Contents

Protein modelir /ith Blue Gene/

I.J. Christopher Ward and Ruhong Zhou

Published on June 09, 2009

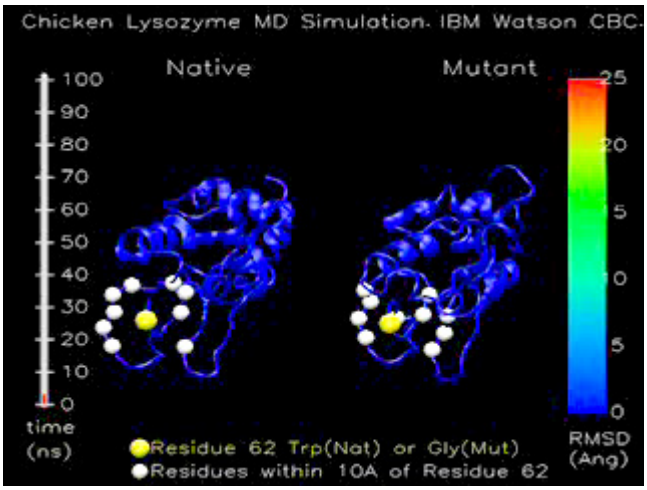
Real-world scientific advances thro deling and data visualization on a

In 2001, IBM's research scientists starte sign of a new family of servers, now ma
Gene®. These servers have been availabl since 2004—first the Blue Gene/L (wh
then the Blue Gene/P.

The Blue Gene family of supercomputers ned to deliver ultra-scale performance
environment, they're also designed to pr iciencies in power, cooling, and floor-s
universities, government, and commercia ch labs use Blue Gene for computation
protein folding, climate research, cosmology, and drug development. The system is mak
magnitude, in the way science can be done, because it offers a more cost-effective tool
alternative versions of complex models.

In this article, we present some of the progress that has been made by one of the projec
modeling. Figure 1 shows the scale of work we can do now, thanks to the power of Blue
starts from the lysozyme crystal structure (see Related topics for source).

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



Proteomics: The pr n economy

Contents

Introduction

Proteins are biological macromolecule e an essential component of organisms a within cells. Many proteins are enzyme atalyze biochemical reactions; some are i immune responses; many others have al and mechanical functions for muscles illustrate how pervasive and important s are:

One protein is responsible for the " " of blood; it carries oxygen from the lung

Another protein is responsible for t an body's response to the poison in poiso generally harmful.

There are hundreds of thousands of pr volved in life on Earth. Proteomics is the they interact, and how their diversity a alization evolve among the living organis tour of what proteins are, how they are and how they affect the systems they inha

DNA is the information storage compo very cell in every plant and animal. It sto chemical building blocks (nucleotides) we call **A**, **C**, **T**, and **G** (for adenine, cytosine, thyn replacing thymine in RNA). From a distance, these building blocks look very similar, so e 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 F internal pattern. Rather like pressing a key into putty, you now have an image of the key next presented as a blueprint to the ribosome, a protein that behaves like an all-purpos 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. Th the food we eat (humans cannot synthesize all the amino acids we need and therefore n "essential" amino acids, from food). Each amino acid has a "head" and a "tail." The ribos

acid for each "letter" and assembles them head-to-tail in sequence; other "letters" indic
The resulting linear sequence of amino acids is a newly minted protein molecule, forme
imprinted in the section of DNA that was used.

Stresses and strains between the atoms in the protein molecule, interactions with the sl
random vibrations that you would call *heat* then cause the protein molecule to "fold" int

Protein molecules are quite stable; some of them can exist unchanged for hundreds of y
hundreds of degrees, which would kill the organism that made them. They stay roughly
denatured by strong chemicals, high pressure, heat or cold, or by becoming food for anc

The shape and the way it varies with ti perature, and surrounding molecules det
will do—whether it will transport oxyge ou a poison-ivy allergy, or do any of the o
tiny scale.

Figure 2 demonstrates the familiar bal ick model of DNA (image is a stereo pair;
source):

