

it is much more verbose, requiring 10–1000 times as much network bandwidth because of the XML markup.

In my talk, I described an XML format in which markup is reduced by providing a header that describes the contents of the file, the bulk of which appears in plain text columns with only higher level data containing markup. Although dismissed as sacrilege by XML purists, similar strategies had been considered by other members of the audience who needed to handle very large datasets.

Ontology and semantics: the meaning(s) of words

So, it turns out that the real problem facing data integration is not the technologies involved, but in getting everyone to agree on the meaning of the terms they use. For example, we are all pretty clear about what a protein is ... or are we? Does the term 'protein' mean a poly-peptide chain or an assembled quaternary structure? Are post-translational modifications or cofactors included in a 'protein'? These differences might not seem important when you are talking to a colleague; the exact meaning

becomes obvious from context, but to integrate data resources successfully, such inconsistencies must be removed.

Another problem is the 'not invented here syndrome'. Individuals often believe that they can 'do it better' and this might well be true! However, if we are ever to have fully integrated data resources, individuals must communicate with each other. The data dictionary of the mmCIF format for protein crystallographic data took 10 years to define (and still not everyone is happy with it). We cannot afford to wait that long to have agreed restricted vocabularies and ontologies for the rest of biology and an ontology for genetics has been produced much more quickly (<http://www.geneontology.org>). Alan Robinson (EBI, Hinxton, UK) suggested that there should be many more instances of groups of people being forced to sit in a room together and fed pizza until they agree on something!

Of course, for us to integrate data across the Web in a consistent and automated manner, we must also have confidence in the quality of those data. Luca Toldo (Merck KGaA, Darmstadt, Germany) talked about the

need for quality assurance measures to be linked to raw data and annotation of those data. This is a serious issue and currently very few data resources contain confidence scores.

Conclusions

In conclusion, the feelings of the meeting were very positive:

- the technologies are largely available, but people must be persuaded to use them;
- firewalls can cause problems;
- data providers need to move away from hiding all their data behind point-and-click interfaces and make data available in a machine understandable form;
- measures of data quality must be provided with the data;
- restricted vocabularies, data dictionaries and ontologies are the key factors in integrating resources.

Andrew C.R. Martin

Lecturer in Bioinformatics, School of Animal and Microbial Sciences, University of Reading, PO Box 228, Whiteknights, Reading, UK RG6 6AJ.
e-mail: a.c.r.martin@reading.ac.uk;
andrew@bioinf.org.uk

Proteins in gels, computers, crystals and camels

Konrad Büssow

The IBC's Proteomics 2001 meeting was held in Philadelphia, PA, USA, 14–17 May 2001.

With the introduction of automated technologies in the field of molecular biology and especially microarray technology, genome and gene expression analysis has been accelerated enormously. To complement automated genome analysis, automation and miniaturization is now being introduced to the analysis of the proteomes. Proteomes are analysed using electrophoretic separation and automated chromatography, and protein components are identified using mass spectrometry at an accelerating rate. Three-dimensional structures of large sets of proteins are elucidated by functional genomics projects. Analogous to DNA microarrays, protein arrays offer the opportunity to screen thousands of immobilized biomolecules at a time, using steadily reduced amounts of sample.

The goal of this meeting was to determine the impact of contemporary protein

research on drug development. There was a large focus on 2D-electrophoresis, alternative protein separation techniques and mass spectrometry. Notably, a whole section of the conference was dedicated to sample preparation, which demonstrates the importance of standardization to allow proteome analysis to be reproducible.

'...automation and miniaturization is now being introduced to the analysis of proteomes.'

The laboratory automation industry has discovered a new market – no less than four manufacturers exhibited 'picking' robots for 2D-gel protein spots. The companies Tecan (Männedorf, Switzerland; <http://www.tecan.com>), Qiagen (Hilden, Germany; <http://www.qiagen.com>), Genetix (New Milton, UK; <http://www.genetix.co.uk>) and Marsh (Rochester, NY, USA; <http://www.marshbio.com>) are well known for robot technology in the genome

analysis field and are now moving into proteomics.

However, high-throughput solutions that do not require sophisticated robot technology were also presented. Martin Schürenberg (Bruker Daltonics, Bremen, Germany) presented a novel sample support for MALDI mass spectrometry, named AnchorChip, which owing to its surface chemistry, enables efficient sample concentration and purification, leading to an increase of sensitivity by orders of magnitude.

The representation of living systems in a computer language was addressed by Robert Franza (University of Washington, WA, USA). He presented the Cell Systems Initiative (CSI) to represent and store biological data. The CSI is developing a new scientific notation with symbols for biological terms such as 'interaction' and 'modification', and is also developing computer animation to visualize complex biological pathways.

Protein function

The functional genomics section of the conference covered different general approaches towards understanding protein function. The model organism *Caenorhabditis elegans*, offers several unique advantages, as pointed out by Thierry Bogaert (deVGen nv, Gent, Belgium). Creating transgenic animals only requires feeding the worms with bacteria that carry the gene construct of interest. When the phenotypes of transgenic animals are compared with those of wild-type animals that have been administered drugs, the biochemical pathway targeted by the drugs can be identified.

