Secondary Structure Characterization Based on Amino Acid Composition and Availability in Proteins

Joji M. Otaki,*,†,§ Motosuke Tsutsumi,† Tomonori Gotoh,‡ and Haruhiko Yamamoto§

The BCPH Unit of Molecular Physiology, Department of Chemistry, Biology, and Marine Science, University of the Ryukyus, Nishihara, Okinawa 903-0213, Japan, and Department of Information Science and Department of Biological Sciences, Kanagawa University, Hiratsuka, Kanagawa 259-1293, Japan

Received November 19, 2009

The importance of thorough analyses of the secondary structures in proteins as basic structural units cannot be overemphasized. Although recent computational methods have achieved reasonably high accuracy for predicting secondary structures from amino acid sequences, a simple and fundamental empirical approach to characterize the amino acid composition of secondary structures was performed mainly in 1970s, with a small number of analyzed structures. To extend this classical approach using a large number of analyzed structures, here we characterized the amino acid sequences of secondary structures (12 154 α-helix units, 4592 3_{10}-helix units, 16 787 β-strand units, and 30 811 "other" units), using the representative three-dimensional protein structure records (1641 protein chains) from the Protein Data Bank. We first examined the length and the amino acid compositions of secondary structures, including rank order differences and assignment relationships among amino acids. These compositional results were largely, but not entirely, consistent with the previous studies. In addition, we examined the frequency of 400 amino acid doublets and 8000 triplets in secondary structures based on their relative counts, termed the availability. We identified not only some triplets that were specific to a certain secondary structure but also so-called zero-count triplets, which did not occur in a given secondary structure at all, even though they were probabilistically predicted to occur several times. Taken together, the present study revealed essential features of secondary structures and suggests potential applications in the secondary structure prediction and the functional design of protein sequences.

1. INTRODUCTION

Proteins are a group of important functional molecules in biological systems. The function of a protein molecule is a direct reflection of its three-dimensional structure. Protein tertiary structures can be considered as a collection of secondary structures, notably, α-helices and β-sheets. These secondary and tertiary structures are thought to be determined directly by their primary structures, i.e., amino acid sequences.

Not surprisingly, the characterization of secondary structures was one of the main concerns of protein scientists in 1970s, who proposed the fundamental nature of secondary structures with a limited number of protein structure records.[1–5] Although these studies have been influential in this research field, only 15 analyzed structure records (29 records four years later) were used in the original Chou and Fasman method[1,2] and only 25 in the improved Garnier−Osguthorpe−Robson (GOR) method.[5] To compensate for this incompleteness, the use of amino acid triplet information was proposed, but it still suffered from the lack of a sufficient number of analyzed structures.[6] Moreover, structure records that were available at that point may be necessarily biased from the entire protein population.

Nonetheless, since then, the compositional importance in proteins has been well recognized, although it is widely accepted that the amino acid composition alone cannot directly specify secondary structures. Some of the most popular biochemistry textbooks, such as *Biochemistry*[7] and *Proteins*,[8] mention the relative occurrence of amino acid residues in primary structures as important information that could influence secondary structures. These prestigious textbooks refer to Williams et al.,[9] which, with a limited number of structure records (still fewer than one hundred), is a further improvement and reconfirmation of the classical analyses mentioned above.

In addition to Williams et al.,[9] Nakashima et al.[10] analyzed the amino acid composition of 135 records to predict the folding types of proteins. Similarly, the amino acid composition was used to predict the amount of secondary structures (i.e., secondary structure content).[11–14] Ruan et al.[15] and Lee et al.[16] proposed highly accurate methods for predicting the secondary structure content based on the amino acid composition and positions. The incorporation of amino acid pairs as useful information for predicting the secondary structure itself or for its content has also been investigated.[17–19] However, to our knowledge, a simple and comprehensive compositional analysis of secondary structures has not been performed in this postgenome era.

More than 30 years has passed since the classical analyses, and now partly because of the worldwide effort for genomics and proteomics, publicly available analyzed structure records have been accumulated in the Protein Data Bank (PDB).[20] As of December 12, 2008, there were 50 507 protein structure

* Corresponding author. E-mail: otaki@sci.u-ryukyu.ac.jp.
† University of the Ryukyus.
‡ Department of Information Science, Kanagawa University.
§ Department of Biological Sciences, Kanagawa University.

10.1021/ci900452z
Published on Web 03/08/2010

records in the PDB. Based on these records, a recharacterization of secondary structures can be expected. Moreover, secondary structure analyses from both modern perspectives and computational technologies, such as a comprehensive search for constituent short amino acid sequences,[21–23] are likely to produce new insights into the nature of amino acid sequences in proteins.

In the present study, we collected a statistically rigorous number of representative samples (1641 protein chains) from the PDB and constructed and characterized four kinds of the secondary structure databases (SS-DBs), the α-helix, 3_{10}-helix, β-strand, and "other" databases, from the full-length database (FL-DB, also called the "all" database). Note that the "other" database contained sequences that were left after the removal of the sequences for the α-helices, 3_{10}-helices, and β-strands and that the "other" regions of amino acid sequences may be considered as one of the secondary structures just for convenience in this paper.

We focused not only on the occurrence of single amino acids but also on the occurrence of short amino acid sequences (i.e., doublets [dipeptides] and triplets [tripeptides], consecutive two and three amino acid combinations, respectively) that constitute secondary structures. A similar idea was proposed more than three decades ago by Nagano[6] without a sufficient number of analyzed structures and computational power. We have already pointed out the importance of short constituent amino acid sequences in proteins in general and also in secondary structures.[21–23] Our strategy is to exhaustively search for all 20 species of amino acid singlets and all 400 and 8000 species of amino acid doublets and triplets, respectively, and to examine their absolute counts (occurrence) and relative counts (defined as availability) in a defined structure. We discuss the potential implications of these results to the very fundamental understanding of proteins.

2. MATERIALS AND METHODS

2.1. The PDB, PDB-REPRDB, and FL-DB. 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/)[20] on November 18−19, 2007. At the end of 2007, the RCSB-PDB had accumulated 47 283 protein structure entries. However, these entries contain highly redundant sequences and structures under different experimental conditions. For this reason, we focused on 1590 entries (1643 protein chains) for which PDB-IDs were specified by the PDB-REPRDB (http://mbs.cbrc.jp/pdbreprdb-cgi/).[24] 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. To do this, we set the following conditions: database, all protein chains including membrane proteins; resolution, not more than 2.0 Å; *R*-factor, not more than 0.2; and the number of residues, not fewer than 40 amino acids. Among the 1644 protein chains specified, one sample was not found in the PDB, and thus, the total number of protein chains was 1643. For comparison, we downloaded the nonredundant amino acid sequence database (NR-AA) from the ftp site (ftp://ftp.ncbi.nih.gov/blast/db/) maintained by the National Center for Biotechnology Information (NCBI) in April, 2008.

The downloading process of the structural PDB files was automatically performed by our computer program, and the amino acid sequences were also extracted automatically from these files. This program converted the nonstandard amino acid residues and unknown residues to X. The number of Xs was 1110. Independently, we manually collected the same entries from the structural PDB files ("text/markup" in "Chain Display" enclosed by the "Sequence" tabs), and the amino acid sequences were extracted manually. These two databases contained identical numbers of amino acid residues, demonstrating the validity of the collecting procedure. We decided to use the automatically collected one for the subsequent analyses. We ignored the X residues and X-containing triplets from the analyses, unless they were amended manually (see below). This elimination should not change the results significantly due to the relatively small number of the X residues (0.25% of the total number of amino acid residues). Stretches of amino acid sequences without three-dimensional structural coordinates were cut off from the main parts of protein chains when they were found at the beginning or at the end of the chains. These regions are highly likely to be artificial tags, such as polyhistidine. When found in the middle of protein chains, they were retained but not included in the "all" and "other" databases. Thus, after these manual inspections and amendments, all sequence entries had defined structural coordinates. We called this file the full-length database (FL-DB) or the "all" database, which contained 1641 entries and 421 865 amino acid residues. Our database contained 56 membrane proteins (3.4%), although most of their structural data may be derived from their intra- or extracellular sites.