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

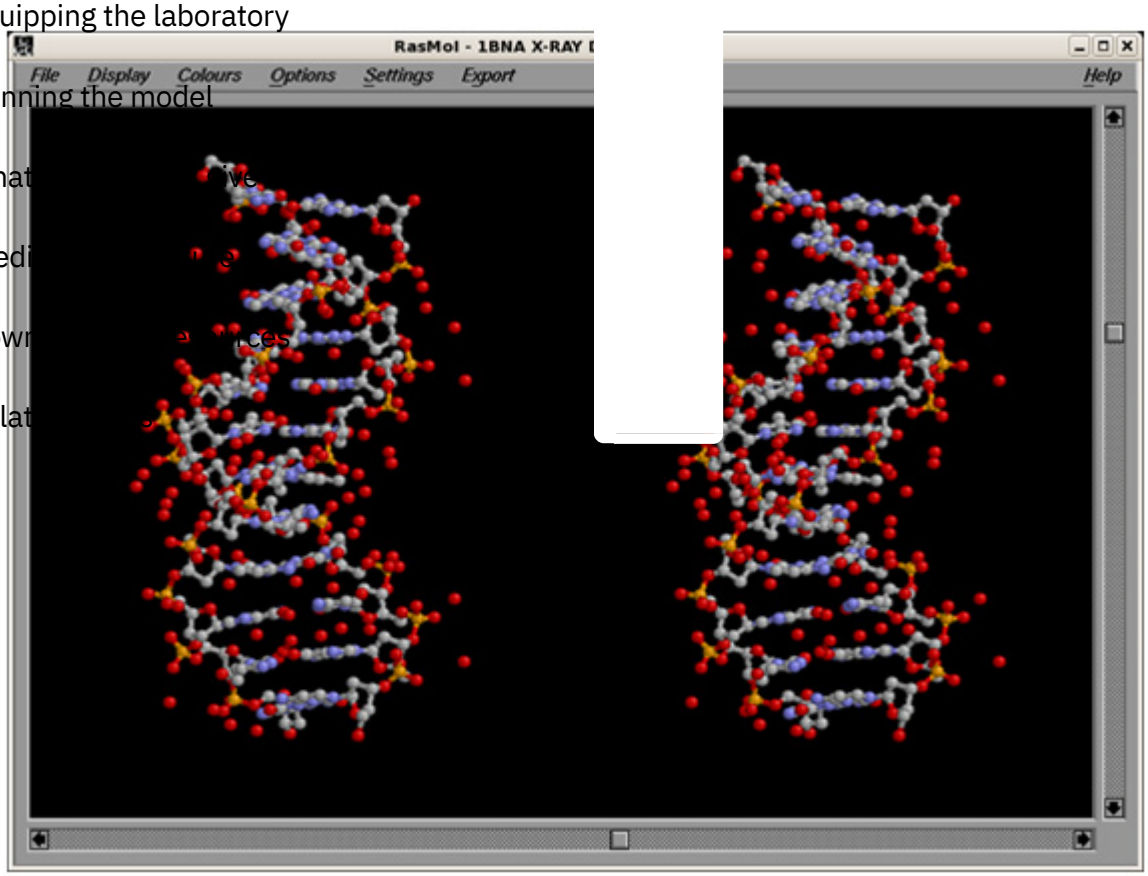
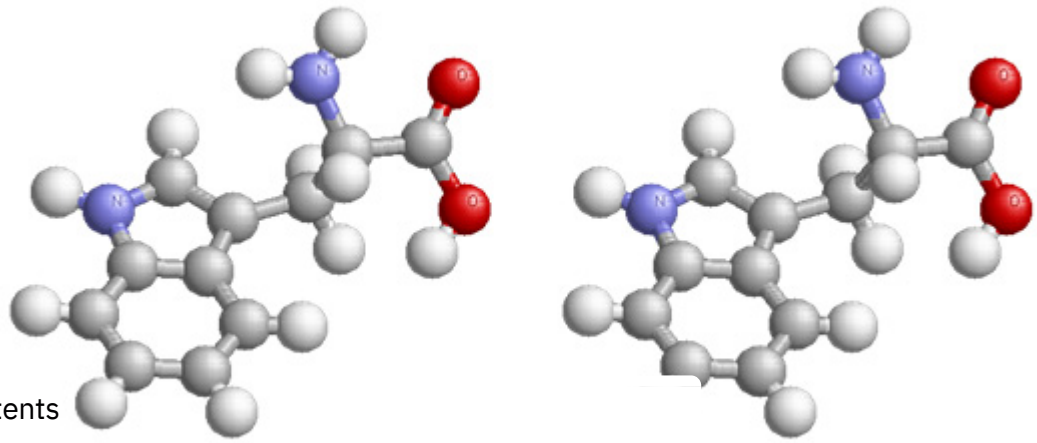


Figure 3 shows tryptophan, one of the 20 standard amino acids (image is a stereo pair;
source).

