Analyzing energy landscapes for folding model proteins

Graham A. Cox and Roy L. Johnston[a)]
School of Chemistry, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom

(Received 22 December 2005; accepted 29 March 2006; published online 31 May 2006)

A new benchmark 20-bead *HP* model protein sequence (on a square lattice), which has 17 distinct but degenerate global minimum (GM) energy structures, has been studied using a genetic algorithm (GA). The relative probabilities of finding particular GM conformations are determined and related to the theoretical probability of generating these structures using a recoil growth constructor operator. It is found that for longer successful GA runs, the GM probability distribution is generally very different from the constructor probability, as other GA operators have had time to overcome any initial bias in the originally generated population of structures. Structural and metric relationships (e.g., Hamming distances) between the 17 distinct GM are investigated and used, in conjunction with data on the connectivities of the GM and the pathways that link them, to explain the GM probability distributions obtained by the GA. A comparison is made of searches where the sequence is defined in the normal (forward) and reverse directions. The ease of finding mirror image solutions are also compared. Finally, this approach is applied to rationalize the ease or difficulty of finding the GM for a number of standard benchmark *HP* sequences on the square lattice. It is shown that the relative probabilities of finding particular members of a set of degenerate global minima depend critically on the topography of the energy landscape in the vicinity of the GM, the connections and distances between the GM, and the nature of the operators used in the chosen search method. © *2006 American Institute of Physics*. [DOI: 10.1063/1.2198537]

I. INTRODUCTION

One of the most important problems in chemical biology is to establish or predict the three-dimensional local spatial arrangement ("secondary structure") and folded conformation ("tertiary structure") adopted by a protein molecule from knowledge of its primary structure: the one-dimensional sequence of amino acid residues.[1] This sequence-structure correlation is of critical importance if we are to understand how proteins fold and, hence, to investigate sequence-activity relationships for proteins. The "protein folding problem" is essentially a search for the biologically active (functional) conformation of a protein (the so-called native state) for a given sequence of amino acid residues.[2] The ability of natural proteins to fold reliably to a unique native state has been attributed to the presence of a "folding funnel" on the folding free energy landscape, so that misfolded states are funneled towards the native state.[3] As well as determining the low energy protein conformations, therefore, it is important to discover the nature of the folding energy landscape (funnels, heights of potential barriers, etc.) in order to gain a better understanding of the dynamics of protein folding.[4]

There are a variety of protein models which differ in the way in which they approximate the protein molecule and how they treat interactions between amino acid residues and solvents (if included). Due to the size and complexity of protein hypersurfaces, simplified models have often been employed to study the protein folding process.[5] One of the simplest protein models is the *HP* lattice bead model,[6–8] which is a minimalist model of a protein, representing the constituent amino acid residues by either hydrophobic (*H*) or polar (hydrophilic) (*P*) beads which lie on a two-dimensional (2D) or three-dimensional (3D) lattice: square and cubic lattices are most common, though other lattices have also been studied. Such coarse grained protein models can actually capture some of the important folding behavior of real proteins, and they have the advantage of being simple, so that their energies may be calculated quickly, making them good for systematic grid searches and for carrying out comparisons of different folding search algorithms.

In a previous study[9] using a genetic algorithm (GA) to find the global minimum (GM) structures for a number of benchmark *HP* sequences (ranging from 20 to 50 beads[10]) on a square lattice, we found that most of the sequences have multiply degenerate global minimum structures. (The GM energies and degeneracies of these sequences—along with the new sequence introduced for this study: *HP*-20a—are listed in Table I.) Subsequent studies have shown that, for benchmark sequences with multiply degenerate GM, our GA finds the degenerate GM often with significantly different probabilities.

The aim of the present study is to determine to what extent the differences with which the GA finds different, degenerate GM conformations depends on: (a) the topography of the potential energy hypersurface for model protein folding,[4] (b) the way in which the GA searches the surface, and (c) the nature of the GA operators utilized. In this way, we hope to obtain some insight into the interconnection between the folding landscape, the search algorithm, and the ease or difficulty in finding the global minima.

[a)]Author to whom correspondence should be addressed. FAX: +44-(0)121 414 4403. Electronic mail: r.l.johnston@bham.ac.uk

TABLE I. Benchmark *HP* sequences investigated in the present study (Ref. 10), including the new benchmark sequence *HP*-20a. The lowest energies found for these sequences are indicated by E(GM) and the GM degeneracies (restricted to the $+x, +y$ quadrant) by D(GM). E(GM) and D(GM) values in bold have been confirmed by systematic grid searching.

Name	Sequence	E(GM)	D(GM)
HP-20	*HPHPPHHPHPPHPHHPPHPH*	**−9**	**2**
HP-20a	*HPHHPPHPHPPHPHPHPHPH*	**−8**	**17**
HP-24	$HHPP(HPP)_6HH$	**−9**	**19**
HP-25	$PPHPP(H_2P_4)_3HH$	**−8**	**16**
HP-36	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$	−14	192
HP-48	$P_2H(P_2H_2)_2P_5H_{10}P_6(H_2P_2)_2HP_2H_5$	−23	285
HP-50	$H_2(PH)_3PH_4P(HP_3)_3P(HP_3)_2HPH_4(PH)_4H$	−21	370

II. METHODOLOGY

A. The *HP* lattice bead model

In the present work, we have adopted the 2D square lattice *HP* bead model,[6,8] where the *H* and *P* beads are constrained to lie on a 2D square lattice and interactions occur only between nonbonded beads that lie adjacent to each other on the lattice ("topological neighbors"), but are not adjacent in the sequence (i.e., they are not directly bonded "sequence neighbors").[6] The values of the *H-H*, *H-P*, and *P-P* interactions (ϵ_{ij}) in the standard *HP* model are[6]

$$\epsilon_{HH} = -1.0, \quad \epsilon_{HP} = 0.0, \quad \epsilon_{PP} = 0.0. \tag{1}$$

The energy of the model protein is obtained by summing over these local interactions as follows:

$$E = \sum_{i<j} \epsilon_{ij}\Delta_{ij}, \tag{2}$$

where

$$\Delta_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are topological neighbors,} \\ & \text{but are not sequence neighbors} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

It should be noted that the effective attractive (stabilizing) interaction between the *H* beads reflects the fact that in aqueous solution the hydrophobic interaction (i.e., the repulsion of hydrophobic residues and water molecules) is the driving force for protein folding and that the native structures of many proteins are compact, with cores which are relatively rich in hydrophobic residues.[6,11] The reasons for studying the 2D, rather than the 3D lattice bead model are twofold:[6] first, the surface-to-volume ratio of the 2D model approaches realistic "protein values" for smaller sequences than in 3D; and second, the computational requirements are greatly reduced. The 2D analogs of protein secondary structure features, such as α helices and β sheets, naturally arise in the compact cores of such models, implying that secondary structure formation is a consequence of the compactness of the core and the presence in the core of hydrophobic groups.[12]

In this work, we define the folding conformation of the protein using a local coordinate system in which the position of a bead j is defined relative to its predecessors ($j-2$ and $j-1$).[11,13–15] Thus, in two dimensions, the direction of the bond joining the $(j-1)$th and jth beads can be left (0), right (1), or straight ahead (2) relative to the bond joining the $(j-2)$th and $(j-1)$th beads. Each protein conformation is therefore represented by a conformation vector **c**, which is a string of 0's, 1's, and 2's. As the energy of each conformation is invariant to the rotation of the whole molecule, we fix the positions of the first two beads in the chain, such that bead 1 lies at the origin (0,0) and bead 2 lies along the x axis (1,0). Thus, for an N-bead sequence, **c** has $(N-2)$ elements.

B. The genetic algorithm

Despite the reduction in complexity inherent in the minimalist *HP* lattice bead model, it has been shown to belong to the set of problems that are "NP-hard."[13,16] This means that there is no algorithm that can solve the protein folding problem exactly in polynomial time. For this reason, researchers have adopted heuristic and approximation algorithms. For the *HP* lattice bead model and other minimalist models, the approaches adopted include Monte Carlo,[17–21] chain growth algorithms,[22,23] simulated annealing,[24] genetic algorithms,[9,13–15,25–27] and ant colony optimization.[11,28–30]

Our GA program, its parameters, and operators have been described in detail previously,[9] so only a brief description is presented here.

1. Generating the initial population

The initial population corresponds to the starting set of individuals which are to be evolved by the GA. In our GA, the individuals are a set of conformation vectors (strings of 0's, 1's, and 2's, as described above). The initial population is formed by the constructor operator, which generates a number of valid conformations at random. In lattice bead models, valid protein conformations correspond to self-avoiding walks on the 2D or 3D lattice. In contrast, invalid/infeasible conformations correspond to non-self-avoiding walks, where two or more beads occupy one or more sites on the lattice. This is clearly unphysical, and such conformations should be eliminated. We have adopted a "recoil growth" algorithm, which involves growing the chain one bead at a time, checking the validity of the incomplete conformation at each step, and backtracking when an invalid subconformation is generated.[11,31]

2. Fitness

In our GA, the fitness of the ith individual (conformation), which determines the likelihood of it surviving and taking part in crossover, is simply related to its energy as follows:

$$F_i = -E_i + 0.01. \tag{4}$$

Thus, the fitness is a positive quantity, with high fitness corresponding to a large negative energy.

3. Selection

Selection refers to the way in which individual members of the population are chosen to pass into a temporary "parent population," which is subsequently subjected to a number of

genetic operators. We have adopted roulette wheel selection, whereby individuals are chosen for crossover with a probability proportional to their fitness.

