Genetic Annealing to Determine Protein Structures

Hrishikesh Belagali

This manuscript was compiled on October 20, 2024

Abstract

This project proposed a hybrid genetic annealing algorithm for optimizing a 3 dimensional AB off-lattice model. A population of solutions was created. The fitness of each solution was evaluated and parents were chosen through tournament search. The new population was subjected to mutation in the form of simulated annealing. This process was repeated for the required number of iterations. The algorithm was tested on artificial Fibonacci sequences and a benchmark set of 5 small proteins from the Protein Data Bank and yielded models with RMSD ranging from $1.875 - 3.468$. The predicted models were represented using the C_α space-filling model. The results of this project also raised questions regarding the algorithm's use in prion energetics.

Keywords: *protein folding conformation, genetic annealing, off-lattice model, simulated annealing, C_α space-filling model*

1. Introduction

The study of biological mechanisms requires the 3D structure of proteins (Dill & MacCallum, 2012). These conformations can be obtained through physical methods such as X-ray crystallography and nuclear magnetic resonance spectroscopy (Bagaria et al., 2013). However, these methods are time-consuming and resource-intensive to execute (Güntert, 2004). Computational models have become necessary due to these constraints (Javidpour, 2012). Many proteins exist whose conformations are not similar to any known protein structures. Free modeling is based on the general mechanisms behind protein folding (Simons et al., 1999). This technique involves the primary amino acid sequence and solvent as input parameters (Anfinsen et al., 1961). Anfinsen's thermodynamic hypothesis details that protein folding is determined by the minimum surface energy (Anfinsen, 1973). The convex nature of the energy landscape, i.e. the presence of a global minimum makes it suitable for computational simulation methods (Tzul et al., 2017). However, highly-simplified models are used due to the complexity of real 3D protein structures.

Stillinger et al. proposed an off-lattice model which details interactions between residues of varying hydrophilicity. The model resembles real protein characteristics due to free-to-rotate bonds between residues (Zhang et al., 2020). Irbäck extended the model to three dimensions with three energy terms, backbone bending energy, torsion energy, and Lennard-Jones potential energy (Irbäck et al., 1997).

Simulated annealing is a Monte Carlo optimization method used to determine global optima. Previous implementations have yielded significant results (Zhang et al., 2020). This project aims to implement a hybrid genetic annealing algorithm to improve computation times. A simulated annealing algorithm consists of an initial solution, a neighbor function, and a cooling schedule. The hybrid genetic annealing algorithm runs multiple annealers in parallel and determines the fitness of annealers at regular intervals. Based on the fitness, daughter annealers are produced and unfit annealers are culled. As the algorithm progresses, the annealer population approaches the ideal parameters.

2. Irbäck's off-lattice model

Irbäck's off-lattice model states that the primary interactions determining protein conformations are hydrophobic (Irbäck et al., 1997). Amino acid residues are reduced to beads and the bonds are represented as rigid rods. Hydrophobic residues are denoted by "A" and hydrophilic interactions are denoted by "B". For a protein of N residues, the structure is determined by 2N-5 independent angles, namely N-2 bond angles $(\alpha_1, \alpha_2 \cdots \alpha_{N-2})$ and N-3 torsion angles $(\beta_1, \beta_2 \cdots \beta_{N-3})$, both in the interval $[-\pi, \pi]$. For a sequence of N residues, the energy can be represented as

$$E = -k_1 \sum_{i=1}^{N-2} \cos\alpha_i - k_2 \sum_{i=1}^{N-3} \cos\beta_i + \sum_{j=1}^{N-2} \sum_{i=j+2}^{N} 4C(\xi_i, \xi_j)\left(\frac{1}{r_{ij}^{12}} - \frac{1}{r_{ij}^{6}}\right)$$

In 3 dimensions, this equation is classified as an NP-hard problem, eliciting the need for methods like simulated annealing (Unger & Moult, 1993). The three terms are backbone bending energy, torsion energy, and Lennard-Jones energy respectively. k_1 and k_2 are parameters controlling the strength of each term and are set to $(-1, 0.5)$. This ensures structural stability (Zhang et al., 2020). The first two terms are independent of the residue sequence and depend only on the bond and torsion angles. The term $C(\xi_i, \xi_j)$ gives the coefficient of interaction between two residues ξ_i and ξ_j. It is defined as

$$C(\xi_i, \xi_j) = \begin{cases} 1; \mathrm{AA} \\ 0.5; \mathrm{AB\ or\ BB} \end{cases}$$

