Reports: ~2500 words including refs; no more than 4 fig+table. Abstract, introductory paragraph, ~30 refs. Materials and methods in supplemental

Abstract:

Body:

Protein coding sequences evolve in response to a mixture of natural selection, mutational bias, and drift ([*1*](#_ENREF_1)*,* [*2*](#_ENREF_2)). While this has been known since the Modern Synthesis, and in spite of the explosion of sequence data from genomics studies and other studies using next gen sequencing approaches, the methods used to analyze coding sequences generally ignore one or more of these processes. Most models of nucleotide sequence evolution assume constant, reversible substitution rates over a specified set of sites. Goldman and Yang ([*3*](#_ENREF_3)) developed a codon model that incorporated different transition rates between synonymous and nonsynonymous sites. Nevertheless, this model and its descendants ([*4-7*](#_ENREF_4)), which incorporate heterogeneity across sites and taxa, retain the assumption that the substitution rate from codon *i* to codon *j* equals the reverse rate, even though in reality these two codons are expected to have unequal fitness, especially if they differ in amino acid. Various models that incorporate mutation and selection on codons and amino acids ([*8*](#_ENREF_8)*,* [*9*](#_ENREF_9)) also inherit this limitation of equal rates. Two models have been advanced that deal with non-time-reversible models. Seoighe et al. ([*10*](#_ENREF_10)) developed a non-time-reversible model for evolution that allows a different rate of evolution to a specified optimal amino acid. Kosakovsky Pond et al. ([*11*](#_ENREF_11)) developed a model based on this that also allows biased substitutions towards an optimal amino acid. Amino acid models are used less frequently for phylogenetic inference but typically share this assumption that the rate of going from amino acid *i* to amino acid *j* equals the reverse rate. Many of these matrices are fit once from empirical data (PAM, BLOSSUM\_\_\_\_\_\_\_\_); some are estimated anew from each dataset (\_\_\_\_\_\_), but these all share this symmetry assumption. Another assumption is that sites in a given, usually prespecified set, share a transition matrix. This is relaxed in some models by summing likelihoods across multiple matrices differing by scaling (\_\_gamma\_\_\_)\_\_ or individual rates (PAGEL\_\_\_\_), or by site specific models. Many models also describe patterns without getting at the underlying process. Here we develop a family of amino acid models that mechanistically include drift, nucleotide mutation, and selection on amino acids. These models fit far better than do competing models based on model selection, do a better job predicting data, and allow inference of meaningful parameter values.

Our model includes a mutation rate matrix and a fixation rate matrix to get at a net substitution rate matrix that incorporates both components. Mutation occurs at individual nucleotides; nonsynonymous changes lead to changes in amino acids. Fixation probability depends on the relative fitness of two alleles, and is based on the model of Sella and Hirsh ([2005](#_ENREF_24)). The fitness of a genotype is based on the idea of that the products of its translation can be described using a cost-benefit function ([*12-14*](#_ENREF_12)). We assume there is a certain level of functionality (such as reactions catalyzed per second) required. Proteins must be produced to meet this functionality and have a fixed cost of production based on their length; less efficient proteins must be produced at a greater rate to achieve this level of functionality and thus incur more cost (cost is based on protein length and the rate of production). We assume this functionality is an inverse function of the physiochemical distance between the observed and the optimal amino acid (*di*) as well as the sensitivity of the protein’s functionality to such distances (*g*). Physiochemical distance is measured using the three factors Grantham ([*15*](#_ENREF_15)) developed: composition (the relative amount of non-carbon atoms), polarity, and molecular volume. While Grantham assigned weights to these factors, we allow these weights to be estimated from the data. This cost-benefit function is scaled by the target protein production rate for that gene: the more of the protein product that is required, the stronger the selection on the sequence to reduce its cost-benefit function. Rates of mutation between amino acids is based on rates of mutations between codons based on a GTR \_\_\_\_ substitution matrix. Fixation probability is based on the

**1. W. H. Li, Maintenance of genetic variability under the joint effect of mutation, selection and random drift. *Genetics* 90, 349 (1978).**

**2. M. Kimura, J. F. Crow, The number of alleles that can be maintained in a finite population. *Genetics* 49, 725 (1964).**

**3. N. Goldman, Z. H. Yang, Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA-Sequences. *Molecular Biology and Evolution* 11, 725 (Sep, 1994).**

**4. Z. H. Yang, W. J. Swanson, Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular Biology and Evolution* 19, 49 (Jan, 2002).**

**5. Z. H. Yang, R. Nielsen, Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* 19, 908 (Jun, 2002).**

**6. Z. H. Yang, R. Nielsen, N. Goldman, A. M. K. Pedersen, Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431 (May, 2000).**

**7. R. Nielsen, Z. H. Yang, Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929 (Mar, 1998).**

**8. Z. Yang, R. Nielsen, Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution* 25, 568 (Mar, 2008).**

**9. S. Miyazawa, Selective Constraints on Amino Acids Estimated by a Mechanistic Codon Substitution Model with Multiple Nucleotide Changes. *PLoS ONE* 6, (Mar 18, 2011).**

**10. C. Seoighe *et al.*, A Model of Directional Selection Applied to the Evolution of Drug Resistance in HIV-1. *Molecular Biology and Evolution* 24, 1025 (2007).**

**11. S. L. Kosakovsky Pond, A. F. Y. Poon, A. J. Leigh Brown, S. D. W. Frost, A Maximum Likelihood Method for Detecting Directional Evolution in Protein Sequences and Its Application to Influenza A Virus. *Molecular Biology and Evolution* 25, 1809 (2008).**

**12. M. A. Gilchrist, Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Molecular Biology and Evolution* 24, 2362 (2007).**

**13. M. A. Gilchrist, P. Shah, R. Zaretzki, Measuring and Detecting Molecular Adaptation in Codon Usage Against Nonsense Errors During Protein Translation. *Genetics* 183, 1493 (Dec, 2009).**

**14. P. Shah, M. A. Gilchrist, Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proceedings of the National Academy of Sciences* 108, 10231 (2011).**

**15. R. Grantham, Amino acid difference formula to help explain protein evolution. *Science* 185, 862 (Sep 6, 1974).**

**16. G. Sella, A. E. Hirsh, The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences of the United States of America* 102, 9541 (2005).**