Assembly of protein tertiary structures from secondary structures using optimized potentials

Trinh Xuan Hoang,[1] Flavio Seno,[2] Jayanth R. Banavar,[3] Marek Cieplak[4] and Amos Maritan[5]

[1]*The Abdus Salam International Center for Theoretical Physics (ICTP), Strada Costiera 11, 34100 Trieste, Italy*

[2]*INFM-Dipartimento di Fisica "G. Galilei," Università di Padova, Padova, Italy*

[3]*Department of Physics, 104 Davey Laboratory, Pennsylvania State University, University Park, Pennsylvania*

[4]*Institute of Physics, Polish Academy of Sciences, Aleja Lotnikow 32/46, 02-668 Warsaw, Poland*

[5]*International School for Advanced Studies and INFM, Via Beirut 2-4, 34014 Trieste, Italy*

Correspondence to:

Trinh Xuan Hoang,

Condensed Matter Section,

The Abdus Salam International Center for Theoretical Physics (ICTP),

Strada Costiera 11, 34100 Trieste, ITALY.

E-mail: hoang@sissa.it

Tel. +39-040-2240460

Fax. +39-040-224163

Key words: LINUS; protein structure prediction; perceptron learning; ab initio prediction; protein folding; simulated annealing

Abstract

We present a simulated annealing based method for the prediction of the tertiary structures of proteins given knowledge of the secondary structure associated with each amino acid in the sequence. The backbone is represented in a detailed fashion whereas the sidechains and pair-wise interactions are modeled in a simplified way, following the LINUS model of Srinivasan and Rose. A perceptron based technique is used to optimize the interaction potentials for a training set of three proteins. For these proteins, the procedure is able to reproduce the tertiary structures to below 3$\mathring{A}$ in root mean square deviation (rmsd) from the PDB targets. We present the results of tests on twelve other proteins. For half of these, the lowest energy decoy has a rmsd from the native state below 6$\mathring{A}$ and in 9 out of 12 cases, we obtain decoys whose rmsd from the native states are also well below 5$\mathring{A}$.

I. INTRODUCTION

The protein folding problem entails the determination of the secondary and tertiary structure of a protein based on the knowledge of its sequence.[1,2] The tertiary structure determines the functionality of the protein and thus the prediction of the folded structure is a central challenge. The protein folding problem can be successfully tackled by using an "engineering approach" which may involve comparative modeling, investigation of statistical correlations in a protein data base, and a tool box of other techniques which work with varying degrees of success. One successful example of this approach is based on building native structure-like conformations from protein-like fragments which are database derived.[3,4] This allows for a sensible narrowing of the conformational search because the number of preferred conformations for short peptide fragments is limited.[5]

From a fundamental point of view, however, it is desirable to develop scientific methods that are purely ab initio, and that are easy to understand.[6] One the most interesting, simple, and truly ab initio microscopic models is LINUS that was introduced by Srinivasan and Rose.[7–9] This model

incorporates the most important structural features of proteins including the backbone representation, simple descriptors of the side-chain geometry, excluded volume interactions, interactions between side-chains, and hydrogen bonding. LINUS has been demonstrated to be valuable[7–9] for protein structure prediction, particularly for determining the propensities of various segments to form particular secondary structures. This determination is done at a fixed moderate temperature that can be chosen optimally.[10] This is because LINUS embodies an interplay between energy and entropy: helices form as a result of minimizing the energy whereas strands are favored by the entropy.

In this paper, we build on LINUS in the context of ab initio tertiary structure prediction given a sequence of amino acids and their secondary structure. We demonstrate that LINUS has many features that are appealing for this task. First, based on an appropriate Monte Carlo dynamics, it is able to fold an extended chain to a globular form without violating any steric constraints. Second, the dynamics allow for an efficient exploration of the low energy states in a way that singles out the native state provided one parametrizes the energy functions in a proper manner. We determine the parameters through a learning procedure.

As noted earlier, we simplify the task of tertiary structure prediction by assuming an a priori knowledge of the location of the secondary structures of the proteins under study. It should be noted that many of the current methods of proteins structure prediction rely, to various extents, on being preceded by prediction of the secondary structures. Indeed, by using a combination of techniques such as multiple sequence alignment and neural networks, one can predict the basic secondary structural features, e. g. α-helices and β-strands, with a confidence that exceeds 70%.[11–14] Our simplified task could be thought of as a step in the ultimate goal of ab initio tertiary structure prediction. As we shall demonstrate, an important attendant advantage of such a simplified challenge is that it allows one to glean fundamental insights into the protein folding problem.

In this paper, we adopt a simulated annealing procedure[16] to study a LINUS-based model at low temperatures. In order to accommodate this shift in focus from intermediate to low temper-

atures, we work with a modified version of the LINUS model in which the original simplified interaction potentials are endowed with more amino acidic specificity. The basic idea of our learning procedure is to generate low energy conformations, which preserve the pre-assigned secondary structure of proteins with known tertiary structure with a set of carefully chosen potential energy parameters. These low energy conformations are then used as decoy conformations for carrying out a refinement of the potential energy parameters in order to ensure that the known native state structure is indeed lower in energy than the decoy conformations. This procedure, when iterated leads to an estimation of the optimal potential energy parameters. Strikingly, our procedure can be used to assess whether the form chosen for the potential energy is adequate to the task of successfully discriminating between the true native state structure and the realistic competing decoy conformations.

Our studies lead to the conclusion that the potential energy parameters for a commonly used pairwise interaction can be optimized to simultaneously ensure that the native state structures of three distinct proteins can be successfully discriminated against decoy conformations. The parameters obtained by learning the three proteins provide an adequate description of these proteins: the LINUS native states are very close to the experimentally determined structures, the energy landscape is funnel shaped, steric clashes are avoided and the native state is kinetically accessible. Furthermore, the parameters are also found to be adequate for several other proteins.

II. METHODS

Models

A comprehensive description of LINUS can be found in the original papers of Srinivasan and Rose[7,8] and in our independent assessment of LINUS[10] which follows the particular details of reference.[8] Briefly, in LINUS, the atoms are modeled as hard spheres with predefined radii and they are not allowed to overlap. The coordinates of all backbone heavy atoms are represented exactly whereas those of the sidechains are represented in a simplified manner. Specifically, glycine

is effectively considered to have no sidechain, the side chains of alanine, valine, serine, threonine and cysteine are considered explicitly by their heavy atoms (no more than three in each case), and the remaining amino acids are represented by C_β and one or two pseudo C_γ atoms, depending on whether the sidechain is branched out or not.

The Hamiltonian contains local and nonlocal terms. The former correspond to the fixed-distance tethering by the peptide bonds and the torsional energies. The latter correspond to the hydrogen bonds, and to the pairwise contact interactions. A torsional energy is included to prevent formation of conformations with positive Ramachandran ϕ angles by associating a positive cost, ε_{tor}, with them. The exception is glycine for which positive values of ϕ are rewarded with a negative energy of $-\varepsilon_{tor}$.

Hydrogen bonds can be formed between the backbone N atoms and either the backbone O atoms (backbone-to-backbone H-bonds) or the sidechains of the amino acids from the set [Ser, Thr, Asn, Asp, Gln, Glu] (sidechain-to-backbone H-bonds). The distance between the donor and acceptor must be smaller than $5\mathring{A}$ and $4\mathring{A}$ for these two cases respectively. It is also required that the out-of-plane dihedral angle $O(k)$–$N(l)$–$C_\alpha(l)$–$C(l-1)$ should be larger than 140^o, where k and l are the indices of the amino acids involved along the sequence. The contact interaction arises between the sidechain atoms of the amino acids. A contact is declared to be formed if the distance, r_{ij}, between the two atoms, i and j, is smaller than $R_c \equiv R_i + R_j + 1.4\mathring{A}$, where R_i and R_j are the contact radii[8] of the atoms. The energy of a contact decreases linearly from zero to its minimal value as the distance between the atoms decreases from R_c to $R_i + R_j$ and it remains constant at smaller distances. This constant depends on the specificity of the sidechains. It should be noted that the contact radii of the atoms are larger than their hard sphere radii so that atoms can be fully in contact without steric clashes.

Our implementation of LINUS incorporates two crucial changes compared to the original formulation.[8] The first is that we distinguish between the short range backbone-to-backbone H-bonds that correspond to the H-bonds within helices and turns and the long range backbone-to-backbone H-bonds (the ones that are further than 4 residues apart along the sequence). We have

found that using a lower energy (more favorable) score for the long range H-bonds is essential for the assembly of the β-strands into a sheet.

