論文

Frequency Distribution of the Number of Amino Acid Triplets in the Non-Redundant Protein Database
重複を除いた蛋白質データベースにおける3アミノ酸組の出現数の頻度分布

Joji M. OTAKI[*,‡]　**Tomonori GOTOH**[†]
and Haruhiko YAMAMOTO[*,§]

大瀧 文二 [*,‡]　**後藤 智範** [†]　**山本 晴彦** [*,§]

Protein molecules are polymers of amino acids linked by peptide bonds, and they play various roles in innumerable biological functions. This remarkable functional diversity of biological proteins originates from linear sequences of 20 different amino acid residues. Their sequence information, which is encoded in genes as DNA sequences, is a product of molecular evolution at the genetic level. Upon completion of many genome projects, amino acid sequence records of proteins in databases, which include conceptually translated sequences from DNA, have already been accumulated over 1.24 million, and more than ever, the number of records is still increasing rapidly. Although these sequence databases have been mainly used for similarity searches, fundamental characters of these databases have not been examined thoroughly. Here we investigated biological significance of 8000 combinatorial sets of three amino acids (triplets) in proteins. Defining the number of each triplet in a database as "triplet count", we constructed a histogram for the frequency distribution of triplet counts in the non-redundant protein (nr-aa) database downloaded from the National Center for Biotechnology Information as of November 2002. Distribution range of the histogram was shown to be larger than that of the theoretical histogram generated randomly from the population having the amino acid composition of the nr-aa database, although overall shapes of these histograms were similar to each other. The difference between these two distributions was more dramatically highlighted in histograms showing the ratio of the original triplet counts in the nr-aa database or of the theoretical triplet counts generated randomly to the expected triplet counts derived from the amino acid composition in the database. Whereas the theoretical distribution well fitted the normal error curve due to the random fluctuations inherently associated with the sampling procedure itself, the distribution for the existing triplets in the nr-aa database peaked much less and skewed much more toward higher values than the theoretical one, indicating a non-random and possibly biological nature of triplet counts in the nr-aa database. We also performed the same procedure in five phylogenetically distinct species: human (*Homo sapiens*), mouse (*Mus musculus*), fruit fly (*Drosophila melanogaster*), soil nematode (*Caenorhabditis elegans*), and a colon bacterium (*Escherichia coli*). We found similar trends in all species examined here, largely excluding the possibility that the characteristic trend of the triplet-count distribution that was found using the entire database records might have resulted solely from some "sampling artifacts" of the database itself. In other words, the existence of the species-independent distribution trend mostly ruled out the possibility that the nr-aa database over-represents or under-represents particular kinds of proteins simply because of the arbitrary research history of biological sciences. Taken together, this study suggested the existence of non-random and species-independent biological preferences for particular triplets in proteins at the population level, which might have been "fixed" either accidentally or for functional reasons early during the course of biological evolution.

蛋白質のアミノ酸配列情報は，近年顕著に増加している．この研究では，重複を除いたアミノ酸配列データベース中に存在する3アミノ酸組 (トリプレット) の出現数 (トリプレット

数) について統計的に調べた．実際のトリプレット数の頻度分布は，データベースのアミノ酸組成を基礎としてランダムに発生させた理論的な分布よりも分布範囲が広いが，全体の傾向には大きな違いは見られなかった．これら二つの頻度分布の相違は，それぞれのトリプレット数とデータベースのアミノ酸組成から期待される期待トリプレット数との比を求めることで明確となった．理論的な分布が無作為な標本抽出過程から生じる正規分布を示したのに対し，実際のトリプレット数の分布はかなり幅広い歪んだ分布を示した．同様な頻度分布は生物種別に調べても得られた．このことは，これらの分布傾向は人為的なデータの偏りに起因するのではなく，生物学的なデータの性質に起因することを示唆している．

Keywords: non-redundant protein database, protein sequence, amino acid composition, amino acid triplet, triplet count, frequency distribution
非重複蛋白質データベース，蛋白質配列，アミノ酸組成，3 アミノ酸組 (アミノ酸トリプレット)，トリプレット数，頻度分布

1 Introduction

Among various types of biological molecules, two chemically distinct groups of macromolecules are of great importance in molecular biology: nucleic acids (DNA and RNA) and proteins. They are informational molecules whose sequences of monomers primarily determine their structural and functional specificities. DNA and RNA molecules are polymers of 4 different nucleotides linked by phosphodiester bonds, whereas protein molecules are polymers of 20 different amino acids linked by peptide bonds. Protein sequence information is encoded in genes as DNA sequences. Once known, a DNA sequence of a particular gene can conceptually be translated into an amino acid sequence of "hypothetical" protein, according to the semi-universal genetic code.

Proteins are an almost exclusive source for innumerable catalytic reactions and a main source for structural measures in biological systems. This remarkable functional diversity of biological proteins originates from linear sequences of 20 different amino acids, and it is a product of long evolutionary history of biological systems at the genetic level. Since sequences of amino acids are of primary importance in protein research, algorithms for sequence similarity searches have been developed extensively[1–4], resulting in several programs such as BLAST (basic local alignment search tool), one of the most popular web-based search system[5][6]. Since early 1980s, these algorithms use the "word hit" strategy, defining a "word" for two or three amino acid residues as a seed for further alignment[7].