2.2. The SS-DBs. From the FL-DB, we first produced four secondary structure databases (SS-DBs), the α-helix (12 154 entries, 154 129 amino acids), 3_{10}-helix (4592 entries, 21 497 amino acids), β-strand (16 787 entries, 90 650 amino acids), and "other" (30 811 entries, 155 658 amino acids) databases, through our computer program. The "other" database contained sequences that were not identified as α-helices, 3_{10}-helices, or β-strands. We did not make a turn database because of ambiguity in identifying turns in the PDB files. In the present study, we simply followed the PDB-specified amino acid sequences for these secondary structures, assuming that their specifications are correct. In accepting the three-dimensional coordinates, the PDB automatically carries out secondary structure assignment by PROMOTIF,[25] which basically executes the DSSP (dictionary of secondary structure of protein) algorithm.[26] In this algorithm, the hydrogen bonds of NH and CO groups are used to identify secondary structures. Very short helices are highly likely to be algorithmic artifacts (as confirmed by a PDB curator), but we did not exclude them from our analyses just for simplicity, which should not significantly bias the final results due to their rarity. If a more dihedral-angle centric definition is used, some minor differences from the DSSP algorithm may show up especially in small units and in boundaries. However, considering the large number of samples we collected, such differences would not significantly change our results presented here.

To examine the validity of these original SS-DBs, we manually extracted the HELIX and SHEET sections from

the original PDB files and compared them to the original SS-DBs, using the exact function of the Microsoft Excel 2007. We examined every residue that was not identical between these two data sets, including insertions, lack of residues, and Xs in the known residues, and corrected them all manually. Note that the PDB original file contains information on the length of a given secondary structure together with its amino acid sequences. This information is separately written from the secondary structure assignment. However, the length information 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 with the molecular graphics software RasMol 2.7.4.2 (2007), assuming that the structural coordinates must be correct. We found 363 helices and 197 strands that had to be corrected. These were partly because of the overlap of two secondary structure entities. The corrected SS-DBs were used for the subsequent analyses.

2.3. The Rank Order Distance Scores. Rank order distance (ROD) scores, which indicate distances of each secondary structure from the "all" database in terms of the compositional rank order, were calculated as follows:

$$\mathrm{ROD} = \sqrt{\frac{1}{20}\sum_{n=1}^{20}({}_{\mathrm{all}}R_n - {}_{\mathrm{SS}}R_n)^2} \quad (1)$$

where ${}_{\mathrm{all}}R_n$ is the rank number of a given amino acid n in the "all" database (FL-DB), ${}_{\mathrm{SS}}R_n$ is the rank number of the same amino acid n in the secondary structure database (SS-DB), and n indicates a given amino acid species.

2.4. The Availability Scores for Doublets and Triplets. There are 20 species of amino acid residues and, hence, 400 ($=20^2$) doublet and 8000 ($=20^3$) triplet species. For each SS- and FL-DB, the number of occurrences of each amino acid was counted, and similarly, the number of occurrences of each amino acid doublet and triplet was counted. They were called the absolute count, or real count R. The probability P that a given amino acid appears at a given position can be derived from this occurrence. Below, we discuss the case of triplets rather than doublets, but both of them can be treated similarly. With the total number of the existing triplets Q, we can obtain the expected triplet count E for each triplet as follows:

$$E = Q \cdot P_1P_2P_3 \quad (2)$$

The total number of the existing triplets in a database can be derived as $Q = X - 2N$ ($X \geq 3N$), where X is the total number of amino acid residues of proteins whose individual chains are greater than three amino acids in length, and N is the total number of protein chains. Thus, the eq 2 can be rewritten as

$$E = (X - 2N) \cdot P_1P_2P_3 \quad (3)$$

A more complex but operationally sound way to derive the numerical value of E is to produce randomized imaginary databases in which the protein chains in a database were broken into individual amino acids and then recombined randomly to produce the same number of protein chains. From a randomized imaginary database, each species of triplets is counted.

We defined the difference between the probabilistically estimated triplet count E and the real triplet count R for each triplet in a database as the relative triplet count or "availability" for a given triplet. Availability A can be expressed as follows:

$$A = (R - E)/E = (R/E) - 1 \quad (4)$$

In the present study, we first derived E from both methods, the probabilistic calculations based on eq 3 and the randomized simulations. For the latter, we used the random generator that was built in the Microsoft Excel 2007 and simulated the imaginary database, which was generated 100 times. The E value was derived as the mean of these 100 trials. We confirmed that both operations produced essentially equal results, as was observed in the previous study.[22] Simply because of the operational simplicity, we showed the results obtained from the E values derived from the eq 2 in this study.

We understand that the triplet MXX or doublet MX would occur in high frequency at the beginning of a protein chain, as discussed in the previous study.[27] In our samples, there were only 337 protein chains that started with MXX, which is much smaller than the entire number of MXX (8350) in our FL-DB. Thus, the starting MXX comprises 4.0% of the entire MXX. This is partly because proteins are usually processed upon expression and also before crystallization. We indeed removed artificial tags at the ends of protein chains, as mentioned above. Other than that, we did not perform any particular operations for the starting MXX correction simply because their influence on the statistical outputs should be minimal.

3. RESULTS AND DISCUSSION

3.1. The Fundamental Statistics and Unit-Size Distributions of Secondary Structures. We constructed the SS-DBs from the FL-DB that contained 1641 representative protein chains from the PDB-REPRDB. In order to confirm the representative quality of our FL-DB in terms of the length distribution, we first constructed the histogram for the full-length distribution of proteins using the NR-AA database, which contained 6 405 498 nonredundant protein entries (Figure 1A). The length was distributed broadly with the peak around 100−150 amino acids. The highest peak was found at 99 amino acids, but similar peaks were repeatedly found roughly every 50−100 amino acids, which could indicate the average size of protein domains that may be connected with each other in tandem. We then examined the full-length distribution of the FL-DB (Figure 1B). The distribution pattern was reasonably similar to that of the NR-AA, justifying the use of this database for the subsequent analyses as representative protein samples. The amino acid compositional patterns were also similar to each other (see below), further justifying the use of our FL-DB.

We next examined the length distribution patterns of secondary structures. Each secondary structure unit (a continuous stretch of amino acid sequence that forms a single type of secondary structure as defined by the PDB) was treated as a single entry. The length distribution histogram for α-helices showed that it spanned from 1 to 65 amino acids in length (Figure 1C). Note that, just for simplicity, we followed the specification of secondary structures given

