Protein Folding: Uniform Random Sampling using the HP-Model

Submitted on: **30-09-2023**

Jesse Kommandeur
j.kommandeur@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Daan van den Berg
d.van.den.berg@vu.nl
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

1 INTRODUCTION

Proteins, complex structures composed of amino acid chains, are crucial in executing a variety of vital processes within the body. These molecules are typically stored in a 'folded' state within cells, a configuration essential for their functional efficacy [1]. However, the misfolding of proteins is implicated in numerous health issues, including cancer, Alzheimer's, and cystic fibrosis, and understanding their multi-dimensional structures is essential for deciphering their functions [1–3].

The artificial synthesis of proteins, a practice initiated in the 1960s, has been pivotal for developing treatments targeting diseases associated with protein misfolding [4]. The identification of native and stable conformations is an integral aspect of this process but is notably time-consuming to do experimentally, leading researchers towards computational and algorithmic methodologies [5].

Yet, the complexity intrinsic to protein folding, especially in predicting the multifaceted structures that proteins adopt, presents formidable challenges when trying to find stable conformations [6, 7]. Proteins' folding pathways are characterized by a staggering level of combinatorial variability. Proteins themselves are diverse, with small peptides composed of around 50 amino acids to gigantic structures like titin, containing more than 27.000 amino acids. The 'common' protein, however, typically consists of approximately 200 to 300 amino acids, with various proteins containing more than 2.000 amino acids [8, 9].

This vast array of potential amino acid sequences and combinations leads to an exponential number of possible conformations, a scenario that already tests the limits of contemporary computational capabilities [10]. While tools like AlphaFold have made remarkable strides, they are not exempt from delivering erroneous folding patterns that diverge from natural occurrences [11, 12].

In the realm of exact algorithms, simplified models like Dill's hydrophobic-polar (HP) model have gained prominence [13]. In an HP model, hydrophobic amino acids (H) like to lie 'adjacently', polar amino acids (P) do not have that preference. When two hydrophobic amino acids lie next to each other, an 'H- bond' is formed due to the attractive forces between the two. And the more bonds, the more stable the protein [13]. The HP model, containing just two types of amino acids, arranges these in a grid or a lattice as shown in Figure 1.

Protein folding is fundamentally an energy-minimization endeavour. Proteins adopt energetically favourable conformations. However, the NP-completeness and NP-hardness of the protein folding problem, even within the simplified HP model, underscore the computational intensity of identifying or verifying maximally stable conformations, particularly given the expansive length of proteins [14, 15].

In light of these complexities, the focus has often shifted towards (meta)heuristic methods. These approaches, however, are not without their encumbrances [16]. The initiation of metaheuristic algorithms with randomly chosen individuals, a practice advocated by various metaheuristic methods such as Genetic Algorithms, Simulated Annealing, Particle Swarm- and Ant Colony Optimization, encounters the obstacle of a search space saturated with invalid conformations. The aspiration for uniformly random and unbiased initial populations for algorithms is challenged by the saturation of the conformation space with invalid structures, raising questions about the adaptability of genetic algorithms to the HP-protein folding conundrum [17]. This leads us to the following research question: How can the uniform randomness in the initial populations of metaheuristic algorithms, such as genetic algorithms, hill climbing, or simulated annealing, be ensured to enhance the accuracy and efficiency of protein folding, and what is the impact of different resampling techniques on the quality and diversity of solutions in the context of the HP-protein folding problem?"

2 RELATED WORK

The HP-protein folding problem, a well-known challenge in computational biology, has seen various methodological approaches since Unger and Moult's seminal work in 1993 as described in Janssen et al. (2023) [17]. Their study, one of the first to employ genetic algorithms (GA) in this domain, compared Monte Carlo methods with GAs, although it left certain technical aspects, like the treatment of collisions and initialization, ambiguous [18]. Subsequent works echoed this limitation, offering valuable but partial insights into the management of invalid conformations and the need for resampling during mutation, crossover, and initialization.

Patton et al. (1995) made significant strides, pioneering a GA that accommodated and penalised collision-rich conformations, thus facilitating quicker identification of lower-energy conformations [19]. The subsequent work of Custodio et al. (2004) built upon this, introducing diversity preservation in the selection and an innovative fitness function favouring compact, natural conformations. However, this method still hinged heavily on extensive evaluations [20].

