Comment

Model behavior

Gregory A Petsko

Address: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454-9110, USA.
E-mail: petsko@brandeis.edu

Published: 4 July 2001

Genome ***Biology*** 2001, **2(7)**:comment1009.1–1009.2

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/7/comment/1009

"I am not a doctor, and you are not ill;" says Selim to Osmin in Voltaire's *Philosophical Dictionary* (1694), "but it seems to me I should be giving you a very good prescription if I said to you: 'Put not your trust in all the inventions of charlatans ... and believe that two and two make four.' " I've always liked this advice, but it seems to me that most scientists are prone to believing - or perhaps hoping is a better word - that two and two can, sometimes, with the aid of the right technology, make five.

Three examples of this misguided faith have caught my attention recently. The first was a report that the extremely sophisticated (and extremely expensive) 'stealth technology' that supposedly allows US military aircraft to evade detection by enemy radar can be defeated by inexpensive - and commonly available - networks of mobile phone towers. The second is a fascinating book by Bruce Schneier: *Secrets and Lies: Digital Security in a Networked World* (John Wiley & Sons, US and UK; 2000). Schneier, an authority on computer security, wrote the book because he had come to the conclusion that, contrary to the belief of those who invest a fortune in fancy encryption and authentication methods, it is impossible to build a totally secure system. Computer security systems rely on bug-ridden or unstable hardware and software, and their creators and users are fallible and unreliable human beings. Schneier's advice is, "put not your trust in mathematics." He would have been delighted with Samuel Johnson's remark that there is no problem the mind of man can set that the mind of man cannot solve. His is an entertaining - and sobering - look at a catalog of disasters, foulups and frauds, and a marvelous record of a journey that took him from idealist to a pragmatist.

The third example is more subtle, but perhaps equally telling. It also reminds us that, when it comes to depending on technology, we would all do well to be more pragmatic. It comes from the burgeoning science of structural genomics. In the latest issue of *Nature Structural Biology*, Vitkup *et al.* (*Nat Struct Biol* 2001, **6:**482-484) attempt to calculate how many protein structures will need to be determined in order to meet the stated goal of obtaining useful, three-dimensional models of all proteins by a combination of experimental structure determination and comparative model building. They evaluate different strategies for optimizing information return for effort invested, and conclude that the strategy that maximizes structural coverage requires about seven times fewer structure determinations compared with the strategy in which targets are selected at random. With a choice of reasonable model quality and the goal of 90% coverage, they extrapolate that it would take approximately 16,000 carefully selected structure determinations to provide information allowing the construction of useful atomic models for the vast majority of all proteins. They further point out that, in practice, unless there is global coordination of target selection, the total effort is likely to increase by a factor of three.

This is a nice analysis and its conclusions are very important for the field, but that isn't the point I want to make here. What I found striking is their assumption, which I think is right, that the goal of the structural genomics initiative will be seen to have been met when enough structures have been done to allow all other structures to be modeled from them. In other words, we are putting our trust in homology modeling.

Given the two examples I mentioned earlier, I think it is worth considering whether this trust in technology is justified. Homology modeling aims to produce a reasonable approximation to the structure of a protein using the known structure of a homolog: a protein related to it by divergent evolution from a common ancestor. Structures that have diverged too far cannot be modeled reliably; the arrangements in space of their secondary structure elements tend to shift too much. In practice, structures with more than about 40% amino-acid sequence identity, and with no large insertions or deletions in their aligned sequences, can usually be used to produce homology models roughly equivalent to a

medium-resolution (about 3 Angstroms resolution) experimental structure. Vitkup *et al.* aim for approximately that degree of reliability.

What can be done with such models? Well, it is more instructive to consider what cannot be done with them. They can be used to determine which amino acids are in the catalytic site or molecular recognition site if those sites are in the same place in the modeled and experimentally observed protein structures, but they cannot be used to find new binding sites that have been added by evolution. At present, there is no reliable way to interrogate a purely modeled structure and locate such sites from first principles. Further work in this area is urgently needed. Homology models cannot be used to study conformational changes induced by ligand binding, pH changes, or post-translational modification. At present, computational tools to generate such changes from a starting model are not robust. Again, more work is needed here. Homology models also cannot be docked together to produce good structures of protein-protein complexes; not only are the docking algorithms unreliable, but the likelihood of significant conformational changes when proteins associate makes it impossible to know whether one is docking the right structures. In short, many, if not most, of the things that biologists want to do with a protein structure cannot be done with confidence using homology models alone.

This is not to say that such models are useless. But it is meant to inject a cautionary note to the frenetic salesmanship that surrounds genome-wide structure determination. There will still be a huge amount of structure-based work to be done on important proteins, real work using real proteins and based on experimentally determined structures. The homology models will be very helpful in determining those structures, but will not replace them for most things of interest. And it is unclear that the structural-genomics initiatives will produce those structures: structures with ligands or cofactors bound, structures at different pH values, structures of modified proteins and structures of protein-protein complexes. Such work will be done by individual investigators, who need the support of research funding that may be siphoned off into genome-wide programs if we put too much trust in high-throughput technology.

In the US we are dealing with the after-effects of a bubble economy in internet company stocks. In Japan, a similar bubble of asset price inflation, chiefly in property, has long since burst with consequences that are still being felt. Scientific research has its bubbles too, but they come not from inflated prices but from inflated expectations. If we put too much trust in any one technology and neglect the important things that are less glamorous and require harder and longer efforts, we are in danger of failing to follow Selim's prescription. As the euphoria over all things genomic continues, we would all do well to remind ourselves that two plus two still equals four, and always will.