Eur Biophys J (2018) 47:31–38
https://doi.org/10.1007/s00249-017-1226-6

ORIGINAL ARTICLE

EBSA
Biophysics in Europe

Distribution of dipeptides in different protein structural classes: an effort to find new similarities

Mahin Ghadimi[1] · Emran Heshmati[1] · Khosrow Khalifeh[1]

Received: 14 February 2017 / Revised: 15 April 2017 / Accepted: 29 May 2017 / Published online: 13 June 2017

Abstract Finding any regularity in the sequences of proteins and determining their correlation with structural features are of great interest for an understanding of molecular biology. We statistically analyzed the relative frequencies of all 400 possible dipeptides in a data set containing randomly selected proteins of different defined structural classes including all-alpha, all-beta, alpha + beta and alpha/beta families. We found that the distribution of dipeptides is not the same for different structural classes, and some of them are significantly far from a random distribution. A tendency of a given amino acid to localize in the first or second position of a dipeptide depending on the structural class of protein was also found. Interestingly, some amino acids may be substituted for each other in the first or second positions of specific dipeptides in each structural class. This finding apparently contrasts with the routine expectation from the viewpoint of amino acid properties, as classically understood.

Keywords Regularity · Protein · Dipeptide · Structural class · Random distribution

Electronic supplementary material The online version of this article (doi:10.1007/s00249-017-1226-6) contains supplementary material, which is available to authorized users.

✉ Emran Heshmati
heshmati@znu.ac.ir

✉ Khosrow Khalifeh
khalifeh@znu.ac.ir

[1] Department of Biology, Faculty of Science, University of Zanjan, University Blvd, Zanjan, Islamic Republic of Iran

Introduction

According to Anfinsen's theory, the primary sequence of a protein as an ordered string of amino acids contains all the information required for it to gain its final functional three-dimensional structure (Anfinsen 1973). In the 45 years since this argument was made, important questions have been raised concerning the origin and identity of protein fold information (Dill and MacCallum 2012). These questions resulted in significant efforts toward elucidation of information hidden in protein sequences. Statistical analysis of databases containing protein sequences indicates that the 20 naturally occurring amino acids do not occur with equal frequency (Rani et al. 1995), while in other studies the relative frequency of each amino acid in a group of similar proteins has been determined (Schwartz et al. 2001). However, a single residue in a sequence has limited information, and the context of any residue can play a crucial role in its structural and/or functional properties, e.g., because of its neighbors (Fu et al. 2014). Hence, finding any regularity in protein sequences including dipeptides is of great importance, but in spite of much argument this need remains unaddressed (Hermans 2011). The frequency of motifs in proteins was first investigated in the context of protein primary structure sourced from whole-protein sequence databases (Unger and Sussman 1993; Aitken 1999), while Vondervisztet et al. analyzed the frequency of dipeptides in the sequences of known proteins (Vondervisztet et al. 1986). However, the total data set in protein databases we limited at that time and their input data contained the primary sequence of protein with no reference to secondary structures. It has been reported that some dipeptides may play a critical role in intracellular protein stability (Guruprasad et al. 1990), and Reddy et al.

have analyzed some representative dipeptides and found that stabilizing and destabilizing dipeptides have different patterns of interactions (Reddy 1996). Furthermore, analyzing three-dimensional structure databases has revealed that Cys residue oxidation is affected by neighboring residues (Fiser et al. 1992). By encoding dipeptide features and selecting a subset of dipeptide compositions, Nakariyakul et al. developed an interaction predictor tool and reported that selected dipeptide features have important roles in the specificity of protein domain interactions (Nakariyakul and Liu 2011). In other studies, it was found by statistical database analysis of the four major structural classes of protein including all-alpha, all-beta, alpha/beta and alpha + beta proteins that the propensities of each amino acid for the secondary structure are related to the structural class of the protein overall (Costantini et al. 2006; Ismail and Chowdhury 2010), but this analysis concerned the propensities of single amino acids rather than dipeptides. The outputs of such studies on the single amino acids led to several important assumptions in protein science that formed the basis for applications such as substitution matrices (Henikoff and Henikoff 1992) and structural prediction algorithms (Lim 1974). However, in the majority of these applications, the direct effect of neighbor residues was ignored. For instance, in the construction of substitution matrices based on multiple sequence alignment of protein superfamilies, the identity of only a single amino acid in an alignment file is considered despite the fact that it seems the conservation of a single residue may be affected by adjacent amino acids (Anishetty et al. 2002; Betancourt and Skolnick 2004).

