# Protein modeling with Blue Gene/L

## Real-world scientific advances through modeling and data visualization on a supercomputer

T.J. Christopher Ward (tjcw@uk.ibm.com)
Advisory Software Engineer
IBM

Ruhong Zhou, Ph.D. (ruhongz@us.ibm.com)
Research Staff Member
IBM

09 June 2009

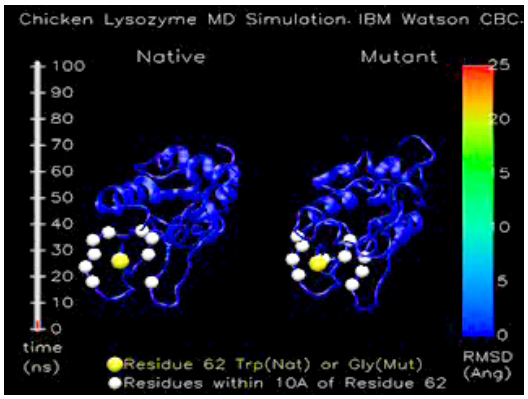
The Blue Gene®/L supercomputer provides scientists with the cutting-edge computing power and complex data-visualization tools they need to stay at the forefront of their disciplines. Learn how this technology lets computational molecular biologists create protein folding and misfolding simulations to better understand these complex molecules.

In 2001, IBM's research scientists started the design of a new family of servers, now marketed as the IBM System Blue Gene®. These servers have been available for use since 2004—first the Blue Gene/L (which we talk about in this article), then the Blue Gene®/P.

The Blue Gene family of supercomputers is designed to deliver ultra-scale performance with a standard programming environment; they're also designed to provide efficiencies in power, cooling, and floor-space consumption. Many universities, government, and commercial research labs use Blue Gene for computational studies in radio astronomy, protein folding, climate research, cosmology, and drug development. The system is making quite a change, by orders of magnitude, in the way science can be done, because it offers a more cost-effective tool for designing and running alternative versions of complex models.

In this article, we present some of the progress that has been made by one of the projects having to do with protein modeling. Figure 1 shows the scale of work we can do now, thanks to the power of Blue Gene/L. The initial configuration starts from the lysozyme crystal structure (see Resources for source).

**Figure 1. Part of the total ten microseconds of life inside a living cell; watch the video**




## Proteomics: The protein economy

Proteins are biological macromolecules that are an essential component of organisms and participate in every process within cells. Many proteins are enzymes that catalyze biochemical reactions; some are involved in cell signaling and immune responses; many others have structural and mechanical functions for muscles and cytoskeletons. Two examples illustrate how pervasive and important proteins are:

- One protein is responsible for the "redness" of blood; it carries oxygen from the lungs to all the other parts of the body.
- Another protein is responsible for the human body's response to the poison in poison-ivy; extremely irritating, but not normally harmful.

There are hundreds of thousands of proteins involved in life on Earth. Proteomics is the study of how proteins work, how they interact, and how their diversity and specialization evolve among the living organisms around us. This article is a short tour of what proteins are, how they are made, and how they affect the systems they inhabit.

DNA is the information storage component in every cell in every plant and animal. It stores information as a sequence of chemical building blocks (nucleotides) we call **A**, **C**, **T**, and **G** (for adenine, cytosine, thymine, and guanine in DNA, and uracil replacing thymine in RNA). From a distance, these building blocks look very similar, so every piece of DNA you look at has the same overall shape—the famous Watson-Crick Double Helix.

To read out the information in the DNA, the DNA untwists and another molecule called RNA is formed by presentation of the internal pattern. Rather like pressing a key into putty, you now have an image of the key in the putty. This RNA molecule is next presented as a blueprint to the ribosome, a protein that behaves like an all-purpose factory. The ribosome reads the A/C/T/G code in groups of three, allowing us to derive a 64-letter "alphabet."

