Computational challenges for modeling and simulating biomacromolecular assemblies

Ed Uberbacher, Philip LoCascio, Sergey Passovets, Pavan Ghattyvenkatakrishna, Pratul Agarwal, Nikita Arnold, Andrew Bordner, Andrey Gorin

Life Sciences Division and Computational Sciences Division
Oak Ridge National Laboratory, Oak Ridge TN 37831-6164

ube@ornl.gov

Abstract. Understanding the structure and dynamics of large biomolecular assemblies requires the development of new computational methods for (i) accurate structure prediction, (ii) molecular docking and (iii) long time-frame molecular simulation, and implementation on massively parallel computing infrastructure. This paper reviews our progress in these areas and applications on important molecular systems.

1. Context and Outline

A major next step in understanding and utilizing complex biological systems is a capability to rapidly model and simulate the structure, assembly and dynamics large assemblies of macromolecules. Molecular machines are the basis for life's chemistry. They malfunction in human disease, can be used to produce energy in microbes and from biomass, and provide the possibility of innovative new drug therapies at the interface of nanoscience and biology.

The ORNL Computational Biology Program is focusing on the key steps necessary to build and simulate large molecular machines: (1) *Computing Accurate Building Blocks*: Building accurate models of molecular machines relies on a capability to build accurate component protein or nucleic acid structures. (2) *Putting the pieces together*: Given relatively accurate starting models, the components must be docked properly to create models for the molecular assembly. (3) *Seeing how they work*: Once models are built, simulating complex molecular machines at biologically meaningful time scales requires thoughtful problem design and current and next generation capability computers. These three key areas are addressed in the following sections.

2. Computing accurate starting protein models

Structural Genomics Programs are investing huge sums with the assumption that good structures can be created by improved homology modeling systems [1]. Currently, while approximate structures can be obtained using existing infrastructure with techniques such as homology modeling, creating accurate computational models of protein and nucleic acids is generally beyond current capabilities. We are developing conformational search methods, constrained by multiple homology examples, which can greatly improve the accuracy of derived protein models.

Even with proper alignment, homology models usually have about 4A RMS error [2]. This error is too large to permit meaningful computational simulation / molecular dynamics. The goal is to reduce the errors in initial models. Our approach is to multi-align the group of evolutionary neighbors using sequence and structure information in order to find the most stable parts of the structure. The extracted stable core structure is typically 20-30% closer in terms of RMSD to a given prediction target than any single original template. Once this core structure has been established, more flexible parts of the target are modeled locally by choosing most sequence similar loops from the library of local segments found in all the homologues. Thus the prediction target is modeled using specific parts of multiple known structures as well as structural elements they all share.

The procedure typically reduces the RMS error in the predicted structure from about 3.4-4A to 2-2.5A. Figure 1 shows a typical computed structure compared to the target, improved from 3.4A using standard homology modeling methods, to 2.3A

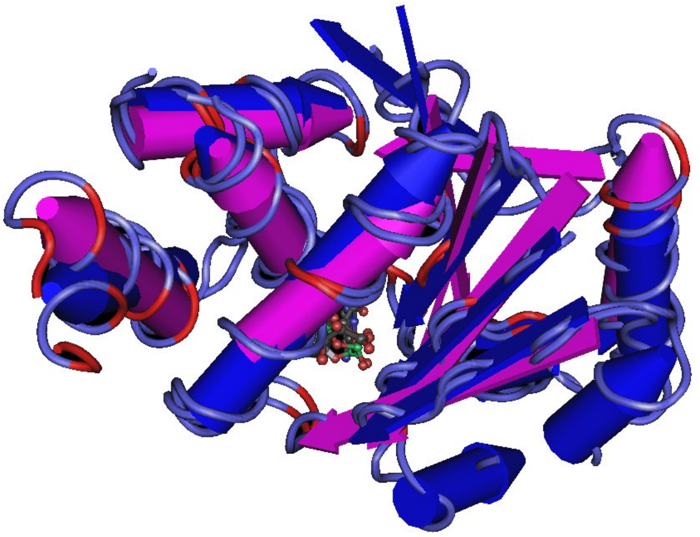

Figure 1. Computed model and crystallographic 1fq1 target protein at 2.3A RMSD

Our plan is to further improve these Intermediate accuracy models using a genetic algorithms conformation search strategy which uses backbone and sidechain angles as genetic parameters. This involves very large conformational searches in the neighborhood of the intermediate accuracy structure that will utilize shared memory space to house the population of molecular conformations. A key issue in this development is the production of code libraries for backbone and sidechain rotations (under control of the GA), which do not accumulate errors and have sufficient accuracy to avoid Cartesian errors at distances far away from the axis of rotation. Libraries for this have been developed and tested on the ORNL NCCS Cray XT3.

