Meeting report

From genomes to systems

David I Ellis, Steve O'Hagan, Warwick B Dunn, Marie Brown and Seetharaman Vaidyanathan

Address: School of Chemistry, University of Manchester, Sackville Street, Manchester M60 1QD, UK.

Correspondence: David I Ellis. E-mail: D.Ellis@umist.ac.uk

Published: 28 October 2004

Genome ***Biology*** 2004, **5:**354

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/11/354

A report on the 2nd Conference of the Consortium for Post-Genome Science (CPGS) 'Genomes to Systems', Manchester, UK, 1-3 September 2004.

The second conference of the Consortium for Post-Genome Science aimed to portray the breadth of studies in this field, including genomics, transcriptomics, proteomics and metabolomics. The emphasis was on the transition from understanding at the level of each '-omics to a more integrated approach - that of systems biology. Systems biology aspires to be more comprehensive than previous approaches and to build predictive models that require powerful computation and the inclusion of meta data (information that describes the attributes of the data, such as experimental conditions). Six hundred delegates heard talks on diverse topics ranging from structural genomics through protein dynamics to pharmacogenetics and healthcare.

The plenary talk on networks by Albert-László Barabási (University of Notre Dame, Indiana, USA) was both thought-provoking and stimulating, dealing with fundamental principles and network function in both biological and non-biological systems. The building blocks of networks consist of nodes and connections, or links, and while certain networks can be modeled effectively by connecting their nodes with randomly placed links (according to the so-called random network theory), others do not fit this model. The highway system in the US could be considered a random network, for example, in which nodes correspond to cities and links to the connecting roads. In this network, each node has approximately the same number of links and is therefore 'typical' of the others; the network can therefore be 'scaled'. In contrast, the air-traffic system consists of a few important nodes that are well connected (so-called hubs) and many others that are sparsely connected. The nodes in this network are therefore 'atypical' and the network is 'scale-free'. Accidental failure of a number of nodes in a random network can fracture the system into dysfunctional units. But scale-free networks are more robust to such failure while being highly vulnerable to attacks at the hubs. Many real-world networks have this scale-free architecture, as they continuously expand by the addition of new nodes. The new nodes attach preferentially to already highly connected nodes and there is a power-law distribution of connectivity.

Barabási pointed out that biological networks such as protein-gene interactions in the genome, protein-protein interactions in the proteome and biochemical reactions within the metabolome, are scale-free. In a metabolic network, for example, the nodes are the chemicals (substrates) and the links the biochemical reactions, and metabolic networks have been shown to be scale-free in all three domains of life - archaea, bacteria and eukaryotes. In the proteome, where proteins (the nodes) are preferentially attached to other proteins, the origin of this scale-free topology is gene duplication, which has the consequence that a new protein tends to interact with the same proteins as the protein product of the original duplicated gene. Proteins with more interactions are therefore more likely to get a new link from other proteins. Barabási described how these complex networks are robust and maintain their basic functions even when subjected to errors and failures, such as cell mutations. He also discussed modularity in cellular networks, based on the hypothesis that biological functions are 'bolted' together into modules, an idea that is supported by metabolic data, for example in *Escherichia coli*, and suggests a hierarchical organization of metabolic networks. Barabási then went on to describe system-level experimental analysis of the entire spectrum of *E. coli* metabolism, including the global flux organization of the metabolic network, and

concluded with a description of life's complexity pyramid (see [http://www.nd.edu/~networks]).

Proteomics and mass spectroscopy

Two sessions on proteomics dealt with quantitative aspects, protein dynamics and post-translational modifications. Peter Roepstorff (University of Southern Denmark, Fredericia, Denmark) presented highlights from studies in his laboratory applying two-dimensional gel electrophoresis and matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) to the study of protein expression. In his experience, offline coupling of protein or peptide separation and MALDI-MS (and tandem MS, MS/MS) identification are currently more efficient for expression proteomics than online coupling of separation and MALDI-MS (for example, direct coupling of liquid chromatography and mass spectrometry, LC-MS). And tandem MS with MALDI-QToF (quadrupole time-of-flight) is more efficient than MALDI-ToFToF for detecting and analyzing protein modifications. Roepstorff presented studies on serum glycoproteins, in which 62 glycosylation sites on 37 glycoproteins were detected using MALDI-MS/MS coupled with lectin-affinity purification. Describing the application of proteomic approaches to the study of flagellar proteins in *Trypanosoma brucei*, Simon Gaskell (UMIST, Manchester, UK) showed how selective fragmentation of carboxyl-containing residues and subsequent analysis by MS/MS can be useful in protein identification.

