The Protein Journal (2020) 39:21–32
https://doi.org/10.1007/s10930-020-09880-6

Propensities of Amino Acid Pairings in Secondary Structure of Globular Proteins

Cevdet Nacar[1]

Published online: 13 January 2020

Abstract
A class of secondary structure predictionalgorithms use the information from the statistics of the residue pairs foundin secondary structural elements. Because the protein folding process isdominated by backbone hydrogen bonding, an approach based on backbonehydrogen-bonded residue pairings would improve the predicting capabilities ofthese class algorithms. The reliability of the prediction algorithms depends onthe quality of the statistics, therefore, of the data set. In this study, it was aimed to determinethe propensities of the backbone hydrogen-bonded residue pairings for secondarystructural elements of α-helix and β-sheet in globular proteins using a new andcomprehensive data set created from the peptides deposited in Worldwide ProteinData Bank. A master data set including 4882 globular peptide chains withresolution better than 2.5 Å, sequence identity smaller than 25% and length ofno shorter than 100 residues were created. Separate data sub sets also were createdfor helix and sheet structures from master set and each sub set includes 4594and 4483 chains, respectively. Backbone hydrogen-bonded residue pairings inhelices and sheets were detected and the propensities of them were representedas odds ratios (observed/[random or expected]) in matrices. Propensitiesassigned by this study to the residue pairings in secondary structural elements(as helix, overall strands, parallel strands and antiparallel strands) differfrom the previous studies by 19 to 34%. These dissimilarities are importantand they would cause further improvements in secondary structure predictionalgorithms.

Keywords Secondary structure prediction · Residue pairing · Residue propensity · Hydrogen bonding

1 Introduction

Studies on protein functions mostly require tertiary structure of the protein. Due to the technical limitations, tertiary structure of the many proteins could not be determined by experimental methods such as X-ray diffraction, NMR spectroscopy, cryo electron microscopy or be determined in poor quality. In such cases, computational methods (homology/comparative, threading or ab initio modelling) are valuable approaches to obtain the tertiary structure. In existence of a known structure similar to the query sequence as a template, tertiary structure of an unknown protein chain could be modelled with a great success using homology modeling [1]. The huge number of protein structures in Worldwide Protein Data Bank (wwPDB) [2] is an important factor in this success [3]. If the similar template sequence is not available, de novo or ab initio based prediction methods [4–8] are the main alternative approaches to the homology modeling. De novo prediction methods are mainly based on Anfinsen's thermodynamic hypothesis, which states that the Gibbs free energy of the conformation of a native protein in physiological condition is lowest [9]. Therefore, the main goal of the de novo methods is to find out the global free-energy minimum in conformational energy landscape [10]. However, there are so many local minima in vast conformational energy landscape [11] and it requires enormous amount of time to search the global free-energy minimum among them. Therefore, a qualified starting conformation which corresponding to neighborhood of global minimum in energy landscape and which leading the algorithms to the nearest local minimum is extremely important to overcome this intrinsic limitation [12]. Two of Critical Assessments of

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10930-020-09880-6) contains supplementary material, which is available to authorized users.

✉ Cevdet Nacar
cevnacar@marmara.edu.tr

[1] Department of Biophysics, School of Medicine, Marmara University, Istanbul, Turkey

Methods of Protein Structure Prediction (CASPs), CASP12 and CASP13 [13, 14], have reported great improvements in de novo or template free modeling. Despite these successes, template free modeling requires further improvement, especially, for longer chains.

The qualified starting conformation can be constructed using secondary structure prediction methods (SSPMs) [15–31]. The main goal of SSPMs is to identify the secondary structural elements of the protein in peptide sequence: helices [32], sheets [33] and coils. SSPMs are classified in many ways depending on their approaches to the problem [28, 34]. Some statistical methods or studies those investigating the spatial aspects of β-strands mainly are based on the occurrence of amino acid pairings in α-helices and/or in partner strands of β-sheets in the chain [17, 35–43]. However, because the backbone hydrogen bonding is the dominant factor of the protein folding process, as proposed by Rose et al. [44], the SSPMs based on the frequencies of the hydrogen-bonded amino acid pairings are more reasonable candidates for constructing the qualified starting conformation for de novo methods. The success of latter class SSPMs directly depends on the reliability of the information gathered from statistics of the backbone hydrogen-bonded amino acid pairings of secondary structural elements.

In this study, all backbone hydrogen-bonded residue pairings in α-helices/β-sheets of globular proteins in appropriate data sub sets were determined. The master data set was prepared from the chains deposited in Worldwide Protein Data Bank according to the criteria stated in Sect. 2. The master data set includes 4882 globular, non-homolog protein chains. Two data sub sets also created for helix and sheet structures from master data set. Using the residue pairing frequencies, the propensities of hydrogen-bonded residue pairings in secondary structural elements were calculated as odds ratios.

The helix/sheet propensities of residue pairings were studied by many researchers to some extent in this context [17, 35–43, 45, 46]. However, this study differs from those in both size of the data set and protein type homogeneity. Membrane proteins, fibrous proteins, immunoglobulins, proteins related to extremophile organisms and homolog chains/domains were excluded from the master data set to attain this homogeneity.

Some findings of this study on propensities of the residue pairs are not in consistent with findings of the previous studies. As discussed in Sect. 10 in details, these inconsistencies could be important and make valuable contributions to the secondary structure prediction algorithms.

2 Material and Methods

2.1 Protein Data Sets

150,037 protein structure files in pdb format were downloaded from ftp site of Worldwide Protein Data Bank [47, 48]. The protein structure files those do not include any peptide or secondary structural elements (α-helix or β-sheet) and those do not meet those criteria were excluded from the data set:

Resolution value : ≤ 2.00 Å
Free R-value : ≤ 0.250
R-value : ≤ 0.200 (if Free R-Value not available)
Sequence length : ≥ 100 residues.

Membrane proteins, fibrous proteins, immunoglobulins, and proteins related to extremophile organisms were removed from the data set. Membrane proteins and extremophile organisms were determined according to the lists (see Supp_MembraneProteins.pdf and Supp_ExtremophileOrganisms.pdf, respectively) prepared using the data provided by Stephen White Laboratory at UC Irvine [49] and Wikipedia [50], respectively. Structure files including keywords of *membrane, transmembrane, immunoglobulin, collagen, fibroin, keratin, fibrous, keratous* in COMPND, SOURCE, HEADER, KEYWDS and TITLE record types of their PDB files also were removed. A match ratio higher than 90% between keywords and target word accepted as perfect match. Because proteins are classified according to their type as globular, membrane, fibrous and non-globular in SCOP2 database [51], remaining chains were checked against the list (see Supp_SCOP2.pdf) prepared from SCOP2 web site [52], including membrane, fibrous and non-globular chains; no match found.

2.2 Pairwise Alignment

Amino acid sequences of peptide chains in remaining PDB files were extracted using information from SEQRES entry of the PDB files and identical sequences were removed. If identical sequences are from different PDB files, the sequence has a better resolution left. After the removal of protein tags, remaining 18,384 chain sequences were aligned against to each other, as all possible pairs, using pairwise alignment algorithms in two stages. In first stage, global pairwise alignments were completed using Needleman and Wunsch algorithm [53] in order to detect the homolog chains. In second stage, local pairwise alignments were completed using Smith and Waterman algorithm [54] in order to detect the homolog domains in sequence pairs. In case of an identity value higher than 25%, the longer sequence in length was kept and the other

sequence was excluded. The alignment parameters for both of algorithms are below:

Open gap penalty : 10

Extension gap penalty : 1

Substitution matrix : BLOSUM62 [55, 56]

At the end of the alignments, 4882 chains in 4782 PDB files left (see Supp_MasterDataSet_Chains.pdf and Supp_MasterDataSet_PDBFiles.pdf, respectively).

