Improved models of biological sequence evolution

by

Benjamin Murrell

Dissertation approved for the degree of Doctor of Philosophy
in Computer Science at Stellenbosch University

Computer Science Division
Department of Mathematical Sciences
Stellenbosch University,
Private Bag X1, Matieland 7602, South Africa.

Promoter: Prof. Konrad Scheffler

August 2012

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: ....
B. S. Murrell

Date: 1/08/2012 ....

Abstract

Improved models of biological sequence evolution

B. S. Murrell

Computer Science Division
Department of Mathematical Sciences
Stellenbosch University,
Private Bag X1, Matieland 7602, South Africa.

Dissertation: PhD (Computer Science)

August 2012

Computational molecular evolution is a field that attempts to characterize how genetic sequences evolve over phylogenetic trees – the branching processes that describe the patterns of genetic inheritance in living organisms. It has a long history of developing progressively more sophisticated stochastic models of evolution. Through a probabilist's lens, this can be seen as a search for more appropriate ways to parameterize discrete state continuous time Markov chains to better encode biological reality, matching the historical processes that created empirical data sets, and creating useful tools that allow biologists to test specific hypotheses about the evolution of the organisms or the genes that interest them. This dissertation is an attempt to fill some of the gaps that persist in the literature, solving what we see as existing open problems. The overarching theme of this work is how to better model variation in the action of natural selection at multiple levels: across genes, between sites, and over time. Through four published journal articles and a fifth in preparation, we present amino acid and codon models that improve upon existing approaches, providing better descriptions of the process of natural selection and better tools to detect adaptive evolution.

Uittreksel

Verbeterde modelle van biologiese sekwensie-evolusie

("Improved models of biological sequence evolution")

B. S. Murrell

Afdeling Rekenaarwetenskap
Departement Wiskundige Wetenskappe
Universiteit Stellenbosch,
Privaatsak X1, Matieland 7602, Suid-Afrika.

Proefskrif: PhD (Rekenaarwetenskap)

Augustus 2012

Komputasionele molekulêre evolusie is 'n navorsingsarea wat poog om die evolusie van genetiese sekwensies oor filogenetiese bome – die vertakkende prosesse wat die patrone van genetiese oorerwing in lewende organismes beskryf – te karakteriseer. Dit het 'n lang geskiedenis waartydens al hoe meer gesofistikeerde waarskynlikheidsmodelle van evolusie ontwikkel is. Deur die lens van waarskynlikheidsleer kan hierdie proses gesien word as 'n soektog na meer gepasde metodes om diskrete-toestand kontinuë-tyd Markov kettings te parametriseer ten einde biologiese realiteit beter te enkodeer – op so 'n manier dat die historiese prosesse wat tot die vorming van biologiese sekwensies gelei het nageboots word, en dat nuttige metodes geskep word wat bioloë toelaat om spesifieke hipotesisse met betrekking tot die evolusie van belanghebbende organismes of gene te toets. Hierdie proefskrif is 'n poging om sommige van die gapings wat in die literatuur bestaan in te vul en bestaande oop probleme op te los. Die oorkoepelende tema is verbeterde modellering van variasie in die werking van natuurlike seleksie op verskeie vlakke: variasie van geen tot geen, variasie tussen posisies in gene en variasie oor tyd. Deur middel van vier gepubliseerde joernaalartikels en 'n vyfde artikel in voorbereiding, bied ons aminosuur- en kodon-modelle aan wat verbeter op bestaande benaderings – hierdie modelle verskaf beter beskrywings van die proses van natuurlike seleksie sowel as beter metodes om gevalle van aanpassing in evolusie te vind.

Acknowledgements

My sincerest gratitude to: Konrad Scheffler for his careful supervision, paying attention to detail while always maintaining a clear view of the big picture; Sergei Kosakovsky Pond, for his brilliant unofficial supervision, expert collaboration, and amazingly swift responses to all of my questions; Chris Seebregts, and the Medical Research Council, for being the best employer imaginable; Tulio de Oliveira for his excellent co-supervision on the MSc that was upgraded to this PhD. Without them I would most likely be squandering my time in an uninteresting field writing obscure papers that nobody reads.

I would also like to express my gratitude to all the students that have worked on the projects comprising this dissertation, either during the computational biology research workshops, or as research assistants. It was truly a pleasure to work with you.

My everlasting thanks to my parents for their encouragement and confidence throughout my life, and to Sasha, for her love, support and help spotting badly worded sentences.

Funding: I gratefully acknowledge financial support from Europeaid Grant number SANTE/2007/147-790 from the European Commission.

Foreword

This is a dissertation by publication. It begins with a general introduction to the field, followed by four chapters which very briefly summarize five papers, concluding with a synopsis chapter. The five papers themselves are appended, for the convenience of the reader, to the end of the dissertation. Four of the papers are included in their final published form, but the one paper is unpublished as of submission time. Should this paper be accepted, the published version may differ from the one included in this dissertation. The bibliographies for each paper are self contained, and only the references cited during the introduction and synopsis of the dissertation itself are included in the dissertation bibliography.

Like most endeavors in science, these papers are collaborations with multiple authors. I have attempted to clarify my contribution to each paper after the brief summary section outlining the papers.

Contents

Declaration **i**

Abstract **ii**

Uittreksel **iii**

Acknowledgements **iv**

Foreword **v**

Contents **vi**

List of Figures **viii**

List of Tables **ix**

Abbreviations **x**

1 Introduction **1**

1.1 Dissertation outline 1
1.2 Preliminaries 2
1.3 Briefest biology 2
1.4 Phylogenetic models of evolution 3
1.5 Specific models 7
1.6 Literature overview: accounting for variation 9
1.7 Implementation 12

2 Random effects models allowing rate variation over branches **14**

2.1 Summary 14
2.2 Contribution statement 15

3 Modeling HIV-1 Drug Resistance as Episodic Directional Selection **16**

3.1 Summary 16
3.2 Contribution statement 17

4 FUBAR : An efficient analysis of the molecular footprint of natural selection 18

4.1 Summary 18

4.2 Contribution statement 19

5 Non-Negative Matrix Factorization for Learning Alignment-Specific Models of Protein Evolution 20

5.1 Summary 20

5.2 Contribution statement 21

6 Synopsis 22

6.1 Introduction 22

6.2 Methodological contributions 23

6.3 List of problems addressed 25

6.4 Future work 27

Appendices 30

A Approaches to inference 31

A.1 Inference and model selection 31

A.2 Empirical Bayes 33

List of References 34

List of Figures

1.1 *Phylogeny and alignment.* A coding sequence is depicted, with taxon names $tax1 \ldots tax5$, and the phylogeny describing the ancestry of the sequences. For example, $tax1$ and $tax2$ are the most closely related, since the total branch length separating them is the shortest. 4

List of Tables

6.1 A summary of the models introduced in this dissertation. 23

Abbreviations

Biology

AA - Amino acid
ARS - Antigen-recognition sites
DNA - Deoxyribonucleic acid
DRAM - Drug resistance associated mutation
HIV - Human immunodeficiency virus
MHC - Major histocompatibility complex
RNA - Ribonucleic acid

Models

BS-REL - Branch-site random effects likelihood
DEPS - Directional evolution in protein sequences
EDEPS - Episodic directional evolution in protein sequences
FEEDS - Fixed effects episodic directional selection
FEL - Fixed effects likelihood
FUBAR - Fast unconstrained Bayesian approximation
GTR - Generalized time reversible
NNMF - Non-negative matrix factorization
MEDS - Model of episodic directional selection
MEME - Mixed effects model of evolution
REL - Random effects likelihood

Other

HyPhy - Hypothesis testing using phylogenies (software package)
LRT - Likelihood ratio test
MCMC - Markov chain Monte Carlo
PAML - Phylogenetic analysis by maximum likelihood (software package)

Chapter 1

Introduction

1.1 Dissertation outline

The central theme of this dissertation is incorporating variation into models of sequence evolution. This mirrors the history of progress in the field, where the restrictive assumptions of earlier models are incrementally relaxed to better account for the staggering variability of biological reality. The selective forces influencing the evolution of living organisms are never constant. Different genes have different functions, so selective pressures will vary from one gene to another. Indeed, different amino acid sites within each gene have specific roles, facilitating different interactions, and thus a concomitant heterogeneity of selective forces governs which specific amino acids at each site yield better adapted organisms with greater replicative fitness. Finally, the environments of genes – including other genes which influence their contribution to the fitness of the organism, as well as the environment exogenous to the organism – are frequently in flux, and selective pressure is seldom constant over time. After chapter 1, which briefly introduces the material upon which the rest of the dissertation builds, the remaining chapters all present techniques to better model variation in natural selection.

Chapter 2 describes a new way of modeling heterotachy, where an evolutionary rate is allowed to vary across the phylogeny. The process on every branch is modeled as a random effect, using a mixture of Markov substitution processes, allowing the rate at each site to vary from one branch to another. This technique is applied to two problems: detecting lineages where some sites are under episodic selection, and detecting sites where some lineages are under selection. Both demonstrate substantial improvements over previous models.

Chapter 3 continues the theme of selective pressure varying over time, but deals with the case where a known rapid exogenous event triggers a sudden shift in the fitness landscape. This model is applied to HIV-1(Human Immunodeficiency Virus Type 1) drug resistance, where current models fail to appropriately capture the dynamics of the scenario.

Chapter 4 explores variation from site to site – a niche with a rich history in this field – and demonstrates that substantial improvement is still possible. By employing innocuous computational shortcuts, we can develop an approximate Bayesian approach which captures far richer site to site heterogeneity in selective pressure, enabling the models to detect sites with greater accuracy and at a fraction of the computational cost, allowing the analysis of larger datasets.

Finally, models of protein evolution require a large number of parameters: typically too large to estimate for specific genes. Chapter 5 proposes and tests a method for allowing gene to gene variation in models of protein evolution. Non-negative matrix factorization – a dimensionality reduction technique – is employed to efficiently parameterize amino acid rate matrices, where the final model is a mixture of "basis" rate matrices discovered through non-negative matrix factorization.

1.2 Preliminaries

Understanding models of biological sequence evolution requires an understanding of basic probability theory, statistical inference, and elementary biology. This chapter attempts to introduce just enough of each to make the subsequent chapters comprehensible, but is invariably too short to serve as a comprehensive introduction. If a more detailed introduction to biology is required, please see Chapter 1 of Hunter (1992) (this chapter is available online). If probability theory is required, see the first chapter of Durbin *et al.* (1998) for an appropriately brief introduction. We will also introduce the components of a phylogenetic model of sequence evolution: 1) the alignment, 2) the phylogeny, and 3) the transition probability matrices, and describe how substitution processes are parameterized. If more depth is necessary, we refer the reader to the excellent and comprehensive Yang (2006) and Salemi and Vandamme (2003). We will proceed to describe the necessary parts of the history of the field of modeling molecular evolution, followed by more recent developments that form the foundation for the novel work presented in later chapters.

1.3 Briefest biology

This dissertation is about modeling the evolution of genes that code for proteins. DNA (deoxyribonucleic acid) is composed of nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). In non-viral organisms, stretches of nucleotides (with specific properties) will be transcribed to messenger RNA (ribonucleic acid), which is in turn translated into chains of amino acids (AAs), which fold into proteins. Although there are some exceptions, the functional importance of such protein coding genes is overwhelmingly determined by the particular string of amino acids that determines the structure (through complex folding dynamics) and physicochemical properties of the resulting protein. In some viruses, genetic information is encoded directly as RNA, which has Uracil (U) in place of Thymine.

If we want to model how gene sequences change over evolutionary history, then we have to capture the mutational forces affecting the nucleotide sequences, and the selective forces acting on the resulting proteins. We thus need to understand the genetic code, which determines the amino acid sequence from the nucleotide one. Nucleotides are arranged in a triplet code, where each successive nucleotide triplet is called a "codon". Each of these codons is translated into one of 20 amino acids. Given a triplet of 4 possible characters, there are 64 codons. In the standard "universal" genetic code (as always, there are exceptions), 61 of these triplets code for amino acids (the "sense codons"), and 3 encode the "stop codons" which terminate translation, and so do not themselves occur in coding sequences. There are 61 possible sense codons and only 20 amino acids. The genetic code is thus a many-to-one mapping from codons to amino acids.

Mutations are changes to the nucleotide sequence. Because of the many-to-one genetic code, some of these mutations modify the resulting amino acid and thus modify the protein (called "non-synonymous" changes), but some leave the amino acid (and thus the protein) unaltered ("synonymous changes"). An assumption made throughout this dissertation is that natural selection acts predominantly on the protein sequence, and thus synonymous changes have little to no impact on the replicative capacity or fitness of the organism (although there are counter-examples: see Cuevas *et al.*, 2011, for experimental evidence from viruses). We can thus build models that treat nucleotide substitutions differently depending on whether they are synonymous or non-synonymous.

Evolution is change in gene frequency over time. Mutations occur by chance, and can either die out, go to fixation (where the entire population comes to possess that mutation), or be maintained as stable polymorphisms (where a stable proportion of individuals posses the mutation). The factors influencing the fate of any particular mutation are vast. Even if the mutation has no fitness consequences, stochastic fluctuations may still drift it to fixation, or, if the mutation has fitness benefit, chance may see it eradicated. The process we seek to model is thus far from deterministic, and this is why we turn to stochastic models of evolution.

1.4 Phylogenetic models of evolution

Gene sequencing technology provides an abundance – rapidly growing (Rothberg and Leamon, 2008) – of protein coding sequences from related organisms. Phylogenetic models of evolution allow us to infer the structure of the ancestral relationships between these organisms and details of the evolutionary

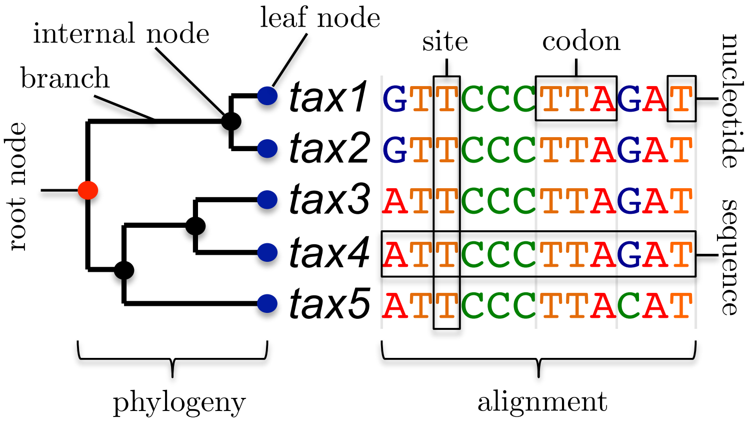

Figure 1.1: *Phylogeny and alignment.* A coding sequence is depicted, with taxon names $tax1 \ldots tax5$, and the phylogeny describing the ancestry of the sequences. For example, $tax1$ and $tax2$ are the most closely related, since the total branch length separating them is the shortest.

processes that led to these organisms.

1.4.1 The alignment

For our purposes, an alignment is a set of characters arranged by taxon and site, including gap characters which handle insertions and deletions. This is the data that the model needs to explain. As illustrated in figure 1.1, each row is a sequence that has been obtained from an organism, and a multiple alignment algorithm (e.g. Löytynoja and Goldman, 2008) has determined where to place gap characters to best align the sequences.

Multiple sequence alignment algorithms typically use heuristic search techniques to find gap placements. One popular multiple alignment approach, called "progressive alignment", identifies the two most closely related sequences, and performs a pairwise alignment. Then the next closest sequence is aligned to this alignment, and so on. More sophisticated algorithms use alignment scoring systems, and perform a more thorough search of the space of possible alignments. Recent developments (e.g. Löytynoja and Goldman, 2008) take the phylogenetic relationships between sequences into account when deciding on gap placement. Sequence alignment is not dealt with in this dissertation: the alignment is treated as a fixed entity throughout, and gap characters are treated as missing data.

Another simplifying assumption that will be made throughout this dissertation (and is indeed made throughout most of the literature) is that each alignment column evolves independently of the rest – an unrealistic assumption required for computational tractability. An alignment column (often called a "site") represents the genetic information for a single position in every sequence, but differs depending on whether the model is capturing nucleotide, amino acid, or codon evolution. If, for example, one is neglecting the effects of the genetic code and using a simple nucleotide model, a site will be one nucleotide per taxon, but will instead be one codon (a nucleotide triplet) if codon data is being modeled.

1.4.2 The phylogeny

For our purposes, the phylogeny is simply the specification of the branching structure of nodes, intended to represent the ancestral relationships between the observed taxa. Besides the root node – which is an orphan – every node must have a single parent (see figure 1.1). This ensures that the phylogeny has a tree structure. We refer to the nodes that lack children – corresponding to the sequenced taxa – as leaf nodes, and the internal nodes represent the most recent common ancestors of their child nodes. We will aim to model evolution over this phylogeny. Since each site is assumed to have evolved independently, we may refer to the "state" of the process at a single site: the genetic character (nucleotide, codon or amino acid) present at that node for that site; observed for the leaf nodes but uncertain for the internal nodes.

Each branch of a phylogeny represents a homogeneous population. One of the assumptions made by phylogenetic models of evolution is that the time it takes for a mutation to either die out or go to fixation is negligible compared to the rate at which mutations occur. A mutation occurs in an individual, and then instantly tends to fixation or extinction across the whole population that is represented by that branch. This assumption is innocuous when leaf nodes represent distinct species, but may be violated when the data is sampled from a more homogeneous intermixing population, as is the case with within-host viral data. In such cases, the phylogeny may be thought of as a partial genealogy where branches represent individuals rather than populations, although the behavior of phylogenetic models in these contexts have been questioned (Kryazhimskiy and Plotkin, 2008).

Techniques for estimating phylogenies from sequence data abound. "Distance methods" first compute a matrix of pairwise evolutionary distances between all sequence pairs, and then use heuristic algorithms to attempt to find phylogenies that capture the structure of the distance matrix. Maximum likelihood methods search through the space of possible phylogenies, attempting to find the branching structure that best explains the data, where "explains" is precisely defined in the next section. Bayesian methods similarly rely on explicit probabilistic models of sequence evolution, but, rather than finding a single tree, they find a set of credible phylogenies, quantifying the uncertainty in the tree structure. While all the methods presented in this dissertation require the existence of a phylogeny, this is not the focus, and standard phylogeny estimation tools are used throughout.

1.4.3 The likelihood function

If we denote the i^{th} site as D_i, then a phylogenetic model of evolution with parameters θ will assign $P(D_i|\theta)$, the probability of the site given the model. Because sites are assumed to be independent, we can compute the probability of the whole alignment, $P(D|\theta) = \prod_i P(D_i|\theta)$. This probability is called the likelihood function, since it assigns a likelihood to the data as a function of the model parameters. Phylogenetic models of evolution calculate this likelihood by parameterizing a Markov process over a tree, where the state at any node is conditionally dependent on its parent state, and conditionally independent of the state at any other node.

1.4.3.1 Transition probability matrices

A transition probability matrix captures the conditional dependence from a parent to a child node. Let b denote a node in the phylogeny, with $pa(b)$ the parent of that node and A_b the state at node b. If the set of allowable states is S, then an element of the transition probability matrix T is $T_{ij} = P(A_b = S_j|A_{pa(b)} = S_i)$, the probability of a transition from state i to state j over that branch. Depending on exactly what we are attempting to model, this transition probability matrix could be unique to branches or sites or both, and would then require appropriate indices.

1.4.3.2 Calculating the likelihood over a phylogeny

Let $A = \{A_0 \ldots A_B\}$ denote the vector of states for each node, indexing from 0 (the root) to the total number of branches, B. Further, partition this vector into terminal states, D, and internal states, A^*. Then, if the ancestral states for each node were known, the joint probability of the terminal and ancestral states could be calculated as

$$P(D, A^*|\theta) = P(A|\theta) = P(A_0|\theta) \prod_{b=1}^{B} P(A_b|A_{pa(b)}, \theta) \qquad (1.4.1)$$

where θ determines the full specification of the transition probability matrix along each branch. Since the internal ancestral states A^* are not typically known, we can compute the likelihood, $P(D|\theta)$, by marginalizing over the unknown ancestral states:

$$P(D|\theta) = \sum_{A^*} P(D, A^*|\theta) \qquad (1.4.2)$$

where the sum in the expression is taken over all possible arrangements of the unknown ancestral states. Naively, the complexity of the sum grows exponentially with the number of branches, but fortunately Felsenstein's algorithm

(Felsenstein, 1981) can compute this quantity efficiently using dynamic programming, the resulting computation being linear in the number of branches.

Thus, given an alignment, a phylogeny, and a transition probability matrix for each branch, we can compute the likelihood of the data.

1.4.4 A substitution process along a single branch

Transition probability matrices are obtained by specifying a substitution process for each branch (but see Barry and Hartigan, 1987, for a different approach). We will be dealing with transition probability matrices that are parameterized as discrete state, continuous time Markov processes. An underlying rate matrix $Q = \{q_{ij}\}$ is specified, where all off-diagonal entries are non-negative ($q_{ij,i \neq j} \geq 0$) and the rows sum to 0 ($q_{ii} = -\sum_{\forall k \neq i} q_{ik}$). The entries in Q are the instantaneous substitution rates. To obtain a transition probability matrix after time t, we use the matrix exponential:

$$T(t) = e^{Qt} \tag{1.4.3}$$

Once again, the rate matrix Q can differ by branch or by site or both, and may need to be indexed as such. We will refer to t as the branch length parameters – they are typically shared across sites but differ from branch to branch.

1.5 Specific models

1.5.1 Nucleotide models

When the alignment is considered at the nucleotide level, substitutions between the 4 character states are modeled with a 4×4 nucleotide rate matrix. There is a large literature – motivated by computational considerations – surrounding models with tractable analytic forms of the matrix exponential required to compute the transition probability matrix from the rate matrix (see Posada and Crandall, 1998, for a popular approach comparing such models). With current computational power it is inexpensive to exponentiate rate matrices numerically, so more complex and flexible models may be adopted. For the purposes of this dissertation, knowledge of the Generalized Time Reversible (GTR) model will suffice, which has 6 rate parameters $n_{ij} = n_{ji}$ $(i \neq j)$, and 3 equilibrium frequency parameters π_j (there are only 3 because of the usual stochastic constraint: $\sum_j \pi_j = 1$). The rate matrix is the product of a symmetric matrix of rate parameters (with the diagonal elements constrained to ensure row sums of 0, which leaves 6 parameters) with equilibrium frequency parameters multiplied column-wise:

$$q_{ij} = \begin{cases} n_{ij}\pi_j, & i > j \\ n_{ji}\pi_j, & i < j \\ -\sum_{k \neq i} q_{ik}, & i = j. \end{cases} \tag{1.5.1}$$

Since this is a Markov process, as long as no changes are impossible the expected frequencies of each state will tend to a constant limit as time tends to infinity, and the expected frequencies at this limit are controlled by the equilibrium frequency parameters. This model is called "time reversible" because, at equilibrium, the process is indistinguishable if run forwards or backwards, which is true of any symmetric matrix of rates with frequencies multiplied column-wise (Yang, 2006). This dissertation will not deal with nucleotide models in depth. They are used to provide computationally expedient estimates of branch proportions, and as the backbone upon which codon models are built.

