

Simple Sequences are Rare in the Protein Data Bank

Melanie A. Huntley and G. Brian Golding*

Department of Biology, McMaster University, Hamilton, Ontario, Canada

ABSTRACT A simple sequence is abundant in the proteins that have been sequenced to date. But unusual protein features, such as a simple sequence, are not present in the same high frequency within structural databases. A subset of these simple sequences, a group with a highly repetitive nature has been shown to be abundant in eukaryotes but not in prokaryotes. In this study, an examination of the eukaryotic proteins in the Protein Data Bank (PDB) has revealed a large deficiency of low complexity, highly repetitive protein repeats. Through simulated databases of similar samples of eukaryotic proteins taken from the National Center for Biotechnology Information (NCBI) database, it is shown that the PDB contains a significantly less highly repetitive, simple sequence than artificial databases of similar composition randomly derived from NCBI. When the structural data for those few PDB sequences that did contain a highly repetitive simple sequence is examined in detail, it is found that in most cases the tertiary structure is unknown for the regions consisting of a simple sequence. This lack of a simple sequence both in the PDB database and in the structural information suggests that this type of simple sequence may produce disordered structures that make structural characterization difficult. *Proteins* 2002;48:134–140. © 2002 Wiley-Liss, Inc.

Key words: low complexity; databases; protein structures; amino acid repeats

INTRODUCTION

The sequence of any protein does not consist of randomly chosen amino acids. Beyond even the specific sequence necessary to encode its function, there are a large number of nonrandom features that protein sequences display.¹ One of these unusual features is the presence of an excess of simple sequences in proteins.² In a series of articles, Wootton and his colleagues showed that most proteins in the publicly available sequence databases had localized excesses of simple sequences.^{2,3} They developed a method, the SEG algorithm, based on information content to detect such low-complexity, simple-sequence regions. The sequence complexity, or more generally, information complexity, is a measure of how much information is actually encoded in a protein sequence. It has been shown that these low-complexity sequences are present in both eukaryotic and prokaryotic proteins.⁴

One subset of these low complexity sequences are perfect repeats of a single amino acid. Such perfect repeats have been shown to be present in large excess in eu-

karyotes but not in prokaryotes.⁵ These homopolymers are composed primarily of residues of the amino acids Q, N, S, T, P, H, G, A, D, and E. Their length is generally less than 20 residues long. Many of the proteins found to contain multiple, long homopeptides are essential developmental proteins in *Drosophila*, and have homeotic homologs in humans and mice. Humans tend to have fewer homopolymers than *Drosophila*, but the expansion of repeated glutamine residues are the cause of several human neurological disorders.^{6,7}

Another subset of low-complexity sequences are regions composed primarily of a single amino acid but not necessarily uninterrupted by other amino acids. These low-complexity, highly repetitive regions are quite lengthy, and can cover over 100 residues. They are also common in eukaryotic proteins but absent from prokaryotic proteins.⁸ Indeed, within eukaryotes, they form the most commonly shared peptide segments and are rapidly evolving.^{8,9}

It is important to determine the role of these low-complexity protein segments within the proteome of organisms. A large quantity of information about the functional role of proteins can be determined from its tertiary structure. Unfortunately, the Protein Data Bank (PDB) is not a random collection of protein sequences. The time and energy required to structurally characterize a protein imposes a limit on the number and the type of proteins for which structures can be solved. The first proteins characterized tend to be ones that have known functions, and are of great interest to the scientific community. This is probably what causes much of the difference between the PDB, and other public sequence databases. In addition, other factors that influence the composition of the PDB include the structural stability of crystallized proteins, solubility of proteins, availability of large, relatively pure quantity, and so on.

Wootton³ and Saqi¹⁰ analyzed the structures in the PDB for the presence of low-complexity sequences. They found that although there was a deficiency of low-complexity sequences in the structural database in comparison to the protein sequence database, it was nevertheless present within many protein structures. When a low-complexity

Grant sponsor: Natural Sciences and Engineering Research Council of Canada (to G.B.C.); Grant sponsor: NSERC summer scholarship (to M.A.H.).

*Correspondence to: G. Brian Golding, Department of Biology, McMaster University, Hamilton, Ontario, L8S 4K1, Canada. E-mail: Golding@McMaster.CA

Received 23 October 2001; Accepted 1 March 2002

sequence was present in the protein sequence, it formed well-ordered structures, and the low-complexity region was often a helical structure.

