Protein Science (1998), 7:1368–1379. Cambridge University Press. Printed in the USA.

Deciphering the structural code for proteins: Helical propensities in domain classes and statistical multiresidue information in α-helices

JOSÉ A. NEGRETE, YOLANDA VIÑUALES, AND JAUME PALAU

Unitat de Biotecnologia Computacional, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Tarragona 43005, Catalonia, Spain

(RECEIVED September 23, 1997; ACCEPTED March 2, 1998)

Abstract

We made several statistical analyses in a large sample of nearly 4,000 helices (from 546 redundancy-controlled PDB protein subunits), which give new insights into the helical properties of globular proteins. In a first experiment, the amino acid composition of the whole sample was compared with the composition of two helical sample subgroups (the "mainly-α" and the "$(\alpha/\beta)_8$ barrel" domain classes); we reached the conclusion that composition-based helical propensities for secondary structure prediction do not depend on the structural class.

Running a five-residue window through the whole sample, the positional composition revealed that positive and negative residues are located throughout the helices and tend to neutralize the macrodipole effect. On this basis, we analyzed charged triplets using a running five-residue window. The conclusion was that only mixed charged residues [positive (+) and negative (−)] located at positions 1–2–5 and 1–4–5 are clearly favored. In these locations the most abundant are (− − .. +) and (− .. + +), and this shows the existence of side chain microdipoles, which neutralize the large macrodipole of the helix.

We made a systematic statistical analysis of charged, dipolar, and hydrophobic + aromatic residues, which enabled us to work out rules that should be useful for modeling and design purposes.

Finally, we analyzed the relative abundance of all the different amphipathic double-arcs that are present in helices formed by octapeptides (8) and nonapeptides (18). All of the double-arcs that make up Schiffer and Edmundson's classical helical wheel are found in abundance in the sample.

Keywords: α-helix; helical patterns; hydrophilic residues; hydrophobic residues; medium-range interactions

Since Chou and Fasman's pioneering work (Chou & Fasman, 1974), it has been accepted that the bulk of consecutive amino acid residues within a polypeptide sequence, with a high average intrinsic propensity, defines the nucleus of secondary structure segments in a protein (α-helices, β-strands, and reverse turns). Under this assumption, a number of secondary structure predictive procedures, based on sets of propensities, have been worked out, all of which are merely refinements of the Chou and Fasman (1974) initial method. The secondary structure still cannot be accurately predicted because a number of factors derived from short- and medium-range interactions among neighboring residues and between residues and the solvent are not well understood, and therefore, they are not considered within the algorithms that are presently used for structural predictions.

Reprint requests to: Jaume Palau, Unitat de Biotecnologia Computacional, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Tarragona 43005, Catalonia, Spain; e-mail: palau@quimica.urv.es

Lotan et al. (1966) discovered the effect of hydrophobic side-chain interactions on stabilizing the α-helix. In addition, stereochemical approaches by Schiffer and Edmundson (1967) shed light on the architecture of α-helices, in the sense that polypeptide segments, when in helical conformation, tend to segregate hydrophobic and hydrophilic residues. Palau and Puigdomenech (1974) and Lim (1974a) found an accumulation of hydrophobic triplets at positions 1–2–5 and 1–4–5, which helped to stabilize α-helices. In a further contribution, Palau et al. (1982) extended the analysis of hydrophobic triplets at positions 1–2–5 and 1–4–5 in α-helices to the four main classes of protein domains (mainly alpha, mainly beta, alternating alpha/beta, and alpha + beta); from the Palau et al. (1982) results, it can be concluded that the 1–2–5 and 1–4–5 hydrophobic clustering in helices is a universal feature found in proteins, whatever their architecture may be. More recently, a number of authors have focused their attention on hydrophobic clustering in helices (Muñoz & Serrano, 1994; Padmanabhan & Baldwin, 1994a, 1994b; Creamer & Rose, 1995). Creamer and Rose (1995) studied stabilizing interactions by leucine triplets at various spac-

ings on a polyalanyl α-helix model; they found that in some triplets, the free energy of interaction is greater than the pairwise sums, because of an improvement in side-chain contacts.

In earlier reports, the characteristic distribution of hydrophobic and hydrophilic residues in different secondary structure segments was the basis on which methods for predicting secondary structure were tested (Lim, 1974b; Cid et al., 1982). More recently, the effect on helix formation of patterns of hydrophobic and nonhydrophobic side chains in protein sequences has been studied in different ways (Torgerson et al., 1991; Kamtekar et al., 1993; Vazquez et al., 1993a; 1993b; West & Hecht, 1995; Xiong et al., 1995). Torgerson et al. (1991) predicted quadrant orientations of amino acids in most α-helices, and reported that the template-predicted configurations closely match crystallographic data on α-helices. Vazquez et al. (1993b) reported the presence of favored or suppressed side-chain patterns within protein sequences in relation with α-helices and β-strands and also developed an α-helix predictor (Vazquez et al., 1993a), which was based on the identification of a longitudinal, hydrophobic strip-of-helix pattern. Kamtekar et al. (1993) described a successful general strategy for the de novo design of proteins based on sequence locations of hydrophobic and hydrophilic residues which caused polypeptide chains to collapse into globular α-helical folds. West and Hecht (1995) studied the binary patterning of polar and nonpolar amino acids, in order to get a better understanding of the design of new proteins, and Xiong et al. (1995), from the same research group, concluded that the major determinant for self-assembling oligomeric peptides is the polar/nonpolar periodicity throughout their amino acid sequence. Finally, recent papers focus on the use of amino acid patterns (Zhu & Blundell, 1996), or of binary word encoding (Kawabata & Doi, 1997) to improve protein secondary structure predictions.

As recently suggested by Kawabata and Doi (1997), it seems desirable to carry out statistical studies to obtain multiresidue information (i.e., information that depends on more than one residue). The main purpose of this information is to find out combinatorial features such as periodicity, residue pair interaction, and residue triplet interaction, as well as other undefined knowledge-based properties. This paper gives a comprehensive picture of triplet interactions within an α-helical pentapeptide, whatever the chemical characteristics of the amino acid residues may be. Statistically significant data about higher order polar/nonpolar binary patterns (in octapeptides and nonapeptides) are also provided. For this purpose, we imported a large database of nearly 4,000 helices present in 546 redundancy-controlled protein subunits from the Brookhaven PDB (Bernstein et al., 1977), and we grouped amino acid residues according to a number of chemical characteristics. Our results are important for a better understanding of side-chain relationships within an α-helix, and they lead to important rules concerning the stabilizing groupings that may be used to model α-helices.

Results

General and positional propensities of residues within an amino acid sequence may merely define (with uncertainties) the existence of an α-helix and some of its properties

We analyzed the amino acid composition of the whole α-helical sample (43,607 amino acid residues), as well as two subsamples formed by α-helices in the two domain groups of "mainly-alpha" and of "alternating alpha/beta (TIM barrels)." We used a simple program written in Fortran 77 (PERCENT). The results, shown in Figure 1, indicate that helical samples, if large enough, have an amino acid composition that does not vary much with the folding type of proteins. Our observation is coherent with evidence showing that the bulk of consecutive helical amino acid residues is responsible for the nucleation of helices (Chou & Fasman, 1974).

We also calculated the five positional amino acid compositions (grouped as indicated in the Rationale section) for the 28,448 pentapeptide windows, which slide across 3,863 α-helical segments. Although the set of consecutive pentapeptide windows overlap (which in principle should have a randomizing effect on the positional amino acid compositions, which, in turn, should make each of these compositions more similar to the composition of the whole sample), nonrandom overall tendencies can be seen, as is shown in Figure 2. In Figure 2A, G1 (positively charged residues) and G2 (negatively charged residues) percentages increase and decrease, respectively, from position 1 to position 5; in Figure 2B, the uncharged polar G3 and G4 groups prefer outside positions within the window; and in Figure 2C, G5 (aliphatic residues) and G6 (aromatic residues) prefer internal positions.

Pentapeptide grouping within α-helices reveals permissive triplet combinations formed either by hydrophilic or by hydrophobic residues

Figure 3 shows a complete set of patterns for combined charged residues. Figure 3A (or Fig. 3B) reveals that any triplet formed by three positively (or three negatively) charged residues remain below the mean value. These results indicate that such triplets are suppressed patterns, i.e., very uncommon within α-helices. However, Figure 3C shows that combinations of G1 and G2 (charged groups, formally represented by C) increase enormously for the patterns CC..C and C..CC. Of these patterns, 22..1 (i.e., −−..+) and 2..11 (i.e., −..++) are the most common ones (Fig. 3D,E).

