Experimentally informed phylogenetic models are biased towards laboratory conditions and can be improved upon by mechanistic models of stabilizing selection.

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: December 13, 2018

Introduction

- Numerous attempts to incorporate selection into phylogenetic models have been made.
  - Phylogenetic inference of sequence relationship has long been focused on substitution rates and fixation probabilities.
  - However, the importance of site specific equilibrium frequencies has long been noted.
  - Models of site specific equilibrium frequencies tend to be unfeasible as they are very parameter rich.
  - Incorporating selection from experimental sources, therefore, seems like an attractive option.
    - Site specific selection on amino acid allows to incorporate heterogeneity of selection along the protein sequence into phylogenetic models.
- Independent fitness estimates have potential to greatly reduce number of parameters estimated from phylogenetic data.
  - DMS generates estimates of site specific selection on amino acids for large amount of mutations in a single experiment.
  - This allows for the fitting of complex site specific models to smaller data sets.
    - Site specific selection on amino acids improves model fits.
  - Empirical selection estimates are not always available, and their application for phylogenetic inference is questionable.
    - DMS experiments are limited to proteins and organisms that can be manipulated under laboratory conditions, greatly limiting their application in phylogenetics.

        - Estimates depend on factors like initial library of mutants, leading to heterogeneous competing populations.
        - The applied selection between the wild and the laboratory is likely to differ.
        - Hilton et al. (2017) showed that the variation between DMS experiments can have a significant effect on their utility.
    - To understand the reliability of selection on amino acids infered by DMS to inform phylogenetic studies, we utilize a DMS experiment by Stiffler et al. (2016).
        - TEM is found in gram-negative bacteria like *E. coli*.
        - The applied selection pressure was limited to ampicillin and focused on the sequence variant TEM-1.
        - TEM, however, can confer resistance to a wide range of antibiotics, causing it to be of wide interest.

- Mechanistic model of site specific stabilizing selection on amino acids rooted in population genetics are an improvement over DMS.
    - We used *phydms* and *SelAC* to assess the realibility of DMS estimates of selection to inform phylogenetic models.
        - model selection prefered *SelAC* over DMS estimates of selection.
    - Assesment of model adequacy via simulations highlights the inadequacy of experimentally inferred selection.
        - Sequences similarity of optimal DMS sequence to observed consensus sequence lower than expected.
        - Genetic load of observed sequences higher than expected.
    - Models fits informed by experimentally inferred selection improve model fit over conventional codon and nucleotide models but can be improved upon by *SelAC*.

        - We also compare model fit of the two codon models of site specific stabilizing selection to models fits of 227 other codon and nucleotide models using IQTree.

Results

Site Specific Stabilizing Selection on Amino Acids Improves Model Fit

- We evaluate model fits of models of site specific selection and 227 other codon and nucleotide models to 49 observed TEM sequences.
    - All models of site specific selection improve model fit.
    - Number of parameters estimated from phylogenetic data differs between *SelAC*, and *SelAC*+DMS and *phydms*, resulting in slightly worse AICc for *SelAC*.
    - However, *SelAC* outperforms *phydms* (Table 1).

Table 1: Model selection, shown are the three models of stabilizing site specific amino acid selection (*SelAC*, *SelAC*+DMS, *phydms*) and the best performing codon and nucleotide model (**??**). Reported are the log-likelihood $\log(L)$, the number of parameters estimated n, AIC and ΔAIC values. See Table X for results from all models we tested.

Model	log(L)	n	AIC	ΔAIC
SelAC	-1498	374	3744	0
SelAC+DMS	-1768	111	3758	14
phydms	-2061	102	4326	582
SYM+R2	-2230	102	4663	919
GY94 +F1X4+R2	-2243	102	4690	946

- We observe differences in topology between the model fit of *phydms* and *SelAC*.
    - *SelAC* is too slow for a topology search, however, due to the improved model fit of *SelAC*+DMS over *phydms*, it is likely that the *phydms* infered topology is inadequate due to the biased laboratory conditions.

