More benign random peptides in E. coli are enriched for similar amino acids as young animal but not plant genes

Luke Kosinski[1]*, Nathan Aviles[2]*, Kevin Gomez[3], Joanna Masel[4]

[1] Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721

[2] Graduate Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ 85721

[3] Graduate Interdisciplinary Program in Applied Math, University of Arizona, Tucson, AZ 85721

[4] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721

* These authors contributed equally to this work.

Corresponding author: Joanna Masel, masel@email.arizona.edu +1 520 626 9888

Abstract

Proteins are the workhorses of the cell, yet they carry great potential for harm via misfolding and aggregation. Despite the dangers, proteins are sometimes born *de novo* from non-coding DNA. Proteins are more likely to be born from non-coding regions that produce peptides that do little to no harm when translated than from regions that produce harmful peptides. To investigate which newborn proteins are most likely to "first, do no harm", we estimate fitnesses from an experiment that competed *Escherichia coli* lineages that each expressed a unique random peptide. A variety of peptide metrics significantly predict lineage fitness, but almost all this predictive power stems from simple amino acid composition. Amino acids that are smaller and that promote intrinsic structural disorder have more benign fitness effects. Our amino acid composition-based predictions also have validity for an independent dataset using small random N-terminal tags. The same amino acids that predict high fitness in *E. coli* are enriched in young Pfams in animals, but not in plants. To modify Jacques Monod's famous quote, what makes peptides benign in *E.coli* also makes them benign in elephants, but not in eucalyptus.

Keywords: Experimental evolution, evolvability, fitness estimation, preadaptation, de novo gene birth

Introduction

Proteins are the workhorses of the cell, but they are dangerous. The polypeptide backbone is the key structural feature of amyloids, putting all proteins at risk of forming insoluble aggregates (Chiti and Dobson 2017). The risks of aggregation are exacerbated by the fact that most proteins are expressed at or just beyond their solubility limits (Vecchi, et al. 2020). Despite these dangers, new protein-coding genes are nevertheless born *de novo* from essentially random sequences (McLysaght and Guerzoni 2015; Van Oss and Carvunis 2019; Vakirlis, Carvunis, et al. 2020). To be beneficial enough for *de novo* birth, a peptide must first do no serious harm, i.e. it must not be strongly deleterious (detrimental to the basic functioning of a cell). Here we ask what properties are predictive of whether a peptide is harmful.

Neme et al. (2017) competed over 2 million *Escherichia coli* lineages, each containing a plasmid designed to express a unique random peptide, and tracked lineage frequencies over four days using deep DNA sequencing. This study has been criticized for including too little evidence in support of the beneficial nature of the top candidates (Weisman and Eddy 2017; Knopp and Andersson 2018). But these criticisms do not detract from using the dataset to identify statistical predictors of serious harm versus relatively benign effect. Neme et al. used a strong promoter (Neme, et al. 2017), so evaluating whether a peptide is harmful or benign is largely an evaluation of tolerance of high expression. Given this high expression, fitness of these lineages might plausibly be influenced by properties of the peptide itself, not just variation in expression e.g. due to auto-downregulation at the RNA level (Knopp and Andersson 2018). Here we pursue analyses based on this hypothesis that the properties of the random peptides comprise some appreciable portion of the variation in fitness among lineages in this experiment.

Conveniently, some properties, such as intrinsic structural disorder and aggregation propensity, can be predicted computationally from peptide sequences alone. Because insoluble proteins have been implicated in toxicity and disease (Chiti and Dobson 2017) and peptides with high intrinsic structural disorder are less prone to forming insoluble aggregates (Linding, et al. 2004; Angyan, et al. 2012), we hypothesize that highly disordered peptides are least likely to be strongly deleterious. Random sequences with high predicted disorder are well-tolerated *in vivo* (Tretyachenko, et al. 2017). Existing mouse (Wilson, et al. 2017) and Drosophila (Heames, et al. 2020) proteins, which are the product of evolution, are predicted from their amino acid sequences to be more disordered than what would be translated from intergenic controls.

Younger protein-coding sequences should be particularly constrained to first do no harm, as they have had little time to evolve more sophisticated harm-avoidance strategies (Foy, et al. 2019). In support of the idea that high intrinsic structural disorder is an accessible way to avoid harm, young animal and fungal domains (James, et al. 2020) and genes (James, et al. 2020), and novel overprinted viral genes (Willis and Masel 2018) have higher predicted disorder than their older counterparts.

High predicted intrinsic structural disorder in young genes is driven primarily by amino acid composition, with amino acid order playing a more minor role (Wilson, et al. 2017). Fortunately, the dataset of Neme et al. (2017) is large enough to look at the frequencies of each amino acid as predictors, rather than assume that existing prediction programs such as IUPred (Dosztányi, et al. 2005; Meszaros, et al. 2018) or Tango (Fernandez-Escamilla, et al. 2004; Linding, et al. 2004; Rousseau, et al. 2006) integrate amino acid composition (and more) in the best possible way. We can then test whether such programs have additional ability to predict peptide fitness, above and beyond the influence of amino acid content.

How each amino acid frequency predicts peptide harm may also be useful in resolving disputes about the technique of phylostratigraphy, which assigns ages to genes. Detecting homologs is more difficult for fast-evolving sequences, which may be erroneously scored as young (Alba and Castresana 2007; Moyers and Zhang 2015, 2016). Disordered proteins tend to be fast evolving (Chen, et al. 2011), casting doubt over findings that young genes are more disordered. If the amino acid preferences of higher fitness random peptides match those of purportedly young protein-coding sequences, this would be evidence that the *de novo* gene birth process, rather than homology detection bias alone, causes trends in protein properties as a function of apparent gene age.

Monod and Jacob (1961) famously claimed that “anything found to be true of *E. coli* must also be true of Elephants.” We test whether amino acids that are good for random peptides in *E. coli* are enriched in young animal and plant proteins from a recent phylostratigraphy dataset (James, et al. 2020) of over 400 different eukaryotic species. We also test whether published, sequence-based estimators of disorder, aggregation propensity, water solubility, hydrophobic residue clustering etc. predict fitness information not already predicted by amino acid frequencies.

Results

Estimating the fitness effects of random peptides

Assessing predictors of the fitness effects of random peptides requires those fitness effects to be measured accurately and precisely. Neme et al. (2017) tracked lineage frequencies over four days, and categorized a peptide as increasing or decreasing in frequency by comparing the counts of day 4 to day 1 using DESeq2 (Love, et al. 2014). Out of the 1082 peptides with at least five sequencing reads, they found that 277 were increasing, while 436 were decreasing.

