Meeting report

Systems biology: where it's at in 2005

Ben Lehner, Julia Tischler and Andrew G Fraser

Address: The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

Correspondence: Andrew G Fraser. E-mail: agf@sanger.ac.uk

Published: 1 August 2005

Genome ***Biology*** 2005, **6:**338 (doi:10.1186/gb-2005-6-8-338)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/8/338

A report on the joint Keystone Symposia on Systems and Biology and Proteomics and Bioinformatics, Keystone, USA, 8-13 April 2005.

Recent developments in high-throughput biology mean that we can now study the functions of hundreds or thousands of genes in parallel. Systems biology is the discipline that aims to make sense of the resulting deluge of data, in order to provide a comprehensive molecular description of biological processes. A recent joint Keystone meeting provided an opportunity for reflection on the current state of play and future directions for systems biology.

Mapping networks

Currently one of the largest subsets of systems biologists are the 'molecular cartographers' - researchers who are systematically mapping huge datasets of, for example, protein-protein or protein-DNA interactions. Although generating such networks *de novo* is extremely important, another vital aspect of network construction is the incorporation of data already available from the scientific literature. Mike Tyers (University of Toronto, Canada) described how a group of about ten people were able to extract about 30,000 protein-protein and 11,000 genetic interactions for the yeast *Saccharomyces cerevisiae* from the literature in a period of about ten weeks, and he strongly encouraged other communities to engage in similar activities. Analysis of the resulting dataset revealed some interesting differences between interaction maps derived from the literature and maps derived from high-throughput screens. For example, whereas high-throughput genetic-interaction and physical-interaction maps show only a minimal overlap, the two kinds of map derived from the literature share a much greater fraction of edges (interactions). In addition, essential proteins and highly connected proteins do not tend to interact with each other in high-throughput protein-interaction datasets, whereas they do in the literature-derived datasets. Although these conclusions may be partially explained by a bias in the interactions published in the literature, when combined with observations recently published by Michael Stumpf and colleagues showing that sampled subsets of networks often have very different properties to their parent networks, the conclusions show the importance of caution before inferring global properties of networks from our current incomplete datasets.

Genetic interactions identify functional connections between genes that often transcend physical interactions. Charlie Boone (University of Toronto, Canada) described how he and his collaborators are using hypomorphic or conditional alleles of genes in order to expand their systematic identification of genetic interactions in *S. cerevisiae* to include essential genes. Interestingly, essential genes seem to make many more genetic interactions than non-essential genes, but a smaller proportion of these interactions make intuitive mechanistic 'sense' to a biologist. A future challenge will be to provide a mechanistic explanation for the plethora of observed genetic interactions between seemingly functionally unrelated genes.

Edward Marcotte (University of Texas, Austin, USA) set out a rational approach for assessing the quality of high-throughput datasets as a key first step before combining them to provide a global view of the functional relationships between the genes of a eukaryotic cell. Clearly there is still a long way to go for network mappers - although their current high-quality yeast protein interaction map incorporates about 80% of yeast proteins, a similar map for humans contains less than one third of human proteins and is estimated to be under 10% complete. Moreover, over one quarter of the 'human protein interactions' derive solely from predictions from model organism datasets and lack experimental verification. Although we can expect a flood of metazoan protein-protein and genetic interaction data over the coming years,

we also need to encourage the development of new methods that target classes of proteins that are not well represented in the current maps. For example, Igor Stagljar (University of Zurich, Switzerland) described how a modified version of the yeast two-hybrid system can be used to identify protein interactions for transmembrane proteins, a class comprising many metazoa-specific and vertebrate-specific proteins.

Perturbing networks

A good starting point for the systematic understanding of a biological process is the comprehensive identification of genes that function in that process. One of the most powerful methods for genome-scale perturbation analysis is RNA interference (RNAi). David Sabatini (Whitehead Institute, Cambridge, USA) discussed his group's use of RNAi and *Drosophila* cell arrays, in combination with automated image analysis, to dissect the pathways regulating cellular growth on a genome-wide scale. For example, they were able to identify a previously mysterious kinase responsible for phosphorylating protein kinase B (Akt) using an immunofluorescence-based screen. He also described the progress of a Boston-based consortium aiming to create genome-wide collections of mouse and human RNAi libraries in lentiviral vectors. To date, approximately 35,000 short hairpin RNAs targeting 7,000 human genes and approximately 12,000 hairpins targeting 2,000 mouse genes have been constructed. Pilot screens were successful in identifying previously unknown mitotic regulators, and the field of cell-based RNAi screens seems certain to greatly expand in the future.

