Overview of molecular dynamics techniques and early scientific results from the Blue Gene project

F. Suits
M. C. Pitman
J. W. Pitera
W. C. Swope
R. S. Germain

The Blue Gene® project involves the development of a highly parallel supercomputer, the coding of scalable applications to run on it, and the design of protein simulations that take advantage of the power provided by the new machine. This paper provides an overview of analysis techniques applied to scientific results obtained with Blue Matter, the software framework for performing molecular dynamics simulations on the Blue Gene/L computer. Blue Matter is a portable environment that runs on several platforms ranging from single-processor to massively parallel machines. Since the Blue Gene/L computer has become available only recently, this work describes analysis techniques applied to a range of experiments of increasing complexity on a corresponding range of machine sizes, concluding with a membrane protein simulation currently running on a 512-node Blue Gene/L computer.

Introduction

The Blue Gene* project was originally conceived in 1999 with the goal of building a petaflops computer to address the grand challenge problem of protein folding [1]. Since then, the machine design has evolved, as have the scientific objectives, but the goal of gaining insight into protein science and the mechanisms behind protein folding remains. We chose classical molecular dynamics as the method for simulating protein systems and designed a software framework—called *Blue Matter*—for doing these simulations on highly parallel machines consisting of thousands of nodes [2].

It is important to emphasize that the Blue Gene project involves not only the development of a powerful, highly parallel computer, but the design of software and mathematical approaches [3] that take advantage of this power, and the application of the software to do research in protein science. Since the Blue Gene/L (BG/L) hardware has been available for production only within the last year preceding this writing, this paper includes scientific results obtained on traditional IBM computers and describes some of the techniques for validating the software as it was developed. In short, a great deal of BG/L science was done prior to the arrival of the computer hardware, both in the development of a highly parallel application framework and in molecular simulations performed on conventional hardware. In this paper, we describe the progression of experiments performed with the Blue Matter framework, with an emphasis on the analysis applied to the results. The intent is to provide general insight into the range of studies performed, without requiring a detailed protein science background.

Background on molecular dynamics simulations

Biomolecular systems, which comprise one or more molecules of biological interest surrounded by some amount of solvent (e.g., a protein in water), can be studied by a variety of computational methods. One of these is molecular dynamics, which simulates the movement of all of the particles of a molecular system by iteratively solving Newton's equations of motion. This calculation is based on the instantaneous coordinates of all of the particles of the system to evaluate their energies and forces of interaction. Given the coordinates, velocities, and resulting forces, one can compute the coordinates and velocities that the particles would have a short time later. This process is then repeated many times, yielding the motion of all of the particles over the time of the simulation. The small timesteps are usually of the order of one femtosecond (10^{-15} s), and a typical simulation might perform ten million of these steps,

0018-8646/05/$5.00 © 2005 IBM

resulting in a simulation of 10 nanoseconds (10^{-9} s). The maximum size of the small timesteps is limited by the fastest motions in the system, which, for biomolecules, corresponds to the vibrations between chemically bonded atoms.

Molecular dynamics allows one to monitor the simulated system as it moves from one conformational state to the next and to deduce the timescales for those transitions compared with experimental data [4]. In addition, a number of variations on molecular dynamics allow one to also simulate the behavior of a molecular system under different conditions of temperature, pressure, and other parameters [5]. These variations allow the simulation of biomolecules in environments that correspond to typical experimental conditions. Under these conditions, properties such as the relative populations of two conformations can be determined from the simulation and compared directly with experimental data. A drawback of these variations is that they alter the temporal evolution of the simulated system, so they cannot be used to directly compare time-dependent phenomena between simulations and experiments. Two important forms of molecular dynamic simulations are *thermodynamic* and *kinetic*. Both techniques are important and provide complementary insight into protein structure and dynamics, and both are applied in the Blue Gene science program.

Thermodynamic questions involve the structure and stability of conformational states of proteins. To carry out their biological function, most proteins must be in a specific folded state. This state is generally their "native state," the most thermodynamically stable conformation for the protein under biologically relevant conditions of temperature, pressure, and pH. As a result, thermodynamic studies of protein folding attempt to answer questions such as the following:

- *What is the most stable structure for the protein at a given temperature?*
- *Is there only one stable structure, or are there multiple ones with comparable stability?*
- *How do the populations of these stable states change with temperature?*

Experimental techniques, such as X-ray crystallography and nuclear magnetic resonance spectroscopy, provide a partial picture of the folding thermodynamics, which can be complemented by detailed simulations [6]. By understanding what governs the stable states of proteins, we will have the potential to design proteins of novel structure [7] and function [8].

Kinetic questions are concerned with time-dependent phenomena: specifically, the questions of how proteins fold so quickly, how much time they spend in intermediate states along the folding pathway, and what interactions govern the overall rate of folding. The first question arises from the *Levinthal paradox* [9]—there are an astronomical number of possible conformations for a flexible protein chain [a chain of m residues could have $O(10^m)$ distinct conformations]—but only a few correspond to the folded state. If the folding process involved some sort of random motion through this space of conformations until the folded state was "found," proteins would fold much more slowly than observed. Therefore, specific chemical interactions, along with topological constraints, must guide the protein to fold faster than it could by a simple random search. As it folds, the time each protein spends in intermediate states is also important. Intermediate states may be partially folded or misfolded and hence prone to degradation [10] or aggregation [11], both of which can affect the ability of the protein to function, or which may themselves trigger disease [12]. Finally, a full understanding of kinetic phenomena in protein folding must enable us to alter those phenomena in a deliberate fashion. By determining which specific interactions control the overall folding rate, we can design proteins that fold more rapidly or more slowly by making specific mutations. Rapidly folding protein variants could have therapeutic uses [13] or serve as functional or structural nanomaterials.

Simulation output from Blue Matter

The first step in analyzing results from a simulation is the retrieval of data from the running program. The traditional view of a computer simulation writing values to a file becomes more complex when that computer contains thousands of independent processors working together on the same problem. To provide a more scalable solution that minimizes the need for synchronized, cooperative communication among the nodes, Blue Matter reports simulation results from individual nodes via binary packets sent over sockets. This allows individual nodes to report only their contribution to simulation observables, such as total energy, and relies on external analysis routines to organize and integrate the packets to determine the full simulation state over time. Although there are many observables to study and many modes of analysis, they all begin with the examination of a raw datagram stream containing a sequence of packets representing partial quantities from the simulation. Note that the frequency of the output for various quantities can be set so that it is appropriate to the need. For example, energy terms are usually output much more frequently than the full positions and velocities of the atoms, which represent a much larger volume of data.

Numerous forms of analysis can be performed on the output of a molecular simulation, but many of them

represent a reduction of a large amount of information to a simpler form that is easier to interpret. Sometimes this process involves a visualization component based on two-dimensional (2D) or 3D representation; at other times, it is a strictly quantitative reduction of a value that can be matched to experiment. One of the simplest 3D representations of a system is a direct visualization of the configuration, which can be useful both as a check that the system is behaving properly and to provide insight into dynamic processes occurring either within or between molecules. There are many examples of familiar single-value reductions of a molecular system, such as total energy, temperature, and pressure, but there are additional quantities that are specific to a system, such as the number of native hydrogen bonds. Each of these reduced forms can then be studied over time, allowing additional analysis on the time series to be performed, such as autocorrelations. Specific examples of these reduction and visualization techniques are described in more detail in the sections that follow.

