Meeting report

An international showcase of bioinformatics research

Todd Vision

Address: Department of Biology, University of North Carolina, Chapel Hill, NC 27599, USA. E-mail: tjv@bio.unc.edu

Published: 28 August 2003

Genome ***Biology*** 2003, **4:**337

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/9/337

A report on the 11th International Conference on Intelligent Systems for Molecular Biology, Brisbane, Queensland, Australia, 29 June - 3 July 2003.

The blossoming of bioinformatics around the world was clearly in evidence at the 11th annual meeting of the International Society for Computational Biology, the first 'ISMB' to be convened outside Europe and North America. Given that there are now well over 100 published whole-genome sequences from cellular organisms, it was no surprise to see many new developments in comparative genomics at the conference. Michael Brudno (Stanford University, USA) discussed methodological advances for aligning very long genomic sequences (for example, the length of mammalian chromosomes) that differ as a result of not only substitutions and indels but also rearrangement events, such as inversions and translocations. This approach was dubbed 'glocal' alignment as it provides a high-level global alignment of smaller local alignments. Jens Lagergren (The Royal Institute of Technology, Stockholm, Sweden) presented a method for probabilistically distinguishing between orthologs and paralogs in gene trees, a problem that had previously been addressed only by optimizing under the unrealistic assumption of maximum parsimony.

The newly completed draft mouse genome sequence has provided a key point of comparison for those studying the human genome. Jim Kent (University of California, Santa Cruz, USA) reported on the frequency of local rearrangements between the human and mouse genomes, estimating that there are approximately two inversions and five local duplications per aligned megabase of human DNA. He also reported that there are apparent hotspots of rearrangement. Perhaps not surprisingly, one such hotspot is the immunoglobulin locus. While Kent addressed the small scale, the clustering of larger-scale chromosome rearrangement breakpoints between human and mouse has recently been reported in the literature. David Sankoff (University of Ottawa, Canada) took issue with the published results by showing that the methods that have been used lead to a conclusion of clustered rearrangement breakpoints even for simulated data in which chromosome breakage is random. David Haussler (University of California Santa Cruz, USA), in his keynote address, discussed the 22% of the human genome that is covered by retrotransposons conserved between human and mouse. The rate of interspecific sequence divergence in these elements varies along the chromosomes. Assuming that retrotransposons reflect the neutral rate of interspecific sequence divergence, Haussler estimated that 5% of the human genome may be under purifying selection. Interestingly, this is approximately two-fold higher than the proportion of the genome that is thought to be coding sequence. The surprising amount of conserved, and therefore presumably functional, non-protein-coding DNA may have a number of explanations, including conserved *cis*-regulatory elements and novel classes of transcribed RNA molecules.

In another keynote address, John Mattick (University of Queensland, Brisbane, Australia) made a forceful argument that non-coding RNA plays a greater role than commonly appreciated in allowing organisms to navigate the combinatorial complexity of development. Both he and Yoshihide Hayashizaki (RIKEN Genomic Sciences Center, Yokohama, Japan) pointed to an analysis of the RIKEN full-length mouse cDNA collection in which it has been estimated that 47% of transcripts do not contain any substantial open reading frame. As evidence for their functionality, many of these non-coding transcripts show evidence of differential expression, and almost a third of them are spliced. In addition to the importance of non-coding RNAs for natural processes within the cell, their use in RNA interference (RNAi) is now also an important functional genomics tool, as reflected by the number of posters on the computational design of small interfering RNA probes.

More long-standing problems in transcriptional regulation also saw many new advances reported at the conference. A flurry of presenters addressed the statistical analysis of mRNA abundance data and the use of these data in the identification of *cis*-regulatory elements, operons, splicing variants, and other genomic features. Among them was Jung Kyoon Choi (Advanced Institute of Science and Technology, Daejeon, Korea), who suggested the normalization and combination of effect sizes for expression changes across experiments as an approach to the difficult, and increasingly important, problem of expression data meta-analysis.

One of the themes to emerge from the conference was the challenge of integrating multiple sources of experimental evidence. One type of evidence to receive a great deal of attention was that from high-throughput protein-interaction studies - despite a protracted open discussion at a preconference satellite meeting about the myriad problems with this type of evidence. Eran Segal (Stanford University, USA) presented a machine-learning approach for inferring sets of proteins belonging to the same pathway from combined analysis of gene expression and protein-interaction data. The rationale is that proteins in the same pathway are more likely than random proteins both to be expressed under the same conditions and to interact physically with one another. Simultaneous consideration of both datasets leads to improved predictions relative to either one alone. An example of a novel prediction from Segal's approach is a potentially new member of the cytoplasmic exosome complex involved in the 3′ processing of pre-rRNAs in yeast. Similarly integrated views of experimental data are likely to be *de rigueur* in the future, even for 'traditional' areas of bioinformatics such as gene prediction and protein-structure prediction.

Another major theme of the meeting was the increasing reliance on the controlled vocabulary of functional assignments known as the Gene Ontology (GO). Segal, along with many other presenters, used the assignment of GO terms to yeast proteins in the *Saccharomyces* Genome Database [http://www.yeastgenome.org/] as a source of data against which to validate his method. Other presenters used GO assignments directly for data mining. In my view, it is somewhat ironic that the manual curation of functional terms should play such a large role in a field that is typically more inclined to high-throughput, automated, and machine-learning techniques. Although GO is clearly of immense value, it is important that researchers remain vigilant against the uncritical acceptance of GO functional assignments: controlled vocabularies cannot be perfect, the experimental evidence on which assignments are based is not infallible, and there is a real risk of error propagation and circularity in some applications.

Expert knowledge, such as that used to make GO assignments, also plays an important role in the success of literature-mining strategies. That was the conclusion of Alexander Yeh (Mitre Corporation, Bedford, USA), who reported on the results of the Knowledge Discovery and Data Mining Challenge Cup, an exercise modeled on the popular CASP (critical assessment of techniques for protein structure prediction) competition among protein-structure predictors. The goal in this case was to use computational text-mining techniques to flag scientific articles that contain information that should be included in FlyBase [http://flybase.bio.indiana.edu/]. Competing teams had a training set of articles to learn from, and their success was evaluated relative to a test set manually categorized by FlyBase. Despite the organizers' attempts to limit the role of technical biological knowledge in the competition, the most successful teams did employ biological experts to identify feature lists in the training data manually. Another major lesson was that successful teams took advantage of the linguistic structure of the articles (for example, distinguishing between figure captions and literature citations), rather than the classic approach of treating the text as a 'bag of words'. Such objectively evaluated competitions among teams of researchers using different methods have the potential to greatly accelerate progress on difficult computational problems.

ISMB also plays host to a number of affiliated Special Interest Groups, including ones on open-source software, text mining, biological pathways, ontologies, and education. At some of these satellite meetings, practitioners could be seen leaving talks in twos and threes, laptops in tow, to hack away at each other's code. An additional feature of ISMB is the set of half-day tutorial sessions that are held in conjunction with the conference. This year, there were over a dozen different topics covered, including molecular evolution, statistical analysis of microarray data, and homology modeling of protein structures. The disciplinary breadth of ISMB continues to be remarkable in this age of specialization. The role that computation plays in diverse biological fields is visibly increasing, as is the sophistication with which computational techniques are being employed to help generate and test experimental predictions. One can expect the same combination of breadth and depth at next year's ISMB in Glasgow, Scotland, to be held in conjunction with the European Conference on Computational Biology. Proceedings of this year's meeting are published in supplement 1 to volume 19 of *Bioinformatics*.