Methods

Input format and preprocessing

We analyze a codon alignment in PHYLIP sequential format. The first line declares the number of taxa and sites; subsequent blocks consist of a taxon name line followed by wrapped sequence lines. Sequences are uppercased on read. Characters outside {A,C,G,T} (e.g., N, ?, -) are treated as missing and excluded from per-site calculations. The script verifies that all sequences have equal length and that the total length is divisible by three.

Per-site and pooled composition statistics

For each alignment column (nucleotide site), we count observed A/C/G/T across taxa after discarding missing characters. Sites with fewer than two observed bases are ignored for entropy computations. We also compute pooled base frequencies across the entire matrix (all taxa × sites), again ignoring missing values; these frequencies (p_A, p_C, p_G, p_T) are used to parameterize expectations under full substitution saturation (FSS).

Entropy and the Index of Substitution Saturation (Iss)

Per-site Shannon entropy is computed as:

$$H = -\sum_{b \in \{A,C,G,T\}} p_b \log p_b$$

where p_b is the observed frequency of base b at the site (natural log units). We average over usable sites to obtain H̄. Following Xia’s formulation, the script defines:

$$\mathrm{Iss} = \frac{\bar{H}}{H_{\mathrm{FSS}}}$$

where H_FSS is the expected per-site entropy under full substitution saturation given the number of taxa and the pooled base frequencies.

Exact expectation under full substitution saturation

To obtain H_FSS and its variance, the script evaluates the exact multinomial expectation over all possible site compositions for N taxa. Specifically, for all non-negative integer quadruples (n_A, n_C, n_G, n_T) with sum n_b = N, it weights the entropy of the composition:

$$-\sum_{b}\left(\frac{n_b}{N}\right)\log\left(\frac{n_b}{N}\right)$$

by the multinomial probability:

$$\frac{N!}{\prod_b n_b!}\prod_b p_b^{n_b}$$

It also accumulates E[H^2] to report Var(H_FSS)=E[H^2]-E[H]^2. This yields the theoretical mean and variance of per-site entropy under FSS without simulation. Computationally this is O(N^3) states; practical for typical N.

Position-specific Iss

To characterize heterogeneity across codon positions, sites are partitioned by position (1st, 2nd, 3rd) and their mean entropies are normalized by the same H_FSS to yield Iss_pos1/2/3.

Codon-level contributions

To localize saturation-like behavior, the script computes a codon contribution score for each codon k. It averages the entropies of its three nucleotide sites, then normalizes by H_FSS:

$$\mathrm{Iss\_contrib}_k = \frac{\frac{1}{3}\sum_{i\in\{3k-2,\,3k-1,\,3k\}} H_i}{H_{\mathrm{FSS}}}$$

Higher values indicate codons whose within-column variability is closer to the FSS expectation.

Species (taxon) contributions via leave-one-out

To identify taxa that inflate saturation, the script performs leave-one-taxon-out (LOTO) analyses.

For each taxon t, it recomputes Iss on the reduced alignment and reports:

$$\Delta\mathrm{Iss}_t = \mathrm{Iss}_{(-t)} - \mathrm{Iss}_{\mathrm{full}}$$

Large positive ΔIss indicates that removing t reduces saturation (i.e., t had been pushing Iss upward).

Monte Carlo FSS check (optional)

As an empirical complement to the exact expectation, the script can run a Monte Carlo test by simulating independent sites under FSS: for each site, it draws N nucleotides i.i.d. with the pooled base frequencies, computes the site entropy, and averages over the number of usable sites to obtain H̄_sim. Repetition yields an empirical distribution of H̄ under FSS; the script reports p_hi = P(H̄_sim ≥ H̄_obs) and p_lo = P(H̄_sim ≤ H̄_obs). Small p_lo indicates the observed mean entropy is well below FSS (i.e., unsaturated).

Thresholding and filtered alignment output

Two user-supplied thresholds enable targeted cleaning: Codon filter: drop codons with Iss_contrib ≥ τ_codon. Taxon filter: drop taxa with ΔIss ≥ τ_taxon. Filters can be applied separately or jointly. When filtering, entire codons (all three sites) are removed to preserve the reading frame. The script writes PHYLIP-sequential alignments for the codon-filtered, taxon-filtered, and combined-filtered datasets, using the same wrapping width as the input. Manifests listing removed codons and taxa are also written.

Outputs

A summary table reporting N, total and usable sites, H̄, H_FSS, Var(H_FSS), Iss, Iss_pos1/2/3, and pooled base frequencies. A codon-wise table of Iss_contrib (and flags, if thresholded). A taxon-wise table of ΔIss (and flags, if thresholded). Optional Monte Carlo FSS results (p_hi, p_lo). Optional filtered PHYLIP alignments ready for downstream inference.

How this differs from DAMBE—and why the results remain valid

DAMBE implements Xia’s Iss test together with critical thresholds Iss_c (for symmetric vs. asymmetric trees and varying numbers of taxa), derived from extensive simulations, and uses those Iss_c tables to make a formal decision (“significantly saturated or not”). This script (i) computes Iss with the same core definition H̄/H_FSS; (ii) obtains H_FSS and Var(H_FSS) by exact multinomial expectation given the observed base composition and taxon count; and (iii) provides an empirical FSS Monte Carlo check rather than consulting Iss_c tables tied to tree-shape assumptions. Thus, while it does not label results using DAMBE’s Iss_c decision framework, its Iss values are directly comparable and its “distance from FSS” assessments are grounded in the same entropy logic that underpins Iss. Moreover, the script adds diagnostically useful localization (codon contributions) and influence (leave-one-out taxa) analyses that aid targeted data curation prior to phylogenetic inference.