doi:10.1016/S0022-2836(03)00323-1 J. Mol. Biol. (2003) 329, 159–174

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

A Novel Approach to Decoy Set Generation: Designing a Physical Energy Function Having Local Minima with Native Structure Characteristics

Chen Keasar* and Michael Levitt

Department of Structural Biology, Stanford School of Medicine, Stanford CA 94305, USA

**Corresponding author*

We suggest a new approach to the generation of candidate structures (decoys) for *ab initio* prediction of protein structures. Our method is based on random sampling of conformation space and subsequent local energy minimization. At the core of this approach lies the design of a novel type of energy function. This energy function has local minima with native structure characteristics and wide basins of attraction. The current work presents our motivation for deriving such an energy function and also tests the derived energy function.

Our approach is novel in that it takes advantage of the inherently rough energy landscape of proteins, which is generally considered a major obstacle for protein structure prediction. When local minima have wide basins of attraction, the protein's conformation space can be greatly reduced by the convergence of large regions of the space into single points, namely the local minima corresponding to these funnels. We have implemented this concept by an iterative process. The potential is first used to generate decoy sets and then we study these sets of decoys to guide further development of the potential. A key feature of our potential is the use of cooperative multi-body interactions that mimic the role of the entropic and solvent contributions to the free energy.

The validity and value of our approach is demonstrated by applying it to 14 diverse, small proteins. We show that, for these proteins, the size of conformation space is considerably reduced by the new energy function. In fact, the reduction is so substantial as to allow efficient conformational sampling. As a result we are able to find a significant number of near-native conformations in random searches performed with limited computational resources.

Keywords: ab initio; energy function; optimization; local minima; conformation

Introduction

Ab initio protein structure prediction: global optimization *versus* divide and conquer

After almost four decades of intense research, the knights of computational structural biology are still on the roads. The Holy Grail they seek is a deciphering mechanism for the enigmatic coding of protein structures by their amino acid sequence. One strategy, adopted in the pioneering attempts[1,2] and still widely used,[3,4] is global optimization of an energy function that approximates the actual free energy landscape governing protein folding. This is a very attractive scheme, as it promises to give not only the predicted structure, but also insight into the folding process. Unfortunately, the derivation of appropriate energy functions is still a major problem. Further, global optimization of the energy functions appears to be very difficult and resists even powerful optimization methods such as simulated annealing (SA),[3,5] genetic

Present address: C. Keasar, Departments of Computer Science and Life Sciences, Ben-Gurion University of the Negev, Be'er-Sheva, Israel.

Abbreviations used: RMS, root-mean-square; NMR, nuclear magnetic resonance; BA, basin of attraction.

E-mail address of the corresponding author: keasar@cs.bgu.ac.il

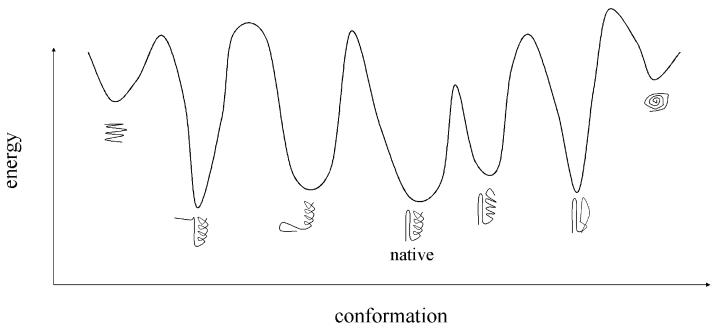

Figure 1. An idealized cartoon of the one-dimensional energy function we try to formulate for decoy set generation. It is dominated by wide and smooth catchment basins leading to local minima. Each of these minima has native structure characteristics like compactness, hydrophobic interior and hydrophilic surface, and extensive secondary structure. The native structure of the protein also has these properties, and as such is expected to be a minimum, not necessarily the global one. Random sampling of conformation space followed by energy minimization is used to generate the decoy set. Due to the wide basins of attraction of the local minima, a manageable decoy set is expected to provide a good representation of the space of all local minima. As such, the decoy set is likely to include the native conformation. Note that as we do not use global optimization techniques, the entire landscape does not need to be funneled towards the native conformation, which is a much stronger requirement.

algorithm (GA)[6] and energy surface smoothing.[7] These two obstacles are related, as global optimization methods perform best when the hyperspace on which they operate is funneled towards the optimal solution. None of the available energy functions for proteins has been shown to have this property.

In an attempt to face these difficulties, an alternative "divide and conquer" scheme emerged.[8] According to this scheme the quest is broken into two almost independent tasks. The first is to create a set of alternative conformations (decoys) for the protein. The second is to choose one of these decoys as a model for the protein's structure. Both of these tasks are hard and do not yet have a satisfactory solution. One may hope, though, that dealing with them separately, possibly using different methods, will make the whole quest easier. The current work concentrates only on the first task, namely creating sets of near-native decoy structures.

Current methods for the generation of decoy sets

The essence of generating decoy sets is reducing the size of the proteins' conformation space. Instead of the overwhelming and non-enumerable space of all possible conformations of a protein, we want a small and enumerable set of conformations that include the native fold or at least an approximation of it. This set should be small enough, so that one can evaluate the conformations by one's favorite scoring function in order to pick the most probable model for the protein structure.

One way to accomplish this task is to exhaustively search the conformation space spanned by a reduced representation of the protein structure,[9–12] or by the structures of proteins that have already been solved.[13] Alternatively, energy-guided sampling of conformation space can be done by a series of optimization simulations, starting from different random conformations.[14–16]

The scope of our work; energy-guided conformation search using local minimization

A straightforward method for energy-guided decoy set generation is to take advantage of the very efficient local minimization procedures available[17] and minimize each of the initial random conformations to the closest local energy minimum. This pathway, however, has not been traversed much since first used by one of us some 20 years ago.[18] Due to the rough energy landscape characteristic of proteins, downhill minimization is expected to reach only shallow local minima. Thus, the resulting decoy set is expected to be of very limited value, including only non-compact structures with low secondary structure content. Instead, non-local optimization procedures (e.g. simulated annealing) were employed.[14–16] Non-local procedures are capable of reaching deep energy wells resulting in compact structures with high secondary structure content. This capability, though, has a high computational price tag with most of the computational effort invested in barrier crossing.

In this work we return to the forsaken path of local minimization by trying to attack the source of its weakness, the apparently inherent roughness of protein energy landscapes.

Local minima with native-like characteristics and wide basins of attraction

Our goal is to design an energy function that obeys two criteria (Figure 1).

(1) The native or at least a close-to-native conformation is a local energy minimum.

(2) The energy landscape is dominated by local minima with wide basins of attraction (BAs).

While the first criterion is a minimal requirement for any potential used for energy-guided conformational search, the second is unique. The

Table 1. Proteins used in the current work

Protein	Number of residues	Number of decoys generated	SCOP class	Comments
A. *Learning set*				
1bba	36	10,000	Peptides	
2cr	65	10,000	α	
1ctf	68	10,000	(α + β)	
1fc2	43	10,000	α	
1gpt	47	10,000	Small	4 SS-bonds
1igd	61	10,000	(α + β)	–
2ovo	56	10,000	Small	3 SS-bonds
4pti	58	10,000	Small	3 SS-bonds
1shf	59	10,000	β	–
1ubi	76	100,000	(α + β)	–
B. *Test set*				
1f0a	80	10,000	(α + β)	–
1jwe	114	100,000	α	–
1e68	70	10,000	α	Cyclic backbone
1d3b	72	100,000	β	–

desirable shape of the energy landscape is obviously one that is funneled towards the global minimum, allowing the global optimization methods to reveal their full strength; finding such a potential is considered to be very difficult indeed. We believe, however, that a systematic derivation of an energy function whose local minima have native structure properties is a feasible task. The current work is intended to demonstrate a systematic derivation of such an energy function and its use for decoy set generation.

Current implementation

Here we have iteratively formulated a novel energy function characterized by local minima with native attributes and large catchment basins. To test the usefulness of the approach we generate libraries of alternative conformations (decoy sets) that hopefully include native-like structures.

As a starting point we begin with a united atom version of the ENCAD forcefield,[18] gradually adding long-range and cooperative terms. Protein structures with native structure characteristics (including of course the native structure) are either minima of the ENCAD forcefield or very close to minima. The inverse is not true in that all minima are not native-like structures: the overwhelming majority of the local minima are neither compact nor have considerable secondary structure content. The ENCAD energy function includes only two-body, short-range terms resulting in very narrow basins of attraction for local minima including the native fold. Any two conformations that differ in as little as one or two favorable interactions are likely to belong to different basins of attractions. This is because moving from one of them to the other requires breaking of one favorable interaction before making another one; such a step is most likely to involve barrier crossing. High-energy non-compact minima with no favorable contacts may have much wider basins of attractions. Indeed, local minimization of extended chains does tend to converge to non-compact minima, which are unlikely to be similar to the compact native fold.

