GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

MAX-PLANCK-INSTITUT FÜR MULTIDISZIPLINÄRE NATURWISSENSCHAFTEN

Evolutionary Genomics: Consequences of Biodiverse Reproductive Systems (EvoReSt)

EvoReSt

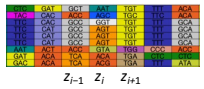
C2

Novel methods for analysis of selection using mutation-selection

Johannes Söding

State of the art: codon substitution models

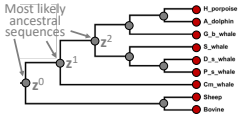
Codon-based MSA



z_{i-1} z_i z_{i+1}

External tool

Phylogenetic tree for protein family



Probability to go from codon *a* to b in time Δ*t*:

$p(z_i=b \mid z_i=a, \Delta t) = r_i\, \mathbf{S}(\Delta t)_{ab}$ $\quad z_i=a \xrightarrow{\Delta t} z_i=b$

r_i are site-specific rates

$\mathbf{S}(\Delta t) = \mathbf{I} + \mathbf{R}^{\omega}\,\Delta t$: 61×61 substitution matrices

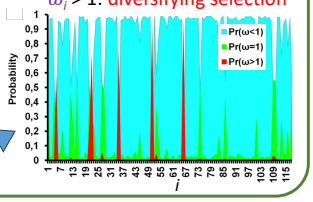
$\mathbf{R}^{\omega}$ is 61×61 rate matrix (with parameter ω).

$$R^{\omega}_{ab} = \begin{cases} R_{ab} & \text{if a}\to\text{b synonymous} \\ \omega R_{ab} & \text{if a}\to\text{b nonsynonymous} \end{cases}$$

Fitting parameters $\omega_i = dN/dS$ on MSA yields selection pressures at each position *i* (e.g. ref[1])

$\omega_i < 1$: purifying selection

$\omega_i > 1$: diversifying selection



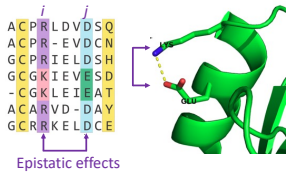
Objective: More realistic model to analyse selection on proteins [2,3]

- Model ***nearly neutral theory of evolution***
- ***Epistatic fitness function***
- ***Heterogeneous codon frequencies***
- ***Rate heterogeneity*** as emergent property [2]
- ***Heterotachy*** as emergent property [2]
- ***GC bias*** (in nucleotide mutation matrix μ_{ab})
- Estimate selection pressure (~dN/dS) for each protein position and each tree node *relative to rest of phylogenetic tree*

PhD 1 - Inference method for epistatic MutSel model of molecular evolution

Goal: Develop 'realistic' epistatic MutSel models, train and use to sample ancestral sequences

Proteins need to maintain stably folding structre. This requires many residue-residue interactions. So, fitness effect of mutations depends strongly on entire sequence ⇒ **strong epistasis**



Probability for mutation a→b

MutSel: $p(\mathbf{z}' \mid \mathbf{z}, \Delta t) = \mu_{ab}\,\Delta t \quad p_{\text{fix}}(s_{\mathbf{z}\to\mathbf{z}'})$

$\mathbf{z} = (z_1 \ldots z_{i-1}\, a\, z_{i+1} \ldots z_L)$

Δt ↓ Mutate a→b

$\mathbf{z}' = (z_1 \ldots z_{i-1}\, b\, z_{i+1} \ldots z_L)$

↓ **Fix b** in pop.

