Protein Science (1995), *4*:2032–2039. Cambridge University Press. Printed in the USA.

Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins

MICHAEL W. WEST AND MICHAEL H. HECHT
Department of Chemistry, Princeton University, Princeton, New Jersey 08544-1009

(RECEIVED June 14, 1995; ACCEPTED July 26, 1995)

Abstract

Protein sequences can be represented as binary patterns of polar (○) and nonpolar (●) amino acids. These binary sequence patterns are categorized into two classes: Class A patterns match the structural repeat of an idealized amphiphilic α-helix (3.6 residues per turn), and class B patterns match the structural repeat of an idealized amphiphilic β-strand (2 residues per turn). The difference between these two classes of sequence patterns has led to a strategy for de novo protein design based on binary patterning of polar and nonpolar amino acids. Here we ask whether similar binary patterning is incorporated in the sequences and structures of natural proteins. Analysis of the Protein Data Bank demonstrates the following. (1) Class A sequence patterns occur considerably more frequently in the sequences of natural proteins than would be expected at random, but class B patterns occur less often than expected. (2) Each pattern is found predominantly in the secondary structure expected from the binary strategy for protein design. Thus, class A patterns are found more frequently in α-helices than in β-strands, and class B patterns are found more frequently in β-strands than in α-helices. (3) Among the α-helices of natural proteins, the most commonly used binary patterns are indeed the class A patterns. (4) Among all β-strands in the database, the most commonly used binary patterns are not the expected class B patterns. (5) However, for solvent-exposed β-strands, the correlation is striking: All β-strands in the database that contain the class B patterns are exposed to solvent. (6) The bias of class A patterns for α-structure over β-structure and the bias of class B patterns for β-structure over α-structure are significant, not merely when compared to other binary patterns of polar (○) and nonpolar (●) amino acids, but also when compared to the full range of sequences in the database. The implications for the design of novel proteins are discussed.

Keywords: binary code; de novo proteins; polar/nonpolar patterns; protein design

Globular, water-soluble proteins invariably fold into structures that bury extensive hydrophobic surface area and expose polar side chains to solvent (Creighton, 1993). This observation, coupled with extensive experimental and theoretical work has supported the view that burial of hydrophobic surface area is the dominant force in protein folding (Kauzmann, 1959; Dill, 1990).

If a protein partitions polar and nonpolar residues between its surface and core, and if it simultaneously forms abundant secondary structure, then its amino acid sequence should be capable of forming α-helices and/or β-strands that are amphiphilic. In order to form an α-helix or a β-strand with one hydrophobic face and one hydrophilic face, the linear sequence of amino acids must possess a pattern of polar and nonpolar residues matching the structural repeat for that type of secondary structure (Kaiser & Kezdy, 1983; DeGrado & Lear, 1985; Bowie & Sauer, 1989; Bowie et al., 1990; Xiong et al., 1995). Indeed, several studies of known protein structures have shown the linear sequences of α-helical regions tend to have a hydrophobic periodicity of 3 or 4, in agreement with the structural repeat of 3.6 residues/turn typical of α-helices. Likewise, the linear sequences of β-strands in known structures often have hydrophobic periodicities of 2, in agreement with the structural repeat of β-strands (Eisenberg et al., 1984).

Recently, we demonstrated that the sequence pattern of polar and nonpolar amino acids is important not only for the structures of natural proteins, but also can serve as a cornerstone for the design of proteins de novo (Kamtekar et al., 1993). This patterning was used to develop a "binary code" for protein design, wherein individual sequence positions are specified as either polar or nonpolar (i.e., the code is binary), but the precise identity of each residue is allowed to vary. By using combinatorial methods to generate a large collection of amino acid sequences, we showed that the relatively simple information in the "binary code" is sufficient for the design of enormous numbers of de novo proteins that fold into compact α-helical structures (Kamtekar et al., 1993). More recently, we have attempted to use the

Reprint requests to: Michael H. Hecht, Department of Chemistry, Princeton University, Princeton, New Jersey 08544-1009; e-mail: hecht@berthaw.princeton.edu.

binary code to design de novo β-sheet proteins as well (M.W. West & M.H. Hecht, unpubl. results).

Having used a binary code to design de novo proteins, we were curious to see whether nature has used a similar binary code to design native proteins. Thus, we wished to determine whether the amino acid sequences of known protein structures conform to the binary patterns of polar and nonpolar residues used in our designed proteins.

Previous work by Humphreys and coworkers showed that in the sequences of 48 known protein structures, the amino acids Leu, Ile, Val, Phe, and Met are distributed in both favored and suppressed patterns (Vazquez et al., 1993). In our current work, we search a much larger collection of proteins and focus not merely on the presence or absence of these five amino acids but more explicitly on the binary patterns of polar (Arg, Lys, Asp, Glu, Asn, Gln, and His) and nonpolar (Phe, Leu, Ile, Met, Val) residues similar to those patterns used in our design of de novo proteins. For each occurrence of a particular binary sequence pattern, we ascertain whether the sequence occurs in α-helical, β-sheet, or "other" secondary structure. By comparing the actual secondary structure of a particular stretch of polypeptide chain with the structure expected from the binary code design strategy, we assess whether particular binary sequence patterns are favored in particular types of secondary structure. Analysis of the frequency of binary patterns in natural α-helices and β-strands enables us to assess whether the design of natural proteins is consistent with the binary code strategy.

Results and discussion

A binary code of polar and nonpolar amino acids

The binary pattern of polar and nonpolar amino acids consistent with the design of a four-helix bundle is shown in Figure 1A as a helix wheel representation (Schiffer & Edmundson, 1967). This pattern specifies amphiphilic α-helices and was used in our design of a large collection of de novo α-helical proteins (Kamtekar et al., 1993). The helix wheel pattern of a single amphiphilic α-helix can be represented as a linear sequence by the following binary pattern of polar (○) and nonpolar (●) residues: ○●○○●●○○●○○●●○. This sequence encompasses seven different pentapeptide sequence patterns: ○●○○●, ●○○●●, ○○●●○, ○●●○○, ●●○○●, ●○○●○, and ○○●○○. We will refer to these seven pentapeptide patterns as class A binary sequence patterns.

