JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

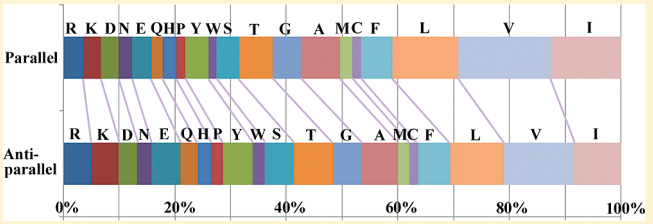
Parallel and Antiparallel β-Strands Differ in Amino Acid Composition and Availability of Short Constituent Sequences

Motosuke Tsutsumi and Joji M. Otaki*

The BCPH Unit of Molecular Physiology, Department of Chemistry, Biology and Marine Science, University of the Ryukyus, Nishihara, Okinawa 903-0213, Japan

Supporting Information

ABSTRACT: One of the important secondary structures in proteins is the β-strand. However, due to its complexity, it is less characterized than helical structures. Using the 1641 representative three-dimensional protein structure data from the Protein Data Bank, we characterized β-strand structures based on strand length and amino acid composition, focusing on differences between parallel and antiparallel β-strands. Antiparallel strands were more frequent and slightly longer than parallel strands. Overall, the majority of β-sheets were antiparallel sheets; however, mixed sheets were reasonably abundant, and parallel sheets were relatively rare. Notably, the nonpolar, aliphatic hydrocarbon amino acids, valine, isoleucine, and leucine were observed at a high frequency in both strands but were more abundant in parallel than in antiparallel strands. The relative amino acid occurrence in β-sheets, especially in parallel strands, was highly correlated with amino acid hydrophobicity. This correlation was not observed in α-helices and 3_{10}-helices. In addition, we examined the frequency of 400 amino acid doublets and 8000 amino acid triplets in β-strands based on availability, a measurement of the relative counts of the doublets and triplets. We identified some triplets that were specifically found in either parallel or antiparallel strands. We further identified "zero-count triplets" which did not occur in either parallel or antiparallel strands, despite the fact that they were probabilistically supposed to occur several times. Taken together, the present study revealed essential features of β-strand structures and the differences between parallel and antiparallel β-strands, which can potentially be applied to the secondary structure prediction and the functional design of protein sequences in the future.



1. INTRODUCTION

Two frequent secondary structures in proteins are α-helices[1] and β-sheets;[2] therefore, their understanding in protein three-dimensional structure and function cannot be overemphasized. Scientists have been attempting to characterize and predict secondary structure regions in protein chains since the early 1970s.[3−6] Whereas α-helices are relatively simple in structure, with a repetitive helical arrangement of amino acids, β-sheets are composed of a pleated collection of β-strands with either a parallel or an antiparallel arrangement of amino acids. Because a single sheet can be composed of several strands that are not necessarily adjacent parts of a protein chain, structural analyses of β-sheets are usually more complex than those of α-helices. To add further complexity, the formation of β-sheets is context-dependent and is not entirely based on intrinsic amino acid sequences.[7−9]

Energetically, a single β-strand is unstable. However, as a collection of strands, a β-sheet is stabilized by hydrogen bonds between strands. Such a hydrogen bond is formed between polar CO and NH groups of different strands, as commonly described in major biochemistry textbooks.[10−12] On the other hand, β-sheets are often observed in the hydrophobic core of β-domain proteins.[12] Consistent with this, at the population level, compositional analyses have shown that there is an increased frequency of hydrophobic amino acids in β-sheets.[4,13,14] In the β-sheet configuration, amino acid residues are arranged vertically to the sheet plains. Thus, the surface of a β-sheet is likely to be mainly hydrophobic.

In addition to the general abundance of hydrophobic residues in β-sheets, the differences between parallel and antiparallel strands in amino acid composition have been indicated empirically.[15,16] Furthermore, determinants for parallel or antiparallel strands have been sought to predict parallel and antiparallel orientation in a protein chain; several studies have focused on interstrand amino acid pairs[18−21] and interactions between side chains[22−24] that contribute to stabilization of β-sheets. However, the original studies to differentiate parallel and antiparallel strands were performed in the 1970s with a very small number of protein structure samples (e.g., 30 samples in Lifson and Sander[16]), and the reliable Kyte-Doolittle hydrophobicity index was invented later, in 1982.[17]

Presently, in the postgenome era, publicly available structure records have accumulated in the Protein Data Bank (PDB).[25] As of December 12, 2008, there were 50,507 protein structure records in the PDB. In our previous study, we collected a statistically rigorous number of representative samples (1641 protein chains) from the PDB and constructed and characterized

Received: January 20, 2011
Published: April 26, 2011

four kinds of the secondary structure databases from the full-length database (FL-DB): α-helix, 3_{10}-helix, β-strand, and "other" databases.[14] Based on these records, we have characterized secondary structures based on compositional analysis at the amino acid level. Moreover, we performed secondary structure analyses using modern computational technologies, that is, a comprehensive search for constituent short amino acid sequences (amino acid doublets and triplets).[26−29] Our strategy is to exhaustively search for the 400 possible amino acid doublets and the 8000 possible amino acid triplets in the protein sequences and to examine their absolute counts (occurrence) and relative counts (defined as "availability") in a defined structure.

Here, we characterized parallel and antiparallel β-strands in terms of strand length, amino acid composition, and availability to depict the whole picture of β-strands. Importantly, we discovered a clear tendency toward more hydrophobic residues in parallel than in antiparallel β-strands. We also discovered amino acid triplets that could serve as markers for parallel or antiparallel strands.

2. MATERIALS AND METHODS

2.1. PDB and PDB-REPRDB. As described in our previous study,[14] we downloaded the structural files from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB-PDB, or simply PDB in this paper; http://www.pdb.org/)[25] on November 18−19, 2007. To avoid redundant structural information, we focused on 1590 entries (1643 protein chains) of which PDB-IDs were specified by the PDB-REPRDB (http://mbs.cbrc.jp/pdbreprdb-cgi/).[29] These entries all had X-ray resolution of 2.0 Å or higher.[14] The PDB-REPRDB can specify a collection of representative PDB entries in which similar entries in terms of amino acid sequence and three-dimensional structure were eliminated. Thus, each PDB-REPRDB entry is supposed to be unique. Among the 1644 protein chains specified, one sample was not found in the PDB, and two files had wrong specifications on secondary structures. These three files were eliminated from our analysis. Therefore, we continued our analysis with 1641 protein sequences.

