Exceptional pairs of amino acid neighbors in α-helices

Bahram Goliaei*, Zarrin Minuchehr

Laboratory of Biophysics and Molecular Biology, Institute of Biochemistry and Biophysics, University of Tehran, P.O. Box 13145-1384 Tehran, Iran

Received 13 January 2003; revised 27 January 2003; accepted 27 January 2003

First published online 6 February 2003

Edited by Gunnar von Heijne

Abstract **Amino acids seem to have specific preferences for various locations in α-helices. These specific preferences, called singlet local propensity (SLP), have been determined by calculating the preference of occurrence of each amino acid in different positions of the α-helix. We have studied the occurrence of amino acids, single or pairs, in different positions, singlet or doublet, of α-helices in a database of 343 non-homologous proteins representing a unique superfamily from the SCOP database with a resolution better than 2.5 Å from the Protein Data Bank. The preference of single amino acids for various locations of the helix was shown by the relative entropy of each amino acid with respect to the background. Based on the total relative entropy of all amino acids occurring in a single position, the N_{cap} position was found to be the most selective position in the α-helix. A rigorous statistical analysis of amino acid pair occurrences showed that there are exceptional pairs for which, the observed frequency of occurrence in various doublet positions of the α-helix is significantly different from the expected frequency of occurrence in that position. The doublet local propensity (DLP) was defined as the preference of occurrences of amino acid pairs in different doublet positions of the α-helix. For most amino acid pairs, the observed DLP (DLP_O) was nearly equal to the expected DLP (DLP_E), which is the product of the related SLPs. However, for exceptional pairs of amino acids identified above, the DLP_O and DLP_E values were significantly different. Based on the relative values of DLP_O and DLP_E, exceptional amino acid pairs were divided into two categories. Those, for which the DLP_O values are higher than DLP_E, should have a strong tendency to pair together in the specified position. For those pairs which the DLP_O values are less than DLP_E, there exists a hindrance in neighboring of the two amino acids in that specific position of the α-helix. These cases have been identified and listed in various tables in this paper. The amount of mutual information carried by the exceptional pairs of amino acids was significantly higher than the average mutual information carried by other amino acid pairs. The average mutual information conveyed by amino acid pairs in each doublet position was found to be very small but non-zero.**

Key words: Relative entropy; Mutual information; Sequence analysis; Amino acid pair; Doublet position; α-Helix

1. Introduction

The α-helix is a major structural element in protein architecture. It constitutes the most frequent element of secondary structure in globular proteins, and almost one quarter of all residues are found in helices [1]. A residue in the α-helix is defined either by the values of torsion angles, hydrogen bonds with (*i*) and (*i*+4) residues or the location of the carbon alpha. The first four and last four residues of helix are unique in their lack of one hydrogen bond [2]; therefore these two sides have been carefully analyzed for their effects on the stability of the α-helix. It has been found that the presence of the negative charge of the N_{cap} adds some stabilization energy because of the interaction with the macroscopic dipole of the helix [3,4]. It was also shown that aspartate at the amino end and arginine at the carboxyl end are excellent helix stabilizers [5]. A series of short helix models also showed that specific side-chain interactions with the main-chain stabilized isolated α-helical structures [3–9]. Early studies showed that some amino acids such as Glu, Ala and Leu are strong α-helix formers and Met, Val and Ile are strong β-sheet formers [10–13]. These works were extended for secondary structures in proteins such as α-helix and were followed by propensity studies for different residues at helical positions. In a recent study, capping motifs were identified using propensity for different helical positions [14]. It has been shown that Ser, Asp, Thr, Asn, Gly and Pro are preferred N_{cap} residues while Val, Ile, Phe, Ala, Lys, Arg, Glu, Met and Gln are being strongly avoided. It was also shown that X-Pro with Pro at C′ position has been a common structural motif in helix C-terminal [15]. The first turn of the α-helix was also analyzed and good N_2 amino acids such as Gln, Glu, Asp, Asn, Ser, Thr and His were found to hydrogen bond to the back bone of the helix [16]. Recent studies have also shown that different helical positions such as N_1 and N_2 have special effect on α-helix stability [17–19].

As mentioned above, NH donors of the first four residues and CO acceptors of the last four residues lack inter-helical hydrogen bond partners. Presta and Rose hypothesized that a necessary condition for helix formation is the presence of residues flanking the helix termini that have side chains to supply hydrogen-bond partners for unpaired main-chain NH and CO groups of the helix [2].

The Protein Data Bank [20,21] now contains a large enough number of solved protein structures to provide a statistically meaningful analysis of paired amino acid preferences for various locations of secondary structure elements. In this study, from the non-homologous subset of PDB we selected 343 proteins. Each protein represented a unique superfamily from the SCOP [22,23] database. The database contained 2177 α-helices. We analyzed the frequency of occurrences, the relative entropies, local propensities, and mutual information of single and pairs of amino acids for the two terminals region of α-helices.

*Corresponding author. Fax: (98)-21-6404680.
E-mail address: goliaei@ibb.ut.ac.ir (B. Goliaei).

