

Computational Biology and Biophysics

With the recent successful completion of the Human Genome Project and related attempts to determine whole genomes it has become obvious that the obtained wealth of data needs to be matched by information on the structure and function of the huge number of proteins encoded in the various genomes (“proteomics”). As proteins are only functional if they assume specific shapes, it is important to explore how these structures emerge from a protein’s chemical composition (the sequence of amino acids as specified in the genome). Such knowledge could not only lead to the rational design of novel drugs, but also to a deeper understanding of various diseases that are caused by misfolding of proteins, and enable simulation of disease pathways to predict the best target for treatment.

Given a sufficiently accurate description of the forces between the atoms in a protein, and between a protein and the surrounding environment, it is theoretically possible to simulate the folding and thus predict the biologically active configuration of a protein. However, the complex type of the interactions containing both repulsive and attractive terms leads to a very rough energy landscape with a huge number of local minima separated by high energy barriers. Hence, a typical thermal energy of the order $k_B T$ is much lower than the energy barriers that the protein has to overcome at room temperature. As a consequence, the simulation of proteins becomes a hard computational task. For instance, Carlos Simmerling’s (State University of New York, Stony Brook) ongoing studies on the stability of the tiny 20-residue Trp-cage protein and three of its mutants require already up to 1000 processors of the LeMieux system (a 6 TFlops cluster comprised of 750 Compac Alphaserver ES45 4-processor nodes) to simulate the native form of the protein and three mutants for 20-50 ns [1]. Typical folding times however are on the order of \approx seconds, i.e. more than 10^7 times longer, and stable domains in proteins usually contain 50-200 residues. As the computational effort in canonical molecular dynamics and Monte Carlo simulations increases exponentially with the size of the molecule, folding simulations using all-atom models are presently restricted to molecules with up to 30-40 residues and sub- μ s trajectories. Extending the size of molecules that can be studied in computer experiments is one of the defining challenges in high performance computing requiring both new hardware and new algorithms.

The immense amount of data involved multiplies the computational difficulties. Even if only 1% of the DNA sequences deposited in GenBank encode proteins without closely homologous structures, by 2008 there will be 300,000 protein sequences/year whose structures need to be predicted. Assuming that with fold recognition techniques such as threading (i.e. searching a fold library for the structure which fits the sequence best) and improved Monte Carlo techniques the time for predicting a structure could be reduced to 1 day on an average processor (2GFlops), this would still require 40 days on a 15 TFlops supercomputer.

Various European groups study folding, misfolding and aggregation by computer

experiments. Examples are the groups lead by Prof. Berendsen (University of Groningen, Netherlands), Prof. van Gunsteren (ETH Zürich, Switzerland), Prof. Kolinski (U. Warsaw, Poland), Prof. Irbäck (University of Lund, Sweden), Dr. Wenzel (FZ Karlsruhe, Germany) and Prof. Hansmann (starting July 2005, NIC, Jülich, Germany). Structure prediction using data mining is also a prominent approach investigated at the European Molecular Biology Laboratory (EMBL) in Heidelberg and Hinxton, the Max-Planck-Institut für Informatik (Prof. Lengauer), and various other places. Polish groups (Ginalski, Kolinski and Bujnicki) were extremely successful in last years CASP6 competition of protein-structure prediction groups and techniques. However, most computational protein research today originates in the USA, with the centers in San Diego (UCSD, Scripps Institute), Harvard University, University of Illinois at Urbana Champaign, Cornell University and the Thomas J. Watson Research Center (IBM). This US preeminence is at least in part due to better computational resources. For instance, Jeff Skolnick's group (Center for Excellence in Bioinformatics, Buffalo) alone has exclusive access to 4500 processors with a total of about 7 TFlops [2]. Utilized by the US protein research community is also the Helix system of the National Institutes of Health (NIH) consisting of the shared memory multiprocessor SGI 3400 supercomputer Nimbus and 2500+ processor Biowulf Linux cluster [3]. Other accessible computational resources include the data-intensive IBM Linux cluster (4 TFlops peak performance and 540 terabytes of network disk storage) and the DataStar IBM 10-TFlops supercomputer at the San Diego Supercomputing Center that will also receive in 2005 a 5.7 TFlops IBM eServer Blue Gene system. A total of 40 TFlops of computing power with a Petabyte-scale data storage is available since October 2004 with TeraGrid that operates a 40 gigabyte-per-second network and pools computational resources from Argonne National Laboratory, Center of Advanced Computing Research at Caltech, Indiana University, National Center for Supercomputing Applications, Oak Ridge National Laboratory, Pittsburgh Supercomputing Center, San Diego Supercomputing Center and Texas Advanced Computing Center. TeraGrid is used currently by the group of Prof. K. Schulten (University of Illinois at Urbana-Champaign) for the simulation of membrane proteins and the recognition of DNA by proteins [4].