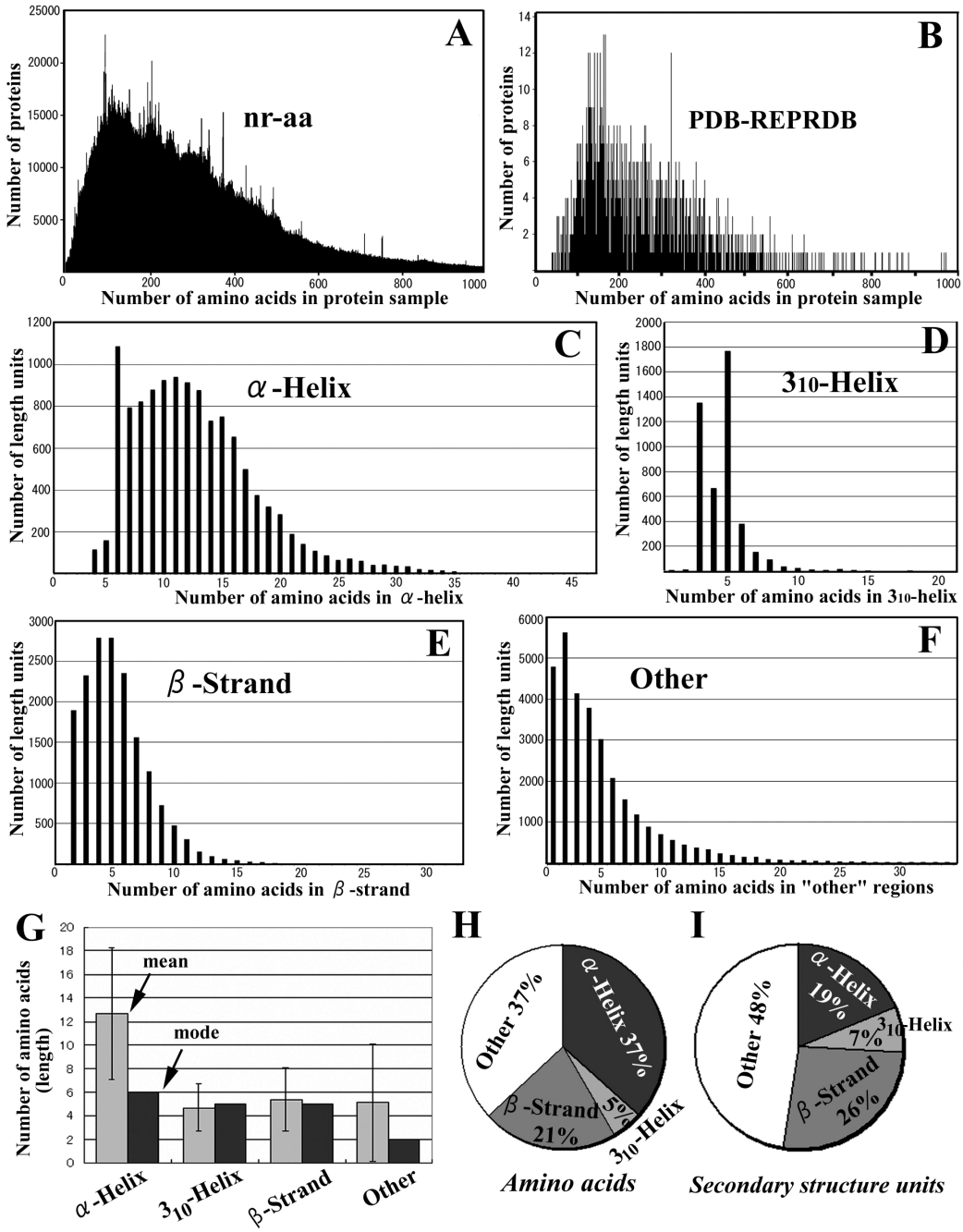

Figure 1. Length characterization and compositional analysis of secondary structures. Right-hand side of *X*-axis was truncated in the histograms because of the lack of visually detectable bars. (A) Full-length distribution of proteins in the NR-AA database. (B) Full-length distribution of proteins in the FL-DB (derived from the PDB-REPRDB). (C) Length distribution of α-helices. Actual range of sample distribution: 1−65. (D) Length distribution of 3_{10}-helices. Actual range of sample distribution: 1−34. (E) Length distribution of *β*-strands. Actual range of sample distribution: 1−38. (F) Length distribution of the "other" regions. Actual range of sample distribution: 1−189. (G) Mean (lightly-colored bars) and mode (darkly-colored bars) lengths of each secondary structure. Length is expressed as the number of amino acid residues. Error bars indicate standard deviations, although the sample distributions are not normal. (H) Percentages of amino acid residues that occupy secondary structures in the FL-DB. (I) Percentages of the number of the secondary structure units in the FL-DB.

by the PDB, even though one residue cannot form any helical structure by itself. Visual inspections of randomly chosen samples of short α-helices (1−4 amino acids, 15 samples) and 3_{10}-helices (1−3 amino acids, 32 samples) revealed that they form intramolecular hydrogen bonds with other amino acids nearby.

The length histogram for α-helices showed that the most frequent length was found at 6 amino acids, but another peak was found at 11 amino acids (Figure 1C). In contrast, 3_{10}-helices spanned from 1 to 34 amino acids in length, but mostly 3 to 5 amino acids (Figure 1D), indicating that 3_{10}-helices are generally shorter than α-helices. The two-peak patterns were found for both the α-helices and the 3_{10}-helices. On the other hand, *β*-strands spanned from 1 to 38 amino acids in length relatively smoothly (Figure 1E). Most of the "other" regions of proteins were small in size and peaked at 2 amino acids, although they were widely distributed from 1 to 189 amino acids (Figure 1F).

To understand the relationships among secondary structures, the mean and mode length of secondary structures are shown (Figure 1G), confirming the distribution patterns discussed above. Further, the number of amino acid residues

that belong to a given secondary structure was expressed as a percentage in a pie chart to depict the picture of an average protein (Figure 1H). It readily revealed that α-helix-constituting residues (37%) and residues that constitute the "other" region (37%) comprised the highest percentage of amino acids, followed by the β-strand-constituting (21%) and 3_{10}-helix-constituting (5%) residues. The relatively large proportion of α-helices and β-strands and the relatively small proportion of 3_{10}-helices were not surprising, but the significant proportion of 3_{10}-helices should not be ignored, and the compositional data for 3_{10}-helices should contribute to a more accurate understanding of protein structure. In addition, the fact that the "other" regions occupied a significant percentage might imply the functional importance of these less well-characterized chains in proteins.

The numbers of the secondary structure units were also expressed as a percentage in a pie chart to depict the picture of an average protein (Figure 1I). The α-helix units represented 19%, smaller than that of the amino acid percentage in Figure 1H (37%), suggesting that the unit length of α-helices is relatively large, as shown in Figure 1C and G. The largest number of units was seen for the "other" regions (48%), followed by β-strands (26%), suggesting that the unit lengths of the "other" regions and β-strands are relatively small. From these results, we obtained the picture of a typical protein of approximately 100 amino acids in length that was made of small stretches of secondary structures and of "other" segments in this proportion.

It is well-known that α-helices are stabilized by the intrahelical hydrogen bonds between the amide group (>N−H) of the *i*-th amino acid and the carbonyl group (>C═O) of the $(i - 4)$-th amino acid (often written as $i \rightarrow i - 4$), as predicted by the Pauling and Corey model. In 3_{10}-helices, the helical turns are more tight with the $i \rightarrow i - 3$ intrahelical hydrogen bonds. Obviously, small helix units cannot form such intrahelical hydrogen bonds. Nevertheless, the unit size was relatively small, and we even found very small helix units, as discussed above, assuming that the PDB definition is correct. For the existence of these small helix units to be possible, nonintrahelical hydrogen bonds or unconventional stabilizing interactions may be necessary, being similar to the end effects to form and to stabilize helices.[28–30]

3.2. The Amino Acid Occurrence in Secondary Structures. We next examined the amino acid occurrence (synonymously called composition, frequency, or count, in various cases) in each secondary structure and expressed it as a percentage (Figure 2A and C). We examined the amino acid occurrence of the FL-DB or "all" and also that of the NR-AA database (Table 1). Although we detected some differences especially in S, G, and L, these two data sets were largely similar to each other. The amino acid occurrence of secondary structures was then divided by that of either the FL-DB or the NR-AA to produce the relative occurrence (Figure 2B; Table 2).

Some conspicuous amino acids in a certain secondary structure were found. To name amino acids with the relative occurrence ≥1.2 (high) and ≤0.8 (low) in the FL-DB-based data, α-helices had A, L, E, Q, R, and M with high values and P and G with low values. The well-known strand-breaking nature of P and G was evident. And β-strands showed high values for V, I, Y, F, W, T, and C and low values for P, D, N, E, G, Q, and K. The "other" regions showed high values for P, G, and D and low values for I, L, V, M, F, and A. On the other hand, 3_{10}-helices showed high values for P, D, W, and S and low values for I, V, M, and T, which was in contrast to α-helices and β-strands, where P was shown to be low.

3.3. Rank Order Differences of Amino Acid Occurrence among Secondary Structures. We next examined the rank order differences of amino acid occurrences among secondary structures because, even if the amino acid occurrence was similar between two secondary structures at first glance, their rank order among the 20 amino acids could be different. This might also signify a given secondary structure. The rank orders of amino acid occurrences in each secondary structure were compared to those in the "all" database (FL-DB) by simply subtracting the order of a given amino acid in a SS-DB from that of the same amino acid in the FL-DB. This simple operation revealed the structure-specific amino acids in terms of the rank order (Figure 2C).

