conference report

Microarrays, databases and hard, hard sums

David Bousfield, david@ganesha-associates.com

Therapeutic applications of computational biology
Hinxton, UK
4–6 September 2005
Organizers:
European Bioinformatics Institute (EBI) and Nature Biotechnology

Computational biology is of key strategic importance to the healthcare industry. It is used at all stages of the drug discovery and development pipeline, from target identification through to regulatory approval. The conference, *Therapeutic applications of computational biology* held on 4–6 September in Hinxton, UK, was designed to show not only how computational biology can expedite the development of new therapeutic, diagnostic and prognostic approaches, but also how we can work towards solving current challenges in the field.

The main program sessions were organized along canonical lines, entitled 'target to lead optimization', 'cheminformatics meets bioinformatics', 'toxicoinformatics' and 'into the clinic'. The definition of computational biology used by the organizers in choosing topics and speakers was quite broad subsuming bioinformatics and systems biology, therefore the conference provided a useful overview of the field.

Using genetics to dissect disease complexity

Many common diseases, such as heart attack, asthma, stroke and cancer, result from the interplay of multiple genes with environmental and other health factors. Genetics offers a tool for unraveling some of this complexity when information about disease incidence can be correlated with genetic variation across human populations in which the generational lineage is known. Kari Stefansson (DeCODE Genetics, Iceland) described the development of a database derived from 100,000 Icelandic volunteers for whom genealogical records exist since before 1800. High-throughput genotyping enables haplotype and linkage disequilibrium studies to isolate key biomarkers and genes associated with disease risk.

For example, a genome-wide scan of 296 families revealed 713 that had suffered a myocardial infarction (MI). Further linkage analysis revealed that a 4 single nucleotide polymorphism (SNP) haplotype in the gene *ALOX5AP* conferred an almost twofold increase in risk of MI and stroke. The encoded lipoxygenase-activating protein, FLAP, regulates the genesis of leurkotriene inflammatory mediators, already implicated in atherosclerosis in the mouse apolipoprotein E knockout model and in human studies [1,2].

Biomarkers can also be used to guide clinical prognosis and treatment strategies. Yixin Wang and his colleagues at Veridex (New Jersey) together with collaborators at the Erasmus Medical Center in the Netherlands have developed a microarray assay that can identify which patients with lymph node negative (LNN) breast cancer are most at risk of developing metastatic disease [3]. Currently, 25% of LNN patients will suffer a relapse after surgery and over 80% receive adjuvant chemotherapy, which can have extremely toxic side effects.

The assay is based on a 76-gene signature and Veridex is now working with chip manufacturer Affymetrix to develop a commercially robust version by reducing overall assay complexity, automating data analysis and resolving IP and regulatory issues. A pivotal clinical trial will be underway shortly.

Getting a signature is relatively easy, but what we really need is to be able to read and understand the fine print and move from correlative to robust causative explanations. Stefansson and Wang pointed to the lability of specific gene-disease correlations (i.e. different gene combinations are responsible for a single clinical profile), although at the pathway level the associations appear to be more robust. Significant geographical and racial differences in haplotype frequencies exist and the reproducibility of linkage studies can be challenged even when the experimental methodology is robust. 'Nevertheless', Stefansson remarked, 'Icelanders are reasonable models for *Homo sapiens*'.

Systems and *in silico* approaches to understanding complexity

There is a view among some biologists that systems biology is simply joining up all the things we have learnt in our reductionist past. David Searls (GSK, Philadelphia) challenged this notion with a review of metaphors for biological complexity drawn from electronic circuitry, neural networks and linguistics. These conceptual tools enable an appreciation

1359-6446/05/$ – see front matter PII: S1359-6446(05)03663-9

conference report

of the importance of emergent properties in biology [4–8].

For example, how much contextual information is required to define unambiguously the pronunciation of 'ough', examples being tough, ought, wrought, through, thought and so on. The importance of contextual information extends to the scale of the sentence (e.g. 'does a buck like does') and beyond. A parallel in biology is the ability of amino acid heptamers (e.g. ASVKQVS) to exhibit different conformations in different proteins (e.g. the β-sheet in an aminopeptidase and the α-helix in a guanylate kinase). Secondary structure can be determined by nonlocal interactions and the existence of identical hepta- or even octa-meric amino acid sequences does not imply identical tertiary structure or function.