Twenty of these "letters" correspond to amino acids, the building blocks for proteins. These amino acids come mainly from the food we eat (humans cannot synthesize all the amino acids we need and therefore must obtain the others, called "essential" amino acids, from food). Each amino acid has a "head" and a "tail." The ribosome finds the appropriate amino acid for each "letter" and

assembles them head-to-tail in sequence; other "letters" indicate when to start and when to stop. The resulting linear sequence of amino acids is a newly minted protein molecule, formed precisely according to the code imprinted in the section of DNA that was used.

Stresses and strains between the atoms in the protein molecule, interactions with the slightly salty water in the cell, and random vibrations that you would call *heat* then cause the protein molecule to "fold" into a characteristic shape.

Protein molecules are quite stable; some of them can exist unchanged for hundreds of years and sustain temperatures of hundreds of degrees, which would kill the organism that made them. They stay roughly the way they are until they are denatured by strong chemicals, high pressure, heat or cold, or by becoming food for another living thing.

The shape and the way it varies with time, temperature, and surrounding molecules determine what the protein molecule will do—whether it will transport oxygen, give you a poison-ivy allergy, or do any of the other things that can happen at that tiny scale.

Figure 2 demonstrates the familiar ball-and-stick model of DNA (image is a stereo pair; see Resources for the image source):

## Figure 2. The ball-and-stick model of DNA

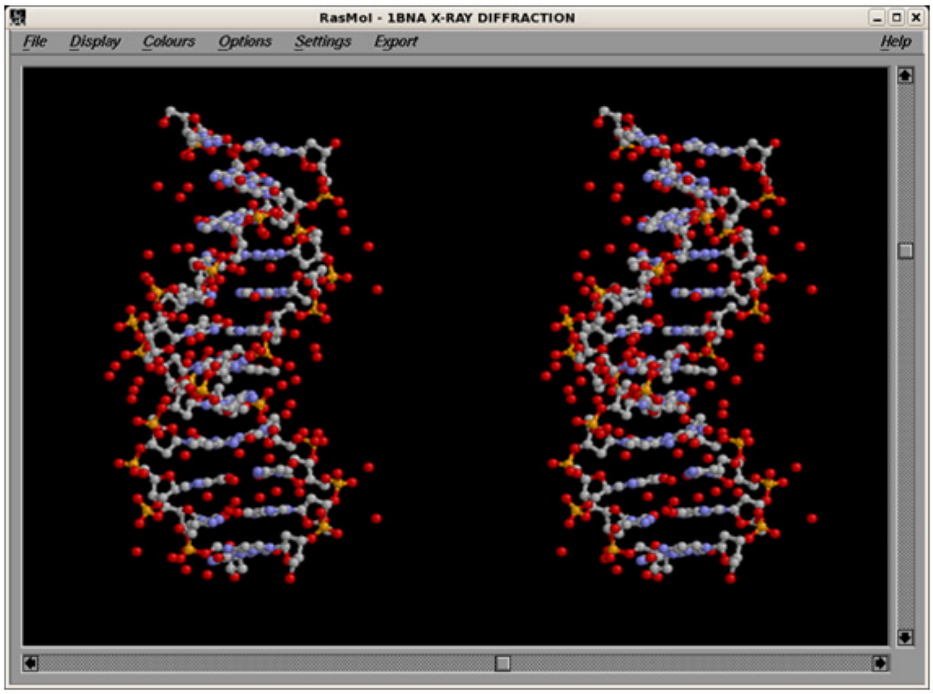
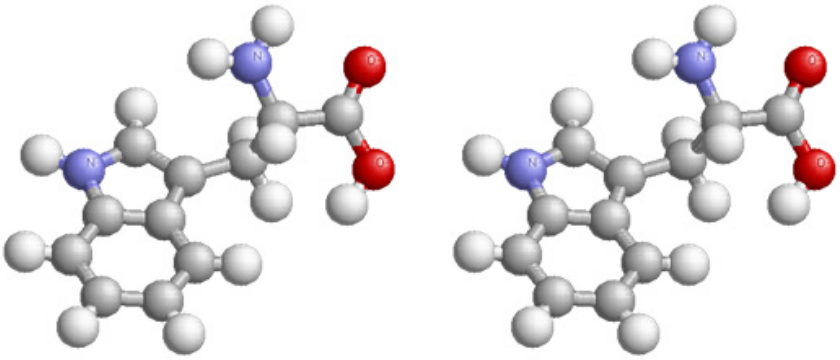