For the reasons discussed, it is important to identify regular patterns of di- and/or tri-peptides (motifs), which are specific for a group of protein families and may have similar structural and functional consequences to each other. To shift the concept of the neighbor effect from sequence-based information to include a three-dimensional structural element, we first investigated the frequency of different mono- and dipeptides in defined structural classes of proteins including all-alpha, all-beta, alpha + beta and alpha/beta proteins. We found that the frequency of dipeptides is not the same in different structural classes. Additionally, we found that in structurally similar proteins some dipeptides are not randomly distributed, and the first or second position of these motifs is occupied by specific amino acids. We conclude that the microenvironment of an amino acid can be considered as an evolutionary driving force in dictating the structural properties of a protein, which leads to directional selection of amino acids for structural and functional purposes.

Materials and methods

Data

All structures were selected from the Protein Data Bank (Berman 2000) under the advanced search menu. The structure of all selected proteins was resolved by X-ray crystallography with a resolution better than 3.0 Å. All structures with ligands and more than 30% identity have been omitted. Structural classes were filtered in the search menu using both the ScopTree and CathTree options. Based on these criteria, we found that there were 499, 587, 626 and 670 structures for all-alpha, all-beta, alpha + beta and alpha/beta protein classes, respectively, at the end of 2015. Among them, 125 structures were sampled randomly for each structural class. Note that any structure that has unusual, unknown or missing amino acids was discarded. Thus, our data set consists of 400 protein structures containing 152,474 residues. All structures were converted from PDB to DSSP file format using the Linux-based mkDSSP program (Kabsch and Sander 1983; Joosten et al. 2011). They were then analyzed by PARS software (Fathinavid et al. http://www.znu.ac.ir/members/newpage/702) to calculate the frequency of any of the 20 residues (or monopeptides) and 400 dipeptides. The output was further analyzed by MS-Excel software. All analysis was performed separately for each structural class as well as for the total data set.

Normalized frequency distribution

Based on the results of the PARS software, the total number of monopeptides and dipeptides for any structural class as well as for the total data set was calculated. As proposed by Vondervisztet al., the normalized frequency distribution for any dipeptide (S_{ij}) formed by the ith and jth monopeptides in the first and second positions, respectively, was calculated by the following equation (Vondervisztet al. 1986):

$$S_{ij} = \frac{O_{ij}}{E_{ij}} \quad (1)$$

where O_{ij} and E_{ij} are the observed and expected values of occurrence of dipeptide ij in the data set, respectively. The values of E_{ij} for each dipeptide in each respective data set (total or any structural class) were calculated by Eq. 2:

$$E_{ij} = P_i \times P_j \times N \quad (2)$$

Here, P_i and P_j are the relative frequencies of individual amino acids in the first and second positions of a given dipeptide, and N is the total number of dipeptides in the corresponding data set. The values of P_i and P_j are provided in different columns of Table 1.

Table 1 Relative abundance (%) of monopeptides in the total data set and different structural classes

Monopeptide	All-alpha	All-beta	Alpha + beta	Alpha/beta	Total
A	9.62	6.98	8.28	9.07	8.51
C	1.28	1.25	1.21	1.23	1.24
D	5.49	6.00	5.92	5.80	5.81
E	8.15	6.41	7.27	7.03	7.20
F	4.06	4.14	3.70	3.62	3.86
G	5.26	7.77	7.51	7.48	7.05
H	2.35	2.03	2.46	2.09	2.23
I	5.48	6.04	5.79	6.53	5.99
K	6.67	5.56	5.90	5.90	5.99
L	11.08	8.27	8.71	10.35	9.62
M	1.95	1.54	1.75	1.41	1.65
N	3.80	4.74	4.25	3.79	4.13
P	3.84	4.91	4.47	4.49	4.43
Q	4.85	3.63	3.55	3.67	3.90
R	5.69	5.14	5.70	5.02	5.37
S	5.56	6.37	5.56	5.70	5.79
T	4.65	6.72	5.52	5.19	5.51
V	6.13	7.49	7.50	7.76	7.26
W	1.03	1.46	1.34	1.11	1.23
Y	3.04	3.56	3.63	2.74	3.22

Based on these criteria, an $S_{ij} = 1.0$ means a completely random association of an *ij* pair in the primary sequence, while values 1.5× greater than unity (1.5) and 1.5× less than unity (0.67) are regarded as non-random distributions indicating preferential association and avoidance in the primary structure, respectively. So, 1.5 times greater and less than 1.0 have been written in red and blue font, respectively.

Cross-correlation coefficient

Cross-correlation coefficients were extracted from the S_{ij} matrices by determining the correlation row-wise and column-wise for the *i*th and *j*th positions, respectively.