On a different trajectory, Bui and Sundarraj (2005) concentrated on evolving the secondary structures of the hydrophobic subsection independently, leaning on a robust library of structures [21]. The necessity for numerous repair mechanisms hinted at the nuanced challenge of managing collisions. Lin and Hsieh (2009) presented a hybrid model, blending the Taguchi method and particle swarm optimization with GAs, showing improved performance, yet leaving collision treatment and uniform random sampling opaque [22].

Garza-Fabre et al. (2015) navigated towards multi-objective optimization, striving for a balance between collision numbers and conformation stability, unveiling foundational aspects of the issue but resulting in a large state space dominated by mostly invalid conformations [23]. In a similar vein, Wang et al. (2016) ventured into cloning and 'chaotic mutation', yet the collision treatment and sampling method remained pickwickian. Other recent contributions, like those from Boumedine and Bouroubi (2021) and Atari and Majd (2022), integrated advanced concepts like hill climbing and quantum genetic algorithms but left gaps in presenting substantial data on collision frequency and how random initial populations were sampled [24, 25].

This review by Janssen et al. (2023) consistently unveiled a recurring constraint: the limitation of protein lengths to 64 or 100, which is less than half of the average protein length[8, 17] This aspect underscored the inherent challenge of random sampling and raised questions on the broad applicability of iterative algorithms for extensive HP-protein folding scenarios. It therefore serves as a beacon, elucidating the nuanced challenges of solution sampling. Their inquiry dives into the core issue of uniform sampling from the entire array of valid conformations, with an emphasis on deterministic time algorithms and the expected number of resamples. They spotlight two contrasting strategies - strict inclusion of valid solutions and a lenient approach permitting but penalizing collisions.

Their experiment, involving the random folding of 'neutralized' proteins, revealed a linear increase in collision numbers with protein length, casting doubts over the feasibility of uniformly sampling valid initial solutions for extensive proteins. This insight propels the discourse beyond the technical and methodological terrain navigated by predecessors like Unger and Moult (1993), and Patton et al. (1995), to the foundational quagmire of solution sampling [18, 19]. It accentuates the complexity intrinsic to the dance between computational capability and biological fidelity, as echoed in prior works.

Furthermore, the work of Janssen et al. (2023) produces conformations that are uniform random samples. This signifies that if all valid conformations were to be enumerated, the probability of selecting any one of these conformations would be uniformly random. This is of significant importance since genetic algorithms, hill climbing, or simulated annealing inherently have the potential to initiate from a uniform random starting condition [16]. This observation is intriguing, particularly because a review of the literature suggests a lack of attention or consideration given to this aspect by those employing genetic algorithms within the realm of protein folding. Yet, this is a critical step; if selection is not uniformly random, there exists a probability of consistently locating the same local optimum.

Additionally, for certain problems such as the travelling salesman problem, it is easier to employ uniform random sampling. One can simply select a specific number of random cities, sequence them, and be assured of a uniformly random solution. However, this appears to be an unattainable approach in protein folding. Consequently, an approach undertaken by Janssen et al. (2023) involved evaluating the frequency of required resampling to attain a valid conformation without violations [17]. The advantage of this method is its adherence to uniform randomness. However, a significant drawback is the extensive repetition of resampling, particularly since the likelihood of encountering a violation, especially for large proteins, is exceedingly high. While resampling can potentially be performed with backtracking, it remains uncertain if this yields a uniformly random sample.

In this context, the uniform randomness of conformational selections is essential to ensure the reliability and robustness of computational methods like genetic algorithms, hill climbing, or simulated annealing in exploring the vast conformational space of proteins and identifying their native structures. This lack of uniform randomness could potentially bias the exploration and trapping the optimization process into local optima. The criticality of addressing this issue underscores a pivotal frontier in enhancing computational efficiency and accuracy in the field of protein folding.

The eclectic mix of insights from these studies paints a rich, multifaceted landscape of HP-protein folding. The journey from the pioneering works of Unger and Moult (1995) to the intricate insights of Janssen et al. (2023) unveils a narrative rife with complexities. The nuances of GA application, collision management, and the foundational enigma of solution sampling coalesce into a compelling narrative that sets the stage for future endeavours.

