'Managing large compound collections and handling massive amounts of screening data would simply not be possible without well-designed informatics infrastructures'

editorial

Christian N. Parker
Caroline E. Shamu
Brian Kraybill
Christopher P. Austin
Jürgen Bajorath*

Measure, mine, model, and manipulate: the future for HTS and chemoinformatics?

Chemoinformatics has developed into a multifaceted research area [1] with significant potential to interface with experimental disciplines including, among others, analytical chemistry, combinatorial chemistry, medicinal chemistry and HTS. In a recent Editorial in *Drug Discovery Today* [2], Ricardo Macarron provided a well-balanced perspective on the role of HTS in drug discovery – its promises, accomplishments and caveats. Here, we expand on this theme by discussing another aspect that is highly relevant to the current and future potential of compound screening – the interface between HTS and chemoinformatics. There are many reasons to address this issue including a very practical one; with the introduction of the McMaster University Data Mining and Docking Competition [3,4] a forum has been established to bring experimental and virtual screeners together and systematically assess the predictive performance of computational methods. The McMaster competition was held for the first time in 2004 and now goes into its second round, providing an opportunity to introduce it to a wider audience. It is hoped that this will help establish this initiative on a long-term basis. Whether or not the term 'competition' is the most meaningful designation for this event is probably a matter for debate. Ultimately, there are no losers, because all the contributors help to further advance the field by allowing unbiased analyses of experimental and theoretical protocols and the prospective testing of predictions, just as in many other areas of science.

Synergies and road blocks

Managing large compound collections and handling massive amounts of screening data would simply not be possible without well-designed informatics infrastructures. Moreover, many recently developed methods for molecular similarity analysis and chemical database mining have a high degree of complementarity to experimental screening approaches [5]. These advances suggest that iterative experimental and computational screening campaigns have the potential to contribute to the discovery process, possibly even reducing the resources needed for success [6]. Integration of experimental and computational efforts is feasible at many different levels including library design, compound prioritization and de-prioritization, or data analysis. In fact, experimental and computational approaches are often concerned with similar questions. For example: how to omit flawed compounds from consideration and reduce false-positive rates; how to balance chemical diversity and target focus; and how to ensure probe-, lead- or drug-likeness of candidate molecules. However, as we have pointed out elsewhere [7], the integration of experimental and computational methods is currently much less advanced than one might think (or perhaps wish for) despite these obvious synergies. Clearly, there are some technical hurdles to overcome, but equally, if not more, relevant for the current situation are differences in mind sets between experimental and computational screeners that are, at least in part, supported by different reward structures. Scientists in the HTS area are typically

1359-6446/06/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2006.08.006

challenged with increasing the efficiency of the screening process and maximizing the throughput of their system. By contrast, the primary goal of virtual screeners is to rationalize compound selection and select small numbers of compounds for further investigation. Given this situation, initiatives like the McMaster competition play a role that goes well beyond the comparative evaluation of computational methods and predictions; they also provide a forum for a dialogue between scientists from different disciplines and a basis for a meaningful integration of experimental and computational strategies.

Lessons learned from the first McMaster competition

The first event was made possible because HTS datasets for deriving computational models and making predictions were made freely available by the screening group at McMaster University, directed by Eric Brown. As an established and well-described drug target, dihydrofolate reductase (DHFR) was chosen to allow the application of both ligand- and structure-based virtual screening methods. The primary objective of the study was simple: given the HTS results for a learning set of 50,000 compounds, predict active compounds from a testset of another 50,000 small molecules. A secondary objective of the study, only made possible by the enthusiastic support of the McMaster group, was to identify and re-test possible false-negatives identified by computational analyses (that is, compounds predicted to be active that were not HTS hits).

The scenario was complicated by two unexpected findings. First, the compound libraries for learning and testing displayed significant differences in chemical composition and compound diversity. Second, screening the test compounds did not yield any validated competitive inhibitors [8]. The first complication made it very difficult for data mining approaches to develop models of predictive value. Only a few of the participating groups recognized the discrepancies between the compound datasets and pointed them out in their evaluation. By contrast, docking methods also predicted many active compounds, although they would not be expected to suffer from such discrepancies between compound learning and testsets owing to their primary focus on the DHFR structure. Thus, high false-positive rates emerged as a major limitation of virtual screening, regardless of the strategies applied. The second complication not only nicely illustrated the difficulties associated with validating primary HTS hits but also made judging the competition a rather difficult task so that only general trends could be evaluated [4]. In summary, the outcome of the first McMaster screening competition was far from optimal, but it did highlight the need for further advanced methods for validating models and evaluating their appropriateness for different testsets. Furthermore, the event provided several valuable scientific insights for the screening field and a real life HTS dataset for further chemoinformatics analyses. It also provided a basis for setting up the next competition.

