Protein Structure Prediction With Evolutionary Algorithms

Natalio Krasnogor
Intelligent Computer
System Centre
Univ of the West
of England
Bristol, United Kingdom
natk@ics.uwe.ac.uk

William E. Hart
Sandia National Laboratories
PO Box 5800, MS1110
Albuquerque, NM 87185 USA,
wehart@cs.sandia.gov

Jim Smith
Intelligent Computer
System Centre
Univ of the West
of England
Bristol, United Kingdom
jim@ics.uwe.ac.uk

David A. Pelta
Dept de Ciencias
de la Computacion e
Inteligencia Artificial
Universidad de Granada
17801 Granada, España
dpelta@platon.ugr.es

Abstract

Evolutionary algorithms have been successfully applied to a variety of molecular structure prediction problems. In this paper we reconsider the design of genetic algorithms that have been applied to a simple protein structure prediction problem. Our analysis considers the impact of several algorithmic factors for this problem: the conformational representation, the energy formulation and the way in which infeasible conformations are penalized. Further we empirically evaluate the impact of these factors on a small set of polymer sequences. Our analysis leads to specific recommendations for both GAs as well as other heuristic methods for solving PSP on the HP model.

1 INTRODUCTION

A protein is a chain of amino acid residues that folds into a specific *native* tertiary structure under certain physiological conditions. A protein's structure determines its biological function. Consequently, methods for solving protein structure prediction (PSP) problems are valuable tools for modern molecular biology. Exhaustive search of a protein's conformational space is not a feasible algorithmic strategy for PSP even for small protein sequences. Furthermore, recent computational analyses of PSP have shown that this problem is intractable on simple lattice models [1, 3]. Consequently, heuristic optimization methods seem the most reasonable algorithmic choice to solve PSP problems. In particular, evolutionary methods have been used by a variety of researchers [14, 13, 10, 7, 8, 11, 12].

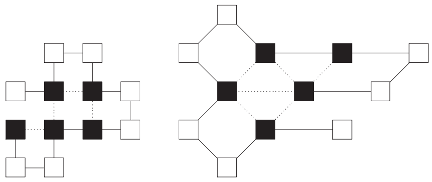
In this article we examine the basic design principles that have guided prior work with GAs on the PSP problem for the HP model [4]. We focus on a simple lattice model because lattice models can capture many global aspects of protein structures, they are inexpensive to use, and it is possible to design test problems for which the best conformational structure is known (for small protein sequences). The PSP problem for the HP model is a good test problem for evaluating GAs because its complexity is well understood and there has been a lot of prior work developing heuristics and global optimization methods for this problem.

We consider three basic algorithmic factors that affect how GAs are applied to this PSP problem. First, we evaluate the representations commonly used for this problem and describe equivalences between different operators across these search domains. Next, we propose a new method for formulating the energy potential for the HP model that makes the energy potential more continuous while preserving the integer rank order of conformations in the search domain. Finally, we describe how penalty methods can be used to safely enforce self-avoiding constraints.

2 THE HP PROTEIN FOLDING MODEL

One of the most studied simple protein models is the hydrophobic-hydrophilic model (HP model) proposed by Dill [4]. HP models abstract the hydrophobic interaction process in protein folding by reducing a protein to a heteropolymer that represents a predetermined pattern of hydrophobicity in the protein; nonpolar amino acids are classified as hydrophobic and polar amino acids are classified as hydrophilic. A sequence is $s \in \{H, P\}^+$, where H represents a hydrophobic amino acid and P represents a hydrophilic amino acids.

The HP model restricts the space of conformations to self-avoiding paths on a lattice in which vertices are labeled by the amino acids. The energy potential in the HP model reflects the fact that hydrophobic amino acids have a propensity to form a hydrophobic core. To capture this feature of protein structures, the HP model adds a value ϵ for every pair of hydrophobics that form a topological contact; a topological contact is formed by a pair of amino acids that are adjacent on the lattice and not consecutive in the sequence. The value of ϵ is typically taken to be -1. Figure 1 shows sequences embedded in the square and the triangular, with hydrophobic-hydrophobic contacts (HH contacts) highlighted with dotted lines. The conformation in Figure 1a has an energy of -4 and the conformation in Figure 1b has an energy of -6.



