



**HAL**  
open science

## Evolution is not Uniform Along Coding Sequences

Raphaël Bricout, Dominique Weil, David Stroebel, Auguste Genovesio,  
Hugues Roest Crolius

► **To cite this version:**

Raphaël Bricout, Dominique Weil, David Stroebel, Auguste Genovesio, Hugues Roest Crolius. Evolution is not Uniform Along Coding Sequences. *Molecular Biology and Evolution*, 2023, 40 (3), 10.1093/molbev/msad042 . hal-04051951

**HAL Id: hal-04051951**

**<https://hal.science/hal-04051951>**

Submitted on 30 Mar 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Evolution is not uniform along coding sequences

Raphaël Bricout<sup>1</sup>, Dominique Weil<sup>2</sup>, David Stroebel<sup>1</sup>, Auguste Genovesio<sup>1\*</sup>, Hugues Roest Crollius<sup>1\*</sup>

1. Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL ; 46 rue d'Ulm, 75005 Paris, France

2. Sorbonne Université, CNRS, Institut de Biologie Paris-Seine (IBPS), Laboratoire de Biologie du Développement, F-75005 Paris, France.

\* Corresponding authors: [auguste.genovesio@ens.psl.eu](mailto:auguste.genovesio@ens.psl.eu), [hrc@bio.ens.psl.eu](mailto:hrc@bio.ens.psl.eu)

## Abstract

Amino acids evolve at different speeds within protein sequences, because their functional and structural roles are different. Notably, amino-acids located at the surface of proteins are known to evolve more rapidly than those in the core. In particular, amino-acids at the N- and C-termini of protein sequences are likely to be more exposed than those at the core of the folded protein due to their location in the peptidic chain, and they are known to be less structured. Because of these reasons, we would expect that amino-acids located at protein termini would evolve faster than residues located inside the chain. Here we test this hypothesis and found that amino acids evolve almost twice as fast at protein termini compared to those in the centre, hinting at a strong topological bias along the sequence length. We further show that the distribution of solvent-accessible residues and functional domains in proteins readily explain how structural and functional constraints are weaker at their termini, leading to the observed excess of amino-acid substitutions. Finally, we show that the specific evolutionary rates at protein termini may have direct consequences, notably misleading *in silico* methods used to infer sites under positive selection within genes. These results suggest that accounting for positional information should improve evolutionary models.

## Introduction

Since the early days of structural biology, protein termini have somewhat unsurprisingly often been found at the protein surface (Kendrew et al. 1960). Today, tens of thousands of protein structures have been resolved (Berman et al. 2000), and protein termini have indeed quite consistently been found to be less structured and exposed outside of the protein core (Carugo 2011), often leading to difficulties in including them in crystals for X-ray

37 crystallography. It is now well established that residues located at the surface of a folded  
38 protein and accessible to the solvent evolve faster than those in the core of the protein  
39 (Moutinho et al. 2019). We can therefore ask if residues at the extremities of peptidic chains  
40 evolve faster than those in the central region of the sequence. Within proteins, evolutionary  
41 rates are already known to be heterogeneous, because they are influenced by a residue's  
42 implication in functional domains and by structural constraints in the folded protein (Echave  
43 et al. 2016). Accounting for such heterogeneity in models of molecular evolution is critical to  
44 accurately infer phylogenies and estimate cases of positive selection. Elaborate models have  
45 been developed to achieve this (Halpern and Bruno 1998), generally by estimating site-  
46 specific rates in a maximum likelihood framework employing Markov models of sequence  
47 evolution (Yang et al. 2000; Kosakovsky Pond and Frost 2005; Baele et al. 2021).  
48 However, the impact of the position of a given amino acid in the sequence relative to protein  
49 start and end on the rate of molecular evolution has received little attention so far, and is  
50 therefore not accounted for in such models.

51 We therefore tested the hypothesis that residues located at protein termini might evolve  
52 faster than residues in the core. Towards this, we measured the rates of amino-acid  
53 changing substitutions (non-synonymous, dN) and silent substitutions (synonymous, dS) at  
54 individual codons positions and averaged them over thousands of coding sequences (CDS)  
55 in animal or in plants. We found that rates of molecular evolution are almost twice as fast at  
56 CDS ends as in the centre. We next showed that structural constraints of the folded proteins,  
57 characterised by the relative solvent accessibility (RSA) of residues explain well why  
58 evolutionary rates differ depending on residue positions along the sequence. Finally, we  
59 showed that this faster rate of evolution at protein termini has probable consequence on the  
60 prediction of positive selection.

61

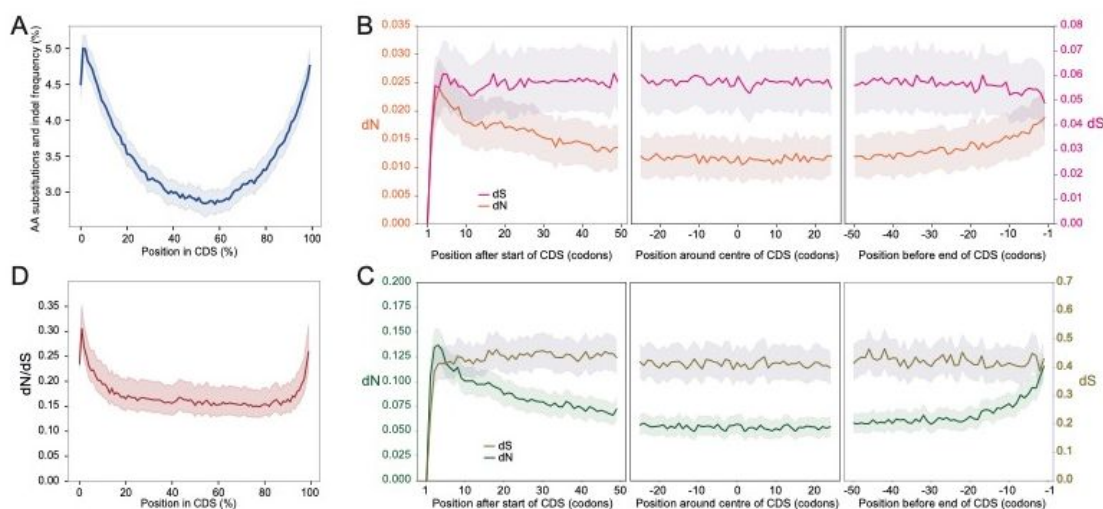
## 62 **Results**

### 63 **Evolution is faster at protein termini**

64 To examine the dependency between codon position in CDS and sequence variation during  
65 evolution, we first took a global view of multiple sequence alignments in 15,828 primate gene  
66 families to identify fixed mutations (substitutions, insertions and deletions) that took place in  
67 these sequences during the evolution of 26 primate species. The results showed a strong  
68 excess of such changes towards the sequence extremities (Fig. 1A), leading to a distinctive  
69 U-shaped pattern. To better understand this result, we computed separately the dN and dS  
70 rates in this collection of primate sequences. Computing rates of molecular evolution in  
71 eukaryote CDS is conventionally performed on multiple sequence alignments of CDS  
72 belonging to the same gene family. At each site considered independently, results rely on the  
73 number of aligned sequences at this position and on the small minority that show

74 substitutions. The robustness of measures at individual sites is therefore variable and  
 75 generally low. Here we computed average rates at specific codon positions but from  
 76 collections of CDS comprising thousands of gene families. This enabled us to concatenate  
 77 aligned codons from the same position sampled from each gene family. dN and dS rates  
 78 were then computed from virtual sequences composed exclusively of codons from the same  
 79 position in their respective CDS (Figure S1), thus providing strong and uniform statistical  
 80 power across sites.

