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 10, 2018

Introduction

- Numerous attempts to incorporate selection into phylogenetic models have been made.
  - Phylogenetic inference of sequence relationship was long only 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.
  - Independent fitness estimates have potential to greatly reduce number of parameters estimated from phylogenetic data.
- Incorporating selection from experimental sources therefore seems like an attractive option.
  - Site specific amino acid preferences acknowledge the heterogeneity of selection along the protein sequence.
  - It allows for the fitting complex site specific models to smaller data sets.
  - DMS allows to estimate empirical selection on amino acids for large amount of mutations on a single experiment.
- However, selection between the wild and the laboratory likely differ.
  - For example, selection estimated with DMS depends on many factors like initial library of mutants and applied selection, which leads to heterogeneous competing population when using extensive mutation libraries.
- In addition, DMS experiments are also limited to proteins and organisms that can be manipulated under laboratory conditions, greatly limiting application of experimentally informed phylogenetic models.

  - Even when empirical selection estimates are available, their application for phylogenetic inference is questionable.

- To understand the applicability of selection on amino acids infered by DMS to hylogenetic studies, we utilize the class A β-lactamase TEM.

  - TEM is found in gram-negative bacteria like *E. coli* and of wide interest due to its role in the decradation of lactam based antibiotocs.
  - We utilize a DMS experiment by (**?**) where 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.

- We compare experimentally inferred site specific selection to inform phylogenetic models using *phydms* to *SelAC*, a mechanistic model of site specific stabilizing selection on amino acids rooted in population genetics.

  - 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.
  - Models fits informed by experimentally inferred selection improve model fit over conventional codon and nucleotide models but can be improved upon by *SelAC*.

- Simulations highlight the inadequacy of experimentally inferred selection.

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, ΔAIC, AICc, and ΔAICc values. See Table X for results from all models we tested.

Model	log(L)	n	AIC	ΔAIC	AICc	ΔAICc
SelAC+DMS	-1768	111	3758	14	3760	0
SelAC	-1498	374	3744	0	3766	6
phydms	-2061	102	4326	582	4328	568
SYM+R2	-2230	102	4663	919	4694	934
GY94 +F1X4+R2	-2243	102	4690	946	4821	1061

- 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, 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 17.88 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.
    - For the 52 sites where both, DMS and *SelAC*, estimate a non-zero genetic load we find a correlation of $\rho = 0.247$, explaining 6% of the variation in the empirical selection estimates.
- A closer look at the 100 *SelAC* does estimate a genetic load of 0 but DMS does estimate a non-zero load revealed that XXXX
    - How many have significant differences in likelihood between the two different optima AA?

  - what is the difference in genetic load at these sites?
  - Where are these sites?

Discussion

- We evaluated how well experimental selection estimates from DMS experiments explain natural sequence evolution and compared 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.
  - Model selection favored the SelAC model fit and the corresponding fitness estimates over the DMS estimates using both, SelAC and phyDMS (Table 1).
- Adequacy of the DMS selection has previously not been assessed.
  - The amino acid with the cumulative highest fitness experimentally estimated with DMS only has 49% concordance with the observed alignment.
  - In contrast, the SelAC estimate has 99% concordance (Figure **??**).
  - Estimates of selection coefficients do not represent evolution.
    - 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 mal-adapted or where unable to reach a fitness peak.

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

- Estimates of selection coefficients do not represent evolution.
  - 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).
  - Still better than models without site specific equilibrium frequencies.

- DMS estimates of the observed TEM variants predict them to be mal-adapted while SelAC predicts most TEM variants to be well adapted.
  - Given *E. coli*'s large effective population size, the efficacy of selection should be very large.
  - We therefore expect the observed sequence variants to be at the selection-mutation-drift barrier, which in turn can expected to be near the optimum.
  - We find the majority of sequences near the optimum, therefore the SelAC estimates are consistent with theoretical population genetics results.
  - In contrast, finding strong selection against the observed TEM variants indicates that DMS is not consistent with theoretical population genetics expectations.

  - This is consistent when thinking about that DMS only reflects the selection on the TEM sequence with regards to one antibiotic, which seems appropriate to model selection in modern hospital environments but not when the interest lies in the natural evolution of TEM.

- We find that SelAC produces similar selection against the observed TEM variants if we assume the fitness peaks (optimal AA) that are estimated by DMS.
  - This shows that DMS and SelAC can provide consistent estimates of selection against amino acids.
  - SelAC has the advantage that it can be applied to any protein coding sequence alignment.
  - This removes the need for extrapolation e.g. from TEM to SHV.
- SelAC has the advantage that it can be applied to any protein coding sequence alignment.
  - This removes the need for extrapolation e.g. from TEM to SHV.
- Difference in selection parameters between TEM and SHV indicate that extrapolation is not a good idea.
  - The difference in the site specific strength of selection shows that TEM and SHV are facing different selection pressures.
  - this is also highlighted by the differences in physicochemical weightings between the two proteins.
- SelAC outperforms DMS, but 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 negative frequency dependent selection.
  - SelAC assumes the same G distribution across all sites.
    - Different G distribution for each type of secondary structure
    - active sites may not follow distribution.
  - SelAC assumes that selection is proportional to distance in physicochemical space.
    - We used Grantham (1974) properties, however many other distances are available which may an even better 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, provided our simulations support that TEM is actually under stabilizing selection

- 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.
  - This highlights the limitations of DMS to explain natural evolution.