J Mol Model (2008) 14:769–775
DOI 10.1007/s00894-008-0313-0

ORIGINAL PAPER

A reexamination of the propensities of amino acids towards a particular secondary structure: classification of amino acids based on their chemical structure

Saša N. Malkov · Miodrag V. Živković · Miloš V. Beljanski · Michael B. Hall · Snežana D. Zarić

Received: 7 December 2007 / Accepted: 8 April 2008 / Published online: 27 May 2008

Abstract The correlation between the primary and secondary structures of proteins was analysed using a large data set from the Protein Data Bank. Clear preferences of amino acids towards certain secondary structures classify amino acids into four groups: α-helix preferrers, strand preferrers, turn and bend preferrers, and His and Cys (the latter two amino acids show no clear preference for any secondary structure). Amino acids in the same group have similar structural characteristics at their Cβ and Cγ atoms that predicts their preference for a particular secondary structure. All α-helix preferrers have neither polar heteroatoms on Cβ and Cγ atoms, nor branching or aromatic group on the Cβ atom. All strand preferrers have aromatic groups or branching groups on the Cβ atom. All turn and bend preferrers have a polar heteroatom on the Cβ or Cγ atoms or do not have a Cβ atom at all. These new rules could be helpful in making predictions about non-natural amino acids.

S. N. Malkov · M. V. Živković
Department of Mathematics, University of Belgrade,
Studentski trg 16,
11000 Belgrade, Serbia

M. V. Beljanski
Institute of General and Physical Chemistry,
Studentski trg 16,
11000 Belgrade, Serbia

M. B. Hall
Department of Chemistry, Texas A&M University,
College Station, TX 77843-3255, USA

S. D. Zarić (✉)
Department of Chemistry, University of Belgrade,
Studentski trg 16,
11000 Belgrade, Serbia
e-mail: szaric@chem.bg.ac.yu

Keywords Amino acid · Protein · Protein secondary structure · Statistical correlation

Introduction

The conformational preferences of amino acids are very important for understanding conformational interactions in proteins. Moreover, when used as propensities they can be helpful in predicting secondary and tertiary structures of proteins. There are many methods of addressing protein folding, and many of them use information regarding the protein secondary structure [1–15]. The structural preferences of amino acids were introduced and calculated a long time ago, and it is known that different amino acids have distinct preferences for the adoption of helical, strand, and random coil conformation [16–23]. Levitt [18] observed that 19 of the 20 naturally occurring amino acids have preferences for only one of the several types of secondary structure, leading to a very clear classification of amino acids by their preferences. It was shown that these preferences and classifications correlate with the chemical structure of amino acids. More recently, the position-dependent amino acids propensities have been studied [24–27].

Although much is known about secondary and tertiary protein structure and folding, the process of folding is not understood completely. The molecular mechanism of protein self-assembly is still an open question [28]. It is believed that the energetics of side chain interactions dominate protein folding processes. However, it was shown that secondary structure can determine native protein conformation, devoid of side chains [29, 30]. Recently, a backbone-based theory of protein folding was proposed, where the protein folding mechanism is based on backbone hydrogen bonding [31], while α-helix and β-sheet propen-

sities are closely connected with the energetics of peptide hydrogen bonds [32].

In this work, we studied the preferences of amino acids for secondary structures in terms of statistical correlation using a large data set from the Protein Data Bank (PDB). Although many of our results with this much larger data set are in accord with results obtained in the 1970s on a very small number of proteins [18, 33, 34], we have identified a number of important differences. Furthermore, our new results allow us to show more clearly how the chemical structure of amino acids plays a major role in determining their preferences for specific secondary structures. Important differences discovered here enabled us to determine rules for predicting the preference of an amino acid towards a particular secondary structure type based only on the chemical structure of its substituents at the Cβ or Cγ atoms. To the best of our knowledge, this is the first improvement in connecting amino acid preferences with their chemical structures since 1978 [18].

