

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. Details of talks on yeast genetic networks and methods for building gene networks and detecting regulatory elements from microarray data are available with the complete version of this article, online.

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.

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 exons (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.

Details of talks on the errors in genome annotations, methods to predict protein function from sequence, clustering of domains of *Escherichia coli* enzymes, how duplicated genes (including odorant receptor and mitogen-activated kinase genes) change after duplication, and how individual amino-acid positions evolve are available with the complete version of this article, online.

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.



.reports

The complete version of this article, available online at <http://genomebiology.com/2002/3/7/reports/4022>, includes further details of talks on genetic networks and their reconstruction from microarray data, functional annotation of genome sequences and the evolution of enzyme domains, of genes after duplication, and of individual amino acids.