Hypothesis

Structural features of single amino acid repeats in proteins

Juan A. Subirana[a,*], Jaume Palau[b]

[a] *Department d'Enginyeria Química, ETSEIB, Universitat Politècnica de Catalunya, Av. Diagonal 647, E-08028 Barcelona, Spain*
[b] *Unitat de Biotecnologia Computacional, Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, E-43005 Tarragona, Spain*

Received 25 February 1999

Abstract Single amino acid repeats are found in different kinds of proteins. Some of these repeats are pathogenic. It is striking that some amino acids are able to form such repeats, but other amino acids are not. We suggest an explanation for this fact based on the different tendency of each amino acid to form aggregates. Aggregation may be due to the formation of incipient lamellar crystals as they have been described in poly-α-amino acids and in most synthetic polymers.

Key words: Amino acid sequence; Protein structure; Nucleotide sequence; Glutamine; Neurological disease; Protein aggregation

1. Introduction

In a recent review one of us pointed out [1] the strong relationship between polymer crystallization and protein folding/aggregation. Synthetic polymers, including poly(amino acids), fold as lamellar crystals, the thickness of which is similar to the size of globular proteins. An example is shown in Fig. 1. In this paper we suggest that the fibrillar aggregation of single amino acid repeats is basically related to the process of polymer crystallization in lamellar crystals. We also analyze why some amino acid single repeats are found more frequently in nature.

2. The role of amino acid repeats in proteins

Many proteins have been described which contain long runs of single amino acid repeats. Runs of different amino acids may be found in the same protein [2]. Long regions with unusual compositions are also common [3]. It has been suggested that some of these repeats, in particular in transcription factors, may have a positive role in evolution [4] and be involved in species-specific regulatory factor interactions [5,6]. The case of the TFIID factor [5] is most striking, since it contains a $(Gln)_{34}$ repeat in the human protein, which is absent in related proteins in other species.

In contrast with these reports, it is clear that $(Gln)_n$ and other repeats are involved in some neurological diseases [7,8]. In many cases the repeats give rise to aggregates which build abnormal depositions in the cell. The accepted view is that these depositions are pathogenic, although this view has been recently challenged [9]: toxic behavior has been observed in the absence of aggregates.

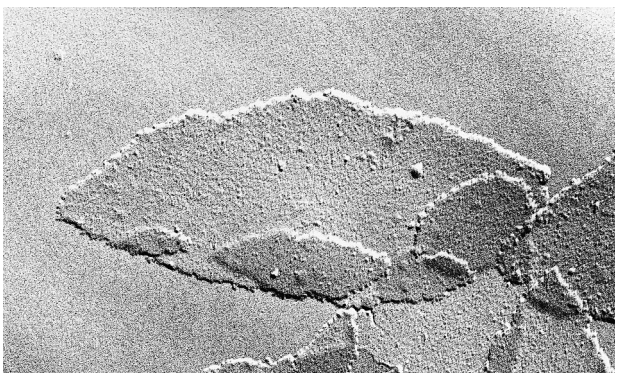

Fig. 1. Example of lamellar crystals obtained from synthetic polymers. The figure shows an electron micrograph of a polyamide which contains glycine [30]. The crystals have a thickness of about 60 Å. Polymer chains fold regularly back and forth following the short dimension of the crystals, with a bend present approximately every 60 Å. Such a folding process is similar to that found in proteins, as described in the text.

At the DNA level, trinucleotide repeats may occur both in expressed and in silent regions of the genome. When expressed they will give rise to single amino acid repeats. Some of the trinucleotide repeats may undergo expansion, probably due to alternative DNA structures [10]. Above a certain threshold [11] some of these expansions are associated with diseases.

3. A bias in amino acid composition

A paradox of triple repeat expansion in coding sequences is that usually the same amino acid is read. For example, the repeat sequence $(CAG)_n$·$(CTG)_n$ could be translated as $(Gln)_n$, $(Ser)_n$, $(Ala)_n$, $(Cys)_n$ and $(Leu)_n$, depending on the strand read and its reading frame. However, in practice $(Gln)_n$ is by far the most common repeat found, as shown in Table 1, whereas $(Leu)_n$ is seldom found and $(Cys)_n$ has never been found. From a statistical point of view one would expect $(Ala)_n$ to be the most common repeat, since it could be coded by both $(GCA)_n$ and $(GCT)_n$. However, long $(Ala)_n$ repeats are not frequent, as is apparent from Table 1.

