Meeting report

The salmon genome (and other issues in bioinformatics)

Lena EF Milchert, David A Liberles and Arne Elofsson

Address: Department of Biochemistry and Biophysics and Stockholm Bioinformatics Center, Stockholm University, 10691 Stockholm, Sweden.

Correspondence: David A Liberles. E-mail: liberles@sbc.su.se

Published: 24 June 2002

Genome ***Biology*** 2002, **3(7)**:reports4022.1–4022.4

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/7/reports/4022

© BioMed Central Ltd (Print ISSN 1465-6906; Online ISSN 1465-6914)

A report on the fourth annual conference of the Society for Bioinformatics in the Nordic Countries (SOCBIN), Bioinformatics 2002, Bergen, Norway, 4-7 April 2002.

In the land of fjords and salmon, the Nordic bioinformatics conference included sessions on the mechanisms by which functional proteins are generated from gene sequences, genomes and metabolism, genetic networks, molecular evolution, and data-mining and machine learning in biology. A session of local interest focusing on the salmon genome project was also included. Two significant themes emerged from the talks: systems biology, particularly the elucidation of gene-interaction networks from microarray data, and the fate of genes and genetic networks through evolution.

Genetic networks

The keynote address from Simon Easteal (Australian National University, Canberra, Australia) was a good introduction to a conference with heavy emphasis on genetic networks. Easteal emphasized that bioinformatics should ultimately be driven by human health concerns, working to scale up from DNA to genes to proteins to networks to tissues to health, keeping evolutionary considerations and epigenetics in mind. To show how useful knowledge of human genetic networks will be, he presented the cases of actinin-3 and BRCA1. Actinin-3, a muscle-specific protein, is highly conserved between human and mouse, but a polymorphism creating a premature stop codon is present in the corresponding gene in 12-25% of non-African populations; it shows no apparent correlation either with human disease or, as might be expected, with elite athleticism. This may be explained either by postulating that the gene has duplicated in humans, so that the prematurely terminated protein is compensated for by the duplicated version, or by the effects of interacting genes, which may compensate for the prematurely terminated protein in other ways. BRCA1 is a protein involved in DNA repair and cell-cycle regulation that affects the risk of breast and ovarian cancer in women. The *BRCA1* gene has been shown to be under positive selection (as shown by the facts that a high ratio of nonsynonymous to synonymous substitutions (K_a/K_s) is seen both between humans and chimpanzees and between them and their last common ancestor with gorillas, and that the gene is found to be in linkage disequilibrium with its neighbors). Comparison of the reconstructed sequence of the *BRCA1* gene in the common ancestor of chimpanzees and humans with the known mutations in cancer indicates that the evolutionarily derived states of the gene (states that have arisen recently) may correlate with high disease risk. This surprising phenomenon may be accounted for when we have a better understanding of the system-level network of interacting genes that includes *BRCA1*.

Peter Uetz (University of Karlsruhe, Germany) presented an analysis of genetic networks that was based on large-scale two-hybrid screens, which detect protein-protein interactions, and mass spectrometry; these two approaches generated partially overlapping results. From the networks that were discovered, some sub-networks, such as that of cell-cycle control, had more connections than others, such as those of membrane fusion or cytokinesis. Uetz pointed out that as there are only 1,900 structures in the protein database (PDB [http://www.rcsb.org/pdb/]) and only 26 measured dissociation constants currently available with which to analyze 10,432 known interactions, we have a long way to go.

Trey Ideker (Whitehead Institute, Cambridge, USA) has analyzed gene-expression profiles of yeast strains that have deletions of individual galactose-metabolism (*GAL*) genes

and were grown in the presence or absence of galactose. The model generated from the gene-expression profiles gave a 70-80% agreement with predictions made from the classical model of how the *GAL* genes interact, although the strains in which *GAL7* (encoding galactose-1-phosphate uridylyltransferase) and *GAL10* (encoding UDP galactose-4-epimerase) had been deleted gave some unexpected results; these were used to propose a new feedback system from these two genes. Ideker put forward a way of building an integrated molecular-interaction network using automatic methods to compare gene-expression levels with metabolic pathways derived from protein-protein, protein-DNA, and small-molecule interactions. He also proposed novel techniques to score how well a proposed network reproduced the experimental data and new methods using simulated annealing to find the best network.