Outline and objectives of the second event

The second competition is being organized in association with the Society for Biomolecular Sciences and has expanded to offer three datasets for analysis and prediction. However, the objectives of the study remain the same: first, to introduce the work done in this field to a larger audience, helping to bring screeners and computational modelers together in a common forum; second, to evaluate the success, or failure, of diverse methods applied to the problems presented; and, third, to highlight areas where significant improvements could be made to current methods for modeling and predicting HTS data.

The first dataset has been supplied by researchers from the NIH Chemical Genomics Center (NCGC) and is the result of a biochemical assay searching for modulators (inhibitors or activators) of *Bacillus stearothermophilus* pyruvate kinase. The training set consists of just over 51,000 compounds tested at at least seven different concentrations within the range 4 nM to 57 μM, producing concentration–response data for all compounds in the primary screen (termed quantitative HTS, or qHTS, by the NCGC). For this assay there is a short timeline to complete the analysis and prediction of the activity of the 21,000 testset compounds, because the results of the screen have just been published [9] – note that similar time constraints apply to many practical HTS applications. The submission deadline for this dataset will therefore be 1 September 2006, because all screening results reported in the paper will be made publicly available in PubChem on that date (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=361). The second dataset, also from the NCGC, is the result of a qHTS screen for inhibitors or activators of human glucocerebrosidase (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=360). For this assay, the training set consists of concentration–response data on 48,125 compounds, and the testset consists of ~8700 compounds, to be used for prediction. A crystal structure is available for glucocerebrosidase and pyruvate kinase crystal structures are available from related species, for use in docking applications.

The third and final dataset represents a cell-based screen carried out at the ICCB-Longwood Screening Facility at the Harvard Medical School to identify inhibitors of the type III secretion pathway of the human pathogen *Pseudomonas aeruginosa*. The assay identified hits in a screen of >100,000 compounds. Some of the hit compounds have been shown to inhibit the activity of the *P. aeruginosa* phospholipase A enzyme ExoU, but other hits could potentially interact with every component of this biological system. Thus, one of the challenges here will be to identify target-specific compounds. Also, the nature of this dataset precludes the exclusive use of docking methods. The training set for this screen consists of data for >66,000 compounds. The data will be presented with the plate and layout information as well as the primary reader data. This will allow the modelers to test normalization and data modeling algorithms as well as their prediction methods against the testset of >65,000 compounds (the datasets are available through these links: http://www.sbsonline.org/datamining/ and http://ncgc.nih.gov/pub/ncgcsbs/).

Although submission of the pyruvate kinase predictions closes 1 September 2006, the submission dates for the other datasets are in September 2006. We hope that many groups will participate and help to further advance this field and the integration of biological and computational screening efforts.

References

1 Bajorath, J. (2004) Understanding chemoinformatics: a unifying approach. *Drug Discov. Today* 9, 13–14

2 Macarron, R. (2006) Critical review of HTS in drug discovery. *Drug Discov. Today* 11, 277–279

3 Parker, C.N. (2005) McMaster University data-mining and docking competition: computational models on the catwalk. *J. Biomol. Screen.* 10, 647–648

4 Lang, P.T. *et al.* (2005) Evaluating the high-throughput screening computations. *J. Biomol. Screen.* 10, 649–652

5 Stahura, F.L. and Bajorath, J. (2004) Virtual screening methods that complement HTS. *Comb. Chem. High Throughput Screen.* 7, 259–269

6 Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1, 882–894

7 Parker, C.N. and Bajorath, J. Towards unified compound screening strategies: a critical evaluation of error sources in experimental and virtual high-throughput screening. *QSAR Combinat. Sci., Special Issue on Challenges in Virtual Screening* (in press)

8 Elowe, N.H. *et al.* (2005) Experimental screening of dihydrofolate reductase yields a ‘‘test set’’ of 50,000 small molecules for a computational data-mining and docking competition. *J. Biomol. Screen.* 10, 653–657

9 Inglese, J. *et al.* (2006) Quantitative high-throughput screening (qHTS): a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U. S. A.* 103, 11473–11478

Christian N. Parker
Novartis Pharma AG, WSJ-088.2.06, CH-4002 Basel, Switzerland

Caroline E. Shamu, Brian Kraybill
Harvard Medical School, ICCB-Longwood, Seeley G. Mudd – 604, Boston, MA, USA

Christopher P. Austin
National Institutes of Health, National Human Genome Research Institute, NIH Chemical Genomics Center, 9800 Medical Center Drive, MSC: 3370, Bethesda, MD 20892-3370, USA

Jürgen Bajorath*
Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany
bajorath@bit.uni-bonn.de