2.3 Hydrogen Bond Detection

Two different data sub sets were created for each secondary structural element (α-helix, β-sheet) from 4882 chains. Each chain member of the data sub set for helices includes at least one helix as secondary structural element. The same is true for the data sub set for sheet. The residue names, their sequence numbers and boundaries of helices/strands were obtained from HELIX and SHEET entries of the PDB files and residues only within boundaries were involved in hydrogen bond calculations. Residues those modified (information from MODRES entry), those have link with the other atoms (information from LINK entry) and those have missing backbone atoms (C, CA, N, O) (PDB file convention used for representing atoms) were discarded. Because of the resolution limitations of the experimental methods, hydrogen atoms rarely are found in PDB files. The coordinates of missing hydrogen atoms bound to the backbone nitrogen atom (NH) were determined according to the geometrical properties of the peptide bond (Fig. 1). The double-line joining C and O atoms of the nth residue was regarded parallel to the line joining N and H atoms of the (n+1)th residue and the bond length of N–H was accepted as 1.00 Å.

Baker and Hubbard's hydrogen bonding criteria were used to detect the hydrogen bonds [57]. The COH angle is defined as the angle between the lines passing through the C=O and O···H atoms of the mth and nth residues, respectively. Likewise, the NHO angle is the angle between the lines passing through the O···H and H–N atoms of the mth and nth residues, respectively (Fig. 2). Because of the limited number of peptide chains in Baker and Hubbard's study and the existence of peptide chains with worse resolution value in this study, slightly relaxed criteria were chosen for detecting hydrogen bonds.

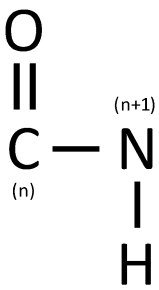


Fig. 1 Geometry of the peptide bond. All atoms are coplanar and line joining the O and C atoms is parallel to the line joining the N and H atoms. The length of bond between the N and H atoms is approximately 1 Å.

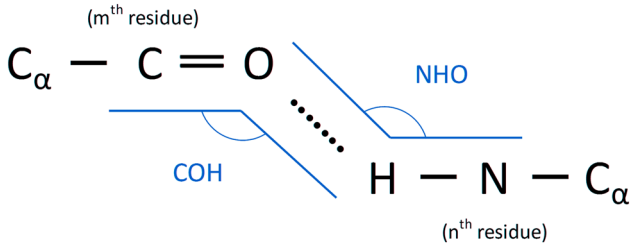


Fig. 2 Depiction of the COH and NHO angles. The hydrogen bond is represented by dotted line between the O atom of the mth residue and the H atom of the nth residue, respectively.

2.3.1 In α-Helices

The helix data sub set includes 4594 chains (see Supp_DataSubSet_Helix.pdf) and each chain includes at least one α-helix. Peptide chains including unusual helices in length (that is, comprise more than 40 residues) were removed from the data sub set to avoid the involvement of the fibrous or extremophile related peptides. The hydrogen bond between the O atom of the nth residue and the HN atom of the (n + 4) th residue in α-helices were traced using those criteria for the bond length of the O···H and COH/NHO angles (Fig. 2).

Bond length : 2.000 ± 0.400 Å

COH angle : 150.0 ± 25.0°

NHO angle : 155.0 ± 25.0°

Any amino acid pair in sequential order of n and (n+4) in helical segment, listed in the HELIX entry of the PDB file, satisfying these criteria accepted as a backbone hydrogen-bonded residue pairing in α-helix. Because the proline residue cannot be a hydrogen bond donor, in such cases, i.e. XXX:PRO (XXX represents any residue), the hydrogen bond calculations were skipped.

2.3.2 In β-Sheets

The data sub set for sheets includes 4483 chains (see Supp_DataSubSet_Sheet.pdf) and each chain includes at least one β-sheet structure. Because of the conformational strains, the residues in partner strands are not aligned one-to-one despite it is depicted in textbooks as it is. The position of the residues in strands may shift a few residue back and forth and a bulb may occur in the strand. Therefore, the backbone hydrogen bonds between O and NH atoms in partner strands were traced by considering all the probable residue matches between the partner strands and those satisfied those criteria were accepted as a backbone hydrogen-bonded residue pairing in β-sheet.

Bond length : 2.000 ± 0.400 Å

COH angle : 150.0 ± 25.0°

NHO angle : 160.0 ± 25.0°

- Depending on the orientation of the strand, these pairings were grouped as parallel, antiparallel and overall. Overall group includes all parallel and antiparallel pairings.

2.4 Odds ratios

If a hydrogen bond was determined between the O and HN atoms of the main chain of two different residues (sequential order of residues for helices and sheets were described in Sects. 2.3.1 and 2.3.2, respectively), these residues were counted as an amino acid pairing. Because of the topology of the antiparallel strands of the sheet, an amino acid pairing may have two hydrogen bonds between their O and HN atoms. In such case, this pairing was counted up twice. Relative abundance or odds ratios of amino acid pairings were calculated as the ratio of observed occurrence to random (or expected) occurrence in peptide chain and were represented by $\mathbf{M_H[i, j]}$ and $\mathbf{M_{SO,\ SP,\ SA}[i, j]}$ matrices for helices and sheets, respectively (H:helix, SO:sheet-overall, SP:sheet-parallel, SA:sheet-antiparallel). The data used to calculate the odds ratios were represented by $\mathbf{A_{H,\ S}[i]}$ and $\mathbf{F_{H,\ SO,\ SP,\ SA}[i, j]}$ matrices. The residue location within the pairing in the strands of the sheet is not preferential, that is, XXX_1:XXX_2 and XXX_2:XXX_1 residue pairings are regarded as identical for sheet in context of matrices. Therefore, $\mathbf{M_{SO,\ SP,\ SA}[i, j]}$ and $\mathbf{F_{SO,\ SP,\ SA}[i, j]}$ matrices are symmetric with respect to the main diagonal but, $\mathbf{M_H[i, j]}$ and $\mathbf{F_H[i, j]}$ matrices are non-symmetric.

Definitions $\mathbf{N_{AAPairs\_H,SO,SP,SA}}$ = Total number of amino acid pairings detected in helices (H), in overall strands (SO), in parallel strands (SP) and in antiparallel strands (SA), respectively.

$\mathbf{N_{AA\_H,\ S}}$ = Total number of amino acids in chains in helix data sub set and in sheet data sub set, respectively.

$\mathbf{A_{H,\ S}[i]}$ = Matrix representing the number of each amino acids in helix data sub set and in sheet data sub set, respectively.

$\mathbf{F_{H,\ SO,\ SP,\ SA}[i, j]}$ = Matrix representing the number of each amino acid pairings detected in helix, in overall strands, in parallel strands and in antiparallel strands, respectively.

$\mathbf{P_{o\_H,\ SO,\ SP,\ SA}(i, j)}$ = Probability of observed occurrence of amino acid pairing i and j in helix, in overall strands, in parallel strands and in antiparallel strands, respectively.

$\mathbf{P_{r\_H,\ SO,\ SP,\ SA}(i, j)}$ = Probability of random occurrence of amino acid pairing i and j in helix, in overall strands, in parallel strands and in antiparallel strands, respectively.

