Meeting Report

Integrated bioinformatics – high-throughput interpretation of pathways and biology

Eric Jain and Kewal Jain

The Cambridge Healthtech Institute's third annual conference on Integrated Bioinformatics was held in Zurich, Switzerland, 24–26 January 2001.

The monitoring of gene expression and the identification of protein interactions provide useful information to aid the understanding of gene pathways and regulatory circuits. Massive amounts of data are being produced from this field of research and computational methods are needed to analyse it in a high-throughput manner. Algorithms, data management, analysis software and databases were the focus of this conference; only a selection of the numerous presentations at the meeting are mentioned in this brief report.

Computational analyses of gene expression data

DNA microarrays enable a global view of the cellular processes by monitoring transcription levels of numerous genes simultaneously. Ron Shamir (Tel Aviv University, Tel Aviv, Israel) presented a system for the analysis of raw data from DNA microarrays – CLICK (cluster identification via connectivity kernels), an algorithm for creating more homogenous gene clusters from expression data and at a higher speed than comparable algorithms. He described a prototype of PARVE (pathway reconstruction and visualization engine) for expanding biological networks from various input data and discussed briefly how universal DNA chips, originally designed for sequencing, can be used to reconstruct the sequence of single nucleotide polymorphisms (SNPs) or pathogens.

Jan Michel (LION Bioscience, Heidelberg, Germany), Deepak Thakker (Rosetta Inpharmatics, Kirkland, WA, USA), David Nick (Spotfire Inc, Cambridge, MA, USA) and Frank White (InforMax, N. Bethesda, MD, USA) described the tools that their companies use for analysing and, especially for visualizing, microarray gene expression data. All have desktop applications and graphical display options, enable the import and export of data in various formats and provide a data-filtering function. The analysis tools of LION Bioscience can speed up the process of identifying and prioritizing potential target genes. Rosetta's Gene Expression Mark-up Language (GEML) facilitates interchange of gene expression data from multiple DNA technologies. Spotfire provides integrated informatics by simple access to data and ability to clearly share analyses and conclusions. The integrated data analysis and visualization platform of Informax, combines analytical tools for sequence, expression, genomic DNA and protein function data to facilitate natural paths of research leading to drug discovery.

Frederique Lisacek (Geneva Bioinformatics, Geneva, Switzerland) presented ProteomeSystems GlycoSuiteDB, a curated database of glycan structures (http://www.glycosuite.com/). This is the first relational database of protein glycosylation and includes a diverse group of protein modifications that are important in protein function and diseases such as cancer. The GlycoSuiteDB is a natural extension of the SWISS-PROT protein database and, in combination, these will provide researchers with unparalleled opportunities to address commercially important problems in the field of proteomics.

Andrew Whiteley (Amersham Pharmacia Biotech, Sunnyvale, CA, USA) highlighted the importance of an efficient and well-integrated Laboratory Workflow System capable of keeping track of variables that range from machine throughput to data quality, to ensure high quality data in large-scale drug development.

Gene function prediction

Christos Ouzounis (European Bioinformatics Institute, Cambridge, UK) presented RAGE (http://www.ebi.ac.uk/research/cgg/services/rage/), an algorithm for clustering large protein databases into families by taking into consideration single protein domains and therefore minimizing typical inaccuracies. Steen Knudson (Technical University of Denmark, Lyngby, Denmark) then demonstrated how regulatory patterns in promoter regions of bacteria, yeast and human genomes can be detected using expression data[1].

Thomas Werner (Genomatix Software, Munich, Germany) showed that by using promoter models instead of simple alignments, the function of genes could be predicted with a higher accuracy and fewer false positive results. He pointed out the importance of distinguishing between coexpressed and coregulated genes when analysing expression data given that related functions can only be assumed for coregulated genes.

Genomatix also provides a database of promoters (http://www.genomatix.de/). Stephan Heymann (Kelman Gesellschaft für Geninformation, Berlin, Germany) presented a software for visualizing gene interaction networks, graphs with nodes representing gene products and their mutant versions, each having a separate connection point for various connecting lines representing protein – protein interactions, phylogenetic links, coexpression or any other relations (http://www.kelman.de/frameset_de.html?+16000).

Rajan Kumar (Sarnoff Corporation, Princeton, NJ, USA) discussed the building of pathway models. Given that biological systems are too complex for most approaches, statistical methods must be used. So far, simulations have been run with an apoptosis pathway model and have simulated experimental observations.

Gene profiling for target identification

Functional attributes are currently assigned to genes either by alignment to well-characterized sequences or by comparison to a model. Using this approach, only a small percentage of all newly identified genes can be categorized accurately and too many conclusions are

http://tibtech.trends.com 0167-7799/01/$ – see front matter PII: S0167-7799(01)01606-7

drawn from too little experimental data, as shown by frequent errors in automatic annotation. Jean-Michel Claverie (Centre National de la Recherche Scientific, Marseille, France) highlighted the need for more experimental data besides sequencing. He described analyses of multi-conditional gene-expression data and its application in tumour classification.

