Protein Folding in the Two-dimensional Hydrophobic Polar Model based on Cellular Automata and Local Rules

Alia Madain , Abdel Latif Abu Dalhoum , and Azzam Sleit

Department of Computer Science King Abdulla II School for Information Technology The University of Jordan, Amman

Summary
Cellular Automata are discrete computational models that rely on local rules. The main focus of this paper is to build a model of proteins based on simple and local rules of a cellular automaton. Research in this direction depend mainly on combining cellular automata with other paradigms. Many schemes in literature rely on different evolutionary algorithms to support the use of cellular automata and some depend on combining protein parameters with parameters extracted from a cellular automaton image. The aim here is to keep the simplicity of cellular automata as much as possible. It is not known yet if a set of local rules that can solve the protein folding problem does exist. So far, research depend on some sort of searching or a global view of the sequence in order to find a reasonable confirmation. This paper discusses what simple rules can be like. The proposed cellular automaton rules and states depend on a well-known simple exact model and the basic principles governing protein folding. In the proposed cellular automaton, the cell state can be a hydrophobic amino acid, a polar amino acid, an empty cell, or a control cell. The argument of local rules is supported by graphical examples of applying the proposed rules.

Key words:
Protein Folding, Cellular Automata, 2D HP Model, Local Rules, Moore Neighborhood.

1. Introduction

Cellular Automata (CA) is a discrete model studied in computability theory. The model is discrete in terms of time and space. CAs rely mainly on the concept of local rules. When the local rules are applied in iterations, the cellular automaton exhibit a certain behavior. The concept was first proposed in the field of computer science in the forties by John von Neumann as formal models of self-reproducing organisms [1].

A cellular automaton pattern can be found in nature and has its roots in biology, which made the concept of CA naturally linked to building accurate models of the central dogma of molecular biology processes. Consequently, CAs were used in studying the behavior of gene networks [2], DNA sequence evolution [3] [4] [5], DNA duplication, and mRNA transcription from DNA [6].

Based on the central dogma of molecular biology, the result of the DNA decoding process is proteins. Proteins are lead performers of the cell functions where each protein has its own task [7]. The processes that results in a functioning protein are quite complex and full of details [8].

Building models of any biological phenomenon usually consists of finding a certain abstraction of some type. Models can be used to understand a problem from a certain point of view. In general, Models help in further experimentation [9]. Since the proteins are complex there are many simplified models or simple exact models (SEMs) of proteins. The most commonly used SEM is the HP model where amino acids are classified to be either a hydrophobic amino acid or a polar amino acid [10].

Modeling proteins based on CA was studied before; previous studies depend on combined models, where CA parameters are added to other parameters. In addition, there are models where CA rules depend on an evolutionary algorithm to find an optimal confirmation.

In this paper, we seek to map the general rules involved in protein folding to a two-dimensional CA. The model proposed is simple, deterministic and rely on local rules. No additional searching is required.

The remaining of the paper is organized as follows: Section 2 discusses related work; Section 3 and section 4 give background information on cellular automata and proteins simplified models; Section 5 presents the proposed scheme; Section 6 discusses the results and finally section 7 concludes the work done and provides possible future work.

2. Related Work

There are many attempts to model proteins based on cellular automata. We will start with those models that combine cellular automata with evolutionary algorithms as in the case of [11] where a genetic algorithm is used with CA to predict proteins secondary structures. A CA like structure or what is called a neural CA was proposed in two and three-dimensional HP models [12] [13].

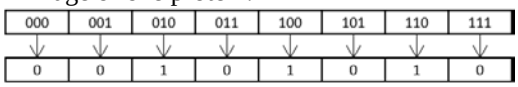
A three dimensional CA model was proposed in [14], which is a theoretical model that presents the use of

Manuscript received September 5, 2016
Manuscript revised September 20, 2016

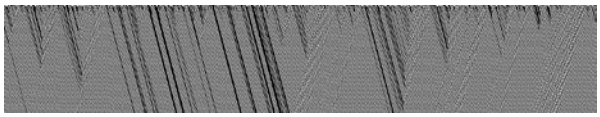
heuristic rules of biochemistry and thermodynamics in a 3D cubic space.

The PseAAC protein composition given in [15] combined with elementary rule 84 were proposed to classify proteins based on their structure classes [16] [17] and to predict transmembrane regions [18] and G-protein-coupled receptor functional classes [19].

Simply, the model converts the protein to a binary representation using the digital coding proposed in [20] [21]. After the protein is converted to a binary sequence it is considered as the initial configuration of the CA 84. the CA runs for 100 generations. Figure 1 shows the rule and the CA image of one protein.