(a) (b)

Figure 1: HP sequences embedded in (a) the square lattice and (b) the triangular lattice.

3 PSP METHODS FOR THE HP MODEL

Despite the simplicity of the HP model, it is powerful enough to capture a variety of properties of actual proteins [5]. The PSP problem for the HP model has been shown to be NP-complete on the square lattice [3] and cubic lattice [1], and performance guaranteed approximation algorithms have been developed for a variety of lattice models (e.g. see Hart and Istrail [6]).

A wide variety of global optimization techniques have been applied to PSP (e.g. see the papers in Biegler *et al.* [2] and Pardalos, Shalloway and Xue [9]). In particular, GAs have proven a particularly robust and effective global optimization technique for PSP. For these methods, the embedding of a sequence in a lattice may be represented in a number of ways. Three common methods are Cartesian coordinates (the location of each acid on the lattice is specified independently), internal coordinates (the protein is specified as a sequence of moves taken on the lattice from one acid to the next), and as a distance matrix (amino acid locations are inferred from inter-amino acid distances).

An early application of GAs to PSP was that of Unger and Moult [14, 13]. Their GA uses internal coordinates that specify an *absolute* direction on a square or cubic lattice. Thus individuals are coded with a sequence in $\{U, D, L, R, F, B\}^{n-1}$ (which correspond to up, down, left, right, forward and backward moves in a cubic for a length n protein). Additionally, their GA only considers feasible conformations that are self-avoiding paths on the lattice. When mutation and crossover are applied, their GA iterates until a feasible conformation is generated.

Patton *et al.* [10] describe a standard GA that significantly outperforms the GA used by Unger and Moult [13]. They also employed an internal coordinate representation that uses *relative* offsets from the current position, together with a a chain-growth method to help the GA search through feasible conformations. A standard form of relative offsets represents a conformation as a sequence in $\{F, L, R, U, D\}^{n-2}$. In this case the direction is interpreted relative to the direction of the previous move, rather than relative to the axes defined by the lattice. This has the advantage of guaranteeing that all solutions are 1-step self–avoiding (since there is no "back" move). The authors use a penalty method to enforce the self-avoiding constraints. Their objective function adds a penalty if two or more amino acids lie at the same position on the lattice. Further, any hydrophobic amino acid which lies at the same position as another amino acid does not add hydrophobic-hydrophobic contacts to the potential energy.

Khimasia and Coveney [7] considered the performance of Goldberg's Simple Genetic Algorithm (SGA) using internal coordinates with absolute moves. The objective function was defined as a hybrid between the Random Energy Model and the HP model. This included two penalty terms: a penalty for each lattice site that has two amino acids on it, and a penalty for each lattice site that has three or more monomers on it.

Krasnogor *et al.* [8] empirically evaluate what mix of evolutionary operators (mutations, macromutations, crossover) were most useful for solving the PSP problem for the HP model. Their experiments evaluated GAs that applied these operators with different combinations of probabilities. Their results strongly suggest that (1) one point crossover was not able to transfer building blocks and (2) macromutation was acting like powerful local search. For the instances studied, the best combination of parameters had a small crossover probability and high mutation and macromutation probabilities.

4 ALGORITHMIC DESIGN

In this section we critique three algorithmic factors that impact the performance and general applicability of GAs for PSP problems: the energy potential, the method of constraint management, and the conformational representation.

