ENERGY LANDSCAPING FOR ANALYTICAL OPTIMIZATION: PROTEIN FOLDING PROBLEM

Temsiri Sapsaman

King Mongkut's University of Technology North Bangkok

ABSTRACT

Predicting the shape of an unknown protein, the protein folding problem is one of the most difficult global-optimization problems. Due to the complex energy landscape, locating the protein conformation with the lowest energy is a complicated and time-consuming process. This article presents the background of the protein folding problem, discusses its difficulty, and introduces a method of energy landscaping that improves the efficiency of the analytical optimization in the protein conformation prediction process.

KEYWORDS: Analytical Optimization, Energy Landscaping, Protein Folding

1. Introduction

In drug design, researchers search substances reacting to proteins that are important to diseases. These substances interrupt proteins' function, which can disrupt the growth or even kill diseases. This search is traditionally done by conducting numerous experimental trials. However, when the protein conformation is known, the number of experimental trials can be reduced by pretesting substance-protein interaction using computer simulation. This process could greatly increase the success rate of pharmaceutical trials. Determining the shape of unknown protein molecule is called protein folding problem and it is a global minimization problem.

This article is intended to give the basic knowledge of the protein molecule, the folding process, and the protein conformation prediction, which are presented in the Background section. Moreover, it will introduce the method of energy landscaping applied to the conformation prediction and give some usage suggestion. Energy landscaping approach aims to strategically change the energy function, which is the objective function, to assist the analytical optimization.

2. Backgrounds

2.1 Protein Structure

A protein molecule is a chain of amino acid units known as residues with the sequence of amino acids known as a protein sequence. Twenty types of amino acids are identified by their side chains; the examples of Glycine and Tryptophan are shown in figure 1. These side chains connect to alpha-carbon atoms of the amino acids with zero to four rotation angles, known as the Rotamer angles χ_i. The configurations of the side chains depend on the values of these angles. When two amino acids combine, they release a molecule of water and form a strong bond called peptide plane, as illustrated in figure 2 (in which side chains are omitted). Dihedral angles ϕ and ψ between peptide planes mainly dictate the shape of protein main chain or backbone.

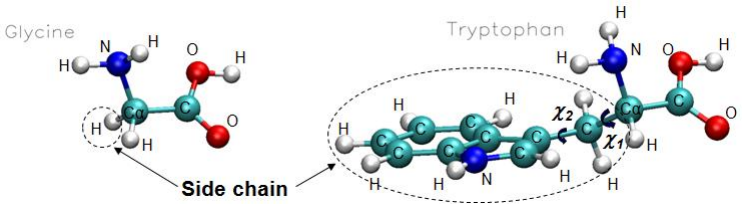

Figure 1 Amino Acids and Their Side Chains

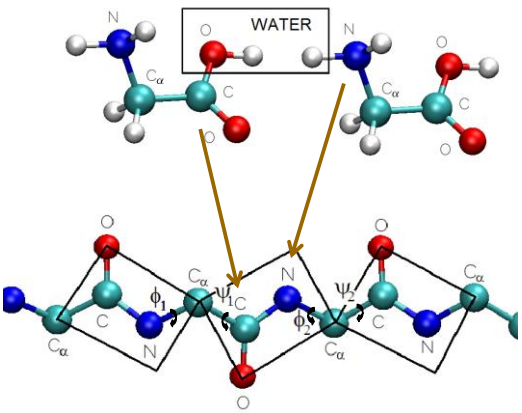

Figure 2 Peptide Plane

To reduce computational difficulty, the protein molecular model is simplified in most numerical simulation. Assumed the absolute flatness and rigidity of the peptide plane, the values of the dihedral and the Rotamer angles give the position of each atom relative to one another using the kinematic of the protein chain. Sometimes the side chains are omitted in the calculation and only the dihedral angles are left in calculation.

