Article

Progressively more subtle aggregation avoidance strategies mark a long-term direction to protein evolution

Authors: S.G. Foy[1,2], B.A. Wilson[1], M.H.J. Cordes[3], J. Masel[1]*

Affiliations:

[1]Department of Ecology and Evolutionary Biology, University of Arizona.

[2]present address: St. Jude Children's Research Hospital, Memphis, Tennessee.

[3]Department of Chemistry & Biochemistry, University of Arizona.

*Correspondence to: masel@email.arizona.edu

Short title: Long-term directionality of protein evolution

Keywords: phylostratigraphy, gene age, aggregation propensity, protein folding, protein misfolding

Abstract

To detect a direction to evolution, without the pitfalls of reconstructing ancestral states, we need to compare "more evolved" to "less evolved" entities. But because all extant species have the same common ancestor, none are chronologically more evolved than any other. However, different gene families were born at different times, allowing us to compare young protein-coding genes to those that are older and hence have been evolving for longer. To be retained during evolution, a protein must not only have a function, but must also avoid toxic dysfunction such as protein aggregation. There is conflict between the two requirements; hydrophobic amino acids form the cores of protein folds, but also promote aggregation. Young genes have a hydrophilic amino acid composition, which is presumably the simplest solution to the aggregation problem. Young genes' few hydrophobic residues are clustered near one another along the primary sequence, presumably to assist folding. Later evolution increases hydrophobicity, increasing aggregation risk. This risk is counteracted by more subtle effects in the ordering of the amino acids, including a reduction in the clustering of hydrophobic residues until they eventually become more dispersed than if distributed randomly. This dispersion has previously been reported to be a general property of proteins, but here we find that it is restricted to old genes. Quantitatively, the index of dispersion delineates a gradual trend, i.e. a decrease in the clustering of hydrophobic amino acids over billions of years.

Introduction

Proteins need to do two things to ensure their evolutionary persistence: fold into a functional conformation whose structure and/or activity benefit the organism, and also avoid folding into harmful conformations. Amyloid aggregates are a generic structural form of any polypeptide, and so pose a danger for all proteins (Monsellier and Chiti 2007). Several lines of evidence suggest that aggregation avoidance is a critical constraint during protein evolution. Highly expressed genes are less aggregation-prone (Tartaglia et al. 2007), and evolve more slowly due to greater selective constraint against alleles that increase the proportion of mistranslated variants that misfold (Drummond et al. 2005; Drummond and Wilke 2008). Genes that homo-oligomerize or are essential (Chen and Dokholyan 2008) or that degrade slowly (De Baets et al. 2011) are also less aggregation-prone. Aggregation-prone stretches of amino acids tend to have translationally optimal codons (Lee et al. 2010), and be flanked by “gatekeeper” residues (Rousseau et al. 2006). Disease mutations are enriched for aggregation-promoting changes (De Baets et al. 2015; Reumers et al. 2009), and known aggregation-promoting patterns are underrepresented in natural protein sequences (Broome and Hecht 2000; Buck et al. 2013). Thermophiles, whose amino acids need to be more hydrophobic, show exaggerated aggregation-avoidance patterns (Thangakani et al. 2012).

Here we ask whether and how proteins get better at avoiding aggregation during the course of evolution. Absent a fossil record or a time machine, biases introduced during the inference of ancestral protein states (Trudeau et al. 2016; Williams et al. 2006) make it difficult to assess how past proteins systematically differed from their modern descendants. We have therefore developed an alternative method to study protein properties as a function of evolutionary age, one that does not rely on ancestral sequence reconstruction.

While all living species share a common ancestor, all proteins do not. It has become clear that protein-coding genes are not all derived by gene duplication and divergence from ancient ancestors, but instead continue to originate de novo from non-coding sequences (McLysaght and Guerzoni 2015). Different gene families (i.e. sets of homologous genes) therefore have different ages, and the properties of a gene can be a function of age.

The age of a gene can be estimated by means of its “phylostratum”, which is defined by the basal phylogenetic node shared with the most distantly related species in which a homolog of the gene in question can be found (Domazet-Lošo et al. 2007). Failure to find a still more distantly related protein homolog (i.e. failure of a gene to appear older) can have multiple causes. First, more distantly related homologs might not exist, as a consequence of de novo gene birth either from intergenic sequences or from the alternative reading frame of a different protein-coding gene (the latter yielding nucleotide but not amino acid homology). Second, apparent age might indicate the time not of de novo birth but of horizontal gene transfer (HGT) from a taxon for which no homologous genes have yet been sequenced. Third, independent loss of the entire gene family in multiple distantly related lineages can yield a pattern of apparent gain. Fourth, divergence between gene duplicates might be so extreme that homology can no longer be detected.

The diversity of sequenced taxa now available makes the second possibility (HGT) increasingly unlikely, especially outside microbial taxa that experience high levels of HGT; here we minimize this possibility by focusing on the set of mouse genes. The same wealth of sequenced taxa also makes the third possibility (phylogenetically independent loss of the entire gene family) unlikely, given the large number of independent loss events implied. More importantly, neither HGT nor independent loss are likely to drive systematic trends in protein properties as a function of apparent gene age; instead, they are likely to dilute any underlying patterns resulting from other determinants of apparent gene age.

Most critiques of the interpretation of phylostratigraphy in de novo gene terms therefore focus on the fourth possibility, specifically the concern that trends may be driven by biases in the degree to which homology is detectable (Albà and Castresana 2007; Moyers and Zhang 2016, 2017, 2015). In particular, homology is harder to detect for shorter and faster-evolving proteins, which might therefore appear to be young, giving false support to the conclusion than young genes are shorter and faster-evolving. The problem of homology detection bias extends to any trait that is correlated with primary factors, such as length or evolutionary rate, that directly affect homology detection. We previously studied such a trait, intrinsic structural disorder (ISD), and found that statistically correcting for evolutionary rate did not affect the results, and that statistically correcting for length made them stronger (Wilson et al. 2017). This suggested that the pattern in ISD was likely driven by time since de novo gene birth, rather than by homology detection bias.

