is known to be affected by both adaptation/response habituation[8] and attention[9], [10]. Therefore, it is an ideal candidate to investigate attention during serial stimulation.

To sum up, this study aimed at investigating how attention works for temporal sequences of complex visual objects, more precisely, how the N1 evoked component is modulated on the single trial level by attention during serial stimulation.

This work was supported by grant from the Hungarian Scientific Research Fund to Zoltán Vidnyánszky (CNK80369)

B. Knakker, "Attentional modulation of visual cortical responses to sequential stimuli - a Single-Trial approach,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 87-91.

II. Materials and Methods

A. Subjects

20 healthy young adults participated in this study. All of them had normal or corrected-to-normal vision; none of them had any history of neurological or psychiatric diseases. All participants gave their informed consent prior to starting the experiments. Four of the subjects were discarded because of inadequate task performance or excessive noise, so the data from 16 subjects was analysed.

B. Stimuli and Procedure

Subjects were seated in a dark room, their head supported by a chin rest in a distance of 50 centimetres from the screen, which had a resolution of 1600 x 1200 pixels. A blue, 0.1°fixation disc was always present in the middle of the screen. The background was mid-grey.

The stimulus material consisted of male and female face images with equalized luminance and contrast and printed word images from sub-categories 'fruit' and 'animal'. In each of the 480 trials, six stimuli (S1-S6) were presented with an SOA of 683 ms, but without ISI. That is, the offset of stimulus 1was the onset of stimulus 2 at the same time, 683 milliseconds after the onset of the first stimulus.

In half of the trials, the stimulus was *simple:* either a face or a word was presented centrally.In the other half, it was *compound,* that is, a face was presented centrally with a word overlaid on it slightly above fixation. A second before each trial, a cue was displayed which instructed the subject to perform the task with (and thus attend to) either faces (*attend-face* conditions) or words (*attend-word* conditions), regardless of the presence or absence of the other stimulus type. These add up to a total of four experimental conditions: the simple conditions *word* and *face,* and the two visually equivalent compound conditions *word-face* and *face-word.* In one third of the subjects, the stimulus sub-category (male vs. female faces, animal vs. fruit words) was alternating during stimulation. In the remaining two third of the trials, one or two one-back repetitions of stimulus sub-category occurred. The task of the participants was to count these one-back events and indicate them with a three-button mouse after each trial. For example, a male-female-male-female-male-female sequence would count as no (zero) one-back repetition, a fruit-animal-fruit-fruit-animal-fruit would count as one repetition, and so on. This task was designed to sustain the attentional state of the subject throughout the whole trial as much as possible.

Stimulus presentation and subject response registration was implemented in MATLAB using PsychToolbox version 3[11], [12].

C. Electrophysiological and Behavioral Measurements

EEG was acquired using 64 electrodes (Brain Products ActiCap; amplifier: BrainAmp MR) mounted on an elastic cap according to the extended 10/20 system. The sampling rate was 500 Hz and the signal was digitized using an external D/A converter supplied by Brain Products and recorded by the Brain Vision Recorder software. Eye movements were recorded using IView X Hi-Speed (SensoMotoric Instruments) at a sampling rate of 240 Hz.

D. Analysis

Basic preprocessing of the EEG signal was done in Brain Vision Analyzer. The signal was bandpass filtered (Butterworth zero-phase filter, 0.1Hz-30Hz, 12 dB/octave) and segmented. Segments containing artefacts were marked using amplitude, amplitude difference and voltage step thresholds and by visual inspection; these segments were not used in further analyses.

Data were imported to MATLAB, and surface Laplacian approximations of the scalp current density was calculated using the CSD Toolbox[13] (spline flexibility m=4, smoothing $\lambda=10^{-5}$, maximal degree of Lagrange polynomials = 10). Data were segmented so that each segment contained one stimulus onset; baseline correction was done on a [-100 0] interval. Subsequent analyses steps were conducted on two electrode clusters, OTL (Occipito-temporal Left; consisting of PO7, PO9 and P7) and OTR (Occipito-temporal Right: PO8, P8, PO10); within these clusters, signals were averaged. In the case of stimulus 1, for each subject and condition clean segments were averaged and peaks were manually picked on both sides. Stimulus 2 to 6 were pooled for better signal-to-noise ratio, peak latencies were found in each subject and condition by visual inspection. The corresponding amplitudes were then gathered for each stimulus based on this pooled latency estimate.

Single-trial peaks were detected as follows: single trial P1 was defined as the largest positive sample in the window [-50 50] ms around the manually picked peak latency for the current condition, subject, and hemisphere. The windows in which N1 was defined as the most negative sample were defined in each segment individually, starting at the P1 detected previously in the same segment and ending at +120 ms from it. Some putative N1s were found at the terminal edge of this window; these segments (~5-10%) were not used in subsequent analyses. From the acquired distributions, average amplitudes and the interquartile ranges of latency were calculated for each subject, condition, stimulus and hemisphere.Normality was checked by means of the Kolmogorov-Smirnov test where applicable; latency distributions were non-normal, hence the interquartile range was used instead of the standard deviation as a scale parameter.

For the N1 component ERP amplitudes, mean single trial amplitudes and interquartile range parameters of single trial latency distributions were entered into a repeated-measures analysis of variance (ANOVA) using within-subject factors CATEGORY (2 levels: *word-attended* and *face-attended*), DISTRACTOR (2 levels: distractor absent for *simple*, present for *compound* stimuli), STIMULUS (5 levels: *S2* to *S6*) and HEMISPHERE (2 levels: *OTL* and *OTR*). The Huynh-Feldt correction[14] for violation of sphericity was applied where necessary (indicated by H-F).Planned comparisons were conducted as one-sample t-tests with linear contrast weights [0 -2 -1 0 1 2] along the factor STIMULUS (to be called

'habituation contrast' from now on) in every condition and over both hemispheres separately, to assess the hypothesis that responses show gradual linear changes.

Correlations between single-trial summary statistics (mean amplitude and median latency) and conventional average ERP peak-based parameters were assessed by means of Spearman correlation coefficient and test.

Subjectwise comparisons of single-trial distributions were conducted with two-sample t-tests in the case of comparing conditions. For comparing stimuli in the sequence, one-sample t-tests with a linear contrast akin to the ANOVA analysis detailed above. In the latter analysis, only those trials could be used in which all stimuli corresponding to non-zero contrast weights were present (i.e. not discarded because having found the putative peak at the edge of the window), which might have substantially reduced the power of the test.

The FDR procedure of Benjamini and Yekutieli[15] was used for correction for multiple comparisons where applicable.

III. Results

A. Conventional ERP analysis

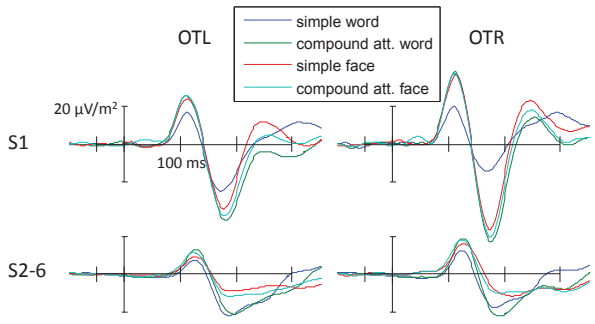


Figure 1. Grand average ERPs of 16 subjects on the two electrode clusters. Top row shows averages for stimulus 1., bottom row for stimulis 2 to 6, averaged.

Repeated measures ANOVA involving stimuli 2 to 6 revealed significant main effects of CATEGORY ($F(1, 15)=22.562$, $p=0.00026$) and STIMULUS ($F(4, 60)=8.1083$, $p_{H\text{-}F}=0.00036$).The differences underlying these are that responses were larger in word-attended conditions than face-attended conditions; and that across S2 to S6 responses show a decreasing trend that is further analyzed below. The interaction of these two factors was also significant (CATEGORY x STIMULUS, $F(4, 60)=2.8068$, $p_{H\text{-}F}=0.043$). No other effects, including the main effect and interactions of the factor DISTRACTOR, were significant ($p>0.2$).

Based on the interaction, habituation contrasts on the STIMULUS factor were tested separately for levels *word-attended* and *face-attended* of factor CATEGORY. The contrast indicated significant linear trend in the *face-attended* case ($t(15)=3.73$, $p=0.0019$), but only as a marginal trend in the *word-attended* case ($t(15)=2.03$, $p=0.06$). Planned contrast analyses conducted for each condition over each hemisphere

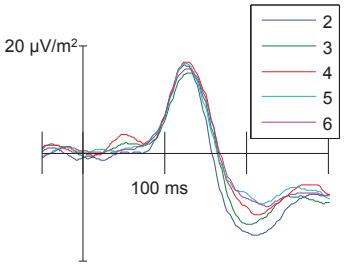


Figure 2. Grand average ERP waveforms of responses to simple face stimuli number 2 to 6 on the cluster OTR. Gradual response habituation is seen in the N1 component.

were consistent with this ($p\approx0.1$ for all *face-attended,* $p=0.055$ for the *simple word* condition and $p>0.1$ for all remaining *word-attended* contrasts tested).

B. Group-level analysis of summary statistics of single-trial distributions

To assess how single-trial and conventional ERP data are related to each other, means of amplitude distributions and medians for latency distributions were calculated and subjected to correlation analysis with their respective ERP parameters within each ANOVA cell. For mean amplitudes, all correlations were significant (FDR correction was applied) with a minimum Spearman's rho of 0.6, the median of p values was 0.00018. For median latencies, 96% of the correlations tested were significant after FDR correction, with a median p of 10^{-5}.

For each ANOVA cell, mean single trial amplitudes and latency interquartile ranges were computed and analyzed with the same procedure as the ERP results to facilitate their comparison.

For single-trial amplitudes, the main effect of CATEGORY($F(1, 15)=23,892$, $p=0.00020$) and STIMULUS ($F(4, 60)=3,3297$, $p_{H\text{-}F}=0.043$) were significant, just like in the case of ERP amplitudes. The pattern was slightly different, stimulus 3 tending evoke a larger N1 than stimulus 2. The CATEGORY x STIMULUS interaction was only marginal ($F(4, 60)=2.5435$, $p_{uncorrected}=0.04870$, $p_{H\text{-}F}=0.054$); post-hoc comparisons with Tukey's method confirmed that for *word-attended* conditions the S3 response was indeed larger (more negative) than the S2 response ($p=0.043$). Most probably due to this trend, the habituation contrast was only significant marginally for the *simple face* condition on the right side ($p=0.046$). To confirm the presence of habituation trends after S3, a modified, exploratory habituation contrast analysis was performed in a post-hoc manner, which yielded results consistent with the ERP results.

Interquartile range statistics of latency distributions showed partly consistent trends: latency jitter of the cortical responses was smaller for *word-attended* stimuli than for *face-attended* stimuli (main effect of CATEGORY, $F(1, 15)=8.9060$, $p=0.00926$), and jitter showed an *increasing* trend throughout stimulation (STIMULUS, $F(4, 60)=3.2440$, $p_{H\text{-}F}=0.018$). Despite the visible trend, none of the planned habituation

contrasts reached significance (*simple face* OTR: p=0.088, OTL: p=0.067, *simple word* OTR: p=0.067, all others p>0.2).

C. Within-subject comparisons of single-trial distributions

		Number of subjects with sign. differences				
		S2	S3	S4	S5	S6
simple w>f	OTL	8	7	5	7	9
	OTR	4	4	6	6	6
	either	8	9	8	8	9
comp. attw>attf	OTL	5	7	6	7	9
	OTR	4	6	8	7	10
	either	6	8	9	9	10

Table 1. Number of subjects with significant effects on a given comparison; FDR correction was applied. The subject is counted in the 'either' rows if in at least one of the clusters the effect is significant. The top three rows indicate in how many subjects simple-word responses were greater than simple-face responses; the bottom three rows show the number of subjects in whom compound stimuli evoked greater responses when the word component was attended.

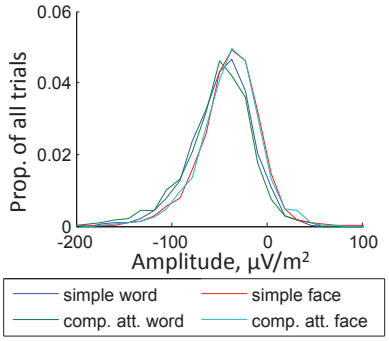


Figure 3. Single trial N1 amplitude distributions pooled across all subjects; Occipitotemporal right cluster. Note the difference between attend-left and attend-right conditions.

Single trial distributions were compared within each subject. The category effect was evaluated by comparing simple face to simple word responses, and then compound word-attended responses to compound face-attended responses for each hemisphereand stimulus. As a simple quantification of the prevalence and strength of the effect in the subject group, we counted in how many subjects each comparison appear significant, which is summarized in Table 1. For the second stimulus, only 4-5 of the 16 subjects show a difference between word-attended and face-attended compound conditions, but the number of subjects with significant effect increases up to 9-10, i.e. 60% of the whole sample. No significant effects were found for the stimulus factor on the between-trial level.

IV. Discussion

This study was aimed at investigating response habituation and neural correlates of attention in a sequential stimulus paradigm.The N1 component, as the earliest electrophysiological correlate of higher-level visual processing, is the focus of this report.First, the results of the conventional ERP analysis, then, after some methodological considerations, the single-trial results will be discussed.

The response to the first stimulus is dominated by the face stimulus: both P1 and N1 are large in the *face, face-word* and *word-face* conditions, and *word* responses are smaller. For subsequent stimuli, there is a general decrease in response amplitudes, but more delicate effects arise, which are described below. This general amplitude difference most probably originates from the fact that the background-to-stimulus transition involves a huge change of contrast, but for the between-stimulus transition gross contrast change is negligible. That is, this amplitude drop between the first and the second stimulus is due to simple low-level contrast adaptation.The forthcoming discussion is about between-stimulus transition responses for stimuli 2. to 6.In this contrast-adapted state, the conventional ERP analysis shows an effect of category, which is not surprising in the *word* and *face* conditions given the large differences in physical stimulus parameters. But this difference is also present in the conditions where the other type of stimulus is also present as a distractor:a word overlaid on a face evokes a *greater* N1 if the word is attended than in the case when the face component is attended. Note that in the two conditions, the stimulus is physically the same, so all differences can be attributed to endogenous factors: that is, attention assures the processing of the attended stimulus regardless of what other stimuli are present. The direction of the effect – *words* evoking greater N1 responses than *faces* – might be unexpected, as in the literature the face N170 is considered largest and most robust ERP component.

As the stimulus sequence proceeded, the amplitude of the N1 component gradually diminished further. This small but steady gradual response habituation effect was only significant for the face and the face-attended compound stimuli; for word-stimuli there was only a decreasing trend. This is in agreement with results from previous studies aimed at characterizing response habituation for word and face stimuli [16], [17]. The strength of response habituation was characterized by the attended stimulus, i.e. it was under attentional control. The main effect of category and the fact that solely face-responses show gradual decrement together also imply that the effect of attentionis not diminished but increasing as stimulation proceeds.

P1 peaks were detected in each segment in windows based on latencies determined in the peak-based average ERP analysis, and N1 peaks were sought for in the 120ms window beginning at the P1 of each trial. Comparing mean single-trial amplitudes and median latencies to respective ERP data reassured that these distributions plausibly underlie conventional ERP peaks.

To facilitate comparison with results of conventional ERP analysis, summary statistics were generated from the single-trial latency and amplitude distributions, on which the same analysis of variance procedure could be performed. These analyses revealed that the ERP amplitude effects are not reflected only in either of the two but in both of them: for *word* stimuli, single trial response amplitudes are *greater* on average than for*face*stimuli, and also, latency variability is *smaller* for *word*s than *face*s.

We found significantsingle-trial effects partly compatible with ERP habituation. Note that the gradual habituation effects we describe here are a composite of very small (i.e. generally not significant in pairwise comparisons), but consistent trends from one stimulus to the next. Most of these trends were compatible across methods, but the trend between S2 and S3 was different in this respect. Instead of exhibiting the decrement (or stagnation) that was expected based on the ERP results, single trial amplitudes were increasing (i.e., more negative for S3 than S2). From S2 on, the pattern was consistent with the ERP habituation effect, as confirmed by contrast analyses conducted in a post-hoc manner. According to the statistics, latency jitter differed in the stimulation sequence, but no consistent linear trend was found, however, responses in the face-only condition exhibited a non-significant ERP-compatible trend.

These results suggest that ERP amplitude differences that we measured are mainly determined by actual amplitude differences on the single-trial level, and latency jitter may only play a secondary role. This is not necessarily true in every case, for example, Bankó et al.[6] describe an ERP amplitude effect which arises solely from increased jitter of the timing of the responses.

One main advantage of single-trial methods is that the effects can be characterized within each individual subject[18], [19]. For the comparison between stimuli, (i.e. habituation) no significant effects were found; possibly due to the reasons described in the methods section. Within each subject, the amplitude distributions of *word* vs. *face (simple stimulus)* and *word-face* vs. *face-word (compound stimulus)* responses were compared against each other, respectively.The results for S2 to S6 are in line with the previous analyses in terms of progressive response habituation effectively *strengthening* attention effects: as stimulation proceeded, more and more subjects displayed significant differences. Also, the number of subjects with significant effects in the compound stimulus case was comparable to (or even larger than) that of the simple stimulus case, which is indicative of effective attentional filtering of distracting stimuli.

V. Conclusion

The aim of the present research was to characterize visual cortical responses to sequential high-level stimuli, and also the effect of attention on these responses. Stimulus sequences consisted of six simple or compound word and face stimuli; in the compound case the stimuli were physically equivalent, thus all differences observed can be attributed to attentional cueing. After the first stimulus, an attentional effect appeared which became stronger until the sixth, last stimuli. Together with this, in conditions in which faces were attended gradual response habituation was observed. According to a simple single-trial approach, these differences are mainly due to actual amplitude changes in the signal, but the increased jitter of the timing of the responsesmay also play a role in them. In terms of single trial distributions, 10 of the 16 participants showed attentional effects individually.

References

[1] A. Delorme and S. Makeig, “EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis,” *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[2] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, “FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data,” *Comput. Intell. Neurosci.*, vol. 2011, p. 156869, 2011.

[3] C. R. Pernet, N. Chauveau, C. Gaspar, and G. A. Rousselet, “LIMO EEG: a toolbox for hierarchical LInear MOdeling of ElectroEncephaloGraphic data,” *Comput. Intell. Neurosci.*, vol. 2011, p. 831409, 2011.

[4] W. Klimesch, P. Sauseng, S. Hanslmayr, W. Gruber, and R. Freunberger, “Event-related phase reorganization may explain evoked neural dynamics,” *Neurosci. Biobehav. Rev.*, vol. 31, no. 7, pp. 1003–1016, 2007.

[5] G. Turi, S. Gotthardt, W. Singer, T. A. Vuong, M. Munk, and M. Wibral, “Quantifying additive evoked contributions to the event-related potential,” *Neuroimage*, vol. 59, no. 3, pp. 2607–2624, Feb. 2012.

[6] E. M. Bankó, J. Körtvélyes, J. Németh, B. Weiss, and Z. Vidnyánszky, “Amblyopic deficits in the timing and strength of visual cortical responses to faces,” *Cortex J. Devoted Study Nerv. Syst. Behav.*, Apr. 2012.

[7] N. Kanwisher, J. McDermott, and M. M. Chun, “The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception,” *J. Neurosci.*, vol. 17, no. 11, pp. 4302 –4311, Jun. 1997.

[8] G. Kovács, M. Zimmer, É. Bankó, I. Harza, A. Antal, and Z. Vidnyánszky, “Electrophysiological Correlates of Visual Adaptation to Faces and Body Parts in Humans,” *Cereb. Cortex*, vol. 16, no. 5, pp. 742 –753, May 2006.

[9] W. Feng, A. Martinez, M. Pitts, Y.-J. Luo, and S. A. Hillyard, “Spatial attention modulates early face processing,” *Neuropsychologia*, vol. 50, no. 14, pp. 3461–3468, Dec. 2012.

[10] C. Aranda, E. Madrid, P. Tudela, and M. Ruz, “Category expectations: a differential modulation of the N170 potential for faces and words,” *Neuropsychologia*, vol. 48, no. 14, pp. 4038–4045, Dec. 2010.

[11] D. H. Brainard, “The Psychophysics Toolbox,” *Spat. Vis.*, vol. 10, no. 4, pp. 433–436, 1997.

[12] D. G. Pelli, “The VideoToolbox software for visual psychophysics: transforming numbers into movies,” *Spat. Vis.*, vol. 10, no. 4, pp. 437–442, 1997.

[13] J. Kayser and C. E. Tenke, “Principal components analysis of Laplacian waveforms as a generic method for identifying ERP generator patterns: I. Evaluation with auditory oddball tasks,” *Clin. Neurophysiol.*, vol. 117, no. 2, pp. 348–368, Feb. 2006.

[14] H. Huynh and L. S. Feldt, “Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs,” *J. Educ. Behav. Stat.*, vol. 1, no. 1, pp. 69–82, Mar. 1976.

[15] Y. Benjamini and D. Yekutieli, “The control of the false discovery rate in multiple testing under dependency,” *Ann. Stat.*, vol. 29, no. 4, pp. 1165–1188, Aug. 2001.

[16] E. Mercure, K. Cohen Kadosh, and M. H. Johnson, “The n170 shows differential repetition effects for faces, objects, and orthographic stimuli,” *Front. Hum. Neurosci.*, vol. 5, p. 6, 2011.

[17] U. Maurer, B. Rossion, and B. D. McCandliss, “Category specificity in early perception: face and word n170 responses differ in both lateralization and habituation properties,” *Front. Hum. Neurosci.*, vol. 2, p. 18, 2008.

[18] C. R. Pernet, P. Sajda, and G. A. Rousselet, “Single-trial analyses: why bother?,” *Front. Percept. Sci.*, p. 322, 2011.

[19] G. A. Rousselet, C. M. Gaspar, K. P. Wieczorek, and C. R. Pernet, “Modeling Single-Trial ERP Reveals Modulation of Bottom-Up Face Visual Processing by Top-Down Task Constraints (in Some Subjects),” *Front. Psychol.*, vol. 2, p. 137, 2011.

Compressive Sensing in Digital In-line Holography

Péter Lakatos
(Supervisors: Dr. Szabolcs Tőkés and Dr. Ákos Zarándy)
lakatos.peter@itk.ppke.hu

Abstract— **Compressive sensing (aka compressed sensing or sampling) is a novel signal reconstruction or sampling model, which enables significantly less measurement than reconstructed data for a class of signals. It also offers algorithmic solutions via the linear inverse problem. We use these models and algorithms to solve the reconstruction problem of digital in-line hologram of sparse or otherwise redundant images.**

Keywords - **compressed sensing; compressive sensing; digital holography; in-line holography; holographic tomography; sparsity; inverse problem; linear inverse problem;**

I. Introduction

Compressive sensing is a novel signal reconstruction or sensing model which enables significantly less measurement than reconstructed data. Not in general but for some wide class of signals. It applies the fact that most of the signals are sparse or redundant in some way. For example most of the images can be represented in some wavelet basis with only a few significant coefficients.

Compressive sensing grew up from questions raised up by medical imaging techniques (like MRI [1]) and after some theoretical groundwork [2-4] it produces a lot of practical (mainly in different imaging techniques) or simply fun (single pixel camera, [5]) results.

In the second section we introduce compressive sensing with some theoretical foundations and the linear inverse problem which is essential in the practical usage of it. In the third section we take a fast look to digital in-line holography. In the fourth section we show how we can adopt the philosophy and practice of compressive sensing to holography.

II. Compressive Sensing

There is lot of different aspect of compressive sensing. It can be introduce from the direction of signal sampling theorems, denoising functions [3] or random matrixes [2]. Here we will use a linear algebraic approach [15].

A. Linear algebra aproach of compressive sensing

In information theory and its related subjects almost every measuring or sensing process can be write in the form of a linear equation system:

$$g = \Phi f \qquad (1)$$

where f is the subject of the sensing, Φ represents the sensing process and g is the outcome of the sensing. Here f and g are real (or complex) valued vectors with size N and M, respectively, and ϕ is an N by M real (or complex) valued matrix. M is the number of measures. We know ϕ and g and we are interested in f. If a measuring is not in this form, discretization, linear approximation or some other processes (tricks) usually can help.

Such a linear system is easily solvable if $M \geq N$, i.e. we have at least as many equations as variables. On the other hand, if $M<N$, there is impossible to solve the equation, because there is infinitely many solutions. Unless we have some additional information or constraints on the variables (f).

Compressive sensing is dealing with the case of $M<N$ when some redundancy or sparsity on the subject of the sensing (f) is assumed or a priori known.

In the sparse case we can formalize the problem as

$$\hat{f} = \underset{f}{\operatorname{argmin}} \|f\|_0 \quad \text{subject to} \quad g = \Phi f \qquad (2)$$

where $\|f\|_0 = |\{i: f_i \neq 0\}|$ is the number of nonzero element of f. $\|f\|_0$ is also known as the l_0-norm of g (but in fact it is not a norm, because it is not scalable). So we search for the sparsest solution.

Redundancy in f means there is some basis (Ψ) in what f is sparse. Let α be the representation of f in this basis: $f = \Psi \alpha$. In this case we can formalize the problem as

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|\alpha\|_0 \quad \text{subject to} \quad g = \Phi \Psi \alpha \qquad (3)$$

and then take $\hat{f} = \Psi \hat{\alpha}$.

The problem with the above mentioned l_0-norm is that it is numerically hard to handle and extremely sensible to noise. Compressive sensing suggests that instead of the l_0-norm, we can recover f or α by using of the l_1-norm $\|\alpha\|_1 = \sum_{i=1}^{N} |\alpha_i|$. In this case we can formalize the problem as

$$\hat{\alpha}_1 = \underset{\alpha}{\operatorname{argmin}} \|\alpha\|_1 \quad \text{subject to} \quad g = \Phi \Psi \alpha. \qquad (4)$$

Compressive sensing guarantees that the solution of problem (2) and problem (3) are the same (i. e. $\hat{\alpha}_1 = \hat{\alpha}$), if there is incoherence (dissimilarity) between the sensing and the

P. Lakatos, "Compressive sensing in digital in-line holography,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 93-96.

sparsifying matrix, and the number of measures is not too small:

$$M \geq C \cdot \mu^2(\Phi, \Psi) \cdot K \cdot \log_{10} N \quad (5)$$

where C is a small positive constant, K is the maximal number of nonzero elements of $\hat{\alpha}$ and μ is the above mentioned similarity of Φ and Ψ, called mutual coherence:

$$\mu(\Phi, \Psi) = \sqrt{N} \cdot \max_{i,j} |\langle \Phi_i, \Psi_j \rangle| \quad (6)$$

where Φ_i and Ψ_j denote the i-th and j-th column vector of Φ and Ψ, respectively.

Notice that if $\mu(\Phi, \Psi)$ and K are not too big (and in a lot of theoretically or practically important cases they are not), then M is enough to be much smaller than N, unlike in the well known Nyquist-Shannon sampling theorem or in the linear algebraic considerations in the beginning of this section, where $M \geq N$ is required. It is not a contradiction since the redundancy or sparsity constraints.

B. The linear inverse problem

Compressive sensing states that we can solve problem (2) by solving problem (3). They can be reformulate as

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}}(\|g - \Phi \Psi \alpha\|_2^2 + \|\alpha\|_0) \quad (7)$$

$$\hat{\alpha}_1 = \underset{\alpha}{\operatorname{argmin}}(\|g - \Phi \Psi \alpha\|_2^2 + \|\alpha\|_1) \quad (8)$$

Both of them can be considered as a special case of the linear inverse problem:

$$\hat{x} = \underset{x}{\operatorname{argmin}}(\|y - K x\|_2^2 + \tau \rho(x)) \quad (9)$$

where x and y are vectors, K is a matrix with proper size, τ is a nonnegative constant called the regularization parameter and ρ is a $\mathbb{R}^N \to [0, \infty[$ function called the regularizer function. Commonly used regularizer functions are for example:

- the l_0-norm
- the l_1-norm
- the Euclidean or l_2-norm $\|x\|_2 = \left(\sum_{i=1}^{N} |x_i|^2\right)^{\frac{1}{2}}$
- the general l_p-norm $\|x\|_p = \left(\sum_{i=1}^{N} |x_i|^p\right)^{\frac{1}{p}}$
- if x represents an image the total variation norm $\|x\|_{TV}$ which we will introduce in the fourth section

One of the advantages of this reformulation that if we choose λ carefully, the effects of noise can be reduced [6].

For the solution of the linear inverse problem a lot of algorithms were developed recently thanks to the general interest for the compressive sensing. The best of them are the SpaRSA (sparse reconstruction by separable approximation, [7]), the IST (iterative shrinkage/thresholding [8]) and the TwIST (two-step IST, [9]). These are all special cases of the

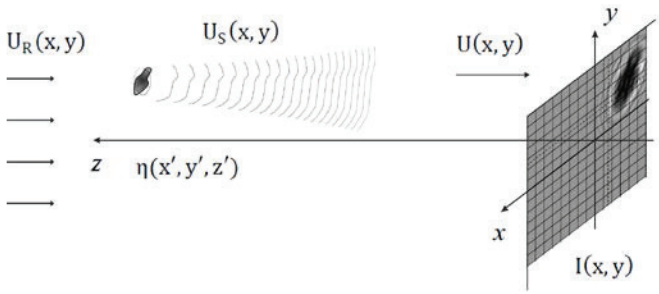


Figure 1: In-line hologram model

so-called proximal forward-backward splitting algorithm ([19]), which is provides solution for the

$$\hat{x} = \underset{x}{\operatorname{argmin}}(f_2(x) + f_1(x)) \quad (10)$$

problem, where f_1 and f_2 are proper (i.e. never equals to $-\infty$ and not the constant function with value $+\infty$ everywhere), convex and lower-semicontinuous (i.e. if it jumps, than the value of the function at the jump is equal to the lower limit point) and f_2 is also differentiable and has a Lipschitz-continuous gradient.

In our case

$$f_2(x) = \|y - K x\|_2^2, \quad (11)$$

$$f_1(x) = \tau \rho(x). \quad (12)$$

The proximal forward-backward splitting algorithm is an iterative algorithm. It takes two steps in turns. The first step is minimizing f_2 by moving x in the direction of $\nabla f_2(x)$. The second step is minimizing f_1 by moving x in the direction of

$$prox_{f_1}(x) = \underset{y}{\operatorname{argmin}}\left(f_1(y) + \frac{1}{2}\|x - y\|^2\right), \quad (13)$$

the so-called proximity operator, which is an extension of the projector operator. The proximity operator has a simple closed form for a lot of f_1 functions, i.e. for a lot of regularizer functions. For example,

$$prox_{\|\cdot\|_0}(x) = sign(x)\max\{|x| - 1, 0\} \quad (14)$$

the soft-threshold function.

The efficiency of the proximal forward-backward splitting algorithm is highly effected by the tuning of the algorithm.

III. Digital In-line Holography

A. Holography

Holography is an imaging technique based on the capture of coherent fields scattered from objects. It was introduced by Gabor in 1947 [10] and it became common after the development of the laser by Leith and Upatnieks in 1962. Gabor earned the Nobel Prize in Physics in 1971.

In holography [16] there is always a reference beam with complex amplitude $U_R(x,y)$ and an object beam scattered from the object $U_S(x,y)$, and we capture the interference

$U(x,y) = U_R(x,y) + U_S(x,y)$ of them in a photographic plate or in digital photometric sensor. Both of these devices can capture the intensity of the field:

$$I(x,y) = |U(x,y)|^2 = U(x,y) \cdot U^*(x,y) = \\ = |U_R(x,y)|^2 + |U_S(x,y)|^2 + \\ + U_R^*(x,y) \cdot U_S(x,y) + U_R(x,y) \cdot U_S^*(x,y) \quad (16)$$

If after the capture of I we light the photographic plate with the reference beam (or in the digital case simulate it), we get

$$U_R(x,y)I(x,y) = U_R(x,y)(|U_R(x,y)|^2 + |U_S(x,y)|^2) + \\ + |U_R(x,y)|^2 \cdot U_S(x,y) + U_R(x,y) \cdot U_S^*(x,y) \quad (17)$$

The first term is the reference beam with slightly modified amplitude. The second is the object beam, which forms a real image of the object. Finally the third term is called the "conjugate object beam" which forms an artifact called the "twin image".

B. In-line holograhy

There are plenty of holographic processes, but we can easily group them by the route of the reference beam compared to the scattered beam. In the off-axis holography the two beams are not parallel when they arrive to the sensor. In the on-axis holography the two beams are parallel, but this is achieved by a beam splitter. Finally in the in-line holography the two beams are also parallel and the reference beam arrives to the sensor among the scattering objects. The last one works only if there are a few and little objects in a transparent volume. It also suffers from the effect of the twin image, but it is easy and cheap to realize it.

In in-line holography we usually use a plane waves with high amplitude as reference beam, so it can be considered as constant $U_R(x,y) = U_R$ with high intensity compare to the scattered beam:

$$I(x,y) = |U_R|^2 + U_R^* U_S(x,y) + U_R U_S^*(x,y) \quad (18)$$

$$I(x,y) = |U_R|^2 + 2\,Re\left(U_R^* U_S(x,y)\right) \quad (19)$$

With the Born approximation the scattered beam can be considered as

$$U_S(x,y) = \iiint \eta(x',y',z') \cdot h(x-x', y-y', z-z')\,dx'dy'dz' \quad (20)$$

where η is the scattering density of the measured volume, z is the distance of the sensor and h is the point spread function (aka impulse response function).

C. Digital in-line holograhy

After discretization and consider the finite aperture we get

$$U_{S,n_x,n_y} = U_S\left(n_x \cdot \Delta p, n_y \cdot \Delta p\right) = \\ \sum_{m_x} \sum_{m_y} \sum_{m_z} \eta\left(m_x \cdot \Delta y, m_y \cdot \Delta x, m_z \cdot \Delta z\right) \cdot h(m_x \cdot \Delta x - n_x \cdot \\ \Delta p, m_y \cdot \Delta y - n_y \cdot \Delta p, z - m_z \cdot \Delta z) \quad (21)$$

where Δy, Δx and Δz are the size of a voxel (3D volume pixel) and Δp is the size of a pixel in the sensor [17]. We can rearrange (11) in the form of

$$U_S = H \cdot \eta \quad (22)$$

with the vectors U_S and η and the matrix H. With this we get

$$I = |U_R|^2 + 2\,Re(U_R^* \cdot H \cdot \eta) \quad (23)$$

which is, if H and U_R are real valued,

$$d = c \cdot 1 + H \cdot \eta \quad (24)$$

where d is the measured intensity data, the 1 is a vector containing only ones and c is a constant.

D. The Gerchberg-Saxton-Fineup method

Since with an optic sensor we can only measure the intensity of the light and we can't measure the phase of it, we are losing information. These results the so-called twin image problem. One of the solution for the twin image problem is the Gerchberg-Saxton algorithm ([20]) and its variants, the Gerchberg- Saxton- Fineup algorithms ([21]).

The Gerchberg-Saxton algorithms are iterative algorithm. These take two steps in turns. The first step is minimizing the twin image effect by some a priori information of the image (for example the scattering density is almost everywhere 0). The second step is to minimize the error we caused with the first step by modify the phase of the hologram.

The Gerchberg-Saxton algorithms also can be viewed as a complex optimization method ([22]).

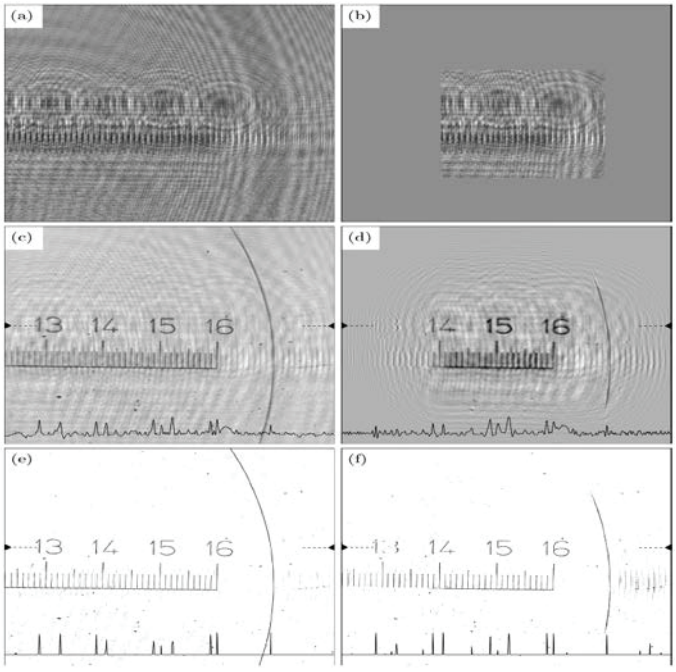


Figure 2: (a) hologram, (b) hologram with missing pixels on the side, (c-d) classical reconstructions, (e-f) compressed sensing reconstruction

IV. COMPRESSIVE HOLOGRAPHY

In the previous section we saw that holographic imaging can be representing as a linear system. This observation clear the way for the compressive sensing. In this section we introduce two case studies which both realize a compressive sensing based reconstruction method of digital in-line holography.

A. Hologram reconstruction with sparsity constraints

In [11] Denis et al. used a 2D reconstruction, so it reconstructs objects from a fixed z depth in one run. However they ran their algorithm for a series of depths, one after other, just like the classical holographic reconstruction methods.

They chose the Fresnel approximation as the point spread function and assumed that the objects are located sparsely in the examined volume, so their regularizer function was the l_1-norm:

$$\hat{\eta}_1 = \underset{\eta}{\operatorname{argmin}}(\|\mathrm{H}\cdot\eta - \mathrm{d}\|_2^2 + \tau\|\eta\|_1) \qquad (25)$$

They used a modified version of the IST [9] to solve it with changing τ in every step.

They achieved axial resolution of 60 μm, and their algorithm worked well with objects outside of the image. Results showed in Fig. 2.

B. Hologram reconstruction with smoothness constraints

In [12] Brady et al. used the approach of the angular spectrum method [13] to reformulate the problem in the form of $d = H\cdot\eta$ with $\mathrm{H} = \mathrm{G}_{-1}\mathrm{Q}\mathrm{G}_1$. Here G and G_{-1} are the DFT and its inverse and Q represents the light propagation.

They handled the redundancy of the images not by finding a proper basis but by the minimalization of the total variation:

$$\hat{\eta}_{\mathrm{TV}} = \underset{\eta}{\operatorname{argmin}}\|\eta\|_{\mathrm{TV}} \quad \text{subject to} \quad \mathrm{d} = H\cdot\eta \qquad (26)$$

with $\|\eta\|_{\mathrm{TV}} = \sum_{\mathrm{m_x}}\sum_{\mathrm{m_y}}\sum_{\mathrm{m_z}}\left|\nabla\left(\eta_{\mathrm{m_z}}\right)_{\mathrm{m_x},\mathrm{m_y}}\right|$ and used the TwIST algorithm [9].

They give a wider overview of the subject in [14].

V. CONCLUSION AND FUTURE PLANS

We presented the compressive sensing and its application for digital in-line holography, both in a theoretic way and via case studies. Our plan is to adapt and optimize these algorithms to the water quality measuring digital holographic microscopy [18].

We want to investigate the relation between the Gerchberg-Saxton algorithms and the proximal forward-backward splitting algorithm in this scenario.

Further plan is to extend these models to use multiple nonparallel detectors to reduce axial resolution.

