Arrangement of genes involved in communication and cooperation in known bacterial genomes

The genes of the AHL quorum sensing regulatory system.

Theses of the Ph.D. dissertation

Pázmány Péter Catholic University

Faculty of Information Technology and Bionics

Multidisciplinary Technical Sciences Doctoral School

Zsolt Gelencsér

Supervisor: Prof. Sándor Pongor

2014

1. Introduction

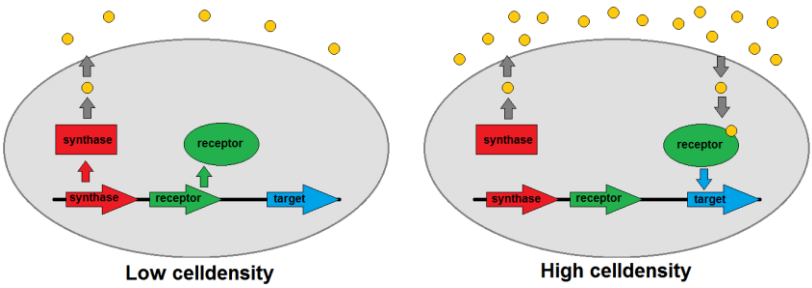
The number of the new DNA sequences is growing exponentially due to the rapid development and lower costs of current sequencing methods: there are hundred times more DNS sequences available than ten years ago and the accelerating growth is continuing. Besides merely sequencing the genomes, it is important to define the biological function of the sequences. The most important part of this process is an informatics problem: only powerful computers can “annotate a genome”, i.e. compare it with the growing databases, predict the location and the function of the genes, using semi-automatic annotation algorithms.

Beside the analysis of human genome one of the important areas of research is the analysis of the bacterial genomes. It turned out that the bacteria can form complex communities - these play significant role in human health, infection spreading and balance of soil- and marine biosphere. Cooperation and communication between cells and species is the basis of community formation, and the underlying molecular mechanism has large practical significance.

One of the most well studied communicational mechanisms is the so-called *quorum sensing* that use external chemical transmitters to deliver information. During my work I analyze the genes of the best known *quorum sensing* mechanism, the so called AHL system in which is N-acylhomoserine lactone is the signaling molecule. As first shown by this dissertation, this system is used by many hundred

species and shows a very large variety. It is based on two protein families: a *LuxI*-family protein synthesizes the signal that interacts with a *LuxR*-family protein which in most cases result in molecular complexes that bind to specific promoter DNA sequences located in the promoter region of themselves and of various target genes. In terms of regulation there is positive feedback between two proteins. There are bacteria which have a third player, a negative regulator protein whose gene is located close to *LuxR* and *LuxI* proteins. There are two important regulator proteins: *RsaL* and *RsaM*.

The purpose of my work was to analyze the genes of the AHL system in the known bacterial genomes, to design a computer program suitable for this problem and systematically analyze the resulting predictions.



Mechanism of the quorum sensing

2. Methods

I developed program scripts in order to create a pipeline for finding and analyzing quorum sensing genes using a collection of specific bioinformatics algorithms: the output of one program is retrieved reformatted as required by the next algorithm. With this method I created a workflow that needs little user interaction.

The genome annotation pipeline developed during my work is based on the so called subsystem approach: instead of a genome we analyze a defined subsystem in many genomes. The subsystem is defined by a rule collection extracted from experimental surveys that can be used to search genome sequences. In my work this subsystem included quorum sensing genes well as a few rules related to their mutual arrangements.

The search was based on Hidden Markov Models which is often used in bioinformatics for the description of multiple alignments of protein families. The so called HMM profiles are capable to generate a quantitative similarity measurer for each record of the examined sequences-database and we are able to select the genes which have both significant similarity and fulfill the rules of the subsystem.

3. New scientific results

I. Thesis: I applied an automated algorithm which is capable to analyze the topology of neighboring genes and subsystems and is capable to recognize non-annotated genes in genome-databases.

The requirements of the algorithm are the HMM profiles of the protein-families or genes and the list of the validation attributes. (e.g.: maximum gene-length, maximum gene-distance … etc.). First the program runs a search based on profiles in the NCBI (National Center for Biotechnology Information) gene-database. This search is relatively relaxed so we need to further validate the results using different NCBI databases. The validation uses more methods:

1) *Based on previous results:* for example the database already contains the function of the gene and this function is similar to the searched one.

2) *Using other search algorithms:* for example we run a BLAST algorithm in the neighborhood of our results then compare it with the results of the HMM search

3) *Gene-attributes:* for example if we know the average length of the gene then we reject every candidate which differs substantially from it.