0014-5793/03/$22.00 © 2003 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00105-4

2. Materials and methods

2.1. Database

We used the May-1999 list of non-homologous (sequence identity ≤25%) protein chains compiled by Hobohm and Sander [24,25] to select a subset of globular proteins for which the crystallographic structure has been solved to a resolution of 2.5 Å or better. Based on the version 1.59 of the SCOP database, we selected proteins which represented a unique superfamily in the SCOP database. The final result was a database of 343 proteins. We selected α-helices that contained at least seven amino acids. The resulting database contained a total of 2177 α-helices with unique sequences and 36727 amino acids.

2.2. Helix position nomenclature

For each α-helix we have considered the following doublet positions: $X_1 = N'N_{cap}$, $X_2 = N_{cap}N^1$, $X_3 = N^1N^2$, $X_4 = N^2N^3$, $X_5 = N^3N^4$, $X_6 = C^4C^3$, $X_7 = C^3C^2$, $X_8 = C^2C^1$, $X_9 = C^1C_{cap}$, $X_{10} = C_{cap}C'$. We call each doublet position X_i, $i = 1$–10, for future calculations. We call each single location of the helix N_i.

2.3. Helix ends criteria

Helix termini were assigned based on the assignments in the Definition of Secondary Structure of Protein [1] without further modification. N_{cap} and C_{cap} are the residues with non-helical Φ and ψ values immediately preceding and following N-terminus and C-terminus of an α-helix respectively [26].

2.4. Statistical analysis

2.4.1. Test of independence. We used the maximum likelihood (ML) statistic to test the Null hypothesis that the occurrence of two amino acids in a doublet position have occurred independently. i.e. $p(a_ia_j) = p(a_i)p(a_j)$. The ML statistic has, asymptotically, a χ^2 distribution [27] and is given by:

$$\chi^2_{\mathrm{ML}}(ab, X_i) = 2\sum O(ab, X_i)\ln\frac{O(ab, X_i)}{E(ab, X_i)} \quad (1)$$

where $O(ab,X_i)$ and $E(ab,X_i)$ represent the observed and expected frequencies of the amino acid pair ab in the doublet position X_i respectively. If X_i is the doublet which comprises positions N_{i-1} and N_i, then it is always assumed that amino acid a is occurring in position N_{i-1} and amino acid b is occurring in position N_i. The contingency table was made of all possible combinations of (a,b), $(a,-b(not\ b))$, $(-a(not\ a),b)$, and $(-a,-b)$. $E(ab,X_i)$ is defined as:

$$E(ab, X_i) = p(a^{N_{i-1}})p(b^{N_i})n(X_i) \quad (2)$$

$p(a^{N_{i-1}})$ and $p(b^{N_i})$ represent the probability of occurrences of amino acids a and b in the positions N_{i-1} and N_i respectively. $n(X_i)$ represents the number of available doublet seats in the doublet position X_i of the α-helix.

We rejected the Null hypothesis of independence at a significance level of 0.01.

2.4.2. Relative entropy. Relative entropy of an amino acid a_i at a given position N_j of the α-helix with respect to the background is given by:

$$D(p(a_i^{N_j})||q(a_i)) = p(a_i^{N_j})\log\frac{p(a_i^{N_j})}{q(a_i)} \quad (3)$$

where $p(a_i^{N_j})$ represents the probability of occurrence of amino acid a_i in the position N_j of the helix and $q(a_i)$ represent the background probability of occurrence of the amino acid a_i in the database of α-helices.

We defined the relative entropy for the location N_j of an α-helix as the sum of the relative entropies of all amino acids occurring in that position, given by:

$$D(p(a^{N_j})||q(a)) = \sum_i p(a_i^{N_j})\log\frac{p(a_i^{N_j})}{q(a_i)} \quad (4)$$

2.4.3. Mutual information. The mutual information carried by a pair of amino acid (a,b) occurring in the doublet position X_i of the helix is given by:

$$\mathrm{MI}(a,b) = \sum p(a,b)\log_2\left(\frac{p(a,b)}{p(a)p(b)}\right) \quad (5)$$

The summation extends over all possible combination of (a,b), $(a,-b)$, $(-a,b)$, and $(-a,-b)$.

2.4.4. Local propensities. $\mathrm{SLP}(a_i^{N_j})$, singlet local propensity for an amino acid a_i, to occupy the position N_j of the helix is defined as:

$$\mathrm{SLP}(a_i^{N_j}) = \frac{p(a_i^{N_j})}{p(a_i^{\mathrm{helix}})} = \frac{n(a_i^{N_j})/n(N_j)}{n(a_i^{\mathrm{helix}})/n(\mathrm{helix})} \quad (6)$$

$p(a_i^{\mathrm{helix}})$, represent the probability of that amino acid to occur any where in the helix.

$n(a_i^{\mathrm{helix}})$, represents the total occurrences of amino acid in the helix database.

n(helix), represents total number of amino acids in the helix database.

$\mathrm{DLP_O}(a_ja_k,X_i)$, the observed doublet local propensity for the pair of amino acids a_ja_k to occupy the doublet position X_i of the helix is defined as:

$$\mathrm{DLP_O}(a_ja_k, X_i) = \frac{p(a_ja_k, X_i)}{p(a_ja_k, \mathrm{helix})} = \frac{n(a_ja_k, X_i)/n(X_i)}{n(a_ja_k, \mathrm{helix})/n(\mathrm{doublets})} \quad (7)$$

$p(a_ja_k,X_i)$ represents the probability of the pair of amino acids a_ja_k to occur in the doublet position X_i of the helix.

