Estimating the genetic load of natural protein coding sequences using a phylogenetic framework.

Abstract

CEDRIC LANDERER[1,2,*], BRIAN C. OMEARA[1,2], AND MICHAEL A. GILCHRIST[1,2]

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

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

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

Version dated: September 25, 2018

Abstract

Protein production is a very costly process every cell performs, resulting in selection for proteins that can perform their function most efficiently. The efficacy of selection is limited by the effective population size N_e, leading to genetic load via the introduction of mutations. As all proteins have to face this selection-mutation-drift barrier, we expect to find proteins near a fitness peak, but never at the peak. Here, we assess the efficacy of selection on individual proteins and quantify the genetic load by phylogenetic inference of the optimal amino acid at each site using SelAC. We compare our estimates of site specific amino acid fitness and genetic load to empirical estimates from deep mutation scanning experiments. Our work demonstrates the shortcomings of empirical fitness estimates obtained under laboratory conditions and highlights the general applicability of SelAC. Using model selection, we show that phylogenetic estimates of fitness are preferred over empirical estimates (ΔAIC = 586). Using the empirical estimates, the genotype with the highest fitness only shows 49% sequence identity with the observed TEM variants. Simulations reveal that the empirical fitness values do not adequately reflect natural evolution as we would not expect to observe the natural TEM sequence variants. Furthermore, we demonstrate the generality of SelAC by estimating the genetic load of cytochrome B in whales. These results indicate that genetic load varies greatly along cytochrome B.

Introduction

- NOTE: Use Substitutional Load instead (Kimura and Maruyama 1968)? Initial frequency implicit 1/2Ne?
- Genetic load is a measure of distance between the average genotype's fitness and the genotype with the highest fitness.
  - The genotype with the highest fitness is assessed based on the set of observed genotypes.

  - Mutation constantly introduces new, potentially deleterious mutations increasing the genetic load of a population.
  - Genetic drift limits the efficacy of natural selection.
  - Therefore, the optimal genotype is likely not among the observed genotypes.
  - To remedy this, experimental procedures like deep mutation scanning can be employed to asses the fitness of genotypes.

- Deep mutation scanning (DMS) requires a library of mutants for which the fitness should be assessed.
  - This limits the application of DMS experiments to organisms that can be manipulated under laboratory conditions, and have a sufficiently short generation time.
  - It also requires that artificial selection can be applied.
  - This limits DMS experiments even further to proteins for which we can assume they respond to a singular stress factor.
  - While it is save to ignore effects of mutation, the low population size does severely limit the efficacy of selection.
  - It is therefore required to apply extremely strong selection pressure.

- In this study, we assess how well DMS experiments are suited to assess genetic load produced by natural evolution.
  - It has previously been demonstrated that incorporation of DMS experiments into phylogenetic approaches improve model fit when compared to classical approaches like GY94.
  - However, model adequacy has not been assessed.

- First we show that the models fits achieved by the incorporation of DMS experiments can be improved upon using a novel phylogenetic framework, SelAC.
- We find that we would not expect to observe the natural TEM variants when simulating under the DMS inferred fitness landscape.
- We then compare the genetic load of natural TEM variants according to DMS and SelAC and show that DMS predicts an increased genetic load.

- Having shown that SelAC provides more adequate inference of genetic load we further demonstrate its generality by assessing the genetic load of cytochrome B in whales, an organisms for which DMS experiments are not possible.

Results

- We used SelAC and phyDMS to compare model fits to TEM sequence variants.
  - AIC values showed that SelAC provided an improved model fit (Tabel 1).
  - Ignoring the phylogenetic relationship, sequence comparison reveals that the sequence with the highest cumulative fitness according to DMS only shows $\sim 49\%$ agreement with the consensus of the observed TEM variants (Figure 1).
  - In contrast, the optimal amino acid sequence inferred by SelAC shows 99% sequence similarity.

