

**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: October 2, 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 unfeasible as they are very 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.
 - 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 competing genotypes are a potential source of bias.
 - 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 using DMS inferred site specific selection on amino acids show that observed TEM variants are unexpected; revealing the inadequacy of DMS.
 - Models fits achieved by the incorporation of DMS experiments can be improved upon using a hierarchical phylogenetic framework of stabilizing selection; SelAC.

- Extrapolation 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 XXX AICc 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 (Figure 1).
 - Observed sequences unlikely under the lab inferred fitness landscape (Figure 2a,b).
 - * 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.
 - * Lack of repeatability between labs introduces further problems (Firnberg et al 2014 vs. Stiffler et al. 2016).
- SelAC better explains observed sequences than DMS and other models.

- Model selection shows that SelAC outperforms phydms (Table 1).
- 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 (Figure 3).
 - * 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 (Figure 4a).
 - * Greatest difference is observed in the physicochemical properties, specifically α (Figure 4b).

Discussion

- Incorporating selection into phylogenetic frameworks is already a long lasting endeavor.
 - As the type of selection on a protein is not always clear, or differs between proteins phylogenetic models have to make generalizing assumptions.
 - Incorporating selection from experimental sources therefore seems like an attractive option.
 - Incorporating empirical fitness has some important features.
 - * It allows for site specific amino acid preferences, acknowledging the heterogeneity of selection along the protein sequence.

- * It greatly reduces the number of parameters that have to be estimated from the data.
- However, the incorporation of empirical fitness also has some important shortcomings.
 - * DMS experiments are limited to proteins and organisms that can be manipulated under laboratory conditions.
 - * But even in the case of TEM, the applied selection pressure is limited to the defense against a specific antibiotic.
 - * TEM, however, has evolved to compete against con-specifics using secreting metabolites to gain an advantage.
 - * Furthermore, DMS relies on a library of mutants and therefore on a heterogeneous population with competing genotypes.
 - * Therefore, it is important to ask how adequate such experiments reflect natural evolution.
- We evaluated how well experimental selection estimates from DMS experiments explain natural sequence evolution and compared it to a novel phylogenetic framework, SelAC.
 - * Previous work has shown that DMS selection estimates can improve model fit over classical approaches like GY94 and our work confirms this.
 - * Model selection favored the SelAC model fit and the corresponding fitness estimates over the DMS estimates using both, SelAC and phyDMS (Table 1).
- Adequacy of the DMS selection has previously not been assessed.
 - The amino acid with the cumulative highest fitness experimentally estimated with DMS only has 49% concordance with the observed alignment.
 - In contrast, the SelAC estimate has 99% concordance (Figure 1).

- Assuming that the DMS selection inference adequately reflects natural evolution, the observed TEM sequences are either mal-adapted or where unable to reach a fitness peak.
 - *E. coli* has a large effective population size, estimates are on the order of 10^8 to 10^9 (Ochman and Wilson 1987, Hartl et al 1994).
 - The large N_e would allow *E. coli* to effectively "explore" the sequence space, thus suggesting that the TEM sequences are mal-adapted according to the DMS estimates.
 - Our simulations of sequence evolution with various N_e values and the DMS fitness values in contrast show that we would expect higher adaptation even with much smaller N_e (Figure 2).
- DMS estimates of the observed TEM variants predict them to be mal-adapted while SelAC predicts most TEM variants to be well adapted.
 - Given *E. coli*'s large effective population size, the efficacy of selection should be very large.
 - We therefore expect the observed sequence variants to be at the selection-mutation-drift barrier, which in turn can expected to be near the optimum.
 - We find the majority of sequences near the optimum, therefore the SelAC estimates are consistent with theoretical population genetics results.
 - In contrast, finding strong selection against the observed TEM variants indicates that DMS is not consistent with theoretical population genetics expectations.
 - This is consistent when thinking about that DMS only reflects the selection on the TEM sequence with regards to one antibiotic, which seems appropriate to model selection in modern hospital environments but not when the interest lies in the natural evolution of TEM.

