# Annotated Bibliography - Inferring Selection

Jenna Baughman & Betsabé Castro

IB 290 - Dec. 6, 2017

**Goldman, N., & Yang, Z. (1994). A Codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol. 11*(5):725-736. https://doi.org/10.1093/oxfordjournals.molbev.a040153**

This paper introduces for the first time the codon-based model for evolution of protein-coding DNA using a continuous-time Markov process and a maximum likelihood approach. Contrary to previous models like the nucleotide-based model, the codon models allows the use of parameters for transition/transversion rates bias, codon usage bias, and selective restraints related to the physicochemical distances between amino acids in the codons.

The codon model seeks to build more realistic models where previously ignored effects (e.g lack of independence among neighboring sites in a codon triplet; evolutionary rates differences between codon positions) and significant biological processes (e.g. amino acid differences, transition/transversion rates, variability in a gene indicated via synonymous/non-synonymous rates ) are incorporated. In addition, the codon model produces a better fit to coding sequence data than previous models based on Goldman's test (1993). The model also allows for separating the effects of within and between codons. Two data sets are used to test this model where (1) mammalian $\alpha$ - and $\beta$ -globin genes (e.g. primate, marsupial, rodent, lagomorph, and artiodactyl) and ADP-glucose pyrophosphorylase genes (e.g. wheat, potato, rice, maize). Goldman & Yang point out that their model seems to produce more accurate estimation of phylogenetic relationships.

Although not explicitly stated, a major disadvantage of the codon model to this date was that it averages effects across all codon sites within the sequence and for all lineages, thus it could mask specific selective forces for sites and/or branches within a phylogeny.

**Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics 148*: 929–936. http://www.genetics.org/content/148/3/929.short**

This paper builds on Goldman & Yang's codon model (1994) to introduce the first site-specific codon model, which allows for variable selection intensity among sites. They introduce two flavors of their site specific codon model: neutral and positive-selection. The neutral model assumes two categories of sites (e.g. neutral and conserved sites) at which amino acid replacements can be either neutral or deleterious. The positive model is nested within the neutral model but assumes an additional category of positively selected sites at which nonsynonymous substitutions have a higher rate of occurrence than synonymous substitutions. The positive-selection model is also useful for identifying target sites experiencing positive selection. They implement their site specific codon model by using a maximum likelihood framework and likelihood ratio tests of neutral evolution. Nielsen & Yang apply their model to a data set of V3 region of the HIV-1 envelope gene, that were sampled and sequenced across different years for the same individual. Their results point towards rejecting their neutral model over their positive-selection model and show strong support for variable selection intensity among amino acid sites. Furthermore, prior to previous models that relied to $d_N/d_S$ ratio counts and the original codon model, this site-specific codon model can provide a more in depth look at the evolutionary process of protein-coding sequences. A slight shortcoming is that there is no reference yet to include lineage specific variation in these models of selection.

**Yang, Z. (1998). Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution.** ***Mol. Biol. Evol.*** **15(5):568–573. https://doi-org.libproxy.berkeley.edu/10.1093/oxfordjournals.molbev.a025957**

Here Yang introduces the branch-specific codon model as a follow-up to Goldman & Yang (1994) codon model. The motivation for developing this model was to be able to compare a lineage or a group of lineages that might have a different selective history than the other branches. The model allows for comparison of two groups of lineages, or one lineage vs. the rest, and is user defined *a priori.* Thus, you have to have an explicit hypothesis of which lineage(s) might be under different selection than the rest and use this model to test it. Since this is a modification of the Goldman & Yang (1994) model with an additional parameter, in other words that one is nested within this one, you can compare the original codon model vs. this branch model using a likelihood ratio test (LRT). A problem with this model is that it still averages across all codon sites within the sequence (for each lineage set) and so is quite prone to failing to detect positive selection. This is because it seems that positive selection often occurs on a few sites of a gene against a backdrop of purifying selection on the rest of the sites. The average $\omega$ of the gene, therefore, will come out looking negative/purifying or neutral. Still, this model opened up much more interesting question possibilities, such as whether selection on a particular gene or genes occurred differently on the branch leading to hominoids vs. other primates, which they tested with lysozyme genes in this paper.

***Yang, Z., & Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.** ***Mol Biol Evol*****,** ***19*****(6), 908–917. https://doi.org/10.1093/oxfordjournals.molbev.a004148**

Yang & Nielsen introduce the branch-site codon model which simply combines previously developed branch and site models (Nielsen & Yang, 1998; Yang, 1998) by allowing the $d_N/d_S$ rate ratio ($k_a/k_s$ or $\omega$, omega) to vary both among sites and among lineages. This is an important advancement because branch models alone appear lack power to detect a lot of positive selection present in data sets/evolution and site models would fail to detect in certain circumstances. Specifically, if positive selection occurs a few time points and at a few codon sites against a background of purifying selection, previous models would likely fail to detect it due to averaging $\omega$ over sites and/or lineages. This model is probably almost exclusively used now, and can anyway be tested and compared to previous two versions (site and branch independently), using a likelihood ratio test. In fact, in this paper the authors did just that on three data sets and found improved ability to detect selection in most cases. One exception is that by splitting the lineages you can sort of reduce total number of nucleotide substitutions in each part of the tree which can reduce statistical power in inferring selection on those branches.