The second change is an increase in the variety of the contact interactions. In the original LINUS, the amino acids were divided into three categories: hydrophobic, amphipathic and polar. The contact interactions were allowed only between the hydrophobic and the amphipathic amino acids.[8] We have verified that the interactions provided by considering just three kinds of amino acids are not enough to make the native state the global energy minimum. Thus we consider 20 distinct types of amino acids. Because glycine, as defined above, does not participate in contact interactions, there are 190 different energy parameters corresponding to different contacts between the remaining nineteen amino acids. A special case corresponds to the contact between the S atoms of two cysteines. A disulfide bond is said to be formed if two such S atoms are closer than $3.6\mathring{A}$ and the energy of the bond decreases linearly to its minimal value as the distance decreases to $2.65\mathring{A}$. This energy remains constant when the distance decreases to $2\mathring{A}$, below which value there is a steric clash. Since the disulfide bond has a different nature than other kinds of interactions, we consider its minimal energy to be 5 times lower (more favorable) than the energy of a short range backbone-to-backbone H-bond. Each cysteine is allowed to have no more than one disulfide bridge so when it comes to a situation in which several S-S contacts are formed with a given cysteine then only the one with the lowest energy is selected as a disulfide bridge, while the others are regarded as usual sidechain-sidechain contacts. Overall we have 194 adjustable energy parameters: three for hydrogen bonds, 190 for contact interactions, and one for the torsional energy. We attempt to determine these 194 parameters through learning.

The move set

Since our goal is to predict the tertiary structure based on the knowledge of the secondary structure and not the prediction of the latter, we need to modify the original set of Monte Carlo moves so the secondary structures are maintained. We define the local moves in the Ramachandran space of the torsional angles. Specifically, three consecutive residues are moved at a time and the changes in the ϕ and ψ angles are drawn from a Gaussian distribution with a zero mean and a

dispersion that ranges from 1^o to 5^o (the value of the dispersion is chosen in proportion to the acceptance rate). The residues assigned to α-helices or β-strands are allowed to have adjustable φ and ψ but the adjustments are constrained to lie within appropriate regions in the Ramachandran plot as shown in Figure 1. Thus for α-helices $\phi = -64 \pm 7^o$ and $\psi = -43 \pm 7^o$. For β-strands $\phi = -130 \pm 15^o$ and $\psi = 135 \pm 15^o$. On the other hand, the residues in the loops can have arbitrary φ and ψ values.

The sidechains can be rotated around the C_α–C_β bond and the corresponding torsion angle χ may take any value. However, in order to enhance the acceptance rate the values related to the rotamer positions are chosen with a higher probability. Thus χ is chosen randomly in one of the three windows: -60 ± 20^o, 60 ± 20^o, and 180 ± 20^o with a probability of 0.9 and the values outside the windows are picked with a probability of 0.1. All other angles and bond lengths are kept fixed during the simulation except that the torsion angle ω about the peptide bond can vary within 180 ± 10^o, and the N–C_α–C bond angle can vary within 110 ± 5^o.

The simulation starts from an open conformation with the α-helices and β-strands built in, based on the experimental structure file from the Protein Data Bank (PDB).[15] The moves proceed from the N-terminus and a complete progression to the C-terminus is called a cycle. A new conformation selected in this way is rejected if it leads to steric clashes (up to 50 attempts are made to decide on a move that does not generate steric clashes), otherwise it is accepted with a probability $\mathcal{P} = min\left\{1, e^{-\Delta E/T}\right\}$, where ΔE is the energy difference compared to the previous conformation and T is a fictitious temperature.

The low energy conformations are obtained through an annealing scheme[16] in which the temperature is decreased in steps according to the formula $T_{k+1} = 0.97\ T_k$, where k is a step index. The starting temperature is chosen so that the acceptance rate, r, is larger than 0.1, and once r falls below 0.001 we carry out a zero temperature quench. The typical length of a simulation is between 50 000 and 100 000 cycles.

Perceptron method for learning the interaction potentials

The principal requirement for the stability of the native state is that the native state has a lower energy than any other viable conformation.[17–25] This requirement is also a necessary condition for the simulated annealing to lead to the native state from an open conformation. Thus

$$E(\Gamma_0) < E(\Gamma_D) \quad \forall\, D\,, \tag{1}$$

where Γ_0 denotes the native conformation and Γ_D denotes the conformation of a decoy. We have found that, even with the secondary structures incorporated in the decoys, the conditions given in Eq. (1) cannot be satisfied by using unrefined interactions such as those used in the original version of LINUS.

The optimal stability perceptron algorithm[26] is an iterative procedure which allows one to find a solution that satisfies a given set of linear inequalities optimally. We follow one of the several learning schemes discussed in Ref.[22] (see also Ref.[23] and, in the context of determining environmental scores for the amino acids, Ref.[24]). For a given set of decoys and a known native conformation, our learning procedure deals with a set of inequalities of the type

$$\frac{E(\Gamma_D) - E(\Gamma_0)}{d(\Gamma_D, \Gamma_0)} \;>\; 0\,, \tag{2}$$

where $d(\Gamma_D, \Gamma_0)$ denotes the root mean square deviation (rmsd) of the decoy structure from the native state. The use of the rmsd in the denominator helps to shape the energy landscape into a funnel and ensures that a conformation with a large rmsd is less preferred than conformations which are more native like. Figure 2 illustrates that our procedure indeed leads to the required features.

A brief description of the perceptron method is as follows. The inequalities (2) can be rewritten in the form

$$\vec{\mathbf{A}}_k \cdot \vec{\varepsilon} > 0, \quad k = 1, 2, \ldots, M\,, \tag{3}$$

where $\vec{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{194}\}$ is a 194-component vector defined by the energy parameters to be optimized, $\vec{\mathbf{A}}_k$ is a vector of the corresponding coefficients of the k-th inequality, M is the total

number of inequalities (the combined number of decoys for all proteins in the training set), and the dot denotes a scalar product. For a given training set of parameters $\vec{\varepsilon}$, the stability of the k-th inequality is defined by $Q_k = \vec{\mathbf{A}}_k \cdot \vec{\varepsilon}$, while the perceptron stability is defined as $Q \equiv Q_l$ with l such that $Q_l < Q_k$, $\forall k \neq l$. The stability Q can be maximized by iteratively replacing $\vec{\varepsilon}$ by a new vector given by $\vec{\varepsilon}' = \vec{\varepsilon} + \delta \cdot \vec{\mathbf{A}}_l$, where δ is a small number. Each iteration step is followed by a normalization of $\vec{\varepsilon}$ so that $\Delta = \sum_i \varepsilon_i^2$ remains constant while Q is recomputed. A convergence is guaranteed to be reached in a finite number of steps. In our case this typically takes tens of thousands of iterations and this number depends on how small the value of δ is. The starting values for $\vec{\varepsilon}$ can be arbitrary. If $Q > 0$ the problem is learnable, which means that there exists a set of energy parameters such that all of the inequalities are satisfied. Indeed, if many solutions exist, the perceptron method selects the best among them.

Generation of LINUS native conformation

Since LINUS is an approximate model – it uses idealized bond angles and bond lengths for the backbone and a simplified representation of the sidechains – it can not be expected to provide a perfect fit to the PDB structures. However, in order to make use of inequalities (2) in the learning process we need to find a LINUS-based conformation that plays the role of the native state. We determine it by arranging the backbone to have the PDB-derived values of ϕ and ψ and then perform moves in order to reduce the rmsd to the PDB backbone as much as possible. In practice, we easily arrive at conformations with the backbone rmsd as small as $0.1\mathring{A}$ for the small size proteins considered here. Subsequently, the LINUS sidechains are built in and the sidechains are rotated in order to eliminate steric clashes. At this stage, the energy is minimized by the Monte Carlo quenching process without imposing any constraints on the segments that correspond to the secondary structures. The energies here are assigned treating the proteins as homopolymers, i.e. all of the energy parameters, but the torsional energy, are assumed to be uniform and negative. The final best native-like conformation usually has a rmsd of about 0.5–1.0$\mathring{A}$ from the PDB structure.

III. RESULTS AND DISCUSSION

Learning the interaction potentials from three training proteins

Three proteins have been chosen for learning the energy parameters: the B1 domain of protein G (the PDB code is 1GB1), ribosomal protein L7 (1CTF) and crambin (1CRN). They are all α/β proteins and their lengths do not exceed 70 residues. Crambin contains 3 disulfide bonds while the other two do not contain any cysteines. Their native conformations are shown in the left-hand side of Figure 3. For each protein, we have generated the LINUS-based native-like conformations using the fine tuning procedure described in the previous section. Their rmsd's relative to the corresponding PDB structures are $0.56\mathring{A}$, $0.68\mathring{A}$ and $0.72\mathring{A}$ for 1GB1, 1CTF and 1CRN respectively.

The decoys for those proteins are generated with built in constraints on the secondary structures. The identification of the secondary structures is obtained from the PDB files whenever it is available. For protein G, this kind of information is not present in the PDB file for 1GB1. In this case we take the assignment of the secondary structure from another file, corresponding to 1PGA, which contains the X-ray determined coordinates for the same protein. The residues at the end of the α-helices or β-strands can sometimes be confused with turns or coils. Because of this, we have effectively shortened some secondary structure fragments by one or two residues from their ends. An example of this situation is the second β-strand of 1CTF (residues 92-98 according to the PDB assignment), which is significantly distorted at positions 94, 97 and 98. For this case, we impose the strand geometry, in the Monte Carlo process, only on the residues from site 92 to 96. The α-helices usually show much smaller irregularities than the β-strands and the adjustments are much less severe. In general, adjustments are made in order to accommodate residues that lie too far from the Ramachandran regions associated with a given type of secondary structure. Such assignment changes have been made for 3 proteins in the learning set and 1 protein in the test set. Table 1 indicates all the modifications that we have made in the secondary structure assignment given in the PDB file. Note that the PDB information is needed only for α-helices and β-sheets

and the records on turns and other type of secondary motifs such as the 3_{10}-helix are irrelevant for our calculations. It should be noted that, even with these irregularities removed, the native values of the ϕ and ψ angles for some residues in the secondary structures may still lie outside the expected regions in the Ramachandran plot within which they are dynamically constrained. This situation is illustrated in Figure 1 for the 1CTF protein. However, we have found that, for most of the proteins considered here, we can generate conformations within $3\mathring{A}$ in rmsd from the PDB structures with the constraints imposed on the secondary structures.