81 Results showed that while the dS remains remarkably constant along the CDS length  
 82 (average dS=0.052), the dN increases significantly in the region spanning the first and last 50  
 83 codons (Fig. 1B). We observed a similar bias when computing dN and dS along the CDS of  
 84 6,459 plant (Fabids) gene families, which were subjected to an approximately 8-fold higher  
 85 divergence rate than primates (Fig. 1C). Effect sizes are major, with a more than two-fold  
 86 decrease in dN between the first 10 codons and the middle 50 codons (0.117 vs 0.054, a  
 87 54% decrease) while dS remains essentially constant (0.400 vs 0.413, a 3% increase). In  
 88 summary, the dN appears to be driving the distinctive U-shaped pattern of total substitutions  
 89 and dN/dS in gene CDS (Fig. 1D).



103 **Fig. 1. (A).** Frequency of amino acid substitutions, insertion and deletions computed in 15,828  
 104 primate multiple sequence (CDS) alignments (MSA), rescaled from 0 to 100% of the CDS  
 105 length. **(B).** Distribution of silent (dS) and non-synonymous (dN) substitution rates computed  
 106 at each codon position from random pairs of sequences sampled from 14,186 primate MSA  
 107 without alignment gaps and shown here for the first (left panel), middle (middle panel) and last  
 108 (right panel) 50 codons. **(C).** Same as in B but for pairs of CDS sampled from 6,459 MSA of  
 109 plant (Fabids) genes. **(D).** Distribution of dN/dS ratio for dN and dS values shown in B but  
 110 across the entire CDS length rescaled from 0 to 100%. In all panels the shaded area represents  
 111 the 95% confidence interval.

113 Computing substitutions in a multiple alignment of CDS is a multi-step process, with many  
114 potential sources of technical biases which could potentially explain this pattern (Schneider  
115 et al. 2009; Prosdocimi et al. 2012). We conducted a serie of experiments to exclude  
116 annotation errors, multiple alignment artefacts and compositional biases (Supplementary  
117 material, Fig. S2 and S3), showing that our observations were robust to controls designed to  
118 address possible technical artefacts in the process, from CDS annotation to substitution  
119 calculations.

120

### 121 **Evolution is faster for solvent-exposed amino acids**

122 We next examined biological or evolutionary explanations. We first eliminated the possibility  
123 that the increased dN at CDS extremities would be caused by a stronger local mutation rate  
124 because the dS, which would be much more sensitive to the mutation rate, is essentially  
125 constant along CDS length (Fig. 1B,C). We next reasoned that weaker negative selection at  
126 protein termini might be caused by weaker functional constraints. The functional role of  
127 proteins depends both on the structural architecture of the folded sequence and on specific  
128 domains present at key positions within the sequence. We investigated how structural  
129 constraints, or lack thereof, may influence evolutionary rates at protein termini. Within the  
130 core of a folded protein, amino-acids are involved in a tight network of interactions that  
131 shapes the molecular function and role of the macromolecule. Amino acids located at the  
132 surface of the folded protein are interacting with the solvent and are typically less important  
133 for the protein function. It is indeed well-established that evolutionary rates differ between  
134 residues depending on their solvent accessibility (Franzosa and Xia 2009; Ramsey et al.  
135 2011; Moutinho et al. 2019). The precise relationship between solvent accessibility,  
136 evolutionary rates and amino acid position along the sequence has however never been  
137 ascertained. Such investigation has long been impaired by the quasi-systematic absence of  
138 N- and C-terminal regions in protein structure files, because those were either poorly  
139 resolved in electron density maps, genetically modified or removed prior to protein  
140 production. To circumvent this issue, we used the recently released dataset of 22,613  
141 complete protein structures predicted by Alphafold (Jumper et al. 2021) from the human  
142 genome, which have been shown to provide remarkably accurate Relative Solvent  
143 Accessibility (RSA) estimates on individual proteins (Bæk and Kepp 2022). We computed the  
144 RSA of each residue for all proteins (Fig. S6) and we noted that RSA values follow a bimodal  
145 distribution, which coincide with a strong enrichment (RSA < 0.3) or depletion (RSA > 0.6) in  
146 PFAM functional domains. Computing the distribution along protein sequences of residues  
147 from these two categories, we discovered that solvent accessibility increases sharply at  
148 protein termini (Fig. 2A), consistent with these regions being largely unstructured. This could  
149 theoretically be caused by the existence of structures unknown to Alphafold in protein

