19

Data Mining in the Molecular Biology Era – A Study Directed to Carbohydrates Biosynthesis and Accumulation in Plants

Renato Vicentini and Marcelo Menossi
Universidade Estadual de Campinas, Departamento de Genética e Evolução
Brazil

1. Introduction

The last revolutionary advance in biological research was driven by the concepts of molecular biology, which links information about genetic traits to DNA and proteins (Katagiri, 2003). The central dogma of molecular biology is that information stored in DNA is processed through RNA to produce proteins that execute various cellular functions. At the basic hierarchical level, information flow is from the genes, to mRNA, and proteins. The current adopted approaches are based in the attribution of the biological phenomena to the actions of one or a few genes. Unfortunately with these approaches it is difficult to reconstitute a model for a whole biological system by simple combining the information they generate. Recently, ideas about linear flow of information have been revised to permit the development of a more integrated view of cellular functions as being distributed among groups of elements that all interact within large networks.

Over the last years, there has been an explosion of information in biology. The sequencing of more than a hundred genomes has detailed thousands of genes. High-throughput technologies, especially DNA arrays, generate information about the expression of these genes under different conditions (Ideker et al, 2001). The next step towards a more comprehensive understanding of the biological system is to integrate these data into a conceptual framework. The ultimate goal is to understand biological systems in sufficient detail to enable accurate and quantitative predictions about the behaviors of biological systems, including predictions of the effects of modifications of the system (Katagiri, 2003).

Systems biology applies the methods of biology, mathematics, computer science, engineering, and physics to understanding living systems. Developing accurate predictive methods capable of scaling from genotype to phenotype can be approached through systems biology coupled with genomics and gene expression data (Fig. 1). The way to build bridges from molecular biology to physiology is to recognize that a network of interacting genes and proteins is a dynamic system evolving in space and time according to laws of reaction, diffusion and transport (Tyson, 2007).

A system can be generally defined as a network of interacting elements receiving certain inputs and producing certain outputs. Quantitative models are generated as tools to aid understanding and prediction of system output in response to the environment and system inputs. Models are simplified representations of system dynamics, usually in mathematical

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

form (Janes & Yaffe, 2006). In biology, a system can be described equally at the level of gene action, biochemical pathway, an organelle, a cell, an organ, a whole organism, or a community (Hammer et al., 2004).

2. Genomics and others 'omics' approaches

The use of high-throughput technologies in recent year has generated extensive information on the various levels of cellular and developmental process in many organisms. The major challenge, however, remains in the integration of this information towards a broad understanding on how the different biological layers interact to form higher functional units (Girke, 2003).

