Assessment of protein folding potentials with an evolutionary method

David de Sancho and Antonio Rey[a)]
Departamento de Química Física I, Facultad de Ciencias Químicas, Universidad Complutense, E-28040 Madrid, Spain

(Received 29 March 2006; accepted 11 May 2006; published online 5 July 2006)

Many different protein folding potentials have been developed in the last decades, based upon knowledge of experimentally determined protein structures. Decoy-based techniques are frequently used to assess these force fields, but other methods can explore different features in the performance of the interaction schemes, thus helping in their evaluation. Here, we propose an evolutionary strategy to efficiently assess folding potentials. We apply it to three potentials with different characteristics, taken from the bibliography. A search for minimum energy protein topologies, treated as arrangements of rigid protein fragments, is performed. The method, applied to a set of helix bundle proteins, shows the different behavior of the studied potentials, providing a reasonably fast tool to evaluate their advantages and limitations. © *2006 American Institute of Physics*. [DOI: 10.1063/1.2210931]

I. INTRODUCTION

Certain amino acid sequences acquire their three dimensional native conformations in a process known as protein folding. Complex interactions, among the protein residues and with the solvent, take part in the events that guide a polypeptide chain towards the functional form of a protein. For most proteins, this state is reached reproducibly and in a short period of time.[1] The native conformation of a protein is believed to correspond to the global minimum in the free energy of the system, and it may be surrounded by several local minima.[2] The biopolymer is assumed to be funneled through a rugged energy surface and driven to its native conformation, avoiding to get trapped into the local minima.[3]

Considering this, a straightforward approach to predict the native conformation of a given amino acid sequence would be to locate the structure corresponding to its global energy minimum. In order to do so, both an accurate protein interaction potential and a reliable sampling algorithm are necessary. Protein folding potentials can be developed using different approaches, including statistical treatments.[4] These take advantage of the structural information deposited by experimentalists in the Protein Data Bank (PDB).[5] The so-called knowledge-based potentials are thus derived from a statistical analysis of the contacts between interaction centers of the protein, which is frequently represented in a reduced (coarse-grained) manner, to permit a wider sampling of the conformational space. The statistical treatment permits the derivation of contact matrices, which provide values for pairwise interaction energies. A wide range of this kind of potentials, with different levels of resolution and distance dependencies, has been developed in the last years.[6,7]

After being designed, any energy function needs some evaluation, in order to check whether an energy minimum is unmistakably defined and, moreover, whether it corresponds or not to the native structure. Generally, decoy-based techniques have been used for this assessment.[8] For a given sequence, these methods generate a large ensemble of plausible conformations (decoys) by means of gapless threading, molecular dynamics, or exhaustive enumeration. The energy of each resulting conformation is calculated according to the potential. The real native structure must have the lowest energy from the set, and thus *z-score* measurements are used to evaluate how the potential function distinguishes among the different structures. Decoy-based techniques are useful as they allow the comparison of the native conformation of a protein with many other folds.

In spite of their interest, they lack the possibility to test an interaction potential as it can be done in folding or energy minimization experiments, where a huge amount of decoys is automatically built by the searching technique itself. These methods, such as Monte Carlo or molecular dynamics annealing simulations and the more specific evolutionary algorithms, allow a broader search in the conformational space of the protein. Therefore, these minimization methods represent an important contribution to the evaluation of folding potentials.[9,10] A problem appears because, when analyzing the performance of the method, it is difficult to detach the effects of the searching strategy from the accuracy of the force field,[11] as it usually happens with protein folding simulation techniques. In addition, the large computational cost of the calculations involved in these annealing strategies makes it difficult to use them in detailed force field analyses.

Evolutionary algorithms are able to fast and efficiently locate optimal or near optimal solutions of a given function and, therefore, they have been also used in protein science.[11,12] These algorithms, based upon the mechanisms observed in natural evolution, handle a population of solutions for the system, which are encoded as chains of variables called *chromosomes*.[13] A series of genetic operators, mainly *reproduction*, *crossover*, and *mutation*, transforms the population of chromosomes along a number of generations. They combine and modify the solutions at one stage to gen-

[a)]Author to whom correspondence should be addressed. Electronic mail: jsbach@quim.ucm.es

erate new individuals, in such a way that new information, more adapted to the "environment," can be produced. The extent to which a given solution is adapted is given by a *fitness function*. The process continues until (hopefully) the optimal solution of the fitness function for the system is reached.

When dealing with a search for the minimum energy structure of a protein, a set of possible conformations is encoded in chromosomes according to the chosen representation, and the genetic operators successively build new structures. The value of the fitness function of a given solution will be related to the energy of its corresponding conformation, consistent with the interaction potential employed.

We have recently developed an evolutionary method for the evaluation of protein folding interactions with a reduced representation of the molecule.[14] The algorithm handles fragments of the protein, which are considered as rigid bodies, and tries to locate the arrangement that optimizes the energy of the collapse interaction. In that work, we used a structure based (i.e., Gō type) potential, which constitutes a good test for a minimization algorithm, since it presents a clearly defined minimum for every protein structure but yet involves a search in a complex energy landscape. The algorithm proved efficient in finding the best (native) packed structure for a series of fragments representing the protein, provided that the energy surface was sufficiently well defined.[14]

In this present work we use essentially the same evolutionary algorithm, with just a few variations, to check the performance of three representative interaction potentials: the Nanias potential posed by the Scheraga group,[15] the TE-13 potential by Tobi and Elber,[16] and the DFIRE-SCM potential developed by Zhang *et al.* in the Zhou laboratory.[17] It is important to stress that we are not presenting here a protein folding algorithm, much less a structure prediction strategy. The three force fields chosen are therefore considered just as examples of interaction potentials resulting from different derivations, to check the possibilities of our method to efficiently assess the capabilities of a given force field. Obviously, our results also represent a source of valuable information about the characteristics of the selected interaction definitions.

