A Novel Method for Sampling Alpha-Helical Protein Backbones

Boris Fain and Michael Levitt

Department of Structural Biology, Stanford University School of Medicine,

Stanford, CA 94305

Abstract

We present a novel technique of sampling the configurations of helical proteins. Assuming knowledge of native secondary structure, we employ assembly rules gathered from a database of existing structures to enumerate the geometrically possible 3-D arrangements of the constutuent helices. We produce a library of possible folds for 25 helical protein cores. In each case ur method finds significant numbers of conformations close to the native structure. In addition we assign coordinates to all atoms for 4 of the 25 proteins. In the context of database driven exaustive enumeration our method performs extremely well, yielding significant percentages of structures (0.02% - 82%) within 6A of the native structure. The method's speed and efficiency make it a valuable contribution towards the goal of predicting protein structure.

I. INTRODUCTION

A. Outline of current prediction methodology

Prediction of protein structure from sequence is one of the most enticing goals of scientific inquiry today. Currently the most reliable method of determining a protein's shape is to search for a close homolog in the database of solved protein structures. Despite the fact that the number of solved structures increases daily, it is estimated that in the near future

at least 40% *cite someone...* of proteins of interest bear no discernible sequence resemblance to a known macromolecule. Therefore *ab initio* prediction of structure from sequence remains an important challenge.

When sequence homology cannot be used to construct a 3-D model, the current *modus operandi* for predicting structure is composed of three separate (yet interdependent) steps: sampling, searching, and ranking. One first picks a formalism to index and sample the possible shapes, then searches through the shapes, and finally uses some ranking criterion to pick out a structure as close to native as possible. To clarify this classification let us provide some examples of each step.

sprinkle some sitations here. very delicate situation, since we basically can site the whole field in the next three paragraphs.

The first requirement in sampling structures is that of representation, e.g. all-atom, reduced 'pseudo' amino-acid interaction centers, lattice, etc. Once the representation is chosen, one then picks the degrees of freedom to use. Researchers vary Cartesian coordinates of all or a subset of the atoms, dihedral angles, relative distances, and so on. The decisions of which, how many, discretized or continuous, and so on, are motivated by the amount of detail one wants to capture, as well as by computational complexity of the forthcoming search. One notable example, similar in spirit to this work, is ref. [6] in which the authors assemble helical sub-units using consensus distance geometry.

Having picked a representation, one then selects an appropriate search technique, or a mixture of search techniques to march through the enormous space of possible 3-dimensional structures. As we've mentioned above, the three complementary steps of prediction are highly interdependent. The most important factors in picking a search method are computational complexity and the nature of the energy function and of the representation. Lattice models, for instance, lend themselves easily to exhaustive enumeration. All atom cartesian coordinate representations usually employ a physical potential, and search for a minimum with either an intelligent minimization routine, or a copy of Nature's algorithm, molecular

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

dynamics. [1]

The remaining tool employed in protein structure prediction is a potential function constructed to pick out native or near-native conformations from a vast number of alternatives. If the search method employed is a guided one (for example, minimization) then the potential function also provides a landscape that will allow the procedure to converge to the correct answer. Just like sampling and search techniques, potential functions are many and varied: physical potentials with/without water representation, knowledge based potentials of various detail, hydrophobic contact potentials, and many others. Frequently more than one potential is used to evaluate candidate structures.

B. Desirable features of search methods

In this article we present a novel sampling technique for helical proteins. We shall review the features that make a sampling method effective thus providing a context in which one can judge the merits and drawbacks of our procedure effectively.

A search method should have several desirable properties to be useful in *ab-initio* structure prediction. Our assertion is that the proportion of sampled conformations that are native-like should be statistically significant. Admittedly, the definitions of both "statistically significant" and "native-like" are imprecise - but clearly, the more and the closer, the better.

Let us first explain the need to get close to the native structure. An energy functions can only be effective in a slice of conformational space close to the actual protein . Simply stated, structures can be 'wrong' in a vast number of ways, many of which are low scoring alternatives to the native fold in a very large space. Only a small subspace close to the actual structure can be expected to exhibit the properties of an energy funnel; i.e. a region where the energy decreases uniformly as one approaches the native, and where a majority

---

[1] Computational complexity is the quintessential burden of protein structure prediction.

of conformations are more favorable than all the others in the available space.

How close is close enough? To make the definition more quantitative, one can compare one's sampling techique to random sampling. Reva *et. al.* [11] suggest a 'native-like' target value of about 6 Å CRMS deviation relative to the actual structure. Other authors [10,12] also propose a similar distance of 4-6 Å. [2]

Thus a selection method cannot predict if it is not given a chance, namely at least one member of the sample is 'native-like'. The requirement that a 'statistically significant' fraction of the sample is native-like comes from a different obstacle. The need to enrich the sampled structures with near-native conformations arises mainly from the inadequacies of the energy functions we currently use. A perfect energy function should be able to tell apart a native structure from *all* other alternatives - this, in fact, is what nature is able to do with striking consistency. [3] Today's best scoring functions need all the help they can get. The current generation of potentials *site many people, ram, others* often mis-identifies incorrect folds as correct predictions. By increasing the number of 'correct' answers we increase the statistical likelyhood of making a successful prediction. In addition some current techniques use consesus information - selecting a subset and building a consesus model from it [6]. These methods clearly need a sizeable native-like population to work. And, all the above non-withstanding, having more, rather than less, native-like structures can't possibly be bad!

