Annealing genetic algorithm for protein folding simulations in the 3D HP model

Jingfa Liu, Jun Wang, Yu Zheng, Yonglei Yao, Zhaoxia Liu

1Network Information Center, Nanjing University of Information Science & Technology, Nanjing 210044, China; School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China; jfliu@nuist.edu.cn

2Network Information Center, Nanjing University of Information Science & Technology, Nanjing 210044, China; wangjun@nuist.edu.cn, zaochenhua@163.com

3School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China; Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044, China; yzheng@nuist.edu.cn, ylyao@nuist.edu.cn

Abstract
The protein folding problem, i.e., the prediction of the tertiary structures of protein molecules from their amino acid sequences is one of the most important problems in computational biology. This problem has been widely studied under the HP model in which each amino acid is classified, based on its hydrophobicity, as a hydrophobic (H) residue or a hydrophilic (or polar, P) one. The protein folding problem in the HP model is in fact to find conformations with the lowest energies for some benchmark amino acid sequences. A genetic algorithm (GA) is used to find the lowest energy conformation in this paper. Each time for a newly produced offspring individual, which is originated from the selection, crossover and mutation operator of two parent individuals, we adopt the new acceptance criteria based on the annealing strategy to let it pass into the next generation, and propose a so-called annealing genetic algorithm (aGA) to predict efficiently the protein folding conformations in the three-dimensional (3D) HP model. Eleven benchmarks are tested to verify the effectiveness of the proposed approach and the computational results show that aGA explores the conformation surfaces more efficiently than other methods, and finds new lower energies in several cases, which means that aGA is an efficient tool for the protein folding simulations.

Keywords: protein folding, HP model, genetic algorithm, annealing strategy

1. Introduction
One of the most important problems in computational biology is to predict the three-dimensional (3D) folding conformations (“tertiary structures”) of protein molecules from their primary structures: the one-dimensional sequences of amino acid residues. The protein folding problem is essentially to search for the biologically active (functional) conformations (the so-called native states) of proteins for the given sequences of amino acid residues. But design of appropriate energy functions which can generally distinguish the native states from the non-native states of protein molecules and the extremely difficult optimization problems which arise from energy functions used in folding simulations prevent the solving of the folding problem.

To address this problem, it is a common practice to use models that simplify the search space of possible conformations. These models try to generally reflect different global characteristics of protein structures. One of the most popular models of protein folding is the so-called hydrophobic-hydrophilic (or hydrophobic-polar, HP) lattice model [1,2], where only two types of residues, hydrophobic (H) and hydrophilic (or polar, P) ones, are considered. The free energy of a conformation is defined as the negative number of non-consecutive hydrophobic-hydrophobic contacts, where a contact denotes two non-consecutive monomers in the chain occupying adjacent sites in the lattice.

In spite of its apparent simplicity, finding the conformation with the lowest energy of the HP model on a cubic lattice remain NP-complete [3], i.e., no approach is able to gain the global optimal solution of the problem in polynomial time. So, a wide variety of approximate computational methods are
generally proposed to simulate and analyze this model, including genetic algorithm (GA) [4,5] and its variations (the modified genetic algorithm (mGA) with a different selection scheme and multiple-point crossover [6] and the improved genetic algorithm (iGA) with the number limit of clonal individual, brood selection and a local search [7]), the Monte Carlo (MC) method [5], the elastic net (EN) algorithm with a local search [8], the pruned enriched Rosenluth method (PERM) [9] and its variations [10,11], the simulated annealing algorithm (SA) [12], the ant colony optimization (ACO) [13], the particle swarm optimization (PSO) [14], the artificial immune system (AIS) [2], and others.

In this paper, we put forward a new folding algorithm, the so-called annealing genetic algorithm (aGA), which combines the genetic algorithm with an annealing strategy, to predict proteins’ structures on 3D HP model. The computational results, on eleven 27-mer sequences, show that aGA is an effective approach for protein folding simulations.

