A shift in aggregation avoidance strategy marks a long-term direction to protein evolution

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

Affiliations:

[1]Department of Ecology and Evolutionary Biology, University of Arizona, 1041 E Lowell St Tucson AZ 85721 USA.

[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 direction to 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 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.

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 (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. Absent 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 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

61 and Guerzoni 2015). Different gene families (i.e. sets of homologous genes) therefore have
62 different ages, and the properties of a gene can be a function of age.

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

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

83 Most critiques of the interpretation of phylostratigraphy in de novo gene terms therefore focus
84 on the fourth possibility, specifically the concern that trends may be driven by biases in the
85 degree to which homology is detectable (Albà and Castresana 2007; Moyers and Zhang 2015;
86 Moyers and Zhang 2016; Moyers and Zhang 2017). In particular, homology is harder to detect
87 for shorter and faster-evolving proteins, which might therefore appear to be young, giving false
88 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, 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

117 composition (Fig. 2, middle panels on the fraction of hydrophobic residues and on intrinsic
118 structural disorder, the latter previously reported by Wilson *et al.* (2017)), with an increase in
119 hydrophobicity taking place over ~200-400 million years. Previously reported differences in the
120 aggregation propensity (Tartaglia *et al.* 2005) and hydrophobicity (Mannige *et al.* 2012) of
121 proteomes from different organisms might therefore be accounted for by systematic variation
122 among species in the composition of old vs. young genes; in our analysis, all proteins were
123 taken from the same mouse species, removing this confounding factor. Analyses focused on a
124 set of ancestral reconstructed sites also find a trend of recently increasing hydrophobicity in
125 Drosophilid genomes (Yampolsky and Bouzinier 2010) that is ongoing even for ancient gene
126 families (Yampolsky *et al.* 2017), although this data is subject to the bias of observing slightly
127 deleterious substitutions more often than the reverse (Hurst *et al.* 2006; McDonald 2006).

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

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

145 and those that increase the stability of the aggregated or otherwise misfolded form (Sánchez *et*
146 *al.* 2006).

147 We find that young genes do show hydrophobic clustering, while very old genes show
148 interspersion of hydrophobic amino acid residues (Fig. 4), and that this accounts for much of
149 the excess aggregation propensity of young genes relative to scrambled controls (Fig. 3 blue
150 points are closer to zero than orange points). Previous reports have suggested that the danger
151 of aggregation selects against hydrophobic clustering (Monsellier *et al.* 2007). In other words,
152 among consecutive blocks of amino acids, the variance in hydrophobicity is lower than the
153 mean, i.e. the index of dispersion is less than one in proteins overall (Irbäck *et al.* 1996;
154 Schwartz *et al.* 2001) and in the core of protein folds (Patki *et al.* 2006). In the present analysis,
155 this holds true only for old, highly evolved proteins. Younger proteins not only appear less
156 evolutionarily constrained to intersperse polar and hydrophobic residues, but to the contrary,
157 their hydrophobic residues show excess concentration near one another along the sequence,
158 increasing aggregation propensity. Our results are extremely robust when we control for
159 protein length, evolutionary rate, and expression level (Fig. S1). Similar results, albeit not
160 extending quite as far back in time, are found using the normalized mean length of runs of
161 hydrophobic amino acid lengths FLIMVW (Fig. S2) as by using the more sophisticated published
162 metric of the degree to which these amino acids are clustered (Irbäck *et al.* 1996; Irbäck and
163 Sandelin 2000) shown in Fig. 4.

164 We investigated whether the difference might be explained by differences in the frequencies of
165 transmembrane proteins as a function of gene age. Unfortunately, sequence-based prediction
166 of transmembrane status is likely directly confounded with clustering, and only 137 mouse
167 proteins have been experimentally verified as transmembrane. The increased clustering of
168 transmembrane proteins was barely significant as a fixed effect within our linear model
169 (p=0.042). Because annotations are available for so few genes, we do not know whether
170 transmembrane proteins are more likely to be old or young, and hence whether
171 transmembrane status helps explain the trend or make it still more puzzling. The weak effect
172 size suggests that transmembrane status might explain at best only a small portion of the trend.

173 We checked whether this trend in clustering is also found in the proteins of *Saccharomyces*
174 *cerevisiae* (Fig. S3), which is the other species for which homologous gene family annotation
175 was combined with gene age annotation (Wilson *et al.* 2017). The very youngest 499 putative
176 gene families (unique to *S. cerevisiae*, and which might therefore contain non-coding sequences
177 annotated in error, although to minimize this problem, genes annotated as "dubious" are
178 excluded) had a clustering value of 1.035 (66% CI 1.024-1.047; central tendency and CI
179 backtransformed from the central tendency estimate +/- one standard error derived from a
180 linear model with and gene family as a random effect). The oldest 1,966 gene families (with
181 homologs in prokaryotes) had clustering 0.890 (66% CI 0.886-0.895), even lower than clustering
182 of 0.943 (66% CI 0.939-0.946) found in mouse gene families of the same age. Among the 2,467
183 gene families allocated to eight phylostrata of intermediate age, we found no significant
184 differences among the phylostrata (p=0.6, likelihood ratio test of linear model with gene family
185 and random effect and phylostratum as putative fixed effect), which range from genes shared
186 only with *S. paradoxus* to genes shared with distantly related eukaryotes. The clustering in all
187 these phylostrata was lower than we expected from our mouse results, at 0.951 (66% CI 0.945-
188 0.958). These results, shown in Figure S3, suggest that low clustering evolves far more rapidly,
189 at least in the earlier stages, in unicellular yeast with short generation times and large
190 population sizes than it does in the ancestral lineage of mice. However just as for the mouse
191 lineage, saturation is not reached for gene families dating back "only" to an early eukaryote;
192 genes with prokaryotic homologs have even lower clustering values than those with homologs
193 in distantly related eukaryotes but not prokaryotes.