$p(a_ja_k$,helix) represents the probability of the pair of amino acids a_ja_k to occur in any doublet position in the helix.

$n(a_ja_k,X_i)$ is the occurrence of the pair of amino acids a_ja_k in the doublet position X_i.

$n(X_i)$ is the total number of amino acid pairs in the doublet position X_i.

$n(a_ja_k$,helix) is the occurrence of the pair of amino acids a_ja_k in the helix database.

n(doublets) represents the total number of doublets in the helix database. This quantity is equal to the number of single positions of the helices, $n(N_i)$, minus the number of helices in the database.

$\mathrm{DLP_E}(a_ja_k,x_i)$, the expected doublet local propensity for the pair of amino acid a_ja_k to occupy the doublet position X_i of the helix, can be shown to be the product of the SLP of amino acids a_j and a_k:

$$\mathrm{DLP_E}(a_ja_k, X_i) = \mathrm{SLP}(a_j^{N_i}) \times \mathrm{SLP}(a_k^{N_{i+1}}) \quad (8)$$

A totally random distribution of amino acids in the helix would result in a local propensity (LP) of 1 in any position (doublet or single). Therefore, an LP value of larger than 1 would indicate preference of the amino acid for that location and an LP value of smaller than 1 would indicate avoidance of the amino acid for that position.

3. Results

3.1. Relative entropies

Relative entropies of all amino acids in various locations of

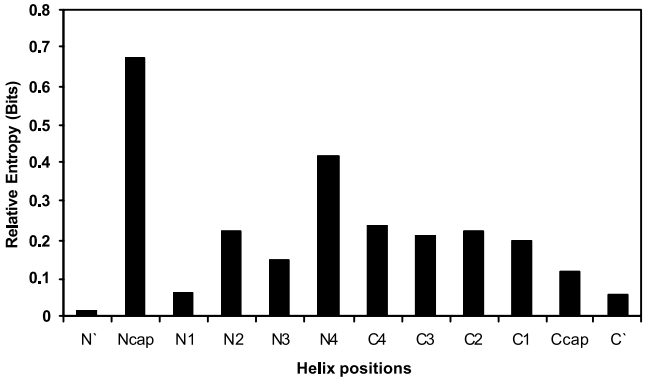

Fig. 1. The relative entropy of each single location on the N-terminal and C-terminal of the α-helix in Bits units.

Table 1
The occurrence and the relative entropy of amino acids in different positions of the α-helix

