High resolution, high throughput protein structure prediction using IBM Blue Gene supercomputers: Predicting CASP targets in record time

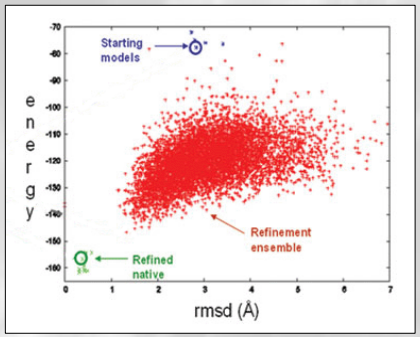
Ross C. Walker[1], Srivatsan Raman[2] & David Baker[2,3]

Introduction

With genome-wide sequencing completed for a number of organisms including humans, the next challenges are to functionally characterize the proteins encoded by these genes and to understand their roles and interactions in cellular pathways. High-resolution 3-dimensional structures of proteins can help in functional annotation and also explain the underlying molecular mechanisms of protein interactions. This could constitute a powerful tool towards target drug development. Despite considerable technical advances, however, the experimental determination of protein structures by nuclear magnetic resonance and x-ray diffraction techniques remains slow, expensive, and arduous. In particular, the rate at which protein structures are experimentally solved is lagging far-behind the explosive rate at which protein sequence information is being gathered by high-throughput genome sequencing efforts. Thus, a high throughput methodology to computationally predict protein structures with atomic level accuracy from protein structure is highly desirable and has been a major focus of the Baker research group. The Rosetta software suite has been under development in Prof Baker's research group for the last 15 years and currently there are over 80 developers across 6 university campuses actively improving and extending the code.

The Scientific Challenge

It is well understood that the three dimensional structure of a protein is encoded in its primary amino acid sequence. With severable notable exceptions, most proteins are known to occupy the global minimum of the free energy landscape. Computational protein structure prediction involves varying the degrees of freedom of the protein structure in a constrained manner and evaluating the energy until it converges. To ensure that the structure is not trapped in a local minimum, we carry out a large number of independent trajectories each starting from different random seeds. The lowest energy model from this large population is predicted to be the native structure of the protein.



A structure prediction starts with a series of initial models being generated from a sequence comparison against proteins of known structure. These initial guesses are then subjected to a Monte Carlo refinement where each structure is minimised and it's energy calculated using energy functions developed for this purpose. The ensemble of structures with the lowest energy are taken to be the native state.

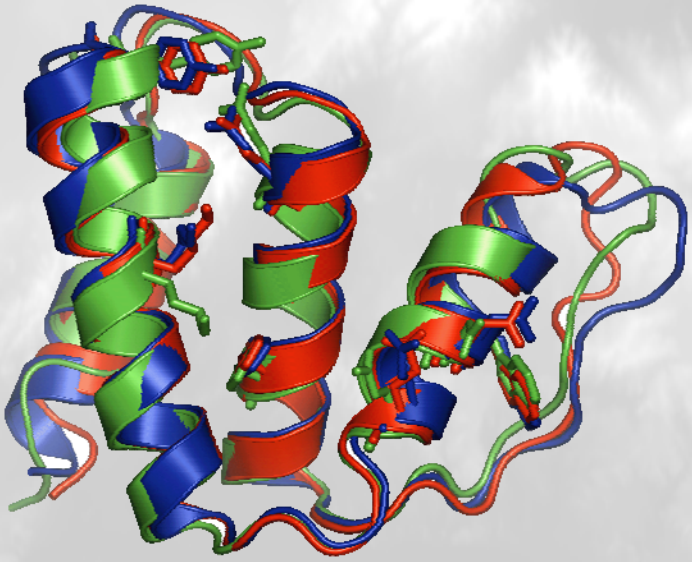
The problem is challenging because the size of the conformational space to be searched is vast, requiring large computing power to sample diverse structural conformations. The difficulty is further compounded by the presence of many local minima in the free energy landscape of a protein. As a result, if the sampling is insufficient, many of the conformations get trapped in these local minima without representing the full structural diversity. Thus, to guarantee that we see conformations close to the global minimum, we need a large ensemble of structures. Hence the more computing power available the better the prediction is likely to be since statistically more of phase space is sampled.

[1]San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA
[2]Dept. of Biochemistry, University of Washington, Seattle, WA
[3]Howard Hughes Medical Institute and Dept. of Biochemistry, University of Washington, Seattle, WA

Acknowledgements
The authors would like to thank IBM for graciously providing access to their Blue Gene Watson machine and San Diego Supercomputer Center and Argonne National Lab for providing computational resources to the Baker team during CASP 7.

The CASP Competition

Critical Assessment of Techniques for Protein Structure Prediction (CASP) is a biennial community-wide experiment for protein structure prediction. It is often referred to as the "world cup" of protein structure prediction and provides the research groups an opportunity to assess the quality of their methods for protein structure prediction compared to those of other groups across the world. Thus, the outcome of CASP showcases the state of the art in this field to the scientific community.
The most recent CASP competition (CASP 7) ran from May 10th to August 29th.



