Research on protein folding of cannabinoid receptor type II based on HP model

Zhengfu Li¹, Fan Liang¹, Hongyan Zhuang¹, Jiaojiao Li²*, Cunhua Li¹

¹ School of Computer Engineering, Huaihai Institute of Technology, Lianyungang 222005, China
² School of Pharmacy, Huaihai Institute of Technology, Lianyungang 222005, China
2778527252@qq.com

Abstract. Cannabinoid receptor type II (CB2 for short) is a subtype of cannabinoid receptor. At present, the three-dimensional conformation of CB2 protein has not been determined. In this paper, a protein folding method of CB2 based on HP model and genetic algorithm are proposed, which can help to study the folding mechanism and pharmacological properties of CB2. By using the backtracking mechanism, the genetic algorithm, which can help to solve this optimization problem, can help to avoid local optimization. The folding results obtained by the algorithm are analysed and compared. The experimental results demonstrate that the CB2 conformation obtained by genetic algorithm has a lower energy value, which is a good basis for further research on CB2 folding mechanism.

1. Introduction

The prediction problem of protein folding [1] is to solve the three-dimensional spatial structure problem of protein directly through sequence in the case of known protein amino acid sequence, and then calculate the minimum energy value of protein. Structural biology's studies have already verified that the spatial structure of proteins determines the specific functions of proteins [2]. Protein folding has become one of the key issues in bioinformatics, whose solution is an important task in post-genomic protein research [3]. Although the three-dimensional conformation of some amino acids can be obtained by special technical methods, it is extremely time-consuming and more restrictions are needed. So far, most proteins have no actual three-dimensional concept, and thus it is an urgent need to predict the three-dimensional conformation of protein amino acid sequences by theoretical methods.

The protein folding prediction model based on HP grid [4] can effectively explore and explain the way of protein folding by gradually reducing the amino acid position of fixed protein and making continuous sequence discretization. Although protein folding based on the HP model is already a simplified prediction model that removes many constraints, whereas it is still a NP hard problem.

Cannabinoid receptor type II (CB2) is a subtype of cannabinoid receptor with 360 amino acids. Amino acid sequence analysis showed that the structure of CB2 receptor includes seven times lipophilic transmembrane α helical structures [5], which is a typical G-protein coupled receptor. The distribution and expression of its receptors lead to the unique physiological and pharmacological effects. In addition, CB2 receptors are also associated with a large number of physiological phenomena and diseases, such as embryonic development, atherosclerosis, pancreatitis, urethral infarction, arthritis, chronic cough and so forth. At present, the spatial conformation of CB2 protein...
has not been determined, and there are few studies on CB2 folding. In the paper, some work has been carried out to explore this problem. There are two types of CB2 receptors, named CB2A and CB2B [6].

2. Materials and methods

2.1. Protein folding

The primary objective of protein folding is to infer the protein-folding pattern in space based on the amino acid sequence of protein, and thus the conformation of proteins with specific functions in three-dimensional space was determined. In Anfinsen’s renaturation experiments on bovine pancreatic ribonuclease, it was found that the amino acid sequence of a protein uniquely determined its conformation in three-dimensional space, and a famous thermodynamic hypothesis was proposed, that is, the minimum conformation of the free energy of the protein, namely its natural conformation. Current theories and methods of protein folding structure prediction are based on this hypothesis. The study of protein folding is still very difficult in computer simulation and experimental research [7].

2.2. HP folding model

Hydrophobic interaction within proteins is the main factor that promotes protein folding. The folded structure of spherical protein is usually deposited by hydrophobic nucleus, forming a specific spatial structure [8]. Based on the HP model, the model is more focused on computer simulation research, because it only studies the protein folding in two-dimensional space, and ensures the effective rate of the solution in the case of relative folding accuracy.

