Proteins in motion

Biochemists are systematically simulating the unfolding pathways of all known protein folds, hoping to discover the general principles of protein folding

Proteins do the work of life; they are essential to the structures and functions of all living cells. Some proteins play structural or mechanical roles, others catalyze chemical reactions, and still others are involved in storage or transportation or immune response.

A protein's biological function depends on its shape. Proteins are synthesized as long, non-branching chains of amino acids; but in order to perform their proper functions, each type of protein must assume a unique, stable, and precisely ordered three-dimensional folded structure, which biologists call its *native state.* Protein folding might be compared to a shoelace tying itself, except the shapes of proteins are much more complicated than knots, and the folding process is much faster, being completed only microseconds to seconds after the protein is created.

Incorrectly folded proteins are responsible for many illnesses, including Alzheimer's disease, cystic fibrosis, Huntington's disease, Parkinson's disease, Type II diabetes, Creutzfeldt-Jakob disease (the human counterpart of mad cow disease), and many cancers. The degradation of protein folding may even play a key role in aging. But despite the importance of protein folding, the phenomenon remains largely a mystery.

"Protein folding is one of the fundamental unsolved problems in molecular biology," said Valerie Daggett, Professor of Medicinal Chemistry and Adjunct Professor of Biochemistry, Biomedical and Health Informatics, and Bioengineering at the University of Washington. "Although we know a lot about of the structural details of the native folded conformation of proteins, very little is known about the actual folding process. Experimental approaches only provide limited amounts of information on the structural transitions and interactions occurring during protein folding. Given that protein folding is of such widespread importance to human health, we are using computer simulation methods in an attempt to delineate the important forces acting during this process."

Daggett is principal investigator of an INCITE project called "Molecular Dynameomics," which was awarded 2 million processor-hours at NERSC. Just as genomics studies the complete genetic makeup of an organism, and *proteomics* studies the complete protein structure and functioning of an organism, *dynameomics* is an ambitious attempt to combine molecular dynamics and proteomics, using molecular dynamics simulations to characterize and catalog the folding/unfolding pathways of representative proteins from all known protein folds. There are approximately 1,130 non-redundant protein folds, and Daggett's group has used their INCITE grant to simulate proteins from the 151 most common folds, which represent about 75% of all known protein structures (Figure 25).

"Structure prediction remains one of the elusive goals of protein chemistry," Daggett said. "In order to translate the current deluge of genomic information into knowledge about protein functions, which we can then use for drug design, we have to be able to successfully predict the native states of proteins starting with just the amino acid sequence."

The novelty of her group's computational approach, however, is that they work in the opposite direction, starting not with a protein's amino acid sequence but with its folded structure, as determined by NMR spectroscopy and X-ray crystallography experiments. A protein can easily be unfolded or denatured, in both experiments and simulations, by simply raising the temperature. (Anyone who has ever cooked an egg has seen the results of denaturation: the heat causes the bonds within the egg white proteins to break, unfolding the individual protein molecules and allowing them to form bonds with other molecules, thus solidifying the egg white.) Daggett's research over the last several years suggests that protein unfolding follows the same pattern as folding, moving in reverse through the same transition states and intermediate structures. The high temperature of the simulations, 498 Kelvin (437° F), just speeds up the process without modifying the overall pathway, she says.

To perform these simulations, Daggett's group uses a code called "*in lucem* Molecular Mechanics" or *il*mm, which was developed by David Beck, Darwin Alonso, and Daggett. The code solves