With the advent of very-large-scale cellular computer systems such as BG/L, there is the prospect of producing massive volumes of data. Molecular dynamics simulations on a 512-node BG/L partition currently generate approximately 6 GB/day of data (2.2 TB/year) when taking a snapshot of the simulation state every picosecond of simulation time. On a 64-rack system, the extrapolated volume of data might be 300 TB/year, assuming linear scaling of the data volume because multiple simulations could be running on the 64-rack system. There are also significant volumes of data generated by analysis, and there may be simulations that require more frequent sampling of the simulation data. Much of the data generated is "reference" data that is accessed for analysis and then put down without any further access. Although access to the raw data is unlikely to be needed, there is an understandable reluctance to actually discard it. Instead, we have adopted a hierarchical storage approach using Tivoli* Space Manager that migrates data to tape according to user-defined policies.

Analysis for code validation

Analysis is useful not only for scientific insight into the results of a simulation, but to help establish its validity and scientific correctness [14]. One example of this is the demonstration that the integrator, which determines the motion of the atoms on the basis of applied forces, is faithfully representing the forces and energies involved. This appears most directly in the conservation of energy in the system, but also in more subtle ways that can be recognized by running a simulation at different timestep sizes.

A molecular system under simulation usually has a conserved quantity, related to its total energy, that should not change over time.[1] Energy may change from potential to kinetic, allowing the system to undergo significant changes in its configuration and behavior, but the conserved quantity should remain constant. In actual simulations, the conserved quantity changes slightly on each timestep because of the finite timestep of the calculation, but there should be little or no drift over time. Thus, one of the first analysis tests for the validity of a simulation is to confirm that the conserved quantity is not drifting excessively. An additional test of the validity of the integrator involves running multiple short runs of the same simulation over the same time interval, but with a series of shorter timesteps. For short simulations with small timesteps, the change in energy over time should increase quadratically with timestep in a manner that is very sensitive to integration errors [15].

Figure 1(a) shows a plot of the conserved quantity during a short molecular simulation, along with its constituent energy terms. On this scale, the conserved quantity appears constant, and fluctuations of the individual terms appear to cancel out to keep the sum constant. **Figure 1(b)** shows a closeup of a series of conserved quantity plots for the same simulation, but at different timesteps, showing that the conserved quantity does vary slightly, and that the fluctuations increase with increasing timestep. Note that the vertical scale is greatly magnified to reveal fluctuations that are small on the scale of the actual conserved quantity. In this view, the curves appear roughly the same shape, but the quadratically increasing scale is not evident. **Figure 1(c)** shows the same plots, but with the vertical scale decreasing quadratically with the timestep size. This view allows a direct comparison of the behavior of the simulations over time, and the quadratic relationship is evident. There is a slight departure of the curves at later time due to the slightly different trajectory taken by the simulations, but this is a small effect on short simulations.

A more complex simulation involving constant temperature and pressure appears in **Figure 2(a)**, which shows abrupt changes in conserved quantity that appear to scale with timestep, although quadratic behavior is not directly evident.[2] When each plot is scaled quadratically with timestep, as shown in **Figure 2(b)**, the departure is immediately evident and linked to the code related to temperature control. Although this test does not guarantee that the simulation is working correctly, it is a

[1]The conserved quantity can be simply the total energy, in the case of a constant volume and energy simulation, or, more generally, it is the Hamiltonian, which can include terms related to temperature and pressure control. Some molecular dynamics simulations do not allow for the calculation of such a conserved quantity, but Blue Matter provides a conserved quantity for all of its simulation modes, which allows for important validation tests.

[2]Here, the conserved quantity includes terms for the energy transferred to the heat bath and the piston that controls pressure.

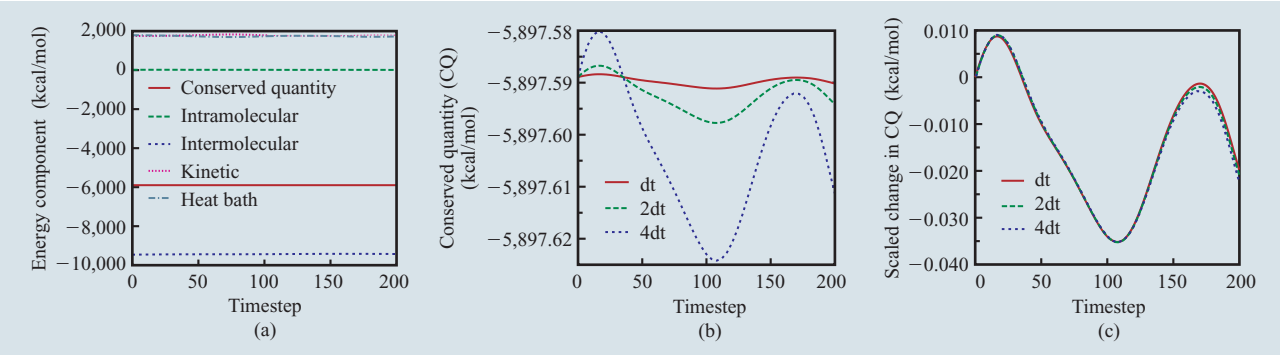

Figure 1

(a) Energy components during a short molecular dynamics simulation, with the conserved quantity equal to the sum of four constituent energy terms. On this scale, there is some slight variation in the individual terms, but the conserved quantity appears nearly constant, and the fluctuations in the terms amount to very little net change. (b) Closeup of the conserved quantity shown in (a) for three repeats of the same simulation, but with successively doubled timesteps. Note that the initial energy is identical in the three runs, but the fluctuations increase dramatically with timestep. Also note that the seemingly large fluctuations in this plot are closeups of the conserved quantity curve in (a), which does reveal structure on this finer scale. (c) Another plot of the information in (b), but with the energy change scaled by the square of the inverse timestep. If the energy fluctuations were proportional to the square of the timestep and if each simulation followed the same configurations in time, the three curves would line up exactly. Here they line up very well, indicating a properly implemented integrator, but with slight departure toward the end due to slight differences in configurations over time.

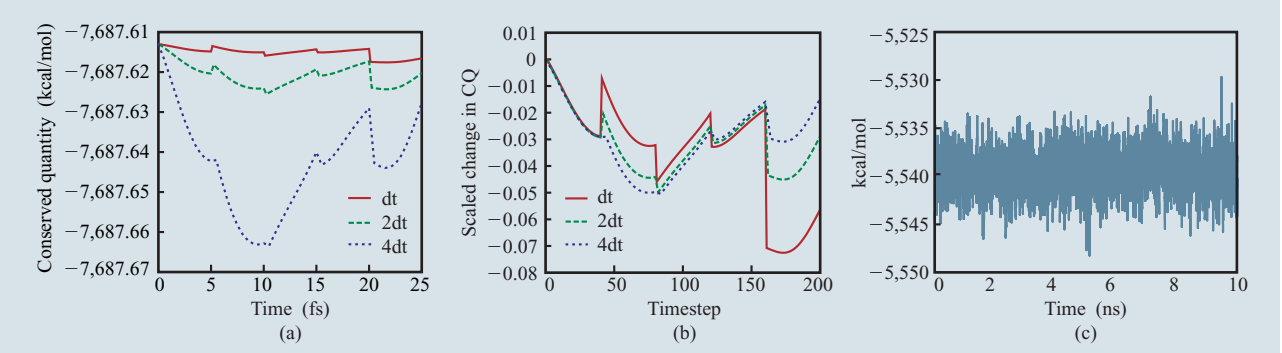

Figure 2

(a) Plot similar to Figure 1(b), but for a more complicated simulation involving temperature and pressure control. The periodic jumps are due to sudden randomizations of the atom velocities to maintain constant temperature. Despite the abrupt changes to the system, the conserved quantity should still show the quadratic behavior with timestep, and on this constant scale, it is difficult to tell whether this holds. (b) Quadratic scaling of the changes in conserved quantity, similar to Figure 1(c), reveals that the simulations are initially behaving quadratically with timestep, but the temperature control causes sudden changes that are not quadratic. This not only indicates that there is an error in the integrator but helps identify its location in the code. (c) Energy trace of a multinanosecond simulation shows virtually no drift over time. This is the result of careful analysis of the behavior of the conserved quantity as a sensitive probe of integration errors, and it provides confidence in the long simulations necessary for biologically relevant results.

sensitive test of subtle errors that would otherwise be hard to find, and it serves as a necessary, though not sufficient, sign of a correctly implemented algorithm.