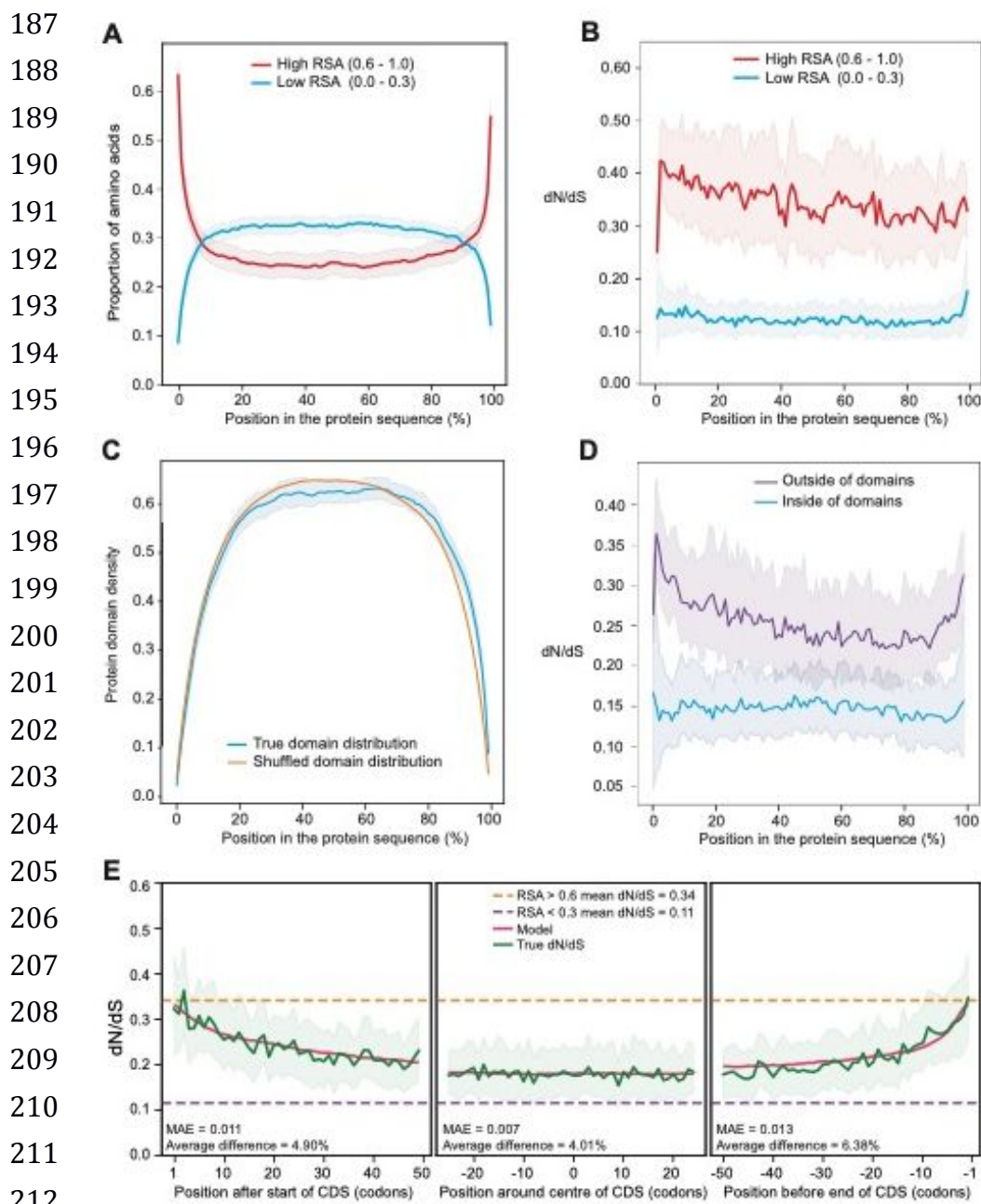
150 termini, resulting in the absence of predicted structure, but several reasons argue against  
151 this. First, this result is in agreement with a general pattern seen in structures obtained  
152 experimentally (Carugo 2011). Second, AlphaFold is able to properly fold proteins with  
153 unknown structure (Jumper et al. 2021). Therefore, the observed increased accessibility is  
154 consistent with weak structural constraints at protein termini (Ruff and Pappu 2021; Bæk and  
155 Kepp 2022; Wilson et al. 2022). Critically, the dN/dS rate is low and constant along protein  
156 length in sites with low accessibility, while it is elevated in highly accessible regions (Fig. 2B).  
157 In these controls, the marked increase in dN/dS at sequence extremities shown in Figure 1 is  
158 absent, indicating that solvent accessibility is likely a strong marker of the decrease in  
159 selective pressure observed in the N- and C-terminal region of proteins.

160

### 161 **Evolutionary rates correlate with protein domain density**

162 In order to better characterize these results, we analysed the contribution of protein domains  
163 in the observed evolutionary profile. Predicted protein domains indeed capture a large  
164 fraction of amino acids involved in structural and functional roles in protein sequences, and  
165 their prediction relies on sequence similarity and structural information (Wang et al. 2021).  
166 Both of these features make them also good proxies for sites under evolutionary constraints.  
167 We computed, for several predicted protein domains databases, the distribution of protein  
168 domains by calculating the frequency at which a given position along the protein sequence is  
169 found inside a domain. We found that domains are strongly depleted at protein termini (Fig.  
170 2C and fig. S4A-C), following a distribution consistent with a mechanical exclusion caused by  
171 the physical impossibility for a domain to overlap protein edges. The distribution of domains  
172 decreases sharply towards protein edges regardless of the length of the protein sequences  
173 (Fig. S4D-E), supporting a scenario where all proteins are similarly affected by a deficit of  
174 domain-induced evolutionary constraints at their edges. The dome-shaped distribution of  
175 domains mirrors the distinctive U-shaped distribution of the dN/dS ratio (Fig. 1A), consistent  
176 with our initial hypothesis that a depletion of domains at the edges of proteins would make  
177 them more permissive to non-synonymous changes and indels because of weaker selective  
178 constraints. To test this more directly, we distinguished codons that code for amino acids  
179 involved in a domain from those that do not, and computed the dN/dS for each category  
180 separately (Fig. 2D). In line with the above expectation, the dN/dS bias could not be  
181 observed when computed exclusively inside domains. Again, the difference in dN/dS  
182 behaviour is largely caused by the dN, since the dS remains constant both inside and outside  
183 of domains and is almost identical in both categories throughout the protein length (Fig. S5).  
184 These results are consistent with conclusions from the structure-based analysis, strongly  
185 supporting a model in which selective constraints are significantly weaker at protein edges.

186



213 **Fig. 2.** (A) The distribution of frequency of amino acids with high (red line) and low (blue line)  
 214 Relative Solvent Accessibility (RSA) computed by pCASA on 3D structure predicted by  
 215 AlphaFold on 22,613 human protein sequences, rescaled to 0-100% of the length. (B) dN/dS  
 216 computed on 7,144 sequences common to the AlphaFold and Ensembl primate CDS datasets,  
 217 where sites with high (red line) and low (blue line) RSA are distinguished. (C) The distribution  
 218 of protein domains from the PFAM database in 9,073 human proteins rescaled from 0-100%  
 219 of their length (blue line). The orange line shows the distribution of the same domains in  
 220 random non-overlapping positions in the same sequences. (D) dN/dS computed in 14,186  
 221 alignments from 26 primate genomes, where sites inside (blue line) and outside (purple line)  
 222 PFAM domains are distinguished. (E) Proposed model where the mean dN/dS in high (RSA >  
 223 0.6) and in low (RSA < 0.3) accessibility regions are weighted according to the percent of  
 224 codons in each RSA category. This model is compared to true observations (green line) from  
 225 data without signal peptides. The Mean Absolute Error (MAE) and percent average error  
 226 between the model and the true observations are indicated for each panel. In all panels the  
 227 shaded area represents the 95% confidence interval.  
 228

**229 Solvent exposure largely explains the increase in evolutionary rate at protein termini**

230 Because both RSA and functional domains correlate with dN/dS variation at protein termini,  
231 we designed a model to estimate their respective influence. The model takes as parameter a  
232 dN/dS value computed as an average across all domains, or across all low RSA ( $RSA < 0.3$ )  
233 or all high RSA ( $RSA > 0.6$ ) regions, and weighs it by the proportion of residues in the  
234 corresponding category at a given position (Supplementary Material). When applied solely  
235 with RSA as parameter, the model reproduced the observed dN/dS with remarkable  
236 accuracy in human proteins (mean difference between model and data = 5.09%, Figure  
237 S7A), suggesting that RSA is sufficient to explain the bias in average dN/dS along protein  
238 sequences. When the model uses only functional domains as parameter (Figure S7B), the fit  
239 to the data at protein termini is degraded compared to the RSA-only model (mean difference  
240 at both termini: 19%) and mean difference between model and observation is higher at 14%.  
241 The model combining both RSA and domain as parameters (Figure S7C) confirms the lower  
242 impact of domains in explaining the dN/dS shift, because the average difference with the  
243 data is the same as with the RSA-only model (Supplementary Material).

