



**HAL**  
open science

## Evolution is not uniform along protein 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 protein sequences. 2022. hal-03863169

**HAL Id: hal-03863169**

**<https://hal.science/hal-03863169v1>**

Preprint submitted on 21 Nov 2022

**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 protein sequences

Raphaël Bricout<sup>1</sup>, Dominique Weil<sup>2</sup>, David Stroebe<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. Institut de Biologie Paris Seine (IBPS)

\* 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. However, the position of an amino-acid within the sequence is not known to influence this evolutionary speed. Here we discovered that amino-acid evolve almost twice faster at protein termini than in their centre, hinting at a strong topological bias along the sequence length. We further show that the distribution of functional domains and of solvent-accessible residues in proteins readily explain how functional constraints are weaker at their termini, leading to the observed excess of amino-acid substitutions. Finally, we show that methods inferring sites under positive selection are strongly biased towards protein termini, suggesting that they may confound positive selection with weak negative selection. These results suggest that accounting for positional information should improve evolutionary models.

## Main text

Rates of evolution vary greatly between protein-coding gene families, for example in correlation with their expression levels (1, 2), their function (3) or with translational selection (4) effects. Within a given gene family, molecular rates of evolution can also vary within lineages (5) and between lineages (6). Within proteins themselves, rate heterogeneity among amino-acid sites is influenced by their implication in functional domains and by structural constraints in the folded protein (7). Accounting for such heterogeneity in evolutionary models is critical to accurately infer phylogenies and estimate cases of positive selection, and elaborate models have been developed to achieve this (8), generally by estimating site-specific rates in a maximum likelihood framework employing Markov models of sequence evolution (9–11).

38 Surprisingly, the impact of the position of a given amino-acid in the sequence relative to the  
39 protein start and end on the rate of molecular evolution has not yet been investigated. To  
40 address this question, we measured the rates of amino-acid changing substitutions (non-  
41 synonymous, dN) and silent substitutions (synonymous, dS) at individual codon positions  
42 and average them over thousands of CDS sequences. Specifically, we computed multiple  
43 sequence alignments (MSA) of CDS from 16,810 primate gene families to identify fixed  
44 mutations (substitutions, insertions and deletions) that took place in these sequences during  
45 the evolution of 26 primate species. The results show a strong excess of such changes  
46 towards the sequence extremities (Fig. 1A), leading to a distinctive U-shaped pattern.  
47 Looking further into substitutions at each individual codon position, we computed position-  
48 specific codon average evolutionary rates (Fig. S1) to examine separately the dN and the dS.  
49 While the dS remains remarkably constant along the CDS length (average dS=0.052), the dN  
50 increases significantly in the region spanning the first and last 50 codons (Fig. 1B). We  
51 observe a similar bias when computing dN and dS along the CDS of 7,513 plant (Fabids)  
52 gene families, which were subjected to an approximately 8-fold higher divergence rate than  
53 primates (Fig. 1C). In summary, the dN appears to be driving the distinctive U-shaped  
54 pattern of total substitutions and dN/dS in gene coding sequences (Fig. 1D).

