Meeting report

The modular era of functional genomics

Eran Segal* and Stuart K Kim†

Addresses: *Computer Science Department, Stanford University, Stanford, CA 94305, USA. †Department of Developmental Biology and Genetics, Stanford University Medical School, Stanford, CA 94305, USA.

Correspondence: Stuart K Kim. E-mail: kim@pmgm2.stanford.edu

Published: 14 April 2003

Genome ***Biology*** 2003, **4**:317

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/5/317

A report on the Keystone Symposium 'Functional Genomics: Global Analysis of Complex Biological Systems', Santa Fe, USA, 20-24 February 2003.

Studying the full parts list of a car could provide a clue to how it works, but in order to get a fuller understanding, we would first need to know how these parts are assembled into the radiator, the water pump, the transmission, and so on. More importantly, we would also need to know how these higher-order functional units interact with one another to generate a fully functioning automobile. This conference clearly showed how the field of functional genomics is endeavoring to produce this kind of qualitative leap in our understanding of how cells and organisms work. The research described by the speakers goes beyond classical genetic approaches - focusing on studying single proteins in great detail - to incorporate large-scale functional assays measuring nearly all the genes of an organism and tracking them through space, time and diverse environmental conditions. The diversity of the high-throughput data presented was striking and included measurements of mRNA transcripts, protein-protein interactions, protein-DNA interactions, protein-lipid interactions, comparisons of sequence data from related species, and large-scale arrays of cells with different phenotypes. Three common methodological threads were apparent throughout many of the talks: the use of high-throughput data to detect underlying functional modules (groups of proteins that work together to execute a function, as defined by Harley McAdams, Stanford University Medical School, USA); integration of two or more types of genome-scale information; and comparisons between the genomes of multiple species in order to identify conserved sequences or expression profiles.

An exciting view of functional modules in the transcriptional regulatory networks of *Saccharomyces cerevisiae* was presented by Richard Young (Whitehead Institute for Biomedical Research, Cambridge, USA) and David Gifford (Massachusetts Institute of Technology, Cambridge, USA). Young has developed a high-throughput method to identify many *in vivo* target genes of most yeast transcription factors. These data provided insights into the yeast transcriptional network, suggesting the existence of several regulatory structures, including auto-regulation, feed-forward loops, and multi-component loops. Gifford has combined Young's promoter-binding data with gene-expression data to extract functional modules; in this context, a module is defined more specifically as a set of genes plus the set of transcription factors that control them. The key advantage of Gifford's algorithm is that it can use the expression data to confirm or refute whether genes are true target genes for each transcription factor and can add new genes to the modules. Gifford showed how the modules discovered can be automatically combined to accurately recover the temporal relationships between key regulatory events in the *S. cerevisiae* cell cycle.

The rationale behind comparative genomics is that evolutionary conservation of a feature implies that it has been retained by selection, which means it is likely to have a function. Mark Johnston (Washington University School of Medicine, St. Louis, USA) has used comparative genomics to identify potential regulatory regions in *S. cerevisiae*. His group has sequenced the genomes of five different *Saccharomyces* species, aligned the sequences upstream of orthologous genes, and thereby identified hundreds of sequences in the yeast genome that are conserved and thus potentially functional. They found that conserved sequence motifs are typically found between 125 and 250 base-pairs upstream of the translation-initiation codon. Johnston estimates that there are about 5,500 different conserved upstream motifs, and that 73% of these are made up of combinations of the known binding sites of 37 transcription factors.

A different approach for identifying functional non-coding sequences was presented by Michael Eisen (Lawrence Berkeley National Laboratory, Berkeley, USA). He relied on the assumption that, in higher eukaryotes, regulatory sequences are organized into relatively short modular units, each containing multiple binding sites for multiple transcription factors. He has used these characteristics to train an algorithm to recognize regulatory sequences, and was able to identify 28 new potential regulatory regions in the *Drosophila* genome. Some of the regulatory regions predicted using this method were confirmed using RNA *in situ* hybridization, and one was identified as the enhancer responsible for controlling posterior expression of the *giant* gene in the developing *Drosophila* embryo.

One of the most exciting aspects of functional genomics is the opportunity to use high-throughput data to track the activity of whole genomes temporally and spatially through complex biological processes. Matthew Scott (Stanford University Medical School) presented his work on the use of microarrays to track the expression of large numbers of genes through the life cycle of *Drosophila* - from fertilization, through the embryonic, larval and pupal periods, and into the first 30 days of adulthood. Scott found that some developmental stages that are morphologically very different from each other in fact have remarkably similar expression profiles; the largest changes in gene-expression profile occur during the more morphologically active stages of development, such as embryonic and pupal development. Scott also found that genes from the same functional group tend to be expressed at the same times in development - for example, most cell-cycle genes are expressed at the earlier time stages.

In an example of how complex expression patterns can be tracked in a prokaryote, Lucy Shapiro (Stanford University Medical School, USA) described the modular architecture that her group found in the transcriptional program of *Caulobacter crescentus* during the cell cycle, measured using microarrays. Shapiro was able to show that the CtrA response regulator, which controls several cell-cycle functions, is periodically activated by phosphorylation and cleared from stalked cells by temporally regulated proteolysis. Shapiro also showed that certain other proteins are spatially regulated within the cell so that they are at the right location when needed.

Julie Ahringer (Wellcome/Cancer Research UK Institute, Cambridge, UK) showed the utility of genome-wide RNA interference (RNAi) screens for identifying the function of previously uncharacterized genes in *Caenorhabditis elegans*. Her group constructed a library of 17,757 bacterial strains, each capable of expressing a double-stranded RNA designed to correspond to a single gene; 86% of all predicted *C. elegans* genes are covered by these strains. Animals fed these bacteria induce RNAi, resulting in knock-down of the targeted gene. Ahringer's group has used this library to identify novel genes for which RNAi results in sterility, longevity, embryonic lethality or larval lethality, and has also screened for particular phenotypes such as DNA-repair problems or early embryonic defects. Examination of RNAi phenotypes during early embryogenesis identified several new genes involved in cell polarity.

Finally, Mike Snyder (Yale University, New Haven, USA) showed how impressive proteomics can be. His group has developed several protein microarrays (chips) that can be used to assay protein-protein interactions, protein-lipid interactions, and interactions of proteins with small molecules. In an earlier version, the chip was designed with tiny wells ('nanowells'), each having one of the 119 protein kinases of yeast covalently attached inside it. Snyder used these chips to analyze *in vitro* the substrate specificity of all 119 kinases, using 17 different substrates; this provided clues to which kinases might phosphorylate these substrates *in vivo*. Snyder also presented a protein chip consisting of 5,800 yeast proteins (the products of almost all yeast genes), and showed the results of *in vitro* protein-protein and protein-lipid interaction assays. For example, 150 proteins were found to bind lipids - including, surprisingly, 17 kinases.

Judging from the broad range of topics covered at this meeting, it seems that a new field has emerged in which traditional genetics has been scaled up to produce a diverse, genome-wide view of living organisms. The challenge now is to bring together scientists from genetics, computer science, and statistics to assemble the cellular parts lists into functional units.