Simple Representation of Protein Structure and Folding Techniques

Jordan Yang
Advisor: Dr. Sorin Istrail

Submitted in partial fulfillment of the requirements for the
Master's of Science

Department of Computer Science
Brown University
April 23, 2019

1 Introduction

The ultimate goal of protein folding is to predict and obtain a high-accuracy structure comparable to that determined via experiment. This would enable users to confidently use rapidly generated *in silico* protein models in all the contexts where currently only experimental structures provide a solid basis, including structure-based drug design, analysis of protein function, and rational design of proteins with increased stability or novel functions. Protein modeling is the only way to obtain structural information if experimental techniques fail. A variety of different proteins are simply too large for NMR analysis or cannot be crystallized for X-ray diffraction[3] Although the approach of using the concept of local structure, in which the conformation of a short segment of polypeptide chain is supposed to depend almost entirely on the sequence of that segment, helped understand local secondary structure, it did not provide sufficient information regarding the process of distant residues interact with each other to form the overall structure. Warshel introduced a simplified representation of a protein structure by averaging over the fine details. The rational of this simplified representation is due to the fact that minor details of a protein structure do not contribute much to the formation of the overall structure. The simplified representation also speeds up the simulation process and drastically reduce the computational time. Besides the simplified protein representation, Warshel also introduced his simulation process, in which he used the combination of convergent energy minimization and normal mode normalisation, essentially accelerates the process by avoiding the many non-productive random fluctuations that occur in nature[8].

2 Simple representation of protein structure

Warshel made two assumptions when introducing his simplified representation to overcome the problems during simulation such as the large number of degrees of freedom for even a small protein, the interaction with the rapidly moving solvent molecules, and the thermal motion of parts of the protein itself[8]. His two assumptions were that much of the protein's fine structure can be eliminated by averaging, and that the overall chain folding can be obtained by considering only the most effective variables[8].

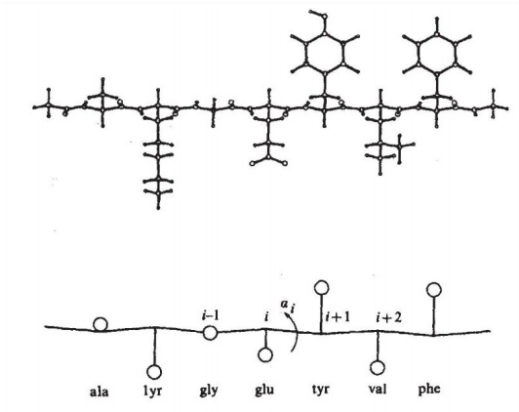

Figure 1: Relationship between the simplified model of protein structure and the all-atom structure of proteins. The two reference points for each residue in the simplified model correspond to the centroid of the side chain and the C^{α}. Each residue is only allowed one degree of freedom: the torsion angle α between the 4 successive C^{α}s of residues(i-1,i+1,i+2). All the side chains of a given type have the same simplified geometry. The bond lengths, bond angles, and torsion angles used to define the geometry of the simplified molecule were taken as the average values found in eight protein conformations, though they could just as well have been taken from amino-acid model compounds.

Based on his assumptions, each residue is represented by only two centers, the C^{α} atom, and the centroid of the side chain. Interactions are assumed to occur only between side chains, while the C^{α} positions define the chain path1. Each amino acid residue only has one degree of freedom, the torsion angle about the line joining two adjacent C^{α}s[8].

3 Energy Potential of The Simplified Representation

Based on the ergodic theorem that averaging over all possible behaviors of an ergodic system over time gives the global minimum behavior of such system[9], Warshel introduced the idea that the effective potential can be obtained by averaging the energy over all those conformations of a dipeptide that have a particular value of the torsion angle α. To reduce the computation burden of calculating the effective potential of all 400 different dipeptides, Warshel did calculations on six most representative dipeptides: ala-ala, ala-gly, ala-pro, gly-gly, gly-ala, and pro-ala. Based on the results obtained from the effective potentials, the effective potential only depended on the nature of the second amino acid, giving different potentials for the α preceding ala, gly, and pro. Because aspartic acid and asparagine were found to occur as frequently in the reverse turns of known protein conformations as glycine, the same potential was used for all three. The alanine-type potential was used for all other remaining amino acids except proline.

