ASSESSING AND IMPROVING THE RELIABILITY OF MODELS OF MOLECULAR EVOLUTION

by

Joseph Mingrone

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
November 2021

Table of Contents

List of Tables

List of Figures

Abstract

Selection pressure that has acted upon proteins and amino acids can be revealed through a ratio of nonsynonymous to synonymous codon substitutions (ω). Markov models of codon substitution fitted to an alignment of homologous protein-coding DNA sequences estimate an ω mixing distribution in a likelihood framework in order to detect positive selection. Publications describing codon substitution model implementations have nearly $10,000$ citations from research in a variety of fields, from vaccine design to mammalian physiology.

The site models of codon substitution used to detect positive selection at amino acids sites first use a pre-screening likelihood ratio (LR) test for positive selection at the level of the protein. Due to statistical irregularity, the large-sample distributions of the LR statistic are often not justified and thresholds determined from the distributions can give larger than expected type I error rates. Presented in Chapter 2 is a modified LR test for protein-level selection. The modified LR test is shown to restore statistical regularity to give tractable LR statistic distributions. No matter the parameter settings of the underlying null hypothesis, the modified LR gives approximately correct type I error probabilities when the number of codon sites is not too small. Simulation results show that type I error rates are closer to expectations without loss of power. Under certain data-generation settings, very different estimated ω distributions can give nearly identical site likelihoods when the number of taxa and the total tree length are not large enough.

After the pre-screening LR test, most codon substitution models use an empirical Bayes approach to detect positive selection at individual amino acid sites. After model parameters are estimated via maximum likelihood, they are passed to Bayes formula to compute the posterior probability that a site evolved under positive selection. A difficulty with the empirical Bayes approach is that estimates with large errors can negatively impact classification. Presented in Chapter 3 is a new technique called smoothed bootstrap aggregation (SBA) that uses bootstrapping and kernel smoothing to accommodate uncertainty in the estimates. Simulation results show that SBA balances accuracy and power at least as well as Bayes empirical Bayes (BEB), and when parameter estimates are unstable, the performance gap between BEB and SBA can widen in favour of SBA.

Branch-site models of codon substitution, like the site models, can detect positive selection at a subset of amino acid sites. Unlike the site models however, the

branch-site pre-screening LR test limits positive selection to prespecified branches on the phylogeny. Chapter 4 includes new simulation studies, which show limitations to these widely used models. The branch-site LR distributions under the null hypothesis are sometimes poorly approximated by those predicted by theory and can vary heavily according to factors such as the branches considered for positive selection and irregularity of certain parameter estimates. Moreover, false positives are shown to be common when positive selection has occurred in the tree but not along on the prespecified branches.

List of Abbreviations and Symbols Used

ω the ratio of nonsynonymous to synonymous codon substitutions.

BEB Bayes empirical Bayes.

CDF cumulative distribution function.

DNA deoxyribonucleic acid.

JC69 Jukes and Cantor model.

KL Kullback-Leibler divergence.

LR likelihood ratio.

ML maximum likelihood.

MLE maximum likelihood estimate.

mRNA messenger RNA.

NEB Naïve empirical Bayes.

RNA ribonucleic acid.

ROC receiver operator characteristic.

SBA smoothed bootstrap aggregation.

tRNA transfer RNA.

Acknowledgements

I struggle to find the words to express my gratitude to Edward Susko and Joe Bielawski for remarkable supervision and seemingly infinite patience. I acknowledge Ed's contribution to sections 6.2 and 6.4. Thank you to Andrew Roger, Chris Field, and Stéphane Guindon for insightful review of the manuscript. Thank you to Stuart Carson, Lingyun (Peter) Ye, and Edward Reddick for engaging conversations about statistics, science, and life.

I am grateful for the support of my parents, my wife Mary, and my friend Filipp.

Chapter 1

Introduction

Science must often be conducted without direct observation of that which we wish to explain. The electron is hidden from the chemist, the cosmologist formulates theories about the genesis of the universe based clues that are billions of years old, and so on. Likewise, as the evolution of the heritable characteristics of biological populations has occurred over billions of years and continues at a pace that spans generations, its direct observation is typically not possible. Whether it be by studying the fossil record or comparing the physiology of extant populations, when studying evolution, one must also use inference techniques to understand the processes that gave rise to biological populations. Molecular evolution is a subdiscipline of evolutionary biology in which statistical models and computational algorithms are used to make inferences about evolutionary processes. The field is termed *molecular* evolution because the questions are related to organic molecules such as proteins and amino acids using the molecules that store the genetic information for all life, nucleic acids. The topic of this thesis is molecular protein evolution and the aims are twofold. First, I aim to describe strengths and limitations of some commonly used models of molecular evolution. Second, I build upon some of these models in order to improve the reliability of detection of positive selection at the level of proteins and amino acids.

Access to the data that underpins the field, genetic sequence data, became available with sequencing techniques developed in the 1970s (Gilbert and Maxam, 1973; Sanger et al., 1977) and culminated in the Human Genome Project (Watson, 1990). The project to sequence the entire human genome and to identify all its genes was completed in 2003 after 13 years and a multi-billion dollar budget (Hood and Rowen,

2013). Today, faster and cheaper sequencing techniques put whole genome sequencing within reach of small laboratories and there is an abundance of genetic data to study (Goodwin et al., 2016). With more data and increasingly sophisticated models, the computational demands to model molecular evolution have increased along with the data. These demands are being met by exponential growth in computing power that has lasted for over half a century (Mack, 2011). With an abundance of new data and access to powerful tools, it is an exciting time to study the evolutionary history of our world's remarkable biological diversity. Along with the excitement, many also feel a sense of urgency to study the history of life (e.g., Forest et al., 2015). The urgency is due to human activities that have altered the land, oceans, and atmosphere so profoundly that life is being re-ordered in ways not seen for millions of years (Lewis and Maslin, 2015). Our extraction of resources, direct harvesting of species, fragmentation of habitats, introduction of non-native species, and spreading of pathogens have possibly hastened the sixth mass extinction (Barnosky et al., 2011) and led to proposals for a new human-induced geochronological epoch called the Anthropocene (Crutzen, 2006).

1.1 Beyond The Modern Synthesis

In the mid-nineteenth century, foundational ideas were described about biological inheritance and the evolution of the heritable characteristics in biological populations. Darwin presented compelling evidence of adaptive evolution, i.e., heritable traits which increase reproductive success become more common in populations. Through his experiments with pea plants, Mendel showed that phenotype can be determined by the inheritance of discrete trait units, which we now know are variants of genes called alleles. It wasn't until a half century later when these ideas were reconciled in what has been referred to as the *Modern Synthesis* (Huxley et al., 1942). Key to the development of the *Modern Synthesis* was the new field of population genetics. The development and application of statistical models of evolution to biological population data helped to further understand the forces that drive changes in the allele frequencies in biological populations, i.e., evolution: mutation, gene flow, non-random mating, stochastic factors due to finite population size, and adaptive evolution (Fisher, 1923, 1931; Haldane, 1927; Wright, 1931, 1942; Kimura, 1957, 1962). In this thesis, the primary focus is on the stochastic forces called random genetic drift

and adaptive evolution. More specifically, the aim is to detect evidence of adaptive evolution in a background of random genetic drift.

1.2 A Brief Introduction to Molecular Evolution

1.2.1 Biochemical Fundamentals

Proteins are organic macromolecules that are a fundamental component of life. They participate in nearly all cellular processes from catalyzing chemical reactions to transporting molecules. They effect muscle contraction, form various support structures, and can act as toxins. All proteins are composed of one or more long, linear chains containing different forms of a molecular unit or monomer called an amino acid. The ordering of amino acids in a chain determines how a protein folds into a functioning three-dimensional structure (Anfinsen, 1972).

The information about the precise order of a protein's amino acids is stored in another class of biological macromolecule called nucleic acids. Both types of nucleic acids involved in protein synthesis, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), are composed of chains of monomers called nucleotides. Each nucleotide contains a 5-carbon sugar, a phostphate group, and a nitrogen-containing base. Linear chains of nucleotides are formed via bonds between the phostphate of one nucleotide and the sugar of another to form a sugar-phosphate backbone. The information stored in nucleic acids is encoded by the ordering of four different forms of nucleotides in the chain with each form having a different nitrogen-containing base. RNA molecules are composed, most often, of single chains of nucleotides containing four different bases: adenine (A), uracil (U), guanine (G), or cytosine (C). DNA molecules are composed of two nucleotide chains that are bonded via specific nucleotide base pairings, C with G and A with thymine (T), the DNA analog of RNA's U. The two linked nucleotide chains of DNA form the well known double-helix structure.

The central dogma of molecular biology originally conveyed the idea that once the protein-building information contained in nucleic acids was transferred to a protein, that information could no longer be recovered from the protein. Nowadays, the central dogma often refers to a more detailed flow of the information, i.e., organisms replicate DNA, transcribe DNA to RNA, and translate RNA to protein. During transcription, it is the protein-encoding unit of DNA, the gene, that is used as a

template to synthesize single-stranded RNA called messenger RNA (mRNA). During translation, three-nucleotide sequences within the mRNA called codons are bound to by the complementary anti-codon of transfer RNA (tRNA). The tRNA molecules continue to bind to codons along the length of mRNA, each time carrying a particular amino acid to transfer to an elongating polypeptide chain that will become a protein.

1.2.2 The Genetic Code

The genetic code refers to the mapping of codons to amino acids during protein synthesis. The code (table 1.1) was believed to be universal in that the same codon to amino acid mappings were always employed by all organisms, however some exceptions have been discovered. For example, there are some nonstandard codons in vertebrate mitochondrial DNA, bacteria, and the nuclear genes of protozoans. Aside from the codon to amino acid mappings there are two types of special codons. In the standard genetic code, the codon ATG is referred to as a start codon, because it signals cellular machinery to start reading a gene for translation and also to begin the polypeptide chain with the amino acid Methionine. No tRNA molecules have anti-codons for three codons, TAA, TAG, and TGA. These three codons are called stop codons, because they signal the end of the polypeptide chain.

1.2.3 Mutation, Fixation, and Selection

Suppose all individuals in a population carry the same version of a gene, wild-type allele A, when a heritable error occurs in the gene of one individual introducing mutant allele a into the population. It is generally rare for errors called mutations to occur in genes (Nachman and Crowell, 2000), however when heritable mutations do occur, they provide the ultimate source of variation for evolution. Mutations occur in different forms such as insertions, deletions, or substitutions of single nucleotides to whole chromosomes. However, for the models discussed here, only single-nucleotide changes are considered such as the one shown in table 1.2 where a nucleotide substitution changes codon GAG in wild-type allele A to GTG to create mutant allele a. Because GAG and GTG encode different amino acids, Glutamate and Valine in the universal genetic code, such a nucleotide substitution will cause a change in the protein.

In the absence of mechanisms that direct evolution such as selection, the allele composition of the next generations can be considered the result of random sampling

First Codon Position		Second Codon Position: T		Second Codon Position: C		Second Codon Position: A		Second Codon Position: G	
	T	TTT	F	TCT	S	TAT	Y	TGT	C
		TTC	F	TCC	S	TAC	Y	TGC	C
		TTA	L	TCA	S	TAA	*STOP*	TGA	*STOP*
		TTG	L	TCG	S	TAG	*STOP*	TGG	W
	C	CTT	L	CCT	P	CAT	H	CGT	R
		CTC	L	CCC	P	CAC	H	CGC	R
		CTA	L	CCA	P	CAA	Q	CGA	R
		CTG	L	CCG	P	CAG	Q	CGG	R
	A	ATT	I	ACT	T	AAT	N	AGT	S
		ATC	I	ACC	T	AAC	N	AGC	S
		ATA	I	ACA	T	AAA	K	AGA	R
		ATG	M/*START*	ACG	T	AAG	K	AGG	R
	G	GTT	V	GCT	A	GAT	D	GGT	G
		GTC	V	GCC	A	GAC	D	GGC	G
		GTA	V	GCA	A	GAA	E	GGA	G
		GTG	V	GCG	A	GAG	E	GGG	G

Table 1.1: The Standard Genetic Code. Table rows represent the first nucleotide position, columns the second position, and cell lines the third position. The codon CAG, which codes for the amino acid Glutamine (Q), is found in the second row (C), third column (A), and the fourth line of the (2,3) cell. The codon ATG is both a *START* codon and codes for the amino acid Methionine (M). The codons TAA, TAG, and TGA are *STOP* codons. The single letter amino acid abbreviations are: A Alanine, B Asparagine, C Cysteine, D Aspartate, E Glutamate, F Phenylalanine, G Glycine, H Histidine, I Isoleucine, K Lysine, L Leucine, M Methionine, N Asparagine, P Proline, Q Glutamine, R Arginine, S Serine, T Threonine, V Valine, W Tryptophan, Y Tyrosine, and Z Glutamine.

of alleles from the current generation. Thus, whether the frequency of mutant allele a in the population eventually goes to 0 (elimination) or to 1 (fixation) is determined by random genetic drift, i.e., chance. Selection is one mechanism of evolution that causes sampling of alleles for subsequent generations to be non-random. For example, when individuals carrying mutant allele a have reduced fitness to pass on the allele, relative to carriers of A, purifying selection acts to reduce the frequency of a in the population. When individuals carrying a have increased fitness, positive selection makes fixation of a more probable.

The time between the occurrence of a new mutation and its fixation can vary,

Wild-type Allele A									
ATG	GTG	CAC	CTG	ACT	CCT	**GAG**	GAG	AAG	TCT
Start	Val	His	Leu	Thr	Pro	**Glu**	Glu	Lys	Ser
Mutant allele a									
ATG	GTG	CAC	CTG	ACT	CCT	**GTG**	GAG	AAG	TCT
Start	Val	His	Leu	Thr	Pro	**Val**	Glu	Lys	Ser

Table 1.2: The first ten codons of a gene sequence and their associated amino acids. At the top is wild-type allele, A. Mutant allele a is a copy of A, except that a single-nucleotide mutation changed codon GAG to GTG, substituting the amino acid Glutamate with Valine in the protein.

but it is generally very small relative to the overall time considered in the models of evolution discussed here, i.e., only evolution over longer time scales (macro-evolution) is considered. Thus, polymorphisms, two or more concurrently existing alleles in a population, are ignored and the genetic data for a taxon is considered representative of the population.

1.3 Modelling Molecular Evolution

1.3.1 The Data

The data for the models of evolution considered in this thesis are nucleotide sequences from protein coding genes and a bifurcating phylogenetic tree. The order of the nucleotides for some number of homologous taxonomic units (taxa) are either obtained directly using genetic sequencing technology or indirectly from a genetic database such as GenBank (Benson et al., 2012). The sequences are arranged so that the data for each taxon is a row in a data matrix, $\boldsymbol{X}$. A goal of the alignment is to arrange homologous characters, either nucleotides or codons, in the columns of $\boldsymbol{X}$. With short and highly conserved sequences this is a straightforward task, but software-implemented alignment algorithms (e.g., Altschul et al., 1990; Buchfink et al., 2021) are usually required because different accumulated mutations in each of the sequences, such as deleted or inserted nucleotides, make manual alignment impractical. For all models considered in this thesis, the aligned data at each site $\boldsymbol{x}_h$, a column in $\boldsymbol{X}$, is assumed to be an independent observational unit. An example four-taxon alignment is shown in table 1.3.

Phylogenetic trees, such as those shown in figure 1.1, are structures that represent

the inferred evolutionary relationships among taxa. Their two components are nodes and branches. Each taxon is represented by a node, labelled 0 to 6 in figure 1.1, and each evolutionary path between taxa is represented by a branch. The branch lengths, labelled t_1 through t_6 in figure 1.1, typically represent the genetic distance between adjoining taxa. Rooted trees have a unique internal node called the root node, which is interpreted as the common ancestor of all other nodes. The root node is labelled 0 in subfigure 1.1a and is absent from the unrooted tree in subfigure 1.1b as unrooted trees do not have a root node and do not define the direction of evolution. When the tree is bifurcating, all other internal nodes have two branches that each connect to a descendant node and a third branch that connects to an ancestral node. Sequence data in the rows of $\boldsymbol{X}$ are only observed for external nodes (also referred to as tip or leaf nodes). For the models considered in this thesis, evolution along a branch of the tree is assumed to be independent of evolution along any other branch, conditional upon some value for any unknown states at the ends of the branch.

With the assumptions of independence across both sites and branches of the tree, the unit of evolution is simplified to substitution between states along a branch. In addition, a property called time reversibility means, roughly, that the probability of the data at a site is equal regardless whether either end of the branch is considered the ancestor. Time reversibility will be discussed in further detail below.

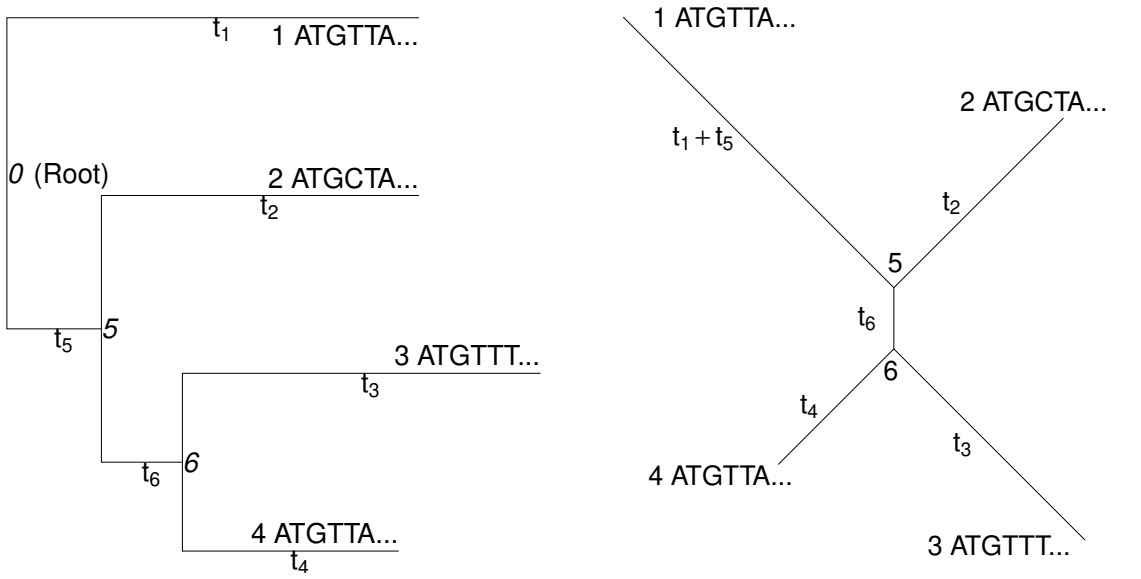
Taxon	Codon Site														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	A	T	G	T	T	A	T	T	T	A	G	T	T	G	G
2	A	T	G	C	T	A	T	T	T	A	G	T	T	G	A
3	A	T	G	T	T	T	T	T	A	G	T	T	G	A	T
4	A	T	G	T	T	A	T	T	T	A	G	T	T	G	A
1	T	T	C	A	T	T	T	A	T	A	T	A	A	T	T
2	A	T	G	G	T	G	T	T	A	T	T	G	A	T	T
3	A	T	T	A	T	T	T	A	T	T	A	A	G	G	T
4	T	T	A	G	T	A	T	T	A	G	T	A	G	T	A
1	A	T	A	T	A	A	T	A	C	A	T	T	T	T	T
2	A	T	G	G	T	G	T	T	A	T	T	G	T	T	T
3	A	T	A	T	A	A	T	T	A	G	G	A	T	T	T
4	A	T	A	G	T	A	C	T	A	C	T	A	C	T	T
1	A	T	A	T	A	A	T	A	C	A	T	T	T	T	T
2	A	T	G	G	T	G	T	T	A	T	T	G	T	T	T
3	A	T	A	T	A	A	T	T	A	G	G	A	T	T	T
4	A	T	A	G	T	A	C	T	A	C	T	A	C	T	T

Table 1.3: An example four-taxon sequence alignment.

1.3.2 Modelling Substitution as a Markov Process

A stochastic process is a collection of random variables that are indexed by a set T, which often represents time. If $X(t) = i$, the process X is said to be in state i at time t. For the values i and j from some finite set of states, all $t \geq 0$, and all $s \geq 0$, if

$$P[X(t+s) = j | X(s) = i, X(u) = x(u), 0 \leq u < s] = P[X(t+s) = j | X(s) = i] \quad (1.1)$$



(a) Rooted tree

(b) Unrooted tree

Fig. 1.1: A four-taxon, rooted tree (a) and the corresponding unrooted tree (b). The rooted tree has 6 branches with lengths labelled t_1 to t_6, 3 internal nodes labelled 0, 5, and 6 and 4 external nodes labelled 1, 2, 3, and 4. Node 0 is the root node, which is interpreted as the common ancestor all other nodes. The unrooted tree has no root node. Partial sequence data is shown for the extant taxa at the external nodes.

holds, then the stochastic process is a continuous-time Markov process and equation (1.1) is referred to as the Markov property, i.e., the conditional distribution of future states given present and past states depends only on the present state. This makes Markov processes well suited for modelling nucleotide substitution as only the present nucleotides states are observable. If equation (1.1) is independent of s, the Markov process is said to be time-homogeneous and the probability of transitioning from state i to j in time t can be expressed as $p_{ij}(t)$. Because transition probabilities satisfy the Chapman-Kolmogorov theorem,

$$p_{ij}(t_1 + t_2) = \sum_k p_{ik}(t_1) p_{kj}(t_2), \tag{1.2}$$

the probability of transitioning from state i to state j in time $t_1 + t_2$ is equal to the probably of first transitioning to any intermediate state in time t_1 before transitioning to state j in time t_2. Thus, using a Markov process to estimate, e.g., genetic divergence accounts for unobserved nucleotide transitions.

The Jukes and Cantor model (JC69) uses a continuous time Markov process to model nucleotide substitution (Jukes and Cantor, 1969). The model is useful for understanding properties that are shared with other models of molecular evolution, including more sophisticated models that will be presented below. It assumes that any nucleotide, i, has the same instantaneous rate, λ, of transitioning to any other nucleotide state, j. The rate matrix for the JC69 model is

$$Q = \{q_{ij}\} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{array}. \tag{1.3}$$

The probability of transitioning from nucleotide state i to state j within some small time interval h is $p_{ij}(h) = \lambda h + o(h)$ with $o(h)$ representing some function $g(h)$ such that $g(h)/h \to 0$ as $h \to 0$. The probability of remaining in state i is $p_{ii}(h) = 1 - \sum_{j \neq i} p_{ij}(h) = 1 - 3\lambda h + o(h)$, so the transition probabilities within the small time interval h can be expressed as

$$P(h) = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} 1-3\lambda h & \lambda h & \lambda h & \lambda h \\ \lambda h & 1-3\lambda h & \lambda h & \lambda h \\ \lambda h & \lambda h & 1-3\lambda h & \lambda h \\ \lambda h & \lambda h & \lambda h & 1-3\lambda h \end{pmatrix} \end{array} + o(h) = I + Qh + o(h). \tag{1.4}$$

When h is 0, the current nucleotide state can not change and the transition probability matrix is the identity matrix.

The transition probability matrix $P(t)$, the probability of transitioning from nucleotide i to j in any time $t > 0$, can be obtained by directly solving the Kolmogorov

backward equation.

$$\begin{aligned} P(h+t) &= P(h)P(t) \\ &= [I + Qh + o(h)]P(t) \\ P(t+h) - P(t) &= QhP(t) + o(h)P(t) \\ [P(t+h) - P(t)]/h &= QP(t) + o(h)P(t)/h \\ P'(t) &= QP(t) \qquad (\text{as } h \to 0) \end{aligned} \tag{1.5}$$

A more general solution involves exponentiation of Qt. Using the Chapman-Kolmogorov theorem, the transition probabilities can be expressed as

$$\begin{aligned} P_t &= P_{t\frac{n-1}{n}} P_{t\frac{1}{n}} \\ &= \left(P_{t\frac{n-2}{n}} P_{t\frac{1}{n}}\right) P_{t\frac{1}{n}} = P_{t\frac{n-2}{n}} \left(P_{t\frac{1}{n}}\right)^2 \\ &= \left(P_{t\frac{n-3}{n}} P_{t\frac{1}{n}}\right) \left(P_{t\frac{1}{n}}\right)^2 = P_{t\frac{n-3}{n}} \left(P_{t\frac{1}{n}}\right)^3 \\ &\vdots \\ &= \left(P_{t/n}\right)^n, \end{aligned}$$

where the argument is indicated as a subscript for clarity. When n is large, the time interval t/n is small, and from equation (1.4), the transition probabilities are $P(t) = lim_{n\to\infty}[I + Qt/n + o(t/n)]^n$. Because t is fixed, $o(t/n) = o(1/n)$, and $P(t)$ can be expressed as $\sum_{k=0}^{\infty}(Qt)^k/k!$, a Taylor series expansion of an exponential, and thus

$$P(t) = e^{Qt}. \tag{1.6}$$

A common approach for solving equation 1.6 is by eigen decomposition. An $N \times N$ matrix, A, with N linearly independent eigenvectors can be expressed in its eigen decomposition, $A = U\Lambda U^{-1}$, where U is the $N \times N$ matrix whose i^{th} column is the eigenvector u_i of A and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues. Note that $A^2 = (U\Lambda U^{-1})(U\Lambda U^{-1}) = U\Lambda(U^{-1}U)\Lambda U^{-1} = U\Lambda^2 U^{-1}$ and in general $A^n = U\Lambda^n U^{-1}$. The eigen decomposition of the rate matrix, $Q = U\Lambda U^{-1}$, allows $P(t)$ to be obtained by exponentiating the diagonal matrix

entries.

$$\begin{aligned} P(t) &= \sum_{k=0}^{\infty} \frac{(Qt)^k}{k!} \\ &= \sum_{k=0}^{\infty} \frac{(U\Lambda U^{-1}t)^k}{k!} \\ &= U\left[\sum_{k=0}^{\infty} \frac{(\Lambda t)^k}{k!}\right] U^{-1} \\ &= Ue^{\Lambda t}U^{-1} \\ &= U\text{diag}\{e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_c t}\}U^{-1} \end{aligned} \tag{1.7}$$

The solution of $P(t)$ for the JC69 model is

$$P(t) = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix} \end{array}, \text{ with } \begin{cases} p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} \\ p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} \end{cases} \tag{1.8}$$

Each row of the transition-probability matrix is a probability distribution, and thus sums to 1. When $t = 0$ the transition-probability matrix is the identity matrix, i.e., over time $t = 0$ the current nucleotide state can not change. The limiting distribution when $\lim_{t\to\infty} p_{ij}(t) = \pi_j$ represents the probability that the process is in state j after infinite time. For the JC69 model, the limiting probabilities, $\boldsymbol{\pi} = (\pi_T, \pi_C, \pi_A, \pi_G)$ are $(1/4, 1/4, 1/4, 1/4)$, i.e., when enough time has passed and so many substitutions have occurred, the probably of observing any nucleotide at a site is equal, regardless of the nucleotide state at time $t = 0$. When the vector of states, $\boldsymbol{\pi}$, satisfies the $\boldsymbol{\pi} = \boldsymbol{\pi}P(t)$ for all $t \geq 0$, $\boldsymbol{\pi}$ is referred to as the stationary distribution and if such a distribution of a Markov process exists, it is unique. The stationary distribution for the JC69 model is the limiting distribution, $\boldsymbol{\pi} = (\pi_T, \pi_C, \pi_A, \pi_G) = (1/4, 1/4, 1/4, 1/4)$. Another notable property of the JC69 model is that time, t and rate, λ are present as a product, and thus only distance, the expected number of nucleotide substitutions per nucleotide site, $d = \lambda t$, can be estimated.

Most conventional models of evolution are time reversible. A Markov process is said to be time reversible if the long-run rate of transitioning from state i to state j is equal to the long-run rate of transitioning from state j to state i, i.e., $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$, for all states, i and j. There are two useful implications for time-reversibility with respect to models of evolution. First, a time-reversible Markov process is guaranteed to have real eigenvalues, so the transition probability matrix can be obtained using equation (1.7). The second implication relates to site probabilities and the evolutionary relationship of the sequences. Given two sequences, the site probabilities are equivalent when either sequence is the ancestor, or if both sequences are descendants of some ancestral sequence. With more than two sequences and their evolutionary relationship described by a phylogenetic tree, time-reversibility means the probabilities of the data do not depend on the root node.

1.3.3 An Excess of Nonsynonymous Codon Substitutions is a Signature for Positive Selection

The JC69 and related models (e.g., Kimura, 1980; Tamura and Nei, 1993) describe substitutions between nucleotides using a Markov process. Other models of molecular evolution describe substitutions between either amino acids or codons. Those models of amino acid or codon substitution follow the same Markov theory described previously, but the state space of their Markov processes are either the 20 amino acids in proteins or the 61 sense codons of the universal genetic code. Models of codon evolution are often more powerful for detecting positive selection, because they can harness additional information contained in the DNA sequences about whether a codon substitution resulted in a change in the encoding amino acid or not.

With four possible nucleotides in each of the three positions in a codon, there are $4^3 = 64$ different codons. Of these 64 codons, the 61 codons that code for amino acids are referred to as sense codons. There are 20 different amino acids commonly found in proteins, thus the genetic code is redundant. In the standard code, only tryptophan and methionine are encoded by single codons and all other amino acids are encoded by two, four, or six different codons (table 1.1). Because of this redundancy in the genetic code, a nucleotide substitution within a codon can either result in a change in the amino acid (nonsynonymous substitution) or no change in the amino acid (synonymous substitution).

Under the assumption that natural selection acts only on proteins, all synonymous substitutions must be selectively neutral and will be fixed in the population by chance alone. By contrast, because nonsynonymous substitutions cause changes to proteins, selection may affect their fixation. Thus, a comparison of nonsynonymous and synonymous substitutions can be used to detect selection. Consider two models, an alternative model that permits positive selection and a null model that is nested within the alternative model and does not permit positive selection. Under both the alternative and null models, the ratio of the long-run proportions of nonsynonymous and synonymous substitutions, p_N/p_S and p_N^0/p_S^0, can be determined. The ratio of these ratios, $(p_N/p_S)/(p_N^0/p_S^0)$ over a fixed period of time, often referred to as d_N/d_S or the ratio of nonsynonymous to synonymous codon substitutions (ω), can be used to infer the strength and direction of selection (Yang, 2014, pp. 43-65). If under the alternative model ω is less than it would be under a comparable null model, this is a signature of purifying selection (Kimura, 1986). If under the alternative model ω is larger than it would be under a comparable null model, this is a signature of positive selection. This is a fundamental concept considered throughout this thesis.

1.3.4 A Model of Codon Evolution

The model of codon evolution described in (Nielsen and Yang, 1998) defines the relative, instantaneous substitution rate between codon i and j ($i \neq j$) at site h as

$$Q_{ij} \propto \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition} \end{cases} \tag{1.9}$$

where Q is the rate matrix of a continuous-time, stationary, time-reversible Markov process. The π_j parameters are the stationary frequencies of codon j, which can be estimated using different methods:

- Fequal: All sense codons have the frequency 1/61.
- F61: Each codon frequency is a parameter with the constraint that all frequencies sum to 1 (60 free parameters)

- F1x4: Use nucleotide frequencies such that the frequency of, e.g., codon ACG is $\pi_{ACG} = (1/C)\pi_A^*\pi_C^*\pi_G^*$ where C is a scale factor to ensure the frequencies sum to 1 and π_A^*, π_C^*, π_G^*, and $\pi_T^* = 1 - \pi_A^* - \pi_C^* - \pi_G^*$ are the nucleotide frequencies (3 free parameters)

- F1x4MG: Use separate nucleotide frequencies for each of the three positions within a codon such that the frequency of the target codon is the frequency of the target nucleotide. As an example, the frequency for a substitution from ACA to ACG would be π_G^3 (9 free parameters) (Muse and Gaut, 1994).

- F3x4: Use separate nucleotide frequencies for each of the three positions within a codon such that the frequency of, e.g., codon ACG is $\pi_A^1\pi_C^2\pi_G^3$ (9 free parameters) (Yang, 1997).

Nucleotides can be categorized into two different groups based on whether their nitrogen-containing base has one or two rings. The pyrimidines, C and T, have a nitrogen-containing base with a single ring, whereas the base of purines, A and G, have two fused rings. Substitutions within each group, i.e., $C \leftrightarrow T$ or $A \leftrightarrow G$, is referred to as a transition, whereas substitutions between groups, i.e., $C \leftrightarrow A$ or $G \leftrightarrow T$, is called a transversion. The κ parameter, the transition to transversion rate ratio, accounts for the different rates of the two types of substitutions. In some models, κ is extended to a full generalized time reversible process (GTR) (Tavaré et al., 1986). The ω parameter, discussed above, is the key parameter for the inference of selection pressure.

1.3.5 Parameter Estimation using Maximum Likelihood

Maximum likelihood (ML) is a widely used method of statistical inference that is used with all models discussed in this thesis. With data $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)^T$ sampled from a population with a known distribution, the likelihood function, $L(\theta)$, is proportional to the joint density function of the data, but treated as a function of the unknown parameter, θ. A goal of ML estimation is to determine the parameter value, $\hat{\theta}$, that maximizes $L(\theta)$, i.e., to determine the parameter value that makes the observed data most probable. The parameter may be a single value, or a vector of values, $\theta = (\theta_1, \theta_2, \ldots, \theta_k)^T$. The value of θ, $\hat{\theta}$, that maximizes $L(\theta)$ is the maximum likelihood estimate (MLE) of θ. Some of the favourable properties of estimation by

ML include the likelihood principle, efficiency, consistency, and asymptotic normality (Kalbfleisch, 1985; Bickel and Doksum, 2006).

Using ML to estimate, e.g., the distance between two sequences under a JC69 model involves estimation of the single parameter, d. The probability that a site differs between two homologous sequences separated by some distance d is $p = 3p_1 = 3/4 - 3/4e^{-4d/3}$, where the distance between two sequences separated by time t is $d = 3\lambda t$. The probability of observing x out of n sites showing polymorphism is the binomial probability, $L(d) = \binom{n}{x}\left(3/4 - 3/4e^{-4d/3}\right)^x\left(1/4 - 3/4e^{-4d/3}\right)^{n-x}$. The estimate of d, $\hat{d}$, is the value of d that maximizes $L(d)$. Setting $L'(d)$ (or equivalently $log[L'(d)]$) to 0 and solving for d gives $\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{x}{n}\right)$. Thus, under a JC69 model, two aligned sequences of 350 nucleotides that differ at 29 positions have an ML estimated distance of $\hat{d} = 0.0878$.

Likelihood calculation with more than two taxa is an extension of likelihood calculation with a pair of taxa. The evolutionary history of the taxa is described using a phylogenetic tree whose topology is considered fixed in the models discussed here. Consider a codon model, such as the one described in equation (1.9), fitted to an alignment of DNA sequences with n codon sites. Denote the codons in the sequences at site h ($h = 1, \ldots, n$) as $\mathbf{x_h}$, the site pattern at site h. Because sites are assumed to evolve independently, the likelihood of the data is the product of site probabilities, $f(\boldsymbol{x}_h; \theta)$,

$$L(\theta) = \prod_{h=1}^{n} f(\boldsymbol{x}_h; \theta). \tag{1.10}$$

It is equivalent and usually more convenient to maximize the log transformation of the likelihood,

$$\ell = \log(L) = \sum_{h=1}^{n} \log\{f(\boldsymbol{x}_h; \theta)\}. \tag{1.11}$$

Consider calculation of $f(\boldsymbol{x}_h; \theta)$ over tree topology in figure 1.1a. To calculate site probabilities, it is necessary to condition on all possible unknown ancestral states. Under a codon model, this requires summation over all 61 possible codons for each internal node. Given the ancestral state, substitution along a branch of the tree is assumed to be independent of the substitution processes along the other branches. This means calculation of $f(\boldsymbol{x}_h; \theta)$ is the product of the substitution probabilities over

branches of the tree,

$$f(\boldsymbol{x}_h;\theta) = \sum_{x_0}\sum_{x_5}\sum_{x_6}\left[\pi_{x_0}p_{x_0,TTA}(t_1)p_{x_0,x_5}(t_5)p_{x_5,CTA}(t_2)p_{x_5,x_6}(t_6)p_{x_6,TTT}(t_3)p_{x_6,TTA}(t_4)\right]. \tag{1.12}$$

Here x_i represents the ancestral state and the h subscript is omitted. At the root node, the stationary distribution of the substitution process is assumed. Using equations (1.7) and (1.9) along with the universal genetic code in table 1.1, transitioning from, e.g., GTA at node 5 to CTA at node 2 over branch length t_2 is given by $p_{GTA,CTA}(t_2) = \pi_{GTA}e^{\omega\pi_{CTA}t_2}$. The quantity in the exponent is not scaled by κ because G is a purine and C is a pyrimidine, but is scaled by ω because GTA and CTA do not code for the same amino acid, i.e., this codon substitution is a nonsynonymous transversion.

Because the substitution process described in equation (1.9) is reversible and branch length estimation is unconstrained, the likelihood calculation over the tree is invariant to the placement of the root node (Felsenstein, 2004, p. 256). Consider the tree in figure 1.1a and a set of model parameters. If the root node x_0 were shifted, the tree topology and branch length parameters would change, but the likelihood of the data will remain unchanged. When the root node is shifted in either direction until either branch length t_1 or t_5 is 0, the tree becomes unrooted.

Algorithms designed to obtain MLEs, such as $\hat{\theta} = (\hat{\kappa}, \hat{\omega}, \hat{t}_i)$ from the model described in equation (1.9), explore parameter space by adjusting parameter values and recalculating the likelihood. The goal is to ascend a metaphorical likelihood landscape to reach the peak where the parameter values maximize the likelihood function. Each such iteration requires recalculation of the likelihood of the data (equation 1.11) which in turn requires recalculation of the transition probabilities over the tree. In the four-taxon tree in figure 1.1a, there are 61^3 unobserved codon states per site to sum over. Increasing the number taxa quickly makes likelihood maximization a computationally expensive task.

To economize on the computation of the likelihood over phylogenetic trees, the pruning algorithm was developed (Felsenstein, 1973), an application of Horner's method. Horner's method describes an optimal algorithm for evaluation of polynomials such that a polynomial of degree n is calculated with n multiplications and n additions. For example, calculation of $a_0 + a_1x + a_2x^2 + a_3x^3$ as $a_0 + a_1 \cdot x +$

$a_2 \cdot x \cdot x + a_3 \cdot x \cdot x \cdot x$ requires three additions and six multiplications. Calculation following Horner's method, $a_0 + x \cdot (a_1 + x \cdot (a_2 + a_3 \cdot x))$, requires only three additions and three multiplications. The number of computations can be similarly reduced in equation 1.13 by pushing the summations as far right as possible,

$$f(\boldsymbol{x}_h; \theta) = \sum_{x_0} \Bigg[\pi_{x_0} p_{x_0,TTA}(t_1) \Bigg[\sum_{x_5} p_{x_0,x_5}(t_5) p_{x_5,CTA}(t_2) \Bigg[\sum_{x_6} p_{x_5,x_6}(t_6) p_{x_6,TTT}(t_3) p_{x_6,TTA}(t_4) \Bigg] \Bigg] \Bigg]. \tag{1.13}$$

For m terminal nodes, the number of computations is exponential in m using a naive calculation as in equation 1.13, but is reduced to linear in m when the pruning algorithm is followed.

Formal hypothesis tests can be conducted using the likelihoods of two competing models, a null model and an alternative model. The likelihood under the alternative model is found by maximizing over the entire parameter space, whereas constraints are imposed on the null model making it a special case of the alternative model. For example, a test for positive selection could be formulated from equation (1.9) with the null model imposing the constraint $\omega = 1$ and the alternative model allowing $\omega \geq 1$. A standard likelihood ratio (LR) test calls for twice the difference in log likelihoods between the null and alternative model to be compared to thresholds from a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters, Δ_p, between the models,

$$2\Delta\ell = 2[\log(L_a) - \log(L_0)] \sim \chi^2_{\Delta_p}, \tag{1.14}$$

where L_0 and L_a are the likelihood scores under the null and alternative models.

1.3.6 Selection Pressure at Amino Acid Sites

Because proteins subjected to positive selection must still maintain the capacity to fold into complex structural and functional domains, most amino acid sites in a protein will be subjected to purifying selection pressure, but a large point mass of neutral nonsynonymous mutations is nonetheless expected in the distribution of mutational effects (e.g., Kim et al., 2017). This means a single ω estimated as an average over

all sites of such a protein would rarely be large enough to detect positive selection. Including a separate ω parameter for each site would require a large number of taxa, which is often not practical. An alternative approach is to use site classes that are subject to different levels of selection pressure such as $\omega < 1$, $\omega = 1$, $\omega > 1$ (Nielsen and Yang, 1998). The ω at a site is treated as coming from a probability distribution, with various distributional forms allowed (Yang et al., 2000a). A null model includes some number of $\omega \leq 1$ site classes, but does not permit ω values larger than 1, i.e., positive selection is not permitted. The probability of site pattern $\boldsymbol{x}_h$ can be expressed as a mixture over choices of ω,

$$p(\boldsymbol{x}_h; \theta) = \sum_i p_i p(\boldsymbol{x}_h | \omega_i; \zeta) \tag{1.15}$$

with θ denoting all model parameters including branch lengths and ζ denoting all parameters other than those describing the ω distribution, namely p_i and ω_i. The alternative model includes the ω site classes of the null model and also a site class for $\omega > 1$,

$$p(\boldsymbol{x}_h; \theta) = \sum_i p_i p(\boldsymbol{x}_h | \omega_i; \zeta) + \left(1 - \sum_i p_i\right) p(\boldsymbol{x}_h | \omega > 1; \zeta). \tag{1.16}$$

As the null model is nested within the alternative model, a standard LR test calls for twice the difference in log likelihoods to be compared to thresholds from a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between the models. For the models in equations (1.15) and (1.16), one additional $\omega > 1$ site class in the alternative model gives two additional parameters, $p_{\omega>1}$ and $\omega > 1$, which would suggest the LR statistic follows a χ^2_2. However, because certain regularity conditions of the test are not satisfied, the specific distribution is unknown and in practice either a $\chi^2_0/2 + \chi^2_1/2$ or a χ^2_1 is used (Wong et al., 2004; Yang, 1997).

If the LR test for positive selection within a gene is rejected, evidence of positive selection at each site can be gathered. To calculate the posterior probability that site h evolved under ω site class i, the Naïve empirical Bayes (NEB) approach passes the

MLEs to Bayes formula,

$$Pr(\omega^{(h)} = \omega_i | \boldsymbol{x}_h; \zeta) = p_i f(\boldsymbol{x}_h | \omega_i; \zeta) / \sum_{j=1}^{k} p_j f(\boldsymbol{x}_h | \omega_j; \zeta). \tag{1.17}$$

A large posterior probability that site h evolved under the $\omega > 1$ site class is interpreted as evidence that site h evolved under positive selection.

Because the site posterior probabilities always depend on the fitted values of the model parameters (shape parameters of the distribution, branch lengths, etc.), the reliability of NEB inference depends on the accuracy of the fitted values. If they have been accurately estimated, as is often the case with large, information-rich datasets, they can simply be treated as known without errors. However, when the fitted values are subject to large errors, the detection of positive selection according to the posterior probabilities can be negatively impacted and in some cases the false positive rate can be unacceptably high (e.g., Wong et al., 2004). Bayes empirical Bayes (BEB), is used to adjust for uncertainty in the parameters of the ω distribution by assigning priors to those parameters and using numerical integration to average over the uncertainty.

1.4 Thesis Outline

Standard likelihood theory calls for the LR test for positive selection within a gene used by mixture models of codon evolution to compare the LR statistic to thresholds determined from a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between the null and alternative models. This is not justified due to a lack of statistical regularity (Anisimova et al., 2001), so the LR test is often applied with thresholds determined from χ^2 or mixture of χ^2 distributions with degrees of freedom selected to make the test more conservative with the hope that not much power is lost.

