HIGHLIGHTED ARTICLE
GENETICS | INVESTIGATION

A Shift in Aggregation Avoidance Strategy Marks a Long-Term Direction to Protein Evolution

Scott G. Foy,*[,1] Benjamin A. Wilson,* Jason Bertram,* Matthew H. J. Cordes,† and Joanna Masel*[,2]
*Department of Ecology and Evolutionary Biology, and †Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona 85721

ORCID IDs: 0000-0002-3992-2876 (B.A.W.); 0000-0001-5374-6912 (J.B.); 0000-0002-7398-2127 (J.M.)

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 avoid strongly hydrophobic amino acids, which is presumably the simplest solution to the aggregation problem. Here we show that young genes' few hydrophobic residues are clustered near one another along the primary sequence, presumably to assist folding. The higher aggregation risk created by the higher hydrophobicity of older genes 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 interspersed than if distributed randomly. This interspersion 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.

KEYWORDS phylostratigraphy; gene age; aggregation propensity; protein folding; protein misfolding

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 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 to be flanked by "gatekeeper" residues (Rousseau *et al.* 2006). Disease mutations are enriched for aggregation-promoting changes (Reumers *et al.* 2009; De Baets *et al.* 2015), 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. In the absence of a fossil record or a time machine, biases introduced during the inference of ancestral protein states (Williams *et al.* 2006; Trudeau *et al.* 2016) make it difficult to assess how past proteins systematically differed from their modern descendants. We have therefore developed

doi: https://doi.org/10.1534/genetics.118.301719
Manuscript received October 23, 2018; accepted for publication January 25, 2019; published Early Online January 28, 2019.
Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7597616.
[1] Present address: St. Jude Children's Research Hospital, Memphis, TN 38105.
[2] Corresponding author: Department of Ecology and Evolutionary Biology, University of Arizona, 1041 E Lowell St., Tucson, AZ 85721. E-mail: masel@email.arizona.edu

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 noncoding 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 2015, 2016, 2017). 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.

Methods

Mus 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 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 because of the danger that they were falsely annotated as protein-coding genes (McLysaght and Hurst 2016), 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 for mouse in Supplemental Material, Table S1 of Wilson *et al.* (2017). "Cellular Organisms" contains all mouse gene families that share homology with a prokaryote. Yeast gene family and phylostratum annotation is taken from Table S7 of Wilson *et al.* (2017).

For greater resolution at shorter timescales, we used the recently sequenced *M. pahari* genome (Thybert *et al.* 2018) to compile a younger phylostratum, using Ensembl's orthology annotation (Herrero *et al.* 2016) to find homologs in *M. musculus*. Of the 789 putative proteins excluded in Wilson *et al.* (2017) as being unique to *M. musculus*, 155 also had homologs in *M. pahari*. Nine of these also had Ensembl ortholog assignments among members of older gene families and were excluded. BLASTp detected only one pair hitting each other among the genes with *E*-value < 0.001; these were placed together while each of the others was placed in its own gene family, collectively forming the youngest phylostratum to be analyzed. Note also that Ensembl ortholog annotation is not as rigorous a filter to remove false positives as the rat *vs.* mouse dN/dS measures used by Wilson *et al.* (2017) for older phylostrata. We therefore do not expect this youngest Mus phylostratum to be entirely free of false positives. This likely explains why its hydrophobicity metrics are lower than those of *Rattus*. The fact that hydrophobicity is still significantly elevated above that of controls (especially as

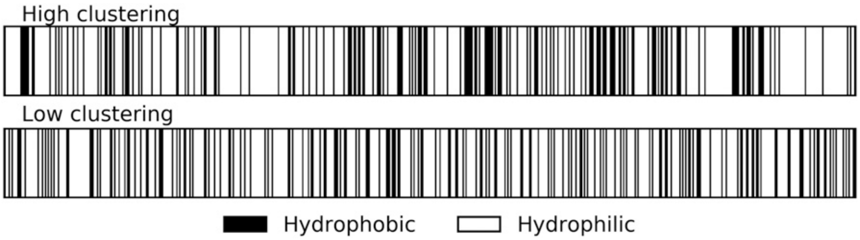

Figure 1 Illustration of the distribution of hydrophobic residues along the primary sequence of proteins with high *vs.* low clustering, of similar lengths and net hydrophobicities. The high clustering gene Fzd5 has length 585 amino acids, 31.5% hydrophobicity, and clustering of 1.58. The low clustering gene Farsb has length 589 amino acids, 31.6% hydrophobicity, and clustering of 0.69.

measured by ISD and by predicted aggregation propensity of scrambled sequences) suggests that the problem of contamination with sequences that are not protein-coding genes is not so profound as to exclude the phylostratum. However, it should be interpreted with caution.

Intergenic control sequences were also taken from previous work (Wilson *et al.* 2017). Briefly, one intergenic control sequence per gene was taken 100 nt downstream from 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 100 nt further downstream. This choice of location ensures that control sequences are representative of genomic regions in which protein-coding genes are found. One version of the control sequences used all intergenic sequences for this procedure a second used only RepeatMasked (Smit *et al.* 2015) intergenic sequences.

