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 25, 2018

1 Introduction

- 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
  - Advantages
    * 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.
  - Shortcomings
    * Empirical selection estimates are not always available.
      - DMS experiments are limited to proteins and organisms that can be manipulated under laboratory conditions, greatly limiting their application in phylogenetics.

    * Application for phylogenetic inference is questionable.
        · 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 have a reproducibility problem and the resulting 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* also uses site categorization approach similar to Lartillot and colleagues by using a simplistic nested model of amino acid distances in physicochemical space.
        · *SelAC* is rooted in population genetics, like Lartillot work.
        · *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: A) Results consistent with previous work, but unnatural supplementary data causes poor model adequacy, clearly demonstrating that better models are more informative.
  - 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*).
  - *SelAC* has higher model adequacy and provides more realistic estimates of genetic load.

- 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.
    * *SelAC* does not rely on supplementary data.
    * *SelAC* can be expanded to test other hypothesis.

2 Results

2.1 *SelAC* Outperforms Experimentally Informed Models

for

- 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 data from Stiffler et al. (2016) 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 AA as discrete 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, but estimated a different one, 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?

2.2 Shortcomings of DMS Data

Below implies that DMS environment of lab 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.
- 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
- 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.

2.2.1 DMS Predictions Inconsistent with Observed Genetic Variation in TEM

- 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, at 107 of those sites DMS and *SelAC* agree in their estimated optimal amino acid.
  - 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.
  - 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 , 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 find 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

2.2.2 DMS Implies Unrealistic Genetic Loads

Quantitative comparison

- Estimates of genetic load differ greatly between the *SelAC* and experimentally esti-

mated fitness landscape.

  - The site specific selection estimated by DMS for the observed TEM sequences represent an average site specific load of 0.065 which is an average sequence specific genetic load of 17.12.
  - In contrast, the site specific selection estimated by *SelAC* for the observed TEM sequences represent an average site specific load of 2.4×10^{-7} which is an average sequence specific genetic load of 6.4×10^{-5}..

- 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}.

3 Discussion

- We compared the performance of two site specific codon level phylogenetic models, *phydms* and *SelAC*, and XXX more commonly used codon and nucleotide models in explaining TEM data.

- Brief statement about *E. coli* TEM data.
- Using AIC as a measure of model fit, we found...
  - While both site specific models, *phydms* and *SelAC*, perform substantially better than the alternative models, including the classic *GY94* model.
  - Further, *SelAC* clearly outperforms *phydms* ($\Delta AIC = XXXX$).
  - The improved performance of *phydms* and *SelAC* presumably results from their ability to more realistically describe the effects of natural selection on sequence evolution.
  - However, this realism comes at a cost.
    * *phydms* requires fitness estimates for each amino acid at every site, which necessitates experimental work.
    * *SelAC* uses a nested modeling approach, which avoids the necessity of amino acid specific fitness estimates, but greatly increases the computational cost of model fitting.
- In order to better understand the difference in model performance of *phydms* and *SelAC*, we evaluted their model adequacy.
- Define model adequacy.
- *phydms*'s model adequacy is a direct function of DMS measurements. As a result, we will focus on these measurements directly.
- Highlight how this property is often ignored relative to model fit.
- Model adequacy results strongly favors *SelAC*
  - Sequence similarity

    - 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.
  - Genetic Load
    - 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).

- Why we should believe our main claims [CUT EXTENSIVELY]
  - 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 maladapted 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.

- Estimates of selection obtained from *SelAC*, in contrast, show the observed sequences to be have high fitness.
  - 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.
  - 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.
  - This sweep would only accelerate with reduced N_e in isolated populations.

- Besides poor performance in terms of model fit and adequacy, there are other shortcomings to using DMS data for phylogenetic inference.
  - Lack of repeatability between labs introduces further problems (Firnberg et al 2014 vs. Stifler et al. 2016).
  - Cost
  - Only applicable to fast growing microorganisms.
  - Laboratory environment may not represent evolution in the wild.
    - Due to artificial selection environment; Heterogeneous population, very large s.

    * Unlikely such a strong and singluar selective force commonly encountered in wild. While this shortcoming may be the easiest to overcome by altering the laboratory conditions to include multiple and weaker selective forces ...

- Advantages of *SelAC*. 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 aligned protein coding sequences from organisms of any size and any growth rate.

- *SelAC* in current form has numerous shortcomings. However, mechanistic nature provides avenue for overcoming these via model expansion.