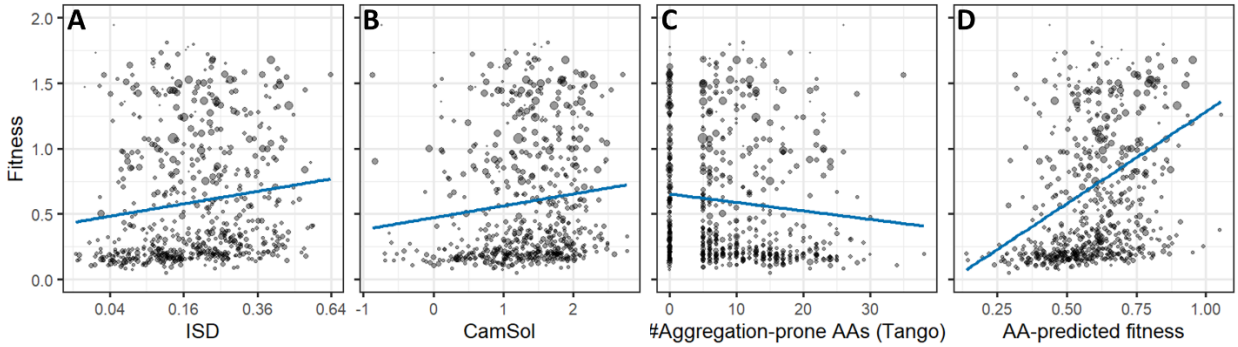
We reanalyze the same data, instead using a custom maximum likelihood framework (see Materials & Methods) to quantitatively estimate "fitness" and its associated confidence interval / weight. "Fitness" here refers to allele frequency changes over an entire cycle of population growth and dilution, rather than per generation. Our method classifies peptides quantitatively rather than qualitatively. It accounts for the fact that mean population fitness increases over the four days (see Materials and Methods). Our use of all available data within an appropriate maximum likelihood framework should make our method more sensitive and specific for identifying benign vs harmful peptides. We find statistical support for increases in frequency (i.e. fitness – 1.96 × standard error > 1) for 475 peptides (Dataset S1). We exclude 22 of the previously reported 277, at least some of which arose from a bug we located in Neme et al.'s (2017) pipeline. 509 peptides had statistical support for decreases in frequency, with support lost for 9 of the previously reported 436 deleterious peptides.

Note that some peptides are pseudoreplicates (see Materials & Methods). There were 646 total clusters, of which there was statistical support for increases in frequency for the highest-weighted peptide in 138, and for decreases in 488. Some of our statistics use cluster as a random effect within a linear mixed model. When interpretable R^2 values are needed, we instead collapse each cluster into a single pseudo-datapoint with value given by the weighted mean and weight given by the sum of weights, and then use fixed effect models.

Amino acid counts are the best predictor of random peptide fitness effects

We estimated peptide disorder using IUPred (Dosztányi, et al. 2005; Meszaros, et al. 2018), water solubility using CamSol (Sormanni, et al. 2015), and general aggregation propensity using Tango (Fernandez-Escamilla, et al. 2004; Linding, et al. 2004; Rousseau, et al. 2006). Fewer than 6% of the

113 random peptides have a predicted transmembrane helix (Dataset S1) from TMHMM (Krogh, et al. 2001),
114 so our choice of these predictors is guided by our assumption that the random peptides are
115 predominantly located in the cytosol. Having a predicted transmembrane helix did not in itself predict
116 random peptide fitness effects (P = 0.2, likelihood ratio test relative to mixed model with only the
117 intercept as a fixed effect). In contrast, each of our cytosol-solubility-inspired metrics significantly
118 predicted random peptide fitness (Fig. 1A – 1C), with effects in the predicted direction (more disorder
119 and more solubility are good, more aggregation propensity is bad). Another aggregation predictor, Waltz
120 (Maurer-Stroh, et al. 2010), that specializes in β aggregates, was in the right direction but did not quite
121 meet statistical significance (P = 0.06).

122 

123 Fig. 1. **Many metrics, on their own, predict peptide fitness effects, but amino acid composition**
124 **predicts the most.** A) ISD (on a square root scale), B) CamSol, and C) the number of amino acids in
125 Tango-predicted aggregation-prone regions each significantly predict peptide fitness on their own (P = 7
126 × 10^{-4}, 0.003, and 0.02, respectively, likelihood ratio test in mixed model). D) The best model is fitness
127 predicted by amino acid composition; weighted R^2 values for predicting cluster fitness (n = 646) from the
128 ISD, CamSol, Tango, and amino acid composition models are 0.029, 0.031, 0.017, and 0.15, respectively.
129 Each point shows a cluster with area is proportional to summed weights across that cluster. Seven
130 clusters with fitness greater than 2 are not shown here for ease of visualization; a complete y-axis is
131 shown in supplemental fig. 1. Log-transforming fitness would remove high fitness skew, but creates
132 systematic heteroscedasticity, and so was not done (supplemental fig. 2). The lack of systematic
133 heteroscedasticity can be seen here in the form of similar point size across fitness values. Blue lines
134 come from weighted linear mixed regression models (see Materials and Methods), where a random
135 effect term controls for pseudoreplication within clusters.

136

137 While IUPred, CamSol, and Tango are all statistically supported predictors of fitness, all pale in
138 comparison to amino acid composition. Weighted R^2 values for IUPred, CamSol, and Tango are 0.029,
139 0.031, 0.017, respectively. However, when the frequencies of each of the twenty amino acids were also
140 used as predictors (Fig. 1D), IUPred, CamSol, and Tango were no longer significant predictors of fitness
141 (P = 0.2, 0.2, and 0.3, respectively, likelihood ratio test in mixed model, see Supplemental Table S1),
142 suggesting that their predictive power in our dataset came merely from being metrics of amino acid
143 composition. Amino acid composition also explains a much greater amount of the variation in fitness
144 than any of the other predictors, with a weighted adjusted R^2 of 0.15. We are surprised that such a high
145 proportion of fitness can be predicted merely by a linear function of the amino acid content of a single
146 highly expressed peptide – both protein folding and organismal function are generally assumed not to
147 be so easily reduced.

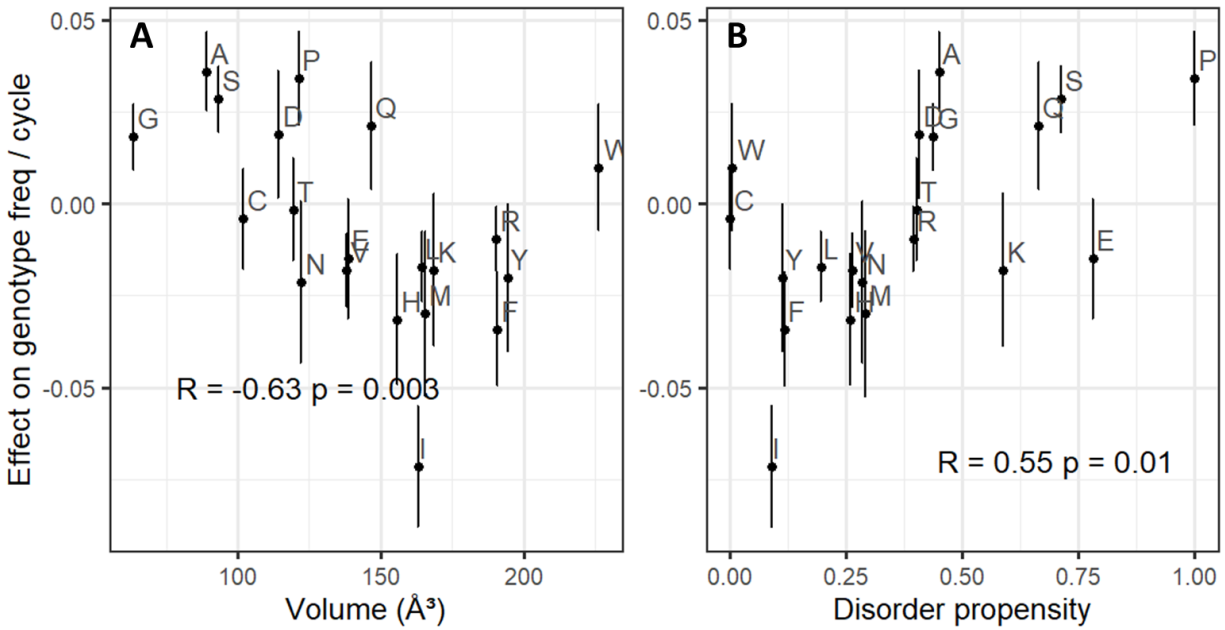
148 We next tried two predictors that focus exclusively on the order rather than the frequency of amino
149 acids in a peptide. First, the contribution of amino acid order to ISD (calculated as ΔISD = original
150 sequence ISD – scrambled sequence ISD), is higher in younger mouse genes (Wilson, et al. 2017).
151 However, this was not the case for the very youngest mouse genes, and including ΔISD did not
152 significantly improve our model (P = 0.2).

