LARGE SCALE ANNOTATION OF BIOMOLECULAR DATA USING INTEGRATED DATABASE MANAGEMENT TOOLS

Theses of the Ph.D. dissertation

Pázmány Péter Catholic University

Faculty of Information Technology and Bionics

Roberto Vera Alvarez

Supervisor: Prof. Sándor Pongor

2014

Introduction

The use of information technologies for managing biomolecular data has profoundly changed the way biological research is done today. This sweeping change was triggered mainly by technological advances in data collection technologies, such as low-cost genomic DNA sequencing known as next-generation sequencing (**NGS**), proteomics and robotic biological assays. These and other high throughput technologies are producing an avalanche of noisy data on different aspects of living systems that have to be interpreted by integrating knowledge from disciplines as varied as medicine, computer sciences, biology, physics, chemistry and engineering. The integrated use of these information sources, also called Systems Biology, harbors four major challenges: (i) system-wide identification and quantification of components ("OMICS" data evaluation); (ii) experimental identification of physical component interactions, especially for information processing networks; (iii) computational inference of structure, type, and quantity of component interactions from data; and (iv) rigorous integration of heterogeneous data.

Integrating heterogeneous bioinformatics data is perhaps the most challenging tasks in biomedical research today. Data integration is one of the oldest themes in computer science that takes another face when applied to biomolecular data. The fast accumulation, frequent updating and the inherently noisy and redundant nature of molecular data are only one side of the problem. The most important source of difficulties is the complex nature of biological knowledge itself which makes biomolecular data integration different from those of other fields. All the more since much of this complex knowledge has to be discovered by the tools of bioinformatics itself.

The traditional tools, algorithms and databases of bioinformatics have been used for decades as background support for biological research projects in distinct areas of biological research. For analyzing the large amount of data provided by current data collection techniques, larger computer power is simply not sufficient. One needs advanced, cross-disciplinary integration of various knowledge domains in order to reach high quality interpretations based on the noisy, mass produced data.

One example of the application is metagenomics when DNA from a microbial community consisting of thousands of bacterial species is sequenced at the same time, and computational tools are necessary to identify the species and the gene functions that are present in a sample. This is often carried out by shotgun next generation sequencing that produces millions of short sequence reads that correspond to various, random parts of the genomes present. Evaluation is a complex task since the genomes of many bacteria contain identical segments,

so the origins of many reads cannot be unequivocally determined. Also, some species are present in very low quantities, but in the case of pathogenic species, identifying of low abundance species becomes very important. This task can be tackled only by first integrating the microbial genomes into a comprehensive database and then make it amenable to taxonomic and functional querying.

Another example of OMICs applications is the in depth analysis protein sequences, a frequent task in genome annotation. Often, some proteins encoded by a newly sequenced microbial genome are not sufficiently similar to any known protein so that the function could be inferred. In this case, one needs to query a variety of databases in a recurrent manner and combine the answers into new queries. This can be best done by integrating various databases into one common framework and then crating sub-databases amenable for large scale querying.

In order to address the above problems, I started to develop a common framework that allows the integration of biological databases with newly determined experimental results. The **JBioWH** framework focuses on the integration of biomolecular data in an efficient computational environment which can be used as integrative data supplier to other bioinformatics tools or to answer complex questions arising in System Biology projects. I also contributed to the annotation of metagenomic sequencing data, specifically, the taxonomic identification in unknown samples and the mapping of reads to functions. In order to address the above issues we have developed **Taxoner**, a simple, parallelizable pipeline that allows one to align millions of reads against a large, comprehensive nucleotide sequence database, using a standard personal computer or laptop.

Data and Methods

During the course of these projects, I developed several tools, pipelines and computational environments using various programming languages and third party libraries.

Relational schemas were developed using standard SQL language. The Database Management System used was MySQL community server version 5.6. Also, MySQL Workbench Tool was used for design and visualization of the relational schemas. All databases developed have an associated MySQL Workbench project freely available for users.

The programming languages more used were Java, C and bash shell scripting.

Oracle Java Standard Edition (SE) version 7 was used for the Java programs. Netbeans IDE was used as integrated development environment for writing the Java codes. The source code building process is executed by the Maven tool. Several Java technologies and libraries have been used in the projects. The most important technologies used are EclipseLink for the Java Persistence Model (JPA), Red Hat JBoss Middleware for the webservices development, JGraph for graph computing and visualization, Mojarra JavaServer Faces using the Java Server Faces (JSF) technology and Primefaces as JSF component for web interfaces.

