PHILIPPINE INFORMATION TECHNOLOGY JOURNAL. VOLUME 1, NUMBER 1, FEBRUARY 2008

# Semi-Automated Reconstruction of Biological Circuits

Rui Alves
Dept. Ciències Mèdiques Bàsiques,
Universidade de Lleida
Montserrat Roig 2
25008 Lleida, Spain
0034 973 702406
ralves@cmb.udl.cat

Ester Vilaprinyo
Dept. Ciències Mèdiques Bàsiques,
Universidade de Lleida
Montserrat Roig 2
25008 Lleida, Spain
0034 973 702406
evilaprinyo@cmb.udl.cat

Albert Sorribas
Dept. Ciències Mèdiques Bàsiques,
Universidade de Lleida
Montserrat Roig 2
25008 Lleida, Spain
0034 973 702406
Albert.sorribas@cmb.udl.cat

## ABSTRACT
Large amounts of molecular data regarding most aspects of cellular functioning are accumulating. These data range from sequence and structural data to gene and protein regulation data, including time dependent changes in the concentration of all cellular molecules. Integration of the different datasets through computational methods is required to efficiently organize and extract biological information that is relevant from a Systems Biology perspective.

In this paper we discuss how different computational tools and methods can be made to work together integrating different types of data, mining these data for biological information, and assisting in pathway reconstruction and biological hypothesis generation. We propose an algorithm for the integration of data and discuss an example of its application. This algorithm can then be used to generate testable biological hypothesis, creating an iterative theoretical/experimental loop.

## Categories and Subject Descriptors
D.3.3: Algorithms

## General Terms
Algorithms, Management, Measurement, Documentation, Performance, Design.

## Keywords
Biological circuit reconstruction. Iron-sulfur cluster Biogenesis.

## 1. INTRODUCTION
A consequence of the different high throughput (HTP) data sets that accumulate daily is that it is impossible for any one person to analyze and integrate them all. However, such integration is fundamental to reconstruct the molecular networks involved in cellular processes and obtain a systemic perspective of how those networks work. Thus, researchers need tools that assist with processing the information, filtering, organizing, and displaying it with appropriate representations.

The increasing availability of computational power and storage capabilities described for example by Moore's law makes such machines essential for the analysis and integration of large amounts of information. The increase in computer power facilitates the creation of CPU-demanding algorithms, software, and applications that can potentially manipulate the different available datasets, integrate the information they contain, and display the end result of the analysis in a user friendly format. Web based applications play an important role in making all these software tools available to most scientists. Efficient information mining and an appropriate display of the results facilitates *in silico* reconstruction of the molecular network that regulates and executes relevant Molecular Biology processes. *In silico* reconstruction ultimately generates hypothesis regarding the connectivity and dynamic behavior of pathways and circuits that must be validated against experimental predictions. Thus, pathway reconstruction is a fundamental step in applying a Systems Biology perspective in Molecular Biology, Biotechnology, and other areas.

Different types of reconstruction problems can be addressed *in silico*:

**(1) Identifying the pathways, genes and processes/functions that exist in a cell.** This provides information regarding possible ***qualitative*** systemic responses of the cell. For example, this type of problem includes whole genome network reconstruction (e.g. [1-2]). Placing the genes of a genome into pre-existing maps of metabolism, gene circuits and signal transduction, may reveal what a cell type can in principle do.

**(2) Reconstructing the detailed reaction network that exists within specific pathways or circuits** (e.g. [3-6]). This is an important problem because it may allow for more precise and ***quantitative*** predictions regarding how specific parts of the cellular response are regulated and executed. With the sequencing of metagenomes, many genes of unknown function and several previously unknown pathways are being discovered. Furthermore, novel pathways and new components of classical pathways are still being found in well studied organisms.

**(3) Reconstructing regulatory networks.** For example, identifying regulatory motifs in DNA suggests what transcription factors may regulate the expression of different pathways, and genes, thus identifying modules that are involved in specific processes (e. g. [7]). At the metabolic and signal transduction level, reconstruction of the detailed regulatory networks for the enzymes in a pathway is a requirement for accurate and quantitative prediction of the dynamic cellular behavior in response to environmental changes.