2.2 Protein Folding Problem

Protein folding is a process that protein molecule changes from an incompact, unstable, nonfunctional de-natured state to a compact, stable, functional natured state (figure 3a.). Protein folding problem is protein conformation prediction; that is to find the shape of an unknown protein from its sequence (figure 3b.). Each alphabet represents the amino acid residue. Since the protein native conformation and the global minimum energy configuration highly correlate, predicting the protein conformation is a global minimization problem. It is a computationally intensive task due to the high-dimensional and the complex energy landscape. Figure 4 illustrates this complexity as it shows the potential energy of butane, a small amino acid, in different shapes. In atomic level, each pair of atoms has potential energy [1] described as follows:

$$E_{ij} = \sum_{vdW}\left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right) + \sum_{elec}\frac{q_i q_j}{\varepsilon R_{ij}} \tag{1}$$

where E_{ij} is potential energy between atoms i and j, A and B are experimentally known constants, R is the distance between the two atoms, q_i and q_j are charges on each atom, and ε is the dielectric constant. In a simplified form, the total energy function is estimated as the sum of the potential energy between all atoms as in equation (2).

$$E = \sum_{i}\sum_{j, j\neq i} E_{ij} \tag{2}$$

While an average protein molecule has hundreds molecules of amino acids, or thousands atoms, the energy landscape becomes larger and much more complex. Hence, an extremely intensive search is generally required. In optimization, which is discussed more in the next section, the total energy function is used as an objective function. Dependent on the molecular model, the decision variables usually are the dihedral angles ϕ and ψ and the Rotamer angles χ.

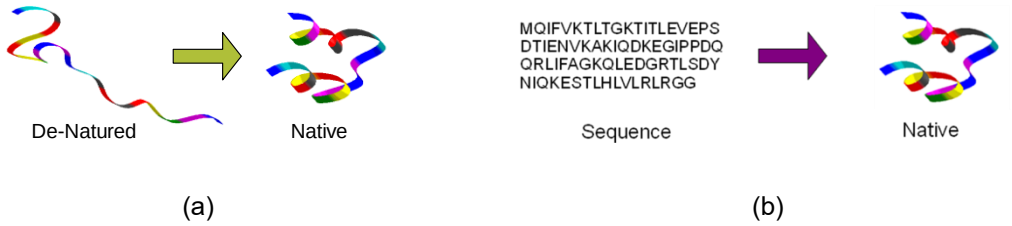

Figure 3 (a) Protein Folding (b) Protein Conformation Prediction

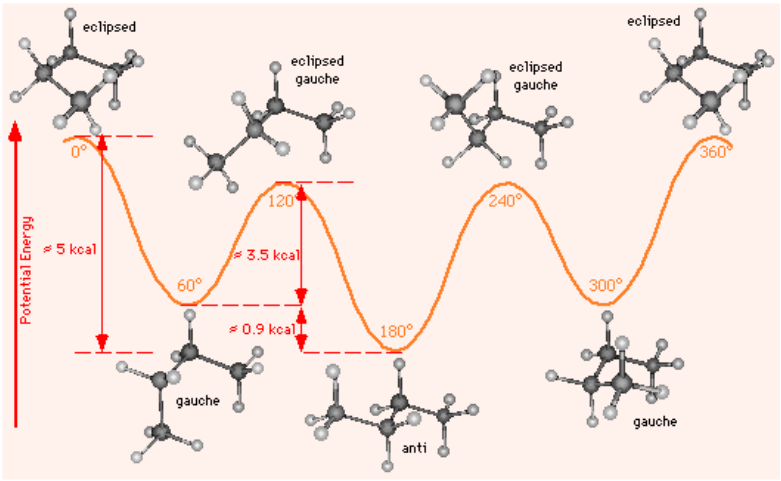

Figure 4 Potential Energy of Butane [2]

3. Methods of Protein Conformation Prediction

The challenge of determining the native conformation of a new protein is approached with several methods of conformation prediction, including molecular dynamics simulation and optimization.