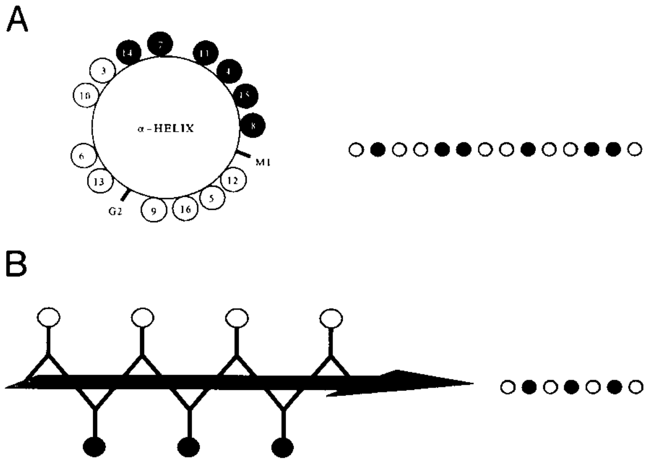

Fig. 1. A: Helix wheel representation of an amphiphilic α-helix. The sequence is the same as that used by Kamtekar et al. (1993) for the design of de novo four-helix bundle proteins. The binary code of amino acids is represented as ○, polar residues Asp, Asn, Glu, Gln, Lys, His, and Arg; ●, nonpolar residues Phe, Leu, Ile, Met, and Val. **B:** Schematic representation of an amphiphilic β-strand.

In contrast to this α-helical pattern, the binary pattern of polar and nonpolar residues consistent with an amphiphilic β-strand (Fig. 1B) is much simpler: It is merely an alternating sequence of polar (○) and nonpolar (●) residues. Thus, only two binary pentapeptide patterns are required for the de novo design of amphiphilic β-strands. These two patterns, which we will call class B sequence patterns, are ○●○●○ and ●○●○●.

Within these sequence patterns, polar (○) and nonpolar (●) amino acids were classified by two criteria: (1) the hydrophobicity scale of Fauchere and Pliska (1983), which measures the partitioning of amino acids between octanol and water; and (2) the binary organization of polar and nonpolar residues in the genetic code. In general, these two criteria led to the same choices of polar (○) and nonpolar (●) residues. The residues included in the polar (○) category are Arg, Lys, Asp, Glu, Asn, Gln, and His. These are the seven most polar residues in octanol/water partitioning experiments (Fauchere & Pliska, 1983). Furthermore, all of these residues, except for Arg, are encoded by the degenerate codon NAN (where N represents any of the four bases, except T is excluded from the first position to prevent stop codons or tyrosines). Thus, these polar amino acids are readily substituted for one another in natural proteins and are accessible to combinatorial genetic methods for the design of de novo proteins.

The residues included in the nonpolar (●) category are Phe, Leu, Ile, Met, and Val. These are the residues encoded by the degenerate codon NTN (where N represents any of the four bases). In the hydrophobicity scale of Fauchere and Pliska (1983), Trp and Cys are also hydrophobic. However, they occur rarely in our database, and an initial search showed our results were not significantly affected by the inclusion of Trp and Cys. Ultimately, we chose to exclude Trp and Cys from our study because they play ambiguous roles: Trp has an indole NH capable of forming hydrogen bonds and, thus, cannot simply be included into the hydrophobic core of a protein without consideration of possible H bonding partners. Cys can form disulfide bonds and, thus, plays a structural role more complex than the simple binary decision of polar versus nonpolar. The remaining five nonpolar (●) residues (Phe, Leu, Ile, Met, Val) are the same as those used by Humphreys and coworkers (Vazquez et al., 1993) in their search for favored and suppressed patterns of hydrophobic amino acids.

By restricting our study to these seven polar (○) and five nonpolar (●) residues and excluding residues in the middle of the hydrophobicity scale, we explore the possibilities of protein design with a simplified code of amino acids.

Class A patterns are common and class B patterns are rare in the linear sequences of native proteins

We surveyed a database of 177 protein structures comprising a combined sequence of 40,667 residues (for a description of the

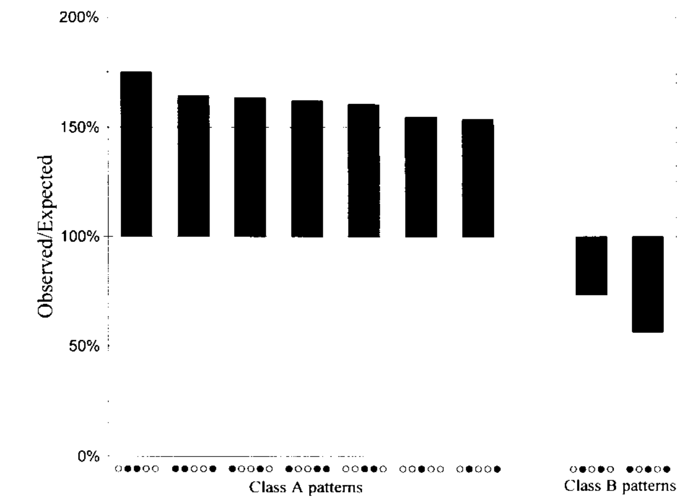

Fig. 2. The relative occurrence of each class A and class B binary pattern in the amino acid sequences of the database. The number of occurrences of each pattern is divided by the expected value. The expected frequency is the number of times a given pentapeptide binary pattern would be expected to occur in the database given that some amino acids occur in natural proteins more frequently than others (see the Materials and methods). For the class A sequence patterns, the expected and observed occurrences are: O●●OO (exp = 102, obs = 179); ●●OO● (exp = 83, obs = 137); ●OO●O (exp = 102, obs = 167); ●OO●● (exp = 83, obs = 135); OO●●O(exp = 102, obs = 164); OO●OO (exp = 125, obs = 195); and O●OO● (exp = 102, obs = 157). The sum over all class A sequence patterns is exp = 699, obs = 1,134. For the class B sequence patterns, the expected and observed occurrences are as follows: O●O●O (exp = 102, obs = 75) and ●O●O● (exp = 83, obs = 47). The sum over all class B sequence patterns is exp = 185, obs = 122.

