

# Theory and simulation

## Can theory challenge experiment?

Editorial overview

Patrice Koehl\* and Michael Levitt†

### Addresses

Department of Structural Biology, Fairchild Building, Stanford University School of Medicine, Stanford, CA 94305, USA

\*e-mail: koehl@zen.stanford.edu

†e-mail: levitt@zen.stanford.edu

**Current Opinion in Structural Biology** 1999, 9:155–156

<http://biomednet.com/elecref/0959440X00900155>

© Elsevier Science Ltd ISSN 0959-440X

### Abbreviation

MD molecular dynamics

As the end of this millennium approaches, understanding protein structure and dynamics remains one of the most challenging problems in biology; solutions do not appear to be imminent. This statement is not meant to detract from the ongoing theoretical efforts in structural biology. It serves to emphasize the fact that this is a very difficult problem and that it is important to recognize it as such. Until the problem is solved, it can be difficult to assess whether progress is being made. This year's reviews for the Theory and simulation section have consequently been chosen to reflect the fact that simulations are becoming increasingly successful and that we can, indeed, look forward to a brighter future. Although the success of a theory is usually defined by its ability to explain and predict experimental results, it is also important that simulations lead to new experiments. Recent developments, reviewed in this section, clearly indicate that this is the case. In fact, simulations can now generate results that lead to questions concerning the accuracy of an experimental measurement [1].

Simulations of protein dynamics are now routinely performed; a simple search of MEDLINE with the key words 'molecular dynamics' and 'protein', restricted to the past two years, recovers 836 different publications. Theoretically, following the trajectory in time of any of these molecular dynamics (MD) simulations should provide explicit information on the process of folding, identifying the final conformation as the native state and defining thermodynamic properties by computing averages over the sampled set of conformations. In practice, however, we are far from this goal. The length of the simulation remains a major limitation; most trajectories are computed over tens of nanoseconds, which is still orders of magnitude smaller than the fastest folding times for proteins. It is clear that longer simulations are needed and these may now be within reach. By combining the use of highly parallel computer hardware with a careful rewriting of the MD code, Duan and Kollman [2] were able to compute a 1  $\mu$ s all-atom trajectory of a villin headpiece

subdomain (a 36-residue protein fragment) in the explicit presence of water.

Sampling is another problem, in that MD simulations only sample a small region of conformational space. This is not necessarily improved by increasing the length of the simulation. In fact, it was shown that several short MD simulations of a protein provide a better coverage of its phase space than a single simulation lasting longer than the sum of the simulation times of the short trajectories [3].

Finally, there is the problem of potential energy functions, which is probably best revealed by homology modeling. The goal is to build a structural model of a protein on the basis of close sequence similarity to a template protein of known structure. Most of the methods for comparative modeling are criticized for their inability to improve the model obtained by directly transposing elements of the template. Even more embarrassing is the fact that the energy minimization or MD simulation applied to this initial model generally leads to a model that is less like the experimental structure [4]. Several perspectives on how to solve these problems are presented in the following reviews.

Doniach and Eastman (pp 157–163) survey recent progress in MD simulations of macromolecules. Although part of this progress resulted from an increase in computer power, their review clearly shows how much new and improved techniques have contributed as well. For example, techniques have been proposed to remove the resonance problems from multiple time step dynamics, which should lead to much longer simulation times. A major issue associated with MD remains the sampling of conformational space. This has led to the notion of 'essential dynamics', in which motions are constrained either to move along the essential modes detected in a classical MD simulation or to the 'jumping-among-minima' algorithm, reviewed in detail by Kitao and Go (pp 164–169). The conformational space explored by both techniques is considerably larger than that explored by classical MD. Doniach and Eastman report a puzzling result from the work of Chatfield *et al.* [1], who describe an 18 ns MD simulation of staphylococcal nuclease. From this simulation, the authors were able to extract information on the range of motion of  $C_{\alpha}$ - $C_{\beta}$  and  $C_{\alpha}$ - $H_{\alpha}$  vectors in alanine residues, showing that their order parameters are very similar. Although this is physically reasonable, it contradicts experimental results derived from NMR relaxation studies of staphylococcal nuclease. This led to questions concerning the accuracy of the measurement technique, in order to explain the discrepancy. This is a first for theory, which is normally subordinate to experiment in structural biology.

Functionally important motions often occur along the direction of a few collective motions. This was shown for the T4 lysozyme, for which hinge-bending motions, characterized by the two largest amplitude collective modes, were found to be similar to variations of conformation observed in a different crystal environment. Kitao and Go review the progress made in extracting these collective motions, as well as their applications to investigating protein dynamics, to sampling of the conformational space and to the analysis of experimental data.