Thus, the essence of converting the ENCAD forcefield to the desired energy function is pruning of local minima. We destabilize the majority of the local minima by applying novel cooperative and/or long-range terms, so that the space of possible conformations is mapped to the relatively few remaining minima. As a result, local minimizations of extended chains are likely to converge to compact local minima, including native-like ones.

The iterative derivation of the current potential

A rather unique aspect of our work is the iterative methodology used to derive the energy function. At each iteration of the derivation process we manually study the local minima of an energy function that is somewhat modified compared with the one used in the previous iteration. For each of the proteins in our learning set (see Table 1) 10,000 to 100,000 decoy conformations are generated. If the fraction of native-like structures among the decoys increases at least for some of the proteins and does not decrease for the others, we accept this new version of the potential and use it as the starting point for the next iteration.

More often than not, manipulation of the energy function resulted in unexpected artifacts, and in most cases we rejected the modified energy function and retreated to the previously accepted energy function as the basis for the next attempt. Eventually, however, the accepted modifications resulted in local minima with wider basins of attraction yet include conformations that are similar to the native structures.

Studying the decoy sets was not only used for the evaluation of the energy functions, it also suggested possible directions for further manipulations. We tried to identify general characteristics (such as compactness, organized sheets, non-polar cores, etc.), which are not as pronounced in the

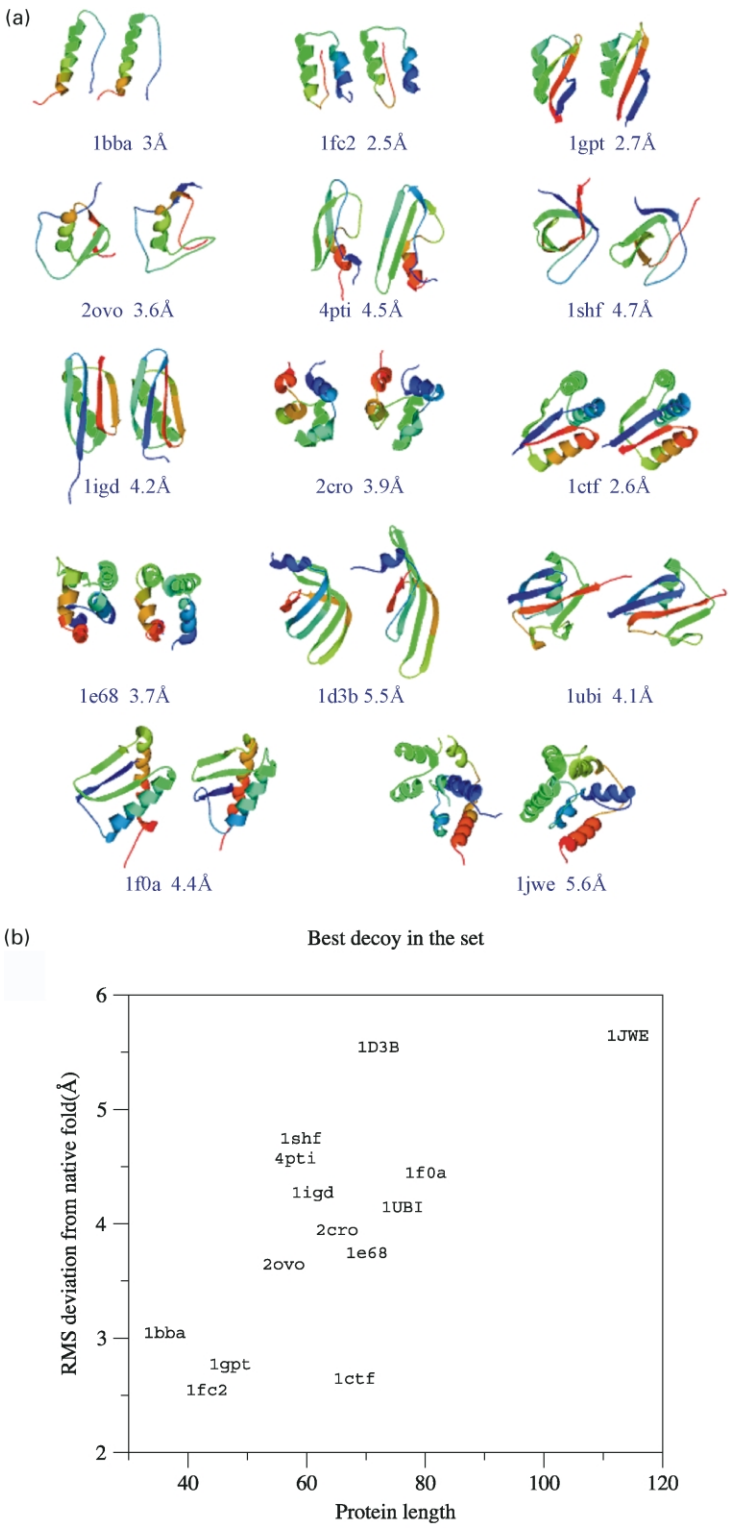

Figure 2 (*legend opposite*)

decoy sets as in native structures. An attempt was made to bias the potential towards these features. On the other hand, we often found peculiarities in the decoy sets that are rare in native structures, such as backbone hydrogen bond patterns that are incompatible with ordered secondary structures. The energy function was designed to penalize these undesirable features.

A similar approach of designing an energy function by iterations of decoy set generation and parameter optimization has been suggested recently.[19] In that work, the parameter optimization

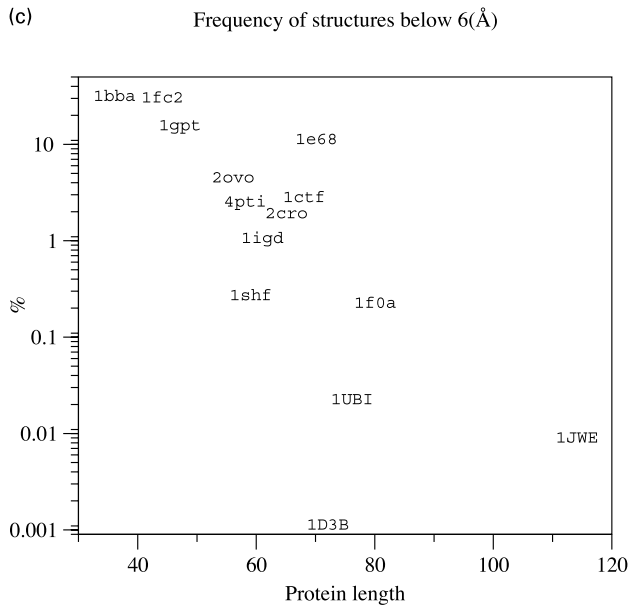

Figure 2. A summary of the 14 decoy sets generated here (Table 1). Each set is comprised of 10,000 (lower case) or 100,000 (upper case) decoys. (a) The best (closest to native) decoy from each set is shown (right) alongside the native structure (left). The chains are color coded by residue number from the N terminus (blue) to the C terminus (red) and we give the coordinate RMS deviation in each case. (b) The coordinate RMS deviation of the best decoy from the native fold is roughly correlated with the length of the protein. (c) The fraction of decoys that are within a 6 Å deviation from the native fold decreases with protein length.

was done by an automatic method and the parameters optimized in an objective and consistent way. We believe that our approach, though less objective, is more flexible and allows the incorporation of human intuition and knowledge.

Results

Decoy sets

Fourteen decoy sets were generated for proteins of diverse sizes and folds (see Table 1). Initially 10,000 decoys were generated for each protein. The decoy sets of three proteins 1ubi, 1jwe and 1d3b included a very small number of native-like structures. For those proteins the sizes of the decoy sets were enlarged to 100,000.

Figure 2 summarizes the main features of the decoy sets. The best (in terms of RMS deviation from the native conformation) decoy of each set is presented in Figure 2(a). They are all within 6 Å RMS deviations from the native fold and the deviation is roughly correlated with protein size (Figure 2(b)). Also, a rough inverse correlation exists between the protein length and the overall fraction of decoys within 6 Å RMS deviations from the native fold (Figure 2(c)).

Correlation between the energy values and the quality of the decoys (in terms of similarity to the native fold) ranges from weak to non-existent and in most cases the lowest energy conformations were not native-like (data not shown). Thus, the energy values cannot serve as a criterion for choosing a reliable model out of the decoy set. Nevertheless, the energy values of the decoys can serve at least to restrict the problem of selection. The subsets of lowest energy decoys are usually enriched in good decoys compared with the whole sets (Figure 3(a)). Further, the best decoys within these subsets are typically comparable to the best decoys in the whole sets (Figure 3(b)). The proteins 1jwe and 1d3b are exceptions with no decoys below 6 Å in their lowest energy subsets.

Conformation space reduction

The aim of our work is to reduce the conformation spaces of proteins to a manageable size. To test whether we made progress in this direction, we compared three decoy sets for each of two "representative" proteins, 2cro (all alpha) and 1shf (all beta). The decoy sets tested were the set of the initial random conformations, a set of local minima generated with the original ENCAD energy function, and the set of local minima generated with the current energy function.