In the present study, we simply assumed that PDB-specified amino acid sequences for β-sheets are fundamentally correct. In accepting the three-dimensional coordinates, the PDB automatically carries out a secondary structure assignment by PROMOTIF,[30] which executes the DSSP (Dictionary of Secondary Structure of Proteins) algorithm.[31] In this algorithm, the hydrogen bonds of NH and CO groups are used to identify secondary structures.

2.2. Parallel and Antiparallel Databases. We manually extracted the SHEET sections from the original PDB files, which were then compared to the β-strand database made in the previous study,[14] using the exact function of the Microsoft Excel 2007.

Based on the residue numbers at the beginning and the end of a given secondary structure in the original PDB file, we obtained the length of every possible secondary structure by simple subtraction. This information can separately be obtained from the secondary structure assignment in the original PDB file. However, the length and the secondary structure assignment in a given file were not always consistent with each other. In such cases, we visually inspected their three-dimensional structures shown by a molecular graphics software RasMol 2.7.4.2 (2007), assuming that the structural coordinates are correct. We corrected 197 strands, but 8 samples could not be corrected because of incomplete information in the files. Thus, 1633 protein chains were subsequently studied for β-strand statistics.

Table 1. Fundamental Characteristics of β-Structures in the FL-DB[a]

	protein chain	β-containing chain	β-sheet	parallel strand (aa)	antiparallel strand (aa)
n	1633	1486	4198	5478 (26,764 aa)	11,149 (62,991 aa)

[a] *n*, number of samples; aa, amino acid. Amino acid X was excluded from the number of amino acids in parallel and antiparallel strands.

To sort each strand into either the parallel or the antiparallel category, we assigned a direction classifier of 1 (parallel) or −1 (antiparallel) to every strand. In the original PDB files, the first strand is defined as 0, and the second strand is defined based on the direction of the first strand. The directions of the third and later strands are defined based on the previous strand. We followed these rules and further defined the direction of the first strand based on the second strand. That is, the first strand assignment is always identical to the second strand.

Other amendments from the original PDB files were necessary. In the case where information on more than two strands overlapped, some strands were found to be incorrectly classified as 0 in the PDB file when they should have been 1 or −1. In some files, 0 was assigned to the last strand of a barrel sheet, probably because the last strand is identical to the very first strand that is previously defined as 0. In some files, all the β-strands in a sheet were designated as 0, or there were more than two 0 strands in a single sheet. In these cases, we referred to the three-dimensional structures using RasMol to assign a correct identity to each strand. When only one stretch of amino acids was indicated as a β-strand in a protein chain, we graphically confirmed that it formed a β-sheet with other protein chains that were crystallized together. We also found some PDB files that contained no information on strand directions and hydrogen bonds. They were considered incomplete files and thus were eliminated from our analysis. Moreover, we found some files that could not be read by RasMol. For these files, we had no way of confirming the strand direction, and thus, they were eliminated from our analysis.

Among 4198 β-sheets found in the protein samples, 3 sheets had only one strand in the PDB file. These sheets were eliminated from the calculation of the proportion of sheet classifiers (parallel, antiparallel, or mixed). We assigned direction classifier 1 or −1 to each strand and recombined them into sheets so that we could identify every sheet as parallel, antiparallel, or mixed. We could not include another 9 sheets in the proportion statistics because, in these 9 samples, one strand was assigned to two or more sheets.

2.3. Availability Scores for Doublets and Triplets. Definition and calculation of availability scores for doublets and triplets were described elsewhere.[14,26−28] Briefly, we defined the difference between the probabilistically estimated count *E* and the real count *R* for each doublet or triplet in a database as the relative count or availability for a given doublet or triplet. Availability can be expressed as follows

$$A = (R - E)/E = (R/E) - 1 \quad (1)$$

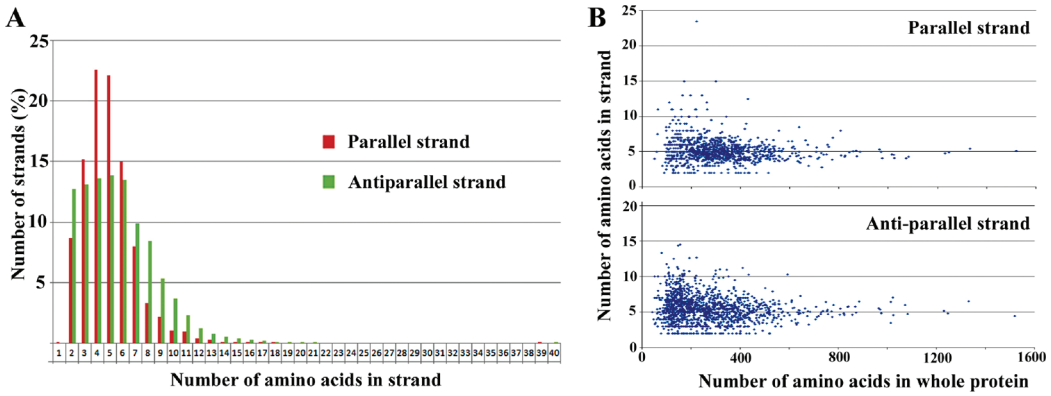


Figure 1. Length relations of β-strands. (A) Length distribution of parallel and antiparallel strands. *X*-axis is the number of amino acids, and *Y*-axis is the number of strands expressed as a percentage in the PS-DB and AS-DB. (B) Scatter plots of length of parallel and antiparallel strands against the whole protein length. The whole protein length showed no correlation with strand length.