55

56

57

58

59

60

61

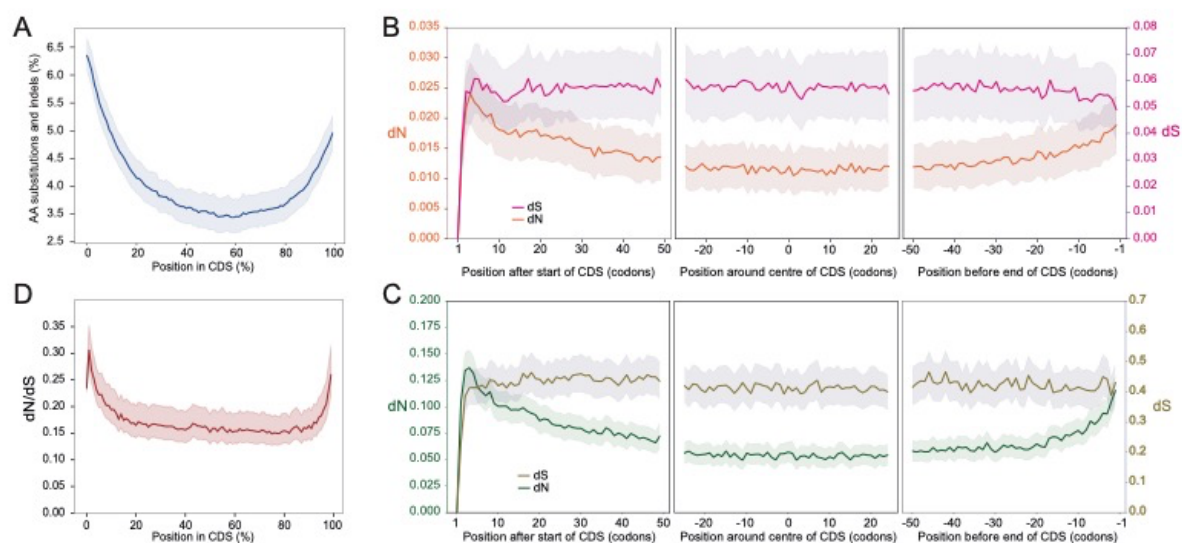
62

63

64

65

66



67 **Fig. 1. (A).** Frequency of amino acid substitutions, insertion and deletions computed in 16,810  
68 primate multiple sequence (CDS) alignments (MSA), rescaled from 0 to 100% of the coding  
69 sequence length. **(B).** Distribution of silent (dS) and non-synonymous (dN) substitution rates  
70 computed at each codon position from random pairs of sequences sampled from 16,248  
71 primate MSA without alignment gaps and shown here for the first 50 codons (left panel), the  
72 central 50 codons (middle panel) and the last 50 codons (right panel). **(C).** Same as in B but  
73 for pairs of CDS sampled from 7,513 plant MSA. **(D).** The distribution of dN/dS ratio for dN and  
74 dS values shown in B but across the entire CDS length rescaled from 0 to 100%. In all panels  
75 the shaded the area represents the 95% confidence interval.

