PROTEINS: Structure, Function, and Genetics 48:85–97 (2002)

Distributions of Beta Sheets in Proteins With Application to Structure Prediction

Ingo Ruczinski,[1,2]* Charles Kooperberg,[2] Richard Bonneau,[1] and David Baker[1]
[1]*Department of Biochemistry, University of Washington, Seattle, Washington*
[2]*Fred Hutchinson Cancer Research Center, Seattle, Washington*

ABSTRACT **We recently developed the Rosetta algorithm for ab initio protein structure prediction, which generates protein structures from fragment libraries using simulated annealing. The scoring function in this algorithm favors the assembly of strands into sheets. However, it does not discriminate between different sheet motifs. After generating many structures using Rosetta, we found that the folding algorithm predominantly generates very local structures. We surveyed the distribution of β-sheet motifs with two edge strands (open sheets) in a large set of non-homologous proteins. We investigated how much of that distribution can be accounted for by rules previously published in the literature, and developed a filter and a scoring method that enables us to improve protein structure prediction for β-sheet proteins. Proteins 2002; 48:85–97.** © 2002 Wiley-Liss, Inc.

Key words: beta sheets; Rosetta; structure prediction; protein folding

INTRODUCTION

The ab initio protein folding algorithm Rosetta[1,2] carries out a simulated annealing algorithm to search through the conformation space of three-dimensional structures. From the protein structure database, fragment libraries for three and nine residue segments of the chain are generated, utilizing a sequence profile comparison method. A move set in the annealing algorithm is defined by substituting local segments in the chain with fragments from this library. One part of the scoring function used in the annealing procedure favors the assembly of strands into sheets. However, it only governs how many sheets will be formed given the number of strands, but does not influence how those strands get arranged in the sheets. After generating many decoys using Rosetta, we found that the folding algorithm predominantly generates very local sheets.[3] With the intent to correct the observed biases towards local structures in Rosetta populations, we analyzed the three-dimensional structures in the currently available database of non-homologous proteins (http://www.fccc.edu/research/labs/dunbrack/culledpdb.html), and estimated the distributions of sheet motifs with two edge strands (open sheets) for various sizes in native structures.

Previous studies of β-sheet architecture have involved classification of the β-sheet topologies found in nature and the development of rules and principles that account for the observed distributions. Richardson[4] and Chothia and Finkelstein[5] classify proteins by tertiary structure patterns. Holm et al.,[6], Orengo,[7] Orengo et al,[8], and Murzin et al.[9] classify the proteins into structural families. Some authors investigate structural motifs in a specific subset of proteins (for example, see Orengo and Thornton[10] on α + β folds), or common motifs such as the Greek key.[11] Further, topological features such as the handedness of crossover connections between strands have been described in the literature.[12,13] In addition to the more descriptive studies, some authors analyze folds with β-sheets and report rules that reduce the number of possibilities for the sheet motifs (for example, see refs. 14–17; for β-sandwiches, see refs. 18,19. Only two rules, the absence of knots (no crossing loops) and no left-handed connections,[12,20] significantly reduce the number of possible β-sheet topologies. King et al.[21] use a machine learning approach to automatically generate rules for β-sheets in α/β-domain proteins, compare their rules to the more hand-crafted rules from the above cited literature, and assess the predictive power of their rules. Other authors[15,20] report statistics, such as the frequencies of connection types, or the number of strands per sheet and the number of residues in the strands and their connecting regions, respectively, and describe the patterns that emerge from their analysis. Richardson[15] uses the occurrence frequencies of consecutive connection types to generate pseudo probabilities, which gives a relative ordering among possible β-sheet topologies. We carried out our own analysis since we were interested in modeling the distribution of sheet motifs explicitly, conditioning on other known variables such as loop lengths between strands in the sheets and the proportion of helix residues in the structures. In addition, several hundred structures have been added to the database in recent years, which we use in our analysis. In this article, we describe the derivation and utilization of a new scoring function for β-sheet structures, which incorporates many of the insights of the studies mentioned above in a manner appropriate for evaluating Rosetta models.

Grant sponsor: Packard Foundation; Grant sponsor: NIH; Grant number: 74841.

*Correspondence to: Ingo Ruczinski Johns Hopkins University, Department of Biostatistics, Bloomberg School of Public Health, 615 N Wolfe Street, Baltimore, MD 21205-2179. E-mail: ingo@jhu.edu

Received 28 June 2001; Accepted 29 January 2002

METHODS

Each pair of adjacent residues in a β-strand is represented as a dimer, i.e., we define one landmark per pair of adjacent residues. We define two strands to be neighbors if there is a pair of dimers, one dimer from each strand, within a given distance in space (usually chosen as 5.5 Å). The strands in the sheets are labeled by their numbers in sequence along the backbone, starting with the N-terminus of the protein. A motif is described by the sequence of positions the strands have in the motif, and their directions. For an n-stranded sheet, the position information therefore is a permutation of the numbers 1 through n (sequence). Neighboring strands in the sheet are either parallel or anti-parallel, and we describe this feature (orientation) by a sequence of zeros and ones (up/down). There are two axes of symmetry for sheet motifs, and since it is irrelevant from which angle we look at the protein, four different motifs actually describe the same sheet, of which only one needs to be modeled. To uniquely characterize the sheet, we require the first strands in sequence to be on the left side of the sheet (for sheets with an odd number of strands we require the second strand to be on the left side if the first strand is the center one), and that the first strand points up. For example, Figure 2 shows all possible motifs for three-stranded sheets subject to those requirements.

Modeling the arrangement of strands in sheets in later sections, we assume that the secondary structure is known. Although the arrangement of strands into sheets depends on many characteristics of the protein under consideration, we decided after an initial exploratory data analysis to use only two features of proteins in our model that are given with known secondary structure.

1. α/β vs. all β proteins: We consider a protein to be α/β if at least 20% of its residues are part of a helix. Therefore, the helical status is a binary variable.
2. The lengths of the loops between strands: A loop in this context is the sequence of amino acids that connects the strands under consideration. Therefore, these loops can also contain residues that are in other secondary structures than coils. A loop between two strands is defined as short if the number of residues was ten or less, and long otherwise. For an n-stranded sheet, the loop lengths are, therefore, recorded as $n - 1$ binary variables.

We considered the use of other known properties of proteins in our model, such as the length of the strands in the protein, the protein length (number of residues), and an indicator whether or not there is a helical structure between two strands. We only used the two features described above, since they capture most of the information the other characteristics provide, and because the inclusion of more features was prohibited by the limited amount of data available.

There are $n!$ ways to position the strands in a sheet of size n, and 2^n possibilities for their orientations, ignoring the axes of symmetry. After taking those into account, there are $\frac{1}{4} \times n! \times 2^n = n! \times 2^{n-2}$ possible n-stranded motifs (assuming that all cross-overs between parallel strand pairs are right-handed, and thus not modeling the connections between strands). Modeling the distributions of sheet motifs up to size four without major assumptions and simplifications was feasible since there are only 2 motifs for 2-stranded sheets, 12 motifs for 3-stranded sheets, and 96 motifs for 4-stranded sheets.

RESULTS AND DISCUSSION

We surveyed the distribution of β-sheet motifs with two edge strands (open sheets) in the database of non-homologous proteins. In agreement with previously published investigations, we found that in general pure parallel and pure anti-parallel β-sheet motifs are preferred, as well as motifs with high numbers of sequentially adjacent strands that are spatially adjacent as well. In addition, we found absolute rules that eliminate some motifs from the distribution of possible motifs. In the first part of this section, we show some general results of our survey and discuss a list of deterministic and probabilistic rules for sheet motifs. These rules however are not fully sufficient to describe the distribution of sheet motifs in native proteins. For improved prediction of β-sheet protein structure a scoring function that reflects the entire motif distributions of sheets in native proteins is valuable. We show more detailed results of our survey in the second part of this section, and use those results to model a scoring function for β-sheet motifs. In the last part of this section, we show some applications of our methods for structure prediction of β-sheet proteins.