In spite of the primarily linear nature of proteins, they are three-dimensional existence in their functional forms. From the primary sequences, it has been difficult to deduce three dimensional structures of proteins with limited success in 1970's[8–10]. Since then, more experimental three-dimensional data have been accumulated, and accordingly many efforts have been made to computationally extract structural information from linear protein sequences through intensive similarity searches for proteins with three-dimensional data[3][4][11][12]. These structure predictions

* Department of Biological Sciences, Kanagawa University
神奈川大学理学部生物科学科
‡ E-mail: otaki@bio.kanagawa-u.ac.jp
§ E-mail: yamamoto@bio.kanagawa-u.ac.jp
† Department of Information and Computer Science, Kanagawa University
神奈川大学理学部情報科学科
E-mail: gotoh@info.kanagawa-u.ac.jp

including those of 1970's take advantage of structural redundancy in proteins, especially, small secondary structural units called α-helix and β-strand[11][12]. These efforts yielded several prediction programs such as PredictProtein[13] and PSIpred[14].

These computer-aided biological sciences have recently shown remarkable advancement with the expansion and maintenance of the web-based databases for biological research community. Upon completion of many genome projects including those of human and mouse, amino acid sequence records of proteins in databases have already been accumulated over 1.24 million, and more than ever, the number of records is still exponentially increasing, although a significant proportion of records are "hypothetical proteins" that are products of conceptual translation of DNA. Much attention is being paid to these protein databases in this "post-genome" era, as proteome research advances[15][16]. However, fundamental characteristics of these databases have not been examined thoroughly.

Here we considered 8000 combinatorial sets of three amino acids (triplets) as a unit of information, on the assumption that the amino acid triplets in proteins could be of biological significance. This reflects the fact that key residues in "active sites" of proteins are often composed of small number of amino acids, although these sites must be three-dimensionally supported by other residues to form characteristic structures[17][18]. Many hairpin loops, which often participate in active sites, are composed of just a few residues, usually 3–5 residues in length. Packing of two α-helices is made between the ridge of one helix and groove of other helix, which are mostly made of 3 or 4 residues[19]. Functional and evolutionary significance of such short stretch of amino acid sequences among the existing proteins is of our interest in this study.

For convenience, we here define the number of each triplet in a database as "triplet count". We also interchangeably use the term "triplet composition". Similarly, we define the number of each amino acid residue in a database as "amino acid count" or "amino acid composition". Triplet count and amino acid count can be expressed as absolute number of count, percentage, or probability. According to this definition, we statistically examined the frequency distribution of triplet counts in the non-redundant protein (nr-aa) database downloaded from the National Center for Biotechnology Information[20]. We constructed a histogram for triplet counts in the database, together with a histogram for the randomly generated theoretical triplet counts. These two frequency distributions were compared to each other, and we concluded that their difference seemed to originate not from an inherent bias of the database itself but from some biological consequences of either accidental or functional nature.

2 Methods

2.1 Database and Sample Records

We analyzed about 1.24 million of all entry records of the "non-redundant" protein (nr-aa) database maintained by NCBI (National Center for Biotechnology Information)[20]. "Non-redundant" indicates that identical sequence entries are represented by one entry when they have identical lengths and identical residues at every position. We downloaded the "nr.Z" FASTA file from the NCBI FTP site, `ftp://ftp.ncbi.nih.gov/blast/db/` as of November 2002. The number of total records was 1,242,001 (Table 1). This file contains all non-redundant records of PDB (Protein Data Bank, Research Collaboratory for Structural Bioinformatics), Swiss-Prot (Swiss Institute of Bioinformatics and European Bioinformatic Institute), PIR (Protein Information Resource, National Biomedical Research Foundation, Georgetown University Medical Center), and conceptual translations of GenBank coding sequences. Biological existence of many *in-silico*-generated proteins deduced from DNA sequences has not directly confirmed *in vivo* or *in vitro*.

All data were then converted to XML (Extensible Markup Language) file with several tags[21]. Sample records with annotation of “mutant”, “Mutant”, “mutation”, “Mutation”, or “Engineering” in the “definition” section were deleted (Table 1). Although this exclusion of artificially created sequences may not be complete, the remaining artificial records, if any, would be insignificant in terms of the data analyses performed afterwards.

2.2 Definitions and Operations

We analyzed 8000 combinatorial sets of three amino acids (triplets). A three-amino-acid window that defines a triplet in a large linear sequence is conceptually slid one by one along the protein chain so that a given amino acid residue is an overlapping part of three different triplets unless it is located at the ends of the chain. Thus, the total number of existing triplets in all sample records (defined as S below) can be written as:

$$S = \sum_{j=1}^{N} (n_j - 2) = A - 2N \qquad (1)$$

where n_j is the number of amino acid residues in a given protein j, N is the number of protein records in the database, and A is the total number of amino acid residues in the database. Alternatively, based on triplet count for each triplet $a_k a_l a_m$ or α in the database, $T_{k \cdot l \cdot m}$ or T_α, the total number of existing triplets (S) can be expressed as follows, considering there are 8000 different triplets:

$$S = \sum_{\alpha=1}^{8000} T_\alpha \qquad (2)$$

Conversely, from the probabilistic expression of amino acid count for each amino acid (p, q, or r) in the database, P_p, P_q, or P_r , the expected triplet count, E_α, for each triplet $a_p a_q a_r$ or α is given as follows:

$$E_\alpha = S \cdot P_p P_q P_r \qquad (3)$$

Difference between theoretically-estimated triplet count E_α and the real triplet count T_α for each triplet in the database is expressed as follows:

$$D_T = \frac{T_\alpha - E_\alpha}{E_\alpha} = \frac{T_\alpha}{E_\alpha} - 1 \qquad (4)$$

