doi:10.1006/jmbi.2000.3514 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 296, 961–968

JMB

COMMUNICATION

Nature Disfavors Sequences of Alternating Polar and Non-polar Amino Acids: Implications for Amyloidogenesis

Bede M. Broome and Michael H. Hecht*

Department of Chemistry
Princeton University
Princeton, NJ 08544-1009
USA

**Corresponding author*

Recent experiments with combinatorial libraries of *de novo* proteins have demonstrated that sequences designed to contain polar and non-polar amino acid residues arranged in an alternating pattern form fibrillar structures resembling β-amyloid. This finding prompted us to probe the distribution of alternating patterns in the sequences of natural proteins. Analysis of a database of 250,514 protein sequences (79,708,024 residues) for all possible binary patterns of polar and non-polar amino acid residues revealed that alternating patterns occur significantly less often than other patterns with similar compositions. The under-representation of alternating binary patterns in natural protein sequences, coupled with the observation that such patterns promote amyloid-like structures in *de novo* proteins, suggests that sequences of alternating polar and non-polar amino acids are inherently amyloidogenic and consequently have been disfavored by evolutionary selection.

Keywords: amyloid; protein aggregation; fibril; protein design; binary patterning

The two common types of secondary structure in proteins each have a distinct structural periodicity. For α-helices the periodicity is 3.6 residues per repeat, while for β-strands it is ~2 residues per repeat (Creighton, 1993). Consequently, for a segment of secondary structure to present one face that is entirely hydrophilic and an opposing face that is entirely hydrophobic, this inherent structural periodicity must be matched by a corresponding periodicity of polar and non-polar amino acids in the linear sequence of the polypeptide. Hence, for amphiphilic β-strands the sequence periodicity of polar (○) and non-polar (●) amino acids must be arranged with an alternating pattern (...○●○●○●○...), whereas for amphiphilic α-helices, the sequence periodicity must place a non-polar residue every three or four positions (e.g. ○●●○○●●○○●○○●●●○).

Amphiphilic secondary structures composed of appropriately patterned amino acid sequences are driven by the hydrophobic effect to bury their non-polar faces against one another. Such interactions favor the self-assembly of amphiphilic peptides into multimeric structures (Xiong *et al.*, 1995) and play an important role in the folding of globular proteins into tertiary structures with hydrophilic surfaces and hydrophobic cores (Bowie *et al.*, 1990).

Polar/non-polar sequence patterning can be represented by a ''binary code'', which specifies the type of amino acid (polar *versus* non-polar) at each position, but not its precise identity. Because several different polar and non-polar side-chains are allowed, a given binary pattern is consistent with many different amino acid sequences. This potential for diversity has allowed binary patterning to serve as the primary constraint in designing large combinatorial libraries of *de novo* proteins (Kamtekar *et al.*, 1993). In such libraries the sequence locations of polar and non-polar residues are designed *a priori*. However, since the exact identity of the polar or non-polar residue at each location is not specified, combinatorially diversity is facilitated. This constrained diversity is made possible by the organization of the genetic code in which polar residues are encoded by the degenerate DNA codon NAN, and non-polar residues are

Present address: B. M. Broome, UCLA School of Medicine, Los Angeles, CA, USA.

E-mail address of the corresponding author: hecht@princeton.edu encode by the degenerate codon NTN (N denotes a mixture of DNA bases).

We previously reported the use of binary patterning to design combinatorial libraries of four-helix bundles (Kamtekar *et al.*, 1993; Roy *et al.*, 1997; Rojas *et al.*, 1997; Moffet *et al.*, unpublished results). Our designed binary pattern used four copies of the α-helical pattern, ○●○○●●○○●○○●●○. These were separated by short sequences designed to form helix caps and inter-helical turns.

Characterization of proteins isolated from that library demonstrated that they indeed folded into α-helical structures. None of the proteins isolated from the α-helical library formed large aggregates (Kamtekar *et al.*, 1993; S. Roy & M.H.H., unpublished results), suggesting that the designed amphiphilic helices bury their non-polar faces by intramolecular folding rather than by inter-molecular aggregation.