The learning process is implemented in an iterative manner by starting from some generic potentials (as derived from references[7] and[8]) and generating the corresponding decoys. We identify decoys that compete with the native state in that they have energies comparable to or better than the native state energy. We adjust the energy parameters to destabilize the decoys. At the next stage, we use the just learned potentials and generate new corresponding decoys, and so on. The set of decoys keeps expanding through accumulation and all of them are used to generate the inequalities. Furthermore, the inequalities derived for different proteins are combined together. Decoys, which are found to be within $2\mathring{A}$ from the PDB structure, are discarded because these conformations are too close to the native state to provide unbiased inequalities that could properly shape the folding funnel (the number of such discarded decoys is typically one or two for a given protein). In total, we have generated approximately 1200, 1200, and 500 decoys for the 1GB1, 1CTF, and crambin proteins respectively. In the final rounds of learning, approximately 50 new decoys are generated for each protein. We stop when the predicted structures in the training set have a rmsd which is smaller than $3\mathring{A}$.

Figure 4 shows the plot of energy vs. rmsd for all the decoys for 1GB1 using the energy parameters that have been learned at the final round of learning. It can be seen that the decoys cluster in a way that suggests an emergence of a funnel-like energy landscape. The energy decreases with the rmsd at the lower boundary of the cluster, as indicated by the dashed line in the figure, and the native state is the global energy minimum. The perceptron stability Q is the slope of the dashed line. Note that the bigger the Q, the more funnel-like the energy landscape. On the other

hand, the learning process ought to result in a systematic decrease of Q because more and more inequalities are taken into account unless convergence is achieved. This can be seen in Figure 5, which shows Q/Δ as a function of the number of decoys where $\Delta = \sqrt{\sum_i \varepsilon_i^2}$, the rms value of the energy parameters, provides a normalization. Q/Δ is seen to decrease with the number of decoys used for learning. A sharp decrease at around 2400 decoys corresponds to arriving at a situation at which the set of decoys is sufficiently large to provide a significant reduction in the size of the relevant parameter space. After this decrease there is a plateau in Q which suggests a convergence of the energy parameters.

On optimizing the energy parameters, as described above, our annealing procedure is able to predict the tertiary structures for the three proteins that were used in learning with a very good accuracy. Among about 50 decoys generated for each of the three proteins after the learning was completed, the lowest energy state is found to be below $3\mathring{A}$ in rmsd from the PDB structure (see Table I and Figure 3). Furthermore, as shown in Figure 4 for 1GB1 and 1CTF, the decoys appear to be in a well developed funnel-shaped energy landscape. Figure 3 compares the PDB native conformations to the predicted conformations. Although the rmsds seem to be still somewhat large, e.g. $2.94\mathring{A}$ in the case of 1CTF, the predicted conformations exhibit striking similarities to the PDB structures. The best success in prediction is for 1GB1, for which the rmsd of the lowest energy decoy is $1.90\mathring{A}$ away from the PDB structure.

After optimizing the potentials, we rescaled our parameters to set the energy of a short range backbone-backbone H-bond equal to -1.0. In these units, the energy of a long range backbone-backbone H-bond is found to be -1.4952 and the sidechain-backbone H-bond has a value of -0.7779. In contrast, the original LINUS parameters for these two situations were equal to and twice the short range backbone-backbone H-bond respectively. The torsion energy remains positive and the hydrophobic interactions are better differentiated. However, we find that, on average, the contact energies between groups of sidechains of different hydrophobicity are in rough accord with the scale introduced by Srinivasan and Rose.[7,8]

Prediction of tertiary structures using the optimized potentials

Once the optimized potentials from three training proteins were determined, we carried out prediction of the tertiary structures for a dozen additional proteins assuming knowledge of the secondary structure. These proteins are listed in Table I. Seven of them are of the α/β type and five of the α type. Their lengths vary between 28 and 72. 50 decoys are generated for each protein and the one which was lowest in energy was taken as a representation of the native state. The results of our simulations are compiled in Table I, in which we present the rmsd of the lowest energy decoy with respect to the PDB target and also the lowest rmsd among all the decoys obtained through the simulated annealing. It can be seen that we are able to predict the tertiary structure with an accuracy of below 6$\mathring{A}$ for half of the new proteins, and in the four best cases the rmsd of the predicted conformation is below 5$\mathring{A}$. We have checked that if we use the standard LINUS energy parameters without the adjustments obtained through learning then the predicted structures are further than 8$\mathring{A}$ away from the targets.

Figure 6 compares our predictions to the PDB targets for three α/β-proteins: the zinc-finger (1ZAA), protein L (1HZ6) and cro repressor (1ORC). Although the rmsds obtained are not as good as for the training proteins, the close similarities between the predicted and targets structures can be easily recognized. The overall topologies are correct in all cases. The best prediction in terms of the rmsd belongs to the zinc-finger which is also the smallest protein in the set. However, there are some noticeable flaws in the predicted structure. The first one is due to the irregularity of the single helix in the PDB structure. The N-terminal part of this helix is twisted in a tighter way than the regular helix. Our procedure fails to reproduce the correct twist because the Ramachandran regions used for the helices correspond only to perfect helices. The second flaw is that the β-hairpin is not as close to the helix as it should be. This can be due to the presence of a zinc atom which is covalently bonded to the sidechains of two histidines from the helix and two cysteines from the β-hairpin in the real structure. These bonds stabilize the structure significantly but they are not taken into account in our model.

The second best predicted protein in the set is protein L. Its native topology is close to that of protein G which has been used for learning. However, protein L is 6 residues longer and

shares little sequence similarity with protein G and yet our procedure is able to predict the tertiary structure with a correct topology and a rmsd of nearly 4$\mathring{A}$. The cro repressor is also an interesting case. The formation of the β-sheet and the relative positions of the helices are very well predicted. The relatively large rmsd of 4.79$\mathring{A}$ is primarily due to poor prediction of the loops at the two terminals.

The prediction is bad for four α/β-proteins in the set: the mercuric transport protein (1AFI), the C-terminal domain of the E-coli reca (1AA3), hyperthermophile protein (1SAP), and the chey-binding domain of chea (1FWP). For these proteins the lowest rmsd and the rmsd of the lowest energy decoy are all larger than 8$\mathring{A}$ from the PDB targets. This indicates that the learned potentials are not adequate for those proteins. Another possibility is that their native conformations depend on the binding cofactors, which are not taken into account in our simulations. Note that both 1AA3 and 1SAP bind to DNA, whereas 1FWP binds to another substrate.

The predictions for the α-proteins are also reasonably good even though the potentials were learned using α/β-proteins. For the ROP protein (1RPO), a two-helix bundle with a simple topology, we obtained the lowest energy decoy with a rmsd of 4.63$\mathring{A}$. For the next three-helix bundles: human P8MTCP1 (1HP8), staphylococcus protein A (1BDD) and pheromone ER-1 (1ERC), the predicted rmsds are 5.74$\mathring{A}$, 5.85$\mathring{A}$ and 7.89$\mathring{A}$ respectively. For the five-helix bundle 434 repressor (1R69), the prediction is a failure since the lowest energy decoy is 11.31 $\mathring{A}$ away from the native state. It should be noted that for all helical proteins the lowest rmsd obtained is quite low (in the range from 3.3$\mathring{A}$ to 4.4$\mathring{A}$) in comparison to some α/β proteins. This is due to the fact that the α-helix bundles are easily accessible dynamically.

It should be noted that the overall rmsd is not the only measure for the assessment of the predictions. As we have seen in Figure 6 for Cro repressor, the large rmsd, 4.79$\mathring{A}$, of the predicted structure primarily arises from the misfolding of the two tails of this protein, while the rest is correctly formed. Another example is the case of the hyperthermophile protein (1SAP) shown in Figure 7. For this protein, in the predicted structure, the helix is misfolded and it causes a very large overall rmsd of 11.9$\mathring{A}$. If one considers only a fragment of this protein which excludes the

helix, the rmsd of the predicted structure with respect to the native structure is reduced to 4.89Å. For this fragment, the N-terminal β-hairpin and the central β-sheet with three strands are correctly formed.

In order to assess the quality of our predictions, in Figure 8 we plot the rmsd vs. length of the best predicted fragment for all the proteins considered. This kind of plot is usually seen in CASP assessments. For a given number of consecutive residues, L, we find a fragment in the predicted conformation that has the smallest rmsd from the PDB structure. Typically, the rmsd monotonically increases with the fragment's length, L. For the case of 1SAP the increase of rmsd vs. L is relatively small when L is less than 50. The rapid increase of rmsd after $L = 50$ corresponds to the situation when the misfolded helix is included in the fragment. Figure 8 shows that, for fragments of length 50 or shorter, our procedure can predict the structure with an accuracy below 4Å for seven proteins out of 15, and below 5Å for 10 proteins out of 15. These results are encouraging and indicate that our potential parameters may be useful for proteins other than in the training set.