The cross-correlation coefficient can be used to reveal similarities among preferred sequential environments of various amino acids and also to determine the tendency of each residue to localize in the first or second position of an *ij* pair. In our correlation coefficient analysis and for prevention of any statistical fluctuations, *P* values less than 0.01 were considered statistically significant. Values of ±0.1 would correspond to perfect correlation between two number series, while 0.0 means no correlation.

Results and discussion

The sampling and choice of data set are very important because proteins that belong to the same family may have similar evolutionarily conservative dipeptides which could bias our analysis. Hence, stringent search criteria were used so that the selected proteins have only a maximum value of 30% sequence identity. It is also notable that several proteins may contain highly homologous domains or repetitive sequences, leading to the problem of redundancy. We minimized this effect by using the largest available data set. It should be noted that the observed and expected frequencies (O_{ij} and E_{ij}) are not listed, and only the S_{ij} and correlation coefficient values are addressed directly here.

The relative frequencies of all 20 monopeptides in the total data set and for all structural classes are provided in Table 1. These data show high correlation (CC = 0.97) with the result of the study by Xia and Xie in which more than 7343 protein sequences were analyzed (Xia and Xie 2002). The data in Table 1 indicate that the frequency distribution of all amino acids for different structural classes of proteins is not the same. Furthermore, our calculated and expected frequencies of dipeptides show good correlation (CC = 0.94 and 0.96) with the report provided by Shen et al. (2006).

In the next step of analysis, the normalized frequency distribution of dipeptides (S_{ij}) for the total data set as well as all-alpha, all-beta, alpha + beta and alpha/beta structural classes were calculated (Fig. 1). Quantitative data are provided as Tables 2–6 in the supplementary material. The values of S_{ij} are in the range of 0.13 (indicating avoidance)–3.97 (indicating favorable association). Since S_{ij} is equal to the ratio of observed to expected values of dipeptides, an $S_{ij} \geq 1.5$ indicates the tendency of a given dipeptide to occur more than 1.49× relative to expected values and is considered a boundary for a high tendency for association. Similarly, the values of $S_{ij} \leq 0.67$ are considered as a measure of the avoidance. These critical values are shown in the red- and blue-colored spectrum in Fig. 1 and corresponding tables in the supplementary data.

As shown in Fig. 1a, in the all-alpha structural class, there are 19 dipeptides with extremely high S_{ij} including Cys-Cys, Met-Cys, Trp-Cys, His-Phe, Cys-His, His-His, Arg-His, His-Met, Ser-Met, Met-Asn, His-Pro, Pro-Pro, Trp-Pro, Tyr-Pro, Cys-Arg, Ala-Trp, Ser-Trp, Trp-Trp and Asp-Tyr, while 29 dipeptides including Cys-Ala, His-Cys, Asn-Cys, His-Asp, Met-Gly, His-His, Asn-His, Thr-His, Trp-His, Cys-Ile, Trp-Ile, HisLys, Pro-Lys, Cys-Leu, Cys-Met, Gly-Met, Pro-Met, Ala-Asn, Ala-Pro, Glu-Pro, His-Arg, Arg-Ser, Cys-Trp, Ile-Trp, Leu-Trp, Met-Trp, Tyr-Trp, Ala-Tyr and Trp-Tyr have extremely low S_{ij} values. These

