WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Amino acid pairing at the N- and C-termini of helical segments in proteins

Nuno A. Fonseca,[1] Rui Camacho,[2] and A. L. Magalhães[3]*

[1] IBMC and LIACC, R. Campo Alegre, 1021/1055, 4169-007 Porto, Portugal

[2] LIACC and FEUP, R. Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal

[3] REQUIMTE, Universidade do Porto, R. Campo Alegre, 687, 4169-007 Porto, Portugal

ABSTRACT

A systematic survey was carried out in an unbiased sample of 815 protein chains with a maximum of 20% homology selected from the Protein Data Bank, whose structures were solved at a resolution higher than 1.6 Å and with a R-factor lower than 25%. A set of 5556 subsequences with α-helix or 3_{10}-helix motifs was extracted from the protein chains considered. Global and local propensities were then calculated for all possible amino acid pairs of the type (i, i + 1), (i, i + 2), (i, i + 3), and (i, i + 4), starting at the relevant helical positions N1, N2, N3, C3, C2, C1, and N-int (interior positions), and also at the first nonhelical positions in both termini of the helices, namely, N-cap and C-cap. The statistical analysis of the propensity values has shown that pairing is significantly dependent on the type of the amino acids and on the position of the pair. A few sequences of three and four amino acids were selected and their high prevalence in helices is outlined in this work. The Glu-Lys-Tyr-Pro sequence shows a peculiar distribution in proteins, which may suggest a relevant structural role in α-helices when Pro is located at the C-cap position. A bioinformatics tool was developed, which updates automatically and periodically the results and makes them available in a web site.

Proteins 2008; 70:188–196.

Key words: protein helices; amino acid pairs; propensities; N-terminus; C-terminus.

INTRODUCTION

Helices are the most common secondary structural motif observed in folded proteins. To understand completely the helix formation and stability, the contributing factors have to be assessed thoroughly. One of those relevant factors is the formation of intramolecular hydrogen bondings between main-chain C═O and N—H groups, specially belonging to residues spaced *i*, *i* + 4 apart. As a consequence, the N-terminus has unsatisfied hydrogen-bond donors, whereas the C-terminus has unsatisfied hydrogen-bond acceptors. Another important characteristic is the presence of a net dipole moment along the helix axis with the N-terminus polarized positively and the C-terminus polarized negatively. These two characteristics are already enough to induce different amino acid occurrences in the helix. In fact, the different tendency of the amino acids to occur in α-helices is known for some decades.[1–4] Moreover, it is reasonable to accept that each individual helical position has its own role in helix stabilization and, thus, show distinct amino acid distributions. Factors such as side chain–main chain hydrogen bonds, solvent exposure, conformational entropy, and side chain–side chain interactions contribute to the diversity of positional preferences.

It has been shown that the helical occurrence of the 20 type of residues is highly dependent on the position, with a clear distinction between N-terminal, C-terminal, and interior positions.[5–16] It has also been remarked the importance of the first nonhelical positions at both termini, namely, C-cap and, specially, N-cap.[17–26]

More recent studies have shown that short sequences of residues may have an important role in protein folding and stability, namely, specific pairs found in parallel β-sheets,[27] loops[28] or inter-domain linkers,[29] helix-stabilizing *i*, *i* + 4 pairs,[30] and triplets of charged residues in helices showing a cooperative effect in their stabilization.[31,32]

All of these studies suggest that different local propensities of the residues may be correlated with each other, and that their presence is important not only as individual residues but as part of small sequences (pairs, triplets, etc.) that induce and stabilize the helical structure. Statistics on known protein structures provide crucial information about that correlation, and can be very useful to reveal patterns of amino acid interactions. Moreover, the data of

Grant sponsor: FCT; Grant number: SFRH/BPD/26737/2006
*Correspondence to: A. L. Magalhães, REQUIMTE, Universidade do Porto, R. Campo Alegre, 687, 4169-007 Porto, Portugal. E-mail: almagalh@fc.up.pt
Received 24 January 2007; Revised 8 March 2007; Accepted 15 March 2007
Published online 24 July 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21525

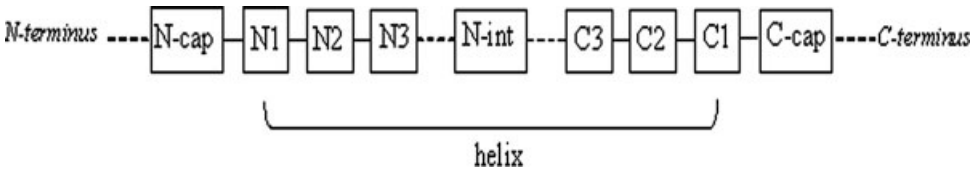


Figure 1

Relevant positions considered in this study according to Richardson and Richardson's notation.[6]

known protein 3D structures are continuously increasing which makes obligatory the updating of the information provided by the statistical studies.

In this work, we present an exhaustive analysis of the occurrence of amino acid pairs in a recent and updated set of helical protein segments, taking into account a few factors such as the type of helix (α-helices or 3_{10}-helices), residue separation, position, and orientation in the helix. Based on the data collected we have also carried out a search for the most probable initial and final sequences in each type of helices. The main goal of this work was to extract information from a systematic survey of helices in proteins that support some empirical rules for their formation and stabilization. Moreover, this study also presents a web site (http://www.ncc.up.pt/∼nf/rps) where the results are automatically and periodically updated, following the growth of the Protein Data Bank (PDB) database.

