Meeting report

Open-source software accelerates bioinformatics

John Quackenbush

Address: The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. E-mail: johnq@tigr.org

Published: 1 September 2003

Genome ***Biology*** 2003, **4:**336

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/9/336

A report on the Wellcome Trust/Cold Spring Harbor Genome Informatics meeting, Cold Spring Harbor, USA, 7-11 May 2003.

The wide availability of genome sequence data has created a wealth of opportunities, most notably in the realm of functional genomics and proteomics. This quiet revolution in the biological sciences has been enabled by our ability to collect, manage, analyze, and integrate large quantities of data. In the process, bioinformatics has itself developed from something considered to be little more than information management and the creation of sequence-search tools into a vibrant field encompassing both highly sophisticated database development and active pure and applied research programs in areas far beyond the search for sequence homology. The nearly 250 participants at this meeting represented not only hard-core computational scientists but also laboratory biologists who are increasingly moving from being users of software to developing it themselves.

Databases and ontologies

One of the most tangible products of genome projects is the vast body of data that has been generated, and this was reflected in the two sessions on the databases that are, with increasing sophistication, providing the scientific public with access to the data. The challenge is not collecting the data but identifying and annotating features in genomic sequence and presenting them in an intuitive fashion. The general-purpose sequence databases provide uniform access to the data and a consistent annotation for an increasing number of organisms - examples include the EMBL database, GenBank and the DNA database of Japan (DDBJ) and genome databases such as Ensembl and the National Center for Biotechnology Information (NCBI) Genome Views - but species-specific databases, such as the *Saccharomyces* Genome Database [http://www.yeastgenome.org] and the Mouse Genome Database [http://informatics.jax.org], provide much richer and more complex information about individual genes. Other resources, such as the University of California Santa Cruz Genome Browser, have democratized genome annotation by allowing specialists from around the world to present their own view of genomic features. Reflecting the maturing of these utilities, however, the database sessions instead focused on other issues.

Increasingly, we are coming to realize that protein-coding genes are not the only important transcribed sequences in the genome. Sam Griffith-Jones (Sanger Institute, Hinxton, UK) described the development of Rfam [http://www.sanger.ac.uk/Software/Rfam], a database of non-coding RNA families developed in collaboration with Sean Eddy's group at Washington University, St. Louis, USA. Rfam provides users with covariance models - which flexibly describe the secondary structure and primary sequence consensus of an RNA sequence family - as well as multiple sequence alignments representing known non-coding RNAs and provides utilities for searching sequences for their presence, including entire genomes.

Several other useful databases were also presented. David Torrents (European Molecular Biology Laboratory, Heidelberg, Germany) described work underway to identify pseudogenes on a whole-genome scale. Laurens Wilming (Sanger Institute) provided a brief summary of the strange gene structures that his group have identified in their curation of annotated vertebrate genomes as part of the Human and Vertebrate Analysis and Annotation project (HAVANA; [http://www.sanger.ac.uk/HGP/havana/]). Sohrab Shah (University of British Columbia, Vancouver, Canada) described the open-source PeGASys system that he and his colleagues have developed for the rapid annotation and curation of genomes. Finally, Alexi Sharov (National Institute on Aging, National Institutes of Health, Bethesda, USA) described efforts to sequence expressed sequence tags (ESTs) and measure expression levels using microarrays, with a focus on understanding the genes that are expressed in embryonic stem cells during differentiation.

Another major topic discussed was the need for standardizing the language for describing genes and their functions through the use of ontologies. Ontologies provide hierarchical, controlled vocabularies for describing biological entities. The most highly developed at present is the Gene Ontology [http://www.geneontology.org] system, which provides functional assignments for genes and their products within three categories: molecular function, biological process, and cellular localization. Within each category, the assignments follow a hierarchy with increasing functional specificity as the level of assignment increases. Christopher Mungall (Berkeley Drosophila Genome Project, Berkeley, USA) described the Slot'n'GO system, which allows the rapid creation and assignment of new functional classes, avoiding confusing compound terms. For example, the term 'actin binding' combines terms from separate ontologies for physical process ('binding') and protein ('actin') ontologies. These 'cross-product' terms between ontologies can generate a large number of complex interrelationships that can make ontologies unwieldy. The Slot'n'GO approach would supplement functional classes with attributes that can be used to classify them. S, an annotator would classify a gene product as 'protein binding' and then 'fill in a slot' for the term 'binds' with the term 'actin' from the protein ontology. This would represent a transition from a 'phrase-based ontology' to a 'property-based ontology.'

