ER 25134

DOE/ER/25134-72

High Performance Computing and Communications Grand Challenges Program

Computational Structural Biology
(Caltech Component)

Jerry E. Solomon (P.I.)
The Beckman Institute
California Institute of Technology
Pasadena, CA 91125
(October, 1994)

Co-Investigators
Prof. Alan Barr
Prof. K. Mani Chandy
Prof. William A. Goddard III
Dr. Carl Kesselman

Professional Staff
Huy Cao
David Liney
John Garnett

Graduate Students/PostDocs
Dr. Michael Holst (Research Fellow, CRPC)
Moira Regelson (Graduate student, Applied Mathematics)

Collaborators
Prof. David Baker (University of Washington)
Prof. Leroy Hood (University of Washington)
Argonne National Laboratory
Center for Research in Parallel Computing (NSF STC)

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Grand Challenge Project
Computational Structural Biology

I. Background. The so-called protein folding problem has numerous aspects, however it is principally concerned with the *de novo* prediction of three-dimensional (3D) structure from the protein primary amino acid sequence, and with the kinetics of the protein folding process. Our current project focuses on the 3D structure prediction problem which has proved to be an elusive goal of molecular biology and biochemistry. Traditionally there have been two major approaches to this problem, the first being attempts to solve the (essentially deterministic) equations of motion which drive the multi-atom protein system to an equilibrium configuration. This is usually referred to as the molecular dynamics (MD) approach. The details of this approach are described quite well in McCammon and Harvey (1987). The second major approach is stochastic in nature and utilizes Monte Carlo methods to explore the space of possible protein conformations to find the one having the global free energy minimum. Both methods are plagued by essentially the same problem, which is that the number of local energy minima is exponential in the number of amino acids in the protein. Thus, all current methods of 3D structure prediction attempt to alleviate this problem by imposing various constraints that effectively limit the volume of conformational space which must be searched. The work of Skolnick's group (*cf* Kolinski and Skolnick (1994a), (1994b)) is typical of the stochastic approach to the protein 3D structure prediction problem.

Our Grand Challenge project consists of two elements: (1) a hierarchical methodology for 3D protein structure prediction; and (2) development of a parallel computing environment, the Protein Folding Workbench, for carrying out a variety of protein structure prediction/modeling computations. During the first three years of this project we are focussing on the use of selected proteins from the Brookhaven Protein Data Base (PDB) of known structures to provide validation of our prediction algorithms and their software implementation, both serial and parallel. Two proteins in particular have been selected in collaboration with Prof. David Baker (University of Washington, Biochemistry Department) and Prof. Lee Hood (University of Washington, Molecular Biotechnology Department) to provide our project with direct interaction with experimental molecular biology. Prof. Baker, in cooperation with the Molecular Biotechnology Science and Technology Center (NSF), is carrying out a variety of site-specific mutagenesis experiments on these two proteins to explore the many-to-one mapping characteristics of sequence to structure. Both proteins, protein

L from *peptostreptococcus magnus*, and *streptococcal* protein G, are known to bind to IgG; and both have an $\alpha + \beta$ sandwich conformation. Although both proteins bind to IgG, they do so at different sites on the immunoglobulin and it is of considerable biological interest to understand structurally why this is so.

II. Protein Structure Prediction Approach Our approach to the protein structure prediction problem is a hierarchical one which combines a coarse level stochastic Monte Carlo algorithm with a fine-grained molecular dynamics calculation. A block diagram description of this approach is shown in the attached figure. The coarse-grain level utilizes a lattice model to represent the protein C^{α} backbone structure and also incorporates a side-chain representation which is illustrated in the accompanying figure labelled Figure A-9. In this figure, the small red spheres represent C^{α} lattice site positions, while the blue-hatched larger shperes represent side-chain sites. This model allows one to automatically account for excluded volume and side-chain steric hinderance effects which are present in real proteins. Since it is well-known that hydrophobic effects provide the major driving force for protein folding (Dill (1990)), we represent the protein sequence as a copolymer consisting only of hydrophobic (**H**) and polar (**P**) elements. This is essentially the model devised by Chan and Dill (1991). The effective topological contact potentials for the resulting intramolecular interactions are derived from distance distribution statistics of the PDB.

Utilizing this lattice model we then apply a novel guided replication Monte Carlo chain-growth algorithm developed by Garel and Orland (Garel and Orland (1990); Garel, *et.al.* (1991)), and recently reviewed in Bascle, *et.al.* (1993), to generate a large ensemble of candidate lattice chains from the input primary amino acid sequence. Details of this method are described in Solomon and Liney (1993). The novelty of this method lies in the use of a "guide field" which can be used to insert folding constraints that modify the Boltzmann weight associated with prospective moves in the lattice chain growth process. This allows the introduction of *a priori* knowledge regarding the folding characteristics of real proteins in a way that cannot be achieved through the use of the system Hamiltonian alone. We have used this technique to introduce such constraints as: (1) hydrophobic residues prefer the protein interior; (2) charge residues have a strong tendency to appear on the protein surface; and (3) postulated secondary structure is introduced through the use of partial contact maps.

