Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of β -sheet formation in proteins

HONGYAO ZHU AND WERNER BRAUN

Sealy Center for Structural Biology, Department of Human Biological Chemistry and Genetics,
University of Texas Medical Branch, Galveston, Texas 77555-1157

(RECEIVED April 10, 1998; ACCEPTED October 15, 1998)

Abstract

A statistical analysis of a representative data set of 169 known protein structures was used to analyze the specificity of residue interactions between spatial neighboring strands in β -sheets. Pairwise potentials were derived from the frequency of residue pairs in nearest contact, second nearest and third nearest contacts across neighboring β -strands compared to the expected frequency of residue pairs in a random model. A pseudo-energy function based on these statistical pairwise potentials recognized native β -sheets among possible alternative pairings. The native pairing was found within the three lowest energies in 73% of the cases in the training data set and in 63% of β -sheets in a test data set of 67 proteins, which were not part of the training set. The energy function was also used to detect tripeptides, which occur frequently in β -sheets of native proteins. The majority of native partners of tripeptides were distributed in a low energy range. Self-correcting distance geometry (SECODG) calculations using distance constraints sets derived from possible low energy pairing of β -strands uniquely identified the native pairing of the β -sheet in pancreatic trypsin inhibitor (BPTI). These results will be useful for predicting the structure of proteins from their amino acid sequence as well as for the design of proteins containing β -sheets.

Keywords: β -strand alignment recognition; β -strand pairing; self-correcting distance geometry calculations; statistical potentials; tertiary structure

As the number of different protein folds determined from X-ray or NMR structures is reaching a saturation level, predicting the fold of a new sequence may be reduced to correctly matching of the amino acid sequence to one of a finite set of three-dimensional (3D) protein folds. Statistical potentials, derived from observed frequencies of contacts between nonbonded pairs of amino acids in known protein structures (Tanaka & Scheraga, 1976; Miyazawa & Jernigan, 1985, 1996; Sippl, 1990, 1993; Maiorov & Crippen, 1992), can be used to determine which folds are compatible with a given sequence (Sippl, 1990, 1995; Kocher et al., 1994; Park & Levitt, 1996; Mirny & Shakhnovich, 1996; Skolnick et al., 1997). Recent computer tools, such as our self-correcting distance geometry (SECODG) method (Hänggi & Braun, 1994), have simplified modeling the 3D structure of a protein if its sequence has sufficient

identity to that of a protein of known structure. In the absence of such a match, statistical potentials can greatly aid the calculation of the fold of a protein from sequence alone.

The SECODG-based program DIAMOD calculates the topology of segments of regular secondary structures using empirically calibrated distance constraints for buried and solvent exposed residues. The program, using only the amino acid sequence and the location of helical segments, determined the correct helical packing for six of eight small globular proteins with root-mean-square deviation (RMSD) of the calculated to the native fold of <3 Å (Mumenthaler & Braun, 1995). However, to extend the method to proteins containing β -sheets, more specific information was needed on preferential alignments of amino acid pairs between strands in β -sheet structures. We started from the assumption that there would be a nonrandom distribution of residues in contact across neighboring β -strands, which could be detected by statistical analysis of representative proteins in a training set.

We first determined the statistical pairwise potential for residue specific cross-strand interactions in native protein structures. The

Reprint requests to: Werner Braun, Sealy Center for Structural Biology, Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, Galveston, Texas 77555-1157; e-mail: werner@newton.utmb.edu.

first question was to predict the native alignment of two neighboring β -strands from the amino acid sequence of the strands. We analyzed the propensity of residue pairs shifted by one or two residues along the strands from the nearest contacts. These statistical potentials were combined in a simple pseudo-energy function that recognized the native β -sheets in 63% of β -sheets in 67 test proteins not included in the training set. Further, for a significant fraction of the cases tested, the energy function gave lowest values for the interaction of tripeptides with their native partners in β -sheets. The β -strand-pairing potentials presented in this study are intrinsically different from the contact potentials of Miyazawa and Jernigan (1985, 1996) and their extensions (Sippl, 1990, 1995; Maiorov & Crippen, 1992; Kocher et al., 1994; Mirny & Shakhnovich, 1996; Park & Levitt, 1996; Skolnick et al., 1997), in which the pairwise potentials were derived for amino acid packing in the interior of a protein without regard to the secondary structure. Based on a larger and more diverse data set, the values of our cross-strand potential are similar to the recent study (Wouters & Curmi, 1995) for side-chain interactions and pair correlations within antiparallel β -sheets. However, the formulation of amino acid pairing potentials for β -sheet-forming were extended, beyond the cross-strand pairing, to the local neighborhood in 3D structure of proteins.

The pseudo-energy functions based on these statistical pairwise potentials were also applied to β -sheet recognition in the context of 3D structure modeling with distance geometry calculations. Although the energy function did not identify the native β -strand pairs in all cases as the lowest energy conformation, superposition of distance constraints selects for the correct 3D structure from several possible β -strand alignments with low energies. In test calculations with BPTI, correct β -strand alignment significantly improved the calculated 3D structures while high residual distance constraints violations indicated incorrect β -strand pairing.

A β -strand is cooperative with its neighbors in four directions, which makes β -sheet structures more difficult to design than α -helices (Hecht, 1994; Beasley & Hecht, 1997). The cross-strand preferences in β -sheets are consistent with the experimental results in the thermal stability measurement of double mutant in the B1 domain of streptococcal protein G (Smith & Regan, 1995). Therefore, the statistical potentials and the detected preferential occurrence of tripeptides and their favorite partners, formulated in this study, will aid not just in protein modeling but also in the de novo design and architecture of β -sheet structures (Smith & Regan, 1997).

Results

Statistical propensity of amino acids to occur in α -helices or β -sheets

Table 1 lists the average frequency of occurrence of each amino acid in α -helix, f_α^X , and β -sheet structures, f_β^X , in the proteins of our training set as described in Methods. The data for α -helices demonstrate that selecting mainly β -sheet proteins in our training set did not dramatically change the frequency of amino acids occurring in other regular secondary structures. The conformational propensities for each amino acid in α -helix and β -sheet structures, P_α^X and P_β^X , calculated by dividing the corresponding frequency in those conformations for a residue by the global frequency of this residue in the dataset (Chou & Fasman, 1978; Levitt, 1978) are also included in Table 1. The average frequency of individual residues f_β^X in β -sheet structures is used to obtain the pair corre-

Table 1. Preferences of the 20 amino acids for secondary structures in proteins of the training dataset^a

Amino acids	α -Helix		β -Sheet	
	f_α^X	P_α^X	f_β^X	P_β^X
Gly	4.12	0.49	6.02	0.71
Ala	12.19	1.44	6.38	0.76
Val	6.14	0.85	13.08	1.82
Leu	10.74	1.32	9.97	1.23
Ile	5.63	1.04	9.41	1.73
Phe	3.91	0.99	5.83	1.48
Tyr	3.59	1.00	5.14	1.43
Trp	1.68	1.19	1.75	1.24
Met	2.50	1.31	2.40	1.26
Cys	1.38	0.74	2.37	1.26
Pro	2.50	0.55	1.99	0.43
Ser	5.27	0.80	5.40	0.82
Thr	4.32	0.71	7.26	1.19
Asn	3.57	0.80	2.36	0.53
Gln	4.50	1.29	2.71	0.78
His	1.79	0.86	2.12	1.01
Lys	6.91	1.12	5.07	0.82
Arg	5.24	1.23	3.68	0.86
Asp	5.58	0.92	3.12	0.52
Glu	8.42	1.45	3.93	0.67
Total	11,423		9,030	

^a f_α^X is the average frequency of residue X in α -helical structures, i.e., $f_\alpha^X = N_\alpha^X / \sum_Y N_\alpha^Y$. The propensities for each amino acid in α -helix P_α^X and β -sheet P_β^X are calculated by dividing the corresponding frequency in the conformations for a residue X by the global frequency, $f^X = N^X / \sum_Y N^Y$, as suggested by Levitt (1978).

lation $p_{i,j}^{XY}$ for each residue pair X and Y as defined in Equation 2 in Methods. In common with other studies (Levitt, 1978; Lifson & Sander, 1980; Williams et al., 1987; Munoz & Serrano, 1994; Wouters & Curmi, 1995; Park & Levitt, 1996), we find that the β -branched and aromatic residues and cysteine [Val, Ile, Phe, Tyr, Trp, Thr, Cys] have the highest tendency to form β -sheets, and [Glu, Asn, Asp, Pro] the lowest. The derived statistical preferences of β -strand forming for the 20 amino acids are consistent with most experimental scales of β -strand propensities in host-guest studies (Kim & Berg, 1993, 1996; Minor & Kim, 1994a; Smith et al., 1994; Otzen & Fersht, 1995). Exceptions are the $\Delta\Delta G$ values for β -sheet formation measured at an edge β -strand of the IgG-binding domain of protein G (Minor & Kim, 1994b). They display no correlation with our statistical preferences of β -sheet-forming, which suggests that β -strand propensities depend on the context of both secondary and tertiary structures. It is our goal to resolve these discrepancies for β -sheet formation through a combination of statistical analysis and 3D modeling with self-correcting distance geometry calculations (Hänggi & Braun, 1994).

The cross-strand contact potential ϵ^0

There are 5,749 cross-strand contact pairs ($i \sim j$) of amino acids found in 1,274 pairs of β -strands (888 in antiparallel and 386 in parallel β -sheets with 4,146 and 1,603 contact pairs, respectively) in the training dataset proteins. The cross-strand contact potentials ϵ^0 (Fig. 1) were obtained from Equation 4 for all 210 types of

Fig. 1. The relative locations of pairs between two strands in a β -sheet. The statistical potentials were derived with $(i \sim j)$, $(i \sim j \pm 1)$, and $(i \sim j \pm 2)$ pairs shown as ϵ^0 , ϵ^1 , and ϵ^2 , respectively.

amino acid pairs (Table 2). The values, in units of $k_B T$, range from -1.77 to 1.45 , where a negative value means that the residue pair was in cross-strand contact more often than expected from its frequency of occurrence. The general properties of these contact potentials are consistent with grouping the 20 amino acids into nonpolar [Gly, Ala, Val, Leu, Ile, Phe, Tyr, Trp, Met, Cys, Pro], polar [Ser, Thr, Asn, Gln, His], positively charged [Lys, Arg], and negatively charged [Asp, Glu] groups. The negative contact energies within the first group, except for the rarely occurring Pro residue, indicates the preferential pairing of nonpolar side chains, especially the branched residue pairs (Val-Val, Val-Ile, Ile-Ile, and Leu-Leu) across β -strands. The mainly positive values of the contact energies between nonpolar, polar, and charged groups indicate electrostatic interactions: residues of the same charges repulse one another, oppositely charged groups attract. The only exception is the small negative value for the Arg-Arg pair, presumably arising from the large nonpolar part of the Arg side chain, which makes the pair interaction slightly favorable.

