Computer Simulation of Protein Folding Using the 3D Hydrophobic-Polar Lattice Model

Richard Wang

Dr. Siqian He

National Center for Biotechnology Information/NLM/NIH

Abstract

Protein folding is a natural phenomenon that guides the structure and function of proteins involved in nearly all cellular processes. The Hydrophobic-Polar(HP) model is used to help researchers study the mechanics of protein folding in a simplified and computationally inexpensive way. I developed a protein folding simulation in Python using the HP model on a 3D cubic lattice, including features such as the Monte Carlo algorithm, simulated annealing, and anti-annealing. A contact-based energy function and a distance-based energy function conducted folds on identical HP sequences and were compared with the native structure and each other to reveal information about the shape of the energy landscape and foldability of certain sequences. The contact-based energy function outperformed the distance-based energy function by producing a much more stable range of folds. However, the lowest observed energy was still out of reach of the known native state for complex sequences. Anti-annealing experiments revealed a jagged funnel-shaped energy landscape, with only 7% of folds returning to native state post-experiment, and with folds landing on certain energies more frequently than others. These results suggest clues about the complex shape of the energy landscape and the possibility of new developments to this critical problem.

Introduction

Proteins are complex polymer structures composed of a chain of amino acids folded together into a unique, stable shape. They participate in nearly all cellular functions within organisms, such as enzymes, cell signaling, and immune response. The native structure of proteins are critical to their overall function. When proteins do not fold correctly, they may not function effectively or may result in new unfavorable functions that can lead to cancers and diseases such as Alzheimer's, Parkinson's and Type 2 diabetes [1,2,3]. Due to our limited knowledge for most proteins and their structures, research in areas such as structure-based drug design is majorly hindered as protein structure information is key in developing those drugs [4].

Protein folding is a grand puzzle that is lightly scratched by researchers even after decades of research. There are multiple facets of the problem that researchers are currently trying to decode [5]. The first question is identifying the physical balance that dictates the precise, folded structure of a protein given an amino acid chain. The second question deals with finding out how proteins fold so quick. Lastly, the third question deals with predicting the final protein structure given only the amino acid sequence.

There are several factors theorized to stabilize the structure of a folded protein. These factors include hydrogen bonds, hydrophobic interactions, Van Der Waals electrostatic forces, and temperature to name a few. While all of these factors are responsible for ensuring a stable structure, hydrophobic interactions are by far the most significant contributors. The entropic penalty associated with hydrophobic residues and surrounding water molecules is much more significant than those of other interactions, causing hydrophobic cores to form as the hydrogen bonding order is limited, resulting in a lower net energy and a more stable fold.

Simulating actual protein folding is extremely difficult and computationally expensive. The protein folding problem is NP-complete [6], meaning that while a solution can be tested in polynomial time, the problem cannot be solved in polynomial time. This is further described by Levinthal's paradox introduced by Cyrus Levinthal in 1969. Given a 150 amino acid sequence, Levinthal estimated there to be 10^{150} possible folding conformations. For a protein to search through all of those conformations, it would take 10^{130} years, yet proteins are able to correctly fold in a matter of milliseconds to minutes. Thus, there must be a clear biological pathway for real protein folding.

One way researchers simulate protein folding in polynomial time is by using low-resolution coarse-grained models of the complex protein to quickly produce a near-optimal approximation of the protein fold. A widely used coarse-grain model is the Hydrophobic-polar(HP) model introduced by Ken Dill in 1989 [7], of which each amino acid is represented as a single bead that is either hydrophobic or polar. When folding, the hydrophobic beads congregate as close as possible with other hydrophobic beads towards the center of the structure. Simplified representation methods can also include lattices, which are often used in conjunction with the HP model. The lattices dictate precise locations where the beads must be placed when folding, similar to points on a grid. The most commonly used lattices in modeling are square, triangular, and cubic lattices (Figure 1.)

