A phylogenetic and population genetic model of amino acid substitution

March 25, 2013

1 Abstract

We introduce a new, mechanistic Markov model for studying the evolution of amino acid sequences within a phylogenetic context. Our model links together genotype, phenotype, and fitness of proteins, by calculating the fixation probability of a newly arisen mutant using a model of allele substitution from population genetics. As a result, our model explicitly includes the effects of mutation bias, genetic drift, and natural selection favoring an optimal amino acid sequence for a given protein. We assume that the strength of purifying selection is a function of the physiochemical differences between a given amino acid and the optimal amino acid for the site, how a protein's functionality declines with this distance, and the target expression level of the gene. Analysis of a multi-locus yeast data set using AIC shows that our new model provides a substantially better fit to data than the standard empirical models and allows researchers to estimate biologically meaningful parameters such as the sensitivity of a protein's function to an amino acid substitution and the optimal amino acids at a given site. Further, because our model is based on explicit models of various biological processes, unlike empirical models it can easily be modified in the future to include other important biological phenomena such as selection on codon usage or test alternative hypotheses about the relationship between amino acid sequence and protein functionality.

2 Introduction

TO ADD: Importance of building accurate model for protein evolution.

In phylogenetics, models for evolution of protein-encoding sequences are usually formulated at three levels: mono-nucleotide level in DNA (e.g., see Jukes and Cantor, 1969; Kimura, 1980; Felsenstein, 1981; Hasegawa et al., 1985), codon level (see Goldman and Yang, 1994; Muse and Gaut, 1994; Yang et al., 1998; Yang and Nielsen, 2008), and amino acid (AA) level (see Kishino et al., 1990). The nucleotides, codons, or amino acids are assumed to evolve independently. DNA- and codon-based models use more data and are often the most powerful in terms of their ability to distinguish between closely related sequences. Of these two, codon-based models use all the information in DNA and know the product of amino acid in addition. On the other hand, AA-based models ignore synonymous differences between sequences by focusing not on the codons themselves but the amino acids they code for. Since synonymous codon usage is largely driven by mutation bias in low expression genes and selection on translational efficiency for high expression genes (see Gilchrist, 2007)[CHECK CITATIONS], ignoring this aspect of the data has the advantage of reducing the noise in sequence data for low expression genes but at the cost of losing potentially useful information held in the high expression genes.

Based on how the substitution rates are formulated, models of amino acid substitution fall into two categories: empirical models and mechanistic models. In empirical models, the substitution rates are based solely on analysis of large quantities of sequence data complied from databases. Commonly used models in this category include Dayhoff (Dayhoff et al., 1978), JTT (Jones et al., 1992), WAG (Whelan and Goldman, 2001), LG (Le and Gascuel, 2008) for nuclear proteins; mtREV, MTMAM (Adachi and Hasegawa, 1996; Yang et al., 1998) for mitochondrial proteins; and cpREV (Adachi et al., 2000) model for chloroplast proteins, etc. (see Cao et al., 1994; Henikoff and Henikoff, 1992; Gonnet et al., 1992) In contrast, mechanistic models consider the actual biological processes thought to drive sequence evolution, such as mutation bias in DNA, translation of codons into amino acids, and natural selection.

Goldman and Yang (1994) implemented a mechanistic model (GY) at the level of codons and explicitly modeled the biological processes involved, including different mutation rates between nucleotides (transition vs. transversion bias), the translation of the codon triplet into an amino acid (synonymous vs. non-synonymous rates), and the acceptance or rejection of the amino acid due to selective pressure on the protein. The selective restraints at the amino acid level was accounted for by multiplying the substitution rate by a factor $\exp(d_{aa_i,aa_j}/V)$ where d_{aa_i,aa_j} is the distance between amino acids aa_i and aa_j given by Grantham (1974) (i.e. Grantham Distances) and V is a parameter representing the variability of the gene or its tendency to undergo non-synonymous substitution. The model in common use is a simplified version of this model that ignores the effect of selection. Yang et al. (1998) implemented a few mechanistic models on the codon level and found from analysis of mitochondrial genomes of 20 mammalian species that they fit the data better than empirical models.