The Nanias potential has been chosen as an example of very simple potentials that use a relatively low number of parameters. In this case, the widespread Miyazawa-Jernigan (MJ) matrix of energy terms,[18] which is the result of a statistical analysis of contacts, is combined with a Lennard-Jones-type function. This function uses only the positions of the alpha carbons of the protein, and the Lennard-Jones function is for sure a very crude estimation of the potential dependency on distance, even more when taking into account that the original MJ matrix was based on contacts among side chains, not backbone atoms. We have chosen this potential, however, since it has apparently provided reasonable results for their authors in a coarse-grained model similar to ours.[15] Moreover, it provides a distance dependency to the Miyazawa-Jernigan contact matrix, yet nowadays a cornerstone in statistical protein contact potentials.[7] The crudeness of the mathematical representation may somehow preclude a good performance for the model (see below), but for our purposes of potential testing it is a good candidate to be checked against other interaction definitions.

On the contrary to the Miyazawa-Jernigan potential, the distance dependency of the contact frequencies was considered in detail in the original statistical treatment leading to the design of both the TE-13 and the DFIRE-SCM potentials. Like others before,[19,20] Tobi *et al.* derived their TE-13 potential using a linear optimization procedure with a huge number of constraints to overcome some of the problems they found in previous mean force potentials.[21] Zhang *et al.* also used a physics-based approach (as MJ did) for their potential design, but in their statistical protocol they established a new reference state that makes the force field to perform apparently better than other interaction models in decoy testing.[22]

All three potentials represent long range (along the sequence) interactions in the protein core. The level of coarse graining of our model[14] seems appropriate to calculate the energy of protein conformations as inter-residue interactions with these potentials and to evaluate how well they represent the interactions that actually stabilize the three dimensional native structure of a protein. In this work we use our evolutionary algorithm to analyze the behavior of these folding potentials and comment on what one can expect from them when used in folding simulations.

II. MODEL AND METHODS

A. Protein model and searching algorithm

The model is a minor extension of the one previously described.[14] Protein conformations are generated as arrangements of rigid fragments, whose geometry is kept frozen. This approach has been also used by other authors before.[10,15,23,24] See also the work in Ref. 25 and references therein. In the present work, we have considered helix bundle proteins, which we split at the turns that join well defined secondary structure regions, resulting in a series of rigid fragments. We have only considered all-alpha proteins since they are better suited for the three interaction potentials considered, which mainly represent the packing of hydrophobic residues in the protein core. In beta sheets, the hydrogen bonds among backbone atoms play a central role, but they are not explicitly considered in the potentials we are analyzing here. These bonds also participate in the stabilization of helices, a factor not explicitly included in the considered force fields either, which we avoid through the use of rigid fragments.

Thus, we take the PDB file and every single alpha helix, as defined in the file header, is included in its corresponding individual fragment of the model. Only one virtual bond between alpha carbons separates a protein fragment from the next one. For each protein, the sequence and coordinates of the interaction centers are taken from the PDB. For the calculations carried out with the Nanias potential, only alpha-carbon positions are required, since the authors use them as interaction centers. For the calculations using the TE-13 and DFIRE-SCM potentials, we need to know both the alpha

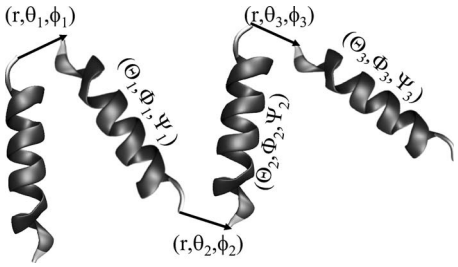


FIG. 1. Schematic representation of the variables used to completely describe a conformation for a protein divided in four mainly helical fragments. The value of r is set to 3.8 Å.

carbon and the side chain centroid positions. These are computed as the average over the coordinates of the side chain heavy atoms for every residue.

The evolutionary algorithm requires that each individual in the population (in this case, the configuration defined by the group of rigid fragments) has to be encoded as a chain of variables. A set of five variables is required to define the position and orientation of each fragment, except the first one that is kept fixed as a reference frame. Then, for a protein divided in n fragments, $5\times(n-1)$ variables are aligned as a chromosome and describe a possible topology of the protein model (see Fig. 1). The first two variables of set i encode the angles (θ_i,ϕ_i) of the spherical coordinates for the alpha carbon of the first residue in fragment $(i+1)$ from the alpha carbon of the last residue in fragment i (the separation r between both alpha carbons is fixed at 3.8 Å). The remaining three variables in the set encode the orientation of protein fragment $(i+1)$, through the Euler angles (Θ_i,Φ_i,Ψ_i). The reference frame for the orientation of a fragment can be external or be settled in the orientation of the previous fragment, a craft that improved the versatility of the method.[14] Thus, for each individual in the chromosome population, we decode two different topologies. The energy is computed for both, and we consider only the lowest energy from each pair as the fitness value for that individual.

