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 24, 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 demonstrate the assessment of the genetic load proteins impose for TEM in *E. coli* and for cytochrome B in whales. We quantify the genetic load for 49 TEM sequences and 12 cytochrome B sequences. We compare the inferred optimal TEM amino acid sequence to fitness estimates from deep mutation scanning experiments. We find that the observed TEM sequences have a 3 to 20 fold increased genetic load when compared to the DMS estimates instead of the SelAC inference. Furthermore, we find that the DMS inference only shows 49 % sequence agreement with the consensus sequence of the observed alignment. We also observe a higher genetic load in CytB than in TEM, which is to be expected given the difference in N_e between whales and *E. coli*.

Introduction

- 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 assed 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.

- NOTE: Use Substitutional Load instead (Kimura and Maruyama 1968)? Initial frequency implicid 1/2Ne?
  - 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 then compare the genetic load of natural TEM variants according to DMS and SelAC and show that DMS predicts an increased genetic load.
  - We find that we would not expect to observe the natural TEM variants when simulating under the DMS inferred fitness landscape.

- 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.

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 acroding 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 populaiton size as small as 100, we would expect a significantly lower genetic load than observed.

- In contrast, SelAC estimates a much lower genetic load ($3 - 20$ fold) (Figure 4a).

- Using SelAC, we estimated the genetic load each site carries from the alignemnt.
  - The alignement 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 apear 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.

- 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 homogenous 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 apear to be clustered, but spread out over the whole sequence, with the highest load located within the 5th akpha helix.
  - Genetic load apears to decrease closer to the active sites, with the exception of the binding site at the end of the 4th alpha helix.

Discussion

- We demonstrate the inference of site specific selection from protein coding sequence data using phylogenetics.
  - We find that the fitness landscape estimated by SelAC better explains the observed sequences than DMS.
  - Simulations show that the observed sequences would not arrise under the imposed selection during the DMS experiments.
  - While DMS allows for the inferece of properties such as substitutions confering antibiotic resistance, it does not allow to explain natural sequence variation.
  - In addition, as researrchers show more and more interests in non-model organisms, the limitation to proteins and organsism that can be manipulated under laboratory conditions limits its uses across the tree of life.
- We estimate the genetic load of natural occurring proteins relative to an inferred optimal amino acid sequence.
- The optimal amino acid at each site was inferred from the observed proteins and their phylogenetic relationship.
- In both cases, TEM and CytB, we find high agreement between the consensus sequence inferred by ignoring the phylogenetic relationship and the optimal sequence inferred using SelAC (TEM: 99%, CytB: 95%).
  - The strong agreement between consensus sequence and estimated optimal sequence for both proteins can be seen as an indication that the phylogenetic relationship does not play a large role in the examined cases.
  - However, such an assumption should not be made a priory.

- The similarity between consensus and predicted optimal sequence could be because the proteins are under stabilizing selection like the model assumes, because rate of shifts in the optimal amino acid sequence is low, or because not enough time has passed for shifts to occur, despite diversifying selection.
- The used alignments contain a high amount of homogeneous sites (TEM: 75%, CytB: 78%), thus these sites do not allow for the inferred optimal amino acid to deviate from the observed consensus.

- In contrast, the experimentally inferred optimal amino acid sequence for TEM only has 49% agreement with the observed consensus.
  - Assuming that this inferred sequence is free of any bias introduced by the experimental conditions, we could only come to the conclusion that the observed TEM sequences show either strong mal-adaptation or did not have enough time to evolve towards the optimal sequence.
  - However, *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.
  - On the other hand, each mutation in the library used for the DMS experiments starts of with only a few copies, potentially biasing the results due to strong genetic drift.

- The genetic load of the observed sequences was inferred relative to the optimal amino acid sequence estimated by SelAC.
  - Both, CytB and TEM show variation in the genetic load represented by each observed sequence, CytB represents a higher genetic load than TEM.
  - Most TEM sequences show a small genetic load, likely due to the high selection pressure on TEM due to its usage in chemical warfare between microorganisms.

    * If the experimental sequence is assumed to be most optimal, the observed TEM proteins represent a high genetic load to the organism.
    * This is would be in conflict with a large effective population size and therefore high efficacy of selection.
    * However, while this would make fixation unlikely, it would not be impossible.
    * In addition, the experimental sequence was inferred based on small population sizes for each genotype and artificial selection pressure.