---

[2]Clearly, this number depends on the size of the protein.

[3]Nature, most likely, does not sample every possible structure, but only those lying within thermal fluctuations' reach from the proteins' preferred folding pathway - still a very large and diverse ensemble.

II. DESCRIPTION OF THE SAMPLING METHOD

A. Graph-Theoretical representation of secondary structure

We represent alpha-helical proteins as connected graphs. Each helix is represented by a vertex; an edge is drawn between two vertices if the corresponding helices are in contact. To reproduce a protein core one builds a connected subgraph, adding one helix at a time, until the protein is assembled. Figure 1 shows the graph-theoretical representation and the construction of myoglobin.

We are, in essence, building an off-lattice model for secondary structure segments, but our degrees of freedom are not in the frequently chosen [10,12,3] (ϕ, ψ) loop residues (or their subsets), but in the relative geometric position of the helices themselves. The disadvantage of this approach is that a significant fraction of visited structures violate loop constraints - simply put, the ends of helices are too far apart to be joined together by an intervening loop. There are, however, several advantages. First, the structures we generate have a much better tendency to be compact. Second, we are able to exploit the correlation between the sequence patterns of helix-helix contact to significantly enrich our sampling with native-like structures. (A simple example of this is two helices with small residues in the contact area are more likely to be close to each other than those with large residues between them.) Third, we are able to sample possible packing much more finely than when using loop torsion angles, as is usually done in off-lattice models.

B. Exhaustive enumeration of helix-helix contacts

In this work our principal aim is to achieve coarse-grained sampling of the helical protein core. The way we choose to travel through our space is by exhaustive enumeration of a discrete version of our representation. We can choose to sample helix-helix packing ('link')

in several ways. [4] To clarify this we, once again, draw an analogy to dihedral sampling: there one can discretize a particular dihedral angle into a (not necessarily) uniform spectrum of values suggested by geometric considerations; e.g. sampling a dihedral angle in, say, 30 deg. intervals. Alternatively, one can extract the local moves from a database of existing structures, thus sampling the space more efficiently. One could, for example, assign three possible values - helix, loop, sheet - to a dihedral angle. *site someone here* Another possibility would be to bias the assigment with sequence matching. [13] or secondary structure prediction. We choose to derive the preferences for helix-helix packing from the distribution of such orientations in known proteins. Our hypothesis is that the contact patches on each helix influence the way the helices pack. If we are incorrect, then our packing strategy will position helices randomly - no harm done.

C. Definition of helix to helix contact ('link') and the contacts' database

We extract the possible relative orientations of two helices from a database of touching helix-helix pairs obtained from a subset of the SCOP [9] database. The 1305 folds in our library have a sequence identity to each other no greater than 35%. We then parse our database to get touching helix-helix pairs. We define two helices X and Y to be in contact if a) the shortest distance d_{min} between any two CB's, CB_x and CB_y, located respectively on helix X and helix Y, is less than 7.3 Å; and b) each helix has at least 3 CB atoms within $d = d_{min} + 2.5$ Å. [5] [6] Additionally the CB's on both helices that are within 2.5 Å of

---

[4] We call a specific realization of an edge a 'link'.

[5] The three CB on each helix are necessary to define a relative rotation from one helix to another.

[6] The specific values of $d_{min} = 7.3$Å and 2.5Å worked best in our tests. Lowering the d_{min} parameter to below 5Å caused some of the proteins in our database to be represented by disconnected graphs, thus making it impossible to ever reproduce them with our technique. Conversely, making the d_min cutoff much larger than $7.3\AA$ decreases the influence the contact sequence has on the

$d_{min} = 7.3$ are defined to be in a contact 'patch'. Our definition picks out the residues in the contact region by assigning a 'patch' of contact residues between helices. Residues on the far side of each helix, which have very little influence on the relative orientation, are ignored. The 'patch' concept is illustrated in Figure 2 does the job well by showing the patches and the structure of each link in the database.

D. The Enumeration Procedure

Once again, figure 1 illustrates the method. The build-up of a helical protein proceeds as follows: Given a protein sequence and native secondary structure assignment we construct idealized helices on the chain. (In a predictive scenario the native secondary structure assignment will be replaced by a prediction, or possibly several alternative predictions.) We then go to the library of links and perform a sequence alignment between the patches on each link and all possible pairs of helices on our target sequence. The residues between key 'patch' residues serve as spacers to fix the position of the influencial patch residues. If the sequence match is high then this particular link will be used to bring the pair of helices together. The actual threshold for matching depends on the number of structures we want to sample. The sequence match is scored using the Blosum62 *cite Blosum people* matrix. The comparison threshold typically varies from 0.1 to 0.6 and is adjusted to give anywhere from 10 to 1000 possible links for a given pair of helices in the target protein. The number chosen depends on how many final structures we wish to generate in the available amount of computer time. The relative orientations of the helices are then loaded into memory and the buildup of structures begins.

We use each topologicaly distinct pathway to build the target - (refer to figure 1 b)) - each specific pathway corresponds to a minimal subgraph spanning the protein graph. (Dashed and solid lines, respectively, in figure 1.) As seen in Fig. 3 for a structure of 4 helices this

---

relative orientation of the helices.

would result in $4^{(4-2)} = 16$ possible topologies. Note that the graph corresponding to each final structure might (indeed should) posess other edges not used in construction; however it's sufficient to follow a minimally connected subgraph to construct it.