Collective variable approaches are still making important contributions to the field of dynamics simulations of nucleic acids. Their main advantage is still the significant decrease in the number of variables. In conjunction with the use of simplified models, they offer ways of looking at very large nucleoprotein complexes. These issues have been carefully reviewed by Lafontaine and Lavery (pp 170–176).

Hansmann and Okamoto (pp 177–183) describe applications of new Monte Carlo techniques for studying the conformational space that is accessible to a protein. For many years, the emphasis in protein folding studies has been set on finding the conformation with the global minimum potential energy. Recently, these interests have broadened to include more global knowledge of the phase space, including intermediate states and denatured states. Monte Carlo methods have proven to be valuable for this purpose, leading to the new algorithms, as well as the new powerful optimization techniques, that are reviewed by Hansmann and Okamoto.

Traditionally, potential energy functions are derived from general physical concepts and the folding of a protein is expected to be a natural consequence of its properties. New approaches for designing potentials have appeared that either extract information from the growing database of known protein structures (leading to statistical or knowledge-based potentials) or optimize the potential parameters in order to meet the foldability criteria, such as the presence of a large stability gap between the native and unfolded states of a protein. In order to be successful, any of these potentials should provide the correct shape for the free energy landscape of a protein. Simplified pair potentials for lattice-type protein chain models have been shown to fail and current studies focus on identifying the quantitative factors that affect the energy landscape. Hao and Scheraga (pp 184–188) provide a thorough review of these issues.

Recent experimental results suggest that protein folding mechanisms are largely determined by the overall topology of the native state of the protein and are relatively insensitive both to sequence and to the fine details of interatomic interactions. In particular, Alm and Baker (pp 189–196) review a series of experiments in which a significant correlation was found between the folding rate of a protein and the average separation along the sequence of residues that are geometrically close in the three-dimensional structure (contact order).

This is good news for theory, as it suggests that folding can be explained in terms of simple physical principles.

The reviews by Alm and Baker, and by Thirumalai and Klimov (pp 197–207) both provide examples of the meeting of theory and experiment. In particular, for proteins that fold efficiently, following a two-state model, the process of collapse and the acquisition of the native structure are nearly simultaneous; the same behavior was observed in simulation studies of the nucleation/collapse mechanism in minimal off-lattice models.

Knowing the structure of a protein is essential to understanding its function. It is expensive to determine experimentally the structure of every protein, however, and this issue is becoming increasingly important with the completion of several genome projects. This has led to the development of structural genomics [5], which focuses on determining the structures of a well-chosen subset of proteins that should put all other protein sequences within the range of comparative modeling. This was discussed at a recent meeting held in Avalon, New Jersey [6]. The results of the latest Critical Assessment of Structure Prediction (CASP) meeting, held in December 1998 [4], confirmed that it is possible to build a reasonable model when a proper template can be identified. Fischer and Eisenberg (pp 208–211) review recent work on the latter aspect, namely the computational assignment of folds to genome sequences. They show, in particular, that current techniques are presently assigning up to 30% of all sequences and that this number should cross the 50% barrier by the year 2003.

From these reviews, it is clear that simulation studies are progressing towards their goal of providing a picture of the properties of macromolecules in solution. This progress is a consequence of both the increase in computer power and the development of new techniques for sampling and minimization. It should be mentioned also that there is an increased awareness in the field of structural biology that solutions to the difficult problems we are facing will come from combined efforts among the thousands of scientists working in this field.

## References

1. Chatfield DC, Szabo A, Brooks BR: **Molecular dynamics of staphylococcal nuclease: comparison of simulation with  $^{15}\text{N}$  and  $^{13}\text{C}$  NMR relaxation data.** *J Am Chem Soc* 1998, **120**:5301-5311.
2. Duan Y, Kollman PA: **Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.** *Science* 1998, **282**:740-744.
3. Caves LSD, Evanseck JD, Karplus M: **Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin.** *Protein Sci* 1998, **7**:649-666.
4. Koehl P, Levitt M: **A brighter future for protein structure prediction.** *Nat Struct Biol* 1999, **6**:108-111.
5. Kim SH: **Shining a light on structural genomics.** *Nat Struct Biol* 1998, **5**(synchrotron suppl):643-645.
6. Sali A: **100,000 protein structures for the biologist.** *Nat Struct Biol* 1998, **5**:1029-1032.