2. The 3D HP model

The 3D HP model, which was proposed by Dill [1], is a free-energy model based on the belief that interactions between hydrophobic amino acids greatly contribute to the free energy of the natural conformation of a protein. In 3D HP model, 20 kinds of amino acids are divided into two classes according to their hydrophobicity: hydrophobic (H) and hydrophilic (or polar, P) residues. Based on this abstraction, a protein sequence can be regarded as a string with binary characters, H and P, and we define \( s = s_1 s_2 \ldots s_n \) as a protein chain with \( n \) amino acids, where \( s_i \) is H if the \( i \)th amino acid in the sequence is hydrophobic and P if it is hydrophilic. On the other hand, a protein sequence will be arranged as a 3D self-avoiding walk chain, where adjacent residues in the sequence occupy adjacent grid points and no grid point in the lattice is occupied by more than one residue. Two amino acids are topological adjacent if they are neighbors in the lattice but are not adjacent in sequence. A topological \( H-H \) bond is formed between two topological adjacent hydrophobic amino acids. The values of the \( H-H, H-P, \) and \( P-P \) interactions (\( \epsilon_{ij} \)) in the 3D HP model [1] are

\[
\epsilon_{HH} = -1.0, \quad \epsilon_{HP} = 0, \quad \epsilon_{PP} = 0. 
\]

The free energy of a conformation \( s \) is obtained by summing over these local interactions as follows:

\[
E(s) = \sum_{i=1}^{n} \sum_{j>i} \epsilon_{ij} \Delta r_{ij},
\]

where

\[
\Delta r_{ij} = \begin{cases} 1 & \text{if } s_i \text{ and } s_j \text{ are topological adjacent} \\ 0 & \text{otherwise} \end{cases}
\]

The HP lattice protein folding problem can be formally defined as follows: given an amino acid sequence \( s = s_1 s_2 \ldots s_n \), we try to find an energy-minimizing conformation of \( s \), that is, to find \( c^* \in T(s) \) so that \( E(c^*) = \min\{E(c) \mid c \in T(s)\} \), where \( T(s) \) is the set of all the valid conformations of \( s \). So the protein folding problem in 3D HP model is transformed into the optimization calculation of the minimal free energy of the protein folding conformations.

3. The annealing genetic algorithm

Genetic algorithm (GA) [15,16] was proposed by Holland in 1975 and was applied to the protein folding problem by Unger and Moult in 1993 [4,5]. GA imitates the process of the biological evolution. First, some individuals are generated randomly. For the 3D HP model, an individual represents a protein conformation and different individuals possess different conformations. The population includes a fixed number of individuals. Then, all of the individuals are evaluated by the fitness (in this paper we set the free energy of the conformation to be the fitness). The smaller the fitness is, the better the individual will be and the greater the possibilities to survive will be. Next, the operators of selection, crossover, and mutation are used for generating of the individuals in the next generation. In this paper, two parent individuals are chosen each time, and one new offspring conformation is produced by using a crossover operator, and maybe accompanying by mutation operator to make the new conformation valid. For the newly produced offspring conformation, the new acceptance criteria based on the annealing strategy is introduced. The crossover operation is...
repeated on \( n-1 \) pair of parent conformations until \( n-1 \) new offspring are created. The parent conformation with the lowest energy is copied to the next generation as offspring individual \( n \). After enough iterations (in this paper we set the iterative step number to be 1000), an optimal or approximate optimal solution will be obtained. The main flowchart of aGA is shown in Figure 1.

3.1. Generating initial individuals

In the 3D HP model, Cartesian coordinates are used for describing the 3D spatial positions of amino acids. For a given protein sequence with \( n \) amino acids, the first two amino acids are fixed at \((0, 0, 0)\) and \((1, 0, 0)\), respectively. Obviously, in the process of generating initial individuals, if more than one residue occupies a same point in the lattice, this protein conformation is not permitted. To eliminate the generation of such invalid protein conformations, for the other \( n-2 \) amino acids, we adopt a “recoil growth” algorithm [17,18], which involves growing the chain one residue at a time, checking the validity of the incomplete conformation at each step, and backtracking when an invalid sub-conformation is generated. Thus, a number of valid initial individuals (in this paper we set the initial population size to be 100 for each sequence) are generated at random.