In this equation, E is calculated in the case of the doublet as follows

$$E = Q \cdot P_1 P_2 \qquad (2)$$

where Q is the total number of existing doublets in a database, and P_1 and P_2 are the probabilities that a given amino acid appears at a position, which are derived from the occurrence of each amino acid in that database.

2.4. Statistical Analysis. Pearson correlation coefficients and *p*-values between the Kyte-Doolittle hydropathy index[17] and relative amino acid occurrence[14] were calculated using JSTAT10.0 (2006).

3. RESULTS AND DISCUSSION

3.1. The Fundamental Statistics and Size Distribution of β-Strands. We obtained representative data for 1641 protein structures from the PDB-REPRDB, as in our previous compositional study.[14] From these protein chains, we constructed the parallel strand database (PS-DB) and the antiparallel strand database (AS-DB). Fundamental statistical numbers of β-structures in the full-length database (FL-DB), from which the PS-DB and AS-DB were derived, are summarized in Table 1. From these numbers, information about β-strands in proteins was obtained.

Antiparallel strands were 2.0 times as frequent as parallel strands in proteins in general. On average, there were 2.8 β-sheets in a single protein chain, and one β-sheet was composed of about 4.0 strands. Furthermore, 91.2% of protein chains contained at least one β-sheet, and there were 10.2 strands per protein chain. Note that these 10.2 strands do not form a single sheet, but instead, they form at least a few different sheets. There were 3.4 parallel strands per protein chain on average, whereas there were 6.8 antiparallel strands per protein chain. Among the protein chains that contain one or more β-strands, 11.1 strands per protein chain were found. These figures are useful for depicting a picture of an average protein in terms of β-strand components.

We next made a histogram in which the distribution of strand length in the PS-DB and the AS-DB (expressed as percentage of the number of amino acids) were shown (Figure 1A). We found that an antiparallel strand (mean $\pm$ SD = 5.7 $\pm$ 2.9 amino acid residues) was larger than a parallel strand (mean $\pm$ SD = 4.9 $\pm$ 2.0 amino acids residues) at the population level. The antiparallel distribution peaked at 5, but the peak of the distribution was relatively flat, with 2−6 amino acids being almost equally frequent. However, the parallel distribution exhibited a relatively sharp peak at 4−5 amino acids.

There was no correlation between size of a whole protein molecule and size of β-strands that are contained in that protein (Pearson correlation coefficient $r = -0.12$ for parallel and $r = -0.14$ for antiparallel strands) (Figure 1B).

3.2. Proportion of Parallel, Antiparallel, and Mixed β-Sheets. It has been previously stated that parallel and antiparallel strands are not mixed frequently in a single sheet.[12] We tested this statement using our databases. We found that antiparallel sheets occupied 61.0% of the β-sheet samples, and parallel sheets occupied only 14.9%. The mixed sheets were 24.2% of the β-sheet samples. It seems that the majority of β-sheets are found in the antiparallel configuration, and mixed sheets are not very rare. The least frequent are parallel sheets. This may be partly because the antiparallel configuration is intrinsically favored.[11]

3.3. Amino Acid Occurrence of β-Strands. We next examined amino acid occurrence (synonymously called composition, frequency, or count in various cases) in the PS-DB and the AS-DB and expressed it as percentage (Table 2; Figure 2A). Amino acid occurrence was then divided by the occurrence in the full-length database (FL-DB) obtained in the previous study,[14] producing the relative occurrence (Table 2; Figure 2B).

In both parallel and antiparallel strands, nonpolar, aliphatic hydrocarbon amino acids, I, V, L, and A were more abundant than others (Table 1; Figure 2A). Even in the relative occurrence, which is normalized against the overall amino acid usage throughout proteins, the abundance of I and V were extraordinarily large. Importantly, only six amino acids, namely, I, V, L, M, A, and H, were more abundant in parallel strands than in antiparallel strands. Indeed, the P/A ratio (i.e., ratio of the PS-DB occurrence to the AS-PD occurrence) was highest in I, V, L, M, and A, all of which are nonpolar, hydrophobic chains (Table 1). In contrast, K, Q, W, and E had the lowest P/A ratios, which are polar and less hydrophobic, with the exception of W.

A gradual increase of relative occurrence was observed among nonpolar residues from G to I (Figure 2B), possibly indicating the importance of hydrophobicity in forming β-strands. By

relative occurrence, Y and W were frequent, especially in antiparallel strands, whereas P was the lowest (<0.5) in both strands. The uniqueness of P in secondary structures has been well documented.[14,32,33]

Table 2. Amino Acid Occurrence in PS-DB and AS-DB[a]

amino acid	FL-DB	PS-DB	AS-DB	P/A
A (Alanine, Ala)	8.72	6.96(0.80)	6.45(0.74)	1.08
E (Glutamic acid, Glu)	6.62	3.45(0.52)	5.09(0.77)	0.68
L (Leucine, Leu)	8.94	11.68(1.31)	9.74(1.09)	1.20
Q (Glutamine, Gln)	3.68	2.04(0.55)	3.16(0.86)	0.65
M (Methionine, Met)	2.00	2.22(1.11)	1.95(0.98)	1.14
R (Arginine, Arg)	5.00	3.62(0.72)	5.06(1.01)	0.72
K (Lysine, Lys)	5.51	4.91(0.89)	3.12(0.57)	0.62
D (Aspartic acid, Asp)	5.87	3.19(0.54)	3.26(0.56)	0.98
W (Tryptophan, Trp)	1.49	1.39(0.93)	2.09(1.40)	0.66
S (Serine, Ser)	5.79	4.01(0.69)	5.25(0.91)	0.76
H (Histidine, His)	2.38	2.41(1.01)	2.34(0.98)	1.03
V (Valine, Val)	7.16	16.58(2.32)	12.50(1.75)	1.33
I (Isoleucine, Ile)	5.46	12.62(2.31)	8.39(1.54)	1.50
F (Phenylalanine, Phe)	4.00	5.60(1.40)	5.68(1.42)	0.99
Y (Tyrosine, Tyr)	3.49	4.25(1.22)	5.30(1.52)	0.80
C (Cysteine, Cys)	1.29	1.56(1.21)	1.63(1.26)	0.96
T (Threonine, Thr)	5.58	5.97(1.07)	7.02(1.26)	0.85
P (Proline, Pro)	4.74	1.53(0.32)	2.14(0.45)	0.71
G (Glycine, Gly)	7.87	5.14(0.65)	5.19(0.66)	0.99
N (Asparagine, Asn)	4.21	2.44(0.57)	2.62(0.62)	0.91

