doi:10.1016/S0022-2836(02)00973-7 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 323, 453–461

JMB

On the Significance of Alternating Patterns of Polar and Non-polar Residues in Beta-strands

Yael Mandel-Gutfreund and Lydia M. Gregoret*

Department of Chemistry and Biochemistry, University of California, Santa Cruz, CA 95064, USA

A common assumption about protein sequences in β-strands is that they have alternating patterns of polar and non-polar residues. It is thought that such patterns reflect the interior/exterior geometry of amino acid residue side-chains on a β-sheet. Here we study the prevalence of simple hydrophobicity patterns in parallel and antiparallel β-sheets in proteins of known structure and in the sequences of amyloidogenic proteins. The occurrence of 32 possible pentapeptide binary patterns (polar (P)/non-polar (N)) is computed in 1911 non-homologous protein structures. Despite their tendency to aggregate in experimentally designed proteins, the purely alternating hydrophobic/polar patterns (PNPNP and NPNPN) are most frequent in β-sheets, typically occurring in antiparallel strands. The overall distribution of the pentapeptide binary patterns is significantly different in strands within parallel and antiparallel sheets. In both types of sheets, complementary patterns (where the hydrophobic and polar residues pair with one another) associate preferentially. We do not find alternating patterns to be common in amyloidogenic proteins or in short fragments involved directly in amyloid formation. However, we do note some similarities between patterns present in amyloidogenic sequences and those in parallel strands.

*Corresponding author

Keywords: β-sheet; parallel; antiparallel; binary patterns; aggregation

Introduction

The sequestration of non-polar amino acid residues in the cores of globular proteins is generally agreed to be a dominant force in protein folding and stability.[1,2] Soon after the first structures of the myoglobin and hemoglobin proteins were solved, it was noted that hydrophobic residues tended to segregate in the interiors of the proteins, while the polar and charged residues were exposed to the aqueous solvent.[3,4] It was also noted that most of the α-helices in these structures were of an amphiphilic nature in which one surface of the helix projects mainly hydrophilic side-chains while the other side contains hydrophobic residues.[5]

On the basis of the observed packing in polypeptide chains, Lim suggested general principles for identifying α-helices and β-sheet structures from sequence.[6] Later, Eisenberg and co-workers introduced the "hydrophobic moment" to measure the amphiphilicity of protein segments and showed that residue hydrophobicity patterns match the periodicity of secondary structures.[7] Sequences forming α-helices tend to have a strong periodicity of three or four residues while β-strands show periodicity of ~two residues, typical of the alternating period of β structures. Although individual amino acid residues possess an intrinsic propensity to form either α-helical or extended structures,[8–13] these preferences are modulated by the sequence segments within which they reside.[14–16]

The recognition that protein sequences contain hydrophobicity patterns led to the development of theoretical models of proteins in which the 20 amino acid residues are reduced to a binary hydrophobic/polar code. Binary lattice models, in particular, have been instrumental in developing theories of protein stability and folding.[2,17] The simplicity of these theoretical models, in turn, led experimentalists to attempt to build real proteins using restricted amino acid alphabets so as to better understand the minimal information necessary to specify a protein fold.[18–22] Using simple binary patterns, Kamtekar and co-workers[18] succeeded in designing monomeric α-helical proteins. However, proteins designed to adopt β-sheet folds

Abbreviations used: PDB, Protein Data Bank.

E-mail address of the corresponding author: gregoret@chemistry.ucsc.edu

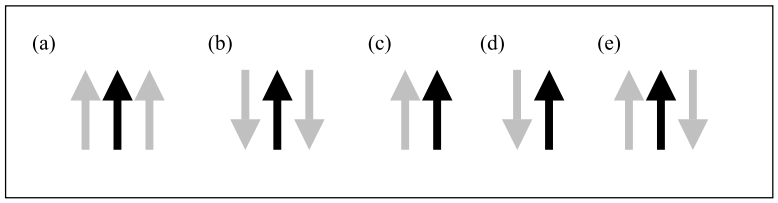


Figure 1. A schematic diagram of the five different types of β-strands analyzed: (a) interior parallel strand, (b) interior antiparallel strand, (c) parallel edge strand, (d) antiparallel edge strand, (e) mixed strand with one parallel and one antiparallel partner. Arrows define the direction of each strand.

did not form monomers. Instead, they assembled into large oligomers.[19,20] Other investigators have also observed a tendency of alternating polar/non-polar patterns to aggregate *in vitro*.[23–25] Broome & Hecht[26] recently suggested that sequences with alternating binary patterns are prone to form fibril structures resembling β-amyloid and thus are disfavored in natural proteins.

Here we report on the occurrence of alternating polar/non-polar binary patterns in a large data set of crystallographically determined protein structures. β-Strand windows of five residues were divided into five different groups according to their interactions with the neighboring strands (parallel, antiparallel, edge-parallel, edge-antiparallel and mixed; illustrated in Figure 1). The distribution among the possible binary patterns was analyzed in each group. We also studied the preferences of these different patterns to pair with one another in sheet structures. Finally, we analyzed the sequences of amyloid-forming proteins and found differences between the patterns within these sequences and those in the large set of proteins.

