Protein Science (2000), *9*:1935–1946. Cambridge University Press. Printed in the USA.

Constraint-based assembly of tertiary protein structures from secondary structure elements

KAIZHI YUE and KEN A. DILL
Department of Pharmaceutical Chemistry, University of California at San Francisco, Box 1204, San Francisco, California 94143

(Received May 9, 2000; Final Revision July 31, 2000; Accepted July 31, 2000)

Abstract

A challenge in computational protein folding is to assemble secondary structure elements—helices and strands—into well-packed tertiary structures. Particularly difficult is the formation of β-sheets from strands, because they involve large conformational searches at the same time as precise packing and hydrogen bonding. Here we describe a method, called Geocore-2, that (1) grows chains one monomer or secondary structure at a time, then (2) disconnects the loops and performs a fast rigid-body docking step to achieve canonical packings, then (3) in the case of intrasheet strand packing, adjusts the side-chain rotamers; and finally (4) reattaches loops. Computational efficiency is enhanced by using a branch-and-bound search in which pruning rules aim to achieve a hydrophobic core and satisfactory hydrogen bonding patterns. We show that the pruning rules reduce computational time by 10^3- to 10^5-fold, and that this strategy is computationally practical at least for molecules up to about 100 amino acids long.

Keywords: conformational search; protein folding; structure prediction; tertiary structure assembly

Computer algorithms are beginning to succeed at predicting the topologies of proteins in blind tests (Bates et al., 1997; Simons et al., 1997; casp3, 1998; Ortiz et al., 1998). It is of interest to continue refining and improving such folding models until they reach sufficiently high resolution to help in ligand design.

A popular approach to computational protein structure prediction has been to divide the problem into two parts: first, parse the sequence into appropriate secondary structure elements, then assemble those elements into a tertiary fold (Cohen et al., 1981; Harris et al., 1994). Recently, there has been a significant advance in the first part of this strategy. Whereas past prediction methods identified secondary structures one residue at a time, Srinivasan and Rose (1995) and Baker and his colleagues (Simons et al., 1997) have recognized that whole peptides, typically 4–15 amino acids long, often have a relatively small ensemble of conformers. This insight has led to a substantial reduction of conformational searching in their folding models, and it provides considerable justification for the strategy noted above, of parsing the sequence into relatively rigid conformational elements, and then trying to assemble them into tertiary structures.

The step that remains challenging is to assemble secondary structures into tertiary structures at high resolution, particularly in β-sheet proteins. Docking large rigid elements together causes either severe steric clashes or loose packing, or both, whereas real native structures involve snug fits and a high degree of steric complementarity. Real proteins are pliable and accommodating. The problem with β-sheet proteins is that interstrand contacts can be nonlocal, often involving a large number of intervening degrees of freedom along the chain, and yet the hydrogen bonds and side chains must line up with a high degree of precision. In terms of energy landscapes (Dill & Chan, 1997), β-structures are represented by deep narrow holes surrounded by large flat plains, in contrast to the large bowls that characterize local structures, such as turns, helices, and strands.

To predict tertiary folds, a key strategy has been to identify a canonical set of secondary structural elements as building blocks and to try to assemble them into a finite number of canonical tertiary packings. For example, a large literature describes the "knobs into holes" and "ridge into groove" models of helical packing (Bowie, 1997, and references therein). Also extensively explored are the twists of β-strands, the twists in intrasheet strand packing (Chothia & Janin, 1981; Chothia, 1984; Chothia & Finkelstein, 1990); tight turns (Richardson, 1981; Richardson & Richardson, 1989), helical packing in 4-helix bundle proteins (Harris et al., 1994), coiling of β-hairpins, packing of α-helices onto β-pleated sheets (Janin & Chothia, 1980), packing of α/β barrels (Lesk et al., 1989) and of β-sandwiches (Cohen et al., 1981), residue pairing in antiparallel strand packing (Hutchinson et al., 1998), $\beta\alpha\beta$ turns (Edwards et al., 1987), and side-chain organization in type I tight turns (Richardson, 1981). Statistics have been collected on packing angles and distances (Sklenar et al., 1989; Brown, 1992; Reddy & Blundell, 1993; Kurochkina & Privalov, 1998). Comprehensive categorization algorithms exist, such as CATH (Michie et al., 1996). There is a rich foundation for assem-

Reprint requests to: Ken A. Dill, Department of Pharmaceutical Chemistry, University of California at San Francisco, Box 1204, San Francisco, California 94143; e-mail: dill@maxwell.ucsf.edu.

bling tertiary structures from secondary structures in canonical ways.

Our work here is also based on assembling canonical secondary structure elements into canonical tertiary structures. Our approach follows most closely that of Simons et al. (1997); they try different orientations of fragments/secondary structures to achieve optimal packing. However, while Simons et al. accomplish this by interchanging different fragments randomly, we explicitly calculate the ϕ/ψ angles for the connecting loops and systematically try all the possible packing positions. But despite our extensive search of packings, we are able to treat subtle side-chain variations because of the efficiency achieved by pruning of the conformational search tree using rules of the types described below.

First, orientational restrictions apply to the packing of β-strands into sheets. In a helix, hydrogen bonding is internally satisfied, so when a helix packs against another structural element, orientational preferences involve only the side chains and steric interactions. A β-strand can be represented as having four sides, as if it were a square. The "east" and "west" sides are polar, because of the alternation of pairs of amides and carbonyls, while the "north" and "south" sides have the side chains, also alternating. A strand cannot pack its east–west polar sides either against helices or against north–south sides of other strands, without wasting a whole row of hydrogen bonds. Therefore, the only energetically viable option for strand–strand packings is east–east or east–west. These two options define parallel and antiparallel sheets. The exceptions are rare cases of β-strands on the edge of a sheet, where a polar side can pack against a helix or a loop because it can form hydrogen bonds with water. Once the coordinates for one strand are known, it defines the plane of any sheet that can be formed from it. This information helps constrain the placement of other strands.

Second, two adjacent strands cannot be too far out of register with each other, because this will either lead to unsatisfied H-bonding groups, or poor packing if a third strand fills the void.

Third, the north–south direction of the side chains usually defines the vector that points toward the inside and outside of the protein. Because protein cores are hydrophobic, if one group of alternating side chains (say the even-numbered ones) is more hydrophobic than the other group (the odd-numbered ones), then the former are more likely to point toward the inside, provided there are no kinks or bulges. This directional information can also be used to prune a conformational search tree. These are illustrations of pruning rules that can limit conformational searching in tertiary assembly algorithms.

The algorithm: Geocore 2

We have previously developed an ab initio algorithm called Geocore that attempts to find the native structures of peptides (Yue & Dill, 1996; Ishikawa et al., 1999). Geocore generates protein conformations by a chain growth process, i.e., by adding one amino acid at a time, each in one of a small number of discrete possible conformers, depending on the amino acid type. Conformations are discriminated by a simple physical potential function, rather than by a database-derived potential. Geocore uses a branch-and-bound conformational search method that is sufficiently comprehensive to ensure reaching the globally optimal conformation, unlike trajectory-based methods, like Monte Carlo, which have no such guarantees. Tests show that Geocore captures native-like aspects of the 18 peptides on which it has been tested (Ishikawa et al., 1999). In the present paper, we refer to that algorithm as Geocore-1 (G1).

In Geocore-2 (G2), described here, we add a method to assemble secondary structural elements into tertiary structures. We assume that some pre-processing algorithm has parsed the sequence into secondary structure elements. In the present paper, we focus only on the assembly step, and not on the parsing of the sequence into secondary structural elements. We describe how G2 searches the possible packings of those elements.