database see the Materials and methods). Within this database, the observed frequency of each class A and class B binary pattern was compared to its "expected frequency." (The "expected frequency" is the number of times a given pattern would occur if the 40,667 amino acids in the database were arranged in random sequences; it is normalized for the fact that some amino acids occur more frequently than others—see the Materials and methods.) The results shown in Figure 2 demonstrate that the seven class A binary patterns occur considerably more frequently than would be expected at random. Depending on which pattern is considered, the class A patterns occur between 154% and 175% as often as expected. In contrast, the class B patterns occur considerably less often than expected. The two class B patterns, O●O●O and ●O●O● occur only 73% and 56%, respectively, as often as expected.

Why do class A patterns occur more frequently and class B patterns occur less frequently than would be expected at random? We can account for this observation in three ways. First, α-helical segments occur in our database more frequently than β-strand segments. As described in the Materials and methods, we classify any pentapeptide having four or five consecutive α residues to be an α-helical segment, whereas a β-strand segment is defined by having three or more consecutive β residues. Despite the more lenient definition of β segments, there are still 45% more α-helical segments in the database than β-strand segments (12,103 versus 8,371). Second, the α-helices in the database are typically longer (average length = 10 residues) than the β-strands (average length = 5 residues), and so a search for binary patterns is likely to find several pentapeptides wholly contained within a contiguous helix more frequently than within a contiguous β-strand. A third reason for the abundance of the class A patterns relative to the paucity of class B patterns stems from nature's preference for hydrophobic sequences in β-strands and will be discussed below.

Class A patterns occur in α-helical structure and class B patterns occur in β-structure

To probe whether nature has incorporated binary patterning of polar (O) and nonpolar (●) residues in the design of native proteins, we asked: How frequently are the class A and class B sequence patterns actually found in the α-helices and β-strands of natural proteins? In this analysis, a segment of protein structure was classified as α-helical if it contained at least four consecutive residues in α conformation and was classified as β-structure if it contained at least three consecutive residues in β conformation. If a segment of structure satisfied neither of these criteria, then it was labeled "other." Using these definitions, secondary structure was assigned to every occurrence of each of the class A and class B binary patterns. The number of times each pattern occurred in a given secondary structure is summarized in Figure 3. The observed occurrences are compared to the number of times that would be expected at random given that the database as a whole contains 31% α-helical structure, 21% β-structure, and 48% "other" structure.

The number of times each class A and each class B pattern occurs in α-helices versus β-strands is shown schematically in Figure 4. The results demonstrate that all of the class A patterns occur far more frequently in α-helices than in β-strands. For example, the class A sequence pattern O●●OO occurs a total of 179 times in our database. Of these occurrences, 130 (i.e., 73%) are in α-helical secondary structure, 14 (i.e., 8%) are in β-strands, and 35 (i.e., 19%) are in "other" structure. Moreover, this class A sequence pattern occurs in α-helical secondary structure at a significantly higher percentage (73%) than the overall percentage of α-helical secondary structure in the database as a whole (31%). This pattern occurs in β-strands only 8% of the time, which is lower than the overall percentage of β-structure in the database as a whole (21%).

The class B sequence patterns occur in β-strands far more frequently than they occur in α-helical or "other" secondary structure (see Fig. 4). For example, ●O●O● occurs 47 times in our database. Twenty-six of these occurrences (55%) are in β-strands, whereas only 7 (15%) are in α-helices, and 14 (30%) are in "other" structure. Moreover, the occurrence of this class B sequence pattern in β-structure is significantly more probable than the overall probability of β-structure in the database as a whole: 55% of the occurrences of the ●O●O● pattern are in β-strands, whereas only 21% of the database as a whole is in β-structure. Furthermore, this same pattern occurs in α-structure at a lower frequency (15%) than the overall frequency of α-structure in the database as a whole (31%). Finally, it should be noted that even though all of the pentapeptide patterns occasionally occur in "other" secondary structures, in all cases the observed percentage is lower than the percentage of "other" secondary structure in the database as a whole.

These results indicate that nature uses binary sequence patterns preferentially in those secondary structures that accommodate the amphiphilicity of the pattern. Thus, the pentapeptide

Class A	Alpha Helices		Beta Strands		Other	
	Expected	Observed	Expected	Observed	Expected	Observed
○●●○○	55	**130**	38	**14**	87	**35**
●●○○●	42	**82**	29	**14**	66	**41**
●○○●○	51	**84**	35	**25**	81	**58**
●○○●●	41	**97**	29	**14**	65	**24**
○○●●○	50	**112**	35	**13**	79	**39**
○○●○○	59	**117**	41	**11**	94	**67**
○●○○●	48	**88**	33	**23**	76	**46**
SUM=	*346*	***710***	*240*	***114***	*548*	***310***

Class B	Expected	Observed	Expected	Observed	Expected	Observed
○●○●○	23	**10**	16	**33**	36	**32**
●○●○●	14	**7**	10	**26**	23	**14**
SUM=	*37*	***17***	*26*	***59***	*59*	***46***

Fig. 3. Frequency of each class A and class B sequence pattern within α-helical, β-strand, or "other" secondary structure. For each binary sequence pattern, the observed number of occurrences is compared to the number of times expected at random given the database as a whole contains 31% α-helical structure, 21% β-structure, and 48% "other" structure. For example, there are 137 occurrences of the pattern ●●○○●. Thus, the expected values for this pattern are 31% × 137 = 42 α-helical segments, 21% × 137 = 29 β-strand segments, and 48% × 137 = 66 "other" segments.

patterns ○●○○●, ●○○●●, ○○●●○, ○●●○○, ●●○○●, ●○○●○, and ○○●○○ are found preferentially in α-helices, whereas the alternating patterns ●○●○● and ○●○●○ are found preferentially in β-strands. Examples of each of the nine sequence patterns in the corresponding secondary structures of natural proteins are shown in Figure 5.