Pedro Mendes (Virginia Bioinformatics Institute, Blacksburg) described how traditional bottom–up approaches to modeling metabolic pathways (Gepasi, www.gepasi.org; COPASI, www.copasi.org), which require an understanding of the kinetics of each reaction in a network, can still give rise to nonintuitive input–output behavior. In addition, individual enzymes can be multifunctional, and cross-talk between signaling pathways can give rise to emergent and nonintuitive properties [9].

Visualizing chemical space

Harren Jhoti (Astex, Cambridge) described how a fragment-based discovery strategy is being used to develop novel compounds for targets such as cyclin-dependant kinases, key proteins involved in cancer. A shortcoming of traditional HTS is caused by the complexity and the relatively large size of the compounds being screened. An alternative approach developed by Astex uses very small fragments to keep low the overall complexity and molecular weight of each drug candidate. Typically, bioassays are unable to detect such small fragments because of their low-potency binding to the protein target. But their detection and binding properties can be measured using biophysical techniques such as X-ray crystallography and NMR spectroscopy. The final drug candidate will be constructed from a series of fragments, each binding to neighboring sites within the protein crystal. Astex's lead drug candidate, AT7519, is a potent cell-cycle inhibitor that targets key cyclin-dependent kinases, which entered clinical development this year [10–12].

Advances in automated synthesis and HTS have led to the need for better library design. Dimitris Agrafiotis (Johnson and Johnson, Exton) described a range of *in silico* diversity profiling and visualization techniques for optimizing compound diversity and maximizing the hit rates [13,14].

Making the most of what we know

As Janet Thornton (Director of EMBL-EBI) explained in her welcoming remarks to the delegates, EBI's mission is to provide freely available data and bioinformatics services to enable industry to quickly adapt to and maximize the benefit from bioinformatics innovations and to foster the development of synergies between medical- and bio-informatics. This has led the EBI to invest heavily in the development and integration of databases with increasing degrees of functional sophistication.

Ewan Birney (EBI, Hinxton) described Reactome, one of the most recent additions to the fold. Reactome (www.reactome.org) is an NIH-funded collaboration between the EBI and Cold Spring Harbor Laboratory and is a curated database of biological processes in humans taking the 'parts list' of proteins, RNA molecules and metabolites and putting them together to create pathways. The basic information in Reactome is provided by bench biologists expert on a particular pathway and ranges from basic processes of metabolism to complex regulatory pathways such as hormonal signaling. Although Reactome is targeted at human pathways, it also includes many individual biochemical reactions from nonhuman systems, such as rat, mouse, pufferfish and zebrafish. Birney demonstrated how the graphical interface could be used to tackle evolutionary issues, such as lineage deletion rates and the modular structure of insulin signaling pathways.

Steve Bryant (NCBI, Bethesda) provided a complimentary review of recent database developments at NIH. The new PubChem (http://pubchem.ncbi.nlm.nih.gov/) database is organized as three linked databases within the NCBI's Entrez information retrieval system. These components are PubChem Substance, PubChem Compound, and PubChem BioAssay and are now being fed with data from a network of ten national screening centers across the USA. This screening centers' network will enable academic and government researchers to contribute in a much more vigorous way to an understanding of the mechanisms of disease and even to the identification of potential targets for new therapies.

As Brynn Williams-Jones (Pfizer, Sandwich) remarked, 'biology is now too big and complicated to fit in our heads' and it is essential that the proliferation of database resources is matched by an equal effort to integrate this information and more importantly enable the disclosure of new knowledge.

Williams-Jones reported on the use of refined SAR methods based around a library of 1567 structurally diverse molecules and 92 ligand-binding assays – a subset of the data found in Cerep's BioPrint database (www.cerep.com, [15]). This battery of assays provides the basis for a 'biospectrum' profiling a molecule's biological activity. Echoing Searls' presentation, the Pfizer approach makes use of the modularity of protein structure design to predict target affinities for new compounds, as well as potential off-target secondary pharmacology. The effectiveness of this approach has been demonstrated by using the system to predict adverse secondary pharmacology for a WHO list of 153 withdrawn drugs. Unfortunately, the proprietary nature of the work meant that there could be no real discussion of the details of the Pfizer system.