The same energy function is used in G2 as in G1. The main difference lies in the conformational generation and refinement scheme. G1 adds one monomer at a time to a growing chain conformation. G2 can also add one secondary structure at a time. G2 treats each secondary structure element as a rigid body that is docked using the six translational and rotational degrees of freedom to the existing structures. The conformational docking is made efficient by pruning constraints, such as those described above. When a helix or strand is added to the growing chain, the algorithm keeps track of the location of its H-bonding groups and hydrophobic residues and the algorithm keeps track of other secondary structural elements that have already been fixed in space and that are located nearby. To facilitate its docking attempts, we use canonical secondary structure elements and canonical dockings, taken from the Protein Data Bank (PDB), so the method has a discrete set of architectures it explores at first. Docking attempts are guided by a branch-and-bound search that, as with G1, is comprehensive in retaining the lowest energy structures and discarding the poor ones.

Methods

Canonical structure elements

Choosing a canonical structure to represent an α-helix is straightforward. An α-helix is a well-defined conformation that can be modeled by a single small region of ϕ/ψ space. But β-strands are more challenging. β-Strands come in about four varieties. They can have low or high twist and be straight or curled, depending on the patterns of repetition of the ϕ/ψ angles (Chothia, 1983; Richardson & Richardson, 1989; Michie et al., 1996). The different twists and curlings of strands cause the hydrophobic and polar side chains to have different spatial distributions.

Here we model only a single type of β-strand. Our model strand is a straight conformation having low twist, defined by $(\phi,\psi) = (-120,130)$. The twist is $\sim 11°$. We chose this model strand because it has sufficient symmetry that two such strands can hydrogen bond to form long β-sheets. But there are tradeoffs. H-bonds are better and more uniform along the chain in curled strands than in straight strands, where H-bonding deteriorates along the chain. But a straight canonical strand can have identical H-bonding geometric parameters for both east and west polar groups, and provide for a better allowance for third and subsequent β-strands to form a sheet. The slight twist in our canonical strand leads to good hydrophobic contacts among the side chains. For this model of a strand, we calculated the intrastrand separation and packing angle that will give maximal hydrogen bonding; this defines the canonical tertiary packing for two strands. The standard packing parameters for two strands in a sheet are shown in Table 1.

Geocore-1: The underlying model

Geocore-1 is described elsewhere (Yue & Dill, 1996); here we just give an overview. Amino acids are represented at the united-atom

Table 1. *Packing angles and distances for standardized packing*[a]

Packing type	Packing angles	Average distances
Helix/helix	−160, 20, 50, −130, 90, −90	12.64
Intersheet strand/strand	0, 170*	12
Intrasheet strand/strand	−10	9.68
Strand/helix	90, 70, 110, −90*	12.3

[a] For all the packing types, the distance between two secondary structures is the close approach distance between the two cylinders and computed based on the actual average side chain sizes on the interface. The entries in the table only show average distances for the categories. In helix–helix packing, each packing pair would assume one of the listed packing angles. In intersheet strand or strand–helix packing, when the two secondary structures are not near parallel or antiparallel, the dihedral can be whatever that is from the original packing position, which is indicated by "*" in the table entries. There is no tilt between two packing structures, i.e., the tilt angle is zero. There is a shift between two strands approximately along the direction of their axes. This is determined by the need of optimal H-bonding in the case of intrasheet packing or by the "knob" and "hole" match in the case of intersheet packing.

level. Backbone conformations are represented by discrete sets of dihedral angles (ϕ/ψ). Standard values are used for bond lengths, bond angles, and other geometric parameters. Each atom is a hard sphere with its appropriate van der Waals (vdW) radius. Because vdW radii are larger than those implied by the minimum contact distances observed in proteins (Cantor & Schimmel, 1980), we soften the potential using a steric violation allowance. The steric allowance affects the compactness of the conformations and the number of conformations that are retained as viable. The selection of values of ϕ/ψ angles for each residue in the molecule determines which region of the conformational space is most thoroughly searched. (Although the options are discretized, subtle adjustments are allowed at later stages of the search.) The user can specify the value of steric allowances and the values of ϕ/ψ angles. The default ϕ/ψ angle preferences of the different amino acids are extracted from the PDB. For loops that interconnect secondary structures, G2 uses a continuum of ϕ/ψ angles for a stretch of three or four residues, whereas the secondary structure elements themselves—β-strands and α-helices—have the fixed standard geometric parameters noted above.

The energy function in Geocore includes hydrophobic and hydrogen-bonding interactions. Each hydrophobic contact of two united atoms with each other is favored by −0.7 kcal/mol. Our energy function does not count hydrogen bonds. Rather, because polar groups can hydrogen bond either to water or other polar groups in the protein, Geocore assigns an energy penalty to the burial of carbonyl or amide groups in the core that are not hydrogen bonded. Each polar group that is buried but not H-bonded has an energy penalty of 1.5 kcal/mol.

Geocore "grows" conformations by adding one residue at a time. By adding residues with different allowed ϕ/ψ angles, Geocore exhaustively considers all the conformations, but not in a brute force way. Instead, a branch-and-bound method is used that guarantees that all globally optimal and near-globally optimal conformations will be found, while neglecting less important conformations. The search is done in depth-first order (Aho et al., 1974). On the search tree, the nodes represent each added amino acid and the different branches are the ϕ/ψ choices. When all the monomers are added or a dead end is reached, the search backtracks. Geocore performs a complete search, subject to the two constraints that no steric overlap is permitted, and that the chain must be compact enough to lead to at least a near maximal number of nonpolar contacts. Geocore gives the user the option to specify if there is a bound on the shape of the conformation and, if so, the dimensions of that shape.

Conformational searching and structure assembly

G2 begins with an amino acid sequence that has already been parsed into *chain segments*, local runs of amino acids that have a given conformation, such as a canonical helix or strand. Figure 1 shows the steps for assembling them into a tertiary structure. We call any particular parsing of the sequence into segments a *chain segment assignment*.

1. For a given chain segment assignment, the first step is to make an estimate of the optimal shape of the hydrophobic core, following a strategy developed in lattice models (Yue & Dill, 1993). The purpose is to establish a lower bound on the possible energy (best possible fold) that could be achieved by the chain. This bound estimate is then used to assess growing conformations to discard them at the earliest possible step, if they are not going to be viable.

2. The chain is then grown either by adding one residue at a time in loop regions, or a whole strand or helix at a time as a rigid body, each in its canonical conformation. The initial position and orientation of a rigid body is defined by six parameters—three translational coordinates and three rotation angles. The ϕ/ψ angles within the segment are fixed, either because a segment is in its canonical conformation or because the segment has been already assigned spatial coordinates in preceding growth steps and is therefore fixed. These initial positions and orientations of helices and strands will generally be poor, because no account has yet been taken of good packing at this stage.

   To achieve better packing, each new strand or helix is treated as a cylinder of appropriate proportions (see below), which is translated and rotated rigidly to come into a better packing configuration with its potential packing partner strand or helix. Having identified whether the potential packing partner is a helix or strand, canonical packing values are chosen appropriately. At this adjustment step, the loops are disconnected so that the secondary structures have the opportunity to achieve optimal pairwise packing.

3. Once this packing is determined, G2 now computes appropriate loop conformations, by calculating continuous ϕ/ψs that would cause the loop to connect the end of one segment to the beginning of the next.