  - Easy

    * HMM extension would allow for

      · frequency dependent selection (c.f. GY94) Particularly appropriate here because TEM plays a role in the chemical warfare with conspecifics and other microbes. As a result, some sites may be under negative frequency dependent selection. However, *GY94* fit and high number of invariant sites suggests this is likely only a small fraction of sites.

      · Shifts in optimal aa along branches.

  - Addition of selection between amino acid synonyms.

  - Use of more realistic functions to map amino acid sequence to protein function

    * Higher order dependence on physicochemical distance.

    * Use explicit molecular models

- Hard

  - Allowing site sensitivity term to < 0 and, thus, model diversifying selection.

  - Extend to mixture model where model parameters vary between categories of sites.

- Incorporating epistasis. *SelAC*, like all of the other models considered here, assumes site independence and, thus, ignores epistatic interactions.

- 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.
  - Better models avoid many of these problems.
  - Ability to expand *SelAC* as outline above make it a natural framework for hypothesis testing. Go forth and build better models!

Cut 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 observed sequences.

- 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%.

Cut from Discussion

- 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, 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.
- However, population genetics indicate the newly introduced mutations would sweep rapidly through the population if they provide a strong fitness advantage.

4 Materials and Methods

4.1 Phylogenetic Inference and Model selection

TEM and SHV sequences were obtained from Bloom (2017) already aligned. We separated the TEM and SHV sequences into individual alignments. Experimentally fitness values for TEM were taken from Stiffler *et al.* (2016). We followed (Bloom, 2017) to convert the experimental fitness values into site specific equilibrium frequencies for *phydms*. *phydms* (version 2.5.1) was fitted to the site specific selection from Stiffler *et al.* (2016) using python

(version 3.6). *SelAC* (version 1.6.1) was fitted to the TEM alignment using R (version 3.4.1) (R Core Team, 2013) with and without experimental site specific selection. We assumed the physicochemical properties estimated by Grantham (1974). We choose the constraint free general unrestricted model (Yang, 1994) as mutation model for *SelAC*. All other models were fitted using IQTree (Nguyen *et al.*, 2015). We report each model's $\log(L)$, AIC, and AICc. Models were selected based on the AICc values.

4.2 Sequence Simulation

Sequences were simulated by stochastic simulations using a Gillespie algorithm (Gillespie, 1976) that was model independent. To calculate fixation probabilities during the simulation we followed Sella and Hirsh (2005). The fitness values were estimated using *SelAC* or taken from Stiffler *et al.* (2016). We choose the fitness values resulting from the highest concentration (2500 $\mu g/mL$) treatment of ampicillin for our comparison. We rescaled the experimental fitness such that the amino acid with the highest fitness at each site has a value of one. Mutation rates for the simulations were taken from the *SelAC* or *SelAC*+DMS fit, respectively. The initial sequences were either a random sequence sampled with uniform codon probabilities or the ancestral sequence reconstructed using FastML (Ashkenazy *et al.*, 2012) (last accessed: 30.09.2018). Each sequence was simulated 10 times and we report average genetic load and sequence similarity and the standard error. The sequences were sampled at times 0.01, 0.1, 1, and 10 expected mutations per site.

4.3 Estimating site specific efficacy of selection G

SelAC does not by default estimate site specific values for G but assumes G values follow a Γ-distribution (Felsenstein, 2001). Site specific values for G were optimized by fixing all estimated parameters and performing a maximum likelihood search without the integration over G. In contrast to *SelAC* that assumes G to be purely positive, we allowed negative values for G but constraint the search to values between -300 and 300 to ensure numerical

stability.

4.4 Estimating site specific fitness values w_i

Following Beaulieu *et al.* (ress) w_i is proportional to

$$w_i \propto \exp(-A_0 \eta \psi) \tag{1}$$

where A_0 describes the decline in fitness with each high energy phosphate bond wasted per unit time, and ψ is the protein's production rate. η is the cost/benefit ratio of a protein (see (Beaulieu *et al.*, ress) for details). However, *SelAC* only estimates a composition parameter $\psi' = A_0 \psi N_e$ thus

$$\psi = \frac{\psi'}{A_0 N_e q} \tag{2}$$

SelAC assumes that the effective population size $N_e = 5 \times 10^6$ and that $A_0 = 4 \times 10{-}7$ (Gilchrist, 2007).

