New Challenges in Computational Biochemistry

B. HONIG
Department of Biochemistry and Molecular Biophysics
Columbia University
630 West 168 St.
New York, NY 10032

The application of physical chemical methods to the simulation of biological macromolecules has become an important area of scientific research (which I will term computational biochemistry). Molecular dynamics simulation methods, continuum electrostatics and a variety of empirical solvation models have become widely used tools in physical biochemistry and structural biology. Algorithms and computer programs that have been developed for the purposes of simulation have become an integral part of the methodological repertoire of experimental scientists in these fields. In biologically oriented research, the adaptation of theoretical methods and concepts by experimentalist is often taken as a sign of success and by this criterion, theory has indeed had significant impact. The existence of software companies that market the wares of computational biochemists provides further evidence of the acceptance of simulation methodologies. This is a non-trivial accomplishment given the fact that as recently as ten to fifteen years ago, the results of simulations on proteins, membranes and nucleic acids were not taken very seriously by the experimental community.

The successes of computational biochemistry have resulted from a number of factors. First and foremost in my view was the willingness of many investigators to defy the conventional wisdom which posited that large molecules could not be gainfully studied until simpler systems could be better understood. From a practical standpoint, essential accomplishments included the development of potential functions that could effectively mimic physical reality and the parallel development of simulation methods that could relate these potential functions to physical observables. Improved representations of the aqueous phase, ranging from atomic level descriptions to continuum approximations, have played a crucial role in this process. Another factor has been the ability of workers in the field to relate their results to experimental questions that were under concurrent investigation in the larger biochemical and biophysical communities. In this way, the results of simulations have been extensively tested and refined. The many successful correlations between theory and experiment that have been reported in the literature testify to the effectiveness of this approach as does the integration of many theoretical concepts into the every day language of biochemistry and structural biology.

Despite the enormous progress that has been made, a number of challenges have been elusive. Perhaps the most glaring of these is the inability of existing

methods to consistently predict the relative binding free energies of different substrates to the same protein. A number of impressive successes have been reported but, in a general sense, the problem remains unsolved. The same can be said about the prediction of the relative free energies of different protein or nucleic acid conformations. For example, the prediction of the conformation of loops which connect two fixed secondary structure elements is an extremely important problem for which anecdotal successes have been reported but for which no general solution is available. It is useful to consider possible sources of the difficulties inherent in these problems but before doing so, it is of interest to consider challenges of a very different nature to the field of computational biochemistry. These come from computational biology at the one extreme, and combinatorial chemistry at the other.

Combinatorial chemistry poses a "threat" of sorts to the entire field of structure based drug design. As has been widely discussed, the ability to simultaneously test a large number of different compounds for binding affinities reduces the need for the precise design of a compounds that binds tightly to a specific target. With regard to computational methods, one might inquire as to the point of carrying out a complicated calculation of a binding free energy when, at least in principle, it is straightforward to test thousands or even millions of potential substrates. Clearly the "binding problem" will remain one of great theoretical interest but is it of practical interest as well?

The challenge of computational biology arises from an entirely different source. The vast quantity information now accumulating in biological databases provides an approach to the prediction of biological structure and function that in many ways bypasses approaches based on physics and physical chemistry. A clear example is provided by the protein folding problem. Despite the range of methods that have been applied to understanding the physical and chemical principles upon which protein folding is based, the fact remains most of the successes to date in actual fold prediction have resulted from database mining of one type or another. It is somewhat sobering for example that despite the extraordinary progress that has been made in understanding the physical basis of secondary structure formation, the most successful secondary structure prediction methods are based solely on database analysis, for example using neural networks. The same can be said about tertiary structure prediction based on "threading" and 3-D profile fold recognition methods.

One response to these comments is to point out that not all science need be immediately practical. Database mining does not necessarily lead to the type of deep understanding that is satisfying to physical chemists while the existence of combinatorial chemistry does not really detract from the interest and importance in elucidating the principles of molecular recognition. On the other hand, there are many practical elements to academic science of which we are all aware; such as the difficult in obtaining funding, or in finding jobs. I find it difficult to ignore the enormous disparity in job availability for individuals trained to carry out comparative analysis of amino acid sequences and individuals trained in simulation

methods. Both fields involve computational research but the former has become so "hot" that the absence of trained researchers has become a frequent subject of commentary in Science and other journals.

How can computational biochemistry respond to these theoretical and practical challenges. The difficulties in calculating accurate binding and conformational free energies may arise from a number of sources. First, it is possible that available potential functions are simply not good enough. There is some evidence to this effect; for example it has been shown recently in a number of studies that the relative solvation free energies associated with a number of functional groups cannot be explained with existing force fields. The problem seems to result from non-electrostatic contributions to hydrogen bonding which are difficult to reproduce with the reliance of all widely used force fields on atomic partial charges. Moreover, it is essential to find a way to properly account for electronic polarizability. There has been much effort in this direction for some time but a new generation of "polarizable force fields" has not yet emerged.

A second major problem concerns conformational sampling. There is a need to develop methods that effectively sample conformational space based on a physically meaningful energy function. The integration of continuum solvent representations and molecular mechanics methods may prove useful in this regard and has indeed been attempted on a number of occasions. However here again, much work remains to be done. Of course, there may be problems that are simply intractable. If for example, the difficulties in calculating relative binding free energies are due in large part to different entropy changes distributed over an entire protein associated with the binding of different substrates, the prospects for a general solution to the binding problem are dim.

It is likely that much theoretical effort will be devoted to these problems in the coming years and it is likely that the effort will be worthwhile. However, even in the absence of methodological progress, there are clearly many ways to exploit existing techniques in computational biochemistry in the study of biological macromolecules; for example in the analysis of the structure and function of the large numbers of proteins and nucleic acids whose three dimensional structures have been determined. Enzyme mechanisms, the nature of ligand induced conformational changes, the effects of mutations on protein stability and protein folding pathways are examples of problems of great current interest that will not soon disappear. On the other hand, computational biochemistry may also have answers to the two practical challenges mentioned above.

Regarding the challenge of combinatorial chemistry, a more optimistic view is that this technology effectively relieves the pressure to calculate binding free energies at better than kcal/mole accuracy. Rather, developing approximate methods to screen combinatorial libraries based on their ability to bind to a specific target or to possess properties similar to a known pharmacaphore has become an important

goal. This in turn requires an accurate description and understanding of the properties of isolated molecules in aqueous solution and existing methods are already quite successful in this regard.

An optimistic view of the challenge of biological database mining suggests even greater opportunities for computational biochemistry. Specifically, in the absence of a well-defined physical model, statistical methods used in the analysis of databases, though extremely powerful, are ultimately limited in their ability to extract meaningful information. Their effectiveness can be enormously enhanced if they are combined with the tools and insights of computational biochemistry. As an example, it seems clear that threading methods could be significantly improved if the scoring functions they use to distinguish correct from incorrect protein folds were based on meaningful physical rather than statistical potentials.

The integration of computational biochemistry and computational biology offers many exciting opportunities. One can envision the day where the process of multiple sequence analysis, structural prediction, the design of combinatorial libraries and binding free energy calculations will be carried out in a single group by researchers who understand the intricacies of each of these problems. For this to occur will require new training modes for graduate students and postdoctoral fellows and, perhaps, new modes of thinking for their more senior colleagues.