We turn now to an examination of the kinetic accessibility of the low rmsd structures through our annealing procedure. Figure 9 shows the distributions of the rmsd of the generated decoys relative to the PDB targets for all of the proteins studied here. Except for 1CRN, the low rmsd structures do not correspond to peaks in the histograms. This suggests that the annealing used here may be too fast to avoid large rmsd conformations. On comparing the histograms of different proteins, one finds that for some proteins, such as 1GB1, 1CRN, 1ZAA and 1RPO, the native state or more precisely the low rmsd conformations are more easily kinetically accessible unlike others such as 1SAP and 1FWP. Poor accessibility to the native state could arise for different reasons including a poor annealing scheme, inefficiency of the dynamics, bad energy potential parametrization or from factors intrinsic to the geometry of the native conformation itself. Interestingly, we are able to obtain low rmsd conformations consistently for α-proteins.

It should be noted that, for each of the three proteins in the learning set and a large number of the proteins in the test set, the lowest rsmd we obtain is of the same order as that found using

Baker's ROSETTA method.[4] Note that ROSETTA uses predicted secondary structures whereas we use experimentally determined ones. Even though the total number of decoys generated for each protein in our study is much smaller than in ROSETTA, we are able to obtain lower rmsd decoys in several cases. It is interesting to note that for the two proteins, 1SAP and 1FWP, for which our scheme performs most poorly, ROSETTA was also unable to generate decoys with rmsds lower than 6Å.[4]

In none of the tested proteins, the native-like LINUS conformation has the lowest energy compared to the energies of the decoys. Thus any poor predictions are not due to a lack of dynamical accessibility of the native state but due to its poor thermodynamic stability. One might expect that one ought be able to improve the potentials and, thence, the thermodynamic stability by enlarging the training set. However we have found that special care should be taken when choosing proteins for learning and also a good preparation of the LINUS-based native conformation is needed. This is due to the nature of the perceptron learning for which one bad inequality may lead to a destruction of the entire set of potential energy parameters.

It has been recently pointed out that the protein backbone in the native state behaves like a tube with a strikingly uniform radius.[27,28] This reflects special properties of the structural and chemical components of the protein sequence and the nature of its compact state. Thus it is interesting to check whether the decoys generated by our methods behave like a tube and whether one can use this criterion to select the structure that is most protein-like in this respect. Given the C_α backbone of a decoy, the local radius of the tube at position i is defined by

$$\mathcal{R}_i = \min_{j,k}\{\mathcal{R}_{ijk}\}, \tag{4}$$

where $\mathcal{R}_{ijk}$ is the radius of the circle drawn through 3 points given by the positions of the C_α atoms of indices i, j and k in the sequence. It has been found that in real protein structures $\mathcal{R}_i$ has a very small fluctuation around $2.7\mathring{A}$. In order to take this into account we define a tube parameter which is given by

$$R_0 = \frac{1}{L}\sum_{i=1}^{L}(\mathcal{R}_i - 2.7)^2\,, \tag{5}$$

where L is the number of residues in the protein. The R_0 values for protein structures are typically smaller than 0.02. It is found that many of our decoys have R_0 significantly larger than this PDB value, but the best (lowest energy) decoy is found to have a R_0 value which is very reasonable. Figure 10 shows R_0 versus rmsd for the decoys obtained for 1CTF with the optimized potentials. Interestingly, the lowest rmsd decoy (which is also the lowest energy decoy for this protein) is also the decoy with the lowest R_0. This indicates that R_0 can be used to weed out a lot of decoys that are not protein-like. In Table I we show the rmsd of the decoy selected with the lowest R_0 obtained for the proteins studied. Though they are not as good as those obtained by using the lowest energy criterion, the tube-based criterion appears to be quite efficient. One can see that, in addition to 1CTF, for a number of other proteins such as 1GB1, 1CRN, 1ZAA and 1ROP, R_0 can be used to select a structure with rmsd lower than 5$\mathring{A}$.

A recent study[10] analyzed the ability of LINUS to predict different kinds of secondary structures based on conformational biases,[8] i.e. the propensities of the chain to form a given kind of secondary motif during a fixed temperature run. The potentials used in this study were the unrefined ones given by Srinivasan and Rose. We have tried to use our optimized potentials obtained from the present study to predict the secondary structures of several proteins studied in Ref.[10]. We found that the optimal temperature for the prediction is different from that found in Ref.[10] due to changes in the energy parameters but the rate of successful prediction at this temperature remains roughly unaffected. This indicates that the procedure used to predict the secondary structures is quite robust against fine-tuning of the interactions. It also confirms that the sterics play an important role in the formation of the secondary structures.[8]

IV. CONCLUSIONS

The approach presented in this study involves both potential design and folding, and it relies on one of the most important aspects of protein folding: the belief that proteins fold to their native conformations by searching for the global minimum in the energy landscape. With *a priori* knowledge of the secondary structures, proper folding is obtained for all three proteins used

in the training set for which the potentials have been optimized. Despite some limitations, the transferability of the potentials to other proteins is shown to work reasonably well.

The energy function, that we employed, is rather simple with, basically, only two kinds of non-local interactions: hydrogen bonding and side-chain contact interactions. The success of the methods tested here indicates that these interactions are of primary importance in folding. Our work demonstrates that it is possible to fold a protein with pre-formed secondary structures without violating steric constraints in a well-defined microscopic model. This supports a scenario for the folding mechanism in which the secondary structures are formed early during the folding process.

ACKNOWLEDGMENTS

We are indebted to Raj Srinivasan and George Rose for their very helpful advice on LINUS. We also thank Giuseppe Zanotti for helpful discussions. This work was supported by NASA, INFM and COFIN-2001.

REFERENCES

[1] Anfinsen C. Principles that govern the folding of protein chains. Science 1973;181:223-230.

[2] Creighton T. Proteins, structure and molecular properties. W.H. Freeman and Company publishers; 1993.

[3] Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209-225.

[4] Simons KT, Strauss C, Baker D. Prospects for *ab initio* protein structural genomics. J Mol Biol 2001;1191-1199.

[5] Micheletti C, Seno F, Maritan A. Recurrent oligomers in proteins: An optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. Proteins 2000;40:662-674.

[6] Osguthorpe DJ. Ab initio protein folding. Curr Opi Struct Biol 2000;10:146-152.

[7] Srinivasan R, Rose GD. LINUS: a hierarchic procedure to predict the fold of a protein. Proteins 1995;22:81-99.

[8] Srinivasan R, Rose GD. A physical basis for protein secondary structure. Proc Nat Acad Sci USA 1999;96:14258-14263.

[9] Srinivasan R, Rose GD. Ab initio prediction of protein structure using LINUS. Proteins 2002;47:489-495.

[10] Hoang TX, Cieplak M, Banavar JR, Maritan M. Prediction of protein secondary structures from conformational biases. Proteins 2002;48:558-565.

[11] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584-599.

[12] Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure prediction. Proteins 1997;27:329-335.

[13] Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34:508-519.

[14] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195-202.

[15] Bernstein FC, Koetzle TF, Williams GJB, E. F. Meyer Jr. EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535-542.

[16] Kirpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science 1983;220:671-680.

[17] Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876-888.

[18] Crippen GM. Easily searched protein folding potentials J Mol Biol 1996;260:467-475.

[19] Seno F, Maritan A, Banavar JR. Interaction potentials for protein folding. Proteins 1998;30:224-248.

[20] Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;24:11101-11108.

[21] Van Mourik J, Clementi C, Maritan A, Seno F and Banavar JR. Determination of interaction potentials of amino acids from native protein structures: test on simple lattice models. J Chem Phys 1999;110:10123-10133.

[22] Dima RI, Settanni G, Micheletti C, Banavar JR, Maritan A. Extraction of interaction potentials between amino acids from native protein structures. J Chem Phys 2000;112:9151-9165.

[23] Micheletti C, Seno F, Banavar JR, Maritan A. Learning effective amino acid interactions through

iterative stochastic techniques. Proteins 2001;42:422-431.

[24] Chang I, Cieplak M, Dima RI, Maritan A, Banavar JR. Protein threading by learning. Proc Nat Acad Sci USA 2001;98:14350-14355.

[25] Dobbs HT, Orlandini E, Bonaccini R and Seno F Optimal potentials for predicting inter-helical packing in transmembrane proteins. Proteins 2002;49:342-349.

[26] Diederich S, Opper M. Learning of correlated patterns in spin-glass networks by local learning rules. Phys Rev Lett 1987;58:949-952.

[27] Maritan A, Micheletti C, Trovato A, Banavar JR. Optimal shapes of compact strings. Nature 2000;406:287-290.

[28] Banavar JR, Maritan A, Micheletti C, Trovato A. Geometry and physics of proteins. Proteins 2002;47:315-322.