Nir Friedman (Hebrew University, Jerusalem, Israel) works on a Bayesian approach to building networks from gene-expression data. He presented a novel scoring function and a heuristic search method to find protein interactions in network space, which correctly identified six well structured sub-networks from yeast data, including one for mating, and performed significantly better than traditional clustering methods. The interactions that his method missed were those involving genes such as the main transcription factor involved in mating, which show very little variation in expression across all samples.

Harmen Bussemaker (Columbia University, New York, USA) used a motif-finding approach to analyze microarray data and to predict conserved upstream regulatory elements. Using a simple linear model that ignores much of the complexity of transcription, he was able to identify 30% of the systematic variance in mRNA abundance (the changes in gene expression correlated with the presence of certain upstream elements). Alvis Brazma (European Bioinformatics Institute, Hinxton, UK) also examined this problem and found 60 sequence groups in coexpressed promoters, of which 40 are already known to be transcription-factor-binding sites. He emphasized the need to progress towards more quantitative dynamic networks rather than looking only at scale-free networks of interactions (networks that define the connections but not their behaviors. He also advocated the use of an international standard for sharing microarray data between groups.

The last talk to focus on genetic networks came from Zoltan Szallasi (Children's Hospital, Boston, USA), which served as a statistical warning shot. Emphasizing the small number of measured data points compared with the complexity of the problem (when the size of the network and the number of gene-expression states are very large), he stressed the importance of careful statistical design and interpretation of results. Practically, data points without at least a twofold difference in expression level should not be considered significant.

How functional proteins are generated from genomes

Genetic networks represent the end step of a long process, and there are many variables and variations in the process of generating functional proteins from a genome. Laura Landweber (Princeton University, USA) presented the complex story of ciliates, which generate a macronucleus (from which all transcription takes place) by splicing a fraction of the DNA from the micronucleus. Genes destined for the macronucleus are amplified 1,000 times and spliced from segments on both strands, in both orientations and in different loci, in a process that depends on proper DNA bending and folding. Ciliates also use four alternative genetic codes, which can be tied to genetic-code-specific residues in the translational release factor eRF1. The selective pressures behind such variation and complexity are unknown, and only a tip of the ciliate biology iceberg may be known at this point.

Moving from DNA to RNA, Peter Arctander (University of Copenhagen, Denmark) profiled the vast diversity of sequences that may be generated by alternative mRNA splicing. For example, the human *slo* gene produces over 500 potential gene products from at least eight differential splice sites, resulting in variation in ion sensitivity in the encoded ion-regulated channel. In *Drosophila*, the *Dscan* gene has over 38,000 potential gene products, and the *para* gene has 13 alternative exons giving 1,536 mRNAs in addition to 11 RNA-editing sites, which together provides the potential for over a million different gene products. Overall, there are estimated to be 500 different post-transcriptional and post-translational modification mechanisms in the human cell, in addition to variation generated epigenetically. It is unclear how stable splicing patterns are in evolution, but it is known that 32-50% of human genes are alternatively spliced and that 15% of human genetic diseases are caused by mis-splicing.

Francine Perler (New England Biolabs, Beverly, USA) moved us from mRNA splicing to protein splicing, in which exteins (protein sequences analogous to exons) create a protein by excising an intein (analogous to an intron and frequently consisting of a homing endonuclease), in a very rapid reaction. She has generated a database, InBase, containing 130 inteins from 56 organisms across Archaea, Eubacteria, and single-celled eukaryotes, and has identified several trends in the sequences at protein splice junctions from these data, although polymorphism is known. This has allowed her to develop an intein-specific search program that can be used for identifying inteins in newly sequenced genes and genomes. From an evolutionary perspective, protein splicing may have evolved as an early kind of recombination. Subsequently, coevolution of specific intein-extein pairs appears to have occurred in many cases, perhaps dictated by protein-folding requirements for the reaction.

Zoran Obradovic (Temple University, Philadelphia, USA) discussed the importance of disorder in protein structures

and showed that, like structure, disorder can be predicted from sequence alone. The observed disorder falls into three categories that he calls 'flavors': V, sequences that form ordered helices upon binding of a ligand; C, polysaccharide- or oligosaccharide-binding domains; and S, leucine-rich regions. Long stretches of disorder appear to be common in the proteins in the SwissProt database [http://www.expasy.ch/sprot/] and in those encoded by completed genome sequences, and disorder is the dominant sequence component of many oncogene products. Flavor V predominates in Archaea, whereas flavor S predominates in Eubacteria and in Eukaryota, which have the most disorder.