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. TANGO scores were Box–Cox transformed ($\lambda = 0.362$, optimized using only coding genes not controls, Q-Q plots shown in Figure S6A, B). Box–Cox λ values were determined using maximum-likelihood estimation (Box and Cox 1964) as implemented in geoR (https://CRAN.R-project.org/package=geoR). Central tendency estimates and confidence intervals derived from these models 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.

"Clustering" was assessed as a normalized index of dispersion, *i.e.*, by comparing the variance in hydrophobicity between blocks of consecutive amino acids to the mean hydrophobicity (Irbäck *et al.* 1996). Examples of high and low clustering are shown in Figure 1. We used $s = 6$, with different values of s yielding qualitatively similar results. Where the amino acid length was not divisible by six, a few amino acids were neglected at one or both ends, yielding a truncated length of N, and we used the average clustering measure ψ across different phases for the blocking procedure. We averaged over all phases using the maximum number of blocks, *e.g.*, only one phase for values of N divisible by 6. Results when we average over all six phases are very similar. Following past practice, we transformed amino acid sequences into binary hydrophobicity strings by taking the six amino acids FLIMVW as hydrophobic (+1) and scoring all the other amino acids as −1. We summed hydrophobicity scores to a value σ_k for each block $k = 1, \ldots, N/s$ and $M = \sum_{k=1}^{N/s} \sigma_k$ overall (Irbäck and Sandelin 2000). Our clustering score is a 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 clustering at the nucleotide level, blocks of length $s = 18$ rather than 6 were used. Nucleotide clustering values 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 clustering values ψ were Box–Cox transformed ($\lambda = -0.29$ for mouse, $\lambda = -0.008$ for yeast) prior to use in linear models, with the mouse Q-Q plot shown in Figure S6C,D.

To generate a scrambled control sequence that is paired to each gene, we simply sampled its amino acids without replacement. To generate clustering-controlled scrambled sequences, 1000 scrambled sequences of each protein were produced, and the one that most closely matched the clustering value 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. The mean value of each property was used across 50 scrambled sequences, but this led only to ~20% reductions in confidence interval width relative to using a single scrambled control. Because generating well-matched clustering-controlled scrambled sequences is computationally expensive, we used only a single matched-clustering scrambled control sequence per gene.

A protein was designated as transmembrane if TMHMM (Sonnhammer *et al.* 1998; Krogh *et al.* 2001) version 2.0c predicted that >18 of its amino acids lay within transmembrane helices.

Data availability

Source data for the statistical analyses and figures are provided in Tables S1–S6, available at Figshare and captioned in the main Supplemental Materials file. Code associated with generating and analyzing these tables is publicly available at https://github.com/MaselLab. Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7597616.

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 (Figure 2), 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 (Figure 3, top), 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 [Figure 3, middle panels on the fraction of hydrophobic residues and on ISD, the latter previously reported by Wilson *et al.* (2017)], with an increase in hydrophobicity taking place over ~200–400 MY. 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. Analyses focused on a set of ancestral reconstructed sites also find a trend of recently increasing hydrophobicity in drosophilid genomes (Yampolsky and Bouzinier 2010) that is ongoing even for ancient gene families (Yampolsky *et al.* 2017), although these data are subject to the bias of observing slightly deleterious substitutions more often than the reverse (Hurst *et al.* 2006; McDonald 2006).

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 (Figure 4, orange confidence intervals for genes shared with prokaryotes lie below 0). But surprisingly, the opposite was true for young genes (Figure 4, orange values 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 (Gunasekaran *et al.* 2004; Davey *et al.* 2012). 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 (Figure 5), and that this accounts for much of the excess aggregation propensity of young genes relative to scrambled controls (Figure 4 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 <1 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 (Figure S1). Similar results, albeit not extending quite as far back in time, are found using the normalized mean length of runs of hydrophobic amino acids FLIMVW (Figure S2) as by using the more sophisticated published metric of the degree to which these amino acids are