We are using the evolutionary algorithm that we set up in previous work,[14] but here some details have been changed, as described below. The core of the method is a genetic algorithm with a population of n_c chains of variables (chromosomes). All variables are real and given an initial random value in the interval [0,1]. Afterwards, they are normalized to the usual intervals: $[0,\pi]$ and $[0,2\pi]$ for θ_i and ϕ_i of the spherical coordinates, respectively, and $[0,2\pi]$ for all the Euler angles. The fitness value is calculated for the n_c individuals, and the genetic operators are applied to produce the offspring.

(1) The reproduction operator copies all the chromosomes to the offspring population, in order to avoid any loss of valuable information in the initial set of chromosomes for every generation.

(2) Couples of individuals in the parent population are selected following an elitist procedure,[13] and the *crossover* operator is applied to combine the information of matched pairs. This way, n_c additional individuals are generated. A *roulette wheel* mechanism is used for the selection. The probability of an individual being selected is directly related to the absolute value of its fitness function, given that only individuals with a negative value of the energy can be selected. The exchange starts from a randomly selected variable of the chromosome. An input parameter allows us to control the kind of crossover to be performed: *single* or *double point crossover*, with p_{cross} and $(1-p_{\text{cross}})$ probabilities, respectively. With the former, big chromosome fragments may be exchanged that can be very useful at the first stages of the optimization. The double point crossover exchanges a single variable, and therefore it is more appropriate for a fine tuning of yet good solutions.

(3) Copies of individuals that come from steps 1 and 2 are mutated with a given probability p_{mut}. This mutation operator allows the creation of new information, hence increasing the diversity of the population, and may also improve good solutions further. $2n_c$ new individuals are generated with the mutation operator. This way, there is no information prone to be lost, independently of its origin. We have implemented the mutation operator as a random change in a single chosen variable in its corresponding range.

After the three operators have been applied, there are $4n_c$ individuals in the offspring. All of them are scored with the energy function and n_c individuals are selected to resize the population. For an energy minimization like the one we are carrying out, a decreasing order in the absolute value of the energy is mandatory as a selection criterion. However, when it is considered alone, it can lead to problems of premature convergence of the population to a local minimum, because many closely related or identical individuals may be selected. Instead of just imposing a threshold energy difference between individuals, we have found it better to match a structural criterion with the energy order, which is the major novelty of the algorithm in this work. We have chosen to use the *root mean square deviation* (RMSD), since it reflects differences between rigid bodies and it has been widely applied to the protein structure field.[26] Then, the selection is performed as follows: The best energy individual is selected first. From then on, increasing energy order is followed, but a threshold RMSD from all the previously selected conformations is required for every new individual of the offspring to be passed to the next generation.

The full procedure is repeated until a number of generations (n_{gen}) are completed.

We showed in our previous work that the genetic algorithm was a good optimization tool at the first stages but was likely to get stuck at local minima.[14] This problem cannot be easily solved by tuning the parameters that control the genetic operators. We found that a dramatic loss in the diversity of the population was the reason why the search tends to become growingly inefficient, as it usually happens with genetic approaches.[11] In order to overcome this serious drawback of the method, we devised a tool that resembles the island model, based on local independent searches that share part of the information from time to time.[27] Local best solutions from n_s program runs for a common set of parameters

(except seed numbers for random number generation) are brought to an intermediate population. Next, n_s new local searches are started from that intermediate population, which is completed at the beginning of each run with (n_c-n_s) random individuals. This *merging* procedure seems to refresh the optimization with new "genetic raw deal" that increases the diversity of the search and allows local best solutions to be mixed up and refined. After n_{gen} generations, a new merging process occurs, and the minimization continues until no noteworthy improvement is observed. With this refinement in the search algorithm, the performance gets significantly better, as we showed in our previous results.[14]

In this paper, we have implemented the method described in a series of programs with the different force fields mentioned in the Introduction. The programs have been used to minimize the global energy of a group of proteins, represented according to the coarse-grained model described above. The parameters that we have used for the algorithm runs are basically the same as those we used in our previous work, though certain changes have been introduced in view of meaningful improvements in performance. The size of the population n_c is fixed at 100 individuals. The values for the probabilities controlling the type of crossover and the mutation are $p_{\mathrm{cross}}=0.5$ and $p_{\mathrm{mut}}=0.1$. Ten different seed numbers are used for each run of the evolutionary strategy (n_s), and n_{gen} is fixed at 500 generations. We choose 1 Å as threshold RMSD.

Although we have checked the correct performance of the minimization procedure in our previous work,[14] the changes included in the protein model used here, mainly the reduction in the remaining degrees of freedom when we "delete" a single peptide bond among contiguous fragments in the model, deserve a preliminary control. Therefore we have also used, in a series of initial tests, a fitness function based on the radius of gyration of the sampled structures, plus a simple repulsive core among interaction centers, to see whether a nonspecific collapsing scheme could provide nativelike structures for our model. We have found that, with a model which only considers the alpha-carbon coordinates, as it happens with the one designed by their authors to be used with the Nanias potential, several topological arrangements of the helix bundles are possible with similar compact structures (schematic representations of some of those arrangements also appear in the calculations with the energy potentials being assessed and will be shown below). The inclusion in the model of the side chain centroids, as required by the TE-13 and DFIRE-SCM potentials, partially restricts these alternative topologies for proteins represented by a small number of fragments, surely due to the more stringent packing requirements, but the models with medium or large number of fragments still show degeneracy regarding the topological arrangements of the helices. Therefore, we conclude that the protein model used in this work is still flexible enough to allow a reasonably wide set of different compact structures (and a huge number of noncompact ones), which a properly defined interaction potential should be able to discriminate.