The end result of these and many other validation procedures appears in **Figure 2(c)**, which is a plot of energy over time for a 20K-atom simulation of lipid molecules. The ability to do extended runs of large systems without significant drift of the conserved quantity provides confidence in the simulation results and allows long runs with constant energy to yield results that provide useful kinetic information, as described in the next section.

Protein kinetics study

As described in the Introduction, two common but very different ways to study a protein are via thermodynamics

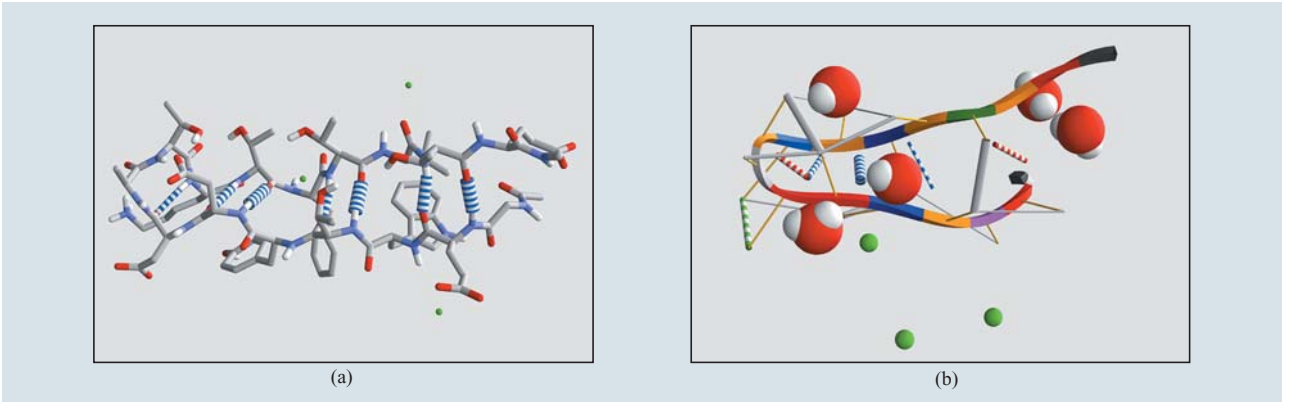

Figure 3

(a) Simple "stick" view of the beta-hairpin, but with striped tubes to show the hydrogen bonds. The hairpin shape is evident, with the hydrogen bonds linking the two "legs" together and the "turn" of the hairpin on the left side. This depiction includes all of the atoms of the protein except the nonpolar hydrogens, which do not play a role in hydrogen bonding. The green spheres represent the counter-ions of sodium in the simulation. The actual simulation has 1,660 water molecules surrounding the protein, but these are not shown here to avoid obscuring the hairpin. (b) A more abstract representation of the hairpin designed to convey some of the forces that define the configuration, with the protein shown as a ribbon colored according to the chain of amino acid residues. The green striped tube on the left is a salt bridge near the turn; native hydrogen bonds are blue, while nonnative are red, and both show a thickness proportional to the energy of the bond. The gray bonds represent sidechain–sidechain contacts that provide additional structural linkage. The five nearest waters are also shown.

and kinetics. Established methods exist to generate ensembles of proteins at a given temperature on the basis of molecular dynamics [16], and the resulting distribution of conformations represents the "canonical ensemble" at that temperature. Kinetic information tends to be more elusive, largely because the timescales of protein processes are so much longer (microseconds) compared with the timestep of the simulation (femtoseconds)—a factor of one billion.

As our first experiment with Blue Matter molecular dynamics, we chose a novel approach to determine kinetic information about a small protein by running many short, independent simulations in parallel in order to piece together the behavior of a typical protein over a long period of time. The technique involved starting a number of independent runs (237) of a protein from a wide range of initial configurations and tracking their changing shape over time. Different configurations were divided into bins, and the time spent in each bin, along with the probability of transitioning to other bins, allowed a Markov model of the folding process that, in theory, could predict the folding rate and the equilibrium (canonical) ensemble.

The protein in the study was the beta-hairpin from the C-terminus of protein G, which is a 16-residue peptide commonly used in protein simulations because of its rich folding behavior despite its small size [17]. An early step in analyzing a protein simulation is to determine observables that characterize the state of the protein with regard to its degree of foldedness. Two common observables are the number of hydrogen bonds formed and the radius of gyration, but many others can be defined [18]. Hydrogen bonds are part of what gives a protein its shape, and the folded configuration consists of a number of native bonds. An arbitrary protein configuration may consist of some native bonds and other non-native bonds that may be obstructing or assisting the path to the folded configuration. The radius of gyration captures the extent of the protein, with a small value indicating a compact structure. For the hairpin, the number of native bonds and the radius of gyration together define a 2D space into which a given protein configuration can be plotted.

Figure 3(a) is a rendering of a folded configuration of the beta-hairpin created with the Prototype Protein Viewer [19]—the visualization component of the Blue Matter framework. This is a standard "stick" view of the protein, but with a novel representation of the hydrogen bonds as striped tubes with thickness proportional to their energies. This is a relatively direct view of the molecule layout and provides precise location of the individual atoms, but it does not convey the underlying forces responsible for the shape, except for the hydrogen bonds. In contrast, **Figure 3(b)** shows the ribbon shape of the peptide pioneered by Richardson [20], along with novel representations of other structural components, such as salt bridges and sidechain–sidechain contacts. This is a reduced, abstract view of an otherwise opaque arrangement of atoms; it conveys in one image many of the forces dictating the shape of the protein and their spatial relationship. These representations were very

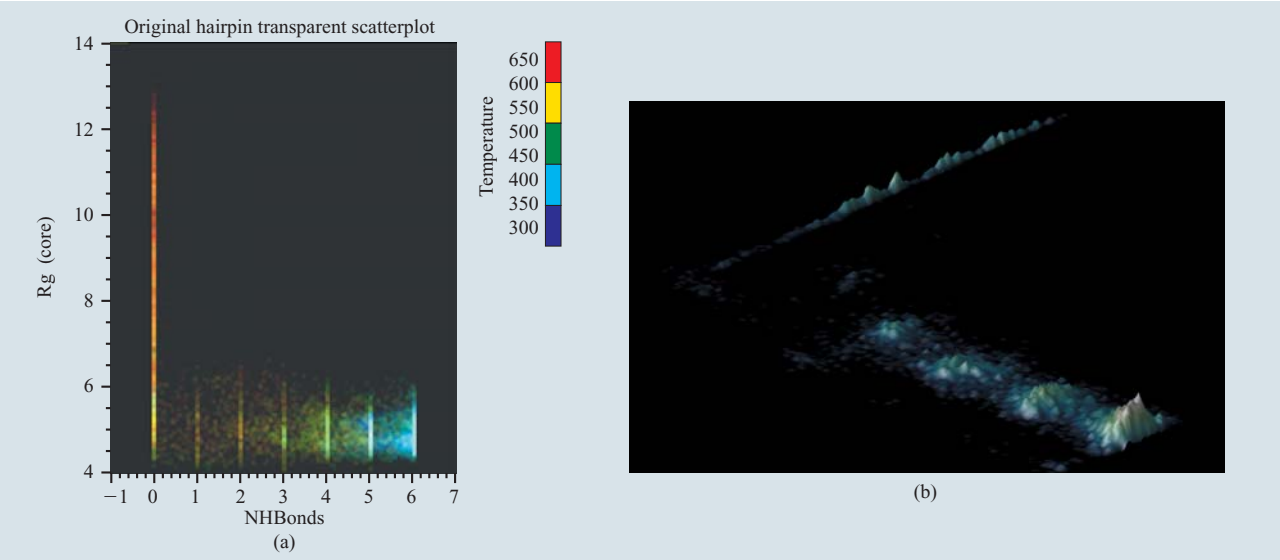

Figure 4

(a) Scatterplot of configurations from a replica-exchange simulation of the hairpin showing the increasing spread of points with temperature. Each configuration appears as a small transparent square colored according to temperature. The vertical axis represents the size of the central region of the hairpin; the horizontal axis corresponds to the number of native hydrogen bonds. The upper left is fully unfolded; the lower right is fully folded. (b) Free-energy landscape of (a) as a 3D surface based on data from a single temperature. The relative occupation of the bonded areas is evident, as is the separation of bonded from nonbonded configurations.

