Beyond the genome: turning data into knowledge

Joanna Milburn, joanna.milburn@current-trends.com

Since the publication of the draft sequence of the human genome in April 2001[1,2], the pressure on life scientists to annotate and gain useful knowledge from the millions of As, Gs, Cs and Ts, is ever increasing. Now that the initial hype has passed, the life sciences field demands the technologies and computer power that can produce and analyze data in a high-throughput manner for drug discovery and disease diagnosis. Cambridge Healthtech Institute's *Beyond Genome 2001* conference[a] (17–22 June 2001, San Francisco, CA, USA) brought together the latest research in these technologies in a 'tri-conference' that encompassed the *10th Annual Conference of Bioinformatics and Genome Research*, the *3rd Annual Conference of* In Silico *Biology* and the *5th Annual Conference of Proteomics*.

Bioinformatics and genome research

Bioinformatics has exploded and evolved since its inception in 1987. The *10th Annual Conference of Bioinformatics and Genome Research* reflected just that, attracting both experts in the field and those inevitably interested in expanding their research capabilities. The meeting focused on the exploitation of genomic information, and the development of computational tools as an aid to understanding the underlying mechanisms in biology.

Mining gene expression data

Several presentations focused on the analysis of gene expression data. Jim Golden (CuraGen, Branford, CT, USA) described a transcript profiling method called GeneCalling®, which processes mRNA into fragments and makes comparisons between not only diseased and normal tissue, but also drug treated versus untreated tissues. Each differentially expressed fragment is queried against gene databases for the identification of known and novel genes. The aim of the technology is to measure gene activity within the context of a specific disease or drug treatment for use in therapeutics. Golden estimates that of the ~30,000 genes in the human genome, there are ~8200 that have potential for therapeutic use. CuraGen already has 5500 of these on a 'pharmaceutically-tractable genome chip'. The technology has also been used to map pathways, and the company plans to release the complete pathway map for *Drosophila melanogaster* in the near future.

The TANGO (Transcription ANalysis of GenOmes) system, used for microarray gene-expression data analysis, was described by Terry Gaasterland (Rockefeller University, New York, NY, USA). The system integrates multi-genome information about gene-sequence family conservation, genome organization, metabolic pathways and putative promoter sites into gene expression clusters. The system not only examines single clusters but also generates explanations of gene expression associated with a biological entity, such as expression patterns of genes as part of a specific metabolic pathway.

Daniel Shoemaker (Rosetta Inpharmatics, Kirkland, WA, USA) described exon-based expression profiling, using ink-jet oligonucleotide arrays. The method can define full-length transcripts based on the co-regulated expression of component exons. Rosetta has already generated a chromosome-22 exon array in 10 bp steps in what is known as a tiling array. To put this into context, the array used two-thirds of a glass slide and represents 1% of the human genome, meaning the entire genome could be displayed as exons on 200–300 glass slides. Rosetta hopes to design the ultimate exon array for use under 60 different conditions.

Phage display libraries

A novel alternative to microarrays was described by Lee Makowski (Argonne National Laboratory (Argonne, IL, USA). This technique uses phage-display peptide libraries for the generation of consensus sequences for small-molecule binding to proteins. The technique uses libraries of up to 10^9–10^{10} phage, each displaying a different peptide on their surface. The main advantage of this technique over microarrays is that the peptide of interest can be isolated and amplified in large amounts for characterization. However, the method has the disadvantage of having to encounter the dynamics of phage–host biology. Nevertheless, the technique has been used successfully to map contact-residues in several known targets of the anticancer drug paclitaxel (Taxol™), and to identify a novel paclitaxel receptor.

In silico biology: modelling life

At the *3rd Annual Conference of* In Silico *Biology*, the presentations aimed to define the concept of systems biology, the techniques involved, and implications

[a] Conference proceedings will be published for the first time this year, available through Cambridge Healthtech Institute.

1359-6446/01/$ – see front matter PII: S1359-6446(01)01953-5

for the next decade. The keynote presentation by Leroy Hood (Institute for Systems Biology, Seattle, WA, USA) gave an excellent overview of the field as it stands in the post-genomic era. He suggests that the concept of systems biology is somewhere between hypothesis driven and drug-discovery driven. His views on the long-term impact of the Human Genome Project were to generate a 'periodic table of life': a lexicon of genomic and proteomic motifs. He envisions the systems approach to biology in four steps: (1) define all elements; (2) perturb the system to establish the functional and measurable relationships of elements; (3) integrate information and model; and (4) re-iteration of steps 1–3. Hood stressed the importance of mathematical tool development even without the biological data; although initially they might appear of little use, in three years the amount of data that needs processing will have increased, by which time the models will be well established.

The need to develop complementary theoretical *in silico* approaches that enable the experimental investigation of the systematic properties of biological systems was highlighted by Bernhard Palsson (University of California, Menlo Park, CA, USA). He described 20th-century biology as taking a 'reductionist approach' (that is, investigation of individual cellular components and their properties), and how this must change in the 21st century to an integrative approach (bioinformatics, systems biology, mathematical modelling and computer simulation). Palsson focused on reconstructing cellular functions, and on how well *in silico* models correspond to *in vivo* experimentation. With full knowledge, the exact solution to theory-based models of cells can be determined. However, with the currently incomplete knowledge about cellular biochemical reaction networks, the exact solutions cannot be determined, but a solution space can be generated given several currently available factors (from research and literature) and governing constraints (such as mass balance and thermodynamics). This solution space contains all the possible functions of the cell that are consistent with the defined factors and constraints. This procedure then forms the basis for data-driven iterative model building within a framework of applying successive constraints to continually confine possible biological functions.