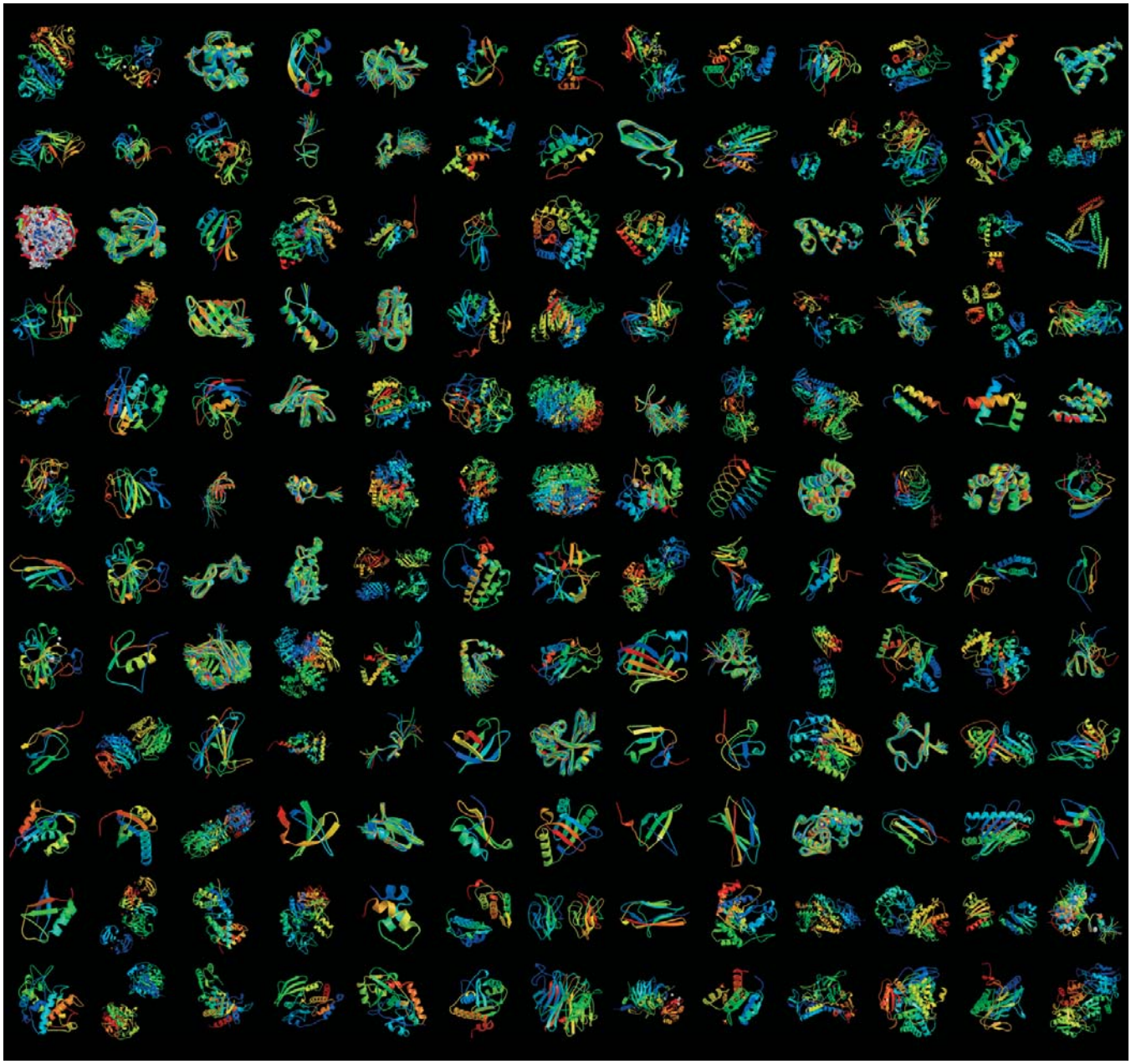

FIGURE 1. The first 156 dynameomics simulation targets.

Newton's equations of motion for every atom in the protein molecule and surrounding solvent environment, determining the trajectories of the atoms over time. Obtaining a reasonably good picture of protein unfolding pathways requires six simulations for each protein: one 20 nanosecond (ns) simulation of the native (folded) molecular dynamics at a temperature of 298 K or 77° F (Figure 26); and five simulations at 498 K, three at 2 ns duration and two at 20 ns (Figure 27).

Thus the 151 proteins studied in the INCITE project required a total of 906 simulations and produced 2 terabytes of compressed data. Even with the allocation of 2 million processor-hours, the researchers had to limit the size of the chosen proteins to no more than 300 amino acids. Fortunately, the larger, more complex proteins are generally structured as if they were assembled from a number of smaller shapes, so understanding the smaller shapes should reveal principles that are applicable to all proteins.

In fact, Daggett hopes that cataloging the folding strategies for every type of protein fold will enable the discovery of general rules of protein folding, which so far have eluded researchers, and that those rules will be used to improve algorithms for predicting protein structure. The results of the protein unfolding simulations will be made publicly available on the Dynameomics website (www.dynameomics.org). Just as the Protein Data Bank has become an important repository of experimental results that has enabled many further discoveries, Daggett envisions the Dynameomics simulation database as a resource for the whole research community, in which data mining may produce unexpected findings and a dynamic description of proteins will aid in understanding their biological functions.