MATERIALS AND METHODS

The Pisces server[33] was used to select a subset of protein chains from the PDB[34,35] with structures solved at a resolution higher than 1.6 Å, with a *R*-factor lower than 25%, and showing a maximum of 20% homology. Secondary structure assignments were automatically done by PDB using the Kabsch and Sander algorithm.[36] However, to avoid the use of incorrectly classified motif sequences, the HELIX and ATOM records in the pdb files were checked using our own programs, and those presenting inconsistencies were rejected. In addition, all the pdb entries that contain more than 10% of nonstandard or undefined residues were also discarded. Consequently, the original set of 1125 protein chains was reduced to 815 as the working set (available as Supplementary Material at www.ncc.up.pt/∼nf/rps) which corresponds to a total number of 186,301 amino acid residues. Two subsets were then formed with 5388 and 168 sequences corresponding to α-helices and 3_{10}-helices, respectively, containing at least seven residues.

The relevant helical positions considered in this work are the first three residues and the first nonhelical residue at the N-terminus side of the helices N1, N2, N3, and N-cap, respectively, and the equivalent positions at the C-terminus side C3, C2, C1, and C-cap according to Richardson and Richardson's notation.[6] All other helical residues between N4 and C4 are classified as interior, N-int (see Fig. 1). Four types of amino acid vicinities were analyzed, namely, $(i, i + 1)$, $(i, i + 2)$, $(i, i + 3)$, and $(i, i + 4)$.

The preference of a particular amino acid to be included in helical motifs was evaluated by means of two different statistics called *global* ($P^{g}_{X_i}$) and *local* ($P^{l}_{X_i}$) *propensities*.[11] In a similar fashion we defined the global ($P^{g}_{X_iY_{i+k}}$) and local ($P^{l}_{X_iY_{i+k}}$) propensities, for all the amino acid pairs as:

$$P^{g}_{X_iY_{i+k}} = \frac{n^{helix}_{X_iY_{i+k}}}{\sum_{A,B} n^{helix}_{A_iB_{i+k}}} \Big/ \frac{n^{all}_{XY_k}}{\sum_{A,B} n^{all}_{AB_k}},$$

$$P^{l}_{X_iY_{i+k}} = \frac{n^{helix}_{X_iY_{i+k}}}{\sum_{A,B} n^{helix}_{A_iB_{i+k}}} \Big/ \frac{n^{helix}_{XY_k}}{\sum_{A,B} n^{helix}_{AB_k}}$$

which evaluate the tendency of the particular pair of amino acids XY to integrate helices with X in position i and Y in position $i + k$ (i = N-cap, N1, N2, N3, N-int, C3, C2, C1, C-cap; k = 1, 2, 3, 4). $P^{g}_{X_iY_{i+k}}$ is the ratio of the relative frequency of the pair XY appearing at a particular position $(i, i + k)$ in helices and the relative frequency of occurrence of that pair in all of the protein sequences. Therefore, the larger/smaller the value of $P^{g}_{X_iY_{i+k}}$, the higher/lower the preference of the pair to occur in helical position i (a value close to 1 means no preference at all, reflecting similar distributions inside and outside the helices). On the other hand, the local propensity $P^{l}_{X_iY_{i+k}}$ is defined in the subset of all the helical pairs by the ratio of the percentage of occurrence of the pair XY in a particular position i and the percentage of occurrence of that pair in the helices regardless its position. This statistics is thus a measure of the preference of the pair for a particular position inside the helix. The term $n^{helix}_{X_iY_{i+k}}$ is the number of occurrences of the particular pair XY found in the helices set with X in position i and Y in position $i + k$; in $\sum_{A,B} n^{helix}_{A_iB_{i+k}}$ the summation is extended to all amino acids and it represents the total number of pairs in positions $(i, i + k)$ found in helices; $n^{all}_{XY_k}$ and $n^{helix}_{XY_k}$ are the number of occurrences of the particular pair XY observed in all the protein sequences and in all helices, respectively, where Y is always

Table I
Global Propensities of Individual Amino Acids in α-Helices and in 3_{10}-Helices

	Ncap		N1		N2		N3		Nint		C3		C2		C1		Ccap	
	α	3_{10}	α	3_{10}	α	3_{10}	α	3_{10}	α	3_{10}	α	3_{10}	α	3_{10}	α	3_{10}	α	3_{10}
C	0.9	0.5	1.1	0.5	0.5	0.0	0.5	0.5	0.9	0.9	0.9	0.0	0.8	1.8	1.0	0.9	0.7	1.4
P	1.0	1.2	1.3	3.7	2.6	4.0	1.1	1.2	0.1	0.4	0.0	1.2	0.0	0.0	0.0	0.1	1.9	1.8
A	0.9	1.0	0.5	0.6	1.1	0.8	1.2	1.2	1.6	1.2	1.5	1.4	1.5	0.5	1.2	0.6	0.7	0.5
T	1.0	1.2	2.0	1.0	0.8	0.8	0.8	0.9	0.8	0.8	0.7	0.5	0.9	1.0	0.7	0.8	0.8	1.0
G	1.0	1.2	1.2	0.9	0.7	1.0	0.7	0.8	0.4	0.7	0.2	0.5	0.3	0.6	2.2	0.8	2.7	2.1
S	0.9	1.6	2.5	1.1	0.9	2.0	1.0	1.4	0.7	0.9	0.7	1.1	1.0	1.0	1.0	0.5	1.1	1.9
D	1.0	0.8	2.4	2.0	0.9	0.9	1.6	1.0	0.8	0.8	0.7	1.0	0.6	0.9	0.7	0.6	1.0	0.9
N	0.8	1.1	2.1	1.6	0.5	0.7	0.8	1.6	0.7	0.9	0.7	0.7	1.1	2.1	1.7	0.6	1.3	1.0
E	0.9	0.7	0.6	0.7	1.4	1.3	2.6	2.1	1.3	1.0	1.5	1.0	1.5	0.9	0.8	0.7	0.7	0.8
Q	0.9	1.0	0.7	1.6	1.2	0.8	1.5	1.1	1.3	1.1	1.3	1.5	1.4	1.5	1.3	0.8	1.0	0.7
K	0.9	1.2	0.7	0.6	1.0	1.0	1.0	0.4	1.1	1.2	1.6	1.8	1.5	1.9	1.2	1.1	1.1	0.3
R	0.9	1.2	0.7	0.8	1.0	0.6	0.9	0.6	1.3	0.8	1.5	1.1	1.3	0.8	1.2	1.3	1.0	0.9
Y	1.1	0.5	0.7	0.5	0.9	0.2	0.7	0.9	1.0	1.2	0.8	1.2	1.0	2.4	1.0	2.1	0.6	0.5
H	1.1	0.5	1.0	2.1	0.8	0.8	1.1	0.3	0.8	0.8	1.2	1.3	1.3	0.8	1.3	1.3	0.9	2.4
W	0.9	0.0	0.7	0.4	1.2	2.3	1.0	3.1	1.0	2.4	0.9	0.8	0.5	2.3	0.5	2.3	0.6	0.4
F	0.9	1.2	0.5	0.5	0.9	0.3	0.7	1.1	1.1	0.8	0.7	1.4	1.0	1.5	0.9	2.0	0.6	0.8
M	1.5	1.3	0.6	1.0	0.7	0.3	0.6	0.0	1.4	1.6	1.5	0.6	1.2	1.0	0.9	1.9	0.6	0.3
L	1.0	0.8	0.4	0.4	1.0	0.9	0.6	1.2	1.5	1.8	1.6	1.4	1.5	1.2	1.0	1.5	0.7	1.0
I	1.2	1.0	0.3	1.0	0.8	0.6	0.5	0.4	1.2	0.9	1.2	0.8	0.8	0.3	0.5	1.5	0.6	0.7
V	1.1	0.6	0.3	0.5	0.8	0.7	0.6	0.5	1.0	0.7	0.8	0.4	0.7	0.5	0.5	1.0	0.6	0.7