The interaction potential between a pair of identical amino-acid side chains were also calculated by spatial averaging. Each side chain was assumed to be spherically symmetrical with a radius equal to the average radius of gyration of that side chain based on equation 1.

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - r_{cm})^2} \tag{1}$$

Warhsel calculated the effective potential at various distances apart as a sum of the interaction energy of all atoms any where in one sphere with all atoms anywhere in the other sphere. The effective potential between identical side chains was approximated by a Lennard-Jones type function based on equation

2, and potentials between pairs of different side chains were obtained by a geometric mean combining law[8].

$$V_{LJ} = \epsilon[(\frac{r_m}{r})^2 - 2(\frac{r_m}{r})^6] \tag{2}$$

Warshel also included solvent interactions in his calculation as well because proteins fold in water not a vacuum, interactions with the solvent are considered by assigning to each side chain a hydrophobic energy taken from the solubilities of amino acids in water and in ethanol. In the calculations, the energy of transfer between these two solvents were taken as the difference in energy of the side chain when isolated in water and when completely surrounded by other residues.

4 Molecular Dynamics

The folding of the idealised protein can be simulated by solving the equations of molecular dynamics at sufficiently small time intervals. Molecular dynamics simulation needs a potential energy function[7]. Equation 3.

$$\begin{aligned}
&E_p(x) = E_{bonded} + E_{non-bonded} \\
&E_{bonded} = E_{bonds} + E_{angles} + E_{dih,imp} \\
&E_{bonds} = \sum_{bonds} k_b(l - l_0)^2 \\
&E_{angles} = \sum_{angles} k_\theta(l - l_\theta)^2 \\
&E_{dih,imp} = \sum_{dihedrals} k_\phi(1 - cos(n_\phi)) + k_\theta(\theta - \theta_0) + \sum_{improper} k_w(w - w_0) \\
&E_{non-bonded} = E_{vdw} + E_{electrostatic} \\
&E_{vdw} = \sum_{i,k}(\frac{A_{i,k}}{r_{i,k}^2} + \frac{G_{i,k}}{r_{i,k}^6})
\end{aligned} \tag{3}$$

In a viscous medium like water, these equations of motion can be approximated by Langevin equations, which is a stochastic differential equation describing the time evolution of a subset of the degrees of freedom[2], where the change in variables is directed down the energy gradient with a random deflection due to Brownian motion[8, 6]. For computational efficiency, Warshel

neglected these thermal fluctuations while the chain folds[8], and the end point of the trajectory is the potential energy minimum accessible from the starting conformation. He minimized the energy of the idealized protein chain with respect to all the α angles using a quadraically convergent method[8]. After a minimum was reached, he reintroduced thermal fluctuations and the conformation was considered to be vibrating about the minimum so that each normal mode had average kinetic energy $\frac{kT}{2}$. He chose a new staring conformation for the next pass of energy minimization by suddenly stopping the thermal vibration[8]. The idea of using noraml-mode thermalization is to avoid non-productive changes in conformation for it knows which combinations of angle changes should cause the greatest change in conformation for a given energy increase[8].

5 Testing The Simplified Representation

Warshel tested his simplified representation by minimizing from near the native folded conformation[8]. He picked Bovin pancreatic trypsin inhibitor because it is the only small protein with less then 100 resiudes of known conformation at that time that has a single polypeptide chain and non additional prosthetic group[8]. As a first step, Warshel obtained a simplified native PTI conformation by taking the C^{α} positions and side-chain centroids from the X-ray coordinates[8]. Next and idealized chain, based on the PTI sequence and having the same geometry for all side chains of the same type, was made to fit the simplified n ative coordinates by adjusting the α torsion angles. This conformation, known as the idealized native strucutre, deviates by 1.1 A r.m.s. from the simplified native structure. The RMSD is based on equation 4.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{4}$$

where δ_i is the difference, in the two structures, of the distance between side-chain centroids[8]. Warshel carried out energy minimization from this starting conformation to reproduce the stability of native PTI. A perfect minimum was reached after 558 cycles at an energy of -52.0 kcal/mol and a RMSD of only 3.7 A from the simplified native conformation.