4.1 ENCODINGS FOR INTERNAL COORDINATES

When working with lattice models, proteins are often represented using internal coordinates. However no comparative studies have evaluated whether an absolute (as per Unger above) or relative (as per Patton *et al.*) representation is more effective for GAs; prior researchers selected an encoding without explicit numerical comparisons [10, 13, 8]. Since other algorithmic parameters were also chosen differently, it is difficult to assess the impact that the choice of encoding has on a GA's performance. In this section we illustrate how the fitness landscapes induced by these encodings and the standard genetic operators can have important differences that may affect the global search behavior of the GA. Our discussion considers the two dimensional square lattice, but an extension to other discrete lattices is straightforward.

Mutation on the Relative Encoding Consider the effect of one-point mutation on the structure in Figure 2(a). It's relative encoding is $S_{rel} = FLLFR\mathbf{R}LRLLR$ when viewed from the H amino acid. A one point mutation in the sixth position could produce either of $S^1_{rel} = FLLFR\mathbf{F}LRLLR$ or $S^2_{rel} = FLLFR\mathbf{L}LRLLR$, which are shown in Figures 2(b) and 2(c) respectively.

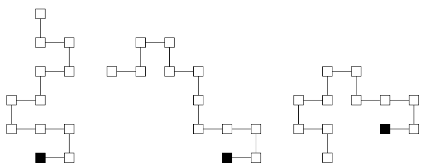


Figure 2: In (b) a one point mutation of the structure in (a) at the sixth gene. An 'R' was mutated to an 'F' producing a lever effect of 90 degrees counterclockwise. In (c) an 'R' was mutated to an 'L' producing a lever effect of 180 degrees counterclockwise

We can see from this example that a one point mutation in the relative encoding produces a rotation effect in the structure at the mutated point. To produce the same effect in the absolute encoding we must perform a macromutation, that is, several genes need to be simultaneously mutated to produce the same change in the structure. We define a rotation operator in the absolute encoding as, given a point where to produce the rotation, all the remaining genes will be changed according to a mapping that depends on the angle to be rotated[1]. In the working example, Figure 2(a) is encoded as $S_{abs} = RULLU\mathbf{RURULU}$ while the first mutated structure (Figure 2(b)) is $S^1_{abs} = RULLU\mathbf{ULULDL}$ and Figure 2(c) is $S^2_{abs} = RULLU\mathbf{LDLDRD}$.

Mutation on the Absolute Encoding A One-point mutation in an absolute encoding leaves the orientation of the rest of the structure unchanged. To achieve the same effect in the relative encoding, it is necessary to change two subsequent values in the encoding. There are, however, restrictions in the mapping from a one-point mutation under the absolute encoding to a two-point mutation in the relative encoding. Specifically, point mutations in an absolute encoding can produce structures that are not one-step self-avoiding, which have no equivalent in a relative encoding.

4.2 POTENTIAL ENERGY FORMULATION

Figure 3 illustrates two conformations of a hydrophobic sequence that are formed from two domains connected by a hydrophilic chain. Since the HP model only rewards direct hydrophobic-hydrophobic contacts, only the compact subconformations contribute energy to these conformations. However, it is clear that Figure 3a is closer to forming the optimal conformation than Figure 3b.

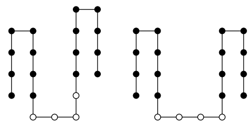


Figure 3: Two conformations with equal energy for the HP model. Figure (a) has lower energy for the modified HP model.

This type of disparity between the energy value and the "closeness" of the conformation can be remedied by augmenting the energy function to allow a distance-dependent hydrophobic-hydrophobic potential. Since the distances between amino acids form a countable set, it is possible to construct a distance-dependent potential that preserves the rank order of the conformations in the HP model while enabling a finer level of distinction between conformations with the same number of hydrophobic-hydrophobic contacts. For example, if d_{ij} is the distance between two hydrophobic

[1]I.e. to rotate 90 degrees clockwise $U \mapsto R, D \mapsto L, R \mapsto D, L \mapsto U$

amino acids H_i and H_j, then we can use

$$\hat{E}_{H_iH_j}(d_{ij}) = \begin{cases} -1 & ,d_{ij}=1 \\ -1/(d_{ij}^k N_H) & ,d_{ij}>1 \end{cases}, \quad (1)$$

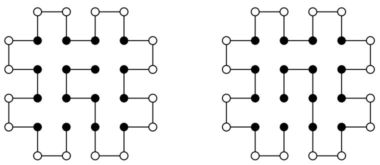
where N_H equals the number of hydrophobics in the polymer sequence, and where $k = 4$ for the square lattice and $k = 5$ for the triangular and cubic lattices.