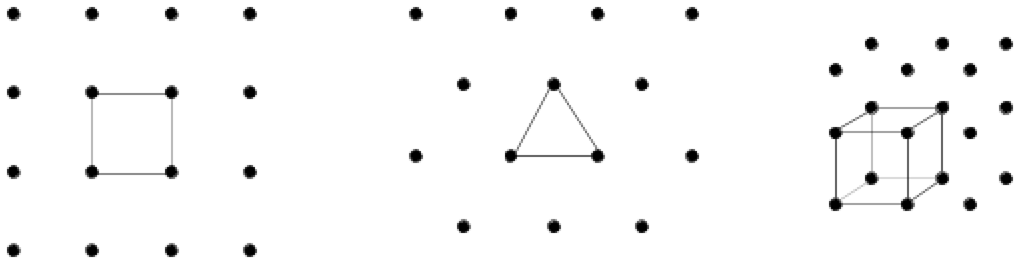

Figure 1: Commonly used lattice models. From left to right: Square, Triangular, Cubic.

Other ways of increasing efficiency of a simulation include the usage of simulated annealing. Simulated annealing occurs when the simulation gradually decreases in temperature while running. As temperature dictates the level of activity of the simulation, a high temperature will make the simulation accept riskier folds and increase the chance of a fold jumping out of a local energy minimum. Likewise, a low temperature will cause the simulation to reject many moves that do not lead to an immediate energy decrease. However, this may cause the simulation to be stuck in a local energy minimum, unable to reach the global minimum at all. The steady decrease in temperature encourages the protein to explore conformations in the beginning of folding while maintaining a near-optimal or stable shape if achieved near the end of the simulation.

Methods

First, an HP cubic lattice protein folding simulation was coded using Python 2.7.2. It takes a primary structure HP chain sequence and conducts a three-dimensional lattice folding simulation while enforcing self-avoiding walk and using the Monte Carlo method. It also incorporates features such as simulated annealing, a contact-based energy function, and a distance-based energy function.

At each simulation folding step, the protein structure folds across a single residue in a random direction. The program then checks that no single location on the lattice is occupied by two unique residues, maintain the self-avoiding walk requirement. If the fold is not possible, then the program selects a new location on the residue chain for the fold and repeats the folding process on the new location.

If a fold is possible, the simulation will then compare the energy of the previous state and the new state. This simulation uses a Markov chain, meaning that the future state will only depend on the current state, not on any of the previous states. With the Monte Carlo method, a random number between 0 and 1 is generated and compared directly with a value calculated using the Boltzmann factor[8] in Eq. (1.1), where *ΔE* is the change in energy from the original conformation to the proposed conformation, *k* is the Boltzmann constant at 1.987 * 10^{-3} kcal/(mol*K), and *T* is the current temperature of the simulation.

$$p = e^{-\frac{\Delta E}{kT}} \tag{1.1}$$

If the generated random number is less than the calculated Boltzmann factor value, the program will accept the fold and begin looking for a location for the next fold of the protein. This is done to replicate the effects of water molecule bombardment that occurs when a protein folds. The energy from the constant motion of surrounding solvent molecules displaces parts of the protein, causing the conformation shape to change [9]. In the simulation, this allows the protein to jump out of dips in the energy landscape when folding, increasing the chances of finding the global energy minimum.

Two different energy functions were tested using this program: an H-H contact based function, and an H-H distance based function. In the H-H contact-based function, for every direct side-by-side contact of H beads, the energy of the given conformation is subtracted by one. Trials were conducted using both energy functions separately. The end conformations from both energy functions were compared with each other and to the known native state of the protein sequence. In the H-H distance based function, the energy of a conformation was the negative sum of the inverse distance between every H bead, shown in Eq. (1.2).

$$E = -\sum \frac{1}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}} \tag{1.2}$$

As the two energy functions provide results on widely different ranges for identical proteins, to compare the energy functions, all calculations were made using the respective energy function, but the distance-based trials also outputted a comparison value using the contact-based energy equation. This conversion did not impact the folding process and the folding path that created the final structure, as folding algorithm calculations were restricted to the original energy function type, and only the numerical energy value used for comparison was calculated separately using the contact energy equation. The contact-based energy was chosen for the standardized comparison energy type over distance-based because contact-based energy more accurately mimics the tendency for hydrophobic amino acids to cluster together when folding, and multiple other factors come into play when considering distance, such as the dielectric effect of water for electrostatics [10]. The units for the contact-based energy values are in terms of contacts, while there are no units for distance-based energy values as it is calculated entirely on unitless lattice coordinate values.

2D and 3D folding was conducted using HP sequences with known native states and energy values previously used by researchers in their work [11,12]. However, emphasis in this paper is placed on 3D folding as this is most similar to the folding mechanics of real proteins as they can fold in 3D, although not constricted to a lattice.