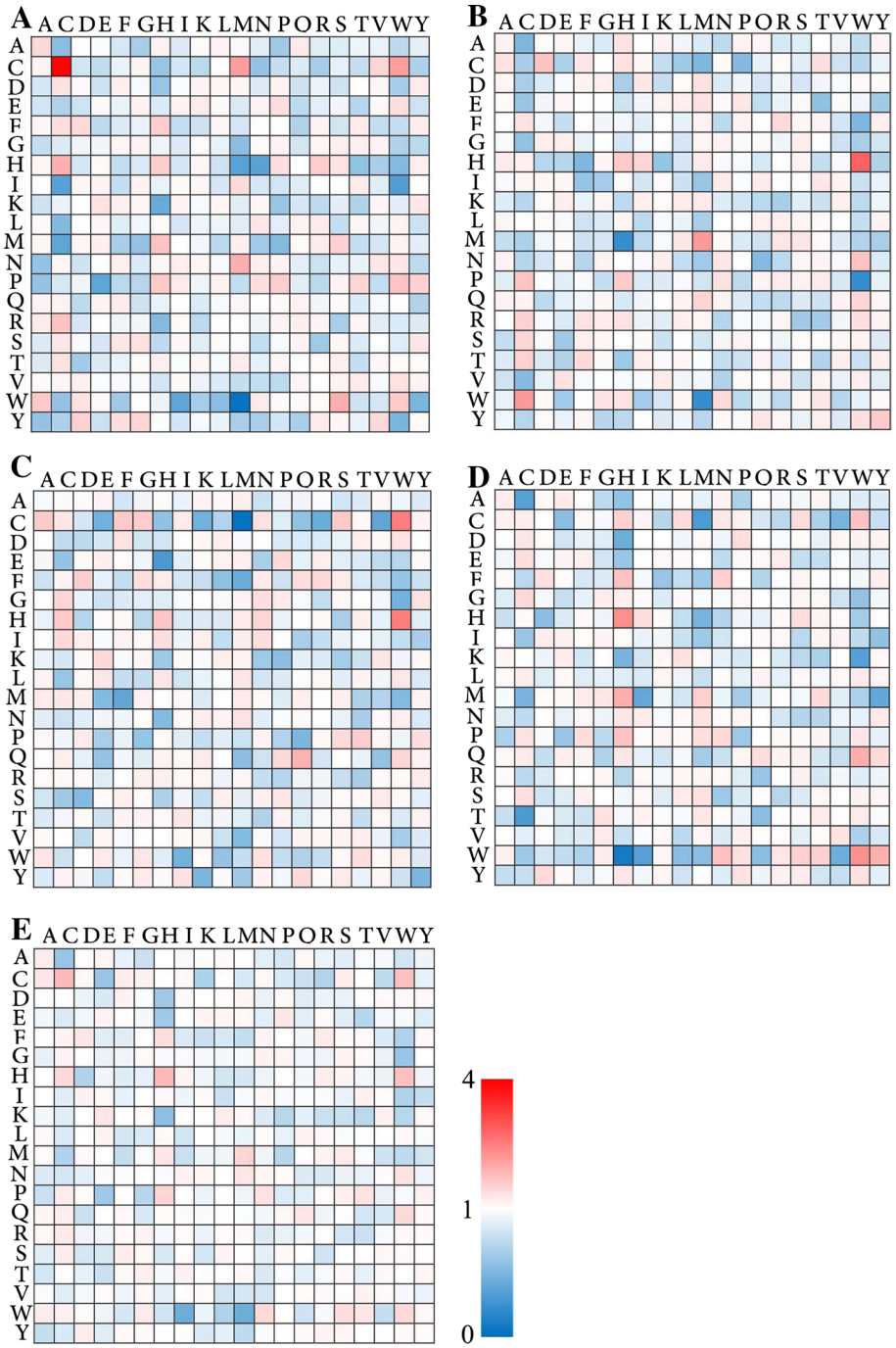

Fig. 1 Graphical representation of the normalized frequency distribution matrix, S_{ij}, of the dipeptide fragment for all-alpha (**a**), all-beta (**b**), alpha + beta (**c**), alpha/beta (**d**) class and our total data set (**e**). Each panel contains the first position of a dipeptide (*i*-position) in the *horizontal line*, while that of the second position (*j*-position) is shown in the *vertical line*. For better clarification in finding the differences between specific cells, the numerical values are also provided in the supplementary material 1

data together demonstrate that Cys, His and Trp are the most selective amino acids in their sequential association; a number of 12 Cys-containing, 14 His-containing and 13 Trp-containing dipeptides are characterized by extremely high or low S_{ij} values. In contrast, Gln and Val appear to be virtually neutral showing a nearly random association with other amino acids in the all-alpha class. Other amino acids have a moderate tendency to be selective. Additionally, there are four Ala-containing dipeptides with extremely low S_{ij} values and only one with an extremely high S_{ij}. This means that the selectivity of Ala is toward association rather than avoidance. It was also found that Ala is more selective when it localizes in the first position of an *ij*-pair.

S_{ij} values for the all-beta structural class are provided in Fig. 1b showing 11 dipeptides have extremely high S_{ij} values (Asp-Cys, His-His, Trp-His, Met-Met, Trp-Asn, Cys-Pro, Cys-Arg, His-Pro, Cys-Thr, Cys-Trp, Tyr-Tyr), while 16 dipeptides have extremely low S_{ij} values (Cys-Ala, Met-Cys, Pro-Cys, CysE, ThrE, Trp-Phe, Cys-Gly, Phe-His, Lys-His, Phe-Ile, Met-Ile, HisMet, Gln-Asn, Trp-Pro,

Cys-Val, Met-Trp). These data demonstrate that Cys is relatively selective in its sequential association; 11 Cys-containing dipeptides are characterized by extremely high or low S_{ij} values. Leu and Ser appear to be virtually neutral, while others have a moderate tendency to be selective. As for Ala in the all-alpha structures, Cys is more selective when located in the *i*th position of a dipeptide in the all-beta structural class.

