Online Resource 1

Journal of Molecular Evolution

A hydropathic asymmetry in the second codon position separates two otherwise matched transversions

Lucas Giovani Ribeiro, Marcos André Simonssini

Independent Researcher, Uruana de Minas, Minas Gerais, Brazil

Corresponding author: Lucas Giovani Ribeiro — coin26@mail.com

Contents

Supplementary Note S1 — Derivation of the axis, in order

The axis was found, not hypothesized. The order was:

1. R regressed on log κ and the CpG>T shift over the 4,972 species. 2. The residual correlated against each of the 96 canonical substitution categories, within Fungi and globally. 3. The twelve categories with the

strongest correlations were all C>G (positive, +0.71 to +0.76) or T>A (negative, −0.71 to −0.76), with no dependence on trinucleotide context. 4. Collapsing context, C>G correlated +0.79 with the fungal residual and T>A −0.78; the remaining four classes fell below 0.23 in absolute value. 5. A first candidate axis, `log[(C>G + T>A) / (C>A + T>G)]` — strength-preserving over strength-switching transversions — reduced residual variance by 6.5%. It fails because the two classes act in opposite directions and summing them cancels the signal. It is not used. 6. The ratio `log(C>G / T>A)` reduced residual variance by 63.5% and is the axis reported.

Steps 1–6 were carried out on the same 4,972 species on which the results are reported. The protocol recording this, frozen before the validation, is deposited as `protocol/protocol_v1.md`.

Table S1 — The six substitution classes across 4,972 species

Fraction of the normalized spectrum, context collapsed.

class	type	median	minimum	maximum
C>T	transition	0.4784	0.1207	0.8030
T>C	transition	0.2063	0.0657	0.6694
C>A	transversion	0.0871	0.0206	0.2468
C>G	transversion	0.0765	0.0088	0.3691
T>A	transversion	0.0573	0.0072	0.2433
T>G	transversion	0.0547	0.0090	0.2901

The axis `A_s = log(C>G / T>A)` ranges from −2.81 to +3.15: the underlying ratio runs from 0.06 to 23.3, a span of nearly 400-fold, with median +0.34.

Table S2 — Cost of every substitution class by codon position

Mean squared polar requirement over sense→sense edges, weighted by the reference exposure.

Classes collapsed by reverse complement, as supplied.

class	overall	position 1	position 2	position 3	synonymous	nonsense
C>T	2.872	—	—	—	54.5%	4.2%
C>G	3.185	5.069	4.149	0.213	17.4%	2.1%
C>A	5.823	—	—	—	21.8%	7.4%
T>G	5.823	—	—	—	21.8%	2.2%
T>A	10.669	5.810	25.646	0.205	22.0%	7.9%

C>A and T>G are mutually reverse under the collapse and therefore identical in cost; C>G and T>A are each self-reverse. Synonymous fractions are taken over sense→sense edges, nonsense fractions over all outgoing weight.

Table S3 — Within-clade fits

Coefficient on `A_s` per standard deviation, fitted within each clade, and the reduction in residual variance relative to the two-predictor model.

clade	n	coefficient	reduction
Plants	903	+0.01986	78.7%
Reptiles	134	+0.01307	75.8%
Insects	489	+0.01583	73.9%
Other	212	+0.01712	66.0%
Fungi	619	+0.00972	63.4%
Birds	897	+0.01306	53.5%
Fish	1,049	+0.01273	51.2%
Mammals	669	+0.01203	47.8%

Removing all clade-level variation from both response and predictors, the pooled reduction is 62.2% with a coefficient of +0.01168, against 63.5% and +0.01175 before removal.