In the work presented here we discuss how semi automated reconstruction of different types of biological circuits can be achieved and present an example of *in silico* reconstruction for an ill-characterized pathway, that of iron-sulfur cluster (ISC) biogenesis in yeast.

## 2. DATA INTEGRATION FOR RECONSTRUCTION

Circuit reconstruction requires integration of knowledge at many different levels. We shall briefly characterize each of the various types of datasets that are available for *in silico* reconstruction of Molecular Biology circuits.

### 2.1 Literature Analysis

Bibliographic data has been accumulating now for more than a century. Databases such as MEDLINE or the Web of Science Citation Index collect and organize data from this published literature. Automated search engines can be used to identify documents in these databases that contain results on genes, pathways, and networks of interest in relevant cell types.

Recently, the application of automated literature analysis in metabolic reconstruction has become feasible. Existing public access servers can automatically identify genes that are referred to in the same papers [8-9]. This generates a network of co-occurrence of genes in papers that is often interpreted as implying functional interactions between genes. However, such a network should be viewed as a low level reconstruction of the molecular network that is involved in the processes for which those specific genes are important.

### 2.2 Sequence Data, Functional Data, and Structural Data in gene annotation for function

Databases of annotated gene and protein sequences facilitate the functional annotation of new genomes, through the use of homology sequence comparisons. If high sequence homology exists between a gene in a new genome and genes with clear functional annotation in other organisms, inferring the function of the new gene is almost trivial. The accumulated knowledge for the function of many genes and proteins has facilitated the creation of metabolic maps, signal transduction maps and gene circuit maps for the different genomes [e.g. 10-11]. In such maps, the individual function of a protein is superimposed onto the particular steps of the maps where that protein is active. Those maps facilitate reconstructing the molecular circuits of new genomes.

While annotating a genome and finding genes, gene circuit reconstruction can also be done. By searching for regulatory motifs upstream of gene promoters one can identify regulatory units within a genome. Furthermore, if those motifs are phylogenetically conserved in fairly close organisms, this provides additional support for the accuracy of predictions.

When no structural or sequence homology exists between a gene/protein and other of known function, sequence information can still be used to infer some functional information. For example, one can use phylogenetic conservation to investigate possible functions of the genes. The logic underlying phylogenetic conservation analysis is as follows. If a set of homologous genes with unknown function is present (absent) in the same set of genomes than other genes with known function, then it is likely that evolution acted simultaneously on that set of genes because somehow they share a function. A similar method for inferring function is that of finding gene fusion events.

### 2.3 Genomics, Proteomics and Metabolomics

Either high or low throughput gene expression data can be used to infer functional information for the genes. For example, if a gene/protein of unknown function is differentially regulated during the response to some stress, then one can infer that this gene is involved in the response to that stress. Although gene expression data can be used to infer functional involvement of a given gene in specific processes, in many cases the data will not reveal the specific function of that gene. Exceptions might be for example a situation where only one protein is missing in a circuit or pathway that is well known. In such cases, if only a gene of unknown function is identified in the expression data, its function is likely to be the one that is missing.

Furthermore, by identifying the genes whose expression is regulated under a given set of conditions, one can reconstruct the gene circuits that regulate the cellular response to those conditions. By using different types of mathematical analysis such data can be analyzed to predict which genes are involved in large transcriptional units that compose gene circuits from microarray data [e.g. 12].

HTP proteomic experiments can generate data for direct physical interaction between proteins. Finding which proteins interact physically with those of unknown function sheds light upon the processes in which the later proteins may be involved in, thus facilitating the reconstruction of their roles in the cell. Other proteomic experiments in which measurements of protein levels and activity are made can also assist in network reconstruction. This kind of information is still lacking for most organisms.

Metabolomics data can also, in principle, be used to reveal information regarding pathway and circuit connectivity [e.g. 13]. By measuring the different time scale of the changes in concentration of metabolites or signaling molecules, it may be possible to infer which reactions and which steps causally precede others in the connectivity of a network [14]. This information is difficult to derive from the mining of other types of databases discussed in the previous sections.

