J Comput Aided Mol Des (2012) 26:69–72
DOI 10.1007/s10822-011-9515-0

PERSPECTIVE

Computation, experiment and molecular design

Peter W. Kenny

Received: 29 November 2011 / Accepted: 2 December 2011 / Published online: 11 December 2011

Perhaps the best way to start a Perspective like this is to say what we mean by 'Molecular Design' and one definition is control of properties and behaviour of compounds and materials through manipulation of molecular properties [1]. Molecular interactions [2] occupy a special place in the framework of Molecular Design because the functional behaviour of a compound is determined by the strengths of the interactions of its molecules with the different environments (e.g. solution; crystal lattice; bound to protein) in which they exist [1]. Molecular Design can be hypothesis-driven or prediction-driven and the objectives of Computer-Aided Molecular Design (CAMD) are to provide physicochemical insight for framing design hypotheses and to build predictive models with which to evaluate compounds. Energy, geometry and connectivity form the basis of the tools used in CAMD. For example, a method for predicting crystal structure might use geometry to pack molecules into prospective unit cells and energy to rank the resulting crystal lattices. The molecular connection table is the defining cheminformatic data structure and we can use connectivity to infer physicochemical characteristics such as interaction potential (e.g. strong hydrogen bond acceptor) and likelihood of being ionised in solution.

From the CAMD perspective, Drug Discovery is a process in two steps. Lead Identification is a search for active compounds that can be optimised in the second step to compounds with properties and behaviour consistent with dosing in humans as therapeutic agents. Although the second step is frequently described as a process of multi-objective optimisation, it can be argued that lead optimisation actually has minimisation of therapeutic dose as its principal, if not sole, objective. The last 25 years have seen adoption of a number of technologies by pharmaceutical and biotechnology industries and much CAMD activity is a reaction to these developments. Typically, each new technology is introduced with the promise that it will revolutionise Drug Discovery and much of the hyperbole spills over into the CAMD arena. Some of the difficulty that CAMD scientists face in gaining acceptance for their approaches can be traced to extravagant claims made earlier by other CAMD scientists. Assessing the value added by a new technology is not always easy. When a pharmaceutical company spends a large amount of money to acquire new capability, it is in the interests of both vendor and customer that purchases and collaborations are seen in the most favourable light. Over-selling of technologies leads to panacea-centric thinking, which is especially dangerous in CAMD because success frequently depends on bringing together diverse computational tools to both define and solve problems. One important lesson from the last 25 years of Drug Discovery is that technology is a good servant but a poor master.

The software tools of CAMD are in many cases interconnected and the CAMD scientist must recognise these connections in order to exploit them. I will attempt to illustrate this point with reference to computational methods for identifying compounds with interesting biological activity, which has become known as virtual screening. Pharmacophore searching, molecular shape-matching and docking are essentially geometry-based approaches to virtual screening. The first two of these are ligand-based and complement each other in virtual screening campaigns. Despite algorithmic differences, pharmacophore searching and shape-matching are actually more similar than first appearances might suggest in that each method aligns hit molecules with the respective query. We can rank

P. W. Kenny (✉)
Macclesfield, UK
e-mail: pwk.pub.2008@gmail.com

structures aligned with a pharmacophore by overlap and excluded volumes with respect to the overlay of structures from which the pharmacophore has been abstracted. Analogously, shape-matching hits can be prioritised according to whether or not they are able to correctly position molecular recognition elements (e.g. cation) that are conserved in active compounds. Pharmacophoric criteria can also be used to select docked poses that display features of experimentally determined binding modes. All three virtual screening approaches require 3D structures for potential ligands and the software for generating 3D coordinates (and in some cases conformations) is an integral, although often invisible, part of virtual screening. Connectivity-based cheminformatic tools are used both to organise the output of virtual screens (e.g. by chemotype) and to prepare the databases (e.g. setting ionisation and tautomeric states). Often these will be the same tools used for analysis of high throughput screening results and for compound library design. For example, one can use the same fingerprint-based measure of molecular similarity both to eliminate virtual screening hits that are too similar to what is already available in house and to characterise neighbourhoods around high throughput screening hits.