Here, α-helices contained high-ranked K and R and low-ranked G and P, whereas 3_{10}-helices contained high-ranked P and low-ranked V and I. The compositional rank order of these two helices was very different from each other especially in P, consistent with the results from the relative occurrence above. The high-ranked K in α-helices was somewhat unexpected from an initial look at the relative occurrence. On the other hand, β-strands contained high-ranked I, F, and Y and low-ranked E, D, and P, showing a unique compositional ranking. The "other" regions showed high-ranked P and low-ranked L and V, which was somewhat similar to the rank order of 3_{10}-helices.

The rank order deviations from the "all" database as indicated by the ROD score for each structure revealed that β-strands were most deviated from the "all" database, followed by "other", 3_{10}-helices, and α-helices.

3.4. The Amino Acid Assignment to Secondary Structures. We further evaluated how each amino acid was "assigned" to each secondary structure. While the simple compositional analysis is intended to reveal sequence characters from the point of secondary structures, assignment is a reversed concept from the point of amino acid. The concept of assignment could be especially useful in empirically classifying the 20 amino acid species. Indeed, each amino acid showed a unique assignment pattern (Figure 2D). For example, as expected from the previous results, A was mainly used in α-helices, whereas P and G were mostly used in "other" regions. V and I showed high assignments in β-strands. Note that it is not correct to make such pie charts or similar ones with the occurrence or the relative occurrence discussed above. From the point of the relative occurrence, D appeared to prefer 3_{10}-helices, for example. However, simply because 3_{10}-helices are relatively rare, D is more likely to be found in other secondary structures, which is clearly seen in the assignment. The assignment necessarily reflects the abundance and length of secondary structure units in proteins, which sounds statistically reasonable only when a large number of protein samples are used.

To systematically evaluate these individual assignments, a dendrogram was produced together with a heat map based on the four percentage values for each amino acid (Figure 2E). Because of the use of the four percentage values, the results were not influenced by the frequency

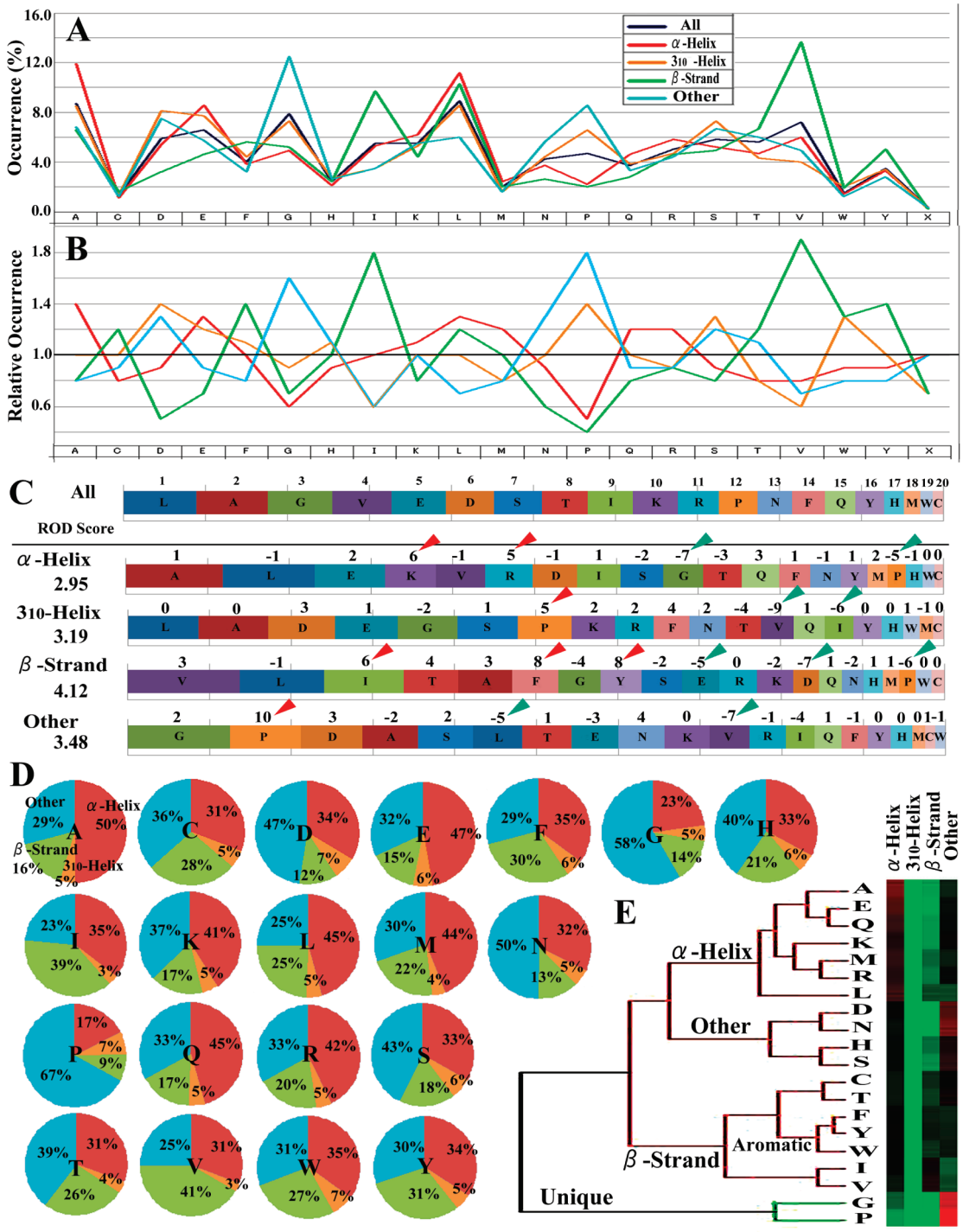

Figure 2. Amino acid occurrence and assignment in secondary structures. (A) Amino acid occurrence in four SS-DBs and FL-DB ("all"). (B) Relative amino acid occurrence in four SS-DBs and in FL-DB ("all"). (C) Percentages of amino acids and rank order comparison among secondary structures. The rank number from 1 to 20 are indicated along the bar obtained from the "all" database. In other bars obtained from the SS-DBs, the numbers along the bars indicate the rank order differences between the rank of an amino acid in a given secondary structure and that of the same amino acid in the "all" database. Red arrowheads indicate amino acids with the increased rank (5 or more up-ranked) in comparison to the "all", whereas green arrowheads indicate amino acids with decreased rank (5 or more down-ranked). Rank order distance (ROD) scores are indicated at the left end of the bars. (D) Amino acid assignment in 20 pie charts. (E) Characterization of amino acid relationships based on assignment values by cluster analysis with a heat map. The dendrogram shows uniqueness of G and P. In the heat map, red indicates higher values and green lower values.

differences among amino acids in proteins. This analysis revealed the assignment-based amino acid relationships. The uniqueness of P and G was conspicuous; they were mostly assigned to the "other" regions, as seen from the heat map. The cluster of A, E, Q, K, M, R, and L was composed of the α-helix-preferring amino acid residues, whereas the cluster of C, T, F, Y, W, I, and V was composed of the β-strand-preferring ones with similar preferences for α-helices. The cluster of D, N, H, and S appeared to be preferred by the "other" regions and also, to some extent, by α-helices.

The clustering pattern appeared to parallel the physicochemical nature of amino acids, to some extent. The aromatic amino acids, F, Y, and W, were clustered together, and they

Table 1. Amino Acid Occurrence in Proteins[a]

amino acid	FL-DB ("all")	NR-AA
L (leucine)	8.94	9.81
A (alanine)	8.72	8.38
G (glycine)	7.87	7.00
V (valine)	7.16	6.59
E (glutamic acid)	6.62	6.20
D (aspartic acid)	5.87	5.26
S (serine)	5.79	6.99
T (threonine)	5.58	5.56
K (lysine)	5.51	5.34
I (isoleucine)	5.46	5.77
R (arginine)	5.00	5.56
P (proline)	4.74	4.96
N (asparagine)	4.21	4.15
F (phenylalanine)	4.00	3.99
Q (glutamine)	3.68	4.00
Y (tyrosine)	3.49	3.01
H (histidine)	2.38	2.25
M (methionine)	2.00	2.37
W (tryptophan)	1.49	1.31
C (cysteine)	1.26	1.44

[a] Note: Numbers (%) were calculated based on either the FL-DB or the NR-AA.