3.2. Selection

Selection refers to the way in which individual members of the parent population may be chosen to pass into the offspring. We adopt the roulette wheel selection, i.e., the selected probability \( P_i \) of the \( i \)th individual is proportional to the absolute value of its free energy:

\[
P_i = \frac{E_i}{\sum_{j=1}^{m} E_j}
\]

where \( E_i = \sum_{j=1}^{n} e_j \Delta \alpha_j \) and \( m=100 \) is the population size. By equation (4), each time we select two good individuals \( s_l \) and \( s_r \) from parent population to execute the crossover operator.

3.3. Crossover

Crossover is the way in which the genetic information from two parent individuals is combined to generate offspring individuals. In this paper, we use one-point crossover, where the two parent conformations are cut at one point and their complementary portions are exchanged to produce the offspring conformations.

As an example, we show the case of one-point crossover at location \( s^{(i)} \) \((1 \leq i \leq n-1)\) as follows:

\[
\begin{align*}
\left( s_1^{(1)}, s_2^{(1)}, \ldots, s_i^{(1)}, s_{i+1}^{(1)}, \ldots, s_{n-1}^{(1)} \right) & \Rightarrow \left( s_1^{(2)}, s_2^{(2)}, \ldots, s_i^{(2)}, s_{i+1}^{(2)}, \ldots, s_{n-1}^{(2)} \right) \\
\left( s_1^{(1)}, s_2^{(2)}, \ldots, s_i^{(2)}, s_{i+1}^{(1)}, \ldots, s_{n-1}^{(1)} \right) & \Rightarrow \left( s_1^{(2)}, s_2^{(2)}, \ldots, s_i^{(1)}, s_{i+1}^{(1)}, \ldots, s_{n-1}^{(1)} \right)
\end{align*}
\]

In order to connect these two parts, a random direction is chosen for the gluing \((90^\circ \text{ left, } 90^\circ \text{ right, } 90^\circ \text{ up, } 90^\circ \text{ down, or continue ahead})\). The offspring are checked for self-avoiding walk. We choose the valid offspring conformation with the lowest energy as the new offspring conformation. If all 5 random gluing directions lead to colliding structures, we execute the following mutation operator.

3.4. Mutation

In mutation, we choose the first collided amino acid of a conformation with the lowest energy as a pivot, and the subsequent residues of the chain as a rigid body, which are rotated around the pivot. The rotation can either be \(90^\circ \text{ right, } 90^\circ \text{ left, or } 180^\circ \) along one of the axes planes \(XY, XZ, \text{ and } YZ\). Once we produce a self-avoiding offspring conformation \( s_m \), its energy \( E_m \) will be evaluated. In case none of all rotations leads to a valid conformation, another non-self-avoiding conformation from crossover operator is selected for mutation and the process is repeated until a valid conformation is found.
Figure 1. The main flowchart of the annealing genetic algorithm

3.5. Acceptance criteria based on annealing strategy

For newly produced valid offspring conformation $s_m$, its energy $E_m$ is computed and compared to the average energy value of its two parent conformations ($\bar{E} = \frac{(E_x + E_y)}{2}$). The conformation $s_m$ will be accepted if $E_m \leq \bar{E}$. If $E_m \geq \bar{E}$, the conformation is still accepted if

\[ \text{random} \left(0, 1\right) < \exp \left(\frac{\bar{E} - E_m}{T_k}\right) \]

where random(0, 1) is a random number 0 and 1, and $T_k$ is gradually decreased (cooled) during the simulations. In our simulations, we set $T_0=100$, and $T_k=0.9T_{k-1}$ until $T_k<0.01$ for each iterative step.
4. Computational results and analysis

We implement the annealing genetic algorithm (aGA) in Java language and run it on a Notebook PC with Intel Core i5, 2.4GHz processor and 4.0 GB of RAM. We test a set of instances from [5-8], consisting of eleven 27-mer sequences, which is listed in Table 1. We will focus ourselves on finding the lowest-energy states for this set of 3D proteins, and compare aGA’s performance with other methods.