Here we trace a number of other protein properties as a function of apparent gene family age, including aggregation propensity and hydrophobicity, and find a particularly striking trend for the degree to which hydrophobic residues are clustered along the primary sequence. This trend, as with the previous ISD work, experiences negligible change after correction for length, evolutionary rate, and expression, and is thus not a result of homology detection bias. Our results point to a systematic shift in the strategies used by proteins to avoid aggregation, as a function of the amount of evolutionary time for which they have been evolving.

Results

We assigned mouse genes to gene families and to times of origin, and assigned a protein aggregation propensity score to each protein on the basis of its amino acid sequence (see Methods). No clear trend is seen in aggregation propensity as a function of gene age (Fig. 1), although all genes (black) show lower aggregation propensity than would be expected if intergenic mouse sequences were translated into polypeptides (blue). Note that intergenic sequences represent not only the raw material from which de novo genes could emerge, but also the fate of any sequence, e.g. a horizontally transferred gene, that is subjected to neutral mutational processes.

However, striking patterns emerge when we decompose aggregation avoidance into the effect of amino acid composition (with hydrophobic amino acids making aggregation more likely), and the effect of the exact order of a given set of amino acids. The contribution of amino acid composition alone can be assessed by scrambling the order of the amino acids (Fig. 2, bottom), revealing that young genes make greater use of amino acid composition to avoid aggregation. The pattern is mirrored by other measurements of the hydrophobicity of the amino acid composition (Fig. 2, top and middle, intrinsic structural disorder as per (Wilson et al. 2017) shown in Fig. S1), with the decline in hydrophilicity taking place over ~200 million years. Previously reported differences in the aggregation propensity (Tartaglia et al. 2005) and hydrophobicity (Mannige et al. 2012) of proteomes from different organisms might therefore be accounted for by systematic variation among species in the composition of old vs. young genes; in our analysis, all proteins were taken from the same mouse species, removing this confounding factor.

The contribution of amino acid ordering alone, independent from amino acid composition, can be assessed as the difference between the aggregation propensity of the actual protein and that of a scrambled version of the protein. We expected real proteins to be less aggregation-prone than their scrambled controls (Buck et al. 2013), and confirmed this for the very oldest proteins (Fig. 3, orange confidence intervals for genes shared with prokaryotes lie below 0). But surprisingly, the opposite was true for young genes (Fig. 3, orange confidence intervals for phylostrata from metazoa onward lie above 0). In other words, they are more aggregation-prone than would be expected from their amino acid composition alone.

One possible source of increased aggregation propensity is if young genes, struggling to achieve any kind of fold at all given their low hydrophobicity (Dill 1990), cluster their few hydrophobic amino acid residues closer together along the sequence. Such clustering could allow proteins to evolve small, foldable, potentially functional domains within an otherwise disordered sequence (Uversky et al. 2000). Alternatively and still more primitively, very highly localized clustering could produce short peptide motifs that cannot fold independently but acquire structure conditionally through binding or oligomerization (Davey et al. 2012; Gunasekaran et al. 2004). Hydrophobic clustering also increases the danger of aggregation (Monsellier et al. 2007); indeed, there is significant congruence between mutations that increase the stability of a fold and those that increase the stability of the aggregated or otherwise misfolded form (Sánchez et al. 2006).

We find that young genes do show hydrophobic clustering, while very old genes show interspersion of hydrophobic amino acid residues (Fig. 4), and that this accounts for much of the excess aggregation propensity of young genes relative to scrambled controls (Fig. 3 blue points are closer to zero than orange points). Previous reports have suggested that the danger of aggregation selects against hydrophobic clustering (Monsellier et al. 2007). In other words, among consecutive blocks of amino acids, the variance in hydrophobicity is lower than the mean, i.e. the index of dispersion is less than one in proteins overall (Irbäck et al. 1996; Schwartz et al. 2001) and in the core of protein folds (Patki et al. 2006). In the present analysis, this holds true only for old, highly evolved proteins. Younger proteins not only appear less evolutionarily constrained to intersperse polar and hydrophobic residues, but to the contrary, their hydrophobic residues show excess concentration near one another along the sequence, increasing aggregation propensity. Our results are extremely robust when we control for protein length, evolutionary rate, and expression level (Fig. S2). We also attempted to control for experimentally verified transmembrane status (use of sequence-based prediction would be problematically confounded), but found only 10 mouse transmembrane proteins plus 37 mouse proteins with human transmembrane homologs in the "Membrane Proteins of Known 3D Structure" database (Stansfeld et al. 2015) (http://blanco.biomol.uci.edu/mpstruc/ accessed July 16, 2017) Unsurprisingly given their small number, the increased clustering of transmembrane proteins was not significant as a fixed effect within our linear model (p>0.05). Transmembrane proteins showed the same trend in clustering as a function of age as did mouse genes as a whole.

Dispersion/clustering is a metric for which genes that have been evolving for longer have different properties from genes that are "less evolved", creating a consistent direction of evolution over billions of years. This directionality of evolution can be interpreted as a slow shift from a primitive strategy for avoiding misfolding in young genes to more subtle strategies in old genes.

The primitive aggregation avoidance strategy used by young genes is simply to have a hydrophilic amino acid composition (Fig. 2), creating intrinsic structural disorder (Linding et al. 2004; Thangakani et al. 2012; Wilson et al. 2017). Given such an amino acid composition, young genes might form an early folding nucleus by concentrating hydrophobic amino acids in localized regions of the sequence (Fig. 4, right), while still keeping total hydrophobicity and hence aggregation propensity within tolerable limits (Figs. 1-2). Such a folding nucleus would not necessarily be an entire independently folded domain. In particular, some origin theories posit that ancient proteins first achieved folding by becoming structured only upon binding to some interaction partner (Soding and Lupas 2003; Zhu et al. 2016). In contemporary proteins, potential representatives of nascent structure are found in intrinsically disordered proteins that contain peptide-length binding motifs (small linear interaction motifs; SLiMs), many of which become ordered when bound to a partner (Davey et al. 2012). We do not, however, find that young genes have more known SLiMs (Fig. S3).