Figure 4 shows a complete set of patterns for all those combined hydrophilic residues, except the patterns that are only charged already shown in Figure 3. Figures 4A and 4B show that dipolar residues from group G3 and from G4, separately, do not have a

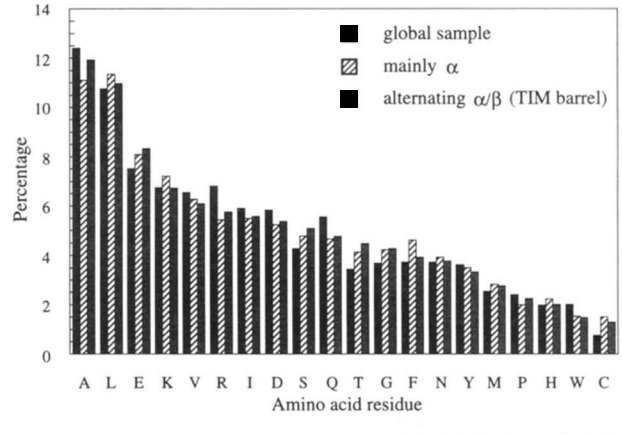

Fig. 1. Comparison of the amino acid composition of helical samples taken either from the whole sample or from particular foldings found in the Brookhaven PDB: ■ whole sample; ▨ mainly alpha folding group; ■ alternating alpha/beta folding group.

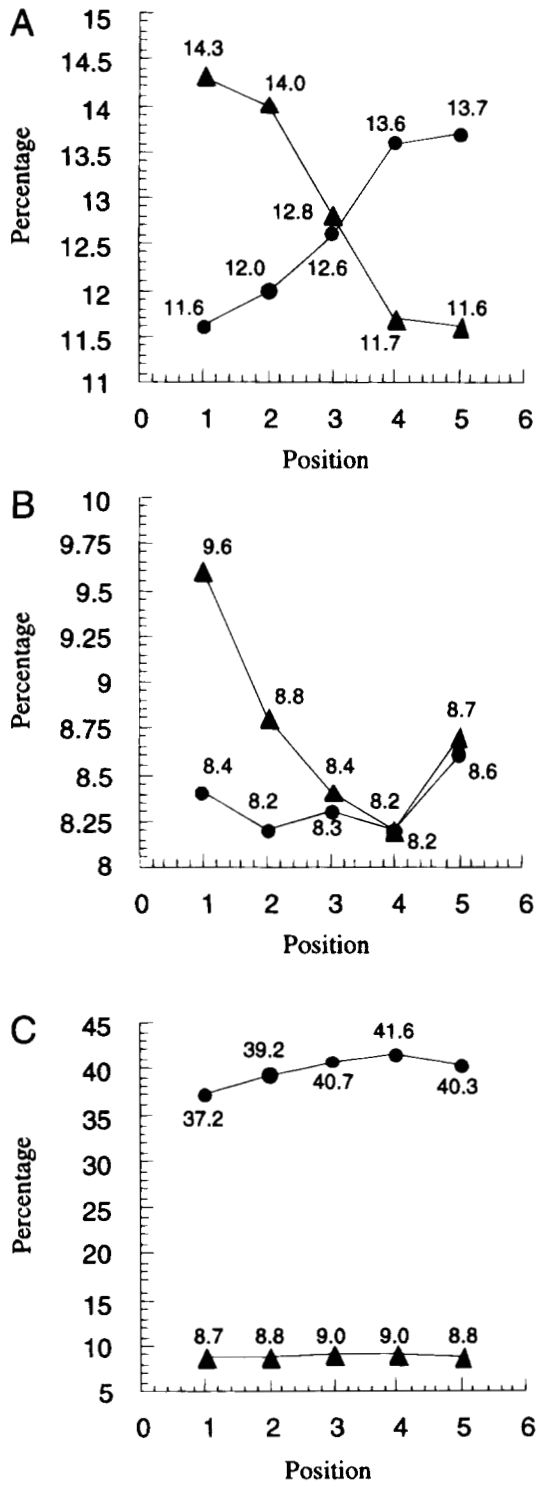

Fig. 2. Frequency of grouped amino acids as a function of the position within a pentapeptide window moving through all alpha-helical segments of the whole protein sample. A, G1 (Arg and Lys) ● and G2 (Asp and Glu) ▲. B, G3 (Asn and Gln) ● and G4 (Ser and Thr) ▲. C, G5 (Ala, Val, Leu, Ile and Met) ● and G6 (Phe, Tyr and Trp) ▲.

preferential pattern, with the exception of 3..33, 44..4, 4..44. However, combined dipolar residues -G3 + G4- (formally represented by D) have a spectacular difference between DD..D/D..DD and all the other patterns, the former being by far the most populated within the sample (Fig. 4C) and the pattern 34..4 being the most frequent (cf. Fig. 4D,E). Figure 4F–I shows triplet combinations of polar residues (P), which are made up simultaneously of C (1 or 2 charged ones) and D (1 or 2 dipolar ones). In general, all combinations are poorly represented in helices in comparison to the expected values. Two exceptions are worth mentioning: P..PP in Figure 4F (which contains elements from G1 and G3); and PP..P in Figure 4H (which contains elements from G2 and G3). Figure 4J shows that the patterns PP..P and P..PP (which contains elements from G1/G2/G3/G4) are more represented than the other patterns (PP.P., .PP.P, P.PP., .P.PP, PPP.., .PPP., ..PPP., P.P.P).

As shown in Figure 5, our statistical analysis in a very extensive sample of proteins validates and extends earlier results on stabilizing hydrophobic triplets using only a reduced number of proteins (Palau & Puigdomenech, 1974; Palau et al., 1982). Triplets 55..5, 5..55, 66..6, HH..H, and H..HH (H being a residue that belongs either to G5 or to G6) show large deviations from the expected values. It is worth mentioning, because it is the first time the result has been reported, that other triplet patterns such as .55.5, 5.55., .5.55, ..555, and .H.HH have deviations around 2σ or higher. This shows that there is an increasing presence of hydrophobic helices (i.e., nonamphipathic helices) in the PDB, sandwiched in the interior of protein domains. Positional permutations of elements in group G5 and G6 (i.e., in a ratio of two to one) for the HH..H, H..HH, and .H.HH triplet patterns do not enhance any particular combination (results not shown).

Octapeptide and nonapeptide groupings of hydrophilic and hydrophobic residues within α-helices reveal permissive and nonpermissive patterns

We define "helical double-arc pattern" as a concept that describes a helical multiresidue arrangement, which is antithetical in nature (i.e., residues of the same character are in the same half of the helix). Octapeptide and nonapeptide helices have 8 and 18 different double-arc patterns, respectively. None of the schematic arrangements are identical because the number of residues per turn of an α-helix is not an integer, but 3.6. It is interesting to note that a Schiffer and Edmundson (1967) helical wheel is made up of a combination of double-arc patterns. The question is whether all patterns can be used to make a helical wheel. Our very extensive set of protein α-helices from PDB enables us to give a statistical answer to such an open question.

Figure 6A shows four octapeptide patterns. Depending on the nature of the residue, either polar (P) or hydrophobic (H), for a given color (white or grey), there may be eight different helical double-arc patterns. Figures 6B and 6C show the statistics for the presence of these eight patterns within the whole sample of helices. In all cases, the positive deviations are highly significant. As Figure 7 shows, the helical double-arc octapeptide patterns are also favorable if instead of a P residue there is a C residue (Fig. 7A), and instead of an H residue there is a G5 residue (Fig. 7B). Exception appears to be the CC55C55C pattern, which is not significantly different from the theoretical mean. Many other tested combinations of eight consecutive residues with patterns that are different from helical double-arc patterns have negative deviations (results not shown).

The upper part of Figure 8A shows nine nonapeptide patterns. By proceeding in the same way as for octapeptides, we can see that there are 18 different helical double-arc patterns. As with octapep-

