Paper submitted for publication as a Letter (Methods)

A codon model of nucleotide substitution with selection on synonymous codon usage

Laura Kubatko[1,2,*], Premal Shah[3], Radu Herbei[1], and Michael A. Gilchrist[4]

[1] Department of Statistics, The Ohio State University, Columbus, OH 43210

[2] Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210

[3] Department of Biology, University of Pennsylvania Philadelphia, PA 19104

[4] Department of Ecology and Evolutionary Biology, University of Tennessee - Knoxville, Knoxville, TN 37996-1610

[*] Author for correspondence: Laura Kubatko, lkubatko@stat.osu.edu

ABSTRACT

The quality of phylogenetic inference made from protein-coding genes depends, in part, on the realism with which the codon substitution process is modeled. Here we propose a new mechanistic model that combines the standard M0 substitution model of Yang (1997) with a simplified model from Gilchrist (2007) that includes selection on synonymous substitutions as a function of codon-specific nonsense error rates. We tested the newly proposed model by applying it to 104 protein-coding genes in brewer's yeast, and compared the fit of the new model to the standard M0 model and to the mutation-selection model of Yang and Nielsen (2008) using the AIC. Our new model provided significantly better fit in approximately 85% of the cases considered for the basic M0 model and in approximately 25% of the cases for the M0 model with estimated codon frequencies, but only in a few cases when the mutation-selection model was considered. However, our model includes a parameter that can be interpreted as a measure of the rate of protein production, and the estimates of this parameter were highly correlated with an independent measure of protein production for the yeast genes considered here. Finally, we found that in some cases the new model led to the preference of a different phylogeny for a subset of the genes considered, indicating that substitution model choice may have an impact on the estimated phylogeny.

Introduction

Successful phylogenetic inference based on protein-coding genes relies on use of an appropriate model of sequence evolution. Models in current use are of one of two classes: models that use information in the nucleotides only, without taking into account information about codons or amino acids (e.g., the GTR model (Tavare, 1986; Yang, 1994; Zharkikh, 1994) and its submodels), and codon-based models that are designed to incorporate information about rates of change between pairs of amino acids (e.g., Yang et al. (1998); Kosiol et al. (2007)). Codon-based models can be further classified into empirical and mechanistic models (Kosiol et al., 2007). Empirical models use the observed frequencies of changes in state observed within large data sets to specify the rate of change used in the model. These estimated rates are assumed to be applicable to a broad set of sequence data sets, and thus parameters are not generally estimated separately for a particular data set. Empirical models have been widely used for modeling the amino acid substitution processes (Dayhoff and Eck, 1968; Dayhoff et al., 1972, 1978; Jones et al., 1992; Whelan and Goldman, 2001; Jones et al., 1994; Goldman et al., 1996, 1998; Adachi and Hasegawa, 1996; Adachi et al., 2000; Dimmic et al., 2002; Yang, 1994).

Mechanistic models, on the other hand, specify an explicit model for the evolutionary process using features such as selective pressures acting on certain types of changes and the varying frequencies of codons in the data. Nearly all codon models in common use are mechanistic (but see Kosiol et al. (2007) and references therein). In general, codon substitution models are based on Markov models of the rates of nucleotide substitution and typically include a parameter to quantify differences in the rates of synonymous versus nonsynonymous substitutions. The magnitude of this parameter, $\ll 1$ or $\gg 1$, is often taken as evidence for protein-level stabilizing or diversifying selection, respectively (Goldman and Yang, 1994; Yang and Nielsen, 1998; Yang and Bielawski, 2000; Yang et al., 2000). Extensions of this approach have been proposed to test for selection at specific sites in the sequence, in specific lineages of the phylogeny, or both (Yang et al., 2000; Wong et al., 2004; Massingham and Goldman, 2005; Yang and Nielsen, 1998, 2002).

Within the last ten years, these basic models have been extended in several important directions to attempt to capture the complexities in the codon substitution process in the presence of selection. For example, Kosakovsky Pond and Muse (2005) allowed the rates of synonymous and nonsynonymous substitutions to vary across position in the sequence, while Mayrose et al. (2007) constructed a family of models to allow both variation in synonymous and nonsynonymous rates across sites and site-to-site dependence in rates via a first-order Markov process. Nielsen et al. (2007) and Zhou et al. (2010) used the idea of "preferred" and "non-preferred" synonymous substitutions, where a synonymous substitution is either preferred or non-preferred at a site depending on selective forces specific to that location. A primary emphasis in both of these studies is the development of ways to measure selection along the genome and across branches of a phylogeny. Several recent models have also separated the substitution rate into components due to the mutational process and the selection process. For example, Yang and Nielsen (2008) introduced the FMutSel model, in which each codon is assigned a fitness parameter; differences in the fitness parameters between two codons are used to specify the substitution rates in the Markov matrix, by modifying the rates specified by the standard mutation models. Similarly, Rodrigue et al. (2010) developed a model in which amino acid propensity scores are used to estimate scaled selection coefficients that are then used to specify substitution rates.

In this paper, we propose a new mechanistic codon-based substitution model that takes into account selection on codon usage of a gene. Our model is based on the M0 model in PAML (Yang, 1997) but where the substitution rate between codons, including synonymous changes, is modified by the substitution's effects on protein production costs and the average protein production rate of the gene. We calculate effects of a codon substitution on the production cost of a protein using a model of protein translation based on the movement of the ribosome along the mRNA of a gene (Gilchrist and Wagner, 2006; Gilchrist, 2007). The production cost includes the effects of nonsense (a.k.a. processivity) errors which result in premature termination of the translation of a protein. More specifically, changes in protein production costs are due to presumed inter-codon variation in nonsense error rates.