Recently, we have extended the binary code strategy by designing libraries of *de novo* β-strand structures (West *et al.*, 1999). In this case, our design was based on the alternating pattern, ○●○●○●○. Six copies of this pattern were interspersed with segments designed to form turns. Proteins isolated from this library exhibit properties dramatically different from those of the α-helical library: (i) they form structures dominated by β-sheet secondary structure; and (ii) these β-structures do not result from intra-molecular folding into monomers or small oligomers, but occur in the context of large inter-molecular aggregates. Electron microscopy and atomic force microscopy of these aggregates revealed long thin fibrils. Further characterization demonstrated that these fibrils recapitulate many properties of the amyloid structures found in Alzheimer's disease and the prion-related diseases (West *et al.*, 1999).

Why did the first library yield α-helical structures that fold intra-molecularly into small globular domains, while the second library yielded β-strands that assemble inter-molecularly into large aggregates resembling amyloid? The lengths of the sequences encoded by these two libraries are not dramatically different (74 *versus* 63 residues), nor are their overall amino acid compositions. The major difference is the binary patterning itself. In contrast to the ○●○○●●○○●○○●●○ pattern, which favored folding into α-helical globular proteins, the alternating pattern, ○●○●○●○, apparently predisposes sequences to form β-strands that self-assemble into long fibrils.

If alternating patterns favor amyloid-like structures, one might expect such patterns to be toxic, and therefore disfavored by natural selection. To test this possibility, we analyzed a database of 250,514 natural protein sequences comprising 79,708,024 residues, and calculated the frequencies of alternating patterns relative to other patterns with similar compositions. The results of this search demonstrate that for "windows" ranging from five to ten residues, alternating patterns occur significantly less frequently than would be expected from the overall composition of the database.

The protein sequences used in this study were derived from the OWL database (Bleasby *et al.*, 1994) as obtained from the Bioinformatics Group at the University of Leeds. This database emphasizes stringent non-redundancy, which is extremely important for assessing the relative frequencies of various sequence patterns.

Binary patterns were defined as any continuous string of polar and non-polar amino acids with lengths ranging from five to ten residues. Classification of a residue as either polar or non-polar was based upon two criteria: (i) the hydrophobicity scale described by Fauchere & Pliska (1983), which is derived from the partitioning of amino acids between octanol and water; and (ii) the binary organization of the genetic code. In general, these two criteria led to the same choices of polar and non-polar amino acids. Residues included in the polar category were Arg, Lys, Asp, Glu, Asn, Gln and His. These are the seven most polar amino acids in octanol/water partitioning experiments (Fauchere & Pliska, 1983), and all of them except the Arg residues are encoded by the degenerate codon XAN. (N represents any base. X represents C, A, or G. Exclusion of T from the first position prevents stop codons and tyrosine residues.) Residues included in the non-polar category were Leu, Ile, Val, Phe and Met. These residues are encoded by the degenerate codon NTN. The polar and non-polar amino acids used in this study are the same as those used in the earlier work by West & Hecht (1995).

The expected probability for any binary pattern of polar and non-polar amino acids can be calculated from the amino acid composition of the database as a whole. In the OWL database the polar residues Arg, Lys, Asp, Glu, Asn, Gln and His occur 26,426,173 times and represent 33.15 % of the total number of residues. The non-polar residues Leu, Ile, Val, Phe and Met occur 22,337,911 times and represent 28.02 % of the total number of residues. Therefore at any given position, the expected probability of finding a polar residue (P_p) is 0.3315, and the expected probability of finding a non-polar residue (P_n) is 0.2802. For a sequence pattern of length L, with X non-polar residues, the expected probability of that pattern is $[(P_n)^x(P_p)^{L-X}]$. For example, for the pattern ○●○●○●○, $L = 7$ and $X = 3$. Thus the expected probability of this pattern is $(0.2802)^3 (0.3315)^4 = 0.000266$. All cassettes containing the same number of polar and non-polar residues (e.g. four polar and three non-polar in this example) are expected to occur with the same probability.

The number of times a given pattern would be expected to occur in the database is simply the probability of that pattern multiplied by the number of windows of length L in the database. For a database represented as a continuous string of J

residues, the number of windows of length L is $[J - L + 1]$. When $J \gg L$, this approaches J. Thus, the expected number of occurrences for a pattern is $[(P_{\mathrm{n}})^X(P_{\mathrm{p}})^{L-X}(J)]$. For example, for any pattern with three non-polar residues and four polar residues (e.g. ○●○●○●○), the expected number of occurrences is $(0.2802)^3$ $(0.3315)^4$ 79,708,024 = 21,236. All patterns containing three non-polar and four polar residues would be expected to occur with this same frequency.