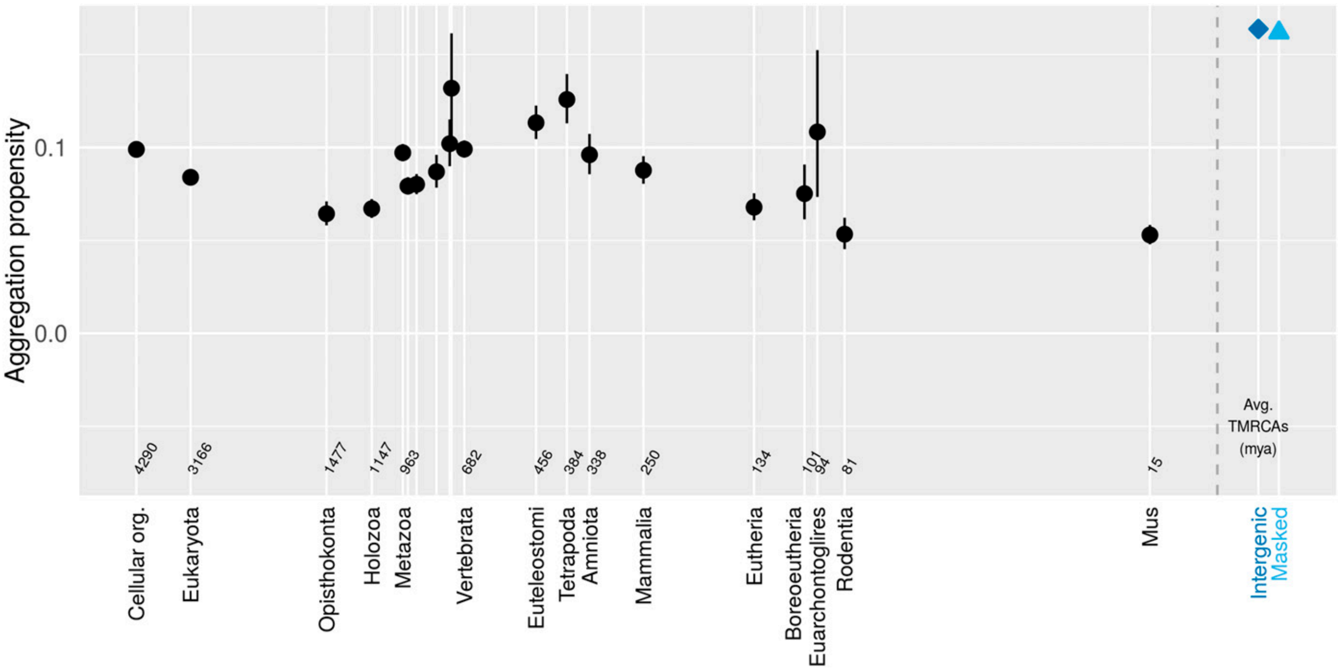

Figure 2 Mouse genes show little pattern in aggregation propensity (assessed via TANGO) as a function of age. Genes (black) show less aggregation propensity than intergenic controls (blue). Back-transformed central tendency estimates ± 1 SE come from a linear mixed model applied to transformed data, 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) for most phylostrata were taken from TimeTree.org (Kumar *et al.* 2017) on February 18, 2016 and that for *M. pahari* was taken May 7, 2018. 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). Cellular organism age is shown as the midpoint of the last universal common ancestor and the last eukaryotic common ancestor. Taxon names, some of which are omitted for space reasons, follow the sequence Metazoa, Eumetazoa, Bilateria, Deuterostomia, Chordata, Olfactores, Vertebrata, Euteleostomi, Tetrapoda, Amniota, Mammalia, Eutheria, Boreoeutheria, Euarchontoglires, Rodentia, Mus. The gray dashed line shows the 0 time, with control sequences to the right of it.

clustered (Irbäck *et al.* 1996; Irbäck and Sandelin 2000) shown in Figure 5.

We investigated whether the difference might be explained by differences in the frequencies of transmembrane proteins as a function of gene age. Given limited experimental annotation of transmembrane status, we used TMHMM (Sonnhammer *et al.* 1998; Krogh *et al.* 2001) to predict transmembrane status on the basis of protein sequence. Predicted transmembrane sequences had higher clustering (effect size of 0.16 in transformed space corresponds for example to clustering values of 1.18 *vs.* 1 as a function of transmembrane status in the linear model, different with $P < 0.0001$). But correcting for this slightly strengthened rather than weakened the trend in clustering (Figure S1).

We checked whether this trend in clustering is also found in the proteins of *Saccharomyces cerevisiae* (Figure S3), which is the other species for which homologous gene family annotation was combined with gene age annotation (Wilson *et al.* 2017). The very youngest 499 putative gene families (unique to *S. cerevisiae*, and which might therefore contain noncoding sequences annotated in error, although to minimize this problem, genes annotated as "dubious" are excluded) had a clustering value of 1.035 (66% C.I. 1.024–1.047; central tendency and C.I. back-transformed from the central tendency estimate ± 1 SE derived from a linear model with gene family as a random effect). The oldest 1966 gene families (with homologs in prokaryotes) had clustering 0.890 (66% C.I. 0.886–0.895), even lower than clustering of 0.943 (66% C.I. 0.939–0.946) found in mouse gene families of the same age. Among the 2467 gene families allocated to eight phylostrata of intermediate age, we found no significant differences among the phylostrata ($P = 0.6$, likelihood ratio test of linear model with gene family and random effect and phylostratum as putative fixed effect), which range from genes shared only with *S. paradoxus* to genes shared with distantly related eukaryotes. The clustering in all these phylostrata was lower than we expected from our mouse results, at 0.951 (66% C.I. 0.945–0.958). These results, shown in Figure S3, suggest that low clustering evolves far more rapidly, at least in the earlier stages, in unicellular yeast with short generation times and large population sizes than it does in the ancestral lineage of mice. However, just as for the mouse lineage, saturation is not reached for gene families dating back "only" to an early eukaryote; genes with prokaryotic homologs have even lower clustering values than those with homologs in distantly related eukaryotes but not prokaryotes.