In contrast to young genes, older genes have higher hydrophobicity, which must be offset by the evolution of other aggregation-avoidance strategies (Thangakani et al. 2012). For such changes to occur through descent with modification probably happens only slowly. Changing the amino acid composition of a protein takes ~200 million years (Figs. 2 and S1); changing the index of dispersion requires such a large number of changes that it is extraordinarily slower, with a consistent direction to evolution visible over the entire history of life back to our common ancestor with prokaryotes.

Note that our very youngest phylostratum, of mouse genes shared only with rats, shows less clustering than other young genes, suggesting that rapid change in the index of dispersion may be possible (in the other direction) after all, on short and recent timescales. However, very young gene families are subject to significantly higher death rates than other gene families (Palmieri et al. 2014). With gene family loss so common at first, it is possible that the rapid initial increase in clustering is due to differential retention of gene families with highly clustered amino acids. This interpretation of the data is consistent with explaining how slow the later fall in clustering is, by positing that descent with modification is constrained to change clustering values slowly.

The youngest genes show similar clustering to what would be expected were intergenic sequences to be translated (Fig. 4, blue). Clustering of amino acids translated from non-coding intergenic sequences is a direct consequence of the clustering of nucleotides; indices of dispersion at the nucleotide level are all above the expectation of one from a Poisson process, in the range 1.2-1.9 for intergenic sequences and 1.1-1.8 for masked intergenic sequences, depending on which nucleotides are considered. (The lowest indices are found for the GC vs. AT contrast, presumably due to avoidance of CpG sites causing a general paucity of clusters of G and C.) Very short tandem duplications, e.g. as may arise from DNA polymerase slippage, automatically create segments in which the duplicated nucleotide is overrepresented; observed nucleotide clustering values greater than one can therefore be interpreted as a natural consequence of mutational processes. The consequence of this mutational pattern is therefore a small and fortuitous degree of preadaptation, i.e. intergenic sequences have a systematic tendency toward higher clustering than “random”, in a manner that facilitates the de novo birth of new genes.

Discussion

As discussed in the Introduction, apparent gene family age can be a function of time since i) gene birth, ii) HGT, iii) divergence from other phylogenetic branches all of which have independently lost all members of the gene family, or iv) rapid divergence of a gene made homology undetectable. In all cases, our results describe evolutionary outcomes as a function of time elapsed since that event. In the case of our primary result on clustering, this means that genes appear with clustering values similar to those expected from intergenic sequences, are retained only if their clustering is exceptionally high, and then show gradual declines in clustering after that.

We believe that gene birth is the most plausible driver of our results. HGT is rare in more recent ancestors of mice, simultaneous loss in so many branches is unlikely, and statistical correction for evolutionary rate, length and expression (Fig. S2) has, in contradiction to the predictions of homology detection bias, a negligible effect on our results. However, our results on the evolution of protein properties following a defining event remain of interest under all scenarios of what the gene-age-determining event is.

There are three ways to explain subsequent patterns as a function of gene family age. The two mentioned so far are biases in retention after birth, and descent with modification. The third possibility is that the conditions of life were significantly different at different times, and hence so were the biochemical properties of proteins born/transferred/rapidly diverged at that time. Specifically, ancestral sequence reconstruction techniques have been used to infer that proteins in our ancestral lineage became progressively less thermophilic (Gaucher et al. 2008).

This might explain why young genes are more hydrophilic; they were born at more permissive lower temperatures. However, ancestral reconstruction techniques are likely biased toward consensus amino acids that are fold-stabilizing (Bloom and Glassman 2009; Godoy-Ruiz et al. 2004; Lehmann et al. 2000; Steipe et al. 1994) and hence may be more hydrophobic (Trudeau et al. 2016; Williams et al. 2006). Alarmingly, ancestral reconstruction also suggests that the ancestral mammal was a thermophile (Trudeau et al. 2016). What is more, the main trend that we see of hydrophobicity/thermophilicity as a function of gene age is on shorter timescales; billions of years of common evolution has erased the differences in starting points. It is the more subtle signal of hydrophobic amino acid dispersion that shows the long-term pattern. However, variation in the conditions of life at the time of gene origin remains a plausible explanation for the idiosyncratic differences between phylostrata, i.e. for the remaining, statistically meaningful deviations of individual phylostrata from the trends reported here.

We have already invoked differential retention as a possible driver of the short-term evolutionary increase in the clustering values of young genes. It is logically possible that the long-term trend in clustering values is also a result of differential retention; if gene families with higher clustering values are more likely to be lost, different gene ages represent different spans of time in which this loss has had an opportunity to occur. Given the billion year time scales and thus enormous number of lost gene families this implies, this seems at present a less plausible scenario than descent with modification for different durations following different dates of origin. In other words, descent with modification seems the most plausible of the three possible drivers of biochemical patterns as a function of gene age, independently of what exactly “gene age” means.

Note that our findings go in the opposite direction to those of Mannige et al. (2012), who used more speciation-dense branches as a proxy for longer effective evolutionary time intervals, to infer an evolutionary trend away from, rather than toward, hydrophobicity. Part of this discrepancy (“oiliness” in Fig. 2 is the same metric as used in their work) may arise from differences in which proteins are present in which species, which could be a confounding factor when Mannige et al. (2012) attributed proteome-wide trends to descent with modification. Mannige et al. (2012) also confirmed their results for single genes, but did not, in that portion of their analysis, also confirm that results were not sensitive to the difficulty of scoring speciation-density in prokaryotes.

We propose that our findings may be best explained by three phases of protein evolution under selection for proteins that both avoid misfolding and have a function. First, a filter during the gene birth process gives rise to low hydrophobicity in newborn genes (Wilson et al. 2017), as the simplest way to avoid misfolding. Second, young genes with their few hydrophobic amino acids clustered together are more likely to have functional folds that remain adaptive for some time after birth, and so are differentially retained in the period immediately after birth (when young genes are subject to very high rates of attrition (Palmieri et al. 2014)). Finally these two initial trends are both slowly reversed by descent with modification, continuing over billions of years of evolutionary search for better solutions for exceptions to the intrinsic correlation between propensity to fold and propensity to misfold.