Cys preferentially forms cross-strand contact pairs with other cystine residues. All disulfide-rich proteins included in our training dataset were examined in more detail. Of the 19 Cys-Cys cross-strand pairs in the dataset, eight are disulfide-bonded to each other and the remaining 11 are in a disulfide linkage with residues outside the pair (Harrison & Sternberg, 1996) since free cysteines occur rarely in native β -sheets. In addition to the critical role of long-range disulfide-bridges in the folding and stabilization of these proteins (Seckler & Jaenicke, 1992; Creighton et al., 1996), our data indicate the importance of cysteine pairs in aligning β -strands. Preferential pairing of oppositely charged amino acids and Cys residues were also observed in other studies of inter-residue contact energies (Miyazawa & Jernigan, 1985, 1996) and pair correlation within antiparallel β -sheets (Wouters & Curmi, 1995).

Several side-chain repulsive interactions, such as between Ile-Pro, Ile-Asn, His-Met, Cys-Arg, and Glu-Trp, cannot be explained by polarity considerations. While the aromatic side-chain Phe avoids pairing with Pro ($\epsilon^0 = 0.65$) or Lys ($\epsilon^0 = 0.75$), Tyr may pair with either amino acids. Recognition of such repulsive interactions can aid in designing β -sheet structures.

The first-order cross-strand potential ϵ^1

The first-order cross-strand potentials ϵ^1 (Table 3) correlate residues i and $j \pm 1$, i.e., those following next in the amino acid sequence to the cross-strand contact potentials ϵ^0 (Fig. 1). The side chains of these pairs, i and $j \pm 1$, are located in opposite faces of a regular β -sheet. We found 8,950 pairs (6,516 in antiparallel, 2,434 in parallel β -sheets) for the first-order cross-strand potentials ϵ^1 in our training dataset. The absolute values of these potentials are generally smaller than those for cross-strand (ϵ^0), which indicates a weaker specificity for the amino acid types at the positions i and $j \pm 1$ compared to the pairing correlation of $(i \sim j)$. The absolute value of ϵ^1 exceeds 0.4 for about 50 pairs of amino acids. Most of the first-order potentials for the nonpolar amino acids are weakly attractive, independent of the polarity of the partner residue, and strongly repulsive for pairs where both residues are polar or charged (independent of the sign of charges). Both faces of a β -sheet are seldom solvent-exposed; the characteristic amphipathic alignment of residues means that one face is usually solvent-exposed with the other contributing to the hydrophobic core, or both faces contribute to two different hydrophobic cores. However, there are exceptions to this general property, e.g., the strong repulsive pairs involving proline, Pro-Pro, Pro-Ser, Pro-Thr, Pro-Asp, Pro-Glu, and the weak attractive pair, His-His, which indicates that histidine behaves more like a nonpolar residue. The corresponding contact potentials ϵ^0 for proline are quite different from the ϵ^1 , showing that the position of proline is crucial in the correct alignment for β -sheet structures. There are no disulfide-bonded bridge S-S in 14 $(i \sim j \pm 1)$ pairs Cys-Cys found in the dataset, even though a strong attractive potential is observed with statistical significance.

The second-order cross-strand potential ϵ^2

The second-order cross-strand potentials ϵ^2 (Table 4), derived for the pairs between residues i and $j \pm 2$, are based on 6,430 pairs (4,766 in antiparallel and 1,664 in parallel β -sheets). The side chains of these $(i \sim j \pm 2)$ pairs, which are similar to those of ϵ^0 , point to the same face of a regular β -sheet. As expected, ϵ^2 values are similar to ϵ^0 values but have smaller absolute values. Most of the $(i \sim j)$ pairs with attractive interactions are also attractive for the $(i \sim j \pm 2)$ pairs, and the pattern of ϵ^2 for the charged residues is the same as those of ϵ^0 , i.e., negative values for opposite charges and positive values for the same sign charges. However, the β -strand breaking amino acids Asn and Pro are strongly repulsive for the $(i \sim j \pm 2)$ pairing, but attractive for the $(i \sim j)$ pairing.

The reliability of statistical results

The reliability of the statistical potentials depends on how well the training set represents proteins in general. To estimate the numerical stability of pairwise potentials, ϵ^0 values were calculated for eight datasets where in each case 20 different proteins were randomly removed. Only 17 out of 210 potential pairs have standard

Table 2. The residue-specific cross-strand potential $\epsilon^0 - (i \sim j)$ pairing in β -sheets

	G	A	V	L	I	F	Y	W	M	C	P	S	T	N	Q	H	K	R	D	E
G	-0.29																			
A	-0.14	-0.18																		
V	-0.10	-0.15	-0.48																	
L	-0.04	-0.24	-0.29	-0.43																
I	0.27	-0.25	-0.31	-0.45	-0.48															
F	-0.09	-0.16	-0.31	-0.28	-0.05	-0.50														
Y	-0.21	-0.18	0.00	-0.10	-0.34	-0.27	-0.11													
W	-0.34	-0.01	0.18	-0.18	-0.28	0.16	-0.30	-0.53												
M	0.25	-0.02	-0.02	-0.32	0.21	-0.36	0.01	-0.73	-0.75											
C	-0.42	0.08	0.08	0.36	-0.16	-0.28	0.69	-0.74	0.27	-1.77										
P	0.06	0.28	0.76	0.30	0.99	0.65	-0.02	0.70	-0.78	0.31	-0.78									
S	0.04	0.38	0.18	0.30	0.57	0.15	-0.03	0.44	-0.00	0.12	0.21	-0.68								
T	0.28	0.06	0.19	0.57	0.34	0.25	0.23	0.74	0.43	0.28	0.04	-0.23	-0.58							
N	0.49	-0.04	0.48	0.25	1.45	0.12	-0.14	0.46	-0.52	0.07	0.59	-0.21	-0.06	-0.45						
Q	0.54	0.35	0.41	0.35	0.44	-0.04	-0.06	-0.09	0.07	0.39	0.73	0.19	-0.31	0.20	-0.17					
H	-0.09	0.44	0.37	0.10	0.24	0.25	0.33	-0.34	1.07	-0.45	-0.21	-0.13	-0.22	-0.56	0.28	-0.15				
K	0.56	0.28	0.53	0.37	-0.00	0.75	-0.00	0.02	0.44	0.68	0.26	-0.05	-0.26	-0.27	0.05	0.57	0.21			
R	0.40	0.59	0.43	0.37	0.05	0.31	0.03	-0.20	0.53	0.92	0.34	0.24	-0.31	-0.00	-0.56	-0.11	0.58	-0.03		
D	-0.26	0.24	0.51	0.80	0.26	0.33	0.61	0.74	0.21	0.53	0.87	-0.03	0.32	-0.43	-0.03	-0.61	-0.43	-0.79	0.11	
E	0.21	0.53	0.37	0.51	0.53	0.38	0.25	1.37	0.44	0.17	0.41	0.10	-0.27	0.76	-0.20	-0.14	-1.12	-0.85	0.86	0.58
	G	A	V	L	I	F	Y	W	M	C	P	S	T	N	Q	H	K	R	D	E

deviations of the pairwise potentials ϵ^0 larger than 0.2 on the reduced statistical samples (Table 5). The potential parameters obtained are stable; interactions are significant where the statistical potentials exceed 0.4 (2σ) in absolute value.

Statistical energies in possible alignments of two neighbor strands in β -sheets in the training dataset

We tested whether the statistical energy functions, defined in Equations 5A–C in Methods, would predict energetically preferred align-

ments of two β -strands, and could identify the native alignment as the lowest energy configuration. All possible alignments of a given pair of neighbor strands were exhaustively searched by shifting the two strands relative to each other systematically, and calculating the statistical energy functions using Equations 5A–C. For two β -strands m and n with N_m and N_n residues, there are $(N_m + N_n - 3)$ alignments to be screened assuming that the overlap between the two strands should be at least two residues long. This procedure is illustrated for the protein CD8 (1CD8), a typical Greek key topology with antiparallel β -sheets, and triose phos-

Table 3. The residue-specific first-order cross-strand potential $\epsilon^1 - (i \sim j \pm 1)$ pairing in β -strand

	G	A	V	L	I	F	Y	W	M	C	P	S	T	N	Q	H	K	R	D	E
G	-0.33																			
A	0.01	-0.34																		
V	0.07	-0.23	-0.26																	
L	0.00	-0.13	-0.08	-0.03																
I	0.08	-0.13	-0.16	-0.20	-0.28															
F	0.15	0.07	-0.10	-0.20	-0.14	0.01														
Y	0.30	-0.13	-0.01	-0.06	-0.04	-0.01	-0.20													
W	0.10	-0.05	0.10	0.04	-0.20	0.07	-0.17	0.31												
M	0.55	-0.15	-0.01	-0.07	-0.04	0.13	-0.12	-0.18	-0.30											
C	-0.06	0.59	0.14	0.31	0.22	-0.13	-0.50	-0.07	-0.08	-1.02										
P	0.07	0.35	0.35	0.31	0.42	0.33	0.08	-0.11	-0.25	0.53	1.27									
S	-0.03	0.03	-0.12	-0.29	0.09	-0.26	-0.19	-0.06	-0.11	-0.30	1.01	0.62								
T	0.04	-0.07	-0.17	-0.24	-0.04	-0.19	-0.15	-0.05	-0.14	-0.16	0.95	0.51	0.60							
N	-0.10	0.07	0.00	0.13	0.10	-0.29	0.30	0.90	0.37	-0.40	0.52	0.73	0.53	0.91						
Q	-0.06	1.04	0.06	-0.23	-0.01	-0.24	-0.25	-0.57	0.38	0.05	0.88	0.63	0.23	0.83	0.09					
H	0.13	-0.03	-0.07	0.41	0.05	-0.37	0.02	0.28	-0.50	0.25	-0.06	0.31	0.14	-0.11	0.94	-0.22				
K	0.24	0.02	-0.11	-0.13	-0.08	0.06	-0.09	-0.33	0.09	0.50	0.59	0.80	0.43	1.27	0.50	0.65	0.65			
R	0.21	-0.23	-0.07	-0.04	-0.05	-0.20	-0.09	-0.04	0.20	-0.43	0.01	0.48	0.34	0.35	0.24	0.24	0.46	0.30		
D	0.18	0.29	-0.05	-0.06	-0.01	-0.01	0.22	0.49	-0.11	-0.25	2.41	0.41	0.66	0.79	0.64	-0.01	0.57	1.41	1.47	
E	0.19	0.07	-0.13	-0.32	-0.24	-0.12	0.37	0.11	-0.17	-0.18	1.03	0.64	0.68	0.73	0.55	-0.01	0.14	0.42	0.79	0.57

Table 4. The residue-specific second-order cross-strand potential $\epsilon^2 - (i \sim j \pm 2)$ pairing in β -strand