- Simulations of sequence evolution using the site specific DMS fitness estimates show that DMS does not reflect natural sequence evolution.
  - Assuming reasonable, but still small effective population sizes for *E. coli* ($10,000-1,000,000$), we would expect to observe a sequence similarity of $\sim 70\%$ (Figure 2a).

- We also expect to only observed half the genetic load (Observed mean: ~ 22 v Expected mean: ~ 10) (Figure 4a and 2b)
- However, even with an effective population size as small as 100, we would expect a significantly lower genetic load than observed.

- SelAC estimates a much lower genetic load ($3 - 20$ fold) than DMS (Figure 4a).
  - Using the DMS estimated optimal amino acid sequence, SelAC estimates similar genetic loads to DMS.

- Using SelAC, we estimated the genetic load each site carries from the alignment.
  - The alignment of observed TEM variants has high homogeneity; 68% of sites had only one codon present; 75% of sites encoded the same amino acid.
  - Increases of genetic load appears to be clustered, mostly between secondary structure elements but not limited to unstructured regions (Figure 5).
  - We find that the DMS genetic load is always greater than the genetic load inferred by SelAC.
  - We also find an increase in genetic load at the catalytic triad.

- Highlighting the generality of a phylogenetic approach, we estimated the genetic load of Cytochrome B in a small set of whales.
  - The optimal amino acid sequence inferred by SelAC shows 95% sequence similarity.
  - This is a slightly lower agreement than in the TEM case, however, CytB is less homogeneous as well; 22% of sites had only one codon present; 78% of sites encoded the same amino acid.
  - Genetic load and variation carried by CytB sequences is higher than for TEM variants (Figure 4b).

- Genetic load also does not appear to be clustered, but spread out over the whole sequence, with the highest load located within the 5th alpha helix.
- Genetic load appears to decrease closer to the active sites, with the exception of the binding site at the end of the 4th alpha helix.

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 short comings.
    * 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 population with very low N_e.
    * 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.
    * Adequacy of the DMS selection has not been assessed.

- Model selection favored the SelAC model fit and the corresponding fitness estimates over the DMS estimates using both, SelAC and phyDMS (Table 1).
  - 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 them to be near the optimum.
  - 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.
  - Further, we find that SelAC can recreate the DMS genetic loads if the DMS optimum is assumed.
- Mapping of the genetic load on to the TEM sequence revealed clusters of increased variation mostly located between secondary structure elements.
  - Unstructured regions tend to be more robust to amino acid substitutions.
  - Despite the clustering in the unstructured regions, the highest peak in genetic load is observed in the last beta-sheet (Anything special about this?).
- CytB has a higher genetic load in whales than TEM in *E. coli* (Figure 4)
  - No surprise here as whales have a much lower N_e, therefore the efficacy of selection is weaker.
  - CytB shows more variation in genetic load across the sequence.
  - Some binding site have increased, genetic load, some none (Figure 6).
- Caveats
  - DMS and SelAC assume site independence.

Model	L	n	AIC	ΔAIC
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 ΔAIC.

- SelAC assumes stabilizing selection, reasons why that might not be the case?
- SelAC assumes selection is proportional to distance in physicochemical space, a lot of different properties, maybe we used the wrong ones.
- SelAC is limited by observed sequence variation

- 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.

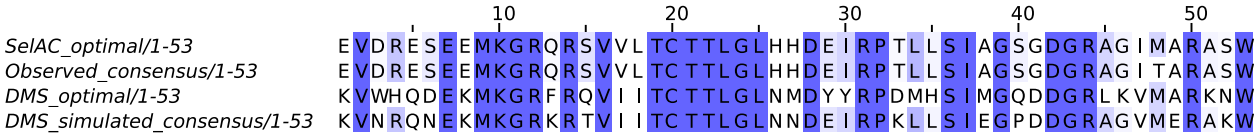

Figure 1: Every 5th residue. DMS and simulation based on DMS do not reflect natural sequences

