IWPLS'09

Sys-Bio Gateway: a framework of bioinformatics database resources oriented to systems biology

Luciano Milanesi[1*], Roberta Alfieri[1], Ettore Mosca[1], Federica Viti[1], Pasqualina D'Ursi[1], Ivan Merelli[1]

[1]Institute for Biomedical Technologies-CNR, via Fratelli Cervi 93, 20090 Segrate (Mi),

[*]Corresponding author

Associate Editors: Sandra Gesing and Jano van Hemert

ABSTRACT

In this paper we present SysBio-Gateway, a framework of standard solution for data integration in bioinformatics and Systems Biology. In the context of several research projects, we developed a set of databases and related web interfaces to cover many levels of the biological system complexity. Furthermore several analysis tools have been developed and integrated in the web resources in order to cope with problems of data integration and mining in a systems perspective. The SysBio-Gateway, freely accessible at the URL http://www.itb.cnr.it/sysbio-gateway, offers access to all the presented resources, which concerns different levels of organization of biological living systems, from genes to organs, passing through proteins, protein families, cellular processes, tissues, and pathologies.

1 INTRODUCTION

Data integration is nowadays an essential task to accomplish in order to achieve a view of the biological knowledge as much complete as possible. This is very important in a discipline, such as the bioinformatics one, in which data are growing at lightning speed thanks to novel bio-molecular high-throughput techniques. In particular, considering the systems biology field, the integration of biological knowledge related to different levels - such as genomics, transcriptomics, proteomics, and network interactions – is crucial in order to support the mathematical modelling and the computer simulation of biological pathways.

Data integration can be defined as the process of combining information, residing at different sources, to provide the user with a unified view of these data for enabling the possibility to achieve real knowledge. Experimental researchers and computer scientists can discover through data integration new and interesting relationships that allow better and faster experimental decisions, for example about protein targets and drug molecules. Moreover, the achievement of interesting results in most bioinformatics and systems biology-related activities, from functional characterization of genomic and proteomic data to the development of mathematical models of biological processes, requires an integrated view of all relevant information useful to accomplish those tasks.

Data integration problem can be described by three levels of complexity. The first layer is the integration of information from heterogeneous resources by collecting data between different databases to allow a unified query schema. The second level consists in identifying correlative associations across different datasets, generally using an ontology support, to provide a comprehensive view of the same objects in light of different data sources. The third layer is mapping the information gained about interacting objects into networks and pathways that may be used as basic models for the underlying cellular systems.

In this paper we show an overview of solutions, developed by our bioinformatics laboratory, for data integration in the context of bioinformatics and Systems Biology, where many levels of complexity are covered: from proteins to cellular processes, from tissues to disease and organs. A set of databases and web resources oriented to different biological topics are presented here in a unified framework, to provide some practical examples of our experience in this field. In particular we present resources related to specific a) biological processes, such as cell cycle, b) pathologies, such as breast cancer, c) organs, such as brain, d) protein families, such as protein kinases, e) protein mutations, and f) tissues. The here proposed solutions are based on a common methodology of data integration suitable we adoptedto support different biological projects. Several tools have been developed ad hoc to cope with different problems arising both from the increasing number of experimental data and from the need to improve the knowledge in a systems perspective.

Web resources are presented here in a unified framework, to provide some practical examples of our experience in this field. We present resources related to specific knowledge about a) biological processes, such as cell cycle, b) pathology, such as breast cancer, c) organs, such as brain, d) protein family, such as protein kinases, e) protein mutations, and f) tissue.

2 METHODS

* To whom correspondence should be addressed.

All the presented databases rely on a data warehouse approach, which requires collection and transformation of heterogeneous data coming from different sources to make them accessible by the scientific community though a unified query schema. This model is typical of data integration and differs from the normalized databases, designed to support data integrity, which are widely used to maintain primary resources. While many data warehouses solutions used in bioinformatics provide generic query interfaces applicable to all the data they contain, our system allows the construction of queries that extract and filter information derived from the original resources. This integration improves the usability of the information, but data must be fitted into unique format that takes into account the relationships between the different sources.