4. Crossover

Crossover is the way in which "genetic" information from two parent structures is combined to generate "offspring." In this study, the variable mating rate is defined as the percentage of parents in the parent population which undergo crossover. The two offspring produced from each crossover operation overwrite their parents. The offspring and unmated parents then pass into the "offspring population." In this study, we have considered one-point (1pt) and two-point (2pt) crossovers, whereby the conformation vectors corresponding to the two selected parents are cut at either one or two points and complementary portions are exchanged to produce the offspring.

5. Mutation

While the crossover operation leads to a mixing of genetic material in the offspring, no new genetic material is introduced. The GA mutation operator helps to increase population diversity by introducing new genetic material, using the following mutation operators:[9]

In-plane Rotation=single-point mutation. This involves a ±90 or 180° rotation, in the xy plane, of the subchain following a randomly selected bond (say, between beads $j-1$ and j). In terms of the conformation vector, this corresponds to a change of the local coordinate direction of bead $j+1$, with the rest of the conformation vector being unchanged, i.e., a single bit change. This mutation, therefore, leaves most of the local structure intact.

Out-of-plane rotation. This involves a 180° rotation, in either the xz or the yz plane, of the subchain following a randomly selected bond (say, between beads $j-1$ and j). [The rotation plane depends on whether the $(j-1)-j$ bond points along the x or the y axis.] In terms of the conformation vector, this corresponds to all of the 0's being changed to 1's and all of the 1's being changed to 0's (with the 2's left unchanged) for the entire subchain starting at bead $j+1$. This mutation, therefore, leads to an inversion of the rotated fragment, thereby generating a diastereoisomer of the original conformation.

Crank shaft rotation. This involves a 180° rotation, in either the xz or the yz plane, of a crank shaft local structure motif (corresponding to the four digit strings…0110…or…1001…in the conformation vector), which leads to the interconversion of these four digit strings, with the rest of the conformation vector left unchanged.

Kink motion. This involves the inversion of a kink (or bend) local structure motif, where the kink bead (say, bead j) is moved diagonally across a lattice square, such that it is still bonded to its two neighbors (beads $j-1$ and $j+1$). This only leads to a change of the local coordinate directions of the $(j-1)-j$, $j-(j+1)$, and $(j+1)-(j+2)$ bonds, with the rest of the conformation vector left unchanged.

Snake motion. This involves the movement of the end of the protein to a neighboring vacant lattice site (if available), with each of the remaining beads moving to the position of its predecessor. This is analogous to the process of reptation in polymers and is one way in which a dense structure can be mutated with a low likelihood of creating an invalid mutant. In terms of the conformation vector, this mutation corresponds to shifting the vector along by one place and placing the first component of the vector at the end.

The variable mutation rate is defined as the probability of a selected individual undergoing mutation.

6. The corrector operator

Since the mutation operator often generates invalid (non-self-avoiding) conformations, a correction operator has been introduced to generate valid conformations from any invalid conformations resulting from mutation. Our corrector operator is based on the approach introduced by Schmygelska and Hoos in their ant colony optimization study of protein folding for the *HP* bead model.[28] An invalid conformer can undergo refolding at points of infeasibility (i.e., where two beads lie on top of each other), ensuring that a valid conformer results. The operator starts at the first nonfixed bead and cycles through the conformer placing beads using their corresponding value in the conformation vector. If the placement of the jth bead results in an infeasibility, the bead is randomly repositioned to a valid site; if the bead cannot occupy a valid site, the operator returns to the $(j-1)$th bead and attempts a valid repositioning. The operator continues in this fashion, backtracking as much as necessary until a valid conformation vector is obtained, which is as closely related to the initial invalid conformer as possible.

7. Elitism

In the context of genetic algorithms, an "elitist strategy" corresponds to allowing the best individuals in a population to survive unchanged from one generation to the next, thereby ensuring that the best member of the population cannot get worse. In our GA, elitism is accomplished by specifying the fraction of the best individuals within the jth population which are to be appended to the mutant population, prior to the generation of the $(j+1)$th population.

8. "Natural" selection

In biological evolution the concept of the "survival of the fittest" (or best adapted to the environment) is a strong evolutionary driving force. In the case of a GA, although the selection is clearly not "natural," individuals (be they parents, offspring, or mutants) are likewise selected to survive into the next generation on the basis of their fitness (their quality with regards to the quantity being optimized). The GA program generally continues for a predetermined number of generations (each generation corresponding to a cycle of crossover, mutation, and elitism) or until some convergence

criterion is reached. In the calculations reported here, however, as we know the GM from previous grid searches, the GA program terminates once one of the degenerate GM structures has been found.

9. Duplicate predator

In recent work, we have extended the analogy between GAs and natural evolution by considering the use of "predators" to remove unwanted individuals or traits from a population.[32] In our protein folding GA studies, we have investigated the application of a "duplicate predator," which deletes ("predates") identical conformations.[9] We define the duplicate predator limit (DPL) to be the maximum number of times that a given structure is allowed to appear in the population in any particular generation. It has previously been shown that DPL=1 yields the highest success rates.[9] The duplicate predator serves to increase the diversity (proportion of unique structures) of the population in order to prevent premature convergence of the population on a nonoptimal solution ("stagnation").

10. Local search

In problems where the search space is continuous, offspring and mutants invariably occupy states which are not minima, but which lie within an energy well. In such cases, performing a local minimization will relax each individual to its corresponding local minimum. Although, due to the discrete nature of the conformation space of the *HP* lattice bead model, it is not possible to perform gradient-driven energy minimizations, it is possible to perform a local search whereby a given conformation undergoes a number of folding changes, testing a number of closely related conformations.

In this study, we have implemented local searching using the "long range move" Monte Carlo-type approach introduced by Schmygelska and Hoos,[28] though it should be noted that Unger and Moult also used a Monte Carlo mutation in their GA study.[13] In our application, a conformation c_1, with energy E_1, is folded at a randomly chosen position (as in the in-plane rotation mutation) by randomly changing one of the digits in the conformation matrix **c**. The new conformation c_2 is accepted if its energy $E_2 \leqslant E_1$. For conformation changes where $E_2 > E_1$, the conformational change is accepted with a probability

$$p = \frac{E_2}{15E_1}, \tag{5}$$

where the factor of 15 was found to give reasonable acceptance rates (approximately 25%). Each local search corresponds to 36 of these Monte Carlo steps, with a new random fold carried out at a random position each time, starting from the current conformation.

C. The *HP*-20a sequence

A new 20-bead benchmark sequence (*HP*-20a ={*HPHHPPHPHPPHPHPHPHPH*}) has been chosen for this study. This sequence has 17 degenerate, but nonsymmetry-equivalent, GM [with energy $E(\mathrm{GM})=-8$]. From Table I, it can be seen that *HP*-20a has the same *H-P* composition ($H_{10}P_{10}$) as the *HP*-20 benchmark sequence, being related to *HP*-20 by four *H-P* point mutations (*H-P* swaps) at loci (beads) 4, 6, 15, and 16.

The GM degeneracy of *HP*-20a is low enough for the GA to find all of the GM with reasonable probability, while being large enough to show significant variation between the GM. The sequence size chosen is small enough to allow systematic grid searching of all possible conformations, thereby ensuring that the GM energy and degeneracy are known precisely. The number of GM is restricted to 17 by fixing the first two beads in the chain at positions (0,0) and (1,0) and constraining the first bead that lies off the x axis to lie in the $+x,+y$ quadrant (though subsequent beads are allowed to fold around into the other quadrants). The 17 GM structures (labeled GM1-GM17) are shown in Fig. 1. The numerical order of the GM conformers follows the numerical order of their conformation vectors **c**, written as a series of 0's, 1's, and 2's.

Figure 1 shows that, although all the GM have eight topological *H-H* contacts (hence $E=-8$), not all of the ten *H* beads have to be involved in *H-H* interactions. Thus, GM3–GM11 and GM14 have one noninteracting *H* bead (bead 1, 3, 5, 9, or 14 may be noninteracting). In these cases, the other nine *H* beads compensate by forming additional *H-H* contacts.

In some of the studies below we have relaxed the constraint on the first off-axis bead, so that it can lie in either the $+x,+y$ or the $+x,-y$ quadrant. This results in 17 enantiomeric pairs (nonsuperimposable mirror image conformations) of GM which are related by reflection in the xz plane. Enantiomeric GM can be interconverted by exchanging 0's for 1's (and vice versa) in the conformation vector **c**, while keeping the 2's unchanged.

We define the sequence *HP*-20a′ {*HPHPHPHPHPPHPHPPHHPH*} as the reverse of sequence *HP*-20a. (The "forward" and "backward" sequence vectors are not identical as *HP*-20a is not a "palindromic" sequence.) The set of global minima for *HP*-20a′ are, of course, the same as those for *HP*-20a (being independent of which end of the chain is taken as the start of the sequence). For any particular GM of *HP*-20a, the isostructural GM of *HP*-20a′ is simply generated by reversing the conformation vector **c** and interchanging 0's and 1's.

In order to determine the success rate of the GA and the probabilities of it finding different GM conformers, the GA was run 20 000 times, with the program terminating when a conformer with the global minimum energy (as determined from the grid search) is found (up to a maximum of 200 generations). For successful GA runs, the first GM conformer found is reported. The GA parameter set is as follows: population size=200, maximum number of generations=200, crossover types=1pt and 2pt, crossover rate=1.0, mutation rate=0.1, elitism=0.3, Monte Carlo local search (LS) rate (when implemented)=100%, Monte Carlo steps=36, and duplicate predator limit=1.