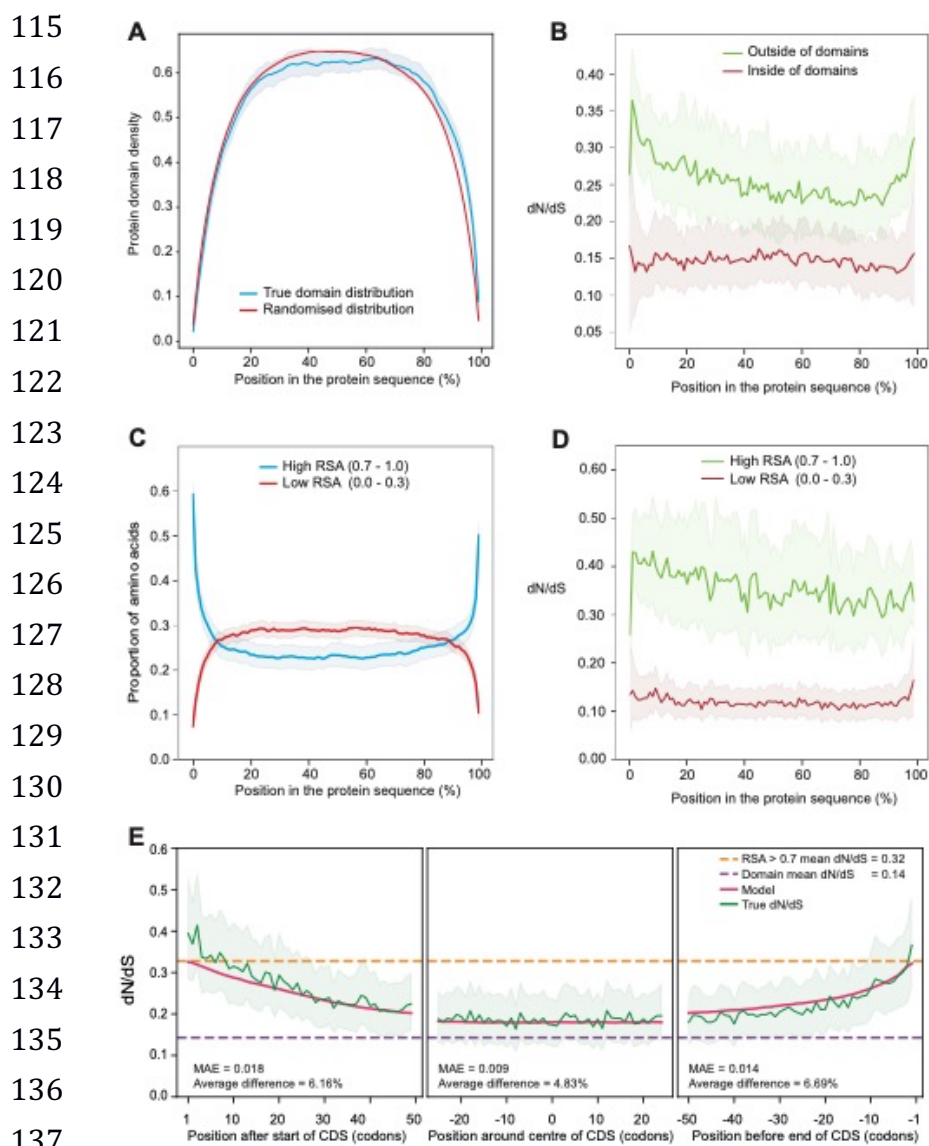
76

77

78 Computing substitutions in a multiple alignment of coding sequences is a multi-step process,  
79 with many potential sources of technical biases which could potentially explain this pattern  
80 (12, 13). We conducted a series of experiments to exclude annotation errors, multiple  
81 alignment artefacts and compositional biases (Supplementary material, Fig. S2), showing  
82 that our observations are robust to controls designed to address possible technical artefacts  
83 in the process from CDS annotation to substitution calculations.

84 We next examined biological or evolutionary explanations. We eliminated the possibility that  
85 a stronger mutation rate at CDS extremities would fuel the increased dN because the dS,  
86 which would be much more sensitive to the mutation rate, is essentially constant along CDS  
87 length (Fig. 1B,C). We next reasoned that a weaker negative selection at protein termini  
88 might be caused by weaker functional constraints. Predicted protein domains capture a large  
89 fraction of amino acids involved in structural and functional roles in protein sequences, and  
90 their prediction relies on sequence similarity and structural information (14). Both of these  
91 features make them good proxies for sites under evolutionary constraints. We computed the  
92 distribution of protein domains predicted by different methods along the 12,067 human  
93 protein sequences involved in our set of primate gene families, and we show that domains  
94 are strongly depleted at protein edges (Fig. 2A and fig. S3A-C). This dome-like shape is  
95 caused by the lower probability of domains overlapping amino acids immediately adjacent to  
96 these termini, since they cannot physically overlap the termini themselves. The distribution of  
97 domains decreases sharply towards protein edges regardless of the length of the protein  
98 sequences (Fig. S3D-E), supporting a scenario where all proteins are similarly affected by a  
99 deficit of domain-induced negative constraints at their edges. The dome-shaped distribution  
100 of domains mirrors the distinctive U-shaped distribution of the dN/dS ratio (Fig. 1A),  
101 consistent with our initial hypothesis that a depletion of domains at the edges of proteins  
102 would make them more permissive to non-synonymous changes and indels because of  
103 weaker selective constraints. To test this more directly, we distinguished codons that code  
104 for amino acids involved in a domain from those that do not, and computed the dN/dS for  
105 each category separately (Fig. 2B). In line with the above expectation, the dN/dS bias  
106 disappears when computed exclusively inside domains. Again, the difference in dN/dS  
107 behaviour is largely caused by the dN, since the dS remains constant both inside and outside  
108 of domains and is almost identical in both categories throughout the protein lengths (Fig. S4).  
109 These results strongly support a model in which selective constraints are significantly weaker  
110 at protein edges.

111 We next investigated how structural constrains, or lack thereof, may also influence  
112 evolutionary rates at protein termini. A protein sequence is folded in space through both local  
113 and distant amino-acid interactions. Amino acids which are free of those interactions are  
114 conversely accessible to solvents, and typically found on the surface of the folded protein.



