J Comput Aided Mol Des (2012) 26:103–105
DOI 10.1007/s10822-011-9520-3

PERSPECTIVE

The character of molecular modeling

Anthony Nicholls

Received: 5 December 2011 / Accepted: 7 December 2011 / Published online: 17 December 2011

Abstract In asking what progress might occur in molecular modeling in the next 25 years it is worth asking what progress has been made in the last twenty-five. In doing so it is hard to be optimistic for the future of the field unless a greater commitment is made to basic science.

Keywords Structure-based design · Docking · Proteins · Scientific method

In a series of lectures for the BBC in 1964, Richard Feynman [1] commented that it was a great time to be in particle physics. The first hints of the quark-like nature of the proton and neutron were appearing, the sense of deep connections between fundamental forces was being discussed and every meeting was full of exciting new findings. In Feynman's mind he already saw a time when a theory, such as the current Standard Model, would emerge and he even recognized that further advances would be difficult because increasingly large energies would be required to challenge theory. To be there, at that time, to be a part of that great leap forward, to be antediluvian but knowing the great flood was coming, was, he believed, the best any scientist could ever hope for. I have told this story to many young scientists to encourage them to stay in molecular modeling even though it is a very slowly evolving field. "It won't always be so", I say, "Please stay and be a part of the inevitable leap forward". Sometimes they do, some times I lose them to information sciences, or wider pursuits. A special issue such as this is an opportunity to honestly consider whether the ones who stay will enjoy Feynman's reward, to see the field revolutionized, or whether those who leave will be glad they did.

The first consideration is whether we should expect much progress over the next 25 year since there has been precious little over the last twenty-five. Back in 1986, you might have been able to look at a crystal structure of a protein–ligand complex, see where the ligand might be improved, make some compounds, and develop some useful SAR. What has changed? At a Gordon Conference from a few years ago the presentations on successful molecular modeling consisted of little else than "spot the pocket, fill the pocket". These successes are just anecdotes. We don't hear the anecdotes of failure, which are not uncommon. There are major pharmaceutical companies, heavily invested in structural biology, who have struggled to ascertain the real benefit of protein structures relative to the more traditional approach of 'blindly' developing SAR via synthesis and testing. This is because structure enhances what economists call the "sunk cost fallacy": when we have made a substantial investment in a task we are reluctant to give up, even if that is the most cost effective course of action. In structure-based design we are held captive by the picture crystallography gives us, even though we know it is only a snapshot of reality. Just as it sometimes really helps design, that picture can also prevent cost-effective project termination, an essential part of the drug discovery process.

There have been advances, in particular Dave Weininger's invention of SMILES and SMARTS stand out as a signature moment for the field, one still appreciated mostly only in industry not academia. Pharmacophores and shape analysis, made possible by regular increases in computing power, have increased the usefulness of molecular similarity at the 3D level. But, as Willard Quine commented [2], the use of similarity is the sign of an immature science

A. Nicholls (✉)
OpenEye Scientific Software, Inc., Santa Fe, NM, USA
e-mail: anthony@eyesopen.com

and the application of (what ought to be) "mature" science, i.e. physics, has had minimal impact on the field. Consider docking. For all the effort and application inside and outside of industry, docking has had relatively little impact on drug discovery. It does have utility, mostly helping to predict poses of ligands in active sites, something it can do about 30% of the time from a "standing start", i.e. knowing nothing other than the structure of the protein. As this can circumvent the need for co-crystallization this is no little achievement, even though other, simpler methods (e.g. alignment by maximum common substructure) are often competitive and faster, although requiring some prior knowledge. However, the broader fantasies of supplanting physical screening, let alone the holy grail of predicting binding energies is as distant now as it was when docking software first became available. The reasons are simple enough to illustrate by looking at the actual, as opposed to the claimed, utility of 'flexible' docking, i.e. where protein side chain or main chain motions are allowed. Such techniques can help pose *known* binders, but make virtual screening of to-be-determined binders worse. This is because we cannot predict the relative energies of different protein states. If we know something binds then increasing the search space of possible active site configurations can help. But if we don't know if something binds then more protein configurations merely increases the false positive rate. It is our inability to accurately assess the energy of proteins that prevents us from making progress towards useful predictions in docking.

