ZILELE ACADEMICE CLUJENE 2018, INFORMATICA, Volume **LX**, Number 1, 2018
DOI: ...

PARALLEL ALGORITHM FOR PROTEIN FOLDING SIMULATION ON HP MODEL. EXPERIMENTAL STUDY

IOAN SIMA

Abstract. Proteins are the essential components of the living systems which ensures their functioning. The predicting of the protein function requires both the knowing of the sequence of the amino acids and their three-dimensional conformation. The protein folding problem consists in finding the native conformation from a huge number of conformations. The backtracking algorithm is not suitable for parsing of some spaces of exponential dimensions. However, in this paper, by applying a parallel backtracking algorithm and by applying techniques to reduce conformational space, have been obtained all the conformations for a few known sequences in literature and for k-hephutoxin, a real protein from scorpion venom.

1. Introduction

Protein macromolecules, important cell components, are the basis of the living. They are made of chains of tens to hundreds of thousands of amino acids (shortly aa). Amino acids are molecules with chemical opposite characteristics (acid and alkaline). In the living organisms are found approximate one hundred types of amino acids. From these, 20 amino acids, called proteinogenic amino acids, form the proteins[5]. Because globular proteins are immersed in the aqueous cell cytoplasmic environment, the ability of amino acids to attract (hydrophilic amino acids) or to repel (hydrophobic amino acids) water molecules is essential for the understanding of the way proteins are folding. Thereby, they can be classified into: H aa and P aa, where H means Hydrophobic and P means polar (or Hydrophilic).

Among the important factors which determine the protein function can be mentioned: the amino acid sequence of the protein and the protein conformation (the three-dimensional or the folded form). Given a DNA sequence, the determination of amino acids sequence of the protein is a simple process. Instead, the determination of the three-dimensional conformation for a given protein is a process incompletely solved yet.

Received by the editors: oct 2018.
Key words and phrases. Protein Folding Simulation, Backtracking, 2D HP Model.

Protein folding is the physical process by which their amino acid sequence, called primary structure (1D), is transformed into native conformation, named tertiary or quaternary structure (3D). The phenomenon is important because proteins have biological activity only in folded form (called conformer). The protein folding problem consists in predicting or finding the native conformation from the primary structure[1]. Addressing this problem is a challenge because the number of possible conformations is huge [8]. For insulin, the size of the conformational space (number of conformation) is $\approx 3^{300}$ or 10^{143}. The thermodynamic hypothesis, advanced in the 1960s, states that proteins fold in the conformation that has the minimum energy.

2. HP Model

Several types of models have been proposed to simulate protein folding. Thus, there are all-atom models, in which the units are the atoms, or coarse-grain models, in which the units are the amino acids. In another possible classification, there are off-lattice or on-lattice models.

The HP Model [7], proposed by Lau and Dill in 1989, is a simple and coarse-grained on-lattice model in which the 20 types of amino acids are reduced to 2 types: **H** aa and **P** aa. The angles under which the amino acids can be placed, side by side, are 90 degrees, 180 degrees and 270 degrees, respectively. The lattice on which aa can be placed can be 2D (rectangular) or 3D (cubic) type.

Each conformation has associated an arbitrary value, called energy. The energy of a conformation is given by the sum of the energy of a contact between two H aa which are topological neighbors but are not neighbors in the protein sequence. Arbitrary, the energy of a H..H contact is considered -1. Conformation with the smallest energy will form in the protein center an hydrophobic kernel, the same with the one of the real proteins. The energy function is briefly presented in the Table 1. Berger, in 1998, showed that the

aa	**H**	**P**
H	-1	0
P	0	0

Table 1. Energy function

protein folding problem in the HP model is NP-complete[3] and Bahi et al.[2], in 2011, reveals a chaotic behavior.

Shortly, the characteristics of the HP Model are:

- Unit: amino acids
- Alphabet: {H,P}

- No of connections: 4
- Connections type: HH, HP, PH, PP
- Lattice type: rectangular (2D), cubic (3D)
- NP-complete problem
- Chaotic behavior

3. Backtracking algorithm

3.1. **Why is necessary the backtracking algorithm?** Backtracking is an algorithm whose purpose is finding all solutions (or some solutions) from space possible solution, applicable to some computational problems. It incrementally builds candidates to the solutions and abandons a candidate which cannot be completed to a valid solution[4].

