Predicting amino acid functionality from sequence data 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: June 18, 2018

Introduction

- The introduction of selection into phylogenetic frameworks has been a long going effort.
  - Many models have been developed (Yang and Nielsen, Halpern and Bruno, ...)
  - Insert brief review of methods.
    * Models provided great theoretical inside.
    * Still often assume uniform stationary distribution of amino acids across sites.
- So far these models find limited application as these frameworks are very parameter rich.
  - The most popular models/tools, however, are still based purely on the mutation process (RaxML, RevBayes).
- A more recent take on the incorporation of selection on a protein is the independent estimation of fitness effects.
  - Deep Mutation Scanning (DMS) experiments provide site specific fitness values on synonymous and non-synonymous mutations (focus on non-synonymous).
    * This limits the number of estimated parameters greatly and allows for computationally feasible models.
  - However, the information on selection gained by DMS experiments is limited to single proteins of organisms that can be manipulated in the laboratory with short generation times.
    * This greatly limits its application in phylogenetics
- SelAC is a mechanistic model that utilizes the idea of site specific selection, and estimates it from sequence data.
  - SelAC has multiple advantages over other methods incorporating selection.

    - It does not assume a uniform stationary amino acid distribution across sites.
    - Does not depend on experimental data, and can therefore be applied to all codon sequences.
    - Clearly states model assumption and provides interpretable parameter estimates beyond branch length and nucleotide transition rates.
  - Due to SelACs hierarchical model structure it can also be parameterized with relatively few parameters.

- In this study, we compare the quality of phylogenetic estimates obtained utilizing DMS experiments to estimates from SelAC.
  - We utilize DMS experiments from Firnberg (2014) and Stifler (2016) for the TEM β-lactamase of *E. coli*.
  - We use phydms, a tool explicitly designed to utilize selection information from DMS experiments for an independent assessment of the SelAC estimated stationary amino acid distribution.
  - We compare model fit and adequacy of the DMS and SelAC amino acid preferences using SelAC and phydms.
    - We show that DMS experiments can have trouble accurately reflecting natural evolution of protein sequences.
    - We find that amino acid preferences estimated with SelAC provides better model adequacy than DMS experiments.
    - We show that information about amino acid preference can be extracted from sequence data using SelAC.

Results

- 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?).