3. Predicting interaction interfaces and binding partners solely from backbone structural information

Understanding of the physical principles governing binding interfaces is key to a capability for predicting novel protein interactions and elucidating important cellular mechanisms. A significant body of work has been accumulated in the field on residue conservation [3], biased interface composition [4], and pair-wise residue preferences [5], but major contradictions remain regarding what are the real features of protein interfaces and how knowledge of those features can be used for *in silico* interface discovery.

In our study, we explored one important issue: how well prediction of the protein interfaces and docking partners can be done when only protein backbone information available. This question remains poorly explored while having major practical implications. First, homology models on the backbone resolution level are available for significant parts of the sequenced genomes, and robust methods based on backbone structure information could lead to extraction of the significant functional insights. Second, backbone-based methods will almost necessarily be fast enough to dock thousands of potential candidate structures, providing an opportunity for obtaining genome wide functional insights from such computational predictive studies.

Our approach includes three major components: (i) a new algorithm for the exhaustive sampling of all possible docked configurations for two subunits (Figure 2); (ii) Bayesian potentials for the scoring of the obtained configurations; and (iii) calibration curves predicting the likelihood that a given configuration is a native one from its score value [6]. In the new sampling procedure, the protein surface is defined using a series of normal vectors that are spaced 1.5 A apart. Docking is realized by anti-aligning the surface normal vectors on each protein surface and rotating about their common axis [7]. This ensures efficient sampling of the relative conformational space since the protein surfaces are touching at a point with common normals. For a medium size complex, the number of the generated "trial" configurations for this 1.5A grid is in the range of 100 or so. The Bayesian framework for the potentials can accommodate fairly simple approaches, such as a simple count of specific contact pairs, or very sophisticated ones, involving relative orientations of C_{alpha}-C_{beta} bonds for the contacting residue pairs, higher order motifs of many contacting residues, etc.

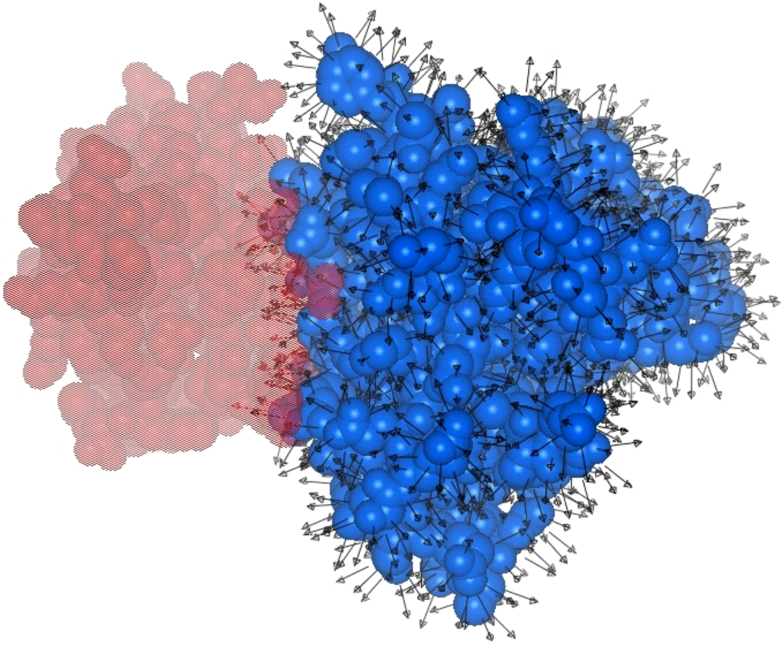

Figure 2. Diagram showing surface normals for two proteins which are combinatorially anti-aligned to generate the set of candidates for the docked structure. A rotational search is carried out around the axis of each pair of anti-aligned vectors with each result scored as described in the text.

High performance computing implementations ideally fit our approach as more sophisticated parameter schemas require higher fidelity of the decoy complexes, and the increased fidelity can be guaranteed by a finer grid on the surface of the corresponding proteins. The resulting applications are naturally parallelizable up to 1000s processors. The protein surface can be divided up into small patches with every computing thread given its own group of decoys to score, while reporting to a centralized score stack where the calibration curves are maintained and analyzed.