**Bamshad, M., & Wooding, S. P. (2003). Signatures of natural selection in the human genome. *Nature Genetics*, *4*, 99–111. https://doi.org/10.1038/nrg999**

This a good review to go over methods on how and where to look for signatures of selection at the genome level. Using the human genome as an example, Bamshad and Nielsen highlight different types of selection (e.g. positive, neutral, purifying, balancing) and their effects on variation in DNA. They proceed to explain relative advantages and disadvantages of different strategies to detect signatures of selection at a given locus. In addition, they outline proposed methods for scanning genomes for evidence of selection. Last, they discuss issues associated with identifying signatures of selection and making inferences about the nature of selective process.

**Ree, R., Citerne, H. L., Lavin, M., Cronk, Q. C. B. (2004). Heterogeneous selection on LEGCYC paralogs in relation to flower morphology and the phylogeny of Lupinus (leguminosae). *Mol. Biol. Evol. 21*(2):321–331. https://doi.org/10.1093/molbev/msh022**

In this application paper, Ree and colleagues conducted an analysis of molecular evolution for two paralog genes of *LEGCYC* group I (e.g. *LEGCYC1A* and *LEGCYC1B*) in New World *Lupinus* species in relation to flower morphology. Here they reveal varied history for site-specific and lineage-specific evolutionary rates, as well as selection for both within and between loci. Parting from the branch-site model proposed earlier by Yang and Nielsen (2002), they introduce a modification using an additional $\omega$ parameter to account for effects of greater positive and greater purifying selection at different codon sites along a branch. For this paper, Ree and colleagues were interested in (1) gaining a more fined tuned look at the molecular evolution of genes that are likely to be important in flower development and (2) using *LEGCYC* to reconstruct the phylogeny for *Lupinus* species using rapidly evolving nuclear genes loci that may provide more phylogenetic signal than molecular makers traditionally used for phylogeny estimation (ITS). Their phylogenetic estimates for non-synonymous/synonymous substitution rate ratio ($\omega$) find support for positive selection ($\omega > 1$) along the *L. densiflorus* lineage at some codon sites of one of the paralogs genes (*LEGCYC1B*) while finding greater purifying selection ($\omega < 1$) at some sites in the other paralog gene (*LEGCYC1A*). Results support that *LEGCYC1A* might be evolving faster than *LEGCYC1B,* while both paralogs are evolving faster than

ITS. These higher rates of evolution and congruent phylogenetic signal seem promising to use as markers for phylogenetic reconstruction of low taxonomic levels in the family Genisteae.

***Kosakovsky Pond, S. L., Frost, S. D. W., Grossman, Z., Gravenor, M. B., Richman, D. D., & Brown, A. J. L. (2006). Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Computational Biology*, *2*(6), 0530–0538. https://doi.org/10.1371/journal.pcbi.0020062**

This is an interesting application of the branch-site codon model. The authors investigate selective pressures at different time scales and among different populations of HIV. Specifically, they looked at two HIV genes from two genetically distinct human host populations. Using slight modifications of the original branch-site model, they find that patterns of selection are different in the two populations and, interestingly, positive is higher on tip branches than deep branches. This result, along with comparison of specific amino acid transitions, suggests that positive selection for new forms may be beneficial in the short term but deleterious in the long term. This makes sense in terms of HIV biology: within a host, changes in HIV can help it evade the host's immune system so would therefore be positively selected. However, some of those changes may turn out to be deleterious during transmission to another host and are therefore likely to experience more purifying selection.

**Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., Pupko, T. (2007). Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Res. 35,* W506–W511. https://doi.org/10.1093/nar/gkm382**

This paper introduces a web-based program that takes alignments and calculates the $d_N/d_S$ rate ratio ($\omega$, omega) on-the-fly at each site, showing whether each is under positive, neutral, or purifying selection. They also implemented a variety of evolutionary site-specific codon models to test hypotheses either with a neighbor-joining tree, built by Selecton, or with a user-supplied tree. Though they note that more accurate parameter estimates result from using a more accurate tree, they also argue that, based on comparison studies, use of neighbor-joining trees do not alter inference selection that much in most cases and is very fast. Still, if you have other means of building a tree and can supply it it is preferred--the program can infer branch lengths. An interesting model developed by this group for this tool is one that takes into account the biochemistry and physical properties of the amino acids in the alignments, even showing a predicted 3D structure of the protein. Selecton actually incorporates a model that accounts for the differences between amino acid replacement rates. In other words, no all non-synonymous substitutions occur at the same rate (ie, are equally likely) and one of the models here incorporates the rate differences. They use a 61x61 codon rate matrix when $k_a$ (non-synonymous substitution rate) is inferred the replacement probabilities between the amino acids are taken into account. Thus, in this model $\omega$ is not directly equivalent to $k_a/k_s$.