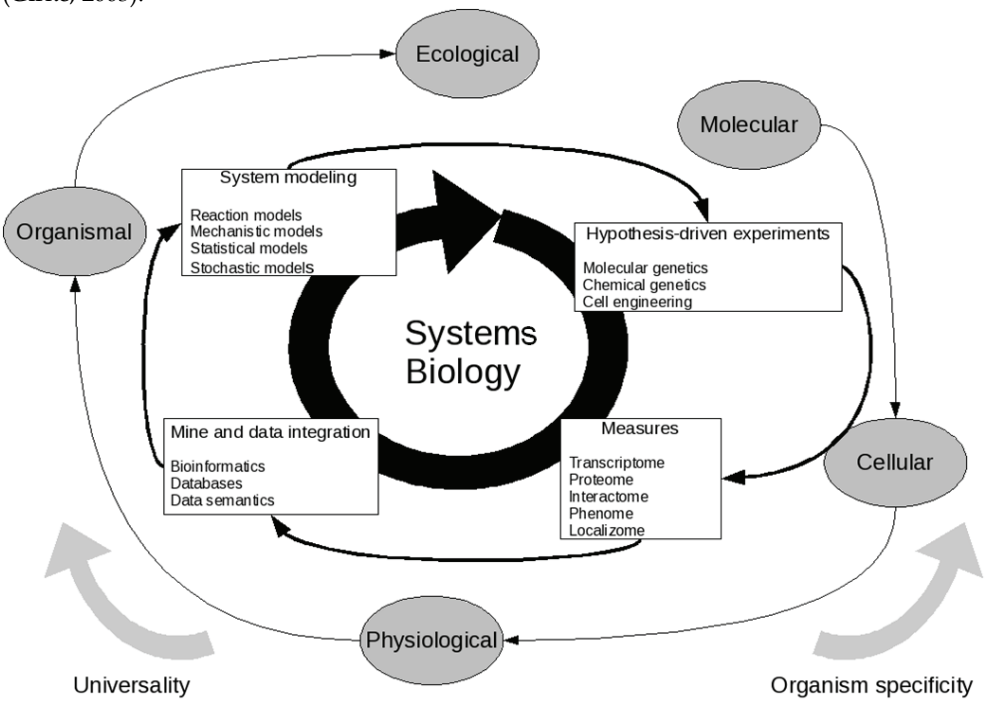

Fig. 1. Systems biology viewed as a combination of 'omic' approaches, mine and data integration, system modeling and hypothesis-driven experiments. The goal of systems biology is to generate a model of the whole organism that describes processes across the layers of biological organization (molecular, cellular, physiological, organismal, and ecological). With the availability of complete genome and transcriptome sequences, functional genomics and proteomic approaches are used to map the transcriptome, proteome, interactome, phenome, and localizome of a given organism. Computational methods can be then used to model biological process based on integrated data. The hypotheses lead to the design of new experiments that start a new round of the cycle. Although the individual components (in molecular level) are unique to a given organism at a particular time, the grouped components (cellular, physiological, and organismal) share similarities with other large-scale processes.

Using genomic techniques, we can now identify all the genes in an organism. Moreover, using microarray and proteomic techniques, we now have the ability to resolve which genes are activated or inactivated during development or in response to an environmental change. The DNA arrays approach has allowed investigators to evaluate simultaneously thousands of genes, measuring which ones are turned on or turned off in a genome in response to an experimental treatment. Proteomic methods reveal the proteins translated from the mRNA molecules that are the direct result of gene expression, and can be used to determine not just whether a protein is present but how much of the protein is present and in some cases how active it is. On a genome-wide scale, combining data from several unrelated measure profiling experiments can result in more detailed and informative module assignments (Ge et al., 2003). Such integration should not only improve functional annotation but also help to formulate biological hypotheses. However, identifying all the genes and proteins in an organism is comparable with listing all the parts of a machine. Although such a list provides a catalog of the individual components, it is not sufficient to understand the complexity underlying the engineered object (Minorsky, 2003). The massive acquisition of data in molecular and cellular biology has led to the simulations of biological systems. Simulations, increasingly paired with experiments, are being successfully and routinely used by computational biologists to understand and predict the quantitative behavior of complex system, and to drive new experiments (Di Ventura, 2006).

3. Data mining and modeling

Data mining refers to the analysis of large-scale data sets for the purpose of general inference and the extraction of specific information that relates to some initial data of interest (Kersey, 2006). Data mining for large data sets can be quite different from extracting data about individual sequences. The types of analysis performed are generally similar, but the development of automated procedures is usually essential as the data volume is increased. Large data sets enable 'knowledge discovery' through the identification of patterns within the data.

The analysis of biological systems requires extensive use of bioinformatics resources for data management, mining, and modeling. An additional dimension of complexity will be added by the incoming data from new technologies. The importance of automatic methods for linking gene, protein and literature data has grown as the number of known sequences has increased (Kersey, 2006; and Fig. 2). To manage these multidimensional data sets, it will be necessary to develop a new generation of integrated databases to allow complex queries across diverse types combined with new algorithms and flexible software for mining and simulating network architectures (Girke, 2003).

3.1 'Omics' data mining

DNA array analysis is exploratory and very high dimensional, and the primary purpose is to generate a list of differentially regulated genes that can provide insight into the biological phenomena under investigation. While it is possible to interpret DNA array experiments a single gene at a time, most studies generate long lists of differentially expressed genes whose interpretation requires the integration of prior biological knowledge. This prior knowledge is stored in various public and private databases (Fig. 2) and covers several aspects of gene function and biological information (Coulibaly & Page, 2008). Below are described the main features of the types of bioinformatics tools and analysis that permit mining the microarray data (Coulibaly & Page, 2008).