Although all patterns with the same polar/non-polar composition are expected to occur with the same frequency, our search of the database demonstrated that such equal representation is not observed: some patterns occur far less frequently than others. As shown in Figure 1, the alternating binary patterns occur significantly less often than other patterns with the same polar/non-polar composition. For cassettes between five and nine residues, the alternating patterns are the absolute rarest (Figure 1(a)-(h)). For example, there are 20 possible ways of arranging three polar and three non-polar residues in a string of six residues. The two alternating patterns ○●○●○● and ●○●○●○ rank 19th and 20th (see Figure 1(c)). Likewise, there are 70 ways of arranging four polar and four non-polar residues in a string of eight residues. The two alternating patterns ○●○●○●○● and ●○●○●○●○ rank as 69th and 70th (Figure 1(f)). For windows of ten residues, the alternating patterns are not the absolute rarest. Yet even for these longer cassettes, the alternating patterns are among the least common (Table 1).

The number of occurrences of an alternating pattern can be compared with the mean number of occurrences for all patterns with the same composition. For all windows ranging from five to ten residues, alternating patterns invariably occur at least one standard deviation below the mean (Table 1).

The results shown in Figure 1 and Table 1 demonstrate that in a large database of protein sequences derived from many different organisms, alternating patterns of polar and non-polar residues are significantly underrepresented relative to other patterns with the same polar/non-polar composition.

Why has evolution selected against alternating binary patterns? It is difficult (and often impossible) to determine conclusively why evolution disfavors some options relative to others. However, one can study the properties of disfavored options ("the road less traveled") by constructing artificial systems based on the very features that evolution has rejected. For the alternating binary patterns, such an experiment has been done. The alternating seven-residue pattern ○●○●○●○, which ranks 35th among the 35 possible patterns in Figure 1(e), was used in the design of two combinatorial libraries of *de novo* sequences (West, 1997; West *et al.*, 1999). One library contained six repeats of the ○●○●○●○ pattern punctuated by turns, while the other library contained eight repeats of this pattern punctuated by turns. *De novo* proteins from both libraries were purified and shown to form fibrils resembling those seen in amyloid diseases (West *et al.*, 1999; Wang & M.H.H., unpublished results). Because these proteins were derived from combinatorial libraries, their individual sequences differed from one another considerably. The key feature these diverse sequences shared in common was the alternating pattern of polar and non-polar amino acids. Hence, it appears that this patterning drives their assembly into amyloid-like structures.

Table 1. Alternating patterns of polar and non-polar residues occur less frequently than other patterns with the same polar/non-polar composition

Length	Composition	Alternating pattern	Number of patterns possible with this composition	Rank of alternating pattern[a]	Mean number of occurrences for all patterns with this composition	Number of occurrences of the alternating pattern	Standard deivation below mean[b]
5	2 ○ and 3 ●	●○●○●	10	10	196,961	138,995	1.51
5	3 ○ and 2 ●	○●○●○	10	10	283,105	210,866	1.6
6	3 ○ and 3 ●	●○●○●○	20	20	77,825	51,235	1.42
		○●○●○●		19	77,825	51,975	1.38
7	3 ○ and 4 ●	●○●○●○●	35	35	19,767	10,546	1.62
7	4 ○ and 3 ●	○●○●○●○	35	35	30,876	20,349	1.19
8	4 ○ and 4 ●	●○●○●○●○●	70	70	8142	4039	1.49
		○●○●○●○●○		69	8142	4085	1.47
9	4 ○ and 5 ●	●○●○●○●○●	126	125	1966	897	1.44
9	5 ○ and 4 ●	○●○●○●○●○	126	126	3277	1652	1.30
10	5 ○ and 5 ●	●○●○●○●○●○	252	232	817	466	1.02
		○●○●○●○●○●		231	817	469	1.01

Each line represents a "window" of a given length and a particular composition of polar (○) and non-polar (●) residues.

[a] Rank of binary patterns is shown graphically in Figure 1.