Results

Binary patterns in parallel and antiparallel β-strands

The occurrences of the 32 possible binary patterns (P = polar, N = non-polar) of length five was tabulated in all β-strands in a large data set of 1911 protein structures. As illustrated in Figure 2, the most frequent patterns in all β-strands (when including both parallel and antiparallel strands together) were the two pure alternating patterns (PNPNP and NPNPN). These two patterns were also found most frequently when analyzing antiparallel β-strands separately (including strands that have two antiparallel strand partners and edge strands; see Figure 3(a)). This held true when we analyzed the data using different definitions for N and P.

As can be observed in Figure 3(b), in the parallel strands, the alternating patterns were much less frequent. When we considered all possible binary patterns in parallel β-strands, the alternating patterns were ranked 14th and 15th. The most frequent pattern among the parallel strands was the purely non-polar pattern (NNNNN). This pattern was ranked fourth among all β-strands as well as among antiparallel strands considered separately. Other non-polar rich patterns ranked highly as well. The histograms in Figure 3 strongly show that antiparallel strands tend to contain patterns with an excess of polar residues while the hydrophobic patterns are preferred in parallel strands. These results were again insensitive to the exact N and P definitions used.

The abundance of non-polar patterns in parallel β-strands correlates well with the overall tendency of parallel β-sheets to be buried.[27] The distribution of binary patterns in antiparallel strands is similar to the distribution for all β-strands combined (antiparallel, parallel, mixed and edge strands). This is expected on the basis of their prevalence in the

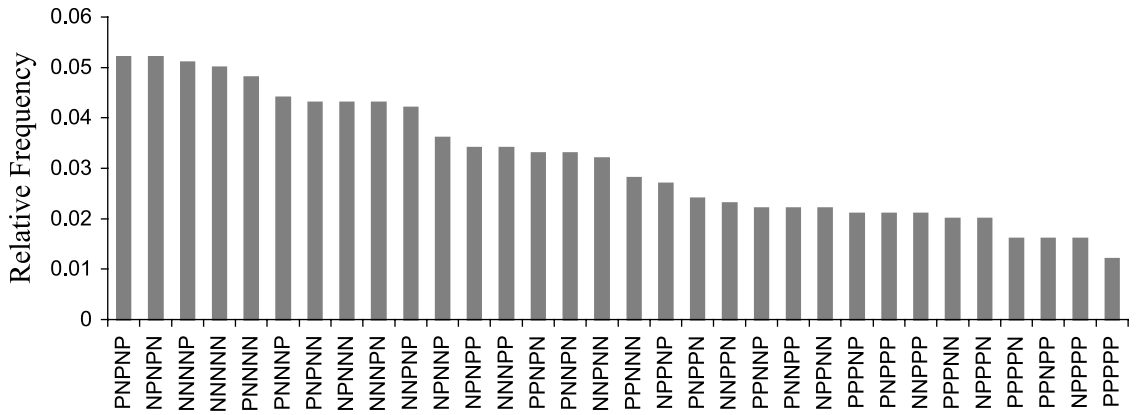
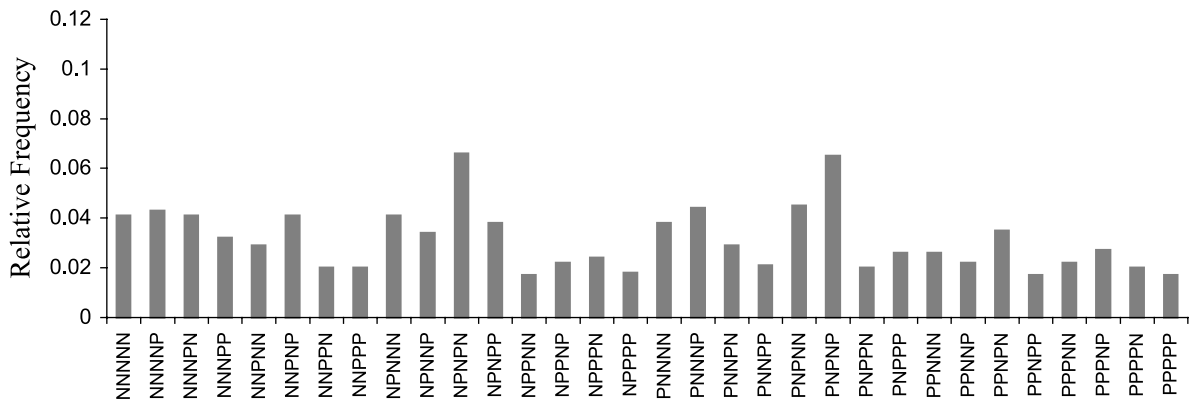


Figure 2. The relative frequency of 32 binary patterns (P, polar; N, non-polar) in β-strands (including parallel and antiparallel strands).