found k positions after X, independently of its position in the protein chain; in $\sum_{A,B} n^{all}_{AB_k}$ and $\sum_{A,B} n^{helix}_{AB_k}$ the summation is extended to all amino acids, and it represents the total number of pairs observed in all the protein sequences and helices, respectively, where both amino acids are k positions apart independently of the position in the protein chain. From the above definitions, the ratio of the global and local propensities gives the overall propensity of a pair to occur somewhere in a helix.

The tables with the global and local propensities calculated in our study are available in our web site at http://www.ncc.up.pt/~nf/rps in two formats: as a Comma Separated Value (CSV) file and as a PDF file. We developed a program that automatically and periodically updates this information in three steps. In a first step, the program retrieves an updated list of protein chains with a maximum of 20% homology, whose structures were solved at a resolution higher than 1.6 Å and with a *R*-factor lower than 25%. That list may be obtained from the Dunbrack Lab website.[33] In the second step, the retrieved list of protein chains is used to download the corresponding files from the PDB website,[34,35] where each chain segment is identified by its secondary structure motif. Finally the program uses each amino acid pattern occurrence frequency to compute the global and local propensities and to update the tables in our website.

RESULTS AND DISCUSSION

Individual propensities in α-helices

The large and unbiased database used in this work, which includes almost 6000 helices selected with very tight criteria from the PDB,[34,35] gives a strong statistical support to the observations. In general, there is good agreement between our findings and previously reported statistical analysis in protein α-helices with smaller protein samples.[5–12,14,37]

Table I shows the global propensities of each individual amino acid for different positions calculated in our working data set in a way similar to Penel *et al.*[11] At a first sight it seems that the amino acid distribution is highly dependent on the position in the helix. In fact, this was confirmed quantitatively for each of the helical positions by means of a χ^2 test with 19 degrees of freedom and 0.5% level of significance. The χ^2 values obtained were indeed highly significant at this level. Moreover, the χ^2 values for the distribution of each amino acid over the eight extreme positions (N-cap, N1, N2, N3, C3, C2, C1, and C-cap) were also evaluated. Again, high significant χ^2 values were obtained for the same level of significance, which confirm that, at this level, none of the amino acids have a uniform distribution over these eight positions, with the exception of Cys.

A close inspection of Table I shows that the results corroborate previous studies on this type of protein secondary structure.[5–9,11,37] On one hand, the extreme positions show a wide range of global propensities. As paradigmatic examples we refer the positions N1, for which the values go from 0.3 (Val and Ile) to 2.5 (Ser), and C1 with the range 0.0 (Pro) to 2.2 (Gly). On the other hand, some amino acids show a particularly high dependence on the position within α-helices, as is the case of Pro (2.6 in N2 versus total absence in C1, C2, and C3), as remarked before by Duncan *et al.*,[13] and also Gly (2.2 in C1 versus 0.2 in C3). The particular case of Pro is note-

DOI 10.1002/prot

worthy. It is the residue with the lowest preference for helices ($P^{g}_{Pro}/P^{l}_{Pro} \approx 0.6$) but it has a very high local propensity differentiation. It occurs almost always at N-terminus positions rather than interior or C-terminus positions because it does not possess a free amide hydrogen to establish *i*, *i* + 4 backbone–backbone hydrogen bonding.[11] Interesting to notice is also its behavior at the C-terminus. It has high propensity values in the four positions immediately after the end of the helix which contrasts with the last four positions within the helix (see Supplementary Material at www.ncc.up.pt/~nf/rps). However, our results clearly show a higher preference for N2 than for N1 position. Moreover, Penel *et al.*[11] identified Gln, Glu as the residues with highest global propensities at N2, but from our results, Pro and Trp also have to be included in that list. Ala was found to be the amino acid residue with highest helical occurrence ($P^{g}_{Ala}/P^{l}_{Ala} \approx 1.8$), slightly more frequent at interior positions, and this was previously justified by a low side-chain configuration entropy lost upon chain folding.[38] In addition, it seems that N1 is a position particularly avoided by Ala when compared with the others.