A method for identifying novel protein ligands was presented by David Nelson (Anadys Pharmaceuticals, Inc., San Diego, CA, USA; <http://www.anadyspharma.com>) and Raymond Salemme (3-Dimensional Pharmaceuticals, Exton, PA, USA; <http://www.3dp.com>). The stabilizing effect of ligand binding increases the denaturation temperature of the protein. Highly sensitive fluorescence assays for protein denaturation allow the screening of large numbers of compounds in a microtiter plate format. The method can be used to identify lead compounds but also identifies natural ligands of orphan receptors or other proteins of unknown function.

Aled Edwards (University of Toronto, Toronto, Canada) showed how recombinant proteins produced for structural analysis can be used for protein-protein interaction studies. He pointed out the limitations of genome-wide interaction screenings and described the problem of inaccurate data in the literature. He proposed the use of highly purified and well-characterized proteins to not only determine the absence or presence of interaction but also to titrate and carefully elucidate the reversibility, stoichiometry and affinity of the interactions.

Protein structure

The goal of structural genomics is to obtain structural models for all proteins, either determined experimentally or calculated from structures of closely related sequences. Pharmaceutical companies will benefit from structural genomics by structural models of target proteins at a very early stage of drug development. Peter Rose (Pfizer Global R&D, San Diego, CA, USA and La Jolla/Agouron Pharmaceuticals Inc.) showed how structural genomics will influence structure-based drug development.

Structures obtained using structural genomics will help in target selection and give hints on protein expression and crystallization. At a later stage, structures of homologs of target proteins enable rapid structure determination by homology modeling and molecular replacement.

Protein crystallization was traditionally performed in a 24-well format but the 96-well microtiter plate format is now becoming the standard. George DeTitta (Hauptman-Woodward Medical Research Institute, Buffalo, NY, USA) presented his highly automated protein crystallization laboratory. Crystallization is carried out under paraffin oil in 0.4 µl drops, which are dispensed by a 96-needle pipetting device (Robbins Scientific Hydra, Sunnyvale, CA, USA; <http://www.robsci.com>) into 1536-well microtiter plates. Thus, 1536 conditions are screened in one experiment.

In my talk I presented the Berlin Protein Structure Factory, a structural genomics initiative focussed on human proteins of medical interest (PSF; <http://www.fu-berlin.de/psf>). The Protein Structure Factory combines expertise of several Berlin academic research institutes and companies in the fields of cDNA cloning, protein expression and crystallization, biophysical characterization and NMR.

A systematic protein expression approach was presented by Shane Taremi (The Schering-Plough Research Institute, Kenilworth, NJ, USA). Open reading frames are cloned into vectors for expression in *Escherichia coli*, *Sf9* or yeast expression vectors. Protein solubility during extraction from cells and fusion partner cleavage is affected by buffer conditions and suitable additives. Optimal buffers are empirically determined using a sparse matrix screen. For proteins that remain insoluble, mutants with improved yield and solubility are produced. Beneficial point mutations are identified using a colony blot procedure and light scattering is used to monitor protein solubility and aggregation. Alternatively, as Cheryl Arrowsmith (Ontario Cancer Institute, Toronto, Canada) demonstrated, ¹⁵N NMR spectroscopy can rapidly determine the folding status of a protein.

Protein arrays

There was no dedicated session on protein and antibody arrays but a growing interest in these new formats was notable. Antibody arrays hold great potential for the simultaneous monitoring of thousands

of proteins from biological samples, for example, blood serum.

Lawrence Cohen (Zyomix, Inc., Hayward, CA, USA) presented the design of his company's protein microchips. Highly developed surface chemistry prevents nonspecific binding and eliminates the need for the extensive washing steps that are required in standard immunoassays.

The production of antibody arrays is currently hampered by a lack of highly specific and stable antibodies against large numbers of proteins or even complete proteomes. Karen Silence (Free University of Brussels, Brussels, Belgium) demonstrated a strategy for generating antibodies against large numbers of proteins simultaneously that takes advantage of a characteristic of the immune system of cameloids. Of a camel's antibodies, 50% lack a light chain and consist only of heavy chains. This facilitates the cloning of variable domains of the antibody and subsequent recombinant expression and selection by phage display. Usually, very large repertoires of antibody presenting phage have to be generated to represent all combinations of antibody heavy and light chains.

Conclusion

A general theme of the meeting was the standardization and automation of the study of protein function. Several approaches have to be integrated to ultimately understand and make use of protein function, including the following:

- separation of whole proteomes using 2D-electrophoresis and protein identification by mass spectrometry;
- systematic identification and refinement of ligands in drug development;
- high throughput protein structure determination;
- development of repertoires and arrays of recombinant proteins and antibodies;
- systematic investigation of protein-protein interactions.

Advances in the separation and identification of the proteins of a cell were demonstrated, as well as approaches towards structure determination and generation of antibodies for large numbers of proteins. Drug development benefits from new ligand screening technologies and from rapid protein structure determination.

Konrad Büssov

Max-Planck Institute of Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany.
e-mail: buessow@molgen.mpg.de