Conveniently, the coefficient matrix can be represented as a scaled outer product of the protein sequence with itself-

$$\mathbf{C} = 0.5(\mathbf{P} \otimes \mathbf{P}) + 0.5\mathbf{J}_{nxn}$$

where $\mathbf{J}$ is square matrix with each element equal to 1.
r denotes the Euclidean norm between the positions of the ith and jth residue.

$$r_{ij} = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|$$

The first three residues lie in a plane. Their coordinates are

$$\mathbf{x_1} = (0, 0, 0)$$
$$\mathbf{x_2} = (0, 1, 0)$$
$$\mathbf{x_3} = (\cos\alpha_1, \sin\alpha_1 + 1, 0)$$

The positions of the remaining residues are obtained from the recursive relation

$$\mathbf{x_n} = \mathbf{x_{n-1}} + (\cos\alpha_{n-2}\sin\beta_{n-3}, \sin\alpha_{n-2}\sin\beta_{n-3}, \sin\beta_{i-3})$$

The optimal protein conformation is the set of angle vectors α^* and β^* which yield the minimum value of the energy function.

3. Annealing

3.1. Serial Simulated Annealing

Simulated annealing is a metaheuristic global optimization method (Kirkpatrick et al., 1983). It uses random search to explore the solution space and has the ability to escape local minima. It is comprised of three main components, an initial solution, a neighbor function, and a cooling schedule.

3.1.1. The initial solution

The initial angle vectors can be determined using the equation

$$\phi = -\pi + 2\pi x$$

where ϕ is a component of an angle vector and $x \in \chi \backsim U([0,1))$

3.1.2. The neighbor function

The neighbor function performs a transformation $\phi \to \phi^*$ where ϕ is a random component of either angle vector.

$$\phi^* = \phi + (x_1 - 0.5)(x_2)\left(1 - \frac{k}{k_{max}}\right)^{\lambda}$$

where $x_1, x_2 \in X \backsim U([0,1))$
k is the current iteration of the annealing cycle and k_{max} is the total number of iterations. λ is called the tuned constant and is a hyperparameter of the algorithm. It determines the heterogeneity of the perturbations introduced.

3.1.3. Traditional cooling schedules

Serial simulated annealing often utilizes nonadaptive cooling schedules such as

$$T_{k+1} = \gamma T_k$$

where $0.8 < \gamma < 1.0$ is the cooling constant. Linear or logarithmic cooling schedules can also be used depending on the nature of the problem.

3.1.4. Algorithm description

- Obtain a protein sequence.
- Generate an initial solution and compute its energy.
- Set the temperature.
- Generate a neighbor and compute the energy difference of the two states.
- If $\Delta E < 0$ or $e^{-\frac{\Delta E}{T}} > x \in \chi \backsim U([0,1))$ set the neighbor to the optimal solution. Repeat steps 4 and 5 for the desired Markov chain length.
- Update the temperature. Repeat steps 3-5 for the desired number of iterations k_{max}.

3.2. Genetic annealing

Genetic annealing is a hybrid of genetic algorithms and simulated annealing. First, an initial population is randomly generated. The candidate solutions are represented as genotypes, and the fitness of each genotype is evaluated. Then, a parent selection method is used. In this implementation, tournament search was used with tournament size 3. Once parents are selected, they crossover with some probability p_c. This is repeated until a new population is created. Then, mutation is performed in the form of simulated annealing to each genotype. Population generation, fitness evaluation, crossover, and mutation are repeated for a set number of iterations. The hybrid algorithm combines the rapid convergence of genetic algorithms with the guarantee of reaching a global minimum provided by simulated annealing.

3.2.1. Cooling schedule

The cooling schedule used is the same as the serial simulated annealing algorithm.

$$T_{k+1} = \gamma T_k$$

3.2.2. The fitness model

The nature of the problem does not require a complicated fitness equation. Fitness is defined as the negative of the energy of a candidate solution-

$$F_i = -E_i$$

This implementation employs selection bias towards the most optimal solution of a population. This is to prevent premature convergence to suboptimal minima.

3.2.3. The crossover model

A single point crossover is employed in this implementation. The bond angle vectors of the two parents are cut at some point and exchange components. The same process occurs for the torsion angle vectors. To prevent population homogenization, crossover occurs with a probability p_c.