Simulated annealing was incorporated to minimize the energy of the folds further. In Simulated annealing, the temperature of the simulation is brought from a high starting temperature that allows the protein structure to overcome large energy barriers when folding, to a near zero final temperature to prevent large conformational changes and to encourage low-risk movement. Temperature is decreased every step in increments inversely proportional to the length of the simulation and the temperature range.

After each folding step, the program outputs coordinate data to a .xyz file that is viewable through an external program such as Visual Molecular Dynamics. The final .xyz file allows for 3D visualization of the path the protein took to fold to the final conformation. Each trial in this simulation was conducted with 100,000 calculation iterations.

Results and Discussion

Over a series of tests utilizing a range of configurations (2D, 3D, multiple lengths, simulated annealing), the results of the folding suggested an energy barrier when comparing the energy of the folds produced and the energy of the known native state of more complex proteins. The energies of the best folds produced using the simulation on more complex proteins (length > 30) were not as low as the energy of the native state. However, similar energy distributions were exhibited in many cases, such as on a protein sequence of length 48 (Figure 2). The center of the energy distribution was shifted to higher energies when using the distance-based energy function, compared to the more stable contact-based energy function.

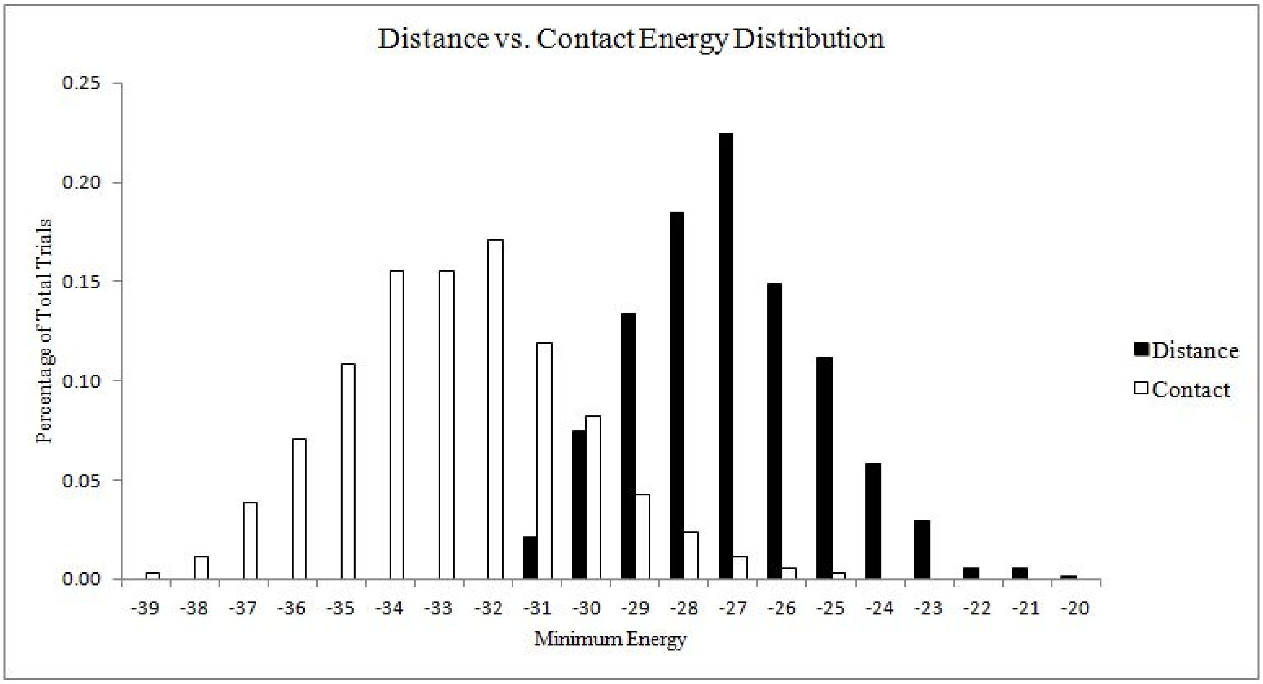

Figure 2: Comparison of frequencies of minimum energies found from 708 trials for each energy function. Sequence: PHHPPPPPPHHPPPHHHPHPPHPHHPPHPPHPPHHPPHHHHHHHPPHH Native energy of sequence is -46.

