Iterative Assembly of Helical Proteins by Optimal Hydrophobic Packing

G. Albert Wu[1,3], Evangelos A. Coutsias[2] and Ken A. Dill[1,*]

[1]Department of Pharmaceutical Chemistry, University of California in San Francisco,
San Francisco, California 94143-2240.

[2]Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131.

[3]Present address: Physical Biosciences Division, Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley, California 94720

[*]Correspondence: dill@maxwell.compbio.ucsf.edu
Phone: (415) 476-9964 Fax: (415) 502-4222

Running title: Iterative assembly of helical proteins

SUMMARY

We present a method for the computer-based iterative assembly of native-like tertiary structures of helical proteins from alpha-helical fragments. For any pair of helices, our method, called MATCHSTIX, first generates an ensemble of possible relative orientations of the helices with various ways to form hydrophobic contacts between them. Those conformations having steric clashes, or a large radius of gyration of hydrophobic residues, or with helices too far separated to be connected by the intervening linking region, are discarded. Then, we attempt to connect the two helical fragments by using a robotics-based loop-closure algorithm. When loop closure is feasible, the algorithm generates an ensemble of viable interconnecting loops. After energy minimization and clustering, we use a representative set of conformations for further assembly with the remaining helices, adding one helix at a time. To efficiently sample the conformational space, the order of assembly generally proceeds from the pair of helices connected by the shortest loop, followed by joining one of its adjacent helices, always proceeding with the shorter connecting loop. We tested MATCHSTIX on 28 helical proteins each containing up to 5 helices and found it to heavily sample native-like conformations. The average RMSD of the best conformations for the 17 helix-bundle proteins that have 2 or 3 helices is less than 2 Å; errors increase somewhat for proteins containing more helices. Native-like states are even more densely sampled when disulfide bonds are known and imposed as restraints. We conclude that, at least for helical proteins, if the secondary structures are known, this rapid rigid-body maximization of hydrophobic interactions can lead to small ensembles of highly native-like structures. It may be useful for protein structure prediction.

INTRODUCTION

For predicting the native structures of proteins, a useful computational strategy is to assemble known secondary structures into putative native tertiary structures, and then to use a scoring function to seek the best such chain packings. Our interest in this approach was motivated by our recent use of an all-atom physical force field, Amber 96 (Cornell et al., 1995) with implicit solvent (Onufriev et al., 2004), for scoring conformations that have been generated via a folding-mechanism-inspired search method called Zipping and Assembly (ZAM) (Ozkan et al., 2007). When limited to these putative folding routes, ZAM found the native structures of a test set of 8 out of 9 small globular proteins to within about 2 Å root-mean-square-deviation (rmsd) of their experimental structures. More recently, we also tested ZAM in the 7th community wide experiment on the Critical Assessment of techniques for protein Structure Prediction (CASP7) (Moult, 2005). ZAM found secondary structural elements relatively efficiently, but was slow to assemble those secondary structures into tertiary native-like conformations (Shell et al., unpublished). Assembly was often bottlenecked by side chain packing and re-arrangements (Bromberg and Dill, 1994). Our interest here is in more efficient ways to sample different possibilities of assembling secondary structures into native-like tertiary structures. We consider here only water-soluble α helical proteins, but we believe a similar approach with an appropriate scoring function should also be useful for other types of secondary structure assemblies.

There has been much previous work in assembling tertiary structures from secondary structural fragments (Fain and Levitt, 2003; Fleming et al., 2006; Hoang et al., 2003; Kolodny and Levitt, 2003; Simons et al., 1997; Yue and Dill, 2000), especially in helix packing (Bowie and Eisenberg, 1994; Cohen et al., 1979; Crick, 1953; Fain and Levitt, 2001; Huang et al., 1999; Kohn et al., 1997; Lupas et al., 1991; McAllister et al., 2006; Mumenthaler and Braun, 1995; Nanias et al., 2003; Narang et al., 2005; Wolf et al., 1997; Zhang et al., 2002). Yue and Dill (Yue and Dill, 2000) used a set of discrete helix-helix packing angles for tertiary structure assembly. Zhang et al. (Zhang et al., 2002) used torsion angle dynamics and predicted interhelical contacts as restraints for fold prediction. The Floudas group (McAllister et al., 2006) has predicted primary and helical-wheel interhelical contacts and then generated interhelical distance restraints in alpha-helical globular proteins. Fain and Levitt used a packing algorithm based on graph theory and database-generated contact information (Fain and Levitt, 2001). Using a Cα-only protein model, the Scheraga group (Nanias et al., 2003) generated native-like folds of alpha-helical proteins by the global optimization of a Miyazawa-Jernigan-based contact potential function (Miyazawa and Jernigan, 1996). More recently, Narang et al. (Narang et al., 2005) have used knowledge-based biophysical filters of persistence length and radius of gyration for pruning out unlikely conformational candidates, followed by Monte Carlo optimization of loop dihedrals, to bracket native-like structures for small helical proteins.

Our approach is different, and has the following features:

1) We do not rely on database-derived packing information, such as helix-helix packing angles. Instead, we start with canonical helices (backbone φ = -57° and ψ = -47°) to represent the helical fragments in the native structure, with side chain dihedrals sampled from a rotamer library (Dunbrack, 2002; Dunbrack and Karplus, 1993).

2) To seek optimal hydrophobic packing, we align the helices as rigid-body cylinders by matching up every pair of inter-helical hydrophobic residues subject to certain restraints.

3) We then connect the two helices via their linking chain using a fast robotics-based analytic loop closure algorithm (Coutsias et al., 2004; Coutsias et al., 2005) that generates an ensemble of loop conformations for a given pair of aligned secondary structures. Our method incorporates probability-weighted *Sobol quasirandom sampling* (Bratley and Fox, 1988) of the Ramachandran accessible regions for the $\phi - \psi$ torsions, which further enhances the efficiency in finding loop-closure solutions.

4) The iterative assembly of additional helices is further optimized by ordering the choices: adjacent helices connected by short loops are assembled before helices separated by long loops. And, in later iterations, an adjacent helix is joined to the pre-existing assembly (in case of two adjacent helices, the one with the shorter loop is chosen). This process is repeated until all helices are assembled.

This iterative assembly of given secondary structures, in the two steps of combining the helices then linking the loops, implemented in the algorithm called MATCHSTIX, is much more efficient in sampling native-like conformations than other methods, such as the backbone dihedral rotation of the loop residues (Narang et al., 2005; Ozkan et al., 2007) or anisotropic-network-model sampling (Atilgan et al., 2001; Ozkan et al., 2007). MATCHSTIX follows a greedy conformational search strategy; this largely circumvents the multi-component combinatorial explosion problem and brings into feasibility the assembly of multiple helices even in all-atom representations of proteins.

Details of the method are described in the Experimental Procedures section.

RESULTS

Assembly of Multi-Helical Protein Structures

We have tested MATCHSTIX on a set of 28 helical proteins each consisting of up to 5 helices. Five of these proteins contain disulfide bridges. This set of proteins partially overlaps with previous test sets (Nanias et al., 2003; Narang et al., 2005; Zhang et al., 2002). Hence we can make some comparisons of our method with those. To evaluate the quality of the sampling and scoring, the top 1, 5, 20 and 50 structures from the last assembly step are analyzed and the RMSD of the most native-like structure among them is calculated relative to the native conformation for $C\alpha$ atoms of the helical residues.