Either a satisfactory packing is achieved or a dead end is reached. A dead end can arise from an unresolvable steric conflict or when the energy of a conformation is too suboptimal to satisfy the current branch-and-bound criterion. At this step, G2 has now fixed this secondary structure element, and it then moves to the next element in the chain, which now becomes the working mobile element.

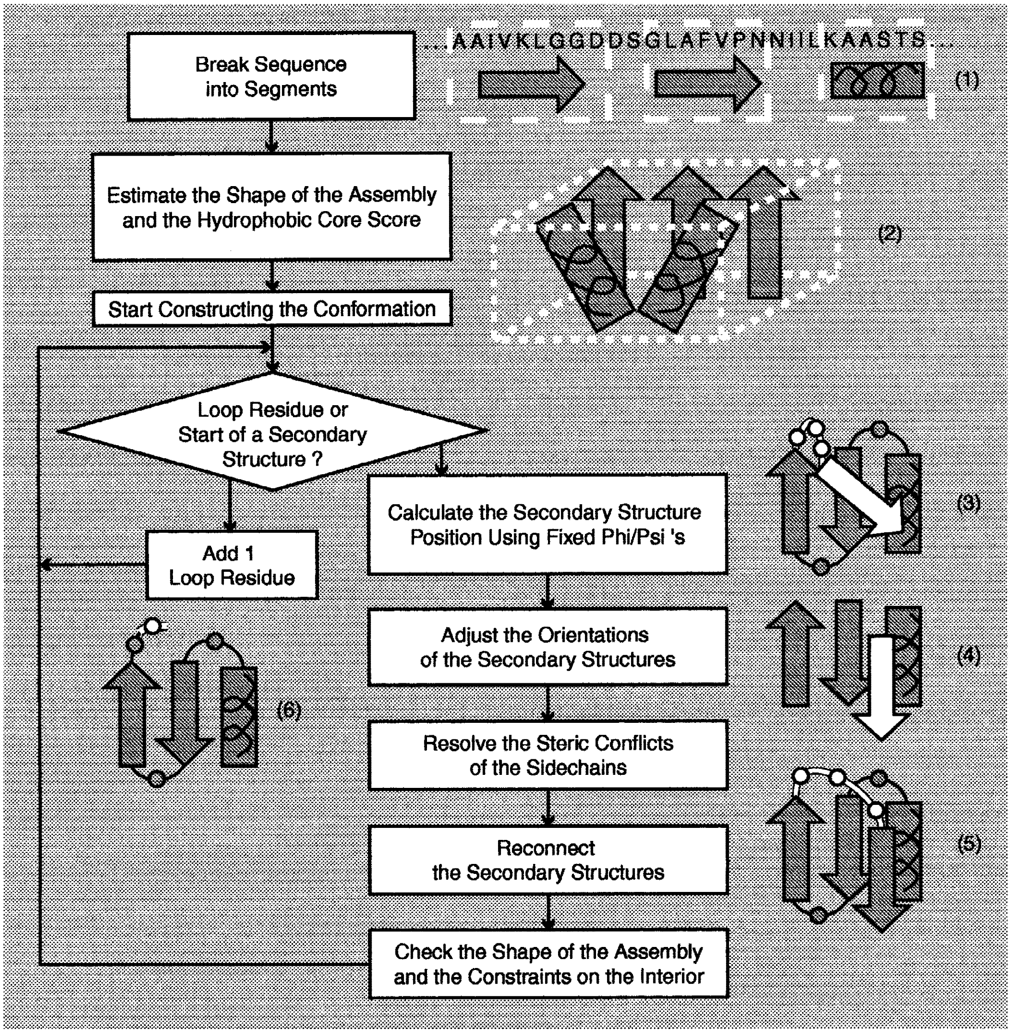
Break Sequence into Segments
...AAIVKLGGDDSGLAFVPNNIILKAASTS...
(1)
Estimate the Shape of the Assembly and the Hydrophobic Core Score
(2)
Start Constructing the Conformation
Loop Residue or Start of a Secondary Structure ?
Add 1 Loop Residue
Calculate the Secondary Structure Position Using Fixed Phi/Psi 's
(3)
Adjust the Orientations of the Secondary Structures
(4)
(6)
Resolve the Steric Conflicts of the Sidechains
Reconnect the Secondary Structures
(5)
Check the Shape of the Assembly and the Constraints on the Interior

Fig. 1. Flowchart of the search procedure based on secondary structure assembly. At step 1, a protein sequence is parsed into alternating secondary structures and loops. At step 2, we estimate the shape and size of the secondary structure assembly, the frame. At step 3, three or four residue look ahead is used to calculate the tentative position of the next helix or strands. The standard packing position for the new helix or strand is calculated/adjusted at step 4. At step 5, a system of equations is solved to find the exact ϕ/ψ angles for the connecting residues. Step 6 shows that if the next residue is on the start of a long loop, add one residue.

Our current procedure has a few limitations. First, we can only handle mid-sized proteins with chain lengths less than about 120 residues. This is due to our current inability to treat structures with more than a single hydrophobic core. A single core means there is only one pocket of buried side chains in which any two side chains are either in direct contact or in contact with other side chains that are in direct contact with each other. For example, a β-sandwich is a single core structure, while a double wound α/β structure has two cores, formed on the two sides of the β-sheet with the flanking helices.

Second, when each segment is added to a growing structure, adjustments are made only on the one mobile secondary structure element, not on the whole molecule. A segment cannot displace other pairing partners already formed. Now we describe each step in more detail.

Finding lowest energy bounds, and estimating optimal packings

For a given parsing of the chain into helical and strand segments, we first estimate an approximate "best possible" packing. Our first estimate is based on nonpolar burial. Because it involves packing secondary structures, hydrogen bonding is included implicitly but approximately. Briefly, each rigid body helix or strand can be regarded as having an inside, based on hydrophobicity, and an outside. (Only in a structure having a single hydrophobic core can we assign a single inside or outside to a helix or strand. In a structure having multiple hydrophobic cores, there are two insides.)

For a strand, it is simple to determine an inside and outside. East–west directions are polar, so only the north and south directions involve side chains that can distinguish the inside from the outside. For helices, it is more difficult. Because there is no hydrogen-bonding requirement for helices with docking partner structures, helices can be rotated through a continuum of angles. With small rotations about the helical axis or slight shifts along the axis, a different pattern of side chains will face the core. For example, inward-pointing side chains may be i, $i+1$, $i+4$, $i+7$, and $i+8$ or, with a slight shift, they may be i, $i+3$, $i+4$, $i+7$, $i+10$, $i+11$.

To estimate an optimal packing energy, we rotate all secondary structure elements (as rigid bodies without atomic detail) to orient their most hydrophobic parts toward the core center, and compute an energy based on standard hydrophobic burial scores (Miyazawa & Jernigan, 1985; Engelman et al., 1986; Roseman, 1988). At this stage, no attempt is made at detailed packing. This gives an estimated bound on the best possible hydrophobic score for the secondary structure assembly. This estimate is sufficient to begin the branch-and-bound process.

Knowing the numbers and sizes and shapes of canonical secondary structures allows us to create a *frame* (Yue & Dill, 1993), which is a minimal box that can contain them. Growing conformations that can fit their secondary structures into the frame will have low energies, whereas conformations that substantially spill over the frame boundaries will ultimately be poor conformations and can be eliminated from further assembly. The following describes the individual docking step that brings a new ("mobile") secondary structure element into a tertiary arrangement with other elements that have already been laid down and fixed in space.