4.3 CONSTRAINT MANAGEMENT

Two broad classes of constraints need to be enforced to define a feasible conformation: (1) the connectivity of the polymer chain and (2) the self-avoidance of the conformation. Perhaps the strongest motivation for using internal coordinates is that they handle the first constraint implicitly, whereas this must be done explicitly if Cartesian coordinates are used.

Two basic approaches have been taken to manage self-avoiding constraints when internal coordinates are used. First, the search is constrained to only consider feasible, self-avoiding conformations. This method is not well suited to PSP problems, though, because the shortest path from one compact feasible conformation to another may be very long when compared with the shortest path through the space of infeasible conformations. For example, in Figure 4 the HP sequence can move from conformation (a) to conformation (b) through three single-point changes. These changes generate infeasible conformations, but the sequence of feasible conformations which perform this move would clearly be much larger than this.

The second approach to enforcing constraints uses penalties to guide the GA toward feasible solutions. Two penalty methods have been used to solve PSP for the HP model. First, a penalty is added for every pair of amino acids that lie at the same lattice point. Using this method, there may be $O(n^2)$ penalties. Second, a penalty is added for every lattice point at which there are two or more amino acids. Using this, there may be $O(n)$ penalties. Patton *et al.* [10] extend this further to prevent hydrophobic amino acids from contributing to the objective function if they lie on a lattice point with other amino acids.



(a) (b)

Figure 4: Two compact conformations which are "close" if infeasible moves are allowed.

When evaluating these penalty methods, it is important to consider whether they correctly bias the search strategy to feasible regions. The GAs discussed in Section 3 that use a penalty method apply a fixed constant penalty term C per violation. This policy can cause problems if the second penalty method is applied without the extension of Patton *et al.* . For fixed values of C is possible to construct examples where the structure with optimal energy with the penalty method does not correspond to the optimal energy for the HP model. It is also important to consider the efficacy of the penalty method to understand how well they facilitate optimization. For example, we believe that the extended formulation proposed by Patton *et al.* may lead to a less effective search than other methods. When the hydrophobic amino acids are prevented from contributing to the objective function because they overlap, the fitness landscape may have large flat regions, which can make the optimization problem more difficult.

These considerations recommend the use of a fixed penalty approach that is adapted based on the number of hydrophobics available in the protein sequence, N_H. The idea is that we can select C sufficiently large to ensure the validity of the fixed-penalty constraint formulation. For example, on the square lattice if $C = 2N_H + 2$ then the penalty is large enough that any infeasible conformation has positive energy while all feasible conformations have nonpositive energy. Thus the optimal conformation of the HP model is strictly better than the best penalized conformation.

5 METHODS AND RESULTS

The GAs used in our experiments had a $(500 + 500)$ selection strategy, and mutation was applied to each structure with probability 0.3. One-Point mutation was used to change one value in an absolute encoding, and two-point mutation was used to change two consecutive values in a relative encoding. One-Point, Two-Point and Uniform crossover operators were used with probability of 0.8. Each run of the GA consisted of 200 generations. For the comparison of the encodings the proposed modified energy potential was used. Furthermore, every pair of amino acids mapped to the same lattice position was penalized with a constant penalty, C, dependent on the length of the instance. We used five polymer sequences in our experiments, which have a relatively short length (less than 50 monomers). The instances used can be found at `ttp://www.ics.uwe.ac.uk/staff/Natk`.tml.