As noted ('Corpora non agunt nisi fixate') by Ehrlich a century or more ago, a drug must bind in order that its effects be observed and affinity prediction is of great interest in CAMD. Prediction of affinity is difficult, even when a high-resolution structure of the target is available, primarily because the association of drug with target occurs in aqueous media and, in some cases, because the target is flexible. The cohesiveness of liquid water that drives hydrophobic association in aqueous media is a consequence of strong, cooperative hydrogen bonds between water molecules and I like to think of the hydrophobic force as a non-local, indirect, electrostatic interaction. While a limited number of solvation free energy measurements are available for small molecules, we know much less about solvation of proteins and, crucially, the deep, narrow pockets on the protein surface that serve as hot spots for ligand recognition. Although simulation methods are used to compute binding free energies, the calculations are computationally demanding and typically can only be used to address the effects of conservative structural changes on affinity. Other approaches to affinity prediction include molecular mechanics with implicit solvation and summation of contributions from protein–ligand contacts in the complex, although results to date have been disappointing and neither approach has been shown to represent a general solution to the problem. The contribution to affinity of a particular interaction, such as a hydrogen bond, is not strictly an experimentally observable quantity although this does not prevent assertions that a drug is getting specificity from hydrogen bonds and affinity from hydrophobic interactions. There is no shortage of Voodoo Thermodynamics in the literature of Drug Discovery.

No discussion of CAMD in pharmaceutical research would be complete without reference to Quantitative Structure Activity/Property Relationships (QSAR/QSPR). The predictive model of 2012 is typically built using descriptors selected from a large pool [3] and over-fitting [4] is an ever-present concern. Model building and validation work best when compounds are evenly distributed in the space spanned by training and test sets. When large clusters are present in the data, validation can give an optimistic view of model quality, which can lead to over-fitting, and models described as 'global' may in fact be ensembles of local models. Assessing quality and scope of published QSAR/QSPR models can be difficult since data and molecular structures (and in some cases even the models themselves) are often not made available. QSAR/QSPR models represent a data-hungry way to do molecular design and the danger is that projects will have moved on before enough measured data becomes available for modelling. Some screening assays sacrifice dynamic range for throughput and making full use of these measurements may require new modelling approaches that can handle mixed continuous and categorical data. In the future, we can expect to see more prediction to be based on relationships between structures. For example, we might test the hypothesis that chloro-substitution is bad for aqueous solubility by searching for pairs of molecules in the appropriate database for which the only difference is the chloro-substitution. The assumption underlying this approach, which has been termed Matched Molecular Pair Analysis [5], is that it is easier to predict differences in the values of a property than it is to predict the value of the property itself. Observing an effect across multiple chemotypes increases our confidence that the effect is indeed real. Although relationships between structures can be used predictively, they are also an integral part of representing activity landscapes [6].

We use experimental measurements to parameterise, train and validate the majority of the models used in CAMD. Sometimes the measurements are of properties such as affinity and solubility that are directly relevant to Drug Discovery but, in other cases, the properties are primarily of interest for building physical models of molecular recognition. The change in free energy associated with transfer of a compound from gas phase to a solvent such as water is the most direct measure of solvation. However, these are difficult measurements to make and are simply not feasible for many compounds. One challenge to those developing solvent models would be to consider how better use could be made of more accessible measurements, such as partition coefficients and dissociation constants, that are less directly related to solvation. Equilibrium constants measured in non-aqueous solvents also yield information that is useful for model building. For example, correlations of hydrogen bond acidity and basicity with molecular electrostatic potential are useful in assessing different schemes for distributing charge in molecules. Given that accurate prediction of affinity is the ultimate aim of much CAMD, measured affinities that are linked to crystal structures for protein–ligand complexes provide, at very least, the means to validate models. Isothermal titration calorimetry is a particularly direct method for measuring affinity and changes in both enthalpy and entropy associated with binding can be quantified. Although some researchers assert the desirability of binding that is enthalpy-driven, it is not clear how isothermal physiological systems sense this and one challenge to CAMD is how best to exploit measurements of changes in enthalpy, entropy and volume associated with binding. Thermodynamic measurements would be especially valuable if the associated crystal structures showed destabilising features in the bound complex such as strained conformations or contacts between polar and non-polar surface.

So what does gazing into the crystal ball reveal? Future Drug Discovery scientists will have a lot more data, some of which may be relevant to Drug Discovery, and much faster computers at their disposal. Force fields and implicit solvation treatments should become more accurate and it will become feasible to study larger systems using ab initio and density functional electronic structure tools, leaving less room for semi-empirical molecular orbital methods. Improved force fields and more computational power will allow a more realistic treatment of conformational flexibility of proteins and this should lead to more efficient virtual screening and more accurate affinity prediction. Quantum mechanical methods will be used routinely to an increasing extent for assessing molecules and the descriptors used in QSAR/QSPR studies will become both more physical and less numerous. We will see an increasing amount of prediction being based on relationships between structures. The emergence of Fragment-Based Drug Discovery is likely to renew interest in both De Novo Structure-Based Design and techniques for using molecular probes to map surfaces of proteins. I could go on but instead I will take a more detailed look at alkane/water partition coefficients which have been largely ignored by the CAMD community.