A natural measure of conformation space size is the root-mean-square (RMS) value of the distances between randomly sampled representative structures. This measure of the diameter of the space indeed reduces when we move from the set of initial structures to the minimized ones and reduces further when we move to the local minima of the current energy function (Figure 4).

Yet another way of representing the different conformation spaces is to map them on a

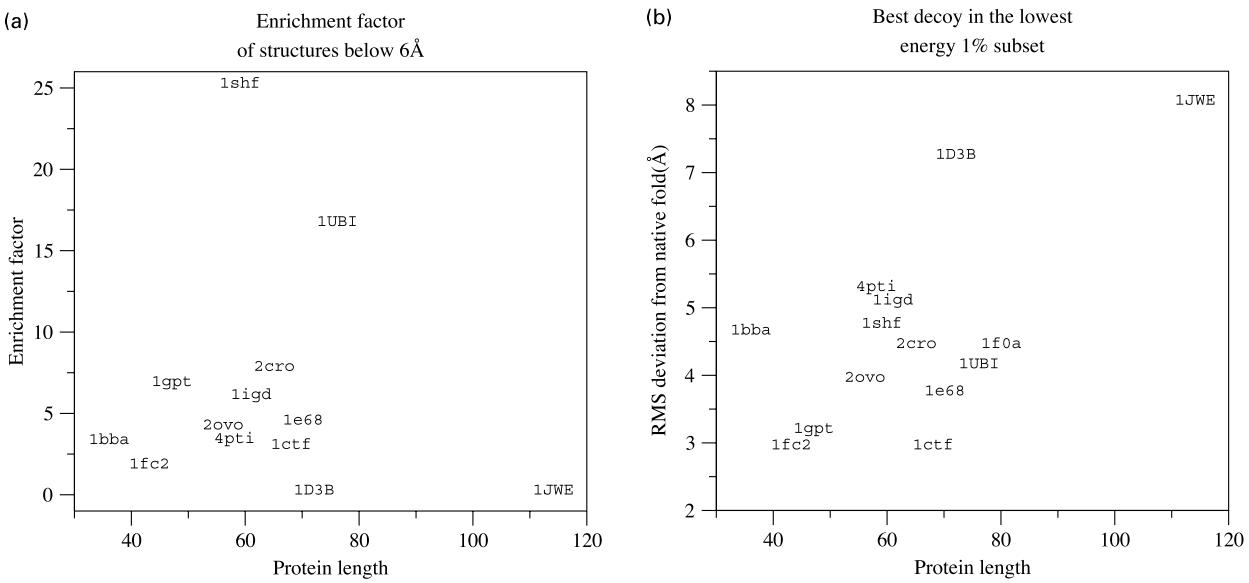

Figure 3. The lower energy decoys are enriched in near-native structures relative to the entire set. In (a) we show the enrichment factor expressed as a percentage and calculated as:

$$\left(\frac{\text{Fraction of decoys below 6 Å in the 1\% lowest energy subset}}{\text{Fraction of decoys below 6 Å in the entire set}}\right)\times 100$$

In (b) we show that the best decoys (smallest coordinate RMS deviation) within the 1% lowest energy subsets are in many cases as close to the actual native structure as the best decoys in the whole sets.

two-dimensional space. The mapping to the space defined by radius of gyration and contact order, is rather telling (Figure 5). It shows a similar trend as shown in Figure 4: the local minima are much closer to the native structure characteristics than the initial random structure, and the local minima with the current energy function are distributed more narrowly and more biased towards the native structure than for the local minima of the original energy function.

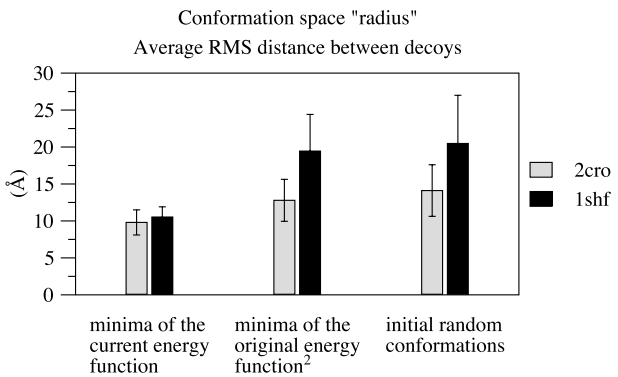

Figure 4. The average RMS deviation between pairs of decoys is used to estimate the sizes of conformation spaces sampled by sets of 10,000 decoys. Three decoy sets for each of 2cro (filled bars) and 1shf (striped bars) were examined. Note that as the conformation spaces are multi-dimensional the rather small changes in the "radius" presented here correspond to very large changes in "volume". The estimation of the dimensionality of the space, however, is beyond the scope of the current work. For both proteins, the initial random conformations span the largest conformation spaces and the local minima of the current energy function span the smallest. Local minima generated by the original ENCAD potential span a conformational space almost as large as that of the initial random conformations.

Discussion

General

This work is a snapshot of a continuous process devoted to the derivation of an energy function with local minima that have native structure characteristics. Fourteen proteins were used in this study. For all of them, local minimization of random, generally extended conformations converged to compact low-energy minima. For all proteins some of these minima were native-like. Thus, near-native structures are minima in our new energy function and their basins of attraction, as well as the basins of attraction of many other compact structures, are wide enough to include extended chains.

With the current potential, up to 100,000 minimization experiments are needed in order to sample the local minima space properly. Thus, due to the efficiency of the minimization process itself, the sampling process consumes manageable computational resources.

The set of proteins used

A method in the field of protein structure

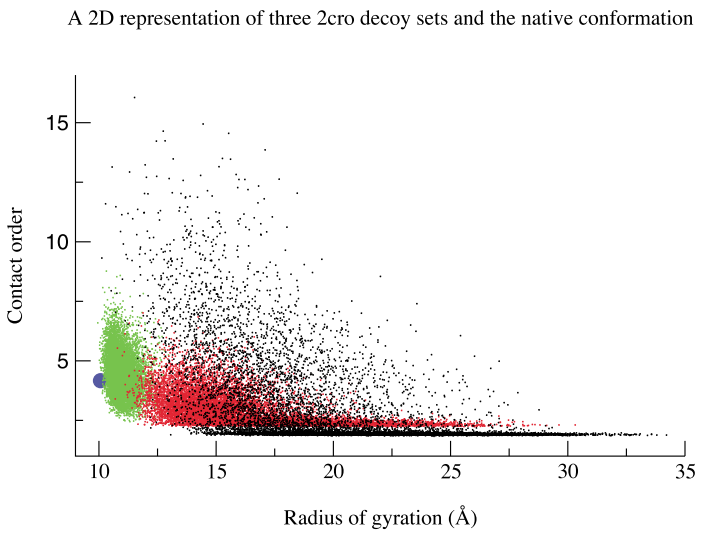

Figure 5. Mapping three sets of 2cro conformations onto the radius of gyration *versus* contact order space. The black dots represent 10,000 random, generally extended initial structures. The red and green dots represent the local minima reached by energy minimizing these initial conformations under the original ENCAD potential and our new energy function, respectively. Clearly, the local minima of the current energy function spans a much smaller space then the local minima of the original energy function and are, on average, closer to the native fold. Here, the radius of gyration is the root-mean-square of C^{α} distances from their center of mass, and contact order is the sequence separation between contacting residues. Residues are considered in contact if the distance between their C^{α} atoms is less then 6.5 Å.

prediction can be considered valid only if it has been tested on a large number of proteins of different folds. Recent work has set very high standards for the size and diversity of the protein set.[14] The 14 proteins used in the current work (Table 1) are more limited: in allocating resources between better testing and further development of the energy function, we mainly emphasized the latter. We believe, however, that within the domain of small water-soluble proteins, our set of 14 proteins is large and diverse enough to ensure that overall our approach is not tailored for the specific peculiarities of the chosen proteins. On the other hand, due to the small number of proteins used, we consider this work mainly a feasibility test for our novel approach and not yet a ready-to-use recipe for protein structure prediction.

The use of predefined secondary structure

A major characteristic of our approach is that secondary structure is predefined and effectively fixed along the simulation. The initial structures are generated with the predefined secondary structure elements and the energy function depends strongly on them.

(1) The backbone torsion angles of predefined strands and helices are restricted to their characteristic values.
(2) The cooperative hydrogen bond term applies only to predefined secondary structure elements. Further, it is applied differently to strands and helices.
(3) The hydrophobic term ignores side-chains outside the secondary structure elements.

The advantage of this approach is clear when the secondary structure is available before the tertiary structure, as is often the case for 2D NMR experiments. The advantage is less clear when predicted secondary structure is used. We have not studied thoroughly the sensitivity of our results to the secondary structure assignment accuracy.