The protein folding problem is notoriously hard. Here we see that it isn’t just hard for human biochemists – it’s so hard that evolution struggles with it too. Proteins evolve to find stable folds despite the correlated and ever-present danger of aggregation. They do so via a slow exploration of an enormous sequence space, a search that has yet to saturate after billions of years (Povolotskaya and Kondrashov 2010). Given the enormous space that has already been searched, existing protein folds, especially of older gene families, may therefore be a highly unrepresentative sample of the typical behaviors of polypeptide chains. Protein folds are best thought of as a collection of corner cases and idiosyncratic exceptions, which are hard to find even for evolution, let alone for our “free-modeling” techniques to predict ab initio.

Materials and Methods

M. musculus proteins from Ensembl (v73) were assigned gene families and gene ages as described elsewhere (Wilson et al. 2017). To briefly outline this previous procedure, BLASTp (Altschul et al. 1997) against the National Center for Biotechnology Information (NCBI) nr database with an E-value threshold of 0.001 was used for preliminary age assignments for each gene, followed by a variety of quality filters. Genes unique to one species were excluded due to the high rate of sequences falsely annotated as protein-coding genes, leaving Rodentia as the youngest phylostratum. Paralogous genes were clustered into gene families, and a single age was reconciled per gene family, which filtered out some inconsistent performance of BLASTp. Numbers of genes and gene families in each phylostratum can be found in Table S1 of Wilson et al. (2017). “Cellular Organisms” contains all mouse gene families that share homology with a prokaryote.

Intergenic control sequences were also taken from previous work (Wilson et al. 2017), including the Masked Control sequences taken only from RepeatMasked (Smit et al. 2015) intergenic sequences. Briefly, one intergenic control sequence per gene was taken 100nt downstream from the end of the 3' end of the transcript, with stop codons excised until a length match to the neighboring protein-coding gene was obtained. A second control sequence per gene began 100nt further downstream. This choice of location ensures that control sequences are representative of genomic regions in which protein-coding genes are found.

Aggregation propensity was scored using TANGO (Fernandez-Escamilla et al. 2004) and Waltz (Maurer-Stroh et al. 2010). We counted the number of amino acids contained within runs of at least five consecutive amino acids scored to have >5% aggregation propensity, added 0.5, and divided by protein length to obtain a measure of the density of aggregation-prone regions. For those scores derived using TANGO, we then performed a Box-Cox transformation (λ=0.362, optimized using only coding genes not controls) prior to linear model analysis in Figs. 1 and S1. Central tendency estimates and confidence intervals were then back transformed for the plots. Paired differences in TANGO scores or Waltz scores between genes and scrambled controls were not transformed. Results were qualitatively indistinguishable when runs of at least six consecutive amino acids were analyzed instead of runs of at least five.

The index of dispersion was assessed by comparing the variance in hydrophobicity between blocks of $s = 6$ consecutive amino acids to the mean hydrophobicity (Irbäck et al. 1996). Result for different values of s yielded qualitatively similar resuls. Where the amino acid length was not divisible by six, an average was taken over all phases for the blocking procedure, with a few amino acids neglected at each end yielding a truncated length of N. Following past practice, amino acid sequences were transformed into binary hydrophobicity strings by taking the six amino acids Leu, Ile, Val, Phe, Met, and Trp as hydrophobic (+1) and the other amino acids as hydrophilic (-1), summing to a value σ_k for each block $k = 1, \dots, N/s$ and $M = \sum_{k=1}^{N/s} \sigma_k$ overall (Irbäck and Sandelin 2000). The normalized index of dispersion

$$\psi = \frac{s}{N} \sum_{k=1}^{N/s} \frac{1}{K} (\sigma_k - sM/N)^2,$$

where the normalization factor for length N and total hydrophobicity M of a protein is

$$K = s \frac{N^2 - M^2}{N^2 - N} \left(1 - \frac{s}{N}\right).$$

For randomly distributed amino acids of any length N and hydrophobicity M, this normalization makes the expectation of ψ equal to 1. For nucleotide dispersion, blocks of length $s = 18$ rather than 6 were used. Nucleotide dispersion scores were calculated for each possible permutation as to which nucleotides were scored as +1 and which as -1 (e.g. G and C as +1 and A and T as -1 constitutes one permutation). Amino acid dispersion values ψ were Box-Cox transformed (λ=-0.295) prior to use in linear models.

To generate a scrambled control sequence that is paired to each gene, we simply sampled its amino acids without replacement. To generate dispersion-controlled scrambled sequences, 1000 scrambled sequences of each protein were produced, and the one that most closely matched the index of dispersion of the focal gene was retained. This left the average gene with a clustering value 0.0035 higher than its matched control, with the mean difference of the absolute deviation between a gene and its matched control equal to 0.0057, showing a close match with little directional bias.

Source data for the statistical analyses and figures are provided in Supplementary Tables S1-S6. Code associated with generating and analyzing these tables is publicly available at https://github.com/MaselLab.

Acknowledgments: This work was supported by the John Templeton Foundation (39667), and the National Institutes of Health (GM104040). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Rafik

Neme for insightful discussions and Joost Schymkowitz and Rob van der Kant of the VIB Switch Laboratory for providing us with a stand-alone Waltz script.