Table S4 — Robustness across amino-acid metrics

metric	residual variance, base	with A_s	reduction
squared polar requirement	1.876×10^{-4}	6.857×10^{-5}	63.5%
squared Kyte–Doolittle	7.264×10^{-4}	2.860×10^{-4}	60.6%
Grantham	4.714×10^{-5}	3.431×10^{-5}	27.2%

Table S5 — Leave-one-clade-out validation

Out-of-fold RMSE with and without the axis, fitting on all clades but one.

held-out clade	n	base	with A_s	gain
Reptiles	134	0.00886	0.00371	58.1%
Fungi	619	0.02196	0.01153	47.5%
Plants	903	0.01439	0.00848	41.1%
Fish	1,049	0.01223	0.00732	40.2%
Birds	897	0.00929	0.00621	33.2%

held-out clade	n	base	with A_s	gain
Other	212	0.02685	0.01802	32.9%
Mammals	669	0.00579	0.00492	15.1%
Insects	489	0.01762	0.01606	8.9%

Size-weighted: 0.01458 to 0.00952, a gain of 34.7%. Improves in 8 of 8.

Table S6 — Nested and simple validation compared

Every partition of the eight clades into a held-out set of one, two or three clades and the remainder as training: 7 + 28 + 35 = 70 partitions. Held-out sets range from 212 to 2,849 species.

Nested — the contrast is reselected from the fifteen available using the training clades alone, then evaluated on the held-out ones:

quantity	value
partitions evaluated	70
partitions selecting C>G / T>A in training	70
partitions where held-out error falls	69
median gain	31.6%
gain range	−3.8% to 42.7%

The contrast is reselected, not assumed, and it is selected in every partition. Held-out error falls in 69 of the 70; the exception is a partition holding out a single small clade.

Table S7 — Specificity against the randomized ensemble

Each code selects its own best contrast from the fifteen pairwise ratios among the six substitution classes, over the full 10,000-code ensemble.

	value
standard code, best contrast	C>G / T>A
standard code, reduction	63.5%
ensemble median	16.9%
ensemble range	0.1% to 85.1%
codes matching or exceeding the standard code	131 of 10,000
percentile of the standard code	98.7

Under the looser control, with the axis held fixed for every code, the ensemble median is 8.9% and the percentile 99.0 over a 600-code sample. The stricter figure is reported in the manuscript. Per-code values are deposited as `data/fair_null_gains.npy` and `data/specificity_gains.npy`.

Table S8 — Phylogenetic generalized least squares

Fitted on the calibrated tree pruned to the 4,972 confirmatory species, with Pagel's λ estimated and the per-species measurement variance entering the terminal edges. Coefficients are per standard deviation of the predictor.

model	τ^2	λ	log L	ΔBIC
κ + CpG	1.840×10^{-3}	0.9926	17,594.1	—
κ + CpG + axis	6.832×10^{-4}	0.9882	19,462.4	−3,728

predictor	coefficient	95% CI
κ	+0.01531	+0.01498 to +0.01564
CpG shift	−0.00545	−0.00596 to −0.00493
axis	+0.01275	+0.01241 to +0.01308

The latent variance falls by 62.9%, against 63.5% under ordinary least squares and 62.2% after removing clade means. The estimate of λ near 0.99 in the base model indicates that `R` tracks the tree closely; the axis does not act by proxying that structure, since correcting for it leaves the reduction almost unchanged.

The tree is labelled with reduced binomials, which differ from the CORAL labels for 254 species; joining on the wrong column silently drops those rows. See `data/DATA_DICTIONARY.md`.

Table S9 — Pure exposure to each substitution class

The standard code evaluated against an exposure restricted to one class at a time, against the same 10,000-code ensemble. Costs are in squared polar requirement over sense-to-sense edges, weighted by the reference exposure.

class	standard code	ensemble mean	advantage	codes doing better
C>G	3.185	10.443	+69.5%	42
C>T	3.094	8.213	+62.3%	11
T>C	3.094	8.213	+62.3%	11
C>A	5.823	9.869	+41.0%	146
T>G	5.823	9.869	+41.0%	146
T>A	10.669	9.872	−8.1%	6,276

C>T and T>C are mutually reverse under the collapse, as are C>A and T>G; C>G and T>A are each self-reverse.