$\mathbf{M_{H,\ SO,\ SP,\ SA}[i, j]}$ = Matrix representing the odds ratio of each amino acid pairings in helices, in overall strands, in parallel strands and in antiparallel strands, respectively.

$$M_{H,SO,SP,SA}[i,j] = \frac{P_{o\_H,SO,SP,SA}(i,j)}{P_{r\_H,SO,SP,SA}(i,j)}$$

$$P_{o\_H,SO,SP,SA}(i,j) = \frac{F_{H,SO,SP,SA}[i,j]}{N_{AAPairs_H,SO,SP,SA}} (i \geq jforstrands)$$

$$P_{r\_H}(i,j) = \frac{A_H[i]A_H[j]}{N_{AA\_H}(N_{AA\_H}-1)}$$

$$P_{r\_SO,SP,SA}(i,j) = \frac{A_S[i](A_S[i]-1)}{N_{AA\_S}(N_{AA\_S}-1)} (ifi = j)$$

$$P_{r\_SO,SP,SA}(i,j) = 2\frac{A_S[i]A_S[j]}{N_{AA\_S}(N_{AA\_S}-1)} (ifi \neq j)$$

$$\sum P_{r\_H,SO,SP,SA} = 1 and \sum P_{o\_H,SO,SP,SA} = 1$$

$$N_{AAPairs\_SO} = N_{AAPairs\_SP} + N_{AAPairs\_SA}$$

$$F_{SO}[i,j] = F_{SP}[i,j] + F_{SA}[i,j]$$

2.5 Single Amino Acid Propensities

Single amino acid propensities to helix and strand were determined using $\mathbf{M_H[i, j]}$ and $\mathbf{M_{SO}[i, j]}$ matrices, respectively. Single amino acid propensities were calculated by normalizing the sum of the values of the cells including the same residue in the matrix (e.g. for ALA residue, all ALA:XXX cell values in $\mathbf{M_H[i, j]}$ matrix or ALA:XXX and XXX:ALA cell values in $\mathbf{M_{SO}[i, j]}$ matrix were summed) according to the normalization condition. Normalization condition is the sum of whole cell values in the related matrix.

Pairwise alignments, chain data extraction from PDB files and calculations for hydrogen bond detection and matrices were done using programs written by author in QB64 v1.2 [58].

3 Results

3.1 Amino Acid Pairing Propensities in α-Helices

$\mathbf{M_H[i, j]}$ matrix for α-helices is shown in Fig. 3 (see Supp_AH_and_FH_Matrices.pdf for $\mathbf{A_H[i]}$ and $\mathbf{F_H[i, j]}$ matrices). Odds ratios of homopairs corresponding to diagonal of the $\mathbf{M_H[i, j]}$ matrix are shaded in gray. An odds ratio higher than

i

j	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	**2,734**	1,648	1,224	1,769	1,448	1,777	1,900	1,155	1,238	1,275	1,575	1,297	1,580	1,143	1,276	1,484	1,297	1,342	1,069	1,197
ARG	1,724	**1,466**	1,013	1,804	0,893	1,483	2,523	0,770	1,107	1,117	1,374	1,016	1,244	0,918	1,000	1,244	1,192	1,244	1,206	1,063
ASN	0,817	0,886	**0,882**	0,976	0,644	1,433	1,460	0,370	0,627	0,908	0,761	0,957	0,996	0,688	0,421	0,839	0,865	1,046	0,794	0,699
ASP	0,721	1,197	0,669	**0,729**	0,445	1,954	0,873	0,386	0,570	0,423	0,519	1,551	0,542	0,335	0,410	0,632	0,536	0,419	0,338	0,356
CYS	1,221	0,697	0,873	0,775	**1,800**	0,822	0,794	0,606	1,088	1,640	1,705	0,566	1,795	1,884	0,405	0,888	1,036	1,803	1,393	1,322
GLN	1,524	1,593	1,329	1,557	0,724	**1,943**	2,144	0,602	1,148	1,109	1,250	1,545	1,237	0,977	0,986	1,303	1,153	1,496	1,135	0,924
GLU	1,326	2,216	1,066	1,292	0,696	2,046	**1,806**	0,512	0,900	0,856	1,008	2,264	0,994	0,644	0,855	0,945	0,863	0,977	0,730	0,783
GLY	0,764	0,517	0,338	0,389	0,486	0,526	0,501	**0,370**	0,365	0,432	0,495	0,381	0,569	0,343	0,238	0,439	0,384	0,447	0,361	0,338
HIS	1,114	0,902	0,663	0,820	0,855	0,980	1,146	0,453	**0,913**	0,947	1,201	0,724	1,206	1,089	0,492	0,845	0,821	1,246	1,196	0,820
ILE	1,251	0,983	0,935	0,927	1,051	1,045	1,063	0,550	1,017	**1,690**	1,690	0,928	1,719	1,165	0,531	1,010	1,332	1,181	1,162	1,386
LEU	1,782	1,404	1,075	1,201	1,480	1,429	1,451	0,825	1,264	1,946	**2,268**	1,077	2,031	1,637	0,726	1,211	1,598	1,434	1,526	1,793
LYS	1,316	1,005	1,076	1,615	0,716	1,428	2,490	0,524	0,881	1,228	1,200	**1,465**	1,191	0,860	0,898	1,216	0,989	1,133	1,362	1,054
MET	1,637	1,248	1,138	1,325	1,575	1,462	1,315	0,819	1,403	1,731	1,909	1,061	**2,878**	1,844	0,673	1,324	1,513	1,704	1,603	1,615
PHE	1,301	0,870	0,822	0,988	1,057	1,087	1,015	0,590	0,749	1,311	1,496	0,817	1,571	**1,328**	0,516	1,019	0,999	1,309	1,024	1,004
PRO	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	**0,000**	0,000	0,000	0,000	0,000	0,000
SER	1,036	0,799	0,741	0,859	0,688	1,012	1,074	0,507	0,676	0,619	0,766	0,787	0,756	0,625	0,442	**0,804**	0,579	0,777	0,636	0,563
THR	0,799	0,771	0,689	0,815	0,558	0,906	0,914	0,457	0,736	0,753	0,778	0,692	0,752	0,682	0,398	0,716	**0,653**	0,819	0,671	0,692
TRP	1,218	1,003	0,981	1,127	1,276	1,382	1,141	0,577	0,885	0,988	1,175	0,751	1,281	1,132	0,559	0,960	0,983	**1,328**	0,899	0,874
TYR	1,329	0,876	0,841	0,882	0,998	0,974	0,967	0,551	0,701	1,186	1,501	0,865	1,429	1,391	0,486	0,876	0,875	1,197	**1,074**	0,952
VAL	1,107	0,728	0,699	0,756	0,814	0,704	0,763	0,514	0,994	1,167	1,238	0,553	1,197	1,283	0,471	0,712	0,995	1,053	1,072	**1,062**

j

$M_H[i,j]$

Fig. 3 $\mathbf{M_H[i, j]}$ matrix represents the odds ratios of amino acid pairings (n, n+4) in helices as [observed]/[random]. A value greater than unity implies the tendency of the residue pairing to helical structure and a higher value corresponds to a higher tendency. Odds ratios of homopairs are shaded in gray

unity implies a higher abundance than expected. Therefore, it reflects the propensity of the pair in helices. 212 of the 400 amino acid pairs have an odds ratio greater than unity and 10 pairs of them have an odds ratio value greater than 2.000. The latter pairs are ALA:ALA, GLU:ARG, ARG:GLU, GLU:GLN, GLN:GLU, GLU:LYS, LYS:GLU, LEU:LEU, MET:LEU and MET:MET. The pairs including ALA, except ALA:ASN, ALA:ASP, ALA:PRO and ALA:THR, have an odds ratio greater than unity. Also, most of the XXX:[GLN, MET, ARG, LEU] pairs have the tendency to exist in helices. Odds ratio of MET:MET, 2.878, is the highest value in the matrix. In contrary, PRO:XXX, XXX:PRO, GLY:XXX, XXX:GLY, XXX:SER and XXX:THR residues, except PRO:ALA, PRO:ARG and GLY:ALA, have smaller odds ratios than unity. 53 of them have a value smaller than 0.500. Because PRO residue cannot act as a donor in hydrogen bonding, scores for XXX:PRO pairs are zero.

There are limited number of studies on α-helical segment in proteins using (n, n+4) pairing [17, 22, 35, 36, 59]. Studies by Gibrat et al. [22], Frishman et al. [17] and Periti et al. [59] include small number of peptides in their data sets and study by Fonseca et al. [36] deals only with residue pairs at the N- and C-termini of the helical segments. Therefore, comparing the findings of this study to these ones would not be conclusive.