    - *GY94* is outperformed by several nucleotide model e.g. *SYM*+R2, potentially indicating that frequency dependent selection is inappropriate for TEM.
    - *SelAC* model fit shows 84% of all evolution happening at the tips, while this reduces to 77% in the *phydms* model fit.

Assessing Adequacy of Laboratory and *SelAC* Inferences of Site Specific Selection

- Assessing model adequacy as sequence similarity sequence of selectively favored amino acids and observed consensus sequence.
    - Experimentally inferred selection is inconsistent with observed sequences.
    - Experimentally inferred sequence of selectively favored amino acids has only 52% sequence similarity with the observed consensus sequence.
    - *SelAC* inferred sequence of selectively favored amino acids has 99% sequence similarity with the observed consensus sequence.
        - It is tempting to assume that the consensus sequence will allways fair best, however, this would implicitly assume indepence between sequences.
        - However, the high sequence similarity of the consensus sequence and the sequence of selectively favored amino acids is likely due to the high average sequence similarity between the 49 observed sequences of 98%.
        - Furthermore, in addition to providing an estimate of the selectively favored amino acid, *SelAC* also allows to estimate the genetic load of observed amino acids.

Comparing Laboratory and *SelAC* Inferences of Genetic Load

- Laboratory estimates predict large genetic load

  - Simulations under DMS and *SelAC* inferred selection were used to establish a baseline expectation.
  - Assuming the site specific selection estimated by DMS, the observed TEM sequences represent an average sequence specific genetic load of 17.12 or, equivalently, an average site specific load of 0.065.
  - This load is significantly larger than the simulated sequences with an average sequence specific load of 6.68 or, equivalently, an average site specific genetic load of 0.025
  - In contrast, assuming the site specific selection estimated by *SelAC*, the observed TEM sequences represent an average sequence specific genetic load of 6.4×10^{-5} or, equivalently, an average site specific load of 2.4×10^{-7}.
  - Again, the simulated sequences show a decreased genetic load with an average sequence specific load of 1.3×10^{-5} or, equivalently, an average site specific genetic load of 4.8×10^{-8}.

Comparing Laboratory and *SelAC* Inferences of Site Specific Selection

- Distribution of genetic load differs between DMS inferred site specific selection and *SelAC* inferred site specific selection.
  - Assuming the site specific selection estimated by DMS, 111 sites have a genetic load of 0.
  - Assuming the site specific selection estimated by *SelAC*, 207 sites have a genetic load of 0.
    - In general, it is not surprising to find a large number of sites with 0 genetic load as many sites (X %) show no variation in the observed amino acid.

  - The selection estimates from DMS and *SelAC* agree for 107 sites at which no genetic load is found.
  - Thus, for 100 sites *SelAC* does estimate a genetic load of 0 but DMS does estimate non-zero genetic load, the inverse is true for four sites.
    - A closer look at the 100 sites for which *SelAC* does estimate a genetic load of 0 but DMS does estimate a non-zero load revealed that all 100 sites display a significant difference in likelihood between the *SelAC* and DMS estimated optimal amino aicd.
    - These 100 sites show a significantly ($p = 3 \times 10^{-13}$) higher mean genetic load under the DMS estimates than the remaining 163 sites of 0.0157 and 0.003, respectively, indicating that DMS represents the evolution of TEM particularly badly at these sites.
  - For the 52 sites where both, DMS and *SelAC*, estimate a non-zero genetic load we a correlation of $\rho = 0.247$, explaining 6% of the variation in the empirical selection estimates, when compared on the log scale.
    - In 26 cases *SelAC* and DMS estimate the same optimal amino acid.
    - The remaining cases all show a significant difference in likelihood between the *SelAC* and DMS infered optimal amino acids.
    - The 26 cases in which the infered optimal amino acid differs, we observe a significantly higher mean genetic load ($p = 2 \times 10^{-5}$) than in the remaining 26 sites of 0.0158 and 0.004, respectively, for which *SelAC* and DMS estimate the same optimal amino acid

Discussion

- We evaluate how well experimental selection estimates obtained by DMS explain natural sequence evolution and compare it to a novel phylogenetic framework, *SelAC*.