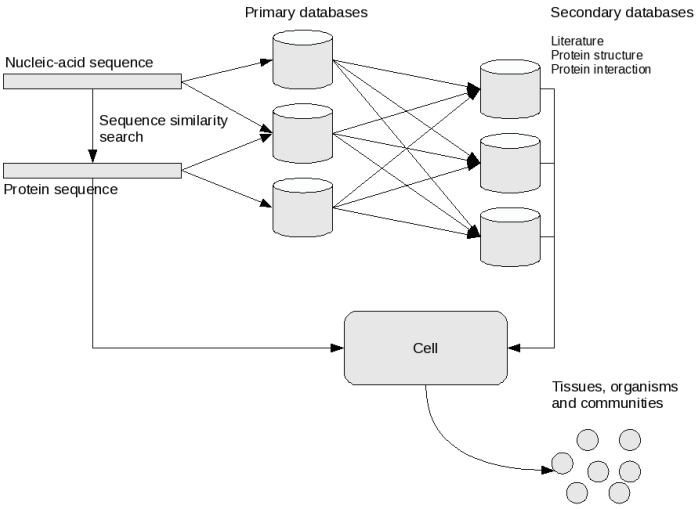

Fig. 2. A schematic representation of a workflow for bioinformatics analysis. The workflow performs the analysis from sequence to functional annotation, protein structure and literature. The complete analysis may be carried out entirely within a bioinformatics warehousing system, or as a sequence of separate operations performed in different environments. Starting with the sequence of a gene or protein, identical and/or similar sequences are identified in the primary databases. The database records describing these sequences also contain general information about the sequence, and accurate links to secondary database. Finally, modeling analyzes are performed to build a cell model than can be expanded to a complex systems biology modeling of tissues, organisms or communities.

Functional annotation tools. The goal of these tools is to relate the expression data to other attributes such as cellular localization, biological process, and molecular function. The most common way to functionally analyze a gene list is to gather information from the literature or from databases covering the whole genome.

Gene coexpression analysis tools. In most DNA array studies, gene expressions are measured on a small number of arrays or samples; however, large collections of arrays are available in several databases. These tools provide the opportunity to analyze the transcriptome by pooling gene expression information from multiple data sets. It has been demonstrated that genes which protein products cooperate in the same pathway.

Gene network analysis. Genes and their protein products are related to each other through a complex network of interactions. By using systems biology approach we can analyze the behavior and relationships of all the elements in a particular biological system to arrive at a more complete description of how the system functions. These analyses permit the development of models for gene regulatory networks, the display of a simplified view of large amount biological components, and the associate network nodes and edges with biological information.

*Biological pathway resource*s. One of the downstream applications of the reconstruction of a gene regulatory network or the identification of clusters of functionally related genes is to associate the genes and their interconnections with known metabolic pathways.

3.2 Literature data mining

The systematic application of automated high-throughput molecular biology techniques has led to the generation of an immense quantity of data. However, the interpretation of these data is still dependent on inference drawn from hypothesis-driven experimentation, the details of which reside in free-text articles. The value of bioinformatics data is thus utterly dependent on the ability to make the correct links from the sequences to the scientific literature (Kersey & Apweiler, 2006).

Text mining refers to computational methods for the automatic analysis of semi-structured text, and has gained considerable attention in recent years in the molecular biology field (Hakenberg et al., 2004). Literature data mining has progressed from simple recognition of terms to extraction of interaction relationships from complex sentence (Hirschman et al., 2002). The current research focused on tasks needing limited linguistic context and processing at the level of words, like identifying protein names or on tasks relying on word co-occurrence and pattern matching.

3.3 System modeling

Biologists commonly use the term 'model' for verbal or graphical description of a mechanism underlying a cellular process. However, the mathematical modeling and description of complex biological process has become more important in the last years. Knowledge about the system is essential and needs to be formalized for the chosen framework (Di Ventura, 2006).

Clearly, useful modeling will depend on having a large amount of high-quality quantitative information about all aspects of biological processes, and many new types of data have to be systematically determined. Recently, Bayesian network, a probabilistic graphic model representation, has been widely used to analyze expression data. Compared with clustering analysis, Bayesian network has the advantage of uncovering conditional independency among genes, which provides a promising way to survey direct interaction of gene regulation (Chen et al., 2006). A complete understanding of regulation requires quantitative information about kinetic laws and the concentrations of metabolites and enzymes. This quantitative knowledge in combination with the known network of metabolic pathways allows the construction of mathematical models that describe the dynamic changes in metabolite concentrations over time. A variety of pathways modeling tools such a CellDesigner (Funahashi et al., 2003) has been developed which simplify model construction and analysis. Most of these tools are able to store and exchange models in the Systems Biology Markup Language (SBML, Hucka et al., 2003) and to fit parameters for a given set of experimental data.

4. Systems biology

The emerging field of systems biology is a new branch of biology that attempts to discover and understand biological properties that emerge from the interactions of many elements (Minorsky, 2003). Systems biology examines the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism. These systems may be gene expression networks, signal transduction pathways, metabolic networks, or combinations of them. In contrast to previous approaches, systems biology endeavors to quantitatively model and simulate complex biological process and systems comprising thousands of chemical compounds and reactions.

