

Comment

The Grail problem

Gregory A Petsko

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

Published: 9 June 2000

Genome **Biology** 2000, 1(1):comment002.1-002.2

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2000/1/1/comment/002>

© Genome**Biology**.com (Print ISSN 1465-6906; Online ISSN 1465-6914)

The Holy Grail is a familiar metaphor in science. A current Holy Grail is the complete sequence of the human genome, but there seems to be one for every field of biology. In biophysics, it is the prediction of the three-dimensional structure of a protein from its amino acid sequence alone. But what if, when someone claims that this Grail is found, we - like the hero of the film *Indiana Jones and the Last Crusade* - can't be sure it's the right one? Something like that could happen in structure prediction unless we refine our measure of correctness for a predicted structure.

Genome sequencing projects have created a heavy demand for protein structure prediction. Structure prediction at present relies on modeling based on data collected from the many proteins for which both sequence and structure are known (reviewed by Baker, *Nature* 2000, **405**:39-42). When the sequence identity between a protein of known structure and the putative homolog is high (about 50% or greater), most existing modeling methods work well. The difficulty arises in the most interesting cases, when sequence identity to proteins of known structure is low or absent. No completely reliable methods for structure prediction exist for these cases at present. New methods and claimed improvements to existing ones are always evaluated on test systems in the same way: the predicted structure is superimposed onto the true one so as to minimize the root-mean-square deviation in atomic coordinates - a measure of the difference in position - between all pairs of equivalent atoms (which may be alpha carbons or all backbone atoms; side-chains are usually excluded). This single number, the root-mean-square or rms deviation, is then reported as the measure of how well the predicted and actual structures agree.

The use of the rms deviation as a measure of the quality of a structure prediction has its origins in the early days of protein crystallography, when there was considerable interest in the precision of experimentally determined protein structures. Two different structures of the same protein solved, for

example, in two different laboratories, or by the same laboratory in two different crystal forms, would be superimposed and the rms deviation would be calculated. Well-determined structures at high resolution often yield rms deviations of less than 0.5 Angstroms in such a comparison.

But predicted structures are not experimental ones, and the rms deviations between models of homologous protein structures and real ones are typically between 2 and 4 Angstroms, even in the best cases. And for the far more difficult problem of taking an arbitrary polypeptide chain and folding it up into the correct structure *ab initio*, by brute force calculation, the best available methods usually produce numbers even larger. All of which raises the same question as in the *Indiana Jones* Grail situation: how do we tell the true Grail from a false one? What constitutes 'good enough' agreement between a predicted structure and the real one to demonstrate that the prediction method works? No one expects *de novo* folding to get within 0.5 Angstroms rms deviation, but is 2 Angstroms good enough? What about 3?

I believe that a fundamental difficulty faced by the whole folding field - one shared with structural biology in general - is that it has never solved this Grail problem. The use of a single number to represent the disagreement between hundreds or thousands of pairs of numbers is of no statistical validity. Consider two predicted-observed structure pairs, each of which has an rms deviation of 4 Angstroms. Are they of equal quality? Suppose one pair has a roughly 4 Angstrom difference between every one of its superimposed sets of atoms, while the other has most of the equivalent atoms about 1 Angstrom apart except for a small number (say in a few loops) where the deviation is 10 Angstroms, making the overall value 4. We would certainly prefer the latter prediction, but the rms deviation alone would never allow us to decide that. Yet this is usually the only measure that is reported. That is just silly.

There are a few simple changes to this custom that would help give the field of structure prediction (and structure comparison) some much-needed numerical credibility. One is for referees to reject, out of hand, manuscripts that report only the rms deviations between pairs of structures. The maximum and minimum deviations in the whole set should be given, and I would recommend reporting the most commonly observed deviation as well. Best of all would be a histogram of the deviations; I see no reason why we should not enforce that as a requirement in all publications.

But even with these improvements, I don't think any number or set of numbers is the best indication that the Grail of always being able to predict a protein structure from its sequence has been found, because we still have no good sense of what number constitutes 'close enough'. But there is an obvious method of evaluation that will allow any structure prediction method to be assessed. It is simply to demand that the method produce a model that can be used to solve the corresponding protein crystal structure by the method of molecular replacement.

Molecular replacement is a common crystallographic tool for solving the structures of proteins that are similar in fold to ones that have already been determined. The crystallographer calculates the diffraction patterns expected for the known structure when it has been placed in all possible orientations in the unit cell of a theoretical crystal of the unknown protein, and compares the observed and calculated diffraction patterns. A likely solution is defined as one where the two patterns agree within some specified numerical criteria. But there is a further, absolute test as well: the correct solution allows the unknown structure to be completed (that is, refined to crystallographic convergence, when the observed and calculated diffraction patterns match each other as closely as possible) by automatic refinement methods combined with manual model rebuilding.

Such a test can be set up for any computational method that claims to be able to solve the *ab initio* folding problem or to improve on existing methods of modeling weakly similar structures. It is well-defined and easy to carry out. And when a computational procedure comes along that passes this test for helical proteins, and for all-beta-sheet proteins, and for proteins with mixed secondary structures and proteins with multiple domains, we will know that the true Grail has been found at last.