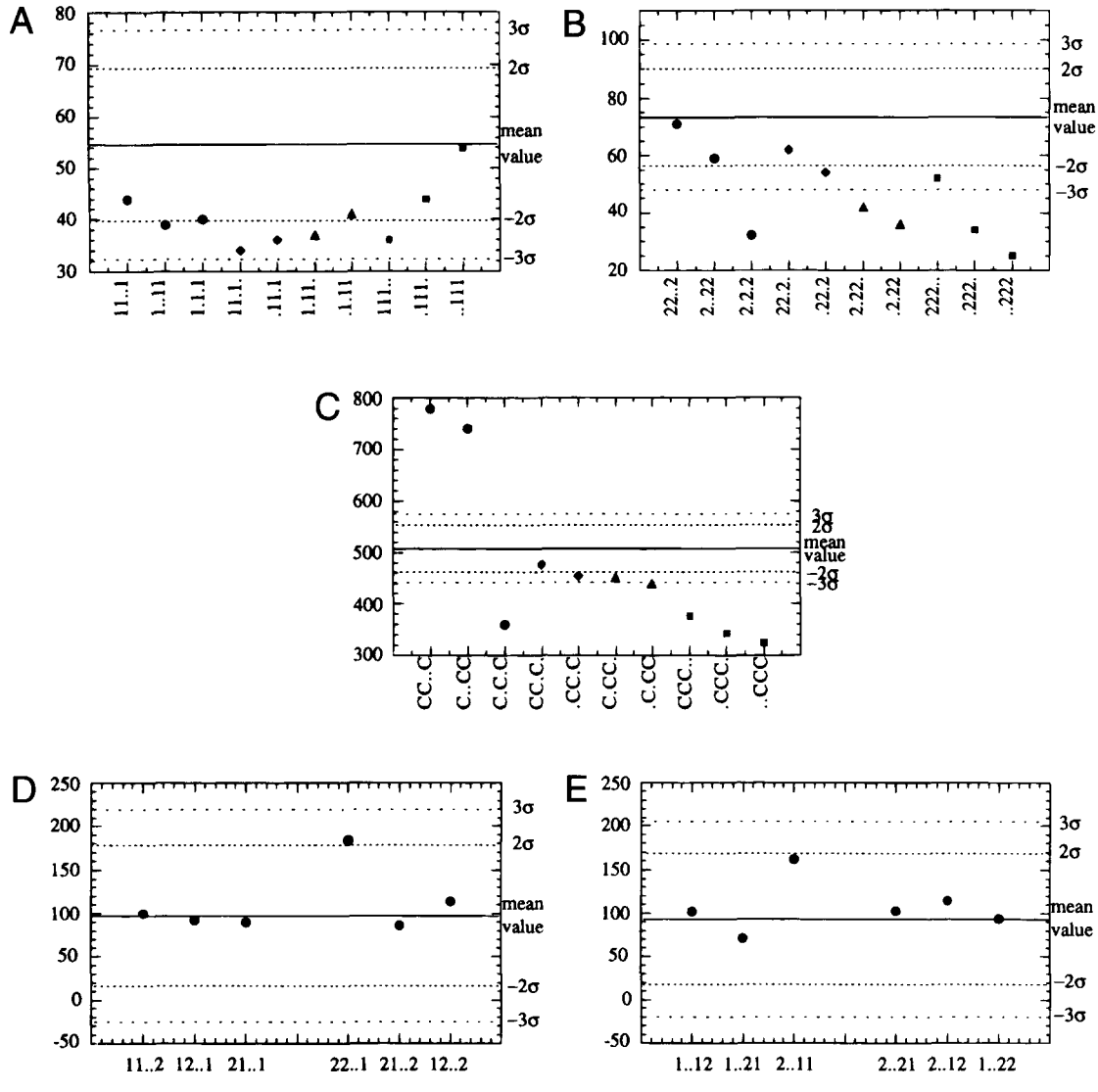

Fig. 3. Hydrophilic triplet frequencies found in α-helical regions from different groupings (first set). **A:** G1 (Arg and Lys). **B:** G2 (Asp and Glu). **C:** G1 + G2. **D:** Specific 1–2–5 subpatterns for all combinations of G1 and G2. **E:** Specific 1–4–5 subpatterns for all combinations of G1 and G2. Continuous lines show the statistical probability (mean value) of finding a given triplet, whereas the two kinds of broken lines "dense and spaced" indicate standard deviation thresholds of 2σ and 3σ, respectively. Equivalents patterns are represented by the same symbol: 1-2-3, 2-3-4, and 3-4-5 by ■; 1-2-4 and 2-3-5 by ◆; 1-3-4 and 2-4-5 by ▲. For the rest of the patterns, 1-2-5, 1-4-5, and 1-3-5 by ●.

tide patterns, when H and P residues are combined, all the statistical deviations are positive and highly significant (Fig. 8B,C). However, Figure 9 shows that combinations of H and C residues are the most plentiful. Exceptions appear to be the HHCCHCCCH and CCHHCHHCC patterns, although in these cases, the experimental statistical values are not very significant. As in the case of octapeptides, tested combinations of nine consecutive residues with patterns that are different from helical double-arc patterns have negative deviations (results not shown).

Discussion

Our analysis of the amino acid composition for a large sample of α-helices, not only for the undifferentiated whole sample, but also for the "mainly-alpha" and "alternating alpha/beta (TIM barrels)" samples, provides solid evidence against any definition of specific propensities related to domain classes, as has been proposed in the past (Palau et al., 1982; Chou, 1989). Since the Chou and Fasman (1974) method has a success rate of approximately 50% (in general) or higher (in "mainly-alpha" domains), there is no doubt that the method is useful for predicting the helical nucleation of most—although not all—sequence segments. However, other predictive methods can give greater accuracy in correctly predicting N- and C-terminal residues or in relation with definite pattern enchainments.

Argos and Palau (1982) compiled distributions of each amino acid at given positions within and around α-helical secondary structures, including the N- and C-terminal positions; Richardson and Richardson (1988) also studied amino acid preferences for specific

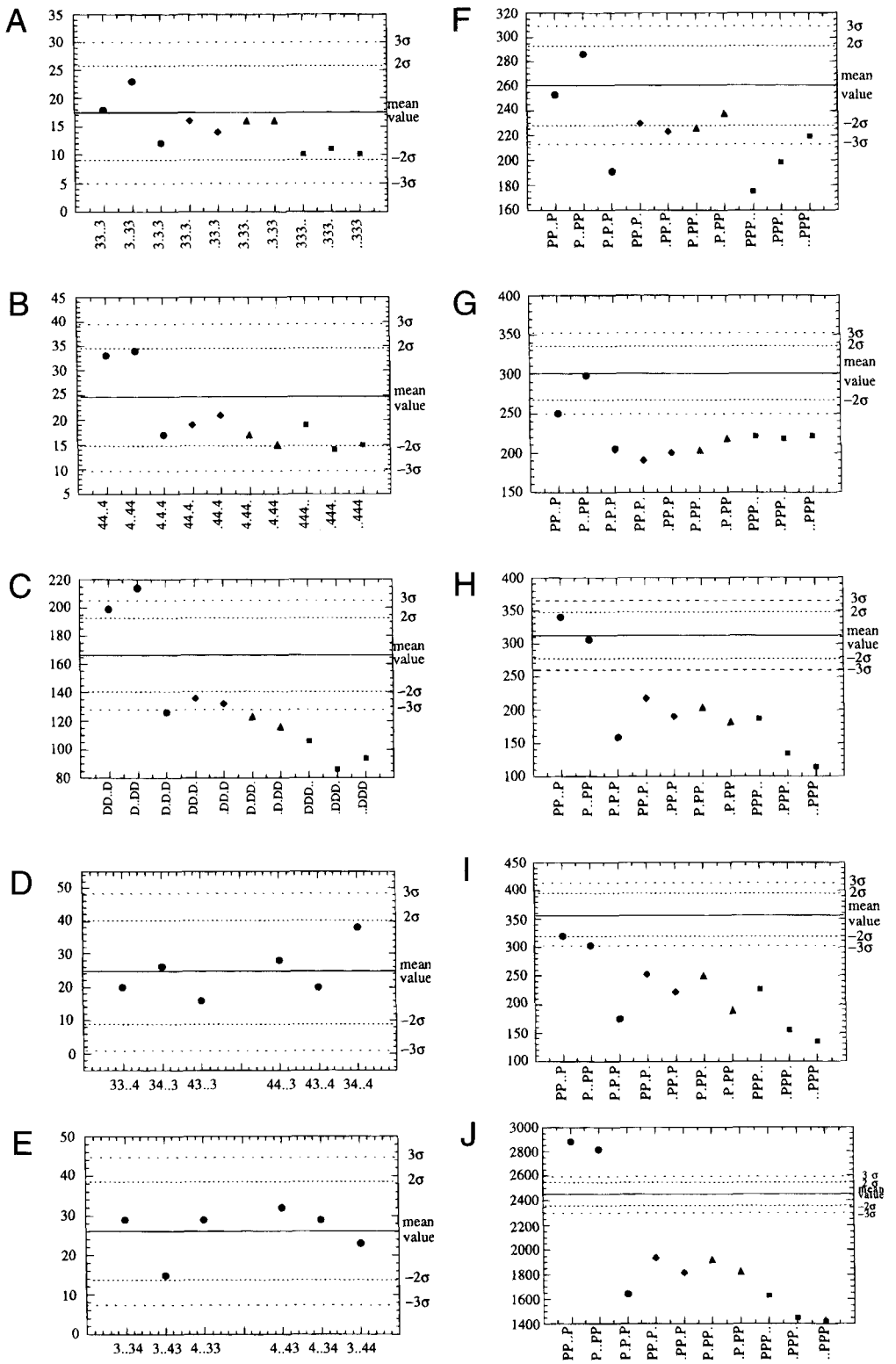

Fig. 4. Hydrophilic triplet frequencies found in α-helical regions from different groupings (second set). **A:** G3 (Asn and Gln). **B:** G4 (Ser and Thr). **C:** G3 + G4. **D:** Specific 1–2–5 subpatterns for all combinations of G3 and G4. **E:** Specific 1–4–5 subpatterns for all combinations of G3 and G4. **F:** G1 + G3 (Arg, Lys, Asn, and Gln). **G:** G1 + G4 (Arg, Lys, Ser, and Thr). **H:** G2 + G3 (Asp, Glu, Asn, and Gln). **I:** G2 + G4 (Asp, Glu, Ser, and Thr). **J:** G1 + G2 + G3 + G4. Continuous lines show the statistical probability (mean value) of finding a given triplet, whereas the two kinds of broken lines "dense and spaced" indicate standard deviation thresholds of 2σ and 3σ, respectively. Equivalents patterns are represented by the same symbol: 1-2-3, 2-3-4, and 3-4-5 by ■; 1-2-4 and 2-3-5 by ♦; 1-3-4 and 2-4-5 by ▲. For the rest of the patterns, 1-2-5, 1-4-5, and 1-3-5 by ●.