TABLES

Name	PDB code	length	class	lowest rmsd	lowest energy decoy	most tube-like decoy
[a]Protein G	1GB1	56	α/β	1.90Å	1.90Å	4.99Å
[b]Ribosomal protein L7	1CTF	68	α/β	2.94Å	2.94Å	2.94Å
[c]Crambin	1CRN	46	α/β	2.04Å	2.15Å	3.68Å
[d]Zinc-finger	1ZAA	28	α/β	2.11Å	2.98Å	3.28Å
[e]Protein L	1HZ6	62	α/β	4.03Å	4.03Å	16.58Å
Cro repressor	1ORC	64	α/β	4.79Å	4.79Å	6.00Å
Merp	1AFI	72	α/β	4.96Å	6.83Å	10.37Å
E-coli Reca	1AA3	63	α/β	8.17Å	8.17Å	10.71Å
Hyperthermophile Sac7D	1SAP	66	α/β	8.39Å	11.91Å	11.58Å
Chea	1FWP	69	α/β	9.29Å	12.09Å	12.29Å
Cole1 ROP protein	1RPO	61	α	3.37Å	4.63Å	4.43Å
Human P8-MTCP1	1HP8	68	α	4.14Å	5.74Å	9.87Å
[f]Staphylococcal protein A	1BDD	52	α	3.67Å	5.85Å	8.66Å
Pheromone ER-1	1ERC	40	α	4.38Å	7.89Å	8.29Å
434 repressor	1R69	63	α	4.21Å	11.31Å	6.30Å

[a] Secondary structure assignment: helix (23-35), strands (1-8, 13-18, 42-46, 51-56)

[b] Secondary structure assignment: helices (65-75, 81-87, 101-112), strands (54-59, 92-96, 116-120)

[c] Secondary structure assignment: helices (7-17, 23-30), strands (1-4, 32-35)

[d] Fragment studied: 33-60

[e] Secondary structure assignment: helices (26-39, 41-44), strands (4-11, 17-24, 47-52, 57-62)

[f] Fragment studied: 6-57

TABLE I. The name, PDB code, length and class of the proteins studied and the rmsds of the decoys obtained by our prediction using the optimized potentials. The 4-th, 5-th and 6-th columns show the lowest rmsd, the rmsd of the lowest energy decoy and the rmsd of the decoy selected by using a criterion based on the tube picture of protein structures (see text). The first three proteins have been used for learning, and the prediction is made for the other 12 proteins in the list. The secondary structure assignment is taken as that given in the PDB file for all proteins, except the ones marked with superscripts *a*, *b*, *c* and *e*. The assignment used for these proteins are indicated in the Table.

FIGURE CAPTIONS

Fig. 1. The ϕ and ψ angles of the residues in the α-helices (solid squares) and β-sheets (star points) of the ribosomal protein L7 (PDB code 1CTF). The secondary structure assignment for this protein is shown in Table I. The two rectangles indicate the constraint regions for α and β secondary structure.

Fig. 2. The energy vs. rmsd of all the decoys that have been generated for protein G plotted with the potentials obtained in the final round of learning. The open circle at a small rmsd denotes the LINUS conformation (obtained by relaxing the PDB structure) that plays the role of the native conformation in the learning process. The slope of the dashed line indicates the perceptron stability, Q.

Fig. 3. The PDB native conformation and the lowest energy conformation obtained with optimized potentials for the three proteins that have been used in the learning procedure. The numbers in Angstroms denote the rmsd from the native state structure.

Fig. 4. Energy vs. rmsd for decoys generated using the optimized potentials for 1GB1. The open circle indicates the LINUS conformation used as the native state in the learning. A similar plot for 1CTF is shown in the inset.

Fig. 5. The perceptron stability, Q, normalized by the square root of energy parameters, $\Delta = \sqrt{\sum_i \varepsilon_i^2}$, plotted as a function of the number of decoys, M, used in the learning procedure.

Fig. 6. The structure prediction for three proteins in the test set using the potentials that were optimized by learning.

Fig. 7. The structure prediction for 1SAP. The darker regions show the first 50 residues for which the rmsd is 4.89 $\AA$. The overall rmsd is 11.91 $\AA$ because the helix is misfolded.

Fig. 8. The rmsd vs. the length of the fragment that was predicted the best when compared to the PDB target. The solid lines are for the three proteins that were used in the learning

procedure, the dashed lines are for α/β proteins from the testing set and the dotted lines are for the α proteins. The thicker line corresponds to 1SAP.

Fig. 9. Histograms of decoys generated using the optimized potentials as functions of the rmsd to the native state. Approximately 50 decoys have been generated for each protein.

Fig. 10. The tube parameter, R_0, versus rmsd of the decoys obtained for the ribosomal protein L7 (1CTF) with optimized potentials. The continuous horizontal line indicates the value of R_0 computed from the PDB structure of this protein.

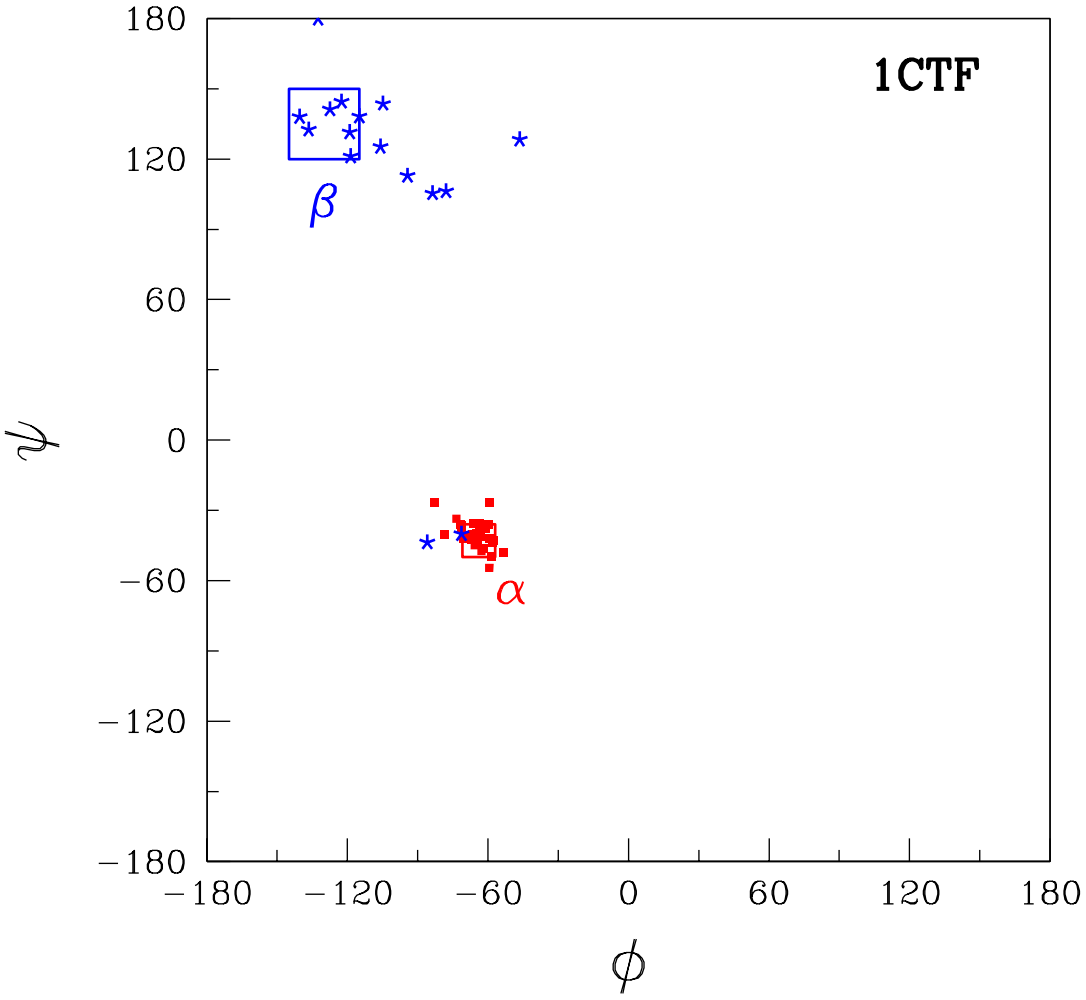

FIG. 1.