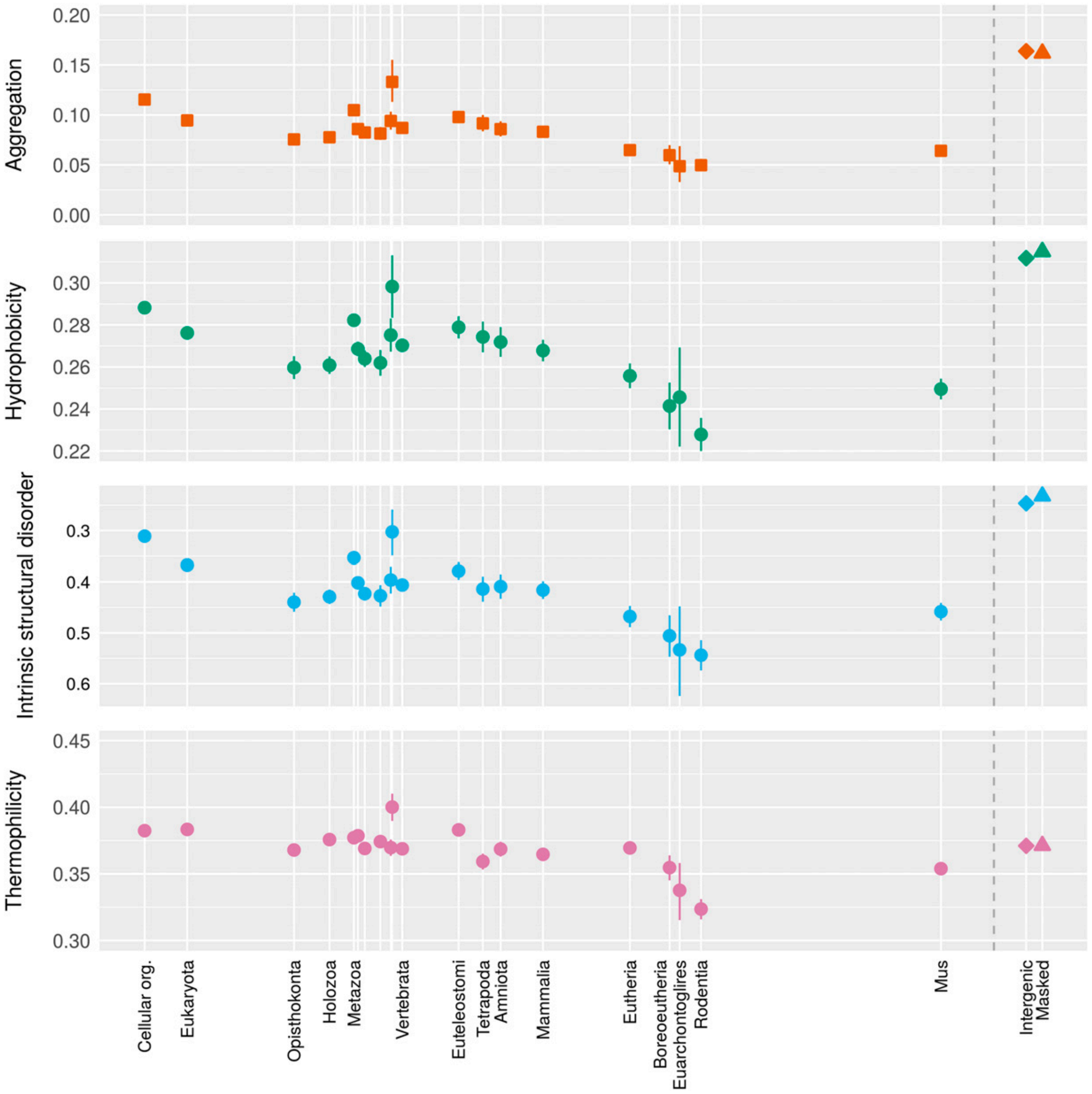

Figure 3 Four different measures for the hydrophobicity of the amino acid content as a function of gene family age. "Aggregation" represents the average TANGO results from 50 scrambled versions of each gene, and hence captures the effect of amino acid composition on TANGO's estimate of β-aggregation propensity. 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. Hydrophobicity gives the fraction of amino acids that are FLIMVW. The "oiliness" measurement of Mannige *et al.* (2012), namely content of FLIV, is similar. Intrinsic structural disorder scores are as previously reported in Wilson *et al.* (2017), shown here for more phylostrata, and inverted for easier comparison with other metrics. Thermophilicity represents the content of ILVYWRE, as analyzed by Boussau *et al.* (2008), subjected to a Box–Cox transform with $\lambda = 2.412$ prior to model fitting; thermophilicity is dominated by the same general hydrophobicity trend as the other measures. 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 ± 1 SE come from a linear mixed model, where gene family and phylostratum are random and fixed terms, respectively; $\lambda = 0.93$ is used for hydrophobicity, other transforms are described in the *Methods*. The *x*-axis is the same as for Figure 2.