REFERENCES

[1] Michael Lustig, David Donoho, and John M. Pauly, Sparse MRI: The application of compressed sensing for rapid MR imaging, Magnetic Resonance in Medicine, 58(6), pp. 1182 - 1195, 2007

[2] David Donoho, Compressed sensing, IEEE Trans. on Information Theory, 52(4), pp. 1289 - 1306, 2006

[3] E. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications on Pure and Applied Mathematics, 59(8), pp. 1207-1223, 2006

[4] Emmanuel Candès and Terence Tao, Near optimal signal recovery from random projections: Universal encoding strategies?, IEEE Trans. on Information Theory, 52(12), pp. 5406 - 5425, 2006

[5] Abdorreza Heidari, D. Saeedkia, A 2D Camera Design with a Single-pixel Detector, Int. Conf. on Infrared, Millimeter and Terahertz Waves, Busan, South Korea, 2009

[6] R. Tibshirani, Regression shrinkage and selection via LASSO, Journal Royal Statistical Society B, vol 58, pp 267-288, 1996

[7] Stephen J. Wright, Robert D. Nowak, Mário A. T. Figueiredo, Sparse reconstruction by separable approximation, Journal IEEE Transactions on Signal Processing, Volume 57 Issue 7, Pages 2479-2493, 2009

[8] José M. Bioucas-Dias , Mário A. T. Figueiredo, Two-step algorithms for linear inverse problems with non-quadratic regularization, IEEE International Conference on Image Processing ICIP', 2007

[9] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Communications on Pure and Applied Mathematics, V 57, I 11, pp 1413-1457, 2004

[10] D. Gabor, "A new microscopic principle", Nature, 161, pp 777-778, 1948

[11] Loïc Denis, Dirk Lorenz, Eric Thiébaut, Corinne Fournier, Dennis Trede, "Inline hologram reconstruction with sparsity constraints.", Opt.Lett. 34, pp 3475-3477, 2009

[12] David J. Brady, Kerkil Choi, Daniel L. Marks, Ryoichi Horisaki, and Sehoon Lim, Compressive Holography, Optics Express, Vol. 17, Issue 15, pp. 13040-13049, 2009

[13] Kyoji Matsushima, Tomoyoshi Shimobaba, Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields, Opt. Express 17, 19662-19673, 2009

[14] Sehoon Lim, Daniel L. Marks, and David J. Brady, Sampling and processing for compressive holography, Applied Optics, Vol. 50, Issue 34, pp. H75-H86, 2011

[15] Yair Rivenson, Adrian Stern and Bahram Javidi, Compressive Fresnel Holography, IEEE/OSA Display Technology, Journal of , vol.6, no.10, pp.506-509, 2010

[16] J. W. Goodman, Introduction to Fourier optics, 3rd Ed., Roberts and Company Publishers, 2005

[17] Corinne Fournier, Loic Denis, Eric Thiebaut, Thierry Fournel, Mozhdeh Seifi, Inverse Problem Approaches for digital hologram reconstruction, Proc. SPIE 8043, 80430S, 2011

[18] Z. Göröcs, L. Orzó, M. Kiss, V. Tóth, Sz. Tőkés, In-line color digital holographic microscope for water quality measurements, Proceedings of the SPIE, Volume 7376, pp. 737614-737614-10, 2010

[19] Patrick L. Combettes and Valérie R. Wajs, Signal recovery by Proximal Forward-Backward Splitting, Multiscale Modeling & Simulation 4.4 (2005): 1168-1200.

[20] R.W. Gerchberg and W.O. Saxton, A practical algorithm for the determination of phase from image and diffraction plane pictures, Optik, 35, pp. 237-246, 1972

[21] J. R . Fineup, Phase retrieval algorithms: A comaprison, Appl. Opt., 21, 2758-2769, 1982

[22] Bauschke, Heinz H., Patrick L. Combettes, and D. Russell Luke, Phase retrieval, Gerchberg-Saxton algorithm, and Fienup variants: A view from convex optimization, (2001).

Multiset Reordering for Efficient Large-scale Unstructured Grid Simulation

Endre László
(Supervisor: Ph.D. Péter Szolgay)
laszlo.endre@itk.ppke.hu

***Abstract*—In the present paper the effect of reordering mesh and data in the memory is investigated. The meshes and related data structures in focus are used in CFD (Computational Fluid Dynamic) simulations. The present application considered is the Volna [1] tsunami simulation code. The Volna code is transformed to utilize the OP2 framework [2], [3], [4] features of parallel computing. Using the OP2 abstraction the simulation can run on multi-core CPU (with OpenMP and MPI), many-core GPU (with CUDA) and a cluster of these processors with MPI. The massively parallel nature of the computation raises extra complexity by means of data arrangement and access pattern. These issues are tackled by reordering mesh elements to improve data locality and increase the efficiency of multiset partitioning. An overall 30 times speed up is achieved on a single GPU and a scalable performance across MPI processes is ensured by a new multiset reordering algorithm.**

***Keywords*-multi-core; many-core; HPC; tsunami; reordering**

I. Introduction

Tsunami waves are among the most disastrous natural phenomena in the world. The waves are initiated by underwater activity, usually the displacement of huge tectonic layers. The rapid change of the bathymetry modifies the water level and the wave begins to propagate. The amplitude of the travelling wave changes according to the sea depth. In deep sea (where the tide is generated) the amplitude of the wave is in the order of centimetres (usually less than 100 cm), the length of the wave is in the order of hundreds of kilometres (in large tsunamis around 300 meters) and the velocity can reach up to as extreme as 900 km/h, but usually is in hundreds of km/h. As the wave approaches the dryland, where the stationary sea level is shallow, the vast amount of water slows down to tens of km/h and piles up high over the sea level (around 10 meters), while the wavelength decreases to tens of kilometres. The wave enters the dry-land and destroys it with the debris picked up. Every inundation scenario is different as it depends on the properties of the wave. Therefore simulating tsunami waves to get inundation maps or to do simulation in real-time is of high importance. Wave propagation predictions based only on seismic activity fails with high probability, making such predictions unreliable. Thus recent researches focus on numerical predictive modelling coupled with deep sea pressure measurements.

Dutykh et al. [1] developed VOLNA, a robust numerical model and a C++ implementation for performing such predictions. The underlying physical model is based on Nonlinear Shallow Water Equation (NSWE). The numerical model relies on the Finite Volume method and applies triangulated unstructured mesh to represent the real-world landscape and bathymetry. The latter is important because coastal lines tend to have fractal shape [5] that can be more correctly modelled with triangular meshes. The numerical model covers the total life cycle of the tsunami from the generation through the propagation to the inundation. The model is based on previous theoretical works of Bermudez [6], Anastasiou [7], Vázquez-Cendón [8], Alcrudo [9], Barthélemy [10].

The VOLNA numerical model has been abstracted by the OP2 framework. The OP2 abstraction makes it possible to run the code on various computing platforms (multi and many core clusters) with high efficiency and scalability. The code written using OP2 API is transformed to utilize either OpenMP, CUDA or MPI on a single node, or transformed to the combination of MPI and OpenMP, or the combination of MPI and CUDA. The translator is a MATLAB based code which does source-to-source translation. To make the VOLNA to OP2 transition easy for the researchers, parts of the original code was kept to interface the simulation setup to the new environment. Thus the VOLNA OP2 uses the same configuration (*.vln extension) and mesh files (GMSH file format) as input and the same mesh (VTK file format) and text files as output where appropriate. The *volna2hdf5* tool is used to transform the input configuration and data files to the HDF5 file format used by OP2. The HDF5 format make the distributed data handling possible for OP2.

When performing calculation with OP2 on an unstructured mesh using a parallel computer architecture, there are three issues that might ruin the performance. 1)Using unstructured mesh in a computation introduces analogous problem which exist in the field of sparse linear algebra, namely the irregular memory access pattern when reading data. In the case of large meshes, that are used in CFD simulations, the size of the L2 or L3 cache and the pre-fetching mechanism is not enough to hide the main memory access latency. The portion of data outside the interval of the cache involves a cache misses and thus extra latency. Improving data locality therefore is essential. 2) OP2 uses a graph coloring algorithm to overcome the issues of data race conditions when performing reduction or increments of variables in parallel. A set of elements assigned with the same color can be processed in parallel. The sets of different colors can be processed only one after the other. The number of color in a mesh therefore is important, because the more colors are assigned the more the problem is serialized. The

E. László, "Multiset reordering for efficient large-scale unstructured grid simulation,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 97-100.

number of colors produced by the algorithm is influenced by the order of visiting the elements in the mesh. 3) In case of using MPI in the simulation the partitioning is also influenced by the quality of the ordering of the mesh elements. Since the partitioning is done in parallel on different MPI processes the mesh is distributed among the processes. The initial distribution of the partitions is done on the basis of mesh ordering. The partitioners (ParMETIS or PT-Scotch) use the initial distribution to create the optimized partitions. Thus the quality of partitioning is indirectly influenced by the element ordering.

II. Numerical model

The underlying numerical model of VOLNA is briefly summarized in the present section. The detailed description of the physical and numerical model can be found in [1] [11] and [12]. Two parameters are important from the perspective of validating the applicability of the current tsunami model. The characteristic values in these parameters are a_0, h_0 and l which are the wave length the average depth and the characteristic wave length. In case of a tsunami the typical values are: $a_0 \approx 0.5m$, $h_0 \approx 4km$ and $l \approx 100km$. This implies that the two parameters (ϵ - non-linearity and μ^2 - dispersion) in such cases have negligible contribution to the Serre-type equations in offshore conditions:

$$\epsilon = \frac{a_0}{h_0} \ll 1 \ , \ \mu^2 = \left(\frac{h_0}{l}\right)^2 \ll 1$$

As the wave approaches the shore, the average depth h_0 tends to get smaller and the dispersion effect tends to get more significant. The authors of VOLNA claim that this effect can be neglected and the equations reduce to the Nonlinear Shallow Water Equations (NSWE):

$$\begin{aligned} \frac{d}{dt}H + \nabla \cdot (H\mathbf{u}) &= 0 \\ \frac{d}{dt}(H\mathbf{u}) + \nabla \cdot (H\mathbf{u} \times \mathbf{u} + \frac{g}{2}H^2) &= gH\nabla h \end{aligned} \tag{1}$$

Here $H = h + \eta$ denotes the total depth of the water, $\mathbf{u} = (u, v)(\mathbf{x}, t)$ is the depth averaged horizontal velocity function, $h(\mathbf{x}, t)$ is the time-dependent bathymetry and g is the gravitational acceleration constant. The importance of $h(\mathbf{x}, t)$ being time-dependent lies in the models' capability of generating waves by underwater movements.

The above equation is rewritten into a conservation law form by introducing the conservative variable w:

$$\frac{\partial \mathbf{w}}{\partial t} + \nabla \cdot \mathcal{F}(\mathbf{w}) = \rho(\mathbf{w}) \tag{2}$$

where $\mathbf{w}$ is defined as:

$$\mathbf{w}(\mathbf{x}, t) : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^3, \ w = (w_1, w_2, w_3) = (H, Hu, Hv) \tag{3}$$

The rest of the formulation with the spatial and time discretization of the conservation law form and the boundary conditions is detailed in [1].

III. OP2

OP2 is a unity of an abstraction, framework and function library written in C and FORTRAN, a successor of OPlus (Oxford Parallel Library for Unstructured Solvers) [13]. The aim of the framework is to: 1) enable *general* data representation in arbitrary dimensions, 2) provide good *performance* on parallel systems with 3) *portability* across platforms, 4) a *single source* code that is translated to a variety of different platforms and 5) code longevity.

OP2 uses the concept of sets (`op_set`), datasets (`op_dat`), set maps (`op_map`) between sets and loops over sets (`op_par_loop`). These concepts are the basis of the abstraction that aims hiding the implementation details from the application programmer. All these concepts are defined by the particular solution of the problem. A set of nodes and a set of cells in a graph might be an `op_set`. One might define an `op_map` between the `op_set` of nodes and the `op_set` of cells, which defines the connectivity between these elements. Flow data might be associated with an `op_set`, e.g. every cell in the mesh contains the amount of fluid in the cell and the velocity at the cell center or every node in the mesh might contain its own position (x, y). An `op_par_loop` is a function that iterates though a sets elements and performs a given "kernel" calculation.

IV. Reordering of elements

Transforming the original C++ code to the OP2 framework lead to a more efficient simulation. On the other hand the new code gave itself to new optimisation tasks. One of the key optimization is increasing data locality by reordering the sequence of processing elements. By having the original loops transformed into OP2 parallel loops the order of calculation execution (on parallel backend) is shuffled by the different execution paths. Therefore reordering the set elements for better data locality doesn't introduce significant deviation from the transformed code and thus from the original code. The code utilizing the reordering passes the validation tests. Also, some reordering of the execution code is permitted as long as it doesn't corrupt the validation.

Reordering data to achieve better performance in numerical codes has well established strategies [14]. *Matrix bandwidth reduction* is the method to reorder data elements in a way that it minimizes the adjacency matrix bandwidth. This method increases data locality in an unstructured (sparse) system, thus increasing performance. The better performance comes from the increased cache hit rate, which is a consequence of better data reuse and the higher probability that a neighbouring element is also loaded into the cache by a previous fetch. Such reordering algorithms are the Cuthill-McKee [15], Reverse Cuthill-McKee (RCM) [16], Gibbs-Poole-Stockmeyer (GPS) [17] etc.

In the present work the GPS bandwidth reduction algorithms is used as the basis of a multiset reordering. The GPS algorithm is implemented in the PT-Scotch library, thus using it means no extra library dependence compared to the base OP2 code. Reordering set elements involves reordering map

and data structures related to the primary set. The primary set in the present case is the set of cells. The reordering of all the related structures is done in multiple steps:

Reordering cell set related structures:

1) The cells-to-cells map is converted into an adjacency matrix of CSR (Compressed Sparse Row) format.
2) The CSR matrix is processed by the GPS algorithm, which results in a permutation and inverse permutation vector. These vectors specify the new ordering. The p_i element of the permutation vector specifies the new number of the ith element, whereas the p_i^{inv} value tells the opposite: the old value of cell i.
3) The elements of maps related to cell sets are renumbered according to cell permutation vector and map rows are shuffled according to the inverse cell permutation vector.
4) Data structures are reordered according to the inverse cell permutation. The data elements are shuffled according to inverse cell permutation.

Reordering edge set related structures:

1) The cells-to-edges map is used to create a permutation and inverse permutation vector for edge set. Edges are renumbered by visiting the reordered cells and assigning new numbers to the referenced edges. The new number of an edge is the iteration number of visiting it.
2) The elements of maps related to edge sets are renumbered and reshuffled the same way as cell-related maps.
3) Data structures related to edges are reordered the same way as cell-related data structures.

Reordering node set related structures:

1) The cells-to-nodes map is used to create a permutation and inverse permutation vector for node set. Maps related to nodes are renumbered the same fashion as edge-related maps.
2) The elements of maps related to node sets are renumbered and reshuffled the same way as edge and cell related maps.
3) Data structures related to edges are reordered the same way as edge and cell related data structures.

The reordering of data in OP2 Volna has a serious impact on the performance:

1) On a single node with sequential execution: the increased the data locality increases cache hit rate, thus the performance.
2) On a single node with parallel (OpenMP, CUDA) execution: beside the increased cache hit rate the efficiency of colouring threads is increased significantly. Having less colors means having less sequentially processed elements, therefore higher parallelism.
3) On multiple nodes: the better initial distribution helps partitioners to create partitions with much smaller halo size, thus much less communication is necessary for inter-node data exchange. The efficient partitioning makes the OP2 Volna code scale with the number of nodes in the system.

The GPS-based reordering is implemented in the *volna2hdf5* tool and it is done by default at every execution. The reordering can be disabled by the *no-reorder* switch, see help for details.

V. Results

The effect of reordering is significant in the presented cases. The reason for the poor ordering of the original meshes is the way they are generated. Usually these meshes are generated by refinement which inherently shuffles numbering. In Table I properties of four meshes are compared. The landslide mesh is generated in a structured way and thus has an acceptable numbering. The rest of the meshes are generated by GMSH mesh generator [18]. The effects of reordering is detailed in the following.

a) Effect on data locality: The reordering significantly improved the matrix bandwidth, the average distance, number of block colors and partition halo sizes. The definition of matrix bandwidth can be found in [14]. The average distance is the measure of the average distance from the diagonal in the adjacency matrix. This is similar to average matrix bandwidth.

b) Effect on block coloring: Coloring of elements in a block and coloring of parallel blocks is done to prevent data write conflicts during concurrent code execution. Block coloring is done by the OP2 library with the greedy coloring algorithm. This is the most significant performance barrier in the original mesh numbering. On Figure 1 the effect of traversal order on the coloring can be seen. The improper traversal results in high number of colors.

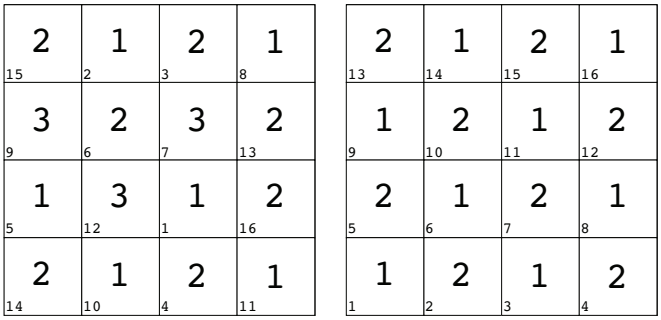


Fig. 1. The example of greedy block coloring clearly show the benefit of ordering. The grid of blocks on the left has a bad ordering with bandwidth 31 and 3 block colors. The grid of blocks on the right has a good ordering with bandwidth 9 and 2 block colors.

c) Effect on partitioning: From the point of simulations in a cluster the halo size is the other limiting factor. The values present in the table are ratio of the average partition size and the average halo size. The averages are computed on the partitions. Thus it can happen that the average halo size is greater than the average partition core size. In case of bad ordering this value is usually around 1. Clearly, the GPS reordering has a significant effect on the partitioning.

The reordering has three-fold effect on the performance which involve 1) improving data locality, 2) decreasing the number of block colors and 3) decreasing halo sizes of partitions. The latter one ensures that the computation can be distributed among the nodes of a cluster. The result is a

Mesh name	No. cells in mesh	Bandwidth		Avg. distance		Block colors		Halo sizes	
		Original	GPS	Original	GPS	Original	GPS	Original	GPS
landslide	20000	399	201	66.99	66.33	2	2	0.67	0.07
catalina	98198	196295	723	13166.6	181.64	63	2	1.32	0.03
matane	232992	465935	1149	46241.6	294.97	N/A	2	1.45	0.02
conical_island	1291056	2560911	2607	74828.8	797.2	93	3	1.28	0.01

TABLE I
MESH STATISTICS BEFORE (ORIGINAL) AND AFTER (GPS) THE REORDERING.

Example	No. cells in mesh	Total execution time [sec]			Speedup
		Original	OP2 OpenMP	OP2 CUDA	Original vs. CUDA
Landslide	500k	5 hours*	21.8	8.7	2070
Catalina	98k	68.9	3.4	2	34
Newrat2	171k	63	3.77	1.94	33

TABLE II
EXECUTION TIMES FOR DIFFERENT RUN CASES USING THE ORIGINAL VOLNA CODE AND VOLNA OP2 WITH OPENMP AND CUDA ON ONE NODE. * EXTRAPOLATED VALUE

scalable simulation across nodes. The performance benefits of data locality and block coloring can be seen on Table II. OP2 with reordered mesh provides significant speedup on various platforms. Speedup with CUDA implementation is above 30 for real-world applications. The specific case of the landslide example is exceptional and reaches the 2070 times speed up.

VI. CONCLUSION

In the present paper a new, GPS-based, multi set reordering algorithm is presented. The reordering algorithm improves the data locality, block coloring and partition halo size aspects of a real-world scientific model, the Volna tsunami simulation model. As a consequence of a 30 times speedup on one GPU card is achieved. The simulation is now capable of running on many nodes in a cluster with MPI. If the node in the cluster contains GPU than the computations scale in the multi GPU environment with the use of the OP2 library and the new reordering.

ACKNOWLEDGMENT

The author is thankful for the collaboration of Michael B. Giles, Gihan R. Mudalige from University of Oxford and István Reguly from Pázmány Péter Catholic University. The founding of TÁMOP-4.2.1./B-11/2-kmr-2011-0002 and TÁMOP-4.2.2./B-10/1-2010-0014 projects are gratefully acknowledged.

REFERENCES

[1] D. Dutykh, R. Poncet, and F. Dias, "The VOLNA code for the numerical modeling of tsunami waves: Generation, propagation and inundation," *European Journal of Mechanics - B/Fluids*, vol. 30, no. 6, pp. 598 – 615, 2011. Special Issue: Nearshore Hydrodynamics.

[2] M. B. Giles, G. R. Mudalige, Z. Sharif, G. Markall, and P. H. Kelly, "Performance analysis of the OP2 framework on many-core architectures," *SIGMETRICS Perform. Eval. Rev.*, vol. 38, pp. 9–15, Mar. 2011.

[3] M. Giles, G. Mudalige, B. Spencer, C. Bertolli, and I. Reguly, "Designing OP2 for GPU architectures," *Journal of Parallel and Distributed Computing*, no. 0, pp. –, 2012.

[4] G. Mudalige, M. Giles, I. Reguly, C. Bertolli, and P. H. J. Kelly, "OP2: An active library framework for solving unstructured mesh-based applications on multi-core and many-core architectures," in *Innovative Parallel Computing (InPar), 2012*, pp. 1–12, 2012.

[5] B. Sapoval, A. Baldassari, and A. Gabrielli, "Self-stabilized Fractality of Sea-coasts Through Damped Erosion," *AGU Spring Meeting Abstracts*, p. A6, May 2004.

[6] A. Bermúdez, A. Dervieux, J.-A. Desideri, and M. Vázquez, "Upwind schemes for the two-dimensional shallow water equations with variable depth using unstructured meshes," *Computer Methods in Applied Mechanics and Engineering*, vol. 155, no. 1–2, pp. 49 – 72, 1998.

[7] K. Anastasiou and C. T. Chan, "Solution of the 2D shallow water equations using the finite volume method on unstructured triangular meshes," *International Journal for Numerical Methods in Fluids*, vol. 24, pp. 1225–1245, June 1997.

[8] M. E. Vázquez-Cendón, "Improved treatment of source terms in upwind schemes for the shallow water equations in channels with irregular geometry," *Journal of Computational Physics*, vol. 148, no. 2, pp. 497 – 526, 1999.

[9] F. Alcrudo and P. Garcia-Navarro, "A high-resolution godunov-type scheme in finite volumes for the 2d shallow-water equations," *International Journal for Numerical Methods in Fluids*, vol. 16, no. 6, pp. 489–505, 1993.

[10] E. Barthélemy, "Nonlinear shallow water theories for coastal waves," *Surveys in Geophysics*, vol. 25, no. 3-4, pp. 315–337, 2004.

[11] Y. Kervella, D. Dutykh, and F. Dias, "Comparison between three-dimensional linear and nonlinear tsunami generation models," *Theoretical and Computational Fluid Dynamics*, vol. 21, pp. 245–269, July 2007.

[12] F. Dias and P. Milewski, "On the fully-nonlinear shallow-water generalized serre equations," *Physics Letters A*, vol. 374, no. 8, pp. 1049 – 1053, 2010.

[13] D. Burgess, P. Crumpton, and M. Giles, "A parallel framework for unstructured grid solvers," in *Programming Environments for Massively Parallel Distributed Systems* (K. Decker and R. Rehmann, eds.), Monte Verità, pp. 97–106, Birkhäuser Basel, 1994.

[14] N. E. Gibbs, W. G. Poole, Jr., and P. K. Stockmeyer, "A comparison of several bandwidth and profile reduction algorithms," *ACM Trans. Math. Softw.*, vol. 2, pp. 322–330, Dec. 1976.

[15] E. Cuthill and J. McKee, "Reducing the bandwidth of sparse symmetric matrices," in *Proceedings of the 1969 24th national conference*, ACM '69, (New York, NY, USA), pp. 157–172, ACM, 1969.

[16] A. George and J. W. Liu, *Computer Solution of Large Sparse Positive Definite*. Prentice Hall Professional Technical Reference, 1981.

[17] N. E. Gibbs, J. Poole, William G., and P. K. Stockmeyer, "An algorithm for reducing the bandwidth and profile of a sparse matrix," *SIAM Journal on Numerical Analysis*, vol. 13, no. 2, pp. pp. 236–250, 1976.

[18] C. Geuzaine and J.-F. Remacle, "Gmsh: A 3-d finite element mesh generator with built-in pre- and post-processing facilities," *International Journal for Numerical Methods in Engineering*, vol. 79, no. 11, pp. 1309–1331, 2009.

Prioritization of cancer drug combinations by integrating drug-drug interaction measures

Balázs Ligeti
(Supervisor: Dr. Sándor Pongor)
ligeti.balazs@itk.ppke.hu

***Abstract*—Drug combinations are known to be efficient in treating complex diseases such as cancer, diabetes, arthritis and hypertension. However, most combinations were found in empirical ways so there is a need for efficient computational methods. In this paper I will present a novel, easily usable method that can efficiently prioritize known cancer drug combinations (AUC=0.92). It considers not only the network phenomena such as crosstalks, feedback and feed forward loops (identified via perturbation analysis of the constituent drugs), but also therapeutic and functional similarities between the components identified by analysing the network of gene ontology data and therapeutic informations. The method is based on the assumption that those drugs can form efficient combinations that are linked to a large number of common perturbed proteins and share some therapeutic and functional properties (i.e. they regulate the same biological process, etc). We compared our predictions with the outcome of recently finished clinical trials (carried out onTrastuzumab, a well known and widely used cancer drug). The aggregated scores of the combinations containing Trastuzumab and different cytotoxic drugs show good correlation with the outcome of clinical trials, both with the objective response (OR - 0.62) and the progression free survival (PFS - 0.67).**

***Index Terms*—drug combination;drug interaction;**

I. Introduction

In the past few decades the number of novel marketed drugs have fallen much below the expectations despite the growing resources invested in this area [1], [2], [3]. Drugs designed by one drug - one target paradigm often fail during the clinical trials usually because of the unexpected side effects or the low therapeutic efficiency [1], [4]. In general, biological systems are robust against various kind of perturbations such as toxins, chemical compounds, mutations [3], [5]. For example, biological pathways are often redundant, diverse and modular. Furthermore they are rich in negative feedbacks, positive feedbacks, feed-forward and other regulatory loops that can compensate the effect of perturbations. Multitarget drugs or drug combination can overcome the problem of robustness since parallel modulation of multiple sites can more efficiently influence a system. Ágoston et al. [6], [7], [8] showed that multiple partial knockout of targets is more efficient than single knockout. In addition drug combinations have lower toxicity and therapeutic selectivity [9]. Instead of developing highly selective pharmacons, one should try to use multitarget drugs or drug combinations as a drug discovery paradigm and to base the design process on a broad scope of information [10]. However. finding efficient combinations is not easy since given the complexity of the underlying biological system. Nevertheless, the number of approved drug combinations is increasing, even though most of them were found by experience and intuition [12], [13]. Several experimental methods, even high throughput methods, have been developed for measuring the efficiency of drug combinations, such as Bliss independence or Loewe additivity [14], [15], but this kind of exhaustive search is impractical. Wong et al. used a stochastic search algorithm [16] while Calzoari et al. used a sequential decoding algorithms for finding the best combinations [17]. Yang et al. use differential equations to find a perturbation pattern that can revert the system from a disease state to a normal state [18]. Jin et al. employed a Petri net based model to microarray data in order to predict the synergism of drug pairs [19]. The common in these computational methods is that they require a large number of experiments or deep knowledge of the kinetic parameters of the pathways even if the search space is small. Others use data mining methods to integrate pharmacological and network data [20], [21], [22]. Li et al. used the concept of network centrality and disease similarity to prioritize drug combinations [23]. Wu et al. used the microarray profile of the individual drugs for the predictions [24], while others use the concept of synthetic lethality and the available gene interaction data [25], [26]. In this paper I present a novel drug combination prediction algorithm which is partly based on the assumption that the perturbations generated by the drugs propagate through the possible interactions between proteins. I also assume that the components of the combination have to share some therapeutic and functional properties that can be measured by using the Anatomical Therapeutic Chemical (ATC) Classification System and the Gene Ontology [34] annotations.

II. Methods

Figure 1 shows the general workflow of the method. Four different drug-drug interaction strength measures were used. Two are based on the analysis of the perturbation made by the drugs individually, the other two measure the functional and therapeutic similarities between the components. Each measure can be seen as a feature that describes one aspect of the drug - drug interaction. Then a logistic regression modell was trained using different features of known and random combinations. Finally the trained model ranked the candidate combinations.

B. Ligeti, "Prioritization of cancer drug combinations by integrating drug-drug interaction measures,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 101-105.

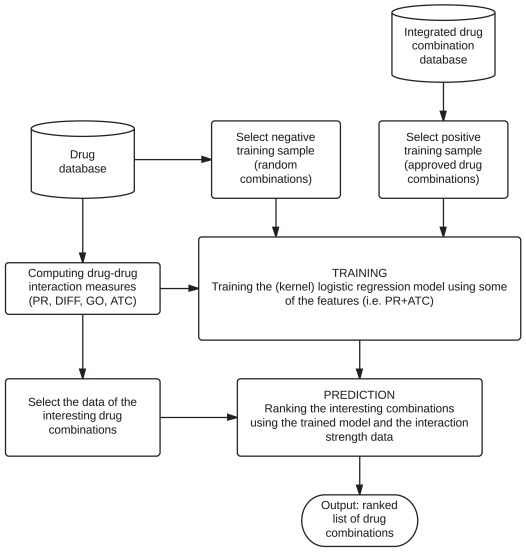


Fig. 1: *Each interaction measure (PR, DIFF, GO, ATC) can be seen as a predictor variable. A logistic regression model was trained using known and random generated drug combination data. Then the trained model was used to make predictions. The names refer to the variables used in training and prediction. PR+ATC means that both the ATC level similarity and the network based interaction measure were used.*

A. PageRank with prior (PR)

The propagation of the perturbation generated by a drug is modeled by a random walk initiated from the drug target proteins (PageRank with prior, which was successfully used to prioritize disease candidate genes based on a similar hypothesis [27], [28], [29], [30]). I define the subnetwork affected by a drug as the set of proteins significantly perturbed by the drug. These proteins are the nodes of the subnetwork. I assume that those drug combinations are strong that share many drug-affected proteins, in other words, their subnetworks substantially overlap. This overlap can be measured by the Jaccard measure (similarity measure between sets), where the elements of the sets are the nodes of the significantly perturbed subnetworks. The significance levels were computed with Monte Carlo simulations.

The network is a graph $G(V, E)$ where V, E are the set of nodes and edges, respectively. In this case the nodes represent genes or proteins, and the edges are the associations between them. The edges may have a weight, which can be interpreted as an association strength. Let A be the adjacency matrix of the graph. The element a_{ij} is the weight of the edge between node i and j, if there is no edge then it is 0. One could define a random walk on that graph by rescaling the edges to transition probabilites. Let M be a stochastic matrix of the graph $G(V, E)$, then m_{ij} is the probability of going to node j from node i.

$$M = G^{-1}A$$

Where G is a diagonal matrix, where $g_i = \sum_{j=1}^{|V|} A_{ij}$. The PageRank with prior [31] is a modified random walk, where in each step the random walker jumps back to one of the initial nodes or continues travelling with a certain probability.

$$P^{(i+1)} = (1-\alpha)\left(M^T P^{(i)}\right) + \alpha P^0 \tag{1}$$

$$p_i^0 = \begin{cases} \frac{1}{|N_T|}, & \text{if protein } i \text{ is drug target} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where N_T is the number of drug targets, P^i is a probability distribution, so p_k^i is the probability of being at node k in step i. P^0 is the initial probability distribution vector, which are the probabilities of starting the random walk at a given a node.

B. Regularized Laplacian Exponential Diffusion Kernel

Graph kernels can reveal important feutares of the graph structures, thus they are widely used in network analysis. The drug affected proteins can also be determined by using the Regularized Laplacian Exponential Diffusion Kernel ($K_{\mu,\alpha}$) [32] (DIFF). The formula of that kernel is:

$$K_{\mu,\alpha} = \sum_{k=1}^{\infty} \frac{\alpha^k}{k!}(-L_\mu)^k = e^{-\alpha L_\mu}$$

where L_μ is the regularized Laplacian of the graph:

$$L_\mu = \mu G - A$$

The ith drug (D_i) perturbation can be expressed with vector:

$$S_{DIFF}(D_i) = K_{\mu,\alpha} p_0 \tag{3}$$

where p_0

$$p_0 = \begin{cases} 1, & \text{if the protein } i \text{ is drug target} \\ 0, & \text{otherwise} \end{cases}$$

The jth element of $S(D_i)$ measures the disruption effect of D_i on protein j.

C. Randomizations and the drug affected proteins (DAP)

In order to find the subset of the drug affected proteins a Monte Carlo simulation procedure was used. In protein interaction network there are nodes which are more central or more important thus more likely to be reached by chance. To avoid this situation randomization procedure (10000 times) was applied to estimate statistical significance of each gene [33]. If we have p-values then we can define the set of drug affected proteins (DAPs) as follows:

$$DAP = \{v_j | v_j \in V, p_j < 0.05\}$$

I assumed that the sets of DAPs of the interacting drugs largely overlap, which is measured by the Jaccard coefficient, thus the PR and DIFF drug-drug interaction strength is:

$$S_{PR}(D_i, D_j) = \frac{|DAP_i \cap DAP_j|}{|DAP_i \cup DAP_j|}, \tag{4}$$

D. Gene ontology

For each drug a GO vector (g_i) was built, where each entry of the vector represents the presence or the absence of a GO term annotated to the drug targets. The ith entry is 1 if the ith term is annotated to the target protein, 0 otherwise. Then the cosine similarities between drugs can be computed.

$$S_{GO}(D_i, D_j) = 1 - \frac{g_i^T g_j}{\|g_i\|\|g_j\|} \quad (5)$$

E. ATC

To assess the therapeutic similarities between the components the Anatomical Therapeutic Chemical Classification (ATC) was used. System classifies drugs into groups at five levels. For each level the similarities between drugs were computed. Then the five similarities were aggregated. At each level l the similarity is determined by the Jaccard measure of that level of the code:

$$S^l_{ATC}(D_i, D_j) = \frac{\#\text{shared codes}}{\#\text{codes annotated to } D_i \text{ or } D_j} \quad (6)$$

F. Logistic regression model

The logistic regression (LR) is able to predict how successful an unknown drug combination will be using the four drug-drug interaction measures (S_{PR}, S_{DIFF}, S_{GO}, S_{ATC}). The combined score is:

$$S(D_i, D_j) = \frac{1}{1 + e^{\beta_0 + \beta_1 S_{PR} + \beta_2 S_{DIFF} + \beta_3 S_{GO} + \beta_4 S_{ATC}}} \quad (7)$$

Where the regression parameters β_i were estimated by the glmfit MATLAB function.

G. Score of drug regimens

All the drug-drug interaction measures are only applicable in pairwise cases, where the combinations have only two components. In the multicomponent cases one could simply aggregate the score of all possible pairwise interactions in the interaction. Let the combination DC have n components $DC = \{D_1, D_2, \ldots, D_n\}$ then the score of DC is:

$$S(DC) = \frac{\sum_{(D_k, D_n) \in DC \times DC} S(D_k, D_n)}{n} \quad (8)$$

III. Description of the experiments

All the algorithms were implemented in MATLAB 2012b. The used network was STRING [35]. The drug related data (drug targets, synonyms, aliases, ATC codes) were downloaded from the Drugbank [36], Stitch [37] and TTD databases [39]. The drug combination data were taken from the DCDB [38], TTD database [39] as well as collected from scientific articles.

A. Clinical trial data

This dataset contains information about the recently conducted clinical studies that include Trastuzumab. The dataset was created by the group of Balázs Győrffy. The measured response variables were the objective response (OR) and the progression free survival (PFS). In the cases where these values were missing, they were manually computed. The combination is identified by its components, thus the different combinations studied in the same trial were treated as independent entities, but no distinction has been made if the combinations had different dose or time schedule (these combinations have different OR, PFS values).

Features	AUC
ATC	0.6308
GO	0.8008
GO+ATC	0.8021
DIFF	0.8825
DIFF+ATC	0.9206
DIFF+GO	0.8808
DIFF+GO+ATC	0.9014
PR	0.6989
PR+ATC	0.7606
PR+GO	0.7797
PR+GO+ATC	0.7669
PR+DIFF	0.8894
PR+DIFF+ATC	0.8867
PR+DIFF+GO	0.8622
PR+DIFF+GO+ATC	0.8744

TABLE I: *The table shows the prediction performance of the different classifiers. The AUC values were computed by 10 fold cross validations where the negative drug combinations were generated randomly. The procedure was repeated 10 times to obtain the final average AUC.*

1) Measuring the classification performance: I used the AUC (area under roc curve) for measuring the ranking performance. The output is a ranked list. Whether a drug combination belongs to a positive (good combination) or negative class (not a good combination) depends on a variable treshold [40]. If the rank of the drug combination score is lower than the given treshold then it is considered to be positive otherwise negative, thus a FPR (false positive rate) and a TPR (true positive rate) can be determined. One could generate TPR, FPR for every possible ranking treshold, thus we got a ROC curve and an AUC value.

2) Experiments: Since a low number of true negative, unsuccessful DCs are available I used random combinations (selected from the database) as negative samples. The size of the negative and positive training set was the same. The target distribution of the drug components in the negative sample was the same as in the known set. Only the known and FDA approved cancer combinations were used as positive sample. Multicomponent combinations were also included by generating all the possible pairwise combinations from the components.

The propagation parameter α in PageRank, and the number of steps taken by the random walker (k) were 0.9 and 2, respectively. The diffusion paramater α was 0.01 and the regularization parameter μ was chosen as 0.2. The AUCs for all possible feature combinations (PR, DIFF, GO, ATC) were computed using 10fold cross-validation process. The training set selection and cross-validation procedure was repeated 10 times to obtain an average test AUC.

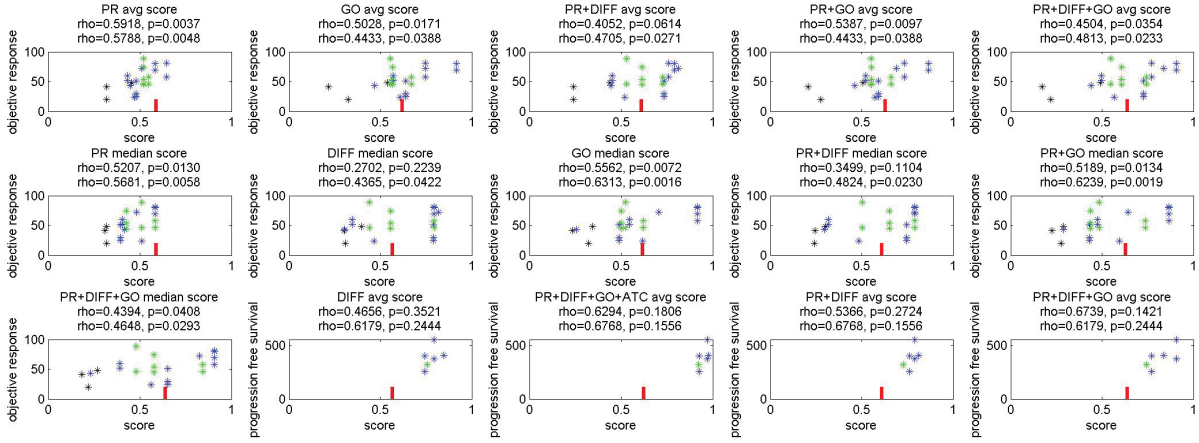


Fig. 2: *The figure shows the relation between the prediction scores of multicomponent drug combinations containing Trastuzumab and the results of the clinical trial. Only the significant (p-value* < 0.05*) or strong (*$\rho \geq 0.4$*) associations are showed. The scores are on the x axes, the responses (objective response, progression free survival) are on the y axes.*
black stars - combinations having neither taxan nor anthracycline
component red stars - combinations having anthracycline component
green star - combinations having taxan and anthracycline
blue star - combinations having a taxan component
red line - average score of known combination
The titles' 2nd and 3rd rows show the linear (Pearson) and the Spearman rank correlation coefficient and the corresponding p-values.