N′		N_{cap}		N_1		N_2		N_3		N_4		C_4		C_3		C_2		C_1		C_{cap}		C′	
Pro *134*	0.08	Ser *345*	0.25	Pro 264	0.26	Glu *333*	0.14	Glu *285*	0.09	Val *217*	0.1	Leu *287*	0.07	Leu *327*	0.07	Lys *229*	0.07	Ala 280	0.03	Gly *396*	0.31	Gly *394*	0.31
Gly *166*	0.03	Asp 334	0.22	Glu 205	0.02	Asp *214*	0.08	Asp *189*	0.05	Ile *181*	0.07	Ala 258	0.04	Ala *303*	0.05	Glu *230*	0.04	Leu 270	0.03	Asn *141*	0.05	Pro *210*	0.18
Thr 125	0.02	Thr *251*	0.15	Val 148	0.01	Pro *117*	0.05	Gln *156*	0.04	Leu *283*	0.06	Ile 136	0.03	Met *92*	0.03	Arg *160*	0.03	Thr 130	0.02	His 73	0.02	Asn 110	0.02
Ser 125	0.01	Asn *210*	0.12	Lys 37	0	Ser 147	0.02	Thr 121	0.01	Phe *121*	0.04	Phe 102	0.02	Lys 165	0.02	Leu 261	0.02	Lys 169	0.02	Ser 129	0.01	Lys 152	0.01
Asp 131	0.01	Pro *120*	0.06	Phe 71	0	Gln 115	0.01	His 61	0.01	Arg 138	0.02	Tyr 83	0.02	Arg 134	0.01	Gln 116	0.01	Ser 136	0.01	Pro 1	0	Arg 117	0
Val 146	0.01	Gly *196*	0.06	Tyr 19	0	Lys 149	0.01	Phe 99	0.01	Ala 239	0.02	Trp 46	0.02	Asn 80	0	Ala 248	0.01	Asn 101	0.01	Gln 108	0	His 45	0
Ile 131	0.01	His 54	0	Trp 137	0	Ala 242	0.01	Pro 55	0	Tyr 85	0.02	Arg 122	0.01	Gln 102	0	Ile 125	0.01	Glu 194	0.01	Lys 148	0	Cys 13	0
Phe 103	0.01	Cys 22	0	Cys 79	0	Trp 41	0.01	Ser 118	0	Met 59	0.01	Val 129	0.01	Glu 166	0	Thr 103	0	Arg 141	0.01	Phe 80	0	Ser 106	−0.01
Tyr 80	0.01	Tyr 49	−0.01	Gly 97	−0.01	His 54	0	Val 133	0	Trp 40	0.01	Met 68	0.01	His 42	0	His 53	0	His 64	0.01	Tyr 66	0	Thr 84	−0.01
Met 76	0.01	Met 25	−0.01	Thr 38	−0.01	Cys 12	0	Trp 28	0	Pro 0	0	Cys 31	0.01	Ile 116	0	Val 122	0	Phe 105	0.01	Cys 24	0	Gln 90	−0.01
Asn 88	0	Trp 14	−0.01	Gln 114	−0.01	Gly 111	−0.01	Cys 16	0	Gln 92	0	Gln 92	0	Phe 88	0	Met 54	0	Tyr 81	0.01	Thr 80	−0.01	Asp 111	−0.01
His 45	0	Phe 45	−0.02	Asp 100	−0.01	Thr 90	−0.01	Tyr 54	−0.01	Cys 22	0	Pro 7	−0.01	Tyr 74	0	Trp 26	0	Pro 0	0	Asp 103	−0.01	Phe 61	−0.01
Trp 28	0	Gln 45	−0.03	Arg 86	−0.01	Asn 75	−0.01	Met 49	−0.01	Thr 80	−0.01	Thr 78	−0.01	Trp 35	0	Cys 23	0	Gln 98	0	Arg 104	−0.01	Tyr 50	−0.01
Cys 19	0	Glu 107	−0.03	His 215	−0.01	Arg 109	−0.01	Gly 94	−0.02	His 25	−0.01	Glu 139	−0.01	Cys 26	0	Pro 3	−0.01	Met 50	0	Ala 214	−0.01	Met37	−0.01
Lys 123	−0.01	Lys 74	−0.03	Ala 103	−0.01	Phe 58	−0.01	Asn 53	−0.02	Asn 41	−0.02	Lys 112	−0.01	Pro 6	−0.01	Ser 108	−0.01	Trp 26	0	Met 45	−0.01	Trp 14	−0.01
Leu 217	−0.01	Arg 59	−0.03	Ile 34	−0.01	Tyr 53	−0.01	Lys 101	−0.02	Lys 102	−0.02	His 34	−0.02	Ser 106	−0.01	Asn 71	−0.01	Cys 23	0	Trp 23	−0.01	Glu 132	−0.02
Gln 59	−0.02	Val 42	−0.03	Met 102	−0.01	Met 24	−0.01	Arg 93	−0.01	Arg 61	−0.02	Gly 61	−0.03	Ser 75	−0.02	Val 117	−0.01	Phe 66	−0.01	Ile 103	−0.01	Leu 199	−0.02
Arg 88	−0.02	Ile 35	−0.03	Ser 88	−0.02	Val 82	−0.02	Ala 205	−0.02	Ser 51	−0.02	Asn 47	−0.03	Asn 47	−0.02	Thr 67	−0.02	Tyr 61	−0.02	Asp 77	−0.01	Val 79	−0.02
Glu 107	−0.03	Ala 70	−0.06	Asn 49	−0.02	Ile 60	−0.03	Ile 89	−0.02	Asp 41	−0.03	Gly 52	−0.03	Asp 82	−0.02	Asp 75	−0.02	Val 82	−0.02	Val 60	−0.03	Glu 122	−0.03
Ala 160	−0.04	Leu 80	−0.06	Leu 191	−0.02	Leu 91	−0.06	Leu 178	−0.03	Glu 80	−0.04	Asp 60	−0.03	Gly 49	−0.03	Gly 43	−0.03	Gly 47	−0.03	Ile 61	−0.03	Ile 73	−0.02
																						Ala 134	−0.05
																						Leu 135	−0.05

For each position of the α-helix termini, the occurrence in the database, of all 20 amino acids occurring in that position, is given right in front of the name of that amino acid. The relative entropy (in Bits units) of the amino acid for that location is given next to the number of occurrences of that amino acid. For each position of the helix, the data is sorted according to the relative entropy value of amino acids in that position.

the N-terminal and C-terminal of the α-helix were calculated and are shown in Table 1. For each location, the data for all 20 amino acids has been sorted descending according to the relative entropy of each amino acid in that location. The relative entropy has been calculated with respect to the background. Also, the frequency of occurrence for each amino acid is given in the tables. Therefore, for example, in the N_1 position, proline has the highest relative entropy value with respect to other amino acids occurring in the N_1 location. Relative entropy, when calculated with respect to the background, is a convenient measure of the preference of an amino acid to occur in a specific position in the helix with respect to its general occurrence anywhere in the helix. For those amino acids which do not have any preference for a specific position, the relative entropy in that location is zero. Positive values indicate preference for the location, while negative values indicate reluctance for that location.

We have defined the relative entropy for a given location of the α-helix as the sum of the relative entropies of all amino acids occurring in that location. Fig. 1 shows the relative entropies of all singlet positions in the N-terminal and C-terminal sides of the α-helix. N_{cap} has the highest value of relative entropy and thus, it is the most selective site in the α-helix. N′ had the least value of the relative entropy and therefore, was the least selective site in the α-helix.