Clustering is a metric for which genes that have been evolving for longer have different properties from genes that are "less evolved." There must either be a long-term trend in the clustering values of newborn genes as a function of the time at which they are born, or else there has been a long-term direction to evolution over billions of years. We consider the latter possibility more plausible than the former. 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 avoid the most hydrophobic amino acids (Figure 3), creating ISD (Linding *et al.* 2004; Thangakani *et al.* 2012; Banerjee and Chakraborty 2017;

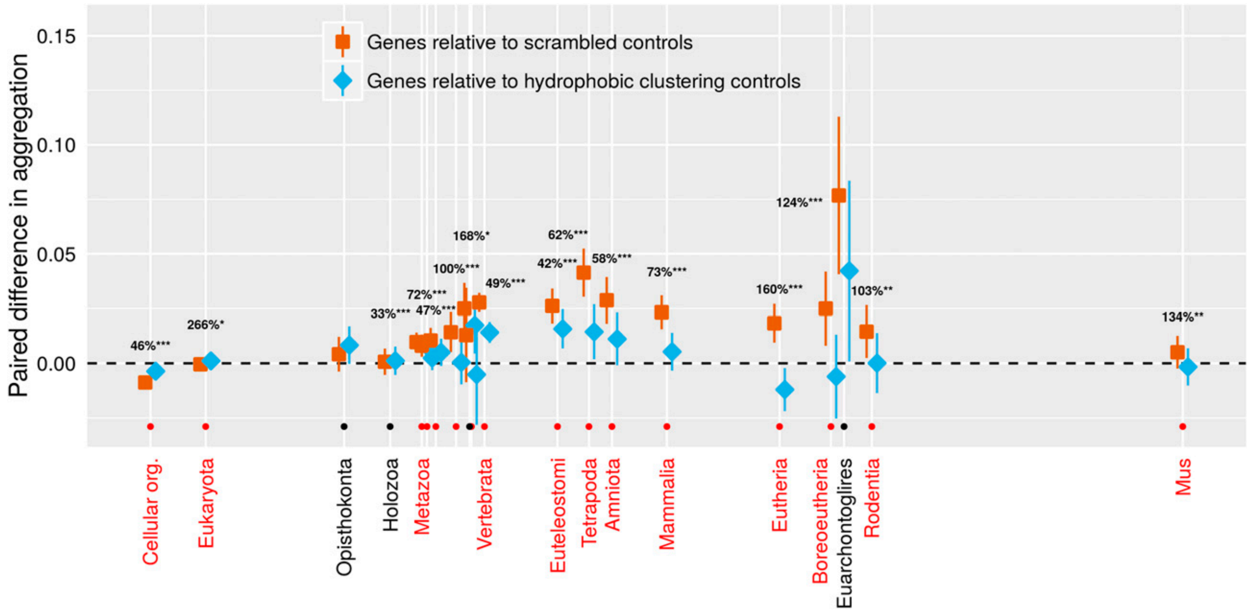

Figure 4 Only very old genes have aggregation propensities lower than that expected from their amino acid composition alone (orange < dashed line expectation of 0). This puzzling finding is reduced when we account for clustering (blue is closer than orange is to the 0 dashed line) using a scrambled sequence that is controlled to have a similar clustering 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 nonsignificant (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 (Figure S5). Orange values come from the mean of 50 scrambled sequences per gene, blue from a single scrambled sequence with a closely matched clustering value. The *x*-axis is the same as for Figure 2.

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 (Figure 5, right), while still keeping total hydrophobicity and hence aggregation propensity within tolerable limits (Figure 2 and Figure 3). 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 (Söding 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 (Figure S4).

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. Under the assumption that amino acid composition at birth does not vary systematically as a function of the time of birth, we could conclude that changing the amino acid composition of a protein takes ~200–400 MY (Figure 3). In contrast, changing the index of dispersion might require 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 two youngest phylostrata, the Mus phylostratum of *M. musculus* genes shared only with *M. pahari*, and the Rattus phylostratum of *M. musculus* genes shared with rats, show 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 (Figure 5, blue). Clustering of amino acids translated from noncoding intergenic sequences is a direct consequence of the clustering of nucleotides; indices of dispersion at the

