Non-occurring and rare quads in PDB and translated introns from XPro with possible applications in nanostructure design

G. Sampath
g_sampath0@lycos.com

James Ten Eyck
Department of Computer Science
Marist College, Poughkeepsie, NY 12601
james.teneyck@marist.edu

Abstract

Exhaustive search over 17313 unique protein sequences in the database PDB indicates the absence of 4036 of the 160000 possible subsequences of four residues (quads). When the polypeptides obtained by translating 100000 prion sequences in the database XPro are searched the number drops to 424, which still exceeds what would be obtained by pure chance. More generally there are 11444 quads that occur 3 or fewer times in PDB. Using the Kyte-Doolittle hydrophobicity index, the 4036 quads (including the 424 absent in XPro) are divided into 16 groups, five of which can form unbroken helices or sheets by repetition. Most of the 16 groups are evenly distributed, one exception being quads with all-apolar residues, which are significantly less frequent. The helical and sheet structures so formed are artificial polypeptides not observed in nature. By using patterns from the other 11 groups more complex structures can be formed. Such structures could potentially serve as tubules and substrates in nanostructure design.

1. Introduction

We have performed a search of protein sequences in the protein database PDB to find the frequency of occurrence of all distinct subsequences of four residues. The purpose was to learn if there are quads that never, or very rarely, occur in nature. Identifying such rare or non-occurring subsequences provides essential information for addressing the question of why particular patterns are not preferred. Conversely, identifying sequences that occur more frequently than predicted by chance provides information for studying evolutionary pathways.

In our analysis we have identified 4036 quads that are not observed in PDB, a larger number that occur three or fewer times, and 424 that do not occur in the translated sequences from the intron database XPro. The identification of such 'unnatural' sequences motivates one to learn whether structures made from them are stable and useful as tubular or substrate-like structures in nanotechnology.

2. Rare subsequences in proteins

Exhaustive search over 17313 unique protein sequences in the protein database PDB [1] indicates the absence of 4036 of the 160000 possible subsequences of four residues (quads). Additionally they are also absent in prion sequences. When a similar search (in both directions and with frame shifts of 1 and 2 as well) for these 4036 quads over the polypeptides obtained by translating more than 100000 introns in the database XPro [2] the number absent drops to 424, a number that is in excess of what would be obtained by pure chance. That the absence of these 424 quads is not just a chance occurrence is suggested by results obtained when the same sequences are randomly permuted and subjected to the search process: the absences virtually disappear. This latter result was strengthened when a search was done for the 424 quads over runs of randomly generated sequences up to a million residues long and fewer than 5 of them were found to be missing in any run. More generally rare quads are defined as 4-mers occurring no more than K (a small number) times in the entire PDB database. For K = 3, the number of rare quads is 11444.

The choice of 4 for the length of the subsequence to search for was dictated by the pitch of helices in known protein secondary structure (3.4 to 3.6 residues per helical turn). Given the tendency of hydrophilic and hydrophobic residues to be found respectively on the exposed and sheltered (from water) sides of a helix motif the relative propensity of the absent quads to form the three types of secondary structures 'A' (helix), ‘B’ (sheet), and ‘G’ (neither-A-nor-B or ‘other‘) was studied via their polarity properties based on the widely used Kyte-Doolittle hydrophobicity index [3].

If the 4036 non-occurring quads from PDB are coded for polarity with 1 for apolar and 0 for polar, they can be divided into 16 groups. Of these, 3 groups (with codes 1001, 0010, and 0100) can form unbroken helices by repetition in the order given and 2 groups (0101 and 1010) can each form unbroken sheets by repetition (other variations are possible). Most of the groups have frequencies that are evenly distributed, the most notable exception is group 1111, which suggests a relative (but significant) paucity of rare quads with all-apolar residues. Analysis of the 424 quads that are absent in both PDB and the bidirectionally translated XPro reveals the following: even when only amino acids with hydrophobicity ≥ +3 and ≤ -3 are considered, there still are small numbers of quads that can form a helix turn, a sheet, or neither.

Table 1. Frequencies of polarity-based groups among 4036 non-occurring quads in PDB

Polarity-based group	Frequency
0000	330
0001	355
0010	347
0011	261
0100	326
0101	293
0110	259
0111	170
1000	357
1001	224
1010	285
1011	154
1100	259
1101	144
1110	176
1111	96

3. Artificial polypeptides and possible applications in nanostructure design

Long helical structures and long sheets can be formed using the 5 patterns listed above, all of them can be viewed as artificial polypeptides with specific motifs that have not been observed in nature. When quads containing cysteine are excluded (because of the latter’s tendency to bond with another cysteine in the polypeptide and thereby imparting to the polypeptide a tendency towards a tertiary fold), the numbers are lower but remain significant. The other 11 groups are essentially in the ‘other‘ category, although some sequences from them could be sparsely interspersed among helices since the pitch value of 3.4 to 3.6 is not rigidly observed in naturally occurring helical structures. They could also be used to induce turns in direction to form tertiary level structures. It remains to be seen whether these ‘unnatural’ structures are stable and useful (for example, as tubular or substrate-like nanostructures [4, 5]), something that can be established in the lab and/or by computation and simulation. The increasing number of stable artificial polypeptide structures being synthesized in the laboratory [6, 7] suggests that such an expectation might not be far-fetched.

4. References

[1] Protein Database. http://www.rcsb.org/pdb/

[2] Intron Database. http://origin.bic.nus.edu.sg/xpro

[3] Kyte J and Doolittle RA. “Simple method for displaying the hydropathic character of a protein.” *J. Mol. Biol.* 157, 105-132 (1982).

[4] Seeman NC and Belcher AM. **“**Emulating biology: Building nanostructures from the bottom up.” *Proc. Natl. Acad. Sci. (USA)* 99, 6451-6455 (2002).

[5] Scheibel T, Parthasarathy R, Sawicki G, Lin, X-M, Jaeger H, and Lindquist, SL "Conducting nanowires built by controlled self-assembly of amyloid fibers and selective metal deposition." *Proc. Natl. Acad. Sci. (USA)* 100, 4527-4532 (2003).

[6] Wei Y and Hecht MH. “Enzyme-like Proteins From An Unselected Library of Designed Amino Acid Sequences. *Protein Engineering, Design and Selection* 17, 67-75 (2004).

[7] Forster AC, Tan Z, Nalam MNL, Lin H, Qu H, Cornish VW, and Blacklow SC. “Programming peptidomimetic syntheses by translating genetic codes designed de novo.” *Proc. Natl. Acad. Sci. (USA)* 100, 6353-6357 (2003).