J Comput Aided Mol Des (2012) 26:121–124
DOI 10.1007/s10822-011-9512-3

PERSPECTIVE

Can we really do computer-aided drug design?

Matthew Segall

Received: 15 November 2011 / Accepted: 2 December 2011 / Published online: 11 December 2011

Abstract In this article, we discuss what we mean by 'design' and contrast this with the application of computational methods in drug discovery. We suggest that the predictivity of the computational models currently applied in drug discovery is not yet sufficient to permit a true design paradigm, as demonstrated by the large number of compounds that must currently be synthesised and tested to identify a successful drug. However, despite the uncertainties in predictions, computational methods have enormous potential value in narrowing the range of compounds to consider, by eliminating those that have negligible chance of being a successful drug, while focussing efforts on chemistries with the best likelihood of success. Applied appropriately, computational approaches can support decision-makers in achieving multi-parameter optimisation to guide the selection and design of compounds with the best chance of achieving an appropriate balance of properties for a drug discovery project's objectives. Finally, we consider some approaches that may contribute over the next 25 years to improve the accuracy and transferability of computational models in drug discovery and move towards a genuine design process.

Keywords Drug discovery · Quantitative structure activity relationship · Multi-parameter optimisation · Quantum mechanics

Introduction

The title may seem like a strange question in a special issue celebrating 25 years of the Journal for Computer-Aided Molecular Design. However, it is an important question to address and to do so we first need to define 'design'. The Collins Dictionary defines the verb 'design' as "to work out the structure or form of (something), by making a sketch or plans". In a design process representations of an object are created, perhaps on paper or, more commonly, on a computer. From these representations the characteristics of the different designs are calculated, on the basis of which a small number of prototypes may be chosen and built before identifying the final object.

Beresford et al. [1] contrasted the design of the Boeing 777, which was designed entirely on a computer before a single prototype was built and successfully flown, with the drug *discovery* process in which it is common to test thousands of compounds prior to selection of a development candidate [2]. Even after selection of a candidate compound, the chance of this compound reaching the market is less than 10% [3]. This does not appear to be consistent with the earlier definition of design.

We would argue that a claim to be doing drug design is not justified because the predictive models currently used in drug discovery do not yet have sufficient accuracy. It is not possible to identify a small number of compounds and say with confidence that at least one will achieve the requirements for a good drug, including sufficient target potency, good physicochemical, absorption, distribution, metabolism, and elimination (ADME) properties and appropriate safety.

For example, there are many ways to seek to identify a potent compound against a therapeutic target [4]. Similarity search techniques, which use 2- or 3-dimensional molecular structure to search for compounds that are similar to a known active, achieve a 'hit rate' of only approximately 20–30% for the most similar compounds [5]. Techniques which use the 3-dimensional structure of the target protein, such as docking,

M. Segall (✉)
Optibrium Ltd, 7226 IQ Cambridge, Beach Drive,
Cambridge CB25 9TL, UK
e-mail: matt.segall@optibrium.com

achieve a similar hit rate among the highest scoring compounds [6]. Quantitative structure activity relationship (QSAR) methods, which predict numerical values of the potency of a compound, typically have an uncertainty of 0.5–1.0 log units unless working within a tight congeneric series for which detailed data are already available.

Predictions of physicochemical and other properties have similar levels of uncertainty; for example, in a study of predictive models of aqueous solubility, Dearden found that the standard errors of 17 commercial models for aqueous solubility ranged from 0.47 to 1.96 log units on a set of 122 drugs [7].

Furthermore, these individual molecular properties must be combined to assess the ultimate efficacy, disposition and safety of a compound in vivo. Thus, the uncertainties in the individual properties are compounded, resulting in a much larger uncertainty in the overall suitability of a compound for a given therapeutic objective.

If we accept that we are not working in a design paradigm, the next question is, "What *are* we doing and is it useful?" The computational methods developed to support drug discovery have enormous potential value in narrowing the range of compounds to consider, by eliminating those that have negligible chance of being a successful drug, while focussing efforts on chemistries with the best likelihood of success (in [8] the author used the analogy of card counting in the game blackjack to illustrate this). Computational predictions can also help to prioritise the most relevant experiments to perform on the selected compounds to quickly make a more accurate decision about their suitability. However, the value of computational methods can only be fully realised if they are applied appropriately; trusting a model too little wastes effort exploring compounds that are highly likely to fail, while too much trust may lead to missed opportunities by incorrectly rejecting good compounds. Understanding the confidence of a predictive model and the chemistry to which it may be usefully applied is critical; hence the necessary focus on validation of predictive models and understanding their domains of applicability [9].