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

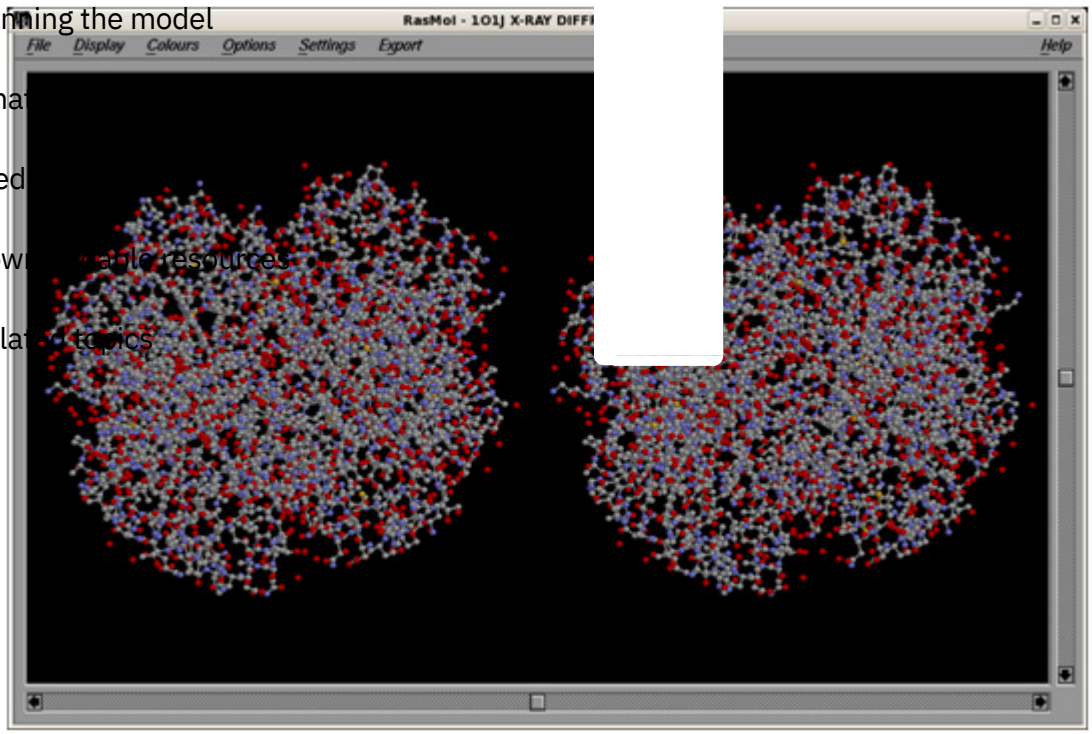


Contents

Amino acids are assembled into protei tting off the O-H group (right side of Figu away from the N (top of Figure 3) of an olecule, and splicing the molecules togetl water molecule. All amino acids have t acteristic atomic grouping (top right of Fig

Figure 4 offers a visual look at the prot oglobin (image is a stereo pair; see Relate

Figure 4. The protein hemoglobin



Hemoglobin is a total of 574 amino acid molecules in 4 subunits. Hemoglobin, with its a assembled into the protein is beyond the scope of this article), transports oxygen aroun transport system is possible with just the iron atoms, but it is very much more effective of hemoglobin provides. If you put this image into a stereo viewer, you can pick out the

more complex than this, we need a different way to visualize what is going on.

Commercial and academic motivations

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. current topic is *protein misfolding and aggregation*—if a protein folds into a shape other 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 (M , Huntington's and Parkinson's, cystic fibr and other amyloidosis.

Understanding what can cause protein les to change from their useful folded for a different folded form is an active res pic in the search for a treatment for these significant diseases. Recent experimer ered by Chris Dobson and co-workers at Cambridge University (see Related top link) have shown that amyloids and fibrils be formed not only from the traditiona myloid peptides but also from almost any proteins (such as lysozyme) given the ate conditions. In fact, a single mutation (W62A) on lysozyme protein can cause tein to be in a much less stable state as compared to the wildtype (see sidebar also cause it to misfold and form possible amyloids in urea solution due to the lo "long-range hydrophobic interactions."

Scientists do not yet know how this sir residue can play a key role in the hydrophobic long-range interactions d folding process and then shift to the sur from presumably a nucleation site for ing reasons. This offers a unique opportu for better understanding of the single effects, as well as the mechanism behin the aforementioned diseases related to protein misfolding and aggregation.

The Blue Gene/L technology offers a powerful way to study these types of diseases, bec 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 a single mutation. Lysozyme is a protein that is part of the human immune system; wher 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 differe lysozyme molecule. The theory is that this different amino acid affects the shape that th differently shaped lysozyme molecule is differently effective in puncturing bacterial cell change, we may be able to design pharmaceuticals or other forms of therapy that will as recovering from bacterial disease.

As part of the work, we store the positions and velocities of every atom in one molecule approximately 10,000 water and urea molecules (this simulation is done in an 8 molar u experiments), in the computer's memory. There are many ways to model the forces betv *ball and spring* model for bonded forces with an inverse-square-law model for electrost and an attract/repel model for atoms t ıear each other but not covalently bonde

At each time step, we calculate the for ach atom, then we update the velocities a Newton's second law.