Some Rules for β-sheet Motifs

Previously described rules reduce the number of possibilities for the sheet motifs.[14–17] Using only two rules, namely no knots/crossing loops and no left-handed connections, the number of possible topologies for the β-sheets can be substantially limited. In addition to those deterministic rules, some authors have pointed out probabilistic rules. For example, Richardson[15] pointed out the preference for pure parallel and pure anti-parallel β-sheets, and Cohen et al.[14] reported the preference of sequentially adjacent strands in the sheet to be spatially adjacent as well. Below, we summarize some of those rules, and illustrate them for four-stranded sheets. Although there are 96 possible motifs for four-stranded sheets in theory, we observed only 48 of those in the database. Among those, 17 motifs were observed only once or twice. We saw a total of 872 four-stranded sheets in the database, but only 19 different motifs were observed eight times or more. In Table I, motifs are classified into three groups based on their frequency of occurence in our database: frequent (eight or more times), rare (at least one but less than eight times), or never. Figure 1 shows the motifs that we did not observe at all in the database, and Figure 3 shows the most likely four-stranded motifs, for both α/β and all β proteins and various loop lengths between the strands.

Table I shows the extent to which the various rules hold up in the current database of β-sheet proteins. There are a

TABLE I. Absolute and Probabilistic Rules for Four-Stranded β-Sheet Proteins†

Feature	Total	Frequent	Rare	Never
Absolute rules				
clashes	4	0	0	4
2413s	8	0	0	8
pretzels	4	0	0	4
spirals	2	0	0	2
Probabilistic rules				
3 jumps	8	0	0	8
2 jumps	40	3	11	26
1 jump	40	13	14	13
0 jumps	8	3	4	1
3 parallel pairs	12	3	2	7
2 parallel pairs	36	0	10	26
1 parallel pair	36	8	14	14
0 parallel pairs	12	8	3	1

†The features are described in detail in Results. The sheet motifs are classified into three groups based on their frequency of occurence: frequent (eight or more times), rare (at least one but less than eight times), or never. The motif in panel 46 in Figure 1 is the only motif subject to two absolute rules ("2413s" and "pretzels").

total of 17 motifs that violate one of the four of the absolute rules, and none of these are observed in proteins. Assuming that all loops between parallel strand pairs are right-handed, there can be a clash between two crossings connecting pairs of parallel strands in some motifs (panels 4, 13, 26 and 31 in Figure 1, Table I "clashes"). Cohen et al.[14] noted the absence of what they call "pretzels." The authors report that the spatial strand sequence 2413 never occurs in sheets (Fig. 1, panels 41–48; Table I "2413s"). We use this term for different motifs, namely those that have crossing loops (Fig. 1, panels 17, 23, 39 and 46; Table I "pretzels"). For obvious reasons, we named the motifs in panels 29 and 36 spirals. As Table I shows, none of those motifs ever occurs in native four-stranded proteins. In addition to those absolute rules, probabilistic rules can be formulated to assess the preferred status of other motifs. Also in Table I, we illustrate the distributions for the number of strand pairs adjacent in sequence that are not neighbours in the sheet (referred to as "jumps"), and the distributions for the number of parallel pairs in four-stranded sheets. Motifs with 3 jumps never occur, as for four-stranded sheets these motifs coincide with the 2413 pretzels reported by Cohen et al.[14] Clearly discouraged are also motifs with two jumps, and simpler motifs with none or only one jump are preferred. We can also see from Table I that there is no frequently occuring four-stranded sheet motif with two parallel strand pairs. The preference for low numbers of chain reversals and the preference for purely parallel and purely anti-parallel β-sheets becomes even more obvious when we regard the absolute numbers how often the motifs occur, and relate those numbers to what would be expected to if all sheet motifs would occur equally often (Tables II and III).

The above rules help us eliminate or discourage certain structures that we see in Rosetta decoy sets. However, they do not explain why, for example, we never see the all parallel 1243 motif (panel 2 in Fig. 1) in native proteins, especially since the related all parallel 2134 motif occurs very frequently in α/β proteins. For some of the motifs that never occur, we do not know if there are physical constraints that prohibit those sheets or if those were simply never sampled by evolution. To overcome the fact that these rules do not account for the entire distribution of sheets motifs in native proteins, we develop a scoring function that captures information beyond those rules.

Scoring Function for β-Sheet Motifs

We distinguish between sheets of up to four strands, and sheets of five strands or more. For up to four strands, the counts from the database were sufficient to determine the probability distribution of the motifs, using the raw counts and small pseudo counts for each motif. For larger sheets this is not the case, and the sheet distributions were modeled using insights gained by the studies described in the previous section. The model, as written in equation (1) in Scoring Function for Larger Sheets, is applicable to all sheets of size five or larger.

Sheet motifs with four or less strands

There are only two ways for two strands to pair: parallel or anti-parallel. Table IV shows the counts of parallel and anti-parallel pairs of strands in the database (2,000 + non-homologous structures) and the estimated probability of being (anti-)parallel, conditioning on the loop length between the two strands and the helical status of the protein. If the loop between the two strands is ten residues or less, chances are 99% that the sheet is anti-parallel for both α/β and all β proteins. If the loop has more than ten residues, α/β proteins are about twice as likely to have a parallel sheet than all β proteins (27% compared to 13%).

There are twelve motifs for three-stranded sheets (Fig. 2). We classify the loop lengths between the three strands as short-short (L_1), short-long (L_2), long-short (L_3), and long-long (L_4). For most bins (we use the term "bin" to refer to the class of structures that have a specific motif, loop length distribution and helical status), the initially fitted probabilities were very similar, comparing α/β and all β proteins. Using χ^2 tests for bins with sufficient counts, we determined which bins we could combine for α/β and all β proteins. We removed single counts from bins and used small pseudo-counts to fit the motif probabilities, which are shown in Table V. Adding pseudo counts avoids completely ruling out motifs which are not in the database, but still strongly discourages them. For most bins we found no difference between α/β and all β proteins. For each type of protein however, there are vast differences between the class of loop lengths. If both loops are short, then the up-down-up motif M_3 is by far the most likely (probability $\approx$ 90%). Only motifs M_6 and M_{12} were observed in the database as well, all other motifs have a very small chance, which is only nonzero due to the pseudo counts. If not both loops are short, especially if both loops have more than ten residues, motifs M_6 and M_{12} become even more likely alternatives to M_3.