Results from Chapter 2 show that these commonly used thresholds need not yield conservative tests, but instead give larger than expected type I error rates. Statistical regularity can be restored by using a modified LR test. Theoretical results are provided to prove that, if the number of sites is not too small, the modified LR test gives approximately correct type I error probabilities regardless of the parameter settings of the underlying null hypothesis. Simulations show that modification gives type I

error rates closer to those stated without a loss of power. The simulations also show that parameter estimation for mixture models of codon evolution can be challenging in certain data-generation settings with very different mixing distributions giving nearly identical site pattern distributions unless the number of taxa and tree length are large. Because mixture models are widely used for a variety of problems in molecular evolution, the challenges and general approaches to solving them presented here are applicable in a broader context.

To mitigate problems with classification of selection pressure at sites when parameter are estimated with large errors, BEB assigns prior probabilities to some parameters. However, as implemented, it imposes uniform prior probabilities, which causes it to be overly conservative in some cases. When standard regularity conditions are not met and parameter estimates are unstable, inference, even under BEB, can be negatively impacted.

In Chapter 3, an alternative to BEB called smoothed bootstrap aggregation (SBA) is presented. SBA uses bootstrapping of site patterns from an alignment of protein coding DNA sequences to accommodate the uncertainty in the parameter estimates. Deriving the correction for parameter uncertainty from the data in hand along with kernel smoothing techniques improves site specific inference of positive selection. Included is a comparison of BEB and SBA by simulation and real data analysis. Simulation results show that SBA balances accuracy and power at least as well as BEB, and when parameter estimates are unstable, the performance gap between BEB and SBA can widen in favour of SBA. SBA is applicable to a wide variety of other inference problems in molecular evolution.

Chapter 2

A Modified Likelihood Approach to Explore and Restore Regularity when Testing for Positive Selection

This work was previously published in the journal *Bioinformatics* (Mingrone et al., 2018).

2.1 Introduction

Detecting positive selection within proteins is important for understanding the processes of molecular evolution (Nielsen and Yang, 1998). The likelihood methods used in codon-based models developed in Yang et al. (2000a) are among the most widely used to test for positive selection. An important component of these models is the LR test, which is used to test for evidence of positive selection within a gene or as a filter before testing for positive selection at amino acid sites. Standard likelihood theory gives that, when certain regularity conditions are satisfied, the distribution of an LR statistic under the null hypothesis is that of a chi-square random variable with degrees of freedom equal to the difference between the number of parameters fit under the alternative hypothesis and the number under the null hypothesis. LR tests of positive selection usually employ two additional parameters under the alternative model, often an $\omega > 1$ parameter to quantify the positive selection and another parameter for the proportion of sites evolving under $\omega > 1$. This suggests the LR statistics follows a χ^2_2 null distribution. However, it has long been recognized that the regularity conditions required for standard likelihood theory are not satisfied for such LR tests of positive selection (Anisimova et al., 2001).

Simulations suggest that a χ^2_2 distribution will give 5% thresholds for the LR test that are too large (Anisimova et al., 2001; Wong et al., 2004). Drawing upon the

non-standard likelihood theory of Self and Liang (1987), Swanson et al. (2003) indicate that, for model comparison they describe as M8a vs M8, theory supports a 50:50 mixture of a point mass at 0 and a χ^2_1 distribution or, more concisely, a $\chi^2_0/2 + \chi^2_1/2$ distribution. However, Wong et al. (2004) and Anisimova et al. (2001) raised concerns about whether this is the appropriate distribution for comparison. Nevertheless, the $\chi^2_0/2 + \chi^2_1/2$ distribution and, to be more conservative, the χ^2_1 distribution are the most frequently used distributions. While there have been some simulation studies indicating that the χ^2_1 distribution is indeed conservative in the sense that LR statistics generated under the null tend to be smaller than predicted by a χ^2_1 distribution (Anisimova et al., 2001; Wong et al., 2004; Berlin and Smith, 2005), some of these same studies have found settings where the false positive rates are larger than 5% (Wong et al., 2004; Berlin and Smith, 2005).

That the null distribution of the LR statistic is neither χ^2 nor a mixture of χ^2 distributions is a theoretical possibility, even with large samples, because of a regularity condition violation in both standard likelihood theory and the non-standard likelihood theory of Self and Liang. If the only regularity condition violation were that parameters are on the boundary of the parameter space, the Self and Liang theory would hold and the LR test using the mixture of χ^2 distributions would be conservative. An explanation is provided for why a lack of identifiability under the null hypothesis makes it possible that the LR test will be anti-conservative (under the null, LR statistics tend to be larger than is predicted by a $\chi^2_0/2 + \chi^2_1/2$ distribution). The anti-conservative behaviour is confirmed through simulation.

A dramatic illustration of how a lack of this type of identifiability with mixture models can cause LR tests to be anti-conservative is provided by Hartigan (1985) (see also Chen, 2017). The setting was a test of $N(0, 1)$ against the alternative hypothesis of a mixture of $N(0, 1)$ and $N(\theta, 1)$. The testing problem suffers from a similar irregularity problem to the one described here: a lack of identifiability of the full model under the null hypothesis. When a mixing weight is 0 any θ gives the null hypothesis and Hartigan (1985) shows that the LR statistic approaches ∞ with probability 1.

To obtain tractable limiting $\chi^2_0/2 + \chi^2_1/2$ null distributions, a modified LR test was developed. The modified likelihood, a type of penalized likelihood, borrows from

similar methods in mixture model tests of heterogeneity (Chen et al., 2001). The test statistic is obtained as it is for the standard LR test, but with the likelihood replaced by one that penalizes small mass on $\omega > 1$ relative to $\omega = 1$. This strategy has been effective under a variety different mixture settings (cf. Chen et al., 2001, 2004; Fu et al., 2009, and references therein).

2.2 Theory and Methods

The base model of Yang et al. (2000a) is a conventional stationary time-reversible Markov model of codon sequence evolution described in Goldman and Yang (1994) with instantaneous rate matrix for transitions from codon i to j given by

$$Q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ at two or three nucleotide positions} \\ \pi_j & \text{if } i \text{ and } j \text{ differ by one synonymous transversion} \\ \kappa\pi_j & \text{if } i \text{ and } j \text{ differ by one synonymous transition} \\ \omega\pi_j & \text{if } i \text{ and } j \text{ differ by one nonsynonymous transversion} \\ \omega\kappa\pi_j & \text{if } i \text{ and } j \text{ differ by one nonsynonymous transition} \end{cases}$$

where κ is the transition/transversion parameter, π_j is the stationary frequency of codon j and ω is the parameter quantifying selection pressure as purifying ($\omega < 1$), neutral ($\omega = 1$) or positive ($\omega > 1$). To model varying selection pressure at sites, the ω at a site is treated as coming from a probability distribution, referred to here as the mixing distribution, with various distributional forms allowed (Yang et al., 2000a). The null hypothesis of interest is that there is no positive selection, which corresponds to the distribution of ω having all of its mass between 0 and 1. The alternative is that the distribution allows some positive probability of an $\omega > 1$. For example, following the naming conventions of Yang et al. (2000a) and Berlin and Smith (2005), null model M1a uses a distribution with mass at an $\omega_0 < 1$ and at $\omega_1 = 1$. The corresponding alternative model, M2a, adds an $\omega_2 > 1$ to the M1a distribution.

For any of the models considered, the probability of a site pattern x can be expressed as a mixture over choices of ω of the following form

$$p(x; \beta, p_+) = p_0 p(x|\omega < 1; \zeta, \lambda) + (1-p_+)(1-p_0)p(x|1; \zeta) + p_+(1-p_0)p(x|\omega_+; \zeta) \quad (2.1)$$

with β denoting all model parameters other than p_+. Theoretical derivations are simpler with this unconventional parameterization and it allows consideration of different models listed in supplementary table 6.1 in the same setting. Usually the weights on ω values are parameters. For instance, model M2a replaces $(1-p_+)(1-p_0)$ and $p_+(1-p_0)$ with p_1 and p_2. For both models M1a and M2a there is a single $\omega_0 < 1$, so $p(x|\omega < 1; \zeta, \lambda) = p(x|\omega_0; \zeta)$. Here ζ denotes parameters common to each ω and includes edge-lengths and substitution model parameters. The parameter ω_+ is restricted to be at least 1 and the parameters in λ are those involved in the mixture model under purifying selection. For instance, for model M8a, λ gives the parameters of the beta distribution. Let $\psi = (\zeta^T, \lambda^T, p_0)^T$ be the parameters that are common to both null and alternative models. The LR statistic is

$$2\{l(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\hat{\psi}_H)\} \tag{2.2}$$

where l and l_H denote the log likelihoods under the alternative and null models, and $\hat{p}_+$, $\hat{\omega}_+$, $\hat{\psi}$, and $\hat{\psi}_H$ denote the MLEs under the alternative and null hypotheses.

The likelihood theory of Self and Liang (1987) gives appropriate null distributions in a number of cases where usual regularity conditions do not hold, but it does not generally apply to (2.2). This is because there can be multiple parameter values under the alternative hypothesis that give the null model. Any $\omega_+ > 1$ and $p_+ = 0$ gives the null model. Also, if the alternative model allows mass at $\omega = 1$, then $\omega_+ = 1$ and any p_+ gives the null model. The M8a vs M8 comparison considered in Swanson et al. (2003) finesses this difficulty by not allowing the alternative model to have mass at both an $\omega = 1$ and an $\omega_+ > 1$. Because of this restriction, whenever the true null generating model has mass on $\omega = 1$, the only alternative model parameterization giving the generating distribution has $\omega_+ = 1$; $p_+ = 0$ and $\omega_+ > 1$ no longer gives the generating model. The Swanson et al. (2003) approach restores regularity, but may make it more difficult to model settings where the alternative is true but there is also appreciable mass near $\omega = 1$. In what follows, the model is allowed to have mass at $\omega = 1$ under both the null and alternative model, with additional mass at an $\omega_+ > 1$ under the alternative hypothesis.

The regularity problems for the Self and Liang theory do not arise if $\omega_+ > 1$ is

fixed, in which case the LR statistic is

$$2\{l(\hat{p}_+(\omega_+), \omega_+, \hat{\psi}(\omega_+)) - l_H(\hat{\psi}_H)\} \tag{2.3}$$

where $\hat{p}_+(\omega_+)$ and $\hat{\psi}(\omega_+)$ denote the MLEs of p_+ and ψ holding ω_+ fixed. With ω_+ fixed, the only parameter giving a null model is $p_+ = 0$. Because that value is on the boundary of the parameter space, standard chi-square results for the limiting distribution of the LR statistic do not apply. However, case 5 of Self and Liang (1987) gives that the large sample distribution is $\chi^2_0/2 + \chi^2_1/2$. This allows something to be said about the distribution of the usual LR statistic (2.2). Because (2.2) can be obtained by maximizing (2.3) over $\omega_+ \geq 1$, it is sure to be larger than any test statistic (2.3) that uses a fixed ω. Thus, since (2.3) has a $\chi^2_0/2 + \chi^2_1/2$ distribution, usual LR statistic values (2.2) will tend to be larger than values predicted by the $\chi^2_0/2 + \chi^2_1/2$ distribution. How much larger LR statistic values tend to be depends upon how much (2.3) tends to vary over $\omega_+ > 1$ which in turn likely depends on how much of the mass of the generating distribution is near $\omega = 1$. Thus, using a $\chi^2_0/2 + \chi^2_1/2$ distribution to calculate thresholds for the LR test can generally be expected to give an anti-conservative test: the null hypothesis is rejected too frequently when it is true.

The main reason that the null distribution of the LR statistic is intractable is that $p_+ = 0$ and any $\omega_+ > 1$ gives the null model. A similar difficulty arises when testing for mixture structure or heterogeneity in mixture models. The distribution for the data, x, is $\gamma p(x;\theta_1) + (1-\gamma)p(x;\theta_2)$ where $p(x;\theta)$ is a parametric distribution. A hypothesis of particular interest is that the data corresponds to a single distribution $p(x;\theta)$. If this is the case, the population might be considered homogeneous when it is otherwise heterogeneous with $\gamma \times 100\%$ of the individuals having parameter θ_1 and the rest having parameter θ_2. As with tests for positive selection, the reason for a non-standard LR statistic distribution in mixture models is that multiple parameter settings correspond to the null hypothesis: (i) $\gamma = 0$ and any θ_1 or (ii) $\theta_1 = \theta_2$ and any γ. To restore simple limiting distributions while maintaining a test statistic similar to the LR statistic, Chen et al. (2001) replace log likelihoods with modified log likelihoods that add a term, $C\log[\gamma(1-\gamma)]$ where $C > 0$ is a tuning parameter. Because this term gets very large in magnitude but negative when γ is close to 0 or 1,

the modified log likelihood is maximized by values with γ away from these boundaries, implying that the only way modified MLEs under the null can approach true values is if $\hat{\theta}_1 \approx \hat{\theta}_2$, which restores the sort of regularity needed for chi-square or mixture of chi-square distributions. The strategy has been effective in a number of different settings (cf. Chen et al., 2001, 2004; Fu et al., 2009, and references therein) and a similar approach here is presented here.

The modified log likelihood under the alternative hypothesis is

$$\tilde{l}(p_+, \omega_+, \psi) = l(p_+, \omega_+, \psi) + C \log(p_+) \tag{2.4}$$

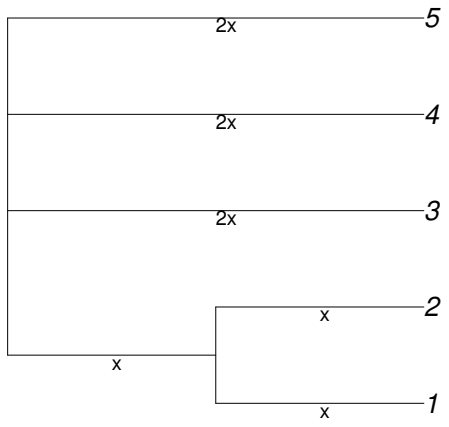
and the modified LR statistic is then

$$2\{\tilde{l}(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\hat{\psi}_H)\} \tag{2.5}$$

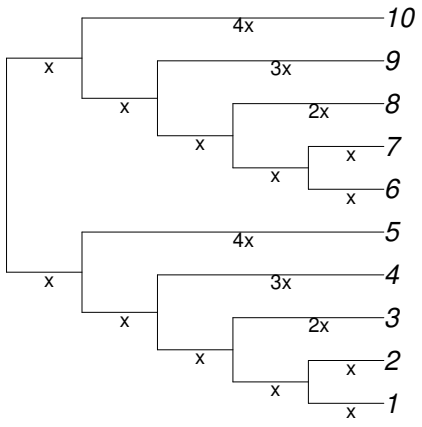
where now the estimates denote the maximizers of the modified log likelihood. Shown in Appendix II (section 6.2) is that for $C > 0$ the large sample distribution of (2.5) under the null hypothesis is $\chi_0^2/2 + \chi_1^2/2$. Here $C > 0$ is a tuning parameter. While the theory holds for any $C > 0$, choosing C too small makes the modified LR statistic too similar to the LR statistic, leading to similar difficulties in behaviour. The sensitivity to C through simulations is investigated.

Simulation is used to estimate LR and modified LR statistic cumulative distribution functions (CDFs) under the null hypothesis. For each of six simulation scenarios, 10,000 sequence alignments 500 codons long were generated using 5-, 10-, and 32-taxon trees with branch lengths summing to 3, 6, and 9. The 5-taxon tree (figure 2.1a) was the same one used in the simulation studies of Wong et al. (2004) and Mingrone et al. (2016) and the 10- and 32-taxon trees have caterpillar (figure 2.1b) and balanced (figure 2.1c) topologies. Sites were simulated to evolve under the M1a model (described in Wong et al., 2004; Yang et al., 2005), which places weight p_0 on a single $\omega_0 < 0$, with the remaining weight, $1 - p_0$, placed on $\omega = 1$; thus, the mixing distribution is determined by (p_0, ω_0). Each simulation scenario used $\kappa = 1$ and equal codon frequencies, but (p_0, ω_0) varied over scenarios.

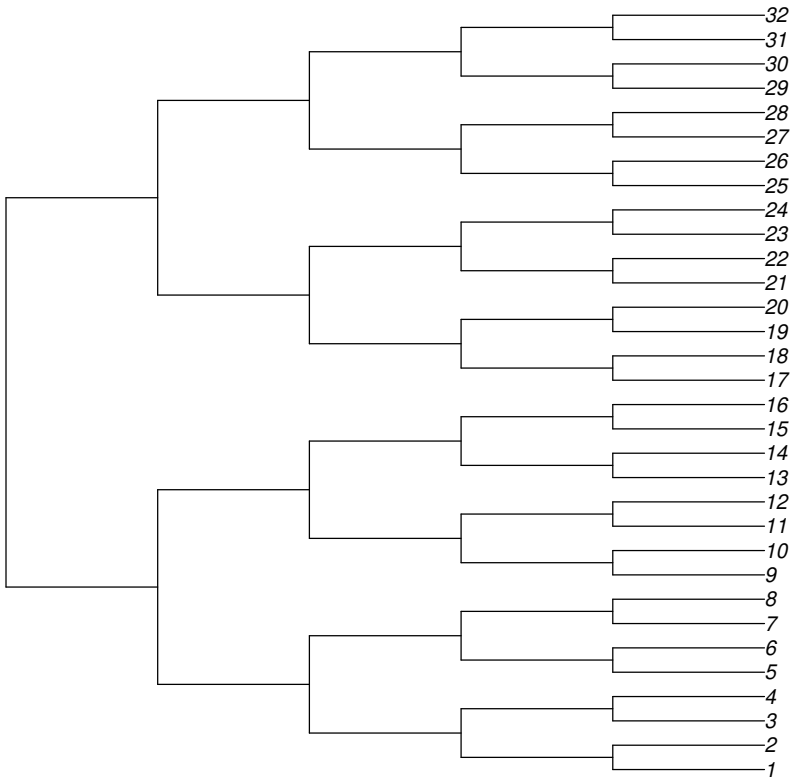
To determine the effect of likelihood modification on power, sequence alignments 500 codons long were simulated under the M2a alternative model (Wong et al., 2004;



(a) 5-taxon tree



(b) 10-taxon tree



(c) 32-taxon tree

Fig. 2.1: Phylogenetic tree topologies used in simulation studies, with relative edge lengths shown for the 5- and 10-taxon trees. All edge lengths are equal in the rooted 32-taxon tree.

Yang et al., 2005), which, by comparison with the M1a mixing distribution, has an additional component, $\omega_2 > 1$. The mixing distributions used in simulations had $(p_0, \omega_0) = (0.45, 0.5)$ with (p_2, ω_2) varying over simulation settings. Codon frequencies were 1/61 and $\kappa = 1$, as was the case for simulations under the null hypothesis and the tree toplogies also matched those used in the simulations under the null. For each ω-distribution scenario, $10,000$ alignments were generated for the 5- and 10-taxon trees and 1000 alignments for the 32-taxon tree. To ensure that comparisons of power with and without likelihood modification corresponded to the same false positive rate, the thresholds for significant LR statistics were calibrated to an error rate of 0.05. For this, $10,000$ sequences were generated under the null with the weight on $\omega > 1$ under the alternative settings added to $\omega = 1$. The 95th percentiles of these LR statistic distributions under both M1a/M2a (C=0) and M1a/M2a (C=2) were used as the thresholds for calculating power.

2.3 Results and Discussion

2.3.1 Modified LR Distribution Approximations are Accurate for Most Settings

Figure 2.2 shows the estimated LR and modified LR statistic CDFs for the M1a/M2a nested model pair for the simulations under the null using the 32-taxon tree with branch lengths summing to 9. With likelihood modification, a tuning parameter of $C = 2$ was used. Other tuning parameters were tested, but the LR statistic CDFs for values of C between 2 and 5 were indistinguishable from those with $C = 2$, and CDFs for values of $C < 2$ were always between the one for $C = 2$ and the one for the unmodified LR statistics. CDFs for $\chi_0^2/2 + \chi_1^2/2$ are also included in each plot. For all of the CDFs in Figure 2.2, the modified LR statistic distributions are better approximated by a $\chi_0^2/2 + \chi_1^2/2$ distribution than the corresponding distributions without likelihood modification. Tree topology made little difference as both the LR statistic and modified LR statistic CDFs were similar when data were simulated with different topologies. Figure 2.3 and supplementary figures 6.1 - 6.7 contain the LR statistic CDFs for the remaining simulation scenarios.

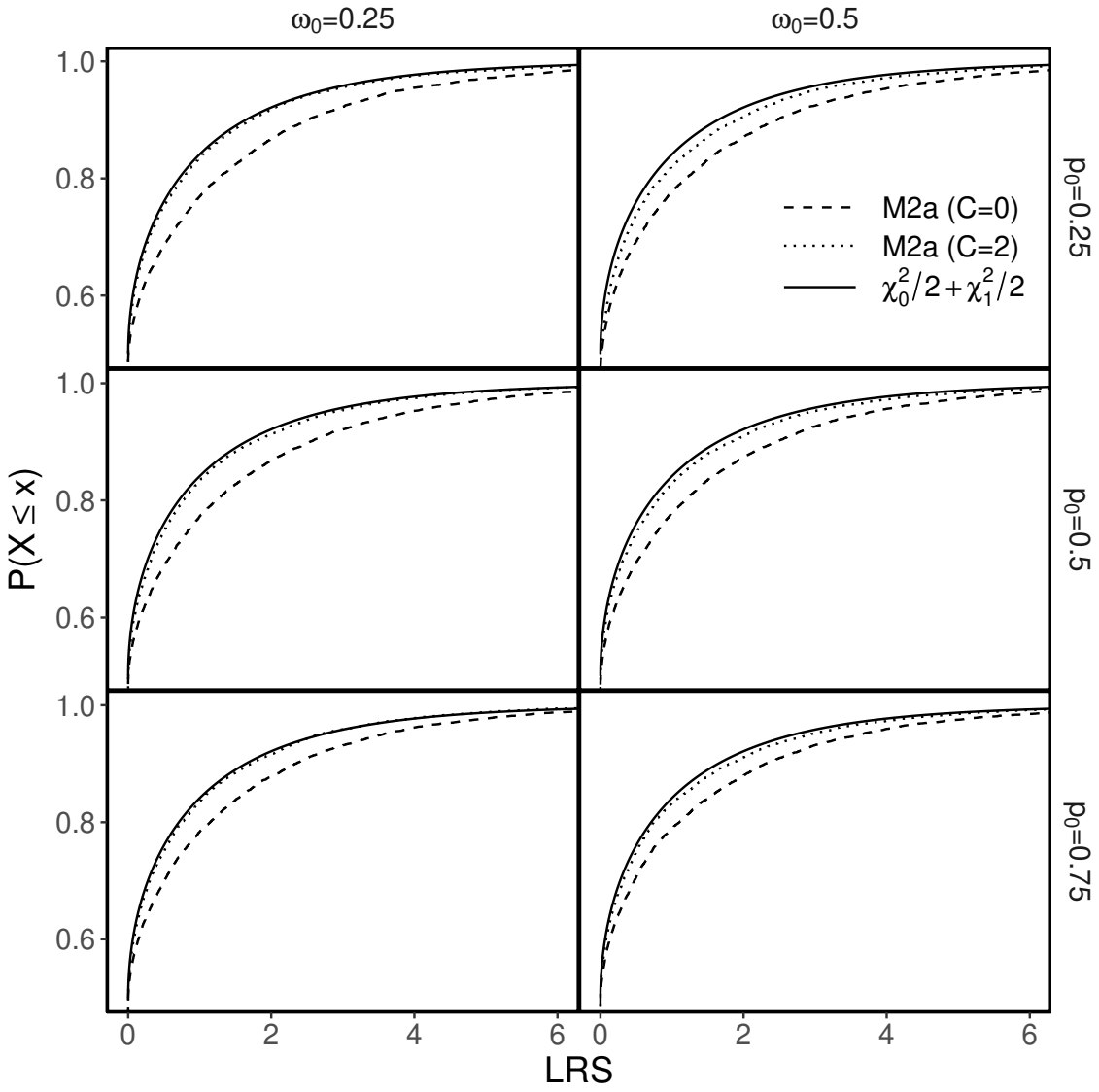


Fig. 2.2: CDFs of LR (C=0) and modified LR (C=2) statistics under M1a/M2a nested model pairs for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a balanced, 32-taxon tree topology with branch lengths summing to 9. The value of ω_0 and its weight, p_0, used to generate the data are shown as column and row labels. CDFs for $\chi_0^2/2 + \chi_1^2/2$ are also included.

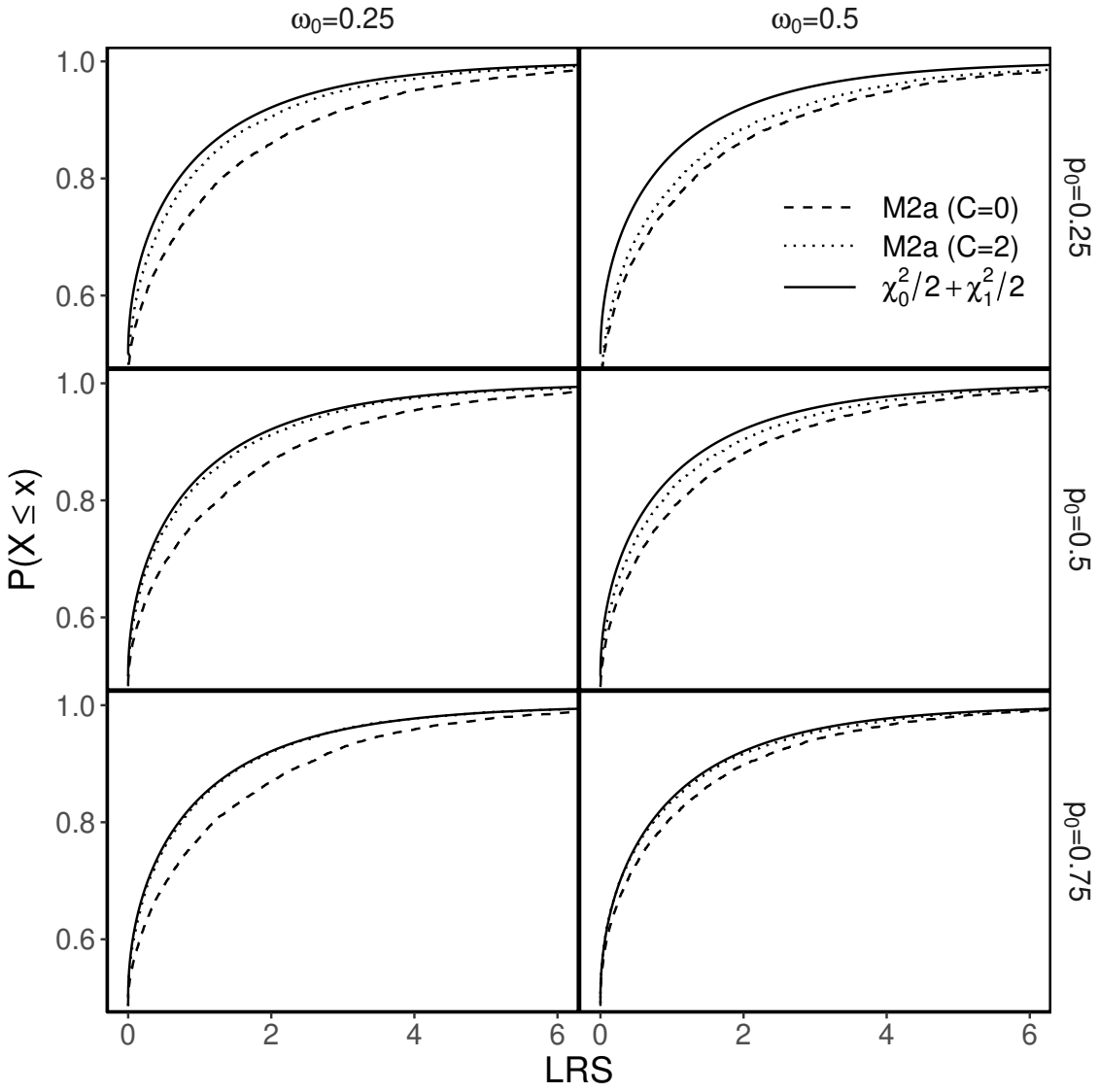


Fig. 2.3: CDFs of LR (C=0) and modified LR (C=2) statistics under M1a/M2a nested model pairs for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a 5-taxon tree topology with branch lengths summing to 3. The value of ω_0 and its weight, p_0, used to generate the data are shown as column and row labels. CDFs for $\chi^2_0/2 + \chi^2_1/2$ are also included.

2.3.2 False Positive Rates are too Large Without Modified LR Tests

The false positive rates for each of the LR tests of positive selection under nested models M1a/M2a with and without likelihood modification are shown in Table 2.1. The threshold used to reject each LR test was determined from the 95th percentile of the $\chi_0^2/2 + \chi_1^2/2$ distribution. Thus, when the $\chi_0^2/2 + \chi_1^2/2$ does well to approximate the LR statistic distribution, the expected false positive rate is 0.05. For each simulation setting under the null hypothesis, the rates were closer to the expected value using the modified likelihood than with the unmodified likelihood. Excluding the simulation scenario with 5 taxa and $(p_0, \omega_0) = (0.25, 0.5)$ where parameters are almost unidentifiable (discussed below), the false positive rates were between 0.06 and 0.1 (average 0.09) without likelihood modification and between 0.05 and 0.07 (average 0.06) with likelihood modification. While the false positive rate of the modified LR statistic was usually close to 0.05, there is a small sample bias using sequences of length 500. Analyzing datasets simulated under the same settings, but with sequences 1500 codons long confirms this bias. All but one of the false positive rates that were 0.06 with sequences 500 codons long dropped to 0.05 with sequences 1500 codons long and the false positive rate for the simulation setting with $(p_0, \omega_0) = (0.25, 0.5)$ dropped to 0.06 with the longer sequences.

2.3.3 Power of the Modified LR Tests is Comparable to Re-calibrated LR Tests

LR tests are generally expected to have power that is in some sense optimal (cf Section 5.4.4 of Bickel and Doksum, 2006). By modifying the LRs, it is possible that some loss of power will accrue. Figure 2.4 shows the power curves, using a threshold calibrated to have Type I error rate 0.05, with and without likelihood modification. The plots suggest that likelihood modification has minimal impact on power.

2.3.4 Modified Likelihood Improves Estimation for Difficult Real Data Settings

The same 16 genes described and analyzed in Mingrone et al. (2016) were analyzed and the results are summarized in Table 2.2. For each of the genes Mingrone et al. (2016) described as *regular* cases, meaning there were no indications of departures from the limiting properties predicted by ML theory, estimation showed no evidence of instabilities, and bootstrap parameter distributions had low variance (lysin, *nuoL3*,

Table 2.1: False positive rates.

		Tree Length 3																	
		5 taxa						10 taxa						32 taxa					
		$\omega_0 = .25$			$\omega_0 = .5$			$\omega_0 = .25$			$\omega_0 = .5$			$\omega_0 = .25$			$\omega_0 = .5$		
Model	$p_0 =$	.25	.5	.75	.25	.5	.75	.25	.5	.75	.25	.5	.75	.25	.5	.75	.25	.5	.75
M2a (C=0)		.10	.09	.09	.10	.08	.07	.09	.09	.08	.09	.08	.06	.09	.08	.08	.08	.08	.07
M2a (C=2)		.06	.06	.05	.08	.06	.06	.06	.06	.05	.07	.06	.05	.06	.06	.05	.07	.06	.06
		Tree Length 6																	
		5 taxa						10 taxa						32 taxa					
		$\omega_0 = .25$			$\omega_0 = .5$			$\omega_0 = .25$			$\omega_0 = .5$			$\omega_0 = .25$			$\omega_0 = .5$		
	$p_0 =$	.25	.5	.75	.25	.5	.75	.25	.5	.75	.25	.5	.75	.25	.5	.75	.25	.5	.75
M2a (C=0)		.10	.09	.08	.10	.09	.08	.10	.09	.08	.10	.08	.07	.09	.10	.09	.09	.09	.07
M2a (C=2)		.06	.06	.05	.08	.06	.07	.06	.05	.05	.07	.06	.05	.05	.06	.06	.07	.07	.05
		Tree Length 9																	
		5 taxa						10 taxa						32 taxa					
		$\omega_0 = .25$			$\omega_0 = .5$			$\omega_0 = .25$			$\omega_0 = .5$			$\omega_0 = .25$			$\omega_0 = .5$		
	$p_0 =$	.25	.5	.75	.25	.5	.75	.25	.5	.75	.25	.5	.75	.25	.5	.75	.25	.5	.75
M2a (C=0)		.10	.09	.08	.10	.09	.07	.09	.09	.08	.10	.08	.06	.09	.09	.08	.09	.09	.08
M2a (C=2)		.07	.05	.06	.08	.07	.06	.06	.06	.05	.06	.06	.05	.05	.05	.05	.06	.06	.06

False positive rates for LR tests of positive selection under nested models M1a/M2a with and without likelihood modification. For each of six simulations scenarios with varying weights and values for two site classes, $\omega < 1$ and $\omega = 1$, 10,000 sequence alignments 500 codons long were generated using 5-, 10-, and 32-taxon tree topologies with branch lengths summing to 3, 6, or 9. The value of ω_0 and its weight, p_0, used to generate the data are shown in column and row labels. Modified likelihood tuning parameters of $C = 0$ (no likelihood modification) and $C = 2$ were used. The LR statistics were compared to the 95th percentile of the $\chi^2_0/2 + \chi^2_1/2$ distribution.

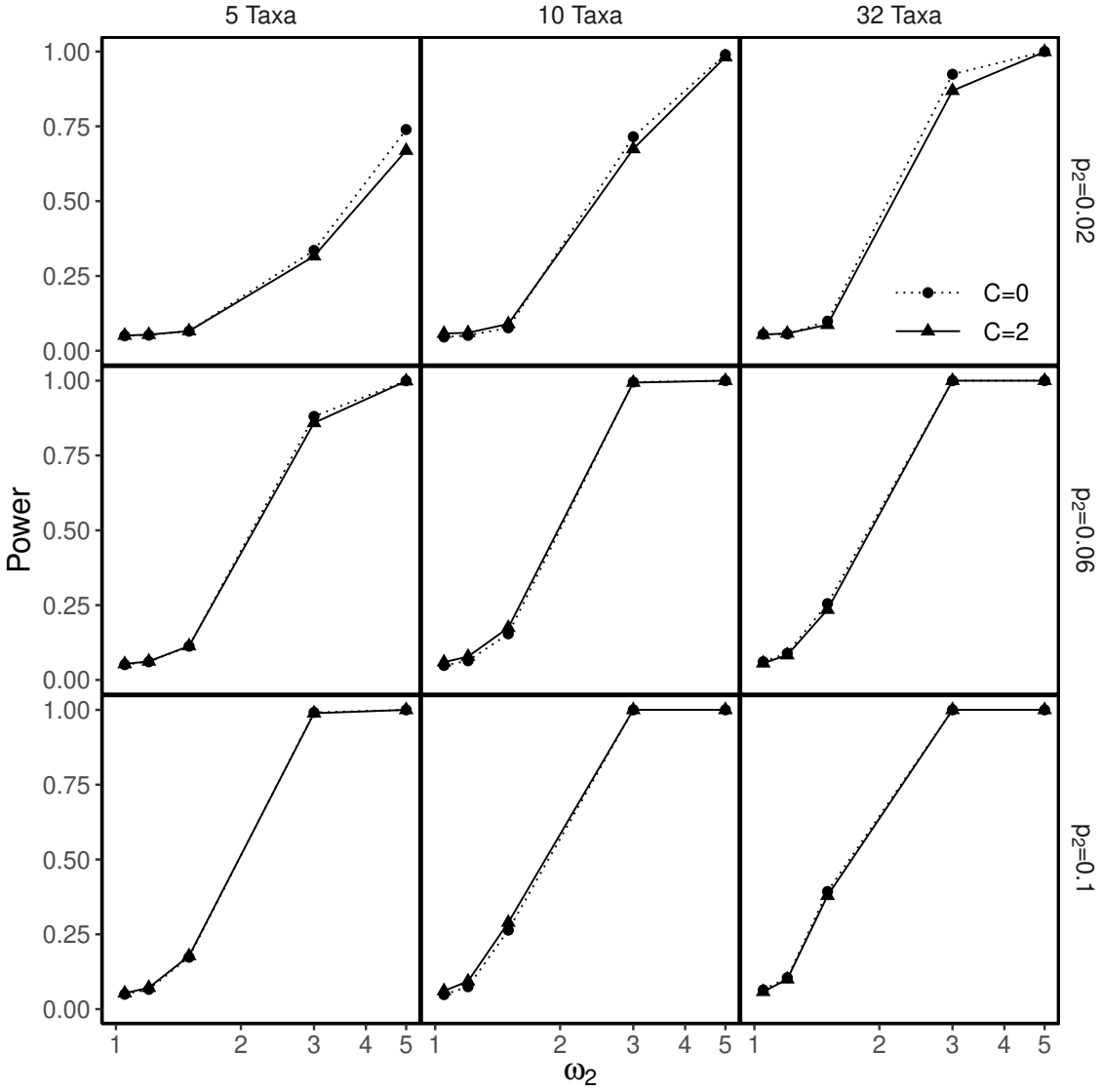


Fig. 2.4: Comparison of power under model M2a without (C=0) and with (C=2) likelihood modification. For each simulation setting, 10,000 (5 and 10 taxa) or 1,000 (32 taxa) alignments were generated with 500 codons and 45% weight on $\omega = 0.5$, p_2 weight on ω_2 and the remaining weight on $\omega = 1$.

pol, *RafL*, *TrbL-VirB6_3*, and *vif*), the LR statistics, $\hat{p}_2$, and $\hat{\omega}_2$ are comparable with and without likelihood modification. On the other hand, for 4 of the 5 *irregular* genes, genes for which the ω distribution had been poorly estimated in Mingrone et al. (2016) (*CDH3*, *mivN*, *pgpA*, *tax*, and *TrbL-VirB6_2*), the results are very different with and without likelihood modification. Without likelihood modification, an estimated $\hat{p}_2 = 0.006$ of the sites in *pgpA* were estimated to have evolved under $\hat{\omega}_2 = 34.7$ and the LR test was rejected. With likelihood modification, (p_2, ω_2) was estimated to be $(0.09, 1.00)$, the likelihoods under both the null and alternative models are the same, and the LR test was not rejected. With the exception of *tax*, the estimates of p_2 were always larger using modified likelihoods and the corresponding estimates of w_2 were always smaller with average decreases in the estimated ω_2 equal to 16.85, 2.22, and 0.29 for the genes described in Mingrone et al. (2016) as *irregular* (excluding *tax*), *uncategorized*, and *regular*, respectively. Differences in the branch length and κ estimates were minor in all cases. The only *irregular* gene with estimates that did not vary between the two likelihood approaches was the well-known *tax* gene (Suzuki and Nei, 2004; Yang et al., 2005). Its highly unusual site-pattern distribution gives extreme MLEs with 100% weight ($\hat{p}_2 = 1$) placed on $\omega > 1$. Because the modified likelihood penalizes against small weight on $\omega > 1$, it is not surprising that likelihood modification has no impact on likelihood estimation for the *tax* gene.

Table 2.2: Genes analyzed under models M1a and M2a without (C=0) and with (C=2) likelihood modification.

			p-value		Tree Length		$\hat{p}_2/\hat{\omega}_2$	
Gene	N_t	N_c	C=0	C=2	C=0	C=2	C=0	C=2
CDH3	11	176	1.40e-04	8.39e-03	0.56	0.54	0.00/24.57	0.08/2.01
mivN	5	504	1.54e-01	5.00e-01	1.62	1.60	0.00/5.95	0.07/1.00
pgpA	5	198	2.33e-02	5.00e-01	2.93	2.06	0.01/34.70	0.09/1.00
TrbL-VirB6_2	5	657	4.03e-01	5.00e-01	2.12	2.11	0.00/6.17	0.11/1.00
lysin	25	134	0.00e+00	0.00e+00	8.81	8.92	0.26/3.25	0.27/3.24
nuoL3	5	499	8.26e-14	9.63e-14	4.58	4.75	0.04/12.53	0.04/12.03
pol	23	947	4.33e-15	5.61e-15	1.31	1.32	0.02/5.59	0.02/5.14
RfaL	5	403	6.20e-06	7.89e-06	3.46	3.50	0.07/4.34	0.08/3.94
TrbL-VirB6_3	5	938	2.05e-09	2.36e-09	3.06	3.12	0.03/5.99	0.04/5.76
vif	29	192	2.86e-13	3.47e-13	2.90	2.95	0.08/3.56	0.10/3.43
β-globin	17	144	3.69e-03	5.84e-03	8.40	8.62	0.03/2.94	0.05/2.72
ccmF	5	635	2.54e-05	4.40e-05	3.41	3.28	0.01/15.47	0.03/8.41
ENAM	11	1142	7.66e-04	9.73e-04	0.46	0.46	0.02/5.69	0.08/3.41
env	13	91	2.59e-05	1.33e-04	2.04	2.03	0.18/3.63	0.33/2.79
perM	5	351	1.71e-01	2.16e-01	1.78	1.77	0.02/2.57	0.04/1.89
tax	20	181	4.17e-03	4.17e-03	0.13	0.13	1.00/4.87	1.00/4.87

N_t: number of taxa; N_c: sequence length in number of codons; p-value of the LR test for the presence of positive selection using a $\chi^2_0/2+\chi^2_1/2$ distribution; estimated total tree length; estimated proportion of sites evolving under $\omega > 1$: $\hat{p}_2/\hat{\omega}_2$. The top genes represent *irregular* estimation, the middle *regular*, and the bottom genes are uncategorized.

2.3.5 Real Data Results Show that Using Modified Likelihood Improves Estimation and Detection of Sites Under Positive Selection

Although site classification was not a focus of this study, evidence of positive selection at individual sites was checked in order to assess differences using the two likelihood approaches and three site classifiers. Spearman correlations for the site posteriors are summarized in Table 2.3 for NEB, BEB (Yang et al., 2005) and SBA,

Table 2.3: Spearman rank correlations of site posterior probabilities for different methods of classification under model M2a.

Gene	N*/N	N*/B	N*/S	B*/N	B*/B	B*/S	N/B	N/S	B/S
CDH3	0.40	1.00	1.00	0.40	1.00	1.00	0.40	0.40	1.00
mivN	0.76	0.99	0.97	0.77	1.00	0.96	0.77	0.78	0.96
pgpA	0.71	0.99	0.99	0.72	1.00	0.98	0.72	0.73	0.98
TrbL-VirB6_2	0.72	1.00	0.98	0.72	1.00	0.98	0.72	0.72	0.98
lysin	1.00	1.00	0.99	0.99	1.00	0.99	1.00	0.99	0.99
nuoL3	1.00	0.99	0.90	0.99	1.00	0.93	0.99	0.90	0.93
pol	0.94	0.96	0.79	0.91	1.00	0.85	0.91	0.76	0.85
RfaL	1.00	1.00	0.97	1.00	1.00	0.97	1.00	0.96	0.97
TrbL-VirB6_3	0.98	0.98	0.91	1.00	1.00	0.93	1.00	0.93	0.93
vif	1.00	1.00	0.97	1.00	1.00	0.98	1.00	0.97	0.98
β-globin	0.96	0.94	0.85	0.90	1.00	0.90	0.90	0.82	0.90
ccmF	0.93	0.93	0.87	0.86	1.00	0.96	0.86	0.80	0.96
ENAM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
env	0.99	1.00	0.99	1.00	1.00	0.99	1.00	0.98	0.99
perM	0.99	1.00	0.92	0.99	1.00	0.93	0.99	0.91	0.93

N: NEB; B: BEB; S: SBA; *: parameter estimation with modified likelihood. The top genes represent *irregular* estimation, the middle *regular*, and the bottom genes are uncategorized.

the smoothed bootstrap method (Mingrone et al., 2016, Chapter 3), each with and without likelihood modification. Site classification was nearly identical using BEB with both likelihood approaches. This is to be expected since BEB integrates over the uncertainties in the estimates of the ω distribution using discretized uniform and Dirichlet priors. Thus, the only parameters under BEB that differ with or without modified ML estimation are the edge-lengths and some parameters in the rate matrix, which tended to change much less than the parameters of the mixing distribution. By contrast, NEB directly uses the ML estimates of the mixing distribution, which differ considerably with and without likelihood modification. Consequently, site classification differs substantially under NEB with and without the modified likelihood. Given that previous studies have indicated that BEB and SBA do better than NEB at balancing accuracy and power for identifying sites under positive selection (e.g., Anisimova et al., 2002; Mingrone et al., 2016), the stronger agreement between BEB and SBA with NEB using modified likelihood than NEB without modified likelihood suggests modified likelihood is beneficial for detecting sites under positive selection.