[a] Numbers are expressed as percentage of total amino acids. Numbers in the FL-DB were taken from Otaki et al. (2010). Numbers in parentheses are relative amino acid occurrence, which was derived by dividing occurrence in the PS-DB or AS-DB by that of the FL-DB. The P/A is a ratio of the PS-DB occurrence to the AS-DB occurrence. Amino acid residue X in the PDB files is not included in this table, and it comprised 0.22% in the PS-DB and 0.23% in the AS-DB.

In the P/A ratio, 14 amino acids (R, K, D, N, E, Q, P, Y, W, S, T, G, C, and F) were <1.0 (more frequent in antiparallel strands), whereas 6 amino acids (H, A, M, L, V, and I) were >1.0 (more frequent in parallel strands). Except for H, the latter are all nonpolar, aliphatic amino acids. To reach the 50% level of occurrence, the sum of the occurrences of 5 amino acids (V, L, I, A, and T) was sufficient in the PS-DB, whereas the sum of the occurrences of seven amino acids (V, L, I, T, A, F, and Y) was required in the AS-DB. That is, parallel strand composition is more biased toward these amino acids.

3.4. Correlation between Amino Acid Occurrence and Hydrophobicity. Together with other secondary structure databases that were made from the same representative protein set,[14] we examined scatter plots and obtained Pearson correlation coefficients between amino acid composition and the hydrophobicity scale defined by Kyte and Doolittle[17] (Table 3; Figure 3). There was no significant correlation in whole protein sequences (FL-DB) and α-helices; although a weak negative coefficient value was obtained in 3_{10}-helices. In contrast, β-strands showed a relatively high coefficient value, which was more significant in parallel than in antiparallel β-strands. The P/A ratio also had a high coefficient.

Generally speaking, from the point of amino acid composition, α-helices are likely to be composed of relatively hydrophilic and aliphatic residues, and β-strands relatively hydrophobic and aromatic residues in addition to aliphatic ones.[14] Furthermore, the amino acid composition of β-strands was more deviated from the whole protein composition than any other secondary structure.[14] This is at least partly because of the high content of hydrophobic amino acids in β-strands, which is consistent with the present study. Structurally, the present results most likely indicate that hydrophobic interactions between strands or between one strand and other structures play an important role in stabilizing β-sheets, especially parallel sheets. We speculate that hydrophobicity may contribute to strand pairing tenden-

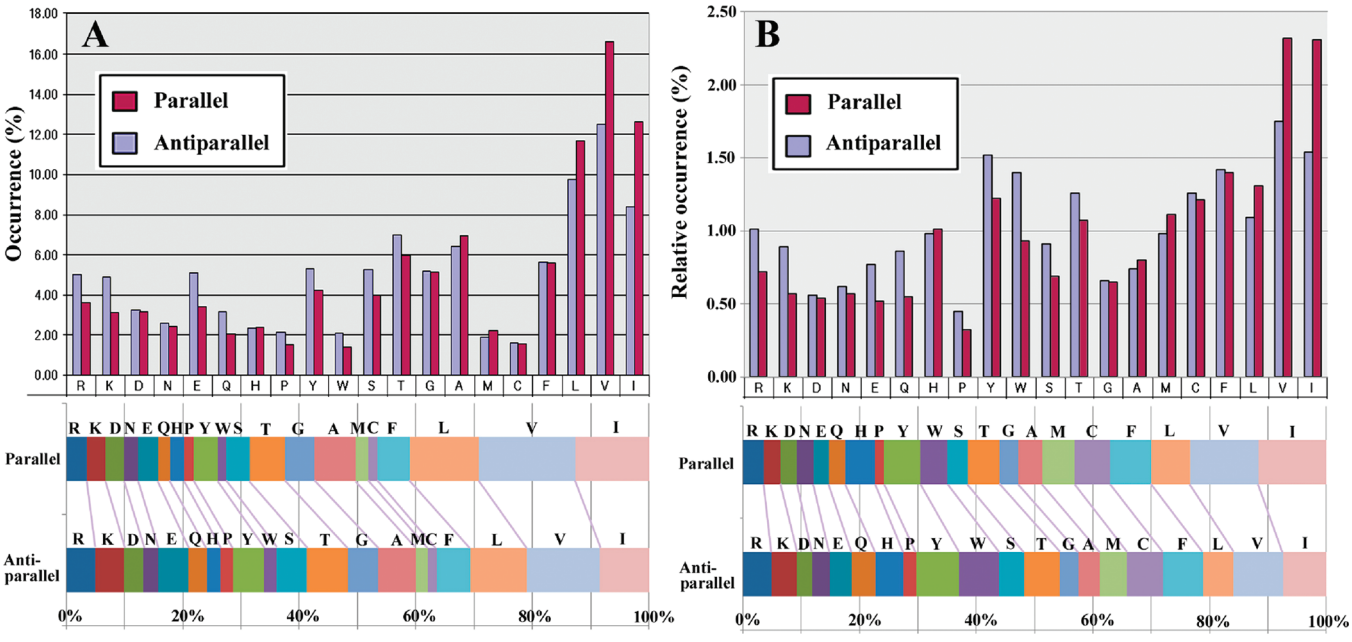


Figure 2. Amino acid composition in the PS-DB (parallel) and the AS-DB (antiparallel). Amino acids are listed in the order of hydrophobicity. (A) Occurrence (%). (B) Relative occurrence (%), which was derived by dividing occurrence in the PS-DB or AS-DB by that of the FL-DB.

dx.doi.org/10.1021/ci200027d |J. Chem. Inf. Model. 2011, 51, 1457–1464

Table 3. Correlation Analysis between Relative Occurrence and Hydrophobicity[a]

DB	FL-DB	α-helix	3_{10}-helix	β-strand	parallel (P)	antiparallel (A)	P/A
r	0.18	−0.02	−0.52	0.69	0.79	0.59	0.78
p	0.44	0.93	0.019	0.0008	<0.0001	0.0067	<0.0001

[a] Pearson correlation coefficients are listed, and *p*-values are obtained from regression analysis to fit the best straight line. DB: database, FL-DB: full-length database containing all amino acids of 1641 samples.

