JOURNAL OF CHEMICAL INFORMATION AND MODELING

EDITORIAL

pubs.acs.org/jcim

A Call to Arms: What You Can Do for Computational Drug Discovery

This issue of the *Journal of Chemical Information and Modeling* has a section devoted to the 2010 Scoring Exercise organized by the Community Structure−Activity Resource (CSAR) at the University of Michigan, Ann Arbor. Sixteen groups—representing academia, pharma, and software developers—contributed scores for the exercise. Some provided papers for this issue; others gave talks at the symposium (240th ACS National Meeting, Boston, MA, August 22−28, 2010), and a few only submitted scores. Also, many people have contributed excellent suggestions and valuable feedback on various issues. This has been a rewarding experience, and we greatly appreciate everyone's participation. We particularly thank CSAR's Advisory Board and the National Institute of General Medical Science (U01 GM086873) for their support and guidance. We also thank Editor William L. Jorgensen and the staff of the *Journal of Chemical Information and Modeling* for their significant efforts to make this issue possible.

CSAR is a platform for computational chemists to work together to push our field forward. Our goal is to provide the public with data sets of the highest quality to create a firm foundation for method development and testing. We promote collaboration over competition, and we ask you to join us in this effort with the following activities.

PARTICIPATE IN THE CSAR EXERCISES

The 2010 exercise involved the difficult task of scoring diverse ligands bound to diverse proteins based on crystal structures from the PDB. Our next exercise will be based on blinded data donated from the pharmaceutical industry, and participants will be asked to tackle the more practical task of docking and scoring congeneric series of ligands against a limited set of protein targets. That exercise is slated to start in late 2011, and the exact timeline will depend on when final legal approval is given to release the data to the public.

Poorly designed exercises are run like contests and do little to improve our field. They focus solely on who wins, and frequently, the outcome is heavily influenced by instinctual knowledge. Someone who has a "good eye" can usually discern proper docked poses better than most methods can with scoring. We strive for more concrete insights from CSAR's exercises. Participants are asked to turn in two or more sets of scores/poses, each generated with alternative approaches. The goal is to have participants use the exercise to test hypotheses and share those outcomes with the entire community. The power of testing approaches in this manner will be strongest using blinded data. There are no limits to the options a participant might compare: (1) Does the inclusion of water molecules in the binding site improve correlation to experiment?; (2) does minimizing the protein−ligand structures after docking help?; (3) which treatment of electrostatics is better?; (4) does a particular desolvation metric improve results?; etc. The approaches will be evaluated by their enrichments in virtual screening, the prediction of crystal-structure poses from docking, and the correlation between scores and experimental binding affinities.

DONATE DATA

The ideal data sets have ∼50 active compounds and ∼10 chemically similar inactives. The actives should be congeneric series that span at least three orders of magnitude in binding affinity (preferably spanning the critical range of low nM to mid μM), and roughly a dozen should have bound crystal structures available. The actives should have a range of sizes, rotatable bonds, and hydrogen-bonding features. Of course, different protein targets will have different limits for ligand properties, but obtaining the widest range possible for each target is desired. We work closely with depositors to help with the analysis and the selection of ligand data.

The desired breadth and depth of unpublished data spurred us to pursue donations from the pharmaceutical industry (though it may be possible to obtain this type of data from academic researchers or services centers). We are asking for "old" data, not necessarily the latest hot target. We need data on well-established systems with high-quality crystal structures with well-resolved ligands. Data from abandoned projects or from the development of successful drugs on the market (well-covered intellectual property) tend to be easier to obtain. CSAR currently has confidentiality agreements with Abbott, GlaxoSmithKline, and Genentech; Roche has donated crystal structures without a contract. We are negotiating with several more companies to join the effort. After establishing a confidentiality agreement, we can visit the site and help with curating the data, which significantly reduces the burden on our pharma colleagues who have little spare time for side projects. We also have crystallographers to help complete structure refinement and eventually submit the structures to the PDB.

Why might a company want to release its proprietary data? At this time, the greatest limitation of structure-based drug discovery (SBDD) affects pharma every day; it is the inability to distinguish tight binding from moderate binding (nM- vs μM-level inhibitors). If this were overcome, then SBDD would identify candidate compounds over mere leads and significantly reduce false positives. Also, more accurate modeling takes us one step closer to predicting selectivity, toxicity, and druggability. Pharma has the data to help solve the problem but not necessarily the time and the staff to develop the new tools. Sharing that information effectively outsources the method development, which benefits the company in the form of better tools from academics and software companies. The company can also increase its visibility in the field by having its data form the basis of an exercise. For old data, companies have nothing to lose and so much to gain.

Special Issue: CSAR 2010 Scoring Exercise

Published: September 26, 2011

USE CSAR DATA IN METHOD DEVELOPMENT AND ASSESSMENT

Exceptional data are fundamentally much more than just test sets for exercises. They provide benchmarks that can be used as standards for method development. Over the course of an exercise, the data are processed and analyzed by many groups: the donators, CSAR, and every subsequent user. This creates an exceptionally stringent curation process, a trial-by-fire that continually improves the set as more feedback is received from the community.

CSAR is all about data: better data, more data, and specific data to solve SBDD problems. Our predictions will always be limited by the uncertainty in the data used to develop our methods. We must have confidence in its accuracy. We aim to keep the error as small as possible and to describe how the error can impact the subsequent use of the data. The data must be complete and diverse so that it best represents the proper chemical space of a particular congeneric series for a target. Furthermore, negative data for chemically similar inactives are absolutely essential to bound the problem. Lastly, and most importantly, the data must help us address significant deficiencies in our field. We emphasize a wide range of hydrogen-bonding characteristics and number of rotatable bonds for ligands because we found that these features were associated with difficult systems in the 2010 scoring exercise.

DREAM BIG

The ultimate evolution of SBDD techniques would provide methods that accurately design nM- and pM-level inhibitors based on their binding sites, rather than docking and ranking libraries of compounds. If SBDD were perfect, then large libraries of in-house compounds would not be needed. SBDD software would accurately design candidate compounds, and only a handful of compounds would be synthesized to move forward with pharmacological and toxicological studies. It will take a lot of work to get there, and it is impossible unless we work together. The future is in your hands.

Heather A. Carlson* and James B. Dunbar, Jr.

Department of Medicinal Chemistry, University of Michigan, 428 Church Street, Ann Arbor, Michigan 48109-1065, United States

AUTHOR INFORMATION

Corresponding Author

*E-mail: carlsonh@umich.edu. Telephone: (734) 615-6841.