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 17, 2018

Introduction

Phylogenetic inference is of ever increasing importance across biology. Most commonly used models used for phylogenetic inference are incorporated into powerful software packages (Stamatakis, 2014; Höhna *et al.*, 2016; Nguyen *et al.*, 2015) and therefore fast and easy to use. However, these models come with significant shortcomings as they lack biological realism.

Phylogenetic models focused on the nucleotide composition such as GTR, or UNREST (**??**) are limited to mutation effects and are agnostic to any higher level selection on codons or amino acids. Amino acid models attempt to capture selection, however, these do not properly account for mutations as they occur on the nucleotide level. In an attempt to remedy the issues of nucleotide and amino acid models, codon models combine mutation between nucleotides and selection on amino acids. However, all types of models have in common that they describe the same equilibrium frequencies at each site.

- Phylogenetics plays an ever increasingly important role in biology.
  - Co-Expression
  - species relationship across all fields of biology
  - protein evolution
  - cancer
- Most commonly used methods
  - Strengths
    - Fast
    - Easy to use (software packages)
  - Weaknesses
    - Many ignore key forces in evolution.

    * Nucleotide models account for mutation but not selection
        · Mutation only: UNREST, GTR, JC.
        · Mutation rates can vary between nucleotide positions.
        · Use the same matrix for all site
    * Amino Acid models try to capture selection but mutation happens on nucleotide level.
        · Selection strictly phenomenological: PAM, BLOSSUM, and WAG?
        · Use same matrix for all sites
        · Can also be applied with categorization approach introduced by Lartillot and colleagues.
    * Codon models to remedy problems of nucleotide and amino acid models
        · Most popular one that includes selection (GY94 and derivatives) which is commonly misinterpreted and restricted selection scenario: freq dependence.
        · codon models allow to capture the mutation process on the nucleotide level and the selection on amino acids.
    * Mutation, AA, and codon models all end up with same AA equilibrium frequency for all sites.
    * Biologists have long recognized that equilibrium frequencies, and thus the substitution matrix responsible, can vary substantially between sites.
  – Halpern and Bruno (1998) provide general model.
    * Can have distinct substitution matrix for each site.
    * As a result requires $19 \times n$ parameters.
    * Large number of parameters makes implementation unfeasible

- Potential solutions to parameterization issue

- Use additional information: experiments via DMS
    * 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.
- Use better models
    * Lartillot and colleagues mitigate this issue using a site categorization approach. (Mention in discussion as potential next step to avoid reviewers asking you to do this.)
    * *SelAC* continues the site categorization approach introduced by Lartillot and colleagues by using a simplistic model of amino acid distances in physicochemical space.
        · *SelAC* is rooted in population genetics
        · *SelAC* uses distance in physicochemical space between amino acids to describe decline in fitness.

- Ideally, we would use better models and additional data.

- We assess the reliability of selection on amino acids inferred by DMS to inform phylogenetic studies.
  - we utilize a DMS experiment by Stiffler et al. (2016) for TEM.
  - 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.

- Main Findings: Results consistent with previous work, but clearly demonstrates that better models are more informative than un-natural supplementary data
  - Model selection preferred *SelAC* over *phydms*.
  - Evidence that DMS data does not describe conditions in the wild
    * Poor model adequacy (c.f. *SelAC*)
    * Optimal aa under DMS not consistent with genetic variation in TEM observed in wild (c.f. *SelAC*).
    * Genetic loads implied by DMS very large (c.f. *SelAC*).

- Conclusion:
  - Better models more informative and applicable than un-natural supplementary data
  - *SelAC* provides additional, biologically meaningful information such as site specific optimal amino acid and fitness landscape.