helpful in interpreting the many protein simulations performed for this kinetics study.

If one selects thousands of hairpin configurations randomly from an ensemble at a given temperature, a distribution appears as shown in **Figure 4(a)**. When this scatterplot is instead shown as a log histogram, it represents the "free-energy surface" at the given temperature and depicts the relative distribution of configurations in the canonical ensemble. **Figure 4(b)** depicts the resulting histogram as a surface in three dimensions, with height proportional to the log of the probability density. In each figure, the strong vertical bands are due to the successive numbers of native hydrogen bonds, and the L-shaped appearance captures extended, nonbonded configurations in the upper left and collapsed, fully folded configurations in the lower right. Note that the distribution of configurations indicates that the hairpins do not simply fold and remain folded, but instead are constantly changing from one configuration to another. However, this thermodynamic picture tells us nothing about how a given hairpin moves about this landscape. That perspective requires a kinetics study.

For the kinetics experiment, we selected 237 starting configurations distributed around the free-energy landscape and launched independent simulations using Blue Matter on an IBM SP* computer. Each simulation involved 256 protein atoms, 1,660 water molecules, and three counter-ions to balance the charge in the system.[3] (It is not uncommon for protein–water simulations to be dominated by the computational costs of the water molecules). As each of the protein systems evolved in time, its trajectory on the free-energy landscape captured the folding and unfolding process of an individual protein. **Figure 5** is a unique visualization that combines the kinetic data from the independent trajectories with the thermodynamic data from the canonical ensemble. By running all systems in parallel and starting them distributed around the free-energy surface, we were able to track the behavior of individual proteins in different sections of the landscape at the same time. This embarrassingly parallel technique was well suited to an IBM SP computer rather than a tightly coupled shared memory machine, since the different simulations were independent and did not communicate with one another while they were running. This is in contrast to other simulations that require a large number of atoms and many processors in parallel working on a single molecular system.

After running each of the trajectories for approximately one nanosecond, most of the free-energy

[3]These were NVE runs using the OPLSAA force field and SPC water in a 38-Å box. P3ME electrostatics provided fully periodic boundary conditions, and heavy-atom hydrogens were rigidly constrained, allowing one-femtosecond timesteps. (An NVE simulation has a constant number of particles, constant volume, and constant energy. OPLSAA = optimized potential for liquid simulation—all atom. SPC = single point charge. P3ME = particle-particle particle-mesh Ewald.)

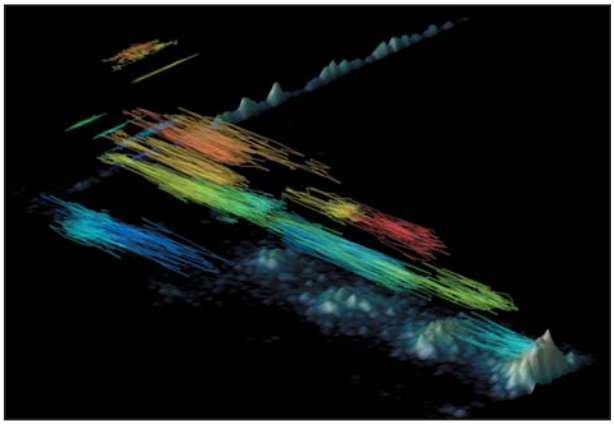

Figure 5

Combination of the static, thermodynamic view of the free-energy landscape from Figure 4(b) with kinetics from several independent trajectories of the hairpin system started from a range of initial configurations. Each trajectory is colored differently and explores its own region of the free-energy landscape. A single, very long simulation would gradually explore the entire landscape, showing local behavior similar to that seen in each of the short simulations.

surface had been visited by at least one trajectory, and together they provided good statistical coverage of the entire region. By dividing the area into bins and statistically determining the mean time in each bin, along with the likelihood of transitioning into the other bins, a transition matrix was defined for the process, which was then modeled using Markov methods. The full theory behind this technique is described in [21], but **Figure 6** is a visualization of the transition matrix as a pattern of directed arcs, with probability flux depicted as thickness along each tube. The combination of the static thermodynamic information in the free-energy landscape with a reduced but easily interpreted view of the path of a protein along the surface is abstractly visualized as hops between bins with corresponding branching probabilities. The full results of the simulation are described in [22], which presents the first publication of simulation work done with Blue Matter as part of the Blue Gene project. Although the behavior of the hairpin turned out to be too complex to be Markovian in our simple binning scheme, the technique shows promise for modeling protein transitions on the basis of data from multiple independent runs. We did not perform the simulation on Blue Gene hardware (since it did not yet exist), but the experiment was consistent with the research goals of the Blue Gene project and exercised Blue Matter along with its analysis framework.

Lipid–cholesterol bilayer membrane system

The next system we studied was a lipid bilayer consisting of 20K atoms, corresponding to a substantial increase

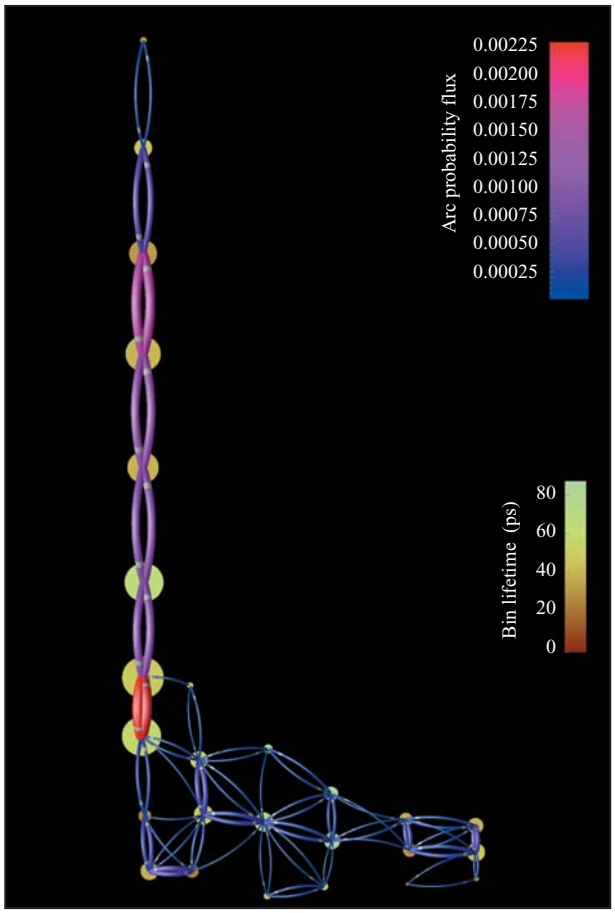

Figure 6

Depiction of a Markov model of the hairpin folding process deduced from observed transitions in many short trajectories. The circles are at the centers of bins defined on the landscape, and the arcs depict the probability flux flowing between the bins. The size of each circle depicts the expected occupancy of the bin, and the color conveys the bin lifetime, with blue to red representing short to long.

in system size from the hairpin. Lipid membranes are biologically important because they form the cell wall and control the transfer of material into and out of the cell. They are also important as the location where many proteins perform their function, as described in the final section.

