Hypothesis

Are non-functional, unfolded proteins ('junk proteins') common in the genome?

Simon C. Lovell*

School of Biological Science, University of Manchester, 2.205 Stopford Building, Oxford Rd, Manchester M13 9PT, UK

Received 19 August 2003; revised 25 September 2003; accepted 27 September 2003

First published online 27 October 2003

Edited by Felix Wieland

Abstract It has recently been shown that many proteins are unfolded in their functional state. In addition, a large number of stretches of protein sequences are predicted to be unfolded. It has been argued that the high frequency of occurrence of these predicted unfolded sequences indicates that the majority of these sequences must also be functional. These sequences tend to be of low complexity. It is well established that certain types of low-complexity sequences are genetically unstable, and are prone to expand in the genome. It is possible, therefore, that in addition to these well-characterised functional unfolded proteins, there are a large number of unfolded proteins that are non-functional. Analogous to 'junk DNA' these protein sequences may arise due to physical characteristics of DNA. Their high frequency may reflect, therefore, the high probability of expansion in the genome. Such 'junk proteins' would not be advantageous, and may be mildly deleterious to the cell.

Key words: Low-complexity sequences; Natively unfolded proteins; Junk proteins; Evolution

1. Background

There has recently been a realisation that many proteins are unfolded in their functional state. These proteins have a wide range of functions, including transcription, signalling, translation and RNA binding. The degree of folding can vary from completely unfolded to proteins being active in the molten globule state, and through to the conventionally studied proteins which approximate having a unique, active, native state. Comprehensive reviews are available [1–3]. The surprising discovery of these so-called 'natively unfolded' proteins has challenged conventional views on the relationship between protein structure and function [1].

Natively unfolded proteins would appear to be very common. It has been estimated that fully 25% of residues in proteins in a range of different genomes are in regions unlikely to fold into globular structures [4]. This proportion will represent some completely unfolded proteins, and some folded proteins with unfolded insertions. Dunker and co-workers predict that up to 15 000 proteins in the SwissProt database are either predicted as being disordered over their entire length, or are predicted as having large (longer than 40 amino acids) disordered regions within an otherwise ordered protein [5].

Several authors [1,4] have suggested that this commonness indicates that they are likely to have a function. In effect they ask: 'if they are not functional, why would they exist?' In contrast I propose that these sequences may be common because they expand rapidly in the genome due to genetic instability. Moreover, this rapid expansion makes it likely that the majority of these unfolded proteins will not have a specific cellular function. Because these proteins are likely to be unfolded and lacking a specific function, I term them 'junk proteins'.

2. Low-complexity sequences

One of the characteristics of disordered proteins is that their sequences often have relatively low complexity. Strictly 'low complexity' is defined through information theory approaches [6]. In practice, it means that there is a bias in favour of only a few amino acids with a larger than normal degree of repetition. The lowest possible complexity sequence would be a run of a single amino acid.

For example the transcriptional coactivator CREB binding protein (SwissProt entry Q92793) has several poly-glutamine repeats with half the sequence having low complexity. Sec9, the yeast homologue of SNAP-25 (SwissProt P40357), which is involved in vesicle fusion, similarly has runs of poly-glutamine with some poly-glutamate. The human amyloid precursor NACP (SwissProt P37840) has tandem sequence repeats over a third of its length. All of these proteins are disordered in solution. More generally, the highest-scoring sequences that are predicted to be disordered by the programme PONDR (predictor of natural disordered regions) are of this type [7]. These proteins have a reduced amino acid alphabet, typically enriched for hydrophilic amino acids and proline, with lower levels of hydrophobic amino acids.

3. Genetic instability

Low-complexity sequences at the protein level are likely to be coded for by low-complexity DNA sequences [8]. These low-complexity DNA sequences tend to be genetically unstable, that is they tend to expand rapidly in the genome over time. In extreme cases this expansion can be over a very small number of generations and lead to disease states such as Hun-

*Fax: (44)-161-275 5082.
E-mail address: simon.lovell@man.ac.uk (S.C. Lovell).

