

Meeting report

Bioinformatics 2000

David W Ussery

Address: Center for Biological Sequence Analysis, Institute of Biotechnology, The Technical University of Denmark, DK-2800 Lyngby, Denmark. E-mail: dave@cbs.dtu.dk

Published: 1 September 2000

Genome **Biology** 2000, 1(3):reports4014.1-4014.2

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2000/1/3/reports/4014>

© Genome**Biology**.com (Print ISSN 1465-6906; Online ISSN 1465-6914)

A report from the Bioinformatics 2000 conference, held in Elsinore, Denmark, 27-30 April, 2000.

Protein folding

Despite the explosion in sequence information, the problem of modeling and prediction of protein folding still remains unsolved. As Jeff Augen (Life Sciences Division of IBM) pointed out, there are two problems manifested by the sequence information explosion: the first is the continuing need to solve computationally intensive biochemical problems such as tertiary protein structure prediction. IBM [<http://www.ibm.com/news/1999/12/06.phtml>] is attempting to build the largest and fastest computer ever (capable of more than one petaflop; that is, more than one quadrillion - 10^{15} - operations, per second), and have plans to use it to try and solve the computationally intensive 'protein folding problem', by modeling the folding of a small protein. They have not yet decided which particular protein they will choose for this analysis.

The second problem resulting from the sequence explosion is the need to manage, store, search, study, and compare enormous amounts of genetic and proteomic data. Clare O'Donovan (EMBL Outstation, Cambridge, UK) described the human proteomics initiative of the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. This is a major project that will provide information about the structure, function and subcellular location of every known protein encoded in the human genome. More information about this, as well as the current status of the project, can be found at the Human Proteomics initiative web page: <http://www.ebi.ac.uk/swissprot/hpi/hpi.html>. O'Donovan also talked about attempts to deal with the explosion of sequence information. At the moment, to find all the available information about a given protein, a large number of databases must be searched. There is a real need for some sort

of centralized, curated (and reliable) source of information about all human proteins identified by the human genome.

On the subject of *ab initio* protein structure prediction, David Baker (University of Washington, USA) described the development of a fast computational method to approximate protein folding. To simulate protein structure, Baker uses 'mini-threading', in which small protein segments of known structure are joined together to build the complete protein, using an optimization scheme to obtain the best fit. This approach has proven quite successful in the recent Critical Assessment of Structure Prediction - CASP3 [<http://predictioncenter.llnl.gov/casp3/>] competition. Some of the best *ab initio* protein structure prediction methods (including Baker's) still give structures in the four to eight Angstrom resolution range; unfortunately, this level of resolution usually does not allow the assignment of functions to individual residues. It is, however, occasionally possible to use these structure predictions to assign a function, or to provide useful information for drug discovery. This is possible because the known 'local structures' from well characterized proteins can sometimes be stitched together in a different context to provide a reasonable estimate of the structure in a novel protein.

RNA structures

Extensive sequence information has led to the discovery of new RNA-encoding genes as well as new protein-encoding ones, and two speakers based their talks on RNA structures. Sean Eddy (Washington University, St Louis, USA) described the families of small nucleolar RNAs involved in splicing. In some cases, mRNA codes for a 'protein' which is never made, and the 'introns' in this RNA turn out to be essential RNAs. Eddy found small nucleolar RNAs, or 'snoRNAs', in such a setting. Putative snoRNA genes have now been found in *Archaea*, but not in bacteria, by sequence analysis alone.

The Intron Sequence Information System (ISIS) website [<http://isis.bit.uq.edu.au/>] described by Soeren Schandorff (University of Copenhagen, Denmark) contains information on more than 170,000 spliceosomal introns. ISIS contains phylogenetic and protein homology categories, information about individual sequences and various bioinformatics analyses of taxonomical groupings of sequences using non-redundant subsets of the data. Schandorff has found that at least 42% of all human genes are alternatively spliced, and considers this to be a conservative estimate. The ISIS web site is based at the Department of Mathematics, University of Queensland, Australia, where Schandorff's collaborators are working.

Algorithms

Support vector machines (SVMs) represent a general, non-linear machine-learning method, which is commonly used for classification but can also be used for the analysis of mRNA expression data from DNA chip microarrays. David Haussler (University of California, Santa Cruz, USA) gave a clear overview of the theory behind support vector machines (for more information see <http://www.kernel-machines.org>), and presented results of protein sequence alignments using a joint hidden marker model (HMM; a probabilistic method) and SVM approach (Fischer kernel), which performs better than existing approaches, such as PSI-BLAST and stand-alone HMMs.

Whole genome analysis

'Life on the Edge' was the theme of Bernhard Palsson's (University of California, San Diego, USA) talk about modeling metabolic genotypes of sequenced microbes. Using a combination of all the predicted proteins in the *Escherichia coli* K-12 genome (isolate MG1655), and an exhaustive knowledge of the biochemistry, Palsson has modeled metabolism under various growth conditions, and could accurately predict the behavior of 73 out of 80 mutants. This model also predicts that there are relatively few critical gene products in central metabolism, and some of the predicted non-essential genes have been experimentally verified.

Gene Meyers (Celera Genomics, USA) gave a detailed explanation of Celera's whole genome assembly of *Drosophila*. Meyers listed a set of requirements (for example, at least 6x coverage of the genome) that he had asked for in order to be able to complete assembly, and compared them with the material that he was given; Table 1 illustrates that, in nearly all categories, the final sequence had exceeded his list of requirements. The assembler was given a collection of 3.2 million paired reads, called mates, that were sequenced from the ends of a number of insert libraries. All low copy portions of the genome were identified by the assembler, built into contigs and ordered into scaffolds spanning each of the chromosomes. Of course, the *Drosophila* genome was

Table 1

A comparison of the quality of material expected from the Celera *Drosophila* sequence to that achieved for use in assembly

Description	Expectation	Final result
Shotgun coverage	10x	14.6x
Basepairs/read	500	551
Reads in pairs	70%	72%
Insert length distance	+/-3% variation	+/-3% variation
False positive rate on mate pairs	<1%	<0.34%
Map-coverage (BAC pairs)	15x	13.18x
Long-to-short ratio	1-4	1-1.32

considered a 'warm-up' project for the human genome project, which should be publicly available soon. Meyers seemed optimistic that the same strategy would work on the human genome project, and his description of the methodology put forward a convincing argument that this will indeed be the case. The poor quality of human gene finders, however, means that annotation of the coding regions will remain a greater, and potentially longer lasting challenge.