Stepping from disorder in structures to disorder in genome functional annotation, Steven Brenner (University of California, Berkeley, USA) compared three independent annotations of *Mycoplasma genitalium*, one performed automatically using GeneQuiz and two performed manually, by Eugene Koonin (National Center for Biotechnology Information, Bethesda, USA) and by workers at The Institute for Genomics Research (TIGR, Rockville, USA). He showed several examples where these assignments disagreed, both linguistically and functionally. From the disagreements between these annotations, he assumed that the minimum error rate was 8% and is probably as high as 20%. These errors are caused by poor sequence comparison, incorrect inferences of function from homology, and propagation of erroneous data. He proposed that structural genomics is the solution, through analysis of structural homology, the ligands revealed in bound structures, and conservation of surfaces (for example in DNA-binding proteins). One remaining problem, differentiating between analogy and homology, still lacks an effective computational approach.

Søren Brunak (Danish Technical University, Lyngeby, Denmark) presented a machine-learning approach, the ProtFun method, to get from sequence to feature prediction to a neural network integrating the features, resulting in a prediction of biological (not biochemical) function. Although the correlations found were not good enough for making predictions about unknown proteins in genomes, the method does allow tracking of the way that features such as tyrosine phosphorylation change through metabolism or the cell cycle.

Sarah Teichmann (Medical Research Council Laboratory of Molecular Biology, Cambridge, UK) has analyzed the evolution of domains in yeast and *Escherichia coli*. Initially focusing on small-molecule metabolism in *E. coli*, she clustered 722 domains from 510 enzymes into 213 families. A tendency was observed for families to be spread across different pathways, showing that the pathways have evolved through duplication of genes from these families. Conservation of general enzyme chemistry was commonly found in duplicated enzymes; conservation of the binding site of a cofactor or minor substrate was less common; and conservation of the main binding substrate very rare. This argues against a previously proposed model suggesting that duplicated enzymes go on to act at neighboring steps within the same pathway. Comparing yeast and *E. coli*, 54-65% of genes have homologs in the other species; two thirds of shared enzymes have identical domain structures and only a sixth show non-orthologous displacement, in which a structurally unrelated protein takes over a function. Teichmann observed 40 gene duplications and 20 gene fusions in this comparison.

Michael Lynch (Indiana University, Bloomington, USA) presented a systematic analysis of the process of gene duplication. From an analysis of gene number, he has shown that three complete genome duplications are the minimum needed to reach metazoan genome sizes from the last common ancestor of yeast. Vertebrates require at least two additional complete genome duplications, which may also have enabled the increased speciation in vertebrate lineages. Lynch described the process of subfunctionalization, in which a gene with various functions duplicates and the duplicated genes each take on a subset of the original functions. He said that duplication of small regions of the order of a kilobase has been the dominant process in genome expansion, rather than whole genome duplication (polyploidization). Subfunctionalization appears to 'buy time' after gene duplication, so that the two genes are both retained and can later take on new functions ('neofunctionalization'); it also allows independent optimization of the different functions of the original gene. The most recent duplicates in genomes are complete copies, but over time these diverge into chimeric copies with differential loss of sequences and move to different regions of the genome. Although the rapid process of birth and death of genes (by duplication and deleterious mutation, respectively) may be complex, there is clearly an increase in the strength of purifying selective pressure with time after each duplication event in eukaryotic genomes. The overall average is a birth rate of 0.019 genes per million years and a half-life of duplicated genes of five million years. Lynch went on to conclude that this rapid birth and death process may be related to speciation: gene duplication and movement cause microchromosomal rearrangements, ultimately resulting in a reproductive isolating barrier.

Doron Lancet (Weizmann Institute, Rehovot, Israel) continued the discussion of gene duplication, focusing on the olfactory receptor (OR) family of seven-transmembrane helix proteins. Phylogenetic analysis shows that their genes began on chromosome 11, spreading to chromosomes 1 and 3 and then ultimately throughout the genome, with the exceptions of chromosomes 20 and Y. Whereas 60% of human OR genes are pseudogenes, only 20% of the approximately 1,300 mouse and dog OR genes are pseudogenes. Mouse and dog also have species-specific expansions in OR gene number. In the human population, 35 segregating pseudogenes (which are functional in some individuals but not in others) have been identified. At a more general level, residues that are

conserved in orthologs (direct homologs) but not in paralogs (which are related by duplication) have been identified in an attempt to better understand the odorant-binding region, although little can be said about specificity so far.