3.2. Exceptional pairs of amino acids

There are 400 possible pairs of amino acids which can occur in any doublet position in the α-helix. The frequency of occurrence of each pair of amino acids *ab* in a given doublet position, X_i, was determined and saved as $O(ab,X_i)$. The expected frequency of occurrence of the same pair of amino acids in the position X_i, $E(ab,X_i)$, was calculated as defined by the Eq. 2. Subsequently, for the occurrence of each possible pairs of amino acids in every doublet position, the χ^2_{ML} value was calculated. Those amino acid pairs for which the Null hypothesis of independence could be rejected significantly were identified for each doublet locations. These are exceptional pairs of amino acids which their frequencies of occurrences were significantly either higher or lower than the expected values. There were 125 such pairs among 4000 possible pairs of amino acids occurring in all 10 doublet positions of the α-helix terminals.

3.3. DLP values

We used the DLP_O and DLP_E as a convenient measure to define the tendency or hindrance for two amino acids to neighbor as an exceptional pair in a doublet position in the α-helix. If $DLP_O > DLP_E$, it implies the existence of a tendency between the two amino acids for neighboring in a given doublet position. On the other hand, if $DLP_O < DLP_E$, it means that there is a hindrance for the neighboring of the two amino acids in a given doublet position. We defined the 'excess DLP (EDLP)' as $(DLP_O - DLP_E)/DLP_E$ as a convenient measure of the difference of DLP_O and DLP_E. Positive values of EDLP indicate a tendency of the two amino acids to neighbor in a given doublet position X_i; negative values indicate hindrance and zero indicate no preference. We classified the exceptional amino acid pairs described above, on the basis of their EDLP values into two groups. Tables 2a and 2b show the exceptional amino acid pairs which have a tendency for neighboring in each doublet position in the α-helix N-terminal

Table 2a
The distribution of exceptional amino acid pairs with tendency for neighboring at N-terminal sites of the α-helix

$N'–N_{cap}$		$N_{cap}–N_1$		$N_1–N_2$		$N_2–N_3$		$N_3–N_4$	
AA	EDLP MI	AA	EDLP MI	AA	EDLP MI	AA	EDLP MI	AA	EDLP MI
Gln–Tyr	3.01 0.0032	Gly–Gly	1.89 0.0054	Cys–Ser	4.42 0.0035	Ile–Pro	4.93 0.0078	Trp–Asn	7.08 0.0052
Lys–Gln	2.20 0.0028	Ser–Gln	0.45 0.0036	His–Ile	3.34 0.0028	Phe–Pro	2.45 0.0040	His–Leu	1.04 0.0044
His–Ile	4.03 0.0026	Trp–Arg	6.01 0.0029	Gly–Arg	2.37 0.0028	Gly–Cys	3.79 0.0037		
				Gln–Gln	0.79 0.0025	Cys–His	6.46 0.0028		
				Thr–Gln	1.44 0.0025				
				Ile–Pro	1.01 0.0023				
				Gly–Leu	1.90 0.0022				

For each pair the EDLP value and the mutual information in Bits are given. For all these pairs the observed frequency is significantly larger than the expected frequency.

and C-terminal respectively. The $DLP_O = 1$ indicates no preference for the given doublet position. Therefore, we omitted from Tables 2a and 2b any amino acid pair which had a DLP_O value of less than 1 (even though the difference between the DLP_O and the related DLP_E value was significant according to the χ^2 test of independence). We further filtered the results and accepted only those pairs which had $EDLP > 1$.

Tables 3a and 3b show the exceptional amino acid pairs which, there is a hindrance for their neighboring the specified doublet positions in N- and C-terminals of the α-helix. We filtered the results and accepted only those pairs which had $DLP_E > 1$.

3.4. Mutual information

The mutual information conveyed by each amino acid pairs in all possible doublet positions was calculated as described in Section 2. The average mutual information carried by an amino acid pair was found to be 0.000567 Bits. This value shows the background information or the noise in the system. The mutual information carried by the exceptional pairs was an order of magnitude larger than the background noise. Therefore, we disregarded this noise and provided the calculated values of the mutual information of exceptional pairs of amino acids in the Tables 2a,b and 3a,b.

It is possible to calculate the average mutual information contained in helix doublet positions by summing the mutual information carried by all amino acid pairs occurring in a given doublet position of the α-helix. This quantity was calculated from the mutual information values obtained above and is plotted in the Fig. 2 for all possible doublet positions of the α-helix termini.

Table 2b
The distribution of exceptional amino acid pairs with tendency for neighboring at C-terminal sites of the α-helix

$C_4–C_3$		$C_3–C_2$		$C_2–C_1$		$C_1–C_{cap}$		$C_{cap}–C'$	
AA	EDLP MI	AA	EDLP MI	AA	EDLP MI	AA	EDLP MI	AA	EDLP MI
Pro–Leu	6.42 0.0053	Thr–Leu	0.89 0.0038	Lys–His	1.74 0.0053	Thr–Gly	0.58 0.0071	Ala–Gly	0.74 0.0149
Trp–Phe	2.01 0.0044	Leu–Arg	0.26 0.0030	Glu–Lys	0.32 0.0038	Leu–Gly	0.19 0.0062	Gly–Ile	1.16 0.0106
Tyr–Thr	1.76 0.0029	Pro–Gly	27.88 0.0028	Pro–Met	33.13 0.0038	Ala–Ala	0.40 0.0055	Gly–Tyr	1.16 0.0069
Ile–Asn	1.24 0.0027	Ser–Ile	0.86 0.0023	Met–Ile	2.62 0.0037	Lys–Tyr	1.31 0.0049	Phe–Pro	0.96 0.0063
Thr–Leu	0.37 0.0027			Ser–Gln	1.14 0.0036	Tyr–Gly	0.71 0.0037	Cys–Ala	4.39 0.0055
Glu–Arg	0.76 0.0027			Thr–Ile	1.43 0.0028	Glu–Arg	0.82 0.0031	Gly–Val	0.83 0.0039
Leu–Asn	0.58 0.0026					Gly–Asp	2.24 0.0024	Gly–Phe	0.59 0.0031
								Gly–Leu	0.89 0.0024
								Ser–Gln	0.80 0.0023
								Thr–Asn	1.68 0.0023