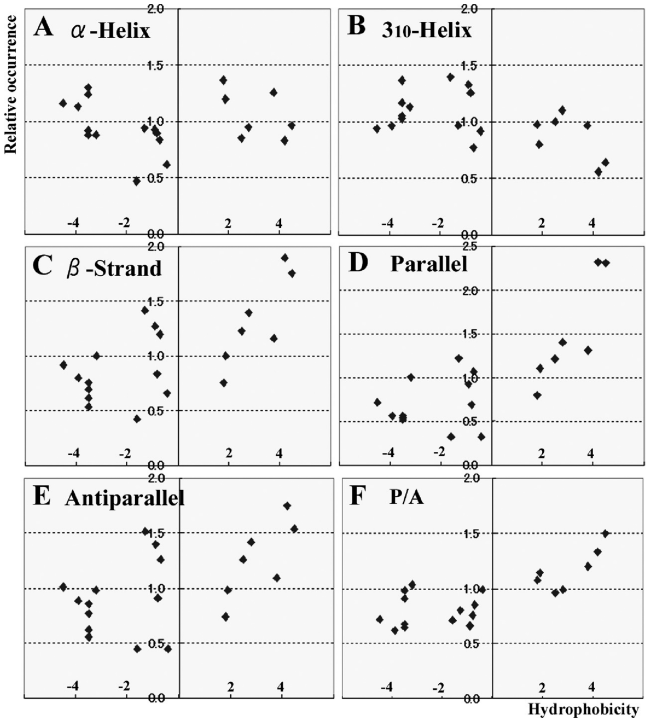


Figure 3. Scatter plots of relative occurrence of amino acids (*Y*-axis) against hydrophobicity (*X*-axis). (A) α-Helix (B) 3_{10}-Helix. (C) β-Strand. (D) Parallel strand (PS-DB). (E) Antiparallel strand (AS-DB). (F) P/A ratio.

cies in stabilizing the pleated sheet, as pointed out in other studies.[15,16,18−20] Alternatively, functional hydrophobic cores may often be made by β-sheets, and the high frequency of hydrophobic amino acids is a simple reflection of the functionality of β-sheets.

3.5. Distributions of Doublet and Triplet Scores in β-Strands. We have previously shown that usage of short constituent amino acid sequences in proteins are biologically biased, and, in extreme cases, some of the constituent sequences are completely absent from any proteins.[14,26−28] This approach to protein decoding has also been expanded by other groups.[34−36] Here, we focused on doublets and triplets, two or three consecutive amino acid sequences in a protein chain, in β-strands.

The number of doublets and triplets, excluding those containing X residue, are shown in Table 4. In the parallel database excluding X residues, we had 21,246 doublets and 15,760 triplets. In the antiparallel database, we had 51,746 doublets and 40,544 triplets. Because there are 400 doublet species and 8000 triplet species, assuming that all triplet species appear in the database

Table 4. Number of Doublets and Triplets in the PS-DB and AS-DB[a]

	singlet (amino acid)	doublet	triplet
combinatorial set ($C = 20^n$)	20	400	8000
number in the PS-DB (Q_P)	26,764	21,246	15,760
Q_P/C	1341	53	2
number in the AS-DB (Q_A)	62,991	51,746	40,544
Q_A/C	3157	129	5

[a] X-containing doublets and triplets are excluded.

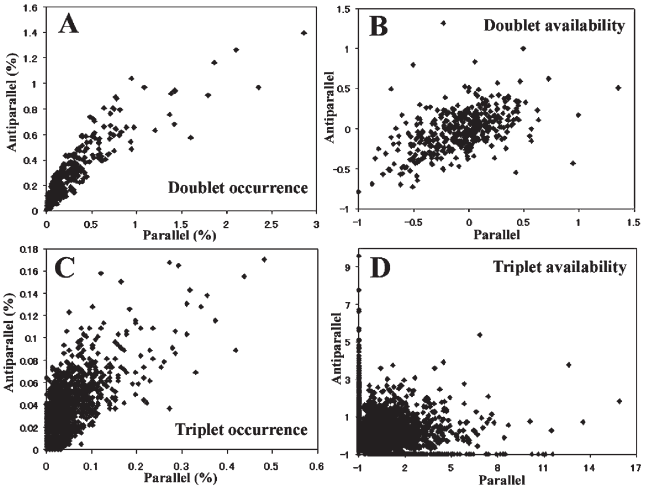


Figure 4. Parallel versus antiparallel scatter plots in occurrence and availability. (A) Doublet occurrence. (B) Doublet availability. (C) Triplet occurrence. (D) Triplet availability.

equally frequently, we expect each doublet and triplet to appear 53 times and 2 times, respectively, in the PS-DB. Similarly, we expect each doublet and triplet to appear 129 times and 5 times, respectively, in the AS-DB. In reality, the assumption of equal frequency is simply not true, but these numbers can be used to probe high score doublets and triplets below.

To determine differences in the overall usage of doublets and triplets in parallel and antiparallel strands, we constructed scatter plots of 400 doublets and 8000 triplets in the combinations of parallel and antiparallel structures in terms of occurrence (%) and availability scores (Figure 4). The overall distribution of doublets was found mostly along the diagonal line, indicating they were found equally frequently in both strands (Figure 4A, B). In contrast, the overall distribution of triplets was more dispersed than the doublet distributions, indicating that some triplets were preferably used in either strand. Indeed, some of them appeared to be found exclusively in one of two structures. There were many triplets with an availability score of −1, some of which were identified in the subsequent sections.