Figure 7 shows a cross section of the lipid–cholesterol simulation, with two layers of lipid forming the bilayer, a layer of water on the outside, and cholesterols interspersed among the lipids as they would be in a real membrane.[4] A key role of analysis for this system is to

[4]This simulation consisted of 72 SDPC lipid molecules and 24 cholesterols in two leaflets, plus 2,174 TIP3P water molecules, using the CHARMM27 force field; NVE ensemble with a 2-fs timestep and rigid bonds to hydrogen. (CHARMM = chemistry at Harvard molecular mechanics [23]. TIP3P = transferable intermolecular potentials with three point charges. SDPC = 1-stearoyl-2-docosahexaenoyl-sn-glycero-3-phosphocholine.)

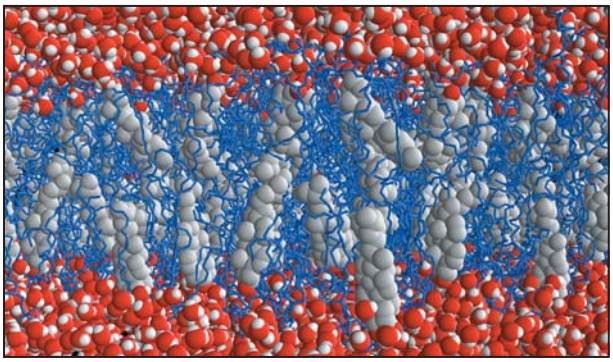

Figure 7

Cross-sectional view of a lipid–cholesterol bilayer membrane with water molecules on each side. Each blue strand is one chain of a lipid, and each lipid has two chains plus a head group. Cholesterols are interspersed in the lipids and shown as the gray molecules. The lipid–cholesterol system is split horizontally into two leaflets, consistent with the structure found in a cell wall, which acts as a permeable barrier between two solutions.

characterize the way in which cholesterols interact with the lipids. One way to do this is with histograms of the atoms in the vicinity of the cholesterol, which can then be sliced and shown as the isolines in **Figure 8(a)**. Three-dimensional visualization helps here by avoiding the need to slice the volume and instead showing a three-dimensional isosurface of the histogram [**Figure 8(b)**]. This immediately conveys the density of neighboring atoms around the cholesterol and reveals preferential associations at an atomistic level that can be accessed only via simulations such as these. Additional details of membrane simulation and analysis are discussed in the next section, which describes a lipid–cholesterol membrane and its interaction with an embedded membrane protein, rhodopsin.

Rhodopsin simulation on Blue Gene/L

Membrane proteins are the focus of more than half of the contemporary drug targets [24]. It is difficult to obtain structural detail for membrane proteins at atomic resolution because of the challenges of growing crystals of proteins in a membrane environment [25, 26]. In addition to structure, the functional dynamics are accessible via spectroscopic techniques [27–29]. The added complexity of the membrane environment must be included in membrane protein studies, since the composition of the membrane can have dramatic effects on the function of a protein [30–32].

Simulation of membrane protein environments has received a direct benefit from recent advances in supercomputing. The ability to produce tens to hundreds of nanoseconds of a fully atomistic simulation of membrane environments has produced detailed characterization of their structure and dynamics to a degree previously unavailable and at a scale inaccessible by experiment. As a result, simulation can offer additional insight into the structure, dynamics, and environment of the membrane proteins themselves [33].

Rhodopsin is a member of the signaling protein family known as G-protein-coupled receptors (GPCRs) [34–37]. It functions as the first step in the signal cascade

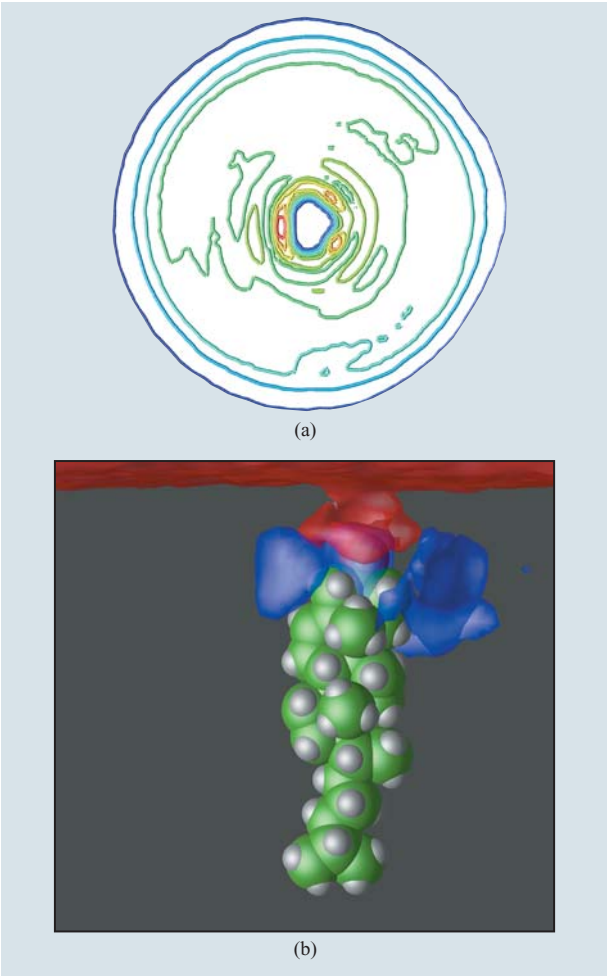

Figure 8

(a) Average neighborhood of a cholesterol molecule showing the density of adjacent atoms in a horizontal slice through each cholesterol. This 2D view shows preferential associations along each "face" of the molecule, but is hard to interpret, since it represents only a single slice through a complicated 3D interaction. (b) The same information as in (a), but shown as 3D isosurfaces in the context of a model of the cholesterol. In addition, the proximity of water molecules is shown as the red isosurface above. This use of 3D analysis immediately conveys the three dimensions of the cholesterol neighborhood in ways that would be difficult to assemble from multiple 2D slices as shown in (a).

that results in the perception of light [38], but has the distinction of being the only GPCR whose structure is known with atomic resolution [25]. Although other workers have used molecular dynamics to study rhodopsin in a membrane environment [33], we have chosen to apply the capabilities of BG/L to simulate rhodopsin in a more complex, native-like environment of polyunsaturated lipids with cholesterol.

Setting up a large-scale simulation is technically demanding and requires a lengthy equilibration protocol. This is complicated by the fact that it is difficult to determine the point in time at which equilibrium has been reached. For a system as complex as rhodopsin in a two-component lipid matrix with cholesterol, the equilibration time alone can extend to tens of nanoseconds. Until recently, this would have represented the full time allotted to a complete simulation, because of the limited computing resource available.

Figure 9(a) shows a top view of the rhodopsin/lipid/cholesterol system, looking down on the membrane surface. The rhodopsin appears as a red ribbon; the green and blue respectively represent the SDPE and SDPC[5], and the cholesterol atoms are shown as gray spheres. This snapshot is from early in the simulation, before the cholesterol distribution has had time to fully equilibrate. Experimental diffusion rates for cholesterol are of the order of 1×10^{-7} cm^2/s [39], which means that many nanoseconds of simulation time are required for the cholesterol distribution to reach a steady state.