1.5.2 Amino acid models

When the alignment data consists of strings of amino acids, even with the reversibility assumption imposing symmetry on the rate matrix, 209 parameters must be specified. The matrix has exactly the same definition as equation 1.5.1, but i and j go from 1 to 20 instead of from 1 to 4. 190 values are required for the symmetric off-diagonal elements and 19 for the equilibrium frequencies. This is over-parameterized for most datasets, and most applications involving amino acid models resort to using fixed-value models trained on databases of large protein family alignments. This problem is discussed and addressed in chapter 5.

1.5.3 Codon models

Codon models (introduced simultaneously by Goldman and Yang, 1994; Muse and Gaut, 1994) exploit the structure of the genetic code to efficiently parameterize substitution models at the codon level. Under the universal genetic code, there are 61 sense codons, which, even with reversibility constraints, would make for a large number of parameters (1830 off-diagonal elements, with 60 frequency parameters). Instead, codon models distinguish only two types of nucleotide substitutions: those that change the amino acid, and those that do not (but see Delport *et al.*, 2010*b*). The process along a branch is defined by its instantaneous rate matrix, $Q = \{q_{ij}\}$, with elements that describe the rate of substitution of codon i with codon j:

$$q_{ij}(\alpha,\beta,\Pi,\mathcal{N}) = \begin{cases} \alpha\pi_{ij}n_{ij}, & \delta(i,j)=1,\ AA(i)=AA(j), \\ \beta\pi_{ij}n_{ij}, & \delta(i,j)=1,\ AA(i)\neq AA(j), \\ 0, & \delta(i,j)>1, \\ -\sum_{k\neq i} q_{ik}, & i=j. \end{cases} \tag{1.5.2}$$

$\delta(i,j)$ counts the number of nucleotide differences between codons i and j, and instantaneous changes requiring more than one nucleotide substitution are disallowed (although they are possible in any finite amount of time). α and β parameterize the rates of synonymous and non-synonymous substitutions respectively. n_{ij} (comprising $\mathcal{N}$) are the nucleotide mutational biases, parameterized as in the nucleotide model in equation 1.5.1. π_{ij} (comprising Π) denote the equilibrium frequency parameters. There is a large literature surrounding the equilibrium frequency parameters, with the Muse and Gaut (1994) approach differing from the Goldman and Yang (1994) approach, and with some more sophisticated new developments (Kosakovsky Pond *et al.*, 2010; Yap *et al.*, 2010). In this dissertation, the approach of Kosakovsky Pond *et al.* (2010) is adopted. Traditionally, the frequencies of the 61 sense codons are estimated using the product of the position specific nucleotide frequencies (the so-called "F3×4" estimator), invoking an independence assumption to reduce the number of required parameters. Kosakovsky Pond *et al.* (2010) demonstrate that this approach neglects the nucleotide composition of stop codons, causing the estimate to be biased, and they propose a corrected estimator that accounts for stop codon nucleotide composition. Readers are referred to Kosakovsky Pond *et al.* (2010) for details.

1.6 Literature overview: accounting for variation

With the future chapters in mind (particularly chapters 2, 3 and 4), we will attempt to give an overview of the existing literature and the relevant techniques used to incorporate variation. The discussion will mostly be in the context of codon models, since that is what we will be dealing with in later chapters, although similar developments have occurred for nucleotide and amino acid models. See Delport *et al.* (2009) and Anisimova and Kosiol (2009) for comprehensive reviews.

The earliest codon models allowed no variation in the parameters governing the relative synonymous and non-synonymous rates (Goldman and Yang, 1994; Muse and Gaut, 1994). These were fixed across the sites in the alignment and across the branches in the phylogeny. One of the goals of these codon models was to detect instances where nucleotide changes that modified the protein were more likely to go to fixation than changes that did not – where the non-synonymous rate was larger than the synonymous rate. This suggests that natural selection was acting to change the protein. The problem for these models is that natural selection seldom favors changes to all amino acids in a protein, and seldom on all branches in a phylogeny – over the entire evolutionary history of that protein.

1.6.1 Variation over branches

The simplest way of incorporating variation over branches is to allow each branch to have its own set of selection parameters. This was proposed by Yang (1998). One can then test whether $\beta > \alpha$ on any particular branch by introducing a constrained null model with $\beta \leq \alpha$ for each branch in turn, and comparing these null models to the unconstrained alternative model. Inference with likelihood ratio tests (LRT – see A.1.2) can assess the evidence for rejecting the null model.

A second method for incorporating variation over branches was introduced in the form of a "covarion" model Tuffley and Steel (1998). A number of substitution models can be combined, such that the overall substitution model allows switching between any of the component models at any point along a branch. The switching process itself is also modeled as a continuous time Markov chain, and switching rate parameters need to be estimated from the data. While some attempts have been made to use covarion models to capture variation in selection parameters (Guindon *et al.*, 2004), this approach has not been widely adopted, perhaps because the tests for selection based on this model do not outperform the tests that assume constant rates over branches. In chapter 2 we describe an approach that addresses this problem, and which has substantially greater power to detect sites under episodic selection.

1.6.2 Variation over sites

There is generally no reason to expect the selective pressures that guide sequence evolution to be identical from one amino acid site to another. Allowing the selective pressure to vary from site to site has thus been one of the most important developments in the history of codon models. There are two strategies: fixed and random effects.

1.6.2.1 Fixed effects models

Fixed effects models partition the alignment into regions that we would expect *a priori* to share the same selection parameters. For example, Yang and Swanson (2002) model the evolution of the major histocompatibility complex (MHC) by partitioning it into antigen recognition sites (ARS – which bind to foreign antigens) and non-ARS regions. The ARS regions all share one ω (which is equal to β/α, using a single parameter governing the non-synonymous to synonymous rate ratio), and the non-ARS regions share another. This model is able to show that the ω value for the ARS regions was significantly greater than 1, using an LRT.

With many genes, such *a priori* partitioning is not available. In such circumstances, a fixed effects approach can allow each site to possess a unique set of selection parameters. Such approaches have been proposed by Kosakovsky Pond and Frost (2005*b*) and Massingham and Goldman (2005) and have shown some success at detecting individual sites subject to positive selection.

1.6.2.2 Random effects models

Random effects models allowing rate variation over sites were first proposed in the context of nucleotide models (Yang, 1993). Assume, for illustration, that a single parameter, ω, varies across sites. If we allow the value of ω at each site to be one of a set of K discrete rate categories, indexed $\omega_1, \ldots, \omega_K$, then, introducing probabilities for each rate category, $P(\omega_i)$, we can calculate the marginal likelihood for a single site as

$$P(D) = \sum_{i=1}^{K} P(D|\omega_i)P(\omega_i). \tag{1.6.1}$$

The marginal likelihood is thus calculated as the sum of the likelihoods for each rate category, weighted by the probability of each category. The parameters governing the distribution of ω are shared across sites, but the value of ω at each site varies as a random draw from this distribution. This allows ω to vary over sites, while incurring a small number of parameters relative to the fixed effects likelihood approach, which requires a different parameter for each site.

Inference over entire alignments using such models allows one to infer where there was positive selection on a small proportion of sites, where constant rate models would suggest that selection was, on average, purifying. As an example, the earliest such model allowed 3 ω categories (Nielsen and Yang, 1998; Yang *et al.*, 2000): $\omega_0 \leq 1$, $\omega_1 = 1$ and $\omega_2 \geq 1$ for the alternative model. The null model only possessed the first two categories. This allows an LRT to assess the evidence for a proportion of sites evolving under positive selection.

An empirical Bayes (see A.1.3) procedure is used to estimate the posterior probabilities for the ω categories for each site. Using the maximum likelihood parameter estimates for our prior distribution, we use the conditional likelihoods for each category to compute the posteriors:

$$P(\omega|D) = \frac{P(D|\omega)P(\omega)}{\sum_{\forall\omega} P(D|\omega)P(\omega)} \tag{1.6.2}$$

We can use these posterior probabilities directly, or compute Bayes factors (Kosakovsky Pond and Muse, 2005) to assess the strength of evidence for positive selection at each site.

Much of the literature surrounding random effects models can be characterized as finding more parsimonious ways to parameterize distributions over selection parameters, as well as increasing the list of parameters that should vary from site to site. Kosakovsky Pond and Muse (2005), for example, show that allowing the synonymous rate to vary from site to site results in better fitting models, and that not doing so can lead to false positives. In chapter 4, we demonstrate a computationally inexpensive approach that flexibly accounts for rich site to site variation in both synonymous and non-synonymous rates without assuming any parametric form for their joint distribution.

1.6.2.3 Branch-site models

An additional class of models are the so-called "branch-site" models, introduced in Yang and Nielsen (2002), and refined in Zhang *et al.* (2005) and Yang and Reis (2011). These models allow a subset of branches on the phylogeny to be designated as foreground, and the rest as background. A random effects model is created with this partition, allowing each site to belong to one of a number of categories, where each category allows the foreground and background branches to be treated differently. For example, in Yang and Nielsen (2002) there are 3 ω values: $\omega_0 \leq 1$, $\omega_1 = 1$ and $\omega_2 \geq 1$. One category has both foreground and background branches with $\omega_0 \leq 1$, another has both with $\omega_1 = 1$, and yet another has foreground branches with $\omega_2 \geq 1$, but background branches with $\omega_0 \leq 1$, while the final category has foreground branches with $\omega_2 \geq 1$, but background branches with $\omega_1 = 1$. These models can use LRTs to detect branches (or sets of branches) where selection is positive at only a small proportion of sites, and use empirical Bayes to infer which sites were likely under diversifying selection. We show, however, in chapter 2, that this model is sensitive to departures from an overly restrictive null model, which can lead to both false positive and false negative rates being uncontrolled. We then propose an approach that addresses this problem by relaxing the restrictive constraints on the background branches.

1.7 Implementation

There are multiple software packages implementing phylogenetic models of evolution. PAML (Phylogenetic Analysis by Maximum Likelihood) (Yang, 1997) contains the models implemented by Ziheng Yang's group, which are predominantly maximum likelihood random effects models. MrBayes (Huelsenbeck and Ronquist, 2001) allows for the implementation of a range of phylogenetic models in a fully Bayesian framework. HyPhy (Hypothesis Testing using Phylogenies)(Kosakovsky Pond *et al.*, 2005) has a useful graphical user interface, and possesses a rich scripting language, allowing the implementation of custom models of molecular evolution. All of the models in this dissertation are implemented in HyPhy, and some have been included in the HyPhy group's webserver, Datamonkey (Kosakovsky Pond and Frost, 2005*a*).

Chapter 2

Random effects models allowing rate variation over branches

2.1 Summary

The two papers comprising this chapter both exploit a novel technique for allowing the rate class on each branch to be a random draw from a discrete distribution. While random effects models over sites calculate marginal likelihoods for each site as a weighted mixture of conditional likelihoods of the data at each site given the rate class, our approach takes this mixing further inside Felsenstein's algorithm, mixing the transition probability matrices themselves. We show in Kosakovsky Pond *et al.* (2011) that doing this is equivalent to marginalizing over all possible assignments of rate classes to branches. The model in Kosakovsky Pond *et al.* (2011) allows each branch to have a set of 3 selection parameters and 2 mixture proportions, and constructs a likelihood ratio test for selection affecting individual branches, using the transition probability mixture approach to avoid the overly restrictive assumptions of previous branch-site tests. These previous tests are shown to behave poorly when such assumptions are violated, leading to loss of power or inflated false positive rates, while our random effects test is well behaved.

In Murrell *et al.* (2012*b*), we define a Mixed Effects Model of Evolution (MEME) that allows each site to have two non-synonymous rates, a synonymous rate, and a mixture proportion (interpreted as the proportion of branches evolving under the larger non-synonymous rate at this site). A likelihood ratio test is used to detect episodic selection at individual sites. This test can detect sites even where only a small proportion of branches are evolving under positive selection. Existing tests which assume that selection is constant across branches, effectively relying on an averaged selection pressure, identify only purifying selection at such sites. Using MEME on 16 empirical alignments, we show that the number of sites with detectable selection is approximately 4 times greater than previous tests would suggest, and conclude that the number of sites evolving under positive selection may have been greatly underestimated.

Papers for this chapter:

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*, vol. 28, no. 11, pp. 3033–3043. ISSN 1537-1719.
Available at: `http://dx.doi.org/10.1093/molbev/msr125`

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K. and Kosakovsky Pond, S.L. (2012). Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLoS Genet*, vol. 8, no. 7, pp. e1002764+.
Available at: `http://dx.doi.org/10.1371/journal.pgen.1002764`

2.2 Contribution statement

My contribution to Kosakovsky Pond *et al.* (2011): The idea for weighted mixtures of probability transition matrices was mine, but Sergei Kosakovsky Pond implemented and tested the Branch-Site random effects model (BS-REL) that first used these mixture matrices. The authorship order reflects Sergei's leading role here. I wrote and edited some sections of the paper. I have included this paper, partly because it is an important prelude to the other paper in this chapter, and partly because the mixture idea which allows random effects models over branches may turn out to be my largest contribution to this field.

My contribution to Murrell *et al.* (2012*b*): The idea for the model and test was mine. I wrote the HyPhy code implementing the test, and ran simulations. Sergei Kosakovsky Pond ported the code to the Datamonkey web server, and refined the test statistic distribution. I wrote the first draft of the paper, and refined it along with my co-authors. Joel Wertheim contributed biological expertise and interpreted the detected sites for the empirical datasets.

Chapter 3

Modeling HIV-1 Drug Resistance as Episodic Directional Selection

3.1 Summary

When exposed to treatment, HIV-1 and other rapidly evolving viruses have the capacity to acquire drug resistance mutations (DRAMs), which limit the efficacy of antivirals. There are a number of experimentally well characterized HIV-1 DRAMs, but many mutations whose roles are not fully understood have also been reported. In Murrell *et al.* (2012*a*) we construct evolutionary models that identify the locations and targets of mutations conferring resistance to antiretrovirals from viral sequences sampled from treated and untreated individuals. While the evolution of drug resistance is a classic example of natural selection, existing analyses fail to detect the majority of DRAMs. We show that, in order to identify resistance mutations from sequence data, it is necessary to recognize that in this case natural selection is both episodic (it only operates when the virus is exposed to the drugs) and directional (only mutations to a particular amino-acid confer resistance while allowing the virus to continue replicating). The new class of models that allow for the episodic and directional nature of adaptive evolution performs very well at recovering known DRAMs, can be useful at identifying unknown resistance-associated mutations, and is generally applicable to a variety of biological scenarios where similar selective forces are at play.

The paper for this chapter:
Murrell, B., de Oliveira, T., Seebregts, C., Kosakovsky Pond, S.L., Scheffler, K., (2012). Modeling HIV-1 Drug Resistance as Episodic Directional Selection. *PLoS Comput Biol*, vol. 8, no. 5, pp. e1002507+.
Available at: `http://dx.doi.org/10.1371/journal.pcbi.1002507`

3.2 Contribution statement

The episodic directional selection model was designed by myself and Konrad Scheffler, and implemented by myself. The datasets and simulations were constructed by myself. Sergei Kosakovsky Pond contributed a modified DEPS (Directional Evolution in Protein Sequences) model for an episodic directional model of amino acid evolution. The paper was written by myself and edited by Konrad Scheffler and Sergei Kosakovsky Pond, with suggestions from the other co-authors.

Chapter 4

FUBAR : An efficient analysis of the molecular footprint of natural selection

4.1 Summary

Model-based selection analyses that attempt to detect sites evolving under non-neutral selection constraints often model the site-to-site variation in selection parameters as a random effect. Random effects methods can be slow, and are restricted to using a relatively small number of discrete rate categories (see section 1.6.2.2), placing unrealistic constraints on the distribution of selection parameters over sites. Such methods are also prohibitively slow for large alignments. We present an approximate Bayesian method that allows rich, flexible site-to-site variation, which improves the statistical performance of the method, while still detecting selection much faster than current methods.

By exploiting some commonly used approximations, FUBAR (Fast Unconstrained Bayesian AppRoximation) can accurately identify positive and purifying selection orders of magnitude faster than existing random effects methods and 3 to 20 times faster than fixed effects methods (with the disparity increasing for larger alignments). We introduce a fast Markov chain Monte Carlo (MCMC) routine that allows a flexible distribution over the selection parameters to be learned from the data (see A.1.3), with no parametric constraints on the shape of this distribution. This allows information to be shared between sites, yielding greater power to detect positive selection than that of fixed effects methods, but without the potential bias introduced by the overly restrictive distributions used by current random effects models.

The flexibility and speed is achieved using a precomputed grid of conditional likelihoods, which means the the shape of the distribution over the synonymous and non-synonymous rates can be inferred, using MCMC, without having to recompute the likelihood function each iteration. From a practical

perspective, selection analyses of smaller alignments that typically required hours of computation time now take just minutes. Very large alignments, which were previously intractable, can now be analyzed in reasonable time. We demonstrate this on a large influenza Haemagglutinin dataset (3142 sequences), which took just 1.5 hours to complete.

4.2 Contribution statement

The method was conceived by Konrad Scheffler and myself. The first version of the code for FUBAR was written by myself, assisted by Sasha Moola and Thomas Weighill. Amandla Mabona assisted with some theoretical arguments about the optimal grid scaling (which did not make it into the final version of the paper) and with the PAML comparison. Sergei Kosakovsky Pond parallelized and ported the code to Datamonkey. Daniel Sheward provided a biological interpretation of the influenza analysis. Konrad Scheffler and Sergei Kosakovsky Pond suggested changes to the manuscript.

Chapter 5

Non-Negative Matrix Factorization for Learning Alignment-Specific Models of Protein Evolution

5.1 Summary

Models of protein evolution are used to align protein sequences, construct phylogenies and infer details about the evolutionary process. In Murrell *et al.* (2011), we consider a class of models that quantify the exchangeability of each of 190 amino acid pairs. Originally, these were constructed from large datasets involving different proteins, and were intended to describe protein evolution generally. More recently, researchers have found that amino acid exchangeabilities can be very different for different genes or organisms, and that models constructed from gene-specific or organism-specific datasets outperform generalist models. Large, specific datasets are seldom available, however. We propose a method, based on a mathematical technique called non-negative matrix factorization (NNMF), that achieves a compromise between the generalist and specialist approaches. Our method uses a large, general dataset to estimate a set of basis matrices, and then learns a small number of parameters from a single alignment of interest. The resulting model of protein evolution is specialized to match a single alignment, with the degree of specialization adapted to suit the richness of the data. Our new models outperform existing approaches in terms of model fit, quantify the degree of conservation of different amino acid properties, and lead to improved inference of phylogenies.

The paper for this chapter:
Murrell, B., Weighill, T., Buys, J., Ketteringham, R., Moola, S., Benade, G., du Buisson, L., Kaliski, D., Hands, T. and Scheffler, K. (2011). Non-Negative

Matrix Factorization for Learning Alignment-Specific Models of Protein Evolution. *PLoS ONE*, vol. 6, no. 12, pp. e28898+.
Available at: `http://dx.doi.org/10.1371/journal.pone.0028898`

5.2 Contribution statement

This project arose from a computational biology workshop. Along with Konrad Scheffler, I posed the problem, conjectured a solution, and then worked with a team of undergraduate student assistants to implement it. The components of the solution were implemented during the workshop. A collection of training rate matrices were estimated from a large collection of Pandit alignments, assisted by Gerdus Benade and Lise du Buisson. Matlab's NNMF routine was used to perform the factorization, assisted by Jan Buys and Tristan Hands. Code to infer the mixture weights was written in HyPhy, assisted by Thomas Weighill and Sasha Moola. Examination of the amino acid properties of the basis matrices was assisted by Daniel Kaliski. Running the comparisons on the UCSD cluster was assisted by Robert Ketteringham. The estimates of the phylogenetic impact of the method were assisted by Sasha Moola. Analyzing results was performed after the workshop, by myself. I wrote the first draft of the paper, and refined it along with Konrad Scheffler.

Chapter 6

Synopsis

6.1 Introduction

The papers comprising this dissertation have contributed a number of methodologies and models for the analysis of biological sequence data. This chapter will attempt to summarize the contributions, clarifying the relationships between the models, and suggesting profitable avenues for future research.

We distinguish between modeling contributions and methodological contributions: the former is where a new model is created using modeling techniques already available in the field, but the latter is the introduction of novel techniques themselves. MEDS and EDEPS are examples of models created from existing techniques to address a previously unattended biological scenario. Using a mixture of continuous time Markov chains to construct a random effects model of branch to branch rate variation is an example of a methodological contribution, and BS-REL and MEME are the models that employ this new technique.

Table 6.1 provides a list of the models contributed by this dissertation, as well as the key properties of each model. They are characterized by: the kind of data they describe, codon or amino acid (AA); by how they account for heterotachy, using mixtures to incorporate random effects over branches (Mixture), or using a fixed *a priori* partition over branches (Fixed); by how they model variation over sites, with site-specific parameters that are optimized (Fixed), or in the random effects framework (Random), or even relying on the branch mixtures to implicitly allow site heterogeneity (Implicit); or by what kind of selection they model, whether it be in the form of $\omega = \beta/\alpha$ (Diversifying) or elevated rates towards specific amino acids (Directional), or in the form of aggregated exchangabilities between amino acids (Implicit).

Table 6.1: A summary of the models introduced in this dissertation.

	Data	Heterotachy	Site variation	Selection
NNMF	AA	None	None	Implicit[a]
BS-REL	Codon	Mixture	Implicit[b]	Diversifying
MEME	Codon	Mixture	Fixed	Diversifying
MEDS	Codon	Fixed	Fixed	Directional
EDEPS	AA	Fixed	Random	Directional
FUBAR	Codon	None	Random	Diversifying

[a] NNMF captures selection implicitly by accounting for different substitution rates between different amino acids

[b] BS-REL allows site-to-site variation implicitly - the process is identical from one site to another, but each branch at each site is a random draw, allowing site to site variation

6.2 Methodological contributions

This section describes the methodological contributions made by this dissertation - new ways to model evolutionary processes.