For each topology we construct each possible combination of links (gathered from the sequence-matching procedure above) that can realize a given edge. To improve performance we used branch and cut filters for loop and clash constraints. The clash filter eliminates a conformation if more than 3 residues on one helix are closer than 2 $\mathring{A}$ to residues on another helix. The loop filter eliminates conformations for which the distance for the loops necessary to connect the helices is longer than the maximum available loop length. The reason for the branch-cut approach is simple: if in a given six-helix enumeration helix one clashes with helix three there is no need to cycle through and build helices four, five, and six. Finally, each geometrically viable structure is tested for compactness. All the tests and filters are extremely fast because whenever possible we use the coarse segment representation of the structure and thus escape having to visit each amino acid's coordinates. We have also incorporated other filtering information, most notably disulfide bond locations, into the build-up procedure. In it's current incarnation the method is able to sample roughly 10^3 conformations per second on a 400MHz Pentium workstation. The ultimate speed of the procedure will, in the future, be limited by scoring function evaluations.

III. RESULTS

In this paper we present the results of several applications of the sampling method.

To test the performance of our technique we have used the coordinates of helical cores for 25 proteins. The molecules' sizes range from 31 to 172 amino acids, and the number of helical residues we've assigned coordinates to ranged from 23 to 130.

The position of disulfide bonds is easily obtained through chemical methods, and one can often rely on knowing their location prior to prediction of structure. To this end, for 4 of the 25 proteins in our set that have disulfide links, we have also evaluated the performance

of our method with and without *a priori* knowledge of these bonds.

The fact that we assign residues only to the helical backbone leaves open the possibility that our results will be degraded by subsequent assignment of coordinates to all remaining atoms. We investigated this effect by building all atomic coordinates for 5 of our proteins using the program SegMod [8]. Prior to reconstructing full coordinates we scored the helical cores with a slightly modified Sippl-like function [14] leaving 500 to 4500 lowest energy structures for each protein. (We left more decoys for larger proteins).

A. Sampling of helical cores

We used the method to generate conformations for 25 helical proteins, ranging in size from 31 to 172 residues and containing anywhere from 2 to 6 helices. To check how much the quality of our sampling will be degraded when all the coordinates are reconstructed from the helical core, we have included several proteins with significant loop content. The assigments of helices to the structure were made identital to the native structure. The program STRIDE [4] was used to determine the secondary structure of the native protein.

The results are summarized in table I. Overall the results look very good. For nearly every protein a sizeable proportion of the sampled structures is within 3Å of the native. For 9 proteins the best structure produced is closer than 1Å CA RMSD to the actual structure, and is virtually identical to the protein itself. Our procedure clearly gives a suitable potential the opportunity to make a successful prediction.

Our method is not only very fast, it is also very efficient, in the sense of being native-like rich. We can access the efficiency by estimating how many random protein-like compact structures are needed to obtain the RMSD of our best structure, and compare that number to the number of conformations we have visited. We use the estimate from [11]

$$\frac{1}{\left(\sigma\sqrt{2\pi}\right)}\int_{-\infty}^{R}\exp\left[\frac{-\left(x-<R>\right)^2}{2\sigma^2}\right]dx \tag{3.1}$$

Where, following Reva *et al* we set $\sigma = 2.0$ and $< R >= 3.333N^{1/3}$, where N is the number of

residues in the protein core that we have assigned coordinates to. The proximity of our best structures to native exposes the weakness in estimating probability of low RMSD's, namely that the Gaussian distribution cannot be used to effectively describe sets of conformations which are very close to native. Having said that, we still feel that equation 3.1 gives a good estimate of efficiency. The last column in table I shows the ratio of the number of structures we sample to the number of random structures given by formula 3.1.

As the last column in table I shows, our method is efficient - the sampling is strongly biased towards the native conformation. The worst performance is shown in protein 1res, where we would have done slighly better if we chose relative helical conformations at random. However for most cases we are beating the random selection by factors of 10^4 to 10^6, with efficiency actually increasing with the size of the proteins.

B. Features of the sampled conformations

Figure 4 shows a typical distribution of the sampled conformations. One particularly interesting feature is the non-gaussian tail extending towards the native structure. This is most likely the result of using sequence matching to select the possible orientations of the helix-helix pairs. The use of sequence information enriches the set with native-like structures and makes the curve decidedly non-gaussiian.

C. Using disulfide bond information

To test how well the sampling is aided by supplementary information, we took proteins with disulfide bonds and subjected the sampling for those structures to an additional filter. [7] Our filters rejected conformations in which the CYS CB-CB distance was greater than 8Å.

---

[7] We did not screen 1cc5 because although it does have disulfide bonds, they occur in the loop regions and do not help filter the helical cores.

This is not a very strict filter; it simply weeds out grossly incorrect topologies. The results both with and without [8] disulfide bond information are displayed in table II. The number of structures was cut by approximately a factor of 8. The filtering did not significantly improve the average RMSD of the emsemble; however the proportion of 'good' structures was raised approximately fivefold.

D. Completing the structure

The structures described in the previous two sections posess only backbone coordinates for residues in the helical core. It is quite possible that the successes of the sampling can be washed out when coordinates are assigned to the remaining residues and sidechain atoms. To check how the quality of the sampling changes during completion we ran our sets of conformations through the program SegMod. [8] We did not construct full coordinates for all twenty five proteins because we simply wanted to show that reconstruction does not significantly degrade the quality of the sampling procedure. To make the test credible, we selected proteins of varying size, and with long unassigned loops. The results are shown in table III.