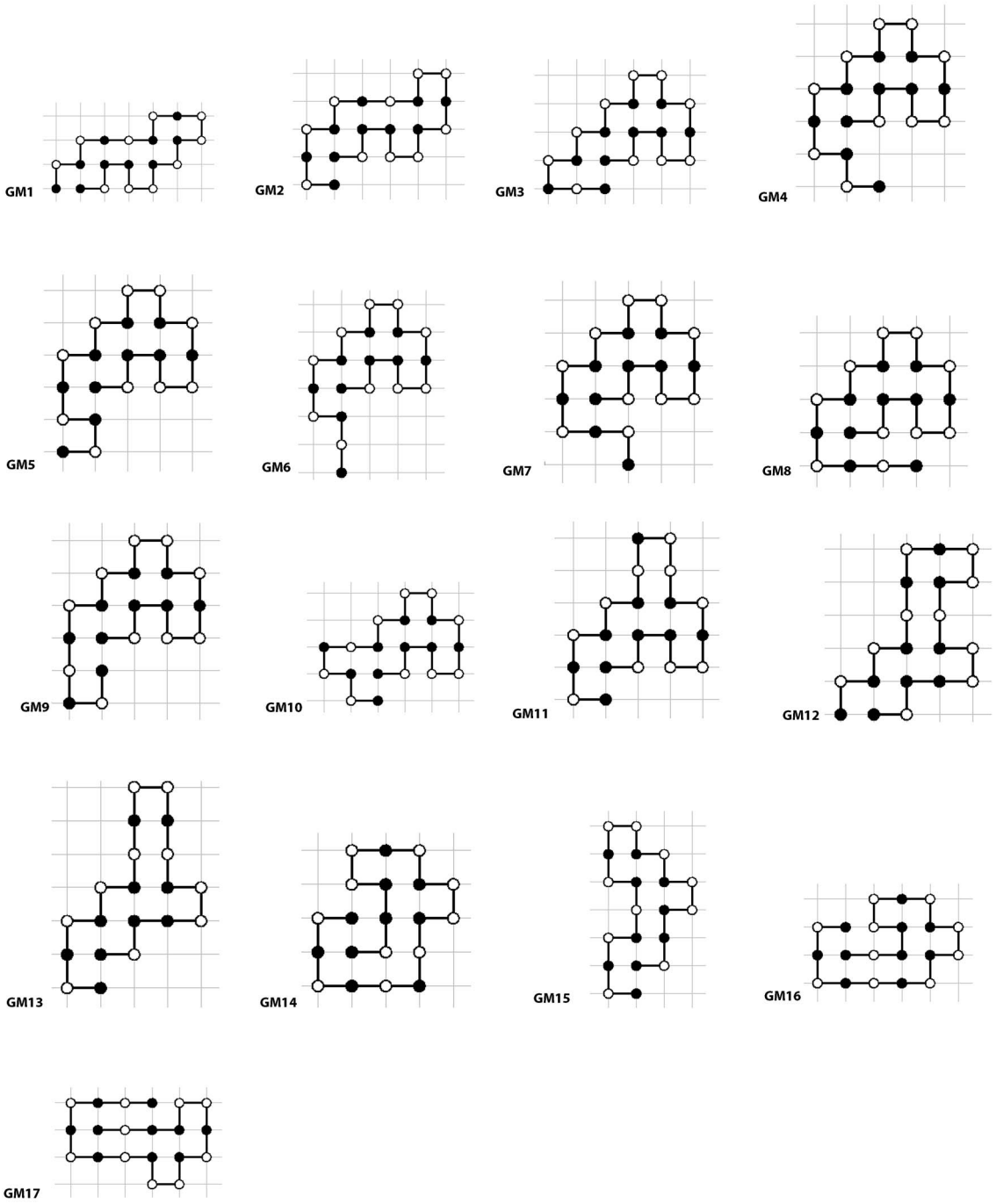

FIG. 1. GM conformations for *HP*-20a (black=*H*, white=*P*).

III. RESULTS AND DISCUSSION

A. GM probability distributions

1. Dependence on GA search methodology

The probability distribution of finding GM1–GM17 using the GA are shown in Fig. 2, which compares the results for 1pt and 2pt crossovers, with and without Monte Carlo LS. In each case, 20 000 GA runs were performed. In this study, initial and subsequent populations were allowed to sample both the $+x,+y$ and the $+x,-y$ quadrants in order to avoid possible problems due to crossover and mutation, generating offspring in the "unallowed" quadrant. (However, Fig. 2 only shows the relative probabilities of obtaining the various GM in the $+x,+y$ quadrant.) Our previous work has shown that 1pt crossover is generally more efficient than 2pt, because it leads to less disruption of "schemata" related to low energy structures.[9] For the *HP*-20a sequence, a comparison of 1pt and 2pt crossovers (with and without local searching) again shows that 1pt crossover has comparable success rates to 2pt, but generally with fewer structures having to be searched before finding one of the GM.

The distributions in Fig. 2 show that there is little difference between 1pt and 2pt crossovers as regards the relative

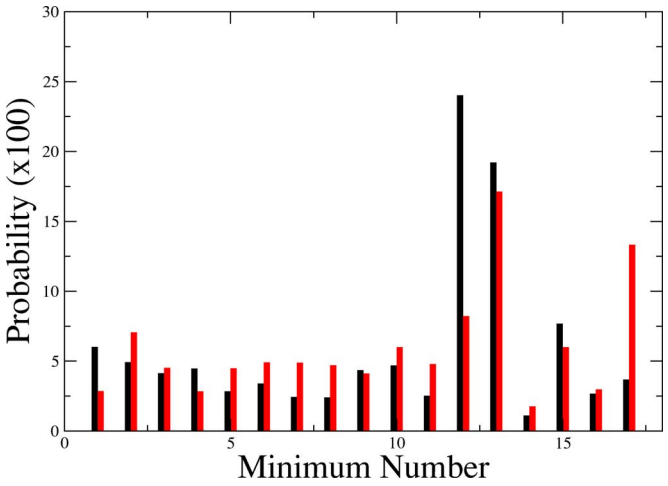

FIG. 2. Probability distribution of minima found by 20 000 GA runs for both 1pt with (red) and without (black) the incorporation of a Monte Carlo local search (LS).

probability of finding each global minimum in successful GA runs. What is apparent, however, is that, for both crossover types, some GM are found with considerably higher likelihood than others—with the maximum ratio between "likely" and "unlikely" GM being approximately 34:1 for GM12:GM14.

Figure 2 also shows that the introduction of Monte Carlo local searching leads to a significant change in the probability distribution. Thus, when the local search is not implemented, GM 12 and 13 are most likely to be found, whereas, with local search, minima 13 and 17 are the most probable. The change in probability distribution when local search is implemented shows that the distributions depend to a certain extent on how the GA operations search the energy surface.

2. Mirror image GM

Figure 3 shows the probability distribution for all 17 pairs of mirror image GM, obtained from 20 000 GA runs with 1pt crossover and no local search. The figure clearly shows that the differences in the probability of finding enantiomers are statistically insignificant when compared to differences between nonenantiomeric minima. Similar results are found for 2pt crossover and when local search is implemented.

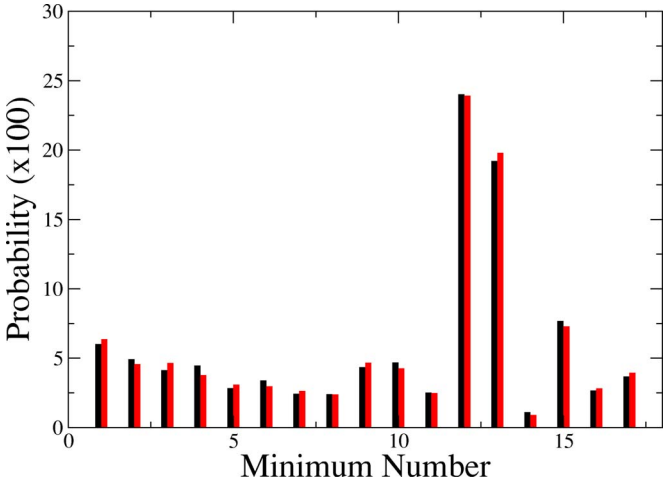

FIG. 3. Probability distribution of enantiomeric pairs of GM found by 20 000 GA runs for a 1pt crossover, without incorporating any local search, for *HP*-20a. Red $(+x,+y)$ quadrant, black $(-x,+y)$ quadrant.

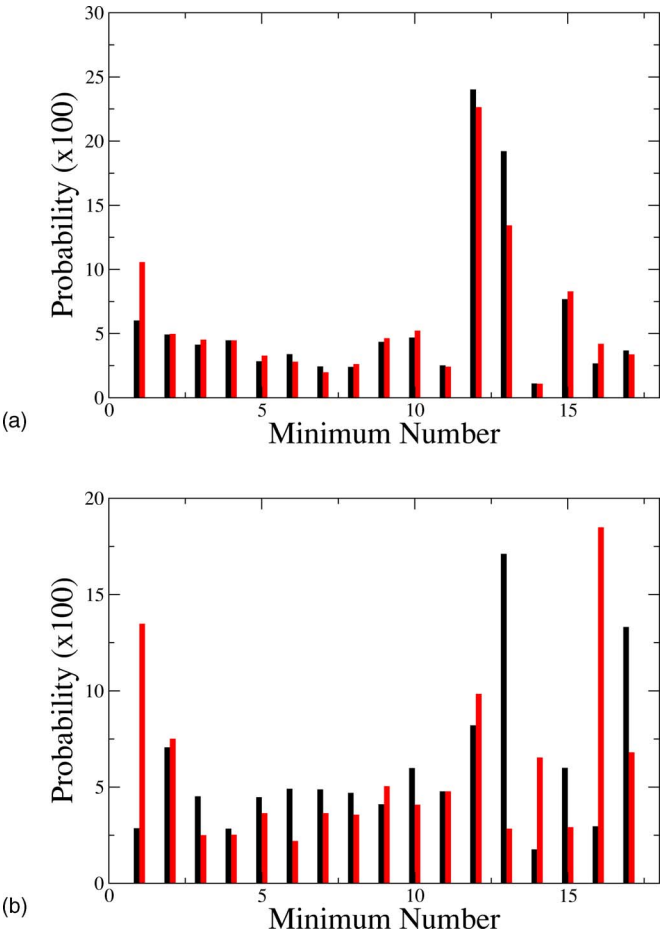

FIG. 4. Probability distribution of minima found by 20 000 GA runs using 1pt crossover, (a) without and (b) with the incorporation of a Monte Carlo local search (LS) for both *HP*-20a (black) and *HP*-20a′ (red).

