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 1, 2018

Abstract

Protein production is a very costly process that every cell performs, resulting in selection for proteins that can perform a function optimal. The efficacy of selection is limited by the effective population size N_e, resulting in genetic load due to the introduction of mutations. As all proteins have to face this selection-rift 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 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 was to be expected given the difference in N_e between whales and *E. coli*.

Introduction

- Natural selection favors proteins with greater functionality.
  - However, selection can be overpowered by mutation or drift causing proteins to move away from the optimum.
- Genetic load is usually assessed relative to a predefined wild-type.
  - One can, however, assess genetic load relative to the genotype encoding a function of interest most optimally.

- However, this requires to assess the fitness of each genotype.

- Previously, deep mutation scanning (DMS) experiments have been utilized to assess site specific amino acid fitness for a variety of proteins.
  - However, these experiments are limited to fast growing organisms that can be manipulated under laboratory conditions, and proteins where a specific selection pressure can be applied.
  - Furthermore, DMS experiments utilize prepared libraries containing each genotype (ignoring epistatis), causing extremely low effective population sizes.
  - Thus, while mutation does not play a role, genetic drift reduces the efficacy of selection dramatically making it necessary to apply extremely high selection pressures.

- We utilize SelAC, a phylogenetic framework, to assess the genetic load of naturally occurring sequence variation on the species level.
  - SelAC is a mechanistic phylogenetic model rooted in population genetics, and estimates site specific selection from sequence data.
  - SelAC does not assume a uniform stationary amino acid distribution across sites, thus allowing it to estimate the optimal amino acid for each position given the available sequence data.
  - Furthermore, SelAC is applicable to the whole tree of life and not limited to fast growing organisms that can be manipulated under laboratory conditions.

- We predict the site specific optimal amino acid from sequence alignments of TEM, a β-lactamase in E. coli and cytochrome b (CytB), a mitochondrial transmembrane protein in whales.

- We then assess the genetic load of naturally occurring sequences TEM and CytB relative to the predicted functionally optimal amino acid sequence.

- We compare our genetic load estimates for TEM to empirical DMS estimates and find an increase of genetic load.

- Furthermore, we will illustrate how the strength of selection varies along the analyzed proteins.

Results

- We predicted the functionally optimal amino acid at each site from the observed sequence variation using SelAC.
  - The observed TEM alignment shows a high percentage of homogeneous sites.
    * 68% of sites had only one codon present.
    * 75% of sites encoded the same amino acid.
  - The observed CytB alignment shows a more codon heterogeneity but a similar homogeneity in amino acids.
    * 22% of sites had only one codon present.
    * 78% of sites encoded the same amino acid.
  - We find that the predicted optimal amino acid sequence has high agreement with the observed consensus sequence of the alignment (TEM: 99%, CytB: 95%).
  - In contrast, the experimentally obtained sequence estimate only has an agreement of 49% with the observed TEM consensus sequence.

- We assessed the genetic load of the observed sequences.
  - We find that the genetic load of TEM differs greatly depending on the optimal amino acid sequence assumed.

    * The genetic load of the observed sequences increases $3 - 20$ fold when using the experimentally inferred optimal sequence compared to the SelAC inferred optimal sequence.
    * Besides the great variation that arises from the usage of different optimal amino acid sequences, we also find variation within each optimal amino acid sequence.
    * E.g. sN_e varies between ~ 0 to ~ -10 for the SelAC optimal sequence and between ~ -20 to ~ -27 for the optimal sequence obtained from the DMS experiment.
  - We lack the ability to compare our estimates of genetic load for CytB as DMS experiment can not be performed on whales.
  - We find a higher genetic load and greater variation in sN_e for the CytB (not taken into account: sequences differ in length).
    * sN_e varies between ~ -10 to ~ -35.