    - Previous work has shown that DMS selection estimates can improve model fit over classical approaches like GY94 and our work confirms this.
    - However, model selection shows that the *SelAC* model fit and the corresponding fitness estimates are favored over DMS estimates (Table 1).

- Adequacy of the DMS selection has previously not been assessed.
    - The amino acid sequence with the highest fitness estimated using DMS has only 49% sequence similarity with the observed consensus sequence.
    - In contrast, the SelAC estimate has 99% sequence similarity (Figure **??**).
    - In addition, we find evidence that experimental estimates of selection do not represent evolution in the wild.
        - Due to artificial selection environment; Heterogeneous population, very large s.
        - Only one antibiotic used, maybe a mixture of antibiotics would better reflect natural evolution.
        - Lack of repeatability between labs introduces further problems (Firnberg et al 2014 vs. Stifler et al. 2016).

- Assuming that the DMS selection inference adequately reflects natural evolution, the observed TEM sequences are either maladapted or were unable to reach a fitness peak.
    - 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, thus suggesting that the TEM sequences are mal-adapted according to the DMS estimates.
        - With a mutation rate of $2.54 \times 10^{-10} \times 789 = 2 \times 10^{-7}$ mutations per generation for TEM (Lee et al. 2012), we expect between $\mu N_e = 10^1$ and 10^2 new

mutations per generation of which on average XXX % are advantages per site.

    - Our simulations of sequence evolution with various N_e values and the DMS fitness values show that we would expect higher adaptation even with much smaller N_e (Figure **??**).

  - In addition, with an average site specific selection 0.085, we would expect that mutations fix on average between $(4/|s|) \times \ln(2N_e) \approx 1200$ and 1300 generations assuming N_e to be on the order of 10^8 to 10^9 (Crow and Kimura 1970).

  - As *E. coli* doubles every 15 hours in the wild (Gibson et al. 2018), we would therefore expect that a mutation with an average $s = 0.085$ sweeps through the population of size 10^9 in ~ 1.5 years.

    - This sweep would only accelerate with reduced N_e due to e.g. isolation between populations.

- The evidence derived from population genetics theory has us expecting the observed sequences to be at the selection-mutation-drift equilibrium, which is not the case if we assume the DMS inference of selection.

  - Estimates of selection obtained from *SelAC*, in contrast, show the observed sequences to be have high fitness.

    - The average site specific genetic load estimated by *SelAC* is four orders of magnitude lower than the average site specific load esimated using DMS (2.4×10^{-7} vs. 0.065).

  - We find the majority of sequences near the optimum, indicating that the *SelAC* estimates are consistent with theoretical population genetics results.

  - Taken together, it appears that DMS reflects the selection on the TEM sequence with respect to only one antibiotic, which seems appropriate to model selection in a hospital environments but not when the interest lies in the evolution of TEM in the wild.

- In addition to the result that *SelAC* better explains the evolution of observed sequences in the wild, *SelAC* has the advantage that it can be applied to any protein coding sequence alignment, however, is not without flaws itself.
  - Like DMS and most phylogenetic models, *SelAC* assumes site independence.
  - *SelAC* is a model of stabilizing selection, in contrast to e.g. GY94 which is a model of frequency dependent selection.
    - Since TEM plays a role in the chemical warfare with conspecifics and other microbes, some sites may be under frequency dependent selection.
  - In addition *SelAC* assumes that selection follows the same distribution for all sites.
    - However, the distribution of selection could differ for sites in the different secondary structure types.
    - Similarly, active sites may not follow the assumed distribution.
  - *SelAC* also assumes that selection is proportional to the distance of amino acids in physicochemical space.
    - In this study, we defaulted to the properties described by Grantham (1974) polarity, composition, and molecular volume, however, many other distances are available which may improve model fit.
- Low sequence variation in the TEM may be cause for concern as it could be misinterpreted by the model as stabilizing selection because of the short branches.
  - However, population genetics indicate the newly introduced mutations would sweep rapitly through the population if they provide a strong fitness advantage.

- In conclusion, 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 using a carefully constructed model of stabilizing selection rooted in first principles.
  - Further, we highlight the bias of laboratory inferences of selection and suggest to focus effors in improving phylogenetic inference on the development of more realisitc models.