Exercises in prognostication: Crystal structures and protein folding

Jack D. Dunitz* and Harold A. Scheraga**

*Organic Chemistry Laboratory, Swiss Federal Institute of Technology, Eidgenössische Technische Hochschule–Hönggerberg, CH-8093 Zurich, Switzerland; and *Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301

Contribution by Jack D. Dunitz and Harold A. Scheraga, August 5, 2004

Given the molecular structure of a chemical compound, can one predict its crystal structure (1–3)? Given the amino acid sequence of a protein, can one predict its three-dimensional structure? Both questions are not only of theoretical interest but of practical interest as well. As far as the crystal structure problem is concerned, we know that polymorphic forms of the same compound may differ widely with respect to physical properties, such as melting point, vapor pressure, density, solubility, photochemical stability, etc., differences that may affect the usefulness of the compound in industrial and pharmaceutical applications. One need only think of the difference between graphite and diamond. Because the function of a protein depends on its three-dimensional conformation, the protein folding problem has obvious implications for deciphering the chemical and biological significance of the linear code expressed in the amino acid sequence. An understanding of protein folding will also help in combating misfolding diseases, e.g., Alzheimer’s disease and cystic fibrosis, and even in design of proteins with new or enhanced function. The two problems, crystal structure and protein folding prediction, are very different, but they share important aspects. In particular, both problems depend on our ability to calculate the most stable arrangements of large molecular aggregates from general principles. Indeed, both challenges are usually regarded as global optimization problems to identify the structure(s) of lowest potential energy or of free energy with entropy contributions included.

Both problems have been subjected to objective blind tests in which knowledge of the experimental structures was not made available to those preparing predictions until the predicted structures had been submitted to the judges. Three such crystal structure prediction tests (Crystal Structure Prediction Workshops) have been organized by Sam Motherwell and were held at the Cambridge Crystallographic Data Centre in 1999 (4), 2001 (5), and 2004 (W. D. S. Motherwell and J.D.D., unpublished work). The evaluation of the predictions of protein structures, designated Critical Assessment of Techniques for Protein Structure Prediction (CASP), proposed by John Moult and under the aegis of the Lawrence Livermore National Laboratory, has been carried out every 2 years, the most recent (sixth) test being conducted in 2004 (http://predictioncenter.llnl.gov/Casp6.html).

In crystal structure prediction, energy calculations are usually based on a force field involving summation of intermolecular atom–atom interaction energies; different force fields involving different kinds of energy parameterization are in use, and none of them appears to be strikingly superior to the others. The main problem seems to be not so much a matter of generating stable crystal structures but rather of selecting one or more possible structures from very many almost equiequilibrium candidates. For example, even for such a simple molecule as benzene, with only one known crystal structure at normal pressure, calculations yield at least 30 possible crystal structures with estimated lattice energies within a 10-kJ-mol⁻¹ range of the global minimum (6–8). Indeed, the one with the lowest calculated energy turned out to correspond to the experimental structure but only by a fraction of a kJ-mol⁻¹. For a given molecule, the concentration of several different packing arrangements within a small energy range appears to be fairly typical and is consistent with the widespread occurrence of polymorphs with very similar energies. Various kinds of search routines are in use, and most of them seem fairly reliable in locating a similar (but by no means identical) collection of possible low-energy structures, provided there are not too many degrees of conformational freedom. The fine tuning of the force-field parameterization, of course, can influence the energy ranking. Introduction of entropy contributions tends to make calculated free energy differences even smaller than potential energy differences. The tighter the packing, the lower the potential energy tends to be, but also the smaller the vibrational entropy, because the molecules are held more firmly in their equilibrium positions and orientations. Structures with similar energies may have quite different packing arrangements in different space groups, and a small change in the force-field parameterization can easily alter the energy ordering.

Although the three Crystal Structure Prediction Workshops (refs. 4 and 5; W. D. S. Motherwell and J.D.D., unpublished work) have not been marked by spectacular success in prediction, they have been most useful in underlining the limitations of our methods. The molecules in question, illustrated in Fig. 1, range from the simple (propane and azetidine) to the moderately complex (structure VI). Although none of the participating groups was consistently successful in predicting the experimental crystal structures, some groups did score individual hits, and many groups found the experimental structure somewhere in their list of possible low-energy structures. Thus, present methods are capable, if not of predicting the experimental structure(s) a priori, then at least of providing a set of structures as possible polymorphs. This means that, if relevant partial information is available, such as knowledge of the space group, unit cell dimensions, or an experimental powder diffraction pattern, then present-day methods have a good chance of finding the correct crystal structure.