Alan Marshall (Florida State University, Tallahassee, USA) highlighted the capabilities of Fourier-transformed ion cyclotron resonance (FT-ICR) MS, the fastest-growing MS sector (growing 25% per year in terms of output), in proteomic studies, with resolutions that allow a 1000 Da peptide to be distinguished from a peptide of 1000.0003 Da. He showed its utility in identifying biomarkers for Alzheimer's disease and renal disease, with MS/MS strategies based on electron capture dissociation and infrared multiphoton dissociation.

A detailed account of current proteomic concepts was presented in the plenary session by Ruedi Aebersold (Institute for Systems Biology, Seattle, USA), who described the application of quantitative isotope labeling strategies in protein-complex studies for understanding protein-interaction networks. He discussed the current limitations of protein-complex analysis and showed how for any protein, the probability of its being complex-specific can be calculated from its isotope-coded affinity tags (ICAT) ratio. Aebersold highlighted the current limitations of liquid chromatography MS/MS approaches and the complexities involved in analyzing multiple peptides for protein identifications, suggesting that perhaps we should look for "proteotypic peptides", with the aim of identifying and quantifying precisely one peptide per protein that uniquely identifies that protein. He discussed the construction of a nonredundant database of proteotypic peptides for the human proteome; this database currently covers 20-25% of human genes and consists of 2.5 million MS/MS spectra. He also highlighted the potential for the development of arrays of proteotypic peptides.

Most proteomic investigations currently concentrate on protein expression patterns. But this is now changing with the realization that there is more to proteomics than mere expression data. Rob Beynon (University of Liverpool, UK) highlighted the importance of studying the nuances of protein dynamics, with respect to the synthesis and degradation of proteins in the cell. He showed, using studies in yeast and chicken as examples, how stable-isotope-labeled amino acids and MS can be used to study protein dynamics.

The importance of employing mathematical tools in the biosciences, in particular in studying protein-protein interactions, was highlighted by Ian Humphrey-Smith (University of Utrecht, The Netherlands). He presented data illustrating the complexities of comprehending and analyzing proteomes with respect to protein interactions - there are an estimated 9.6 million protein-protein interactions in *E. coli* alone. He discussed how biomolecules are never specific for a single target and how a single peptide will never be enough to always identify a unique protein, immunologically speaking.

Mathias Uhlen (The Royal Institute of Technology, Stockholm, Sweden) discussed systematic exploration of the human proteome using a combination of high-throughput generation of affinity-purified antibodies, involving the cloning and expression of protein epitope signature tags (PrESTs), with protein profiling using tissue arrays. Analysis of the putative gene products of human chromosome 21 using this strategy was described. The results suggest that the strategy can be used in profiling proteomes and in subcellular localization studies on a genome-wide scale.

A particularly well attended session on metabolomics heard the announcement of a new journal, *Metabolomics*, to be published in January 2005 by the Springer Verlag (New York, USA). Douglas Kell (University of Manchester, UK) described a range of metabolomic investigations in different organisms, including the application of gas chromatography MS and machine learning in the search for new diagnostic biomarkers for human disease states. He concluded that metabolomics has great potential in both clinical and non-clinical applications and will play an important and integral part in systems biology.

Bioinformatics and modeling

A strong motivation in informatics research is the retrieval of buried information from both new and existing data, thus providing inductive insights leading to new knowledge. Use of existing, often analogous, data also helps avoid duplication

of experimental work, in that new experiments can be more focused. The other trend in informatics is the need to integrate diverse information to obtain a more holistic description of the biology under investigation.

