Meeting report

Modeling molecular networks: a systems biology approach to gene function

Alessandro Guffanti

Address: FIRC Institute of Molecular Oncology, Via Adamello 16, 20154 Milano, Italy. E-mail: guffanti@ifom-firc.it

Published: 16 September 2002

Genome ***Biology*** 2002, **3(10)**:reports4031.1–4031.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/10/reports/4031

A report on the European Science Foundation Workshop on Modeling of Molecular Networks, Granada, Spain, 11-14 June 2002.

Bioinformatics has for a long time been a discipline based on the comparison of gene and protein sequences, with the aim of discovering evolutionary relationships (and hence comparable function). There is now a growing interest in a more systems-based approach, where the focus is on how molecules and pathways interrelate and on the identification of the 'minimal building blocks', that is, molecules that exchange a common substrate or participate in the formation of large supramolecular complexes. This ESF workshop on modeling molecular networks gathered an ensemble of speakers who illustrated to the audience both theoretical and experimental approaches to systems bioinformatics, highlights of which are presented here.

Protein-protein interactions

An innovative experimental approach to the determination of protein-protein interactions in yeast cells was presented by Anne-Claude Gavin (Cellzome AG, Heidelberg, Germany). Assemblies of ten or more proteins form protein complexes that dominate biological processes. Cellzome started with 6,229 *Saccharomyces cerevisiae* open reading frames (ORFs) and systematically purified the multiprotein complexes formed around these gene products using the Tandem Affinity Purification (TAP) tagging strategy. TAP permits a high-affinity purification step followed by very specific and mild elution of multiprotein complexes. A total of 2,700 different proteins were identified by matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS) in complexes containing from 2 to 83 proteins. The data cover around 75% of the known components of protein complexes, and 90% of the characterized complexes have one or more novel components. By using the 'guilt by association' concept, it is possible to propose a function for 820 proteins that had no previous functional annotation. Around 80% of the analyzed proteins are part of multiprotein complexes. The analysis of protein complexes that include yeast proteins that have a human ortholog suggest that conservation between species extends from proteins to their molecular environment, and that the overall proteome architecture is probably conserved. Many of the identified proteins are found in multiple complexes and it would be exciting to try to correlate these ubiquitous factors with essential genes using *S. cerevisiae* genetics. Complexes are highly connected through the sharing of components, revealing an unexpected level of functional organization.

Alfonso Valencia (National Centre for Biotechnology, CNB-CSIC, Madrid, Spain) focused mainly on systems for theoretical predictions of protein-protein interaction networks. The AbXstract system [http://www.pdg.cnb.uam.es/blaschke/cgi-bin/abx] starts from Medline abstracts and systematically extracts terms that may indicate interaction. On this basis the user can try and extract rules (frames) to identify the interactions. Valencia illustrated two different, sequence-based systems for prediction of protein-protein interactions that differ from the well-established 'gene neighbors' and 'Rosetta stone' approaches, namely 'mirror trees' and '*in silico* two hybrid'. The basis for all this work is careful multiple alignments between homologous sequences belonging to different organisms. In particular, the *in silico* two-hybrid system is based on the search for pairs of positions between two proteins that show a correlated mutational behavior (detected by multiple alignments) that could be due to those two positions interacting *in vivo*. The results of the predictions for *Escherichia coli* are available from the Predicted protein interactions database [http://www.pdg.cnb.uam.es/i2h].

Modeling and representing gene networks

Duncan Davidson (MRC Human Genetics Unit, Edinburgh, UK) gave a really holistic view of systems biology by