The iterative approach to energy function development

The knowledge-based potentials that currently dominate the field of protein structure prediction are automatically derived from the proteins' sequence and structure databases. The iterative scheme we use to develop the energy function is rather unique as we manipulate the formula "by hand" without relying on any data mining and analysis. This process is inherently imprecise and may be biased by our preconceptions. We believe, however, that the risk is more than compensated for by the opening of large space for human intuition and knowledge. A similar approach has been shown useful in the related field of cyclic peptide structure prediction.[20]

The route presented here is a very general one. Taken by other people it may well lead to different functional forms and hopefully to better results than those we present here. This should not, however, be interpreted as if only the people who derived the energy function can use it in a consistent way. Once a certain version of the energy function is set, its usage is as straightforward as any other empirical energy function.

Energy *versus* similarity to the native fold

An ideal potential for protein structure prediction would have good correlation between energy and the similarity to the native fold. Our current approach emerged from recognizing that such a potential is currently not at hand. Still it might be important to identify aspects of the current energy

function that consistently reward some of the non-native structures, making them lower in energy then native-like structures.

(1) Native protein structures do not maximize the number of hydrogen bonds within beta-sheets. Thus, due to the cooperative nature of our hydrogen bond potential, structures with slightly more hydrogen bonds than the native tend to be much lower in energy. These structures are typically less compact then the native-like structures and tend to have higher hydrophobic energy values.

(2) Both the torsion constraint and the hydrogen bond potential favor planar beta-sheets while the beta-sheets in native structures tend to be twisted.

(3) The long-range hydrophobic potential favors an overall spherical shape. Thus, spherical decoys have lower hydrophobic energy than native-like decoys when the native structure is more elongated.

(4) Burial of loop regions and helix ends is not penalized in the current energy function, allowing very compact non-native structures that are favored by the hydrophobic term.

We hope to solve some of these problems in the next versions of our potential and thus achieve somewhat better energy–RMS correlation.

Computational considerations

An important advantage of our approach is the natural way it can be parallelized. The task of decoy set generation can be distributed among as many computers as are available simply by allocating different ranges of random number seeds to each computer. As no communication is required between these processes there is practically no overhead to the parallelization. This allows the use of large arrays of cheap computers and the efficient utilization of existing computational resources.

Conclusion

An energy function whose local minima have wide basins of attraction, and at the same time have native structure characteristics, has been derived, and may serve as a useful tool for the sampling of protein conformations. Currently other approaches have been shown to generate better decoy sets than the ones presented here.[14] We believe, however, that further improvements in the potential may result in an interesting alternative to the current leading methods.

While the problem of conformation space sampling is still far from an optimal solution, the current work together with other recent publications,[11,14,21,22] suggests that currently this is not the limiting factor for *ab initio* protein structure prediction, at least in the case of small water-soluble proteins. The complementary problem of picking the close to native structure out of the generated decoy set appears to be much harder.[15,23,24]

Methods

Proteins used

A total of 14 small (36–114 residues), water-soluble proteins were used here (Table 1). The proteins were added gradually as work developed. Four of them, however (1jwe, 1f0a, 1e68 and 1d3b), were added after the current functional forms and parameterization were set. They allowed us to test the current energy function on proteins for which it was not "tailored". This separation between learning and test sets is somewhat arbitrary. In the future steps of energy function development, the results for all these proteins will be taken into consideration and more proteins will be used for testing.

Generation of initial random structures

The initial structures were built using standard bond lengths and angles. Torsion angles were assigned in the following way.

(1) All peptide bonds were built in a *trans* configuration.

(2) χ_1 torsion angles were set to $-60°$ and all the other χ torsion angles to 180°.

(3) The predefined helices and beta-sheet strands were generated with the ideal (Φ,Ψ) torsion angles of ($-60°$, $-40°$) and ($-120°$, 150°), respectively.

(4) Other residues were built with (Φ,Ψ) values randomly distributed in the ranges: $-120(\pm 60)°$ and $150(\pm 90)°$, respectively. This choice of values for the loop (Φ,Ψ) angles biased the set of initial conformations towards rather extended ones.

It should be noted that the differences in loop (Φ,Ψ) values between the initial conformations were the only source of diversity in our system. As such, a wider distribution could be expected to result in better sampling of conformation space. In practice, however, wider ranges of (Φ,Ψ) angles resulted in tangled initial conformations and the subsequent minimization tended to freeze in non-native knots.

Energy minimization

Energy minimization was performed in torsion angle space[18] using the Fletcher's VA09D variable metric routine.[17] We find that this routine is both efficient and robust.

(1) On average it converges after a small number of energy evaluations.

(2) We have not encountered round-off problems.

(3) Most important, the routine is very sensitive to errors in the derivation of the energy function, which made it easier to debug the program.

The minimization trajectories typically began with a fast collapse followed by slower rearrangement to the final structure (Figure 6). Due to the long-range nature of the potential (see below), stabilizing contacts created

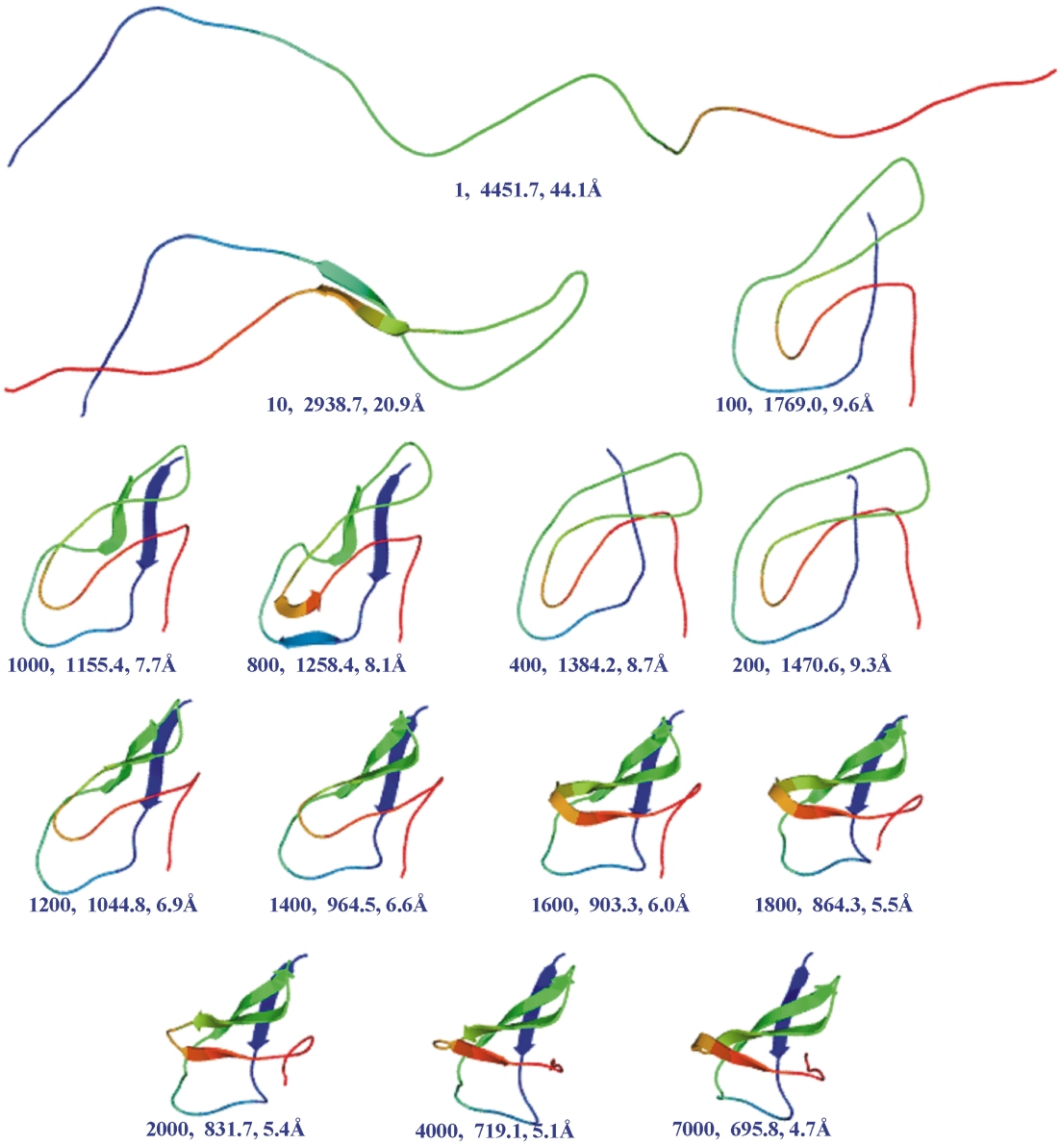

Figure 6. A minimization trajectory of the all-beta protein 1shf is presented by 14 snapshots. The chains are color coded by residue number from the N terminus (blue) to the C terminus (red); secondary structure involving hydrogen bonds between beta-strands is indicated by the wide ribbons. Each snapshot is labeled by the number of minimization steps, the energy (in kcal/mol) and the RMS deviation from the native fold. During the first 1200 steps, a 98% reduction in energy is accompanied by a rapid collapse of the extended initial conformation to a rather compact structure. This rapid collapse is followed by a much slower phase; 5800 more steps are required for the chain to reach the 0.0085865 (kcal/mol)/radian gradient, which we consider convergence. The modest 30% reduction in energy during this phase is accompanied by a rather significant rearrangement of the chain leading to the compact and native like final structure. These 7000 steps take a total of five minutes on a Pentium II 400 MHz CPU.

during the collapse phase may be broken during the minimization to allow the creation of other, stronger or more numerous interactions (Figure 7).