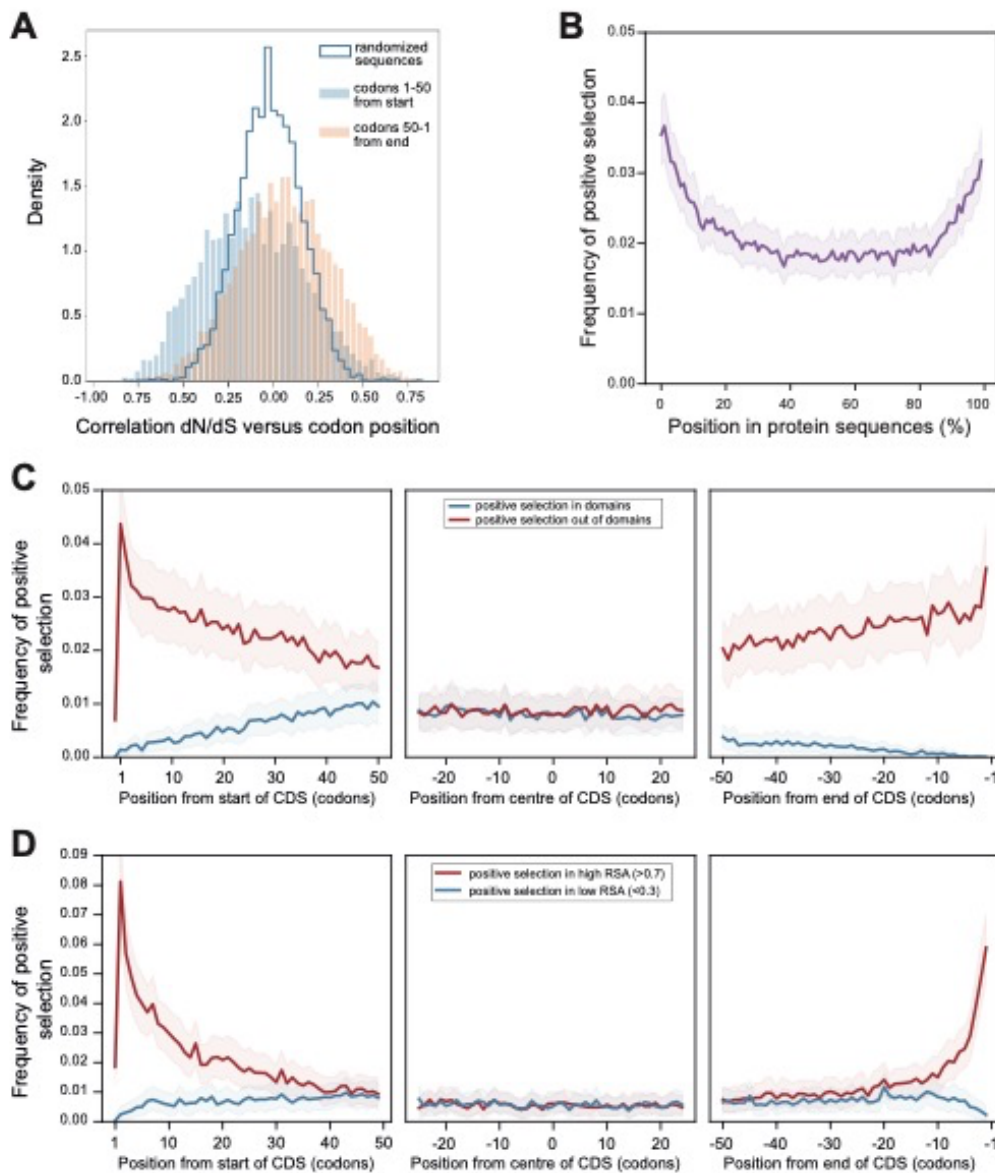
**Fig. 2.** (A) The distribution of protein domains from the PFAM database in 12,067 human proteins rescaled from 0-100% of their length (blue line). The red line shows the distribution of the same domains in random non-overlapping positions in the same sequences. (B) dN/dS computed in 16,248 alignments of at least 2 CDS sequences from 26 primate genomes, where sites inside (dark red line) and outside (green line) PFAM domains are distinguished. (C) The distribution of frequency of amino acids with high (blue line) and low (red line) Relative Solvent Accessibility (RSA) computed by pCASA on 3D structure predicted by AlphaFold on 23,391 human protein sequences, rescaled to 0-100% of the length. (D) dN/dS computed on 7,614 sequences common to the AlphaFold and Ensembl primate CDS datasets, where sites with high (green line) and low (dark red line) RSA are distinguished. (E) Proposed model where the mean dN/dS in domains and in highly accessible regions (RSA > 0.7) are weighted according respectively to the percent of codons overlapping domains and residues with RSA > 0.7. The Mean Absolute Error (MAE) and percent average error are indicated for each panel. In all panels the shaded area represents the 95% confidence interval.