244 The RSA-only model, however, noticeably deviates from the observed data in the first 10  
245 codons at the N-terminus. We hypothesised that signal peptide, which are known to be highly  
246 variable in sequence (von Heijne 1985), might play a role in causing this deviation. Indeed,  
247 the shift largely disappeared when we removed the 15.8% of proteins with a N-terminal  
248 signal peptide (Fig. 2E). Signal peptides are therefore likely to provide additional relief from  
249 the evolutionary pressure measured in this small region, that is not accounted for with RSA.  
250 The impact of peptide signals is particularly notable when comparing the N-terminus, which  
251 contain some, and the C-terminus, which contain none (Figure S8A). Differences in average  
252 dN/dS distribution between the two protein ends are greatly reduced when signal peptide are  
253 excluded (Figure S8B).

254

**255 Faster evolution can be measured at the termini of individual proteins**

256 Until now, we measured the dN/dS bias at protein ends by averaging thousands of sites at  
257 each position. Is the bias also significant at the level of individual sequence alignments? This  
258 is important if evolutionary models applied to single gene families are likely to be affected. To  
259 address this, we computed a correlation between codon position and dN/dS for 7,800  
260 multiple sequence alignments, separately for the 50 codons at the beginning and at the end  
261 of CDS (Fig. 3A).

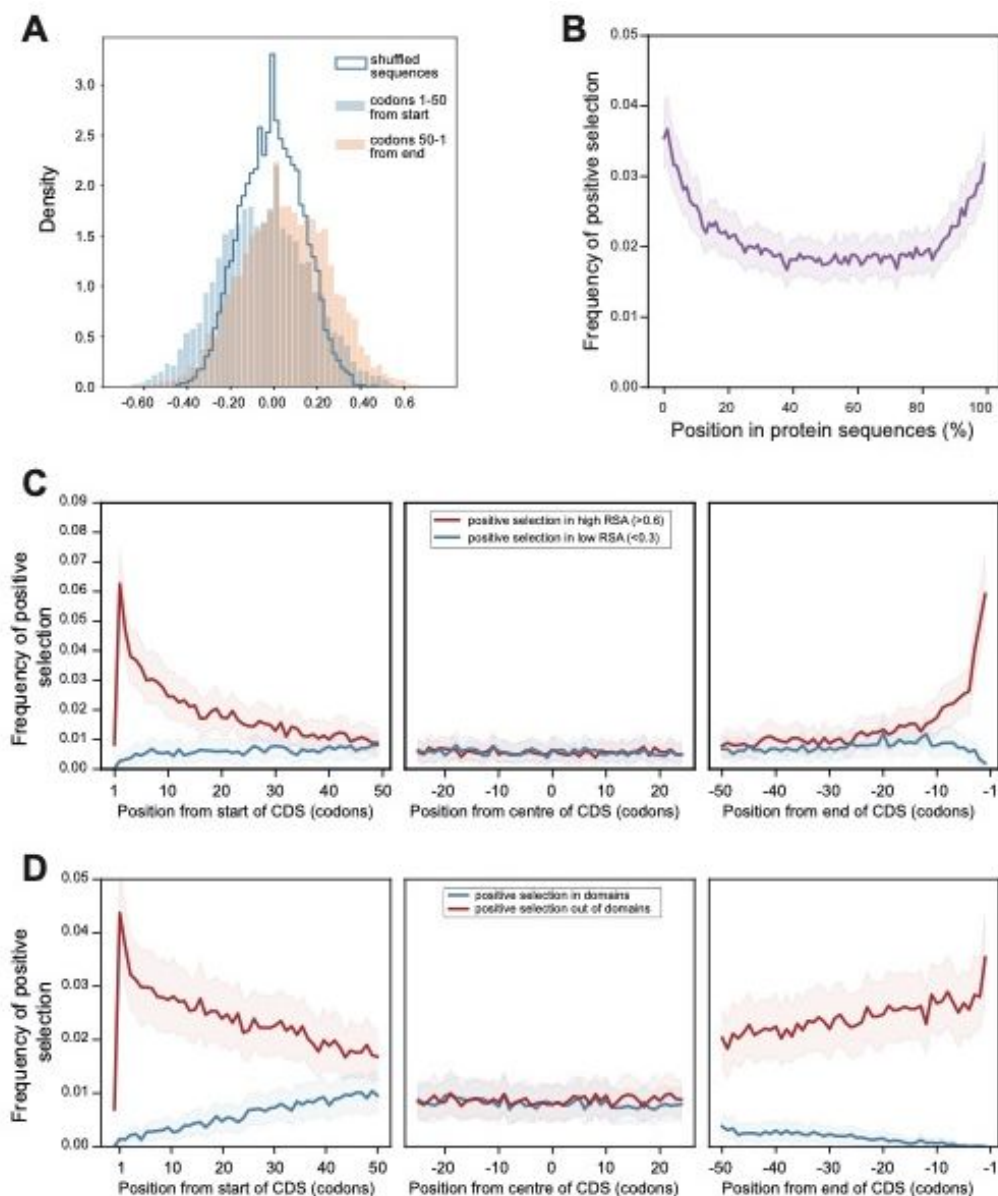
262

263

264

265





**Figure 3.** (A) Distribution of Pearson correlation coefficients between dN/dS values and their position in mouse CDS from 7,800 alignments of at least 3 rodent sequences. The distributions correspond either to the first (filled blue bars) or the last (filled orange bars) 50 codons. The blue line shows the distribution for the first and last 50 codons in the same mouse sequences with shuffled positions. (B) Distribution of the frequency of sites under positive selection in 13,301 rodent CDS rescaled to 0-100% of their length (C) Distribution of the frequency of sites under positive selection in the first, middle and last 50 codons of 8,371 rodent CDS, distinguishing sites with high (red line) and low (blue line) RSA. (D) Distribution of the frequency of sites under positive selection in the first, middle and last 50 codons of 13,020 rodent CDS, distinguishing sites inside (blue line) and outside (red line) PFAM domains.

306

307 Compared to a control where amino acid positions were randomised, the distribution of  
308 correlation coefficients were significantly shifted towards negative ( $p$ -value= $3.10^{-63}$ ; t-test)  
309 and positive ( $p$ -value =  $7.10^{-79}$ ; t-test) values for the start and end regions of CDS,  
310 respectively. This reflects the existence of a measurable increase in dN/dS towards CDS  
311 edges, even in individual sequences. We propose that this pattern is caused by the same  
312 factors as for the average sequences analysed previously (Fig. 1B), the biased solvent  
313 accessibility and, to a lesser extent, the domain distribution.

314

### 315 **Faster evolutionary rates at termini confound tests of positive selection**

316 These results immediately raised questions for the identification of positive selection in protein  
317 sequences, because a significantly elevated dN compared to some background rate is  
318 generally taken as evidence of adaptive changes (Nei and Kumar 2000). The U-shaped bias  
319 in dN observed in our study suggests that relaxation of constraints at protein edges might  
320 confound tests of positive selection. To investigate this, we estimated sites under positive  
321 selection in a set of 13,301 rodent gene trees using a site model (methods). We found that  
322 sites estimated under positive selection are strikingly enriched at protein edges (Fig. 3B), and  
323 that this enrichment can be specifically attributed to residues with high solvent accessibility  
324 (Figure 3C) and to sites located outside of functional domains (Fig. 3D). Notably, the same  
325 bias towards CDS extremities can be observed in several published scans for positive selection  
326 (Figure S9). Interestingly, while this bias is conspicuous for sites estimated to have been  
327 subject to positive selection using bioinformatic methods, it is not the case for experimentally  
328 verified sites, although our compilation of cases for this category is too small to draw general  
329 conclusions.