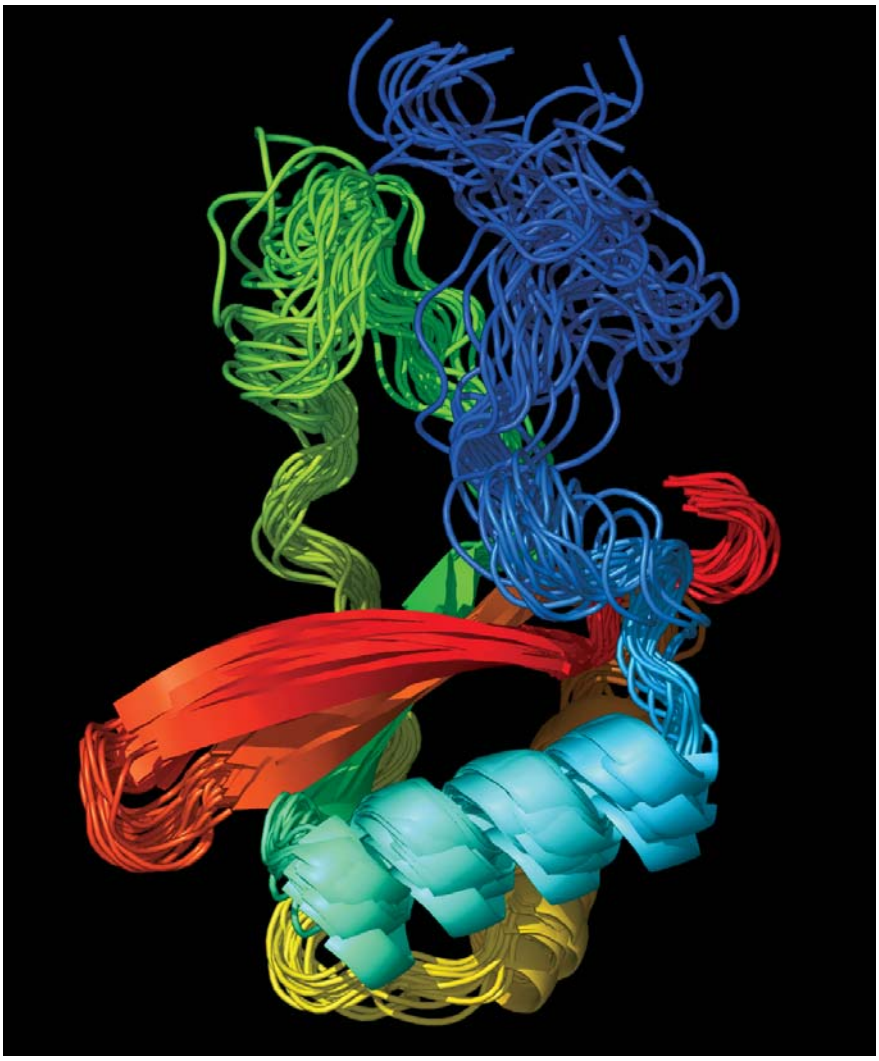

FIGURE 2. Protein target 3rub: A schematic representation of the secondary structure taken from a native state simulation of the enzyme RuBisCO, the most abundant protein in leaves and possibly the most abundant protein on Earth. Water, hydrogen, and protein atoms included in the simulation are not shown. The highly dynamic region (image top) is responsible for finding the protein's binding partner (not included in the simulation). RuBisCO catalyzes the first major step of carbon fixation, a process by which atmospheric carbon dioxide is made available to organisms in the form of energy-rich molecules such as sucrose.

David Beck, a graduate student in the Daggett lab, worked with NERSC consultants to optimize *il*mm's performance on Seaborg. "The INCITE award gave us a unique opportunity to improve the software, as well as do good science," Beck said. Improvements included load balancing, which sped up the code by 20%, and parallel efficiency, which reached 85% on 16-processor nodes. The INCITE award enabled the team to do five times as many simulations as they had previously completed using other computing resources.