Analyzing the data for the alpha + beta structural class (Fig. 1c) shows that there are 11 dipeptides with extremely high S_{ij} values (Ala-Cys, Phe-Cys, Gly-Cys, Ser-Cys, Trp-Cys, Asp-Phe, Cys-His, His-His, Trp-His, ThrP and Gln-Gln) and 31 dipeptides with extremely low S_{ij} values (Cys-Glu, Cys-Leu, Asp-Ser, Glu-Cys, Glu-Met, Glu-Gln, PheMet,Gly-Pro, His-Cys, His-Glu, His-Lys, His-Asn, Ile-Trp, Lys-Cys, Lys-Trp, Leu-Phe, Leu-Trp, Met-Cys, Met-Phe, Met-Gln, Met-Val, Pro-Lys, Gln-Cys, Gln-Pro, Arg-Cys, Val-Cys, Val-Gln, Trp-Phe, Trp-Gly, Trp-Met and Tyr-Trp). In this structural class, Cys is the most highly selective residue; in total, nine and six Cys-containing dipeptides have extremely low or high S_{ij}, respectively. We also found that five, four and seven dipeptides contain Glu, Lys and Met, respectively, with extremely low S_{ij} values. However, these residues have no significant values of S_{ij} for association. So, the selectivity of these amino acids is toward avoidance for pairing with other amino acids in the alpha + beta structural class.

In Fig. 1d, the S_{ij} values for the alpha/beta structural class are shown. According to these data, there are 14 dipeptides with high S_{ij} values (His-Cys, Trp-Cys, His-Phe, Asn-Phe, His-His, His-Met, Met-Met, His-Pro, Trp-Gln, Asn-Trp, Ser-Trp, Thr-Trp, Trp-Trp and Tyr-Trp) and 30 dipeptides with low S_{ij} values (His-Trp, Ile-Trp, Leu-Trp, Met-Trp, Gln-Trp, Val-Trp, Cys-Ala, His-Ala, Glu-Cys, Met-Cys, Val-Cys, His-Asp, HisE, Lys-Phe, Met-Phe, Trp-Gly, Asp-His, Met-His, Met-Ile, Tyr-Ile, Cys-Ile, HisLys, TrpLys, CysMet, Ile-Met, Tyr-Met, Glu-Pro, Gln-Arg, Cys-Thr and Gln-Thr) revealing the selectivity for Trp, His and Met. Indeed, 15 Trp-containing, 13 His-containing and 11 Met-containing dipeptides have extremely high or low S_{ij} values. Also these data show that His prefers to locate in the *i*th position, while the preference of Trp is for the *j*th position.

Interestingly, in the total data set (Fig. 1e), there are only five dipeptides, including Cys-Cys, Cys-Trp, His-His, His-Trp and His-Pro, which have extremely high S_{ij} values, while six of them, including Cys-Ala, Glu-Cys, Trp-Gly, His-Lys, Ile-Trp and Met-Trp, have extremely low values of S_{ij}.

The mean values of S_{ij} for homo-dipeptides for all-alpha, all-beta, alpha + beta, alpha/beta and the total data set were 1.29, 1.08, 1.04, 1.22 and 1.21, respectively. This finding indicates that homo-peptides have a nearly random distribution. However, we found that some of them, including His-His, Pro-Pro, Gln-Gln, Met-Met and Cys-Cys, show some degree of frequency significance, which is in good agreement with the available data (Xia and Xie 2002). However, Xia and Xie reported that asymmetry between dipeptides is not significant, that is, the frequency of *ij* is nearly equal to that of *ji*, while as can be seen in Fig. 1, nearly all dipeptides show asymmetry in their amino acid positions.

Comparing our results with other work, particularly that of Vonderviszt et al. (1986), shows that cysteine is a specific amino acid in its selectivity for pairing with other amino acids. However, we observed similar behavior for other residues in the context of protein structural classes. Since cysteine is observed as a special residue in association or avoidance propensity, it appears that this observation may be related to its oxidation state in the structures of proteins, which needs a separate detailed structural study.

The above-mentioned results indicate that some amino acids have S_{ij} values representing their occurrence far from randomness and that they are sensitive to pairing with or avoidance of other amino acids. We also show that positioning in the first or second position of a dipeptide may act as a determinant structural factor for the selection of a given amino acid. The preferences of residues for the first or second positions will be further discussed below. It was also found that dipeptides with association or avoidance far from random distribution are not the same for different structural classes. This indicates that dipeptide selectivity is determined mainly by structural factors rather than simply primary sequence. Since the differences of these structural classes originate from their secondary structures, it may be concluded that the effective parameters for different secondary structures play critical roles in this selectivity.

Other factors in our data are related to the difference in the number of avoided and associated dipeptides. While in the total data set this difference is not significant, using structural class as input data, the number of avoided dipeptides increases compared with associated ones. This fact demonstrates that the unique identity of a single residue is reflected in its pairing characteristics.