The ANSI C language was used for programs with a high computing demand. The GCC compiler for GNU/Linux operating systems was used as compiler for the C code. The projects use the Make program to build the executables and libraries using the well-known Makefile. The parallelism in the programs was implemented using the POSIX Thread library. This library allows our programs to take advantage of the multicores and multiprocessors architectures available today. A C library named **BioC** was developed to provide highly optimized and fast methods for dealing with demanding tasks in terms of computational power and highly loaded data access. This library includes an implementation of a B+ Tree index to facilitate fast access to the elements of a larger body of data, such as the entries in a database. Also, modules to work with the FASTA file format and the NCBI Taxonomy database are included.

All the source codes developed are included in Projects that are freely available using the Google Code, a free project hosting service that provides a free collaborative development environment for open source projects (Table 1).

Table 1: List of Google Code Projects developed

Name	URL	Language
Taxoner	http://code.google.com/p/taxoner/	C
JBioWH	http://code.google.com/p/jbiowh/	Java
BioC	http://code.google.com/p/bioc/	C

Finally, the Google Cloud Platform was used for computations which require a high performance computing using virtual machines *in-house* modified to create a virtual Beowulf-like Cluster inside the cloud platform. Several bash scripts were developed for administration of the virtual cluster remotely.

New Scientific Results

I. **Biological database integration. I developed an open-source framework for biological data integration. The framework is a computational system freely available that can be used to answer complex biological questions, or just, as supplier system of integrative data to others client applications [1].**

The Java BioWareHouse project that I developed provides an open-source platform-independent programming framework that allows a user to build his/her own integrated database from the most popular data sources. **JBioWH** can be used for intensive querying of multiple data sources and the creation of streamlined task-specific data sets on local PCs. It is based on a MySQL relational database and provides four kind of access to the integrated data: a) direct access to the relational schema (SQL), b) programmatically access through the Java API (java persistence model and search classes), c) graphical access through the Desktop Client and d) html access through the webservices (JSON and XML). The system has a modular design that can be easily modified accordingly to the biological context of the problem. Therefore, **JBioWH** can be tailored for use in specific circumstances, including the handling of massive queries for high-throughput analyses or CPU intensive calculations. At present, JBioWH contain parsers for retrieving data from 24 public databases (e.g. NCBI, KEGG, etc) [1].

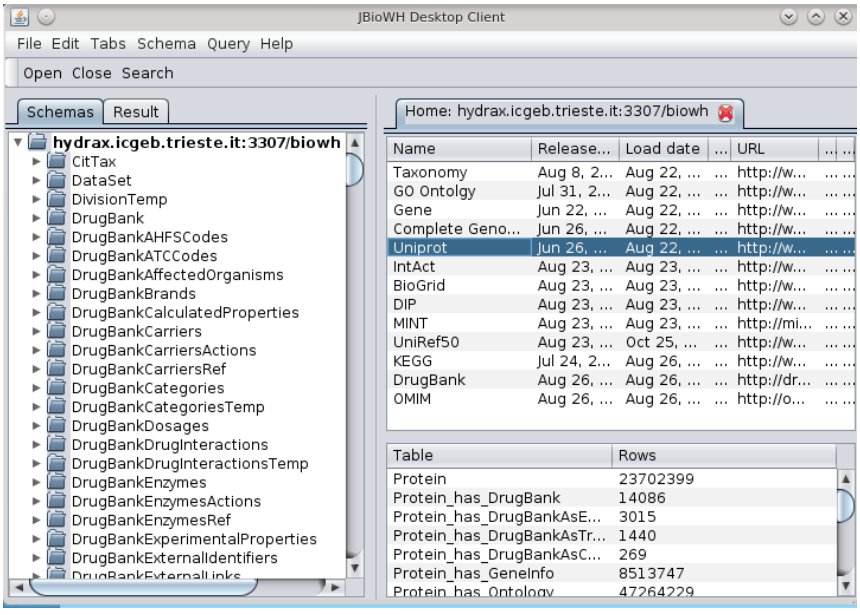

Figure 1: JBioWH Desktop Client

II. Prediction of bacterial taxa and gene functions from next generation sequencing results. I developed an algorithm and a pipeline of programs for the detection of bacterial species and gene functions from metagenomic datasets. The pipeline uses a fast aligner and an integrated database to assign reads to bacterial strains, species and higher taxa, and to gene functions, using an integrated database [2].