We fit our model to empirical data for 104 genes from 8 yeast species, and compare the fit our model to several of the codon models in common use, such as the M0 model in PAML and the FMutSel model of Yang and Nielsen (2008). We also examine the effect of the substitution model on phylogenetic inference by considering which of two competing phylogenies for these 8 yeast species is preferred by various models.

New Approaches

Background: Codon Substitution Models

To motivate development of our method, we review the codon substitution models in common use. Specifically, we give the details of the M0 model implemented in the program PAML. This model uses a continuous-time Markov model for the substitution process between codons in a protein-coding gene. The states in the Markov process are the 61 sense codons (stop codons are not included). The model is then specified by a 61×61 matrix $\mathbf{Q}$, whose entries Q_{ij} give the instantaneous rate of substitution of codon i with codon j and satisfy the constraint $Q_{ii} = -\sum_{j \neq i} Q_{ij}$. The probabilities of substitution of codon i with codon j over time t can then be found by solving the matrix differential equation $\mathbf{P}^{'}(t) = \mathbf{Q}\mathbf{P}(t)$ with initial condition $\mathbf{P}(0) = \mathbf{I}$, which yields $\mathbf{P}(t) = \exp\{\mathbf{Q}t\}$. Thus specification of $\mathbf{Q}$ and the stationary frequencies of the codons is sufficient to compute the substitution probabilities that will be used in modeling the codon mutation process.

Define codon i to be $i_1 i_2 i_3$ and codon j to be $j_1 j_2 j_3$, where $i_k, j_k \in \{A, C, G, T\}$ for $k = 1, 2, 3$. The M0 model specifies $\mathbf{Q}$ as follows:

$$Q_{ij} = \begin{cases} 0, & \text{if 2 or 3 of the pairs } (i_1, j_1), (i_2, j_2), (i_3, j_3) \text{ are different} \\ \mu\kappa\pi_j, & \text{if exactly 1 of the pairs } (i_1, j_1), (i_2, j_2), (i_3, j_3) \\ & \text{is different, and that difference is a synonymous transition} \\ \mu\kappa\omega\pi_j, & \text{if exactly 1 of the pairs } (i_1, j_1), (i_2, j_2), (i_3, j_3) \\ & \text{is different, and that difference is a nonsynonymous transition} \\ \mu\pi_j, & \text{if exactly 1 of the pairs } (i_1, j_1), (i_2, j_2), (i_3, j_3) \\ & \text{is different, and that difference is a synonymous transversion} \\ \mu\omega\pi_j, & \text{if exactly 1 of the pairs } (i_1, j_1), (i_2, j_2), (i_3, j_3) \\ & \text{is different, and that difference is a nonsynonymous transversion} \end{cases}$$

In the above expression, the parameter κ allows for a different rate of substitution for transitions versus transversions. The ω parameter is used to specify a different rate for synonymous and nonsynonymous substitutions. These parameters are typically estimated from the data, and are often used to provide insight into the mode of evolution of a particular gene. For example, an $\omega > 1$ indicates positive selection, while an $\omega < 1$ indicates negative selection. When ω is not significantly different than 1, there is no evidence that the gene is under selection. The π_j parameters give the frequency of each of the 61 possible codons at equilibrium. There are several options for setting these parameters: they may all be set to be equal (the 'Fequal' model in PAML), they may be estimated as $\pi_i = \pi'_{i_1}\pi'_{i_2}\pi'_{i_3}$, where π'_l refers to the empirical frequency of nucleotide l for that gene (the 'F1 × 4' model in PAML), they may be estimated as $\pi_{i,} = \pi'_{i_1,1}\pi'_{i_2,2}\pi'_{i_3,3}$, where $\pi'_{l,j}$ refers to the empirical frequency of nucleotide l at codon position j ($j = 1, 2, 3$) for that gene (the 'F3 × 4' model in PAML), or empirical estimates of each of the 61 codons may be used (the 'Fcodon' model in PAML). Finally, the parameter μ is set so that $-\sum_{i=1}^{61} \pi_i Q_{ii} = 1$, which scales time along the tree to be in units of expected numbers of nucleotide substitutions per codon. Finally, we note that the model above satisfies the condition of time reversibility, i.e., $\pi_i Q_{ij} = \pi_j Q_{ji}$ for all i and j. Given a particular $\mathbf{Q}$ matrix, $\mathbf{P}(t)$ can be calculated using standard numerical algorithms (see (Yang, 2006; Moler and Van Loan, 2003; Golub and van Loan, 1996) for details).

Background: Modeling Protein Translation

Using the approach developed in Gilchrist (2007), we can calculate the cost of producing a complete and functional protein product in the face of nonsense errors. Briefly, the production cost of a protein is equivalent to the ratio of the expected cost to the expected benefit, i.e. functionality, of a protein produced from a given coding sequence (Gilchrist et al., 2009). Functionality is defined on a 0 to 1 scale relative to the functionality of a complete and error-free protein produced from the coding sequence. This allows us to avoid having to consider the specific biological role a protein plays.

Let q_j represent the probability of elongation of a codon of type j, where j ranges over the 61 sense codons $(AAA, \ldots, TTT)$. Let the codon at position i of a gene of length n be represented by c_i, where $i = 1, 2, \ldots, n$. The probability that the codon at the i^{th} position, will be successfully translated is given by $p_i = q_{c_i}$, and thus $1 - p_i$ represents the probability that a nonsense error occurs at that codon. It follows that the expected cost of translating a coding sequence $\vec{c} = \{c_1, c_2, c_3, \ldots c_n\}$ of length n, is

$$\mathbb{E}\left(\text{Cost}\,|\vec{c}\right) = \sum_{i=1}^{n+1} \beta_i \left(\prod_{j=1}^{i-1} p_j\right)(1 - p_i). \tag{1}$$

where β_i represents the energetic investment, in Adenosine Tri-Phosphate molecules (ATPs), of translating i codons successfully. Note that the summation is taken up to $n+1$ to account for the stop codon and $(1 - p_{n+1}) = 1$ by definition. In general, $\beta_i = a_1 + a_2 i$ where a_1 and a_2 represents the cost of translation initiation and protein elongation, respectively, and $a_1 = a_2 = 4\text{ATP}$.