The results are summarized in Table 1 and Fig 1. For calculating RMSD and for calculating a quantity we call R_h, the all-atom radius of gyration of hydrophobic residues, we consider only the helical residues, because loops, especially the longer ones, are often floppy and not well defined in the known structures, and it turns out that their detailed structure doesn't strongly affect the performance of the packing algorithm. We find that for 2-helix and 3-helix bundles, sampling often explores low RMSD conformations (about 2 Å or smaller) that tend to rank in the top 20 or better. For 4- and 5-helix-bundle proteins, native-like conformations are also frequently sampled, but the errors are somewhat worse, with the lowest RMSDs being in the 3-5 Å range and among the top 50 conformations.

The final ranking of the conformations corresponding to cluster centroids depends on the cutoff distance for the clustering. A larger cutoff gives a smaller number of clusters and generally improves the positional ranking of native-like conformations. However, the lowest RMSD structures may be filtered out as a result of using a larger cutoff. As an example, for the 69-residue 3-helix bundle 2A3D, a 2 Å cutoff gives 784 conformations with the lowest RMSD (2.29 Å) ranking at position 30. Increasing the cutoff to 3 Å reduces the ensemble size from 784 to 134, and the lowest RMSD increases from 2.29 to 2.50 Å with its improved ranking at position 9. For both cutoffs, the lowest Rh conformation has an RMSD of 9.86 Å. In Table 1, we use the clustering cutoff n-1 Å for an n-helix protein which does not have any SS bonds. For the 5 disulfide-containing proteins, due to the significant reduction in population by imposing the Cβ-Cβ distance restraint between the disulfide-bonded pair, we have used smaller clustering cutoffs of 1.5 Å and 2.0 Å for 3-helix and 4-helix proteins respectively.

Comparison with Other Methods

Other groups have also previously developed helix packing methods (Nanias et al., 2003; Narang et al., 2005; Zhang et al., 2002). We cannot make a full comparison because of the incomplete overlap of their test sets with ours. But we are able to make a few comparisons. First, the loop torsion sampling (LTS) method (Narang et al., 2005) samples the backbone torsion angles of the loop residues to generate a diverse set of relative orientations of the helices. Since it works with all-atom protein models, a direct comparison can be made with our approach. Whereas the performance of our method improves for longer loops, the LTS method works best with short loops. Table 2 compares them. For LTS, the RMSDs tend to be in the range of 4 Å, whereas RMSDs from the present method tend to be in the range of less than 2 Å. In addition, the present method is more efficient computationally. For the 3-helix bundle protein 1GVD, with 2.8 GHz Xeon processors, our method takes about 20 CPU hours, compared to about 200 CPU hours for LTS.

In another approach, the Scheraga group packed helices using a coarse-grained potential (Nanias et al., 2003), where each amino acid is represented by its Cα atom. A simplified energy function was used to capture the pairwise interaction between two residues from two helices. Their treatment of the loop was limited to requiring that the ends of the two helices to be linked must be smaller than the maximal loop length. The helices were treated as rigid bodies, and best helix packing orientations are generated by global optimization of the potential energy. Given the simplicity of their protein model and energy function, it is remarkable that their method could reproduce native-like folds of dozens of helical proteins as local energy minima of the energy function. A direct comparison between their method and ours is difficult both because of the different protein models used (coarse-grained vs. all-atom) and because of the different treatment of loop residues (implicit vs. explicit). Nonetheless, the results from the two methods on a set of seven helical proteins that we tested in common are listed in Table 2 for reference. For the five three-helix bundles, the average RMSD of the most native-like structures is 2.37 Å and 3.1 Å for our method and theirs respectively, while their method is better for the 4- and 5-helix proteins. Hence, in this limited test, the quality of predictions appears to be equivalent. A useful aspect to our approach is that it retains full atomic detail, including in the loops.

Proteins Containing Disulfide Bonds

We also tested our method on five proteins having disulfide bonds. For these proteins, we have imposed SS bond restraints as described in the EXPERIMENTAL PROCEDURES section, and the final assembly results are summarized in Table 3. Some general observations can be made about these tests: 1) Near-native configurations are sampled even more densely when native SS bond restraints are imposed. 2) These near-native structures appear among the top 20 or 50 conformations indicating that Rh ranking can still serve a useful filter.

To assess the effects of SS bond restraints, we have also run tests by ignoring the SS bonds and not imposing any restraints. The results are summarized in Table 3. The absence of the SS bond restraints leads to a much larger conformational space which also makes the search for near-native structures more difficult. As a result, we observe bigger RMSD values for the best structures sampled in the absence of SS bond restraints. To find the lowest RMSD and its Rh ranking as shown in Table 3, the final conformations from the last iterative assembly step have been clustered with clustering cutoffs of 1.5 and 2.0 Å for the 3-helix and 4-helix proteins respectively, and the centroids are kept as representative conformations and are ranked by Rh.

DISCUSSION

Not surprisingly, our computational assembly and sampling method performs better on proteins having fewer helices (2-3) than on proteins having more (4-5). The sampling quality is not very sensitive to the number of amino acids in the protein, but it decreases significantly with the number of helices that are assembled, because of the exponential growth in conformations with helix number. The predictions also depend, to some extent, on the fold. We obtain good structures for the 4-helix bundle protein 2MHR (91 residues), which has the up-down-up-down motif with parallel helices. The lowest-RMSD structure is 2.1 Å away from the native conformation and ranks number 8 among more than 400 conformations.

It is possible that the assembly order that we have adopted might bias our structures away from certain topologies, by excluding for example some arrangements where a certain helix is wedged between two helices that our algorithm might preferentially bring into contact first. Such topologically frustrated arrangements might be better explored by a fully combinatorial approach based on exhaustive exploration of all possible contact graphs, like the approach of Fain and Levitt (Fain and Levitt, 2001).

The performance of Rh is shown in Fig. 2. The figure shows that while the Rh criterion is not good enough to uniquely pick out native structures, it is a useful filter for identifying relatively small ensembles within which the native structure can be found. One example is the 3-helix bundle 2A3D. Among the sampled conformations, the most compact structure (9.86 Å rmsd from native) has helices 1, 2 and 3 packed in a counter-clockwise fashion when viewed from the N-terminal along the 1st helix. The lowest rmsd conformation (2.3 Å relative to native) however, has the three helices packed in a clockwise fashion. And the difference in Rh between the two conformations is only 2%. Another example is the 5-helix bundle protein 2ICP, whose lowest Rh conformation has a non-native packing. Though the near-native conformations may not have the lowest Rh score, they generally appear among the top-ranking conformations in our test set.

Further improvements in our method may be possible by going beyond Rh as a simple measure of initial quality. This becomes necessary when dealing with proteins with more than one hydrophobic core, or when the protein structure is held together predominantly by forces other than the hydrophobic effect as in the case of inter-helical SS bonds. This is illustrated in Fig. 2 by the results on 3 proteins containing SS bonds. One example is the 3-helix bundle 1HP8, which has 3 inter-helical disulfide bridges. The disulfide bonds hold the protein in a non-optimal conformation, relative to a simple compactness criterion, with the third helix pointing away from the other two helices. Another example is the 4-helix bundle 1J0T. As can be seen from the figure, although the native state of 1J0T ranks poorly compared to the sampled conformations, the most native-like conformation (rmsd 2.7 Å) can still rank near the top at position 28. In this work, we have imposed SS bond restraints to cut down on the conformational search space, and are able to sample native-like conformations. But a better scoring function is clearly needed to improve the ranking of the best sampled structures.

Another challenge to the simple Rh scoring function is posed by proteins with long loops. For example, the 3-helix bundle protein 1FEX has a 10-residue loop connecting helix 1 and helix 2. This long loop is structured in its native conformation with several hydrogen bonds. In this work, we have not explored the diverse conformations of long loops, but have focused only on the packing of helices. Our calculation of RMSD and the Rh does not include loop residues, but an improved method might result from including them.