In this paper, we will start by extending the work of Janssen et al. (2023) as shown in Figure 2a. Instead of focusing on the inclusion of valid solutions and a lenient approach permitting but penalizing collisions, we will stop once we encounter collisions for different values of n, and look at their distribution. Subsequently, we deploy a backtracking algorithm on the same set of proteins to discover to what extend also produces uniform randomness of the conformational population of proteins as displayed in figure 2b.

3 METHODOLOGY

In the initial phase of our experiment, we performed a random folding of 1000 'neutralized' 1D strings of proteins, each with lengths n 5,10,15,...,95,100, on a 2D lattice, void of any prior assumptions. We termed the process 'neutralization' as it does not take into account the 'hydrophobic' and 'polar' labels from the aminos, underscoring our focus on valid conformations rather than optimal ones, a method similar to the work of Janssen et al. (2023) [17]. The amino acids were placed sequentially, with each subsequent placement randomly selecting a position from the set left, right, straight, in adherence to a chain-relative representation, while avoiding backward collisions. We kept a count of the total number of amino acids - n_{aminos} - each time an amino was positioned on the 2D lattice, continuing until no valid adjacent locations remained as showed on figure 2b.

The range for n_{aminos} in any given conformation is established with a minimum value of 6, a scenario that arises in the unique instance of a circle collision occurring as a result of random placement, and a maximum value equating to the total length of the respective amino string. A notable trend observed is the decrement in the number of valid conformations concomitant with an increase in amino length. This is attributed to the escalating likelihood of collisions as the 1D protein strings elongate.

In the course of our experiment, for each specified length n, we meticulously documented the amino length associated with each of the 1,000 conformations generated. This data was then represented in a histogram, offering a visual and quantitative depiction

of the distribution and frequency of conformational lengths. This graphical representation serves as an instrumental tool, not only in validating our preliminary observations but also in providing intricate insights into the nuanced dynamics governing protein folding and conformational variability.

In the subsequent phase of the experiment, we mirrored the initial steps of the first phase. The same 1000 'neutralized' 1D strings of proteins were randomly folded on a 2D lattice, adhering to the same lengths and conditions. However, a significant divergence was introduced at the point where no valid adjacent locations remained. Rather than concluding the process, we implemented backtracking to navigate through previously explored paths and explore alternative configurations.

The integration of backtracking aimed to enhance the depth and breadth of our exploration within the configurational space of the protein strings. This process involved retracing our steps to previous positions where alternative paths existed and then proceeding along these unexplored paths. The main impetus behind this was to unearth a wider array of valid conformations, offering a comprehensive insight into the diverse structural possibilities inherent within the protein strings.

4 RISK ASSESSMENT

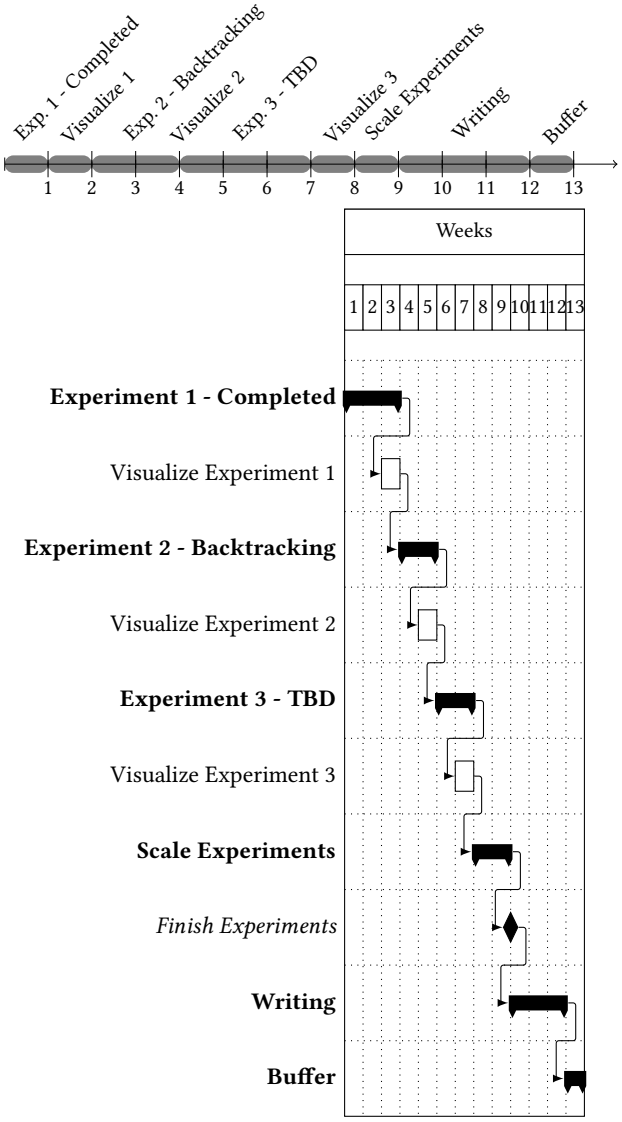
One of the challenges of this research emanates from the computational limitations inherent in addressing the intricate and expansive conformational space of proteins. The intricate nature of protein folding, exacerbated by the number of potential conformations, demands substantial computational power and optimized algorithms to facilitate efficient runtime for both experiments 2, 3 and the scale-up of all experiments. Mitigation of these challenges is instrumental in ensuring the project's success. Adaptation to emerging insights and continuous enhancement of algorithms, coupled with strategic allocation and resource augmentation, form the approach's cornerstone.