There is some dispute as to the types of structures formed by homopolymers. Polyalanine is often used in structural studies. Under aqueous conditions, short stretches of polyalanine will form helical structures.¹¹ The formation of specific secondary structures can, however, be nucleation dependent and with an appropriate sequence of residues (e.g., Ac-KA₁₄K-NH₂) they can form a stable β -pleated sheet structure.¹²

Polyglutamine is known to bind to other proteins, possibly interfering with their normal function, and this may be the cause of human pathologies when polyglutamine is abnormally expanded.¹³ In aggregation, peptides of polyglutamine form a β -sheet that are linked together by hydrogen bonds between their amide groups and may function as a polar zipper.^{14,15} On the other hand, it has been suggested by Sharma¹⁶ and Brahmachari¹⁷ that when interspersed with histidines (Q₈HQH₈, Q₁₀HQH₁₀, Q₈HQ₄HQ₈) a polyglutamine β -structure will also form under normal physiological aqueous conditions. Altschuler and colleagues note that this does not occur in other polyglutamine peptides without the interspersed histidines, and that instead, a random coil is formed for peptides with Q₉ or Q₁₇ repeats (surrounded by other amino acids).¹⁸

In contrast to artificial peptides, homopolymer repeats such as those studied by Karlin and Burge⁵ are present in naturally occurring peptides and are surrounded by many different amino acids. Similarly, the larger, highly repetitive, low-complexity regions, are usually interspersed with many other amino acids. How this would influence their potential abilities to form secondary or tertiary structures is not known.

In this study we compare the amount of low-complexity sequences found in the PDB to random databases created to match the PDB sequences for both taxonomic origin and sequence length. We confirm Saqi¹⁰ and Wootton's³ results that a simple sequence is underrepresented in the database despite the growth of the PDB database from 3091 to 15,213 structures characterized in the intervening years between 1994 and 2001. Hence, despite the much larger number of structures now known and the broader range of proteins that are now characterized, they are still a very biased sample. We also examine the PDB database for sequences with highly repetitive, low-complexity segments, and find these to be nearly absent.

METHODS

The PDB protein sequence database was downloaded from NCBI's ftp site (ftp://ftp.ncbi.nlm.nih.gov/pub/blast/db/) on May 30, 2000. Only eukaryotic proteins were selected from this file, because a highly repetitive, low-complexity sequence is found virtually exclusively within eukaryotes.⁸ In addition, due to their prokaryotic evolutionary origins, all mitochondrial and chloroplast proteins were discarded. There are 4458 protein sequences remaining within the database after these others are removed.

To promote comparisons between the sequences in NCBI and in PDB, the collected eukaryotic proteins were subdivided into 14 representative taxonomic groups: (1) human with 1623 sequences; (2) mouse with 489; (3) rat with 228; (4) mammalian (excluding human, mouse, or rat) with 597 entries; (5) aves with 154; (6) reptiles/amphibians with 96; (7) fish with 88; (8) *Drosophila* with 46; (9) arthropods (not including *Drosophila*) with 104; (10) *Arabidopsis* with 13; (11) plants (excluding *Arabidopsis*) with 306; (12) *Saccharomyces* with 250; (13) fungi (excluding *Saccharomyces*) with 197; and (14) others with 267 sequences. The number and length of proteins in each group were recorded.

To compare the PDB sequences with NCBI sequences, artificial databases were constructed from the NCBI sequences. These artificial databases were constructed by randomly sampling sequences from NCBI with the same taxonomic origins of the PDB sequences (within the groups listed above). This information was used to collect protein sequences for each category from NCBI to create new artificial databases, similar to the (remaining eukaryotic) PDB database. Three types of artificial databases were constructed, and 50 instances of a database for each type were generated (giving a total of 150 databases). The first type of database, which we denote as "Five Percent," consisted of full-length proteins randomly chosen from NCBI that were within 5% of the length of the PDB proteins. The second type of database, denoted "Ten Percent," was composed of full-length proteins within 10% of the actual PDB proteins lengths. Finally, the third type of database, denoted as "Fragment," consisted of peptide fragments that were identical in length to the PDB proteins. These fragments were constructed by choosing a larger NCBI sequence from the correct taxonomic group and then randomly picking a peptide fragment of the exact length from within this larger sequence. This was done to avoid fragments biased toward the amino- or carboxy-terminus. The first database is quite stringent in the proteins that can be chosen, and hence, has a small amount of overlap in samples within each of the 50 simulated databases. The second permits a greater choice among proteins. The third differs mainly by lacking sequences that are characteristic of amino- and carboxy-termini but adheres strictly to the observed lengths.