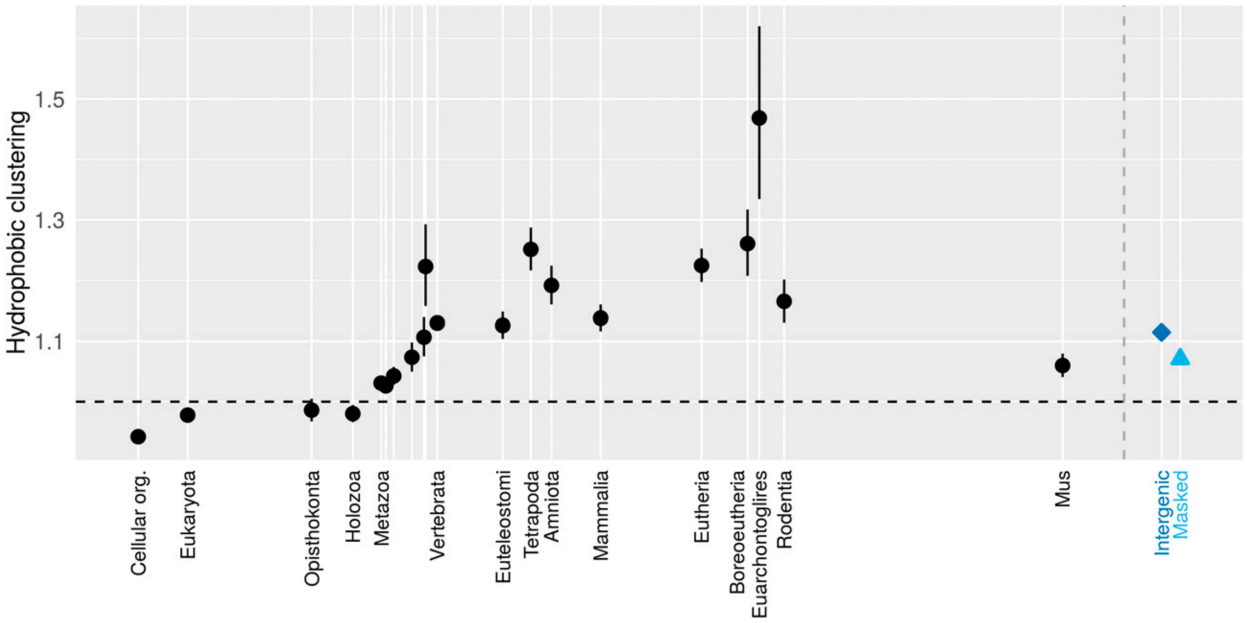

Figure 5 Clustering initially follows that of its raw material, and evolves rapidly upward at first, but then decays downward extremely slowly, indicating a long-term direction of evolution. Only the oldest genes have hydrophobic amino acids spread out from each other, as previously reported; young genes have clustered hydrophobic amino acids. Back-transformed central tendency estimates ± 1 SE 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 2.

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 >1 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, or (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 (Figure S1) 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 have fewer strongly hydrophobic amino acids: they were born at more permissive lower temperatures. However, ancestral reconstruction techniques are likely biased toward consensus amino acids that are fold-stabilizing (Steipe *et al.* 1994; Lehmann *et al.* 2000; Godoy-Ruiz *et al.* 2004; Bloom and Glassman 2009) and hence may be more hydrophobic (Williams *et al.* 2006; Trudeau *et al.* 2016). Alarmingly, ancestral reconstruction also suggests that the ancestral mammal was a thermophile (Trudeau *et al.* 2016), although drosopholid reconstructions are compatible with a trend in the opposite direction to reconstruction bias, toward greater hydrophobicity with time (Yampolsky and Bouzinier 2010; Yampolsky *et al.* 2017).

The main trend that we see of hydrophobicity/thermophilicity as a function of gene age is on shorter timescales; for older gene families, billions of years of common evolution has erased the differences in starting points. It is the subtler signal of hydrophobic amino acid interspersion that shows the long-term pattern in our analysis. 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 may arise from differences in which proteins are present in which species, which could be a confounding factor when Mannige *et al.* attributed proteome-wide trends to descent with modification. Mannige *et al.* 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 is not just hard for human biochemists – it is 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*.

Acknowledgments

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. This work was supported by the John Templeton Foundation (39667, 60814), and the National Institutes of Health (GM-104040). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions: JM conceived the general approach. MHJC conceived the clustering metric, SLiM, and transmembrane protein analyses. JB produced Figure 1 and Figure S2, BAW fitted statistical models, and the other final figures. SGF conducted all other upstream data analysis, and JM wrote the paper. The authors declare no competing financial interests.

Literature Cited

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

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Banerjee, S., and S. Chakraborty, 2017 Protein intrinsic disorder negatively associates with gene age in different eukaryotic lineages. Mol. Biosyst. 13: 2044–2055. https://doi.org/10.1039/C7MB00230K

Bloom, J. D., and M. J. Glassman, 2009 Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. PLoS Comput. Biol. 5: e1000349. https://doi.org/10.1371/journal.pcbi.1000349

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

Box, G. E. P., and D. R. Cox, 1964 An analysis of transformations. J. R. Stat. Soc. B 26: 211–252.