B. Energy calculations

The fitness function for every one of the interaction definitions we want to test is just equal to the value of the global energy of the conformation. This energy is calculated as the sum of all the pairwise terms between residues belonging to different protein fragments in the model. Thus, for a given conformation, distances between interaction centers that represent the residues are calculated. Intrafragment interactions are not computed because the protein fragments are kept frozen, and therefore their contribution to the energy is the same for any arrangement. In all the cases, interactions between bonded or neighboring amino acids at the linked ends of consecutive rigid fragments are not taken into account, since the three potentials considered in this work represent long range (along the sequence) interactions.

The calculations of pairwise energies have been performed in a particular way for each potential. Here, we describe these special features, which follow the definitions given by the authors.

(1) Nanias potential. The Nanias potential[15] uses the 20×20 matrix of contacts derived by Miyazawa and Jernigan.[18] As suggested by the authors, the values of e_{ij} used have been obtained by subtracting e_{rr} to the parameters listed in Table 3 of Ref. 18. These values are included in a Lennard-Jones-type equation for the distance dependency:[15]

$$U(r_{ij})=\frac{e_{ij}}{14\pm 15}\left[14\left(\frac{\sigma}{r_{ij}}\right)^{15}\pm 15\left(\frac{\sigma}{r_{ij}}\right)^{14}\right]. \tag{1}$$

The signs in Eq. (1) are adjusted according to the value of the contact term e_{ij}: positive if $e_{ij}>0$ and negative if $e_{ij}<0$. Therefore, for each amino acid pair ij, the distance r_{ij} between alpha carbons is calculated, the equation is applied, and a value of the energy $U(r_{ij})$ is obtained. The value of σ was set to 7.5 Å by the authors[15] for all the interaction pairs, and thus it has been kept here (although it does not seem to be too reasonable, as the results included in the next section show).

(2) TE-13 potential. For each pair of amino acids, the TE-13 potential[16] is defined in the distance interval of 2–9 Å between side chain centroids, and this range is divided in 13 bins. We have observed that, with the TE-13 potential alone, the minimization of the energy results in overlapping of the fragments. To solve this problem, two extra contributions are added to the potential: a soft sphere repulsive potential between alpha carbons for distances lower than 3.8 Å and another soft sphere contribution for distances between centroids lower than 2 Å. The aim of the latter is to compensate for the lack of statistics at very small distances in PDB structures. These distances do not appear in rigid decoys either, and therefore they are neither included nor missed in the original TE-13 potential derivation. They are, however, possible in a conformational sampling carried out in continuous space. The same happens with the repulsive term among alpha carbons, which just tries to avoid *phantom chains* which could yield unphysical conformations, therefore representing an unfair test of the interaction potentials.

(3) DFIRE-SCM potential. For each residue pair, the DFIRE-SCM potential of mean force[17] is defined for dis-

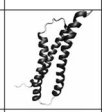
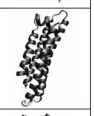
PDB code	# Protein fragments	# Residues in the Model	
1rpo	2	58	
1i6z	3	116	
2a3d	3	69	
1ktm	4	128	
1le4	5	138	
1ls4	5	155	

FIG. 2. Protein structures included in this study. The protein fragments have been defined using the guide of the secondary structure definition in the PDB files. Helices are shown in dark gray.

tances between centroids from 2 to 15 Å. This distance range is divided into 20 bins. The inter-residue pairwise energy is calculated as in the TE-13 potential. Again, soft sphere potentials between alpha carbons and side chain centroids are added in this case, for the same reasons previously commented on. As a matter of fact, a repulsive core has been also very recently included in one of the potentials of the DFIRE family (not the one considered here) by its authors.[10]

III. RESULTS AND DISCUSSION

In this work we have applied the method to helix bundle proteins that range from two to five fragments (see Fig. 2), with the three force fields chosen as tests for the evolutionary procedure. As previously explained, we have selected this kind of proteins because, with our rigid fragments model, they appear as an appropriate system to test potentials that mainly represent the interactions among non-neighboring residues in the protein core. In Table I we show the results for the lowest energy minimum encountered by the algorithm, with the RMSD of the corresponding fragment arrangement from the native conformation (computed, in all the cases, from the alpha-carbon coordinates).