In order to calculate the expected functional benefit of a protein produced from a coding sequence $\vec{c}$ we begin with the simplifying assumption that truncated proteins have zero functionality. Because we measure functionality on a relative scale, such that a complete, error free protein has a functionality of 1, it follows that this expected functionality of a gene is simply equal to the probability of producing a complete protein such that,

$$\mathbb{E}\left(\text{Benefit}\,|\vec{c}\right) = \prod_{i}^{n} p_i. \tag{2}$$

Then the expected cost-benefit of protein production for a given codon sequence, which we call the protein production cost for brevity, is $\eta(\vec{c}) = \mathbb{E}\left(\text{Cost}\,|\vec{c}\right)/\mathbb{E}\left(\text{Benefit}\,|\vec{c}\right)$. By including the denominator into the summation term, we can re-write this as

$$\eta(\vec{c}) = \sum_{i=1}^{n} \beta_i \left(\prod_{j=i+1}^{n} \frac{1}{p_j}\right) \frac{1-p_i}{p_i} + \beta_{n+1}, \tag{3}$$

(see Gilchrist et al. (2007) for more details).

Due to the non-linear nature of the protein production cost η, the effect of changing codon c_k to codon c'_k at position k on η will depend on the codons at the other sites. In other words, the effect of a codon substitution on η is not independent between sites. To take such inter-dependence explicitly into account when trying to formulate the substitution rate matrix $\mathbf{Q}$ would be unfeasible computationally. However, given that the probability of a nonsense error at a codon, $1-p_i$, is much smaller than the elongation probability p_i(Gilchrist, 2007) , the first-order approximation of Equation 3 can be written as

$$\eta(\vec{c}) = \sum_{i=1}^{n} \beta_i \left(\frac{1-p_i}{p_i}\right) + \beta_{n+1} \tag{4}$$

Since different codons have different elongation probabilities, $\eta(\vec{c})$ will vary between alleles of a coding sequence. More specifically, if two alleles, $\vec{c} = \{c_1, c_2, c_3, \ldots, c_k, \ldots, c_n\}$ and $\vec{c'} = \{c_1, c_2, c_3, \ldots, c'_k, \ldots, c_n\}$, differ at codon k and $\eta(\vec{c}) > \eta(\vec{c'})$ then allele $\vec{c'}$ should be *favored* by natural selection because of its lower protein production cost. Based on Equation 4, the cost of substituting codon c_k with c'_k depends only on the position k and the elongation probabilities of the two codons involved, p_k and p'_k:

$$\Delta\eta_{c'_k,c_k} = \eta(\vec{c'}) - \eta(\vec{c}) \tag{5}$$

$$= \beta_k \left(\frac{1-p_{c'_k}}{p_{c'_k}} - \frac{1-p_{c_k}}{p_{c_k}}\right) \tag{6}$$

The goal of our work is to incorporate these differences in protein production costs due to inter-codon variation in elongation probabilities into our substitution model.

A New Codon Substitution Model: MutNSE

The main idea behind our new model, which we call MutNSE to indicate that the model incorporates both the mutation process and nonsense errors, is to combine the features of the models described in the previous two subsections in order to incorporate two separate features of the process of the codon substitution in an explicit manner. These features are (1) the typically modeled rate biases (transition/transversion and synonymous/nonsynonymous), and (2) the change in the probability of the protein being ultimately produced following substitution of one amino acid by another. For (2), we note that the probability of successful protein production following codon substitution varies throughout the sequence. We take this aspect of the model into account by specifying a different instantaneous rate matrix $\mathbf{Q}$ for each codon position k in the sequence as follows:

$$Q_{ij}^k = Q_{ij} \exp\left(a\Delta\eta_{c_k',c_k}\right) \qquad (7)$$

where Q_{ij} refers to the substitution rates in the standard codon substitution model of choice and $\Delta\eta_{c_k',c_k}$ is defined in (6). Selection on codon usage has been known to vary between genes. In particular, codon usage in genes with low expression is driven by patterns of mutation biases, while codon usage in high expression genes is primarily driven by natural selection (Sharp and Li (1986), Bulmer (1991), Shah and Gilchrist (2011), Wallace et al. (2013)). The parameter a scales the contribution of selection on codon usage to the overall substitution rate and is a free parameter in our model that we estimate. When there is no selection on codon usage, then $a \approx 0$ and the model reduces to the standard codon substitution model. When a is large, the substitution probabilities depend primarily on changes in protein production costs η.

Our model makes certain simplifying assumptions about the evolution of tRNA copy number and expression levels of genes. Similar to other codon-based models of protein evolution that incorporate selection on individual codons, we assume that the selection on synonymous codons remains fairly constant across the phylogenetic breadth of organisms under consideration. In our mechanistic framework, this translates to the assumption that the variation in tRNA copy numbers and expression levels of genes is quite small across the phylogeny.