3.1 Molecular Dynamics Simulation

Molecular dynamics (MD) simulation is used in finding the protein conformation by simulating the motion of a protein molecule from denatured state to natured state. The time step of an accurate simulation must be very small, typically on the order of femtoseconds (10^{-15} s). However, since the folding process takes a much longer time, typically on the order of milliseconds or seconds, MD simulations require numerous time steps and become computationally expensive [3]. This technique can locate the final conformation for a very small molecule but impractical for average protein molecules. Addressing this problem, many studies use the protein molecule model with reduced degrees of freedom of [4, 5]. In this model peptide planes are treated as rigid bodies and the *n*-residue main chain has 2*n* degrees of freedom. Although this can reduce computational costs to an order of magnitude, the MD simulation is still not practical for average proteins. Moreover, since peptide planes are not perfectly flat and can be slightly bended by atomic forces, errors unavoidably exist in position calculation and get carried over to energy and force calculation.

3.2 Optimization

At the native state proteins are in a minimum energy configuration; therefore, most research on protein conformation prediction uses optimization algorithms to search the conformational space for the optimal conformation. Since on average a protein has 638 residues [6], this high-dimensional space requires an extremely intensive search. Additionally, the imprecise and highly complex force field model adds more difficulties to the search.

Numerous algorithms for the protein-folding problem are the variations of analytical or probabilistic optimization algorithms, or a combination of both. Probably the most popular probabilistic optimization algorithm used in the protein folding problem, the Monte Carlo (MC) method [7] is easily implemented. By adding random changes to the current configuration, MC simulation generates a large collection of statistical configurations (or decoys). By considering their energies, these configurations are accepted or rejected under the Metropolis criterion [7]. The MC algorithm can be inefficient for a large protein because most generated configurations will have energies higher than the criterion and be rejected.

Therefore, in practice a combination of MC and other optimization techniques is generally used.

3.3 *De Novo* Conformation Prediction

Compared to *ab initio* techniques for solving the protein-folding problem, *de Novo* or knowledge-based methods use information from the PDB and have better performance. Here a *de Novo* method, Rosetta, is discussed. Rosetta [8, 9] performs the protein conformation prediction in two main stages as depicted in figure 5. First stage, called decoy generation, constructs several configurations called “decoys” from pre-generated fragments. Second stage, called refinement protocol, optimizes the predicted conformation. Two steps are in decoy generation: fragment generation and probabilistic search. Using protein database, fragments are generated from naturally available shapes for each protein section. Then initial configurations or decoys are built from fragments using Monte Carlo to find decoys with lowest energy. Knowing the shapes of decoys, the position of each atom can be found and the energy can be calculated from the distance between each pair of atoms. The decoy generation is the process of global search.