However, scope of this study is similar to the one by de Sousa et al. [35] and a meaningful comparison could be obtained. Propensities of homopairs proposed by this study, which correspond to main diagonal of matrix $\mathbf{M_H[i,j]}$, coincide with ones represented in matrix of "Table 1: Global propensities for the (i, i + 4) pairing." by de Sousa et al., except CYS:CYS and TYR:TYR pairs. While this study gives a helical tendency to CYS:CYS and TYR:TYR homopairs by assigning matrix scores of 1.800 and 1.074, respectively, they look neutral and non-helical in the global propensities matrix by de Sousa et al. [35], respectively. There are also 75 heteropair dissimilarities between these two propensity matrices (Here, the word of "*dissimilarity*", implies that propensity score of a residue pair from one study is greater than unity while the corresponding score from other study is smaller than unity or vice versa. Likewise, *"similarity"*, implies that both of propensity scores from different studies are greater or smaller than unity). All these 77 dissimilarities are represented in Fig. 4. The degree of dissimilarity for some pairs, such as VAL:TYR and TYR:TYR, is so small but for some pairs it is not negligible. Because the total number of dissimilarities in propensities of residue pairings for α-helices in these two studies corresponds to 19% of the pairings, these dissimilarities could be crucial when assigning helical secondary structure to primary peptide structure. Therefore, the propensity matrix proposed by this study for α-helical structure could be valuable for secondary structure prediction algorithms.

Last issue of this comparison on matrices is about XXX:PRO residue pairs. Study by de Sousa et al. [35] determines the (n, n + 4) pairings by just considering the position of the residues in helical region, not using hydrogen bonding information. Therefore, in their matrix, "Table 1: Global

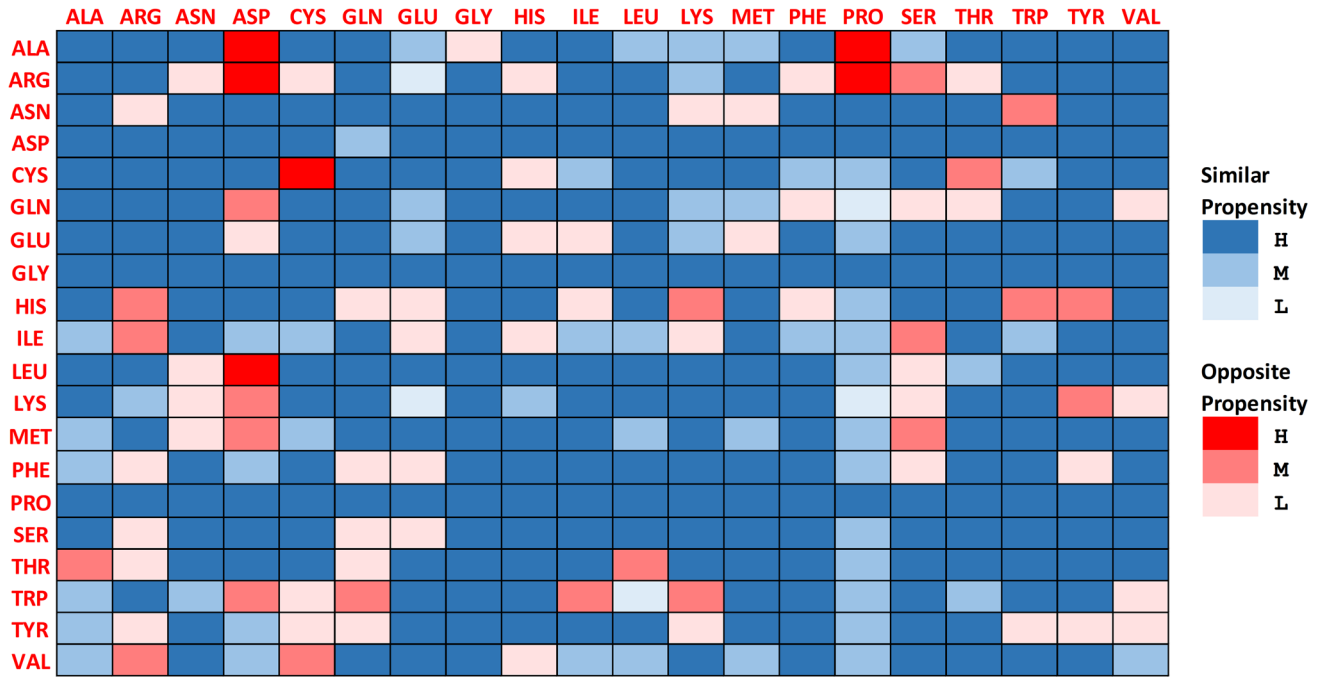


Fig. 4 Comparison of two propensity matrices ($\mathbf{M_H[i, j]}$ matrix and matrix from the study by de Sousa et al. [35]; see the text) for helix structure represented in shades of blue and of red colors. While shades of blue color represent similar propensity, shades of red color do opposite propensity. Color shades are graded as H (high), M (moderate) and L (low) (Color figure online)

propensities for the (i, i + 4) pairing."[35], they have scores greater than zero for XXX:PRO residue pairs. But, because this study is mainly based on the assumption proposed by Rose et al. [44], residue pairings were determined by taking into account the presence of backbone hydrogen bond between the pairs in sequential order of (n, n + 4). Residues at position (n + 4) are hydrogen bond donors and because proline cannot act as a hydrogen bond donor, XXX:PRO residue pairs in $\mathbf{M_H[i, j]}$ matrix cannot have a backbone hydrogen bond. Therefore the scores of XXX:PRO pairs in $\mathbf{M_H[i, j]}$ matrix are zero. This important difference between the matrices could be worth consideration, especially when using secondary structure prediction algorithms based on residue pairings.

3.2 Amino Acid Pairing Propensities in β-Sheets

Residue pairings in β-sheets were grouped as parallel and antiparallel depending on the orientation of the strand or as overall without noticing the orientation. $\mathbf{M_{SO}[i, j]}$, $\mathbf{M_{SP}[i, j]}$ and $\mathbf{M_{SA}[i, j]}$ matrices represent propensities of pairs and are shown in Figs. 5, 6 and 7, respectively (see Supp_AS_and_FSO_Matrices.pdf, Supp_AS_and_FSP_Matrices.pdf and Supp_AS_and_FSA_Matrices.pdf for $\mathbf{A_S[i]/F_{SO}[i, j]}$, $\mathbf{A_S[i]/F_{SP}[i, j]}$ and $\mathbf{A_S[i]/F_{SA}[i, j]}$ matrices, respectively). Because there is no preferential order for the position of the residues in the peptide sequence for sheet structure (that is, ALA:XXX and XXX:ALA pairings are identical in sense of probability calculations in sheet), $\mathbf{M_{SO}[i, j]}$, $\mathbf{M_{SP}[i, j]}$ and $\mathbf{M_{SA}[i, j]}$ matrices are symmetric with respect to the diagonal.

β-sheet propensities of pairs for each matrices are summarized in Table 1 by showing the number of pairs in the group (i.e. ALA:XXX) those have a score greater than unity and those have a score smaller than unity. Because matrices are symmetric, ALA:XXX represents both ALA:XXX and XXX:ALA pairs and so on. $\mathbf{M_{SO}[i, j]}$, and $\mathbf{M_{SA}[i, j]}$ matrices almost have the same tendency profile in general. In $\mathbf{M_{SO}[i, j]}$ matrix, [ILE, TYR, VAL]:XXX pairs, in $\mathbf{M_{SP}[i, j]}$ matrix, [ILE, VAL]:XXX pairs and in $\mathbf{M_{SA}[i, j]}$ matrix, [ILE, TYR, VAL]:XXX pairs have a tendency for corresponding β-strands. In contrary, [ASN, ASP, GLN, GLU, LYS, PRO, SER]:XXX pairs in $\mathbf{M_{SO}[i, j]}$ matrix, [ARG, ASN, ASP, GLN, GLU, GLY, LYS, PRO, SER]:XXX pairs in $\mathbf{M_{SP}[i, j]}$ matrix and [ASN, ASP, GLU, PRO]:XXX pairs in $\mathbf{M_{SA}[i, j]}$ matrix mostly avoid from hydrogen bonding in corresponding β-strands. Due to the limited hydrogen bonding capacity of proline, PRO:XXX pair scores are extremely low.