Likewise, difference between theoretically-estimated triplet count E_α and randomly-generated triplet count R_α from the population with the identical amino acid composition is expressed as follows:

$$D_R = \frac{R_\alpha - E_\alpha}{E_\alpha} = \frac{R_\alpha}{E_\alpha} - 1 \qquad (5)$$

We call D_T and D_R the relative triplet-counts. The frequency distribution of D_R is supposed to show random fluctuations of the sampling procedure itself around a central value, resulting in the normal error curve. Distribution histograms for D_T and D_R were compared to each other.

2.3 Computer Programs

We developed a JAVA program to count

the number of each amino acid and each triplet in the database, and to subsequently execute several operations. The output data were exported to the Microsoft Excel 2000 and processed numerically and graphically.

To generate a theoretical random distribution from the population with the identical amino acid composition, we used the Mass.Random program in JAVA. Sampling procedure was exhaustively repeated as many times as the number of amino acid residues in the database, which was equivalent to having a random reconstitute of theoretical proteins from all the real database records that are conceptually broken into pieces of amino acid monomers. The randomness of the sampling procedure was confirmed by comparing the result of sampling repeated much less times to that of the exhaustive method (data not shown).

To demonstrate the operational accuracy using the JAVA program developed by ourselves and the one for the random sampling procedure, we employed the "ABC model" in which three imaginary amino acids represented by letters, A, B, and C, were treated using these programs with a given composition and an imaginary population of 100 million letters. In this case, only 27 triplets exist, making the system simpler and amenable to calculations by hand. The output data generated by the programs were compared to the hand-calculated ones. We found these outputs were virtually identical except for unavoidable fluctuations from the random sampling procedure itself, confirming the operational accuracy (data not shown).

3 Results

3.1 Frequency Distribution of Triplet Counts in the Database

From the nr-aa database that contains about 1.24 million sample records and about 394 million residues (Table 1), we obtained "amino acid count" for each chemical species of 20 amino acids (Fig. 1). As expected, each amino acid count varied, ranging from the lowest count, triptophan (W; 1.35%) to the highest count, leucine (L; 9.68%). We note that this count order of amino acids has some aspects of similarity to the order of molar concentration of amino acids in human plasma[22–24]with notable exceptions of glutamine (Q; the highest plasma concentration but relatively small amino acid count), cysteine (C; very small amino acid count but reasonably abundant in plasma), leucine (L; the most frequent amino acid count but not so abundant in plasma), and asparatic acid (D; very low concentration in plasma but not so small amino acid count). There seemed to be no other conspicuous tendency of physiological or physicochemical properties in this count order.

Similarly, we obtained "triplet count" for each of 8000 triplets as shown in Equations (1) and (2), from which we constructed a histogram for the frequency distribution of triplet counts. That is, since each triplet has one numerical value of triplet count, their distribution can be examined in a histogram. This histogram skewed extensively toward higher values (Fig. 2). For comparison, we also produced a histogram for the theoretical distribution of triplet counts randomly generated from the population of the identical amino acid composition. Although distribution range of the real triplet counts in the

Table 1　Numbers of records in the nr-aa database[a].

Species	Total records	Mutant excluded[b]	Total residues	Total triplets
all species	1,242,001	1,238,414	394,228,622	391,508,133
human	99,032	98,891	33,676,936	33,474,218
mouse	76,821	76,782	22,444,370	22,284,395
fruit fly	23,917	23,900	12,735,836	12,678,652
nematode	23,598	23,597	10,843,796	10,796,287
colon bacterium	8,781	8,760	2,507,399	2,489,619

[a]: As of November, 2002.

[b]: Records of artificial sequences were excluded, on which "Total residues" and "Total triplets" were based (see Methods).

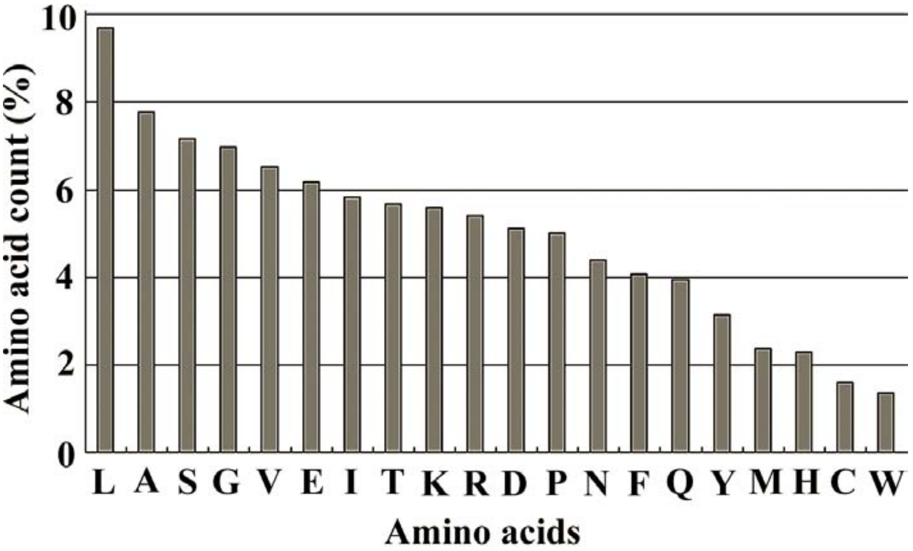

Fig. 1　Amino acid counts in the nr–aa database. X-axis shows 20 different amino acids expressed as single-letter codes. Y-axis shows amino acid counts in percentage. Amino acids are ranked from the highest to the lowest counts.

database was larger than that of the theoretical one, overall shapes of these two histograms were similar to each other. Likewise, ranked orders of triplets according to the triplet counts in these two populations were similar to each other, although not identical (Table 2).