In the HP model, the amino acids in the protein sequence are divided into two kinds of amino acids: hydrophobic amino acids (H) and hydrophilic amino acids (P). Amino acids in protein sequences can be expressed in letters H and P respectively, in order to form a simplified form of protein amino acid sequences. Based on the HP folding model, the problem of protein folding can be simplified as: two-dimensional space is divided into equidistant grid space, and each amino acid is simplified as the node placed in the grid. The spatial conformation of legitimate protein sequences needs to meet the following three conditions simultaneously [9]:

1) Any amino acid (hydrophobic amino acids or hydrophilic amino acids) must be placed on the coordinates of the integer points in two-dimension space;
2) The adjacent nodes in the chain sequence are still adjacent after placement (distance is 1);
3) No more than one amino acid node can be placed on two-dimensional coordinates.

In the protein-folding model based on HP model, the energy function corresponding to its amino acid sequence can be expressed as:

\[ E = \sum_{i<j} E_{ij} \Delta (r_i - r_j) \]

When the node \( r_i \) on the amino acid sequence is the same as its previous node \( r_1 \), and they are both H type amino acid. They are not connected in the first order sequence, but \( r_i \) is adjacent to \( r_j \) in the two-dimensional lattice space, then \( E_{ij} \Delta (r_i - r_j) = -1 \); else \( E_{ij} \Delta (r_i - r_j) = 0 \). The energy value of a two-dimensional conformation can be obtained simply by representing the protein sequence in the HP model.

2.3. Genetic algorithm and Monte Carlo

Genetic algorithm is a random search method derived from the evolutionary rule of "survival of the fittest" followed by biological evolution. It is mainly characterized by direct manipulation of structural objects, free from the restrictions of derivative and function continuity. It can automatically obtain and guide the optimal search space, and adaptively adjust the search direction without certain rules by using the probabilistic optimization method. These characteristics of genetic algorithms have been widely used in the fields of combinatorial optimization, machine learning, signal processing, and adaptive control artificial life. Genetic algorithm is one of the key techniques in modern intelligent
computing. It is a search algorithm with the iterative process of "survival plus detection". Genetic algorithms take all individuals in a population as objects and use random techniques to guide efficient searching of encoded parameter spaces. Among them, the genetic algorithm consists of selection, crossover and mutation. It has the characteristics of simple universality, strong robustness, fast parallel processing, high efficiency and strong practicability. It can get good results and gradually become one of the important intelligent algorithms.

Monte Carlo method is a random simulation method, also known as statistical simulation or random sampling. It is a method based on probability theory and mathematical statistics to solve specific calculation problems by generating random numbers. Probabilistic models are often used to describe problems that need to be solved. After obtaining the probability model of the same solution of the problem, the probability model is simulated by a random experiment. The final statistical value obtained by experiment is an approximate feasible solution to this problem.

3. Experimental method

3.1. Lattice position of CB2 amino acid sequence
When the space grid position of the CB2 amino acid sequence is created, planar space information is empty, and the position of the first amino acid was first fixed in the centre of the grid space. Then, the placement of the following amino acids is divided into four kinds of circumstances: placed on the upward side of the last amino acid, placed on the right side of the last amino acid, placed on the bottom side of the last amino acid, placed on the left side of the last amino acid. When amino acids are placed in the appropriate location, they are randomly generated except for the first amino acid. Through a continuous cycle, all the amino acids are placed in order as required and the position information of the amino acids are placed in a specific array (when the amino acid move upward a position, put 1 in the array; when the amino acid move one position to the right, store 2 in the array; when the amino acid move down a position, the array stores 3; when the amino acid move one position to the left, the array is stored in 4). When all the amino acids are laid out, a legal protein placement sequence is formed.

3.2. Backtracking mechanism
When amino acid sequences are placed sequentially, the amino acid may overlap with the previous placed amino acids, resulting in the formation of illegal amino acid sequences. As the number of amino acids increases, so does the probability of conflict. In order to reduce the occurrence of this situation, this paper adopts the backtracking mechanism [10]. When the overlay of amino acids occurs, the conflicting amino acids cancel this placement, and the amino acid sequence will return to the position of the previous amino acid, record the position that cannot be placed, and re-select the position to place, remove the position that cannot be placed, and if there is still one position left, place the next amino acid: if there are two remaining positions, one of them will be randomly selected to place the next amino acid; if there are three remaining positions, one of them will be randomly selected to place the next amino acid; if there is no position, the current amino acid placement will be cancelled and go back to the position of the front amino acid and repeat the above steps.