### 2.4 Mathematical Models

The *in silico* reconstruction of circuits, with a causal connectivity, may generate a conceptual model for the processes of interest. However, due to the extensive non-linearity of biological dynamics, adequateness of such hypothetical schemas to explain experimentally observed dynamics of a process can not often be evaluated by simple logic. Without this evaluation, one may be unable to distinguish the compatibility of alternative schema to explain available data. To avoid such a situation, the conceptual schemas can be used to create mathematical models whose dynamical behavior can be rigorously analyzed and compared to what is experimentally observed. This validation process is fundamental in testing the hypothesis generated during the reconstruction of networks. As discussed in the previous section, metabolomic data may play a very important role to finally identifying a reasonable network structure.

Statistical methods can be used for comparing how well alternative network connectivity can reproduce experimentally observed behavior for a given pathway, if sufficient experimental information is available. Such methods can either be numerical optimization algorithms that will decide which network best fits the known quantitative data or, they can be classification

algorithms that facilitate decision regarding comparison with qualitative experimental data.

As more complex problems are addressed, interest in using structured and systematic mathematical formalisms that a) facilitate automated model building and parameter estimation and b) facilitate model expansion and recycling is likely to increase.

## 2.5 Challenges to the Automated Computational Integration of HTP Data

Integrating the different datasets that are now available for pathway and circuit reconstruction in Molecular Biology is not trivial. The design of a global solution for this integration requires careful consideration of the goals, challenges, and limitations of both data sets and data mining methods. At a first glance, one may argue that the amount of data being generated by HTP methods is the most difficult challenge for the integration process. However, from the computational point of view, the available storage space and analytical capacity is far from being exhausted by the data generated through HTP approaches. Thus, from the computational point of view, data accumulation may not be a major issue. In our perspective, the major challenges are more likely to be:

**1) The information content**, organization, and deposition of a given type of data into databases. It is crucial that the relevant information needed to address a given biological problem is well organized and easily accessible.

**2) The formatting of the data**. Developing and applying standards in data formatting plays a central role in facilitating integrative approaches, because these standards facilitate the development of integrative computational methods.

**3) Lack of universal standards for the reporting.** The development and acceptance by the community of such standards may facilitate not only the analysis of experimental data that is deposited in databases but also the automation of literature data mining.

## 2.6 An integrated algorithm

Figure 1 outlines a flow chart for the integration of the different datasets in order to reconstruct different types of pathways and circuits. This process should be highly parallelized, allowing for flexible exclusion of one or more types of datasets and for an appropriate interaction with the user. A first layer of sequence data, structural data, literature data, genomics, proteomics and metabolomics data can be integrated to derive a second layer of functional data, pathway and gene circuit data, interaction data, and parameter data. Functional data, pathway and gene circuit data, interaction data, and parameter data can then be integrated to generate alternative conceptual schemas of how the pathways and circuits of interest may function. This second layer of data can also have direct experimental inputs, for example with new gene circuit maps that may have been experimentally reconstructed. The data from this second layer can then be itself integrated into schemas that can be used to interrogate pre-existing mathematical models or to generate new ones. Such models can be used to predict the dynamic behavior of the alternative schemas. That behavior can be compared to experimental behavior, thus validating some of the alternative conceptual maps.

