Meeting report

Genome informatics: taming the avalanche of genomic data

Erik LL Sonnhammer

Address: Center for Genomics and Bioinformatics, Karolinska Institutet, 171 77 Stockholm, Sweden. E-mail: Erik.Sonnhammer@cgb.ki.se

Published: 17 December 2004

Genome ***Biology*** 2004, **6:**301

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/6/1/301

A report on the fourth Cold Spring Harbor Laboratory/Wellcome Trust Conference on Genome Informatics, Hinxton, UK, 22-26 September 2004.

The pace of genomic data generation has recently accelerated substantially, particularly for higher eukaryotes. The genome sequences of more than a dozen multicellular eukaryotes are now finished, and complementary information such as expressed sequence tags (ESTs), and data on gene expression, protein-protein interactions, single-nucleotide polymorphisms (SNPs) and proteomics are being generated in high-throughput mode. To make sense of this flood of heterogeneous data and make it useful to scientists, powerful informatics systems are being put in place. The Cold Spring Harbor Laboratory/Wellcome Trust conference on genome informatics focused both on informatics technologies for data storage and processing as well as on the latest discoveries gained by analyzing genomic data.

The many blessings of multiple genome sequences

Once the human genome sequence was underway, scientists realized that the only way to decode it fully was to sequence a number of related genomes as well. The mouse and chimpanzee genomes have now been completed, as well as those of some other animals, and this has helped genome annotation tremendously. In particular, functional but non-protein-coding parts of the human genome could be identified by looking for sequence conservation between different genomes. For example, the presence of 'ultraconserved' regions was completely unexpected (thousands of regions of more than 100 base-pairs (bp) are 100% conserved between different mammals), and their function is still largely a mystery. There is also evidence that a large proportion of the non-protein-coding regions of the genome is transcribed, suggesting that much of the genomic sequence previously thought of as junk may well be functional. With the aim of identifying all functional elements in the human genome sequence, the National Human Genome Research Institute in the USA has launched a public research consortium named ENCODE, the ENCyclopedia Of DNA Elements [http://www.genome.gov/10005107]. Many projects presented at the conference come from members of this consortium.

Rotem Sorek (Compugen and Tel Aviv University, Israel) presented evidence that some non-coding sequences conserved between human and mouse contain signals that govern alternative splicing. About 80% of the alternatively spliced exons he analyzed are flanked on both sides by regions of around 100 bp (on average) that are more conserved than other intronic sequences. Finding such conserved regions around splice sites thus indicates the presence of alternative splicing. Sorek claimed that more than 1% of all functional sequences in the human genome are involved in the regulation of alternative splicing.

It has long been known that *cis*-regulatory elements are often conserved between mammalian genomes and can thus be detected by conservation analysis, even though such elements tend to be short. Remo Sanges (Telethon Institute for Genetics and Medicine (TIGEM), Naples, Italy) described a whole-genome analysis looking for sequences upstream and downstream of protein-coding genes that are conserved in human, mouse and rat. Using the Ensembl Compara database [http://www.ensembl.org/Multi/martview], he found 72,000 conserved regions across these mammals. A problem is, of course, to determine which of these are functional *cis*-regulatory elements. An approach for enriching for functional elements was presented by Steven Jones (Michael Smith Genome Sciences Centre and British Columbia Cancer Research Centre, Vancouver, Canada), who described a pipeline for using coexpression calibrated by Gene Ontology annotations to select co-regulated genes for defining regulatory motifs in the cisRED database [http://www.cisred.org/].

MicroRNAs (miRNAs) are a class of RNA genes that can now be detected thanks largely to the availability of a number of eukaryotic genome sequences; miRNAs are regulatory RNAs of 20-22 nucleotides that are cleaved from stem-loop folding RNA precursors of around 70 nucleotides, and are usually conserved between vertebrates. The binding between an miRNA and its mRNA target is characterized by the presence of a perfectly complementary heptamer at the 5′ end of the miRNA (the seed). Anders Krogh (Copenhagen University, Denmark) presented a computational miRNA-identification pipeline based on their known characteristics. When applied to the *Arabidopsis thaliana* genome, it predicts about 1,000 novel miRNAs. Although some of these are probably false, the method had high specificity on a test set of 127 known miRNAs.

