In: Proceedings of 3rd. Brazilian Workshop on Bioinfomatics, Brasília (DF), 20-22/october, [CD-ROM], 2004.

Protein Structure Prediction using an enhanced Genetic Algorithm for the 2D HP model

Marcos Paulo Scapin, Heitor Silvério Lopes

Laboratório de Bioinformática / CPGEI
Centro Federal de Educação Tecnológica do Paraná - CEFET-PR
Av. 7 de setembro, 3165 – 80230-901 Curitiba (PR) – Brazil
tel. +55-41-310-4694, fax. +55-41-310-4683
http://bioinfo.cpgei.cefetpr.br
e-mails: {mpscapin, hslopes}@cpgei.cefetpr.br

Abstract. This paper presents an enhanced genetic algorithm for the protein structure prediction problem. A new fitness function, that uses the concept of radius of gyration, is proposed. Also, a novel operator called partial optimization, together with different strategies for performance improvement, are described. Tests were done with five different amino acid chains from 20 to 85 residues long and better results were obtained, when compared with those in the current literature. Results are promising and suggest the suitability of the proposed method for protein structure prediction using the 2D HP model.

1 Introduction

A protein is a chain of amino acid residues that folds into a specific native tertiary structure under natural conditions. The task of predicting this tertiary structure is called the protein structure prediction (PSP) problem and its resolution is of great importance for modern molecular biology.

Although many other techniques have already been applied to this problem, heuristic optimization methods seem to be the most reasonable algorithmic choice to solve this problem due to its NP-completeness, and, amongst them, many evolutionary computation approaches have been proposed [1], [2], and [3].

The 2D HP model was introduced by [4] and is the most widely studied discrete model in recent literature. It models the concept that the major contribution to the free energy of the native conformation of a protein is due to interactions among hydrophobic residues, which tend to form a core in the protein structure while being surrounded by hydrophilic residues that interface the environment. In this model, the chain is embedded in a 2D square lattice and each lattice point can be occupied by only one residue. In a legal conformation, the adjacent residues in the sequence must be adjacent in the lattice. The free energy of a conformation is inversely proportional to the number of hydrophobic non-local bonds (or H–H contacts) where a H–H contact occurs if two hydrophobic residues occupy adjacent grid points in the lattice but are not consecutive in the sequence.

3 Implementation

The dynamics and effectiveness of a GA is strongly influenced by the way solutions are represented. Based on the results presented by [5], our implementation uses internal coordinates, where the definition of an amino acid position takes into account the position of the previous one, with relative movements where there are only three possible moves: (R)ight, (L)eft and (F)orward.

Our implementation uses a penalty method, which allows the existence of unfeasible conformations during the evolution, but a penalty is added for every lattice point at which there is a collision of more than one amino acid. Based on results of [6], a different strategy was used to create the initial population that tends to minimize the collisions while generating a greater initial genetic diversity. This strategy divides the population into two parts that are generated differently. The proportion of each part is established by a user-defined parameter called *PopIniFull*. The first part of the population (indicated by parameter *PopIniFull*) is randomly generated, as usual, and the second is generated considering each individual as totally unfolded, and then applying a number of random mutations that varies between 3 and the total number of genes in the chromosome, uniformly distributed.

To evaluate an individual it is necessary to translate its genotypical encoding so as to obtain its Cartesian coordinates. In this work, we propose a new fitness function composed of three terms, as shown in Equation 1:

$$Fitness = HnLB \times RgH \times RgP \ , \tag{1}$$

where *HnLB* is the number of hydrophobic non-local bonds of the conformation and *RgH* and *RgP* are terms computed using the radius of gyration of the hydrophobic and hydrophilic residues, respectively, as explained below.

It is believed that hydrophobic non-local bonds are the main force that drives the protein folding process. We are considering the problem as the maximization of the H–H contacts, thus for every hydrophobic non-local bond, *HnLB* is added by 1. Since we are using a penalty method, *HnLB* is decreased by a penalty term whenever a collision occurs. The penalty term is composed by the number of grid points which are occupied by more than one residue, multiplied by the penalty weight which, in turn, is set according to the chain length: the longer the chain, the higher it is.

The original HP model uses only the hydrophobic non-local bonds term to evaluate an individual but, according to [5], without a modified energy function, there will exist large plateaus in the energy landscape on which local search cannot find a descent direction, leading to a random search. Thus, we propose the use of a new concept, called radius of gyration (*Rg*) that, in the scope of the PSP problem, estimates the compactness of a set of amino acids: the more compact a conformation is, the smaller is its radius of gyration. In this term of the fitness function, only hydrophobic residues were considered. This term is presented in Equation 2, where x_i and y_i are the Cartesian coordinates of the *i*-th hydrophobic residue, $\overline{X}$ and $\overline{Y}$ are the mean values of all hydrophobic x_i and y_i, respectively; *NH* is the number of hydrophobic residues in the chain; and *MaxRgH* is the radius of gyration of the chain totally unfolded. The second part of Equation 2 represents the radius of gyration of hydrophobic residues related to the point given by the mean coordinates.