3.2.4. The mutator

Artificially selecting optimal genotypes improves computation times but introduces the possibility of premature convergence to a local minimum. To avoid this, a mutation in the form of simulated annealing is carried out.

3.2.5. Algorithm description

- Create an initial population.
- Evaluate the fitness of the population and determine parents through tournament search.
- Perform crossover of parents till a new population of appropriate size is created.
- Perform mutation in the form of simulated annealing.
- Update the temperature.
- Repeat steps 2-5 for the desired number of iterations.

4. Results and discussion

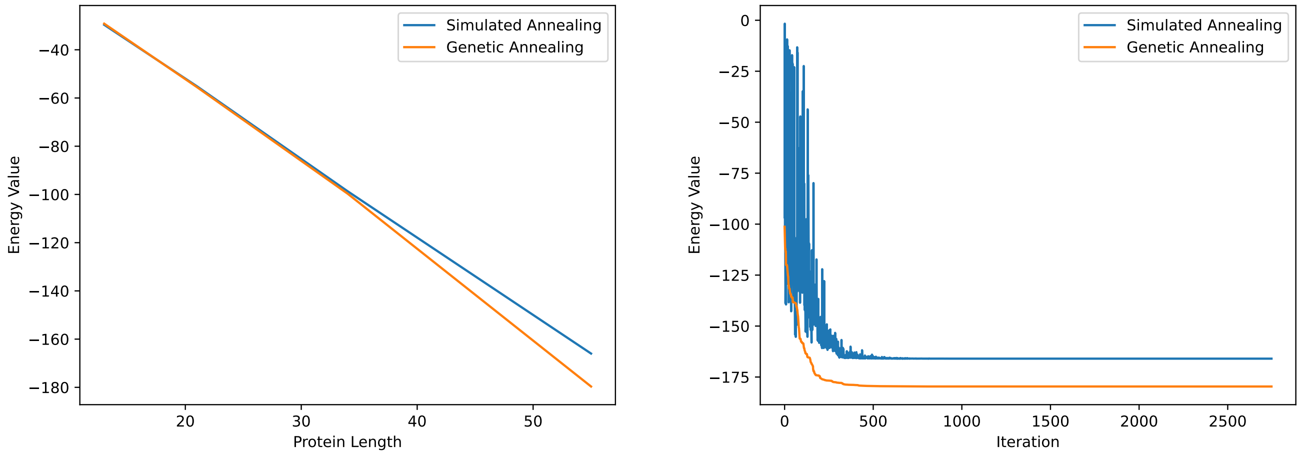
Comprehensive simulations of artificial and real protein structures were carried out. The Fibonacci sequences are used for benchmarking. Real proteins were obtained from the Protein Data Bank (https://www.rcsb.org/). Both algorithms were implemented using Python 3.12.6. The simulations were run on MSU's High Performance Computing Cluster. Parallel processing was leveraged to perform genetic annealing. The hyperparameters of the serial simulated annealing algorithm are as follows: initial temperature $T_0 = 1.0$, cooling coefficient $\gamma = 0.99$, tuned heterogenous constant $\lambda = 3.0$. A Markov chain length of 50000 was used. The mutation in genetic annealing was carried out with the same hyperparameters. In addition, a crossover probability of $p_c = 0.8$ and a population size $P = 50$ was used. The two algorithms were allowed to run for the same amount of time in order to determine their relative efficiency.

4.1. Testing on artificial proteins

Fibonacci protein sequences are usually used for benchmarking (Kim et al., 2005). The first two protein sequences are $S_0 = A$ and $S_1 = B$. Further sequences are defined recursively by

$$S_{n+2} = S_n^* S_{n+1}$$

where * is the concatenation operator. The protein sequences of lengths 13, 21, 34, and 55 were used for benchmarking. The comparative results are seen below, see Figure 1 and Table 1.