The implementation of our relational databases is always managed by a MySQL server. The primary data are collected with a series of Perl scripts which retrieve data from external resources, transform them in a compliant format and load them into the warehouse data model. The developed resources are all freely accessible through web interfaces which are made up of a set of HTML pages dynamically generated from PHP scripts, in order to provide information in specific reports created for responding to specific requirements related to the biological problem of interest.
Besides the integration problem, the analysis of large quantity of experimental data in a Systems Biology perspective has been tackled by developing a number of specific tools integrated in the web resources. Some examples available through the SysBio-Gateway are:

- a high performance tool for the simulation of Ordinary Differential Equation relying on mathematical models based on Xppaut
- a tool for the visualization of the Protein Data Bank protein structure and correlated Connolly surfaces
- a set of tools for the analysis of protein-protein interactions network (search for the first neighbourhood, search for shortest path and common annotations)
- a tool for modelling the protein mutant starting from Single Nucleotide Polymorphism data which relies on Modeller
- a tool for pathological image processing oriented to support tissue microarray analysis.

In order to guarantee data sharing and structuring, developed databases are enriched by a underlying crucial feature: the ontological support. Exploited ontologies concern all levels of molecular biology, from genes to proteins to pathways, even covering tissues and diseases aspects. As example, we used Gene Ontology for genes annotation and KEGG Pathway Ontology (derived from the hierarchical organization of KEGG pathways) for biological networks. For their intrinsic structure, that is more complex than a simple recognized vocabulary, ontologies are used in the developed resources to enrich to the informative content. Ontologies provides not only the availability of a commonly accepted vocabulary, which facilitates data sharing and information querying, but also increases the performance of statistical and analytical studies. The hierarchical graphs, which represent ontologies backbone, can in fact support the generation of novel knowledge, leading to the creation of new relationships among biological entities or to the deduction of new associations.

In the context of Systems Biology, another service provided by our infrastructure is a high performance framework for the numerical simulation of molecular models and an associated parameters values estimation tool. In fact, uncertainty of parameters values is one of the greatest problem in the development of new cellular models and *in silico* parameters estimation is one of the most common adopted solutions. Our system provides a global optimization algorithm, which relies on an evolution strategy, that can be employed on the top of the simulation engine to accomplish parameter estimation of complex mathematical models. Due to the high computational load needed for the parameter estimation of large networks, we implement this system using a distributed paradigm which can also be used in the context of grid computing technologies. Relying on a data parallel approach it is possible to handle the parameters estimation of complex models using large computational facilities, which is very useful for example in the case of cell cycle related models.

3 RESULTS

We propose here a standard solution for the implementation of bioinformatics framework, by collecting our experiences in the SysBio-Gateway, which embraces different levels of organization of biological living systems by linking, integrating and analysing heterogeneous data that vertically lead from genes to organs, passing through proteins, protein families, cellular processes, tissues, and pathologies. This approach has been tested for different databases implementations, such as:

- Cell Cycle
- G2S Breast Cancer
- Gene Nerve Cell
- Kinweb
- ProCMD
- TMA Rep

SysBio-Gateway is freely accessible at the URL, http://www.itb.cnr.it/sysbio-gateway. In this page (Fig.1) the user can directly access the resources held in the gateway web page, handle and analyse data according to specific demands coming from the bioinformatics and the Systems Biology communities.

3.1 Cell Cycle Database

The cell cycle is one of the biological processes most frequently investigated in systems biology studies and it involves the knowledge of a large number of genes and networks of protein interactions. A deep knowledge of the molecular aspect of this biological process can contribute to making cancer research more accurate and innovative. In this context the mathematical modelling of the cell cycle has a relevant role to quantify the behaviour of each components of the systems. The mathematical modelling of a biological process such as the cell cycle allows a systemic description that helps to highlight some features such as emergent properties which could be hidden when the analysis is performed only from a reductionist point of view. Moreover, in modelling complex systems, a complete annotation of all the components is equally important to understand the interaction mechanism inside the network: for this reason data integration of the model components has high relevance in systems biology studies.

The Cell Cycle Database [1], intended to support systems biology analysis on the cell cycle process, starting from two organism, yeast and mammalian, that present a high evolutionary molecular conservation between them. The database integrates information about genes and proteins involved in cell cycle process, stores complete models of the interaction networks and allows the mathematical simulation over time of the quantitative behaviour of each component. To accomplish this task, we developed, on the top of the database, a web interface for browsing information related to cell cycle genes, proteins and mathematical models. In this framework, we implemented a pipeline which allows users to deal with the mathematical part of the models, in order to solve under different conditions the Ordinary Different Equation systems that describe the biological process. In this way the resource is useful both to retrieve information about cell cycle model components and to analyze their dynamical properties.