The most common binary sequence patterns in the α-helices and β-strands of natural proteins

As demonstrated above, we have established that when a particular class A or class B binary sequence pattern occurs, it is found preferentially in the expected secondary structure. However, the converse question must also be addressed: When a particular secondary structure occurs, how often does its sequence match one of the patterns expected from the binary code strategy? For example, pentapeptide segments of α-helical secondary structure occur in our database 12,103 times. Among these α-helical segments, how often are the class A patterns actually used? Likewise, pentapeptide segments of β-structure occur 8,371 times. How many of these have class B patterns?

If we consider all possible polar/nonpolar patterns, not merely the class A and class B patterns described above, then there are 32 binary pentapeptide patterns ($2^5 = 32$). The frequencies of each of these patterns in α-helical and β-strand secondary structures are shown in Figure 6A and B, respectively. For α-helical structure, the seven most common pentapeptide sequence patterns are indeed class A patterns. This finding is consistent with the structural environment of α-helices. In natural proteins, α-helices typically have a buried face and an exposed face. For example, among the nine protein domain superfolds defined by Thornton and coworkers, all of the α-helices pack along the sur-

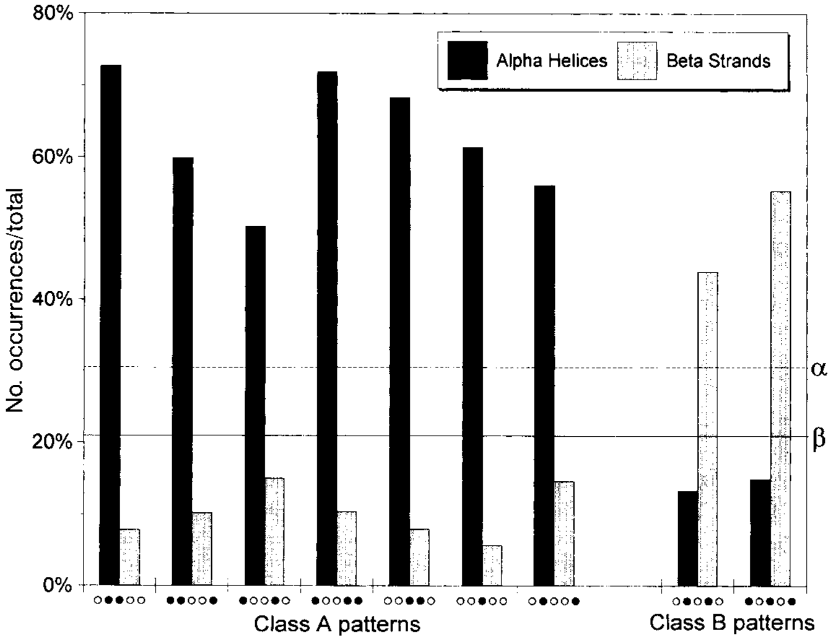

Fig. 4. The structural environment for each of the class A and class B patterns. The number of occurrences of each binary pattern within a given secondary structure was divided by the total number of occurrences of that pattern in the database. Horizontal lines represent the percentages of secondary structure in the database as a whole (i.e., irrespective of sequence). The database as a whole contains 31% α-helical structure, 21% β-structure, and 48% other structure.

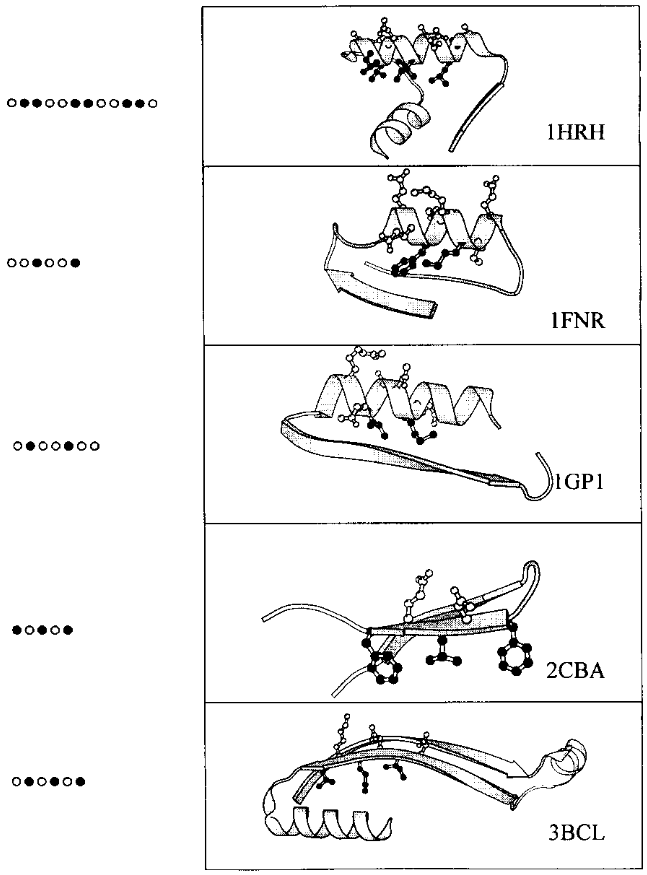

Fig. 5. Examples from protein structures in the database depicting particular occurrences of class A patterns in α-helical secondary structure and class B patterns in β-strand secondary structure. Patterns are shown for residues 516–528 of ribonuclease H (1HRH); residues 214–221 of ferredoxin reductase (1FNR); residues 54–60 of glutathione peroxidase (1GP1); residues 66–70 of carbonic anhydrase (2CBA); and residues 141–146 of bacteriochlorophyll-*a* (3BCL). Figures were generated using MOLSCRIPT (Kraulis, 1991)