330

## 331 **Discussion**

332 We revealed a pattern of evolutionary rate along CDS that has so far remained concealed: the  
333 average amino acid substitution rate (dN) increases towards the extremities of the sequence.  
334 We found this pattern by assembling observations that were sometimes known quantitatively or  
335 intuitively in the field but never connected with respect to codon or amino acid positions in  
336 sequences. This pattern provides insights into the elusive mechanism driving evolutionary rate  
337 heterogeneities (Echave et al. 2016). First noted by Perutz and colleagues on haemoglobin  
338 (Perutz et al. 1965) and confirmed by many studies since then, protein surfaces evolve faster  
339 than their interior, where structural constraints, residue interactions and functional sites are  
340 most enriched and solvent accessibility is lower (Franzosa and Xia 2009; Ramsey et al. 2011).  
341 Attempts at explaining evolutionary rate heterogeneity have thus mainly focused on this  
342 paradigm, that structural constraints governed by complex spatial interactions create a range

343 of selective pressures on amino acids, but that are hard to predict from the sequence itself.  
344 Here, we show that the average dN/dS varies along the length of protein sequences following  
345 a U-shape distribution, and that a simple model based on high and low RSA values explains >  
346 93% of this distribution over the extremities of protein sequences, with a small contribution by  
347 signal peptides in the first codons at the N-terminus.

348 It is worth noting that in the present study, molecular rates are not dependent on a substitution  
349 model, as they do not rely on ancestral state inferences in CDS, and molecular rates are  
350 computed on MSA without gaps, which are known to introduce biases. This may explain that  
351 while we see the same increase in dN/dS at the 5' end of CDS as in a recent study on  
352 *Drosophila* genes (Davydov et al. 2019), we do not see the same decrease in substitution rates.  
353 On a related matter, we wish to point out that our finding that the average dS is constant along  
354 CDS length (Figs. 1B, 1C) should not be interpreted as meaning that dS does not vary or is  
355 not subject to site heterogeneity in individual genes, as this has been shown in many previous  
356 studies (Rubinstein and Pupko 2012).

357 Protein domains in the PFAM database are unlikely to cover all biological domains (Sammut  
358 et al. 2008), even in well-studied vertebrate genomes. That human protein sequences still  
359 contain non-annotated domains (false negatives) would be consistent with the observation that  
360 the dN/dS measured outside domains is lower in the centre of proteins (Fig. 2D), as if some  
361 feature (e.g. non-annotated domains) would exert negative selection. However, the important  
362 point for the present study is that the PFAM database contains few false positives, i.e. regions  
363 incorrectly annotated as functional domains. This is supported by the observation that dN/dS  
364 inside domains is remarkably constant across the length of proteins (Fig. 2D), suggesting that  
365 domain annotation by PFAM is highly specific.

366 We may ask why, if protein termini are under low evolutionary pressure, are they not cropped  
367 by micro-deletions in the course of evolution? Requirements for reduced RNA secondary  
368 structures at the beginning of the CDS (Gu et al. 2010), presence of signal sequence with  
369 limited folding constraints but with key roles in the processing and traffic of the protein, such  
370 as signal peptides (von Heijne 1985), but also the necessary display of amino acids carrying  
371 specific post-translational modifications such as epigenetic marks in histones (e.g. methyl or  
372 acetyl groups) (Ghoneim et al. 2021), illustrate how protein extremities can fulfil specific  
373 important roles linked to their intrinsic evolutive and structural flexibility.

374 Methods designed to identify positive selection are sensitive to false positives, potentially  
375 caused by factors such as variable effective population size (Rouselle et al. 2018), biased  
376 gene conversion (Ratnakumar et al. 2010), multi-nucleotide mutations (Venkat et al. 2018)  
377 and punctual relaxation of selective pressure in a lineage (Zhang et al. 2005; Hughes 2007).  
378 Here we show that sites inferred as having experienced a period of positive selection are  
379 conspicuously enriched in regions with high dN caused by low selective pressure, suggesting

380 that they may contain a high proportion of false positives. This is consistent with the  
381 observation that experimentally tested positively selected sites are, on the contrary, depleted  
382 at sequence extremities. Of note, we observed that the synonymous rate  $dS$  is constant  
383 along protein length, thus providing little leverage for background model adjustments to  
384 counteract this effect in statistical tests of positive selection. Considering the excess of  
385 positive selection inferences at protein extremities as false positives would also be consistent  
386 with expectations that selection for advantageous traits would operate predominantly where  
387 functional domains and structural constraints are most frequent, i.e. far away from the  
388 extremities (Slodkowitz and Goldman 2020). Altogether, we propose that weaker structural  
389 constraints and weaker functional constraints lead to lower selective pressure at protein  
390 termini. Accounting for this bias in models of molecular evolution should improve their  
391 handling of site heterogeneity and accuracy of adaptive evolution inference.

392

## 393 **Material and Methods**

394 **Coding sequences, protein sequences and alignments.** The sequence data comprises 4  
395 sets covering different taxonomic groups: primates, rodents, plants, and Human-Mouse  
396 orthologs. Primate and rodent sequences were downloaded from the Ensembl database  
397 (Cunningham et al. 2021) as follows:

398 *Primates.* The genomes are from the following species: *Otolemur garnettii*, *Microcebus*  
399 *murinus*, *Propithecus coquereli*, *Prolemur simus*, *Saimiri boliviensis boliviensis*, *Cebus*  
400 *capucinus*, *Aotus nancymaae*, *Cercocebus atys*, *Mandrillus leucophaeus*, *Papio anubis*,  
401 *Theropithecus gelada*, *Macaca mulatta*, *Macaca fascicularis*, *Macaca nemestrina*,  
402 *Chlorocebus sabaeus*, *Colobus angolensis palliatus*, *Ptilocolobus tephrosceles*,  
403 *Rhinopithecus roxellana*, *Rhinopithecus bieti*, *Pongo abelii*, *Gorilla gorilla*, *Pan troglodytes*,  
404 *Pan paniscus*, *Homo sapiens*, *Nomascus leucogenys*, *Carlito syrichta*, *Mus musculus*  
405 (outgroup).

406 *Rodents.* The genomes are from the following species: *Tupaia belangeri*, , *Dipodomys ordii*,  
407 *Jaculus jaculus*, *Rattus norvegicus*, *Mus musculus*, *Mus spicilegus*, *Microtus ochrogaster*,  
408 *Cricetulus griseus*, *Mesocricetus auratus*, *Peromyscus maniculatus bairdii*, *Nannospalax*  
409 *galili*, *Octodon degus*, *Cavia porcellus*, *Chinchilla lanigera*, *Sciurus vulgaris*, *Marmota*  
410 *marmota marmota*, *Urocitellus parryii*, *Ictidomys tridecemlineatus*, *Ochotona princeps*  
411 (outgroup), *Oryctolagus cuniculus* (outgroup).