Within the 50 replicates of the "Five Percent" databases, there is a 5% overlap of sequences. For the "Ten Percent" databases there is a 4% overlap, and for the "Fragment" databases the overlap is 2%. The "Five Percent" and "Ten Percent" databases have less than 4% of their sequences in common. The "Fragment" databases share less than 2% of their sequences with either of the "Five Percent" or "Ten Percent" databases.

Each database was then searched, using BLAST¹⁹ with the SEG filter turned off. Homopolymers of each of the 20 amino acids were used as a query for these searches. These queries find not only homopolymers (indeed, few of these are ever found), but also sequences with highly repetitive low-complexity regions. Query sequences of length 50 and 100 residues were used in turn. When homopolymers of length 50 were used as queries, hits with BLAST expect

TABLE I. Examples of Low-Complexity Sequence

Category	Protein	Residues	Complexity value	Sequence
A <i>Marginal low complexity</i>	PDB 1B80 Chain A	316–350	2.33	TIKDVEQACAETPFPTLTLLPGPETSVQRI PPPPG
B <i>Highly repetitive low complexity</i>	PDB 1QLX Chain A	10–94	1.92	GWNTGGSRYPGQSPGGNRYPPQGGGGWGQPHGGG WCQPHGGGGWQPHGGGGWQPHGGGGWQGGTTHSQW NKPSPKPTNMKHMAG
C <i>Highly repetitive low complexity</i>	NP_015309	317–400	1.82	NSTSNANTVFSERAAMFAALQQKQQRFQALQQQQ QQQQNQQQNQPPQQQQQQQNPKELQSQRQQQQR SILQSLNPALQEKI

The shaded text corresponds to the low-complexity region detected by SEG. A structure has been resolved for the low-complexity region of PDB protein 1B80 (chain A),²³ while a structure has not been resolved for the low-complexity region of PDB protein 1QLX (Chain A).²⁴ NP_015309²⁵ is from the protein database of NCBI, and not found within the PDB.

values of 0.05 or less had a range of 30 to 80% identity to the query sequence (a mean of 51.5 ± 13.6). For query sequences of length 100, the percent identity for the significant BLAST hits was 21 to 80 (a mean of 45.1 ± 16.3). For comparison, the percent identity between human myoglobin and human β -hemoglobin is 23.4%. For each database and for each amino acid used in the query sequence, the number of significant BLAST hits was recorded. This number is the number of sequences within the PDB or simulated database that contain a segment with significant similarity to the query.

We also examined the structural data for all those PDB sequences that showed simple sequence (as determined by SEG). A window length of 15 and K2(1) of 1.9 were used as parameters for the program SEG as opposed to the default values (12 and 2.2, respectively). These values were chosen because trial and error had demonstrated that they would preferentially detect the longer and more repetitive repeats such as those observed in yeast and other eukaryotes.⁸ The parameter K2(1) is an initial cutoff complexity value such that when SEG initially calculates the complexity of a subsequence, it must not exceed the cutoff complexity value. Table I gives an example of sequences detected as low complexity using the default parameters for SEG, and using the more stringent parameters. The *marginal low-complexity* sequence in Table 1A has a complexity value of 2.33, and would not be detected using the more stringent parameters. When SEG detected a low-complexity sequence in a PDB protein sequence, the corresponding PDB structural file was examined to determine the nature of the predicted structure in this region.

An identical search of each of the simulated databases was conducted using the same parameters in SEG. Again, the number of protein sequences containing a low-complexity sequence was recorded for each of the databases.

RESULTS

BLAST Similarity

To determine the degree to which proteins with highly repetitive segments are under- (or over-) represented we compared the PDB database to artificially constructed databases with the same number of proteins, with proteins of similar length and with similar taxonomic origin (as

TABLE II. Number of Significant BLAST Hits for a Homopolymer Query Sequence of Each Amino Acid (of Length 50) in the PDB Database (Containing 4458 Sequences), and the Percent of Simulated Databases with as Few or Fewer Hits for Each Type of Generated Database

Amino acid	PDB	Fragment	Five Percent	Ten Percent
A	3	0	4	2
C	5	100	4	0
D	3	0	0	0
E	1	0	0	0
F	0	62	40	38
G	2	4	6	2
H	2	0	0	0
I	0	98	88	90
K	3	0	0	0
L	0	94	94	90
M	0	94	74	78
N	0	0	0	0
P	2	0	4	4
Q	0	0	0	0
R	1	0	0	0
S	0	0	0	0
T	0	0	0	0
V	0	100	66	82
W	0	100	98	98
Y	0	70	56	56

For each type of database (Fragment, Five-Percent, and Ten Percent), BLAST searches were done for each amino acid using homopolymers of length 50.

described above). Lengths of proteins were matched to within a 10, 5, or 0% error of the lengths in PDB. To achieve 0%, fragments from full-length proteins were used.