Arun Holden (University of Leeds, UK) described an ambitious project aimed at modeling the probability of death from sudden cardiac arrest due to mutations in cardiac ion channels, in which modeling is based on a representation of cellular protein dynamics, a histologically detailed tissue model and a detailed anatomical model of the heart. This complex system was used to understand the mechanism of onset, persistence and termination of arrhythmias. Arrhythmias were shown to correspond to large-scale spiral waves propagating throughout the cardiac tissue, and the extent of meander of the spiral-wave focal point was found to be of prime importance in determining the duration of the arrhythmia. This model was successful in explaining the qualitative effects on the likelihood and duration of arrhythmias of different gene mutations corresponding to K^+ and Na^+ channel inhibition. The simulation/modeling approach used, which has become known as 'virtual tissue engineering', appears to be one of the first attempts to understand the effects of processes at the level of cellular biochemistry on macro-scale function through integrating knowledge of structural, anatomical, tissue and cell characteristics and behavior.

David Westhead (University of Leeds) presented his group's work on the design of the metabolic reconstruction software metaSHARK, together with case studies on the shikimate pathway in *Plasmodium* and on the metabolism of *Eimeria tenella*, the cause of coccidiosis in poultry. Metabolic reconstruction seeks to elucidate an organism's complement of enzymes and metabolic reactions from knowledge of the genome sequence, using information from protein databases and known metabolic mechanisms. The knowledge base need not be confined to the particular organism under study, as the software is designed to search genome data from related species for protein and enzyme analogs while taking into account sequence differences. The metaSHARK suite of programs includes SHARKdb, an object-orientated database of metabolic information stored in a form (Petri Net) that makes computation easy; SHARKbuilder, an automated system for metabolic reconstruction from unannotated genomic data, which uses amino-acid conservation information to search for genes with significant similarity to model enzyme sequences from other organisms; and SHARKview (see [http://bioinformatics.leeds.ac.uk/shark]), a web-based graphics tool enabling visualization of both database contents and metabolic reconstruction results.

Using annotated SWISS-PROT data, metaSHARK performed well compared to existing reconstruction software. In the shikimate pathway study in *Plasmodium*, only one of the seven pathways is annotated in the published *P. falciparium* genome. A metaSHARK search gave evidence for a bifunctional enzyme encoding both 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase and shikimate kinase activities, corresponding to the two precursor pathways of the already identified chorismate synthase. *Eimeria tenella* metabolism is currently poorly characterized: using raw *Eimeria* DNA data, metaSHARK gave better identification of potential metabolic function than did methods that are based on PRIAM, an existing method for automated enzyme detection from genome sequences, based on the classification of enzymes. Thus, metaSHARK should become an important tool for obtaining new knowledge about metabolic processes from unannotated eukaryotic genomes.

The final plenary lecture was given by Steve Oliver (University of Manchester) who discussed the integration of the different '-omics' technologies highlighted during the conference and their application to the study of *Saccharomyces cerevisiae*. He described the generation of commercially available single-gene knockout mutants, highlighting the replacement of single genes with a *kanMX* deletion cassette that has no genetic effect on the organism and is marked by two unique 20 bp 'barcodes'. In competition experiments using such mutants, heterozygotes mutant for each of the essential genes (1,500 mutants) were studied to determine the effect on fitness, as determined by growth rate, of each of these genes. Haploinsufficient genes were identified and were shown to exert a high level of control over the pathways in which they participate or regulate. Oliver described work following on from this, using continuous chemostat cultures to study the effect on different nutrient-limiting conditions (carbon, nitrogen, phosphate and sulfate) and dilution rates on the yeast metabolome, proteome and transcriptome. He described analysis of the endometabolome (intracellular metabolites) and exometabolome (metabolites secreted into the culture medium) using direct injection electrospray MS, analysis of the proteome by two-dimensional gel electrophoresis/MALDI-MS and analysis of the transcriptome by microarrays, and showed how the effects of different nutrient-limiting conditions and dilution rates could be separated for all these systems using principal components analysis. Application of a systems biology approach using all 'ome data can be assumed to be the next stage. The next conference of the CPGS is scheduled for Spring 2006 in Manchester; details are available online at [http://www.postgenomeconsortium.com/conference_2004/index.html].