Backtracking mechanism is a better way to solve conflicts, but if the frequency of backtracking is not limited, in extreme cases, it will degrade to the initial amino acid sequence, which will greatly limit the efficiency of the algorithm. Configuration of experimental machines: CPU, Intel (R) Core (TM) i5 CPU M 460 @ 2.53GHz; RAM, 4.00GB; 64 bit operating system. For the backtracking times of 10,000 random sequence and average energy value, the experimental results of the algorithm are shown in Tab.1:

| Tab.1 Comparison of backtracking times |
|---------------------------------------|
| backtracking time | average energy value | average calculating time(ms) |
| 10               | -69.5               | 30941.6                      |
As the number of backtracking times increases, the average energy decreases at the beginning and then increases. When the backtracking time is 50, the algorithm gets the best average energy value.

### 3.3. Energy assessment of sequence placement

Method of evaluating the energy value of amino acid sequence: Starting from the first amino acid, when the amino acid is not H-type amino acid (hydrophobic amino acid), continue to determine whether the next amino acid is H-type amino acid. When the amino acid is H-type amino acid (hydrophobic amino acid), the position of this amino acid is fixed. The following amino acid according to the array storing the putting position of amino acids is placed behind this amino acid, and all the amino acids behind it (two amino acids are not connected) are judged weather H-type amino acid or not. When the latter amino acid is H-type amino acid, the distance between the next amino acid and the first fixed amino acid is determined. If the coordinates of the two amino acids on the X-axis are the same, the difference between the coordinates on the Y-axis is 1, or if the coordinates of the two amino acids on the Y-axis are the same, the difference between the coordinates on the X-axis is 1, then energy 1 (this is a unit of energy value) is obtained. By traversing all the amino acids in the sequence, the corresponding energy of CB2 under a certain emission mode can be obtained.

### 3.4. Genetic algorithms optimize sequences

Starting from the second sequence, the cumulative fitness is equal to the relative fitness of the sequence plus the cumulative fitness of the previous sequence. Finally, for each sequence, a random number from 0 to 1 is generated. If the random number is less than the cumulative fitness of the first sequence, the amino acid sequence of the first sequence is inherited to the next generation. Otherwise, the random number is compared with the cumulative fitness of the sequence itself and the cumulative fitness of the subsequent sequence. If the random number is greater than or equal to the cumulative fitness of the sequence itself and less than the cumulative fitness of the latter sequence, the amino acid sequence of the latter sequence is inherited to the next generation of the sequence. If the random number is not greater than or equal to that of the sequence itself and is not less than that of the following sequence, the amino acid sequence of the sequence itself is directly passed on to the next generation of the sequence. Through the above method, the amino acid sequence with good quantifiable value in each generation was transmitted to the next generation.

In the case of cross-inheritance of two sequences, the position of crossed nodes in the sequence is determined by the number randomly generated, that is, the position of crossed nodes is randomly get. The mutation rules 1-4 refers to the method in literature [3]. In order to ensure that the sequence in which the node is located is still legal after mutation, all nodes after mutation are rearranged. In order to ensure that all nodes after mutation are placed legally, a backtracking mechanism is also added.
Rule 1: When there is a diagonal vacancy in the mutant amino acid, it can be moved to the diagonal vacancy.

Rule 2: When the four consecutive amino acids in which the mutated amino acids are located form a u-shape, with the two endpoints as the axis, and the symmetrical sites of the other two points have vacancies, the two points can be moved to the corresponding vacancies.

Rule 3: When nodes C and L are empty, the mutated amino acids are moved from vertex i to node L and node i-1 to node C.

Rule 4: When nodes C and L are empty, the mutated amino acids are first moved from vertex i to node L, from node i-1 to node C, from node i-2 to original node i, and from node i-3 to original node i-1.

When there is vacancy around the mutant amino acid, move it to any vacancy around it, and then the amino acids are rearranged. Otherwise, look for the next amino acid, see if there is a vacancy around the next amino acid, until you find the amino acid around the vacancy, and then carry out the above mutation of this amino acid.

