FEBS Letters 586 (2012) 2468–2472

FEBS Letters

journal homepage: www.FEBSLetters.org

Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: Is aggregation the main bottleneck?

Annamária F. Ángyán [a], András Perczel [a,b], Zoltán Gáspári [c,*]

[a] Eötvös Loránd University, Institute of Chemistry, Pázmány Péter s. 1/A, H-1117 Budapest, Hungary
[b] HAS-ELU Protein Modelling Group, Pázmány Péter s. 1/A, H-1117 Budapest, Hungary
[c] Pázmány Péter Catholic University, Faculty of Information Technology, Práter u. 50/A, H-1083 Budapest, Hungary

ARTICLE INFO

Article history:
Received 18 April 2012
Revised 4 June 2012
Accepted 5 June 2012
Available online 20 June 2012

Edited by Robert B. Russell

Keywords:
Protein evolution
De novo protein
Aggregation propensity
Random sequence
Structure prediction

ABSTRACT

Present-day proteins are believed to have evolved features to reduce the risk of aggregation. However, proteins can emerge de novo by translation of non-coding DNA segments. In this study we assess the aggregation, disorder and transmembrane propensity of protein sequences generated by translating random nucleotide sequences of varying GC-content. Potential de novo random-sequence proteins translated from regions with GC content between 40% and 60% do not show stronger aggregation propensity than existing ones and exhibit similar tendency to be disordered. We suggest that de novo emerging proteins do not mean an unavoidable aggregation threat to evolving organisms.

1. Introduction

The emerging consensus on protein aggregation is that it is an inherent property of any polypeptide chain and, regardless of their amino acid sequences, the amyloid fibril might be the most favored thermodynamic state of all proteins [1–3]. Even so, proteins display sequence-specific aggregation propensities that can be estimated by in silico methods [4,5]. Thus, proteins can evolve to reduce the risk of aggregation and detailed studies of selected proteins revealed a number of such mechanisms [6]. However, proteins continuously emerge de novo by transcription and translation of previously non-coding DNA segments [7–9]. This poses the question whether novel proteins that did not yet have the chance to reduce their aggregation load by selection can seriously hinder molecular evolution: if the aggregation propensity of de novo proteins is generally high, leading to the aggregation of practically all de novo polypeptides, that might render the chances of the emergence of such proteins negligible.

De novo origin of coding sequences from non-coding ones is a rare but not improbable event, there are e.g. human- and primate-specific proteins thought to have arisen by this mechanism [9–11]. The overall low-level transcriptional activity of the human genome provides a plausible basis for such events [12]. Thus, the aggregation propensity of such proteins is worth to be explored. As the number of known genuine de novo proteins is fairly low, in this study we chose to use an in silico study on random, translated DNA sequences to (i) have a dataset of sufficient size to observe trends, (ii) assess the aggregation propensity before any – however short-time – selection could take place at the protein level and (iii) have a standardized way to assess and compare trends for sequences with different GC-content that can be used as a benchmark for real de novo proteins.

2. Materials and methods

Detailed description of all the methods used and datasets can be found in the online Supplementary data. Random DNA sequences of varying GC-content were generated with the restriction that in-frame STOP codons were avoided. Translated protein sequences were subjected to different algorithms (Table S1) to assess their tendency for aggregation (TANGO [13], WALTZ [14] and FoldAmyloid [15,16]), forming disordered (IUPred [17,18], RONN [19] and VSL2B [20,21]) or transmembrane structures (HMMTOP [22], DASTMfilter [23] and TMHMM [24]). The number of residues predicted to be in the given structural classes by the algorithms were averaged and used as a consensus prediction. The same algorithms were applied to a number of databases representing folded (ASTRAL40, version 1.75), unfolded (DISPROT, version 5.7), aggregation-prone (AmyPDB, last update on 7th April, 2008) and

[*] Corresponding author.
E-mail address: gaspari.zoltan@itk.ppke.hu (Z. Gáspári).

0014-5793/$36.00 © 2012 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.febslet.2012.06.007

transmembrane proteins (PDBTM, version 2.3) as well as the complete human and mouse proteomes (from Uniprot release 2011_05). The obtained one- and two-dimensional distributions at the three properties (disorder, aggregation and transmembrane tendency defined as the percentage of residues falling to these categories in the consensus prediction) were compared by the appropriate variants of the Kolmogorov–Smirnov test. In addition, the area spanned by the sequences in the two- and three-dimensional plots and the overlap between the distributions obtained for different databases were estimated using a grid-based approach. The coding sequences of human *de novo* proteins were obtained by comparing the translated mRNA sequences to the available protein sequences and extracting the nucleotide sequences in the matching region.