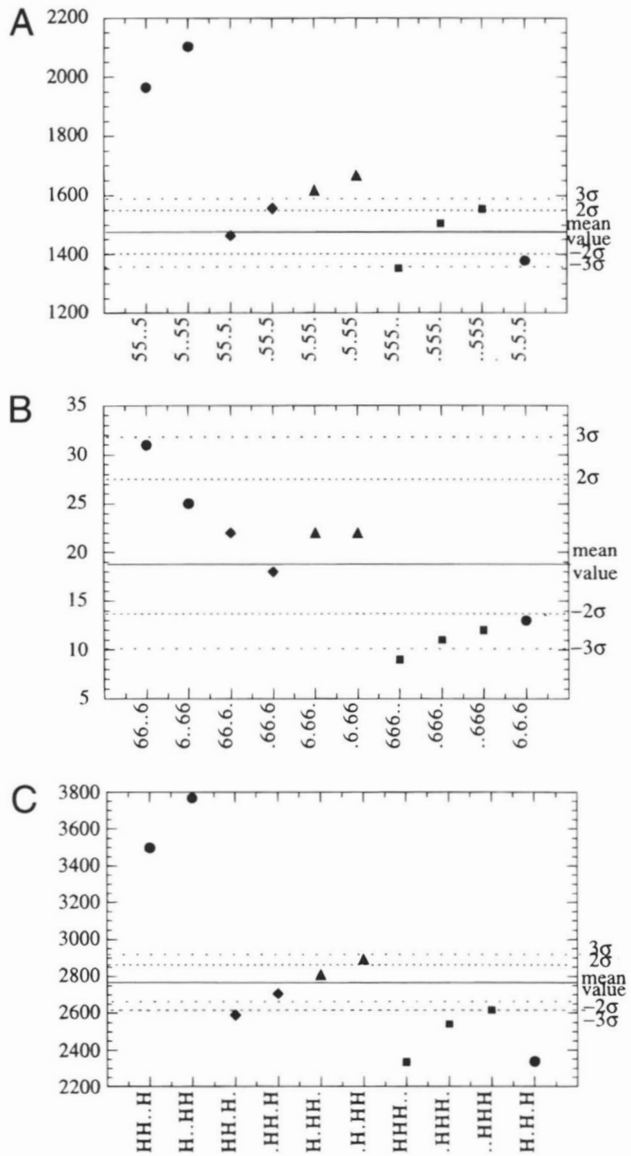

Fig. 5. Hydrophobic triplet frequencies found in α-helical regions from different groupings. **A:** G5 (Ala, Val, Leu, Ile, and Met). **B:** G6 (Phe, Tyr, and Trp). **C:** G5 + G6. Continuous lines show the statistical probability (mean value) of finding a given triplet, whereas the two kinds of broken lines "dense and spaced" indicate standard deviation thresholds of 2σ and 3σ, respectively. Equivalents patterns are represented by the same symbol: 1-2-3, 2-3-4, and 3-4-5 by ■; 1-2-4 and 2-3-5 by ◆; 1-3-4 and 2-4-5 by ▲. For the rest of the patterns, 1-2-5, 1-4-5, and 1-3-5 by ●.

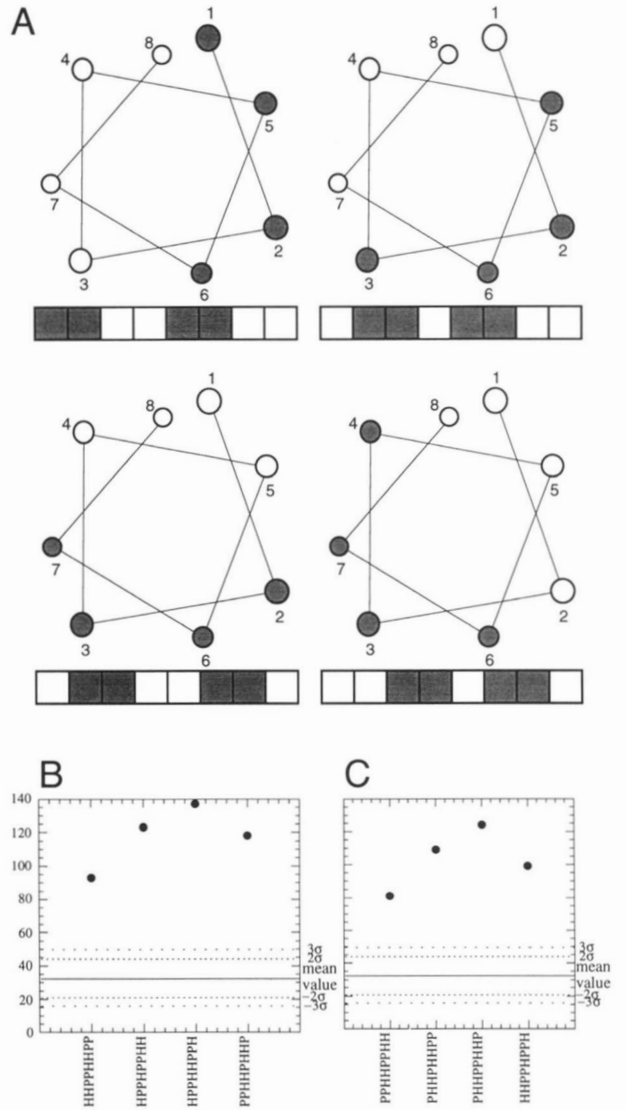

Fig. 6. A: Four helical wheels for octapeptide patterns combining polar (P = G1 + G2 + G3 + G4) (filled circles) and nonpolar (H = G5 + G6) residues (open circles). A second set of four helical wheels can be created if open circles represent P and filled circles represent H. For the sake of clarity, underneath each wheel a lineal scheme of the polar/nonpolar sequence is also shown. **B** and **C:** Statistical octapeptide frequencies are shown, in the same way that Figures 3, 4, and 5 show the triplets.

locations at the ends of α-helices in a subsequent data sample, and they found trends that were similar to those reported by Argos and Palau (1982). Our results on the positioning of amino acid distributions (Fig. 2) confirm that negatively and positively charged residues tend to accumulate inside the α-helix near the N- and C-cap residues, respectively, and they describe earlier results by Argos and Palau (1982). Such nonsymmetrical distribution of charged residues would help to neutralize the effect of the helical macrodipole. Robinson and Sligar (1993) determined for 4-helix bundle cytochrome *b*-562 from *E. coli* the contribution of indirect electrostatic effects of opposite charges located at the termini of adjacent anti-parallel α-helices, which simulate an anti-parallel pair; this system was the first experimental evidence for electrostatic interactions, such as those between partial charges, due to helix macrodipole charges. In addition to this previously observed effect (i.e., the existence of an accumulation of charges at the ends of helices), our results from a large sample of helices demonstrate that there is also an accumulation of charged pentapeptide segments, a large number of which are oriented in the form of microdipoles-oriented from negative to positive, throughout the length of the helix. To our knowledge, this is the first report on the existence of such microdipoles counterbalancing the action of the helical macrodipole.