153 Second, previous work found that young proteins have their hydrophobic amino acids more clustered
154 together along the primary sequence than expected by chance, while old proteins have them more
155 interspersed (Foy, et al. 2019; James, et al. 2020). This hydrophobic dispersion metric (Irback, et al.
156 1996; Irback and Sandelin 2000), which is normalized to exclude any influence from amino acid
157 composition, did not significantly predict the fitness of our random peptides on its own (P = 0.2,
158 likelihood ratio test in mixed model), nor did including it significantly improve our model with amino
159 acid composition (P = 0.2, Supplemental Table S1).

160 Young animal sequences are enriched for amino acids that increase fitness in 161 random peptides

162 For ease of interpretation, we calculated the marginal effect of each amino acid on fitness (see
163 supplementary text and Table S2, displayed in fig. 2, y-axis), instead of simply using its slope in a
164 regression model. Fitness effects are correlated with both amino acid volume (Tsai, et al. 1999) and with
165 a metric of how often an amino acid tends to be found in known disordered versus ordered proteins
166 (Theillet, et al. 2013) (fig. 2A and 2B; P = 0.01 for both disorder propensity and volume, likelihood ratio

167 test for dropping either term from weighted regression containing both). Volume and disorder
168 propensity together explain over half the weighted variation in marginal fitness effect (weighted
169 adjusted R^2 = 0.52). Other properties of amino acids, such as stickiness (Levy, et al. 2012), relative
170 solvent accessibility (Tien, et al. 2013), amino acid cost in *E. coli* (Akashi and Gojobori 2002), and
171 isoelectric point (Liu, et al. 2004) did not provide significant explanatory power on top of disorder
172 propensity and volume (all P > 0.1, likelihood ratio test).



173

174 Fig. 2. **Amino acids that are small and are associated with intrinsic structural disorder promote higher**
175 **fitness**. The y-axis shows each amino acid's marginal effect on fitness, which is the change in fitness
176 when one amino acid of the focal type replaces one randomly chosen amino acid of a different type in a
177 random peptide (see Supporting Information). Error bars are +/- one standard error. P-values and
178 correlation coefficients come from weighted Pearson's correlations, where weights for marginal effects
179 are calculated as 1 / s.e.(marginal fitness effect)^2, and volume and disorder propensity are unweighted.
180 Amino acid fitness effects are correlated with (A) amino acid volume and with (B) disorder propensity.

181 Tryptophan is an outlier for amino acid effects on fitness, predicting minor positive effects on fitness
182 despite both its large volume and its underrepresentation in disordered regions. Removing tryptophan
183 from a weighted regression model of volume and disorder propensity predicting marginal effect changes
184 the weighted adjusted R^2 from 0.52 to 0.68. Tryptophan, encoded only by UGG, is nearly 60% more
185 common among peptides with at least 5 sequences reads than we expect from the 58% GC content of

our dataset. With the confidence interval for its marginal fitness effect including 1, this provides further evidence that tryptophan is advantageous, making it a distinct outlier, for reasons that are not clear to us.

Isoleucine also stands out, as even more harmful than expected by its size and order propensity. Isoleucine's harmful effects may be exacerbated by its role in amyloid formation. For example, familial amyloid cardiomyopathy is most commonly caused by a valine to isoleucine mutation (Jacobson, et al. 1997; Dubrey, et al. 2015), suggesting isoleucine has potential to form dangerous amyloids where other hydrophobic amino acids do not. Isoleucine, valine, and leucine are all hydrophobic amino acids with a branched carbon, but only raised isoleucine levels are associated with a higher risk of Alzheimer's disease (Larsson and Markus 2017), further suggesting that isoleucine may be especially prone to amyloid formation.

Next we tested whether the amino acids that are relatively benign in random peptides expressed under laboratory conditions in *E. coli* are similar to the amino acids that are found in purportedly young protein-coding sequences in natural populations. We took the latter from James et al. (2020) who quantified how amino acid frequencies depend on the age of the protein domain, across over 400 eukaryotic species. As predicted by the de novo gene birth interpretation of phylostratigraphy trends, amino acids that are good for random peptides are enriched among the youngest animal Pfams (fig. 3A).

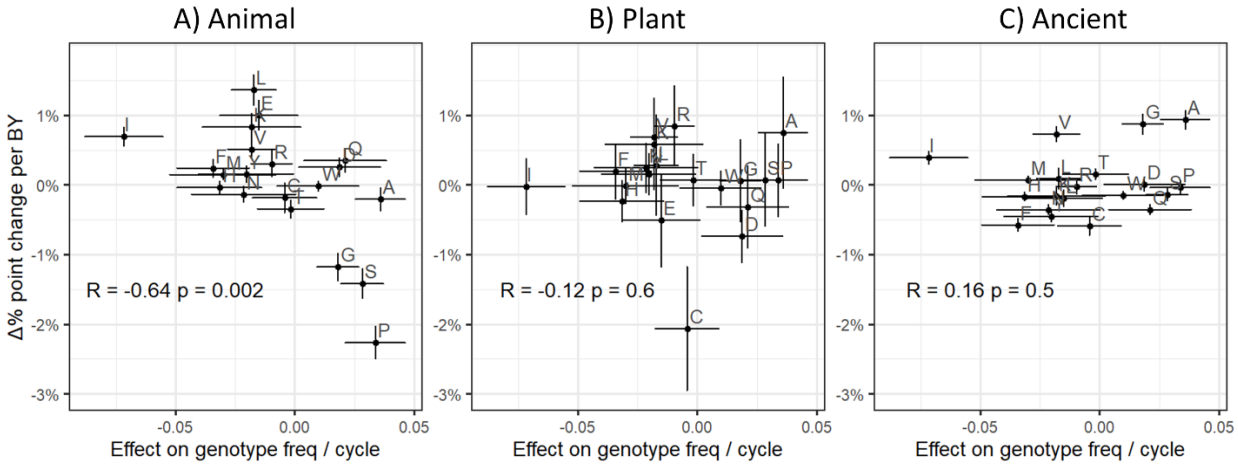


Fig. 3. **Purportedly young animal Pfams are enriched for amino acids that predict high fitness in random peptides.** The y-axis represents the change in the frequency of each amino acid per billion years (BY) for non-transmembrane Pfam domains (James, et al. 2020). Change is in number of percentage points, not percentage increase or decrease. For example, a change in glutamic acid content from 5% to

208 6% is a change of 1 percentage point. The x-axis shows each amino acid's marginal effect on fitness,
209 which is the change in fitness when one amino acid of the focal type replaces one randomly chosen
210 amino acid of a different type in a random peptide (see Supporting Information). Error bars are +/- one
211 standard error. Fitness effects predict A) animal, but not B) plant, or C) ancient (older than 2.1 billion
212 years) Pfam phylostratigraphy slopes. Correlation coefficients and P-values come from weighted
213 Pearson correlations. Note that the P-value for animal phylostratigraphy slopes vs marginal effects
214 survives a conservative Bonferroni correction ($P = 0.002 < 0.05/3 = 0.017$).

