Protein Folding and the IBM BlueGene Supercomputer

Shawn Douglas

December 13, 2002

Abstract

Understanding how proteins fold is an important step in uncovering the mysteries of protein function. To date, the computational resources needed to simulate even the shortest and simplest protein folds have been astronomical. Simulation of larger and more complex folds has been prohibitively expensive. In 1999, IBM announced the start of an effort to build a supercomputer, the BlueGene, to be applied to biomolecular problems such as protein folding. This paper provides an overview of the problem of protein folding, a description of IBM's BlueGene project, and what we may hope to gain once the machine is fully operational.

1 Introduction

Proteins are the most versatile macromolecules in living organisms and are essential to nearly all biological functions [1]. Understanding how proteins function is necessary to comprehend the cellular processes in which they participate. The mechanisms of catalysis, molecular transport, locomotion, immune response, signal transduction, cell growth, differentiation, and reproduction are all the result, in whole or in part, of protein interactions. When these processes malfunction, they can lead to illness or death. If we hope to design drugs to remedy such malfunction, our best bet is to focus on studying the underlying mechanisms at the level of protein function.

We know that the function of a protein is inextricably related to its structure [2]. Effective drug treatments often target specific proteins. As we continue to solve three-dimensional structures for proteins, it is possible to use the these structures as a basis for rational drug design [3]. It is relatively easy to determine the amino acid sequence of a protein. Unfortunately most proteins do not easily lend themselves to the art of X-ray crystal structure determination. The bridge between a freshly translated, unfolded protein and its functional, three-dimensional native structure (aside from post-translational modification) is, of course, folding. One possibility for molecular simulations is to start with only the primary structure of a protein, simulate its folding pathway, and end up with a close approximation of its native three dimensional structure. Such simulations would not be hindered by the same difficulties encountered in the laboratory by crystallographers. Additionally, some diseases are caused by protein misfolding, such as mad cow disease [4] and cystic fibrosis [5]. So in addition to gaining insight into the folding of proteins that we would not otherwise be able to crystallize, fold simulations can offer insight into treatment of folding-related illnesses.

The scientists at IBM Research have set out to accomplish two primary goals as part of the BlueGene project. First, they hope to advance the state of the art of biomolecular simulation. The second aim is to advance the state of the art in computer design and software for extremely large scale systems [6]. These are substantial endeavors, and here I hope to provide some context to these statements as well as the significance of their realization.

2 The Protein Folding Problem

Proteins are linear polymers of amino acids. Amino acids consist of a central carbon atom linked to an amino group, a carboxylic acid group, a hydrogen atom, and a side chain. The side chains, twenty of which occur most prominantly in nature, vary in shape, size, polarity, and charge. There are several ways to model proteins and protein folding, each with varying levels of complexity. As Levinthal noted, the fact that proteins fold reliabily and quickly despite the extremely high number of possible conformations indicates that pathways are involved [7]. A significant challenge in the study of protein dynamics is to understand these pathways.

One way to view protein folding involves characterizing the amino acid side chains according to charge and hydrophobicity. In the simplest picture, the folded protein is arranged in such a way that hydrophobic side chains are positioned near the core, while the charged and hydrophilic side chains are near the surface. The stability of the protein can be described in terms of Gibbs free-energy change ΔG.[1] The idea behind a folding simulation is to start with the unfolded polypeptide chain and, in infinitesimal timesteps, iteratively calculate the energetics associated with each atom and assign each atom a new position according to the most favorable change in free energy.

It should be noted that the BlueGene effort is focused on the exploration of the problem of protein folding, which is not precisely the same as tertiary structure prediction. The general folding simulation described above, if only examined for its end result, may be considered a kind of tertiary structure prediction called *ab initio* prediction. This name implies that there is no reliance on known structures in the prediction - it is solely based on molecular dynamics. In practice, most tertiary structure prediction methods make extensive use of information from X-ray crystallography or nuclear magnetic resonance (NMR) experiments. The method of "homology modeling" is based on the idea that when sequences are similar, structures will typically share this correspondence. Several other methods are used, independently or in conjunction, such as fold recognition and threading [8]. Given the amount of work that is currently invested in alternative methods of protein structure prediction, it has been deemed "unnecessary" to spend a petaflop-year on the prediction of a single protein structure [9].

3 IBM's BlueGene Project

In 1999, IBM announced at 5 year, $100 million effort to build a petaflop scale supercomputer, the BlueGene/C. Two years later, IBM also announced a second machine based on similar technolgy, BlueGene/L. BG/L is expected to be completed in 2004, about 18 months before BG/C. By the time BG/C is fully operational, IBM can expect to have the first and second fastest supercomputers in the world.

To provide some frame of reference, it is useful to examine the current offerings in high performance parallel computing. The most obvious example is the "Beowulf Cluster." These machines are built from commodity hardware components for a fraction of the cost of traditional supercomputers. An index of the fastest Beowulf clusters is available online.[2] With 5120 processors, the fastest Beowulf cluster is the Earth Simulator supercomputer installed earlier this year in Yokohama, Japan, and has been benchmarked at 35.86 Tflop/s. The combined performance of all 500 computers on the list (222161 processors total) is 293 Tflop/s - less than one third the performance IBM expects from BlueGene.

Assuming the goal is one million processors, the space, power, and financial resources needed to implement a 1 Pflop/s Beowulf cluster make such a project probitively expensive at the current time. In other words, if IBM has any hope to achieve such performance, the machine's architecture must be vastly different from that of typical personal computers or clusters.