Of course, the experimental crystal structure may not be selected on thermodynamic grounds alone but may depend on kinetic factors. The rate-determining step in the formation of a crystal may be the formation of particular clusters that act as viable nuclei for a subsequent crystal growth process. Unfortunately, very little is known about crystal nucleation and, because it is a nonequilibrium process (crystallization occurs only under conditions of supersaturation of a solution or supercooling of a neat liquid), it is not easy to study it experimentally or theoretically. Our lack of knowledge about crystal nucleation may be the main hurdle on the way to reliable crystal structure prediction.

The analogous protein-structure prediction problem is being treated by two different approaches. The first approach is a knowledge-based one, in which ex-
Experimental data on known structures are used as constraints in the global optimization. Such experimental data include knowledge or prediction of the location of secondary-structure elements, i.e., $\alpha$-helices and $\beta$-structures; knowledge of the structure of a related protein with a similar but not identical amino acid sequence (homology modeling); and matching of fragments against a database of known structures (threading and fragment coupling). More challenging is an ab initio (or physics-based) approach, based on empirical force fields analogous to those used in crystal structure prediction and relying only on the validity of the applied potential energy function and the efficiency of the global optimization procedure. The approach of choice depends on the objective. If this is simply to obtain the structure, then of course one should use as much knowledge-based information as is available. On the other hand, if the objective is to gain an understanding of how the physical interactions determine the structure, then no use should be made of knowledge-based information in the conformational search. As in crystal structure prediction, in both approaches, the problem of selecting the most stable tertiary structure from many almost equienergetic forms is a severe one.

The Critical Assessment of Techniques for Protein Structure Prediction exercises have demonstrated that both knowledge-based and ab initio procedures seem to be on the right track; the knowledge-based approach is of course far more successful (9), but the ab initio approach is making progress (10). It is necessary to continue to focus on improvement of the potential energy functions and of the treatment of the solvent environment of the protein, and on the procedures to search conformational space to locate the thermodynamically most stable structures.

As also in the crystal structure prediction problem, kinetic factors are certainly important in determining the folding pathways, but their specific role in determining the three-dimensional structure of a globular protein has yet to be identified. In fact, attempts are being made to identify protein folding pathways by molecular and Langevin dynamics and stochastic difference equation approaches. It is especially important to elucidate the roles of chaperones and of agents that catalyze specific processes such as cis-trans isomerization of peptide bonds involving proline residues and sulfhydryl-disulfide interchange in the rate-determining step.

The next few years will undoubtedly produce improvements in potential functions and search procedures. It would be comforting to think that they will also produce new ideas and unexpected solutions to some of the problems. Possibly the role of individual atomic nuclei as loci of intermolecular attractions and repulsions will need to be replaced by better models based on more delocalized quantum mechanical charge density distributions. In any case, the Cambridge Crystallographic Data Centre and Critical Assessment of Techniques for Protein Structure Prediction tests need to be maintained so they can continue to document progress and monitor excessive claims.

Fig. 1. Molecular diagrams of the compounds presented to the participants in the three Crystal Structure Prediction Workshops (refs. 4 and 5; W. D. S. Motherwell and J.D.D., unpublished work). Below each diagram are the number of hits (successful predictions) and of groups that were engaged in that particular problem, each group being allowed three proposals. Compound I was found to exist in two polymorphic crystal forms; the less stable form was correctly predicted by four groups, the more stable form by none. After compound VIII had been set as one of the targets for the 2004 Workshop (W. D. S. Motherwell and J.D.D., unpublished work), it was noticed that its crystal structure had already been described in a conference abstract in 2002. Nevertheless, because several groups had already made extensive calculations for this compound, it was retained as a target molecule. It may seem surprising that no participant succeeded in predicting the crystal structure of such a simple molecule as XI. The reason may be that this compound crystallizes (at 170 K) with two symmetry-independent molecules in the unit cell.