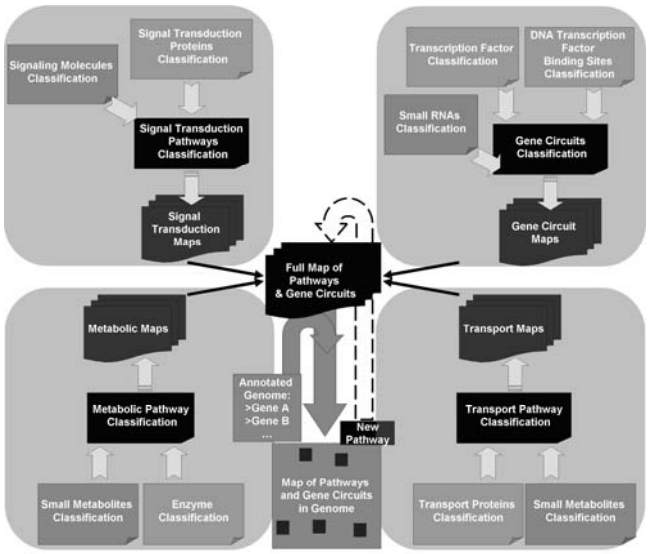


Figure 1: **Creating and integrating maps for pathways and gene circuits and applying them to the reconstruction of pathways in a genome.**

In some cases, the prediction that new genes or proteins are involved in a set of alternative networks for a given process may be tested directly without the need for a mathematical model. If experimental disruption of the protein activity or of gene expression affects the process in which the protein is involved, then it is likely that the prediction of involvement is correct. When more detailed hypothesis are being posed regarding the connectivity or dynamic behavior of the gene in the network, then the analysis of a mathematical model may often be necessary due to the non linear dependency of fluxes on proteins and metabolites in molecular biology processes. If the hypotheses regarding the reconstructed pathways and circuits are validated, they can then be refined. Otherwise, one must go back and reanalyze the data to derive new hypothesis.

## 2.7 Reconstruction of Fe-S Cluster (ISC) Biogenesis in Yeast

*S. cerevisiae* is the eukaryotic organism in which the ISC biogenesis has been more extensively studied. The following proteins are known to be involved in this biogenesis: Arh1, Yah1, Yfh1, Isu1, Isu2, Isa1, Isa2, Nfu1, Nfs1, Isd11, Mge1, Ssq1, Jac1, Atm1 and Grx5. The exact role of many of these is unclear. Combining automated literature analysis, phylogenetic profiling, structural bioinformatics, *in silico* docking and mathematical modeling we have generated alternative mathematical models for the causal network of this pathway. Extensive parameter scanning followed by comparison of model behavior to experimental results suggests that the network shown in Figure 2 is the most likely for the pathway. Most of the network predictions made in

this work agree with the experimental data available for the system [4].

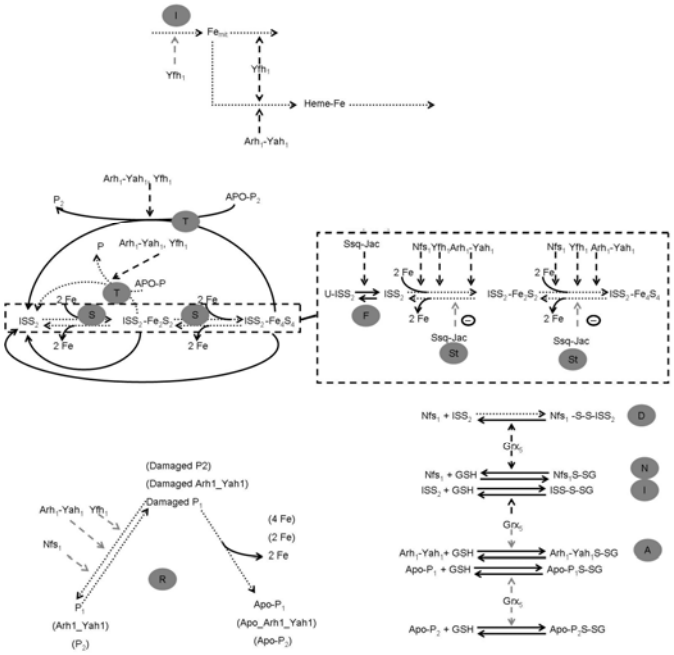


Figure 2: **Network model for ISC biogenesis in *S. cerevisiae*. A - recovery of Arh1 by Grx5. D - recovery of the dead end complex between Nfs1 and the scaffold by Grx5. F - folding. FI - Fe import. I - recovery of the scaffolds by Grx5. N - Nfs1 recovery by Grx5. R - repair of the clusters. S - synthesis of ISC. St - stabilization of the ISC assembled on the scaffolds. T - transfer of the ISC to apo-proteins. Dotted arrows represent reactions that have been observed to occur experimentally. Dashed arrows represent the alternative modes of regulatory action for the different proteins. Light grey arrows represent the network interactions that are not likely to exist in the pathway, according to our analysis.**

## 3. DISCUSSION

In this work we propose an *in silico* algorithm for reconstruction of cellular pathways. We discuss types of available data sets that can be used for such reconstruction. We present a specific example of how those datasets have been integrated to build a model for ISC biogenesis in yeast.