Energy function

Energy terms inherited from ENCAD

The four energy terms of the united atom ENCAD potential[18] are also used in our energy function.

(1) Soft atom van der Waals term.
(2) Periodic torsion angle term.
(3) Harmonic torsion constrains for (Φ,Ψ) angles in predefined secondary structure elements.
(4) Linear bond term for SS bonds (the energy increases linearly as the bond length deviates from the target value of 2.8 Å).

There is no explicit treatment of electrostatic forces and hydrogen bonds are favorable due to the van der Waals parameters of hydrogen and oxygen atoms.

As these four terms have been thoroughly discussed elsewhere,[18] they will not be described in detail here. A major characteristic of the first three terms should, however, be emphasized: they are all short-range terms. Together with the ideal bond lengths and angles inherent

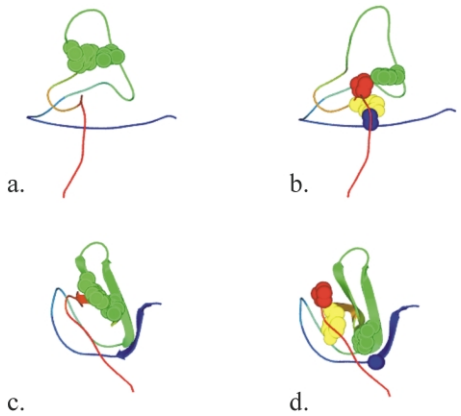

Figure 7. With our energy function, a few favorable contacts do not make a local minimum as the energy surface is dominated by long-range energy terms. This is demonstrated by following the fate of three hydrophobic contacts, which were formed at the beginning of the minimization trajectory presented in Figure 6. After as few as 100 minimization steps, the initial extended structure has already converged into a rather compact structure with three clear favorable hydrophobic contacts. In a there is a Leu28–Trp37 contact shown as green space-filling atoms. In b there are contacts between Ala6–Tyr49 (blue and yellow, respectively) and Phe26–Pro51 (green and red, respectively). In c we see that of these three different contact pairs, only one (Leu28–Trp37) survives. In d we see that the other two break during the minimization allowing further rearrangements.

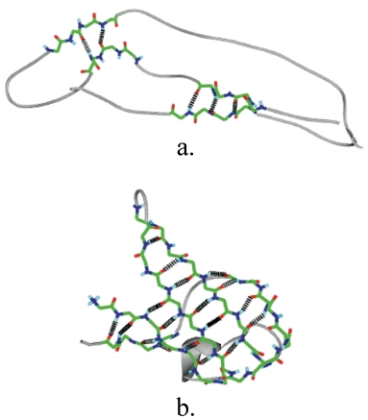

Figure 8. A demonstration of the cooperative hydrogen bond term. a, The lowest energy of 10,000 randomly generated decoys of 1shf using the original ENCAD potential. Ordered arrays of hydrogen bonds are rare and very short. Once a hydrogen bond is created, it cannot be broken and the system quickly freezes. A structure with more hydrogen bonds will be lower in energy, but it is very unlikely to be found as the result of a minimization. b, The lowest energy of 10,000 randomly generated decoys of 1shf using the current energy function. Now there are many more hydrogen bonds that form the regular patterns characteristic of antiparallel beta-sheet structure.

in torsion angle minimization, these terms assure that all generated structures are chemically reasonable at least in the local sense (good atom contacts, correct geometry, etc.). Native structures have, in general, low energy values for these terms, though they are typically not the global minima.

The fourth (SS-bonding) term is the exception. It is both long-range and soft. That is, even when the two sulfur atoms are far away the energy associated with the bonds is neither zero (like in the case of van der Waals), nor high enough to dominate the minimization (like a "normal" harmonic bond term). The efficiency of this "non-physical" term in forming SS-SS bonds[18] encouraged us in the derivation of the somewhat similar global structure terms described below.

Global structure terms

The other set of terms is presented here for the first time and will be described in detail. They are intended to capture some of the more global aspects of "being a native structure". The derivation of these terms is at the heart of our approach. Thus, before the formal definition of each term we present the motivation behind it as well as a brief history of its derivation. It should be noted that our method of choice for structure generation, namely minimization, restricts us to functional forms that have continuous analytical first derivatives.

Zipper-like hydrogen bond term for beta-sheets

Motivation. Hydrogen bonds are favored by the original ENCAD forcefield due to a careful parameterization of the van der Waals interactions between oxygen and polar hydrogen atoms. Thus, structures with a high secondary structure content (including native structures) are low in energy. However, the short-range nature of the hydrogen bonds tends to result in many local minima with narrow basins of attraction. During an energy minimization of a beta-sheet-containing protein, the first few hydrogen bonds appear randomly. In most cases they do not constitute an organized pattern. Once formed, they are unlikely to break and the minimization freezes in a high-energy local minimum with small and sparse beta-sheet regions (Figure 8a).

Derivation. Our first naïve attempts to solve this problem involved explicit hydrogen bond terms making them more favorable than the original function does. Contrary to our expectation the number of hydrogen bonds decreased, probably since the numerous redundant local energy minima deepened. Making the hydrogen bonds more "long-range" by widening their energy wells also appeared to be a bad direction: when a large fraction of the energy gain is reached while the hydrogen and oxygen atoms are still far away, multiple hydrogen bonds with non-physical geometry become more energetically favorable than a single proper one.

We started to get somewhat better results only when we moved to a cooperative term that operates on pairs of hydrogen bonds. With this term, a single hydrogen bond does not contribute to the energy. It does, however, favor the formation of other hydrogen bonds that belong to the same secondary structure element. Thus, the formation of a single hydrogen bond is likely to result in the formation of other bonds, which then give rise to the formation of yet other ones. As a result, local minima of the current energy function are likely to include large beta-sheet regions (Figure 8b).

An unanticipated result of this term was the emergence of non-physical structures (double and triple helices for example) alongside with the normal parallel and antiparallel beta-sheets. Fortunately, we were able to identify within these structures patterns of hydrogen bond pairs that are either rare or totally absent in native protein structures. Penalizing these pairs increased the energy of the non-physical structures and they are now very rare in the sets of local minima.

Lists of hydrogen bonds and hydrogen bond pairs for beta-sheets. In the initialization phase of each simulation four lists are built: a list of hydrogen bonds between segments of predefined beta-strands; a list of favorable antiparallel pairs of hydrogen bonds, a list of favorable parallel pairs of hydrogen bonds; and a list of unfavorable pairs. It should be noted that these lists include only a subset of the hydrogen bonds in the protein and only a small fraction of the possible hydrogen bond pairs.

The lists are built using the following set of rules.

(1) Currently, we take into consideration only hydrogen bonds between amide hydrogen atoms and carbonyl oxygen atoms that belong to different beta-strands. Each hydrogen bond is associated with a donor residue and an acceptor residue and is said to connect two beta-strand segments. A residue is involved in a hydrogen bond pair if it is a donor or acceptor residue of at least one of the hydrogen bonds. Note that each of the residues in the predefined beta-strands is actually involved in many pairs.

(2) A pair of hydrogen bonds is considered a favorable antiparallel bond if it satisfies the following conditions (Figure 9a):

(i) The two hydrogen bonds connect the same two beta-strands. On each of these strands, one or two residues may be involved.

(ii) These strands are either consecutive, or are separated by more then one secondary structure element.

(iii) If in one of the segments only one residue is involved in the pair, then this is also the case in the other strand (see the pair of hydrogen bonds between residues i and $j+8$ in Figure 9a).

(iv) If in one of the segments residues m and $m+l$ are involved in the pair (where $l = 2$, 4, 6 or 8), then they are bonded to residues $n+l$ and n, respectively, on the other strand.

(3) A pair of hydrogen bonds is considered a favorable parallel bond if it satisfies the following conditions (Figure 9b):

(i) The two hydrogen bonds link the same two beta-strands. On each of these strands, one or two residues may be involved.

(ii) These strands are non-consecutive.

(iii) If in one of the segments only one residue is involved in the pair, then two residues are involved in the other segment. Of these two residues the one that is a hydrogen bond acceptor is two sequence positions before the other (see the hydrogen bonds between residues i, j and $j+2$ in Figure 9b).