(a) Rule 84



(b) Rule 84 after 100 Generations Representing One Protein Sequence

Fig. 1 CA Rule 84.

The resulting image parameters are extracted using a separate algorithm such as GLCM texture features [22] and Hu geometric moments [23].

Finally, there are systems different from cellular automata but proven to be equivalent in terms of computational power to cellular automata such as L-systems [24] [25]. These systems were also used in protein modeling in [26], [27], and [28].

In this paper, we propose the use of simple deterministic rules and represent the problem using the HP model. The proposed model does not depend on any searching and represent the problem as simple as possible. The results section gives some examples of protein sequences folded based on the rules proposed.

3. Cellular Automata

Cellular automata are discrete models that depend on local and simple rules to produce some overall global behavior. According to Wolfram the overall behavior of a CA can be classified into one of four main classes (stable, periodic, chaotic, and complex) [29] [30] [31].

There is no superior behavior of a CA. If the CA is used to model a natural phenomenon, the best behavior of the CA is the one describing the phenomenon in a meaningful way. The same applies to other applications, where different CAs with different behavioral classes can be applied to the same application [32] [33].

CAs have many variants. Figure 2 shows some of the ways CAs may differ. The spatial distribution of the CA cells can be in any dimension and the cells may be in any shape. The shape of the cells effects the neighborhood, which effects the complexity and behavior of the CA.

Theoretically, the spatial space is assumed infinite. In terms of implementation, some assumptions are usually made such as the use of null boundaries and periodic boundaries. Sometimes the assumption of infinite grid can be realized when the CA considers a specific initial configuration that may move to a limited distance, so the CA may start with a two-dimensional grid of a proper size that there will be no need to configure the boundaries.

Fig. 2 CA Variants.

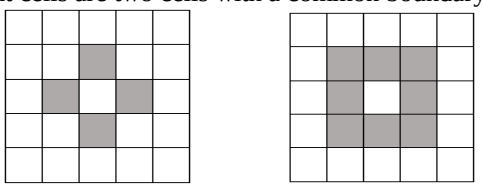
Another parameter that effects the cell states from one generation to the other is the cell neighborhood considered. The neighborhood of a core cell is those cells considered when the core cell changes its state. The actual neighborhood is crucial for the global behavior of a CA [34]. Two common neighborhood types are von Neumann and Moore neighborhood shown in Figure 3.

4. Proteins Problem Simplified

The process of building computational models of any natural phenomenon starts with proper abstraction. Any model represents an aspect of a reality but not all of the reality [9].

The hydrophobic-polar protein folding model proposed by dill [35][36], usually referred to as the HP model, is a popular and a simplified model that emphasizes the hydrophobic and hydrophilic properties of proteins. Simple exact models can account for the properties that

characterize protein folding [36]. The complexity of the simplified HP model was studied for a long time [37] [38].
Based on the HP model assumptions, the primary chain of the protein contains two types of amino acids, namely, the hydrophobic (non-polar) amino acids and the polar (hydrophilic) amino acids. The way to evaluate a certain confirmation is to assign an energy value of -1 to each H-H contact. The lower the energy the better so the best confirmation is the one with maximum number of H-H contacts.
Not all H-H contacts are counted. If the core cell has a hydrophobic amino acid, the H-H contacts counted are only those contacts that are placed in one of the four directions (up, down, left, right) surrounding the core cell. These cells must have a hydrophobic amino acid and must not be part of the primary chain as given in definition 1.
Definition 1: H-H contact can be defined as two hydrophobic amino acids that are not part of the protein primary sequence and placed in two adjacent cells where adjacent cells are two cells with a common boundary.



(a) von Neumann (b) Moore

Fig. 3 Different CA Neighborhood

5. Proposed CA

This section describes the proposed CA in terms of its parameters, rules and initial configuration. In subsection 5.1 the main parameters are given and in subsection 5.2 the initial configuration assumed and the CA rules proposed are described.

5.1 CA Parameters

CAs differ in their spatial distribution, cell neighborhood, transition rules, cell possible states, boundary, number of generations (iterations), cells shape, and the initial configuration from where the CA starts. Table 1 summarizes the parameters used in the proposed model.

Table 1: CA Parameters

No.	*Parameter*	*Value*
1	Neighborhood	Moore
2	Dimension	Two-dimensional
3	Possible States	H, P, Empty, control
4	Boundary	Empty cells

The CA proposed is a two-dimensional one with a homogeneous regular grid. The CA states can be one of the following: amino acid state (either H or P), empty state or a control state. The neighborhood assumed is Moore neighborhood, which includes the diagonals of a core cell.
The empty cells in the proposed scheme takes part in the CA emergence. If the CA considers only those cells with amino acids it would be easy to avoid the boundaries by adjusting the size of the grid. In this work, the boundary is configured to be cells of empty state.