In the present study, our secondary structures were given as input and taken to have canonical alpha-helical structures. This is essential for the purpose of testing an assembly algorithm. However, had our purposes been different, starting secondary structures could have been obtained, instead, from other sources. The starting secondary structures could also be obtained from all-atom molecular dynamics simulations (Ho and Dill, 2006; Ozkan et al., 2007) or from secondary structure prediction servers (Cuff and Barton, 1999; Jones, 1999; Rost et al., 2004), both of which can successfully predict helices. A previous study (Nanias et al., 2003) showed that the final assembled structures are not very sensitive to the secondary structure assignments. Our tests also indicate that, at least for the set of proteins we considered, the assembly performance is mainly determined by the hydrophobic core and not sensitive to the conformational details of the loop residues.

Conclusions

We have presented an iterative assembly algorithm for constructing native-like tertiary structures from individual helical fragments. We show that the method is much faster and more efficient at sampling native-like structures for two- and three-helix bundles than the previous methods for which we can make a direct comparison. Moreover, the present method can be used directly with all-atom physical forcefields, as we have done here, and does not require a first coarse-grained step.

The best structures (i.e., lowest RMSD) among the top 1, 5, 20, 50 and all sampled conformations average respectively 4.7, 3.6, 2.2, 2.1 and 2.0 Å RMSD for the Cα atoms of the helical residues for the 17 2- and 3-helix bundles. Errors are somewhat larger for proteins with more helices, where there may be advantage to coarse-graining on a simpler energy landscape (Nanias et al., 2003).

Our method is robust in the following respects. First, its performance is not sensitive to small variations in the secondary structure assignments. For example, the length of a long loop may be shortened by assigning helical conformations to some loop residues, and this in general does not change the final structures at the end of the assembly. Similarly for one- or two-residue loops (e.g. 1X9B), one can extend the loop length by a few residues and still sample native-like structures in its final, top-ranked conformations. This is consistent with the fact that the tertiary structure is largely determined by the hydrophobic core of residues (hydrophobic effect), and can allow fluctuations in secondary structures of certain residues not participating in the hydrophobic core. Second, the helical packings are generally insensitive to the structures of the loops generated between them. We believe that these computational methods may be useful in all-atom physical protein structure prediction and refinement for helical proteins.

EXPERIMENTAL PROCEDURES

In order to assemble helical fragments into tertiary folds that have a compact hydrophobic core, we start with *n* helical fragments to be assembled along with *n*-1 connecting loops. We assume canonical backbone torsions for the helical residues. Since the loop closure algorithm requires a minimum of 6 variable backbone torsions (i.e. at least 3 loop residues), it is necessary to extend tight turns of 1 or 2 residues to a loop of at least 3 residues by shuffling the adjacent helical residues to the loop. Given a loop with k backbone $(\phi - \psi)$ torsion pairs, k-3 of these are chosen to lie in the Ramachandran regions of their corresponding residues, while the remaining 3 pairs, belonging to residues used as *pivots* for loop closure, are set by the algorithm to satisfy closure constraints. Since these must also be screened for Ramachandran compatibility in order for the resulting loops to be viable, we do not allow any of the pivot residues to be a Proline. Thus, a loop that is closable by our algorithm needs to include at least 3 non-Proline residues. The iterative assembly starts with the 2 helical fragments connected by the shortest loop. For each subsequent iteration, one adjacent helix is chosen along with the connecting loop. If there are 2 adjacent helices, the one with the shorter connecting loop is chosen. The process is repeated till all helical fragments are assembled and all loops are joined. For *n* helical fragments, the assembly will finish in *n*-1 iterative steps.

The backbone and side chain geometries of the helical fragments are chosen as follows. The bond lengths and bond angles are set to their canonical values as used in the InsightII molecular modeling suite (http://www.accelrys.com/products/insight/). The canonical backbone torsions of φ = -57° and ψ = -47° are used for all helical residues at the outset. For this work, we take secondary structure information from the native structure according to the DSSP definition (Kabsch and Sander, 1983). For NMR ensembles, we use the minimized average structure as the native conformation. The side chain dihedrals of each helical fragment are sampled from a rotamer library (Dunbrack, 2002; Dunbrack and Karplus, 1993); details are given below. The number of side chain conformations of a given helical fragment is mainly determined by its associated loop length and to a lesser degree by the iteration cycle. This is because, for short loops, only a very small percentage of the sampled conformations can be loop-closed, and consequently a large and diverse sample is needed to generate sufficient loop closures. For example, if the first iteration cycle has a 3-residue loop or 4-residue loop including a Proline, we use 7 or 5 side chain conformations for each of the 2 helical fragments (i.e. 49 or 25 pairs for the 2 different loop lengths respectively) before we use MATCHSTIX to generate a diverse set of relative orientations for each helix pair. On the other hand, if the first iteration cycle has a longer loop, 3 side chain conformations for each helix can be used. For later iteration cycles and also longer loops, the side chain conformations for the added single helix is reduced to 2 at the second assembly cycle, and 1 at the third and later iterations for the assembly of 4 or more helices.

The reason for the decreasing number of side chain conformations of the single helix at each subsequent assembly cycle is because of the rapidly increasing number of partially assembled conformations with which the helix must pair up. These partially assembled structures not only have a diverse range of helix backbone arrangements, their side chain dihedrals have also been modified in diverse ways during energy minimization. If we keep more side chain conformations for the single helix, we will need to cut down on the number of partially assembled conformations for sake of computational efficiency. Exactly how many conformations are kept for each assembly cycle depend on the resolution and diversity of the final assembled structures in terms of relative rmsd, and this is explained in the following section.

We find that the native-likeness of the final assembled structures is not sensitive to the number of side chain assignments for each single helix except for very short loops. In fact, for a number of 3-helix bundles, the final ensemble contains low rmsd conformations even if only a single side chain conformation for each of the three helices is used. This may be understood from two aspects. 1) Energy minimization after loop closure has redistributed the side chain conformations. 2) The native state does not adopt a single side chain conformation but rather undergoes dynamic fluctuations. Both X-ray and NMR protein structures exhibit large side chain conformational entropies (Zhang and Liu, 2006).

During the assembly process, we treat these helical fragments as rigid bodies, except for the minor distortion caused by energy minimization after loop closure. The energy minimization is done to remove (mainly minor) atomic steric clashes. Our assembly process for helical proteins, as it is implemented in the algorithm MATCHSTIX is divided into 4 stages: 1) we align two helical fragments, absent the connecting loop, and keep the compact conformations as measured by R_h;   2) we connect the loop;   3)   we minimize the energy, cluster the conformations, and retain a representative set of conformations; 4)   we iterate steps 1 - 3 until all fragments are assembled.

MATCHSTIX: An Algorithm for the Iterative Assembly of Helical Proteins.

A. Rigid-Body Alignment to Match up Hydrophobic Residues

In the first step, we align two helices to achieve good hydrophobic matching between them. The cylindrical geometry of a canonical alpha helix can be specified by the N-Cα-C backbone atoms of any one hydrophobic residue in the helix. The origin of each coordinate system is located at the intersection of the cylindrical axis and the circular cross section containing the Cα atom of the residue (see Fig. 3). The axis of the cylinder defines the x-axis that points from the N-terminal to the C-terminal, the z-axis points from the origin to the given Cα atom, and the y-axis is defined such that x-y-z forms a right-handed Cartesian coordinate system. The initial alignment between the two coordinate systems is such that the two cylindrical axes are parallel with a separation of 10 Å along the z direction, with the two residues facing each other (i.e. the two z axes are anti-parallel). In other words, the origin of the second coordinate system O_2 is at (0, 0, 10 Å) relative to the first coordinate system. This orientation brings the hydrophobic patches from the two helices into contact. From this initial orientation, rotations and translations are used to generate a distribution of relative orientations. For translational moves, O_2 can vary within a cube centered at the initial position of O_2 and of size 10 Å, 12 Å, and 10 Å along the x_1, y_1, and z_1 directions of the first coordinate system respectively. A full range of angular distribution is generated by rotating around x_1, y_1, and z_1 with angles in the range –90 to 90, -45 to 45, and -90 to 90 degrees respectively. By trying to match up every pair of hydrophobic residues from the two helical fragments, a wide range of relative orientations with hydrophobic contacts are generated.

