Evolution is not Uniform Along Coding Sequences

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

Evolution is not uniform along coding sequences

Raphaël Bricout[1], Dominique Weil[2], David Stroebel[1], Auguste Genovesio[1]*, Hugues Roest Crollius[1]*

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, 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

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

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

Results

Evolution is faster at protein termini

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

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

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

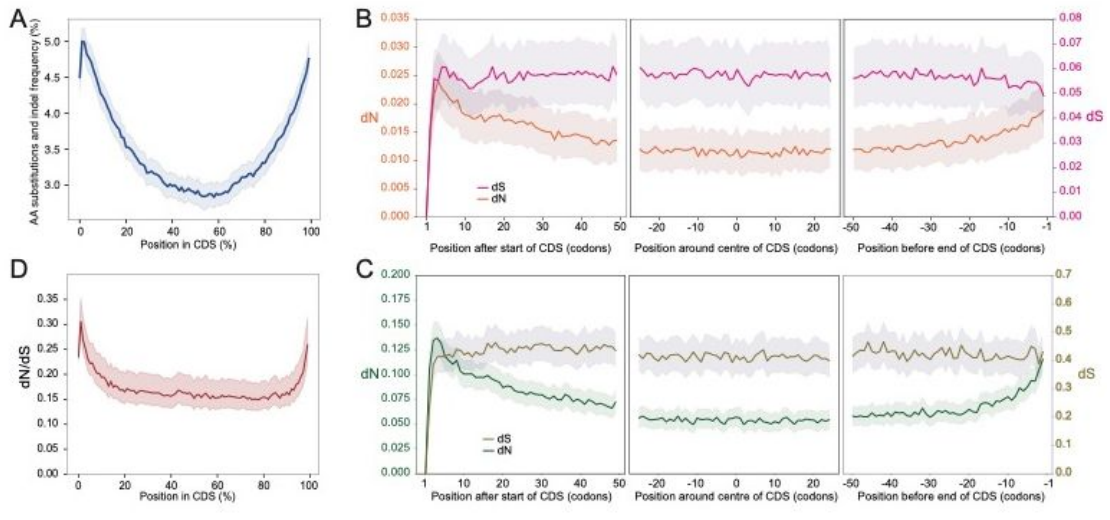


Fig. 1. (**A**). Frequency of amino acid substitutions, insertion and deletions computed in 15,828 primate multiple sequence (CDS) alignments (MSA), rescaled from 0 to 100% of the CDS length. (**B**). Distribution of silent (dS) and non-synonymous (dN) substitution rates computed at each codon position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