Table 2. Relative Amino Acid Occurrence in Secondary Structures in Proteins[a]

amino acid	α-helix	3_{10}-helix	β-strand	other
A (alanine)	1.37(1.43)	0.98(1.02)	0.76(0.79)	0.78(0.81)
E (glutamic acid)	1.30(1.39)	1.17(1.25)	0.70(0.75)	0.86(0.92)
L (leucine)	1.26(1.15)	0.97(0.88)	1.16(1.06)	0.67(0.61)
Q (glutamine)	1.24(1.14)	1.03(0.95)	0.76(0.70)	0.89(0.82)
M (methionine)	1.20(1.01)	0.80(0.68)	1.00(0.84)	0.80(0.68)
R (arginine)	1.16(1.04)	0.94(0.85)	0.92(0.83)	0.88(0.79)
K (lysine)	1.13(1.17)	0.96(0.99)	0.80(0.83)	1.00(1.03)
D (aspartic acid)	0.92(1.03)	1.37(1.53)	0.54(0.60)	1.27(1.42)
W (tryptophan)	0.93(1.06)	1.33(1.51)	1.27(1.44)	0.80(0.91)
S (serine)	0.90(0.75)	1.26(1.04)	0.84(0.70)	1.16(0.96)
H (histidine)	0.88(0.93)	1.13(1.20)	1.00(1.06)	1.08(1.14)
V (valine)	0.83(0.90)	0.56(0.61)	1.90(2.06)	0.68(0.74)
I (isoleucine)	0.96(0.91)	0.64(0.61)	1.76(1.67)	0.64(0.61)
F (phenylalanine)	0.95(0.95)	1.10(1.10)	1.40(1.40)	0.80(0.80)
Y (tyrosine)	0.94(1.09)	0.97(1.12)	1.42(1.65)	0.80(0.93)
C (cysteine)	0.85(0.74)	1.00(0.88)	1.23(1.08)	0.92(0.81)
T (threonine)	0.84(0.84)	0.77(0.77)	1.20(1.20)	1.07(1.07)
P (proline)	0.47(0.45)	1.40(1.34)	0.43(0.41)	1.83(1.75)
G (glycine)	0.62(0.70)	0.92(1.03)	0.66(0.74)	1.58(1.78)
N (asparagine)	0.88(0.89)	1.05(1.07)	0.62(0.63)	1.33(1.35)

[a] Note: Relative occurrence was derived as a ratio of [occurrence in a given secondary structure] to [occurrence throughout proteins]. It, thus, shows relative abundance in secondary structures in comparison to entire protein sequences. Numbers without parentheses were derived from the FL-DB-based data, whereas numbers in parentheses were derived from the NR-AA-based data. The first block amino acids (A−K) have the highest value in α-helices, the second block amino acids (D−H) in 3_{10}-helices, the third block amino acids (V−T) in β-strands, and the fourth block amino acids (P, G, and N) in the "other" regions.

had roughly equal assignments to α-helices, β-strands, and "other" regions. The nonpolar aliphatic V and I were associated with each other, and they were assigned preferably to β-strands. Also noteworthy was the clustering of E and Q as well as D and N.

3.5. Comparison to the Classical Studies. Strictly speaking, a simple comparison between the classical and present studies is not possible because the classical analysis in Williams et al.[9] was performed from different viewpoints and operations, using a relatively small amount of input data. Our analysis here is based on simple ideas of the occurrence, rank order, and assignment derived from the counts of amino acid residues in four different secondary structure regions, using a large amount of input data. Furthermore, the inclusion of 3_{10}-helices here is a novel attempt. Nonetheless, our results here appeared to be largely consistent with the classical view of amino acid preferences for secondary structures,[9] albeit there were some differences. In Williams et al.[9] and in this study, A, E, Q, K, M, R, and L were α-helix-preferring amino acids. The amino acid H, which was considered to be weakly α-helix-preferring in Williams et al.,[9] was more 3_{10}-helix- and "other"-preferring in this study. The β-strand-preferring residues in Williams et al.,[9] V, I, Y, C, W, F,and T, were also consistent with this study. Further studies are expected to differentiate between parallel and antiparallel β-strands. The turn-preferring D, N, and S in Williams et al.[9] were also associated together in the dendrogram, but they were not only "other"-preferring but also 3_{10}-helix-preferring in this study. Also, the uniqueness of P and G was well highlighted in this study, being consistent with their well-known nature as helix breakers.[30–35]

In general, hydrophobicity,[36–38] steric factors,[39] and electronic properties[40,41] are considered to be important factors to determine secondary structures. Although none of these simple categories of amino acids can solely be applicable to a specific secondary structure, to summarize our results very roughly, α-helices are likely to be composed of relatively hydrophilic and aliphatic residues, and β-strands are likely to be composed of relatively hydrophobic and aromatic residues, in addition to aliphatic ones.

We also noted that our simple compositional data (Tables 1 and 2) were quite different from those of classical studies,[27,42] especially for C (2.14[27] versus 1.26 [FL-DB] and 1.44 [NR-AA]) and A (7.75[27] versus 8.72 [FL-DB] and 8.38 [NR-AA]). This point exemplifies the importance of use of the "post-genomic" databases in such compositional analyses.

3.6. The Doublet and Triplet Availability Scores. We have previously shown that the usage of short constituent amino acid sequences in proteins are biologically biased, and in extreme cases, some of them are completely absent in any proteins.[21–23] This approach to protein decoding has also been expanded by other groups.[43–45] Here, we focused on doublets and triplets, two and three consecutive amino acid combinations in protein chains and in secondary structures.

In the SS-DBs excluding X residues, we had 142 871 doublets in the α-helix database, 17 035 doublets in the 3_{10}-helix database, 73 929 doublets in the β-strand database, and 124 847 doublets in the "other" database. The FL-DB (the "all" database) had 419 805 doublets. Because there are 400 doublet species, assuming that all doublet species appear in the database equally frequently, we expect each doublet to appear 357 times in the α-helix database, 43 times in the 3_{10}-helix database, 185 times in the β-strand database, 312 times in the "other" database, and 1050 times in the "all" database. In reality, the assumed equal frequency is not simply true, but these numbers can serve as indicators to judge the feasibility of the analyses performed below.

The absolute and relative doublet counts were defined (see Materials and Methods Section), and we calculated the