For a better understanding of the first step in our analysis, we extended the study by calculating the correlation coefficient between each row- and column-wise pair of S_{ij} matrices as provided in Fig. 2 and Tables 7–11 in the supplementary data. The point of this analysis was to determine the similarity of the different residues localizing in the first or second positions of a given dipeptide. The row- and column-wise correlation coefficients were used to determine how similar the different residues in the first and second positions of a dipeptide were, respectively. Colored font values in the corresponding tables indicate significant low or high correlation between two amino acids that

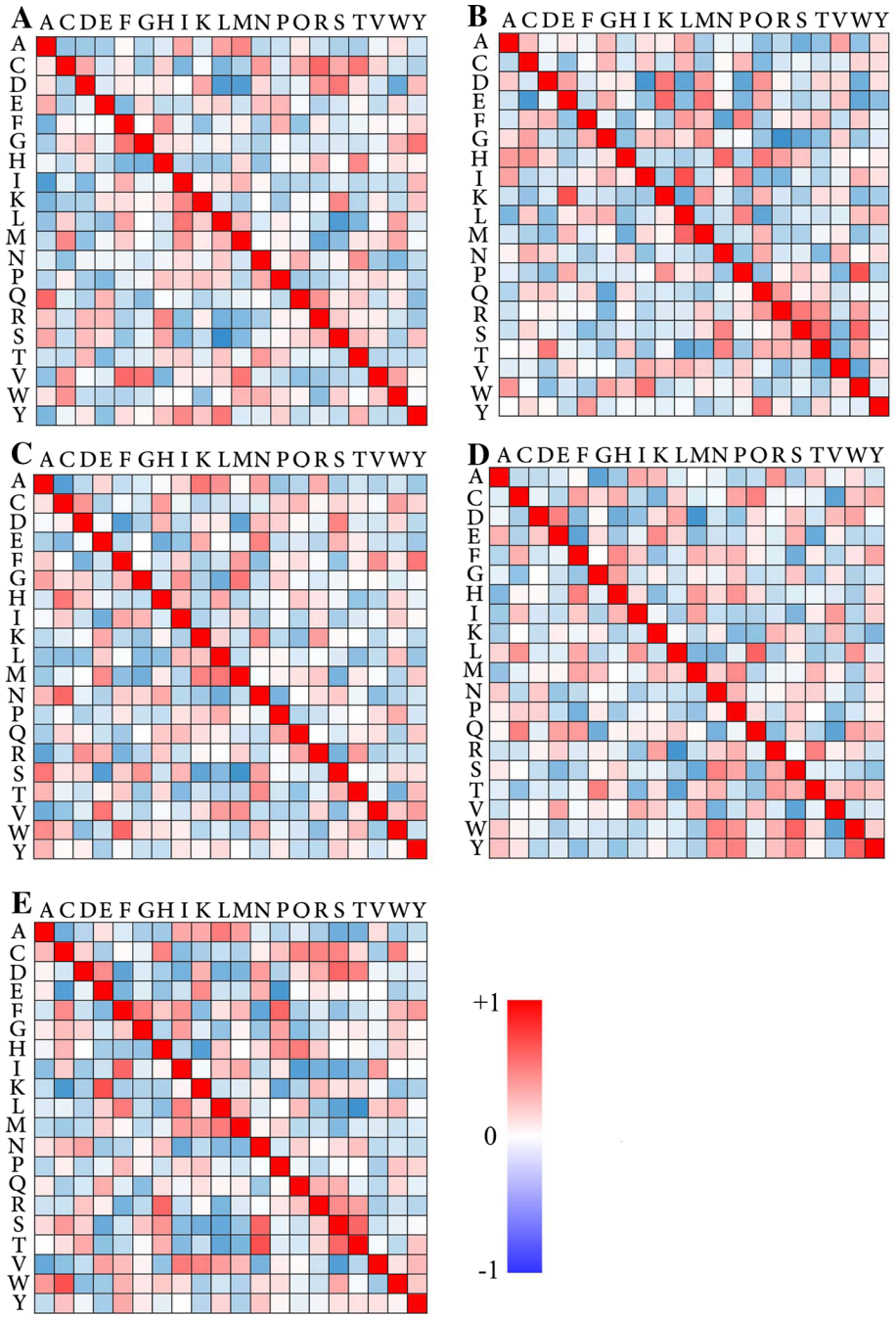

Fig. 2 Graphical representation of correlation coefficients of dipeptide fragment for all-alpha (**a**), all-beta (**b**), alpha + beta (**c**), alpha/beta (**d**) class and our total data set (**e**). Data are analyzed by considering the values in *upper* and *lower part* of the *diagonal line*. The values in the *upper part* refer to the *i*th position and those of the *lower part* are related to the *j*th position of a dipeptide. For better clarification in finding the differences between specific cells, the numerical values are also provided in supplementary material 2

might be substituted for each other, meaning significant dissimilarity or similarity between two given amino acids. Note that the data in Fig. 2 should be analyzed by considering the values above and below the diagonal line. In this analysis, the values above the diagonal line refer to the *i*th position and those in the lower part to the *j*th position of a dipeptide.

