Genetics and Molecular Biology, 27, 4, 611-615 (2004)

www.sbg.org.br

Research Article

Investigation of the three-dimensional lattice HP protein folding model using a genetic algorithm

Fábio L. Custódio, Hélio J.C. Barbosa and Laurent E. Dardenne

Laboratório Nacional de Computação Científica, Petrópolis, RJ, Brazil.

Abstract

An approach to the hydrophobic-polar (HP) protein folding model was developed using a genetic algorithm (GA) to find the optimal structures on a 3D cubic lattice. A modification was introduced to the scoring system of the original model to improve the model's capacity to generate more natural-like structures. The modification was based on the assumption that it may be preferable for a hydrophobic monomer to have a polar neighbor than to be in direct contact with the polar solvent. The compactness and the segregation criteria were used to compare structures created by the original HP model and by the modified one. An islands' algorithm, a new selection scheme and multiple-points crossover were used to improve the performance of the algorithm. Ten sequences, seven with length 27 and three with length 64 were analyzed. Our results suggest that the modified model has a greater tendency to form globular structures. This might be preferable, since the original HP model does not take into account the positioning of long polar segments. The algorithm was implemented in the form of a program with a graphical user interface that might have a didactical potential in the study of GA and on the understanding of hydrophobic core formation.

Key words: HP model, genetic algorithms, protein folding.

Received: August 15, 2003; Accepted: August 20, 2004.

Introduction

Determining the functional conformation of a protein molecule from amino acid sequence remains a central problem in computational biology (Pedersen and Moult, 1996). Even the experimental determination of these conformations is often difficult and time consuming. To address this problem, it is 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 (Lyngsø and Pedersen, 2000).

The hydrophobic-hydrophilic (or hydrophobic-polar, HP) model (Dill and Lau, 1989) describes the proteins based on the fact that hydrophobic amino acids tend to be less exposed to the aqueous solvent than the polar ones thus resulting in the formation of a hydrophobic core in the spatial structure. In the model, the amino acid sequence is abstracted to a binary sequence of monomers that are either hydrophobic or polar. Even though some amino acids cannot be clearly classified, the model disregards this fact to achieve simplicity. The structure is a self-avoiding chain whose monomers are on the vertices of a three-dimensional cubic lattice.

Send correspondence to Fábio Lima Custódio. Laboratório Nacional de Computação Científica, Av. Getúlio Vargas 333, sala 1A19, Quitandinha, 25651-075 Petrópolis, RJ Brazil. E-mail: flc@lncc.br.

The free energy of a conformation is defined as the negative number of non-consecutive hydrophobic-hydrophobic contacts. A contact is defined as two non-consecutive monomers in the chain occupying adjacent sites in the lattice. The HP model's folding process has behavioral similarities with folding in the real system (Dobson *et al*., 1998).

In spite of its apparent simplicity, finding optimal structures of the HP model on a cubic lattice has been classified as a NP-complete problem (Berger and Leighton, 1998). This means that it belongs to a class of problems that are believed to be computationally intractable. Since the free energy on the original model is given only by the number of non-specific hydrophobic contacts, it is a fact that the positions of polar segments are not directly optimized when searching for optimal structures. This may result in unnatural structures if these segments are too long, or located at the ends of the sequences. A modification is proposed to the original HP model's scoring system, to try to obtain more natural-like structures. The modification was based on the assumption that it may be preferable for a hydrophobic monomer to have a polar neighbor than to be in direct contact with the polar solvent. The free energy of the model is now given by a weighted sum of the number of hydrophobic-hydrophobic (H-H) contacts, the number of hydrophobic-polar contacts and the number of hydrophobic-solvent

contacts. As in the original model, monomer collisions are penalized.

This work is an investigation of the HP (hydrophobic-polar) model in a three-dimensional cubic lattice using a genetic algorithm (GA) as a tool to find the optimal conformation for a given sequence. The GA did not use information about the problem in any form. The effects of modifications on the GA and on the original HP model were evaluated using sequences of 27 monomers and 64 monomers from the literature (Unger and Moult, 1993).

Methods

Ten test sequences were run 50 times, allowing a maximum of 3,500,000 function evaluations for each run, in order to access the efficiency of variations of the standard GA (Goldberg, 1989) and to compare the results obtained with the two models. The comparison between the models was made using the compactness (average number of contacts per monomer) and the segregation criteria (standard deviation of the number of contacts per monomer) (Araújo, 1999). Results on the free energy of the original HP model from the literature (Patton *et al*., 1995) are used as a reference and we considered a successful run (an optimal structure was found) to be one where an equal or better value was obtained. The parameters of the algorithm were a population size of 100 individuals, a recombination probability of 0.8 and a mutation probability of 0.05.