Figure 3 shows tryptophan, one of the 20 standard amino acids (image is a stereo pair; see Resources for the image source).

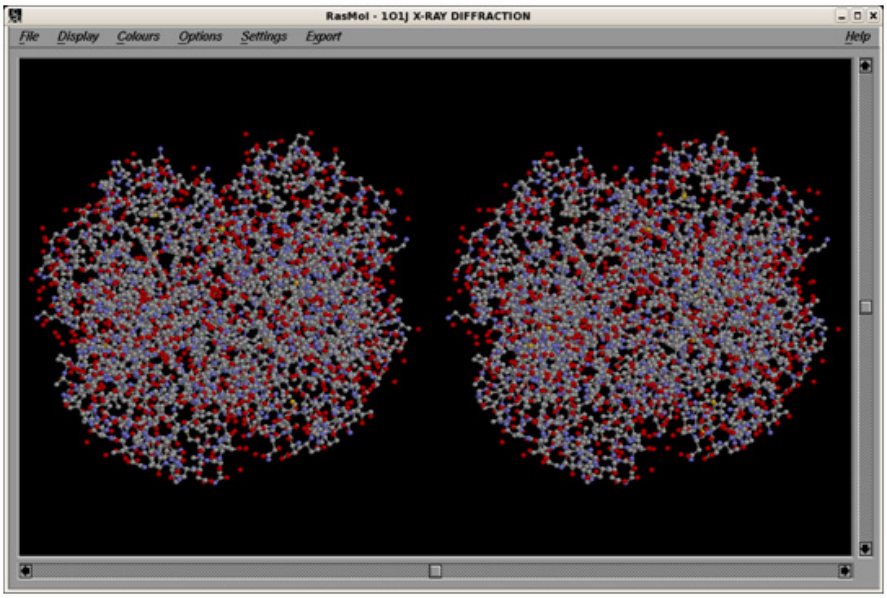
## Figure 3. Tryptophan, one of the 20 standard amino acids



Amino acids are assembled into proteins by cutting off the O-H group (right side of Figure 3) of one molecule, cutting the H away from the N (top of Figure 3) of another molecule, and splicing the molecules together. The H-O-H group left over is a water molecule. All amino acids have this characteristic atomic grouping (top right of Figure 3).

Figure 4 offers a visual look at the protein hemoglobin (image is a stereo pair; see Resources for the image source).

## Figure 4. The protein hemoglobin




Hemoglobin is a total of 574 amino acid molecules in 4 subunits. Hemoglobin, with its associated iron atoms (how they are assembled into the protein is beyond the scope of this article), transports oxygen around the bloodstream. An oxygen transport system is possible with just the iron atoms, but it is very much more effective with the protein "cage" the structure of hemoglobin provides. If you put this image into a stereo viewer, you can pick out the 3D atomic structure; for anything more complex than this, we need a different way to visualize what is going on.

## Commercial and academic motivations

### What's a wildtype?

A wildtype (also written as "wild type" and "wild-type") is the typical form that an organism, gene, strain, or characteristic takes in nature. If referring to the *phenotype* (an organism's observable characteristics, generally the expression of genes and environmental factors), wildtype characterizes the most common traits in the natural population. If referring to the *genotype* (the unobservable genetic composition), it defines the allele at each locus required to produce the wildtype phenotype. Wildtypes are neither dominant nor recessive. A good antonym for wildtype is *mutant*.