194 Clustering is a metric for which genes that have been evolving for longer have different
195 properties from genes that are "less evolved". There must either be a long-term trend in the
196 clustering values of newborn genes as a function of the time at which they are born, or else
197 there has been a long-term direction to evolution over billions of years. We consider the latter
198 possibility more plausible than the former. This directionality of evolution can be interpreted as
199 a slow shift from a primitive strategy for avoiding misfolding in young genes to more subtle
200 strategies in old genes.

201 The primitive aggregation avoidance strategy used by young genes is simply to avoid the most
202 hydrophobic amino acids (Fig. 2), creating intrinsic structural disorder (Linding *et al.* 2004;
203 Thangakani *et al.* 2012; Banerjee and Chakraborty 2017; Wilson *et al.* 2017). Given such an
204 amino acid composition, young genes might form an early folding nucleus by concentrating
205 hydrophobic amino acids in localized regions of the sequence (Fig. 4, right), while still keeping
206 total hydrophobicity and hence aggregation propensity within tolerable limits (Figs. 1-2). Such a
207 folding nucleus would not necessarily be an entire independently folded domain. In particular,
208 some origin theories posit that ancient proteins first achieved folding by becoming structured
209 only upon binding to some interaction partner (Soding and Lupas 2003; Zhu *et al.* 2016). In
210 contemporary proteins, potential representatives of nascent structure are found in intrinsically
211 disordered proteins that contain peptide-length binding motifs (small linear interaction motifs;
212 SLiMs), many of which become ordered when bound to a partner (Davey *et al.* 2012). We do
213 not, however, find that young genes have more known SLiMs (Fig. S4).

214 In contrast to young genes, older genes have higher hydrophobicity, which must be offset by
215 the evolution of other aggregation-avoidance strategies (Thangakani *et al.* 2012). For such
216 changes to occur through descent with modification probably happens only slowly. Under the
217 assumption that amino acid composition at birth does not vary systematically as a function of
218 the time of birth, we could conclude that changing the amino acid composition of a protein
219 takes ~200-400 million years (Fig. 2). In contrast, changing the index of dispersion might require
220 such a large number of changes that it is extraordinarily slower, with a consistent direction to
221 evolution visible over the entire history of life back to our common ancestor with prokaryotes.

222 Note that our two youngest phylostrata, the Mus phylostratum of *Mus musculus* genes shared
223 only with *M. pahari*, and the Rattus phylostratum of *M. musculus* genes shared with rats, show
224 less clustering than other young genes, suggesting that rapid change in the index of dispersion
225 may be possible (in the other direction) after all, on short and recent timescales. However, very
226 young gene families are subject to significantly higher death rates than other gene families
227 (Palmieri *et al.* 2014). With gene family loss so common at first, it is possible that the rapid
228 initial increase in clustering is due to differential retention of gene families with highly clustered
229 amino acids. This interpretation of the data is consistent with explaining how slow the later fall

230 in clustering is, by positing that descent with modification is constrained to change clustering
231 values slowly.

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

246 **Discussion**

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

255 We believe that gene birth is the most plausible driver of our results. HGT is rare in more recent
256 ancestors of mice, simultaneous loss in so many branches is unlikely, and statistical correction
257 for evolutionary rate, length and expression (Fig. S1) has, in contradiction to the predictions of

258 homology detection bias, a negligible effect on our results. However, our results on the
259 evolution of protein properties following a defining event remain of interest under all scenarios
260 of what the gene-age-determining event is.

261 There are three ways to explain subsequent patterns as a function of gene family age. The two
262 mentioned so far are biases in retention after birth, and descent with modification. The third
263 possibility is that the conditions of life were significantly different at different times, and hence
264 so were the biochemical properties of proteins born/transferred/rapidly diverged at that time.
265 Specifically, ancestral sequence reconstruction techniques have been used to infer that
266 proteins in our ancestral lineage became progressively less thermophilic (Gaucher *et al.* 2008).
267 This might explain why young genes have fewer strongly hydrophobic amino acids; they were
268 born at more permissive lower temperatures. However, ancestral reconstruction techniques
269 are likely biased toward consensus amino acids that are fold-stabilizing (Steipe *et al.* 1994;
270 Lehmann *et al.* 2000; Godoy-Ruiz *et al.* 2004; Bloom and Glassman 2009) and hence may be
271 more hydrophobic (Williams *et al.* 2006; Trudeau *et al.* 2016). Alarmingly, ancestral
272 reconstruction also suggests that the ancestral mammal was a thermophile (Trudeau *et al.*
273 2016), although Drosopholid reconstructions are compatible with a trend in the opposite
274 direction to reconstruction bias, towards greater hydrophobicity with time (Yampolsky and
275 Bouzinier 2010; Yampolsky *et al.* 2017).

276 The main trend that we see of hydrophobicity/thermophilicity as a function of gene age is on
277 shorter timescales; for older gene families, billions of years of common evolution has erased
278 the differences in starting points. It is the more subtle signal of hydrophobic amino acid
279 interspersion that shows the long-term pattern in our analysis. However, variation in the
280 conditions of life at the time of gene origin remains a plausible explanation for the idiosyncratic
281 differences between phylostrata, i.e. for the remaining, statistically meaningful deviations of
282 individual phylostrata from the trends reported here.

283 We have already invoked differential retention as a possible driver of the short-term
284 evolutionary increase in the clustering values of young genes. It is logically possible that the
285 long-term trend in clustering values is also a result of differential retention; if gene families with

286 higher clustering values are more likely to be lost, different gene ages represent different spans
287 of time in which this loss has had an opportunity to occur. Given the billion year time scales and
288 thus enormous number of lost gene families this implies, this seems at present a less plausible
289 scenario than descent with modification for different durations following different dates of
290 origin. In other words, descent with modification seems the most plausible of the three possible
291 drivers of biochemical patterns as a function of gene age, independently of what exactly "gene
292 age" means.