6.2.1 Using dimensionality reduction to reduce model complexity

In Murrell *et al.* (2011), we introduced a method to reduce the statistical complexity of amino acid substitution models. The modeling technique was presented in the context of that specific application, but it should have more general applicability as well. From a large collection of large datasets, we learn particular parameter values for over-parameterized models. From this set of fitted models, we then learn which model dimensions are critical, and which can be ignored. Our particular method of dimensionality reduction, non-negative matrix factorization, also has the benefit of learning which parameters vary together - when parameters vary together, only one degree of freedom is required to accommodate this variation.

This can be seen as an attempt to automate the discovery of parsimonious models. This stands in contrast with human-driven model development: When the history of the field of nucleotide model development is considered, for example, the first models were simple, and complexity was added incrementally, usually guided by biological hypotheses (eg. the step from JC69 to the inclusion of the transition/transversion rate ratio). In contrast to this hypothesis-driven model development, our NNMF approach begins with maximally complex models, and uses a data-driven dimensionality reduction step to achieve a spectrum of models of varying complexity. Wherever there is a large amount of data and a way to construct models of varying complexity in a commensurable manner (where parameters from more complex models are identifiable with parameters from simpler models), this method of data-driven model discovery could prove useful.

6.2.2 Modeling heterotachy using mixtures of Markov processes

The approach for allowing branch to branch rate variation in the random effects framework (Kosakovsky Pond *et al.*, 2011; Murrell *et al.*, 2012*b*) should have applications across phylogenetics. In this dissertation, we applied the technique to codon models to better handle cases where selection is episodic, but it can be used wherever a model requires branch to branch substitution process variation, regardless of the kind of model used (nucleotide, amino acid, codon etc.).

Heterotachy is a ubiquitous feature of evolution (Lopez *et al.*, 2002). It can bias estimates of node dates (Wertheim *et al.*, 2012), and, as we show, obscure selection (Murrell *et al.*, 2012*b*). We have used our mixture of Markov processes along each branch to alleviate the latter problem, but it may yield an improvement with respect to the other concerns too. Heterotachy itself can describe a variety of ways in which the process can change from one branch to another. Our mixture of Markov processes assumes branch-to-branch and site-to-site independence. This works well when modeling selection, but might not work well when modeling variation that violates this independence assumption: strong correlations between neighbouring branches (covarion-like processes) or when a large number of sites all switch to a different process along a single lineage. It should, however, always be better than not modeling branch to branch variation at all.

6.2.3 Building complex models efficiently

The approach employed by FUBAR to efficiently detect $\beta > \alpha$ can also be considered more generally: sacrifice correctness when estimating unimportant parameters to improve model complexity for critical parameters. This is not a novel tradeoff, even within phylogenetics. The CAT approximation (Stamatakis, 2006), for example, yields dramatic improvements in the speed of phylogeny reconstruction, by abandoning site to site variation using a discretized Γ-distribution, and instead estimating a distinct rate for each site. The speed increase comes from forcing these rates to be one of a number of discrete categories, which allows the re-use of the matrix exponential across sites. The CAT approximation allows the efficient discovery of phylogenies that have improved likelihoods, even when evaluated under the Γ-distribution.

FUBAR's use of the α, β grid is very similar, but the approach is used to capture potentially complex distributions of site to site variation in selection parameters with a fixed phylogeny, rather than to estimate the phylogeny itself. The fact that the CAT approximation is so successful in phylogeny estimation

(Price *et al.*, 2010) suggests that grid-based approaches like FUBAR might be useful even when the phylogeny is unknown and needs to be reconstructed under the full model.

Grid-based approximations do have their limitations. If the number of parameters that need to vary from one site to another (1 in the CAT approximation, 2 in FUBAR) is too large, then building a grid of conditional likelihoods becomes infeasible, because the number of evaluations grows exponentially with grid size. MEDS, for example, would probably not benefit from a FUBAR-like implementation, because the synonymous rate, background and foreground non-synonymous rates, and directional rate would all have to vary from site to site, enforcing a very coarse grid. It remains to be seen just how many dimensions these sorts of grid-based approaches can tolerate before their gains are nullified.

6.3 List of problems addressed

This section describes practical biological problems to which this dissertation has contributed, either by addressing new problems that had no existing solutions in the literature, or by improving upon existing methods. These improvements could be in terms of performance in model selection criteria, computational efficiency, or statistical performance in the form of power and false positive rates.

6.3.1 Detecting lineages under selection

Detecting lineages under positive selection has been a popular kind of selection analysis since the first methods were developed by Yang (1998). Prior to this dissertation, there were a number of model-based methods for detecting lineages under adaptive evolution. The earlier approaches assumed that rates of evolution on the target lineage were constant across sites, but this was relaxed with the introduction of branch-site methods (Yang and Nielsen, 2002; Zhang *et al.*, 2005; Yang and Reis, 2011). These branch-site methods partitioned sites into foreground and background, allowing a small number of different patterns (positive on foreground but purifying on background, positive on foreground but neutral on background etc.) at each site. We used our mixture of Markov processes to relax these assumptions, integrating over all possible rate categories on each branch. The resulting method, BS-REL (Kosakovsky Pond *et al.*, 2011), has greater power to detect selection, and does not break down (as the existing branch-site methods do) when selection is variable on the background branches. Since its introduction, this method has been used in a number of papers (9 citations as of July 2012), including selection analyses on a glucose transporter gene in fruit bats (Shen *et al.*, 2012), antifreeze proteins in *Fragilariopsis* (Sorhannus, 2011), Bean Necrotic

Mosaic Virus (de Oliveira *et al.*, 2012), Dengue virus (Costa *et al.*, 2012), and Hepacviruses in primates (Patel *et al.*, 2012).

6.3.2 Estimating alignment specific amino acid models

Models of amino acid substitution are used throughout comparative bioinformatics - for aligning sequences, constructing phylogenies, and as a baseline against which to detect elevated substitution rates. Specifying a model requires specifying 190 symmetric exchangeability rates between all pairs of amino acids. When analyzing a new sequence alignment, a biologist would select an amino acid model from a list of pre-estimated models, either using a model selection tool or, in some packages, using the only available option. Depending on the alignment, this selected model could be very well or very poorly suited to the data.

With the NNMF approach we propose in (Murrell *et al.*, 2011), one can now fit a model to a specific dataset. The model complexity is automatically adapted to suit the amount of data. The resulting amino acid models can be used to refine alignments or phylogenies, or as baseline AA models used to detect directional selection.

6.3.3 Detecting selection in larger alignments

The speed of selection analyses places restrictions on the size of alignments that can be studied. This is especially true of publicly available webservers such as Datamonkey (Delport *et al.*, 2010*a*), where many jobs are being processed (166 jobs per day over July 2012) - computational considerations become critical. For example, FEL and REL analyses on Datamonkey are restricted to 500 and 75 taxa respectively, which makes the analysis of large alignments inconvenient.

FUBAR has extended the range of selection analyses. Datamonkey allows FUBAR to analyze 5000 sequences, which is an order of magnitude greater than any other Datamonkey analysis. The primary practical contribution of FUBAR is the efficient detection of selection in much larger alignments. This will hopefully prove useful in the era of large databases and deep sequencing.

6.3.4 Detecting episodic positive selection masked by pervasive purifying selection

The blind discovery of individual sites under selection - where nothing beyond the alignment and phylogeny is known or provided - is a very common kind of selection analysis, with a large number of approaches addressing this problem. As was demonstrated in Kosakovsky Pond and Frost (2005*b*), once synonymous rate variation is accounted for, the performance of different methods is very similar. The conclusions in Murrell *et al.* (2012*b*) challenge that picture. A critical feature of selection was being overlooked by all existing methods that attempted to detect sites under positive selection: selection is overwhelmingly episodic. Even more importantly, the combination of positive selection and prevalent purifying selection at some sites causes methods that assume constant selection across branches to calculate an "average" β/α, which is often less than 1 simply because purifying selection is more pervasive. MEME detects almost 4 times as many sites as FEL, by relaxing the assumption of a fixed rate across all branches.

The inability to detect episodic positive selection has not previously been addressed in the literature, and - when the prevalence of this mode of selection is considered - is a crippling feature of models that fail to account for this kind of variation. We thus hope MEME will prove useful to the field. MEME became publicly available on the Datamonkey webserver before the paper describing the method was published, and, prior to that publication, results from MEME analyses were already used in two papers, finding sites under selection in circoviruses in the endangered Echo parakeet (Kundu *et al.*, 2012), and identifying episodic diversifying selection in Influenza A H3N2 (Westgeest *et al.*, 2012).

6.3.5 Episodic directional selection

When a sudden environmental shift affects a large number of lineages simultaneously, and when it is known (or at least hypothesized) which lineages were affected, then existing models may fail to adequately capture the pattern of selective forces acting at those sites. This is caused by a combination of two problems. Firstly, the environmental disruption means that selection is not constant, but episodic. Secondly, even if a model allowed branches to be partitioned into background (before the environmental shift) and foreground (after the environmental shift), the manner in which positive selection is usually modeled (β/α) is inappropriate for this scenario, where selection tends to favor a particular amino acid, which is under purifying selection once it becomes fixed. This scenario is exactly what we observe in the emergence of drug resistance, and models such as MEDS or EDEPS, that appropriately account for both the episodic and directional features, are much better at identifying drug resistance mutations than those that do not. We count this kind of scenario among the biological problems that did not previously have an adequate solution.

6.4 Future work

There is much room for model development using the techniques introduced in this dissertation. This section will outline research directions we believe should prove profitable.

6.4.1 FUBAR-DEPS

DEPS and EDEPS detect sites in proteins evolving under directional selection, using a random effects approach. DEPS uses an underlying amino acid model, and detects sites where evolution towards a particular amino acid is elevated across the phylogeny. This is done using a random effects model, where, for each possible target amino acid, there is a single null category and a single category with a rate multiplier greater then 1.

DEPS and EDEPS could likely be improved by a grid-based implementation, in the style of FUBAR. This would have many foreseeable benefits: 1) The very coarse 2 category site to site variation could be enriched with far more categories, which should improve the power to detect weaker directional selection. 2) Site to site variation in the overall (non-directional) substitution rate could be efficiently incorporated, which could prevent false positives due to directional rate parameters compensating for a non-directional rate increase. 3) Rather than considering 20 distinct alternative models, with each having accelerated substitutions to one amino acid, the hierarchical Bayesian approach could consider a composite model that included 20 models at once. 4) The speed of the method could be dramatically increased, allowing a sophisticated directional (or episodic directional) analysis to be performed on very large alignments.

6.4.2 Branch mixtures and site mixtures

It is interesting to compare the manner in which traditional random effects models allow for variation over sites to how our approach allows for random effects variation over branches. If D is the data at a single site, then $P(D) = \sum_{k=1}^{K} P(D|\omega_k)P(\omega_k)$ computes the probability of D when the value of ω is an unknown draw from the discrete distribution $P(\omega_k)$. Thus the marginal likelihood is computed as a weighted mixture of conditional likelihoods. When we want ω to vary from branch to branch, with the process along each branch being one of a number of categories, we use a weighted mixture of transition matrices: $T(t)_{ij} = \sum_{k=1}^{K} P(S = j|S_{pa} = i, t, \omega_k)P(\omega_k)$. $T(t)_{ij}$ is the probability of transitioning from the parent state $S_{pa} = i$ to the descendent state $S = j$.

In both cases we are using weighted mixtures to describe a situation where the overall distribution is generated by an unknown selection from a number of possible processes, but the models differ at the level (site or branch) of the randomness (when thinking of the generative model) or the uncertainty (when thinking of the inferential procedure). In one case, the mixing happens once per site (we call these site-mixtures), and, in the other, once per branch per site (branch-mixtures).

There is no impediment to allowing both kinds of variation. MEME - which used two categories to allow branch to branch variation - used site specific parameters (fixed effects) to achieve variation from site to site. It would be interesting to consider far more parametrically efficient models where branch and site variation are both modeled in the random effects framework.

As an example, consider the following hypothetical model meant to account for heterogenous positive selection combining site to site and branch to branch variation: There are 7 site categories, $M1, \ldots, M7$, with 6 associated mixture weights controlling the site proportions. We model the sites as belonging to these categories using a 7 category site-mixture. There are three ω values, $\omega_1, \ldots, \omega_3$. If a site is $M1$, then all branches evolve under ω_1. Similarly for $M2$ and $M3$ with ω_2 and ω_3. If a site is $M4$, then we use a branch-mixture to let each branch evolve under either ω_1 or ω_2, with a mixture proportion parameter p_4 controlling the branch proportions. If a site is $M5$, then we use a branch-mixture of ω_1 and ω_3, with p_5 for the proportions. If $M6$, then we use a branch-mixture of ω_2 and ω_3, with p_6 for the proportions. Finally, if a site is $M7$, then we use a branch-mixture of ω_1, ω_2 and ω_3, with p_{7a} and p_{7b} controlling the proportions.

Such a model would be very effective at handling data where some sites appear to evolve at constant rates, but others are more episodic, switching between purifying-, neutral-, and positive selection. The proportion of sites with episodic vs. constant positive selection could also be estimated, which would be an interesting metric under which to compare different genes. This model has 6 site level proportion parameters, 5 branch level proportion parameters, 3 omega parameters, but, in a sense, this is the smallest "complete" model with 3 ω values that can be constructed. We might want the values of ω to differ between the different Ms, or we might want to allow site to site variation in the synonymous rate rather than just ω, or a large number of potential variations, all of which complicate matters further.

Combining site and branch process mixing should provide a fruitful ground upon which to explore novel models and strategies to detect selection. Other kinds of variation, such as covarion models, could also be incorporated, and comparisons between different modes of evolution would serve to further elucidate the nature of natural selection.

Appendices

Appendix A

Approaches to inference

A.1 Inference and model selection

Inference plays multiple roles in phylogenetics. Phylogenies themselves can be compared to see if the data supports one possible ancestry over another. This will not be dealt with in this dissertation, which instead concerns itself with which model the data supports, or even what range of values are supported for a model parameter. All probabilistic inference is based on the likelihood function, but it can be used in different ways. Each has vastly different philosophical grounding, but in phylogenetics the choice is often based on practical considerations.

A.1.1 Frequentist

Frequentist inference, for our purposes, proceeds by constructing statistical tests with reliable frequency properties. Typically, null and alternative hypotheses are compared. A test statistic is obtained from the data, and a p-value derived from the distribution of the test statistic under the null model describes the probability of obtaining a test statistic at least that extreme given that the null model was true (ie. given that the data was generated under the null model).

A.1.1.1 Likelihood ratio tests for model comparison

When models are nested - when the alternative model reduces to the null model for specific parameter values - then likelihood ratio tests (LRTs) may be employed. The log-likelihood functions of both the null and alternative models are maximized, yielding $ML_{null} = \max_{\theta_{null}} P(D|\theta_{null})$ and $ML_{alt} = \max_{\theta_{alt}} P(D|\theta_{alt})$. If we define our likelihood ratio statistic (LRS) as

$$LRS = -2\ln(ML_{null}) + 2\ln(ML_{alt}), \tag{A.1.1}$$

then, under some regularity conditions (Self and Liang, 1987), when the null model is true the LRS will be asymptotically χ^2_d distributed with degrees of freedom d equal to the difference in the number of free parameters between the null and the alternative models. Knowing this, we can use the inverse of the cumulative χ^2_d distribution to compute a p-value, telling us how often we obtain a likelihood improvement this large when the null model is used to generate the data. When this p-value is very small, we can reject the null hypothesis.

The χ^2_d distribution of the statistic only holds asymptotically, and in practice such tests are often conservative. Further, if the null models constrains a parameter to be on the boundary of parameter space, a mixture of χ^2 distributions will describe the distribution of the test statistic (Self and Liang, 1987). Simulations under the null model are used to investigate how well the tests behave for different amounts of data.

A.1.2 Information theoretic

When models are not nested, the conditions for LRTs do not obtain. The Akaike information criterion (AIC, Akaike, 1974) provides a convenient alternative, based on information theory. It provides an asymptotically unbiased estimate of the expected information loss (in the form of Kullback-Leibler divergence) incurred relative to the true data generating distribution when one model is used instead of another. If ML is the maximized log-likelihood of a model, then

$$AIC = 2k - 2\ln(ML) \tag{A.1.2}$$

where k is the number of parameters. AIC gets smaller when the maximized likelihood is larger, but gets larger when the number of parameters increases. To compare models within a candidate set, the AIC for each model is calculated, and then difference between AIC values indicates how much support there is for one model over another.

As with LRTs, AIC is only asymptotically valid. With small datasets, a second order correction is recommended (Burnham and Anderson, 2002). One of the components in this correction is the number of observations, which is particularly problematic in phylogenetics, since it is not clear exactly what constitutes a single observation. See Posada and Buckley (2004) for further discussion.

A.1.3 Bayesian

Bayesian methods of inference involve specifying a prior distribution over models and parameters, and using Bayes' theorem to compute the posterior distribution. The data plays its role in updating the posterior through the likelihood function:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{A.1.3}$$

$P(D)$ is a normalizing constant whose exact calculation requires integrating over theta: $P(D) = \int_\theta P(D|\theta)P(\theta)$. Such integrals are often intractable. Approximate Bayesian inference techniques such as Markov Chain Monte Carlo allow samples to be drawn from the posterior when only point-wise evaluation of the likelihood and prior functions is available. The integral in the denominator of Bayes' theorem does not need to be evaluated. Inference can then be performed on the posterior samples.

Bayesian approaches have the advantage that prior information can be included in the modeling process through the specification of the prior distribution over parameters, but the disadvantage that prior distributions are required even in the absence of such knowledge. In such cases vague or non-informative priors can be used. Bayesian inference procedures lack the frequency coverage guarantees of frequentist hypothesis testing and are often slower than frequentist approaches, but are sometimes the only option when a large number of nuisance parameters would hamper frequentist inference (see Rodrigue *et al.*, 2010, for a recent relevant example).

A.2 Empirical Bayes

Phylogenetic models of evolution are often hierarchical - parameters themselves can be unobserved variables that have a prior distribution specified over them. When calculating the model likelihood, this prior must be integrated out. We frequently encounter cases where the prior distribution itself has parameters, but we are more interested in the value of the latent variable. A fully Bayesian treatment would specify a hyper-prior over the prior distribution, and integrate over all parameters. Empirical Bayes, which can be viewed as an approximation to this, finds the maximum likelihood values of the prior parameters, and then uses Bayes' theorem to perform inference over the latent variable. It is often faster than a fully Bayesian treatment.

This approximation is most useful when inference over the unobserved variable is insensitive to changes in the prior parameters. When such inference is sensitive to some prior parameters, but not to others, Bayes Empirical Bayes (Deely and Lindley, 1981; Yang *et al.*, 2005) can be useful, where some prior parameters (the ones that minimally affect inference) are fixed at their maximum likelihood values, but the others are integrated out.

List of References

Akaike, H. (1974 December). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723. ISSN 0018-9286.
Available at: `http://dx.doi.org/10.1109/TAC.1974.1100705`

Anisimova, M. and Kosiol, C. (2009 February). Investigating Protein-Coding Sequence Evolution with Probabilistic Codon Substitution Models. *Molecular Biology and Evolution*, vol. 26, no. 2, pp. 255–271. ISSN 1537-1719.
Available at: `http://dx.doi.org/10.1093/molbev/msn232`

Barry, D. and Hartigan, J.A. (1987 May). Statistical Analysis of Hominoid Molecular Evolution. *Statistical Science*, vol. 2, no. 2, pp. 191–207. ISSN 0883-4237.
Available at: `http://dx.doi.org/10.1214/ss/1177013353`

Burnham, K.P. and Anderson, D. (2002 July). *Model Selection and Multi-Model Inference*. 2nd edn. Springer. ISBN 0387953647.
Available at: `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0387953647`

Costa, R.L., Voloch, C.M. and Schrago, C.G. (2012 March). Comparative evolutionary epidemiology of dengue virus serotypes. *Infection, Genetics and Evolution*, vol. 12, no. 2, pp. 309–314. ISSN 15671348.
Available at: `http://dx.doi.org/10.1016/j.meegid.2011.12.011`

Cuevas, J.M., Domingo-Calap, P. and Sanjuán, R. (2011 July). The Fitness Effects of Synonymous Mutations in DNA and RNA Viruses. *Molecular Biology and Evolution*. ISSN 1537-1719.
Available at: `http://dx.doi.org/10.1093/molbev/msr179`

de Oliveira, A.S., Melo, F.L., Inoue-Nagata, A.K., Nagata, T., Kitajima, E.W. and Resende, R.O. (2012 June). Characterization of Bean Necrotic Mosaic Virus: A Member of a Novel Evolutionary Lineage within the Genus Tospovirus. *PLoS ONE*, vol. 7, no. 6, pp. e38634+.
Available at: `http://dx.doi.org/10.1371/journal.pone.0038634`

Deely, J.J. and Lindley, D.V. (1981 December). Bayes Empirical Bayes. *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 833+. ISSN 01621459.
Available at: `http://dx.doi.org/10.2307/2287578`

Delport, W., Poon, A.F.Y., Frost, S.D.W. and Kosakovsky Pond, S.L. (2010 Oct*a*). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, vol. 26, no. 19, pp. 2455–7.