(a) **Antiparallel strands**



(b) **Parallel strands**

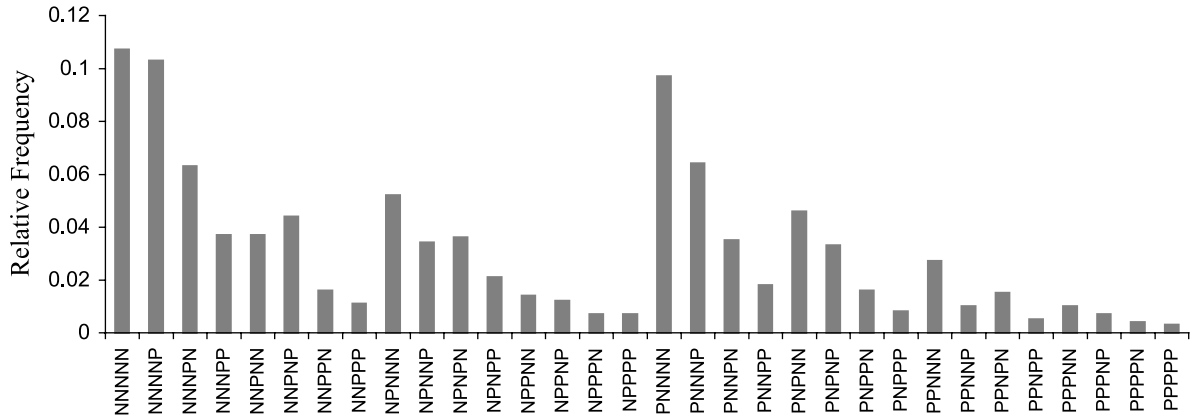


Figure 3. The relative frequency of the 32 binary polar/non-polar patterns in (a) antiparallel and (b) parallel β-strands.

data: there are 19,317 five-residue windows involved in antiparallel interactions and only 4693 windows involved in parallel interactions. These values are also consistent with the frequencies of single residues found in the Protein Data Bank (~78% antiparallel and 22% parallel).

We next calculated the expected probability of each of the 32 binary patterns in parallel and antiparallel strands on the basis of the overall frequency of polar and non-polar residues in each type of strands. For binary patterns of length five, the expected probability of each pattern can be calculated by the formula: $(P_p)^i(P_n)^{5-i}$, where P_p and P_n are the probability of polar and non-polar residues in the data and i is the number of polar residues in the specific pattern. For example, the probability for every one of the ten possible patterns with three polar and two non-polar residues in antiparallel strands is $(P_p)^3 \times (P_n)^2 = 0.40^3 \times 0.60^2 = 0.02304$. When we multiply the probability by the total number of antiparallel windows analyzed, we get the expected number of windows for each pattern with a similar composition (502, in our case). The alternating pattern PNPNP was actually observed 1150 times. The χ-square goodness-of-fit statistical test was further applied to see whether the frequencies with which all patterns were found were significantly different from those expected on the basis of the total frequency of polar and non-polar residues.

In addition to their high relative frequencies in antiparallel β-strands, the alternating patterns (PNPNP and NPNPN) were also significantly more probable than expected on the basis of chance (probability, $P < 10^{-4}$). In parallel β-strands, the pure hydrophobic pattern was most frequent but was not significantly more common than expected on the basis of the overall frequency of the non-polar residues. However, the pattern PNPNP, which was ranked much lower in parallel than in antiparallel strands, was nonetheless found significantly more frequently than expected ($P < 10^{-2}$) in the parallel strands. For comparison, we also calculated the occurrences of the 32

Table 1. The most frequent pentapeptide binary pairs in antiparallel and parallel β-strands scored by the log-odds ratio between the observed number of pairs and the expected number of pairs on the basis of random association

Antiparallel strands				Parallel strands			
Pair	Frequency	Score	# of mismatches	Pair	Frequency	Score	# of mismatches
→ NPNPN ← NPNPN	125	1.7	0	→ NNNNN → NNNNN	46	1.1	0
→ PNPNP ← PNPNP	119	1.6	0	→ NNNNN → PNNNN	33	0.4	1
→ PNNNP ← PNPNP	64	1.4	1	→ PNNNN → PNNNN	32	1	0
→ PNPNP ← PNNNP	62	1.1	1	→ NNNNN → NNNNP	32	0.5	1
→ NNPNP ← PNPNP	60	1.2	1	→ PNNNN → PNPNN	28	1.4	1
→ PNPNP ← PNPNN	59	1.3	1	→ NNNNP → NNNNN	27	0.7	1
→ NNNPN ← NPNPN	55	1.2	1	→ NNNNP → PNNNN	26	0.5	2
→ PNPNN ← PNPNP	53	1.1	1	→ NNNNP → NNNNP	23	0.5	0
→ NPNPN ← NNNPN	52	1	1	→ NNNNN → NNNPN	23	0.5	1
→ PNPNP ← NNPNP	51	0.9	1	→ NNNNN → NPNNN	22	0.5	1

Patterns are aligned as in the β-sheets: inter-strand pairs are positioned one on top of each other, arrows define the direction of each strand in the native protein (N to C (→) and C to N (←)). The overall number of pairs (frequency) and log-odds scores are shown. The number of mismatches in which a polar and a non-polar residue are paired is specified in the last column of each group.

possible pentapeptides in all α-helical regions in our data. Overall we analyzed 85,063 overlapping five-residue α-helical windows (compared to 24,010 β-sheet windows). The most frequent binary patterns in α-helices were PPNNP, NPPNN, and PNNPP. This is consistent with the expected periodicity of ~3–4 for α-helical structures.