A standard GA consists of an algorithm that uses natural selection's principle to optimize an objective. A representation of a possible solution is codified in a chromosome. A population of these solutions is generated (usually randomly) and allowed to recombine (crossover) and mutate (genetic operators) to form a new population. The new population is then evaluated and a process of selection picks the best solutions to form the next generation.

Absolute encoding was used to represent the structures in a vector of real numbers (0-5) of the same length as the HP sequence with each position indicating the global direction of the next monomer *i.e.*, a chromosome containing 0 2 1 would mean: west, south and east. The use of this encoding has the drawback of allowing a larger number of collisions in the shape of redundant moves (east - west, south - north). The initial populations were randomly generated.

Initially we used a standard GA in which the selection scheme copied the newly generated population over the original one ensuring that the best original structure was maintained. A different selection scheme was implemented as follows: for each individual on the new population (sequentially), scan the original population for an equal or worse solution and replace it. The new individuals face competition with their own brothers as the old population is gradually replaced. This new scheme my cause the loss of good individuals, but it has the advantage of maintaining diversity in the population. If we replaced an individual when it is worse as opposed to "equal or worse" the algorithm would converge prematurely.

The selection of individuals for recombination was random so that structures with collisions still had a chance to improve. Individuals with illegal states (collisions) were tolerated in the population, but their score (Piccolboni and Mauri, 1997) was penalized (set to zero). The illegal individuals were represented just as a legal one. The decoding step (prior to using the objective function) generated a list of coordinates, an illegal individual possessed duplicate (or triplicate) coordinates. A standard two-point crossover was used. Later, the recombination algorithm was modified to calculate the number of cut points on-the-fly (multiple--point crossover, or mp crossover), according to the length of the HP sequence. The number of cut points was calculated to compute approximately one point for each 10 monomers, so the 27-monomers sequences had three points and the 64-monomers six points. The position of the cuts were randomly chosen. The mutation operator implemented was that if an individual is selected for mutation (according to the probability), randomly select a base on its chromosome and change it to a randomly selected direction.

An islands' algorithm (Whitley, 1994) was implemented. The idea behind it is to run a series of GA with small populations and fewer generations and use the best structure (individual) resulting from each of these runs (islands) to form an initial population for another GA with more generations. The islands parameters used were: 100 islands, a population size of 10 individuals, 500 generations. The operator probabilities used on the islands were the same as those used on the standard GA.

Results and Discussion

The comparison of the standard GA (Table 1) with its modifications, showed that the mp crossover improved the results in terms of number of H-H contacts found and, as

Table 1 - Original HP model: Performance comparison for variations on the standard GA* A: standard with two-point crossover. B: standard with multiple-points crossover (mp crossover). C: GA with new selection scheme and mp crossover. D: islands' GA with new selection scheme and mp crossover.

Case[1]	A[2]	B	C	D
27.2	9 6.48(1.33)	9 6.68(1.35)	10 9.10(0.36)	10 9.16(0.37)
27.4	12 9.36(1.55)	13 10.27(1.67)	15 14.26(0.75)	15 14.30(0.68)
27.8	4 3.10(0.79)	4 3.68(0.55)	4 4(0.00)	4 4(0.00)
64.3	28 20.22(3.21)	30 21.69(3.87)	34 27.50(2.40)	34 29.95(2.19)

*Results are the number of H-H contacts from 50 runs.
[1]Cases are numbered according to Patton *et al*. (1995).
[2]Best result, average and standard deviation.

expected, the improvement was more pronounced on the longer sequence (Table 1, case 64.3). The different selection scheme enhanced the results not only on the best cases, but especially on the average number of H-H contacts. The islands' algorithm showed slight improvements on the average number of H-H contacts of the 27-monomers sequences. With regard to the 64-monomer sequence (Table 1), the improvement was visible on the average number of H-H contacts across the population.

The islands' algorithm coupled with the new selection scheme and the mp crossover was chosen to obtain the structures for the comparison between the original and the modified HP model. Results depicted in Table 2 show that the modification to the model not only resulted in structures that are more compact and have a higher degree of segregation, but also increased, in some cases (64-monomer sequences, see Table 2), the number of H-H contacts found.