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

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

311 The protein folding problem is notoriously hard. Here we see that it isn't just hard for human
312 biochemists – it's so hard that evolution struggles with it too. Proteins evolve to find stable
313 folds despite the correlated and ever-present danger of aggregation. They do so via a slow

314 exploration of an enormous sequence space, a search that has yet to saturate after billions of
315 years (Povolotskaya and Kondrashov 2010). Given the enormous space that has already been
316 searched, existing protein folds, especially of older gene families, may therefore be a highly
317 unrepresentative sample of the typical behaviors of polypeptide chains. Protein folds are best
318 thought of as a collection of corner cases and idiosyncratic exceptions, which are hard to find
319 even for evolution, let alone for our "free-modeling" techniques to predict ab initio.

320 **Methods**

321 *M. musculus* proteins from Ensembl (v73) were assigned gene families and gene ages as
322 described elsewhere (Wilson *et al.* 2017). To briefly outline this previous procedure, BLASTp
323 (Altschul *et al.* 1997) against the National Center for Biotechnology Information (NCBI) nr
324 database with an E-value threshold of 0.001 was used for preliminary age assignments for each
325 gene, followed by a variety of quality filters. Genes unique to one species were excluded
326 because of the danger that they were falsely annotated as protein-coding genes (McLysaght
327 and Hurst 2016), leaving Rodentia as the youngest phylostratum. Paralogous genes were
328 clustered into gene families, and a single age was reconciled per gene family, which filtered out
329 some inconsistent performance of BLASTp. Numbers of genes and gene families in each
330 phylostratum can be found for mouse in Table S1 of Wilson et al. (2017). "Cellular Organisms"
331 contains all mouse gene families that share homology with a prokaryote. Yeast gene family and
332 phylostratum annotation is taken from Table S7 of Wilson et al. (2017).

333 For greater resolution at shorter timescales, we used the recently sequenced *M. pahari* genome
334 (Thybert *et al.* 2018) to compile a younger phylostratum, using Ensembl's orthology annotation
335 (Herrero *et al.* 2016) to find homologs in *M. musculus*. Of the 789 putative proteins excluded in
336 Wilson et al. (2017) as being unique to *M. musculus*, 155 also had homologs in *M. pahari*. 9 of
337 these also had Ensembl ortholog assignments among members of older gene families, and were
338 excluded. BLASTp detected only one pair hitting each other among the genes with e-value <
339 0.001; these were placed together while each of the others was placed in its own gene family,
340 collectively forming the youngest phylostratum to be analyzed. Note also that Ensembl ortholog
341 annotation is not as rigorous a filter to remove false positives as the rat vs mouse dN/dS

342 measures used by Wilson et al. (2017) for older phylostrata. We therefore do not expect this
343 youngest Mus phylostratum to be entirely free of false positives. This likely explains why
344 its hydrophobicity metrics are lower than those of Rattus. The fact that hydrophobicity is still
345 significantly elevated above that of controls (especially as measured by ISD and by predicted
346 aggregation propensity of scrambled sequences) suggests that the problem of contamination
347 with sequences that are not protein-coding genes is not so profound as to exclude the
348 phylostratum. However, it should be interpreted with caution.

349 Intergenic control sequences were also taken from previous work (Wilson *et al.* 2017). Briefly,
350 one intergenic control sequence per gene was taken 100nt downstream from the end of the 3'
351 end of the transcript, with stop codons excised until a length match to the neighboring protein-
352 coding gene was obtained. A second control sequence per gene began 100nt further
353 downstream. This choice of location ensures that control sequences are representative of
354 genomic regions in which protein-coding genes are found. One version of the control sequences
355 used all intergenic sequences for this procedure, a second used only RepeatMasked (Smit *et al.*
356 2015) intergenic sequences.

357 Aggregation propensity was scored using TANGO (Fernandez-Escamilla *et al.* 2004) and Waltz
358 (Maurer-Stroh *et al.* 2010). We counted the number of amino acids contained within runs of at
359 least five consecutive amino acids scored to have >5% aggregation propensity, added 0.5, and
360 divided by protein length to obtain a measure of the density of aggregation-prone regions. For
361 those scores derived using TANGO, we then performed a Box-Cox transformation (λ=0.362,
362 optimized using only coding genes not controls, Q-Q plot shown in Fig. S6A) prior to linear
363 model analysis in Figs. 1 and S1. Box-Cox λ values were determined using maximum-likelihood
364 estimation (Box and Cox 1964) as implemented in geoR (https://CRAN.R-
365 project.org/package=geoR). Central tendency estimates and confidence intervals derived from
366 these models were then back transformed for the plots. Paired differences in TANGO scores or
367 Waltz scores between genes and scrambled controls were not transformed. Results were
368 qualitatively indistinguishable when runs of at least six consecutive amino acids were analyzed
369 instead of runs of at least five.

370 "Clustering" was assessed as a normalized index of dispersion, i.e. by comparing the variance in
371 hydrophobicity between blocks of consecutive amino acids to the mean hydrophobicity (Irbäck
372 *et al.* 1996). Examples of high and low clustering are shown in Fig. 5. We used $s = 6$, with
373 different values of s yielding qualitatively similar results. Where the amino acid length was not
374 divisible by six, a few amino acids were neglected at one or both ends, yielding a truncated
375 length of N, and we used the average clustering measure ψ across different phases for the
376 blocking procedure. In Fig. 4, we average over all phases using the maximum number of blocks,
377 e.g. only one phase for values of N divisible by 6. Results when we average over all 6 phases are
378 very similar. Following past practice, we transformed amino acid sequences into binary
379 hydrophobicity strings by taking the six amino acids FLIMVW as hydrophobic (+1) and scoring all
380 the other amino acids as -1. We summed hydrophobicity scores to a value σ_k for each block
381 $k = 1, \dots, N/s$ and $M = \sum_{k=1}^{N/s} \sigma_k$ overall (Irbäck and Sandelin 2000). Our clustering score is a
382 normalized index of dispersion

383 $$\psi = \frac{s}{N} \sum_{k=1}^{N/s} \frac{1}{K} (\sigma_k - sM/N)^2,$$

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

385 $$K = s \frac{N^2 - M^2}{N^2 - N} \left(1 - \frac{s}{N}\right).$$

386 For randomly distributed amino acids of any length N and hydrophobicity M, this normalization
387 makes the expectation of ψ equal to 1. For clustering at the nucleotide level, blocks of length
388 $s = 18$ rather than 6 were used. Nucleotide clustering values were calculated for each possible
389 permutation as to which nucleotides were scored as +1 and which as -1 (e.g. G and C as +1 and
390 A and T as -1 constitutes one permutation). Amino acid clustering values ψ were Box-Cox
391 transformed (λ=-0.29 for mouse, λ=-0.008 for yeast) prior to use in linear models, with mouse
392 Q-Q plot shown in Figure S6B.