Delport, W., Scheffler, K., Botha, G., Gravenor, M.B., Muse, S.V. and Kosakovsky Pond, S.L. (2010 August*b*). CodonTest: Modeling Amino Acid Substitution Preferences in Coding Sequences. *PLoS Comput Biol*, vol. 6, no. 8, pp. e1000885+.
Available at: `http://dx.doi.org/10.1371/journal.pcbi.1000885`

Delport, W., Scheffler, K. and Seoighe, C. (2009 January). Models of coding sequence evolution. *Briefings in Bioinformatics*, vol. 10, no. 1, pp. 97–109. ISSN 1477-4054.
Available at: `http://dx.doi.org/10.1093/bib/bbn049`

Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1998 May). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. ISBN 0521629713.
Available at: `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0521629713`

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, vol. 17, no. 6, pp. 368–376. ISSN 0022-2844.
Available at: `http://view.ncbi.nlm.nih.gov/pubmed/7288891`

Goldman, N. and Yang, Z. (1994 September). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution*, vol. 11, no. 5, pp. 725–736. ISSN 0737-4038.
Available at: `http://mbe.oxfordjournals.org/content/11/5/725.abstract`

Guindon, S., Rodrigo, A.G., Dyer, K.A. and Huelsenbeck, J.P. (2004 August). Modeling the site-specific variation of selection patterns along lineages. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 35, pp. 12957–12962. ISSN 1091-6490.
Available at: `http://dx.doi.org/10.1073/pnas.0402177101`

Huelsenbeck, J.P. and Ronquist, F. (2001 August). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)*, vol. 17, no. 8, pp. 754–755. ISSN 1367-4803.
Available at: `http://dx.doi.org/10.1093/bioinformatics/17.8.754`

Hunter, L. (1992). Artificial intelligence and molecular biology. In: *Proceedings of the tenth national conference on Artificial intelligence*, AAAI'92, pp. 866–868. AAAI Press. ISBN 0-262-51063-4.
Available at: `http://portal.acm.org/citation.cfm?id=1867269`

Kosakovsky Pond, S., Delport, W., Muse, S.V. and Scheffler, K. (2010 July). Correcting the Bias of Empirical Frequency Parameter Estimators in Codon Models. *PLoS ONE*, vol. 5, no. 7, pp. e11230+.
Available at: `http://dx.doi.org/10.1371/journal.pone.0011230`

Kosakovsky Pond, S. and Muse, S.V. (2005 December). Site-to-Site Variation of Synonymous Substitution Rates. *Molecular Biology and Evolution*, vol. 22, no. 12, pp. 2375–2385. ISSN 1537-1719.
Available at: `http://dx.doi.org/10.1093/molbev/msi232`

Kosakovsky Pond, S.L. and Frost, S.D.W. (2005 May*a*). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, vol. 21, no. 10, pp. 2531–2533. ISSN 1460-2059.
Available at: `http://dx.doi.org/10.1093/bioinformatics/bti320`

Kosakovsky Pond, S.L. and Frost, S.D.W. (2005 May*b*). Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Molecular Biology and Evolution*, vol. 22, no. 5, pp. 1208–1222. ISSN 1537-1719.
Available at: `http://dx.doi.org/10.1093/molbev/msi105`

Kosakovsky Pond, S.L., Frost, S.D.W. and Muse, S.V. (2005 March). HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, vol. 21, no. 5, pp. 676–679. ISSN 1367-4803.
Available at: `http://dx.doi.org/10.1093/bioinformatics/bti079`

Kosakovsky Pond, S.L., Murrell, B., Fourment, M., Frost, S.D., Delport, W. and Scheffler, K. (2011 November). A random effects branch-site model for detecting episodic diversifying selection. *Molecular biology and evolution*, vol. 28, no. 11, pp. 3033–3043. ISSN 1537-1719.
Available at: `http://dx.doi.org/10.1093/molbev/msr125`

Kryazhimskiy, S. and Plotkin, J.B. (2008 December). The Population Genetics of dN/dS. *PLoS Genet*, vol. 4, no. 12, pp. e1000304+. ISSN 1553-7404.
Available at: `http://dx.doi.org/10.1371/journal.pgen.1000304`

Kundu, S., Faulkes, C.G., Greenwood, A.G., Jones, C.G., Kaiser, P., Lyne, O.D., Black, S.A., Chowrimootoo, A. and Groombridge, J.J. (2012 February). Tracking Viral Evolution During a Disease Outbreak: The Rapid and Complete Selective Sweep of a Circovirus in the Endangered Echo parakeet. *Journal of Virology*. ISSN 1098-5514.
Available at: `http://dx.doi.org/10.1128/JVI.06504-11`

Lopez, P., Casane, D. and Philippe, H. (2002 January). Heterotachy, an important process of protein evolution. *Molecular biology and evolution*, vol. 19, no. 1, pp. 1–7. ISSN 0737-4038.
Available at: `http://view.ncbi.nlm.nih.gov/pubmed/11752184`

Löytynoja, A. and Goldman, N. (2008 June). Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science*, vol. 320, no. 5883, pp. 1632–1635. ISSN 1095-9203.
Available at: `http://dx.doi.org/10.1126/science.1158395`

Massingham, T. and Goldman, N. (2005 March). Detecting Amino Acid Sites Under Positive Selection and Purifying Selection. *Genetics*, vol. 169, no. 3, pp. 1753–1762. ISSN 1943-2631.
Available at: `http://dx.doi.org/10.1534/genetics.104.032144`

Murrell, B., de Oliveira, T., Seebregts, C., Kosakovsky Pond, S.L., Scheffler, K., on behalf of the Southern African Treatment and Consortium, R.N.S. (2012 May*a*). Modeling HIV-1 Drug Resistance as Episodic Directional Selection. *PLoS Comput Biol*, vol. 8, no. 5, pp. e1002507+.
Available at: `http://dx.doi.org/10.1371/journal.pcbi.1002507`

Murrell, B., Weighill, T., Buys, J., Ketteringham, R., Moola, S., Benade, G., du Buisson, L., Kaliski, D., Hands, T. and Scheffler, K. (2011 December). Non-Negative Matrix Factorization for Learning Alignment-Specific Models of Protein Evolution. *PLoS ONE*, vol. 6, no. 12, pp. e28898+.
Available at: `http://dx.doi.org/10.1371/journal.pone.0028898`

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K. and Kosakovsky Pond, S.L. (2012 July*b*). Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLoS Genet*, vol. 8, no. 7, pp. e1002764+.
Available at: `http://dx.doi.org/10.1371/journal.pgen.1002764`

Muse, S.V. and Gaut, B.S. (1994 September). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, vol. 11, no. 5, pp. 715–724. ISSN 1537-1719.
Available at: `http://mbe.oxfordjournals.org/content/11/5/715.abstract`

Nielsen, R. and Yang, Z. (1998 March). Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics*, vol. 148, no. 3, pp. 929–936. ISSN 1943-2631.
Available at: `http://www.genetics.org/content/148/3/929.abstract`

Patel, M.R., Loo, Y.-M., Horner, S.M., Gale, M. and Malik, H.S. (2012 March). Convergent Evolution of Escape from Hepaciviral Antagonism in Primates. *PLoS Biol*, vol. 10, no. 3, pp. e1001282+.
Available at: `http://dx.doi.org/10.1371/journal.pbio.1001282`

Posada, D. and Buckley, T.R. (2004 October). Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Systematic Biology*, vol. 53, no. 5, pp. 793–808. ISSN 1076-836X.
Available at: `http://dx.doi.org/10.1080/10635150490522304`

Posada, D. and Crandall, K.A. (1998 January). MODELTEST: testing the model of DNA substitution. *Bioinformatics*, vol. 14, no. 9, pp. 817–818. ISSN 1460-2059.
Available at: `http://dx.doi.org/10.1093/bioinformatics/14.9.817`

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010 March). FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, vol. 5, no. 3, pp. e9490+. ISSN 1932-6203.
Available at: `http://dx.doi.org/10.1371/journal.pone.0009490`

Rodrigue, N., Philippe, H. and Lartillot, N. (2010 March). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4629–4634. ISSN 1091-6490.
Available at: `http://dx.doi.org/10.1073/pnas.0910915107`

Rothberg, J.M. and Leamon, J.H. (2008 October). The development and impact of 454 sequencing. *Nat Biotech*, vol. 26, no. 10, pp. 1117–1124. ISSN 1087-0156.
Available at: `http://dx.doi.org/10.1038/nbt1485`

Salemi, M. and Vandamme, A.-M. (eds.) (2003 September). *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny.* 1st edn. Cambridge University Press. ISBN 052180390X.
Available at: `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/052180390X`

Self, S.G. and Liang, K.-Y. (1987 June). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 605–610. ISSN 01621459.
Available at: `http://dx.doi.org/10.2307/2289471`

Shen, B., Han, X., Zhang, J., Rossiter, S.J. and Zhang, S. (2012 April). Adaptive Evolution in the Glucose Transporter 4 Gene Slc2a4 in Old World Fruit Bats (Family: Pteropodidae). *PLoS ONE*, vol. 7, no. 4, pp. e33197+.
Available at: `http://dx.doi.org/10.1371/journal.pone.0033197`

Sorhannus, U. (2011). Evolution of antifreeze protein genes in the diatom genus fragilariopsis: evidence for horizontal gene transfer, gene duplication and episodic diversifying selection. *Evolutionary bioinformatics online*, vol. 7, pp. 279–289. ISSN 1176-9343.
Available at: `http://dx.doi.org/10.4137/EBO.S8321`

Stamatakis, A. (2006). Phylogenetic models of rate heterogeneity: a high performance computing perspective. In: *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*, pp. 8 pp.+. IEEE. ISBN 1-4244-0054-6.
Available at: `http://dx.doi.org/10.1109/IPDPS.2006.1639535`

Tuffley, C. and Steel, M. (1998 January). Modeling the covarion hypothesis of nucleotide substitution. *Mathematical biosciences*, vol. 147, no. 1, pp. 63–91. ISSN 0025-5564.
Available at: `http://view.ncbi.nlm.nih.gov/pubmed/9401352`

Wertheim, J.O., Fourment, M. and Kosakovsky Pond, S.L. (2012 February). Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. *Molecular biology and evolution*, vol. 29, no. 2, pp. 451–456. ISSN 1537-1719.
Available at: http://dx.doi.org/10.1093/molbev/msr266

Westgeest, K.B., de Graaf, M., Fourment, M., Bestebroer, T.M., van Beek, R., Spronken, M.I.J., de Jong, J.C., Rimmelzwaan, G.F., Russell, C.A., Osterhaus, A.D.M.E., Smith, G.J.D., Smith, D.J. and Fouchier, R.A.M. (2012 June). Genetic Evolution of Neuraminidase of Influenza A (H3N2) Viruses from 1968 to 2009 and its Correspondence to Hemagglutinin. *Journal of General Virology*. ISSN 1465-2099.
Available at: http://dx.doi.org/10.1099/vir.0.043059-0

Yang, Z. (1993 November). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, vol. 10, no. 6, pp. 1396–1401. ISSN 1537-1719.
Available at: http://mbe.oxfordjournals.org/content/10/6/1396.abstract

Yang, Z. (1997 October). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences : CABIOS*, vol. 13, no. 5, pp. 555–556. ISSN 1460-2059.
Available at: http://dx.doi.org/10.1093/bioinformatics/13.5.555

Yang, Z. (1998 May). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular biology and evolution*, vol. 15, no. 5, pp. 568–573. ISSN 0737-4038.
Available at: http://mbe.oxfordjournals.org/cgi/content/abstract/15/5/568

Yang, Z. (2006 December). *Computational Molecular Evolution (Oxford Series in Ecology and Evolution)*. Oxford University Press, USA. ISBN 0198567022.
Available at: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0198567022

Yang, Z. and Nielsen, R. (2002 June). Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Molecular Biology and Evolution*, vol. 19, no. 6, pp. 908–917. ISSN 1537-1719.
Available at: http://mbe.oxfordjournals.org/content/19/6/908.abstract

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.-M.K. (2000 May). Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics*, vol. 155, no. 1, pp. 431–449. ISSN 1943-2631.
Available at: http://www.genetics.org/content/155/1/431.abstract

Yang, Z. and Reis, M.D. (2011 March). Statistical Properties of the Branch-Site Test of Positive Selection. *Molecular Biology and Evolution*, vol. 28, no. 3, pp. 1217–1228. ISSN 1537-1719.
Available at: http://dx.doi.org/10.1093/molbev/msq303

Yang, Z. and Swanson, W.J. (2002 January). Codon-Substitution Models to Detect Adaptive Evolution that Account for Heterogeneous Selective Pressures Among Site Classes. *Molecular Biology and Evolution*, vol. 19, no. 1, pp. 49–57. ISSN 1537-1719.
Available at: `http://mbe.oxfordjournals.org/content/19/1/49.abstract`

Yang, Z., Wong, W.S. and Nielsen, R. (2005 April). Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular biology and evolution*, vol. 22, no. 4, pp. 1107–1118. ISSN 0737-4038.
Available at: `http://dx.doi.org/10.1093/molbev/msi097`

Yap, V.B., Lindsay, H., Easteal, S. and Huttley, G. (2010 March). Estimates of the Effect of Natural Selection on Protein-Coding Content. *Molecular Biology and Evolution*, vol. 27, no. 3, pp. 726–734. ISSN 1537-1719.
Available at: `http://dx.doi.org/10.1093/molbev/msp232`

Zhang, J., Nielsen, R. and Yang, Z. (2005 December). Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular Biology and Evolution*, vol. 22, no. 12, pp. 2472–2479. ISSN 1537-1719.
Available at: `http://dx.doi.org/10.1093/molbev/msi237`

A Random Effects Branch-Site Model for Detecting Episodic Diversifying Selection

Sergei L. Kosakovsky Pond,*[1] Ben Murrell,[2,3] Mathieu Fourment,[4] Simon D.W. Frost,[5] Wayne Delport,[4] and Konrad Scheffler[2]

[1]Department of Medicine, University of California San Diego, San Diego
[2]Computer Science Division, Department of Mathematical Sciences, University of Stellenbosch, Stellenbosch, South Africa
[3]Biomedical Informatics Research Division, eHealth Research and Innovation Platform, Medical Research Council, Tygerberg, South Africa
[4]Department of Pathology, University of California San Diego, San Diego
[5]Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom

***Corresponding author:** E-mail: spond@ucsd.edu.

Associate editor: Hervé Philippe

Abstract

Adaptive evolution frequently occurs in episodic bursts, localized to a few sites in a gene, and to a small number of lineages in a phylogenetic tree. A popular class of "branch-site" evolutionary models provides a statistical framework to search for evidence of such episodic selection. For computational tractability, current branch-site models unrealistically assume that all branches in the tree can be partitioned a priori into two rigid classes—"foreground" branches that are allowed to undergo diversifying selective bursts and "background" branches that are negatively selected or neutral. We demonstrate that this assumption leads to unacceptably high rates of false positives or false negatives when the evolutionary process along background branches strongly deviates from modeling assumptions. To address this problem, we extend Felsenstein's pruning algorithm to allow efficient likelihood computations for models in which variation over branches (and not just sites) is described in the random effects likelihood framework. This enables us to model the process at every branch-site combination as a mixture of three Markov substitution models—our model treats the selective class of every branch at a particular site as an unobserved state that is chosen independently of that at any other branch. When benchmarked on a previously published set of simulated sequences, our method consistently matched or outperformed existing branch-site tests in terms of power and error rates. Using three empirical data sets, previously analyzed for episodic selection, we discuss how modeling assumptions can influence inference in practical situations.

Key words: episodic selection, random effects model, evolutionary model, branch-site model.

Introduction

The inference of selection from molecular data, both along a sequence (Nielsen and Yang 1998; Suzuki and Gojobori 1999; Yang et al. 2000) and over the evolutionary tree (Yang and Nielsen 2002; Kosakovsky Pond and Frost 2005a), has been an area of active research and unrelenting debate (Suzuki and Nei 2004; Wong et al. 2004; Nozawa et al. 2009). Selective pressures can vary over both sites and time, resulting in bursts of selection localized to a subset of sites and a small number of lineages, for example, Messier and Stewart (1997).

A class of methods, termed "branch-site" tests (Yang and Nielsen 2002), was the first to offer a model-based phylogenetic hypothesis testing framework for deciding whether or not a lineage (or lineages) of interest had undergone adaptive change. Branch-site tests measure selective pressure by ω, the ratio of nonsynonymous (β) to synonymous (α) substitution rates, and if a proportion of sites in the sequence provides statistically significant support for $\omega > 1$ along the lineages of interest, then episodic positive selection is inferred. The original formulation of the method suffered from high rates of false positives when the model assumptions were violated (Zhang 2004) because the model could misidentify relaxed selective constraints as evidence of diversifying selection and was subsequently revised to address that shortcoming (Zhang et al. 2005). Typically, the lineages to be tested ("foreground" lineages) were specified a priori, until a recent extension outlined and benchmarked a sequential testing approach to examine whether any single lineage was under selection (Anisimova and Yang 2007). These branch-site methods have been used extensively, with well over 1,000 citations to date, highlighting the interest of the evolutionary community in being able to identify instances of episodic selection. Alternative approaches to capturing variable selective pressures include the covarion models of Guindon et al. (2004) and a full Bayesian treatment in the framework of Rodrigue et al. (2010).

In the context of codon evolutionary models, the selective profile of site D_s in a multiple sequence alignment can be characterized by the collection of branch-specific ω values, $(\omega_1, \ldots, \omega_B)$, denoted Ω_s, where B equals the total number of branches in the phylogeny. Existing branch-site models use three alignment-wide (i.e., shared by all sites) ratios $\omega^- < \omega^N = 1 \leqslant \omega^+$ to model strong conservation,

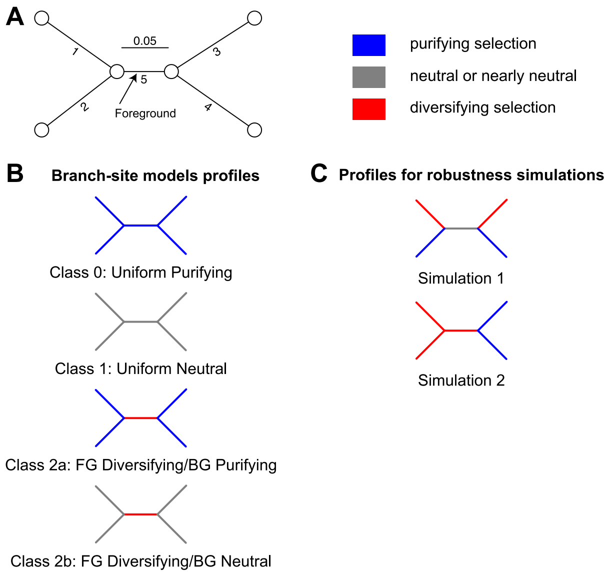
A
0.05
1
2
3
4
5
Foreground
purifying selection
neutral or nearly neutral
diversifying selection
B
Branch-site models profiles
Class 0: Uniform Purifying
Class 1: Uniform Neutral
Class 2a: FG Diversifying/BG Purifying
Class 2b: FG Diversifying/BG Neutral
C
Profiles for robustness simulations
Simulation 1
Simulation 2

FIG. 1. An illustration of episodic selection profiles at a single site with three possible regimes: negative, neutral (or nearly neutral), and diversifying selection along a branch. Panel (*A*) depicts the phylogeny used for discussion in the text and to carry out robustness simulations; Branch 5 is designated as foreground (FG), and the remaining four branches as background (BG). Panel (*B*) illustrates the four a priori selective profiles allowed by the model of Zhang et al. (2005). Panel (*C*) shows 2 of 239 possible selective profiles not modeled by current branch-site models; these profiles are used in robustness simulations (see Methods).

neutral evolution, and diversifying selection, respectively. Assuming these three ω ratios (fig. 1) with no further restrictions, each site can follow one of 3^B possible selective profiles—the number of different ways to assign the B branches to the three different selection rate bins. However, it is unclear how to determine which of these selective profiles or, equivalently, assignments of branches to selection rate bins is the most appropriate at a given site.

One approach (Yang and Nielsen 2002) is to model each site using only four predefined profiles regardless of the size of the phylogeny. More specifically, 1) every branch belongs either to the a priori known foreground class, which is allowed to experience diversifying selection, or the "background" class, which evolves under purifying selection or neutrally and 2) at a given site, there is no variation in selection strength (ω) among background branches, with all foreground branches either sharing the selection strength of the background or being under shared diversifying selection (fig. 1*B*). Clearly, these options are not exhaustive: For example, neither variable strength of selection among background or foreground nor positive selection along background branches is allowed. We refer to this approach as the restricted branch-site (rBS) model because the number of selective profiles is limited to the four a priori defined scenarios. Given a 4-taxon tree (fig. 1), and three selection parameters (as in fig. 1*B*), there are $3^5 = 243$ possible selection configurations, only four of which are accounted for by the branch-site model. The number of ω configurations grows as K^B, where K is the number of rate classes, thus making it unlikely that any four selection profiles chosen a priori are going to be sufficiently representative. Because there are no compelling biological reasons to expect that any two branches in the phylogenetic tree will have the same ω at any given site, we do not expect these four predefined selective profiles to provide an adequate description of complex biological data. This model was likely motivated by the need to avoid overfitting in the case of small sample sizes; however, we argue that if branches with differing selective pressures are incorrectly assigned to the same class, likelihood ratio test (LRT)-based branch-site methods can be positively misleading. In this manuscript, we present one case where they falsely identify positive selection on a neutrally evolving lineage (Type I or false positive error), and another where they fail to detect positive selection on a lineage with $\omega > 1$ (Type II or false negative error). In addition, if several branches are claimed to be under positive selection by setting the foreground to one branch at a time, as is done by the sequential testing procedure of Anisimova and Yang (2007), this creates a logical inconsistency—when a branch is found to be under selection, the model under which this was established implies that no other branch could be under positive selection.

We introduce a new class of models in which substitution rates may vary from branch to branch and from site to site. We incorporate this variation via "random effects"—unobserved strengths of selection at sites and branches are incorporated using a discrete or a discretized parametric probability distribution. Parameters defining the distribution are estimated jointly from all sites using maximum likelihood. Random effects likelihood (REL) and complementary fixed effects likelihood (FEL) models are standard tools in statistical modeling. Both types of model have been used to allow sitewise rate variation in phylogenetic models—see Kosakovsky Pond and Frost (2005b) for an overview. Nucleotide REL models were first introduced in Yang (1994), where rates over sites in a nucleotide alignment followed a discretized unit-mean gamma distribution (the now ubiquitous $+\Gamma_4$ model). Nielsen and Yang (1998) and Yang et al. (2000) applied REL models to codon data in order to identify signatures of natural selection, whereas Kosakovsky Pond and Frost (2005b) and Massingham and Goldman (2005) used FEL models for the same purpose. For all these models, likelihoods of individual sites are computed by Felsenstein's pruning algorithm (Felsenstein 1981). However, as we show later, the direct application of the pruning algorithm is intractable for REL models with branchwise as well as sitewise rate variation. It is presumably for this reason that, to date, branch models (Yang 1998; Kosakovsky Pond and Frost 2005a) have only been implemented in the FEL framework and branch-site models only as a four-category sitewise REL model. Our solution involves a simple extension of the pruning algorithm which makes it feasible to implement not only the model proposed here but also several other branch-site REL models.