At each time step (very small, on the o . femtosecond), there are in principle hun calculated. This large number of calcu nd the fact that we also want to be able t (microseconds) to model motions of in ıeans this approach has only recently bec computers that we know how to build. e details on what we do and some alterna "Destruction of long-range interaction: ıgle mutation in lysozyme" in Related top

Equipping the labor y

At IBM Watson Research Lab in Yorkto / York, we have 20 racks of BlueGene/L se PowerPC® dual-core microprocessor c ch microprocessor is attached to 512MB *compute lattice*, there is an additional ocessor connected to a 1Gbps Ethernet lir connected through standard Ethernet s to standard IBM Power Systems machin compilers, and job-control software.

This lysozyme modeling work has used an average of four racks of BlueGene/L processc an aggregate of more than 10 microseconds of molecular dynamics data. Periodically, th positions and velocities of all the atoms under simulation (part of this stream of informa synthetic video mentioned above). Whenever it is necessary to restart the simulation ru and velocities can be reloaded. Restarting may be needed after a planned shutdown, aft in order to replay a model event of scientific interest with a different time step granulari

Running the model

The application is booted onto the Blue Gene/L nodes by a mechanism similar to MPICH available, portable implementation of MPI, the message-passing interface; see Related the cluster provides a POSIX-file-system environment to the application. Data can be se System (GPFS) file system for the application to read; when the application writes result for external use.

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

What does all this give us?

Contents

The video is a glimpse into a world tha ver before been visible. Of course, we dor scientists always need to compare wh el shows with what they can see in the re misfolds in reality is still a dream; ever ;" part of the "fixed" conformations mean them under an electron microscope or / even causing large numbers of lysozyme X-ray diffraction spectroscopy. Howev experimental techniques typically do not might move.

Therefore, the current large-scale simu offer a unique window to look into the de critical changes involved in disease-re sfoldings. Hopefully, the availability of the happen will push the envelope and ad e state of the art in amyloidosis studies. I generation of scientists to solve these problems in this new way as their primar research.

Predicting the futur

Actually, we would not be so bold as to attempt to divine tomorrow, but we would ventu computing will continue to follow a development path (we use version L; the available B processors per chip, 10Gbps Ethernet, and a host of other improvements). The cost of d arithmetic and the cost of more and faster storage (both heavily associated with the dat this article) will most likely continue to fall—as they must, because there is several worl scientists need to be doing, both for public research and for businesses to bring product

The lysozyme model we describe only scratches a molecule off the surface of the new fi are more than 50,000 proteins whose structures are catalogued in the public Protein Da link); there are millions of potential pharmaceutically useful compounds to be analyzed;

diseases known to be related to proteins and their defects. And we're not even consider that can benefit from modeling on this scale. Blue Gene's work has just begun.

Downloadable resources

PDF of this content

Related topics

Contents

A roundup of Blue Gene/L-related can be found at the IBM Blue Gene proje components and resources include

The IBM Blue Gene/P solution

The IBM General Parallel File S

Everything you could want to k ut the IBM XL C/C++ Advanced Edition fo

IBM Redbooks on Blue Gene te ies

And a picture of the Blue Gene one of the authors

The RCSB Protein Data Bank (PDB) chive for the study of biological macromo experimentally determined structu oteins, nucleic acids, and complex assem include such cool things as the Mo the Month.

Source data for Figure 1 is from the ne 1.33 A structure of tetragonal hen egg

Source data for Figure 2 is from the tructure of a B-DNA dodecamer: conform

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,H5

"Destruction of long-range interactions by a single mutation in lysozyme" (R. Zhou, N Berne; Proc. Natl. Acad. Sci., 2007) gives more information about the modeling appr

"Parallel implementation of the replica exchange molecular dynamics algorithm on E Rayshubski, J. W. Pitera, B. G. Fitch, R. Zhou, R. S. Germain; IEEE, 2006) explains so used for the simulation.

MPICH2 is the next stage of MPICH, the high-performance, widely portable (and free Passing Interface (MPI) standard.

The Argonne Leadership Computing Facility has a collaborative program that provide

computational science community.

"High-performance Linux clustering" is a two-part series providing background on hi
Linux. Part 1 (developerWorks, September 2005) covers HPC fundamentals, types o
choosing a cluster configuration, and the role of Linux in HPC. Part 2 (developerWork
programming using MPI, covers cluster management and benchmarking, and shows
open source software.

"Port Fortran applications" (developerWorks, April 2009) helps you overcome comm
applications among various high performance computing systems.

Some of the tools integrated into the applications described in this article include th
for Blue Gene/L and author Chris W ustom Math Functions for High Performan

In the developerWorks Linux zone, re resources for Linux developers, and sc
tutorials.

See all Linux tips and Linux tutorial eloperWorks.

Comments

Sign in or register to add and subscril nments.

Subscribe me to comment notificatio

Select a language

English

中文

日本語

Русский

Português (Brasil)

Español

홑

Code Patterns

Articles

Tutorials

Recipes

Open Source Projects

Videos

Newsletters

Events

Cities

Developer Answers