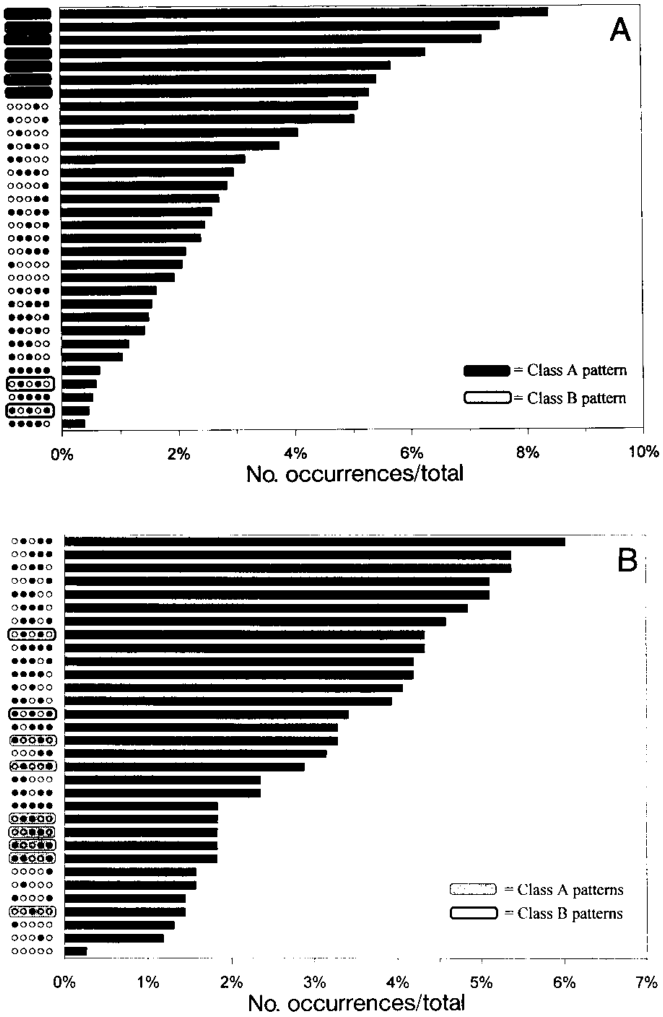

Fig. 6. **A:** Frequencies of each of the 32 binary patterns in α-helical structure. **B:** Frequencies of each of the 32 binary patterns in β-strand structure. The binary code of polar (○) and nonpolar (●) amino acids allows for 32 possible pentapeptide patterns ($2^5 = 32$). Class A patterns are shaded and class B patterns are circled. For each pattern, the number of occurrences within a given secondary structure (α-helical or β-strand) is divided by the total number of occurrences of all 32 patterns within that secondary structure. For example, the class A pattern ○●●○○ occurs 130 times in α-helical structure. The total number of times all 32 patterns occur in α-helical structure is 1,547. Therefore we plot 130/1,547 = 8.4%.

face and thereby expose one face to solvent while burying the other (see Fig. 3 in Orengo et al. [1994]). Thus, it is reasonable that the seven pentapeptide patterns classified originally as class A because of their compatibility with amphiphilic α-helical structure (see Fig. 1) are indeed favored in the α-helices of natural proteins. This finding is reassuring because the seven class A patterns are the same as those used in our experimental work on the de novo design of four-helix bundles (Kamtekar et al., 1993).

However, the situation for β-structure is more complex. Surprisingly, in the β-strands of natural proteins, 12 of the 14 most commonly used binary patterns are *not* class B patterns (see Fig. 6B). Instead, they are often hydrophobic pentapeptide sequences typically containing either three or four nonpolar (●) amino acids. Indeed, among the 11 most common patterns in Figure 6B, nine of them contain either three or four nonpolar (●) residues in their pentapeptide sequences. Thus, the two class B patterns, ●○●○● and ○●○●○, which were chosen originally because of their complementarity with amphiphilic β-structure, are in fact *not* the predominant pentapeptides in the β-strands of natural proteins.

Why does nature prefer hydrophobic patterns rather than the alternating polar/nonpolar sequences expected for the design of amphiphilic β-strands? Presumably, the reason stems from the structural environment of many β-strands. β-Strands, more so than α-helices, are often completely buried in the 3D structures of natural proteins. For example, the two most common protein domain superfolds, the α/β doubly wound (e.g., flavodoxin) and the TIM barrel (e.g., triose phosphate isomerase), are structures in which the β-strands have both faces buried (Orengo et al., 1994). Given the preponderance of such structures, it is not surprising that nature frequently constructs β-strands from sequences rich in nonpolar residues. This is consistent with the overall composition of β-strands in our database, which is significantly enriched in nonpolar (●) residues (39.0%) relative to the database as a whole (26.9%).

Clearly, the class B patterns ●○●○● and ○●○●○ are not among the most commonly used binary patterns in β-structure (Fig. 6B). However, when they do occur in natural proteins, class B patterns are typically found in β-strands (Fig. 4). Is the structural environment of those β-strands that contain class B patterns somehow distinct? We examined the 3D structures of all occurrences of ●○●○● and ○●○●○ in β-structure and found that 58 of 59 occur in β-strands on the surface of a protein (for examples, see 2CBA and 3BCL in Fig. 5). The one exception occurs in β-amylase (1BTC) in a pocket lined with bound water. Thus, although ●○●○● and ○●○●○ may not be suitable for β-strands in all environments, nature has found these patterns ideal for the construction of solvent-exposed β-strands.