The changing cholesterol distribution around rhodopsin as it progresses toward equilibrium appears in **Figure 9(b)** as successive histograms of the radial distribution of cholesterol over time. The counts of cholesterol molecules at a given distance are shown on the vertical scale, where the light color indicates greater likelihood at that distance. Each vertical histogram is an average of the cholesterols over a 250-picosecond time window. The most likely range, which appears as a band of gray across the time axis, is fairly stable, suggesting that the setup protocol produced a configuration near equilibrium. The minimum and maximum ranges of each histogram, however, show a general trend downward. The minimum distances decrease, which indicates that equilibrium includes some cholesterols closely associated with rhodopsin.

Another, simpler, way to characterize the equilibration of the system is to study the mean radius of the cholesterols over time, which amounts to a further reduced representation of the histogram strip chart. **Figure 10(a)** shows the time evolution of the mean radius of cholesterol from the rhodopsin center of mass. The

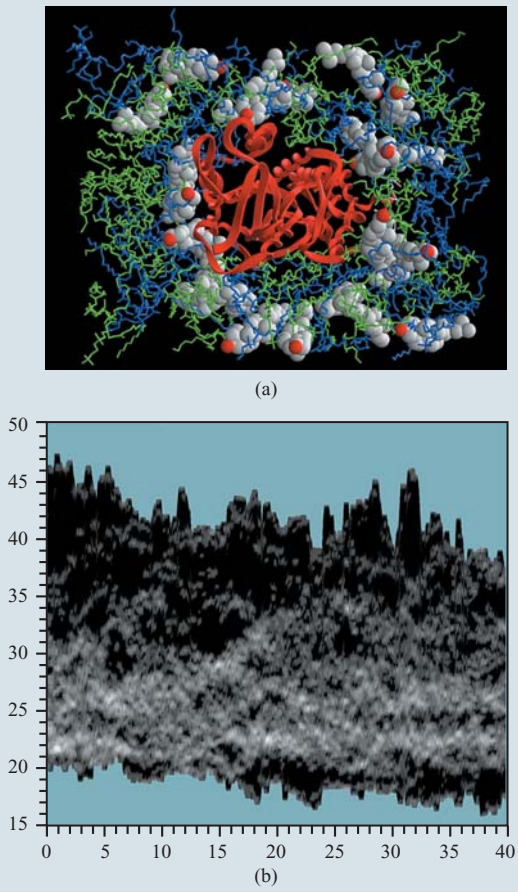

Figure 9

(a) Top view of the rhodopsin system, consisting of the rhodopsin protein, shown as the red ribbon, surrounded by the two-component lipid system, shown in green and blue, and the cholesterol, shown in gray. (b) This plot captures the equilibration of the rhodopsin system by showing the changing radial distribution of cholesterol over time. Each vertical slice is a histogram at that point in time of the radial distribution of cholesterol from the center of the protein. There is a fair amount of mixing evident, but with an overall declining trend that is largely equilibrated after 25 ns.

clear oscillatory pattern indicates fluctuations in the cholesterol distribution on a timescale of several nanoseconds. The amplitude of the oscillations is damping to a converged value, suggesting significant progress toward equilibrium in approximately 25 ns.

Figure 10(b) suggests that in the first 30 ns, part of the rhodopsin structure underwent a rotation of approximately 10 degrees from the initial setup conditions and then stabilized during the 30-ns to 40-ns time frame. Further monitoring is needed to establish stability in the overall rhodopsin orientation, since protein motion occurs on a timescale of tens of

[5]SDPE and SDPC are abbreviations for the chemical makeup of the lipids: SDPE is 1-stearoyl-2-docosahexaenoyl-sn-glycerophosphoethanolamine, and SDPC is 1-stearoyl-2-docosahexaenoyl-sn-glycero-3-phosphocholine.

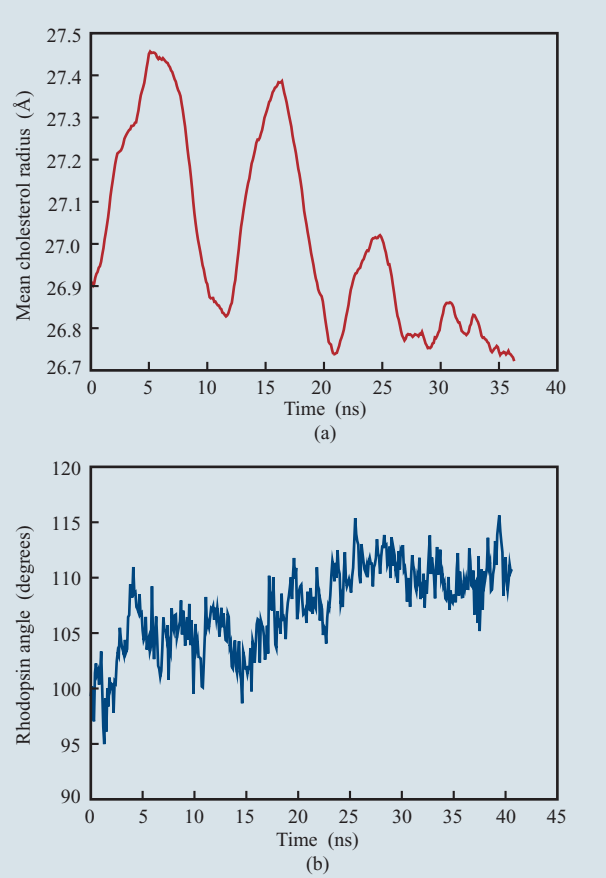

Figure 10

(a) Plot reduced from the information in Figure 9(b), showing simply the mean cholesterol radial distance over time. Here, initial oscillations are evident, which are largely damped out after 25 ns. (b) Changing angle of rhodopsin over time, indicating only a slight shift over many nanoseconds. The lack of rotation simplifies the interpretation of the local behavior of lipid and cholesterol near the protein.

nanoseconds. It is not clear how much of this rotation is due to an internal conformational change in rhodopsin, or whether it represents an overall rigid body rotation as it adjusts to its environment; however, the fact that the protein is not making significant angular changes greatly simplifies the analysis of the neighboring lipid–cholesterol environment.

Care must be taken to validate interpretations of the rich detail offered by fully atomistic simulations of complex environments, such as the one described here. Each line of analysis should make contact with what can be experimentally observed. For example, some measurable spectroscopic quantities can be calculated from the simulation and compared with experiment. Once validated, the atomic detail can be examined further for mechanistic insight that is inaccessible experimentally. Simulation, once validated with experiment, can help interpret experimental results by providing a greater level of detail than is available with contemporary techniques.

Supercomputing resources such as BG/L have expanded the level at which complex biological phenomena may be studied. Most significantly, the newly harnessed capability enables investigations that could not have been attempted a few years ago. In this light, the potential advances with BG/L are in kind, rather than in degree. The rhodopsin simulations are ongoing and will be investigated in detail using techniques such as those described above.

Conclusion

The Blue Gene/L project has been on a dual path to produce a massively parallel machine and associated scalable software to make use of its power. As more nodes of the machine become available, the scientific possibilities increase tremendously, since the simulations can then access more biologically interesting scales of time and complexity. Much insight has already emerged from Blue Matter simulations on an IBM SP computer and on a BG/L computer of up to 512 nodes, and we look forward to the completion of our work on rhodopsin and the continuation onto more complex systems as the node count expands.

Acknowledgments

The authors gratefully acknowledge contributions from Scott E. Feller, valuable discussions with Bruce Berne, and the development effort of the Blue Matter team—Blake G. Fitch, Alex Rayshubskiy, T. J. Chris Ward, Yuri Zhestkov, and Maria Eleftheriou.