Once an ensemble of candidate lattice chains is generated, the next major step is to "filter" the ensemble for chains which are likely to be close to a natively folded conformation. Thus, given a chain ensemble of size $\mathcal{N}$, our problem is to select a subset, of size $m \ll \mathcal{N}$, based on some set of "native-state" properties. There are two, basically equivalent, ways of accomplishing this: (1) minimize a property vector, $\vec{u} = \{\alpha_1 u_1, \alpha_2 u_2, ..., \alpha_K u_K\}$; and (2) minimizing a linear combination of native-state properties,

$$v = \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_K u_K. \tag{1}$$

In both cases the $\{\alpha_i\}$ represent scale factors which control the weight given to any particular property, u_i. Note that the $\{u_i\}$ are constructed from the native-state properties such that smaller values correspond to being "close" to the native-state values. The subset selection algorithm is then simply,

$$min[m]\{|\vec{u}_i|^2\}, i = 1, 2, ..., \mathcal{N}, \tag{2}$$

and

$$min[m]\{v_i\}, i = 1, 2, ..., \mathcal{N}, \tag{3}$$

We are currently using the following property set to evaluate the likelihood of a candidate structure being "close" to a natively folded state: (1) radius of gyration, R_g; (2) total non-bond energy, E_{nb}; (3) solvent accessible surface area, A_s; (4) number of hydrophobic- hydrophobic contacts, $\mathcal{N}_{hh}$; and (5) the ration of (sequence) local contacts to the total number of topological contacts, $\mathcal{N}_{lc}$. Since MD algortihms generally do not produce structures which are far from their starting conformation, usually no more than a few angstroms *rms*, we subject our selected subset to a structural diversity test. The structural similarity measure which we use is the so-called difference of distance maps measure, based on the distance map matrix, $\{d_{i,j}, i, j = 1, 2, .., \mathcal{N}\}$. In this representation, the $d_{i,j}$ are the pairwise distances between all backbone C^α coordinates, and the distance measure is defined as

$$D_d(S, S') = \frac{1}{\mathcal{N}} \left[\sum_{i,j}^{\mathcal{N}} (d'_{i,j} - d_{i,j})^2 \right]^{1/2}. \tag{4}$$

For the purposes of ensuring that the subsample chosen for off-lattice relaxation contains only unique structures separated by some minimum distance, D_{min}, we utilize the distance measure defined above.

Having selected a subset of candidate lattice chains, the structures are relaxed off-lattice to obtain their "minimum energy" free-space backbone conformations. The resulting free-space structures are then subjected to a set of minimization and MD procedures, developed by the Goddard group, to produce full-atom free-space structures. The minimum free energy structure of the resulting set is then taken as the final predicted conformation for the protein.

III. The Protein Folding Workbench Even though the Monte Carlo chain generation algorithm discussed above is computationally quite efficient, the need to generate large (¿ 10^7) ensembles, and the requirement to filter this large set of candidate chains, dictates the use of high performance parallel computing resources. We have taken two approaches to parallelization of our Monte Carlo code: (1) simple domain decomposition, where N processors simply generate N independent ensembles which are merged at the end of the computation; and (2) parallelization of the algorithm, where N processors generate a single ensemble, and the parallelization occurs at the scaling step of the chain-growth process. In order to provide a general parallel computing environment for protein structure prediction/modeling computations we have designed and implemented what we term a "Protein Folding Workbench". The design details of this workbench may be found in Kesselman, *et.al.* (1994). The workbench is implemented in the CC++ (concurrent C++) programming language and provides three levels of access depending on the user's programming sophistication. The top level provides an icon-driven interface that allows unsophisticated users to set up protein folding calculations and specify output formats in graphical fashion. The middle level allows creation of new functionality by combining primitive functions contained in the lowest level class libraries, and is based on the extension language TCL developed at UC Berkeley.

IV. Key Results During FY94 we have had one paper accepted for publication in *Biopolymers*, Solomon and Liney (1993), and have presented a paper, Liney and Solomon (1993) at a national symposium on computational biology. In addition, two technical reports describing the Protein Folding Workbench and utilization of the CC++ programming language have been produced; Kesselman, *et.al.* (1994), and Foster, *et.al.* (1994). During this period we cite the following major accomplishments: (1) validation of the guided replication Monte Carlo method applied to generation of lattice model protein structures; (2) Incorporation of side-chain representations and torsional interactions into the discrete lattice model Monte Carlo code; (3) Demonstration that the native structure is contained in the lattice chain ensemble for the test case of *streptococcal* protein G, Brookhaven designation 1PGX; (4) Implementation of the functional libraries for the Protein Folding Workbench; and (5) Completed initial validation tests of the guided replication Monte Carlo algorithm as implemented in parallel in the Protein Folding Workbench.