Our results corroborate also some experimental studies on Ala-based peptides where the importance of the N1 position is demonstrated,[13] and the role of individual residues is stressed,[12] namely, the moderate tendency of Leu to be found in C2 and C3 positions, and the propensity of Met for C3, in opposite to C1 and C2.

Individual propensities in 3_{10}-helices

The size of the 3_{10}-helices sample extracted from the original dataset is 168, which is substantially smaller than the data set of α-helices (5388). However, after comparison of both cases, we were able to identify some common and distinct features.

A high dependence of the amino acid distribution on the position within this type of helices was also observed. In general the distribution is similar but the particular cases of Cys and Pro seem to differ (see Table I). The former amino acid is far from having a uniform distribution, since some positions, such as C2, are preferred and others are either almost or completely avoided, namely, N2 and C3. As in the case of α-helices, Pro shows a strong preference for N-terminus positions especially for N1 and N2, but not for N3. The case of Trp is noteworthy when compared with its occurrence in α-helices. Its propensity is clearly lower at N1 and, remarkably, at N-cap, but higher at N2, N-int, C2, C1 and, especially, at N3.

In the α-helices case the N-cap position was mainly occupied by big- and low-polarity amino acids ranging from Tyr to Val, whereas in 3_{10}-helices there is a slight shift towards small side chains (from Pro to Asn). However, the case of the voluminous Trp is remarkable, because it seems to decrease in the terminal positions and to increase in the interior of the 3_{10}-helices.

Residue pair propensities

The analysis of the 20 individual amino acid distributions in helices is not consistent enough to result in clear rules for the explanation and prediction of their formation and stabilization. That distribution may depend not only on the characteristics of the different positions but also on the interactions with neighbour residues. Obviously, pairs of amino acids are the next feature to be considered in a progressive complexity analysis. The formation of the helix within the living cell is controlled by signals, and we assume here that the N- and C-termini positions should present particular characteristics that may reflect the ability to start and to stop the folding process.[8,39] The present study puts some emphasis on those helical and nonhelical positions as defined in Figure 1.

The systematic search of pairs in our database resulted in a huge amount of data which is collected in a set of tables, available as Supplementary Material at www.ncc.up.pt/~nf/rps/, from which a few relevant examples are presented here (see Tables II–V). Each table concerns either type of helical motifs (α-helices or 3_{10}-helices) and either type of propensity values (global or local). The first position in a pair is N-cap, N1, N2, N3, N-int, C3, C2, C1, or C-cap. The second position in a pair refers to residues at *i* + 1, *i* + 2, *i* + 3, and *i* + 4, where *i* is the first position of the pair. In all the tables, the first residue in the pair is given by the row and the other one by the column. The 20 amino acids are sequenced from Cys to Val according to their size and polarity. The data are thus organized as a large number of 20 × 20 square matrices, which are periodically updated at our website as explained in the Materials and Methods section. The statistical significance of every propensity value was evaluated by calculating the associated error and *P*-value which are also available for download at the website. To make easier the reading of the propensity tables, a colour scale was introduced and those values with a *P*-value lower than 0.05 appear in bold and white.

An overview of all the matrices shows that they are not uniform neither symmetric. The discrepancy of the values suggest that there are, indeed, different preferences for the position of the pairs and that their orientation in the chain sequence is not meaningless, that is, in general, the frequency of occurrence of a particular pair XZ differs from that of ZX.

Pair propensities in α-helices

Tables II–V concern the global propensities in α-helices of the pairs (*i*, *i* + 1), (*i*, *i* + 2), (*i*, *i* + 3), (*i*, *i* + 4), respectively, with *i* being the N-cap position, and they

Table II

Global Propensities of (i, i + 1) *Pairs in* α*-Helices* (i = *N-cap*)

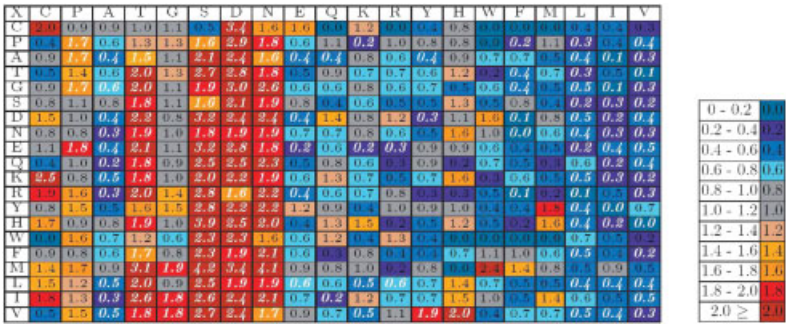


Table III

Global Propensities of (i, i + 2) *Pairs in* α*-Helices* (i = *N-cap*)

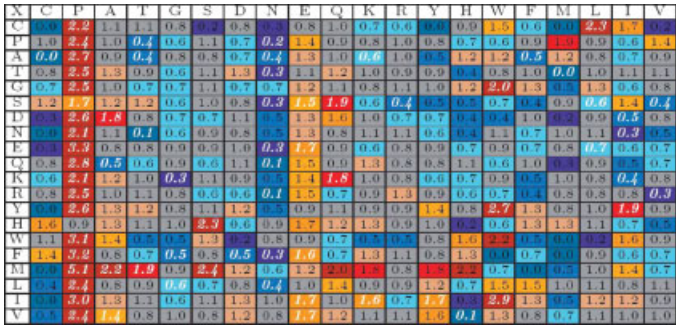


Table IV

Global Propensities of (i, i + 3) *Pairs in* α*-Helices* (i = *N-cap*)