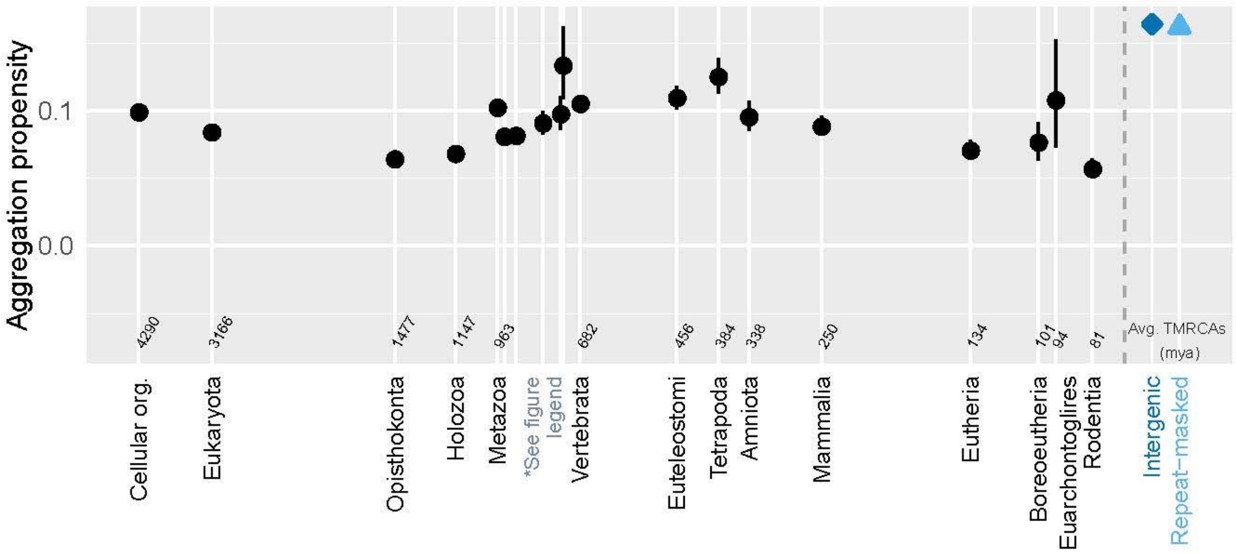

Fig. 1. Mouse genes show little pattern in aggregation propensity (assessed via TANGO) as a function of age. Genes (black) show less aggregation propensity than intergenic controls (blue). Back-transformed central tendency estimates +/- one standard error come from a linear mixed model, where gene family and phylostratum are random and fixed terms respectively. Importantly, this means that we do not treat genes as independent data points, but instead take into account phylogenetic confounding, and use gene families as independent data points. Times to most recent common ancestor (TMRCAs) were taken from TimeTree.org (Kumar et al. 2017) on February 18, 2016. We used the arithmetic means of the TMRCAs of the focal taxon shown on the x-axis and the preceding taxon (i.e. the estimated midpoint of the interior branch of the tree), and these times are displayed on a log scale. Cellular organism age is shown as the midpoint of the last universal common ancestor and the last eukaryotic common ancestor. The taxon names omitted for space reasons follow the sequence Metazoa, Eumetazoa, Bilateria, Deuterostomia, Chordata, Olfactores, Vertebrata.