156 Because such structural interactions are linked to the function of a protein, it is well-  
157 established that evolutionary rates differ between residues depending on their solvent  
158 accessibility(15–17). The precise relationship between solvent accessibility, evolutionary  
159 rates and amino-acid position along the sequence has however never been ascertained,  
160 mostly because N- and C-terminal regions of proteins are often removed to facilitate  
161 crystallisation prior to structure solving, or are generally poorly resolved in electron density  
162 maps. To circumvent this issue, we analysed 23,391 structures predicted by AlphaFold (18)  
163 on complete human protein sequence to compute the relative solvent accessibility (RSA) of  
164 each residue (Fig. S5A). We note that RSA values follow a bimodal distribution with a  
165 distinctive peak above 0.7 depleted in residues included in protein domains. Conversely,  
166 residues with RSA below 0.3 are enriched in functional domains. We took residues from  
167 these two extremes categories to compute their distribution along protein sequences, and we  
168 find that solvent accessibility increases sharply at protein termini, consistent with weak  
169 structural constraints in these regions (Fig. 2C). Critically, the dN/dS rate is low and constant  
170 along protein length in sites with low accessibility, while it is high in highly accessible regions  
171 (Fig. 2D). In both categories, the marked increase in dN/dS at sequence extremities shown in  
172 Figure 1 is absent, indicating that solvent accessibility is likely a strong marker of the  
173 decrease in selective pressure observed in the N- and C-terminal region of proteins.  
174 Because both functional domains and RSA seem to drive dN/dS variation at protein termini,  
175 we applied a model where the average dN/dS at a given residue is the sum of the average  
176 dN/dS in domains and in high RSA (RSA > 0.7), each weighted by the proportion of residues  
177 in their category (Supplementary Material). The model reproduces the observed dN/dS with  
178 remarkable accuracy in human proteins (mean difference = 5.9%, Fig. 2E), suggesting that  
179 functional domains and RSA are two variable that are sufficient to explain the bias in average  
180 dN/dS along proteins sequences. The slight asymmetry in dN/dS bias between N- and C-  
181 termini observed in the model, reminiscent of the asymmetry observed in Figure 1B, largely  
182 disappears when we remove the 15.8% of proteins labelled with a signal peptide at the N-  
183 terminus (Fig. S5B). Signal peptide, which are known to be highly variable in sequence (19)  
184 are therefore likely to provide additional relief from the evolutionary pressure measured in  
185 this region.

186 The dN/dS bias at protein ends is measurable by averaging thousands of sites at any given  
187 position. Is it also significant at the level of individual sequence alignments? This is important  
188 if evolutionary models applied to single gene families are likely to be affected. To address  
189 this, we computed a correlation between codon position and dN/dS for 12,322 multiple  
190 sequence alignments, separately for the 50 codons at beginning and at the end of coding  
191 sequences (Fig. 3A).

192

193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216



**Figure 3.** (A) Distribution of Pearson correlation coefficients between dN/dS values and their position in mouse CDS from 12,322 alignments of 5 to 20 rodent sequences. Positions are either the first 50 (filled blue bars) or the last 50 (filled pink bars) codons. The blue line shows the distribution for the first and last 50 codons in the same mouse sequences but where their positions were randomized. (B) Distribution of the frequency of sites under positive selection in 12,170 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 12,170 rodent CDS, distinguishing sites inside (blue line) and outside (red line) PFAM domains. (D) Distribution of the frequency of sites under positive selection in the first, middle and last 50 codons of 5,893 rodent CDS, distinguishing sites with high (red line) and low (blue line) RSA.

227  
228  
229  
230  
231  
232

Compared to a control where amino-acid positions are randomised, the distribution of correlations are significantly shifted towards negative ( $p\text{-value}=8.10^{-80}$ ; t-test) and positive ( $p\text{-value} = 6.10^{-46}$ ; t-test) values for the start and end regions of coding sequences, respectively. This reflects the existence of a measurable increase in dN/dS towards CDS edges, even in

233 individual sequences. We propose that this pattern is caused by the same factors as for the  
234 average sequences analysed previously (Fig. 1B), respectively the biased domain  
235 distribution and solvent accessibility.