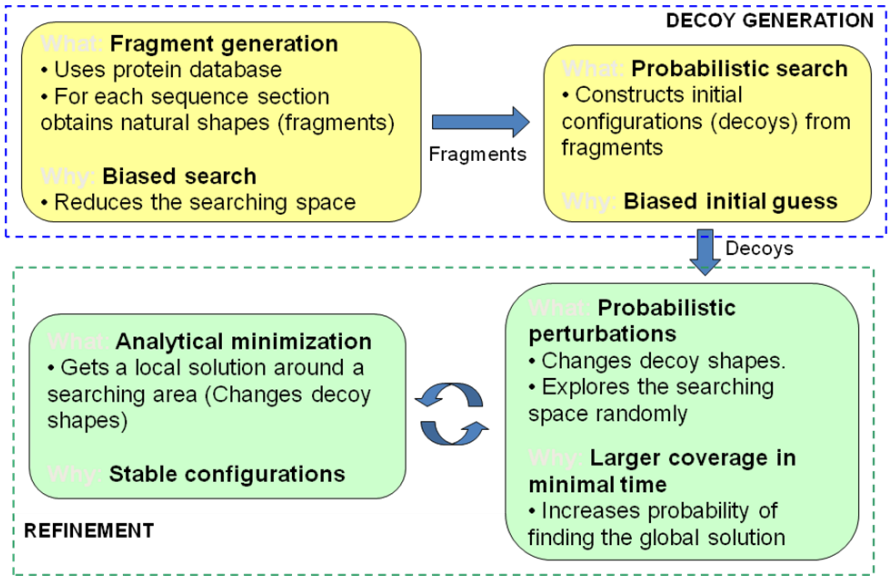

Figure 5 *De Novo* Conformation Prediction Process

Refinement protocol has two recurring steps: probabilistic perturbations and analytical minimization. This is the process of local search. In probabilistic perturbations each decoy is randomly perturbed by slightly changing its configuration. Then the analytical minimization finds a stable configuration near the perturbed decoy and Metropolis criterion is used to decide whether to keep or discard the obtained configuration. The analytical minimization algorithm can be Quasi-Newton algorithm (QNA) or Broyden–Fletcher–Goldfarb–Shanno (BFGS), which are considered good algorithms. The convergence criterion is usually the relative change of the energy but can be set to the number of iterations. As shown in figure 6, this refinement protocol generally has several rounds and takes extremely long computational time. For example, Bradley *et al.* [9] performed high-resolution structure prediction by generating 20,000 to 30,000 decoys and took about 150 CPU days per molecule. Since slowness appears in the analytical minimization, improvement in the efficiency here can significantly reduce the computational time.

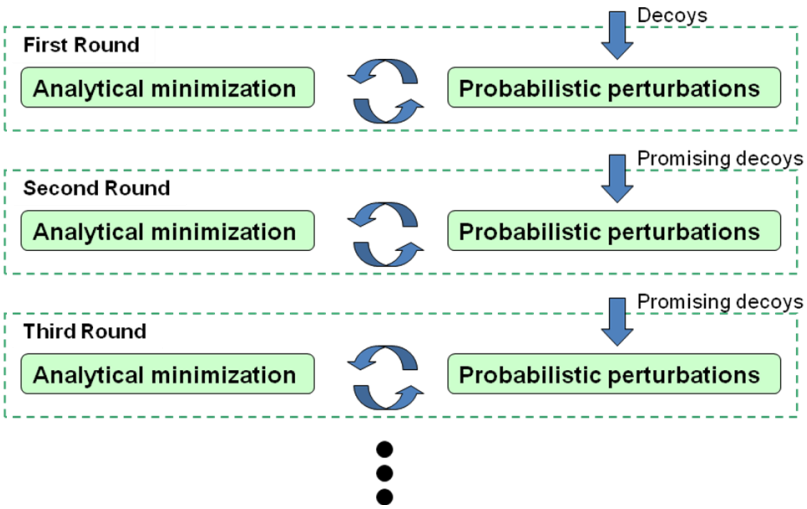

Figure 6 Refinement Process

4. Exponential Energy Landscaping (XEL)

To improve the efficiency of the optimization algorithm, changing the energy landscape is already found in literature. Hypersurface deformation [10, 11] smoothes the energy landscape for analytical minimization. However, it is not applicable to all energy types and it requires remapping to original surface, which is very difficult. Energy flattening

[12] smoothes the energy landscape for probabilistic minimization, which is applicable to all energy types. Although it does not require remapping to original surface, it does not address analytical minimization efficiency. While Exponential Energy Landscaping (XEL) [13] does not require remapping to original surface similar to energy flattening, it addresses the inefficiency of the analytical minimization by improving the refinement process, where is the slowest. Figure 7 shows the *de novo* conformation prediction process with XEL. XEL changes only the magnitude of the energy function but not the locations of minima or maxima, as shown in figure 8. The modified energy function E^* is defined as

$$E^* = sign(E)|E|^n \tag{3}$$

where E^* is described by a nonlinear n^{th}-ordered exponential of the energy function E, n is the exponential order of the energy function, a positive real number. As n is greater than 1, peaks becomes taller and valleys becomes deeper. As n is smaller than 1, peaks becomes shorter and valleys become shallower.