3.5. Elitist policy
Through the challenge algorithm, the sequence with the smallest amount of energy in each generation can be obtained and recorded. By this way, the algorithm can ensure that the population generation is superior to the generation.

4. Results and discussion
The HP folding model of cannabinoid receptor CB2 was expressed by genetic algorithm and Monte Carlo method. The spatial folding conformational arrangement expression of better results obtained by these methods is shown in Fig.2.

In Fig.2, solid points were used to represent h-type amino acids, hollow points were used to
represent p-type amino acids, and bold solid points were used to indicate energy accumulation. As we can be seen intuitively from the figure, the closer the nodes are arranged, the smaller the energy value of amino acids is generally. Conversely, the higher the energy value is. By comparing the results of genetic algorithm and simple Monte Carlo method, it can be found from figure 2 that the results of genetic algorithm optimization are arranged more closely. The result of genetic algorithm is better than that of Monte Carlo method, and the folded energy is lower.

5. Conclusion
In this paper, HP folding prediction model was used to study the folding condition of cannabinoid receptor subtype CB2 protein, and simple Monte Carlo method and genetic algorithm were used to solve the global optimization. By comparing the two methods, the genetic algorithm obtained better folding and arrangement results. In order to ensure the success of each mutation, this paper proposes a variation backtracking mechanism. In order to ensure the correctness after mutation, the rationality of placement of all nodes after variation will be re-examined. This mutation rule can guarantee the stability of mutation operator.

Although the genetic algorithm used in this paper has the ability of global optimization, it also has the problem of slow convergence speed. Moreover, proteins are folded in three-dimensional space in real life, so the study of protein folding in two-dimensional space has limitations. In the next work, the sequential arrangement method will be further optimized and CB2 will be folded in three-dimensional space in order to obtain more accurate conformation energy of CB2.

Acknowledgements
This work was financially supported by Natural Science Foundation of Jiangsu Higher Education Institutions of China (No. 17KJD520002, 17KJB350002), Science Foundation of Huaihai Institute of Technology (No. Z2015012), Jiangsu Key Laboratory of Marine Pharmaceutical Compound Screening (No. HY2015B03).

References
[1] Munoz V, Campos L A, Sadqi M. Limited cooperativity in protein folding [J]. Current Opinion in Structural Biology, 2016, 36:58-66.
[2] Mirceva G, Davcev D. Incorporating several features in the protein ray descriptor for more accurate protein 3D structure retrieval[C]. ACM Workshop on 3d Object Retrieval. ACM, 2010:51-56.
[3] SUN Pengfei,ZHANG Jianpei.A method for predicting protein folding based on a quantum genetic algorithm[J].Journal of Harbin Engineering University,2010,31(1).
[4] LU Zhipeng,HUANG Wenqi.Local Search Algorithm for Solving Protein Structure Prediction Problem[J]. Computer Science.2005, 32(11):148-151.
[5] Lagneux, C.; Lamontagne, D.; Involvement of cannabinoids in the cardioprotection induced by lipopolysaccharide. Br. J. Pharmacol. 2001, 132, 793-796.
[6] Thompson, M. D.; Cole, D. E. C.; et al. Pharmacogenetics of the G Protein-Coupled Receptors. Methods Mol. Biol. 2014, 1175, 189-242.
[7] SHI Xiaohong.Research on protein folding based on HP model[J].Chinese Journal of Bioinformatics.2016, 14(2):112-116.
[8] LIU Yun,WANG Cunxin,WANG Baohan, et al. A Protein Design Procedure Based on The Lattice Model[J]. Prog. Biochem. Biophys.2004, 31(2):172-176.
[9] HUANG Wenqi,LI Zongmin.Solutions for Protein Folding Problem Based on Simulated Annealing Algorithm[J].Computer Engineering and Applications.2005,41(7):40-41.
[10] Johnson C M, Katiikreddy A. A genetic algorithm with backtracking for protein structure prediction[C]. Conference on Genetic and Evolutionary Computation. ACM, 2006:299-300.