Results

Model selection prefers *SelAC* over all other models, including PhyDMS

- Models of site specific selection dramatically improve model fit.
  - Compare *SelAC* and *phydms* to 131 nucleotide and 97 codon models and variations.
  - *SelAC* shows best model fit.
  - *phydms* parameterized by XXX's data was second best. However, problems (discussed below)
  - Best codon model without site specific selection is *GY94*.
  - *GY94* is outperformed by multiple nucleotide models like *SYM*+R2.
  - Caveats
    * Treated discrete aa as parameters (conservative, discuss more later).
    * Topology between the model fit of *phydms* and *SelAC* differs.
      · *SelAC* is too slow for a topology search, therefore we used a topology inferred with the model by Kosiol et al (2007).
      · *phydms* started at Kosiol topology, suggesting that we are being conservative.
      · *SelAC* with *phydms* topology ...
  - Additional observations
    * Statement about evolution inferred from our results with *SelAC* vs *phydms* vs other models (nt, codon, aa).
    * Another statement?

Additional Shortcomings of DMS Data

Below implies that DMS environment is fundamentally different from wild.

DMS Leads to Poor Model Adequacy for TEM

- We define model adequacy as similarity of selectively favored amino acids and observed consensus sequence.
- High adequacy of *SelAC*'s inferences
    - *SelAC* inferred sequence of selectively favored amino acids has 99% sequence similarity with the observed consensus sequence. Perhaps not surprising given this was the only data *SelAC* used.
- Low adequacy of DMS inferences.
    - Experimentally inferred sequence of selectively favored amino acids has only 52% sequence similarity with the observed consensus sequence.
    - Suggests that DMS selection are not informative about selection in wild. Additional support for claim
        - The experimentally inferred optimal amino acid is not observed in nature at X % of sites.
        - Physicochemical properties appear to differ between observed and estimated optimal amino acids

DMS Predictions Inconsistent with Observed Genetic Variation in TEM

Qualitative comparison

- 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 inferred optimal amino acids.
    - The 26 cases in which the inferred 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

Table 1: Genetic load at heterogeneous and Homogeneous sites in the TEM alignement according to DMS and *SelAC*

		Genetic Load	
Sites	# Residues	*SelAC*	DMS
Heterogeneous	66	6.3×10^{-7}	0.01
Homogeneous	197	0	0.007

DMS Implies Unrealistic Genetic Loads

Quantitative comparison

- Estimates of genetic load differ greatly between the *SelAC* and experimentally estimated fitness landscape.
    - 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.
    - 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}.
- Simulations under DMS and *SelAC* inferred selection were used to establish point of reference and further assess model adequacy.
    - Simulations assuming the DMS inferred selection show that the genetic load of the observed sequences is significantly larger than the genetic load of the simulated sequences
        - We find an average sequence specific load of 6.68 or, equivalently, an average site specific genetic load of 0.025.

* Simulations assuming the *SelAC* inferred selection as well show that the genetic load of the observed sequences is significantly larger than the genetic load of the simulated sequences.
    - We find an average sequence specific load of 1.3×10^{-5} or, equivalently, an average site specific genetic load of 4.8×10^{-8}.

Move from Results

- Number of parameters estimated from phylogenetic data differs between *SelAC* and *phydms*. (Methods and Discussion)
- unclear how to deal with number of parameters we, therefore, toked a conservative approach. (Methods and Discussion)
- It is tempting to assume that the consensus sequence will always fair best, however, this would implicitly assume independence 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%.

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.

- 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.
  - 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 $2.54\times10^{-10}\times789 = 2\times10^{-7}$ 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.

- 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 rapidly 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 efforts in improving phylogenetic inference on the development of more realisitc models.

References

Höhna, S., Landis, M., Heath, T., *et al.* 2016. Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Systematic Biology*, 65(4): 726–736.

Nguyen, L., Schmidt, H., von Haeseler, A., and Minh, B. 2015. Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1): 268–274.

Stamatakis, A. 2014. Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.

Tavare, S. 1986. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences*, 17: 57–86.

Yang, Z. 1994. Maximum-likelihood phylogenetic estimation from DNA-sequences with variable rates over sites - approximate methods. *Journal Of Molecular Evolution*, 39: 306–314.