IV. Results

Table I shows the average test AUC-s. The best feature combination is the DIFF+ATC with AUC 0.9206 which is a serious improvement compared to the individual predictions (0.6308, 0.8824), thus the more is not always the better. **Figure 2** shows the significant or strong correlation between the aggregated score (**equation 8**) and the outcome of the clinical studies.

A. Limitation of the method

The main limitation of the method is that it is not possible to determine whether the interaction between two drugs is synergetic, antagonistic or additive, only the existence of the interaction can be presumed. Neither is it possible to decide whether the pathways affected by DC components are upregulated or downregulated. Because of the scantiness of the available information about the interaction between proteins such as the direction, interaction strength, interaction type (inhibiton, activiation, binding, phosphorization, etc.) the above limitations will not go away. An other limitation is that the optimal doses of components cannot be inferred from our predictions.

V. Conclusions

In this paper I presented a novel method that can successfully prioritize candidate drug combinations based on the hypothesis that those drugs can make efficient combinations that a) share therapeutic and functional properties and b) share a large number of perturbed proteins that can be simply measured by the Jaccard coefficient. I also showed that the integration of different kind of drug-drug interaction measures improved the performance compared to the individual classifiers. However, the method has some limitations; for example, the exact nature of the interaction (e.g. synergetic or antagonistic) can not be predicted due to the scantiness of information about protein interactions. On the other hand, the predicted ranking of the candidate combinations showed correlation with the outcome of clinical studies.

Acknowledgment

This project was developed within the PhD program of the Multidisciplinary Doctoral School, Faculty of Information Technology, Pázmány Péter Catholic University, Budapest. Thanks are due to my supervisor, Prof. Sándor Pongor for his help and guidiance throughout the project. I am grateful to Gergely Lukács for his advices and help in creating the drug combination database, Balázs Győrffy for collecting clinical trial data.

References

[1] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," Nature Chemical Biology, vol. 4, 682–690, 2008.
[2] P. Imming, C. Sinning, and A. Meyer, "Drugs, their targets and nature and number of drug targets," Nature Reviews Drug Discovery, vol. 5, p. 821–834, 2006.
[3] K. Hiroaki, "A robustness-based approach to systems-oriented drug design," Nature Reviews Drug Discovery, vol. 5, p. 202–210, 2007.
[4] S. I. Berger, and R. lyengar, "Network analyses in systems pharmacology," Bioinformatics, vol. 25, no. 19, 2466–2472, 2008.
[5] K. Hiroaki, "Biological Robustness," Nature Reviews Genetics, vol. 5, p. 826–837, 2007.

[6] V. Ágoston, P. Csermely, and S. Pongor, "Multiple weak hits confuse complex systems: A transcriptional regulatory network as an example," Physical Review E, vol. 71, no. 5, http://link.aps.org/doi/10.1103/PhysRevE.71.051909, 2005.
[7] P. Csermely, V. Ágoston, S. Pongor, "The efficiency of multi-target drugs: the network approach might help drug design," Trends in Pharmacological Sciences, vol. 26, p. 178–182, 2005.
[8] T. Korcsmáros, M. S Szalay, Cs. Böde, I. A Kovács, P. Csermely, "How to design multi-target drugs: Target search options in cellular networks," Expert Opinion on Drug Discovery, vol. 2, p. 1–10, 2007.
[9] J. Lehár, A. S. Krueger, W. Avery, A. M. Heilbut, L. M. Johansen, E. R. Price, R. J. Rickles, G. F. Short III, J. E. Staunton, X. Jin, M. S. Lee, G. R. Zimmermann, and A. A. Borisy, "Synergestic drug combinations tend to improve therapeutically relevant selectivity," Nature Biotechnology, vol. 27, p. 659–666, 2009.
[10] A-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," Nature Reviews Genetics, vol. 12, no. 1, p. 56–68, 2011.
[11] J. D. Feala, J. Cortes, P. M. Duxbury, C. Piermarocchi, A. D. McCulloch, and G. Paternostro, "System approaches and algorithms for discovery of combinatorial therapies," Wiley Interdisciplinary Reviews: Systems Biology and Medicine, vol.2, no. 2, p. 181–193, 2010.
[12] C. T. Keith, A. A. Borisy, and B. R. Stockwell, "Multicomponent therapeutics for networked systems," Nature Reviews Drug Discovery, vol. 4, p. 71–78, 2005.
[13] G. R. Zimmermann, J. Lehár, and C. T. Keith, "Multi-target therapeutics: when the whole is greater than the sum of the parts," Drug Discovery Today, vol. 12, no. 1-2, p. 34–42, 2006.
[14] W. R. Greco, G. Bravo, and J. C. Parsons, "The search for synergy: a critical review from a response surface perspective," Pharmacological Reviews, vol. 47, no. 2, p. 331–385, 1995.
[15] A. A. Borisy, P. J. Elliott, N. W. Hurst, M. S. Lee, J. Lehar, E. R. Price, G. Serbedzija, G. R. Zimmermann, M. A. Foley, B. R. Stockwell, and C. T. Keith, "Systematic discovery of multicomponent therapeutics," Proceedings of the National Academy of Sciences, vol. 100, no. 13, p. 7977–82, 2003.
[16] P. K. Wong, F. Yu, A. Shahangian, G. Cheng, R. Sun, and C. M. Ho, "Closed loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm," Proceedings of the National Academy of Sciences of the United States of America, vol. 105, no. 13, p. 5105–10, 2008.
[17] D. Calzolari, S. Bruschi, L. Coquin, J. Schofield, J. D. Feala, J. C. Reed, A. D. McCulloch, and G. Paternostro, "Search algorithm as a framework for the optimization of drug combinations," PLoS Computational Biology, vol. 4, no. 12, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2590660/, 2008.
[18] K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang, "Finding multiple target optimal intervention in disease-related molecular network," Molecular Systems Biology, vol. 4:228, Epub 2008 Nov 4.
[19] G. Jin, H. Zhao, X. Zhou, and S. T. Wong, "An enhanced Petri-net model to predict synergistic effects of pairwise drug combinations from gene microarray data," Bioinformatics, vol. 27, no. 13, p. 310–316, 2011.
[20] X. M. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. Noort, and P. Bork, "Prediction of drug combinations by integrating molecular and Pharmacological data," PLoS Computational Biology, vol. 7, no. 12:e1002323, Epub 2011 Dec 29.
[21] H. Fu-Yan, S. Jiangning, and Z. Xing-Ming, "Exploring Drug Combinations in a Drug-Cocktail Network, " IEEE International Conference on Systems Biology (ISB), Zhuhai, China, September 2 - 4, 2011, p. 382–387.
[22] X. Ke-Jia, S. Jiangning and Z. Xing-Ming, "A network biology approach to understand combination of drugs," The Fourth International Conference on Computational Systems Biology (ISB2010); 2010, p. 347–354.
[23] S. Li, B. Zhang, and N. Zhang, "Network target for screening synergistic drug combinations with application to traditional Chinese medicine," BMC Systems Biology vol. 5, Suppl 1:S10, 2011.
[24] Z. Wu, X. M. Zhao, and L. Chen, "A system biology approach to identify effective cocktail drugs," BMC Systems Biology, vol. 4, Suppl 2:S7, 2010.
[25] M. Cokol, H. N. Chua, M. Tasan, B. Mutlu, Z. B. Weinstein, Y. Suzuki, M. E. Nergiz, M. Costanzo, A. Baryshnikova, G. Giaever, C. Nislow, C. L. Myers, B. J. Andrews, C. Boone, and F. P. Roth, "Systematic exploration of synergistic drug pairs," Molecular Systems Biology, vol. 7:544. doi: 10.1038/msb.2011.71, 2011.
[26] J. Xiong, J. Liu, S. Rayner, Z. Tian, Y. Li, and S. Chen, "Pre-Clinical Drug Prioritization via Prognosis-guided genetic interaction networks," PLoS One, vol. 5, no. 11:e13937, 2010.
[27] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," The American Journal of Human Genetics, vol. 82, no. 4, p. 949–58, 2008.
[28] D. Nitsch, J. P. Gonçalves, F. Ojeda, B. de Moor, and Y. Moreau, "Candidate Gene Prioritization by Network Analysis of Differential Expression using Machine Learning Approaches," BMC Bioinformatics, vol. 11:460, 2010.
[29] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating Genes and Protein Complexes with Disease via Network Propagation," PLoS Computational Biology, vol. 6, no.1:e1000641, 2010.
[30] J. Chen, B. J. Aronow, and A. G. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks," BMC Bioinformatics, vol. 10:406, 2010.
[31] W. Scott, P. Smyth, "Algorithms for Estimating Relative Importance in Networks," International Conference on Knowledge Discovery and Data Mining -Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, p. 266 –275, 2003.
[32] T. Ito, M. Shimbo, T. Kudo, Y. Matsumoto. "Application of kernels to link analysis.," International Conference on Knowledge Discovery and Data Mining archive Proceedings of the eleventh ACM SIGKDD Pages: 586 - 592.–, 2005.
[33] P. N. Westfall, and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, Wiley, New York, 1993.
[34] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology, " Nature Genetics, vol. 25, no. 1, p. 25–29, 2000.
[35] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," Nucleic Acids Research, vol. 39 (Database issue), p. 561–8, 2010.
[36] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R Eisner, AC. Guo, DS. Wishart., "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.," Nucleic Acids Research, vol. 39 (Database issue), p. 1035–8, 2011.
[37] M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, LJ. Jensen, P. Bork., "STITCH 3: zooming in on protein-chemical interactions. ," Nucleic Acids Research, vol. 40 (Database issue), p. 876–8, 2012.
[38] Y. Liu, B. Hu, C. Fu, and X. Chen, "DCDB: Drug combination database " . Bioinformatics, vol. 26, no. 4, p. 587–588, 2010.
[39] F. Zhu, Z. Shi, C. Qin, L. Tao, X. Liu, F. Xu, L. Zhang, Y. Song, X. Liu, J. Zhang, B. Han, P. Zhang, Y. Chen., "Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery," Nucleic Acids Research, vol. 40 (Database issue), p. D1128–8, 2012.
[40] P. Sonego, A. Kocsor, and S. Pongor, "ROC analysis: applications to the classification of biological sequences and 3D structures," Briefings in Bioinformatics, vol. 9, no. 3, p. 198–209, 2008.

Improving the accuracy of morphological annotation

Attila Novák
(Supervisor: Gábor Prószéky)
novak.attila@itk.ppke.hu

***Abstract*—In this paper, we present some tools and algorithms that contribute to the improvement of the efficiency of the morphosyntactic annotation process while also improving its accuracy. We present a browser-based manual annotation environment that improves efficiency by presenting text and annotation in an intuitive and transparent manner. It employs methods to ensure that manual annotation work done does not get lost, corrupted or inconsistent when changes are made to the annotation scheme in the course of an annotation project. Finally, we present methods that we used to partially automate the process of the massive extension of the lexical database of a Hungarian morphological analyzer with new lexical items when the resources needed to be adapted to a new domain, thus reducing the error rate of an automatic morphological annotation tool by more than one third.**

***Keywords*-morphosyntactic annotation, morphological analysis, lexicon induction**

I. Introduction

Annotated corpora are important resources for supervised training and testing of natural language processing tools, for linguistic research and as sources of structured knowledge. The quality of such resources affects the quality of tools or resources derived from them. In addition, errors in gold standard corpora, which are used for testing the performance of NLP tools, may compromise the validity of the results thus obtained. Achieving as low error rates as possible in annotated corpora is thus an important goal.

The creation of annotated corpora is, in general, a relatively costly and time-consuming enterprise, as it usually involves the labor of human annotators. Even if the annotation is primarily generated using automatic tools, painstaking manual checking of at least part of the annotated data is necessary to guarantee a low error rate for at least parts of the data that are used as training and especially testing resources for automatic annotation tools. And, in the case of a lack of an automatic annotation tool for the given language, domain and/or type of annotation, massive human involvement in the annotation effort is inevitable.

Given the high cost of human annotation and the fact that the monotonous nature of the task results in humans also producing erroneous annotation, any help in reducing the burden of the human annotators is welcome. Automating as much of the process as possible and minimizing error rates from the very beginning is important.

In this paper, we present some tools and algorithms that contribute to the improvement of the efficiency of the morphosyntactic annotation process while also improving its accuracy. We present a browser-based manual annotation environment that improves efficiency by presenting text and annotation in an intuitive and transparent manner. The corpus annotation tool also employs methods to ensure that manual annotation work done does not get lost, corrupted or inconsistent when changes are made to the annotation scheme in the course of an annotation project, or when the lexicon of the morphological analyzer is updated in the background. Finally, we present methods that we used to partially automate the process of the massive extension of the lexical database of a Hungarian morphological analyzer with new lexical items when the resources needed to be adapted to a new domain.

II. The quality of annotated corpora

Annotated corpora consisting of texts with various markup can be used as sources of structured knowledge. However, before massive amounts of annotated text can be generated fully automatically, the same type of annotations need to be created on a smaller scale as resources for supervised training and testing of natural language processing tools that aim at generating the same kind of markup.

A. Errors in an annotated corpus

There are two types of errors in an annotated corpus. The first type is noise in the text to be annotated. The level of this inherent noise differs widely among different types of text. The consistency of orthography might be unusually low and its deviation from some orthographic standard unusually high for types of text that were created without control, such as in the case text collected from web fora, historical documents, transcribed field recordings in languages that lack a standard orthography or medical records packed with ad hoc abbreviations and semi-Latin medical expressions typed in a haste and without the assistance of any proofing tools. Simply annotating noisy input as if it were noiseless is not an ideal solution.

For noisy corpora, the text to be annotated needs to be normalized during the process so that the annotation can be based on a consistent representation of the text. Recording both the original and a normalized version of the text in the annotated corpus is important not only for the sake of philological preciseness but also because this way the corpus can be used for training of a robust automated annotation system that can handle noisy input by deriving error models from the training data that can be used to normalize the text before or during annotating it. It is also important to note that there may be differences between the original and the normalized version of the text concerning the tokenization of

A. Novák, "Improving the accuracy of morphological annotation,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 107-110.

the text into individual words. The annotation system should be capable to handle these differences of tokenization.

In addition to inherent noise (most of which gets normalized if things go well), there are errors introduced in the course of the annotation process. One type of these errors is failing to normalize or tokenize the text correctly. The rest are errors in the annotation generated.

In general, a document describing normalization and annotation guidelines is created at the beginning and updated in the course of the annotation project. One source of errors can be these guidelines themselves. Nonsensical specifications are ignored by some human annotators while they are followed by others, which results in inconsistency and erroneous annotation. A gap or indeterminacy in the guidelines also results in inconsistency.

Another source of errors is the tools used to generate the primary annotation in the corpus. A gap in the model of the tool due to a gap or erroneous annotation in the lexical resources of a rule-based or statistical tool results in annotation errors. E.g. a statistical tagger does not generate an annotation never seen for words encountered in its training data. A morphological analyzer never returns an annotation that is not in its lexicon. This results in a systematic neutralization of relevant distinctions. In the case of a lexicon gap, the right analysis is not even among the annotation candidates. The tool also may fail to generate correct annotation if the model implemented in the tool is not capable to capture some relevant generalization, e.g. a second-order HMM model may not capture long-distance agreement constraints. This results in random noise.

The third source of errors is the human annotators. They may fail to identify and normalize errors in the input text. They may fail to identify and correct errors introduced by automatic tools. A lack of intuitiveness and transparency of annotations makes it much harder for humans to detect errors. Human annotators themselves may generate erroneous annotation due to misunderstanding of the guidelines, fatigue/boredom, errors in the guidelines, lack of intuitiveness and transparency of annotations or incapability to make the necessary distinctions reliably. The annotations should be as intuitive and transparent as possible to maximize the performance of human annotators both in terms of accuracy and speed. The same is true for the user interface of the manual annotation system. For example, using an interlinear annotation format instead of a vertical tabular format is preferable, as this much more resembles the text representation that humans are used to.

B. Morphological annotation

Annotation may describe properties of individual atomic or composite linguistic units, from tokens (morphemes or words), through phrases and clauses to texts or even subcorpora, or it may represent syntactic or semantic relations between units, such as grammatical function, argument role etc. In this paper, we discuss only non-relational annotation of morphological and semantic properties of tokens (words).

Although it has been assumed for a long time by mainstream NLP authors speaking English, a quasi-isolating language, as a mother tongue that simply annotating word tokens with a single part-of-speech tag learnt from some annotated corpus will do, for languages of a morphology more complex than that of English, this is rarely enough: a more detailed analysis including at least a lemma in addition to a morphological tag is necessary for most applications. Moreover, for agglutinating languages like Hungarian, due to the much higher variation of different word forms,[1] even part-of-speech tagging requires the integration of a morphological analyzer into the tagging process in order to achieve good accuracies even if training corpora on the scale of a million words are available (which themselves cannot in practice be created without using a morphological analyzer). In our experiments with the automatic morphological annotation of Hungarian medical records, PoS tagging accuracy of the HMM trigram PurePos tagger [6] trained on the over-1-million-word Szeged Corpus [3] was only 83.82%. Using the Hungarian Humor morphological analyzer [5] , the tagging accuracy increased to 90.55%. Domain adaptation of the morphological analyzer as described in detail in section IV further improved the accuracy of the tagger to 93.77%. Adaptation of the tagger to handle the massive amount of abbreviations in this corpus further improved accuracy to 94.49%.

While incorporating a morphological analyzer dramatically improves the accuracy of PoS tagging, especially in the case of little training data and a morphologically complex language, it also provides lemmas and lists alternative analyses from which a human annotator can easily select the right analysis if the one selected by the tagger is not correct in the given context. Using a morphological analyzer as a source of annotations has another advantage: there is the possibility of changing the granularity of the annotation even for resources that have been created using less detailed tag set using the morphological analyzer to do the mapping. Previous Hungarian morphosyntactic annotation projects, for example, such as the Szeged Corpus and treebank project used a tag set of relatively low granularity,[2] which did not map derived members of the verbal paradigms, such as participles or factitives to the verb stem, instead, they were annotated as adjectives, adverbs or verbs unrelated to the base verb.

Resources created using this annotation scheme that makes less linguistic distinctions and relations explicit can be mapped to a richer representation using the morphological analyzer to reanalyze the word forms, and replace them if a richer analysis is available. We used a similar approach machine-generating a set of regular expressions and applying them to an already manually checked disambiguated subcorpus of historical documents containing a high number of passive (e.g. *mosatik* ‘to be washed’) and factitive (*mosat* ‘make somebody

[1] See e.g. [2] for a comparison of how the number of different word forms encountered as a function of corpus size for English and agglutinating languages like Finnish, Estonian or Turkish

[2] In the case of Hungarian even this means a tag set of a cardinality of over a thousand

wash/have something washed') verb forms[3] that had first been analyzed using an annotation of lower granularity to one that relates these forms to the base verb and identifies them as passive or factitive.

This latter example illustrates an interesting and important fact: it is often assumed that an annotation of lower granularity is easier to generate consistently and correctly than one that is more fine-grained. This assumption, however, is false. While a tag set distinguishing passives and factitives from other verb forms makes more important linguistic distinctions and relations explicit than one that does not distinguish them and does not relate them to the same lemma, as soon as lemmatization is part of the task, the richer and more valuable annotation can actually be generated much more accurately using the same HMM-based morphological disambiguation algorithms than the less-detailed one.

III. A MANUAL DISAMBIGUATION INTERFACE

Considering the factors above, we created a manual disambiguation interface to facilitate a workflow for disambiguation of morphosyntactic annotation as a semi-automatic process. The interface can be used to check and correct automatically pre-annotated text manually.

We created a web-based interface using JavaScript and Ajax where disambiguation and normalization errors can be corrected very effectively. The system presents the document to the user using an interlinear annotation format that is easy and natural to read. An alternative analysis can be chosen from a pop-up menu containing a list of analyses applicable to the word that appears when the mouse cursor is placed over the word. Note that the list only contains grammatically relevant tags and lemmas for the word returned by the morphological analyzer with the tag selected by the tagger ranked as the top candidate. This is very important since, due to the agglutinating nature of Hungarian, there are thousands of possible tags.

The original and the normalized word forms as well as the analyses can also be edited by clicking them, and an immediate reanalysis by the morphological analyzer running on the web server can initiated by double clicking the word. We use Ajax technology to update only the part of the page belonging to the given token, so the update is immediate. Afterwards, a new analysis can be selected from the updated pop-up menu.

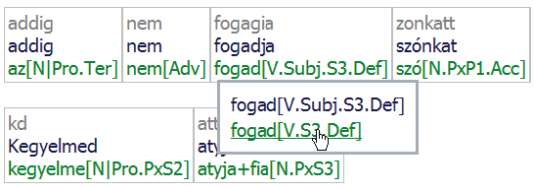


Fig. 1. The web-based disambiguation interface

As there may be a difference between the original and normalized tokenization, and because, even after thorough proofreading of the normalized version, there may remain tokenization errors in the texts, it is important that tokens and clauses can also be split and joined using the disambiguation interface.

The automatic annotation system behind the manual interface was created in a way that makes it possible that details of the annotation scheme be modified in the course of work. The modified annotation can be applied to texts analyzed and disambiguated prior to the modification relatively easily. This is achieved by the fact that, in the course of reanalysis, the program chooses the analysis most similar to the previously selected analysis based on a letter trigram similarity measure. The similarity measure is defined as the ratio number of letter trigrams in common and the number of all trigrams. Nevertheless, the system highlights all tokens the reanalysis of which resulted in a change of annotation, so that these spots can be easily checked manually. For changes in the annotation scheme where the simple similarity-based heuristic cannot be expected to yield an appropriate result (e.g. when upgrading the annotation scheme to use a more detailed analysis of derived verb forms), a more sophisticated method of updating the annotations using automatically generated regular expressions to replace old analyses can be used.

IV. ADAPTING THE MORPHOLOGICAL ANALYZER TO A NEW DOMAIN

We used the Humor morphological analyzer [5] in our annotation experiments. In this section, we describe how this morphology was adapted to the task of annotating anonymized medical records.

To improve the coverage of the morphological analyzer, we decided to expand its stem lexicon in the first round with items from relatively reliable sources. One of the sources that we used was the Dictionary of Orthography of Hungarian Medical Language [4]. This glossary contains no explicit information concerning either the PoS category or the language or pronunciation of the words included in it. However, this information was necessary to add them to the morphology. In addition, we needed to determine compound boundaries in compound words. Since several tens of thousands of words needed to be annotated, we decided to aid categorization and generation of the required annotation using automatic methods.

For part-of-speech categorization, we could rely on simple formal features for some word classes, such as for distinguishing names and abbreviations from the rest of the words. The rest was categorized using the suffix guesser algorithm implemented in the PurePos [6] automatic morphological disambiguation tool.

After categorizing a subset of the words by hand and training the suffix guesser algorithm on that data, we iteratively applied it to new words, manually corrected the errors, and retrained the model for the next round. For certain Greek or Latinate endings, it is difficult to decide if a specific word with that ending is used as an adjective, a noun or both. These had to be checked individually, which was rather time-consuming. For this reason, we took another piece of

[3]Passive is not used in present-day standard Hungarian, but it was extensively used in Old and Middle Hungarian

information into account at the automatic PoS categorization. In the case of Latinate multiword expressions in the dictionary, the last word is usually an adjective (unless the expression is a possessive construction), while the first word is generally a noun. Grouping words to be checked using this information helped to make manual checking of categorization much more efficient.

In addition to a PoS categorization of words, we had to decide which words are written using Hungarian orthography and which ones are foreign. For the latter, we had to add pronunciation so that they would be inflected correctly by the morphology. A subset of corresponding word Hungarian—foreign word pairs was listed in the dictionary as cross-references. Unfortunately, not all cross-referenced orthographical variants are in fact Hungarian—foreign pairs. Many of them are different variants of foreign orthography. Following a partial manual categorization, we used an adapted variant of the TextCat algorithm [1] that could be applied to short strings as well and provided an acceptably accurate foreign vs. Hungarian categorization. The situation was rather clear-cut in cases where the algorithm classified one item in a pair as rather Hungarian and the other as rather foreign. The algorithm helped us a lot in finding pairs (tuples) where all items were foreign spelling variants. The dictionary contains a high amount of foreign (mostly Greek-Latin and quite a few English and French) words that has no equivalent listed written in Hungarian orthography. We needed to recognize these as well, and in this case we could not rely on the kind of information provided by cross-referenced pairs.

For foreign words, we needed to add pronunciation. In the case of Hungarian—foreign cross-referenced pairs, some pronunciation information was provided by the Hungarian member of the pair, however, we needed to add a more Latinate pronunciation in addition to the completely Hungarianized one for many words (among others for most that end in the letter *s*). Since we had to add pronunciations to tens of thousands of words, we did not do this manually. Instead of using some machine-learning-based grapheme-to-phoneme algorithm, we quickly created a heuristic script based on regular expressions the output of which had rarely to be corrected after some initial tweaking. This could even be invoked for blocks of words from within the editor that we used to edit lexicon which was very useful to quickly add pronunciation to words that had been missed by the foreign word identifier.

Another task was the identification of compound boundaries in compound words as well as that of productively compounded elements. Prioritizing the addition of such elements to the lexicon early made it possible for us to reduce the number of words that had to be processed manually, as we could leave it to the morphology to handle compounds composed of these elements. In addition, this reduction of redundancy minimized the risk of inconsistent data entry. To discover compound members, we used the following procedure.

Words appearing in a general and the medical spelling dictionary consisting of at least two letters containing a vowel were stored in a trie and matched against the end of words in the dictionary. Prefixes of words so segmented were marked by the following features: the prefix is shorter than 4 letters; it is listed in either of the dictionaries; it contains a hyphen inside; it is itself a suffix of another word in the dictionary. Segmentations were ranked based on frequencies of prefixes/suffixes and these features. The compound member candidates were checked manually and the most frequent ones were added to the morphology at an early stage. Reanalyzing the rest with the thus extended morphology, and also keeping valid compounds output by the initial step, we managed to add segmented compounds to the lexicon.

The dictionary contains many nouns and participles derived from Latinate verbs with the verbs themselves not being listed. In these cases, we added the verbs, as the nominal derivatives would be thus added too anyway. In addition, it was needless to add a high number of derived adjectives, as adding their base nouns ensured the coverage of these items too.

In addition to the dictionary, we added names of drugs and active ingredients in them listed in a database downloaded from the web site of the National Institute of Pharmacy. Categorization and deciding the part of speech was not a problem for these. We needed to adapt our grapheme-to-phoneme rules, however, as in the names of drugs and ingredients, the English way of spelling basically the same Greek-Latinate vocabulary is predominant with lots of mute final *e*-s.

The third source of lexicon enhancement was the corpus itself. The corpus also played a pivotal role when adding words from the spelling dictionary: we prioritized adding words from the dictionary that appeared in the corpus. Words frequent in the corpus that still remained unanalyzed after adding a massive amount of words from the dictionary and the drug list were added in a second phase. Most of these were abbreviations.

Adding words form the spelling dictionary and the most frequent abbreviations from the corpus increased the size of the stem lexicon with more than 36000 items. In addition, we added 4860 drug names and active ingredients. Adaptation of the morphology reduced tagging error rate by one third.

References

[1] William B. Cavnar and John M. Trenkle. N-Gram-Based Text Categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

[2] Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, and Andreas Stolcke. Morph-based speech recognition and modeling of out-of-vocabulary words across languages.

[3] Dóra Csendes, János Csirik, and Tibor Gyimóthy. The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora LINC 2004 at COLING 2004*, pages 19–23, 2004.

[4] Pál Fábián and Péter Magasi. *Orvosi helyesírási szótár*. Akadémiai Kiadó, Budapest, 1992.

[5] Attila Novák. What is good humor like? [milyen a jó humor?]. In *I. Magyar Számítógépes Nyelvészeti Konferencia*, pages 138–144, Szeged, 2003. SZTE.

[6] György Orosz and Attila Novák. Purepos – an open source morphological disambiguator. In *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science.*, Wroclaw, Poland, 2012.

Hungarian Medical Text Processing - Spelling Correction, Structuring and Distributional Methods

Borbála Siklósi
(Supervisor: Dr Gábor Prószéky)
siklosi.borbala@itk.ppke.hu

***Abstract*—This paper presents the current state of processing Hungarian clinical texts. The main characteristics of such documents created at clinical settings are the non-standard use of language consisting of short fragments instead of proper sentences, usage of Latin words, many acronyms and the very frequent misspellings. Also these documents lack any kind of structure and are presented in text files containing only basic formatting. Thus the first steps of processing these documents are described, i.e. creating a robust structure capable of storing domain specific and grammatical information, then an automatic spelling correction system is presented in order to normalize the texts. Finally some methods were implemented that lead towards a higher level representation of clinical concepts and events.**

***Keywords*-clinical texts; spelling correction; distributional similarity; event extraction; natural language processing**

I. Introduction

Processing medical texts is an emerging topic in natural language processing. There are existing solutions mainly in English to extract knowledge from medical documents, which thus becomes available for researchers and medical experts. However, locally relevant characteristics of applied medical protocols or information relevant to locally prevailing epidemic data can be extracted only from documents written in the language of the local community.

In Hungarian hospitals, clinical records are created as unstructured texts, without any automated proofing control (e.g. spell checking). Moreover, the language of these documents contains a high ratio of word forms not commonly used, such as Latin medical terminology, abbreviations and drug names. Many of the authors of these texts are not aware of the standard orthography of this terminology. Thus processing such documents is not an easy task and automatic correction of the documents is a prerequisite of any further linguistic processing.

Even if having the texts corrected, the language of these clinical narratives are composed of short, ungrammatical sentence fragments or phrases instead of grammatically correct natural language sentences. Thus traditional syntactic parsing cannot succeed without defining a sublanguage grammar for this language as described in [1]. The first step towards building a representation of these documents is to retrieve prevailing patterns and relations from the texts. In order to model the distributional behaviour of this language, a distributional analysis was performed for frequent patterns and a method for building a domain-specific distributional thesaurus is proposed.

II. Correcting Spelling Errors

A characteristic of clinical documents is that they are usually created in a rush without proofreading. The medical records creation and archival tools used at most Hungarian hospitals provide no proofing or structuring tools. Thus the number of spelling errors is very high and a wide variety of error types occur. These errors are not only due to the complexity of the Hungarian language and orthography, but also to characteristics typical of the medical domain and the situation in which the documents are created. The most frequent types of errors are described in [2] and [3].

A. Context-aware Spelling Correction

My goal was to improve a baseline system presented in [2] that is able to generate correction suggestions for misspelled words considering them as isolated single words ignoring their context. The ranking in this baseline system is based on statistics built from domain-specific and general corpora in addition to grammaticality judgement of a wide coverage Hungarian morphological analyzer [4], [5]. The system is parametrized to assign much weight to frequency data coming from the domain-specific corpus, which ensures not coercing medical terminology into word forms frequent in general out-of-domain text. The baseline system was able to recognize most spelling errors and the list of the ten highest ranked automatically generated corrections contained the actually correct one in 98% of the corrections in the test set.

To improve the first-rank accuracy of the system, lexical context also needs to be considered. To satisfy this requirement, I applied Moses [6], a widely used statistical machine translation (SMT) toolkit. During "translation", the original erroneous text is considered as the source language, while the target is its corrected, normalized version. The probability of a target phrase is derived from statistical translation and language models by the decoder.

1) Translation Models: In my system, I applied three translation (correction) models according to three categories of words and errors. The first one handles possible abbreviations, the second one can split erroneously joined words and the third one handles all other errors. The translation model in each case is based on the suggestion generation system, where the ranking scores are normalized as a quasi-probability distribution. I applied this method instead of learning these probabilities from a parallel corpus as no such corpus is available. It should be noted, that though suggestions are

B. Siklósi, "Hungarian medical text processing - spelling correction, structuring and distributional methods,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 111-114.

TABLE I
DETAIL OF THE TRANSLATION MODEL FOR A WRONG COMMON WORD, ITS POSSIBLE CANDIDATE CORRECTIONS AND THEIR PROBABILITY.

hosszúságu	hosszúsági	0.01649
hosszúságu	hosszúságú	0.01560
hosszúságu	hosszúsága	0.01353
hosszúságu	hosszúságuk	0.01317
hosszúságu	hosszúságul	0.01292
hosszúságu	hosszúságé	0.01284
hosszúságu	hosszúság	0.01034

generated for each word in the sentences, these suggestions usually include the original form. The scoring ensures that if the original form was correct, then it will receive a high score, thus the decoder will not modify the word. Table I contains a common word that is misspelled in the input text. The word *hosszúságu* should be written as *hosszúságú* 'of length ...'. Another word form, *hosszúsági* 'longitudinal' is ranked higher by the original context-insensitive scoring algorithm, since it is also a correct and more frequent Hungarian word, and since the *u:i* correspondence is also a frequent error beside *u:ú* since *u* and *i* are neighboring letters on the keyboard. Though the rest of the words in the example are also correct candidates, they received a lower score. In theory without considering the context, all the others would be correct at the word level. Our language model will be responsible for making the contextually optimal choice.

Instead of applying this method on abbreviations (which are of high ratio in the texts), I collected possible variations of each potential abbreviation from the corpus together with their frequencies and used these values into probability estimates. An alternative translation model was created this way. These abbreviations are not present in the first translation model in order to prevent the system transforming them to other words. I then applied the decoder of the SMT system so that the translation model of the abbreviations is given a priority ensuring that abbreviations are transformed to their correct form rather than to other words.

2) Language Model: The language model is responsible for taking the lexical context of the words into account. In order to have a proper language model, it should be built on a correct, domain specific corpus by acquiring the required word n-grams and the corresponding probabilities. Since the only manually corrected portion of our corpus was the test set, I could not build such a language model. Though there are orthographically correct texts of other, mostly general domains, the n-gram statistics of these would not correspond to the characteristics of the clinical domain. However, I found that our clinical corpus contains several very frequent word sequences, but there are a relatively smaller amount of different n-grams compared to general texts. I assumed that the frequency of correct occurrences of a certain word sequence can be expected to be higher than that of the same sequence containing a misspelled word.

3) Decoding: To carry out decoding, we used the widely used Moses toolkit. The parameters of decoding can be set in the Moses configuration file, thus they can be changed easily in order to adapt the system to new circumstances and weighting schemes. During decoding, each input sentence is corrected by creating the translation models based on the suggestions generated for the words occurring in the actual sentence, and using the pre-built abbreviation translation model and the language model.

TABLE II
EVALUATION RESULTS OF THE CONTEXT-AWARE SYSTEM AND THE 1-BEST BASELINE

System	Accuracy
Baseline (1-best)	72.5%
SMT-based context-aware	88.28%

4) Results: In order to evaluate our system, a manually corrected test set of clinical documents was necessary. We randomly selected 2000 sentences and sentence fragments from the corpus, from various clinical departments and corrected these texts regarding both tokenization and spelling. The remaining part of the corpus contained 978,000 sentences and that was used for creating the language model. Both sets of sentences only contained free text parts of clinical reports. Tabular laboratory data, measurement results, headers, ICD coding and other structured content were previously filtered out. In spite of this, there were still a high number of sentences both in the training and test sets that contained hardly any real words, consisting of sequences of abbreviations and numbers.

I evaluated how well the baseline (i.e. the first-best suggestion generator) and the new systems performed on correcting erroneous words in the test set. The size of the test set was 19148 tokens (6744 types), of which 1289 tokens (847 types) were misspelled or in a non-standard form. More than half of these erroneous words are potential abbreviations of length less than 4 characters. Table II shows the performance of each system. The overall accuracy (i.e. the ratio of the well corrected words relative to the number of erroneous words) of the baseline system was 72.5%, while the SMT system improved the results significantly to 88.28%. I performed several experiments with different parameter settings that resulted in worse overall quality, but handled some special phenomena better. Some example sentences with their automatic correction is shown in Table III.

III. CREATING STRUCTURED TEXTS

A widely-accepted format for representing annotated corpora is xml. The structure of the clinical documents is represented at two levels. First an automatic segmentation of content parts were carried out as described earlier in [2]. Digging deeper into the textual contents of the documents, a more detailed representation of these text fragments was necessary. That is why I store each word in each sentence in an individual data tag, augmented with several information. Such information are the original form of the word, the corrected form, its lemma and part-of-speech tag, and some phrase level information such as different types of named entities. The lemma and PoS information are produced by PurePos ([7]), the named entities are produced by the system created for the

TABLE III
ORIGINALLY ERRONEOUS SENTENCES (ORIG) WITH THE AUTOMATIC CORRECTION OF THE BASELINE (BL) AND THE SMT SYSTEMS AND THE MANUALLY CORRECTED REFERENCE (REF)

Orig	csppent előírés szerint ,
Bl	cseppent előír és szerint ,
SMT	cseppent előírás szerint ,
Ref	cseppent előírás szerint ,
Origi	th: mko tovább 1 x duotrav 3 ü-1 rec , íb: 2 x azoipt 3 ü-1 rec
Bl	th: mko tovább 1 x duotrav 3 ü-1 sec , kb: 2 x azoipt 3 ü-1 sec
SMT	th.: mko tovább 1 x duotrav 3 ü-1 rec , kb: 2 x azopt 3 ü-1 rec
Ref	th.: mko. tovább 1 x duotrav 3 ü-1 rec , kb.: 2 x azopt 3 ü-1 rec
Orig	/alsó m?fogsor .
Bl	/alsó műfogsor .
SMT	alsó műfogsor .
Ref	alsó műfogsor .
Orig	vértelt nyállkahártyák , kp erezett conjuctiva , fehér sclera .
Bl	vértelt nyálkahártyák , kp erezett conjunctiva , fehér sclera .
SMT	vértelt nyálkahártyák , kp. erezett conjunctiva , fehér sclera .
Ref	vértelt nyálkahártyák , kp. erezett conjunctiva , fehér sclera .

master thesis of Emánuel Pirk ([8]). The sentence "Azarga th. kezdünk" is represented as the following example shows:

```
<sent>
     <surf>Azarga th. kezdunk </surf>
     <w NE="b-MED" id="102.0.0" type="">
          <orig>Azarga</orig>
          <corr>Azarga</corr>
          <lemma>Azarga</lemma>
          <pos>[FN][NOM]</pos>
     </w>
     <w NE="" id="102.0.1" type="">
          <orig>th.</orig>
          <corr>th.</corr>
          <lemma>th</lemma>
          <pos>[FN][NOM]</pos>
     </w>
     <w NE="" id="102.0.2" type="">
          <orig>kezdunk</orig>
          <corr>kezdunk</corr>
          <lemma>kezd</lemma>
          <pos>[IGE][t1]</pos>
     </w>
</sent>
```

IV. TOWARDS SEMANTIC REPRESENTATION

A. Distributional Similarity

Applying natural language processing techniques in the domain of clinical narratives in English usually involves the usage of handmade medical ontologies, or at least manually curated word lists. Such resources are very hard and expensive to create in Hungarian. That is why I started to experiment with methods based on distributional similarity, which is an unsupervised method for automatically acquiring variant ways of expressing relevant words and measuring their similarity.

1) Computing the Similarity of Two Words: The main idea is that semantically similar words tend to occur in similar contexts ([9]). Thus the similarity of two concepts is determined by their shared contexts. The context of a word is represented as a set of features, each feature consisting of a relation (r) and the related word (w'). In related works these relations are usually grammatical relations, however in the case of clinical texts, their grammatical analysis performs poorly, making a rather noisy model. In [10] Carroll et al suggest using only the occurrence of surface word forms within a small window around the target word as features. In my work I applied a mixture of these ideas by applying the following relations to determine the features for a certain word:

- prev_1: the previous word
- prev_w: words preceding the target word within a distance of 2 to 4
- next_1: the following word
- next_w: words following the target word within a distance of 2 to 4
- pos: the part-of-speech tag of the actual word
- prev_pos: the part-of-speech tag of the preceding word
- next_pos: the part-of-speech tag of the following word

Each feature is associated with a frequency determined from the corpus. From these frequencies the amount of information contained in a tuple of (w,r,w') can be computed by using maximum likelihood estimation. This is equal to the mutual information between w and w'. Then to determine the similarity between two words (w_1 and w_2) I used the similarity measure described in [11], i.e.:

$$\frac{\sum_{(r,w)\in T(w_1)\bigcap T(w_2)}(I(w_1,r,w)+I(w_2,r,w))}{\sum_{(r,w)\in T(w_1)} I(w_1,r,w)+\sum_{(r,w)\in T(w_2)} I(w_2,r,w)}$$

where T(w) is the set of pairs (r,w') such that I(w,r,w') is positive.