[b] Standard deviation for the number of occurrences of the alternating pattern relative to the mean number of occurrences for all binary patterns with the same specified composition.

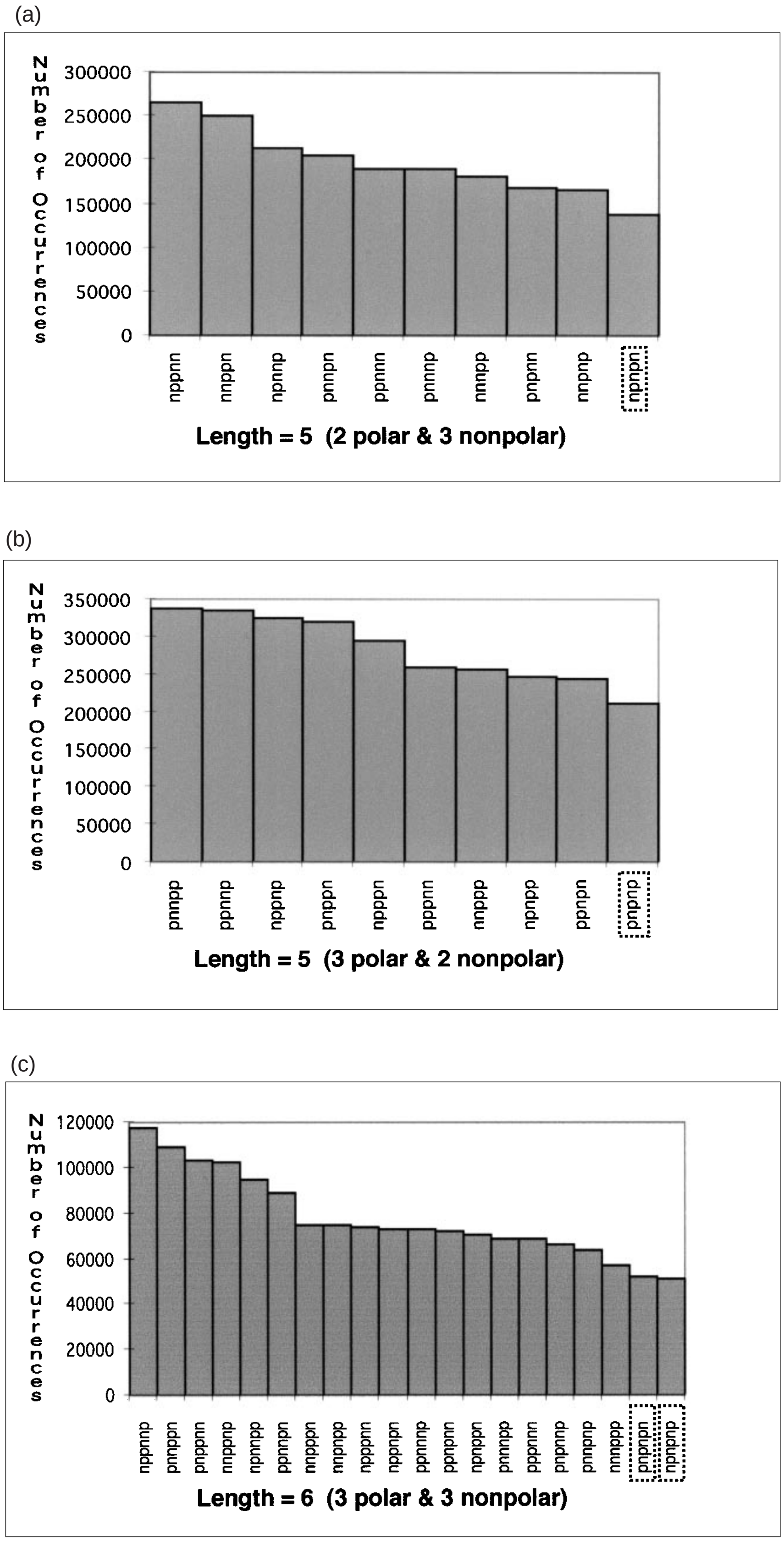

Figure 1 (*legend shown on page 966*)