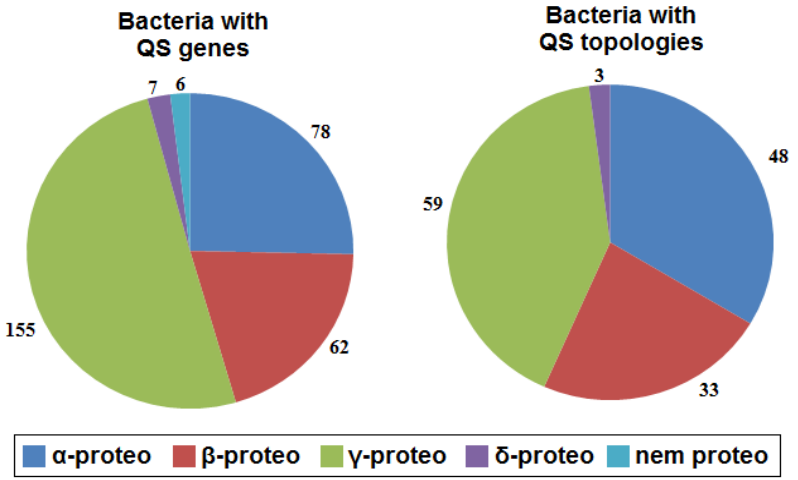
After the validation, the program determines the exact topology of found genes with the help of an algorithm developed in these theses. Then it creates a web-base report about newly found genes the topology-types.

II. Thesis: I showed that the 12% of the currently available proteobacterial genomes have AHL quorum sensing genes which is in agreement with previous biological estimates, on the other hand I observed over 50 of probable AHL gene-group which ones were not known till now.

The arrangement-analysis of the N-AHL based *quorum sensing* gene-topologies started with the bacteria from *Pseudomonadales* order. Later I extended the search to all the currently available bacterial genomes. (The source of the bacterial genomes was the NCBI GenBank database.) The bacteria found during the analysis that contain both of the core genes (luxI and luxR) are all members of the proteobacteria phylum. Since there are very similar protein families to both *LuxR* and *LuxI* but have different function so I individually validated the non-obvious hits via length and sequence-coverage. During the validation I used strict thresholds to get as a reliable result as possible. Only those non-annotated genes counted as valid results which were members of a known arrangement.

The question arises how well the results reflect true the frequency of quorum sensing genes. I think there are several reasons

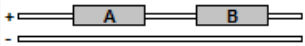
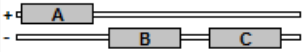
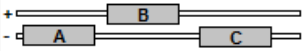
for cautiousness in this regard. First I only examined the cases where the *luxI* and *luxR* genes were near to each other. Second the search was based on the similarity to known *LuxR* and *LuxI* proteins. So there are for instance "solo" *luxR* genes which are out of this search, because they stand alone in the chromosome or linked to other signals. Third I ran the analysis on the whole-genome database which is a biased dataset, not necessarily representative of all bacteria found in nature. With these restrictions I found *quorum sensing* genes in the 12% of the proteobacteria, and this is in agreement with the estimated frequency of AHL positive strains in the proteobacteria (6-12%).



III. Thesis: I developed simple a notation-system for the description of quorum sensing systems and other small gene-subsystems.

I examined the arrangement according to two attributes: how much the genes overlap and what is their orientation. (The orientation is connected to the expression of the genes that it depends, on which DNS strand it is located.)

In the beginning of my work I defined a simple, formal description that includes the direction and the relative order of the genes. The following table shows examples for graphic and text-based examples:

DNA	Arrows	Notation	Cleartext
		$\vec{A}\vec{B}$	AB++
		$\vec{A}\overleftarrow{B}\overleftarrow{C}$	ABC+--
		$\overleftarrow{A}\vec{B}\overleftarrow{C}$	ABC-+-

IV. Thesis: I showed that the quorum sensing genes, i.e. the *luxR-luxI* pairs and the adjacent the negative regulator genes (*rsaM* and *rsaL*) have typical topological arrangements. In addition I determined the frequency of the individual topology types in the current available whole-annotated bacterial genomes.

Gene topology is a broad term that can include the arrangement of genes within chromosomes, with respect to the replication origin or other chromosomal elements. In this work I use the words “topological arrangement” or briefly “topology” to denote the arrangement within a close neighborhood of the *quorum sensing* regulatory genes. To illustrate these, I developed a concise notation based on a PROSITE-like syntax. The *luxR*, *luxI*, *rsaL* and *rsaM* genes were abbreviated as R, I, L and M. An arrow above each gene symbol then shows the direction of transcription.