	G	A	V	L	I	F	Y	W	M	C	P	S	T	N	Q	H	K	R	D	E
G	-0.22																			
A	-0.18	-0.26																		
V	0.12	-0.34	-0.27																	
L	-0.11	-0.13	-0.24	-0.42																
I	0.20	-0.21	-0.20	-0.26	-0.43															
F	-0.23	-0.38	-0.24	-0.43	-0.27	-0.10														
Y	-0.23	-0.37	-0.11	-0.12	-0.03	-0.06	0.35													
W	0.04	0.47	0.01	-0.29	-0.13	-0.14	-0.64	-0.93												
M	-0.17	-0.52	-0.19	-0.04	0.04	-0.37	-0.28	0.99	-1.09											
C	-0.03	0.40	-0.03	-0.05	0.18	-0.26	-0.08	0.06	0.05	-0.91										
P	0.34	0.49	0.52	0.24	0.48	0.40	0.50	0.40	1.13	-0.14	0.94									
S	0.07	0.05	0.08	0.16	0.44	0.17	0.24	0.20	-0.07	0.50	0.43	-0.43								
T	0.37	0.37	-0.01	0.39	0.07	0.31	0.37	0.31	0.28	0.32	0.15	-0.30	-0.57							
N	0.13	0.13	0.25	0.27	0.41	1.26	0.11	0.28	0.04	0.18	1.80	0.09	-0.13	0.58						
Q	0.00	0.27	0.21	0.25	0.31	0.37	0.49	0.42	0.52	1.01	-0.26	-0.50	-0.03	0.32	-0.24					
H	0.23	-0.44	0.17	0.00	0.25	-0.07	-0.14	0.46	0.49	-0.22	0.59	-0.31	0.28	0.07	0.39	0.36				
K	0.34	0.36	0.06	0.17	0.13	0.42	0.33	0.13	0.27	0.25	0.00	0.13	-0.09	0.17	0.39	-0.37	0.61			
R	-0.05	0.46	0.05	0.19	0.54	0.43	-0.42	0.50	0.03	0.11	1.14	0.20	0.17	-0.07	0.35	-0.41	0.23	-0.03		
D	0.14	0.35	0.43	0.19	0.20	0.38	0.19	0.16	0.66	0.64	0.29	0.13	0.08	-0.24	0.09	-0.06	-0.24	-0.49	0.23	
E	0.20	0.25	0.13	0.21	0.31	0.20	-0.18	0.23	0.10	0.87	0.52	0.01	0.02	-0.01	0.67	-0.34	-0.56	-0.51	0.12	0.35
	G	A	V	L	I	F	Y	W	M	C	P	S	T	N	Q	H	K	R	D	E

phate isomerase (5TIM), a typical eight stranded barrel topology with parallel β -sheet structure. Table 6 lists the statistical energy E_2 (Equation 5C) for each alignment. The native alignments are marked with *, and the ranks of the native alignments in the ordered list of energy values are given in the table. Many, but not all of the native alignments, have the lowest E_2 in the list.

Figure 2 shows the distribution of the ranks of the native alignment for all β -sheets in all proteins of the training dataset for all three statistical energies. The native alignments are found as the

lowest energy configuration in 34, 40, and 40% of the cases for E_0 , E_1 , and E_2 , respectively. If we consider the alignments with the three lowest energies as a possible native alignment, the native structures are found in 67, 72, and 73% of the 1,274 β -sheets. All three types of energy display similar distributions for the energy rank. While the cross-strand contact potentials ϵ^0 are primarily responsible for the success of correct prediction of β -sheet alignment, including ϵ^1 and ϵ^2 improves the discriminatory performance of the energy function.

Table 5. The standard deviation of potential ϵ^0 values over eight statistical samples

	G	A	V	L	I	F	Y	W	M	C	P	S	T	N	Q	H	K	R	D	E
G	0.10																			
A	0.05	0.09																		
V	0.04	0.04	0.03																	
L	0.06	0.08	0.03	0.04																
I	0.08	0.04	0.03	0.02	0.05															
F	0.04	0.05	0.03	0.04	0.05	0.04														
Y	0.06	0.07	0.05	0.06	0.05	0.04	0.10													
W	0.07	0.09	0.06	0.09	0.08	0.14	0.05	0.17												
M	0.15	0.05	0.07	0.08	0.07	0.07	0.10	0.16	0.11											
C	0.07	0.09	0.09	0.08	0.10	0.08	0.14	0.13	0.20	0.07										
P	0.13	0.11	0.12	0.06	0.16	0.11	0.09	0.33	0.10	0.14	0.29									
S	0.09	0.15	0.07	0.05	0.08	0.06	0.08	0.28	0.08	0.09	0.15	0.11								
T	0.06	0.06	0.04	0.08	0.05	0.04	0.07	0.11	0.11	0.09	0.10	0.07	0.06							
N	0.12	0.09	0.08	0.09	0.18	0.10	0.10	0.19	0.08	0.14	0.13	0.10	0.11	0.12						
Q	0.30	0.16	0.07	0.11	0.07	0.09	0.06	0.08	0.12	0.18	0.19	0.13	0.09	0.15	0.24					
H	0.09	0.21	0.13	0.06	0.11	0.15	0.11	0.13	0.06	0.16	0.15	0.06	0.09	0.12	0.12	0.34				
K	0.21	0.06	0.10	0.06	0.05	0.08	0.08	0.12	0.11	0.16	0.12	0.06	0.07	0.10	0.13	0.16	0.15			
R	0.19	0.08	0.07	0.13	0.06	0.20	0.10	0.10	0.14	0.25	0.15	0.07	0.07	0.12	0.07	0.12	0.12	0.16		
D	0.10	0.07	0.07	0.12	0.07	0.12	0.12	0.16	0.09	0.21	0.21	0.09	0.12	0.09	0.13	0.10	0.08	0.08	0.13	
E	0.08	0.11	0.08	0.11	0.11	0.10	0.10	0.25	0.14	0.15	0.06	0.05	0.04	0.20	0.09	0.17	0.07	0.07	0.24	0.27
	G	A	V	L	I	F	Y	W	M	C	P	S	T	N	Q	H	K	R	D	E

Table 6. Alignments of β -strand and their corresponding energy expectations

β -sheet alignments ^a	E_2	Rank	β -sheet alignment ^a	E_2	Rank	β -sheet alignment ^a	E_2	Rank	β -sheet alignment ^a	E_2	Rank
1CD8						1CD8 (continued)					
VELKCQV	1.13		FSGKRL	-1.18		GYFCSALS	2.43				
SVRF			LTLVFT			QFLWSC					
VELKCQV	-2.45		FSGKRL	1.11		GYFCSALS	-0.19				
SVRF			LTLVFT			QFLWSC					
VELKCQV	3.13		FSGKRL	-1.36		GYFCSALS	-0.51				
SVRF			LTLVFT			QFLWSC					
VELKCQV	-1.67		FSGKRL	2.24		GYFCSALS	1.28				
SVRF			LTLVFT			QFLWSC					
VELKCQV	4.47		* FSGKRL	-0.73	3	GYFCSALS	-0.71				
SVRF			* LTLVFT			QFLWSC					
* VELKCQV	-2.72	1	FSGKRL	2.25		GYFCSALS	0.93				
* SVRF			LTLVFT			QFLWSC					
VELKCQV	1.91		FSGKRL	0.34							
SVRF			LTLVFT								
VELKCQV	-0.35		FSGKRL	-0.52		5TIM					
SVRF			LTLVFT			IAAAN	-0.85		MVIACI	-0.64	
			FSGKRL	1.25		QCVVA			WIVL		
			LTLVFT			IAAAN	-2.14		MVIACI	-1.59	
						QCVVA			WIVL		
VELKCQV	-0.87		GYFCSALS	-0.13		IAAAN	-1.63		* MVIACI	-3.27	1
LTLVFT			FYMI			QCVVA			* WIVL		
VELKCQV	0.46		GYFCSALS	0.18		* IAAAN	1.12	4	MVIACI	-0.61	
LTLVFT			FYMI			* QCVVA			WIVL		
VELKCQV	-1.99		GYFCSALS	-0.50		IAAAN	2.29		MVIACI	-1.79	
LTLVFT			FYMI			QCVVA			WIVL		
VELKCQV	2.73		GYFCSALS	-1.94		IAAAN	1.88		MVIACI	-1.29	
LTLVFT			FYMI			QCVVA			WIVL		
* VELKCQV	-3.50	1	GYFCSALS	-0.71		IAAAN	1.84		MVIACI	-1.20	
* LTLVFT			FYMI			QCVVA			WIVL		
VELKCQV	4.17		GYFCSALS	-0.90							
LTLVFT			FYMI			IAAAN	-0.89		QCVVA	1.78	
VELKCQV	-2.41		GYFCSALS	-1.99		GFLV			FVIA		
LTLVFT			FYMI			IAAAN	-1.50		QCVVA	0.68	
VELKCQV	2.16		* GYFCSALS	-0.89	5	GFLV			FVIA		
LTLVFT			* FYMI			* IAAAN	-1.98	1	QCVVA	-0.72	
VELKCQV	-1.33		GYFCSALS	-1.19		* GFLV			FVIA		
LTLVFT			FYMI			IAAAN	0.47		* QCVVA	-2.74	1
VELKCQV	0.69		GYFCSALS	-0.62		GFLV			* FVIA		
LTLVFT			FYMI			IAAAN	-0.39		QCVVA	-2.37	
						GFLV			FVIA		
CSWLFQ	-0.92		GYFCSALS	0.21		IAAAN	-0.05		QCVVA	-0.65	
LYLLFT			VPV			GFLV			FVIA		
CSWLFQ	1.45		* GYFCSALS	0.03	2						
LYLLFT			* VPV			MVIACI	-0.37		VVIAY	-0.37	
CSWLFQ	-0.33		GYFCSALS	-0.29		VVIAY			RILY		
LYLLFT			VPV			MVIACI	-1.44		VVIAY	-1.92	
CSWLFQ	-0.59		GYFCSALS	0.99		VVIAY			RILY		
LYLLFT			VPV			MVIACI	-2.98		VVIAY	-1.28	
* CSWLFQ	-3.40	1	GYFCSALS	1.63		VVIAY			RILY		
* LYLLFT			VPV			MVIACI	-1.57		* VVIAY	-0.97	4
CSWLFQ	-2.17		GYFCSALS	0.42		VVIAY			RILY		
LYLLFT			VPV			* MVIACI	-4.57		VVIAY	-1.67	
CSWLFQ	-1.54		GYFCSALS	0.87		* VVIAY			RILY		
LYLLFT			VPV			MVIACI	-1.86		VVIAY	0.02	
CSWLFQ	-1.23		GYFCSALS	1.28		VVIAY			RILY		
LYLLFT			VPV			MVIACI	-0.98				
CSWLFQ	-0.57		GYFCSALS	0.12		VVIAY					
LYLLFT			VPV			MVIACI	-0.25				
						VVIAY					
GYFCSALS	0.48		TFLLYL	1.31							
QFLWSC			AK			RILY	-0.21				
GYFCSALS	1.24		TFLLYL	0.13		GFLV					
QFLWSC			AK			RILY	-1.18				
GYFCSALS	-1.98		TFLLYL	-0.13		GFLV					
QFLWSC			AK			RILY	-0.92				
GYFCSALS	-3.10		* TFLLYL	-0.50	1	GFLV					
QFLWSC			* AK			RILY	-0.77				
GYFCSALS	-0.92		TFLLYL	-0.03		GFLV					
QFLWSC			AK			* RILY	-0.21				
* GYFCSALS	-3.25	1				* GFLV		4			
* QFLWSC											

^aThe asterisk indicates the native alignment in β -sheets.