215

216 This prediction was not, however, supported for trends among recent plant domains (fig. 3B) nor among
217 ancient (fig. 3C) domains older than 2.1 billion years. Plant and ancient trends reflect a *de novo* gene
218 birth process that enriches for the most abundant amino acids in their respective lineages, rather than
219 for amino acids that promote intrinsic structural disorder (James, et al. 2020). We find that intrinsic
220 structural disorder still predicts harmlessness in *E. coli*, even though we do not find evidence it shaped
221 *de novo* gene birth in its distant ancestors. We also note that intrinsic structural disorder does shape
222 recent *de novo* gene birth in viruses (Willis and Masel 2018).

223 The correlation in fig. 3A is nevertheless an important external validation that patterns observed using
224 phylostratigraphy can be indicative of the gene birth process, at least in animals. As Monod suggested,
225 what makes a peptide benign in *E. coli* does also make a peptide benign in elephants – with the caveat
226 that this might not be the case in eucalyptus.

227 Determining which sequences are young is difficult, and may be subject to systematic error (homology
228 detection bias) (Alba and Castresana 2007; Moyers and Zhang 2015, 2016). Our ability to externally
229 validate animal phylostratigraphy slopes against an independent dataset generated in *E. coli* is a
230 powerful demonstration that these slopes represent more than mere bias. This conclusion is also
231 supported by the fact that these slopes differ among taxa and are not correlated with amino acid
232 changeability (James, et al. 2020).

233 Fitness is better predicted by amino acid composition than by GC content

234 Long et al. (2018) proposed that selection acts directly on GC content, perhaps due to the three
235 hydrogen bonds of G-C pairs. Amino acids encoded by Gs and Cs tend to promote higher intrinsic
236 structural disorder (Angyan, et al. 2012), making it difficult to distinguish between selection for high GC
237 content and selection for disorder-promoting amino acids. To do this, we compare amino acids that

always have G or C to those that always have A or T, at both the first and second nucleotide positions in the codon. If selection were for GC nucleotides, we would expect GC to predict high marginal amino acid fitness effects at both positions. But if results are dramatically different at the two positions, this would show that it is selection on amino acid content that drives GC as a correlated trait. Results are statistically significant in the predicted direction at the second position (fig. 4A, P = 0.001, weighted Welch's t-test), and in the predicted direction but not statistically significant at the first (fig. 4B, P = 0.2). The effect size of GC content on fitness could not be statistically distinguished between the first and second position (fig. 4C), with wide and hence inconclusive error bars.

However, amino acid composition is a far stronger predictor of fitness than GC content. While GC content is a predictor, offering a significant improvement over an intercept-only model ($P = 6 \times 10^{-11}$, likelihood ratio test for nested fixed-effect models, weighted adjusted $R^2 = 0.06$), it predicts considerably less than amino acid composition ($P = 10^{-18}$, weighted adjusted $R^2 = 0.15$). Adding GC content to the amino acid frequencies-only model offers a modest improvement (P = 0.004, weighted adjusted R^2 values improves from 0.15 to 0.16), while adding amino acid composition to a GC content only model offers a major improvement ($P = 10^{-11}$, weighted adjusted R^2 improves from 0.06 to 0.16). These findings suggest that while there may be some direct selection on GC content, the strength of this selection pales in comparison to selection on amino acid composition.

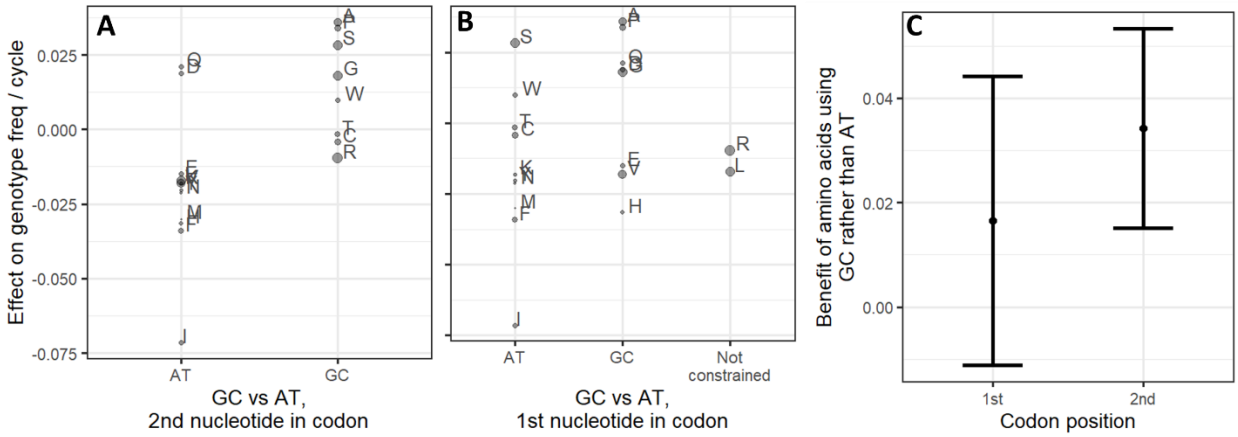


Fig. 4. **Amino acids that are constrained to use Gs and Cs tend to have higher marginal effects on fitness than those constrained to use As and Ts**. The difference is significant for constraints at the second nucleotide position of a codon (A) (P = 0.001, weighted Welch's t-test), but not at the first (B) (P = 0.2). Point area is proportional to weight, which is calculated as 1 / s.e.(marginal fitness effect)^2, as described in Supporting Information. The y-axis is the same as the fig. 2 y-axis and fig. 3 x-axis. C) The

261 mean advantage of amino acids constrained to use GC rather than constrained to use AT is not
262 distinguishable in size between the first and second codon positions. Y-axis gives the difference in the
263 two weighted means of marginal fitness effects from A) and B). Error bars represent 95% confidence
264 intervals on the difference between the means (calculated as difference +/- $t_{crit} \times se$), where $t_{crit} \approx 2.1$ is
265 the critical value of the t-statistic with the appropriate degrees of freedom. Weighted Welch's t-test
266 statistic and the corresponding standard error of the difference in means were calculated using the
267 "wtd.t.test" function from the "weights" R package, version 1.0.1.