The remarkable pairing groups are ARG:XXX, GLN:XXX, LYS:XXX, THR:XXX, TRP:XXX and TYR:XXX in parallel and antiparallel strands. While ARG:XXX, GLN:XXX and LYS:XXX pairs are rarely found in parallel strand, THR:XXX, TRP:XXX and

$M_{SO}[i,j]$

i \ j	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	**0,833**	0,702	0,335	0,310	1,215	0,546	0,488	0,559	0,764	1,590	1,106	0,474	0,966	1,302	0,022	0,569	0,841	1,052	1,264	1,834
ARG		**0,739**	0,357	0,732	0,931	1,177	1,619	0,415	1,050	1,365	0,972	0,498	0,871	1,028	0,005	0,893	1,168	1,021	1,292	1,741
ASN			**0,398**	0,283	0,541	0,492	0,493	0,348	0,532	0,750	0,425	0,354	0,406	0,659	0,010	0,602	0,775	0,456	0,594	0,760
ASP				**0,216**	0,510	0,417	0,285	0,295	0,551	0,695	0,419	0,568	0,416	0,451	0,009	0,437	0,712	0,319	0,474	0,787
CYS					**3,203**	0,957	0,785	1,102	1,431	2,395	1,689	0,847	1,524	2,482	0,021	1,095	1,216	2,101	2,622	2,724
GLN						**0,875**	0,495	0,389	0,812	1,125	0,824	0,882	0,795	0,938	0,019	0,759	1,065	0,879	1,038	1,491
GLU							**0,335**	0,290	0,789	1,020	0,572	1,260	0,556	0,620	0,010	0,525	0,933	0,655	0,789	1,159
GLY								**0,461**	0,636	1,164	0,689	0,335	0,650	1,430	0,017	0,437	0,646	1,230	1,209	1,470
HIS									**1,245**	1,814	1,133	0,675	0,889	1,502	0,016	0,854	1,217	1,572	1,515	2,033
ILE										**4,331**	2,764	1,302	2,304	3,075	0,054	1,190	1,780	2,428	2,963	3,741
LEU											**1,936**	0,791	1,447	1,982	0,035	0,819	1,160	1,640	1,960	2,792
LYS												**0,715**	0,656	0,725	0,010	0,815	1,062	0,652	1,020	1,382
MET													**1,771**	2,065	0,030	0,816	1,017	1,409	1,811	2,483
PHE														**3,253**	0,037	0,973	1,387	2,414	2,838	3,379
PRO															**0,000**	0,004	0,008	0,034	0,023	0,049
SER																**0,914**	1,103	0,945	1,052	1,396
THR																	**1,678**	1,013	1,389	2,030
TRP																		**1,342**	2,419	2,668
TYR																			**2,717**	3,324
VAL																				**4,084**

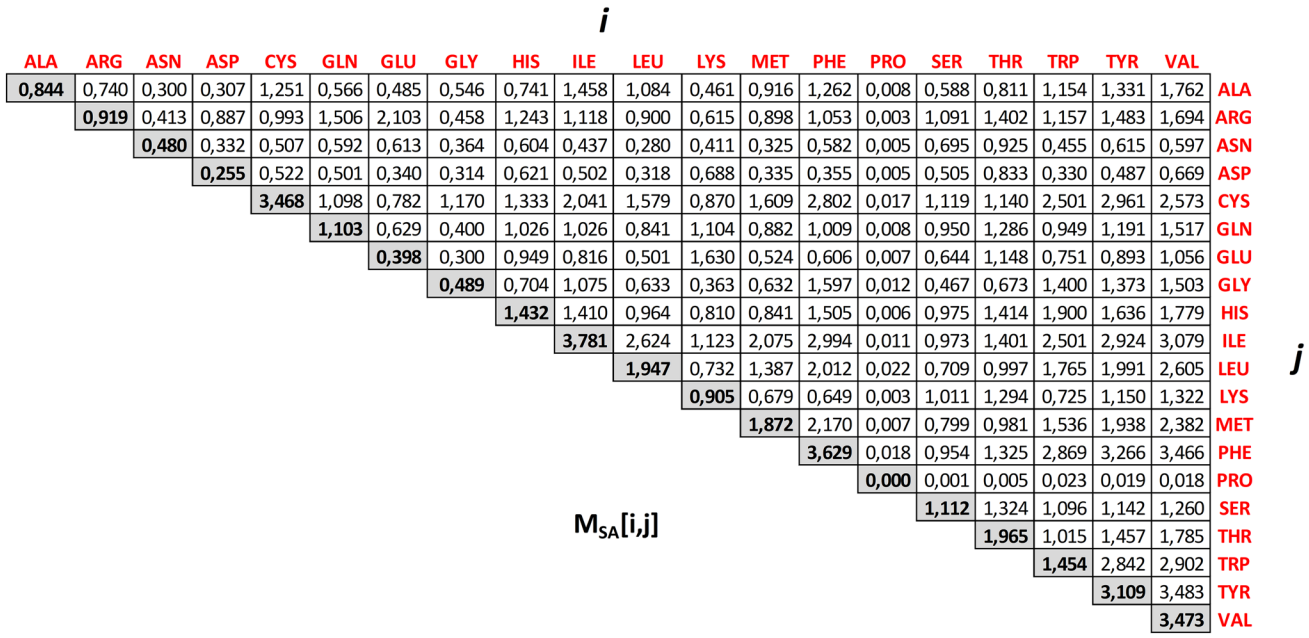
Fig. 5 $\mathbf{M_{SO}[i, j]}$ matrix represents the odds ratio of amino acid pairings in overall strand as [observed]/[random]. A value greater than unity implies the tendency of the residue pairing to sheet structure and a higher value corresponds to a higher tendency. Propensity matrices of sheet strands are symmetric with respect to the diagonal (see the text). Odds ratios of homopairs shaded in gray

$M_{SP}[i,j]$

i \ j	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	**0,799**	0,593	0,435	0,320	1,110	0,487	0,498	0,594	0,832	1,974	1,169	0,509	1,110	1,418	0,064	0,514	0,931	0,753	1,071	2,045
ARG		**0,214**	0,194	0,279	0,750	0,220	0,211	0,292	0,487	2,086	1,180	0,158	0,791	0,955	0,012	0,317	0,488	0,623	0,735	1,876
ASN			**0,157**	0,140	0,643	0,201	0,145	0,302	0,325	1,660	0,846	0,186	0,643	0,884	0,024	0,330	0,337	0,462	0,530	1,237
ASP				**0,104**	0,475	0,172	0,125	0,239	0,350	1,255	0,712	0,216	0,651	0,731	0,021	0,237	0,360	0,287	0,436	1,129
CYS					**2,428**	0,547	0,797	0,903	1,717	3,424	2,010	0,779	1,278	1,551	0,033	1,025	1,438	0,936	1,636	3,163
GLN						**0,210**	0,106	0,357	0,188	1,415	0,774	0,236	0,540	0,728	0,050	0,205	0,425	0,677	0,591	1,412
GLU							**0,150**	0,259	0,325	1,612	0,777	0,184	0,651	0,662	0,020	0,180	0,307	0,377	0,485	1,459
GLY								**0,379**	0,439	1,422	0,852	0,254	0,702	0,943	0,030	0,349	0,568	0,734	0,728	1,373
HIS									**0,701**	2,991	1,623	0,282	1,030	1,492	0,044	0,502	0,643	0,615	1,164	2,773
ILE										**5,934**	3,174	1,822	2,972	3,310	0,179	1,820	2,883	2,213	3,075	5,666
LEU											**1,905**	0,962	1,623	1,894	0,072	1,140	1,634	1,276	1,871	3,336
LYS												**0,162**	0,589	0,947	0,030	0,246	0,387	0,439	0,643	1,556
MET													**1,476**	1,761	0,097	0,865	1,120	1,040	1,443	2,778
PHE														**2,157**	0,093	1,030	1,566	1,091	1,589	3,127
PRO															**0,000**	0,013	0,018	0,066	0,034	0,139
SER																**0,337**	0,460	0,507	0,790	1,790
THR																	**0,844**	1,008	1,192	2,744
TRP																		**1,016**	1,187	1,989
TYR																			**1,575**	2,863
VAL																				**5,865**