- We are able to map variation in selection along the sequence and determine sites with higher contribution to genetic load.
  - Increases in genetic load appear to be locally confined to a few regions among the TEM alignment but do not appear to be associated with any particular structural features.
  - In contrast, CytB shows variation of genetic load across its whole sequence with a particularly strong increase in genetic load within the 5th transmembrane helix.

- Previous work highlighted the advantages of DMS experiments for phylogenetic inferences.

- However, our estimates of genetic load of observed TEM sequences show that natural sequences would actually represent a large genetic load.

  - The SelAC estimated optimal amino acid sequence outperformed the consensus sequence and the experimentally sequence explaining the data.
  - A second model selection was performed using phydms as an independent comparison.
    * Model selection revealed that the main advantage of the DMS experiment comes from the fact that the input alignment is not needed to estimate amino acid preferences.
    * While the experimentally inferred optimal sequence does a worse job explaining the observed sequences, model selection reveals that the improvement in likelihood does not justify the increased number of parameters required to run phydms with the SelAC or the consensus amino acid preferences.

Discussion

- We demonstrate the inference of site specific selection from protein coding sequence data using phylogenetics.
- 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.

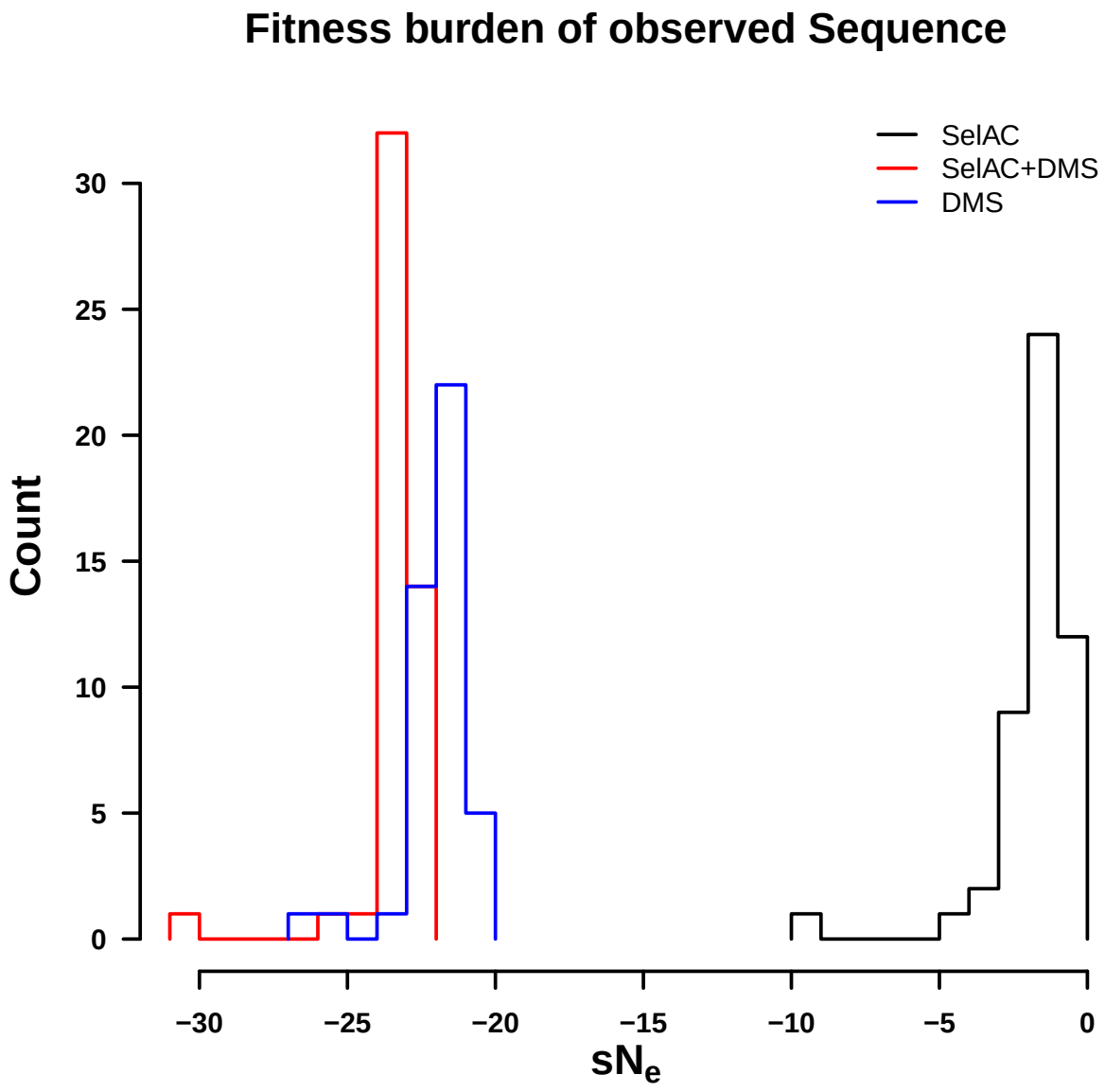

Figure 1: TEM, sNe of whole sequence, variation across tips.