There are no special operators (Lesh *et al*., 2003, Clote *et al*., 2000) or special encodings in our algorithm, but our results are comparable to the well-known reference (Unger and Moult, 1993) and to Patton *et al*. (1995) who used a preference order encoding that had the effect of avoiding collision states. An equivalent or better number of contacts was found in six out of seven 27-monomer sequences and one out of three 64-monomer sequences.

The genetic algorithm implemented in this work is capable of folding 27-monomer sequences of the HP model, in a three-dimensional cubic lattice, without using knowledge of the problem in any form.

The multiple-point crossover showed a good potential in improving the performance, especially when dealing with longer sequences. Given the nature of the problem, a small modification on the chromosome can have drastic effects on the structure, *i.e.*, changing the direction of a monomer changes the position of all following monomers, which is worse in longer sequences. The mp crossover tries to lessen the drastic effects that a single point (or a two-point) crossover can have on a large structure by interchanging smaller fragments and thus reducing chances of creating collisions or by conserving parts that already have a good fold.

The new selection scheme coupled with the random process of selection of individuals for the genetic operators improved the overall performance. Every individual had an equal chance of generating descendents. They suffered competition from their own population and from the population of their parents. The algorithm places its evolving pressure uniquely on the transition of newly generated individuals to the next generation; this scheme keeps the diversity high and prevents premature convergence, as even individuals with low scores can participate in the formation of the new generation. This selection scheme does not simply choose the best individuals amongst two populations, as

Table 2 - Results from 50 runs of the 10 cases. *a*, best result, *b*, average and *c*, standard deviation. The runs allowed for a maximum of 3,500,000 function evaluations using the islands' algorithm with the modified selection scheme and the mp crossover.

	# H-H Contacts[2]			Compactness[2]		Segregation[2]	
Case1	HP	HP Mod.	U&M[3]	HP	HP Mod.	HP	HP Mod.
27.1	9 8.56(0.5)	8 7.76(0.46)	9	1.55 1.11(0.18)	1.62 1.49(0.1)	1.12 1.00(0.05)	1.36 1.21(0.06)
27.2	10 9.16(0.37)	10 9.15(0.61)	9	1.55 1.12(0.18)	1.70 1.44(0.12)	1.35 1.10(0.09)	1.36 1.23(0.07)
27.4	15 14.30(0.68)	15 14.31(0.81)	15	1.7 1.43(0.13)	1.85 1.56(0.13)	1.37 1.22(0.08)	1.39 1.24(0.07)
27.5	8 8(0.00)	8 7.70(0.5)	8	1.33 0.97(0.14)	1.63 0.44(0.15)	1.22 1.04(0.07)	1.40 1.23(0.08)
27.7	13 11.94(0.37)	13 11.96(0.94)	12	1.63 1.38(1.16)	1.85 1.20(0.09)	1.34 1.18(0.07)	1.40 1.21(0.09)
27.8	4 4(0.00)	4 4(0.00)	4	1.26 0.82(0.20)	1.63 1.36(0.14)	1.13 0.91(0.12)	1.42 1.27(0.07)
27.9	7 7(0.00)	7 6.84(0.41)	7	1.56 0.99(0.15)	1.70 1.39(0.14)	1.29 1.09(0.07)	1.42 1.26(0.06)
64.1	20 17.92(1.59)	22 18.14(0.17)	27	1.66 1.33(0.16)	2.09 1.80(0.11)	1.31 1.14(0.08)	1.37 1.28(0.05)
64.2	25 21.50(1.54)	26 21.48(2.13)	29	1.84 1.35(0.17)	2.06 1.84(0.09)	1.29 1.16(0.07)	1.44 1.29(0.06)
64.3	34 29.95(2.19)	35 30(2.52)	35	1.84 1.51(0.16)	2.22 1.91(0.11)	1.36 1.21(0.07)	1.37 1.25(0.07)

[1]Cases are numbered according to Patton *et al*. (1995).
[2]Best result, average and standard deviation.
[3]Maximum number of hydrophobic-hydrophobic contacts obtained by Unger and Molt, 1993 *apud* Patton *et al*., 1995.

such a situation has the potential to quickly decrease the diversity of a population.