Implications for de novo protein design

A central goal of the current work is to determine the optimal sequence patterns for the design of novel proteins. Our previous experimental work is based on the premise that a simplified code of amino acids arranged with the appropriate binary patterning of polar and nonpolar residues can serve as a cornerstone for de novo protein design (Kamtekar et al., 1993; M.W. West & M.H. Hecht, unpubl.). In light of the current analysis of binary patterns in natural proteins, we ask: are the class A and class B binary sequence patterns well-suited for the design of novel proteins?

For protein design it is important that a sequence not only favor the desired structure but also disfavor alternative structures. "Positive design" is not sufficient; "negative design" must also be considered (Hecht et al., 1990). Therefore, we have compared the ratio of α-helical occurrences relative to β-strand occurrences for all of the 32 possible binary pentapeptides (see Fig. 7). From this perspective, it is clear that both class A and class B sequence patterns are well suited for de novo design: Class A patterns strongly favor α-structure and disfavor β-structure, whereas class B patterns strongly favor β-structure and disfavor α-structure (Fig. 7). Among those patterns having the largest ratio of α-helix/β-strand, 7 of the top 12 are class A patterns. The other five are polar sequences suitable for highly exposed α-helices. Even more striking, the two class B patterns, ●○●○● and ○●○●○, clearly have very strong biases for β-structure over α-structure (see Fig. 7, top). They occur in the desired β-structure at a frequency 7.5-fold higher than in the undesired α-structure. The only binary pentapeptides having a stronger bias for β-structure are ○●●●● and ●●●●○. These patterns presumably would be suited only for the design of β-strands predominantly buried in a hydrophobic core.

These findings suggest that the periodicity of polar (○) and nonpolar (●) amino acids can predispose a sequence to form either an α-helix or a β-strand. Indeed, studies of synthetic model peptides have shown when the periodicity of polar and nonpolar residues in an amino acid sequence matches the repeat pattern of a secondary structure, then the sequence will form that secondary structure—regardless of the intrinsic propensities of the component amino acids (Xiong et al., 1995). Specifically, for two peptides having similar amino acid compositions, but different periodicities, the one with its polar and nonpolar residues arranged in a class A pattern self-assembles into oligomers of amphiphilic α-helices, whereas the other peptide, which has a class B periodicity, self-assembles into oligomers of amphiphilic β-strands (Xiong et al., 1995). Thus, both for synthetic peptides and natural proteins, the periodicity of polar and nonpolar amino acids can play a significant role in dictating whether a sequence will form an α-helix or a β-strand.

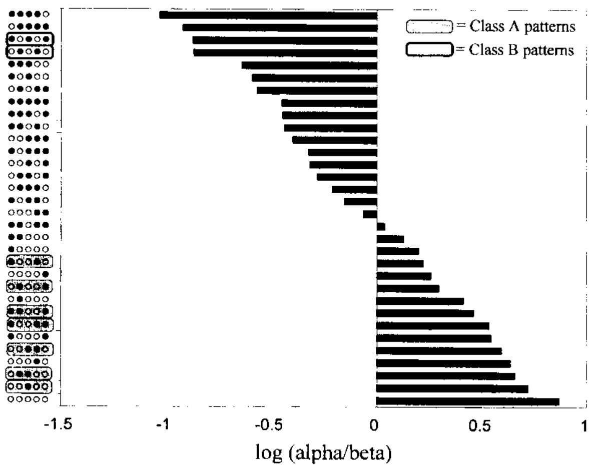

Fig. 7. For all 32 possible pentapeptide patterns, the preference for α-structure relative to β-structure is shown. Positive values indicate a preference for α-helical structure and negative values indicate a preference for β-structure. Class A patterns are shaded and class B patterns are circled. Plots are based on normalized frequencies. Thus, the number of times a given pattern occurs in α-structure is divided by the total number of occurrences of α-structure for all 32 patterns. The normalized β-strand frequency was calculated in the same way. The plot shows the logarithm of the ratio of the normalized α-helical frequency divided by the normalized β-strand frequency. For example, the pattern ○○●○○ occurs 117 times in α-structure, and the total number of α-helical segments for all 32 binary patterns is 1,547. Likewise, ○○●○○ occurs 11 times in β-structure, and the total number of β-strand segments for all 32 binary patterns is 764. Thus, for this pattern we plot log [(117/1,547)/(11/764)] = 0.72.

Despite the predispositions of some binary patterns to form α-helices and others to form β-strands, the suitability of a particular binary pattern for de novo design will depend upon the intended tertiary environment of an α-helix or β-strand in the designed structure. For example, although the class A patterns favor α-structure over β-structure by substantial margins, this is also true for several strongly polar patterns such as ○○○○○ and ○○○●○ (see Fig. 7). Clearly, when devising a sequence for de novo α-helices, one must consider whether the desired α-helix will occur in an amphiphilic environment or be completely exposed to solvent.

Similarly, the choice of a binary pattern for design of β-structure will depend on the intended tertiary environment. Figure 7 demonstrates that the four binary patterns with the strongest bias for β-structure over α-structure are the two hydrophobic patterns ●●●●○ and ○●●●●, followed by the two class B patterns ●○●○● and ○●○●○. Clearly, the class B patterns would be appropriate for designing β-strands in some contexts and not others. Indeed these patterns would probably fail in environments where the β-strands are completely buried and must have nonpolar residues on both sides, such as in the TIM barrel or the flavodoxin structure. Likewise, they would be inappropriate at the interface of a dimeric protein, where one hydrophobic face points toward the interior of the monomer, and another hydrophobic face mediates dimerization. In these environments the hydrophobic patterns would be more suitable than the class B patterns. However, the two class B patterns, ●○●○● and ○●○●○, would be well-suited for the de novo design of small (≤80 residues) monomeric β-barrel proteins in which each β-strand has an exposed face and a buried face (M.W. West & M.H. Hecht, unpubl. results).