To clarify their difference, these distribution data were operationally transformed based on Equations (3), (4), and (5), and the relative triplet-count D_T or D_R was used in X-axis (Fig. 3). By comparison, it is clear that the distribution of the real triplet-count (D_T) had much smaller and wider single peak and much larger distribution range than the random fluctuations of triplet count (D_R). Ranked orders of triplets according to the relative triplet-counts also exhibited a clear difference (Table 3). It is likely that this characteristic distribution cannot be explained by the random fluctuations of triplet formation. Rather, certain triplets exist much more in the database than their random expectations.

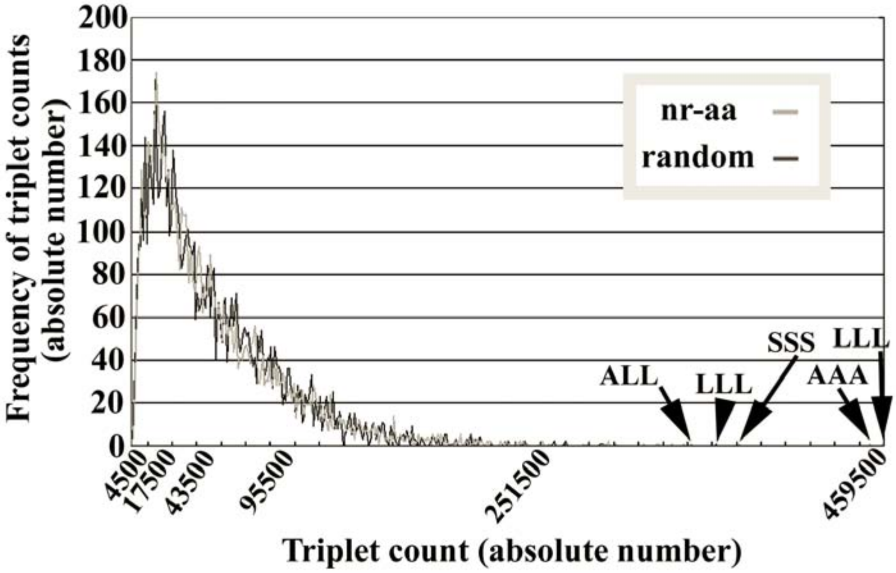

Fig. 2 Triplet-count distribution in the nr-aa database. X-axis shows triplet count in absolute number, and Y-axis shows its frequency in absolute number. In constructing this histogram, bar width was set at 1000, and each frequency point was connected by a line. Total number of triplets is 8000. Arrows and an arrowhead indicate most deviated samples in the nr-aa and random distributions, respectively.

Table 2 Ranked orders of triplets with the highest and lowest counts in absolute numbers.

nr-aa	(1st)LLL-AAA-SSS-ALL-LAA-LLA-ALA-AAL— —CHW-MWW-WMW-MCW-MWC-CMW-WWC-WCM(8000th)
random	(1st)LLL-LLA-LAL-ALL-SLL-LSL-LLS-LLG— —WHW-CCW-CWC-WCC-WWC-WCW-CWW-WWW(8000th)

Note: The highest rank was indicated as 1st, and the lowest rank as 8000th. The triplet counts of the nr-aa database and random sampling range from 459769(LLL) to 1222(WCM) and from 357081(LLL) to 850(WWW), respectively. The presented result of the random procedure is merely one example because the precise order has random fluctuations.

Among the 8000 triplets here examined, some triplets such as WLT ($D_T = D_R = 0.001$) and EIQ ($D_T = D_R = 0.002$) had the same values of D_T and D_R, whereas some other triplets such as HCN ($D_T = 3.015$; $D_R = -0.024$) and YYC ($D_T = 4.796$; $D_R = -0.001$) had totally different values. The former can be considered to have theoretically reasonable counts in the nr-aa database, whereas the latter to have theoretically-deviated counts.

3.2 Species-Independent Distribution of Triplet Counts

The more the database records increase, the less the frequency distribution of triplet counts is biased by database characters, because the sample population becomes closer to the parent population, a collection of all protein species on the earth. Since there are already 1.24 million records from more than 130 thousand biological species in the nr-aa database, it is unlikely that the character-