By far the most technologically developed organism for systematic perturbation analysis is *S. cerevisiae*. A complete collection of gene knockouts (deletion strains) has been available for several years and has been used in many reverse-genetic screens, as well as in the genetic interaction mapping project described by Boone. Marcotte described a new method for screening the collection of deletion strains, in which the yeast are printed at very high density onto a glass slide using a standard microarrayer. In a pilot screen they were able to use these 'cell chips' to identify half of the known and 36 novel regulators of the yeast mating response. The same group has also been using two-dimensional nuclear magnetic resonance (NMR) to quantify the 100-200 most abundant metabolites in different yeast deletion strains, so providing a molecular fingerprint of the state of a cell. Strikingly, removing a single gene often results in the cell switching to an entirely different metabolic regime. It was suggested that cells navigate a complex metabolic energy landscape, where basins of stability are found by adjusting enzyme concentrations, rates and metabolite levels. It seems very likely that applying systematic phenotypic measurements such as those made by Marcotte and colleagues on a genomic scale using gene deletion or RNAi libraries will greatly enlighten our understanding of many areas of biology.

Networks in space and time

Most currently known biological networks derived from high-throughput data provide a purely static view of a cell - they lack any spatial or temporal information. Three researchers - Wolfgang Baumeister (Max Planck Institute for Biochemistry, Martinsried, Germany), Peer Bork and Luis Serrano (both from the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany) - described an ambitious collaborative project that aims to bridge the gap between abstract molecular networks (as in Figure 1a) and the physical cellular architecture using a combination of computational modelling and cryo-electron tomography imaging (Figure 1b). In such an approach, the structures of protein complexes are first reconstructed *in silico* using the high-resolution structures of individual components (such as X-ray or NMR structures), together with protein interaction data (from high-throughput datasets) and lower-resolution structures of entire complexes or organelles (such as electron microscopy structures). These complex structures will then be fitted together into images of entire cells derived from cyro-electron tomography. In turn, these cellular models can then be combined with gene or protein expression data in order to model the dynamics of the cellular architecture.

Most currently known networks also lack any indication of the direction of information flow within the network and any description of cause and effect relationships between nodes. Dana Pe'er (Harvard Medical School, Boston, USA) described a project that aimed to reconstruct the flow of information through a cellular signaling cascade by simultaneously measuring the quantities of multiple phosphoproteins and phospholipids in primary human T cells under nine different perturbation conditions. The measurements were made simultaneously on single cells using multicolor flow cytometry, and the ordering of connections between pathway components was inferred using a Bayesian network framework. They were able to identify many of the previously known network causalities, and several novel inferred relationships were subsequently experimentally verified. An important feature of the approach is that Bayesian network inference yields the most concise models - components are not marked as being connected directly to each other if an indirect connection already exists that can explain the observed correlations.

Several other talks described approaches to mapping the cascades of phosphorylation events that occur within cells. One approach, described by several speakers including Matthias Mann (University of Southern Denmark) and Alejandro Wolf-Yadlin (Massachusetts Institute of Technology, Cambridge, USA), is to purify phosphopeptides with or