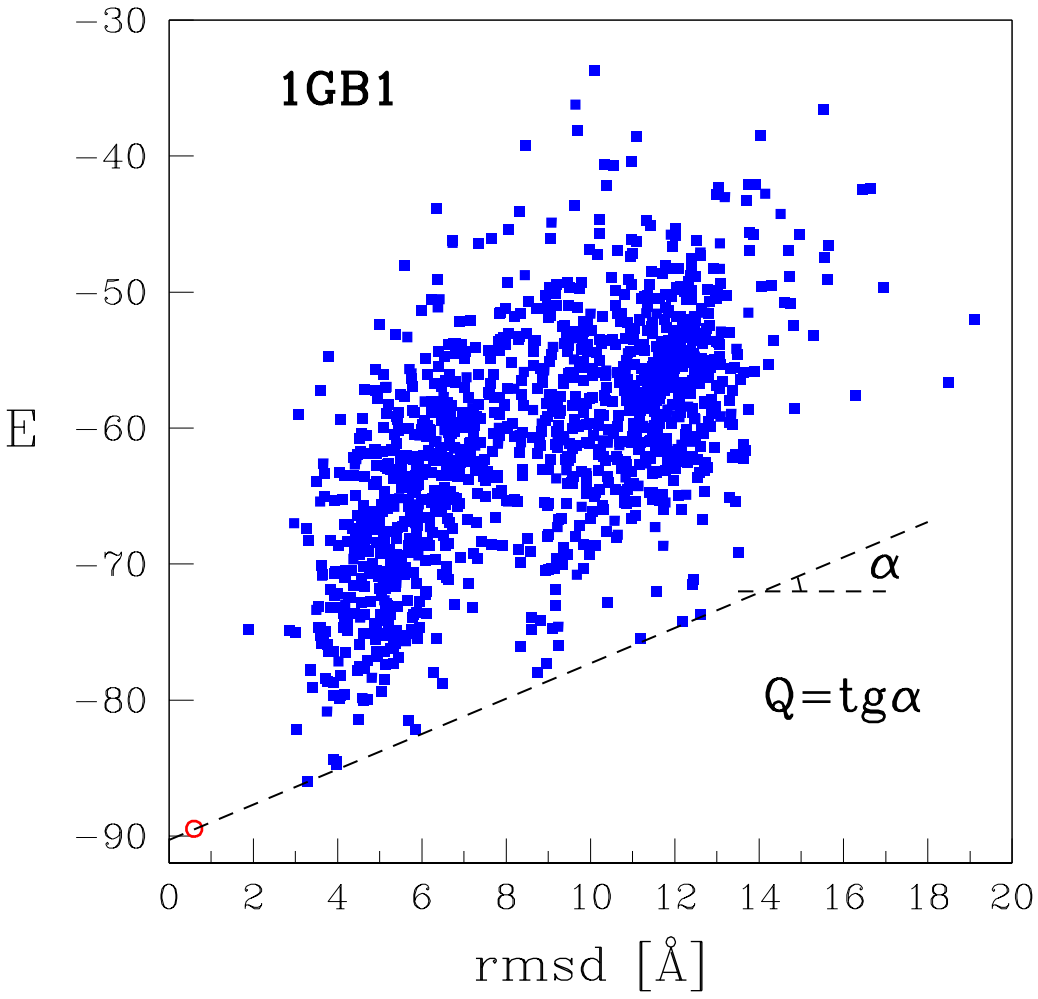

FIG. 2.

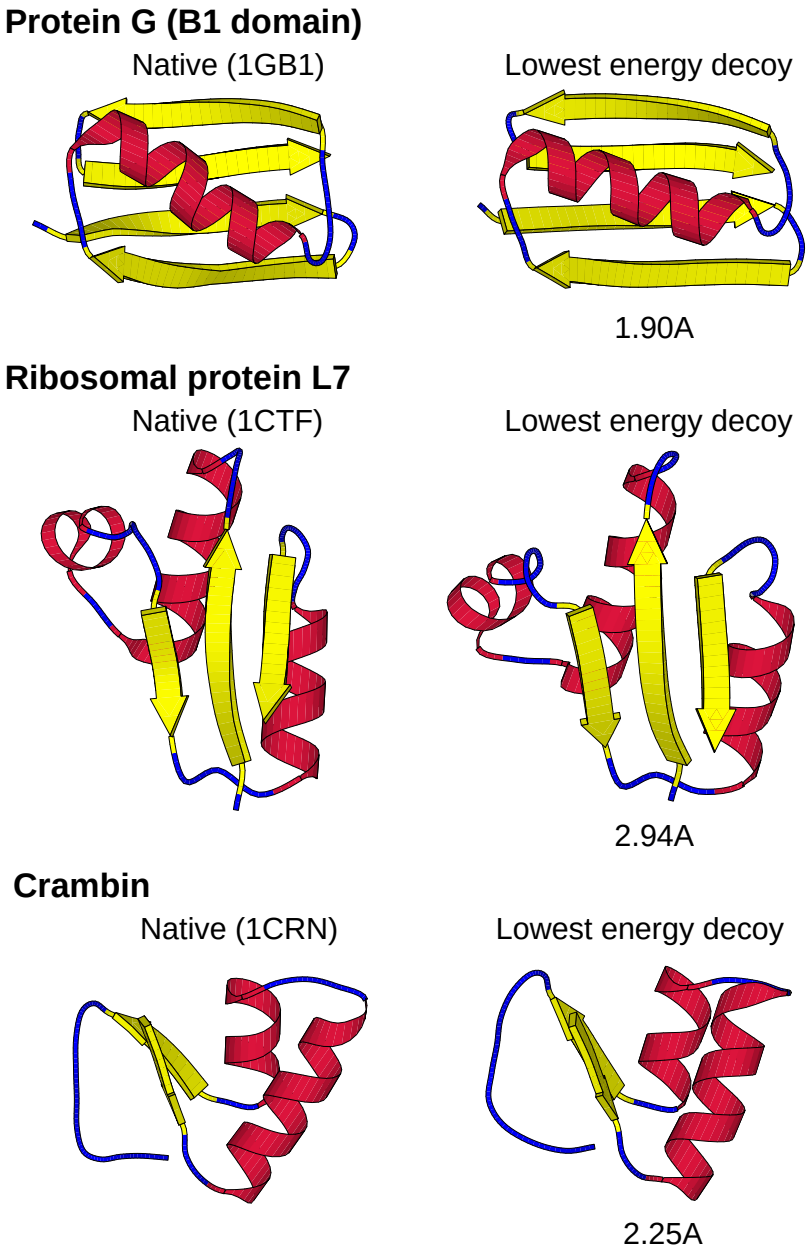
Protein G (B1 domain)
Native (1GB1)
Lowest energy decoy
1.90A
Ribosomal protein L7
Native (1CTF)
Lowest energy decoy
2.94A
Crambin
Native (1CRN)
Lowest energy decoy
2.25A

FIG. 3.

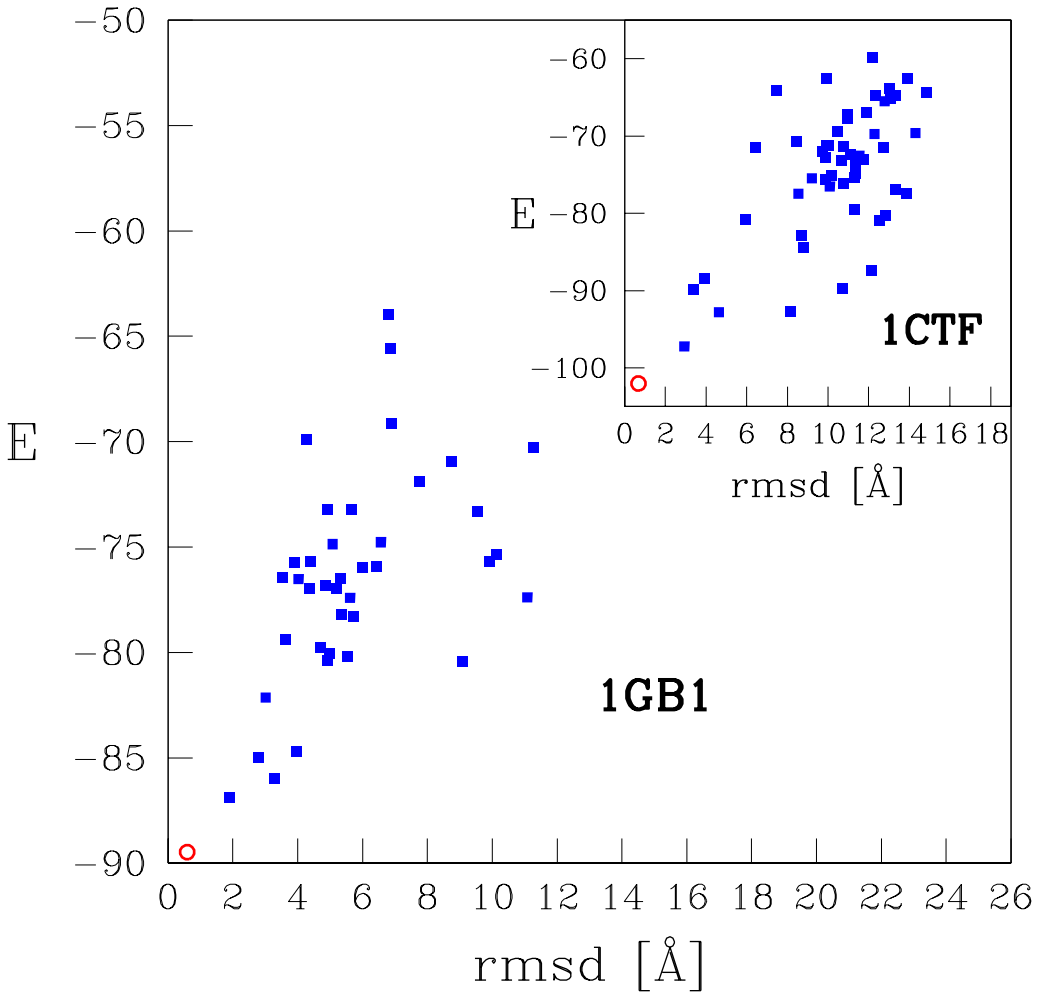

FIG. 4.