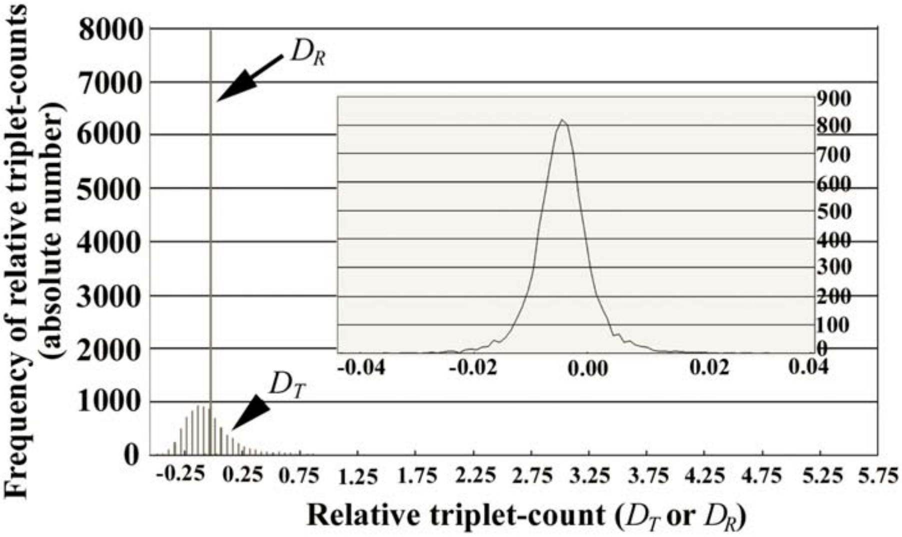

Fig. 3 Triplet-count distribution after operational transformation in the nr-aa database. X-axis shows relative triplet-count, and Y-axis shows its frequency in absolute number. Theoretical triplet-count distribution generated randomly (D_R) is expressed as almost a single bar in this histogram as indicated by an arrow. In contrast, real triplet-count distribution (D_T) is markedly different as indicated by an arrowhead. In constructing this histogram, bar width was set at 0.05. Inset shows the theoretical distribution (D_R) with much smaller bar width, indicating its normality.

Table 3 Ranked orders of triplets with the highest and lowest relative counts.

nr-aa	(1^{st})QQQ-HHH-YYC-NNN-CCC-HCN-WWN-PPP— —PCM-WPK-WPM-KCA-EWP-IMW-KWP-EPN(8000^{th})
random	(1^{st})QWW-WYW-WWE-YQW-HRW-WCG-HWM-MYW— —CCQ-CTW-ECH-CCH-FCC-WCF-CHC-CWW(8000^{th})

Note: The highest rank was indicated as 1^{st}, and the lowest rank as 8000^{th}. The relative triplet-counts of the nr-aa database and random sampling range from 5.743 (QQQ) to −0.500 (EPN) and from 0.060 (QWW) to −0.052 (CWW), respectively. The presented result of the random procedure is merely one example because the precise order has random fluctuations.

istic triplet-count distribution shown above is simply because of a database bias itself. However, the possibility still exists that the characteristic distribution might have resulted from over-representation or under-representation of particular proteins in the database. Accordingly, we further performed similar triplet analysis in five phylogenetically distinct biological species, human (*Homo sapiens*), mouse (*Mus musculus*), fruit fly (*Drosophila melanogaster*), soil nematode (*Caenorhabditis elegans*), and a colon bacterium (*Escherichia coli*), whose genome sequences have already been known.

We first obtained amino acid counts in each species (Fig. 4). Although they were slightly different from one another, their overall trend of the ranked counts seemed to

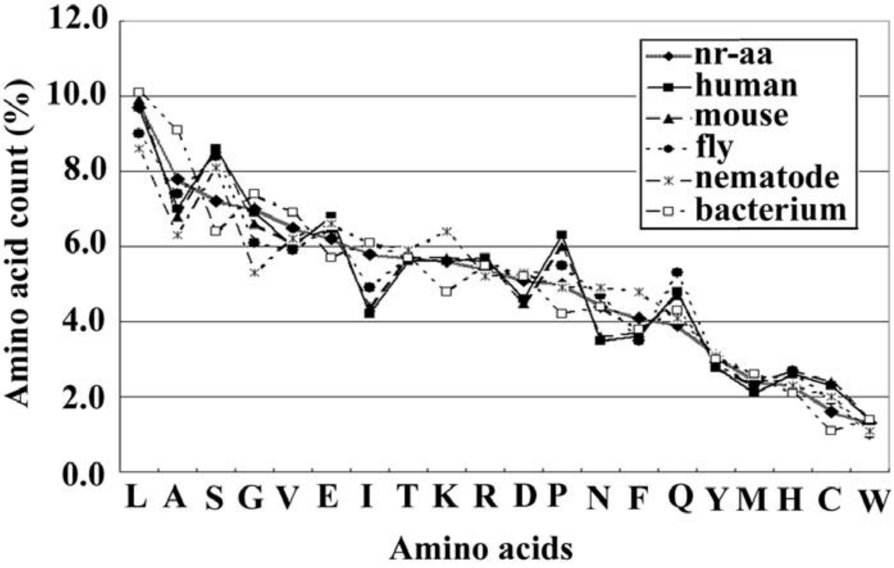

Fig. 4 Amino acid counts in five species. X-axis shows 20 different amino acids expressed as single-letter codes. Y-axis shows amino acid counts in percentage. Amino acids are ranked from the highest to the lowest counts of the nr-aa database.

be almost invariable throughout species. According to Equations (3) and (4), we further obtained histograms for the relative triplet-count (D_T) distributions in each species. They were all essentially similar to that of the whole nr-aa database, compared to the randomly generated one (Fig. 5). This result showed that the characteristic distribution of the whole nr-aa database was unlikely to be biased by the database records themselves.