0014-5793/03/$22.00 © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)01223-7

tington's [9], mytonic dystrophy [10] and fragile-X syndrome [11]. Unstable sequences associated with these diseases tend to be tri-nucleotide repeats of the form GNC or CNG. These represent runs of single amino acids at the protein level. The high GC content of the DNA and its low complexity lead to the formation of stem-loop structures that cause the replication machinery make errors, leading to the further expansion of these sequences [12]. More generally, low-complexity sequences in which the DNA forms secondary structure (but not those which do not form secondary structure) expand due to unequal crossing over and slippage in replication [13].

The genome of *Plasmodium falciparum*, the causative agent of malaria, has an extremely large number of low-complexity sequences [14]. As many as 94% of open reading frames on chromosome 3 are of low complexity or have large, low-complexity insertions. The DNA sequences that code for these low-complexity proteins have been predicted to form stem-loop conformations in a manner similar to those found in disease states [8] (although, interestingly for at least one low-complexity stretch of protein, in the Epstein–Barr nuclear antigen, the reverse is true). It is likely, therefore that these sequences exist due to DNA level adaptations.

If low-complexity protein sequences do correspond to DNA sequences that are genetically unstable, we would expect that they would show indications of arising recently in evolutionary time. In fact, this is the case. Nishizawa and Nishizawa [15] have found that proteins with repetitive stretches of amino acids are unlikely to be found in 'ancient' proteins, that is, those that are conserved from prokaryotes through to higher eukaryotes. In this study of yeast sequences, they showed that those repetitive sequences are found in those proteins unique to the subclass of organism studied. Nandi et al. [16] have found that low-complexity proteins differ substantially not only between *Escherichia coli* and *Mycobacteria tuberculosis*, they also find differences between two species of *Mycobacteria* (*M. tuberculosis* and *Mycobacteria leprae*) and, most importantly, between closely related strains of *E. coli*, and between closely related strains of *M. tuberculosis*. It seems likely, then, that well-known genetic processes are responsible for the production of low-complexity DNA sequences, which in turn code for low-complexity protein sequences, and these sequences do indeed arise in periods of time that are evolutionarily short.

4. The likelihood of function

If this is the case, what are the implications for function? Undoubtedly many of these sequences are functional [1,3]. I think it is likely, however, that large numbers of low-complexity, natively unfolded proteins have no specific, protein-mediated, cellular function.

Biological function, and protein function in particular, may be characterised as 'specific', 'general' or 'conditional'. Specific functions are those for which a given protein is specifically adapted. A general function is one that can be ascribed to all proteins due to their general character, such as maintenance of the Donnan equilibrium or as suggested by Forsdyke et al. [17] acting as intracellular 'immunoreceptors'. Conditional functions are those that are only utilised under a specific set of conditions (for example environmental conditions) that may or may not be encountered. Generally when a protein's function is discussed it is the specific function that is meant, although it is important to bear in mind that other functions may also be important.

In order for a gene product to have a specific function, it must firstly arise in the genome and secondly it must acquire this function. The proportion of sequences that have a specific function will depend on the relative frequency of these two processes. It should be noted that new proteins may immediately acquire general functions with no additional selection. The mechanisms of the acquisition of specific function by new gene products are not well understood. Intuitively it would seem to be an unlikely event. Because low-complexity sequences arise quickly it is unlikely that the acquisition of specific function can keep pace with the rate at which the sequences arise. By analogy, the large teetering piles of paper on my desk are not present because I view the collection of paper as a useful pastime, but merely because the mechanisms for paper accumulating are more efficient than my mechanisms for removing it. The amount of paper is dependent only on the balance between these two processes. How much of this paper ends up being useful (functional) is dependent on a third variable, namely my ability to convert it to a useful purpose, that is, to deal with paper work. Since this last conversion is slow, the majority of things on my desk end up being junk, that is, they have no useful function.