6 Simulation of PTI folding

Warshel tried to simulate the actual process of folding with most tests done with two open starting conformations, one fully extended (all α=180), and one extended apart from the C-terminal helix (α=180, except for resiudes 48 to 58 where $\alpha = 45$)[8]. The reason that Warshel retained the terminal helix from the native structure was because they were more concerned with the process of folding than with prediction of the native conformation of an unknown protein. Figure 2 indicats the iteration history of minimization from the second of these starting points, which was the most successful run[8].Thermal randomization about the first minimum, an irregular but extended conformation, raised the energy and in this case caused the chain to bend back on to itself decreasing the RMSD. From this point, minimization first opened the molecule again, but then reached a new minimum where the chain now had kinks that could become the bendsof β hairpins[8]. After a second randomization, minimization rapidly folded the molecule bringing the terminal helix close to the β hairpin centred on residue 27. The final folded conformation of Figure 2 remarkably resembled the native molecule[8]. Conformations that deviated more from the native structure always had higher energies, which gave an independent criterion for choosing the best conformation and indicated that more passes of thermalization and minimization could lead to a conformation closer to the native one[8].

7 Ramifications of The Simplified Model

The simulation results of Warshel's simplified model demonstrated the accuracy and precision of the model to account for the stability and folding of a molecule[8]. Based on his observation, Warshel made the claim that folding must depend on a very rare random fluctuation that happened to bring the right residues close together with sufficient precision for the short-range forces to take effect and it is unlikely that these short-range forces could direct the folding from an open disordered structure. The time average of short-range forces, which are weak, fairly long range, and not too dependent on orientation, restrict the number of low energy conformations severely; they cause the chain to fold into the approximate shape rapidly without having to pass through many local minima[8]. Based on his simplified model. Warshel proposed that starting from the initial configuration, when the chain