To solve the protein folding problem on the HP model, deterministic and non-deterministic algorithms were applied. A lot of non-deterministic algorithms, especially AI approach, have a large share. It is known that AI techniques find optimal local solutions, but there is no guarantee that these are the optimal general solution . Moreover, because protein folding on the HP model has chaotic behavior [2], AI algorithms have difficulties in finding the optimal conformations.

Hence it is necessary to apply deterministic algorithms that go through the whole solution space to establish with certainty the minimum energy conformations for a given sequences of amino acids. These conformations could be used to test the quality of non-deterministic algorithms. Applying the backtracking algorithm has been avoided because that the solution space is huge. Moreover, solution space increases exponentially with increasing the length of the amino acids sequence. Increasing of computing power of the processors in recent years, as well as the existence of parallel architectures, justify the application of the backtracking algorithm in the case of relatively short sequences of amino acids. In this paper, after applying of some methods to eliminate the areas that contain unfeasible solutions from the solution space, has been applied a parallel backtracking algorithm with 3 and 9 threads (processors), respectively.

3.2. **The reducing of the computational effort in the 2D HP model.** To increase the efficiency of the backtracking algorithm and reduce the search time of the optimal conformations, several pruning techniques have been applied.

3.2.1. *The reducing of the combinatorial space.* The directions in which amino acids can be placed relative to each other can be encoded absolutely (L - left; R - right; U - Up; D - Down) or encoded relative (S-straight; L-left; R- right).

In absolute codification, the size of the space solution for the brute force algorithm is 4^{n-1}, where n is the number of amino acids. The first amino acid is fixed in the center of the lattice. The second amino acid can be placed in one of the four directions (L, R, U, D). Because the four directions generate symmetric conformations, the second amino acid can be fixed on an arbitrary chosen position, too. In this work, the second amino acid is fixed on the R direction from the first amino acid (figure 1). Thus, the solution space is reduced four times, from 4^{n-1} to 4^{n-2}. For relative directions the reduction is

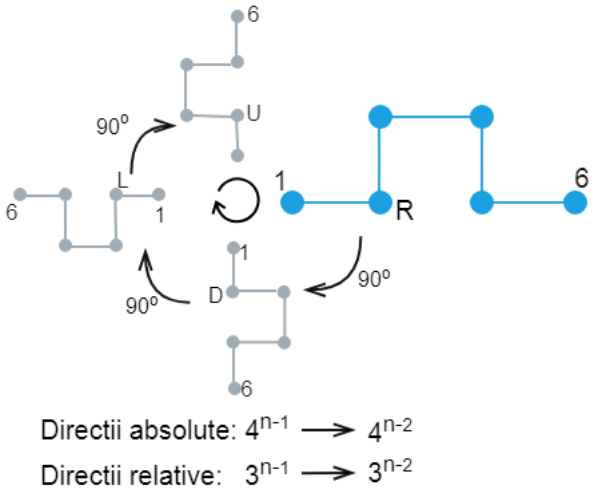

FIGURE 1. The elimination of the three directions

from 3^{n-1} to 3^{n-2}, because there are only 3 directions. But the use of relative directions implies a longer processing time than the use of absolute directions. In this paper, have been used absolute directions for data processing on the lattice and relative directions for generating a new direction required by the steps of the backtracking algorithm. In this way, the search space used has the size of 3^{n-2} and the processing time decreases.

3.2.2. *The parallelization of the backtracking algorithm.* Because of from each node are generated three other nodes, the searching process can be easily parallelized using a number of threads (processors, respectively) powers of 3: 3, 3^2, 3^3, 3^4, 3^5, etc. In this paper the experiments ran with 3 and 9 threads.