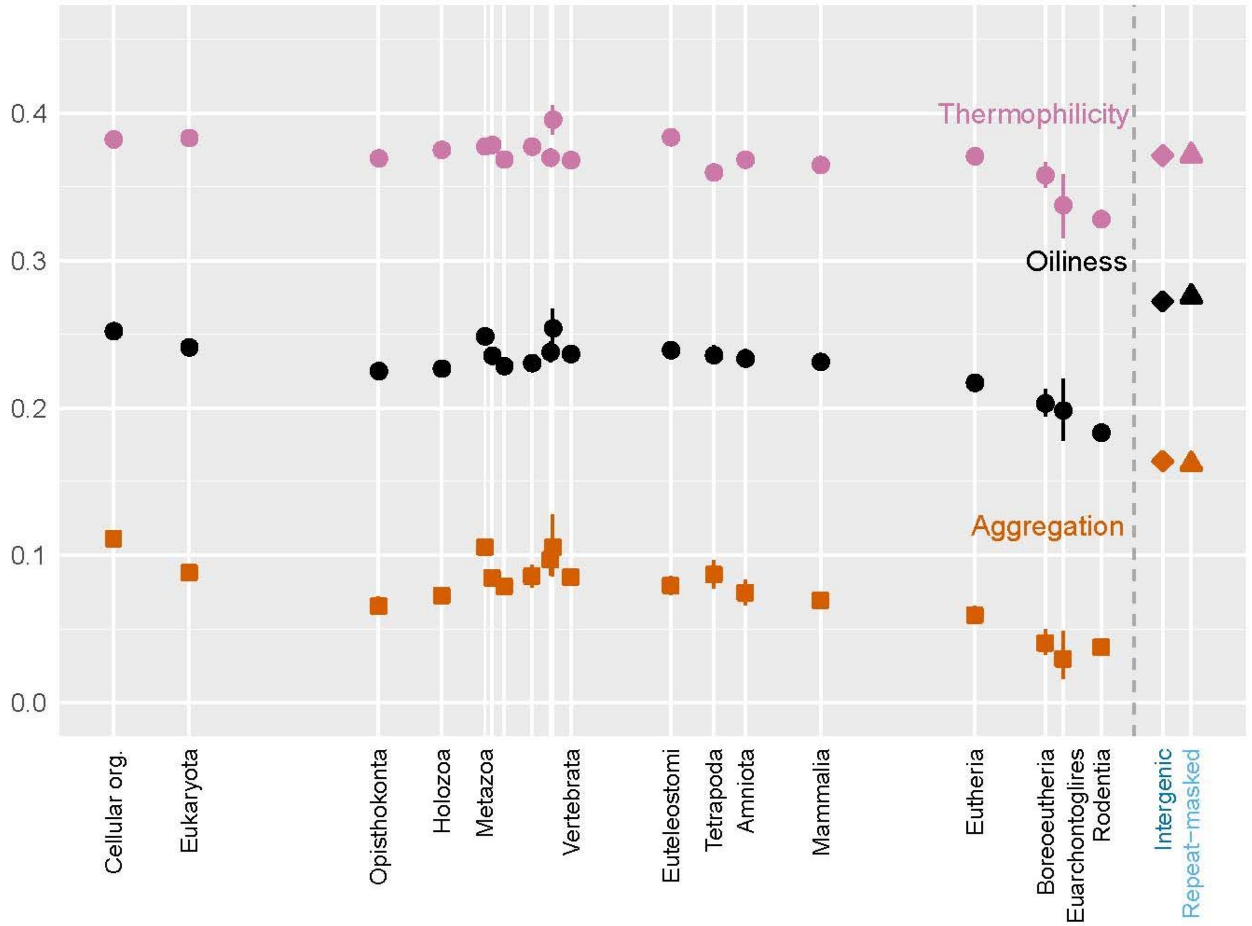

Fig. 2. Three different measures for the hydrophobicity of the amino acid content as a function of gene family age. "Aggregation" represents the TANGO results from scrambled versions of genes, and hence captures the effect of amino acid composition on whatever TANGO captures. The use of scrambled genes is indicated by squares, with unscrambled genes as circles and intergenic controls as diamonds or triangles depending on whether repeat sequences are excluded. Oiliness represents the content (between 0 and 1) of the four most hydrophobic amino acids, FILV, as used in the analysis of Mannige et al. (2012), subjected to a Box-Cox transform with λ= 0.869 prior to model fitting. Thermophily represents the content of ILVYWRE, as analyzed by Boussau et al. (2008), subjected to a Box-Cox transform with λ= 2.412 prior to model fitting; thermophily is dominated by the same general hydrophobicity trend as the other two measures. The hydrophobicity measurement of Irbäck et al. (1996), namely content of FILVMW, is not shown, but is indistinguishable from the FILV oiliness measure. While the trend as a function of gene age is similar in each case, the aggregation measurement shows the most striking deviation from intergenic control sequences. Back-transformed central tendency estimates +/- one standard error come from a linear mixed model, where gene family and phylostratum are random and fixed terms respectively. The x-axis is the same as for Figure 1.

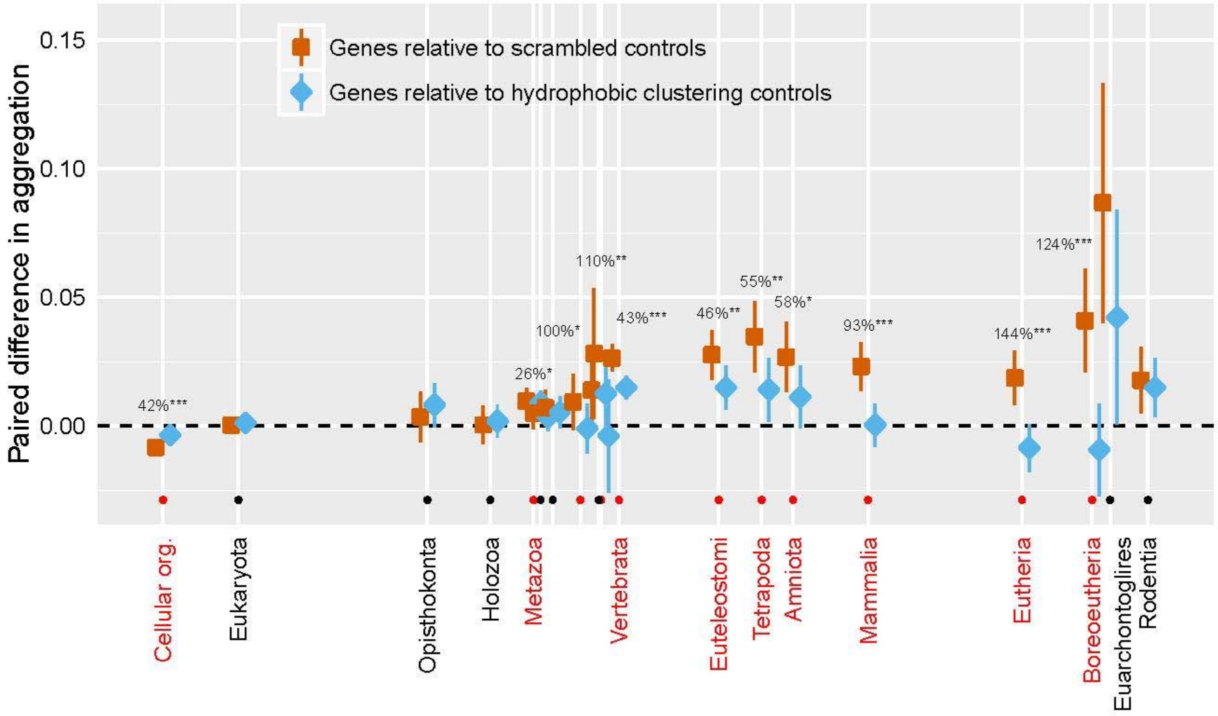

Fig. 3. Only very old genes have aggregation propensities lower than that expected from their amino acid composition alone (orange < dashed line indicating expectation of 0). This puzzling finding is reduced when we account for dispersion (blue is closer than orange is to the 0 dashed line) using a scrambled sequence that is controlled to have a similar dispersion value. The clustering of hydrophobic amino acids in young genes acts to increase their aggregation propensity. 95% confidence intervals are shown, based on a linear mixed model where gene family and phylostratum are random and fixed terms respectively. Note that blue and orange confidence intervals should be compared only to the reference value of zero, and not to each other, due to the paired nature of the data. For phylostrata shown in red and indicated by an orange dot, the difference between blue and orange was significant (**p*<0.01, ***p*<0.001, ****p*<0.0001), and the percentage of deviation from 0 accounted for by the control is shown. For most phylostrata where the difference between blue and orange was non-significant (indicated by a black dot and black text), the orange deviated little from 0, so there was little or nothing for the blue clustering control to account for. Results are shown for TANGO; results for Waltz trend in the same direction but are weaker (Fig. S4). The x-axis is the same as for Figure 1.