4 Discussion

Here we observed the characteristic triplet-count distribution with a non-random and species-independent nature in the nr-aa database. Together with the large number of sample records from various species in the database, it is unlikely that the characteristic distribution is a product of a database artifact.

Some of these fundamental database characters have still remained obscure in spite of intensive research efforts in bioinformatics, whose main interest is to develop suitable algorithms for similarity searches and structural predictions. To be sure, these algorithms are extremely valuable in molecular biology[1−14]. Yet, a different approach such as the one performed here may shed light on a new biological aspect that can intellectually be derived from database searches. With some practical limitations in terms of main memory capacity we could examine not only three but also four, five, and more residues of amino acid sets in this procedure, in which case the procedure becomes increasingly similar to similarity searches.

Amino acid compositions of a given protein or entire proteins in a given species has been used to identify proteins from non-model organisms[25], to classify thermophilic species[26], to infer environmental status of a species[27][28], and to recapitulate prebiotic

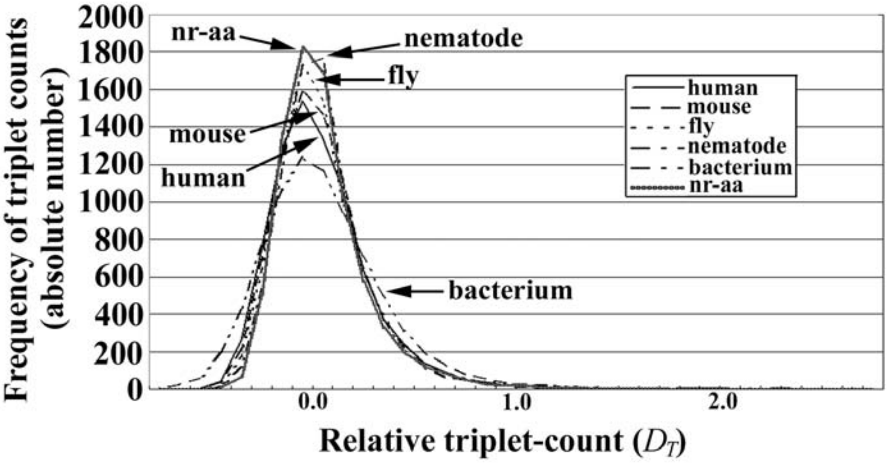

Fig. 5 Triplet-count distributions after operational transformation in the nr-aa database. X-axis shows relative triplet-count, and Y-axis shows its frequency in absolute number. Data were treated as in Figure 3, except that they were separately examined according to species which original records belong to. In constructing this histogram, bar width was set at 0.05, and each frequency point was connected by a line. Both right and left sides of this graph were truncated.

molecular evolution[29]. Although we also observed some species differences in amino acid compositions in five species, here we paid more attention to their overall similarities. Their similarities were more obvious in the triplet-count distributions after the operational transformation, compared to the randomly generated one.

Although biological significance of the characteristic triplet-count distribution is obscure at this point, it is reasonable to consider the fact that structural and functional similarities among proteins are not always observed in a long stretch of amino acids but sometimes detected in a few amino acids. A notable example can be drawn from the G-protein-coupled receptor (GPCR) superfamily, in which little sequence similarity can be found unless two receptors are very closely related[30][31]. This makes the conventional similarity search much less useful than one might expect. To alleviate this problem, we have previously performed length analyses, a collection of statistical analyses of particular lengths of GPCRs, which clearly showed that biological information can computationally be extracted from lengths of terminal and loop regions of GPCRs[32–34].

The triplet analysis performed here can be placed in this line of research. In GPCRs, there are several sites with just a single or a few residues that are relatively conserved among a group of GPCRs, one of which is called “DRY sequence” (triplet of aspartic acid, arginine, and tyrosine) located at the boundary between the third transmembrane domain and the second intracellular loop[30][31]. Regardless of the fact that this DRY sequence of GPCRs cannot readily be identified by the conventional similarity searches, it is highly important in the functional integrity of GPCRs[35][36], together with other key residues[37][38]

Functional importance of triplets or just small sets of residues is also exemplified in other proteins. A potassium channel, another transmembrane molecule, also uses triplets to interact with other molecules[39]. Triplet repeats in collagen, a major structural molecule in biological systems, stabilizes the triple helix structure of collagen fibers[40][41], and a similar triplet in collagen constitutes a protein-protein interaction site[42]. Hypervariable regions of immunoglobulins are composed of clusters of relatively short loops from 4 to 15 residues[43–45].

Although some physicochemical interactions between amino acids and its associated free energy cost may be a cause of the skewed triplet-count distribution shown in this study, more likely but not exclusive explanation would be that its origin could be traced to biological evolutionary history. The biological fixation of the triplet composition in proteins may be either for functional reasons or simply for accidental reasons during the course of biological evolution.

The number of triplets and their corresponding proteins is so large that it is not trivial to examine them thoroughly. Here we only point out some interesting examples: the relative triplet-counts of WGQ and YEC in the human records (3.682 and 1.868, respectively) and those in the bacterium ones (−0.506 and −0.502, respectively) were highly deviated in the positive and negative directions, depending on species, from the theoretically expected values that ranged between 0.060 and −0.052. Conversely, the relative triplet-counts of KPW and YWH were deviated in the negative direction in the human records (−0.257 and −0.303, respectively) but in the positive direction in the bacterium ones (1.174 and 1.285, respectively). More systematic and quantitative analyses will be necessary to clarify differences between species.