The extended pruning algorithm computes the likelihood of each site, treating the selection site profile Ω_s as an unobserved variable, under the assumption that the probability of observing a substitution rate at a branch is independent of all other branches. Computationally, our algorithm is equivalent to replacing the standard Markov evolutionary model at a single phylogenetic branch with a mixture of three Markov models (one each for ω^-, ω^N, and ω^+), where the mixing coefficients and ω rates are inferred for each branch along with branch lengths, nucleotide substitution biases, and other alignment-wide parameters. Just like existing branch-site methods (Anisimova and Yang 2007), we use sequential likelihood ratio testing to identify which branches support a model with episodic diversifying selection. Unlike existing methods, however, our approach is unrestricted and considers every possible site profile, thus avoiding some of the prominent issues posed by model misspecification and further allows ω rates to vary independently from branch to branch and site to site.

Using an extensive collection of simulated sequences from Anisimova and Yang (2007), we perform a direct comparison of the unrestricted branch-site (uBS) model with the existing, restricted, approach (rBS) to evaluate Type I error and power. We also reinvestigate three empirical data sets that had been previously analyzed with the standard or sequential branch-site method and discover that many, but not all, of the original inferences are supported by our mixture model. Lastly, we report selective episodes not previously detected.

Methods

Codon Model Specification

To facilitate our presentation of episodic selection methods, we first briefly review maximum likelihood codon phylogenetic models (although see Delport et al. 2009 and Anisimova and Kosiol 2009 for detailed reviews). These models assume that substitutions along a branch of a phylogenetic tree can be described by an appropriately parameterized continuous-time stationary Markov process, defined by its instantaneous rate matrix, Q, with elements that describe the rate of substitution of codon i with codon j:

$$q_{ij} = \begin{cases} r(A_i, A_j)\theta_{ij}\pi_{ij}, & \delta(i,j) = 1, \\ 0, & \delta(i,j) > 1, \\ -\sum\limits_{k \neq i} q_{ik}, & i = j. \end{cases} \quad (1)$$

Here, $\delta(i,j)$ is the number of nucleotide differences between codons i and j, π_{ij} denote the equilibrium frequency parameters (e.g., $\pi_{AAA,AAC} = q_C^3, \pi_{ACC,AAC} = q_A^2$), θ_{ij} are the nucleotide mutational biases, and $r(A_i, A_j) = r(A_j, A_i)$ are the relative substitution rates between amino acids encoded by codons i and j. In the most general model, each of these $r(A_i, A_j)$'s can be independently estimated (see Delport et al. 2010), but here we follow the common approach of allowing only two rates: α for synonymous ($A_i = A_j$) and β for nonsynonymous ($A_i \neq A_j$) substitutions. Their ratio, β/α, is the familiar selection parameter, ω. The equilibrium frequency parameters may be estimated empirically either as the product of position-specific nucleotide frequencies (Goldman and Yang 1994) or as the position-specific frequency of the target nucleotide (Muse and Gaut 1994). Because we have previously identified biases using such empirical approaches (Kosakovsky Pond et al. 2010), we use corrected estimates (CF3 × 4) of nucleotide frequency parameters. Given a phylogenetic tree T (fig. 1), with B branches and branch lengths $t_i, i = 1, \ldots, B$, the likelihood of changing from state i to j at a site along branch b in time t_b is given by the (i,j) element of the transition matrix $P_Q(t_b) = e^{Qt_b}$. Subsequently, the likelihood of observing the alignment is evaluated as the product of site-likelihoods (with sites ranging from 1 to the number S of sites in the alignment), each of which is calculated using the standard pruning algorithm (Felsenstein 1981) given the data, a phylogenetic tree, T, and instantaneous rate matrix, Q.

Sitewise REL Models

Before extending Felsenstein's pruning algorithm, we first summarize how it is used in the context of the commonly used class of sitewise REL models. We pick our notation to allow extension to other types of REL models in the sections that follow. Throughout, we consider only the case of a finite number of discrete categories; extension to continuous-valued unobserved variables is straightforward, but computationally impractical, at least in the standard frequentist phylogenetic framework.

In a sitewise REL model, we think of each site as belonging to a site category, with the possible site categories ranging from 1 to K. For notational convenience, we present the special case where the categories differ only in terms of their ω values—allowing us to denote the category for site s by ω_s. Considering all sites simultaneously, the configuration of categories over all sites is a vector $\Omega_{\forall b} = (\omega_1, \ldots, \omega_S)$, where the subscript makes it explicit that this configuration is shared by all branches. We model the joint probability of the configuration as the product of independent factors:

$$P(\Omega_{\forall b}) = \prod_{s=1}^{S} P(\omega_s). \quad (2)$$

The individual category probabilities $P(\omega_s)$ are shared across all sites. Although the independence of sites is a standard assumption in the literature and allows for a particularly efficient likelihood calculation, it is not necessary. For example, $P(\Omega_{\forall b})$ has been modeled as a Hidden Markov process to permit spatial correlations among site categories (Felsenstein and Churchill 1996).

Another alternative to the model assumption of equation (2) would have been to allow only a small number of configurations. For example, we could imagine a model where sites are divided a priori into "buried" and "exposed" residues (e.g., Yang and Swanson 2002) and propose the following four configurations: 1) all sites conserved; 2) all sites evolving neutrally; 3) buried sites conserved and exposed sites under positive selection; and 4) buried sites evolving neutrally and exposed sites under positive selection. One could calculate the alignment-wide likelihood under each configuration and infer which of the configurations fits the data best. We mention this not because we think it is a good model (surely, it would not be biologically realistic to assume such a limited number of possible configurations) but because it is directly analogous to the existing branch-site model of Zhang et al. (2005). Our contribution in this manuscript is to upgrade from a branch-site model with four prechosen configurations such as these to one that is analogous to a REL model where the categories of different sites are independent.

Returning to standard sitewise REL models, the likelihood of the data D_s observed at site s (conditioned implicitly on non-ω model parameters) is

$$P(D_s) = \sum_{\omega_s} P(\omega_s) P(D_s|\omega_s) \quad (3)$$

$$= \sum_{\omega_s} P(\omega_s) \sum_{A} P(D_s, A|\omega_s), \quad (4)$$

where the first sum is over all site categories, A denotes a vector of ancestral node states, and the sum over A is taken over all possible such vectors. Labeling each nonroot node with the number of its parental branch, and the root node as 0, we can write this out more fully using

$$P(D_s, A|\omega_s) = P(A_0) \prod_{b=1}^{B} P(A_b|A_{\mathrm{pa}(b)}, \omega_s, t_b), \quad (5)$$

where A_b denotes the state at node b and pa(b) is the parent node of b. The task of Felsenstein's pruning algorithm is to calculate the sum

$$P(D_s|\omega_s) = \sum_{A_0} \sum_{A_1} \cdots \sum_{A_B} P(D_s, A|\omega_s), \quad (6)$$

which, because each of the terms $P(A_b|A_{\mathrm{pa}(b)}, \omega_s, t_b)$ in equation (5) depends only on a local part of the tree (a child and parent node and the branch connecting them), can be factorized efficiently and calculated by means of a postorder tree traversal. In what follows, we retain this property so that the same tree traversal remains an efficient way to calculate the desired likelihood.

Branch-Site REL Models

To define a branch-site REL model, we replace our sitewise category variable ω_s with a branch-site category variable ω_{bs}. Each branch-site combination is considered to belong to one of our K categories. We still aim to calculate the likelihood for a single site s, so we consider the configuration $\Omega_s = (\omega_{1s}, \ldots, \omega_{Bs})$ of branch categories. Our new approach is based on the observation that if the branch categories are independent, so that

$$P(\Omega_s) = \prod_{b=1}^{B} P(\omega_{bs}), \quad (7)$$

then the likelihood at a site can be computed efficiently without the need to apply the pruning algorithm for every possible value of Ω_s. By definition,

$$P(D_s) = \sum_{\Omega_s} P(\Omega_s) P(D_s|\Omega_s) \quad (8)$$

$$= \sum_{\Omega_s} \prod_{b=1}^{B} P(\omega_{bs}) \sum_{A} P(D_s, A|\Omega_s). \quad (9)$$

Changing the order of summations, this can be written as follows:

$$P(D_s) = \sum_{A} P(A_0) \sum_{\omega_{1s}} \sum_{\omega_{2s}} \cdots \sum_{\omega_{Bs}} \prod_{b=1}^{B} P(\omega_{bs}) P(A_b|A_{\mathrm{pa}(b)}, \omega_{bs}, t_b). \quad (10)$$

This is identical to the quantity calculated by Felsenstein's algorithm except for the presence of the $P(\omega_{bs})$ terms and the summations over ω values. Thinking algorithmically, and as indicated in equation (10), the entire space of K^B values of Ω_s can be traversed by B nested loops, where the outermost loop iterates over ω_{1s}, the second loop over ω_{2s} etc. Note that each product term $P(\omega_{bs})P(A_b|A_{\mathrm{pa}(b)}, \omega_{bs}, t_b)$ depends on only one branch. Hence, the sum computed by B nested loops (O(K^B) operations) is equivalent to a product of B sums ($O(KB)$ operations):

$$\sum_{\omega_{1s}} \sum_{\omega_{2s}} \cdots \sum_{\omega_{Bs}} \prod_{b=1}^{B} P(\omega_{bs}) P(A_b|A_{\mathrm{pa}(b)}, \omega_{bs}, t_b)$$

$$= \prod_{b=1}^{B} \sum_{\omega_{bs}=1}^{K} P(\omega_{bs}) P(A_b|A_{\mathrm{pa}(b)}, \omega_{bs}, t_b).$$

Consequently, we can rewrite equation (10):

$$P(D_s) = \sum_{A} P(A_0) \prod_{b=1}^{B} \left[\sum_{\omega_{bs}=1}^{K} P(\omega_{bs}) P(A_b|A_{\mathrm{pa}(b)}, \omega_{bs}, t_b) \right]. \quad (11)$$

The summation in parentheses can be viewed as the transition probability matrix of a mixture of K Markov substitution models, with $P(A_b|A_{\mathrm{pa}(b)}, \omega_{bs}, t_b)$ being the model-specific likelihoods at branch b, and $P(\omega_b)$ being the mixing proportions. If $Q_{\omega_{bs}}$ is the rate matrix associated with ω_{bs} (as in equation (1)), then this transition probability matrix can be computed as

$$p^{bs}(t) = \sum_{\omega_b=1}^{K} P(\omega_{bs}) e^{Q_{\omega_{bs}} t}. \quad (12)$$

The sum over A in equation (11) can be carried out efficiently using Felsenstein's pruning algorithm, with the transition matrices along each branch defined as K-process mixtures as above. In other words, in order to compute the likelihood of an alignment site, we first assume that the probability of a particular selective regime at a branch is independent of that at any other branch, and apply the pruning algorithm as usual, except that the substitution model along each branch is given as the mixture of equation (12).

Depending on how the mixing coefficients and the transition matrices in equation (12) are parameterized, we can obtain different types of branch-site models. In principle, for every branch-site combination (b, s), there could be K independently estimated mixing proportions $P(\omega_{bs})$ and selection parameters ω_{bs}. However, this approach will yield a model with considerably more parameters than observations. Three simpler model types appear promising.

Nonspecific Branch-Site REL.
ω_{bs} and $P(\omega_{bs})$ for each category K are shared by all branches and sites. There are K alignment-wide ω parameters (Ω_k), and the probability that $P(\omega_{bs} = \Omega_k) = q_k$ is described by an alignment-wide frequency parameter $q_k, \sum_k q_k = 1$. This is a simple model with $2K - 1$ parameters estimated from the entire alignment but may not incorporate enough biological realism. We used it as the first step of the optimization process for our more complex model to obtain initial parameter estimates.

Site-Specific Branch-Site REL.
$P(\omega_{bs})$ is a function of s, that is, every site (or more precisely site pattern) has its own set of mixing coefficients, shared across all branches. ω_{bs} are shared by all sites and branches. This model has $KS + K - S$ parameters: $K\Omega_k$ parameters estimated jointly from the alignment and S sets of q_{sk} mixing parameters, with $\sum_k q_{sk} = 1, \forall s = 1, \ldots, S$, so that $P(\omega_{bs} = \Omega_k) = q_{sk}$. Because the number of parameters grows with the size of the alignment, the model will be asymptotically ill behaved. However, for fixed length alignments with many sequences, it may be possible to learn site-specific mixing parameters reliably.

Branch-Specific Branch-Site REL.
ω_{bs} and $P(\omega_{bs})$ are functions of b, that is, every branch has its own set of model parameters (ω_b^k) and mixing coefficients ($q_b^k, \sum_k q_b^k = 1$), but they are estimated jointly from all sites. This model has $(2K - 1)B$ parameters and is investigated in the present manuscript. It has the attractive property that the model parameters we learn include, for every branch, the proportion of sites belonging to every selection category.

A New Test for Episodic Selection

We define and fit a branch-specific branch-site REL model (termed unrestricted branch site or uBS). For consistency with several existing REL models, we restrict ω at every branch to take on one of $K = 3$ values $\omega_b^- \leqslant \omega_b^N \leqslant 1 \leqslant \omega_b^+$, representative of strong and weak conservation and positive diversifying selection. In our experience (e.g., see Kosakovsky Pond et al. 2010), models that permit multiple classes of sites with $\omega < 1$ fit protein-coding sequence alignments much better than those with one of the ω values fixed at 1. We denote their mixing proportions q_b^-, q_b^N, and q_b^+ (subject to $q_b^- + q_b^N + q_b^+ = 1$), respectively. All model parameters are estimated by maximum likelihood. Next, we fit B models (one for each branch), where model $b = 1, \ldots, B$ differs from the unrestricted model by the additional constraint of $\omega_b^+ = 1$. Each of these models, therefore, disallows diversifying selection along a single branch while leaving all other background branches unrestricted. Compare this with the requirement that all background branches have uniform neutral or negative selection regimes in the standard branch-site model (Zhang et al. 2005). As described most recently in Anisimova and Yang (2007), the evidence for positive selection along branch b can be evaluated by a LRT using the asymptotic distribution of the LR statistic defined by $(\chi_1^2 + \chi_0^2)/2$ (Self and Liang 1987). If B branches are tested in sequence, it is necessary to correct the nominal significance level for each individual test to control the cumulative (or family wise) error rate of the tests. Anisimova and Yang (2007) compared multiple such corrections in the context of branch-site methods and reported that their performance was broadly similar. With that in mind, we settled on the correction procedure due to Holm (1979), which is more powerful and as easy to compute as the simple Bonferroni correction. Briefly, if the desired Type I error for the event "any of the B tests is a false positive under the null model" is α, then the testing procedure first ranks p values for each individual test in increasing order $p^{(1)} \leqslant p^{(2)} \leqslant \cdots \leqslant p^{(B)}$ and rejects first k hypotheses if $p^{(i)} \leqslant \alpha/(B - i + 1)$ for $i = 1, \ldots, k$ and $p^{(k+1)} > \alpha/(B - k)$. Our testing procedure uses a single alternative hypothesis and requires that $B + 1$ model fits be performed, whereas the testing procedure of Anisimova and Yang (2007) demands the fitting of $2B$ models because a different null and alternative pair must be evaluated for each branch.

Evaluating the Robustness of the rBS Model

We simulated data according to two selection scenarios along a 4-taxon tree (fig. 1A) using the codon substitution model defined above, with equal codon equilibrium frequencies ($\pi = 1/61$) and the HKY85 (Hasegawa et al. 1985) nucleotide substitution biases (i.e., $\theta_{ac} = \theta_{at} = \theta_{cg} = \theta_{gt} = 2; \theta_{ag} = \theta_{ct} = 1$). This choice of base frequencies and nucleotide substitution biases will deemphasize the differences in how frequency parameters and nucleotide substitution biases are modeled in rBS and uBS.

First (robustness simulation 1, RS1), we designated branch 5 (fig. 1 A) as a neutrally evolving foreground, that is, the one to be tested for episodic diversifying selection by the models), branch ($\omega = 1$), whereas background branches 1 and 3 were simulated under strong diversifying selection ($\omega = 10$), and background branches 2 and 4—under strong purifying selection ($\omega = 0.1$). This scenario was crafted to include variable selection along background branches which is not handled by any of the four classes of the branch-site model, and hence the standard branch-site test of selection along branch 5 will be fitting the data using two incorrect models. Second (RS2), we designated branch 5 as a positively selected foreground branch ($\omega = 2$), whereas background branches 1 and 2 are under strong diversifying selection ($\omega = 10$) and background branches 3 and 4 are under strong purifying selection ($\omega = 0.05$). These two scenarios are designed to explore the asymptotic behavior of the tests and use sequences longer than most genes. A test with poor asymptotic properties when a specific model assumption is violated may appear to behave acceptably on smaller samples due to, for example, lack of power. If test errors increase with sample size, this may point to fundamental issues with the approach.

Evaluating the Performance of the Unrestricted Branch-Site Model

Anisimova and Yang (2007) generated several thousand alignments under seven selective regimes, three of which included no positive selection (to test for Type I error or false positives) and four included varying extents of diversifying selective pressure (to assess Type II error or power). These simulation alignments were kindly provided by the authors, and we reanalyzed the data for a direct comparison with our approach. For complete details on these simulations, we refer the reader to table 2 and text in Anisimova and Yang (2007). Briefly, either 4 or 8 taxon balanced trees were used for simulations, with 1,000 (4 taxa) or 200 (8 taxa) 300-codon long replicates/scenario.

In addition, we test our approach in a high information content setting, using sequences with 1,000 codons simulated along a 16-taxon balanced tree (supplementary fig. S3, Supplementary Material online). We subdivide the length of the sequence into three partitions, such that a site is simulated under one of three potential selection models. The first two models are homogeneous with respect to the tree and encompass purifying selection ($\omega = 0.1$) and neutrality ($\omega = 1$) with proportions, $p_1 = 0.8$ and $p_2 = 0.05$, respectively. Finally, the third model, with proportion $p_3 = 0.15$, is heterogeneous with respect to the tree, comprising neutral evolution ($\omega = 1$) at all branches, except a set of three branches at which strong diversifying selection is simulated ($\omega = 5$). We considered two modifications of this scenario: a lower proportion of selected sites ($p_2 = 0.15, p_3 = 0.05$) or weaker selection ($\omega = 2$ in the third model).

Finally, we reexamine three empirical alignments previously analyzed for evidence of episodic selection: a data set consisting of 19 lysozyme *c* sequences ($S = 130$ codons) from primates, initially analyzed by Messier and Stewart (1997); CD2 gene sequences ($S = 187$ codons) coding for a cell adhesion molecule located on the surface of certain type of lymphocyte, isolated from 10 mammalian species and originally analyzed by Lynn et al. (2005); and 10 mammalian sequences ($S = 1,162$ codons) of the tumor suppressor gene BRCA1 (Zhang et al. 2005).

Implementation

The model is implemented as a collection of HyPhy (Kosakovsky Pond et al. 2005) Batch Language scripts and is distributed as a part of HyPhy v2.0020110306 or later as *BranchSiteREL.bf* file in the *Positive Selection* rubrik of standard analyses.

Results

Test Performance on Simulated Data

We applied our uBS sequential selection test to parametric replicates generated under seven different selection profiles previously used by Anisimova and Yang (2007) to evaluate the original sequential branch-site test for detecting episodic selection (Zhang et al. 2005) and to two additional sets robustness simulations. Details of simulation results are collated in table 1.

1. When sequences are simulated under rBS assumptions (fig. 1), that is, those which conform to the null or the alternative model of Zhang et al. (2005), both uBS and rBS perform comparably (NC1, NC2, and SC in table 1), with similar family wise error rates (FWER) and power. It is encouraging that our unrestricted method does not appear to be strongly underpowered compared with rBS, even when the data are simulated to favor the former (38% vs. 44% power on SC with one sequence). The same holds for data generated under models which deviate from rBS assumptions but not too strongly (NI, SI1 in table 1).
2. The advantages of uBS over rBS become apparent when the assumptions of the latter are inappropriate for the data (SI2 and SI3). Already, in the SI2 scenario, where two branches are experiencing episodic diversifying selection, uBS provides a considerable boost in power for 8-taxon trees (63% vs. 48.5%). The greatest difference between our approach and rBS is revealed in the SI3 simulation scenario, when four background branches in a 4-taxon tree were simulated under episodic selection, whereas the single foreground branch was evolved neutrally or under purifying selection. The intent of SI3 in Anisimova and Yang (2007) was to violate the assumptions of the rBS model as much as conceivably possible and investigate how this would reflect on Type I errors. Although the rBS model controlled the rates of false positives (FWER 1.7%), it suffered a severe loss of power—the cumulative power was reported at only 35.3%, despite pervasive episodic selection in this case. In contrast, uBS achieved 92.5% power while maintaining FWER of 6.0%.
3. Given sufficient deviations from modeling assumptions (RS1, RS2 in table 1), rBS tests for selection on foreground branches can be severely misleading. For RS1, the null model ($\omega_2 = 1$) is rejected in favor the alternative model ($\omega_2 \geqslant 1$), implying positive selection along the neutral lineage five with frequencies much higher than the nominal error rate of the tests, and a very skewed distribution of the p-values (supplementary fig. S1, Supplementary Material online). The null hypothesis rejection rate increases as the length (S codons) of the alignment is increased. For example, at test $p = 0.05$, the null model was rejected 12/100 times for $S = 1,000$, 31/100 times for $S = 2,000$, 74/100 times for $S = 5,000$, and in 97/100 cases for $S = 10,000$. Nominal p-values are commonly interpreted as the acceptable rate of false positives of the test, hence $p = 0.05$ should result in about 5/100 false rejections of the null. Lowering $p = 10^{-4}$ still yields 34/100 false positives for $S = 10,000$, suggesting

Table 1. uBS Performance on Simulated Data.