The fusion of interdisciplinary insights from computer science and bioinformatics, continuous learning, and adaptation to emerging knowledge is intrinsic to my strategy to bridge these gaps and augment the research's robustness. In the intricate dance of unfolding the mysteries of protein folding, unforeseen challenges are not just possible but are a certainty. A contingency plan, characterized by flexibility, is in place to navigate the unpredictable, ensuring that the research remains on course amidst complexities.

5 PROJECT PLAN

The project was initiated with the successful completion of the first experiment, and we are currently immersed in the phase of data visualization, slated for completion within a week. Experiment 2, marked by the integration of a backtracking algorithm, is next, projected to span two weeks, with an additional week earmarked for analyzing and visualizing the results. Experiment 3, scheduled immediately afterwards for two weeks, is designed to be adaptive and therefore TBD, its structure and focus refined based on insights from Experiment 2.

Following these experimental phases, a two-week comprehensive evaluation is planned to distil key findings and insights. The scaling of experiments is then embarked upon; over two additional weeks, the validated findings are extrapolated and assessed for broader applications and implications. The writing phase is allocated four weeks, a period dedicated to the research findings, discussions, and conclusions. A two-week buffer is strategically incorporated to mitigate unforeseen challenges and delays, ensuring that the final thesis submission is both timely and of exemplary quality.



6 APPENDIX

This section refers to the hosted source code and figures. All code is open source and publicly available on GitHub.

6.1 Github

The source code of this Protein Folding research project is hosted on GitHub using the MIT License. Under the public *Protein* repository, I have several code repositories which will be updated on a weekly basis:

(1) Code - This folder contains all notebooks with codes for each of the experiments.
https://github.com/jessekommandeur/Protein-Folding
(2) Data - This folder contains all generated datasets that were used in this research project
https://github.com/jessekommandeur/Protein-Folding
(3) Thesis - Here you can view and/or read the thesis itself. It's also available on:
https://github.com/jessekommandeur/ProteinFolding
(4) Presentation slides - These are all presentations that were given during the weekly thesis sessions under the supervision of Daan van den Berg at the VU Campus in Amsterdam
https://github.com/jessekommandeur/ProteinFolding/

6.2 Figures

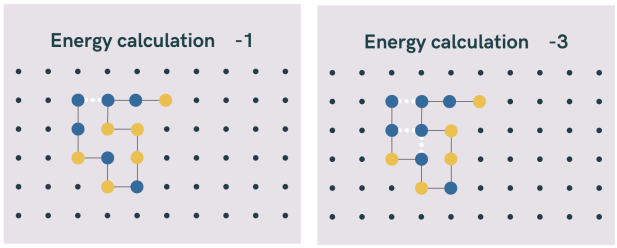


Figure 1: HP-model
2 amino acid samples for length n=12, with a stability of -1 and -3

Figure 2: Sampling Methods
Figure 2a refers to the collision sampling. A method demonstrated in Jansen et al (2023) [17]. Figure 2b refers to the first experiment of this research, Figure 2c refers to the third experiment of this research.