Table 1. Eleven 27-mer sequences

<table>
<thead>
<tr>
<th>No.</th>
<th>HP sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PHPHPIHHPPPIHHHHPPP</td>
</tr>
<tr>
<td>2</td>
<td>PHPHPPPPPPPPPPHIHPPHHHP</td>
</tr>
<tr>
<td>3</td>
<td>HHHHPPPPPIHHPPHHPPHPPPPP</td>
</tr>
<tr>
<td>4</td>
<td>HHHHPPPPIHHHPPHHPPHPPPPP</td>
</tr>
<tr>
<td>5</td>
<td>HHHHPPPPPPPIHHPPHHPPPPP</td>
</tr>
<tr>
<td>6</td>
<td>HHHPPPPPPPPPIHHPPHHPPPPPPPP</td>
</tr>
<tr>
<td>7</td>
<td>HHHPPPPPPPPPIHHPPPPPPP</td>
</tr>
<tr>
<td>8</td>
<td>HHHPPPPPPPPPIHHPPPPPPP</td>
</tr>
<tr>
<td>9</td>
<td>HHHPPPPPPPPPIHHPPPPPPP</td>
</tr>
<tr>
<td>10</td>
<td>HHHPPPPPPPPPIHHPPPPPPP</td>
</tr>
<tr>
<td>11</td>
<td>HHHPPPPPPPPPIHHPPPPPPP</td>
</tr>
</tbody>
</table>

Table 2. Comparison of the results of different methods for the 3D HP protein folding problem. NA means data not available. The numbers in parentheses are the numbers of energy evaluation (i.e. the numbers of valid conformations scanned) before the lowest-energy values are found.

<table>
<thead>
<tr>
<th>No.</th>
<th>GAa</th>
<th>MCa</th>
<th>mGA°</th>
<th>iGA°</th>
<th>ENb</th>
<th>aGA°</th>
<th>Time(s)f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-9(1227964)</td>
<td>-7(1921887)</td>
<td>-8</td>
<td>-9(22354)</td>
<td>-9</td>
<td>-9(11688)</td>
<td>1.28</td>
</tr>
<tr>
<td>2</td>
<td>-9(1225281)</td>
<td>-9(1922018)</td>
<td>-10</td>
<td>-10(61300)</td>
<td>-10</td>
<td>-10(11167)</td>
<td>1.18</td>
</tr>
<tr>
<td>3</td>
<td>-8(1247208)</td>
<td>-6(1932764)</td>
<td>NA</td>
<td>-8(9756)</td>
<td>-8</td>
<td>-8(3116)</td>
<td>0.58</td>
</tr>
<tr>
<td>4</td>
<td>-15(1207688)</td>
<td>-11(1889807)</td>
<td>-15</td>
<td>-15(95231)</td>
<td>-15</td>
<td>-15(15460)</td>
<td>1.74</td>
</tr>
<tr>
<td>5</td>
<td>-8(1118202)</td>
<td>-7(1923899)</td>
<td>-8</td>
<td>-8(7236)</td>
<td>-8</td>
<td>-8(183)</td>
<td>0.18</td>
</tr>
<tr>
<td>6</td>
<td>-11(1226090)</td>
<td>-9(1913399)</td>
<td>NA</td>
<td>-11(34517)</td>
<td>-11</td>
<td>-12(30552)</td>
<td>2.06</td>
</tr>
<tr>
<td>7</td>
<td>-12(1239519)</td>
<td>-10(1904467)</td>
<td>-13</td>
<td>-13(65398)</td>
<td>-13</td>
<td>-13(27063)</td>
<td>1.91</td>
</tr>
<tr>
<td>8</td>
<td>-4(1248118)</td>
<td>-4(1941168)</td>
<td>-4</td>
<td>-4(2103)</td>
<td>-4</td>
<td>-4(1254)</td>
<td>0.21</td>
</tr>
<tr>
<td>9</td>
<td>-7(1198945)</td>
<td>-6(1923287)</td>
<td>-7</td>
<td>-7(6357)</td>
<td>-7</td>
<td>-7(678)</td>
<td>0.36</td>
</tr>
<tr>
<td>10</td>
<td>-11(1174297)</td>
<td>-9(1900935)</td>
<td>NA</td>
<td>-11(14682)</td>
<td>-11</td>
<td>-11(6056)</td>
<td>0.92</td>
</tr>
<tr>
<td>11</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>-14</td>
<td>-16</td>
<td>(45176)</td>
<td>2.49</td>
</tr>
</tbody>
</table>

a Values are from Ref. [5].
b Values are from Ref. [6].
° Values are from Ref. [7].
c Values are from Ref. [8].
 Values are from the present work.
 Average CPU times (seconds) needed to find the lowest-energy values using the annealing genetic algorithm (aGA) in this study.

