THE ROSENBLUTH SAMPLING CALCULATION OF HYDROPHOBIC-POLAR MODEL

Marcin Wierzbiński
Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw, S. Banacha 2,
02-097 Warsaw and
Sano
Centre for Computational Medicine
Czarnowiejska 36 building C5
30-054 Cracow, Poland
m.wierzbinski@sanoscience.org

Alessandro Crimi
Sano
Centre for Computational Medicine
Czarnowiejska 36 building C5
30-054 Cracow, Poland
a.crimi@sanoscience.org

ABSTRACT

Lattice proteins are models resembling real proteins. They comprise an energy function and a set of conditions specifying the interaction between elements occupying adjacent lattice sites. In this paper we present an approach examining the behavior of chains of a large number of molecules. We investigate this by solving a restricted random walk problem on a cubic lattice and square lattice. More specifically, we apply the *hydrophobic-polar* model to examine the spatial characteristics of protein folds using the Monte Carlo method. This technique is the so-called Rosenbluth sampling method for solving restricted random walk problems. Specifically, by solving such walks we obtain plausible folds. In addition, this method can be extended to solve the hydrophobic-polar model. In this paper, we describe this method as an algorithm that calculates the energy spectrum for the hydrophobic-polar model, and the related formula for estimating the number of folds. Moreover, we estimate the number of folds for each sequence using hydrophobic-polar model energy estimation. On test sequences the predicted protein folds were obtained with a mismatch of one unit according to the energy. We also observe that the estimated number of folds depends only on the length and not on the type of sequence. This promising strategy can be extended to quantify other proteins in nature.

1 INTRODUCTION

The search for a more efficient algorithm of protein folding in the hydrophobic-polar (HP) model is an important aspiration in many disciplines ([12], [9]). Knowing how proteins fold can help elucidate their three-dimensional structure-function relationship, which is crucial to the understanding of enzymes and to the treatment of misfolded-protein diseases such as Alzheimer's, Huntington's, and Parkinson's disease. The numerical simulation focused on those proteins is particularly useful for drug design, as it allows to test different physical characteristics using models of various complexities. Indeed, if high-resolution chemical structure is used, leading to precise molecule representations, dynamical simulation showing atomic interactions can be reached. This might ultimately provide more effective and personalized drugs.

It has been shown that the HP protein folding model is NP-Hard ([1]), which means it is difficult to solve efficiently for longer protein sequences. In order to overcome this obstacle, many heuristic algorithms have been proposed ([4], [13]). Besides heuristics mostly based on optimization, other approaches are based on the idea that co-operativity of folding occurs, as local conformational choices which constraints the optimization space in which solutions are searched. Those assumption-based methods include hydrophobic zipper method [3], which assumes that once a hydrophobic contact is created it cannot be broken. And the core-directed chain growth method [2] which constrains the optmization search within the space of solutions having a hydrophobic core with a square (in 2D) or a cube (in 3D).

In this context, there is theoretical and experimental evidence of the advantage of solving a restricted random walk problem (RRW) on cubic and square lattices. One of the earliest proposed numerical algorithms which apply the RRW paradigm is the one designed by M. Rosenbluth and A. Rosenbluth ([11]). In this report, we present a benchmark implementation of Rosenbluth methods for the HP model with an additional extension to estimate the number of possible sequence configurations.

1.1 HYDROPHOBIC-POLAR MODEL

In the hydrophobic-polar model, the set of twenty standard amino acids is reduced to two: H (hydrophobic amino acid) and P (hydrophilic amino acid). Moreover, the amino acids are approximated to be located in a 3D lattice space where bonds are either $90\circ$ or $180\circ$. More formally, the model relies on *embedding* a given finite polypeptide *sequence* $s = (s_1, \ldots, s_i, \ldots, s_k)$ where $s_i \in \{H, S\}$, into a given infinite graph G. In this article, the graph G will primarily be the three-dimensional cubic lattice $G = \mathbb{Z}^3$ and square cubic lattice $G = \mathbb{Z}^2$ over integer numbers $\mathbb{Z}$. A *fold* of length k for s in G is an injective mapping $\omega : [1, \ldots, k] \mapsto G$ such that adjacent integers map to adjacent points of G. In addition, each point is assigned one letter from the polypeptide sequence at position i. Such neighboring points form a *bond*. Each point of $\mathbb{Z}^3$ has six neighbors $(x \pm 1, y \pm 1, z \pm 1)$. The energy of the fold of s is expressed as

$$\mathrm{E}(s, \omega) = \sum_{1 \le i < j \le k} E_{s_i, s_j} \Delta(\omega(i), \omega(j)) \text{ and} \tag{1}$$

$$\Delta(p, q) = \begin{cases} 1 & \text{if p and q are adjacent but do not connect amino acids,} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

with energy equation:

$$E_{(s_i, s_j)} = \begin{cases} -1 & s_i = H \text{ and } s_j = H \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The above equation for calculating the energy of fold s in G can also be expressed as negation of the number of $H - H$ bonds in the fold, where a bond is a pair of symbols corresponding to adjacent points, except for those H's which are adjacent to pairs of sequences s. The goal of the HP model is to minimize the energy $E(s, \omega)$.