It would be valuable to examine relationship between triplets and secondary structures, especially for triplets with extremely high or low counts. As some residues are preferred in a given secondary structure[8–10], some triplets may be favored in a given secondary structure. Non-randomness of the triplet-count distribution may be a reflection of the number of these secondary structures in proteins at the population level. Another possibility is that proteins containing a particular triplet may preferably belong to a particular protein family, and the family composition in the database may be a cause of the non-random nature of triplet count. It is reasonable that protein records for a given triplet are to be examined with respect to structural and functional protein classifications using specialized databases such as PDB (Protein Data Bank)[46] and SCOP (structural Classification of Proteins)[47].

Acknowledgements

We are grateful to Ms. Misato Wada for performing several analyses on this project. We also thank other laboratory colleagues in Kanagawa University.

References

[1] Jonassen, I.: "Methods for discovering conserved patterns in protein sequences and structures", In: "Bioinfromatics: Sequence, Structure, and Databanks. A Practical Approach", Higgins, D., Taylor, W. eds., Oxford University Press, pp.143–166, 2000.

[2] Yona, G., Brenner, S.E.: "Comparison of protein sequences and prac-

tical database searching", In: "Bioinfromatics: Sequence, Structure, and Databanks. A Practical Approach", Higgins, D., Taylor, W. eds., Oxford University Press, pp.167–190, 2000.

[3] Baldi, P., Brunak, S.: "Bioinformatics: The Machine Learning Approach", Second Edition, The MIT Press, 452p., 2001.

[4] Mount, D.W.: "Bioinformatics: Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press, 564p, 2001.

[5] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: "Basic local alignment search tool", Journal of Molecular Biology, Vol.215, pp.403–410, 1990.

[6] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Research, Vol.25, pp.3389–3402, 1997.

[7] Schuler, G.D.: "Sequence alignment and database searching", In: "Bioinformatics: A practical Guide to the Analysis of Genes and Proteins", Baxevanis, A.D., Ouellette, B.F.F. eds., Wiley-Liss, New York, pp.187–214, 2001.

[8] Chou, P.Y., Fasman, G.D.: "Prediction of protein conformation", Biochemistry, Vol.13, p.222–245, 1974.

[9] Lim, V.I.: "Algorithms for prediction of α-helical and β-structural regions in globular proteins", Journal of Molecular Biology, Vol.88, pp.873–894, 1974.

[10] Garnier, J., Osguthorpe, D.J., Robson, B.: "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins", Journal of Molecular Biology, Vol.120, pp.97–120, 1978.

[11] Heringa, J: "Predicting secondary structure from protein sequences", In: "Bioinfromatics: Sequence, Structure, and Databanks. A Practical Approach", Higgins, D., Taylor, W. eds., Oxford University Press, pp.113–142, 2000.

[12] Banerjee-Basu, S.: "Predictive methods using protein sequences", In: "Bioinformatics: A practical Guide to the Analysis of Genes and Proteins", Baxevanis, A.D., Ouellette, B.F.F. eds., Wiley-Liss, New York, pp.253–282, 2001.

[13] Rost, B., Sander, C.: "Prediction of protein secondary structure at better than 70% accuracy", Journal of Molecular Biology, Vol.232, pp.584–599, 1993.

[14] Jones, D. T., Taylor, W. R., Thornton, J. M.: "A model recognition approach to the predication of all-helical membrane protein structure and topology", Biochemistry, Vol.33, pp.3038–3049, 1994.

[15] Tyers, M., Mann, M.: "From genomics to proteomics", Nature, Vol.422, pp.193–197, 2003.

[16] Boguski, M.S., Mclntosh, M.W.: "Biomedical informatics for proteomics", Nature, Vol.422, pp.233–237, 2003.

[17] Ramachandran, G.N., Sassiekharan, V.: "Conformation of polypeptides and proteins", Advances in Protein Chemistry, Vol.28, pp.283–437, 1968.

[18] Chothia, C.: "Principles that determine the structure of proteins", Annual Reviews in Biochemistry, Vol.53, pp.537–572, 1984.

[19] Chothia, C., Levitt, M., Richardson, D.: "Helix-to-helix packing in proteins", Journal of Molecular Biology, Vol.145,

pp.215–250, 1981.

[20] Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L., Rapp, B.A.: "Database resources of the National Center for Biotechnology Information: 2002 update", Nucleic Acids Research, Vol.30, pp.13–16, 2002.

[21] Inomata, Y., Osanai, T., Gotoh, T., Yamamoto, H.: "Experimental construction of amino acid sequence databases", Proceeding of the 10th Annual Conference for Japan Society of Information and Knowledge, pp.47–51, 2002.

[22] Marlis, E.B., Aoki, T.T., Pozefsky, T., Most, A.S., Cahill, G.F. Jr.: "Muscle and splanchnic glutamine and glutamate metabolism in postabsorptive andstarved man", Journal of Clinical Investigation, Vol.50, pp.814–817, 1971.

[23] Sherwin, R.S., Hendler, R.G., Felig, P.: "Effect of ketone infusions on amino acid and nitrogen metabolism in man", Journal of Clinical Investigation, Vol.55, pp.1382–1390, 1975.

[24] Simmons, P.S., Miles, J.M., Gerich, J.E.: "Increased proteolysis. An effect of increases in plasma cortisol within the physiologic range", Journal of Clinical Investigation, Vol.73, pp.412–420, 1984.