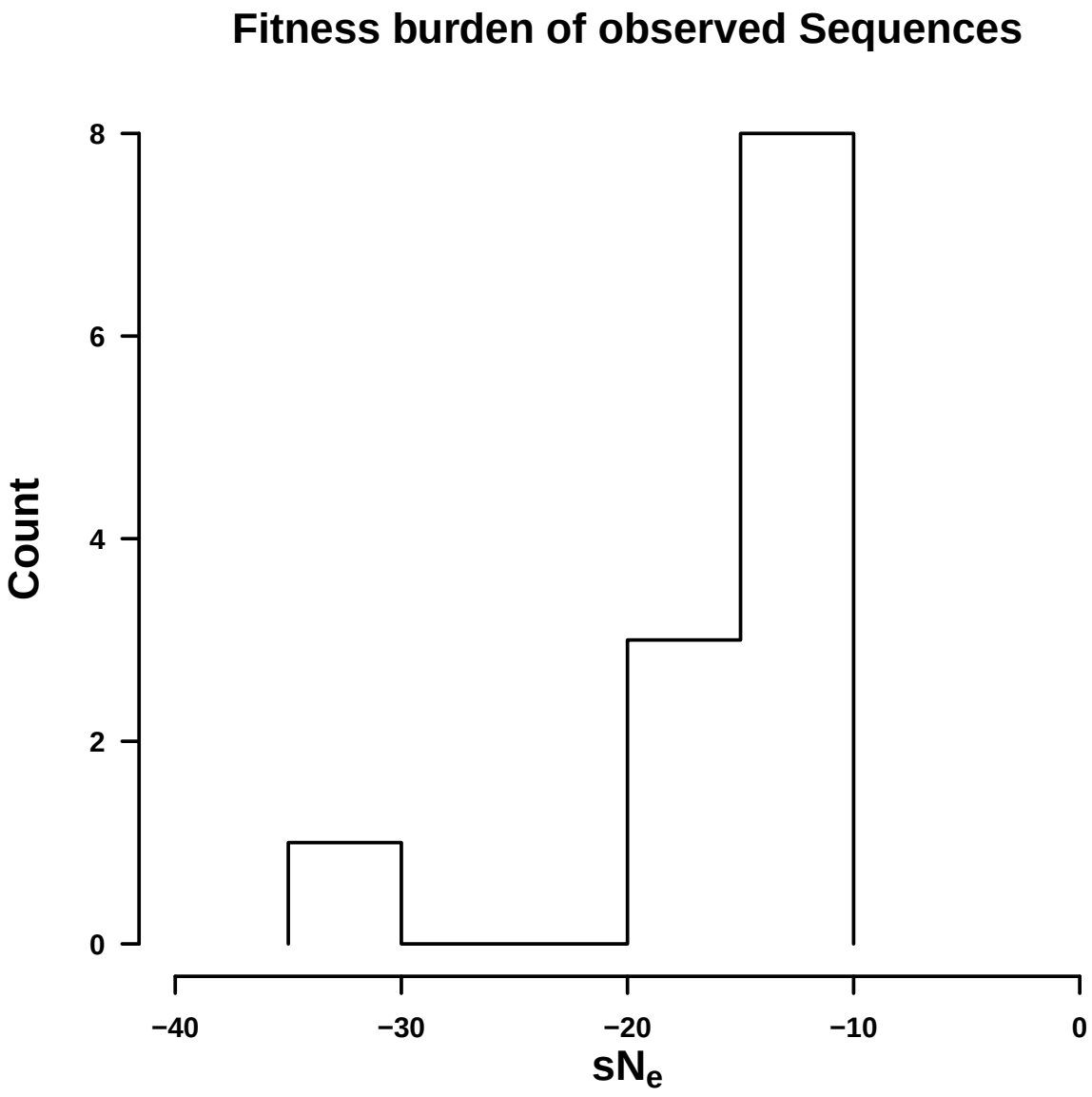

Figure 2: CytB

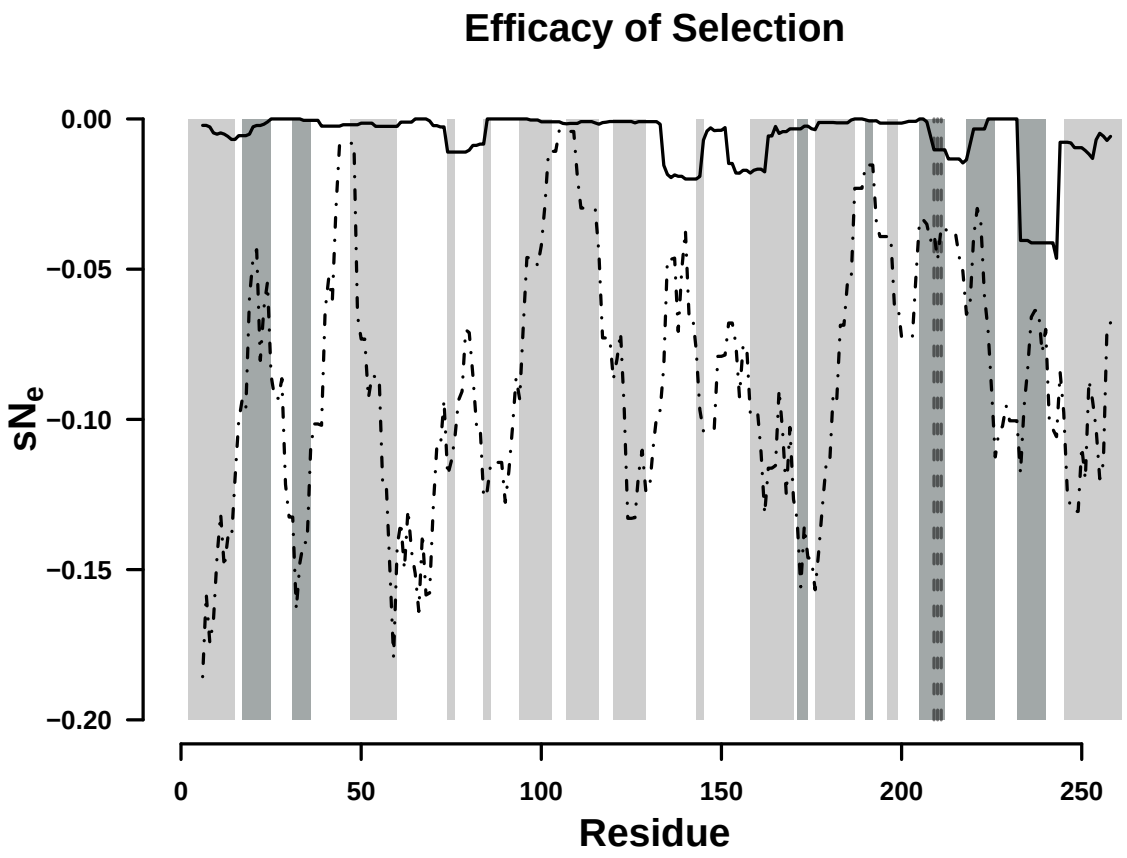