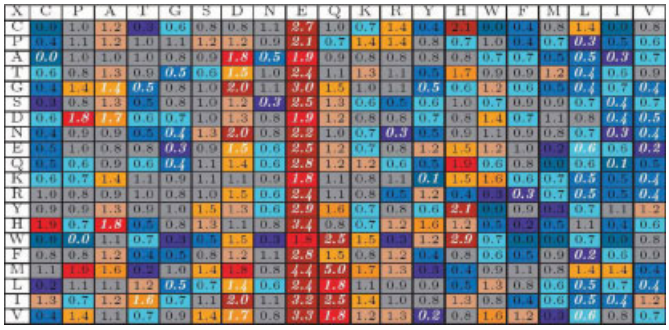
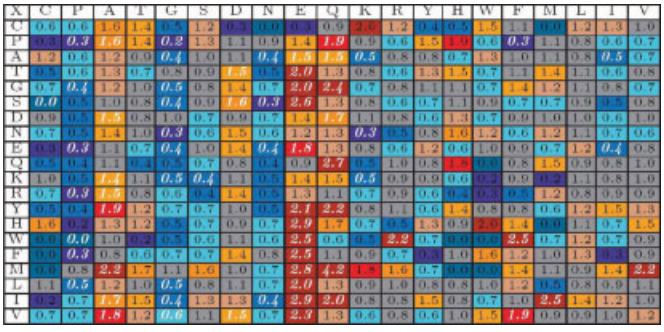


Table V

Global Propensities of (i, i + 4) *Pairs in* α*-Helices* (i = *N-cap*)



are presented here as representative examples. A preference of small and polar residues such as Thr, Ser, Asp, and Asn for position N1 is noticeable regardless the residue at N-cap (see Table II). The Cys residue at N-cap is the only exception as it pairs almost exclusively with Asp ($P^{g}_{C_{Ncap}D_{N1}} = 3.4$)*. In general, the values of Asn are compared with those of Thr, but Ser and Asp show higher propensities to occur at N1. On the other hand, that position seems to be avoided by big and hydrophobic residues, namely, Trp, Phe, Met, Leu, Ile, and Val.

These results have a reasonable agreement with a previous experimental work on polyalanine-based peptides[10] but new features have to be referred. In fact, Thr is more favorable at position N1 than in the helix interior but not in N2; in this latter position the highest propensity for Thr occurs only when Met is in N-cap (see Table III).

When the second position in the pair is N2, then a clear preference for Pro is observed, except if His is in N-cap, $P^{g}_{H_{Ncap}P_{N2}} = 0.9$ (see Table III). We are currently trying to find a chemical/structural explanation for the particular behavior of the $H_{N\text{-}cap}P_{N2}$ pair.

Some interactions between residues have already been recognized as important structural factors of helix formation and stabilization. A typical example is the *capping box*, where the side chain of a residue in N-cap makes a H-bond with the backbone amino group of the N3 and, reciprocally, the side chain of N3 establishes a H-bond with the backbone amino group of the N-cap.[37,39] The pair $S_{N\text{-}cap}E_{N3}$ was identified in a previous analysis on a smaller dataset as particularly relevant to define a capping box motif, presenting a value of 13 for its global propensity.[39] Concerning this particular point, the results that come out of our analysis are based on the data shown in Table IV. In fact, all types of residues show a high propensity to appear at N-cap associated with Glu at N3. However, the highest pair propensity does not appear with Ser ($P^{g}_{S_{Ncap}E_{N3}} = 2.5$) but with Met ($P^{g}_{M_{Ncap}E_{N3}} = 4.4$). These results, on one hand, are reflections of the high ability of Glu in N3 position to be involved in the formation of capping box motifs due to its long side-chain terminating in a carboxylate group. On the other hand, Asp, which has got a shorter side-chain, shows lower pair propensities for the same position as a consequence of its lower potential to form a capping box motif, as referred in a previous study.[39]

Considering pairs starting at N1, it is found, as expected from the results above, predominance of small and polar residues in the first position (Asn, Asp, Ser, and Thr). They show a great variability in propensities, but the strongest couplings appear for Pro in position N2, and Glu in N3 and N4.

At the interior of the helices the pairs are mainly formed by large and polar residues and by Ala as well, but almost never by Pro. On the other hand, Gly appears almost exclusively in positions $i + 3$ and, especially, $i + 4$ of these inner pairs, and with different preferences to the residue in the first position i. Interesting is the high global propensities of the pairs (C_{Nint}, G_{i+4}), (A_{Nint}, G_{i+4}), (M_{Nint}, G_{i+4}), and (L_{Nint}, G_{i+4}) when compared with the corresponding low values for (Nint, $i + 3$) pairs.

The findings are based on the high statistical significance of the pair propensities as shown in the corresponding tables available in the Supplementary Material at www.ncc.up.pt/∼nf/rps.

Pair propensities in 3_{10}-helices

Concerning the 3_{10}-helices, the dimension of the sample is substantially smaller (168 helical sequences) than the previous case (5388), which prevents us to present here an accurate statistical comparison. However, since the website is being periodically updated, the analysis will be done in the follow-up work of this study. Nevertheless, with the whole set of tables stored at present it is possible to compare both cases with some care (see Supplementary Data at www.ncc.up.pt/∼nf/rps). A clear distinction from the previous case is easily perceived at first sight, which is the large number of cells with zero value. In the propensity tables, a value of zero means that the pair is observed somewhere in the chain but never at that particular position. Although this may be a consequence of the small dimension of the sample, the distribution of the propensity values among all the 400 possible pairs is very irregular, and consequently some preliminary conclusions may be inferred.