Other tendencies shown in Figure 2, such as the preferential positioning of dipolar residues at the ends of pentapeptides and of

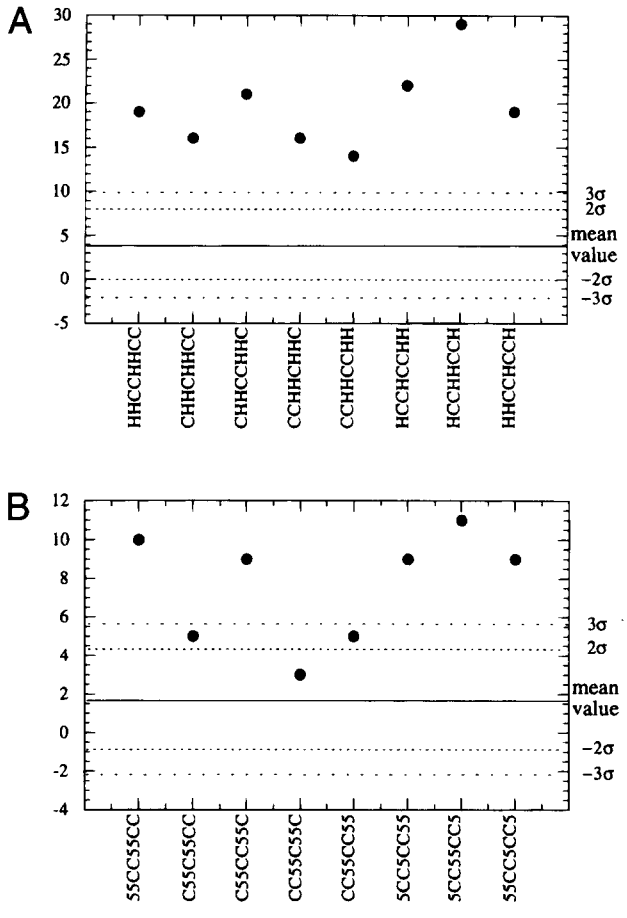

Fig. 7. Statistical frequencies for the two sets of octapeptide patterns presented in Figure 6. **A:** C is G1 + G2 and H is G5 + G6. **B:** C is G1 + G2, and 5 is G5.

hydrophobic residues in the middle, suggest that it would be appropriate to define propensities related to positions within and surrounding the α-helix (Chakrabartty et al., 1993; Doig & Baldwin, 1995). However, the fact that 3_{10}-helices are frequently located at both ends of α-helices, that there are numerous frayed ends on both sides and distortions in the middle of α-helices (Chakrabartty & Baldwin, 1995), and that the α-helices and 3_{10}-helices may undergo conformational transitions (Smythe et al., 1995) make those propensity measurements, which are to determine the limits of α-helices, very uncertain.

Our studies also give valuable clues about pattern distributions in pentapeptides (see Figs. 3, 4). In this respect, the following rules are formulated about polar residues (which we have called P rules): (1) triplet patterns formed exclusively of acid or basic residues are poorly represented; (2) however, if positive (G1) and negative (G2) residues are combined, charged triplets of the type CC..C and C..CC (but not the others) are much favored; (3) the 22..1 triplet with a CC..C pattern, as well as the 2..11 with C..CC pattern are the most frequent, which agrees with the hypothesis that the helical macrodipole is neutralized inside the helices, as discussed above; (4) triplet patterns formed exclusively of amide or hydroxylic residues are also poorly represented; (5) patterns DD..D and D..DD, formed by combining amide and hydroxylic residues (D), are the most common ones; and (6) polar patterns PP..P and P..PP, which are formed of a combination of acid, basic, amide and hydroxylic residues, are highly represented.

The following hydrophobic rules (H rules) are also derived from our results (see Fig. 5). (1) The hydrophobic triplet distribution is greatly enhanced for HH..H and for H..HH. This validates earlier results, which were obtained with fewer samples of proteins (Palau & Puigdomenech, 1974; Palau et al., 1982); (2) triplets, which are formed exclusively of aliphatic residues (G5), are greatly favored, not only for 55..5 and 5..55, but also for 5.55., .5.55, .55.5, and ..555; (3) triplets formed exclusively of aromatic residues (G6) are scarce, and statistical enhancement is only observed for 66..6; and (4) no particular combination of aliphatic and aromatic residues enhance values found for HH..H and for H..HH (results not shown). In principle, the enhanced triplets 5.55., .5.55, .55.5, and ..555 (and also .H.HH) were unexpected. However, there are two type of patterns: those combining positions 1–2–5 and 1–4–5, and those combining hydrophobic groups at different relative positions. In the latter case, one should expect these helices to be located in the interior of the protein tertiary structure, sandwiched in between other structures. As the PDB collection of protein structures increases, the existence of sandwiched helices is becoming more apparent [one recent example is the helices H4-5 and H8 in the ligand-binding domain of three different nuclear receptors (Wurtz et al., 1996)].

From our studies, it is evident that there is a hydrophilic/hydrophobic binary patterning. In accordance with other recent studies (Vazquez et al., 1993b; West & Hecht, 1995; Xiong et al., 1995), we define for the first time, two sets of statistical rules (P and H) that enable the hydrophilic and hydrophobic residues to be assembled in an ordered way. There are some exceptions to the P and H rules in proteins from living cells, but the "failing-rule" patterns are not common and may be explained of some folded domains and/or to certain medium- and long-range interactions within the protein scaffold.

In order to get greater insight into the reasons for the existence of binary patterning, we carried out statistical calculations on patterns formed of eight and nine residues (see Figs. 6–9); these patterns correspond to a segment of helix in between two and three turns. In both cases (see Figs. 6, 8) all possible combinations of polar and hydrophobic residues in the style of a Schiffer and Edmundson (1967) helical wheel are statistically favored at a very high level. Eight- or nine-residue constructs other than those corresponding to Schiffer and Edmundson (1967) helical wheels are poorly represented within the sample (results not shown).

We observed similar, but not identical, patterns when studying the different helical wheels and the relative distribution of hydrophilic and hydrophobic residues. This should be taken into account for design purposes. The P and H rules can be of great help in the design of a polypeptide chain that, in principle, should acquire an α-helix as secondary structure. However, there is no doubt that other more refined or specific rules may be formulated in the future when medium- and long-range interactions may be able to be considered. This is the aim of some of the work that is being carried out at present in our laboratory. However, the statistical scaling up of the triplet analysis presented in this paper, using differentiated amino acid residues, must still wait for some time. In addition, the binary patterning of octa- and nona-peptides are on the limits of statistical validation when only polar and nonpolar residues are used. Protein science is, in most cases, a knowledge-based science; therefore, we have to wait for a drastic expansion of the Brookhaven PDB before we can study patterns based only on single amino acid residues.