The computational resources in the USA will see an unprecedented increase over the next few years. For instance, IBM expects to install a 100 Teraflop eServer Blue Gene system at the Thomas J. Watson Research Center dedicated to life sciences by March 2005 [5]. As part of their *Roadmap for medical research* the National Institutes of Health (NIH) attempts to implement in the USA the core of a universal computing infrastructure for the life sciences. Four centers specialized in the area of biomedical computing have been funded in September 2004. These centers are supported by 5 year research grants with a total of 79.7 Million US\$. Funding of three more centers is planned for 2005. The centers aim at "paving a future information highway dedicated for advancing medical research" and training a new generation of multidisciplinary biomedical computer scientists.

In order to stay competitive, computational biology/biophysics in Europe will need to match the American resources. Access to supercomputers with around 10-50 Teraflops

peak performance as a minimum need to be available between now and 2007. This will allow the European groups to study the folding process in small proteins below 50 amino acids. Especially interesting is the simulation of misfolding and aggregation of small peptides in a detailed representation. A candidate for such research could be the 42-residue peptide β A. This peptide is involved in Alzheimer's disease whose impact on the health care systems and economies will likely grow even larger in the near future with the increased longevity and resulting aging of societies in Europe. Computational resources of this size could enable also substantial progress in the comparison and analysis of protein structure. For instance, Shindyalov & Bourne report a performance of 24,000 CPU hours for an all vs. all comparison of all 11,000 PDB structures using 2000 representatives on a Cray T3 (900MFlops/processor) [6]. Assuming the current size and growth rate of the Protein Data Bank (PDB, the worldwide repository for protein configurations), it will reach 60,000 structures by 2008. An all vs. all comparison with e.g. 8000 representative structures would still need about 1 day on a 15 TFlops supercomputer, but would allow now a much more through clustering of protein structures to study structure-function relationships. Problems such as structure motif detections may require even larger resources. Current structural genomics initiatives are producing an increasing number of structures with unknown function, and structural motifs provide a very sensitive technique to identify protein function. Altman et al. reported an estimated performance of 177 processor hours/feature for recognition of structural motifs by their FEATURE program when screening 11,000 protein structures of the PDB using a Sun 300MHz processor (600MFlops). Assuming again 60,000 structures in the PDB by 2008, a 15 TFlops Supercomputer will need only 139 s per feature for screening the PDB [7]. However, given the size of current sequence motif databases the estimated number of features will be in the range of 1000-5000 for a comprehensive annotation. In addition, predicted structures are likely to have achieved sufficient quality for feature annotation as well, raising the actual number of structures to be screened to several hundred thousands. An increase in computing power to 50-250 Teraflops until 2009 may allow tackling these tasks. With the anticipated algorithmic improvements such increase in computational resources would also allow study of proteins with 80-100 residues in computer experiments. Simulation of transport of proteins through membrane proteins (in atomistic detail), re-folding of proteins, flexible docking and the assembly of nanostructures by proteins are other topics of investigation of practical importance. For instance, re-folding studies could lead to better treatment of severely burns, and many diseases are related to improper transport of proteins through membranes. Assuming access to a European supercomputer center with a peak performance on the order of 1 Petaflop available after 2010, it would be possible to extend these studies to even larger systems (proteins of order 100-150 residues) which would include many potential drug targets and to the systematic investigation of ensembles of proteins. A possible application would be the collection of a library of structural changes in proteins through mutations that are observed in humans. This would be an important step on the way toward a "personal medicine", where a drug and its dosage are chosen according to the specific mutation of a protein in a patient.

Besides tackling these scientific questions, establishment of a network of 3 European and various national supercomputer centers in Europe will also lead to an increased collaboration of the European groups which will be crucial for successfully competing with US research. As in the roadmap initiative of the NIH, the European centers will have both to establish an information highway for life sciences and to offer training in biomedical computer science. The computational resources of this network will allow for the transition from the study of fundamental questions in protein physics toward an engineering approach utilizing its powers e.g. to help in the design of novel proteins with customized properties for use as drugs, novel materials, sensors etc. Hence, these centers and the specialists trained there will become an important factor in maintaining the competitiveness of European pharmaceutical and medical research in both industry and academia.

References

- [1] <http://www.psc.edu/science/simmerling.html>
- [2] Jeff Skolnick, private communication
- [3] <http://helix.nih.gov>
- [4] http://www.psc.edu/publicinfo/news/2004/2004-10-08_teragrid.html
- [5] <http://www.research.ibm.com/bluegene/>
- [6] Ilya N. Shindyalov and Philip E. Bourne (2000), PROTEINS: Structure, Function, and Genetics 38:247260
- [7] A. Waugh, G. A. Williams, L. Wei, & R. B. Altman. Using Metacomputing Tools to Facilitate Large Scale Analyses of Biological Databases. Pacific Symposium on Biocomputing 2001, Mauna Lani, 360-371. 2001.