268

269 Our fitness predictor is validated on random N-terminal tags

270 In order to validate our fitness predictor against an external dataset, we tested the ability of our amino-
271 acid-only model to predict the fitness effects of shorter random peptides present as N-terminal tags.
272 Goodman et al. (2013) tagged the N-prime end of green fluorescent protein (GFP) with 137 different
273 short random sequences (11 amino acids long), allowing random peptide expression level to be
274 measured via fluorescence. Frumkin et al. (2017) measured the fitness effects of these random peptide-
275 tagged GFPs in *E. coli* using FitSeq (Li, et al. 2018). For 89 of them, Frumkin et al. (2017) were able to
276 calculate a "fitness residual" based on the deviation from the fitness expected from the level of GFP
277 expression. Note that this fitness residual combines the cost of inefficient expression (as per the
278 intention of these papers) with the fitness effect of the peptide itself. Frumkin et al. (2017) interpreted
279 the low fitness residuals of hydrophobic and expensive-to-synthesize amino acids as caused by low
280 expression efficiency. However, these findings are consistent with our own estimates of direct peptide
281 effects, as hydrophobic amino acids tend to be order-prone (Linding, et al. 2004; Angyan, et al. 2012),
282 and amino acid size is highly correlated with synthesis cost in *E. coli* (Pearson's correlation coefficient =
283 0.85, $P = 2 \times 10^{-6}$, cost for amino acid synthesis in *E. coli* taken from (Akashi and Gojobori 2002)). Indeed,
284 our fitness predictions were significantly correlated with their fitness residuals (fig. 5). The consistency
285 between our results and the findings of Frumkin et al. (2017), who used a different dataset and
286 methodology, provide an external validation of the accuracy of our fitness predictor, even in cases
287 where a random segment makes up only a small portion of a total protein.

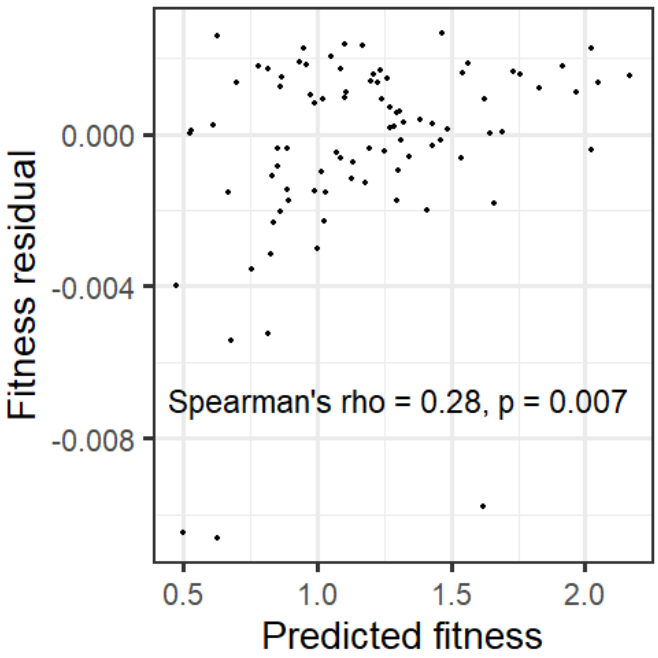


Fig. 5. **Fitness predictions trained on the random peptides of Neme et al. (2017) also work for short random tags attached to the N-terminus of GFP**. Predicted fitness comes from our amino acid composition-only mixed model. "Fitness residuals" of N-terminal tags are from Frumkin et al. (2017), and represent the difference between the fitness of the construct and the expected fitness from expression level. *n* = 89.

Discussion

We found that amino acid composition is the primary predictor of the fitness effects of random peptides expressed in *E. coli*. Amino acids that are small and promote intrinsic structural disorder predict high fitness. Amino acids that are good for random peptides in *E. coli* are enriched in young proteins in animals, suggesting that harm is avoided in similar ways across diverse lineages. What is truly astonishing is that simply knowing how many of which amino acids are present in a single highly expressed random peptide can account for 15% of the variance in fitness. This indicates both the success of our statistical method for minimizing the noise in our fitness estimates, and that mere amino acid content without amino acid order can be highly informative of peptide properties.

Most studies of random peptides have focused on finding peptides that have specific binding or function (e.g. Kaiser, et al. 1987; Keefe and Szostak 2001; Frulloni, et al. 2009). Some were motivated as proof-of-

305 concept that random peptides can exhibit properties of native proteins, such as folding (Davidson and
306 Sauer 1994; Chiarabelli, et al. 2006; LaBean, et al. 2011) and being soluble (Prijambada, et al. 1996).
307 Others focus on how to increase the percentage of native-like random peptides, e.g. by showing that
308 more hydrophilic random peptide libraries have a higher percentage of stable and soluble peptides
309 (Davidson, et al. 1995). Our work has a different intent, identifying properties that make a peptide less
310 likely to be harmful. Neme et al.'s (2017) experiment was perfect for this purpose because it used a
311 large library of peptides with diverse properties, and selected under relatively permissive conditions,
312 simply for relative growth rates in the presence of peptide expression. In contrast, a study design such
313 as that of Knopp et al. (2019), who selected random peptides that rescue viability in the presence of
314 antibiotics, is less useful for our purposes.

315 Having a higher proportion of random peptides do no harm is expected to increase the success rate of
316 future screens for peptide with specific properties. Nucleotide sequences with high %GC content tend to
317 encode peptides with more benign fitness effects, suggesting that higher %GC should be used in future
318 random peptide libraries. However, very high GC content will yield low complexity sequences, which our
319 predictor has not been trained on. The marginal fitness effects of each amino acid might be different in
320 this very different context.

321 While the library used by Neme et al. (2017) was designed to have equal frequencies of each nucleotide
322 in the random region, and thus 50% GC content, the over two million random peptides that had at least
323 one sequencing read had a GC content of ~59% in their random portion. The mean GC content of the
324 peptide clusters we analyzed (see Materials and Methods) was similar, at ~58%, with higher fitness
325 peptides within this group having higher %GC, as discussed in the Results. This might be because many
326 lower GC content sequences were so harmful that lineages that carried them went extinct prior to
327 detection via sequencing. Note that it might also reflect a bias toward GC in sequencing methods
328 (Benjamini and Speed 2012; Choudhari and Grigoriev 2017) – a bias that affects all timepoints equally
329 and so should not affect our fitness estimates.

330 Long et al. (2018) proposed that there is direct selection for high GC content, as evidenced in part by a
331 preference for amino acids with G or C at the second position of codons, in excess of that predicted from
332 mutation accumulation experiments. Our findings cannot exclude this hypothesis, but show stronger
333 selection on amino acid composition, selection that is capable of driving increased GC content in coding
334 regions as a correlated trait. In intergenic regions, elevated %GC is likely driven mostly by GC-biased
335 gene conversion. However, if driven in part by selection, it is possibly not just on GC content, but also on

336 peptides from non-coding regions translated by error (Rajon and Masel 2011; Wilson and Masel 2011).
337 Selection on translation errors is for example strong enough to shape non-coding sequences beyond
338 stop codons in *Saccharomyces cerevisiae* (Kosinski and Masel 2020).

339 The fact that IUPred, CamSol, and Tango have predictive power only when amino acid composition is
340 not controlled for suggests that both predictors serve here primarily as summary statistics of amino acid
341 content. Metrics like hydrophobic clustering, which are independent of amino acid composition by
342 construction, do not add predictive power.

343 While driven by amino acid composition, our findings still strongly support the hypothesis that peptides
344 with low structural disorder tend to be harmful. Disorder-promoting amino acids may help a peptide
345 remain soluble even if unfolded. Small amino acids also tend to be benign, perhaps because they are
346 hydrophobic enough to promote some amount of folding but flexible enough to avoid too much
347 hydrophobic residue exposure.