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

Evolution is faster for solvent-exposed amino acids

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

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

Evolutionary rates correlate with protein domain density

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

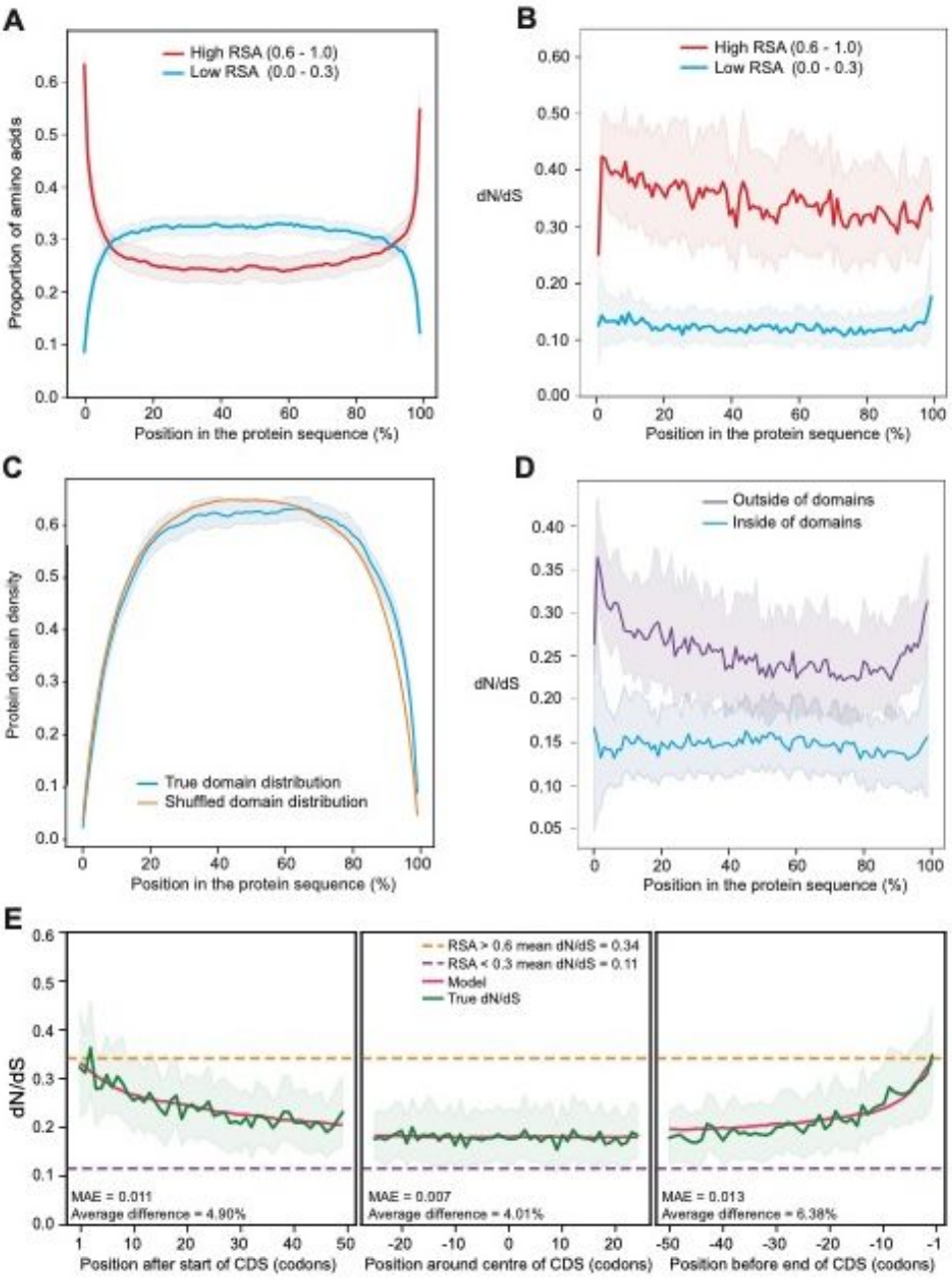


Fig. 2. (**A**) The distribution of frequency of amino acids with high (red line) and low (blue line) Relative Solvent Accessibility (RSA) computed by pCASA on 3D structure predicted by AlphaFold on 22,613 human protein sequences, rescaled to 0-100% of the length. (**B**) dN/dS computed on 7,144 sequences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