- Genetic load varies across the sequence.
    - For both proteins, variation of sN_e across the sequence is not associated with any particular structural features but mostly with variation in the alignment.
    - However, TEM shows increased genetic load near the binding site, and the highest genetic load is found in the last beta sheet of the protein.
    - The genetic load is generally higher for CytB than for TEM, and like for TEM genetic load appears to increase around the binding sites.
    - However, for both proteins, increases in genetic load are not limited to the binding sites.

- DMS experiments have been incorporated into phylogenetic studies to supplement information on selection on amino acids.
    - In contrast, this study shows that information on selection can be extracted from alignments of protein coding sequences.
    - To no surprise, model selection clearly favored the optimal sequence inferred by SelAC when using SelAC, however, when using this sequence in phydms we find that the inferences from SelAC still explained the data better, but the increase in parameters did not merit the increase in likelihood.
    - This highlights the limitations of DMS sequences to explain natural evolution.

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.

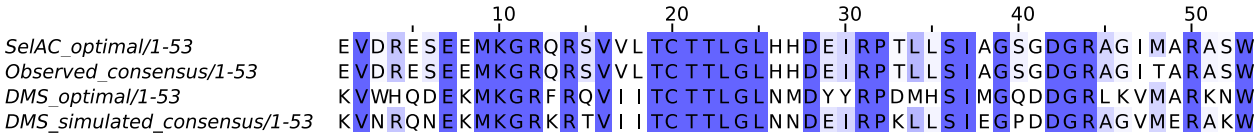

Figure 1: Every 5th residue. DMS and simulation based on DMS do not reflect natural sequences

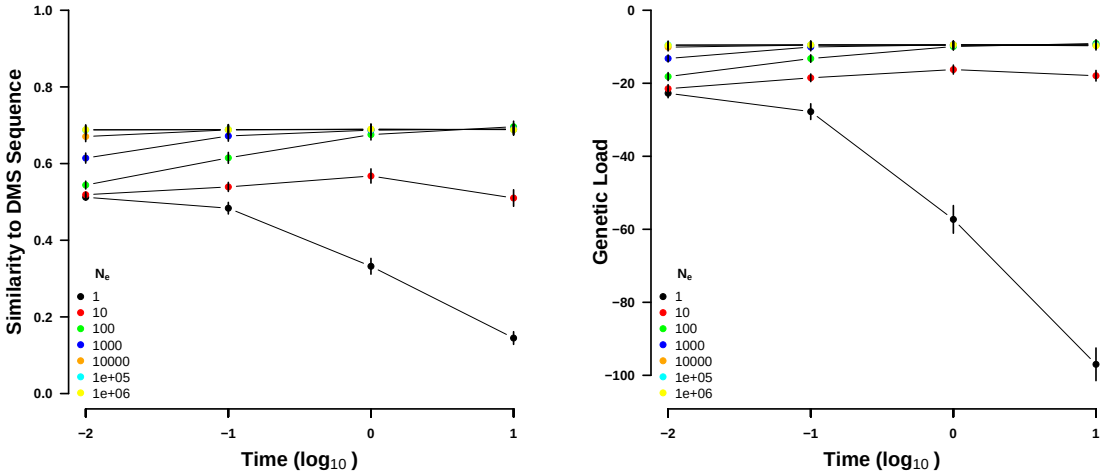

Figure 2: Sequences simulated under various values of N_e and for various times.

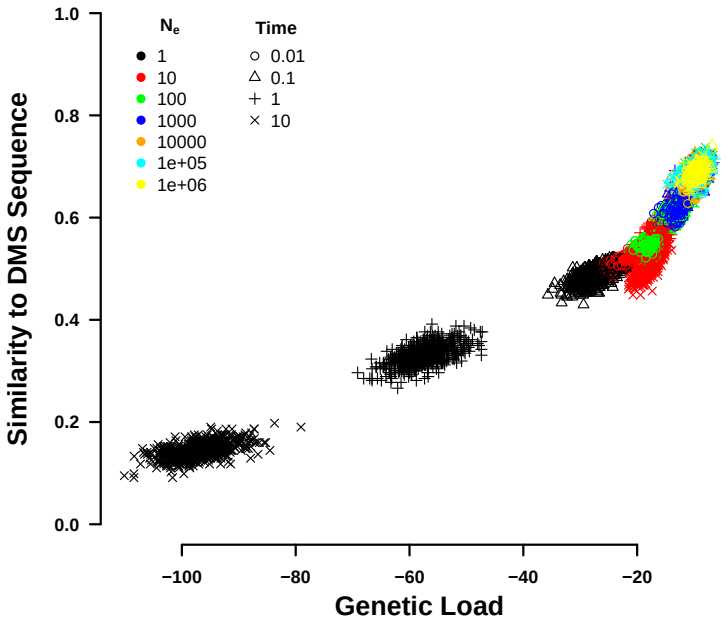

Figure 3: Suppl: Sequences simulated under various values of N_e and for various times. TODO: replace clouds by mean+sd bars