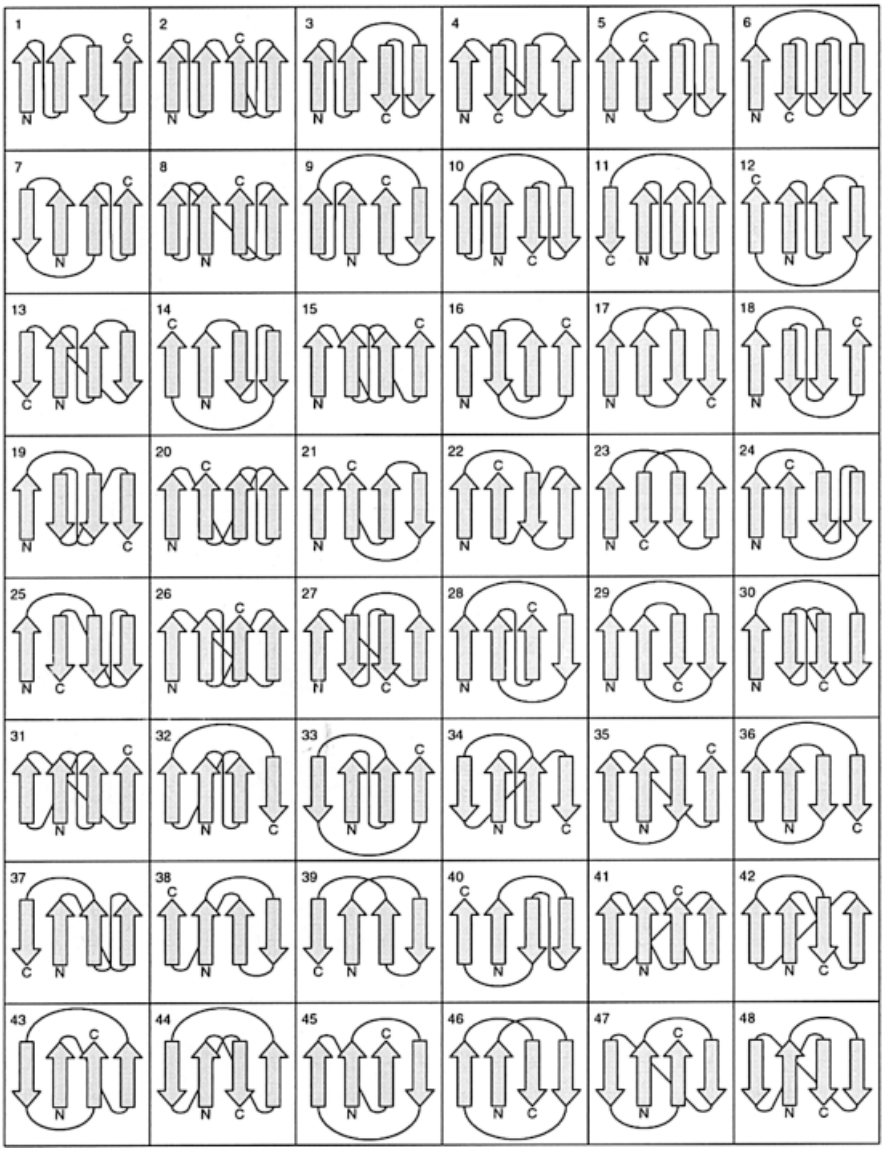


Fig. 1. A complete list of four-stranded motifs that did not occur in the Dunbrack database, the motifs are sorted by the number of jumps, strand sequence, and strand orientation (in that order). Motifs with clashes are in panels 4, 13, 26, 31, motifs with 2431s are in panels 41–48, motifs with pretzels are in panels 17, 23, 39, 46, and motifs with spirals are in panels 29 and 36.

There are eight classes for the loop lengths between the four strands, labeled as follows:

L_1	short	short	short
L_2	short	short	long
L_3	short	long	short
L_4	short	long	long
L_5	long	short	short
L_6	long	short	long
L_7	long	long	short
L_8	long	long	long

As for the three-stranded motifs, we used χ^2-tests to determine which bins we can combine for α/β and all β proteins, and used small pseudo-counts to fit the motif probabilities of the four-stranded sheets. Displaying all fitted probabilities would be excessive (96 motifs, 8 length classes, and 2 classes for the helical status equals 1,536 bins). Instead, we show the motifs that have a probability of 5% or more in their respective length classes for both α/β and all β proteins in Figure 3. The preference for pure anti-parallel β-sheets is quite striking for four-stranded sheets. Pure parallel motifs only occur frequently in α/β proteins when all loops have more than ten residues. In that case, the pure parallel 2134 motif is the most likely [Fig. 3(a), top right). It occurred 44 times in the database, although the very similar all parallel 1243 motif (Figure 1, panel 2) never did. Also noteworthy are the Greek key motifs (anti-parallel 2341 and 1432 motifs), previously discussed,[11,15] which occur in both α/β and all β proteins.

TABLE II. Distribution of Parallel Pairs in β-Sheets With Two Edge Strands[†]

	Number of parallel pairs							
n_{strand}	0	1	2	3	4	5	6	7
Percentages derived from the database								
3	81.3	15.4	3.3					
4	70.0	19.4	3.1	7.5				
5	33.4	27.6	13.4	3.1	22.5			
6	24.2	19.5	11.6	6.8	16.8	21.1		
7	24.2	12.6	7.8	4.8	22.5	10.8	17.3	
8	23.1	15.6	9.1	5.4	7.5	15.1	7.5	16.7
Ratios of observed and expected percentages								
3	3.3	0.3	0.1					
4	5.6	0.5	0.1	0.6				
5	5.4	1.1	0.4	0.1	3.6			
6	7.8	1.3	0.4	0.2	1.1	6.7		
7	15.5	1.3	0.3	0.2	1.0	1.2	11.1	
8	29.6	2.9	0.6	0.2	0.3	1.0	1.4	21.4

[†]Shown are the parallel strand pair distributions for the sheet motifs of sizes 3–8, respectively, in the database and the ratio of the observed percentages and the percentages calculated under the assumption that all motifs are equally likely. A ratio larger than one, therefore, indicates that the motif type is observed more often than expected under the above assumption. This shows a clear preference for all parallel and all anti-parallel motifs.

TABLE III. Distribution of Sequentially Adjacent Strands That Are Not Adjacent in the Sheet (Jumps) in β-Sheets With Two Edge Strands.[†]

	Number of jumps							
n_{strand}	0	1	2	3	4	5	6	7
Percentages derived from the database								
3	57.7	42.3						
4	20.6	59.1	20.3	0.0				
5	8.6	46.2	38.0	6.2	1.0			
6	10.0	33.4	42.1	13.2	1.0	0.3		
7	7.8	23.8	38.1	24.2	6.1	0.0	0.0	
8	6.4	28.5	16.1	30.1	15.1	3.8	0.0	0.0
Ratios of observed and expected percentages								
3	1.7	0.6						
4	2.5	1.4	0.5	0.0				
5	5.2	3.5	1.0	0.2	0.1			
6	36.0	11.0	2.5	0.4	0.1	0.0		
7	196.4	42.9	8.5	1.4	0.2	0.0	0.0	
8	1300.7	337.9	17.8	5.5	0.8	0.1	0.0	0.0

[†]Shown are the the distribution of jumps for the sheet motifs of sizes 3–8, respectively, in the database and the ratio of the observed percentages and the percentages calculated under the assumption that all motifs are equally likely. A ratio larger than one, therefore, indicates that the motif type is observed more often than expected under the above assumption. This shows a clear preference for motifs with few jumps.

Scoring Function for Larger Sheets

In Methods, we established that there are $n! \times 2^{n-2}$ possible motifs for n-stranded sheet. In our model, we also take into account whether the protein is α/β or all β (2 classes) and the loop lengths between the strands (2^{n-1} classes). Hence, to model a probability distribution for the motifs of an n-stranded sheet we have to consider $n! \times 2^{n-2} \times 2 \times 2^{n-1} = n! \times 2^{2n-2}$ bins. For up to four strands, the counts from the database were sufficient to model the probability distribution of the motifs, using the raw counts for each motif. For larger sheets, this is no longer the case. Five-stranded sheets can have 960 different motifs, and

TABLE IV. Counts and Fitted Probabilities for Parallel (P) and Anti-Parallel (AP) Pairs of Strands in α/β and All β Proteins[†]

	α/β		All β	
	S	L	S	L
Counts				
P	8	127	3	32
AP	609	338	278	207
Probabilities				
P	0.01	0.27	0.01	0.13
AP	0.99	0.73	0.99	0.87

[†]S denotes a short loop (ten or less residues), L denotes a long loop (more than ten residues).

taking loop lengths and whether the protein is α/β or all β into account, there are 30,720 bins to model. In addition, the counts of sheets in the database rapidly decline with sheet size (Fig. 4). We already established that there is a preference for low numbers of jumps and a preference for purely parallel and purely anti-parallel β-sheets. This can also be seen in Figure 5. Exploratory data analysis also showed that the position of the first strand in the motif is not random. To proceed with the modeling of the sheet configurations, we made the assumption that the likelihood of an individual motif can be modeled by global features, such as the number of parallel pairs and the number of sequentially adjacent neighbors in the sheet. Analyzing the data, we decided to describe a motif in a sheet of size five or larger by five different variables: (1) The number of parallel neighbour strands in a motif, and (2) how many of those are connected by short loops; (3) The number of strand pairs adjacent in sequence that are not neighbors in the sheet (jumps), and (4) how many of those jumps there are with a short loop between the strand pair; (5) The position of the first strand in the motif. These features were not assumed to be independent. The fact that several different motifs [as described by features (1)–(5) was taken into account.] The features were combined into a distribution function for scoring individual sheets. Below, we give a rough outline of the model, and refer the reader to the thesis of Ruczinski[22] for details. The relevant chapter is also available online (http://www.jhsph.edu/biostats/research/ruczinski.html).