4.5 Model Adequacy

Model adequacy was assessed by comparing the observed sequences and simulations under the site specific selection inferred by the deep mutation scanning experiment or *SelAC*. First, similarity between the sequence of selectively favored amino acids and the observed TEM sequences was assessed. Sequence similarity was measured as the number of differences in the aligned amino acid sequences. Second, the genetic load of the observed and the simulated sequences was calculated using either the site specific selection inferred by the deep mutation scanning experiment or *SelAC*. The average genetic load for site i in the alignment was calculated as

$$L_i = \frac{w_{max,i} - \overline{w_i}}{w_{max,i}} \tag{3}$$

where $w_{max,i}$ is the fitness of the selectively favored amino acids at position i, either estimated using the site specific selection inferred by DMS or *SelAC*. We, however, rescaled all fitness estimates such that $w_{max,i} = 1$ $\overline{w_i}$ represents the average fitness of the residues observed at position i. The average sequence specific genetic load L was calculated as the sum of the site specific genetic loads $L = \frac{1}{n}\sum_{i=1}^{n} L_i$ where n is the number of amino acid sites.

5 Acknowledgments

This work was supported in part by NSF Award and DEB-1355033 (BCO, MAG, and RZ) with additional support from The University of Tennessee Knoxville. CL received support as a Graduate Student Fellow at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation through NSF Award DBI-1300426, with additional support from UTK. The authors would like to thank Russel Zaretzki, Jeremy Beaulieu and Alexander Cope for their helpful criticisms and suggestions for this work.

References

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., *et al.* 2012. Fastml: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, 40(Web Server Issue): W580–4.

Beaulieu, J., O'Meara, B., Zaretzki, R., *et al.* in press. Population genetics based phylogenetics under stabilizing selection for an optimal amino acid sequence: A nested modeling approach. *Molecular Biology and Evolution*, X: NA.

Bloom, J. 2017. Identification of positive selection in genes is greatly improved by using experimentally infromed site-specific models. *Biology Direct*, 12: 1.

394 Felsenstein, J. 2001. Taking variation of evolutionary rates between sites into account in
395 inferring phylogenies. *Journal of Molecular Evolution*, 53(4): 447–455.

396 Gilchrist, M. 2007. Combining models of protein translation and population genetics to pre-
397 dict protein production rates from codon usage patterns. *Molecular Biology and Evolution*,
398 24(11): 2362–2372.

399 Gillespie, D. 1976. A general method for numerically simulating the stochastic time evolution
400 of coupled chemical reactions. *Journal of Computational Physics*, 22(4): 403–434.

401 Grantham, R. 1974. Amino acid differences formula to help explain protein evolution. *Sci-*
402 *ence*, 185(4154): 862–864.

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

406 R Core Team 2013. *R: A Language and Environment for Statistical Computing*. R Founda-
407 tion for Statistical Computing, Vienna, Austria.

408 Sella, G. and Hirsh, A. 2005. The application of statistical physics to evolutionary biology.
409 *Proceedings of the National Academy of Sciences of the United States of America*, 102:
410 9541–9546.

411 Stiffler, M., Hekstra, D., and R, R. 2016. Evolvability as a function of purifying selection in
412 tem-1 β-lactamase. *Cell*, 160: 882–892.

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

6 Appendix: Supplementary Material

Table 1: Model selection of 230 models of nucleotide and codon evolution.