illustrating molecular networks in the context of the whole organism, with his Edinburgh Mouse Atlas project. The molecular functions of genes and their products can be coupled in physical-chemical or control networks. The realization of these networks is, of course, constrained by the fact that their components are compartmentalized in space and time, both at the subcellular level and between the cells in a multicellular organism. Ultimately, we need to understand or simulate the operation of a network across a field of cells or even in the context of a whole organism. The Edinburgh Mouse Atlas [http://genex.hgu.mrc.ac.uk/] provides a bioinformatic framework for recording the spatio-temporal compartmentalization of gene function at the organism level, for example in databases of gene expression in mouse development. The Mouse Atlas Query Interface to the Jackson Lab GXD database [http://genex.hgu.mrc.ac.uk/Resources/GXDQuery1] illustrates the dynamic waves of gene expression in the developing mouse embryo. The textual part of the Edinburgh Mouse Atlas (Standard Anatomical Nomenclature Database) that is used by the GXD consists of a controlled anatomical vocabulary of mouse development. This is linked to three-dimensional digital models of the developing mouse embryo. A gene-expression database that uses both text and spatial parts of the Atlas is also being built to collate *in situ* hybridization data and, in the future, data from spatially defined microarray samples.

Alvis Brazma (The European Bioinformatics Institute (EBI), Cambridge, UK) gave a talk on the building and analysis of a genome-scale expression network for yeast. Brazma introduced a simple graph-based notation for depicting interactions between genes. Distinction between in- and out-degree of genes, that is, the number of incoming and outgoing edges on a graph node, is important for identifying pathways converging on a gene or emerging from a gene. Brazma proposed a model of the *S. cerevisiae* global gene interactions as a scale-free network, high out-degree associated with transcription factors and high in-degree associated with metabolism components. This graph notation can easily represent many situations: the binding of a gene product G1 to the promoter of a gene G2; the fact that the disruption of gene G1 alters the expression level of gene G2; or that the two genes G1 and G2 are associated in literature.

Shoshana J. Wodak (Free University of Brussels, Belgium) presented the aMAZE database [http://www.amaze.ulb.ac.be] originally developed at the EMBL-EBI and now in Brussels. This database manages information on the molecular functions of genes and proteins, their interactions and the biochemical processes in which they participate. Wodak described the data model, which embodies general rules for associating molecules and interactions into large complex networks that can be analyzed using graph theory methods. The processes represented include metabolic pathways, protein-protein interactions, gene regulation, transport and signal transduction. These processes are mapped according to their subcellular localization. A distinct feature of aMAZE is its object-oriented, modular and open user interface. Wodak also illustrated how gene expression data can be interpreted in terms of the metabolic pathways in which some of the co-regulated genes are involved. In the near future aMAZE will also feature tools that will enable scientists to input their own data on pathways and processes.

Simulating cells

Cell function can be described in terms of a network of chemical reactions, and consequently a mathematical model of cell function consists of a set of ordered differential equations. Gene products don't simply float around in the cell: they are distributed in specific districts, according to the laws of localization and diffusion. Engineering stability in gene networks by auto-regulation and investigating regulatory feedback loops in model systems is the focus of Luis Serrano (EMBL, Heidelberg, Germany). Serrano proposes 'Smartcell', a general framework for whole-cell (prokaryotic system) modeling and simulation. It is an XML-based model, which considers reaction kinetics, uses a graphical representation of the cell and takes into consideration the cellular localization of the molecular components. Serrano also illustrated the possible preliminary validation of his model by using the chemotaxis signaling pathway in *E. coli*. Smartcell is employed to predict the location of receptors and the diffusion of response from signal site to signal integration. The cell model is able to reproduce the equilibria between the phosphorylated and unphosphorylated forms of the chemotaxis response regulator CheY. The diffusion dynamics should be remodeled, however, because the time scales of the simulation and the real phenomena are quite different.

The bioinformatics of intracellular networks may yet lead us to understanding the functioning of living cells. Hans V. Westerhoff (BioCentrum Amsterdam, The Netherlands) described integrative bioinformatics, which integrates all the necessary and available data from physical chemistry and biochemistry in order to calculate functions in living cells. Westerhoff showed how this approach led to a possible understanding of the elusive function for the glycosome, an extra organelle in *Trypanosoma brucei*, and therefore a potential drug target. It also led to the discovery of a regulatory function for a yeast protein that was thought to be engaged only in storage. For certain parts of living cells he showed that we already have a good set of data on the kinetics of metabolism (for example, for glycolysis in *S. cerevisiae*) and how this body of knowledge can be integrated in models he calls 'silicon cells'. On the Silicon Cell site [http://www.siliconcell.net], maintained by Jacky L. Snoep, it is possible to perform a number of fascinating *in silico* simulations of metabolic pathways and visualize the corresponding metabolic dynamics. The use of this facility led Westerhoff and coworkers to discover that baker's yeast responds in an adaptive and frequency-dependent way to

oscillatory sugar concentrations, a phenomenon with biotechnological implications.