The analysis of the global pair propensities revealed a preference for small amino acids to occur at N-terminus positions. The particular case of Cys seems to be distinct from the others as we already pointed out while analysing the values of individual propensities. When Cys is at N-cap position there is a clear preference to couple with Cys in N1. Reciprocally, it appears at N1 position only when Cys, and also Lys and Phe but in less extent, is preceding it in N-cap.

In the previous section it was referred that Pro prefers the N2 position in α-helices, but in the case of the 3_{10}-helices it occurs almost exclusively at N1 and N2 with similar preferences. Big and aromatic amino acids show also a distinct behavior, but the particular case of Tryptophan is noteworthy. It never appears at N-cap position, and at N1 it only pairs with Val, Asp, Ile, and Lys at N2, N3, N4, and N5, respectively. On the other hand, at C-terminus, it appears at C-cap only paired with Asn, Lys, and Tyr at C1, C2, and C3, respectively.

Relevant short amino acid sequences at N-terminus and C-terminus

The positions at the N- and C-termini of the helices are nowadays recognized as to play an important role in

*In order to simplify the notation of the pairs, each amino acid is identified by its one-letter code.

Table VI
Number of Occurrences of Some Short Sequences Selected From the Analysis of the Global Pair Propensities

Sequence										α-Helix		3_{10}-Helix	
Ncap	N1	N2	N3	...	C3	C2	C1	Ccap	In protein chains	In local	In helix	In local	In helix
_	S	P	E	...	_	_	_	_	51	24	29	0	0
_	T	P	E	...	_	_	_	_	64	24	28	1	1
_	D	P	E	...	_	_	_	_	60	23	24	1	1
_	S	R	E	...	_	_	_	_	44	10	23	0	0
_	_	_	_	...	A	L	A	_	138	14	134	0	0
_	_	_	_	...	_	E	N	P	33	8	0	0	0
_	_	_	_	...	E	A	A	_	92	8	64	0	1
_	_	_	_	...	_	L	A	K	86	4	62	0	0
_	_	_	_	...	E	K	Y	P	5	4	0	0	0
_	_	_	_	...	I	A	R	_	53	3	32	0	0
_	_	_	_	...	L	E	S	G	8	3	1	0	0
_	_	_	_	...	A	L	M	_	30	0	22	0	1
_	_	_	_	...	W	L	K	_	13	0	8	0	1
V	_	P	D	...	_	_	_	_	46	1	0	0	0
M	T	E	E	...	_	_	_	_	0	—	—	—	—
M	S	W	P	...	_	_	_	_	0	—	—	—	—
G	N	E	D	...	_	_	_	_	0	—	—	—	—
_	_	_	_	...	Q	K	G	H	0	—	—	—	—
_	_	_	_	...	W	M	G	V	0	—	—	—	—

their stabilization. We assume here that the sequences of amino acids at both termini may be highly correlated with the start and stop processes of the helix formation, as well as with its thermodynamic stabilization. Therefore, we looked for relevant short sequences at both termini of the helices, including the first nonhelical positions N-cap and C-cap, from which a few representative examples are reported in Table VI. To save computation time, we did not consider all possible sequences but only those that had the potential to be more frequent based on the high propensity values of the pairs they include. Starting at a particular position, the methodology employed here was able to select the next amino acid in the sequence to which a high pair propensity is associated, and continued forward using the same criterion in order to establish a high probable pathway. Then, an automatic search of those selected sequences was carried out in all protein chains of the working sample.

Table VI shows that the sequence $M_{N\text{-}cap}T_{N1}E_{N2}E_{N3}$ (**MTEE--**) never appears even though the partial pair propensities in α-helices are particularly high, namely, $P^{g}_{M_{Ncap}T_{N1}} = 3.1$, $P^{g}_{T_{N1}E_{N2}} = 4.0$, and $P^{g}_{E_{N2}E_{N3}} = 3.1$. The same happens to many other sequences (like the next four examples in the table) meaning that high propensities of the constituent pairs do not guarantee a high probable sequence.

Other sequences occur with reasonable frequency but rarely or never inside a helix, which is the case of **V_PD--** at the N-terminus (in this notation the broken line indicates the direction of the remaining helical sequence and a dash means a position that may be occupied by any residue).

On the other hand, high propensity pairs may give a clue to find positive examples as, for instance, the sequences _ **SPE--**, **_TPE--**, **_DPE--**, and **_SRE--**. These initial triplets of amino acids appear almost exclusively within helices and approximately half of the observations occur exactly at N1 despite the large number of other positions available in the helices. This is not surprising considering the known ability of Glu to be involved in the formation of a capping box motif at N3, as discussed earlier, but only these four triplets appear more frequently.

At the C-terminus we have found triplets with high incidence but low local preference within the helices, which are the cases of **--ALA_**, **--EAA_**, **--_LAK**, **--IAR_**, **--ALM_**, and **--WLK_**.

There is another group of short sequences that present low overall occurrences but very high preferences to positions at the interior of helices or their termini. For example, **--_ENP** appears eight times in α-helices, out of a total of 33, and exactly with Pro in C-cap.

The frequencies of occurrence of these short sequences in our protein working set suggest a probability function which is far from a uniform distribution because helical motifs represent only 36% of the total length of all the protein chains considered (66,340 amino acid residues out of a total of 186,301).