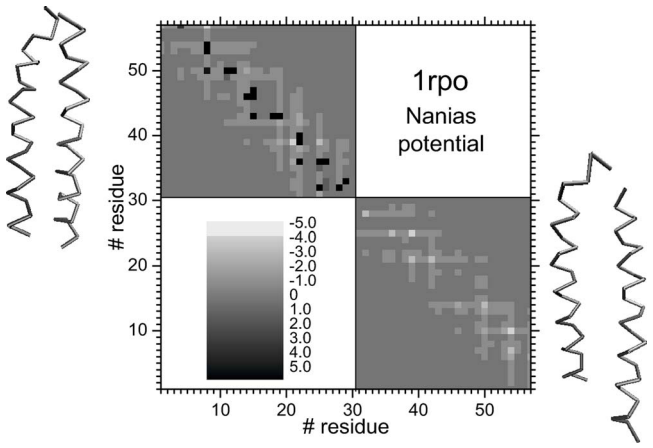


FIG. 3. Contact map with energy terms for the native conformation (top left) and the lowest energy conformation obtained with the evolutionary method and the Nanias potential (bottom right), together with representations of their respective topologies. Energies in the legend in *RT* units.

One of the facts that we want to point out in our results is the value of the energy for native structures that we also show in Table I. As the potentials have been derived from different approaches, they are expressed in different units, and therefore they are not directly comparable. The values of the energy for the native conformation obtained with the Nanias potential are highly positive. They are just the consequence of the high, unique value of the potential distance parameter σ chosen by the authors of the potential and the large value of the exponent for the repulsive part in Eq. (1). On the other hand, with the TE-13 and DFIRE-SCM potentials, we can see that the native energy is negative for all the proteins in Table I, although with different trends.

Taking as an example the simplest case, protein 1rpo, with only two fragments, there are several interactions that yield a positive energy value with the Nanias potential. In Fig. 3 we show the contact map that corresponds to the native conformation (top left panel), where we represent the values of the energy for each pair of residues, according to the Nanias potential. We can see several black squares that correspond to the repulsive interactions. Different amino acid pairs raise such large energies, although mainly

TABLE I. Energy of the native (E_N) and optimized (E_{min}) conformations of the six proteins for the three potentials, with the RMSD of the optimized structure from native in angstroms. Energies are expressed in *RT* units for the Nanias potential, arbitrary units for the TE-13 potential, and kcal/mol for the DFIRE-SCM potential.

	Nanias			TE-13			DFIRE-SCM		
PDB ID	E_N	E_{min}	RMSD	E_N	E_{min}	RMSD	E_N	E_{min}	RMSD
1rpo	6239	−39.9	3.7	−102.3	−155.2	1.0	−23.0	−24.2	0.4
1i6z	50141	−106.8	2.0	−95.2	−296.2	1.2	−52.6	−65.4	1.0
2a3d	1655	−76.4	3.0	−163.5	−260.5	0.9	−33.4	−42.5	0.7
1ktm	16463	−108.5	11.0	−235.7	−439.1	2.8	−89.1	−104.9	0.5
1le4	20411	−158.3	4.2	−362.8	−566.4	1.5	−101.8	−112.0	1.0
1ls4	41889	−127.9	24.0	−163.6	−452.5	18.5	−12.7	−128.2	1.2

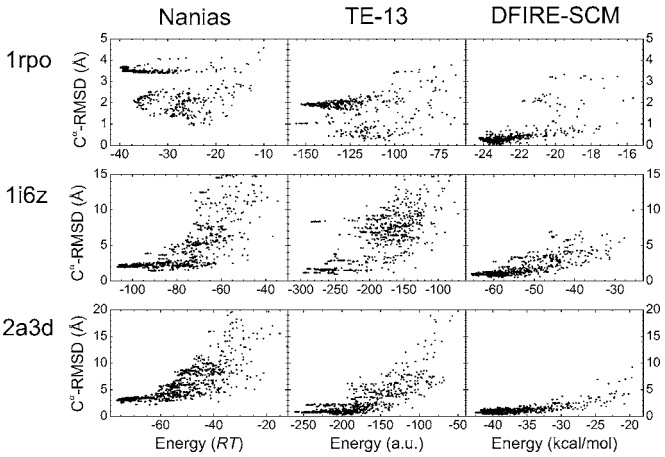


FIG. 4. Scatter plots of the energy vs RMSD (in Å) from native of the best conformations explored along the optimization generations for proteins 1rpo, 1i6z, and 2a3d with the three potentials. Results of five different runs of the algorithm are shown in each graph. The three plots in the same column correspond to a given force field, as indicated at the top; the three plots in the same row correspond to the same protein, whose PDB code is shown at the left.

hydrophobic-hydrophobic and hydrophobic-charged pairs are involved in these interactions. These residues are closely packed in the protein, but due to their short inter-residue distance, their interaction is considered repulsive by the potential. This occurs as a consequence of the functional form of the potential, which has its minimum at 7.5 Å, regardless of the type and size of the side chains involved, and reaches very high values at shorter distances due to the large exponents considered [see Eq. (1)]. The same situation is found for the rest of the proteins, with a different number of repulsive interactions, depending on the native topology of the protein.

Despite this evident drawback of the Nanias potential, the results of the minimization are fair, though the TE-13 and DFIRE-SCM potentials behave clearly better. In Table I we show the best value of the energy for each potential, with the RMSD of its corresponding structure from the native conformation. From the values of RMSD one can infer that, in all the studied cases, the minimum energy conformation is defined closer to the native structure for the DFIRE and TE-13 potentials than for the Nanias potential. The differences between E_N and $E_{\min}$ values for the TE-13 potential are often large, even though the RMSD between native and optimized structures is marginal in most of the proteins considered. This is a direct consequence of the very rough behavior of some of the contributions defined in this potential, with individual pairwise energies between two residues frequently varying $\sim$10 energy units in contiguous distance bins, especially in the short distance region.[16] The differences between E_N and $E_{\min}$ values are much smaller for the DFIRE-SCM potential in most of the cases, with minimal RMS deviations between both structures, reflecting at least its smoother character and probably also its better global behavior (see below).