Fig. 2. The rank distributions of native β -sheets in the entire conformation space for proteins of training dataset. The open, hatched, and solid bars indicate the results of E_0 , E_1 , and E_2 , respectively.

Using pairwise potential ϵ^1 alone as the statistical energy function, the native β -sheets among possible alternative pairing within the three lowest energies are recognized at 59% in the training dataset. Therefore, although the side chains responsible for ϵ^1 are on opposite faces of a β -sheet, it suggests that the amino acid pairs ($i \sim j \pm 1$) are also useful for recognizing the correct pairing of β -strand.

Prediction of the native alignment in the test dataset

The statistical energy functions were then used to predict the alignment of β -sheets in proteins not used to derive the parameters. The native alignments are found as the lowest energy configuration in 31% (E_0), 35% (E_1), and 33% (E_2) of the cases and 62% (E_0), 65% (E_1), and 63% (E_2) within the three lowest energies, values only slightly less than those of proteins in the training set. Figure 3

Fig. 3. The rank distributions of native β -sheets in the entire conformation space for proteins of test dataset. The open, hatched, and solid bars indicate the results of E_0 , E_1 , and E_2 , respectively.

shows the distribution of the rank of native β -sheet alignments in all proteins of the test dataset. This result demonstrates that the derived parameters are not strongly biased toward the proteins of the training set and can be generally used, for example, in the design of novel β -sheet proteins.

Using Z-score to predict native alignments

The Z-score measures the standard deviation of the energy of the native alignment from the mean value over all possible alignments. A good discriminating energy function should have negative Z-scores in most cases. Figure 4 displays the correlation between Z-scores and the ranks of the native alignment in proteins of the training and test datasets. All Z-scores of native β -sheets with rank 1 are negative, and most of them are below -1 . A Z-score criterion, e.g., $Z \leq -1$, serves as another index to recognize the native β -strand pairing from the ensemble of alternative alignments. For the E_2 energies, almost all native β -sheets with $Z \leq -1$ are ranked within the three lowest energies.

Fig. 4. The Z-score vs. rank of native β -sheets with energy E_2 for (A) training dataset and (B) test dataset.

Recognizing tripeptide y - X - y' in native β -sheets with the energy functions

We then tested whether our energy functions could identify tripeptide fragments, which occur frequently in β -sheets of proteins in the training dataset. For each tripeptide y - X - y' there are $20^3 = 8,000$ possible combinations to be a partner in β -sheets. The β -sheet-forming energies E_2 for the whole ensemble were calculated (Fig. 5) and the distribution averaged over all of the frequently occurring tripeptides y - X - y' . The average distribution for native β -sheets found in proteins in the training dataset is also obtained with E_2 . Only a very small fraction of tripeptides in the whole ensemble are in the most favorable energy range where the majority of native β -sheets fall. Table 7 lists the tripeptides y - X - y' for each amino acids X , which have higher frequency occurring in native β -sheets of proteins in training dataset, their corresponding partners in native β -sheets, and favorite candidates as partners, which are obtained with β -sheet-forming energy E_2 through a complete search in the whole ensemble of possible tripeptides. Many partners are similar to those found in native β -sheets.

These statistical results for β -sheet structures are consistent with recent experimental results in host-guest study of β -sheet stability (Smith & Regan, 1995, 1997). The thermal stability measurement of double mutant in the B1 domain of streptococcal protein G showed that pairs Glu-Arg, Glu-Lys, Phe-Phe, Ile-Tyr, and Phe-Tyr are most energetically favored, and pairs Thr-Trp, Thr-Phe, Thr-Tyr, and Thr-Ile least interactive. Our statistical potentials are strongly correlated to the experimental observations. The cross-strand potentials ϵ^0 for the two pairing groups fall in the ranges of $-0.3 \sim -1.1$ and $+0.3 \sim +0.7$, respectively.

Through fully automated sequence selection, a novel protein has been successfully designed with the desired tertiary structure by using zinc finger DNA binding module ($\beta\beta\alpha$ motif) as a template of the de novo design (Dahiyat & Mayo, 1997). The antiparallel β -sheet in the fully designed protein is QYTA/PFTR. Our data show that the pairing patterns **YT** x /**FT** x occur frequently in native β -sheets and were favored energetically in our complete search of tripeptide β -sheets as shown in Table 7. This table should provide

Fig. 5. The energy distributions of β -sheet forming for the tripeptides, which occur most frequently in native proteins. The bar chart is for tripeptide β -sheets found in training dataset. The solid line is the average distribution for all 20^3 possible β -sheets of a give tripeptide, averaged over all frequently occurring tripeptides.

a useful starting point for de novo design of β -sheet architecture. However, as other tertiary interactions must be accounted for, low energy pairings based on the knowledge-based potentials do not guarantee the final fold of the designed protein.

Using β -strand-pairing information in distance geometry calculations to model protein structure

We used the distance constraints for hydrogen bonding based on the β -pairing information to improve the 3D modeling of BPTI with a SECODG-based approach similar to that used for predicting the tertiary structure of small globular helical proteins (Hänggi & Braun, 1994; Mumenthaler & Braun, 1995). This example illustrates that correct β -pairing information significantly improves the 3D modeling of β -sheet proteins, and allows the detection of incorrect β -strand pairing.

The simulated dihedral angle and distance constraints were obtained from X-ray crystal structure as detailed in Methods. For each of the six datasets I–VI described in Methods, 50 structure calculations were started from random initial conformations. The final minimization level, $L = 58$, covers all distance constraints.

The effects of β -strand pairing information on distance geometry calculations of BPTI are illustrated in Figure 6, where 10 structures with lowest target function values are superimposed with their backbone atoms for datasets I, II, and III. Figure 7 shows the correlation between the target function (TF) value and the maximum RMSD of all pairs in the given set of structures, which demonstrates (Widmer et al., 1993) convergence of distance geometry calculations with datasets I, II, and III. For the calculations with dataset I, convergence is low but the correct fold compared to the X-ray structure is obtained. Distance constraints for the correct β -strand pairing gave convergence to a unique and well-defined cluster of correct structures with $\langle \text{RMSD} \rangle \sim 1.7$ Å in the best 10 structures out of the ensemble of 50 structures (dataset II). In the dataset III, with deliberately incorrect β -sheet alignment information, a divergent result is observed in the distance geometry calculations. The incorrect β -strand pairing was effectively indicated by the large number of violated distance constraints. Therefore, a feedback structure-based filter of this type can be used to check the β -strand pairing in β -sheet proteins. Table 8 shows that the results using datasets IV–VI are similar to those of I–III.

Discussion

Our statistical study has indicated several novel facts about the preferential pairing and alignment of proteins in β -sheet proteins. The pairwise potentials are formulated at three different levels, $(i \sim j)$, $(i \sim j \pm 1)$, and $(i \sim j \pm 2)$. The individual pairwise potentials ϵ^0 , ϵ^1 , and ϵ^2 recognize native β -sheets among possible alternative pairings within the three lowest energies in 67, 59, and 57%, respectively, of the cases in the training dataset. The pseudo-energy function based on the sum of these statistical pairwise potentials (Equation 5C) achieved 73 and 63% recognition of the cases in the training and test datasets, respectively. Thus, although the derived statistical potentials and accompanying pseudo-energy functions are related neither to the mean force potentials nor to the free energy change for protein folding, these knowledge-based parameters can discriminate native β -sheets and strand alignments. Our basic assumption that, for a large sample structures, the effects of specific amino acids pairs are averaged out, and the relative frequency of cross-strand residue–residue pairing represents the

Table 7. Tripeptides occurring frequently in native β -sheets, their partners, and favorite candidates of partners obtained through a complete search of all possible tripeptides with energy E_2