John Overington (InPharmatica, London, www.inpharmatica.com) followed with a description of their DrugStore and StARLITe database tools. StARLITe contains reported screening information for more than 300,000 compounds, their molecular targets and their biological activities. DrugStore is a knowledge-base covering key information on all FDA-approved drugs. Overington described how the system had been used successfully to profile the secondary pharmacology of several compounds, for example the identification of vitamin K epoxide reductase as a key target of warfarin [16–18].

Toxico-genomics comes of age

Robert Glen (Unilever Center for Molecular Sciences, Cambridge) said that all small

conference report

molecules are toxic, although the mechanisms (e.g. metabolism, insolubility, genetic predisposition, dose, tissue distribution) are diverse. The multiplicity of these mechanisms makes it hard to predict acute toxicity from QSAR or by using animal models. Chronic toxicity is even less tractable, with organ specificity, genetics, environment, age, sex and weight, all playing a role. It is no surprise then that toxicity is often unexpected and not discovered until clinical trials or later. Drug–drug, drug–food interactions and comorbidities add further dimensions to this complexity. Of the 548 new chemical entities (NCE's) approved between 1975 and 1999, nearly 20% now carry 'black box' warnings.

Han van de Waterbeemd (AstraZeneca, Macclesfield) provided a methodical overview showing how a hierarchical combination of *in vitro* and *in silico* approaches (which he has dubbed *in combo*) could be used to cover the absorption, distribution, metabolism and excretion (ADME) spectrum to optimize on effectiveness and cost [19–22], and he referred to the need for pharmaceutical companies to share their failure data. A similar conclusion was reached last year at a meeting organized by Richard Okita at the NIH and a series of recommendations for removing obstacles to collaborations between academia, government and industry has just been published by the FDA. According to Frederick Marcus (European Commission, Brussels), the EU's new innovative medicine initiative will also address this issue of sharing proprietary data.

An alternative to the mechanistic bottom–up approach is to use a biomarker approach. Paul Cornwell (Rosetta Inpharmatics, Seattle) described in a series of case studies how a compendium of expression profiles of 120 toxic compounds with known molecular mechanisms of action could be used to predict secondary pharmacology, such as the hepatotoxicity of a Merck development compound, MrkA, originally identified as an anti-HIV CCR-5 inhibitor drug. In another example, the sensitivity of the expression assay to the mitochondrial toxin usnic acid was shown to be much greater than a battery of conventional assays. Several companies are working on similar technologies but these new microarray assays will require extensive validation before they are approved for use within the highly conservative regulatory environment.

The rise of regulomics

In 2003, the FDA announced that it would encourage voluntary submission of pharmacogenomic datasets in an effort to develop standards for dealing with this type of information and the reduction of ADME–toxicology problems was a central component of its 2004 Critical Path document (www.fda.gov/oc/initiatives/criticalpath/, [23]).

Weida Tong from the FDA's National Center for Toxicological Research (Jefferson) described how their system was being developed to enable electronic submission of pharmacogenomic data (www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/). ArrayTrack provides an integrated solution for managing, analyzing and interpreting microarray gene expression data. It is MIAME (minimum information about a microarray experiment – a metadata standard developed by the EBI) compliant, storing microarray data and experiment parameters associated with a pharmacogenomics study. It is being integrated with SAS-SRS, as well as Jubilant BioSys and Ingenuity pathway analysis tools. Ultimately, the FDA aims to develop robust protocols so that microarray assays can become central to the regulatory assessment process.

This approach to the standardization of high-throughput toxicological analysis of compounds is essential. Toxic chemical law will undergo a major change in 2006, when the European Union is expected to enact legislation known as the 'registration, evaluation and authorization of chemicals' (REACH, http://ecb.jrc.it/REACH/, [24]). REACH will replace more than 40 existing directives and regulations and will require registration of chemicals made in or imported to the EU, assessment of the risks arising from chemical use and implementation of measures to manage risks. Under REACH, chemical producers will be required to provide basic toxicity and exposure information. Without this information the chemical will not be allowed on the market.