2.3.6 Investigation of a Problematic Setting

Estimation and inference becomes more challenging with smaller evolutionary distances or fewer taxa, but, perhaps surprisingly, the results show that the true mixing distribution is at least as important for determining whether a setting is challenging. This is most evident in the CDFs for the null simulation settings using a 5-taxon tree of length 3 (figure 2.3). Note that the mixing distribution for all of these scenarios is determined by (p_0, ω_0). Overall, except for the $(p_0, \omega_0) = (0.25, 0.5)$ case, the modified LR statistic distribution is still well approximated by a $\chi_0^2/2 + \chi_1^2/2$ distribution, but for this one setting, neither the LR statistic nor the modified LR statistic distribution is well approximated by the $\chi_0^2/2 + \chi_1^2/2$ distribution.

Histograms of the ω_0 estimates under models M1a and M2a with modified likelihood show the largest variation when $(p_0, \omega_0) = (0.25, 0.5)$ (figure 2.5). Of the $10,000$ sets of modified likelihood MLEs, under M2a 2315 had 90% or more weight on an $\hat{\omega}_0 \geq 0.65$. Since the true mixing distribution had two well-separated ω values, $\omega_0 = 0.5$ and $\omega_1 = 1$, the expectation was that the estimated distribution would also have well-separated components with appreciable weight. The theory leading to the $\chi_0^2/2 + \chi_1^2/2$ approximation relies on this being highly likely with sufficiently large sequence lengths. It is clear from the simulations that sequence lengths of 500 are not long enough to guarantee well-separated components, which lead to the discrepancy for $(p_0, \omega_0) = (0.25, 0.5)$ in Figure 2.3. After removing the 2315 sets of modified MLEs that had 90% or more weight on an $\hat{\omega}_0 \geq 0.65$, the $\chi_0^2/2 + \chi_1^2/2$ CDF provides a good approximation to the actual CDF of the modified LR statistic (figure 2.6). This indicates that the estimates with $\hat{p}_0 \geq 0.9$ and $\hat{\omega}_0 \geq 0.65$ were the source of anomalously larger than expected LR statistics.

Whether a pre-screen would be useful for filtering out datasets with $\hat{p}_0 \geq 0.9$ and $\hat{\omega}_0 \geq 0.65$ was tested. The pre-screen considered was to ignore datasets failing to reject M1a in an M0 versus M1a test. While this was effective in that it filtered all the 2315 datasets described above, it also filtered many other datasets. Consequently, the distribution of M1a/M2a LR statistics remaining after the pre-screen was not as well approximated by the $\chi_0^2/2 + \chi_1^2/2$ CDF as the one with the 2315 datasets manually filtered (supplementary figure 6.9). As a second check, the data was re-simulated under the same settings, but with codon frequencies derived from abolone sperm lysin

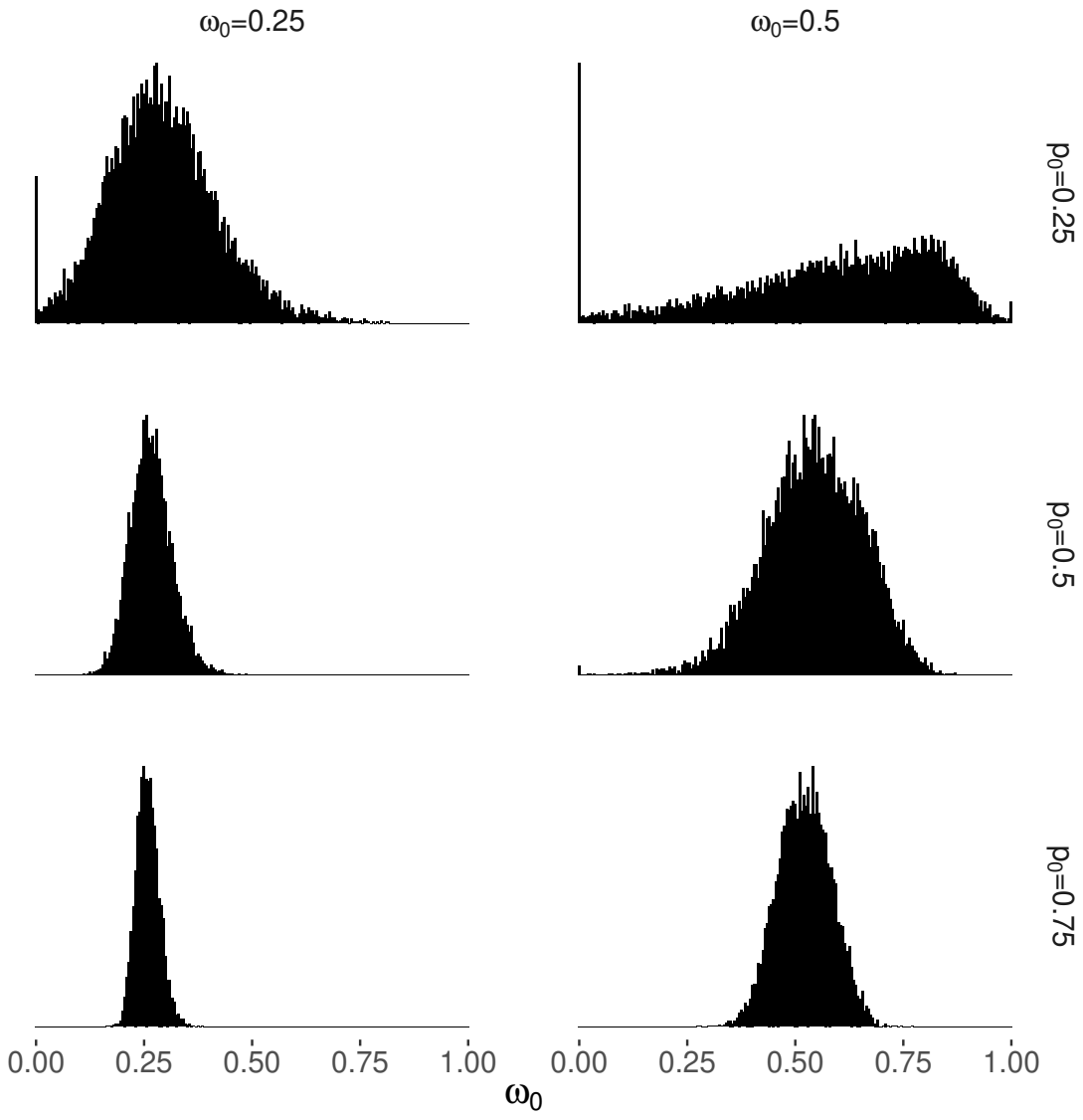


Fig. 2.5: MLEs of the ω_0 parameter under model M2a using a modified likelihood parameter of $C = 2$ for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a 5-taxon tree topology with branch lengths summing to 3. The value of ω_0 and its weight, p_0, used to simulate the data are shown as column and row labels.

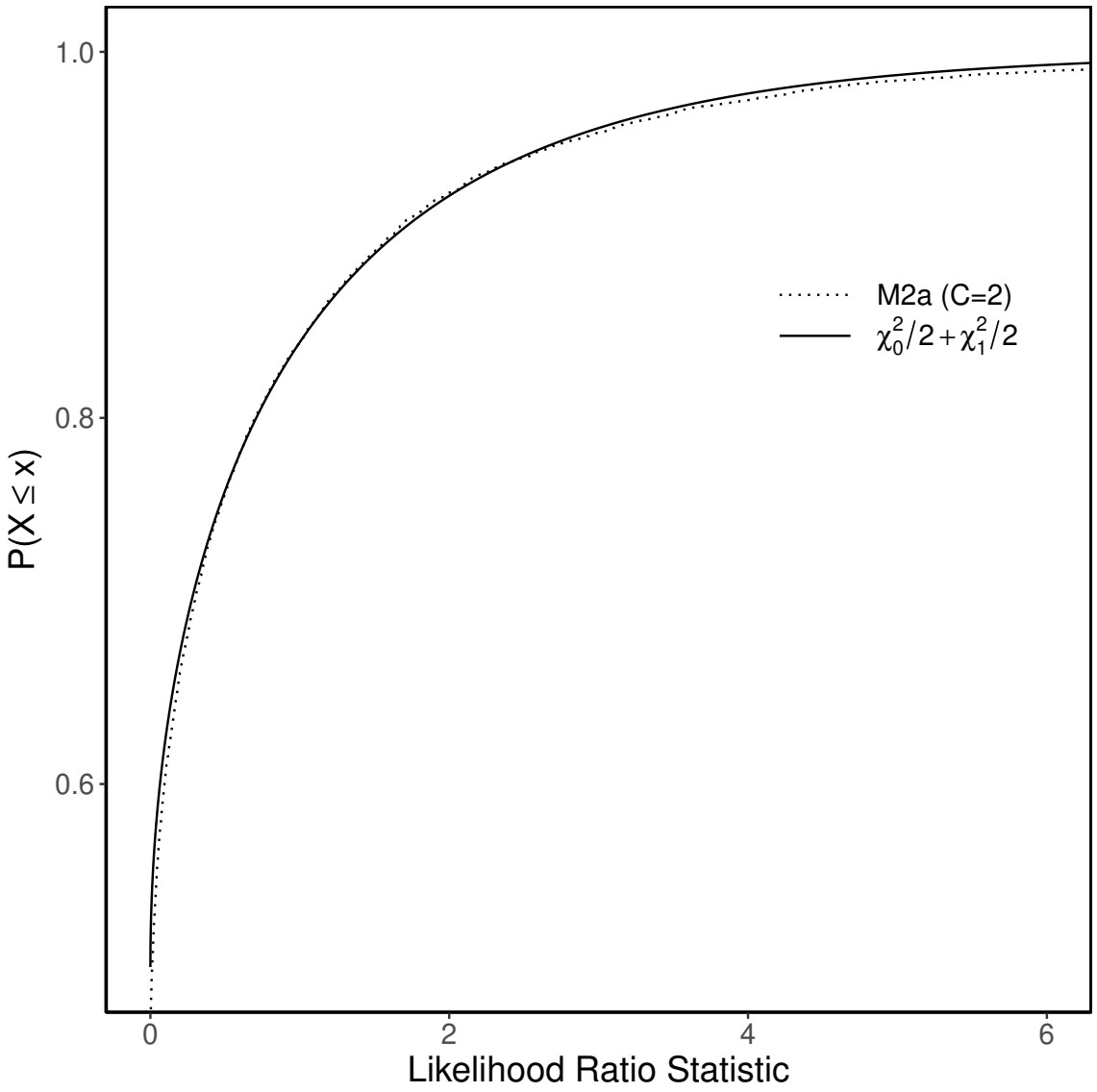


Fig. 2.6: CDF of filtered, modified LR statistics (C=2). The modified LR statistics were calculated under the nested model pair M1a/M2a for 10,000 simulated sequence alignments. The alignments were simulated with 25% of the sites evolving under $\omega = 0.5$ and the remaining sites evolving under $\omega = 1$ using a 5-taxon tree topology with branch lengths summing to 3. A modified likelihood tuning parameters of $C = 2$ was used and 2315 LR statistics associated with ML estimates with greater than 90% of the sites estimated in the $\omega < 1$ site class were excluded from the plot. A $\chi_0^2/2+\chi_1^2/2$ CDF is also included.

(Yang et al., 2000b), however the M1a/M2a LR statistic distribution was, again, not well approximated by a $\chi_0^2/2 + \chi_1^2/2$ CDF.

2.3.7 Parameters can be Almost Unidentifiable for Codon Models

To further investigate why $(p_0, \omega_0) = (0.25, 0.5)$ was a difficult setting, Kullback-Leibler divergences (KLs) were approximated between 5-taxa pattern distributions coming from $(p_0, \omega_0) = (0.25, 0.5)$ and pattern distributions from other mixing distributions (figure 2.7). When $KL = 0$, two mixing distributions give exactly the same pattern probabilities and the mixing distributions are said to be unidentifiable. When $KL > 0$ but small, distinguishing between the two mixing distributions will be difficult. Calculation of site likelihoods for all 61^5 site 5-taxa patterns is not feasible, so the KL values were approximated with $10,000$ simulated sites. Some of the approximated KLs in Figure 2.7 are close to 0, including those for $(p_0, \omega_0) = (0.5, 0.7)$ and $(p_0, \omega_0) = (0.75, 0.8)$. To determine whether the KL was indeed 0, attention was restricted to all site patterns for pairs of taxa to give tractable calculations. Calculation of all 61^2 site patterns is feasible and the KL_s, KL calculated using site pattern distributions for a subset of the 5 taxa, satisfies $KL_s \leq KL$. Thus, if any pair of taxa gives $KL_s > 0$, then the KL for all 5 taxa must be positive. For $(p_0, \omega_0) = (0.5, 0.75)$ this gives $KL > 0.00018$, the maximum KL over pairs.

Consideration of KLs for mixing distributions $(p_0, \omega_0) = (1, 0.75)$ and $(p_0, \omega_0) = (0.5, 0.5)$ allows it to be shown that there are ranges of distributions that are almost unidentifiable. The maximum KL over pairs of taxa from the 5-taxon tree was small (0.00085), thus

$$p(x; \omega = 0.75, \zeta) \approx p(x; \omega = 0.5, \zeta)/2 + p(x; \omega = 1, \zeta)/2.$$

Multiplying this equation by $p_0^{'}$ and rearranging, one can show

$$p(x; \omega = 0.75, \zeta)p_0^{'} + p(x; \omega = 1, \zeta)(1 - p_0^{'}) \approx p(x; \omega = 0.5, \zeta)p_0^{'}/2 + p(x; \omega = 1, \zeta)/(1 - p_0^{'}/2).$$

Thus, mixing distributions $(p_0, \omega_0) = (p_0^{'}, 0.75)$ give pattern probabilities that are difficult to distinguish from $(p_0, \omega_0) = (p_0^{'}/2, 0.5)$. This holds for the range of mixing distributions with $0 \leq p_0^{'} \leq 1$ (figure 2.7).

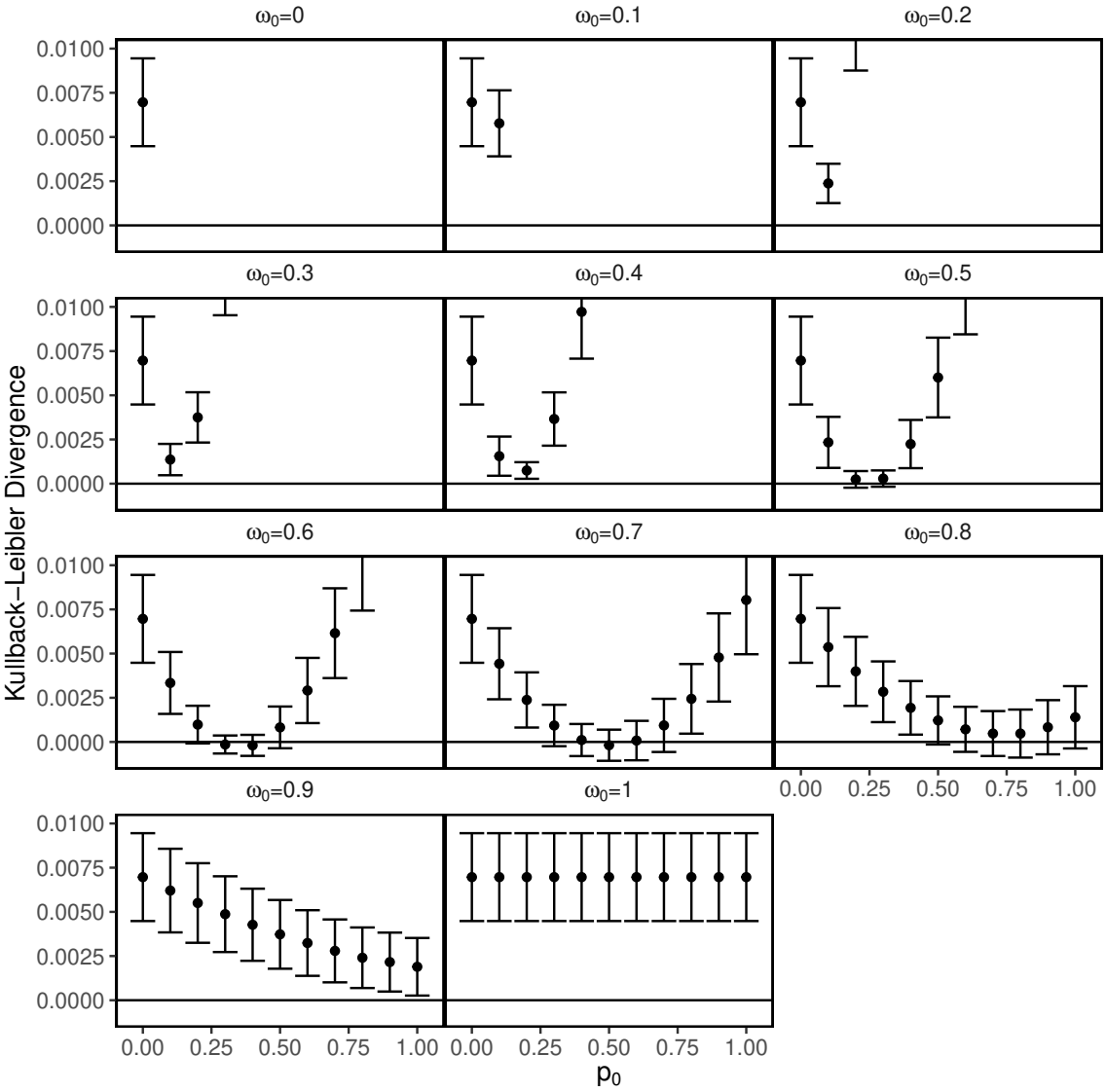


Fig. 2.7: Approximations of the Kullback-Leibler divergences between the distributions of site likelihoods for the generating model and other mixing distributions. The approximations were obtained as the mean lnL difference between 10,000 site patterns generated under model M1a using a 5-taxon tree with branch lengths summing to 3 and the mixing distribution $(p_0, \omega_0) = (0.25, 0.5)$, and other mixing distributions with varying weights on values of ω ranging from 0 to 1. Error bars for two standard errors $(s_{KL}/\sqrt{10000})$ above and below each Kullback-Leibler estimate are included. Points missing from each plot are above the visible range.

While it has been shown that there are regions of mixing-distribution parameter space that can make estimation and inference difficult, the results also show that there are regions where distinguishing between mixing distributions is not difficult, even for the $(p_0^{'}, 0.75) \approx (p_0^{'}/2, 0.5)$ comparison above. This is because $p_0^{'}/2 \leq 0.5$, so pattern probabilities generated from any $(p_0, \omega_0) = (p_0^{'}, \omega = 0.75)$ can not be consistent with e.g., $(p_0, \omega_0) = (0.75, \omega = 0.5)$ (supplementary figure 6.10). Finally, the good behaviour of the modified LR statistic when used with trees with more taxa and longer branch lengths (e.g., figure 2.2) suggests these problems are likely restricted to trees with fewer taxa and shorter branch lengths.

2.3.8 Concluding Remarks

Challenging issues have been described with mixture models of codon evolution that result in null distributions of LR statistics that are not tractable when testing for positive selection. A common violation of the regularity conditions under the widely employed M2a model is for small weight to be placed on $\omega > 1$. This results in LR statistics that, when compared to thresholds predicted by a $\chi_0^2/2 + \chi_1^2/2$ distribution, tend to give inflated false positive rates. By including a penalty in likelihood calculations for small weight on $\omega > 1$, in most cases, LR statistic distributions are well approximated by a $\chi_0^2/2 + \chi_1^2/2$ distribution and false positive rates are adequately controlled. Simulations under the alternative hypothesis show that modifying the LR statistic has minimal impact on power.

The likelihood modification may introduce small positive bias to the estimated weight on $\omega > 1$. But, the results here show that when there is signal in the data for the weight to be close to 0 and the sample size or the number of taxa are large, the contribution from the likelihood overwhelms the penalty term. In small sample cases, it may be difficult to overwhelm the penalty term in the modified LR test. However, when the weight on $\omega > 1$ is estimated to be close to 0 and the sample size is small, the standard LR tests are unreliable. Thus, whether samples sizes are small or large, and whether the weight on $\omega > 1$ is actually close to 0 or not, the modified LR test is appropriate.

For values of the tuning parameter, C, substantially smaller than 2, the behaviour of the modified LR statistic was too similar to that of the standard LR statistic. As values of C larger than 2 and up to 10 gave indistinguishable LR statistic CDFs to

those with $C = 2$, a simple, default value of $C = 2$ was used. Further improvements might be obtained with different data-dependent choices of C; the optimal choice is a topic for future research. A possible approach for making an optimal, data-dependent choice is cross-validation. One would choose C to make the average likelihood on test samples as large as possible.

The problematic behaviour of LR statistics discussed here arises more broadly with mixtures and is not usually amenable to solutions like those discussed in Self and Liang (1987). For instance, Chernoff and Lander (1995) outline a class of problems using a mixture of binomial distributions to evaluate genetic markers for genes representing heterogeneous traits. Similar to how any $\omega_2 > 1$ with a weight of $p_2 = 0$ gives the null under model M2a, any mixing distribution gives the null when the parameters of the two binomials are estimated to be the same. For many of the settings considered, no closed form expression for the distribution was available and, by contrast with usual χ^2 distributions, depended on unknown parameters in the true model under the null hypothesis. Generally, it appears that when dealing with mixtures in molecular evolution settings, simple limiting distributions like those of Self and Liang (1987) are not likely without some modification to the likelihood.

A possible alternative to using a modified likelihood is to estimate the LR statistic distribution using a parametric bootstrap. Parameters estimated from the data are used to simulate B parametric bootstrap samples. The LR statistic distribution is then estimated from the B bootstrap samples and $P(X_{LR} \leq x; \hat{\theta}_0)$, the probability that the LR statistic is less than a threshold determined from the estimated LR statistic distribution is calculated, where $\hat{\theta}_0$ are the MLEs under the null hypothesis. As a function of θ, $P(X_{LR} \leq x; \theta)$ is expected to be continuous because the probabilities of individual site patterns are continuous functions of θ. So if the null hypothesis is true and θ_0 is the true parameter, then when $\hat{\theta}_0 \approx \theta_0$, then $P(X_{LR} \leq x; \hat{\theta}_0) \approx P(X_{LR} \leq x; \theta_0)$, the appropriate null distribution for the LR statistic. There are potential issues using the parametric bootstrap. With small samples, $\hat{\theta}$ might occasionally be far from θ_0. The extent to which this is a problem depends upon how much $P(X_{LR} \leq x; \theta)$ varies as a function of θ. An additional, more serious difficulty when $P(X_{LR} \leq x; \theta_0)$ varies with θ_0, is that when the alternative hypothesis is true, the value of $\hat{\theta}_0$ is not clearly meaningful. This is one reason for

preferring a test statistic like the modified LR that is expected to have a distribution that does not vary too much under the null hypothesis.

The simulation results expose an additional difficulty. For certain generating mixing distributions like $(p_0, \omega_0) = (0.25, 0.5)$ under model M2a, there are other mixing distributions that can give very similar pattern probabilities when the number of taxa is small and edge lengths are short. Models that are almost unidentifiable will give flat likelihood surfaces for subsets of parameters. Perturbations caused by unwanted influences such as sequencing error and model misspecification will be more pronounced than on peaked likelihood surfaces. Thus, the identifiability problems exposed when edge lengths are short might, for instance, help to explain the elevated false positive rates found by Schneider et al. (2009) in the presence of sequencing error and short edge lengths, and the finding of Venkat et al. (2018) that model-misspecification in the human lineage led to higher false positives.

That some models were found to be almost unidentifiable suggests that there may be mixing distributions and trees that lead to a complete lack of identifiability. There has been some work to explore identifiability of mixture models of molecular evolution (e.g., Allman et al., 2008; Allman and Rhodes, 2009; Chai and Housworth, 2011), but none of these results apply directly to codon models. Determining the extent to which these types of issues affect codon models of evolution will be a topic for future research.

Implementation of a similar modified likelihood approach to other models is straightforward and can be expected to offer similar advantages. As there is an abundance of mixture models used to solve a variety of problems in the field of molecular evolution, potential candidates are numerous and include models used to 1. identify functionally divergent protein residues, 2. detect pattern-heterogeneity in gene sequence or character-state data, 3. infer protein phylogenies, and 4. identify across-site heterogeneities in the amino-acid replacement process (Gaston et al., 2011; Pagel et al., 2004; Wang et al., 2008; Lartillot and Philippe, 2004).

Chapter 3

Smoothed Bootstrap Aggregation

This work was published in the journal *Molecular Biology and Evolution* (Mingrone et al., 2016).

3.1 Introduction

Identifying positively selected amino acid sites is a challenging statistical task that is important for investigating the functional consequences of molecular change (Yang, 2005). Several approaches have been developed to detect positive selection within a protein (reviewed in Pond and Frost, 2005; Anisimova and Kosiol, 2009), but their reliability varies according to the properties of the data in hand. The most widely used methods employ a codon model to detect an excess in the rate of nonsynonymous substitutions relative to synonymous substitutions ($dN/dS = \omega > 1$), which is an indication of evolution by positive selection. Proteins evolving under positive selection must retain the capacity to fold into complex structural and functional domains, so the majority of amino acid substitutions will be subject to purifying selection pressure, with $\omega < 1$ (Kimura, 1968). From extensive surveys of positive selection in real genes, the expectation is that only a small fraction of amino acid sites will be subject to adaptive change and exhibit an $\omega > 1$ (e.g., Anisimova et al., 2007; Ge et al., 2008). The sparseness of these sites makes them challenging to identify.

Two general categories of methods for detecting positively selected amino acid sites include counting and fixed-effect methods. Counting methods employ ancestral reconstruction of codon states for all internal nodes of a phylogenetic tree to obtain counts of the synonymous and nonsynonymous changes along each of its branches. The counts inferred for a given site are used to test if $\omega \neq 1$. Some counting methods

use parsimony (Fitch et al., 1997; Bush et al., 1999; Suzuki and Gojobori, 1999), and others likelihood (Suzuki, 2004; Nielsen, 2002; Nielsen and Huelsenbeck, 2002; Suzuki and Nei, 2004; Pond and Frost, 2005) to infer the ancestral codon states. The reconstructions are often similar, but under the likelihood approach uncertainty about the inference can be summarized via the posterior probabilities of the ancestral states. Thus, the parsimony based methods must assume that these uncertainties are irrelevant to the statistical test. While this makes the approach attractive for very large datasets where reliable reconstructions can be obtained relatively quickly (Lemey et al., 2012), widespread use is hindered by a lack of power when the level of divergence is too low or by the negative impact of substitutional saturation when the level of divergence is too high (Pond and Frost, 2005).

An alternative approach is to treat each site as independently relevant to the question of evolution by positive selection, and attempt to fit an ω parameter to the data at each site. Thus, the effect of each site on the task of ω inference is fixed. Model based testing for $\omega \neq 1$ can be carried out via a standard LR test, and no assumptions are required about the distribution of selection pressure, ω. Although ω is treated as a site-specific variable, other important variables in the codon model (e.g., branch lengths) are shared among sites, with their values estimated jointly from the complete set of sites. Results obtained by using these modelling ideas (Pond and Frost, 2005; Massingham and Goldman, 2005) are encouraging, and it is expected that this family of methods will continue to have a role in real data analyses (Scheffler et al., 2014). However, χ^2 approximations to the distribution of the test statistic assume relatively large numbers of taxa, which is often not the case. The lack of independence of data across taxa that is due to phylogeny creates further difficulties for χ^2 approximations.

A third approach for detecting positive selection at amino acid sites, which is the focus of this chapter, treats the value of ω at a site as the realized value of a random variable. A particular model for the distribution of ω is chosen and ML is used to fit the distribution to the data as part of an explicit model of codon evolution. There are recommendations (e.g., Yang and Nielsen, 1998) to use a pre-screen that fits two models: one with a distribution that excludes values of $\omega > 1$, and another with the same distribution, except with weight on values of $\omega > 1$ permitted. This nested-model pre-screening is used to test if the data conveys any evidence of positive

selection. When the null hypothesis of no positive selection is rejected using a LR test, site-wise analysis is warranted. Site-wise analysis is carried out using Bayes rule to calculate the posterior probability that a site h evolved under some estimated value of ω, given the data at site h. This approach is referred to as empirical Bayes (EB) because the marginal distribution of ω is determined from the data. Conclusions regarding the evolution at a site are made based on the estimated ω-values along with their associated posterior probabilities conditioned on the data at the site. For example, when the largest posterior probability for a site is associated with a value of $\omega > 1$, this is taken as evidence of positive selection at that site.

Because the marginal distribution of ω is determined from the data and the site posterior probabilities always depend on the fitted values of the model parameters (shape parameters of the distribution, edge lengths, etc.), the reliability of EB inference depends on the accuracy of the fitted values. If they have been accurately estimated, as is often the case with large, information-rich datasets, they can simply be treated as known without errors. This approach is known as the NEB (Nielsen and Yang, 1998). However, when the fitted values are subject to large errors, the detection of positive selection according to the posterior probabilities can be negatively impacted and in some cases the false positive rate can be unacceptably high (Wong et al., 2004). BEB has been used to adjust for uncertainty in the parameters of the ω distribution by assigning priors to those parameters and using numerical integration to average over the uncertainty represented by the priors (Yang et al., 2005). Because this tactic can substantially reduce the false positive rate relative to NEB in problematic datasets, BEB has become a popular method for inferring the action of selection at individual sites. A fully Bayesian approach that also assigns priors to edge-lengths and other parameters is available for the inference of positive selection at sites (Aris-Brosou, 2003; Huelsenbeck and Dyer, 2004), but it is not as widely employed as EB because it is available for a limited set of models.

BEB does have limitations. As currently implemented, the BEB approach only accommodates uncertainty in the parameters of the ω distribution, leaving all others fixed to their fitted values. Furthermore, only uniform priors are used, which means the adjustment for uncertainty is independent of the signal in the data. Although these will not be serious limitations for many analyses of real data, it is shown here

through simulation and real data analysis that deriving the adjustment for parameter uncertainty from the data can improve inference for some datasets. To avoid the need for priors, a new approach is presented here that uses bootstrapping (Efron, 1979, 1982) of site patterns to simulate dataset variability and adjust for the uncertainty in the data. From bootstrap datasets, the distribution of the MLEs can be estimated. The posterior probabilities for positive selection at a site is then obtained using an aggregate value coming from MLEs over bootstrapped data sets, rather than according to a single posterior probability obtained under NEB or BEB. In principle, bootstrap-based methods should use as many replicates as possible to approximate the infinite-sample bootstrap distribution. As this is computationally expensive, smoothing techniques borrowed from kernel density estimation (Silverman and Young, 1987; Davison and Hinkley, 1997, Section 3.4) are used to obtain an approximation with less computational cost. I refer to this new approach as SBA. Simulation results show that SBA balances accuracy and power at least as well as BEB.

The behaviour of ML estimation is also investigated when standard regularity conditions, such as the requirement for true parameter values to be in the interior of the parameter space, are not met. Codon models fit ω distributions that, for some data-generating settings, violate regularity conditions, which leads to substantial instability in parameter estimation. These instabilities have a negative impact on the inference of positive selection under EB, and it is shown here that the new approach is an improvement over both NEB and BEB in such cases. Also shown here is that results previously reported for the *tax* gene of HTLV (Suzuki and Nei, 2004) are likely a consequence of such instabilities. The *tax* gene is a well known example where EB is widely considered unreliable, and the results obtained for the *tax* gene have been used to criticize the overall approach. An explanation is provided for the past results obtained under EB methods for the *tax* gene. The SBA method can help diagnose such dubious inferences.

3.2 New Approaches

A new approach for classifying sites called SBA is presented here. SBA uses bootstrapping and kernel smoothing techniques to accommodate uncertainties in MLEs. Site patterns from a sequence alignment are sampled with replacement to create a number of bootstrap sequence alignments. For each of the bootstrap sequence

alignments, MLEs are calculated. The usual bootstrap distribution is the empirical distribution of the calculated MLEs. To avoid difficulties due to 1) low information content in the data, 2) necessarily limited bootstrap sampling and 3) instabilities in the parameter estimates, a kernel density estimate of the bootstrap distribution coming from the MLEs is instead used. The smoothness of the distribution is controlled by a bandwidth parameter, which is set larger than conventional values to give greater smoothing.

While typical applications of bootstrapping use MLEs to calculate confidence intervals and standard errors, we, instead, use the bootstrap to accommodate uncertainty in the posterior probabilities of positive selection at sites. For any given site in the original sequence alignment, many parameter values are generated from the smoothed bootstrap distribution and substituted into posterior probability formulas to give a distribution of posterior probabilities which reflects parameter uncertainty. The mean or median of these posteriors is a more stable estimate of the true posterior and is used for classification. See figure 3.1 for an overview of SBA.

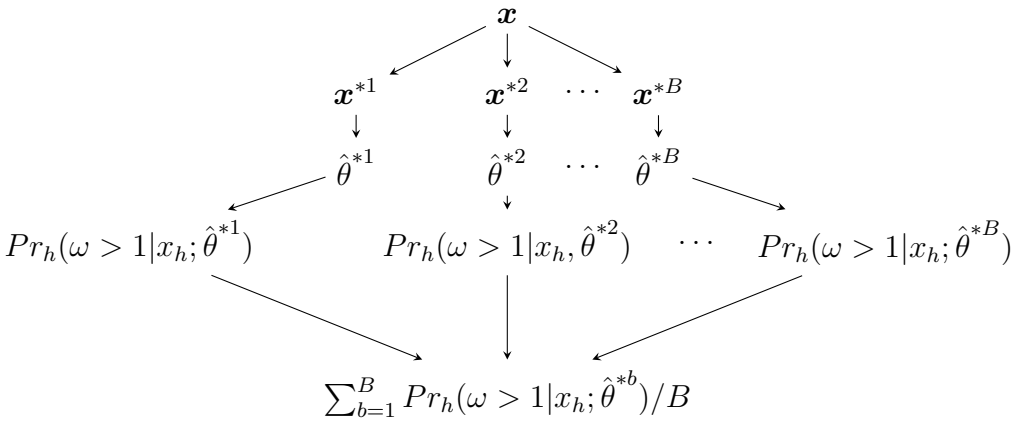


Fig. 3.1: Bootstrapping site patterns in a codon sequence alignment to classify selection pressure at codon sites. From an alignment of protein coding DNA sequences, $\boldsymbol{x}$, with n codon sites, site patterns are randomly sampled with replacement to obtain a bootstrap sample, $\boldsymbol{x}^{*b}$ with n sites. MLEs, $\hat{\theta}^{*b}$, are then estimated for bootstrap sample $\boldsymbol{x}^{*b}$. Using $\hat{\theta}^{*b}$ and $\boldsymbol{x}$, the posterior probability $Pr_h(\omega > 1|x_h; \hat{\theta}^{*b})$, that site h is under positive selection is calculated. These steps are repeated B times to calculate B sets of posterior probabilities. An aggregate posterior probability that site h is under positive selection is calculated by, for instance, averaging posterior probabilities over bootstrap replicates, $\sum_{b=1}^{B} Pr_h(\omega > 1|x_h; \hat{\theta}^{*b})/B$.

3.3 Results

3.3.1 Non-standard ML Estimation Behaviour

Parameter estimation by ML has attractive statistical properties, including consistency, efficiency, and asymptotic normality, when certain regularity conditions hold (Kalbfleisch, 1985; Bickel and Doksum, 2006). For settings where regularity conditions hold, I verified that I could obtain well-behaved estimates of the parameters of the ω distribution under two commonly used codon models: M2a (Nielsen and Yang, 1998; Yang et al., 2005) and M8 (Yang et al., 2000a). I simulated 100 datasets representing a *regular* estimation problem with an ω distribution having at least 10% weight on each site class (45% $\omega = 0$, 45% $\omega = 0.5$, and 10% $\omega = 5$). As expected, MLEs obtained from these data under both M2a and M8 have unimodal and symmetric distributions (figure 3.2a,b). For the estimates in this *regular* case, there are no indications of departures from the limiting properties predicted by ML theory.

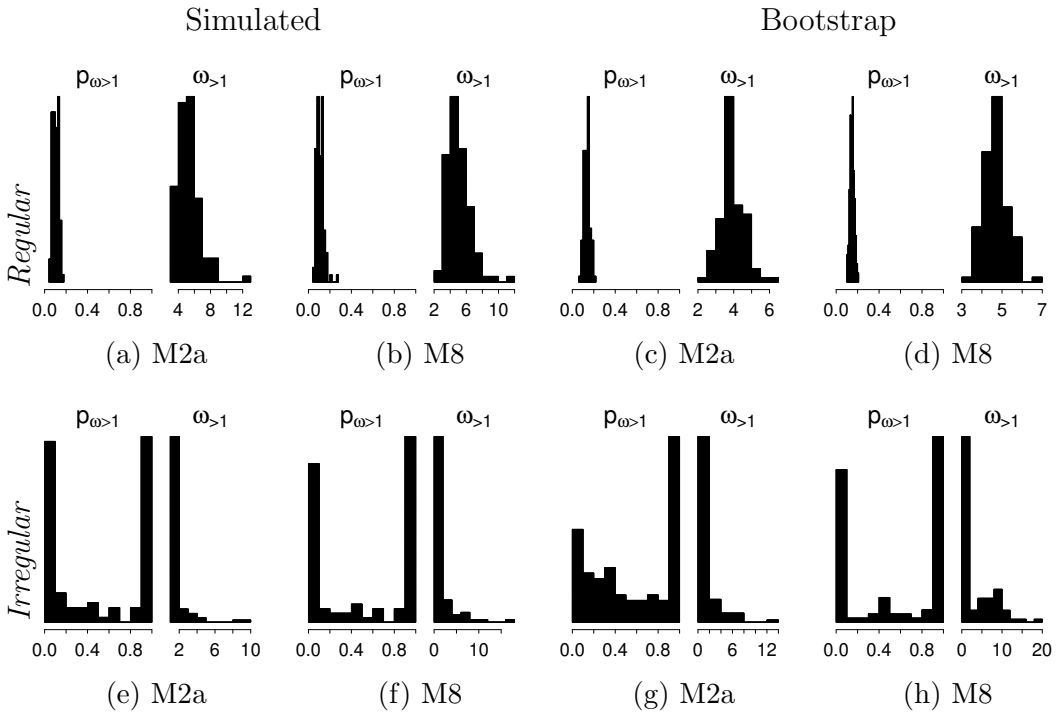


Fig. 3.2: MLE distributions of the $p_{\omega>1}$ and $\omega_{>1}$ parameters under M2a and M8. Histograms are over 100 simulated (a,b,e,f) and bootstrap (c,d,g,h) datasets with the bootstrap datasets generated by sampling from one simulated dataset. Data were simulated under *regular* (a - d) and *irregular* (e - h) conditions.
regular simulation conditions: 5 taxa, 45% ω=0, 45% ω=0.5, and 10% ω=5
irregular simulation conditions: 5 taxa, 100% ω=1

The regularity condition requiring true parameter values to be in the interior of the parameter space is sometimes violated when using codon models. For such parameter settings, instabilities or departures from the expected limiting properties of ML estimation can arise including non-Gaussian and over-dispersed distributions of estimates. To investigate instabilities under models M2a and M8, I simulated 100 datasets representing an *irregular* estimation problem with sparse information, i.e., 100% of the sites at the threshold for positive selection, $\omega = 1$. In figure 3.2e,f, in contrast to the results presented in figure 3.2a,b, there are instabilities in the MLEs for the parameters representing the proportion of sites under positive selection, $p_{\omega>1}$. The $p_{\omega>1}$ parameter distributions under both models have mass concentrated on both the lower and upper boundaries of the parameter space, and the distributions of the corresponding $\omega_{>1}$ parameters are concentrated on the lower boundary. Application of the LR test to filter datasets that convey no evidence of positive selection did not prevent instabilities. The null hypothesis of no positive selection was rejected for 10 datasets under M2a and 9 under M8, however, the MLE distributions after applying this pre-screening step remained unstable (supplementary figure 6.11).

Some of the model M2a MLE instabilities shown in figure 3.2e,f are due to the discrete ω distribution. True discrete distributions of interest can lie on the boundary of the parameter space, which is a regularity condition violation that gives rise to MLE instabilities. For instance, consider data generated from an ω distribution with no mass on $\omega > 1$. Estimates of the ω distribution will tend to approximate the true distribution and one way this can occur under M2a is when $\hat{\omega}_{>1} \approx 1$. When this happens, the likelihood will remain approximately constant over all choices of $p_{\omega=1}$ and $p_{\omega>1}$, giving a sum, $p_{\omega>1} + p_{\omega=1}$, that is approximately the same as that of the MLE. Consequently, estimates of $p_{\omega=1} + p_{\omega>1}$ are stable, but estimates of $p_{\omega=1}$ and $p_{\omega>1}$ are not, because many different choices give the same sum. Likewise, when a $p_{\omega>1}$ parameter is estimated near 0, the corresponding $\omega_{>1}$ can take on almost any value without changing the likelihood. For example, two M2a and six M8 biologically unrealistic estimates of the $\omega_{>1}$ parameter (e.g. $\omega_{>1} = 999$) occurred when the corresponding $p_{\omega>1}$ parameters were estimated to be 0. These estimates were excluded from the $\omega_{>1}$ histograms. For the data representing an *irregular* estimation problem with all sites simulated with $\omega = 1$, two other problematic M2a parameterizations

that fit the data equally well occurred often. First, all the weight was put on the w_1 category and second, all the weight was put on the $w_{>1}$ category when it was estimated very close to 1. Although there is virtually no difference in the likelihood scores between the two parameterizations, the NEB posterior probabilities for positive selection were 0 and 1 respectively. These different MLE instabilities arose with two general types of simulation settings: 1) when fewer site classes were simulated than exist in the fitted model, and 2) when different site classes were simulated with similar levels of selection pressure.

When working with real data, often only a single sample is available and alternative techniques must be used to approximate distributions of parameter estimates. One such technique is the bootstrap. I used the new bootstrap-based approach with sequence alignments to investigate properties of the MLE distributions and to detect settings where inference tends to be problematic (see Methods). While sampling with replacement from a single sample leads to a bootstrap parameter distribution that is a jagged estimate of a smooth distribution, I found the bootstrap, in many cases, can effectively estimate the distributions of MLEs. Figure 3.2c,d shows the distribution of the ω MLEs associated with positive selection generated over 100 bootstrap samples of a *regular* dataset. Note the resemblance of the bootstrap distributions in figure 3.2c,d to the analogous distributions over simulated datasets in figure 3.2a,b. A comparison of figures 3.2e,f and 3.2g,h illustrates that when the distribution over multiple samples is problematic, so too is the distribution over bootstrap samples. Among the 100 bootstrap MLE distributions obtained from the datasets simulated under *irregular* model conditions, I identified 91 of the M2a and 95 of the M8 $p_{\omega>1}$ parameter distributions as unstable using the criterion that at least 5% mass lies both below 0.2 and above 0.8. These distributions indicate that the mixture distribution for ω "flip-flops" between few and many sites in a positive selection class. Recall that under the generating model for these data, no sites are under positive selection. Plots of the other parameters of the ω distributions can be found in supplementary figure 6.12. Scenarios when the bootstrap distribution is not a good estimate of the true distribution of parameter estimates has been described in other settings, e.g., Efron and Tibshirani (1994, p. 81). So, while the bootstrap alone can be helpful for identifying problems, it is not always a robust solution for deriving a correction for

parameter uncertainty.