Classical biochemical approaches to the study of metabolism can be integrated fruitfully in modeling approaches of pathways. For example, tumors have a common metabolic profile (high rates of glucose uptake and macromolecule synthesis) that may confer a common selective advantage. Marta Cascante (University of Barcelona, Spain) has a working hypothesis that the study of substrate flow changes during metabolic adaptation of cells to different phenotypes (metabolic profiling) together with the integration of data in computer models can give clues to identify differences between normal and tumor cells, which can be exploited in cancer therapy. Using this strategy, she illustrated how her group identified the ribose-5-phospate synthetic pathways as a new target in the treatment of cancer. Moreover, using an integrative bioinformatic tool (metabolic control analysis) she predicts that the best targets for inhibiting these pathways are transketolase and glucose-6-phosphate dehydrogenase, because these are the two enzymes with higher control of ribose-5-phosphate synthesis flux. Using specific inhibitors she demonstrated that inhibition of these enzymes results in a potent inhibition of tumor cell proliferation *in vitro* and *in vivo*.

Data mining as a foundation

Jan Komorowski (Norwegian University of Science and Technology, Trondheim, Norway) presented practical applications of mathematics and knowledge engineering to systems biology. The first application is PubGene [http://www.PubGene.org], a freely available service for automatic Medline abstract data mining, which is very useful for reconstructing (or inferring) links between proteins on the basis of common occurrences of gene names in abstracts. Komorowski also introduced the application of supervized learning techniques and 'rough sets' in molecular biology, with the task of classifying genes and gene functions. This approach handles complexity better than traditional engineering approaches such as control theory. Rough sets produce a classification (belongs to, does not belong to) with minimal added knowledge from the learning set. We can apply this technique to gene-expression data, for instance, when associating the expression of given genes in given tissues with processes. This procedure (EUGENES) consists of mining functional classes from an ontology, and extracting features for learning and classification. Information on the application of rough sets to the prediction of gene function from gene expression and ontologies is available from Jan Komorowski's homepage [http://www.idi.ntnu.no/~janko/].

Victor de Lorenzo (Centro Nacional de Biotecnologia, Madrid, Spain) works in the field of microbial bioremediation of soils, which seems to be a nightmare for the systems modeler. In fact, a microbial consortium or strain that degrades toluene in a given soil under certain physico-chemical conditions can be entirely different from a second consortium or bacterium that degrades the same substance a few meters away. In addition, data on biodegradation of toxic pollutants is dispersed in myriad publications that deal with the multifaceted genetics, biochemistry and ecology of a given compound and microorganism in an entirely non-systematic form. As a result we are still unable to predict the fate and effects of many toxic compounds, let alone to rationally design bacteria able to metabolize many chemicals in contaminated ecosystems. In the first instance de Lorenzo discussed the development of systems for automatic extraction of biological information from published scientific texts in biodegradation. The target in this case is to describe networks of biologically catalyzed reactions with sufficient accuracy. A second solution is the development of neural networks dedicated to prognosticating biodegradation and microbial ecology processes.

This workshop gave a clear view of how systems biology is gaining a primary role as a new and fruitful branch of high-throughput biological investigations. There is space both for computer-based and lab-based approaches and the merging of the two seems to be the most useful way of giving a global view of the intricacies of gene and protein networks.

Acknowledgements

Warm thanks to all the speakers who commented on my description of their contributions, with a special mention to Duncan Davidson for suggesting a suitable title.