In the first row of Fig. 4, we show the comparison of the results of the minimum energy values explored along our evolutionary search for protein 1rpo in five independent runs of the algorithm with every potential. In the graphics contained in this figure, each dot represents the fittest individual in one of the generations. Thus, the ensemble of dots represents the evolution of the searching procedure. Although related to the energy landscape, these plots cannot be considered as a proper representation of it, since the minimization technique employed lacks the thermodynamic characteristics of a molecular dynamics or Metropolis Monte Carlo sampling.

In the case of the Nanias potential, Fig. 4 shows for 1rpo two groups of compact structures with low energy values: one that is closely related to the native conformation (RMSD$\simeq$2 Å) and another one that is less similar (RMSD $\simeq$3.5 Å). The lowest energy conformation belongs to this second group. In the contact map representing the minimized conformation in Fig. 3 (bottom right panel), we see that the repulsive interactions defined by the Nanias potential in the native conformation do not appear anymore. Moreover, we can see in the backbone representation of the lowest energy conformation that the two helices are out of register. To decrease the value of the energy, the optimization algorithm has increased the distance between residues with respect to their separation in the native structure.

With the TE-13 potential, the algorithm is able to locate for 1rpo a lowest energy protein conformation of RMSD =1 Å. In addition, another structure is located at RMSD =2 Å that seems to be better defined and apparently more accessible to our searching algorithm than the best energy conformation. Both the 1 Å and the 2 Å solutions are closely related to the native conformation of the protein. With the DFIRE-SCM potential, a low energy structure with RMSD =0.4 Å from native is reproducibly found for this protein. This means that, for this force field, the minimum is unmistakably defined in the native conformation. Thus, even for this simple protein, where the restrictions of the coarse-grained model do not permit a lot of freedom, the different potentials begin to show distinct characteristics.

Two additional examples of the behavior of the Nanias potential are the results that we obtain for the proteins fragmented in three bodies: 1i6z and 2a3d (see Fig. 4). In both cases, despite the high energy value assigned to the native conformation (see Table I), we see that the algorithm reproducibly reaches minima located at only 2–3 Å from the native structure and correspond to nativelike arrangements of the fragments. The main differences are that some inter-residue distances need to be shifted again to avoid repulsions. With the TE-13 potential, the search reaches the native topology for these three-fragment proteins. But in the case of 1i6z, in addition to the nativelike arrangement, another minimum with approximately the same value of the energy but RMSD from the native structure of more than 8 Å is found. In Fig. 5 we show the backbone representations and topological diagrams of both the native conformation and the alternative minima from the algorithm (these "topological mirrors" are the type of structures also resulting from the unspecific compacting test explained at the end of the section describing our model). The minimum with RMSD=8 Å corresponds to an alternative arrangement of the helices, counterclockwise around an imaginary bundle axis instead of

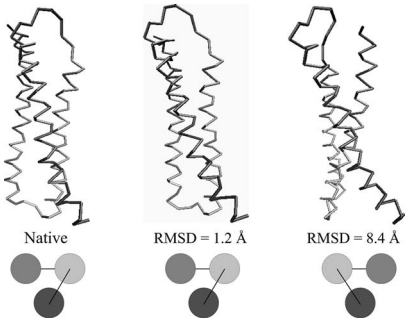


FIG. 5. Backbone representations and topology diagrams of the native conformation of 1i6z and two alternative minima found with the TE-13 potential. Different shades of gray correspond to different protein fragments in the model.

clockwise. Besides, the nativelike minimum corresponds to a rather scattered ensemble of packed fragments, similar to the double minima that we observe for 1rpo (see Fig. 4). The fuzzy definition of the minimum appears again in the case of 2a3d, for which several optimizations get stuck, with this potential, in a topology that has a value of RMSD of 2 Å from native. On the contrary, the DFIRE-SCM potential has a very well defined minimum that is reproducibly located for both three-fragment proteins.

In Fig. 6 we show plots of the energy values for the lowest energy conformations found along the minimizations for proteins with four and five fragments. The Nanias potential offers worse results for more difficult search problems such as 1ktm or 1ls4, although it happens to behave a little better for 1le4. In the four fragment case, 1ktm, the algorithm finds a few low energy structures far from the native one. They correspond to assemblies in which only three of the four helices are packed together, while the remaining one only partly interacts with the external side of the three helix bundle. This may occur due to the difficulties in packing the four helices together, given the poorly defined repulsive part of the potential we have already discussed. For 1le4, the best