Methods

Secondary structure types are assigned by DSSP [35], and are denoted using letters: H for α-helix, B for isolated β-bridge, E for extended strand, G for 3-helix, I for 5-helix, T for hydrogen bonded turn, and S for bend. All other structural elements not belonging to these secondary structure types are considered coil and are denoted by C. Secondary structure types are often reduced to only three; H, E, and C [36, 37]. Here, we consider all eight secondary structure types, including coils.

Computational model

Consider a set P of n protein chains. Primary structures of these protein chains are described by sequences a_1,... a_n. If len(i) denotes the length of a sequence a_i, then residues of the sequence a_i are $a_{i,1}$,... $a_{i,len(i)}$, $1 \le i \le n$. The corresponding assigned secondary structures are described by sequences b_1,..., b_n, where b_i is a sequence of residues $b_{i,1}$,..., $b_{i,len(i)}$, $1 \le i \le n$.

If A is a logical expression, then the indicator variable $I(A)$ is defined by:

$$\mathrm{I(A)} = \begin{cases} 1 & , \quad A = true \\ 0 & , \quad A = false \end{cases}$$

Let $X_{ij}(s) = I(b_{ij} = s)$ and $Y_{ij}(p) = I(a_{ij} = p)$ denote binary random variables corresponding to events that the secondary structure type assigned to residue a_{ij} is s, and that a_{ij} is the amino acid p, respectively. Then, let

$$Z_{ij}(s,p) = X_{ij}(s) \cdot Y_{ij}(p) \tag{1}$$

denote the random variable corresponding to the joint event that the residue a_{ij} is the amino acid p, and the secondary structure type s is assigned to it.

Let us now introduce some notation. $NPS(s, p)$ is the number of times the residue of amino acid p has the secondary structure type s:

$$\mathrm{NPS(s,p)} = \sum_S Z_{ij}(s,p);$$

$NP(p)$ is the number of occurrences of the amino acid p:

$$\mathrm{NP(p)} = \sum_s NPS(s,p);$$

$NS(s)$ is the number of occurrences of the secondary structure type s:

$$\mathrm{NS(s)} = \sum_p NPS(p,s).$$

The total count of observed residues is

$$\mathrm{N} = \sum_{p,s} NPS(s,p). \tag{2}$$

The correlation coefficient of random variables X and Y is defined by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}},$$

(the reader is referred to e.g. the book by Samuels and Witmer [38]). If both variables are binary, then

$$\rho(X,Y) = \frac{\overline{XY} - \overline{X}\ \overline{Y}}{\sqrt{\overline{X}(1-\overline{X})\overline{Y}(1-\overline{Y})}}.$$

The correlation coefficient is always in the range [−1, 1]. It is 0 if X and Y are independent. The correlation coefficient is 1 or −1 if and only if the random variables are linearly dependent.

Consider the correlation of random variables $X_{ij}(s)$ and $Y_{ij}(p)$,

$$\rho\left(X_{ij}(s), Y_{ij}(p)\right) = \frac{\overline{Z_{ij}(s,p)} - \overline{X_{ij}(s)}\ \overline{Y_{ij}(p)}}{\sqrt{\overline{X_{ij}(s)}\left(1-\overline{X_{ij}(s)}\right)\overline{Y_{ij}(p)}\left(1-\overline{Y_{ij}(p)}\right)}},$$

where $Z_{ij}(s, p)$ is defined by Eq. 1. Assuming that the distributions of $X_{ij}(s)$ and $Y_{ij}(p)$ depend only on p and s (i.e. they are independent of the choice of sequence and position inside the sequence), we estimate the correlation coefficients by

$$\rho(s,p) = \frac{\overline{Z(s,p)} - \overline{X(s)}\ \overline{Y(p)}}{\sqrt{\overline{X(s)}\ \left(1-\overline{X(s)}\right)\overline{Y(p)}\left(1-\overline{Y(p)}\right)}}.$$