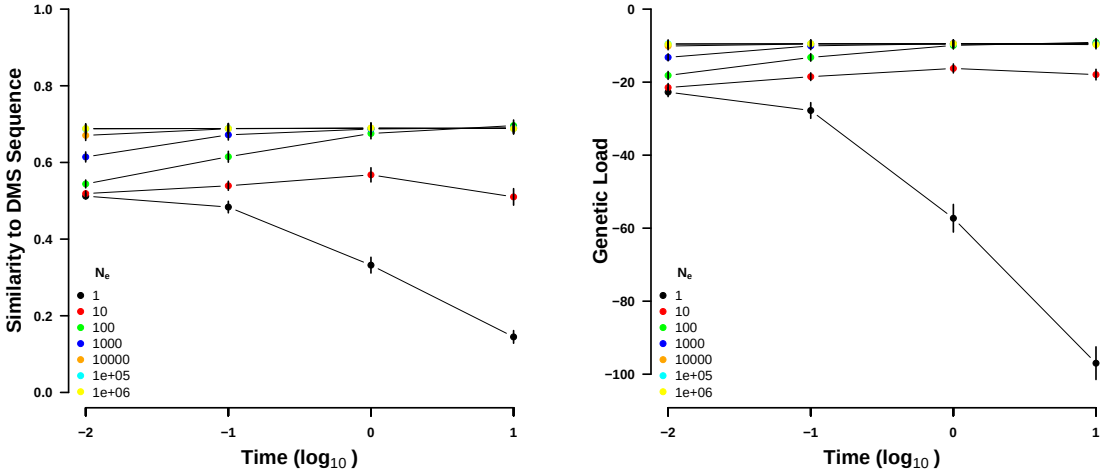

Figure 2: Sequences simulated under various values of N_e and for various times.

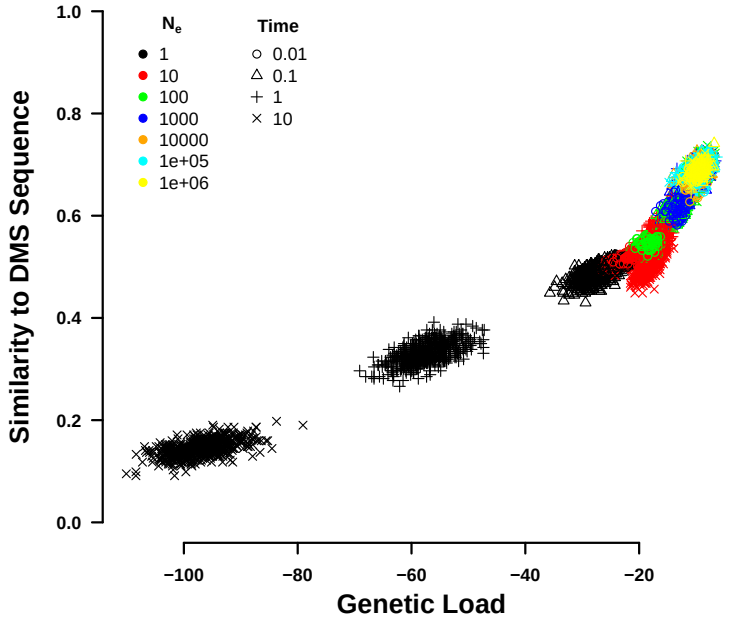

Figure 3: Suppl: Sequences simulated under various values of N_e and for various times. TODO: replace clouds by mean+sd bars