The conformations that are generated in this way are pruned based on three criteria:

1) No severe steric clashes. The minimal heavy-atom distance between the two peptide fragments is required to be no less than 2.5 Å, which is slightly smaller than a typical hydrogen bond length.
2) The loop can close. The distance of the connecting ends of the two fragments must be smaller than the maximal loop length for the given sequence of loop residues.
3) There is sufficient hydrophobic compactness. We determine the hydrophobic radius of gyration, Rh, for all atoms of the hydrophobic helical residues. We keep only those structures having hydrophobic amino acids tightly clustered in space, in order to ultimately lead to a hydrophobic core for the whole protein. For this purpose, we impose an upper cutoff of 5 Å for the minimal heavy atom distance between the two peptide fragments. Note that optimal hydrophobic packing for a final assembled structure allows for less than optimal packing for the partial structures. For example, to assemble a 4-helix-bundle protein in 3 iterative steps, the top 20%, 15%, 10% of the most Rh compact conformations are retained for the 1st, 2nd, and 3rd iterations respectively.

To match up a pair of hydrophobic residues from 2 helical fragments by rigid body translation and rotation, we use 6-dimensional Sobol quasirandom sampling to generate trial orientations. Not every pair of hydrophobic residues can be matched up and satisfy the above 3 constraints, especially for short loops. For this reason, up to a maximum of 100 trial conformations for each residue pair are examined till one feasible structure is found. This will avoid wasting too much time on many un-bridgeable residue pairs. We cycle through all hydrophobic residue pairs for one or more times till a specified number of relative orientations are obtained.

The number of feasible relative orientations thus generated depends on the loop length, due to the low closure rate for short loops. We typically generate up to ten thousand conformations from all helix pairs and all possible hydrophobic residue pairs for a 3-residue loop or 4-residue loop with a Proline loop residue for the first assembly iteration. For later iterations, two hundred conformations are generated for every pair of helical fragments.

For proteins containing inter-helical disulfide bonds, further pruning is possible by requiring that the Cβ-Cβ distance of the bonded residues be smaller than 8 Å.

These structures produced by MATCHSTIX all have good hydrophobic compactness and exhibit a diverse arrangement of side chain packing. Next, they are subjected to loop closure.

B. Closing the Loops

After assembling the helices into an ensemble of favorable structures, we then connect the two helices via the linking loop region of the chain, using a loop closure method we have described previously (Coutsias et al., 2004; Coutsias et al., 2005). Our loop-closure algorithm follows previous work (Dodd et al., 1993; Go¯ and Scheraga, 1970; Wedemeyer and Scheraga, 1999) but is more general in allowing for loops of arbitrary length (≥3 peptides) and arbitrary nonplanar peptide bond structure. Our method requires that there must be at least six intervening torsions whose axes form three distinct coterminal pairs and whose values are degrees of freedom for the loop. All other internal degrees of freedom of the loop (bond lengths, angles, and remaining torsions) can be fixed to any arbitrary value. For this study, bond lengths and bond angles are set to their canonical values as in the InsightII molecular modeling suite (http://www.accelrys.com/products/insight/), while the remaining torsions can be sampled, and they are restricted to the Ramachandran-accessible
(Lovell et al., 2003; Ramachandran et al., 1963) regions.

Our algorithm is considerably simpler to program than more general robotic algorithms, such as (Lee and Liang, 1988) that remove the coterminal axes restriction. Like (Lee and Liang, 1988), our method leads to a robust formulation in terms of multivariate polynomials, that are solved by converting to an ideally dimensioned 16×16 generalized eigenvalue problem. However, because of the simplicity, our method is preferable for most situations related to modeling protein backbones for which, with the exception of Proline, each residue adds two flexible torsions, φ and ψ, at each of the Cα atoms.

In our scheme, we assume that the loop, $N-1$ residues long, is to bridge two residues (the *anchors*), R_0 and R_N, whose positions are fixed in space. Then:

1. Select 3 residues, R_a, R_b, R_c, with $1 \le a < b < c \le N-1$. These are the *pivots* for loop closure and their ϕ, ψ torsions will be chosen automatically to close the loop. None of these may be a Proline.
2. Break the loop into four segments, $R_1 \cdots R_a$, $R_a \cdots R_b$, $R_b \cdots R_c$ and $R_c \cdots R_{N-1}$ (here R_i stands for i-th residue but can be also thought as the Cartesian coordinate vector for the Cα atom of that residue). For each of these, set all of their internal degrees of freedom to predetermined values. The 6 torsions and 3 bond angles about the pivots are not introduced at this stage.
3. Attach the first and last segments to the corresponding anchor residues elongating the end chains to $R_0 \cdots R_a$ and $R_c \cdots R_N$. These two chains are now fixed in space, and their end residues, the pivot residues R_a and R_c, are the new anchors.
4. With the residues R_a and R_c now fixed, form a triangle whose three sides have lengths $L_a = \|R_b - R_a\|$, $L_b = \|R_c - R_b\|$, $L_c = \|R_a - R_c\|$. If this triangle is feasible (i.e. the three sides obey the various triangle inequalities), then the loop closure problem is solvable in principle, else the particular combination of the free parameters is rejected.
5. If the triangle above is feasible, we proceed with formulating the generalized loop closure equations. The details of this step can be found in (Coutsias et al., 2004; Coutsias et al., 2005). In our formulation the atom R_b lies in a circle about the axis (R_c, R_a). Assuming its location is known, the other two chains can be rotated about their respective axes, (R_a, R_b) and (R_b, R_c) so that the bond angles (N, Cα, C) at each of the three pivot atoms have prescribed values. Thus our algorithm involves three unknown angles and three constraints. Setting these, completely fixes all atoms in space. In general, there can be as many as 16 alternative conformations produced by this algorithm. As the solutions appear in the form of the roots of a real polynomial of degree 16, there can be at most 16 real roots, corresponding

to physically realizable conformations. If any real solutions exist there is always an even number of them, often considerably fewer than the maximum 16.

6. The torsions at the pivot dihedrals are now screened, and only loops all of whose torsions are in the Ramachandran regions (Lovell et al., 2003; Ramachandran et al., 1963) are kept as possible leads.
7. The loops that satisfy Ramachandran conditions are fit with sidechains from the proability-sorted, backbone-dependent Dunbrack rotamer library, **bbdep02.May.sortlib**, freely available at http://dunbrack.fccc.edu/bbdep/bbdepdownload.php (Dunbrack and Karplus, 1993). The χ angles for each sidechain are chosen with probabilities from the rotamer library's values. If the assignment leads to a steric clash the rotamers are resampled until the clash is removed or until a preset limit is reached. For this study we allowed up to 50 resamplings. Setting that limit to higher values had no appreciable effect on producing viable structures. The resulting complete protein is screened for steric clashes among loop atoms or between the loop and either of the protein fragments that it connects to.
8. Conformations that pass the steric test are kept as possible alternatives for energy minimization.