There is an emphasis on complexity and large data sets, which are typically produced by a variety of high-throughput genomic, proteomic and metabolomic techniques. The major reason behind the increasing interest in systems biology is that progress in molecular biology, particularly in genomics, proteomics, and high-throughput measurements, is enabling scientists to collect comprehensive data sets on the mechanisms underlying responses to perturbations on a biological system (Minorsky, 2003).
A systems approach to understanding biology can be described as an interactive process that includes (1) data collection and integration of all available information, (2) system modeling, (3) experimentation at a global level, and (4) generation of new hypotheses (Gutiérrez et al., 2005; and Fig. 1).

5. Case study of carbohydrates biosynthesis and accumulation in plants

Recently, genes encoding the enzymes of carbohydrates biosynthesis in plants have been isolated, cloned, and used in experiments to transform the plant to increase or decrease expression of the enzyme with the goal of altering the carbohydrates accumulation. However, results of this reductionist approach towards understanding sucrose accumulation have fallen short of expectations, mainly because of the complex interactions among the multitude of simultaneous processes. Insights into the complex interactions will require systems-level approaches, from the molecular, biochemical, and physiological levels. Carbohydrates accumulations in plants are the products of a large complex network of interactions that can be analyzed from several perspectives. Increasing evidence suggests that sucrose is involved in signaling to modulate expression of genes controlling transporters and storage proteins, division and differentiation of cells, and accumulation of storage products (Lunn and MacRae 2003). Each of the reactions involved is controlled by activation of specific genes in response to an interaction of the genotype of the plant and the environment.

5.1 Approach

Availability of abundant, high-quality data sets from DNA array expression experiments has stimulated rapid progress in gene networks analysis for a variety of plant species. By examining correlated expression patterns between genes, we can infer new functions for previously uncharacterized genes and identify potential causal relationships between regulators and their targets (Srinivasasainagendra et al., 2008). The idea that correlated expression implies biologically relevant relationships between gene products were use in the present study.
We use the AraCyc database to select the sucrose biosynthesis and degradation pathways. The AraCyc database (Table 1) is a reference database for visualization of *Arabidopsis thaliana* biochemical pathways. With the selected enzymes from the sucrose AraCyc pathways, we performed a search by coexpressed genes in the *Arabidopsis* mDNA arrays experiments storage in the Nottingham Arabidopsis Stock Center (NASC). For this purpose we utilized the Cress-express mining tool (Table 1) with the RMA processing method in four specific carbohydrates experiments. Cress-express estimates the coexpression between an user-provided list of genes and all genes from Affymetrix Ath1 platform using up to 1779 arrays. Cress-express also performs pathway-level coexpression (PLC). PLC identifies and ranks genes based on their coexpression with a group of genes. The tool has the data processed with a variety of image processing methods: RMA, MAS5, and GCRMA (Srinivasasainagendra et al., 2008).

With the identified coexpressed genes, we constructed a Bayesian networks by using the BNArray tool (Table 1). It allows the reconstruction of significant submodules within regulatory networks using an extended subnetwork mining algorithm. Under this framework and the assumption of parameter independence, an initial Bayesian network structure is learned from the training data and a user specified prior network. From this initial network, greedy search algorithm with random restarts was performed to get the highest score posterior network to avoid local maxima. Finally, we obtained an optimized Bayesian network that maximizes the Bayes factor.
We utilized the Kinetikon and SABIO-RK softwares (Table 1) for mining kinetic reactions in the biological literature. Finally, the network representation of carbohydrates biosynthesis pathways of coexpressed genes was created by Cytoscape network tool (Table 1), and the simulation of the kinetic dynamics of the network created was modeling by the CellDesigner (Table 1).

Tools and databases	Description	Reference
AraCyc	*Arabidopsis thaliana* database of metabolic pathways and enzymes.	(Mueller et al., 2003)
Cress-express	Coexpression analysis tool for Arabidopsis microarray expression data that computes patterns of correlated expression between user-entered query genes and the rest of the genes in the genome.	(Srinivasasainagendra et al., 2008)
BNArray	R package for construct gene regulatory networks from microarray data by using Bayesian network.	(Chen et al., 2006)
CellDesigner	Structured diagram editor for drawing gene-regulatory and biochemical networks based on standardized technologies.	(Funahashi et al., 2003)
Cytoscape	Software environment for the large scale integration of molecular interaction network data.	(Shannon et al., 2003)
KEGG	Kyoto Encyclopedia of Genes and Genomes that link lower-level information (i.e. proteins) with higher-level information (i.e. pathways).	(Kanehisa et al., 2000)
Kinetikon	A collection of detailed knowledge about biochemical reaction kinetics.	(http://kinetikon.molgen.mpg.de)
SABIO-RK	Web-based application that contains information about biochemical reactions, their kinetic equations with their parameters.	(Wittig et al., 2006)

Table 1. Software and databases adopted in this study.