$\mathbf{z}' = (z_1 \ldots z_{i-1}\, b\, z_{i+1} \ldots z_L)$

Fixation probability:

$$p_{\text{fix}}(s_{\mathbf{z}\to\mathbf{z}'}) = \frac{1-\exp(-2\, s_{\mathbf{z}\to\mathbf{z}'})}{1-\exp(-2N\, s_{\mathbf{z}\to\mathbf{z}'})}$$

N = effective population size

Selection coefficient: $s_{\mathbf{z}\to\mathbf{z}'} = f_{v,w}(\mathbf{z}') - f_{v,w}(\mathbf{z})$

Epistatic effects

Fitness function: $f_{v,w}(\mathbf{z}) \propto \left(\exp \sum_i v_{ia}\,[z_i=a] + \sum_{i<j} w_{ia,jb}\,[z_i=a,\, z_j=b]\right)$

Learn $20L + 20^2L^2$ parameters v and w from large **protein language model** (e.g.MSA transformer), fix w, fit only v on MSA. (using *stochastic variational inference* and *Gibbs sampling*)

Sampling ancestral sequences allows for computing uncertainties.

GC bias: set μ^{s}_{ab} such that equilibrium GC frequency is equal to mean in codon sequences of species s. Estimate for ancestors.

PhD 2 - Methods for analysis of selection based on MutSel model with epistatic fitness

Goal: New Bayesian methods for position- and time-resolved analysis of selection on proteins

Current tools for MutSel–based selection analysis not much used

- Not well calibrated: they always yield $\omega \le 1$ (no pos. sel.!) [4]
- No time-resolved analysis over tree nodes
- No epistasis

Calibration is a conceptual issue with common MutSel tools. e.g.

tree nodes codons

$$\omega_i = dN_i/dS_i = \mathrm{N/N_{mut}} = \frac{\sum_n \sum_a \sum_{b\ \text{nonsyn.}} \mu_{ab}\,\Delta t\; p_{\text{fix}}(s_{a\to b})\; p(z_i^n=a)}{\sum_n \sum_a \sum_{b\ \text{nonsyn.}} \mu_{ab}\,\Delta t\; p(z_i^n=a)} \le 1$$

Here

- Include estimates of uncertainties by posterior sampling over ancestral sequences
- Assess selection pressure for each node *n* and position *i*
- Calibrate ω values relative to (constant) fitness model over entire tree:

Realizations of z_i^m given leaf nodes

Expectation values over many samples of ancestral sequences

$$\omega_{ni} = dN_{ni}/dS_{ni} = \frac{\mathbb{E}_{\mathbf{z}}[\sum_a \sum_{b\ \text{nonsyn.}} p(z_i^m=b \mid z_i^n=a, \Delta t)\,]}{\mathbb{E}_{\mathbf{z}}[\sum_a \sum_{b\ \text{syn.}} p(z_i^m=b \mid z_i^n=a, \Delta t)]} \Big/ \frac{\mathbb{E}_{\mathbf{z}}[\sum_a \sum_{b\ \text{nonsyn.}} p(z_i=b \mid z_i^n=a, \Delta t)\,]}{\mathbb{E}_{\mathbf{z}}[\sum_a \sum_{b\ \text{syn.}} p(z_i=b \mid z_i^n=a, \Delta t)\,]}$$

Realizations of z_i^m ignoring leaf nodes

- Open source software

Collaborations with A1, A2, A4, B3, B4, B5

- Develop visualizations / analysis plots "on the job"
- Improve and extend evolutionary model

Collaborations with all RTG projects

- Assistance with deep annotation and seq/structure analysis (Foldseek, Collabfold, HHpred, MMseqs,…)

References

1. Dasmeh P et al. 2013. Positively Selected Sites in Cetacean Myoglobins Contribute to Protein Stability. PLOS Comput Biol 9:e1002929.
2. de la Paz J *et al*. 2020. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. PNAS 117: 5873-82.
3. Vorberg S, Seemayer S, Söding J. 2018. Synthetic protein alignments quantify noise in residue-residue contact prediction. PLOS CB 14 e1006526.
4. Spielman SJ, Wilke CO. 2015. The relationship between dN/dS and scaled selection coefficients. MBE 32:1097-1108.

RTG 2984/1