3.3.2 Kernel Smoothing Improves the Bootstrap-based Method for Approximating MLE Distributions

To avoid results that are a consequence of randomness due to bootstrapping, it is beneficial to choose the number of bootstrap samples, B, large enough so that the finite-sample bootstrap distribution approximates the infinite-sample bootstrap distribution well. However, when regularity conditions are violated there is no guarantee that even the infinite-bootstrap distribution provides an adequate assessment of the variability of an MLE. I tested this assertion under codon models where the distributions of the $p_{\omega>1}$ parameters were unstable over simulated and bootstrap datasets. For the data representing *irregular* model conditions described above, I generated 10,000 bootstrap datasets for each of the first 10 simulated datasets. The instabilities that characterize these 10 bootstrap distributions were largely unchanged by increasing B (supplementary figure 6.13). Similar difficulties arise in a variety of bootstrap applications. As a simple example of the phenomenon, suppose interest is in θ from a binomial distribution with small n and small θ. It is possible to sample almost all zeros, in which case the variance of the bootstrap distribution of θ estimates will be too small. Such boundary issues related to small samples can similarly be problematic for ω distributions when estimated weights are close to 0.

I used kernel smoothing along with bootstrapping to characterize the uncertainty in MLEs under *difficult* estimation conditions. Kernel smoothing is typically used to approximate the infinite-sample distribution more effectively when using a smaller number of bootstrap samples. However, the standard application of this technique (Davison and Hinkley, 1997, p. 79) was not sufficient when the MLEs were unstable. For such cases, *over smoothing* (i.e., using a larger than typically considered optimal bandwidth) was necessary to obtain conservative estimates of the MLE distributions with larger variance that suppressed the influence of the instabilities (supplementary figure 6.14). By over-smoothing the p parameters of codon models M2a and M8 with a uniform kernel I compensated for 1) low information content in the data, 2) fewer bootstrap samples, and 3) instabilities in the parameter estimates. For this reason I included over-smoothing of the p parameters in all applications of SBA.

3.3.3 Simulation Results

I used simulation to compare the performance of SBA with BEB and NEB. The design of the studies was motivated by the more challenging schemes of Wong et al. (2004) and Yang et al. (2005), however it extends theirs to investigate performance under progressively more model misspecification. The design is divided into three scenarios covering three levels of model misspecification. The *Correct Model Scenario* is comprised of four simulation studies (studies 1-4) where the nuisance parameters of the generating model were freely estimated by the fitted model. The ω distributions used to generate the datasets are listed in the third column of table 3.1.

Table 3.1: Simulation design and false positive rates under NEB, BEB, and SBA each with models M2a and M8.

Study	Misspec.	ω distribution	NEB		BEB		SBA	
			M2a	M8	M2a	M8	M2a	M8
1	None	100% 1	**0.34**	**0.35**	0.00	0.00	0.00	0.00
2	None	50% 0.5, 50% 1	0.00	0.00	0.00	0.00	0.00	0.00
3	None	50% 1 50% 1.5	**0.35**	**0.37**	0.00	**0.05**	0.00	0.02
4	None	45% 0, 45% 1, 10% 5	0.00	0.00	0.00	0.01	0.00	0.00
5	Mild	100% 1	**0.20**	**0.37**	0.00	**0.24**	0.00	**0.13**
6	Mild	50% 0.5, 50% 1	0.00	**0.13**	0.00	**0.11**	0.00	0.02
7	Mild	50% 1, 50% 1.5	**0.30**	**0.30**	0.00	**0.39**	0.00	**0.12**
8	Mild	45% 0, 45% 1, 10% 5	0.00	0.04	0.00	**0.12**	0.00	0.00
9	Heavy	100% 1	**0.71**	**0.71**	**0.55**	**0.62**	**0.13**	**0.52**
10	Heavy	50% 0.5, 50% 1	**0.53**	**0.50**	0.00	0.00	0.00	0.01

Each study used 100 simulated alignments and a 5-taxon tree with branch lengths summing to 3. A posterior probability threshold of 0.95 was used for classifying sites to be under positive selection. Under SBA, $B = 100$ bootstrap samples were generated and smoothing was carried out using a uniform kernel with a bandwidth parameter $h = 0.4$.

This scenario design matches selected schemes in Yang et al. (2005). The *Mild Misspecification Scenario* uses the same ω distribution as the first scenario as the basis of four additional studies (studies 5-8), but includes mild misspecification of the nuisance parameters (see Methods). Lastly, the *Heavy Misspecification Scenario*, includes two studies (studies 9-10) with heavy misspecification for the fitted model, which represents a more plausible scenario for the analysis of real sequences. In one

study (study 9) the data were simulated using the highly biased codon frequencies from the Drosophila *GstD1* gene (Bielawski and Yang, 2005). In the second study (study 10), the generating model is based on a 50/50 mixture of two heterogeneous classes of sites. One class was generated using equal codon frequencies, $\kappa = 1$, and $\omega = 0.5$, while the other used the Drosophila *GstD1* gene codon frequencies, $\kappa = 8$, and $\omega = 1$. For all 10 simulation studies I simulated 100 alignments, each having 500 codons, using the same 5-taxon tree from Wong et al. (2004). Under SBA, $B = 100$ bootstrap samples were generated and smoothing was carried out using a uniform kernel with a bandwidth parameter $h = 0.4$. The studies in the *Correct Model Scenario* were repeated under model M2a with the 30-taxon tree from the same paper.

Table 3.1 lists the false positive rates (proportion of sites inferred positively selected among those that are not) using a posterior probability cutoff of 0.95 for NEB, BEB, and SBA under models M2A and M8. In study 1 (no misspecification of the nuisance parameters and all sites simulated using $\omega = 1$) under NEB there is false positive detection of positive selection, whereas under BEB and SBA there is none. This is expected; NEB is known to yield unreliable posterior probability calculations in small datasets (e.g., Anisimova et al., 2002; Yang et al., 2005). Because the conditions of study 1 yield unstable parameter estimates (figure 3.2e-h), the false positives under NEB reflect more than mere sampling errors. MLE instabilities cause large $p_{\omega>1}$ to occur too often and these values lead to high posterior probabilities for positive selection under M2a and M8. The posterior probability calculations under SBA and BEB are reliable because those approaches do not assume the MLEs have been estimated without error. Yang et al. (2005) suggests that with more data, the problems with NEB controlling false positives can be mitigated. However, the MLE instabilities persisted in study 1 using a tree topology with 30 taxa (supplementary figure 6.15), indicating that large sample sizes do not always ensure accurate predictions.

Relative to simulations with a single $\omega = 1$ (study 1), when the ω distribution was 50% $\omega = 0.5$ and 50% $\omega = 1$ (study 2), the overall signal for positive selection was diminished and all false positive rates were 0. Conversely, when the ω distribution was 50% $\omega = 1$ and 50% $\omega = 1.5$ (study 3) there was a slight increase in the NEB false positive rates relative to study 1. Under M2a the false positive rates were 0

using BEB and SBA, but under M8 they increased to 0.05 using BEB and to 0.02 using SBA. For study 4, because the simulated ω values for the three sites classes were far enough apart, the false positive rates were well controlled.

The introduction of mild model misspecification of the nuisance parameters did not result in higher false positive rates under M2a, but did under M8. For studies 5-8, the BEB false positive rates (using a 0.95 posterior probability threshold) under M8 increased in all four cases relative to the corresponding studies (1-4) in the *Correct Model Scenario*. The same SBA false positive rates only increased in two cases and by smaller amounts than with BEB. When heavy model misspecification was introduced in the third scenario, NEB failed to adequately control false positives with rates between 50 and 71% under both M2a and M8. BEB and SBA also did not control the false positive rates in study 9, but did in study 10.

The results in table 3.1 are over all sites in all simulated datasets. After applying LR tests at the 0.05 level to filter datasets that convey no evidence of positive selection, none of the false positive rates under BEB or SBA changed. Supplementary table 6.3 gives the false positive rates under NEB after the adjustment. With the exception of two cases, the effect is minimal. Interestingly, under the null hypothesis, the false positive rates of the LR tests were larger than expected, particularly with model misspecification.

When testing for positive selection, we aim for large true positive rates, the proportion of sites truly under positive selection that are correctly identified, sometimes referred to as power. A difficulty in comparing methods for detecting positive selection is the choice of threshold. Lower thresholds tend to increase the true positive rate, but tend to also increase the false positive rate. To ensure that comparisons of power for different methods correspond to the same false positive rate I used receiver operator characteristic (ROC) curves, a convenient way to visualize the balance between accuracy and power for classification problems. Each point on a curve represents a threshold for the posterior probability of positive selection. Figure 3.3 shows ROC curves for each of the simulations that included positive selection in the generating model (studies 3, 4, 7, and 8). Curves are also included for the classification of sites using the generating parameters, i.e., the MLEs are fixed to the simulated

values. These curves represent an expected upper limit in performance of site classification (supplementary section 6.4). The lower limit for classification, when each site is randomly identified to be under positive selection, is represented by a $y = x$ line.

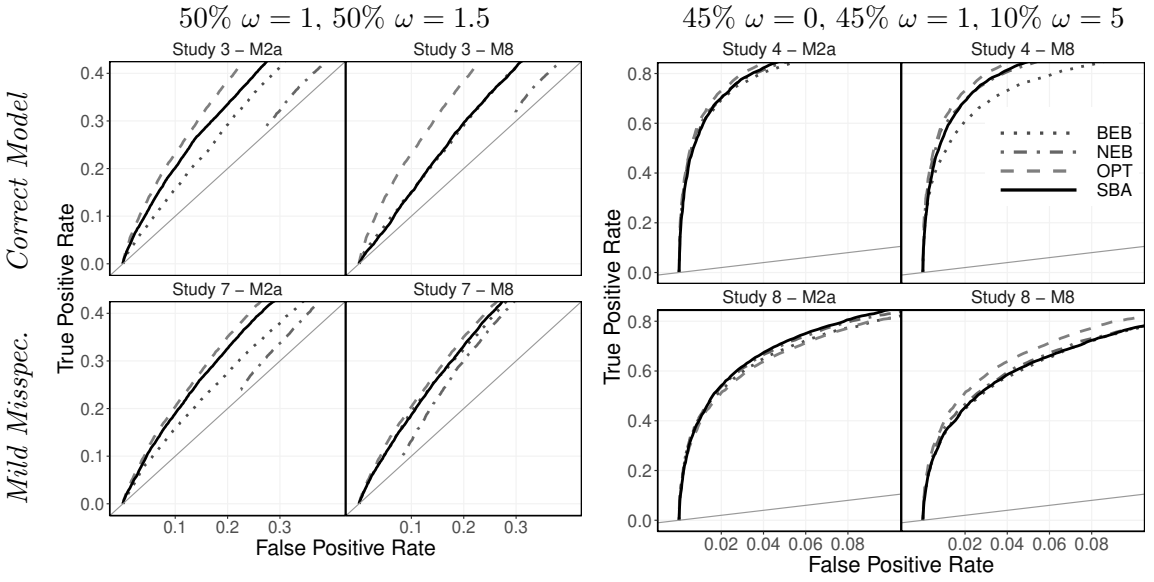


Fig. 3.3: ROC curves for the detection of sites under positive selection for BEB, NEB, and SBA analyses of data generated under two different simulation scenarios: without model misspecification (*Correct Model*, studies 3 and 4) and with mild model misspecification (*Mild Misspecification*, studies 7 and 8). The data were simulated using a 5-taxon tree topology. In studies 3 and 7, 50% of the sites were simulated under neutral evolution ($\omega = 1$) and 50% of the sites under positive selection ($\omega = 1.5$). In studies 4 and 8, 45% of the sites were simulated under purifying selection ($\omega = 0$), 45% under neutral evolution ($\omega = 1$) and 10% under positive selection ($\omega = 5$). Each plot includes a line for the lower bound (y=x) and an expected upper bound (OPT) when classification is made using the generating model parameters. Curves for NEB do not always cover the whole range of false positive rates, because NEB sometimes estimates the ω distribution with all mass on $\omega > 1$. In these cases, even with a posterior probability cut-off of 1, NEB still incorrectly classifies sites to be under positive selection.

The introduction of mild misspecification made the task of detecting sites under positive selection more difficult in study 8. This is evident from the shifting of the ROC curves down and to the right (lower rates of true positives for a given false positive rate) in study 8 relative to the corresponding simulations without the misspecification of the nuisance parameters in study 4. The same effect was not observed

between the ROC curves of studies 3 and 7.

In all cases, the SBA curves were at least as close as the BEB curves to the expected upper limit. In studies 3 and 7 (50% $\omega = 1$, 50% $\omega = 1.5$), under M2a, where the estimates of the $p_{\omega>1}$ and $\omega_{>1}$ parameters were unstable (supplementary figure 6.17), the gaps between the curves for BEB and SBA were the largest, even when the number of taxa was increased from 5 to 30 (figure 3.4). This indicates that SBA, for a given false positive rate, had more power to detect sites under positive selection than BEB.

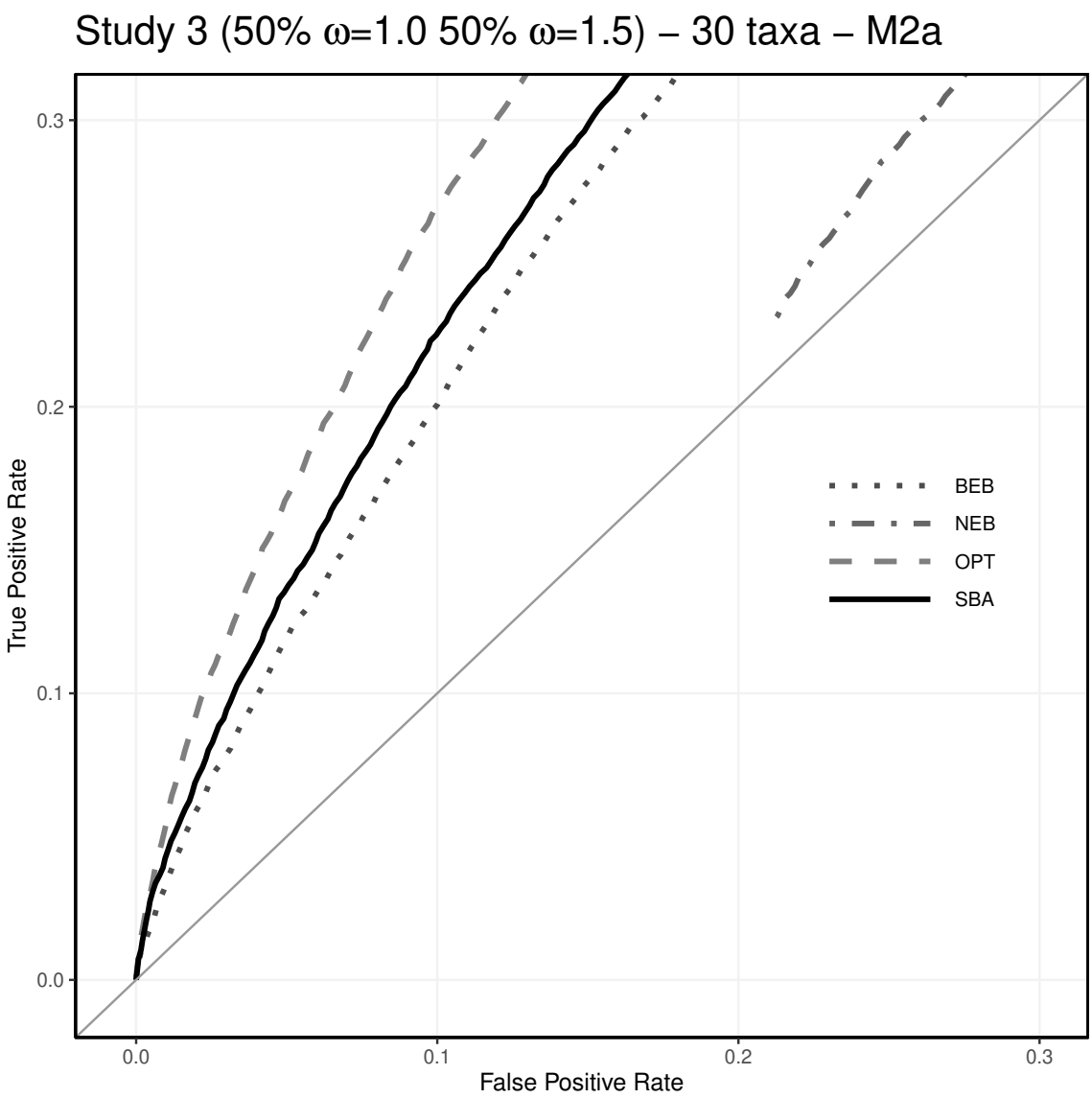


Fig. 3.4: ROC curves for the detection of sites under positive selection for BEB, NEB, and SBA analyses of data generated under *Correct Model*, study 3 (50% $\omega = 1$, 50% $\omega = 1.5$). The data were simulated using a 30-taxon tree topology. The plot includes a curve for the lower bound (y=x) and an expected upper bound (OPT) when classification is made using the generating model parameters. The curves for NEB do not always cover the whole range of false positive rates, because NEB sometimes estimates the ω distribution with all mass on $\omega > 1$. In these cases, even with a posterior probability cut-off of 1, NEB still incorrectly classifies sites to be under positive selection.

In studies 4 and 8 (45% $\omega = 0$, 45% $\omega = 1$, 10% $\omega = 5$), where the parameters of the ω distribution were well estimated, all approaches (NEB, BEB, and SBA)

performed well and the ROC curves were all close to the expected upper limit. Taken together, the results suggest that SBA balances accuracy and power at least as well as BEB and may be preferable to BEB when parameter estimates are unstable.

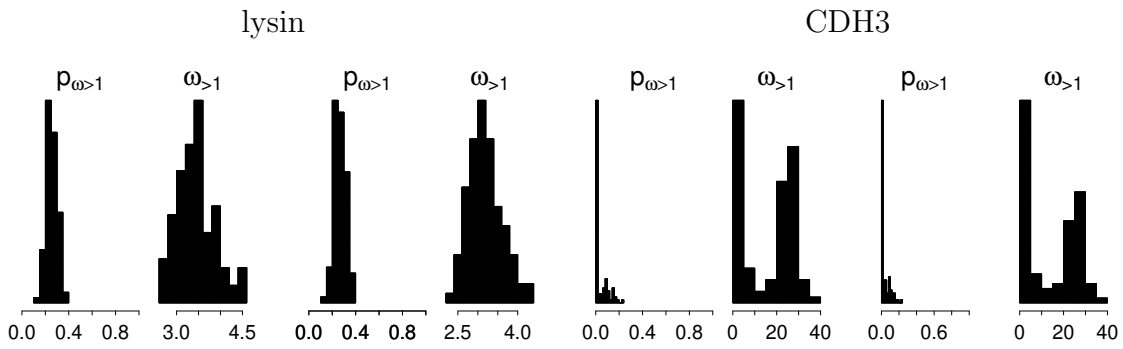
3.3.4 Real Data Analysis

I began the analysis of the 16 real datasets (described in Methods and summarized in table 2.2) by using the bootstrap distributions of the MLEs to investigate their properties. I examined the unsmoothed distributions of the parameters of the ω distribution. These distributions indicate that the MLEs for a given model can have very different properties in different real datasets (supplementary figures 6.18, 6.19, 6.20, 6.21). Although the real data represent different degrees of *regular* and *irregular* model properties, I was able to identify groups of genes that represent both extremes. The *regular* cases had no clear evidence of MLE instabilities and low bootstrap variance (e.g., lysin; figure 3.5a,b). I determined that the ω distributions had been well estimated for 6 genes (*pol*, *vif*, lysin, *nuoL3*, *RafL*, and *TrbL-VirB6_3*). In contrast, I uncovered evidence of MLE instabilities in other genes (e.g., *CDH3*; figure 3.5c,d). I determined that the ω distributions had been poorly estimated for 5 genes (*CDH3*, *mivN*, *pgpA*, *tax*, and *TrbL-VirB6_2*) under at least one model. Because no single summary statistic (number of taxa, sequence length, tree length) was generally predictive of *irregular* model properties, I recommend visual inspection of the bootstrap distributions for all real data analyses (supplementary figures 6.20, 6.21).

Table 3.2: Genes analyzed under models M2a and M8 using NEB, BEB, and SBA approaches for site classification.

Gene	N_t	N_c	-lnL		p-value	TTL	N_s
			M1a/M2a	M7/M8			
β-globin	17	144	3716.14/3712.55	3697.22/3686.13	0.0275/1.53e-5	8.40/8.57	0(0)/3(4)
ccmF	5	635	6121.78/6113.57	6127.62/6116.48	2.72e-4/1.46e-5	5.60/3.03	3(2)/3(5)
CDH3	11	176	5629.97/5623.37	5630.66/5623.88	1.35e-3 /1.14e-3	0.56/0.56	1(1)/1(1)
ENAM	11	1142	7514.30/7509.28	7609.16/7605.74	6.61e-3/0.0327	0.46/0.56	1(1)/2(1)
env	13	91	1114.64/1106.45	1115.40/1106.39	2.76e-4/1.23e-4	2.04/2.04	2(2)/2(4)
lysin	25	134	4472.65/4410.28	4472.16/4410.57	2.86e-14/0.00	8.81/8.82	22(22)/23(23)
mivN	5	504	3383.45/3832.93	3834.69/3831.44	0.595/0.0388	1.62/1.60	0(0)/1(1)
nuoL3	5	499	5006.16/4978.97	5011.37/4977.19	1.56e-12/1.44e-15	4.58/4.49	9(8)/10(10)
perM	5	351	2619.88/2619.43	2621.64/2617.94	0.638/0.0247	1.78/1.80	0(0)/2(0)
pgpA	5	198	1541.27/1539.29	1542.65/1538.91	0.138/0.0238	2.93/2.23	1(0)/1(1)
pol	23	947	9394.05/9363.96	9405.74/9365.88	8.52e-14/0.00	1.31/1.30	6(6)/10(13)
RfaL	5	403	3964.89/3955.34	3970.38/3955.44	7.16e-05/3.23e-7	3.46/3.46	2(1)/4(3)
tax	20	181	895.50/892.02	895.50/892.02	0.0309/0.0309	0.13/0.13	181(0)/181(21)
TrbL-VirB6_2	5	657	5492.55/5492.52	5301.23/5286.43	0.976/3.74e-7	2.12/2.10	0(0)/1(0)
TrbL-VirB6_3	5	938	8305.65/8288.36	8307.06/8269.09	3.09e-8/0.00	3.06/3.02	3(2)/18(11)
vif	29	192	3393.83/3367.86	3400.45/3370.66	2.29e-06/1.16e-13	2.90/2.91	10(8)/10(10)

N_t: number of taxa, N_c: sequence length in number of codons, -lnL: -log likelihood for each nested model pair, p-value of the LR test for the presence of positive selection, TTL: total tree length estimated under M2a/M8, N_s: number of sites classified to under positive selection using a posterior probability threshold of 0.95 under M2a/M8 for NEB(BEB).



(a) M2a (b) M8 (c) M2a (d) M8

Fig. 3.5: MLE distributions over bootstrap datasets for the lysin and *CDH3* genes. The distributions of the $p_{\omega>1}$ and $\omega_{>1}$ parameters associated with positive selection were estimated under models M2a and M8 for each of 100 bootstrap datasets.

Next I investigated the degree to which the real data results obtained under BEB, NEB, and SBA were consistent with each other. This is challenging, because the posterior probability thresholds for site classification are not calibrated to give comparable false positive rates. One way to compare the results is to measure the rank correlations of the site-specific posterior probability scores for positive selection between methods (BEB, NEB, and SBA). As there are a large number of pairwise comparisons, I took the mean relationship between methods for both the genes representing *regular* and *irregular* model estimation (table 3.3).

Table 3.3: Spearman rank correlations between site posterior probabilities for different forms of classification.

	Regular		***Irregular***	
	mean	SD	mean	SD
M2a				
NEB/BEB	0.98	0.04	0.65	0.17
NEB/SBA	0.94	0.09	0.66	0.17
BEB/SBA	0.96	0.05	0.98	0.02
M8				
NEB/BEB	0.99	0.01	0.84	0.30
NEB/SBA	0.96	0.04	0.81	0.27
BEB/SBA	0.98	0.03	0.98	0.02

The mean and standard deviation (SD) of the correlations are for real genes displaying *regular* and *irregular* estimation properties.

I found that when MLEs are well estimated (*regular* genes), there is stronger agreement among all three methods in the ranking of sites according to the signal for positive selection. In contrast, when the ω distributions are poorly estimated (genes representing *irregular* estimation), BEB and SBA are generally consistent in their rankings, but differ from NEB. These results suggest that NEB's inability to accommodate MLE uncertainty in such datasets has the largest effect on the posteriors. However, the problem of calibration remains. The simulation studies reveal that using a common posterior probability threshold for classification does not guarantee a similar trade-off between accuracy and power for different methods. Indeed, I see evidence of this in the real data. Comparing the counts of positively selected sites identified in the genes using thresholds of 0.50 and 0.95 reveals differences between BEB and SBA (table 3.4), despite large rank correlations.

Table 3.4: Number of sites identified to be under positive selection for the real data.

	M2a			M8		
Gene	NEB	BEB	SBA	NEB	BEB	SBA
CDH3	1/1	12/1	46/0	1/1	22/1	117/5
mivN	1/0	7/0	1/0	4/1	12/1	28/0
pgpA	1/0	4/0	4/0	5/1	5/1	17/0
tax	181/181	181/0	181/0	181/181	181/21	181/21
TrbL-VirB6_2	0/0	16/0	0/0	11/1	18/0	59/0
pol	12/6	19/6	94/4	22/10	33/13	83/16
lysin	33/22	32/22	42/5	37/23	37/23	41/11
nuoL3	18/9	18/8	85/18	19/10	20/10	83/20
RfaL	20/2	20/1	70/1	33/4	41/3	74/3
TrbL-VirB6_3	28/3	27/2	73/9	45/18	44/11	134/48
vif	13/10	13/8	31/6	15/10	19/10	37/10
β-globin	4/0	5/0	11/0	8/4	8/4	17/4
ccmF	7/1	11/1	112/0	15/3	79/5	114/5
ENAM	9/1	21/1	184/0	44/2	31/1	78/1
env	14/3	16/3	21/3	16/3	22/5	24/3
perM	4/0	6/0	0/0	6/2	6/0	36/3

The posterior probability thresholds are 0.5/0.95. The top genes represent *irregular* estimation, the middle *regular*, and the bottom genes are not categorized.

Under M2a, there was a stark difference between the *irregular* genes and all other genes. ROC curves for simulations studies are better suited for comparing methods, because they give direct comparisons of power at the same false positive rate.

I also used rank correlation to investigate the robustness of the methods (BEB, NEB, and SBA) to the chosen model (M2a versus M8). I did this by computing the rank correlation, between models, of the site posterior probabilities obtained by the same method (table 3.5).

Table 3.5: Spearman rank correlations between site posterior probabilities for models M2a and M8.

	Regular		***Irregular***	
	mean	SD	mean	SD
NEB	0.98	0.04	0.81	0.13
BEB	0.99	0.01	1.00	0.01
SBA	1.00	0.00	0.99	0.00

The mean and standard deviation (SD) of the correlations are for real genes displaying *regular* and *irregular* estimation properties.

For the *regular* genes, all three methods had high correlations with low variability. For the genes representing *irregular* estimation, the correlation was lower and the variability larger for NEB as compared to BEB and SBA. The similarity across models that I observed for SBA may be a consequence of using nonparametric bootstrapping, which should show robustness to model misspecification. It seems that BEB's application of uniform priors to the ω distribution achieved a similar effect.

Up to this point, bootstrapping has been used to obtain surrogates for posteriors. An alternative use of bootstrapping is to construct posterior probability confidence intervals to quantify the uncertainty at any given site about the true posterior probability of positive selection. Constructed from the 5 and 95% quantiles of the site posterior probabilities for positive selection, these 90% confidence intervals differed substantially between M2a and M8, highlighting differences between the two modelling frameworks. For sites having a posterior of at least 0.9 under one or more methods, the M8 confidence intervals for those sites were never wider than the corresponding M2a intervals (table 3.6). The interval widths are strongly dependent on the MLE estimates of the ω distribution. In the extreme case, if weight on the $\omega > 1$ site class is 0 for some bootstrap samples and 1 for others, the site posteriors will also be 0 and 1.

Table 3.6: Average width of 95% confidence intervals for SBA posterior probabilities. Only sites with at least one method having a posterior probability of at least 0.9 are included.

Gene	M2a	M8	Difference
CDH3	0.95	0.46	0.49
mivN	1.00	1.00	0.00
pgpA	1.00	1.00	0.00
tax	0.87	0.31	0.56
TrbL-VirB6_2	1.00	1.00	0.00
pol	0.78	0.78	0.00
lysin	0.70	0.49	0.20
nuoL3	0.26	0.21	0.05
RfaL	0.68	0.48	0.19
TrbL-VirB6_3	0.66	0.10	0.57
vif	0.36	0.14	0.21
β-globin	1.00	0.22	0.78
ccmF	1.00	0.49	0.51
ENAM	0.53	0.43	0.10
env	0.51	0.27	0.24
perM	0.91	0.14	0.77

The top genes represent *irregular* estimation properties, the middle *regular*, and the bottom genes are not categorized.

This result reflects broad differences between the MLE distributions obtained under these two models; MLE distributions under M8 tend to be tighter, and more likely located away from a boundary (supplementary figures 6.20, 6.21). I believe this represents empirical support for the commonly held notion that M8 is more powerful than M2a (Wong et al., 2004). However, this relationship should not be assumed to hold when the MLEs are poorly estimated. Confidence interval widths were at the maximum (1.0) for both M8 and M2a in three of the five genes representing *irregular* estimation. These findings highlight the importance of (1) inspecting bootstrap distributions to gain insights into the challenges posed by the data in hand, and (2) using SBA to accommodate MLE uncertainties (especially when they are poorly estimated).

Lastly, I interpret the results for the *tax* gene of the human T-cell lymphotropic virus. This gene warrants special attention because it has a highly unusual site-pattern distribution, extreme MLEs, and has been employed as a boundary case in several studies of the NEB and BEB classifiers (Suzuki and Nei, 2004; Yang et al., 2005). The dataset has 20 taxa and 181 sites, 158 (87%) of which are invariant across all 20 lineages. At each of the 23 variable sites, there is just one codon that differs from all the others with 21 of the 23 codon changes coding for a different amino acid. This atypical site-pattern distribution corresponds to a relatively large number of nonsynonymous substitutions over very short branch lengths (mean branch length: 0.0064 under both M2a and M8). A very high probability of positive selection (i.e., large values for both the $p_{\omega>1}$ and $\omega_{>1}$ parameters) is required to account for the nonsynonymous substitutions when the branch lengths are so short. In fact, both models M2a and M8 estimate 100% of the sites to be in the $\omega > 1$ class. This result belies the fact that considerable instability is associated with those parameter estimates, as revealed by bootstrapping (supplementary figures 6.20, 6.21). Since NEB ignores parameter value uncertainty, it must assign a conditional posterior probability of $\omega > 1$ ($Pr = 1.0$) for all sites, including those that are invariant. In contrast, the site posteriors for BEB and SBA were similar and depended on the site patterns (supplementary table 6.4). As expected, the SBA signal for positive selection was strongest at the 21 sites with nonsynonymous changes (M2a: $0.87 < Pr < 0.89$; M8: $0.99 < Pr < 0.99$), as compared to all other sites (M2a: $0.55 < Pr < 0.60$; M8: $0.76 < Pr < 0.80$). The SBA confidence intervals under M8 revealed that the estimates of Pr for the 21 sites with a nonsynonymous change were more reliable (average width: 0.028) than for the invariant sites (average width: 0.418). I suggest this result is appropriate for these data. Almost all the signal in this dataset is contained in those 21 sites, and it is difficult to reconcile this amount of nonsynonymous change over such short branches without strong positive selection. Moreover, when branch lengths are very short, an invariant site can only be viewed as carrying no signal about whether the ω value would be small or large over longer evolutionary periods. This leads to very wide 95% SBA Pr confidence intervals for these sites.

3.4 Discussion

I have presented an approach, based on an unconventional use of the nonparametric bootstrap, for evaluating MLE instabilities and improving site-specific inference of positive selection. For any given site in an alignment, conclusions about positive selection are based on the aggregation and distributions of many estimates of ω and many posterior probabilities. An important step in the approach involves smoothing the bootstrap distributions of the parameter estimates using techniques borrowed from kernel density estimation. This step is critical for overcoming instabilities in parameter estimation. Kernel smoothing also has the benefit of reducing computational costs relative to procedures that use full bootstrap sampling to obtain comparable numbers of MLEs.

Application of BEB, NEB, and SBA using models M2a and M8 to 100 simulated datasets in each of 10 different simulation scenarios showed that, under difficult simulation conditions when regularity conditions have not been met, NEB often poorly controls false positive classification of sites, even when the number of taxa is large. This is in contrast to past recommendations, which suggested NEB does well at controlling false positive rates when analyzing large datasets (many taxa and long sequences) (e.g., Yang et al., 2005). By accounting for variability of estimation, both BEB and SBA achieve better control of the false positive rates. However, SBA provided consistently better control under M8 when there was mild model misspecification (studies 5-8 under in table 1), and this was unaffected by pre-screening via the LR test. I note that all real data are expected to be affected, to some degree, by model misspecification.

By accounting for variability of estimation, both BEB and SBA achieve good power relative to NEB as is evidenced by their tending to be closer to the expected upper limit of performance in the ROC curves. Some of the simulation results suggest that M2a is a better-performing model. For instance, M2a gave 1) ROC curves closer to the expected upper bound in some cases (figure 3.3) and 2) lower false positive rates (table 3.1). This may, however, be a consequence of the simulations conditions being more suitable for M2a than M8. For example, in studies 3 and 7, half the sites were simulated with $\omega = 1$, and M2a has a site class with $\omega = 1$ fixed. On the other hand, considering sites with larger posteriors in the real data analysis, the

95% posterior confidence intervals were usually narrower (and never wider) for M8 than M2a. This supports previous results that suggest M8 has more power to detect sites under positive selection (Wong et al., 2004). The β-globin gene serves as a good example. Of the five sites in this gene where either NEB or BEB gave a posterior of at least 0.9, the SBA confidence interval widths were all 1 for M2a, but averaged 0.129 for M8. Moreover, the $\omega_{>1}$ parameter distributions tended to be wider for M2a than M8, particularly for the genes that displayed properties suggesting regularity conditions were met. This is probably because the beta distribution used by M8 to model $\omega < 1$ has more flexibility in real data conditions compared to an M2a model with the same number of parameters.

An appealing attribute of BEB, relative to SBA, is its limited use of computational resources. Each SBA bootstrap analysis may use similar computational resources as BEB does for the one original dataset. However, SBA's greater computational requirements is a trade-off for a more rigorous assessment of the parameter estimation. For example, SBA adjusts for the uncertainty in all model parameters, including branch lengths, while BEB does not. A new BEB implementation that integrated over branch lengths would require costlier techniques because numerical integration does not scale well with higher dimension. Moreover, because SBA estimates each set of bootstrap parameters independently, they can be estimated in parallel. On a computing cluster with as many cores as bootstrap samples generated, the wall-clock times for BEB and SBA are comparable.

There are a limited number of BEB implementations for different models. By contrast it is comparatively trivial to apply SBA to new models once the basic capacity for bootstrapping and parameter smoothing are in place. This could facilitate the application of SBA to a wider variety of inference problems in molecular evolution than has occurred with BEB. SBA for the popular branch-site codon model A (Yang and Nielsen, 2002; Zhang et al., 2005) was implemented as a demonstration of the feasibility of SBA implementations for new models. A new, preliminary implementation, which was completed within a few hours, can be found at https://github.com/Jehops/codeml_sba. An overview of the analysis of the *NR1D1* gene (Baker et al., 2016) under SBA can be found in the supplementary section 6.6.

There are useful by-products of the SBA approach for classifying sites. The histograms of the distributions of the MLEs over bootstrap samples provide insight into the degree of irregularity of the estimation. For several of the datasets, most notably the *tax* gene dataset, these histograms provided a clear indication that the MLEs were unstable. In such cases, site classifications should be accepted with caution. Even when regularity conditions have been met, the confidence intervals of the posteriors provide an additional tool for assessing the certainty about the strength of the signal for positive selection at an individual site. I suggest that future analyses of real data should include both visual inspection of bootstrap distributions and reporting of SBA-derived confidence intervals of the posterior probabilities associated with positive selection.

Bootstrapping has been shown to provide effective adjustments to EB methods in other settings. For example, Laird and Louis (1987) studied the application of bootstrapping with EB methods for random effects models where both the observations and random effects distributions were Gaussian. They argued that confidence intervals produced from bootstrap posteriors were frequently narrower than they should be and that bootstrap averaging helped to ameliorate problems. They speculated that bootstrapping would produce good EB inferences for a broad class of EB problems. In a prediction setting, a procedure that aggregates predictors generated from bootstrap replicates was proposed by Breiman (1996), which was shown to move some unstable predictors closer to optimality. The bagging procedure used in that paper is equivalent to using the median posterior to classify sites under SBA. My experiments (data not shown) indicated that the average is a better measure of the middle of the distribution of site posterior probabilities.

While using the data in hand to account for errors in MLE estimation is helpful for detecting sites under positive selection, refinements of the SBA approach are warranted. Like other approaches, I have avoided the difficult process of calibrating for type I errors in real data. Choosing an optimal bandwidth parameter for smoothing a distribution is also a difficult process. Under-smoothing will leave spurious bumps and irregularities in the distribution and over smoothing will remove useful information and increase bias. There are different theoretical suggestions for the size of the bandwidth parameter, but these can be challenging to apply as they may depend

on the unknown density (Venables and Ripley, 2013, p. 176). SBA uses bootstrap distributions to highlight problems when MLEs fall on or close to their boundaries. It does well to accommodate the variance in a parameter estimate, however, when estimates are very small, the variance, even under bootstrapping, may be underestimated. This may be a problem encountered with the branch lengths of the *tax* gene. Some preliminary experiments show that perturbing the very small branch length estimates of the *tax* gene can cause large differences in the MLEs of the parameters of the ω distribution. This suggests that applying kernel smoothing to parameters other than those defining the ω distribution may be helpful.

SBA can be applied to a wide variety of problems in molecular evolution where uncertainties or instabilities in MLEs impact inference based on empirical Bayes. Examples where the method can be directly applied, with little or no modification, include: classification of sites into general rate categories (e.g., Mayrose et al., 2004), identification of positively selected sites in non-coding DNA (e.g., Haygood et al., 2007), identification codon sites subject to episodic change in selection pressure (e.g., Yang and Nielsen, 2002), detection of Type-I functional divergence in protein sequences (e.g., Gaston et al., 2011), detection of amino acid sites having shifts in the pattern of exchangeabilities (e.g., Le et al., 2012), and detection of amino acid sites evolving under a covarion-like evolutionary process (e.g., Penn et al., 2008). With some modification, SBA could be applied to the task of ancestral state reconstruction. As the field moves towards increasingly more complex models, there will be increasing demand for methods such as SBA that can account for parameter-estimate uncertainties.

3.5 Theory and Methods

3.5.1 Bootstrap Methods to Adjust for Uncertainty

To construct confidence intervals for a parameter, θ, and correct bias, Efron (1979) devised the bootstrap. A bootstrap sample, $\boldsymbol{x}^*$, is obtained by drawing the values, $x_1^*, \ldots, x_n^*$, with replacement from a random sample, $\boldsymbol{x}$. For each of $b = 1 \ldots B$ bootstrap samples, the bootstrap estimate can then be calculated, $\hat{\theta}^{*b}$, to obtain the bootstrap distribution of $\hat{\theta}$. Bootstrap distributions are commonly used with phylogenetic data to test the topology of a proposed tree. I applied the bootstrap

to site patterns in a sequence alignment to adjust for the uncertainty in parameter estimates in EB classification. The procedure is illustrated in figure 3.1:

1. From an alignment of protein coding DNA sequences, $\boldsymbol{x}$, with n codon sites, randomly sample site patterns with replacement to obtain a bootstrap sample, $\boldsymbol{x}^{*b}$, with n sites.

2. Estimate the MLEs, $\hat{\theta}^{*b}$, for bootstrap sample $\boldsymbol{x}^{*b}$.

3. Use $\hat{\theta}^{*b}$ and $\boldsymbol{x}$ to calculate posterior probabilities, $Pr_h(\omega > 1|x_h; \hat{\theta}^{*b})$, that each site, h, is under positive selection.

4. Repeat steps 1 through 3 B times to calculate B sets of posterior probabilities for each codon site.

5. Calculate an aggregate posterior probability that each site is under positive selection by, e.g., averaging posterior probabilities over bootstrap replicates, $\sum_{b=1}^{B} Pr_h(\omega > 1|x_h; \hat{\theta}^{*b})/B$.

A preliminary implementation of the SBA method supporting codon models M2a, M8, and branch-site model A, built upon the codeml application from the PAML package (Yang, 2007), can be found at
https://github.com/Jehops/codeml_sba.

3.5.2 Kernel Smoothing to Approximate the Bootstrap Distribution

Kernel smoothing (Akaike, 1954; Parzen, 1962; Rosenblatt et al., 1956; Wand and Jones, 1994) is class of nonparametric techniques that can improve estimation of a distribution. The kernel density estimator for a continuous density f, $\hat{f}(x; h) = (nh)^{-1} \sum_{i=1}^{n} K([x - X_i]/h)$, includes a kernel density (probability) function, K, to locally average or smooth observations and the amount of smoothing is controlled by a bandwidth parameter, h. For small h, each of the $h^{-1}K([x - X_i]/h)$ contributions are large only for x close to some X_i giving rise to a bumpy distribution, whereas for h large the $h^{-1}K([x-X_i]/h)$ contributions overlap giving a much smoother distribution (Silverman and Young, 1987). I used kernel density estimation to create smoothed bootstrap distributions for the p parameters of the ω distributions under models M2a and M8 using a uniform kernel.

Kernel density estimation requires a bandwidth parameter as input. One method for determining h is using leave-one-out cross validation (Venables and Ripley, 2013, p. 184),

$$\hat{f}_{(-k)}(x;h) = (n-1)^{-1}h^{-1}\sum_{i\neq k} K([x-X_i]/h).$$

In this approach, h is chosen to maximize the sum of the logged density estimates $\sum_k \log \hat{f}_{(-k)}(x_k;h)$, where $\hat{f}_{(-k)}(x;h)$ is the kernel density estimate constructed from all of the x_i except x_k. However, my experiments using leave-one-out likelihood to choose an optimal bandwidth parameter for the p parameters of M2a and M8 merely resulted in smoothed estimates of the biased bootstrap distributions. To obtain conservative estimates of the p parameters that suppressed the influence of instabilities I chose to over smooth by using a bandwidth parameter of $h = 0.4$ for all applications of SBA.

Adding kernel smoothing to the bootstrap algorithm increases the number of parameter estimates used in step 5 of the unsmoothed algorithm by sampling from a smoothed bootstrap distribution. The adjustment is in step 2 of the algorithm. The ML parameters estimated from bootstrap sample b, $\hat{\theta}^{*b}$, are replaced by θ^{sb} sampled from the smoothed bootstrap distribution. The rest of the algorithm proceeds as in the unsmoothed version, but using θ^{sb} in place of $\hat{\theta}^{*b}$.

For model M8, the step 2 adjustment is as follows. For each $\hat{\theta}^{*b}$, $p^{sb}_{\omega<1}$ samples are repeatedly drawn from a univariate uniform distribution centered at $\hat{p}^{*b}_{\omega<1}$ with width $2h$. If necessary, the minimum and maximum points of the distribution are truncated to 0 and 1. Let θ^{sb} denote $\hat{\theta}^{*b}$ with $p^{sb}_{\omega<1}$ replacing $\hat{p}^{*b}_{\omega<1}$ ($p^{sb}_{\omega>1} = 1 - p^{sb}_{\omega<1}$). The same procedure is used under model M2a, however, with three weight parameters, the sampling is done on a bivariate uniform distribution with the following additional restrictions: i) $p^{sb}_{\omega<1}+p^{sb}_{\omega=1} \leq 1$, ii) $(\hat{p}^{*b}_{\omega<1}-h) \leq p^{sb}_{\omega<1} \leq (\hat{p}^{*b}_{\omega<1}+h)$, and iii) $(\hat{p}^{*b}_{\omega=1}-h) \leq p^{sb}_{\omega=1} \leq (\hat{p}^{*b}_{\omega=1}+h)$. As with M8, if necessary, the minimum and maximum points of the distribution are truncated at 0 and 1, and $p^{sb}_{\omega>1} = 1 - p^{sb}_{\omega<1} - p^{sb}_{\omega=1}$.