1.2 ROSENBLUTH SAMPLING METHOD

The method proposed by [11] involves drawing successive steps of a random walk only from among *acceptable points*, which are points previously not visited. In this section, we describe the random procedure in more detail. We will focus on the 3D case, i.e $G = \mathbb{Z}^3$, but the method is easily transferable to the 2D case.

For a given number of adjacent points k in the fold, any configuration consisting of k adjacent points laid out joined in succession on a cubic lattice $\mathbb{Z}^3$ is considered. Regarded as a random walk problem, for any walk consisting of m adjacent points and ending at position $(x, y, z)_m$, all six positions are a priori equally likely at iteration $m + 1$. The excluded volume effect is simulated by the requirement that a fold is not allowed to cross itself or back up on itself at any iteration. Consequently, at any iteration, there are at most five possible positions to move to. For simplicity, we assume that the first link originates from $(0, 0, 0)$. Any satisfactory set of m adjacent points start from the origin $(x, y, z)_m$ is associated with a weighting function W_m of possible positions calculated at each step according to the procedure described below. At any iteration m where the most recent link terminates at $(x, y, z)_m$ and 5 potential positions $(x \pm 1, y \pm 1, z \pm 1)_m$ must be considered, while position $(x, y, z)_{m-1}$ is ruled out immediately. All five remaining potential positions at $m + 1$ may be associated with values $(x, y, z)_i$ for $i = m - p$ where p is an odd number greater than 1. If the comparison reveals this to be the case, a modification of the weight W_m must be made, obtaining W_{m+1}.

Below we present all possible cases at iteration m:

1. All six new position $(x \pm 1, y \pm 1, z \pm 1)_{m+1}$ are occupied. The process is then terminated with weight $W_{m+1} = 0$.

2. Only l new positions are unoccupied, with $0 < l \leq 5$. Then

$$W_{m+1} = W_m \cdot l \tag{4}$$

During this process, an embedding is generated. If the embedding is equal to the length of the sequence, we can calculate energy according to the presented formula.

Estimation of fold numbers In this section, we present a mathematical justification for the estimation of folds.

Let us assume in general terms that when constructing fold f_i of length k the following is true:

$$w_1 = 6, w_2, \ldots, w_m. \tag{5}$$

It is not excluded that at a certain step we may have no further possibilities for continuation, i.e., $w_{m+1} = 0$.
We then say that a *non-extendable* fold of length m has been formed. Let $\mathcal{Y}$ denote the set of all folds of length $m = k$ and non-extendable folds of length $m < k$. The set of all folds of length $m = k$ is denoted as $\mathcal{Z}$. Clearly, $\mathcal{Z} \subset \mathcal{Y}$.

The probability of picking a random fold $f_i \in \mathcal{Y}$ of length m is equal to:

$$P(f_i) = \frac{1}{w_1} \cdot \frac{1}{w_2} \cdot \ldots \cdot \frac{1}{w_m} \tag{6}$$

with a weight function for the specific fold f_i. Let $f_i \in \mathcal{Y}$, so

$$W(f_i) = \begin{cases} w_1 \cdot w_2 \cdot \ldots \cdot w_k & \text{if } f_i \in \mathcal{Z}_k (\text{ so } m = k) \\ 0 & \text{if } f_i \notin \mathcal{Z}_k \left(\text{ so } l < k \text{ i } w_{m+1} = 0\right). \end{cases} \tag{7}$$

One can interpret $W(f_i)$ as the weight of fold f_i. Let us now repeat the draw using the growth method n times. There are n random folds $f_1, f_2, \ldots, f_n$ from set Y.