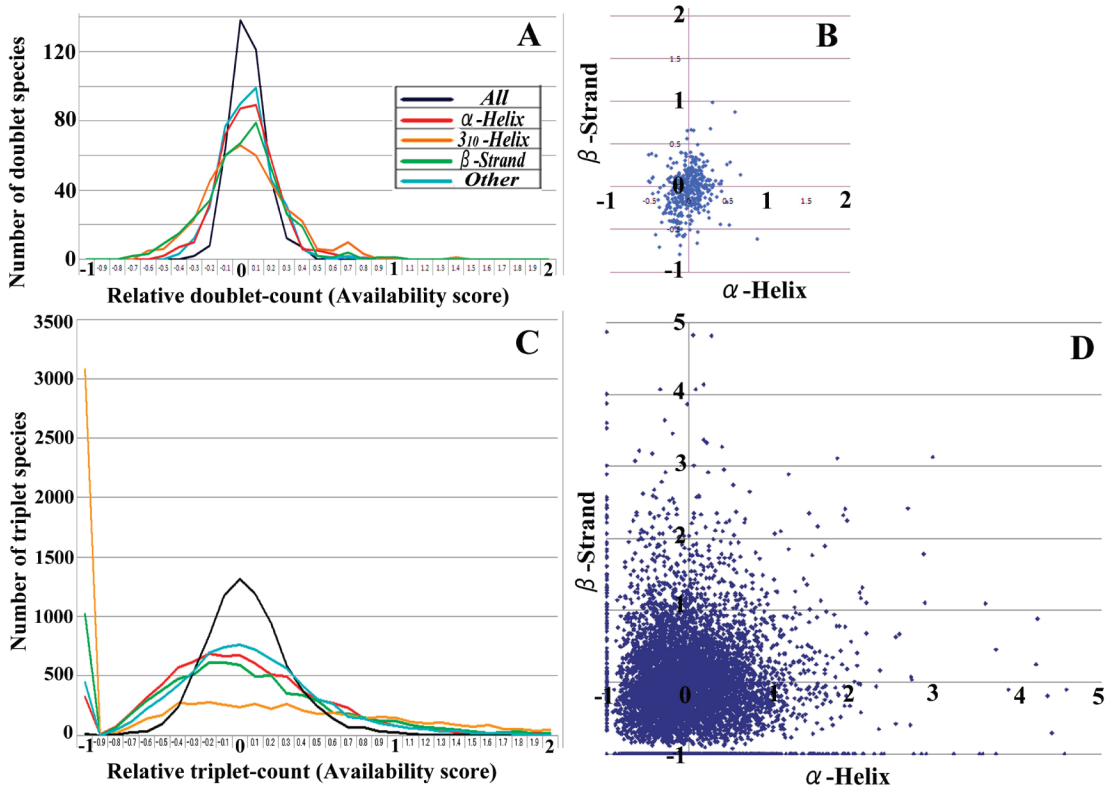

Figure 3. Doublets and triplet availability scores in secondary structures. (A) Distribution of relative doublet count (availability score). (B) Scatter plot of 400 doublets between α-helix and β-strand DBs. (C) Distribution of relative triplet count (availability score). (D) Scatter plot of 8000 triplets between α-helix and β-strand DBs. Note the existence of zero-count triplets (availability score −1) in the scatter plot.

doublet availability scores in each database, and the distribution patterns of the doublet availability scores were examined (Figure 3A). The distribution patterns of the SS-DBs were evidently different from those of the FL-DB, although they all peaked around zero. Note that an even sharper distribution pattern around zero is expected if there is no biological bias in doublet usage.[21–23] Differences in the distribution patterns among the SS-DBs may be attributed at least partly to the differences in the number of doublets in each database. Nevertheless, these non-normal distribution patterns are consistent with the notion that the doublet usage is more biased in secondary structures than in the entire protein.

We constructed scatter plots of the 400 doublets in the combinations of two secondary structures (Figure 3B). The scatter plots between the α-helix and the β-strand databases showed that most doublets were found around zero, which means that most doublets show up in any secondary structures.

A similar analysis was performed for triplets. In the SS-DBs excluding X residues, we had 130 620 triplets in the α-helix database, 12 500 triplets in the 3_{10}-helix database, 56 804 triplets in the β-strand database, and 98 293 triplets in the "other" database. The FL-DB (the "all" database) had 414 283 triplets. Because there are 8000 triplet species, assuming that all triplet species appear in the database equally frequently, we expect each triplet to appear 16.3 times in the α-helix database, 1.6 times in the 3_{10}-helix database, 7.1 times in the β-strand database, 12.3 times in the "other" database, and 51.8 times in the "all" database.

The distribution patterns of the triplet availability scores of the SS-DBs were evidently different from those of the FL-DB, although they peaked at around zero (Figure 3C). The distribution patterns were flatter than those of doublets. Differences in the distribution patterns among the SS-DBs may be attributed at least partly to the difference in the number of triplets in each database. Nevertheless, these non-normal distribution patterns are consistent with the notion that the triplet usage is more biased in secondary structures than in the entire protein, and therefore, there seem to be triplets that are preferred and those that are "avoided" or "forbidden" in a given secondary structure.

We constructed scatter plots of the 8000 triplets in the combinations of two secondary structures. As was shown in the preliminary study,[23] the scatter plots between the α-helix and the β-strand DBs showed that most triplets were found in any secondary structures, but some triplets appeared to be found exclusively in one of the two secondary structures (Figure 3D). We observed many spots with an availability score of −1, which indicates the lack of existence in a database. Similar results were obtained for any combination of two SS-DBs (not shown). These results indicate the triplet usage bias of secondary structures and the existence of secondary structure-specific triplets, which encouraged us to identify unique triplets that exist in one secondary structure but not in others.

3.7. High-Score Doublets and Triplets and Zero-Count Triplets. We here, first, list the doublets with high occurrences or with high availability scores in a given secondary structure (Table 3). For the occurrence, α-helices were dominated by a group of doublets composed of A, L, and E, which was similar to the "all" database but different from the rest of the SS-DBs. And β-strands were dominantly occupied by a group of doublets with V, L, A, and I. These features were mostly consistent with the compositional analysis discussed before in this paper. For the availability

Table 3. High-Score Doublets with Absolute Count (Occurrence) and Relative Count (Availability) in Secondary Structures[a]

	α-helix		3_{10}-helix		β-strand		other		all	
rank	occurrence	availability	occurrence	availability	occurrence	availability	occurrence	availability	occurrence	availability
1	AA(2330)	PE(0.87)	PE(264)	PE(2.03)	VV(1337)	WQ(0.98)	GG(1488)	HH(2.31)	AA(3716)	HH(0.98)
2	AL(2244)	TP(0.65)	DL(165)	HC(1.40)	VL(1113)	CC(0.87)	DG(1380)	CC(0.94)	AL(3572)	CC(0.78)
3	LA(2082)	CC(0.58)	AA(162)	DI(0.80)	VI(1007)	WN(0.68)	GA(1293)	LP(0.68)	LA(3412)	CH(0.40)
4	LL(1780)	MM(0.52)	EL(161)	DW(0.78)	LV(1006)	WD(0.66)	PG(1226)	IP(0.62)	LL(3237)	WQ(0.38)
5	EA(1650)	DP(0.51)	PA(154)	CW(0.72)	IV(856)	ID(0.66)	GS(1183)	YW(0.46)	AG(2907)	MM(0.35)
6	EL(1448)	HF(0.48)	LA(148)	KY(0.70)	LL(795)	QW(0.65)	SG(1104)	VP(0.42)	LG(2741)	HP(0.35)
7	LE(1447)	DW(0.48)	EE(140)	WQ(0.68)	VA(793)	FD(0.56)	AG(1094)	NW(0.35)	GA(2693)	HC(0.33)
8	AE(1435)	HH(0.41)	GG(134)	NW(0.64)	AV(775)	WH(0.49)	LP(1076)	DW(0.33)	VA(2686)	PE(0.31)
9	EE(1253)	CH(0.41)	SA(132)	WN(0.64)	TV(745)	MN(0.40)	GK(1042)	HM(0.33)	GL(2683)	NP(0.30)
10	LK(1240)	QQ(0.41)	SL(132)	CT(0.63)	VT(735)	FS(0.40)	GD(1022)	WD(0.32)	VL(2667)	QQ(0.30)

[a] Absolute count and relative count are shown in parentheses. These counts were obtained from a given SS-DB.