Generating and adjusting positions of helices and strands

When a rigid body helix or strand is added to the current end of the growing chain, its initial position and orientation are determined at first by the preceding fixed residue. From this position and orientation, G2 calculates various properties of this structural element: the direction of the axis of the mobile rigid body, the projections of the starting C_α and ending C_α carbon coordinates on that axis, the direction of the perpendicular from the starting C_α to the axis, and related quantities.

The last few residues in the connecting loop for any strand or helix are considered tentative, subject to adjustment. A "look-ahead" step will calculate the exact ϕ/ψ angles of these residues that can allow the succeeding secondary structures to have optimal packing parameters. For example, if a helix or strand starts at residue j, the previous helix or strand ends at residue i, and $j-i>4$, residues $j-1$, $j-2$, and $j-3$ will not be searched. In calculating the look-ahead, the connecting residues have fixed ϕ/ψs, just as in G1.

When a new helix or strand is added to the growing chain, G2 checks whether there are nearby helices or strands with which the new unit is already in steric conflict, or with which it is close enough to pack. If there is a severe steric conflict, the new unit is rotated or translated to avoid conflict. When no hydrogen bonding can be gained by orientation, the coordinates of the mobile element are adjusted to remove steric conflicts or voids between two neighboring structures. When the conflicts are too severe or the voids too large, implying that no adjustment would succeed, no adjustment is even attempted, and that particular conformation of the mobile element is discarded.

To attempt a docking of a secondary structural element, G2 discretizes the search options by finding an *anchor point*, where a hydrogen bond can be made or where a knob fits into a hole. Anchor points are either the centroid of a side chain or a point equidistant between nearby side-chain centroids. The terms knob and hole are used here in a general way; they apply not only to helices, but also to β-strands, where a protruding carbonyl oxygen (the knob) can fit with an amide nitrogen (the hole).

For example, if two helices are docking, and if side chain i is the knob for a hole near residue j, then i and j will be the *matching residues*. But if instead two knobs come into contact, the anchor position will be shifted to dock a knob with a hole. The same strategy is applied for docking two strands together. If their closest heavy atoms are side chains, intersheet packing is attempted. If instead the closest heavy atoms are main chains, intrasheet packing is attempted. In either case, the anchor point is shifted to achieve appropriate steric fit or hydrogen bonding. The degree of sliding is local and limited, because other parts of the search tree will include attempts at other pairings of the same strands at different positions.

Once it is determined that two secondary structures should attempt to pair, and their anchor points are located, G2 attempts a canonical packing. The appropriate docked separation between the two secondary structures is calculated based on the average sizes of the side chains of both elements. The result is a new set of coordinates for the C_α carbon of the first residue of the mobile helix or strand, a new axis vector for the mobile element, and a new vector that connects the C_α of the first residue with its projection on the new mobile helix or strand axis. The latter two define a new orthogonal coordinate system in space, which we designate M. M is a matrix or a set of orthogonal vectors.

If enough of the protein structure has already been fixed when G2 is laying down a new mobile element—a strand or helix—G2 must decide which fixed secondary structure element should be treated as a possible pairing partner for the mobile element. This choice is made based on priorities. The first priority is proximity. Close partners are attempted, but distant ones are not. The second priority depends on the structure type. An intrasheet strand packing has higher priority than a strand-helix packing, if the distances are about the same, because of the greater possible gain of hydrogen bonding. When a mobile element comes into contact with several packing partners, its position and orientation are computed pairwise with each of the possible docking partners. Then the mobile element is located at the average of those positions.

Once the secondary structures are docked together, G2 then computes torsional angles for loops that could connect them. To find the ϕ/ψ angles for the interconnecting residues between two secondary structure elements, we use the method of Go and Scheraga (Go & Scheraga, 1970; Bruccoleri & Karplus, 1985). We use a three-residue look-ahead: we compute four pairs of ϕ/ψs, therefore, eight total unknowns. The method of Go and Scheraga gives five equations for them. We generate all possible combinations of the values of three variables in the range of $[-\pi,\pi]$ and pass each combination to the Go–Scheraga equation. The solution of Go–Scheraga method reduces the system of equations to a trigonometric equation that is solved by a numerical procedure. The ϕ/ψ angle solutions of the Go–Scheraga equations are then tested for consistency with the allowed Ramachandran values.

If G2 fails to find a set of ϕ/ψ pairs for the interconnecting loop, it attempts another round of adjustment of the paired secondary structure elements with a smaller adjustment, i.e., smaller internal rotations or translations or both. This repeats until a satisfactory set of ϕ/ψs is found, or until all possible rescalings have been exhausted.

If a set of allowable ϕ/ψ pairs is found, then adjustment has succeeded. The ϕ/ψ pairs will be used in calculating the final coordinates of the atoms on the connecting loop. To calculate the coordinates of the atoms for the helix or strand, we compare the original orthogonal system M_0, defined by the original axis vector and the vector that is perpendicular to the axis vector and the helix or strand axis, with M. A rotational transformation can be derived. G2 then applies this transformation to reposition the mobile helix or strand.

At this point, G2 switches from the rigid body level of modelling to the united atom level. For strands in intrasheet packing, side-chain configurations are searched to make sure that no steric conflict exists between side chains on the partner strands and within the mobile strand. If such steric conflicts cannot be removed through side-chain adjustment alone, the mobile helix or strand conformation is discarded.

If all the above succeeds, both the connecting residues and the mobile helix or strand will be checked for steric conflicts, and if successful, the mobile element plus loop is added to the growing chain conformation.

Hydrophobic core-based pruning in conformational searching

During the process of conformational assembly, we use heuristic pruning rules to discard poor conformations at the earliest possible stage. We call them heuristic because they do not guarantee retention of the globally optimal conformation, but these rules impose so little filtering on good conformations that it is very unlikely that they interfere with finding the best conformations (as we show below).

One of our pruning strategies is based on estimating the location of the hydrophobic core in a growing conformation. For example, when the first β-strand is fixed in space, the polar hydrogen bonding plane is already approximately defined. Laying down a second strand then defines an inside and outside of the β-sheet, depending on which side of the east–west plane has more hydrophobic side chains. Adding a third secondary structural element, either a helix or strand, specifies the direction toward the core more definitively. The location of the hydrophobic core becomes increasingly well defined as the growth process proceeds.

Our pruning rule violations are shown in Figure 2. Figure 2A shows a helix sandwiched between two strands on their polar sides, both of which lose the potential for a string of H-bonds. Figure 2B shows a strand that is incorrectly oriented to hydrogen bond to other strands. Figure 2C shows strands that are translationally out of register along the axial direction. If this offset is sufficiently large, the penalty for loss of hydrogen bonding leads to a termination of the conformation. Figure 2D shows a strand that is perpendicular to an existing sheet. This strand cannot ultimately become a part of a sheet because the interior cannot be closed. In Figure 2E, two strands are hydrophobic on opposite sides, which is incommensurate with a hydrophobic core.

Figure 2F shows how G2 prunes the conformational search tree using two frames: the smaller frame contains the secondary structure packing, and the larger frame contains the whole molecule, including loops. If the secondary structures spill out of the smaller frame as the assembly proceeds, G2 prunes those conformations. There are vastly more conformations in which secondary structures spill over the smaller frame, but they are poorer than conformations that fit inside it. The use of the smaller frame for secondary structure assembly provides considerable pruning power.