Edge and mixed strands

The edge strands of β-sheets are expected to have unique features to prevent them from aggregating with other strands.[27] Upon examining edge strands separately (i.e. strands that are hydrogen bonded to only one other strand), we found the distribution of the binary patterns in antiparallel edge strands to be similar to the distribution of patterns in central antiparallel strands: the most common binary patterns in antiparallel edge strands are still the two pure alternating patterns PNPNP and NPNPN. However, we also found that patterns with three consecutive polar residues, corresponding to patterns in which a polar (or charged) residue is placed where a hydrophobic residue is expected, were much more frequent in the antiparallel edge strands. This agrees with the recent findings of Richardson & Richardson.[27] Also in accordance with these findings, we did not observe the edge strands to be more hydrophobic than other antiparallel strands as was suggested much earlier.[28]

The distribution of binary patterns in parallel edge strands did not resemble that of the interior (non-edge) parallel strands: the purely alternating patterns NPNPN and PNPNP were ranked much higher in edge parallel strands (third and sixth rank *versus* 11th and 17th in non-edge parallel strands). The pattern NNNNN (common in parallel strands in general) was ranked ninth in parallel edge strands.

In mixed strands that interact with one parallel and one antiparallel strand, the binary pattern distribution resembled the overall distribution in β-strands. In these strands, the alternating patterns were ranked second and third while the purely hydrophobic pattern was ranked fourth.

Pairing preferences of β-sheet strands

To further investigate if the abundance of alternating binary patterns that we observed in β-stands is an intrinsic property of the secondary

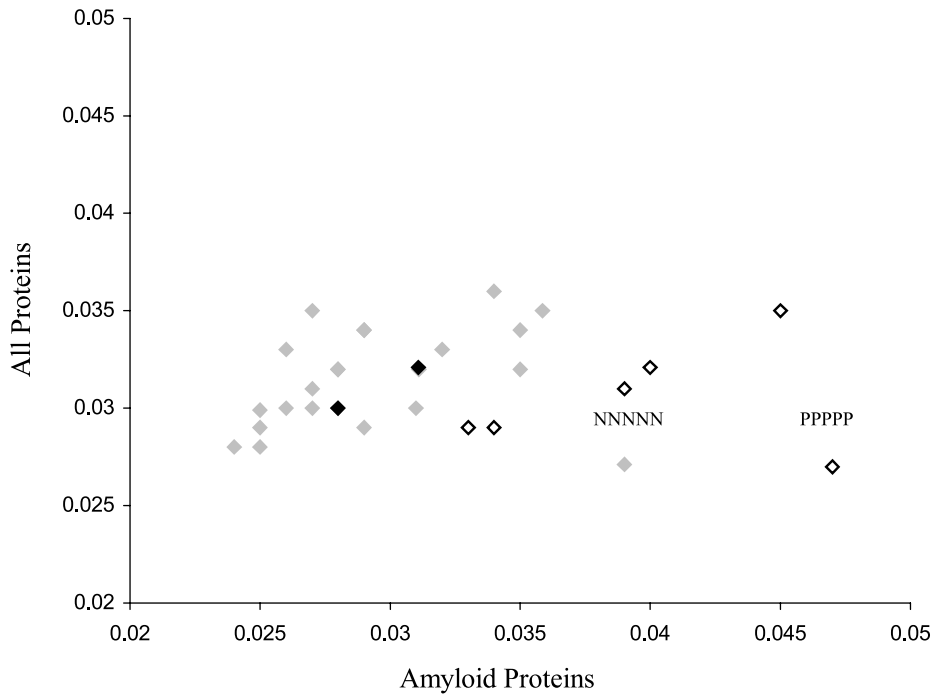


Figure 4. A scatter diagram showing the correlation between the binary pattern frequencies in amyloid proteins compared to their frequencies in all proteins. The two alternating binary patterns (PNPNP, NPNPN) are highlighted in black. The rich polar patterns with four or five polar residues are shown as open diamonds.

structure, we examined whether there is any preference for association between specific binary patterns in β-sheets. If such preferences could be observed, it would further support the theory that the binary patterns are involved in guiding natural β-sheet formation. To see if there is complementarity between adjacent strands, we studied the mutual information between associated or paired patterns (including hairpins and long range sheet association). The mutual information was calculated over all observed pairs:

$$\sum_{ij} P(x_i y_j) \log_2 \frac{P(x_i y_j)}{P(x_i)P(y_j)}$$

i, j runs from one to 32 for all possible binary pentapeptide patterns, $P(x_i y_j)$ is the probability of each specific pair and $P(x_i)P(y_j)$ is the expected probability of the pair on the basis of random association between the patterns. To estimate if the preferences observed for β-strand pairs are statistically significant, we created artificial sets of pairs for parallel and antiparallel strands by randomizing the strand-strand associations in each original set. The randomization procedure was repeated 10,000 times. To evaluate the significance of the pairing preferences in the natural proteins, their calculated mutual information was compared to the average for the 10,000 random experiments. The mutual information for the natural antiparallel β-sheet pairs (0.35 bits) was found to be outstandingly higher than the mutual information of the randomization experiments (0.080 ± 0.004 bits), indicating that the preferences observed for the correct pairings are very unlikely to occur by chance ($P \ll 10^{-4}$).