Our model requires calculation of separate $\mathbf{Q}$ and $\mathbf{P}(t)$ matrices for each codon position k in the sequence. Because this is computationally intensive, we have developed a Graphical Processing Unit (GPU) implementation of this step of the method. In the past few years, GPUs have made a significant contribution to scientific computing due to their ability to perform massive parallel calculations. GPU-based algorithms have now made their way into mainstream computing in many fields, see for example, Suchard and Rambaut (2009), Suchard et al. (2010), Lee et al. (2010), Cron and West (2011), Herbei and Kubatko (2013). The work described in this paper has been implemented and tested initially on a NVIDIA Tesla C2075 GPU while the bulk of the computing was done on the OAKLEY cluster at the Ohio Supercomputing Center (`www.osc.edu`), which has 128 Tesla M2070 GPUs.

For this work we require fast, parallel evaluation of the transition matrix $\mathbf{P}(t) = \exp(t\mathbf{Q})$. Due to the complexity of our approach and the size of the problem (a typical gene has ∼300-400 codons), a sequential evaluation/approximation for each of the required matrix exponentials is far too inefficient. The speed gained by distributing this computing aspect to the GPU comes from the ability of the user to pre-specify a computing array of independent threads. Each thread will evaluate a matrix exponential for a given combination of site/branch length/rate matrix. It is well known that exact evaluation of a matrix exponential can only be done in a few particular cases, while in general, an approximation algorithm is required. A suite of algorithms and their performance is discussed in Moler and Van Loan (2003); Golub and van Loan (1996). For this work, we implemented Method 3 described in Moler and Van Loan (2003), see also Algorithm 11.3-1 of Golub and van Loan (1996). Note that, for a matrix A,

$$e^A = \left(e^{A/m}\right)^m \ ,$$

for any integer $m \geq 1$. When m is selected to be a power of 2, the exponential e^A is then obtained through repeated squaring. For the matrix $e^{A/m}$, we use the Padé approximation, see Moler and Van Loan (2003), page 9. This approach is characterized as "the only generally competitive series method" (Moler and Van Loan, 2003) and requires basic matrix algebra (matrix scaling/multiplication), thus it is very suitable for parallel computing. The GPU implementation results in an approximately 40-fold reduction in computation time over a standard CPU application.

Model Comparison

We compare the fit of our proposed model with several models in PAML using the AIC (Burnham and Anderson, 2002). Denoting the maximized likelihood under our new model by $\hat{L}_{MutNSE}$ and the maximized likelihood under the model in PAML by $\hat{L}_{PAML}$, the AIC for each model is computed using

$$AIC = -2\ln(\hat{L}_i) + 2r_i \tag{8}$$

where i refers to either the MutNSE model or one of the models in PAML, and r_i is the number of parameters in model i.

The models we consider are listed in Table 1. In particular, we consider the MutNSE model with four different choices for the codon frequencies ('Fequal', 'F1x4', 'F3x4', and 'Fcodon'), as well as the M0 model in PAML with the same four choices for the codon frequencies. Finally, we consider the mutation-selection model of Yang and Nielsen (2008) as implemented in PAML (FMutSel). We compare the MutNSE model against the corresponding M0 model, as well as against the FMutSel model, resulting in seven separate comparisons (Table 2).

Application to Yeast Protein-Coding Genes

We applied our model to the 104-gene data set of Rokas et al. (2003). We fixed the tree topology to be the ML tree found by these authors (see Figure 1(a)), and then estimated the MLEs of all model parameters (e.g., κ, ω, a and the branch lengths) along this fixed tree. We compared the fit under various models using the AIC, as described above. We also compared the parameter estimates for κ and ω under the new model and under the M0 model.

Because the parameter a in the MutNSE model can be interpreted as a measure of the extent to which selection on codon usage contributes to the overall substitution rate, we hypothesized that a should be correlated with the rate of protein production. To examine this, we obtained protein production rates, ϕ, for all *S. cerevisiae* genes from Yassour et al. (2009). These values indicate the average rate of protein production and were estimated using mRNA abundances (MacKay et al., 2004) and ribosome occupancy on mRNAs (Arava et al., 2003). We assume strong purifying selection on protein production rate of the genes considered here. Thus ϕ values estimated in yeast are used as proxies for ϕ across the phylogeny.

Finally, we wanted to examine whether use of the new model would impact the preferred topology under the maximum likelihood criterion. To examine this, we considered two candidate topologies (Figure 1). Various studies using these data in different modeling frameworks have preferred one or the other of these trees (see, for example, Edwards et al. (2007)). For each gene, we obtained the maximized value of the likelihood under both the MutNSE model and the models in PAML. We counted the number of genes for which a different tree was preferred based on the likelihood under the various models.

Results

The model comparison results for each of the seven comparisons are shown in Table 2. The MutNSE model is preferred at least 85% of the time when it is compared to the corresponding M0 model with a relatively simple model for estimating the codon frequencies (i.e., 'Fequal', 'F1x4', and 'F3x4'). However, when codon frequencies are estimated using the empirical frequencies, the MutNSE model is only preferred about 25% of the time. When the mutation-selection model FMutSel is used, the MutNSE model no longer provides better fit, except for a couple of genes.

Figure 2 shows various relationships between parameter estimates for both the MutNSE and the M0 models in the case when the 'F3x4' estimates of codon frequencies were used (results for the other cases are similar and are not shown here). Figures 2(a) and 2(b) compare the estimates of κ and ω under the two models, with the line indicating equality of the estimates. In both cases, the estimates under the MutNSE model and under the M0 model are highly correlated, with a slight bias toward higher estimates of both parameters under the M0 model. In particular, there are several genes for which the estimates of ω are larger under the M0 model. This may be an important finding, as this parameter is often used as a test for and measure of the strength of either purifying or diversifying selection across a gene.

Figure 2(c) shows that the estimated value of the MutNSE model parameter a is highly correlated with observed protein production rate (ϕ) of yeast genes. This suggests that selection against nonsense errors plays a significant role in affecting the evolutionary rate of highly expressed genes.