The purpose of these steps is neither to find native loop conformations, nor to sample extensively, but merely to generate loop conformations that are closed and sterically viable. Hence, unlike a search for native loop conformations, our loop closure problem gets easier for longer loops. Smaller loops can have constraints that are challenging to satisfy. Hence, for short loops, we allow flexibility in the ψ angle at R_0 and/or the φ torsion at R_N, thus enlarging the set of end poses and increasing the probability of choosing values for which the loop is closable. Sometimes even for larger loops, it can be difficult to find acceptable leads, if there are partial confinements (e.g. proteins 1FEX and 1HP8). It is therefore desirable to sample the space of the free torsions uniformly and at ever-increasing resolution, until all components of the solution set are located. Here we use Sobol quasi-random sampling (Bratley and Fox, 1988), a number-theoretic algorithm that generates a sequence of k points that is nearly uniformly distributed in an (N-4)-dimensional unit hypercube, independent of k. Its key advantage is that to increase sampling resolution one simply adds new points to the existing ones, without affecting the near-uniformity and quasi-randomness of the sequence. In our implementation, the Ramachandran regions for each residue corresponding to (φ,ψ) pairs with higher than 5% probability for each residue are pixelated into 5 degree squares. These squares are rearranged along a linear dimension, so that to each pixel there corresponds an interval of length $p(\phi,\psi)$/M, with M the total number of pixels and p a measure of the probability of finding a torsion pair at a given position in the Ramachandran plot (Lovell et al., 2003). A unit hypercube of dimension equal to the number of sampled residues is constructed in this way, and points in it (pixel (*N*-4)-tuples) are chosen with the Sobol algorithm. We use a maximum of 200 trial backbone loop conformations for each loop closure. A larger number of trial conformations can be used, at the expense of more computing time wasted on non-closable loops. Although we could close more loops if we were to allow perturbations of omega torsions or bond angles, we would be introducing strains which might lead to significant distortions when we minimize energy. For canonical backbones, we find that the shortest loop closure problem, i.e. for 3 residue loops, imposes severe restrictions on the relative poses of the end bonds (N_1-$C\alpha_1$ and $C\alpha_3 - C_3$) for which closed loops can be found at all: fixing the distance of the two end Cα atoms $(C\alpha_1 - C\alpha_3)$ to a range where closure is possible in principle, we find solutions for at most 20% of the end poses at best (when $C\alpha_1 - C\alpha_3$is in the range of 5.5 to 6.5 Å), and this number falls off to zero quickly outside this range. Allowing a 10-20 degree strain in the ω torsions does not alter this result considerably. Of course, for longer loops this restriction becomes gradually less significant, however it is still a lot easier to close loops if the end points are at a distance that is a certain fraction of the maximum length attainable by the loop in extended conformation.

C. Energy Minimization and Clustering

Such closed-loop conformations found in this way generally still have minor steric clashes or energetically unfavorable side chain conformations. So, we then subject these conformations to energy minimization. We use the energy minimizer in the Amber9 molecular modeling software package (Case et al., 2005). We use the Amber ff96 all-atom forcefield (Cornell et al., 1995) with the generalized Born implicit solvent model (Onufriev et al., 2004). We use 30 steps of steepest descent followed by 30 steps of conjugate gradient minimization for each conformation.

For proteins with disulfide bonds, the pruned conformations from MATCHSTIX based on a Cβ-Cβ distance cutoff generally do not have the correct disulfide-bridge (SS) geometry. This can be corrected by Amber energy minimization, whose energy function has terms associated with SS bridges.

The number of loop-closed, energy-minimized conformations grows rapidly for each subsequent iteration, due to the exponentially growing conformation space with the helix number. To keep a manageable size of seed conformations for the next iteration, we cluster the top 1000 most compact structures and use the cluster centroids as representative structures. The compactness is measured by Rh, the radius of gyration of all atoms of the hydrophobic helical residues of the energy-minimized structures. The clustering procedure is used to remove highly similar conformations.
For efficiency, we use an approximately linear clustering method whose pseudo-code is as follows: for a given cutoff and an ordered list L of the conformations,

```
the first conformation is assigned to the first cluster and removed from L
while L not empty:
    c = 1st conformation from L
    for cluster k of the existing clusters:
        if distance between c and 1st member of k < cutoff:
            add c to cluster k as its last member
            break out of the loop
        end
    end
    if c is not added to any of the existing clusters:
        assign c to  a new cluster
    end
    remove c from L
end
```

The clustering time is roughly proportional to the number of conformations to be clustered, if most of them resemble one another within the cutoff distance. We measure the distance between two conformations by the Cα rmsd of the helical residues. As a rule of thumb, the distance cutoff for the assembly of n helices can be taken as n-1 Å. Slightly smaller rmsd cutoffs of 1.5 and 2 Å are used for disulfide-bridge-containing 3 and 4 helix bundles respectively, to compensate for the smaller sample size after the Cβ-Cβ distance screening. The cluster centroids, defined as the conformation with the smallest Rh within each cluster, are fed as seed conformations for the next iteration. Note that the side chains of these seed conformations could have quite different torsion angles after energy minimization.

D. Iteration until All Components Are Assembled

Having determined how two particular helices are assembled with each other , we then bring in each additional helix, one-at-a-time, and repeat the process above. The order of assembly can directly affect the quality of the final assembled structures. The way we choose which helices should start the process at the outset is by finding the neighboring helices that have the shortest connecting linker between them, as the conformational search space associated with a short loop is relatively small. In the same way, for later assembly iterations, the helix that is connected to the partially assembled structure with a shorter loop is chosen.

We have used Rh as a simple metric to determine the 'native-ness' of the assembled structures. While assessing native-ness in beta-sheets may also require a measure of hydrogen bonding, alpha-helical packings are simpler. We simply measure their hydrophobic cores, using the radius of gyration of hydrophobic residues Rh. An alternative measure previously proposed is simply the radius of gyration Rg (Fleming et al., 2006; Narang et al., 2005). We compare Rh to Rg here, to assess their discrimination power. Fig. 4 shows the running average and running minimum of Cα RMSD plotted against the number of top-ranked structures for a 2-helix bundle protein (PDB ID 1RPO). It is clear that Rh is a better discriminator for these helical packings than Rg for selecting near-native conformations. A related recent study (Lin et al., 2007) found that including hydrophobic potential of mean force in the AMBER force field can significantly improve the predictive power of the energy function.

ACKNOWLEDGMENTS

We thank B. Ho, B. Ozkan, M.S. Shell, V. Voelz and J. Chodera for helpful discussions. G.W. acknowledges support from an NIH fellowship. E.C. acknowledges the hospitality of the Dill Lab during several visits to UCSF. This work was also supported by NIH grant GM 34993 and an Opportunity Award from UCSF and the Sandler Family Foundation.

REFERENCES

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J *80*, 505-515.

Bowie, J.U., and Eisenberg, D. (1994). An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. Proc Natl Acad Sci U S A *91*, 4436-4440.

Bratley, P., and Fox, B.L. (1988). ALGORITHM 659: implementing Sobol's quasirandom sequence generator. ACM Transactions on Mathematical Software (TOMS) *14*, 88-100.

Bromberg, S., and Dill, K.A. (1994). Side-chain entropy and packing in proteins. Protein Sci *3*, 997-1009.

Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The Amber biomolecular simulation programs. J Comput Chem *26*, 1668-1688.

Cohen, F.E., Richmond, T.J., and Richards, F.M. (1979). Protein folding: evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. J Mol Biol *132*, 275-288.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. Journal of the American Chemical Society *117*, 5179-5197.

Coutsias, E.A., Seok, C., Jacobson, M.P., and Dill, K.A. (2004). A kinematic view of loop closure. J Comput Chem *25*, 510-528.

Coutsias, E.A., Seok, C., Wester, M.J., and Dill, K.A. (2005). Resultants and Loop Closure. International Journal of Quantum Chemistry *106*, 176-189.