Due to the smaller number of parallel *versus* antiparallel pairs (1822 *versus* 8856), the average mutual information for the randomized parallel β-sheet pairs was much higher than for the antiparallel randomizations (0.360 ± 0.013 *versus* 0.080 ± 0.004 bits), though was still significantly lower then the mutual information calculated for the real parallel β-sheet pairs (0.632 bits). This again implies that there is a strong preference for specific binary patterns to pair with one another in parallel β-strands.

Table 1 lists the ten most frequent pairs observed for antiparallel and parallel strands. The log-odds ratios (scores) calculated for each of these pairs are also shown to give an indication of the significance of the pairings. As expected, in both parallel and antiparallel pairs, the preferred associations were those in which the hydrophobic residues and the polar residues matched each other across the two-stranded sheet. In the antiparallel pairs we found the two complementary alternating patterns (PNPNP × PNPNP and NPNPN × NPNPN) to be the most common. These pairs scored significantly higher than expected if pairing was merely random. For the parallel strands, the most frequent pair was the purely hydrophobic one. The pair NNNNN × NNNNP was less frequent than the completely hydrophobic pair but was found much more frequently than expected. Though the pairing preferences were very different between the parallel and antiparallel associations, in both types of pairs the most frequent and significant pairings

Table 2. Binary patterns in short peptide sequences involved in amyloid formation

Protein name	Core peptide sequence	Binary pattern	Reference
Gelsolin	SFNNGDCFILD	PNPPNPPNNNP	[43]
Transthyretin	1. CPLMVKVLDAV 2. YTIAALLSPYS	PNNNNPNNPNN PPNNNNNPNNP	[44]
Lactaherin	NFGSVQFV	PNNPNPNN	[45]
Islet amyloid polypeptide	1. SNNFGAILSS 2. TNVGSNTY	PPPNNNNNPP PPNNPPPP	[46]
Serun amyoild A	SFFSFLGEAFD	PNNPNNNPNNP	[47]
β-Amyloid peptide	QKLVFFAEDVGSNK	PPNNNNNPPNNPPP	[48]
PrP(CJD)	PHGGGWGQ	NNNNNNNP	[49]
PrP Scrapie	1. AGAAAGA 2. DCVNITIKQHTVTT 3. DIKIMERVVEQMCTTQY	NNNNNNN PPNPPPNPPNPNPP PNPNNPPNNPPNPPPPP	[50]
Sup35p	PQGGYQQYN	NPNNPPPPP	[51]

were between patterns resulting in zero, one, and in one case, two mismatches (e.g. N residue paired with a P residue). The lowest scoring associations were between patterns with four and five mismatches (data not shown).

Binary patterns in amyloidogenic sequences

To study the possible involvement of alternating binary patterns in pathological aggregation, we investigated their prevalence in a small set of 15 proteins known to be involved in human diseases (summarized in Table II of Sipe & Cohen[29]). Since the three-dimensional structures for most of these proteins are unknown, for fair comparison we analyzed the overall frequency of all binary patterns in the full sequences of the proteins in our structural data set, ignoring the secondary structure affiliation of the sequence segments. Overall we found that the frequencies of the binary patterns in the amyloidogenic proteins resemble those found in our large set of proteins with two notable exceptions: the patterns PPPPP and NNNNN, which were ranked 31st and 32nd (penultimate and last) in the large protein set, were very common in the small set of amyloidogenic sequences (first for PPPPP and fourth for NNNNN).

When we compared the relative frequencies of the binary patterns in amyloidogenic proteins to all other proteins (Figure 4), excluding the purely polar and purely non-polar patterns, we found a weak correlation ($r = 0.36$) between the two sets of proteins. When we also excluded all five binary patterns with four polar and one non-polar residue that are over-represented in the amyloid proteins (plotted as open diamonds in Figure 4), the correlation of the pattern frequencies between the two sets of proteins increased significantly ($r = 0.65$). Among the patterns that showed strong correlation between amyloidogenic and other proteins were the alternating binary patterns (highlighted in black in Figure 4). These were ranked 14th (PNPNP) and 21st (NPNPN) among amyloidogenic proteins and 12th and 20th in the much larger set of other proteins. In other words, the amyloidogenic proteins have about the same number of alternating patterns as other proteins.

While alternating patterns may not be especially common within the complete sequences of amyloidogenic proteins, it is still possible for them to be over-represented within the sequence segments thought to be primarily responsible for aggregation. We therefore analyzed the minimal sequences that have been shown experimentally to be involved directly in fibril formation in pathological situations (summarized by the CIBA Syp.[30] and Azriel & Gazit[31]). The amino acid sequences of these short polypeptides and their P/N binary translations are shown in Table 2. The simple alternating patterns again did not appear to be a common feature of this set. While sequences with runs of polar residues were not as common in this set as in the full-length amyloid-forming protein sequences, hydrophobic stretches were, similar to what is seen in parallel strands in general. We could not identify a single binary pattern that was found in common to short peptides in this set.