Simulation Scenario	Sequences/Codons	Branch 1	Branch 2	Branch 3	Branch 4	Branch 5	FWER		Power	
							rBS	uBS	rBS	uBS
NC1	4	0.008	0.006	0.01	0.007	0.005	0.043	0.036	—	—
	8	0.005	0.005	0.005	0.015	0.00	0.044	0.03	—	—
NC2	4	0.014	0.01	0.016	0.007	0.07	0.053	0.053	—	—
	8	0.005	0.015	0.01	0.005	0.000	0.045	0.035	—	—
NI	4	0.006	0.012	0.009	0.001	0.005	0.051	0.033	—	—
	8	0.03	0.025	0.01	0.005	0.005	0.08	0.07	—	—
SC	4	0.005	0.008	0.004	0.004	*0.101*	0.026	0.02	0.084	0.101
	8	0.015	0.015	0.000	0.005	*0.38*	0.045	0.035	0.44	0.38
SI1	4	0.007	0.007	0.005	0.007	*0.103*	0.033	0.025	0.082	0.103
	8	0.00	0.015	0.005	0.015	*0.435*	0.06	0.035	0.495	0.435
SI2	4	*0.116*	0.004	0.008	0.009	*0.07*	0.033	0.021	0.166	0.176
	8	*0.53*	0.01	0.01	0.00	*0.195*	0.02	0.02	0.485	0.630
SI3	4	*0.295*	*0.484*	*0.599*	*0.667*	0.06	0.017	0.06	0.353	0.925
RS1	1,000	*1*	0.01	*1*	0.00	0.00	0.12	0.01	1.00	1.00
RS1	2,000	*1*	0.00	*1*	0.00	0.08	0.31	0.08	1.00	1.00
RS1	5,000	*1*	0.01	*1*	0.00	0.03	0.74	0.03	1.00	1.00
RS1	10,000	*1*	0.00	*1*	0.01	0.03	0.97	0.03	1.00	1.00
RS2	1,000	*1*	*1*	0.00	0.00	*0.44/0.03**	0.00	0.00	1.00	1.00
RS2	2,000	*1*	*1*	0.00	0.00	*0.83/0.02**	0.00	0.00	1.00	1.00
RS2	5,000	*1*	*1*	0.00	0.00	*0.98/0.03**	0.00	0.00	1.00	1.00
RS2	10,000	*1*	*1*	0.00	0.00	*1.00/0.05**	0.00	0.00	1.00	1.00

RS1 and RS2 are described in the text and figure 1. Simulations NC1, NC2, NI, SC, SI1, SI2, and SI3 are taken from Anisimova and Yang (2007) (see table 2 therein for complete details of simulation parameters). The first three simulations (NC1, NC2, and NI) do not include any lineages under positive selection, whereas the last four include one or more lineages under selection at some sites in the alignment. Branches that experience positive selection are typeset in italic. Entries for Branch 1–Branch 5 columns show the proportion of replicates where any branch from this class was found to be under positive selection at $p \leqslant 0.05$. FWER is the proportion of replicates where at least one branch was falsely classified as undergoing positive selection. The Power column lists the proportion of replicates for which *at least one branch* under positive selection was correctly classified as such. *: the second number reports the proportion of replicates where Branch 5 was reported under positive selection by rBS.

that the rate of false positives is difficult to control. The estimate of ω along lineage 5 is biased, with mean $\hat{\omega} \approx 1.4$ and variance inversely proportional to sample size. On the same data, uBS had well-controlled rates of false positives, which did not correlate with the length of the alignments. For RS2, the rBS test now performs as if the null model ($\omega = 1$ on branch 5) were correct—the rate of rejections is similar to the rate expected under the null model and the ω_2 estimate is now biased downward to $\omega_2 \approx 1.0$ and very low power (2–5%) to detect selection along branch 5 (table 1). We observed shrinking estimator variances for larger sample sizes (fig. S2), showing that the lack of power is not due to insufficient sample sizes. In contrast, uBS showed very low rates of false positives on the negatively selected branches (0%) and power ranging from 44% ($S = 1,000$) to 100% ($S = 10,000$) on the interior branch of the tree simulated to be under diversifying selection.

Test Performance as a Function the Strength and Extent of Episodic Selection

For the 16-taxon tree and 1,000-codon long sequences with lineages A, B, and AB (supplementary fig. S3, Supplementary Material online) are under positive diversifying selection, we observed the following test performance.

15% of Sites under Selection with $\omega = 5$.
uBS achieved 100% power and FWER of 2%, demonstrating that larger and more informative alignments allow the test to be more discriminative and accurate, as expected. For the same data set, rBS was surprisingly conservative with 0% FWER, but only 6% power.

5% of Sites under Selection with $\omega = 5$.
uBS achieved only 9% power at FWER of 2%, demonstrating that if too few sites are under selection, the ability of the test to detect episodic selection is severely impacted.

15% of Sites under Selection with $\omega = 2$.
uBS attained 8% power at FWER of 3%, indicating that a weak selection signal is considerably more difficult to identify.

Empirical Data Applications

First, we analyzed CD2 gene sequences coding for a cell adhesion molecule located on the surface of certain types of lymphocytes. These sequences were isolated from ten mammalian species and were previously analyzed by Lynn et al. (2005) using a branch (no site-to-site variation) method (Yang 1998) and more recently by Anisimova and Yang (2007) with a branch-site method. Lynn and colleagues found that lineages leading to pig, cow, horse, cat, the (pig and cow) ancestor (lineage 3 in fig. 2A), and the primate clade ancestral lineage (13) were under positive selection because the mean point estimate of ω at those branches exceeded one and the branch heterogeneity test (Yang 1998) rejected the hypothesis that all lineages were under the same selective pressure. Anisimova and Yang (2007) identified positive selection along lineages leading to cow, cat, and the ancestor of (pig, cow, horse, and cat) clade using a sequential rBS test; and pointed out that comparing

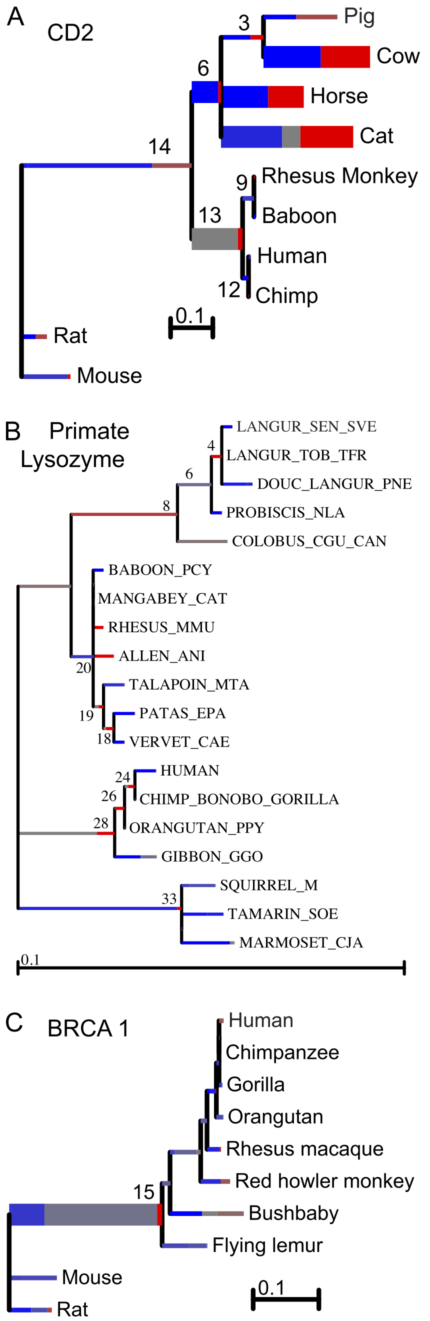

FIG. 2. Empirical data sets analyzed for episodic selection. Each tree is scaled on the expected number of substitutions/nucleotide. The hue of each color indicates strength of selection, with primary red corresponding to $\omega > 5$, primary blue to $\omega = 0$, and grey to $\omega = 1$. The width of each color component represents the proportion of sites in the corresponding class. Thicker branches have been classified as undergoing episodic diversifying selection by the sequential test at $p \leqslant 0.05$.

the value of point estimate of ω to 1 was only suitable for exploratory analyses and did not constitute a valid statistical test. Our uBS model confirms (at $p \leqslant 0.05$) episodic selection along the same three lineages reported by Anisimova and Yang (2007) but also identifies two additional lineages—the horse lineage and the most recent common ancestor of the primate clade. Neither of these lineages approached significance in the analysis of Anisimova and Yang, but because CD2 appears to have undergone extensive episodic selection at multiple lineages, the assumptions of the rBS test are likely to be violated in these data, for example, leading to loss of power by rBS (as was shown in SI3 simulations). The patterns of episodic selection were complex (fig. 2A and table 2), with marked differences in the extent (proportion) and strength (ω^+) of selection along different lineages. Interestingly, Branches 6 (not reported by Lynn et al. 2005) and 13 (not reported by Anisimova and Yang 2007) appear to experience very strong selective forces ($\omega_6^+ = 37.2, \omega_{13}^+ = 39.7$) on a small percentage of sites ($q_6^+ = 0.094, q_{13}^+ = 0.092$), whereas the other three selected branches (cow, horse, and cat) each have approximately 40% of sites under relatively weaker positive selection ($\omega = 5.2$–10.7).

Next, we reexamined a data set consisting of 19 lysozyme *c* sequences from primates initially analyzed by Messier and Stewart (1997) and more recently by Zhang et al. (2005). The authors suspected positive selection along the lineage leading to the colobine monkeys and hominoids for which the lysozyme protein may have acquired a different digestive function that allows them to lyse symbiotic bacteria. Yang (1998) confirmed positive selection along the hominoid lineage (and elevated ω compared with background on the colobine lineage) using codon models that permitted no site-to-site rate variation. Indeed, it appears that if one assumes negative or neutral selection elsewhere on the phylogeny, the "average" strength of selection along the lineages of interest exceeds or approaches one. It was therefore somewhat unexpected that more sensitive rBS models did not find evidence of episodic diversifying selection along the two lineages (Zhang et al. 2005). uBS reached the same conclusion—no single lineage had sufficient statistical support for episodic diversifying selection under a sequential (branch at a time) test. The inferred selective mixture for the hominoid ancestral lineages (28 in fig. 2*B*) showed 18.2% of sites under very strong selection $\omega > 100$ and an uncorrected p-value of 0.008, that is, were we to test only for selection only along this lineage based on a priori information, we would find episodic diversifying selection at $p < 0.05$. For the colobine ancestral lineage (8 in fig. 2*B*), 100% of sites were allocated to the positive selection regime ($\omega = 3.4$), yet the test p-value was only 0.10.

The last data set we analyzed contains ten mammalian sequences of the tumor suppressor gene BRCA1. Zhang et al. (2005) previously analyzed eight of these sequences as the chimpanzee and human lineages are suspected to be under positive selection but found no evidence of positive selection along any lineages. Our sequential analysis found evidence of episodic diversifying selection on the lineage ancestral to primates and lemurs (Branch 15 in fig. 2C) with 3.3% of sites in the $\omega^+ = 17.3$ class. The human lineage shows borderline (uncorrected) significance with $p = 0.076$ (all sites under weaker positive selection, $\omega = 2.26$), whereas the chimpanzee lineage is not significant (uncorrected

Table 2. uBS on the CD2 Data Set.

Branch	Mean ω	ω^-	q^-	ω^N	q^N	ω^+	q^+	LRT	p	Corrected p
Pig	1.341	0.000	0.443	0.919	0.000	2.811	0.557	3.276	0.035	0.352
Cow	1.914	0.000	0.025	0.000	0.513	10.732	0.462	23.465	0.000	0.000
3	1.480	0.000	0.328	0.000	0.370	7.824	0.303	5.989	0.007	0.079
Horse	1.244	0.000	0.001	0.000	0.569	5.190	0.430	11.463	0.000	0.005
Cat	1.598	0.252	0.463	1.000	0.137	6.544	0.400	13.309	0.000	0.002
6	0.664	0.000	0.906	0.118	0.000	37.328	0.094	7.432	0.003	0.038
RHmonkey	22.503	1.000	0.007	1.000	0.316	113.398	0.677	1.196	0.137	0.822
Baboon	0.000	0.000	0.550	0.000	0.336	0.000	0.113	0.000	1.000	1.000
9	0.400	0.047	0.000	0.443	1.000	0.009	0.000	0.000	1.000	1.000
Human	0.002	0.126	0.468	0.215	0.384	2.963	0.148	0.000	0.500	1.000
Chimpanzee	24.634	0.313	0.000	0.812	0.000	47.512	1.000	0.630	0.214	1.000
12	0.368	0.000	0.149	0.000	0.803	12.624	0.048	1.393	0.119	0.952
13	1.915	1.000	0.020	1.000	0.888	39.772	0.092	8.823	0.001	0.019
14	0.432	0.156	0.039	0.162	0.730	2.581	0.232	1.315	0.126	0.880
Rat	1.093	0.000	0.552	0.002	0.000	2.998	0.448	0.367	0.272	1.000
Mouse	0.524	0.400	0.947	0.799	0.000	22.217	0.053	2.240	0.067	0.605

Mean ω is estimated under the free-ratio MG94 × REV model (no site-to-site rate variation). ω and q values reflect the branch-level mixture of negative, (nearly) neutral, and positive selection models. LRT: likelihood ratio test statistic, p: uncorrected p-value obtained using the mixture of χ_0^2 and χ_1^2 distributions; corrected p: after an application of Holm's multiple testing correction. Internal branches are numbered concordantly with figure 2. Branches found by uBS to be under positive diversifying selection are shown in italic.

$p = 0.16$). These findings are in qualitative agreement with previous analyses (Zhang et al. 2005).

Discussion

This work demonstrates that current branch-site methods can have excessive Type I and Type II errors when the data strongly deviate from model assumptions. These models enforce uniform selective pressure on all background branches, thus biasing the estimate of ω along foreground branches. We have demonstrated this behavior to be positively misleading, with decreasing variance for larger sample sizes. The nature of the bias will depend on the distribution of selective pressures along background branches, nucleotide substitution biases, and branch lengths. More critically, the sequential rBS approach (Anisimova and Yang 2007) to test each branch in a phylogeny for evidence of positive selection, while specifically postulating that no other branches in the phylogeny are subject to positive selection, is likely an oversimplification of biological reality. Furthermore, when one branch is found to be under selection by this method, it automatically implies that no other branch (in the background) can be under selection, hence the sequential testing procedure that finds multiple selected branches by setting the foreground to one branch at a time is logically inconsistent.

We have developed and validated a new random effects branch-site model (uBS) to detect positive selection in protein-coding sequences that do not require partitioning lineages into foreground and background branches. This model considers all possible assignments of three selective regimes to the branches in a phylogeny at a given site. If the selective behavior along a branch is independent of that along other branches, our model can be efficiently evaluated in the standard phylogenetic framework. This is accomplished by replacing the standard substitution model along a branch with a mixture of three Markov models: one for purifying, one for nearly neutral, and one for diversifying selection. To detect episodic diversifying selection, we adopt the familiar hypothesis testing framework (Anisimova and Yang 2007) to identify the lineages in a phylogeny that could have undergone episodic selection, and we measure the strength (ω) and extent (proportion of sites) of such selection independently (but jointly) for each branch. uBS is approximately twice as computationally efficient as the current branch-site approach because it tests a series of nulls (no positive selection on a given branch) versus a universal alternative (no constraints on any branches), whereas the sequential rBS approach constructs a separate null and alternative model for each branch. The new approach is more computationally attractive than the family of codon-based covarion models (Guindon et al. 2004), where the addition of each evolutionary modality incurs an expansion of the character state space and the corresponding quadratic-to-cubic (in terms the number of ω classes) increase in algorithmic complexity. However, some aspects of covarion models are more flexible, for example, the switchpoints in the evolutionary process are not delineated by branches in the tree as they are in uBS, hence the two approaches are complementary.

Because our testing procedure does not limit the number and type of site configurations at a site, we expect it to demonstrate improved performance on data that do not conform to the restrictive assumptions of the rBS model. Using the same set of simulations as in Anisimova and Yang (2007), we demonstrate that uBS has notably higher power and lower error rates than the sequential rBS method when the assumptions of the latter method are strongly violated (scenarios SI2 and SI3). Encouragingly, on the data that do meet rBS restrictions, our approach delivers comparable performance, suggesting that it is not necessary to make a priori assumptions about the patterns of episodic selection. uBS attains 100% power if sufficient data (e.g., 16 sequences, 1,000 codons, and 15% of sites under selection) are supplied. Our reanalysis of three benchmark biological data sets revealed slight differences from published results and confirmed the lower power of sequential rBS methods to detect short bursts of strong selection in a data set subject to pervasive episodic selection.

Much future work remains, however. First, there is no clear understanding of what extent and strength of selection, data sizes, and divergence levels are necessary for episodic selection tools to be appropriately powered, yet not subject to excessive false positive rates. Even based on our limited 16-taxon simulations, it is apparent that uBS rapidly loses power when the proportion of sites under selection is too small or when selective pressures are relaxed. Second, does the location of lineages under selection in the phylogeny (e.g., tips vs. deep internal branches) influence our ability to infer selection? Simulations in this study suggest that there may be more power to detect recent episodic selection at terminal branches, but a more systematic exploration is necessary. Third, how does one go about automatically pooling branches together to boost the power to detect weaker selection that affects the same set of sites in multiple lineages—a good example would be HIV evolution to independently acquire drug-resistance mutations in lineages that represent patients on treatment (Seoighe et al. 2007). Fourth, much of episodic selection is likely to be directional rather than diversifying, hence models must be adapted to include this type of selection as well (e.g., Delport et al. 2008; Kosakovsky Pond et al. 2008). Fifth, might it be beneficial to relax the assumption of constant synonymous rates (Kosakovsky Pond and Muse 2005)? Sixth, naive, or Bayes empirical Bayes approaches developed for rBS for detecting individual sites subject to episodic diversifying selection (Yang et al. 2005), need to be adapted to and evaluated in the context of uBS.

Based on the results, theoretical considerations and computational feasibility presented in this manuscript, we advocate our mixture approach over current tools for the detection of episodic diversifying selection (Anisimova and Yang 2007). Unlike Nozawa et al. (2009), who propounded a severely underpowered (and difficult to extend) counting method for lineage-specific selection detection and made a number of strong claims recently refuted by Yang and dos Reis (2011), we espouse the view that likelihood model-based approaches are a much more appealing way forward. We are convinced that continued improvements in biological realism of evolutionary models, underpinned by gains in computing power and algorithmic development, will provide evolutionary biologists with the tools to better characterize fundamental adaptive processes. uBS demonstrates the potential for continued extension of classical frequentist and hypothesis testing approaches to parallel recent seminal developments in Bayesian approaches to fitting complex substitution models (e.g., Rodrigue et al. 2010).

Supplementary Material

Supplementary figures S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This research was supported in part by the National Institutes of Health (AI43638, AI47745, and AI57167, GM093939), the University of California University-wide AIDS Research Program (IS02-SD-701), the Joint DMS/NIGMS Mathematical Biology Initiative through Grant NSF-0714991 and by a University of California, San Diego Center for AIDS Research/National Institute of Allergy and Infectious Diseases Developmental Award to S.D.W.F. and S.L.K.P (AI36214). S.D.W.F. is supported in part by a Royal Society Wolfson Research Merit Award. B.M. is supported by Europeaid grant number Sante/2007/174-790 from the European Commission. The authors are grateful to Dr Maria Anisimova for providing simulated sequence alignments used for benchmarking uBS and rBS.

References

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

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol*. 24:1219–1228.

Delport W, Scheffler K, Botha G, Gravenor MB, Muse SV, Kosakovsky Pond S. 2010. Codontest: modeling amino acid substitution preferences in coding sequences. *PLoS Comput Biol*. 19:e1000885.

Delport W, Scheffler K, Seoighe C. 2008. Frequent toggling between alternative amino acids is driven by selection in HIV-1. *PLoS Pathog*. 4:e1000242.

Delport W, Scheffler K, Seoighe C. 2009. Models of coding sequence evolution. *Brief Bioinform*. 10:97–109.

Felsenstein J. 1981. Evolutionary trees from DNA-sequences—a maximum-likelihood approach. *J Mol Evol*. 17:368–376.

Felsenstein J, Churchill GA. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol*. 13:93–104.

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

Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci U S A*. 101:12957–12962.

Hasegawa M, Kishino H, Yano TA. 1985. Dating of the human ape splitting by a molecular clock of mitochondrial-DNA. *J Mol Evol*. 22:160–174.

Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 6:65–70.

Kosakovsky Pond S, Delport W, Muse SV, Scheffler K. 2010. Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One* 30:e11230.

Kosakovsky Pond SL, Frost SDW. 2005a. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol Biol Evol*. 22:478–485.

Kosakovsky Pond SL, Frost SDW. 2005b. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*. 22:1208–1222.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. Hyphy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

Kosakovsky Pond SL, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*. 22:2375–2385.

Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW. 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol Biol Evol*. 25:1809–1824.

Kosakovsky Pond SL, Scheffler K, Gravenor MB, Poon AFY, Frost SDW. 2010. Evolutionary fingerprinting of genes. *Mol Biol Evol*. 27:520–536.

Lynn DJ, Freeman AR, Murray C, Bradley DG. 2005. A genomics approach to the detection of positive selection in cattle: adaptive evolution of the T-cell and natural killer cell-surface protein cd2. *Genetics* 170:1189–1196.

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

Messier W, Stewart CB. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* 385:151–154.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. 11: 715–724.

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

Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci U S A*. 106:6700–6705.

Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc Natl Acad Sci U S A*. 107: 4629–4634.

Self SG, Liang KY. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc*. 82:605–610.

Seoighe C, Ketwaroo F, Pillay V, et al. (11 co-authors). A model of directional selection applied to the evolution of drug resistance in HIV-1. *Mol Biol Evol*. 24:1025–1031.

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

Suzuki Y, 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*. 21:914–921.

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

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39:306–314.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*. 15:568–573.

Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol*. 28: 1217–1228.

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

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 22:1107–1118.

Yang ZH, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

Yang ZH, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol*. 19:49–57.

Zhang J. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol*. 21:1332–1339.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*. 22: 2472–2479.

Detecting Individual Sites Subject to Episodic Diversifying Selection

Ben Murrell[1,2], Joel O. Wertheim[3], Sasha Moola[2], Thomas Weighill[2], Konrad Scheffler[2,4], Sergei L. Kosakovsky Pond[4]*

1 Biomedical Informatics Research Division, eHealth Research and Innovation Platform, Medical Research Council, Tygerberg, South Africa, **2** Computer Science Division, Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa, **3** Department of Pathology, University of California San Diego, La Jolla, California, United States of America, **4** Department of Medicine, University of California San Diego, La Jolla, California, United States of America

Abstract

The imprint of natural selection on protein coding genes is often difficult to identify because selection is frequently transient or episodic, i.e. it affects only a subset of lineages. Existing computational techniques, which are designed to identify sites subject to pervasive selection, may fail to recognize sites where selection is episodic: a large proportion of positively selected sites. We present a mixed effects model of evolution (MEME) that is capable of identifying instances of both episodic and pervasive positive selection at the level of an individual site. Using empirical and simulated data, we demonstrate the superior performance of MEME over older models under a broad range of scenarios. We find that episodic selection is widespread and conclude that the number of sites experiencing positive selection may have been vastly underestimated.