It is important to note that whether empirical or mechanistic, most phylogenetic models, including the GY model, are time-reversible. In time-reversible models, the relative substitution rate q_{ij} from state i to state j is assumed to satisfy the detailed balance condition $\pi_i q_{ij} = \pi_j q_{ji}$ for any $i \neq j$. While this assumption provides substantial mathematical and computational advantages, the substitution rate matrices are difficult to interpret biologically if one assumes natural selection is acting consistently on a given site. Surprisingly this lack of inconsistency has gone largely unrecognized.

By definition, if the substitution from state i to j is favored by natural selection, in the absence of mutation bias it must occur at a faster rate than the reverse substitution. While mutation bias can alter this requirement when selection is weak, it can only do so when the assumption of time reversibility is violated. (JJ - is this true? substitution rates are different from exchange rates, even in time-reversible models. Under time-reversible model, the higher equilibrium state frequency indicates higher mutation rate to this state. In the discussion follows, consider the case where amino acid i has a higher equilibrium frequency than amino acid j, even if in the exchange rate matrix synonymous rates are bigger than non synonymous rates, it's still possible that the substation rates reflect the "optimality" of amino acids.) For example, consider a time-reversible codon substitution model where synonymous substitutions occur at a faster rate than non-synonymous substitutions. For any given state of the system, such a model implies that the current amino acid is optimal since synonymous substitutions occur at a faster rate than non-synonymous ones. However, once a non-synonymous substitution has occurred (and as time goes to infinity it will), the time reversible aspects of the model now imply that the new state is the optimal state and the old state is sub-optimal. Thus, the only reasonable way to interpret such a time-reversible model is that the substitution matrix is actually describing the rate at which the optimal state switches at a given site and that once such a switch has occurred the system instantaneously shifts to the new state. If, in contrast, one were to assume the converse, that non-synonymous substitution occur at a faster rate than synonymous, then the interpretation of time-reversible models becomes even more problematic from a biological perspective. In such a scenario, not only is the optimal state constantly changing, the current state of any given site is always sub-optimal.

While time-reversible models have played an important role in molecular phylogenetics for the last decade (Tavaré, 1986), in order to model natural selection and mutation bias in a realistic manner the assumption of time reversibility must be relaxed. In this study we develop an amino-acid based model in which we assume that for each individual site i of a protein there is a corresponding optimal amino acid a'_i. The optimal state can be assigned or, as we demonstrate, estimated from the data itself. As with the GY model, we assume that the substitution rate between amino acids at a given site is a function of their Grantham distances from optimal amino acids and, assume genes can vary in their sensitivities to such deviation from optimal amino acids. Here the sensitivity to amino acid changes is calculated using a cost-benefit framework we have developed previously for studying the evolution codon usage bias (Gilchrist, 2007; Gilchrist et al., 2009; Shah and Gilchrist, 2011). Furthermore, unlike most models in phylogenetics, we define the fitness of a given protein explicitly. We then use a model from population genetics to calculate the substitution rate between any two genotypes by explicitly taking into account the fitness differences between them as well as the effects of mutation bias and genetic drift. We illustrate our model by fitting it to the Rokas et al. (2003)'s data set of 106 genes sequenced from 8 different species of yeast with a maximum likelihood approach. When fitting our model, we estimate the phylogenies of the yeast species, the Grantham sensitivity g of a gene (roughly comparable to $1/V$ in the GY model), as well as state optimal amino acid $\vec{a}'$ for any given site in a coding sequence. We compare our model's fit to the Rokas data with other commonly used AA-based models using AIC criterion (Akaike, 1973, 1974, 1981). [LIST MODELS – IF WE STICK TO AA MODELS, WHY DO

WE SPEND SO MUCH TIME DISCUSSING THE GY MODEL WHICH IS A CODON LEVEL MODEL? –ANSWER: WE DON'T HAVE A CODON MODEL, WE HAVE AN AA MODEL WHERE THE MUTATION RATES ARE BASED ON DNA DATA. SO WE TRY TO EXPLAIN THE AA DATA AND COMPARE IT TO AA MODELS – our model is similar to GY model in that grantham distance is used, but we also include the effect of natural selection explicitly.].