3. Sequence reversal

The GM probability distributions for the forward (*HP*-20a) and backward (*HP*-20a′) sequences (using 1pt crossover) are compared in Fig. 4, with and without local search, for 20 000 GA runs. (Again, although the search was not restricted, only the relative probabilities of finding GM in the $+x,+y$ quadrant are shown, with the probabilities for the mirror images being almost identical.) Isostructural GM of *HP*-20a and *HP*-20a′ are paired together, though in the following discussion they are distinguished as GM1–GM17 (when the GA was performed on sequence *HP*-20a) and GM1′–GM17′ (when the GA was performed on sequence *HP*-20a′).

It is immediately apparent from Fig. 4 that the *HP*-20a′ distribution without local search is similar to that of *HP*-20a, but with differences for some conformers (especially GM1/GM1′, GM13/GM13′, and GM16/GM16′) which are clearly greater than those observed between mirror images. It has already been shown that implementing a local search leads to a significant change in the probability distri-

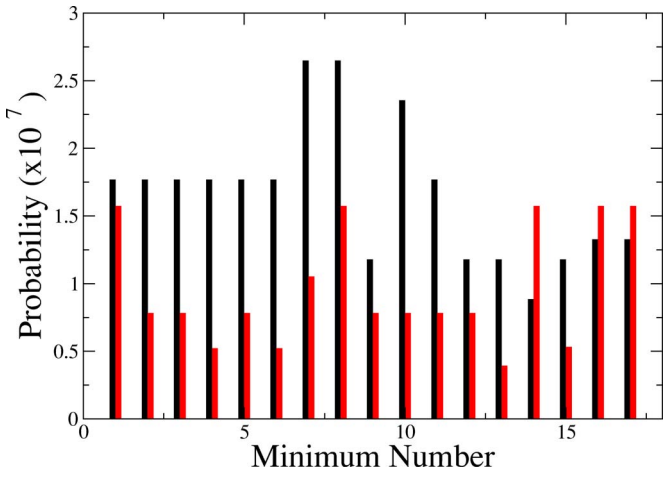

FIG. 5. Theoretical constructor probabilities for the *HP*-20a (black) and *HP*-20a′ (red) GM.

bution for the GM of *HP*-20a. Figure 4 shows that this is also true for the reverse sequence *HP*-20a′ and, more importantly, that the differences between the probability distributions of isostructural GM are generally greater when a local search is implemented. Thus, GM1′, GM14′, and GM16′ are found with significantly increased probability relative to GM1, GM14, and GM16. Conversely, GM13′, GM15′, and GM17′ are found significantly less frequently than GM13, GM15, and GM17. Considering conformers GM13′–GM17′, Fig. 1 shows that these structures have embedded tail *H* beads, which contribute to the overall energy of the structure. Once the tail becomes embedded and the correct outer structure has been formed, there is a set local structure that the tail must form in order to remain feasible, which the correction operator within the local search algorithm will find.

These initial studies show that there are significant differences between GM probability distributions depending on whether or not local searching is included. It has also been shown that reversing the sequence can lead to significant changes in the GA distribution. In the following sections, we rationalize the above distributions in terms of the nature of the GA and of the model protein folding surface.

B. The constructor probability

There are two occasions within a GA run when conformations are generated from scratch: (a) the generation of the initial population and (b) generating new structures to replace those removed by the duplicate predator. In both cases, new structures are generated using the constructor operator, which uses a recoil growth algorithm[11,31] to build up a structure, one bead at a time, backtracking when necessary.

This type of growth results in high probabilities of construction (P_c) for more compact structures. $P_c=\Pi_{i=1}^{N}P_i$, where $P_i=\frac{1}{3}$ if all local coordinates for the next bead are feasible, $\frac{1}{2}$ if only two are feasible, and 1 when there is only one feasible local coordinate.

The construction probabilities for the GM structures of sequences *HP*-20a and *HP*-20a′ (compared in Fig. 5) are different, as the feasible chain growth directions at each step will differ depending on which end of the chain is defined as the origin. In contrast, enantiomeric GM have identical construction probabilities.

It is interesting to note that, with the exception of GM14/14′, GM16/16′, and GM17/17′, the GM for the reverse sequence (*HP*-20a′) have lower constructor probabilities than those (with the same structure) for *HP*-20a. Inspection of the GM structures in Fig. 1 reveals that for *HP*-20a′, these exceptional GM have longer embedded tails, which results in a higher construction probability.

Comparison of Fig. 5 with the GM probability distributions obtained from GA runs (with or without local search, Fig. 4) shows that there is no correlation between the construction probability distribution and the GA probability distribution. In the absence of local searching, because the average number of generations required to find one of the GM conformations (54.5) is high, any biasing of the initial population arising from the constructor is altered significantly by the GA operations before the GM is found. This can be

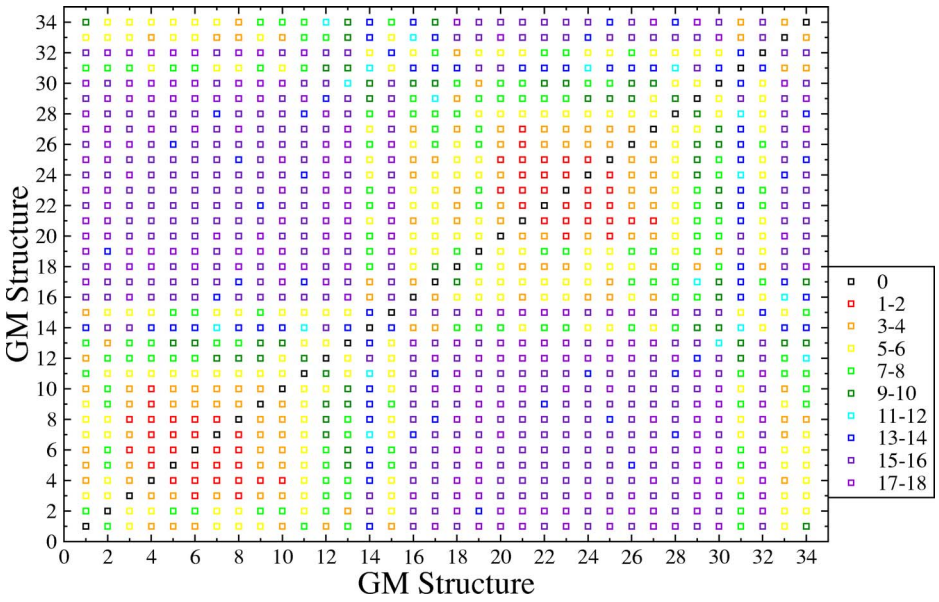

FIG. 6. Hamming distances between all GM (including mirror images, 18–34) for the *HP*-20a sequence. GM(N+17) is the mirror image of GM(N), with N=1–17.