Figure 2(d), (e), and (f) examine the relationship between the estimated values of a and of other characteristics of the model, specifically the estimated values of κ, of ω, and the sequence length, respectively. We expect no relationship among these quantities and that is, in fact, what is observed in these plots.

To examine consistency of model preference across genes, we compared the model selected across all genes. Figure 3 shows genes (x-axis) for which the MutNSE model is preferred (indicated with a colored dot) across the various model comparisons (y-axis; height of the points corresponds to the comparison number in Table 2). It is clear that there is consistency across comparisons overall, though there are also differences. In particular, results for the first three comparisons (in which the MutNSE model is preferred the majority of the time) are very consistent, while there is less consistency across comparisons 4 and 5 (in which the MutNSE model is only preferred 25% of the time).

We also compared the MutNSE model to the M0 model with 'F3x4' in terms of which of the two topologies in Figure 1 was preferred under each model using the likelihood value. The MutNSE model had a higher likelihood for the tree in Figure 1 (a) for 42 of the 104 genes, while the M0 had a higher likelihood for only one of the 104 genes. This means that for 41 of the 104 genes, the likelihood criterion would order the two trees in Figure 1 differently under the MutNSE model versus the M0 model. Thus, model choice can impact estimation of the phylogeny.

Finally, we note that because the MutNSE model is site-specific, likelihood computations using the model are non-trivial. The GPU computing machinery used here is crucial to obtaining phylogenetic estimates in reasonable time. Using the OAKLEY Cluster at the Ohio Supercomputer Center, each likelihood evaluation takes under a minute, and full optimization of all model parameters (including branch lengths) along a fixed tree requires between 5 minutes and 2 hours for most genes. We point out that the only step of our implementation that takes advantage of GPU computing is the computation of transition probabilities for each site. Likelihood computation across a tree has also been implemented in a GPU framework (Ayres et al., 2012) and this would speed computations even further.

Discussion

Overall, our results indicate that incorporating selection on synonymous codon usage is an important component of a codon substitution model, as has been noted by others (Yang and Nielsen, 2008; Nielsen et al., 2007; Zhou et al., 2010; Rodrigue et al., 2010). We found that when simple models of codon frequencies were used, our MutNSE model was preferred over the M0 model for the majority of data sets ($> 85\%$ of genes in the yeast data set). However, when empirical codon frequencies were used in both models, the new model was preferred for only about 25% of the genes, and when a model that incorporates both mutation and selection was used (the FMutSel model), our model was generally not preferred using the AIC. This is not completely unexpected, because the FMutSel model is parameterized so that estimates of the selection parameters are obtained empirically, while our MutNSE model incorporates the effect of codon usage via the inclusion of elongation probabilities that are obtained independently and are fixed across genes. However, the set of comparisons made here highlights the importance of realistic models for both codon frequencies and for the process of selection. This is particularly apparent by noting that the MutNSE model preferred the tree in Figure 1(b) over that in Figure 1(a) for 42 of the 104 genes, while the corresponding M0 model only preferred the tree in Figure 1(b) for a single gene. Thus the choice of substitution model can have an important impact on phylogenetic inference.

An important feature of our MutNSE model is that it is able to accurately predict the level of protein production. For genes with high expression (ϕ), we find that selection on codon usage against translation errors is a significant determinant of evolutionary rate (see also Drummond et al. (2006)). This observation is particularly important given that genes used in building phylogenies tend to have a broad phylogenetic breadth, and are highly expressed (Nei et al., 1997, 2000; Eirín-López et al., 2004). Thus, it is essential to develop models of codon substitution that explicitly take into account the effects of selection on synonymous codons and how they change with gene expression. We expect that such models should improve both the reliability and accuracy of the parameters estimated as part of phylogenetic analyses, especially in terms of evaluating whether the ratio of synonymous to non-synonymous substitutions is consistent with stabilizing vs. diversifying selection on the amino acid sequence of a gene.

The model presented here takes into account selection on synonymous codon usage against premature termination. However, patterns of codon usage are also under selection pressures for translation accuracy (Akashi, 1995; Drummond and Wilke, 2008, 2009) and efficiency (Bulmer, 1991; Plotkin and Kudla, 2011; Shah and Gilchrist, 2011). Although the relative importance of these pressures are actively debated in the field, the selective advantage of a synonymous codon for both efficiency and accuracy has been shown to be correlated with its tRNA abundance (Shah and Gilchrist, 2011; Wallace et al., 2013), as has been assumed here. Because we incorporate selection on codon usage in a mechanistic manner, expanding our model to include these additional selective forces is possible in future implementations. Such extensions should not only improve our ability to reconstruct evolutionary relationships based on DNA sequence data, but also potentially extract additional information on key parameters related to the protein translation process itself such as codon-specific nonsense error rates or ribosome pausing times.

Acknowledgements

We thank Ziheng Yang for providing information concerning the parameterization of the models in PAML. This work was supported by the National Science Foundation (DEB0918195 to L.S.K.; MCB1120370 to M.A.G.; and EF0832858 to the National Institute for Mathematical and Biological Synthesis (NIMBioS), which supported M.A.G. and P.S.); the University of Tennessee Knoxville, NIMBioS, and a Tennessee Science Alliance Joint Directed Research and Development grant (to M.A.G.); and the Alfred P. Sloan Foundation, Burroughs Wellcome Fund Career Award, David & Lucille Packard Foundation Fellowship (to J. B. Plotkin and used to support P.S.). We also acknowledge support from the Ohio Supercomputer Center and the Mathematical Biosciences Institute (MBI) at The Ohio State University.

