drugs and other molecules from the literature. They variously reported specificity figures from 60–90%. The sensitivity of these systems remains an issue. Moreover, these performance studies were conducted on sample sets that were small in size and also different sample sets were used by different researchers. For an impartial assessment and comparison of the performance of these systems, as well as to understand what works and what does not, it is crucial to do a systematic evaluation. Such an evaluation should be based on a biologically important challenge problem, should have extensive training- and blind-test data, and should have a clear repeatable evaluation metric. To date, this issue seems to have started to catch the attention of researchers – the 'KDD Cup 2002' competitions (held in conjunction with ACM SIGKDD *International Conference on Knowledge Discovery and Datamining*) have included a task on 'Information Extraction from Biomedical Articles', jointly organized by Alex Yeh of MITRE (http://www.mitre.org) and the Flybase group of Harvard (see http://www.biostat.wisc.edu/~craven/kddcup/tasks.html).

- It is reasonable to assume that the completeness and reliability of the outcome of text-based knowledge discovery in biology are dependent on the input documents. Should the selection of input documents be based on keywords, as well as on papers chosen by expert biologists and well-cited articles and their cited references therein, or based on some other methods? To date, it seems that no single group has considered it in this context.
- It is also unclear how well a text-based knowledge discovery system has to perform for it to be useful in biology. To know how good a system has to be, working systems must be given to biologists in user-centered evaluations. To date, it appears that no single group has conducted such a study in any extensive way. We acknowledge, however, that, from experience with previous evaluations in the information retrieval community [6], it is hard to extrapolate from results of batch experiments to predict complex issues of use and user acceptance.

Many such issues have remained unaddressed to date. Nevertheless, text-based knowledge discovery for biology has significant potential, because even imperfect tools are useful if they give improved functionality at low cost.

References

1 Boston Consulting Group (2001) A revolution in R&D – the impact of genomics. *BCG Focus* June, 1–15

2 Hirschman, L. *et al.* Accomplishments and challenges in literature data mining for biology. *Bioinformatics* (in press)

3 Mack, R. and Hehenberger, M. (2002) Text-based knowledge discovery: Search and mining of life-sciences documents. *Drug Discov. Today*, 7 (Suppl. 11), S89–S98

4 Park, J.C. *et al.* (2001) Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *Proc. Pac. Symp. Biocomput.* 396–407

5 Putejovsky, J. and Castano, J.M. (2002) Robust relational parsing over biomedical literature: Extracting inhibit relations. *Proc. Pac. Symp. Biocomput.* 362–373

6 Hersh, W. *et al.* (2001) Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Inf. Process. Manag.* 37, 383–402

Limsoon Wong
Laboratories for Information Technology
21 Heng Mui Keng Terrace
Singapore 119613

Cheminformatics – decision making in drug discovery

Roger Lahana, Vice President R&D, Computational Drug Discovery, Synt:em, Parc Scientifique G. Besse, Nimes 30000, France;
tel: +33 466 048 668, fax: +33 466 048 667, e-mail: rlahana@syntem.com

Cambridge Healthtech Institute's 6th Annual *Cheminformatics* Conference was one of three conferences grouped together as *Intelligent Drug Discovery and Development 2002* (6–8 May 2002, Philadelphia, PA, USA). The other two conferences were dedicated to lead optimization and screening.

Building smarter libraries

Several aspects of library design were addressed in the first session, chaired by Roger Lahana (Syntem; http://www.syntem.com). Robert Sheridan (Merck Research Laboratories; www.merck.com) shared his experience on searching databases using chemical similarity and described three variations of tricks with chemical similarity that extend the applicability of similarity searching using topological descriptors. The first involves representing mixtures by their descriptor average, which works surprisingly well; the second concerns ways to derive nonpeptide molecules from an

1359-6446/02/$ – see front matter ©2002 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(02)02368-1

active peptide; and the third, LaSSI (latent semantic structural indexing), is an adaptation to chemical descriptors of a document analysis technique aimed at getting partial synonymy of words.