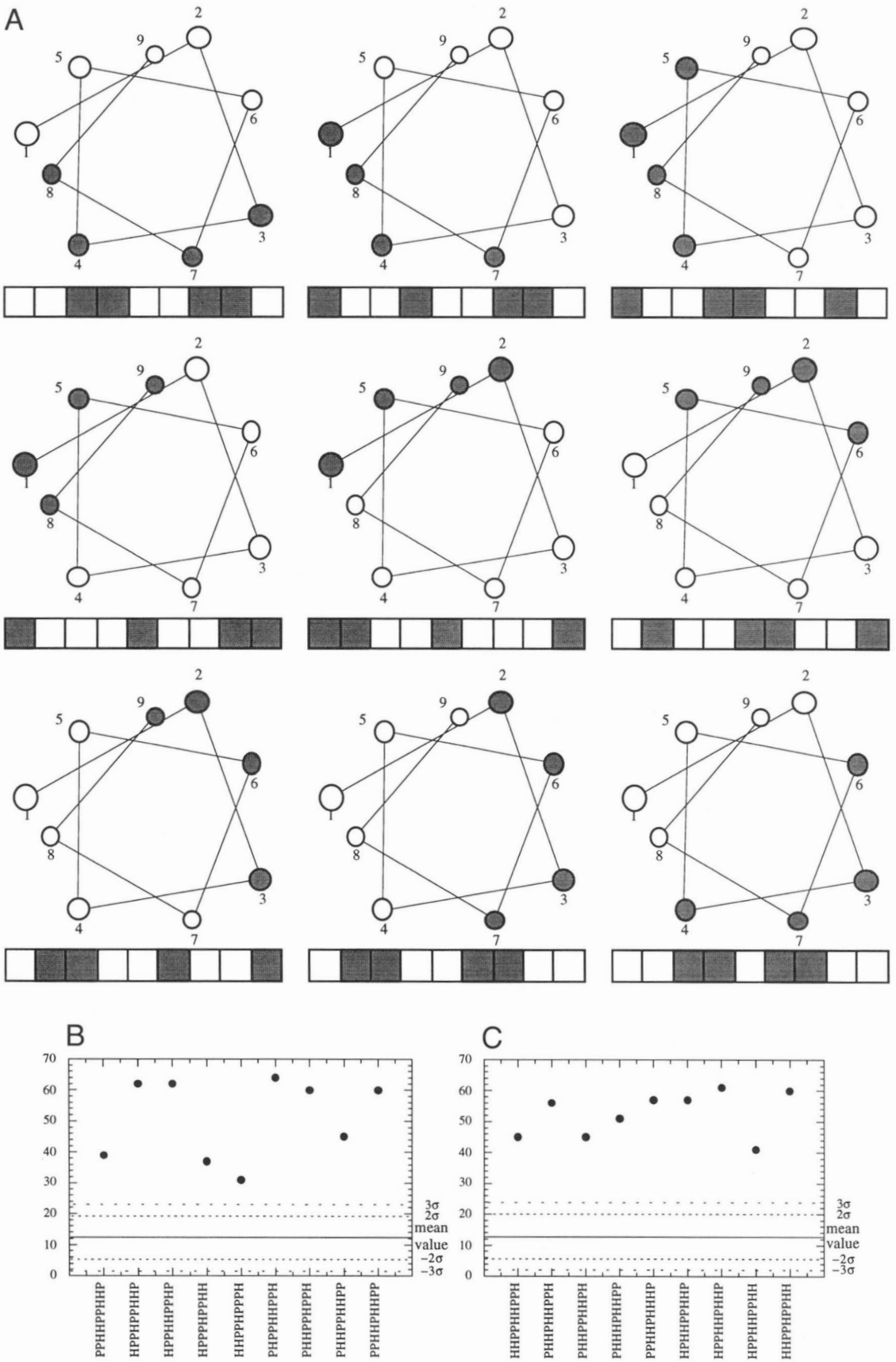

Fig. 8. **A:** Nine helical wheels for nonapeptide patterns combining polar (P = G1 + G2 + G3 + G4) (filled circles) and nonpolar (H = G5 + G6) residues (represented as open circles) is shown. A second set of nine helical wheels can be created if open circles represent P and filled circles represent H. For the sake of clarity, underneath each wheel a lineal scheme of the polar/nonpolar sequence is also shown. **B** and **C:** Statistical nonapeptide frequencies are shown, in the same way that Figures 3, 4, and 5 show the triplets.

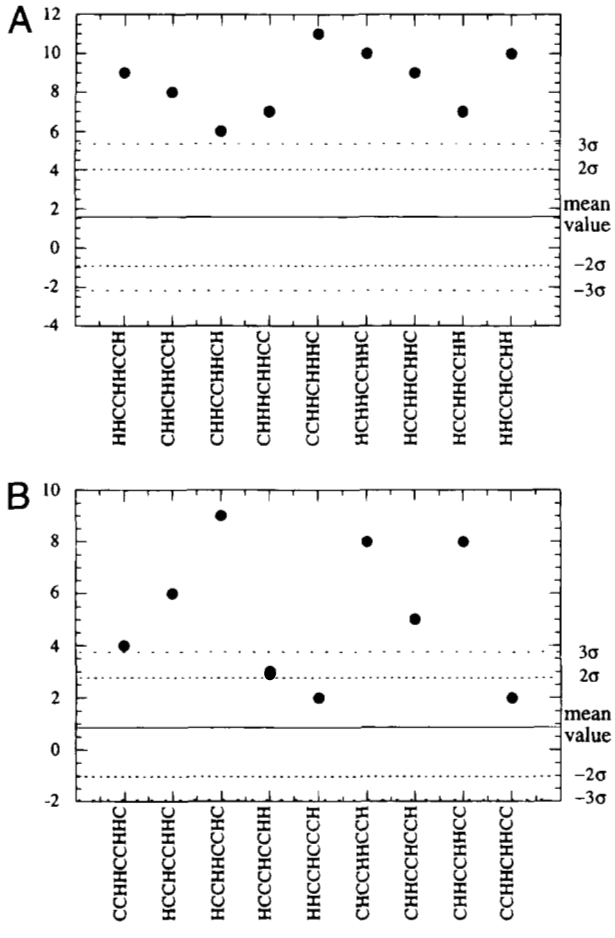

Fig. 9. **A** and **B:** Statistical frequencies for the two sets of nonapeptide patterns presented in Figure 8. In both cases, C is G1 + G2 and H is G5 + G6.

By inspecting constructs of polar and hydrophobic residues (see the schemes in Figs. 6, 8), the following charged subpatterns become apparent: (1) two single polar residues separated by two or three hydrophobic residues; (2) two pairs of joined polar residues separated by one or two hydrophobic residues; (3) one single polar residue and one pair of joined polar residues separated by two hydrophobic residues; and (4) one pair of joined polar residues and one single polar residue separated by two hydrophobic residues. Some of these patterns cover pentapeptide subpatterns that follow the rules P(2) (i.e., favored combined charged triplets are of the CC..C and C..CC types) and P(3) (i.e., the most frequent charged triplets belong to the 22..1 and 2..11 types), as defined above. Obviously, doublets of the type C...C and .C..C are found not only in (1), but also in (2), (3), and (4). The statistical abundance of C...C and .C..C within the octapeptides and nonapeptide constructs is the result of a linear combination of all four types of subpatterns. In the present work on charged triplets, we reveal the presence in helices of complex patterns which appear to have a physical meaning (i.e., counterbalancing the macrodipole, as oriented doublets of the type $-\ldots+$ and $.-..+$ may do, as well).

An interesting question arises when asking why $--..+$ and $-..++$ are statistically more frequent than other mixed charged triplets (those CC..C and C..CC triplets that cover all the variants formed by either two negative charges and one positive charge, or one negative charge and two positive charges, and in both cases are ordered from negative to positive). Although at present the scarcity of the PDB sample may make statistical calculations difficult, we have performed a preliminary work (J.A. Negrete & J. Palau, unpubl. obs.) that suggests that our results on triplets are coherent even at the level of individual charged residues: (1) DD..K, EE..K, ED..K, EE..R, DE..R, ED..R, D..KK., D..RR, D..KR, E..KK., E..RK, E..R., E..KR show positive deviations around 3σ or even higher; (2) DD..R, DE..K, and D..RK. show positive deviations of low significance; (3) E..EK, EK..K, and ER..K show positive deviations around 3σ or even higher; (4) D..EK, DK..K, D..DR, DK..R, E..DK, E..DR, ER..R, and ER..R show positive deviations of low significance; (5) D..DK, DR..K, D..ER, DR..R, and E..ER show negative deviations. Summing up: (a) all 16 patterns of the types $-..++$ or $--..+$ show positive deviations, and in 13 cases such deviations are of high significance; and (b) from the other 16 mixed charged patterns of the type C..CC or CC..C, three show positive deviations of high significance, eight show positive deviations of low significance, and five show negative deviations. The main conclusion is, therefore, that the supremacy of $-..++$ and $--..+$ patterns is not merely the result of combining a large number of charged and uncharged residues, but a general property for all 16 types of $-..++$ and $--..+$ triplets.

A preliminary analysis (J.A. Negrete & J. Palau, unpubl. obs.) shows that around 20% of the $-..++$ and $--..+$ triplet-containing patterns are included in constructs of the type [D,E][D,E].[K,R][K,R], which fulfill the requirements of Schiffer and Edmundson helical arcs that cover type 2 groupings (see two paragraphs above). About 80% of the $-..++$ and $--..+$ triplet containing patterns are found to be isolated. Our analysis also shows that the abundance of patterns $-..++$ and $--..+$ in our sample of helical segments is 4.2% and 4.8%, respectively. Considering some overlapping of both patterns (at least 20%) for a given helix, an estimation of the percentage of helices within the sample that contain triplets of the type $-..++$ and/or $--..+$ is around 6–7%. The first negative charge of $-..++$ and $--..+$ triplets across the helices is estimated to be 4.9% and 11.4% at the N_{cap}, 21.0% and 23.9% at the N_{cap} + 1, 14.8% and 6.0% at the N_{cap} + 2, 4.9% and 2.7% at the N_{cap} + 3, and 54.3% and 56.0% at the $\geq N_{cap}$ + 4, respectively.

Using a current molecular visualization program (see Rationale), we inspected several $-..++$ and $--..+$ triplet-containing patterns of some protein subunits, in an attempt to find structural facts that might afford some clues about the geometry that may govern the side-chain residue interactions. At this preliminary stage of our analysis, we can state that triads of oppositely charged groups (in our case, $-..++$ or $--..+$), distributed on the surface of the helical backbone, have diverse dielectric descriptions. As examples, we describe a few model schemes: (1) DE..R in helix ADELRRT (1oxy%) shows a moderate neighborhood between Asp_{477} and Arg_{481} (with charge distribution schemes in which the two nearest opposite partial charges are located at a distance of 7.76 Å), whereas Glu_{478} formal charges are opposite the Asp_{477} and Arg_{481} residues; (2) DD..K in helix TEELRVRLASHLRKL RKRLRDADDLQKRLAVYQA (1lpe%) shows an interdigitation of Lys_{157} between Asp_{153} and Asp_{154} (with the nearest opposite partial charges being located at distances of 3.76 Å and 5.28 Å, respectively); (3) EE..R in helix VEEMLRSDLALELDGAKNL REAIGYADSV (1bcfA) has a very strong interplay of two pairs of opposite partial charges between Arg_{88} and Glu_{84} (at distances of

Table 1. *List of the 546 PDB codes used in our analysis*[a]