Figure 3: TEM, bars are different seconday structure elements, dashed line is DMS sNe, solid is SelAC., horizontal lines are active/binding sites.

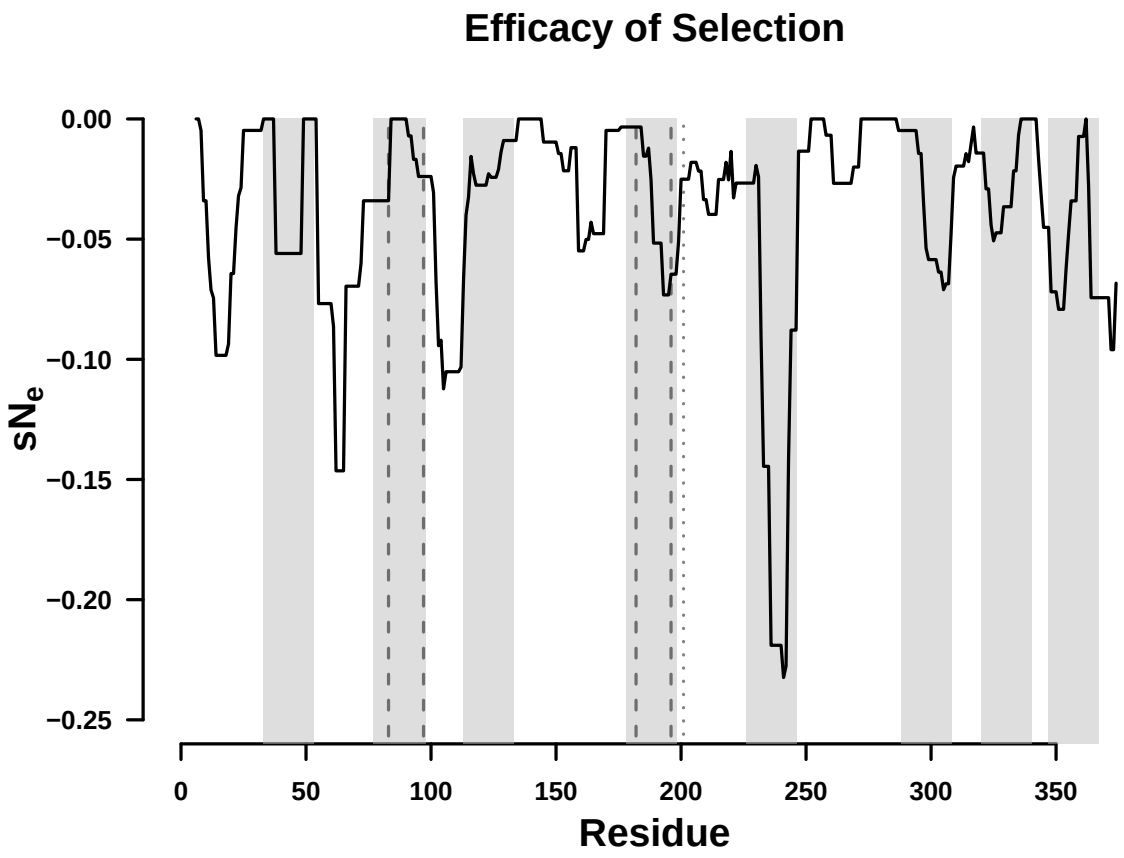

Figure 4: CytB