No.	Model	LnL	n	AIC	ΔAIC	AICc	ΔAICc
1	*SelAC*+DMS +G4	-1768	111	3758	14	3760	0
2	*SelAC* +G4	-1498	374	3744	0	3766	6
3	*phydms*	-2060.85	102	4326	582	4328	568
4	SYM+R2	-2229.616	102	4663.232	919.232	4693.862	933.862
5	TIMe+R2	-2232.406	100	4664.811	920.811	4694.172	934.172
6	TVMe+R2	-2232.838	101	4667.677	923.677	4697.668	937.668
7	TIM3e+R2	-2234.332	100	4668.664	924.664	4698.024	938.024
8	TIM2e+R2	-2234.381	100	4668.763	924.763	4698.123	938.123
9	K3P+R2	-2235.777	99	4669.553	925.553	4698.291	938.291
10	TNe+R2	-2236.078	99	4670.155	926.155	4698.892	938.892
11	SYM+R3	-2229.616	104	4667.232	923.232	4699.162	939.162
12	TIM+F+R2	-2230.958	103	4667.915	923.915	4699.191	939.191
13	TIMe+R3	-2232.404	102	4668.808	924.808	4699.437	939.437
14	GTR+F+R2	-2228.537	105	4667.073	923.073	4699.665	939.665
15	K3Pu+F+R2	-2232.617	102	4669.234	925.234	4699.864	939.864
16	TVM+F+R2	-2230.105	104	4668.21	924.21	4700.14	940.14
17	TVMe+R3	-2232.838	103	4671.676	927.676	4702.952	942.952
18	K2P+R2	-2239.424	98	4674.847	930.847	4702.969	942.969
19	TIM3e+R3	-2234.332	102	4672.664	928.664	4703.293	943.293
20	TIM2e+R3	-2234.381	102	4672.762	928.762	4703.391	943.391
21	TIM3+F+R2	-2233.064	103	4672.127	928.127	4703.403	943.403
22	TIM2+F+R2	-2233.114	103	4672.227	928.227	4703.503	943.503
23	K3P+R3	-2235.777	101	4673.553	929.553	4703.545	943.545
24	TN+F+R2	-2234.624	102	4673.249	929.249	4703.878	943.878
25	TPM3u+F+R2	-2234.673	102	4673.347	929.347	4703.977	943.977
26	TPM3+F+R2	-2234.674	102	4673.348	929.348	4703.978	943.978
27	TPM2u+F+R2	-2234.681	102	4673.363	929.363	4703.993	943.993
28	TPM2+F+R2	-2234.683	102	4673.365	929.365	4703.995	943.995
29	TNe+R3	-2236.077	101	4674.155	930.155	4704.146	944.146
30	TIM+F+R3	-2230.958	105	4671.915	927.915	4704.507	944.507
31	HKY+F+R2	-2236.266	101	4674.531	930.531	4704.522	944.522
32	GTR+F+R3	-2228.536	107	4671.073	927.073	4705.011	945.011
33	K3Pu+F+R3	-2232.617	104	4673.234	929.234	4705.163	945.163
34	TVM+F+R3	-2230.105	106	4672.21	928.21	4705.471	945.471
35	K2P+R3	-2239.192	100	4678.384	934.384	4707.745	947.745
36	TIM3+F+R3	-2233.063	105	4676.127	932.127	4708.718	948.718
37	TIM2+F+R3	-2233.113	105	4676.227	932.227	4708.818	948.818
38	TN+F+R3	-2234.624	104	4677.249	933.249	4709.178	949.178