REFERENCES

[1] Cheolju Lee, Soon-Ho Park, Min-Youn Lee, and Myeong-Hee Yu. Regulation of protein function by native metastability. *Proceedings of the National Academy of Sciences*, 97(14):7727–7731, 2000.
[2] Philip J Thomas, Bao-He Qu, and Peter L Pedersen. Defective protein folding as a basis of human disease. *Trends in biochemical sciences*, 20(11):456–459, 1995.
[3] Christopher M Dobson. Protein misfolding, evolution and disease. *Trends in biochemical sciences*, 24(9):329–332, 1999.
[4] Max F Perutz, Michael G Rossmann, Ann F Cullis, Hilary Muirhead, Georg Will, and ACT North. Structure of hæmoglobin: a three-dimensional fourier synthesis at 5.5-å. resolution, obtained by x-ray analysis. *Nature*, 185(4711):416–422, 1960.
[5] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *science*, 338(6110):1042–1046, 2012.
[6] Thomas E Creighton. The protein folding problem. *Science*, 240(4850):267–267, 1988.
[7] Ken A Dill, S Banu Ozkan, M Scott Shell, and Thomas R Weikl. The protein folding problem. *Annu. Rev. Biophys.*, 37:289–316, 2008.
[8] Jianzhi Zhang. Protein-length distributions for the three domains of life. *Trends in Genetics*, 16(3):107–109, 2000.
[9] Nigel Chaffey. Alberts, b., johnson, a., lewis, j., raff, m., roberts, k. and walter, p. molecular biology of the cell. 4th edn., 2003.
[10] Okke Van Eck and Daan Van Den Berg. Quantifying instance hardness of protein folding within the hp-model. In *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7. IEEE, 2023.
[11] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.
[12] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
[13] Kit Fun Lau and Ken A Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, 1989.
[14] Bonnie Berger and Tom Leighton. Protein folding in the hydrophobic-hydrophilic (hp) is np-complete. In *Proceedings of the second annual international conference on Computational molecular biology*, pages 30–39, 1998.
[15] William E Hart and Sorin Istrail. Robust proofs of np-hardness for protein folding: general lattices and energy potentials. *Journal of Computational Biology*, 4(1):1–22, 1997.
[16] Agoston E Eiben and James E Smith. *Introduction to evolutionary computing*. Springer, 2015.
[17] Reitze Jansen, Ruben Horn, Okke van Eck, Sarah L Thomson, and Daan van den Berg. Can hp-protein folding be solved with genetic algorithms? maybe not. 2023.
[18] Ron Unger and John Moult. Genetic algorithms for protein folding simulations. *Journal of molecular biology*, 231(1):75–81, 1993.
[19] Arnold L Patton, William F Punch III, and Erik D Goodman. A standard ga approach to native protein conformation prediction. In *ICGA*, pages 574–581, 1995.
[20] Fábio L Custódio, Hélio JC Barbosa, and Laurent E Dardenne. Investigation of the three-dimensional lattice hp protein folding model using a genetic algorithm. *Genetics and Molecular Biology*, 27:611–615, 2004.
[21] Thang N Bui and Gnanasekaran Sundarraj. An efficient genetic algorithm for predicting protein tertiary structures in the 2d hp model. In *Proceedings of the 7th annual conference on Genetic and evolutionary computation*, pages 385–392, 2005.
[22] Cheng-Jian Lin and Ming-Hua Hsieh. An efficient hybrid taguchi-genetic algorithm for protein folding simulation. *Expert systems with applications*, 36(10):12446–12453, 2009.
[23] Mario Garza-Fabre, Eduardo Rodriguez-Tello, and Gregorio Toscano-Pulido. Constraint-handling through multi-objective optimization: the hydrophobic-polar model for protein structure prediction. *Computers & Operations Research*, 53:128–153, 2015.
[24] Nabil Boumedine and Sadek Bouroubi. A new hybrid genetic algorithm for protein structure prediction on the 2d triangular lattice. *arXiv preprint arXiv:1907.04190*, 2019.
[25] Moein Atari and Nayereh Majd. 2d hp protein folding using quantum genetic algorithm. In *2022 27th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–8. IEEE, 2022.