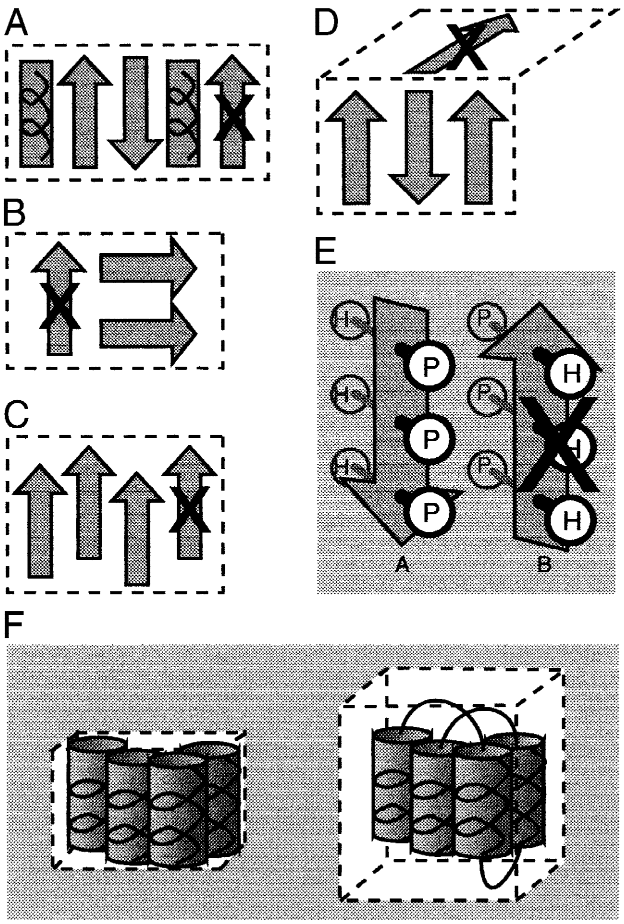

Fig. 2. Pruning situations. **A:** A helix separates two strands in a sheet. **B:** A strand is inside the plane of an existing sheet and is blocked by the fixed strands from forming H-bonds. **C:** The strands within a sheet are substantially offset, limiting the number of hydrogen bonds. **D:** The interior cannot be closed, as when a mobile strand is perpendicular to a fixed sheet. **E:** Hydrophobic faces are incompatible with a hydrophobic core on one side. **F:** Each helix or strand should be within the estimated frame.

Results and discussion

The efficiency enhancements due to pruning

In this section, we show tests of the G2 algorithm. For the five proteins shown in Table 2, having chain lengths from 68 to 118 residues long, we took the known native secondary structure elements as given. G2 then follows the procedures described above: it attempts to assemble all the secondary structures, adjusts the side chains, and inserts and finds appropriate conformations for the intervening loops. Conformational assembly is subject to the pruning rules described above. For each helix or strand, G2 uses four ϕ/ψ choices for three or four residues (i.e., four or five ϕ/ψ pairs) on the preceding connecting loop. Thus, there are 256 or 1,024 possible sets of starting coordinates for each mobile strand or helix. The slow computational step is loop reattachment, which solves a set of equations numerically, for each such set.

Table 2 compares the best G2 structures with the native structures from the PDB. The structures are generated by a systematic

Table 2. *Comparison of best generated structures with native structures*

Protein	Chain length	Native t_{HH}	Native non-H-bonding penalty	Native main-chain H-bonds	Modeled length	t_{HH}	Non-H-bonding penality	Main-chain H-bonds	RMSDs (Å)
7pcy	97	792	7	125	98*	669	9	90	5.87
2mhr	118	1,142	6	175	108	865	11	140	6.18
1ctf	68	500	0	99	99	477	9	74	5.20
1ubq	74	626	3	100	72	603	10	70	6.53
1hoe	74	595	3	87	47	527	11	59	5.9

search in which the conformational space is restricted by the frame size of the hydrophobic cores and the limits on the repulsive terms of Lennard–Jones potentials (i.e., limit on the radii of hard sphere models of atoms). The number of hydrophobic contacts in the native protein is given in the column "native t_{HH}" and the number of such contacts found in the best model conformation is in the column "t_{HH}". In all cases, the model does not create quite as many hydrophobic contacts as in the native protein. Two corresponding columns show that there are also fewer main-chain hydrogen bonds in the model than in the native structure. The overall errors, in terms of C_α root-mean-square deviation (RMSD), range from 5.2 to 6.5 Å.

Table 3 shows the nature of the search and pruning in an arbitrarily chosen subtree, to illustrate the reduction in searching. This comparison cannot be made for the full conformational search because the absence of pruning makes such full searches impossible. Each node on the search tree corresponds to the addition of one monomer or secondary structure in one particular conformation. The same residues can be revisited due to backtracking on the search tree. The table shows the number of nodes visited in this particular subtree. "Secondary structures generated" indicates the total number of secondary structure positions that have been attempted. The "conformations completed" column shows the total number of chain conformations that remain viable when all the residues have been added.

There are two core-checking rows for each protein. The "yes" row indicates the node counts when all the pruning constraints shown in Figure 2 are applied. The "no" row indicates the node counts when the same search and assembly procedure is used, but without the pruning rules. The comparison of these two rows indicates the enhancement due to the pruning rules.

The last two rows of Table 3 (indicated by *) show the subtree that contains the near native structure of 1UBQ. The main point here was to check that the pruning rules do not eliminate good conformations, and that is what these rows show. Table 4 shows the statistics for the full search trees, with pruning constraints applied, to compare to the subtree results shown in Table 3. It is currently impossible to search the full tree without pruning.

Table 5 illustrates the relative pruning power of each of the constraints for one particular subtree for 1UBQ. The numbers for "no core" are lower bounds based on extrapolation from an incomplete run. The actual number could be an order of magnitude higher. Note also that the pruning factors would be greater if the protein is bigger.

At present, for given secondary structures, and using an average of 3.2 ϕ/ψ choices per residue, G2 searches all possible packings in about 240 h for 1UBQ, when the program is compiled in GCC, but not optimized, and running on a Pentium II 450 PC.

Figure 3 shows the distribution of RMSDs of generated structures vs. their conformational energies. All conformations have higher energy than native (-438 kcal/mol).

Figure 4 shows superimposed wire diagrams of folds generated by G2 and the corresponding native structures from the PDB.

Table 3. *Core checking-based pruning on subtrees*

Protein	Chain length	Core checking?	Nodes visited	Secondary structures generated	Conformations completed	Pruning factor
1ubq	72	No	4×10^{12}	2.6×10^{10}	2.8×10^{11}	—
1ubq	72	Yes	9,264,202	2,857,927	1,479	4.3×10^{5}
7pcy	98	No	7.1×10^{8}	1.7×10^{7}	1.1×10^{8}	—
7pcy	98	Yes	781,603	141,599	26	908
1ctf	68	No	7.1×10^{8}	1.1×10^{7}	1.3×10^{6}	—
1ctf	68	Yes	1,097,351	33,557	690	659
1ubq†	72	No	557,010	77,841	66,563	—
1ubq†	72	Yes	1,975	24	0	282
1ubq*	72	No	17,358	2,733	1,935	—
1ubq*	72	Yes	17,318	2,733	1,935	1

Table 4. *Statistics for the conformational search*

Protein	Chain length	Secondary structures generated	Conformations completed	Nodes visited
1ubq	72	91,828,219	5,089	246,853,468
7pcy	98	43,739,440	25,777	157,575,870
2mhr	108	6,525,623	11,616	35,490,029
1ctf	68	3,377,192	4,020	19,304,867

Figure 5 shows some conformations that are pruned at the beginning of the search for 1UBQ. Figure 6 shows some finished conformations for 1UBQ.