Some single amino acid repeats will not be found in nature because trinucleotide expansion may not be possible due to an anomalous DNA structure [10]. Problems at the level of RNA structure, availability of tRNA, etc. may also prevent some DNA sequences from being transcribed in the cell. However, the results shown in Table 1 cannot be explained at this level

*Corresponding author. Fax: (34) (3) 401 6688.
E-mail: subirana@eq.upc.es

PII: S0014-5793(99)00310-5

Table 1
Single amino acid repeats found in proteins

Amino acid	Number of proteins	
	$n \geq 30$	$n \geq 20$
Ala	0	2
Asp	1	7
Glu	1	8
Gly	0	7
Asn	13	33
Gln	14	60 (approx.)
Ser	4	20
Thr	1	3
Pro	1	8

Protein sequences and ORFs containing homorepeats were taken from the SwissProt protein sequence data bank (release 34) and its supplement, TrEMBL (release 51) [28] in the version of DNASTAR, Inc. for McIntosh personal computers. Redundant scores have been eliminated. In the case of Gln for $n = 20$ some of the eliminations were doubtful, the number given is only approximate. A related table, using a smaller database, has appeared elsewhere [29]. The distribution of single amino acid repeats is essentially the same.

only. Why is Gln, followed by Asn and Ser, the most common repeat found? Leu, Ile, Ala, Val, etc. are much more frequent in proteins than Gln, but long repeats are only seldom (Ala, Leu) or never found (Ile, Val). A striking feature of the data given in Table 1 is that most $(Asn)_n$ repeats are found in proteins from *Dictyostelium discoideum*, a fact for which we have no explanation. The other repeats are found in proteins from many different unrelated species.

4. Amino acid repeats and polymer crystallization

It appears to us that a likely explanation for these observations resides in the structure to be expected for single amino acid repeats. It has been established [12–16] that poly-α-amino acids fold into thin lamellae, which have a thickness between 40 and 180 Å depending on the amino acid considered. The amino acid chain can be either in the α, β or other conformation depending on the amino acid and the conditions of crystallization.

Our interpretation of Table 1 is that many single amino acid repeats will tend to crystallize as incipient lamellar structures and will thus give rise to insoluble aggregates which will make cells unviable. The formation of such aggregates has been established for $(Gln)_n$ repeats [17], which might have a β-helical structure related to that suggested for amyloid fibrils [18,19] and prion particles [20,21]. In the latter case the β-helical structure is not generated by single amino acid repeats. The structure of these aggregates is reminiscent of that described in some insect silks [22]. The related question of protein misassembly has been analyzed from different points of view in a recent volume of Advances in Protein Chemistry [23].

5. Aggregation is related to amino acid properties

Unfortunately no detailed structural data are available for the 3D organization of any protein containing long single amino acid repeats. The longest repeat available in the Protein Data Base (PDB) is $(Ala)_7$, found in a small α-helical antifreeze protein [24]. In any case it is apparent that only those amino acid sequences which aggregate with difficulty will give viable proteins. From the data reproduced in Table 1 it can be concluded that sequences of Gln, Asn and Ser with a moderate length fall into this category and can thus be found in functional proteins. The amphipathic nature of the side chains in these amino acids might explain their comparatively greater resistance to aggregation. In fact the insertion of different lengths of Gln, up to $(Gln)_{10}$, has minimal effects on the stability of some proteins [25]. In conclusion, it is likely that long repeats of single amino acids are avoided in proteins since they will have a tendency to either aggregate (hydrophobic amino acids) or have strong electrostatic interactions (charged amino acids). The latter will tend to precipitate/aggregate proteins with an opposite charge. The length of tolerated repeats will depend on the properties of each amino acid side chain.

6. Conclusion

Finally we should point out that simulation methods developed for protein folding [26] have been shown to be adequate to predict polymer crystallization [27]. Both processes (protein folding/aggregation and homopolymer crystallization) should be considered to rest on the same conceptual ground. This realization should result in a cross-fertilization of both fields of endeavor.