By codon position:

class	position	standard code	ensemble mean	advantage	codes better
C>G	1	5.069	12.616	+59.8%	13.5%
C>G	2	4.149	12.627	+67.1%	10.9%
C>G	3	0.213	5.942	+96.4%	0.0%
T>A	1	5.810	12.635	+54.0%	14.9%
T>A	2	25.646	12.650	−102.7%	98.1%
T>A	3	0.205	4.529	+95.5%	0.1%

Table S10 — Anisotropy and the fitted coefficient

A code's anisotropy is `G = log[cost(T>A) / cost(C>G)]` under pure exposure, computed from the code table alone. The coefficient is the fitted effect of the axis on `R` across the 4,972 species, with that code substituted for the standard one. Over 1,500 codes drawn from the ensemble.

quintile of G	mean G	mean coefficient
1	−0.800	−0.01406
2	−0.336	−0.00417
3	−0.029	+0.00022
4	+0.235	+0.00536
5	+0.642	+0.01113

Correlation between `G` and the coefficient: **+0.707**. Linear fit: coefficient = +0.0171 × `G` + 0.0007.

The standard code has `G` = +1.209, at the 99.53rd percentile of the ensemble. The fit predicts +0.0213 for a code in that position; the observed coefficient is +0.0118. The prediction is of the right sign and order, not of the exact value: the relationship is not perfectly linear, and the standard code lies in the sparse upper tail where the fit is least constrained.

Coefficients on absolute cost and on `R` are not related by a constant, because the ensemble mean in the denominator of `R` varies by species. All values here are on the scale of `R`. Per-code values are deposited as `data/anisotropy.npz`.

Table S11 — NCBI translation tables

Every table that reassigns codons relative to table 1 and whose assignment this model can represent. Costs in squared polar requirement over sense-to-sense edges, weighted by the reference exposure. "C>G p2" and "T>A p2" are costs at the second codon position; "ratio" is their quotient there; "overall" is the same quotient over all three positions.

table	name	C>G p2	T>A p2	ratio	overall
2	Vertebrate Mitochondrial	3.930	25.360	6.45	3.35
3	Yeast Mitochondrial	4.220	22.689	5.38	2.93
4	Mold, Protozoan and Coelenterate Mitochondrial	4.220	25.646	6.08	3.26
5	Invertebrate Mitochondrial	3.540	25.360	7.16	3.37
6	Ciliate, Dasycladacean and Hexamita Nuclear	4.149	24.151	5.82	3.13
9	Echinoderm and Flatworm Mitochondrial	3.540	25.572	7.22	3.39
10	Euplotid Nuclear	4.345	25.646	5.90	3.24
12	Alternative Yeast Nuclear	4.149	24.754	5.97	3.05
13	Ascidian Mitochondrial	3.650	25.360	6.95	3.36
14	Alternative Flatworm Mitochondrial	3.540	23.884	6.75	3.33
15	Blepharisma Nuclear	4.149	24.849	5.99	3.16
16	Chlorophycean Mitochondrial	4.149	23.936	5.77	3.46
21	Trematode Mitochondrial	3.540	25.292	7.14	3.36
22	Scenedesmus obliquus Mitochondrial	4.149	23.936	5.77	3.62
23	Thraustochytrium Mitochondrial	4.149	25.646	6.18	3.52
24	Rhabdopleuridae Mitochondrial	4.255	25.646	6.03	3.14

table	name	C>G p2	T>A p2	ratio	overall
25	Candidate Division SR1 and Gracilibacteria	3.899	25.646	6.58	3.34
26	Pachysolen tannophilus Nuclear	4.149	24.851	5.99	3.14
29	Mesodinium Nuclear	4.149	22.471	5.42	3.36
30	Peritrich Nuclear	4.149	29.660	7.15	3.19
32	Balanophoraceae Plastid	4.149	23.942	5.77	3.44
33	Cephalodiscidae Mitochondrial	4.255	23.953	5.63	3.08

The constraint on C>A and T>G holds in all 22, to floating-point error. The separation of C>G and T>A holds in all 22, in the same direction, with the ratio at the second position between 5.38 and 7.22 (median 6.01) against a randomized ensemble median of 1.07.

Regenerate with `python3 code/variants.py`.