Table 1: Models considered in this study, along with number of parameters estimated. The notation π_i refers to nucleotide frequencies, and involves 3 parameters since $\sum_{i \in \{A,C,G,T\}} = 1$; $\pi_{i,j}$ refers to nucleotide frequencies at each codon position $j = 1, 2, 3$, and involves 9 free parameters; π_J refers to codon frequencies, and involves 60 free parameters, corresponding to the 61 sense codons. In the number of parameters to be estimated listed in the table below, we do not include branch length parameters, since this number will be the same under all models.

Model	Parameters estimated	Total number of parameters
M0, equal frequencies	κ, ω	2
MutNSE, equal frequencies	a, κ, ω	3
M0, F1x4	π_i, κ, ω	5
MutNSE, F1x4	π_i, a, κ, ω	6
M0, F3x4	$\pi_{i,j}$, κ, ω	11
MutNSE, F3x4	$\pi_{i,j}$, a, κ, ω	12
M0, Fcodon	π_J, κ, ω	62
MutNSE, Fcodon	π_J, a, κ, ω	63
FMutSel	π_i, π_J, κ, ω	65

Table 2: Number of yeast genes (of 104 total genes) for which the newly proposed model (MutNSE) was preferred over an existing model. The second two columns list the models being compared (see Table 1) and the fourth column gives the number of yeast genes for which the MutNSE model was preferred using AIC.

			Number of Times Model 1
Comparison	Model 1	Model 2	Was Preferred
1	MutNSE, equal frequencies	M0, equal frequencies	93
2	MutNSE, F1x4	M0, F1x4	91
3	MutNSE, F3x4	M0, F3x4	88
4	MutNSE, F3x4	M0, Fcodon	26
5	MutNSE, Fcodon	M0, Fcodon	27
6	MutNSE, Fcodon	FMutSel	0
7	MutNSE, F3x4	FMutSel	2

Figure Legends

Figure 1: Two phylogenetic trees for the yeast data. The tree in (a) was found by Rokas et al. (2003) to be the ML tree for the concatenated data. The tree in (b) has been proposed by several authors (see, e.g., Edwards et al. (2007)) to be a plausible species-level phylogeny for yeast.

Figure 2: Parameter estimation results for the yeast genes for the MutNSE, F3x4 model and the M0, F3x4 model. (a) Comparison of the values of the parameter κ estimated for the MutNSE and for the M0 model. (b) Comparison of the values of the parameter ω estimated for the MutNSE model and for the M0 model. In both (a) and (b), the line has slope 1 and represents equality of parameter estimates in the two models. (c) Plot of the estimated value of a, which determines the relative importance of codon usage in driving sequence evolution, versus an independent estimate of the rate of protein production, ϕ (see text for details on how the estimates of ϕ were obtained). (d) Plot of the estimated values of a versus the estimated value of κ in the MutNSE model. (e) Plot of the estimated values of a versus the estimated value of ω in the MutNSE model. (f) Plot of the estimated values of a versus sequence length in base pairs (bp).

Figure 3: Comparison of model selection results across genes (x-axis) and across comparisons (y-axis; see Table 2 for comparison numbers). For all comparisons, a colored dot indicates preference for the MutNSE model over the relevant existing model.

References

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

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

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at silent sites in *Drosophila* DNA. *Genetics*, 139:1067–1076.

Arava, Y., Wang, Y. L., Storey, J. D., Liu, C. L., Brown, P. O., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, 100:3889–3894.

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., Rambaut, A., and Suchard, M. A. (2012). BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology*, 61:170173.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129:897–907.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical-theoretic approach*. Springer-Verlag, 2nd edition.

Cron, A. and West, M. (2011). Efficient classification-based relabeling in mixture models. *The American Statistician*, 65:16–20.

Dayhoff, M. and Eck, R. (1968). *A model of evolutionary change in proteins*, pages 33–41.

Dayhoff, M., Eck, R., and Park, C. (1972). *A model of evolutionary change in proteins*, pages 89–99.

Dayhoff, M., Schwarz, R., and Orcutt, B. (1978). *A model of evolutionary change in proteins*, pages 345–352.

Dimmic, M. W., Rest, J. S., Mindell, D. P., and Goldstein, R. A. (2002). rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *Journal of Molecular Evolution*, 55(1):65–73.

Drummond, D. A., Raval, A., and Wilke, C. O. (2006). A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution*, 23(2):327–337.

Drummond, D. A. and Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134:341–352.

Drummond, D. A. and Wilke, C. O. (2009). The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics*, 10:715–724.

Edwards, S. E., Liu, L., and Pearl, D. K. (2007). High resolution species tree without concatenation. *Proceedings of the National Academy of Sciences of the United States of America*, 104(14):5936–5941.

Eirín-López, J. M., González-Tizón, A. M., Martínez, A., and Méndez, J. (2004). Birth-and-death evolution with strong purifying selection in the histone H1 multigene family and the origin of orphon H1 genes. *Molecular Biology and Evolution*, 21(10):1992–2003.

Gilchrist, M. A. (2007). Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Molecular Biology and Evolution*, 24:2362–2373.

Gilchrist, M. A., Qin, H., and Zaretzki, R. (2007). Modeling sage tag formation and its effects on data interpretation within a bayesian framework. *BMC Bioinformatics*, 8:403.

Gilchrist, M. A., Shah, P., and Zaretzki, R. (2009). Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics*, 183:1493–1505.

Gilchrist, M. A. and Wagner, A. (2006). A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *Journal of Theoretical Biology*, 239:417–434.

Goldman, N., Thorne, J. L., and Jones, D. T. (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *Journal of Molecular Biology*, 263:196–208.