348 Our findings suggest that the easiest way to avoid harm is through disorder and small size, but do not
349 rule out other, more order-prone strategies. Indeed, BCS4, a de novo evolved protein in *Saccharomyces*
350 *cerevisiae*, has a hydrophobic core and is capable of folding (Bungard, et al. 2017). Vakirlis et al. (2020)
351 found that de novo proteins can emerge as transmembrane proteins, which need to be lipid soluble,
352 presumably requiring different harm-avoidance strategies than peptides that are located in the cytosol.

353 Plants have different trends in amino acid composition change over time than animals, with young
354 genes seeming to prefer readily available amino acids, rather than amino acids that promote intrinsic
355 structural disorder (James, et al. 2020). Nevertheless, our finding of consistency between what is benign
356 in *E. coli* and what is benign in animals suggests a deep concordance in what makes a peptide harmful
357 between two apparently disparate branches of life. The forces that drive protein birth therefore appear
358 to share a key similarity between bacteria and Animalia. To modify Monod's famous quote, what is
359 harmful in *E. coli* is also harmful in elephants, but not necessarily in eucalyptus.

360 Methods

361 Data retrieval

362 Neme et al. (2017) performed seven experiments where *E. coli* lineages, each with a plasmid containing
363 a unique random peptide, were grown and tracked using deep DNA sequencing. We downloaded

sequencing counts from Dryad at http://dx.doi.org/10.5061/dryad.6f356, and obtained amino acid and nucleotide sequences directly from Rafik Neme. Experiment 7 was by far the largest, containing all the peptides that the other six experiments classified as "increasing" or "decreasing," and more. We thus chose to focus on experiment 7 in our work. Experiment 7 consists of the numbers of reads of each random peptide sequence in 5 replicate populations of *E. coli* at 4 time points. We summed across all 5 replicates to obtain a total number of reads for each polypeptide at each time point. Following Neme et al. (2017), we took the 1061 peptides out of over one million that had ≥5 reads across all 5 replicates of experiment 7. We further excluded the six peptides that, while meeting the criterion of ≥5 reads, had all of those reads at the same timepoint, leaving 1055 peptides for analysis.

Estimating lineage fitness from random peptide sequencing counts

The expected number of reads λ_{it} of peptide i at times t=1,2,3,4 was modeled as:

$$\lambda_{it} = N_t p_{i0} \prod_{k=1}^{t} \frac{\omega_i}{W_{k-1}},$$

where N_t is the observed total number of reads, p_{i0} is the initial frequency of peptide i at the beginning of the experiment (prior to the round of selection used to produce the first measured timepoint $t = 1$), $\frac{\omega_i}{W_t}$ is the fitness of bacteria with peptide i at time t (i.e. their propensity to contribute to the next time point), and W_k is population mean fitness at time k.

The likelihoods of observed peptide counts were estimated from this expectation and two different error models. A Poisson distribution, which captures sampling error alone, was used to generate our initial estimates of p_{io}, ω_i, and W_k (collectively yielding λ_{it}) because it is analytically tractable. Under a Poisson error function, the likelihood of observing n_{it} reads of peptide i at time t is

$$f_{Poiss}(n_{it}|\lambda_{it}) = \frac{\lambda_{it}^{n_{it}} e^{-\lambda_{it}}}{n_{it}!}.$$

To also capture variance inflation κ due to PCR amplification, we used a negative binomial distribution in the Polya form:

$$f_{NBP}(n_{it}|\, \lambda_{i,t}, \kappa) = \left(\frac{\Gamma\left(n_{it} + \frac{\lambda_{i,t}}{\kappa - 1}\right)}{n_{it}!\, \Gamma\left(\frac{\lambda_{i,t}}{\kappa - 1}\right)} \right) \left(\frac{1}{\kappa}\right)^{\frac{\lambda_{i,t}}{\kappa - 1}} \left(1 - \frac{1}{\kappa}\right)^{n_{it}}$$

388 where $\Gamma(\cdot)$ is the gamma function. We used the initial estimates of p_{io}, ω_i, and W_k to numerically fit
389 the negative binomial model. For the specifics of fitting the Poisson and negative binomial models, see
390 Supporting Information. Weights were calculated, for use in downstream linear models, from this
391 likelihood inference procedure, as the inverse of Fisher information (see Supporting Information).

392 An existing software package for estimating lineage fitness from sequencing counts is Fit-Seq (Li, et al.
393 2018), which captures the amplification of PCR error through a more sophisticated distribution for the
394 number of reads that is derived in the supplementary information of Levy et al. (2015). However, Fit-Seq
395 assumes that mean fitness is a simple average of all measured lineages' fitness, requiring all individuals
396 to be tagged and measured. But Neme et al.'s (2017) experiment included lineages carrying an empty
397 plasmid, i.e. with the selectable marker but no random peptide. Worse, the proportion of cells with an
398 empty vector can be presumed to increase over time. In the absence of a reliable way to quantify these,
399 we instead consider mean population fitness over time to be a set of independent parameters to be
400 fitted.

401 Clustering non-independent sequences

402 Upon visual inspection, we found that some peptide sequences were extremely similar, with only one or
403 two amino acid differences; these data points will not contain independent information about the
404 relationship between sequence and fitness. To account for non-independence, we clustered peptides by
405 their Hamming distance, and either took only the peptide whose fitness had the highest weight within
406 its cluster, or took weighted means within clusters, or included cluster in our regression models as a
407 random effect term. Single-link clustering with Hamming distance cutoffs of 6 to 29 amino acids all
408 produced an identical set of 646 clusters for our 1055 peptides. The largest cluster had 228 random
409 peptides, and the second largest had only 13. The vast majorities of clusters contained only 1 sequence
410 (Dataset S1). A few peptides had mutations in their non-random regions; these mutations were counted
411 in our Hamming distance measurements.

412 Such similar sequences are highly unlikely to arise by chance if the peptides were truly random; $20^{50} \approx$
413 10^{65} peptides are possible, far more than the ~2×10^6 observed. Because we analyze only peptides with
414 at least 5 reads, replicated sequencing error is an unlikely cause. We see the same nearly-identical
415 sequences appearing in every experimental replicate, suggesting either that mutations occurred during
416 Neme et al.'s (2017) initial growth phase, or that the "random" peptides synthesized for the experiment

417 are not entirely random. We note that construction of the “random” peptide library involved ligations of
418 a smaller set of “seed” sequences, introducing non-randomness at this stage.

419 Predictors of fitness

420 All peptides are exactly 65 amino acids long with 50 amino acids of random sequence, so there was no
421 need to control for length.

422 *GC content*

423 Many amino acid sequences mapped to several possible nucleotide sequences, as part of the same
424 problem of mutation or non-random construction discussed above. To calculate one GC content for each
425 random peptide, we calculated a simple average of GC content across all the nucleotide sequences in
426 the dataset that map to the peptide with the largest weight in the cluster.

427 To calculate GC content for the over two million peptides with at least one sequencing read, we took a
428 simple average of the GC content from the random portion of the peptides.

429 *Disorder*

430 IUPred (Dosztányi, et al. 2005; Meszaros, et al. 2018) returns an ISD score between zero and one for
431 each amino acid in a sequence, with higher scores indicating greater intrinsic disorder. To calculate an
432 ISD score for each random peptide, we took the average of the scores for the whole sequence (i.e.
433 including non-random parts). We used a square root transform because it produced a more linear
434 relationship with fitness than no transform. IUPred2 was used for all analyses except ΔISD, which used
435 the original IUPred program – differences between the two are minimal (Meszaros, et al. 2018).