Comparison of the binary code of polar and nonpolar residues with the full range of sequences in the database

By restricting our analysis to seven polar (○) and five nonpolar (●) residues, we have intentionally excluded residues in the middle of the hydrophobicity scale. How does our analysis based upon this simplified code of amino acids compare to what might be observed with the full range of amino acid sequences in the database?

The eight residues (Ser, Pro, Thr, Ala, Tyr, Trp, Cys, Gly) not included in our binary code can be represented as "other" (□). We can now consider all possible pentapeptides in the database of natural proteins as combinations of polar (○), nonpolar (●), or "other" (□) residues. In this representation there are $3^5 = 243$ possible pentapeptide patterns. Relative to this full range of all possible patterns, how strong are the structural predispositions of our class A and class B patterns?

We calculated the ratio of α-helical occurrences relative to β-strand occurrences for all 243 possible pentapeptides and ranked them according to their preference (see Fig. 8). All seven class A patterns are among those pentapeptides with a strong bias for α-structure over β-structure (Fig. 8). Indeed, the class A patterns represent 5 of the 13 patterns with the highest α/β ratio. Even more striking, when the 243 possible pentapeptides are ranked according to their preference for β-structure (right side of Fig. 8), the class B patterns, ●○●○● and ○●○●○, were both among the six patterns with the smallest α/β ratio. Thus, even when compared to all possible sequence combinations of polar (○), nonpolar (●), or "other" (□) residues, class A and class B binary patterns seem well suited for the design of α-helices and β-strands in novel proteins.

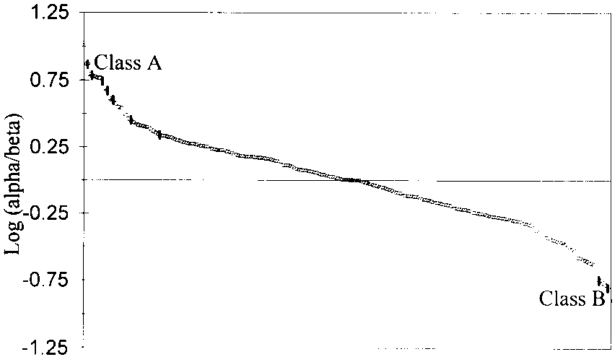

Fig. 8. The preference for α-structure relative to β-structure is shown for all 243 pentapeptide combinations of polar (○), nonpolar (●), or "other" (□) residues. Positive values indicate a preference for α-helical structure and negative values indicate a preference for β-structure. Each point represents a particular pentapeptide pattern. The seven class A and two class B patterns are indicated by tick marks. The plots are based on normalized frequencies. Thus, the number of times a given pattern occurs in α-structure is divided by the total number of occurrences of α-structure for all 243 patterns. Normalized β-strand frequency was calculated in the same way. The plot shows the logarithm of the ratio of the normalized α-helical frequency divided by the normalized β-strand frequency.

Materials and methods

A database of unique protein structures

We obtained a list of 197 proteins representing the unique structures known at a resolution of 2.5 Å or better within the Protein Data Bank (Bernstein et al., 1977) as of the summer of 1994. Files containing only α-carbon coordinates were deleted from this original group, leaving a data set of 177 protein structures. These proteins are listed by their PDB accession code as follows: 135l, 1aaj, 1aak, 1aaz, 1abe, 1abk, 1acp, 1ads, 1ak3, 1ala, 1aoz, 1aps, 1arb, 1atp, 1ayh, 1bgc, 1bgh, 1bia, 1btc, 1bw4, 1cau, 1cc5, 1cde, 1cmc, 1col, 1cpc, 1cse, 1dhr, 1dri, 1eaf, 1eco, 1egr, 1end, 1ezm, 1fba, 1fia, 1fkf, 1fnr, 1fus, 1fxd, 1gal, 1gau, 1gd1, 1gmf, 1gmp, 1gp1, 1hfi, 1hip, 1hmy, 1hom, 1hrh, 1hsb, 1hyp, 1ipd, 1ith, 1lct, 1lga, 1lis, 1lmb, 1lpe, 1mbd, 1min, 1mns, 1nrd, 1nsc, 1nxb, 1opb, 1ova, 1pda, 1pgd, 1phg, 1phh, 1pk4, 1poa, 1poc, 1prc, 1rbp, 1rcb, 1rhd, 1rip, 1rnb, 1rop, 1rtc, 1sbp, 1sha, 1sim, 1ten, 1tfg, 1tfi, 1thg, 1tie, 1tpl, 1tta, 1ubq, 1utg, 1wsy, 1ycc, 1zaa, 256b, 2abd, 2aza, 2cba, 2ccy, 2cdv, 2cmd, 2cpl, 2csc, 2cyp, 2gst, 2had, 2hhm, 2hmz, 2hpd, 2lh7, 2liv, 2mad, 2msb, 2ohx, 2pia, 2por, 2reb, 2rn2, 2rsp, 2scp, 2sga, 2sic, 2sn3, 2sns, 2sod, 2stv, 2trx, 2ts1, 2tsc, 2yhx, 3adk, 3b5c, 3bcl, 3chy, 3cla, 3eca, 3gap, 3gf1, 3grs, 3il8, 3pgk, 3rub, 3sc2, 3sdh, 3sdp, 3tgl, 4enl, 4fgf, 4fxn, 4gcr, 4i1b, 4lzm, 4mdh, 4pfk, 4ptp, 4sgb, 4xis, 5cpa, 5hvp, 5pti, 5tim, 6taa, 7aat, 7rsa, 8acn, 8atc, 8cat, 8dfr, 8rxn, 9icd, 9ldt, 9pap, 9wga. Our database consists of the unique subunits of these 177 proteins. The total number of independent subunits in the database is 183, and the total number of residues is 40,667.