Table 1 Continued

No.	Model	LnL	n	AIC	ΔAIC	AICc	ΔAICc
39	TPM3u+F+R3	-2234.673	104	4677.347	933.347	4709.277	949.277
40	TPM3+F+R3	-2234.674	104	4677.348	933.348	4709.277	949.277
41	TPM2u+F+R3	-2234.681	104	4677.363	933.363	4709.293	949.293
42	TPM2+F+R3	-2234.682	104	4677.364	933.364	4709.294	949.294
43	HKY+F+R3	-2236.074	103	4678.148	934.148	4709.424	949.424
44	SYM+I+G4	-2243.212	102	4690.424	946.424	4721.054	961.054
45	TVMe+I+G4	-2244.533	101	4691.066	947.066	4721.057	961.057
46	TIMe+I+G4	-2246.457	100	4692.914	948.914	4722.275	962.275
47	K3P+I+G4	-2248.166	99	4694.332	950.332	4723.069	963.069
48	TVM+F+I+G4	-2241.853	104	4691.707	947.707	4723.636	963.636
49	TIM3e+I+G4	-2247.379	100	4694.758	950.758	4724.119	964.119
50	K3Pu+F+I+G4	-2245.156	102	4694.311	950.311	4724.941	964.941
51	GTR+F+I+G4	-2241.484	105	4692.968	948.968	4725.559	965.559
52	TIM+F+I+G4	-2244.418	103	4694.836	950.836	4726.112	966.112
53	TPM3u+F+I+G4	-2246.03	102	4696.06	952.06	4726.69	966.69
54	TPM3+F+I+G4	-2246.069	102	4696.138	952.138	4726.768	966.768
55	TIM2e+I+G4	-2248.934	100	4697.868	953.868	4727.228	967.228
56	TNe+I+G4	-2250.587	99	4699.174	955.174	4727.911	967.911
57	TIM3+F+I+G4	-2245.534	103	4697.068	953.068	4728.344	968.344
58	K2P+I+G4	-2252.181	98	4700.362	956.362	4728.484	968.484
59	TPM2u+F+I+G4	-2247.579	102	4699.158	955.158	4729.788	969.788
60	TPM2+F+I+G4	-2247.685	102	4699.371	955.371	4730	970
61	HKY+F+I+G4	-2249.065	101	4700.13	956.13	4730.121	970.121
62	TIM2+F+I+G4	-2247.009	103	4700.018	956.018	4731.294	971.294
63	TN+F+I+G4	-2248.511	102	4701.023	957.023	4731.652	971.652
64	TVMe+I	-2254.804	100	4709.608	965.608	4738.968	978.968
65	K3P+I	-2257.72	98	4711.439	967.439	4739.561	979.561
66	SYM+I	-2254.11	101	4710.221	966.220	4740.212	980.212
67	TIMe+I	-2257.074	99	4712.149	968.149	4740.886	980.886
68	TVM+F+I	-2252.157	103	4710.315	966.315	4741.591	981.591
69	K3Pu+F+I	-2254.856	101	4711.712	967.712	4741.704	981.704
70	TIM3e+I	-2257.796	99	4713.592	969.592	4742.33	982.33
71	TPM3+F+I	-2255.771	101	4713.543	969.543	4743.534	983.534
72	TPM3u+F+I	-2255.771	101	4713.543	969.543	4743.534	983.534
73	K2P+I	-2261.218	97	4716.436	972.436	4743.949	983.949
74	GTR+F+I	-2252.067	104	4712.133	968.133	4744.063	984.063
75	TIM+F+I	-2254.783	102	4713.566	969.566	4744.195	984.195
76	TNe+I	-2260.579	98	4717.158	973.158	4745.28	985.28
77	TIM3+F+I	-2255.684	102	4715.368	971.368	4745.998	985.998
78	HKY+F+I	-2258.352	100	4716.703	972.703	4746.064	986.064
79	TIM2e+I	-2259.878	99	4717.757	973.757	4746.494	986.494
80	TVMe+G4	-2258.853	100	4717.705	973.705	4747.066	987.066

Table 1 Continued

No.	Model	LnL	n	AIC	ΔAIC	AICc	ΔAICc
81	SYM+G4	-2257.573	101	4717.146	973.146	4747.137	987.137
82	TPM2+F+I	-2257.712	101	4717.423	973.423	4747.415	987.415
83	TPM2u+F+I	-2257.712	101	4717.423	973.423	4747.415	987.415
84	K3P+G4	-2261.922	98	4719.844	975.844	4747.966	987.966
85	TIMe+G4	-2260.683	99	4719.365	975.365	4748.103	988.103
86	TN+F+I	-2258.28	101	4718.561	974.561	4748.552	988.552
87	TIM3e+G4	-2261.255	99	4720.51	976.51	4749.247	989.247
88	TVM+F+G4	-2256.108	103	4718.216	974.216	4749.492	989.492
89	TIM2+F+I	-2257.643	102	4719.286	975.286	4749.915	989.915
90	K3Pu+F+G4	-2258.971	101	4719.941	975.941	4749.933	989.933
91	TPM3u+F+G4	-2259.716	101	4721.433	977.433	4751.424	991.424
92	TPM3+F+G4	-2259.717	101	4721.434	977.434	4751.425	991.425
93	GTR+F+G4	-2255.75	104	4719.5	975.5	4751.43	991.43
94	TIM+F+G4	-2258.638	102	4721.276	977.276	4751.906	991.906
95	K2P+G4	-2265.454	97	4724.907	980.907	4752.421	992.421
96	TNe+G4	-2264.219	98	4724.437	980.437	4752.559	992.559
97	TIM3+F+G4	-2259.366	102	4722.732	978.732	4753.361	993.361
98	TIM2e+G4	-2263.57	99	4725.141	981.141	4753.878	993.878
99	JC+R2	-2266.233	97	4726.466	982.466	4753.98	993.98
100	F81+F+R2	-2262.327	100	4724.654	980.654	4754.015	994.015
101	HKY+F+G4	-2262.499	100	4724.999	980.999	4754.359	994.359
102	TPM2+F+G4	-2261.915	101	4725.829	981.829	4755.82	995.82
103	TPM2u+F+G4	-2261.915	101	4725.829	981.829	4755.82	995.82
104	TN+F+G4	-2262.169	101	4726.338	982.338	4756.329	996.329
105	TIM2+F+G4	-2261.585	102	4727.17	983.17	4757.8	997.8
106	F81+F+R3	-2262.028	102	4728.056	984.056	4758.685	998.685
107	JC+R3	-2265.997	99	4729.994	985.994	4758.731	998.731
108	F81+F+I+G4	-2274.845	100	4749.69	1005.69	4779.05	1019.05
109	JC+I+G4	-2279.318	97	4752.636	1008.636	4780.149	1020.149
110	F81+F+I	-2283.56	99	4765.119	1021.119	4793.857	1033.857
111	JC+I	-2287.984	96	4767.968	1023.968	4794.881	1034.881
112	F81+F+G4	-2287.834	99	4773.669	1029.669	4802.406	1042.406
113	JC+G4	-2292.095	96	4776.19	1032.19	4803.103	1043.103
114	*GY94* +F1X4+R2	-2242.963	102	4689.926	945.926	4821.251	1061.251
115	MGK+F1X4+R2	-2243.111	102	4690.221	946.221	4821.546	1061.546
116	*GY94* +F1X4+R3	-2238.022	104	4684.043	940.043	4822.271	1062.271
117	MGK+F3X4+R2	-2229.923	108	4675.846	931.846	4828.729	1068.729
118	*GY94* +F1X4+I+G4	-2247.179	102	4698.359	954.359	4829.684	1069.684
119	MGK+F1X4+I+G4	-2247.292	102	4698.583	954.583	4829.908	1069.908
120	MGK+F1X4+R3	-2241.989	104	4691.978	947.978	4830.206	1070.206
121	MGK+F3X4+R3	-2224.78	110	4669.559	925.559	4830.217	1070.217
122	*GY94* +F1X4+G4	-2251.144	101	4704.287	960.287	4832.263	1072.263