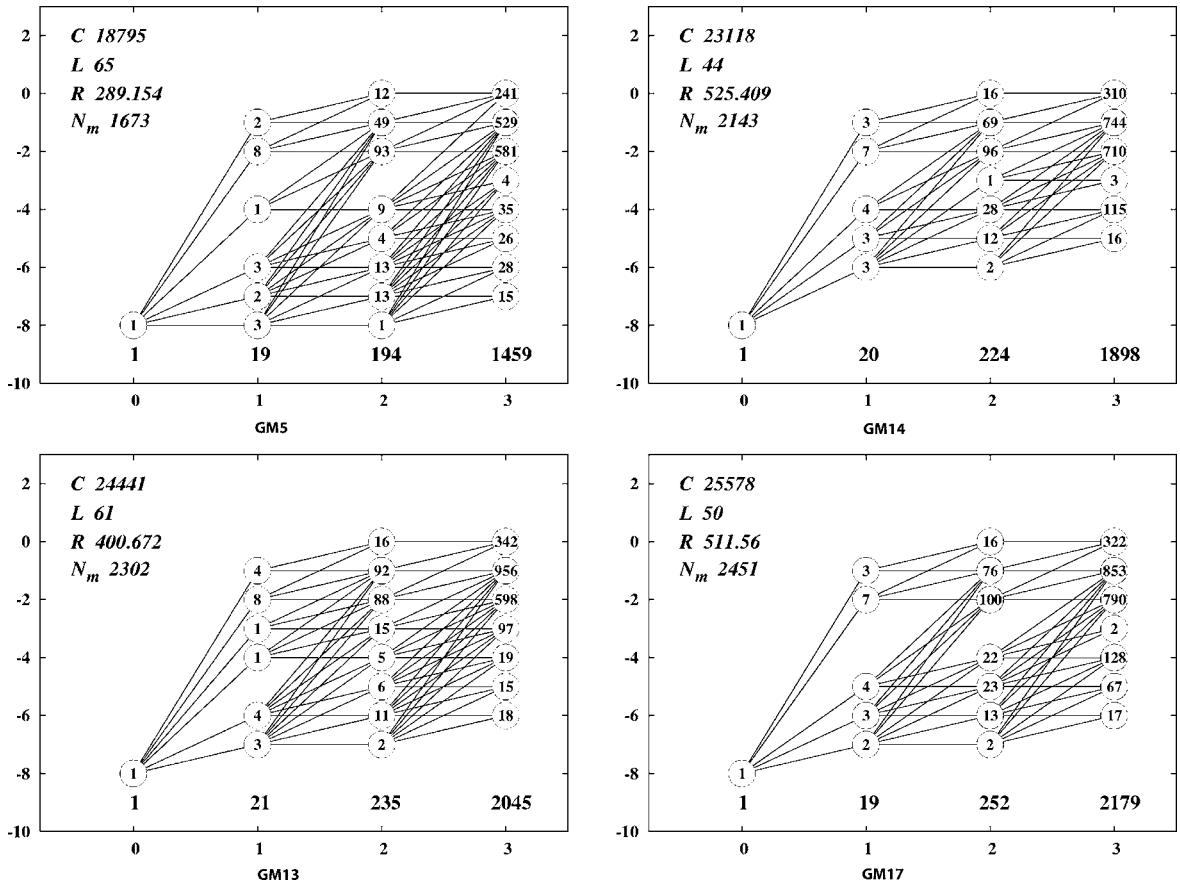

FIG. 7. Valid uphill pathways between a starting GM structure and structures three folds away, using the point mutation move class, for selected GM of *HP*-20a. *x* axis: number of point mutations; *y* axis: energy.

seen by analyzing a smaller sequence (e.g., *HP*-10a $=\{PHPPHHPPHH\}$) which has higher construction probabilities for the GM. For this sequence, we have found that the GM distribution found by the GA within the first few generations is virtually identical to that from the constructor probabilities. However, for longer searches the similarity is lost. Incorporating local searching for *HP*-20a lowers the average number of generations required to find a GM (3.7), but the local search operator itself changes the distribution of structure types significantly from that generated by the constructor.

This investigation shows that, while one of the GM might not be generated by the constructor operator, one of the many structurally related higher energy structures (with similar individual construction probabilities) may be constructed. If the GA finds the GM from this initial population within a small number of steps (which is possible for short sequences), then the GM distribution pattern will be similar to the theoretical construction probability distribution. For longer GA runs (before successfully finding a GM conformation), which will tend to be the case for longer sequences, the other GA operators (especially crossover and mutation) will skew the distribution from that of the constructor.

C. Hamming distances between global minima

The Hamming distance (d_H) is a simple measure of the dissimilarity of two structures, represented as bit strings.[33] For lattice bead models, using a local coordinate system, the Hamming distance between two conformations c_i and c_j is simply the number of positions in the conformation vectors c_i and c_j which have different values (local coordinate directions). Structures which are closely related, i.e., which are in close proximity on the conformational hypersurface, have a low d_H. The Hamming distances between all 17 enantiomeric pairs of GM for *HP*-20a are shown graphically in Fig. 6. GM1–GM17 are defined such that the first off-axis bead lies in the $+x,+y$ quadrant, while GM18–GM34 have the first off-axis bead in the $+x,-y$ quadrant. The numbering of the mirror images is such that GM(N+17) is the mirror image of GM(N).

Figure 6 highlights that conformations GM3–GM10 have very similar conformation vectors ($d_H \leqslant 4$). As can be seen from Fig. 1, in this family of GM, the first 16 beads adopt the same subchain conformation. For GM3–GM10, the first 16 beads are responsible for all but one of the favorable *H-H* interactions and there is a flexible $\{PHPH\}$ tail which folds round in different loops to form the final *H-H* interaction.

GM12 and GM13 have relatively high Hamming distances ($d_H > 6$) from most of the other global minima (including the mirror images), therefore, assuming the GA searches all low energy areas with equal probability, the lower competition for which global minimum the GA evolves to once in the region of GM12 and 13 could explain

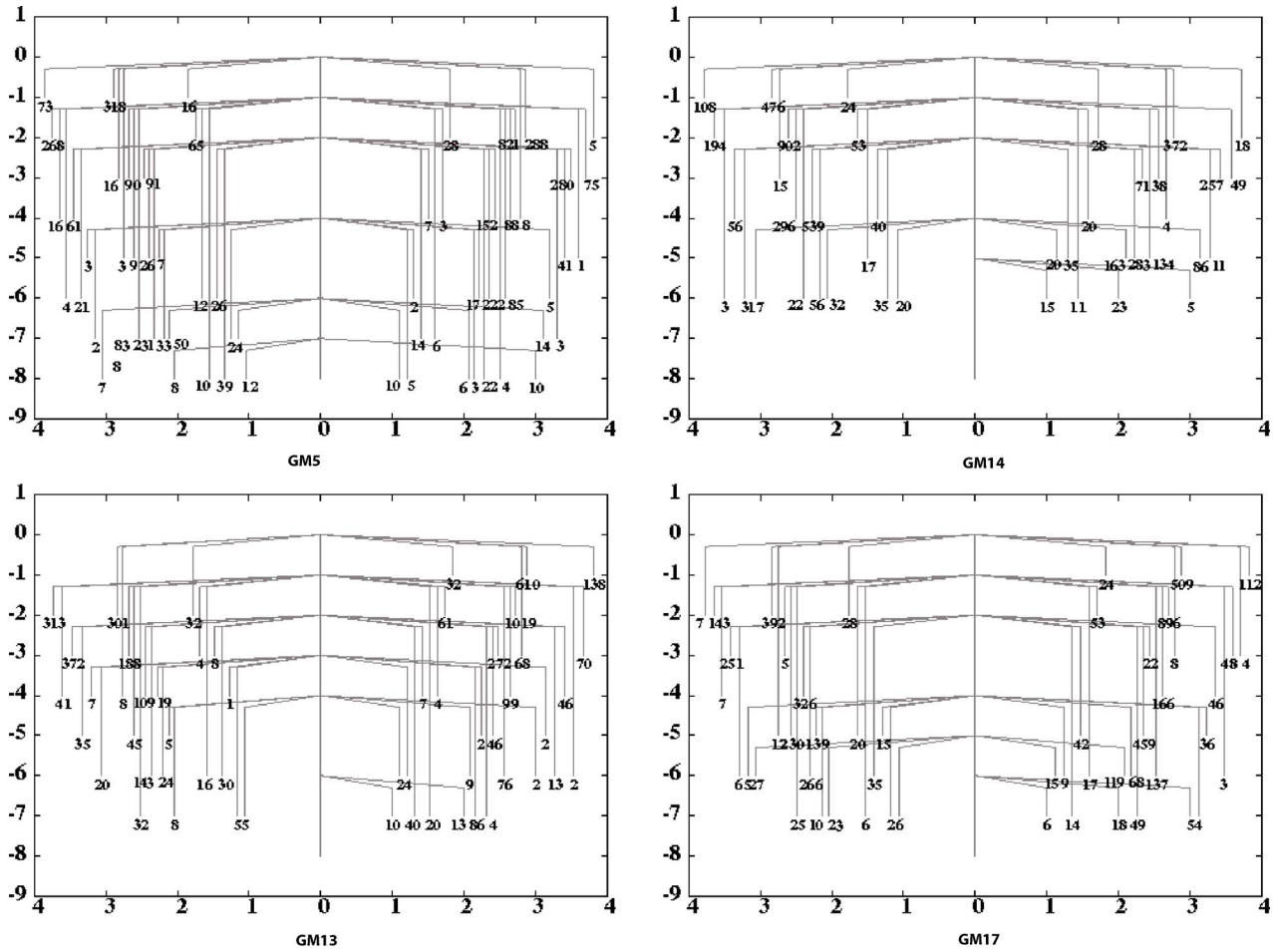

FIG. 8. Disconnectivity graphs for selected GM of *HP*-20a. *x* axis: Hamming distance from starting GM structure; *y* axis: energy.

the high probabilities with which these GM are found, and also the relatively low probabilities associated with the aforementioned family GM3–GM10.

D. Connectivity of global minima

To answer whether or how strongly the GA probability distribution depends on the topography of the energy land-

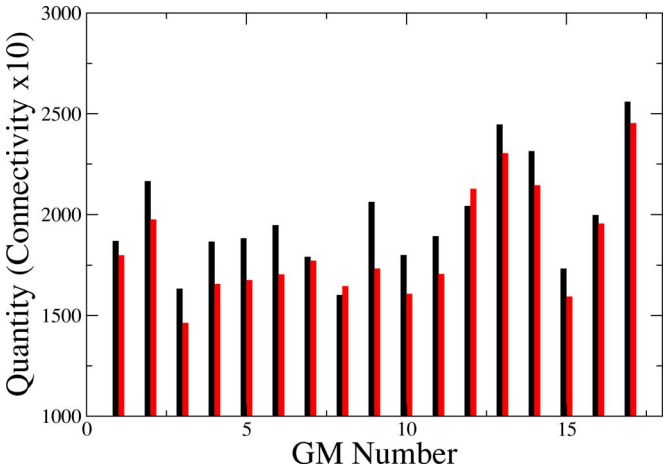

FIG. 9. Distribution of connectivity (C, black) and number of minima (N_m, red) for all valid paths.