Let n_s denote the number of drawn elements f_i for which $W(f_i) = s$ and the set of these elements $\mathcal{W}_s = \{f \in \mathcal{Y} : W(f) = s\}$. Then, based on the large numbers law:

$$\frac{n_s}{n} \approx P(\mathcal{W}_s). \tag{8}$$

Therefore, the above expression can be written as the average weight of the drawn folds. We note that:

$$\frac{1}{n} \sum_{i=1}^{n} W\left(x_i\right) = \sum_s s \frac{n_s}{n} \approx \sum_s sP\left(\mathcal{W}_s\right) = \sum_s s \sum_{f \in \mathcal{W}_s} P(x) = \sum_f P(f) W(f). \tag{9}$$

Finally, the expression for $\bar{W}$ can be written as:

$$\bar{W} = \frac{1}{n} \sum_{i=1}^{n} W\left(f_i\right) \approx \sum_f P(f) W(f) = \sum_{f \in \mathcal{Z}_k} 1 = |\mathcal{Z}_k|. \tag{10}$$

We introduce the following notation for fold estimators of length k:

$$\hat{\mathcal{Z}_k} = \bar{W} \approx |\mathcal{Z}_k|. \tag{11}$$

To validate this fold estimator we test sequences of different length and type and the results are reported in the following section.

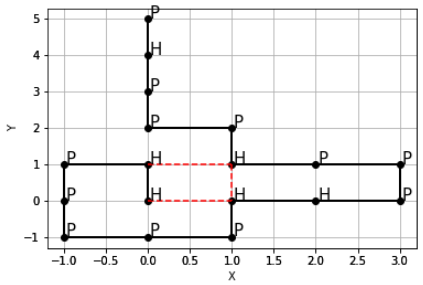
2 RESULTS

The experiments were run on Google's Colab platform on Intel(R) Xeon(R) CPU @ 2.20GHz with 13GB RAM. We investigated 2 different dataset one 2D and one 3D.

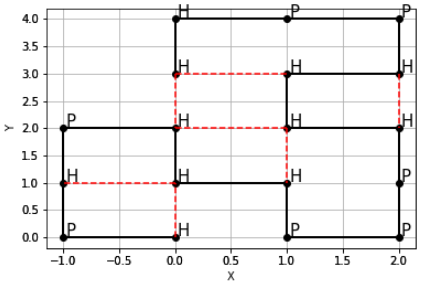
2.1 BENCHMARK FOR 2D

Using the method proposed above, we calculate $\hat{\mathcal{Z}}_k$ with a statistical error. The algorithm was initially tested for several sequences of dimension 2 (for $\mathbb{Z}^2$ from site [6].

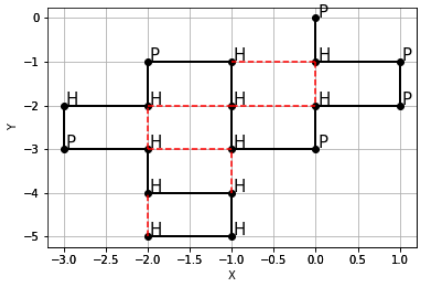
The method was run for $n = 10^5$ of suitable configurations folds with a specific sequence s of length k. This first experiment lasted 5 minutes. The algorithm code, written in Python, can be found at the following website: [8]. Symbols H^i and P^i in the table correspond to i repetitions of sequence characters. The results related to this dataset are reported in Table 1, with 4 examples of resulting predicted folding depicted in Figure 1.



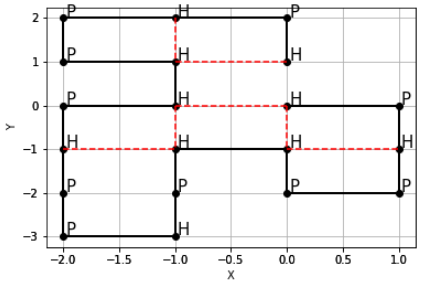
(a) Embedding of sequence $s = HH(P)^5HH(P)^3H(P)^3HP$ with the minimum energy $\hat{E_{min}}(s,\omega) = -3$



(b) Embedding of sequence $s = HPHPHHH(P)^3H^4PPHH$ with the minimum energy $\hat{E_{min}}(s,\omega) = -7$.



(c) Embedding of sequence $s = (PHP)^2(H)^3(PHH)^2H^3$ with the minimum energy $\hat{E_{min}}(s,\omega) = -8$.



(d) Embedding of sequence $s = (HP)^2PHH(PHP)^2HHP(PH)^2$ with the minimum energy $\hat{E_{min}}(s,\omega) = -7$.

Figure 1: In each figure we embed a given finite polypeptide sequence in the square lattice $G = \mathbb{Z}^2$. The energy in the presented diagrams can be easily deduced. Each red line indicates a bond. The number of these edges corresponds to energy $\hat{E_{min}}$.