The estimates of means of X, Y and Z are

$$\overline{X(s)} = NS(s)/N,$$
$$\overline{Y(p)} = NP(p)/N,$$
$$\overline{Z(s,p)} = NPS(s,p)/N.$$

Hence the correlation coefficient estimate is

$$\rho(s,p) = \frac{NPS(s,p)\cdot N - NP(p)\cdot NS(s)}{\sqrt{NP(p)\cdot(N-NP(p))\cdot NS(s)\cdot(N-NS(s))}}. \quad (3)$$

The value of $\rho(s, p)$ is positive (negative, zero) if the pair (p, s) occurs more (less, equally) frequently than it would occur if p and s were independent.

In order to evaluate the significance of the correlation coefficient, we compute the statistic t_s

$$t_s = \rho\sqrt{\frac{N-2}{1-\rho^2}},$$

Under the assumption that the correlation coefficient is 0, the distribution of the statistic t_s is t-distribution with N–2 degrees of freedom (see Samuels and Witmer [36], for examples).

If the sample size N is large, then the t-distribution is approximated by the normal distribution $N(0,1)$. Let the null hypothesis be that X and Y are independent, i.e. that there is no dependence of secondary structure type s and amino acid p. The null hypothesis is considered false if it implies that the probability of obtaining a correlation coefficient estimate with an absolute value greater than calculated is less than 0.05. If the null hypothesis is true, using the normal distribution approximation we obtain that the probability of the event "$|t_s|$ is greater than t_{lim}" is 0.05 for t_{lim}=1.96. Hence, the correlation coefficient is significant, and we consider that X and Y are dependent if $|t_s| \geq$ 1.96. If we denote the corresponding value of the correlation coefficient by ρ_{lim}, then the correlation coefficient is significant if

$$|\rho| \geq \rho_{lim} = \frac{t_{lim}}{\sqrt{t_{lim}^2 + N - 2}} = \frac{1.96}{\sqrt{3.84 + N}}. \quad (4)$$

Data sets

As a source of protein data we used PDB release #103 from January 2003, containing 18,482 proteins [39]. The secondary structure assignment was performed by the program DSSP [35]. There are many families of proteins that are over-represented in PDB. The full set of protein sequences was filtered to eliminate redundant data—we used the PDBSELECT list of nonredundant protein chains [40], with a threshold of 25% [1]. PDBSELECT is designed to both reduce internal homologies and preserve the selection representativeness. It contains sequences with best overall quality, considering sequence source and technique, resolution, completeness and length. The resulting set contains 1,737 sequences with 282,329 amino acid residues. The corresponding value of ρ_{lim} is 0.0037.

Results and discussion

The values for the correlation of amino acids with secondary structure types were computed with the PDBSELECT subset of protein sequences using Eq. 3; 160 correlation values were calculated (8 types of secondary structures, 20 amino acids). Based on these correlation values, the amino acids were classified into four groups according to their preferences to participate in a particular secondary structure (Table 1). With some important exceptions, the classification is in agreement with previous results [18].

Amino acids and secondary structure types

The correlation values for the secondary structure type at the position of amino acid are presented in Table 1. Amino acids are classified according to their preference for particular secondary structures. Most amino acids (exceptions are Thr, Cys, His) show clear preference for one particular secondary structure. The groups are: α-helix preferrers, strand preferrers, turn and bend preferrers, and the fourth group consists of amino acids that show no preference for any of the secondary structure types. Amino acids in each group are ordered by the correlation values.

α-Helix preferrers

The amino acids from the first group in Table 1 (Ala, Leu, Glu, Gln, Arg, Met and Lys) are helix preferrers, showing a preference for building α-helices. Because of the high level of their correlations with α-helices, three of these amino acids (Ala, Leu, Glu) could be further classified as strong helix preferrers.