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

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

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

Faster evolution can be measured at the termini of individual proteins

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

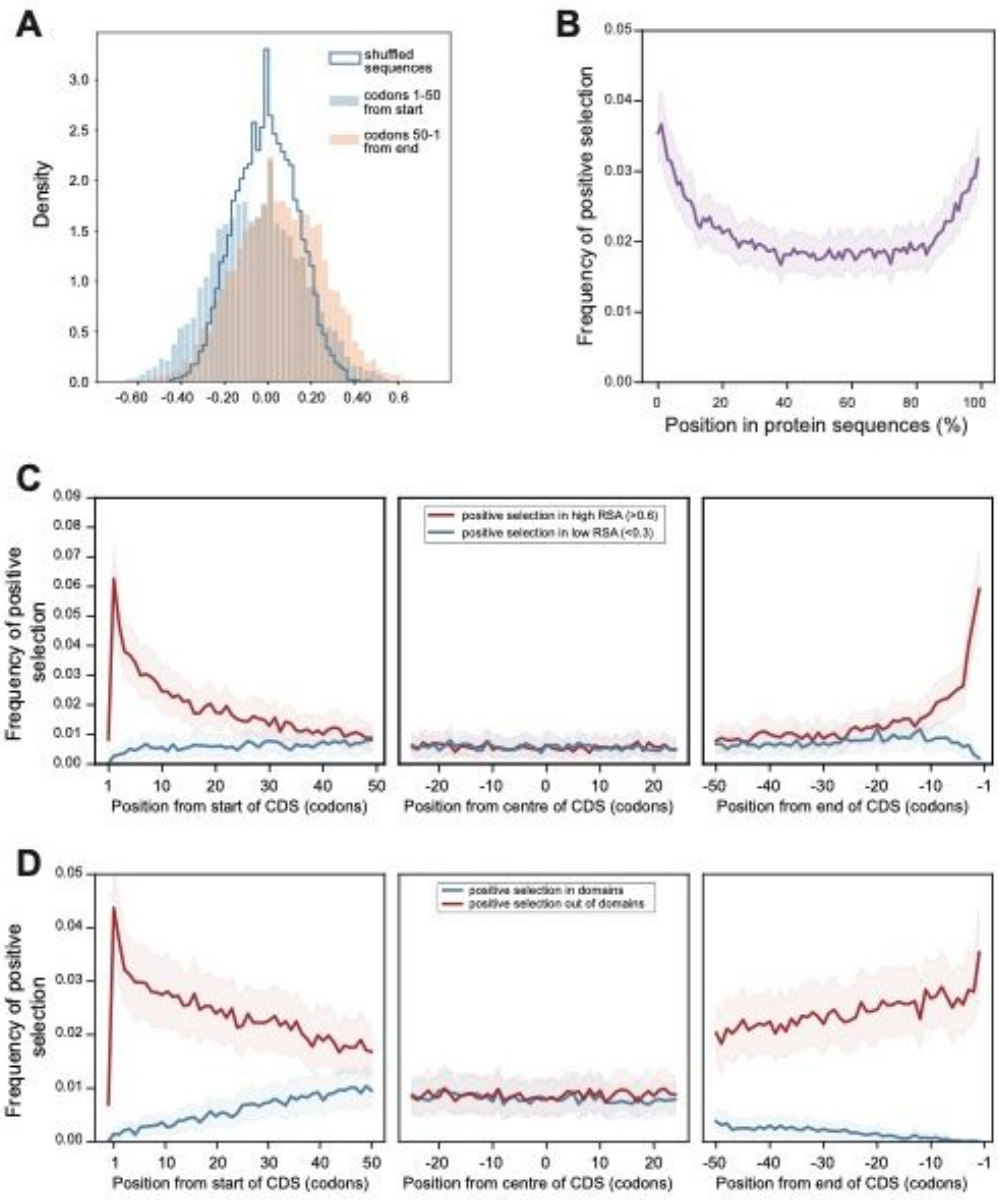


Figure 3. (**A**) Distribution of Pearson correlation coefficients between dN/dS values and their position in mouse CDS from 7,800 alignments of at least 3 rodent sequences. The distributions correspond either to the first (filled blue bars) or the last (filled orange bars) 50 codons. The blue line shows 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.

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

Faster evolutionary rates at termini confound tests of positive selection

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

Discussion

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

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

It is worth noting that in the present study, molecular rates are not dependent on a substitution model, as they do not rely on ancestral state inferences in CDS, and molecular rates are computed on MSA without gaps, which are known to introduce biases. This may explain that while we see the same increase in dN/dS at the 5' end of CDS as in a recent study on Drosophila genes (Davydov et al. 2019), we do not see the same decrease in substitution rates.

On a related matter, we wish to point out that our finding that the average dS is constant along CDS length (Figs. 1B, 1C) should not be interpreted as meaning that dS does not vary or is not subject to site heterogeneity in individual genes, as this has been shown in many previous studies (Rubinstein and Pupko 2012).

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

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

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

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

Material and Methods

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

References and Notes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Ghoneim M, Fuchs HA, Musselman CA. 2021. Histone Tail Conformations: A Fuzzy Affair 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 A.
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, Cuche 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 Slodkowicz 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.

Wei S, Brooks CL, Frank AT. 2017. A rapid solvent accessible surface area estimator for coarse grained molecular simulations. *J Comput Chem* 38:1270–1274.

Wilson CJ, Choy W-Y, Karttunen M. 2022. AlphaFold2: A Role for Disordered Protein/Region Prediction? *International Journal of Molecular Sciences* 23:4591.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

Yokoyama S. 2008. Evolution of dim-light and color vision pigments. *Annu Rev Genomics Hum Genet* 9:259–282.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22:2472–2479.