Michael Farnum (3D Pharmaceuticals; http://www.3dp.com) presented a novel optimization paradigm based on particle swarms, and its use in the design of combinatorial libraries. Particle swarms explore the search space through a population of individuals, which adapt by returning stochastically to previously successful regions, influenced by the success of their neighbours. Each particle corresponds to a potential solution of the optimization problem at hand; this method was compared with other stochastic optimization schemes.

Richard Cramer (Tripos; www.tripos.com) spoke about virtual screening using topomeric comparative molecular field analysis (CoMFA). Using topomerically aligned fragments of the input structures, robust starting points were obtained for automated studies, hence breaking the main input bottleneck of CoMFA. Good results were obtained in repeats of 15 published studies, thereby making this a general high-throughput design methodology, capable of shortening high-quality, information-based, design–test–synthesis cycles.

Does size matter?

How big should a library be to ensure that it will produce a significant hit rate? To address this question, Lahana revisited some key aspects of rational design, such as the strengths, weaknesses and complementarities of structure-based versus property-based design. Two case studies were used to demonstrate that it is not the size of the library that matters but the level of initial knowledge. Examples were taken from both peptide and organic molecule series.

The use of combinatorial optimization approaches to design focused, diverse and target-class-oriented libraries was presented by Weifan Zheng (Eli Lilly; http://www.lilly.com). Many computational techniques have been developed to address these issues, including monomer frequency analysis, multicriteria combinatorial optimization and schemes enhanced by heuristics. The goals are often to obtain diverse libraries as well as target-focused libraries. A recent trend is to design libraries loosely targeted against a target class, rather than against a particular target.

High-throughput conformational search

Paul Labute (Chemical Computing Group; http://www.chemcomp.com) presented a new methodology for the conformational search of large libraries of drug-like compounds. The key component of the method is an algorithm that is based on the charge of breaking each molecule down to overlapping substructures, for which a conformational analysis is performed and stored. If the library is homogeneous in terms of fragments, most molecules lead to previously found fragments. Therefore, corresponding data do not need to be re-calculated but simply retrieved from the database.

Virtual screening

Paul Lyne (AstraZeneca; http://www.astrazeneca.com) opened the second session with an evaluation of several virtual screening approaches, both structure-based and ligand-based, against several targets. Several docking algorithms were compared on the basis of 180 different complexes from public databases. Recent products seemed to behave reasonably well in reproducing the actual structures with an RMS of <2 Å. However, the performance was reduced when using internal data; in this case, the known structures of five proteins, each with 10 ligands.

For Ingo Muegge (Bayer; http://www.bayer.com), virtual screening is a cost-effective alternative for lead identification, especially for smaller pharmaceutical companies that cannot invest in large-scale HTS. Concepts of virtual screening, including structure- and pharmacophore-based library design, as well as small-molecule property-based approaches were discussed.

Novel methods, recurrent problems

An alternative method for scoring function creation was presented by Andrew Good (BMS; http://www.bms.com). All current development on scoring functions derives from known crystal structures, thus providing information on why things bind, but little on why they do not. The method presented attempts to combine structural information with data derived from within structure-based virtual screening calculations. By applying such a strategy, together with pharmacophore constraints, search speed and hit rates can be improved.

In a complementary frame of mind, Diane Joseph-McCarthy (Wyeth Research; http://www.wyeth.com) described a pharmacophore-based docking method in which conformers of the same or different molecules are superimposed and simultaneously docked by their largest 3D pharmacophore. Using this approach, target-derived theoretical pharmacophore points were directly matched to the database pharmacaphores, improving enrichment rates and database generation methods.

Exploiting drug discovery data integration

An integrated cheminformatic environment has been developed by Michael Lajiness at Pharmacia (http://www.pharmacia.com). The functionality of this environment includes the ability to identify structure–activity discontinuities, as well as other features, such as filter hits and clustering by structural classes. It was decided to develop and customize a proprietary system rather than acquire commercial tools – a rather unusual route in big pharma.