While there was a clear difference in the folding capabilities of the two energy functions, the native energy was not obtained even after extended data collection of 60,000 contact-based trials. The most stable conformation observed were a few folds with energies of -41 (Figure 3), compared to the native state with an energy of -46 (Figure 4).

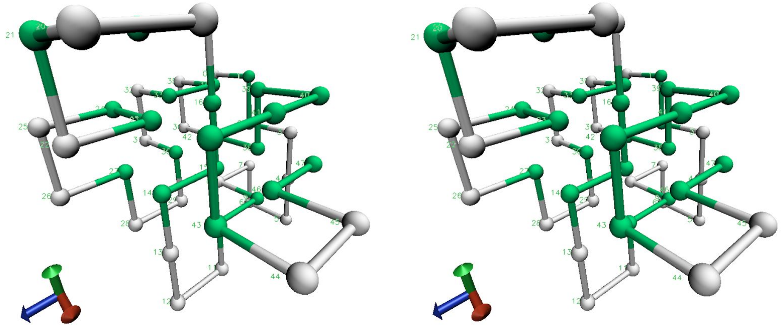

Figure 3: Stereo diagram of simulated fold with lowest contact energy of -41. Green represents hydrophobic beads, white represents polar beads.

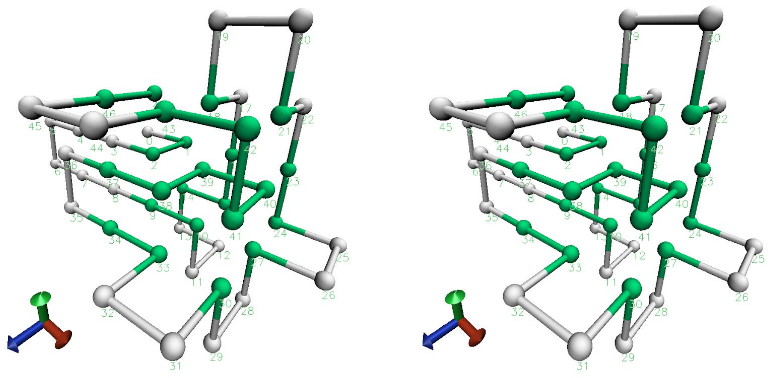

Figure 4: Stereo diagram of known native state with contact energy of -46. Green represents hydrophobic beads, white represents polar beads.

To investigate the folding "barrier" further, anti-annealing was performed on the native state of the protein. This would determine if the folding barrier was a result from the energy landscape of the designed protein, or was a result of the folding algorithm. The temperature of the native state was slowly raised from zero until the protein jumped out of the native state on its own, where the protein was allowed to move for 100 folding steps. The temperature was then slowly lowered back to zero to allow continued folding towards an optimal stable structure. If the protein structure returned to the native conformation after cooling down, the observed barrier would be caused by a folding capability barrier inherent in the folding algorithm used. However, if the protein often deviated from the native state after cooling down, the cause of the observed energy barrier would be due to a very small frequency of low-energy conformations near the native state and to the overall shape of the energy landscape for the protein folded. Data collected from the anti-annealing experiment showed that the folds often parted away from the native state, sometimes even completely unraveling (Figure 5).