The success of ontologies in facilitating biological inquiries was reflected in presentations on other ongoing efforts. Pankaj Jaiswal (Cornell University, Ithaca, USA) described plant and phenotype ontologies being developed by the Plant Ontology Consortium [http://www.plantontology.org]. Winston Hide (South Africa National Bioinformatics Institute, Belville, South Africa) outlined ongoing work to develop a controlled vocabulary for gene-expression data called eVOC, which provides terms for describing the anatomical system, cell type, pathology, and developmental stage necessary to understand and interpret expression data. The eVOC system is being developed in collaboration with EnsMart [http://www.ensembl.org/EnsMart], which Arek Kasprzyk (European Bioinformatics Institute, Hinxton, UK) described; EnsMart extends the Ensembl database to include expression data.

Functional genomics

The availability of genomic resources in an increasing number of species is reflected in the growing prevalence of functional genomics and proteomics; it is difficult to open an issue of almost any journal without seeing one or more papers that use these approaches to investigate biological phenomena. The increasing sophistication of these studies is reflected in the software systems that have been developed to deal with the growing body of data. A number of talks focused on methods that cells use to regulate gene expression. Steven Brenner (University of California, Berkeley, USA) described analysis of experiments underway to uncover the role played by alternative splice forms of genes. His analysis indicates that many of these are targeted for nonsense-mediated mRNA decay (NMD), which is an RNA quality surveillance system. Others are subjected to regulated unproductive splicing and translation (RUST), a general mechanism for controlling protein expression that has been established for a number of genes. Brenner and his group have developed computational methods for identifying candidates for NMD and RUST and are working to validate their predictions experimentally.

Fatemeh Haghighi (Columbia University, New York, USA) is developing novel experimental and computational methods to map the methylation of the human genome. Their analysis indicates a striking pattern of methylation, with clustering of CpG-rich sequences in kilobase-sized unmethylated regions and with Alu elements at the boundaries of these regions. As one might expect, highly methylated regions contain large numbers of transposons, whereas unmethylated areas of the genome contain few transposons other than those severely degraded by mutation. Haghighi's group is developing methods to predict with high confidence whether a particular gene is likely to be methylated. This is an important goal, because methylation of genes is increasingly implicated in regulation of genes involved in a range of human diseases, including cancer.

In proteomics, the identification of protein mass tags (cleaved peptides from all proteins of interest) and the association of these with known genes are important for understanding patterns of protein expression. Brian Halligan (Medical College of Wisconsin, Milwaukee, USA) described an algorithm that uses the amino-acid composition of a peptide rather than its amino-acid sequence to identify its parent protein. As a first test of this approach, Halligan and his collaborators have created a database of tryptic digests of the proteins encoded in the *Saccharomyces cerevisiae* genome in which the frequencies of the various amino acids are used to construct a weight vector. *K*-means clustering of the weight vectors organizes the data into classes and provides features that can rapidly be used to identify new peptides. In a test of this system, only one of 11,735 peptides was incorrectly identified in a search of the indexed database.

Comparative genomics

With genome sequence and expression data from a growing number of species available in well-curated databases, comparative genomics is rapidly maturing as a field. Jack Chen (Cold Spring Harbor Laboratory, Cold Spring Harbor, USA) is using the completed genome sequences of *Caenorhabditis elegans* and *Caenorhabditis briggsae* to examine the olfactory genes identified in these related species. Although *C. elegans* has nearly 700 olfactory genes and *C. briggsae* has 500, only about 330 are clear orthologs. The many additional olfactory genes in *C. elegans* appear to fall into only two of the six olfactory subfamilies listed in the Pfam database. What the overrepresentation of these classes means in terms of the biology of *C. elegans* remains to be determined, but clearly the representation of these families has an effect on how these species adapt to their environments and accommodate their abilities to feed, mate, and communicate effectively.

Saurabh Sinha (Rockefeller University, New York, USA) described the development of software to identify *cis*-regulatory modules in metazoan genomes. Using a combination of hidden Markov models and an expectation maximization algorithm, his group's method uses phylogenetic comparisons between homologous sequences from multiple species and positional correlations between binding sites to improve discrimination of regulatory motifs. The identification of such regulatory modules is crucial for a thorough understanding of gene regulation and development, providing the link between genotype and phenotype.

Emphasizing the value of easy-to-use graphical presentations of synteny and ortholog data, Inna Dubchak (Lawrence Berkeley National Laboratory, Berkeley, USA) described work by her group to create a system for the multiple alignment of whole genomes. Using the MLAGAN multiple alignment algorithm as the engine of their system, Dubchak and colleagues have engineered the Berkeley Genome Pipeline to handle large numbers of eukaryotic genomes. The processed data are cached and presented to users through Vista, an intuitive visualization system they have developed [http://www-gsd.lbl.gov/vista].