Goldman, N., Thorne, J. L., and Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149:445458.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736.

Golub, G. H. and van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins University Press.

Herbei, R. and Kubatko, L. (2013). Monte Carlo estimation of total variation distance of Markov chains on large spaces, with application to phylogenetics. *Statistical Applications in Genetics and Molecular Biology*, 12(1):39–48.

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

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1994). A mutation data matrix for transmembrane proteins. *FEBS Lett.*, 339(3):269–275.

Kosakovsky Pond, S. and Muse, S. V. (2005). Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*, 22(12):2375–2385.

Kosiol, C., Holmes, I., and Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Molecular Biology and Evolution*, 24(7):1464–1479.

Lee, L., Yau, C., Giles, M. B., Doucet, A., and Homes, C. C. (2010). On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 19:769–789.

MacKay, V. L., Li, X. H., Flory, M. R., Turcott, E., Law, G. L., Serikawa, K. A., Xu, X. L., Lee, H., Goodlett, D. R., Aebersold, R., Zhao, L. P., and Morris, D. R. (2004). Gene expression analyzed by high-resolution state array analysis and quantitative proteomics - response of yeast to mating pheromone. *Molecular & Cellular Proteomics*, 3:478–489.

Massingham, T. and Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, 169:1753–1762.

Mayrose, I., Doron-Faigenboim, A., Bacharach, E., and Pupko, T. (2007). Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics*, 23(13):i319–27.

Moler, C. and Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49.

Nei, M., Gu, X., and Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci USA*, 94(15):7799–7806.

Nei, M., Rogozin, I. B., and Piontkivska, H. (2000). Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc Natl Acad Sci USA*, 97(20):10866–10871.

Nielsen, R., DuMont, V. L. B., Hubisz, M. J., and Aquadro, C. F. (2007). Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. *Mol Biol Evol*, 24:228235.

Plotkin, J. B. and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12:32–42. 10.1038/nrg2899.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, 107(10):4629–4634.

Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425:798–804.

Shah, P. and Gilchrist, M. A. (2011). Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25):10231–10236.

Suchard, M., Wang, Q., Chan, C., Frelinger, J., Cron, A., and West, M. (2010). Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics*, 19:419–438.

Suchard, M. A. and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics*, 25:1370–1376.

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

Wallace, E. W. J., Airoldi, E. M., and Drummond, D. A. (2013). Estimating selection on synonymous codon usage from noisy experimental data. *Molecular Biology and Evolution*, 30(6):1438?1453.

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

Wong, W., Yang, Z., Goldman, N., and Nielsen, R. (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, 168:1041–1051.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39:105–111.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS*, 13(5):555–556.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press.

Yang, Z. and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, 15(12):496–503.

Yang, Z. and Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, 46:409–418.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, 19:908–917.

Yang, Z. and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, 25(3):568–579.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449.

Yang, Z., Nielsen, R., and Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution*, 15:1600–1611.

Yassour, M., e. a. (2009). Ab initio construction of a eukaryotic transcriptome by massively parallel mrna sequencing. *Proc Natl Acad Sci U S A*, 106(9):3264–3269.

Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution*, 39:315–329.

Zhou, T., Gu, W., and Wilke, C. O. (2010). Detecting positive and purifying selection at synonymous sites in yeast and worm. *Mol Biol Evol*, 27:19121922.

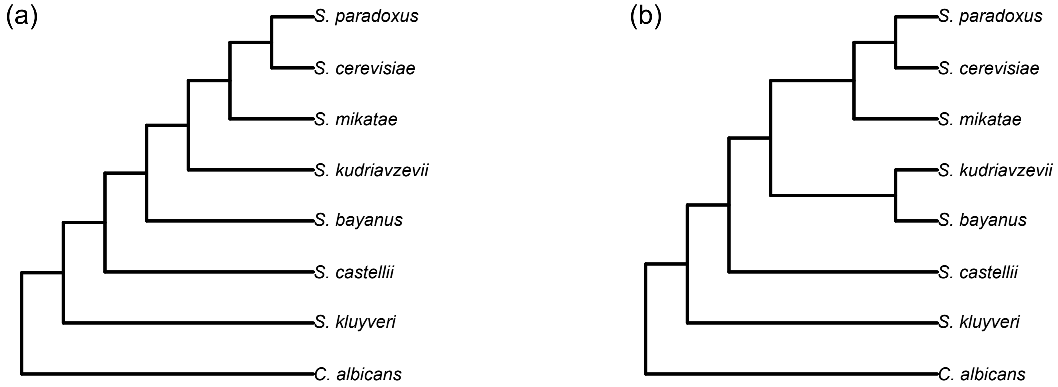

Figure 1: Two phylogenetic trees for the yeast data. The tree in (a) was found by Rokas et al. (2003) to be the ML tree for the concatenated data. The tree in (b) has been proposed by several authors (see, e.g., Edwards et al. (2007)) to be a plausible species-level phylogeny for yeast.