Broome, B. M., and M. H. Hecht, 2000 Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. J. Mol. Biol. 296: 961–968. https://doi.org/10.1006/jmbi.2000.3514

Buck, P. M., S. Kumar, and S. K. Singh, 2013 On the role of aggregation prone regions in protein evolution, stability, and enzymatic catalysis: insights from diverse analyses. PLoS Comput. Biol. 9: e1003291. https://doi.org/10.1371/journal.pcbi.1003291

Chen, Y., and N. V. Dokholyan, 2008 Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. Mol. Biol. Evol. 25: 1530–1533. https://doi.org/10.1093/molbev/msn122

Davey, N. E., K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar *et al.*, 2012 Attributes of short linear motifs. Mol. Biosyst. 8: 268–281. https://doi.org/10.1039/C1MB05231D

De Baets, G., J. Reumers, J. Delgado Blanco, J. Dopazo, J. Schymkowitz *et al.*, 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. https://doi.org/10.1371/journal.pcbi.1002090

De Baets, G., L. Van Doorn, F. Rousseau, and J. Schymkowitz, 2015 Increased aggregation is more frequently associated to human disease-associated mutations than to neutral polymorphisms. PLoS Comput. Biol. 11: e1004374. https://doi.org/10.1371/journal.pcbi.1004374

Dill, K. A., 1990 Dominant forces in protein folding. Biochemistry 29: 7133–7155. https://doi.org/10.1021/bi00483a001

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

Drummond, D. A., and C. O. Wilke, 2008 Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352. https://doi.org/10.1016/j.cell.2008.05.042

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold, 2005 Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA 102: 14338–14343. https://doi.org/10.1073/pnas.0504070102

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

Gaucher, E. A., S. Govindarajan, and O. K. Ganesh, 2008 Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451: 704–707. https://doi.org/10.1038/nature06510

Godoy-Ruiz, R., R. Perez-Jimenez, B. Ibarra-Molero, and J. M. Sanchez-Ruiz, 2004 Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations. J. Mol. Biol. 336: 313–318. https://doi.org/10.1016/j.jmb.2003.12.048

Gunasekaran, K., C.-J. Tsai, and R. Nussinov, 2004 Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. J. Mol. Biol. 341: 1327–1341. https://doi.org/10.1016/j.jmb.2004.07.002

Herrero, J., M. Muffato, K. Beal, S. Fitzgerald, L. Gordon *et al.*, 2016 Ensembl comparative genomics resources. Database (Oxford) 2016: bav096. https://doi.org/10.1093/database/bav096

Hurst, L. D., E. J. Feil, and E. P. C. Rocha, 2006 Causes of trends in amino-acid gain and loss. Nature 442: E11–E12. https://doi.org/10.1038/nature05137

Irbäck, A., and E. Sandelin, 2000 On hydrophobicity correlations in protein chains. Biophys. J. 79: 2252–2258. https://doi.org/10.1016/S0006-3495(00)76472-1

Irbäck, A., C. Peterson, and F. Potthast, 1996 Evidence for nonrandom hydrophobicity structures in protein chains. Proc. Natl. Acad. Sci. USA 93: 9533–9538. https://doi.org/10.1073/pnas.93.18.9533

Krogh, A., B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, 2001 Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305: 567–580. https://doi.org/10.1006/jmbi.2000.4315

Kumar, S., G. Stecher, M. Suleski, and S. B. Hedges, 2017 TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. 34: 1812–1819. https://doi.org/10.1093/molbev/msx116

Lee, Y., T. Zhou, G. G. Tartaglia, M. Vendruscolo, and C. O. Wilke, 2010 Translationally optimal codons associate with aggregation-prone sites in proteins. Proteomics 10: 4163–4171. https://doi.org/10.1002/pmic.201000229

Lehmann, M., L. Pasamontes, S. F. Lassen, and M. Wyss, 2000 The consensus concept for thermostability engineering of proteins. Biochim. Biophys. Acta. 1543: 408–415. https://doi.org/10.1016/S0167-4838(00)00238-7

Linding, R., J. Schymkowitz, F. Rousseau, F. Diella, and L. Serrano, 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. https://doi.org/10.1016/j.jmb.2004.06.088

Mannige, R. V., C. L. Brooks, and E. I. Shakhnovich, 2012 A universal trend among proteomes indicates an oily last common ancestor. PLoS Comput. Biol. 8: e1002839. https://doi.org/10.1371/journal.pcbi.1002839

Maurer-Stroh, S., M. Debulpaep, N. Kuemmerer, M. Lopez de la Paz, I. C. Martins *et al.*, 2010 Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat. Methods 7: 237–242. https://doi.org/10.1038/nmeth.1432

McDonald, J. H., 2006 Apparent trends of amino acid gain and loss in protein evolution due to nearly neutral variation. Mol. Biol. Evol. 23: 240–244. https://doi.org/10.1093/molbev/msj026

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

McLysaght, A., and L. D. Hurst, 2016 Open questions in the study of de novo genes: what, how and why. Nat. Rev. Genet. 17: 567–578. https://doi.org/10.1038/nrg.2016.78

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

Monsellier, E., M. Ramazzotti, P. P. de Laureto, G.-G. Tartaglia, N. Taddei *et al.*, 2007 The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution. Biophys. J. 93: 4382–4391. https://doi.org/10.1529/biophysj.107.111336