4. RESULTS

In this paper an parallel backtracking algorithm was applied for following sequences showed in Table 2. The sequence by 22 aa is k-hefutoxin1, a real

No of aa	Sequence
20	HPHPPHHPHPPHPHHPPHPH
22	PPHPHPPPHPPPPPPPPPPPPP
24	HHPP HPPH PPHP PHPP HPPH PPHH
25	PPHP PHHP PPPH HPPP PHHP PPPHH

TABLE 2. Benchmark

protein which is found in scorpion venom[9], translated from primary structure in HP string in this work. She is a short protein, consisting of a chain which has 22 aa, with a molecular weight of 2664.91[6]. The software was written in Java 11 and the experiments were run on a computer with the following configuration:

- CPU: Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz, 4 cores, 8 logical processors
- HDD: 1 TB
- RAM: 16 GB
- GPU: NVIDIA GeForce GTX 1060, 6 Gb RAM
- OS: Windows 10 Professional 64-bit.

For 3 threads were obtained following conformation:

- for 20 aa sequence:
  1-R-D: 27,239,226 feasible conformation
  1-R-U: 27,239,226 feasible conformation
  1-R-R: 27,239,226 feasible conformation
- for 21 aa sequence:
  1-R-D: 72.988.592 feasible conformation
  1-R-U: 72.988.592 feasible conformation
  1-R-R: 78.447.107 feasible conformation
- for 22 aa sequence (k-hefutoxin1):
  1-R-D: 195.839.752 feasible conformation
  1-R-U: 195.839.752 feasible conformation
  1-R-R: 210.522.003 feasible conformation

The fact that the number of feasible conformations found for branches "1-R-D" and "1-R-U" are equal, suggests that these two are symmetrical and, consequently, searching in one of these is useless. For the sequence of 20 aa, one of the optimal conformations is represented in Figure 2. The same conformation was found by the AI algorithms.

The red unit is Hydrophobic aa and the blue unit is Polar aa. It can be seen that H aa forms a hydrophobic core.

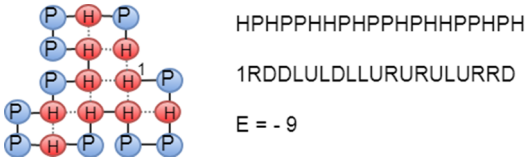

FIGURE 2. The optimum conformation for the 20 aa sequence

5. CONCLUSION

In this study, the backtracking algorithm was applied for the folding protein problem on the 2D HP Model. Even if the conformational space is very large, for short protein chains, it is applicable, because the processors power has increased. By applying of some conformation space reduction techniques and using relative directions were found optimum conformations for known proteins sequence and for kappa-hefutoxin1, a protein from scorpion venom.

REFERENCES

[1] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.
[2] J. M. Bahi, N. Cote, and C. Guyeux. Chaos of protein folding. *Neural Networks (IJCNN)*, pages 1948–1954, 2011.
[3] B. Berger and T. Leight. Protein folding in the hydrophobichydrophilic (hp) model is np-complete. *Journal of Computational Biology*, 5(1):27–40, 1998.
[4] V. Cristea and et al. *Tehnici de programare*. Ed Teora, București, 1998.
[5] V. Dinu, E. Trutia, E. Popa-Cristea, and A. Popescu. *Biochimie medicală – mic tratat*. Ed. Medicală, București, 2006.
[6] k-hefutoxin. https://www.rcsb.org/structure/1HP9. Accessed: 2018-09-25.
[7] K. F. Lau and K. A. Dill. A lattice statistical mechanics model of the conformation and sequence space of proteins. *Macromolecules*, 22:3986–3997, 1989.
[8] C. Levinthal. How to fold graciously. *Mossbauer Spectroscopy in Biological Systems Proceedings*, pages 22–24, 1969.
[9] K. N. Srinivasan and et al. K-hefutoxin1 - a novel toxin from the scorpion heterometrus fulvipes with unique structure and function. *The Journal of Biological Chemistry*, 277(33):30040–30047, 2002.

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE, BABEŞ-BOLYAI UNIVERSITY, CLUJ-NAPOCA, ROMANIA
Email address: sima.ioan@cs.ubbcluj.ro