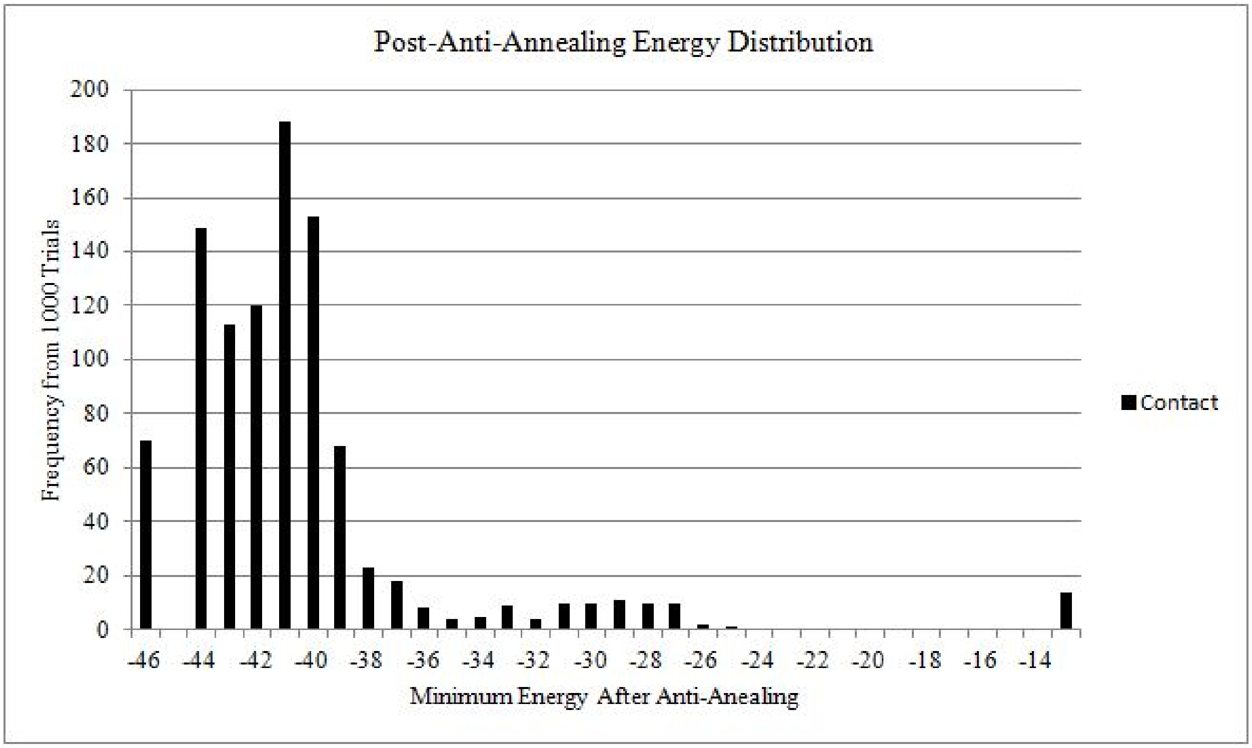

Figure 5: Histogram of minimum energy after the anti-annealing experiment using the contact energy function.

On the same protein sequence used in the contact versus distance experiment, only 7% of the anti-annealing trials returned to the native state energy of -46 after cooling down, with the average energy at -40. There were 14 instances in the 1000 trials where the protein completely unraveled to the maximum energy of -13, with no additional contacts possible within a single fold. This demonstrates the inherent flaw with the contact-based energy function, as the program has no real incentive to fold if no immediate energy change is apparent. This further led me to explore the distance-based energy function. With long-range H-H attraction, the distance-based energy function almost always guarantees a change in energy when the energy of a proposed fold is evaluated. In real protein folding, while the final shape is dependent on the sequence, the hydrophobic effect and the surrounding aqueous solution essentially "squeezes" the protein into a globular shape [13], hence there is a long-range hydrophobic attraction involved that is not present when using the contact-based energy function.

Another observation from the anti-annealing data is that there are no conformations found with an energy of -45. This is due to the shape of the energy landscape not allowing many conformations with that specific energy. With the limited number of low energy conformations possible in the energy landscape, none of them found through the anti-annealing experiment had an energy of -45 for this specific protein sequence. In addition, folds more frequently landed on certain energies than others. Best conformations on 33.7% of the trials were for energies -44 or -41. However, best conformations for only 23.3% of the trials were for energies of -43 or -42. This reveals the increased abundance of conformations with energies of -44 or -41 instead of -43 or -42. The energy landscape surrounding -44 and -41 may have a wider funnel leading to them compared to the energy landscape surrounding -43 and -42, causing more protein fold trials to fall into the -44/-41 basin.

The anti-annealing data suggests that the native energy location in the energy landscape can be likened to a golf hole on a large, uneven golf course. The golf hole represents the native state of the protein, unique from the surrounding area in terms of height. It is difficult for a golf ball to roll into the hole, especially if the location of the hole is not known. If the golf ball "jumps out" of the golf hole, simulated by the anti-annealing, it is not likely that the golf ball will return into the golf hole. Often, the golf ball will end up farther away from the golf hole.