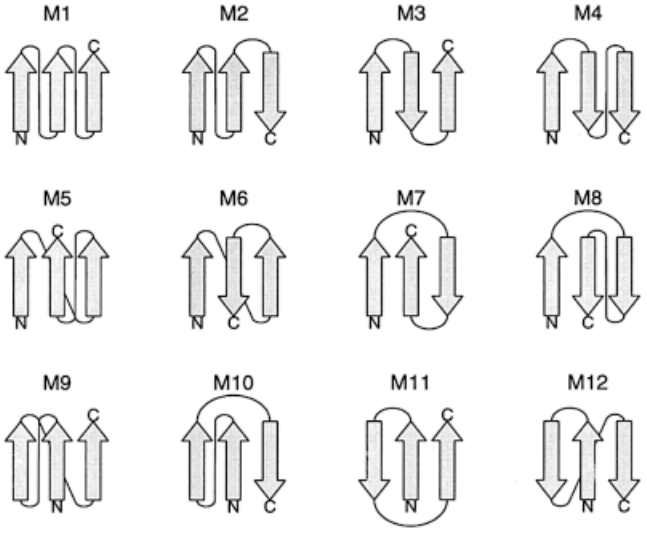


Fig. 2. The twelve possible configurations of three stranded sheets, taking the two axes of symmetry into account. The probability distributions for those motifs are shown in Table V.

Let H be the helical status of the protein and let L be the loop length distribution between the n strands. Let P_p be the number of parallel neighbour strands in a motif, $P^s{}_p$ the number of parallel neighbour strands in a motif with a short loop in between, J the number of jumps, J^s the number of jumps with a short loop between the strand pair, and F the position of the first strand in the motif. Then, using rules for conditional probabilities, we get

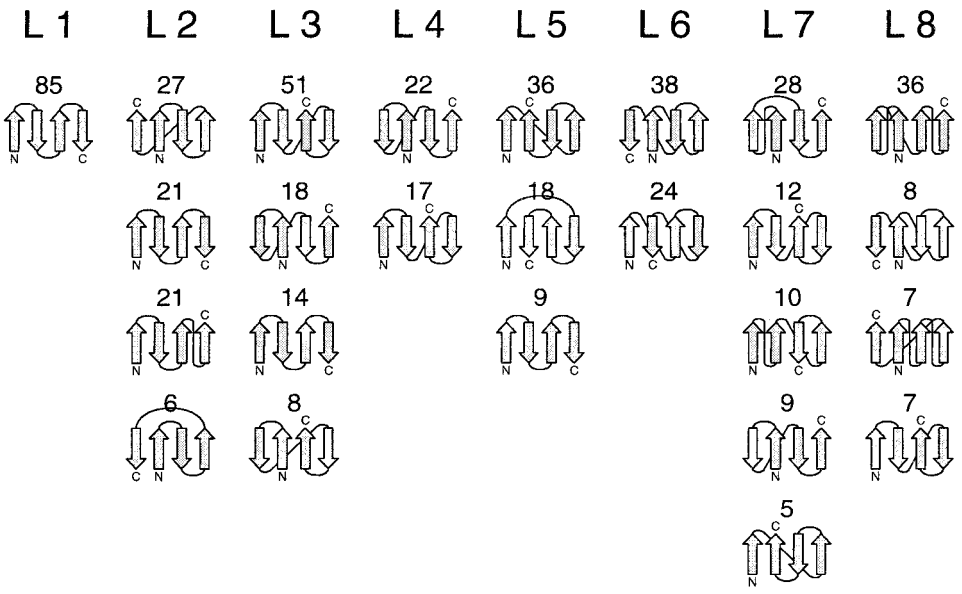
$$\begin{aligned} &P(P_p,P^s_p,J,J^s,F|n,H,L) \\ &= P(F|,H,L) \times P(P_p,P^s_p,J,J^s|,H,L,F) \\ &= P(F|,H,L) \times P(P_p,J|,H,L,F) \times P(P^s_p,J^s|,H,L,F,P_p,J) \end{aligned} \quad (1)$$

After carrying out some exploratory data analysis and simple statistical tests, we found that $P^s{}_p$ and J^s can be taken to be conditionally independent, given n, H, L, F, P_p, and J. The first two terms on the right-hand side of equation (1) are estimated non-parametrically using counts from the database, the two terms arising from the last term of the right-hand side of equation (1) are estimated using binomial models.

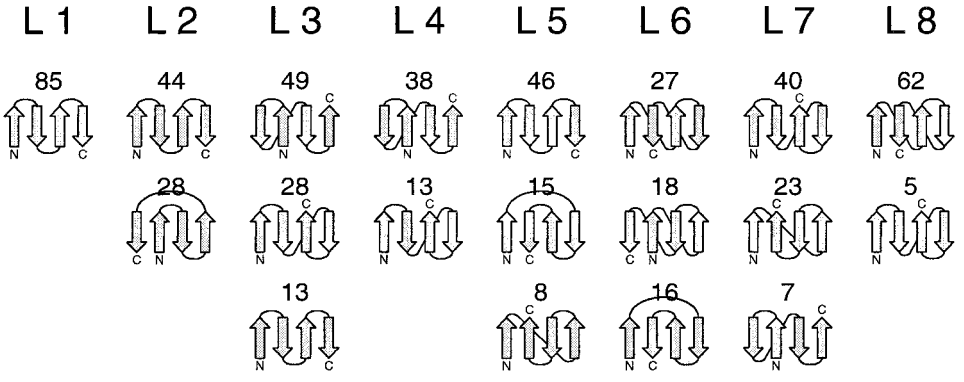
TABLE V. Fitted Probabilities for Three-Stranded Motifs in α/β and All β Proteins, Conditional on Loop Lengths†

	α/β				all β			
	L_1	L_2	L_3	L_4	L_1	L_2	L_3	L_4
M_1	.004	.006	.005	.049	.004	.006	.005	.042
M_2	.004	.006	**.083**	**.080**	.004	.006	**.083**	**.068**
M_3	**.897**	**.401**	**.276**	**.162**	**.897**	**.611**	**.422**	**.252**
M_4	.004	**.262**	.005	.029	.004	.042	.005	.024
M_5	.004	.006	.005	.019	.004	.006	.005	.016
M_6	.036	.012	**.547**	**.282**	.036	.012	**.401**	**.239**
M_7	.004	.006	.048	.032	.004	.006	.048	.027
M_8	.004	.006	.005	.014	.004	.006	.005	.012
M_9	.004	.006	.005	**.114**	.004	.006	.005	.012
M_{10}	.004	.006	.005	.035	.004	.006	.005	.030
M_{11}	.004	.027	.005	.032	.004	.028	.005	.027
M_{12}	.028	**.259**	.010	**.153**	.028	**.267**	.010	**.252**

†A short loop has ten or less residues, a long loop more than ten residues. The loop lengths between the strands are short-short (L_1), short-long (L_2), long-short (L_3), and long-long (L_4). The motifs (M_1–M_{12}) are shown in Figure 2. Probabilities of more than 5% are highlighted in bold.



(a) The most common four-stranded motifs for α/β proteins.



(b) The most common four-stranded motifs for all β proteins.

Fig. 3. Four-stranded motifs with probabilities larger than 5%. The actual probabilities (rounded, in percent) are indicated above the motifs. L1 through L8 refers to the loop length classes, defined in Results.

Application to Structure Prediction for β-Sheet Proteins

Sheet filter

In Methods, we described the procedure necessary to extract the sheet information from the structures in the currently available database of non-homologous proteins. This procedure will not be able to extract the sheet information when the sheet under consideration is not properly formed in a Rosetta structure. This happens, for example, when a strand has more than two neighbor strands according to our neighbor definition (see Methods), or if a strand is unpaired. We transformed the previously described procedure to also evaluate the quality of sheets in the predicted structures, filtering out structures with poorly assembled sheets. After a more detailed description of this filter, we show how it actually improves the overall quality of decoy sets.