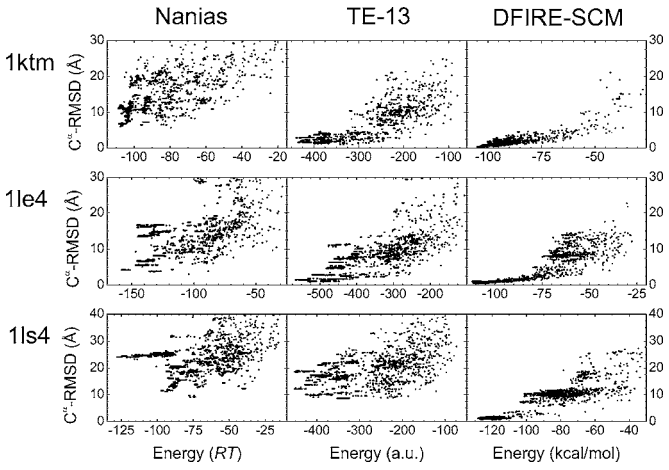


FIG. 6. Scatter plots of the energy vs RMSD (in Å) from native of the best conformations explored along the optimization generations for proteins 1ktm, 1le4, and 1ls4 with the three potentials. Results of five different runs of the algorithm are shown in each graph. The three plots in the same column correspond to a given force field, as indicated at the top; the three plots in the same row correspond to the same protein, whose PDB code is shown at the left.

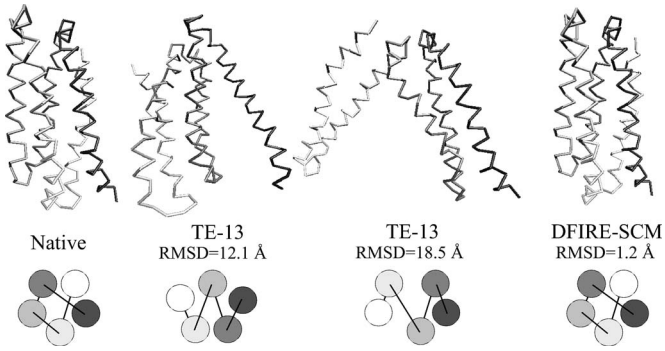


FIG. 7. Backbone representations and topology diagrams of the native conformation of 1ls4, two alternative minima found with the TE-13 potential, and the minimum for the DFIRE-SCM potential. Different shades of gray correspond to different protein fragments in the model.

energy conformation has a value of RMSD of 4.2 Å from native, which is significantly better, though far from optimal. In addition, the search gets easily trapped in local minima as a consequence of the apparently rough definition of the energy surface. The local minima in which the search sometimes gets stuck correspond to four helix compact arrangements in which the remaining fragment is barely interacting with the bundle. Finally, the minimum energy conformation located for 1ls4 is formed by two domains, one with two and one with three fragments. Due to the repulsive part of the potential, the five fragments simply cannot be packed together forming the five helix bundle of the native structure.

Both the TE-13 and DFIRE-SCM potentials succeed in locating the energy minimum of 1ktm and 1le4 in the native topology (see Fig. 6 and Table I). In both cases, the RMSD from native of the best conformation is lower for the DFIRE-SCM potential than for the TE-13 potential, and the minimum is better defined in the former than in the latter. The case in which the differences between both potentials become more remarkable is 1ls4. Again, the DFIRE-SCM potential succeeds in its definition of the energy minimum at the native conformation. On the other hand, two different topologies are ranked best with the TE-13 potential. We show the backbone representations for these structures in Fig. 7. One of these minima obtained for the TE-13 potential is a five helix bundle with an alternative disposition of the helices (again, one of the topological alternatives also found by a generic collapse), and the other one corresponds to a conformation in which the fragments are arranged in groups of independent domains.

IV. SUMMARY AND CONCLUSIONS

In a previous work,[14] we designed an evolutionary method that efficiently searches for minimum energy conformations for a coarse-grained model of a protein. Given the problems that algorithms for conformational sampling usually face, due to the link between the searching method and the fitness function that is being optimized, we used a Gō-type potential to test the performance of the algorithm. Here, it has been applied to different fitness functions, in order to check the performance of our strategy in real cases and therefore to assess, beyond standard decoy testing, three

knowledge-based force fields with different features: the Nanias potential from the Scheraga group[15] that uses the Miyazawa-Jernigan contact terms,[18] the TE-13 potential developed by Tobi and Elber using linear optimization,[16] and the DFIRE-SCM potential of mean force of Zhang *et al.*[17]

The results provided by our method show that the Nanias potential works worse than the other two force fields. We have checked that this poor performance is a direct consequence of the repulsive part posed by the authors in the very simple distance dependent energy function that assigns high values of the energy for contacts at short but feasible distances. Also, the value of σ in the function is rather problematic, as no considerations about size or residue type have been adopted. Still, the potential keeps some good features of the Miyazawa-Jernigan contact matrix, and reasonably good optimized structures are obtained when a small number of fragments take part in the formation of the protein core.

The TE-13 potential works noticeably better than the Nanias potential, as a more detailed parametrization in the distance dependency definition has been assumed. To reach this good behavior, our methodology indicates that some corrections must be introduced in the potential to make it compatible with a folding simulation procedure. A repulsive potential is needed at short distances between alpha carbons and between side chain centroids to avoid overlapping of the protein fragments. This repulsive contribution was not needed in previous decoy-based approaches to the assessment of potentials, as a set of static conformations was used, but we have found here that it is required in folding or energy minimization numerical experiments. For most of the proteins we have considered here, the structures with the lowest values of energy for the TE-13 potential correspond to nativelike arrangements of the protein fragments. But we also observe for this interaction scheme a tendency to find scattered minima, maybe as a consequence of the free functional form of the potential. This can be understood if we take a look at plots of the pairwise contributions as those in Fig. 1 of Ref. 16. Large energy jumps appear between one distance bin and the next that can have dramatic effects in minimization experiments. When dealing with protein-size problems, many residue pairs contribute to the global energy, and with this erratic variation in the energy, it can be sometimes a matter of pure chance that once a local minimum has been found, the appropriate change of variables towards the global minimum happens. Besides, in some cases such as 1i6z or 1ls4, problems of degeneracy are found. The potential shows to be effective in building a well packed hydrophobic core, but in some occasions it can hardly discriminate among the different accessible topologies, something that we have also found when using a generic optimization algorithm which maximizes packing without considering the sequence.