References

1. F. Allen, G. Almasi, W. Andreoni, D. Beece, B. J. Berne, A. Bright, J. Brunheroto, C. Cascaval, J. Castanos, P. Coteus, P. Crumley, A. Curioni, M. Denneau, W. Donath, M. Eleftheriou, B. Fitch, B. Fleisher, C. J. Georgiou, R. Germain, M. Giampapa, D. Gresh, M. Gupta, R. Haring, H. Ho, P. Hochschild, S. Hummel, T. Jonas, D. Lieber, G. Martyna, K. Maturu, J. Moreira, D. Newns, M. Newton, R. Philhower, T. Picunko, J. Pitera, M. Pitman, R. Rand, A. Royyuru, V. Salapura, A. Sanomiya, R. Shah, Y. Sham, S. Singh, M. Snir, F. Suits, R. Swetz, W. C. Swope, N. Vishnumurthy, T. J. C. Ward, J. Warren, and R. Zhou, "Blue Gene: A Vision for Protein Science Using a Petaflop Supercomputer," *IBM Syst. J.* **40**, No. 2, 310–327 (2001).
2. R. S. Germain, Y. Zhestkov, M. Eleftheriou, A. Rayshubskiy, F. Suits, T. J. C. Ward, and B. G. Fitch, "Early Performance Data on the Blue Matter Molecular Simulation Framework," *IBM J. Res. & Dev.* **49**, No. 2/3, 447–455 (2005, this issue).
3. R. F. Enenkel, B. G. Fitch, R. S. Germain, F. G. Gustavson, A. Martin, M. Mendell, J. W. Pitera, M. C. Pitman, A. Rayshubskiy, F. Suits, W. C. Swope, and T. J. C. Ward,

"Custom Math Functions for Molecular Dynamics," *IBM J. Res. & Dev.* **49**, No. 2/3, 465–474 (2005, this issue).
4. C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, *Nature* **420**, No. 6911, 102–106 (2002).
5. D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Second Edition, Academic Press, Inc., New York, 2001.
6. W. Y. Yang, J. W. Pitera, W. C. Swope, and M. Gruebele, "Heterogeneous Folding of the trpzip Hairpin: Full Atom Simulation and Experiment," *J. Molec. Biol.* **336**, No. 1, 241–251 (2004).
7. B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, "Design of a Novel Globular Protein Fold with Atomic Accuracy," *Science* **302**, 1364–1368 (2003).
8. L. L. Looger, M. A. Dwyer, J. J. Smith, and H. W. Hellinga, "Computational Design of Receptor and Sensor Proteins with Novel Functions," *Nature* **423**, No. 6936, 185–190 (2003).
9. C. J. Levinthal, "Are There Pathways for Protein Folding?" *J. Chem. Phys.* **65**, 44–45 (1968).
10. P. Bross, T. J. Corydon, B. S. Andresen, M. M. Jorgensen, L. Bolund, and N. Gregersen, "Protein Misfolding and Degradation in Genetic Diseases," *Human Mutation* **14**, No. 3, 186–198 (September 1999).
11. C. M. Dobson, "Principles of Protein Folding, Misfolding and Aggregation," *Seminars in Cell & Developmental Biol.* **15**, No. 1, 3–16 (2004).
12. D. M. Walsh and D. J. Selkoe, "Oligomers in the Brain: The Emerging Role of Soluble Protein Aggregates in Neurode Generation," *Protein & Peptide Lett.* **11**, No. 3, 213–228 (2004).
13. S. Sato, C. L. Ward, M. E. Krouse, J. J. Wine, and R. R. Kopito, "Glycerol Reverses the Misfolding Phenotype of the Most Common Cystic Fibrosis Mutation," *J. Biol. Chem.* **271**, No. 2, 635–638 (January 1996)
14. W. F. van Gunsteren and A. E. Mark, "Validation of Molecular Dynamics Simulation," *J. Chem. Phys.* **108**, No. 15, 6109–6116 (1998).
15. R. D. Skeel, G. Zhang, and T. Schlick, "A Family of Symplectic Integrators: Stability, Accuracy, and Molecular Dynamics Applications," *SIAM J. Sci. Comp.* **18**, 203–222 (1997).
16. Y. Sugita and Y. Okamoto, "Replica-Exchange Molecular Dynamics Method for Protein Folding," *Chem. Phys. Lett.* **314**, 141–151 (1999).
17. R. Zhou, B. J. Berne, and R. Germain, "Free Energy Landscape of a Beta-Hairpin Folding in Explicit Water," *Proc. Natl. Acad. Sci. USA* **98**, No. 26, 14931–14936 (December 2001).
18. A. E. Garcia and K. Y. Sanbonmatsu, "Exploring the Energy Landscape of a Beta Hairpin in Explicit Solvent," *Proteins* **42**, No. 3, 345–354 (February 2001).
19. D. Gresh, F. Suits, and Y. Sham, "Case Study: An Environment for Understanding Protein Simulations Using Game Graphics," *Proceedings of the IEEE Visualization Conference*, 2001, pp. 445–448.
20. J. S. Richardson, "The Anatomy and Taxonomy of Protein Structure," *Adv. Protein Chem.* **34**, 167–339 (1981).
21. W. C. Swope, J. W. Pitera, and F. Suits, "Describing Protein Folding Kinetics by Molecular Dynamics Simulations. Part 1: Theory." *J. Phys. Chem. B* **108**, 6571–6581 (2004).
22. W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. J. C. Ward, Y. Zhestkov, and R. Zhou, "Describing Protein Folding Kinetics by Molecular Dynamics Simulations. Part 2: Example Applications to Alanine Dipeptide and a Beta-Hairpin Peptide," *J. Phys. Chem. B* **108**, 6582–6594 (2004).
23. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "A Program for Macromolecular Energy, Minimization, and Dynamics Calculations, *J. Comp. Chem* **4**, 187–217 (1983).
24. *Gpcrs: The Targets of Today's Drugs and Tomorrow's Blockbusters*, Drug and Market Development Publishing, January 2002; see *http://www.researchandmarkets.com/reportinfo.asp?report_id=4043*.
25. K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, B. I. Le Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, and M. Miyano, "Crystal Structure of Rhodopsin: a G Protein-Coupled Receptor," *Science* **289**, 739–745 (2000).
26. D. C. Teller, T. Okada, C. A. Behnke, K. Palczewski, and R. E. Stenkamp, "Advances in Determination of a High-Resolution Three-Dimensional Structure of Rhodopsin, a Model of G-Protein-Coupled Receptors (gpcrs)," *Biochemistry* **40**, 7761–7772 (2001).
27. J. Isele, T. P. Sakmar, and F. Siebert, "Rhodopsin Activation Affects the Environment of Specific Neighboring Phospholipids: An FTIR Spectroscopic Study," *Biophys. J.* **79**, No. 6, 3063–3071 (2000).
28. D. C. Mitchell, S.-L. Niu, and B. J. Litman, "Optimization of Receptor-G Protein Coupling by Bilayer Lipid Composition I. Kinetics of Rhodopsin–Transducin Binding," *J. Biol. Chem.* **276**, No. 46, 42801–42806 (November 2001).
29. S.-L. Niu, D. C. Mitchell, and B. J. Litman, "Optimization of Receptor-G Protein Coupling by Bilayer Lipid Composition II. Formation of Metarhodopsin II–Transducin Complex," *J. Biol. Chem.* **276**, No. 46, 42807–42811 (November 2001).
30. K. Burger, G. Gimpl, and F. Fahrenholz, "Regulation of Receptor Function by Cholesterol," *Cell Molec. Life Sci.* **57**, 1577–1592 (2000).
31. B. J. Litman and D. C. Mitchell, "A Role for Phospholipid Polyunsaturation in Modulating Membrane Protein Function," *Lipids* **31**, Supplement S, 193–197 (March 1996).
32. A. Polozova and B. J. Litman, "Cholesterol Dependent Recruitment of di22:6-PC by a G Protein-Coupled Receptor into Lateral Domains," *Biophys. J.* **79**, No. 5, 2632–2643 (November 2000).
33. P. S. Crozier, M. J. Stevens, L. R. Forrest, and T. B. Woolf, "Molecular Dynamics Simulation of Dark-Adapted Rhodopsin in an Explicit Membrane Bilayer: Coupling Between Local Retinal and Larger Scale Conformational Change," *J. Molec. Biol.* **333**, 493–514 (2003).
34. P. J. Reeves, J.-M. Kim, and H. G. Khorana, "Structure and Function in Rhodopsin: A Tetracycline-Inducible System in Stable Mammalian Cell Lines for High-Level Expression of Opsin Mutants," *Proc. Natl. Acad. Sci. USA* **99**, No. 21, 13413–13418 (October 2002).
35. P. J. Reeves, R. L. Thurmond, and H. G. Khorana, "Structure and Function in Rhodopsin: High Level Expression of a Synthetic Bovine Opsin Gene and Its Mutants in Stable Mammalian Cell Lines (G-Protein Coupled Receptor/11-cis-retinal/Glycosylation/Metarhodopsin II/Immunoaffinity Chromatography)," *Proc. Natl. Acad. Sci. USA* **93**, No. 21, 11487–11492 (October 1996).
36. P. J. Reeves, J. Hwa, and H. G. Khorana, "Structure and Function in Rhodopsin: Kinetic Studies of Retinal Binding to Purified Opsin Mutants in Defined Phospholipid–Detergent Mixtures Serve as Probes of the Retinal Binding Pocket (G-Protein-Coupled Receptors/Signal Transduction/11-cis-retinal/Binding Pocket/Retinitis Pigmentosa)," *Proc. Natl. Acad. Sci. USA* **96**, No. 5, 1927–1931 (1999).
37. P. J. Reeves, N. Callewaert, R. Contreras, and H. G. Khorana, "Structure and Function in Rhodopsin: High-Level Expression of Rhodopsin with Restricted and Homogeneous N-Glycosylation by a Tetracycline-Inducible N-Acetylglucosaminyltransferase I-Negative HEK293S Stable Mammalian Cell Line," *Proc. Natl. Acad. Sci. USA* **99**, No. 21, 13419–13424 (October 2002).
38. P. L. Yeagle, G. Choi, and A. D. Albert, "Structure of the G-Protein Coupled Receptor Rhodopsin Including the Putative G-Protein Binding Site in Dark Adapted and Activated Forms," *Biochemistry* **40**, 11932–11937 (2001).