- We find that SelAC produces similar selection against the observed TEM variants if we assume the fitness peaks (optimal AA) that are estimated by DMS.
 - This shows that DMS and SelAC can provide consistent estimates of selection against amino acids.
 - SelAC has the advantage that it can be applied to any protein coding sequence alignment.
 - This removes the need for extrapolation e.g. from TEM to SHV.
- Difference in selection parameters between TEM and SHV indicate that extrapolation is not a good idea.
 - The difference in the site specific strength of selection shows that TEM and SHV are facing different selection pressures.
 - this is also highlighted by the differences in physicochemical weightings between the two proteins.
- SelAC outperforms DMS, but is not without flaws itself
 - Like DMS, SelAC assumes site independence and therefore rather than providing a complete fitness landscape provides a projection of the fitness landscape.
 - SelAC is a model of stabilizing selection, which is in contrast to e.g. GY94 which is a model of frequency dependent selection.
 - * Since TEM plays a role in the chemical warfare with con-specifics, some sites may be under diversifying selection.
 - * However, we do not observe any evidence even when allowing site specific estimates of G to be negative.
 - * Low sequence variation in the TEM may be cause for concern as it could be misinterpreted by the model as stabilizing selection because of the short branches.

Model	L	n	AIC	Δ AIC
SelAC	-1498	374	3744	0
SelAC+DMS	-1768	111	3758	14
phyDMS	-2060	105	4331	586

Table 1: L , number of model parameters n , AIC, and Δ AIC., Full table has > 200 models

- * However, provided our simulations we can be confident that TEM is actually under stabilizing selection
- SelAC assumes that selection is proportional to distance in physicochemical space.
- * We used Grantham (1974) distances, however many other distances are available which may an even better model fit.
- In conclusion, DMS experiments have been proposed to supplement information on selection on amino acids in phylogenetic studies.
- This study shows that information on selection can be extracted from alignments of protein coding sequences.
- This highlights the limitations of DMS to explain natural evolution.

Tables

Figures

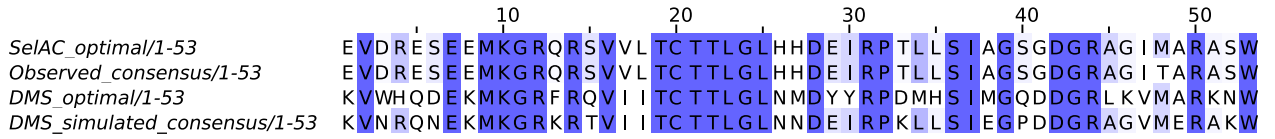


Figure 1: Every 5th residue. DMS and simulation based on DMS do not reflect natural sequences

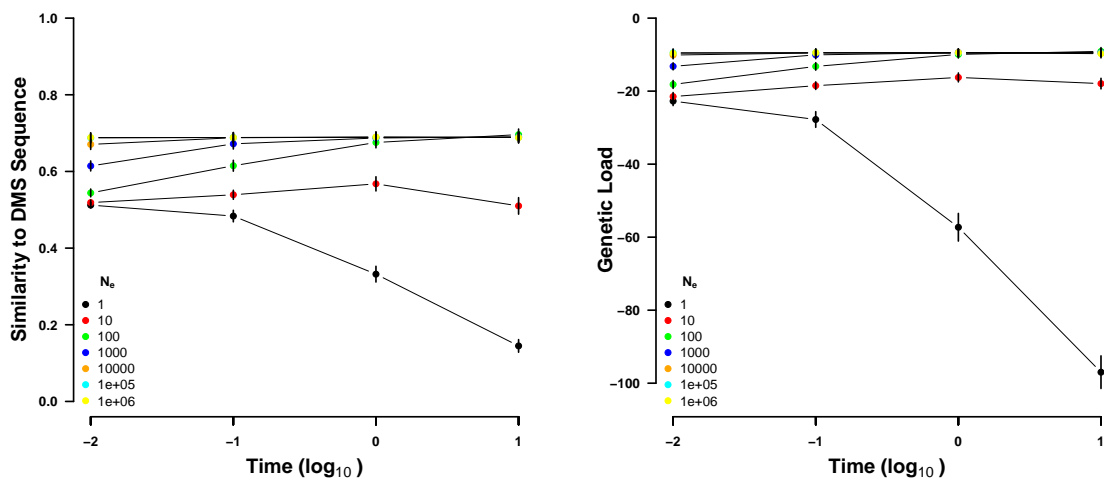


Figure 2: Sequences simulated under various values of N_e and for various times.

