

*This paper is the introduction to the following papers, which were presented at the colloquium "Computational Biomolecular Science," organized by Russell Doolittle, J. Andrew McCammon, and Peter G. Wolynes, held September 11–13, 1997, sponsored by the National Academy of Sciences at the Arnold and Mabel Beckman Center in Irvine, CA.*

## Computational biomolecular science

PETER G. WOLYNES

School of Chemical Sciences, University of Illinois, Urbana–Champaign, Urbana, IL 61801

In this century, the study of the *molecules* of life has transformed the practice of biology as a whole. Molecular thinking now influences the research agenda for scientists studying both the behavior of individual cells and organisms, and the relationships between organisms as in natural history. Even ecology and anthropology are being influenced by this molecular revolution. It is impressive that this transformation has, to a large extent, been made possible by simply identifying (with very clever strategies!) active biological molecules and cataloging their information content through their sequences. One result of all this activity is that raw data about life at the molecular level have become abundant, but understanding its biological meaning remains, in many if not most respects, perplexing. Fortunately, just at this stage, new approaches to understanding the connection between biomolecular sequence and physiological behavior are coming forward. Computation, theory, and novel experimental approaches that utilize the combinatorial power of the genetic code allow us to begin to understand biomolecular function from both the bottom-up atomistic point-of-view of the physical sciences and the top-down view usually associated with the evolutionary perspective.

The goal of this colloquium was to bring together some of the workers from different scientific disciplines who are approaching these problems by using quantitative methods. Because computation plays such a large part in exploiting the information content of sequence data, the conference was entitled "Computational Biomolecular Science," although some of the essential input of new experiments to this emerging discipline was covered too.

From the bottom-up perspective, the first event to consider on the road from sequence to the biological behavior of an organism is the folding of a linear polymer into a three-dimensional structure. Once a molecule is properly folded, a variety of motions still go on in the folded state. It is through these motions that the biological molecule can function. These dynamical aspects represent complex problems in chemistry and physics. But it is the aptness with which these functions are carried out that at last determines whether the organism containing that molecule can survive in the struggle with other organisms. Quantitatively understanding molecular behavior sufficiently well for understanding this final biological goal requires much work from both the theoreticians and the experimentalists.

The top-down interpretation of molecular data appears to proceed quite differently. Avoiding the complexity of molecular theory, the evolutionary perspective takes inheritance, perhaps the most self-evident aspect of "living" things, as its central concept. Comparing sequences between different organisms then provides clues to their molecular function. In this study, dominant use is made of features of molecules that do

not change an organism's fitness, thus allowing markers of inheritance to be reliably assigned. In a sense then the non-functional parts of a molecule's structure and dynamics are the most useful to the phylogenetically inclined scientist. Convergent evolution is hard to establish by such studies but is critically important to those who wonder whether, from the atomistic perspective, there are indeed general themes to the scheme of life. Despite its sometimes "life as a blackbox" character, the top-down viewpoint has achieved a myriad of successes in the practical applications of biomolecular science.

A gap exists between the two different vantage points of looking at biomolecular information, but there are a surprising number of common concepts. In understanding the folding, motions, and function of biological molecules, for example, a powerful new viewpoint that describes the entire energy landscape of a biomolecule in a statistical fashion is proving essential. Understanding and differentiating between those parts of the energetics and dynamics that are biologically significant and those that can be thought of as random noise is the hallmark of this approach. Similarly, in the comparative top-down approach to understanding sequence data, a tremendous amount of statistical thinking must be done to understand whether a perceptible similarity between two sequences really means the molecules have comparable function or structure or whether the similarity is just an accident. Just as in energy landscape theory, extracting signal from noise is the crucial point to understanding molecular evolution. Such frankly statistical viewpoints must also be brought together when planning modern molecular biology experiments that now begin to allow the study of a huge number of variants of a biomolecule in the laboratory simultaneously at one time.

It became apparent in the meeting that, apart from the general common interest in biomolecules and the common but general theoretical concepts based on statistics, there were many specific problems where the top-down and bottom-up viewpoints can profitably be merged. For example, surveys of genomes reveal widespread structural themes that may be clues to folding thermodynamics and kinetic folding routes. For the atomists, several studies show how the structures of specific sequences can be predicted if knowledge of the sequences of many widely different but evolutionary related molecules is available. On the other hand, for the evolutionist, an *a priori* knowledge of structural and energetic patterns in molecules leads to refined algorithms for comparing sequences to obtain reliable phylogenies. Also, convergent evolution can be recognized if both comparative and physical studies are available for proteins in the same family. This breaks evolutionary explanation out of the mold of sophisticated Kipling "just-so" stories into the quantitative mode, most prized by natural scientists.

The papers in this colloquium give a partial snapshot of computational biomolecular science today. The organizers of the meeting, J. A. McCammon, R. F. Doolittle, and I, hope these papers give the readers of the *Proceedings* an idea of what is going on in a branch of science that is destined to grow much larger in the coming years.