5.2 Initial Configuration and CA Rules

We simply assume that the amino acids of the primary sequence are located at (x=0) in the initial configuration of the CA, and we assume that the row located at (x=0) corresponds to the core of the folded protein. So assuming that water is surrounding the protein (as it is in nature) then polar amino acids will surround the hydrophobic core.
This is the first rule: If it is a polar amino acid then it should surround the hydrophobic core and the hydrophobic amino acids stay in place. If the polar amino acid move in the direction from the core of the protein to the water environment, the sequence will move accordingly since the protein is a connected chain of amino acids.
In other words, if the cell in the cellular automata is in state P, and P is going to move over the 2D grid, then all cells to the right are going to move in a cascade or a recursive manner. Assuming x on the vertical axis and y on the horizontal axis, and one cell is going to move, the following rules apply:

1. If the state in the next cell to the right is (P) then it will move over both dimensions (x and y)
2. If the cell to the right has an (H) state, it will move over the y axis only. If the position is occupied then the H state will move over (x and y), just as in the case of (P) state.

Moreover, in the core of the protein, if there are an even number of spaces or empty cells between any two cells, a compactness rules applies. It is applied after all amino acids move to their new positions.

6. Results

This section presents some examples of the use of the simple rules. The examples are given as a graphical representation of the amino acids states. There are no moves in certain generations of the CA, so only generations that contain significant movements are presented. The hydrophobic amino acids are presented in black and polar amino acids are white. In Table 2, the sequences used in the examples are given. Those

sequences are part of the benchmark introduced in [39] and used by [40].

Table 2: Sequences used in the Examples.

No.	*Sequence*	*Length*
1	$(HP)_2PH(HP)_2(PH)_2HP(PH)_2$	20
2	$H_2P_2(HP_2)_6H_2$	24
3	$P_2HP_2(H_2P_4)_3H_2$	25

Figure 4 shows the first example of employing the simple rules, the sequence is of length 20. The first amino acid of the protein primary structure is hydrophobic then the state of the first cell in the cellular automata will stay the same and the movement will start from the second polar amino acid. The compactness rule is applied twice in this sequence as shown in Figure 4g.
The second example is a sequence of length 24, this example does not require any compactness rules as shown in figure 5. The third example starts with a polar amino acid and contains 25 amino acids as given in Figure 6.

7. Conclusion and Future Work

This work studies the use of local rules in protein folding. A cellular automaton is proposed with controllers and compactness rules. The cellular automaton is two-dimensional and the proteins presentation used is the well-known HP simplified model.
The rules keep the hydrophobic amino acids in the core of the protein and moves the polar amino acids from the core, then a compactness rule may apply, depending on the availability of even spaces between amino acids.
Unlike work done in the field of protein modeling based on cellular automata, the model proposed does not need any searching algorithms nor does it depend on any special representation or evolutionary algorithms. The work can be enhanced with more local movement rules, which is left for future work.

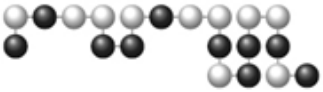
(a) Primary Structure at x=0

(b) Amino Acid in the 2nd Position of the Primary Structure

(c) Amino Acid in the 8th Position of the Primary Structure

(d) Amino Acid in the 13th Position of the Primary Structure

(e) Amino Acid in the 16th Position of the Primary Structure



(f) Amino Acid in the 19th Position of the Primary Structure

(g) Compactness Rule

Fig. 4 Sequence Number 1 of Length 20

(a) Primary Structure at x=0

(b) Amino Acid in the 3rd Position of the Primary Structure

(c) Amino Acid in the 6th Position of the Primary Structure

(d) Amino Acid in the 9th Position of the Primary Structure

(e) Amino Acid in the 12th Position of the Primary Structure

(f) Amino Acid in the 15th Position of the Primary Structure

(g) Amino Acid in the 18th Position of the Primary Structure

(h) Amino Acid in the 21th Position of the Primary Structure

Fig. 5: Sequence Numb

(a) Primary Structure at x=0

(b) Amino Acid in the 1st Position of the Primary Structure

(c) Amino Acid in the 4th Position of the Primary Structure

(d) Amino Acid in the 8th Position of the Primary Structure

(e) Amino Acid in the 14th Position of the Primary Structure

(f) Amino Acid in the 20th Position of the Primary Structure

(g) Compactness Rule

Fig. 6: Sequence Number 3 of Length 25

References

[1] P. Sarkar, "A brief history of cellular automata," ACM Comput. Surv., vol. 32, no. 1, pp. 80–107, 2000.