y-X-y'	Partners of tripeptides in native β -sheets of training set ^a	Five favorite partners of tripeptides and their corresponding energies E_2
VGI	VIL, DFA, VVA, IVV, TEV, VIS, PAK, DVL	IVL, IAL, VVL, VAL, IWL (-1.88~-1.79)
IGV	IAC, DYV, LLM, ALA, GLV, VEY, QLI, AVF, ECV, VYG, SGV, IKI, TFI	LVI, LAI, LVV, LAV, LWI (-1.88~-1.79)
IGI	LLF, GSL, GGI, YIA, LGI, ATK, FGI, KTN, FDQ, VQV	IWI, LWI, IWL, LWL, ILI (-2.41~-2.12)
VAV	LYA, GIN, VIN, YEV, AVL, VHL, GLF, IAK, YVI, VTR, IFA, CLL, VFA, IDI, IFF, IAV, VDI, LYS	VVV, VAV, VIV, AVV, AAV (-2.63~-2.48)
IAL	FQL, ILV, VCA, RII, YCG, KIK, VQV, KAL, LAL, YAL, ATW	LII, III, LIL, VII, LIA (-2.76~-2.52)
SAS	KLT, RAS, EFV, RGV, AIG, TAS, SAS, ALS, MAS	SLS, SFS, SYS, SCS, SVS (-2.99~-2.54)
AVV	AKY, VIG, LVI, EII, MVV, LVV, FAV, LLE, LVV, LLA, IVV, VST, FVV, IVV, RVV	AVV, VVV, IVV, LVV, FVV (-2.81~-2.56)
VVG	GVI, LGI, LTE, TVI, KYD, LYV, VVL, IAG, VIG, GII, LCV, ILT, IAS	FVA, VVA, FVV, AVA, VVV (-2.03~-1.92)
LVL	VVV, VLV, ALI, KVA, VIA, AVL, FCN, EVL, CFL, LVL, VVL, LTV, GDV, FTH, FVL, AKL	LIL, LFL, LII, IIL, LFI (-2.57~-2.51)
VLL	LVV, AVL, VFE, YVL, LIV, VAC, VIA, FKN, LVT, IAT, QLL, FFI, LLL, FLF, GLL, VLL, KLL, FVR, MLL, VLL	FII, FIF, FIA, VII, FIL (-2.61~-2.47)
VLT	VVI, CLL, KLT, KLT, TLT, LLT, ILA, FYA, NLT, TVA, DVT, TLT, THT, TFT, TLT, AET	TLT, TVT, VLT, VVT, TIT (-2.17~-2.07)
KLT	SAS, EIN, TLT, EKT, SLT, ELT, DFV, KLT, SEK, KII	ELE, EIE, EVE, EWE, ELT (-3.37~-3.12)
VIG	AVV, MIT, VVG, EIV, LIA, LTS, GVP, IVE, LIK, IIT, VIG	FIA, FLA, FVA, FAA, LIA (-1.87~-1.78)
VIA	VLQ, LIG, VVN, VLL, AFL, MVV, VIF, VTL, LVL, VVA, IVL, VVI, IAC, RIL, FVI, SVV, GIA, SIA, SVT, RIA, AIA	VAI, VVI, VAL, VII, VVL (-2.54~-2.47)
TIL	IIV, HVL, CII, LFA, DIL, RIL, LRI, IIL	FLI, FII, ELI, EII, TLI (-1.70~-1.60)
LFV	YQL, IVI, ILD, SIM, AVI, VLP, III, SIV, YVE, IFV, IVV	LFL, LFV, IFL, LVL, IFV (-2.58~-2.42)
LFL	MLG, LDF, IAG, GIE, VAI, AFT, LAF, CFL, RFL, LFL	LFL, LFI, IFL, LFF, FFL (-3.00~-2.64)
TFV	QFV, LFV, TVT, NAE, SYS, MTL, LYV, YFV, DTI	TFV, TFT, TVV, TVT, TLV (-2.16~-1.96)
VYT	TYT, KVG, SYT, SEV, TFS, SVM, RYT, VYT, SYT, SYT	TFT, TIT, VFT, VIT, TAT (-1.83~-1.75)
LYA	VAV, LEF, MYA, LVM, SYA, YYA	MIL, MAL, MIF, MAF, MII (-2.35~-2.19)
LYL	LVV, IQS, LIA, WYL, QFV, LAI, FIN, GYL, AYL, LIK, LYL, SEL	LIL, LFL, LSL, LII, IIL (-2.56~-2.40)
TWK	VSV, TLK, KWK	TWE, EWE, TWT, EWT, TME (-2.62~-2.49)
HWI	AID, QIL, SYQ, TWI	IMA, IMK, CMA, CMK, IMY (-2.40~-2.30)
LMV	YTI, IWI, ISD, GYV, IPF, IGL	MML, MMV, LML, AML, LMV (-2.42~-2.31)
KMV	SMT, FWD, EIV, SMV, VIG	EWI, EWF, EWE, EMH, EWI (-2.61~-2.32)
LCG	FAR, FHS, SGT, FHS	YCC, CCC, FCC, MCC, GCC (-3.84~-3.33)
LCS	AFI, YCS, YVR, SCS	MCC, SCC, YCC, MCS, SCS (-3.19~-3.02)
SCT	LCT, DCT, SAS	SCT, SCS, SCQ, TCT, TCS (-4.53~-4.16)
VPV	YPV, SVV	MMM, VMM, MMV, VMV, FMM (-1.74~-1.50)
LPL	FGL, AII, IPL	MMM, WMM, MMW, WMW, LMM (-2.14~-2.07)
VSV	EHQ, VVT, FSF, TWK, VSV, FSF, VVL, ISV, GTV, HSV, AAV	VSV, VSL, LSV, VSF, FSV (-2.64~-2.59)
VSL	VWI, ISL, FSL, FSL, ISL, DSL, ASL, GGC	FSL, LSL, VSL, FSF, LSF (-3.06~-2.86)
YSF	AFL, GKF, SSF, FSF, VKG, LTV, ISF	LSF, LSM, LSL, LSC, FSF (-2.76~-2.56)
VTV	FSI, VIE, ITQ, SNI, VAY, ETV, HTV, FTV, MTV, LTV, YTV, YTV, IAY, VTF	VTV, VTL, LTV, VTF, FTV (-2.74~-2.58)
VTI	VMV, MND, ITI, GTI, FIV, LTI, DNW, IAL, KTI, AEK	VTL, ITL, LTL, FTL, VEL (-2.56~-2.42)
FTV	LIL, FLS, YTV, TFL, QTV, FVV, VNA, RFM, FTV, YTV, ATV, YTV, ITV, FTV, LAS, VTV, STV	FTL, FTV, LTL, VTL, LTV (-2.83~-2.59)
VNT	VKG, VNT, RCD	VMF, VMA, VMH, VMC, FMF (-1.45~-1.27)
FNV	YYG, NNV, FHF, LAH	FHV, FHF, FHL, FHC, FHI (-2.75~-2.51)
VQV	KTW, VLF, LQV, ITL, LQV, AQV, LDV, LQV, YVT, IQV, LQV	FRF, LRF, FRL, FTF, LRL (-2.29~-2.22)
LQL	GKI, TTI, CQL	LEL, WEL, LEW, LTL, WEW (-3.01~-2.93)
LQI	IKV, LYC, FGL, VYI, IQI	LEW, IEW, LEL, WEW, LEL (-2.82~-2.75)
LHL	VKT, FHL, HHL, IGL, RHL	FEF, FDF, FSF, FTF, FEM (-2.95~-2.73)
SHV	DVH, MHV, AHV	MCM, MCF, MCY, FCM, MWM (-1.90~-1.60)
VKV	IEV, IAA, ATG, FIL, LLV, TKV, LEK, AKV, RAL, YKV, GKV, RKV, FKV	VEV, VEL, LEV, IEV, VEF (-3.09~-2.83)
VKL	IEI, GKL, LKL, LER, CKL, IEA, FTM, MTF	LEL, VEL, LEI, VEI, LEV (-3.21~-3.07)
TKV	HKV, LKV, TVS, EKV, QWV	AEV, VEV, TEV, EEV, AEI (-1.43~-1.16)
VRL	IIV, VEI, VDW, GGC, GTA, ERL, LSE, IDM, WRL, NKV	FEA, FEF, FEL, FEI, VEA (-3.08~-2.92)
LRI	FGC, CIG, VDV, FEV, GDV, IEV	IEL, IEI, FEL, LEL, IEF (-3.25~-3.03)
IRV	IIL, FRV, YDG, TRV, FRV	AEI, AEF, AEV, LEI, IEI (-2.84~-2.74)
VDL	TDL, SVV, CDA, VDL, FDL	VRL, LRL, FRL, VRI, LRI (-2.40~-2.31)
LDI	YKL, ADI, VDD, VVK, KDI	IRL, LRL, IRI, IKL, IRV (-2.69~-2.37)
FDV	MDV, VGE, MDV, LDV, AHG	FRL, FRV, FHL, FHV, MRL (-2.59~-2.49)
AEI	LEI, MEI, IEI, IEI, ALY	IKI, IKL, IRI, IRL, LKI (-3.03~-2.93)
VEI	LIY, VIL, IFI, VRL, LVI, YRL, IEI, VEI, VEA, QEI	IKL, IKI, LKL, VKL, LKI (-3.31~-3.09)
IEV	VKV, VTF, IVA, LEV, RVI, VEV, AEV, AYI, MEV, YKK, EVK	LKI, IKI, LKL, LKV, IKL (-3.31~-3.09)

^aThe italics indicate that the tripeptides form parallel β -sheets with y-X-y'. Because the order of y-X-y' is from N- to C-termini, the tripeptides that form antiparallel β -sheets are from C- to N-termini.

Fig. 6. Effects of including β -strand-pairing information in distance geometry calculations of BPTI. The 10 structures with lowest target function values are superimposed with their backbone atoms: **(A)** I—simulated distance constraints without H-bonding constraints between the β -strands; **(B)** II—simulated distance constraints with correct H-bonding constraints between the β -strands; **(C)** III—simulated distance constraints with wrong β -sheet pairing information.

Fig. 7. The correlation between the target function value (TF) and the maximum RMSD value of all pairs in the given set of structures. It demonstrates the convergence of distance geometry calculations with corresponding data sets: I (solid line), II (dotted line), and III (dashed line).

actual intrinsic inter-residue correlation in β -sheets, appears to be justified by the success of tests to predict the 3D structure of BPTI.

A much simpler potential, in which the pairing energies, analogous to the ϵ^0 , are chosen as -1 for polar–polar and nonpolar–nonpolar amino acid pairs and 0 for others, was also applied to proteins of the test dataset to compare the predictive power. It is found that this simplified potential recognizes native β -sheets in the test dataset as unique lowest energy configuration in 25% of the cases and 41% within the three lowest energies. The success rates of our statistical potentials are 31 and 62%, respectively. The main reason for the lower performance of this simple potential is the high degeneracy in the pseudo-energy spectrum. This illustrates the significance of our energy parameters for the predictive ability of β -sheet formation in proteins.

Formation of a β -sheet requires cross-strand pairwise interaction between two side chains, and is not due to an intrinsic property of a single amino acid (Hecht, 1994; Yang & Honig, 1995; Beasley & Hecht, 1997). Our derived pairwise potentials specify the residue preference of β -strand pairing. However, the derived potential values reflect the side-chain chemistry of the amino acids. The dominant factors determining the pairwise preferences in β -sheets are disulfide bridges, and electrostatic and

hydrophobic compatibility. The potentials were able to recognize the native alignment of β -sheets for representative proteins in both training and test datasets. The potentials can also be used to generate the favorite partners of tripeptides y - X - y' , which occur with high frequency in native proteins in training dataset, providing useful guidance for the de novo design of β -sheet proteins.

Although we made no distinction in this study between parallel and antiparallel β -sheets to obtain pairwise potentials, similar differences of pairing propensities were detected in a study limited to antiparallel β -sheets (Wouters & Curmi, 1995). We found little difference in statistical pairwise potentials derived using data from parallel or antiparallel β -sheets separately (data not included).

Starting from the set of folding patterns of different topologies for antiparallel β -sandwiches, a self-consistent molecular field theory was presented to compute stable folding patterns and conformations for the chains of β -sandwich proteins (Finkelstein & Reva, 1996; Reva & Finkelstein, 1996). The calculated 3D folds were close to the native ones, although some incorrect pairings of β -strands were observed. Native folds for 15 of 17 β -proteins were threaded unambiguously. The native folding pattern was among the most stable ones calculated with the difference in Boltzmann probabilities between “good” and “bad” patterns for a given sequence being approximately four orders of magnitude. The improved computational results could be expected with the use of β -strand pairing potentials in the simulations of protein structures.

Our potentials for pairing amino acid residues in neighboring strands of β -sheets differ in several respects from the contact potentials of Miyazawa and Jernigan derived for amino acid packing in the interior of a protein without regard to the secondary structure by Miyazawa and Jernigan (1996). Their inter-residue contact energies differ intrinsically from our cross-strand potential ϵ^0 . With a small correlation coefficient of 0.23, obviously there is no correlation between these two set of potential energies, e.g., the contacts between opposite charged amino acids are strongly attractive for ϵ^0 , but weakly attractive in the potentials of Miyazawa and Jernigan. This reflects the fact that the charged amino acid pairs are mainly distributed on the surface of 3D structures, and therefore infrequently in contact. However, if charged amino acids occur in β -strands, then their positions across β -strands are highly correlated, as shown in our potentials ϵ^0 .

These potentials are tools to further develop the SECODG programs to predict protein folds ab initio. Although our simple pseudo-energy function detected correct β -sheet structures with a high degree of success, further development is necessary to determine the pairing of strands and distinguish parallel and antiparallel β -sheet structures. The integration of the potentials into our structure based

Table 8. Characterization of the six BPTI datasets described in the text to illustrate the usefulness of β -strand-pairing information and corresponding computational results with SECODG calculations

	I	II	III	IV	V	VI
Upper distance limits	518	540	536	228	250	246
Lower distance limits	707	729	729	228	250	246
Dihedral angle restraints	64	64	64	64	64	64
$\langle \text{RMSD} \rangle$	$2.5 \pm 0.5 \text{ \AA}$	$1.7 \pm 0.2 \text{ \AA}$	$5.8 \pm 2.6 \text{ \AA}$	$2.6 \pm 0.5 \text{ \AA}$	$2.5 \pm 0.5 \text{ \AA}$	$5.8 \pm 2.0 \text{ \AA}$
RMSD to X-ray	1.3–2.7 $\text{\AA}$	1.3–1.8 $\text{\AA}$	3.6–8.8 $\text{\AA}$	1.7–3.2 $\text{\AA}$	1.4–2.6 $\text{\AA}$	2.5–7.9 $\text{\AA}$
Target function value	0.4–3.9	0.9–3.6	17.6–25.5	0.8–3.0	1.2–4.5	5.2–7.9

approach, used here for BPTI, must be further refined and the analysis of violations made more systematic.