All 150 artificial databases were searched using BLAST for proteins with significant similarity to homopolymers of length 50 (Table II) or 100 residues (Table III). The data in these tables highlight the unusual composition of PDB proteins compared to those in NCBI. For example, the PDB database has only 3 BLAST hits significantly similar to D_{50} . In contrast, none of the 150 artificial databases had as few as (or fewer than) 3. All constructed databases have more than 3 proteins that are similar to D_{50} . Although this is true for most amino acids it is not the case with all. The

TABLE III. Number of Significant BLAST Hits for a Homopolymer Query Sequence of Each Amino Acid (of Length 100) in the PDB Database (Containing 4458 Sequences), and the Percent of Simulated Databases (out of 50) with as Few or Fewer Hits for Each Type of Generated Database

Amino acid	PDB	Fragment	Five Percent	Ten Percent
A	5	0	2	0
C	3	86	0	0
D	3	0	0	0
E	4	2	0	0
F	0	52	28	38
G	2	10	8	2
H	1	0	0	0
I	0	82	58	66
K	5	0	0	0
L	0	74	68	22
M	0	88	58	64
N	0	0	0	0
P	2	0	0	0
Q	0	0	0	0
R	1	0	0	0
S	0	0	0	0
T	0	0	0	0
V	0	98	60	82
W	0	100	96	98
Y	0	82	70	68

For each type of database (Fragment, Five Percent, and Ten Percent), BLAST searches were done for each amino acid using homopolymers of length 100.

PDB database has no proteins that match W_{50} or I_{50} . Similarly sequences pulled from NCBI virtually never or seldom have significant similarity to these homopolymers, respectively. In this respect, the two databases are similar.

The three types of databases (“Fragment,” “Five Percent,” and “Ten Percent”) show consistent results for each amino acid examined except cysteine. The “Fragment” databases had fewer proteins that were similar to polycysteine. Table II shows that all 50 of the “Fragment” databases contained five or fewer sequences with significant similarity to polycysteine. Out of the “Five Percent” databases, only 2 of 50 databases contained five or fewer such sequences, and all 50 of the “Ten Percent” databases contained more than five sequences that were significantly similar to polycysteine. This difference between the “Fragment” databases and other two types is explained by the occurrence of cysteine-rich metallothionein proteins. These proteins appeared 6 to 11 times more often in the “Five” and “Ten Percent” databases than in the “Fragment” databases. This bias was due to the method in which the databases were constructed. Metallothioneins are typically only 61 amino acids long, so when a sequence of about 61 amino acids was needed, the “Five” and “Ten Percent” databases were required to find proteins within 5 and 10% of the PDB protein sequence length, and hence, they were more likely to choose metallothionein sequences than any others. The “Fragment” databases were not restricted to choosing from proteins that were of any particular length;

Percent of low complexity sequences using SEG parameters $L = 15$ and $K2(1) = 1.9$

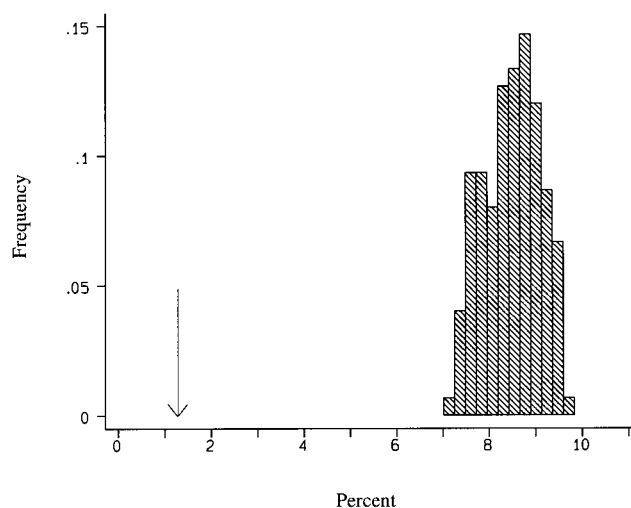


Fig. 1. This histogram shows the distribution for the percentage of protein sequences containing simple sequence repeats in each of the 150 generated databases, the mean of which is 8.5%. The arrow indicates the percentage of protein sequences in the PDB database that contained simple sequence (1.3%). SEG parameters of $L = 15$ (window length) and $K2(1) = 1.9$ were empirically chosen to detect the highly repetitive simple sequence that is observed in NCBI protein sequences.

they could simply choose a fragment from a larger protein. As a result, the “Five” and “Ten Percent” databases, have a similar number of proteins that match polycysteine while the more random, “Fragment” databases do not.