Table 4. High-Score Triplets with Absolute Count (Occurrence) and Relative Count (Availability) in Secondary Structures[a]

	α-helix		3_{10}-helix		β-strand		other		all	
rank	occurrence	availability	occurrence	availability	occurrence	availability	occurrence	availability	occurrence	availability
1	AAA(291)	YCC(5.3)	PEL(37)	CMM(21.2)	VVV(145)	CCN(6.8)	HHH(144)	HHH(82.9)	AAA(408)	HHH(26.6)
2	ALA(283)	GCH(4.6)	PED(24)	CTM(16.0)	VLV(134)	CQC(6.1)	DGS(143)	WCM(11.1)	AAL(377)	CCN(2.6)
3	AAL(280)	HHH(4.5)	EEL(21)	MCH(12.3)	VAV(114)	HWH(5.7)	GAG(120)	CCF(7.4)	ALA(377)	MHC(2.2)
4	LAA(262)	TPE(4.4)	PEE(21)	IHC(11.3)	LVL(112)	WQM(5.5)	PDG(120)	CCV(7.1)	LAA(368)	CCV(2.2)
5	EAL(250)	HFC(4.3)	DPA(20)	MMG(11.3)	LVV(112)	WWH(5.3)	PEG(118)	EHH(5.6)	ALL(328)	CCS(2.1)
6	ALL(229)	SCH(4.2)	GGG(19)	HCG(10.8)	VLL(108)	PCW(4.9)	APG(113)	MQW(5.1)	EAL(322)	SCH(2.0)
7	LLA(222)	CRW(4.2)	NPE(19)	ICT(10.7)	VVL(107)	HWM(4.8)	DGK(104)	NCC(4.9)	LLA(320)	YCC(2.0)
8	EAA(214)	CHG(3.9)	PEA(18)	VCH(9.8)	VVI(106)	QWQ(4.8)	GSG(104)	CMH(4.8)	LAE(297)	RCC(2.0)
9	LAE(205)	CCR(3.7)	DPR(17)	HCW(9.7)	AVV(102)	EWH(4.1)	LPG(101)	MHC(4.8)	AAG(296)	HHM(2.0)
10	ALE(197)	RCC(3.7)	SAL(17)	CYM(9.5)	VIV(102)	DWQ(4.1)	GLP(100)	HHM(4.5)	ALE(292)	CQC(1.9)

[a] Absolute count and relative count are shown in parentheses. These counts were obtained from a given SS-DB.

scores, PE and WQ had the highest score in 3_{10}-helix and β-strand, respectively.

The list of high-score triplets showed similar but slightly different features (Table 4). For the occurrence, α-helices were dominated by a group of triplets composed of A, L, and E, which was similar to the "all" database but different from the rest of the SS-DBs. And β-strands were dominantly occupied by a group of triplets with V, L, A, and I. For the availability scores, all databases showed high-score triplets that were dominated by C. Q-containing triplets were seen only in the β-strand database. However, at this point, it is difficult to point out features that may be specific to a given secondary structure.

We then looked for some doublets and triplets that appeared in one SS-DB (availability score > 0) but not in the other two SS-DBs, except the "other" database. In this category, we detected no doublets, but we did detect 194 triplets for α-helices, 33 triplets for 3_{10}-helices and 97 triplets for β-strands. Among these, the triplets with high availability scores are listed in Table 5, although they showed up just five times or fewer in a given secondary structure. Some of them were not found in the "other" database, either. These triplets may be used for the identification of sequential segments that form a given secondary structure from the primary amino acid sequences.

We noticed that there were some triplets that did not appear in a given secondary structure at all, despite the probabilistic expectation for larger counts. We called these the zero-count triplets and listed them in the order of decreasing probabilistic expectation, expressed as E in eq 4 (Table 6). For example, the zero-count triplet DTI was expected to appear 12 times in the β-strand database but did not appear at all. Interestingly, these zero-count triplets were not always constructed by the amino acids with low compositional percentages. In β-strands, T and I were used very frequently, as shown in Figure 2, which means that the triplet effect cannot be reduced to a single amino acid level. Although they may be found in a secondary structure 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 a given secondary structure. This avoidance could be simply because of physicochemical structural requirements for secondary structures. Alternatively, functional restriction of each secondary structure could have resulted in this avoidance of specific triplets.

Table 5. Triplets Specific to Secondary Structures[a]

rank	α-helix	3_{10}-helix	β-strand
1	CCR(4, 3.7)*	CTM(2, 16.0)	CCN(3, 6.8)
2	CPW(2, 3.6)	MCH(1, 12.3)*	CQC(3, 6.1)
3	CYC(2, 3.2)	SMC(2, 8.9)	ECW(4, 4.0)*
4	GWW(5, 2.8)	FMC(1, 7.3)	WHC(2, 3.9)*
5	CNC(2, 2.7)*	MCN(1, 7.2)	CQH(3,3.9)
6	WWP(2, 2.4)	MNC(1, 7.2)	CKC(3, 3.6)*
7	CHC(1, 2.2),	WCT(1, 5.8)*	MHC(2, 3.5)
8	WWM(2, 2.1)*	PWN(5, 5.7)	WHP(2, 3.0)
9	QCC(2, 2.0)*	NWC(1, 5.6)	WMH(2, 2.9)*
10	MWN(5, 2.0)	PCC(1, 5.6)	CCF(3, 2.6)

[a] Triplets are ranked according to their availability scores (relative counts). Absolute and relative counts are shown in parentheses. These counts were obtained from a given SS-DB. Asterisk (*) indicates their absence also in the "other" database. Note that highly ranked triplets in this table also appear in Table 4.

Table 6. Zero-Count Triplets with the Highest Probabilistic Counts (Expected Counts) in Secondary Structures[a]

rank	α-helix	3_{10}-helix	β-strand	other	all
1	KPR (10.6)	DAP (5.8)	DTI (12.0)	IRK (8.4)	WCK (4.3)
2	KHR (10.1)	SDE (5.8)	FFR (8.4)	IIL (7.1)	CMF (4.2)
3	DSP (8.2)	DDP (5.5)	IQG (8.1)	PCN (5.9)	WNC (3.3)
4	RGH (8.0)	LEP (5.5)	DQV (7.2)	FEQ (5.8)	MWH (2.9)
5	RPT (8.0)	AEP (5.5)	PEV (7.1)	LTM (5.8)	MWM (2.5)
6	QPR (7.8)	LSP (5.3)	SNI (6.9)	QYT (5.5)	MMC (2.1)
7	AHP (7.4)	SLP (5.3)	DTG (6.4)	KML (5.3)	WMC (1.6)
8	HNK (6.4)	DEP (5.2)	GSK (6.3)	SVM (5.3)	WWW (1.4)
9	NRP (6.3)	EDP (5.2)	DDL (6.2)	EQY (5.3)	CWW (1.2)
10	CGE (5.8)	AGP (5.2)	TDS (6.1)	MNE (5.2)	WCW (1.2)

[a] Expected counts in a given secondary structure database are indicated in parentheses.

4. CONCLUSIONS

To our knowledge, the present study is the most comprehensive compositional study performed in this postgenome era. This empirical study employed very simple concepts without complicated mathematical operations, which directly illuminated important features of secondary structures. The recent development of several secondary structure prediction programs has achieved a three-state prediction accuracy (Q_3) score of more than 80%.[44–48] We believe that the use of the empirical information presented in this study, which included 3_{10}-helices as one of the secondary structures, can contribute to an even more accurate prediction of secondary structures from the primary sequences, in combination with other programs. Also, de novo structure predictions without useful templates using a lowest free-energy structure search program, such as Rosetta,[51–53] could be facilitated by our empirical data.

The biological significance of a group of triplets that are specific to a given secondary structure is not clear at this point. We do not know the significance of the zero-count triplets in secondary structures, either, although they can serve as a marker for a given secondary structure in a prediction program. There would be no difficulty in synthesizing zero-count triplets, because we have already demonstrated that even zero-count pentats (five consecutive amino acid sequences that do not exist in the NR-AA database) can be synthesized chemically and biologically without any problem.[22] We speculate that, as discussed in Tuller et al.,[44] specific triplets are preferred for structural rigidity and that the zero-count triplets are avoided or forbidden to avoid breaking the secondary structures. Experimental validation of this hypothesis not only in silico but also in vitro is expected in the future.

ACKNOWLEDGMENT

We thank K. Motomura for programming and examining the NR-AA database, K. Hamano and A. Hiyama for data input and preliminary analysis, and Prof. M. Nakamura and Prof. T. Okazaki for helpful discussion. This research was supported by the Takeda Research Foundation and also by the 21st Century COE program of the University of the Ryukyus, Japan.

Supporting Information Available: The list of protein samples analyzed in the present study. This information is available free of charge via the Internet at http://pubs.acs.org./.