2) Building Distributional Thesauri: Having this metric, I computed the pairwise distributional similarity between the most frequent nouns, verbs and adjectives in the corpus creating a list of words for each target concept ranked according to their similarity. This can be easily transformed into a thesaurus with a tree structure, so that each sense of a word is represented by a subtree. However this requires very long computational time, which is planned to be carried out in the future. Some examples for the most similar words for some concepts are shown in Table IV. In my experiments the words to be ranked were only the 15-20 most frequent verbs, nouns and adjectives.

B. Detecting events

The free text parts of clinical narratives usually contain a few types of information. These are most often measurement results, past events, history and symptoms and present events, present symptoms and suggested therapies. First I applied a method to extract simple events with regard to their temporality (past or present), and polarity (negated or not). By this time I have determined five classes of retrieved information:

TABLE IV
EXAMPLE FOR THE LIST OF THE MOST SIMILAR CONCEPTS FOR SOME TARGET WORDS WITH THE MEASURE OF SIMILARITY FOR A SMALL TEST SET.

kontroll (N)	vizsgálat 0.200137414551; eset 0.164033128256; panasz 0.155728295147; műtét 0.0913003247578
cornea (N)	hátlap 0.14516804352; csarnok 0.137587581556; conj 0.119715252986; pupilla 0.0924046308243; lencse 0.085279167572
javasol (V)	végez 0.0757854615073; történik 0.0727785274109; felír 0.0647341637093; használ 0.0642329428926;
használ (V)	javul 0.0934773088188; cseppent 0.0876840590879; lát 0.0693182060826; fáj 0.067417048954; kezel 0.0656283245778;
tiszta (ADJ)	békés 0.154524127446; sima 0.137185024375; sekély 0.113384183723; sárgás 0.11103999917; ép 0.0983724197695;
rossz (ADJ)	homályos 0.12185879024; piros 0.103178644583; kicsi 0.0754819988997; jó 0.0648128527085; bal 0.0632304528447; száraz 0.0456163404988

1) Examined area: since I used documents only from the department of ophthalmology, the main target area of the examinations documented in the texts are limited to either the left, or the right or both eyes. There might be some additional symptoms or examinations carried out, but the main target can easily be detected by retrieving the target side.

2) Wish of the patient: in most cases, the patient has a direct wish about the purpose of visiting a doctor. They want an examination carried out, or have the doctor prescribe some medication, or especially in the domain I use, they want some glasses or contact lenses. Such wishes are retrieved by looking for some trigger words and their variations (*"szeretne","kér",etc*)

3) History and past events: these include verbs in past tense together with their complement or target. These are not necessarily neighbouring words of the verb, but still I applied a baseline algorithm to find the nearest possible complement candidate. These are then divided into groups of negated and non negated events. Some examples for such events are *"occlusio zajlott", "olvasószemüveget viselt", "károsodás nem igazolódott"*.

4) Present findings and symptoms: similar to the previous category, but in this case the verb is in present tense. For example *"elfogadhatóan lát", "kóros nem látszik"*. These are also grouped into negated and non negated events.

5) Nominal events: a specific characteristic of the clinical narrative language is the short telegraphic phrases, which sometimes though describe an event or a state, they do not include any verb. For example such phrases like *"sárgás magreflex", "ép papilla", "szűkebb artériák"* are used as standalone phrases. In standard language these are less frequent, or at least an existential verb is present along them.

The applied methods I used to extract such events are basic pattern recognition or grammatical structures, which do work with high recall, but low precision. However a significant improvement can be expected from integrating the above described distributional thesaurus to the event extraction methodology.

V. FURTHER PLANS

The medical language processing system has already several modules developed with some baseline results. The above described experiments are carried out by using these modules in a pipelined manner. However a more complex integration and communication between these modules are required, having higher level tools an effect on lower levels of processing. Such a parallel integration of the different processing methods is of crucial importance and is one of the next steps to achieve.

Another aspect besides improving the described modules is developing the missing links, such as disambiguating abbreviations with the help of lexical resources and the distributional methods and processing measurement results included in the texts, which contain one of the most valuable types of information.

REFERENCES

[1] Z. S. Harris, "The structure of science information," *J. of Biomedical Informatics*, vol. 35, no. 4, pp. 215–221, Aug. 2002.

[2] B. Siklósi, G. Orosz, A. Novák, and G. Prószéky, "Automatic structuring and correction suggestion system for hungarian clinical records," *8th SaLTMiL Workshop on Creation and use of basic lexical resources for lessresourced languages*, pp. 29.–34., 2012.

[3] B. Siklósi, A. Novák, and G. Prószéky, "Context-aware correction of spelling errors in hungarian medical documents," *1st International Conference on Statistical Language and Speech Processing*, 2013.

[4] A. Novák, "What is good humor like?" in *I. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: SZTE, 2003, pp. 138–144.

[5] G. Prószéky and B. Kis, "A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ser. ACL '99. Stroudsburg, PA, USA: Association for Computational Linguistics, 1999, pp. 261–268.

[6] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the ACL 2007 Demo and Poster Sessions*. Prague: Association for Computational Linguistics, 2007, pp. 177–180.

[7] G. Orosz and A. Novák, "PurePos – an open source morphological disambiguator," in *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*, B. Sharp and M. Zock, Eds., Wroclaw, 2012, pp. 53–63.

[8] E. Pirk, "Névkifejezések automatikus felismerése orvosi szovegekben," 2013.

[9] J. R. Firth, "A synopsis of linguistic theory 1930-55." vol. 1952-59, pp. 1–32, 1957.

[10] J. Carroll, R. Koeling, and S. Puri, "Lexical acquisition for clinical text mining using distributional similarity," in *Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, ser. CICLing'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 232–246.

[11] D. Lin, "Automatic retrieval and clustering of similar words," in *Proceedings of the 17th international conference on Computational linguistics - Volume 2*, ser. COLING '98. Stroudsburg, PA, USA: Association for Computational Linguistics, 1998, pp. 768–774.

[12] N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. J. Tick, "Natural language processing and the representation of clinical data," *Journal of the American Medical Informatics Association*, vol. 1, no. 2, Mar/Apr 1994.

[13] S. Meystre, G. Savova, K. Kipper-Schuler, and J. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research," *Yearb Med Inform*, vol. 35, p. 128–44, 2008.

A computational workflow for automated genome annotation and result validation

Zsolt Gelencsér
(Supervisor: Prof. Sándor Pongor)
gelzs@digitus.itk.ppke.hu

Abstract **The increasing number of available DNA sequences requires fast automated methods to process the new data and to extract as much information as possible. Genome annotation methods are designed for this task which is especially difficult because of the high diversity of biological organisms. I created an automated, subsystem based genome annotation pipeline that contains multiple, independent methods to validate the results. The pipeline is based on Hidden Markov Model search and requires comprehensive knowledge about the analyzed subsystem.**

Keywords**-component; HMM; genome annotation; validation; topology**

I. INTRODUCTION

In the past few years the speed of sequencing has greatly increased meanwhile the cost of the analysis has dramatically decreased. The number of currently available DNA sequences is over 100 million (NCBI Genbank report 2008 [1]), but the number of those well-characterized in terms of 3D structure and functions is small and their ratio is constantly decreasing. Without any added annotation, a sequence is practically is only a string without any biological meaning.

The method by which we add bonus information to a genome sequence is called genome annotation. Some annotation (e.g.: source origin) is added to the raw data in the phase of data production, but the serious part of gene annotation begins when the data are submitted to a public database. Genome annotation is based on a deep knowledge of biology and bioinformatics and relies to many databases, programs and algorithms. From the computational point of view, raw data refer to a simple string - the DNS sequence of the genome – that consists of only 4 possible characters: A, C, G and T. The theoretical topology of the genome is a linear or circular number line of positive integers; every position corresponds to a nucleotide. The bacterial genome consist one (sometimes a few) long, chromosomal sequence(s) and many short, plasmid sequence. The sequences can be linear or circular.

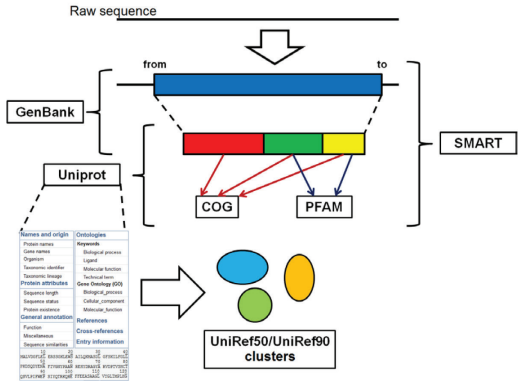


Figure 1. The steps of a genome annotation. In this figure we can see the connection betwween step of the genome annotation and bioinformatic databases.

During the genome annotation we assign attributes (in bioinformatics we called them descriptors) to the structures. There are two types of descriptors; global descriptors refer to to the whole structure (e.g.: name, function, source origin) and local descriptors refer to only a part of it (e.g.: protein segment, domains). There are many sources of the descriptors; human knowledge, computational algorithms, database cross-references …etc. There are many methods to recognize the genes of a genome but we can sort them into two main group: total genome annotation and annotation of a chosen subsystem throughout many genomes.

In the first case we choose a genome and we try to recognize the function of the unknown genes via biological study or database search. The advantage of this method is the proper knowledge of one species is enough to get information about genes but we must rely on the data of the gene-databases that can implicate incorrect annotation. (e.g.: similarity search points to the most similar function in the database that have no data on novel or undocumented biological roles.)

In the case of subsystem annotation throughout several genomes we choose a subsystem that refers to a well-defined biological process or structure in a few, well characterized set of genomes, and characterize it with a set of rules [2]. After we identify the rules that define the genes of the subsystem, we search the known genome sequences using this set of rules. The rules of the subsystem help us to validate the new genes. Even though this method doesn't give us a certain result either because there is a chance for unknown variation of the subsystem or if the subsystem has only a few genes, the identification is relatively robust against the "noise" caused by random similarities

Zs. Gelencsér, "A computational workflow for automated genome annotation and result validation,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 115-117.

II. THE STEPS OF A GENOME ANNOTATION

Now we can sketch a logical outline of subsystem-based DNA sequence annotation. A DNA sequence is considered annotated if the coding genes and other important regions are located. The protein coding genes are linked as many database as possible: primary protein databases (e.g. UNIPROT), function based cluster databases (e.g.: COG[3]) structure based cluster databases (e.g.: PFAM). To achieve this status, first we need to locate the position of the gene. If we know the exact location of the gene, we search for substructures (domains) like binding-sites, HTH structure … etc and associate them with known domain families. We add the recognized protein and the collected information to the Uniprot database. After the new protein is deposited in the database, we analyze the UniRef clusters that contain our protein. From this analysis we can collect some new information regarding our protein.

This outline leads us to two trivial conclusions: i) Database annotations change very fast, because the background databases are updated frequently; ii) Most annotations are incomplete. To reach the idealized state we need a well-organized, updateable, integrated database, but the current public databases are far from this stage.

III. THE WORKFLOW

In my previous work I created an automated protein search algorithm, based on HMM profiles [1, 2]. This method searches all members of the protein families defined by the HMM profiles in the chosen dataset, so it can be considered as part of a subsystems-based annotation workflow. The 0^{th} step is the collection of the necessary knowledge about our subsystem and the selection of the search dataset. Beside the HMM profiles we need some information about protein families. (e.g.: short name, numeric thresholds values of identification) In the first step, the program executes an hmmsearch (algorithm of the hmmer program [3]) on the chosen dataset, and collects all hits as well as their associated significances (e-values). In the next step we carry out a non-strict pre-verification of the hits; we only check the length of the hits and give a high threshold for the e-values. With this method we filter out the certainly false hits so we greatly shorten our list without losing true positives. In the next step we determine the topology of the genes which denotes their relative position in the chromosome. For this we need more data, such as position of the gene in the genome, predicted

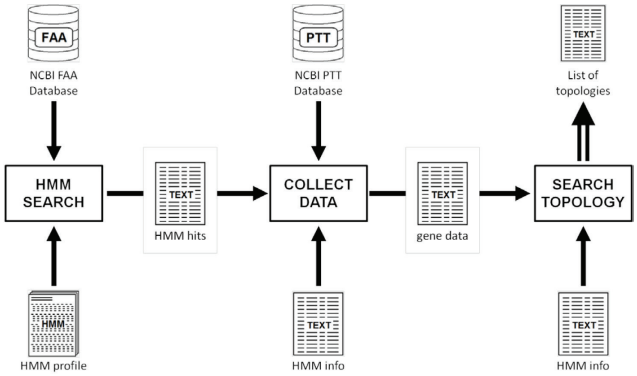


Figure 2. The diagram of the workflow. It show the flow of the data (arrows) during the genome annotation.

COG group… etc). The program collects them from the NCBI's ptt files. If all necessary data is collected, we start to analyze the genes near each other (maximum distance between them is 3000 basepairs) and determine the appeared topologies. In the topologies the position and orientation of the genes are fixed and these topologies help us to validate our result, because the knowledge about the subsystem contains the list of the probable topologies.

IV. THE VALIDATION OF THE RESULTS

The most critical part of genome annotation is the validation of the results. If we get the new information about a genome sequence via manual annotation survey, the reliability of the data is high, but these methods are very slow. The automated annotation programs are out and away faster however there is a high chance of errors because of the natural diversity of genomic sequences. If we want to accept our result we have to validate it. There are many different way to make sure our data are correct. I used the following rules to validate the result of my HMM search based annotation.

The simplest method of validation is the examination of the predicted COG values. The COG database (Clusters of Orthologous Group) is a protein database based on phylogenetic clustering [4]. Proteins included in the same group have the same biological role. If the genes of our subsystem are members of one specific COG cluster, we can easily compare that cluster with the candidate protein's COG value. If they are the same, it increases the credibility of the

TABLE I. THE VALIDATION RESULT OF THE GENOME ANNOTATION.

	genes	COG			product*			BLAST	topology
		OK	Miss	Fail	Ok	Miss	Fail		
luxR	624	595	19	10	587	19	18	621	204
luxI	269	257	6	6	239	26	4	253	206
rsaL	11	0	10	1	4	7	0	9	11
rsaM	36	0	36	0	0	36	0	33	36

This table contains the result of the 4 type of validation. The COG and product based validation can be equal with the expected, not equal or simply no information available. The columns named BLAST and topology show us the number of strengthen genes. (* manually checked)

new protein's prediction. However, a discordant COG value does not necessarily decrease or abolish the prediction's credibility; it simply means that with this method we are unable to increase its reliability. The NCBI COG dataset contains the protein names in natural language, so this database can be used only for manual validation.

Parallel with the HMM based search I execute another search based on the BLAST algorithm [4]. Both algorithms are similarity based methods, but the HMM search examines protein sequences while the BLAST algorithm analyzes the DNA sequences. The BLAST algorithm is less accurate but can eliminate the errors of the translation. If a hit of the BLAST search overlaps with the results of an HMM search, the probability of its correctness is higher.

I tested the annotation and validation method with an AHL-driven quorum sensing circuit subsystem. [5] The subsystem contains 4 types of genes: luxR, luxI, [6, 7] rsaL and rsaM [8, 9]. Table 1 shows the result of a test run performed by the bacterial section of the NCBI bacteria database. In the case of the luxR the annotation was quite efficient. Most of the COG and the product were correct (even if there were two different groups for it) and almost all hits were found also by BLAST search. Only about one third of genes were found to be in known topology, but we know that a) there are solo/orphan luxR genes [10, 11] which are not associated with luxI genes, and also, many transcriptional factors have a structure similar to luxR genes. The luxI genes show almost the same result: there were only a few genes that we could not validate with certainty. The number of "orphan" luxI genes is very low, and most probably these are associated with another type of receptor gene, which is beyond our survey. In the case of the rsaM and rsaL genes there was no proper COG clusters available, and in the current databases almost all products was nominated as "hypothetical protein", but on the other hand, every found gene figured in the known topologies which lends confidence to our findings.

V. Conclusion

We can conclude although the genome annotation seems to be a simple problem there are many difficulties that hamper a full automation of the process. The largest source of difficulty is the quality of public databases; the stored information is often incomplete and some type of data is not ordered properly. (e.g.: GenBank gene product description) The reason is that we new variant of genes continuously emerge, hence the automated genome annotation program has to be robust and prepared to identify new variants.

Using multiple validation methods we can increase the chance of correct results, but the effectiveness of these algorithms depends on the biological knowledge. We have to check the new data we put in a database, because even a little degree of corruption can lead to substantial anomalies in future analyses. We can do fast, non-reliable searches to give fast predictions, but always store a score of reliability so as to guide future analyses.

Acknowledgments

This project was developed within the PhD program of Multidisciplinary Doctoral School, Faculty of Information Technology, Pázmány Péter Catholic University (Budapest).

I thank my supervisor Prof. S. Pongor (PPKE, ICGEB, Trieste) for his help and guidance throughout the project, Kumari Sonal Choudhary, Sanjarbek Hudaiberdiev as well as Dr. Vittorio. Venturi and his group (ICGEB, Trieste) for their advice.

References

[1] S. R. Eddy, "Hidden Markov models," Curr Opin Struct Biol, vol. 6, pp. 361-5, Jun 1996.

[2] S. R. Eddy, "What is a hidden Markov model?," Nat Biotechnol, vol. 22, pp. 1315-6, Oct 2004.

[3] "HMMER hivatalos honlapja: http://hmmer.janelia.org/."

[4] R. L. Tatusov, et al., "The COG database: an updated version includes eukaryotes," BMC Bioinformatics, vol. 4, p. 41, Sep 11 2003.

[5] S. F. Altschul, et al., "Basic local alignment search tool," J Mol Biol, vol. 215, pp. 403-10, Oct 5 1990.

[6] Z. Gelencser, et al., "Classifying the topology of AHL-driven quorum sensing circuits in proteobacterial genomes," Sensors (Basel), vol. 12, pp. 5432-44, 2012.

[7] W. C. Fuqua and S. C. Winans, "A LuxR-LuxI type regulatory system activates Agrobacterium Ti plasmid conjugal transfer in the presence of a plant tumor metabolite," J Bacteriol, vol. 176, pp. 2796-806, May 1994.

[8] W. C. Fuqua, et al., "Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators," J Bacteriol, vol. 176, pp. 269-75, Jan 1994.

[9] M. Mattiuzzo, et al., "The plant pathogen Pseudomonas fuscovaginae contains two conserved quorum sensing systems involved in virulence and negatively regulated by RsaL and the novel regulator RsaM," Environ Microbiol, vol. 13, pp. 145-62, Jan 2011.

[10] V. Venturi, et al., "The virtue of temperance: built-in negative regulators of quorum sensing in Pseudomonas," Mol Microbiol, vol. 82, pp. 1060-70, Dec 2011.

[11] S. Subramoni and V. Venturi, "LuxR-family 'solos': bachelor sensors/regulators of signalling molecules," Microbiology, vol. 155, pp. 1377-85, May 2009.

[12] Y. Lequette, et al., "A distinct QscR regulon in the Pseudomonas aeruginosa quorum-sensing circuit," J Bacteriol, vol. 188, pp. 3365-70, May 2006.

[13] Y. Lequette, et al., "A distinct QscR regulon in the Pseudomonas aeruginosa quorum-sensing circuit," J Bacteriol, vol. 188, pp. 3365-70, May 2006.

Resting-State Functional Connectivity Predicts the Face Selectivity of fMRI Responses in the Fusiform Gyrus

Petra Hermann
(Supervisor: Dr. Zoltán Vidnyánszky)
hermann.petra@gmail.com

Abstract— **Face processing involves a region of the human fusiform gyrus, the fusiform face area (FFA). fMRI responses in the FFA show the highest face selectivity in the visual cortex and this region might play a primary role in coding the structural information of face stimuli. An important unresolved question is whether and to what extent the functional connectivity between the FFA and other visual cortical regions involved in object processing contributes to the face selectivity of fMRI responses in the FFA. Here we addressed this question by measuring both the face selectivity of fMRI responses and the resting-state functional connectivity between visual cortical areas in the same human participants. The results revealed that the strength of the resting-state functional connectivity between the FFA and the lateral occipital complex (LOC) involved in visual object processing showed a strong negative correlation with the face selectivity of fMRI responses in the FFA: the stronger the functional connectivity between these regions during rest, the less face selective the FFA responses. These findings suggest that face selectivity in the FFA is determined in part by its functional connectivity with non-face selective visual cortical areas of the lateral occipital cortex.**

Keywords-face selectivity; fMRI; resting-state functional connectivity; fusiform face area; lateral occipital cortex

I. INTRODUCTION

Object recognition is the ability to separate images that contain one particular object from images that do not. This is the result of the coordinated computational action of neurons / visual areas along the ventral visual processing stream. Yet, object identity can only be decoded from representations in higher level visual areas such as those found in the lateral occipital cortex and temporal cortex. Indeed, these cortical regions abound in areas claimed to be selective for certain object categories such as the extrastriate body area (EBA) in the lateral occipital cortex, most strongly responding to images of headless bodies [1], the parahippocampal place area (PPA) in the parahippocampal cortex, selectively involved in visual scene processing [2], and the fusiform face are (FFA) in the temporal cortex, with a strong selectivity to face images [3]. Human fMRI research played a major role in their identification and characterization [4], [5], as category selectivity is evident in higher fMRI responses to the preferred category compared to other categories irrespective of view-point, lighting conditions, retinal positions etc. However, the computations needed to achieve such invariance and selectivity are still under debate. Category selectivity might be the result of computations performed within these category-selective areas. Alternatively, this information could be present in earlier processing stages but be 'visible' or decodable in later stages of object processing, as suggested by the hypothesis that the ventral visual pathway gradually "untangles" information about object identity through nonlinear selectivity and invariance computations applied at each stage of the ventral pathway [6].

In our research we investigated whether and to what extent intrinsic functional connectivity within the visual object processing network contributes to the face selectivity of evoked fMRI responses in the FFA. Thus, we measured both the face selectivity of fMRI responses and the resting-state functional connectivity between visual cortical areas in the same human participants.

II. EXPERIMENTAL PROCEDURES

A. Subjects

Altogether 17 (one left-handed, ten male, mean ± SD age: 24 ± 4 years) subjects gave their informed and written consent to participate in the study, which was approved by the ethics committee of Semmelweis University. None of them had any history of neurological or psychiatric diseases, and all had normal or corrected-to-normal visual acuity.

B. Experimental Design

We used a region of interest (ROI) approach, in which we localized face- and object-related areas (Localizer runs), and then using an independent set of resting-state data calculated correlations between these predefined category-selective regions (Resting-state runs).

C. Stimuli

In the Localizer runs participants viewed images of human faces and objects and performed a one-back memory task. Face stimuli consisted of front-view grayscale photographs of four male faces with neutral, happy and fearful expressions that were cropped to eliminate the external features (hair, etc.) (see Fig. 1). In their manipulated versions noise was added to the original images by decreasing their phase coherence to 45% (55% noise) using the weighted mean phase technique [7]. In the current study, however, we will present and discuss only

This work was supported by grant from the Hungarian Scientific Research Fund to Z.V. (CNK80369).

P. Hermann, "Resting-state functional connectivity predicts the face selectivity of fMRI responses in the Fusiform Gyrus,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 119-123.

the results obtained with the 100% phase coherence face stimuli, while results obtained with the noisy faces will be presented elsewhere. Object stimuli consisted of grayscale images of three different objects from four categories (cars, mugs, jugs, and fruits) chosen from the Amsterdam Library of Objects Images (ALOI) database [8]. All images were equated for luminance and contrast.

Figure 1. The four male identities with neutral expression used in the fMRI experiments.

Stimuli were presented centrally on a uniform gray background and subtended 3×4 visual degrees. Stimulus presentation was controlled by MATLAB 7.1. (The MathWorks Inc.) using the Psychophysics Toolbox Version 3 (PTB-3) [9], [10]. Stimuli were projected onto a translucent screen located at the back of the scanner bore using a Panasonic PT-D3500E DLP projector (Matsushita Electric Industrial) at a refresh rate of 60 Hz. Stimuli were viewed through a mirror attached to the head coil at a viewing distance of 58 cm. Head motion was minimized using foam padding.

D. Procedure

For the Localizer runs, we used a standard localizer method to identify ROIs. Specifically, participants viewed two runs during which 16 s blocks (8 stimuli per block) of faces (F), noisy faces (NF), and objects (O) were presented interleaved with baseline epochs, which contained only a fixation dot. Stimuli were presented with 0.5 Hz for 500 ms each. Blocks consisted of 6 face, 6 noisy face, 6 object, and 19 baseline blocks, making a total number of 37 blocks per run. During the fMRI experimental session subjects performed a one-back task and reported the total number of one-back repetitions at the end of the run (see Fig. 2).

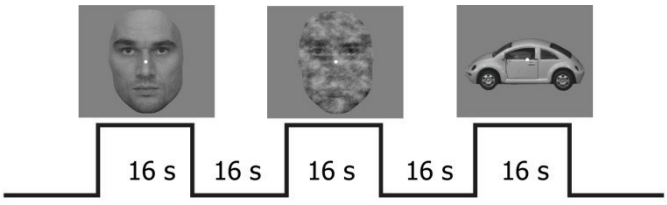


Figure 2. Experimental design in the Localizer runs. 16-s-long epochs of faces, noisy faces, and objects followed each other in random order separated by baseline blocks.

For the Resting-state run, participants were instructed to lie still, with their eyes closed during an eight-minute resting-state scan.

III. Data Analysis

A. fMRI Scanning

Data were collected at the MR Research Center of Szentágothai Knowledge Center (Semmelweis University, Budapest, Hungary) on a 3.0 tesla Philips Achieva scanner equipped with an eight-channel SENSE head coil. High-resolution anatomical images were acquired for each subject using a T1-weighted 3D TFE sequence yielding images with a 1×1×1 mm resolution. Functional images were collected using 31 transversal slices (4 mm slice thickness with 3.5 mm × 3.5 mm in-plane resolution) with a non-interleaved acquisition order covering the whole brain with a BOLD-sensitive T2*-weighted echo-planar imaging sequence (TR=2 s, TE=30 ms, FA=75°, FOV=220 mm, 64×64 image matrix, total scan time= 2×610 s and 1×480 s for Localizer and Resting-state runs, respectively).

B. Data Preprocessing and Analysis

Preprocessing and analysis of the imaging data were performed using SPM8 (Wellcome Department of Imaging Neuroscience). The functional images were realigned to the first image within a session for motion correction and then spatially smoothed using an 8-mm full-width half-maximum Gaussian filter and normalized into standard MNI-152 space. The anatomical images were coregistered to the mean functional T2* images followed by segmentation and normalization to the MNI-152 space using SPM's segmentation toolbox. The resulting gray matter mask was used to restrict statistical analysis on the functional files. To define the regressors for the general linear model analysis of the data, a reference canonical hemodynamic response function was convolved with boxcar functions, representing the onsets of the experimental conditions. Low-frequency components were excluded from the model using a high-pass filter with 128 s cutoff. Movement-related variance was accounted for by the spatial parameters resulting from the motion correction procedure. The resulting regressors were fitted to the observed functional time series within the cortical areas defined by the gray matter mask. Individual statistical maps were then transformed to the MNI-152 space using the transformation matrices generated during the normalization of the anatomical images. The resulting β weights of each current regressor served as input for the second-level whole-brain random-effects analysis, treating subjects as random factors. For visualization purposes, the F vs. O contrast (see Fig. 3) was superimposed with $p_{unc}<10^{-3}$ threshold onto the population average landmark and surface based (PALS-B12) standard brain [11] using Caret 5.62 [12]. Stereotaxic coordinates are reported in MNI space.

C. ROI Selection and Analysis

For the region of interest (ROI) analysis the face and object selective areas were defined individually based on two Localizer runs. Areas matching our anatomical criteria and lying closest to the corresponding reference cluster (i.e., clusters from the random-effects group analysis, $t_{(16)}>4.79$; $p_{unc}<10^{-4}$) were considered to be their appropriate equivalents

on the single-subject level. The location of the fusiform face area (FFA) was determined as area responding more strongly to faces than to objects ($t_{(560)}$>4.79; p_{unc}<10^{-4}). It was possible to define the right FFA (average MNI coordinates ± SD: 41±3, −50±5, −22±3) in all 17 subjects. Object-selective areas were defined as the areas in the dorsal occipito-temporal cortex (DOT) that showed significantly stronger activation ($t_{(560)}$>4.79; p_{unc}<10^{-4}) to objects than to faces. These included three distinct regions which were part of the lateral occipital complex (LOC) [13]: the inferior temporal sulcus (DOT-ITS) (46±3, −63±5, −6±3 and −46±3, −64±4, −6±3 for right and left hemispheres, respectively), the lateral occipital sulcus (DOT-LOS) (43±5, −78±6, 7±5 and −41±5, −80±5, 8±6), which were identifiable in all 17 observers, and the superior occipital sulcus (SOS), which could be defined only in 14 subjects (32±3, −75±7, 26±5 and −28±4, −76±8, 25±6). For the remaining three subjects, the group-average coordinates were taken from the random-effects group statistics (see Table 1 for coordinates). To characterize the magnitude of the signal change, t-values were estimated and averaged within a 7-mm-radius sphere around the local peak of each area of interest for each observer. We performed a one-way repeated-measures ANOVA for right FFA with condition (F vs. O) and a two-way repeated-measures ANOVA for LOC subregions with hemisphere (R vs. L) and condition (F vs. O) as within-subject factors. *Post hoc* t-tests were computed using Tukey honestly significant difference (HSD) tests.

The face selectivity of the FFA was calculated as the average of the t-scores of all voxels within a 7-mm-radius sphere around the local peak of the ROI with the F>O contrast. Thus, the larger the value, the greater the degree of selectivity.

D. *Resting-State Preprocessing and Analysis*

In addition to the aforementioned standard preprocessing (motion correction) of fMRI data, several other preprocessing steps were used to reduce spurious variance unlikely to reflect neural activity in resting-state data. These steps included using a temporal bandpass filter (0.009–0.08 Hz) to retain low-frequency signals only [14], regression of the time course obtained from rigid-body head motion correction, and regression of the mean time course of whole-brain, ventricle, and white matter BOLD fluctuations [15].

After the preprocessing, a continuous time course for each ROI was extracted by averaging the time courses of all voxels in each of the ROIs. Thus, we obtained a time course consisting of 240 data points for each ROI and for each participant. Temporal correlation coefficients between the extracted time course from the right FFA and those from other ROIs (DOT-ITS, DOT-LOS, SOS) located in the right LOC were calculated to determine the extent to which regions were functionally correlated at rest. Relationship between resting-state functional connectivity coefficients (rsFC strength) and individual face-selective fMRI responses was studied by computing between subject correlations. To correct for multiple comparisons (c = 3), significance threshold was set to p_{Bonf} = 0.05 (p_{unc} = 0.017).

IV. RESULTS

A. *Results of the Random-Effects Group Analysis*

Whole-brain random-effects analysis of the fMRI data revealed that face stimuli elicited significantly higher fMRI responses in the FFA compared to object stimuli, while areas in the dorsal occipito-temporal cortex (DOT-ITS, DOT-LOS, and SOS) showed larger responses to objects than to faces ($t_{(16)}$>4.79; p_{unc}<10^{-4}) (see Fig. 3 and Table 1 for more details).

TABLE 1. SIGNIFICANT FMRI CLUSTERS

MNI Coordinates	$t_{(16)}$ Value	Area Label
-42, -68, -2	10.82	Left DOT-ITS
-42, -80, 16	8.11	Left DOT-LOS
42, -76, 10	7.76	Right DOT-LOS
42, -52, -20	6.19	Right FFA
-26, -70, 30	5.82	Left SOS
46, -62, -2	5.24	Right DOT-ITS
34, -78, 20	4.99	Right SOS

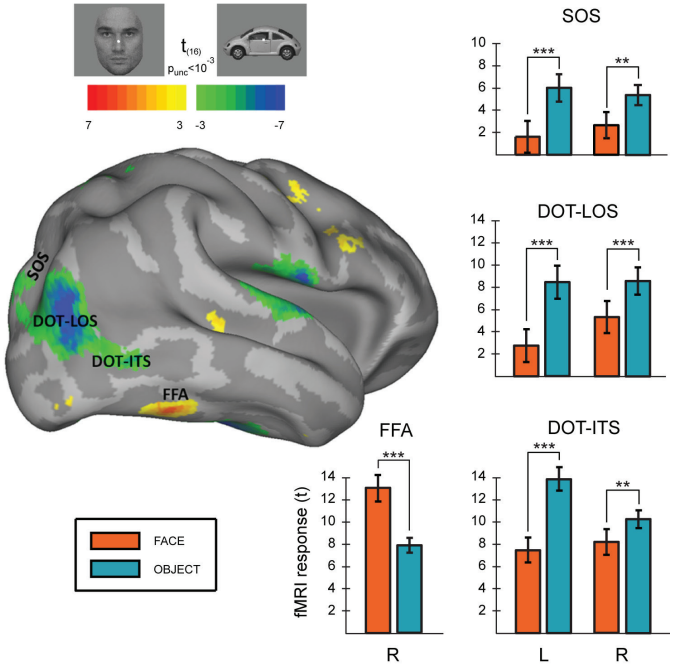


Figure 3. Group-wise (random effects) statistical parametric map of activations to faces vs. objects and region-of-interest (ROI) analysis of fMRI responses to the two different stimulus types from face and object-specific ROIs. Significantly higher fMRI responses were found to faces compared to objects in the right fusiform cortex (FFA), while larger responses were observed to objects than to faces in areas of the dorsal occipitotemporal cortex (DOT-ITS, DOT-LOS and SOS). Maps are displayed with p_{unc}<10^{-3} on the PALS-B12 partially inflated brain [11] (***p<10^{-3}, **p<10^{-2}, *p<5×10^{-2}; R, right; L, left).

B. Results of the ROI Analysis

As previous individual brain analyses highlighted the large amount of interindividual variability in the location of the high-level, shape-specific visual cortical areas, we also performed an ROI-based analysis of the fMRI data in the object-selective (DOT-LOS, DOT-ITS, and SOS) and face-selective (FFA) regions in the visual cortex (for ROI definition, see Section III). In agreement with the results of the whole-brain random-effects analysis, the ROI analysis (Fig. 3) suggested that face stimuli are processed selectively in the right FFA, while object stimuli in bilateral LOC including DOT-ITS, DOT-LOS, and SOS. We found significantly higher fMRI responses to faces relative to objects in the right fusiform cortex (FFA) (main effect of condition: $F_{(1,16)}=48.26$; $p<10^{-5}$). This appears to be in agreement with the large amount of previous results showing lateralization of neural processes associated with face processing to the right hemisphere [3], [16], [17]. In the case of object as compared to face stimuli bilateral LOC showed increased activation (main effect of condition: $F_{(1,16)}=60.37$; $p<10^{-6}$, $F_{(1,16)}=63.90$; $p<10^{-7}$, and $F_{(1,16)}=32.93$; $p<10^{-4}$ for DOT-ITS, DOT-LOS, and SOS, respectively). In addition we also found a hemispheric asymmetry in object processing: the selectivity was more pronounced in the left hemisphere (condition×side interaction: $F_{(1,16)} = 41.35$; $p = <10^{-5}$, $F_{(1,16)} = 14.88$; $p <10^{-3}$, and $F_{(1,16)} = 15.55$; $p = <10^{-3}$ in the cases of DOT-ITS, DOT-LOS, and SOS, respectively).

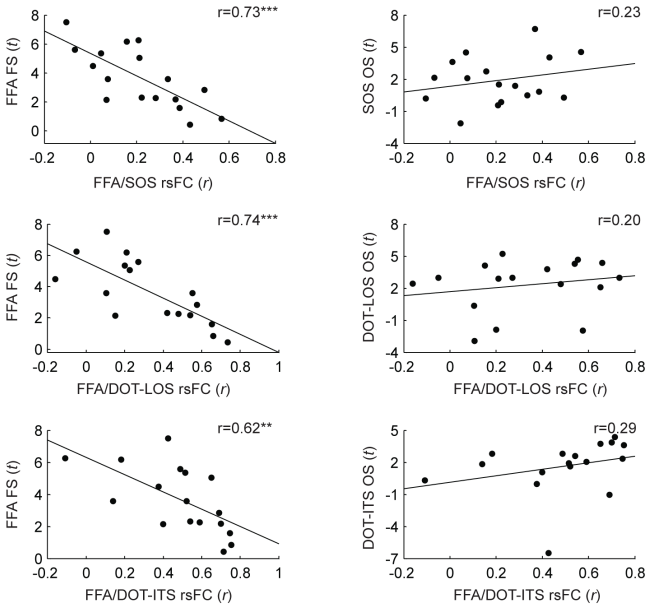


Figure 4. Correlations between FFA/LOC functional connectivity strengths (rsFC) and category selectivity of fMRI responses in regions FFA (left panels) and LOC (right panels). rsFC strength between the FFA and object-related areas in the lateral occipital complex (DOT-ITS, DOT-LOS, SOS) correlated negatively with the face selectivity (FS) of fMRI responses in the FFA. However, object selectivity (OS) of the fMRI responses in the LOC areas did not correlate with the resting-state functional connectivity of these areas with the FFA (***$p<10^{-3}$, **$p<10^{-2}$, *$p<5\times10^{-2}$).

C. Results of the rsFC and Correlation Analysis

A one-sample t-test of rsFC between the right FFA and LOC subregions showed that these regions were functionally connected at rest (FFA/DOT-ITS rsFC: $t_{(16)}=8.45$; $p=3\times10^{-7}$, FFA/DOT-LOS rsFC: $t_{(16)}=5.17$; $p=9\times10^{-5}$, and FFA/SOS rsFC: $t_{(16)}=4.55$; $p=3\times10^{-4}$).

The Pearson correlation analysis between connection strengths and face selectivity (determined by the F>O contrast during the localizer task) revealed that the rsFC strength between the FFA and object-related areas in the lateral occipital complex (DOT-ITS, DOT-LOS, SOS) correlated negatively with the face selectivity of fMRI responses in the FFA (r=−0.62; $p=8\times10^{-3}$, r=−0.74; $p=7\times10^{-4}$, r=−0.73; $p=8\times10^{-4}$ for FFA/DOT-ITS, FFA/DOT-LOS, and FFA/SOS rsFC, respectively). Thus, the stronger the functional connectivity between these regions during rest, the less face selective the FFA responses. However, it was also found that the object selectivity (determined by the O>F contrast during the localizer task) of the LOC was not modulated by the rsFC with the FFA (r=0.29; p=0.26, r=0.20; p=0.45, r=0.23; p=0.37 for FFA/DOT-ITS, FFA/DOT-LOS, and FFA/SOS rsFC, respectively), indicating the feedforward direction of the category-selective information flow (Fig. 4).

V. Discussion and Conclusions

We have found that the strength of the resting-state functional connectivity between the FFA and multiple subregions of the lateral occipital complex (LOC) involved in visual object processing showed a strong negative correlation with the face selectivity of fMRI responses in the FFA: the stronger the functional connectivity between these regions during rest, the less face selective the FFA responses.

These findings suggest that face selectivity in the FFA is determined at least in part by its functional connectivity with non-face selective visual cortical areas located upstream in the lateral occipital cortex.