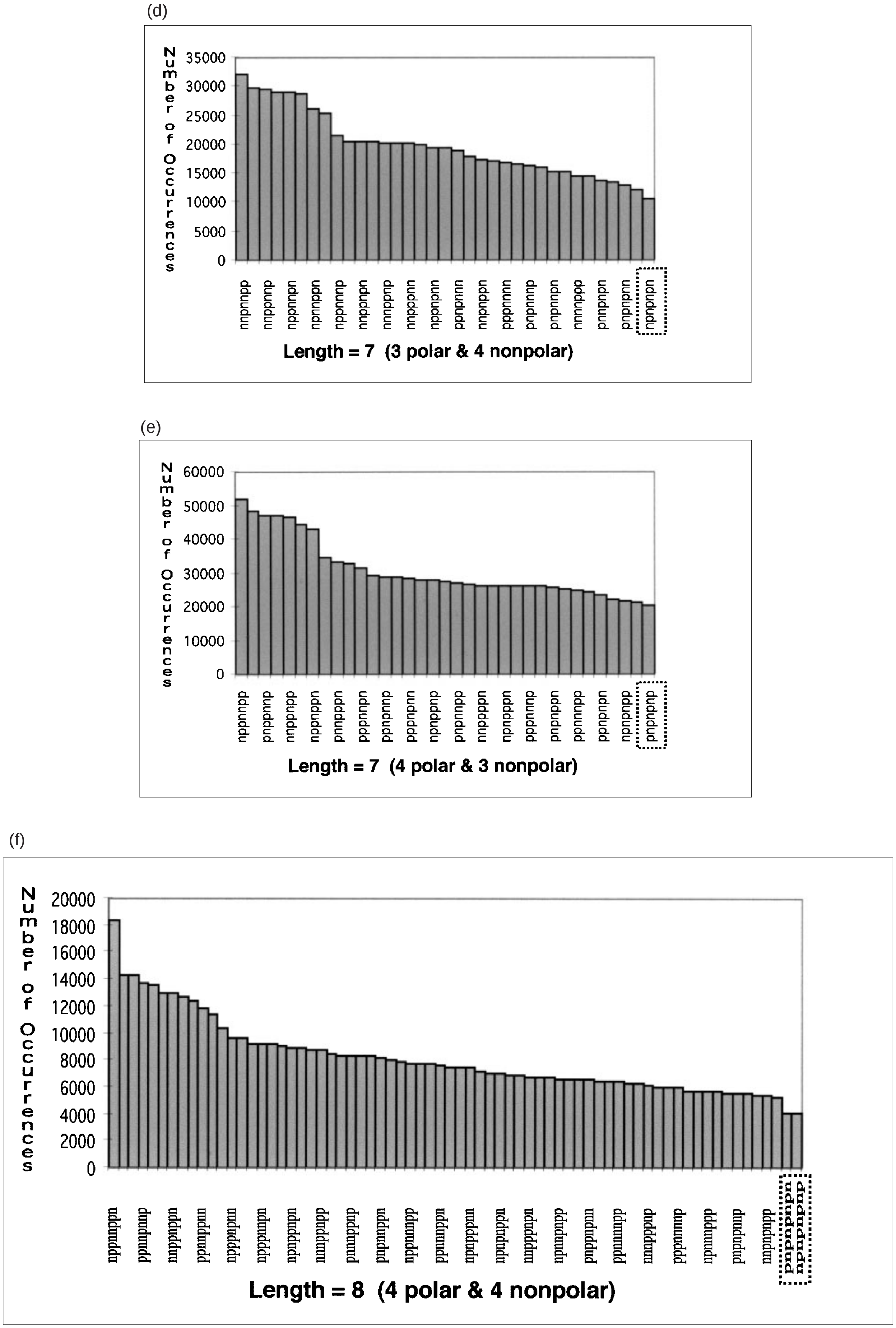

Figure 1 (*legend shown on page 966*)