236 These results immediately raise questions for the identification of positive selection in proteins  
237 sequences, because a significantly elevated dN compared to some background rate is  
238 generally taken as evidence of adaptive changes (20). The U-shaped bias in dN observed in  
239 our study suggests that relaxation of constraints at protein edges might confound tests of  
240 positive selection. To investigate this, we estimated sites under positive selection in a set of  
241 12,170 Rodent gene trees using a site model (methods). We found that sites estimated under  
242 positive selection are strikingly enriched at protein edges (Fig. 3B), and that this enrichment is  
243 specifically attributed to sites located outside of functional domains (Fig. 3C) and to residues  
244 with high solvent accessibility (Figure 3D). Notably, the same bias towards coding sequence  
245 extremities can be observed in several recent published scans for positive selection (Figure  
246 S6). Interestingly, while this bias is conspicuous for sites estimated to have been subject to  
247 positive selection using bioinformatic methods, it is not the case for experimentally verified  
248 sites, although our compilation of cases for this category is too small to draw general  
249 conclusions.

250 We reveal a pattern of evolutionary rate along coding sequences that has so far remained  
251 concealed: the average amino acid substitution rate (dN) increases towards the extremities of  
252 the sequence. We found this pattern by assembling observations that were sometimes known  
253 quantitatively or intuitively in the field but never connected with respect to codon or amino-acid  
254 positions in sequences. This patterns provide insights into the elusive mechanism driving  
255 evolutionary rate heterogeneities (7). First noted by Perutz and colleagues on haemoglobin  
256 (21) and confirmed by many studies since, protein surfaces evolve faster than their interior,  
257 where structural constraints, residue interactions and functional sites are most enriched and  
258 solvent accessibility is very low (15, 16). Attempts at explaining evolutionary rate heterogeneity  
259 have thus mainly focused on this paradigm, that structural constraints governed by complex  
260 spatial interactions create a range of selective pressures on amino acids, but these are still  
261 hard to predict from the sequence itself.

262 We note that in the present study molecular rates are not dependent on a substitution model,  
263 as they do not rely on ancestral state inferences in coding sequences, and molecular rates are  
264 computed on MSA without gaps, which are known to introduce biases. Also, our finding that  
265 the average dS is constant along CDS length (Figs. 1B, 1C) should not be interpreted as  
266 meaning that dS does not vary or is not subject to site heterogeneity in individual genes, as  
267 this has been shown in many studies (22).

268 If protein termini are under lower evolutionary pressure, why are they not cropped by micro-  
269 deletions in the course of evolution? The example of signal peptides, but also of amino-acids



270 carrying specific epigenetic marks (e.g. methyl or acetyl groups) in histones, illustrate why  
271 maintaining structurally flexible regions may be functionally important.  
272 Methods designed to identify positive selection are sensitive to false positives, potentially  
273 caused by factors such as variable effective population size (23), biased gene-conversion  
274 (24), multi-nucleotide mutations (25) and punctual relaxation of selective pressure in a  
275 lineage (5, 26). Here we show that sites inferred as having experienced a period of positive  
276 selection are conspicuously enriched in regions with high dN caused by low selective  
277 pressure, suggesting that they may contain a high proportion of false positives. This is  
278 consistent with the observation that experimentally tested positively selected sites are, on the  
279 contrary, depleted at sequence extremities. Of note, we observed that the synonymous rate  
280 dS is constant along protein length, thus providing little leverage for background model  
281 adjustments to counteract this effect in statistical tests of positive selection. Considering the  
282 excess of positive selection inferences at protein extremities as false positives would also be  
283 consistent with expectations that selection for advantageous traits would operate  
284 predominantly where functional domains and structural constraints are most frequent, i.e.  
285 away from the extremities (27). It has also been previously shown that non-adaptive changes  
286 as well as positively selected sites are significantly enriched on the surface of proteins where  
287 solvent accessibility is high, emphasizing the difficulty in distinguishing them in these regions  
288 (17). Altogether, we propose that intervals between functional domains display a neutrally  
289 evolving size and weaker structural constraints, largely causing lower selective pressure at  
290 protein termini. Accounting for this bias in models of molecular evolution should improve their  
291 handling of site heterogeneity and accuracy of adaptive evolution inference.