The opposite view was presented by Frank Brown (RW Johnson; http://www.rwjpri.com) who favoured the use of off-the-shelf products over developing (and supporting) software. Cheminformatics was defined as the bringing together of data, algorithms and interfaces to assist discovery scientists in their decisions about what to test and make next. Several applications illustrated the efficient use of a combination of standard tools, such as RS3, Diva, Tsar, MOE and Oracle.

Evolutionary chemistry

Evolutionary breeding methods for small molecules were introduced by Lutz Weber (Morphochem; http://www.morphochem.com). These methods have been integrated into a software called MolMind™, which makes use of heuristic evolutionary methods for synthesis planning. It works with Morphochem's proprietary database of multiple component reactions and available starting materials that have been implemented on their robotic systems. MolMind™ is used to discover new multicomponent reactions that enable the synthesis of natural-product-like molecules, as well as the design of compound libraries around any lead structure of interest.

Alexander Tropsha (University of North Carolina; http://www.unc.edu) questioned whether traditional QSAR-model quality techniques, such as leave-one-out cross-validated R^2, could be used as proper indicators of the predictive ability of the models. He proposed novel parameters that characterize the statistical significance of QSAR models in a more rigorous way and their use in screening chemical databases and virtual databases for potentially bioactive molecules.

Requirements for data integration in pharma and biotech R&D present highly complex challenges, as described by Herschel Weintraub (CADDinformatics). A wealth of data, including chemical structures, genome sequences, screening results, Laboratory Information Management Systems (LIMS) data, ADME/Tox data and cheminformatics data, is continuously being collected. These data must be accessible to the research scientists and management in a timely and efficient manner. A key issue is the integrity of the systems that contain validated data and approaches to undertake such a massive challenge to data integration were presented.

Concluding remarks

This conference was rich in interesting talks, highlighting novel ideas and useful views. Expressions such as 'unique proprietary technology' and 'paradigm shift' have been seldom heard and this gave much more weight and interest to all the new ideas that were presented and debated. Several speakers described unsatisfactory or unexpected results as well as successes, stressing the fact that a failure is not necessarily something negative. Indeed, more and more scientists now seem ready to share their experience, even when things do not work as expected. After several years of staggering claims, the age of reason (and truth) seems to have been reached.

New routes for drug discovery

Kewal K. Jain, Bläsiring 7, CH-4057 Basel, Switzerland; tel/fax: +41 61 692 4461, e-mail: jain@pharmabiotech.ch

The recent IBC conference *Proteomics: Delivering New Routes to Drug Discovery* (6–9 May 2002, Philadelphia, PA, USA) continued its tradition of providing a broad coverage of the application of innovative proteomic technologies for drug discovery and clinical applications. Selected presentations are described here briefly. Updated proteomic technologies are described elsewhere [1].

Beyond the genome to clinical proteomic diagnostics

In the keynote address, Emanuel Petricoin (FDA–NCI Clinical Proteomics Program; http://www.fda.gov) described the paradigm shift in molecular medicine as the choice of therapy by category of disease, molecular profiling of tissue cells and therapy tailored to individual patients.

Proteomics has an important role in these concepts as it is not a mere catalogue of proteins but represents cellular network and signal pathways. Cancer is considered to be a proteomic disease at the cellular level and cancer profiles are made from laser capture microdissection (LCM) of cancer tissue taken from patients. The FDA program uses low-profile proteomic technologies for diagnosis and drug discovery in cancer and have justified a prospective population-based assessment of proteomic pattern technology as a screening tool for all stages of ovarian cancer in high-risk and general populations [2]. A five-year follow-up of this study showed that the accuracy rate of diagnosis of carcinoma of cancer was 99%.

David Misek (University of Michigan; http://www.umich.edu) described the proteomic technologies for cancer marker identification with particular application to lung cancer. Major advantages of cancer marker identification are the ability to profile tumor tissue and uncover a variety of markers: overexpressed proteins, post-translationally modified

1359-6446/02/$ – see front matter ©2002 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(02)02354-1