3. Results

3.1. Random sequences and predictions of structural features

We generated 10000 random DNA sequences of 480 nucleotides without in-frame STOP codons for each of GC-content regime from 10% to 90% using steps of 10%. The 160-residue length of the translated polypeptides can be regarded as a reasonable estimate of average domain size in proteins [25,26]. Although not the full GC-range explored is biological relevance, as for example the human genome has an average GC-content of 41% and ranges approximately from 20% to 60% [27], we chose our systematic scan to identify general trends. After translating all of the 9 × 10000 nucleotide sequences, we have used BLAST [28] search to assess the similarity of the resulting random de novo proteins to known sequences. No hits were found below an *E*-value of $1 * 10^{-10}$, and only 30 hits were found below an *E*-value of 0.001 (Table S2). Thus, our random sequence set is sufficiently distinct from extant proteins. Next, we used a set of prediction algorithms to assess their aggregation loads (TANGO [13], WALTZ [14] and FoldAmyloid [15,16]), their disorder (IUPred [17,18], RONN [19] and VSL2B [20,21]) and transmembrane propensities (HMMTOP [22], DASTM-filter [23] and TMHMM [24]). None of the applied methods uses evolutionary information during data processing like today's best-performing secondary structure prediction tools [29]), thus, we expect that they can be used for de novo sequences in an unbiased way. We have performed the same predictions on several databases representing folded, disordered, transmembrane and aggregation-prone proteins as well as the complete human and mouse proteomes. It is important to stress that we do not wish to assess the absolute aggregation propensity of any of the sequence sets, rather, in all evaluations below, we analyze trends and draw conclusions from comparisons of predictions made with the same toolkit.

3.2. General trends

Naturally, the amino-acid composition of our random datasets reflects the standard genetic code organization. At low GC-content, hydrophobic amino acids appear with higher frequency, typically representing 50–70% of all residues. At 90% GC-content, only 10% of all residues are hydrophobic and 20% is arginine (Table S3). In present-day proteomes, acidic amino acids (Glu, Asp) are remarkably more frequent than expected from the codon distribution in the standard genetic code [30,31] (Table S4). At high GC-content, basic amino acids are overrepresented in the standard code-translated dataset relative to present-day natural proteins. The mean net charge of random de novo sequences exhibits a minimum at 40% GC-content and it is still higher than the highest value obtained for present-day proteins, corresponding to IDPs. The mean hydrophobicity shows a decreasing trend with increasing GC-content and covers a wider range than that of present-day proteins (Figs. S1 and S2). According to the averaged structural predictions, the GC-content of the underlying DNA sequences governs the structural preferences of the random proteins with clearly identifiable trends that are much more pronounced than the variations in the simple physico-chemical parameters. Intrinsic disorder is a dominant feature of sequences with coding regions of high GC-content (Table 1). Around 50% GC-content, 25% of all amino acid residues is predicted to be disordered. In this respect, only aggregation-prone and transmembrane present-day proteins have a lower average value. At 60% GC-content and above, random sequences are practically fully disordered containing on average one or two long disordered regions (Fig. 1a).

The propensity to form transmembrane helices is relatively high at low GC-content and decreases rapidly to practically vanish over 60% GC-content. At 40% GC-content, the average ratio of residues in transmembrane segments is comparable to those in the complete human and mouse proteomes (Fig. 1b).

The aggregation load in random sequences is highest at low GC-content and drops quickly to an average 22% of all residues at 50% GC-content. At and above 60% GC, practically all parameters investigated are on average below those of present-day proteins (Fig. 1c).

3.3. Interplay between structural properties

We have investigated whether the predicted structural preferences are independent of each other or there are some associations. We have investigated this aspect at the sequence level, calculating correlations between the percentage of residues predicted to be disordered, transmembrane and aggregation-prone (Tables 2 and S5). Both these values and two-dimensional plots of these features indicate that these features are loosely interdependent and not all regions of the disorder-transmembrane-aggregation space are accessible either for random or for existing protein sequences (Fig. 1d–f).

Sequences with higher percentage of disordered residues tend to have less transmembrane helices and lower aggregation propensity. However, the nature of the interdependence is different, with a large range of variation in transmembrane propensity at low disorder tendency, whereas aggregation load seems to be more strictly negatively associated with disorder. On the other hand, the tendency to form transmembrane helices shows a positive association with aggregation propensity. These trends suggest that amino acid composition plays a decisive role in defining these structural features.

3.4. Comparison to databases

We stress that we do wish to assess the absolute propensity of any sequence set to be disordered, form transmembrane helices and being prone to aggregation, rather use consensus predictions for comparative purposes. Our results allow to compare the trends observed for random-sequence proteins to those observed in natural ones. Below, unless noted otherwise, we will focus on random proteins translated from the physiologically most relevant range of GC-content, between 40% and 60%.

Intrinsic disorder depends heavily on the underlying GC-content of the random sequences, at 60% the random sequences show clearly higher disorder propensity than even DISPROT, whereas at 40 and 50% of the translated proteins are predicted to contain less disordered residues than those in extant proteomes (Table 1).

The tendency to form transmembrane helices is much lower for random sequences translated from DNA of 50% or higher GC content than for extant proteins except globular and disordered ones (Table 1).