In the database of 40,667 residues, the polar amino acids (○) occur with the following frequencies (and percentages): Lys, 2,533 (=6.23%); Glu, 2,429 (=5.97%); Asp, 2,405 (=5.91%); Arg, 1,865 (=4.59%); Asn, 1,770 (=4.35%); Gln, 1,470 (=3.61%); and His, 928 (=2.28%). The nonpolar residues (●) occur with the following frequencies (and percentages): Leu, 3,456 (=8.50%); Val, 2,759 (=6.78%); Ile, 2,201 (=5.41%); Phe, 1,640 (=4.03%); and Met, 870 (=2.14%). Overall, polar (○) residues occur 13,000 times, which represents 32.95% of the database; and nonpolar (●) residues occur 10,926 times, which represents 26.87%. Thus, for example, the expected frequency for the sequence pattern ○●●○○ would be (32.95%) × (26.87%) × (26.87%) × (32.95%) × (32.95%) × 39,651 = 102.

Searches for binary patterns in the sequences and structures of the database

The pattern scanning and processing language *awk* was used to search the amino acid sequences of the proteins in our database for pentapeptides having any of the nine binary patterns described in the text. We chose to search for pentapeptide patterns because: (1) searches for longer patterns would yield a smaller and less representative data set; (2) the average length of a β-strand in natural proteins is five residues; and (3) pentapeptides have been used by several authors in searches for patterns among protein sequences (Kabsch & Sander, 1984; Argos, 1987; Vazquez et al., 1993).

For every occurrence of a binary pattern, the secondary structure of that occurrence was recorded. A segment of protein structure was defined as α-helical if it contained at least four consecutive residues in α conformation and was defined as β-structure if it contained at least three consecutive residues in β conformation. If a segment of structure satisfied neither of these criteria, then it was labeled as "other." The conformations of the residues were assigned by one of two methods. In cases where secondary structure assignments are listed in the PDB file, those assignments were used. In cases where assignments were not available from the PDB file, we used the DSSP algorithm (Kabsch & Sander, 1983). The secondary structure information contained in the header of the PDB files was extracted using an *awk* script and was combined with part of the DSSP output file to create a suitable input file for our search. This input consisted of the protein's PDB designation, residue number, residue name, DSSP secondary structure assignments, and PDB file secondary structure assignments. An *awk* script was written to search our input file for the desired sequence patterns. It produced an output file containing the following information: the binary sequence pattern, the identities and locations of the residues within that pattern, the DSSP and PDB secondary structure assignments for each residue within the pattern, the secondary structure of the pentapeptide based on the DSSP data as well as that based on the PDB assignments. This file was used to analyze the database.

Using other output files created by *awk*, we obtained: (1) the total number of occurrences of each binary sequence pattern; (2) the total number of times that each sequence pattern occurs in a particular secondary structure (i.e., α-helix, β-strand, or other); (3) amino acid analysis of the database (i.e., the total number of each amino acid in the database); (4) the average length of each type of secondary structure; (5) the percentage of α-helical, β-strand, and "other" pentapeptides in the database; and (6) the number of times that a particular pattern occurs within a given type of secondary structure.

Acknowledgments

We are indebted to Robert E. Humphreys for sparking our interest in the patterns of polar and nonpolar amino acids in natural proteins; Tod Klingler for providing a list of unique protein structures within the Protein Data Bank; Peter Kahne and Diana Mallows for help with the early phases of this project; Curt Hillegas for assistance with computations; and Joel A. Ybe, Adam P. Brunet, Satwik Kamtekar, Ellen O. West, Judy Swan, and Mike D. Rozycki for helpful comments on the manuscript. This work was supported by a Beckman Foundation Young Investigator Award to M.H.H. M.W.W. is supported by an NIH predoctoral training grant in Molecular Biophysics (2-T32-GM08309).

References

Argos P. 1987. Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. *J Mol Biol 197*:331–348.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bowie JU, Clarke ND, Pabo CO, Sauer RT. 1990. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins Struct Funct Genet 7*:257–264.

Bowie JU, Sauer RT. 1989. Identification determinants of folding and activity for a protein of unknown structure. *Proc Natl Acad Sci USA 86*: 2152–2156.

Creighton TE. 1993. *Proteins: Structures and molecular properties, 2nd ed.* New York: Freeman Press.

DeGrado WF, Lear JD. 1985. Induction of peptide conformation at apolar/ water interfaces: 1. A study with model peptides of defined hydrophobic periodicity. *J Am Chem Soc 107*:7684–7689.

Dill KA. 1990. Dominant forces in protein folding. *Biochemistry 29*:7133–7155.

Eisenberg D, Weiss RM, Terwilliger TC. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci USA 81*:140–144.

Fauchere J, Pliska V. 1983. Hydrophobic parameters pi of amino-acid side chains from the partitioning of *N*-acetyl-amino-acid amides. *Eur J Med Chem 18*:369–375.

Hecht MH, Richardson JS, Richardson DC, Ogden RC. 1990. De novo design expression and characterization of felix: A four-helix bundle protein of native-like sequence. *Science 249*:884–891.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 18*:2577–2637.

Kabsch W, Sander C. 1984. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations *Proc Natl Acad Sci USA 81*:1075–1078.

Kaiser ET, Kezdy FJ. 1983. Secondary structures of proteins and peptides in amphiphilic environments. *Proc Natl Acad Sci USA 80*:1137–1143.

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

Kauzmann W. 1959. Some factors in the interpretation of protein denaturation. *Adv Protein Chem 14*:1–63.

Kraulis P. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946–950.

Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds *Nature 372*:631–634.

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.

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

Xiong H, Buckwalter BL, Shieh HM, Hecht MH. 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.