Citation: Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, et al. (2012) Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genet 8(7): e1002764. doi:10.1371/journal.pgen.1002764

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, United States of America

Received January 17, 2012; **Accepted** April 26, 2012; **Published** July 12, 2012

Funding: This research was supported in part by the National Institutes of Health (AI47745 and AI57167, AI74621, GM093939); a Joint DMS/NIGMS Mathematical Biology Initiative through Grant NSF-0714991; a University of California San Diego Center for AIDS Research (UCSD CFAR)/NIAID Developmental Award to SLKP (AI36214); a National Institutes of Health Training Fellowship to JOW (AI43638); the Bioinformatics, Statistical Analysis, and Evolutionary Core of the UCSD CFAR; the National Research Foundation of South Africa; and Europeaid Grant number SANTE/2007/147-790 from the European Commission. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: spond@ucsd.edu

Introduction

Following the introduction of computationally tractable codon-substitution models [1,2] nearly two decades ago, there has been sustained interest in using these models to study the past action of natural selection on protein coding genes. Positive selection can be inferred whenever the estimated ratio (ω) of non-synonymous (β) to synonymous (α) substitution rates significantly exceeds one (reviewed in [3] and [4]). In the original models, the ω ratio was shared by all sites in an alignment, providing little power to detect the signature of positive selection. Indeed, even among classical examples of positively selected genes [5,6,7], most substitutions are expected to be neutral or deleterious [8]. Consequently, relatively few genes in which mean ω estimates are significantly greater than one are expected to exist, e.g. only 35/8079 were found in a human - chimpanzee genome-wide comparison [9].

Random effects codon-substitution models [10] permitted ω to vary from site to site, which made it possible to identify instances when positive selection had acted only upon a small proportion of sites. Such site-level models can detect which positions in a sequence alignment may have been influenced by diversifying positive selection, e.g. [11,12]. However, these models posit that diversifying selective pressure at each site remains constant throughout time, i.e. affects most lineages in the phylogenetic tree, (Figure 1A), and there are very few cases where this assumption is biologically justified (see [13,14,15,16] for examples of models that allow selection to vary throughout the tree). When a site evolves under purifying selection on most lineages, site methods which assume ω is constant over time may be unable to identify any episodic positive selection, since they will likely infer $\omega < 1$ [17]. It has been noted that positive selection is more readily identified in smaller alignments: counterintuitively, including additional sequences may cause sites to no longer be detected [18,19]. This phenomenon could be readily explained by purifying selection on some lineages masking the signal of positive selection on others.

We present a mixed effects model of evolution (MEME), based on the broad class of branch-site random effects phylogenetic methods recently developed by our group [20]. MEME allows the distribution of ω to vary from site to site (the fixed effect) and also from branch to branch at a site (the random effect, Figure 1B). Our approach provides a qualitative methodological advance over existing approaches which integrate site-to-site and lineage-to-lineage rate variation, e.g. the branch-site methods [17] or codon-based covarion models [13]. MEME can reliably capture the molecular footprints of both episodic and pervasive positive selection, a task for which current models are not well suited. Using empirical sequence data sets spanning diverse taxonomic categories and gene functions, along with comprehensive simulations, we demonstrate that MEME matches the performance of traditional site methods when natural selection is pervasive, and that MEME reliably identifies episodes of diversifying evolution affecting a small subset of branches at individual sites, where site methods often report purifying selection at the same site. For most

Author Summary

Identifying regions of protein coding genes that have undergone adaptive evolution is important to answering many questions in evolutionary biology and genetics. In order to tease out genetic evidence for natural selection, genes from a diverse array of taxa must be analyzed, only a subset of which may have undergone adaptive evolution; the same gene region may be under stabilizing or relaxed selection in lineages leading to other taxa. Most current computational methods designed to detect the imprint of natural selection at a site in a protein coding gene assume the strength and direction of natural selection is constant across all lineages. Here, we present a method to detect adaptive evolution, even when the selective forces are not constant across taxa. Using a variety of well-characterized genes, we find evidence suggesting that natural selection is generally episodic and that modeling it as such reveals that many more sites are subject to episodic positive selection than previously appreciated.

empirical data sets analyzed here, episodic selection appears to be the dominant form of adaptive evolution. The biological implications of this type of selection are discussed for each specific data set. We conclude by providing practical guidelines for applying MEME to biological data, and argue that while it is possible to reliably identify sites or branches subject to episodic diversifying selection, statistical power to detect individual branch-site pairs evolving adaptively is inherently limited by a small sample size available for such inference.

Methods

At its core, our approach uses phylogenetic models to describe the evolution of codon characters along a branch in a phylogeny by a continuous-time stationary Markov process. Given a phylogenetic tree τ, with B branches and a vector of relative branch length parameters $T=(t_i, i=1\ldots B)$, the probability of changing from codon i to j at a site along branch b in time t_b, is recorded in the (i,j) element of the transition matrix $M_b(t_b)=e^{Qt_b}$, where Q is the rate matrix. The elements $Q=\{q_{ij}\}$ parameterize the instantaneous rate of substitution of codon i with codon j:

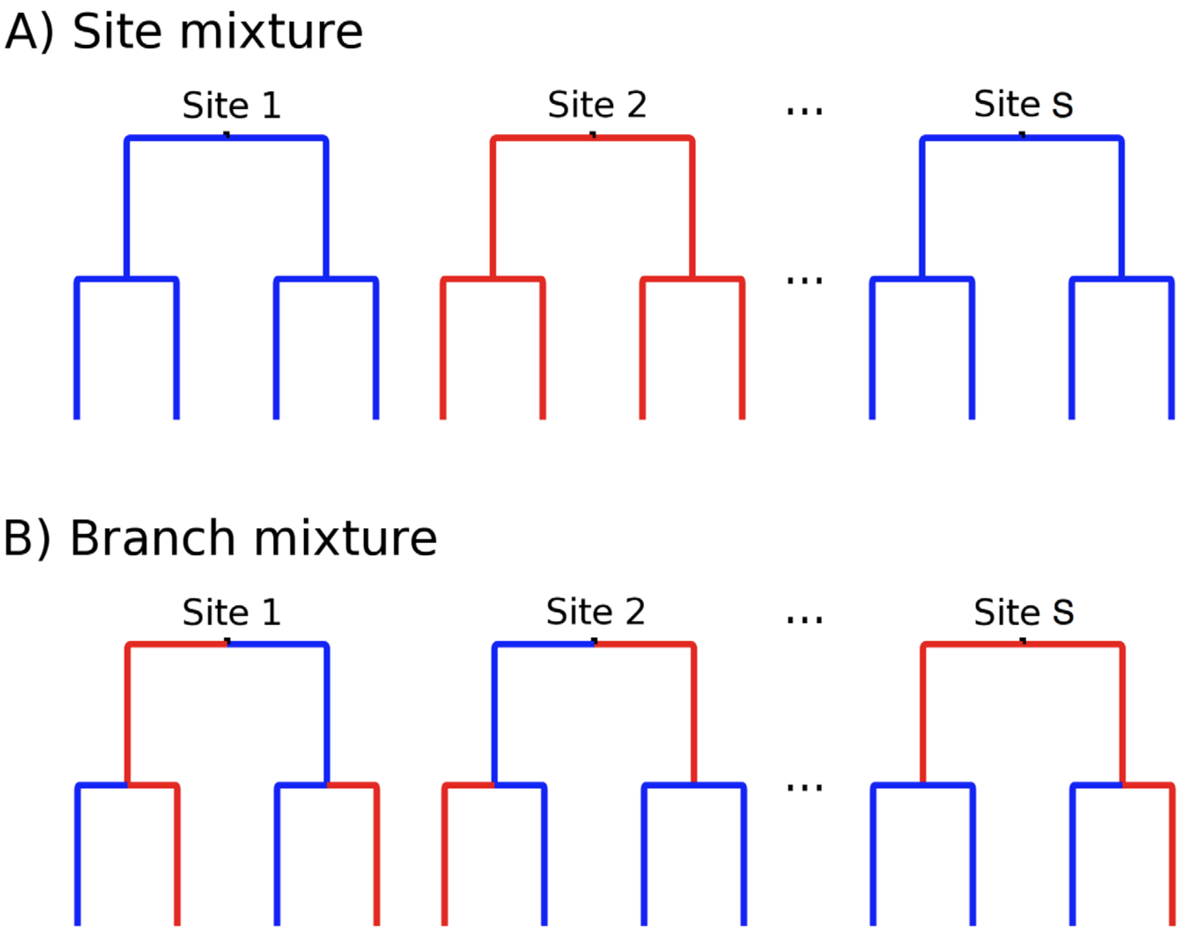

Figure 1. The standard random effects approach and samples. A) The standard random effects approach, in which the rates vary randomly over sites but are constant over branches. Different values of ω are showed in different colors. B) Samples from our new random effects approach [20], used by MEME, in which the rate on each branch is drawn independently of the rate on any other branch. All possible assignments of rates to sites are considered.
doi:10.1371/journal.pgen.1002764.g001

$$q_{ij}(\alpha,\beta,\Pi,\Theta)=\begin{cases}\alpha\theta_{ij}\pi_{ij}, & \delta(i,j)=1,\ \mathrm{AA}(i)=\mathrm{AA}(j),\\ \beta\theta_{ij}\pi_{ij}, & \delta(i,j)=1,\ \mathrm{AA}(i)\neq\mathrm{AA}(j),\\ 0, & \delta(i,j)>1,\\ -\sum_{k\neq i} q_{ik}, & i=j.\end{cases}$$

$\delta(i,j)$ counts the number of nucleotide differences between codons i and j. α and β parameterize the rates of synonymous and non-synonymous substitutions, respectively. θ_{ij} (comprising Θ) are the nucleotide mutational biases, which we model using the 5-parameter general time reversible nucleotide model. π_{ij} (comprising Π) denote the equilibrium frequency parameters. Our estimate (denoted throughout as $\hat{\Pi}$) uses nine position-specific frequency parameters for the target nucleotides [1], corrected for the absence of stop codons using the CF3 × 4 estimator [21]. The likelihood of observing the site is calculated using the pruning algorithm [22] given the data, the tree (τ), the instantaneous rate matrix (Q), and the branch lengths (T).

To model the evolution of a site in an alignment in a manner that treats the non-synonymous rate (β) at each branch b as a random draw from one of K selective categories, we introduce a variable, c_b, which can take values from $1\dots K$. An assignment of categories to all B branches, is described by the configuration vector $C=(c_1,\dots,c_B)$ of branch categories. We assume that the category on each branch is independent of that on all other branches, and that each category has an associated probability, $p(c_b)$, for each branch. Next, we seek to marginalize the likelihood of each site D over all branch configuration vectors:

$$p(D)=\sum_C p(C)p(D|C)$$

Since this sum is over possible configurations, it has B^K terms, and would appear infeasible, unless B is small. However, if we assume that branch categories are independent, $p(C)=\Pi_{b=1}^{B}p(c_b)$, then the sum can be computed directly using the pruning algorithm by replacing the transition matrices with mixtures of transition matrices (see [20] for the derivation). If M_b is the transition matrix on branch b, and we denote Felsenstein's algorithm, which computes the probability of observing D given a transition probability matrix for every branch, as $\mathcal{F}(M_1,\dots,M_B)$, then:

$$\begin{aligned}p(D)&=\sum_C \mathcal{F}(M_1^{c_1},\dots,M_B^{c_B})p(C)\\&=\mathcal{F}\left(\sum_{c=1}^{K}p(c_1)M_1^{c_1},\dots,\sum_{c=1}^{K}p(c_B)M_B^{c_B}\right),\end{aligned}\qquad(1)$$

where $M_b^{c_b}$ associates a transition matrix at each branch with a category. We have thus constructed a tractable model where the process at every branch is a random draw from a set of K categories.

In [20], we used this result to develop a model where each branch had a set of ω values and proportion parameters common to all sites. The goal was to identify lineages with a proportion of sites evolving with $\omega>1$. Here, we let each site have a set of free parameters governing the strength of selection for two discrete categories, and weights for each category, and these parameters are shared for all branches at that site. The goal is to detect sites where a proportion of lineages are evolving with $\omega>1$.

The MEME test for episodic diversifying selection

The fitting of MEME to an alignment of coding sequences proceeds in three stages:

First, the MG94 × REV codon model with an alignment-wide $\omega=\beta/\alpha$ is fitted to the data using parameter estimates under a GTR nucleotide model as initial values. Although in some cases nucleotide branch lengths may be a good approximation to codon branch lengths [23,24], recent results indicate that in other instances, nucleotide models can significantly underestimate branch lengths and possibly bias downstream inference [25]. The resulting maximum likelihood estimates, $\hat{\Theta}$ and $\hat{t}_b$, for each branch $b\in 1\dots B$, are used in the site-by-site analyses in the next two steps. Thus we are assuming that the relative branch length and mutational bias parameters are shared across sites and are well approximated by those estimated under a simpler codon model. However, the absolute branch lengths also depend on the site- and model-specific rate parameters below.

Second, at each site, we first fit the alternative random effects model of lineage-specific selective pressure with two categories of β: $\beta^-\leq\alpha$ and β^+ (unrestricted). The probability ($p(c_b)$ in equation 1) that branch $b\in 1\dots B$ is evolving with $\beta^b=\beta^-$, is $0\leq q^-\leq 1$, and the complementary probability that it is evolving with $\beta_b=\beta^+$ is $q^+=1-q^-$. By equation 1, the phylogenetic likelihood at a site, marginalized over all 2^B possible joint assignments of β_b, is equivalent to computing the standard likelihood function with the following mixture transition matrix for each branch b:

$$\begin{aligned}M_b(\alpha,\beta^-,\beta^+,q^-)=q^-e^{Q\left(\alpha,\beta^-;\hat{\Theta},\hat{\Pi}\right)\hat{t}_b}+\\(1-q^-)e^{Q\left(\alpha,\beta^+;\hat{\Theta},\hat{\Pi}\right)\hat{t}_b}.\end{aligned}\qquad(2)$$

Consequently, the alternative substitution model includes four parameters for each site, inferred jointly from all branches of the tree: β^-,β^+,q^- and α. These form the fixed effects component of the model. Estimating α separately for each site accounts for the site-to-site variability in synonymous substitution rates [26].

Lastly, at every site, we fit the model from the previous step, but with $\beta^+\leq\alpha$: our null model. Using simulated data, we determined that an appropriate asymptotic test statistic for testing most worst-case null of of $\beta^+=\beta^-=\alpha$ is a 0.33 : 0.3 : 0.37 mixture of χ_0^2,χ_1^2 and χ_2^2 (see Text S1). Mixture statistics of this form often arise in hypothesis testing where model parameters take values on the boundaries of the parameter space, and closed-form expressions for mixing coefficients are difficult to obtain [27].

Throughout the manuscript, we compare MEME to the fixed effects likelihood approach, introduced in [24] (see Text S1 for motivation). The procedure used by FEL differs from MEME in that a single pair of α,β rates are fitted at each site (no variation over branches) in Step 2, and the test in Step 3 is to determine if $\alpha\neq\beta$. Positive selection is inferred by FEL when $\hat{\beta}>\hat{\alpha}$ and the p-value derived from the LRT is significant, based on the χ_1^2 asymptotic distribution.

Detecting individual branches subject to diversifying selection at a given site

If the LRT indicates that a particular site (s) is subject to episodic diversifying selection, it may be of interest to explore which branches at that site have undergone diversification. The empirical Bayes (EB) procedure originally used to identify individual sites subject to diversifying selection in random effects models [28], can be readily adapted here. To compute the empirical posterior probability at branch b that $\beta=\beta^+>\alpha$, we apply Bayes' theorem, using D_s to denote the data at site s and $\hat{\Theta}$ to denote all the maximum likelihood parameter estimates from the alternative MEME model fitted to site s:

$$P\left[\beta_b=\beta^+|D_s,\hat{\Theta}\right]=\frac{P(D_s|\beta_b=\beta^+)(1-q^-)}{P(D_s|\beta_b=\beta^+)(1-q^-)+P(D_s|\beta_b=\beta^-)q^-}.$$

To compute the two likelihood terms $P(D_s|\beta_b=\beta^+)$ and $P(D_s|\beta_b=\beta^-)$, we use $q^-=0$ and $q^-=1$, respectively, for the model assigned to branch b in equation 2. The rest of the branches employ the matrices fitted under the alternative model of MEME. Having computed $P\left[\beta_b=\beta^+|D_s,\hat{\Theta}\right]$ for each branch b, we evaluate the empirical Bayes factor for the event of observing positive selection at each branch:

$$EB\left[\beta_b=\beta^+|D_s,\hat{\Theta}\right]=\frac{P\left[\beta_b=\beta^+|D_s,\hat{\Theta}\right]/P\left[\beta_b=\beta^-|D_s,\hat{\Theta}\right]}{(1-q^-)/q^-}.$$

When $EB>K>1$, sequence data increase the prior odds of observing selection at the branch. We do not recommend using this type of inference other than for the purposes of data exploration, even for large values of K (e.g. 100). Intuitively, all the information contributing to the estimate of EB is derived from observing the evolution along a single branch at a single site (i.e. from a sample with size ≈ 1). To quantify this supposition, we simulated sequence data using the vertebrate rhodopsin phylogeny and branch lengths, applied positive selection of varying strength to five branches in the tree selected *a priori* (see Text S1), and applied the EB procedure to infer the identity of selected branches.

Results

Model assessment

To assess the performance of MEME on both simulated and empirical data, we selected the fixed effects likelihood method (FEL [24]) as the most appropriate reference test for pervasive diversifying selection, because FEL most closely matches the assumptions made by MEME (see Text S1). We simulated data sets under a number of scenarios: refer to Text S1 for details of simulation strategies.

Assessing the rates of false positives. Under the scenario where each site was evolved under the worst-case null hypothesis of constant $\omega=1$, MEME had well controlled rates of false positives at test p-value of 0.05 (Figure S1, also see Text S1 for the empirical derivation of the asymptotic distribution of the test statistic for this hypothesis). MEME appears to be conservative for smaller sample sizes (numbers of sequences, N), but not for larger samples. The rates of false positives were <0.01 ($N=8$), 0.01 ($N=16$), 0.03 ($N=32$), 0.04 ($N=64$), and 0.05 ($N=128$ and 256). We also analyzed simulations based on seven large ($N=517-640$) phylogenies downloaded from TreeBase (http://www.treebase.org). The rate of false positives remained well controlled ($0.047-0.053$) at a nominal p-value of 0.05, suggesting that further increasing the number of taxa does not lead to a degradation of Type I error rates.

A further analysis using 36 trees from a variety of published studies downloaded from TreeBase, to simulate 10 replicates from each tree (see Text S1 and Tables S1 and S2 for details), revealed that MEME is generally conservative for alignments of with low pairwise divergence (e.g. <0.1 nucleotide substitutions per site), nominal for those with medium to high pairwise divergence ($0.1-0.4$ nucleotide substitutions per site), and nominal to slightly anti-conservative for higher pairwise divergence (>0.4 nucleotide substitutions per site), although this relationship is influenced by other factors. Overall, we conclude that false positive rates of MEME, are well controlled in the setting of the most pessimistic (strict neutral) null.

Constant selection pressure at individual sites. At nominal $p=0.05$ MEME consistently tracked FEL on sequence alignments simulated under the lineage-constant model assumed by FEL (Table S3), losing several percentage points of power because of its more conservative test statistic. Because each simulated alignment contained a subset of sites generated under the null (neutral model), we could derive empirical estimates of the size of the test and set the nominal p-value to achieve a Type I error rate of 5%. When calibrated to deliver a 5% Type I error rate, MEME held a small edge in power. This finding is not surprising, because at a fixed Type I rate, MEME should find every site found by FEL, and resolve FEL borderline cases affected by stochastic variation in ω throughout the tree.

Variable selection pressure at individual sites. The difference in power between MEME and FEL became stark when selection at individual sites varied among lineages, with each branch evolving under positive selection (ω^+) with probability q^+, and negative selection (ω^-) with complimentary probability $1-q^+$. For every combination of independent simulation parameters (ω^-,ω^+,q^+), MEME had more power to detect sites under episodic diversifying selection (Table 1). Both methods gained power with an increasing proportion of positively selected lineages and/or a greater degree of diversification. The largest differences between MEME and FEL were observed when a small proportion of lineages ($q^+=0.1$) were subjected to diversifying selection. Regardless of the strength of background purifying selection, FEL was effectively powerless (power $0-10\%$) to detect episodes of positive selection under any of the three phylogenetic simulation scenarios, whereas MEME achieved low ($4-53\%$ when $\omega^+=4$), modest ($15-95\%$ when $\omega^+=12$), and excellent ($37-100\%$ when $\omega^+=36$) power. Under these conditions, the power of MEME increased with the alignment size, whereas the power of FEL remained very low. Although FEL gained appreciable power when 25% (or 50%) of the lineages were subject to diversification, its power was on average only $\approx 24\%$ ($\approx 67\%$) of that realized by MEME.

Taken together, the constant and variable selection pressure simulations demonstrate the uniform superiority of MEME over a standard test for diversifying positive selection. MEME has well controlled rates of false positives, has power comparable to FEL when selective forces are uniform at individual sites, and gains a large power advantage when these forces are variable, as is undoubtedly the case in most biological data sets.

Power and accuracy of the empirical Bayes procedure to identify branches subject to diversifying selection at a single site. Our exploratory simulations (see Figure S2) suggest that it is difficult to accurately identify individual positively selected branches at an individual site. We restricted the analysis to only those sites, which were found to be under episodic diversifying selection by MEME ($p\leq 0.05$) and set the threshold of 20 for the empirical Bayes factor to call an individual branch selected. The best results are achieved when selected branches are placed in the background of strongly conserved lineages ($\omega=0.1$) – an individual branch is correctly detected in approximately 25% of cases, while *at least* one selected branch is found in 89.8% of cases (see Figure S3). However, while none of the negatively selected background branches are reported in more than 5% of cases, in

Table 1. Comparative performance of FEL and MEME on simulated data where ω varies along phylogenetic lineages.