In 2012 an EU ban will be imposed on all animal testing in the cosmetics industry. Jos Kleinjans (Netherlands Toxicogenomics Center) explained how REACH would stimulate the development of microarray assay standards and described work being done to develop screens that could discriminate genotoxic from nongenotoxic effects using transcriptomic fingerprinting.

Meanwhile, back in the clinic

Gene expression profiling in acute myeloid leukemia, discussed by Peter van der Spek (Erasmus Medical Center), has demonstrated how powerful microarray signatures can be for clinical decision support and effective risk stratification. The difference in cost of retinoic acid therapy and a bone marrow transplant is US$150,000, but only if you know which patients will benefit from which treatment [25]. But at the moment these tests are expensive, slow and not robust enough for the clinical environment. Leigh Anderson (Plasma Proteome Institute, Washington) explained that, although plasma provides an extremely accessible source of biomarkers for disease, toxicity and therapeutic response, there are some significant problems to be solved if it is to realize this potential. The enormous dynamic range of plasma biomarkers (more than ten orders of magnitude) means that no single technique can be used for their separation and analysis. Furthermore, different experimental platforms seem to expose very different subsets of plasma components.

But, perhaps, the biggest barrier is that proteomics on its own is not capable of creating new clinical tests and the number of FDA CLIA (clinical laboratory improvement act)-approved plasma protein tests has fallen over the past decade. As it currently stands, the economics of test development are not commercially attractive.

The development of drug resistance is a major obstacle to successful treatment of HIV infection. There are three classes of HIV drugs, two are inhibitors of reverse transcriptase, the third protease inhibitors. Highly active anti-retroviral therapy involves at least three drugs from two of these classes.

Different strains of HIV develop resistance to different drugs at different rates, calling for frequent substitutions to the drug cocktail. It is through these extraordinary replication dynamics that HIV facilitates its escape from selective pressure exerted by the human

conference report

immune system and by combination drug therapy. Thomas Lengauer (MPI für Informatik, Saarbrucken) described how he and his colleagues have developed several computational methods to support the design of optimal anti-retroviral therapies based on viral genomic data [26]. These methods predict the evolution of drug resistance and optimize the selection of drug combinations. The result is a rule-based clinical decision support system driven by the patient's viral strain titres. A similar approach could be taken to other viral diseases treated with drug cocktails, such as hepatitis C.

The virtual patient

Seth Michelson (Entelos, Foster City) described how his company was using *in silico* simulation (PhysioLab) of novel targets and pathways to simulate the clinical impact of therapeutic interventions on ensembles of virtual patients. PhysioLab has been developed from a top–down modeling approach, starting with the disease symptoms and drilling down to the cellular mechanisms, signaling pathways and physiological homeostatic processes involved. The model incorporates data derived from the literature, panels of experts, as well as proprietary sources and is being used to optimize clinical trial design and to highlight differences between human and animal disease models.

Zvia Agur (Institute for Medical BioMathematics, Bene-Ataroth) described a similar approach called the 'virtual cancer patient' (VCP) designed to support the later preclinical phases of development, through clinical trials to treatment personalization. VCP takes a systems approach and uses a combination of consensus knowledge about disease processes, organ physiology and pharmacology plus calibration data from the patient to predict disease progression, optimum dosage design and toxicity. Successful applications of the model include prediction of optimal regimens for alleviating the immunogenicity of the thrombocytopenia drug TPO, while maintaining its efficacy.

A fence or an ambulance?

Michael Rawlins (NICE, London) eloquently brought the meeting to a close by focusing on the many problems facing the industry and its payers. Falling numbers of NCE's, rising costs of development and of individual medical care and simplifying the clinical trials process, all figured in his list. But was this the answer the audience was waiting for?

In the subsequent discussion the session chairmen highlighted some of the bottlenecks to progress, as they saw them. For David Searls, the priority was integration (data), integration (applications) and integration (domain knowledge). For Peter van der Spek, it was the huge difference between the sophistication of the research tools he has access to versus the cost, robustness and utility required within a hospital environment. Should the EBI include a network of hospitals in its outreach program? Robert Glen highlighted the need for greater access to data locked away in large pharma, whereas for David Bailey it was the poor funding of biotech (at least in the UK) and the need for a clearer differentiation of the roles of the public versus the private domains that were priority areas for attention.