3.6. Doublets and Triplets with High Availability Scores. We identified doublets with high occurrence and high availability together with the P/A ratio in a given secondary structure to find out doublets that are characteristic to a given strand (Table 5), assuming that those "marker" doublets may have a high availability score, high or low P/A ratio, and high absolute count (occurrence) in a given strand.

Table 5. High-Score Doublets in Parallel and Antiparallel Strands[a]

	parallel (PS-DB)		antiparallel (AS-DB)	
rank	occurrence (P/A)	availability (P/A)	occurrence (P/A)	availability (P/A)
1	VV 607 (2.0)	WN 1.36 (0.9)	VV 722 (2.0)	CC 1.32 (0.3)
2	VI 501 (2.3)	MN 0.99 (1.8)	VL 653 (1.6)	WQ 1.00 (0.3)
3	VL 448 (1.6)	WC 0.95 (2.1)	LV 602 (1.6)	WD 0.83 (0.4)
4	LV 397 (1.6)	ID 0.72 (1.5)	TV 537 (0.9)	QW 0.80 (0.1)
5	IV 381 (1.9)	HF 0.63 (1.4)	VI 502 (2.3)	ID 0.62 (1.5)
6	II 341 (2.7)	LD 0.62 (1.5)	VT 500 (1.1)	FD 0.59 (0.9)
7	LL 303 (1.7)	FW 0.57 (1.2)	LL 490 (1.7)	WH 0.57 (0.5)
8	VA 303 (1.5)	PC 0.57 (1.1)	VA 486 (1.5)	MP 0.53 (0.5)
9	IL 302 (2.0)	MQ 0.56 (1.2)	AV 477 (1.5)	YD 0.51 (0.5)
10	AV 294 (1.5)	YQ 0.51 (0.6)	IV 470 (1.9)	WN 0.51 (0.9)
11	LI 291 (1.8)	WQ 0.49 (0.3)	TL 463 (0.8)	CQ 0.49 (0.1)
12	IA 256 (1.9)	FD 0.47 (0.9)	EV 458 (0.9)	FE 0.48 (0.5)

[a] P/A ratio (occurrence in the PS-DB to occurrence in the AS-DB) is shown for each doublet in parentheses. High P/A is appropriate for parallel markers, whereas low P/A is appropriate for antiparallel markers.

Table 6. High-Score Triplets in Parallel and Antiparallel Strands[a]

	parallel (PS-DB)		antiparallel (AS-DB)	
rank	occurrence (P/A)	availability (P/A)	occurrence (P/A)	availability (P/A)
1	VVV 76 (2.6)	ECW 15.9 (2.4)	VVV 69 (2.6)	CCN 9.6 (0)
2	VLV 69 (2.6)	WQP 13.6 (2.4)	LVL 68 (0.6)	CQC 7.8 (0)
3	VIV 66 (4.3)	WWH 12.6 (1.2)	VAV 67 (0.7)	HWM 6.7 (0)
4	VVL 59 (3.0)	MMQ 11.6 (na)	VEV 64 (0.3)	WQM 6.6 (0)
5	LVV 56 (1.0)	PGG 11.5 (6.3)	VLV 63 (2.6)	PCW 5.7 (0)
6	VVI 54 (1.0)	WCH 11.1 (na)	VTV 61 (0.4)	MHC 5.6 (0)
7	VIL 52 (1.9)	WCN 10.9 (na)	VLL 58 (0.9)	HWH 5.4 (0.8)
8	VLL 50 (0.9)	CCM 10.6 (na)	LVV 56 (1.0)	WHC 5.1 (0)
9	AVV 49 (0.9)	KMN 10.2 (na)	AVV 53 (0.9)	CQH 5.1 (0)
10	VIA 49 (1.2)	QTW 10.2 (na)	VKV 52 (0.3)	PMM 5.1 (0)
11	VAV 46 (0.7)	RPG 10.1 (2.9)	VVI 52 (1.0)	QWQ 4.8 (0)
12	ILV 45 (1.3)	HFH 8.7 (5.9)	LVA 51 (0.6)	QMW 4.7 (0)

[a] P/A ratio (occurrence in the PS-DB to occurrence in the AS-DB) is shown for each triplet in parentheses. High P/A is appropriate for parallel markers, whereas low P/A is appropriate for antiparallel markers. na, not applicable when A (antiparallel count) is zero.

In occurrence, there were no polar or nonhydrophobic amino acids in parallel strands until the 13th doublet VT, whereas there were relatively many of such amino acids in the AS-DB such as the fourth doublet TV. Among the top 12 doublets in availability, WQ, ID, FD, and WN were found in both parallel and antiparallel strands. Thus, they cannot be considered a good marker that is specific for either strand. In parallel strands, MN and WC may be relatively good markers because of their relatively high availability and P/A ratios. On the other hand, in antiparallel strands, most doublets can serve as markers, including CC, WD, QW, WH, MP, YD, CQ, and FE. Among these, CC, QW, and CQ are particularly suitable for this purpose with very low P/A ratios.

Similarly, we listed high-score triplets to find out candidate marker triplets (Table 6). In the case of parallel strands, the top 12 triplets for occurrence contained hydrocarbon side chains only. This is in contrast to antiparallel strands, where E, T, and K were in the top 12 triplets. On the other hand, the high availability triplets contained many polar residues. Interestingly, V, I, and L (the most frequently used amino acids) were not found in the top 12 triplets in availability. Among the high availability triplets, PPG and HFH might be considered to be good markers for parallel strands because their real counts were 8 and 5, respectively, which were the highest in the listed triplets. Note that each triplet is supposed to appear only 2 times on average in the PS-DB if equal frequency is assumed (see Table 4). Among the high availability triplets in the antiparallel strands, none of the top 12 triplets were found in parallel strands except HWH, having a P/A ratio of zero. Among them, QWQ and WQM have the highest real count of 5 and 4, respectively. Note that each triplet is supposed to appear 5 times on average in the AS-DB (see Table 4). Because they contain relatively rare amino acids, their real counts can be considered high. Thus, these triplets could serve as good markers for antiparallel strands.