Table S12 — Inclusion and exclusion of NCBI tables

Taken from the NCBI genetic codes page, last updated 23 September 2024.

table	status	reason
1 Standard	reference	reassigns nothing; the reference against which the others are defined
2, 3, 4, 5, 6, 9, 10, 12, 13, 14, 15, 16, 21, 22, 23, 24, 25, 26, 29, 30, 32, 33	included	assignment is a fixed map from all 64 codons to an amino acid or stop
11 Bacterial, Archaeal and Plant Plastid	excluded	identical to table 1 in codon assignment; differs only in initiation codons
27 Karyorelict Nuclear	excluded	UGA is documented as stop or Trp depending on context
28 Condylostoma Nuclear	excluded	UAA and UAG are Gln or stop, UGA is Trp or stop, depending on context
31 Blastocrithidia Nuclear	excluded	UAA and UAG are Glu or stop depending on context

Twenty-two tables therefore enter the analysis. The exclusions are structural rather than selective: each is excluded because the model used here — a fixed map from all 64 codons to an amino acid or stop — cannot represent it, or because it is not a reassignment at all.

Table S13 — The decomposition, term by term

Weighted with the edge weights of Φ, so the three terms sum to the observed contrast exactly. Regenerate with `python3 code/decomposition.py`.

Squared polar requirement

term	U ↔ A	C ↔ G	contrast
observed Φ	25.646	4.149	+21.497
centroid separation	20.898	0.717	+20.181
internal dispersion	5.861	2.459	+3.401
pairing	−1.113	+0.972	−2.086

Squared Kyte–Doolittle

term	U ↔ A	C ↔ G	contrast
observed Φ	50.540	6.913	+43.627
centroid separation	48.800	2.093	+46.707
internal dispersion	1.188	7.807	−6.619
pairing	+0.551	−2.987	+3.539

The identity holds to 7×10^{-15} under both metrics.

The centroid term accounts for 94% of the observed contrast under polar requirement and 107% under Kyte–Doolittle. The pairing term attenuates the contrast by 8.8% under the first and amplifies it by 8.8% under the second; it is a small correction whose direction is not consistent between metrics.

Prediction for the support-preserving null. That randomization permutes within each second-position column separately for the codons that contribute a sense-to-sense edge and for those whose partner is a stop, so the weighted centroids and dispersions are held exactly fixed and only the pairing term varies. Its median should equal the composition terms alone:

	value
centroids + dispersion, polar requirement	+23.583
median of 2,000 column-wide randomizations (pre-specified)	+23.909
difference	1.4%
median of 2,000 support-preserving randomizations	+23.562
difference	0.09%

Table S14 — Where the residual exposure falls

Among the 211 codes buffered in eleven or twelve of the twelve distinct cells, the cell carrying the largest standardized cost. The last column gives the same share over all 10,000 codes, for comparison. Regenerate with `python3 code/structural.py`.

cell	codes	share among the buffered	share over the ensemble
T>A p2	65	30.8%	12.9%
T>A p1	38	18.0%	10.0%
C>T p3	23	10.9%	11.8%
C>G p1	21	10.0%	9.8%
C>A p1	14	6.6%	7.6%
C>G p2	11	5.2%	8.8%
C>T p2	11	5.2%	6.3%
C>G p3	10	4.7%	5.2%
C>T p1	8	3.8%	7.5%
C>A p2	7	3.3%	7.7%
T>A p3	3	1.4%	10.2%

Eleven of the twelve cells occur. T>A at the second codon position is the most frequent, at 2.4 times its ensemble share (Fisher odds ratio 3.10, p = 5.7e-12), but it is not obligatory: two thirds of these codes carry their residual exposure elsewhere.

The enrichment rises with the number of buffered cells.

cells below the mean	codes	T>A p2 as the worst cell
6	1733	9.4%
7	1933	11.7%
8	1567	14.3%
9	973	20.0%
10	562	31.5%
11	178	33.7%
12	33	15.2%

The fall at twelve is expected rather than informative: in those codes every cell lies below the mean, so the largest standardized cost is not an exposure at all.