James Bassingthwaighte (University of Washington, Seattle, WA, USA) (who, some time ago, coined the popular term physiome), highlighted the need for more human physiological *in silico* models and for more physiological data to be gathered. Currently, there is no physiological database such as those generated for other 'omes'. He described the mission of the recently formed Physiome Project, which aims to define the physiome through databases and development of a sequence of model types (e.g. descriptions of structure and function, logical prediction, and so on). All information generated from the project will be available in the public domain in an open system. The incentives for building such a system are the determination of effective targets for genomic or pharmaceutical therapy, and the design of artificial or tissue engineered biocompatible implants.

Proteomics: from proteins to drugs

The *5th Annual Conference of Proteomics* brought together current efforts in the pharmaceutical industry and academia to assimilate the daily explosion of data into real drug targets and small-molecule drugs.

Protein–protein interactions and pathways

One issue that was given high importance throughout all of the conferences was protein–protein interactions. Donny Strosberg (Hybrigenics, Paris, France) focused on efforts to turn genomic information into knowledge of protein function. He described Hybrigenic's new PIM® (Protein Interaction Map) technology, which uses a high-throughput yeast-two-hybrid approach to decipher protein–protein interactions with low rates of false positives and false negatives. Interactions are determined for their reliability and new domains can be identified. The pathways are then reconstituted, and PIMs are displayed in a bioinformatics program, PIMRider®. This assigns functions and roles to interacting proteins to select targets for validation. Targets are then validated *in vivo* using a molecular tool called SID® (Selected Interacting Domain), which is used, for example, as a surrogate ligand of the target protein. This technology has been used to study the human protein Rac1 (known to interact with several protein partners), which has been studied by research groups for several years. In a single screen, 12 known interactions were confirmed and seven novel protein partners were identified.

The problems of taking a gene-centric view when developing drug therapies was highlighted by Nat Goodman (3rd Millennium, Cambridge, MA, USA). Drugs based solely on targeted genes might be effective *in vitro,* but might not necessarily treat the associated disease in the whole organism. Pathways and interaction data are already being accumulated in several different formats, but no standard means of representing or exchanging these data exists, to date. To prevent the confusion that has occurred in bioinformatics because of the lack of early standards, 3rd Millennium plans to pre-empt this issue by considering software specifications (such as the use of a universal language) when pathway and interactions databases are in their early stages. Goodman described the aim of the recently formed BioPathways Consortium, which hopes to capture, organize and use data to construct representations for systems biology using structural and dynamic information. His conclusion was that interactions are

to pathways as transcripts are to genes. The company is currently developing pathway models for Huntington's disease.

The BioKnowledge Library, described by James Garrels (Incyte, Beverly, MA, USA), is a resource built by comprehensive curation of the research literature from model organisms that enables data mining. Garrels emphasized the importance of the literature in the following context: there are three billion letters in the human genome, but there are 50 billion bytes of information of text generated per year from the literature. One component of the library is GPCR-PD™, which represents in-depth curation of the literature for G-protein-coupled receptors (GPCRs). The BioKnowledge Library is currently being used to annotate the proteins predicted from the human genome to create the complete annotated human proteome.

From gene sequence to protein function

Leonardo Brizuela (Harvard Medical School, Boston, MA, USA) described the ongoing effort of the Harvard Medical School to define the function of all proteins. To this end, they are building the FLEX™ (Full-Length EXpression) database, containing all known genes and predicted open-reading frames from human and other model organisms. The clones in the database are acquired using software to predict PCR primers from available sequence resources. The repository is prepared in a format that enables the transfer of thousands of genes into expression vectors in an overnight step, thus facilitating high-throughput protein expression studies. An advantage of the method is that the cDNAs are configured to enable preparation of proteins with or without the N- and/or C-terminal fusion peptides as required. The repository currently contains >1500 clones.

High-throughput technologies

As models attempt to cope with the sheer volume of data already generated, high-throughput technologies continue to evolve. This was highlighted in a poster presentation describing a versatile high-throughput assay platform. BioTrove (Cambridge, MA, USA) presented The Living Chip™ – a microdevice using high-density arrays for parallel screening and storage or synthesis of organic compounds. Biotrove claims that the chip features the density of a microarray with the functionality of a microtitre plate (Fig. 1). The platform capitalizes on the surface forces that predominate with nanolitre volumes by using these forces to facilitate reagent loading, mixing and recovery. Data were presented for phage-display screening, but the chip is also being developed for yeast-two-hybrid and mammalian cell assays, combinatorial library storage and protein evolution.

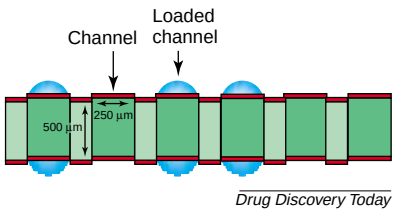

Figure 1. The Living Chip™ is a uniform addressable array of through-holes (250 µm²), precision machined in a silicon wafer (500 µm thick). Each channel typically holds 50–100 nl. Proprietary processes are used to coat the surfaces of the arrays so that they are hydrophobic (red) and hydrophilic (green). Consequently, 10,000 or more separate 50 nl volumes are instantaneously loaded into the chip by dipping into solution, with the liquid held in the channels by surface tension. Parallel initiation of reactions is achieved by stacking the chips, and no cross-talk occurs between channels. Sample recovery is achieved with a blast of air, thereby avoiding syringes and sterilization. Reactions can be monitored in parallel by using colorimetric, fluorometric or luminescent readouts.

The future

The take-home message from the *Beyond Genome 2001* meeting was the importance of data integration and availability in the public domain. Although open databases are clearly the way forward, the issue of quality control and assurance of data needs to be addressed. However, it is clear that only through integration of data will the metamorphosis of data into knowledge really provide results in the post-genomics era.

References

1 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
2 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921