2.2 ESTIMATION FOR 3D

The experiments were performed using $n = 10^5$. The code can be found in [8]. Estimated energy is equal to 0 for all k. This is because it is difficult to fold for $\mathbb{Z}^3$ so that the number of $H - H$ bonds is minimised. The second experiment lasted 4 minutes. Results for this dataset are reported on Table 2, with 2 examples of resulting predicted folding depicted in Figure 2.

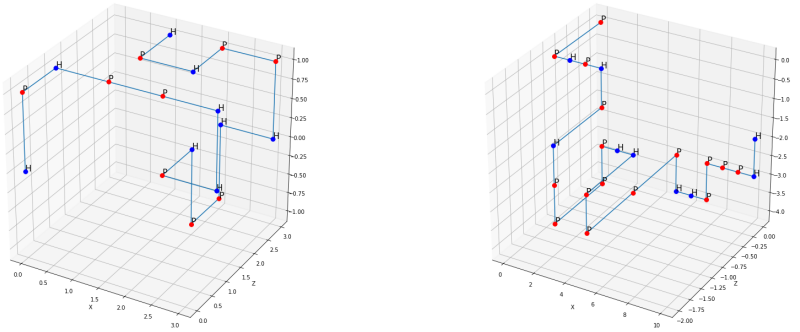
The experiment itself shows how difficult it is to wrap sequences in 3 dimensions. Estimations for sequences 24 and 25 alone show that the number of folds is on the order of 10^{16} and 10^{17}, as shown in Table 1 and 2. However, the energy of the fold is still zero. Therefore, the 3D model is significantly more difficult than the 2D model.

k length fold	$\hat{\mathcal{Z}_k}$ estimated folds	$\hat{E_{min}}(s,\omega)$	E_{bench}	tested sequence s
18	125253209.5	-3	-4	$HH(P)^5HH(P)^3H(P)^3HP$
18	124733316.6	-7	-8	$HPHPHHH(P)^3H^4PPHH$
18	124948202.9	-8	-9	$(PHP)^2(H)^3(PHH)^2H^3$
20	890716733.1	-7	-9	$(HP)^2PHH(PHP)^2HHP(PH)^2$
20	893580948.1	-8	-10	$HHHP(PH)^3PP(HP)^3PH$
24	46020717810.8	-6	-7	$HHPP(HPP)^5HH$
25	123640609088.4	-7	-8	$PP(HHP)^4(PPP)^3HH$
36	5972669674348712.0	-10	-14	$(P)^3(HHPP)^2(P)^2(H)^7PPHH(P)^4H^2PPHP^2$

Table 1: In this table we compare how our method performs against the model from [6]. We can conclude that our energy is minimally different. Having computed an estimation for all folds $\mathbb{Z}^2$ for each sequence, we can conclude that the number of folds $\hat{\mathcal{Z}_k}$ does not depend on the tested sequence s. We can observe that the estimated number of folds depends only on length k and not on the form of s.

k length fold	$\hat{\mathcal{Z}_k}$ estimation of folds	tested sequence s
1	6	H
2	30	HP
3	150	HPH
4	726.0	$HPHP$
5	3533.7	$HPPHP$
6	16928.8	$HPPHPH$
10	8813146.6	$HHPPPPHHHP$
18	2214862342500	$HH(PP)^2PHHPPPHPPPHP$
20	49644802682812.5	$(HP)^2PH(HP)^2(PH)^2HP(PH)^2$
24	2.4776796891665252e+16	$P(PH)^3(PP)^2HH(PP)^2HH(PP)^2H$
25	1.1658696555396226e+17	$P(PH)^3(PP)^2HH(PP)^2HH(PP)^2HH$

Table 2: In the accompanying table we count the fold estimation values for dimension 3 for $\mathbb{Z}^3$. Referring to experiment 1 we see that there is a significant increase in the number $\hat{\mathcal{Z}_k}$ of these folds for each sequence. For values of $k = 24, 25$ we observe particularly large differences.



(a) Embedding of sequence $s = (HP)^2PH(HP)^2(PH)^2HP(PH)^2$ (b) Embedding of sequence $s = P(PH)^3(PP)^2HH(PP)^2HH(PP)^2HH$

Figure 2: In the graph above, we can observe that the fold did not wrap in such a way that the two $H-H$s are next to each other. Therefore, energy is equal to 0. This is because we have significantly more degrees of freedom in 3D space.