In structures generated by Rosetta, we usually allow distances of 6.5Å in the definition of neighbor strands, since we do not model hydrogen bonds explicitly, and strand pairs might not be aligned perfectly. After strand neighbors are identified, the sheet motif can be easily determined, unless the structure does not have a proper sheet. The routine we implemented sequentially checks

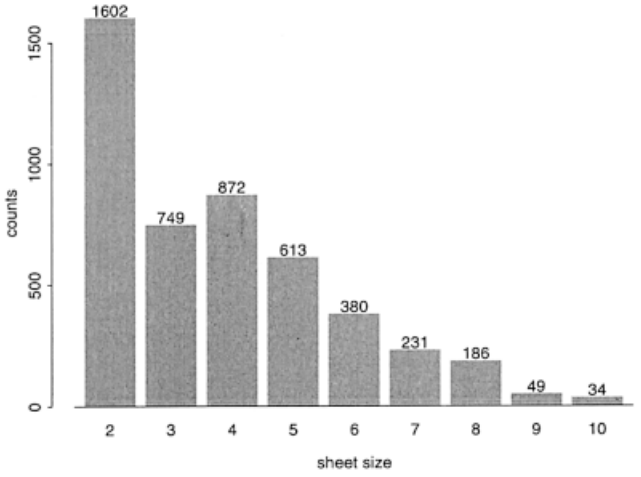


Fig. 4. The counts of sheets of sizes 2–10 observed in non-homologous proteins.

the following criteria, and rejects the structure under consideration for not having a proper sheet if any of those criteria are met: (1) The decoy does not have any strands at all or only has a single strand, but secondary structure predictions indicated that the structure should contain strands. (2) At least one strand has more than two neighbors, or at least one strand has no neighbors, i.e., the strand is included in an improper sheet, or it is unpaired. (3) The connection between two parallel strands is a left-handed connection. (4) Three or four strands in a structure form a barrel type structure, i.e., each strand in that formation has exactly two neighbors. In our database, there are only barrels of size five or larger (predominantly size eight).

We also tried to incorporate other features in the filter, but found that those did not help discriminate good from bad decoys after applying the above-mentioned rules. For example, we also included a subroutine in the filter that checked for poorly aligned neighbor strands, allowing for a twist in the sheet.[23] The fact that this feature did not

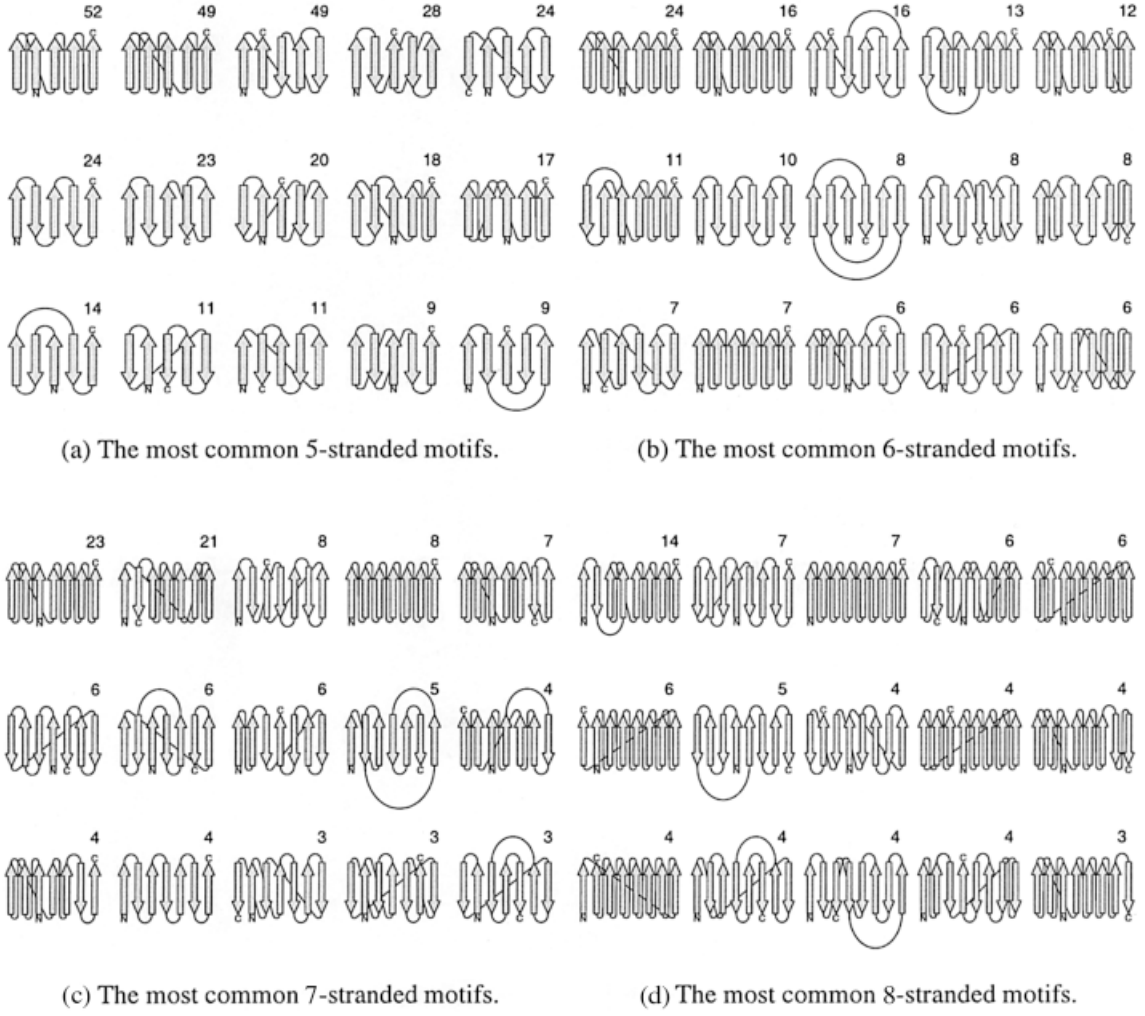


Fig. 5. Larger sheets frequently observed in the database. The sheets clearly have common patterns. There is a preference for purely parallel or purely anti-parallel motifs, and motifs with no or few chain reversals.

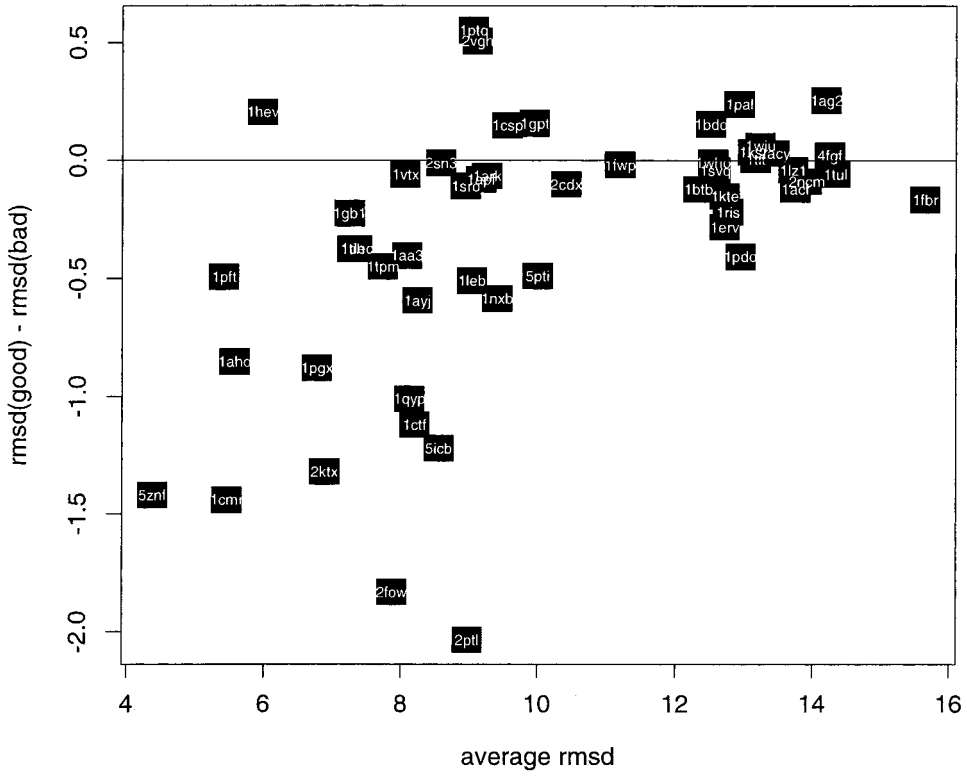


Fig. 6. The improvement in average rmsd achieved for 52 small proteins. On the x-axis, is the average rmsd for the complete generated decoy set: on the y-axis, is the difference in average rmsd between those decoys that passed the filter and those that did not. Improvements are achieved for almost all decoy sets with less than 10Å average rmsd.