[2] J. A. de Sales, M. L. Martins, and D. A. Stariolo, "Cellular automata model for gene networks," Phys. Rev. E, vol. 55, pp. 3262–3270, Mar 1997.

[3] C. Burks and D. Farmer, "Towards modeling dna sequences as automata," Physica 10D, vol. 10, no. 1-2, pp. 157–167, 1984.

[4] G. Sirakoulis, I. Karafyllidis, C. Mizas, V. Mardiris, A. Thanailakis, and P. Tsalides, "A cellular automaton model for the study of dna sequence evolution," Computers in Biology and Medicine, vol. 33, no. 5, pp. 439– 453, 2003.

[5] C. Mizas, G. Sirakoulis, V. Mardiris, I. Karafyllidis, N. Glykos, and R. Sandaltzopoulos, "Reconstruction of dna sequences using genetic algorithms and cellular automata: Towards mutation prediction?" Biosystems, vol. 92, no. 1, pp. 61–68, 2008.

[6] D. Takata, T. Isokawa, N. Matsui, and F. Peper, "Modeling chemical reactions in protein synthesis by a brownian cellular automaton," in 2013 First International Symposium on Computing and Networking, Dec 2013, pp. 527–532.

[7] P. Koehl, Protein Structure Classification. John Wiley & Sons, Inc., 2006, pp. 1–55.

[8] M. Gibson and E. Mjolsness, Computational modeling of genetic and biochemical networks. Cambridge MA: MIT Press, 2004, vol. 8, no. 1, ch. Modeling the Activity of Single Genes, pp. 3–48.

[9] J. C. Wooley and H. S. Lin., Catalyzing Inquiry at the Interface of Computing and Biology. Washington (DC): National Academies Press (US), 2005, ch. Computational modeling and simulation as enablers for biological discovery, pp. 117–202.

[10] E. Ferrada, "The amino acid alphabet and the architecture of the protein sequence-structure map. i. binary alphabets,"