The upper part of the data in Fig. 2a for the all-alpha structural class show that Leu & Asp, Met & Asp, Trp & Asp and Ser & Leu residues have significantly low correlation coefficients, which means that substitution of these amino acids for each other in the *i*th position of a dipeptide is avoided. On the other hand, Arg & Cys has a significantly high correlation coefficient meaning a similarity between these two amino acids for localizing in the *i*th position.

Examining the values of the column-wise correlation coefficient (below of the diagonal line in Fig. 2a) shows that Ile & Ala, Leu & Ser and Ala & Tyr have significantly low correlation coefficients indicating dissimilarity between Ile and Ala for positioning in the *j*th position, while Pro & Ala, Phe & Val and Met & Val have significantly high

correlation coefficients, meaning a similarity between Pro and Ala for localizing in the *j*th position of a dipeptide.

A similar procedure was also used for analyzing the other structural classes. Figure 2b contains the correlation coefficient values for the all-beta structural class and shows significant dissimilarity for Ile & Asp, Leu & Asp, Pro & Asp, Trp & Glu, Asn & Phe, Arg & Gly, Ser & Gly, Gln & Leu and Trp & Val. Furthermore, significant similarity for Lys & Glu, Met & Glu, Asn & His, Leu & Ile, Trp & Pro, Thr & Ser and Trp & Ser was observed for localizing in the *i*th position. A significant dissimilarity for Glu & Cys, Gln & Gly and Thr & Leu together with significant similarity for the Lys & Glu pair in the *j*th position was also observed.

For the alpha + beta structural class (Fig. 2c), dissimilarity was observed for Cys & Ala, Phe & Asp, Met & Asp and pairs in the *i*th position and for Asn & Leu, Arg & Ala, Ser & Gln, Ser & Lys and Ser & Met in the *j*th position. Likewise, similarity for Asn & Cys and Trp & Phe can be seen in the *j*th position.

In Fig. 2d the correlation coefficient values are provided for the alpha/beta structural class, showing significant dissimilarity for Gly & Ala, Val & Cys, Met & Asp, Phe & Glu, Thr & Gln, Ser & Phe and Val & Glu and significant similarity for Glu & Leu to localize in the *i*th position. On the other hand, these data show a significant dissimilarity for Arg & Leu, Ser & Met, Thr & Leu, Val & Ser together with Trp & Val and significant similarity for Trp & Ser for positioning at the *j*th position.

Previous reports emphasized that some residues in helices (known as helix formers) tend to be similar and can be substituted with each other (Xia and Xie 2002) but this insight is not confirmed by our results.

Although we examined the frequency of dipeptides in different structural classes of proteins, each structural class has a different content of secondary structural elements, and more studies, including determining the similarity index for any dipeptide in the context of every secondary structure, is needed. Generally, for both the *i*th and *j*th positions, the number of dissimilar amino acids is significantly greater than that of similar ones. As mentioned above, this fact may originate from the unique properties of amino acids, which lead to more sensitivity in selecting their neighbors.

Unexpectedly, it can be seen that a number of different amino acids have a similar behavior in localizing at the same position of a dipeptide, and they can be substituted with each other. As we know, amino acids are classified based on physico-chemical properties such as hydrophobicity, polarity, size and so on. Our data indicate that upon pairing of amino acids, the characteristics of the individual amino acids matter less than those of the pair such that pairs with quite different physico-chemical properties can confer similar features on equivalent positions in protein structures.

It thus seems that the role of each residue in the context of the secondary structure is not the same when considered alone and in pairs. With respect to the local steric interactions in dipeptides based on their side-chain dihedral angle distributions (Jacobson et al. 2002), their S_{ij} values could be studied further to determine how they correlate with the allowed conformations of the dipeptides using, e.g., hard sphere models (Zhou et al. 2012, 2014).

The significance of this work includes analysis of the frequency of dipeptides in every structural class of proteins. We find that determining the tendency of different dipeptides to be found in different defined secondary structural elements could help researchers in an improved understanding of information stored in the sequences of proteins.

