J Comput Aided Mol Des (2012) 26:113–114
DOI 10.1007/s10822-011-9507-0

PERSPECTIVE

Some thoughts on the "A" in computer-aided molecular design

Matthias Rarey

Received: 15 November 2011 / Accepted: 30 November 2011 / Published online: 13 December 2011

Computer-aided design (CAD) is a widely applied technology in engineering. Neither an airplane, nor a sky scraper or a microprocessor is designed without the substantial aid of computational means. The computational methods available in these fields are impressive. Although the calculations necessary for predicting air flows around planes or the crash behavior of a car are highly complex, a remarkable precision and reliability is achieved.

The situation in molecular design is a completely different one. Although the complexity of a small molecule is comparatively low, we are far away from providing precise predictions. At first sight, the reasons for this are quite obvious: Molecules are designed to influence biological systems—humans, in case of drugs—in a highly specific manner, and the difficulty arises from the complexity of biological systems, not from the complexity of molecules. Furthermore, the basic mechanism of action in biological systems is molecular recognition, which is not accurately calculable even for relatively simple cases.

One of the big challenges that are most central in computer-aided molecular design (CAMD) is the prediction of protein–ligand binding affinity from its complex geometry, the scoring problem. Here, biological complexity is reduced to a single protein. Although substantial progress has been made, a reliable prediction for the general case is out of reach. In 2007, Yvonne Martin stated that, if 'We cannot predict logP, why do we expect to predict protein–ligand binding affinity' [OpenEye, Cup VIII, Feb. 26–28 2007]. This observation still holds. Hence, it is even worse: Already the reliable and precise prediction of solubility is not sufficiently solved—and in this case, the biological element of the problem is reduced to the solvent water. While for engineering applications, the underlying laws of physics are well understood, our knowledge on the foundations of molecular interactions is still a limiting factor for molecular design.

About 25 years ago, our capabilities of computational modeling substantially grew due to the first availability of graphic workstations on the one hand and protein crystal structures on the other hand. In the following years, the foundations of structure-based modeling beyond simulations were laid. Among these were first approaches for molecular docking as well as de novo design. Looking back at these times, one gets the impression that the 'A' in CAMD rather represents 'automation' than 'aided' (see Fig. 1). Based on the initial successes, the hope was that computational methods would be able to reliably detect potent inhibitors from large databases, or—even better—can reliably create synthesizable bioactive compounds. So far, this hope did not hold.

Nevertheless, CAMD is more important than ever today. This circumstance is due to the very simple fact of an overwhelming availability of data. There is so much knowledge buried in more than 70,000 protein structures (PDB), more than 25 million compounds with their physico-chemical properties (ChemSpider), more than billions of affinity data points in databases like Chembl and PubChem. This data is a treasure, which is accessible only through computational means. And if we understand CAMD in its original meaning—namely as a method to aid molecular design—then it truly is a success story.

Our knowledge of molecular properties, interactions and mechanisms is founded on two grounds: Theoretical chemistry and the computational analysis of experimental data. This knowledge is heavily applied in the daily work

M. Rarey (✉)
Center for Bioinformatics, University of Hamburg,
Bundesstraße 43, 20146 Hamburg, Germany
e-mail: rarey@zbh.uni-hamburg.de

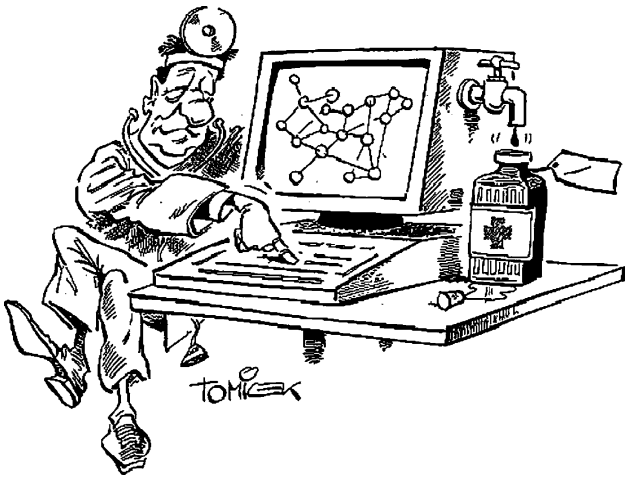

Fig. 1 Tomicek's cartoon—originally explaining personalized medicine—fits very well to the great hope of 'Computer-Automated Molecular Design' (Figure reprint with kind permission from Jürgen Tomicek, http://www.tomicek.de)

of designing new bioactive compounds. Furthermore, even if precise predictions are impossible in many cases, reasonable estimates already have a dramatic positive impact. When it comes to the decision of which molecules to test for bioactivity, the space of compounds to choose from is so huge that we basically have two options: Either we apply computational design and search strategies or we just guess and forego an immense opportunity. Not surprisingly, CAMD has its well-established place in modern design strategies for new bioagents.

Certainly, the question of what comes next in CAMD arises. In this field, there are so many unsolved or not satisfactorily solved problems that it is of primary importance to find the right focus. First of all, there are the basic research questions which have to be resolutely addressed. A deep understanding of molecular interactions is necessary for nearly all problems related to molecular design ranging from molecular superposition to protein structure prediction, docking and scoring. These problems are all well-studied, so it is likely that there are no obvious solutions to them. It is tempting to jump at simpler, less-addressed problems. Instead, it will be important to actively search for alternative views to them as well as for radically new approaches. Several of these new ideas will probably fail to really improve over the state of the art. When statistically sound evaluation and validation is in place, they can, nevertheless, increase our knowledge about the problem substantially.

Coming back to the 'A', we should carefully think of what is really helpful for the chemist at the end of the day. The chemist's knowledge and intuitive thinking will probably never be fully captured in software systems. Therefore, best results can be achieved if software systems focus on what they can do best: operate systematically and at large scale, thereby searching for the needle in large haystacks of data and providing decision support. If the 'A' in CAMD is truly understood as 'aided', the success story of Computer-Aided Molecular Design only just began.