393 To generate a scrambled control sequence that is paired to each gene, we simply sampled its
394 amino acids without replacement. To generate clustering-controlled scrambled sequences,
395 1000 scrambled sequences of each protein were produced, and the one that most closely
396 matched the clustering value of the focal gene was retained. This left the average gene with a

397 clustering value 0.0035 higher than its matched control, with the mean difference of the
398 absolute deviation between a gene and its matched control equal to 0.0057, showing a close
399 match with little directional bias. The mean value of each property was used across 50
400 scrambled sequences, but this led only to very modest reductions in confidence interval width
401 relative to using a single scrambled control, e.g. ~20% in Figure 3. Because generating well-
402 matched clustering-controlled scrambled sequences is computationally expensive, we used only
403 a single matched-clustering scrambled control sequence per gene.

404 Transmembrane protein annotation was taken from both the "Membrane Proteins of Known
405 3D Structure" database (Stansfeld *et al.* 2015) (http://blanco.biomol.uci.edu/mpstruc/,
406 including mouse proteins whose human homolog was experimentally verified as
407 transmembrane, accessed July 16, 2017), and from UniProt (The UniProt Consortium 2017)
408 (transmembrane annotation "experimental" ECO:0000269, "curated" ECO:0000303 and
409 ECO:0000305, or "homology" to a "related experimentally characterized protein" ECO:0000250,
410 accessed November 19).

411 **Data availability.** Source data for the statistical analyses and figures are provided in
412 Supplementary Tables S1-S6, available at Figshare and captioned in the main Supplementary
413 Materials file. Code associated with generating and analyzing these tables is publicly available
414 at https://github.com/MaselLab.

415 **Acknowledgements:** This work was supported by the John Templeton Foundation (39667,
416 60814), and the National Institutes of Health (GM104040). The funders had no role in study
417 design, data collection and analysis, decision to publish, or preparation of the manuscript. We
418 thank Rafik Neme for insightful discussions, and Joost Schymkowitz and Rob van der Kant of the
419 VIB Switch Laboratory for providing us with a stand-alone Waltz script.

420

Author contributions. J.M. conceived the general approach, M.H.J.C. conceived the clustering metric, SLiM and transmembrane protein analyses, J.B. produced Figures 5 and S2, B.A.W. fitted statistical models and produced the other final figures, S.G.F. conducted all other upstream data analysis, and J.M. wrote the paper.

Competing interests. The authors declare no competing financial interests.

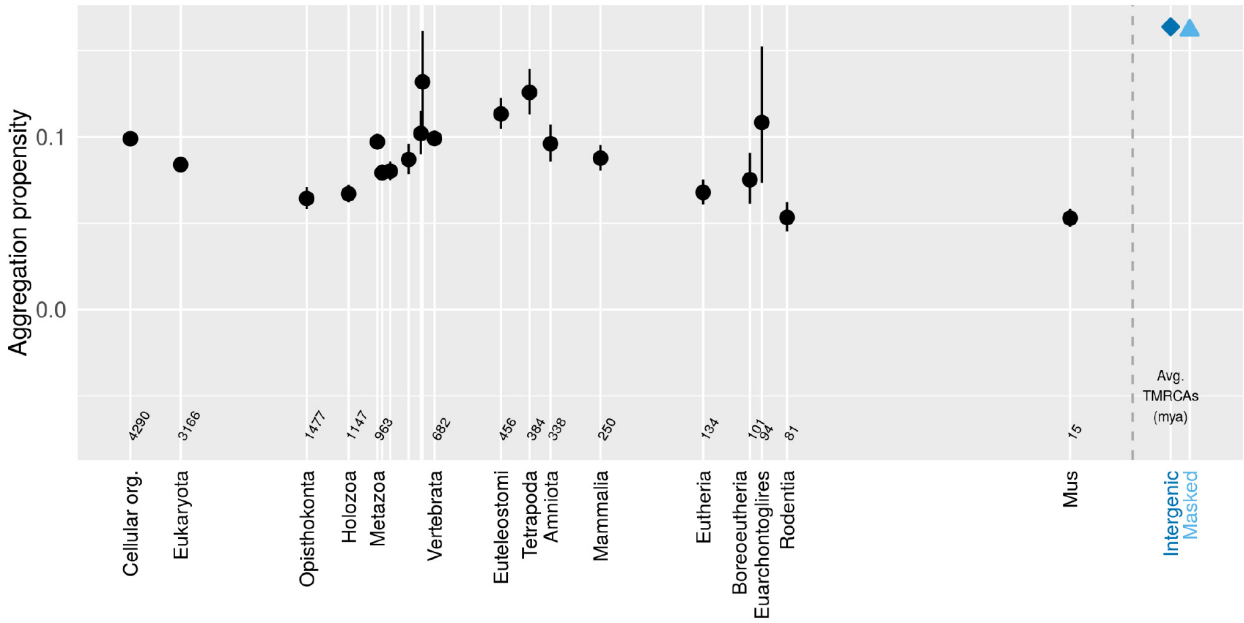


Fig. 1. Mouse genes show little pattern in aggregation propensity (assessed via TANGO) as a function of age. Genes (black) show less aggregation propensity than intergenic controls (blue). Back-transformed central tendency estimates +/- one standard error come from a linear mixed model applied to transformed 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 grey dashed line shows the 0 time, with control sequences to the right of it.