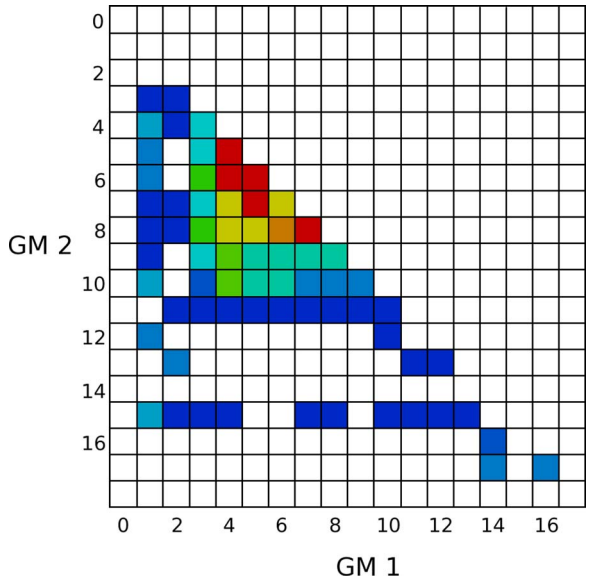

FIG. 10. Number of structures which are common for three fold paths starting at each GM, for all GM pairs. There are 20 different colors, each of width 30, in a linear scale from white (0 common structures) through dark blue, light blue, light green, yellow, orange, and red (583 common structures).

scape and the way in which the genetic operators act, the energy surface in the vicinity of each GM has been investigated. The movement class is defined as a point mutation, i.e., picking the kth element of the conformation vector $\mathbf{c}$ and changing its value. The mutant structure therefore has a Hamming distance, from the starting structure, of 1.

From each GM, all valid uphill/flat paths (i.e., sequences of point mutations which, starting from the GM, either stay constant or go up in energy, while at the same time not resulting in or passing through an infeasible structure) of three point mutation steps in length were calculated. The only restriction imposed was that, after a point mutation has been implemented, the reverse of that mutation cannot be performed in the next step. These represent the pathways most likely to be explored by the local search operator, which always accepts a move if the new fitness is less than or equal to the parent's fitness (though uphill moves are accepted probabilistically). These paths are represented in Fig. 7 for a selection of GM. The starting GM structure is represented in the lower left corner ($x=0$), with energy plotted on the y axis and the number of steps (mutations from the GM) on the x axis. The value at each node indicates the number of unique structures represented by that node. The total number of unique structures for all possible point mutations are indicated at the bottom of each column of nodes.

Other statistics calculated are links (L), the number of connections between all nodes; connectivity (C), the total number of connections between all nodes,

$$C = P \times F, \tag{6}$$

where P is the total number of paths and F is the maximum number of steps (largest d_H); the ratio L/C; and the number of minima (N_m), the total number of unique structures (conformations) represented on the figure.

To investigate the broader topography of the surface surrounding the GM, all valid paths (i.e., paths that can also decrease in energy as a result of a mutation) were also analyzed to find any paths wherein a higher energy "saddle point" structure separates two lower energy minima, one of which is a GM (i.e., defining minimum-saddle-minimum transitions, where the first minimum is one of the GM and the other minimum is a low lying structure). Using these transitions, a modified disconnectivity graph[34] can be created, wherein the x axis shows the Hamming distance between minima in a transition. A selection of these disconnectivity graphs are shown in Fig. 8.

Both representations of local topography clearly highlight the interconversions between GM which are in close proximity on the surface, e.g., the aforementioned GM3–GM10 family. These GM conformations have a high interconversion rate within a few steps.

Within the GA, structures with lower energy are more likely to be chosen to participate in genetic operations and propagate into subsequent generations. The connectivity of structures with an energy of −7 (i.e., one energy unit above the GM energy) is high, suggesting that the GA will easily evolve to these structures. Once a population containing a large number of structures with energy −7 has been created, structures which are higher in energy are less likely to participate in the evolutionary process. Hence GM14, which has no connectivity to a structure with energy −7, as can be seen from both Figs. 7 and 8 (and which has the lowest number of links), is less likely to be found.

To investigate the effect of the topology on the GA distribution, the Connectivity (C) and number of minima (N_m) values have been extracted from each plot in Fig. 7 and are plotted as distributions in Fig. 9. This distribution has a very strong correlation with the GA probability distribution when the local search is applied (see Fig. 2), e.g, GM2, GM12, GM13, and GM17 have high C and N_m and are found most often by the GA. GM14 is an exception to this correlation, which may be attributed to the reasons discussed above and the lack of pathways linking GM14 to other GM. GM14 also has the lowest number of different types of transition, as shown in Fig. 8.

Figure 10 shows the number of structures which appear in the three-mutation pathways for both GM, for all possible pairs of GM. [For a pair GMi and GMj, if square (i,j) is white, this signifies that no structures that appear in the pathways leading from GMi also appear in the pathways leading from GMj.] The highly connected region for GM3—GM10, which was observed in the Hamming distance plot (Fig. 6), is again evident in Fig. 10, with these GM having the highest number of shared structures. The maximum number of common structures is 583, colored red in Fig. 10, between GM4 and GM6.

Figure 10 shows that only a few of the precursor structures that appear in the paths leading to GM14 are also present in paths leading towards other GM, and then only for GM16 and GM17. This is consistent with Fig. 6, which shows that GM16 and GM17 are closer in Hamming distance to GM14 than are any other GM. [The Hamming distances between these three GM are d_H(GM14−GM16) $=d_H$(GM16−GM17)=4 and; d_H(GM14−GM17)=3.]

Inspection of Fig. 1 reveals that GM14, GM16, and GM17 have the same local structure between beads 5 and 13. An obvious difference between GM14 and GM16/GM17 is that, while the latter two structures have embedded $H(1)$ heads, GM14 also has an embedded $H(20)$ tail. The steric requirement imposed by having both ends of the chain embedded (i.e., in four-coordinate sites) may contribute to the reduced probability of the GA finding GM14, though this would not explain the significantly enhanced probability of finding GM14' for the reverse sequence. Finally, it is interesting to note, as mentioned previously, that this family of related structures (GM14/GM14′, GM16/GM16′, GM17/GM17′) are the only conformations for which the constructor probabilities are greater for the reverse (*HP*-20a′) than the forward (*HP*-20a) sequence definition.

E. Previously studied benchmark sequences

The *HP* benchmark sequences that we have previously studied[10] (*HP*-20, *HP*-24, *HP*-25, *HP*-36, *HP*-48, and *HP*-50) are listed in Table I, along with the energies and degeneracies of the global minima for these sequences (restricting solutions to the $+x,+y$ quadrant).[9] (Although these are commonly studied benchmark sequences for the 2D *HP*

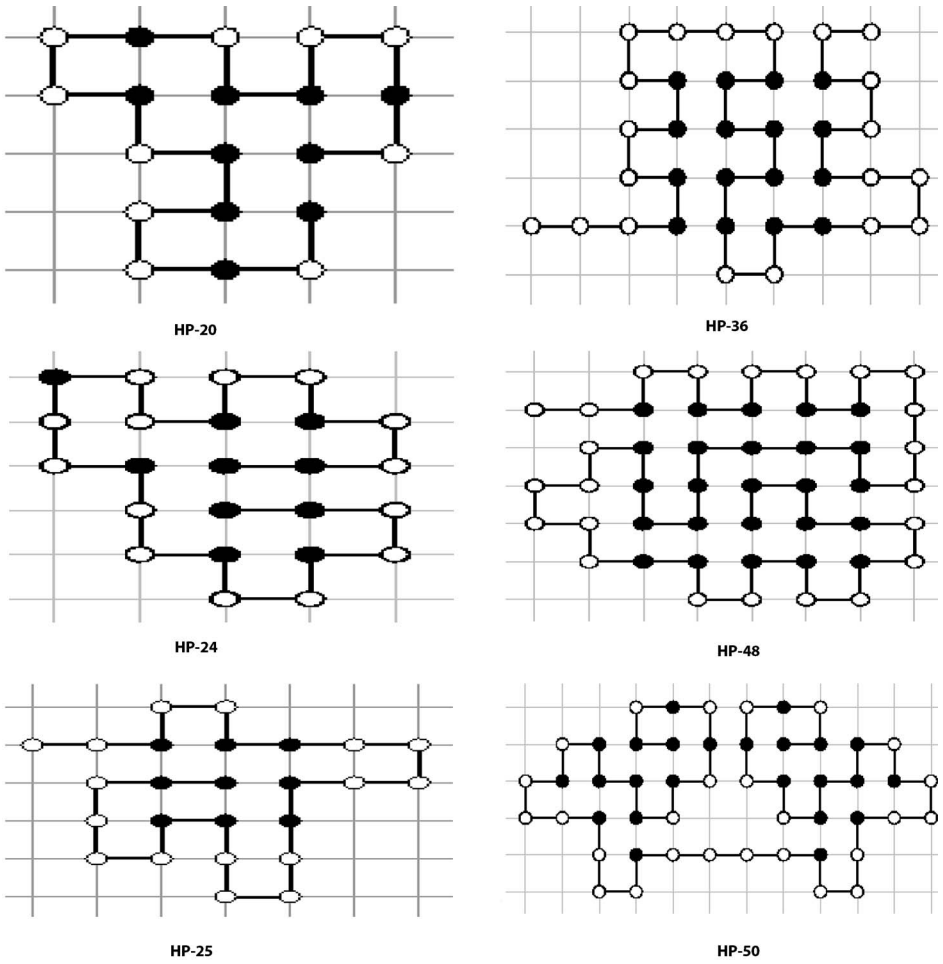

FIG. 11. Example GM structures for benchmark *HP* sequences.