The best performing potential in our test has been the DFIRE-SCM potential, after adding the soft sphere contributions between alpha carbons and side chain centroids, as with the TE-13 potential. For all the proteins considered, well defined minima very close to the native conformation are found with our method and the DFIRE-SCM force field, regardless of the complexity (number of fragments) in the protein model. The statistical thermodynamics based approach used to derive this potential, with a noninteracting reference state, appears to be the most accurate in capturing the interaction propensities between residues. In this potential, jumps in the values of the energy from one distance bin to another are less steep than with the TE-13 potential (see, for example, Fig. 3 of Ref. 17) that makes the solution to the minimization problem more feasible. This way, we find that this potential is the most appropriate of the three considered here for simulation experiments with coarse-grained representations of the protein, since it has been able, in all our tests, to properly discriminate the native structure from other compact conformations available to the model.

With this work we contribute to the need of studies in which force fields developed by different approaches are validated. We have shown that our evolutionary method provides a very efficient tool to properly analyze the behavior and capabilities of interaction potentials as those considered in this paper. Still, much is yet to be done, and other techniques must contribute in this evaluation task. In this case, potentials that mainly represent tertiary interactions in proteins have been tested, but models for other contributions to protein stability need to be covered as well.

ACKNOWLEDGMENTS

This work was supported in part by Grant No. BQU2002-04626-C02-01 from Ministerio de Ciencia y Tecnología, Spain and by Grant No. PR1/06-14424-A from Universidad Complutense de Madrid. One of the authors (D.D.S.) acknowledges a grant from Spanish Ministerio de Educación y Ciencia. The authors would like to thank Harold A. Scheraga, Maurizio Chinchio, Dror Tobi, Ron Elber, Yaoqi Zhou, and Chi Zhang for providing their potentials and kind instructions for energy calculations. Some images were made with VMD support.[28]

[1] C. M. Dobson, A. Šali, and M. Karplus, Angew. Chem., Int. Ed. **37**, 868 (1998).
[2] C. B. Anfinsen, in *Nobel Lectures, Chemistry 1971–1980*, edited by T. Frängsmyr and S. Forsén (World Scientific, Singapore, 1993).
[3] C. L. Brooks III, M. Gruebele, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **95**, 11037 (1998).
[4] M. J. Sippl, J. Comput.-Aided Mol. Des. **7**, 473 (1993).
[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[6] N. Buchete, J. Straub, and D. Thirumalai, Curr. Opin. Struct. Biol. **14**, 225 (2004).
[7] P. Pokarowski, A. Kloczkowski, R. L. Jernigan, N. S. Kothari, M. Pokarowska, and A. Kolinski, Proteins **59**, 49 (2005).
[8] B. H. Park, E. S. Huang, and M. Levitt, J. Mol. Biol. **266**, 831 (1997).
[9] V. Tozzini, Curr. Opin. Struct. Biol. **15**, 144 (2005).
[10] H. Li and Y. Zhou, J. Bioinf. Comput. Biol. **3**, 1151 (2005).
[11] R. Unger, Struct. Bond. **110**, 153 (2004).
[12] D. E. Clark and D. R. Westhead, J. Comput.-Aided Mol. Des. **10**, 337 (1996).
[13] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, Reading, MA, 1989).
[14] D. de Sancho, L. Prieto, A. M. Rubio, and A. Rey, J. Comput. Chem. **26**, 131 (2005).
[15] M. Nanias, M. Chinchio, J. Pillardy, D. R. Ripoll, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **100**, 1706 (2003).
[16] D. Tobi and R. Elber, Proteins **41**, 40 (2000).
[17] C. Zhang, S. Liu, H. Zhou, and Y. Zhou, Protein Sci. **13**, 400 (2004).
[18] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[19] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **227**, 876 (1992).

[20] M. Vendruscolo, R. Najmanovich, and E. Domany, Proteins **38**, 134 (2000).
[21] D. Tobi, G. Shafran, N. Linial, and R. Elber, Proteins **40**, 71 (2000).
[22] H. Zhou and Y. Zhou, Protein Sci. **11**, 2714 (2002).
[23] B. Fain and M. Levitt, J. Mol. Biol. **305**, 191 (2001).
[24] T. X. Hoang, F. Seno, J. R. Banavar, M. Cieplak, and A. Maritan, Proteins **52**, 155 (2003).
[25] X. Li, M. P. Jacobson, and R. A. Friesner, Proteins **55**, 368 (2004).
[26] R. Brüschweiler, Proteins **50**, 26 (2003).
[27] D. Whitley, Inform. Software Tech. **43**, 817 (2001).
[28] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).