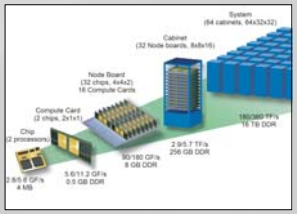
Predicting Protein Structures
Discovering the 3-D structure of proteins opens the door to understanding their function. In a Strategic Applications Collaboration (SAC), SDSC computational scientist Ross Walker helped HHMI investigator David Baker and his group at the University of Washington modify their Rosetta code to utilize 40,960 processors, all working in parallel to compute the high resolution structure prediction shown. The computation run on IBM's Watson Blue Gene supercomputer was part of a large number of Rosetta calculations run for the CASP 7 competition. Red is the native X-ray structure, blue is the Rosetta prediction and green a low resolution NMR model. The entire prediction was completed in under six hours with the computational component taking three hours.

Overcoming Bottlenecks

The organizers of CASP release about 100 targets over a 4 month period. The deadline for submission of the prediction is 3 weeks after the release date of the target. Each target requires a large volume of computing for a reliable prediction to be made. Due to the limited availability of computing resources, we typically have to adopt an iterative approach to predicting the structure which is less reliable than a direct approach. Furthermore, with a large fraction of time spent on computing, there is very little time left for thorough post-production analysis of the models. The data analysis is as important as generating the models themselves. Prediction accuracy could be vastly improved if more time were available for data analysis. Since there is approximately 1 day available per CASP target if one is to carry out predictions without shortcuts that could adversely affect the quality of the predictions then it is necessary to complete all of the required computation in approximately 3 hours or less leaving sufficient time available for analysis and further computation if required.

Utilizing Blue Gene/L for Structure Prediction

As part of the CASP 7 competition we modified the Rosetta protein structure prediction code to allow it to run on the Blue Gene/L architecture (pictured). This architecture is ideal for protein structure prediction since it incorporates a very large amount of computational power in a reliable homogenous package. The memory footprint was reduced and all I/O was done in parallel fashion. Since the time taken to refine each guess is not always constant a job distributor was introduced to control load balancing. Output from the distributor could then be used to monitor the overall progress of prediction and if necessary the job parameters could be changed on the fly. A schematic of the calculation is shown below. This allowed the Baker team to utilize SDSC and ANL Blue Gene systems to aid in the prediction of a number of CASP targets.



A schematic of the Blue Gene/L system construction. (*image courtesy IBM*)

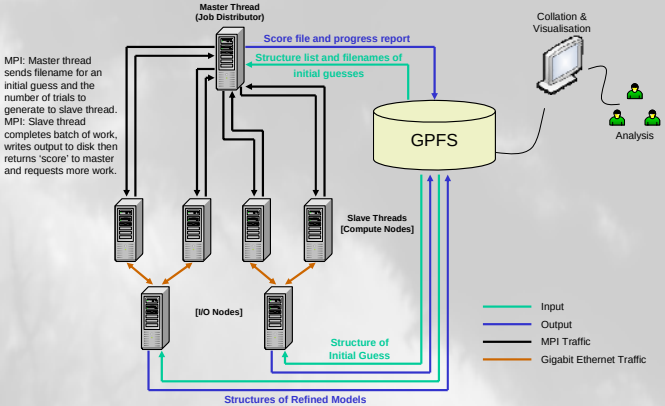
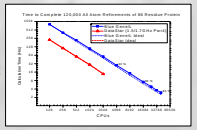


Diagram showing communication patterns within the Rosetta code while carrying out a structure prediction on IBM Blue Gene/L. A job distributor running on the master node controls load balancing as well as producing summary information about the prediction progress. The slave threads receive their computation instructions from the job distributor in batch form over MPI. They each write the results of their structure refinements in parallel to disk and communicate back to the master summary information regarding the batch of refinements and request more work.

Direct CASP Target Prediction in 3 hours

As part of the CASP 7 competition we also decided to carry out a demonstration run on the Blue Gene/L system at the IBM T.J.Watson research laboratories (picture opposite). The scaling of the Rosetta code (shown below) is such that we could carry out a complete, direct, structure prediction of a large CASP target in under 3 hours. In the spirit of the competition we took a target released that morning. Constructed the initial guesses and then submitted the job to all 20 racks (40,960 cpus) of IBM's Blue Gene. The calculation ran for 3 hours generating over 120,000 data points. We then immediately analysed the results. Made our structure prediction and submitted it to the CASP organisers. In all the entire process took less than 6 hours and resulted in the prediction shown in the central image.

Photograph of IBM Blue Gene Watson Machine (*image courtesy IBM*)



Rosetta scaling on Blue Gene/L as a function of cpu count.