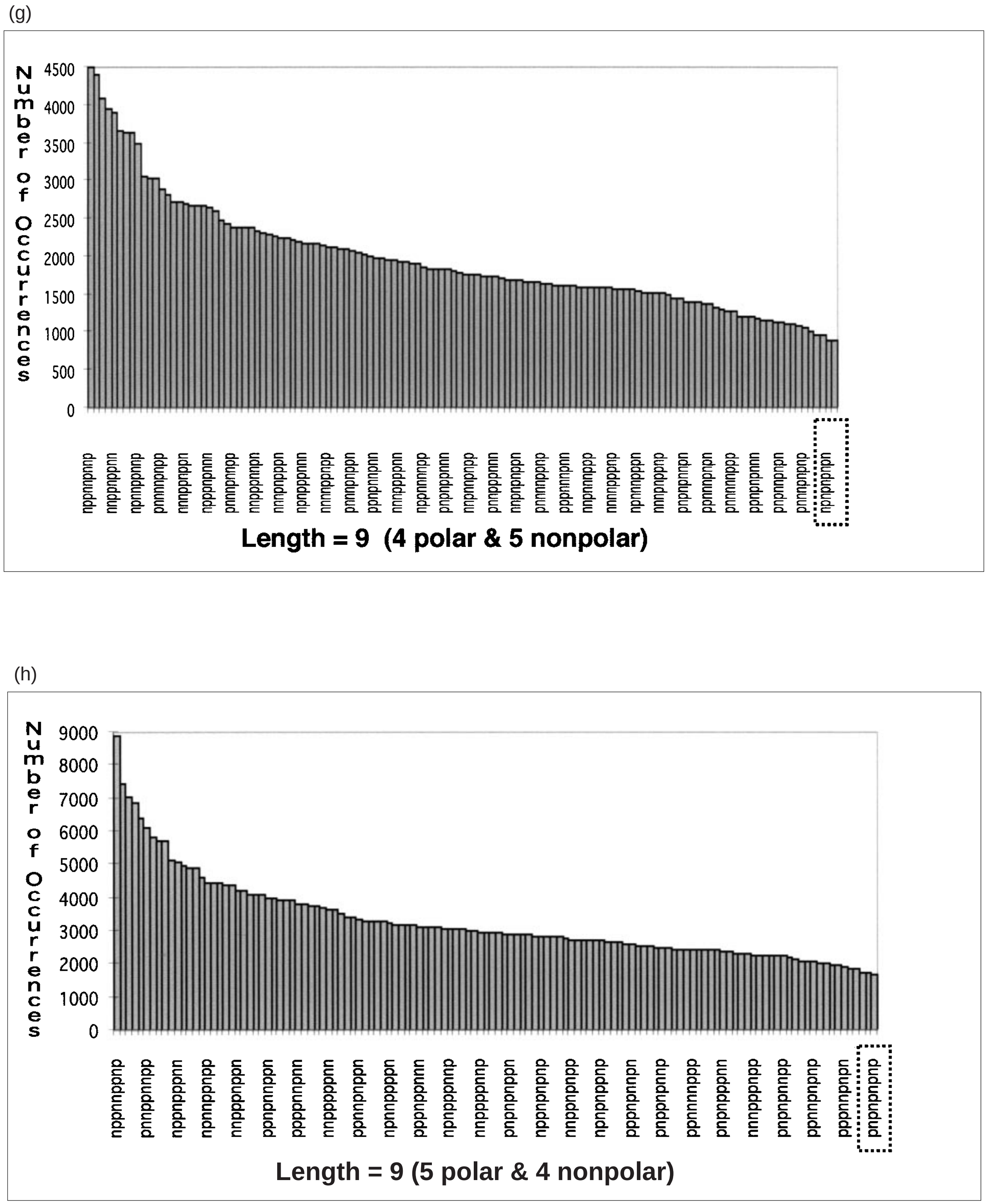

Figure 1. Number of occurrences of various binary patterns of polar and non-polar amino acids. The p and n denote polar and non-polar amino acids, respectively. (a)-(h) Tabulation for a particular length and composition. The alternating patterns are highlighted with a broken rectangle. Composition of database: Arg, 4,147,245 (5.20 %); Lys, 4,634,596 (5.81 %); Asp, 4,087,450 (5.13 %); Glu, 4,961,998 (6.23 %); Asn, 3,644,945 (4.57 %); Gln, 3,162,560 (3.97 %); His, 1,787,379 (2.24 %); Phe, 3,284,591 (4.12 %); Ile, 4,608,150 (5.78 %); Leu, 7,421,844 (9.31 %); Met, 1,849,465 (2.32 %); and Val, 5,173,861 (6.49 %). Overall, the number of polar residues is 26,426,173 (33.15 %), the number of non-polar residues is 22,337,911 (28.02 %), and the number of ''other'' residues is 30,943,940 (38.82 %). To calculate the number of occur-