The results are qualitatively similar whether queries of 50 or 100 residues are used (Table II vs. Table III). In both cases the PDB lacks proteins that are similar to homopolymers despite their common presence in NCBI.

Information Content

To measure the extent to which the PDB lacks these sequences, we analyzed each database using the program SEG. This program will detect low-complexity sequences on the basis of information content, independent of its repetitive nature.² Of the 4458 eukaryotic entries in the PDB, 16 were structures for which the sequence was unknown. This left 4442 PDB entries containing both structural and sequence information. A search of all such proteins in the PDB detected only 59 (out of 4442; 1.3%) that contained a highly repetitive, low-complexity region. In contrast, the smallest percentage found in any of the 150 artificial databases was 7.0% (314 out of 4458). Figure 1 indicates that the number of these proteins in PDB are well beyond the extremes of the distribution described by the artificial databases.

These results were obtained with SEG parameters chosen empirically to improve detection of the highly repetitive segments observed in genome-wide surveys of protein sequences. The use of the suggested default SEG parameters results in a far less stringent search and detects many more proteins with a low-complexity region. Nevertheless, the PDB is also impoverished for these proteins (Fig. 2).

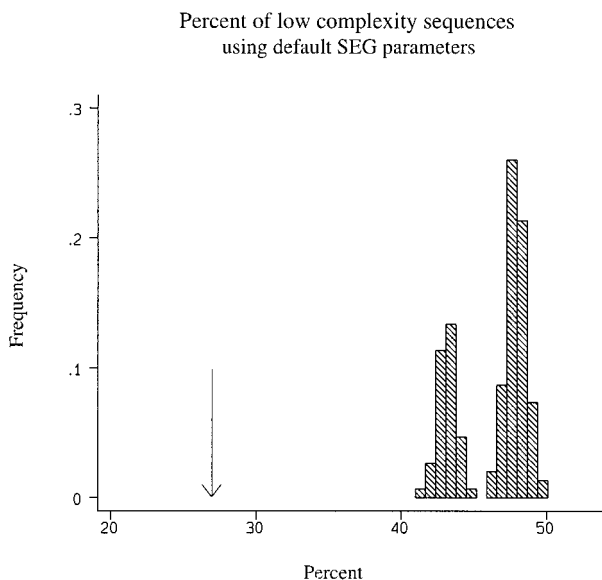


Fig. 2. This histogram shows the distribution for the percentage of protein sequences containing simple sequence repeats in each of the 150 generated databases, the mean of which is 46.4%. The arrow indicates the percentage of protein sequences in the PDB database that contained simple sequence (27.1%). The default SEG parameters of $L = 12$ (window length) and $K2(1) = 2.2$ were used to verify that our previous choice of parameters was not skewing the results. The 50 "Fragment" databases have a distribution that is shifted more to the left than the other two types of databases. This slight shift is the cause of the bimodality in this histogram.

Structure Content

The PDB entries for the 59 proteins identified by SEG were collected and the solved structures were examined. Most of these proteins (39 out of 59; 66%), while they contain a low-complexity region within their protein sequence, do not include this region in the structural information. Hence, these regions were either uncharacterized, post-translationally modified, or disordered in the final structural model. Another 5 of the 59 (8%) had the low-complexity region only partially described in the structure. The remaining 15 of the 59 (25%) were the only PDB entries, which included the low complexity region in its entirety in the solved structure for these proteins.

Four examples of protein structures and their corresponding sequences are shown in Figure 3, with the low-complexity region highlighted in yellow. An example of a protein where all of the low-complexity regions are included in the structure is given in Figure 3(A). A protein where part of the low-complexity region is present is given in Figure 3(D). The majority of entries, even for those with a low-complexity region, do not include a characterized structure for this region of the protein [Fig. 3(B) and (C)]. This means that out of 4442 possible eukaryotic PDB protein sequences, only 15 were found to have simple sequences for which complete structural data existed. This comprises just 0.3% of the total database.