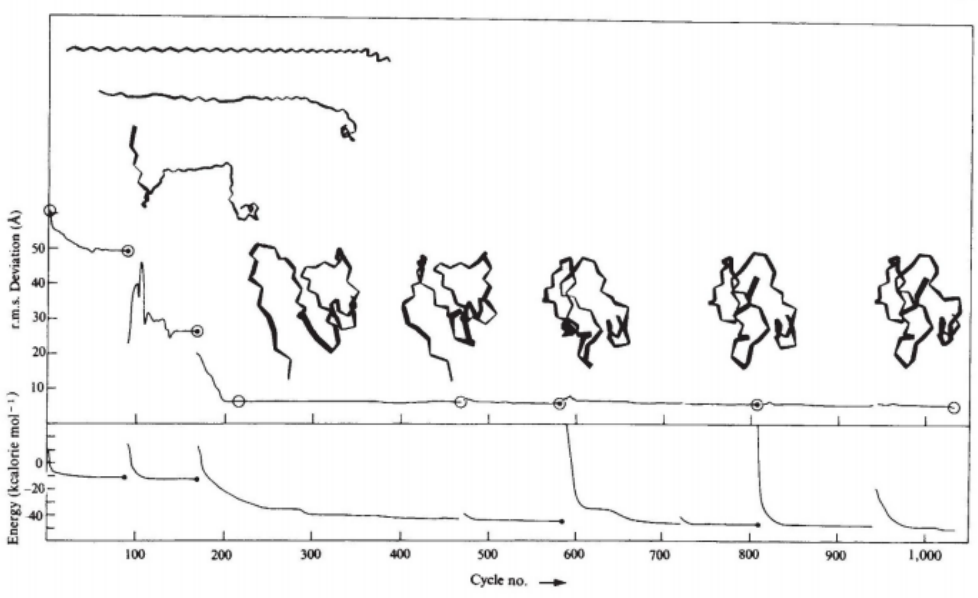

Figure 2: Simulation of PTI folding from an extended starting conformation with the terminal helix ($\alpha = 180$ for all except 48 to 58 where $\alpha = 45$). No knowledge whatsoever about native PTI was used during this simulation (apart from setting the terminal helix). The conformation was thermalized at the end of each minimization except near cycles 490 and 730 when the energy rises slightly because the minimisation was restarted after rounding the torsion angles to one degree. In the first two thermalizations, each normal mode was perturbed in the plus direction to raise the associated energy to $\frac{R(n)kT}{2}$ with $T = 1,000K$. In the other three thermalizations, the perturbations were randomly in the plus and minus directions but always such as to raise the energy by $\frac{kT}{2}$ with $T = 300K$. (Because the random numbers were distributed uniformly rather than exponentially, these temperatures did not correspond to the macroscopic temperature.) The 8 ribbon diagrams, which show the C^{α} chain path, refer from left to right to the 8 conformations at the circled points on the r.m.s. deviation curve, respectively. The last five conformations had progressively lower energies and were each a little closer to the native structure ($E = -43.3$, -45.7, -46.0, -46.9 and -48.9 kcal/mol, respectively; r.m.s. deviation = 6.08, 5.7, 5.6, 5.4 and 5.3 A, respectively). The solid dots at the end of a minimization indicated that a perfect minimum was reached (r.m.s. gradient less than 10^{-6} kcal/mol). One cycle took about 0.6 s an IBM 370/165 computer.

has a flexible open structure, the effective time-averaged forces between the residues play a central role, folding the chain into a compact shape with most groups close to their final positions[8]. Once the chain becomes compact with less freedom of movement, the importance of short-range interactions emerge and they form a precise conformation provided that the resulting gain in enthalpy overcomes the loss in entropy[8]. This idea contributed to the future development of protein folding method using the combination of homology modeling and *ab initio* method[3].

8 Metropolis Algorithm

Metropolis first introduced a method with modified Monte Carlo integration, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules back in 1953[10]. And his method contributed to the development of what is known as the monte carlo protein folding method.

The Metropolis Algorithm is designed to efficiently sample multi-dimensional functions. It is based on the concept of a random walk. In a two-dimensional classical random walk, a point is started at the origin and it is allowed to move one unit in any direction with equal probability. However, a walker is allowed to spend more of its time in those regions where the function is sampled with large probability.

- step 1: start from some initial parameter value θ^c
- step 2: evaluate the unnormalized posterior $P(\theta^c)$
- step 3: propose a new parameter value θ^*, which is randomly drawn from a "jump" distribution centered on the current parameter value
- step 4: evaluate the new unnormalized posterior $P(\theta^*)$
- step 5: decide whether or not to accept this new value using the *Metropolis acceptance criterion*. Accept new value with probability $a = \frac{P(\theta^*)}{P(\theta^c)}$
- step 6: repeat step 3-5

A more realistic example:

- Let one of the molecules move slightly.
- Calculate the change in energy of the system, ΔE (based on the distance between molecules i.e. potential energy).
- If this energy is less than the previous state, accept the change.
- Otherwise accept the change with a probability of $exp(-\delta E/kT)$, where k is the Boltzman constant and T is the temperature of the system.
- repeat until converges.

8.1 Metropolis Criterion

- Let X be the current solution and X' be the new solution, $C(X)$ be the energy state (cost) of X, and $C(X')$ be the energy state of X'.
- Probability $P_{accept} = exp[(C(X) - C(X'))/T]$
- $N = Random(0, 1)$
- Unconditional accepted if $C(X') < C(X)$, the new solution is better
- Probably accepted if $C(X') >= C(X)$, the new solution is worse, accepted only when $N < P_{accept}$.