These results show that even with our most parameter rich model in which we estimate the optimal amino acid at every site, thereby introducing tens of thousands of additional parameters, our model still does a substantially better job fitting the Rokas dataset. So although the computational cost of our model is greater than most time reversible models, our ability to fit the phylogenetic data and extract biologically meaningful information is substantially greater than other models. Furthermore, because our approach explicitly links genotype to phenotype, phenotype to fitness, and fitness to fixation rate, the biological assumptions underlying our model are clearly stated and our ability to incorporate additional biological factors, such as selection on codon usage bias, are greatly enhanced.

References

J. Adachi and M. Hasegawa. Model of amino acid substitution in proteins encoded by mitochondrial dna. *Journal of Molecular Evolution*, 42(4): 459–468, 1996.

J. Adachi, P. J. Waddell, W. Martin, and M. Hasegawa. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast dna. *Journal of Molecular Evolution*, 50(4):348–358, 2000.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory, 2 nd, Tsahkadsor, Armenian SSR*, pages 267–281, 1973.

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

H. Akaike. Likelihood of a model and information criteria. *Journal of econometrics*, 16(1):3–14, 1981.

Y. Cao, J. Adachi, A. Janke, S. Pääbo, and M. Hasegawa. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *Journal of Molecular Evolution*, 39(5):519–527, 1994.

M. Dayhoff, R. Schwartz, and B. Orcutt. A model of evolutionary change in proteins. *In Atlas of Protein Sequences and Structure*, 5:345–352, 1978.

J. Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981. ISSN 0022-2844. doi: 10.1007/BF01734359. URL `http://dx.doi.org/10.1007/BF01734359`.

M. A. Gilchrist. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol Biol Evol*, 24(11):2362–72, 2007.

M. A. Gilchrist, P. Shah, and R. Zaretzki. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics*, 183(4):1493–1505, Dec 2009. doi: 10.1534/genetics.109.108209. URL `http://www.hubmed.org/fulltext.cgi?uids=19822731`.

N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol Biol Evol*, 11(5):725–36, 1994.

G. H. Gonnet, M. A. Cohen, S. A. Benner, et al. Exhaustive matching of the entire protein sequence database. *Science*, 256(5062):1443–1445, 1992.

R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864, 1974. doi: 10.1126/science.185.4154.862. URL `http://www.sciencemag.org/content/185/4154/862.abstract`.

M. Hasegawa, H. Kishino, and T.-a. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, 22:160–174, 1985. ISSN 0022-2844. doi: 10.1007/BF02101694. URL `http://dx.doi.org/10.1007/BF02101694`.

S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences: CABIOS*, 8(3):275–282, 1992.

T. Jukes and C. Cantor. Evolution of protein molecules. *Manmmalian Protein Metabolism*, III:21–132, 1969.

M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980. ISSN 0022-2844. doi: 10.1007/BF01731581. URL `http://dx.doi.org/10.1007/BF01731581`.

H. Kishino, T. Miyata, and M. Hasegawa. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 31(2):151–160, 1990.

S. Q. Le and O. Gascuel. An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25(7):1307–1320, 2008.

S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, 1994. URL `http://mbe.oxfordjournals.org/content/11/5/715.abstract`.

A. Rokas, B. L. Williams, N. King, and S. B. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, Oct 2003. doi: 10.1038/nature02053. URL `http://www.hubmed.org/fulltext.cgi?uids=14574403`.

P. Shah and M. A. Gilchrist. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc Natl Acad Sci U S A*, 108(25): 10231–10236, Jun 2011. doi: 10.1073/pnas.1016719108. URL `http://www.hubmed.org/fulltext.cgi?uids=21646514`.

S. Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lect. Math. Life Sci*, 17:57–86, 1986.

S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5):691–699, 2001.

Z. Yang and R. Nielsen. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol*

Evol, 25(3):568–579, Mar 2008. doi: 10.1093/molbev/msm284. URL `http://www.hubmed.org/fulltext.cgi?uids=18178545`.

Z. Yang, R. Nielsen, and M. Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol*, 15 (12):1600–11, 1998.