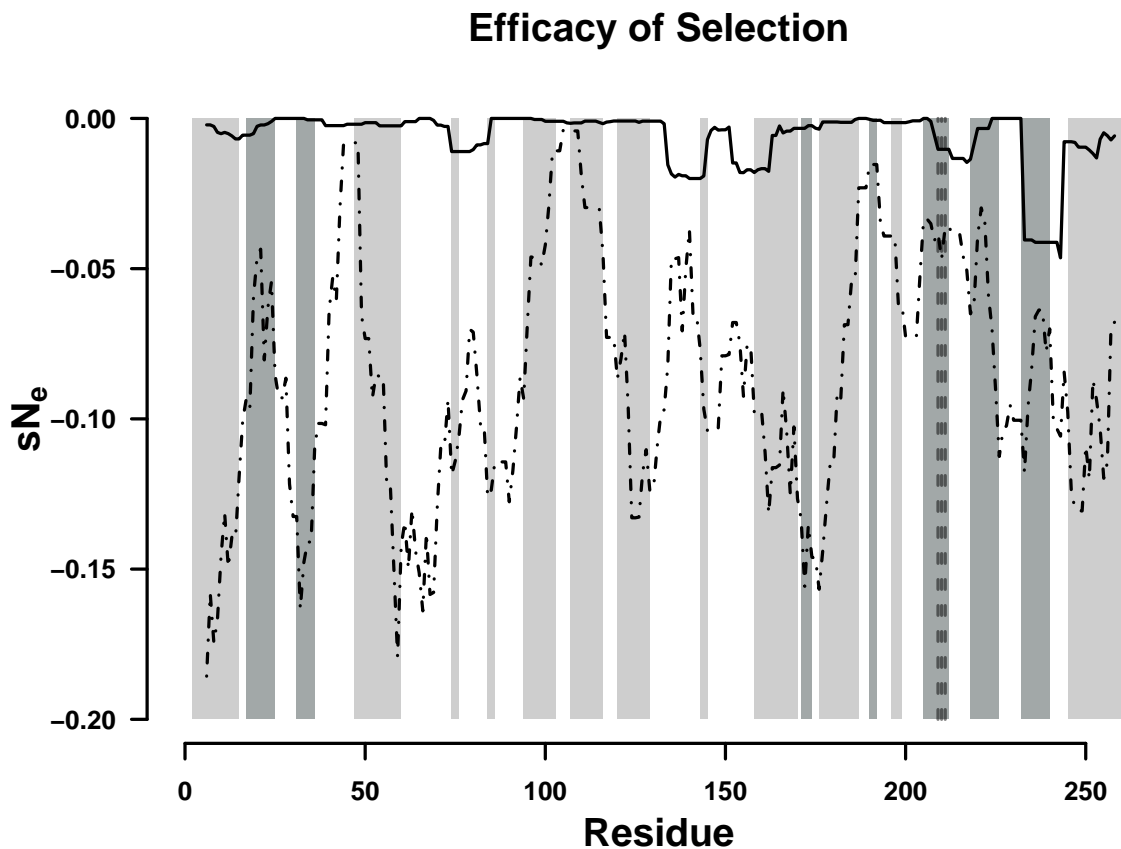


Figure 3: TEM, bars are different secondary structure elements. Dashed dotted line is DMS, solid is SelAC sNe, all lines are means of all sequences, sliding window of 10 sites. vertical lines are active/binding sites.,