For each instance, aGA is run five times independently. In each of five independent runs, the iterative step number is 1000 and the population size is 100, and aGA can find the lowest-energy conformation in each run. Table 2 summarizes aGA’s performance together with those of other methods, including the genetic algorithm (GA) [5] and its variations (the modified genetic algorithm (mGA) with a different selection scheme and multiple-point crossover [6] and the improved genetic algorithm (iGA) with the number limit of clonal individual, brood selection and a local search [7]), the Monte Carlo (MC) method [5], and the elastic net (EN) algorithm with local search [8]. As seen from
Table 2, the results of our algorithm are as good as or better than those of the other five algorithms. For the sequences 6 and 11, it is noteworthy that we find lower energies of -12 and -16, respectively, which are missed by the previous algorithms. For the other nine sequences, the putative native state energies obtained by aGA, iGA and EN are the same. This means that aGA explores the conformation surfaces more efficiently than GA, MC, mGA, EN, and iGA. Moreover, the aGA method scans fewer valid conformations than the MC method [5], the GA method [5], and iGA [7] to obtain the lowest free energy for every sequence. Because the other two algorithms mGA and EN do not report the CPU times needed to find the lowest energies, we cannot compare the times of our algorithm with the other two methods. Figures 2-12 show typical lowest-energy conformations for these eleven sequences.

Figure 2. Typical conformation with an energy of $E_{=-9}$ of sequence 1 found by aGA.

Figure 3. Typical conformation with an energy of $E_{=-10}$ of sequence 2 found by aGA.

Figure 4. Typical conformation with an energy of $E_{=-8}$ of sequence 3 found by aGA.

Figure 5. Typical conformation with an energy of $E_{=-15}$ of sequence 4 found by aGA.

Figure 6. Typical conformation with an energy of $E_{=-8}$ of sequence 5 found by aGA.

Figure 7. Typical conformation with an energy of $E_{=-12}$ of sequence 6 found by aGA.

Figure 8. Typical conformation with an energy of $E_{=-13}$ of sequence 7 found by aGA.

Figure 9. Typical conformation with an energy of $E_{=-4}$ of sequence 8 found by aGA.
Figure 10. Typical conformation with an energy of $E=-7$ of sequence 9 found by aGA.

Figure 11. Typical conformation with an energy of $E=-11$ of sequence 10 found by aGA.

Figure 12. Typical conformation with an energy of $E=-16$ of sequence 11 found by aGA.

5. Conclusions

In this paper, we put forward a new genetic algorithm, the so-called annealing genetic algorithm (aGA), to predict protein folding conformations in 3D HP model. In aGA, two parent individuals are chosen each time, and one new offspring conformation is produced by performing a crossover operator, and potential mutation operator on these two selected parent individuals. For the newly produced offspring conformation, the new acceptance criteria based on the annealing strategy is introduced. Our results on eleven 27-mer sequences, show the aGA method outperforms the original genetic algorithm [5] and its some improved versions [6,7], and other several methods. This could be explained by the fact that the annealing strategy can provide the individuals of the next generation diversity and prevent the genetic algorithm turning early-maturing. It is not hard to see that the proposed method is easy to extend to other protein models and we expect it to be useful for dealing with folding simulations of real proteins.

Acknowledgements

This work is supported by Special Foundation of China Postdoctoral Science Foundation (201104572), the Natural Science Foundation of Jiangsu Province (BK2010570), Jiangsu Planned Projects for Postdoctoral Research Funds (1001030B), the China Postdoctoral Science Foundation funded Project (20100471350), Natural Science Foundation of Education Committee of Jiangsu Province (09KJB520008), and Qing Lan Project.

References