This integrated system aims to become a useful resource for collecting all the information related to actual and future models of this network. The flexibility of the database allows the addition of mathematical data which are used for simulating the behaviour of the cell cycle components in the different models. Coupling structure and dynamical information about models Cell Cycle Database allows to achieve system-level properties, such as stable steady states and oscillations.

3.2 G2S Breast Cancer

The study of breast cancer and its development involves the knowledge of a large number of genes and molecular interactions and thus the systems biology approach is essential to describe the processes related to the pathology and to perform useful predictions. For the effective application of the systemic approach it is essential to arrange information about genes, cellular pathways and interactions that they undertake. These annotations are publicly available in bioinformatics resources and their integration produces an information enrichment, allowing, for example, the clustering of breast cancer genes by common signatures or the suggestion of possible annotations for not yet annotated genes. The Genes-to-Systems Breast Cancer (G2SBC) Database [2] is a bioinformatics resource that collects information about breast cancer genes, proteins and mathematical models and provides a number of tools to analyse the integrated data. Protein-protein interactions data are used to suggest new possible annotations and the link with the Cell Cycle database allows the simulation of cell cycle mathematical models beginning from breast cancer molecular alterations. Taking advantage from the multi-level approach, due to the consideration of both the “building-blocks” level (genes and proteins) and the systems level (molecular and cellular systems), the G2SBC Database enables predictions and new hypothesis formulation.

3.3 Gene Nerve Cell Database

In the past few years the new research field of neuroinformatics has strongly emerged. Two main aspects must be highlighted in his context: the interplay of structural, chemical and electrical signals in nervous tissue and the importance of modelling such signals. The great amount of qualitative experimental data in neuroscience represents the starting point to expand the nerve cell modelling in new directions, especially in the development of gene and protein interaction networks. The aim of this new discipline is thus to gather the application of computational models and analytical tools and the improvement of neuroscience knowledge. Moreover in the nervous system a great importance must be given to the study of the molecular processes, and the knowledge about the key players (genes and proteins) involved in such processes should be as much complete as possible. The study of nerve cells, neurons and, more in general, brain and its development involves information regarding a large number of genes and molecular interactions. Thus the systems biology approach is essential. Starting from the integration of genes and proteins interaction data with experimental data it is possible to develop new discovery strategies in brain studies.

In this context we propose a new data integration system, the “Gene Nerve Cell Database” [3], a resource to support neuroinformatics research which contains up to date information regarding the mouse genes which have brain-specific gene expression patterns. The list of genes specifically expressed in the nervous system was built starting from the Mousebrain Gene Expression Map (BGEM) and the Allen Brain Atlas. Other genes were taken from literature, including the available web resources.

3.4 Kinweb

Protein kinases are a well defined family of proteins, characterized by the presence of a common kinase catalytic domain and playing a significant role in many important cellular processes, such as proliferation, maintenance of cell shape, apoptosys. In many members of the family, additional non-kinase domains con-

tribute to further specialization, resulting in subcellular localization, protein binding and regulation of activity, among others.

About 500 genes encode members of the kinase family in the human genome, and although many of them represent well known genes, a larger number of genes code for proteins of more recent identification, or for unknown proteins identified as kinase only after computational studies.

A systematic in silico study performed on the human genome, led to the identification of 5 genes, on chromosome 1, 11, 13, 15 and 16 respectively, and 1 pseudogene on chromosome X; some of these genes are reported as kinases from NCBI but are absent in other databases, such as KinBase. Comparative analysis of 483 gene regions and subsequent computational analysis, aimed at identifying unannotated exons, indicate that a large number of kinase may code for alternately spliced forms or be incorrectly annotated. An InterProScan automated analysis was perfomed to study domains distribution and combination in the various families. At the same time, other structural features were also added to the annotation process, including the putative presence of transmembrane alpha helices, and the cystein propensity to participate into a disulfide bridge.

The predicted human kinome was extended by identifying both additional genes and potential splice variants, resulting in a varied panorama where functionality may be searched at the gene and protein level. Structural analysis of kinase proteins domains as defined in multiple sources together with transmembrane alpha helices and signal peptide prediction provides hints to function assignment. The results of the human kinome analysis are collected in the KinWeb database [4], available for browsing and searching over the internet, where all results from the comparative analysis and the gene structure annotation are made available, alongside the domain information. The site provides a comprehensive analysis of functional domains of each gene product. For each kinase, GenBank RefSeq and the SwissProt entry names are available along with information about kinase classification (Hanks and Hunter classification). Kinases may be searched by domain combinations and the relative genes may be viewed in a graphic browser at various level of magnification up to gene organization on the full chromosome set.