G2 assembly gives better tertiary structures than the g1 one-monomer-at-a-time method

Assembling protein structures by docking secondary structural elements leads to much better tertiary structures than growing the chain one monomer at a time, as is done in G1. The problem is that small errors in ϕ/ψ angles add up when configurations are grown using mononer units, giving clumsy tertiary packings.

This is shown in Figure 7 for two proteins—1CTF and 1UBQ. The two energy criteria for determining good final conformations are the numbers of hydrogen bonds and of hydrophobic contacts, t_{HH}. Figure 7A shows the best that can be done by G1, Figure 7B shows the best that is done by G2 without core pruning, and Figure 7C shows G2 with core pruning. The true native structure is indicated by "N."

Table 5. *Pruning results of core packing rules for a run of conformational search for 1UBQ*

Pruning rules	Conformations generated	Secondary structures generated	Nodes visited	Approximate pruning factor
No core	92×10^6	3×10^9	8×10^9	1
Rule (a–c)	40×10^6	15×10^8	4×10^9	2
Rule (d–e)	9.1×10^6	2.9×10^8	8×10^8	10
Rule f (1)	15×10^6	5.2×10^8	10^9	8
Rule f (2)	31×10^6	10^9	2×10^9	4
All rules	954	2,717,103	8,691,476	920

We draw the following conclusions from Figure 7. First, as noted above, pruning does not eliminate the good conformations. Second, G1 does not do nearly as well as G2 at forming hydrophobic contacts or a well-packed core. Because helices form well in all cases, when the secondary structure assignment for residues is not fixed, in G1-generated structures where the secondary structures are most helices, the numbers of hydrogen bonds are not appreciably different. However, their RMSDs will be high. On the other hand, for G1 to get good RMSDs, the H-bonding will deteriorate. For example, for a complete G1 search for 1UBQ, in which all residues assume the native secondary structure assignment, we found minimum RMSD to be 4.79 Å. But its total energy is only around −300, as opposed to the energy of around −450 for one of the best conformations found by G2.

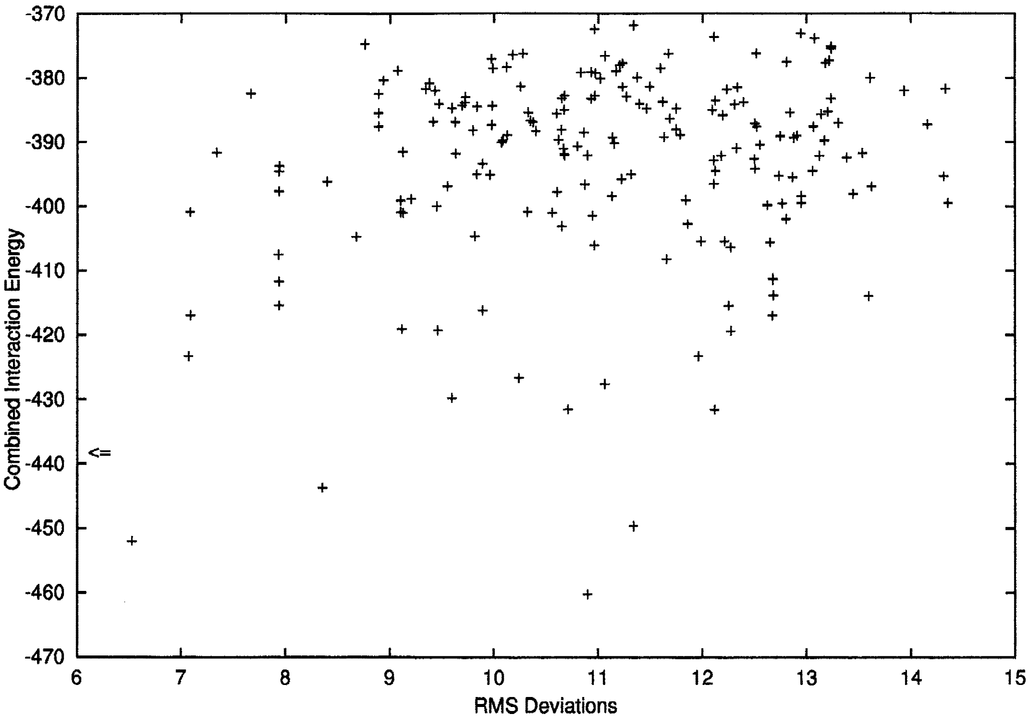

Fig. 3. Energy vs. RMSD for the low-energy conformations of ubiquitin. The *x*-axis is the RMSD and *y*-axis the energy. Each point is a conformation. The native state, which has an energy of −436 kcal/mol, is not shown here.