Discussion

Here we analyzed the occurrence of 32 possible pentapeptide binary patterns in a data set of 1911 protein structures, concentrating on parallel and antiparallel β-sheet strands. We found that the alternating patterns (PNPNP and NPNPN) are common in β-strands in general and are often paired with one another. The purely hydrophobic pattern is common in parallel sheets.

Others have studied the occurrence of polar/non-polar patterns in β-sheet structures in general, coming to somewhat different conclusions. In a set of only 48 proteins and using a very strict

definition for hydrophobic (H) residues (including only L, I, V, F and M), Vazquez *et al.*[32] concluded that the PNPNP pattern is suppressed in β-strands. In a later study on a set of 197 proteins, and again using the same definitions for polar and non-polar residues, West & Hecht[33] found the alternating polar/non-polar patterns in the eighth and 14th ranks among all possible 32 binary patterns in β-strand structures overall. However, when their analysis was restricted to exposed β-strands only, they did find the alternating polar/non-polar patterns to be the most common. Similarly, in a recent study of β-strands located on protein surfaces Palliser *et al.*[34] noted that most surface strands have a pattern of the form $(\text{apolar-X})_n$, where X is any amino acid. The difference in the ranking of the alternating patterns between the previous results[32,33] for all β-sheets (including surface and interior strands) and our results could be due to the small data sets used in the earlier studies and to the different definitions used for the polar and non-polar residues. West and Hecht considered H, K, N, D, Q and E to be polar residues and L, I, V and F to be non-polar residues. Residues with intermediate hydrophobicity (A, C, S, T, Y, W and G) were excluded entirely. Thus, patterns including these residues (some of which have high β-sheet propensity such as S, T and Y) were not counted. Different methods were also used to define secondary structure regions. It is worth pointing out, however, that although the rankings of the alternating binary patterns vary in the different studies, when considering their overall frequency, most studies agree that the alternating binary patterns are common in beta structures (accounting for 8–10% of all binary patterns). In agreement with the other studies,[32,33] the most frequent binary patterns in all protein structures were the patterns corresponding to α-helices: PPNNP, NPPNN, and PNNPP.

We report that the binary patterns common in parallel and antiparallel β-strands differ significantly, and that the purely alternating patterns are more common to antiparallel β-strands. They are less common in parallel strands most likely because parallel sheets tend to be buried. Accordingly, non-polar-rich patterns are the most prevalent in the parallel strands. However, the alternating patterns are still significantly more common in parallel strands than expected on the basis of the relative content of polar and non-polar residues in these strands. Therefore, we propose that purely alternating patterns (PNPNP and NPNPN) are a general feature of β-sheets and probably play a role in determining sheet topology. This argument is supported by the observation that patterns in adjoining strands have complementary patterns. The distributions of patterns that we observe in parallel and antiparallel strands could potentially assist with the refinement of protein structure prediction.

How does nature prevent the self-assembly of these β-strands in natural proteins? Recently, Richardson & Richardson[27] have shown that β-sheet edges exploit many different strategies that should prevent edge-to-edge aggregation. For example, by placing a charged residue within the edge strand, extension of the sheet is disfavored because it would involve burying the charge. Applying this strategy, Wang & Hecht[35] redesigned a combinatorial library of alternating binary patterns of β-strands by replacing one of the edge strands, PNPNPNP, with PNPKPNP, presenting a lysine on the non-polar face of the strand. Characterization of these redesigned proteins showed that in comparison to the oligomers generated from the original combinatorial library, the proteins in this experiment formed monomers. Though our comprehensive analysis of β-sheet structures shows that overall, antiparallel edge strands still favor the alternating polar/non-polar patterns, we also found that patterns with three consecutive polar residues occur frequently.

Even though alternating N/P patterns are common in β-sheets, it is none-the-less possible that they also promote protein aggregation in certain circumstances. When we examined the occurrence of the 32 binary patterns in proteins known to be involved in pathological aggregation, we did not find an overabundance of alternating patterns. However, we did note a higher frequency of the pure polar and the pure hydrophobic patterns in the amyloidogenic proteins. Some amyloidogenic proteins are indeed rich in polar residues (specifically Gln and Asn) that have been postulated to interact *via* "polar zippers".[36–38] The abundance of hydrophobic stretches in the minimal amyloidogenic sequences is reminiscent of what we observe in parallel β-strands and coincides well with the structural evidence that amyloid is composed of parallel β-sheets (e.g. Balbirnie *et al.*[38]). Although the number of sequences that have been isolated from amyloidogenic proteins and verified experimentally to form fibrils is very limited, the lack of any simple alternating binary pattern in these sequences suggests that fibril formation related to amyloid diseases is not related to alternating polar/non-polar patterns. Though alternating patterns tend to form aggregates in isolation, we propose that they are important for natural β-sheet formation and are frequently used by nature to make well-folded globular proteins.