3.5 ProCMD

Activated Protein C (ProC) is an anticoagulant plasma serine protease which also plays an important role in controlling inflammation and cell proliferation. Several mutations of the gene are associated with phenotypic functional deficiency of protein C, and with the risk of developing venous thrombosis. Structure prediction and computational analysis of the mutants have proven to be a valuable aid in understanding the molecular aspects of clinical thrombophilia. A specialized relational database and a search tool for natural mutants of protein C have been built. The database contains 195 entries that include 182 missense and 13 stop mutations. A menu driven search engine allows the user to retrieve stored information for each variant, that include genetic as well as structural data and a multiple alignment highlighting the substituted position. Molecular models of variants can be visualized with interactive tools; PDB coordinates of the models are also available for further analysis. Furthermore, an automatic modelling interface allows the user to generate multiple alignments and 3D models of new variants.

ProCMD [5] is an up-to-date interactive mutant database that integrates phenotypical descriptions with functional and structural data obtained by computational approaches. It will be useful in the research and clinical fields to help elucidate the chain of events leading from a molecular defect to the related disease.

3.6 TMA Rep

Tissue MicroArray technique is becoming increasingly important in pathology for the validation of experimental data from transcriptomics analysis. This approach produces many images which need to be properly managed, if possible exploiting an infrastructure able to support tissue sharing between institutes. Moreover, the nowadays available frameworks oriented to Tissue MicroArray provide good storage for clinical patients, sample treatments and block constructions information, but their utility is limited by the lack of data integration with bioinformatic approaches.

We propose a Tissue MicroArray web oriented system [6] that supports researchers in managing bio-samples and that, through the use of ontologies, enables tissue sharing in order to promote TMA experiments design and results evaluation. Our system provides ontological description both for describing pre-analysis tissue images and for identifying post-process image results, which represents a crucial feature for promoting information exchange. Working on well-defined terms allows to perform queries on web resources for literature articles, in order to integrate both pathology and bioinformatics data.

Through this system, users associate an ontology-based description to each image uploaded into the database and also integrate results with the ontological descriptions of biosequences identified in each tissue. It is even possible to integrate the ontological description provided by the user with a fully compliant gene ontology definition, enabling statistical studies about correlation between the analyzed pathology and the most commonly related biological processes. Finally, the web site embeds a tool oriented to pre-array tissue image analysis, specific for tubular breast cancer affected tissues.

4 CONCLUSION

In this paper we present an integrated solution to explore part of the information gained in the field of life science oriented to systems biology. In order to achieve a systemic perspective of a set interesting topics for our group, we come to this integrated portal, SysBio-Gateway, which combines a bioinformatics approach, i.e. data integration using data warehouse approach, application of tools for the data analysis, study of structural modifications - both for genome and proteins -, and systems biology approach, that is

the study of protein-protein interaction networks, molecular mathematical models, pathological states, under a systemic point of view.

ACKNOWLEDGMENTS

This work has been supported by the NET2DRUG, EGEE-III, BBMRI, EDGE European projects and by the MIUR FIRB LIT-BIO (RBLA0332RH), ITALBIONET (RBPR05ZK2Z), BIOPOP-GEN (RBIN064YAT), CNR-BIOINFORMATICS initiatives. We also acknowledge the support of the e-Science Institute in Edinburgh.

REFERENCES

1. Alfieri R, Merelli I, Mosca E, Milanesi L. (2008) The cell cycle DB: a systems biology approach to cell cycle analysis *Nucleic Acids Res. 36(Database issue): D641–D645.*
2. Mosca E, Alfieri R, Milanesi L, Genes-to-Systems Breast Cancer (G2SBC) Database: a data integration approach for breast cancer research oriented to systems biology, *Sysbiohealth Simposium 2008,*
3. Alfieri R, Mosca E, Milanesi L, Gene Nerve Cell DB: a data integration approach for neuroinformatics research oriented to systems biology, Sysbiohealth Simposium 2008, Bologna, 24-25 November 2008
4. Milanesi L, Petrillo M, Sepe L, Boccia A, D'Agostino N, Passamano M, Di Nardo S, Tasco G, Casadio R, Paolella G. (2005) Systematic analysis of human kinase genes: a large number of genes and alternative splicing events result in functional and structural diversity. *BMC Bioinformatics. 1;6 Suppl 4:S20.*
5. D'Ursi P, Marino F, Caprera A, Milanesi L, Faioni EM, Rovida E. (2007) ProCMD: a database and 3D web resource for protein C mutants. *BMC Bioinformatics. 8;8 Suppl 1:S11*
6. Viti F, Merelli I, Caprera A, Lazzari B, Stella A, Milanesi L. (2008) Ontology-based, Tissue MicroArray oriented, image centered tissue bank. *BMC Bioinformatics. 25;9 Suppl 4:S4*