Table 1 Continued

No.	Model	LnL	n	AIC	ΔAIC	AICc	ΔAICc
123	MGK+F1X4+G4	-2251.472	101	4704.944	960.944	4832.919	1072.919
124	*GY94* +F3X4+R3	-2227.048	110	4674.096	930.096	4834.754	1074.754
125	*GY94* +F3X4+R2	-2233.068	108	4682.136	938.136	4835.019	1075.019
126	MGK+F3X4+I+G4	-2233.539	108	4683.078	939.0781	4835.962	1075.962
127	MGK+F3X4+G4	-2237.512	107	4689.024	945.024	4838.134	1078.134
128	*GY94* +F3X4+I+G4	-2238.243	108	4692.485	948.485	4845.368	1085.368
129	*GY94* +F3X4+R4	-2227.106	112	4678.213	934.213	4846.96	1086.96
130	*GY94* +F3X4+G4	-2242.394	107	4698.789	954.789	4847.899	1087.899
131	*GY94* +F1X4+I	-2260.085	101	4722.169	978.169	4850.144	1090.144
132	MGK+F1X4+I	-2260.345	101	4722.69	978.69	4850.665	1090.665
133	MGK+F3X4+I	-2246.112	107	4706.225	962.225	4855.335	1095.335
134	MG+F1X4+R2	-2268.482	101	4738.963	994.963	4866.938	1106.938
135	*GY94* +F3X4+I	-2252.532	107	4719.064	975.064	4868.174	1108.174
136	MG+F3X4+R2	-2254.453	107	4722.906	978.906	4872.015	1112.015
137	MG+F1X4+I+G4	-2272.057	101	4746.113	1002.113	4874.089	1114.089
138	MG+F1X4+R3	-2267.523	103	4741.047	997.047	4875.789	1115.789
139	MG+F1X4+G4	-2276.171	100	4752.342	1008.342	4877.033	1117.033
140	MG+F3X4+I+G4	-2257.945	107	4729.891	985.891	4879.001	1119.001
141	MG+F3X4+G4	-2261.949	106	4735.898	991.898	4881.309	1121.309
142	MG+F3X4+R3	-2253.514	109	4725.027	981.027	4881.759	1121.759
143	SYM	-2329.878	100	4859.756	1115.756	4889.116	1129.116
144	TIMe	-2333.105	98	4862.21	1118.21	4890.332	1130.332
145	TIM3e	-2333.481	98	4862.961	1118.961	4891.083	1131.083
146	TVMe	-2333.164	99	4864.328	1120.328	4893.065	1133.065
147	GTR+F	-2328.404	103	4862.809	1118.809	4894.085	1134.085
148	K3P	-2336.391	97	4866.783	1122.783	4894.297	1134.297
149	MG+F1X4+I	-2284.946	100	4769.892	1025.892	4894.583	1134.583
150	TVM+F	-2330.086	102	4864.172	1120.172	4894.802	1134.802
151	TIM+F	-2331.48	101	4864.96	1120.96	4894.952	1134.952
152	TNe	-2336.729	97	4867.458	1123.458	4894.972	1134.972
153	K3Pu+F	-2333.162	100	4866.323	1122.323	4895.684	1135.684
154	TIM3+F	-2331.971	101	4865.942	1121.942	4895.934	1135.934
155	TPM3+F	-2333.648	100	4867.297	1123.297	4896.657	1136.657
156	TPM3u+F	-2333.648	100	4867.297	1123.297	4896.657	1136.657
157	TIM2e	-2336.292	98	4868.584	1124.584	4896.706	1136.706
158	MG+F3X4+I	-2270.442	106	4752.885	1008.885	4898.295	1138.295
159	K2P	-2340.015	96	4872.03	1128.03	4898.943	1138.943
160	TN+F	-2335.102	100	4870.204	1126.204	4899.565	1139.565
161	HKY+F	-2336.783	99	4871.566	1127.566	4900.303	1140.303
162	TIM2+F	-2334.7	101	4871.401	1127.401	4901.392	1141.392
163	TPM2u+F	-2336.381	100	4872.761	1128.761	4902.122	1142.122
164	TPM2+F	-2336.381	100	4872.762	1128.762	4902.123	1142.123