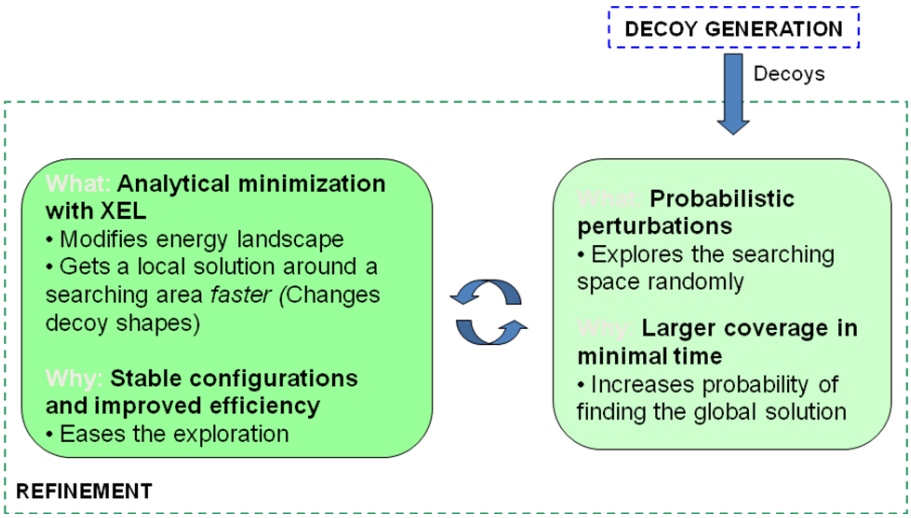

Figure 7 *De Novo* Conformation Prediction Process with XEL

4.1 Implementation and Results

XEL has been invested in simulation with two analytical optimization algorithms [13]: Quasi-Newton algorithm (QNA), and Broyden–Fletcher–Goldfarb–Shanno (BFGS). The simulation is tested on 5 different proteins with 76-304 residues and the values of n between 2^{-9}-2^{-1} for n smaller than 1 and 1-10 for n larger than 1. The tests are done with only 50 decoys and the runtime is between 3:30 and 35:30 hours per simulation on AMD AthlonTM XP 3200+ (2.1 GHz) and 64 Processor 3500+ (2.2 GHz) computers.

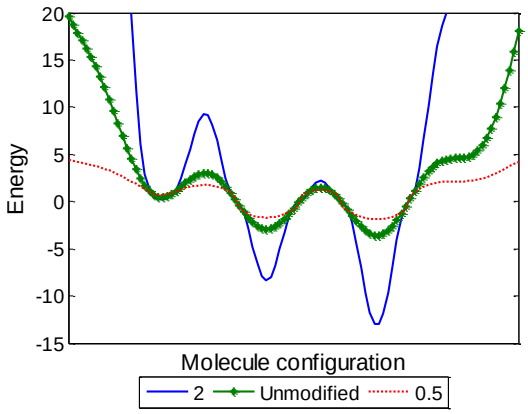

Figure 8 Modified Energy Function

Results show that quality in terms of similarity between configurations obtained from the invested algorithm with and without XEL are the same. However, XEL can achieve better quality in terms of average score improvement, which is the average of energy scores or functions of the obtained configurations. XEL can also achieve better speed in terms of average iteration used in finding optimal solutions. A better quality compared to unmodified case can be achieved with $n > 1$ and a better speed can be achieved with $n < 1$.