One of the biggest challenges in annotating the human genome has been to correctly predict the exon-intron structure of the protein-coding genes. With multiple species this task becomes significantly more accurate. Using the mouse genome, David Carter (Wellcome Trust Sanger Institute, Hinxton, UK) reported that the accuracy increases from around 60% to above 85% on his dataset. Using four species (human, mouse, rat and chicken) improves this to about 90%.

Another use of multiple genome sequences was presented by Paramvir Dehal (DOE Joint Genome Institute, Walnut Creek, USA), who described a new method for clustering orthologous proteins such that gene clusters are consistent with the species tree. This information is collected in the Phylogenetically Inferred Groups database [http://PhIGs.jgi-psf.org]. Dehal reported the use of clusters deriving from individual ancestral genes at the root of the vertebrates to test the controversial ‘2R’ hypothesis - that two genome duplications occurred at the base of vertebrate evolution. As most duplicated genes have been subsequently lost, at random, it has so far been difficult to find strong evidence for the 2R hypothesis, particularly since many genes have been independently duplicated during later stages of vertebrate evolution.

Using the Phylogenetically Inferred Groups database, genes existing at the root of the vertebrates were extracted. These were mapped onto human chromosomes, after which the homologs among them were connected to each other (such homologs are ‘ancient paralogs’ or ‘outparalogs’ when comparing to other vertebrates). These connections tend to form blocks of consistently placed connections on the chromosomes, suggesting a genome duplication. Two rounds of duplications would give a four-way circular arrangement. Support for the 2R hypothesis was given by the observation that 72% of the chromosomes were covered by blocks of four-way circles, while in a situation where the gene positions were randomly chosen, less than 10% of the chromosomes were covered.

More and more databases

Model organism databases (MODs), a hot topic at the conference, exist in many different forms, providing their communities with various capabilities. But most researchers have very similar needs, only for different organisms. There has therefore been convergence towards a generic genome-information system, aptly named the Generic Model Organism Database (GMOD). This is an open-source project, currently used by about 100 institutes worldwide. GMOD [http://www.gmod.org] was described by Lincoln Stein (Cold Spring Harbor Laboratory, USA). It features database management tools as well as user-friendly web visualization tools for a wide range of genomic data. Much of the annotation in a MOD consists of links to other databases, and it is vital to keep these up to date. Stein pointed out that an important next step will be to set up a robust and well-maintained system for linking the MODs to each other via orthologous genes in the different organisms.

As expected, a huge number of databases and websites were presented at the meeting. The emergence of more and more databases is becoming a problem for the end user, who would prefer to have all the relevant information available at one site. To this end, the Distributed Annotation System (DAS) [http://biodas.org] has been created, with the aim of allowing integration of different biological databases across the world. The DAS system is gaining popularity and is now integrated in the Ensembl database [http://www.ensembl.org], said Steve Searle (Wellcome Trust Sanger Institute). Other decentralized bioinformatics systems include the Canadian Chinook [http://smweb.bcgsc.bc.ca/chinook/] system described by Stephen Montgomery (Michael Smith Genome Sciences Centre and British Columbia Cancer Agency), and the industry consortium GeneGrid (http://www.qub.ac.uk/escience/projects/genegrid) described by P.V. Jithesh (Queen’s University, Belfast, UK). These networks are primarily aimed at distributing algorithms and analysis techniques. As the concepts and technologies of decentralized information systems are fairly new to bioinformatics and genomics, it is natural that a number of different solutions are at present being developed in parallel. Probably only a few of them will survive, but those that do will be of great value to the entire community.

Although genomics and bioinformatics have already generated vast amounts of new discoveries, one was left with a feeling that both fields are still in an early phase, and that greater discoveries are ahead. After all, only a few eukaryotic genomes have been sequenced, and sequencing new genomes representing ‘missing’ clades will be invaluable for unraveling evolutionary and functional relationships. A growing trend was evident: a shift of focus from the protein-coding part of genomes to the functional non-protein-coding part. The avalanche of genomic data will no doubt continue to grow and lead to more discoveries of how the genome functions. But to reach this goal, much work will be needed to improve databases and the computational analyses to exploit all the sequence and functional information.