Another example that may illustrate very well the specific features of small sequences in helices is **--EKYP**. It is observed only five times in proteins, but four of them appear exactly with Pro at C-cap position of α-helices. This sequence adopts a typical conformational as shown in Figure 2 (corresponding to 1JR8 entry of PDB), and

DOI 10.1002/prot

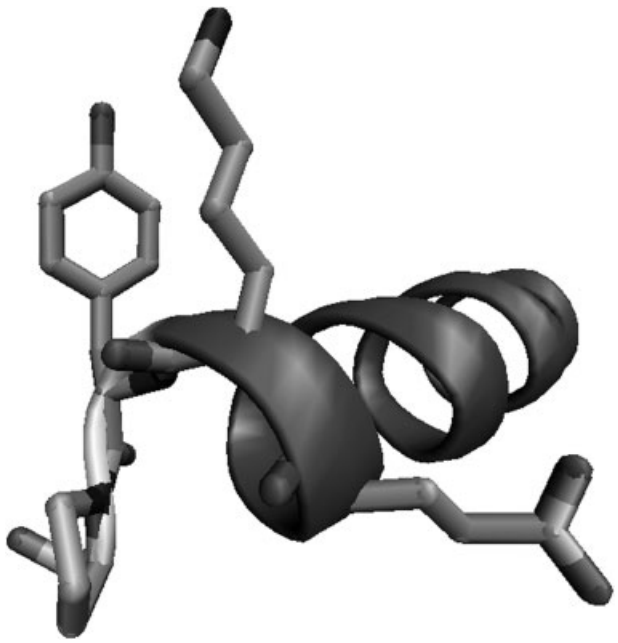


Figure 2
*Geometry of the **--EKYP** sequence in 1JR8 entry of the PDB (residues 66 to 69 of chain A).*

for which the corresponding data are summarized in Table VII. The structure shows an interesting interaction between the two distended side chains of Lys and Tyr, where the average distance (Lys)$N_\zeta \cdots O_\eta$(Tyr) is 3.7 Å. The single case where the sequence does not appear at N-cap position is observed in the 1MIN entry. The four-residue sequence defines a coil structure between two α-helices, where Lys and Tyr adopt a conformation really distinct from that shown in Figure 2, which prevents the interaction between their side chains. To verify the specificity of the conformation, we surveyed the working sample of protein chains for all the pairs $Lys_i Tyr_{i+1}$ and $Tyr_i Lys_{i+1}$. We have found 659 examples of such pairs but only six different cases, not related to those mentioned previously, show geometries of interaction similar to that gathered in Table VII, namely, d[(Lys)$N_\zeta \cdots O_\eta$(Tyr)] $<$ 4 Å, $-165^\circ <$ (Lys)$\chi_{1,4} < +165^\circ$ and $-60^\circ <$ (Tyr)$\chi_{1,2} < +100^\circ$. In addition, these six examples appear in helices or in turns. These results suggest that the sequence **--EKYP** with this particular side-chain interaction between Lys and Tyr has indeed a high specificity to C-terminus position.

CONCLUDING REMARKS AND FUTURE DIRECTIONS

This work presents the most comprehensive analysis of helical motifs in proteins undertaken to date. An exhaustive study of the frequency of occurrence of individual amino acids and all possible pairs was carried out on a set of 5556 helices. Pairs of type (i, $i + 1$), (i, $i + 2$), (i, $i + 3$), and (i, $i + 4$) were considered starting at relevant positions near or within helices: N-cap, C-cap, N1, N2, N3, C3, C2, C1, N-int. The protein sample used in this work was sufficiently large and unbiased which gives confidence to the final results expressed in terms of global and local propensities. Some general features of residue pairs at both N- and C-termini were identified. For example, the Cys residue at N-cap pairs almost exclusively with Asp in N1; and Pro at N2 position seems to pair with whatever amino acid in N-cap except His for which Pro has no preference at all.

Some sequences, although occurring rarely, seem to play a very specific role because they are observed always at the same position of helices. This is the case, for instance, of the sequence EKYP for which four out of five times it is observed at the same C-termini position of α-helices.

In this work, a bioinformatics tool was developed, which can automatically follow the growth of the PDB database and update the propensities values in both α-helices and 3_{10}-helices. The results produced by the statistical analysis are periodically updated and stored as Supplementary Material in our website http://www.ncc.up.pt/∼nf/rps.

The amount of information collected is huge and will need a further automatic analysis using, for instance, Inductive Logic Programming (ILP) algorithms, in order to obtain useful predictive rules. The physico-chemical characteristics of the 20 amino acids and the data concerning their individual and pair propensities generated in this work would be crucial to start the ILP studies.

Table VII
Main Characteristics of the ---EKYP Sequence

		Distance (Å)	Lys side-chain				Tyr side-chain	
PDB entry	Sequence location	(Lys)$N_\zeta \cdots O_\eta$(Tyr)	χ_1	χ_2	χ_3	χ_4	χ_1	χ_2
1H16	(A) 723–726	3.86	−178.9	176.7	174.2	−177.3	−57.5	87.4
1JR8	(A) 66–69	3.39	179.6	174.5	−177.8	−179.7	−55.9	97.3
1T6U	(A) 69–72	3.90	−175.1	169.7	176.6	164.9	−54.7	95.7
1YXY	(A) 129–132	3.82	−173.0	172.4	−176.9	−175.2	−56.7	73.9

With this approach we aim to find some general rules that can be applied to any amino acid sequence in order to predict the stability of helical motifs.

ACKNOWLEDGMENTS

The authors thank Dr. Cristian R. Munteanu for his help in the characterization of the Lys-Tyr interactions in proteins. We are also grateful to the reviewers for their useful comments.