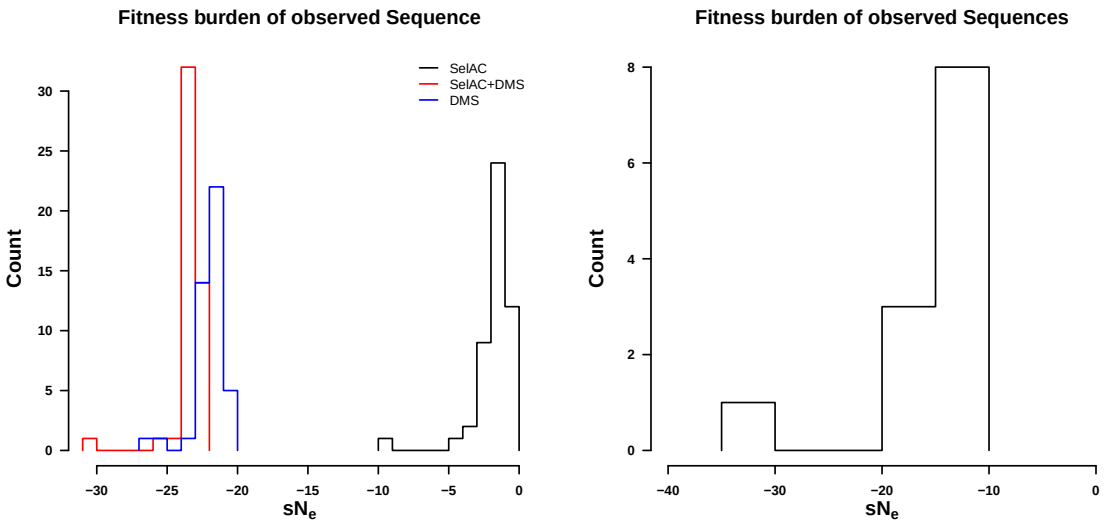

Figure 4: sN_e of whole sequence, variation across tips. TEM(left), CytB(right)