For each pair the EDLP value and the mutual information in Bits are given. For all these pairs the observed frequency is significantly larger than the expected frequency.

Table 3a
The distribution of exceptional amino acid pairs with hindrance for neighboring at N-terminal sites of the α-helix

N'–N_{cap}		N_{cap}–N_1		N_1–N_2		N_2–N_3		N_3–N_4	
AA	EDLP MI	AA	EDLP MI	AA	EDLP MI	AA	EDLP MI	AA	EDLP MI
Arg–Asp	−0.78 0.0048	Asn–Gly	−0.81 0.0041	Pro–Pro	−0.90 0.0078	Lys–His	−1.00 0.0029	Glu–Glu	−0.82 0.0065
Leu–Glu	−0.72 0.0030	Gly–Pro	−0.30 0.0034	His–Ala	−1.00 0.0030	Glu–Pro	−0.65 0.0027	Gly–Ala	−0.61 0.0041
		Asp–Gly	−0.37 0.0032	Asn–Lys	−1.00 0.0023	Lys–Pro	−1.00 0.0026	Glu–Asp	−0.77 0.0039
		Glu–Ser	−1.00 0.003	Glu–Pro	−0.46 0.0022	Glu–Val	−0.47 0.0026	Phe–Tyr	−1.00 0.0037
		Arg–Lys	−1.00 0.0026			Val–Gly	−1.00 0.0024	Ala–Met	−0.85 0.0036
		His–Val	−1.00 0.0026			Lys–Ala	−0.60 0.0022	Asp–Gly	−0.70 0.0033
		Gly–Thr	−0.71 0.0022					Asp–Ala	−0.38 0.0032
								Pro–Val	−0.76 0.0029
								Val–Asn	−1.00 0.0027
								Gln–Ile	−0.41 0.0027
								Ser–Leu	−0.36 0.0025
								Leu–Ile	−0.16 0.0025
								Gln–Gly	−0.72 0.0025
								Lys–Lys	−0.79 0.0025

For each pair the EDLP value and the mutual information in Bits are given. For all these pairs the observed frequency is significantly less than the expected frequency.

4. Discussion

We have studied the frequency of occurrences of single and pairs of amino acids in various positions of the α-helical structures in globular proteins. Several reports have already indicated that amino acid preferences for the helix positions differ greatly between different helical positions [15,16,26,28,29].

Relative entropy (also known as Kullback-Leibler distance) is a measure of the distance between two probability distributions P and Q [30]. When Q represents a uniform background then the relative entropy will show the information content in the P distribution. In this case the relative entropy shows how far away is the P distribution from a totally random situation represented by the background. The concept has been used before to illustrate the conservative regions in DNA sequences [31]. Here, we have used it to illustrate how much the occurrence or the absence of an amino acid in a position of the helix is non-random. The results are in excellent agreement with the propensity data presented by other investigators [14,16,26,32]. Of special interest are hydrophobic amino acids such as Leu and Ala, which are highly disfavored from the N-terminal positions. The relative entropy provides a convenient measure of the preference of an amino acid for a helix location and since it includes negative numbers as well as positives, it is a more sensitive measure than propensity data which covers only positive numbers.

The concept of relative entropy can be expanded to define the relative entropy for each location of the α-helix. We defined the relative entropy of a location as the sum of the relative entropy of all amino acids occupying that location. Relative entropy of a location is an indication that how much that location is selective in accepting amino acids. Our results show that the N_{cap} position is the most selective site in the α-helix termini. This is perhaps due to the fact that amino acids occupying the N_{cap} positions are highly critical in terms of providing a side chain hydrogen bonding to the other vise unsatisfied back bone hydrogen bonds of N-terminal residues [2].

Based on results presented here and by many other investigators before, it is now highly accepted that the occurrence of amino acids in helix positions, especially at helix termini, is not completely random. Now we extended this question to a pair of amino acids which occupy a doublet position in the helix termini. If the occurrence of an amino acid in a given position is a totally random event and independent of its near neighbors, then one would expect that the probability of occurrence of a pair amino acid in a doublet position to be equal to the product of the probability of occurrence of each amino acid in the neighboring singlet position. The results we have presented here show that although for the majority of amino acid pairs the Null hypothesis of independent occurrence holds true, however, there are exceptional pairs for which the Null hypothesis can be rejected significantly. The existence of these exceptional pairs indicates that physico-chemical interactions between near neighbor amino acid in the helix termini are important factors which have to be taken into consideration.

