Ph.D. Candidacy Prospectus

A Homotopy Method for Predicting the Lowest Energy Conformations of Proteins

Daniel M. Dunlavy

April 18, 2003

Research Goals

The main goal of this research is to develop a new algorithm for determining the three dimensional conformation that yields the lowest potential energy of a protein. The goals of this algorithm are to:

1. make use of the existing data in the Protein Data Bank (PDB) to aid in predicting the lowest energy conformation for proteins with unknown structure, and

2. overcome the most severe computational bottlenecks from which current algorithms for solving this problem suffer.

The hypothesis being tested is that it is possible to use known tertiary structural information about a protein to determine the same information for other proteins that share common subsequences of amino acids. A further goal is to determine to what extent the amino acid sequences of the different proteins must match in order to guarantee the accuracy of such structural predictions.

The new algorithm will minimize the potential energy of a protein over all possible conformations, given the lowest energy conformation of another protein and a continuous function mapping the molecular compositions of the two proteins. The proposed algorithm is derived from a class of computational techniques called homotopy methods, and will be referred to as such from this point forward.

Background

Milestones in experimental research in protein structure include the sequencing of insulin (Sanger, 1953), determination of the structure of myoglobin via X-ray crystallography (Kendrew, 1961), and determination that the lowest energy conformation of ribonuclease is its native conformation, i.e. the shape in which it performs its function properly (Anfinsen, 1961). The results of these experiments led Vanderkooi, et. al., in 1966, to embark on using computation and simulation to predict the native conformation of a protein from its amino acid sequence. After four decades of similar work by mathematicians, statisticians, computational biologists and chemists, and computer scientists, the goal of those first computational predictions — to solve the *protein folding problem* — still eludes researchers, despite the prominence of the problem in the area of computational biology.

Any useful description of all possible conformations for a protein exceeds the capacity of contemporary computational resources; using the approximation of three possible conformations per residue (alpha-helix, beta-sheet, and coil), a 100 residue protein would have 3^{100} (about 10^{47}) possible conformations. With a teraflop-class computer (10^{12} floating point operations per second), a direct search of all conformations for the one with lowest energy would take at least 10^{35} seconds $\approx 10^{23}$ years. (To put this in perspective, an estimate of the present age of the universe is 10^{10} years.)

There are currently three major computational approaches to solving the protein folding problem: molecular dynamics simulations, bioinformatics, and potential energy minimization. Molecular dynamics simulations concentrate on the force balance of the atoms within a protein

and the resulting (Langevin) dynamics, which are approximated by a stochastic differential equation. Bioinformatics, or the development of knowledge-based approaches to solving problems in the biological sciences, refers to the use and analysis of experimental protein data for predicting conformations for which no experimental results exist. Threading of homologous proteins (i.e. determining the most probable structure by statistically matching small sections of the protein to sections of slightly different, or homologous proteins with known structure) has been the most successful and widely implemented of the bioinformatics approaches to date. The last approach, potential energy minimization, refers to global optimization techniques which attempt to find the conformation with lowest potential energy by minimizing a model of the potential energy of the protein. Simulated annealing, genetic algorithms, branch-and-bound methods, smoothing methods, and gradient-based methods are several of the more well-studied of the minimization methods.

A very good introduction to the protein folding problem along with details and citations of implementation of the most important algorithmic contributions to the protein folding problem can be found in [3]. Detailed biological descriptions of proteins and the energetics involved in the folding of a protein can be found in [2, 4].

It is possible to assess the performance of protein structure prediction algorithms by direct comparison of the results with experimentally determined structure, so that in principle, progress can be accurately measured. The need for effective methods has recently become much more pressing: many of the proteins coded for in newly sequenced genomes are of unknown function. Effective means of predicting structure would be a major aid to establishing function.

Since 1994, systematic community-wide assessment of prediction methods (known as CASP) has provided a wealth of information on the strengths and weaknesses of existing approaches. Data from this assessment show that the community is a long way from finding a general solution despite substantial progress being made in many areas of protein structure prediction. Furthermore, although existing methods are improving, there are a number of serious computational bottlenecks.