Fig. 6 $\mathbf{M_{SP}[i, j]}$ matrix represents the odds ratio of amino acid pairings in parallel strand as [observed]/[random]. A value greater than unity implies the tendency of the residue pairing to sheet structure and a higher value corresponds to a higher tendency. Propensity matrices of sheet strands are symmetric with respect to the diagonal (see the text). Odds ratios of homopairs shaded in gray

TYR:XXX pairs are mainly found in antiparallel strands. Besides those, some specific pairs such as HIS:HIS, SER:SER, THR:THR, TRP:CYS, ILE:ASN also have opposite tendencies for parallel and antiparallel strands. These distinctions in pairing propensities could provide valuable information for making a discrimination between



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL	
ALA	**0,844**	0,740	0,300	0,307	1,251	0,566	0,485	0,546	0,741	1,458	1,084	0,461	0,916	1,262	0,008	0,588	0,811	1,154	1,331	1,762	ALA
ARG		**0,919**	0,413	0,887	0,993	1,506	2,103	0,458	1,243	1,118	0,900	0,615	0,898	1,053	0,003	1,091	1,402	1,157	1,483	1,694	ARG
ASN			**0,480**	0,332	0,507	0,592	0,613	0,364	0,604	0,437	0,280	0,411	0,325	0,582	0,005	0,695	0,925	0,455	0,615	0,597	ASN
ASP				**0,255**	0,522	0,501	0,340	0,314	0,621	0,502	0,318	0,688	0,335	0,355	0,005	0,505	0,833	0,330	0,487	0,669	ASP
CYS					**3,468**	1,098	0,782	1,170	1,333	2,041	1,579	0,870	1,609	2,802	0,017	1,119	1,140	2,501	2,961	2,573	CYS
GLN						**1,103**	0,629	0,400	1,026	1,026	0,841	1,104	0,882	1,009	0,008	0,950	1,286	0,949	1,191	1,517	GLN
GLU							**0,398**	0,300	0,949	0,816	0,501	1,630	0,524	0,606	0,007	0,644	1,148	0,751	0,893	1,056	GLU
GLY								**0,489**	0,704	1,075	0,633	0,363	0,632	1,597	0,012	0,467	0,673	1,400	1,373	1,503	GLY
HIS									**1,432**	1,410	0,964	0,810	0,841	1,505	0,006	0,975	1,414	1,900	1,636	1,779	HIS
ILE										**3,781**	2,624	1,123	2,075	2,994	0,011	0,973	1,401	2,501	2,924	3,079	ILE
LEU											**1,947**	0,732	1,387	2,012	0,022	0,709	0,997	1,765	1,991	2,605	LEU
LYS												**0,905**	0,679	0,649	0,003	1,011	1,294	0,725	1,150	1,322	LYS
MET													**1,872**	2,170	0,007	0,799	0,981	1,536	1,938	2,382	MET
PHE														**3,629**	0,018	0,954	1,325	2,869	3,266	3,466	PHE
PRO															**0,000**	0,001	0,005	0,023	0,019	0,018	PRO
SER																**1,112**	1,324	1,096	1,142	1,260	SER
THR																	**1,965**	1,015	1,457	1,785	THR
TRP																		**1,454**	2,842	2,902	TRP
TYR																			**3,109**	3,483	TYR
VAL																				**3,473**	VAL

Fig. 7 $\mathbf{M_{SA}[i, j]}$ matrix represents the odds ratio of amino acid pairings in antiparallel strand as [observed]/[random]. A value greater than unity implies the tendency of the residue pairing to sheet structure and a higher value corresponds to a higher tendency. Propensity matrices of sheet strands are symmetric with respect to the diagonal (see the text). Odds ratios of homopairs shaded in gray

Table 1 Distribution of properties of the residue pairs in β-sheet

	Parallel		Antiparallel		Overall	
	> 1	< 1	> 1	< 1	> 1	< 1
ALA:XXX	7	13	7	13	7	13
ARG:XXX	3	17	10	10	9	11
ASN:XXX	2	18	0	20	0	20
ASP:XXX	2	18	0	20	0	20
CYS:XXX	11	9	14	6	13	7
GLN:XXX	2	18	10	10	5	15
GLU:XXX	2	18	4	16	4	16
GLY:XXX	2	18	6	14	6	14
HIS:XXX	7	13	10	10	10	10
ILE:XXX	19	1	15	5	17	3
LEU:XXX	13	7	9	11	11	9
LYS:XXX	2	18	7	13	5	15
MET:XXX	11	9	8	12	9	11
PHE:XXX	12	8	14	6	13	7
PRO:XXX	0	20	0	20	0	20
SER:XXX	5	15	8	12	5	15
THR:XXX	8	12	13	7	14	6
TRP:XXX	8	12	14	6	13	7
TYR:XXX	11	9	16	4	16	4
VAL:XXX	19	1	17	3	17	3

parallel and antiparallel strands when using secondary structure prediction algorithms.

Propensities of amino acid pairings in β-sheet structure were studied by many researchers [17, 37–43]. In the study by Fooks et al. [37], the every residue pairing has one hydrogen bonded residue and one non-hydrogen bonded residue and data on antiparallel pairings are not available. The study by Hutchinson et al. [38] also has such an approach to the pairings in antiparallel strand. In the study by Frishman et al. [17], the criteria for X-ray resolution of peptides in the data set is slightly high, the number of peptides in the data set is low and also propensities of residues are not available. Due to these limitations, findings of this study could not be assessed in the viewpoint of these studies. The study by Wouters et al. [40] on antiparallel strands includes a score matrix for hydrogen bonded pairs. At the first glance, the different scores given to ASP:ASP, ILE:ILE, TYR:TYR and VAL:VAL homopairs by two studies deserve interest. While $\mathbf{M_{SA}[i, j]}$ matrix assigns a score for ASP:ASP pair as low as 0.255, it has a tendency for sheet structure according to the Wouters et al. ILE:ILE, TYR:TYR and VAL:VAL homopairs have a 0 score in their study, but they have higher scores in $\mathbf{M_{SA}[i, j]}$ matrix. Despite ASP:LYS and THR:ASN pairs are being the high scoring pairs in the study of Wouters et al. these pairs have scores smaller than unity in $\mathbf{M_{SA}[i, j]}$ matrix. There are more inconsistencies like these ones between these two matrices assigning opposite propensity for the same pair.

In study by Kim et al. [39], favoured and unfavoured pairs in parallel and antiparallel strands are given in "Tables 4–7". According to these tables, the numbers of residues those are favoured in parallel strands, unfavoured in parallel strands, favoured in antiparallel strands and unfavoured in antiparallel strands are 42, 40, 63, and 67, respectively. Of these, only 12, 12, 42, and 45 are overlapped in $\mathbf{M_{SP}[i, j]}$ and $\mathbf{M_{SA}[i, j]}$ matrices.