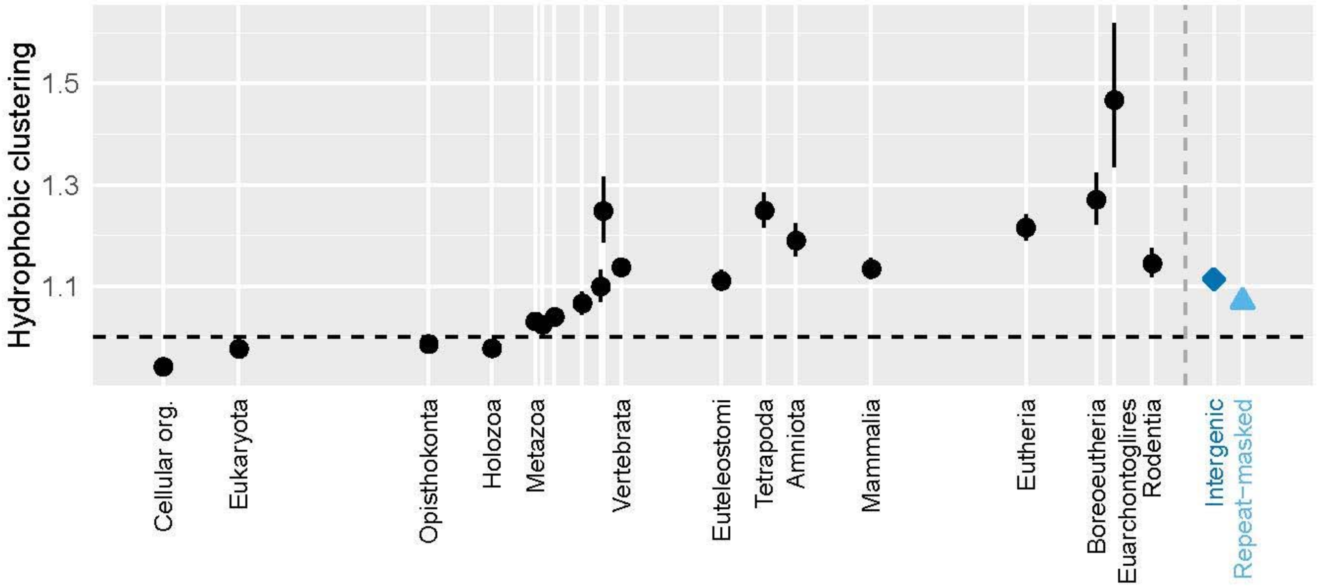

Fig. 4. Clustering initially follows that of its raw material, and evolves rapidly upward at first, but then decays downward extremely slowly, indicating a long-term direction of evolution. Only the oldest genes have hydrophobic amino acids spread out from each other, as previously reported; young genes have clustered hydrophobic amino acids. Back-transformed central tendency estimates +/- one standard error come from a linear mixed model, where gene family and phylostratum are random and fixed terms respectively. The x-axis is the same as for Figure 1.

References

Albà MM, Castresana J 2007. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol 7: 1-8. doi: 10.1186/1471-2148-7-53

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402. doi: 10.1093/nar/25.17.3389

Bloom JD, Glassman MJ 2009. Inferring Stabilizing Mutations from Protein Phylogenies: Application to Influenza Hemagglutinin. PLoS Comput Biol 5: e1000349.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M 2008. Parallel adaptations to high temperatures in the Archaean eon. Nature 456: 942-945. doi: 10.1038/nature07393

Broome BM, Hecht MH 2000. Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis1. J Mol Biol 296: 961-968. doi: 10.1006/jmbi.2000.3514

Buck PM, Kumar S, Singh SK 2013. On the Role of Aggregation Prone Regions in Protein Evolution, Stability, and Enzymatic Catalysis: Insights from Diverse Analyses. PLoS Comput Biol 9: e1003291. doi: 10.1371/journal.pcbi.1003291

Chen Y, Dokholyan NV 2008. Natural Selection against Protein Aggregation on Self-Interacting and Essential Proteins in Yeast, Fly, and Worm. Mol Biol Evol 25: 1530-1533. doi: 10.1093/molbev/msn122

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ 2012. Attributes of short linear motifs. Molecular BioSystems 8: 268-281. doi: 10.1039/C1MB05231D

De Baets G, Reumers J, Delgado Blanco J, Dopazo J, Schymkowitz J, Rousseau F 2011. An Evolutionary Trade-Off between Protein Turnover Rate and Protein Aggregation Favors a Higher Aggregation Propensity in Fast Degrading Proteins. PLoS Comput Biol 7: e1002090. doi: 10.1371/journal.pcbi.1002090

De Baets G, Van Doorn L, Rousseau F, Schymkowitz J 2015. Increased Aggregation Is More Frequently Associated to Human Disease-Associated Mutations Than to Neutral Polymorphisms. PLoS Comput Biol 11: e1004374. doi: 10.1371/journal.pcbi.1004374

Dill KA 1990. Dominant forces in protein folding. Biochemistry 29: 7133-7155. doi: 10.1021/bi00483a001

Domazet-Lošo T, Brajković J, Tautz D 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet 23: 533-539. doi: 10.1016/j.tig.2007.08.014

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA 102: 14338-14343. doi: 10.1073/pnas.0504070102

Drummond DA, Wilke CO 2008. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell 134: 341-352.

Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 22: 1302-1306. doi: 10.1038/nbt1012

Gaucher EA, Govindarajan S, Ganesh OK 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451: 704-707. doi: 10.1038/nature06510

Godoy-Ruiz R, Perez-Jimenez R, Ibarra-Molero B, Sanchez-Ruiz JM 2004. Relation Between Protein Stability, Evolution and Structure, as Probed by Carboxylic Acid Mutations. J Mol Biol 336: 313-318. doi: 10.1016/j.jmb.2003.12.048

Gunasekaran K, Tsai C-J, Nussinov R 2004. Analysis of Ordered and Disordered Protein Complexes Reveals Structural Features Discriminating Between Stable and Unstable Monomers. J Mol Biol 341: 1327-1341. doi: 10.1016/j.jmb.2004.07.002

Irbäck A, Peterson C, Potthast F 1996. Evidence for nonrandom hydrophobicity structures in protein chains. Proc Natl Acad Sci USA 93: 9533-9538.