$$RgH = MaxRgH - \sqrt{\frac{\sum_{i=1}^{NH}\left[(x_i - \bar{X})^2 + (y_i - \bar{Y})^2\right]}{NH}}, \quad (2)$$

The term related to the hydrophilic radius of gyration in the fitness function has the opposite purpose as *RgH*: it fosters the spreading of hydrophilic residues towards the edge of the conformation. This term is calculated in the same way as in Equation 2, except that in this case only hydrophilic residues are considered and it is not subtracted from any other value, generating the hydrophilic radius of gyration value *RgP*. Using the calculated *RgP*, this value is redefined according to Equation 3, in order to penalize conformations where the hydrophobic residues are more spread than the hydrophilic:

$$RgP = \begin{cases} 1 & ,if\,(RgP - RgH) \geq 0 \\ \dfrac{1}{1-(RgP - RgH)} & ,otherwise \end{cases}. \quad (3)$$

After individuals are chosen by the tournament selection method, genetic operators are applied to generate a new population. We used the regular two-point crossover, the regular mutation and an improved mutation. This last operator works as the simple mutation except by the fact that after each mutation is applied, the individual is reevaluated to check if its fitness has increased. In this case, the change is maintained, otherwise it is discarded. Another especially devised operator used in this work is named partial optimization. The idea of this operator is to randomly select two non-consecutive residues of the protein and fix their position in the lattice and find the best path that connects both residues.

A strategy called Decimation was also implemented in order to avoid getting trapped in local minima. After 10 generations without improvement of the best individual, all the population is killed, except the best, and generated again from scratch.

4 Computational Experiments and Results

The tests were performed based on the same instances used by [7], for five amino acid chains with 20, 36, 48, 64, and 85 residues. The known maximum number of H–H bonds for those instances are: 9, 14, 23, 42, and 52, respectively.

The parameters used in the algorithm were chosen after several tests with different combinations of parameters. They were those that performed best among the combinations tested. The main parameters used were: population size (500), number of generations (100), PopIniFull (30%), tourney size (3%), elitism (yes), crossover probability (50% to 70%), mutation probability (5% to 10%), partial optimization probability (4%) and partial optimization size (7 residues). Tests were run 100 times and the best individual of each run was considered. The results are compared in Table 1. For the 20-residue chain, the mean value column represents the average of evaluations to find the global minimum. For the first three chains our results were very similar to [7], and our GA got better results for the longer chains.

Table 2. Comparison of results

Chain length	König and Dandekar		Our implementation	
	Best Score	Mean Value	Best Score	Mean Value
20	9 (100×)	11824	9 (100×)	**10830**
36	14 **(8×)**	**12.40**	14 (4×)	11.89
48	23 (1×)	18.50	23 (1×)	**18.69**
64	37 (1×)	29.30	**39 (1×)**	**31.19**
85	46 (1×)	40.80	**51 (1×)**	**44.18**

5 Conclusions

This paper presented novel strategies for using a genetic algorithm for the protein structure prediction problem. The use of the concept of radius of gyration in the fitness function took smoothness to the fitness landscape, allowing better solutions to be found. Also, the use of the partial optimization and improved mutation operators, together with the decimation strategy have enhanced the GA, allowing it to escape from local minima. Results encourage the continuity of the research towards a more complex lattice model, and further tests with other real-world biological sequences.

6 Acknowledgments

Authors would like to thank CNPq for the financial support to H.S.Lopes (processes 350053/03-0 and 402018/2003-6) and M.P.Scapin (process 131355/2004-0).

References

1. Unger, R., Moult, J.: A genetic algorithm for three dimensional protein folding simulations. In: Proc 5th Int Conf on Genetic Algorithms (1993) 581–588
2. Pedersen, J.T., Moult, J.: Protein folding simulations with genetic algorithms and a detailed molecular description. J. Mol. Biol. 269 (1997) 240–259
3. Day, R.O., Lamont, G.B., Pachter, R.: Protein Structure prediction by applying an evolutionary algorithm. In: Int Parallel and Distributed Processing Symp (2003) 155–162
4. Dill, K.A.: Theory for the folding and stability of globular proteins. Biochemistry 24 (1985) 1501–1509
5. Krasnogor, N., Hart, W.E., Smith, J., Pelta, D.A.: Protein structure prediction with evolutionary algorithms. In: Proc Int Genetic and Evolutionary Comput Conf (1999) 1596–1601
6. Patton, A.L., Punch III, W.F., Goodman, E.D.: A standard GA approach to native protein conformation prediction. In: Proc 6th Int Conf Genetic Algorithms, Morgan Kauffman (1995) 574–581
7. König, R., Dandekar, T.: Improving genetic algorithms for protein folding simulations by systematic crossover. BioSystems 50 (1999) 17–25