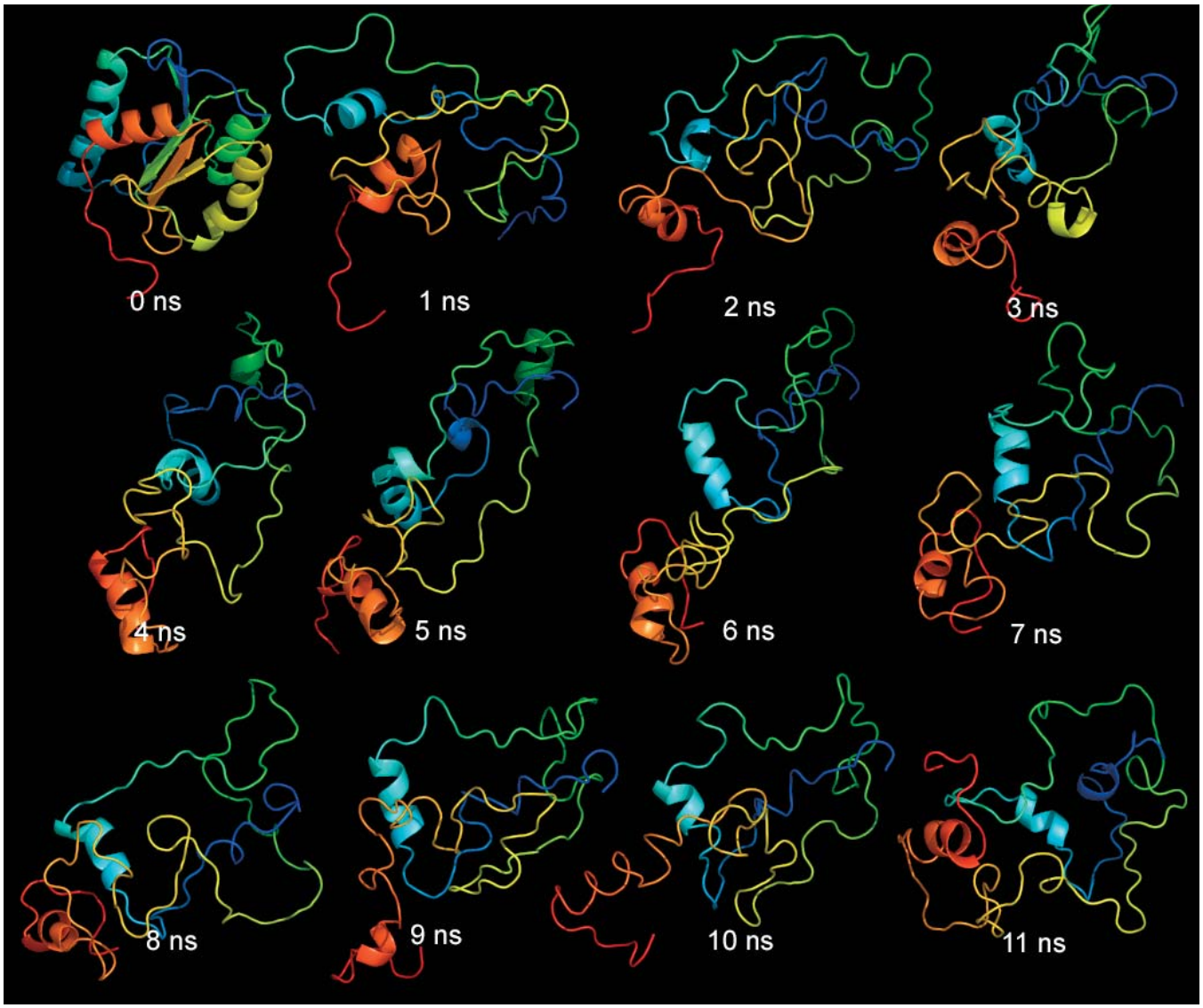

FIGURE 3. Protein target 1imf: Schematic representation of secondary structures taken at 1 ns intervals from a thermal unfolding simulation of inositol monophosphatase, an enzyme that may be the target for lithium therapy in the treatment of bipolar disorder.

The researchers are now analyzing the simulation results and comparing them to experimental results whenever possible. "Our preliminary analysis of this data set focused on the native control simulations and the endpoint of the thermal unfolding simulations—the denatured state," Daggett said. "In particular, we are evaluating the interactions between amino acids. We have already found a shift in the prevalence and types of interactions as the proteins unfold. Characterizing the interactions in the denatured state provides much-needed information about the starting point of folding. This represents the first step towards providing crucial information for prediction of the folding process and ultimately the folded, native structure."

And, of course, there are still hundreds more protein folds to simulate.

Research funding: INCITE, BER
Computational resources: NERSC
This article written by: John Hules