39. C. Gliss, O. Randel, H. Casalta, E. Sackmann, R. Zorn, and T. Bayerl, "Anisotropic Motion of Cholesterol in Oriented DPPC Bilayers Studied by Quasielastic Neutron Scattering: The Liquid-Ordered Phase," *Biophys. J.* **77**, No. 1, 331–340 (July 1999).

Received August 20, 2004; accepted for publication September 28, 2004; Internet publication April 5, 2005

Frank Suits *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (suits@us.ibm.com).* Dr. Suits is a member of the Biomolecular Dynamics and Scalable Modeling Group within the Computational Biology Center at the IBM Thomas J. Watson Research Center. This group is responsible for the software and science involved in the protein simulations that are integral to the Blue Gene project. Although his degree is in optical physics, he has worked on a wide variety of projects at the IBM Thomas J. Watson Research Center, including optical storage, magnetic storage materials, scientific visualization, and queuing systems. At present, Dr. Suits is focusing on the analysis of the protein and membrane simulations currently running on BG/L.

Michael C. Pitman *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (pitman@watson.ibm.com).* Dr. Pitman received his Ph.D. degree in chemistry in 1995 from the University of California at Santa Cruz. He joined the Biomolecular Dynamics and Scalable Modeling Group within the Computational Biology Center at the IBM Thomas J. Watson Research Center soon afterward and continued work in the area of computational drug design methods. He began a leading role in the Blue Gene Protein Science program in 2001, focusing on large-scale membrane and membrane protein simulation. His research interests are focused on understanding the nature of protein–membrane interactions. Dr. Pitman conducts large-scale all-atom simulations of membrane proteins in explicit, biologically relevant environments.

Jed W. Pitera *IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120 (pitera@us.ibm.com).* Dr. Pitera is a Research Staff Member in the Science and Technology Department at the IBM Almaden Research Center. His research focuses on the use of computer simulation to address questions in biology and chemistry, particularly in the areas of protein folding, molecular recognition, and self-assembly. He received undergraduate training in biology and chemistry at the California Institute of Technology, where he worked in Dr. Pamela Bjorkman's Protein Crystallography Group. He subsequently pursued graduate studies in biophysics at the University of California at San Francisco (UCSF) in the laboratory of Dr. Peter Kollman. Dr. Pitera developed an interest in the use of biomolecular simulation and free-energy calculations in the rational design of proteins and pharmaceuticals while in Dr. Kollman's group. He pursued similar work in a postdoctoral position with Prof. Dr. Wilfred van Gunsteren at the Swiss Federal Institute of Technology Zurich (ETH), where his research focused on novel methods of calculating free energies for ligand design. He has worked as a member of the IBM Blue Gene Project Science and Application team since February of 2001. Dr. Pitera is also an adjunct assistant professor in the UCSF Department of Pharmaceutical Chemistry.

William C. Swope *IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120 (swope@almaden.ibm.com).* Dr. Swope has been engaged with the IBM Blue Gene Protein Science program since 2000, with strong emphasis on biomolecular simulation methodology and the development of practical techniques to simulate protein folding kinetics and thermodynamics. He joined the Science and Technology Department in 1992 at the IBM Almaden Research Center, where he has also been involved in scientific software development of computational chemistry applications and in technical data management issues related to life sciences. He began with IBM in 1982 at IBM Instruments, Inc., an IBM subsidiary that developed scientific instrumentation, where he worked in an

advanced processor design group. He also worked for six years at the IBM Scientific Center in Palo Alto, California, where he supported scientific customers of IBM in their development of software for numerically intensive computation. He received his undergraduate degree in chemistry and physics from Harvard University and his Ph.D. degree in quantum chemistry from the University of California at Berkeley. He then performed postdoctoral research on the statistical mechanics of solvation and condensed phases in the chemistry department at Stanford University. Dr. Swope maintains a number of scientific relationships and collaborations with academic and commercial scientists involved in the life sciences, specifically related to drug development.

Robert S. Germain *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (rgermain@us.ibm.com)*. Dr. Germain manages the Biomolecular Dynamics and Scalable Modeling Group within the Computational Biology Center at the IBM Thomas J. Watson Research Center. He received his A.B. degree in physics from Princeton University in 1982 and his M.S. and Ph.D. degrees in physics from Cornell University. He joined the IBM Thomas J. Watson Research Center as a Research Staff Member in the Physical Sciences Department after receiving his doctorate in 1989, and later the VLSI/Scalable Parallel Systems Packaging Department. Dr. Germain was project leader, from 1995 to 1998, for the development of a large-scale fingerprint identification system using an indexing scheme (FLASH) developed at IBM Research. He has been responsible for the science and associated application portions of the Blue Gene project since 2000. His current research interests include the parallel implementation of algorithms for high-performance scientific computing, the development of new programming models for parallel computing, and applications of high-performance computing to challenging scientific problems in computational biology. Dr. Germain is a member of the IEEE and the American Physical Society.