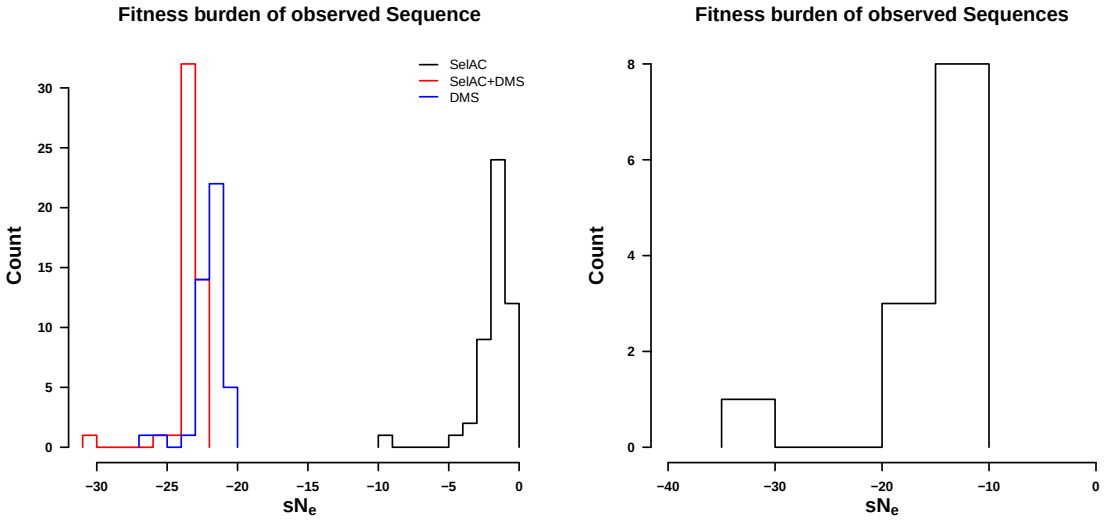

Figure 4: sN_e of whole sequence, variation across tips. TEM(left), CytB(right)