Next generation sequencing (NGS) of metagenomic data is becoming a standard approach to detect individual species or pathogenic strains of microorganisms, see Figure 2. Computer programs used in the NGS community have to balance between speed and sensitivity and as a result, species or strain level identification is often inaccurate and low abundance pathogens can sometimes be missed. Taxoner is an open source taxon assignment pipeline that includes a fast aligner (e.g. Bowtie2) and a comprehensive DNA sequence database. Taxoner performs as well as, and often better than BLAST, but requires two orders of magnitude less running time meaning that it can be run on desktop or laptop computers. When applied to metagenomic datasets, Taxoner can provide a functional summary of the genes mapped and can provide strain level identification as shown Figure 3 [2].

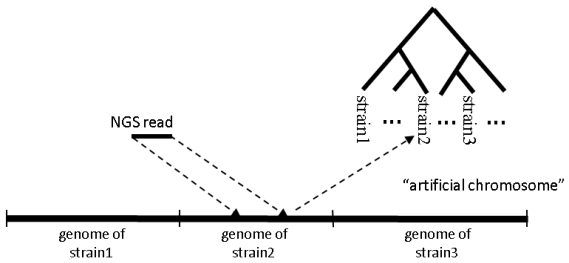

Figure 2: Mapping of reads to bacterial strains using artificial chromosomes. A strain is a segment of the artificial chromosome that is named by a label in the taxonomical hierarchy.

COG/eggNOG Top Classes

Top Classes	Total
INFORMATION STORAGE AND PROCESSING	675
CELLULAR PROCESSES AND SIGNALING	528
METABOLISM	1335
POORLY CHARACTERIZED	564

COG/eggNOG Functional Classification

Functional Classes	One Letter	Total
RNA processing and modification	A	0
Translation, ribosomal structure and biogenesis	J	243
Chromatin structure and dynamics	B	1
Replication, recombination and repair	L	202
Transcription	K	229
Signal transduction mechanisms	T	100
Cell wall/membrane/envelope biogenesis	M	179
Extracellular structures	W	0
Cell motility	N	2
Nuclear structure	Y	0

Figure 3: Function assignment output from Taxoner.

III. A marker database for identification of bacteria. I developed a workflow to identify unique parts of bacterial genomes, and created a comprehensive database for these unique markers suitable for the experimental and computational identification of bacteria. .

DNA markers are unique nucleotide sequences allowing the detection of certain organisms and to distinguish those organisms from all other species, using *in silico* or experimental technologies. Markers can be used as the basis for diagnostic assays to detect microbes in environmental or clinical samples. I have developed a tool that identifies a comprehensive set of markers for any set of target genomes, and screens this set against a background of known genomes (Figure 4). The markers can be browsed by proximal genes, and processed with a PCR primer design program that allows one to select markers for experimental use.