Conclusions and Future Work

A hydrophobic-polar protein folding simulation was developed in Python, through which multiple experiments were conducted to analyze the energy landscape of HP protein sequences to apply to the scientific understanding of real life protein energy landscapes. The program was successful in finding the native state of small (<30 length) protein sequences. However, folding "barriers" were apparent when folding more complex sequences of greater length. Through testing of multiple energy functions, the contact-based energy function was revealed to provide more stable folds compared to the distance-based energy function. Anti-annealing experiments were also conducted to study the energy "barrier" noticed when folding large proteins. The data collected from the experiments allow us to peer into the vast energy landscape that proteins have.

The relationship between the size of the folding barrier noticed and the foldability of the protein sequence tested should be investigated further to identify any possible correlations. An easily foldable protein should have a large and smooth "funnel"-like energy landscape, making it easy for the fold to complete and "funnel" its way to the native state near the end of the simulation [14,15]. The results from the anti-annealing experiment supported the idea of a funnel referenced in previous research. However, while having a funnel, the protein noted above in Figure 5 did not have a very defined funnel, with folds often parting away from the native state or getting stuck in energy "ridges" at energies -44, -41, and higher.

Future work can be done by modifying and improving the folding algorithm to simulate at a higher complexity. The lattice model algorithm may be generalized to an off-lattice algorithm, capable of folding akin to actual proteins. In addition, new features may be added to increase the efficiency and accuracy of the protein folding algorithm, such as replica exchange swapping, where sets of trials using identical sequences are performed at a range of constant temperatures, and semi-stable conformations from one temperature can be copied over to another temperature due to the algorithm's Markov chain properties. Improvements can also be made by utilizing advanced algorithms to calculate pivot locations for each fold, and implementing parallel threading to execute multiple folds simultaneously instead of one at a time. These improvements will allow the simulation to better mimic real world protein folding and create a more accurate representation.

This research helps scientists to understand the mysterious energy landscape that proteins have. Research in protein folding has so far been overwhelmingly an area that researchers are understanding one step at a time, similar to how a group of blind men feels an elephant to learn what it is like. There are still countless questions to be asked about protein folding, and innumerable answers to those questions. However, by answering one question at a time, researchers may fully understand the complex protein folding problem. With this research, scientists are getting closer to using computer models to accurately predict the structure of a protein, opening pathways in biomedical research such as the attractive yet currently unrealizable goal of structure-based drug design.

Addendum

The entirety of the code can be viewed here: https://github.com/CompetitionEntrant/Computer-Simulation-of-Protein-Folding-using-the-3D-HP-Lattice-Model.

References

[1] Ashraf GM, Greig NH, Khan TA, et al. Protein misfolding and aggregation in Alzheimer's disease and type 2 diabetes mellitus. CNS Neurol Disord Drug Targets. 2014;13(7):1280-93.

[2] Gregersen N, Bross P, Vang S, Christensen JH. Protein misfolding and human disease. Annu Rev Genomics Hum Genet. 2006;7:103-24.

[3] Chiti F, Dobson CM. Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem. 2006;75:333-66.

[4] Anderson AC. The process of structure-based drug design. Chem Biol. 2003;10(9):787-97.

[5] Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. Annu Rev Biophys. 2008;37:289-316.

[6] Guyeux C, Côté NM, Bahi JM, Bienia W. Is protein folding problem really a NP-complete one? First investigations. J Bioinform Comput Biol. 2014;12(1):1350017.

[7] Dill KA. Theory for the folding and stability of globular proteins. Biochemistry. 1985;24(6):1501-9.

[8] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57(1):97.

[9] Ben-Naim A. Molecular theory of water and aqueous solutions: Part 1: Understanding water. Singapore, Singapore: World Scientific Publishing Company; April 2009:190.

[10] Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th ed. New York: W H Freeman; 2002. http://www.ncbi.nlm.nih.gov/books/NBK22567/.

[11] Shmygelska A, Hoos HH. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. BMC Bioinformatics. 2005;6:30.

[12] Dill KA, Bromberg S, Yue K, et al. Principles of protein folding--a perspective from simple exact models. Protein Sci. 1995;4(4):561-602.

[13] Huang K. Lectures on statistical physics and protein folding. World Scientific; 2005. http://sciold.ui.ac.ir/~sjalali/book/Huang_Lectures_On_Statistical_Physics_And_Protein_Folding.pdf

[14] Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol. 2004;14(1):70-5.

[15] Oliveira AB, Fatore FM, Paulovich FV, Oliveira ON, Leite VB. Visualization of protein folding funnels in lattice models. PLoS ONE. 2014;9(7):e100861.