3.7. Strand-Specific Triplets. Many of the high availability triplets happened to be specifically found only in either parallel or antiparallel strands in the previous section. Thus, we additionally picked up strand-specific triplets (i.e., triplets that were found exclusively in either PS-DB and AS-DB) with high availability (>2.0) and a high real count (>2 in the PS-DB and >6 in the AS-DB based in Table 4) (Table 7). For example, YSC ranked 55th in availability, and it was found 7 times in the AS-DB but null in the PS-DB.

3.8. Zero-Count Doublets and Triplets in β-Strands. In our previous study,[14] we noticed that some triplets did not appear in a given secondary structure at all, despite their probabilistically expected counts which indicated otherwise. We called them zero-count triplets. Zero-count triplets are the triplets that are "avoided" or "forbidden" for that particular secondary structure. Reasons for this avoidance are not clear at this point, but because they are specific to a given secondary structure, it is likely that they are strong "breakers" for that secondary structure.

There were 3710 zero-count triplets in the PS-DB, but there were only 1321 in the AS-DB. The relatively large number of zero-count triplets in the PS-DB reflects the smaller database size. Among these zero-count triplets, we listed the ones with a high probabilistic expectation, expressed as E in eqs 1 and 2 (Table 8). For example, the zero-count triplet ETA was expected to appear 9.4 times in the AS-DB, but they did not appear at all in reality. Interestingly, these zero-count triplets were not always constructed by the amino acids with low compositional percentages. Instead, amino acids with high compositional percentages, such as V, I, and L, were found in these triplets. This observation points out the importance of the triplet usage in proteins that cannot be deciphered readily from the amino acid composition.

Among the triplets listed in Table 8, 6 triplets (DTI, TNV, FFR, IQG, ESE, and PEV) had zero counts in both PS-DB and AS-DB. DTI, FFR, IQR, and PEV were already identified in our previous study as a negative signature for β-strands.[14] Some zero-count triplets in the PS-DB had relatively high counts in the AS-DB. It is

Table 7. High-Score Strand-Specific Triplets in Parallel and Antiparallel Strands[a]

parallel (PS-DB)				antiparallel (AS-DB)			
rank	triplet	real count	availability	rank	triplet	real count	availability
5	KMN	3	10.20	55	YSC	7	2.78
6	QTW	3	10.19	61	WLN	8	2.68
32	NLC	4	4.67	64	QFN	7	2.65
34	YCG	3	4.54	74	LHW	7	2.59
47	GDM	3	4.20	79	GTW	11	2.53
64	LMM	4	3.39	81	KME	7	2.53
72	DSR	3	3.09	95	GPY	8	2.33
80	CIH	3	2.98	108	RWE	7	2.18
111	FIW	5	2.21	110	MKF	7	2.16
115	QHV	4	2.10	120	DYD	7	2.04
117	IDP	3	2.06				
121	LHM	3	2.03				

[a] Rank is determined among strand-specific triplets according to the availability score, that is, parallel from 1st to 287th, and antiparallel from 1st to 2676th. Among the triplets with an availability score >2.0, those with high real count are listed. That is, triplets with real count >2 in the PS-DB, and real count >6 in the AS-DB, are listed.

Table 8. Zero-Count Triplets with High Probabilistic Counts (Expected Counts)[a]

parallel (PS-DB)				antiparallel (AS-DB)			
rank	triplet	expected	cross count	rank	triplet	expected count	cross count
1	VNI	8.1	17	1	ETA	9.4	2
2	VTK	4.9	16	2	VDE	8.5	2
3	SIF	4.5	10	3	DTI	7.8	0
4	FYL	4.4	13	4	NSV	7.0	1
5	SGI	4.1	5	5	FFR	6.7	0
6	DAL	4.1	1	6	KRA	6.5	2
7	YGL	4.1	10	7	SRG	5.6	1
8	TRL	4.0	11	8	IQG	5.6	0
9	TNV	3.8	0	9	ESE	5.6	0
10	DTI	3.8	0	10	PEV	5.6	0

[a] Triplets are ranked according to the expected count. Cross count means occurrence in other β-strand database. For example, VNI, which did not appear at all in the PS-DB despite its expected count 8.1, appeared 17 times in the AS-DB. Zero-count triplets with $E > 3$ in the PS-DB and $E > 5$ in the AS-DB are shown.

likely that this is partly because the size of the PS-DB is relatively small, and thus, many triplets had no real count. However, there may be a possibility that some of them can serve as a negative signature specific to the parallel strand. Although these zero-count triplets may be found in β-strands in the future with the expansion of data sets, they would be nonetheless low-count triplets that may be "avoided" to a certain extent in β-strands.

4. CONCLUSIONS

To our knowledge, the present study is the most comprehensive compositional study aimed at characterizing β-strands and discriminating parallel and antiparallel strands. This empirical study employed very simple concepts without complicated mathematical operations, which highlighted the important hydrophobic nature and other features of β-strand structures. Hydrophobicity was especially notable in the parallel β-strands, in that more hydrophobic amino acids tend to be incorporated in the parallel than in the antiparallel β-strands. We believe that use of the empirical information presented in this study can contribute to more accurate prediction of secondary structures from the primary sequences in combination with other existing programs.[37]

We found triplets that are relatively specific to either parallel or antiparallel strands. Furthermore, we found zero-count triplets that are not used at all in either parallel or antiparallel strands. At this point, the biological significance of a group of triplets that are specific to a given secondary structure is not clear, nor do we know the significance of zero-count triplets in secondary structures. However, they can at least serve as a marker or signature for a given secondary structure in a prediction program in the future. Importantly, structural and functional studies of these triplets in β-strands are expected. Moreover, the hypothesis that zero-count triplets are avoided or forbidden so as not to break secondary structures must be experimentally validated *in vitro*.

■ ASSOCIATED CONTENT

Supporting Information. The list of β-strand samples (PS-DB and AS-DB) analyzed in the present study. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: otaki@sci.u-ryukyu.ac.jp.