3.5.3 Simulation Studies

Datasets were simulated using *EvolverNSSites* from the PAML 4.8a package (Yang, 2007) and *Indelible* (Fletcher and Yang, 2009) following some of the settings described in Wong et al. (2004). To compare the relative performance of BEB, NEB, and SBA

for predicting sites under positive selection, 10 different simulation studies, divided into three scenarios, were used. Table 3.1 gives an overview of the ω distributions used to simulate the data. The *Correct Model Scenario* included four simulation studies where the nuisance parameters, $\kappa = 1$ and $\pi_i = 1/61$, matched the fitted model. The *Mild Misspecification* and *Heavy Misspecification* scenarios included four simulation studies with mild misspecification and two studies with heavy misspecification of the fitted model, respectively. The data in the *Mild Misspecification Scenario* was simulated using $\kappa = 8$ and empirical codon frequencies derived from application of the general time-reversible model (Yang, 2006, p. 33) to the *TrbL-VirB6-3* plasmid conjugative transfer protein of *Rickettsia*. In the fitted model, κ was estimated, while the misspecification was introduced by using F3x4 (expected codon frequencies calculated using the nucleotide frequencies at the three codon positions). For the *Heavy Misspecification Scenario*, study 9 used the heavily biased codon frequencies from the Drosophila *GstD1* gene and $\kappa = 8$ to simulate the data. In study 10, there were two heterogeneous classes of sites. Half the sites were simulated using equal codon frequencies, $\kappa = 1$, and $\omega = 0.5$, while the other half with the Drosophila gsTD gene codon frequencies, $\kappa = 8$, and $\omega = 1$. For both studies in this scenario, analysis was carried out using a single set of codon frequencies (set equal to 1/61) and a single κ parameter estimated for all sites in the data set. For all studies in the three scenarios, 100 alignments, each having 500 codons, were simulated with the same 5-taxon tree from Wong et al. (2004). The studies in the *Correct Model Scenario* were repeated under model M2a with the 30-taxon tree from the same paper.

3.5.4 Real Data Analysis

Table 2.2 describes the real data sequences analyzed under models M2a and M8 using NEB, BEB, and SBA. Of the 16 genes, eight code for transmembrane proteins in *Rickettsia* (*ccmF*, *mivN*, *perM*, *pgpA*, *RfaL*, *TrbL-VirB6_2*, and *TrbL-VirB6_3*) and were previously analyzed in Bao et al. (2008). Three genes from the HIV-1 virus (*env pol*, and *vif*) and a β-globin gene were described and analyzed in Yang et al. (2000a), two primate genes (*CDH3* encoding cadherin and *ENAM* encoding enamelin), a lysin gene from Yang et al. (2000b), and the *tax* gene from the human T-cell lymphotrophic virus (HTLV) that was analyzed by Suzuki and Nei (2004). All data is available at https://github.com/jehops/sba_real_data.

Chapter 4

Unrecognized Statistical Difficulties with Tests of Positive Selection under the Branch-Site Family of Codon Models

4.1 Introduction

Early models of evolution had limited power to detect positive selection that acted upon proteins and amino acids. A challenge is that positive selection, relative to purifying selection and neutral evolution, is a rare occurrence and often acts only upon a small proportion of sites (Petersen et al., 2007; Studer et al., 2008). A single ω averaged over all sites would rarely be estimated large enough (i.e., $\omega > 1$) to reject the null hypothesis of no positive selection, even for proteins that were subjected to positive selection. By treating ω at a site as the realized value of a random variable, and thus allowing ω to vary over sites, the power to detect positive selection increases (Wong et al., 2004). The site models, however, do not allow ω to vary over time.

Another challenge to detect positive selection is that it often acts episodically (Kosiol et al., 2008; Studer and Robinson-Rechavi, 2009). Early methods to detect episodic positive selection, such as those of Messier and Stewart (1997) and Zhang et al. (1997), inferred ancestral sequences to compute and compare the rates of nonsynonymous and synonymous substitutions at particular lineages of a phylogeny. Yang (1998) avoided the problems with ancestral sequence reconstruction (Collins et al., 1994) in the branch models. The branch models use likelihood methods that condition upon all unknown ancestral states and allowed ω to vary over lineages to capture episodic positive selection. However, the branch models assume no variation in ω among sites, so positive selection is detected along a lineage of the tree only if

the average ω over sites is sufficiently large.

To increase the power to detect positive selection at a subset of sites along pre-specified lineages of a phylogenetic tree, referred to as the foreground of the tree, branch-site codon models were developed (Yang and Nielsen, 2002; Forsberg and Christiansen, 2003; Bielawski and Yang, 2004; Zhang et al., 2005). The first branch-site model A (Yang and Nielsen, 2002) used an LR test that was shown, through simulation, to be susceptible to high false positive rates (Zhang, 2004). Changes were later made (Zhang et al., 2005) and new simulations showed that the updated LR test of branch-site model A did not suffer from high false positive rates when selection was relaxed in the foreground. Suzuki (2008) and then (Nozawa et al., 2009) reported that false positive rates were still excessively high, however most of these claims were dismissed due to faulty statistical interpretation (Yang et al., 2009; Yang and Dos Reis, 2010; Zhai et al., 2012). Yang and Dos Reis (2010) provided new simulation results, which suggest that the asymptotic theory is sound and the large-sample null distribution is reliable. Work has also shown that the branch-site LR test is normally conservative (Gharib and Robinson-Rechavi, 2013), not misled by positive selection in background branches (Zhang et al., 2005; Gharib and Robinson-Rechavi, 2013; Fletcher and Yang, 2010), and robust to insertions and deletions, as long as the sequence alignment is correct (Fletcher and Yang, 2010).

The updated branch-site model A has provided a basis for other codon models of episodic selection. Guindon et al. (2004) developed what they referred to as a stochastic branch-site model, which does not require the phylogenetic tree to be divided *a priori* into background and foreground branches. They employ two Markov processes, the usual process for codon states, and another for the unobservable ω class, which can change along any branch of the tree. Lu and Guindon (2013) showed through simulation that both branch-site models are conservative under null conditions, but branch-site model A is more powerful when the foreground is correctly chosen. On the other hand, when too few or too many foreground branches are chosen, the power of branch-site model A decreases. If prior information about the foreground is unavailable, each branch can be individually tested under branch-site model A using corrections for multiple tests (Anisimova and Yang, 2007). Lu and Guindon (2013) determined that the stochastic branch-site model is better suited for such exploratory

experiments.

Kosakovsky Pond et al. (2011), Murrell et al. (2012), and (Murrell et al., 2015) developed a class of models that relax the constraints of the selection arrangements in branch-sites model A by allowing each branch-site combination to have different ω values. Kosakovsky Pond et al. (2011) showed that branch-site model A can give excessive type I or type II errors when the data strongly deviate from the constrained ω distributions for background and foreground branches. They reported that their random-effects method consistently matched or outperformed branch-site A tests in terms of power and error rates.

Smith et al. (2015) developed an adaptive feature to the random effects method, which determines the optimal number of rate categories for each branch using small-sample Akaike Information Criterion. This is a welcome advancement as others have reported problems with codon models when the number of mixture classes is too large (e.g. Mingrone et al., 2018). Davydov et al. (2019) argued that branch-site model A erroneously detects positive selection in real genes at a rate greater than 70%. They developed a modified branch-site model, which includes a separate site parameter for the synonymous rate, thus accounting for nucleotide sequence selection and mutation rate. Like Baele and Lemey (2013) and Gil et al. (2013), the synonymous rate at each site is the realized value of a discretized unit gamma distribution, but unlike those models, they allow ω to vary over both branches and sites.

Among the branch-site models in this family, branch-site model A employs the most restrictive constraints on its parameter values. The constraints allow the model to serve as an explicit test of positive selection for *a priori* hypotheses about specific branches. Despite the sophistication and claimed improvements of newer methods for detecting episodic positive selection, branch-site model A remains widely used, because such tests are believed to have the highest power when *a priori* information about the foreground branches are tested.

Branch-site model A has recently been used to describe, e.g., the molecular mechanisms of immunity, longevity, and cancer-resistance in bats (Scheben et al., 2020) and the North American beaver (Zhang et al., 2020), tumor suppressor in cetaceans (Martinez et al., 2020), and the antagonistic insect-plant interaction between swallowtail butterflies and birthworts (Allio et al., 2020). With such wide use, it is important

to fully understand its tendencies and limitations and when it or other models are appropriate. Here I perform new simulations studies to reassess the properties of the updated branch-site model A. By assessing the statistical properties of a base form of a branch-site model, I aim to provide insight for the model and related models that implement extensions (e.g., Davydov et al., 2019).

4.2 Theory and Methods

The parameters of the ω distribution under the alternative model of updated branch-site model A described in Zhang et al. (2005) are shown in table 4.1. For the null model, Yang and Nielsen (2002) originally constrained $\omega_0 = 0$ and permitted only sites classes 0 and 1, which is equivalent to sites model M1 (Yang et al., 2000a). Using data simulated under the null hypothesis, Zhang (2004) found positive selection in the foreground was erroneously detected in 19%—54% of cases. Zhang et al. (2005) used an updated null model, which was shown through simulation to resolve the problem of inflated false positive rates. This updated null model includes all four ω site classes of the alternative model, but constrained $\omega_2 = 1$. In addition, Zhang et al. (2005) relaxed the constraint on ω_0 by allowing it to take on values between 0 and 1.

Table 4.1: The ω distribution under the alternative model of branch-site model A, described in Zhang et al. (2005). Under the null model, the constraint $\omega_2 = 1$ is imposed.

Site Class	Proportion	Background	Foreground
0	p_0	$\omega_0 < 1$	$\omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p1)p_0/(p_0 + p_1)$	$\omega_0 < 1$	$\omega_2 > 1$
2b	$(1 - p_0 - p1)p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$

Here I conduct simulation studies to explore how the asymptotic null distribution of the updated branch-site A LR test (Zhang et al., 2005) may vary from the recommended χ^2 distributions. Codon sequence data were simulated using the *Indelible* simulation software (Fletcher and Yang, 2009). For each of three simulation scenarios covering 23 distinct simulations studies, 1000 sequence alignments 500, 5000, or 10000 codons long were generated using a symmetric, 8-taxon tree (figure 4.1) with the total of all branch lengths summing to 3 or 6. All sequence alignments were simulated with

a transition to transversion rate ratio, $\kappa = 2$ and equal codon frequencies. Table 4.2 provides an overview of the simulation conditions, including the ω distributions used to simulate the data.

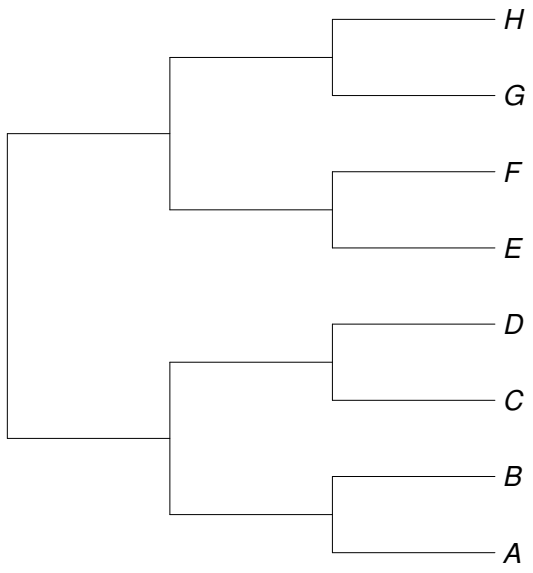


Fig. 4.1: Phylogenetic tree topology used in branch-site model simulations.

The *Single Foreground Branch Scenario* was comprised of 12 studies with each sequence alignment simulated using a single terminal foreground branch leading to taxon A or an internal foreground branch at the base of the tree in figure 4.1. The sequence lengths, total branch lengths, and foreground branch length varied between studies in this scenario. The *Half Tree Foreground Scenario* was comprised of 5 simulations studies, each with half of the 8-taxon tree in the foreground. Four of the studies in this scenario used 5000-codon sequences and one used 10000. The *Misspecification of ω Distribution Scenario* was comprised of 6 studies with sequences generated using codon model M3 (k=3) (Yang et al., 2000a), i.e., the number of simulated ω site classes does not match the number of ω site classes in the fitted model.

Table 4.3 describes simulation studies conducted under the *Confounding Foreground Scenario*. Confounding here means that any foreground branches specified under simulation do not match the foreground branches of the fitted model. For each study in this scenario, data were simulated under both the null ($\omega_2 = 1$ in foreground branch) and the alternative ($\omega_2 \geq 1$ in the foreground) models. This was done to

Table 4.2: Design of branch-site model A simulation studies for assessing estimated LR distribution under null hypothesis conditions.

Study	N_c	N_f	TTL	ω distribution
			Single Foreground Branch	
1a	500	1(t)	3	p_0=0.7 p_1=0.2 ω_0=0.3
1b	5000	1(t)	3	
1c	5000	1(t)*	3	
1d	5000	1(t)**	3	
2a	500	1(t)	3	p_0=0.75 p_1=0.25 ω_0=0.3
2b	5000	1(t)	3	
3a	500	1(t)	3	p_0=0.25 p_1=0.75 ω_0=0.3
3b	5000	1(t)	3	
4a	5000	1(t)*	6	p_0=0.5 p_1=0.5 ω_0=0
4b	5000	1(t)**	6	
4c	5000	1(i)*	6	
4d	5000	1(i)**	6	
			Half Tree Foreground	
5	5000	h	6	p_0=0.375 p_1=0.375 ω_0=0
6	5000	h	6	p_0=0.475 p_1=0.475 ω_0=0
7a	5000	h	3	p_0=0.5 p_1=0.5 ω_0=0
7b	5000	h	6	
7c	10000	h	3	
			Misspecification of ω Distribution	
8a	500	1(t)	3	$[p_0, p_1, p_2]$=[0.4, 0.4, 0.2] $[\omega_0, \omega_1, \omega_2]$=[0.1, 0.5, 0.9]
8b	5000	1(t)	3	
8c	500	h	3	
8d	5000	h	3	
9a	500	h	3	$[p_0, p_1, p_2]$=[0.4, 0.2, 0.4] $[\omega_0, \omega_1, \omega_2]$=[0.1, 0.5, 1]
9b	5000	h	3	

N_c: sequence length in number of codons, N_f: number foreground branches (t: terminal branch, i: internal branch, h: half tree, *: foreground is 1/10 length or other branches, **: foreground branch is 10 times length of other branches), TTL: total tree length. Note, under the Misspecification of ω Distribution scenario, the generating model was M3 (k=3).

compare LR statistic CDFs with and without confounding foreground branches. The goal of the studies in this scenario is to determine whether confounding foreground branches may cause false detection of positive selection under the fitted model.

Table 4.3: Design of branch-site model A simulation studies for assessing estimated LR distribution under confounding foreground branch conditions.

Study	N_c	N_f	TTL	ω distribution
10	5000	1(i)	3	p_0=0.5 p_1=0.4 ω_0=0.75 ω_2=2.0
11	5000	1(i)	6	p_0=0.475 p_1=0.475 ω_0=0 ω_2=3.0
12	5000	h	6	p_0=0.475 p_1=0.475 ω_0=0 ω_2=3.0

N_c: sequence length in number of codons, N_f: number foreground branches (t: terminal branch, i: internal branch, h: half tree, *: foreground is 1/10 length or other branches, **: foreground branch is 10 times length of other branches), TTL: total tree length.

4.3 Results and Discussion

4.3.1 LR Tests Tend to be Conservative when Information Content is Low

Figure 4.2 shows the LR statistic CDFs for branch-site model A when a single branch is specified in the foreground. Studies 1a - 1d differ from the other studies in this scenario in that $p_0 + p_1 < 1$. Consequently, alternative hypotheses parameters that give the true generating model have positive weight on the 2a and 2b classes in Table 4.3. Because they have positive weight, ω_2 must equal 1 to give the true generating distribution. Thus the parameters are identifiable under the alternative model, which implies that the non-standard likelihood theory of Self and Liang (1987) will apply with large samples. That theory suggests that the large sample LR statistic distribution is well approximated by a $\chi^2_0/2 + \chi^2_1/2$ distribution. However, the results of studies 1a and 1c indicate that the sparseness of information contained in a single foreground branch may not be sufficiently influential to draw all or most weight away from p_0 or p_1. Histograms of $1 - p_0 - p_1$ for studies 1a - 1d in figure 4.3 show that too often all, or nearly all, weight is placed on sites classes ω_0 and ω_1 in studies 1a and 1c. With increased information content in the foreground branch, whether by longer sequences lengths as in study 1b, or a longer foreground branch as in study 1d, the $\chi^2_0/2 + \chi^2_1/2$ distribution does well to approximate the large-sample LR statistic distribution.

For the remaining studies in the *Single Foreground Branch Scenario*, with sequence

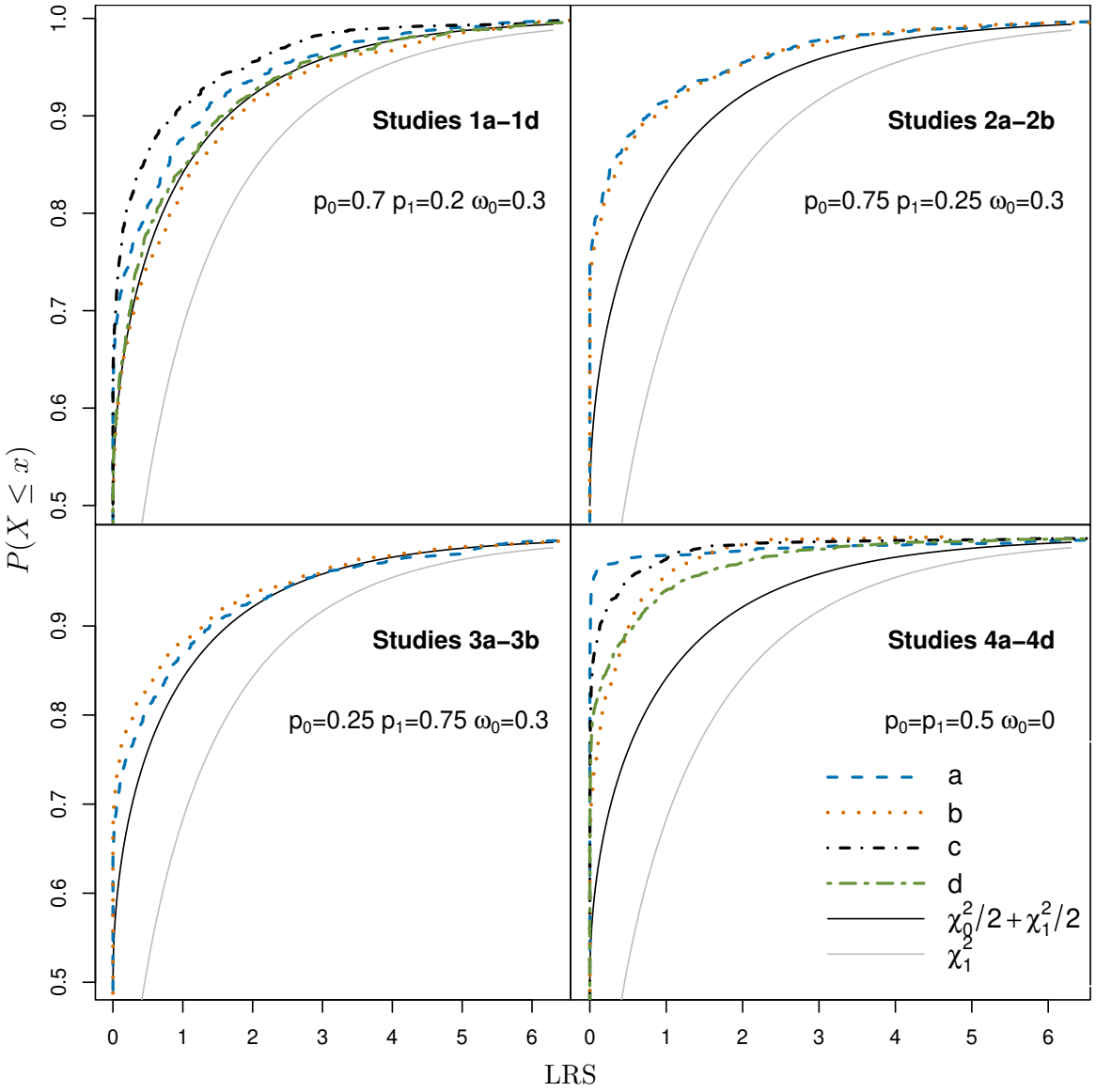


Fig. 4.2: CDFs of LR statistics under branch-site model A for data simulated in the *Single Foreground Branch Scenario*. For each simulation study, 1,000 sequence alignments were generated using a balanced, 8-taxon tree topology with branch lengths summing to 3 or 6. The simulated parameter values of the ω distribution are shown in each panel. Studies 1c, 4a, and 4c have foreground branches that are 1/10 the length of the other branches in the tree. Studies 1d, 4b, and 4d have foreground branches that are 10 times the length of the other branches in the tree. Sequences are 500 codons long in studies 1a, 2a, and 3a and 5000 codons long in all other studies in this scenario. Refer to table 4.2 for detailed simulation conditions. CDFs for $\chi^2_0/2 + \chi^2_1/2$ and χ^2_1 are also included.

data simulated with all weight on the $\omega < 1$ site classes ($p_0 + p_1 = 1$), the non-standard likelihood theory of Self and Liang (1987) is not expected to apply. The

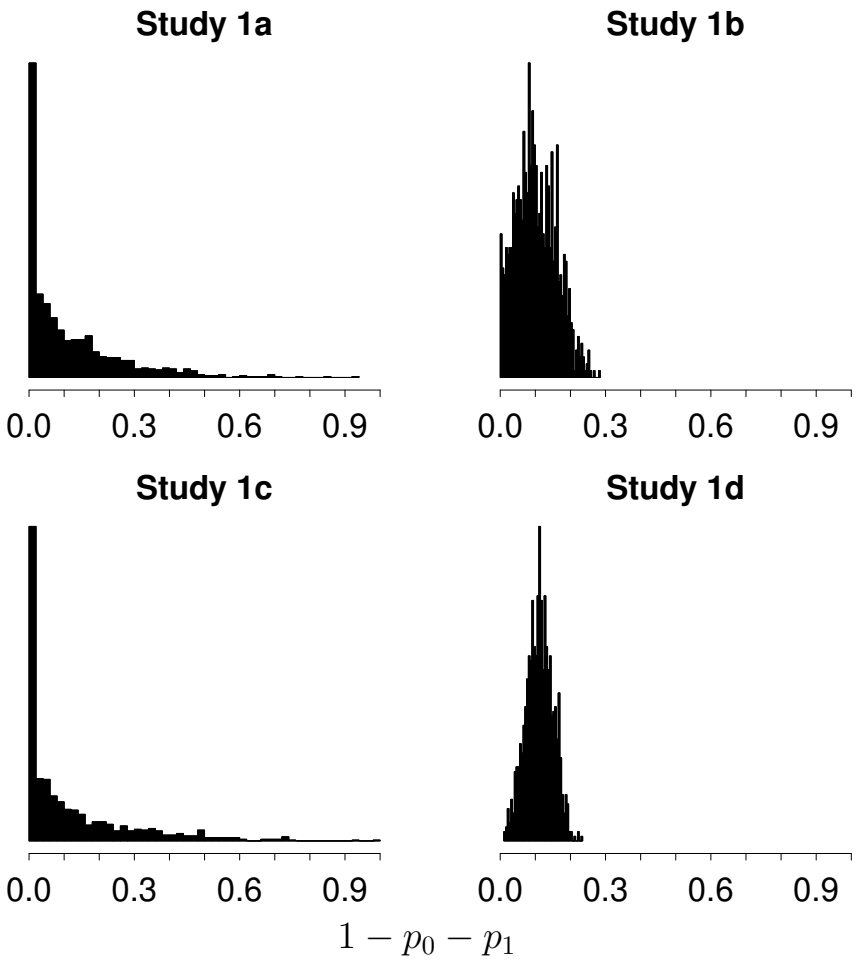


Fig. 4.3: Histograms of the weight MLEs as $1-p_0-p_1$ for studies 1a - 1d of the *Single Foreground Branch Scenario*. Refer to table 4.2 for detailed simulation conditions.

regularity condition of Self and Liang (1987) that is violated is identifiability. Because $p_0 + p_1 = 1$, alternative hypothesis parameters with $p_{2a} = p_{2b} = 0$ and arbitrary w_2 give the true generating model. Unlike the results of Mingrone et al. (2018), which showed violations of the conditions of the non-standard likelihood theory gave anti-conservative LR statistic distributions in site models, studies 2-4 resulted in conservative LR statistic distributions. The conditions are however similar to those described in Mingrone et al. (2018) where it was argued that if ω_2 were fixed under the alternative model, only $p_2 = 0$ would give the null model. These conditions match case 5 of Self and Liang (1987), which gives a $\chi^2_0/2 + \chi^2_1/2$ LR distribution. In Mingrone et al. (2018) the alternative maximized over ω_2 would always give a larger likelihood than with ω_2 fixed, so the theoretical large-sample expectation is anti-conservative

behaviour for LR tests involving sites model M2a. A similar theoretical argument for large-sample anti-conservative behaviour is expected for branch-site model A, but the argument is complicated, because positive weight must remain on ω_2.

4.3.2 LR Tests are Anti-conservative when Information Content is High

In the *Half Tree Foreground Scenario*, the estimated LR distributions are anti-conservative relative to a $\chi^2_0/2+\chi^2_1/2$ and not well estimated by either a $\chi^2_0/2+\chi^2_1/2$ or χ^2_1 distribution, and more so when the total tree length is increased or the number of generated sites is increased to $10,000$ (figure 4.4). With more of the tree in the foreground, the anti-conservative LR distribution behaviour observed in Mingrone et al. (2018) is also observed under branch-site model A. In studies $7b$ and $7c$, 7.1% and 7.4% of the LR statistics were beyond 2.71, the 5% threshold of the $\chi^2_0/2 + \chi^2_1/2$ distribution. I speculate that as the information content increases with more of the tree in the foreground, the alternative model does better to explain nonsynonymous changes evident in several locations. That extra freedom leads to anti-conservativeness. Note that anti-conservativeness increases as p_0+p_1 increases, in line with expectations based on the violation of Self and Liang (1987) regularity conditions when $p_0 + p_1 = 1$ (figure 4.5).

In many circumstances, anti-conservative behaviour is more concerning than lack of power, so recommendations have been to simply use thresholds from a χ^2_1 distribution to make the test conservative. However, if LR statistics become large with larger ω_2 values, tests based on a χ^2_1 distribution could also pose risks of anti-conservative behaviour. For few foreground branches, this is not expected, but with more of the tree in the foreground, the LR statistic could vary more with larger ω_2 values.

4.3.3 LR Tests are Extremely Conservative when the ω Distribution is Misspecified

Data in the simulation studies of the *Misspecification of ω Distribution Scenario* were simulated with two different ω distributions from an M3 k=3 model (Yang et al., 2000a). Studies $8a - 8d$ placed weights $[p_0,p_1,p_2] = [0.4, 0.4, 0.2]$ on $[\omega_0, \omega_1, \omega_2] = [0.1, 0.5, 0.9]$ and studies $9a$ and $9b$ puts weights $[p_0, p_1, p_2] = [0.4, 0.2, 0.4]$ on $[\omega_0, \omega_1, \omega_2] = [0.1, 0.5, 1.0]$. Note that the ω_2 of sites model M2a is not ω_2 of branch-sites model A. Regardless whether a single branch or half of the tree is part of the foreground, LR

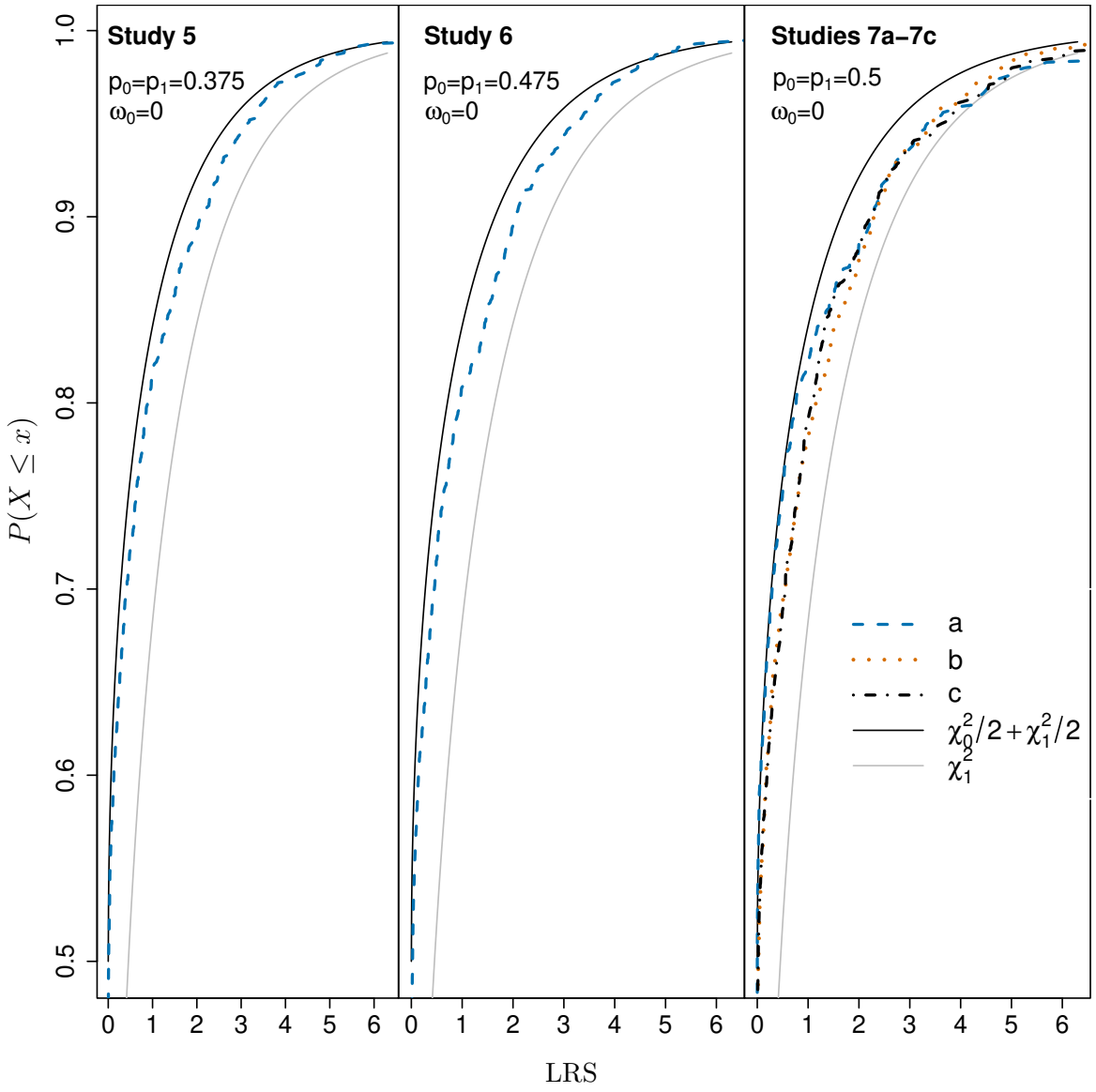


Fig. 4.4: CDFs of LR statistics under branch-site model A for data simulated in the *Half Tree Foreground Scenario*. For each simulation study, 1,000 sequence alignments were generated using a balanced, 8-taxon tree topology with branch lengths summing to 3 or 6. The simulated parameter values of the ω distribution are shown in each panel. Refer to table 4.2 for detailed simulation conditions. CDFs for $\chi^2_0/2 + \chi^2_1/2$ and χ^2_1 are also included.

statistic distributions are highly conservative when the ω is misspecified (figure 4.6).

Under the null model, site class 2a is less constrained relative to the other sites classes, because it can model sites using two ω values, $\omega_0 < 1$ in the background and $\omega = 1$ in the foreground. This additional flexibility and perhaps near unidentifiability

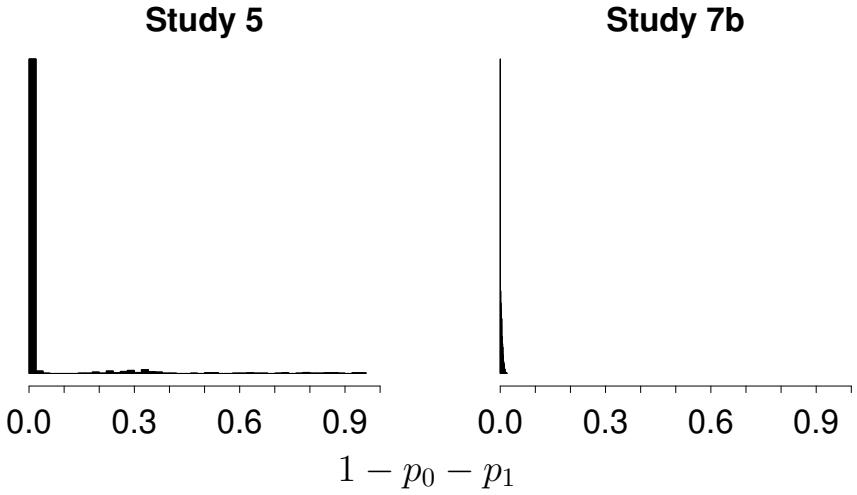


Fig. 4.5: Histograms of the weight MLEs as $1 - p_0 - p_1$ for studies 5 and 7b of the *Half Tree Foreground Scenario*. Refer to table 4.2 for detailed simulation conditions.

results reported in Mingrone et al. (2018) makes site class 2a best able to fit sites generated under $\omega = 0.5$. However, there is little additional flexibility under the alternative model since all the generating ω values are less than 1. Thus, differences in likelihood scores under the null and alternative models are often small, resulting in the highly conservative LR distribution. When half of the tree is in the foreground, only 8 of the 1000 LR tests are rejected with using the 5% $\chi^2_0/2 + \chi^2_1/2$ distribution threshold of 2.71, and only 2 tests are rejected using the 1% threshold of 5.41.

4.3.4 Confounded Foreground Branches May Cause False Detection of Positive Selection

Figure 4.7 shows estimated LR statistic CDFs under branch-site model A for simulation studies 10, 11, and 12 in the *Confounding Foreground Scenario*. In studies 10 and 11, the LR statistics are conservative and distributions are comparable with or without confounding. By contrast, in Study 12, strong anti-conservative behaviour is seen. The more anti-conservative behaviour under the null by comparison with Studies 10 and 11, suggests that part of the reason is simply the additional information content in the alternative model when more of the tree is in the foreground. It is surprising that the results are so anti-conservative with confounding as the non-synonymous changes are expected primarily outside of the foreground specified in the fitted model. Perhaps the ability of the alternative components of the model to

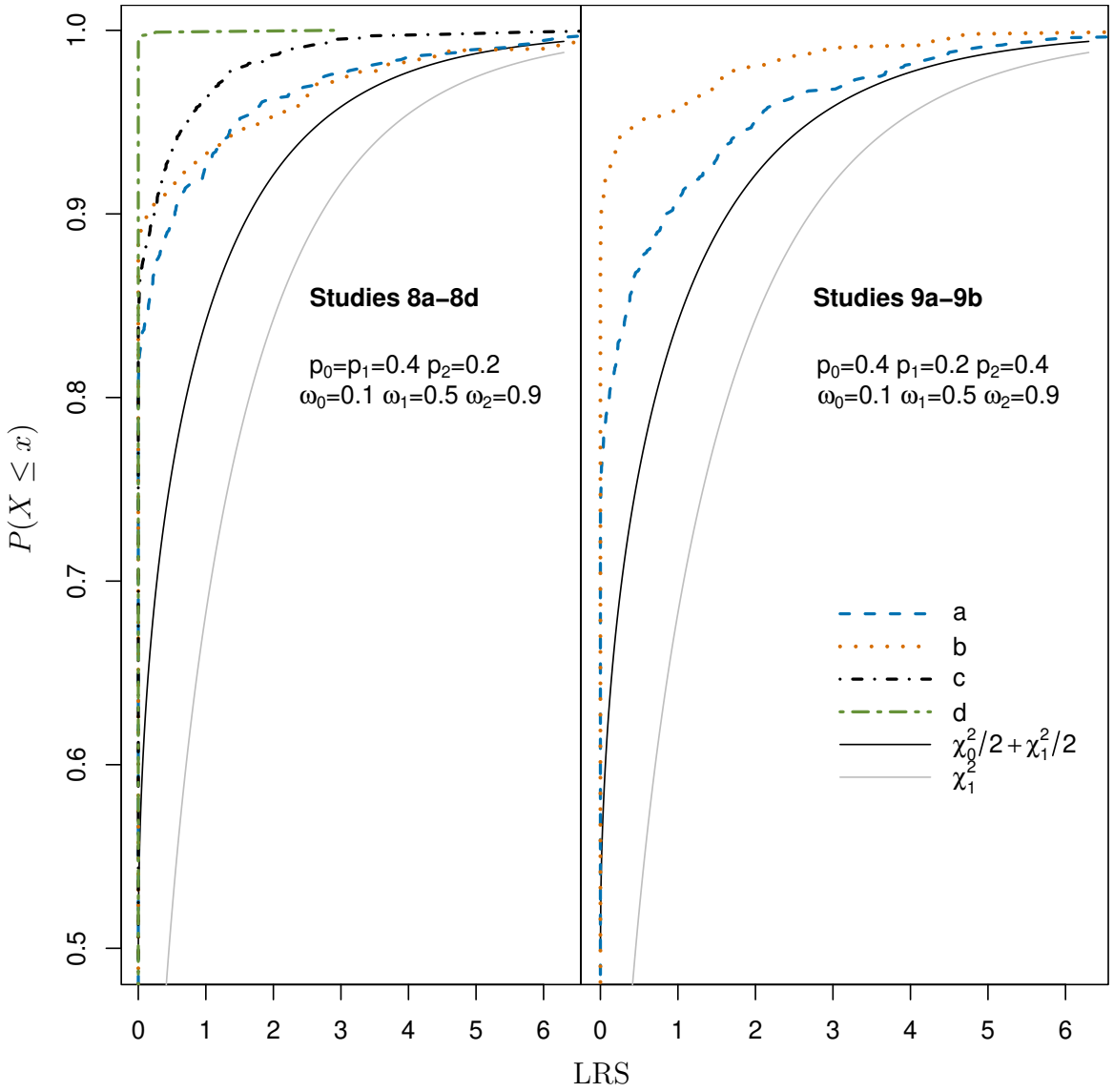


Fig. 4.6: CDFs of LR statistics under branch-site model A for data simulated in the *Misspecification of ω Distribution Scenario*. For each simulation study, 1,000 sequence alignments were generated using a balanced, 8-taxon tree topology with branch lengths summing to 3 or 6. The simulated parameter values of the ω distributions are shown in each panel. Refer to table 4.2 for detailed simulation conditions. CDFs for $\chi^2_0/2 + \chi^2_1/2$ and χ^2_1 are also included.

explain at least one nonsynonymous change when many occur due to the confounding leads to much more explanatory value under the alternative.

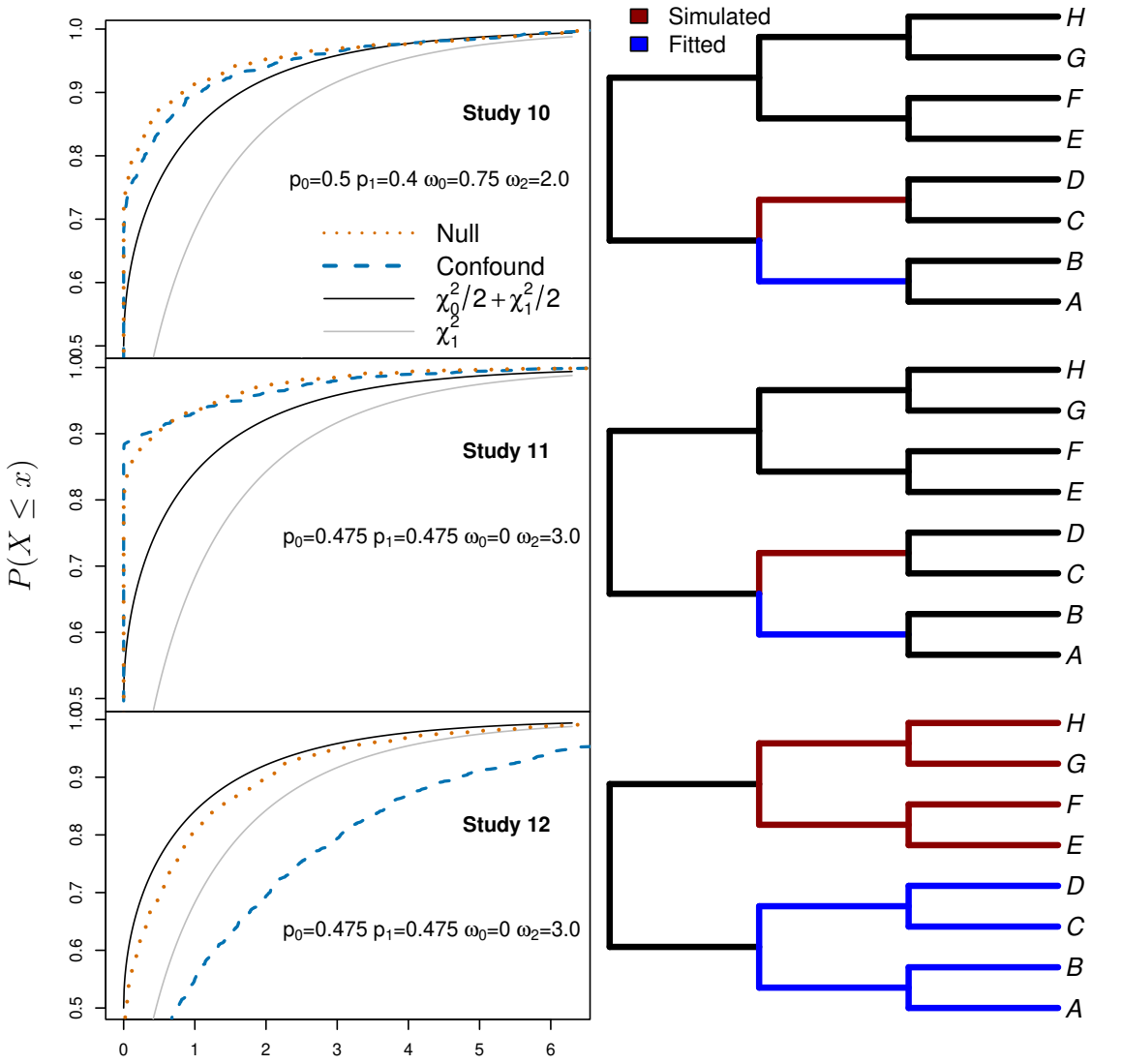


Fig. 4.7: CDFs of LR statistics under branch-site model A for data simulated in the *Confounding Foreground Scenario*. For each simulation study, 1,000 sequence alignments were generated using a balanced, 8-taxon tree topology with branch lengths summing to 3 or 6. All sequences are 5000 codons long. Under the Null simulation conditions no foreground branches were specified and under the confounding simulation conditions, the foreground is shown in the right panel, which does not match the foreground under the fitted model. Refer to table 4.3 for detailed simulation conditions. The simulated parameter values of the ω distribution are shown for each study. CDFs for $\chi^2_0/2 + \chi^2_1/2$ and χ^2_1 are also included.