The results of the islands' algorithm may not yet justify its application due to the extra computational cost. For longer sequences more tests with a wider range of parameters should be made before establishing a conclusion. It is expected that as the length of the sequence increases, the effect of the islands will be enhanced as its not intended to find optimal structures, but rather to accelerate the untangling and the local optimization of the structures. The islands' algorithm can be seen as providing an improved (with fewer collisions) initial population.

The modification on the original model's scoring system proved to be useful in obtaining structures that are more natural-like (globular), more compact and more segregated. It has been shown (Garcia *et al*., 2001) that the more segregated the structure, the more it behaves like a native protein conformation in terms of folding dynamics, *i.e.*, folds cooperatively and exhibits thermodynamics of a two-state system (Barbosa, 2003). The compactness and the segregation were not used in the evaluation of the individuals' fitness and, as shown here, are products of the model. A visual comparison of the results generated by both models indicates that the modification induced the formation of more globular structures while maintaining the same number of H-H contacts (in some actually cases improving that number). These structures (Figure 1, B) do not have the unnatural extended polar loops (Figure 1, A) and ends that the original model yields. By analyzing structures obtained for the sequence [PPPPPPPPHHHHHHHHPPPPPPPP] under the two models this becomes more evident. This sequence is very artificial in the sense that it concentrates all its H monomers in one region, nevertheless, this is intended to maximize the effects of the modification and such a sequence can be easily folded by the algorithm. The structures (Figure 2, A and B) show that the original model does not exercise selective pressure on the positioning of the long polar domains except for avoiding collisions. The

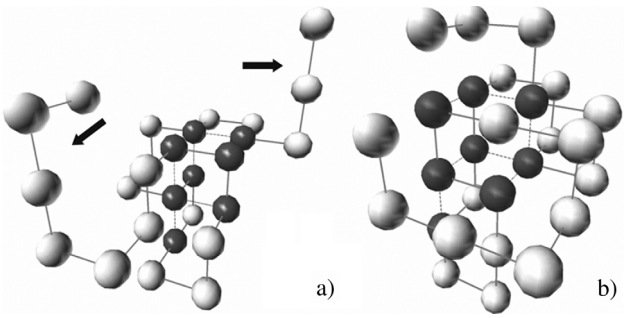

Figure 1 - Structures from case 27.9 under the different models. A: original model, the arrows point to unnaturally extending ends. B: modified model, the ends are now folded around the hydrophobic core. Dark spheres are hydrophobic monomers and white spheres polar ones. The dotted line indicates a hydrophobic contact and the continuous line is the backbone. The number of hydrophobic contacts, compactness and segregation values are given on Table 2.

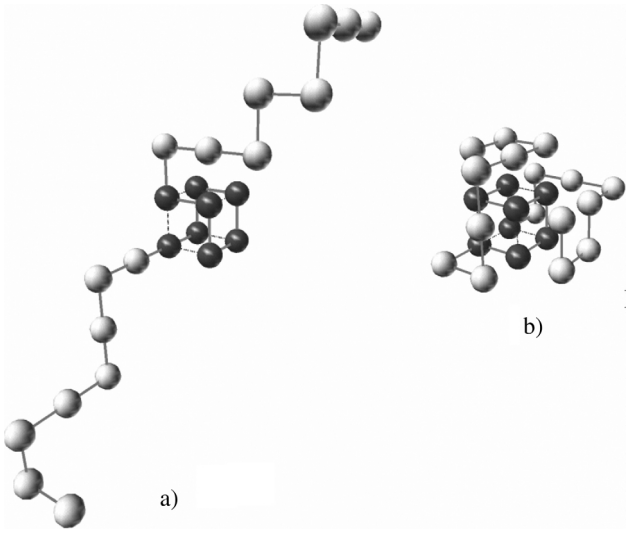

Figure 2 - Structures from sequence [PPPPPPPPHHHHHHHHPPPPPPP] under the different models. Dark spheres are hydrophobic monomers and white spheres polar. The dotted line indicates a hydrophobic contact and the continuous line is the backbone. A: original model, 5 hydrophobic-hydrophobic (H-H) contacts, compactness 0.5 and segregation 0.72. B: modified model, 5 H-H contacts, compactness 1.58, segregation 1.14.

modified model accounts for the long polar domains and tries to "bury" the hydrophobic monomers forming the core.