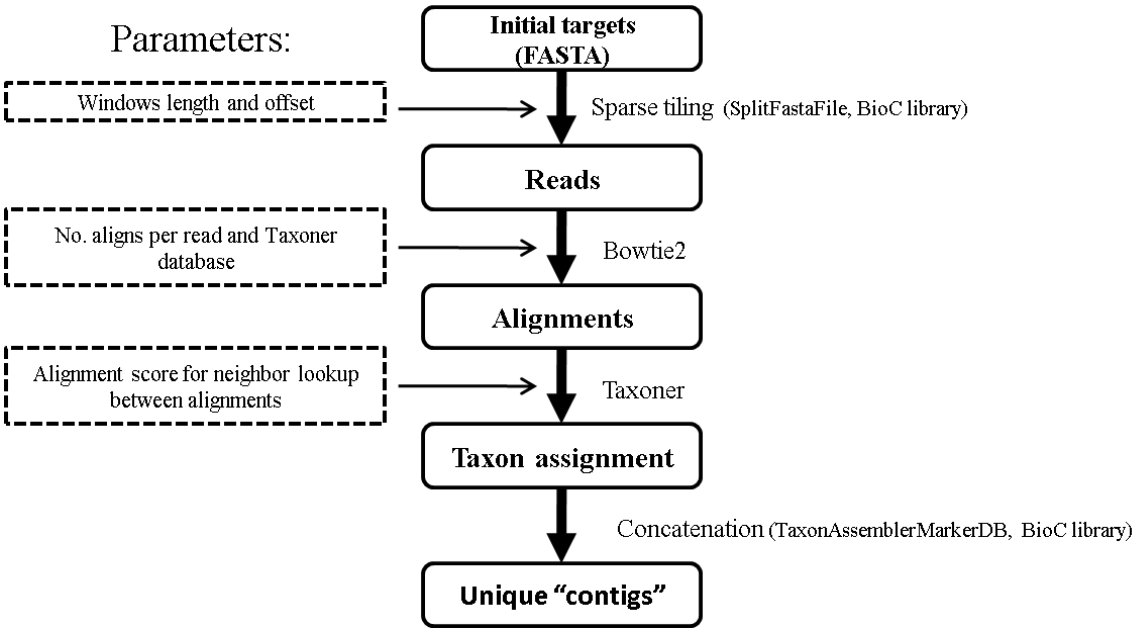

Figure 4: Workflow for unique segments identification.

Application of the results

I applied the above general principles for characterizing quorum sensing genes in bacteria [3], for annotating proteomic data [4-6], for characterizing metabolic pathways, protein targets [7-9] and drugs networks [10].

The annotation of "OMICS" data can be pictured via a generalized scheme derived from the principles outlined in points I and II. Namely, structural similarities are used to map "OMICS" data to existing data that are linked to other data as well as conceptual networks (ontologies, taxonomic hierarchies etc). The annotation emerges then as a validated link between a data item and a concept. A web-server based on this philosophy contains a similarity search module operating on a dedicated database, constructed and pre-annotated with the **JBioWH** system. I built such web servers for proteomics [6] (http://net.icgeb.org/ptmtreesearch/) and taxonomic annotation [2] (http://pongor.itk.ppke.hu/taxoner). Also, proteomics databases designed for protein identification using the Mascot Server was created from **JBioWH**. Specifics criterions for filtering the proteins by sequences composition was used in order to create training datasets for testing proteomics experiments. See the papers [4, 5]. Finally, the integrated database was used in drug design experiments to characterize metabolic pathways, protein targets and drugs networks. The **JBioWH**'s Drug module was used to provide the drug-target relationship used to test the docking programs and the three-dimensional location of the drug inside the protein's active site. See the patents [7, 8] and paper [9, 10].

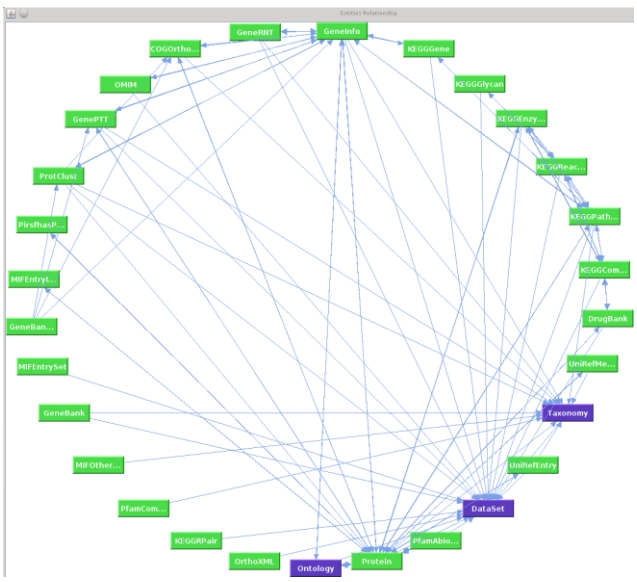

Figure 5: Relations of the integrated data.

Acknowledgments

This work is the result of an exceptionally fruitful collaboration between institutions from Cuba, Italy and Hungary. The project started during my first visit to the ICGEB in 2008, when I was invited by Prof. Sándor Pongor. His support, kindness and, especially, his scientific vision has made possible all these results. Special thanks are due to my friends and close collaborators, Yasset Perez-Riverol (European Bioinformatics Institute), Sonal K. Choudhary and Sanjar Hudaiberdiev (ICGEB, Trieste), Lőrinc S. Pongor (Semmelweis University, Budapest) and Balázs Ligeti (Pázmány University, Budapest). Much of this work is the result of our common efforts so a large part of the credit goes to them.