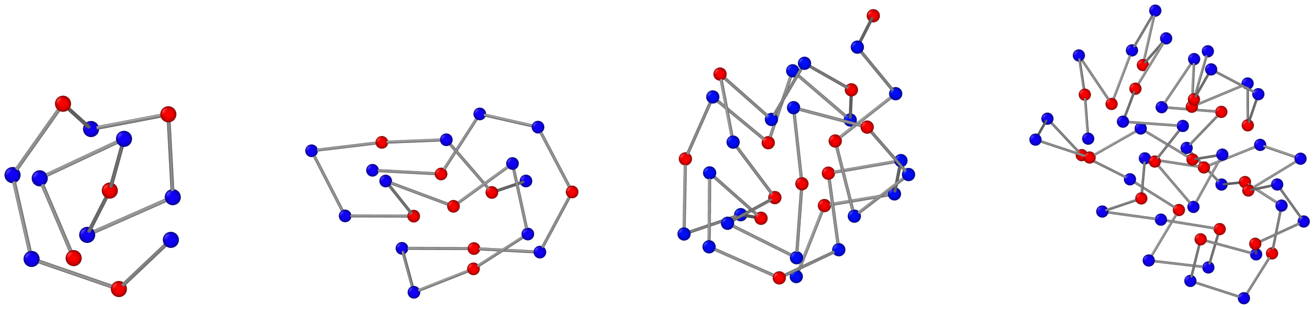
(a) Fibonacci Protein Energy Values

(b) Energy series data from the annealing of S_{55}

Figure 1. Fibonacci Protein Data- rendered in MatPlotLib

Table 1. The lowest energy values reported by various optimization methods and the lowest energy values of SSA and GA

Sequence	$\mathbf{E_{ACMC}}$ (Liang, 2004)	$\mathbf{E_{ELP}}$ (Bachmann et al., 2005)	$\mathbf{E_{ISA}}$ (Zhang et al., 2020)	$\mathbf{E_{SSA}}$	$\mathbf{E_{GA}}$
S_{13}	−26.507	−26.498	−29.474	**−29.643**	−29.166
S_{21}	−51.757	−52.917	**−55.769**	−55.142	−55.482
S_{34}	−94.043	−92.746	**−101.649**	−98.677	−99.686
S_{55}	−154.505	−172.696	**−180.686**	−165.989	−179.653



(a) S_{13}

(b) S_{21}

(c) S_{34}

(d) S_{55}

Figure 2. Fibonacci Protein Structures

4.1.1. Residue topology of artificial proteins

The proteins arrange themselves so that the hydrophobic residues form a core and the hydrophilic residues point outwards, engaging in interactions with the solvent. In the figures 2a, 2b, 2c and 2d, the hydrophobic residues are colored red and the hydrophilic residues are colored blue.

4.2. Testing on real proteins

A benchmark set of 5 proteins from the Protein Data Bank was used for assessing the reliability and comparative efficiency of the different optimization algorithms, see Table 2

Table 2. Prediction results for various proteins

PDB ID	$\mathbf{E_{ISA}}$ (Zhang et al., 2020)	$\mathbf{E_{SSA}}$	$\mathbf{E_{GA}}$
1FCA	−203.330	−205.232	**−214.139**
2OVO	−186.848	−187.252	**−189.624**
2GB1	−176.104	−172.456	**−176.858**
4RXN	−174.612	−175.607	**−184.780**
5PTI	−195.335	−196.318	**−200.731**

4.2.1. Superoptimal conformations

An interesting observation found during the empirical testing of hyperparameters was that the optimal energy conformation did not always correspond to the actual protein structure. Proteins in real life exist in local energy minima, where the energy barriers are too high to escape under normal temperature conditions. It was inferred that due to high annealing temperatures, the solution escaped the local minima corresponding to real protein conformations and achieved a more optimal solution, albeit one that does not exist. This highlighted the importance of proper hyperparameter tuning for annealing algorithms.

4.2.2. Structural alignment via C_α space-filling model

The coordinates of C_α atoms of real protein models were extracted from Protein Data Bank files. However, a direct comparison of predicted and real models was not possible. This is because the predicted models' coordinates were defined in a unitless space, whereas the Protein Data Bank files were in Angstroms. Therefore, the real model was scaled by the equation below

$$\mathbf{x}_{scaled} = \frac{\mathbf{x}}{r_{12}} \forall \mathbf{x} \in \mathbf{P}$$