DISCUSSION

It is clear that the PDB is not a random sample of protein structures. However, as the database grows, it is

hoped that all structurally distinct folds will eventually be found, and that all of the common folds have been (or will shortly be) determined. Certainly, for very common sequence features of proteins, their structural properties should be contained within the rapidly expanding databases.

It is known that perfect repeat sequences are common in eukaryotic proteins.⁵ It is also known that highly repetitive sequences with low-complexity regions are common⁹ in eukaryotes. But the structures that these sequence features might form are not well understood. There is conflicting evidence that suggests that artificial homopeptides of glutamine and alanine do form stable helical structures.^{11,15} When certain other amino acids are introduced to polyglutamine and to polyalanine, making them more similar in composition to sequences found in real proteins, these highly repetitive, low-complexity tracts may instead form stable beta sheets.^{12,17,18}

Highly repetitive, low-complexity sequences are not common in the protein sequences of the PDB. Despite their common occurrence in the protein sequence databases, they are very rare in structural databases. Wootton³ and Saqi¹⁰ carried out a survey of the PDB database for low-complexity sequences and found that such sequences were underrepresented. Saqi¹⁰ described the low-complexity sequences that he found as of "*marginal low complexity*." He concluded that most low-complexity regions are structurally well characterized, and no more or less disordered than the rest of the protein. When present, he found that these low-complexity sequences often formed helical structures.

In the intervening years, the number of protein structures known has grown by fivefold. At the same time, the protein sequence data has also grown and now includes the complete proteome from several species. Our survey of these greatly enlarged databases suggests that they still contain quite different collections of sequences. As Saqi¹⁰ found, the PDB database still contains far fewer low-complexity sequences than do the sequence databases (Fig. 2). We have also found that PDB contains far fewer proteins with highly repetitive low-complexity regions (Fig. 1) despite their prevalence in eukaryotic protein sequences.

Although Wootton³ and Saqi's¹⁰ results, suggest that marginal low-complexity sequences are present and structurally well characterized, we found that when highly repetitive low-complexity regions are present, they do not generally form part of the structural information. In our survey of 4458 eukaryotic protein sequences from the PDB, we found only 59 such sequences. Of these 59, only 15 (25%) were structurally characterized. Of these 15, only 8 formed helical structures and the remaining 7 formed irregular loops.

The differences in sequence composition and structure between marginal low complexity and highly repetitive, low complexity may help to explain why highly repetitive low-complexity tracts, especially polyglutamine, are associated with deleterious conditions, such as Huntington disease. Such highly repetitive amino acid sequences are