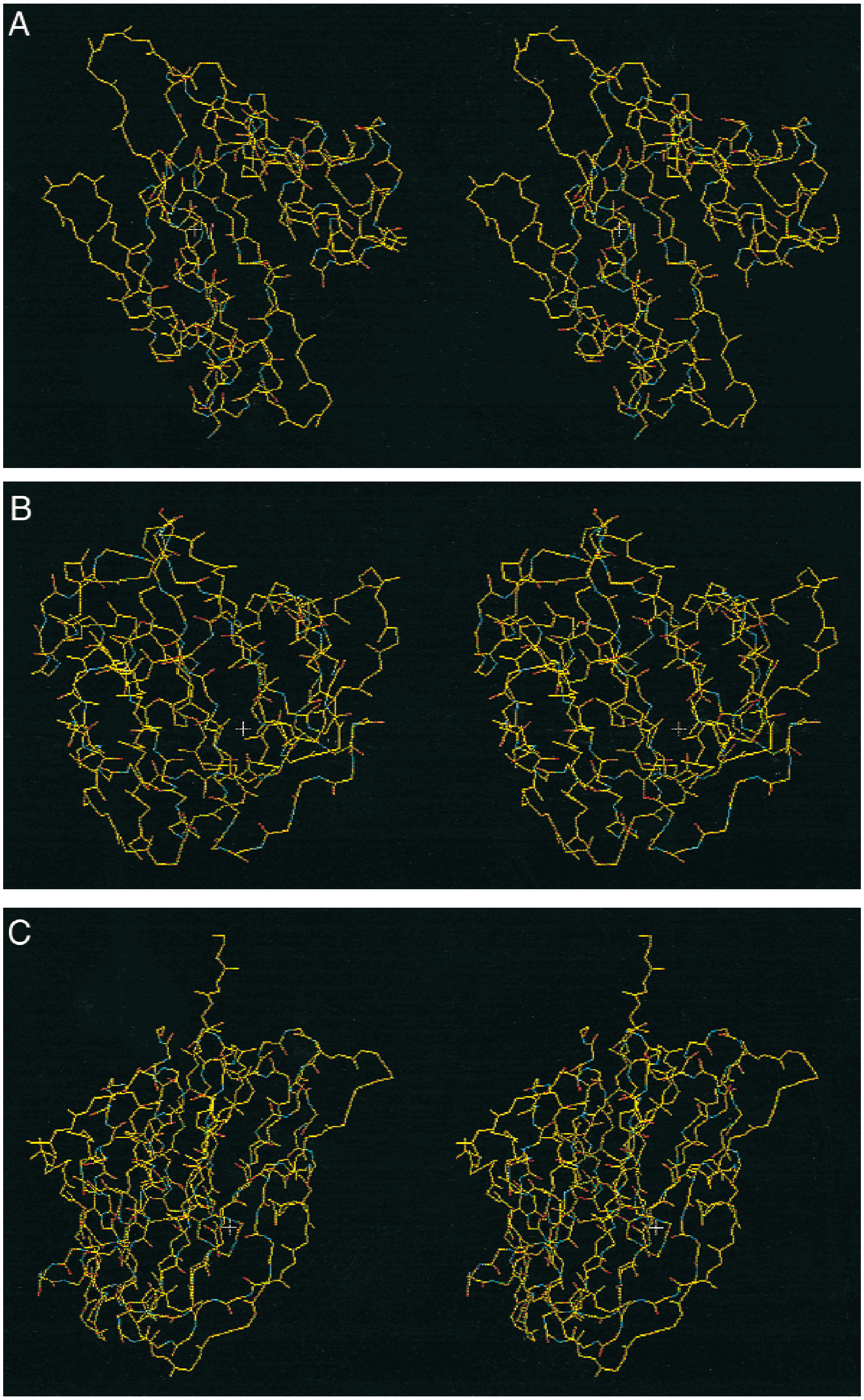

Fig. 4. Superimposed native (in color) and generated (in solid yellow) conformations for 1CTF, 1UBQ, and 7PCY.

Conclusions

We have described a computer algorithm, called Geocore-2, that assembles given helices and strands into low-energy tertiary structures. We show that canonical helices, strands, and packings can be used to give approximate tertiary folds of proteins. The C_α RMSDs range from 5.7 to 6.4 Å for β-proteins having chain lengths between 70 to 98 amino acids. We show that various pruning

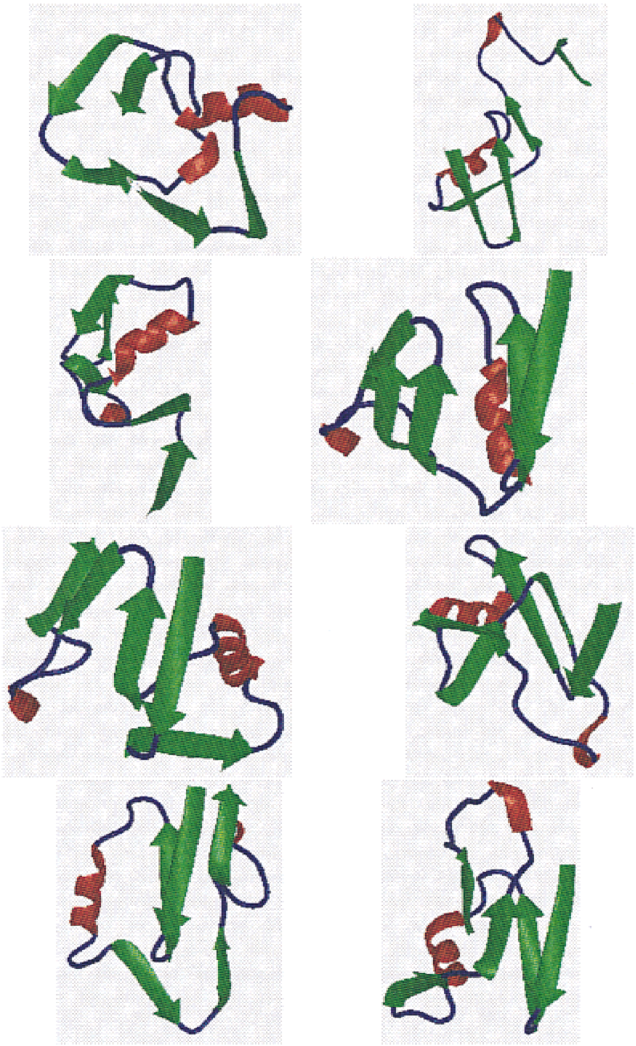

Fig. 5. Pruned 1UBQ conformations. Note that the conformations are shown in a finished form only for the purpose of illustration. The search branch has already been discarded at the third or fourth secondary structure elements.

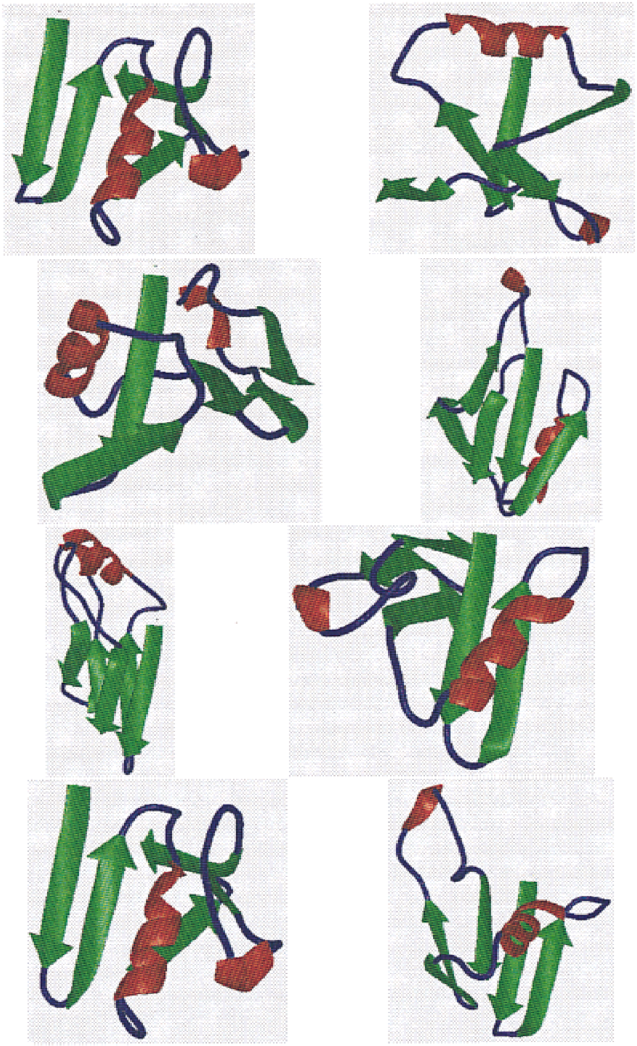

Fig. 6. Completed 1UBQ conformations.

rules can help speed up conformational searching, by factors of 1,000 to 10,000, for proteins the size of ubiquitin. We believe that using pruning rules to guide conformational searches may offer advantages over random trajectory-based search methods for trying to construct native protein folds from given secondary structures.

Acknowledgment

We thank Sarina Bromberg for the help in preparing the illustrations.

References

Aho A, Hopcroft J, Ullman J. 1974. *The design and analysis of computer algorithms*. Reading, MA: Addison-Wesley.

Bates P, Jackson RM, Sternberg MJE. 1997. Model building by comparison: A combination of expert knowledge and computer automation. *Proteins Suppl 1*: 59–67.

Bowie JU. 1997. Helix packing angle preferences. *Nat Struct Biol 4*:915–917.

Brown NP. 1992. Patterns in secondary structure packing—A database for prediction. In: Taylor WR, ed. *Patterns in protein sequence and structure*. Berlin: Springer-Verlag.

Bruccoleri RE, Karplus M. 1985. Chain closure with bond angle variations. *Macromolecules 18*:2767–2773.

Cantor CR, Schimmel PR. 1980. *Biophysical chemistry*. New York: Freeman.

casp3. 1998. *Third meeting on the critical assessment of techniques for protein structure prediction*, Pacific Grove, CA.