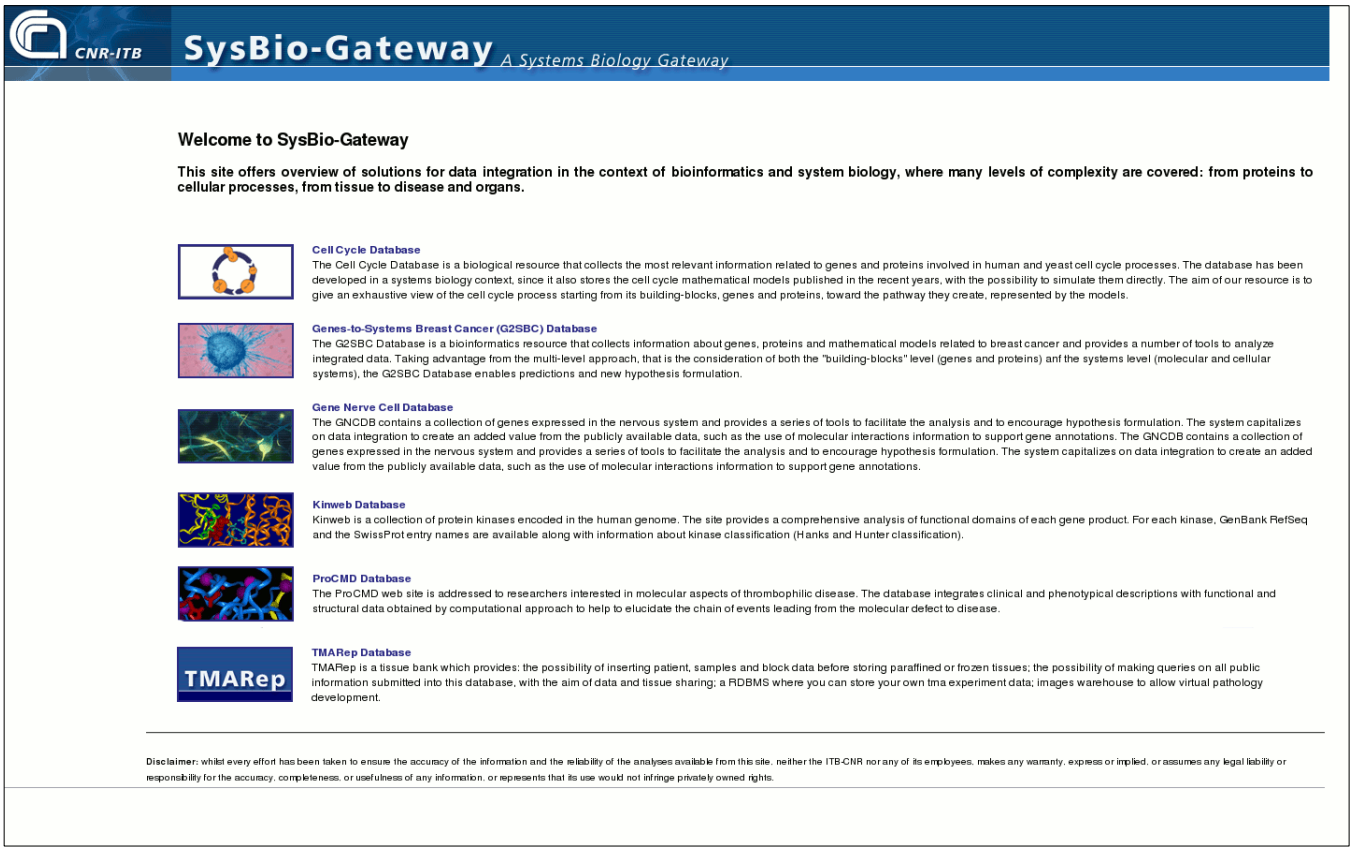

Figure 1: The SysBio-Gateway access page. Both a short description and direct link to each database held in the gateway is provided.