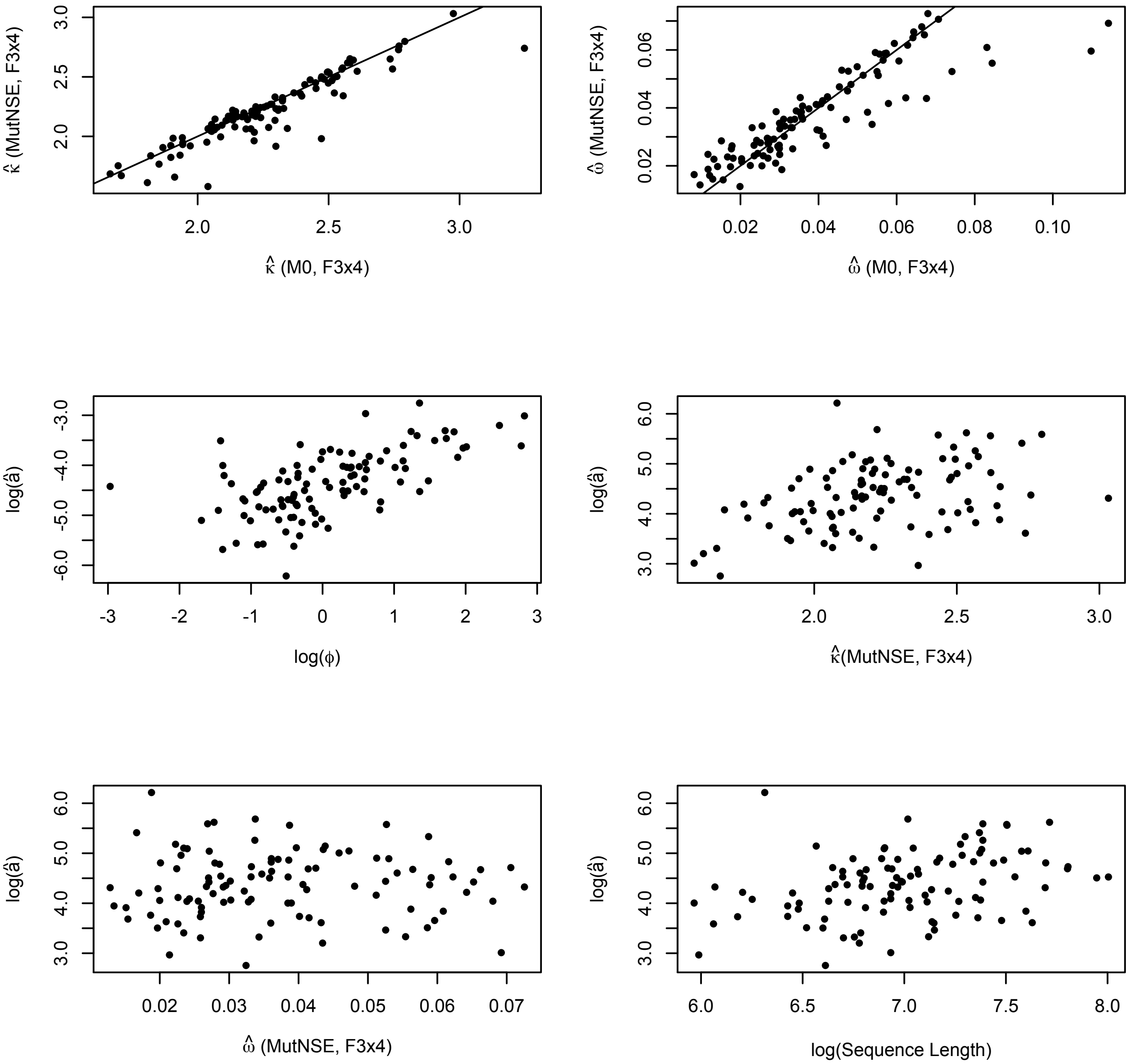

Figure 2: Parameter estimation results for the yeast genes for the MutNSE, F3x4 model and the M0, F3x4 model. (a) Comparison of the values of the parameter κ estimated for the MutNSE and for the M0 model. (b) Comparison of the values of the parameter ω estimated for the MutNSE model and for the M0 model. In both (a) and (b), the line has slope 1 and represents equality of parameter estimates in the two models. (c) Plot of the estimated value of a, which determines the relative importance of codon usage in driving sequence evolution, versus an independent estimate of the rate of protein production, ϕ (see text for details on how the estimates of ϕ were obtained). (d) Plot of the estimated values of a versus the estimated value of κ in the MutNSE model. (e) Plot of the estimated values of a versus the estimated value of ω in the MutNSE model. (f) Plot of the estimated values of a versus sequence length in base pairs (bp).

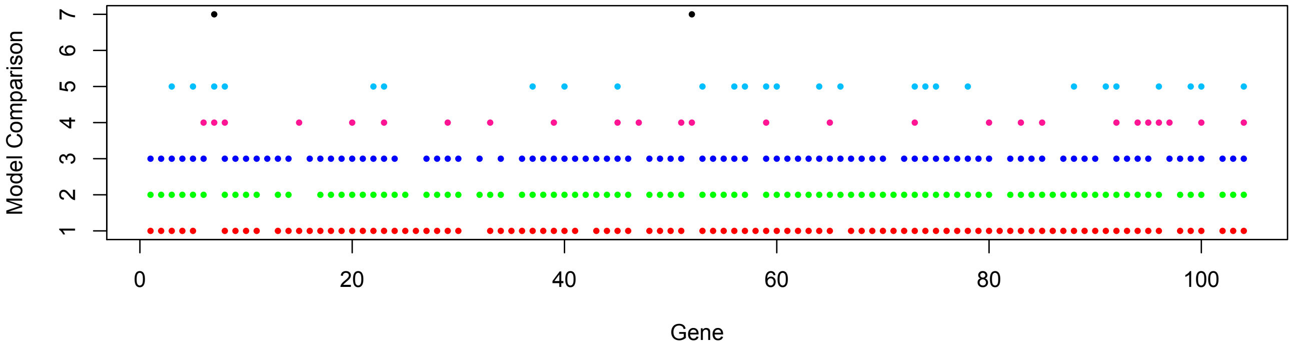

Figure 3: Comparison of model selection results across genes (x-axis) and across comparisons (y-axis; see Table 2 for comparison numbers). For all comparisons, a colored dot indicates preference for the MutNSE model over the relevant existing model.