Several other experimental studies have been performed on artificial systems containing alternating patterns of polar and non-polar residues. In pioneering work, Brack & Orgel (1975) demonstrated that polymers composed of alternating valine and lysine residues form β-structures that aggregate into large assemblies. More recently, Janek *et al*. (1999) showed that alternating valine and lysine residues can serve as model systems for studies of fibril self-assembly. Zhang *et al*. (1993) found that 16-residue peptides composed of sequences that alternate between an alanine residue and a charged residue spontaneously assemble into large oligomeric structures dominated by β-sheet. Finally, Lim *et al*. (1998) described a *de novo* protein, Betabellin-15D, formed by the disulfide-mediated dimerization of a 32-residue peptide. The 32-mer was designed to form four β-strands, each containing six residues of alternating polar and non-polar residues. The resulting proteins formed structures visible in electron microscopy images as long narrow multimeric fibrils.

The experimental results summarized above demonstrate that artificial sequences containing alternating patterns have a strong propensity to aggregate into amyloid-like fibrils. In natural systems, a similar tendency to form amyloid would compete with proper folding into globular protein structures. We propose that because such misfolding would present a severe disadvantage to the host organism, the alternating patterns were disfavored by evolutionary selection.

Based on these considerations one might expect the sequences of natural amyloidogenic proteins to be rich in alternating polar/non-polar patterns. We found, however, that the patterns in these sequences are similar to those in the database as a whole. Although initially surprising, this similarity is exactly what should be expected. The amyloid found in diseased tissues is an off-pathway misfolded structure (Wetzel, 1997; Kelly *et al*., 1997). The amino acid sequences of amyloidogenic proteins did not evolve to form amyloid. They evolved to fold into globular structures capable of performing functions beneficial to the organism. As expected, their amino acid sequences reflect this evolutionary history.

In addition to noting the under-representation of alternating patterns, we also asked which patterns are preferred. Analysis of the 250,514 sequences in the database showed that binary patterns consistent with amphiphilic α-helices occur considerably more often than other patterns with the same polar/non-polar composition. For example, the binary pattern used in the designed α-helical library of Kamtekar *et al.* (1993) was the 14-mer ○●○○●●○○●○○●●○. Analysis of the database showed that for windows ranging from five to ten residues, segments of this amphiphilic α-helical pattern are preferred over other patterns with the same polar/non-polar composition. These findings are consistent with earlier studies on smaller databases by Vazquez *et al*. (1993) and by our own group (West & Hecht, 1995)

In summary, why does nature favor binary patterns consistent with amphiphilic α-helices while disfavoring those consistent with amphiphilic β-strands? Studies of designed combinatorial libraries suggest that for the construction of globular proteins, these two types of patterns have very different properties: patterns consistent with amphiphilic α-helices fold intramolecularly into small soluble structures. Hence, they are well-suited for constructing globular proteins and were favored by evolution. In contrast, patterns consistent with amphiphilic β-strands promote aggregation into amyloid-like fibrils. Such aggregation competes with (and may prevent) intramolecular folding. Consequently, alternating patterns are poorly fit for constructing globular proteins and were disfavored by evolutionary selection.

Figure 1 (*Legend continued*)

rences for any given binary pattern, we treated the entire database as a continuous string of J residues. In a continuous string, the number of possible windows of length L is $(J - L + 1)$. When $J \gg L$, this simplifies to J, and so the number of expected occurrence for a pattern is $[(P_n)^X\ (P_p)^{L-X}\ (J)]$ (see the text). In reality, however, the database is not a single string of amino acids, but 250,514 individual protein sequences. Therefore some of the windows counted in our calculation cannot actually contain patterns (they would span across different protein sequences). To assess whether this might lead to spurious results, we compared the number of windows that should have been excluded due to end effects with the number of windows in our continuous string of 79,708,024 residues. For M individual protein sequences, the number of windows that should be excluded due to end effects is $M(L - 1)$. For example, when $L = 7$, the number of windows that should be excluded is $250{,}514\ (7 - 1) = 1{,}503{,}084$. The number of windows in our continuous string of 79,708,024 residues was $[J - L + 1] = [79{,}708{,}024 - 7 + 1] = 79{,}708{,}018$. Thus the "forbidden" windows included in the continuous string comprise less than 2 % of the total. Software to search the database and format the data was written in C and run on a SGI Indigo R3000 processor. Programs can be obtained from bmbroome@princeton.edu. The search employs a one pass sliding window, which tests the database protein sequences against a target cassette sequence. Only exact matches are recorded. For a collection of sequences (J = total length of database as a single string) the search runs in $K(J)$ time, where K equals the number of patterns tested in a single pass. Several machine specific optimizations were incorporated to increase the actual search speed.