4.3.5 Summary

The simulation results show that LR distributions under the null hypothesis are sometimes poorly approximated by those predicted by theory. The LR distributions

can vary heavily according to factors such as the branches considered for positive selection and the irregularity of certain parameter estimates. In some cases, the test is so lacking in power, that a re-calibration of the null LR distribution should be considered. In other scenarios, when more of the tree is in the foreground, excessively high false positive rates were observed.

Other simulation results show that when positive selection has occurred in background locations of the phylogeny, detection of positive selection in the foreground can be strongly influenced and in some cases, excessive false detection of positive selection may occur.

Chapter 5

Concluding Remarks

Some assumptions made in models of molecular evolution are extremely unrealistic. Assuming equal selection pressure over all amino sites in a protein is one example. As most proteins must maintain the capacity to fold into complex structures to complete some biological function, the majority of amino acid substitutions are selected against. Estimating a single ω parameter averaged over all sites in the protein will usually lack sufficient power to detect the relatively few sites that may be subjected to positive selection. By modelling each site as the realized value from an ω distribution, the increased power to detect a few sites under positive selection made the models practical for wider use. However, increasing model complexity to account for more of the complex evolutionary processes that give rise to the diversity in homologous proteins can have both favourable and unfavourable consequences.

Mixture models of codon evolution, such as those described in Chapter 2, violate regularity conditions required for standard likelihood theory. Indeed, likelihood theory does not support any particular large sample null distribution and the correct distribution is dependent on parameter values of the generating null distribution. Another difficultly with mixture models of codon evolution that have been described in Chapter 2 is near unidentifiability. With few taxa and small branch lengths, two very different ω distributions can give nearly the same site pattern probabilities, making estimation and inference challenging.

The modified likelihood approach, described in Chapter 2, adds a penalty for small weight on $\omega > 1$ to likelihood calculations. Simulation results show that this modification, in many cases, gives tractable LR statistic distributions that are well

approximated by a $\chi_0^2+\chi_1^2/2$ distribution and helps to adequately control false positive rates with minimal impact on power. Potential future work related to the modified likelihood approach is (i) calculation of an optimal likelihood penalty using, e.g., a cross-validation procedure and (ii) application to other models of molecular evolution.

When the null hypothesis of no positive selection at the protein level is rejected by the LR test, a determination whether positive selection has acted upon particular amino acid sites is warranted. This site-wise analysis, carried out using Bayes rule, is dependent on the ML-estimated model parameters. Parameter values that are estimated with high error can lead to unacceptably high levels of false detection of positive selection at amino acid sites. The BEB approach adjusts for uncertainty in some parameter estimates, but the approach has limitations.

In Chapter 3, SBA, a new alternative to BEB that accommodates uncertainty in all MLEs, is described. For each amino acid site, many parameter values are generated from a smoothed bootstrap distribution and substituted into the posterior probability calculation to give a distribution of posterior probabilities which reflects parameter uncertainty. By accounting for errors associated with all model parameters, SBA, relative to BEB, has advantages. Simulations under model M8, using data simulated to reflect many real data conditions with mild model misspecification, showed that SBA provided consistently better control of false positive rates than BEB. Other advantages of SBA are: (i) MLE distributions over bootstrap samples offer the ability to visually inspect the degree of regularity or irregularity of the estimation and (ii) it is comparatively simple to implement for new models. A proof-of-concept SBA implementation for branch-site model was completed within a few hours and can be found at https://github.com/Jehops/codeml_sba/. An implementation in a well developed software package such as PAML (Yang, 1997) is a future goal. Some potential refinements to the SBA approach are left for future work such as determining an optimal bandwidth parameter for smoothing. Under-smoothing will not correct irregularities in the bootstrap parameter distributions and over smoothing will remove useful information and increase bias.

Branch-site models, such as those described in Chapter 4 have been purported to have more power to detect instances of episodic positive selection that has acted upon some sites along particular lineages of the tree. Through simulation studies,

I have identified challenges with the model. With low information content in the foreground of the tree, e.g., when there is a single, shorter branch in the foreground, or when the generating model does not match well to the fitted model, the test for positive selection can be very conservative. Real data may often involve one or few true foreground branches or may have more complex ω distributions over sites. As both of these conditions have been shown to cause the LR test to be conservative, this may explain why the test is viewed as valuable by practitioners. That is, they tend to find that the model does not give spurious results, because it is most often used under conditions that we now expect to be conservative.

On the other hand, when there is more information content in the foreground, the LR test can be anti-conservative. Application of the modified likelihood approach described in Chapter 2 to branch-site model A would likely be beneficial for obtaining tractable LR distributions when more of the tree is in foreground. Perhaps of most concern are the results when the foreground of the generated model does not match with the foreground of the fitted model. Despite past reports that the model is not misled by positive selection in background branches (Zhang et al., 2005; Gharib and Robinson-Rechavi, 2013; Fletcher and Yang, 2010), a simulation result from Chapter 4 strongly contradicts this assertion. While some potential pitfalls have been identified with branch-site model A, more work is warranted to better understand why, e.g., the model can be so strongly misled by positive selection in other parts of the tree.

Supplementary Materials

6.1 Appendix I: Codon Models

6.1.1 Sites Models

Table 6.1: Codon models that allow ω to vary over amino acid sites. These models, implemented in the PAML software program (Yang, 1997), were used in simulations and real data analyses throughout the thesis.

Model	Parameters of ω Distribution	Notes	References
M0	ω	Single ω over sites	Nielsen and Yang (1998)
M1a	p_0, ω_0<1, ω_1=1	Null in M1a/M2a test	Yang et al. (2005)
M2a	p_0, p_1, ω_0<1, ω_1=1, ω_2>1	Alt. in M1a/M2a test	Yang et al. (2005)
M3	p_0, p_1, ω_0, ω_1, ω_2	Unconstrained ωs	Yang et al. (2000a)
M7	p, q	Discretized Beta	Yang et al. (2000a)
M8	p, q, p_0, ω_s>1	Discretized Beta + ω>1	Yang et al. (2000a)

6.1.2 Branch-Site Model A

Table 6.2: The ω distribution under the alternative model of branch-site model A, described in Zhang et al. (2005). Under the null model, the constraint $\omega_2 = 1$ is imposed. Branch site model A was used in the simulation studies of Chapter 4.

Site Class	Proportion	Background	Foreground
0	p_0	$\omega_0 < 1$	$\omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p1)p_0/(p_0 + p_1)$	$\omega_0 < 1$	$\omega_2 > 1$
2b	$(1 - p_0 - p1)p_1/(p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$

6.2 Appendix II: Proof of the Limiting Distribution of the Modified Likelihood

We prove below that the limiting distribution of the modified likelihood ratio statistic is $\chi_0^2/2 + \chi_1^2/2$. Assumed without proof is that the codon model considered is identifiable: for a fixed tree, no two distinct sets of parameters give exactly the same distribution of site patterns. Such results have not been established for codon models. However, there are a number of identifiability results for similar rates-across-sites (Allman et al., 2008) and covarion models (Allman and Rhodes, 2009) that suggest it is a plausible assumption. Also assumed is that third partial derivatives of the probability of a site pattern, over any set of parameters, is bounded in a neighbourhood of the true parameter values. Finally, the covariance matrix V defined below is assumed to be positive definite.

Taylor's Series

Let $\beta = (\omega_+, \psi^T)^T$ and let $\beta^0 = [1, (\psi^0)^T]^T$ where ψ^0 denotes the true generating parameter under the null hypothesis. It follows similarly as in Chen et al. (2004) that $\hat{\beta} \to \beta^0$ where $\hat{\beta}$ is the modified ML estimator. Since the convergence of the modified ML estimator $\hat{p}_+$ of p_+ is at present unclear, modified likelihood ratios are approximated through Taylor's series approximation of the log likelihoods, with respect to β at β^0, holding p_+ fixed:

$$\tilde{l}(p_+, \omega_+, \psi) - l_H(\psi^0) = L(\beta^0; p_+) + Q(\beta^0; p_+) + C(\beta^*; p_+) + C\log(p_+) \tag{6.1}$$

where L, Q and C denote the linear, quadratic and cubic terms and β^* is some value between β and β^0.

Linear Term

The linear term is

$$L(\beta^0; p_+) = (\beta - \beta^0)^T \sum_h \frac{\partial}{\partial \beta} \log p(x_h; \beta^0, p_+)$$

It is not difficult to show that the collection, $S(x_h)_\psi$, of derivatives of $\log p(x_h; \beta^0, p_+)$ with respect to ψ are independent of p_+. The other derivative is

$$\frac{\partial}{\partial \omega_+} \log p(x_h; \beta^0, p_+) = p_+(1 - p_0)\frac{\partial}{\partial \omega} p(x|1; \zeta)/p(x_h; \beta^0, p_+) =: p_+ S(x_h)_{\omega_+} \tag{6.2}$$

so that $S(x_h)^T = [S(x_h)_{\omega_+}, S(x_h)_\psi^T]$ is independent of p_+. Standard likelihood theory gives that $E[S(X_h)] = 0$, so by the Central Limit Theorem, $n^{-1/2} S_n = n^{-1/2} \sum S(X_h)$ is approximately normal with mean 0 and a covariance matrix denoted by V. Let $\delta_n^T = \sqrt{n}[p_+(\omega_+ - 1), (\psi - \psi^0)^T]$. Then

$$L(\beta^0; p_+) = n^{-1/2} S_n^T \delta_n \tag{6.3}$$

Quadratic Term

The quadratic term in (6.1) is

$$Q(\beta^0; p_+) = \frac{1}{2}(\beta - \beta^0)^T l^{(2)}(\beta^0)(\beta - \beta^0) \tag{6.4}$$

where

$$l^{(2)}(\beta^0) = \sum_h Q^{(1)}(x_h; \beta^0, p_+) + \sum_h \frac{\partial}{\partial \beta} \log p(x_h; \beta^0, p_+) \frac{\partial}{\partial \beta} \log p(x_h; \beta^0, p_+)^T \tag{6.5}$$

and $p_H(x_h)Q^{(1)}(x_h; \beta^0, p_+)_{ij}$ is the partial derivative of $p(x_h; \beta^0, p_+)$ with respect to β_i and β_j. It is not difficult to see that $p_H(x_h)Q^{(1)}(x_h; \beta^0, p_+)_{ij}$ is independent of p_+ unless $i = j = 1$, in which case it equals p_+ times a partial derivative of the form $\partial^2 p(x|1; \zeta)/\partial \omega^2$. Standard likelihood theory gives that $E[Q^{(1)}(X_h; \beta^0, p_+)] = 0$. Thus the Central Limit Theorem gives that $Q_n^{(1)} := \sum Q^{(1)}(x_h; \beta^0, p_+) = O_P(n^{1/2})$ for any fixed p_+. Since $Q_n^{(1)}$ depends linearly on p_+, $Q_n^{(1)} = O_p(n^{1/2})$ uniformly in p_+. Substituting in (6.5), then (6.4) and using the relationships between derivatives of $\log p(x_h; \beta^0, p_+)$ and $S(x_h)$ established earlier,

$$Q(\beta^0; p_+) = \frac{1}{2}(\beta - \beta^0)^T Q_n^{(1)}(\beta - \beta^0) + \frac{1}{2n}\delta_n^T \sum_h S(x_h)S(x_h)^T \delta_n \tag{6.6}$$

Let $Q_n^{(2)} = \sum_h S(x_h)S(x_h)^T - nV$. Since $E[S(x_h)S(x_h)^T] = V$, the Central Limit Theorem gives that $Q_n^{(2)} = O_P(n^{1/2})$. Since

$$Q(\beta^0; p_+) = \frac{1}{2}(\beta - \beta^0)^T Q_n^{(1)}(\beta - \beta^0) + \frac{1}{2n}\delta_n^T Q_n^{(2)} \delta_n - \frac{1}{2}\delta_n^T V \delta_n \tag{6.7}$$

for $\delta_n = O_P(1)$ we have that

$$Q(\beta^0; p_+) = -\frac{1}{2}\delta_n^T V \delta_n + O_P(n^{-1/2}) \tag{6.8}$$

Cubic Term

The cubic term in (6.1) is

$$C(\beta^*; p_+) = \frac{1}{6} \sum_{ijk} l^{(3)}(\beta^*; p_+)_{ijk}(\beta - \beta^0)_i(\beta - \beta^0)_j(\beta - \beta^0)_k \tag{6.9}$$

where $l^{(3)}(\beta^*; p_+)_{ijk}$ denotes the third partial derivative of the log likelihood with respect to β_i, β_j and β_k. Since third partial derivatives of $\log p(x_h; \beta^0, p_+)$ are assumed to be bounded in a neighbourhood of β^0 by, say, $M(x_h)$, $|l^{(3)}(\beta^*; p_+)|/n$ is bounded by $n^{-1}\sum_h M(X_h)$. It follows by the Law of Large Numbers that $l^{(3)}(\beta^*; p_+) = O_P(n)$ and consequently that for $\beta - \beta^0 = O_P(n^{-1/2})$, $C(\beta^*; p_+) = O_p(n^{-1/2})$. Combining

(6.3) and (6.8) in (6.1) gives that for $\delta_n = O_P(1)$,

$$\tilde{l}(p_+, \omega_+, \psi) - l_H(\psi^0) = n^{-1/2} S_n^T \delta_n - \frac{1}{2}\delta_n^T V \delta_n + C\log(p_+) + O_P(n^{-1/2}) \qquad (6.10)$$

Approximation with the modified MLE

Since $\hat{\delta}$ has not been shown to be equal to $O_p(1)$, (6.10) does not immediately apply. However, since $b_n = \hat{\delta}/|\hat{\delta}| = O_p(1)$, the argument for (6.10) gives that

$$\tilde{l}(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\psi^0) = |\hat{\delta}|^2\{n^{-1/2} S_n^T b_n/|\hat{\delta}| - \frac{1}{2} b_n^T V b_n + O_P(n^{-1/2})\} + C\log(\hat{p}_+) \quad (6.11)$$

Since V is positive definite, the right-hand side of (6.11) becomes negative if $|\hat{\delta}|$ diverges. However, since $\hat{\beta}$ is a maximizer, the difference in (6.11) is always positive. Thus it must be the case that $\hat{\delta} = O_p(1)$. This implies that $\hat{\psi} - \psi^0 = O_P(n^{-1/2})$ and that $\hat{p}_+(\hat{\omega}_+ - 1) = O_P(n^{-1/2})$. Similarly as in Lemma 1 of Chen et al. (2004), with probability, converging to 1, $\hat{p}_+ \geq \epsilon$ for some $\epsilon > 0$, so that $\hat{\omega}_+ - 1 = O_P(n^{-1/2})$. Thus the approximation (6.10) applies with $\beta = \hat{\beta}$:

$$\begin{aligned} \tilde{l}(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\psi^0) &= n^{-1/2} S_n^T \hat{\delta}_n - \frac{1}{2}\hat{\delta}_n^T V \hat{\delta}_n + C\log(\hat{p}_+) + O_P(n^{-1/2}) \\ &\leq \max_{\delta, p_+}\{n^{-1/2} S_n^T \delta - \frac{1}{2}\delta^T V \delta + C\log(p_+)\} + O_P(n^{-1/2}) \qquad (6.12) \end{aligned}$$

The inequality (6.12) holds when maximization is restricted so that the maximizing δ and p_+ correspond to a valid β and p_+: $\delta_{\omega_+} \geq 0$ and $p_+ \leq 1$. If the corresponding β and p_+ are denoted as $\tilde{\beta}$ and $\tilde{p}_+$, since the maximizing δ and p_+ are $O_P(1)$, the expression in (6.12) is the same as $\tilde{l}(\tilde{p}_+, \tilde{\omega}_+, \tilde{\psi}) - l_H(\psi^0)$ up to the order indicated. Since $\tilde{l}(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\psi^0)$ is larger than $\tilde{l}(\tilde{p}_+, \tilde{\omega}_+, \tilde{\psi}) - l_H(\psi^0)$, the reverse inequality holds in (6.12) as well, implying that it is an equality. The maximized value of $C\log(p_+)$ is $C\log(1) = 0$. Thus (6.12) is

$$\tilde{l}(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\psi^0) = \max_{\delta}\{n^{-1/2} S_n^T \delta - \frac{1}{2}\delta^T V \delta\} + O_P(n^{-1/2}) \qquad (6.13)$$

The log likelihood under the null

No modification of the likelihood is considered under the null, so standard ML results gives that

$$l_H(\psi) - l_H(\psi^0) = (n^{-1/2} S_{n\psi})^T V_{\psi\psi}^{-1}(n^{-1/2} S_{n\psi}) + O_P(n^{-1/2}) \qquad (6.14)$$

The difference between (6.12) and (6.14) gives that the modified likelihood ratio satisfies that

$$\tilde{l}(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\psi) = \max_\delta \{2n^{-1/2} S_n^T \delta - \delta^T V \delta\} - (n^{-1/2} S_{n\psi})^T V_{\psi\psi}^{-1} (n^{-1/2} S_{n\psi}) + O_P(n^{-1/2}) \tag{6.15}$$

Maximization under the alternative

Omitting details, after simplification, maximizing over δ with δ_{ω_+} fixed gives

$$\max_\delta \{2n^{-1/2} S_n^T \delta - \delta^T V \delta\} = (n^{-1/2} S_{n\psi})^T V_{\psi\psi}^{-1} (n^{-1/2} S_{n\psi}) + 2\delta_{\omega_+} n^{-1/2} S^c_{n\omega_+} - \delta^2_{\omega_+} V^c_{\omega_+} \tag{6.16}$$

where

$$S^c_{n\omega_+} = S_{n\omega_+} - V_{\omega_+\psi} V_{\psi\psi}^{-1} S_{n\psi}, \quad V^c_{\omega_+} = V_{\omega_+\omega_+} - V_{\omega_+\psi} V_{\psi\psi}^{-1} V_{\psi\omega_+} \tag{6.17}$$

If $S^c_{n\omega_+} < 0$ in (6.16) the right hand side is decreasing in δ_{ω_+} and so, subject to the restriction that $\delta_{\omega_+} \geq 0$, the maximizing $\delta_{\omega_+} = 0$. Otherwise the maximizer is $n^{-1/2} S^c_{n\omega_+} / \sqrt{V^c_{\omega_+}}$. Substituting in (6.16) and (6.15) gives

$$\tilde{l}(\hat{p}_+, \hat{\omega}_+, \hat{\psi}) - l_H(\psi) = [n^{-1/2} S^c_{n\omega_+} / \sqrt{V^c_{\omega_+}}]^2_+ + O_P(n^{-1/2}) \tag{6.18}$$

Distribution of $S^c_{n\omega_+}$

Because $n^{-1/2} S^c_{n\omega_+}$ is a linear transformation of $n^{-1/2} S_n$ which has an approximate normal distribution with mean 0 and covariance matrix V, it too has a normal distribution. It has mean 0 and variance

$$\begin{aligned} \mathrm{Var}(n^{-1/2} S^c_{n\omega_+}) &= \mathrm{Var}(n^{-1/2} S_{n\omega_+}) - 2V_{\omega_+\psi} V_{\psi\psi}^{-1} \mathrm{Cov}(n^{-1/2} S_{n\psi}, n^{-1/2} S_{n\omega_+}) \\ &\quad + V_{\omega_+\psi} V_{\psi\psi}^{-1} \mathrm{Var}(n^{-1/2} S_{n\psi}) V_{\psi\psi}^{-1} V_{\psi\omega_+} \\ &= V_{\omega_+\omega_+} - 2V_{\omega_+\psi} V_{\psi\psi}^{-1} V_{\psi\omega_+} + V_{\omega_+\psi} V_{\psi\psi}^{-1} V_{\psi\psi} V_{\psi\psi}^{-1} V_{\psi\omega_+} = V^c_{\omega_+} \end{aligned}$$

Thus $n^{-1/2} S^c_{n\omega_+} / \sqrt{V^c_{\omega_+}}$ is approximately standard normal. It follows from $Z \sim N(0,1)$ giving $Z^2 \sim \chi^2_1$ and Z^2 being independent of the event that $Z > 0$, that the limiting distribution of W is $\chi^2_1/2 + \chi^2_0/2$.

6.3 ModL Supplementary Figures

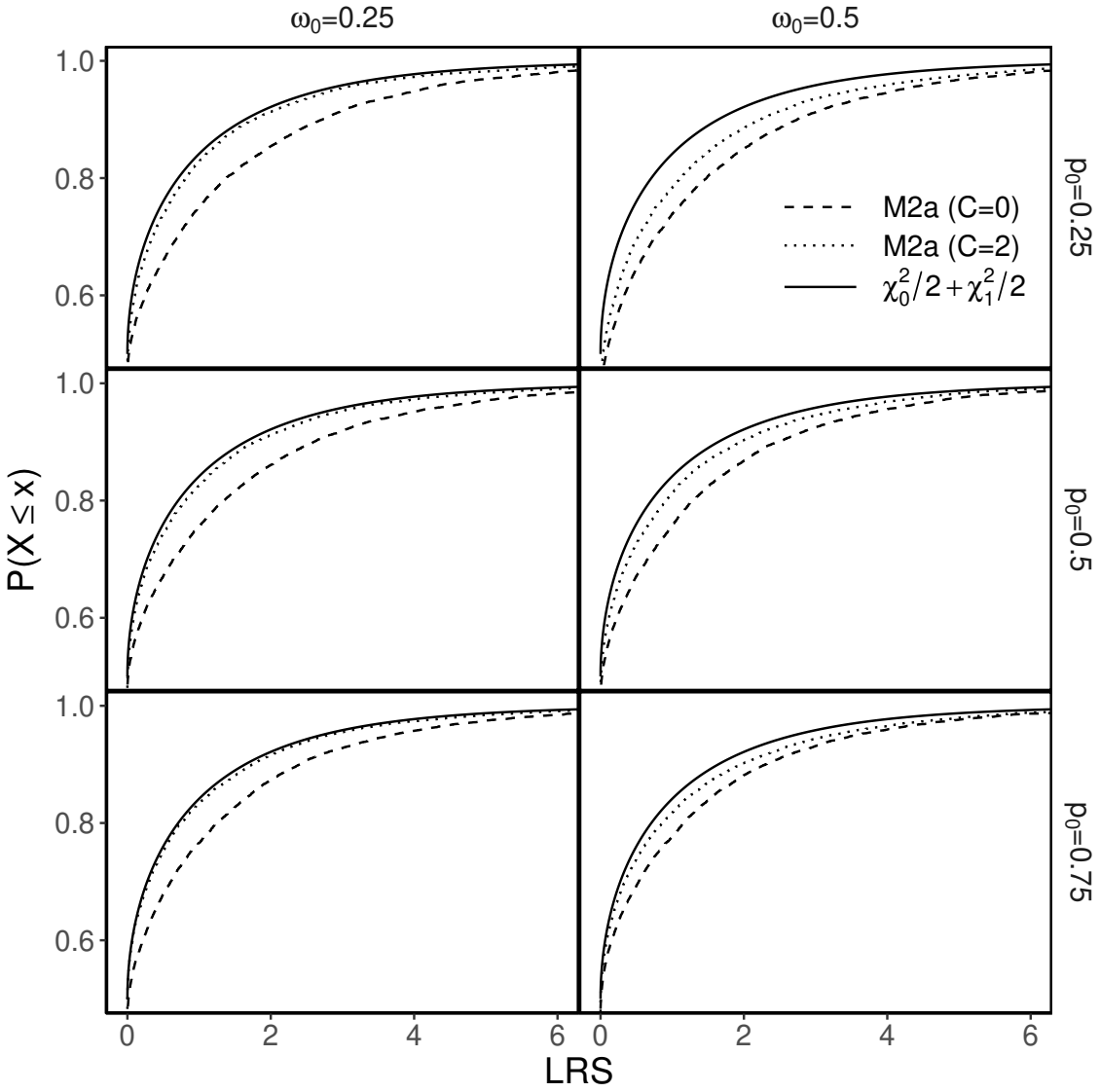


Fig. 6.1: CDFs of likelihood (C=0) and modified likelihood (C=2) ratio statistics under the nested model pair M1a/M2a for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a 5-taxon tree topology with branch lengths summing to 6. The value of ω_0 and its weight, p_0, used to generate the data are shown as column and row labels. CDFs for $\chi^2_0/2 + \chi^2_1/2$ are also included.

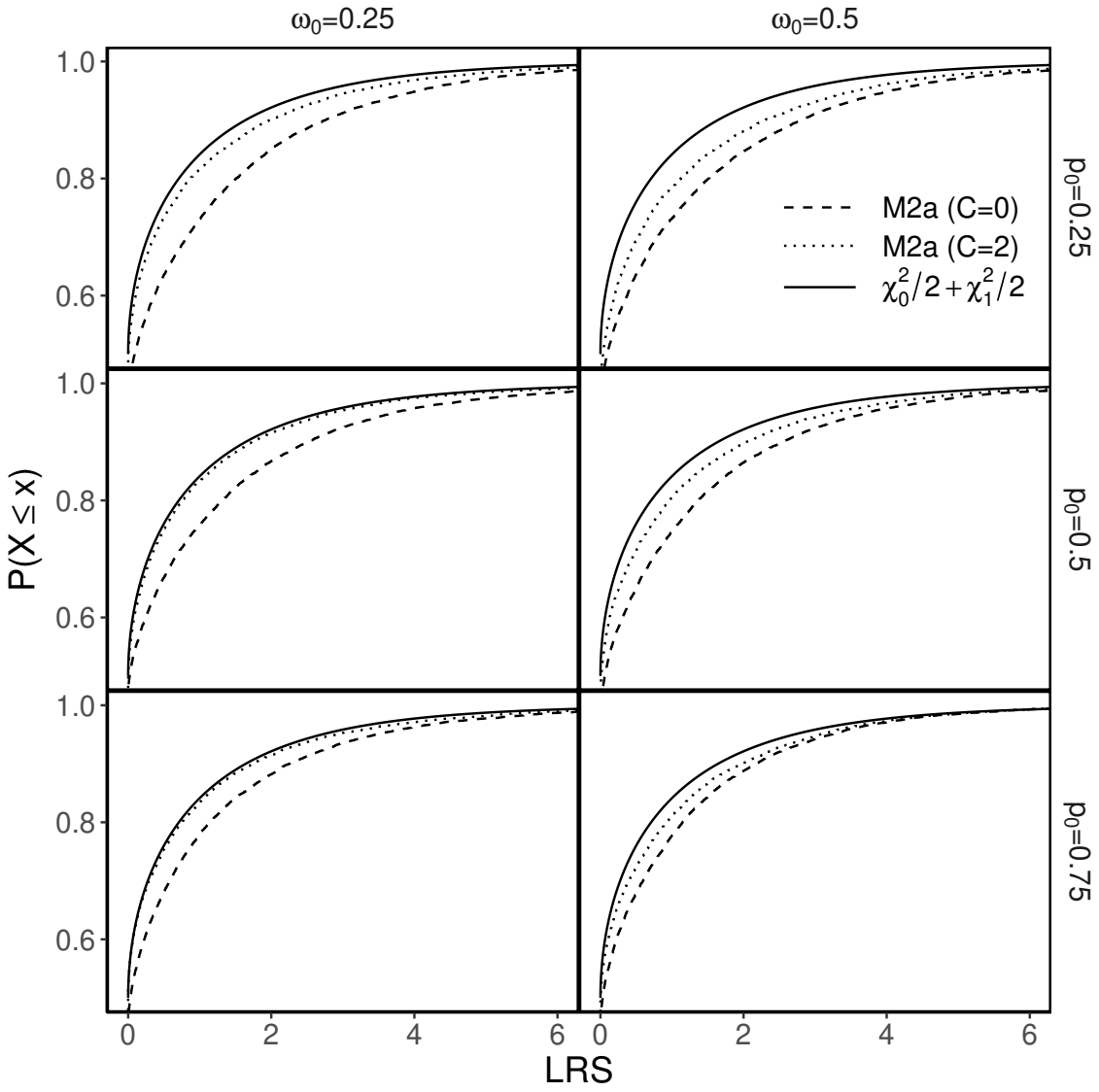


Fig. 6.2: CDFs of LR (C=0) and modified LR (C=2) statistics under M1a/M2a nested model pairs for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a 5-taxon tree topology with branch lengths summing to 9. The value of ω_0 and its weight, p_0, used to generate the data are shown as column and row labels. CDFs for $\chi_0^2/2 + \chi_1^2/2$ are also included.

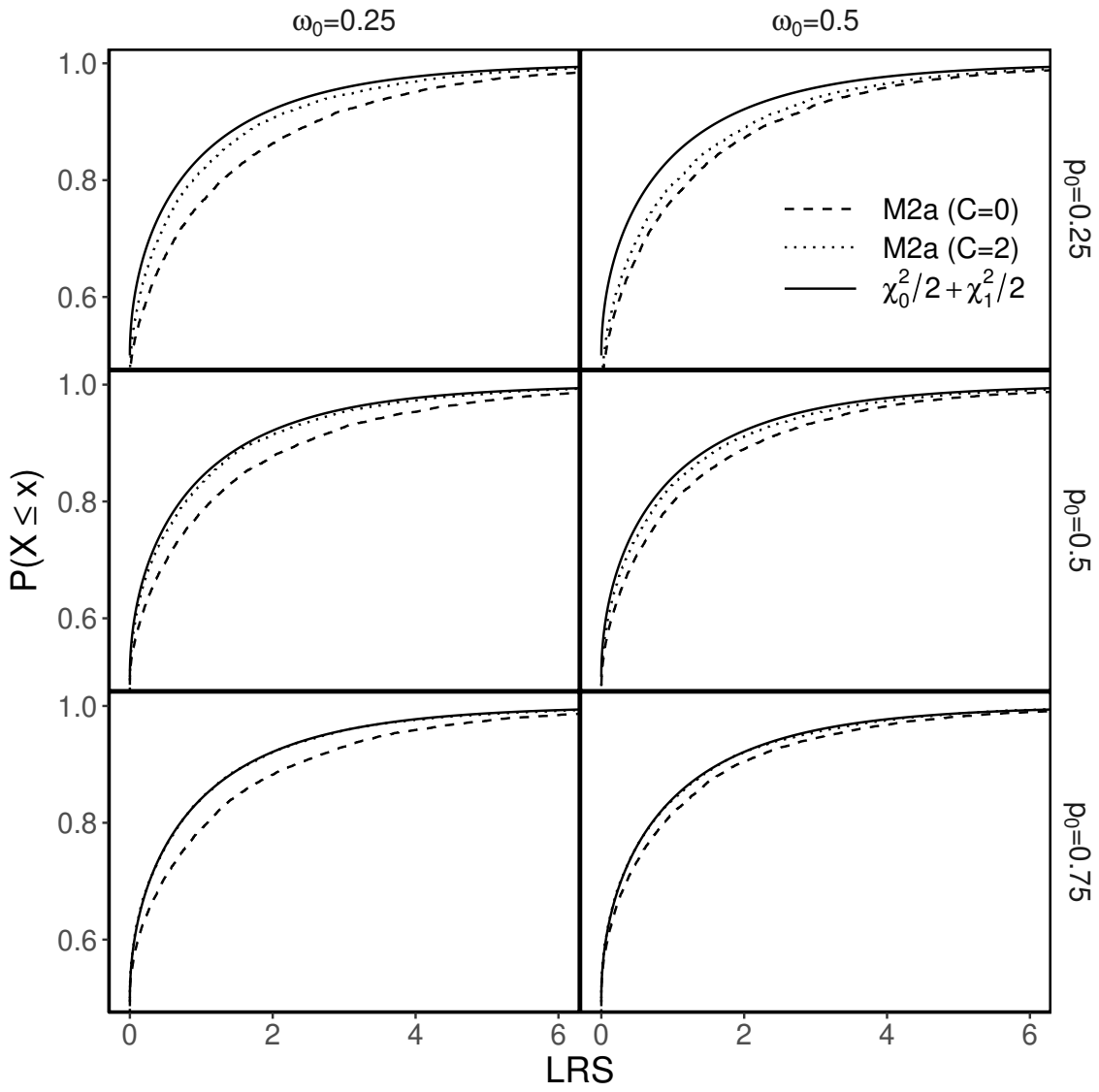


Fig. 6.3: CDFs of LR (C=0) and modified LR (C=2) statistics under M1a/M2a nested model pairs for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a balanced, 10-taxon tree topology with branch lengths summing to 3. The value of ω_0 and its weight, p_0, used to generate the data are shown as column and row labels. CDFs for $\chi^2_0/2 + \chi^2_1/2$ are also included.

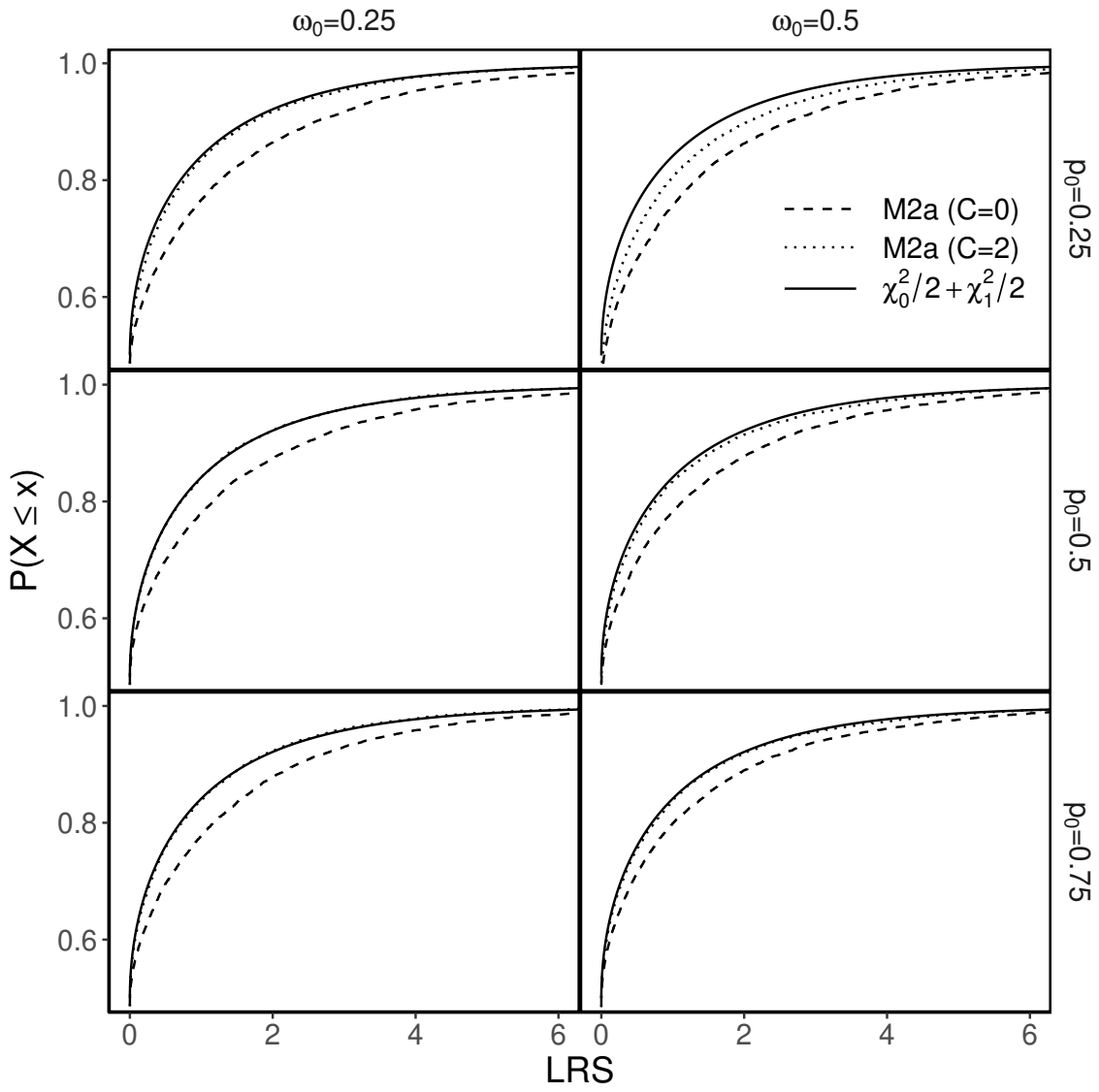


Fig. 6.4: CDFs of LR (C=0) and modified LR (C=2) statistics under M1a/M2a nested model pairs for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a balanced, 10-taxon tree topology with branch lengths summing to 6. The value of ω_0 and its weight, p_0, used to generate the data are shown as column and row labels. CDFs for $\chi^2_0/2 + \chi^2_1/2$ are also included.

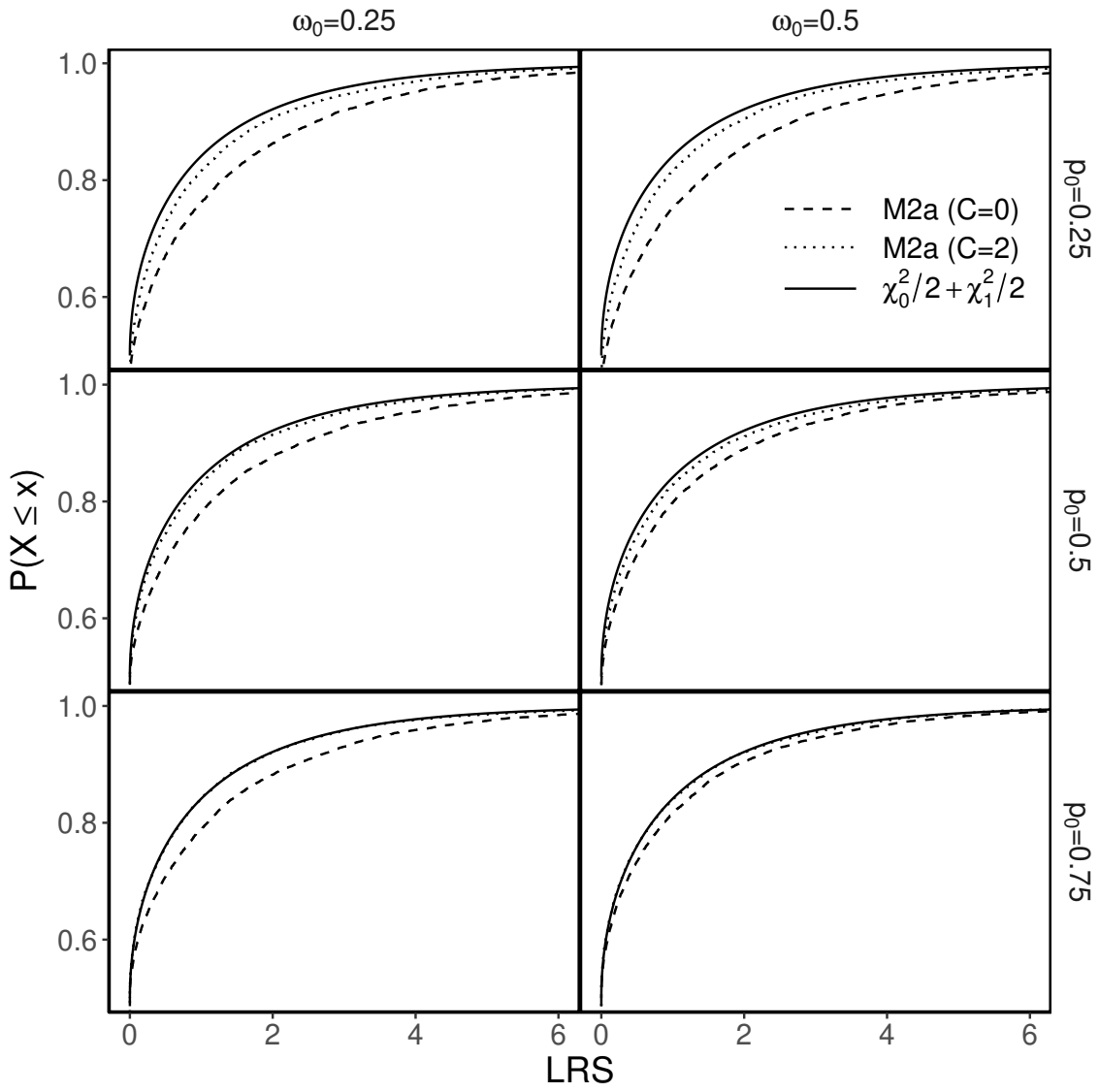


Fig. 6.5: CDFs of LR (C=0) and modified LR (C=2) statistics under M1a/M2a nested model pairs for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a balanced, 10-taxon tree topology with branch lengths summing to 9. The value of ω_0 and its weight, p_0, used to generate the data are shown as column and row labels. CDFs for $\chi_0^2/2 + \chi_1^2/2$ are also included.

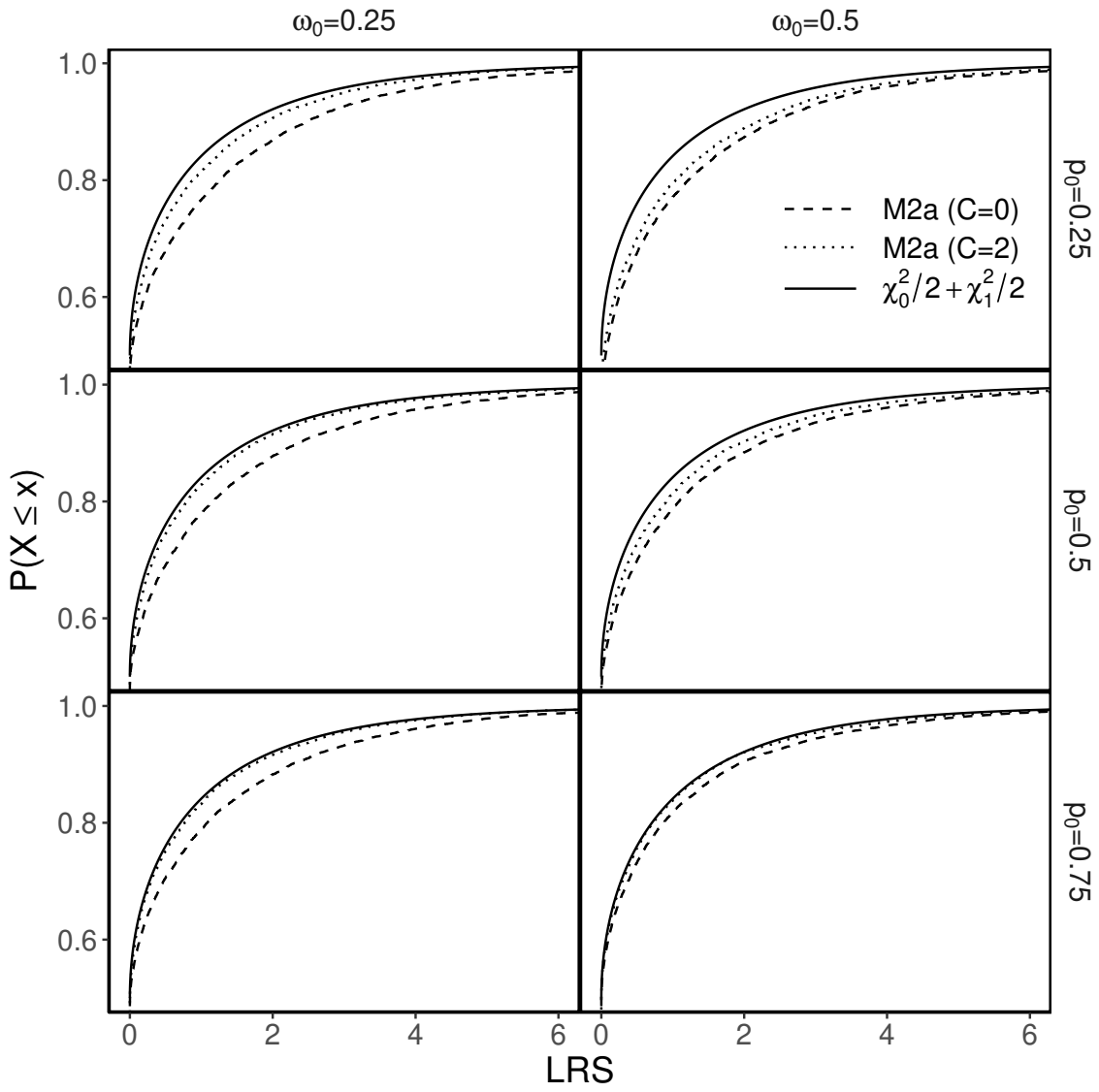


Fig. 6.6: CDFs of LR (C=0) and modified LR (C=2) statistics under M1a/M2a nested model pairs for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a balanced, 32-taxon tree topology with branch lengths summing to 3. The value of ω_0 and its weight, p_0, used to generate the data are shown as column and row labels. CDFs for $\chi_0^2/2 + \chi_1^2/2$ are also included.