This view is in agreement with the previous results showing that human visual system is able to perform object recognition at an incredible speed: the central visual image is processed to support recognition in 150-200 ms, which is termed ‘core object recognition’ [18]. Thus, the mere speed of core object recognition necessitates category- selective information to be present earlier in the processing stream.

Acknowledgment

This work was supported by grant from the Hungarian Scientific Research Fund to Z.V. (CNK80369).

References

[1] P. E. Downing, Y. Jiang, M. Shuman, and N. Kanwisher, “A cortical area selective for visual processing of the human body,” *Science*, vol. 293, no. 5539, pp. 2470–2473, Sep. 2001.

[2] R. Epstein and N. Kanwisher, “A cortical representation of the local visual environment,” *Nature*, vol. 392, no. 6676, pp. 598–601, Apr. 1998.

[3] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," *J. Neurosci.*, vol. 17, no. 11, pp. 4302–4311, Jun. 1997.
[4] U. Hasson, M. Harel, I. Levy, and R. Malach, "Large-scale mirror-symmetry organization of human occipito-temporal object areas," *Neuron*, vol. 37, no. 6, pp. 1027–1041, Mar. 2003.
[5] K. Grill-Spector and R. Malach, "The human visual cortex," *Annu. Rev. Neurosci.*, vol. 27, pp. 649–677, 2004.
[6] J. J. DiCarlo and D. D. Cox, "Untangling invariant object recognition," *Trends Cogn. Sci. (Regul. Ed.)*, vol. 11, no. 8, pp. 333–341, Aug. 2007.
[7] S. C. Dakin, R. F. Hess, T. Ledgeway, and R. L. Achtman, "What causes non-monotonic tuning of fMRI response to noisy images?," *Curr. Biol.*, vol. 12, no. 14, pp. R476–477; author reply R478, Jul. 2002.
[8] J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The Amsterdam Library of Object Images," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, Jan. 2005.
[9] D. H. Brainard, "The Psychophysics Toolbox," *Spatial Vision*, vol. 10, no. 4, pp. 433–436, 1997.
[10] D. G. Pelli, "The VideoToolbox software for visual psychophysics: transforming numbers into movies," *Spat Vis*, vol. 10, no. 4, pp. 437–442, 1997.
[11] D. C. Van Essen, "A Population-Average, Landmark- and Surface-based (PALS) atlas of human cerebral cortex," *Neuroimage*, vol. 28, no. 3, pp. 635–662, Nov. 2005.
[12] D. C. Van Essen, H. A. Drury, J. Dickson, J. Harwell, D. Hanlon, and C. H. Anderson, "An integrated software suite for surface-based analyses of cerebral cortex," *J Am Med Inform Assoc*, vol. 8, no. 5, pp. 443–459, Oct. 2001.
[13] R. Malach, J. B. Reppas, R. R. Benson, K. K. Kwong, H. Jiang, W. A. Kennedy, P. J. Ledden, T. J. Brady, B. R. Rosen, and R. B. Tootell, "Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 92, no. 18, pp. 8135–8139, Aug. 1995.
[14] D. Cordes, V. M. Haughton, K. Arfanakis, J. D. Carew, P. A. Turski, C. H. Moritz, M. A. Quigley, and M. E. Meyerand, "Frequencies contributing to functional connectivity in the cerebral cortex in 'resting-state' data," *AJNR Am J Neuroradiol*, vol. 22, no. 7, pp. 1326–1333, Aug. 2001.
[15] A. Weissenbacher, C. Kasess, F. Gerstl, R. Lanzenberger, E. Moser, and C. Windischberger, "Correlations and anticorrelations in resting-state functional connectivity MRI: a quantitative comparison of preprocessing strategies," *Neuroimage*, vol. 47, no. 4, pp. 1408–1416, Oct. 2009.
[16] A. Puce, T. Allison, J. C. Gore, and G. McCarthy, "Face-sensitive regions in human extrastriate cortex studied by functional MRI," *J. Neurophysiol.*, vol. 74, no. 3, pp. 1192–1199, Sep. 1995.
[17] S. M. Letourneau and T. V. Mitchell, "Behavioral and ERP measures of holistic face processing in a composite task," *Brain Cogn*, vol. 67, no. 2, pp. 234–245, Jul. 2008.
[18] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, no. 3, pp. 415–434, Feb. 2012.

Data Locality Improvement for Mesh Computations

Antal Hiba
(Supervisors: Péter Szolgay, Miklós Ruszinkó)
hiban@digitus.itk.ppke.hu

***Abstract*—Nowadays many-core architectures GPUs and FPGAs have very high theoretical coputational power. In case of many applications, the utilization of processing resources is poor due to irregular memory access. Wide range of simulation tasks (sound, heat, elecrodynamics, fluid dynamics) leads to computations on irregular meshes. The mesh and an ordering on its nodes, defines graph-bandwidth, which is a good indicator of data locality. In this paper the advantages of locality improvement are descussed. The concept of data locality improvement in mesh partitioning is shown, with the results of exsisting solution techniques.**

I. Introduction

Nowadays many-core architectures GPUs and FPGAs have hundreds of processing elements (PEs), which leads to high theoretical computational power (TeraFLOPS/chip). However the utilization of PEs is low in many applications, because these architectures are very sensitive to irregular memory access patterns.

First of all there is not enough theoretical memory bandwidth to feed all PEs simultaneously, from off-chip memory. For utilizing all processing elements, loaded data must be reused several times from on-chip cache. Furthermore the theoretical memory bandwidth can be reached only by sequential bursts (multiple data transfers together), and required data have to fit to the provided access granularity (64 - 256 bit). In case of random 32 bit reads, the real memory bandwidth can be many times lower than the theoretical maximum.

Irregular memory access leads to poor memory bandwidth utilization, and high cashe miss rate, which are the sources of low PE utilization. Preoptimization of input data can increases the regularity of the access pattern. FPGAs are most practicle for the scientific investigation of this problem, because memory access can be fully determined by the designer, including memory interface and the caching mechanism (minimal black-box effect).

Nearly the theoretical bandwidth of the off-chip DRAM can be utilized by moving data in long sequential bursts between the off-chip memory and the PEs in the FPGA. However, optimized input data is necessary, where all dependent data are inside an index-range (in main memory), which can be stored on-chip. If the dependencies are described by a mesh, the result of the optimization is an ordering of nodes, where this index-range is minimized. The maximal difference between the indexes of adjacent nodes is called graph bandwidth (G_BW). Let $BW = 2 * G\_BW + 1$. If the FPGA has $BW * NodeSize$ on-chip memory, every node data needs to be loaded only once, thus the whole theoretical memory bandwidth is utilized, with maximal on-chip data reuse.

Graph Bandwidth Minimization is similar to a well-studied optimization problem, called Matrix Bandwidth Minimization, where the matrix is the adjacency matrix of the mesh. One of the most practical heuristic solutions is GPS(Gibbs, Pole and Stockmeyer) [2], which is fast enough to handle graphs with many million nodes effectively. If the reordered input has grater on-chip memory requirement, than the available resources, or more FPGAs perform the computation, the input mesh must be divided into parts.

Famous partitioning methods, for instance METIS [6], minimize the edge-cut between the parts, and balance the size of the generated parts. The size-balance is important because each part is given to a multi-processor, and the overall runtime is determined by the processor which get the largest part. The edge-cut is proportional to the communication required between the processors. Graph bandwidth of the resulting parts is often smaller than the graph bandwidth of the whole mesh, but the methods do not deal with the graph bandwidth directly. The graph bandwidth of the resulting parts is important, because it determines the minimal size of on-chip memory, which is necessary for maximal data reuse. The edge-cut is also relevant, because it is proportional to the number of random accesses, which appear, when the PE reads data from adjacent parts (ghost nodes). In many cases the boundary surface (set of extremal nodes) of the mesh is also known, which gives information about the geometry, but not used by traditional partitioners.

A novel approach is shown in [5], where boundary (covering) surface is used to detect critical areas of the mesh, and performs a bisection, which decreases the G_BW of the parts. The proposed method has some weak points, but shows the possibility and tools of direct G_BW handling. In [5] a challenging partitioning problem is also presented, which is called Bandwidth-Limited Partitioning.

In this paper the memory bandwidth limitation of different processor architectures is discussed, with some possible solutions. One of them is the preoptimization of input data, which means G_BW reduction in case of mesh computations. The concept of Bandwidth-Limited Partitioning is shown, with the results of exsisting solution techniques.

II. Graph Partitioning

The k-way partitioning problem is the following: Given a graph G(V,E), with vertex set V ($|V| = n$) and edge set E. A partition Q is required, where $Q = \{P_1, P_2, .., P_k\}$, $\bigcup_{i=1}^{k} P_i = V$, $P_i \bigcap P_j = 0$ for $i \neq j$. The subsets have equal size $|P_i| = n/k$, and the number of edges between vertices

A. Hiba, "Data locality improvement for mesh computations,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 125-128.

belongs to different P_i subsets (edgecut) is minimized.
Size balance of P_i subsets provides balanced workload for all processors, and the minimized edgecut minimizes the communication between processors. This objective function can leads to poor real speedup, because the real topology of processors is not taken into consideration [3, 9]. The above objective is the base of all improved partitioning models, and solution techniques of this problem are also building blocks of novel methods.

A. Generalizations of Graph Partitioning

Graph partitioning problem in its original form can not handles many important factors. Generalized partitioning models have been created to stisfiy these needs.

1) Hybrid Architecture: If the processor nodes have different computational capabilities, the workload have to be distributed according to processing powers, thus $|P_i| = pow_i \cdot n$, where pow_i is the normalized computational capability of processor i.

2) Heterogenous Processes: Processes in V can have different computational complexity, a weight function $w_v(v_i)$ can contains this information. In this case the workload of a processor is the sum of weights inside the corresponding subdomain. Real communication needs can be modeled by a weight function on the edges $w_e(e_{ij})$.

3) Multi-Constraint Partitioning: Multiple balancing constraints can be modeled by using weight vectors instead of simple weights. For instance, a weight can be defined for the computation need, and an other for the memory need [7].

4) Skewed Partitioning Model: The model can be improved with adding some penalty functions (skew) to the cost function. Let $p(v_i)$ the set to which vertex v_i is assigned, and $d_{P_k}(v_i)$ is the desire of vertex v_i to be in P_k. The cost function is the following:

$$Min \sum_{e_{ij}} \begin{cases} w_e(e_{ij}) & \text{if p(i)} \neq \text{p(j)} \\ 0 & \text{otherwise} \end{cases} - \sum_{v_i} d_{p(v_i)}(v_i)$$

Desire functions can be used to hold additional knowledge about good solutions [4].

5) Target Graph Reprezentation: Target or architecture graph representation, gives an opportunity to model real communication costs [8]. In this case G is denoted by S as source graph, and target graph T has physical processors as its vertices $V(T)$ and real communication links as its edge set $E(T)$. Both graph has weight functions defined on their vertices $w_{S_v}(v_k)$, $w_{T_v}(v_l)$ and edges $w_{S_e}(e_{ij})$, $w_{T_e}(e_{kl})$. Two functions are required: $\tau_{S,T} : V(S) \rightarrow V(T)$ and $\rho_{S,T} : E(S) \rightarrow \mathcal{P}(E(T))$, where $\mathcal{P}(E(T))$ denotes the set of all simple loopless paths which can be built from $E(T)$. Data exchanges between not adjacent processors, require transmissions through a path from one to the other, which results additional cost. In communication cost function f_C every communication weight is multiplied by the length of its route:

$$f_C(\tau_{S,T}, \rho_{S,T}) = \sum_{e_{ij} \in E(S)} w_{S_e}(e_{ij}) |\rho_{S,T}(e_{ij})|$$

B. Sparse Matrix Reordering

Most of the partitioning tools have built in reordering methods. For G an adjacency matrix A can be given, where $a_{ij} = 1$ if $e_{ij} \in E(G)$ $a_{ij} = 0$ otherwise. Matrix reordering changes the permutation of nodes in A matrix. $A' = PAP^T$, where P is a permutation matrix.
Large $Ax = b$ linear systems, have great importance in many applications. Efficient solvers use the Cholesky factorization $A = L^T L$. The goal of many matrix reordering methods is to minimize the number of nonzero elements in the Cholesky factor L. Other class of reordering methods transforms A to a band matrix.
Mapping assigns a physical processor to each process, but the schedule of processes which belong to the same processor is still not defined. The local memory placement of data structures are also not defined. With reordering techniques, the schedule of processes and the memory placement of data structures can be determined. If the physical processor can run one thread, the memory is random access (random reads takes same time as sequential reads), and the processor has no cashe, these questions are pointless, because the answers have no effect on the execution time. However novel processor architectures, and fastest memory interfaces have extremely different behavior.

III. Memory Bandwidth limitations

A. Processor Architectures and Memory Interfaces

Theoretical computational power of processor architectures is increasing, the theoretical memory bandwidth of memory interfaces is also increasing, however there are some problems in the background, which have to be investigated.
Comparison of different processor architectures is shown in Table I. Intel core i7-2600K is a common desktop CPU with 4 cores, where each core can performs 8 floating-point operations (FLOP) per cycle, which results 108.8 GigaFLOP per second (GFLOPS) computational power. Intel Xeon E7-2860 is used as server CPU, it has 10 cores with decreased operating frequency, and each core provides 4 FLOP per cycle, but it has three times more on-chip memory (24 MB), and better memory interface. BlueGene/Q is the state of the art CPU architecture, which is the building block of power-efficient supercomputing systems. BG/Q has 16 cores with 4-way multithread, which results 204 GFLOPS at only 55 Watt. Geforce GTX 680 represents the family of GPUs. GTX 680 has 1536 cuda cores operating at 1 GHz, so the theoretical computational power is 1536 GFLOPS. Cuda cores do multiply-accumulate (MAC) operations, which is 1 FLOP in our view. The last important class of computing chips is the family of FPGAs. Virtex XC7VX850T is one of the most powerful FPGA in case of floatig-point multiplications, with 3960 DSP slices. The balanced comparison of an FPGA to other processor architectures is very challenging. DSP slices of Vertex 7 FPGAs perform 24*18 bit fix-point MAC, and every FPGA design is a processor architecture, which has its own computational capabilities. Xilinx showed FP32 power

of XC7VX850T in [1], where a 16*16 matrix multiplier design was shown, which has 1145 GFLOPS theoretical computational power. This example design is a lower bound on real theoretical maximum computational power of the FPGA. An upper bound can be given by assuming all DSP slice perform FP32 multiplications, which means 2526 GFLOPS.

TABLE I
BANDWIDTH LIMITATIONS OF DIFFERENT ARCHITECTURES

Chip (cores/threads)	Bandwidth GB/s	Memory Type	L2-L3 cache MB
core i7-2600K 3.4GHz (4/8)	21	DDR3 2*1333	8
Xeon E7-2860 2.26 GHz (10/20)	33	DDR3 4*1066	24
BlueGene/Q 1.6 GHz (16/64)	42	DDR3 1333	32
GTX 680 (1536 cuda cores)	192	GDDR5	0.5
XC7VX850T (3960 DSP slices)	41.65	DDR3 4*1333	62

Chip (cores/threads)	GFLOPS	GFLOPS*	GFLOPS/GFLOPS*
core i7-2600K 3.4GHz (4/8)	108.8	2.62	41.53
Xeon E7-2860 2.26 GHz (10/20)	90.6	4.16	21.78
BlueGene/Q 1.6 GHz (16/64)	204	5.25	38.85
GTX 680 (1536 cuda cores)	1536	24	64
XC7VX850T (3960 DSP slices)	1145** - 2526***	5.20	220.19 - 485.76

FLOP: FP32 Multiplication or MAC
GFLOPS* : when 1FLOP needs 2*4 byte input from main memory (zero cache)
** : synthetized 16x16 FP32 matrix multiplier [1]
***: 3960 DSP @ 638 MHz 25*18 bit multiplications

GPUs and FPGAs have more than 1 TeraFLOPS theoretical computational power per chip, however the available off-chip memory bandwidth (21-192 GB/s) can support input only for 2-24 GFLOPS. The difference between zero-cache GFLOPS* and the theoretical maximum GFLOPS is 64 times for the GPU and more than 100 times for the FPGA, and 21-41 times for the CPUs. It means that the input data have to be reused 20-100 times from on-chip memory to reach 100% utilization of PEs. Memory bandwidth limitation (memory wall) is the reason, why on-chip cache plays important role in many-core processor architectures.
Current DRAM technolgys are DDR3 and GDDR5. These memories are not fully random access, becuse both of them use 8n prefetch, which increases the theoretical memory bandwidth, but also increases the minimal amount of data per transmission. In case of DDR3 the access granularity is 64-128 bit and 256 bit for GDDR5.

B. Possible Solutions

Utilization of processing elements can be increased through many ways.

1) Better Memory Interface: Higher theoretical memory bandwidth is not enough, the access granularity, and latancies between random accesses are also important.

2) Decreased Operating Frequency: Computational power linearly depends on the operating frequency, but the power consumption of a processor chip has quadratic frequency dependence. The main indicator of Green Computing GFLOPS/Watt becomes better if frequency is decreased.

3) Increased On-Chip Memory: The rate of on-chip data reuse can be improved with incresed on-chip memory, which can leads to better PE utilization. However on-chip cache has some side effects, one of them is the cache coherence problem. The caching mechanism is as important as the size of on-chip memory. Increased on-chip memory needs more chip area, thus less PE can be put on the same chip, which results less GFLOPS.

4) Preoptimized Input and Algorithms: Kiloprocessor architectures brought changes in algorithmic design. The efficient mapping of memory to PEs and the wise usage of on-chip memory resources are critical. Preoptimization of input data results an optimized placement of data in main memory. This way is the main topic of this paper. Novel partitioning models are needed, which deal with the architectural parameters of kiloprocessor chips, and their memory interfaces.

IV. BANDWIDTH-LIMITED PARTITIONING

The main goal of partitioning is to give a distribution of computation and data among physical processor nodes, which leads to minimal computation time. Load balance and edge-cut are useful indicators, but more aspects have to be considered in partitioning techniques.

A. BLP definition

Given a graph $G(V, E)$ and MAX_k number of processor nodes which have C cash size in bytes and $COMM\_R$ communication ratio, where $COMM\_R$ equals interprocessor communication bandwidth over memory bandwidth. Partition $Q = \{P_1, P_2, .., P_k\}$ $k \leq MAX\_k$ is required, with the following properties:

- $\bigcup_{i=1}^{k} P_i = V$, $P_i \bigcap P_j = 0$ for $i \neq j$.
- labeling f exists: $(2 \cdot B_f(P_i) + 1) \cdot nodesize \leq C$
- # outgoing edges over # inner nodes $\leq COMM\_R$
- k is maximized
- subsets have equal size $|P_i| = n/k$

BLP has multiple objectives which have to be traded off. It is also possible, that the constraints on bandwidth and communication ratio can not be satisfied at the same time. In this case the goal is the minimization of difference from the given bounds.

V. RESULTS

A. Extended Ordering Method

An extended ordering method (AM1) can be used for the creation of a bandwidth-limited partition. GPS ordering method chooses a starting node, and index nodes according to a breadth-first search tree, which root is that node. Adding a bandwidth estimation to this algorithm, indexing can be stopped before the bandwidth limit is reached. Indexed nodes form a part, and the process started again on the rest node set. The solution quality can be measured by the node reload factor **k**, (k-1)*100% of the nodes have to be reloaded from the external memory, this is also the ratio of random accesses.

Measurements on three meshes with different BW bounds can be found on Table II. The results shows that the BW value of large meshes can be reduced more effectively(with better k factors), with BW_Bound=0.5*AM1_BW, we get k={1.65

TABLE II
RESULTS OF AM1 BOUNDED BANDWIDTH OPTIMIZATION

Case	AM1_BW	BW Bound	num. of parts	N	overall length	k
3d_075	411	412	1	3562	3562	1
3d_075	411	400	3	3562	4489	1.26
3d_075	411	300	8	3562	5241	1.471
3d_075	411	200	15	3562	5899	1.656
3d_035	1893	1894	1	33730	33730	1
3d_035	1893	1800	3	33730	37952	1.125
3d_035	1893	1500	5	33730	39987	1.185
3d_035	1893	1000	15	33730	44439	1.317
3d_015	14985	14986	1	417573	417573	1
3d_015	14985	14000	2	417573	430693	1.031
3d_015	14985	10000	3	417573	439136	1.052
3d_015	14985	7500	7	417573	452510	1.084
3d_015	14985	5000	21	417573	483391	1.158
3d_015	14985	2500	97	417573	577474	1.383

AM1_BW: the bandwidth provided by AM1 for the whole mesh
overall length: length of the generated access pattern
N: number of vertices

1.31 1.08}. BW_Bound is determined by the on the on-chip memory capabilities of the FPGA, which is increasing with every new generation of the technology, furthermore the ratio becomes better for larger problems.

B. DLS-Based Bisection

DLS-Based Bisection is introduced in [5], which uses the additional information of the boundary node set, and uses waves to define vertex separators. A comparison is shown in Table III between the DLS-Based and the METIS-recursive partitioning. For the structured brick-shaped problems, the DLS method provides average 28% better BW partition, with acceptable communication ratio (external reads / inner). The COMM ratio is better for larger problems, in case of sgrid4, which has 1M vertices, the COMM ratio is only 0.0164.

TABLE III
RESULTS OF DLS BASED PARTITION

Problem	N	Orig BW	MET BW	MET COMM	DLS BW	DLS COMM
sgrid1	2200	221	181	0,0388	141	0.1216
sgrid2	16800	841	641	0,0192	479	0.0628
sgrid3	131200	3281	2517	0,0087	1719	0.0323
sgrid4	1036800	12961	9967	0,0045	6599	0.0164
snake100	7821	777	689	0,0254	531	0,079
snake038	158544	5701	5371	0,0074	4941	0.0095
tunnel202	18210	2353	1385	0,0292	1303	0.0262
tunnel100	191592	12525	5675	0,017	5949	0.027
weight045	4899	641	581	0,0169	311	0.1764
weight022	35922	2363	2131	0,0078	1411	0.0559
weight012	230891	8087	8785	0,0037	8785	0.0075

N: number of vertices. Orig BW: GPS bandwidth for the whole mesh.
MET/DLS BW: bandwidth of partitions.
MET/DLS COMM: number of outgoing edges / number of internal edges

The communication ratio (edge-cut ratio) is getting better when the mesh density is increased for all problem instances. This is obvious because the cutting surface has N-1 dimension in case of an N dimensional mesh. This feature is important, because DLS computes a kind of N-1 dimensional surface, which separates the mesh into two parts. DLS-Based solutions can have unacceptable communication need for small meshes, for example weight045, where the COMM ratio is 0.17.
Using DLS-Based bisection the resulting partitions have 40% reduced bandwidth compared to the whole mesh, and create 20% better solutions than METIS. METIS minimizes the edge-cut with providing size-balance, the DLS-Based solutions have same size balance quality, however the edge-cut is several times higher. There is a tradeoff between bandwidth and communication need, and DLS creates partitions with higher COMM ratio to provide reduced bandwidth.

VI. CONCLUSIONS

Kiloprocessor architectures are memory bandwidth limited, heavy on-chip data reuse is necessary to provide input for PEs. Furthermore the memory interface is sensitive to small random accesses, sequential memory access pattern is needed for maximal utilization of the off-chip memory bandwidth.
In this paper a novel partitioning problem is presented, which provides maximal memory-bandwidth and on-chip memory utilization. An extended ordering method (AM1) demonstrates that bandwith limited partitions can be created with optional bandwidth limits. With DLS-Based Bisection, bandwith need can be reduced by 40%, where the bound on communication ratio is satisfied.

REFERENCES

[1] Oliver Garreau and Jack Lo. Scaling up to teraflops performance with the virtex-7 family and high-level synthesis, xilinx wp387 (v1.0). 2011.

[2] Norman E Gibbs, William G Poole Jr, and Paul K Stockmeyer. An algorithm for reducing the bandwidth and profile of a sparse matrix. *SIAM Journal on Numerical Analysis*, 13(2):236–250, 1976.

[3] S.W. Hammond. *Mapping unstructured grid computations to massively parallel computers.* PhD thesis, Rensselaer Polytechnic Institute, Troy, New-York, February 1992.

[4] B. Hendrickson, R. Leland, and R. Van Driessche. Skewed graph partitioning. In *Proceedings of the 8th SIAM Conference on Parallel Processing for Scientific Computing*, 1997.

[5] Antal Hiba, Zoltan Nagy, and Miklos Ruszinko. Memory access optimization for computations on unstructured meshes. In *Proc. 13th International Workshop on Cellular Nanoscale Networks and their Applications*, 2012.

[6] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392, 1998.

[7] George Karypis and Vipin Kumar. Multilevel algorithms for multi-constraint graph partitioning. In *Proceedings of the 1998 ACM/IEEE conference on Supercomputing (CDROM)*, pages 1–13. IEEE Computer Society, 1998.

[8] François Pellegrini. Graph partitioning based methods and tools for scientific computing. *Parallel computing*, 23(1):153–164, 1997.

[9] C. Walshaw, M. Cross, G. Everett, et al. Partitioning and mapping of unstructured meshes to parallel machine topologies. In *Proc. Irregular95, number 980 in LNCS*, pages 121–126, 1995.

Efficient GPU implementation of Lattice-Reduction

Csaba M. Józsa
(Supervisors: Antonio M. Vidal, Gema Piñero, Géza Kolumbán, Péter Szolgay)
jozsa.csaba@itk.ppke.hu

***Abstract*—In this work a brief extract of the paper of Józsa *et al.* [1] is presented, focusing on lattice reduction algorithms and their parallel implementations on GP-GPU architecture. Existing parallel Lattice Reduction (LR) implementations primarily are exploiting the parallel capabilities of multi-core architectures, however significant speed-ups can be achieved by modifying the existing algorithms and mapping them on the GP-GPU architecture.**

***Keywords*-Lattice Reduction, GP-GPU, multi-level parallelism**

I. Introduction

The aim of the LR is to find an orthogonal basis for a given lattice, which from a computational complexity point of view is an NP-hard problem. Therefore several polynomial time LR algorithms have been proposed in the literature, being the most popular the Lenstra-Lenstra-Lóvasz (LLL) algorithm [2]. Different parallel implementations of the LLL algorithm can be found in [3, 4], whereas [5] proposes a very interesting approach to limit the LLL complexity without reducing its performance. Recently, the paper of Wubben *et al.* [6] gives an in-depth review of LR algorithms, analyzing their capacity to achieve a good orthogonality, and evaluating their performance for precoding and detection in wireless applications. In this work the results of the paper [1] are presented, namely the parallelization of the LLL algorithm is presented taking advantage of high performance many-core architectures resulting in a significant decrease of the execution time.

II. Lattice Reduction

A real-valued lattice *L* is a discrete additive subgroup of $\mathbb{R}^n$ defined as $L = \left\{\sum_{i=1}^{n} x_i \cdot \underline{b}_i \;\middle|\; x_i \in \mathbb{Z}, i = 1, \cdots, n\right\}$, where $\underline{b}_1, \underline{b}_2, \cdots, \underline{b}_n \in \mathbb{R}^n$ are linearly independent vectors and $\mathbb{Z}$ denotes the set of integers. Let matrix $\mathbf{B} = (\underline{b}_1, \ldots, \underline{b}_n) \in \mathbb{R}^{n\times n}$ denote the full (column) ranked basis of the lattice. Let $\mathbf{B}^* = (\underline{b}_1^*, \ldots, \underline{b}_n^*) \in \mathbb{R}^{n\times n}$ denote the associated orthogonal basis of $\mathbf{B}$, calculated by the Gram-Schmidt orthogonalization process as follows: $\underline{b}_1^* = \underline{b}_1$, $\underline{b}_i^* = \underline{b}_i - \sum_{j=1}^{i-1} \mu_{i,j}\underline{b}_j$ for $2 \leq i \leq n$, where $\mu_{i,j} = (\underline{b}_i, \underline{b}_j^*)/(\underline{b}_j^*, \underline{b}_j^*)$ for $1 \leq j < i \leq n$, also called the Gram-Schmidt coefficients (GSC), where $(,)$ denotes the ordinary inner product on $\mathbb{R}^n$. With $\mu_{i,i} = 1$ for $1 \leq i \leq n$ and $\mu_{i,j} = 0$ for $i < j$ the following equation holds

The main difference between the various lattice reduction algorithms is the imposed conditions, that will influence the achieved orthogonality, the norm of the lattice basis and the computational cost. The most extensively used lattice reduction technique is the LLL [2] summarized in Algorithm 1.

Definition 1. *Given a lattice* $\mathbf{L} \in \mathbb{R}^n$ *with basis* $\mathbf{B} = (\underline{b}_1, \cdots, \underline{b}_n) \in \mathbb{R}^{n\times n}$*, associated orthogonal basis* $\mathbf{B}^* = (\underline{b}_1^*, \ldots, \underline{b}_n^*) \in \mathbb{R}^{n\times n}$*, and Gram-Schmidt coefficients* $\mu_{i,j}$*,* $\mathbf{B}$ *is called* LLL-reduced *if the following conditions are satisfied: (i.)* $|\mu_{i,j}| \leq \frac{1}{2}$ *for* $1 \leq j < i \leq n$ *and (ii.)* $\|\underline{b}_i^* + \mu_{i,i-1}\underline{b}_{i-1}^*\| \geq \delta\|\underline{b}_{i-1}^*\|$ *for* $1 < i \leq n$ *,* $\frac{3}{4} \leq \delta < 1$

III. CUDA overview

The programming of the GP-GPU devices became popular when Nvidia launched the Compute Unified Device Architecture (CUDA) parallel programming model. An extensive description of CUDA programming and optimization techniques can be found in [7]. The main entry points of a GP-GPU programs are called *kernels*. These kernels are executed N_t times in parallel by N_t different *CUDA threads*. CUDA threads are grouped in *thread blocks* (TB). A *grid* is a collection of thread blocks. Either the threads in the thread block, either the thread blocks in the grid can have a one-dimensional, two-dimensional, or three-dimensional ordering. The ordering is motivated by the problem to be solved, thus when launching a kernel the grid size, the number of threads in a thread block and the ordering dimensions have to be defined by the programmer. In the latest CUDA release new features, such as dynamic parallelism [7], were introduced. Dynamic Parallelism is the ability for kernels to be able to dispatch other kernels. With Fermi only the CPU could dispatch a new kernel, which incurs a certain amount of overhead by having to communicate back and forth with the CPU.

IV. Parallel Lattice Reduction

The LLL algorithm is highly sequential. In order to parallelize this algorithm multiple levels of parallelism have to be identified and explored. As mentioned in the previous sections, existing parallel LR implementations [4, 8, 9] were focusing only on multi-core architectures. The drawback is the low number of the threads and the limited parallelization possibilities offered by these compared to the GP-GPU. In case if the problem could be divided in several sub-problems and these sub-problems could benefit from a multi-threaded environment the low number of threads offered by the multi-cores would mean a significant limiting factor. In case of the GP-GPU the high number of CUDA cores makes possible the parallel execution of a high number of threads which makes feasible the usage of a multi-level parallelism.

The order in which the swaps are applied in Algorithm 1 are too restricting in a parallel framework. Villard in [10] introduced the *any swap reduction* concept, that enables

Cs. M. Józsa, "Efficient GPU implementation of lattice-reduction,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 129-132.

Algorithm 1 The LLL algorithm

Input: $\mathbf{B}$, δ
Output: LLL reduced basis
Compute $\mathbf{B}^*$ and $\mathbf{U}$ with the Gram-Schmidt algorithm
$k = 2$
while $k \leq n$ **do**
  SIZEREDUCE(k,$k-1$)
  if $\|\underline{b}_k^*\|^2 < (\delta - \mu_{k,k-1}^2)\|\underline{b}_{k-1}^*\|^2$ **then**
    SWAP(k)
    $k = \max(k-1, 2)$
  else
    for $l = k-2 \rightarrow 1$ **do**
      SIZEREDUCE(k,l)
    end for
    $k = k+1$
  end if
end while
procedure SIZEREDUCE(k,l)
  if $|\mu_{k,l}| > \frac{1}{2}$ **then**
    $\mu = \lceil \mu_{k,l} \rfloor$, $\mu_{k,l} = \mu_{k,l} - \mu$, $\underline{b}_k = \underline{b}_n - \mu \cdot \underline{b}_l$
    for $j = 1 \rightarrow l-1$ **do**
      $\mu_{k,l} = \mu_{k,l} - \mu \cdot \mu_{l,j}$
    end for
  end if
end procedure
procedure SWAP(k)
  Swap $\underline{b}_k$ with $\underline{b}_{k-1}$
  $\underline{b}_{k-1}^{*'} = \underline{b}_k^* + \mu_{k,k-1}\underline{b}_{k-1}^*$
  $\mu_{k,k-1}^{'} = (\underline{b}_{k-1}^*, \underline{b}_{k-1}^{*'})/\|\underline{b}_{k-1}^{*'}\|^2$
  $\underline{b}_k^{*'} = \underline{b}_{k-1}^* - \mu_{k,k-1}^{'}\underline{b}_{k-1}^{*'}$
  for $j = 1 \rightarrow k-2$ **do**
    Swap $\mu_{k,j}$ with $\mu_{k-1,j}$
  end for
  for $i = k+1 \rightarrow n$ **do**
    $\mu_{i,k-1}^{'} = \mu_{i,k-1} \cdot \mu_{k,k-1}^{'} + \mu_{i,k} \cdot \|\underline{b}_k^*\|^2/\|\underline{b}_{k-1}^{*'}\|^2$
    $\mu_{i,k}^{'} = \mu_{i,k-1} - \mu_{i,k} \cdot \mu_{k,k-1}$
    $\mu_{i,k} = \mu_{i,k}^{'}$, $\mu_{i,k-1} = \mu_{i,k-1}^{'}$
  end for
  $\underline{b}_{k-1}^* = \underline{b}_{k-1}^{*'}$,$\underline{b}_k^* = \underline{b}_k^{*'}$,$\mu_{k,k-1} = \mu_{k,k-1}^{'}$
end procedure

simultaneous basis swaps. This approach served as a basis for future parallel implementations. In case of big matrices a lot of swaps have to be done and the computational requirements are enormous, thus reducing the problem dimensions could lead to further speed-ups. In [3] the block concept of the LLL algorithm was introduced. Instead of applying the LLL algorithm on a high dimension matrix, the matrix is splitted into several sub-problems of lower dimensions and LLL reduction is applied on them.

Further computational cost can be saved by rearranging and delaying the frequently used *Size Reductions* procedures. In [4] the concept of delaying the *Size Reductions* was introduced.

In Algorithm 2 a Cost Reduced All-Swap LLL (CRAS-LLL) algorithm is presented. This combines the any swap strategy with the size reductions strategy. In Algorithm 3 a modified Block LLL (MB-LLL) algorithm is presented, based on the blocking concept. The most important result of the work of Józsa *et al.* presented in [1] besides the improvement of the existing any swap and block concept is that Algorithms 2 and 3 can be used together serving as different levels of parallelism.

The performance of the CRAS-LLL algorithm depends on the efficiency of (i.) *work distribution*, (ii.) *inner product*, (iii.) *size reduction* computation and (iv.) the *column swapping*. In Fig. 1 we give a possible solution for the main parts of the CRAS-LLL algorithm. The *work distribution* is easy to solve and it is highly parallel. The y dimension of thread blocks are defined based of the size of the original basis, namely $y = \max(n/2, 32)$. (A maximum limit is motivated by optimization reasons.) As a result the basis pairs have to be distributed among y number of thread groups. By enabling the usage of $x = \max(n, 32)$ threads in the x dimension the above mentioned *size reductions*, *inner products* and *column swaps* can be computed in a more efficient manner, because the advantages of the caching system and the low latency of the shared memory can be exploited. When computing the inner product the elements of $\underline{b}_k$ are reached in a coalesced pattern and each thread will sum the corresponding elements in the shared memory buffer. After the sum, the parallel prefix sum pattern is applied to the buffer, resulting in the inner product value. In the case of the size reduction, the corresponding elements are reached in a coalesced pattern and the corresponding $\mu_{k,k-1}$ is read from the shared memory.

In Fig. 2 the scheme of the Block-LLL algorithm kernel launches is presented. The MB-LLL algorithm is implemented based on the *dynamic parallelism*, enabling the launch of new kernels from other kernels. Coarse grained parallelism is achieved on this level, because synchronization is not frequent. Moreover if several basis are lattice reduced simultaneously this can be done independently of each other. To achieve this independence in the case of mult-core architectures complex control logic has to be written, that degrades the performance significantly. The main kernels of the LR flow are the **Block-LLL Kernel** that is responsible for passing the correct data, launch the **CRAS-LLL Kernels** for performing the LLL reduction of the sub-problems, launch the **Boundaries Check Kernels** for checking the LLL conditions at the boundaries of the sub-groups and launch the **GSC-Update Kernel** to update the GSC coefficients and to perform the size reductions wherever it is required.