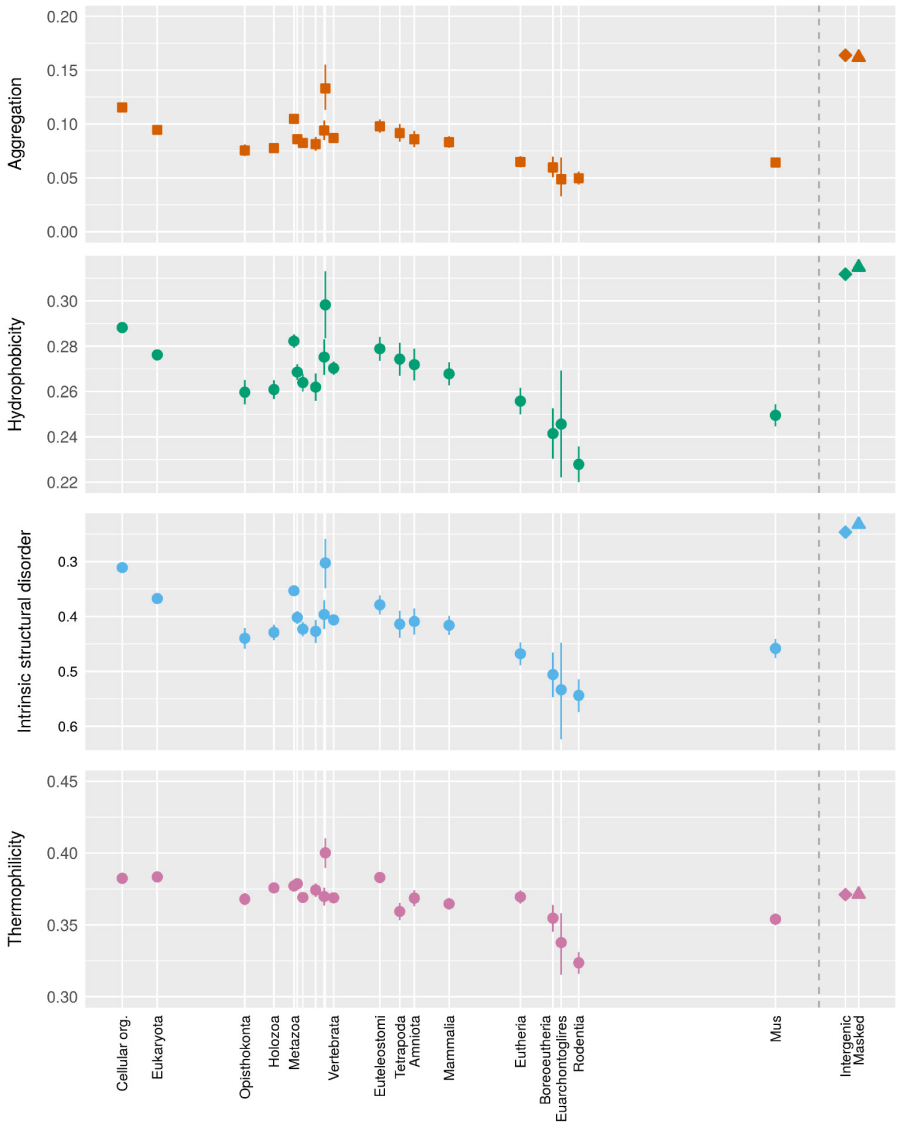


Fig. 2. Four different measures for the hydrophobicity of the amino acid content as a function of gene family age. "Aggregation" represents the average TANGO results from 50 scrambled versions of each gene, and hence captures the effect of amino acid composition on TANGO's estimate of β-aggregation propensity. 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 λ= 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 +/- one standard error come from a linear mixed model, where gene family and phylostratum are random and fixed terms respectively; λ=0.93 is used for hydrophobicity, other transforms are described in the Methods. The x-axis is the same as for Figure 1.