At the DNA level, it is well accepted that non-functional sequences exist. The term 'junk DNA' was first used by Ohno [18] to describe the overwhelming majority of DNA (around 95% in humans) that does not code for proteins. Although much of this has other functions there still remains a large amount of DNA which has no apparent specific function and is thought to exist simply because mechanisms exist to replicate it and there is little selection pressure to remove it. Because it has no cellular function it is likely to be mildly deleterious to the cell due to the time and energetic resources required to maintain it.

Some of these sequences are known to produce protein. Transposable elements in the *Rickettsia* genome are translated in the middle of functional open reading frames [19] with no apparent function. Similarly *Alu* repeats, which are not normally transcribed and translated, have been found in the middle of protein-coding regions [20]. These are likely to have been introduced into the middle of open reading frames by retrotransposition or splicing.

Ohno [18], when first using the term 'junk DNA', did not rigorously define it. He did, however, liken fossils of extinct species with non-functional DNA, suggesting that these sequences are the remains of extinct genes. In contrast, I suggest that instead of being very old, many of these sequences are very new. This appears to be true at least of those sequences in the *M. tuberculosis* and *E. coli* genomes discussed above. The term 'junk' as applied to these sequences seems appropriate, suggesting a similarity to something which currently does not have a function, but which is stored in the attic because it may come in useful one day. Whether the majority of these sequences are pressed into service, or whether they are likely to be the victims of an eventual clear out is, however, unclear.

5. Biases in our knowledge

It is clear that a number of low-complexity, unfolded proteins do, in fact, have well-characterised specific functions. These are likely to be proteins that arose through unequal crossing over and replication slippage and then subsequently acquired a function. Once functional in the cell they would be subject to more stringent evolutionary restraints. This acquiring of a function, first suggested by Orgel and Crick [21], has clearly happened in the examples described above.

Dunker et al. [3] have shown that specific functions can be assigned to 98 out of 115 disordered regions described in the literature. The remaining 17 have no known function. They also point out that proving ruling out a particular function does not rule out all functions for a given protein or disordered region, whether they be specific, conditional or general. Similarly Zuckerkandl [22] suggests that a stretch of DNA that is non-functional on one level may be functional on another or in conjunction with an apparently unrelated set of cellular components (that is, it may have conditional functions). The same arguments can be applied to proteins. Of course, simply because we are ignorant of a function does not mean that they are non-functional.

It is likely that the set of well-characterised proteins, which is essentially the set of proteins that appear in the literature, is likely to be substantially enriched for those proteins that are functional. Gerstein has shown that the database of known protein structures does not represent a typical subset of the protein sequence database [4]. It is also possible that the database of protein sequences does not accurately represent the true proteome. Many gene identification programmes use homology to known genes as a method of identification [23]. If the starting set of known genes is systematically missing low-complexity genes then subsequently low-complexity genes will also fail to be identified. Additionally, it is routine to remove low-complexity sequences before attempting gene prediction, to avoid large numbers of false hits [24]. Thus by studying the known proteome it is possible that the predictions of thousands of disordered proteins by Dunker and co-workers is a serious underestimate. Moreover, the set of proteins for which function, or lack thereof, has been described in the literature is the likely source of the heaviest bias. Biologists, understandably, tend to study biological function and are unlikely to spend the same substantial effort in studying that which is functionless.

We may be unable to find functions for many of these proteins due to experimental limitations, but I suggest that we will be unable to find functions for many of them because they are, in fact, non-functional.

6. Conclusions

I wish to emphasise that I do not attempt to cast doubt on work which shows that there are a large number of proteins that have specific functions while disordered. Nor do I doubt that it is necessary to question the current paradigm that unique, well-ordered structure is an absolute requirement for function. I do suggest, however, that in addition to these functional, disordered proteins there are likely to be a large number of non-functional disordered proteins. I base this suggestion on the likely origins of low-complexity DNA sequences, which give rise to low-complexity protein sequences, which in turn give rise to unfolded proteins. I also suggest that current methods are likely to underestimate the frequency of occurrence of these proteins.