Acknowledgements

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

Funding

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

Author contributions

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

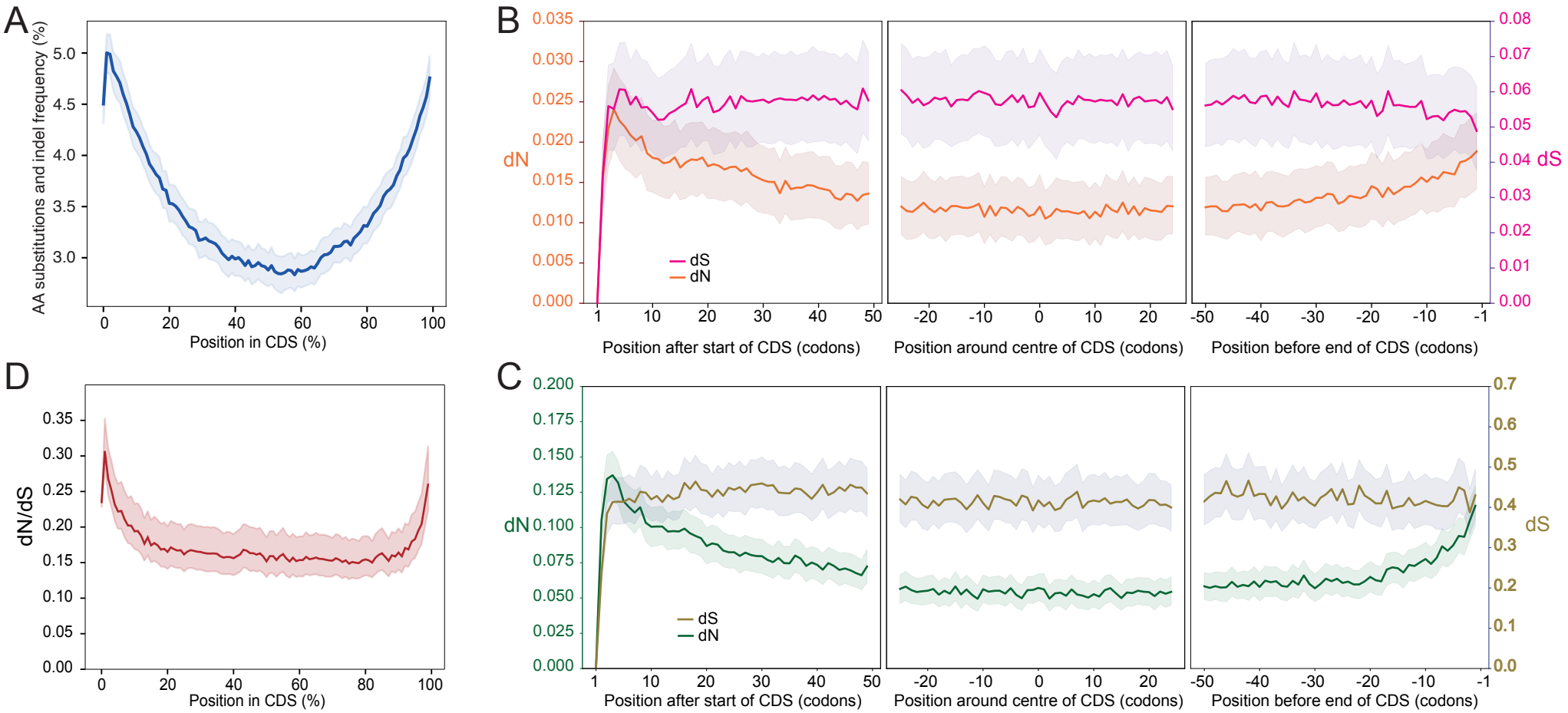
Competing interest

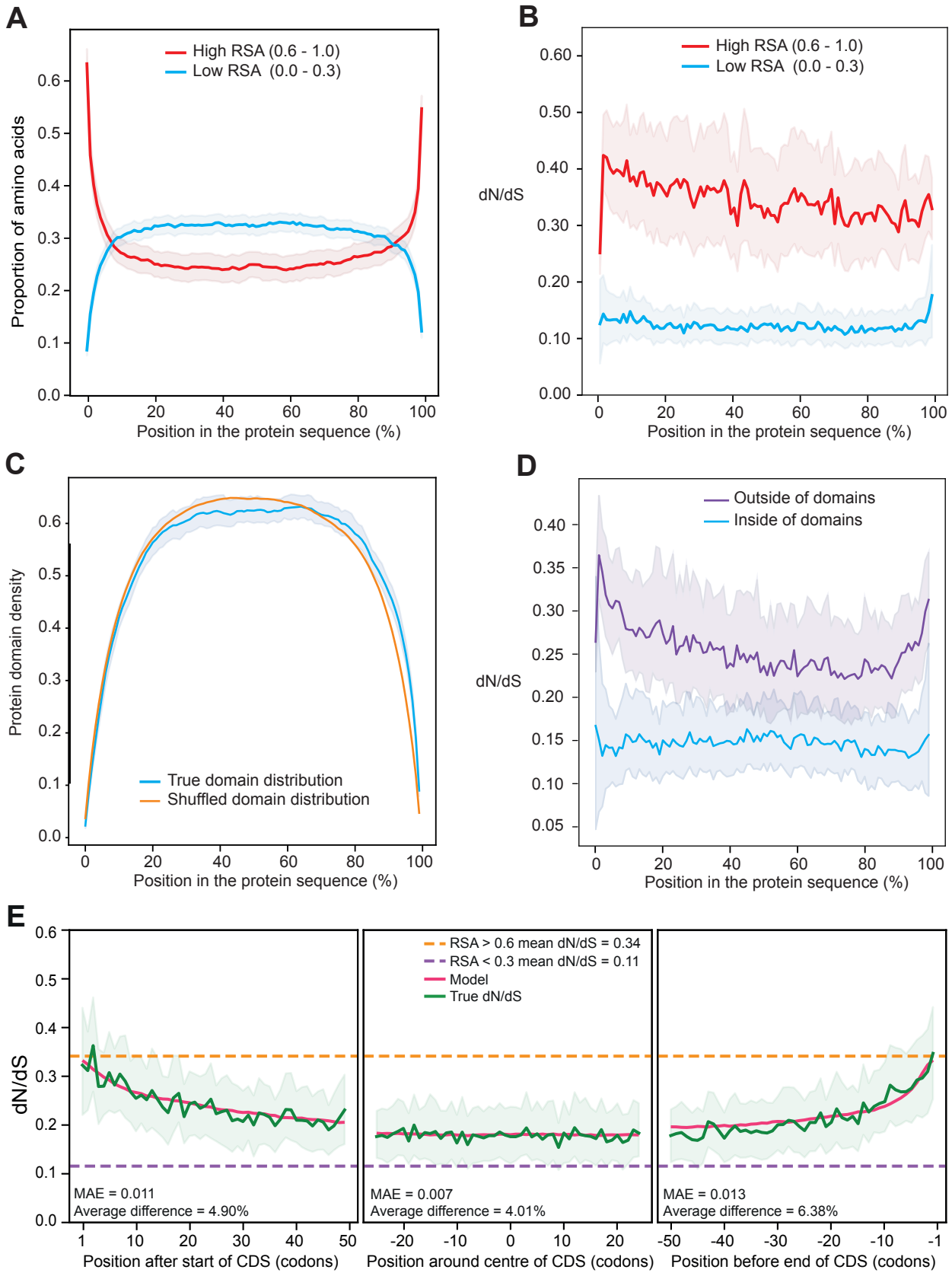
The authors declare no competing interests.