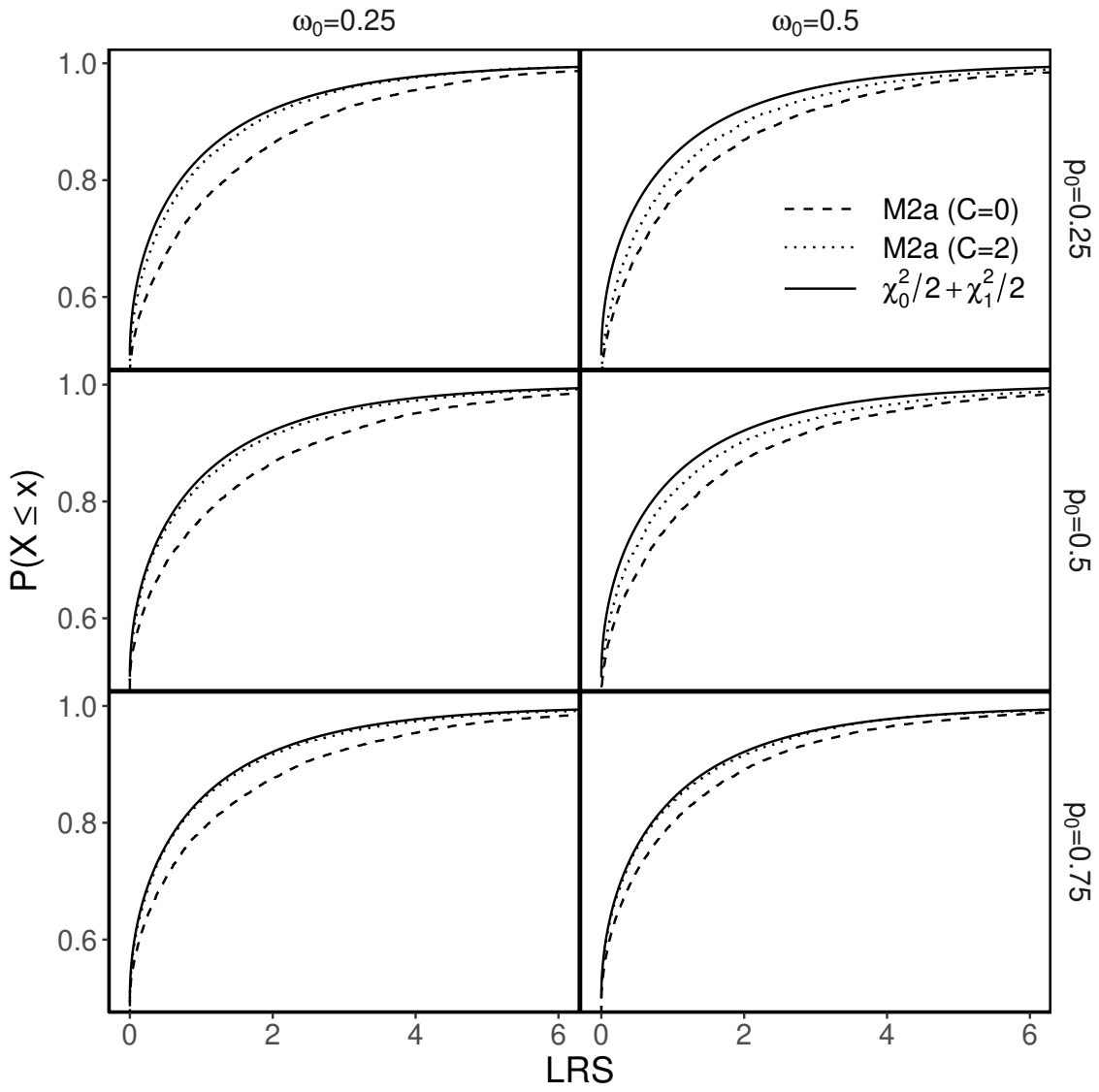


Fig. 6.7: CDFs of LR (C=0) and modified LR (C=2) statistics under M1a/M2a nested model pairs for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a balanced, 32-taxon tree topology with branch lengths summing to 6. The value of ω_0 and its weight, p_0, used to generate the data are shown as column and row labels. CDFs for $\chi_0^2/2 + \chi_1^2/2$ are also included.

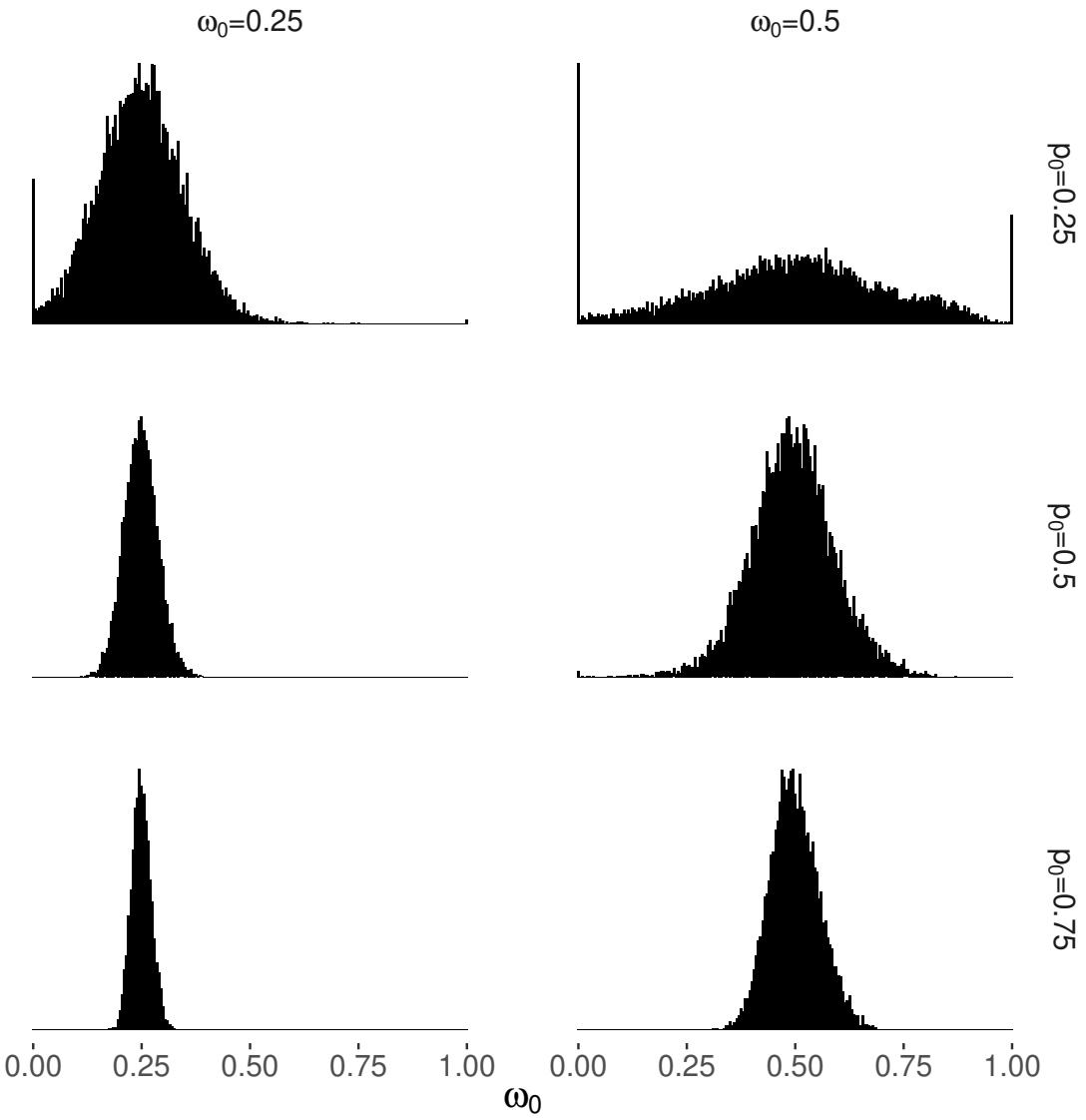


Fig. 6.8: MLEs of the ω_0 parameter under model M1a for six simulation settings. For each simulation setting, 10,000 sequence alignments were generated with two site classes, $\omega < 1$ and $\omega = 1$ using a 5-taxon tree topology with branch lengths summing to 3. The value of ω_0 and its weight, p_0, used to simulate the data are shown as column and row labels.

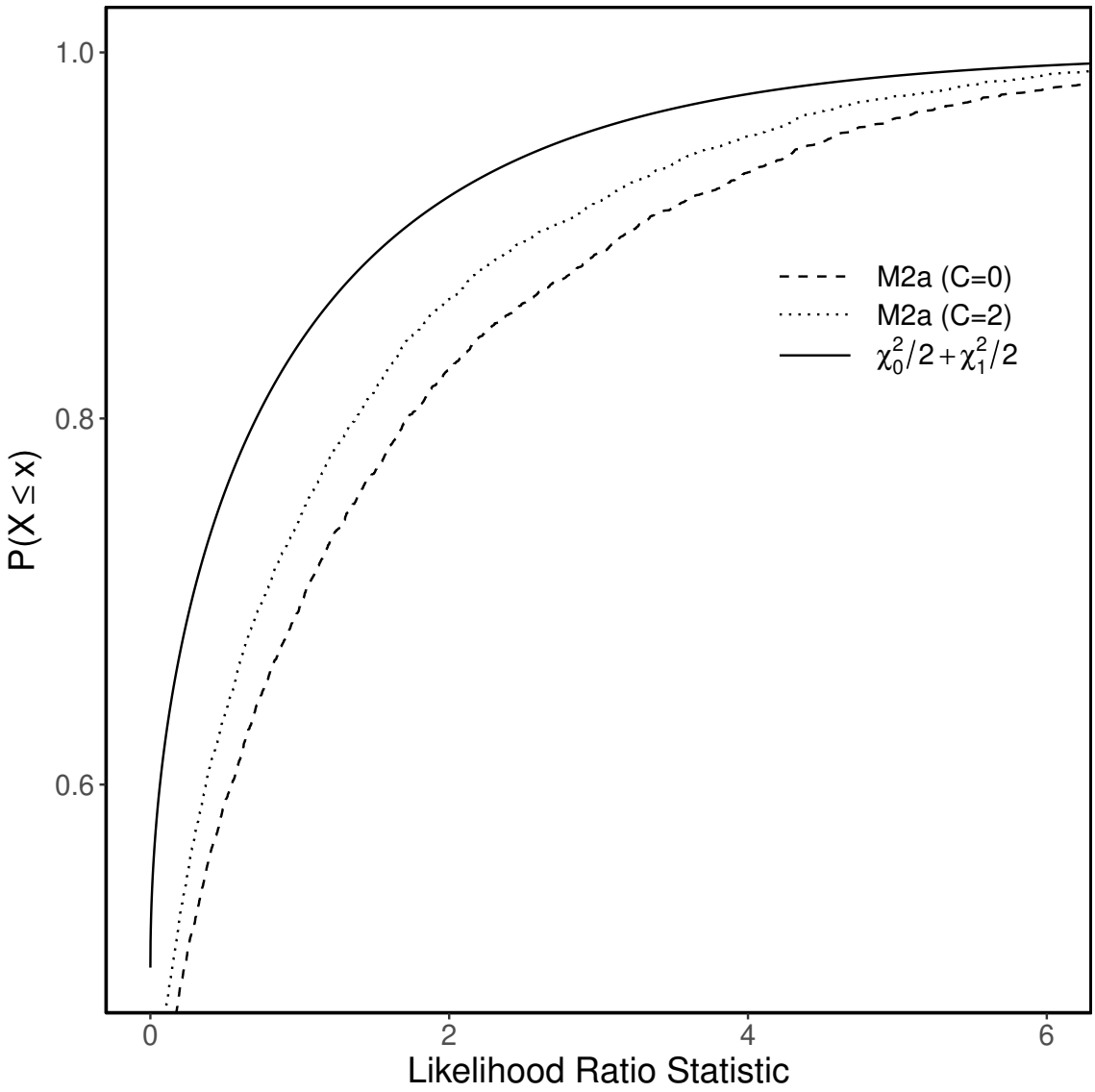


Fig. 6.9: CDFs of LR statistics without (C=0) and with (C=2) likelihood modification after pre-screening the data with M0/M1a LR tests. The modified LR statistics were calculated under the nested model pair M1a/M2a for 4987 simulated sequence alignments that were rejected under the M0/M1a null hypothesis of only one ω site class. The alignments were simulated with 25% of the sites evolving under $\omega = 0.5$ and the remaining sites evolving under $\omega = 1$ using a 5-taxon tree topology with branch lengths summing to 3. A modified likelihood tuning parameters of $C = 2$ was used. A $\chi_0^2/2 + \chi_1^2/2$ CDF is also included.

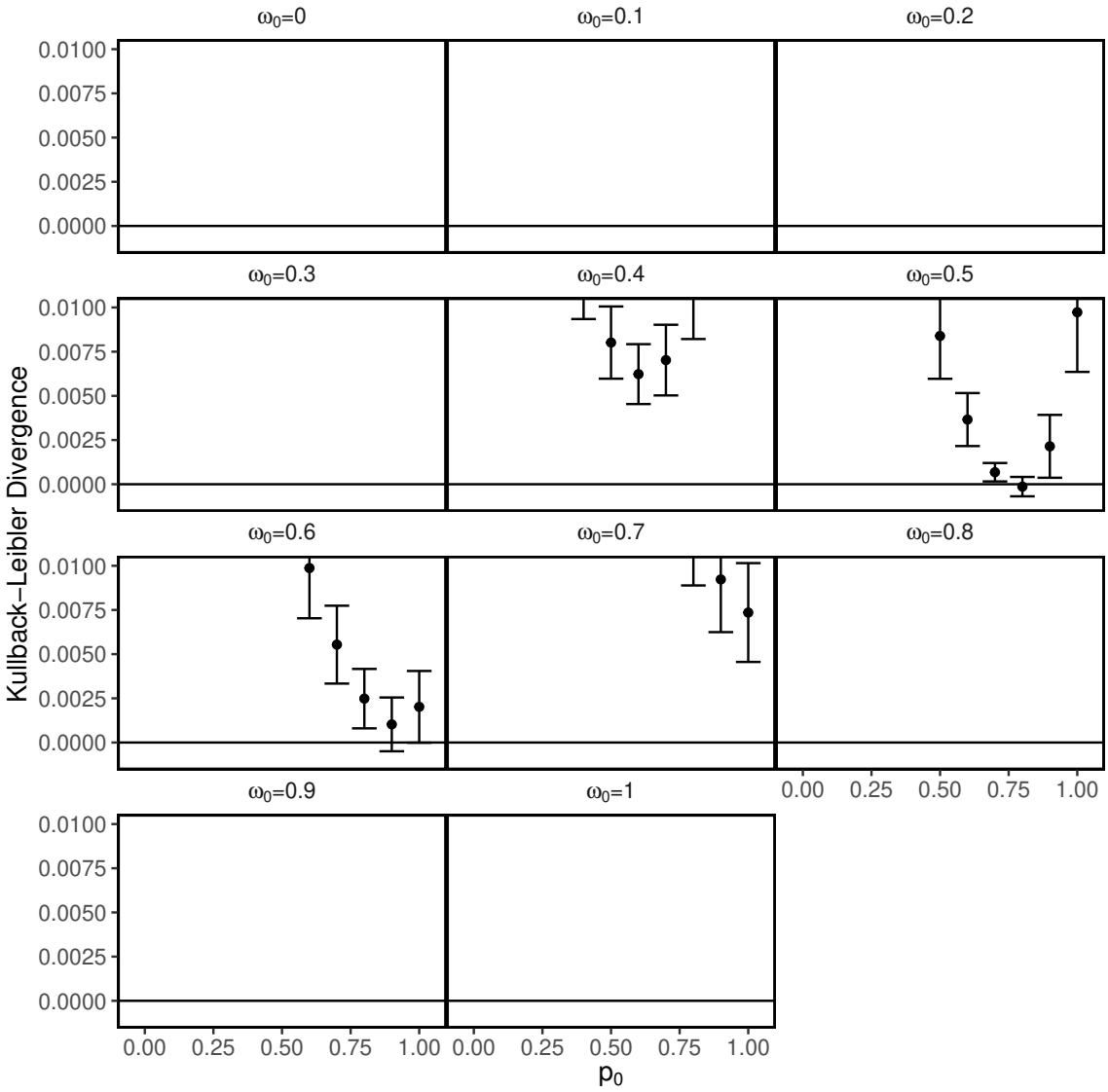


Fig. 6.10: Approximations of the Kullback-Leibler divergences between the distributions of site likelihoods for the generating model and other mixing distributions. The approximations were obtained as the mean lnL difference between 10,000 site patterns generated under model M1a using a 5-taxon tree with branch lengths summing to 3 and the mixing distribution $(p_0, \omega_0) = (0.75, 0.5)$, and other mixing distributions with varying weights on values of ω ranging from 0 to 1. Error bars for two standard errors $(s_{KL}/\sqrt{10000})$ above and below each Kullback-Leibler estimate are included. Points missing from each plot are above the visible range.

6.4 Optimality of ROC curve using the true mixing distribution

Established here is that the ROC curve using posterior probabilities calculated with the true distribution of ω across sites is optimal in the sense that for any x-axis value, its y-axis value is always the largest attainable. The result is a consequence of the Neyman-Pearson Lemma (Neyman and Pearson, 1933) which was used to establish optimality of likelihood ratio tests in the case of simple null and alternative hypotheses where all parameters of the model are known.

For a given site, let $\phi(X, W)$ denote a test for positive selection at that site. Specifically, $\phi(X, W) = 1$ when the test finds in favour of positive selection and is 0 otherwise. It depends on X, which represents the data at the site and possibly on random W, independent of the data at the site; for all tests considered in the paper, W is data from all other sites. For instance, the test, $\phi_1(X)$, using posterior probabilities calculated with the true distribution of ω, uses only the data at the site and sets $\phi_1^{(k)}(X) = 1$ if $P(\omega > 1|X) > k$ for threshold k. In obtaining the ROC curve, each choice of k gives a different false positive rate (the x-axis value). The true positive rate for a given k is the corresponding y-value. Since

$$P(\omega > 1|X) > k \Longleftrightarrow \frac{p(X|\omega > 1)P(\omega > 1)}{p(X|\omega > 1)P(\omega > 1) + p(X|\omega \leq 1)P(\omega \leq 1)} > k,$$

a brief argument can be given to show that the test rejects when

$$p(X|\omega > 1)P(\omega > 1)/\{p(X|\omega \leq 1)P(\omega \leq 1)\} > c. \tag{6.1}$$

where to simplify notation, $c = [1/k - 1]^{-1} > 0$ is set. This test is denoted as $\phi_1^{(c)}$.

The ROC curve for $\phi_1^{(c)}$ is optimal if, for any other test, $\phi_0^{(t)}$, the y-axis value (true positive rate) for $\phi_1^{(c)}$ at least as large as that of $\phi_0^{(t)}$ whenever the x-axis values (false positive rates) for the two tests are the same. Much like with codon frequencies, which can refer to observed or population frequencies, the ROC curve for a test sometimes refers to the random quantity that varies from data set to data set depending upon the number of true and false positives for that data set. According to this definition it is impossible to guarantee that $\phi_1^{(c)}$ always gives a uniformly better ROC curve. For instance, there is always some small probability of unusual data sets where the positively selected sites show only synonymous changes, in which case even poor tests may give better ROC curves. The population version of an ROC curve refers to the curve with the limiting proportions of false positives on the x-axis and true positives on the y-axis; values that can be approximated by averaging over many simulated data sets. It is for the population version of the ROC curve that the optimality property holds.

The limiting false positive rate of $\phi_1^{(c)}$, is the same as $\phi_0^{(t)}$, if c and t are chosen so that

$$P(\phi_1^{(c)}(X) = 1, \omega \leq 1) = P(\phi_0^{(t)}(X, W) = 1, \omega \leq 1). \tag{6.2}$$

The aim is to show that when (6.2) holds, meaning that the x-axis values of the ROC curve are the same, the probability of a true positive for $\phi_1^{(c)}$ (y-axis value) is at least as large as that of $\phi_0^{(t)}$:

$$P(\phi_1^{(c)}(X) = 1, \omega > 1) - P(\phi_0^{(t)}(X, W) = 1, \omega > 1) \geq 0 \tag{6.3}$$

In using the Neyman-Pearson Lemma, it is convenient to express the true positive probability as an expectation of an indicator function. For any event, A, $P(A) = E[I\{A\}]$ where $I\{A\}$ is 1 or 0 depending upon whether A is true or not. Since $\phi_1^{(c)}(X)I\{\omega > 1\} = 1$ or 0,

according to whether the event $\phi_1^{(c)}(X) = 1, \omega > 1$ holds or not, (6.3) can be expressed as

$$E[\phi_1^{(c)}(X) I\{\omega > 1\}] - E[\phi_1^{(t)}(X, W) I\{\omega > 1\}] \geq 0 \tag{6.4}$$

As a final simplification, let $\phi_0^{(*t)}(X) = E[\phi_0^{(t)}(X, W)|X]$. Then

$$\begin{aligned} E[\phi_0^{(t)}(X, W) I\{\omega > 1\}] &= E[E[\phi_0^{(t)}(X, W) I\{\omega > 1\}|X]] \\ &= E[I\{\omega > 1\} E[\phi_0^{(t)}(X, W)]|X]] = E[I\{\omega > 1\}\phi_0^{(*t)}(X)] \end{aligned} \tag{6.5}$$

Substituting (6.5) in (6.4), it is obtained that $\phi_1^{(c)}$ is optimal $\phi_0^{(t)}$ if

$$E[\phi_1^{(c)}(X) I\{\omega > 1\}] - E[\phi_0^{(*t)}(X) I\{\omega > 1\}] \geq 0. \tag{6.6}$$

whenever (6.2) holds. With this simplification the result follows immediately from the proof of the Neyman-Pearson Lemma as now shown.

$$\begin{aligned} E[\phi_1^{(c)}(X) I\{\omega > 1\} - \phi_0^{(*t)}(X) I\{\omega > 1\}] &= \sum_{x,\omega'} \{\phi_1^{(c)}(x) - \phi_0^{(*t)}(x)\} I\{\omega' > 1\} P(X = x, \omega = \omega') \\ &= \sum_{x} \{\phi_1^{(c)}(x) - \phi_0^{(*t)}(x)\} \sum_{\omega' > 1} P(X = x, \omega = \omega') \\ &= \sum_{x} \{\phi_1^{(c)}(x) - \phi_0^{(*t)}(x)\} P(X = x, \omega > 1) \\ &= \sum_{x} \{\phi_1^{(c)}(x) - \phi_0^{(*t)}(x)\} p(x|\omega > 1) P(\omega > 1). \end{aligned} \tag{6.7}$$

For any x such that the test $\phi_0^{(c)}(x)$ rejects, by (6.1)

$$p(x|\omega > 1) P(\omega > 1) > c p(x|\omega \leq 1) P(\omega \leq 1) \tag{6.8}$$

Since $0 \leq \phi_0^{(*t)}(x) \leq 1$ and since the test rejects when $\phi_1^{(c)}(x) = 1$

$$\{\phi_1^{(c)}(x) - \phi_0^{(*t)}(x)\} = 1 - \phi_0^{(*t)}(x) \geq 0. \tag{6.9}$$

Combining (6.8)-(6.9) gives that

$$\{\phi_1^{(c)}(x) - \phi_0^{(*t)}(x)\} p(x|\omega > 1) p(\omega > 1) \geq c\{\phi_1^{(c)}(x) - \phi_0^{(*t)}(x)\} p(x|\omega \leq 1) p(\omega \leq 1) \tag{6.10}$$

Consider now x such that $\phi_0^{(c)}$ does not reject. Then

$$p(x|\omega > 1) P(\omega > 1) \leq c p(x|\omega \leq 1) P(\omega \leq 1) \tag{6.11}$$

and since $\phi_0^{(*t)}(x) \geq 0$ and $\phi_1^{(c)}(x) = 0$,

$$\{\phi_1^{(c)}(x) - \phi_0^{(*t)}(x)\} = -\phi_0^{(*t)}(x) \leq 0. \tag{6.12}$$

Combining (6.11)-(6.12) gives (6.10). In short, (6.10) is satisfied for all x. Substituting in

(6.7),

$$E[\phi_1^{(c)}(X)I\{\omega > 1\} - \phi_0^{(*t)}(X)I\{\omega > 1\}] \geq c\sum_x \{\phi_1^{(c)}(x) - \phi_0^{(*t)}(x)\}p(x|\omega \leq 1)p(\omega \leq 1) \quad (6.13)$$

Now

$$\begin{aligned}\sum_x \phi_1^{(c)}(x)p(x|\omega \leq 1)p(\omega \leq 1) &= \sum_x \phi_1^{(c)}(x)P(X = x, \omega \leq 1) \\ &= \sum_{x,\omega\leq 1} \phi_1^{(c)}(x)p(x,\omega) \\ &= \sum_{x|\phi_1^{(c)}(x)=1,\omega\leq 1} p(x,\omega) = P(\phi_1^{(c)}(X) = 1, \omega \leq 1) \quad (6.14)\end{aligned}$$

Similarly,

$$\sum_x \phi_0^{(*t)}(x)p(x|\omega \leq 1)p(\omega \leq 1) = P(\phi_0^{(t)}(X) = 1, \omega \leq 1) \quad (6.15)$$

Substituting (6.14)-(6.15) in (6.13) and using (6.2) gives the desired result that

$$E[\phi_1^{(c)}(X)I\{\omega > 1\} - \phi_0^{(*t)}(X)I\{\omega > 1\}] \geq c\{P(\phi_1^{(c)}(X) = 1, \omega \leq 1) - P(\phi_0^{(t)}(X) = 1, \omega \leq 1)\} = 0$$

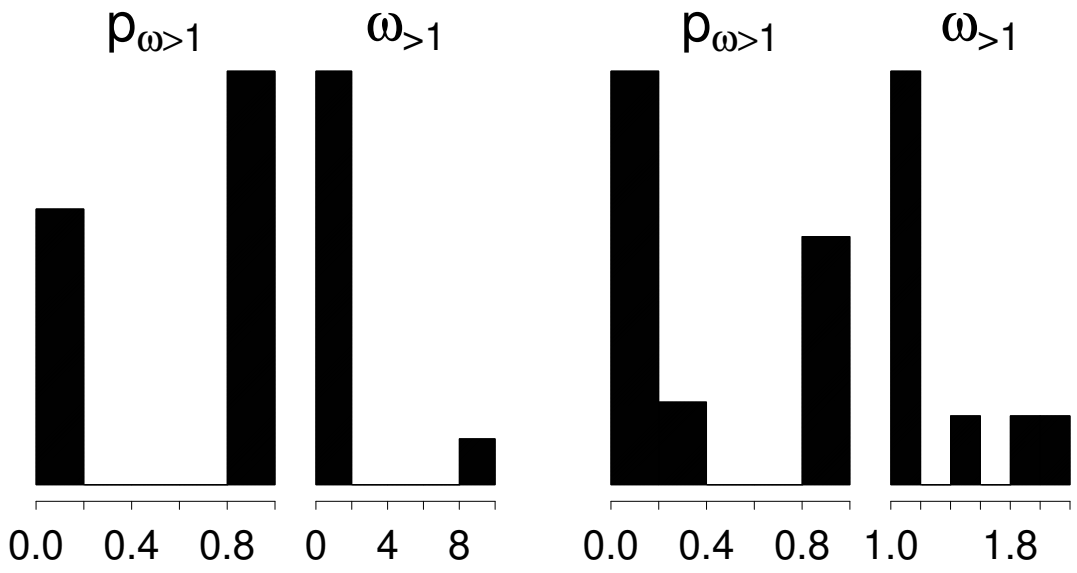
6.5 SBA Supplementary Figures and Tables

Table 6.3: False positive rates for each simulation study after application of the likelihood ratio (LR) test. The *LR Test* column lists the proportion of significant tests. False positive rates are only included when they differ from the corresponding rates without the LR Test. All rates under BEB and SBA remained the same with and without the LR test. Values in parentheses denote the change in the rate after applying the LR Test. A posterior probability threshold of 0.95 was used for classifying sites to be under positive selection.

Study	Misspecification	ω distribution	LR Test		NEB	
			M2a	M8	M2a	M8
1	None	100% 1	0.10	0.09	**0.05** (-0.29)	0.04 (-0.31)
2	None	50% 0.5, 50% 1	0.08	0.17		
3	None	50% 1 50% 1.5	0.94	0.94	**0.33** (-0.02)	**0.35** (-0.02)
4	None	45% 0, 45% 1, 10% 5	1.00	1.00		
5	Mild	100% 1	0.23	1.00		
6	Mild	50% 0.5, 50% 1	0.13	1.00		
7	Mild	50% 1, 50% 1.5	0.92	1.00	**0.26** (-0.04)	
8	Mild	45% 0, 45% 1, 10% 5	1.00	1.00		
9	Heavy	100% 1	1.00	0.00		
10	Heavy	50% 0.5, 50% 1	0.71	0.74	**0.36** (-0.17)	**0.38** (-0.12)

Table 6.4: Analysis of the *tax* gene. Shown are the estimated total tree lengths (TL), maximum likelihood parameters (MLEs), -log likelihoods (-lnL), and the range of site posterior probabilities (Pr) under models M2a and M8 using BEB and SBA to classify sites. The range of posterior probabilities are shown for three categories of sites: invariant, single synonymous substitution (SSS), and single nonsynonymous substitution (SNS).

	M2a	M8
TL	0.128	0.128
MLEs	$p_{\omega>1} = 1.0$, $\omega_{>1} = 4.87$	$p_{\omega>1} = 1.0$, $\omega_{>1} = 4.87$
-lnL	892.0	892.0
Invariant Sites (159)		
BEB Pr Range	$0.552, 0.607$	$0.689, 0.732$
SBA Pr Range	$0.543, 0.596$	$0.761, 0.799$
SSS Sites (2)		
BEB Pr Range	$0.589, 0.590$	$0.718, 0.719$
SBA Pr Range	$0.578, 0.579$	$0.787, 0.787$
SNS Sites (21)		
BEB Pr Range	$0.911, 0.927$	$0.961, 0.968$
SBA Pr Range	$0.871, 0.892$	$0.990, 0.991$



(a) M2a (b) M8

Fig. 6.11: MLE distributions of the $p_{\omega>1}$ and $\omega_{>1}$ parameters under M2a and M8. Histograms are over simulated datasets for which the null hypothesis of no positive selection was rejected by a likelihood ratio test (10 of 100 under M2a and 9 of 100 under M8). Data were simulated under *irregular* conditions: 5 taxa, 100% $\omega = 1$.

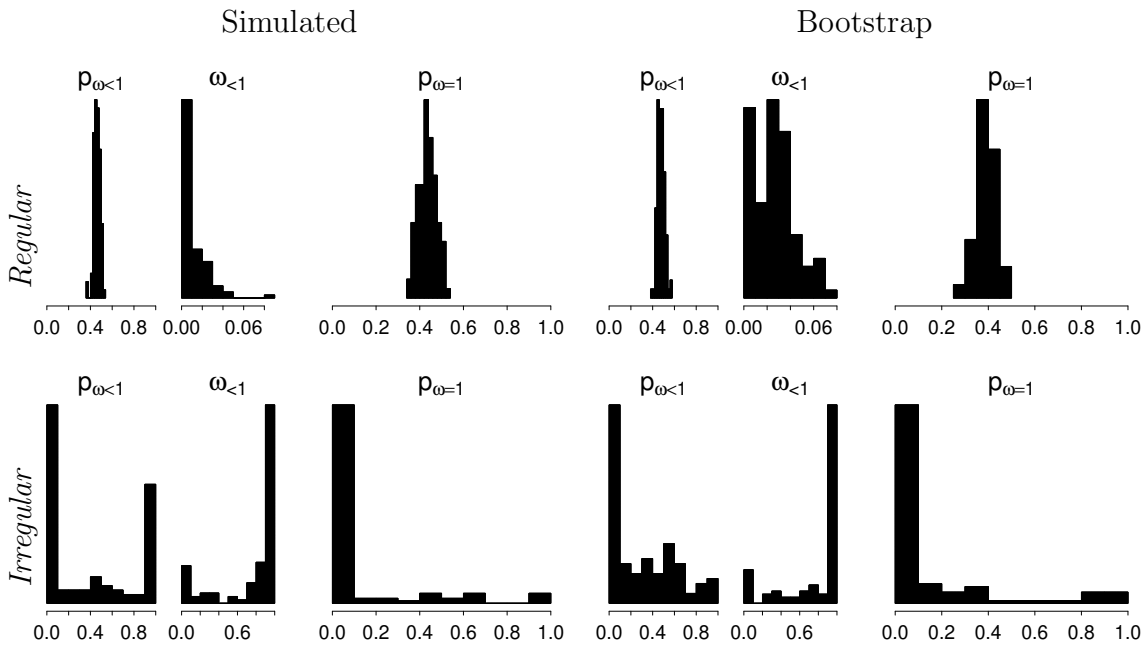
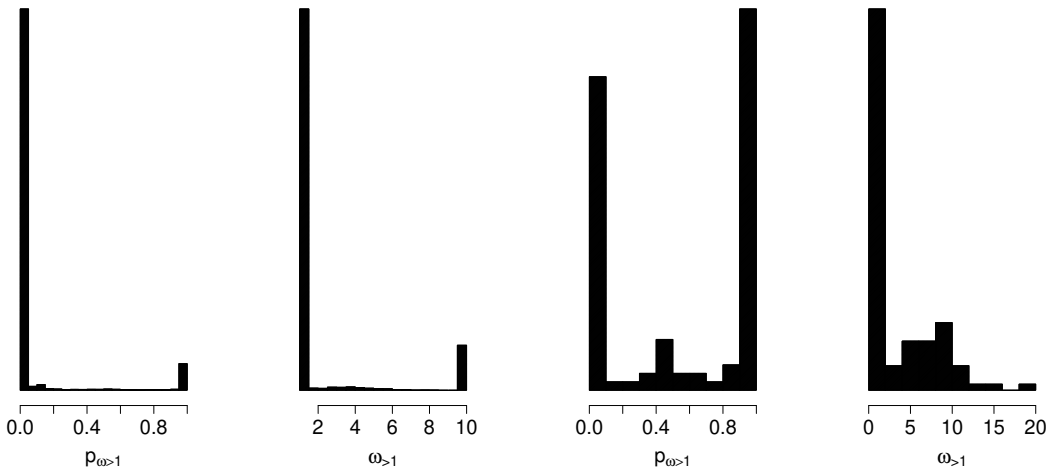


Fig. 6.12: MLE distributions of the $p_{\omega<1}$, $\omega_{<1}$, and $p_{\omega=1}$ parameters under M2a. Histograms are over 100 simulated and bootstrap datasets with the bootstrap datasets generated by sampling from one simulated dataset. Data were simulated under *regular* and *irregular* conditions.

regular simulation conditions: 5 taxa, 45% ω=0, 45% ω=0.5, and 10% ω=5
irregular simulation conditions: 5 taxa, 100% ω=1



(a) M2a (b) M8

Fig. 6.13: Distributions of the $p_{\omega>1}$ and $\omega_{>1}$ parameters associated with positive selection estimated under models M2a and M8. Histograms are over 10,000 bootstrap datasets drawn from a dataset simulated under *difficult* conditions (5 taxa, 100% $\omega = 1$). Under M2a 1000 of the $\omega_{>1}$ values estimated to be biologically unreasonable were capped at 10.

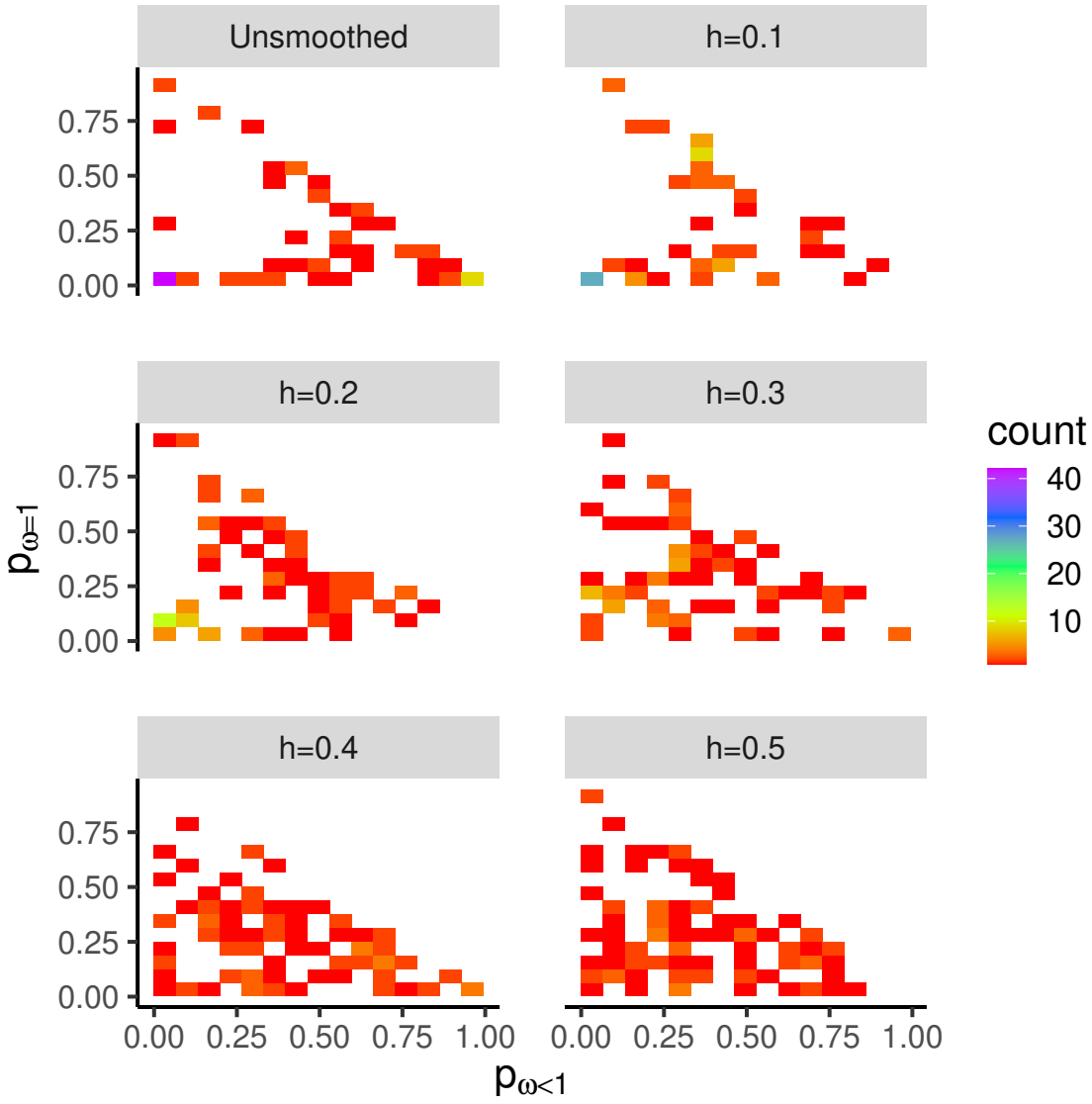


Fig. 6.14: MLEs of the $p_{\omega<1}$ and $p_{\omega=1}$ parameters under model M2a before and after smoothing using a uniform kernel with different bandwidth parameters. The parameters were estimated over 100 bootstrap samples under *irregular* simulation conditions (5 taxa, 100% sites $\omega = 1$).

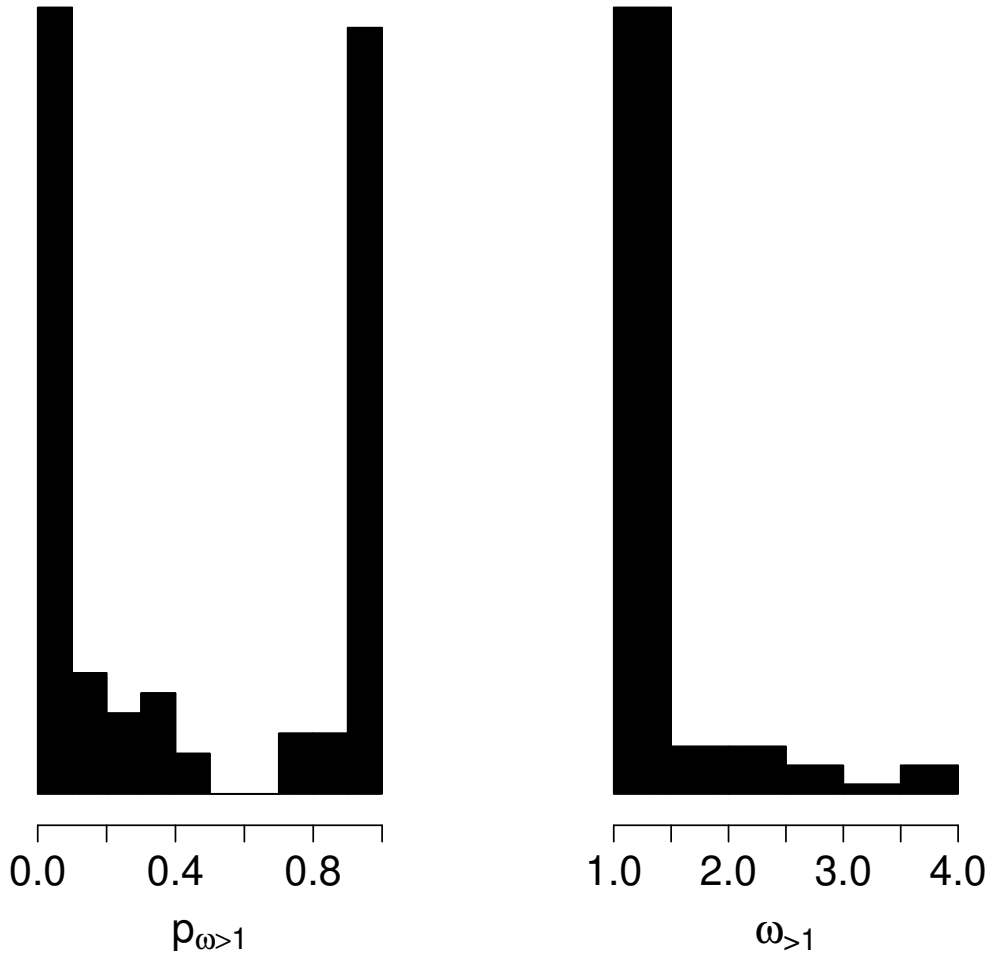
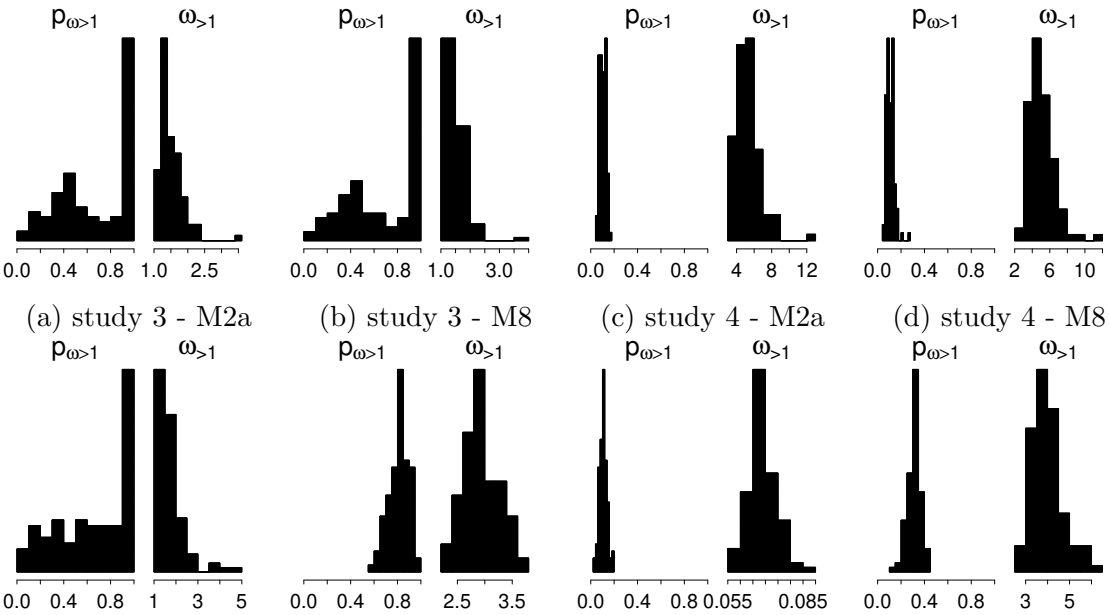


Fig. 6.15: Distributions of the $p_{\omega>1}$ and $\omega_{>1}$ parameters associated with positive selection and estimated under model M2a. Histograms are over 100 simulated datasets simulated under *difficult* conditions (100% $\omega = 1$) and using a 30-taxon tree.