PLOS Computational Biology, vol. 10, no. 12, pp. 1–20, December 2014.
[11] P. Chopra and A. Bender, "Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature," In Silico Biology, vol. 7, no. 7, pp. 87–93, 2006.
[12] J. Santos, P. Villot, and M. Dieguez, "Cellular automata for modeling protein folding using the hp model," in Evolutionary Computation (CEC), 2013 IEEE Congress on, June 2013, pp. 1586–1593.
[13] J. Santos, P. Villot, and M. Di´eguez, "Emergent protein folding modeled with evolved neural cellular automata using the 3d HP model," Journal of Computational Biology, vol. 21, no. 11, pp. 823–845, 2014.
[14] A. Madain, A. L. A. Dalhoum, and A. Sleit, "Computational modeling of proteins based on cellular automata," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 7, pp. 491– 498, 2016.
[15] K.-C. Chou, "Prediction of protein cellular attributes using pseudoamino acid composition," PROTEINS: Structure, Function, and Genetics, vol. 43, pp. 246–255, 2001.
[16] X. Xiao and W. Ling, "Using cellular automata images to predict protein structural classes," in Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on, July 2007, pp. 346– 349.
[17] X. Xiao, P. Wang, and K.-C. Chou, "Predicting protein structural classes with pseudo amino acid composition: An approach using geometric moments of cellular automaton image," Journal of Theoretical Biology, vol. 254, no. 3, pp. 691–696, 2008.
[18] Y. Diao, D. Ma, Z. Wen, J. Yin, J. Xiang, and M. Li, "Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and lempel-ziv complexity," Amino Acids, vol. 34, no. 1, pp. 111–117, 2008.
[19] X. Xiao, P. Wang, and K.-C. Chou, "Gpcr-ca: A cellular automaton image approach for predicting g-protein-coupled receptor functional classes." Journal of Computational Chemistry, vol. 30, no. 9, pp. 1414– 1423, 2008.
[20] X. Xiao, S. Shao, Y. Ding, and X. Chen, "Digital coding for amino acid based on cellular automata," in Systems, Man and Cybernetics, 2004 IEEE International Conference on, vol. 5, Oct 2004, pp. 4593–4598.
[21] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen, and K.-C. Chou, "Using cellular automata to generate image representation for biological sequences," Amino Acids, vol. 28, no. 1, pp. 29–35, 2005.
[22] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEE Transactions on Systems Man and Cybernetics SMC3, vol. 3, no. 6, pp. 610–621, 1973.
[23] M.-K. Hu, "Visual pattern recognition by moment invariants," IRE Transactions on Information Theory, vol. 8, no. 2, pp. 179–187, February 1962.
[24] A. L. A. Dalhoum, A. Ortega, and M. Alfonseca, "Cellular automata equivalent to d0l systems," in 3rd WSEAS International Conference on Systems Theory and Scientific Computation, Special Session on Cellular Automata and Applications, 2003, pp. 15–17.
[25] A. Ortega, A. A. Dalhoum, and M. Alfonseca, "Grammatical evolution to design fractal curves with a given dimension," IBM Journal of Research and Development, vol. 47, no. 4, pp. 483–493, 2003.
[26] G. B. Danks, S. Stepney, and L. S. D. Caves, Folding Protein-Like Structures with Open L-Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 1100–1109.
[27] G. Danks, S. Stepney, and L. Caves, "Protein folding with stochastic l-systems," Artificial Life XI, pp. 150–157, 2008.
[28] G. B. Danks, S. Stepney, and L. S. D. Caves, Cotranslational Protein Folding with L-systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 289–296.
[29] S. Wolfram, A New Kind of Science. Wolfram Media Inc., 2002.
[30] "Universality and complexity in cellular automata," Physica D: Nonlinear Phenomena, vol. 10, no. 12, pp. 1–35, 1984.
[31] "Statistical mechanics of cellular automata," Rev. Mod. Phys., vol. 55, pp. 601–644, Jul 1983.
[32] A. Madain, A. Abu Dalhoum, H. Hiary, A. Ortega, and M. Alfonseca, "Audio scrambling technique based on cellular automata," Multimedia Tools and Applications, vol. 71, no. 3, pp. 1803–1822, 2014.
[33] A. L. Abu Dalhoum, A. Madain, and H. Hiary, "Digital image scrambling based on elementary cellular automata," Multimedia Tools and Applications, pp. 1–16, 2015.
[34] H. Nishio, "How does the neighborhood affect the global behavior of cellular automata?" in Cellular Automata, ser. Lecture Notes in Computer Science, S. El Yacoubi, B. Chopard, and S. Bandini, Eds. Springer Berlin Heidelberg, 2006, vol. 4173, pp. 122–130.
[35] K. A. Dill, "Theory for the folding and stability of globular proteins," Biochemistry, vol. 24, no. 6, pp. 1501–1509, 1985.
[36] K. A. Dill, S. Bromberg, K. Yue, H. S. Chan, K. M. Ftebig, D. P. Yee, and P. D. Thomas, "Principles of protein folding a perspective from simple exact models," Protein Science, vol. 4, no. 4, pp. 561–602, 1995.
[37] B. Berger and T. Leighton, "Protein folding in the hydrophobic hydrophilic (hp) is np-complete," in Proceedings of the Second Annual International Conference on Computational Molecular Biology, ser. RECOMB '98. New York, NY, USA: ACM, 1998, pp. 30–39.
[38] V. Chandru, A. DattaSharma, and V. A. Kumar, "The algorithmics of folding proteins on lattices," Discrete Applied Mathematics, vol. 127, no. 1, pp. 145 – 161, 2003, computational Molecular Biology Series - Issue {IV}.
[39] R. Unger and J. Moult, "Genetic algorithms for protein folding simulations," Journal of Molecular Biology, vol. 231, no. 1, pp. 75 – 81, 1993.
[40] T. N. Bui and G. 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, ser. GECCO '05. New York, NY, USA: ACM, 2005, pp. 385–392.

Alia Madain is a PhD Candidate in the Department of Computer Science at the University of Jordan. She received her B.Sc. and M.Sc. degrees in Computer Science from the University of Jordan in 2009 and 2011, respectively. She has about six years of work experience in governmental and private sectors. Her research interests are security, unconventional computing and complex systems.

Abdel Latif Abu Dalhoum received his PhD degree in computer science from the University Autonoma De Madrid, Spain in 2004. He is currently an Associate Professor of evolutionary algorithms and complex systems at the University of Jordan. He has published about 25 papers in evolutionary algorithms, cellular automata, Fractals and DNA computing. He is a member of GHIA research group at the University Autonoma De Madrid.

Azzam Sleit is a Professor of Computer Science with the University of Jordan and the former Minister of ICT from 2013 to 2015. He is the former Dean of IT and Director of the Computer Center at the University of Jordan. Prof. Sleit has over twenty years of experience and leadership in the IT field working at various levels of government, private and international sectors. Prof. Sleit is the author of more than eighty-five research papers published in reputable journals and conferences.