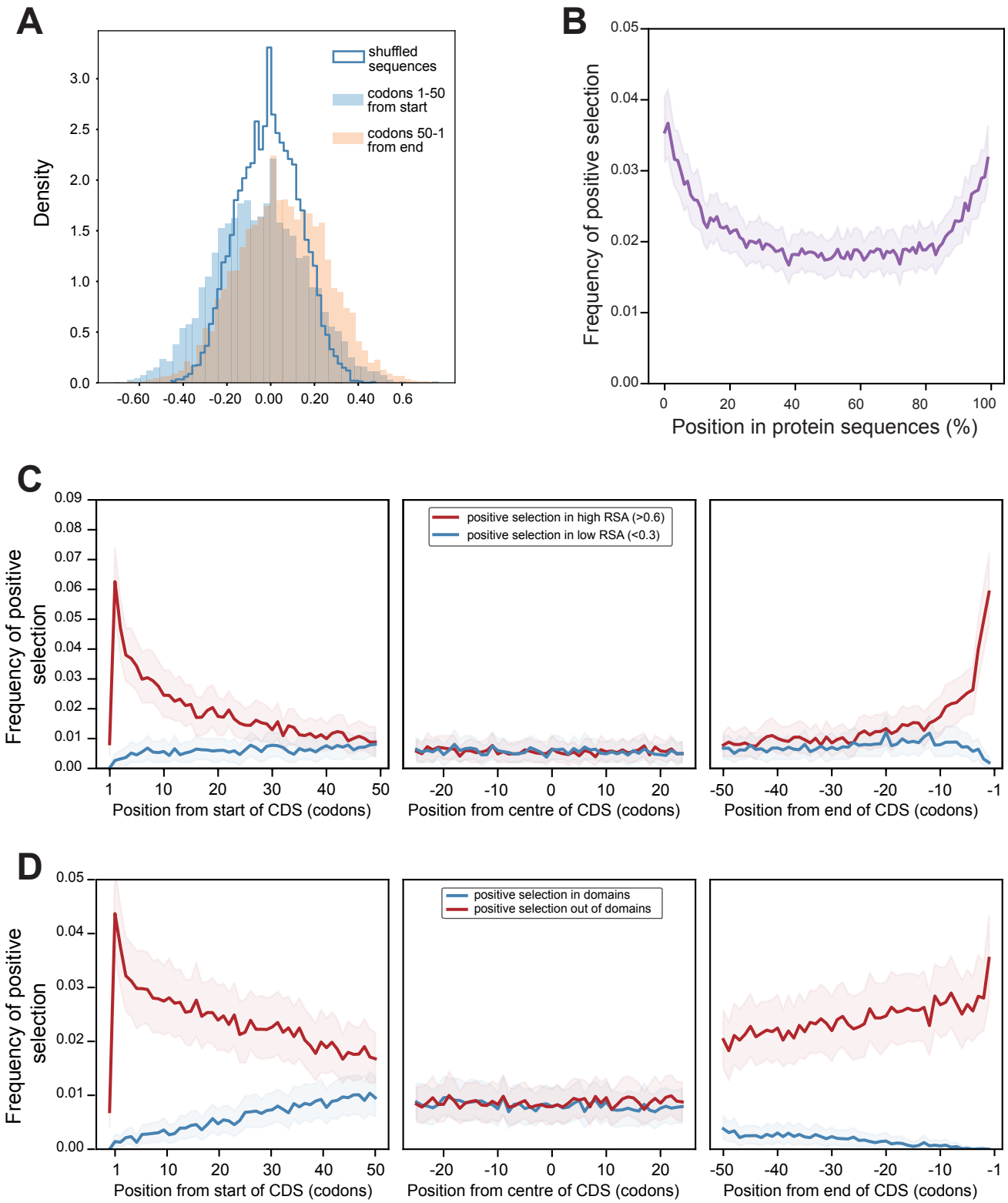
Data and material availability

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

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


A
AA substitutions and indel frequency (%)
Position in CDS (%)
B
dN
dS
Position after start of CDS (codons)
Position around centre of CDS (codons)
Position before end of CDS (codons)
C
dN
dS
Position after start of CDS (codons)
Position around centre of CDS (codons)
Position before end of CDS (codons)
D
dN/dS
Position in CDS (%)


A
High RSA (0.6 - 1.0)
Low RSA (0.0 - 0.3)
Proportion of amino acids
Position in the protein sequence (%)
B
High RSA (0.6 - 1.0)
Low RSA (0.0 - 0.3)
dN/dS
Position in the protein sequence (%)
C
Protein domain density
True domain distribution
Shuffled domain distribution
Position in the protein sequence (%)
D
Outside of domains
Inside of domains
dN/dS
Position in the protein sequence (%)
E
RSA > 0.6 mean dN/dS = 0.34
RSA < 0.3 mean dN/dS = 0.11
Model
True dN/dS
dN/dS
MAE = 0.011
Average difference = 4.90%
MAE = 0.007
Average difference = 4.01%
MAE = 0.013
Average difference = 6.38%
Position after start of CDS (codons)
Position around centre of CDS (codons)
Position before end of CDS (codons)


A
shuffled sequences
codons 1-50 from start
codons 50-1 from end
Density
B
Frequency of positive selection
Position in protein sequences (%)
C
positive selection in high RSA (>0.6)
positive selection in low RSA (<0.3)
Frequency of positive selection
Position from start of CDS (codons)
Position from centre of CDS (codons)
Position from end of CDS (codons)
D
positive selection in domains
positive selection out of domains
Frequency of positive selection
Position from start of CDS (codons)
Position from centre of CDS (codons)
Position from end of CDS (codons)