Finally, although the data, databases, and tools for analysis have evolved significantly in recent years, work still continues on the development of new and better algorithms for the analysis of genomic data. Robert Klein (Washington University, St. Louis, USA) presented RSEARCH, a program designed by Klein and Sean Eddy to identify homologs of single, structured RNA sequences, and demonstrated its utility by finding previously unknown homologs of RNase P in several eukaryotic genomes. Richard LeBlanc (Genome Quebec and McGill University, Montreal, Canada) presented a novel algorithm for mapping gene expression data from microarrays into a low-dimensional discriminant space, simplifying the results of an experiment and allowing the development of robust classification algorithms for assigning hybridization results to various biological classes. And even the 'solved' problems of genome assembly and annotation saw new developments at the meeting, with Zemin Ning (Sanger Institute) presenting an improved genome assembler that is being used to assemble the highly polymorphic zebrafish genome, Masahirl Kasahara (University of Tokyo, Japan) presenting the RAMEN genome assembler for assembling a variety of vertebrate genomes, and Chaochun Wei (Washington University, St. Louis, USA) presenting an approach to using ESTs and genomic sequence to facilitate the sequencing of full-length cDNAs and ultimately to annotating gene structures.

The rise of open-source software

While the scientific presentations at the meeting outlined a rapidly changing and evolving landscape, what was even more interesting was the sociological changes that were evident in the bioinformatics community. In the early days of the genome project, many groups sought a particular advantage over their competitors by carefully guarding their software source code. It was evident in this meeting that this approach is rapidly fading away and being replaced by a commitment among developers to create open-source software tools that can be used, adapted, and improved by the wider scientific community.

Two obvious questions that arise are why anyone would want to release their software code and why others would want to add new utilities and functionality to someone else's software. Aside from the obvious benefits of creating a community resource that can rapidly advance the field, I see a number of advantages to an open-source approach to software development in a scientific environment. These include the fact that it gives full access to the algorithms and their implementation, which allows users to understand what they are doing when they run a particular analysis; it provides the ability to fix bugs and extend and improve the supplied software; it encourages good scientific computing and statistical practice by providing appropriate tools, instruction, and documentation; it provides a workbench of tools, allowing researchers to explore and expand the methods used to analyze biological data; it ensures that the international scientific community is the owner of the software tools needed to carry out research; it encourages support and further development of the tools that are successful; and it promotes reproducible research by providing open and accessible tools with which to carry out that research.

The creation of open-source software is not unique to the scientific community. The best-known example is probably the development of the Linux operating system. For Linux, a world-wide community of developers has allowed the creation of an operating system that now commands a significant portion of the market, particularly for high-end systems. It is apparent that, increasingly, members of the bioinformatics community are trying to create an environment that encourages scientists to develop new applications and to create them in a framework that makes sophisticated tools available and accessible to laboratory biologists. By doing so, my hope is that the same sort of community-based spirit will drive the development of increasingly sophisticated software and so advance the general state of the art in genomics and bioinformatics.

Nowhere is this more apparent than in the creation of software systems for genome annotation and display. There are a growing number of extremely well developed toolkits for genome annotation, many of which were described at the meeting, including the Ensembl project [http://www.ensembl.org] led by Ewan Birney (European Bioinformatics Institute), the Generic Model Organism Database development project (GMOD [http://www.gmod.org]) led by Lincoln Stein (Cold Spring Harbor Laboratory), the Manatee system [http://www.tigr.org/software] engineered by Owen White and his group at The Institute for Genome Research (TIGR, Rockville, USA), and the PeGASys system described above. The open-source fever has spread into the world of expression analysis, with at least three highly refined systems, such as the BioArray Software Environment (BASE [http://base.thep.lu.se]) from the University of Lund, the TM4 system developed at TIGR [http://www.tigr.org/software/tm4], and the BioConductor collection of tools developed in the R statistical language by the BioConductor consortium [http://www.bioconductor.org]. Increasingly, too, we are seeing community-based efforts aimed at developing standards for data reporting, ranging from the Gene Ontology project to the Microarray Gene Expression Data society's efforts to establish standards for expression data [http://www.mged.org] and the Human Proteome Organization's work to establish similar standards for proteomics [http://www.hupo.org].

Ultimately, it will be interesting to see how these efforts will pay off, but already at the 2003 Genome Informatics meeting the changes in the culture of bioinformatics were evident. Clearly, open-source development is becoming an increasing presence in the bioinformatics community and it will be interesting to see its effects at future meetings.