[25] Wilkins, M.R., Williams, K.L.: "Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation", Journal of Theoretical Biology, Vol.186, pp.7–15, 1997.

[26] Krel, D.P., Ouzounis, C.A.: "Identification of thermophilic species by the amino acid compositions deduced from their genomes", Nucleic Acids Research, Vol.29, pp.1608–1615, 2001.

[27] Dumontier, M., Michalickova, K., Hogue, C.W.V.: "Species-specific protein sequence and fold optimizations", BMC Bioinformatics, Vol.3. p.39, 2002, http://www.biomedcentral.com/1471-2105/3/39.

[28] Tekaia, F., Yeramian, E., Dujon, B.: "Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis", Gene, Vol.297, pp.51–60, 2002.

[29] Brooks, D.J., Fresco, J.R., Lesk, A.M., Singh, M.: "Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code", Molecular Biology of Evolution, Vol.19, pp.1645-1655, 2002.

[30] Schwartz, T.W.: "Molecular structure of G-protein-coupled receptors", In: "Textbook of Receptor Pharmacology", Foreman, J.C., Johansen, T. eds., CRC Press, Boca Raton, pp. 65–84, 1996.

[31] Wess, J.: "Molecular basis of receptor/ G-protein-coupling selectivity", Pharmacology & Therapeutics, Vol.80, pp.231–246, 1998.

[32] Otaki, J.M., Firestein, S.: "Length analyses of G-protein-coupled receptors", Journal of Theoretical Biology, Vol.211, pp.77–100, 2001.

[33] Otaki, J.M., Yamamoto, H., Firestein, S.: "Structural feature of odorant receptors inferred from the length analyses of G-protein-coupled receptors", Japanese Journal of Taste and Smell Research, Vol.9, pp.357–360, 2002.

[34] Otaki, J.M., Yamamoto, H.: "Length analyses of *Drosophila* odorant receptors", Journal of Theoretical Biology,

Vol.223, pp.27–37, 2003.

[35] Wilbanks, A.M., Laporte, S.A., Bohn, L.M., Barak, L.S., Caron, M.G.: "Apparent loss-of-function mutant GPCRs revealed as constitutively desensitized receptors", Biochemistry, Vol.41, pp.11981–11989, 2002.

[36] Ohyama, K., Yamamo, Y., Sano, T., Nakagomi, Y., Wada, M., Inagami, T.: "Role of the conserved DRY motif on G protein activation of rat angiotensin II receptor type 1A", Biochemical and Biophysical Research Communications, Vol.292, pp.362–367.

[37] Huttenrauch, F., Nitzki, A., Lin, F.T., Honing, S., Oppermann, M.: "β-arrestin binding to CC chemokine receptor 5 requires multiple C-terminal receptor phosphorylation sites and involves a conserved Asp-Arg-Tyr sequence motif", Journal of Biological Chemistry, Vol.277, 30769–30777, 2002.

[38] Auger, G.A., Pease, J.E., Shen, X., Xanthou, G., Barker, M.D.: "Alanine scanning mutagenesis of CCR3 reveals that the three intracellular loops are essential for functional receptor expression", European Journal of Immunology, Vol.32, pp.1052–1058, 2002.

[39] Dong, K., Xu, J., Vanoye, C.G., Welch, R., MacGregor, G.G., Giebisch, G., Hebert, S.C.: "An amino acid triplet in the NH_2 terminus of rat ROMK1 determines interaction with SUR2B", Journal of Biological Chemistry, Vol.276, pp.44347–44353, 2001.

[40] Pace, J.M., Atkinson, M., Willing, M.C., Wallis, G., Byers, P.H.: "Deletions and duplications of Gly-Xaa-Yaa triplet repeats in the triple helical domains of type I collagen chains disrupt helix formation and result in several types of osteogeneis imperfecta", Human Mutation, Vol.18, pp.319–326, 2001.

[41] Persikov, A.V., Ramshaw, J.A., Kirkpatrick, A., Brodsky, B.: "Peptide investigations of pairwise interactions in the collagen triple-helix", Journal of Molecular Biology, Vol.316, pp.385–394, 2002.

[42] Koide, T., Takahara, Y., Asada, S., Nagata, K.: "Xaa-Arg-Gly triplets in the collagen triple helix are dominant binding sites for the molecular chaperone HSP47", Journal of Biological Chemistry, Vol.277, pp.6178–6182.

[43] Alzari PN, Lascombe M-B, Poljak RJ. Three-dimensional structure of antibodies. Annu. Rev. Immunol. 6, 555–580, 1988.

[44] Davies, D.R., Padlan, E.A.: "Antibody-antigen complexes", Annual Reviews in Biochemistry, Vol.59, pp.439–473, 1990.

[45] Mariuzza, R.A., Phillips, S.E.V., Poljak, R.J.: "The structural basis of antigen-antibody recognition", Annual Reviews in Biophysics and Biophysical Chemistry, Vol.16, pp.139–159, 1987.

[46] 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 Research, Vol.28, pp. 235–242, 2000.

[47] Murzin, A.G., BrennerS.E., Hubbard, T., Chothia, T.: "SCOP: a structural classification of proteins database for the investigation of sequences and structures", Journal of Molecular Biology, Vol.247, pp.536–540, 1995.

(Received April 4, 2003)
(Accepted June 16, 2003)