lattice bead model, to our knowledge, no exploration of their folding landscapes have previously been made.) For the shorter sequences (up to *HP*-25), the GM energies and degeneracies are exact, as they have been obtained by systematic grid searching. For *HP*-36, *HP*-48, and *HP*-50, which are too large to grid search, the "GM" energies and degeneracies have been obtained from multiple GA runs. The energies are consistent with previous studies of these sequences.[9] The degeneracies have not previously been reported and should be regarded as lower bounds on the GM degeneracies for *HP*-36, *HP*-48, and HP-50. Examples of GM structures for the benchmark sequences are shown in Fig. 11.[9]

The GA percentage success rates and average number of structures evaluated (for successful runs) determined in our previous study of the benchmark sequences are summarized in Table II. The search space for the *HP* square lattice model grows as 3^{N-2}, where N is the sequence length. However, Table II shows that our GA program has most difficulty finding the GM for *HP*-36 and *HP*-48, which have compact hydrophobic cores (all of the degenerate GM having 4×4 and 5×5 square hydrophobic cores, respectively as shown in Fig. 11), while the GM of the longer *HP*-50 benchmark sequence (which does not have a compact hydrophobic core, tending to have more open GM structures, often with two clusters of hydrophobic beads) were found relatively easily (with much higher success rates and far fewer structure evaluations).[9] The increased degeneracy of the GM for *HP*-50, as compared with *HP*-36 and *HP*-48 (see Table I), does not explain this observation, as it should be swamped by the larger search space. The significantly higher number of structure evaluations required for *HP*-25, compared with *HP*-24 (an increase of 29%), may similarly be due to the 3×3 square hydrophobic core of the *HP*-25 GM. These results are generally consistent with an earlier GA study by Unger and Moult, though they appear to have required fewer structure evaluations for *HP*-25 than *HP*-24 and far more structure evaluations in the *HP*-50 case.[13]

TABLE II. Percentage success and average number of evaluations (for successful GA runs) for the benchmark *HP* sequences studied previously (Ref. 9).

Sequence	E(GM)	Success (%)	Ave. evaluations
HP-20	−9	100	18 338
HP-24	−9	100	27 278
HP-25	−8	100	35 128
HP-36	−14	70	113 667
HP-48	−23	13	261 311
HP-50	−21	100	97 691

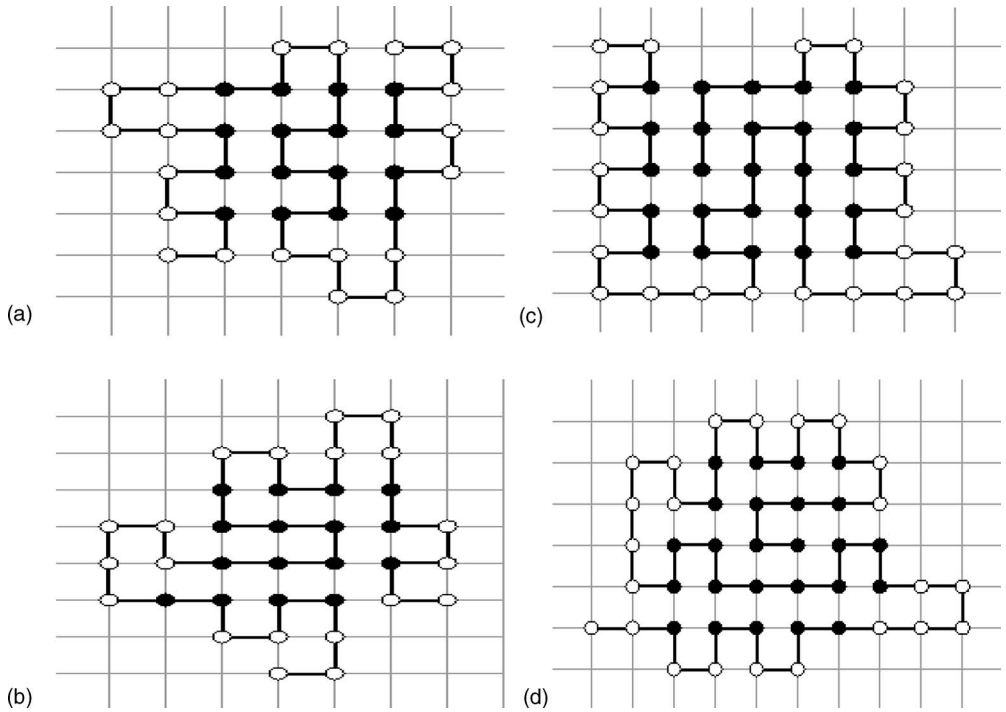

FIG. 12. Typical GM and GM+1 structures for *HP*-36 and *HP*-48: (a) GM *HP*-36, (b) GM+1 *HP*-36, (c) GM *HP*-48, and (d) GM+1 *HP*-48.

For *HP*-36 and *HP*-48, GA runs which are not successful [i.e., do not find a GM, with $E=E(GM)$] nearly always (*HP*-36 100%, *HP*-48 88%) find a metastable structure with energy $E=E(GM)+1$; we will refer to these structures as "GM+1" conformations. Figure 12 shows typical GM and GM+1 conformations for *HP*-36 and *HP*-48. It is apparent that many of the hydrophobic contacts within the maximally compact GM structures are between *H* beads that are far apart in the sequence. This means that a large number of changes in local folds are required to go from a GM+1 structure to a GM. It is, therefore, unlikely that a low energy pathway will exist that will allow a metastable GM+1 conformation to refold into a GM. Given the dense hydrophobic nature of the GM for these sequences, it is also likely that the local minima which are near in conformation space to the GM will have relatively high energies and will again be separated from the GM by high energy barriers. In the absence of a simple pathway between these structures and the GM, the GA frequently fails.

In order to test the above hypothesis, and based on our findings for the *HP*-20a sequence, we have calculated the valid uphill pathways (Fig. 13) and disconnectivity graphs (Fig. 14), involving up to three point mutation steps from the GM, for the GM of the 20–50-bead benchmark sequences. Due to the high GM degeneracies of the benchmark sequences, the valid pathways were calculated for a randomly chosen GM for each benchmark (Fig. 11). Similar plots were observed for other GM for each benchmark.

A number of observations can be made based on these figures:

(1) At one fold (point mutation) from the GM, only *HP*-50 has a structure 1 energy unit above that of the GM (GM+1). The final fold (leading to the GM) for all other benchmark sequences must, therefore, form two or more *H-H* contacts.
(2) *HP*-36 and *HP*-48 have a lower proportion of structures belonging to low energy (longer sequences have a larger possible set *A*).
(3) The theoretical number of paths is given by

$$[2 \times (S-2)]^F, \tag{7}$$

where *S* is the sequence length and *F* is the maximum number of folds (largest d_H). Comparing the theoretical number of paths with the connectivity (*C*) yields the percentage of valid paths, which are listed in Table III for a maximum of three and four point mutation steps. This shows a dramatic decrease in the proportion of valid paths for sequences *HP*-36 and longer, though, interestingly, the percentage is higher for *HP*-50 than for either *HP*-36 or *HP*-48.
(4) *HP*-50 appears to have better funneling towards the GM than either *HP*-36 or *HP*-48.

TABLE III. The percentage of valid paths for the *HP* benchmark sequences for a maximum of three and four point mutation steps.

Name	Three steps	Four steps
HP-20	15.2	8.9
HP-24	15.1	8.8
HP-25	15.8	9.5
HP-36	7.1	3.5
HP-48	5.1	2.4
HP-50	7.9	4.0

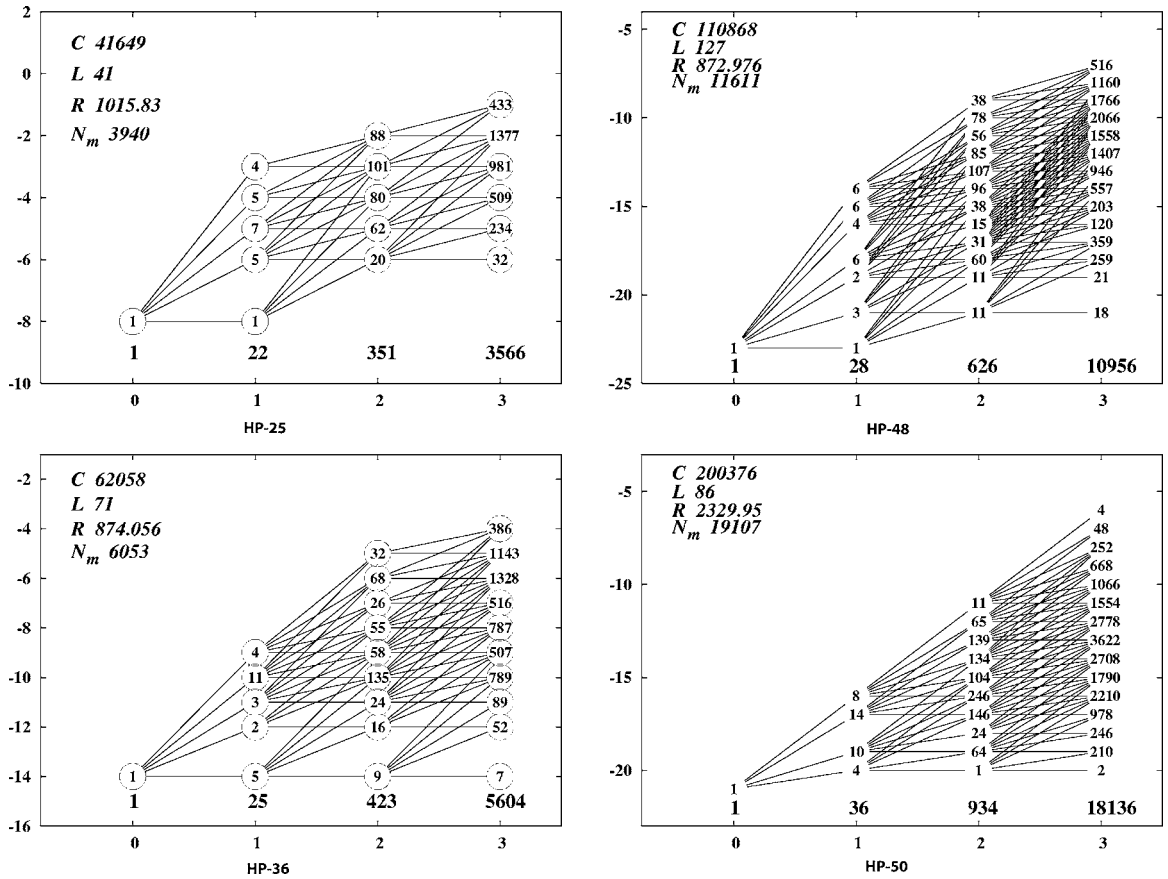

FIG. 13. Valid uphill pathways between a starting GM structure and structures three folds away, using the point mutation move class, for a randomly chosen GM of each benchmark sequence. *x* axis: number of point mutations; *y* axis: energy.