1aak_	1aapA	1ab2_	1aboA	1ack_	1acp_	1adeA	1admA	1adn_	1adr_	1aep_
1agt_	1ain_	1akp_	1aky_	1aliA	1amg_	1aml_	1amp_	1amy_	1ang_	1antI
1aorA	1apa_	1aps_	1arv_	1ash_	1asu_	1aszA	1atlA	1atpE	1babB	1bam_
1bba_	1bbhA	1bcfA	1bco_	1bfmA	1bgc_	1bgh_	1bglA	1bia_	1bip_	1bmtA
1bn21	1bncB	1bnh_	1bovA	1bp2_	1brd_	1briC	1bvp1	1bw4_	1byb_	1c5a_
1cbg_	1cbn_	1cc5_	1ccr_	1cec_	1celA	1cfb_	1cfh_	1cgmE	1cgo_	1chc_
1chd_	1chl_	1cis_	1ckaA	1cksB	1clc_	1cmbA	1cnsA	1colA	1cot_	1cpy_
1crb_	1crl_	1cseI	1csh_	1csn_	1ctf_	1ctl_	1ctm_	1ctn_	1ctt_	1cus_
1cxc_	1cyg_	1cyj_	1cyo_	1cyv_	1d66A	1daaA	1dctA	1ddt_	1deaA	1dhr_
1dhx_	1dih_	1dlhA	1dlhB	1dmaB	1dpb_	1dppA	1dsbA	1dtr_	1dtx_	1dupA
1dyr_	1eca_	1eciB	1ecl_	1ecmA	1ede_	1edt_	1eft_	1ego_	1ehs_	1eny_
1epaB	1erd_	1erg_	1erl_	1erp_	1esc_	1esl_	1etc_	1f3g_	1fca_	1fcdA
1fcdC	1fct_	1fivA	1fkj_	1flp_	1fnc_	1fps_	1ftpA	1ftt_	1ftz_	1fxd_
1gal_	1garA	1gcb_	1gdd_	1gdhA	1gfd_	1ggtA	1ghc_	1glcG	1gln_	1glqA
1gluA	1gmfA	1gmpA	1gox_	1gpb_	1gpc_	1gph1	1gpmA	1gpr_	1gps_	1grj_
1gseA	1gsq_	1gta_	1gtrA	1han_	1har_	1hbq_	1hcnA	1hdcA	1hgeA	1hip_
1hjrA	1hks_	1hlb_	1hma_	1hmcB	1hmpA	1hmt_	1hmy_	1hnr_	1hph_	1hpi_
1hpm_	1hpt_	1hrhA	1hryA	1hsn_	1hstA	1htbA	1htmD	1htp_	1hueA	1hulA
1hurA	1hvd_	1hxn_	1hyhA	1hyp_	1ica_	1iceA	1iceB	1idm_	1idsA	1ifa_
1ifc_	1ifj_	1ikm_	1ilk_	1ilr1	1inp_	1irk_	1irl_	1iscA	1ithA	1kanA
1knb_	1kptA	1krt_	1lba_	1lcpA	1lct_	1ldm_	1leb_	1led_	1lenB	1lfaA
1lfb_	1lgaA	1lgr_	1lis_	1lki_	1lldA	1lmb3	1lpbB	1lpe_	1lpt_	1ltsA
1ltsC	1ltsD	1lvl_	1lxa_	1lybA	1lybB	1lyp_	1mat_	1mdaH	1mdkA	1mdyA
1mhcA	1mhlA	1mhlC	1mldA	1mle_	1mli_	1mls_	1mml_	1mmoB	1mmoD	1mmoG
1mnc_	1mngA	1mnp_	1mntA	1molA	1mrj_	1msc_	1msfC	1mup_	1nal1	1nar_
1ndh_	1ner_	1nfp_	1nhkL	1nhp_	1nif_	1npk_	1nrcA	1onc_	1opr_	1ora_
1ordA	1osa_	1oxa_	1oxy_	1paa_	1paz_	1pbe_	1pbn_	1pbxA	1pbxB	1pce_
1pch_	1pcl_	1pcrH	1pda_	1pdnC	1pfiA	1pfkA	1phg_	1phr_	1pht_	1pii_
1pil_	1pkm_	1pkp_	1plq_	1pls_	1pmlA	1pmy_	1pne_	1pnh_	1pnrA	1poc_
1pod_	1poxA	1ppi_	1ppt_	1prhA	1prr_	1prtA	1prtB	1prtD	1psdA	1psm_
1pspA	1ptf_	1ptq_	1ptx_	1put_	1pvc2	1pvc3	1pvc4	1pyaA	1pyiA	1pyp_
1qorA	1r1a1	1r69_	1rcb_	1rci_	1regX	1ret_	1rfbA	1ris_	1rpa_	1rpo_
1rtc_	1rtp1	1sacA	1safA	1sap_	1sat_	1sbp_	1smnA	1snc_	1spbP	1spf_
1spiA	1std_	1stu_	1svr_	1sxcA	1sxl_	1tahA	1tap_	1tca_	1thg_	1thtA
1thx_	1tif_	1tig_	1tin_	1tlfA	1tml_	1tph1	1tpt_	1trkA	1trrA	1trt_
1tsp_	1tssA	1tupB	1tys_	1ubi_	1udg_	1udpA	1ukz_	1utg_	1vcaA	1vhh_
1vil_	1vsgA	1was_	1wsyA	1wsyB	1xylA	1xyzA	1yptB	1yrnA	1yrnB	1ytbA
1zaaC	2abd_	2abk_	2acg_	2acq_	2ak3B	2apr_	2at2A	2ayh_	2azaA	2bbvC
2bgu_	2bltA	2bopA	2bpa1	2btfA	2cas_	2ccyA	2cdv_	2chr_	2chsA	2cpl_
2ctc_	2cy3_	2cyp_	2cyr_	2dkb_	2dln_	2dnjA	2dri_	2ebn_	2end_	2fal_
2fcr_	2fx2_	2fxb_	2gbp_	2gdm_	2glt_	2gstA	2hbg_	2hft_	2hhmA	2hmx_
2hmzA	2hnq_	2hpdA	2hpqP	2hsp_	2hts_	2ifo_	2kauA	2kauB	2kauC	2lhb_
2liv_	2mev1	2mev2	2mev3	2mnr_	2nacA	2olbA	2omf_	2pcdA	2pcdM	2pde_
2pec_	2pgd_	2phy_	2pia_	2pleA	2plv1	2pnb_	2por_	2prd_	2prk_	2ptl_
2rn2_	2rspA	2sas_	2sblB	2scpA	2sn3_	2spcA	2stv_	2tbvA	2tgi_	2tmdA
2tmvP	2trxA	2uce_	3aahA	3aahB	3c2c_	3cd4_	3chy_	3cox_	3dfr_	3gapB
3grs_	3ladA	3mddA	3pgk_	3pgm_	3pmgA	3pte_	3rubL	3rubS	3sdhA	3sgbI
3sicI	4dfrA	4enl_	4fxn_	4gcr_	4htcI	4icb_	4mt2_	4rhv1	4rhv3	5rubA
5znf_	7apiA	7icd_	7pti_	7rsa_	8abp_	8acn_	8atcA	8atcB	8catA	8dfr_
8tlnE	9rnt_	121p_	131l_	193l_	256bA	451c_				

[a]All proteins were selected with a resolution of 3 Å or less.

2.77 Å and 2.89 Å, which account for strongly coupled groups); etc.

At present, local electrostatic effects caused by charge-charge, charge-solvent, and side-chain–backbone interactions are very difficult to describe in the form of reliable models incorporating continuum electrostatic or dielectric descriptions. From the topological point of view, the charge distribution schemes for $-..++$ and $--..+$ triplet-containing patterns appear to be kaleidoscopic, and need to be studied more deeply in order to find some geometrical clues. Our group is now engaged in the task of gaining insight into such clues by comparing local models found in PDB proteins (J.A. Negrete & J. Palau, unpubl. obs.) and using our specialized rotamer library for α-helices (G. Pujadas & J. Palau, unpubl. obs.).

Rationale

We imported a set of 546 nonredundant protein subunits (homology less than 45%) from the Brookhaven PDB_Select with a res-

olution of 3 Å or less (ftp address: ftp.embl_heidelberg.de/pub/databases/pdb_select). At this resolution, the secondary structure limits are well defined, and therefore, they were taken from the PDB definition, on the basis that none of them had any ambiguously defined residues such as UNK, GLX. We also checked helices in order to eliminate those with any missing residues. When necessary, the molecular visualization program RasMol vs. 2.6 (Sayle, 1996) was used to control the overall quality of helices. Since no statistical study was specifically performed on N- and C-terminal ends, fried ends were subject to no special checking. A list of PDB codes for these proteins is given in Table 1. From this protein domain subbank, 3,863 α-helical segments were selected on the basis that none of these segments should contain fewer than five residues (average size 11.3 residues per segment). All the possible consecutive pentapeptides (28,448 units, 43,607 amino acid residues) were worked out from the helical sample and processed by using the program PATTERNS written in Fortran 77.