References

[1] Subirana, J.A. (1997) Trends Polymer Sci. 5, 321–326.
[2] Karlin, S. and Burge, C. (1996) Proc. Natl. Acad. Sci. USA 93, 1560–1565.
[3] Wootton, J.C. (1994) Curr. Opin. Struct. Biol. 4, 413–421.
[4] Gerber, H.-P., Seipel, K., Georgiev, O., Höfferer, M., Hug, M., Rusconi, S. and Schaffner, W. (1994) Science 263, 808–811.
[5] Hoffmann, A., Sinn, E., Yamamoto, T., Wang, J., Roy, A., Horikoshi, M. and Roeder, R.G. (1990) Nature 346, 387–390.
[6] Cox, G.W., Taylor, L.S., Willis, J.D., Melillo, G., White III, R.L., Anderson, S.K. and Lin, J.-J. (1996) J. Biol. Chem. 271, 25515–25523.
[7] Djian, P. (1998) Cell 94, 155–160.
[8] Wells, R.D., Warren, S.T. and Sarmiento, M. (Eds.) (1998) Genetic Instabilities and Hereditary Neurological Diseases, Academic Press, New York.
[9] Sisodia, S.S. (1998) Cell 95, 1–4.
[10] Pearson, C.E. and Sinden, R.R. (1998) Curr. Opin. Struct. Biol. 8, 321–330.
[11] Perutz, M.F. (1996) Curr. Opin. Struct. Biol. 6, 848–858.
[12] Padden, F.J. and Keith, H.D. (1965) J. Appl. Phys. 36, 2987–2995.
[13] Carr, S.H., Walton, A.G. and Baer, E. (1968) Biopolymers 6, 469–477.
[14] Keith, H.D., Giannoni, G. and Padden, F.J. (1969) Biopolymers 7, 775–792.
[15] Padden, F.J., Keith, H.D. and Giannoni, G. (1969) Biopolymers 7, 793–804.
[16] Muñoz-Guerra, S., Puiggalí, J., Rodríguez-Galán, A. and Subirana, J.A. (1983) J. Mol. Biol. 167, 223–225.
[17] Perutz, M.F., Johnson, T., Suzuki, M. and Finch, J.T. (1994) Proc. Natl. Acad. Sci. USA 91, 5355–5358.
[18] Sunde, B. and Blake, C. (1997) in: Advances in Protein Chemistry (Richards, F.M., Eisenberg, D.S., Kim, P.S. and Wetzel, R., Eds.), Vol. 50, pp. 123–159, Academic Press, New York.
[19] Lazo, N.D. and Downing, D.T. (1998) Biochemistry 37, 1731–1735.
[20] Kelly, J.W. (1998) Proc. Natl. Acad. Sci. USA 95, 930–932.
[21] Fink, A.L. (1998) Curr. Biol. 3, R9–R23.
[22] Geddes, A.J., Parker, K.D., Atkins, E.D.T. and Beighton, E. (1968) J. Mol. Biol. 32, 343–358.

[23] Richards, F.M., Eisenberg, D.S., Kim, P.S. and Wetzel, R. (Eds.), Advances in Protein Chemistry, Vol. 50, Academic Press, New York.
[24] Yang, D.S.C., Sax, M., Chakrabartty, A. and Hew, C.L. (1988) Nature 333, 232–237.
[25] Ladurner, A.G. and Fersht, A.R. (1997) J. Mol. Biol. 273, 330–337.
[26] Toma, L. and Toma, S. (1996) Protein Sci. 5, 147–153.
[27] Toma, L., Toma, S. and Subirana, J.A. (1998) Macromolecules 31, 2328–2334.
[28] Bairoch, A. and Apweiler, R. (1997) Nucleic Acids Res. 25, 31–36.
[29] Green, H. and Djian, P. (1998) in: Genetic Instabilities and Hereditary Neurological Diseases (Wells, R.D., Warren, S.T. and Sarmiento, M., Eds.), pp. 739–759, Academic Press, New York.
[30] Franco, L., Subirana, J.A. and Puiggalí, J. (1999) Polymer 40, 2429–2438.