REFERENCES AND NOTES

(1) Chou, P. Y.; Fasman, G. D. Prediction of protein conformation. *Biochemistry* **1974**, *13* (2), 222–145.
(2) Chou, P. Y.; Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequences. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1978**, *47*, 45–148.
(3) Lim, V. I. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.* **1974**, *88* (4), 857–872.
(4) Lim, V. I. Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.* **1974**, *88* (4), 873–894.
(5) Garnier, J.; Osguthorpe, D. J.; Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **1978**, *120* (1), 97–120.
(6) Nagano, K. Triplet information in helix prediction applied to the analysis of super-secondary structures. *J. Mol. Biol.* **1977**, *109* (2), 251–274.
(7) Berg, J. M.; Tymoczko, J. L.; Stryer, L. *Biochemistry*, 6th ed.; W. H. Freeman: New York, 2007.
(8) Creighton, T. E. *Proteins: Structures and Molecular Properties*, 2nd ed.; W. H. Freeman: New York, 1993.
(9) Williams, R. W.; Chang, A.; Juretić, D.; Loughran, S. Secondary structure predictions and medium range interactions. *Biochim. Biophys. Acta* **1987**, *916* (2), 200–204.
(10) Nakashima, H.; Nishikawa, K.; Ooi, T. The folding type of a protein is relevant to the amino acid composition. *J. Biochem.* **1986**, *99* (1), 153–162.
(11) 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* (10), 2809–2813.
(12) Muskal, S. M.; Kim, S.-H. Predicting protein secondary structure content: A tandem neural network approach. *J. Mol. Biol.* **1992**, *225* (3), 713–727.
(13) Zhang, C. T.; Zhang, Z.; He, Z. Prediction of the secondary structure content of globular proteins based on structural classes. *J. Protein Chem.* **1996**, *15* (8), 775–786.
(14) Eisenhaber, F.; Imperiale, F.; Argos, P.; Frommel, C. Prediction of secondary structural contents of proteins from their amino acid composition alone. I. New analytic vector decomposition methods. *Proteins: Struct. Funct. Genet.* **1996**, *25* (2), 157–168.
(15) Ruan, J.; Wang, K.; Yang, J.; Kurgan, L. A.; Cios, K. Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. *Artif. Intell. Med.* **2005**, *35* (1−2), 19–35.
(16) Lee, S.; Lee, B. C.; Kim, D. Prediction of protein secondary structure content using amino acid composition and evolutionary information. *Proteins: Struct. Funct. Bioinf.* **2006**, *62* (4), 1107–1114.
(17) Gibrat, J.-F.; Garnier, J.; Robson, B. Further development of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.* **1987**, *198* (3), 425–443.
(18) Chou, K. C. Using pair-coupled amino acid composition to predict protein secondary structure content. *J. Protein Chem.* **1999**, *18* (4), 473–480.
(19) Chen, C.; Chen, L.; Zou, X.; Cai, P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein Pept. Lett.* **2009**, *16* (1), 27–31.
(20) 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* (1), 253–242.
(21) 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. Annu. Rev.* **2008**, *14*, 109–141.
(22) Otaki, J. M.; Ienaka, S.; Gotoh, T.; Yamamoto, H. Availability of short amino acid sequences in proteins. *Protein Sci.* **2005**, *14* (3), 617–625.
(23) 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* (3), 25–38.
(24) Noguchi, T.; Matsuda, H.; Akiyama, Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank. *Nucleic Acids Res.* **2001**, *29* (1), 219–220.
(25) Hutchinson, E. G.; Thornton, J. M. PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci.* **1996**, *5* (2), 212–220.
(26) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22* (12), 2577–2637.
(27) Cserzö, M.; Simon, I. Regularities in the primary structure of proteins. *Int. J. Pept. Protein Res.* **1989**, *34*, 184–195.

(28) Presta, L. G.; Rose, G. D. Helix signals in proteins. *Science* **1988**, *240* (4859), 1632–1641.
(29) Richardson, J. S.; Richardson, D. C. Amino acid preferences for specific locations at the ends of α helices. *Science* **1988**, *240* (4859), 1648–1652.
(30) Aurora, R.; Rose, G. D. Helix capping. *Protein Sci.* **1998**, *7* (1), 21–38.
(31) Piela, L.; Nemethy, G.; Scheraga, H. A. Proline-induced constraints in α-helices. *Biopolymers* **1987**, *26* (9), 1587–1600.
(32) O'Neil, K. T.; DeGrado, W. F. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* **1990**, *250* (4981), 646–651.
(33) MacArthur, M. W.; Thornton, J. M. Influence of proline residues on protein conformation. *J. Mol. Biol.* **1991**, *218* (2), 397–412.
(34) Aurora, R.; Srinivasan, R.; Rose, G. D. Rules for α-helix termination by glycine. *Science* **1994**, *264* (5162), 1126–1130.
(35) Chakrabarti, P.; Chakrabarti, S. C-H. . .O hydrogen bond involving proline residues in α-helices. *J. Mol. Biol.* **1998**, *284* (4), 867–873.
(36) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157* (1), 105–132.
(37) Rose, G.; Geselowitz, A.; Lesser, G.; Lee, R.; Zehfus, M. Hydrophobicity of amino acid residues in globular proteins. *Science* **1985**, *229* (4716), 834–838.
(38) Dill, K. A. Dominant forces in protein folding. *Biochemistry* **1990**, *29* (31), 7133–7155.
(39) Creamer, T. P.; Rose, G. D. Side-chain entropy opposes α-helix formation but rationalizes experimentally determined helix-forming propensities. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89* (13), 5937–5941.
(40) Dwyer, D. S. Electronic properties of the amino acid side chains contribute to the structural preferences in protein folding. *J. Biomol. Struct. Dyn.* **2001**, *18* (6), 881–892.
(41) Dwyer, D. S. Electronic properties of amino acid side chains: quantum mechanics calculation of substituent effects. *BMC Chem. Biol.* **2005**, *5*, 2.
(42) Vonderviszt, F.; Mátrai, G. Y.; Simon, I. Characteristic sequential residue environment of amino acids in proteins. *Int. J. Pept. Protein Res.* **1986**, *27*, 483–492.
(43) 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.
(44) Tuller, T.; Chor, B.; Nelson, N. Forbidden penta-peptides. *Protein Sci.* **2007**, *16* (10), 2251–2259.
(45) Bresell, A.; Persson, B. Characterization of oligopeptide patterns in large protein sets. *BMC Genomics* **2007**, *8*, 346.
(46) Rost, B. Prediction in 1D: secondary structure, membrane helices, and accessibility. In *Structural Bioinformatics*; Bourne, P. E., Weissig, H., Ed.; Wiley-Liss: Hoboken, NJ, 2003, pp559−587.
(47) Ward, J. J.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Secondary structure prediction with support vector machines. *Bioinformatics* **2003**, *19* (13), 1650–1655.
(48) Lin, H. N.; Chang, J. M.; Wu, K. P.; Sung, T. Y.; Hsu, W. L. HYPROSP II: a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* **2005**, *21* (15), 3227–3233.
(49) Montgomerie, S.; Sundararaj, S.; Gallin, W. J.; Wishart, D. S. Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics* **2006**, *7*, 301.
(50) Chang, D. T. H.; Ou, Y. Y.; Hung, H. G.; Yang, M. H.; Chen, C. Y.; Oyang, Y. J. Prediction of protein secondary structures with a novel kernel density estimation based classifier. *BMC Res Notes* **2008**, *1*, 51.
(51) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **2004**, *383*, 66–93.
(52) Bradley, P.; Misura, K. M.; Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **2005**, *309* (5742), 1868–1871.
(53) Raman, S.; Vernon, R.; Thompson, J.; Tyka, M.; Sadreyev, R.; Pei, J.; Kim, D.; Kellogg, E.; DiMaio, F.; Lange, O.; Kinch, L.; Sheffler, W.; Kim, B.-H.; Das, R.; Grishin, N. V.; Baker, D. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Struct. Funct. Bioinf.* **2009**, *77* (Suppl 9), 89–99.

CI900452Z