Five CASP experiments have now been completed, spanning the period from 1994 through 2002. Analysis of these data shows that when two proteins share clear sequence homology, their tertiary (three-dimensional) structures are similar. CASP data show that the accuracy of modeling is limited by two factors: obtaining a correct alignment between the target protein sequence and that of available template structures, and refining an initial model obtained by copying the template. There was some improvement in alignment quality between CASP1 and CASP2, but no detectable progress since. Furthermore, there has been no progress in the development of refinement techniques, and conventional molecular dynamics approaches have failed. These are some of the areas where the most important open problems lie.

Several potential energy models (sometimes referred to as *force fields*) have been developed for macromolecular systems and specifically protein systems. Some of the more commonly used models used for protein analysis include AMBER, CHARMM, ECEPP/3, GROMOS, and MM3. Details of the AMBER force field, a characteristic example of these models, can be found in [1].

Research Methods

The potential energy minimization problem is quite difficult: there are many variables (the position of each atom in the protein) and a multitude of local minimizers that are far from the global

minimizer. However, to determine the structure of a protein given one with known structure, the relationship between the two proteins can be exploited by using the known structure as a starting point for determining the unknown structure.

A class of computational techniques called homotopy methods [6, 8] will be used in order to implement this idea. Homotopy methods, or in general continuation methods, have been used in the past in computational bioscience research for exploring potential energy surfaces [5] and as part of a smoothing method for protein structure prediction [7, 9] .

The idea of using a homotopy method for solving the protein folding problem is to create a homotopy function that maps the properties (chemical composition and electrostatic properties) of a template amino acid into a target one. Using this function, a sequence of (imaginary) amino acids can be found, starting with the template and ending with the target, with the intermediate amino acids being slightly changed from the one ahead of it. The conformation of an amino acid in the sequence should usually be a small change from the conformation of the one before it, and thus a very inexpensive computational problem.

One potential drawback of this technique is that there could be discontinuous changes when two rather different conformations suddenly become nearly equal in energy. Both conformations should be considered, and the machinery of homotopy methods (tracking singularities in the Jacobian matrix) gives hope of being able to track these cases, even when many of these such cases arise during a computational run.

Rather than using this on single amino acids, as in some free energy perturbation calculations, a template protein will be continuously deformed into a target one, developing a homotopy that allows insertions, deletions, and substitutions. Successful development of a stable algorithm will require choosing a step size strategy (a priori estimates, adaptive steps, etc.), deriving the appropriate terms to be parameterized in the homotopies, and choosing a sufficiently smooth homotopy to guarantee convergence from the template to the target protein.

In preliminary work under the direction of my advisor, Dr. Dianne O'Leary, and Dr. Ron Unger, I implemented a homotopy method for determining the configuration of a chain of monovalently charged (± 1) particles using a simplified AMBER force field, including only the terms modeling the potential energy of interactions between non-bonded atoms:

$$E(X) = E_{coul} + E_{vdw} = \sum_{i=1}^{m-2} \sum_{j=i+2}^{m} \frac{q_i q_j}{r_{ij}} + \sum_{i=1}^{m-2} \sum_{j=i+2}^{m} \varepsilon_{ij} \left\{ \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right\}$$

where $X \in \mathbb{R}^{3m}$ are the Cartesian coordinates of the m particles in the chain and r_{ij} is the Euclidean distance between particles i and j. The terms in E_{coul} represent the pairwise Coulombic interactions (with charges q_i, q_j), and the terms in E_{vdw} represent the pairwise van der Waal interactions via a Lennard-Jones 6–12 potential. The values ε_{ij} and σ_{ij} are the minimum potential energy and sum of van der Waal radii, respectively, for the pair of particles i and j. Although this is a simplified model of the interactions found in proteins, it can model several of the characteristics of protein interactions that cause the most difficulty for optimization algorithms, namely the multitude of local minimizers separated by high energy barriers. This type of simplification has been used by other researchers (see [9]; Gockenbach, et. al. 1994 for examples). There are also several other instances of protein structure prediction using the AMBER force field (e.g., [7]), so this seemed like a suitable choice for an initial model.