provide any additional discriminatory power is probably because a similar term is used in the scoring function of Rosetta that guides the folding of the decoy. In the future, additional features could be added to the filter. For example, "pretzels" are strongly discouraged by the distribution of sheet motifs we described earlier but can be completely ruled out by using a filter that excludes these motifs.

To quantify how the filter improves the overall quality of decoy sets, we applied the filter to structures generated for 52 β-sheet containing proteins of up to 150 residues. The 52 targets are the β-sheet containing proteins of the Friesner laboratory list of small proteins.[24] For each protein, we generated between 1,000 and 2,000 structures, and divided them into good and bad sets, according to our filter. Figure 6 shows the improvement in average rmsd achieved for the filter. The x-axis is the average rmsd for the decoy set generated, the y-axis is the difference in average rmsd between the decoys that passed the filter and the decoys that did not. The average rmsd for the decoy sets range from 4.4Å for the zinc finger DNA binding domain (5znf) to 15.6Å for a cell adhesion protein (1fbr). Using the filter, we could not achieve any discrimination between better and worse structures for the poor decoy sets (average rmsd more than approximately 10Å). This is not surprising, since the "better" decoys were still poor structures. However, for better quality decoy sets we consistently achieved improved average rmsd for the structures that passed the filter compared to those that did not.

Another interesting feature of the filter is that it might enable us to predict the quality of the decoy sets from the acceptance frequency of the structures. Figure 7 shows the acceptance frequency (percentage of decoys that passed the filter) vs. the rmsd 5th percentile, another measure of quality of the respective decoy set. While not all good decoy sets had a high acceptance ratio, all decoy sets with high acceptance ratios had many low rmsd structures. In other words, if our filter passes many structures, we can be confident that Rosetta produced some near-native structures for the protein.

Elimination of bad structures using the scoring function

Since the Rosetta folding algorithm starts with the extended chain of amino acids and sheets are formed during the annealing process, the sheet scoring function as described earlier is only useful for distinguishing between folded structures in a decoy set. At this point, Rosetta is mainly used to predict ab initio structures of small proteins. While Rosetta still fails to generate a large variety of complex structures and, depending on the secondary structure prediction for the target, sheets with five or more strands are rare in the decoy sets, Rosetta consistently generates good decoys for proteins

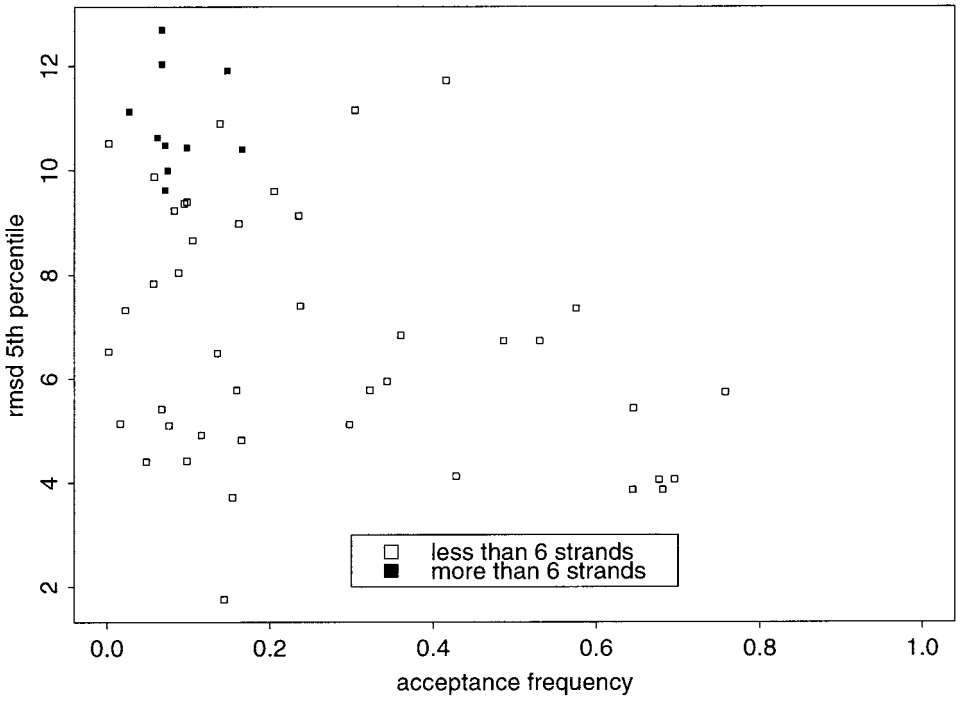


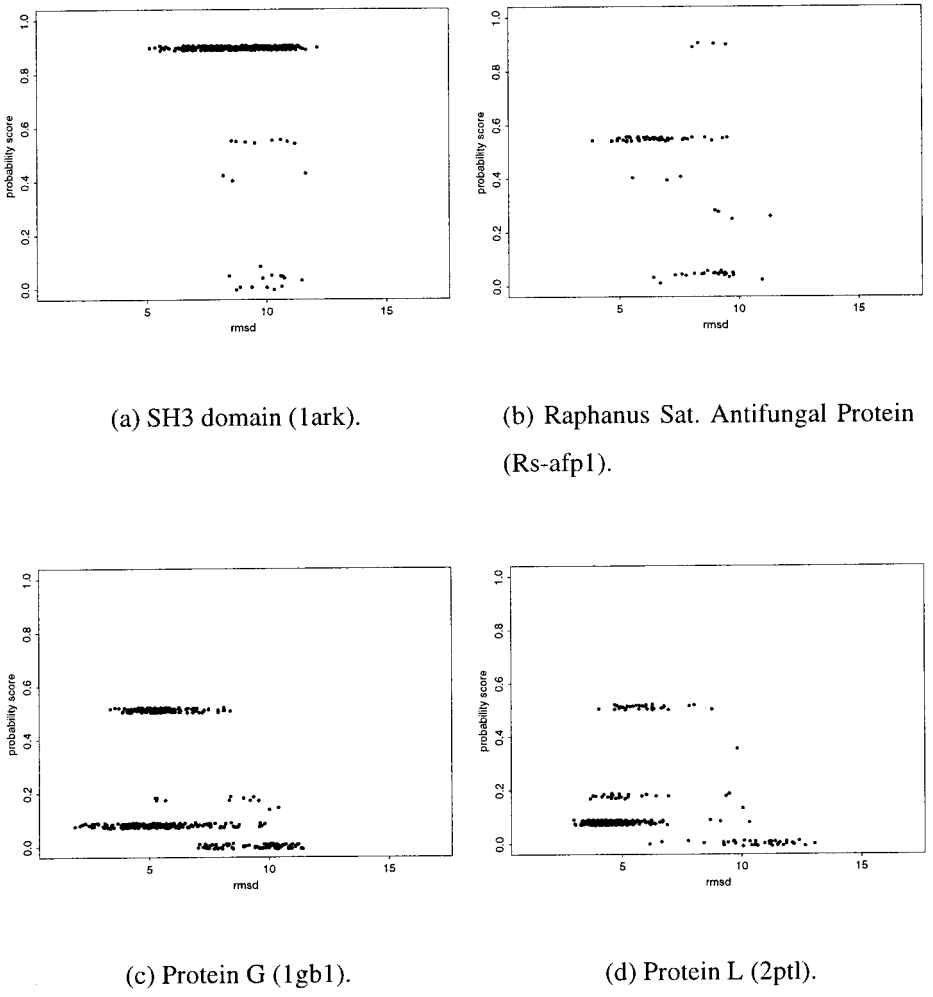
Fig. 7. Acceptance frequencies vs. rmsd 5th percentile for 52 decoy sets. The rmsd 5th percentile is the rmsd that separates 5% of the decoys with the lowest rmsds from the higher rmsd structures. All decoy sets with high acceptance frequencies contain many low rmsd, i.e., near-native structures. However, the converse is not necessarily true. The fact that all proteins with 7 or more strands are in the top left corner illustrates the problems that Rosetta still has folding those structures.