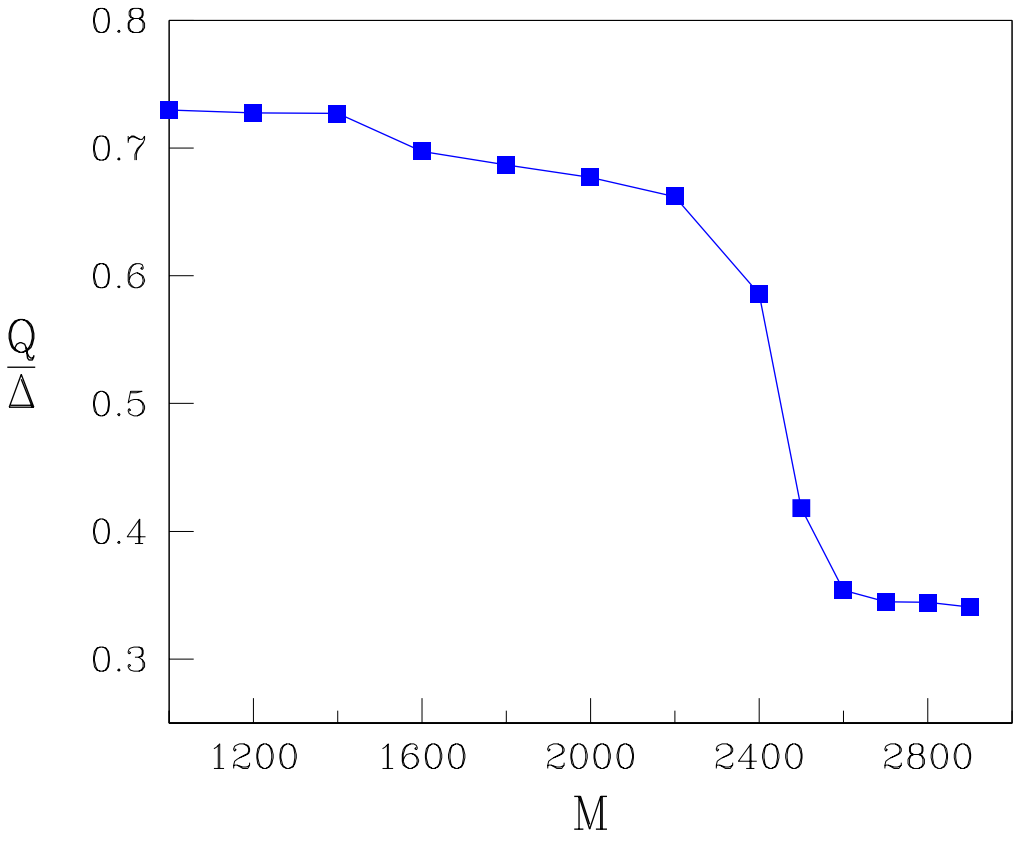

FIG. 5.

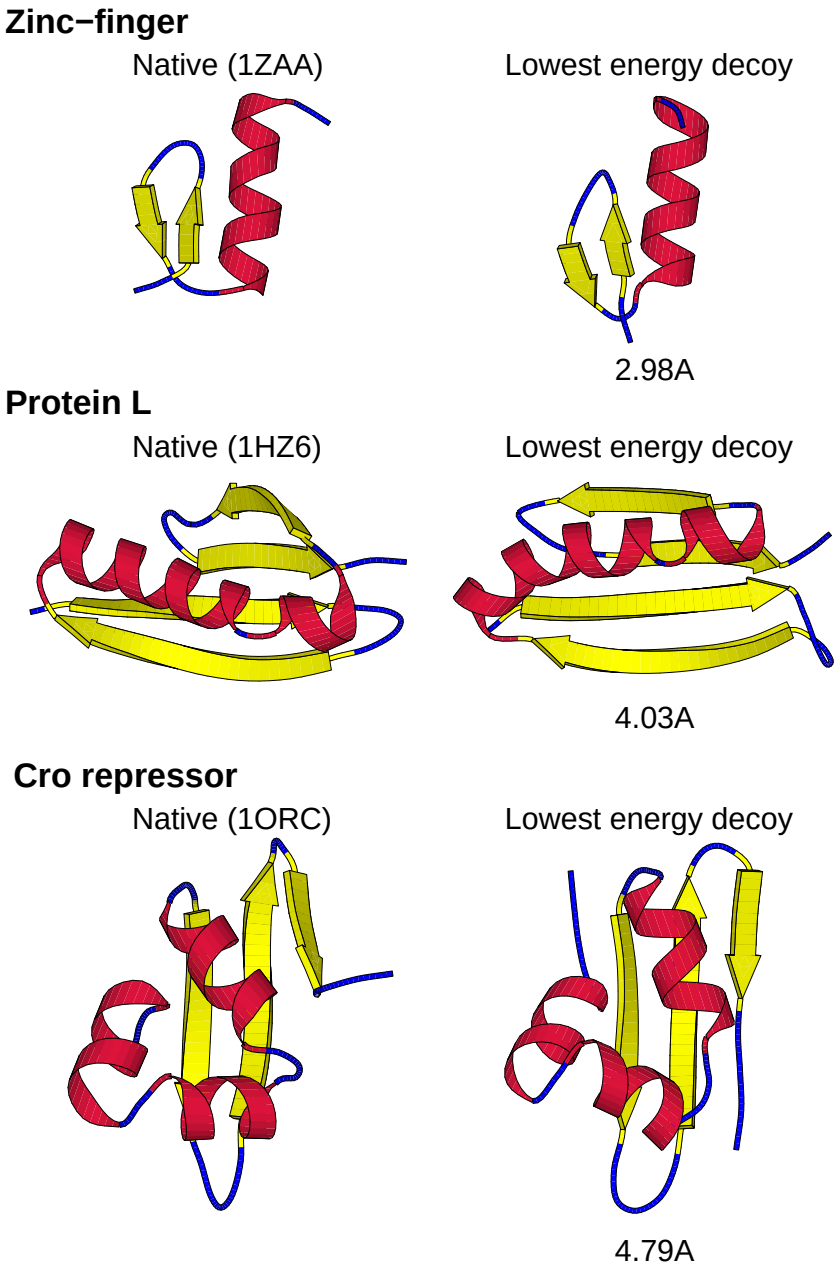
Zinc−finger
Native (1ZAA)
Lowest energy decoy
2.98A
Protein L
Native (1HZ6)
Lowest energy decoy
4.03A
Cro repressor
Native (1ORC)
Lowest energy decoy
4.79A

FIG. 6.

Hyperthermophile protein

Native (1SAP) Lowest energy decoy

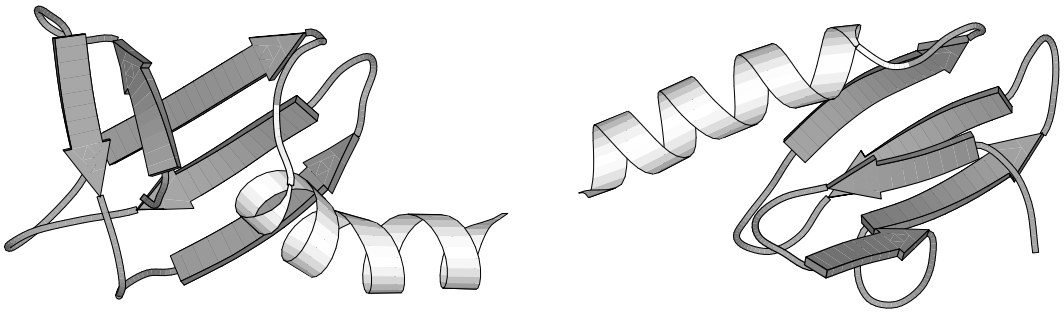

FIG. 7.