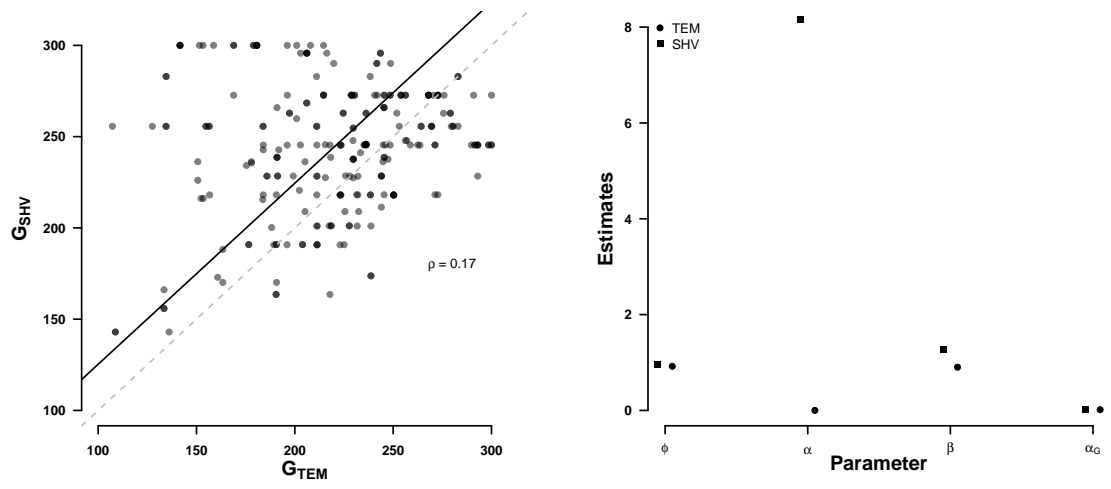


Figure 4: Comparisson of selection related parameters between TEM and SHV.

Suppl. Figures

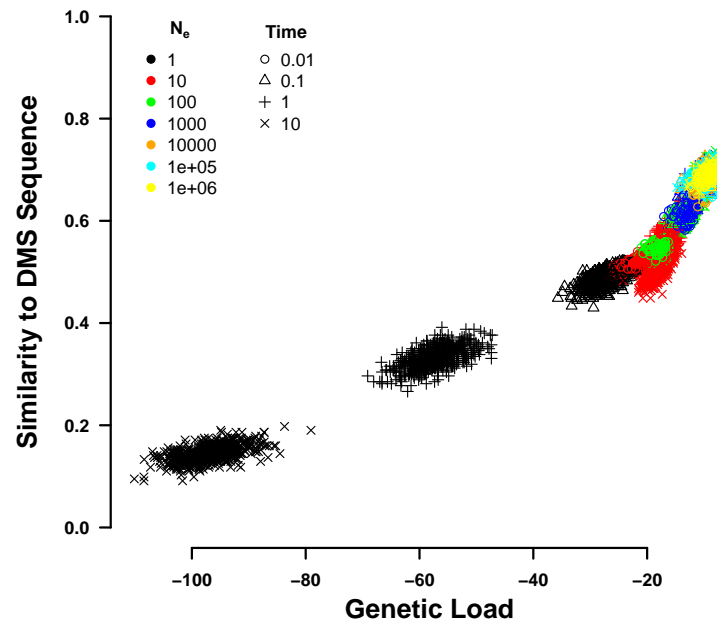


Figure 5: Suppl: Sequences simulated under various values of N_e and for various times.
TODO: replace clouds by mean+sd bars

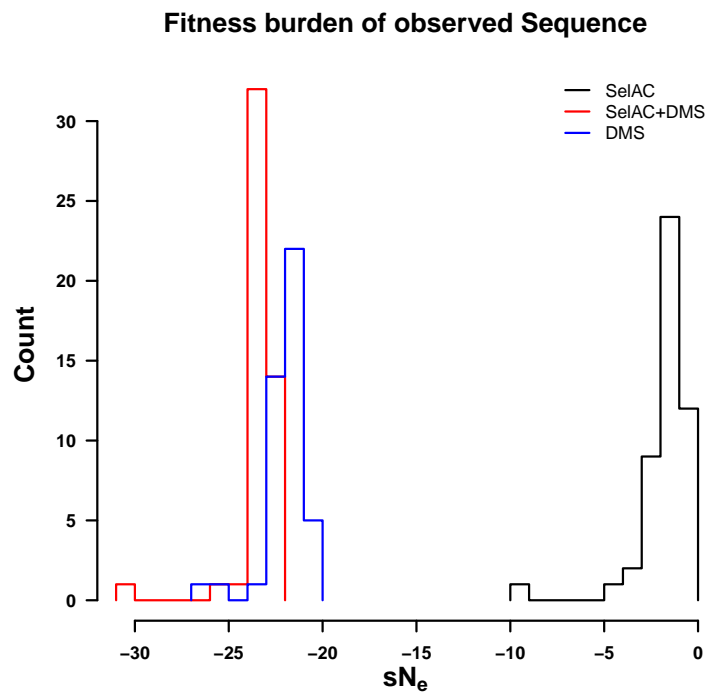


Figure 6: Suppl: sN_e of whole sequence, variation across tips. TEM