Once again, the resulting ensembles of decoys contain a sizeable proportion of near-native decoys. Even though we did not produce any decoys closer than 6$\AA$ for 2fha, we feel that the sampling of such a large (according to *ab initio* prediction standards) protein - 172 residues - was successful. The RMSD of the best decoy, 7.3$\AA$ has log-odds of 23234 (computed from last column of table III. The total ensemble for 2fha has 0.25% of decoys below 10.0$\AA$ rmsd. Both of these results are good for decoy ensembles. Whether they are good enough for prediction remains to be seen.

---

[8]The results without disulfide bond information are the same as in table I

IV. DISCUSSION

A. Comparison with other search methods

Our sampling method is one of the main points of this paper. It has both merits and drawbacks when compared with other search techniques. One alternate approach to sampling is to assemble helices using distance geometry by either embedding distance space in 3-D [1,5], or by minimizing against restraints [6,2,15–18]. In a recent work [6] this method produced decoy libraries with the log-odds of producing the lowest structure in the -4 to -6 range. [9] Our results, which range from log-odds of -4 to -9, are somewhat better. However we must emphasize that the comparison is unfair, since we used actual secondary structure, and Huang et. al. used predicted (albeit well predicted) secondary structure. We are currently examining how well our method will perform with predicted secondary structure. Preliminary indications are that both distance geometry and our method both get sufficiently close to the native fold, with our method having a slight edge due to its speed.

Our method begins to pull away from distance geometry methods in the scaling of CPU time needed when one moves to larger and larger proteins. Currently distance geometry methods cannot be extended to proteins of length over 100 residues [6]. In contrast, because our sampling technique scales with the number of secondary structure segments and not the number of residues, it can easily handle chains of 100 - 200 residues, comparable to single domains of larger proteins. The library for the largest protein in our work - 2fha, 172 residues long - took approximately 15 hours of CPU time on a 400 MHz Pentium II machine. The sampling procedure produces approximately 500 structures per second for two-helical proteins, 100/sec for proteins containing three helices, 10/sec with four, and roughly one

---

[9] Log-odds are $1/N_{best}$, where N_{best} is the number of random structures needed to find at least one structure with RMSD equal to our best structure. N_{best} is the first number in the last column of table I.

structure/second for proteins with five and six helices.

Another sampling techique is to fix the helical segments and then to vary the dihedral angles of the loop regions. This method has been used to produce fold libraries [10], and, combined with a branch and bound algorithm, in folding studies [19,3]. This method carries an advantage over ours because every structure generated automatically satisfies loop constraints, while many of our potential structures do not. Many of our structures have to be rejected because the physical distance between ends of helices violates chain connectivity. Furthermore, we *do not* have loops at the end of construction; therefore we are currently working on an algorithm to sample loop regions of our protein cores that would be a good (i.e. fast) match to our helical core sampling.

On the other hand our method outpaces the loop dihedral angle sampling in two categories: scaling of computational demands with time, and efficient sampling of structures with plausable contacts. Let us illustrage the first difference with a four-helical protein. Assuming an average loop length of seven residues varying loop conformation demands 7(residues)*2(ϕ, ψ)*3 = 42 degrees of freedom. Our procedure needs to position three rigid bodies (the position of the initial helix is arbitrary), requiring 6(rigid body)*3 = 18 degrees of freedom. The second difference stems from the fact each helix in our construction method is guaranteed to have at least one, and possibly two plausible contacts with other helices. When one varies dihedral angles, most of the non-clashing structures have large voids between helices.

Yet another method for generating possible reduced-model conformations for protein structure are lattice and off-lattice models, such as [21,20,22]. Yet again these methods outshine ours because they require no knowledge of either actual or predicted secondary structure. On the other hand it's difficult to see secondary structure at the resolution of these lattice models. In computational performance, for small proteins these approaches compare very favorably with ours - a simplified representation which assigns a lattice point to every second residue can exhaustively sample shapes of proteins of up to 100 residues [20]. However, because of an exponential increase in the number of shapes of a self-avoiding

walk on a lattice, it's difficult at this point to see a generalization of the lattice methods that would apply to larger molecules without a significant sacrifice in resolution.

One final method we want to mention is that of assembling structures from fragments of existing folds [13,12]. We like this method very much. It uses the information in the sequence to bias the assembly much the same way we use the 'patch' information to bias helix-helix orientation. It is reported to produce best structures as good as the lowest RMSD's in our set. In addition, this method - as do the lattice models and the loop angle search - generalizes trivially to beta and mixed alpha/beta proteins: something we cannot yet do. Because we have not had the opportunity to test this method ourselves, we cannot comment on how its efficiency compares with ours. Our feeling is that sequence information in 'contact patches' has more correlation on the geometry of the structure than sequence information located sequentially along the chain. Having said that, we still consider fragment assembly the current method to beat.

B. Advantages and disadvantages

In the beginning of this paper we introduced two desirable qualities of a sampling method: speed, and the ability to produce structures close to the native fold. Our technique fills both of these requirements.