So far we have implemented the simplest Bayesian potential (a count of contacting residue pairs) and extensively tested the whole pipeline in docking studies for a very large benchmark data set with over 1200 non-homologous and manually verified protein-protein complexes. Our method finds near native conformation in 72% of the complexes in the top scoring 1% of decoys (it is 247 decoys per complex on average). As the fraction of such conformations in the top 1% is more than 20%, it is usually enough to draw 5-10 decoys at random from the top percentile in order to get at least one native conformation for these 72% of the complexes. We also have formulated rules that will allow us to determine if a given protein complex falls into the category (those 72%), for which the prediction algorithm works.

Finally, we have assembled a large collection of the independently crystallized subunits corresponding to the protein chains in the tested complexes (with less that 3A RMSD and over 80% sequence identity). The method performance on the independently crystallized subunits was essentially identical to one we have described above. This result validates our strategy and proves that for a certain categories of the protein complexes it is possible to devise algorithms predicting interaction surfaces solely from backbone structural information.

4. Toward dynamic mechanisms of molecular machines

Once models are built, simulating complex molecular machines at biologically meaningful time scales requires thoughtful problem design and significant capability computing. For example, rational engineering to improve cellulases, which are key to biomass energy initiatives, will require a very detailed understanding of molecular mechanisms of catalysis and processivity. Classical molecular dynamics simulations of cellulases are being carried out using models with sizes in the range of 700,000 – 1,000,000 atoms and for periods between 20 and 100 nanseconds. The initial model used was a cellulase CBH I enzyme [8], linker and binding domain, equilibrated in water, built by John Brady (Cornell University) and provided by Mike Himmel (National Renewable Energy Laboratory - NREL). A 20ns simulation was carried out on a system of 711887 atoms including water molecules. Data was collected at a rate of 1ns every 9hrs when calculations were performed on 1024 processors of the CRAY XT3 using the LAMMPS [9] code. The CHARMM force-field [5] was used to perform the calculations and constant volume and temperature simulations were found to be sufficient since the density of the initial system was close to the equilibrium density. Simulations were carried out with CBH I on a fibril of cellulose containing 108 strands with each strand containing 40 glucose units. Initial studies focused on large scale motions of the enzyme and its active site cleft. Figure 3 shows this system.