Within the class of simple topologies, the majority of the cases are made up the $\vec{R}\vec{I}$ (R1) and the $\vec{R}\overleftarrow{I}$ (R2) topologies that Goryachev termed type A and type B. However I found combinations that are outside these two known categories, so all four arrangements that are possible for two vicinal genes appear, but the new types only in a small number. The three-member topologies are less various: In both cases a typical topologies were observed: $\vec{R}\overleftarrow{L}\vec{I}$ and $\overleftarrow{R}\vec{M}\vec{I}$.

The number complete bacterial genomes impressive however it is a negligible fraction of the species that may occurs in nature. Nevertheless we can make some statistical observations: the $\vec{R}\vec{I}$

topology is dominant in α-proteobacteria and the $\vec{R}\overleftarrow{I}$ topology appears most frequently in γ-proteobacteria. The $\vec{R}\overleftarrow{L}\vec{I}$ and $\overleftarrow{R}\vec{M}\vec{I}$ topologies are in both β and γ class but not in the α class.

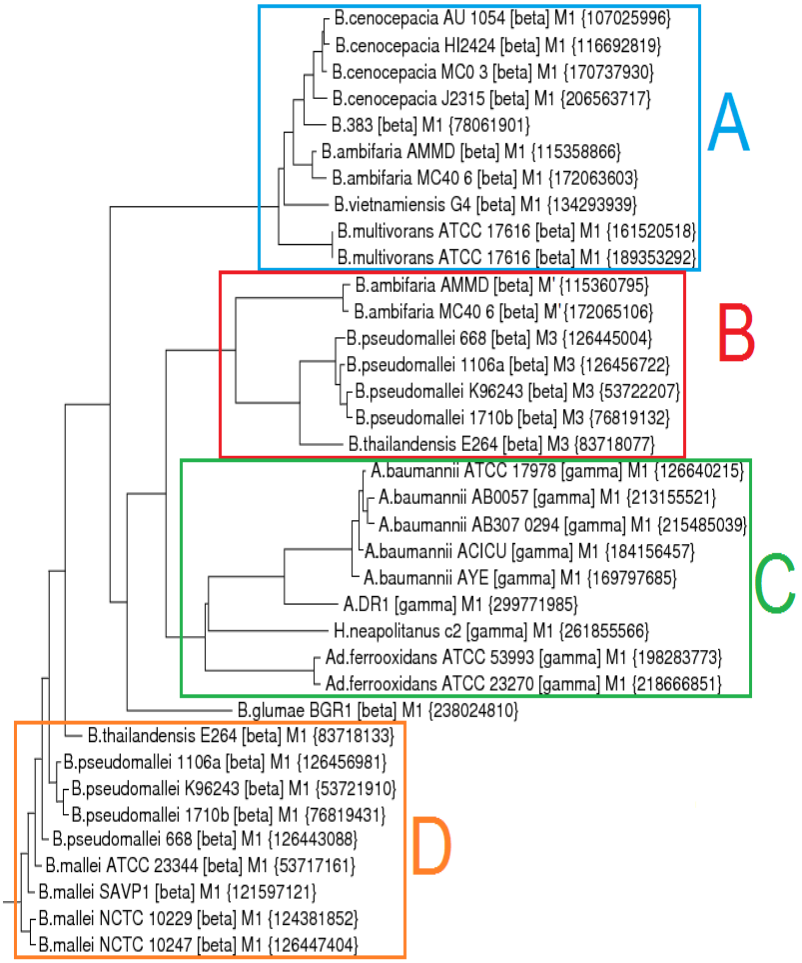
ID	Topology	Occurrence in proteobacteria				
		Total	**α**	**β**	**γ**	**δ**
R1	$\vec{R}\vec{I}$	*96*	71	14	11	0
R2	$\vec{R}\overleftarrow{I}$	*53*	2	2	46	3
R3	$\overleftarrow{R}\vec{I}$	*11*	1	3	7	0
R4	$\vec{I}\vec{R}$	*2*	2	0	0	0
L1	$\vec{R}\overleftarrow{L}\vec{I}$	*15*	0	7	8	0
M1	$\overleftarrow{R}\vec{M}\vec{I}$	*30*	0	20	10	0

Frequency of the topologies (partial table)

V. Thesis: I noticed that the quorum sensing sequences cluster by topology (and by the chemical type of the signal, which is correlated with the topology) and not by the taxonomic relationship. Namely a certain topology's *luxI* gene is more similar to a same topology-type's gene in another genome than a *luxI* gene in its own genome but in different topology type.