5.2 Results

We used the AraCyc database to look up AGI codes for genes associated with the sucrose biosynthesis and degradation pathway. We used Cress-express to determine the degree to

which these genes are coexpressed with other. Using Cress-express tool default parameters, we performed a coexpression analysis of all genes, comparing them both to each other as well as to all other genes represented on the ATH1 array. The Fig. 3 presents the diagrams showing the network obtained by the coexpression genes related with the genes present in the pathways. Each connection in the network represents a pair of connected genes exhibit expression correlation. This study revealed that many genes are highly coexpressed with each other.

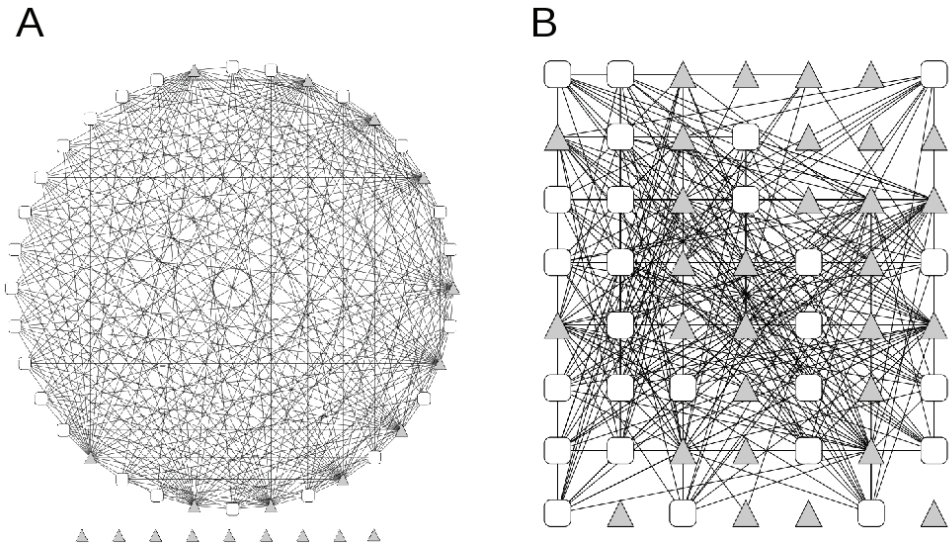

Fig. 3. Network representation of protein coexpressed with proteins in sucrose biosynthesis pathway (A) and sucrose degradation pathway (B). In these networks, the 'bait' genes are represented by triangles.

Bayesian network modeling was used to capture regulatory interactions between genes based on genome-wide expression measurements. To build the network, we integrated expression data from 'bait' genes (from biosynthesis pathway) with others coexpressed genes in the *Arabidopsis* genome. The generated network (Fig. 4) is a set of relationships that defines the probability that genes function together; this regulatory interactions among genes and their directions are derived from expression data. The subnetwork in Fig. 4 is the one that best explains the observed data, and provides the direction of regulatory interactions.

We use text mining systems (Table 1) that supports researchers in their search for experimentally obtained parameters to build our kinetic model. The kinetic modeling of biological reaction networks deals with the question of how the concentrations of substances change over time. The dynamics are determined by (a) the concentrations of substrates and products, (b) the structure of the whole reaction network, and (c) the kinetic parameters of the involved enzymes (Hakenberg et al., 2004).

We based our model (Fig 5) on the knowledge contained in the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa & Goto, 2000). KEGG is a set of databases that constitute a computer representation of biological knowledge at different levels, i.e.