Previous findings [16, 18, 41] agree only partially with our results. Previously, His and Cys [18] and His and Val [16] were classified as very frequent in helical regions, whereas with the larger data set used here all of these amino

[1] The correlations were calculated for some other thresholds with no significant differences. While specific correlation values differ, the trends and general conclusions are the same. The threshold of 25% is subjectively estimated as a good measure because smaller thresholds raise redundancy and larger thresholds reduce the sample size.

Table 1 Values of correlation coefficients of amino acids and secondary structure types, and elements of amino acid structure. Correlation values were calculated using Eq. 3, multiplied by 10,000. To emphasise the most important correlations, significant positive correlation coefficients (ρ> 0.015) are in bold, and significant negative correlation coefficients (ρ <-0.015) are in italics. The three rightmost columns contain information on the structural properties of the amino acids. If an amino acid has branch on Cβ, or an aromatic Cγ atom, the appropriate cell is marked. Otherwise it is empty. If there is a polar heteroatom on Cβ or Cγ, the chemical symbol for the atom is presented in the last column

Preference	Amino acid	α-helix (H)	Strand (E)	Turn (T)	Bend (S)	3-helix (G)	Coil (C)	Branch on Cβ	Aromatic Cγ	Polar hetero atom on Cβ or Cγ
α-helix preferrers	Ala	**825**	*-357*	*-183*	*-279*	90	*-248*			
	Leu	**766**	**174**	*-476*	*-346*	-34	*-413*			
	Glu	**642**	*-392*	32	-58	131	*-358*			
	Gln	**413**	*-247*	-60	-46	57	-159			
	Arg	**297**	-107	-122	-53	-1	-106			
	Met	**265**	-19	*-212*	*-169*	0.2	4			x[a]
	Lys	**242**	*-264*	75	52	6	-101			
Strand preferrers	Val	*-172*	**1,280**	*-581*	*-350*	*-274*	*-281*	X		
	Ile	144	**945**	*-566*	*-371*	*-223*	*-326*	X		
	Tyr	-53	**470**	*-192*	*-150*	9	*-184*		X	
	Phe	91	**458**	*-256*	*-191*	-5	*-163*		X	
	Thr	*-391*	**281**	*-198*	120	*-161*	**287**	X		O
	Trp	87	**157**	-146	-123	66	-89		X	
Turn and bend preferrers	Gly	*-1050*	*-492*	**1,380**	**846**	-112	73			
	Asn	*-379*	*-450*	**567**	**302**	24	**217**			O
	Pro	*-895*	*-664*	**456**	136	**164**	**1,160**			N
	Asp	*-304*	*-623*	**357**	**339**	**157**	**400**			O
	Ser	*-373*	*-203*	61	**249**	**156**	**334**			O
Other	Cys	-42	94	-74	-67	-40	70			S
	His	-118	2	25	66	61	34		X	N

[a] There is a nonpolar sulfur atom on Cγ atom in Met

acids show a negative correlation with α-helix. Also, the finding that Arg is indifferent to helices [18] is negated here, and we classify Arg as a clear helix preferrer. This is in accord with experimental thermodynamics data showing that Arg has high tendency to form helical secondary structure [21].

The amino acids from this group exhibit similar behaviour with respect to other secondary structures, but there are also differences in their behaviour. Amino acids Ala, Glu, Gln, Arg and Lys have negative correlation for strands and coils. In contrast, Leu is a unique amino acid in this group that tends to occur in strands. It prefers short strands and has a negative correlation with longer strands and coils. Met is relatively neutral to its appearance in strands, 3-helices and coils. Ala, Leu and Met show negative correlations with turns and bends, while Glu appears to support the formation of short 3-helices.

Strand preferrers

The amino acids from the second group (Val, Ile, Tyr, Phe, Thr and Trp) prefer strands. Threonine is unique among strand preferrers, and indeed among all amino acids, because it has almost the same correlation value with strands and with coils. We nevertheless put Thr among strand preferrers because of its large negative correlation value for turns. However, it differs from other members of the group.