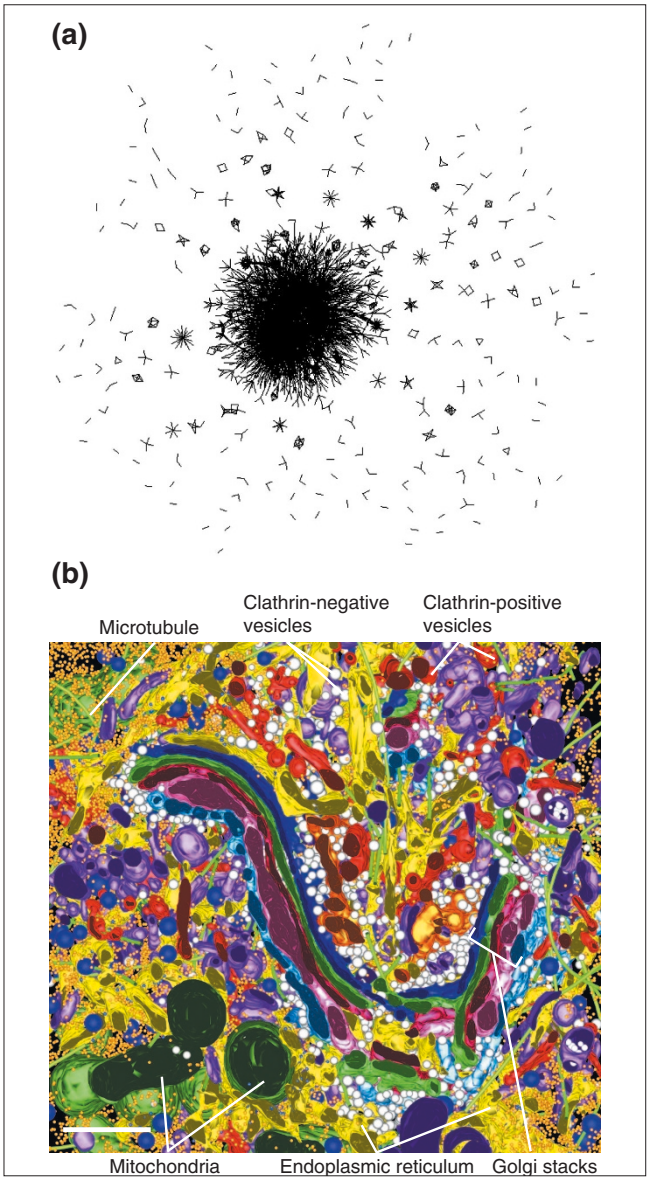

Figure 1
From networks to biology. **(a)** A network representation of a human protein-protein interaction map that we generated by integrating all of the available high-confidence protein interactions from model organism high-throughput protein interaction datasets [http://www.sanger.ac.uk/interactionmap] and visualised using the LGL tool [http://bioinformatics.icmb.utexas.edu/lgl/]. Most nodes (proteins) are connected in one large network (centre), but some are connected in small groups or pairs (outer areas). **(b)** A three-dimensional model of the Golgi region of a pancreatic cell line, as reconstructed by electron tomography. The seven cisternae that comprise the Golgi in the region are false-colored light blue, pink, cherry red, green, dark blue, gold and bright red, respectively. The endoplasmic reticulum is yellow, membrane-bound ribosomes are blue, free ribosomes are orange, microtubules are bright green, dense core vesicles are bright blue, clathrin-negative vesicles are white, clathrin-positive compartments and vesicles are bright red, clathrin-negative compartments and vesicles are purple, and mitochondria are dark green. The scale bar represents 500 nm. Reproduced with permission from Marsh BJ, *et al., Proc Natl Acad Sci USA* 2001, **98:**2399-2406.

without stimulation of a signaling pathway and then to use mass spectrometry to identify the individual phosphorylation sites. A second strategy, described by Philippe Bastiaens (EMBL, Heidelberg, Germany) is to express a library of fluorescently tagged cDNAs *in vivo* using live cell arrays (similar to those described by Sabatini), again with or without stimulation of a signaling pathway. Phosphorylation events are then detected as a fluorescence resonance energy transfer (FRET) signal indicating a very close apposition of the tagged protein and a tagged phosphotyrosine-specific antibody. Although both of these strategies are powerful because they measure phosphorylation events *in vivo*, neither of them is able to identify the exact kinase responsible for each phosphorylation event. This problem is being addressed by Mike Snyder (Yale University, New Haven, USA), who described how his group are using protein chips that represent the majority of the yeast proteome in order to identify all of the potential targets of a protein kinase *in vitro*. The combination of these *in vivo* and *in vitro* approaches should prove a powerful strategy for mapping phosphorylation and other information-processing cascades.

Beyond model organisms

One of the greatest potentials of systems biology may be to allow molecular biologists to move beyond the constraints of studying only a few rather arbitrarily chosen model organisms and out into the diversity of pathologically, agriculturally, or evolutionarily interesting species. To illustrate this point, Elizabeth Winzeler (The Scripps Institute, La Jolla, USA) explained how the application of DNA microarrays, proteomics, yeast two-hybrid analysis, and computational methods are beginning to catalyze research on the malaria parasite *Plasmodium falciparum*. For example, microarrays have been used to reveal evidence for widespread post-transcriptional regulation of gene expression and to identify about 25,000 single-feature polymorphisms amongst 13 worldwide *P. falciparum* isolates.

Bork reviewed recent results showing that it is possible to study the biology of organisms that cannot be cultured in a lab, or even those that have never been physically isolated. Massive shotgun sequence data from microbial communities found in an underground mine biofilm, surface seawater, farm soil, and a deep-seawater vertebrate skeleton were used by various groups to construct 'metagenomes' for these communities, comprising genomic sequences from many species. The proteins encoded in these metagenomes were then assigned to orthologous groups by comparison with known proteins. Strikingly, only half of the open reading frames (ORFs) of the soil microbes could be assigned to orthologous groups. Remarkably, it was also apparent from the sequence data that there are at least 3,000 different bacterial species in half a gram of soil. We look forward to viewing attempts at reconstructing the complete molecular network for this ecosystem at next year's meeting!

So where next for systems biology? Over the next few years, we expect to see the expansion and refinement of protein and genetic interaction maps, a greater concentration on the mapping and modeling of network dynamics, and improved efforts to integrate the network models of biological systems with the observed physical architecture of cells and organisms. Most of all, we anticipate that ever-improving computational analyses will reveal the new and unpredicted areas of biology lurking in the complex hearts of systematically compiled datasets. In short, we anticipate the unexplored and expect the unexpected - what more can one hope for?