RELATIVE VS. ABSOLUTE ENCODING In order to evaluate the effect of the encoding on the ability of the GA to find low energy configurations, a series of experiments were run using the GA to find optimal configurations for a number of proteins under the two encodings. The performance metric was the fitness of the best individual in the final generation of each run. In order to distinguish the effects of the encodings from that of the choice of lattice or operators, experiments were run in three different embedding spaces: two dimensional square and triangular lattices and a three dimensional square lattice. Five protein instances of differing length and difficulty were chosen for each lattice. In each type of lattice separate experiments were run using One-Point, Two-point and Uniform Crossover. As described above the mutation operators used have the same phenotypic effect in each encoding. Twenty nine runs were done for each of the 45 combinations.

Figure 5 summarizes the result of this experiment. For each combination of lattice, crossover operator, and protein sequence, we computed the rank over all trials of the two experiments that use either an absolute or relative encoding of protein conformations. A maximum rank of 58 is possible since there are 29 trials in each of the experiments in each pair. This figure shows boxplots of the relative ranks of the final results for each lattice and for the encodings; maximizing the ranks indicates a better method. This plot clearly indicates that the relative encoding is at least as good in all lattices, and for the square lattice it is much better.

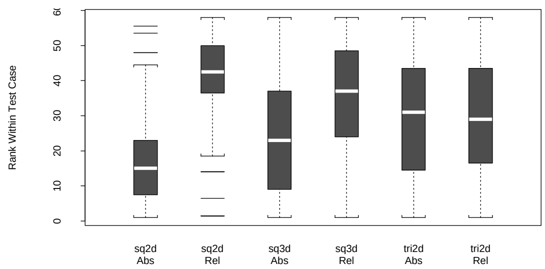


Figure 5: Distribution of relative ranks for relative and absolute encodings on the square, cubic and triangular lattices.

In Table 5 the results of computing the $p - values$ of a $t - test$ are shown, the null hypothesis used was that the mean fitness of both encodings are the same. From this table we can say that: (1) the relative was **almost always** better than the absolute encoding for the square and cubic lattices (at the 95% confidence level) and (2) the robustness of the relative encoding degrades when we look to the triangular lattice (we suspect this might have important consequences in off-lattice modeling).

Lat	Cro	A	B	C	D	E
S2D	1-p	R+	R+	R+	R+	R+
	2-p	R+	R+	R+	R+	R+
	Uni	R+	R+	R+	R+	R+
S3D	1-p	R+	R	-	R+	R+
	2-p	R	R+	R+	R+	R+
	Uni	R+	R+	R+	R	R+
T2D	1-p	-	R	A+	A	A+
	2-p	-	-	-	-	A+
	Uni	R+	R +	R+	A+	A+

Table 1: Summary of t-test analysis: - denotes no significant difference, X denotes Encoding X was better with 90% confidence, X+ denotes 95% confidence, X= R (relative) or A (absolute)

STANDARD VS. DISTANCE ENERGY In this set of experiments we wanted to elucidate if the modified energy potential improves the search capabilities of the GA. For each of the three lattices we used the same first four instances as for the previous section. From the three crossover available we run the simulations for One-point and Two-point crossover. Again, 29 trials were assigned to each of the four instances in each of the three lattices for each crossover tested. The results of these experiments do not show a substantial impact of using the modified energy potential for the test problems. We computed the $p - values$ of a $t - test$ where the null hypothesis was that the final protein configuration were the same using the two energy potential. We found that the two energy potentials had the same performance except for instance B using two-point crossover on the triangular lattice (p = 0.95). There also appeared to be a slight effect for instances C and D on the triangular and cubic lattices, but this was not a significant difference.

6 DISCUSSION

In this paper we have directly compared the encoding of PSP using internal coordinates by means of relative and absolute moves. We have shown the search spaces induced by the relative and absolute encodings to be different and described a mapping of the One-point mutation in the absolute to the relative encoding (the same can be done for other genetic operators). Also, we proposed a modified energy potential that facilitates the GA search while preserving the ranking of the standard HP model. We also identified weaknesses in the standard constraint management strategies and proposed a constraint method that ensures the feasibility of the optimal solution. These algorithmic issues were explored in the three most common lattices being used: two and three dimensional square lattice and two dimensional triangular lattices. Our results support the use of the relative encoding for this problem, which may explain the superior performance of the GA described by Patton et al. [10] as compared to the GA described by Unger.