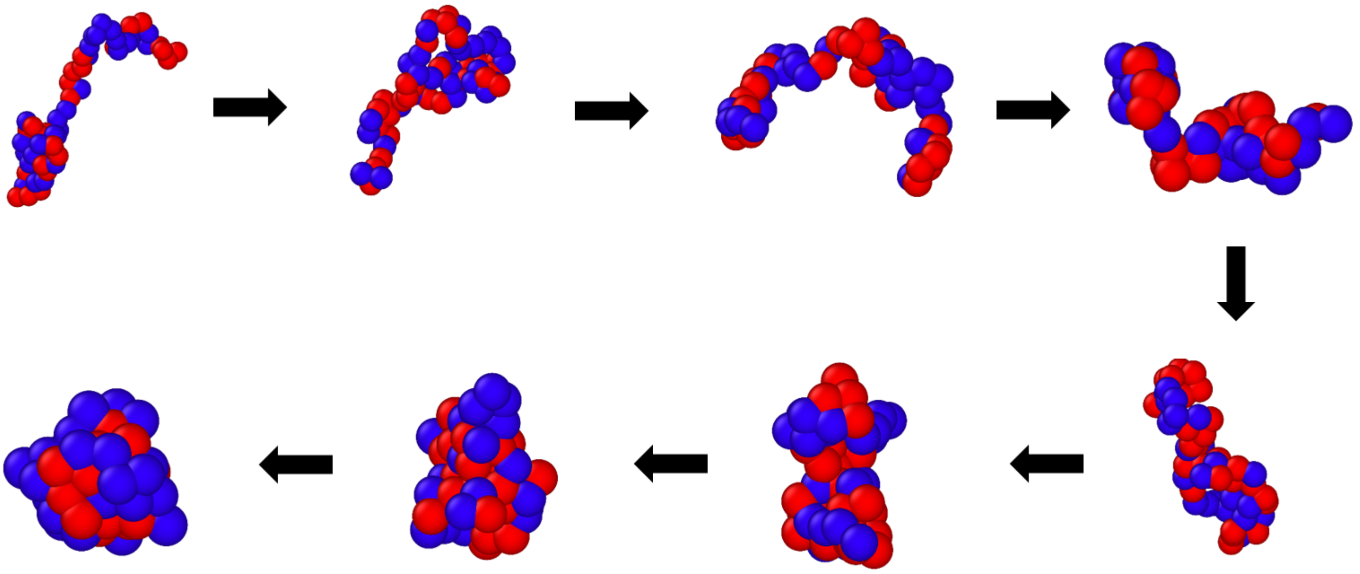


Figure 3. The folding of $5PTI$

where $\mathbf{x}_{scaled}$ is the scaled coordinate, r_{12} is the distance between the first two C_α atoms and $\mathbf{P}$ is the protein conformation.

4.2.3. Protein Folding Simulations

By using the C_α space-filling model, the folding of the protein can be modeled as a temperature series. The evolution of the temperature series of $5PTI$ is shown in Figure 3. The protein gradually folds such that the hydrophobic residues form a core and the hydrophilic residues orient themselves on the surface of the protein.

4.2.4. Accuracy and precision of predicted models

The Root Mean Square Deviation (RMSD) of the actual and predicted protein models was calculated using the equation

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{n} ||\mathbf{x}_{actual} - \mathbf{x}_{real}||^2}$$

The RMSD was calculated using pyMOL (https://www.pymol.org/). The RMSD values for the various annealing runs are listed below. Note that there is no correlation between the energy values presented in Table 2 and the structures here. The reason for this is detailed in Section 4.2.1. The RMSD values lie in the range of $1.875 - 3.468$. The runs with RMSD values less than 3.0 can be considered to be acceptable models (Reva et al., 1998).

The radii of gyration of the predicted models and real models were calculated by the formula

$$R_G = \sqrt{\frac{1}{N}\sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{x}_{CM}||^2}$$

where $\mathbf{x}_{CM}$ is the center of mass of the protein and $\mathbf{x}_i$ is the position of the ith residue. The difference of radii of gyration of the predicted model and real models were calculated and found to be of the order 10^{-3} for serial simulated annealing and order 10^{-5} for genetic annealing.

4.3. Conclusion and future work

The proposed simulated annealing and genetic annealing algorithms successfully predict protein models through the optimization of Irbäck's off-lattice model energy equation. The predicted models are represented using the C_α space-filling model and yield acceptable RMSD values. However, this energy equation only contains three terms. The accuracy of the obtained structures will improve if more terms are added, representing the various other forces at play.

The presence of superoptimal conformations raises the question of the role of annealing algorithms in potentially modeling protein misfolding. Further studies will explore this idea.