The concept of 'propensity' has been used widely to indicate the preference of an amino acid to occupy a given position in

Table 3b
The distribution of exceptional amino acid pairs with hindrance for neighboring at C-terminal sites of the α-helix

C_4–C_3		C_3–C_2		C_2–C_1		C_1–C_{cap}		C_{cap}–C'	
AA	EDLP MI	AA	EDLP MI	AA	EDLP MI	AA	EDLP MI	AA	EDLP MI
Leu–Leu	−0.18 0.0045	Tyr–Ile	−1.00 0.0030	Ser–Asn	−1.00 0.0035	Glu–Gly	−0.33 0.0038	Gly–Gly	−0.67 0.0254
Glu–Leu	−0.36 0.0035	Gln–Glu	−0.69 0.0029	Gln–Thr	−0.85 0.0028	Tyr–Ala	−0.86 0.0035	Gly–Pro	−0.64 0.0138
Gly–Arg	−1.00 0.0033	Leu–Leu	−0.33 0.0029	Gly–Lys	−1.00 0.0023	Leu–Phe	−0.70 0.0035	Leu–Ile	−1.00 0.0049
Gln–Asp	−1.00 0.0033	Ile–Phe	−1.00 0.0024			Gly–Ala	−1.00 0.0033	Ala–Ala	−0.76 0.0035
Met–Ala	−0.51 0.0030					Arg–Gly	−0.32 0.0031	Glu–Thr	−1.00 0.0034
Phe–Ala	−0.32 0.0030					Lys–Gly	−0.39 0.0031	Leu–Val	−0.85 0.0032
Ala–Leu	−0.22 0.0028					Val–Gln	−1.00 0.0028	Ser–Pro	−0.72 0.0031
Glu–Glu	−0.57 0.0028							Asn–Phe	−1.00 0.0028
Arg–Val	−0.67 0.0027							Leu–Met	−1.00 0.0025
Cys–Leu	−0.60 0.0026							Tyr–Asp	−1.00 0.0024
Arg–Gly	−1.00 0.0026								
His–Ala	−0.35 0.0025								
Ala–Thr	−0.57 0.0025								

For each pair the EDLP value and the mutual information in Bits are given. For all these pairs the observed frequency is significantly less than the expected frequency.

the α-helix [26,32]. It is the ratio of the probability of an amino acid to occur in a given position in the α-helix to the probability of the same amino acid to occur either in the α-helix (for LP), or in the database of proteins (for global propensity). We have extended the concept of propensity originally devised for a single amino acid to a broader sense to express the preference of a pair of amino acids to occupy a given doublet position in the α-helix. Based on the relative values of the DLP_O and DLP_E we could divide the set of exceptional pairs of amino acids into two groups. The group for which the DLP_O is larger than DLP_E, $EDLP > 0$ (shown in the Tables 2a and 2b), represent the amino acids which have a strong tendency for neighboring in the given position.

On a similar ground, there should exist a hindrance for neighboring between amino acid pairs listed in Tables 3a and 3b. These are exceptional amino acid pairs which their DLP_O is less than the DLP_E value.

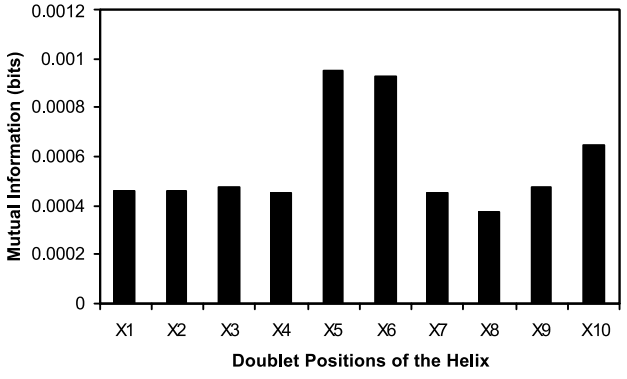

Fig. 2. The mutual information (in Bits units) contained in each doublet position of the α-helix terminals. The locations are $X_1 = N'N_{cap}$; $X_2 = N_{cap}N_1$; $X_3 = N_1N_2$; $X_4 = N_2N_3$; $X_5 = N_3N_4$; $X_6 = C_4C_3$; $X_7 = C_3C_2$; $X_8 = C_2C_1$; $X_9 = C_1C_{cap}$; $X_{10} = C_{cap}C'$.

Mutual information of the two random variables X and Y is a measure of independence of X from Y and vice versa. The concept has been used before to show the mutual information content in various types of pairing of amino acids in proteins [33]. The mutual information for totally independent variables is zero. In our problem of amino acid pair occurrence in helix doublet positions, it shows how much information the occurrence of a conveys regarding the occurrence of b in the doublet position X_i. On a similar ground, the mutual information of a doublet position is an average indicator of how much the amino acid pairs occupying this doublet position may occur independent of one another. The high mutual information at the C-terminal side of the α-helix shows that there should exist a helix signal in α-helix termination.