(iv) If in one of the segments residues m and $m+l$ are involved in the pair (where $l = 2$, 4, 6 or 8), then they are bonded to residues n and $n+k$, respectively, on the other strand (where $k = l-2$, 1, or $l+2$ and $0 < k < 10$).

(4) A pair of hydrogen bonds is considered an unfavorable one if it satisfies at least one of the conditions listed below:

(i) The pair satisfies all the conditions for a favorable antiparallel pair except the segments are separated by a single secondary structure element.

(ii) The pair satisfies all the conditions for a favorable parallel pair, except the segments are consecutive.

(iii) One atom is involved in the two hydrogen bonds (Figure 9c).

(iv) In one of the segments residues m and $m+l$ are involved in the pair (where $l = 0$, 2, 4, 6 or 8) and they are bonded to residues $n+k$ and n, respectively, on another strand (where $0 < k < 8$ and $k \neq l$) (Figure 9c).

(v) In one of the segments residues m and $m+l$ are involved in the pair (where $l = 0$, 2, 4, 6 or 8) and they are bonded to residues n and $n+k$, respectively, on another strand (where $0 < k < 8$, $k \neq l, k \neq l-2$ and $k \neq l+2$) (Figure 9d).

(vi) The two hydrogen bonds connect two strands to a third one and on that strand the separation between the residues involved is 0, 2, 4, 6 or 8 (Figure 9d).

(vii) The two hydrogen bonds connect two consecutive strands to the next consecutive strand (Figure 9e).

Obviously, not all favorable hydrogen bond pairs can have a significant energy contribution at the same time. Different minimizations end up with different patterns (Figure 10) and each pattern is stabilized by a different set of favorable hydrogen bond pairs.

Formal definition of the energy function. The hydrogen bond is a sum of two terms.

(1) The cooperative attractive pair term discussed above.

(2) A non-cooperative core term that prevents the collapse that the first term would have caused:

$$\text{Ehb} = \text{Ehb\_pairs} + \text{Ehb\_core}$$

The hydrogen bond pair term is defined for every pair of hydrogen bonds $(\text{hb}_{i1}, \text{hb}_{i2})$ denoted as hbp_i:

$$\text{Ehb\_pair} = \sum_{\text{hbp}_i} \alpha_{\text{hbp}_i} \beta_{\text{hb}_{i1}} \beta_{\text{hb}_{i2}}$$

where:

$$\alpha_{\text{hbp}_i} = \begin{cases} 20, & \text{if an unfavorable pair} \\ -20, & \text{if a favorable pair} \end{cases} \quad \text{(kcal/mol)}$$

and, for $j = 1$ or 2:

$$\beta_{\text{hb}_{ij}} = \begin{cases} 0, & \text{if } d_{\text{hb}_{ij}} \geq 10 \text{ Å} \\ C(d_{\text{hb}_{ij}} - 10)^2, & \text{otherwise} \end{cases}$$

The constant C equals $0.01506/\text{Å}^2$, resulting in a β value of 1 for an ideal hydrogen bond with $d_{\text{hb}} = 1.85$ Å. The formation of pairs of hydrogen bonds is either strongly rewarded or strongly penalized, whereas a single hydrogen bond makes no contribution to this energy term.

The core term is a step function (the rising part of which implemented by a Gaussian term to make it smoothly differentiable) of d_{hb_i}, the distance between the hydrogen and the oxygen atoms in the

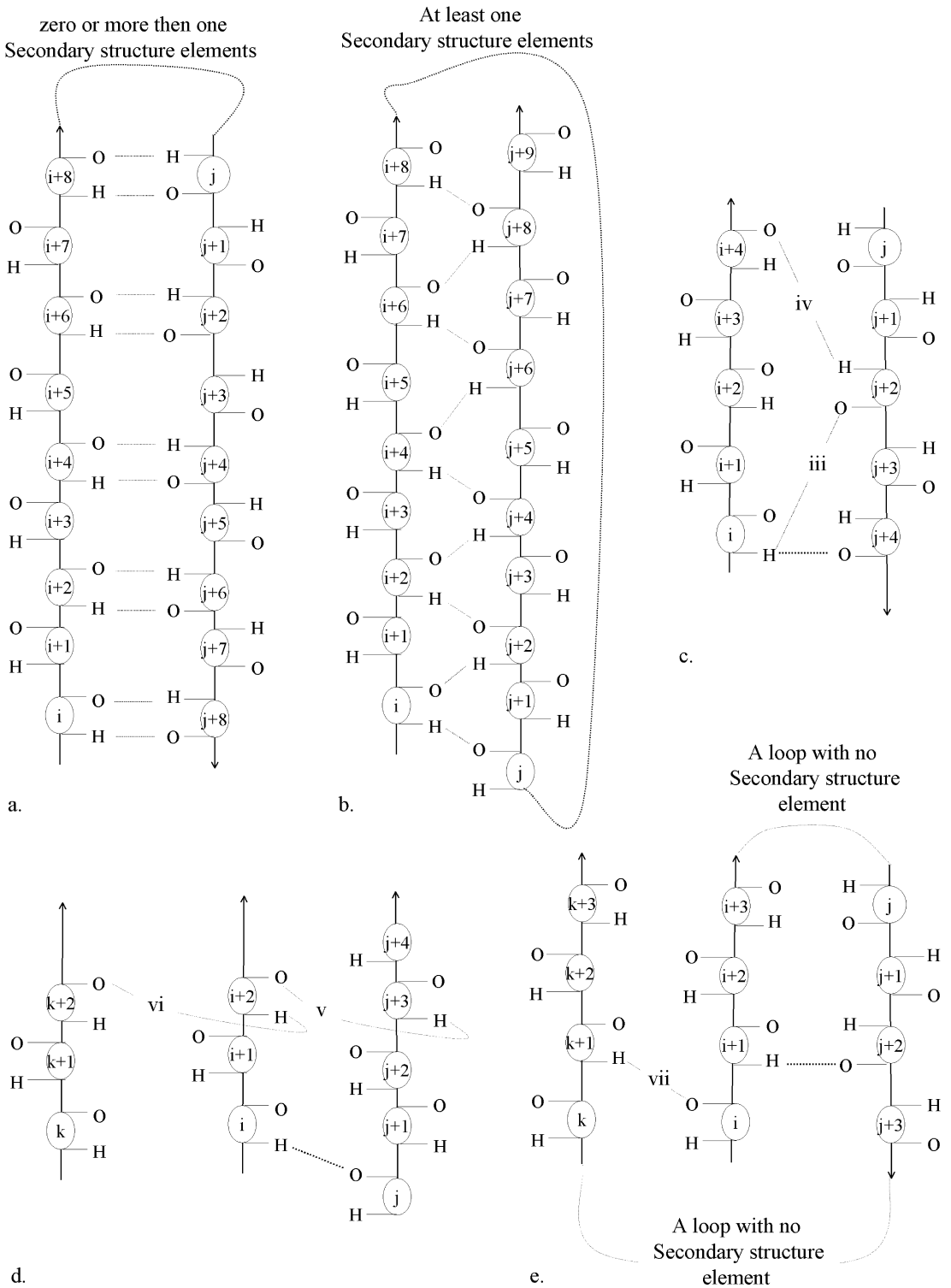

Figure 9. Favorable and unfavorable hydrogen bond pairs. a, Each pair of the hydrogen bonds indicated is a favorable antiparallel pair. Note that for clarity only a few pairs are shown. b, Any pair of the hydrogen bonds indicated is a favorable parallel pair. c–e, Each of the hydrogen bonds indicated by thin line makes an unfavorable pair with the hydrogen bond indicated by a heavy line. The labels on the hydrogen bonds indicate the rules (see the text) that they illustrate.