And yet, having suggested that progress has been minimal, except for advances in the application of similarity, I would state categorically that we are better at molecular modeling than we were 25 years ago. How so? Because we are better at using the tools, even if the effectiveness of the tools has not significantly improved. The field, as a whole, has become 'expert', in the technical sense of having spent enough time, often quoted as 10,000 h of goal-driven practice [3]. Evidence for this can be found in the blind SAMPL challenges run by OpenEye and published on in this journal [4]. At SAMPL humans routinely beat machines in such tasks as pose prediction. That expertise makes the tools, not tools an expert, really just reinforces concern over the lack of fundamental progress in the field. Rather than great leaps forward we have made small strides by looking at data. For example, these days it is unusual not to be able to construct a homology model of some utility, even for membrane proteins. For all practical purposes theorists never solved the protein-folding problem, crystallographers solved it for them. This is another instance of the application of the similarity principle. Expert modelers are not abstracting principles of wide applicability; they are recognizing domains of experience. On the bright side, just as crystallographers have made homology modeling useful, the ever increasing body of knowledge of protein–ligand binding will eventually have impact, if only through the software embodiment of human ability. We shall, in the next 25 years, fumble our way to a much better predictability of proteins and protein–ligand binding. We cannot fail to do so, even if the results are deeply unsatisfying from a scientific standpoint. This has happened in other fields. In voice recognition there used to be schemes to develop formal theories of the human voice that would allow for perfect voice recognition. Instead, machine learning narrowed its focus to being able to recognize whether you said "credit" or "debit"—not an easy problem but easier than constructing Arthur C. Clarke's HAL from "2001".

Niels Bohr once said that all of science is either physics or "stamp collecting". In the next 25 years will see progress as we collect more stamps. Will we see any real breakthroughs, any real application of physics? I think there are two current areas of hope, namely small molecule crystal structure prediction and the folding of small proteins. In the former, Neumann and others have shown in blind challenges run by the University of Cambridge [6] that the application of custom force-fields and parameterized long-range dispersion interactions can actually give reasonable predictions of the lowest energy crystal polymorph- perhaps the first application of physics to an problem of potential pharmaceutical utility. Such techniques are slow- requiring several months of computation, but any problem that is currently difficult but tractable will be straightforward in 10 years, let alone twenty-five. This will fundamentally shift the application of physics within industry, not because the polymorph problem is so vital to drug discovery but because the pathway will then be cleared to the much more central property of solubility. Crystal energy plus vacuum–water solvation equals solubility and for several years we, and others, have been making progress on solvation, again through SAMPL challenges [5].

The other ray of hope is that increases in computing power, whether from custom hardware or harnessing worldwide computing resources, seem to be showing that small protein structures can be folded from the linear sequence [7]. I find this profound for a couple of reasons. The first is simply that because it has taken so long to get to this point it illustrates that empirical methods are just not going to work for proteins—you have to get the physics right. The second is that we *don't* have the physics 'right' and yet we can still fold proteins with a marginal stability of only a few kcal/mol. We can't calculate the solvation of small, drug-like molecules to within one kcal per mole. Merz [8] has been pointing out that even high quality quantum calculations of a pair of interacting groups typically has an error of around half a kcal per mole—which adds up over a protein–ligand interaction. Yet simulations can somehow distinguish the correctly folded state and also

get the kinetics roughly right. My guess is these simulations work partly because we are looking at the relative energy of states, partly because some force field errors must be anti-correlated. Regardless it is a tremendous feat of engineering to be able to simulate such events. It should be noted, however, that 'getting the fold right' is a long way from having side-chains placed such that synthesis can be guided and it is unclear if larger, useful, systems are practical.

So there will be progress in the next quarter century because of more data and more CPU cycles—but will it actually be transformative? I doubt it. Although I make the case for progress in polymorph prediction and protein folding, I am equally convinced progress in most areas will be slight. For instance, I'm not sure more postage stamps will greatly improve the prediction of protein–ligand interactions. I suspect the dimensionality of the problem is sufficiently high that even a vast increase in the quantity of data will not help. And this isn't even considering the quality of that data, something I won't address here but the lack of which is clearly holding back progress. It's entirely possible more data will just mean more extensively parameterized models of limited prospective utility. Similarly, even with respect to crystal structure prediction and protein structure elucidation, we might be at a "Pauling point", i.e. an effective but limited description of systems. Both approaches have addressed fairly simple systems to date and it is entirely possible that as system sizes scale towards the useful both approaches will fail. Can simulation really fold a complex protein? Can quantum calculations really calculate sub-kilojoule energies for large, flexible organic molecules, i.e. drugs? There will be utility there for sure, but a great leap forward? I'm unconvinced.