Methods

All proteins analyzed here were extracted from the Protein Data Bank (PDB).[39] Only structures solved by X-ray crystallography with resolution >2.5 Å were selected. Proteins with more than 35% sequence identity to others in the data set were removed, resulting in a non-redundant set of 1911 proteins. Parallel and antiparallel β-strands were defined using the program DSSP.[40] Only β-strand sequences that consisted of five or more consecutive β-strand residues were considered in the analysis. Strands with β-bulges were excluded from the analysis. Parallel and antiparallel strands

(buried in sheet, each has two β-strand partners), parallel and antiparallel edge strands (containing only one β-strand partner) and mixed strands (containing one parallel and one antiparallel strand partner) (Figure 1) were analyzed both separately and together. The frequencies of each binary pattern were calculated in all overlapping windows of length five along the β-strand sequences.

Amino acids were classified as polar and non-polar as follows: P = {S, T, N, Q, Y, C, K, R, H, D, E}, N = {G, A, V, L, I, M, P, F, W}. This classification corresponds roughly to the octanol-to-water free energy of transfer scale of Fauchere & Pliška,[41] although C and Y were assigned to the polar category since their side-chains possess hydrogen bonding groups. Different partitionings of the N and P categories were also examined (the above grouping but with Y and C in the non-polar group as well as an alternative partitioning suggested by Wang & Wang[42]).

To study the pairing preferences, we enumerated all possible 32 × 32 pairs of pentapeptide patterns in associated antiparallel and parallel β-strands. Pentapeptide pairs were counted only if each of the strands forming the pairs had five consecutive β-sheet residues (with no bulges). The frequencies of the different pentapeptide strand pairs were normalized by the frequencies expected if the associations between the patterns were random (the product of the frequency of each individual binary pattern).

Acknowledgements

We thank Drs Michael Hecht and Ritu Khurana for their helpful comments on the manuscript, Dr Ehud Gazit for his advice regarding amyloid sequences, and Dr Kevin Karplus for advice on statistics. This work was supported by the California Division-American Cancer Society fellowship to Y.M.G. and by NIH grant GM52885 to L.M.G.

References

1. Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Advan. Protein Chem.* **14**, 1–63.
2. Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry*, **24**, 1501–1509.
3. Kendrew, J. C., Dickerson, R. E., Stranberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. (1960). Structure of myoglobin a three-dimensional fourier synthesis at 2 Å resolution. *Nature*, **185**, 422–427.
4. Perutz, M. F. (1965). Structure and function of haemoglobin I. A tentative atomic model of horse oxyhaemoglobin. *J. Mol. Biol.* **13**, 646–668.
5. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). Structure and function of haemoglobin II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **13**, 669–678.
6. Lim, V. I. (1974). Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.* **88**, 857–872.
7. Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl Acad. Sci. USA*, **81**, 140–144.
8. Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advan. Enzymol. Relat. Areas Mol. Biol.* **47**, 45–148.
9. O'Neil, K. T. & DeGrado, W. F. (1990). A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science*, **250**, 646–651.
10. Padmanabhan, S., Marqusee, S., Ridgeway, T., Laue, T. M. & Baldwin, R. L. (1990). Relative helix-forming tendencies of nonpolar amino acids. *Nature*, **344**, 268–270.
11. Kim, C. A. & Berg, J. M. (1993). Thermodynamic beta-sheet propensities measured using a zinc-finger host peptide. *Nature*, **362**, 267–270.
12. Minor, D. L. & Kim, P. S. (1994). Measurement of the beta-sheet-forming propensities of amino acids. *Nature*, **367**, 660–663.
13. Smith, C. K., Withka, J. M. & Regan, L. (1994). A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. *Biochemistry*, **33**, 5510–5517.
14. Minor, D. L. & Kim, P. S. (1994). Context is a major determinant of beta-sheet propensity. *Nature*, **371**, 264–267.
15. Dalal, S., Balasubramanian, S. & Regan, L. (1997). Protein alchemy: changing beta-sheet into alpha-helix. *Nature Struct. Biol.* **4**, 548–552.
16. Xiong, H., Buckwalter, B. L., Shieh, H. M. & Hecht, M. H. (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.
17. Dill, K. A. & Chan, H. S. (1997). From levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10–19.
18. Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science*, **262**, 1680–1685.
19. West, M. W., Wang, W., Patterson, J., Mancias, J. D., Beasley, J. R. & Hecht, M. H. (1999). *De novo* amyloid proteins from designed combinatorial libraries. *Proc. Natl Acad. Sci. USA*, **96**, 11211–11216.
20. Xu, G., Wang, W., Groves, J. T. & Hecht, M. H. (2001). Self-assembled monolayers from a designed combinatorial library of *de novo* beta-sheet proteins. *Proc. Natl Acad. Sci. USA*, **98**, 3652–3657.
21. Davidson, A. R. & Sauer, R. T. (1994). Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl Acad. Sci. USA*, **91**, 2146–2150.
22. Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. (1997). Functional rapidly folding proteins from simplified amino acid sequences. *Nature Struct. Biol.* **4**, 805–809.
23. Brack, A. & Orgel, L. E. (1975). Beta structures of alternating polypeptides and their possible prebiotic significance. *Nature*, **256**, 383–387.
24. Zhang, S., Lockshin, C., Cook, R. & Rich, A. (1994). Unusually stable beta-sheet formation in an ionic self-complementary oligopeptide. *Biopolymers*, **34**, 663–672.
25. Lim, A., Saderholm, M. J., Makhov, A. M., Kroll, M., Yan, Y., Perera, L. *et al.* (1998). Engineering of beta-bellin-15D: a 64 residue beta sheet protein that