[1] See reference [9] for a more detailed explanation.
[2] http://www.top500.org/

So how will BlueGene achieve such phenomenal processing power? The key advance is in designing a chip that integrates multiple processors with their memory and interchip communication logic. With 32 processors per chip, IBM plans to package 64 chips into each board; then eight boards will be stacked in each rack. BG/C's 64 racks should add up to approximately 1 Pflop/s of processing power. BG/L will have a slightly more simplified design: Thirty two 2-processor chips per board, 32 boards per rack, and 64 racks, yielding a grand total of 65,536 nodes that will perform at approximately 200 Tflops/s [10].

4 What will BlueGene be used for?

As mentioned earlier, BlueGene will *not* be used to study a single folding event for a single protein. Instead, wide variety of protein and smaller peptide systems will be studied. Simulations will be run with a high degree of replication in order to gather meaningful statistics [9]. It is hoped that other types of simulations that will be conducted may be useful in refining other methods of structure prediction.

4.1 Pathway Characterization

Folding pathway characterization typically involves optimizing free energy along folding pathways using sophisticated thermodynamic sampling techniques. A free energy "landscape" is mapped out as the protein moves through different conformations during the folding process. Data from this type of simulation may be useful in understanding intermediate states along the folding pathway.

4.2 Folding Kinetics

Folding kinetics studies are designed with the goal of understanding and predicting rates at which a folding protein makes transitions between various conformations. Previous work typically has done calculations with thermodynamic averages, but here we hope to repeatedly simulate the actual dynamics of many systems in order to derive some estimates for actual rates of various fold transitions. One significant result of these studies will be the comparison of observations with recent theories about folding kinetics.

4.3 Force Field Characterization and Solvent Models

Researchers also plan to examine structural stability of experimentally generated native structures using different force field environments. For example, in addition to simply varying the environment and temperature, known protein structures may be partially unfolded in simulation with heat pulses; observations of refolding can provide insight into ability of force fields to correctly reproduce free-energy minima.

Much attention has also been given to solvent models used in molecular simulations. If we are to have any hope of modeling protein structures at they occur *in vivo*, we must first be able to provide an accurate simulation of their natural environments [11].

The C terminus B-hairpin of protein G has been studied frequently due to its small size and fast folding ($\approx 6\mu$s). Theoretical work has been done using both explicit and implicit solvent models, often with quite varied results in the resulting free-energy landscapes. Simulations using explicit solvent models (keeping track of every atom) require enormous amounts of CPU time, so many past studies have used continuum (implicit) solvent models. Recent studies using protein G have found that free energy landscape from continuum solvent models strongly favor non-native folding states [12]. Most significantly, the lowest free energy state is no longer the native β-strand structure. It is expected that BlueGene will be able to provide further our depth of understanding of this and related areas of study.

4.4 Implementation and Analysis of Novel Algorithms

This is an intuitive, but noteworthy topic of consideration. The outstanding advance in computational power that will be applied to protein science will undoubtedly foster further creation and development of techniques and algorithms for studying biological processes such as protein folding.

5 Conclusion

Software always expands to exploit advances in hardware, and as previously impossible pursuits are made possible by lifting technical constraints, the limitations are instead shifted to the scientist's imagination. The BlueGene project represents a unique opportunity to witness a significant step forward in the cutting edge of high performance computing, as well as in our basic understanding of one of the most fundamental units of life, the protein. It will be very fascinating to observe the new directions taken, both scientifically and computationally, as a result of this amazing effort.

References

[1] J. Berg, J. Tymoczko, L. Stryer, *Biochemistry - Fifth Edition* W.H. Freeman, New York, (2002)

[2] A. R. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, W.H. Freeman, New York, (1998)

[3] TL. Blundell, *Structure-based drug design*, Nature 1996 Nov 7;384(6604 Suppl):23-6

[4] J.Collinge, *Variant Creutzfeldt-Jakob Disease*, The Lancet 354, No. 9175, 317-32 (1999)

[5] E. Strickland, B.-H. Qu, and P.J. Thomas, *Cystic Fibrosis: A Disease of Altered Protein Folding*, Journal of Bioenergetics and Biomembranes 29, 483-90 (1997)

[6] IBM Research, *Blue Gene Project Update* Jan 2002, Available at http://www.research.ibm.com/bluegene/

[7] C. Leventhal, *Are There Pathways for Protein Folding?*, Journal de Chimie Physique, 65, 44-45 (1968)

[8] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard, *Critical Assessment of Methods of Protein Structure Prediction (CASP): Round IV* Proteins: Structure, Function, and Genetics Suppl 5:2-7 (2001)

[9] F. Allen et al., *BlueGene: A vision for protein science using a petaflop supercomputer* IBM Systems Journal, Volume 40, Number 2, 2001:310-27

[10] N.R. Adiga et al., *An Overview of the BlueGene/L supercomputer* Supercomputing 2002 Technical Papers, November 2002. Available at: http://www.sc2002.org/paperpdfs/pap.pap207.pdf

[11] M. Gerstein and M. Levitt, *Simulating Water and the Molecules of Life* Scientific American 279:100-105, 1998.

[12] R. Zhou and B. Berne, *Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water?*, Proc Natl Acad Sci U S A. 2002 Oct 1;99(20):12777-82.