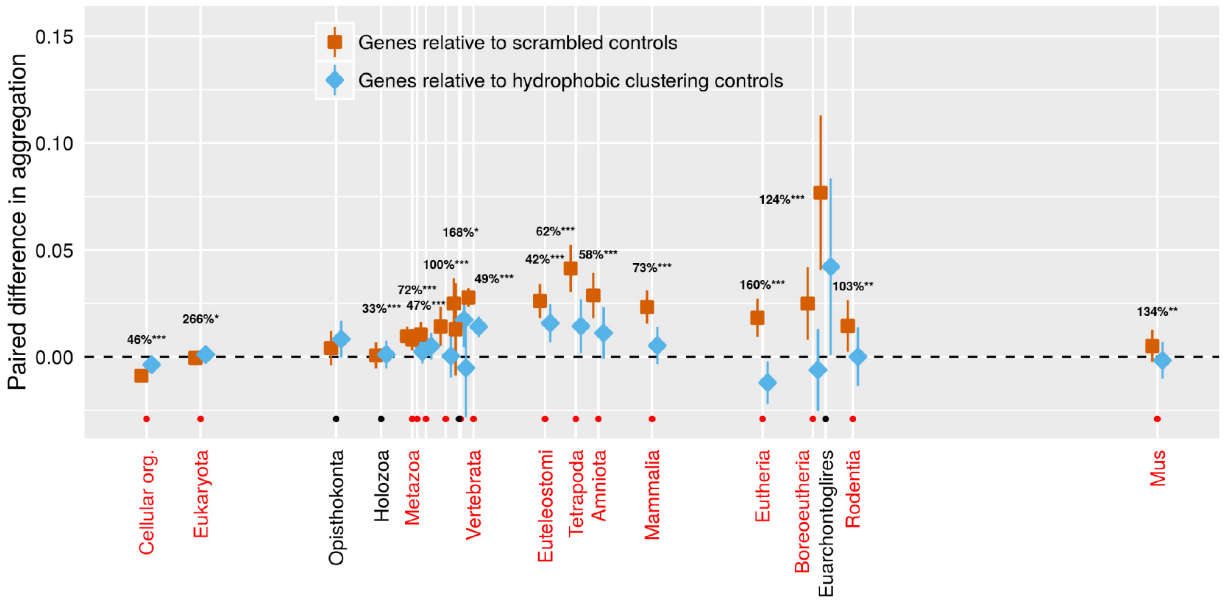
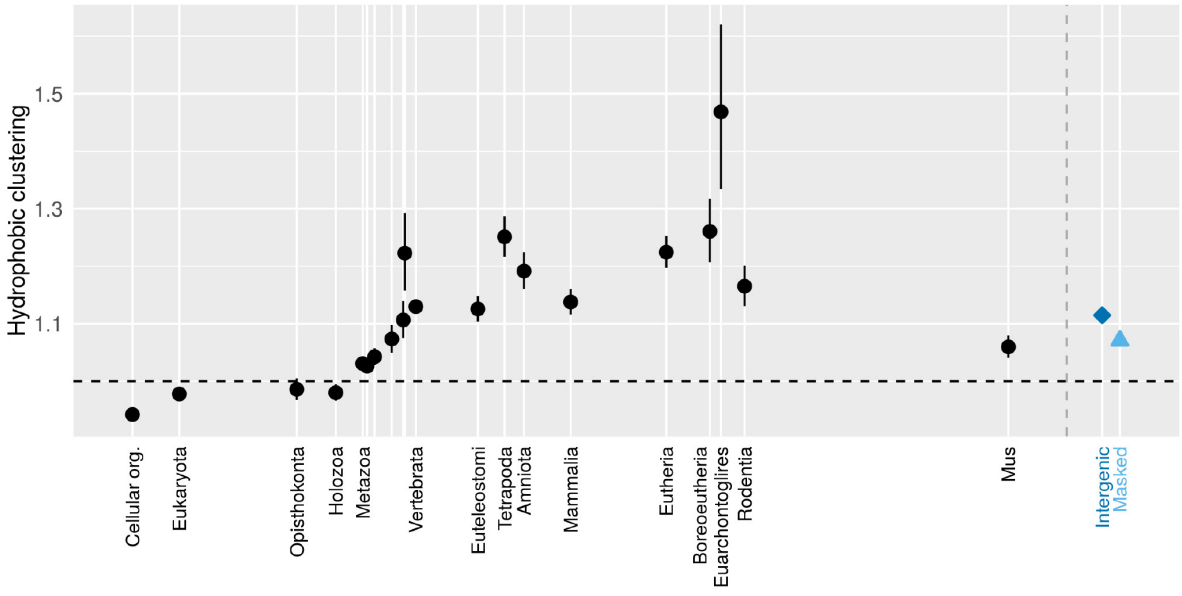


Fig. 3. Only very old genes have aggregation propensities lower than that expected from their amino acid composition alone (orange < dashed line expectation of 0). This puzzling finding is reduced when we account for clustering (blue is closer than orange is to the 0 dashed line) using a scrambled sequence 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 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. 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 1.



493

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

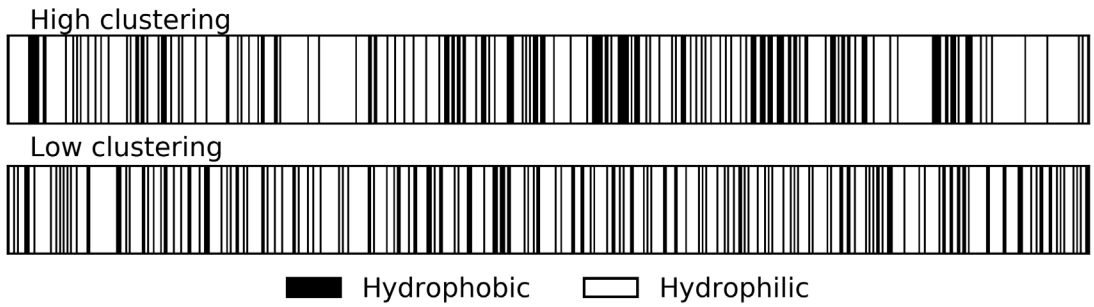


Fig. 5. Illustration of the distribution of hydrophobic residues along the primary sequence of proteins with high vs. low clustering, of similar lengths and net hydrophobicities. The high clustering gene Fzd5 has length 585 amino acids, 31.5% hydrophobicity, and clustering of 1.58. The low clustering gene Farsb has length 589amino acids, 31.6% hydrophobicity, and clustering of 0.69.