Denis Shields (Royal College of Surgeons, Dublin, Ireland) has computationally reconstructed the ancestral sequences of genes encoding the mitogen-activating protein (MAP) kinase family in order to determine the residues that are responsible for differential specificity of members of the family; some predicted residues were verified experimentally to determine specificity. In a larger systematic study of proteins in Pfam [http://www.sanger.ac.uk/Software/Pfam/], he generally found evidence for a period of increased change at constrained sites a short to moderate evolutionary time after duplication events, larger than the changes found after speciation events. Conversely, vertebrate proteins involved in host defense, the brain, and metabolism or synthesis showed more change after speciation than after duplication; they may be subject to positive selective pressure during speciation events.

Richard Goldstein (University of Michigan, Ann Arbor, USA) took a systematic look at the process of substitution during evolution. He found that because positions in a protein sequence do not all evolve under the same constraints during evolution, assuming a simple distribution of rates across sites (such as a gamma distribution) is problematic because the rates are based upon a single substitution matrix that may not be applicable to all sites. Matrices describe the amino-acid changes that are more or less likely, based on structural or functional constraints; matrices built explicitly to deal with structural constraints ignore functional constraints. A more general approach is to build different matrices for different classes of sites with similar evolutionary constraints. Building and applying these matrices to G-protein-coupled receptors had success in predicting transmembrane regions and in detecting the divergence of binding sites. This approach has the potential to improve current methods for a number of applications in molecular-evolutionary and functional-genomic studies.

Phylogeny and evolution

Manolo Gouy (University Claude Bernard, Lyon, France) presented a phylogenetic tree for bacteria by applying principal component analysis to 310 gene trees derived from the HOBACGEN-CG database of proteins in sequenced genomes [http://pbil.univ-lyon1.fr/databases/hobacgen.html]. Informational genes, such as those involved in gene regulation, were found to be more reliable for building this tree than operational genes, such as enzymes and structural proteins; this may reflect a core of informational genes with common ancestry in bacteria. From this phylogeny, spirochetes and chlamydiales were found to be the earliest emerging clades, rather than hyperthermophiles as proposed earlier. Gram-positive bacteria were found not to be monophyletic, contrary to previous studies.

Paul Sharp (University of Nottingham, UK) analyzed the origins of the fast-evolving, highly recombinogenic human immunodeficiency viruses (HIV). From his phylogenetic analysis, HIV-2 appears to be derived from simian immunodeficiency virus (SIV) from sooty mangabeys in West Africa via multiple cross-species transmissions. HIV-1 group M (the most common type) appears to be most closely related to West African chimpanzee SIV. Using a molecular clock based on gamma distances, he estimated that the latter transmission appears to date from around 1931; the urbanization of Africa after this date may be responsible for the spread of HIV. The V3 loop of the surface envelope glycoprotein of HIV-1 appears to have been under positive selective pressure during the emergence of the M and O groups as they diverged from SIV sequences. The general sequence divergence within subtypes, which is much greater than that found in influenza hemagglutinin, is an ominous prospect for vaccine development.

Salmon and Norway

No meeting in Norway could be complete without a local fishy flavor. Bjørn Høyheim (Norwegian School of Veterinary Science, Oslo, Norway) presented an overview of the salmon genome project. In light of all of the discussion of gene duplication, the salmon genome project is daunting, as a genome duplication event has occurred within the past 100 million years, most of which remains, and there is instability in the number of chromosomes. Ongoing mapping studies include 200-500 genotyped markers in the Atlantic salmon *(Salmo salmar*), the rainbow trout *(Oncorhynchus mykiss*), and the brown trout *(Salmo trutta*). Combined with an effort in Canada, 55,000 expressed sequence tags have been generated and have allowed preliminary studies of gene expression. Finn Drabløs (SINTEF Group, Trondheim, Norway) presented a strategy for using structural modeling to identify unknown genes from the salmon genome project, based upon work in other species. He claimed that the most efficient method to detect distantly related proteins was to use PSI-BLAST to search sequence databases iteratively in combination with intermediate sequence searches (linking two homologs through a third sequence).

With the emerging knowledge of gene expression and evolution at many levels and with the recent progress towards an Atlantic salmon genome project, we will all sit down to salmon dinners in the future with a greater degree of appreciation.