V. NUMERICAL RESULTS

Fig. 3 shows the computational time of the CRAS-LLL and the MB-LLL algorithms, evaluated on a Nvidia Tesla K20 GP-GPU. In the case of MB-LLL, different configurations regarding the block size are shown. In this figure, we can see that the parallelization and efficient implementation used in the CRAS-LLL algorithm outperforms the results presented in [11, 12] regarding the computational time. Comparing

Algorithm 2 Cost Reduced All-Swap LLL algorithm

Input: $\mathbf{B}$, δ
Output: LLL reduced basis
Compute $\mathbf{B}^*$ and $\mathbf{U}$ with the Gram-Schmidt algorithm
$oddSwap = true$, $evenSwap = true$, $i = 1$
while $oddSwap$ or $evenSwap$ **do**
  if $i \mod 2 == 1$ **then**
    $oddSwap = false$, $off = 1$
  else
    $evenSwap = false$, $off = 0$
  end if
  for $k = 2 + off$ **to** n **step** 2 **do** ▷ Embarrassingly parallel for all k
    Update $\mu_{k,k-1}$
    SIMPLESIZEREDUCE(k,$k-1$) ▷ Only $\mu_{k,k-1}$ is reduced
    if $\|\underline{b}_k^*\|^2 < (\delta - \mu_{k,k-1}^2)\|\underline{b}_{k-1}^*\|^2$ **then**
      SIMPLESWAP(k) ▷ No GS coefficients are updated
      if $i \mod 2 == 1$ **then**
        $oddSwap = true$
      else
        $evenSwap = true$
      end if
    end if
  end for
  UPDATEGSCOEFFICIENTS ▷ Highly parallel
  $i = i + 1$
end while
procedure SIMPLESIZEREDUCE(k,l)
  if $|\mu_{k,l}| > \frac{1}{2}$ **then**
    $\mu = \lceil \mu_{k,l} \rfloor$, $\mu_{k,l} = \mu_{k,l} - \mu$, $\underline{b}_k = \underline{b}_n - \mu \cdot \underline{b}_l$
  end if
end procedure
procedure SIMPLESWAP(k)
  Swap $\underline{b}_k$ with $\underline{b}_{k-1}$
  $\underline{b}_{k-1}^{*\prime} = \underline{b}_k^* + \mu_{k,k-1}\underline{b}_{k-1}^*$
  $\mu_{k,k-1}^{\prime} = (\underline{b}_{k-1}^*, \underline{b}_{k-1}^{*\prime})/\|\underline{b}_{k-1}^{*\prime}\|^2$
  $\underline{b}_k^{*\prime} = \underline{b}_{k-1}^* - \mu_{k,k-1}^{\prime}\underline{b}_{k-1}^{*\prime}$
  $\underline{b}_{k-1}^* = \underline{b}_{k-1}^{*\prime}$,$\underline{b}_k^* = \underline{b}_k^{*\prime}$,$\mu_{k,k-1} = \mu_{k,k-1}^{\prime}$
end procedure
procedure UPDATEGSCOEFFICIENTS
  for $i = n - 1 \rightarrow 1$ **do**
    for $j = n \rightarrow i + 2$ **do**
      $\mu_{j,i} = (\underline{b}_j, \underline{b}_i^*)/\|\underline{b}_i^*\|^2$
      SIMPLESIZEREDUCE(j,i)
    end for
  end for
end procedure

Algorithm 3 Modified Block LLL algorithm

Input: $\mathbf{B}$, δ, block-size l
Output: LLL reduced basis
Compute $\mathbf{B}^*$ and $\mathbf{U}$ with the Gram-Schmidt algorithm
$m = \lceil n/l \rceil$ ▷ m denotes the number of blocks
for $k = 1 \rightarrow m$ **do** ▷ Create the subgroups $\mathbf{B_{[k]}}$,$\mathbf{B^*_{[k]}}$,$\mathbf{U_{[k]}}$
  $\mathbf{B_{[k]}} = (\underline{b}_{l\cdot(k-1)+1}, \ldots, \underline{b}_{l\cdot k})$
  $\mathbf{B^*_{[k]}} = (\underline{b}^*_{l\cdot(k-1)+1}, \ldots, \underline{b}^*_{l\cdot k})$
  $\mathbf{U_{[k]}} = U_{(l\cdot(k-1)+1\ldots l\cdot k)\times(l\cdot(k-1)+1\ldots l\cdot k)}$ ▷ $\mathbf{U_{[k]}}$ is the $l \times l$ submatrix of U
  $echange[k] = true$
end for
while $\exists k$ such that $exchange[k]$ is **true do**
  for $k = 1 \rightarrow m$ **do** ▷ Embarrassingly parallel
    if $exchange[k]$ is **true then**
      LLL($\mathbf{B_{[k]}}$,$\mathbf{B^*_{[k]}}$,$\mathbf{U_{[k]}}$) ▷ Call CRAS-LLL without performing GS orthogonalization
      $group[k] = true$
    end if
  end for
  for $k = 1 \rightarrow m - 1$ **do** ▷ Checking the boundaries of the groups, embarrassingly parallel
    if $group[k]$ **or** $group[k+1]$ is **true then**
      Update $\mu_{k\cdot l,k\cdot l-1}$
      SIMPLESIZEREDUCE($k \cdot l$,$k \cdot l - 1$)
      if $\|\underline{b}^*_{k\cdot l+1}\|^2 < (\delta - \mu^2_{k\cdot l,k\cdot l-1})\|\underline{b}^*_{k\cdot l}\|^2$ **then**
        for $j = k \cdot l - 1 \rightarrow k \cdot l - l + 1$ **do** ▷ Prepare the GS coef. outside the groups
          $\mu_{k\cdot l+1,j} = (\underline{b}_{k\cdot l+1}, \underline{b}^*_j)/\|\underline{b}^*_j\|^2$
        end for
        for $i = k \cdot l + 2 \rightarrow k \cdot l + l$ **do**
          $\mu_{i,k\cdot l} = (\underline{b}_i, \underline{b}^*_{k\cdot l})/\|\underline{b}^*_{k\cdot l}\|^2$
        end for
        SWAP($k \cdot l + 1$) ▷ Update only the GS coef. inside the groups
        $echange[k] = true$, $echange[k+1] = true$
      end if
    end if
  end for
end while
UPDATEGSCOEFFICIENTS ▷ Only update the GS coefficients outside the groups

CRAS-LLL with MB-LLL, it can also be observed that the block concept used in MB-LLL allows to reduce the computational time for systems with higher dimensions.

VI. CONCLUSIONS

In this work a brief extract of the paper [1] was presented, describing an efficient design and implementation of the parallel LLL algrithm for many-core architectures. In order to achieve peak performance two strategies, CRAS-LLL and MB-LLL were used. The results show that in the case of large matrices, MB-LLL slightly outperforms the CRAS-LLL.

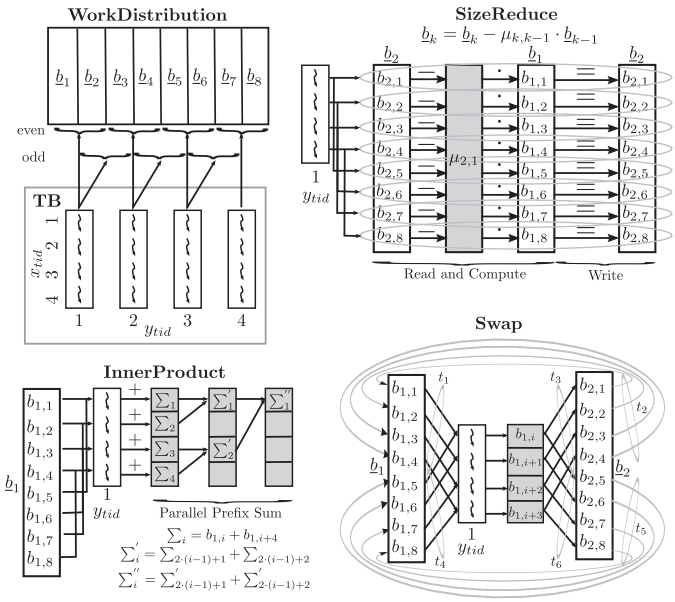


Fig. 1. CRAS-LLL algorithm mapping to GPU architecture.

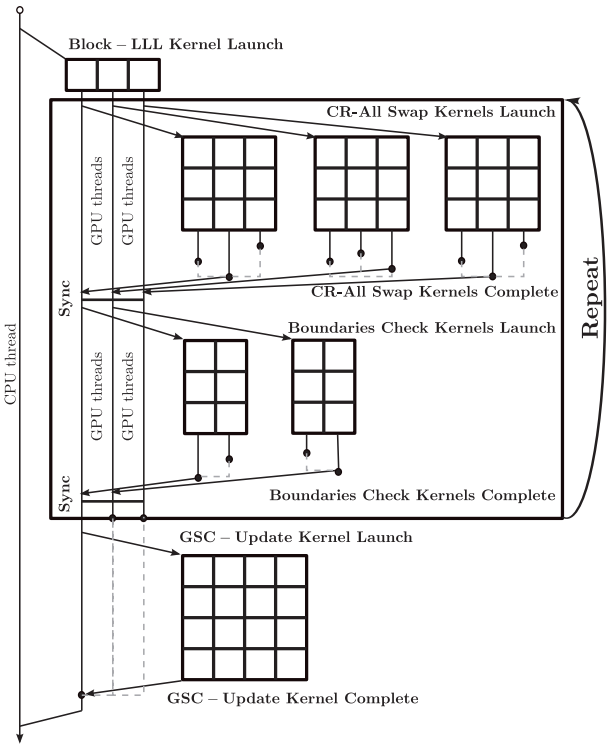


Fig. 2. Kernels Scheduling for the Block-LLL algorithm.

However, both designs show lower execution times compared to previous implementations.

ACKNOWLEDGMENTS

The author would like to thank Antonio M. Vidal, Gema Piñero, Alberto González and Fernando Domene at Univ. Politècnica de València for their help, ideas and support.

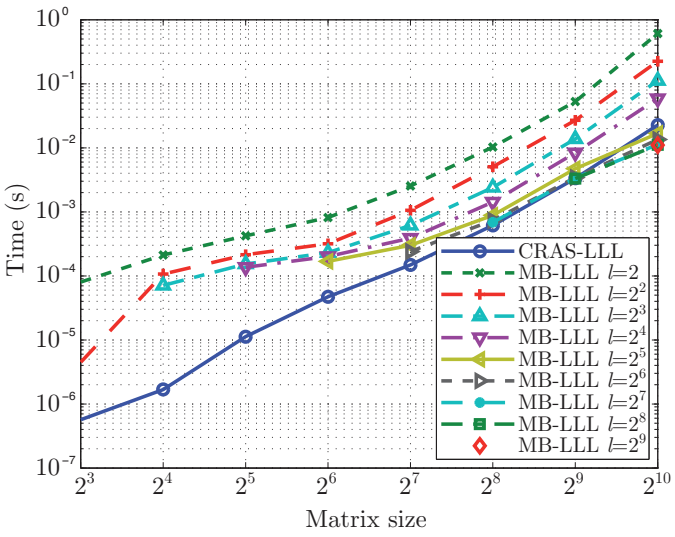


Fig. 3. Computational time of CRAS-LLL and MB-LLL with different block sizes for square matrices of different dimensions.

REFERENCES

[1] C. M. Józsa, F. Domene, G. Piñero, A. Gonzalez, and A. M. Vidal, "Efficient gpu implementation of lattice-reduction-aided multiuser precoding," in *Submitted to Wireless Communication Systems (ISWCS), 2013 10th International Symposium on*.

[2] A. K. Lenstra, H. W. Lenstra, and L. Lovász, "Factoring polynomials with rational coefficients," *Mathematische Annalen*, vol. 261, no. 4, pp. 515–534, 1982.

[3] S. Wetzel, "An efficient parallel block-reduction algorithm," in *Algorithmic Number Theory*. Springer, 1998, pp. 323–337.

[4] Y. Luo and S. Qiao, "A parallel LLL algorithm," in *Proceedings of The Fourth International C* Conference on Computer Science and Software Engineering*, 2011, pp. 93–101.

[5] H. Vetter, V. Ponnampalam, M. Sandell, and P. Hoeher, "Fixed complexity LLL algorithm," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1634–1637, 2009.

[6] D. Wubben, D. Seethaler, J. Jaldén, and G. Matz, "Lattice reduction," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 70–91, 2011.

[7] NVIDIA Corporation, "CUDA C Programming Guide," http://docs.nvidia.com/cuda/cuda-c-programming-guide/, 2012.

[8] W. Backes and S. Wetzel, "Parallel lattice basis reduction - the road to many-core," in *High Performance Computing and Communications (HPCC), 2011 IEEE 13th International Conference on*, 2011, pp. 417–424.

[9] U. Ahmad, A. Amin, M. Li, S. Pollin, L. Van der Perre, and F. Catthoor, "Scalable block-based parallel lattice reduction algorithm for an SDR baseband processor," in *Communications (ICC), 2011 IEEE International Conference on*, 2011, pp. 1–5.

[10] G. Villard, "Parallel lattice basis reduction," in *Papers from the international symposium on Symbolic and algebraic computation*, ser. ISSAC '92. New York, NY, USA: ACM, 1992, pp. 269–277. [Online]. Available: http://doi.acm.org/10.1145/143242.143327

[11] L. Bruderer, C. Studer, M. Wenk, D. Seethaler, and A. Burg, "VLSI implementation of a low-complexity LLL lattice reduction algorithm for MIMO detection," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 2010, pp. 3745–3748.

[12] L. G. Barbero, D. L. Milliner, T. Ratnarajah, J. R. Barry, and C. Cowan, "Rapid prototyping of Clarkson's lattice reduction for MIMO detection," in *Communications, 2009. ICC'09. IEEE International Conference on*, 2009, pp. 1–5.

Multimodal analysis of the human cortical spontaneous synchronous population activity in vitro

Bálint Péter Kerekes
(Supervisor: István Ulbert)
bkerekes@cogpsyphy.hu

Abstract- **Spontaneous synchronous population activity (SPA) emerges from the cortical slices of epileptic and non-epileptic tumor patients maintained in physiological medium in vitro. In order to gain additional spatial information about the network mechanisms involved in the SPA generation, we introduced the two-photon. Human slices were maintained in a dual superfusion chamber of high flow rate physiological incubation medium and otherwise conventional submerged technique to elicit SPA in a two-photon microscope. The population activity was recorded by laminar extracellular electrodes and an extracellular patch electrode. After identifying the active regions of the slice using electrophysiology techniques, bolus loading of OGB-1 and SR101 was applied on the tissue. The neuronal and glial cells took up these dies, thus we were able to image the SPA related Ca-transients in pyramidal cells with two-photon technique, simultaneously with extracellular and whole cell patch measurements. Combining high spatial resolution two-photon Ca-imaging technique and high temporal resolution extra- and intracellular electrophysiology techniques may permit a deeper understanding about the network properties of SPA in the human cortex.**

Index Terms- **2-photon microscopy, spontaneous synchronous population activity**

I. INTRODUCTION

Epilepsy is a common neurological disorder; it's related to hyperactivity of neuronal circuits. Some time the pharmacological treatment is not effective so in these cases the neurosurgeons remove some of the tissue. That gives a possibility to study living human tissue, which could be involved in the generation of the disorder. In vitro studies have shown that these tissues generate SPA[1-6]. In this study we want to describe human neocortical SPA in epileptic patients, and patients with tumor but no epilepsy as control. We made collaboration with the two-photon Imaging Center in PPKE ITK to connect the 2-photon microscopy, with intracellular and extracellular recordings in these studies (Figure 1).

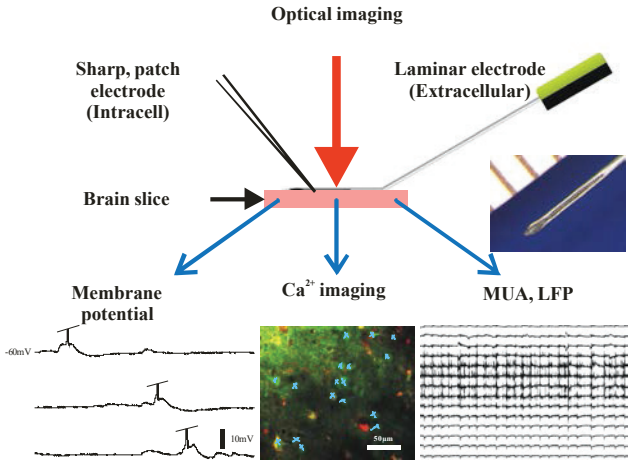


Figure 1. The sematics of the 3 tipes of in vitro neural measuring methods used in our experiments.

II. MATERIALS AND METHODS

A. *Laminar electrode*[7]:

The extracellular laminar electrode is nowadays distributed by Plexon, Plextrode® U-Probe (Figure 2-3.) and we used the brain slice configuration of it. The 24 channel probe was connected to the data transmission systems head stage. We have fixed the electrode to the 2-photon microscopes built-in micromanipulator, to maneuver the probe above the slice.

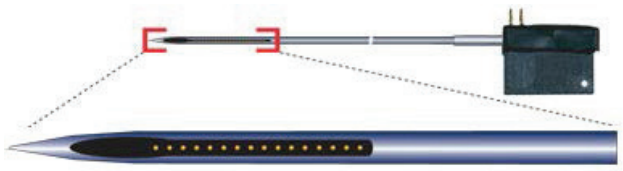


Figure 2. The sematics of Plextrode® U-Probe.

B. P. Kerekes, "Multimodal analysis of the human cortical spontaneous synchronous population activity in vitro,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 133-136.

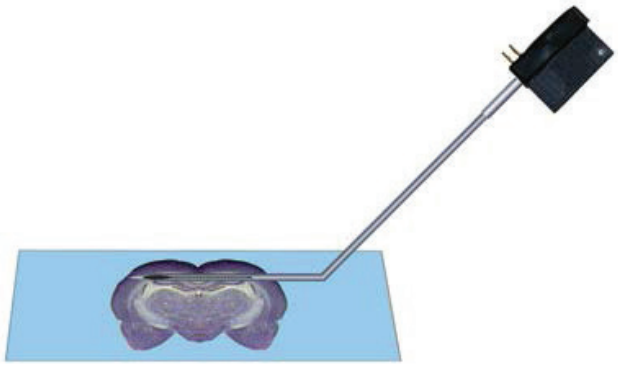


Figure 3. The scematics of Plextrode® U-Probe, and the Brain slice configuration of it under. [28]

B. 2-photon microscope

The 2-photon microscope (Figure 4.) what we used is a Femtonics Kft. design [8]. We are using this device for human in-vitro studies.

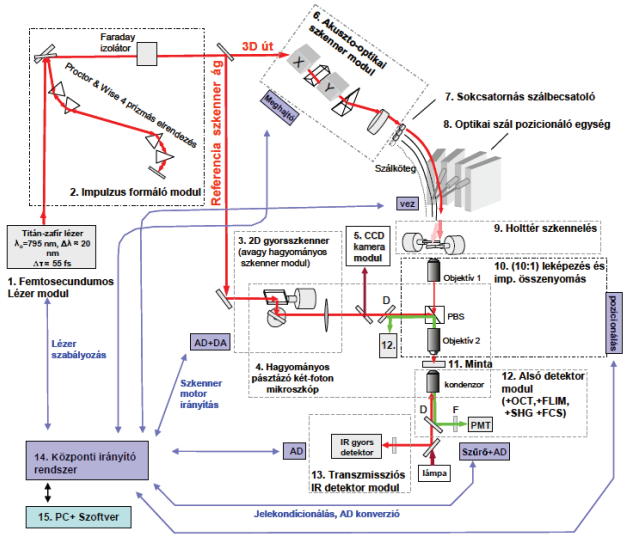


Figure 4. The schematics of the Femtonics 2-photon microscope [8]

C. Human surgery

For the human in-vitro slice preparation we obtained the brain tissue from surgeries of epileptic or tumor (control) patients (in this case the tissue were from the not infiltrated region of the cortex above the tumor). After the surgery the tissue was taken to a nearly frozen state cutting solution (1L solution of: Sucrose: 85.5g, NaHCO3: 2.184g, D-Glucose: 1.802g, from the 1M stock solutions: KCl: 1ml, CaCl2: 1ml, MgCl2 10ml, and 1ml Phenol Red, and dH2O) (Figure 5.). After removing of the pia we cut 500um thick slices perpendicularly from the top of the cortex with vibratom (Leica VT 1000 S), and they were taken into 1hour incubation in ACSF (2L solution of: NaCl 14.384g, NaHCO3: 4.368g, D-Glucose: 3.604g, from the 1M stock solutions: KCl: 7ml, CaCl2: 2ml, MgCl2 2ml, and dH2O) on 36C°. The in-vitro experiments were made in 2-photon microscope in a double superfusion chamber with high flow rate to maintain the good oxygenation level of the tissue.

SOLUTION	substance	molar mass g/mol	end concentration mmol/L	volume of the solution ml	how much to put in g	at the end add from 1 M stock solution (ml)	pH	osm
ACSF	*NaCl*	58	124	2000	**14.384**			314
	KCl	74	3.5	2000		7		
	NaHCO3	84	26	2000	**4.368**		7.485	
	D-Glucose	180.2	10	2000	**3.604**			
	CaCl2	110.8	1	2000		**2**		
	MgCl2	95.1	1	2000		**2**		
CUTTER / CARRIER	*Sucrose*	342	250	1000	**85.5**			336
	KCl	74	1	1000		**1**		
	NaHCO3	84	26	1000	**2.184**			
	D-Glucose	180.2	10	1000	**1.802**			
	CaCl2	110.8	1	1000		**1**		
	MgCl2	95.1	10	1000		**10**		
	Phenol Red					**1**		

Figure 5. The solution what used to carry and cut the brain tissue and the ACSF what used in the 2-photon microscopes chamber [9].

D. Recordings

For the extracellular recordings we managed to combine the 2-photon microscope setup with a 24-channel laminar electrode specially made for in-vitro experiments. The LFP recordings were made with extracellular patch electrodes filled with ACSF. The intracellular patch electrode –used for whole cell recording- were filled with mixture of standard IC, OGB1 60µM (cell impermeable version), Alexa 5uM and biocitin. The used patch electrodes had 5-9MOhm resistance. In current clamp recordings the cells were hold near -70mV. We used Ti:S laser system (80fs long laser impulses, on 80MHz, on 820-840nm wavelengths) for the Ca^{2+} imaging and line scans on the patched cells dendrite and soma and some cells in the field of view for population activity.

III. RESULTS

A. 2-photon microscope and extracellular electrode

Our goal was to combine the two systems (2-photon microscopy, with extra- and intracellular measuring) to make human cortex in-vitro studies, to understand more the phenomena beside the sharp wave oscillation. First we scanned through the slice with the laminar electrode, and if we found good activity, or the oscillation what we needed, than changed into 2-photon mode, and made bolus loading (mixture of oregon green bapta1 and sulforodamin 101) into some sites in the cortex layers, where the activity appeared. After a half hour the cells had taken up the bolus, and we could start to catch the Ca responses with line scans. We put in a LFP micropipette filled with ACSF to measure the field potentials, to correlate it later with the Ca responses, and after we found a good responding cell we tried to patch it, and measured the LFP, intracellular and Ca responses simultaneously. The results were correlated in MatLab (Fig 6-7.).

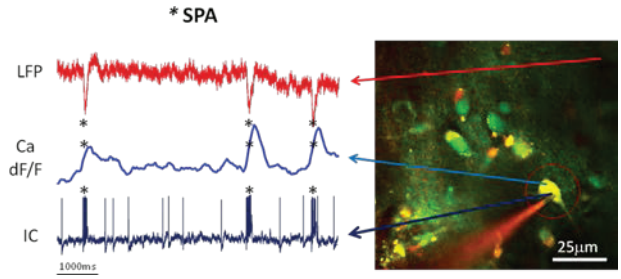


Figure 6. 2-photon picture of the loading site, with the patch pipette (orange), the LFP pipette is on the top right of the screen, the line scan is on the patches cells dendrite.

.

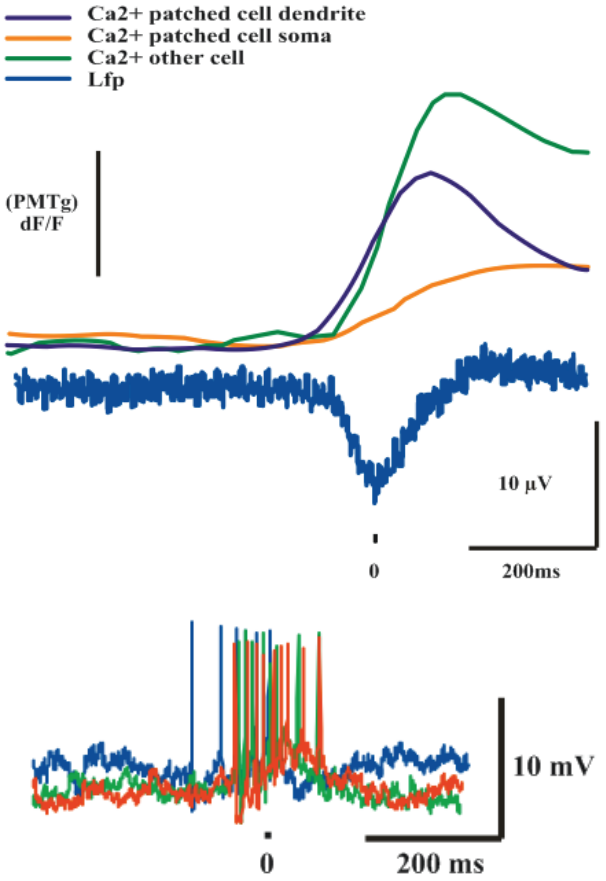


Figure 7. Sharp wave (middle), Ca response (above), cell burst (below), of one cell. The sharp wave was observed with the LFP micropipette, within 100um, the Ca response was observed with 2-photon line scan on the dendrite, and the cell was patched with a micropipette, to observe its potentials.

The patched cells were filled up so after the recordings we prefer to make camera lucida reconstruction for getting more morphological information of the cells (Figure 8).

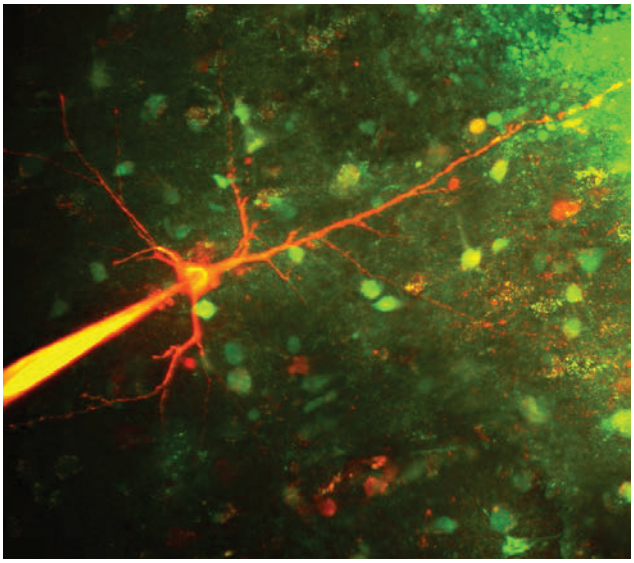


Figure 8. A projection of the Z-stack made with the MES program from a patch electrode and the filled cell.

We are also trying to measure how the cells are taking part of the generation of the SPA and the percentage of the cells which are responsible for it.

IV. CONCLUSION

We are in the middle of the experiment and the data analysis so yet we don't want to make guesses what is the concrete neuronal background of the human SPA, but our primary measurement shows similarities with the previously made rodent experiments.

V. FUTURE PLANS

In the 2-photon experiments we will try some ACSF solutions for investigation of their applicability in our research, and make more experiments in the sharp wave oscillation study.

ACKNOWLEDGEMENT

The author wish to acknowledge Dr. György Karmos, Dr. István Ulbert, Richárd Fiáth, and Domonkos Horváth from the MTA-PKI for the assistance, and advice, and Rózsa Balázs, and the PPKE ITK 2-photon team for the help in the 2-photon microscope lab, and OITI for the human surgeries.

The research was supported by the following grants: OTKA K81354, OTKA PD77864, ANR-TÉT Neurogen, ANR-TÉT Multisca, TÁMOP-4.2.1./B-11/2/KMR-2011-002, TÁMOP-4.2.2./B-10/1-2010-0014

REFERENCES

[1] Rüdiger Köhling, *et al.*, „Does interictal synchronization influence ictogenesis?" , *Neuropharmacology,* 2012

[2] Rüdiger Köhling, *et al.*, „Cellular and molecular mechanisms of epilepsy in the human brain” , *Progress in Neurobiology* vol. 77, 166–200, 2005
[3] Rüdiger Köhling, *et al.*, „Optical imaging of epileptiform activity in experimentally induced cortical malformations” , *Experimental Neurology,* vol. 192, 288– 298, 2005
[4] Rüdiger Köhling, *et al.*, „Methodological approaches to exploring epileptic disorders in the human brain in vitro” , *Journal of Neuroscience Methods,* vol.155, 1–19, 2006
[5] Rüdiger Köhling, *et al.*, " Spontaneous sharp waves in human neocortical slices excised from epileptic patients", *Brain*,vol. 121, pp. 1073–1087, 1998
[6] György Buzsáki, *et al.*, " Relationships between Hippocampal Sharp Waves, Ripples, and Fast Gamma Oscillation: Influence of Dentate and Entorhinal Cortical Activity ", *The Journal of Neuroscience,* vol. 31(23), 8605– 861, 2011
[7] http://www.plexon.com/product
[8] B. Rózsa, "Hippokampális interneuronok dendritikus Ca2+ szignalizációjának mérése 2-foton pásztázó mikroszkóp technológiával ", Semmelwies Egyetem Szentágióothai János Idegtudományi Doktori Iskola, Dokori értekezés.
[9] Wittner Luca – ERC Starting Grant 2013 Research proposal

Digital holographic microscopy for single-shot, volumetric and fluorescent measurements

Márton Kiss
(Supervisor: Dr. Szabolcs Tőkés)
kisma1@digitus.itk.ppke.hu

***Abstract*—A single-shot, volumetric and fluorescent digital holographic microscope setup is introduced here. The aim is to develop a microscope that is able to detect fluorescent and freely flowing microscopical objects. This is why single-shot exposure is needed. The presented setup is based on a Hariharan-Sen interferometer. In this presentation the relation between the place of the target and the quality of the hologram is introduced.**

Keywords: Incoherent or self referenced Digital Holographic Microscope, single-shot exposure, bifocal optical system, Hariharan-Sen interferometer, volumetric imaging

I. Introduction

Nowadays it is a frequently asked question whether the water we use for drinking and swimming is clean enough or not. The quality of water can be measured physically, chemically and biologically. After the first impression that comes from the water's physical properties we usually ask that what kinds of living organisms are in it? Because the living organisms are indicators of the quality of the water, in many cases their presence gives enough information about the water and chemical measurement is not needed. The main indicators in the water are the living bacteria, algae, cells, worms and other micro-organisms that usually can be seen only with microscope. The living cells can also be detected by the help of their fluorescence capability, and also it helps to separate them from the debris. Measurement without any preparation of the sample and volumetric imaging can supports fast, real time and automatic measurements. This is the background of our aim, which is to build a microscope for real time water measurements and monitoring.

The self-referenced digital holographic microscopy is a type of microscopes that can have the advantage of volumetric viewing and fluorescent imaging. The first holographic setup [1] that was invented used a (color and spatial filtered lamp's light as) coherent light, and its theoretical background was also based on the attribute of the coherent light. At the self-referenced or in other name incoherent holography the theoretical basis is the same: with two beams (reference and target), which are coherent with each other, an interference fringe system called hologram is created. If we illuminate the hologram with the known beam (the reference one), the target beam that can draw the image of the target can be produced. At self-referenced holography the reference and target beams light sources are the same target point, which is self-luminous or just reflects the light that has a short coherent length. That is why the optical path difference of the self-referenced setup has to be smaller than the coherent length of the used light. In self -referenced holography color filter is usually used, because decreasing the bandwidth of the scattered or emitted light the coherent length can be increased, and also if many lights with different wavelength create hologram from the same object point to the same plane their interference fringes may disturb each other. And also one hologram is reconstructed numerically with one wavelength.

At the beginning of the self-referenced holography the main idea was to create Fresnel zone-plate (FZP) from the interference fringes, which is the coherent summation of two beams with different radius emitted by the same object point. If this FZP is back lighted, it will focus the light. They used it for example in astronomy [2]. There are mainly two kinds of methods to separate the light, modulate it in different way and interfere them. The first one when an interferometer is used for example Linnik interferometer [3], Hariharan-Sen interferometer [4], and the second one when a bifocal lens [5] or a spatial light modulator is used [6]. The setup with bifocal lens is more compact and stable than the setup with other interferometers.

Digital holography has the degree of freedom to numerically modulate and modify the hologram and also the reconstruction beam, and the advantage to automatically evaluate the hologram that contains the needed information from the viewed volume. This is why at 3D incoherent holographic imaging could grow up. In these new techniques incoherent holographic imaging can be assisted with tomography [7] or scanning [8]. FINCH is a mature technique of nowadays that ignores these possibilities to create a fast setup but also it gives a high quality image [9]. Because FINCH uses three exposures to retrieve the complex hologram through phase shifting, this method is not able to create images from freely moving samples.

M. Kiss, "Digital holographic microscopy for single-shot, volumetric and fluorescent measurements,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 137-140.

Here I present my setup that is based on a Hariharan-Sen interferometer. It uses only one exposure to get an intensity hologram. But two self-luminescence points that are not coherent with each other could be reconstructed from their own holograms that were captured to the same image by a CCD sensor.

II. SELF REFERENCED HOLOGRAPHY WITH AN INTERFEROMETER

A. self-referenced holographic setup based on a Hariharan-Sen interferometer

Hariharan-Sen interferometer is a triangular shaped optical setup, where the entrance and exit gate for the light is one beam splitter cube. This cube divides the incoming light and then the separated lights go around in the optical path (that is puckered to a triangular by two mirrors) on the same path, but in opposite direction, and than this cube combines them too. Because the beams have the same optical path, there is no difference between them. If this triangular is made asymmetric by properly inserting a lens, the exiting beams will have different wave fronts. This will generate interference fringes. Figure 1 shows this asymmetric interferometer completed with an objective (olympus LUCPLFLN 20X), a tube lens (Bi-Convex lens, f=100mm), a polarizer filter that can set the intensity ratio between the two beams, and a detector (Lumenera).

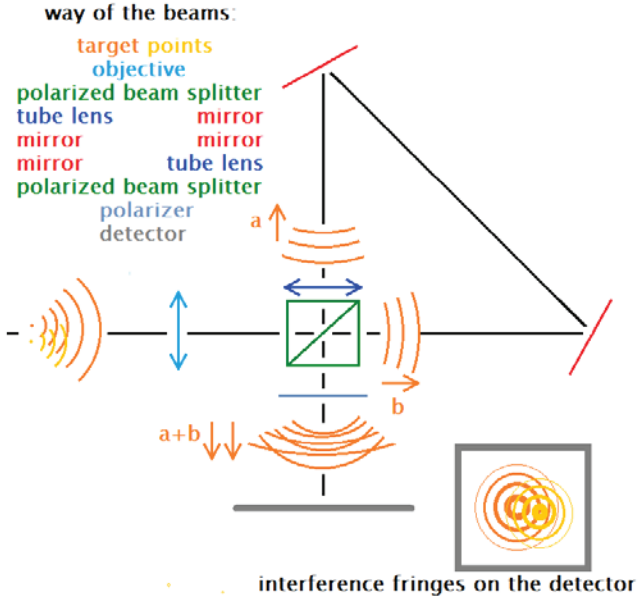


Figure 1. The built self-referenced holographic setup based on a Hariharan-Sen type interferometer. At the way of "a" the beam is going throw an afocal optical system.

One of the two optical ways of the system is afocal. The afocal system has the advantages that the magnification is independent of the target distance, and target and image distance has a linear connection. These are different in a common focal system. Figure 3 displays focal and afocal system's characteristic.

B. Light Source

At the experiment a stabile target was needed. That is why the target points were fiber ends in a same connector with a distance of 128 µm. Light from one red LED was coupled into these fibers, but leaving the fibers they couldn't create any interference fringes, because their coherent length was small enough. This target can be seen in figure 2.

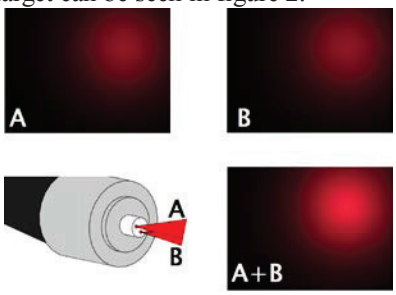


Figure 2. Target was simulated with two fiber coupled red LED light. They couldn't interfere with each other, they added only in intensity.

C. Holograms

In my measurement I was interested in the connection between the target place and the created holograms, and I also wanted to know what kind of image can be reconstruct from them. All the other parameters were fixed. The target was moved from the distance 5 mm from the focal plane of the objective to close to the objective that was 4 mm far from the focal plane. The detector was set after the beam splitter with 20mm, as close to the beam splitter as it was possible. In this case and when the target was in the objective's focal plane, the afocal systems image was before the detector and the focal systems image was after the detector. Moving the target through the above explained area, I founded six sub-areas separated with 5 times. These areas can be seen in figure 3. In the 1st and the 6th the illumination was quite homogenous, because the beams radius was nearly detector size. In this case the holograms were as big as to overlap each other and that is why moiré effect could be seen between the two interference fringes. In the 2nd and 3rd the target is moving a bit, the interference fringes changed so fast, and also they had only a few fringes. At the separating points III. and IV. one of the beams was focused to the detector. So there were no interference fringes and around those separating points the high intensity level disturbs the small holograms.

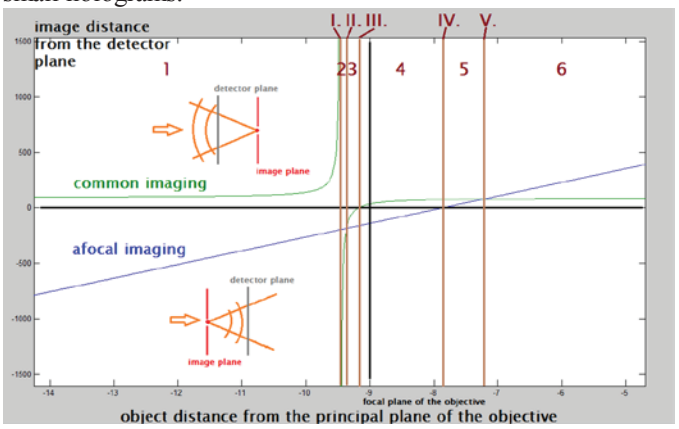


Figure 3. The place of the target before the objective will define the images of the same target point, and also the size and shape of the interference fringes.

In the area of 4 and 5 we got nice interference fringes: they didn't overlap each other and they can be seen clearly. It is shown in figure 4.

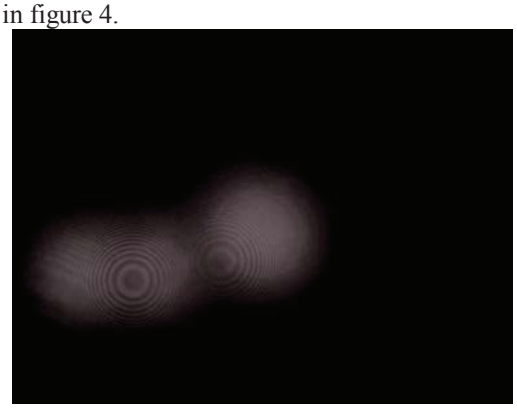


Figure 4. Interference fringes from the 4th area. (See figure 3.)

The separating points II. and V. show that case when a target-point's two images are in the same image plane. In these cases at the detector the curvature of the two beams are the same, so the interference fringes are not concentric but parallel lines.

D. Reconstruction

Angular spectrum method, which is a plane wave propagation method is used to reconstruct the holograms. This method calculates the scalar electric field in this way:

$$E(x,y,z) = F^{-1}\left\{F\left\{E(x,y,0)\right\}\cdot e^{i2\pi\cdot\omega(u,v)\cdot z}\right\},$$

where E is the electromagnetic field, F and F^{-1} are the Fourier and inverse Fourier transforms, ω is the transfer function and the z is the propagating distance.

At the measurement, when the hologram belonged to a target that was at the separating points as it can be seen in figure 3, propagation didn't give any result. In the cases of II. and V. parallel fringes were just moving across the plane because parallel fringes do not focus the plane wave, and in the I. and the III. case the points were already in focus.

At the 2nd to 5th areas of the hologram's reconstruction the problem was that when there were some interference fringes, the reconstructed point cannot be seen at the reconstructed image, because the background intensity overruns it. It can be also possible to compensate the intensity on the hologram (before propagating) to get a higher contrast, but we should see that the better the contrast of the hologram, the better the light efficiency, and at fluorescent imaging, what the final application will be, we should use the light in the best way because it is few.

In the 1st and 6th areas the reconstructed points can be clearly seen as figure 5 and figure 6 shows. When the two point's hologram was propagated at the same time the reconstructed image background was more flat, than when the points hologram were captured and propagated separately, but the contrast became smaller. It also can be seen that the two intensity holograms do not disturb each other, they don't change each other's propagation distance and the place and magnification of the image. The density of the moiré fringes created by the two intensity hologram, gives information about magnification. The closer the fringes are the bigger is the magnification. Two intensity holograms do not disturb each other. It is a question that without any phase retrieves how many point sources can constitue a target. Comparing the 1st and the 6th areas the later has the advantage that it is closer to the objective, so the optical setup can gather more light from the object.

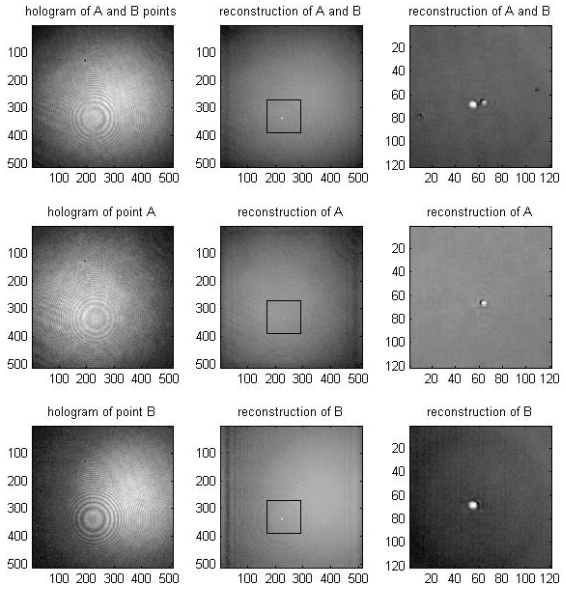


Figure 5. Here a reconstructed image can be seen, where the A and B points that are in the same plane were in the 1st area (see figure 3.)