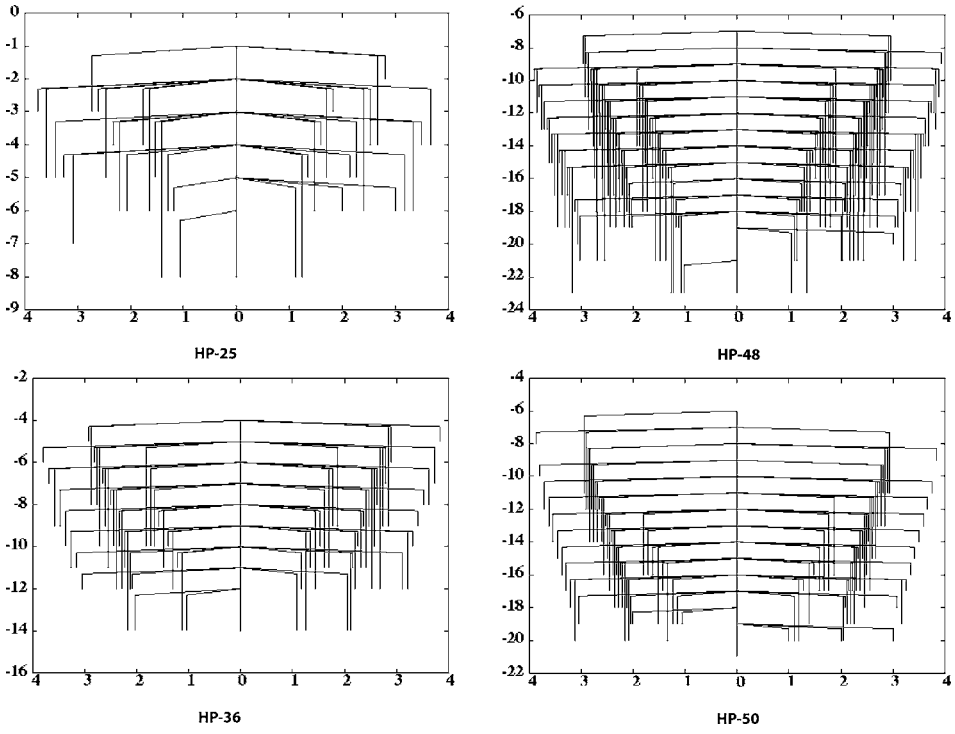

FIG. 14. Disconnectivity graphs for selected benchmark sequences. *x* axis: Hamming distance from starting GM structure; *y* axis: energy.

IV. CONCLUSIONS

A new benchmark 20-bead *HP* model protein sequence (on a square lattice), which has 17 distinct but degenerate global minimum (GM) energy structures, has been studied using a genetic algorithm (GA). The relative probabilities of finding particular GM conformations have been determined and related to the theoretical probability of generating these structures using a recoil growth constructor operator. For longer successful GA runs, the GM probability distribution is generally very different from the constructor probability, as other GA operators have had time to overcome any initial bias in the originally generated population of structures. Structural and metric relationships (e.g., Hamming distances) between the 17 distinct GM have been investigated and used, in conjunction with data on the connectivities of the GM and the pathways that link them, to explain the GM probability distributions obtained by the GA. A comparison has also been made of searches, where the sequence is defined in the normal (forward) and in the reverse direction. The ease of finding mirror image solutions has also been compared. Finally, this approach has been used to rationalize the ease or difficulty of finding the GM for a number of standard benchmark *HP* sequences on the square lattice.

In this study, we have shown that the relative probabilities of finding particular members of a set of degenerate global minima (for *HP* bead model proteins on a square lattice) depend critically on the topography of the energy landscape in the vicinity of the GM, the connections and distances between the GM, and the nature of the operators used in the search method—in this case, a genetic algorithm. While even for a 20-bead sequence the total number of valid conformations (4.19×10^7) is too large to enable the global potential energy surface to be plotted, we have shown that valuable information can be obtained by exploring limited regions around the various global minimum structures.

The work described here is currently being extended to include other lattice configurations, more sophisticated protein models, and alternative search algorithms. We believe that these studies may provide valuable insight into how the foldability of proteins (i.e., how reliably they can fold into a unique native state, which may be termed "natural searching") is related to the topography of the folding energy landscape and how this relates to the ease or difficulty of finding the native structure using "artificial search" algorithms.

ACKNOWLEDGMENTS

One of the authors (G.A.C.) is grateful to the EPSRC and the University of Birmingham for Ph.D. funding.

[1] G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure* (Springer-Verlag, Berlin, 1979).
[2] C. B. Anfinsen, Science **181**, 223 (1973).
[3] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).
[4] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
[5] A. R. Dinner, A. Sali, L. J. Smith, C. M. Dobson, and M. Karplus, Trends Biochem. Sci. **25**, 331 (2000).
[6] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[7] K. F. Lau and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 638 (1990).
[8] H. S. Chan and K. A. Dill, J. Phys. Chem. **95**, 3775 (1991).
[9] G. A. Cox, T. V. Mortimer-Jones, R. P. Taylor, and R. L. Johnston, Theor. Chem. Acc. **112**, 163 (2004).
[10] W. E. Hart and S. Istrail, http://www.cs.sandia.gov/tech_reports/compbio/tortilla-hp-benchmarks.html
[11] A. Shmygelska, R. Aguirre-Hernandez, and H. H. Hoos, Lect. Notes Comput. Sci. **2463**, 40 (2002).
[12] A. R. Leach, *Molecular Modelling: Principles and Applications* (Addison-Wesley-Longman, Harlow, 1996).
[13] R. Unger and J. Moult, J. Mol. Biol. **231**, 75 (1993).
[14] N. Krasnogor, D. A. Pelta, P. Martinez Lopez, P. Mocciola, and E. de la Canal, Proceedings of Engineering of Intelligent Systems 1998 (EIS'98), edited by E. Alpaydin and C. Fyfe (Academic Press, New York, 1998), p. 353.
[15] N. Krasnogor, W. E. Hart, J. Smith, and D. A. Pelta, Proceedings of the Genetic and Evolutionary Computation Conference 1999 (GECCO-99), edited by W. Banzhaf *et al.* (Morgan Kaufman, 1999), p. 1596.
[16] A. Nayak, A. Sinclair, and U. Zwick, J. Comput. Biol. **6**, 13 (1999).
[17] Z. Li and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **84**, 6611 (1987).
[18] E. M. O'Toole and A. Z. Panagiotopoulos, J. Chem. Phys. **97**, 8644 (1992).
[19] R. Ramakrishnan, B. Ramachandran, and J. F. Pekny, J. Chem. Phys. **106**, 2418 (1997).
[20] H. Frauenkron, U. Bastolla, E. Gerstner, P. Grassberger, and W. Nadler, Phys. Rev. Lett. **80**, 3149 (1998).
[21] F. Liang and W. H. Wong, J. Chem. Phys. **115**, 3374 (2001).
[22] T. Beutler and K. Dill, Protein Sci. **5**, 2037 (1996).
[23] H. H. Gan, A. Tropsha, and T. Schlick, J. Phys. Chem. **113**, 5511 (2000).
[24] S. Kirkpatrick, C. D. Gellatt, Jr., and M. P. Vecchi, Science **220**, 671 (1983).
[25] R. König and T. Dandekar, BioSystems **50**, 17 (1999).
[26] J. T. Pedersen, *Evolutionary Algorithms in Molecular Design*, edited by D. E. Clark (Wiley-VCH, Weinheim, 2000), p. 233.
[27] R. Unger, Struct. Bonding **110**, 61 (2004).
[28] A. Shmygelska and H. H. Hoos, Lect. Notes Comput. Sci. **2671**, 400 (2003).
[29] B. C. Curley, *Investigation and Application of Ant Colony Optimisation for Protein Folding*, Undergraduate project, University of Birmingham, 2003.
[30] A. Shmygelska and H. H. Hoos, BMC Bioinf. **6**, 30 (2005).
[31] S. Costa, N. B. Wilding, D. Frenkel, and Z. Alexandrowicz, J. Phys. Chem. **110**, 3220 (1999).
[32] F. R. Manby, R. L. Johnston, and C. Roberts, MATCH **38**, 111 (1998).
[33] R. W. Hamming, Bell Syst. Tech. J. **29**, 147 (1950).
[34] F. Despa, D. J. Wales, and R. S. Berry, J. Chem. Phys. **122**, 024103 (2005).