■ ACKNOWLEDGMENT

We thank K. Motomura, T. Fujita, and other members of the BCPH Unit of Molecular Physiology for discussion. This research was partially supported by the Takeda Research Foundation.

■ REFERENCES

(1) Pauling, L.; Corey, R. B.; Branson, H. R. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 205–211.

(2) Pauling, L.; Corey, R. B. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 729–740.

(3) Chou, P. Y.; Fasman, G. D. Prediction of protein conformation. *Biochemistry* **1974**, *13*, 222–145.

(4) Chou, P. Y.; Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequences. *Adv. Enzymol.* **1978**, *47*, 45–148.

(5) Lim, V. I. Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.* **1974**, *88*, 873–894.

(6) Krigbaum, W. R.; Knutton, S. P. Prediction of the amount of secondary structure in a globular protein from its amino acid composition. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 2809–2813.

(7) Minor, D. L., Jr.; Kim, P. S. Measurement of the β-sheet-forming propensities of amino acids. *Nature* **1994**, *367*, 660–663.

(8) Minor, D. L., Jr.; Kim, P. S. Context is a major determinant of β-sheet propensity. *Nature* **1994**, *371*, 264–267.

(9) Minor, D. L., Jr.; Kim, P. S. Context-dependent secondary structure formation of a designed protein sequence. *Nature* **1996**, *380*, 730–734.
(10) Berg, J. M.; Tymoczko, J. L.; Stryer, L. *Biochemistry*, 6th ed.; W. H. Freeman: New York, 2007.
(11) Creighton, T. E. *Proteins: Structures and Molecular Properties*, 2nd ed.; W. H. Freeman: New York, 1993.
(12) Branden, C.; Tooze, J. *Introduction to Protein Structure*, 2nd ed.; Garland Publishing: New York, 1999.
(13) Williams, R. W.; Chang, A.; Juretić, D.; Loughran, S. Secondary structure predictions and medium range interactions. *Biochim. Biophys. Acta* **1987**, *916*, 200–204.
(14) Otaki, J. M.; Tsutsumi, M.; Gotoh, T.; Yamamoto, H. Secondary structure characterization based on amino acid composition and availability in proteins. *J. Chem. Inf. Model.* **2010**, *50*, 690–700.
(15) Sternberg, M. J. E.; Thornton, J. M. On the conformation of proteins: hydrophobic ordering of strands in β-pleated sheets. *J. Mol. Biol.* **1977**, *115*, 1–17.
(16) Lifson, S.; Sander, C. Antiparallel and parallel β-strands differ in amino acid residue preference. *Nature* **1979**, *282*, 109–111.
(17) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132.
(18) Zhang, N.; Ruan, J.; Duan, G.; Gao, S.; Zhang, T. The interstrand amino acid pairs play a significant role in determining the parallel and antiparallel orientation of β-strands. *Biochem. Biophys. Res. Commun.* **2009**, *386*, 537–543.
(19) Zhang, N.; Duan, G.; Gao, S.; Ruan, J.; Zhang, T. Prediction of the parallel/antiparallel orientation of beta-strands using amino acid pairing preferences and support vector machines. *J. Theor. Biol.* **2010**, *263*, 360–368.
(20) Steward, R. E.; Thornton, J. M. Prediction of strand pairing in antiparallel and parallel β-sheets using information theory. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 178–191.
(21) Koch, O.; Bocola, M.; Klebe, G. Cooperative effects in hydrogen-bonding of protein secondary structure elements: a systematic analysis of crystal data using Secbase. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 310–317.
(22) Yang, A.-S.; Honig, B. Free energy determinants of secondary structure formation: II. Antiparallel β-sheets. *J. Mol. Biol.* **1995**, *252*, 366–376.
(23) Chellgren, B. W.; Creamer, T. P. Side-chain entropy effects on protein secondary structure formation. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 411–420.
(24) Nowick, J. S. Exploring β-sheet structure and interactions with chemical model systems. *Acc. Chem. Res.* **2008**, *41*, 1319–1330.
(25) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 253–242.
(26) Otaki, J. M.; Gotoh, T.; Yamamoto, H. Frequency distribution of the number of amino acid triplets in the non-redundant protein database. *J. Jpn. Soc. Inf. Knowledge* **2003**, *13*, 25–38.
(27) Otaki, J. M.; Ienaka, S.; Gotoh, T.; Yamamoto, H. Availability of short amino acid sequences in proteins. *Protein Sci.* **2005**, *14*, 617–625.
(28) Otaki, J. M.; Gotoh, T.; Yamamoto, H. Potential implications of availability of short amino acid sequences in proteins: an old and new approach to protein decoding and design. *Biotechnol. Ann. Rev.* **2008**, *14*, 109–141.
(29) Noguchi, T.; Matsuda, H.; Akiyama, Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank. *Nucleic Acids Res.* **2001**, *29*, 219–220.
(30) Hutchinson, E. G.; Thornton, J. M. PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci.* **1996**, *5*, 212–220.
(31) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
(32) Piela, L.; Nemethy, G.; Scheraga, H. A. Proline-induced constraints in α-helices. *Biopolymers* **1987**, *26*, 1587–1600.
(33) MacArthur, M. W.; Thornton, J. M. Influence of proline residues on protein conformation. *J. Mol. Biol.* **1991**, *218*, 397–412.
(34) Austin, R. S.; Provart, N. J.; Cutler, S. R. C-terminal motif prediction in eukaryotic proteomes using comparative genomics and statistical over-representation across protein families. *BMC Genomics* **2007**, *8*, 191.
(35) Tuller, T.; Chor, B.; Nelson, N. Forbidden penta-peptides. *Protein Sci.* **2007**, *16*, 2251–2259.
(36) Bresell, A.; Persson, B. Characterization of oligopeptide patterns in large protein sets. *BMC Genomics* **2007**, *8*, 346.
(37) Zimmermann, O.; Wang, L.; Hansmann, U. H. E. BETTY: Prediction of β-strand type from sequence. *In Silico Biol.* **2007**, *7*, 37.