Increasingly, advances in designing pharmaceuticals and protecting public health are coming from a better understanding of the basic building blocks of life such as proteins. One current topic is *protein misfolding and aggregation*—if a protein folds into a shape other than the intended one, the result often produces inactive proteins with different properties, which can lead to such neurodegenerative diseases as Alzheimer's, Creutzfeldt-Jakob, bovine spongiform encephalopathy (Mad Cow), Huntington's and Parkinson's, cystic fibrosis, and other amyloidosis.

Understanding what can cause protein molecules to change from their useful folded form to a different folded form is an active research topic in the search for a treatment for these significant diseases. Recent experiments pioneered by Chris Dobson and co-workers at Cambridge University (see Resources for a link) have shown that amyloids and fibrils can be formed not only from the traditional beta-amyloid peptides but also from almost any proteins (such as lysozyme) given the appropriate conditions. In fact, a single mutation (W62A) on lysozyme protein can cause the protein to be in a much less stable state as compared to the wildtype (see sidebar); it can also cause it to misfold and form possible amyloids in urea solution due to the loss of key "long-range hydrophobic interactions."

Scientists do not yet know how this single W62 residue can play a key role in the hydrophobic long-range interactions during the folding process and then shift to the surface from presumably a nucleation site for functioning reasons. This offers a unique opportunity for better understanding of the single mutation effects, as well as the mechanism behind the aforementioned diseases related to protein misfolding and aggregation.

The Blue Gene/L technology offers a powerful way to study these types of diseases, because it provides a more cost-effective (and faster) way to model the effects of protein folding and misfolding.

## So what are we modeling?

The video from which Figure 1 was captured is a visualization of part of a sequence of a lysozyme protein misfolding due to a single mutation. Lysozyme is a protein that is part of the human immune system; when functioning properly, it punctures the cell wall of an invading bacterium and destroys it.

A single mutation, a different sequence in the DNA, causes the ribosome to use a different amino acid when it builds the lysozyme molecule. The theory is that this different amino acid affects the shape that the lysozyme folds into and that the differently shaped lysozyme molecule is differently

effective in puncturing bacterial cell walls. By understanding this change, we may be able to design pharmaceuticals or other forms of therapy that will assist individuals with this mutation in recovering from bacterial disease.

As part of the work, we store the positions and velocities of every atom in one molecule of lysozyme, as well as those of approximately 10,000 water and urea molecules (this simulation is done in an 8 molar urea solution to mimic the experiments), in the computer's memory. There are many ways to model the forces between atoms; we use a variant of a *ball and spring* model for bonded forces with an inverse-square-law model for electrostatic forces between charged atoms and an attract/repel model for atoms that are near each other but not covalently bonded. The model is run as a time series. At each time step, we calculate the forces on each atom, then we update the velocities and the positions according to Newton's second law.

At each time step (very small, on the order of 1 femtosecond), there are in principle hundreds of millions of forces to be calculated. This large number of calculations and the fact that we also want to be able to run the simulations long enough (microseconds) to model motions of interest means this approach has only recently become practical, even with the largest computers that we know how to build. For more details on what we do and some alternative approaches, see the link for "Destruction of long-range interactions by a single mutation in lysozyme" in Resources.

## Equipping the laboratory

At IBM Watson Research Lab in Yorktown, New York, we have 20 racks of BlueGene/L servers. Each rack contains 1,024 PowerPC® dual-core microprocessor chips; each microprocessor is attached to 512MB of RAM. For every 64 chips in this *compute lattice*, there is an additional microprocessor connected to a 1Gbps Ethernet link. These 320 Ethernet links are connected through standard Ethernet switches to standard IBM Power Systems machines with disks, tapes, language compilers, and job-control software.