The speed of the method depends on the size of the protein and, more specifically, on the number of helices that we are trying to arrange. Each entry in table I takes from a fraction of a second to under an hour to produce. A naive count would estimate a factorial growth in the number of graphs, and an additional geometric growth in the number of structures sampled in each graph; however the actual increase is much less. For most structures of 4 helices or more, cutting the branches that violate self-avoidance results in a significant reduction of the number of conformations we need to sample. In our tests each additional helix increased the time of the runs by approximately a factor of 30. The set of conformations for our largest structure, 1fha, took approximately 4 hours to generate.

In addition to being fast, our method scales well - the use of branch/cut helps because larger proteins are very constrained by self-intersection. Because of this our technique is able to sample molecules comparable in size to small domains.

We chose to re-arrange helices using a 'patch' contact database derived from existing structures. This approach significantly enriches the sampled ensemble with native-like structures. In addition, the graph-theoretical enumeration of relative orientations ensures that we sample all plausible regions of conformation space.

In its current incarnation our method also possesses some drawbacks. The main disadvantage of our technique is the requirement of a specified secondary structure. It's highly desirable for a search technique to also sample alternate secondary structure assignments. Currently we can only do this by specifying different assigments at the beginning of the procedure. Another significant drawback is the absense of a complimentary loop-building method. We need a method with the speed of a loop-library lookup methods, *cite library-loop-building* yet able to get a 3 Å or better approximation of the native loop (so that our near-native cores remain near-native). A final deficiency of our approach is the lack of an obvious generalization to beta and alpha/beta proteins. The definitions of a sub-segment and the contact patch have to be significantly revised to adapt to beta sheets.

V. FUTURE DIRECTIONS

The deficiencies of our technique, outlined in the previous paragraph, point the way to future developments. We are currently working on a fast loop building procedure for short (3-10 residues) loops. Our next project is to enhance sampling of alpha proteins to include variations of the boundaries of the helices.

A slightly more distant goal is the extension of the presented technique to construct β-sheet and mixed α/β proteins. In addition we are working on a fast preliminary discrimination function which would be used prior to the reconstruction of all atoms for each conformation.

VI. ACKNOWLEDGEMENTS

Many thanks to the present and former members of the Levitt group for help and discussions. We wish to thank Patrice Koehl for providing excellent routines for superimposing structures. Thanks to Prof. Joseph Rudnick for hints on graph theory. B.F. wishes to thank the A.P. Sloan Foundation, and the U.S. Department of Energy for financial support.

TABLES

TABLE I. Summary of helical core sampling.

Name	Helices	Assigned / Total residues	CA RMSD range best - worst : ave	% below 3Å	% below 6Å	Random / Generated
T0065	2	23/31	0.483 - 7.503 : 4.988	3.1%	81.4%	291299/2604
1fc2 C	2	23/43	0.759 - 7.937 : 4.813	10.3%	77.4%	153634/2144
1res	3	24/43	2.265 - 9.646 : 6.546	0.055%	29.3%	8392/9028
1erp	3	26/38	0.986 - 9.052 : 6.191	0.26%	41.8%	226495/11188
1mbh	3	30/52	0.954 - 10.412 : 7.123	0.3%	23.3%	773292/11724
1uxd	3	30/59	0.880 - 10.828 : 6.773	0.58%	28.2%	927522/16281
3hdd	3	40/56	0.240 - 12.022 : 8.001	4.0%	11.8%	8.3E7/13380
1trl	3	42/62	0.617 - 12.920 : 8.002	0.24%	12.4%	4.0E7/24151
T0073	2	43/48	0.662 - 8.348 : 4.840	11.9%	78.9%	5.5E7/522
1cc5	4	43/83	3.484 - 13.237 : 9.241	-	0.84%	47664/46725
1r69	5	44/63	3.127 - 13.143 : 9.140	-	1.1%	128078/56520
1lfb	3	44/77	1.584 - 13.668 : 8.973	0.09%	5.4%	5.68E6/41505
2ezh	4	48/65	3.514 - 15.363 : 9.670	-	1.4%	5E6/61981
1c5a	4	49/65	3.774 - 13.731 : 9.064	-	1.7%	78745/98842
1hsn	4	50/79	3.321 - 17.380 : 11.211	-	0.28%	266620/32734
1ropa	2	51/56	2.432 - 13.751 : 6.765	1.3%	40.2%	2.9E6/620
1pou	4	52/71	4.546 - 14.318 : 10.187	-	0.95%	25296/29283
1nre	3	57/81	2.240 - 13.020 : 8.215	0.21%	11.3%	1.6E7/21716
1ail	3	60/70	3.028 - 15.481 : 9.582	-	2.5%	3.6E6/15980
1nkl	4	60/78	3.879 - 14.122 : 9.931	-	1.15%	439759/15486
1aca	4	60/86	0.752 - 15.059 : 10.442	0.13%	1.9%	2.5E9/6827
1flx	4	67/79	3.306 - 14.786 : 10.869	-	0.635%	6.5E6/37141
1aj3	3	88/98	2.594 - 16.309 : 10.125	0.05%	3.1%	1.6E9/8177

1lis	5	91/131	4.345 - 18.696 : 12.430	-	0.07%	2.0E7/32805
2fha	5	130/172	4.855 - 21.671 : 15.912	-	0.023%	1.0E10/4279

Table I summarizes the sampling of helical cores for 25 proteins. Column 3 shows the number of residues that are assigned coordinates as well as the total length of the protein. Column 4 lists the range of CA RMSD of assigned coordinates from native coordinates. Columns 5 and 6 show the percentage of total structures that are closer than, respectively, $3\mathring{A}$ and $6\mathring{A}$ to the native structure. The last column displays the efficiency of our decoy generation method. It shows the ratio of the number of structures one needs to generate randomly in order to produce the best CA RMSD (computed using Eqn. 3.1) to the number of conformations in our ensemble.