*i*th hydrogen bond:

$$\text{Ehb\_core} = \begin{cases} C\sum_{\text{hb}_i} \exp[-\alpha(d_{\text{hb}_i} - 1.75)^2], & \text{if } d_{\text{hb}_i} \geq 1.75 \text{ Å} \\ C, & \text{otherwise} \end{cases}$$

where the constants C and α are set to 500 kcal/mol and 1000/Å^2,respectively. This term is negligible at distances larger than the ideal hydrogen bond distance (1.85 Å) but increases very sharply when the distance between the oxygen and hydrogen atoms is below 1.75 Å. For a favorable hydrogen bond pair, the combined energy

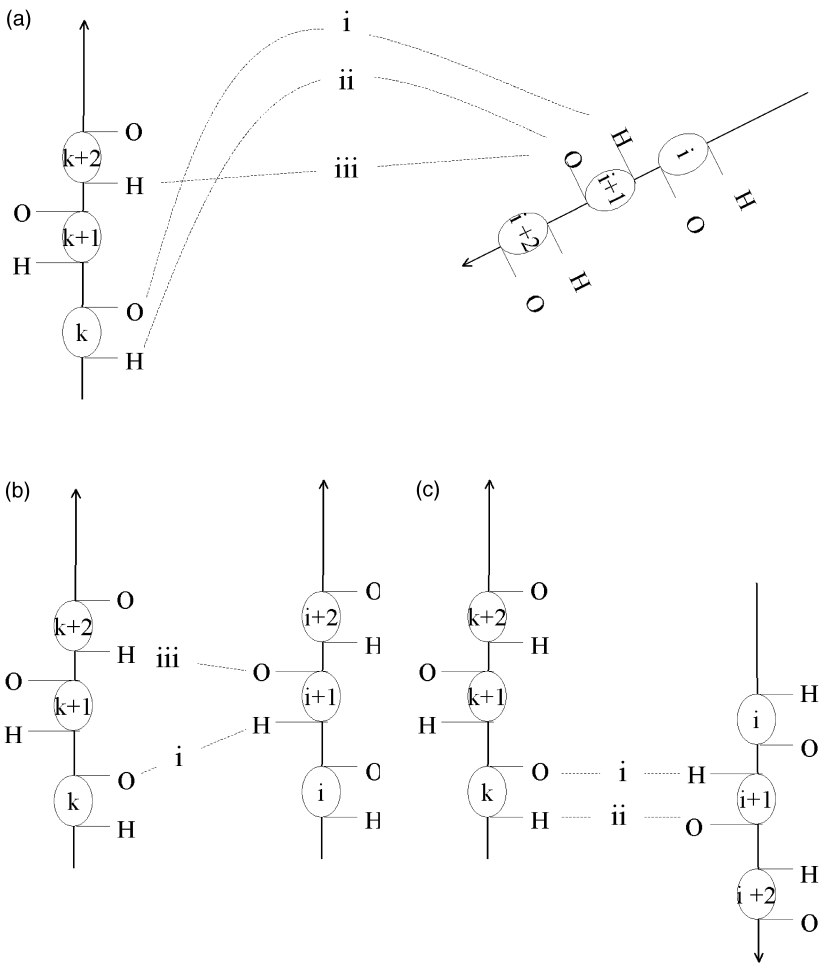

Figure 10. The cooperative hydrogen bonding. Three of the possible hydrogen bonds between two beta-strands are presented. Of the three pairs of these hydrogen bonds (i,ii) is a favorable antiparallel pair, (i,iii) is a favorable parallel pair and (ii,iii) in an unfavorable pair. The lists of favorable and unfavorable pairs are created in the initialization phase of the simulation according to the rules described in the text. (a) At the beginning of the simulation when the conformation is random but rather extended, the distances between the hydrogen and oxygen atoms are large and both the two favorable bonds and the unfavorable bond have only a negligible energy contribution. (b) and (c) Two possible local minima. In each, two of the hydrogen bond distances are close to the ideal, resulting in a stabilizing contribution of one of the favorable pairs. The third hydrogen bond distance is large, causing the remaining two pairs to have only a negligible energy contribution.

term has a minimum when the distances between the hydrogen and oxygen atoms in both hydrogen bonds are 1.85 Å (Figure 11).

Hydrogen bond term for helices

The hydrogen bond term for helices is the same as the one used for sheets, but the definition of the pairs is of course much simpler. In segments that are predefined as helices, hydrogen bonds are considered only between carbonyl oxygen atoms and amide hydrogen atoms that are separated by four residues. Consecutive hydrogen bonds are considered pairs and the term is evaluated as described above.

A logarithmic pair-wise hydrophobic term

Motivation. This term originated from the observation that decoy sets generated by minimization with the ENCAD potential tend to be much less compact than native structures. Indeed, the van der Waals term favors compact structures but its short-range nature results in many high-energy local minima (Figure 12a), which we would have liked to eliminate.

Derivation. High-energy minima can be destabilized by a long-range term enforcing compactness. Long-range terms, however, are typically dominated by the interactions between pairs of atoms that are far away from one another, as these are more numerous than pairs with short distances. Thus, long-range terms are liable to result in structures that are too compact and spherical. To reduce this problem the long-range term should be very "soft", and grow slowly with distance. We tried harmonic and linear terms and decided on a logarithmic term, which still decays most slowly. While this term (described in detail below) gives much better results than its predecessors (Figure 12b) it has at least two problematic features: (1) it is unstable at very short distances; and (2) the compressive "pressure" that it inflicts on the protein depends on the protein size resulting cores that may be too compact for larger proteins. Thus, improving this term is an important direction for further work.

Formal definition. The hydrophobic term, E_{hpb}, is applied to every pair (i,j) of hydrophobic atoms that are not part of the same secondary structure element or in the case of a loop are more than three residues apart:

$$E_{\text{hpb}_{i,j}} = \begin{cases} \text{Cln}\left(\dfrac{D_{i,j}-1}{10}\right), & \text{if } D_{i,j} < 1.1 \text{ Å} \\ \text{Cln}(0.01), & \text{if } D_{i,j} < 1.1 \text{ Å} \end{cases}$$

Where $D_{i,j}$ is the distance between atoms i and j and the constant C is 0.04.

As hydrophobic atoms we consider the following.

(1) C^{α} atoms of the hydrophobic residues (alanine, cysteine, isoleucine, leucine, methionine, phenylalanine, proline, tyrosine, tryptophan and valine) as

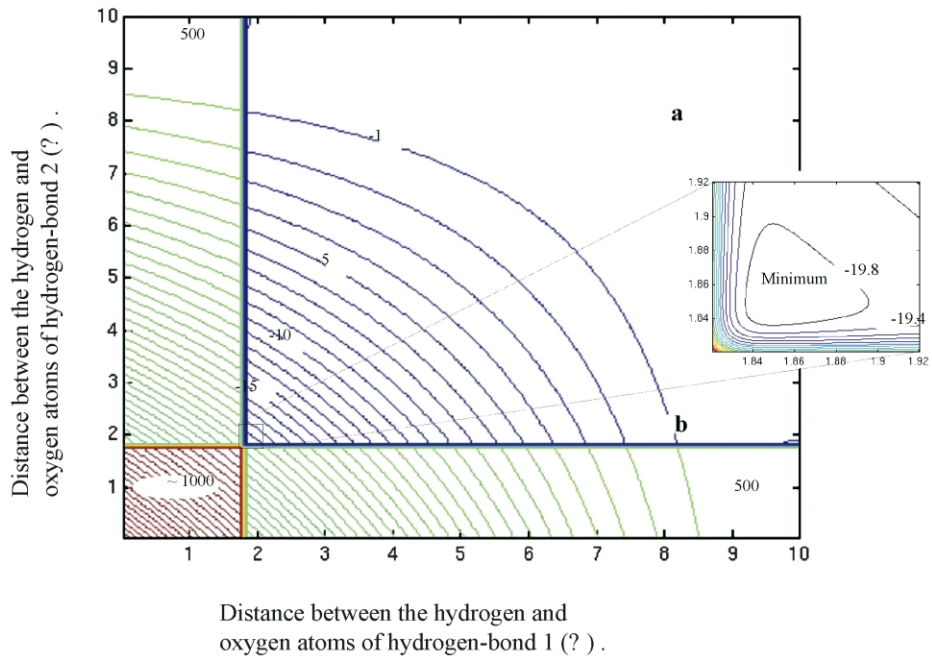

Figure 11. The energy contribution of favorable hydrogen bonding is a four-body function, involving the two atoms in each of the two hydrogen bonds. Most of the space is dominated by the long-range attractive term, with a sharp transition to the repulsive core term when hydrogen-bonded oxygen and hydrogen atoms reach non-physical proximity. A comparison between two pairs of hydrogen bond marked as **a** and **b** in the Figure, reveals the cooperative nature of the energy function. At **a**, the oxygen-to-hydrogen distances in both pairs are large. The energy and the gradient that pull the hydrogen and oxygen atoms towards one another are very small (−0.07 kcal/mol and 0.07 kcal/mol/Å, respectively). This is the case for all hydrogen bond pairs in a typical initial random conformation. Obviously, most pairs remain in this state during the minimization and in the final conformation, since the number of hydrogen bonds that a physical conformation can satisfy is much smaller than the total number of possible hydrogen bonds. At **b** hydrogen bond **2** has already been formed with an O···H separation of 1.8 Å. Both the energy and the driving force to the formation of the other hydrogen bond (**1**) are 16 times larger than in case **a**. If a pair of hydrogen bonds reaches this state, it is rather likely to continue down the steep slope to the minimum where both bonds have proper geometry.

well as C^{α} atoms of non-hydrophobic residues if they are part of a secondary structure element.

(2) Carbon and sulfur side-chain atoms of the hydrophobic residues listed above.

A multi-body hydrophilic term

Motivation. The tendency of charged residues to remain on the surface of native protein structures is mediated by their interactions with the solvent. In a solvent-free system the original ENCAD potential cannot reproduce this affect. As a result a large fraction of the local minima have buried charges (Figure 13). The introduction of the hydrophobic term reduces this problem but does not eliminate it. In structures with buried charged residues the average distance between the hydrophobic residues is larger and the hydrophobic term has high values. These structures may be local minima, which causes a considerable and unnecessary increase in the size of the conformation space. The hydrophilic term is intended to destabilize these local minima and thus reduce the conformation space. We are aware of course that rare, buried charges do play important roles in the structure and/or function of many proteins. Their total exclusion from the current energy function is a "first approximation". In practice buried charged residues can often be deduced from specific conservation patterns and biological data and be treated correctly on a per case basis.