This alternative perspective brings a new challenge; people find it difficult to make good decisions based on complex data in which there is significant uncertainty [10]. However, computational approaches can be used to support decision-makers in achieving multi-parameter optimisation [11, 12]. By combining predicted and experimental data (where available), it is possible to guide the selection and design of compounds with the best chance of achieving an appropriate balance of properties for a drug discovery project's objectives.

Looking forward for the next 25 years, how could we achieve computational drug *design*? The key is to improve the accuracy and transferability of predictions from computational approaches; as confidence in predictions increases, the number of compounds which need to be synthesised and tested to find a good compound will become smaller, until ultimately it becomes possible to design a small number of compounds with confidence that at least one will be a successful drug. But how can we further improve the accuracy of predictive methods?

The computational algorithms for statistical modelling have grown increasingly sophisticated and now include a wide range of powerful non-linear fitting techniques, such as Random Forests [13], Support Vector Machines [14], Artificial Neural Networks [15] and Gaussian Processes [16]. In addition, with the increased availability of large public domain databases, such as PubChem [17] and ChEMBL [18], more data is available than ever before (although a greater quantity of high quality data is always welcome). However, the molecular descriptors commonly used as independent variables in these models have not changed significantly for some time. They typically include simple topological descriptors or whole molecule properties such as lipophilicity, molecular weight, polar surface area or volume.

The information content of these simple descriptors is quite low, which means that each descriptor individually has a low correlation with the observed property. This is illustrated by Fig. 1, which shows the correlation between the potency (pIC_{50}) against the Estrogen receptor alpha and

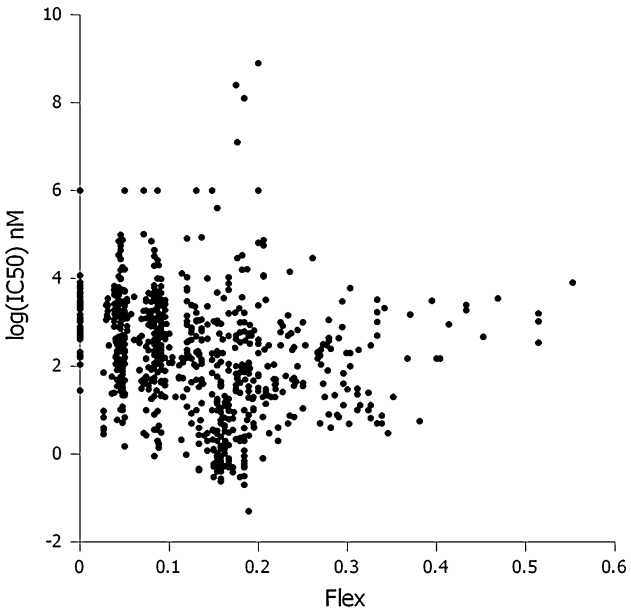

Fig. 1 Graph of observed estrogen receptor alpha log (IC_{50}) against compound flexibility for 794 compounds. Flexibility was the most important descriptor in a 2-dimensional QSAR model for this activity, illustrating the low correlation between biological activity and any individual descriptor contributing to the model. This is a common observation for this type of model

a flexibility descriptor (defined as the proportion of rotatable bonds) for a set of 794 compounds used to train a QSAR model. Flexibility was found to be the most important descriptor in a Gaussian Processes model with a coefficient of determination (R^2) of 0.74 on the training set and an R^2 of 0.67 on an independent test set of 169 compounds. Many descriptors must therefore be used together to obtain a good correlation and this inefficiency necessitates the requirement for a large quantity of high quality data with which to fit the model.

Furthermore, simple structural descriptors such as fragments and fingerprints lack transferability. For example, if a prediction is made for a new compound containing a fragment that did not occur in the training set used to build the model, it is unclear what effect that fragment may have on the predicted property; the new fragment may be irrelevant, in which case the accuracy of the prediction will be unchanged, or it may have a significant effect on the property meaning that the model will not be valid. Methods for determining the domain of applicability of a model can mitigate against this risk, but are still limited by the information content of the descriptors or fingerprints used to represent the domain.

Therefore, one approach to improving the predictivity and transferability of models in the future will be to develop better descriptors that capture the fundamental interactions governing the biological properties of interest, by representing the chemistry and physics of these interactions. As these interactions occur at a molecular level, this suggests the use of quantum mechanical (QM) simulations will be important. Some steps have been made in this direction, such as the estimation of hydrogen bonding acidity using density functional theory (DFT) calculations [19] or prediction of lability to metabolism [20, 21]. However, these calculations are not yet routine and it is likely that similar approaches will be applied more widely in the future.

The same trend is likely to be observed for structural and dynamical simulations. At present, the large majority of these use empirically derived, classical potentials to simulate the interactions between atoms. However, greater accuracy and transferability can be achieved with QM approaches; large scale DFT simulations, involving thousands of atoms, are now achievable with linear scaling techniques [22]. For example, Heady et al. showed how a linear scaling DFT approach can be combined with classical molecular dynamics to improve the accuracy of prediction of binding energies to the CDK2 kinase [23].