The modification also improved, in some cases, the number of hydrophobic contacts. Since the problem exhibits degenerated minimum states (Yue *et al*., 1995) a scoring system that narrows the search space may be a useful tool in obtaining structures with some specific characteristics. The cubic lattice presents some structural limitations, for example, hydrophobic monomers separated by an odd number of monomers will never make a contact (Chandru *et al*., 2003). These limitations may be overcome by using a different lattice with more degrees of freedom, but such sophistications would be a challenge to the algorithm in its present form. It may be difficult to improve the performance of the algorithm without applying knowledge of the problem in some way, but it is expected that the use of a relative over an absolute encoding may yield superior performance (Krasnogor *et al*., 1999). Currently, new operators designed to deal with this problem *(e.g*., a repair operator) and also, as the number of operators increases, an adaptive algorithm (Davis, 1991) are being implemented.

The algorithm was implemented in C++ without using any available GA libraries. The resulting software has a user-friendly graphical interface that permits the visualization of the results and the modification of the parameters with ease. The program might have didactical applications in showing and testing the effects of the GA's parameters on its performance and the basic principle of formation of the hydrophobic core on protein structures. The program is freely available through e-mail contact at flc@lncc.br.

Acknowledgements

We would like to thank CNPq/MCT (grant number 40.2003/2003.9) and FAPERJ (grant number. E26/171.401/01) for their financial support. Special thanks to the Carcara Project at LNCC for the computational cluster resources provided.

References

Araújo AFP (1999) Folding protein models with a simple hydrophobic energy function: The fundamental importance of monomer inside/outside segregation. Proc Natl Acad Sci 96 22:12482-12487.

Barbosa MAA and Araújo AFP (2003) Relevance of structural segregation and chain compaction for the thermodynamics of folding of a hydrophobic protein model. Phys Rev E 67:051919-1–051919-10.

Berger B and Leighton T (1998) Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. J Comput Biol 5:27-40.

Chandru V, Dattasharma A and Kumar, VSA (2003) The algorithmics of folding proteins on lattices. Disc App Math 127:145-161.

Clote P, Backofen R and Will S (2000) Algorithmic approach to quantifying the hydrophobic force contribution in protein folding. In: Altman R (ed) Pacific Symposium on Biocomputing, Honolulu, pp 93-106.

Davis L (1991) Handbook of Genetic Algorithms. Van Nostrand Reinhold, New York, 385 p.

Dill KA and Lau KF (1989) A lattice statistical mechanics model of the conformational sequence spaces of proteins. Macromolecules 22:3986-3997.

Dobson CM, ali A and Karplus M (1998) Protein folding: A perspective from theory and experiment. Angew Chem Int Ed Engl 37:868-893.

Garcia LG, Treptow WL and Araújo AFP (2001) Folding simulations of a three-dimensional protein model with a nonspecific hydrophobic energy function. Phys Rev E 64 011912.

Goldberg DE (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Boston.

Krasnogor N, Hart WE, Smith J and Pelta DA (1999) Protein structure prediction with evolutionary algorithms. In: Banzhaf W, Daida J, Eiben AE, Garzon MH, Honavar V, Jakiela M and Smith RE (eds) Proc of the Genetic and Evo Comp Conf, Morgan Kaufmann, San Francisoco, pp 1596-1601.

Lesh N, Mitzenmacher M and Whitesides SA (2003) Complete and effective move set for simplified protein folding. In: 7th An Int Conf on Res in Comp Mol Biol. RECOMB, Berlin.

Lyngsø RB and Pedersen CNS (2000) Protein folding in the 2D HP model. In: Proceedings of the 1st Journees Ouvertes: Biologie, Informatique et Mathematiques, JOBIM, Montpellier.

Patton A, Punch W and Goodman EA (1995) A standard ga approach to native protein conformation prediction. In: Eshelman L (ed) Proc. Sixth Int Conf Gen Algo, Morgan Kaufmann, San Francisco, pp 574-581.

Pedersen JT and Moult J (1996) Genetic algorithms for protein structure prediction. Curr Opin Struct Biol 6:227-231.

Piccolboni A and Mauri G (1997) Application of Evolutionary algorithms to protein folding prediction. Artificial Evolution, 1363:123-136.

Unger R and Moult J (1993) A genetic algorithm for 3D protein folding simulations. In: Forrest S (ed) Proceedings of the Fifth Annual International Conference on Genetic Algorithms, Morgan Kaufmann, San Francisco, pp 581-588.

Whitley D (1994) A genetic algorithm tutorial. Statistics and Computing 4:65-85.

Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI and Dill KA (1995) A test of lattice protein folding algorithms. Proc Natl Acad Sci 92:325-329.