REFERENCES

1. Davies DR. A correlation between amino acid composition and protein structure. J Mol Biol 1964;9:605–609.
2. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from proteins. Biochemistry 1974;13:211–222.
3. Levitt M. Conformational preferences of amino acids in globular proteins. Biochemistry 1978;17:4277–4285.
4. Williams RW, Chang A, Juretic D, Loughran S. Secondary structure predictions and medium range interactions. Biochim Biophys Acta 1987;916:200–204.
5. Argos P, Palau J. Amino acid distribution in protein secondary structures. Int J Pept Protein Res 1982;19:380–393.
6. Richardson JS, Richardson DC. Amino acid preferences for specific locations at the ends of α-helices. Science 1988;240:1648–1652.
7. Kumar S, Bansal M. Dissecting α-helices: position-specific analysis of α-helices in globular proteins. Proteins Struct Funct Genet 1998;31:460–476.
8. Presta L, Rose GD. Helix signals in proteins. Science 1988;240: 1632–1641.
9. Petukhov M, Munoz V, Yumoto N, Yoshikawa S, Serrano L. Position dependence of non-polar amino acid intrinsic helical propensities. J Mol Biol 1998;278:279–289.
10. Petukhov M, Uegaki K, Yumoto N, Yoshikawa S, Serrano L. Position dependence of amino acid intrinsic helical propensities. II. Non-charged polar residues: Ser, Thr, Asn, and Gln. Protein Sci 1999;8:2144–2150.
11. Penel S, Hughes E, Doig AJ. Side chain structures in the first turn of the α-helix. J Mol Biol 1999;287:127–143.
12. Petukhov M, Uegaki K, Yumoto N, Serrano L. Amino acid intrinsic α-helix dependence at several positions of C terminus. Protein Sci 2002;11:766–777.
13. Duncan AE, Cochran DAE, Penel S, Doig AJ. Effect of the N1 residue on the stability of the α-helix for all 20 amino acids. Protein Sci 2001;10:463–470.
14. Wilson CL, Boardma PE, Doig AJ, Hubbard SJ. Side improved prediction of N-termini of α-helices using empirical information. Proteins Struct Funct Bioinf 2004;57:322–330.
15. Cochran DAE, Doig AJ. Effect of the N2 residue on the stability of the α-helix for all 20 amino acids. Protein Sci 2001;10:1305–1311.
16. Cochran DAE, Penel S, Doig AJ. Effect of the N1 residue on the stability of the α-helix for all 20 amino acids. Protein Sci 2001; 10:463–470.
17. Serrano L, Fersht AR. Capping and α-helix stability. Nature 1989; 342:296–299.
18. Lecomte JT, Moore CD. Helix formation in apocyochrome b5: the role of a neutral histidine at the N-cap position. J Am Chem Soc 1991;113:9663–9665.
19. Bell JA, Becktel WJ, Sauer U, Baase WA, Matthews BW. Dissection of helix capping in T4 lysozyme by structural and thermodynamic analysis of six amino acid substitution at Thr59. Biochemistry 1992;31:3590–3596.
20. Serrano L, Sancho J, Hirshberg M, Fersht AR. α-Helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at N and C-caps and the replacement of alanine by glycine or serine at solvent exposed surfaces. J Mol Biol 1992;227:544–559.
21. Lyu PC, Wemmer DE, Zhou HX, Pinker RJ, Kallenbach NR. Capping interactions in isolated α-helices: position-dependent substitutions effects and structure of a serine-capped peptide helix. Biochemistry 1993;32:421–425.
22. Chakrabartty A, Doig AJ, Baldwin RL. Helix capping propensities in peptides parallel those in proteins. Proc Natl Acad Sci USA 1993;90:1132–1136.
23. Forood B, Feliciano EJ, Nambiar KP. Stabilization of α-helical structures in short peptides via end capping. Proc Natl Acad Sci USA 1993;90:838–842.
24. Yumoto N, Murase S, Hattori T, Yamamoto H, Tatsu Y, Yoshikawa S. Stabilization of α-helix in C-terminal fragment of neuropeptide Y. Biochem Biophys Res Commun 1993;196:1490–1495.
25. Doig AJ, Chakrabartty A, Klingler TM, Baldwin RL. Determination of free energies of N-capping in α-helices by modification of the Lifson-Roig helix-coil theory to include N- and C-capping. Biochemistry 1994;33:3396–3403.
26. Doig AJ, Baldwin RL. N-capping and C-capping preferences for all 20 amino acids in α-helical peptides. Protein Sci 1995;4:1325–1336.
27. Fooks HM, Martin ACR, Woolfson DN, Sessions RB, Hutchinson EG. Amino acid pairing preferences in parallel β-sheets in proteins. J Mol Biol 2006;356:32–44.
28. Crasto CJ, Feng J. Sequence codes for extended conformation: a neighbor-dependent sequence analysis of loops in proteins. Proteins Struct Funct Genet 2001;42:399–413.
29. George RA, Heringa J. An analysis of protein domain linkers: their classification and role in protein folding. Protein Eng 2003;15:871–879.
30. Andrew CD, Bhattacharjee S, Kokkoni N, Hirst JD, Jones GR, Doig AJ. Stabilizing interactions between aromatic and basic side chains in α-helical peptides and proteins. Tyrosine effects on helix circular dichroism. J Am Chem Soc 2002;124:12706–12714.
31. Olson CA, Spek EJ, Shi Z, Vologodskii A, Kallenbach NR. Cooperative helix stabilization by complex Arg-Glu salt bridges. Proteins Struct Funct Genet 2001;44:123–132.
32. Iqbalsyah TM, Doig AJ. Pairwise coupling in an Arg-Phe-Met triplet stabilizes α-helical peptide via shared rotamer preferences. J Am Chem Soc 2005;127:5002–5003.
33. Wang G, Dunbrack RL, Jr. Pisces: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
34. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
35. Bernstein FC, Koetzle TF, Williams GJ, Meyer EE, Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
36. Kabsch W, Sander C. Dictionary of protein secondary structure-pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
37. Aurora R, Rose GD. Helix capping. Protein Sci 1998;7:21–38.
38. Creamer TP, Rose GD. Side-chain entropy opposes α-helix formation but rationalizes experimentally-determined helix-forming propensities. Proc Natl Acad Sci USA 1992;89:5937–5941.
39. Harper ET, Rose GD. Helix stop signals in protein and peptides: the capping box. Biochemistry 1993;32:7606–7609.

 DOI 10.1002/prot