For the initial cases $k = 1, 2$, the results are obvious. If the length of the sequence is 1 ($k = 1$), then we have only 6 possible points. For $k = 2$ we have 36 possible point assignments, 6 of which are forbidden. We have also prepared special sequences for cases $k \in 1, 2, 3, 4$ that will always have zero energy. It is not possible to wrap a sequence in such a way that points with $H - H$ labels touch each other.

3 DISCUSSION

Correctly predicting protein conformations based on the amino acid sequence is of pivotal importance for drug design and other relevant computational chemistry tasks. In this paper we report our computational experiments, where we use HP sequences corresponding to published benchmarks [6] with a 2D lattice in the HP model. Our model successfully estimates the number of folds for a particular sequence; regardless on the type of sequence but only on its length. For small sequences, the method accurately estimates the number of folds. Our experiments show that for sequences of size $k = 24, 25$ the 3D model becomes significantly more complex than the 2D model. It has been observed that adding one dimension significantly affects the solution base. In 2D, the energies are -6 and -7, respectively, when in 3D the energy is zero for both cases. There are too many degrees of freedom to draw consecutive points. Therefore, it is difficult to find a wrap that has non-zero energy even for shorter sequences. However, the Rosenbluth sampling method can be successfully used to estimate the number of all folds, especially those with energy 0. This can be helpful in designing heuristic algorithms based on this hindsight. The estimation itself, according to our mathematical justification, increases in accuracy as we increase the number of iterative executions of the method. The described approach is effective for identifying and sampling configurations on a lattice geometry. This kind of representations can be useful in the context of ab initio protein structure prediction [10]. Expansions as implementations on quantum devices have been proposed, but those have been limited to the 2D case so far [7]. Conversion of the proposed tool into quadratic unconstrained binary optimization (QUBO) [5] using 3D lattices on quantum devices will be investigated as future work.

3.1 ACKNOWLEDGEMENTS

This research is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement Sano no 857533, and by the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund.

REFERENCES

[1] B Berger and T Leighton. "Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete". eng. In: *Journal of computational biology* 5.1 (1998), pp. 27–40. ISSN: 1066-5277.

[2] Thomas C Beutler and Ken A Dill. "A fast conformational search strategy for finding low energy structures of model proteins". In: *Protein Science* 5.10 (1996), pp. 2037–2043.

[3] Ken A Dill, Klaus M Fiebig, and Hue Sun Chan. "Cooperativity in protein-folding kinetics." In: *Proceedings of the National Academy of Sciences* 90.5 (1993), pp. 1942–1946.

[4] Tianzi Jiang et al. "Protein folding simulations of the hydrophobic–hydrophilic model by combining tabu search with genetic algorithms". In: *The Journal of chemical physics* 119.8 (2003), pp. 4592–4596.

[5] Gary Kochenberger et al. "The unconstrained binary quadratic programming problem: a survey". In: *Journal of combinatorial optimization* 28.1 (2014), pp. 58–81.

[6] ISTRAIL LABORATORY. *HP 2D Benchmarks*. `https : / / www . brown . edu / Research / Istrail _ Lab / hp2dbenchmarks . html`. [Online; accessed 02-Feb-2022]. 2011.

[7] Cristian Micheletti, Philipp Hauke, and Pietro Faccioli. "Polymer Physics by Quantum Computing". In: *Physical Review Letters* 127.8 (2021), p. 080501.

[8] *Monte Carlo Calculation of Protein Folding*. https://github.com/marcin119a/Monte-Carlo-Calculation-of-Protein-Folding. [Online; accessed 02-Feb-2022]. 2022.

[9] Vijay S Pande. "Simple theory of protein folding kinetics". eng. In: *Physical review letters* 105.19 (2010), pp. 198101–198101. ISSN: 0031-9007.

[10] Mahmood A Rashid et al. "Guided macro-mutation in a graded energy based genetic algorithm for protein structure prediction". In: *Computational biology and chemistry* 61 (2016), pp. 162–177.

[11] Marshall N Rosenbluth and Arianna W Rosenbluth. "Monte Carlo Calculation of the Average Extension of Molecular Chains". eng. In: *The Journal of chemical physics* 23.2 (1955), pp. 356–359. ISSN: 0021-9606.

[12] Andrej Sali, Eugene Shakhnovich, and Martin Karplus. "How does a protein fold?" In: *Nature (London)* 369.6477 (1994), pp. 248–251.

[13] Nicola Yanev et al. "Protein folding prediction in a cubic lattice in hydrophobic-polar model". In: *Journal of Computational Biology* 24.5 (2017), pp. 412–421.