Methods

Database preparation

The pairwise statistical potentials for residue-specific correlation in β -sheets were derived from a training data set of 169 X-ray or NMR protein structures deposited in the Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1977). We selected representative structures from each family of different protein folds as classified by (Orengo et al., 1993). Proteins containing only α -helical secondary structures or those containing modified amino acids or prosthetic groups (yeast hexokinase B (2YHX), selenomethionyl ribonuclease (1RNH), and bacteriochlorophyll A (3BCL)) were excluded from the training set. Table 9 lists all proteins in the training data set, the number of amino acids, the number of amino acids within β -strands, and the number of β -strands in the proteins.

A different set of proteins, which were not used in the classification of folds by Orengo et al. (1993), selected according to the SCOP classification (Murzin et al., 1995) of proteins deposited in the PDB in the years 1994 and 1995, was used to assess the success of the statistical potential in recognizing the native alignment of β -sheets. The 67 chains of protein included in the test dataset were: 1ADT, 1AFP, 1ALO, 1AOR, 1ATA, 1BCO, 1BGL, 1BMC, 1BNH, 1BVH, 1CBY, 1CHC, 1CHD, 1CTM, 1CTT, 1DAA, 1DCH, 1DEA, 1DHY, 1DNP, 1ECL, 1FBL, 1FRS, 1GPH, 1GPM, 1GRL, 1GTP, 1HXP, 1ICE, 1KBP, 1KNB, 1KPT, 1LEH, 1LFA, 1LZ6, 1MAS, 1MLA, 1MSA, 1OAC, 1ORD, 1OTF, 1PLS, 1PMC, 1PNG, 1PRE, 1PXT, 1RLR, 1RVV, 1SAT, 1SMN, 1SRS, 1SSO, 1SVB, 1UDG, 1VHH, 1VIH, 1VMO, 1XRC, 1YPT, 1YUA, 2CHS, 2KAUA, 2KAUb, 2KAUc, 2PRF, 3MDD, 3PMG.

Secondary structure determination

The assignment of protein segments as regular secondary structures was done with the hydrogen-bonded pattern recognition algorithm DSSP (Kabsch & Sander, 1983).

Determination of contact cross-strand residue pairs i and j in β -sheet

We identified contacting residues i and j between two neighbor strands in native protein structures through the characteristic hydrogen bonding patterns in antiparallel and parallel β -sheets (Fig. 8) and through side-chain contacts. The criteria for the two contacting residues are different for the antiparallel and parallel β -sheets, and for both cases two different hydrogen bond patterns can occur.

A hydrogen bond $HB_{mn}(i,j)$ between the donor (carbonyl oxygen of amino acid i) of the strand m and the acceptor (peptide nitrogen of amino acid j) of strand n is found in the 3D structures of the training set, if the distance between donor and acceptor atoms is <3.6 Å. If the residues i and j are in contact across the two strands in a regular antiparallel β -sheet (Fig. 8A) either the two hydrogen bonds $HB_{mn}(i-1, j+1)$ and $HB_{mn}(j-1, i+1)$ or the two hydrogen bonds $HB_{mn}(i,j)$ and $HB_{mn}(j,i)$ should be present. If one of these hydrogen bond patterns occurs in a native β -sheet, the two residues are a contact pair across the two strands

m and n . For parallel β -sheets, the hydrogen bond patterns $HB_{mn}(i, j+1)$ and $HB_{mn}(j-1, i)$ or $HB_{mn}(i-1, j)$ and $HB_{mn}(j, i+1)$ are used as a criterion to identify i and j (Fig. 8B). As long as one of the patterns of Figure 8 is determined, the amino acids i and j are identified as a corresponding pair of residues between two β -strands m and n .

Because of the structural deviations of β -sheet structures from regular ideal geometry in experimental 3D protein structures, hydrogen bond patterns are not sufficient to detect all pairs residues in cross-strand contact. The partial list of contact pairs in neighboring strands of β -sheets was therefore extended to include short side-chain contacts. Two amino acids in β -sheet regions are considered as a potential pair if one of their side-chain distances is less than 6.5 Å. If the additional side-chain contacts are consistent with these pairs in their sequential residue numbers, the partial list of contact pairs found from the hydrogen bond pattern is extended. We calculated the difference between the sequential numbers of residues, k and l , which are in side-chain contacts, and those residues, i and j , which are found as contact pairs from the hydrogen bond patterns. If the difference in sequence numbers ($i-k$) in strand m is equal to the difference $j-l$ in strand n , the new β -sheet-forming pair ($k \sim l$) is added to the partial list of contact pairs. To avoid bias in the recognition of contact pairs, we checked the sequential differences starting from multiple reference pairs between the two ends of each β -sheet.

Fig. 8. A: Two patterns of antiparallel β -sheets showing the pairing residues in the H-bonded and non-H-bonded sites. **B:** Two patterns of parallel β -sheets showing the pairing residues.

Table 9. Representative proteins in the training set

Proteins	PDB code	N_t^a	N_β	N_s	Proteins	PDB code	N_t^a	N_β	N_s
Beta: Orthogonal barrel					Alpha/beta: Doubly wound				
Intestinal fatty acid binding protein	1lfc	131	77	10	Ras P21 protein	5p21	166	45	6
Retinol binding protein	1rbp	175	82	9	Elongation factor Tu	1etu	177	36	6
Bilin binding protein	1bbpA	173	83	10	Flavodoxin (Clostridium MP)	4fxn	138	30	6
Beta: Super barrel					Flavodoxin (<i>Desulfovibrio vulgaris</i>)	2fx2	147	37	8
Porin	2por	301	172	16	Flavodoxin (<i>Chondrus crispus</i>)	2fcr	173	33	8
Beta: Greek key					Chey protein	3chy	128	22	5
Beta trypsin	4ptp	219	68	13	Carboxypeptidase A	5cpa	307	48	8
Trypsin	1sgt	213	75	14	Thioredoxin (<i>Escherichia coli</i>)	2trx	108	30	5
Proteinase A	2sga	168	93	16	Thioredoxin (human)	3trx	105	27	5
Alpha-lytic protease	2alp	170	82	15	Glutaredoxin	1lego	85	22	4
Sindbis virus capsid protein	1snw	151	73	14	Glutathione peroxidase	1gp1	184	30	7
Immunoglobulin	2rhe	114	49	9	Subtilisin Carlsberg	1cseE	274	47	9
CD8	1cd8	114	53	10	Dihydrofolate reductase (<i>E. coli</i> B)	4dfrA	159	48	8
CD4	3cd4	178	82	20	Dihydrofolate reductase (chicken)	8dfr	186	59	11
Immunoglobulin FAB	221	102	19		Dihydrofolate reductase (<i>Lactobacillus casei</i>)	3dfr	162	53	10
Human class I histocompatibility antigen	3hlaB	270	101	15	Adenylate kinase	3adk	194	25	5
Amylase inhibitor	1hoe	74	34	6	Guanylate kinase	1gky	186	41	9
Actinoxanthin	1acx	107	45	10	Phosphoglycerate mutase	3pgm	230	20	5
Superoxide dismutase	1cobA	151	60	8	Alpha/beta: Two doubly wound domains (multidomain)				
Pseudoasurin	1paz	120	44	8	Rhodanese	1rhd	293	34	10
Plastocyanin	2pcy	99	36	8	Phosphofructokinase	4pfr	319	59	11
Asurin	2azaA	129	46	8	Phosphoglycerate kinase	3pgk	415	55	15
Prealbumin	2pabA	114	58	9	Glucose binding protein	2gbp	309	57	12
Gamma crystallin	2gcr	173	68	14	Arabinose binding protein	8abp	305	63	12
Satellite tobacco necrosis virus	2stv	184	82	9	Leucine/isoleucine/valine binding protein	2liv	344	66	14
Bean pod mottle virus	1bmV1	185	71	12	Alpha/beta: One doubly wound (multidomain)				
	1bmV2	374	136	20	Glutathione reductase	3grs	461	111	23
Poliovirus	2plv1	283	76	9	Thioredoxin reductase	1trb	316	89	21
	2plv2	268	95	14	Catalase	8cat	498	75	9
	2plv3	235	69	12	Lactate dehydrogenase	6ldh	329	53	12
Tumour necrosis factor	1tnf	152	69	10	Malate dehydrogenase	4mdh	333	63	14
Mengo encephalomyocarditis virus coat protein	2mev1	268	68	11	Isopropylmalate dehydrogenase	1ipd	345	65	12
	2mev2	249	94	12	Phosphorylated isocitrate dehydrogenase	4icd	414	74	18
	2mev3	231	70	12	6-Phosphogluconate dehydrogenase	2pgd	473	38	10
Southern bean mosaic virus	4sbv	199	69	10	Alcohol dehydrogenase	8adh	374	89	18
Tomato bushy stunt virus	2tbv	287	115	18	Glyceraldehyde phosphate dehydrogenase	1gd1	332	95	20
Concanavalin A	2cna	237	103	16	Aspartate aminotransferase	7aat	401	56	13
Pea lectin	2ltm	181	60	11	Tyrosyl-transfer RNA synthetase	2ts1	317	28	6
Beta: Complex sandwich					P-hydroxybenzoate hydroxylase	1phh	394	98	22
Endothia aspartic protease	2er7	320	122	28	Alpha/beta: Mainly alpha				
Pepsinogen	2psg	326	137	25	Lysozyme (bacteriophage T4)	3lzm	164	14	3
HIV-1 protease	5hvpA	99	52	7	Lysozyme (human)	1lz1	130	10	3
Rous sarcoma virus protease	2rspA	115	41	9	Prophospholipase	4bp2	117	8	2
Phosphocarrier III	1f3g	150	63	14	Alpha/beta: Sandwich				
16th complement control protein	1hcc	59	19	7	Selenomethionyl ribonuclease	1rnh	148	44	5
Beta: Trefoil					Ribonuclease H domain of HIV reverse transcriptase	1hrh	125	36	5
Basic fibroblast growth factor	2fgf	127	49	10	Eco Rv endonuclease	1rvC	244	65	11
Interleukin 1-beta	8ilb	146	74	13	Ribonuclease A	7rsa	124	41	7
Beta: Propeller					Ribonuclease T1	9rnt	104	28	7
Neuraminidase sialidase	1nsb	390	182	25	Ribonuclease SA	2sar	96	19	5
Beta: Metal rich					Barnase	1rnb	109	27	6
Rubredoxin	5rxn	54	8	3	Mannose binding protein	1msb	115	36	7
High potential iron sulfur protein	2hip	71	12	3	FK506 binding protein	1fkf	107	41	6
Beta: Disulfide rich					Verotoxin	1bov	69	34	7
Bowman Birk proteinase inhibitor	1pi2	61	14	5	Staphylococcal nuclease	1snc	135	42	8
Sea anemone toxin	1atx	46	16	4	Interleukin 8	1il8	71	18	3
Neurotoxin	1sh1	48	25	4	Monocyte chemo-attractant and activating protein	1mca	68	17	3
Erabutoxin	3ebx	62	27	5	Ubiquitin	1ubq	76	26	5
Epidermal growth factor	1epg	53	6	2	Ferredoxin (blue green algae)	1fxi	96	27	5
Transforming growth factor	4tgf	50	16	4	Protein G	2gb1	56	22	4
Alpha/beta: TIM barrel					Crambin	1crn	46	6	2
Triosephosphate isomerase	5tim	249	37	8	L7/L1250 S ribosomal protein	1ctf	68	19	3
Glycolate oxidase	1gox	350	44	12	Acetylphosphatase	1aps	98	37	5
Aldolase	1ald	363	52	10	Ferredoxin II (<i>Desulfovibrio gigas</i>)	1fdx	58	12	4
Flavocytochrome B2	1fcb	494	56	15	Ferredoxin (<i>Bacillus thermoproteolyticus</i>)	2fxb	81	14	4
Anthranyl isomerase	1pii	452	90	22	Ferredoxin (<i>Azotobacter vinelandii</i>)	5fd1	106	14	4
Xylose isomerase	6xia	387	32	10	Ferredoxin (<i>Peptococcus aerogenes</i>)	1fdx	54	4	2
Rubisco	5rub	436	77	15	Subtilisin inhibitor (<i>Bacillus amyloliquefaciens</i>)	2sic	107	34	5
Taka-amylase A	2taa	478	72	20					
Enolase	4enl	436	72	15					