This lysozyme modeling work has used an average of four racks of BlueGene/L processors for several months to generate an aggregate of more than 10 microseconds of molecular dynamics data. Periodically, the application writes out the positions and velocities of all the atoms under simulation (part of this stream of information was used to produce the synthetic video mentioned above). Whenever it is necessary to restart the simulation run, an appropriate set of positions and velocities can be reloaded. Restarting may be needed after a planned shutdown, after an unplanned machine failure, or in order to replay a model event of scientific interest with a different time step granularity.

## Running the model

The application is booted onto the Blue Gene/L nodes by a mechanism similar to MPICH job submission (MPICH is a freely available, portable implementation of MPI, the message-passing interface; see Resources for a link). Each processor in the cluster provides a POSIX-file-system environment to the application. Data can be set up in an IBM General Parallel File System (GPFS) file system for the application to read; when the application writes results, the results would also go there for external use.

For time-series modeling applications such as this, it is normal to read the initial conditions from the file system and then to write periodic "snapshots" of the model state to the file system.

## What does all this give us?

The video is a glimpse into a world that has never before been visible. Of course, we don't know that it represents the truth—scientists always need to compare what a model shows with what they can see in the real world. Seeing how lysozyme misfolds in reality is still a dream; even "seeing" part of the "fixed" conformations means preparing samples and putting them under an electron microscope or possibly even causing large numbers of lysozyme molecules to crystallize, then using X-ray diffraction spectroscopy. However, these experimental techniques typically do not give insight into how the protein might move.

Therefore, the current large-scale simulations offer a unique window to look into the details of molecular movements and critical changes involved in disease-related misfoldings. Hopefully, the availability of the technology that can make this happen will push the envelope and advance the state of the art in amyloidosis studies. It can also be used to train the next generation of scientists to solve these types of problems in this new way as their primary method for doing this type of research.

## Predicting the future

Actually, we would not be so bold as to attempt to divine tomorrow, but we would venture to guess that Blue Gene computing will continue to follow a development path (we use version L; the available Blue Gene/P upgrades to 4 processors per chip, 10Gbps Ethernet, and a host of other improvements). The cost of doing more compute-intensive arithmetic and the cost of more and faster storage (both heavily associated with the data-visualization tasks we describe in this article) will most likely continue to fall—as they must, because there is several worlds' worth of advanced modeling that scientists need to be doing, both for public research and for businesses to bring products to market.

The lysozyme model we describe only scratches a molecule off the surface of the new field of computational biology. There are more than 50,000 proteins whose structures are catalogued in the public Protein Data Bank (see Resources for a link); there are millions of potential pharmaceutically useful compounds to be analyzed; and there are many human diseases known to be related to proteins and their defects. And we're not even considering the myriad other research areas that can benefit from modeling on this scale. Blue Gene's work has just begun.

