

Meeting report

From microarrays to genome duplications

Christian Roth and Timothy Hughes

Address: Computational Biology Unit, Bergen Centre for Computational Science, University of Bergen, 5020 Bergen, Norway.

Correspondence: Christian Roth. E-mail: chregu@cbu.uib.no

Published: 23 July 2003

Genome Biology 2003, **4**:332

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2003/4/8/332>

© 2003 BioMed Central Ltd

A report on the fifth annual conference of the Society for Bioinformatics in the Nordic Countries (SOCBIN), 'Bioinformatics 2003', Helsinki, Finland, 22-24 May 2003.

The annual conference of SOCBIN, held for the first time in Finland, included talks on protein-structure prediction, gene regulation and genome evolution. Speakers described experiments and bioinformatic analyses using a wide variety of techniques, such as microarrays, large-scale sequence comparisons, machine learning and various proteomic techniques.

Medical and pharmacological genomics

Olli Kallioniemi (University of Turku, Finland) gave the keynote talk, in which he presented new approaches to overcoming the existing bottleneck in the process of identifying suitable research targets (such as drug targets), after candidate genes are identified using the fast-growing genomic and proteomic databases and before high-throughput screening techniques are used to find drugs themselves. Two useful techniques are comparative genomic hybridization (CGH) and nonsense-mediated RNA decay (NMD) microarrays, which can both be used to identify genes that have primary genetic alterations in the condition of interest. In CGH, genomic DNA from tumor and control tissue is hybridized to a microarray to identify deletions and insertions (which can cause changes in the copy number of genes). The relative accumulation of transcripts before and after inhibition of NMD, measured as a ratio in both cancer and normal cells, is used to find the genes most likely to have protein-truncating mutations. A third technique is cell-based microarrays: arrays in which cDNAs, small interfering RNAs (siRNAs), drugs or other reagents are printed onto microscope slides and cells are plated on top of the array. These functional cell-based microarray studies provide fundamentally different kinds of data from standard

techniques and establish important cause-and-effect relationships. Finally, sample-based microarray strategies, such as microarrays of tiny pieces of tumor tissue, facilitate the analysis of individual DNA, RNA or protein targets in thousands of samples in order to establish definitive clinical correlations for molecular targets at the population level.

In addition to aiding drug-target identification, the new high-throughput technologies pose novel challenges to practicing biologists. John Weinstein (National Cancer Institute, Bethesda, USA) identified the three most important of these as statistical analysis, biological interpretation of gene lists and integration of microarray data with other molecular and pharmacological information ('integromics'). As an example, he mentioned the tests of more than 100,000 chemical compounds on a set of 60 cancer cell lines available from the National Cancer Institute. This huge amount of data and the above-mentioned challenges motivated the development of a number of software tools in Weinstein's group, including MedMiner, a tool to organize the biomedical literature on genes and drugs, and GEEVS (Genomic Exploration and Visualization System), which integrates multiple types of molecular information. The whole set of tools is available at his website [<http://discover.nci.nih.gov>].

Addressing another key issue in medical research, Ian Humphery-Smith (University of Utrecht, Netherlands) explained the importance of the highly efficient mechanisms in humans that expose developing lymphocytes to the 'self' proteome in order to avoid immune reaction against the body's own tissues. He presented estimates of the number within the human proteome of sites made up of five amino acids within a target peptide, adjacent either in the sequence or in its structure, which could potentially be targets for an antibody. He showed that even at 100% identity, the potential for cross-reactivity between binding proteins was very high, which works in favor of the body if antibodies cross-react to 'non-self' antigens but has catastrophic consequences if they

bind to 'self' antigens. Exposing developing lymphocytes to all possible self antigens is difficult not only because of the high number of epitopes, but also because most proteins in living cells and body fluids occur at very low abundance and are consistently clouded by a small number of dominant species. One solution that the body has found is replication-induced protein synthesis (RIPS): a small amount of transcription occurs on the open DNA just behind the replication machinery, leading to the production of tiny amounts of all proteins contained in the genome. Large numbers of lymphocytes dying in the thymus and in germinal centers, and subsequent cross-priming to complement naturally-occurring endogenously-primed peptides carried by the major histocompatibility complex on antigen presenting cells, provides the necessary concentrations of physiologically relevant self elements during lymphocyte development; cells showing high avidity for self antigens die by apoptosis.