Despite the lack of discrimination between hydrogen-bonded and non-hydrogen-bonded pairings in the study by Zhang et al. [41], the findings of this study were compared with the ones by them because the data sets of both studies are similar in context of the size and criteria (see for comparison results Supp_ComparisonResultsforSheet.pdf). Because the two other studies by Zhang et al. [42, 43] have inadequate criteria for their data sets, findings of these two studies were not used. Within 210 amino acid pairs, of 66

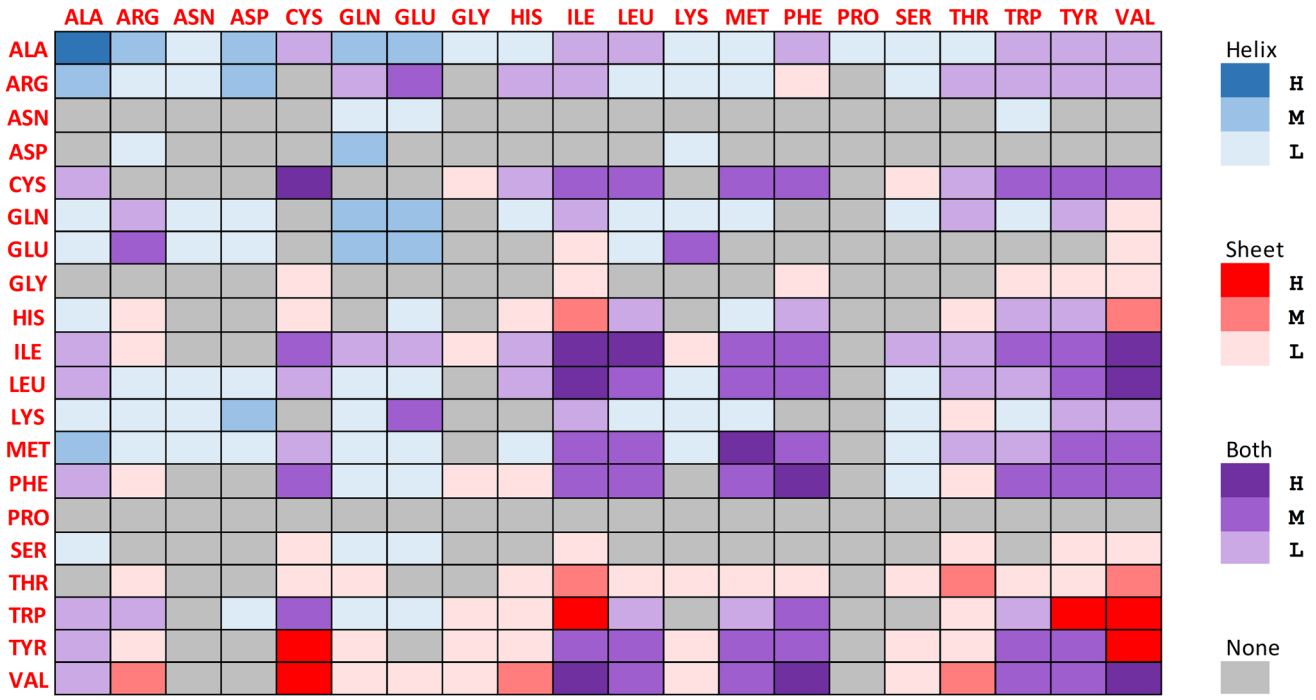


Fig. 8 This combined matrix represents the amino acid pairing propensities to helix and to sheet structures in a single matrix using shades of blue, red and purple colors. Shades of blue color represent the propensity to helix, shades of red color represent the propensity to sheet and shades of purple color represent the propensity to both helix and sheet structures. Therefore, for the same pairing, blue color implies that cell value in the $\mathbf{M_H[i, j]}$ matrix is greater than unity and cell value in $\mathbf{M_{SO}[i, j]}$ matrix is smaller than unity; red color implies that cell value in the $\mathbf{M_H[i, j]}$ matrix is smaller than unity and cell value in $\mathbf{M_{SO}[i, j]}$ matrix is greater than unity; purple color implies that cell values in the both $\mathbf{M_H[i, j]}$ and $\mathbf{M_{SO}[i, j]}$ matrices are greater than unity; gray color implies that cell values in the both $\mathbf{M_H[i, j]}$ and $\mathbf{M_{SO}[i, j]}$ matrices are smaller than unity. Color shades are graded as H (high), M (moderate) and L (low) (Color figure online)

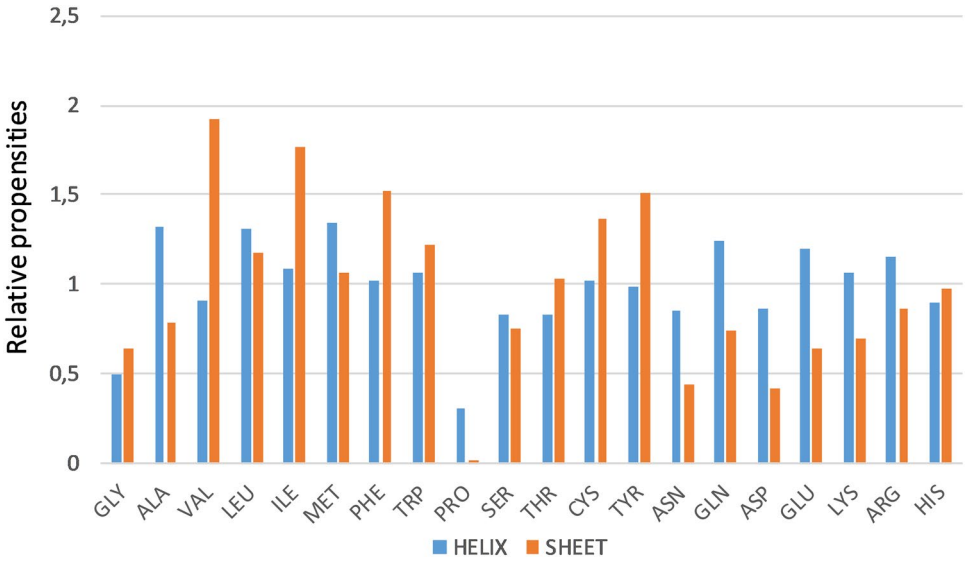


Fig. 9 Single amino acid propensities to helix and sheet structures (see the text)

(31%), 44 (21%) and 72 (34%) pairs have opposite propensity for overall, parallel and antiparallel strands, respectively.

Amino acid pairing propensities to helix and to sheet structures (represented in Fig. 3 and in Fig. 5 as $\mathbf{M_H[i, j]}$ and $\mathbf{M_{SO}[i, j]}$ matrices, respectively) were combined into a single color-coded matrix as in Fig. 8.

3.3 Assessment of the Backbone Hydrogen Bonding Assumption

This study is mainly based on the unproven assumption by Rose et al. [44] which states that energetics of the backbone hydrogen bonding is the dominant factor of the protein folding process. Therefore, if other dominating factors rather than backbone hydrogen bonding are discovered or this assumption is collapsed, the reliability of the findings of this study would reduce partially or completely. In case of the existence of other dominating factors, it is expected that validity of the findings would depend on the weight of backbone hydrogen bonding within the overall factors. But, in case of collapse of backbone hydrogen bonding assumption, the results of this study would become invalid and any consistency between findings of this study and related literature would be accidental.

In a study by Chemmama et al. [60], propensities of amino acid pairings in protein secondary structure were determined using molecular dynamics (MD) simulation. This methodological approach makes their findings free of any single dominant interaction. Therefore, comparing of findings of this study with the ones of Chemmama et al. [60] could be informative, at least to some extent, to assess the reliability of the backbone hydrogen bonding assumption.

Single amino acid propensities were compared using Fig. 9 of this manuscript and Fig. 2 from manuscript of Chemmama et al. [60]. Only propensities to helix and sheet were compared, to coil not included. If an amino acid has same relative propensities to secondary structure in both of these figures, findings for this residue were accepted as in agreement. According to the comparison, 13 of 20 residues (ALA, VAL, LEU, ILE, MET, TRP, THR, ASN, GLN, ASP, GLU, LYS, and HIS) have the same relative tendencies to the secondary structural elements.

This high percentage (65%) in agreement supports the reliability of the backbone hydrogen bonding assumption but, two aspects on methodology of the manuscripts must be taken into account. First, Chemmama et al. [60] used just hexapeptides, which are extremely shorter than an average protein chain. Therefore, in context of protein folding, all potential interactions from distant residues for MD simulation have been ignored. Second, in this study, for propensities to helix, amino acid pairs in a sequential order of (n, n + 4) were traced, and for propensities to sheet, there is no preferential sequential order for residue pairings. But in study of Chemmama et al. [60] only adjacent residue pairs were used.