436 *Aggregation propensity*

437 Tango (Fernandez-Escamilla, et al. 2004; Linding, et al. 2004; Rousseau, et al. 2006) returns an
438 aggregation score for each amino acid in a sequence. At least five sequential amino acids with a score
439 greater than or equal to five indicates an aggregation-prone region. We scored peptide aggregation
440 propensity as the number of amino acids within regions scored as aggregation-prone, including
441 contributions from non-random regions.

Solubility

CamSol (Sormanni, et al. 2015) returns a solubility score for each amino acid in a sequence, as well as a simple average of all scores for a sequence, which CamSol calls a "solubility profile." We used the solubility profile of the full sequences, including non-random regions.

Hydrophobic clustering

We used a Perl script developed by Scott Foy ("calculate_hydrophobic_dispersion.pl" at https://github.com/MaselLab/A-shift-in-aggregation-avoidance-strategy-marks-a-long-term-direction-to-protein-evolution/tree/master/script) to calculate Irback's hydrophobicity dispersion metric (Irback, et al. 1996; Irback and Sandelin 2000). This metric splits a sequence into blocks of equal length and calculates dispersion as the variance of the hydrophobicity per block by the mean hydrophobicity per block, normalized to account for differences in mean hydrophobicity and length. We used blocks of 6 amino acids, considered I, L, M, F, and V hydrophobic, and scored whole peptide sequences including non-random regions.

Amino acid composition

We scored amino acid composition by counting frequencies among the 50 amino acids in the random portion of each peptide.

The values for all the above predictors for each peptide are listed in Dataset S1.

Statistics

All statistical tests were carried out in R version 3.6.3 (R Core Team 2019), with figures generated using "ggplot2" (Wickham 2016). Weighted linear mixed models were implemented using the "lmer" function from the "lme4" package (Bates, et al. 2015), with cluster as a random effect. See Supporting Information for details, including justification of a log-transform for fitness. When R^2 values were needed, we instead averaged peptides within the same cluster into a combined datapoint, allowing us to avoid the use of random effect term, and took R^2 values from the base R "lm" function.

Code availability

All code is available on GitHub at https://github.com/MaselLab/RandomPeptides.

468 ## Acknowledgements

469 This work was supported by the John Templeton Foundation (39667, 60814) and the National Institutes
470 of Health (GM-104040, T32GM-008659, T32GM-084905). We thank Rafik Neme and Diethard Tautz for
471 sharing their data with us and for graciously answering all our questions regarding their analyses, Dvir
472 Schirman and Tzachi Pilpel for sharing their data with us, Joe Watkins for helpful discussions about our
473 likelihood estimation procedure, and Catherine Weibel for driving the GC content analysis forward.

474 ## Works cited

475 Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of
476 Escherichia coli and Bacillus subtilis. Proceedings of the National Academy of Sciences of the
477 United States of America 99:3695-3700.
478 Alba MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the
479 age of genes. BMC Evolutionary Biology 7:53.
480 Angyan AF, Perczel A, Gaspari Z. 2012. Estimating intrinsic structural preferences of *de novo* emerging
481 random-sequence proteins: Is aggregation the main bottleneck? FEBS Letters 586:2468-2472.
482 Bates D, Maechler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. Journal
483 of Statistical Software 67:1-48.
484 Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput
485 sequencing. Nucleic Acids Research 40:e72.
486 Bungard D, Copple JS, Yan J, Chhun JJ, Kumirov VK, Foy SG, Masel J, Wysocki VH, Cordes MHJ. 2017.
487 Foldability of a natural de novo evolved protein. Structure 25:1687-1696.
488 Chen SCC, Chuang TJ, Li WH. 2011. The Relationships Among MicroRNA Regulation, Intrinsically
489 Disordered Regions, and Other Indicators of Protein Evolutionary Rate. Molecular Biology and
490 Evolution 28:2513-2520.
491 Chiarabelli C, Vrijbloed JW, De Lucrezia D, Thomas RM, Stano P, Polticelli F, Ottone T, Papa E, Luisi PL.
492 2006. Investigation of de novo totally random biosequences Part II: On the folding frequency in
493 a totally random library of de novo proteins obtained by phage display. Chemistry & Biodiversity
494 3:840-859.
495 Chiti F, Dobson CM. 2017. Protein misfolding, amyloid formation, and human disease: A summary of
496 progress over the last decade. Annual Review of Biochemistry 86:27-68.

497 Choudhari S, Grigoriev A. 2017. Phylogenetic heatmaps highlight composition biases in sequenced reads.
498 Microorganisms 5:4.
499 Davidson AR, Lumb KJ, Sauer RT. 1995. Cooperatively folded proteins in random sequence libraries.
500 Nature Structural Biology 2:856-864.
501 Davidson AR, Sauer RT. 1994. Folded proteins occur frequently in libraries of random amino-acid
502 sequences. Proceedings of the National Academy of Sciences of the United States of America
503 91:2146-2150.
504 Dosztányi Z, Csizmok V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid
505 composition discriminates between folded and intrinsically unstructured proteins. Journal of
506 Molecular Biology 347:827-839.
507 Dubrey S, Ackermann E, Gillmore J. 2015. The transthyretin amyloidoses: advances in therapy.
508 Postgraduate Medical Journal 91:439-448.
509 Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-
510 dependent and mutational effects on the aggregation of peptides and proteins. Nature
511 biotechnology 22:1302-1306.
512 Foy SG, Wilson BA, Bertram J, Cordes MHJ, Masel J. 2019. A shift in aggregation avoidance strategy
513 marks a long-term direction to protein evolution. Genetics 211:1345-1355.
514 Frulloni L, Lunardi C, Simone R, Dolcino M, Scattolini C, Falconi M, Benini L, Vantini I, Corrocher R,
515 Puccetti A. 2009. Identification of a Novel Antibody Associated with Autoimmune Pancreatitis.
516 New England Journal of Medicine 361:2135-2142.
517 Frumkin I, Schirman D, Rotman A, Li F, Zahavi L, Mordret E, Asraf O, Wu S, Levy SF, Pilpel Y. 2017. Gene
518 Architectures that Minimize Cost of Gene Expression. Molecular Cell 65:142-153.
519 Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial
520 genes. Science 342:475-479.
521 Heames B, Schmitz J, Bornberg-Bauer E. 2020. A continuum of evolving de novo genes drives protein-
522 coding novelty in Drosophila. Journal of Molecular Evolution 38:382-398.
523 Irback A, Peterson C, Potthast F. 1996. Evidence for nonrandom hydrophobicity structures in protein
524 chains. Proceedings of the National Academy of Sciences of the United States of America
525 93:9533-9538.
526 Irback A, Sandelin E. 2000. On hydrophobicity correlations in protein chains. Biophysical Journal
527 79:2252-2258.