Table S15 — The exploratory functional gradient

Cost matrices were built from 209 ProteinGym deep mutational scanning assays. Within each assay, every position was scored by the mean standardized effect of mutations at that position, computed **excluding all mutations of the two focal classes**, so that the focal observations do not contribute to their own ranking. Positions were then split into quintiles and a separate matrix built for each, with each assay contributing equally and each amino-acid pair's mean shrunk toward zero in proportion to how little data supports it.

Each matrix was substituted for the theoretical metric in the second-position contrast, for the standard code and for each of the 10,000 randomized codes.

constraint quintile	T>A − C>G under the standard code
1, most constrained	+0.2488
2	+0.1170
3	+0.0517
4	−0.0334
5, most permissive	−0.0371

	value
slope across the five quintiles, standard code	-0.0722
ensemble median slope	+0.00220
percentile of the standard code	93.8
codes with a slope this negative	615 of 10,000
one-sided p	0.061

A tercile split rather than quintiles gives the same picture: the contrast is +0.207 at constrained positions and −0.046 at permissive ones, and the difference places the standard code at the 94.3rd percentile. Repeating the whole construction on three random halves of the assay collection gives percentiles of 88.9, 95.8 and 96.9.

Limitations. The constraint index is a proxy derived from the same assay scores that supply the cost matrices, not an independent structural measurement. The stratification was chosen after the pattern was seen. The monotonic ordering across five bins is a description of the gradient, not an independent test. The percentile corresponds to a one-sided p of about 0.06 and does not reach conventional significance. We report the analysis because it is the closest this work comes to a functional consequence and because it defines a test that structural annotation would settle.

Regenerate with `python3 code/pg_matrix.py`; the stratified variant is described in the protocol, §8n.

Table S16 Mean cost of the two free classes by codon position.

class	position 1	position 2	position 3
C>G	5.069	4.149	0.213
T>A	5.810	25.646	0.205

Table S17 Mean cost of every exchange available at the second codon position.

exchange	type	cost
A ↔ U (T>A)	transversion	25.646
A ↔ C	transversion	14.530
G ↔ U	transversion	9.166
A ↔ G	transition	7.931
C ↔ G (C>G)	transversion	4.149
C ↔ U	transition	3.486

Table S18 The two second-position columns in the standard code and in the 33 codes that fall below the ensemble mean in all twelve cells.

	T>A at position 2	C>G at position 2	sum over twelve cells
standard code	+2.33	−1.13	−13.75
the 33, mean	−0.73	−0.74	−9.48 (median)

Table S19 The second-position contrast T>A − C>G in each natural code, against its own block-preserving null. Percentiles under squared polar requirement; per- table values in Online Resource 1, Table S11.

	value
variant tables tested	22
randomizations per table	2,000
median percentile	98.5
tables above the 95th percentile	22 of 22
median of the code-specific nulls	within a few units of zero

Table S20 The same test under alternative amino-acid metrics.

metric	median percentile	tables above the 95th
squared polar requirement	98.4	22 of 22
squared Kyte–Doolittle	100.0	22 of 22
Grantham	71.9	0 of 22

Table S21 The second-position contrast of the standard code against four randomization schemes. The two within-column schemes differ in what they hold fixed; see Methods §9.

scheme	null median	percentile of the standard code
block-preserving, amino acids permuted	≈ 0	98.8
sense codons permuted freely	≈ 0	100.0
permuted within columns, column-wide (pre-specified)	+23.91	0.0
permuted within columns, edge support preserved	+23.56	0.0

Table S22 Residual standard deviation by fold, with and without the axis. Residuals are those of the pooled model; the reduction is in standard deviation, not variance, and is not the within-clade refit reported in §11.

clade	n	residual sd, base	with A_s	reduction
Fungi	619	0.02194	0.01057	51.8%
Insects	489	0.01388	0.00782	43.7%
Plants	903	0.01345	0.00782	41.8%
Reptiles	134	0.00630	0.00372	40.9%
Other	212	0.02500	0.01601	35.9%
Fish	1,049	0.00979	0.00677	30.9%
Birds	897	0.00834	0.00665	20.3%
Mammals	669	0.00568	0.00468	17.6%