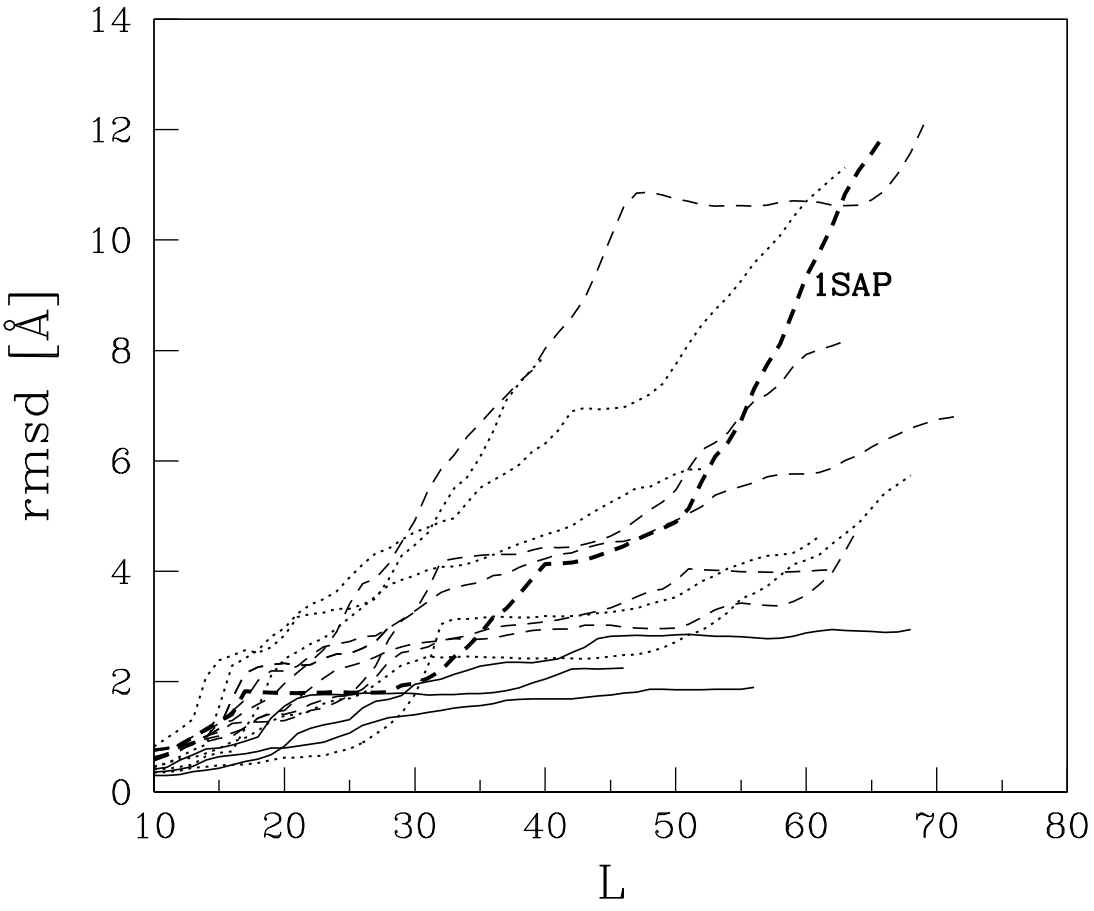

FIG. 8.

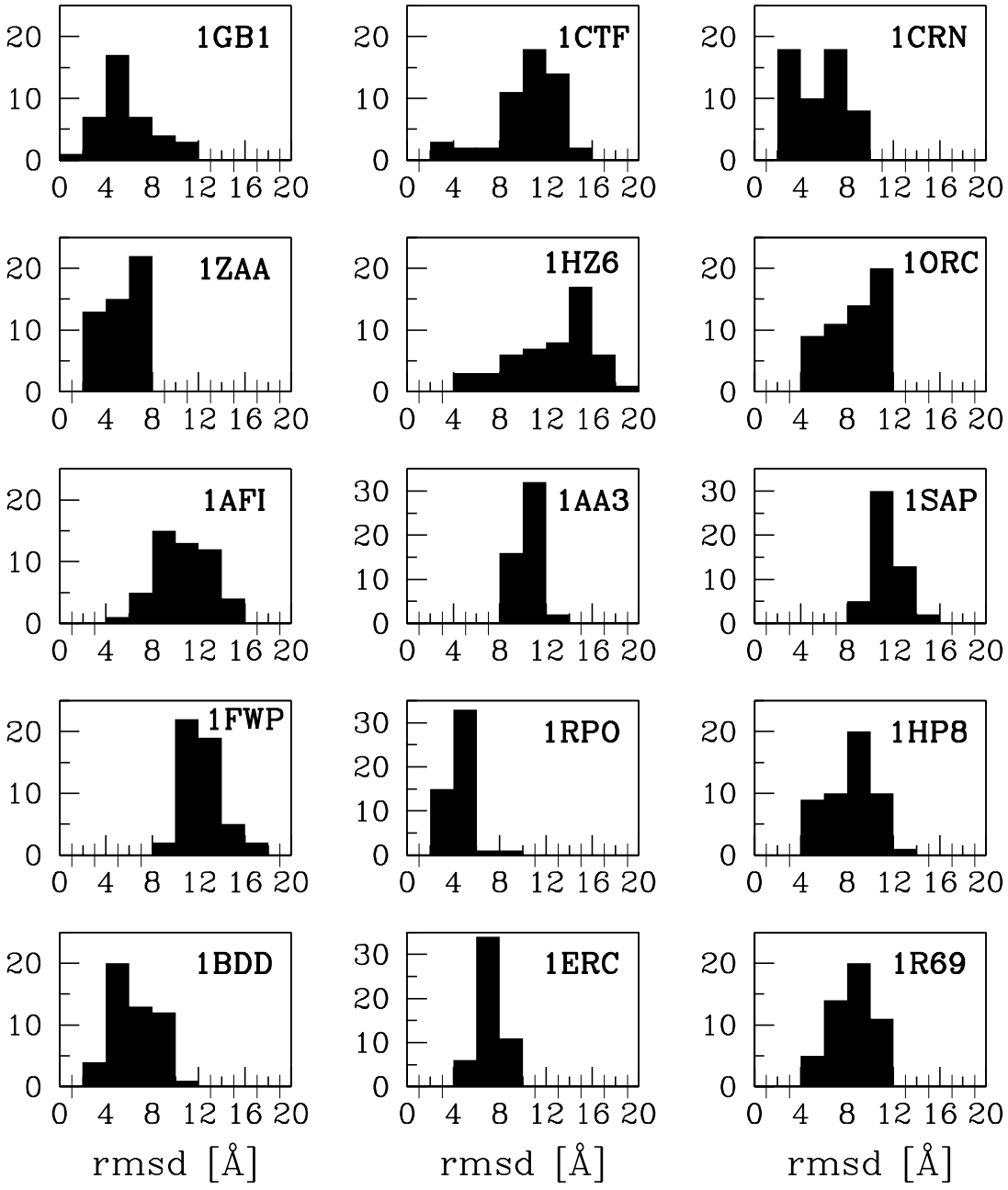
1GB1
1CTF
1CRN
1ZAA
1HZ6
1ORC
1AFI
1AA3
1SAP
1FWP
1RPO
1HP8
1BDD
1ERC
1R69
rmsd [Å]
rmsd [Å]
rmsd [Å]

FIG. 9.

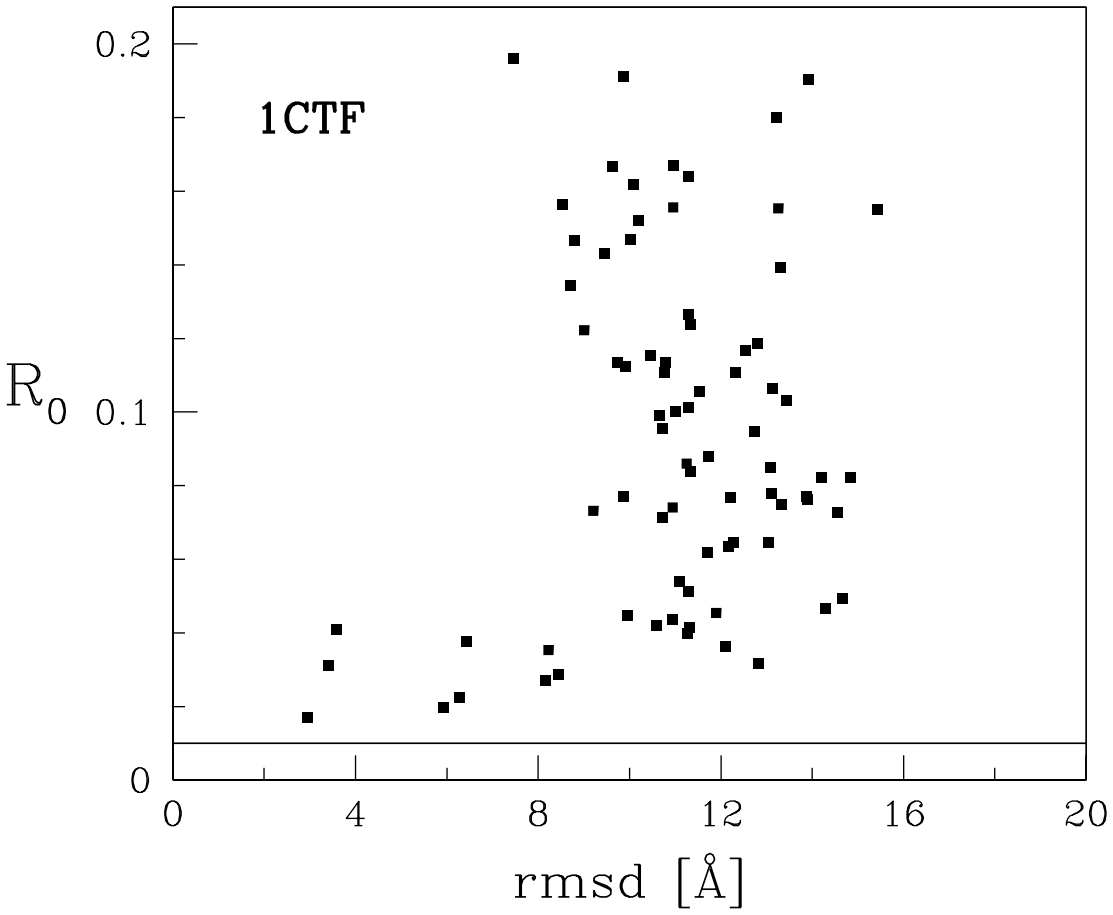

FIG. 10.