Chothia C. 1983. Coiling of β-pleated sheets. *J Mol Biol 163*:107–117.

Chothia C. 1984. Principles that determine the structure of proteins. *Ann Rev Biochem 53*:537–572.

Chothia C, Finkelstein AV. 1990. The classification and origin of protein folding patterns. *Annu Rev Biochem 59*:1007–1039.

Chothia C, Janin J. 1981. Relative orientation of close-packed β-pleated sheets in proteins. *Proc Natl Acad Sci USA 78*:4146–4150.

Cohen FE, Sternberg MJE, Taylor WR. 1981. Analysis of the tertiary structure of protein β-sheet sandwiches. *J Mol Biol 148*:253–272.

Dill K, Chan H. 1997. From levinthal to pathways to funnels. *Nat Struct Biol 4*:10–19.

Edwards M, Sternberg MJE, Thornton JM. 1987. Structural and sequence patterns in the loops of $\beta\alpha\beta$ units. *Protein Eng 1*:173–181.

Engelman DM, Steinz TA, Goldman A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem 15*:321–353.

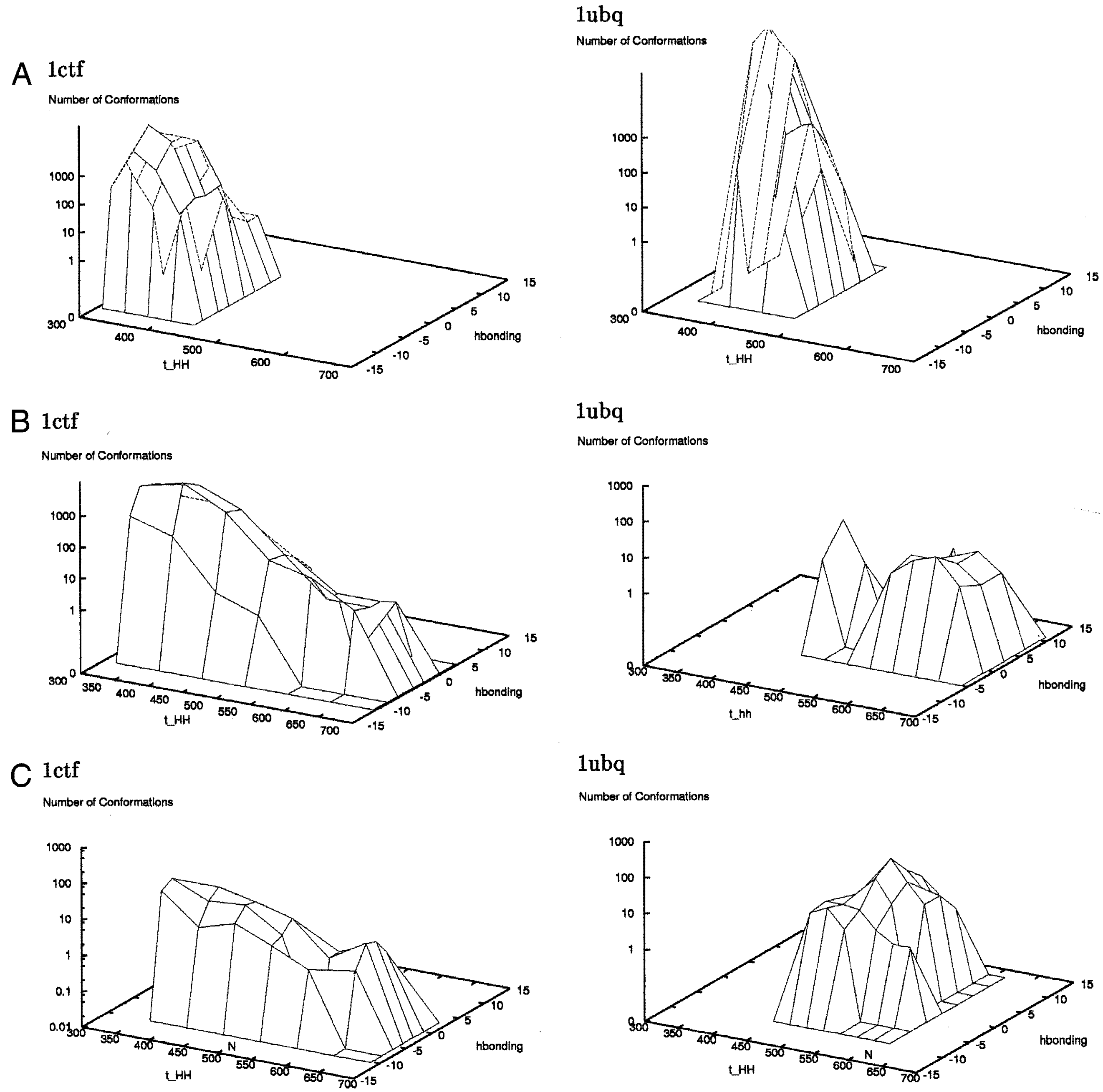

Fig. 7. Energy surfaces for (**A**) G1-generated, (**B**) G2 (without pruning) search-generated, and (**C**) G2 search (with pruning) generated structures.

Go N, Scheraga HA. 1970. Ring closure and local conformational deformations of chain molecules. *Macromolecules 3*:178–186.

Harris NL, Presnell SR, Cohen FE. 1994. Four helix bundle diversity in globular proteins. *J Mol Biol 236*:1356–1368.

Hutchinson EG, Sessions RB, Thornton JM. 1998. Determinants of strand register in antiparallel β-sheets of proteins. *Protein Sci 7*:2287–2300.

Ishikawa K, Yue K, Dill K. 1999. Predicting the structures of 18 peptides using geocore. *Protein Sci 8*:716–721.

Janin J, Chothia C. 1980. Packing of α-helices onto β-pleated sheets and the anatomy of α/β proteins. *J Mol Biol 143*:95–128.

Kurochkina N, Privalov G. 1998. Heterogeneity of packing: Structural approach. *Protein Sci 7*:897–905.

Lesk AM, Branden C, Chothia C. 1989. Structural principles of alpha/beta barrel proteins: The packing of the interior of the sheet. *Proteins 5*:139–148.

Michie AD, Orengo CA, Thorton JM. 1996. Analysis of domain structural class using an automated class assignment protocol. *J Mol Biol 262*:168–185.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecule 18*:534–552.

Ortiz AR, Kolinski A, Skolnick J. 1998. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proc Natl Acad Sci USA 95*:1020–1025.

Reddy BVB, Blundell TL. 1993. Packing of secondary structural elements in proteins. *J Mol Biol 233*:464–479.

Richardson J. 1981. Anatomy and taxonomy of protein structure. *Adv Protein Chem 34*:167–339.

Richardson J, Richardson DC. 1989. Principles and patterns of protein conformation. In: Fasman G, ed. *Prediction of protein structure and the principles of protein conformation*. New York: Plenum Press.

Roseman MA. 1988. Hydrophobicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J Mol Biol 200*:513–522.

Simons KT, Kooperberg C, Huang E, 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.

Sklenar H, Etchebest C, Lavery R. 1989. Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins 6*:46–60.

Srinivasan R, Rose G. 1995. Linus—A hierarchic procedure to predict the fold of a protein. *Proteins 22*:81–99.

Yue K, Dill KA. 1993. Sequence structure relationship of proteins and copolymers. *Phys Rev 48*:2267–2278.

Yue K, Dill KA. 1996. Folding proteins with a simple energy function and extensive conformational searching. *Protein Sci 5*:254–261.