(continued)

Table 9. Continued

Proteins	PDB code	N_i^a	N_β	N_s	Proteins	PDB code	N_i^a	N_β	N_s
Alpha+beta: Sandwich (continued)					Multidomain (continued)				
Thymidilate synthase	2tsc	264	65	10	Acetyl cholinesterase	lace	526	82	16
Chloramphenicol acetyltransferase	4cla	213	61	10	Cholesterol oxidase	3cox	500	102	19
Subtilisin inhibitor (<i>Bacillus subtilis</i>)	1csei	63	21	4	C-Amp-dependent protein kinase	2cpk	336	49	9
Chymotrypsin inhibitor	2ci2	65	14	3	Fructose 1,6-bisphosphatase	1fbp	317	75	13
Inorganic pyrophosphatase	1pyp	281	28	5	Ferredoxin (spinach)	1fnd	296	83	13
Alpha+beta: Beta open sheet disulfide rich					Isoenzyme 3-3 of glutathione S-transferase	1gst	217	20	4
Trypsin inhibitor	5pti	58	14	2	Guanylate kinase	3gly	470	24	9
Kringle 2	1tpk	85	8	2	Leucine aminopeptidase	1lap	481	102	21
Wheat germ agglutinin	9wga	170	16	8	Ovalbumin	1ova	367	118	17
Scorpion neurotoxin	2sn3	65	12	3	Modified alpha-1 antitrypsin	7api	339	112	12
Ovomucoid third domain	2ovo	56	9	3	Variant surface glycoprotein	1vsg	362	27	10
Trypsinogen inhibitor	1tgs	55	9	3	Lactoferrin	1lfi	691	123	27
Proteinase inhibitor	3sgb	50	9	3	Tryptophan synthase	1wsy	385	66	12
Alpha+beta: Metal rich					Cytochrome C peroxidase	2cyp	293	16	5
Cytochrome c3	2cdv	107	10	4	Glutamine synthetase	2gls	468	78	17
Cytochrome b5	3b5c	86	19	5	Phosphoglucomutase	3pmg	561	142	30
Multidomain					Aconitase	8acn	753	143	32
Papain	9pap	212	36	8	RecA protein	2reb	303	77	14
Beta lactamase	3blm	257	48	9	Thermolysin	6tmn	316	58	14
Cytochrome P450Cam	2cpp	405	41	14	Hemagglutinin	1hgea	328	117	29
Aspartate carbamoyl transferase	8atc	310	44	9		1hgeb	175	20	4
Citrate synthase	1csc	429	6	2	Carbonic anhydrase	2ca2	256	76	15
					Gene regulatory protein	3gap	208	25	5
					Photosynthetic reaction centre	1prc	258	54	13

^a N_i , number of amino acids; N_β , number of amino acids within β -strands; N_s , number of β -strands.

For β -sheets containing β -bulge structures (Richardson, 1981), the criterion of sequential differences is not satisfied. Therefore, β -bulges occurring at the ends of a β -sheet are skipped as potential contact pairs, while β -bulges in the middle decompose the β -sheet into two parts. The longer strand in the sheet is treated as two strands with a break in the β -strand at the location of the bulge. This separation does not substantially influence the list of contact pairs in β -sheet structures.

The list of contact pairs found by this automatic procedure was checked by visual inspection in native structures of proteins, and found to be consistent with an intuitive identification of contact residue pairs in antiparallel and parallel β -sheets. The side-chain contacts are specific for interactions *across* two β -strands, and are thus different from other amino acid pair-specific potentials (Kocher et al., 1994; Mirny & Shakhnovich, 1996; Park & Levitt, 1996), which were derived from short through-space contacts in native proteins.

Contact pairs and residue specific potentials

Contact pairs in β -sheets of all 169 proteins in the training set were obtained. The frequency of all contact pairs for each of the 210 amino acids pairs was determined, and compared to the frequencies of occurrence, which one would expect if all amino acids were randomly and independently distributed along all β -strands. The numbers of contact pairs of residues with the amino acid types X and Y at corresponding positions i and j (Fig. 1) are denoted as $N_{i,j}^{XY}$. N_{pairs}^{XY} is the total number of all contact pairs of all proteins in the whole data set. Thus, the frequency of a contact pair for given residues X and Y is

$$f_{i,j}^{XY} = \frac{N_{i,j}^{XY}}{N_{pairs}^{XY}}. \quad (1)$$

We obtain the pair correlation $p_{i,j}^{XY}$ for each residue pair X and Y as

$$p_{i,j}^{XY} = \frac{f_{i,j}^{XY}}{s_{XY} \cdot f_\beta^X \cdot f_\beta^Y} \quad (2)$$

where f_β^X is the average frequency of a residue in all β -strands of our training set (Table 9). The symmetry factor s_{XY}

$$s_{XY} = \begin{cases} 1 & \text{if } X = Y, \\ 2 & \text{if } X \neq Y \end{cases} \quad (3)$$

accounts for the fact that we do not distinguish the order of the pair X and Y , i.e., we include both cases in $N_{i,j}^{XY}$, if residue pair XY or pair YX is present in the locations i and j . The pair correlation $p_{i,j}^{XY}$ shows that a pair of residues X and Y is preferentially in contact in a β -sheet, if $p_{i,j}^{XY} > 1$. If $p_{i,j}^{XY} = 1$ the residue pair X and Y does not have a specific preference to be present as a contact pair in a β -sheet.

We define the statistical contact energy between amino acids X and Y by

$$\epsilon_{i,j}^{XY} = -k_B T \ln(p_{i,j}^{XY}). \quad (4)$$

In contrast to most studies (Bryant & Lawrence, 1991, 1993; Finkelstein et al., 1993, 1995), we do not interpret this contact energy as a potential of mean force between the side chains of both residues, but rather as a convenient statistical tool to discriminate native pairing. The contact energy includes not only the specific physical interactions between the two residues, but also general global properties of the geometry and the physical chemical nature of native β -sheets.

This is evident for other types of contact energies, which we define. In addition to the contact energy ϵ^0 , we also derive the statistical potentials ϵ^1 and ϵ^2 (Fig. 1). The statistical potential ϵ^1 represents correlation between the residues i and $j \pm 1$, which are pairs of residues next in sequence to the contact pair i and j , but not necessarily next in cross-strand distances. The pair correlation and the statistical energies $\epsilon_{i,j\pm 1}^{XY}$ for these first-order “contact” pairs were calculated with the same procedure as described above. Both cases $j \pm 1$ were treated as one group due to scarcity of the data. The interactions of the statistical potential ϵ^0 includes the side chain contacts on one face of a β -sheet, whereas the correlation of the statistical potential ϵ^1 is between side chains on opposite faces of a β -sheet. We therefore expect a qualitatively different preference of amino acid pairs expressed in potentials ϵ^0 than in the statistical potentials ϵ^1 . Similarly, we calculate the second-order contact energies ϵ^2 , which are the interactions of residue i to the (sequentially) third nearest neighbor $j \pm 2$.

The energy of a given alignment in a β -sheet is then calculated as a sum of the pairwise energies, which are based on the pairwise potentials defined by Equation 4. Three statistical potentials with increasing levels of interactions are evaluated for their ability to recognize the native alignment, E_0 , E_1 , and E_2 .

$$E_0 = \sum_{i \sim j} \epsilon_{i,j}^{XY} \quad (5A)$$

$$E_1 = \sum_{i \sim j} \epsilon_{i,j}^{XY} + \sum_{i \sim j \pm 1} \epsilon_{i,j \pm 1}^{XY} \quad (5B)$$

$$E_2 = \sum_{i \sim j} \epsilon_{i,j}^{XY} + \sum_{i \sim j \pm 1} \epsilon_{i,j \pm 1}^{XY} + \sum_{i \sim j \pm 2} \epsilon_{i,j \pm 2}^{XY}. \quad (5C)$$

Using the Boltzmann approach as shown in Equation 4, we assume that the interactions ϵ^0 , ϵ^1 , and ϵ^2 can be treated as independent from each other. These energy values should provide us with an estimate for the tendency of two β -strands to align in a sequence specific way. We describe and analyze these residue-specific pairwise potentials in Results and then assess the “energy functions” (Equations 5A–C) using our training and test protein data sets for recognizing native β -sheet structures.

Statistical fluctuations of the energy potentials

Some of the amino acid pairs occur rarely in the data set. We therefore analyzed the accuracy of the statistical potentials and robustness of the energy potential values derived by randomly removing some proteins from the training set. The statistical analyses were repeated for several data sets where 20 proteins were randomly removed, and the pairwise potential $\epsilon_{i,j}^{XY}$ was determined. The standard deviations of the energy functions are then calculated with

$$s^{XY} = \sqrt{\frac{1}{N-1} \sum_{k=1}^N (\epsilon_k^{XY} - \langle \epsilon^{XY} \rangle)^2} \quad (6)$$

where ϵ_k^{XY} is the pairwise potential obtained from the statistical study on the dataset k , and $\langle \epsilon^{XY} \rangle$ is the mean potential over all statistical samples.