Holger Hiemisch (BASF-LYNX Bioscience AG, Heidelberg, Germany) explained laboratory techniques for transcription profiling. Megaclone technology uses microscopic glass beads, each binding one piece of cDNA generated from cellular mRNA, thereby creating a library of the transcriptome with individual expression levels preserved. Megasort uses competitive hybridization on beads to show how much a gene has been up- or down regulated. Massively Parallel Signature Sequencing (MPSS) immobilizes up to one million beads with bound cDNA in a flow cell and partially sequences them to reveal a signature characteristic of cDNA, thereby providing a profile of the gene activity. These techniques have been used for expression profiling in the brain to discover new targets for drugs to treat neurological disorders.

SNP scoring is used as a tool to detect complex disease loci. Charles Mein (Royal London Medical School, London, UK) described that it is difficult to map complex diseases to a genome because the effects of single mutations can be weak and their loci are unlikely to be known. They used Invader technology (Third Wave Technologies, Madison, WI, USA), which provides a simple method for measuring the ratios of alleles for all loci, thereby enabling comparisons[2]. Invader assay can be automated and the advantages, as well as disadvantages, of this approach were discussed.

Protein expression

The study of protein expression is important because changes in protein expression might provide clues to the role of certain proteins in disease and some of the identified proteins might map to known genetic loci of a disease[3]. The identification of protein–protein interactions is useful in the drug discovery process. Assigning functions to the proteins encoded within a genome has now become important. John Overington (Inpharmatica Ltd, London, UK) presented Biopendium, a database with pre-computed analyses relating protein sequences to structural and ligand data (http://www.biopendium.com/). This is an example of linking bioinformatics and chemoinformatics through protein structure. A transmembrane glycoprotein, nicastrin, which appears to modulate the production of amyloid-beta peptide and might be involved in the development of Alzheimer's disease, was annotated by Biopendium[4].

Jong Park (European Bioinformatics Institute, Cambridge, UK) demonstrated that the Protein Structural Interaction Map (PSIMAP) could be built using structural information derived from PDB and a classification of protein structures. A domain-to-domain interaction can be detected by calculating atom-to-atom distances between pairs of domains.

Alon Amit (Compugen, Tel Aviv, Israel) presented ProLoc, an algorithm that predicts subcellular localization of proteins by taking into consideration various factors such as protein length, amino acid composition, predicted transmembrane regions, signal peptides and Pfam domains. This is an innovation of the LEADS platform of Compugen, an algorithm-driven bioinformatics platform for the analysis of genomic data that is designed to help pharmaceutical, biotechnology and other life science companies accelerate the development of drugs and biological products.

Computational genomics

Douglas Brutlag (Stanford University, Stanford, CA, USA) pointed out that any effort to provide extensively annotated genomes must be based on numerous collaborations because no single company currently provides all the necessary tools for such a task. He discussed the data pipeline for the annotation process of the first comprehensively annotated version of the human genome by DoubleTwist Inc (Oakland, CA, USA; http://DoubleTwist.com) based on both public and proprietary data from more than 30 collaborators.

Barry Robson (IBM Computational Biology Center, Yorktown Heights, NY, USA) highlighted the importance of proper data integration and annotation, automation of complex processes and protein structure modelling to increase drug output and reduce the constantly increasing expenditures on research and development. Improving bioinformatic analysis of gene function is an active research area at IBM. Research into protein modelling is being conducted by hardware research (e.g. the Blue Gene Project) as well as software research. IBM is aware of the challenge of developing personalized medicines of the future based on proteomic and genomic knowledge evaluated by bioinformatics.

Conclusions

This was an excellent conference with a 'state-of-the-art' review of the application of bioinformatics in life sciences. It was a mix of academic as well as commercial bioinformatics representatives with the latter predominating, which corresponds to the more extensive development of this discipline in the industry. To be successful, commercial bioinformatics software has to appeal either to computer science people, who typically prefer programmable and customizable applications, or to biologists, who often prefer simple graphical applications. Even better, of course, would be software that supports both types of users.

With the number of bioinformatics tools increasing, there is a large demand for software that is capable of integrating all relevant tools and thereby of saving the need to spend a large effort investigating which tools are available and should be used.

The major bottleneck in bioinformatics was said to lie in the algorithms being used and not in the power of the hardware. The general consensus was that, although there are a lot of useful solutions available today, something bigger and better must be hiding just behind the next corner; or perhaps the one after that.

References

1 Jenson, L. *et al.* (2000) Automatic discovery of regulatory patterns in regions based on whole cell expression data and functional annotation. *Bioinformatics* 16, 326–333
2 Mein, C.A. *et al.* (2000) Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation. *Genome Res.* 10, 330–343
3 Jain, K.K. (2001) *Proteomics: technologies and commercial opportunities*. Jain PharmaBiotech Publications, Basel, Switzerland
4 Yu, G. *et al.* (2000) Nicastrin modulates presenilin-mediated notch/glp-1 signal transduction and bAPP processing. *Nature* 407, 48–54

Eric Jain
Kewal Jain*
Jain PharmaBiotech, Blaesiring 7,
4057 Basel, Switzerland
*e-mail: jain@pharmabiotech.ch