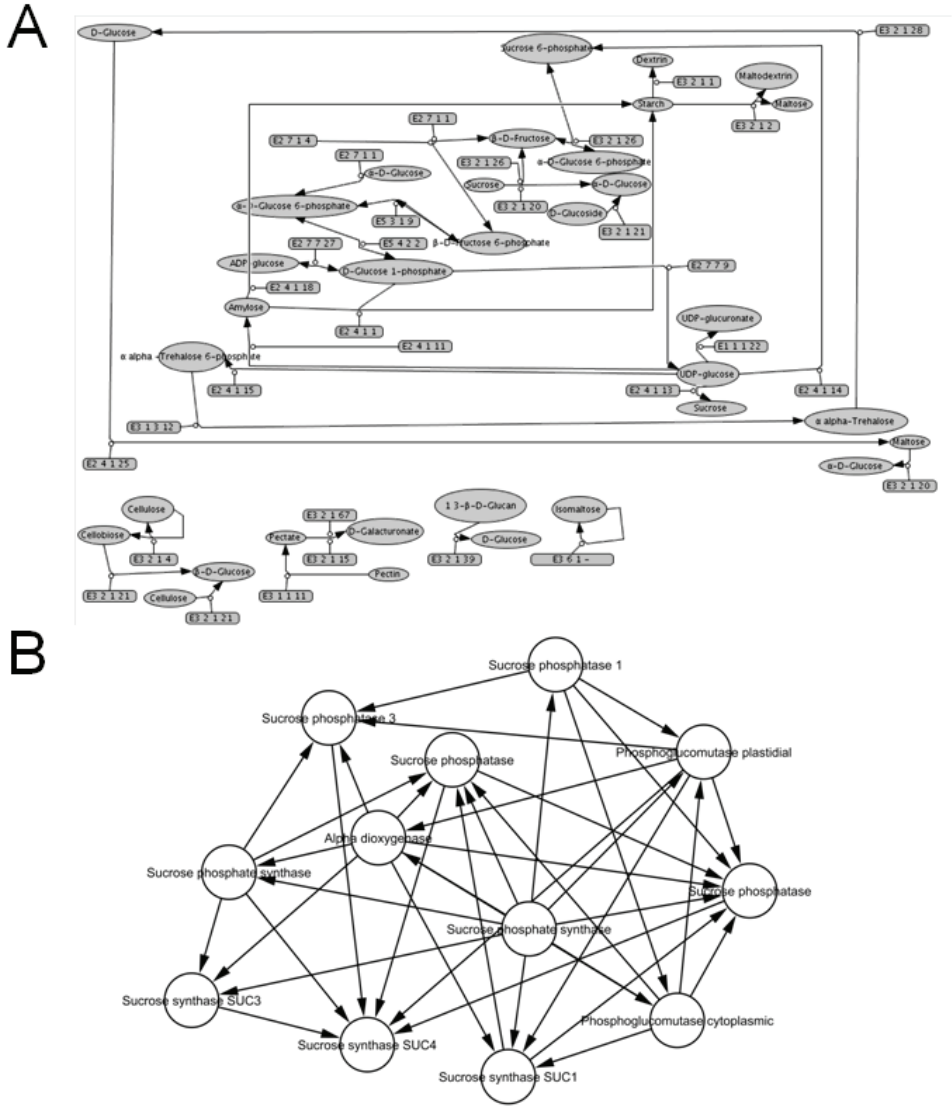

Fig. 4. Model for sucrose biosynthesis (A) and Bayesian subnetwork create from coexpressed genes of this pathway (B).

pathways, reactions, enzymes, compounds and genes. Our metabolic network was built from chemical reactions stored in KEGG (Fig. 4A). We start with a set of reactions (Table 2), with these reactions identifiers we retrieve information on the relevant compounds and enzymes. Each reaction in the model has a reaction identifier, an enzyme identifier and a metabolite identifier associated with it.

UTP + D-Glucose 1-phosphate <=> Pyrophosphate + UDPglucose
UDPglucose + D-Fructose 6-phosphate <=> UDP + Sucrose 6-phosphate
UDP-D-glucose + D-Fructose 6-phosphate <=> UDP + Sucrose 6-phosphate
Sucrose 6-phosphate + H2O <=> Sucrose + Orthophosphate
Sucrose 6-phosphate + H2O <=> Neohancoside D + Orthophosphate
UDPglucose + D-Fructose <=> UDP + Sucrose
UDP-D-glucose + D-Fructose <=> Neohancoside D + UDP
alpha-D-Hexose 1-phosphate <=> alpha-D-Hexose 6-phosphate
D-Ribose 1-phosphate <=> D-Ribose 5-phosphate
alpha-D-Ribose 1-phosphate <=> alpha-D-Ribose 5-phosphate

Table 2. Reactions list used to build the kinetic model.

Kinetic modeling of biological systems depends on sets of different kinetic data and values measured in expensive experiments. Such data are published in thousands of scientific articles. It is infeasible for humans to read and analyze this number of papers with reasonable time constraints (Hakenberg et al., 2004). To simulate our model we mined literature and databases for kinetic data using the SABIO-RK and Kinetikon databases. Finally, we applied the kinetic data to the model of carbohydrate biosynthesis pathway (Fig. 5). The result is the concentration curve showed in Fig. 6.