The exceptional pairs of amino acids provide exceptions and contradictions to currently known rules of helix stabilization. For instance, the N_{cap} residue is known to have torsion angles different form the α-helix but its capability of providing backbone hydrogen bond to i+3 residue is an important factor for helix stabilization. The presence of Gly–Gly pairs in N_{cap}–N_1 position provides an exception to this generally accepted statement. Another example is the presence of positively charged amino acids such as Lys and Arg in N-terminal locations and negatively charged residues at C-terminal locations. It is generally accepted that His and other positively charged residues at the C-terminal locations help to stabilize the α-helix through interaction with the helix dipole [34].

While energetic terms, such as hydrophobic interactions, steric clashes, conformational entropy, helix dipole interac-

tions, electric charge, and Van der Waals interactions are likely to be important in amino acid neighboring in various locations of the α-helix, the number of amino acid substitutions in both mitochondrial and nuclear genes, with amino acids having similar sets of neighbors replacing each other more frequently than those having very different sets of neighbors [35], may also be responsible for such neighboring effects in various positions of the α-helix. Obviously, further studies are required to clearly determine the role and the extent of contribution from each of these terms in positive cooperation or the hindrance observed in amino acid neighboring in α-helices. However, while the contribution of individual factors in amino acid neighboring is quite clear, it may be suggested that the difference between DLP_O and DLP_E value be considered as a quantitative measure of overall contributions to amino acid neighboring.

Acknowledgements: This work was supported by grant numbers 521/1/395, 521/1/450, and 521/1/551 from the Research Council of the University of Tehran. We hereby also thank all the scientists not mentioned but whose work helped to enlarge the PDB data bank used in this study.

References

[1] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[2] Presta, L.G. and Rose, G.D. (1988) Science 240, 1632–1641.
[3] Serrano, L. and Fersht, A.R. (1989) Nature 342, 296–299.
[4] Serrano, L., Sancho, J., Hirshberg, M. and Fersht, A.R. (1992) J. Mol. Biol. 20, 544–559.
[5] Forood, B., Feliciano, E.J. and Nambiar, K.P. (1993) Proc. Natl. Acad. Sci. USA 90, 838–842.
[6] Lyu, P.C., Wemmer, D.E., Zhou, H.X., Pinker, R.J. and Kallenbach, N.R. (1993) Biochemistry 32, 421–425.
[7] Chakrabartty, A., Doig, A.J. and Baldwin, R.L. (1993) Proc. Natl. Acad. Sci. USA 90, 11332–11336.
[8] Doig, A.J. and Baldwin, R.L. (1995) Protein Sci. 4, 1325–1336.
[9] Yumoto, N., Murase, S., Hattori, T., Yamamoto, H., Tatsu, Y. and Yoshikawa, S. (1993) Biochem. Biophys. Res. Commun. 196, 1490–1495.
[10] Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 222–245.
[11] Levitt, M. and Greer, J. (1977) J. Mol. Biol. 114, 181–239.
[12] Levitt, M. (1978) Biochemistry 17, 4277–4285.
[13] Williams, R.W., Chang, A., Juretic, D. and Loughran, S. (1987) Biochim. Biophys. Acta 916, 200–204.
[14] Aurora, R. and Rose, G.D. (1998) Protein Sci. 7, 21–38.
[15] Kumar, S. and Bansal, M. (1998) Proteins 31, 460–476.
[16] Penel, S., Hughes, E. and Doig, A.J. (1999) J. Mol. Biol. 287, 127–143.
[17] Cochran, D.A. and Doig, A.J. (2001) Protein Sci. 10, 1305–1311.
[18] Cochran, D.A., Penel, S. and Doig, A.J. (2001) Protein Sci. 10, 463–470.
[19] Doig, A.J., Andrew, C.D., Cochran, D.A., Hughes, E., Penel, S., Sun, J.K., Stapley, B.J., Clarke, D.T. and Jones, G.R. (2001) Biochem. Soc. Symp. 68, 95–110.
[20] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) Nucleic Acids Res. 28, 235–242.
[21] Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) Eur. J. Biochem. 80, 319–324.
[22] Lo, C.L., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2002) Nucleic Acids Res. 30, 264–267.
[23] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) J. Mol. Biol. 247, 536–540.
[24] Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) Protein Sci. 1, 409–417.
[25] Hobohm, U. and Sander, C. (1994) Protein Sci. 3, 522–524.
[26] Richardson, J.S. and Richardson, D.C. (1988) Science 240, 1648–1652.
[27] Williams, K. (1976) Statistician 25, 49.
[28] Argos, P. and Palau, J. (1982) Int. J. Pept. Protein Res. 19, 380–393.
[29] Penel, S., Morrison, R.G., Mortishire-Smith, R.J. and Doig, A.J. (1999) J. Mol. Biol. 293, 1211–1219.
[30] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1999) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge.
[31] Schneider, T.D. and Stephens, R.M. (1990) Nucleic Acids Res. 18, 6097–6100.
[32] Kumar, S. and Bansal, M. (1996) Biophys. J. 71, 1574–1586.
[33] Cline, M.S., Karplus, K., Lathrop, R.H., Smith, T.F., Rogers Jr., R.G. and Haussler, D. (2002) Proteins 49, 7–14.
[34] Scholtz, J.M. and Baldwin, R.L. (1992) Annu. Rev. Biophys. Biomol. Struct. 21, 95–118.
[35] Xia, X. and Xie, Z. (2002) Mol. Biol. Evol. 19, 58–67.