4 Conclusion

In this study, propensities of amino acid pairings in α-helix and β-sheet structure of globular proteins were determined as odds ratios represented by matrices. Because the reliability of the results mainly depends on the quality of the data set, despite the previous studies on this issue, author has created a new, comprehensive data set using all peptides deposited in Worldwide Protein Data Bank. Only globular protein

chains were included to data set by removing membrane, fibrous, immunoglobulins and extremophile related proteins.

To increase the quality of the data set, both homolog chains and homolog domains in the chains were detected using global and local pairwise alignment algorithms, respectively and were removed from the data set. Because alignment algorithms are heuristic algorithms and alignment parameters has been determined empirically, there is no way to determine the homolog chains or domains as absolutely. Despite this minor drawback, the data set of this study is one of the qualified data set available in the related literature.

Comparison of the findings of this study with the previous studies shows that propensities proposed by this and the other studies for the same residue pairing may differ. The number of such residue pairings corresponds to 19–34% of the all pairings in each secondary structure element. Therefore, findings of this study could provide valuable information to secondary structure prediction algorithms based on hydrogen-bonded residue pairings when predicting secondary structural elements of the peptide.

Funding N/A

Compliance with Ethical Standards

Conflict of interest The author declares that he has no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by the author.

References

1. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815
2. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10(12):980
3. Zhang Y, Skolnick J (2005) The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci USA 102(4):1029–1034
4. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE et al (2001) Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins Suppl 5:119–126
5. Bystroff C, Thorsson V, Baker D (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 301(1):173–190
6. Levitt M, Warshel A (1975) Computer simulation of protein folding. Nature 253(5494):694–698
7. Osguthorpe DJ (1999) Improved ab initio predictions with a simplified, flexible geometry model. Proteins Suppl 3:186–193
8. Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl 3:171–176
9. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181(4096):223–230
10. Bonneau R, Baker D (2001) Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct 30:173–189
11. Scheraga HA (1971) Theoretical and experimental studies of conformations of polypeptides. Chem Rev 71(2):195–217
12. Burgess AW, Ponnuswamy PK, Scheraga HA (1974) Analysis of conformations of amino acid residues and prediction of backbone topography in proteins. Israel J Chem 12(1–2):239–86
13. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2018) Critical assessment of methods of protein structure prediction (CASP)-Round XII. Proteins 86(Suppl 1):7–15
14. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J (2019) Critical assessment of methods of protein structure prediction (CASP)—round XIII. Proteins 87(12):1011–1020
15. Deleage G, Roux B (1987) An algorithm for protein secondary structure prediction based on class prediction. Protein Eng 1(4):289–294
16. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23(4):566–579
17. Frishman D, Argos P (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Eng 9(2):133–142
18. Garnier J, Gibrat JF, Robson B (1996) GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 266:540–553
19. Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120(1):97–120
20. Geourjon C, Deleage G (1994) SOPM: a self-optimized method for protein secondary structure prediction. Protein Eng 7(2):157–164
21. Geourjon C, Deleage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci 11(6):681–684
22. Gibrat JF, Garnier J, Robson B (1987) Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J Mol Biol 198(3):425–443
23. Guermeur Y, Geourjon C, Gallinari P, Deleage G (1999) Improved performance in protein secondary structure prediction by inhomogeneous score combination. Bioinformatics 15(5):413–421
24. King RD, Sternberg MJ (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci 5(11):2298–2310
25. Levin JM (1997) Exploring the limits of nearest neighbour secondary structure prediction. Protein Eng 10(7):771–776
26. Levin JM, Garnier J (1988) Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. Biochim Biophys Acta 955(3):283–295
27. Levin JM, Robson B, Garnier J (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS Lett 205(2):303–308
28. Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232(2):584–599
29. Rost B, Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19(1):55–72
30. Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry 13(2):222–245
31. Chou PY, Fasman GD (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13(2):211–222
32. Pauling L, Corey RB, Branson HR (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 37(4):205–211

33. Pauling L, Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci USA 37(5):251–256
34. Deleage G, Blanchet C, Geourjon C (1997) Protein structure prediction. Implications for the biologist. Biochimie 79(11):681–686
35. de Sousa MM, Munteanu CR, Pazos A, Fonseca NA, Camacho R, Magalhaes AL (2011) Amino acid pair- and triplet-wise groupings in the interior of alpha-helical segments in proteins. J Theor Biol 271(1):136–144
36. Fonseca NA, Camacho R, Magalhaes AL (2008) Amino acid pairing at the N- and C-termini of helical segments in proteins. Proteins 70(1):188–196
37. Fooks HM, Martin AC, Woolfson DN, Sessions RB, Hutchinson EG (2006) Amino acid pairing preferences in parallel beta-sheets in proteins. J Mol Biol 356(1):32–44
38. Hutchinson EG, Sessions RB, Thornton JM, Woolfson DN (1998) Determinants of strand register in antiparallel beta-sheets of proteins. Protein Sci 7(11):2287–2300
39. Kim SB, Tsui KL, Borodovsky M (2006) Multiple testing in large-scale contingency tables: inferring patterns of pair-wise amino acid association in beta-sheets. Int J Bioinform Res Appl 2(2):193–217
40. Wouters MA, Curmi PM (1995) An analysis of side chain interactions and pair correlations within antiparallel beta-sheets: the differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. Proteins 22(2):119–131
41. Zhang N, Duan G, Gao S, Ruan J, Zhang T (2010) Prediction of the parallel/antiparallel orientation of beta-strands using amino acid pairing preferences and support vector machines. J Theor Biol 263(3):360–368
42. Zhang N, Ruan J, Duan G, Gao S, Zhang T (2009) The interstrand amino acid pairs play a significant role in determining the parallel or antiparallel orientation of beta-strands. Biochem Biophys Res Commun 386(3):537–543
43. Zhang N, Ruan J, Wu J, Zhang T (2007) SHEETSPAIR: a database of amino acid pairs in protein sheet structures. Data Sci J 6:S589–S595
44. Rose GD, Fleming PJ, Banavar JR, Maritan A (2006) A backbone-based theory of protein folding. Proc Natl Acad Sci USA 103(45):16623–16633
45. Lifson S, Sander C (1980) Specific recognition in the tertiary structure of beta-sheets of proteins. J Mol Biol 139(4):627–639
46. Petersen SB, Neves-Petersen MT, Henriksen SB, Mortensen RJ, Geertz-Hansen HM (2012) Scale-free behaviour of amino acid pair interactions in folded proteins. PLoS ONE 7(7):e41322
47. ww PDBc (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res 47(D1):D520–D528
48. Worldwide Protein Data Bank. FTP site. http://ftp.wwpdb.org/pub/pdb/data/structures/divided/pdb/. Accessed 16 Apr 2019
49. Stephen White laboratory at UC Irvine. Membrane Proteins of Known 3D Structure. https://blanco.biomol.uci.edu/mpstruc/ Accessed 16 Apr 2019
50. Wikipedia The Free Encyclopedia. Extremophile. https://en.wikipedia.org/wiki/Extremophile Accessed 2 May 2019
51. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (2014) SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res 42(Database issue):D310–D314
52. MRC Laboratory of Molecular Biology. Structural Classification of Proteins 2. https://scop2.mrc-lmb.cam.ac.uk/pt-index.html. Accessed 11 Oct 2019.
53. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453
54. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197
55. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919
56. NCBI National Center for Biotechnology Information. BLOSUM Matrices. ftp://ncbi.nih.gov/blast/matrices/. Accessed 2 May 2019.
57. Baker EN, Hubbard RE (1984) Hydrogen bonding in globular proteins. Progr Biophys Mol Biol 44(2):97–179
58. QB64. https://www.portal.qb64.org/ . Accessed 21 Oct 2019.
59. Periti PF, Quagliarotti G, Liquori AM (1967) Recognition of alpha-helical segments in proteins of known primary structure. J Mol Biol 24(2):313–322
60. Chemmama IE, Chapagain PP, Gerstman BS (2015) Pairwise amino acid secondary structural propensities. Phys Rev E 91(4):042709