Crick, F. (1953). The packing of [alpha]-helices: simple coiled-coils. Acta Crystallographica *6*, 689-697.

Cuff, J.A., and Barton, G.J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins *34*, 508-519.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System. (DeLano Scientific, San Carlos, CA).

Dodd, L.R., Boone, T.D., and Theodorou, D.N. (1993). A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. Molecular physics(Print) *78*, 961-996.

Dunbrack, R.L., Jr. (2002). Rotamer libraries in the 21st century. Current opinion in structural biology *12*, 431-440.

Dunbrack, R.L., Jr., and Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol *230*, 543-574.

Fain, B., and Levitt, M. (2001). A novel method for sampling alpha-helical protein backbones. J Mol Biol *305*, 191-201.

Fain, B., and Levitt, M. (2003). Funnel sculpting for in silico assembly of secondary structure elements of proteins. Proc Natl Acad Sci U S A *100*, 10700-10705.

Fleming, P.J., Gong, H., and Rose, G.D. (2006). Secondary structure determines protein topology. Protein Sci *15*, 1829-1834.

Go¯, N., and Scheraga, H.A. (1970). Ring Closure and Local Conformational Deformations of Chain Molecules. Macromolecules *3*, 178-187.

Ho, B.K., and Dill, K.A. (2006). Folding very short peptides using molecular dynamics. PLoS computational biology *2*, e27.
Hoang, T.X., Seno, F., Banavar, J.R., Cieplak, M., and Maritan, A. (2003). Assembly of protein tertiary structures from secondary structures using optimized potentials. Proteins *52*, 155-165.
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.
Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol *292*, 195-202.
Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577-2637.
Kohn, W.D., Mant, C.T., and Hodges, R.S. (1997). Alpha-helical protein assembly motifs. J Biol Chem *272*, 2583-2586.
Kolodny, R., and Levitt, M. (2003). Protein decoy assembly using short fragments under geometric constraints. Biopolymers *68*, 278-285.
Lee, H., and Liang, C. (1988). Displacement analysis of the general spatial 7-link 7 R mechanism. Mechanism and machine theory *23*, 219-226.
Lin, M.S., Fawzi, N.L., and Head-Gordon, T. (2007). Hydrophobic potential of mean force as a solvation function for protein structure prediction. Structure *15*, 727-740.
Lovell, S.C., Davis, I.W., Arendall 3rd, W.B., de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S., and Richardson, D.C. (2003). Structure validation by Calpha geometry: phi, psi and Cbeta deviation. Proteins *50*, 437-450.
Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science *252*, 1162-1164.
McAllister, S.R., Mickus, B.E., Klepeis, J.L., and Floudas, C.A. (2006). Novel approach for alpha-helical topology prediction in globular proteins: generation of interhelical restraints. Proteins *65*, 930-952.
Miyazawa, S., and Jernigan, R.L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol *256*, 623-644.
Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Current opinion in structural biology *15*, 285-289.
Mumenthaler, C., and Braun, W. (1995). Predicting the helix packing of globular proteins by self-correcting distance geometry. Protein Sci *4*, 863-871.
Nanias, M., Chinchio, M., Pillardy, J., Ripoll, D.R., and Scheraga, H.A. (2003). Packing helices in proteins by global optimization of a potential energy function. Proc Natl Acad Sci U S A *100*, 1706-1710.
Narang, P., Bhushan, K., Bose, S., and Jayaram, B. (2005). A computational pathway for bracketing native-like structures for small alpha helical globular proteins. Phys Chem Chem Phys *7*, 2364 - 2375.
Onufriev, A., Bashford, D., and Case, D.A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins *55*, 383-394.
Ozkan, S.B., Wu, G.A., Chodera, J.D., and Dill, K.A. (2007). Protein folding by zipping and assembly. Proc Natl Acad Sci U S A *104*, 11987-11992.
Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. J Mol Biol *7*, 95-99.
Rost, B., Yachdav, G., and Liu, J. (2004). The PredictProtein server. Nucleic Acids Res *32*, W321-326.

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.

Wedemeyer, W.J., and Scheraga, H.A. (1999). Exact analytical loop closure in proteins using polynomial equations. Journal of Computational Chemistry *20*, 819-844.

Wolf, E., Kim, P.S., and Berger, B. (1997). MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci *6*, 1179-1189.

Yue, K., and Dill, K.A. (2000). Constraint-based assembly of tertiary protein structures from secondary structure elements. Protein Sci *9*, 1935-1946.

Zhang, C., Hou, J., and Kim, S.H. (2002). Fold prediction of helical proteins using torsion angle dynamics and predicted restraints. Proc Natl Acad Sci U S A *99*, 3581-3585.

Zhang, J., and Liu, J.S. (2006). On side-chain conformational entropy of proteins. PLoS computational biology *2*, e168.

FIGURE LEGENDS

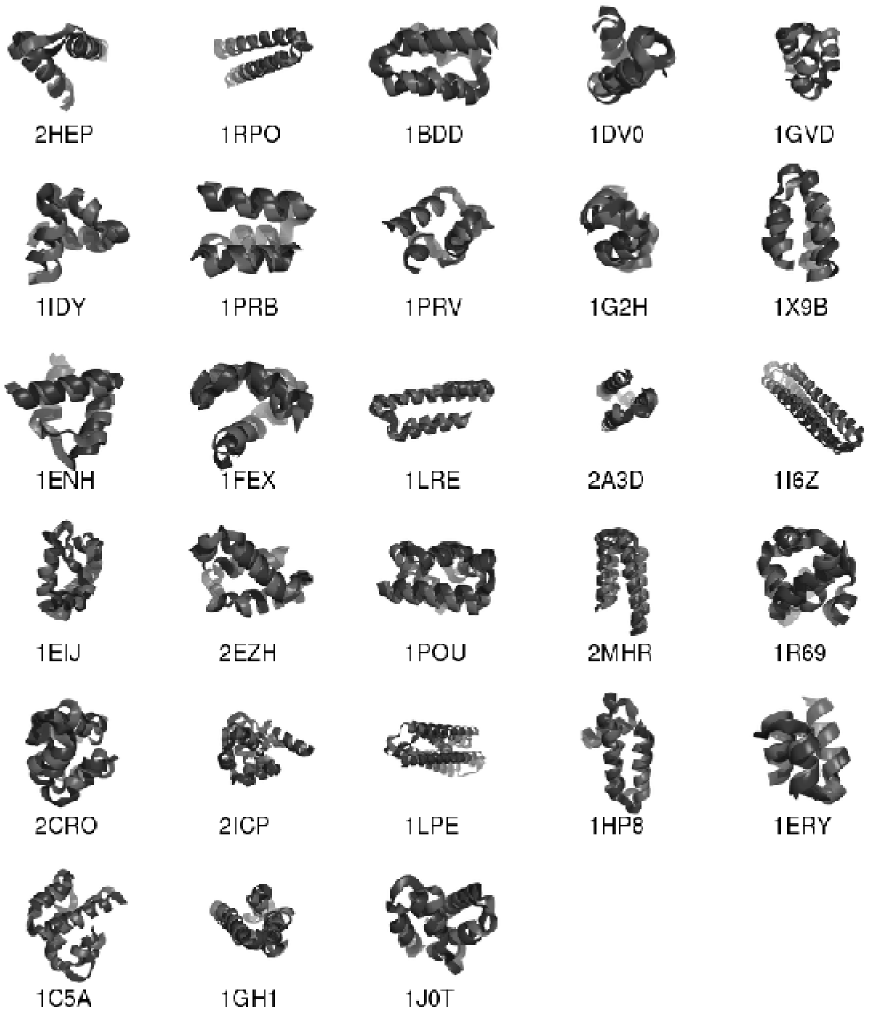

Fig 1. Cartoon representation of the native (red) versus the lowest rmsd structures assembled (blue) for the 28 proteins listed in Table 1. Figure produced with Pymol (DeLano, 2002).