Acknowledgments

We thank Michael West, Laura Landweber and Tzachi Pilpel for comments on the manuscript. This work was supported by the Culpepper Foundation, the Biological Sciences Program of the ARO, the New Jersey Center for Biomaterials, and NSF MRSEC (DMR98-09483).

References

Bleasby, A. J., Akrigg, D. & Attwood, T. K. (1994). OWL: a non-redundant composite protein sequence database. *Nucl. Acids Res.* **22**, 3574-3577.

Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (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.

Brack, A. & Orgel, L. E. (1975). β Structures of alternating polypeptides and their possible prebiotic significance. *Nature,* **256**, 383-387.

Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*, 2nd edit., WH Freeman & Company, New York.

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

Janek, K., Behlke, J., Zipper, J., Fabian, H., Georgalis, Y., Beyermann, M., Bienert, M. & Krause, E. (1999). Water soluble β-sheet models which self-assemble into fibrillar structures. *Biochemistry,* **38**, 8246-8252.

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and non-polar amino acids. *Science,* **262**, 1680-1685.

Kelly, J. W., Colon, W., Lai, Z., Lashuel, H. A., McCulloch, J., McCutchen, S. L., Miroy, G. J. & Peterson, S. A. (1997). Transthyretin quaternary and tertiary structural changes facilitate misassembly into amyloid. *Advan. Protein Chem.* **50**, 161-181.

Lim, A., Saderholm, M. J., Makhov, A. M., Kroll, M., Yan, Y., Perera, L., Griffith, J. D. & Erickson, B. W. (1998). Engineering of betabellin-15D: a 64 residue beta sheet protein that forms long narrow multimeric fibrils. *Protein Sci.* **7**, 1545-1554.

Rojas, N. R., Kamtekar, S., Simons, C. T., McLean, J. E., Vogel, K. M., Spiro, T. G., Farid, R. S. & Hecht, M. H. (1997). *De novo* heme proteins from designed combinatorial libraries. *Protein Sci.* **6**, 2512-2524.

Roy, S., Ratnaswamy, G., Boice, J. A., Fairman, R., McLendon, G. & Hecht, M. H. (1997). A protein designed by binary patterning of polar and non-polar amino acids displays native-like properties. *J. Am. Chem. Soc.* **119**, 5302-5306.

Vasquez, S., Thomas, C., Lew, R. A. & Humphreys, R. E. (1993). Favored and suppressed patterns of hydrophobic and nonhydrophobic amino acids in proteins. *Proc. Natl Acad. Sci. USA,* **90**, 9100-9104.

West, M. W. (1997), *De novo* design of a library of β-sheet proteins. PhD thesis, Department of Chemistry, Princeton University.

West, M. W. & Hecht, M. H. (1995). Binary patterning of polar and non-polar amino acids in the sequences and structures of native proteins. *Protein Sci.* **4**, 2032-2039.

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.

Wetzel, R. (1997). Domain stability in immunoglobulin light chain deposition disorders. *Advan. Protein Chem.* **50**, 183-242.

Xiong, H., Buckwalter, B. L., Shieh, H. M. & Hecht, M. H. (1995). Periodicity of polar and non-polar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc. Natl Acad. Sci. USA,* **92**, 6349-6353.

Zhang, S., Holmes, T., Lockshin, C. & Rich, A. (1993). Spontaneous assembly of a self-complementary oligopeptide to form a stable macroscopic membrane. *Proc. Natl Acad. Sci. USA,* **90**, 3334-3338.

Edited by F. E. Cohen

(Received 11 October 1999; received in revised form 3 January 2000; accepted 3 January 2000)