Much has been written about the importance of appropriate physicochemical properties and the most important of these is lipophilicity, which is usually quantified as the logarithm of the octanol/water partition coefficient ($\log P_{oct}$). Most medicinal chemists will be aware of the celebrated Rule of 5 (Ro5) which is essentially a statement of property distributions for drugs [7]. However, Ro5 also raises questions as well as answering them. Why is Ro5's high polarity limit defined in terms of numbers of hydrogen bond donors and acceptors rather than $\log P_{oct}$? What is the origin of Ro5's asymmetry with respect to hydrogen bond donors and acceptors? In CAMD we use $\log P_{oct}$ as to quantify the ease or difficulty of moving the drug from an aqueous environment to a hydrophobic binding pocket or the interior of a lipid bilayer. Excessive lipophilicity is seen as the root cause of many evils such as promiscuity although the relevant correlations are not always as strong as creative data analysis may make them appear to be. Octanol has a hydroxyl group, which in the context of a shake flask experiment, causes it to become quite wet (2 M water) and the octanol/water partitioning system does not appear to 'see' hydrogen bond donors [8]. The logP of phenol is 1.5 for octanol/water but −0.8 for hydrocarbon/water. Which of these two figures is more relevant if you are assessing the difficulty of burying a phenolic hydroxyl in a hydrophobic pocket? One reason that we do not make more use of alkane/water logP ($\log P_{alk}$) is that the measurements are more difficult but CAMD scientists should not let that blind them to the limitations of octanol/water as a partitioning system. For the foreseeable future, it will simply not be technically feasible to measure $\log P_{alk}$ for many compounds and general access to this property will require the development of predictive models [9]. However, building those models will require high quality measurements of $\log P_{alk}$ for compounds for which measurement is feasible.

This Perspective was written for the 25th anniversary edition of the Journal of Computer-Aided Molecular Design so it is appropriate to conclude with a look to the future of publishing in the CAMD field. As an observer, one can get the general impression that the scientific publishing industry is headed for a Malthusian catastrophe as too many journals pursue too little content of genuinely high quality. One of the primary objectives of a specialist scientific journal should be to shape the scientific debate that drives its field forward and the relevance of Impact Factor in this context is far from clear. What is clear, however, is that scientific journals now share an ecosystem with bloggers and online discussion groups and I direct the reader to a cautionary tale for journal editors [10]. To remain relevant, scientific journals will need to be prepared to publish negative results and articles that are critical of other work such as this survey [11] of optical biosensor literature. Although CAMD is the focus of this journal, it would benefit from being seen as place for sharing high quality experimental measurements. Ultimately, a scientific journal needs to create an environment in which articles are discussed publicly and the authors of those articles are active participants in the discussion.

References

1. Kenny PW (2009) J Chem Inf Model 49:1234–1244. doi:10.1021/ci9000234
2. Bissantz C, Kuhn B, Stahl M (2010) J Med Chem 53:5061–5084. doi:10.1021/jm100112j
3. Doweyko AM (2008) J Comput Aided Mol Des 22:81–89. doi:10.1007/s10822-007-9162-7
4. Hawkins DM (2004) J Chem Inf Comput Sci 44:1–12. doi:10.1021/ci0342472
5. Kenny PW, Sadowski J (2005) Structure modification in chemical databases. Methods and principles in medicinal chemistry. In: Oprea T (ed) Chemoinformatics in drug discovery, vol 23, pp 271–285. doi:10.1002/3527603743.ch11
6. Wassermann AM, Wawer M, Bajorath J (2010) J Med Chem 53:8209–8223. doi:10.1021/jm100933w
7. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Adv Drug Deliv Rev 23:3–25. doi:10.1016/S0169-409X(96)00423-1
8. Lipophilicity teaser. http://fbdd-lit.blogspot.com/2011/06/lipophilicity-teaser.html
9. Toulmin A, Wood JM, Kenny PW (2008) J Med Chem 51:3720–3730. doi:10.1021/jm701549s
10. A short rant about journal editors. http://fbdd-lit.blogspot.com/2011/04/short-rant-about-journal-editors.html
11. Rich R, Myszka D (2010) J Mol Recognit 23:1–64. doi:10.1002/jmr.1004