Definition of Z-scores and rank

We used the Z-score (Bowie et al., 1991; Godzik et al., 1992; Bryant & Lawrence, 1993) to test whether the native pairing of β -strands can be consistently discriminated from all alternative alignments with low energy values. The Z-score gives the deviation of the energy of the native conformation from the mean energy value of all conformations in the ensemble of alternative alignments in the unit of the standard deviation,

$$Z = \frac{E_{\text{native}} - \langle E \rangle}{\sigma} \quad (7)$$

where E_{native} is the energy of the native alignment of two strands in a β -sheet, $\langle E \rangle$ is the mean energy value for all possible alignments between two strands in the β -sheet, and σ is the standard deviation of the energy distribution.

To quantify the performance of energy functions in the discrimination of the native pair of β -strands, we calculate the energies of all configurations for a given β -sheet, and then look for the rank score, the position of the native alignment in the sorted list of energies. The rank score was determined with three different energy levels by using Equations 5A–C, i.e., the energy includes: (1) the cross-strand pairing contributions ($i \sim j$); (2) the cross-strand pairing ($i \sim j$) and the first-order pairing ($i \sim j \pm 1$) contributions; (3) the cross-strand pairing ($i \sim j$), the first-order pairing ($i \sim j \pm 1$), and the second-order pairing ($i \sim j \pm 2$) contributions.

3D modeling with distance geometry calculations

Distance geometry calculations in torsion angle space were done with the program DIAMOD (Hänggi & Braun, 1994) on a SGI/R10000 workstation. Hydrogen bond patterns generate very restrictive distance constraints to be included in distance geometry calculations for a given possible alignment of two strands in a β -sheet (Fig. 8). We incorporated these constraints to calculate the 3D structure of BPTI with the SECODG-based program DIAMOD (Hänggi & Braun, 1994; Mumenthaler & Braun, 1995). The backbone dihedral angles in the α -helical segments Asp3–Leu6 and Ala48–Cys55 and in the β -strands Ile18–Asn24 and Leu29–Tyr35 are constrained by $(-62, -52) = (\varphi, \phi) = (-52, -42)$, and $(-140, 100) = (\varphi, \phi) = (-100, 140)$, respectively. A wide range of dihedral angles in β -strands was used to satisfy the high twisted β -sheet conformation in BPTI. In addition, ϕ angles of all other residues Arg, Glu, Tyr, Thr, and Phe are restricted to $-180 < \phi < 0$, following a recent statistical study (Kamimura & Takahashi, 1994). Distance constraints simulated from the known 3D structure (4PTI) were used to prepare two types of datasets. For the first group (Datasets I–III), distances between reference atoms for side chains separated in the sequence by more than two residues were constrained with a uniform bound of 12 Å. Those distances, which are smaller than 12 Å in the crystal structure, were constrained with an upper bound of 12 Å, otherwise with a lower bound. The reference atoms used to represent the residue side chains are described elsewhere (Hänggi & Braun, 1994; Mumenthaler & Braun, 1995). In addition, three disulfide bridges: Cys5–Cys55, Cys14–Cys38, and Cys30–Cys51 are restricted with standard bond length of $1.9 \text{ Å} < d_{S-S} < 2.1 \text{ Å}$. The input dataset I for the BPTI 3D simulation consists of all these constraints. The calculations of β -sheet-forming energy indicate that the antiparallel β -sheet pattern from Ile18 and Val34 to Tyr23 and Leu29 has the lowest

energy ($E_2 = -1.8$) while the native one (aligned from Ile18 and Tyr35 to Asn24 and Leu29) ranks second ($E_2 = -0.8$). Accordingly, two additional input datasets (II: native β -sheet pattern; III: β -sheet pattern with the lowest energy for β -sheet-forming but not present in the native structure) were generated by adding the distance constraints for β -strand-pairing to dataset I.

Another group of input datasets (IV–VI) was created based on hydrophobic clusters in BPTI (Wüthrich et al., 1980). One hydrophobic core consists of the side chains of residues Arg20, Phe33, Tyr35, and Asn44, and the other consists of the side chains of residues Phe4, Cys5, Tyr23, Cys30, Asn43, Phe45, Cys51, and Cys55. Twenty amino acids of BPTI were determined to be solvent-exposed since their solvent-accessible surface area of side-chain atoms in the tertiary structure was more than 50% of average value in the tripeptide Gly-X-Gly in an ensemble of 30 random conformations for the residue X. All distances between the reference atoms of the residues in each group are then extracted and used as upper and lower limits with ± 0.5 Å gap. Short-, medium-, and long-range distance constraints were defined for residues within the same hydrophobic core, between two hydrophobic cores, or between solvent-exposed residues, respectively. Accordingly, input datasets (IV–VI) were generated: IV—the upper and lower distance constraints with three disulfide-bonded bridges: Cys5–Cys55, Cys14–Cys38, and Cys30–Cys51; V—dataset IV with native β -sheet pattern; VI—dataset IV with incorrect β -sheet pattern. The dihedral angle constraints were identical in the six datasets I–VI.

Acknowledgments

This work was supported by the National Science Foundation: DBI-9632326, Department of Energy: DE-FG03-96ER62267, and John Sealy Memorial Endowment Fund: 454940. We thank Dr. Catherine H. Schein for a careful reading of the manuscript.

References

- Beasley JR, Hecht MH. 1997. Protein design: The choice of *de novo* sequences. *J Biol Chem* 272:2031–2034.
- Bernstein FC, Koetzle TF, Williams GJ, Meyer EFJ, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem* 80:319–324.
- Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253:164–170.
- Bryant SH, Lawrence CE. 1991. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. *Proteins* 9:108–119.
- Bryant SH, Lawrence CE. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins* 16:92–112.
- Chou PY, Fasman GD. 1978. Empirical predictions of protein conformation. *Ann Rev Biochem* 47:251–276.
- Creighton TE, Darby NJ, Kemmink J. 1996. The roles of partly folded intermediates in protein folding. *FASEB J* 10:110–118.
- Dahiyat BI, Mayo SL. 1997. De novo protein design: Fully automated sequence selection. *Science* 278:82–87.
- Finkelstein AV, Badretidinov AY, Gutin AM. 1995. Why do protein architectures have Boltzmann-like statistics? *Proteins* 23:142–150.
- Finkelstein AV, Gutin AM, Badretidinov AY. 1993. Why are the same protein folds used to perform different functions? *FEBS Lett* 325:23–28.
- Finkelstein AV, Reva BA. 1996. Search for the most stable folds of protein chains: I. Application of a self-consistent molecular field theory to a problem of protein three-dimensional structure prediction. *Protein Eng* 9:387–397.
- Godzik A, Kolinski A, Skolnick J. 1992. Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol* 227:227–238.
- Hänggi G, Braun W. 1994. Pattern recognition and self-correcting distance geometry calculations applied to myohemerythrin. *FEBS Lett* 344:147–153.
- Harrison PM, Sternberg MJE. 1996. The disulphide β -cross: From cystine geometry and clustering to classification of small disulphide-rich protein folds. *J Mol Biol* 264:603–623.
- Hecht MH. 1994. De novo design of β -sheet proteins. *Proc Natl Acad Sci USA* 91:8729–8730.
- Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
- Kamimura M, Takahashi Y. 1994. Phi-psi conformational pattern clustering of protein amino acid residues using the potential function method. *CABIOS* 10:163–169.
- Kim CA, Berg JM. 1993. Thermodynamic β -sheet propensities measured using a zinc-finger host peptide. *Nature* 362:267–270.
- Kim CA, Berg JM. 1996. A 2.2 Å resolution crystal structure of a designed zinc finger protein bound to DNA. *Nature Struct Biol* 3:940–945.
- Kocher JP, Rooman MJ, Wodak SJ. 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol* 235:1598–1613.
- Levitt M. 1978. Conformational preferences of amino acids in globular proteins. *Biochemistry* 17:4277–4285.
- Lifson S, Sander C. 1980. Specific recognition in the tertiary structure of beta-sheets of proteins. *J Mol Biol* 139:627–639.
- Maierov VN, Crippen GM. 1992. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol* 227:876–888.
- Minor DLJ, Kim PS. 1994a. Measurement of the β -sheet-forming propensities of amino acids. *Nature* 367:660–663.
- Minor DLJ, Kim PS. 1994b. Context is a major determinant of β -sheet propensity. *Nature* 371:264–267.
- Mirny LA, Shakhnovich EI. 1996. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol* 264:1164–1179.
- Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* 18:534–552.
- Miyazawa S, Jernigan RL. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256:623–644.
- Mumenthaler C, Braun W. 1995. Predicting the helix packing of globular proteins by self-correcting distance geometry. *Protein Sci* 4:863–871.
- Munoz V, Serrano L. 1994. Intrinsic secondary structure propensities of the amino acids, using statistical ϕ - ψ matrices: Comparison with experimental scales. *Proteins* 20:301–311.
- Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.
- Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng* 6:485–500.
- Otzen DE, Fersht AR. 1995. Side-chain determinants of β -sheet stability. *Biochemistry* 34:5718–5724.
- Park B, Levitt M. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol* 258:367–392.
- Reva BA, Finkelstein AV. 1996. Search for the most stable folds of protein chains: II. Computation of stable architectures of β -proteins using a self-consistent molecular field theory. *Protein Eng* 9:399–411.
- Richardson JS. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem* 34:167–339.
- Seckler R, Jaenicke R. 1992. Protein folding and protein refolding. *FASEB J* 6:2545–2554.
- Sippl MJ. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 213:859–883.
- Sippl MJ. 1993. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput-Aid Mol Des* 7:473–501.
- Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 5:229–235.
- Skolnick J, Jaroszewski L, Kolinski A, Godzik A. 1997. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci* 6:676–688.
- Smith CK, Regan L. 1995. Guidelines for protein design: The energetics of beta sheet side chain interactions. *Science* 270:980–982.
- Smith CK, Regan L. 1997. Construction and design of β -sheets. *Acc Chem Res* 30:153–161.
- Smith CK, Withka JM, Regan L. 1994. A thermodynamic scale for the β -sheet forming tendencies of the amino acids. *Biochemistry* 33:5510–5517.
- Tanaka S, Scheraga HA. 1976. Medium- and long-range interaction parameters

- between amino acids for predicting three-dimensional structure proteins. *Macromolecules* 9:945–950.
- Widmer H, Widmer A, Braun W. 1993. Extensive distance geometry calculations with different NOE calibrations. New criteria for structure selection applied to Sandostatin and BPTI. *J Biomol NMR* 3:307–324.
- Williams RW, Chang A, Juretic D, Loughran S. 1987. Secondary structure predictions and medium range interactions. *Biochim Biophys Acta* 916:200–204.
- Wouters MA, Curmi PM. 1995. An analysis of side chain interactions and pair correlations within antiparallel β -sheets: The differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins* 22:119–131.
- Wüthrich K, Wagner G, Richarz R, Braun W. 1980. Correlation between internal mobility and stability of globular proteins. *Biophys J* 32:549–560.
- Yang AS, Honig B. 1995. Free energy determinants of secondary structure formation: II. Antiparallel beta-sheets. *J Mol Biol* 252:366–376.