Finally, Selecton has a model comparison tool which will perform either a likelihood ratio test (LRT) for nested models or the second order Akaike Information Criterion ($AIC_C$) if they are not.

**Rubinstein, N. D., Doron-Faigenboim, A., Mayrose, I., & Pupko, T. (2011). Evolutionary models accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection research article. *Mol Biol Evol*, *28*(12), 3297–3308. https://doi.org/10.1093/molbev/msr162**

This paper introduces new models that address something that has apparently been becoming a bigger and bigger problem with codon models: the fact that synonymous substitutions are not always 'silent' and may be under selection, too. This  is one of the "layers of selection" they refer to in their title. The new model, then, is a version of the codon model that allows for variation in $k_s$ (synonymous substitution rate) as well as $k_a$ (non-synonymous substitution rate). Specifically, they separate selection at what they call the "DNA and RNA level" vs. at the "protein level" in two nested models. They explain that by comparing these models it is possible to identify positive selection (at the protein level) while accounting for variability in the nucleotide substitution rate. They test these claims with simulated data and furthermore analyze mammalian genomes to show that baseline nucleotide substitution rate is highly variable and using $k_a/k_s$ without accounting for variability in synonymous substitution rate leads to erroneous inferences of selection.

***Wilson, D. J., Hernandez, R. D., Andolfatto, P., & Przeworski, M. (2011). A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genetics*, *7*(12), 1–13. https://doi.org/10.1371/journal.pgen.1002395**

This paper represents an interesting and novel approach to combine population genetics statistics with phylogenetic models of evolution to infer selection both within and among lineages, accounting for variation both spatially in the genome and temporally. While it is not modeling the full coalescent process with selection, it is attempting to account for mutation, drift, and selection within polymorphic genes in a species or population. Specifically, the within-lineage inference of selection looks at allele frequencies of polymorphic genes and the between-lineage angle is a branch-site codon model. The key parameter of this new model is the population-scaled selection coefficient, $\gamma$, which is the product of the ploidy, the effective population size, and the fitness advantage relative to the ancestral allele. Over long time scales, the phylogenetic substitution rate of this model converges to the Nielsen & Yang site-specific codon model, allowing for a connection between these two time scales of selection. The idea behind comparing the population allele frequencies and relative rate of non-synonymous to synonymous mutations in the population ($p_n/p_s$) to the $k_a/k_s$ of fixed differences between lineages is based in the McDonald Kreitman test to detect adaptation where either measure alone might not be able to. The spatial analysis used a Bayesian sliding window approach to understand the spatial scale at which selection is acting in different genes and/or different lineages. The authors tested these models on *Drosophila* polymorphism and divergence data for 100 X-linked genes.

**Pentony, M. M., Winters, P., Penfold-Brown, D., Drew, K., Narechania, A., DeSalle, R., Bonneau, R., Purugganan, M. D. (2012). The Plant proteome folding project: Structure and positive selection in plant protein families. *Genome Biology and Evolution*, 4(3), 360–371. https://doi.org/10.1093/gbe/evs015**

This paper applies the site-specific codon model to infer selection in 2,120 plant gene families within 5 species (*Arabidopsis thaliana, Oryza sativa, Populus trichocarpa, Sorghum bicolour, and* and *Vitis vinifera*) and identified the specific sites that were positively selected. What makes this particularly interesting, though, is they compare patterns of selection with structure of the proteins. To do this they developed a model for predicting protein structure and used it to develop a massive proteome database with over 15,000 gene families from these 5 species. They then placed the positively-selected sites in a structural context by visualizing them in predicted 3D protein structure.

**Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). FUBAR : A Fast , Unconstrained Bayesian AppRoximation for inferring selection. *Mol Biol Evol*, *30*(5), 1196–1205. https://doi.org/10.1093/molbev/mst030**

This paper introduces a new site-specific codon model that attempts to improve upon previous models in two main ways: (1) removing site class constraints and (2) increasing speed. The new model, called FUBAR, is a hierarchical Bayesian approximation MCMC approach and the authors test it against fixed effects likelihood (FEL) and random effects likelihood (REL) implementations in HyPhy with simulated data. Where as previous models use a few discrete categories of $\omega$ to approximate the inherently continuous distribution of the actual ratio of synonymous and non-synonymous substitutions, FUBAR uses a much denser grid of values chosen *a priori,* then using MCMC to sample weights at each point. This is an approximation of a true Bayesian posterior probability distribution uses these computational shortcuts to speed up inference of selection, especially in large data sets. The key change is that it precomputes the (many) conditional likelihoods arranged on the *a priori* grid selection of values for the components of $\omega$ instead of shifting location of the parameter classes during optimization depending on the data.

*** Assigned papers**