According to their physico-chemical similarities, the amino acid residues were clustered into six groups: G1 (Arg and Lys, 5,418 residues); G2 (Asp and Glu, 5,980 residues); G3 (Asn and Gln, 3,707 residues); G4 (Ser and Thr, 4,154 residues); G5 (Ala, Val, Leu, Ile, and Met, 16,257 residues); G6 (Phe, Tyr, and Trp, 3,798 residues). In order to avoid statistical dispersion, the remaining miscellaneous residues, grouped as G7 (Gly, Pro, Cys, and His, 4,293 residues), were not considered in our studies. Higher order groupings (G1/G2, 11,398 charged residues; G1/G3, 9,125 basic plus amide residues; G2/G3, 9,687 acid plus amide residues; G1/G4, 9,572 basic plus hydroxylic residues; G2/G4, 10,134 acid plus hydroxylic residues; G3/G4, 7,861 dipolar residues; G1/G2/G3/G4, 19,259 charged plus dipolar residues; G5/G6, 20,055 aliphatic plus aromatic residues) were also studied. Although the positional amino acid composition analysis on sliding pentapeptides could also be carried out with the 20 amino acid residues, for reasons of coherence we kept the same physico-chemical groupings.

All possible triplets within a pentapeptide were considered (1–2–3, 1–2–4, 1–2–5, 1–3–4, 1–3–5, 1–4–5, 2–3–4, 2–3–5, 2–4–5, and 3–4–5). Stereochemically, some patterns are equivalent (1–2–3, 2–3–4, and 3–4–5; 1–2–4 and 2–3–5; 1–3–4 and 2–4–5) and should, in principle, give the same results. The three triplet residues were chosen from a single group, a combination of two groups or, in a few cases, a combination of several groups.

The theoretical probability of finding characteristic triplets in the pentapeptide sample, q, was calculated as follows:

$$q = (N_G/N_T)^3 \quad (1)$$

where N_G is the total number of residues belonging to a group (or combination of groups), and N_T the total number of residues in the helix database.

The experimental frequency for all the triplets was calculated by counting their occurrence within α-helices, with a pentapeptide window moving along the 3,863 helices. A statistical test was used to study the significance of deviations of the triplet experimental frequencies with respect to the theoretical probabilities. For this purpose the normal distribution was regarded as a binomial distribution. The mean value for a given triplet and its standard deviation σ_G is, respectively:

$$M = q * N_s \quad \text{and} \quad \sigma_G = [q(1 - q)N_s]^{1/2} \quad (2)$$

where N_s is the total number of pentapeptides in the helix database (28,448 units). In a binomial distribution, experimental deviations from M of $1.645\sigma_G$, $1.960\sigma_G$, and $2.576\sigma_G$ mean that they are 95.0%, 97.5%, and 99.5% certain of not being simply statistical.

If a pattern formed by two different groups is seen to be stabilizing, we analyze it to find out if the residues in each group occupy a specific place (there are six different arrangements or subpatterns in a specific pattern, formed by doing all combinations considering that one group is placed in two positions of the triplet and the other group is placed in the third position). In this analysis, we calculate the theoretical mean as a function of the number of residues of each group: if the number of residues in each group, for example a and b, is similar, the theoretical mean is calculated as the mean of all values (occurrences of the six subpatterns), and the standard deviation as the standard deviation of all these values; on the other hand, if the groups have different numbers of residues, we divide the subpatterns in two: the subpatterns formed by two a residues and one b residue, and the subpatterns formed by two b residues and one a residue. For each divided pattern, the mean and the standard deviations are calculated as has been shown above for the subpatterns with the same number of residues.

The statistical rationale for octapeptide and nonapeptide patterns was the same as the one described above for pentapeptides.

Acknowledgments

We thank general facilities from our university, although this work has not been awarded grants by any research-supporting institution. One of us (J.A.N.) thanks Gerard Pujadas for the training in Bioinformatics received during the initiation period. We also thank Prof. Enric Querol for computational facilities and John Bates (English Text & Style Correction Service from our university) for his help during the manuscript writing process.

References

Argos P, Palau J. 1982. Amino acid distribution in protein secondary structures. *Int J Pept Protein Res 19*:380–393.

Bernstein F, Koetxle T, Williams G, Meyer E, Brice M, Rodgers J, Kennard O, Shimannouchi T, Tasumi M. 1977. The protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Chakrabartty A, Baldwin RL. 1995. Stability of alpha-helices. *Adv Protein Chem 46*:141–176.

Chakrabartty A, Doig AJ, Baldwin RL. 1993. Helix capping propensities in peptides parallel those in proteins. *Proc Natl Acad Sci USA 90*:11332–11336.

Chou PY. 1989. Prediction of protein structural classes from amino acid composition. In: Fasman GD, ed. *Prediction of protein structure and the principles of protein conformation*. New York: Plenum Press. pp 549–586.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry 13*:222–245.

Cid H, Bunster M, Arriagada E, Campos M. 1982. Prediction of secondary structure of proteins by means of hydrophobicity profiles. *FEBS Lett 150*:247–254.

Creamer TP, Rose GD. 1995. Interaction between hydrophobic side chains within α-helices. *Protein Sci 4*:1305–1314.

Doig AJ, Baldwin RL. 1995. N- and C-capping preferences for all 20 amino acids in α-helical peptides. *Protein Sci 4*:1325–1336.

Kamtekar S, Schiffer JM, Xiong HY, Babik JM, Hecht MH. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science 262*:1680–1685.

Kawabata T, Doi J. 1997. Improvement of protein secondary structure prediction using binary word encoding. *Proteins 27*:36–46.

Lim VI. 1974a. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J Mol Biol 88*:857–872.

Lim VI. 1974b. Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J Mol Biol 88*:873–894.

Lotan N, Yaron A, Berger A. 1966. The stabilization of the α-helix in aqueous solutions by hydrophobic side chain interaction. *Biopolymers 4*:365–368.

Muñoz V, Serrano L. 1994. Elucidating the folding problem of helical peptides using empirical parametres. *Nat Struct Biol 1*:399–409.

Padmanabhan S, Baldwin RL. 1994a. Helix-stabilizing interaction between tyrosine and leucine or valine when the spacing is *i*, *i* + 4. *J Mol Biol 241*: 706–713.

Padmanabhan S, Baldwin RL. 1994b. Tests for helix-stabilizing interactions between various nonpolar side chains in alanine-based peptides. *Protein Sci 3*:1992–1997.

Palau J, Argos P, Puigdomenech P. 1982. Protein secondary structure. Studies on the limits of prediction accuracy. *Int J Pept Protein Res 19*:394–401.

Palau J, Puigdomenech P. 1974. The structural code for proteins: Zonal distribution of amino acid residues and stabilization of helices by hydrophobic triplets. *J Mol Biol 88*:457–469.

Richardson JS, Richardson DC. 1988. Amino acid preferences for specific locations at the ends of alpha helices. *Science 240*:1648–1652.

Robinson CR, Sligar SG. 1993. Electrostatic stabilization in four-helix bundle proteins. *Protein Sci 2*:826–837.

Sayle R. 1996. RasMol vs. 2.6. Glaxo Research & Development; ros@dcs.ed.ac.uk.

Schiffer M, Edmundson AB. 1967. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys J 7*:121–135.

Smythe ML, Huston SE, Marshall GR. 1995. The molten helix: Effects of solvation on the alpha- to 3_{10}-helical transition. *J Am Chem Soc 117*:5445–5452.

Torgerson RR, Lew RA, Reyes VE, Hardy L, Humphreys RE. 1991. Highly restricted distributions of hydrophobic and charged amino acids in longitudinal quadrants of alpha-helices. *J Biol Chem 266*:5521–5524.

Vazquez SR, Kuo DZ, Salomon M, Hardy L, Lew RA, Humphreys RE. 1993a. Prediction of alpha-helices in proteins with the hydrophobic strip-of-helix template and distributions of other amino acids around the hydrophobic strip. *Arch Biochem Biophys 305*:448–453.

Vazquez S, Thomas C, Lew RA, Humphreys RE. 1993b. Favored and suppressed patterns of hydrophobic and nonhydrophobic amino acids in proteins sequences. *Proc Natl Acad Sci USA 90*:9100–9104.

West MW, Hecht MH. 1995. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci 4*:2032–2039.

Wurtz JM, Bourguet W, Renaud JP, Vivat V, Chambon P, Moras D, Gronemeyer H. 1996. A canonical structure for the ligand-binding domain of nuclear receptors. *Nat Struct Biol 3*:87–94.

Xiong HY, Buckwalter BL, Shieh HM, Hecht HM. 1995. Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc Natl Acad Sci USA 92*:6349–6353.

Zhu ZY, Blundell TL. 1996. The use of amino acid patterns of classified helices and strands in secondary structure prediction. *J Mol Biol 260*:261–276.