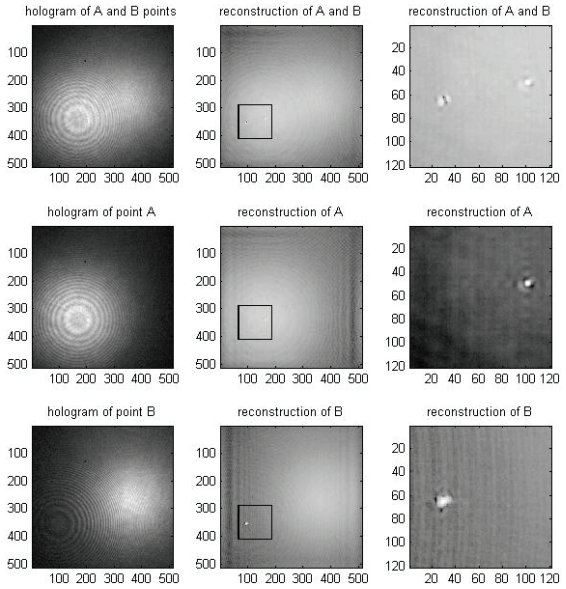


Figure 6. Here a reconstructed image can be seen, where the A and B points that are in the same plane were in the 6th area (see figure 3.)

III. Conclusion

A self-referenced digital holographic microscopy was created with a modified Hariharan-Sen interferometer. The tests showed that this setup is able to create the hologram from a target that is in a large volume and illuminating an incoherent light. Also this can be done without any phase retrieving and scanning, so it can be done with a single exposure. This promising result is fulfilling one of the assumptions to use self-referenced holography at freely moving samples. So to make detection of rare, freely moving and fluorescent object such as algae, the next step will be to increase the luminous power of the optical setup that also has a more sensitive camera.

References

[1] D. Gabor, “A new microscopic principle”, Natue 161,777, (1948).

[2] Mertz and N.O. Young, “Optical Instruments” Proc.ICO Conf. Opt. Instr. (London), (1961).

[3] Kozma Adam and Massey Norman, “Bias level reduction of incoherent holograms”, Applied Optics 8,2, 393-397, (1969).

[4] Gary Cochran, “New method of Making Fresnel Transforms with Incoherent light” JOSA, 56, 11, 1513-1517 (1966).

[5] Lohmann, AW , “Wavefront reconstruction for incoherent objects” JOSA, 55, 11, 1555_1-1556 (1965).

[6] Joseph Rosen, Gary Brooker, “Fluorescence incoherent color holography”, Opt. Express, 15, 5, 2244-2250, (2007).

[7] Y. Sando, M.Itoh, and T. Yatagai, “Holographic three-dimensional display synthesized from three-dimensional Fourier spectra of real existing objects” Optics Letters, 28, 2518, (2003).

[8] Guy Indebetouw, Alouahab El Maghnouji, and Richard Foster, "Scanning holographic microscopy with transverse resolution exceeding the Rayleigh limit and extended depth of focus" JOSA A, Vol. 22 Issue 5, pp.892-898 (2005)

[9] Siegel, Nisan; Rosen, Joseph; Brooker, Gary, "Reconstruction of objects above and below the objective focal plane with dimensional fidelity by FINCH fluorescence microscopy", Optics Express, Vol. 20 Issue 18, pp.19822-19835 (2012)

Improving Hungarian Morphological Disambiguation Quality with Tagger Combination

György Orosz
(Supervisor: dr. Gábor Prószéky)
oroszgy@itk.ppke.hu

***Abstract*—In case of morphologically rich languages full morphological disambiguation is a fundamental task that is known to be more difficult to solve effectively than just providing PoS tags. In our work we overview Hungarian disambiguator tools and present some common tagging combination techniques in order to investigate how these methods and tools could be used together to improve the full annotation accuracy. After analyzing the disambiguators' error analysis, we introduce a method that jointly picks the proper tagger and lemmatizer tool and harmonizes their output, thus achieving a 28.90% error reduction rate compared to a PurePos, a Hungarian state-of-the-art disambiguator.**

***Keywords*-part-of-speech tagging, morphological disambiguation, lemmatization, agglutinative languages, natural language processing**

I. Introduction

Part-of-speech tagging is one of the basic and most studied tasks of computational linguistics. There are several freely available tools and algorithms that work with high precision. However, assigning PoS tags is only a subtask of morphological disambiguation. It is also crucial to identify the lemma, which is not a trivial task for languages having a rich morphology like Turkish or Hungarian. Nevertheless, most of the currently available tools only deal with disambiguating morphosyntactic labels; there are only few that do the whole job. Robust and accurate operation of these tools is important, since they are usually parts of larger linguistic processing chains. Thus errors propagating from this level affect the performance of systems performing more complex language processing tasks.

In our paper, we survey taggers that perform full morphological disambiguation for Hungarian, investigating and comparing their common errors. Lessons learned from the error analysis help us to combine them successfully to gain better performance.

II. Background

First we give a brief overview of full morphological annotation tools for Hungarian. After comparing them, we overview commonly used tagger combination techniques. The experiments described in this paper were performed on the Hungarian Szeged Corpus [3] with PoS annotation automatically converted to morphosyntactic tags used by the Hungarian HuMor morphological analyzer [10], [9]. 10% of the corpus was separated for testing and another 10% is used for development and tuning purposes. Each set contains about 7100 sentences, while the rest, about 57000 sentences, were used for training the systems.

A. Morphological annotation tools

1) PurePos: [8] is an open source hybrid system for full morphological disambiguation. It is based on hidden Markov models, but it can use an integrated morphological analyzer (MA) module as well to tag unseen words and to assign lemmas. The tool uses well-known trigram tagging algorithms, but what distinguishes it from its predecessors is the complete integration of a morphological analyzer, which results in a further boost in its PoS tagging accuracy and also makes high precision lemmatization possible.

2) HuLaPos: [7] is a purely statistical annotation tool based on an SMT[1] decoder. An advantage of applying this methodology to PoS tagging is that it can consider the context in both directions. Moreover, HuLaPos uses a higher order language model than PurePos. On the other hand, HuLaPos has an inferior performance on unseen words, although it utilizes a simple smoothing algorithm that enables it to handle such words to some extent.

3) Magyarlanc: [13], another commonly used tool, is a full processing chain, consisting of a sentence splitter, tokenizer, part-of-speech tagger, lemmatizer, and its latest version even contains a dependency parser. It also contains a built-in morphological analyzer based on morphdb.hu [11]. As a tagger, it is reported to attain 96.33% precision on a random 4:1 split of the Szeged Corpus.

B. Tagger combination schemes

The design process of a combined system of classification or annotation tools involves several steps. First, it needs to be examined whether the errors of each system to be combined are different enough for the aggregate system to be likely to outperform the best individual system significantly. Then an appropriate combining algorithm must be found.

A basic combining scheme, which is often used as a baseline, is majority voting. Other, more advanced, combining schemes involve training a top-level classifier for the task of generating the output of the combined system based on outputs of the individual embedded systems. This class of combination schemes is commonly referred to as stacking learners. The top-level classifier may use various features of both the input

[1] statistical machine translation

Gy. Orosz, "Improving hungarian morphological disambiguation quality with tagger combination,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 141-144.

and the outputs of the bottom-level classifiers when making its decision. The set of features used may have a significant impact on the performance of the combined system.

Finally, decisions to be made by the top-level classifier can be of at least two sorts: it can either always select the output of one of the bottom-level systems, or it can generate an output of its own that may differ from the output of each individual embedded system. When applying the former solution, the errors of the embedded systems determine a theoretical upper limit on the accuracy of the combined system (it can never generate the expected output whenever neither of the embedded classifiers generate it), thus the latter solution seems more beneficial in theory. However, complexity of the annotation task to be performed and the available training data may have an influence on which of these options is feasible and how they perform in practice. If the cardinality of the output annotation and of the features involved in training the classifier is high, there may be either data sparseness or performance problems with the combining classifier, or it may simply become too complicated.

One of the first attempts of combining English PoS taggers was done by Brill and Wu [2]. They propose a memory-based learning system for tagger combination that employs contextual and lexical clues. In their experiments, the solution where the top-level learner always selects the output of one of the embedded taggers outperformed the more general scheme that allowed the output differ from either of the proposed tags.

A comprehensive study by van Halteren et al. [6] presents detailed overview of previous combination attempts using mainly machine learning techniques. Several combination methods are compared and evaluated systematically in the paper. The authors show that cross-validation can be used to train the top-level classifier for an optimal utilization of the training corpus. They found a scheme perform best in their experiments that they characterize as generalized voting, although it is a scheme that can output annotation that may differ from the output of either of the embedded taggers and thus can also be interpreted as a stacking method. However, the cardinality of the tag set and the dimensionality of the feature space was modest compared to that in our case.

A system of different architecture is presented in e.g. Hajič et al. [4]: in contrast to the parallel and hierarchical architecture of the systems above, it employs a serial combination of annotators starting with a rule-based morphological analyzer, followed by constraint-based filters feeding a statistical tagger at the end of the chain.

III. Error analysis

As we mentioned above, it is useful to start the design process with an error analysis of the systems to be combined in order to see whether a system combination is likely to improve performance. We present tagging accuracy values of PurePos (PP) and HuLaPos (HLP) measured on the development set in Table I.[2] Unfortunately, magyarlanc is not directly comparable with the others above, since its built-in annotation scheme is not compatible with the HuMor scheme used by the two other tools.

[2]All other measurements in Sections III and IV were also made on the development set.

TABLE I
BASE SYSTEM ACCURACIES

	Tagging	Lemmatization	Full disambig.
PurePos	98.57%	99.58%	98.43%
HuLaPos	97.61%	98.11%	97.03%

TABLE II
COMPARISON OF PUREPOS AND HULAPOS

	Tagging	Lemmatization	Full disambig.
$OER(PP, HL)$	22.41%	11.66%	21.16%
$OER(HLP, PP)$	53.58%	80.21%	58.24%
Agreement rate	97.60%	98.02%	96.92%
Both are right on agreement	99.30%	99.85%	99.29%
One is right on disagreement	97.53%	98.89%	97.14%
Oracle	99.26%	99.83%	99.22%

It may not be evident from these values why and how combining these tools can boost performance, but deeper investigation on common errors suggests that chances for success are good.

We use the metric[3] $OER(A, B) = (\#errors\ of\ A\ only/\#all\ errors)$ that measures the percentage of the cases where tagger A is wrong but B is correct in proportion of all errors that were made by either A or B. We do not use the complementarity formula proposed by Brill et al.[2], because that gives hard-to-interpret unlimited negative values in cases where there is a significant overlap between the errors made by the two taggers. Although HuLaPos makes more errors than PurePos, own error rates (Table II) indicate that error distribution is fairly balanced between the two tools for tagging and full disambiguation. In addition, we calculated the agreement rate of the tools and the relative percentage of times they agree on the right morphological annotation. Table II also shows that one of them assigns the right annotation most of the time they disagree. Assuming a hypothetical oracle that can always select the better annotation output, the performance of the better tagger can be increased by more than 0.6% corresponding to 72.73% relative error rate reduction on the development set. These results encourage us to combine the two tools.

IV. Annotation tool combinations

It was shown previously [2], [6] that stacking of classifiers can improve PoS tagging accuracy. For an optimal utilization of the training material, we applied training with cross-validation in our experiments, as it was suggested by some authors, e.g. [6]. The training set was split into 5 equal-sized parts and level-0 taggers (PurePos and HuLaPos) were trained 5 times using 4/5 of the corpus, and the rest was annotated by both taggers in each round. The union of these automatically annotated parts of the training corpus was used to train the top-level (or level-1) metalearner. Thus the full training data was

[3]OER=own error rate

TABLE III
FEATURE SETS USED IN THE EXPERIMENTS

ID	Base FS	Additional features
FS1	Brill-Wu	—
FS2	FS1	whether the word contains a dot or hyphen
FS3	FS1	use at most 5-character suffixes
FS4	FS2, FS3	—
FS5	FS1	guessed tags for the second words (right,left)
FS6	FS4	use at most 10-character suffixes

TABLE IV
ERROR RATE REDUCTION USING METALEARNERS WITH DIFFERENT FEATURE SETS SETS

Task:	Tagging		Lemmatization		Full annotation	
ID	NB	IB	NB	IB	NB	IB
FS1	19.03%	24.65%	-6.21%	22.24%	5.06%	22.89%
FS2	18.91%	24.82%	-0.80%	23.85%	4.95%	23.16%
FS3	21.04%	27.60%	0.80%	26.65%	18.42%	25.31%
FS4	20.92%	27.90%	4.01%	26.65%	18.96%	25.20%
FS5	16.37%	17.55%	-19.24%	16.03%	-0.70%	18.47%
FS6	19.27%	27.30%	-17.03%	26.85%	16.16%	25.79%

TABLE V
RELATIVE ERROR RATE REDUCTION ON THE DEVELOPMENT SET

System	Tagging	Lemmatization	Full disamb.
Disamb. combination	23.05%	18.64%	25.79%
Tagger combination	27.90%	6.81%	25.26%
Multiple metalearners	29.85%	30.06%	32.42%

available for level-1 training, yet separating the two phases of the training process. In addition, this workflow made the full training material available also for the level-0 learners.

As the use of a "relatively global, smooth" level-1 learner is suggested in [12], we investigated the naïve Bayes (NB) classifier and instance-based (IB) learners[4] [1], which, in addition to be simple, were shown to perform well in sequence classifier combination tasks. We used the instance-based combinators as follows. Given the high agreement rate of the level-0 taggers on correct events, we decided to use all metalearners only in cases of disagreement. After extracting features for a word on the annotation of which the tools disagree, the classifier finds the most similar previously seen case(s) based on Euclidean distance, and it selects the output of the annotation tool that generated the correct output in the most similar case(s). We decided to use the tagger-picking approach, since Hungarian has a tag set with a cardinality of over a thousand and an almost unlimited vocabulary, which suggests that the tagger-picking approach would not be feasible.[5]

Brill et al. [2] proposed a simple but powerful feature set (FS1 in Table III) that consists of the word to be tagged, its immediate neighbors and all their suggested tags. We intended to extend this feature set systematically to make them better fit languages with a very productive morphology like Hungarian. Several experiments were run[6] in order to investigate whether using word shape (FS2,FS4), word suffix (FS3,FS4,FS6) or wider contextual features (FS5) can improve the performance of tagger selection (see Table III).

In our experiments, the naïve Bayes classifier (NB) performed significantly worse than the instance-based learners (IB) even when using seemingly independent features. Moreover, lemmatizer combination turned out to be an almost insoluble task for it, as error rate reduction data in Table IV show. It is also interesting that accommodating word shape features (FS2) always increased tagging accuracy. The results show that using longer suffix features is beneficial in cases where assigning a lemma is part of the task. However, for combining part-of-speech taggers only, omitting the word form and using at most five-character-long suffix features gives the best result.

We describe below how these combinations were applied to improve automatic Hungarian morphological annotation quality, reporting on relative error rate reduction compared to that of PurePos on the development set.

A. Full disambiguator combination

The most straightforward combination method is to treat annotations as atomic and simply use the full output of the embedded tool selected by the metalearner. Results in Table IV show that the best combination for this approach is the instance-based learner with FS6. Comparison of relative error rate reduction achieved on the development set using this model and alternative models described below is shown in Table V.

B. Combining taggers

Another plausible scheme is to combine the morphosyntactic tagger subsystems of the annotation tools. However, in this case, one has to deal with lemmatization as well. The PurePos lemmatizer performs better than the one built into HuLaPos, so we chose to use that to generate the lemma corresponding to the selected PoS tag. The highest accuracy values for tag selection were achieved by extending the baseline feature set with word shape and at most five-character-long suffix features (FS4).

This algorithm allowed us to achieve higher error rate reduction (see Table V) for the morphosyntactic tags, however, the gain in lemma accuracy is much lower, thus the overall accuracy improvement is inferior to that achieved using the previous method.

C. Multiple metalearners

It is possible to benefit from the strengths of both types of combination using two level-1 learners: one is trained to choose the better lemmatizer and the other to select the optimal tagger for the given case. The combination scheme with best accuracy for lemmatization was the IB classifier with FS6, and that for tagging was the same algorithm with FS4. This configuration may yield incompatible tag-lemma pairs[7]. We

[4]The C4.5 decision tree algorithm was also tested, but it was unable to handle the large amount of feature data involved in our experiments.

[5]We plan to verify this assumption in the future. Nevertheless, all experiments described in this paper used the tagger-picking model.

[6]The WEKA [5] machine learning software was used in the experiments.

[7]A lemma and a tag for a word is incompatible if the MA can analyze the word, but no analysis contains both the lemma and the morphosyntactic label.

TABLE VI
RELATIVE ERROR RATE REDUCTION ON THE TEST SET

System	Tagging	Lemmatization	Full disamb.
Oracle	48.60%	59.42%	51.53%
Disamb. combination	23.23%	23.55%	26.86%
Tagger combination	22.76%	13.77%	23.81%
Multiple metalearners	25.07%	29.89%	28.90%

used the HuMor analyzer to find and fix these cases: if the tag is found to be correct, the lemma is provided by the analyzer and vice versa. With this enhancement, we achieved higher relative error rate reduction and better overall accuracy than with any of the previous two methods.

V. EVALUATION

We present the performance of the best combinations on the unseen test set in Table VI.

In accordance with our results on the development set, the double combination method achieved the best performance with over 98.90% full disambiguation accuracy; that is a 28.90% overall relative error rate reduction. Taking a closer look at the output, we found that the best compound annotation system can partly deal with error types that are typical for PurePos. This combination scheme attained 56.08% of the possible improvement that could be achieved by a perfect oracle.

VI. CONCLUSION

In this paper, we presented a combination of two automatic morphological annotation tools for Hungarian reducing the error rate of the better system by 28.90%. The tools combined, one based on an SMT decoder and the other an HMM-based tool, were found to complement each other well. The combination is based on a machine learning algorithm, but we use a technique that allows the whole training data to be utilized for training all level-1 and level-0 models at the same time. The described combination scheme benefits from utilization of a morphological analyzer during the whole process. The combined system outperforms the known best system for Hungarian, thus it can be used in cases where very high disambiguation accuracy is crucial.

REFERENCES

[1] David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[2] Eric Brill and Jun Wu. Classifier combination for improved lexical disambiguation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, pages 191–195, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[3] Dóra Csendes, János Csirik, and Tibor Gyimóthy. The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora LINC 2004 at The 20th International Conference on Computational Linguistics COLING 2004*, pages 19–23, 2004.

[4] Jan Hajič, Pavel Krbec, Květoň, Karel Oliva, and Vladimír Petkevič. Serial combination of rules and statistics: A case study in Czech tagging. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 268–275. Association for Computational Linguistics, 2001.

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1):10, 2009.

[6] Hans Van Halteren, Jakub Zavrel, and Walter Daelemans. Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational Linguistics*, 27(2):199–229, 2001.

[7] László János Laki. Investigating the Possibilities of Using SMT for Text Annotation. In *SLATE 2012 - Symposium on Languages, Applications and Technologies*, pages 267–283, Braga, Portugal, 2012. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik.

[8] György Orosz and Attila Novák. PurePos – an open source morphological disambiguator. In Bernadette Sharp and Michael Zock, editors, *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*, pages 53–63, Wroclaw, 2012.

[9] Gábor Prószéky. *Industrial applications of unification morphology*. Association for Computational Linguistics, Morristown, NJ, USA, October 1994.

[10] Gábor Prószéky and Attila Novák. Computational Morphologies for Small Uralic Languages. In *Inquiries into Words, Constraints and Contexts.*, pages 150–157, Stanford, California, 2005.

[11] Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of the Fifth conference on International Language Resources and Evaluation*, pages 1670–1673, Genoa, 2006.

[12] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition, 2011.

[13] János Zsibrita, Veronika Vincze, and Richárd Farkas. magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 238–374, Szeged, 2013. Szegedi Tudományegyetem.

Multi-layered Abstractions for an Industrial CFD Application

István Reguly
(Supervisors: András Oláh, Tamás Roska, Mike Giles)
istvan.reguly@itk.ppke.hu

***Abstract*—This work presents a benchmarking, performance analysis of the OP2 "active" library applied to an industrial scale application. OP2 provides an abstraction framework for the parallel execution of unstructured mesh applications, aiming to decouple the scientific specification of the application from its parallel implementation, and thereby achieve code longevity and near-optimal performance through re-targeting the application to execute on different multi/many-core hardware. We present the challenges involved in applying such an abstraction to an industrial application and preliminary performance results on HECToR, up to 4096 cores.**

***Keywords*-Unstructured grid, domain specific language, active library, OP2, OpenMP, GPU, CUDA, computational fluid dynamics**

I. Introduction

Due to the physical limitations to building faster single core microprocessors, the development and use of multi- and many-core architectures has been transforming High Performance Computing (HPC). Increased performance can no longer be achieved through higher clock frequencies, and so after many years the "free" scaling of applications with newer generations of processors is mostly over. The trends show that increased throughput can only be achieved via increased parallelism, modern architectures feature multiple cores on the same piece of silicon; ranging from a few to thousands, depending on the complexity of individual processing elements. At the same time, we see a reduction in clock frequencies to limit power consumption, which is becoming a growing concern.

There is a wide range of many-core hardware used in HPC, and it is not clear which ones suit certain applications the best, much less if there is going to be one most suited for HPC applications in general. A common factor in most of them is importance of vector processing to efficiently use the hardware. On one hand, we have traditional CPUs with large on-chip caches and sophisticated control circuitry, being equipped with more and more cores (up to 10) and growing vector units (SSE/AVX). Accelerator cards, such as GPUs and the Xeon Phi, are becoming common in supercomputers. These devices offer a limited size, very high bandwidth off-chip memory and a processor with 14-64 units, each with its own vector processing cores: the exact ratios vary, but the common factor is that in order to get close to the advertised theoretical performance limit, SIMD (Single Instruction Multiple Data) type processing and certain memory access patterns are required. In the future we can also expect energy-efficient designs from ARM and other companies to be adopted in HPC hardware.

In light of these developments, an applications developer faces a difficult problem. Optimising an application for a target platform requires more and more low-level hardware-specific knowledge and the resulting code is increasingly difficult to maintain. Additionally, adapting to new hardware may require a complete re-write, since optimisations and languages vary widely between different platforms. It can not be expected of domain scientists to gain platform-specific expertise in all the hardware they wish to use. This is especially the case with industrial applications that were developed many years ago and expected to be used for many more years to come. Since these codes frequently consist of tens or hundreds of thousands of lines of code - usually written in Fortran - regularly porting them to new hardware is infeasible. For these industrial applications some of the most important factors are maintainability and longevity.

Active libraries and *Domain Specific Languages* (DSLs) offer a solution to these problems by providing a well-understood abstraction for a specific domain of applications to be used by the application scientists. By doing this, it is possible to separate the implementation details from the abstract description of the computations and generate optimised code for different hardware from the same source. Much research [3], [4] has been done on the multi-layered abstractions approach to developing scientific simulation software. But so far there have been very few examples of the applicability of such frameworks in full scale industrial production applications, demonstrating the expressiveness of the abstraction and its practical viability both in terms of developer productivity and performance.

Unstructured meshes are particularly popular in the domain of computational fluid dynamics (CFD) as they permit correct simulations of highly complex physical problems, by adjusting the resolution of the mesh at the appropriate locations to guarantee accuracy. This implies that the connectivity of the mesh has to be described in an explicit way and computations and data movement have to be carried out through indirections, which may cause issues when using shared memory parallelism. Over the past two years, we have been developing an active library framework, *OP2*, that provides an abstraction for unstructured mesh applications as well as implementations for different, heterogeneous backends. In previous works, we have presented its design and development [4] and its

I. Reguly, "Multi-layered abstractions for an industrial CFD application,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 145-148.

performance in heterogeneous systems [5]. In particular, we have investigated the performance on a standard unstructured mesh finite volume CFD benchmark ("Airfoil") written in C using the OP2 API and parallelised on a range of multi-core and many-core platforms; our results showed considerable performance gains could be achieved on a diverse set of hardware.

This work charts the conversion of an industrial CFD application from Oplus to OP2 and presents key optimisation and development strategies that allowed us to gain near-optimal performance on modern parallel systems. More specifically, we make the following contributions:

1) We present the deployment of OP2, mapping out the key difficulties encountered in the conversion of the legacy application (designed and developed over 15 years ago) with OPlus to OP2. Our work demonstrates the clear advantages in developing future-proof and performant applications through a high-level abstractions approach. This application, to our knowledge, is the first successful industrial application to demonstrate the viability of this research.
2) We analyse the performance of the baseline OP2 CPU implementation on a range of modern HPC platforms, executing on an important standard mesh (NASA ROTOR 37). The most important factors affecting performance are explored: the effect of ordering on cache locality in conjunction with partitioning using ParMetis or PT-Scotch when scaling up to 5000 cores on a large-scale Cray XE6 system. We present new features of the OP2 library that address these issues.
3) The performance of the application is investigated with our shared-memory backends that use source-to-source translation to utilise OpenMP or CUDA over MPI. We discuss optimisations to improve multi-threaded performance as well as performance on NVIDIA Kepler GPUs as compared to previous implementations - purely by making changes to the backend-specific code generators, without having to touch the application source code itself.
4) Finally, we introduce a mixed CPU and GPU backend that makes use of both in order to maximise achieved performance. We discuss the challenges involved in load balancing for such heterogeneous systems and investigate performance on a cluster of 16 NVIDIA Tesla K20 GPUs.

We use highly-optimised code generated through OP2 for all system back-ends, suign the same application code, allowing for direct performance comparison. Re-enforcing our previous findings, this research demonstrates that an application written once at a high-level using the OP2 framework is easily portable across a wide range of contrasting platforms, and is capable of achieving near-optimal performance without the intervention of the application programmer.

II. The OP2 Active Library

OP2, is designed as an *active library* and provides application programmers with the ability to express complex abstractions through an API, analogously to classical software libraries, but with the benefit of compiler support to optimize those abstractions accordingly. An application written once using the OP2 API, which is hosted in C/C++ or Fortran, can be translated using the OP2 source-to-source compiler tools to deliver performance portability across a diverse range of multi-core and many-core architectures. In this section we briefly discuss the OP2 API, its compile and run-time infrastructure as a preamble to presenting the main contributions of this paper.

CFD applications based on unstructured meshes or grids are a key class of applications that require significant computational resources. Often solved across a 3D mesh, these applications repeatedly iterate over a large number of elements forming the mesh to reach the desired solution accuracy of resolution. Unlike unstructured meshes, where regular stencils can be used to describe data access, unstructured meshes use the explicit connectivity between elements during computations. This leads to irregular patterns of data access, usually through indexing arrays.

The OP2 approach to the abstraction of unstructured mesh computations (based on ideas developed in its predecessor OPlus [1]) involves breaking down the problem into four distinct parts: (1) sets, (2) data on sets, (3) connectivity (or mapping) between sets and (4) operations over sets. This leads to an API through which a mesh and operations on the mesh can be completely and abstractly defined.

We present the OP2 programming model and the API through snippets of code of our Airfoil example implemented in Fortran. The programmer first defines the structure of the mesh be declaring sets and the connectivity between the sets as follows:

```
! subroutine op_decl_set(set_size, variable, name)
call op_decl_set(nnode, nodes, 'nodes')
call op_decl_set(nedge, edges, 'edges')
call op_decl_set(nbedge, bedges, 'bedges')
call op_decl_set(ncell, cells, 'cells')

! subroutine op_decl_map(from, to, map_dim, &
!                  & map_data, variable, name)
call op_decl_map(edges, nodes, 2, &
        & edge, pedge, 'edges2nodes')
call op_decl_map(edges, cells, 2, &
        & ecell, pecell, 'edges2cells')
call op_decl_map(bedges, nodes, 2, &
        & bedge, pbedge, 'bedges2nodes')
call op_decl_map(bedges, cells, 1, &
        & becell, pbecell, 'bedges2cells')
call op_decl_map(cells, nodes, 4, &
        & cell, pcell, 'cells2nodes')
```

where *map_data* is an array of size $map\_dim * set\_size$, listing the map_dim number of set elements in to for each set element in $from$. This is followed by the definition of datasets, by specifying the set they are defined on, the dimensionality (i.e. number of state variables per set element

in the given dataset) and passing the array containing the data that is laid out in a similar way maps are.

```
! subroutine op_decl_dat(set, dat_dim, data, variable, name)
call op_decl_dat(nodes, 2, &
            & x, p_x, 'p_x')
call op_decl_dat(cells, 4, &
            & q, p_q, 'p_q')
call op_decl_dat(cells, 1, &
            & adt, p_adt, 'p_adt')
call op_decl_dat(cells, 4, &
            & res, p_res, 'p_res')
```

Given these declarations and definitions, the mesh is fully defined. An alternative to directly passing in arrays containing the values of maps and datasets is to use OP2's HDF5 API, which also enables the application to run in a distributed setting using MPI immediately - otherwise an initial partitioning is required and sizes and data local to each partition have to be passed in to the API. It is important to note that OP2 takes ownership of the map data and the dataset values after these definitions, the programmer only sees opaque handles and any access to them may either fail or return invalid values, and more importantly any direct changes to them that are not carried out through the OP2 API are ignored by OP2 and may lead to an inconsistent state. This is so in order for OP2 to be able to operate in different memory spaces and to apply certain transformations to how the data is laid out.

Operations on the mesh can be described as iterations over a given set, applying a "kernel function" to each element of the set, accessing data either defined on the iteration set or through indirections using maps connecting the iteration set to other sets. A simple example computing the flux residual in Airfoil that iterates over edges, reading and incrementing data on vertices at the two ends of the edge is as follows:

```
subroutine res_calc(x1, x2, q1, q2, &
                    & adt1, adt2, res1, res2)
implicit none
real(8), intent(in) :: x1(2), x2(2), &
         & q1(4), q2(4), adt1, adt2
real(8), intent(inout) :: res1(4), res2(4)
...
end subroutine
...
call op_par_loop_8 ( res_calc, edges, &
& op_arg_dat(p_x, 1, pedge, 2, OP_READ), &
& op_arg_dat(p_x, 2, pedge, 2, OP_READ), &
& op_arg_dat(p_q, 1, pecell, 4, OP_READ), &
& op_arg_dat(p_q, 2, pecell, 4, OP_READ), &
& op_arg_dat(p_adt, 1, pecell, 1, OP_READ), &
& op_arg_dat(p_adt, 2, pecell, 1, OP_READ), &
& op_arg_dat(p_res, 1, pecell, 4, OP_INC), &
& op_arg_dat(p_res, 2, pecell, 4, OP_INC))
```

A dataset argument follows the Access/Execute specification (AEcute [2]) and fully describes the access to a dataset by indicating the following:

1) Access method of the corresponding parameter. The value may be read, overwritten, incremented or read-and-written. This implies certain constraints on execution to avoid race conditions when using shared-memory parallelism.
2) If the data is accessed through an indirection, by specifying the mapping from the execution set to the target set, as well as indicating which element of the relation should be used (e.g. which vertex at the two ends of the edge).

A. OP2 Compilation and Run-Time Support

The OP2 library handles the architecture-specific support through header files, code parsing and source-to-source generation. An OP2 application written using the "sequential" header file can be compiled using conventional compilers (e.g. gcc, icc etc.), linked against the OP2 sequential backend and be executed or debugged as a serial application, which enables quick and easy development. When using HDF5 to define sets, maps and datasets the exact same source code can be linked against the OP2 pure MPI backend, and executed in a distributed system. This MPI backend already provides near-optimal performance on CPUs, making use of techniques such as latency hiding and partitioning using ParMetis or PT-Scotch.

To enable execution with OpenMP or CUDA, OP2 uses source-to-source translation: because of the expressiveness of the OP2 API, it is sufficient to parse the calls to the API, since all access types are described. This enables our "compiler", written in Python, to easily generate target-specific source code which can be then compiled using the appropriate compiler (e.g. nvcc for CUDA) and linked against the corresponding OP2 backend. Our build system is described in Figure 1, which shows the multi-layered nature of our library. Race conditions that occur during shared-memory execution are handled through multiple levels of colouring in conjunction with the creation of an execution plan that is reused between executions of the same parallel loop that use the same arguments. More details on this can be found in our previous papers [4], [5]

OP2's general decomposition of unstructured mesh algorithms, imposes no restrictions on the actual algorithms, it just separates the components of a code. However, OP2 makes an important restriction that the order in which elements are processed must not affect the final result, to within the limits of finite precision floating-point arithmetic. This constraint allows OP2 to choose its own order to obtain maximum parallelism, which on platforms such as GPUs is crucial to gain good performance.

The straightforward programming interface combined with efficient parallel execution makes it an attractive prospect for the many algorithms which fall within the scope of OP2. For example the API could be used for explicit relaxation methods such as Jacobi iteration; pseudo-time-stepping methods; multi-grid methods which use explicit smoothers; Krylov subspace methods with explicit preconditioning; semi-implicit methods where the implicit solve is performed within a set member, for example performing block Jacobi where the block is across a number of PDE's at each vertex of a mesh.

Currently, OP2 supports execution on a single-threaded CPU, a single SMP system based on multi-core CPUs using OpenMP, a single NVIDIA GPU using CUDA, a cluster of CPUs using MPI as well as heterogeneous settings: MPI

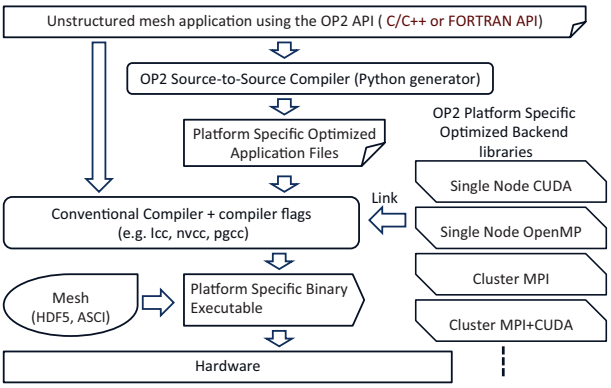


Fig. 1. OP2 build hierarchy

combined with OpenMP or CUDA. In this paper we also introduce a hybrid execution model that utilises CPUs and GPUs at the same time.

III. Baseline Performance Results

The performance of various implementations was benchmarked on a range of hardware, from here on, we refer to them by their name. Ruby, our development machine, was used to investigate scaling within node, the other systems are considered on a per node basis in our figures. Our tests use different variations of the NASA Rotor 37 mesh, at different resolutions and using one or two blade passages. The standard practice for simulations is to assign on the order of a million vertices to a node, hence our basic test case consists of 0.8 million vertices and 2.5 million edges. We use this mesh to investigate strong scaling.

A. Strong scaling

Strong scaling has been an important metric of performance scalability for a long time, it essentially investigates how much faster the same problem can be solved, given more resources. While important to get results quickly, it is not of primary interest to users, to whom weak scaling is more significant.

Figure 2 presents strong scaling figures on the HECToR supercomputer. It clearly shows a wide performance gap between OPlus and the baseline OP2 implementations, and a fairly stable strong scaling at 70

The reason for the performance difference between the OP2 and the OPlus implementations is the ordering of set elements within the partitions; while OPlus performs a sort by coordinates, OP2 by default uses the ordering that it was given during initialisation. Applying a reordering algorithm has a profound effect on performance, as figure 2 shows, the block partitioned version scales very steadily to a high number of cores, outperforming OPlus beyond 512 cores. A strange interaction between the reordering scheme and the ParMetis KWAY partitioner can be observed as well, when the number of partitions is low, OP2 marginally outperforms OPlus, however as the number of partitions increases there is dramatic decrease in scaling. The reason for the slowdown is that ParMetis only optimises for minimal edge-cut, the average

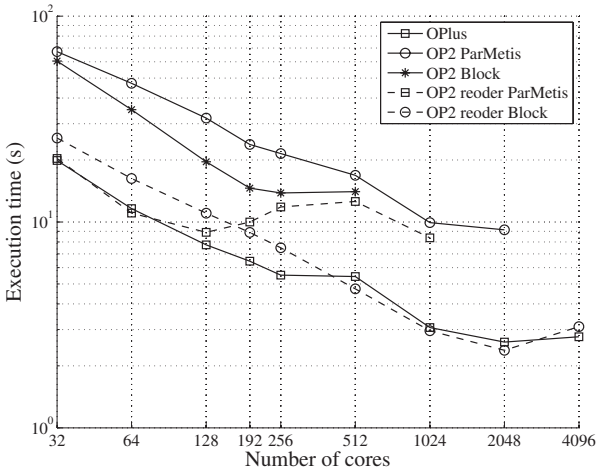


Fig. 2. Strong scaling of Mesh #1 on HECToR

number of neighbors each partition has goes up, and the high number of small messages prevent scaling.

IV. Conclusion

This work shows the viability of domain specific languages in an industrial setting and shows that once the application is converted to using the OP2 API, it enables near-optimal performance on a range of heterogeneous hardware, without any changes to the application. We investigated strong scaling on a supercomputer up to 4096 cores and explored performance bottlenecks.

Acknowledgment

The author has to acknowledge TÁMOP-4.2.1./B-11/2/KMR-2011-002 and TÁMOP-4.2.2./B-10/1-2010-0014. Funding has come from the UK Technology Strategy Board and Rolls-Royce plc. through the Siloet project, the UK Engineering and Physical Sciences Research Council projects EP/I006079/1, EP/I00677X/1 on Multi- layered Abstractions for PDEs and the Natural Environment Research Council project NE/G523512/1.

References

[1] Crumpton, P. I. and Giles, M. B. (1996) Multigrid aircraft computations using the OPlus parallel library. Parallel Computational Fluid Dynamics: Implementations and Results Using Parallel Computers, 339-346, A. Ecer, J. Periaux, N. Satofuka, and S. Taylor, editors, North-Holland.

[2] Howes, L. W., Lokhmotov, A., Donaldson, A. F., and Kelly, P. H. J. (2009) Deriving efficient data movement from decoupled access/execute specifications. Seznec, A., Emer, J., OBoyle, M., Martonosi, M., and Ungerer, T. (eds.), High Performance Embedded Architectures and Compilers, vol. 5409 of Lecture Notes in Computer Science, pp. 168182, Springer Berlin/Heidelberg.

[3] DeVito, Z., Joubert, N., Medina, M., Barrientos, M., Oakley, S., Alonso, J., Darve, E., Ham, F., and Hanrahan, P., Liszt: Programming mesh based pdes on heterogeneous parallel platforms. Presentation given by the Stanford PSAAP Center, Oct 2010 http://psaap. stanford.edu.

[4] M.B. Giles, G.R. Mudalige, Z. Sharif, G. Markall, P.H.J. Kelly. (2011) Performance Analysis and Optimization of the OP2 Framework on Many-Core Architectures. The Computer Journal. ISSN 0010-4620

[5] G.R. Mudalige, I. Reguly, M.B. Giles, C. Bertolli and P.H.J. Kelly. OP2: An Active Library Framework for Solving Unstructured Mesh-based Applications on Multi-Core and Many-Core Architectures. In Proceedings of Innovative Parallel Computing (InPar), 2012, pp.1-12, 13-14 May 2012.

Improved optimization methods for efficient chemical network structure computation

János Rudan
(Supervisor: Gábor Szederkényi, Katalin M. Hangos)
rudan.janos@itk.ppke.hu

***Abstract*—In this report linear programming (LP)-based algorithms are presented to compute alternative realizations of biochemical reaction networks (CRNs) with mass action kinetics. The main new contributions are the following: firstly, an LP-based method having polynomial time complexity is presented that is guaranteed to compute the dense super-structure of a CRN, and secondly, it is shown that dynamically equivalent sparse structures can be computed efficiently and precisely by applying the theory of sparse nonnegative solutions of under-determined linear systems. Sections III and IV of this report are essentially a short summary of the recently submitted paper [8] where the methodological details and the analysis of results can be found.**

***Keywords*-chemical reaction networks; dynamical equivalence; optimization**

I. Introduction

The rigorous structural and dynamical analysis of biologically motivated kinetic systems such as intracellular signalling pathways and gene regulation networks has gained an increased attention in recent years. This naturally imply a need for the parallel improvement of modeling and computational methods to be able to handle the growing amount of data and to analyse more complex, possibly biologically relevant processes and networks.