Table 1 Continued

No.	Model	LnL	n	AIC	ΔAIC	AICc	ΔAICc
165	JC	-2366.286	95	4922.571	1178.571	4948.892	1188.892
166	F81+F	-2362.554	98	4921.108	1177.108	4949.229	1189.229
167	*GY94* +F1X4	-2315.788	100	4831.575	1087.575	4956.267	1196.267
168	KOSI07+FU+R2	-2325.725	97	4845.45	1101.45	4960.675	1200.675
169	MGK+F1X4	-2318.048	100	4836.095	1092.095	4960.787	1200.787
170	KOSI07+FU+R3	-2323.063	99	4844.126	1100.126	4965.599	1205.599
171	MGK+F3X4	-2304.357	106	4820.713	1076.713	4966.124	1206.124
172	*GY94* +F3X4	-2306.17	106	4824.339	1080.339	4969.749	1209.749
173	KOSI07+FU+I+G4	-2335.554	97	4865.108	1121.108	4980.332	1220.332
174	KOSI07+FU+G4	-2339.513	96	4871.026	1127.026	4983.218	1223.218
175	KOSI07+F3X4+R2	-2315.814	106	4843.627	1099.627	4989.038	1229.038
176	KOSI07+F3X4+R3	-2310.509	108	4837.018	1093.018	4989.901	1229.901
177	KOSI07+F1X4+R2	-2333.491	100	4866.983	1122.983	4991.674	1231.674
178	KOSI07+F1X4+R3	-2328.692	102	4861.383	1117.383	4992.708	1232.708
179	SCHN05+FU+R2	-2344.705	97	4883.411	1139.411	4998.635	1238.635
180	KOSI07+F1X4+I+G4	-2337.965	100	4875.93	1131.93	5000.621	1240.621
181	KOSI07+F1X4+G4	-2341.156	99	4880.312	1136.312	5001.784	1241.784
182	SCHN05+FU+R3	-2341.179	99	4880.358	1136.358	5001.831	1241.831
183	KOSI07+FU+I	-2349.617	96	4891.233	1147.233	5003.426	1243.426
184	KOSI07+F3X4+I+G4	-2323.767	106	4859.534	1115.534	5004.944	1244.944
185	MG+F1X4	-2342.797	99	4883.593	1139.593	5005.065	1245.065
186	KOSI07+F3X4+G4	-2327.376	105	4864.751	1120.751	5006.534	1246.534
187	MG+F3X4	-2328.539	105	4867.078	1123.078	5008.861	1248.861
188	SCHN05+F1X4+R3	-2340.927	102	4885.854	1141.854	5017.179	1257.179
189	KOSI07+F1X4+I	-2349.1	99	4896.2	1152.2	5017.672	1257.672
190	SCHN05+F3X4+R3	-2324.472	108	4864.944	1120.944	5017.827	1257.827
191	SCHN05+FU+I+G4	-2354.523	97	4903.046	1159.046	5018.27	1258.27
192	SCHN05+F1X4+R2	-2348.226	100	4896.452	1152.452	5021.143	1261.143
193	SCHN05+F3X4+R2	-2331.916	106	4875.833	1131.833	5021.243	1261.243
194	SCHN05+FU+G4	-2358.682	96	4909.365	1165.365	5021.558	1261.558
195	KOSI07+F3X4+I	-2336.826	105	4883.653	1139.653	5025.436	1265.436
196	SCHN05+F1X4+I+G4	-2351.096	100	4902.192	1158.192	5026.883	1266.883
197	SCHN05+F1X4+G4	-2353.895	99	4905.79	1161.79	5027.263	1267.263
198	SCHN05+F1X4+R4	-2340.593	104	4889.187	1145.187	5027.414	1267.414
199	SCHN05+F3X4+R4	-2324.102	110	4868.203	1124.203	5028.861	1268.861
200	SCHN05+F3X4+I+G4	-2338.345	106	4888.69	1144.69	5034.101	1274.101
201	SCHN05+F3X4+G4	-2341.811	105	4893.621	1149.621	5035.404	1275.404
202	SCHN05+FU+I	-2370.471	96	4932.943	1188.943	5045.135	1285.135
203	SCHN05+F1X4+I	-2363.696	99	4925.391	1181.391	5046.864	1286.864
204	SCHN05+F3X4+I	-2352.81	105	4915.621	1171.621	5057.404	1297.404
205	KOSI07+FU	-2394.782	95	4979.563	1235.563	5088.785	1328.785
206	KOSI07+F1X4	-2398.44	98	4992.88	1248.88	5111.197	1351.197