As Rolf Apweiler concluded, the EBI has a crucial role to play in lubricating translational processes between academe, government and the healthcare industry and, as such, still has some way to go before its contribution could be said to be optimal. However, what this meeting achieved was to show that there are significant synergies between these different needs. Perhaps these areas should be the focus for a subsequent meeting?

References

1 Wang, Q. (2005) Molecular genetics of coronary artery disease. *Curr. Opin. Cardiol.* 20, 182–188
2 O'Donnell, C.J. (2005) Translating the human genome project into the prevention of myocardial infarction and stroke – getting close? *JAMA* 293, 2277–2279
3 Wang, Y. *et al*. (2005) Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365, 671–679
4 Searls, D.B. (2003) Pharmacophylogenomics: Genes, evolution and drug targets. *Nat. Rev. Drug Discov*. 2, 613–623
5 Kim, J.W. and Dang, C.V. (2005) Multifaceted roles of glycolytic enzymes. *Trends Biochem. Sci*. 30, 142–150
6 Barabasi, A-L. and Oltvar, Z.N. (2004) Network biology: understanding the cell's functional organisation. *Nat. Rev. Genet*. 5, 101–113
7 Bhalla, U.S. and Iyengar, R. (1999) Emergent properties of networks of biological signalling pathways. *Science* 283, 381–387
8 Papin, J.A. *et al*. (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.* 6, 99–111
9 Westerhoff, H. and Snoep, J. (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? *Eur. J. Biochem*. 267, 5313–5329
10 Gill, A. *et al*. (2005) The discovery of novel prpotein kinase inhibitors by using fragment-based high-throughput x-ray crystallography. *ChemBioChem* 6, 506–512
11 Jhoti, H. (2005) A new school for screening. *Nat. Biotechnol*. 23, 184–186
12 Hartshorn, M.J. *et al.* (2005) Fragment-based lead discovery using x-ray crystallography. *J. Med. Chem*. 48, 403–413
13 Agrafiotis, D.K. (1998) The diversity of chemical libraries. In *The Encyclopedia of Computational Chemistry*, John Wiley and Sons
14 Izrailev, S. and Agrafiotis, D.K. (2004) A method for quantifying and visualizing the diversity of QSAR models. *J. Mol. Graph. Model*. 22, 275–284
15 Fliri, A.F. *et al.* (2005) Biological spectra analysis: Linking biological activity profiles to molecular strucutres. *Proc. Natl. Acad. Sci. U. S. A*. 102, 261–266
16 Hopkins, A.L. and Groom, C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov*. 1, 727–730
17 Sconce, E.A. *et al*. (2005) The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen. *Blood* 106, 2329–2333
18 Michalovich, D. *et al*. (2002) Protein sequence analysis in silico: application of structure-based bioinformatics to genomic initiatives. *Curr. Opin. Pharmacol*. 2, 574–580
19 Beresford, A.P. *et al.* (2004) In silico prediction of ADME properties: are we making progress? *Curr. Opin. Drug Discov. Devel*. 7, 36–42
20 Butina, D. *et al.* (2002) Predicting ADME properties in silico: methods and models. *Drug Discov. Today* 7 (Suppl. 11), S83–S88
21 van de Waterbeemd, H. and Gifford E. (2003) ADMET in silico modeling: towards a predictive paradise? *Nat. Rev. Drug Discov*. 2, 192–204
22 Dickins, M. and van de Waterbeemd, H. (2004) Simulation models for drug disposition and drug interactions. *Drug Discov. Today: Biosilico* 2, 38-45
23 Anon. (2004) *Drug Development Science: Obstacles and Opportunities for Collaboration Among Academia, Industry and Government*. AAMC/FDA publication
24 Anon. (2004) *REACH in brief*. European Commission
25 Grimwade, D. and Haferlach, T. (2004) Gene expression profiling in acute myeloid leukaemia. *N. Engl. J. Med*. 350, 1676–1678
26 Beerenwinkel, N. *et al*. (2005) Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics* DOI:10.1093/bioinformatics/bti654

David Bousfield
73 de Freville Avenue,
Cambridge CB4 1HP,
UK
e-mail: david@ganesha-associates.com