In this report, by Chemical Reaction Networks (CRNs) we mean deterministic kinetic systems obeying the mass action law. It is known that such systems form a wide class of smooth nonlinear systems that are able to produce all important qualitative phenomena in nonlinear dynamics such as oscillations, multiplicities and even chaos [17].

It has been known at least from the 1970's that structurally/parametrically different reaction schemes might produce exactly the same dynamics in the concentration space. This phenomenon is usually called *macro-equivalence* [6] or *dynamical equivalence*. For the computation of dynamically equivalent structures with preferred properties such as detailed or complex balance, (weak) reversibility, the inclusion of minimal/maximal number of reactions or complexes, optimization-based procedures have been published in [10], [12], [13], [14]. Biologically relevant examples for the structural non-uniqueness of CRNs were shown in [11].

Linear programming (LP) can be defined as the problem of minimizing (or maximizing) a linear function with respect to linear constraints, where all the variables are real-valued. Currently applied solution methods usually guarantee the polynomial-time solution of the LP problems. Mixed integer linear programming (MILP) is different from LP in that some of the decision variables are integer valued. The emerging MILP problem is a combinatorial optimization problem which is generally NP-hard.

In [10] the so-called *dense* and *sparse* dynamically equivalent realizations of CRNs were defined containing the maximal and minimal number of reactions, respectively. Using the fact that certain propositional logic problems can be transformed into MILP problems by introducing appropriate logical variables [7], [1], the computation of dense and sparse realizations were straightforwardly written in an MILP framework in [10]. However, the original MILP approach seriously limits the network size that can be treated computationally within a reasonable time interval.

Therefore, the aim of this work is to propose and analyse computationally efficient methods preferably not containing integer variables for determining dynamically equivalent dense and sparse realizations.

II. Structural and dynamical description of CRNs

In this Section, the general notations and definitions for representing CRNs are introduced. The applied notations are based on the introductory parts of [10], [12], [13].

A. Basic notions

We consider CRNs as closed deterministic chemical systems under isothermal and isobaric conditions. The fundamental elements of the system are the so-called chemical species X_i, $i = 1, ..., n$ taking part in r reactions that obey the mass action law. The state vector is built up from the concentrations of the species, i.e. $x_i = [X_i]$, $i = 1, ..., n$ the values of which are nonnegative.

Complexes are formally linear combinations of the species, i.e. $C_m = \sum_{i=1}^{n} \alpha_{ij} X_i$ for $j = 1, ..., r$. An elementary reaction step between the source and product complexes, $C_i = \sum_{i=1}^{n} \alpha_{ij} X_i$ and $C_j = \sum_{i=1}^{n} \beta_{ij} X_i$, respectively, is denoted by

$$\sum_{i=1}^{n} \alpha_{ij} X_i \rightarrow \sum_{i=1}^{n} \beta_{ij} X_i, \;\; j = 1, ..., n \qquad (1)$$

In eq. (1), α_{ij} and β_{il} are the stoichiometric coefficients of the source and product complexes, respectively. In the

J. Rudan, "Improved optimization methods for efficient chemical network structure computation,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 149-152.

models we consider, stoichiometric coefficients are nonnegative integers.

According to the mass action law, the reaction rate corresponding to reaction (1) is given by

$$\rho_{ij}(x) = k_{ij} \prod_{i=1}^{n} x_i^{\alpha_{ij}}, \tag{2}$$

where $k_{ij} > 0$ is the reaction rate coefficient.

It can happen that both reactions $C_i \to C_j$ and $C_j \to C_i$ are present in the network. In this particular framework, these reactions are handled as separate elementary reactions.

B. Graph representation

We can assign the following directed graph to the reaction network: the directed graph $D = (V_d, E_d)$ consists a finite nonempty set V_d of vertices and finite set of E_d of ordered pairs of distinct vertices called directed edges. The vertices represent the complexes: $V_d = \{C_1, ...C_m\}$ and the edges stand for the reactions: $(C_i, C_j) \in E_d$ if complex C_i is transformed to C_j in one of the reactions in the network. The reaction rates appear as nonnegative weights on the edges in the directed graph.

By the structure of a given CRN we mean the unweighted graph of the reaction network.

C. ODE-based description

CRN systems can be represented with differential equations and vica versa. In this report - similar to [10] - the following representation is used which was defined in [5].

$$\dot{x} = Y \cdot A_k \cdot \psi(x) \tag{3}$$

where $x \in \mathbb{R}^n$ is the vector of the concentrations of the species, $Y \in \mathbb{R}^{n \times m}$ is the matrix containing the stoichiometric coefficients of the complexes, $A_k \in \mathbb{R}^{m \times m}$ contains the structure of the weighted directed graph of the CRN and $\psi : \mathbb{R}^n \to \mathbb{R}^m$ is a vector mapping defined by:

$$\psi_j(x) = \prod_{i=1}^{n} x_i^{Y_{ij}}, \;\; j = 1, ..., m \tag{4}$$

The structure of matrix Y is the following: the ith column of Y contains the composition of complex C_i, so the Y_{ij} value is the stoichiometric coefficient of C_i corresponding to the specie X_j.

The matrix A_k has the following elements:

$$[A_k]_{ij} = \begin{cases} -\sum_{l=1, l \neq i}^{m} k_{il} & if\, i = j \\ k_{ji} & if\, i \neq j \end{cases} \tag{5}$$

It should be noted that the sum of the elements in each column is zero, hence A_k is often called as the Kirchhoff matrix of the CRN.

Using the notation

$$M = Y \cdot A_k, \tag{6}$$

Eq. (3) becomes

$$\dot{x} = M \cdot \psi(x) \tag{7}$$

where M contains the coefficients of the monomials in the polynomial ODE describing the time-evolution of the state variables.

D. Dynamically equivalent realizations of CRNs

The matrix pair (Y, A_k) is called a realization of a CRN described by M if the following conditions are fulfilled: Eq. (6) is fulfilled while all the elements of Y are nonnegative integers, and A_k is a column conservation matrix having nonpositive diagonal and nonnegative off-diagonal elements.

$$M = Y^1 \cdot A_k^1 = Y^2 \cdot A_k^2 \tag{8}$$

is fulfilled, then (Y^1, A_k^1) and (Y^2, A_k^2) are called dynamically equivalent.

E. Optimization framework for computation of dynamically equivalent structures

The classical method of computing alternative realizations is based on the natural formulation of the problem as a MILP problem [9]. However, some alternative LP-based methods (see e.g. [11] or [16]) have also appeared in the literature recently.

1) Mixed Integer Linear Programming approach: The detailed description of the LP-problem formulation and MILP-problem formulation can be found in [9].

Let us formulate the CRN alternative realization problem as an MILP. This formulation is not only used in this algorithm but it has also a crucial role in one of the new LP-based methods presented in this report (see Section III-B). As it is described in Section II-C, for a reaction network containing m complexes A_k is the following:

$$A_k = \begin{bmatrix} -a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & -a_{22} & & a_{2m} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & -a_{mm} \end{bmatrix} \tag{9}$$

Now in addition to Eq. (6). we are able to formulate the following constraints:

$$\sum_{i=1}^{m} [A_k]_{ij} = 0, \quad j = 1, \dots, m \tag{10}$$

$$[A_k]_{ij} \geq 0, \quad i, j = 1, \dots, m, \quad i \neq j \tag{11}$$

$$[A_k]_{ii} \leq 0, \quad i = 1, \dots, m, \tag{12}$$

Let us call Eq. (6). and Eq. (10)-(12) altogether as *kinetic constraints* because they clearly characterize the dynamic behaviour of the network in a kinetic form and they can be considered as a constraint set in an optimization problem.

From now we are able to formulate the MILP structure based on [10].

It should be noted that solution of a large size MILP problem is usually a heuristic-driven, computationally very intensive task. This means that MILP problems can not

scale well to networks which are large enough to describe real-life systems.

2) Computing sparse and dense realizations with LP-based techniques: In [16] an LP-based algorithm is presented to search for a sparse realization of gene regulation networks. The proposed iterative method can be exactly fitted to our problem. The algorithm uses an iterative method to approximate the edge weights of the directed graph representing the CRN. The method is quite efficient: as it is mentioned in [16], the algorithm usually doesn't need more than a few iterations. In the following, we will refer to this algorithm as *Iterative LP*.

To find the dense realization of a CRN, the following method can be found in [11]. The method uses $m \cdot (m-1)$ LP computation steps to generate the result. We will use this method for comparison with our novel approach. In the following, we will refer to this algorithm as *Element-wise LP*.

III. Improved methods for computing CRN structures

In this Section two new methods will be proposed that are fast, LP-based techniques to generate sparse and dense realizations of CRNs.

A. Computing sparse realizations

The results in [4] show that in case of large size, undetermined system of linear equations formulated as $Ax = b$, the l^1-norm based minimization can produce a sparse solution out of the infinitely many possible solutions of the problem. This stands if the solution vector x with length n is sparse enough: it should not have more than $\rho \cdot n$ nonzero elements. In [3] an empirical limit $\rho = 0.3$ is suggested.

This method is applied in our case as follows. Recall that in a given CRN there are n species and m complexes. The problem matrices were built up as the *kinetic constraints*. For a single column of A_k which contains m elements the number of kinetic constrains will be $n+1$. Hence, in case of $n+1 < m$ the emerging equality-type constraints as a system of linear equations remains undetermined.

Therefore, a column-wise l^1-norm based minimization is completed and the resulting vectors are considered as the column vectors of the sparse realization of the CRN:

$$\min \sum_{i=1}^{m} abs([A_k]_{i,j}) \; for \; \forall j = 1, \dots m \tag{13}$$

Let us denote the ratio of non-zero and zero elements in $[A_k]_{.,j}$ as τ. If $\tau < \rho \cdot m$ the minimization will successfully find the sparse solution.

In the following, we will refer to this algorithm as the l^1-norm based algorithm.

B. Computing dense realizations

The proposed method is based on the construction of the MILP problem formulation given in sub-section II-E1. The main idea was to formulate the problem by relaxing the integral constraints from the MILP problem and then solve the remaining LP problem. By maximizing the sum of the real valued auxiliary variables, the number of nonzero edge weights are forced to be maximal in the weighted graph of the CRN.

The problem matrices were built up based on the *kinetic constraints*. Similarly to the MILP case, a set of auxiliary variables were defined: a σ_i variable for all state variable x_i. All these auxiliary variables are real valued. The original constraints were formulated as $\sigma_i = 1 \longleftrightarrow x_i > \epsilon_e$ where ϵ_e is a minimal (naturally positive) edge weight under which the edge is excluded from the network. After the relaxation of the integrality constrains the formula becomes:

$$\epsilon \cdot \sigma_i \leq x_i \tag{14}$$

where $\sigma_i \in [0; 1]$ and $\epsilon > 0$ is a sufficiently small number, but $\epsilon > \epsilon_e$. By considering ϵ as a scaling factor, we obtain after short reformulation:

$$-x_i + \sigma_i \leq 0 \tag{15}$$

$$\sigma_i \in [0; \epsilon] \tag{16}$$

where Eq. (15) is formulated as a constraint and Eq. (16) is formulated as lower and upper bounds in the LP problem.

Now the optimization problem is defined as follows:

$$\max \sum_{i=1}^{m} \sigma_i \tag{17}$$

$$\begin{cases} Y \cdot A_k = M & \\ \sum_{i=1}^{m} [A_k]_{i,j} = 0 & j = 1, \dots m \\ 0 \leq \sigma_i \leq \epsilon & i = 1, \dots m \\ -x_i + \sigma_i \leq 0 & i = 1, \dots m \end{cases} \tag{18}$$

In the following, we will refer to this algorithm as the *LP-MAX* algorithm.

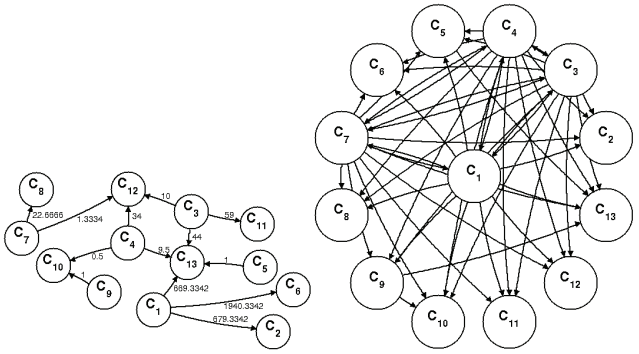
IV. Results

In this report, two new algorithms were presented to compute alternative realizations of CRNs. Both methods were involved in a comparative study in which extensive simulations were completed to evaluate the performance of the proposed methods. Large scale problems were investigated to present the capability of dealing with biologically relevant problems, too.

A. Computing alternative realizations of a CRN describing a classical 3-dimensional Lorenz system

Using the 3-dimensional Lorenz system introduced in Section II-D, the validation of the proposed methods on small-scale CRNs can be completed. It is known from [15] that this system can be represented by CRNs having 3 species and 13 complexes with sparse realizations containing 13 off-diagonal non-zero elements, and its dense realization contains 51 off-diagonal non-zero elements.

Computing the sparse realization takes 17.172 seconds in case of the MILP based method. *Iterative LP* were able to find a valid sparse realization in 0.0144 seconds while



(a) A sparse realization (b) The dense realization

Figure 1: Different, dynamically equivalent realizations of the Lorenz-system.

the proposed l^1-norm based sparse search can complete this task in 0.0166 seconds.

The dense realization was also successfully computed by all the three algorithms. The MILP based algorithm can complete it in 5.3477 seconds, *Element-wise LP* consumes 0.2524 seconds and the *LP-MAX* algorithm needs 0.0446 seconds to compute the solution.

Two different, dynamically equivalent realization of the Lorenz-system can be seen in Fig. 1.

B. Computing alternative realizations of the ErbB network

As it was mentioned before, we would like to use our methods to study the possible structures of large scale, biologically relevant networks. As a case study the *ErbB* network described in [2] was investigated.

In our representation the *ErbB* signalling pathway model consists of 504 species, 1082 complexes and 1654 reactions. The model description was originally a sparse representation. With the help of the *LP-MAX* algorithm introduced in Section III-B the dense realization is computed. It contains 1683 reactions: 29 mathematically possible extra reactions compared to the published model originating from 15 different complexes. The overall computational time was 4993 seconds.

The sparse realization was also extracted from the dense realization with the help of the l^1-norm based sparse search. The resulting network had the same structure as the original sparse representation. The computational time was around 430 seconds.

V. Conclusion

In this work linear programming based methods were presented to compute alternative realizations of CRN. By analysing the properties of the system model and the MILP-based description, simplified algorithms were developed which have polynomial complexity. The algorithms can be easily applied in a parallel framework, too.

Acknowledgements

This research has been supported by the Hungarian National Research Fund through grant NF104706. The first and second authors were also supported by the projects TÁMOP-4.2.1./B-11/2/KMR-2011-002 and TÁMOP-4.2.2./B-10/1-2010-0014.

References

[1] A. Bemporad and M. Morari. Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35:407–427, 1999.

[2] William W. Chen, Birgit Schoeberl, Paul J. Jasper, Mario Niepel, Ulrik B. Nielsen, Douglas A. Lauffenburger, and Peter K. Sorger. Input-output behavior of erbb signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular Systems Biology*, 5, January 2009.

[3] David L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math*, 59:797–829, 2004.

[4] David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.

[5] M. Feinberg. *Lectures on chemical reaction networks.* Notes of lectures given at the Mathematics Research Center, University of Wisconsin, 1979.

[6] F. Horn and R. Jackson. General mass action kinetics. *Archive for Rational Mechanics and Analysis*, 47:81–116, 1972.

[7] R. Raman and I.E. Grossmann. Modelling and computational techniques for logic based integer programming. *Computers and Chemical Engineering*, 18:563–578, 1994.

[8] J. Rudan, G. Szederkényi, and K. M. Hangos. Efficiently computing alternative structures of large biochemical reaction networks using linear programming. *MATCH Commun. Math. Comput. Chem.*, 2013. submitted.

[9] Thomas L. Magnanti Stephen P. Bradley, Arnoldo C. Hax. *Applied Mathematical Programming.* Addison-Wesley, 1977.

[10] G. Szederkényi. Computing sparse and dense realizations of reaction kinetic systems. *Journal of Mathematical Chemistry*, 47:551–568, 2010.

[11] G. Szederkényi, J. R. Banga, and A. A. Alonso. Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC Systems Biology*, 5:177, 2011.

[12] G. Szederkényi and K. M. Hangos. Finding complex balanced and detailed balanced realizations of chemical reaction networks. *Journal of Mathematical Chemistry*, 49:1163–1179, 2011.

[13] G. Szederkényi, K. M. Hangos, and T. Péni. Maximal and minimal realizations of reaction kinetic systems: computation and properties. *MATCH Commun. Math. Comput. Chem.*, 65:309–332, 2011.

[14] G. Szederkényi, K. M. Hangos, and Zs. Tuza. Finding weakly reversible realizations of chemical reaction networks using optimization. *MATCH Commun. Math. Comput. Chem.*, 67:193–212, 2012.

[15] Z. A. Tuza, G. Szederkényi, K. M. Hangos, and J. R. Banga A. A. Alonso. Computing all sparse kinetic structures for a Lorenz system using optimization methods. *International Journal of Bifurcation and Chaos*, accepted:to appear, 2013.

[16] Michael M. Zavlanos, A. Agung Julius, Stephen P. Boyd, and George J. Pappas. Inferring stable genetic networks from steady-state data. *Automatica*, 47(6):1113–1122, 2011.

[17] P. Érdi and J. Tóth. *Mathematical Models of Chemical Reactions. Theory and Applications of Deterministic and Stochastic Models.* Manchester University Press, Princeton University Press, Manchester, Princeton, 1989.

Complex Electrophysiological Analysis of the Effect of Cortical Electrical Stimulation in Humans

Emília Tóth
(Supervisor: Dr. István Ulbert)
totem@digitus.itk.ppke.hu

***Abstract*— Electrical stimulation is frequently performed in concurrence with electrocorticogram recording for functional mapping (or electrical stimulation mapping-ESM) of the cortex and identification of critical cortical structures. In medically refractory epilepsy surgical candidates, intracranial electrodes are necessary to localize the epileptogenic focus prior to surgical resection. This electrodes are used to record the underlying brain activity and also for electrical stimulation of the cortex. Electrical stimulation mapping (ESM) is the gold standard for identifying functional and pathological areas of the brain. Although the procedure remains unstandardized, and limited data support its clinical validity nevertheless, electrical stimulation mapping for define language areas has likely minimized postoperative language decline in numerous patients, and has generated a wealth of data elucidating brain-language relations [3]. Our aim was to study another way of cortical stimulation, so called single pulse electrical stimulation (SPES) to map pathological and functional networks in the brain.**

***Keywords*-component; biomedical signal processing, electrodes, brain networks, electrocorticography, epilepsy, in vivo, human**

***Abbreviations*- ESM=electrical stimulation mapping; SPES=single pulse electrical stimulation; CT=computed tomography; DCES= direct cortical electrical stimulation; CCEP=cortico-cortical evoked potential; BA=Brodmann area; ROC curves=receiver operating characteristic curves**

INTRODUCTION

Mapping of functional areas in the human brain is crucial in epilepsy and tumor surgery. There are several non-invasive methods to identify eloquent cortices, such as functional Magnetic Resonance Imaging (fMRI) or Positron Emission Tomography (PET), but the gold standard is direct high frequency cortical electrical stimulation. In this study we used single pulse electrical stimulation evoked late responses to map language and motor networks and to better understand the electrophysiological mechanisms of the cortico-cortical evoked potentials.

Single pulse electrical stimulation is a new method to investigate the cortico-cortical connections in vivo in the human language, motor and sensory system which can provide insight into the mechanisms of higher-order cortical functions and the connections between functional areas [1]. When using a crown configuration, a handheld wand bipolar stimulator may be used at any location along the electrode array. However, when using a subdural strip, stimulation must be applied between pairs of adjacent electrodes due to the nonconductive material connecting the electrodes on the grid. Electrical stimulating currents applied to the cortex are relatively low, between 2 to 4 mA for somatosensory stimulation, and near 15 mA for cognitive stimulation. The functions most commonly mapped through DCES are primary motor, primary sensory, and language. The patient must be alert and interactive for mapping procedures, though patient involvement varies with each mapping procedure. Language mapping may involve naming, reading aloud, repetition, and oral comprehension; somatosensory mapping requires that the patient describe sensations experienced across the face and extremities as the surgeon stimulates different cortical regions.[2]

High frequency electrical stimulation is the gold standard in neurosurgery for mapping brain functions, but the exact mechanisms behind the effect and parameters used need to be further studied. There is also some risk associated with the stimulation, due to its proepileptic effect and the limits imposed by the fact that the cortex has to be exposed using some type of surgery.

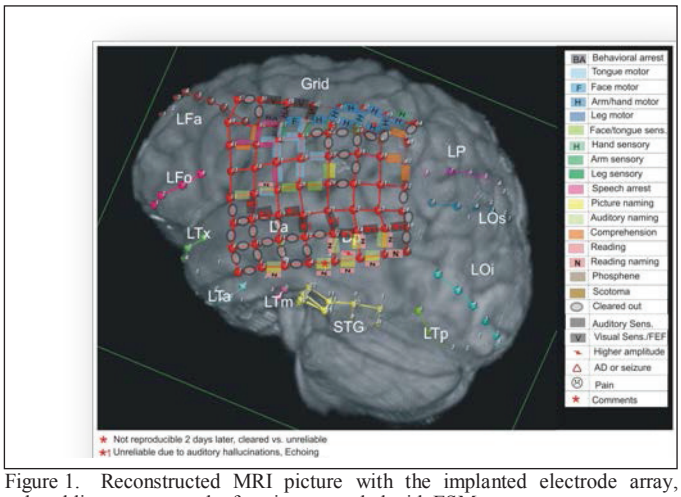


Figure 1. Reconstructed MRI picture with the implanted electrode array, colored lines represent the functions revealed with ESM.

During the development of epilepsy, the connections between nerve cells are also strengthen or weakening because of various reasons (neuronal cell death, proliferation, brain stem injury, etc). We hypothesized, this cause changes in the number of significant evoked potentials between the areas showing epileptic activity compared to other areas.

E. Tóth, "Complex electrophysiological analysis of the effect of cortical electrical stimulation in humans,"
in *Proceedings of the Interdisciplinary Doctoral School in the 2012-2013 Academic Year*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2013, vol. 8, pp. 153-156.

Our aim with this study was to find other ways to map functional networks in the brain, using a less invasive method and analyze the network features with this new approach. Single pulse electrical stimulation (0,5Hz) is much less invasive in terms of seizure generation, and the distribution of the evoked potentials may reveal the intracortical pathways between cortical regions.

METHODS

A. *Clinical electrodes and recordings*

The electrode implantations and recordings, along with ESM and SPES took place at two well established epilepsy surgical centers in Budapest (National Institute of Neuroscience) and New York (North Shore-LIJ Health System). Patients were implanted with intracranial subdural grid, strip, and in some cases depth electrodes for 5– days. They were monitored to identify the seizure focus, at which time the electrodes were removed and, if appropriate, the seizure focus was resected. Continuous intracranial EEG was recorded with standard recording systems with sampling rates 1000 or 2000 Hz. The microelectrodes were implanted in eleven cases, perpendicularly to the cortical surface to sample the width of the cortex. This 24 contact laminar electrode has been described previously [4]. Differential recordings were made from each pair of successive contacts to establish a potential gradient across the cortical lamina.

B. *Functional Stimulation Mapping*

For localization of functional cortical areas, electrical stimulation mapping was carried out according to standard clinical protocol (bipolar stimulation: 2–5 s, 3–15 mA, 20–50 Hz). Areas were defined as expressive language sites when stimulation resulted in speech arrest. When stimulation resulted in a naming deficit based on auditory or visual cues, or an interruption in reading or comprehension, the area was deemed a nonexpressive language site. Sensory and motor areas were identified when stimulation caused movement or changes in sensation.

C. *Cortical Electrical Stimulation and Cortico-Cortical Evoked Potentials.*

Following implantation of intracranial electrodes, patients were monitored for epileptic activity and during this time, CCEP mapping was performed using single-pulse stimulation. Systematic bipolar stimulation of each pair of adjacent electrodes was administered with single pulses of electrical current (3 mA-15 mA, 0.5 Hz, 0.2-ms pulse width, 20-25 trials per electrode pair). The associated evoked responses (CCEPs) were measured at all other electrode sites. The current amplitude of 10 mA activated the maximal number of neuronal elements without epileptic afterdischarges or other clinical signs. The 2 seconds interstimulation interval was used to minimize the effect of overlapping evoked responses and to leave enough restitution time for the cortex. Patients were awake and at rest at the time of CCEP recording

D. *Analysis of CCEPs.*

Electrophysiological data analyses were performed using Neuroscan Edit 4.5 software (Compumedics) and own developed MATLAB scripts. Evoked responses to stimulation were divided into 2-s epochs (-500 ms to 1,500 ms) time-locked to stimulation pulse delivery. The CCEP consists of two usually negative peaks termed N1, timed at ~10–30 ms, and N2, which exhibits a broader spatial distribution and occurs between 70 and 300 ms [1]. To quantify the magnitude of the CCEPs in the time window of the N2, the data were low-pass filtered (30 Hz), and baseline correction (−450 to −50 ms) was performed. The SD was computed for each electrode separately using all time points in the -450 to -50 time window, CCEPs were considered significant if the N2 peak of the evoked potential exceeded the baseline amplitude by a threshold of ±6 SD as determined from the receiver operating characteristic (ROC) curves.

E. *Electrode localization*

To co-register the electrodes to anatomical structures, we used sophisticated imaging techniques, developed by our co-operational research team. We used intraoperative pictures and a postoperative CT scan to localize the electrodes in the skull. This was co-registered to a high resolution preoperative MRI where we could precisely localize the anatomical structures. Using these scans and freely available softwares (Bioimagesuite, Freesurfer, FSL, AFNI) we developed a semi – automated co-localizing each electrode to the underlying Brodman area of the brain. Determination of the seizure onset zone was performed by epileptologists [5].

F. *Patients*

Twentyfive patients (ages 6–53 years, 28±14.84, 14 females) with medically intractable focal epilepsy were enrolled in the study after informed consent was obtained. These procedures were monitored by local Institutional Review Boards, in accordance with the ethical standards of the Declaration of Helsinki.

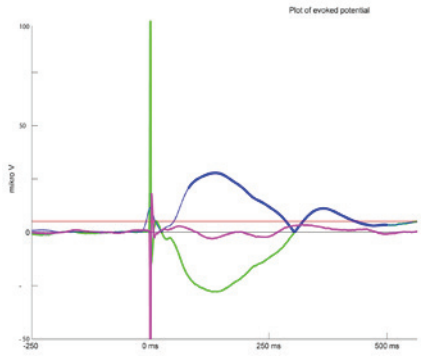


Figure 2. This figure shows averaged responses time locked to the bipolar stimulation artefact (-250-600ms). Green line is the significant response, blue line is the absolute value of the significant response, pink line is a non significant response and the red horizontal line is the threshold for the two responses.

G. *Pathological and physiological networks*

Neurologist defined pathological and non pathological electrode groups. Pathological electrodes were those which showed seizure onset, or early spread (in the first 10 s) The number of significant evoked potentials was divided into four groups, according to pathological or non pathological classification of the stimulate and the recording electrode.

In addition, two type of seizure spreading mechanism were distinguished, according to the consistency of seizure spread. Consistent seizure spread is when seizure starts always at same places, inconsistent if there were more than one typical seizure spreading mechanism. Network connections were examining from these two aspects.

RESULTS

H. Analysis of the significant signal features.

Due to the artifact caused by the stimulation we only focused on the N2 response, which seemed very reliable and reproducible. The variance of both time and amplitude of the N2 peak was high, but it the largest number of peaks occurred around 150 ms, and showed quasi-normal distribution, with two smaller deflections at around 180 -190 ms and 210-250ms. Analysis of 892 peaks, the average latency was 152.84 ms, with 58.7 ms standard deviation.

I. Create a graph.

A significant evoked response indicates the relationship between the electrodes which were stimulated and which showed the significant response. Significant CCEPs were converted to a distance matrix and transformed to a graph using multidimensional scaling (a toolbox from Matlab)

On the one hand the result shows that the functional areas which are close to each other are tightly connected (above somatosensory cortex BA40, BA3, BA2; visual cortex BA17, BA18, BA19 and motor cortex BA6, BA4). On the other hand, those regions which are physically more distant from each other seemed also connected, such as Broca's (BA 45) and Wernicke's (BA 21, BA20, BA22) area. Using this methodology we tried to map as many areas of the brain as possible, to be able to map all the connections between regions which were covered with electrodes.

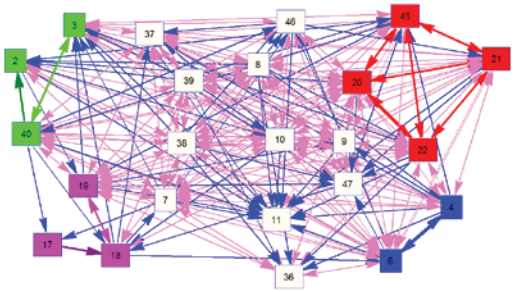


Figure 3. Significant CCEPs were converted to a distance matrix and transformed to a graph using multidimensional scaling. Numbers in squares represent Brodmann areas and lines represent connections. Functional networks are color coded: green sensory, pink visual, red language, blue motor. Lines color coded: thin light pink bidirectional, thin blue unidirectional, darker lines between the elements of functional networks is unidirectional, same color is bidirectional. Stimulating electrodes over Broca's area showed significant responses in electrodes part of the language network as defined with functional stimulation mapping. Responses to stimulation of the primary motor cortex revealed connections to major hubs involved in motor processing.

J. Analysis of changes taking place in cortical layers.

After processing the data from the laminar microelectrode and the implanted macroelectrodes, it can be concluded that after the stimulus, there is a decrease in the power of 15-100 Hz frequency band, and the stimulus elicit deactivation in the middle cortical (3th-5th) layers. This finding is in correlation with previous animal studies, which showed wide band decrease in oscillatory power after stimulation was induced.

K. Analysis of the physiological networks.

We calculated the incoming connections and the outgoing connections of all BAs across all patients (n=25) included in the study. BAs localized on the convexity of the brain were densely covered with electrodes, including primary motor and sensory cortex, temporal lobe and the majority of the frontal lobe.

To visualize the connections between BAs, we created a matrix showing the presence of a significant connection between two BAs across all subjects. Figure 4 represents the percentage of patients having a significant evoked CCEP response in the recording BA after stimulation. Due to different electrode coverage of patients we also depicted the percentage of patients in which each BA was sampled. This matrix highlights the densely covered fronto-parieto-temporal lateral areas and also the strong intra- and inter-lobar connectivity. Stimulating the motor cortex (Area 4) or the premotor cortex (Area 6) resulted in significant CCEPs at almost every other recorded BA, which strongly supports the notion of a central role of that region in executing the motor response of the body. Connectivity between premotor and motor cortex were consistent across subjects, significant CCEP connection was found from motor to premotor cortex (9 / 10 subjects) and from premotor cortex to motor cortex (8 / 10 subjects). Both premotor cortices showed highly consistent outgoing connections to BA 9 (10 / 11 subjects for premotor; 6 / 8 subjects for motor), BA 10 (8 / 10 subjects for premotor; 6 / 9 subjects for motor), BA 46 (7 / 10 subjects for premotor; 9 / 10 subjects for motor) and Broca's area (5 / 7 subjects for premotor; 6 / 7 subjects for motor).

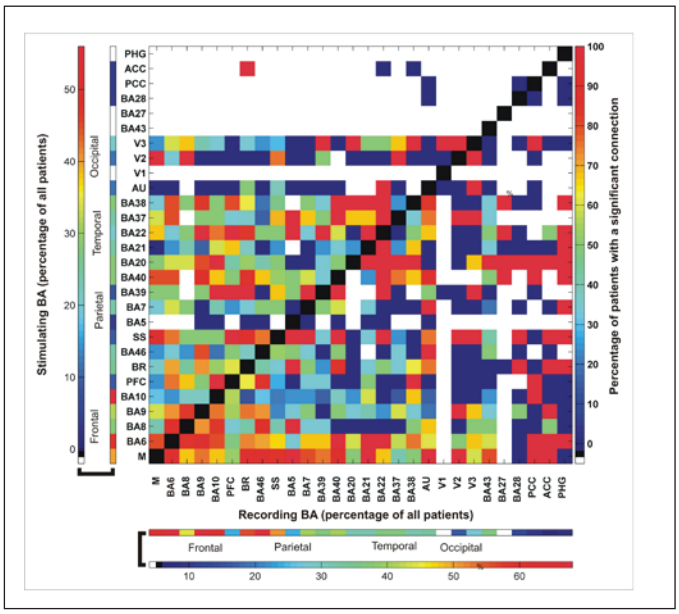


Figure 4. Percentage of patients with significant connections between specific BAs. Y axis: stimulating BAs and small regions; X axis: recording BAs and small regions. The side bar next to and below the BA labels represents the percentage of the specific BAs from all patients being a stimulating or a recording electrode with the appropriate color codes. Frontal and temporal connections were reproducible in most of the patients.

Association areas such as BA: 9, 10, 40, 21 showed a proportionally larger number of incoming connections, while areas involved in sensorimotor and language function such as BA: 4, 6, SS and 20 showed more outgoing connections. One notable exception was BA38, the temporo-polar region, which showed higher outdegree than indegree.

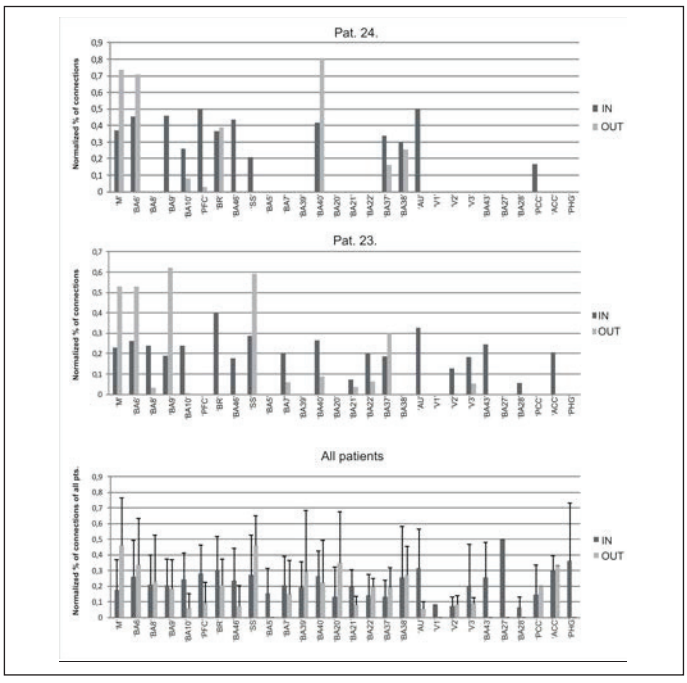


Figure 5. Two samples and the grand mean average of patients BA's in and outgoing connections. Connections are normalized with possible connections, grand mean average is normalized with possible connections from each patient. BA1-3: somato-sensory (SS); BA 44,45: Broca's area (BR); BA 11,12,25,47: prefrontal cortex (PFC); BA 23,26,29,30,31: posterior cingular cortex (PCC); BA 24,32,33: Anterior cingular cortex (ACC); BA 13,14,43,52:; BA 34,35,36: parahippocampal gyrus (PHG); BA 41,42: auditory cortex (AU). We have found statistically significant difference between in- and outgoing connections at 'M', 'BA10' and 'BA21'.

Further Aims

We would like to make a pathological, non pathological network identifier algorithm to facilitate the neurologists work and take measurable the pathological or non pathological connections.

To verify these results, we need higher number of patients involved to increase the statistical significance of the study.

Conclusion

The results suggest that single pulse electrical stimulation evoked potentials may reveal connections of functional areas and functional networks of the human brain. Other studies also report that direct cortical stimulation has a suppressive effect on fast cortical activity and epileptic spikes [7], or can help to clarify the size of the area to be removed[8].

We conclude that single pulse electrical stimulation is a promising technique in delineating eloquent cortex and might be a useful tool to identify pathological networks.

References

[1] R. Matsumoto, D.R. Nair, E. LaPresto, I. Najm, W. Bingaman, H. Shibasaki, and H.O. Lüders, "Functional connectivity in the human language system: a cortico-cortical evoked potential study.," Brain, vol. 127, (no. Pt 10), pp. 2316-30, Oct 2004.

[2] L. Schuh and I. Drury, "Intraoperative Electrocorticography and Direct Cortical Electrical Stimulation.," Seminars in Anesthesia vol. 16, pp. 46-55, 1996.

[3] M.J. Hamberger, "Cortical language mapping in epilepsy: a critical review.," Neuropsychol Rev, vol. 17, (no. 4), pp. 477-89, Dec 2007.

[4] I. Ulbert, E. Halgren, G. Heit, and G. Karmos, "Multiple microelectrode-recording system for human intracortical applications.," J Neurosci Methods, vol. 106, (no. 1), pp. 69-79, Mar 2001.

[5] D. Kovalev, J. Spreer, J. Honegger, J. Zentner, A. Schulze-Bonhage, and H.J. Huppertz, "Rapid and fully automated visualization of subdural electrodes in the presurgical evaluation of epilepsy patients.," AJNR Am J Neuroradiol, vol. 26, (no. 5), pp. 1078-83, May 2005.

[6] C.J. Keller, S. Bickel, L. Entz, I. Ulbert, M.P. Milham, C. Kelly, and A.D. Mehta, "Intrinsic functional architecture predicts electrically evoked responses in the human brain.," Proc Natl Acad Sci U S A, Jun 2011.

[7] M. Kinoshita, A. Ikeda, R. Matsumoto, T. Begum, K. Usui, J. Yamamoto, M. Matsuhashi, M. Takayama, N. Mikuni, J. Takahashi, S. Miyamoto, and H. Shibasaki, "Electric stimulation on human cortex suppresses fast cortical activity and epileptic spikes.," Epilepsia, vol. 45, (no. 7), pp. 787-91, Jul 2004.

[8] A. Valentín, G. Alarcón, M. Honavar, J.J. García Seoane, R.P. Selway, C.E. Polkey, and C.D. Binnie, "Single pulse electrical stimulation for identification of structural abnormalities and prediction of seizure outcome after epilepsy surgery: a prospective study.," Lancet Neurol, vol. 4, (no. 11), pp. 718-26, Nov 2005..

Appendix

PhD-studies started in 2010-2011:
- Zsolt Gelencsér
- Petra Hermann
- Antal Hiba
- Csaba Máté Józsa
- Bálint Péter Kerekes
- Márton Zsolt Kiss
- György Orosz
- István Zoltán Reguly
- János Rudan
- Emília Tóth

PhD-studies started in 2011-2012:
- István Endrédy
- Anna Horváth
- Balázs Jákli
- Balázs Knakker
- Péter Lakatos
- Endre László
- Balázs Ligeti
- Attila Novák
- Borbála Siklósi

PhD-studies started in 2012-2013:
- Vamsi Kiran Adhikarla
- Dóra Bihary
- Bence József Borbély
- Erzsébet Farkas
- Katharina Hofer
- Balázs Indig
- Attila Gyula Jády
- Mátyás Jani
- Imre Benedek Juhász
- András József Laki
- Gábor Zsolt Nagy
- Dénes Pálfi
- Ágnes Polyák
- Norbert Sárkány
- Máté Sipos
- Ádám Vály