505 **References**

506 Albà, M. M., and J. Castresana, 2007 On homology searches by protein Blast and the characterization of
507 the age of genes. BMC Evol. Biol. 7: 1-8.
508 Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a
509 new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.
510 Banerjee, S., and S. Chakraborty, 2017 Protein intrinsic disorder negatively associates with gene age in
511 different eukaryotic lineages. Molecular BioSystems 13: 2044-2055.
512 Bloom, J. D., and M. J. Glassman, 2009 Inferring Stabilizing Mutations from Protein Phylogenies:
513 Application to Influenza Hemagglutinin. PLoS Comput. Biol. 5: e1000349.
514 Boussau, B., S. Blanquart, A. Necsulea, N. Lartillot and M. Gouy, 2008 Parallel adaptations to high
515 temperatures in the Archaean eon. Nature 456: 942-945.
516 Box, G. E. P., and D. R. Cox, 1964 An Analysis of Transformations. Journal of the Royal Statistical Society.
517 Series B (Methodological) 26: 211-252.
518 Broome, B. M., and M. H. Hecht, 2000 Nature disfavors sequences of alternating polar and non-polar
519 amino acids: implications for amyloidogenesis. J. Mol. Biol. 296: 961-968.
520 Buck, P. M., S. Kumar and S. K. Singh, 2013 On the Role of Aggregation Prone Regions in Protein
521 Evolution, Stability, and Enzymatic Catalysis: Insights from Diverse Analyses. PLoS Comput Biol 9:
522 e1003291.
523 Chen, Y., and N. V. Dokholyan, 2008 Natural Selection against Protein Aggregation on Self-Interacting
524 and Essential Proteins in Yeast, Fly, and Worm. Mol. Biol. Evol. 25: 1530-1533.
525 Davey, N. E., K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar *et al.*, 2012 Attributes of short linear motifs.
526 Molecular BioSystems 8: 268-281.
527 De Baets, G., J. Reumers, J. Delgado Blanco, J. Dopazo, J. Schymkowitz *et al.*, 2011 An Evolutionary
528 Trade-Off between Protein Turnover Rate and Protein Aggregation Favors a Higher Aggregation
529 Propensity in Fast Degrading Proteins. PLoS Comput Biol 7: e1002090.
530 De Baets, G., L. Van Doorn, F. Rousseau and J. Schymkowitz, 2015 Increased Aggregation Is More
531 Frequently Associated to Human Disease-Associated Mutations Than to Neutral Polymorphisms.
532 PLoS Comput Biol 11: e1004374.
533 Dill, K. A., 1990 Dominant forces in protein folding. Biochemistry 29: 7133-7155.
534 Domazet-Lošo, T., J. Brajković and D. Tautz, 2007 A phylostratigraphy approach to uncover the genomic
535 history of major adaptations in metazoan lineages. Trends Genet. 23: 533-539.
536 Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke and F. H. Arnold, 2005 Why highly expressed
537 proteins evolve slowly. Proc. Natl. Acad. Sci. USA 102: 14338-14343.
538 Drummond, D. A., and C. O. Wilke, 2008 Mistranslation-Induced Protein Misfolding as a Dominant
539 Constraint on Coding-Sequence Evolution. Cell 134: 341-352.
540 Fernandez-Escamilla, A. M., F. Rousseau, J. Schymkowitz and L. Serrano, 2004 Prediction of sequence-
541 dependent and mutational effects on the aggregation of peptides and proteins. Nat. Biotechnol.
542 22: 1302-1306.
543 Gaucher, E. A., S. Govindarajan and O. K. Ganesh, 2008 Palaeotemperature trend for Precambrian life
544 inferred from resurrected proteins. Nature 451: 704-707.
545 Godoy-Ruiz, R., R. Perez-Jimenez, B. Ibarra-Molero and J. M. Sanchez-Ruiz, 2004 Relation Between
546 Protein Stability, Evolution and Structure, as Probed by Carboxylic Acid Mutations. J. Mol. Biol.
547 336: 313-318.
548 Gunasekaran, K., C.-J. Tsai and R. Nussinov, 2004 Analysis of Ordered and Disordered Protein Complexes
549 Reveals Structural Features Discriminating Between Stable and Unstable Monomers. J. Mol. Biol.
550 341: 1327-1341.

551 Herrero, J., M. Muffato, K. Beal, S. Fitzgerald, L. Gordon *et al.*, 2016 Ensembl comparative genomics
552 resources. Database 2016: bav096.
553 Hurst, L. D., E. J. Feil and E. P. C. Rocha, 2006 Causes of trends in amino-acid gain and loss. Nature 442:
554 E11-E12.
555 Irbäck, A., C. Peterson and F. Potthast, 1996 Evidence for nonrandom hydrophobicity structures in
556 protein chains. Proc. Natl. Acad. Sci. USA 93: 9533-9538.
557 Irbäck, A., and E. Sandelin, 2000 On Hydrophobicity Correlations in Protein Chains. Biophysical Journal
558 79: 2252-2258.
559 Kumar, S., G. Stecher, M. Suleski and S. B. Hedges, 2017 TimeTree: A Resource for Timelines, Timetrees,
560 and Divergence Times. Mol. Biol. Evol. 34: 1812-1819.
561 Lee, Y., T. Zhou, G. G. Tartaglia, M. Vendruscolo and C. O. Wilke, 2010 Translationally optimal codons
562 associate with aggregation-prone sites in proteins. Proteomics 10: 4163-4171.
563 Lehmann, M., L. Pasamontes, S. F. Lassen and M. Wyss, 2000 The consensus concept for thermostability
564 engineering of proteins. BBA-Protein Struct. M. 1543: 408-415.
565 Linding, R., J. Schymkowitz, F. Rousseau, F. Diella and L. Serrano, 2004 A Comparative Study of the
566 Relationship Between Protein Structure and β-Aggregation in Globular and Intrinsically
567 Disordered Proteins. J. Mol. Biol. 342: 345-353.
568 Mannige, R. V., C. L. Brooks and E. I. Shakhnovich, 2012 A Universal Trend among Proteomes Indicates
569 an Oily Last Common Ancestor. PLoS Comput. Biol. 8: e1002839.
570 Maurer-Stroh, S., M. Debulpaep, N. Kuemmerer, M. Lopez de la Paz, I. C. Martins *et al.*, 2010 Exploring
571 the sequence determinants of amyloid structure using position-specific scoring matrices. Nature
572 Methods 7: 237-242.
573 McDonald, J. H., 2006 Apparent Trends of Amino Acid Gain and Loss in Protein Evolution Due to Nearly
574 Neutral Variation. Mol. Biol. Evol. 23: 240-244.
575 McLysaght, A., and D. Guerzoni, 2015 New genes from non-coding sequence: the role of de novo
576 protein-coding genes in eukaryotic evolutionary innovation. Phil. Trans. R. Soc. B 370: 20140332.
577 McLysaght, A., and L. D. Hurst, 2016 Open questions in the study of de novo genes: what, how and why.
578 Nature Reviews Genetics 17: 567–578.
579 Monsellier, E., and F. Chiti, 2007 Prevention of amyloid-like aggregation as a driving force of protein
580 evolution. EMBO Rep. 8: 737-742.
581 Monsellier, E., M. Ramazzotti, P. P. de Laureto, G.-G. Tartaglia, N. Taddei *et al.*, 2007 The Distribution of
582 Residues in a Polypeptide Sequence Is a Determinant of Aggregation Optimized by Evolution.
583 Biophysical Journal 93: 4382-4391.
584 Moyers, B. A., and J. Zhang, 2015 Phylostratigraphic Bias Creates Spurious Patterns of Genome
585 Evolution. Mol. Biol. Evol. 32: 258-267.
586 Moyers, B. A., and J. Zhang, 2016 Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene
587 Birth in Genome Evolution. Mol. Biol. Evol. 33: 1245-1256.
588 Moyers, B. A., and J. Zhang, 2017 Further Simulations and Analyses Demonstrate Open Problems of
589 Phylostratigraphy. Genome Biol. Evol. 9: 1519-1527.
590 Palmieri, N., C. Kosiol and C. Schlötterer, 2014 The life cycle of *Drosophila* orphan genes. eLife 3: e01311.
591 Patki, A. U., A. C. Hausrath and M. H. J. Cordes, 2006 High Polar Content of Long Buried Blocks of
592 Sequence in Protein Domains Suggests Selection Against Amyloidogenic Non-polar Sequences. J.
593 Mol. Biol. 362: 800-809.
594 Povolotskaya, I. S., and F. A. Kondrashov, 2010 Sequence space and the ongoing expansion of the
595 protein universe. Nature 465: 922-926.
596 Reumers, J., S. Maurer-Stroh, J. Schymkowitz and F. Rousseau, 2009 Protein sequences encode
597 safeguards against aggregation. Hum. Mutat. 30: 431-437.