# Resources

## Learn

- A roundup of Blue Gene/L-related research can be found at the IBM Blue Gene project page. Other Blue Gene solution components and resources include:
  - The IBM Blue Gene/P solution page
  - The IBM General Parallel File System
  - Everything you could want to know about the IBM XL C/C++ Advanced Edition for Blue Gene compiler
  - IBM Redbooks on Blue Gene technologies
  - And a picture of the Blue Gene used by one of the authors
- The RCSB Protein Data Bank (PDB) is an archive for the study of biological macromolecules with information about experimentally determined structures of proteins, nucleic acids, and complex assemblies. Educational resources include such cool things as the Molecule of the Month.
- Source data for Figure 1 is from the PDB, The 1.33 A structure of tetragonal hen egg white lysozyme.
- Source data for Figure 2 is from the PDB, Structure of a B-DNA dodecamer: conformation and dynamics.
- Figure 3 is courtesy of the MathMol library hosted at New York University.
- Source data for Figure 4 is from the PDB, Deoxy hemoglobin (A-GLY-C:V1M,L29F,H58Q; B,D:V1M,L106W).
- Chris Dobson's group posts links to more research in molecular biology.
- "Destruction of long-range interactions by a single mutation in lysozyme" (R. Zhou, M. Eleftheriou, A. Royyuru, B. J. Berne; Proc. Natl. Acad. Sci., 2007) gives more information about the modeling approach used in these simulations.
- "Parallel implementation of the replica exchange molecular dynamics algorithm on Blue Gene/L" (M. Eleftheriou, A. Rayshubski, J. W. Pitera, B. G. Fitch, R. Zhou, R. S. Germain; IEEE, 2006) explains some of the mathematical techniques used for the simulation.
- MPICH2 is the next stage of MPICH, the high-performance, widely portable (and free) implementation of the Message Passing Interface (MPI) standard.
- The Argonne Leadership Computing Facility has a collaborative program that provides Blue Gene/P time to the computational science community.
- The modeling application is available for demonstration at IBM Innovation Centers worldwide.
- "High-performance Linux clustering" is a two-part series providing background on high-performance computing with Linux. Part 1 (developerWorks, September 2005) covers HPC fundamentals, types of clusters available, reasons for choosing a cluster configuration, and the role of Linux in HPC. Part 2 (developerWorks, October 2005) discusses parallel programming using MPI, covers cluster management and benchmarking, and shows how to set up a Linux cluster using open source software.
- "Port Fortran applications" (developerWorks, April 2009) helps you overcome common hurdles when porting Fortran applications among various high performance computing systems.

- In the developerWorks Linux zone, find more resources for Linux developers, and scan our most popular articles and tutorials.
- See all Linux tips and Linux tutorials on developerWorks.
- Stay current with developerWorks technical events and Webcasts.

### Get products and technologies

- Open Discovery is a Fedora Core-based, Live Linux distribution of bioinformatics software tools, licenced under Academic Free license (AFL), that can handle anything from sequence analysis to molecular dynamics tasks. It can be booted from DVD or USB storage key and features data persistence. Many thanks to the Department of Bioinformatics, SRM University, Ramapuram campus, Chennai, India.
- Some of the tools integrated into the applications described in this article include the 3D Fast Fourier Transform Library for Blue Gene/L and author Chris Ward's Custom Math Functions for High Performance Computing.
- With IBM trial software, available for download directly from developerWorks, build your next development project on Linux.

### Discuss

- Get involved in the My developerWorks community; with your personal profile and custom home page, you can tailor developerWorks to your interests and interact with other developerWorks users.

## About the authors

### T.J. Christopher Ward

Chris Ward joined the IBM UK Development Laboratories in Hursley, England in 1982 with a degree in Engineering from Cambridge University. He has worked on the development of many products for IBM, from disk files to branded middleware. He is privileged to be working on technology that is as valuable to IBM's future customers as IBM WebSphere Software and IBM Lotus Software are to the IBM customers of today.

---

### Ruhong Zhou, Ph.D.

Ruhong Zhou is a Research Staff Scientist at the Computational Biology Center/ IBM Thomas J. Watson Research Center and an Adjunct Professor at the Columbia University Department of Chemistry. He received his Ph.D. with Bruce Berne in chemistry from Columbia University in 1997. He joined IBM Research in 2000 after spending two and a half years working with Richard Friesner (Columbia) and William Jorgensen (Yale) on polarizable force fields and protein-ligand binding mechanisms. He has authored 80 journal publications and 7 patents, delivered numerous invited talks at major conferences and universities, and chaired several conferences in computational biology and chemistry and biophysics. He won the Hammett Award in 1997 from Columbia, the DEC Award in 1995 from the American Chemical Society on Computational Chemistry, and the Outstanding Technical Achievement Award in 2005 and 2008 from IBM. His current research interests include development of novel methods and algorithms for computational biology and bioinformatics and large-scale simulations for protein folding, ligand-receptor binding, and protein structure prediction.