TABLE II. Disulfide bond information included

Protein	ave without → with	%below 3A	%below 6A	num/numact
1c5a	9.064 → 8.047	-	1.7% → 7.4%	98842 → 6325
1erp	6.191 → 6.017	0.26% → 1.3%	41.8% → 46.9%	11188 → 761
1nkl	9.931 → 9.254	-	1.15% → 5.8%	15486 → 889

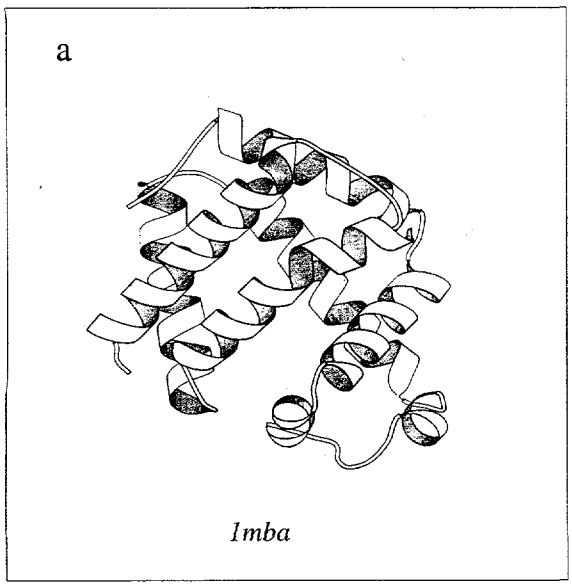
Table II displays the results of pruning the decoy ensembles with disulfide bond information. The columns show shifts in the average Ca RMSD, proportions with RMSD below $3\mathring{A}$ and $6\mathring{A}$, and the reduction in the number of structures in each set.

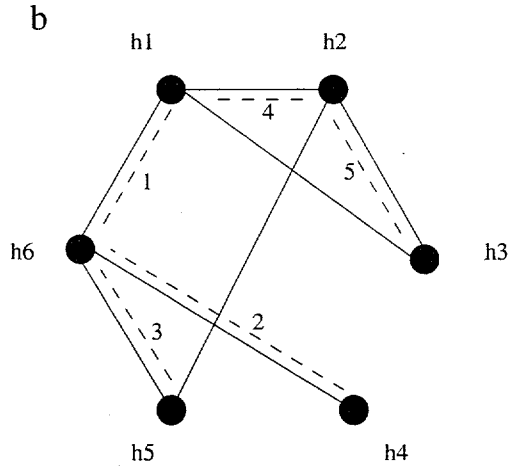
TABLE III. RMSD's for full structures

Name	Size, Helices	CA RMSD range best - worst : ave	% below 4Å	% below 6Å	Random/ Generated
T0073	48, 2	1.435 - 8.320 : 4.840	24.5%	73.7%	2.14E7/522
1ropa	56, 2	2.446 - 15.041 : 7.074	6.6%	34.0%	7.79E6/712
1ail	70, 3	3.200 - 15.624 : 9.730 •	0.26	3.50%	1.45E7/1917
1flx	79, 4	3.282 - 14.886 : 11.329	0.06	0.9%	5.56E7/3342
1aj3	98, 3	2.67 - 16.02 : 10.396	1.0%	4.0%	9.17E9/1880
2fha	172, 5	6.644 - 26.323 : 18.668	-	0.25 %$\leq 10\mathring{A}$	7.27E8/4279

Table III summarizes the results for decoy sets with all atoms' coordinates assigned. The columns are similar to table I. The number of decoys in each ensemble is smaller than in table I because of pruning with a statistical pairwise potential. 2fha had no conformations below $6\mathring{A}$, however 0.25 % of structures were closer than $10\mathring{A}$ to the native.