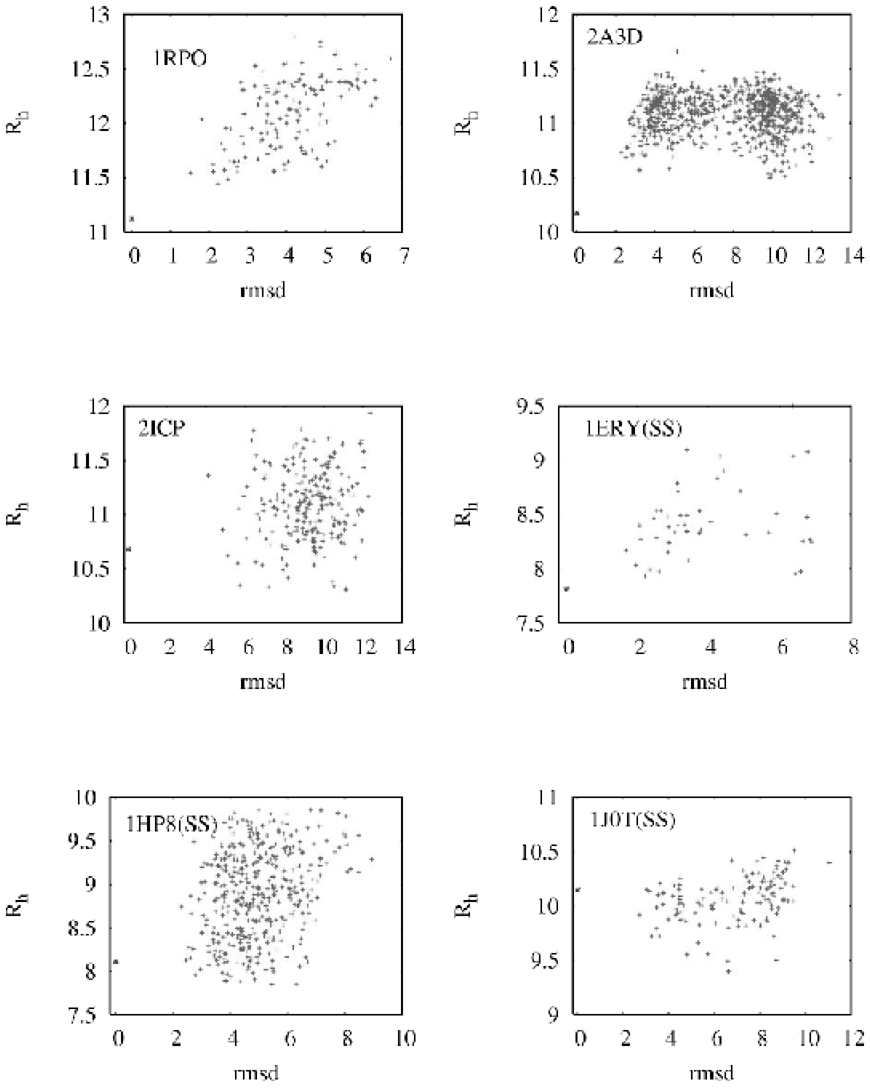

Fig 2. The scoring function Rh vs. rmsd for a 2-helix bundle (1RPO), three 3-helix bundles (2A3D, 1ERY, 1HP8), a 4-helix bundle (1J0T), and a 5-helix bundle (2ICP). Three proteins contain disulfide bridges (1ERY, 1HP8, 1J0T), and SS-bond restraints have been imposed during conformational sampling. Red dots are sampled conformations; blue dot denotes the native conformation at rmsd zero. Note that for most cases, Rh of the native is among the smallest of all sampled conformations, with the exception of certain proteins containing disulfide bonds. Both Rh and rmsd are in units of Angstroms.

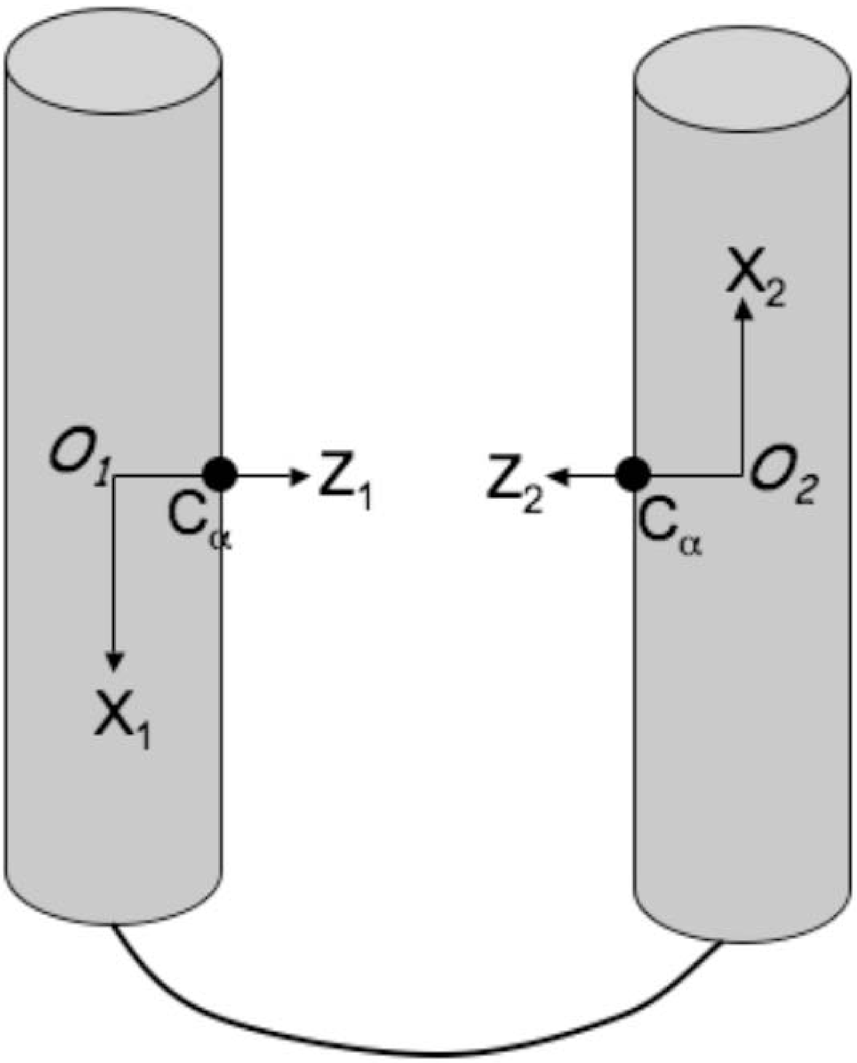

Fig 3. Starting point for MATCHSTIX: the Cα carbons of two hydrophobic residues are placed 10 Å apart, facing each other. The cylinders are aligned, and coordinate axes are defined from this configuration. The cylinders are then translated and rotated rigidly and randomly. This procedure is then performed for every possible different hydrophobic pairing.