forms long narrow multimeric fibrils. *Protein Sci.* **7**, 1545–1554.
26. Broome, B. M. & Hecht, M. H. (2000). Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. *J. Mol. Biol.* **296**, 961–968.
27. Richardson, J. S. & Richardson, D. C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA*, **99**, 2754–2759.
28. Sternberg, M. J. & Thornton, J. M. (1977). On the conformation of proteins: hydrophobic ordering of strands in beta-pleated sheets. *J. Mol. Biol.* **115**, 1–17.
29. Sipe, J. D. & Cohen, A. S. (2000). Review: history of the amyloid fibril. *J. Struct. Biol.* **130**, 88–98.
30. The nature and structure of amyloid structures (1996). Ciba Foundation Symposium, London, UK.
31. Azriel, R. & Gazit, E. (2001). Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. An experimental support for the key role of the phenylalanine residue in amyloid formation. *J. Biol. Chem.* **276**, 34156–34161.
32. Vazquez, S., Thomas, C., Lew, R. A. & Humphreys, R. E. (1993). Favored and suppressed patterns of hydrophobic and nonhydrophobic amino acids in protein sequences. *Proc. Natl Acad. Sci. USA*, **90**, 9100–9104.
33. West, M. W. & Hecht, M. H. (1995). Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci.* **4**, 2032–2039.
34. Palliser, C. C., MacArthur, M. W. & Parry, D. A. (2000). Surface beta-strands in proteins: identification using an hydropathy technique. *J. Struct. Biol.* **132**, 63–71.
35. Wang, W. & Hecht, M. H. (2002). Rationally designed mutations convert *de novo* amyloid-like fibrils into monomeric beta-sheet proteins. *Proc. Natl Acad. Sci. USA*, **99**, 2760–2765.
36. Perutz, M. F., Johnson, T., Suzuki, M. & Finch, J. T. (1994). Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc. Natl Acad. Sci. USA*, **91**, 5355–5358.
37. Michelitsch, M. D. & Weissman, J. S. (2000). A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. *Proc. Natl Acad. Sci. USA*, **97**, 11910–11915.
38. Balbirnie, M., Grothe, R. & Eisenberg, D. S. (2001). An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated beta-sheet structure for amyloid. *Proc. Natl Acad. Sci. USA*, **98**, 2375–2380.
39. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
40. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
41. Fauchere, J. & Pliska, V. (1983). Hydrophobic parameters of amino acid-side-chain from partitioning of n-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369–375.
42. Wang, J. & Wang, W. (1999). A computational approach to simplifying the protein folding alphabet. *Nature Struct. Biol.* **6**, 1033–1038.
43. Maury, C. P. & Nurmiaho-Lassila, E. L. (1992). Creation of amyloid fibrils from mutant Asn187 gelsolin peptides. *Biochem. Biophys. Res. Commun.* **183**, 227–231.
44. Gustavsson, A., Engström, U. & Westermark, P. (1991). Normal transthyretin and synthetic transthyretin fragments form amyloid-like fibrils *in vitro*. *Biochem. Biophys. Res. Commun.* **175**, 1159–1164.
45. Häggqvist, B., Näslund, J., Sletten, K., Westermark, G. T., Mucchiano, G., Tjernberg, L. O. *et al.* (1999). Medin: an integral fragment of aortic smooth muscle cell-produced lactadherin forms the most common human amyloid. *Proc. Natl Acad. Sci. USA*, **96**, 8669–8674.
46. Padrick, S. B. & Miranker, A. D. (2001). Islet amyloid polypeptide: identification of long-range contacts and local order on the fibrillogenesis pathway. *J. Mol. Biol.* **308**, 783–794.
47. Wouters, M. A. & Curmi, P. M. (1995). An analysis of side chain interactions and pair correlations within antiparallel beta-sheets: the differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins: Struct. Funct. Genet.* **22**, 119–131.
48. Gorevic, P. D., Castano, E. M., Sarma, R. & Frangione, B. (1987). Ten to fourteen residue peptides of Alzheimer's disease protein are sufficient for amyloid fibril formation and its characteristic X-ray diffraction pattern. *Biochem. Biophys. Res. Commun.* **147**, 854–862.
49. Prusiner, S. B., Scott, M. R., DeArmond, S. J. & Cohen, F. E. (1998). Prion protein biology. *Cell*, **93**, 337–348.
50. Gasset, M., Baldwin, M. A., Lloyd, D. H., Gabriel, J. M., Holtzman, D. M., Cohen, F. *et al.* (1992). Predicted alpha-helical regions of the prion protein when synthesized as peptides form amyloid. *Proc. Natl Acad. Sci. USA*, **89**, 10940–10944.
51. Patino, M. M., Liu, J. J., Glover, J. R. & Lindquist, S. (1996). Support for the prion hypothesis for inheritance of a phenotypic trait in yeast. *Science*, **273**, 622–626.

Edited by B. Honig

(Received 18 June 2002; received in revised form 16 August 2002; accepted 30 August 2002)