412 Phylogenetic gene trees and CDS sequences restricted to either primate or rodent genomes  
413 were downloaded from Ensembl Multi compara v101 (primates) and v104 (rodents) via the  
414 Perl API. To retain only strict 1:1 orthologs, for each tree the largest sub-tree that does not  
415 contain duplications was extracted, with a random choice in case of a tie. To avoid  
416 alignments with too few sequences, only the trees of size greater than 4 were selected. A

417 multiple sequence alignment (MSA) on the amino acid sequences was then performed with  
418 MAFFT (Kato and Standley 2013) (--maxiterate 1000 --localpair). Sequences were finally  
419 back-translated (treebest backtrans -t 0.9) with the corresponding CDS to obtain the final  
420 aligned codons in nucleotides.

421 Plants. The genomes were from the following *Fabids* species: *Cucumis sativus*, *Medicago*  
422 *truncatula*, *Lotus japonicus*, *Glycine max*, *Phaseolus vulgaris*, *Phaseolus angularis*, *Lupinus*  
423 *angustifolius*, *Manihot esculenta*, *Populus trichocarpa*, *Prunus persica*.

424 Coding sequences from all 10 genomes were downloaded from the April 2021 release of the  
425 OMA database (Altenhoff et al. 2021) based on Ensembl Plants. A MSA on the amino acid  
426 sequences of each OMA family was then computed with FSA (Bradley et al. 2009) with  
427 default parameters. Sequences were finally back-translated (treebest backtrans -t 0.9) with  
428 the corresponding CDS to obtain the final aligned codons in nucleotides.

429 Human-Mouse orthologs. Gene CDS were directly extracted from mRNA sequences  
430 (transcripts) downloaded from the October 2020 release of AniProtDB (Barreira et al. 2021).  
431 A reciprocal blastp was performed (e-value  $1.10^{-3}$ ) to select orthologs. Pairs of matching  
432 sequences were then aligned using the Needleman-Wunsh algorithm (needle, with -gapopen  
433 10.0 -gapextend 0.5) from the EMBOSS package.

434 (See Table S1 for a summary of data set size (number of sequences) related to figures.)

435

436 **Gap-less alignments.** Algorithms introduce gaps in MSA to accommodate insertion or  
437 deletions of amino acids. They are more frequent at the edges than in the middle of CDS,  
438 and they may bias the counting of substitutions in these regions: for every gap, there is one  
439 less site available to count substitutions in a column of the alignment, thus decreasing  
440 statistical power and increasing noise. To quantify substitutions in an unbiased way, for some  
441 results we removed, in each alignment, sequences that introduce gaps. Therefore, remaining  
442 sequences are aligned with no gaps, only substitutions. Since it was not always possible to  
443 find at least 2 remaining sequences in a given alignment, this procedure resulted in a smaller  
444 set of alignments (see table S1).

445

446 **Position-specific codon alignments and evolutionary rate computation.** To compute  
447 molecular evolutionary rates presented in 3 panels of 50 codons each, the following  
448 procedure was followed. Two random CDS sequences beginning with a start codon ATG  
449 were chosen in each MSA, and the aligned codons at the same position in each pair were  
450 concatenated into a new, position-specific alignment. The procedure was applied for the first,  
451 middle and last 50 codons of each alignment. On each concatenated position-specific  
452 alignment, the dN and dS were computed using the YN00 model in Codeml from the PAML4  
453 package (Yang 2007) using the Bio.Phylo.PAML.codeml library. See Figure S1 for a

454 schematic representation of the procedure. Several codon models were tested (YN00,  
455 LWL85 and NG86), with similar results. In order to limit situations where two codons would  
456 be counted in adjacent 50-codons panels because of short CDS, we restricted these  
457 computations to sequences at least 134 codons (>400 nucleotides) long (Figure S2).

458

459 **Metagenes.** Representations scaled from 0% to 100% of the length of proteins (metagenes)  
460 were computed as follows. For each MSA (e.g. for evolutionary rates, Figure 1A) or each  
461 protein sequence (e.g. RSA value density, Figure 3A), the length of the MSA (respectively  
462 sequences) was divided in 100 intervals. For this reason, only CDS/proteins of at least 100  
463 codons (or aa) are used. For each interval, the variable of interest (e.g. the number of aa  
464 substitutions and indels in Figure 1A) was divided by the number of sites in the interval  
465 (height of the columns in the MSA times the width of the interval). This ratio, now normalized  
466 to account for the varying size of each sample, was averaged on all samples for each  
467 position.

468

469 **Synthetic sequences.** To control for the potential role of errors during MSA constructions,  
470 which may have caused the observed excess of substitutions in CDS edges, we built a set of  
471 14,470 synthetic multiple alignments using INDELible (Fletcher and Yang 2009), which  
472 simulates aligned sequences containing insertions/deletions. Parameters were LAV 2 300  
473 model for gaps, submodel 4 0.175, insertrate 0.010500, deleterate 0.021000, and primate  
474 values for codon stationary frequencies. The template tree was taken from the real dataset,  
475 with one replicate for each gene tree. This resulted in 14,470 simulated alignments which  
476 were considered to be the “truth”. All gaps were then removed to reconstitute the original  
477 CDS, which were translated before being aligned with mafft (--maxiterate 1000 --localpair) in  
478 amino acids and back-translated into codon alignments. There were thus two sets of 14,470  
479 sequences: a “truth” set generated by INDELible and a “realigned” set. It was then possible  
480 to compute the dN/dS ratio for both datasets and compare the values at each position to  
481 measure the impact of the multiple alignment method. The same experiments were  
482 performed by realigning with FSA with and without HMM-Cleaner (Di Franco et al. 2019),  
483 with similar results. Other experiments were conducted by artificially lengthening (by a factor  
484 of up to 50) the length of the branches to obtain more distant sequences, but this did not  
485 change the conclusions.

486

487 **Flanking regions.** To control for a potential bias due to annotation errors, particularly at the  
488 extremities of the CDS which are notoriously difficult to identify precisely, we computed  
489 dN/dS values on position-specific codon alignments that included Untranslated (UTR)  
490 sequences before the start codon and after the stop codon. For this, all *Nomascus*

491 *leucogenys* (Nleu 3.0) and *Pan troglodytes* (Pan tro 3.0) CDS sequences from Ensembl  
492 version 104 were downloaded with upstream and downstream flanking 30 nucleotides  
493 collected via BioMart. Amino acid translation, double blastp (-evalue 1e-3) and reciprocal  
494 best hits selection followed by alignment (mafft --maxiterate 1000 --localpair) and finally  
495 backtranslation were performed to obtain aligned nucleotide data. Only the 13,079 pairs of  
496 orthologs aligned with no gaps were retained in order to compute dN/dS in these CDS as  
497 well as in the 10 flanking pseudo-codon overlapping UTR sequences.