528 Jacobson DR, Pastore RD, Yaghoubian R, Kane I, Gallo G, Buck FS, Buxbaum JN. 1997. Variant-sequence
529 transthyretin (isoleucine 122) in late-onset cardiac amyloidosis in black Americans. New England
530 Journal of Medicine 336:466-473.
531 James JE, Willis SM, Nelson PG, Weibel C, Kosinski LJ, Masel J. 2020. Universal and taxon-specific trends
532 in protein sequences as a function of age. bioRxiv:2020.2003.2026.010728.
533 Kaiser CA, Preuss D, Grisafi P, Botstein D. 1987. Many random sequences functionally replace the
534 secretion signal sequence of yeast invertase. Science 235:312-317.
535 Keefe AD, Szostak JW. 2001. Functional proteins from a random-sequence library. Nature 410:715-718.
536 Knopp M, Andersson DI. 2018. No beneficial fitness effects of random peptides. Nature Ecology &
537 Evolution 2:1046-1047.
538 Knopp M, Gudmundsdottir JS, Nilsson T, Konig F, Warsi O, Rajer F, Adelroth P, Andersson DI. 2019. De
539 novo emergence of peptides that confer antibiotic resistance. Mbio 10:e00837-00819.
540 Kosinski LJ, Masel J. 2020. Readthrough errors purge cryptic sequences, facilitating the birth of coding
541 sequence. Molecular Biology and Evolution 37:1761–1774.
542 Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology
543 with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology
544 305:567-580.
545 LaBean TH, Butt TR, Kauffman SA, Schultes EA. 2011. Protein folding absent selection. Genes 2:608-626.
546 Larsson SC, Markus HS. 2017. Branched-chain amino acids and Alzheimer's disease: a Mendelian
547 randomization analysis. Scientific Reports 7:13604.
548 Levy ED, De S, Teichmann SA. 2012. Cellular crowding imposes global constraints on the chemistry and
549 evolution of proteomes. Proceedings of the National Academy of Sciences of the United States
550 of America 109:20461-20466.
551 Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. 2015. Quantitative evolutionary
552 dynamics using high-resolution lineage tracking. Nature 519:181-186.
553 Li F, Salit ML, Levy SF. 2018. Unbiased fitness estimation of pooled barcode or amplicon sequencing
554 studies. Cell Systems 7:521-525.
555 Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L. 2004. A comparative study of the relationship
556 between protein structure and β-aggregation in globular and intrinsically disordered proteins.
557 Journal of Molecular Biology 342:345-353.

558 Liu HX, Zhang RS, Yao XJ, Liu MC, Hu ZD, Fan BT. 2004. Prediction of the isoelectric point of an amino
559 acid based on GA-PLS and SVMs. Journal of Chemical Information and Computer Sciences
560 44:161-167.
561 Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss C, Stone
562 C, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. Nature
563 Ecology & Evolution 2:237-240.
564 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq
565 data with DESeq2. Genome Biology 15:550.
566 Maurer-Stroh S, Debulpaep M, Kuemmerer N, de la Paz ML, Martins IC, Reumers J, Morris KL, Copland A,
567 Serpell L, Serrano L, et al. 2010. Exploring the sequence determinants of amyloid structure using
568 position-specific scoring matrices. Nature Methods 7:237-242.
569 McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-
570 coding genes in eukaryotic evolutionary innovation. Philosophical Transactions of the Royal
571 Society B-Biological Sciences 370:20140332.
572 Meszaros B, Erdos G, Dosztanyi Z. 2018. IUPred2A: context-dependent prediction of protein disorder as
573 a function of redox state and protein binding. Nucleic Acids Research 46:W329-W337.
574 Monod J, Jacob F. 1961. General conclusions - teleonomic mechanisms in cellular metabolism, growth,
575 and differentiation. Cold Spring Harbor Symposia on Quantitative Biology 26:389-401.
576 Moyers BA, Zhang JZ. 2016. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth
577 in Genome Evolution. Molecular Biology and Evolution 33:1245-1256.
578 Moyers BA, Zhang JZ. 2015. Phylostratigraphic Bias Creates Spurious Patterns of Genome Evolution.
579 Molecular Biology and Evolution 32:258-267.
580 Neme R, Amador C, Yildirim B, McConnell E, Tautz D. 2017. Random sequences are an abundant source
581 of bioactive RNAs or peptides. Nature Ecology & Evolution 1:0127.
582 Prijambada ID, Yomo T, Tanaka F, Kawama T, Yamamoto K, Hasegawa A, Shima Y, Negoro S, Urabe I.
583 1996. Solubility of artificial proteins with random sequences. FEBS Letters 382:21-25.
584 R Core Team. 2019. R: A language and environment for statistical computing: R Foundation for Statistical
585 Computing.
586 Rajon E, Masel J. 2011. Evolution of molecular error rates and the consequences for evolvability.
587 Proceedings of the National Academy of Sciences of the United States of America 108:1082-
588 1087.

589 Rousseau F, Schymkowitz J, Serrano L. 2006. Protein aggregation and amyloidosis: confusion of the
590 kinds? Current Opinion in Structural Biology 16:118-126.
591 Sormanni P, Aprile FA, Vendruscolo M. 2015. The CamSol Method of Rational Design of Protein Mutants
592 with Enhanced Solubility. Journal of Molecular Biology 427:478-490.
593 Theillet F-X, Kalmar L, Tompa P, Han K-H, Selenko P, Dunker AK, Daughdrill GW, Uversky VN. 2013. The
594 alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in
595 intrinsically disordered proteins. Intrinsically Disordered Proteins 1:e24360.
596 Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent accessibilites
597 of residues in proteins. PLoS One 8:e80635.
598 Tretyachenko V, Vymetal J, Bednarova L, Kopecky V, Hofbauerova K, Jindrova H, Hubalek M, Soucek R,
599 Konvalinka J, Vondrasek J, et al. 2017. Random protein sequences can form defined secondary
600 structures and are well-tolerated in vivo. Scientific Reports 7:15449.
601 Tsai J, Taylor R, Chothia C, Gerstein M. 1999. The packing density in proteins: Standard radii and
602 volumes. Journal of Molecular Biology 290:253-266.
603 Vakirlis N, Acar O, Hsu B, Coelho NC, Van Oss SB, Wacholder A, Medetgul-Ernar K, Bowman RW, Hines
604 CP, Iannotta J, et al. 2020. De novo emergence of adaptive membrane proteins from thymine-
605 rich genomic sequences. Nature Communications 11:781.
606 Vakirlis N, Carvunis A-R, McLysaght A. 2020. Synteny-based analyses indicate that sequence divergence
607 is not the main source of orphan genes. eLife 9:e53500.
608 Van Oss SB, Carvunis AR. 2019. *De novo* gene birth. PLoS Genetics 15:e1008160.
609 Vecchi G, Sormanni P, Mannini B, Vandelli A, Tartaglia GG, Dobson CM, Hartl FU, Vendruscolo M. 2020.
610 Proteome-wide observation of the phenomenon of life on the edge of solubility. Proceedings of
611 the National Academy of Sciences of the United States of America 117:1015-1020.
612 Weisman CM, Eddy SR. 2017. Gene Evolution: Getting Something from Nothing. Current Biology
613 27:R661-R663.
614 Wickham H. 2016. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag.
615 Willis S, Masel J. 2018. Gene birth contributes to structural disorder encoded by overlapping genes.
616 Genetics 210:303-313.
617 Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by the
618 preadaptation hypothesis of *de novo* gene birth. Nature Ecology & Evolution 1:0146.
619 Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes.
620 Genome Biology and Evolution 3:1245-1252.