V. Future Research Plans During FY95 we expect two major activities to be that of developing robust and consistent "native-state" filters for our chain ensembles, and completion of the Protein Folding Workbench implementation. We expect to accomplish the following major milestones:

(1) Full parallel implementation of our advanced version of the Monte Carlo chain-growth algorithm on the Argonne National Laboratory IBM SP-1 computers. This will include "production" runs for structure prediction of *streptococcal* protein G, and protein L from *peptostreptococcus magnus.* We will also use this capability to

predict conformational changes associated with site-specific mutagenesis experiments carried out by Prof. David Baker of the University of Washington.

(2) Demonstration of the complete hierarchical prediction methodology (including free-space calculations) on our two target proteins (G and L).

(3) Incorporation of electrostatic effects (solvation) in intramolecular interactions of the folded protein in collaboration with Dr. Michael Holst of the NSF STC Center for Research on Parallel Computation.

(4) Publication of our parallel algorithm and protein structure prediction results in *Biopolymers* and *J. Comp. Chemistry.*

VI. Collaborations with Other Institutions Our Computational Structural Biology project involves significant collaboration and interaction with three organizations: (1) The NSF STC for Molecular Biotechnology, University of Washington; (2) Argonne National Laboratory, Mathematics and Computer Science Division; and (3) The NSF STC Center for Research on Parallel Computation.

NSF STC for Molecular Biotechnology: This collaboration involves Prof. Leroy Hood (Department of Molecular Biotechnology, Unviersity of Washingto), and Prof. David Baker (Department of Biochemistry, University of Washington), and constitutes the major biology and biochemistry input into our project. As pointed out above, Professors Hood and Baker are conducting a series of site-directed mutagenesis experiments on two proteins which are of biological significance due to their unique binding properties with respect to the immunoglobulin IgG. Our project is providing the structural prediction capabilities required in this study to investigate the conformational changes (if any) produced by replacement/deletion of specific amino acids in the primary sequence of these proteins. We will also be using our computational capabilities to explore the probable binding mechanisms involved since it is known that these two proteins bind IgG at different sites.

Argonne National Laboratory: Our project has an ongoing and active collaboration with Drs. Ian Foster and Rick Stevens of Argonne National Laboratory involving the development and application of computational libraries for the CC++ programming language. The Computational Structural Biology project is also scheduled for 12,000 node-hours of IBM SP-1 use during the period of October, 1994 through January, 1995.

NSF STC Center for Research on Parallel Computation: We have established an active collaboration with the CRPC in cooperation with Dr. Michael Holst, a CRPC postdoctoral research fellow. Dr. Holst is involved in developing fast numerical methods for solving the Poisson-Boltzmann equation for both inter-molecular and intra-molecular electrostatic interactions in the presence of a solvating medium. Our project has enabled Dr. Holst to parallelize these calculations for implementation on the CRPC Intel Paragon machine; and to use these calculations to investigate the electrostatic contributions involved in protein-enzyme interactions. During the coming year, Dr. Holst will be working with our group to incorporate his Poisson-Boltzmann solver into a refinement of our structure prediction algorithms which can accomodate solvation effects in the folding process.

References

J. Bascle, T. Garel, and H. Orland (1993), *J. Phys. I (France)*, **3**, 259-275.

H.S. Chan and K.A. Dill (1991), *J. Chem. Phys.*, **95**, 3775-3787.

K.A. Dill (1990), *Biochem.*, **29**, 7133-7155.

I. Foster, C. Kesselman, and S. Tuecke (1994), *Nexus: Runtime Support for Task-Parallel Programming Languages*, Technical Report, CCB-94-04/May, 1994, The Beckman Institute, Center for Computational Biology, California Institute of Technology, Pasadena, CA 91125.

T. Garel and H. Orland (1990), *J. Phys. A*, **23**, L621-L626.

T. Garel, J.C. Niel, H. Orland, and B. Velikson (1991), *J. Chim. Phys.*, **88**, 2473-2478.

C. Kesselman, H.T. Cao, D. Liney, and J.E. Solomon (1994), *An Extensible Toolkit for Protein Folding Experiments*, Technical Report, CCB-94-03/May 4, 1994, The Beckman Institute, Center for Computational Biology, California Institute of Technology, Pasadena, CA 91125.