with smaller sheets. In those cases, the sheet scoring function proves to be quite useful to discriminate good and bad structures. The sheet scoring function reflects the frequency of how often sheet motifs occur in native proteins. Hence, not necessarily are the motifs in a decoy set that get assigned the highest probabilities the correct ones, and we do not expect those decoys to necessarily have the lowest rmsds. However, the motifs that are totally discouraged by the scoring function, i.e., that never or very rarely occur in native proteins, are usually in decoys with very high rmsd (Fig. 8), and the elimination of those further improves the decoy sets and aids in the pick of a final prediction for the targets.

The sheet filter and the scoring function are two methods to discriminate between good and bad decoys. Since the scoring function can only be applied to structures with proper β-sheets, the filter is a prerequisite for the scoring function. The α/β and all β ab initio targets in the Fourth Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP4, described in structure prediction for β-sheet proteins in CASP4) contained mostly large sheets, for which not very many decoys with proper sheets could be generated. Therefore, only the sheet filter was used in this round of CASP for the selection of candidates from the decoy sets. In the future, as Figure 8 suggests, once enough decoys with good sheets for the targets can be generated, the scoring function can be of great help for picking the final predictions by further eliminating structures with unlikely sheet configurations. This could potentially be a problem if the target protein has a rare topology. It, therefore, might be more advantageous, once large decoy sets with good sheets can be generated, to adjust the distribution of β-sheet motifs in the decoy sets according to the distribution of motifs in native proteins, which is described by the scoring function.

Structure prediction for β-sheet proteins in CASP4

For an objective comparison of existing protein structure prediction methods, a bi-annual blind test, the Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP), has been carried out (http://predictioncenter.llnl.gov). For this experiment, the sequences of some newly (by X-ray crystallography or NMR) solved three-dimensional protein structures are distributed prior to publication of the structure. The structure of those targets are not published until after a given deadline, which gives CASP participants the chance to submit their predictions for the 3D protein structures. The above-described procedures were part of our CASP4 protocol.[25–27] As a first step, we generated up to 150,000 independent structures (decoys) for each target. In the second step, the decoy sets were filtered. We corrected for the low contact order bias we observed in previous Rosetta populations.[3] For targets with strands we applied the sheet filter described here and removed structures with non-protein-like strand arrangements, which turned out to be between 30 and 90% of the decoys. Then the decoys (backbone plus one centroid per residue) were expanded by the addition of side chains[26] to an all



(a) SH3 domain (1ark).

(b) Raphanus Sat. Antifungal Protein (Rs-afp1).

(c) Protein G (1gb1).

(d) Protein L (2ptl).

Fig. 8. Rmsd vs. motif distribution score. For ease of visualization, motif scores were jittered slightly. The SH3 domain and Rs-afp1 have a single three-stranded sheet each, Protein G and Protein L have a single four-stranded sheet each. Shown are only the decoys with the correct number of strands in the respective structure. Since the scoring function reflects the frequency how often sheet motifs occur in native proteins, motifs that score close to zero are usually in decoys with very high rmsds, and the elimination of those further improves the decoy sets. Note that all decoys with rmsd smaller than say 5.5Å have probability scores larger than 5%.

atom models. A physically based all-atom potential was applied to identify well-packed models (Tsai, unpublished data). The last step was to cluster the filtered structures for each target.[27] The cluster centers were ranked by size, and, in general, the five largest unique clusters were submitted as predictions for the respective targets (manual inspection of cluster centers was still necessary for larger targets). In some cases, there was not enough computing time available to add side chains to all structures before clustering, so the last two steps were exchanged.

Rosetta's CASP4 ab initio structure predictions were considerably more consistent and accurate than structures produced by ab initio structure prediction methods in the past (including our own), for example those submitted for CASP3.[28] Since participating in CASP3,[29] we altered our prediction method in various ways. Mainly, we improved the basic simulation methods, and added the previously described filters and clustering technique (details given in ref.[25]. One of the motivations for the development of the sheet filter was our failure in CASP3 to consistently predict β-sheet proteins. While it is impossible to separate the effects that the various improvements in Rosetta for the CASP4 predictions had, it appears that the sheet filter together with the correction for the low contact order bias was crucial for the prediction of several β-sheet containing targets. Besides making the clustering procedure feasible by substantially reducing the total number of decoys, the β-sheet filter also improved the overall quality of the decoy sets (Fig. 9). While the only globally correct structures we submitted in CASP3 were α-helical domains,[29] we were able to predict ab initio β-sheet protein structures of unprecedented accuracy in CASP4 (targets T087 ab, T091, T105, T110, T116ab, Fig. 10). To our knowledge, our predictions for the domains A and B for target T087 (164 and 192 residues, respectively) are the largest correct ab initio predictions of β-sheet containing structures to date.

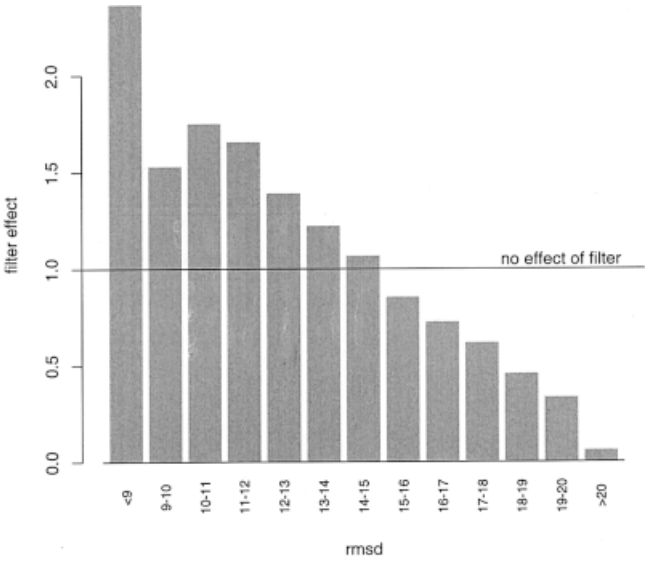


Fig. 9. The effect of the β= sheet filter on the decoy set for domain A of target T087. The filter effect is a relative measure of the efficiency of the filter on a decoy set. This score for each rmsd bin is the percentage of decoys in that bin that passed the filter, divided by the percentage of all decoys in that bin before applying the filter. Hence, the filter effect of a non-informative filter is 1, indicated by the solid line. In this example, the proportion of structures less than 9Å is almost 2.5 times higher in the filtered decoy set compared to the overall population.