References

Aitken A (1999) Protein consensus sequence motifs. Mol Biotechnol 12:241–253. doi:10.1385/MB:12:3:241

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181(80):223–230. doi:10.1126/science.181.4096.223

Anishetty S, Pennathur G, Anishetty R (2002) Tripeptide analysis of protein structures. BMC Struct Biol 2:9

Berman HM (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242. doi:10.1093/nar/28.1.235

Betancourt MR, Skolnick J (2004) Local propensities and statistical potentials of backbone dihedral angles in proteins. J Mol Biol 342:635–649. doi:10.1016/j.jmb.2004.06.091

Costantini S, Colonna G, Facchiano AM (2006) Amino acid propensities for secondary structures are influenced by the protein structural class. Biochem Biophys Res Commun 342:441–451. doi:10.1016/j.bbrc.2006.01.159

Dill KA, MacCallum JL (2012) The protein-folding problem, 50 years on. Science 338:1042–1046. doi:10.1126/science.1219021

Fathinavid A, Khalifeh K, Heshmati E (2017) PARS a software for protein assignment regarding secondary structure. http://www.znu.ac.ir/members/newpage/702. Unpubl. Data. Accessed 5 Jan 2017

Fiser A, Cserzö M, Tüdös É, Simon I (1992) Different sequence environments of cysteines and half cystines in proteins application to predict disulfide forming residues. FEBS Lett 302:117–120. doi:10.1016/0014-5793(92)80419-H

Fu M, Huang Z, Mao Y, Tao S (2014) Neighbor preferences of amino acids and context-dependent effects of amino acid substitutions in human, mouse, and dog. Int J Mol Sci 15:15963–15980. doi:10.3390/ijms150915963

Guruprasad K, Reddy BVB, Pandit MW (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng Des Sel 4:155–161. doi:10.1093/protein/4.2.155

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919. doi:10.1073/pnas.89.22.10915

Hermans J (2011) The amino acid dipeptide: small but still influential after 50 years. Proc Natl Acad Sci USA 108:3095–3096. doi:10.1073/pnas.1019470108

Ismail WM, Chowdhury S (2010) Preference of amino acids in different protein structural classes: a database analysis. 4th International Conference on bioinformatics and biomedical engineering, Chengdu, pp 1–5. doi:10.1109/ICBBE.2010.5514826

Jacobson MP, Kaminski GA, Friesner RA, Rapp CS (2002) Force field validation using protein side chain prediction. J Phys Chem B 106:11673–11680. doi:10.1021/jp021564n

Joosten RP, Te Beek TAH, Krieger E et al (2011) A series of PDB related databases for everyday needs. Nucleic Acids Res. doi:10.1093/nar/gkq1105

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637. doi:10.1002/bip.360221211

Lim VI (1974) Algorithms for prediction of ??-helical and ??-structural regions in globular proteins. J Mol Biol 88:873–894. doi:10.1016/0022-2836(74)90405-7

Nakariyakul S, Liu ZP, Chen L (2011) Protein interaction prediction for mouse PDZ domains using dipeptide composition features. IEEE Int Conf Sys Biol ISB 2011:129–132. doi:10.1109/ISB.2011.6033143

Rani M, Mitra CK, Cserzo M, Simon I (1995) Proteins as special subsets of polypeptides. J Biosci 20:579–590. doi:10.1007/BF02703299

Reddy BVB (1996) Structural distribution of dipeptides that are identified to be determinants of intracellular protein stability. J Biomol Struct Dyn 14:201–210. doi:10.1080/07391102.1996.10508109

Schwartz R, Istrail S, King J (2001) Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. Protein Sci 10:1023–1031. doi:10.1110/ps.33201

Shen S, Kai B, Ruan J et al (2006) Probabilistic analysis of the frequencies of amino acid pairs within characterized protein sequences. Phys A Stat Mech Appl 370:651–662. doi:10.1016/j.physa.2006.03.004

Unger R, Sussman JL (1993) The importance of short structural motifs in protein structure analysis. J Comput Aided Mol Des 7:457–472. doi:10.1007/BF02337561

Vonderviszt F, Matrai G, Simon I (1986) Characteristic sequential residue environment of amino acids in proteins. Int J Pept Protein Res 27:483–492

Xia X, Xie Z (2002) Protein structure, neighbor effect, and a new index of amino acid dissimilarities. Mol Biol Evol 19:58–67

Zhou AQ, O'Hern CS, Regan L (2012) The power of hard-sphere models: explaining side-chain dihedral angle distributions of thr and val. Biophys J 102:2345–2352. doi:10.1016/j.bpj.2012.01.061

Zhou AQ, O'Hern CS, Regan L (2014) Predicting the side-chain dihedral angle distributions of nonpolar, aromatic, and polar amino acids using hard sphere models. Proteins Struct Funct Bioinform 82:2574–2584. doi:10.1002/prot.24621