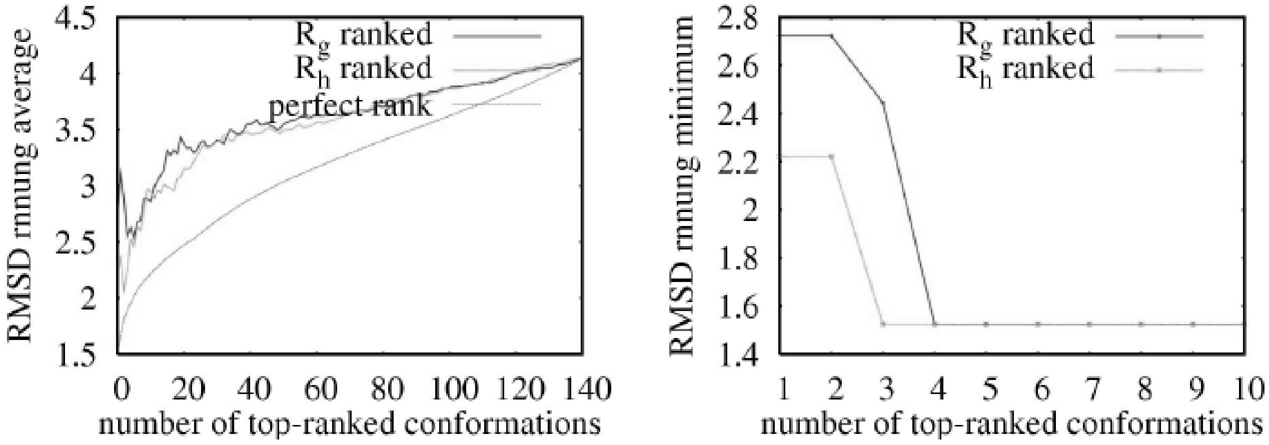

Fig 4. Comparison between Rg (radius of gyration) and Rh (radius of gyration of all the atoms of hydrophobic, helical residues) for a set of 140 simulated compact structures for a 2 helix bundle protein (PDB ID 1RPO). The left figure shows the RMSD running average, which is a measure of the overall native-likeness of top-ranked conformations. The right figure shows the RMSD running minimum or the lowest RMSD in the top-ranked structures. The red and green curves correspond to the Rg and Rh metrics respectively, whereas the blue curve corresponds to a hypothetical, perfect metric by which the conformations rank in ascending order of their rmsd relative to native. It is seen that the Rh metric is closer to the perfect metric than Rg especially in the top-ranked conformations. RMSDs are in Angstroms.

Table 1. Assembly Results

PDB	Chain	#	Assembly	Lowest RMSD (Å)				
code	length	helices	order	Top 1	Top 5	Top 20	Top 50	All
2HEP	36	2	1-2	2.52	1.41(3)	1.41(3)	1.27(32)	1.27(32)
1RPO	56	2	1-2	2.22	1.52(3)	1.52(3)	1.52(3)	1.52(3)
1BDD	47	3	2-3-1	3.58	2.84(2)	2.58(9)	1.58(25)	1.58(25)
1DV0	32	3	1-2-3	4.55	3.77(3)	1.92(19)	1.92(19)	1.89(51)
1GVD	40	3	1-2-3	1.53	1.53(1)	1.51(13)	1.51(13)	1.51(13)
1IDY	39	3	1-2-3	6.61	3.70(5)	1.93(6)	1.93(6)	1.81(67)
1PRB	42	3	2-3-1	1.50	1.50(1)	1.50(1)	1.50(1)	1.50(1)
1PRV	38	3	1-2-3	4.96	3.54(5)	2.03(13)	2.03(13)	2.03(13)
1G2H	32	3	1-2-3	1.80	1.80(1)	1.80(1)	1.80(1)	1.66(53)
1X9B	45	3	1-2-3	8.60	2.21(2)	2.21(2)	2.21(2)	2.21(2)
1ENH	46	3	2-3-1	4.15	4.15(1)	3.97(8)	1.92(34)	1.92(34)
1FEX	50	3	2-3-1	6.07	6.07(1)	2.67(10)	2.67(10)	2.67(10)
1LRE	66	3	1-2-3	8.93	4.16(2)	2.96(19)	2.96(19)	2.96(19)
2A3D	69	3	2-3-1	9.86	9.75(4)	2.50(17)	2.29(30)	2.29(30)
1I6Z	112	3	1-2-3	5.93	5.93(1)	2.91(8)	2.91(8)	2.91(8)
1EIJ	59	4	1-2-3-4	6.98	5.09(2)	5.09(2)	5.09(2)	3.82(154)
2EZH	59	4	1-2-3-4	7.35	4.07(5)	3.56(10)	3.21(47)	2.42(143)
1POU	69	4	1-2-3-4	11.3	9.28(3)	5.87(18)	4.22(42)	3.75(261)
2MHR	91	4	1-2-3-4	8.20	3.14(5)	2.14(8)	2.14(8)	2.14(8)
1R69	60	5	2-3-1-4-5	6.70	5.24(2)	5.24(2)	4.0(43)	3.54(85)
2CRO	60	5	1-2-3-4-5	9.65	7.42(4)	4.01(12)	4.01(12)	4.01(12)
2ICP	72	5	1-2-3-4-5	11.1	5.73(4)	5.10(17)	5.10(17)	4.09(185)
1LPE	138	5	1-2-3-4-5	5.22	5.22(1)	5.22(1)	4.59(23)	4.54(74)
1HP8	54[3]	3	1-2-3	5.44	3.77(5)	2.77(8)	2.45(30)	2.29(149)
1ERY	32[2]	3	2-3-1	2.21	2.21(1)	1.69(9)	1.69(9)	1.69(9)
1C5A	63[3]	4	4-3-2-1	6.72	3.79(3)	2.49(17)	2.49(17)	2.04(103)
1GH1	69[2]	4	2-3-4-1	7.57	3.21(2)	3.21(2)	3.21(2)	3.13(96)
1J0T	58[2]	4	1-2-3-4	6.64	4.83(4)	3.24(8)	2.73(28)	2.73(28)

The last 5 columns list the lowest RMSD structures and their Rh-ranking (in parenthesis) among the top 1, 5, 20, 50, and all the sampled conformations. The 2nd column lists chain lengths excluding termini non-helical residues, with the number of disulfide bridges in square brackets. For the 4th column, each helix is numbered by its relative position to the N terminal.

Table 2. Assembly Performance Comparisons

Proteins			Lowest RMSD (Å)		
PDB code	Chain length	# helices	Present method	torsion sampling	Ca model
1BDD	47	3	1.58	4.21	
1GVD	40	3	1.51	4.89	
1DV0	32	3	1.92	4.74	
1HP8	54	3	2.45	4.20	
1IDY	39	3	1.93	3.36	
1PRV	38	3	2.03	3.87	
2EZH	59	4	3.21	4.40	
1PRB	42	3	1.50	4.08	2.9
1G2H	32	3	1.80		3.4
1FEX	50	3	2.67		3.4
1LRE	66	3	2.96		3.4
1I6Z	112	3	2.91		2.5
1EIJ	59	4	5.09		4.6
1LPE	138	5	4.59		3.4

Performance comparisons among the present assembly method, the loop torsion sampling method (Narang et al., 2005), and a coarse-grained model (Nanias et al., 2003). The last three columns list the lowest rmsd relative to the native among the top 50, 100, and 50 structures respectively.

Table 3. Effect of Disulfide Bond Restraints

PDB code	Chain length	# helices	Assembly order	Lowest RMSD (Å)	
				SS restraint	no restraint
1HP8	54	3	1-2-3	2.29(149)	3.29(335)
1ERY	32	3	2-3-1	1.69(9)	2.04(10)
1C5A	63	4	4-3-2-1	2.04(103)	2.68(389)
1GH1	69	4	2-3-4-1	3.13(96)	3.90(94)
1J0T	58	4	1-2-3-4	2.73(28)	3.83(20)

Effect of SS bond restraints on best assembled structures. The R_h-rankings of the lowest rmsd structures are in parenthesis.