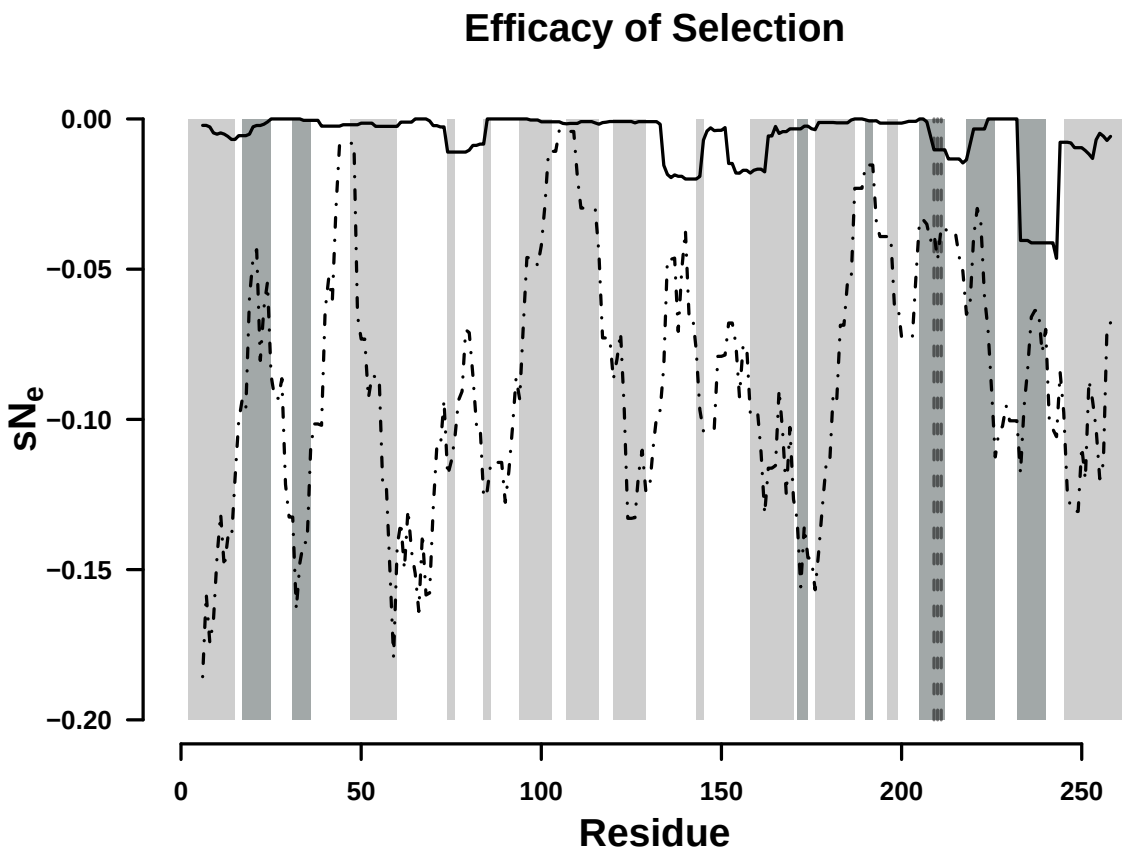

Figure 5: TEM, bars are different seconday structure elements. Dashed dotted line is DMS, solid is SelAC sNe, all lines are means of all sequences, sliding window of 10 sites. vertical lines are active/binding sites.,

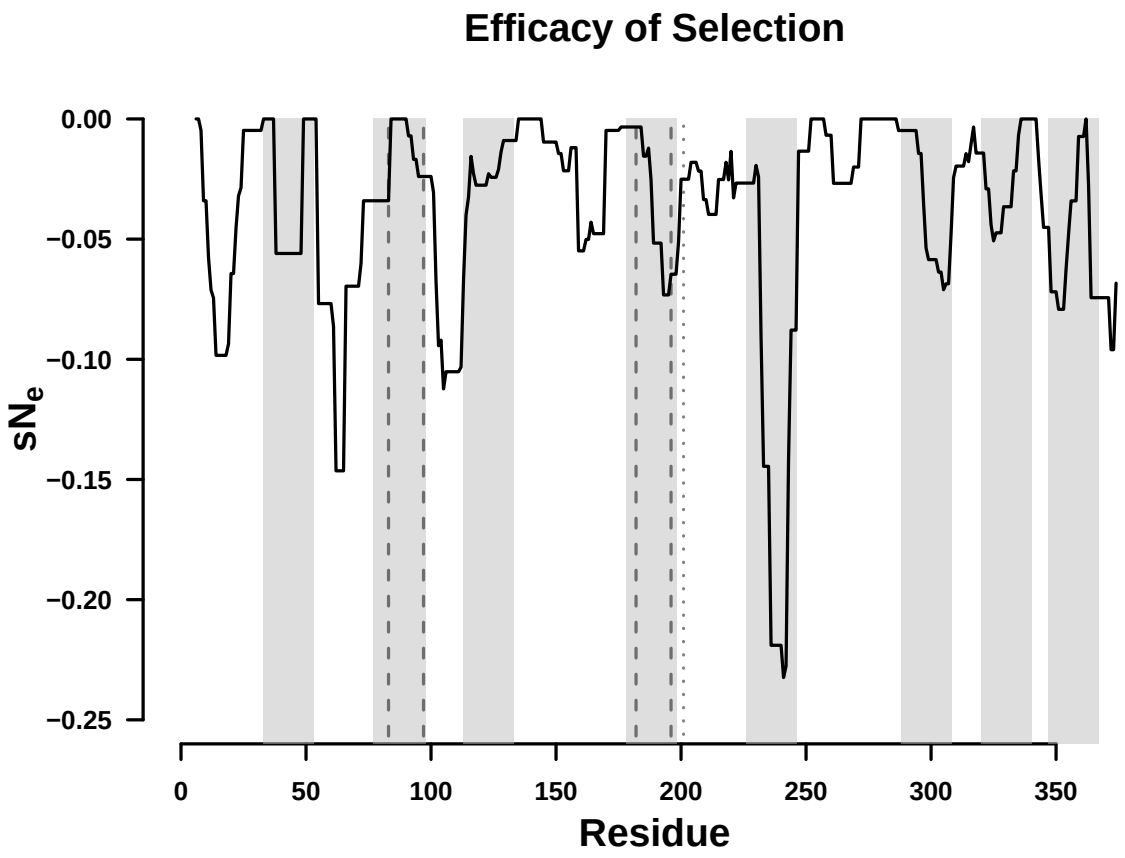

Figure 6: solid lines is average Genetic Load of CytB alignment,. dashed and dotted lines are different types ofbinding sites. Horizontal bars are alpha helices.