Classical molecular dynamics is still required in order to sample the conformation space of complex molecules and their interactions over sufficiently long time-scales. However, even here, there is the potential to use QM methods to improve the accuracy and transferability of classical potentials to achieve quantum accuracy [24]. This approach has so-far been applied to the simulation of crystals, but the extension of similar methods to biological systems could provide very useful improvements in the accuracy of dynamical simulations.

Of course, QM methods are more computationally intensive than those currently used on a routine basis. Therefore, their adoption will be driven by the increased availability of computing power. But, assuming Moore's law [25] continues to hold, there is a good chance that the next 25 years will see a genuine move towards a true design paradigm.

Acknowledgments The author would like to thank Ed Champness, Chris Leeding, Iskander Yusof and James Chisholm for helpful discussions regarding the topics in this article.

References

1. Beresford AP, Selick HE, Tarbit MH (2002) The emerging importance of predictive ADME simulation in drug discovery. Dug Discov Today 7:109–116
2. Oprea TI (2002) Current trends in lead discovery: are we looking for the appropriate properties? J Comput Aided Mol Des 16: 325–334
3. Paul S, Mytelka D, Dunwiddie D, Persinger C, Munos B, Lindborg S, Schacht A (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 9:203–214
4. Ekins S, Mestres J, Testa B (2007) In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br J Pharmacol 152:9–20
5. Bender A, Glen RC (2005) A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. J Chem Inf Model 45:1369–1375
6. Kroemer RT (2007) Structure-based drug design: docking and scoring. Curr Protein Pept Sci 8:312–328
7. Dearden JC (2006) In silico prediction of aqueous solubility. Expt Opin Drug Discov 1:31–52
8. Segall MD (2008) Why is it still drug discovery? Eur Biopharmaceut Rev. May
9. Weaver S, Gleeson NP (2008) The importance of the domain of applicability in QSAR modeling. J Mol Graph Model 26:1315–1326
10. Chadwick AT, Segall MD (2010) Overcoming psychological barriers to good discovery decisions. Drug Discov Today 15: 561–569
11. Ekins S, Boulanger B, Swaan P, Hupcey M (2001) Towards a new age of virtual ADME/TOX and multidimensional drug discovery. J Comp Aided Mol Design 16:381–401
12. Segall MD (2011) Multi-parameter optimization: identifying high quality compounds with a balance of properties. Curr Pharm Des (in press)
13. Svetink V, Liaw A, Tong C, Culberson J, Sheridan R, Feutson B (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:1947–1958
14. Doucet JPBF, Xia H, Panaye A, Fan B (2007) Nonlinear SVM approaches to QSPR/QSAR studies and drug design. Curr Comput Aided Drug Des 3:263–289

15. Devillers J (1996) Neural networks in QSAR and drug design (Principles of QSAR and drug design). Academic Press, London
16. Obrezanova O, Csanyi G, Gola JM, Segall MD (2007) Gaussian processes: a method for automatic QSAR modelling of ADME properties. J Chem Inf Model 47:1847–1857
17. Bolton E, Wang Y, Thiessen P, Bryant S (2008) PubChem: integrated platform of small molecules and biological activities. In: Annual reports in computational chemistry, vol 4. American Chemical Society, Washington DC, pp 217–241
18. Warr WA (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. J Comput Aided Mol Des 23:195–198
19. Kenny PW (2009) Hydrogen bonding, electrostatic potential, and molecular design. J Chem Inf Model 49:1234–1244
20. Jones JP, Mysinger M, Korzekwa KR (2002) Computational models for cytochrome P450: a predictive electronic model for aromatic oxidation and hydrogen atom abstraction. Drug Metab Dispos 30:7–12
21. Zaretzki J, Bergeron C, Rydberg P, Huang T, Bennett KP, Breneman CM (2011) RS-predictor: a new tool for predicting sites of cytochrome P450-mediated metabolism applied to CYP 3A4. J Chem Inf Model 51:1667–1689
22. Skylaris CK, Haynes PD, Mostofi AA, Payne MC (2005) Introducing ONETEP: linear-scaling density functional simulations. J Chem Phys 122:084119
23. Heady L, Fernandez-Serra M, Mancera RL, Joyce S, Venkitaraman A, Artacho E, Skylaris CK, Ciacchi LC, Payne MC (2006) Novel structural features of CDK inhibition revealed by an ab initio computational method. J Med Chem 49:5141–5153
24. Bartok AP, Payne MC, Kondor R, Csanyi G (2010) Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. Phys Rev Lett 104:136403
25. Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38:114–117