My *real* thesis, the one I want to end with, is that unless simulations, or model building built on data dredging, do one simple thing there won't be much progress in the next 25 years. That one simple thing? Science. What amazes me about molecular modeling is how fundamentally un-scientific it is. Galileo had a very simple definition of science, a definition that really founded Western science four centuries ago. His said science was a three-part process: Resolution, Demonstration, and Experimentation. Resolution was having an idea—being able to 'resolve' that idea into a statement intelligible in terms and concepts you knew. Demonstration was thinking through the idea, working out its consequences in the framework of concepts you have accepted. Demonstration was the way your saw what your idea meant and how to test it. And then Experimentation was the actual testing, the step that fed back to the Resolution and to the framework of concepts that supported that idea. We get lots of Resolutions in molecular modeling—although not much novelty considering the field has little sense of prior art. We also get some Demonstration but we don't get much Experimentation. Checking the consequences of our idea against what we already know doesn't count—that is a part of the process of Resolution. The best we usual do is to find and use data we hadn't originally considered. If that data is substantially different we typically think of this as good work. And it is, but it isn't Experimentation. Thinking of crucial, real-world experiments that can disprove an idea—distinguish one idea conclusively from another—we don't do this. In pharma, what matters is making the right molecule- not making molecules to confirm or refute a theory, while in academia the endless granting cycle, the relative new concept of needing to pursue "useful" work, generally prohibits real science. So the field meanders.

There is no reason it has to stay this way. The simple commitment to spend a small percentage of the science budget at the NIH or at pharmaceutical companies on *non*-translational work, providing support for the small cabals of scientists actually interested in making fundamental progress—25 years compound interest on that sum, as calculated through Galileo's process, would be enormous. Reestablishing the contact between theorists and experimentalists, the publishing of high quality data, conferences devoted to the actual testing of ideas—in 25 years we might hope molecular modeling could become a real scientific discipline.

References

1. Feynman RP (2001) The character of physical law (Messenger Lectures, 1964). The MIT Press, Cambridge. ISBN-10: 0262560038
2. Quine WV (1977) Natural kinds in ontological relativity and other essays. Columbia University Press, New York
3. Ericsson KA, Charness N, Feltovich PJ, Hoffman RR (eds.) (2006) The Cambridge handbook of expertise and expert performance. Cambridge University Press, Cambridge. ISBN-10: 0521600812
4. Skillman AG (ed.) (2008) Special issue: SAMPL2: tautomer ratio and solvation energy challenge, JCAMD, vol. 24, No. 4 (2010). J Med Chem 51(4):769–779
5. Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Bayly CI, Cooper MD, Pande VS (2008) Predicting small-molecule solvation free energies: an informal blind test for computational chemistry
6. Day GM, Cooper TG, Cruz-Cabeza AJ, Hejczykm KE, Ammon HL, Boerrigter SXM, Tan JS, Della Valle RG, Venuti E, Jose J, Gadre SR, Desiraju GR, Thakur TS, van Eijck BP, Facelli JC, Bazterra VE, Ferraro MB, Hofmann DWM, Neumann MA, Leusen FJJ, Kendrick J, Price SL, Misquitta AJ, Karamertzanis PG, Welch GWA, Scheraga HA, Arnautova YA, Schmidt MU, van de Streek J, Wolf AK, Schweizerr B, Cryst A (2009) Significant progress in predicting the crystal structures of small organic molecules—a report on the fourth blind test. B65
7. Voelz VA, Bowman GR, Beauchamp K, Pande VS (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9 (1–39). J Am Chem Soc 132(5):1526–1528
8. Ucisik MN, Dashti DS, Faver JC, Merz KM Jr (2011) Pairwise additivity of energy components in protein-ligand binding: the HIV II protease-indinavir case. J Chem Phys 135:085101