The *LuxI* and *LuxR* protein sequence cladograms show us that the proteins in different topology types are separating to clearly perceptible groups. The observation is the same in the case of the negative regulatory proteins.

The similarity trees of *rsaM* sequences we can notice that both the type of the topology and the taxonomy influences the separation of clades (figure below): There are 2 main groups; **M1** and **M3/M'**. It is noticeable that the genes from different topologies don't mix with each other: genes from the **M3/M'** type are located in the red (marked by **B**) group, the genes from the M1 type are located in the other three. The figure show us that the different bacterial groups have influence (though secondary) the localization in the tree. The γ-proteobacteria appears in the **C** group, the β-proteobacteria appears in the **A** and **D** groups. Briefly, the genes belonging to a certain topology behave like orthologs, and the relation between topology groups is reminiscent of paralogy.



The four types of the rsaM genes represented on a cladogram

4. Applications of the results

The program was planned with generality in mind, so the resulting tool is capable to analyze not only quorum sensing systems but any other small-size subsystem, if a sufficient number of protein sequences are known and they are suitable for creating HMM profiles.

Beside the general definition it was a goal to create as automata program as possible, so once the data are collected,, the search algorithm should run independently and return a result in a suitable form for further analysis. So with sufficient computing capacity we can execute more searches in parallel so as to compensate the algorithm's long running time.

Because there are many genes of unknown function in the currently known bacterial genomes, there is room for this kind of analysis. Since my dissertation was finished, out group accomplished the analysis of 10 further communication systems. One of these was published: the bottleneck is human time necessary to prepare the publications. The fact that protein groups that were previously thought as homogeneous orthologous groups, can be consistently subdivided according to the local gene topologies, is observed for a growing number of proteins and may lead to a further refinement of orthology definitions.

Acknowledgment

I would like to thank my supervisor Prof. Sándor Pongor for guiding and supporting me over the years. This work could not have come into existence without him. . I am grateful to Kumari Sonal Choudharynak and Sanjarbek Hudaiberdievnek for the help in biological and computing problems. Special thanks to Prof. Vittorio Venturi and his group because they provide me with necessary data and help me in the biological validation of results.

I thank the help to the PhD students of the Doctoral School, especially to Dóra Bihary and Balázs Ligeti. I am grateful to Borisz Galbáts and Áron Erdei who helped the progress of my work with their BSC thesis.

I am grateful to Prof Tamás Roska and Prof Péter Szolgay, the leaders of the Doctoral School, for the opportunity to work in a stimulating and protected environment. Finally I wish to thank the ICGEB Institute in Trieste, for the possibility to participate at their courses which helped me to deepen my knowledge in various fields of bioinformatics.

Author's Publications:

Zsolt Gelencsér, Borisz Galbáts, Juan F. Gonzalez, K. Sonal Choudhary, Sanjarbek Hudaiberdiev, Vittorio Venturi, and Sándor Pongor "Chromosomal Arrangement of AHL-Driven Quorum Sensing Circuits in Pseudomonas" ISRN Microbiology, vol. 2012, Article ID 484176, 6 pages, 2012.

Dóra Bihary, Ádám Kerényi, **Zsolt Gelencsér**, Sergiu Netotea, Attila Kertész-Farkas, Vittorio Venturi, Sándor Pongor "Simulation of communication and cooperation in multispecies bacterial communities with an agent based model" Scalable Computing: Practice and Experience Volume 13, Number 1, pp. 21–28.

Zsolt Gelencsér, Kumari Sonal Choudhary, Bruna Goncalves Coutinho, Sanjarbek Hudaiberdiev, Borisz Galbáts, Vittorio Venturi, and Sándor Pongor "Classifying the Topology of AHL-Driven Quorum Sensing Circuits in Proteobacterial Genomes" Sensors, vol. 12(5), pp. 5432-5444, 2012.

Kumari Sonal Choudhary, Sanjarbek Hudaiberdiev, **Zsolt Gelencsér**, Bruna Gonçalves-Coutinho, Vittorio Venturi, and Sándor Pongor "The Organization of the Quorum Sensing luxI/R Family Genes in Burkholderia," Int J Mol Sci, vol. 14, pp. 13727-13747, 2013.

Sanjarbek Hudaiberdiev, K. Sonal Choudhary, Roberto Vera, **Zsolt Gelencsér**, Doriano Lamba and Sándor Pongor."Census of solo luxR genes in becteria" 2014 (in preparation)