598 Rousseau, F., L. Serrano and J. W. H. Schymkowitz, 2006 How Evolutionary Pressure Against Protein
599 Aggregation Shaped Chaperone Specificity. J. Mol. Biol. 355: 1037-1047.
600 Sánchez, I. E., J. Tejero, C. Gómez-Moreno, M. Medina and L. Serrano, 2006 Point Mutations in Protein
601 Globular Domains: Contributions from Function, Stability and Misfolding. J. Mol. Biol. 363: 422-
602 432.
603 Schwartz, R., S. Istrail and J. King, 2001 Frequencies of amino acid strings in globular protein sequences
604 indicate suppression of blocks of consecutive hydrophobic residues. Protein Science 10: 1023-
605 1031.
606 Smit, A., R. Hubley and P. Green, 2015 RepeatMasker Open-4.0 version 4.0.5.
607 url=http://www.repeatmasker.org.
608 Soding, J., and A. N. Lupas, 2003 More than the sum of their parts: on the evolution of proteins from
609 peptides. BioEssays 25: 837-846.
610 Stansfeld, Phillip J., Joseph E. Goose, M. Caffrey, Elisabeth P. Carpenter, Joanne L. Parker *et al.*, 2015
611 MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid
612 Membranes. Structure 23: 1350-1361.
613 Steipe, B., B. Schiller, A. Plückthun and S. Steinbacher, 1994 Sequence Statistics Reliably Predict
614 Stabilizing Mutations in a Protein Domain. J. Mol. Biol. 240: 188-192.
615 Tartaglia, G. G., S. Pechmann, C. M. Dobson and M. Vendruscolo, 2007 Life on the edge: a link between
616 gene expression levels and aggregation rates of human proteins. Trends Biochem. Sci. 32: 204-
617 206.
618 Tartaglia, G. G., R. Pellarin, A. Cavalli and A. Caflisch, 2005 Organism complexity anti-correlates with
619 proteomic β-aggregation propensity. Protein Science 14: 2735-2740.
620 Thangakani, A. M., S. Kumar, D. Velmurugan and M. S. M. Gromiha, 2012 How do thermophilic proteins
621 resist aggregation? Proteins: Struct. Funct. Bioinf. 80: 1003-1015.
622 The UniProt Consortium, 2017 UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45:
623 D158-D169.
624 Thybert, D., M. Roller, F. C. P. Navarro, I. Fiddes, I. Streeter *et al.*, 2018 Repeat associated mechanisms of
625 genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes. Genome
626 Res. 28: 448-459.
627 Trudeau, D. L., M. Kaltenbach and D. S. Tawfik, 2016 On the Potential Origins of the High Stability of
628 Reconstructed Ancestral Proteins. Mol. Biol. Evol. 33: 2633-2641.
629 Uversky, V. N., J. R. Gillespie and A. L. Fink, 2000 Why are "natively unfolded" proteins unstructured
630 under physiologic conditions? Proteins 41: 415-427.
631 Williams, P. D., D. D. Pollock, B. P. Blackburne and R. A. Goldstein, 2006 Assessing the Accuracy of
632 Ancestral Protein Reconstruction Methods. PLoS Comput. Biol. 2: e69.
633 Wilson, B. A., S. G. Foy, R. Neme and J. Masel, 2017 Young genes are highly disordered as predicted by
634 the preadaptation hypothesis of de novo gene birth. Nat. Ecol. Evol. 1: 0146.
635 Yampolsky, L. Y., and M. A. Bouzinier, 2010 Evolutionary patterns of amino acid substitutions in 12
636 Drosophila genomes. BMC Genomics 11: S10.
637 Yampolsky, L. Y., Y. I. Wolf and M. A. Bouzinier, 2017 Net Evolutionary Loss of Residue Polarity in
638 Drosophilid Protein Cores Indicates Ongoing Optimization of Amino Acid Composition. Genome
639 Biol. Evol. 9: 2879-2892.
640 Zhu, H., E. Sepulveda, M. D. Hartmann, M. Kogenaru, A. Ursinus *et al.*, 2016 Origin of a folded repeat
641 protein from an intrinsically disordered ancestor. eLife 5: e16761.

642

643