It was previously argued [11, 7, 8] that the representation with internal coordinates and standard crossovers fail to transfer building blocks. In the works studied there were no direct comparisons of the relative and absolute encodings. If alternative representations based on internal coordinates are to be researched or used in heuristics (i.e. G.R.A.S.P., Hill Climbers, etc.), then our results supports the use of the relative encoding over the absolute one. Although the modified energy potential did not provide improved optimization performance in our experiments, we suspect that for longer proteins the difference between the two formulations will be more clear. Furthermore, we believe that a modified energy formulation will be particularly important for the effective use of hybrid GAs that use a local search method. Without a modified energy potential, there will exist large "plateaus" in the energy landscape on which local search cannot find a descent direction and where it must effectively perform a random search.

Acknowledgements

N. Krasnogor and D. Pelta are also at LIFIA, Universidad Nacional de La Plata, Facultad de Ciencias Exactas, Departamento de Informatica. David Pelta is supported by the COMETAS project of the ALFA program of the European Commision-DG-I. This work was supported in part by Sandia National Laboratories, a multiprogram laboratory operate by Sandia corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

References

[1] B. Berger and T. Leight. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J. Comp. Bio.*, 5(1):27–40, 1998.

[2] L. T. Biegler, T. F. Coleman, A. R. Conn, and F. N. Santosa, editors. *Large-Scale Optimization with Applications. Part III: Molecular Structure and Optimization*, volume 94 of *The IMA Volumes in Mathematics and its Applications*. Springer-Verlag, New York, 1997.

[3] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. *J. Comp. Bio.*, 5(3):409–422, 1998.

[4] K. A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24:1501, 1985.

[5] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding: a perspective from simple exact models. *Prot. Sci.*, 4:561–602, 1995.

[6] W. E. Hart and S. Istrail. Lattice and off-lattice side chain models of protein folding: Linear time structure prediction better than 86% of optimal. *J. Comp. Bio.*, 4(3):241–259, 1997.

[7] M. Khimasia and P. Coveney. Protein structure prediction as a hard optimization problem: The genetic algorithm approach. In *Molecular Simulation*, volume 19, pages 205–226, 1997. (to appear).

[8] N. Krasnogor, D. Pelta, P. M. Lopez, P. Mocciola, and E. de la Canal. Genetic algorithms for the protein folding problem: A critical view. In C. F. E. Alpaydin, editor, *Proceedings of Engineering of Intelligent Systems*. ICSC Academic Press, 1998.

[9] P. M. Pardalos, D. Shalloway, and G. L. Xue, editors. *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, volume 23 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, Providence, Rhode Island, 1996.

[10] A. Patton, W. P. III, and E. Goldman. A standard ga approach to native protein conformation prediction. In *Proc 6th Intl Conf Genetic Algorithms*, pages 574–581. Morgan Kauffman, 1995.

[11] A. Piccolboni and G. Mauri. Application of evolutionary algorithms to protein folding prediction. In N. e. a. Kasabov, editor, *Proceedings of ICONIP '97*. Springer, 1998. (to appear).

[12] A. A. Rabow and H. A. Scheraga. Improved genetic algorithm for the protein folding problem by use of a cartesian combination operator. *Protein Science*, 5:1800–1815, 1996.

[13] R. Unger and J. Moult. A genetic algorithm for three dimensional protein folding simulations. In *Proc 5th Intl Conf on Genetic Algorithms*, pages 581–588. Morgan Kaufmann, 1993.

[14] R. Unger and J. Moult. Genetic algorithms for protein folding simulations. *Journal of Molecular Biology*, 231(1):75–81, 1993.