Irbäck A, Sandelin E 2000. On Hydrophobicity Correlations in Protein Chains. Biophysical Journal 79: 2252-2258. doi: 10.1016/S0006-3495(00)76472-1

Kumar S, Stecher G, Suleski M, Hedges SB 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol 34: 1812-1819. doi: 10.1093/molbev/msx116

Lee Y, Zhou T, Tartaglia GG, Vendruscolo M, Wilke CO 2010. Translationally optimal codons associate with aggregation-prone sites in proteins. Proteomics 10: 4163-4171.

Lehmann M, Pasamontes L, Lassen SF, Wyss M 2000. The consensus concept for thermostability engineering of proteins. BBA-Protein Struct M 1543: 408-415. doi: 10.1016/S0167-4838(00)00238-7

Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L 2004. A Comparative Study of the Relationship Between Protein Structure and β-Aggregation in Globular and Intrinsically Disordered Proteins. J Mol Biol 342: 345-353.

Mannige RV, Brooks CL, Shakhnovich EI 2012. A Universal Trend among Proteomes Indicates an Oily Last Common Ancestor. PLoS Comput Biol 8: e1002839. doi: 10.1371/journal.pcbi.1002839

Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, Schymkowitz JW, Rousseau F 2010. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nature Methods 7: 237-242.

McLysaght A, Guerzoni D 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Phil Trans R Soc B 370: 20140332. doi: 10.1098/rstb.2014.0332

Monsellier E, Chiti F 2007. Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep 8: 737-742. doi: 10.1038/sj.embor.7401034

Monsellier E, Ramazzotti M, de Laureto PP, Tartaglia G-G, Taddei N, Fontana A, Vendruscolo M, Chiti F 2007. The Distribution of Residues in a Polypeptide Sequence Is a Determinant of Aggregation Optimized by Evolution. Biophysical Journal 93: 4382-4391. doi: 10.1529/biophysj.107.111336

Moyers BA, Zhang J 2016. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. Mol Biol Evol 33: 1245-1256. doi: 10.1093/molbev/msw008

Moyers BA, Zhang J 2017. Further Simulations and Analyses Demonstrate Open Problems of Phylostratigraphy. Genome Biology and Evolution 9: 1519-1527. doi: 10.1093/gbe/evx109

Moyers BA, Zhang J 2015. Phylostratigraphic Bias Creates Spurious Patterns of Genome Evolution. Mol Biol Evol 32: 258-267. doi: 10.1093/molbev/msu286

Palmieri N, Kosiol C, Schlötterer C 2014. The life cycle of *Drosophila* orphan genes. eLife 3: e01311. doi: 10.7554/eLife.01311

Patki AU, Hausrath AC, Cordes MHJ 2006. High Polar Content of Long Buried Blocks of Sequence in Protein Domains Suggests Selection Against Amyloidogenic Non-polar Sequences. J Mol Biol 362: 800-809.

Povolotskaya IS, Kondrashov FA 2010. Sequence space and the ongoing expansion of the protein universe. Nature 465: 922-926. doi: 10.1038/nature09105

Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F 2009. Protein sequences encode safeguards against aggregation. Hum Mutat 30: 431-437. doi: 10.1002/humu.20905

Rousseau F, Serrano L, Schymkowitz JWH 2006. How Evolutionary Pressure Against Protein Aggregation Shaped Chaperone Specificity. J Mol Biol 355: 1037-1047.

Sánchez IE, Tejero J, Gómez-Moreno C, Medina M, Serrano L 2006. Point Mutations in Protein Globular Domains: Contributions from Function, Stability and Misfolding. J Mol Biol 363: 422-432. doi: 10.1016/j.jmb.2006.08.020

Schwartz R, Istrail S, King J 2001. Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. Protein Science 10: 1023-1031. doi: 10.1110/ps.33201

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. Version 4.0.5.

Soding J, Lupas AN 2003. More than the sum of their parts: on the evolution of proteins from peptides. BioEssays 25: 837-846. doi: 10.1002/bies.10321

Stansfeld Phillip J, Goose Joseph E, Caffrey M, Carpenter Elisabeth P, Parker Joanne L, Newstead S, Sansom Mark SP 2015. MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes. Structure 23: 1350-1361. doi: 10.1016/j.str.2015.05.006

Steipe B, Schiller B, Plückthun A, Steinbacher S 1994. Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. J Mol Biol 240: 188-192. doi: 10.1006/jmbi.1994.1434

Tartaglia GG, Pechmann S, Dobson CM, Vendruscolo M 2007. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. Trends Biochem Sci 32: 204-206.

Tartaglia GG, Pellarin R, Cavalli A, Caflisch A 2005. Organism complexity anti-correlates with proteomic β-aggregation propensity. Protein Science 14: 2735-2740. doi: 10.1110/ps.051473805

Thangakani AM, Kumar S, Velmurugan D, Gromiha MSM 2012. How do thermophilic proteins resist aggregation? Proteins: Struct Funct Bioinf 80: 1003-1015. doi: 10.1002/prot.24002

Trudeau DL, Kaltenbach M, Tawfik DS 2016. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. Mol Biol Evol 33: 2633-2641. doi: 10.1093/molbev/msw138

Uversky VN, Gillespie JR, Fink AL 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41: 415-427. doi: 10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7 [pii]

Williams PD, Pollock DD, Blackburne BP, Goldstein RA 2006. Assessing the Accuracy of Ancestral Protein Reconstruction Methods. PLoS Comput Biol 2: e69. doi: 10.1371/journal.pcbi.0020069

Wilson BA, Foy SG, Neme R, Masel J 2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nature Ecology & Evolution 1: 0146. doi: 10.1038/s41559-017-0146

Zhu H, Sepulveda E, Hartmann MD, Kogenaru M, Ursinus A, Sulz E, Albrecht R, Coles M, Martin J, Lupas AN 2016. Origin of a folded repeat protein from an intrinsically disordered ancestor. eLife 5: e16761. doi: 10.7554/eLife.16761