Table 1 Continued

No.	Model	LnL	n	AIC	ΔAIC	AICc	ΔAICc
207	KOSI07+F3X4	-2383.159	104	4974.318	1230.318	5112.546	1352.546
208	SCHN05+FU	-2419.333	95	5028.665	1284.665	5137.887	1377.887
209	SCHN05+F1X4	-2416.544	98	5029.088	1285.088	5147.405	1387.405
210	SCHN05+F3X4	-2402.838	104	5013.675	1269.675	5151.903	1391.903
211	*GY94* +F+R2	-2208.59	159	4735.181	991.181	5229.161	1469.161
212	*GY94* +F+G4	-2217.694	158	4751.388	1007.388	5234.504	1474.504
213	*GY94* +F+I+G4	-2213.659	159	4745.319	1001.319	5239.299	1479.299
214	*GY94* +F+R3	-2202.599	161	4727.198	983.198	5243.673	1483.673
215	*GY94* +F+I	-2228.346	158	4772.691	1028.691	5255.807	1495.807
216	*GY94* +F+R4	-2202.61	163	4731.219	987.219	5271.26	1511.26
217	*GY94* +F	-2282.254	157	4878.509	1134.509	5351.004	1591.004
218	KOSI07+F+R2	-2291.643	157	4897.286	1153.286	5369.781	1609.781
219	KOSI07+F+G4	-2301.662	156	4915.325	1171.325	5377.438	1617.438
220	KOSI07+F+I+G4	-2298.418	157	4910.835	1166.835	5383.33	1623.33
221	KOSI07+F+R3	-2286.723	159	4891.446	1147.446	5385.426	1625.426
222	KOSI07+F+I	-2311.78	156	4935.559	1191.559	5397.672	1637.672
223	SCHN05+F+R2	-2310.015	157	4934.03	1190.03	5406.525	1646.525
224	SCHN05+F+G4	-2316.684	156	4945.369	1201.369	5407.482	1647.482
225	SCHN05+F+I+G4	-2313.733	157	4941.467	1197.467	5413.962	1653.962
226	SCHN05+F+R3	-2303.732	159	4925.463	1181.463	5419.444	1659.444
227	SCHN05+F+I	-2327.127	156	4966.254	1222.254	5428.367	1668.367
228	SCHN05+F+R4	-2303.45	161	4928.9	1184.9	5445.375	1685.375
229	KOSI07+F	-2357.579	155	5025.157	1281.157	5477.12	1717.12
230	SCHN05+F	-2379.264	155	5068.528	1324.528	5520.491	1760.491