CONCLUSION

We surveyed the distribution of β-sheet motifs with two edge strands (open sheets) in the database of non-homologous proteins, and examined deterministic and probabilistic rules for sheet motif distributions. We used the results of our survey to develop a full scoring function of sheet motifs for both α/β and all β proteins, which also takes the loop lengths between the strands into account. This scoring function, paired with a filter to eliminate structures with poor sheet configurations, proved to be valuable in discriminating good and bad decoys. In ab initio structure prediction, the use of the scoring function is still limited, but will become more important in the future, since we hope to be able to use the increase in computer power to create Rosetta decoys with larger and more non-local sheets. We modeled the distributions of β-sheets, but the physical origins of those distributions remain unclear. However, it is not obvious to us why some motifs were not in the database we investigated, although some very similar motifs occur frequently in nature. For example, the all parallel 1,243 motif (panel 2 in Fig. 1) never occurs in native proteins, but the related all parallel 2,134 motif occurs very frequently in α/β proteins [Fig. 3(a) L8]. Similarly, the motif in panel 1 of Figure 1 never occurs in native proteins, but reversing the order of the first two strands makes it a likely motif in α/β proteins [Fig. 3(a)

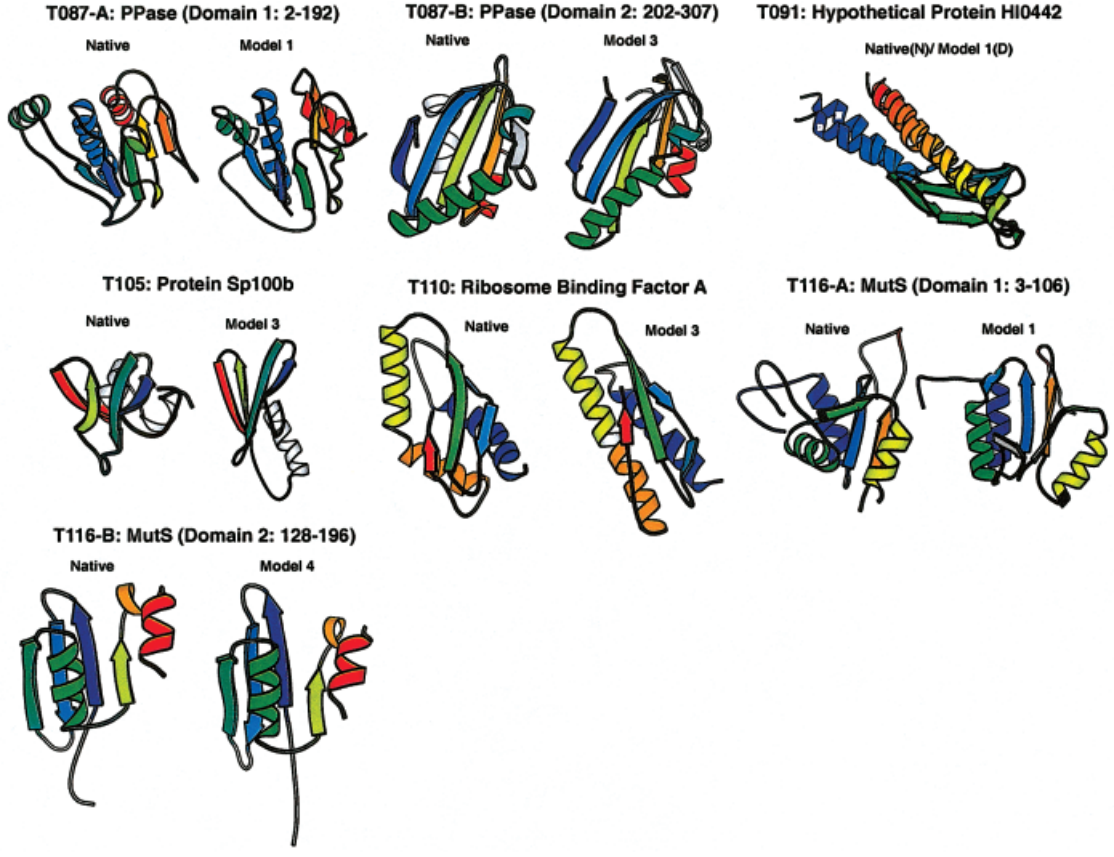


Fig. 10. Predicted and actual structure of β-sheet CASP4 targets. For a detailed description of each target see Bonneau.[25]

L7]. This puzzle poses an interesting challenge for current protein design methods.

Some results of our studies are available online at http://biosun01.biostat.jhsph.edu/iruczins/sheets/sheets.html. We provide a tool to search the pdb database for specific sheet motifs, and an online version of our scoring function.

ACKNOWLEDGMENTS

We thank Brian Kuhlman and Jerry Tsai for helpful comments and suggestions. This work was supported in part by a grant from the Packard foundation to D.B. and NIH grant 74841 to C.K.

REFERENCES

1. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol 1997;268:209–225.
2. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognintion of native-like protein structures using a combination of sequence dependent and sequence independent features of proteins. Proteins, 1999;34:82–95.
3. Bonneau R, Tsai J, Ruczinski I, Baker D. Improvement of ab initio protein structure prediction using insights from experiments. Forthcoming.
4. Richardson JS. The anatomy and taxonomy of protein structure. Adv Prot Chem 1981;34:167–339.
5. Chothia C, Finkelstein AV. The classification and origins of protein folding patterns. Annu Rev Biochem 1990;59:1007–1039.
6. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. A database of protein structure failies with common folding motifs. Prot Sci 1992;1:1691–1698.
7. Orengo CA. Classification of protein folds. Curr Opin Struct Biol 1994;4:429–440.
8. Orengo CA, Flores TP, Taylor WR, Thornton JM. Identification and classification of protein fold families. Prot Eng 1993;6:485–500.
9. Murzin AG, Brenner SE, Hubbard T, Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
10. Orengo CA, Thornton JM. Alpha plus beta folds revisited: some favoured motifs. Structure 1993;1:105–120.
11. Hutchinson EG, Thornton JM. The greek key motif: extraction, classification and analysis. Prot Eng 1993;6:233–245.
12. Richardson JS. Handedness of crossover connections in β sheets. Proc Natl Acad Sci 1976;73:2619–2623.
13. Sternberg MJE, Thornton JM. On the conformation of proteins: An analysis of β-pleated sheets. J Mol Biol 1977;110:285–296.
14. Cohen FE, Sternberg MJE, Taylor WR. Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. J Mol Biol 1982;156:821–862.
15. Richardson JS. β-sheet topology and the relatedness of proteins. Nature, 1977;268:495–500.
16. Taylor WR, Green NM. The predicted secondary structures of the nucleotide-binding sites of six cation-transporting atpases lead to a probable tertiary fold. Eur J Biochem 1989;179:241–248.
17. Clark DA, Shirazi J, and Rawlings CJ. Protein topology prediction through constraint-based search and the evolution of topological folding rules. Prot Eng 1991;4:751–760.
18. Cohen FE, Sternberg MJE, Taylor WR. Analysis and prediction of protein β-sheet structures by a combinatorial approach. Nature 1980;285:378–382.
19. Woolfson DN, Evans PA, Hutchinson EG, Thornton JM. Topological and stereochemical restrictions in β-sandwich protein structures. Prot Eng 1993;6:461–470.
20. Sternberg MJE, Thornton JM. On the conformation of proteins: The handedness of the connection between parallel β-strands. J Mol Biol 1977;110:269–283.
21. King RD, Clark DA, Shirazi J, Sternberg MJE. A database of protein structure failies with common folding motifs. Prot Eng 1994;7(11):1295–1303.
22. Ruczinski I. Logic Regression and Statistical Issues Related to the Protein Folding Problem. PhD thesis, University of Washington, Seattle, WA, August 2000.
23. Chothia C. Conformation of twisted beta-pleated sheets in proteins. J Mol Biol 1973;75:295–302.
24. Eyrich VA, Standley DM, Friesner RA. Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. J Mol Biol 1999;288:725–742.
25. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D. Rosetta in casp4: progress in ab initio protein structure prediction. Proteins 2001;Suppl 5:119–126.
26. Dunbrack RL, Karplus M. Backbone-dependent rotamer library for proteins. application to side-chain prediction. J Mol Biol 1993;230:543–574.
27. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci 1998;95:11158–11162.
28. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins 1999;(Suppl 3):149–170.
29. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of casp iii targets using rosetta. Proteins 1999;37:171–176.