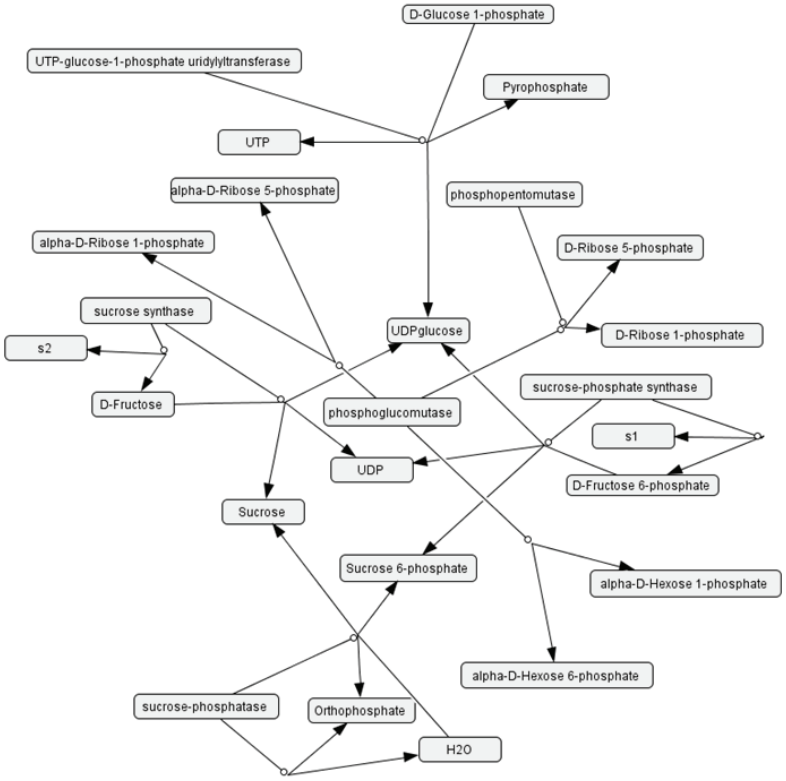

Fig. 5. Kinetic model for reaction present in sucrose biosynthesis.

6. Conclusion

The immensity of the information explosion in biology presents the challenge of how these data sets will be organized and mined in standard and easily accessible forms. To successfully iterate through the cycle described in Fig. 1, we need high-quality quantitative data and a flexible software platform that integrates arbitrary data types and that is coupled to data visualization and analysis tools. It will allow scientists to study and understand biological dynamics, to create a detailed model of cell function, and to provide system level knowledge for the network of signaling that are essential for physiological function. To reach this goal, we must adopt mathematical and computational methods for modeling and simulating complex biological systems (Girke, 2003).

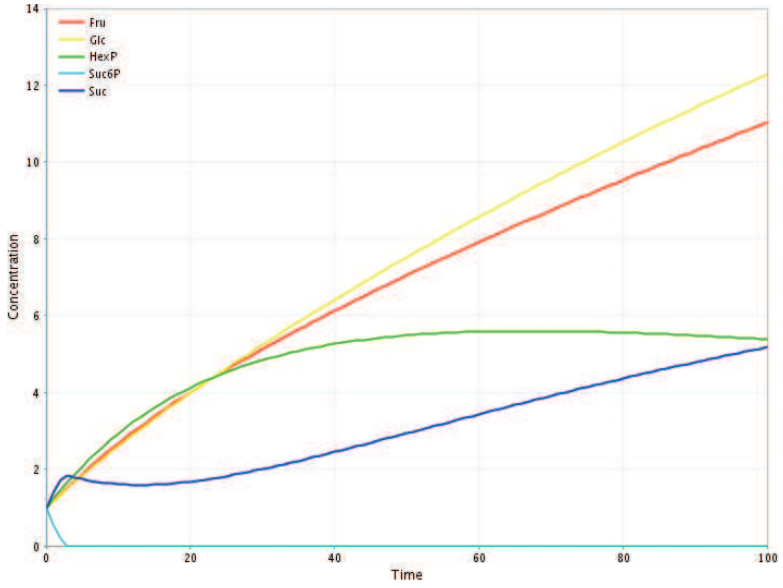

Fig. 6. Simulated concentration curve (in time) for metabolites of sucrose biosynthesis pathway. Fru: fructose; Glc: glucose; HexP: hexose-phosphate; Suc6P: sucrose 6-phosphate; Suc: sucrose.