Given the conformation of a template chain (a chain whose lowest energy conformation is known), several of the charges on the particles in that chain were changed to produce a target chain. The homotopy method was then applied to use the lowest energy conformation of the template chain to predict the lowest energy conformation of the target chain. Results show that the homotopy method proves to be quite robust, clearly outperforming the current state-of-the-art general minimization algorithms when more than 50% of the charges differed between the template and target chains, and matching the results of those algorithms when less than 50% of the charges differed. My plan is to extend this work to proteins.

The remaining steps for developing a general homotopy method for protein structure prediction include the following:

1. *Creating a generic software interface for the potential energy function to be minimized.* By using a generic interface, several different energy models can be tested and used. This will allow the homotopy method to be independent of the energy model, with the advantage being that future improvements in energy models can be incorporated into the method without any modifications to the core algorithm.

2. *Creating a software interface to the Protein Data Bank.* With over 50,000 known protein structures, the PDB will be used as the primary resource for template proteins. Creating a two-way data exchange interface to the PDB will provide a mechanism for exchanging data between a conformational prediction tool (the homotopy method) and visualization tools that use data in the PDB format.

3. *Developing the homotopy tracing algorithm.* Choices for path tracing algorithms include the standard ordinary differential equation (ODE) solvers (implicit Euler, Runge-Kutta, etc.) and predictor-corrector methods (using an ODE solver step for the for the predictor step and an optimization algorithm for the correction step). The choice of the potential energy model will influence the choice of path following algorithm as well as the choice of molecular properties to be parameterized and choices of homotopy, step size, and initial point to be used. Numerical stability and convergence analyses of the algorithm will be performed for each of the models included in the research.

4. *Validating the method.* Pairs of proteins will be selected from the PDB as template and target proteins and the homotopy method will be applied to predict the tertiary structure of the target proteins. The proteins will be chosen such that a full range of the percentage of changes in sequence will be covered in the experiments. This will aid in determining the extent to which the homotopy method can guarantee accurate results.

References and Reading List

Area of Specialization: Protein Structure Prediction

[1] W. D. Cornell, P. Cieplak, C. L. Bayly, I. R. Gould K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. F. J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nuclei acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[2] T. E. Creighton. *Proteins: Structures and Molecular Properties*. W.H. Freeman and Company, 2nd edition, 1993. *Chapters 1 (1.1–1.3), 4, 5 (5.1–5.3), 7 (7.4).*

[3] A. Neumaier. Molecular modeling and mathematical prediction of protein structure. *SIAM Review*, 39:407–460, 1997.

[4] I. Tinoco, K. Sauer, J. C. Wang, and J. D. Puglisi. *Physical Chemistry: Principles and Applications in Biological Sciences*. Prentice Hall, Upper Saddle River, NJ, fourth edition, 2002. *Chapters 4–8, 9 (pp.493–517), 10.*

Related Area: Homotopy Methods

[5] S. Ackermann and W. Kliesch. Computation of stationary points via a homotopy method. *Theoretical Chemistry Accounts*, 99:255–264, 1998.

[6] E. L. Allgower and K. Georg. Continuation and path following. *Acta Numerica*, pages 1–64, 1992.

[7] A. Azmi, R. H. Byrd, E. Eskow, and R. B. Schnabel. Predicting protein tertiary structure using a global optimization algorithm with smoothing. In C. A. Floudas and P. M. Pardalos, editors, *Optimization in Computational Chemistry and Molecular Biology*. Kluwer Academic Publishers, 2000.

[8] L. T. Watson, M. Sosonkina, R. C. Melville, A. P. Morgan, and H. F. Walker. Algorithm 777: Hompack90: A suite of fortran 90 codes for globally convergent homotopy algorithms. *ACM Trans. Math. Softw.*, 23:514–549, 1997.

[9] Z. Wu. The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation. *SIAM Journal on Optimization*, 6(3):748–768, 1996.

Related Area: Optimization Algorithms

[10] J. E. Dennis, Jr. and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. *Chapters 5–9.*

[11] S. G. Nash and A. Sofer. *Linear and Nonlinear Programming*. McGraw–Hill, New York, NY, 1996. *Chapters 10–12.*