(a) study 3 - M2a (b) study 3 - M8 (c) study 4 - M2a (d) study 4 - M8

(e) study 7 - M2a (f) study 7 - M8 (g) study 8 - M2a (h) study 8 - M8

Fig. 6.16: MLE distributions of the $p_{\omega>1}$ and $\omega_{>1}$ parameters under models M2a and M8 for two different simulation scenarios: without model misspecification (*Correct Model*, studies 3 and 4) and with mild model misspecification (*Mild Misspecification*, studies 7 and 8). The data were simulated using a 5-taxon tree topology. In studies 3 and 7, 50% of the sites were simulated under neutral evolution ($\omega = 1$) and 50% of the sites under positive selection ($\omega = 1.5$). In studies 4 and 8, 45% of the sites were simulated under purifying selection ($\omega = 0$), 45% under neutral evolution ($\omega = 1$) and 10% under positive selection ($\omega = 5$).

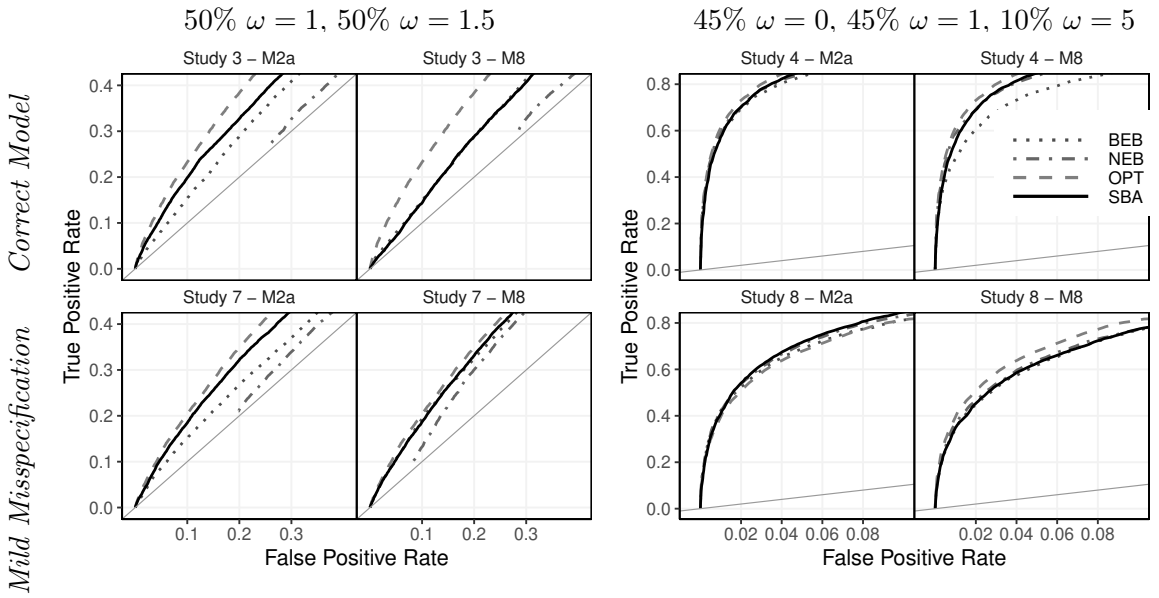
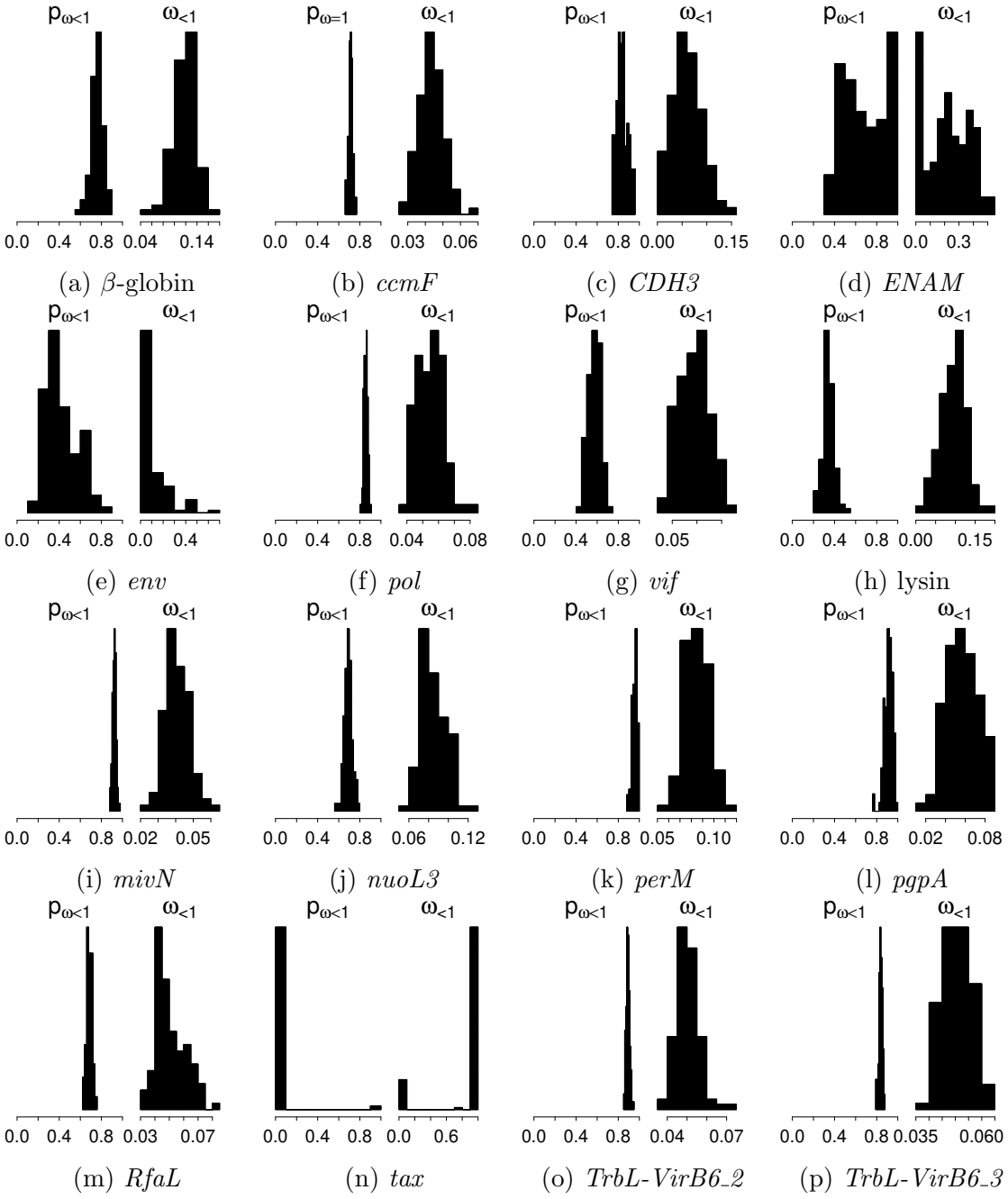


Fig. 6.17: ROC curves for the detection of sites under positive selection for BEB, NEB, and SBA analyses of data generated under two different simulation scenarios: without model misspecification (*Correct Model*, studies 3 and 4) and with mild model misspecification (*Mild Misspecification*, studies 7 and 8). Likelihood ratio tests were performed prior to site-wise analyses. The data were simulated using a 5-taxon tree topology. In studies 3 and 7, 50% of the sites were simulated under neutral evolution ($\omega = 1$) and 50% of the sites under positive selection ($\omega = 1.5$). In studies 4 and 8, 45% of the sites were simulated under purifying selection ($\omega = 0$), 45% under neutral evolution ($\omega = 1$) and 10% under positive selection ($\omega = 5$). Each plot includes a line for the lower bound (y=x) and an expected upper bound (OPT) when classification is made using the generating model parameters. The curves for studies 4 and 8 and study 7 under M8 are identical to those without pre-screening because the null hypotheses of likelihood ratio tests were rejected for all simulated datasets. The curves for NEB do not always cover the whole range of false positive rates, because NEB sometimes estimates the ω distribution with all mass on $\omega > 1$. In these cases, even with a posterior probability cut-off of 1, NEB still incorrectly classifies sites to be under positive selection.



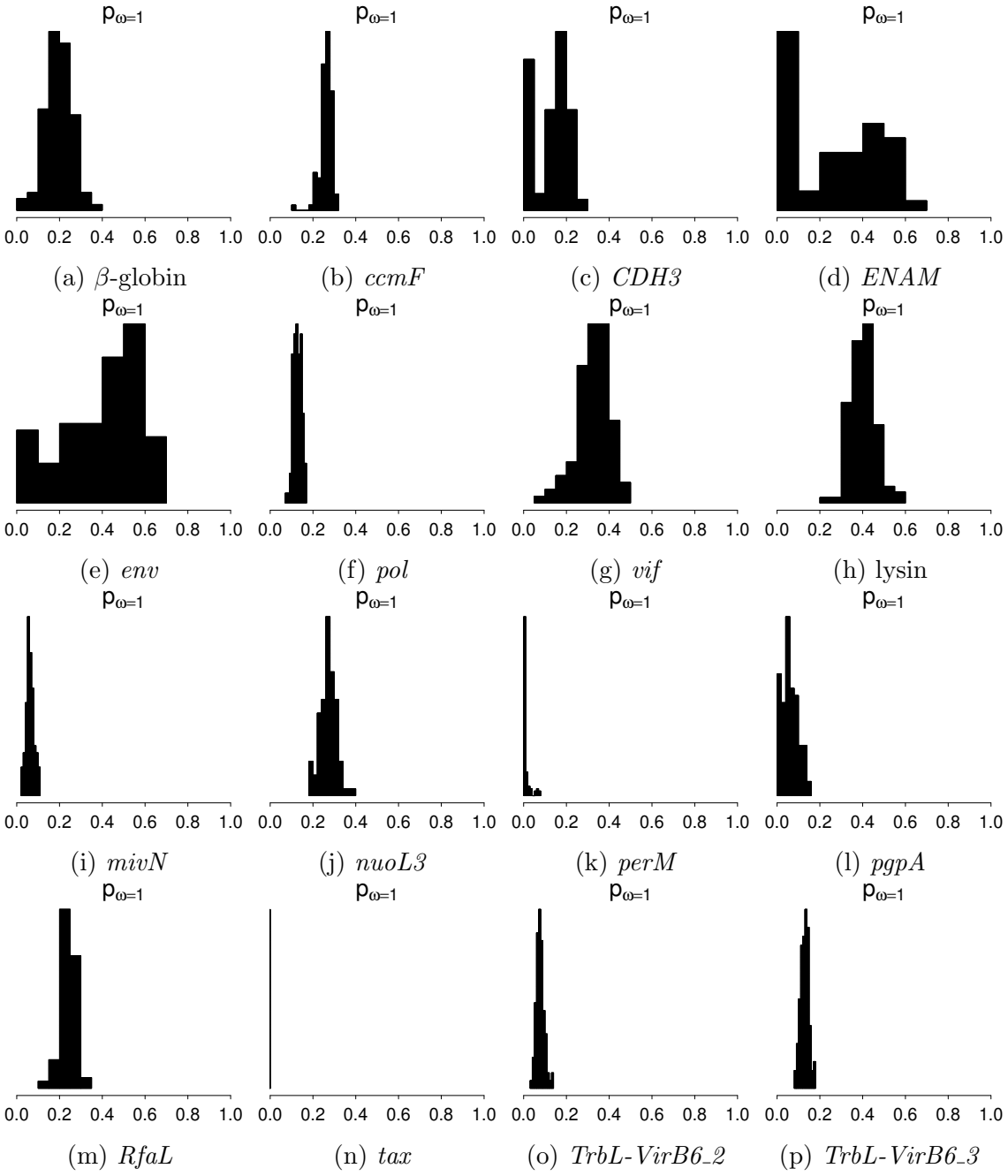
(a) β-globin (b) *ccmF* (c) *CDH3* (d) *ENAM*

(e) *env* (f) *pol* (g) *vif* (h) lysin

(i) *mivN* (j) *nuoL3* (k) *perM* (l) *pgpA*

(m) *RfaL* (n) *tax* (o) *TrbL-VirB6_2* (p) *TrbL-VirB6_3*

Fig. 6.18: Distributions of the $p_{\omega<1}$ and $\omega_{<1}$ parameters for the real data under model M2a. Histograms are over 100 bootstrap datasets.



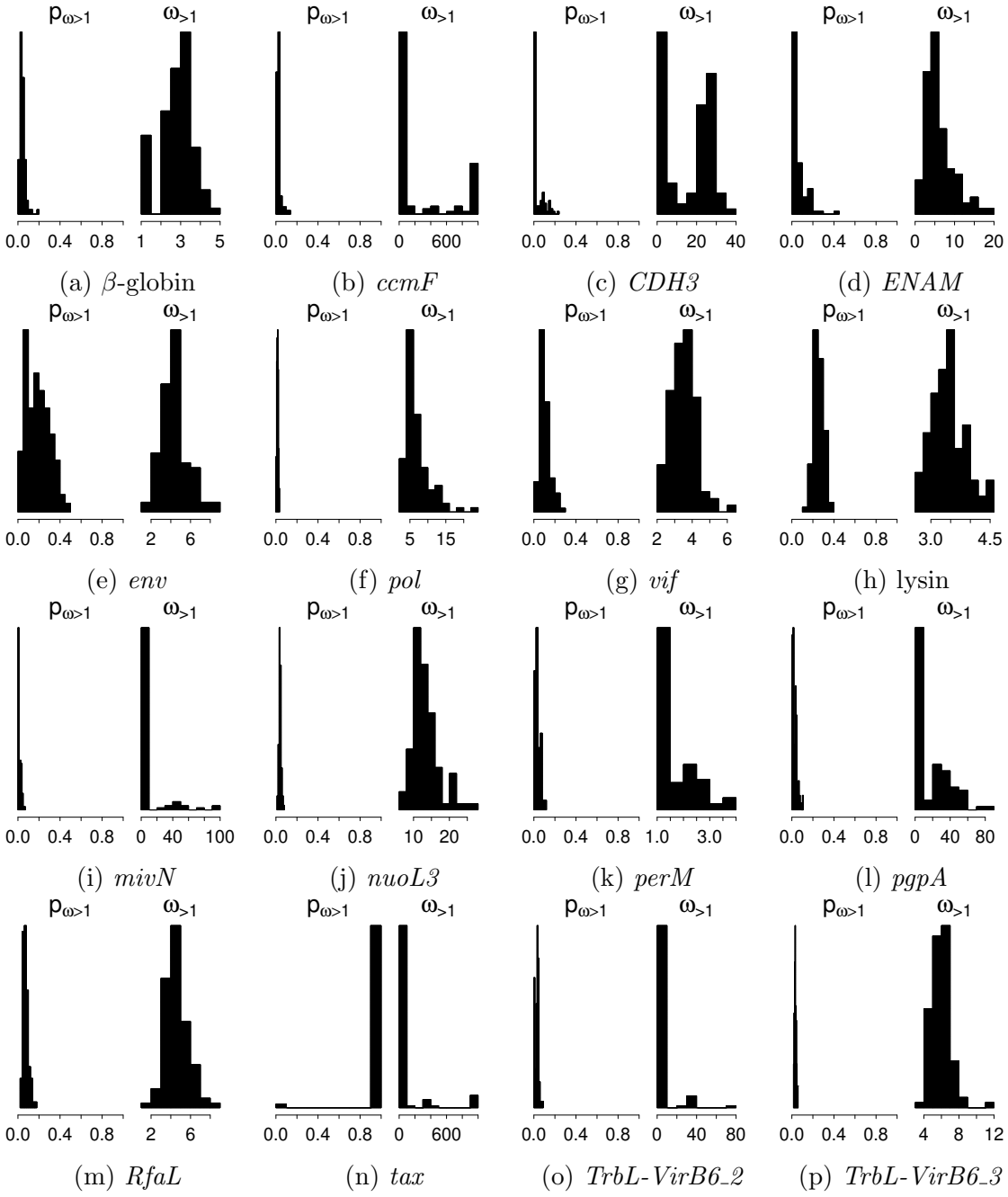
(a) β-globin (b) *ccmF* (c) *CDH3* (d) *ENAM*

(e) *env* (f) *pol* (g) *vif* (h) lysin

(i) *mivN* (j) *nuoL3* (k) *perM* (l) *pgpA*

(m) *RfaL* (n) *tax* (o) *TrbL-VirB6_2* (p) *TrbL-VirB6_3*

Fig. 6.19: Distributions of the $p_{\omega=1}$ parameters for the real data under model M2a. Histograms are over 100 bootstrap datasets.



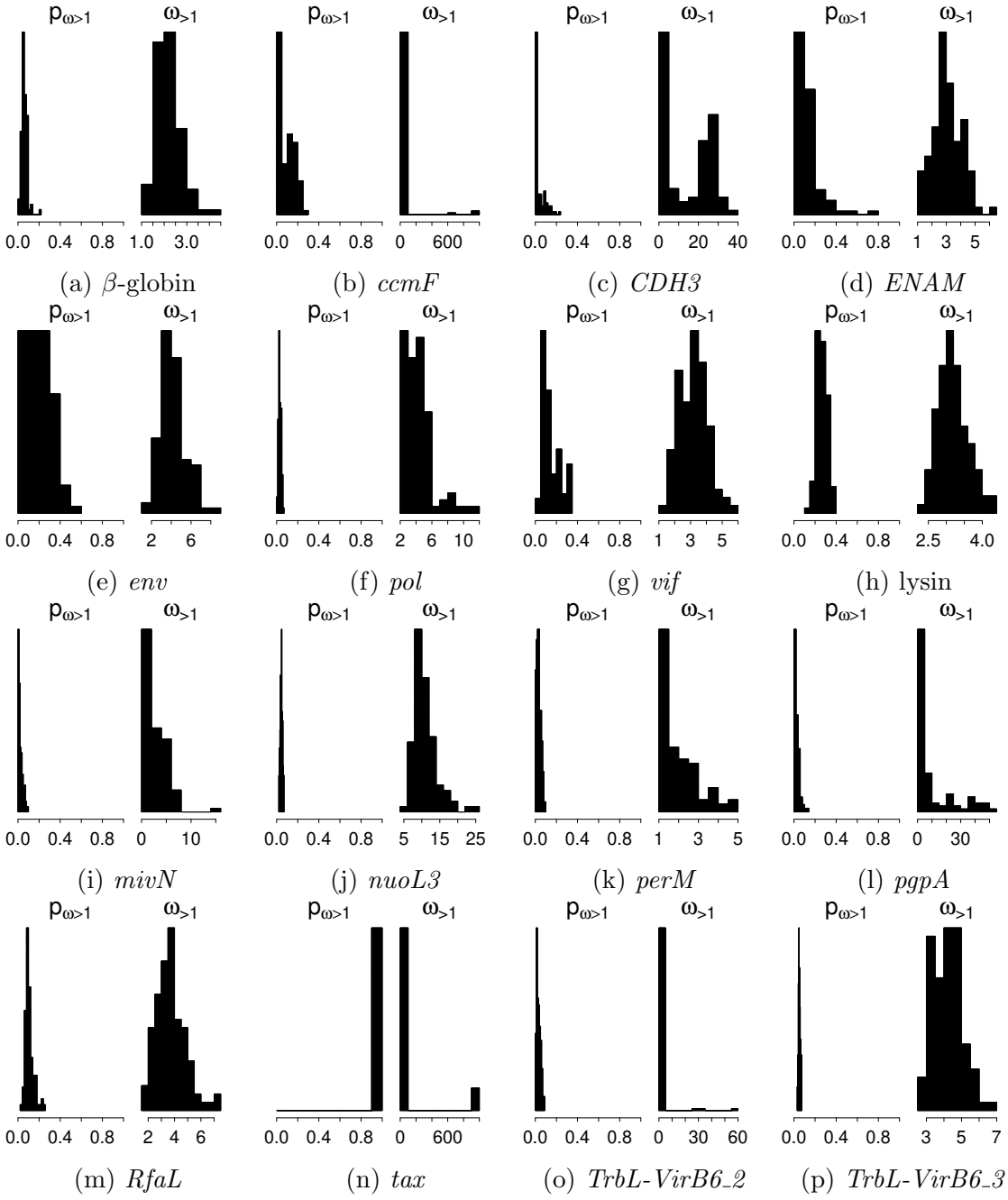
(a) β-globin (b) *ccmF* (c) *CDH3* (d) *ENAM*

(e) *env* (f) *pol* (g) *vif* (h) lysin

(i) *mivN* (j) *nuoL3* (k) *perM* (l) *pgpA*

(m) *RfaL* (n) *tax* (o) *TrbL-VirB6_2* (p) *TrbL-VirB6_3*

Fig. 6.20: Distributions of the $p_{\omega>1}$ and $\omega_{>1}$ parameters for the real data under model M2a. Histograms are over 100 bootstrap datasets.



(a) β-globin (b) *ccmF* (c) *CDH3* (d) *ENAM*

(e) *env* (f) *pol* (g) *vif* (h) lysin

(i) *mivN* (j) *nuoL3* (k) *perM* (l) *pgpA*

(m) *RfaL* (n) *tax* (o) *TrbL-VirB6_2* (p) *TrbL-VirB6_3*

Fig. 6.21: Distributions of the $p_{\omega>1}$ and $\omega_{>1}$ parameters for the real data under model M8. Histograms are over 100 bootstrap datasets.

6.6 SBA Branch-Site Model A - Analysis of *NR1D1*

SBA for branch-site codon model A (Zhang et al., 2005) was implemented to demonstrate the feasibility of SBA implementations for new models. The new implementation, which was completed within a few hours, can be found at https://github.com/Jehops/codeml_sba. The nuclear receptor gene, *NR1D1* (Baker et al., 2016), was analyzed under NEB, BEB, and SBA methods. The branch-site test of positive selection on the foreground branch leading to the human lineage was rejected at the 1% level (LRT test statistic: 10.26612, p-value: 0.00135). The MLEs of the ω-distribution parameters are shown in table 6.5. Because the estimated weights of the positive selection classes are very small, the estimates of $\omega > 1$ were unreasonable.

Under both NEB and BEB, the same site had a posterior probably of positive selection larger than 0.99, whereas the posteriors were well below 0.5 for all other sites. On the other hand, the mean posterior under SBA for the same site was 0.879. Plots of the maximum likelihood estimates (MLES) of the ω-distribution parameters are shown in figure 6.22.

Table 6.5: Estimates of the ω-distribution parameters for the *NR1D1* gene under branch-site model A.

Site Class	0	1	2a	2b
weight	0.95058	0.04751	0.00183	0.00009
background ω	0.03702	1.00000	0.03702	1.00000
foreground ω	0.03702	1.00000	999.00000	999.00000

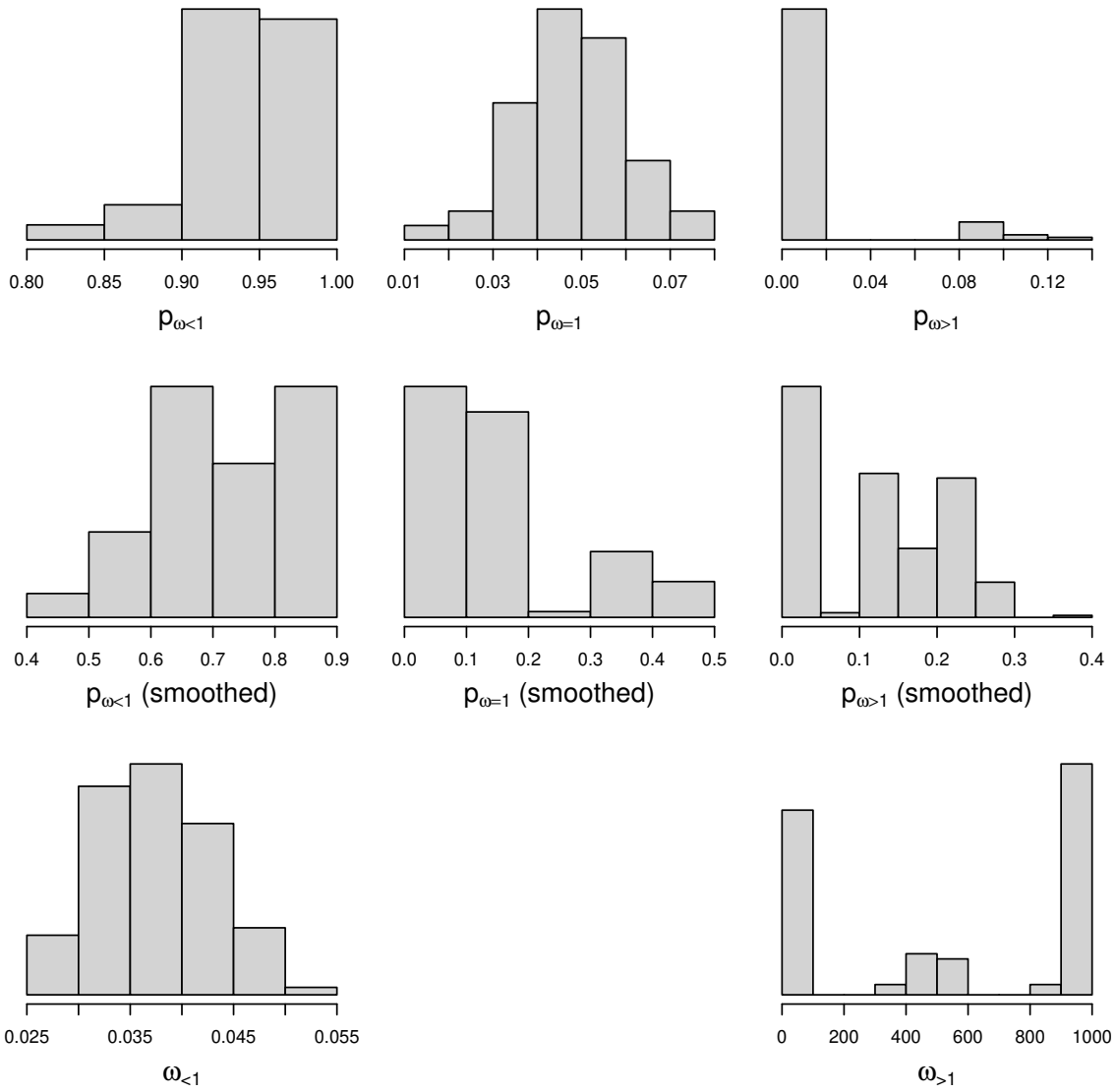


Fig. 6.22: Branch-site model A ω distribution parameter estimates over bootstrap samples. A bandwidth parameter of 0.4 was used to smooth the p estimates.

Bibliography

Akaike, H. (1954). An approximation to the density function. *Ann I Stat Math*, 6(2):127–132.

Allio, R., Nabholz, B., Wanke, S., Chomicki, G., Pérez-Escobar, O. A., Cotton, A. M., Clamens, A.-L., Kergoat, G., Sperling, F. A., and Condamine, F. L. (2020). Genome-wide macroevolutionary signatures of key innovations in butterflies colonizing new host plants. *bioRxiv*.

Allman, E. S., Ané, C., and Rhodes, J. A. (2008). Identifiability of a markovian model of molecular evolution with gamma-distributed rates. *Advances in Applied Probability*, 40(1):229–249.

Allman, E. S. and Rhodes, J. A. (2009). The identifiability of covarion models in phylogenetics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(1):76–88.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Anfinsen, C. B. (1972). The formation and stabilization of protein structure. *Biochemical Journal*, 128(4):737–749.

Anisimova, M., Bielawski, J., Dunn, K., and Yang, Z. (2007). Phylogenomic analysis of natural selection pressure in Streptococcus genomes. *BMC Evol Biol*, 7(1):154.

Anisimova, M., Bielawski, J. P., and Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular biology and evolution*, 18(8):1585–1592.

Anisimova, M., Bielawski, J. P., and Yang, Z. (2002). Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol*, 19(6):950–958.

Anisimova, M. and Kosiol, C. (2009). Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol*, 26(2):255–271.

Anisimova, M. and Yang, Z. (2007). Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Molecular biology and evolution*, 24(5):1219–1228.

Aris-Brosou, S. (2003). How Bayes tests of molecular phylogenies compare with frequentist approaches. *Bioinformatics*, 19(5):618–624.

Baele, G. and Lemey, P. (2013). Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency. *Bioinformatics*, 29(16):1970–1979.

Baker, J. L., Dunn, K. A., Mingrone, J., Wood, B. A., Karpinski, B. A., Sherwood, C. C., Wildman, D. E., Maynard, T. M., and Bielawski, J. P. (2016). Functional divergence of the nuclear receptor nr2c1 as a modulator of pluripotentiality during hominid evolution. *Genetics*, 203(2):905–922.

Bao, L., Gu, H., Dunn, K. A., and Bielawski, J. P. (2008). Likelihood-based clustering (LiBaC) for codon models, a method for grouping sites according to similarities in the underlying process of evolution. *Mol Biol Evol*, 25(9):1995–2007.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., et al. (2011). Has the earth's sixth mass extinction already arrived? *Nature*, 471(7336):51–57.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2012). Genbank. *Nucleic acids research*, 41(D1):D36–D42.

Berlin, S. and Smith, N. G. (2005). Testing for adaptive evolution of the female reproductive protein zpc in mammals, birds and fishes reveals problems with the m7-m8 likelihood ratio test. *BMC evolutionary biology*, 5(1):1.

Bickel, P. J. and Doksum, K. A. (2006). *Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, 2nd Edition*. CRC Press.

Bielawski, J. P. and Yang, Z. (2004). A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *Journal of molecular evolution*, 59(1):121–132.

Bielawski, J. P. and Yang, Z. (2005). Maximum likelihood methods for detecting adaptive protein evolution. In *Statistical methods in molecular evolution*, pages 103–124. Springer.

Breiman, L. (1996). Bagging predictors. *Mach Learn*, 24(2):123–140.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using diamond. *Nature methods*, 18(4):366–368.

Bush, R. M., Fitch, W. M., Bender, C. A., and Cox, N. J. (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol*, 16(11):1457–1465.

Chai, J. and Housworth, E. A. (2011). On rogers' proof of identifiability for the gtr+ γ+ i model. *Systematic biology*, 60(5):713–718.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):19–29.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):95–115.

Chen, J. (2017). On finite mixture models. *Statistical Theory and Related Fields*, 1(1):15–27.

Chernoff, H. and Lander, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of Statistical Planning and Inference*, 43(1-2):19–40.

Collins, T. M., Wimberger, P. H., and Naylor, G. J. (1994). Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Systematic Biology*, 43(4):482–496.

Crutzen, P. J. (2006). The "anthropocene". In *Earth system science in the anthropocene*, pages 13–18. Springer.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, volume 1. Cambridge university press.

Davydov, I. I., Salamin, N., and Robinson-Rechavi, M. (2019). Large-scale comparative analysis of codon models accounting for protein and nucleotide selection. *Molecular biology and evolution*, 36(6):1316–1332.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann Stat*, pages 1–26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet*, 25(5):471.

Felsenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA.

Fisher, R. A. (1923). Xxi.—on the dominance ratio. *Proceedings of the royal society of Edinburgh*, 42:321–341.

Fisher, R. A. (1931). Xvii.—the distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh*, 50:204–219.

Fitch, W. M., Bush, R. M., Bender, C. A., and Cox, N. J. (1997). Long term trends in the evolution of H(3) HA1 human influenza type A. *P Natl Acad Sci USA*, 94(15):7712–7718.

Fletcher, W. and Yang, Z. (2009). Indelible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*, 26(8):1879–1888.

Fletcher, W. and Yang, Z. (2010). The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection. *Molecular Biology and Evolution*, 27(10):2257–2267.

Forest, F., Crandall, K. A., Chase, M. W., and Faith, D. P. (2015). Phylogeny, extinction and conservation: embracing uncertainties in a time of urgency.

Forsberg, R. and Christiansen, F. B. (2003). A Codon-Based Model of Host-Specific Selection in Parasites, with an Application to the Influenza A Virus. *Molecular Biology and Evolution*, 20(8):1252–1259.

Fu, Y., Chen, J., and Kalbfleisch, J. D. (2009). Modified likelihood ratio test for homogeneity in a two-sample problem. *Statistica Sinica*, pages 1603–1619.

Gaston, D., Susko, E., and Roger, A. J. (2011). A phylogenetic mixture model for the identification of functionally divergent protein residues. *Bioinformatics*, 27(19):2655–2663.

Ge, G., Cowen, L., Feng, X., and Widmer, G. (2008). Protein coding gene nucleotide substitution pattern in the apicomplexan protozoa Cryptosporidium parvum and Cryptosporidium hominis. *Comp Funct Genom*, 2008.

Gharib, W. H. and Robinson-Rechavi, M. (2013). The Branch-Site Test of Positive Selection Is Surprisingly Robust but Lacks Power under Synonymous Substitution Saturation and Variation in GC. *Molecular Biology and Evolution*, 30(7):1675–1686.

Gil, M., Zanetti, M. S., Zoller, S., and Anisimova, M. (2013). Codonphyml: fast maximum likelihood phylogeny estimation under codon substitution models. *Molecular biology and evolution*, 30(6):1270–1280.

Gilbert, W. and Maxam, A. (1973). The nucleotide sequence of the lac operator. *Proceedings of the National Academy of Sciences*, 70(12):3581–3584.

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

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333.

Guindon, S., Rodrigo, A. G., Dyer, K. A., and Huelsenbeck, J. P. (2004). Modeling the site-specific variation of selection patterns along lineages. *P Natl Acad Sci USA*, 101(35):12957–12962.

Haldane, J. B. S. (1927). A mathematical theory of natural and artificial selection, part v: selection and mutation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 23, pages 838–844. Cambridge University Press.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. *Proc. Barkeley Conference in Honor of J Neyman and J Kiefer*, 2:807–810.

Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D., and Wray, G. A. (2007). Promoter regions of many neural-and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet*, 39(9):1140–1144.

Hood, L. and Rowen, L. (2013). The human genome project: big science transforms biology and medicine. *Genome medicine*, 5(9):1–8.

Huelsenbeck, J. P. and Dyer, K. A. (2004). Bayesian estimation of positively selected sites. *J Mol Evol*, 58(6):661–672.

Huxley, J. et al. (1942). Evolution. the modern synthesis. *Evolution. The Modern Synthesis.*

Jukes, T. H. and Cantor, C. R. (1969). *Evolution of protein molecules*. Mammalian Protein Metabolism. Academic Press, New York.

Kalbfleisch, J. (1985). *Probability and Statistical Inference: Volume 2: Statistical Inference*. Springer Texts in Statistics. Springer New York.

Kim, B. Y., Huber, C. D., and Lohmueller, K. E. (2017). Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*, 206(1):345–361.

Kimura, M. (1957). Some problems of stochastic processes in genetics. *The Annals of Mathematical Statistics*, pages 882–901.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129):624.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.

Kimura, M. (1986). Dna and the neutral theory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 312(1154):343–354.

Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D., Delport, W., and Scheffler, K. (2011). A random effects branch-site model for detecting episodic diversifying selection. *Molecular biology and evolution*, 28(11):3033–3043.

Kosiol, C., Vinař, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS Genet*, 4(8):e1000144.

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J Am Stat Assoc*, 82(399):739–750.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Le, S. Q., Dang, C. C., and Gascuel, O. (2012). Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol*, page mss112.

Lemey, P., Minin, V. N., Bielejec, F., Pond, S. L. K., and Suchard, M. A. (2012). A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*, 28(24):3248–3256.

Lewis, S. L. and Maslin, M. A. (2015). Defining the anthropocene. *Nature*, 519(7542):171–180.

Lu, A. and Guindon, S. (2013). Performance of Standard and Stochastic Branch-Site Models for Detecting Positive Selection among Coding Sequences. *Molecular Biology and Evolution*, 31(2):484–495.

Mack, C. A. (2011). Fifty years of moore's law. *IEEE Transactions on semiconductor manufacturing*, 24(2):202–207.

Martinez, D. T., de Magalhaes, J. P., and Opazo, J. C. (2020). Positive selection and fast turnover rate in tumor suppressor genes reveal how cetaceans resist cancer. *bioRxiv*.

Massingham, T. and Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, 169(3):1753–1762.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol*, 21(9):1781–1791.

Messier, W. and Stewart, C.-B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature*, 385(6612):151–154.

Mingrone, J., Susko, E., and Bielawski, J. (2016). Smoothed bootstrap aggregation for assessing selection pressure at amino acid sites. *Molecular Biology and Evolution*, 33(11):2976–2989.

Mingrone, J., Susko, E., and Bielawski, J. P. (2018). ModL: exploring and restoring regularity when testing for positive selection. *Bioinformatics*, 35(15):2545–2554.

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D. P., Smith, D. M., et al. (2015). Gene-wide identification of episodic selection. *Molecular biology and evolution*, 32(5):1365–1371.

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Pond, S. L. K. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS genet*, 8(7):e1002764.

Muse, S. V. and Gaut, B. S. (1994). 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.

Nachman, M. W. and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304.

Neyman, J. and Pearson, E. S. (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

Nielsen, R. (2002). Mapping mutations on phylogenies. *Syst Biol*, 51(5):729–739.

Nielsen, R. and Huelsenbeck, J. P. (2002). Detecting positively selected amino acid sites using posterior predictive p-values. In *Pacific Symposium on Biocomputing*, volume 7, pages 576–588.

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

Nozawa, M., Suzuki, Y., and Nei, M. (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proceedings of the National Academy of Sciences*, 106(16):6700–6705.

Pagel, M., Meade, A., and Crandall, K. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53(4):571–581.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann Math Stat*, pages 1065–1076.

Penn, O., Stern, A., Rubinstein, N. D., Dutheil, J., Bacharach, E., Galtier, N., and Pupko, T. (2008). Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. *PLoS Comput Biol*, 4(11):e1000214.

Petersen, L., Bollback, J. P., Dimmic, M., Hubisz, M., and Nielsen, R. (2007). Genes under positive selection in escherichia coli. *Genome research*, 17(9):1336–1343.

Pond, S. L. K. and Frost, S. D. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*, 22(5):1208–1222.

Rosenblatt, M. et al. (1956). Remarks on some nonparametric estimates of a density function. *Ann Math Stat*, 27(3):832–837.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467.

Scheben, A., Ramos, O. M., Kramer, M., Goodwin, S., Oppenheim, S. J., Becker, D. J., Schatz, M. C., Simmons, N. B., Siepel, A., and McCombie, W. R. (2020). Unraveling molecular mechanisms of immunity and cancer-resistance using the genomes of the neotropical bats artibeus jamaicensis and pteronotus mesoamericanus. *bioRxiv*.

Scheffler, K., Murrell, B., and Pond, S. L. K. (2014). On the validity of evolutionary models with site-specific parameters. *PloS One*, 9(4):e94534.

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., and Graur, D. (2009). Estimates of positive darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome biology and evolution*, 1:114–118.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc*, 82(398):605–610.

Silverman, B. and Young, G. (1987). The bootstrap: To smooth or not to smooth? *Biometrika*, 74(3):469–479.

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular biology and evolution*, 32(5):1342–1353.

Studer, R. A., Penel, S., Duret, L., and Robinson-Rechavi, M. (2008). Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome research*, 18(9):1393–1402.

Studer, R. A. and Robinson-Rechavi, M. (2009). Evidence for an episodic model of protein sequence evolution. *Biochemical Society Transactions*, 37(4):783–786.

Suzuki, Y. (2004). New methods for detecting positive selection at single amino acid sites. *J Mol Evol*, 59(1):11–19.

Suzuki, Y. (2008). False-positive results obtained from the branch-site test of positive selection. *Genes & genetic systems*, 83(4):331–338.

Suzuki, Y. and Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. *Mol Biol Evol*, 16(10):1315–1328.

Suzuki, Y. and Nei, M. (2004). False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom Thalassiosira weissflogii and the tax gene of a human T-cell lymphotropic virus. *Mol Biol Evol*, 21(5):914–921.

Swanson, W. J., Nielsen, R., and Yang, Q. (2003). Pervasive adaptive evolution in mammalian fertilization proteins. *Molecular biology and evolution*, 20(1):18–20.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular biology and evolution*, 10(3):512–526.

Tavaré, S. et al. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17(2):57–86.

Venables, W. N. and Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Venkat, A., Hahn, M. W., and Thornton, J. W. (2018). Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nature ecology & evolution*, 2(8):1280.

Wand, P. and Jones, C. (1994). *Kernel Smoothing*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Wang, H.-C., Li, K., Susko, E., and Roger, A. J. (2008). A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC evolutionary biology*, 8(1):331.

Watson, J. D. (1990). The human genome project: past, present, and future. *Science*, 248(4951):44–49.

Wong, W. S., 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(2):1041–1051.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2):97.

Wright, S. (1942). Statistical genetics and evolution. *Bulletin of the American Mathematical Society*, 48(4):223–246.

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

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular biology and evolution*, 15(5):568–573.

Yang, Z. (2005). The power of phylogenetic comparison in revealing protein function. *P Natl Acad Sci USA*, 102(9):3179–3180.

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

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8):1586–1591.

Yang, Z. (2014). *Molecular evolution: a statistical approach*. Oxford University Press.

Yang, Z. and Dos Reis, M. (2010). Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution*, 28(3):1217–1228.

Yang, Z. and Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of molecular evolution*, 46(4):409–418.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19(6):908–917.

Yang, Z., Nielsen, R., and Goldman, N. (2009). In defense of statistical methods for detecting positive selection. *Proceedings of the National Academy of Sciences*, 106(36):E95–E95.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. (2000a). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449.

Yang, Z., Swanson, W. J., and Vacquier, V. D. (2000b). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol Biol Evol*, 17(10):1446–1455.

Yang, Z., Wong, W. S., and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol*, 22(4):1107–1118.

Zhai, W., Nielsen, R., Goldman, N., and Yang, Z. (2012). Looking for darwin in genomic sequences—validity and success of statistical methods. *Molecular biology and evolution*, 29(10):2889–2893.

Zhang, J. (2004). Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular Biology and Evolution*, 21(7):1332–1339.

Zhang, J., Kumar, S., and Nei, M. (1997). Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Molecular Biology and Evolution*, 14(12):1335–1338.

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution*, 22(12):2472–2479.

Zhang, Q., Tombline, G., Ablaeva, J., Zhang, L., Zhou, X., Smith, Z., Xiaoli, A. M., Wang, Z., Lin, J.-R., Jabalameli, M. R., et al. (2020). The genome of north american beaver provides insights into the mechanisms of its longevity and cancer resistance. *bioRxiv*.