Derivation. We first tried to implement a hydrophilic term by pair-wise repulsion between the charged residues and hydrophobic ones. These terms were either ineffective or resulted in non-compact structures. The attempt to overcome this by increasing the hydrophobic term resulted in an "arms race" that was as unfruitful. To solve this problem we abandoned the two-body terms and derived a cooperative term with an all-or-none behavior. It hardly penalizes a small number of hydrophobic–hydrophilic contacts, but inflicts a high penalty when the number of such contacts increases and the charged residue is buried.

Formal definition. The current hydrophilic term is an interaction of every charged atom i, with all the set of hydrophobic atoms $j_1, j_2, \ldots, j_n$:

$$\mathrm{Ehp}l = C \sum_i \exp(\mathrm{Burial\ of}(i))$$

A contact between a charged atom and a hydrophobic one is represented by a Gaussian term and its buried area is approximated by the sum of these Gaussians terms:[1]

$$\mathrm{Burial\ of}(i) = \sum_j \exp(-\alpha D_{ij}^2)$$

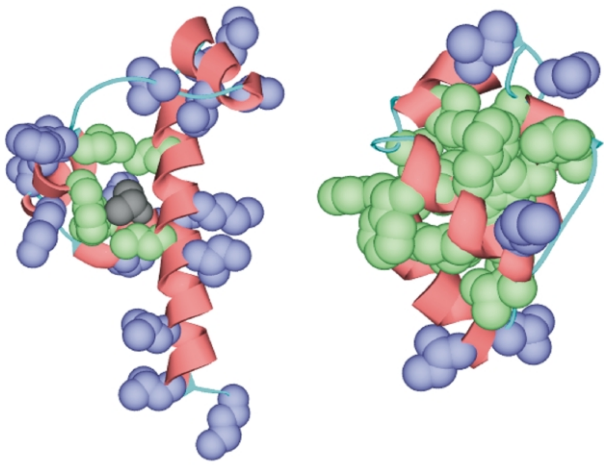

Figure 12. A demonstration of the hydrophobic term. a, The lowest energy structure from 10,000 randomly generated decoys of 2cro using the original ENCAD potential. The structure is mainly stabilized by van der Waals interactions within a core composed of five large hydrophobic side-chains (green) and a threonine residue (gray). The majority of the large hydrophobic side-chains (17 out of 22) are almost fully exposed (blue). A more compact structure would be lower in energy but very unlikely to occur as a result of energy minimization. b, The lowest energy structure from 10,000 randomly generated decoys of 2cro using the current potential with the special hydrophobic term. The structure is much more compact. Most of the large hydrophobic side-chains (17 out of 22) constitute the core (green) and only five are fully exposed (blue).

Figure 13. Burial of charged residues in a solvent-free system. Two arginine residues (blue) and the C-terminal (yellow) are pointing towards the center of this 2cro local minimum (the same as in Figure 12a). This is a rather uncommon feature in native structures. The introduction of the special hydrophilic term used here eliminates structures like this without affecting compactness.

where D_{ij} is the distance between the hydrophilic atoms i and the hydrophobic atom j; C and α were empirically assigned the values 0.0003 kcal/mol and 0.0125/Å^2, respectively. Hydrophobic atoms are defined as described above and charged atoms are the carboxylate oxygen atoms of glutamic and aspartic acid and the side-chain nitrogen atoms of arginine, lysine and histidine.

With this term, the small number of contacts a surface charge makes with hydrophobic residues is hardly penalized. A large number of such contacts, characteristic of a buried charge, are penalized heavily. As a result buried charges are almost absent from our decoy sets without affecting the compactness.

Uncharged polar groups raise a similar but more difficult problem. They are very often buried but almost always part of a network of hydrogen bonds. In our decoy sets buried non-hydrogen-bonding polar groups are, in general, more abundant than in native structures. Penalizing them without preventing the creation of hydrogen bonds could be useful but appears to be non-trivial.

Computational requirements

Simulations were preformed on a loosely coupled cluster of PENTIUM and ALPHA based computers running LINUX. Single minimizations took 20 seconds to five minutes on a single PENTUM II 400 MHz, depending on protein size. The parallelization of the simulations is trivial and efficient with all CPUs running the same code with different values of the random number generator seeds. As no communication is required between the processes, performance grows linearly with the number of CPUs.

Acknowledgements

The authors thank E. Domany for his indispensable support. This work was supported by NIH grant GM-41455 (to M.L.). This work was in great part shaped by numerous stimulating discussions with R. Elber, P. Koehl, R. Samudrala, Y. Xia, B. Fain, J. Tsai, S. Brenner, R. Olender, H. Senderovich and R. Rosenfeld. This work would have not been possible without the availability of free software packages, most notably Linux, the GNU packages, SwissPdb Viewer,[25] POV-ray† and xmgrace‡.

References

1. Levitt, M. & Warshel, A. (1975). Computer simulation of protein folding. *Nature*, **253**, 694–698.
2. Burgess, A. Y. & Scheraga, H. A. (1975). Assessment of some problems associated with prediction of the three-dimensional structure of a protein from its amino-acid sequence. *Proc. Natl Acad. Sci. USA*, **72**, 1221–1225.
3. Osguthorpe, D. J. (1999). Improved *ab initio* predictions

† http://www.povray.org
‡ http://plasma-gate.weizmann.ac.il/Grace

with a simplified, flexible geometry model. *Proteins: Struct. Funct. Genet.*, 186–193.
4. Lee, J., Liwo, A., Ripoll, D. R., Pillardy, J. & Scheraga, H. A. (1999). Calculation of protein conformation by global optimization of a potential energy function. *Proteins: Struct. Funct. Genet.*, 204–208.
5. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, **220**, 671–680.
6. Pedersen, J. T. & Moult, J. (1997). Protein folding simulations with genetic algorithms and a detailed molecular description. *J. Mol. Biol.* **269**, 240–259.
7. Piela, L., Kostrowicki, J. & Scheraga, H. A. (1989). The multiple minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method. *J. Phys. Chem.* **93**, 3339–3346.
8. Park, B. & Levitt, M. (1996). Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* **258**, 367–392.
9. Hinds, D. A. & Levitt, M. (1994). Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* **243**, 668–682.
10. Park, B. H. & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **249**, 493–507.
11. Fain, B. & Levitt, M. (2001). A novel method for sampling alpha-helical protein backbones. *J. Mol. Biol.* **305**, 191–201.
12. Ozkan, B. & Bahar, I. (1998). Recognition of native structure from complete enumeration of low-resolution models with constraints. *Proteins: Struct. Funct. Genet.* **32**, 211–222.
13. Bryant, S. H. & Altschul, S. F. (1995). Statistics of sequence–structure threading. *Curr. Opin. Struct. Biol.* **5**, 236–244.
14. Simons, K. T., Strauss, C. & Baker, D. (2001). Prospects for *ab initio* protein structural genomics. *J. Mol. Biol.* **306**, 1191–1199.
15. Betancourt, M. R. & Skolnick, J. (2001). Finding the needle in a haystack: educing native folds from ambiguous *ab initio* protein structure predictions. *J. Comput. Chem.* **22**, 339–353.
16. Vendruscolo, M. & Domany, E. (1998). Efficient dynamics in the space of contact maps. *Fold. Des.* **3**, 329–336.
17. Fletcher, R. (1970). New approach to variable metric algorithms. *Comput. J.* **13**, 317–322.
18. Levitt, M. (1983). Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.* **170**, 723–764.
19. Vendruscolo, M. & Domany, E. (1998). Pairwise contact potentials are unsuitable for protein folding1. *J. Chem. Phys.* **109**, 11101–11108.
20. Keasar, C. & Rosenfeld, R. (1998). Empirical modifications to the Amber/OPLS potential for predicting the solution conformations of cyclic peptides by vacuum calculations. *Fold. Des.* **3**, 379–388.
21. Xia, Y., Huang, E. S., Levitt, M. & Samudrala, R. (2000). *Ab initio* construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.* **300**, 171–185.
22. Ortiz, A. R., Kolinski, A., Rotkiewicz, P. & Skolnick, J. (1999). *Ab initio* folding of proteins using restraints derived from evolutionary information. *Proteins: Struct. Funct. Genet.*, 177–185.
23. Bonneau, R., Strauss, C. E. M. & Baker, D. (2001). Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins: Struct. Funct. Genet.* **43**, 1–11.
24. Samudrala, R. & Molt, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275**, 895–916.
25. Guex, N. & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.

Edited by J. Thornton

(Received 26 October 2002; received in revised form 3 March 2003; accepted 6 March 2003)