Protein-structure prediction

Several different approaches to protein-structure prediction were presented. Christine Orengo (University College London, UK) presented two protein-structure database resources: CATH and Gene3D. The CATH database [<http://www.biochem.ucl.ac.uk/bsm/cath/>] uses a novel hierarchical classification of protein domain structures to cluster proteins at four major levels: class, architecture, topology and homologous superfamily. Gene3D is a database of pre-calculated structural assignments for putative proteins encoded within sequenced genomes and was developed in order to perform comparative analyses of the distributions of the CATH protein families across the genomes of different species and kingdoms. Gene3D, which currently contains information on 66 genomes, has made it possible to establish that 90% of genes can now be assigned to known domain families, 60% of which are common to all three major kingdoms of life.

Moving from structure prediction by homology to *ab initio* prediction, Leszek Rychlewski (BioInfoBank Institute, Poznan, Poland) discussed 3D-Jury, a fully automated consensus protein-structure prediction system (meta-predictor) that is publicly available [<http://bioinfo.pl/Meta/>]. The 3D-Jury system is similar to other structure meta-prediction servers but it has the highest combined specificity and sensitivity and shows a high correlation between the reported confidence score and the accuracy of the model (as measured by the benchmarking tool LiveBench-6 [<http://bioinfo.pl/LiveBench/>]).

Machine learning can be applied to several structure-prediction problems, as described in an overview by Pierre Baldi (University of California, Irvine, USA). These include predictions of protein secondary structures, the relative solvent accessibility of residues, the contacts between residues, three-dimensional protein structures, and inter-chain β sheet

quaternary structures. Several tools relating to these problems are available at [<http://www.igb.uci.edu/tools.htm>].

Gene regulation

It is clear from recent work in many labs that regulation of genes by siRNAs is an important mechanism. When an siRNA binds to its target mRNA, the mRNA is degraded; this is called RNA interference (RNAi). Ron Unger (Bar-Ilan University, Ramat Gan, Israel) presented a method to systematically scan entire genomes for siRNAs and the genes they regulate. The method is based on suffix trees, a data structure that enables a very compact representation of strings and efficient searching for palindromes. Initial scanning suggests that a large number of genes, about 7% of *Caenorhabditis elegans* genes and 3% of *C. briggsae* genes, have the potential to be subject to natural RNAi control. Comparative studies were used to provide evidence that some of these are real cases of RNAi control; this refinement leaves about 70 genes that are candidates for verification by experimental work.

John Mattick (University of Queensland, Brisbane, Australia) gave an overview of the evidence to suggest that intronic and other non-coding RNAs comprise a second tier of gene-expression regulation in the higher eukaryotes, allowing the integration of complex suites of gene activity. Thus, although proteins are the fundamental components and effectors of cellular structure and function, the programming of eukaryotic complexity and phenotypic variation may be primarily embedded in an endogenous network of trans-acting RNAs that relay information on cellular states that is required for the coordination and modulation of gene expression.

Looking at gene regulation more generally, Daphne Koller (Stanford University, USA) described an automated probabilistic framework for discovering regulatory modules from gene-expression data. The method simultaneously searches for groups of genes that fit into modules (groups of coregulated genes) and for a regulation program for each module that explains the expression behavior of genes in the module as a function of the activity of some set of regulators. The method has been applied to a yeast gene-expression dataset, and the results have been validated using gene annotations, known regulatory relationships, and known and novel binding-site motifs, demonstrating the method's ability to identify highly coherent modules and their regulators.

Gene regulation can be affected by polymorphisms. Yitzhak Pilpel (Weizmann Institute of Science, Rehovot, Israel) presented novel data-mining techniques (assembled in the web application rMotif [<http://bioportal.weizmann.ac.il/~lapidotm/rMotif/html/>]) that make it possible to predict the effect of regulatory single-nucleotide polymorphisms (rSNPs), through the analysis of sequences, expression data and structural data. The main use of the application is in

prioritizing rSNPs according to their ability to appreciably affect gene-expression profiles.

Comparative genomics

The growing amount of available genomic data, especially complete genome sequences, has made comparative genomics an expanding field in bioinformatics. Kenneth Wolfe (Trinity College Dublin, Ireland) compared the complete genomes of the yeast *Saccharomyces cerevisiae* and the plant *Arabidopsis thaliana*. These genomes contain large regions where a series of genes on one chromosome has a series of paralogs on another chromosome, with conserved gene order. In both *S. cerevisiae* and *A. thaliana*, almost the entire genome can be paired in this way, which suggests that polyploidization (duplication of the whole genome), rather than duplication of smaller parts of chromosomes, has occurred. In *Arabidopsis*, polyploidization events are estimated to have occurred 24-40 and 92 million years ago. Only a few genes are retained as duplicates and the rest are deleted soon after polyploidization. Current investigation in Wolfe's group focuses on the factors determining which genes are retained in duplicate.

