Meeting report

Unraveling nature's networks

Lev Soinov and Misha Kapushesky

Address: EBI-EMBL, Microarray Informatics Group, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

Correspondence: Lev Soinov. E-mail: lev@ebi.ac.uk

Published: 16 September 2003

Genome ***Biology*** 2003, **4:**341

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/10/341

A report on the meeting 'Unravelling Nature's Networks: from Microarray and Proteomic Analysis to Systems Biology', Sheffield, UK, 21-22 July 2003.

Understanding complex regulatory biochemical networks in cells, tissues and on an organism-wide scale has become the holy grail of biologists and bioinformaticians alike. The 'Unravelling Nature's Networks: from Microarray and Proteomic Analysis to Systems Biology' meeting put together recent experimental and computational advances, providing a picture of how far we have come in this subject, and outlining ways forward.

Having always been one of the primary goals in biology, the problem of inferring relationships between entities in biochemical networks has been re-defined by the community each time a new experimental technique appears. The high-throughput methods of genomics and proteomics have increased interest in the problems of biological network reconstruction enormously. It has become possible not only to develop sophisticated, albeit abstract, models using classical modeling approaches, but also to apply various novel concepts. One of the most promising examples of such concepts is discovering recurring patterns in large sets of data, allowing effective and direct analysis of the results of modern biological experiments.

For a while, the exciting possibilities of large-scale studies led many researchers to think that general theories could be revealed simply by doing as many large-scale experiments as possible and looking at the results to find common features. That time has passed, and several practical as well as theoretical restrictions on high-throughput methodologies are becoming apparent. It is now well understood and widely accepted that only a combination of different approaches and different types of data and, more importantly, the concentrated efforts of experts with different scientific backgrounds on both 'large' and 'small' scales, can lead us to fruitful results. The predominant idea now is that a fusion of novel techniques with expert knowledge will help to avoid what can otherwise happen: the transformation of high-throughput concepts into scattered attempts to gather a lot of costly experimental information, often with only a vague purpose.

The opening talk was given by Steve Dower (University of Sheffield, UK), who studies gene expression in signaling pathways associated with chronic inflammatory diseases and host defense mechanisms. He showed that, although in general the considered signaling system is highly robust to changes in activator expression and responds in a linear fashion even to many-fold overexpression, it consists of 'noisy' stochastic components. Indeed, he noted that gene expression varies significantly at the single-cell level, suggesting that the pathway dynamics are governed at a higher level, possibly by the formation of large multiprotein complexes from signal transduction components.

The combination of experimental and *in silico* techniques on small and large scales is routine for some laboratories. Rick Livesey (The Wellcome Trust/Cancer Research UK Institute, Cambridge, UK) described his work on the modeling of transcriptional regulatory networks involved in mouse forebrain development. The goal is to develop a means of combining the results of spatial/temporal arrays of expression of candidate transcription factors with the computational analysis of putative regulatory regions of co-regulated genes. He made the point that it is possible to answer some 'simple' questions about the expression of the selected genes at single-cell resolution in a precise and reproducible way. A great deal of noise and natural stochasticity is intrinsic to biological systems, and, as Dower described, the same pathway may produce a wide variety of responses, which differ from cell to cell. The reproducibility of even 'simple' experimental observations of the expression of specified genes when comparing individual cells (for example, overexpressed versus underexpressed) is an exciting development, as this represents progress beyond experiments dealing with averages at the level of cell populations.

One of the main goals of bioinformatics is still the development of manageable and accessible banks of biological information. Terry Speed (The Walter & Eliza Hall Institute of Medical Research, Parkville, Australia) and Patrick Kemmeren (University of Utrecht, The Netherlands) presented two different approaches to this type of development. Speed's example was of a classical bioinformatic methodology focused on particular practical questions such as the experimental problems of how to identify multiple proteins in complex mixtures and the data mining and acquisition issues in the tandem mass spectrometry. He presented a scoring algorithm based on estimating factors that influence the gas-phase fragmentation of protonated peptides, using a recently developed 'relative proton mobility scale' and evaluated and compared it with several other existing algorithms. The proposed scale proved to be an effective basis for the automatic classification and statistical analysis of MS/MS spectra. Taking a 'systems biology' view, Kemmeren introduced a conceptually simple but promising approach to data management in which high-throughput biological information from different sources is integrated in order to derive common observations that would otherwise be hidden or even neglected. The approach involves storing data from different experiments for subsequent analysis and in a qualitative form, for example the presence or absence of a particular protein-protein interaction or a significant or nonsignificant change of gene expression. The underlying philosophy is that, as high-throughput methods contribute more and more data, a larger overall picture should become visible, and newly developed methods may help refine this image.