7. References

Chen, X.; Chen, M. & Ning, K. (2006). BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics*, 22, 2952-2954, ISSN 1367-4803

Coulibaly, I. & Page, G.P. (2008) Bioinformatic tools for inferring functional information from plant microarray data II: Analysis beyond single gene. *International Journal of Plant Genomics*, 893-941, ISSN 1687-5370

Di Ventura, B.; Lemerle, C.; Michalodimitrakis, K. & Serrano, L. (2006). From *in vivo* to *in silico* biology and back. *Nature*, 443, 527-533, ISSN 0028-0836

Funahashi, A.; Tanimura, N.; Morohashi, M. & Kitano, H. (2003). CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1, 159-162, ISSN 1741-8364

Ge, H.; Walhout, A.J.M. & Vidal, M. (2003). Integrating 'omics' information: a bridge between genomics and systems biology. *TRENS in Genetics,* 19, 551-559, ISSN 0168-9525

Girke, T.; Ozkan, M.; Carter, D. & Raikhel, N.V. (2003). Towards a modeling infrastructure for studying plant cells. *Plant Physiology*, 132, 410-414, ISSN 0032-0889

Gutiérrez, R.A.; Shasha, D.E. & Coruzzi, G.M. (2005). Systems biology for the virtual plant. *Plant Physiology*, 138, 550-554, ISSN 0032-0889

Hakenberg, J.; Schmeier, S.; Kowald, A.; Klipp, E. & Leser, U. (2004). Finding kinetic parameters using text mining. *OMICS*, 8, 131-152, ISSN 1536-2310

Hammer, G.L.; Sinclair, T.R.; Chapman, S.C. & van Oosterom E. (2004). On systems thinking, systems biology, and the *in silico* plant. *Plant Physiology*, 134, 909-911, ISSN 0032-0889

Hirschman, L.; Park, J.C.; Tsujii, J.; Wong, L. & Wu, C.H. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18, 1553-1561, ISSN 1367-4803

Hucka, M.; Finney, A.; Sauro, H.M.; Bolouri, H.; Doyle, J.C.; Kitano, H. et al. (2003) The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19, 524-531, ISSN 1367-4803

Ideker, T.; Thorsson, V.; Ranish, J.A.; Christmas, R.; Buhler, J.; Eng, J.K.; Bumgarner, R.; Goodlett, D.R.; Aebersold, R. & Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292, 929-934, ISSN 0036-8075

Janes, K.A. & Yaffe, M.B. (2006). Data-driven modeling of signal-transduction networks. *Nature Reviews Molecular Cell Biology*, 7, 820-828, ISSN 1471-0072

Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research,* 28, 27-30, ISSN 1362-4962

Katagiri, F. (2003). Attacking complex problems with the power of systems biology. *Plant Physiology*, 132, 417-419, ISSN 0032-0889

Kersey, P. & Apweiler, R. (2006). Linking publication, gene and protein data. *Nature Cell Biology*, 8, 1183-1189, ISSN 1465-7392

Lunn, J.E. & MacRae, E. (2003). New complexities in the synthesis of sucrose. *Current opinion in plant biology*, 6, 208-214, ISSN 1369-5266

Minorsky, P.V. (2003). Achieving the in silico plant. System biology and the future of plant biological research. *Plant Physiology*, 132, 404-409, ISSN 0032-0889

Mueller, L.A.; Zhang, P. & Rhee, S.Y. (2003). AraCyc: A biochemical pathway database for Arabidopsis. *Plant Physiology,* 132, 453-460, ISSN 0032-0889

Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B. & Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*. 13, 2498-2504, ISSN 1088-9051

Srinivasasainagendra, V.; Page, G.P.; Mehta, T.; Coulibaly, I. & Loraine, A.E. (2008) CressExpress: A tool for large-scale mining of expression data from Arabidopsis. *Plant Physiology*, 147, 1004-1016, ISSN 0032-0889

Tyson, J.J. (2007). Bringing cartoons to life. *Nature*. 445, 823, ISSN 0028-0836

Wittig, U.; Golebiewski, M.; Kania, R.; Krebs, O.; Mir, S.; Weidemann, A.; Anstein, S.; Saric, J. & Rojas I. (2006) SABIO-RK: Integration and curation of reaction kinetics data. In: *Data Integration in the Life Sciences*, 94-103, Springer Berlin / Heidelberg, ISBN 978-3-540-36593-8

Data Mining and Knowledge Discovery in Real Life Applications

Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Renato Vicentini and Marcelo Menossi (2009). Data Mining in the Molecular Biology Era — A Study Directed to Carbohydrates Biosynthesis and Accumulation in Plants, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-53-0, InTech, Available from: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/data_mining_in_the_molecular_biology_era_-_a_study_directed_to_carbohydrates_biosynthesis_and_accu

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821