Evan Eichler (Case Western Reserve University, Cleveland, USA) compared duplicated segments in the human and rodent genomes. The duplicated regions in the human genome are highly non-random and show a high level of nucleotide identity (>95%) spanning large genomic distances (1-100 kb). Through processes of non-allelic homologous recombination, these regions provide a significant source of genetic disease in the human population. Analysis of other vertebrate genomes indicates that the high amount of segmental duplication might be a relatively unusual property of our genome: the mouse genome, for example, has fewer duplications, and these are mostly intrachromosomal. These duplications in the human genome have an impact on the assembly of draft genome sequences, as duplicated segments can be mistaken for versions of the same sequences.

Comparative genomics can also be used to study one of the traditional explanations for eukaryotic phenotypic complexity - alternative splicing. Christopher Lee (University of California, Los Angeles, USA) reported an analysis of 9,434 orthologous genes in human and mouse that indicates that, whereas most exons in the mouse and human genomes are strongly conserved in both genomes, exons that are included only in alternative splice forms (as opposed to the major transcript form) are mostly not conserved, and thus are the product of recent exon creation/loss events.

Diving into one of the most topical controversies, Dan Graur (Tel Aviv University, Israel) presented an analysis of the functional role of junk DNA in the human genome, focusing on Alu repetitive elements and nuclear sequences of mitochondrial origin ('numts'). Alu elements in introns can be

converted into new exons by changes in splicing. He revealed that there are only a few alternatively spliced exonized Alu elements in the human genome and that there is strong negative selection on constitutively spliced Alu elements favoring the loss of the element. The analysis of the numts showed that there have been only a few actual independent insertions of these sequences into the genome and many post-insertional duplications. In addition, all the numts seem to be devoid of function. He concluded that nonfunctional DNA cannot be easily recruited into new functions.

The sequences that do not match sequences in other genomes are also being studied. Marie Skovgaard (Technical University of Denmark, Lyngby, Denmark) has found that the genome of *Methanopyrus kandleri* AV19, like many other newly sequenced prokaryotic genomes, contains large numbers - up to 30% - of orphan protein-coding genes, that is, genes with no homology to known genes in any other species. Instead of being randomly scattered throughout the genome as in other species, however, a large fraction of these genes occurs within two large regions of the genome. Although they show no homology to known genes except within *M. kandleri*, neither their length nor their codon usage gives any reason to suspect that they are random (artificial) open reading frames. These regions could be considered as candidates for massive lateral transfer, but the bioinformatic analysis suggests that they have evolved by vertical descent.

Two speakers presented new tools related to sequence alignments. A hidden Markov model (HMM) is a computational structure for describing the subtle patterns that define families of homologous sequences; such models are powerful tools for detecting distant relatives. More specifically, HMMs are probabilistic models composed of interconnected states, in this case whether a site in two different sequences is a match or not, or whether there has been a deletion or an insertion. Transition probabilities are associated with the connections between the states, and residue emission probabilities are associated with the match and insert states. Markus Wistrand (Karolinska Institute, Stockholm, Sweden) showed how the method of estimation of transition priors (model initialization values that make sure that overfitting of models is avoided) as well as structural information and negative sequences can help fine tune HMMs by improving the estimation of transition probabilities. The ideas have been tested using the program HMMER and a benchmarking test that is based on detection of remotely related sequences hidden among unrelated sequences. Compared with the default version of HMMER, these improvements decreased the number of false predictions by roughly 10%.

Finally, Robert Giegerich (Bielefeld University, Germany) presented a generalization of an alignment of sequences in terms of tree structures and forests (collections of tree structures): an alignment is considered as a data structure, not an editing

process (a sequence of deletions, insertions and matches), and it is not to be confused with the matrix form that is used as a visualization of the alignment. This generalization is used in RNA-forester [http://bibiserv.techfak.uni-bielefeld.de/bibi/Tools_RNA_Studio.html], an alignment tool for RNA secondary structures that is suitable for global comparison of RNA secondary structures and for finding similar regions between them.

It is clear from this conference that bioinformatics, together with new experimental technologies, is giving a lot of insight into many biological systems.