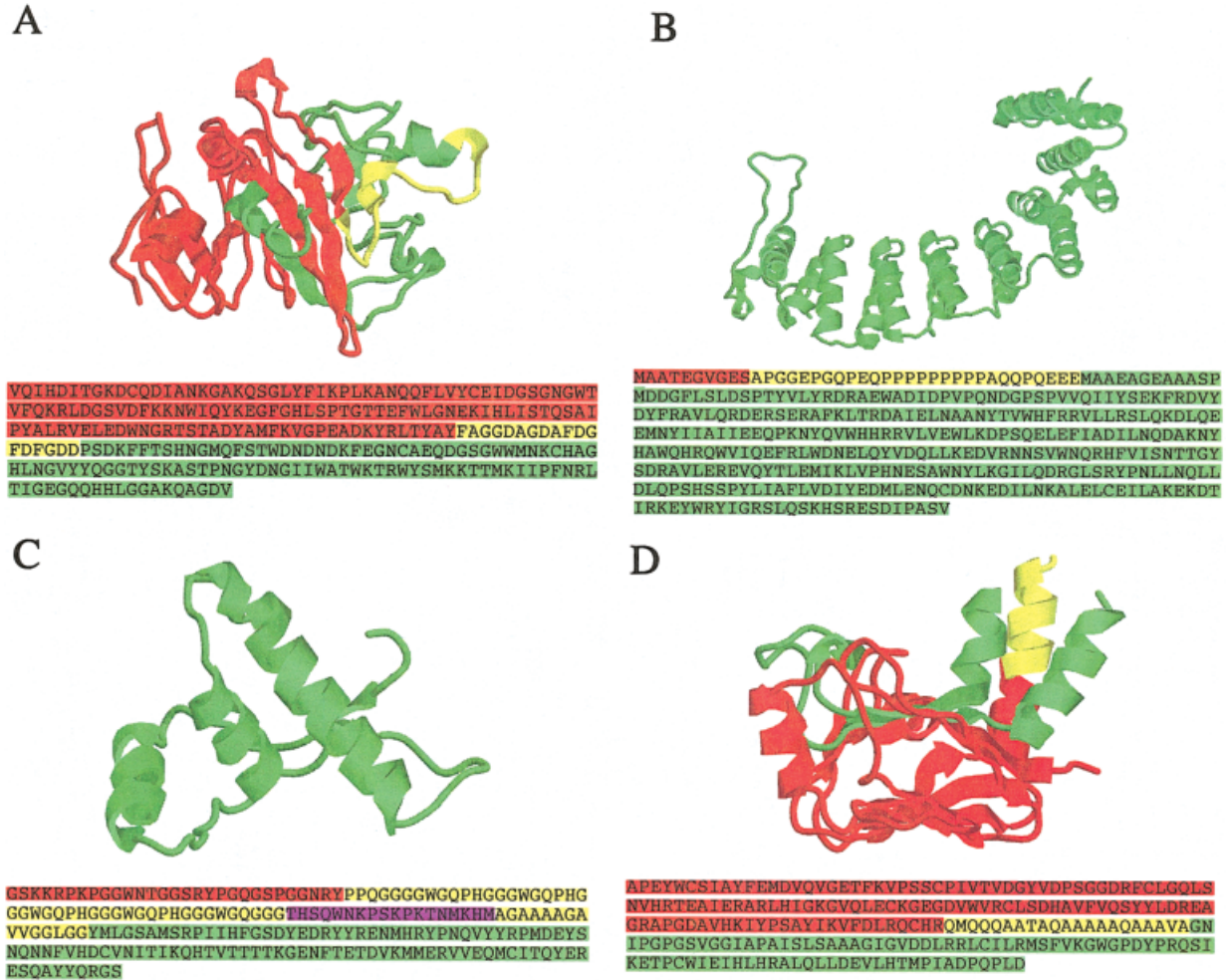


Fig. 3. Structure of four PDB proteins with the amino acid sequence shown below. Yellow indicates low complexity (simple sequence repeats) as determined by SEC^2 using a window length of 15 and $K2(1)$ of 1.9; red indicates the N-terminus, green indicates the C-terminus, and magenta is simply to separate repeat regions if there are more than one. (A) PDB protein 1FID.²⁶ The entire repeat region was included in the structural data. This was the case for 25% (15/59) of the PDB proteins containing simple sequence repeats. (B) PDB protein 1FT1.²⁷ Neither the N-terminus nor the repetitive region were present in the structural data. This complete absence of the protein repeat in the structural data was the case for 66% (39/59) of the PDB proteins containing simple sequence repeats. (C) PDB protein 1QLX.²⁴ As in (B), neither N-terminus nor repetitive regions were found to have structural data. (D) PDB protein 1YGS.²⁸ The simple sequence repeat was only partially present in the structural data. In this case, the repeat (yellow) was supposed to be joined with the C-terminal portion of the protein (green), but it does not because the structural data is missing. Having structural data only partially present for simple sequence repeats occurs 8% (5/59) of the time.

thought to be generated by slipped-strand mispairing at the DNA level. This would generate triplet repeats in the nucleotide sequence, and these have been observed.^{20,21} Starting from a small, marginal low-complexity island that has a defined crystal structure, one can expand the low-complexity region via slipped-strand mispairing at the DNA level until a longer, much more repetitive tract is generated. Once this happens, the sequence becomes predisposed to further tandem duplications by unequal crossing over.²² Such highly repetitive regions have higher rates of evolution than the rest of the protein within which they are embedded.⁸ At this point, the low-complexity region may be so long that the protein can no longer function normally, leading to a clinically recognizable disorder.

New evidence suggests that some repeats may also be caused by positive selection because they do not show any evidence of slippage-like processes at the DNA level. In *Drosophila melanogaster*, polyglutamine repeats were found that showed no evidence of long codon reiterations.²¹ Such repeats would be functionally important, and most likely required for correct tertiary structure of the protein as well.

CONCLUSION

Whatever the true nature of highly repetitive, low-complexity protein sequences, they are common within protein sequences, some of which cause human diseases, but do not seem to form stable structures. Although some proteins with these regions may form structures in labora-

tory experiments, they are not present within the entirety of the PDB entries. We conclude that highly repetitive low-complexity regions most likely form disordered structures in natural proteins. The purpose of these regions in eukaryotic proteins remains a mystery.

ACKNOWLEDGMENTS

This work was supported by a Natural Sciences and Engineering Research Council of Canada grant to GBG and an NSERC summer scholarship to M.A.H.