Mahesan Niranjan (University of Sheffield) moved away from the experimental towards computational methods. The often-encountered problem in large-scale data analysis is that the high dimensionality of the data makes statistical inference difficult and restricts the methods that can be applied. Niranjan reviewed several approaches for feature subset selection in the context of microarray expression data analysis; such selection is crucial when it is necessary to determine a subset of genes that helps to discriminate between different samples or experimental conditions. Using publicly available *Saccharomyces cerevisiae* gene-expression data, he showed that it is indeed possible to significantly reduce the dimensionality of microarray-related classification problems without considerable loss in discrimination ability.

One of the pitfalls in analyzing high-throughput data is that our intuitions about the nature of the networks often color both our approach and our interpretation of the results. The connectivity distributions of various biochemical networks (such as protein-protein interactions or metabolic networks) have been analyzed, and it has been suggested that they follow a power law, which is a distinctive mark of so-called scale-free networks. Scale-free networks have been proven to be one of the fundamental types of interconnected systems, both real and artificial. A considerable error rate is inherent in several of the high-throughput datasets, however; for instance, by some estimates around half of the protein-protein interactions revealed in different studies simply should not exist in real cells. This indicates that, in the worst case, our judgment about the presence of a particular interaction may be as good as tossing a coin, and it also gives rise to doubts about whether the 'scale-freeness' can be claimed to be a distinctive property of biochemical networks. Alun Thomas (University of Utah, Salt Lake City, USA) pointed out that it is possible to observe non-scale-free structures that have properties resembling those of the scale-free ones. Thomas presented a model for the structure of graphs of protein-protein interactions, derived from yeast two-hybrid experiments and supported by the data on human protein-protein interactions, that does not follow power law. The generated interaction graphs were also subsampled, taking in only a fraction of the connections for each node. It was shown that the vertex degree distribution of the subsampled graphs indeed looks approximately like a power law distribution, while not actually being one.

Several speakers at the meeting suggested that one of the ways to analyze high-throughput data is to take the data piecemeal, adopting comprehensible and interpretable approaches to studying relationships inside and/or between moderate-sized groups of genes or proteins, rather than performing system-wide searches. A parallel methodology coupled with the systems biology approach was introduced by Kwang-Hyun Cho (University of Manchester Institute of Science and Technology, UK). Cho uses systems of nonlinear ordinary differential equations to model networks of biochemical reactions in signal transduction pathways. Acknowledging that such an approach requires a large amount of clean and trustworthy data in order to estimate parameters of equations in use, he pointed out that even approximate and general features derived may help to develop more efficient experimental strategies.

Béla Novák (Technical University of Budapest, Hungary) uses the tools of dynamical systems theory to create a comprehensive differential-equation model of the *Schizosaccharomyces pombe* cell cycle. The model suggests that the cell-cycle engine can be broken down into modules responsible for transitions between the consecutive cycle stages. Novák's study has a distinct value, as it describes the transitions between all the major cell-cycle events in a simple way, using only a few key members of the cell-cycle control system.

The Mathematica® package Cellerator™ presented by Eric Mjolsness (University of California, Irvine, USA) allows modeling of biological networks at various levels of scope and complexity, ranging from signal-transduction pathways to cells and tissues, thus forming a hierarchy. The package supports the translation of chemical reactions, regulatory interactions and physical forces into differential equations and can run multicellular developmental models based on such equations. This is an important achievement in creating a complete modeling framework that can operate simultaneously at the molecular signaling level and at the level of multiple cells that communicate and affect each other via their physical movements and dislocations. The methods were illustrated by modeling the shoot apical meristem of *Arabidopsis thaliana*.

All in all, the meeting demonstrated that the field is developing quickly and the means of data acquisition and analysis are still being refined and extended. The task of reconstructing gene networks is, however, daunting. Jaroslav Stark (Imperial College, London, UK) posed an important question in his overview of the problem: what are the limits on network reconstruction with the presently available data and methods? The current strategy for network elucidation and analysis seems to be not to attempt to model the entire system at once, because the paucity of data makes the results impossible to interpret correctly, but to tackle network modules (although identifying these still remains an open question) with simple hypotheses that are verifiable by existing experimental techniques. For additional information on the meeting please refer to the Biochemical Society past meetings [http://www.biochemistry.org/meetings/pastmeet.cfm].