Our results about strands agree with previous results [16,18], with some differences in correlations of amino acids from other groups with strands. In earlier results, Met [16, 18] and Cys [16, 18, 41] were among the strongest β-sheet formers, while our data show that Met has a slightly negative, and Cys exhibits only a small positive, correlation.

Some strand preferrers are weakly correlated with α-helices. All strand preferrers obstruct the formation of turns and bends, except Thr, which supports the formation of bends.

Turn and bend preferrers

The amino acids from the third group (Gly, Asn, Pro, Asp and Ser) exhibit a preference to build bends or turns or coils. This is in full agreement with results of Levitt [18], and in good agreement with the results of Chou and Fasman [16] (who found that Pro, Gly, Asn and Ser are the most frequent coil residues) and of Gibrat et al. [41] (who added Asp to the group).

Turn and bend preferrers, Gly, Asn, Pro, Asp and Ser, have quite large positive correlation values for bends, turns,

3-helices and coils. Only Gly has a negative value for 3-helices, Asn has a very small value for 3-helices, and Ser a small value for turns. All amino acids in this group occur rarely in α-helices and strands. Glycine has very high tendency to build turns and it appears very often at turn ends. Proline tends to initiate turns. Proline also tends to appear in terminating parts of bends and coils. Proline, Asp and Ser support the formation of 3-helices.

Cysteine and histidine

The remaining two amino acids, Cys and His, are relatively weakly correlated with all secondary structure types. The correlation values suggest statistical significance, but these values are substantially lower compared to other amino acids. Cysteine and His do not show clear preference to build any particular secondary structure. With small correlation coefficients, Cys tends to build strands, while His has negative correlation with α-helices.

It is interesting that the results of Levitt [18] show a preference of His and Cys towards α-helices, while the results of Chou and Fasman [16] show a preference of His towards α-helices, and a preference of Cys towards β-sheet structures. These differences are considered below.

Influence of amino acids properties on their preference for particular secondary structures

Polarity and amino acid size

An amino acid's tendency to form a certain secondary structure can be related to its physicochemical properties. For example, strands consist mostly of hydrophobic amino acids [16].

It was recently shown that the propensity of amino acids for certain positions within a helix depends on physico-chemical properties [26]. Polar and nonpolar amino acids show different phase distribution—they usually appear at different positions in helices.

Classification of amino acids as long polar (Glu, Gln, Arg, Lys), short polar (Asn, Asp, Ser), hydrophobic aromatic (Phe, Tyr, Trp), and hydrophobic aliphatic (Leu, Met, Val, Ile) can be related to our results. All long polar amino acids are α-helix preferrers, all aromatic amino acids are strand preferrers, while all short polar amino acids are turn and bends preferrers. However, hydrophobic aliphatic amino acids do not belong to any one group of preferrers (Table 1); some are α-helix preferrers, and some are strand preferrers. Looking at the structures of aliphatic amino acids reveals that amino acids with branching at the Cβ atom are strand preferrers (Val and Ile), while amino acids without branching on the Cβ atom are α-helix preferrers (Ala, Leu, Met). Using our classification, some more general rules related to the structural properties of amino acids can be defined. We will discuss these below.

All strand preferrers are hydrophobic, with the exception of Thr, which is slightly polar. The preference of nonpolar amino acids for strands has been known for a long time [16, 41]. It is also supported by the fact that all polar amino acids that belong to α-helix preferrers and turn and bend preferrers, have a negative correlation with strand structures while the hydrophobic Leu has a positive correlation. This is in agreement with the finding that proteins with increased hydrophobicity are less resistant to misfolding [42].

Among α-helix preferrers there are hydrophobic and long polar amino acids. The preference of α-helix for hydrophobic amino acids is also supported by the fact that hydrophobic aromatic amino acids have a positive (Phe, Trp), or slightly negative (Tyr) correlation with α-helices, while all short polar amino acids show a negative correlation.