498

499 **Functional Domain distribution.** Functional protein domains were downloaded via the  
500 Ensembl v101 API for 12,067 human protein sequences selected from the primate MSA for  
501 the computation of molecular rates, for four different databases: Pfam (Mistry et al. 2021),  
502 Smart (Letunic et al. 2021), SuperFamily (Gough et al. 2001) and Prosite (Sigrist et al. 2013)  
503 patterns. To compute random re-distribution of domains, the sizes of the inter-domain spaces  
504 for each protein were summed, then redivided in the same number of intervals but using  
505 randomly sampled boundaries. These new inter-domain intervals were then re-inserted  
506 between the original domains without changing their order or content.

507

508 **Solvent accessibility.** PDB-formatted 3D structures of the human and mouse proteomes  
509 predicted by AlphaFold (Jumper et al. 2021) were downloaded from the AlphaFold Protein  
510 Structure Database (Varadi et al. 2022) resulting in 23,391 human and 21,615 mouse  
511 structures. To obtain relative solvent accessibilities (RSA), we used pCASA (Wei et al. 2017)  
512 to compute the total accessible surface (m) and the average accessible surface (n) for each  
513 residue, and we took the ratio between the former and the latter:  $RSA = m / n$ .

514

515 **Signal Peptides.** Signal peptides were identified in the primate sequences involved in gap-  
516 less multiple sequence alignments (Table S1) using SignalP (Almagro Armenteros et al.  
517 2019) resulting in 15.82% proteins with a signal peptide.

518

519 **Pearson correlations in individual sequences.** We started from the same 13,668 rodent  
520 phylogenetic trees and sequence families as described above from Ensembl version 104  
521 (Table S1), aligned with MAFFT as described. We restricted the alignments to those of at  
522 least 400 bp, containing sequences from at least three different species and with no gap. The  
523 few MSA where all dN/dS values are null and no correlation can be computed were also  
524 excluded, resulting in 7,812 MSA for the 5' end and 7,790 MSA for the 3' end, respectively.  
525 The dN and dS were computed using the YN00 model in Codeml from the PAML4 package  
526 (Yang 2007) using the Bio.Phylo.PAML.codeml library. A Pearson coefficient from a linear

527 correlation was computed for each alignment between the dN/dS values and their respective  
528 positions.

529

530 **Positive selection.** To estimate sites under positive selection, we used 13,555 rodent  
531 phylogenetic trees and sequence families as described above from Ensembl version 104  
532 (Table S1), aligned with MAFFT as described, where the MSA was at least 300 bp (13,301  
533 MSA, Fig. 3B) or at least 400 bp (13,020 MSA, Fig. 3C). To compute positive selection in  
534 high and low RSA amino acids, we had to restrict the analysis to alignments for which we  
535 could unambiguously find a correspondence between the AlphaFold dataset and the  
536 Ensembl Dataset, thus ending with 8,371 MSA. We used Hyphy MEME (Mixed Effects Model  
537 of Evolution) (Murrell et al. 2012; Kosakovsky Pond et al. 2020) to detect sites evolving under  
538 positive selection under a proportion of branches. In all cases, only positions with a positive  
539 selection p-value lower than 0.05 were retained.

540 We also collected data from the literature (Figure S9) as follows:

- 541 • Table S5 in supplementary data in van der Lee et al. (2017) (van der Lee et al. 2017),  
542 corresponding to 934 Positively Selected Residues (PSR) in 331 primate protein  
543 coding genes out of 11,096 initial gene families tested. Positively selected genes  
544 were identified by Codeml from the PAML package (Yang 2007) using a gene-level  
545 dN/dS test and positively selected residues by selecting sites with a significant  
546 Bayesian posterior probability.
  - 547 • Tables S4-S5-S7-S15-S16-S18 in Murrell et al. (2012) (Murrell et al. 2012): We  
548 retained only 13 positively selected sites identified jointly using HyPhy-MEME and  
549 FEL (M+F+) in 7 animal genes.
  - 550 • Table 1 in Rodrigue et al., (2021) (Rodrigue et al. 2021), corresponding to 51  
551 positively selected sites identified in 6 metazoan genes with the MutSel-M3 model at  
552 a threshold  $p > 0.95$ .
  - 553 • Table 1 from Yokoyama, S. (2008) (Yokoyama 2008) Corresponding to 51  
554 experimentally validated sites in 6 visual pigment proteins (e.g. Rhodopsins).
- 555



556 **References and Notes**

557

558 Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von  
559 Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using  
560 deep neural networks. *Nat Biotechnol* 37:420–423.

561 Altenhoff AM, Train C-M, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, Nevers Y,  
562 Radoykova H-S, Rossier V, Warwick Vesztrocy A, et al. 2021. OMA orthology in  
563 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic  
564 Acids Res* 49:D373–D379.

565 Bæk KT, Kepp KP. 2022. Assessment of AlphaFold2 for Human Proteins via Residue  
566 Solvent Exposure. *J. Chem. Inf. Model.* 62:3391–3400.

567 Baele G, Gill MS, Bastide P, Lemey P, Suchard MA. 2021. Markov-Modulated Continuous-  
568 Time Markov Chains to Identify Site- and Branch-Specific Evolutionary Variation in  
569 BEAST. *Syst Biol* 70:181–189.

570 Barreira SN, Nguyen A-D, Fredriksen MT, Wolfsberg TG, Moreland RT, Baxevanis AD. 2021.  
571 AniProtDB: A Collection of Consistently Generated Metazoan Proteomes for  
572 Comparative Genomics Studies. *Mol Biol Evol* 38:4628–4633.

573 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne  
574 PE. 2000. The Protein Data Bank. *Nucleic Acids Res* 28:235–242.

575 Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009.  
576 Fast Statistical Alignment. *PLOS Computational Biology* 5:e1000392.

577 Carugo O. 2011. Participation of protein sequence termini in crystal contacts. *Protein Sci*  
578 20:2121–2124.

579 Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-  
580 Orimoloye O, Azov AG, Barnes I, Bennett R, et al. 2021. Ensembl 2022. *Nucleic  
581 Acids Res*:gkab1049.

582 Davydov II, Salamin N, Robinson-Rechavi M. 2019. Large-Scale Comparative Analysis of  
583 Codon Models Accounting for Protein and Nucleotide Selection. *Mol Biol Evol*  
584 36:1316–1332.

585 Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment  
586 filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol  
587 Biol* 19:21.

588 Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among  
589 protein sites. *Nat Rev Genet* 17:109–121.

590 Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution.  
591 *Mol Biol Evol* 26:1879–1888.

592 Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive  
593 at the residue level. *Mol Biol Evol* 26:2387–2395.

594 Ghoneim M, Fuchs HA, Musselman CA. 2021. Histone Tail Conformations: A Fuzzy Affair  
595 with DNA. *Trends in Biochemical Sciences* 46:564–578.