Acknowledgements: I thank Drs Jon Moore and Kathy Hentges for valuable suggestions, and Drs David Burke and Kathy Hentges for critical reading of the manuscript. I thank the Wellcome Trust for funding in the form of a Fellowship in Mathematical Biology.

References

[1] Wright, P.E. and Dyson, H.J. (1999) J. Mol. Biol. 293, 321–331.
[2] Dyson, H.J. and Wright, P.E. (2002) Curr. Opin. Struct. Biol. 12, 54–60.
[3] Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M. and Obradovic, Z. (2002) Biochemistry 41, 6573–6582.
[4] Gerstein, M. (1998) Fold. Des. 3, 497–512.
[5] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Garner, E.C., Guilliot, S. and Dunker, A.K. (1998) Pac. Symp. Biocomput. 3, 435–446.
[6] Wootton, J.C. and Federhen, S. (1996) Methods Enzymol. 266, 554–571.
[7] Romero, P., Obradovic, Z., Xiaohong, L., Garner, E.C., Brown, C.J. and Dunker, A.K. (2001) Proteins Struct. Funct. Genet. 42, 38–48.
[8] Xue, H.Y. and Forsdyke, D.R. (2003) Mol. Biochem. Parasitol. 128, 21–32.
[9] MacDonald, M.E., Novelletto, A., Lin, C., Tagle, D., Barnes, G., Bates, G., Taylor, S., Allitto, B., Altherr, M., Myers, R., Lehrach, H., Collins, F.S., Wasmuth, J.J., Frontali, M. and Gusella, J.F. (1992) Nat. Genet. 1, 99–103.
[10] Harley, H.G., Brook, J.D., Floyd, J., Rundle, S.A., Crow, S., Walsh, K.V., Thibault, M.C., Harper, P.S. and Shaw, D.J. (1991) Am. J. Hum. Genet. 49, 68–75.
[11] Richards, R.I., Holman, K., Friend, K., Kremer, E., Hillen, D., Staples, A., Brown, W.T., Goonewardena, P., Tarleton, J., Schwartz, C. and Sutherland, G.R. (1992) Nat. Genet. 1, 257–260.
[12] Richards, R.I. and Sutherland, G.R. (1994) Nat. Genet. 6, 114–116.
[13] Stephan, W. (1989) Mol. Biol. Evol. 6, 198–212.
[14] Pizzi, E. and Frontali, C. (2001) Genome Res. 11, 218–229.
[15] Nishizawa, M. and Nishizawa, K. (1999) Proteins Struct. Funct. Genet. 37, 284–292.
[16] Nandi, T., Kannan, K. and Ramachndran, S. (2003) Curr. Sci. 85, 185–187.
[17] Forsdyke, D.R., Madill, C.A. and Smith, S.D. (2002) Trends Immunol. 23, 575–579.
[18] Ohno, S. (1972) Brookhaven Symp. Biol. 23, 366–370.
[19] Ogata, H., Audic, S., Barbe, V., Artiguenave, F., Fournier, P.-E., Raoult, D. and Claverie, J.-M. (2000) Science 290, 347–350.
[20] Makalowski, W., Mitchell, G.A. and Labuda, D. (1994) Trends Genet. 10, 188–193.
[21] Orgel, L.E. and Crick, F.H.C. (1980) Nature 284, 604–607.
[22] Zuckerkandl, E. (2002) Genetica 115, 105–129.
[23] Burge, C. and Karlin, S. (1997) J. Mol. Biol. 268, 78–94.
[24] Burge, C.B. and Karlin, S. (1998) Curr. Opin. Struct. Biol. 8, 346–354.