All turn and bend preferrers are small polar amino acids, except Gly and Pro. The tendency of small polar amino acids to build turns, bends, and coils is also in agreement with the positive correlation of Thr (which is classified as strand preferrer) with bends and coils (Table 1). Gly and Pro are the only exceptions in this group, since they are not polar.

Structural properties of amino acids

The groupings of amino acids given in Table 1 are based on our results on the preference of an amino acid to be part of a certain secondary structure. However, our classification, based on the correlation in a larger data set, show more clearly than previous classifications [16, 18] that amino acids belonging to the same group have similar structural properties, which differ from the properties of amino acids in other groups. From this recognition, we can develop rules that enable the classification of amino acids as preferrers of certain secondary structures based only on the structural properties of the amino acid's substituents at the Cβ and Cγ atoms.

In all α-helix preferrers there are two hydrogens and one carbon (in Ala there are three hydrogen atoms) on the Cβ atom and there is no branching on the Cβ atom; the Cγ atom is aliphatic (sp^3 hybridisation) and there are no heteroatoms—only hydrogen and carbon atoms are on Cγ. The only exception is Met, with sulfur on the Cγ atom. Met is probably an α-helix preferrer because in this case sulfur does not result in strong polarity. Hence, we can say that there are no polar heteroatoms on the Cγ atom, and classify Met as an α-helix preferrer. There can be polar groups, or polar heteroatoms in the structures of α-helix preferrers, but they are always further away than the Cγ atom.

All aromatic and all amino acids with branching on the Cβ atom are strand preferrers. It was shown previously that

the dominant cause for high preference of aromatic and Cβ-branching amino acids for strands is a consequence of the avoidance of steric clashes between an amino acid side chain and its local backbone [23].

All turn and bend preferrers have polar heteroatoms on Cβ or Cγ atoms. Proline is also among turn and bend preferrers. It has an unusual structure, but it also has a polar heteroatom, N, on Cβ. The example of Thr is very interesting. It has both branching on Cβ and a polar heteroatom on Cβ. Hence, it has the structural properties of both strand preferrers and turns and bend preferrers. At the same time, Thr has indeed almost the same correlation coefficient for strands and coils. This shows clearly that these structural properties are closely connected with secondary structure. Hence, it seems that branching on the Cβ atom is more important for preference for strands than the nonpolarity of the amino acid. Again, this is in accord with the finding that steric clashing is dominant for strand preferrers [23].

Histidine and Cys, which show no preference for any particular secondary structure, also have quite different structures from all other amino acids. Histidine has a polar heteroatom on Cβ, but it is included in the aromatic ring. Because of this polar heteroatom on Cβ, His shares some similarities with turn and bend preferrers. Cysteine has an SH group on the Cβ atom, which is not very polar. Hence, it differs from turn and bend preferrers. Another difference is that SH groups can make disulfide bridges, making this amino acid quite different from the others, and explaining why it does not belong to any of the previous groups.

Based on these observations, it is clear that the tendency of an amino acid to take part in a certain secondary structure type is not defined solely by amino acid polarity or hydrophobicity (although there is some relationship), but that the crucial property of an amino acid is the substituent (s) on the Cβ or Cγ atoms. Thus, substituents closest to the backbone determine the type of secondary structure, while the rest of the side chain is less important. This is in agreement with the observation that propensities are closely connected with the energetics of peptide hydrogen bonds [32], and is evocative of the backbone-based theory of protein folding [31].

A similar classification and its relation to chemical structure were introduced by Levitt [18]. However, there are significant differences in our classifications of Arg, Cys and His, resulting in a more explicit connection of our classification to amino acid chemical structure. As shown above, our data indicate that Arg is an α-helix preferrer, while Cys and His are weakly correlated to any secondary structure. Levitt found that Cys and His prefer α-helices, while Arg has no preference for α-helices. He concluded: (1) amino acids with a bulky side chain, branched on Cβ or with aromatic groups, favour strands, (2) amino acids with short polar side chains or with special side chains (Gly and Pro) prefer turns, (3) all other amino acids prefer α-helix, with the exception of Arg.