FIGURES





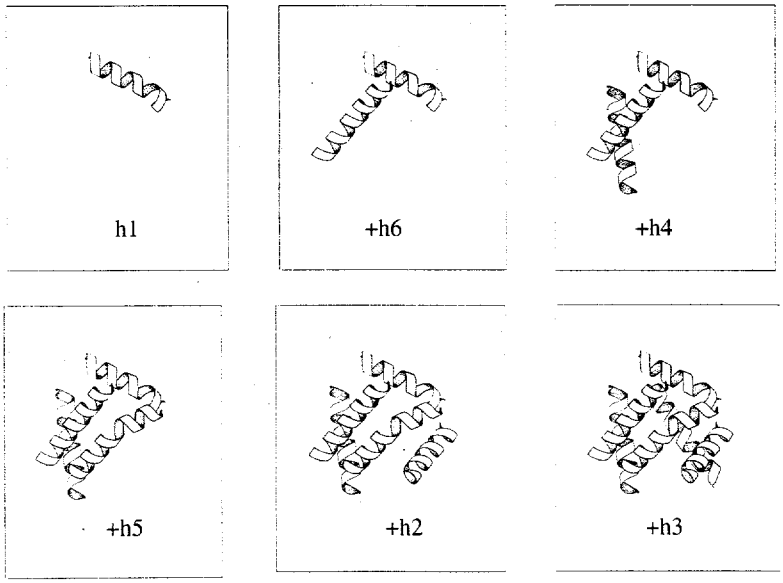


FIG. 1. This figure shows the graph representation of the protein Myoglobin, as well as its construction by our method. a) 1mba. This well known globin consists of 6 major and two minor helices. We sample the 6 major ones. b) The vertices of the graph are helices 1 through 6 of 1mba. (helix1: residues 4-19; h2: 21-35; h3: 59-77; h4: 81-98; h5: 102-119; h6: 126-144.) The vertices represent the 6 major helices; the solid lines are drawn whenever two helices are, according to our definition, in contact. The minimal spanning subgraph, outlined in dashed lines, represents one possible way to reassemble the helices. The dashed lines are numbered in the order the protein is assembled in one of our conformations. c) The order of assembly of the helical core. Helices are assembled in the order 1, 6, 4, 5, 2, 3 using the relative orientation from the dashed lines of b). Figures were made with the aid of MOLSCRIPT [7]. In the process of construction other 'links' may form, however they are not neccesary to reproduce the helical core.

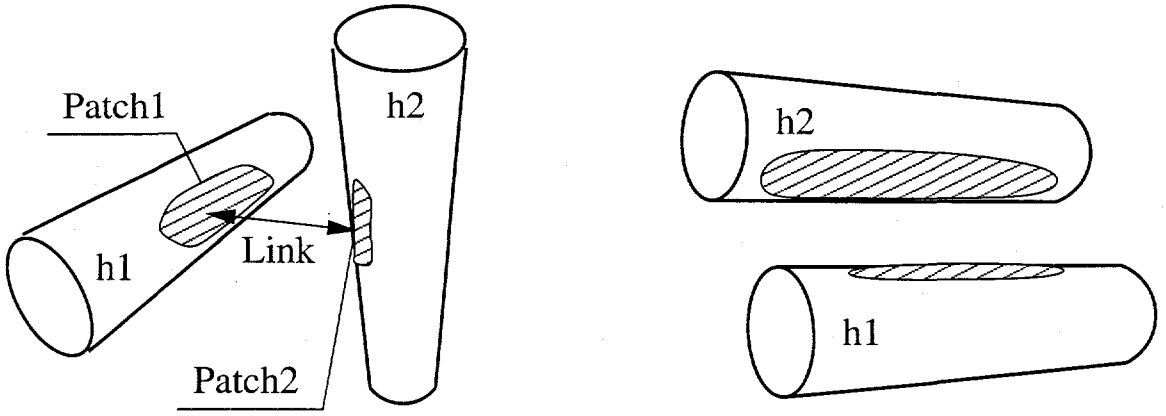


FIG. 2. A diagram of two typical link configurations. The amino-acids in the shaded areas are involved in sequence matching against the target. The configuration on the left produces shorter patches than the configuration on the right.

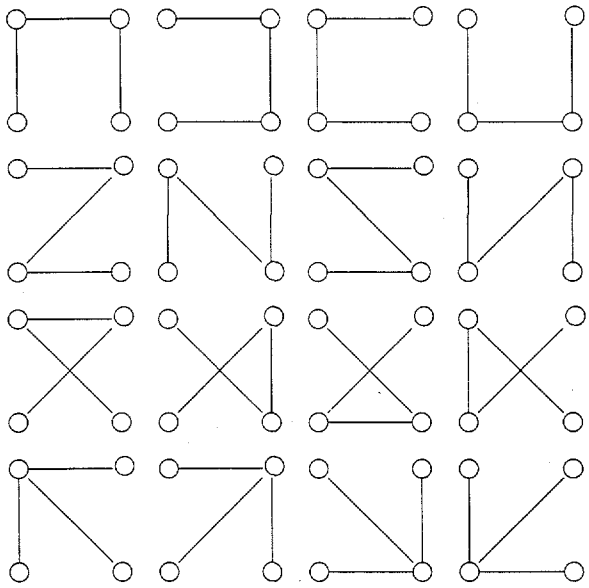


FIG. 3. This figure shows the 16 topologically distinct ways to construct a 4-helix protein. In general there are $N^{(N-2)}$ spanning trees for a N vertex graph.