292

293

## 294 **References and Notes**

295

- 296 1. C. Pál, B. Papp, L. D. Hurst, *Genetics*. **158**, 927–931 (2001).
- 297 2. L. Duret, D. Mouchiroud, *Molecular Biology and Evolution*. **17**, 68–74 (2000).
- 298 3. L. Zhang, W.-H. Li, *Mol Biol Evol*. **21**, 236–239 (2004).
- 299 4. D. A. Drummond, A. Raval, C. O. Wilke, *Mol Biol Evol*. **23**, 327–337 (2006).
- 300 5. J. Zhang, R. Nielsen, Z. Yang, *Mol Biol Evol*. **22**, 2472–9 (2005).
- 301 6. H. Ellegren, *Mol Ecol*. **17**, 4586–4596 (2008).
- 302 7. J. Echave, S. J. Spielman, C. O. Wilke, *Nat Rev Genet*. **17**, 109–121 (2016).
- 303 8. A. L. Halpern, W. J. Bruno, *Molecular Biology and Evolution*. **15**, 910–917 (1998).

- 304 9. Z. Yang, R. Nielsen, N. Goldman, A. M. Pedersen, *Genetics*. **155**, 431–449 (2000).
- 305 10. S. L. Kosakovsky Pond, S. D. W. Frost, *Mol Biol Evol*. **22**, 1208–1222 (2005).
- 306 11. G. Baele, M. S. Gill, P. Bastide, P. Lemey, M. A. Suchard, *Syst Biol*. **70**, 181–189  
307 (2021).
- 308 12. A. Schneider *et al.*, *Genome Biol Evol*. **1**, 114–118 (2009).
- 309 13. F. Prosdocimi, B. Linard, P. Pontarotti, O. Poch, J. D. Thompson, *BMC Genomics*. **13**, 5  
310 (2012).
- 311 14. Y. Wang, H. Zhang, H. Zhong, Z. Xue, *Computational and Structural Biotechnology  
312 Journal*. **19**, 1145–1153 (2021).
- 313 15. D. C. Ramsey, M. P. Scherrer, T. Zhou, C. O. Wilke, *Genetics*. **188**, 479–488 (2011).
- 314 16. E. A. Franzosa, Y. Xia, *Mol Biol Evol*. **26**, 2387–2395 (2009).
- 315 17. A. F. Moutinho, F. F. Trancoso, J. Y. Dutheil, *Mol Biol Evol*. **36**, 2013–2028 (2019).
- 316 18. J. Jumper *et al.*, *Nature*. **596**, 583–589 (2021).
- 317 19. G. von Heijne, *J Mol Biol*. **184**, 99–105 (1985).
- 318 20. M. Nei, S. Kumar, *Molecular Evolution and Phylogenetics* (Oxford University Press,  
319 2000).
- 320 21. M. F. Perutz, J. C. Kendrew, H. C. Watson, *Journal of Molecular Biology*. **13**, 669–678  
321 (1965).
- 322 22. N.D. Rubinstein, T. Pupko, in *Codon evolution: mechanisms and models* (Oxford  
323 University Press Inc., New York, 2012), pp. 218–228.
- 324 23. M. Rousselle, M. Mollion, B. Nabholz, T. Bataillon, N. Galtier, *Biol Lett*. **14**, 20180055  
325 (2018).
- 326 24. A. Ratnakumar *et al.*, *Philos Trans R Soc Lond B Biol Sci*. **365**, 2571–2580 (2010).
- 327 25. A. Venkat, M. W. Hahn, J. W. Thornton, *Nat Ecol Evol*. **2**, 1280–1288 (2018).
- 328 26. A. L. Hughes, *Heredity*. **99**, 364–73 (2007).
- 329 27. G. Slodkowicz, N. Goldman, *PNAS*. **117**, 5977–5986 (2020).

330

### 331 **Acknowledgements**

332 We thank P. Vincens and the informatics service at IBENS for support, and  
333 Alexandra Louis, Guillaume Louvel, François Giudiucelli and Nicolas Lartillot for  
334 helpful discussions.

335

### 336 **Funding**

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

342

### 343 **Author contributions**

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

346

### 347 **Competing interest**

348 The authors declare no competing interests.

349

### 350 **Data and material availability**

351 The scripts and data necessary to generate all figures, including in supplementary  
352 Information, have been deposited in Zenodo (doi:10.5281/zenodo.6472876).

353

### 354 **Supplementary Materials**

355 *Material and Methods*

356 *Fig S1 to S6*

357 *Table S1*

358 *References*

359