Levitt's conclusions about the structure of amino acids that prefer strands and turns are in agreement with our conclusions. However, there are significant differences regarding the amino acids that are α-helix preferrers. While Levitt has no clear structural rules for those amino acids, we concluded that α-helix preferrers have no polar heteroatoms on Cβ and Cγ atoms, and neither branching nor aromatic groups on the Cβ atom. In this way, we clearly excluded Cys and His from α-helix preferrers. Note that Levitt concluded that, based on its chemical structure, Arg would be expected to be an α-helix preferrer, while his data resulted in a different classification of Arg.

To clarify if the differences between our results and those of Levitt are consequences of (1) different statistical methods, (2) different secondary structure assignments, or (3) different data sets, we applied our method to both the entire original Levitt data set (both the sequences and secondary structures published in Levitt [43]) and to contemporary protein structure data for Levitt's proteins, with secondary structures assigned by DSSP. Very similar correlation tables were obtained, leading to the conclusion that potential differences in secondary structure assignments are not substantial. In the following we will discuss only the analysis of the original Levitt data set.

Our results on Levitt's data differ from his in two important aspects: (1) Cys is not strongly correlated to any secondary structure, and (2) we found no significant correlation of Arg, Gln, Trp and Cys to any secondary structure type, while Phe has only a negative correlation with turns. Therefore, the application of our statistical approach allows for stricter conclusions, without misjudgment on the amino acids' preferences.

It is clear that the differences between our results and those of Levitt are the consequences both of different methodology and the availability of more representative data today. However, it is very impressive that, using data for only 66 proteins, Levitt was able to get results in such good agreement with our results, which were obtained using a set of 1,737 proteins.

Summary

The calculated correlations of amino acids with secondary structure types from a larger data set enable us to determine more clearly an amino acid's tendency to participate in a particular secondary structure type. The results show that most amino acids (except His and Cys) have a clear preference to participate in one particular secondary structure type. Based on these preferences, amino acids

can be classified in four groups: α-helix preferrers (Ala, Leu, Glu, Gln, Arg, Met, Lys), strand preferrers (Val, Ile, Tyr, Phe, Thr, Trp), turn and bend preferrers (Gly, Asn, Pro, Asp, Ser), and others (Cys, His). There are several important differences in our conclusions with respect to earlier studies that allow us to better understand how various substituents control these preferences. These amino acid preferences are caused by structural properties at the Cβ or Cγ atoms, while the rest of the side chain is less important. In some way, this is evocative of the backbone-based theory of protein folding [31]. We can specify the rules that classify amino acids as preferrers of certain secondary structures based only on the structural properties of the Cβ or Cγ atoms. The common structural properties of all α-helix preferrers are: no polar atoms on Cβ and Cγ atoms, no branching on Cβ, and an aliphatic (sp^3) Cγ atom. All strand preferrers have aromatic groups or branching on the Cβ atom, while all turn and bend preferrers have a polar heteroatom on Cβ or Cγ atoms, or do not have a Cβ atom. Following these rules, based only on structure, it is possible to determine for which secondary structure type an amino acid will have a preference. Since these rules are based only on the structure of the amino acid, they could help in predicting preferences of non-natural amino acids. The results indicate that polarity, charge and capability for hydrogen bonding do not have a crucial influence on the preference for particular secondary structure types.

Acknowledgements This work was supported under projects, No 142037 and No 144030 by the Ministry of Science of the Republic of Serbia. M.B.H. acknowledges the support of the National Science Foundation, USA (CHE-0518074).