I wish to express my gratitude to ICGEB-Trieste, Italy, to LNCIB, Trieste, Italy, and to Pázmány Péter Catholic University, Budapest, Hungary for the fellowships and the financial support. In particular, I wish to thank the help and assistance of Profs. Francisco E. Baralle (ICGEB) and Claudio Schneider (LNCIB)

I want to thank my labmates at ICGEB, Sonal and Sanjar for being family and for the many English corrections to my thesis and my presentations. I also want thank my colleagies at LNCIB, Vanessa Florit and Raffaella Florit for their trust and support.

My special hanks are due to all my friends in Trieste, especially to Piero, Roberta, Giacomo, Elena, Giorgia, Giampaolo, Simona, Sorrentino, Roberto and Daniela for their friendship and support.

And to my catholic community at the San Giusto Cathedral of Trieste: *Grazie mille per la preghiera e per ricordarci che non dobbiamo avere paura: Dio ci ama e sempre provvede.*

I would like to thank my wife and children for being there for me always and letting me work despite of their needs. And to my parents and my sister for their sacrifices which made me the man that I am: *Gracias mis amores por todo. Este también es un logro de todos ustedes. Los amo infinitamente.*

Author's Publications:

1. **Vera R**, Perez-Riverol Y, Perez S, Ligeti B, Kertesz-Farkas A, Pongor S, Kertész-Farkas A: **JBioWH: an open-source Java framework for bioinformatics data integration.** *Database : the journal of biological databases and curation* 2013, **2013**:bat051.
2. Pongor LS, **Vera R**, Ligeti B: **Fast and sensitive alignment of microbial whole genome sequencing reads to large sequence datasets on a desktop PC: application to metagenomic datasets and pathogen identification**. *PloS one* 2014, **accepted**.
3. Dogsa I, Choudhary KS, Marsetic Z, Hudaiberdiev S, **Vera R**, Pongor S, Mandic-Mulec I: **ComQXPA Quorum Sensing Systems May Not Be Unique to Bacillus subtilis: A Census in Prokaryotic Genomes**. *PloS one* 2014, **9**(5):e96122.
4. Sanchez A, Perez-Riverol Y, González LJ, Noda J, Betancourt L, Ramos Y, Gil J, **Vera R**, Padrón G , Besada V: **Evaluation of Phenylthiocarbamoyl-Derivatized Peptides by Electrospray Ionization Mass Spectrometry: Selective Isolation and Analysis of Modified Multiply Charged Peptides for Liquid Chromatography-Tandem Mass Spectrometry Experiments.** *Analytical chemistry* 2010, **xxx**:552-559.
5. Perez-Riverol Y, Sánchez A, Ramos Y, Schmidt A, Müller M, Betancourt L, González LJ, **Vera R**, Padron G, Besada V: **In silico analysis of accurate proteomics, complemented by selective isolation of peptides.** *Journal of proteomics* 2011, **74**:2071-2082.
6. Kertész-Farkas A, Reiz B, **Vera R**, Myers MP, Pongor S: **PTMTreeSearch: a novel two-stage tree-search algorithm with pruning rules for the identification of post-translational modification of proteins in MS/MS spectra.** *Bioinformatics (Oxford, England)* 2013, **30**:234-241.
7. Mazola Reyes Y, Chinea Santiago G, Guirola Cruz O, **Vera R**, Huerta Galindo V, Fleitas Salazar N, Musacchio Lasa A: **Chemical compounds having antiviral activity against dengue virus and other flaviviruses**. In.: WO/2009/106019; 2009.
8. Rodriguez Fernandez RE, **Vera R**, de la Nuez Veulens A, Mazola Reyes Y, Perea Rodriguez SE, Acevedo Castro BE, Musacchio Lasa A, Ubieta Gomez R: **Antineoplastic compounds and pharmaceutical compositions thereof**. In.: WO/2006/119713; 2006.
9. Perez-Riverol Y, **Vera R**, Mazola Y, Musacchio A: **A parallel systematic-Monte Carlo algorithm for exploring conformational space.** *Current topics in medicinal chemistry* 2012, **12**:1790-1796.
10. Ligeti B, **Vera R**, Lukacs G, Gyorffy B, Pongor S: **Predicting effective drug combinations via network propagation**. In: *2013 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE; 2013: 378-381.