8.2 Incorporating the Metropolis Criterion

We know the positions of the N particles in the square and we can calculate the potential energy of the system.

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} V(d_{ij}) \tag{5}$$

V is the potential between molecules, and d_{ij} is the minimum distance between particles i and j. The Monte Carlo method really arises from circumventing the difficulty of carrying out a several hundred-dimensional integral by the usual numerical methods. Because to calculate the equilibrium value of any quantity of interest F,

$$\bar{F} = \frac{\int F exp(-E/kT) d^{2N}p d^{2N}q}{\int exp(-E/kT) d^{2N}p d^{2N}q} \tag{6}$$

we need to perform the integration over the $2N$-dimensional configuration space and it is evidently impractical to carry out a several hundred-dimensional integral by the usual numerical methods. Equation 1 actually came from countering the impracticality of the naive method for close-packed configurations. The most naive method of carrying out the integration would be to put each of the N particles at a random position in the square, then calculate the energy of the system, and give this configuration a weight $exp(-E/kT)$. This method is not ideal for close-packed configurations, since with high probability we choose a configuration where $exp(-E/kT)$ is very small; hence a configuration of very low weight. So Equation 1 is actually a modified Monte Carlo scheme, where, instead of choosing configurations randomly, then weighting them with $exp(-E/kT)$, we choose configurations with a probability $exp(-E/kT)$ and weight them evenly. We place the N particles in any configuration, move each particles within a reasonable range, and then we calculate the change in energy of the system ΔE, which is caused by the move. If $\Delta E < 0$, we allow the move and put the particle in its new position. If $\Delta E > 0$, we allow the move with probability $exp(-\Delta E/kT)$. This is essentially where did Equation 1 come from and how it relates to the Metropolis Criterion described in class.

9 Future Development

Just as Warshel proposed, when the starting configuration has a flexible open structure, the effective time-averaged forces between the residues play a central role, folding the chain into a compact shape with most groups close to their final positions. Future work can be done in terms of perfecting the sampling technique. Warshel used molecular dynamics simulation in his work. Another famous *ab initio* sampling technique is Monte Carlo method. However, at low temperatures simulations based on canonical Monte Carlo or molecular dynamics techniques will get trapped in one of the multitude of local minima separated by high energy barriers[4]. Multicanonical algorithms and simulated tempring are examples of such an approach to overcome this difficulty[1, 5]. Hansmann introduced another algorithm, parallel tempering, which can improve updates by means of constructing a special generalized ensemble[4]. He showed that his method can be successfully applied to the simulation of molecules with complex energy landscape[4]. A lot of work is also needed to be done on refining the potential energy function. As the

configuration becomes more compact with less freedom of movement, the specific short-range interactomic forces become important; therefore, a more detailed model incorporating more atomic interactions is much needed.

References

[1] Bernd A Berg et al. The multicanonical ensemble: a new approach to computer simulations. *Int. J. Mod. Phys. C*, 3(FSU-HEP-92-08-17):1083–1098, 1992.

[2] Daniel T Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[3] Jenny Gu and Philip E Bourne. *Structural bioinformatics*, volume 44. John Wiley & Sons, 2009.

[4] Ulrich HE Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1-3):140–150, 1997.

[5] Ulrich HE Hansmann and Yuko Okamoto. Monte carlo simulations in generalized ensemble: Multicanonical algorithm versus simulated tempering. *Physical Review E*, 54(5):5863, 1996.

[6] Ioannis Karatzas and Steven E Shreve. Brownian motion. In *Brownian Motion and Stochastic Calculus*, pages 47–127. Springer, 1998.

[7] Georg Kresse and Jürgen Hafner. Ab initio molecular dynamics for liquid metals. *Physical Review B*, 47(1):558, 1993.

[8] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694, 1975.

[9] Donald Allan McQuarrie, Donald Allan McQuarrie, Donald Allan McQuarrie, and Donald Allan McQuarrie. *Statistical thermodynamics*. Harper & Row New York, 1973.

[10] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.