References

1. Bowie JU, Luthy R, Eisenberg DA (1992) Science 253:164–170
2. Chen CC, Singh JP, Altman RB (1999) Bioinformatics 15:53–65
3. Eyrich VA, Standley DM, Felts AK, Friesner RA (1999) Proteins 35:41–57
4. Eyrich VA Standley DM, Friesner RA (1999) J Mol Biol 288:725–742
5. Fischer D, Eisenberg D (1996) Protein Sci 5:947–955
6. Kelley LA, MacCallum RM, Sternberg MJE (2000) J Mol Biol 299:499–520
7. Koretke KK, Luthey-Schulten L, Wolynes PG (1998) Proc Natl Acad Sci USA 95:2932–2937
8. Levitt M, Warshel A (1975) Nature 253:694–698
9. Lomize AL, Pogozheva ID, Mosberg HI (1999) Proteins Suppl 3:199–203
10. Maiorov VN, Crippen GM (1992) J Mol Biol 227:876–888
11. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowsky B, Skolnick J (1999) Proteins Suppl 3:177–185
12. Rost B (1998) Protein structure prediction in 1D 2D and 3D. In: von Rague-Schleyer P et al (eds) Encyclopedia of computational chemistry. Wiley, Sussex, pp 2242–2255
13. Samudrala R, Xia Y, Huang E, Levitt M (1000) Proteins Suppl 3:194–198
14. Samudrala R, Huang E, Koehl P, Levitt M (2000) Protein Eng 13:453–457
15. Solis AD, Rackovsky S (2004) Polymer 45:525–546
16. Chou PY, Fasman GD (1974) Biochemistry 13:222–245
17. Chou PY, Fasman GD (1978) Adv Enzymol Relat Areas Mol Biol 47:45–148
18. Levitt M (1978) Biochemistry 17:4277–4285
19. Kim CA, Berg JM (1990) Nature 362:267–270
20. Minor DL, Kim PS (1994) Nature 367:660–663
21. O'Neil KT, DeGrado WF (1990) Science 250:646–651
22. Padmanabhan S, Marqusee S, Ridgeway T, Laue TM, Baldwin RL (1990) Nature 344:268–270
23. Street AG, Mayo SL (1999) Proc Natl Acad Sci USA 96:9074–9076
24. Penel S, Hughes E, Doig AJ (1999) J Mol Biol 287:127–143
25. Petukhov M, Muñoz V, Yumoto N, Yoshikawa S, Serrano L (1998) J Mol Biol 278:279–289
26. Engel DE, DeGrado WF (2004) J Mol Biol 337(5):1195–1205
27. Mandel-Gutfreund Y, Gregoret LM (2002) J Mol Biol 323 (3):453–61
28. Fitzkee NC, Fleming PJ, Gong H, Panasik N Jr, Street TO, Rose GD (2005) Trends Biochem Sci 30:73–80
29. Gong H, Fleming PJ, Rose GD (2005) Proc Natl Acad Sci USA 102(45):16227–16232
30. Fleming PJ, Gong HP, Rose GD (2006) Prot Sci 15(8):1829–1834
31. Rose GD, Fleming PJ, Banavar JR, Maritan A (2006) Proc Natl Acad Sci USA 103(45):16623–16633
32. Baldwin RL (2007) J Mol Biol 371:283–301
33. Chou PY, Fasman GD (1974) Biochemistry 13(2):211–222
34. Robson B (1974) Biochem J 141(3):853–867
35. Kabsch W, Sander C (1983) Biopolymers 22(12):2577–2637
36. Rost B (2001) J Struc Biol 134:204–218
37. Kloczkowski A, Ting KL, Jernigan RL, Garnier J (2002) Proteins 49:154–166
38. Samuels ML, Witmer JA (2003) Statistics for the life sciences, 3rd edn. Pearson, New Jersey
39. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28 (1):235–242
40. Hobohm U, Sander C (1994) Protein Sci 3:522–524
41. Gibrat JF, Garnier J, Robson B (1987) J Mol Biol 198:425–443
42. Bastolla U, Moya A, Viguera E, van Ham RCHJ (2004) J Mol Biol 343:1451–1466
43. Levitt M, Greer J (1977) J Mol Biol 114:181–239