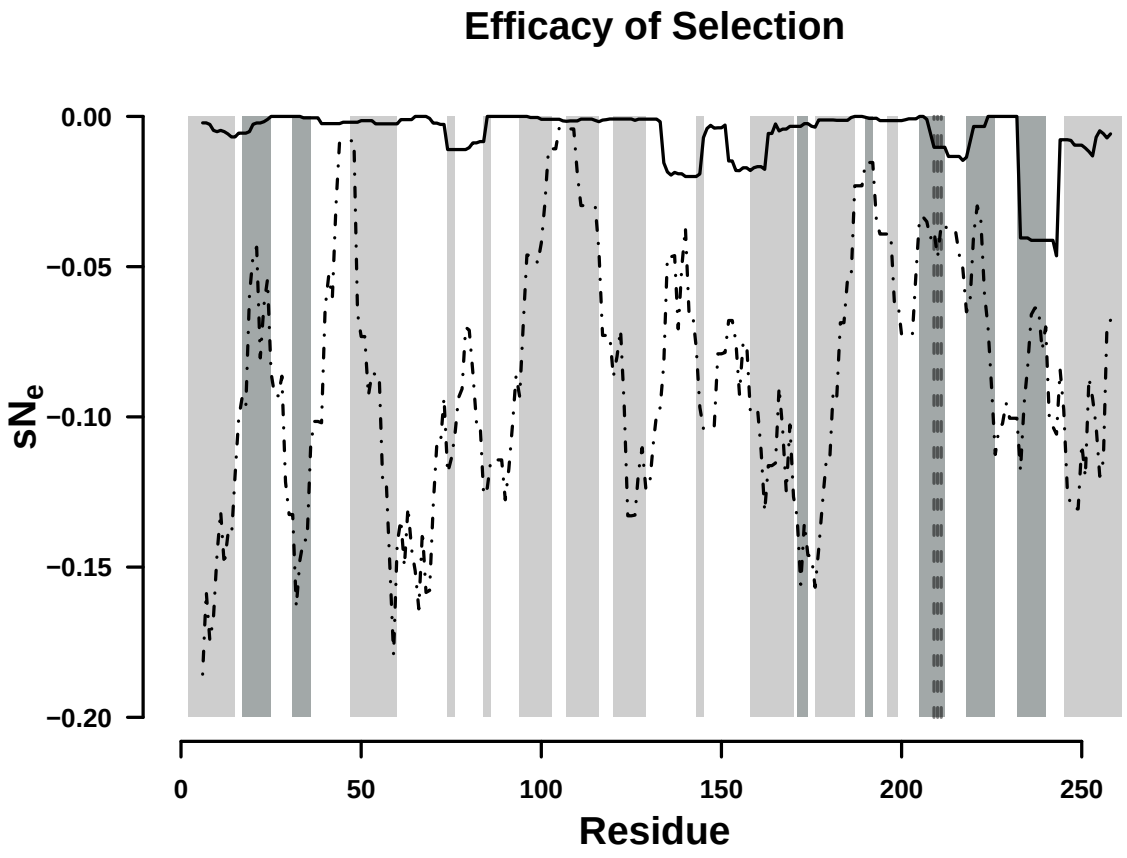

Figure 5: TEM, bars are different seconday structure elements. Dashed dotted line is DMS, solid is SelAC sNe, all lines are means of all sequences, sliding window of 10 sites. vertical lines are active/binding sites.,

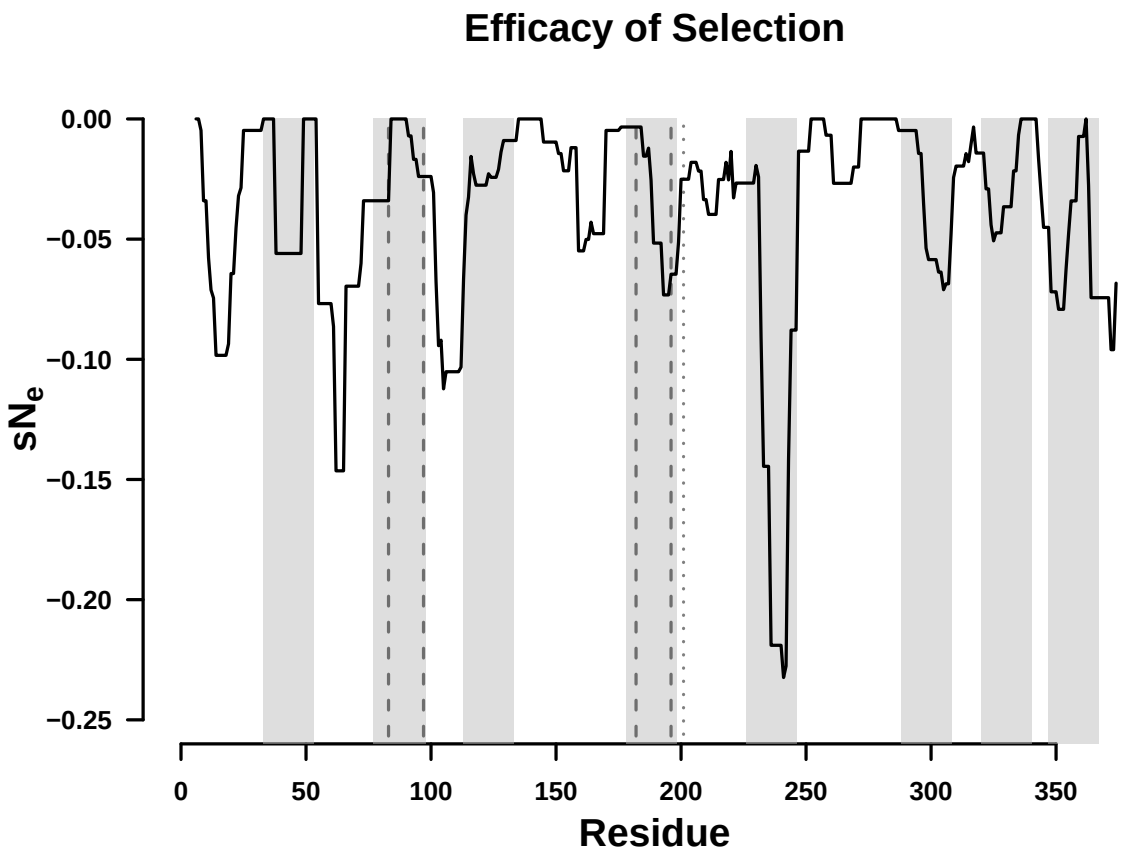

Figure 6: solid lines is average Genetic Load of CytB alignment,. dashed and dotted lines are different types ofbinding sites. Horizontal bars are alpha helices.