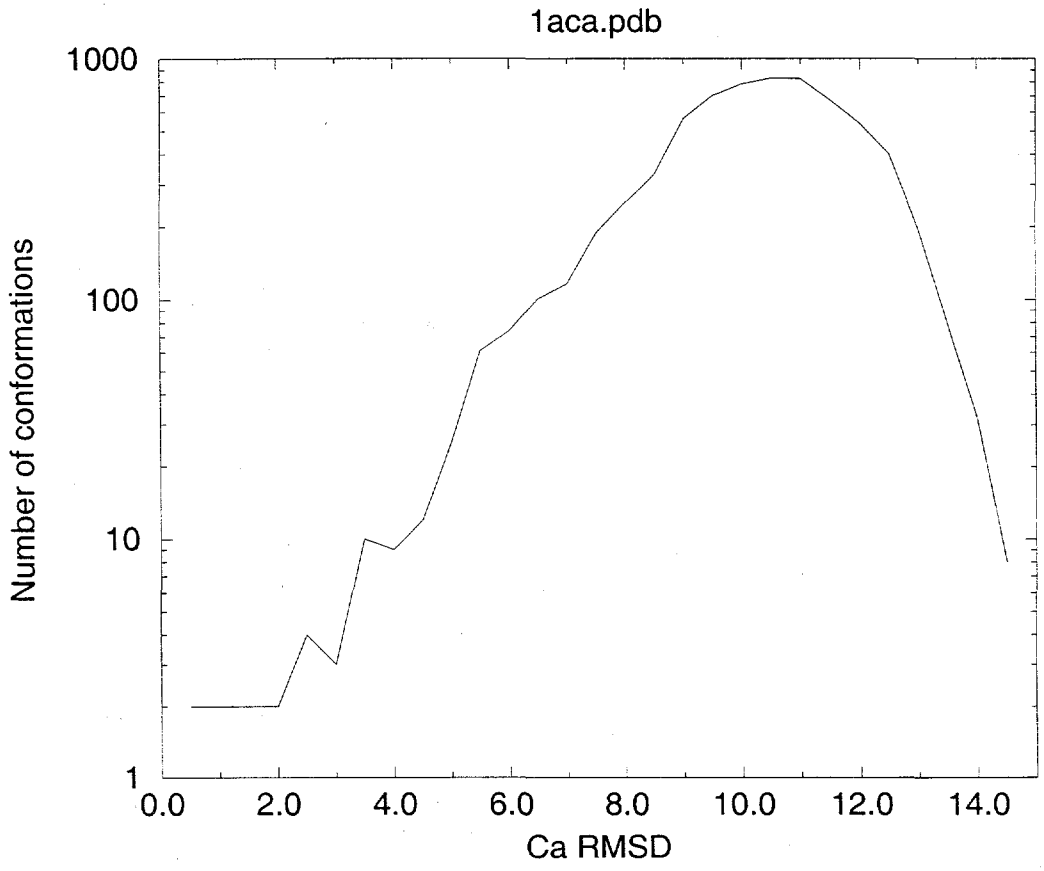


FIG. 4. RMSD from the native for sampled conformations for 1aca.pdb. The distribution appears gaussian with an enriched tail of native-like structures

REFERENCES

[1] Aszodi, A., Gradwell, M. J., Taylor, W. R. (1995). Global fold determination from a small number of distance restraints. *J. Mol. Biol.*, **251**, 308-326.

[2] Chelvanayagam, G., Knecht, L., Jenny T., Benner, S. A., Gonnet, G. H. (1998). A combinatorial distance-constraint approach to predicting protein tertiary models from known secondary structure. *Fold. Design*, **3**, 149-160.

[3] Eyrich V.A., Standley, D.M., Friesner, R.A. (1999). Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J. Mol. Biol.*, **288**, 725-742.

[4] Frishman, D., Argos, P. (1995). Knowledge-based secondary structure assignment. *Proteins: structure, function and genetics*, **23**, 566-579.

[5] Havel, T. F. (1991). An evaluation of computational strategies for for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog. Biophys. Mol. Biol.*, **56**, 668-682.

[6] Huang, E.S, Samudrala, R, and Ponder, J.W. (1999). *ab initio* Fold Prediction of Small Helical Proteins using Distance Geometry and Knowledge-based Scoring Functions. *J. Mol. Biol.* **290**, 267-281.

[7] Kraulis P.J., *J. Appl. Cryst.* (1991), **24**, 946-950.

[8] Levitt, M. (1992). Accurate modeling of protein conformations by automatic segment matching. *J. Mol. Biol.*, **226**. 507-533.

[9] Murzin A.G., Brenner S.E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536-540.

[10] Park, B. and Levitt, M. (1996). Energy functions that discriminate X-ray and near

native folds from well-constructed decoys. *J. Mol. Biol.*, **258**, 367-392.

[11] Reva, B. A., Finkelstein, A. V, and Skolnik, J. (1998). What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å? *Fold. Design*, **3**, 141-147.

[12] Simons, K.T., Ruczinski, I., Koopenberg, C., Fox, B.A., Bystroff, C.A,., and Baker, D. (1999). Improved Recognition of Native-Like Protein Structures Using a Combination of Sequence-Dependent and Sequence-Independent Features of Proteins. *PROTEINS: Structure, Function, and Genetics*, **34** 82-95.

[13] Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209-225.

[14] Sippl, M. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge based prediction of local structures in globular proteins. *J. Mol. Biol.*, **213**, 859-883.

[15] Lund, O., Hansen, J., Brunak, S., Bohr, J. (1996). Relationship between protein structure and geometrical constraints. *Protein Sci.* **5**, 2217-2225.

[16] Mumenthaler, C., Braun, W. (1995). Predicting the helix packing of globular proteins by self-correcting distance geometry. *Protein Sci.*, **4**, 863-871.

[17] Ortiz, A. R., Kolinski, A., Skolnik, J. (1998). Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.*, **277**, 419-448.

[18] Smith-Brown, M. J., Kominos, D., Levy, R. M. (1993). Global folding of proteins using a limited number of distance restraints. *Protein Eng.*, **6**, 605-614.

[19] Eyrich V. A., Standley D. M., Felts A. K., Friesner R. A. (1999). Protein tertiary structure prediction using a branch and bound algorithm. *Proteins*, **35**, 41-57.

[20] Hinds D. A., Levitt M. A (1992). Lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. USA*, **89**, 2536-2540.

[21] Hinds D. A., Levitt M.(1994). Exploring Conformational Space with a Simple Lattice Model for Protein Structure. *J. Mol. Biol.* **243**, 668-682.

[22] Covell D. G. (1994). Lattice model simulations of polypeptide chain folding.*J. Mol. Biol.* **235**, 1032-1043.