Trade-offs between quality and speed must be considered when the XEL is implemented. Results from QNA- and BFGS-XEL algorithms lead to one quantitative recommendations. To improve speed by 15% to 47%, the XEL with n within 2^{-9}-2^{-5} should be used. Quality may be worsened by 4% to 15% on that range of n.

5. Conclusion

Protein conformation prediction is a difficult task and yet very important to the success of drug discovery. Promising results from applying XEL to the *de novo* protein prediction algorithm suggest that XEL could improve efficiency in the analytical optimization. The effect of XEL could be further studied in other optimization problems or algorithms.

Acknowledgement

The author would like express a gratitude to Dr. Harvey Lipkin for his supervision throughout the years at Georgia Institute of Technology and to the Baker Laboratory, University of Washington, for generously providing the Rosetta software.

References

[1] Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. (1995). “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules”. **Journal of the American Chemical Society.** 117: 5179-5197.

[2] Anonymous Author. “Conformations of Alkanes & Cycloalkanes”. Obtained on May 22, 2012. (http://askthenerd.com/ocol/ALKANE/F3.HTM)

[3] Schlick, T. (2002). **Molecular Modeling and Simulation**. Springer.

[4] Rapaport, D. C. (2002) “Molecular Dynamics Simulation of Polymer Helix Formation Using Rigid-Link Methods”. **Physical Review E**. 66 (1) : 011906.

[5] Rapaport, D. C. (2003) “Dynamics of Polymer Chain Collapse into Compact States”. **Physical Review E**. 68 (4) : 041801.

[6] RCSB Protein Data Bank. “Residue count histogram”. Obtained on May 22, 2012. (http://www.pdb.org)

[7] Leach, A. R. (2001). **Molecular Modelling: Principles and Applications.** 2 ed: Prentice Hall.

[8] Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004). “Protein Structure Prediction Using Rosetta”. **Methods in Enzymology**. 383 : 66-93.

[9] Bradley, P., Misura, K. M. S., and Baker, D. (2005). “Toward High-Resolution *de Novo* Structure Prediction for Small Proteins”. **Science**. 309 : 1868-1871.

[10] Azmi, A. M., Byrd, R. H., Eskow, E., Schnabel, R. B., Crivelli, S., Philip, T. M., and Head-Gordon, T. (2000). "Predicting Protein Tertiary Structure Using a Global Optimization Algorithm with Smoothing". **Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches** (Floudas, C. A. and Pardalos, P. M., eds.). Dordrecht/Boston/ London: Kluwer Academic. 1-18

[11] Azmi, A. M., Byrd, R. H., Eskow, E., and Schnabel, R. B. (2006). "A new Smoothing-Based Global Optimization Algorithm for Protein Conformation Problems". **Global Optimization: Scientific and Engineering Case Studies** (Pinter, J. D., ed.). **Nonconvex Optimization and Its Applications** 85 : 73-102 New York: Springer.

[12] Zhang, Y., Kihara, D., and Skolnick, J. (2002). "Local Energy Landscape Flattening: Parallel Hyperbolic Monte Carlo Sampling of Protein Folding". **Proteins: Structure, Function, and Genetics**. 48 : 192-201.

[13] Sapsaman, T. (2009). **An Energy Landscaping Approach to the Protein Folding Problem**. Doctor of Philosophy. The George W. Woodruff School of Mechanical Engineering. Georgia Institute of Technology.

Author's Profile

Temsiri Sapsaman, Ph.D., is a Lecturer at Production Engineering Department, Faculty of Engineering, King Mongkut's University of Technology North Bangkok. The address is 1518 Pibulsongkram Rd., Wongsawang, Bangsue, Bangkok 10800, Tel. (66) 2913-2500 ext. 8208, Fax (66) 2587-0029, Email: tsapsaman@gmail.com. Received her doctorate degree in Mechanical Engineering from Georgia Institute of Technology, GA, USA, in December 2009, Sapsaman is interested in Optimization, Robotics, Mathematical Modeling, and Simulations.