A. Kolinski and J. Skolnick (1994a), *Proteins: Struct., Func., Gen.*, **18**, 338-352.

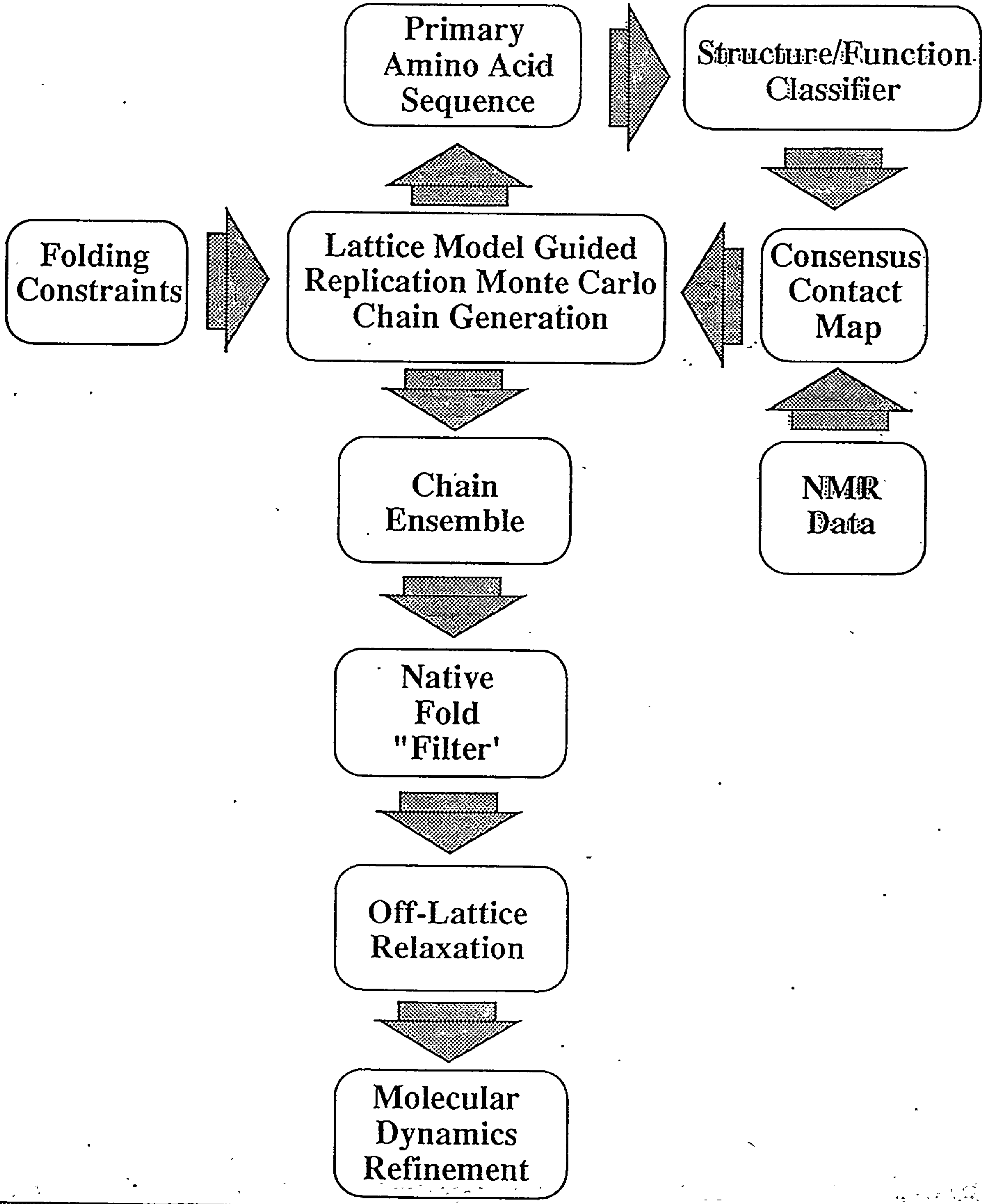
A. Kolinski and J. Skolnick (1994b), *Proteins: Struct., Func., Gen.*, **18**, 353-366.

D. Liney and J.E. Solomon (1993), *Protein Structure Prediction using the Guided Replication Monte. Carlo Method*, Proceedings of the 4th Beckman Symposium on Computational Biology, Pittsburg, PA, October, 1993.

J.A. McCammon and S.C. Harvey (1987) *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press.

J.E. Solomon and D. Liney (1993), accepted for publication in *Biopolymers*; Techical Report CCB-94-06/June 25, 1994, Center for Computational Biology, The Beckman Institute, California Institute of Technology, Pasadena, CA 91125.

HIERARCHICAL PROTEIN STRUCTURE PREDICTION



Primary Amino Acid Sequence

Structure/Function Classifier

Folding Constraints

Lattice Model Guided Replication Monte Carlo Chain Generation

Consensus Contact Map

Chain Ensemble

NMR Data

Native Fold "Filter'

Off-Lattice Relaxation

Molecular Dynamics Refinement