REFERENCES

- Karlin S, Ghandour G, Ost F, Tavare S, Korn LJ. New approaches for computer analysis of nucleic acid sequences. *Proc Natl Acad Sci USA* 1983;80:5660–5664.
- Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 1993;17:149–163.
- Wootton J. Sequences with “unusual” amino acid compositions. *Curr Opin Struct Biol* 1994;4:413–421.
- Marcotte E, Pellegrini M, Yeates T, Eisenberg D. A census of protein repeats. *J Mol Biol* 1998;293:151–160.
- Karlin S, Burge C. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc Natl Acad Sci USA* 1996;93:1560–1565.
- Martin JB. Molecular genetics of neurological diseases. *Science* 1993;262:674–676.
- Ross CA, McInnis MG, Margolis RL, Li SH. Genes with triplet repeats: candidate mediators of neuropsychiatric disorders. *Trends Neurosci* 1993;16:254–260.
- Huntley M, Golding GB. Evolution of simple sequence in proteins. *J Mol Evol* 2000;51:131–140.
- Golding GB. Simple sequence is abundant in eukaryotic proteins. *Protein Sci* 1999;8:1358–1361.
- Saqi M. An analysis of structural instances of low complexity sequence segments. *Protein Eng* 1995;8:1069–1073.
- Rohl CA, Fiori W, Baldwin RL. Alanine is helix-stabilizing in both template-nucleated and standard peptide helices. *Proc Natl Acad Sci USA* 1999;96:3682–3687.
- Blondelle SE, Forood B, Houghten RA, Perez-Paya E. Polyalanine-based peptides as models for self-associated beta-pleated-sheet complexes. *Biochemistry* 1997;36:8393–8400.
- Shimohata T, Onodera O, Tsuji S. Interaction of expanded polyglutamine stretches with nuclear transcription factors leads to aberrant transcriptional regulation in polyglutamine diseases. *Neuropathology* 2000;20:326–333.
- Krull L, Wall J, Zobel H, Dimler R. Synthetic polypeptides containing side-chain amide groups: water-insoluble polymers. *Biochemistry* 1965;4:626–632.
- Perutz MF, Johnson T, Suzuki M, Finch JT. Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc Natl Acad Sci USA* 1994;91:5355–5358.
- Sharma D, Sharma S, Pasha S, Brahmachari SK. Peptide models for inherited neurodegenerative disorders: conformation and aggregation properties of long polyglutamine peptides with and without interruptions. *FEBS Lett* 1999;456:181–185.
- Brahmachari SK, Sharma D, Sharma S, Pasha S, Sen S, Saleem Q. Probing the polyglutamine puzzle in neurological disorders. *FEBS Lett* 2000;472:167–168.
- Altschuler EL, Hud NV, Mazrimas JA, Rupp B. Structure of polyglutamine. *FEBS Lett* 2000;472:166–168.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
- Green H, Wang N. Codon reiteration and the evolution of proteins. *Proc Natl Acad Sci USA* 1994;91:4298–4302.
- Alba MM, Santibanez-Koref MF, Hancock JM. The comparative genomics of polyglutamine repeats: extreme differences in the codon organization of repeat-encoding regions between mammals and *Drosophila*. *J Mol Evol* 2001;52:249–259.
- Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 1987;4:203–221.
- Blodig W, Doyle WA, Smith AT, Winterhalter K, Choinowski T, Piontek K. Autocatalytic formation of a hydroxy group at C beta of trp171 in lignin peroxidase. *Biochemistry* 1998;37:8832–8838.
- Zahn R, Liu A, Luhrs T, Riek R, von Schroetter C, Lopez Garcia F, Billeter M, Calzolari L, Wider G, Wuthrich K. NMR solution structure of the human prion protein. *Proc Natl Acad Sci USA* 2000;97:145–150.
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. Life with 6000 genes. *Science* 1996;274:546–547.
- Yee VC, Pratt KP, Cote HC, Trong IL, Chung DW, Davie EW, Stenkamp RE, Teller DC. Crystal structure of a 30 kDa C-terminal fragment from the gamma chain of human fibrinogen. *Structure* 1997;5:125–138.
- Park HW, Boduluri SR, Moomaw JF, Casey PJ, Beese LS. Crystal structure of protein farnesyltransferase at 2.25 angstrom resolution. *Science* 1997;275:1800–1804.
- Shi Y, Hata A, Lo R, Massague J, Pavletich N. A structural basis for mutational inactivation of the tumour suppressor Smad4. *Nature* 1997;388:87–93.