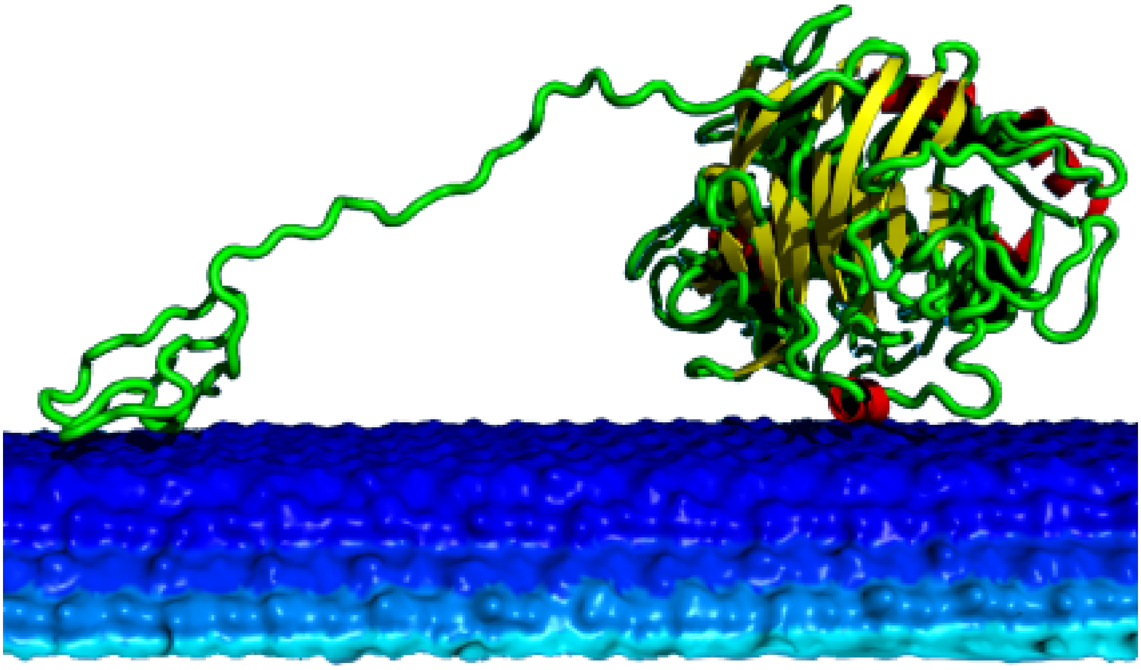

Figure 3. Cellulase CBH 1 enzyme complex with cellulose fibril showing binding domain (left), linker (center) and catalytic domain (right). A single cellulose polymer chain is extracted from the fibril surface and threaded into the catalytic domain for processing into single sugars. The details of this mechanism are not understood and computational simulation is being used to explore this.

As part of a consortium working on cellulase mechanics, continuing work is focused on how the nicked polymer strand is recognized by portions of the enzyme, how the cellulose polymer strand becomes oriented and fed into the catalytic domain (right domain) and how the central linker acts to guide the strand. Other large problems involving molecular complexes including Alzheimer's disease related amyloid fiber assembly [6] and simulation of drug binding to membrane bound receptors are being carried out on the ORNL Cray XT3 using molecular mechanics and quantum mechanics / molecular mechanics approaches.

References

[1] D. Baker, A. Sali. Protein structure prediction and structural genomics. Science 294(5540):93-6 (2001).

[2] C. Chothia, A. M. Lesk. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823-826 (1986).

[3] Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J. and Huang, E.S. (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?, Protein Sci, 13, 190-202.

[4] Jones, S. and Thornton, J.M. (1996) Principles of protein-protein interactions, Proc Natl Acad Sci U S A, 93, 13-20.

[5] Murphy, J., Gatchell, D.W., Prasad, J.C. and Vajda, S. (2003) Combination of scoring functions improves discrimination in protein-protein docking, Proteins, 53, 840-854.

[6] Arnold N.D, Bordner A. J., Tian Y., Uberbacher E.C. and Gorin A. (2006) Analyzing protein-protein complexes using large-scale rigid docking with Bayesian residue-based contact score (in preparation)

[7] Bordner, A. J. and Gorin A. (2006) Protein docking using surface matching and supervised machine learning (submitted to Proteins)

[8] Himmel, M.E.; Adney, W.S.; Baker, J.O.; Elander, R.; McMillan, J.D.; Nieves, R.A.; Sheehan, J.; Thomas, S.R.; Vinzant, T.B.; Zhang, M. (1997). "Advanced Bioethanol Production Technologies: A Perspective." Woodward, J.; Saha, B., eds. *Fuels and Chemicals from Biomass*, ACS Series 666, Washington, DC: American Chemical Society; pp. 2-45.;
Himmel, M.E.; Baker, J.O.; Saddler, J., eds. (2001). "Glycosyl Hydrolases for Biomass Conversion." ACS Symposium Series 769, Washington, DC: American Chemical Society: Distributed by Oxford University Press.;
Sheehan, J.; Himmel, M.E. (1999). "Enzymes, Energy, and the Environment: Cellulase Development in the Emerging Bioethanol Industry." *Biotechnology Progress* (15:3); pp.817-827

[9] S. J. Plimpton, R. Pollock, M. Stevens, "Particle-Mesh Ewald and rRESPA for Parallel Molecular Dynamics Simulations", in Proc of the Eighth SIAM Conference on Parallel Processing for Scientific Computing, Minneapolis, MN, March 1997.;
S. J. Plimpton, "Fast Parallel Algorithms for Short-Range Molecular Dynamics", J Comp Phys, 117, 1-19 (1995).

[10] MacKerell, Jr., et al. All-atom empirical potential for molecular modeling and dynamics Studies of proteins. Journal of Physical Chemistry B, 1998, 102, 3586-3616.;
Brooks, I., C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. In The Encyclopedia of Computational Chemistry; A. D. MacKerell, J., B. Brooks, Ed.; John Wiley & Sons: Chichester,1998; Vol. 1; pp 271.

[11] Ma B, Nussinov R. Stabilities and conformations of Alzheimer's beta -amyloid peptide oligomers (Abeta 16-22, Abeta 16-35, and Abeta 10-35): Sequence effects. Proc Natl Acad Sci U S A. Oct 29;99(22):14126-31. (2002).