Estimating the genetic load of natural sequences in 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: August 26, 2018

Introduction

- Genes evolve with natural selection favoring proteins that encode their function optimally
  - To the extend at which the efficacy of selection becomes to weak and mutation and genetic drift pushes genes away from this optimum.
  - Therefore, in the absence of any compromises between different selection pressures, mutation and genetic drift introduce a genetic load, reducing a proteins fitness.
- Genetic load is usually assessed relative to a predefined wilt-type.
  - One could assess genetic load also relative to the genotype encoding a desired function 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.
- 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 not limited to fast growing organisms that can be manipulated under laboratory conditions and thus applicable along the whole tree of life.

- We predict the site specific optimal amio acid from sequence alignments of TEM, a β-lactamasein E. coli and cytochrome b (CytB), a mitochrondrial 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 estimates of genetic load for TEM to estimates obtained from DMS experiments.

- 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 using SelAC.
  - We find that the predicted amino acid sequence has high agreement with the observed consensus sequence of the alignment (TEM: 99%, CytB: 95%).

- Compare DMS from Firnberg and Stiffler to SelAC and majority under SelAC and phydms

  - Comparison of Frinberg under SelAC for TEM and SHV (three sequences: DMS, Majority, SelAC)
  - Comparison of Frinberg under phydms for TEM and SHV (three sequences: DMS, Majority, SelAC)
  - Comparison of Stiffler under SelAC for TEM and SHV (three sequences: DMS, Majority, SelAC)
  - Comparison of Stiffler under phydms for TEM and SHV (three sequences: DMS, Majority, SelAC)
- Comparison of perfered sequence
  - Simulations of sequences under each prefered sequence.
  - Only majority rule (duh) and SelAC agree with observed sequences.
- SelAC is dependent on choice of PC propertie to produce amino acid rankorder and assumes stabilizing selection.
  - Rankorder of certain sites can not be produced by any of the PC checked (no combination checked)

Discussion

- SelAC sequence outperformes DMS experiments, reflecting evolution better than DMS sequences under artificial selection pressure.
- SelAC only uses prefered state as input, no information about 2nd or third prefered amino acid.
- The reduction of a DMS experiment to this state might be considered an unfair comparison, however, we tested the sequences under phydms (no reduction of information), with the same result.

- This also means that SelAC produces the same information a DMS experiment would, but for naturally evolving sequences and can be applied to any sequence.
- TEM/SHV have not evolved to combate specific human developed antibiotics, but as means of "warfare" between bacteria (need more reading here).
- This could be the cause for the great difference between DMS and observed sequences.
- SelAC, however can not provide any information about antibiotic resistency, making DMS very valuable, but not for phylogenetics.
- but additional tip information could be combined with SelAC to get at this information (out of scope? future directions?).