Table 3. RMSD values for 10 independent runs of annealing on various real proteins

Run Number	4RXN	5PTI	2OVO	2GB1	1FCA
1	2.174	3.261	2.492	2.983	2.715
2	2.352	3.468	3.013	3.009	2.746
3	2.859	2.953	2.928	**2.298**	2.642
4	2.263	3.354	2.579	2.786	**2.150**
5	3.007	3.231	3.017	2.627	2.548
6	2.003	3.129	2.746	2.704	2.749
7	2.668	3.228	**2.478**	2.999	2.937
8	2.671	**2.570**	2.761	2.804	2.572
9	2.684	2.792	2.863	2.917	2.764
10	**1.875**	2.831	2.830	3.200	2.430

References

Anfinsen, C. B., Haber, E., Sela, M., & White, F. H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences*, *47*(9), 1309–1314. https://doi.org/10.1073/pnas.47.9.1309

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223–230. https://doi.org/10.1126/science.181.4096.223

Bachmann, M., Ark ın, H., & Janke, W. (2005). Multicanonical study of coarse-grained off-lattice models for folding heteropolymers. *Phys. Rev. E*, *71*, 031906. https://doi.org/10.1103/PhysRevE.71.031906

Bagaria, A., Jaravine, V., & Güntert, P. (2013). Estimating structure quality trends in the protein data bank by equivalent resolution. *Computational Biology and Chemistry*, *46*, 8–15. https://doi.org/https://doi.org/10.1016/j.compbiolchem.2013.04.004

Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, *338*(6110), 1042–1046. https://doi.org/10.1126/science.1219021

Güntert, P. (2004). Automated nmr structure calculation with cyana. In A. K. Downing (Ed.), *Protein nmr techniques* (pp. 353–378). Humana Press. https://doi.org/10.1385/1-59259-809-9:353

Irbäck, A., Peterson, C., Potthast, F., & Sommelius, O. (1997). Local interactions and protein folding: A three-dimensional off-lattice approach [Cited by: 97; All Open Access, Green Open Access]. *Journal of Chemical Physics*, *107*(1), 273–282. https://doi.org/10.1063/1.474357

Javidpour, L. (2012). Computer simulations of protein folding [Cited by: 2]. *Computing in Science and Engineering*, *14*(2), 97–103. https://doi.org/10.1109/MCSE.2012.21

Kim, S.-Y., Lee, S. B., & Lee, J. (2005). Structure optimization by conformational space annealing in an off-lattice protein model. *Phys. Rev. E*, *72*, 011916. https://doi.org/10.1103/PhysRevE.72.011916

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680. https://doi.org/10.1126/science.220.4598.671

Liang, F. (2004). Annealing contour monte carlo algorithm for structure optimization in an off-lattice protein model [Cited by: 54]. *Journal of Chemical Physics*, *120*(14), 6756–6763. https://doi.org/10.1063/1.1665529

Reva, B. A., Finkelstein, A. V., & Skolnick, J. (1998). What is the probability of a chance prediction of a protein structure with an rmsd of 6 å? *Folding and Design*, *3*(2), 141–147. https://doi.org/https://doi.org/10.1016/S1359-0278(98)00019-4

Simons, K. T., Bonneau, R., Ruczinski, I., & Baker, D. (1999). Ab initio protein structure prediction of casp iii targets using rosetta [Cited by: 470]. *Proteins: Structure, Function and Genetics*, *37*(SUPPL. 3), 171–176. https://doi.org/10.1002/(SICI)1097-0134(1999)37:3+<171::AID-PROT21>3.0.CO;2-Z

Tzul, F. O., Vasilchuk, D., & Makhatadze, G. I. (2017). Evidence for the principle of minimal frustration in the evolution of protein folding landscapes [Cited by: 53; All Open Access, Bronze Open Access]. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(9), E1627–E16322. https://doi.org/10.1073/pnas.1613892114

Unger, R., & Moult, J. (1993). Finding the lowest free energy conformation of a protein is an np-hard problem: Proof and implications. *Bulletin of Mathematical Biology*, *55*(6), 1183–1198. https://doi.org/https://doi.org/10.1016/S0092-8240(05)80169-7

Zhang, L., Ma, H., Qian, W., & Li, H. (2020). Protein structure optimization using improved simulated annealing algorithm on a three-dimensional ab off-lattice model. *Computational Biology and Chemistry*, *85*, 107237. https://doi.org/https://doi.org/10.1016/j.compbiolchem.2020.107237