Moyers, B. A., and J. Zhang, 2015 Phylostratigraphic bias creates spurious patterns of genome evolution. Mol. Biol. Evol. 32: 258–267 [corrigenda: Mol. Biol. Evol. 33: 3031 (2016)]. https://doi.org/10.1093/molbev/msu286

Moyers, B. A., and J. Zhang, 2016 Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. Mol. Biol. Evol. 33: 1245–1256. https://doi.org/10.1093/molbev/msw008

Moyers, B. A., and J. Zhang, 2017 Further simulations and analyses demonstrate open problems of phylostratigraphy. Genome Biol. Evol. 9: 1519–1527. https://doi.org/10.1093/gbe/evx109

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

Patki, A. U., A. C. Hausrath, and M. H. J. Cordes, 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. https://doi.org/10.1016/j.jmb.2006.07.055

Povolotskaya, I. S., and F. A. Kondrashov, 2010 Sequence space and the ongoing expansion of the protein universe. Nature 465: 922–926. https://doi.org/10.1038/nature09105

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

Rousseau, F., L. Serrano, and J. W. H. Schymkowitz, 2006 How evolutionary pressure against protein aggregation shaped chaperone specificity. J. Mol. Biol. 355: 1037–1047. https://doi.org/10.1016/j.jmb.2005.11.035

Sánchez, I. E., J. Tejero, C. Gómez-Moreno, M. Medina, and L. Serrano, 2006 Point mutations in protein globular domains: contributions from function, stability and misfolding. J. Mol. Biol. 363: 422–432. https://doi.org/10.1016/j.jmb.2006.08.020

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

Smit, A., R. Hubley, and P. Green, 2015 RepeatMasker open-4.0 version 4.0.5. Available at: http://www.repeatmasker.org.

Söding, J., and A. N. Lupas, 2003 More than the sum of their parts: on the evolution of proteins from peptides. BioEssays 25: 837–846. https://doi.org/10.1002/bies.10321

Sonnhammer, E. L., G. von Heijne, and A. Krogh, 1998 A hidden Markov model for predicting transmembrane helices in protein sequences, pp. 175–182 in *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, edited by J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff *et al.* AAAI Press, Menlo Park, CA.

Steipe, B., B. Schiller, A. Plückthun, and S. Steinbacher, 1994 Sequence statistics reliably predict stabilizing mutations in a protein domain. J. Mol. Biol. 240: 188–192. https://doi.org/10.1006/jmbi.1994.1434

Tartaglia, G. G., R. Pellarin, A. Cavalli, and A. Caflisch, 2005 Organism complexity anti-correlates with proteomic β-aggregation propensity. Protein Sci. 14: 2735–2740. https://doi.org/10.1110/ps.051473805

Tartaglia, G. G., S. Pechmann, C. M. Dobson, and M. Vendruscolo, 2007 Life on the edge: a link between gene expression levels and aggregation rates of human proteins. Trends Biochem. Sci. 32: 204–206. https://doi.org/10.1016/j.tibs.2007.03.005

Thangakani, A. M., S. Kumar, D. Velmurugan, and M. S. M. Gromiha, 2012 How do thermophilic proteins resist aggregation? Proteins: Struct. Funct. Bioinf. 80: 1003–1015. https://doi.org/10.1002/prot.24002

Thybert, D., M. Roller, F. C. P. Navarro, I. Fiddes, I. Streeter *et al.*, 2018 Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes. Genome Res. 28: 448–459. https://doi.org/10.1101/gr.234096.117

Trudeau, D. L., M. Kaltenbach, and D. S. Tawfik, 2016 On the potential origins of the high stability of reconstructed ancestral proteins. Mol. Biol. Evol. 33: 2633–2641. https://doi.org/10.1093/molbev/msw138

Uversky, V. N., J. R. Gillespie, and A. L. Fink, 2000 Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41: 415–427. https://doi.org/10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7

Williams, P. D., D. D. Pollock, B. P. Blackburne, and R. A. Goldstein, 2006 Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput. Biol. 2: e69. https://doi.org/10.1371/journal.pcbi.0020069

Wilson, B. A., S. G. Foy, R. Neme, and J. Masel, 2017 Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nat. Ecol. Evol. 1: 0146. https://doi.org/10.1038/s41559-017-0146

Yampolsky, L. Y., and M. A. Bouzinier, 2010 Evolutionary patterns of amino acid substitutions in 12 Drosophila genomes. BMC Genomics 11: S10. https://doi.org/10.1186/1471-2164-11-S4-S10

Yampolsky, L. Y., Y. I. Wolf, and M. A. Bouzinier, 2017 Net evolutionary loss of residue polarity in Drosophilid protein cores indicates ongoing optimization of amino acid composition. Genome Biol. Evol. 9: 2879–2892. https://doi.org/10.1093/gbe/evx191

Zhu, H., E. Sepulveda, M. D. Hartmann, M. Kogenaru, A. Ursinus *et al.*, 2016 Origin of a folded repeat protein from an intrinsically disordered ancestor. eLife 5:e16761. https://doi.org/10.7554/eLife.16761

Communicating editor: C. Jones