The example discussed here has been implemented using off-the-shelf software that is freely available, and integrating the results from the different datasets manually. Currently in our lab we are developing a server that will automate this integration. We will still rely on off-the-shelf software for the calculations done in many of the modules described in Fig. 1 (e.g. GRAMM for docking, ROSETTA code for structural modeling, etc.), but some of the modules will have an in-house developed motor (e.g. Literature analysis, Mathematical Modeling set up and analysis). Some of these modules are already implemented and running in beta testing mode at the University of Lleida intranet, where any member of the university can use them. Once each module passes sufficient quality control it will be made available over the www.

## 4. ACKNOWLEDGMENTS

This work has been partially supported by grants BFU2005-0234BMC and BFU2007-62772/BMC of the Spanish Ministerio de Educación y Ciencia. RA was supported by a Ramon y Cajal award from the Spanish Ministerio de Educacion y Ciencia.

## 5. REFERENCES

[1] Feist, AM, Scholten, JC, Palsson, BO, Brockman, FJ,Ideker, T. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol Syst Biol* **2006**; 2: 2006 0004

[2] Notebaart, RA, van Enckevort, FH, Francke, C, Siezen, RJ,Teusink, B. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics* **2006**; 7: 296

[3] Panina, EM, Vitreschak, AG, Mironov, AA,Gelfand, MS. Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. *Fems Microbiology Letters* **2003**; 222: 211-220

[4] Alves, R,Sorribas, A. In Silico Pathway Reconstruction: Iron-Sulfur Cluster Biogenesis in Saccharomyces cerevisiae. *BMC Systems Biology* **2007**; 1: 10

[5] Workman, CT, Mak, HC, McCuine, S, Tagne, JB, Agarwal, M, Ozier, O, Begley, TJ, Samson, LD,Ideker, T. A systems approach to mapping DNA damage response pathways. *Science* **2006**; 312: 1054-1059

[6] Haugen, AC, Kelley, R, Collins, JB, Tucker, CJ, Deng, C, Afshari, CA, Brown, JM, Ideker, T,Van Houten, B. Integrating phenotypic and expression profiles to map arsenic-response networks. *Genome Biol* **2004**; 5: R95

[7] Herrgard, MJ, Lee, BS, Portnoy, V,Palsson, BO. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. *Genome Res* **2006**; 16: 627-635

[8] Hoffmann, R,Valencia, A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* **2005**; 21 Suppl 2: ii252-ii258

[9] Stapley, BJ,Benoit, G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac Symp Biocomput* **2000**: 529-540

[10] Nikitin, F, Rance, B, Itoh, M, Kanehisa, M,Lisacek, F. Using protein motif combinations to update KEGG pathway maps and orthologue tables. *Genome Inform* **2004**; 15: 266-275

[11] Paley, SM,Karp, PD. The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res* **2006**; 34: 3771-3778

[12] Yugi, K, Nakayama, Y, Kojima, S, Kitayama, T,Tomita, M. A microarray data-based semi-kinetic method for predicting quantitative dynamics of genetic networks. *BMC Bioinformatics* **2005**; 6: 299

[13] Marino, S,Voit, EO. An automated procedure for the extraction of metabolic network information from time series data. *J Bioinform Comput Biol* **2006**; 4: 665-691

[14] Vance, W, Arkin, A,Ross, J. Determination of causal connectivities of species in reaction networks. *Proc Natl Acad Sci U S A* **2002**; 99: 5816-5821