- 596 Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome  
597 sequences using a library of hidden Markov models that represent all proteins of  
598 known structure. *J Mol Biol* 313:903–919.
- 599 Gu W, Zhou T, Wilke CO. 2010. A Universal Trend of Reduced mRNA Stability near the  
600 Translation-Initiation Site in Prokaryotes and Eukaryotes. *PLOS Computational*  
601 *Biology* 6:e1000664.
- 602 Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling  
603 site- specific residue frequencies. *Molecular Biology and Evolution* 15:910–917.
- 604 von Heijne G. 1985. Signal sequences. The limits of variation. *J Mol Biol* 184:99–105.
- 605 Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for  
606 positive selection at the nucleotide sequence level. *Heredity* 99:364–373.
- 607 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K,  
608 Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure  
609 prediction with AlphaFold. *Nature* 596:583–589.
- 610 Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7:  
611 Improvements in Performance and Usability. *Molecular Biology and Evolution*  
612 30:772–780.
- 613 Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, Shore VC.  
614 1960. Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å.  
615 resolution. *Nature* 185:422–427.
- 616 Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: a comparison of methods  
617 for detecting amino acid sites under selection. *Mol Biol Evol* 22:1208–1222.
- 618 Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD,  
619 Magalis BR, Bouvier D, Nekrutenko A, et al. 2020. HyPhy 2.5-A Customizable  
620 Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol*  
621 37:295–299.
- 622 van der Lee R, Wiel L, van Dam TJP, Huynen MA. 2017. Genome-scale detection of positive  
623 selection in nine primates predicts human-virus evolutionary conflicts. *Nucleic Acids*  
624 *Res* 45:10634–10648.
- 625 Letunic I, Khedkar S, Bork P. 2021. SMART: recent updates, new developments and status  
626 in 2020. *Nucleic Acids Res* 49:D458–D460.
- 627 Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto  
628 SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families  
629 database in 2021. *Nucleic Acids Res* 49:D412–D419.
- 630 Moutinho AF, Trancoso FF, Dutheil JY. 2019. The Impact of Protein Architecture on Adaptive  
631 Evolution. *Mol Biol Evol* 36:2013–2028.
- 632 Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. 2012. Detecting  
633 Individual Sites Subject to Episodic Diversifying Selection. *PLOS Genetics*  
634 8:e1002764.
- 635 Nei M, Kumar S. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press

- 636 Perutz MF, Kendrew JC, Watson HC. 1965. Structure and function of haemoglobin: II. Some  
637 relations between polypeptide chain configuration and amino acid sequence. *Journal*  
638 *of Molecular Biology* 13:669–678.
- 639 Prosdocimi F, Linard B, Pontarotti P, Poch O, Thompson JD. 2012. Controversies in modern  
640 evolutionary biology: the imperative for error detection and quality control. *BMC*  
641 *Genomics* 13:5.
- 642 Ramsey DC, Scherrer MP, Zhou T, Wilke CO. 2011. The relationship between relative  
643 solvent accessibility and evolutionary rate in protein evolution. *Genetics* 188:479–  
644 488.
- 645 Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010.  
646 Detecting positive selection within genomes: the problem of biased gene conversion.  
647 *Philos Trans R Soc Lond B Biol Sci* 365:2571–2580.
- 648 Rodrigue N, Latrille T, Lartillot N. 2021. A Bayesian Mutation-Selection Framework for  
649 Detecting Site-Specific Adaptive Evolution in Protein-Coding Genes. *Mol Biol Evol*  
650 38:1199–1208.
- 651 Rousselle M, Mollion M, Nabholz B, Bataillon T, Galtier N. 2018. Overestimation of the  
652 adaptive substitution rate in fluctuating populations. *Biol Lett* 14:20180055.
- 653 Rubinstein ND, Pupko T. 2012. Detection and analysis of conservation at synonymous sites.  
654 In: Codon evolution: mechanisms and models. New York: Oxford University Press Inc.  
655 p. 218–228.
- 656 Ruff KM, Pappu RV. 2021. AlphaFold and Implications for Intrinsically Disordered Proteins.  
657 *Journal of Molecular Biology* 433:167208.
- 658 Sammut SJ, Finn RD, Bateman A. 2008. Pfam 10 years on: 10,000 families and still growing.  
659 *Brief Bioinform* 9:210–219.
- 660 Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of  
661 positive Darwinian selection are inflated by errors in sequencing, annotation, and  
662 alignment. *Genome Biol Evol* 1:114–118.
- 663 Sigrist CJA, de Castro E, Cerutti L, Cuče BA, Hulo N, Bridge A, Bougueleret L, Xenarios I.  
664 2013. New and continuing developments at PROSITE. *Nucleic Acids Res* 41:D344–  
665 347.
- 666 Slodkowitz G, Goldman N. 2020. Integrated structural and evolutionary analysis reveals  
667 common mechanisms underlying adaptive evolution in mammals. *PNAS* 117:5977–  
668 5986.
- 669 Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O,  
670 Wood G, Laydon A, et al. 2022. AlphaFold Protein Structure Database: massively  
671 expanding the structural coverage of protein-sequence space with high-accuracy  
672 models. *Nucleic Acids Res* 50:D439–D444.
- 673 Venkat A, Hahn MW, Thornton JW. 2018. Multinucleotide mutations cause false inferences  
674 of lineage-specific positive selection. *Nat Ecol Evol* 2:1280–1288.
- 675 Wang Y, Zhang H, Zhong H, Xue Z. 2021. Protein domain identification methods and online  
676 resources. *Computational and Structural Biotechnology Journal* 19:1145–1153.

- 677 Wei S, Brooks CL, Frank AT. 2017. A rapid solvent accessible surface area estimator for  
678 coarse grained molecular simulations. *J Comput Chem* 38:1270–1274.
- 679 Wilson CJ, Choy W-Y, Karttunen M. 2022. AlphaFold2: A Role for Disordered Protein/Region  
680 Prediction? *International Journal of Molecular Sciences* 23:4591.
- 681 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*  
682 24:1586–1591.
- 683 Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for  
684 heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.
- 685 Yokoyama S. 2008. Evolution of dim-light and color vision pigments. *Annu Rev Genomics*  
686 *Hum Genet* 9:259–282.
- 687 Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method  
688 for detecting positive selection at the molecular level. *Mol Biol Evol* 22:2472–2479.

689

690

### 691 **Acknowledgements**

692 We thank P. Vincens and the informatics service at IBENS for support, and  
693 Alexandra Louis, Guillaume Louvel, François Giudicelli and Nicolas Lartillot for  
694 helpful discussions.

695

### 696 **Funding**

697 This work has received support under the program « Investissements d’Avenir »  
698 launched by the French Government and implemented by ANR with the references  
699 ANR–10–LABX–54 MEMOLIFE and ANR–10–IDEX–0001–02 PSL\* Université Paris.  
700 R.B. received funding from the French Ministry for Education, Research and  
701 Innovation.

702

### 703 **Author contributions**

704 R.B., A.G. and H.R.C. conceived the study and analysed results. D. W. and D.S.  
705 analysed results. All authors contributed to the writing of the manuscript.

706

### 707 **Competing interest**

708 The authors declare no competing interests.

709

### 710 **Data and material availability**

711 The scripts and data necessary to generate all results and figures, including those in  
712 supplementary Information, have been deposited in Zenodo

713 (<https://zenodo.org/record/7650471>). Scripts are provided within a Conda  
714 environment.

