

**Phylogenetic model of stabilizing selection is more
informative about site specific selection than
extrapolation from laboratory estimates.**

CEDRIC LANDERER^{1,2,*}, BRIAN C. OMEARA^{1,2}, AND MICHAEL
A. GILCHRIST^{1,2}

¹Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996-1610

²National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996-3410

*Corresponding author. E-mail: cedric.landerer@gmail.com

Version dated: September 30, 2018

Introduction

- Phylogenetic inference of sequence relationship was long focused on rates of substitutions.
 - Focus has shifted towards site specific equilibrium frequencies (HB98, Bloom2014, ...)in the last 20 years.
 - Such models however, tend to be not feasible as they are too parameter rich.
 - Inference of site specific selection on amino acids from laboratory experiments e.g. DMS is therefore appealing.
- Incorporation of external information on site specific selection on amino acids allows for the fitting more complex models.
 - This comes with a loss of generality as DMS experiments are limited to fast growing organisms that can be manipulated under laboratory conditions.
 - Strong artificial selection and very heterogeneous population with a lot of competing genotypes are a potential source of bias.
 - In the case of TEM, the application of only one very specific antibiotic is unlikely evolutionary history, may reflect modern hospital environments.
- In this study we will assess how adequate DMS inference of site specific selection on amino acids is using TEM and provide an alternative, more generally applicable solution.
 - Simulations under the DMS inferred site specific selection on amino acids show that we would not expect to observe the natural TEM variants; revealing the inadequacy of DMS.
 - We show that models fits achieved by the incorporation of DMS experiments can be improved upon using a hierarchical phylogenetic framework of stabilizing

selection, SelAC.

- We further show that extrapolation even between sequences (TEM and SHV) with related function can be inadequate.

Results

- Model selection shows that DMS can improve phylogenetic inference.
 - phyDMS improved model fit to 49 TEM sequences by 142 log(likelihood) units
 - number of parameters comparable to GY94 and others despite complex description of fitness landscape thanks to experimental estimates.
- Lab inferences of selection (DMS) are inconsistent with natural sequence evolution.
 - The inferred fitness landscape does not reflect observed sequences.
 - * The optimal amino acid sequence inferred by DMS only shows 49% sequence similarity with the observed sequences.
 - Simulations show that one would not expect the observed sequences under the lab inferred fitness landscape.
 - * We would expect about half of the observed fitness burden.
 - * Sequence similarity is expected to be about $\sim 70\%$.
 - Estimates of selection coefficients do not represent natural evolution.
 - * Due to artificial selection environment; Heterogeneous population, very large s .
 - * Only one antibiotic used, maybe a mixture of antibiotics would better reflect natural evolution.
- SelAC better explains observed sequences than DMS and other models.

- Model selection shows that SelAC outperforms phydms (only for AIC).
- Model adequacy shows that SelAC better represents the observed sequences.
- SelAC is a more general approach, applicable to all protein coding sequences.
 - Application of SelAC to TEM, site specific estimates of aa fitness.
 - * most sites show the estimated optimal amino acid.
 - * We find that selection against used amino acids is clustered and locally confined.
 - Comparison between TEM and SHV reveals that extrapolation is not always a good idea.
 - * Site specific G terms for TEM and SHV are only weakly correlated ($\rho = 0.17$), despite similar α_G .
 - * Greatest difference is observed in the physicochemical properties, specifically α .