		Japanese encephalitis virus *env*			Vertebrate rhodopsin			Camelid VHH		
ω^-	q^+	$\omega^+=4$	$\omega^+=12$	$\omega^+=36$	$\omega^+=4$	$\omega^+=12$	$\omega^+=36$	$\omega^+=4$	$\omega^+=12$	$\omega^+=36$
0	0.1	0.00 **0.06**	0.01 **0.25**	0.03 **0.50**	0.00 **0.21**	0.00 **0.53**	0.02 **0.81**	0.00 **0.53**	0.00 **0.95**	0.04 **0.99**
0	0.25	0.01 **0.12**	0.06 **0.32**	0.12 **0.51**	0.01 **0.30**	0.04 **0.68**	0.15 **0.88**	0.00 **0.66**	0.14 **0.98**	0.56 **1.00**
0	0.5	0.06 **0.12**	0.19 **0.29**	0.34 **0.45**	0.09 **0.28**	0.34 **0.59**	0.54 **0.82**	0.23 **0.77**	0.85 **0.98**	0.96 **0.98**
0.2	0.1	0.00 **0.05**	0.01 **0.21**	0.02 **0.41**	0.00 **0.09**	0.01 **0.35**	0.02 **0.67**	0.00 **0.16**	0.01 **0.87**	0.04 **0.98**
0.2	0.25	0.02 **0.08**	0.07 **0.27**	0.14 **0.48**	0.03 **0.17**	0.09 **0.55**	0.17 **0.84**	0.01 **0.42**	0.27 **0.96**	0.62 **0.99**
0.2	0.5	0.05 **0.11**	0.18 **0.29**	0.36 **0.49**	0.13 **0.25**	0.36 **0.60**	0.55 **0.76**	0.30 **0.72**	0.84 **0.99**	0.90 **0.99**
0.4	0.1	0.00 **0.04**	0.01 **0.15**	0.03 **0.37**	0.01 **0.07**	0.02 **0.30**	0.03 **0.57**	0.01 **0.10**	0.04 **0.78**	0.10 **0.97**
0.4	0.25	0.02 **0.06**	0.09 **0.27**	0.15 **0.45**	0.04 **0.16**	0.09 **0.49**	0.21 **0.78**	0.03 **0.32**	0.33 **0.97**	0.63 **0.99**
0.4	0.5	0.07 **0.10**	0.17 **0.26**	0.33 **0.46**	0.17 **0.28**	0.39 **0.58**	0.51 **0.76**	0.40 **0.62**	0.82 **0.94**	0.96 **1.00**

Power to detect sites under selection ($p=0.05$) are reported for FEL and MEME (in **boldface**) for each unique combination of negative selection (ω^-), positive selection (ω^+), and proportion of branches under positive selection (q^+) parameters.

doi:10.1371/journal.pgen.1002764.t001

55% of cases *at least* one background branch was falsely detected as positively selected. In a more difficult case of neutrally evolving background, the EB procedure performs considerably worse: at least one select branch is found in 55.6% of cases, whereas at least one background branch is detected in 86.5% instances. 18 background neutral branches are reported as selected at over 5% frequency, while the 5 positively selected branches are identified at $3.4-26\%$ of selected sites.

Empirical data

To gauge the comparative performance of MEME and FEL when identifying sites subject to pervasive diversifying selection, we used a collection of 16 protein-coding alignments, representing a diverse array of taxa, genes subject to differing levels of conservation, and a range of data set sizes (Table 2). In 12/16 alignments analyzed, MEME identified all the sites inferred by FEL to be under diversifying positive selection and found between

Table 2. Comparative performance of MEME and FEL on 16 empirical alignments (see Results and Text S1 for an extended discussion of each individual case).

Data set	N	S	Mean	Classes of sites detected at $p\leq0.05$				Mean q^+		Sites where
			Div.	M^+F^0	M^+F^+	M^+F^-	M^-F^+	M^+F^{0-}	M^+F^+	MEME>FEL at $p=0.05$
Abalone sperm lysin	25	134	0.43	17	9	0	1 (0.04/0.05)	0.17	0.35	19
Camelid VHH	212	96	0.27	22	6	2	0 (n/a)	0.11	0.50	26
Diatom SIT	97	300	0.54	12	0	36	0 (n/a)	0.05	n/a	82
Drosophila *adh*	23	254	0.26	9	1	0	0 (n/a)	0.09	0.19	7
Echinoderm H3	37	111	0.33	0	0	1	0 (n/a)	0.02	n/a	3
Flavivirus NS5	18	342	0.48	3	0	1	0 (n/a)	0.16	n/a	7
Hepatitis D virus Ag	33	196	0.29	13	7	0	1 (0.05/0.07)	0.08	0.37	10
HIV-1 *rt*	476	335	0.08	12	10	7	0 (n/a)	0.04	0.69	27
HIV-1 *vif*	29	192	0.08	5	2	0	7 (0.04/0.06)	0.11	0.59	3
IAV H3N2 HA	349	329	0.04	7	11	2	3 (0.04/0.06)	0.04	0.73	8
JEV *env*	23	500	0.13	2	1	1	0 (n/a)	0.11	1.00	3
Mamallian β-globin	17	144	0.38	10	2	0	0 (n/a)	0.20	0.31	11
Primate *COXI*	21	510	0.36	3	0	1	0 (n/a)	0.18	n/a	4
Salmonella *recA*	42	353	0.04	1	0	0	0 (n/a)	0.02	n/a	0
Vertebrate rhodopsin	38	330	0.34	13	1	5	0 (n/a)	0.11	0.74	39
West Nile virus NS3	19	619	0.13	1	1	0	0 (n/a)	0.04	1.00	2
Total/Mean				130	51	56	12	0.10	0.59	

N (S) reports the number of sequences (codons) in the alignment. M^+ (M^-) refers sites found by MEME to be positively (negatively) selected ($p\leq0.05$). F^+ (F^-) denote sites found by FEL to be positively (negatively) selected ($p\leq0.05$). F^0 references sites that are classified as neutrally evolving by FEL. Values in parentheses for the M^-F^+ column show the mean p-values for FEL and MEME on this set of sites, respectively. Values reported in the rightmost column count the number of sites where MEME fits significantly better than FEL, based on a 2-degrees of freedom LRT ($p\leq0.05$). Abbreviations: IAV = Influenza A virus, JEV = Japanese encephalitis virus.

doi:10.1371/journal.pgen.1002764.t002

1 (e.g. West Nile virus NS3) and 48 (Diatom SIT) additional sites that were subject to episodic diversifying selection (Table 2). In four data sets, $1-7$ sites identified by FEL with p-values close to 0.05 were missed by MEME. Note that MEME p-values for these sites remained in the $0.05-0.07$ range (Table 2), i.e. marginally significant.

Sites identified by both methods tended to have a greater average proportion of lineages under selection (0.59, measured by the mean of MLE estimates of q^+); sites found only by MEME experienced more episodic selection (0.10). In 9 data sets (Table 2), sites that FEL inferred to be under purifying selection are instead identified by MEME as likely to have been subjected to episodic diversifying selection. Almost universally (Tables S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19), such sites had a smaller estimated proportion of positively selected lineages ($<10\%$). This behavior is consistent with the relative performance of the two tests on simulated data and corroborates the expectation that MEME has greater power to identify sites when only a proportion of lineages evolved under positive selection. Vertebrate rhodopsin, Japanese encephalitis virus *env*, and Camelid VHH are investigated in detail below; for a discussion other genes, see Text S1.

Vertebrate rhodopsin

The vertebrate rhodopsin (a low-light vision protein) data set was previously experimentally investigated for the substitutions that modulate the wavelength of the light absorbed by the molecule (λ_{max}, [18]). The authors asserted that, because none of the 12 sites that they had determined as affecting λ_{max} by site-directed mutagenesis were detected by site-level computational methods, "statistical tests of positive selection can be misleading without experimental support." Other authors reanalyzed the same data set more comprehensively and went even further, questioning the utility of ω-based methods for detecting experimentally validated sites, because "most of the current statistical methods are designed to identify codon sites with high ω values, which may not have anything to do with functional changes. The codon sites showing functional changes generally do not show a high ω value" [29]. The validity of this generalization has been correctly questioned with a simple counter-argument that the sites detected by computational methods may also be functionally important, because the change in λ_{max} is unlikely to be the sole determinant of adaptation [17].

The MEME analysis of this gene suggests another obvious alternative, also expounded by previous studies [17]: the failure of the original computational analysis [18] to identify functionally important sites results from the fact that these sites have been subjected to episodic selection, which is masked by predominantly purifying selection elsewhere in the tree. Indeed, among three sites that alter λ_{max} found by MEME (96, 183 and 195, versus none found by FEL), no more than 13% of the branches exhibited $\omega > 1$ (Table S17); at these sites, the average ω is less than 1. We note that, because adaptive evolution will not always adhere to a single, simple scenario of episodic diversifying selection, we do not expect MEME to find all 12 sites experimentally confirmed to alter λ_{max}. For example, three of the nine missed sites (83,194,292) appear to have been subjected to partial selective sweeps and have been detected using a specialized model of directional evolution [29].

Three sites from this alignment can be used to illustrate how the inclusion of lineage variability modifies inference of selection (Figure 2). Site 54 was inferred to have experienced pervasive non-synonymous substitutions throughout its evolutionary history. Both FEL and MEME detect this site as positively selected ($p = 0.02$). Sixty three percent of the lineages at this site evolved with $\beta^+ > \alpha$, whereas the remainder were conserved ($\alpha = \beta^- = 0$), according to MEME. The log-likelihood of the site is only marginally higher for MEME, which suggests that MEME behaves like FEL at sites with

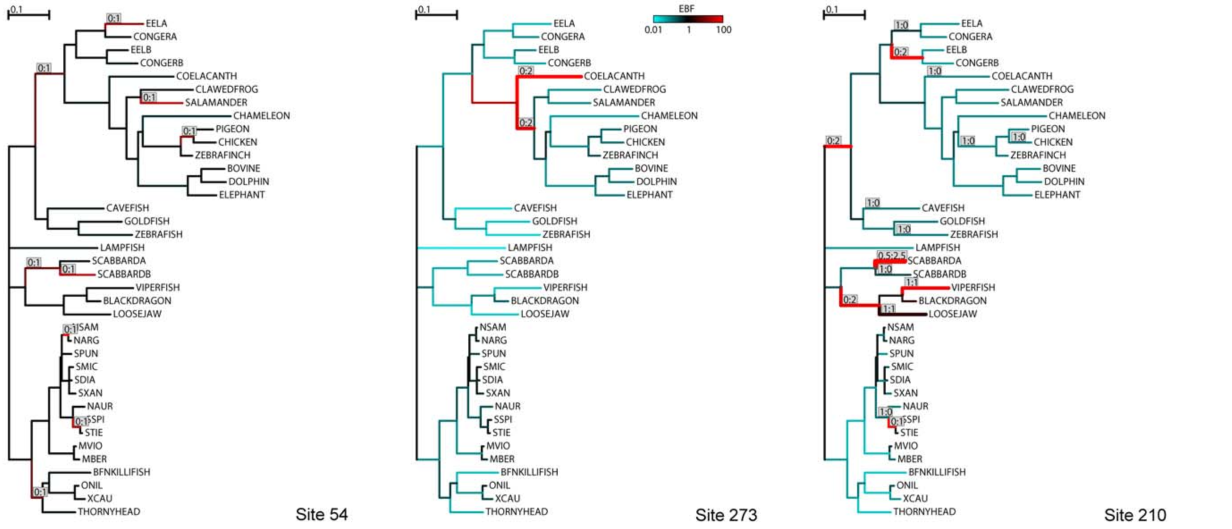

Figure 2. Individual sites of the vertebrate rhodopsin alignment used to illustrate similarities and differences between FEL and MEME. Branches that have experienced substitutions, based on most likely joint maximum likelihood ancestral reconstructions at a given site, are labeled as count of synonymous substitutions:count of non-synonymous substitutions. The thickness of each branch is proportional to the minimal number of single nucleotide substitutions mapped to the branch. Branches are colored according to the magnitude of the empirical Bayes factor (EBF) for the event of positive selection: red – evidence for positive selection, teal – evidence for neutral evolution or negative selection, black –Ê no information. See Methods for more detail. All three sites were identified as experiencing positive diversifying selection by MEME. FEL reported site 54 as positively selected, site 273 as neutral, and site 210 as negatively selected.

doi:10.1371/journal.pgen.1002764.g002

"canonical" patterns of diversifying selection, corroborating the simulation results.

At codon 273, FEL obtained a maximum likelihood estimate of $\beta > \alpha$, but failed to infer positive selection, as the signal was not statistically significant ($p = 0.70$). MEME, on the other hand, allocated 0.04 (0.013–0.10: 95% confidence interval obtained by latin hypercube sampling importance resampling [30]) of branches to a rate class with $\alpha = 0.0, \beta^+ = 9.49$ (2.94–6726) and inferred positive selection ($p = 0.03$). The difference in log-likelihoods between MEME and FEL is 4.9 points: MEME fits significantly better, based on a 2-degrees of freedom likelihood ratio test ($p = 0.007$). The maximum likelihood estimates of individual model parameters have large associated errors (although in all posterior samples we obtained $\beta^+ > \alpha$), as is expected for inference based on a single site. This has also been noted by Yang and dos Reis [17]. The point estimates themselves, however, are immaterial for inferring whether or not a site is positively selected, since the likelihood ratio test is used for that purpose.

Perhaps the most dramatic example of the added power of MEME is illustrated by site 210. At this site, the evolutionary history is replete with non-synonymous substitutions along deep lineages followed by extensive synonymous evolution, indicative of purifying selection. There is also a small clade with repeated synonymous and nonsynonymous substitutions. Averaging over all branches, FEL determined that the site, overall, is under negative selection ($p = 0.01$). MEME reported that 89% of the branches were under a very strong selective constraint ($\alpha = 2.13, \beta^- = 0.0$), but that the remaining $11\%(5.5 - 18.6\%)$ were under strong diversifying selection ($\beta^+ = 26.5(10.1 - 6519)$). The log-likelihood improvement is now 13.4 at the cost of two parameters, which is highly significant ($p < 0.001$). Site 210 is the ideal illustration of why it is undesirable to average ω over all lineages: bursts of diversification followed by conservation will most likely be missed by traditional site methods.

Japanese encephalitis virus *env*

No evidence for selection was found in this envelope gene in previous analyses [28], and FEL found only one site under positive selection. Despite the low levels of divergence among a relatively small number of taxa (23 isolates), MEME found episodic selection at sites called negatively selected by FEL (Table S12). Two of these sites fall within a beta-barrel epitope known to be involved in escape from neutralizing antibodies [31]. Sites 33 and 242 showed evidence of repeated toggling at terminal lineages. Remarkably, site 33 – likely a part of a neutralizing antibody epitope [32] – changed from isoleucine to leucine on 6 terminal lineages; site 242 changed from phenylalanine to serine on 5 terminal lineages. These substitutions co-occur on three terminal lineages. Evidence of recombination was detected in this alignment, and corrected for using a partitioning approach (details on how MEME can correct for recombination are in Text S1).

Camelid VHH

The camelid VHH data set comprises partial variable domain sequences (germline alleles) of llama and dromedary heavy chain only antibodies (Table S3). 11 of 16 sites in the variable complementarity determining regions (CDR) 1 (sites 26–33) and 2 (sites 51–58) were found to be under diversifying selection by MEME (2/16 were detected by FEL and 2 more were marginally significant). Because CDR regions are driven to diversify in order to provide a broad basis of antigen recognition, positive selection is expected to be commonplace in the CDRs [33]. MEME was able to uncover selective signatures at a majority of those sites. Of the remaining 19 sites classified by MEME as positively selected, six were associated with VHH family differentiation [34]. Unlike standard antibodies, which must maintain relatively conserved framework regions (FR) involved in binding heavy and light chains to form functional tetramers, VHH antibodies are free of such functional constraints. A previous analysis of camelid VHH for evidence of positive selection using counting methods [35] reported evidence for positive selection at a single site (14) in FR1 (sites 1–25 in Table S3), but this analysis could find no clear evidence of positive or negative selection at 49 FR sites. In contrast, MEME inferred episodic selection at six sites in FR1, six sites in FR2 (sites 34–50), and 7 sites in FR3 (sites $59 - 96$). The well-known lack of power of counting methods to detect even pervasive selection [17] likely hampered the previous study.

Effect of sequence sampling

Although a previous analysis of 38 vertebrate rhodopsin sequences found no sites under selection at posterior probability $\geq 95\%$ [18], the same authors found 7 selected sites in the subset of 11 squirrelfish sequences, and 2 selected sites when the subset of 28 fish sequences was analyzed. These results run counter to the expectation that more data should provide greater power to detect selection. MEME, on the other hand, detects more selected sites when more sequences are included. One site is identified in the squirrelfish alignment, 9 in the fish alignment, and 19 in the complete rhodopsin alignment. All but 5 sites detected in the subset alignments are also identified in the full alignment (Table S20). Allowing ω to vary over branches at least partially mitigates the pathology of constant-ω models which effectively rely on an average ω for inferring selection. A similar pattern is seen in the analysis of the influenza A virus H3N2 hemagglutinin sequences, where site-level methods also appear to be sensitive to sequence sampling ([19], see Text S1 and Table 23).

Discussion

We have presented a mixed effects model of evolution, MEME, and a statistical test for detecting the signal of past episodic positive selection from molecular sequence data. Our model corrects the biologically unrealistic assumption that selective pressure, as measured by the ω ratio, remains constant over lineages. Based on comprehensive simulations and empirical analysis of an array of taxonomically diverse genes, MEME can be recommended as a replacement for existing site models. MEME matches the performance of older approaches when natural selection is pervasive, but possesses greater power to identify sites where episodes of positive selection are confined to a small subset of branches in a phylogenetic tree.

Our results suggest that it may be necessary to revise previous estimates of the proportion of sites under positive selection in many genes. Using the FEL method, which assumes constant selective pressure at a site, we are able to detect 63 sites across all 16 empirical alignments. MEME identifies 51 of these sites (the remaining 12 are borderline significant) and 186 additional sites – nearly 4 times as many as FEL. For individual data sets (e.g. Drosophila *adh* and Diatom SIT, Table 2), the differences may be even more dramatic. The greater power of MEME indicates that selection acting at individual sites is considerably more widespread than constant ω models would suggest. It also suggests that natural selection is predominantly episodic, with transient periods of adaptive evolution masked by the prevalence of purifying or neutral selection on other branches. We emphasize that MEME is not just a quantitative improvement over existing models: for 56 sites in our empirical analyses, we obtain qualitatively different conclusions. FEL asserts that these sites evolved under significant purifying selection, but MEME is able to identify the signature of positive selection on some branches. Furthermore, MEME is less sensitive to sampling effects that plague existing positive selection detection tools [18,19]. Variable levels of purifying selection pressure across different lineages prevented these older methods from detecting instances of episodic positive selection; MEME is able to peer through the fog of purifying selection.

It is important to bear in mind that the mixture χ^2 statistic used to calculate the p-values reported here is based on a null model under which all sites are evolving neutrally. This, however, is not biologically realistic: the null hypothesis against which sites ideally ought to be screened is one under which sites are evolving *either* neutrally *or* under purifying selection. But the proportion of sites evolving under negative selection and the strength of this selection are unknown and vary from case to case, which means that such a null hypothesis would be very sensitive to modeling assumptions that cannot be justified in general. Instead, the neutral null hypothesis represents a worst case scenario for our inference, so that the p-values we obtain are upper bounds of the true p-values. This ensures that our inference is conservative. Even in the worst (and biologically unrealistic) case for MEME, namely when selective pressures are constant throughout the phylogeny, the loss of power compared to FEL is minimal: a site with FEL p-values between 0.0346 and 0.05 will be missed by MEME, since its p-value will be >0.05 for the same ranges of the likelihood ratio test statistic (LRT). In our simulation scenarios under the assumption of constant ω, this translates to no more a 5% loss in power (Table S3).

Our inference is performed in a site-wise rather than an alignment-wide manner, and we therefore control the site-wise rather than the family-wise error rate. We do not recommend combining the results of multiple site-wise inferences to perform alignment-wide inference. To aid interpretation of the results while taking account of multiple testing, we calculate the false discovery rate [36]; the resulting q-value upper bounds are reported alongside their corresponding p-value upper bounds in Tables S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19. This gives an upper bound on how many of the reported sites can be expected to be false discoveries: for instance, of the 30 sites reported in Table S5 we expect no more than 5 (14%) to be false, and probably far fewer because of the conservative choice of null model. We emphasize that q-values are usually much larger than their corresponding p-values and caution that p-values (regardless of whether they have been corrected for multiple testing) cannot be used to estimate an expected number of false discoveries in the same way.

MEME is a conceptual advance over the first generation of random effects models designed to detect episodic selection (called "branch-site models" in the literature [17]). MEME does not require *a priori* designation of, or an exhaustive search for, the branches under selection, and it allows each site to have its own selective history. Whereas branch-site models make restrictive *a priori* assumptions about how ω values are distributed across the tree – sometimes leading to very poor statistical performance [20] – MEME treats the selective class on each branch as a random effect that is marginalized over in the likelihood calculation.

For computational tractability, MEME assumes that the value taken by ω on each branch is independent of that on any other branch, i.e. selective pressures between branches are uncorrelated. This assumption could potentially be violated: for example, if ω changes very slowly across the phylogeny, then ω values on neighboring branches will be correlated. Further research is needed to understand how inference of selection would be affected if these correlations were directly accounted for, and whether the additional model and computational complexity would be justified. In practice, MEME could be combined with models of directional selection to improve power, e.g. [15,16]. Unlike covarion models [37,13], MEME does not allow ω to change in the middle of a tree branch. The effect of this restriction is unclear, but it could be tested by implementing a mixed effects covarion model, where switching rates and proportion of time spent under $\omega>1$ are estimated at an individual site.

The ability of MEME, or similar substitution model-based methods, to accurately infer the identity of individual branches subject to diversifying selection at a given site seems unavoidably limited. Most of the information that such inference might be based on is limited to character substitutions along a single branch at a single site, i.e. one realization of the Markov substitution process. Selection along terminal branches in the context of negatively selected background can be detected more reliably than selection along interior branches among neutrally evolving background lineages. However, we caution that despite obvious interest in identifying specific branch-site combinations subject to diversifying selection, such inference is based on very limited data (the evolution of one codon along one branch), and cannot be recommended for purposes other than data exploration and result visualization. This observation could be codified as the "selection inference uncertainty principle" – one cannot simultaneously infer both the site and the branch subject to diversifying selection. In this manuscript, we describe how to infer the location of sites, pooling information over branches; previously [20] we have outlined a complementary approach to find selected branches by pooling information over sites.

Finally, although MEME is considerably more powerful than existing methods at detecting bursts of selection, it still requires that a measurable proportion of lineages ($5-10\%$) experience non-synonymous evolution at a site. When a single substitution modifies an adaptive trait and is subsequently fixed, we expect ω based methods to have very little power. Specialized methods which make use of change in allele frequencies [15,16], or between and within-population diversification patterns [38], will be required in such cases.

Supporting Information

Figure S1 Quantile–Quantile plot of three asymptotic distributions (x-axis) for the MEME LRT test versus the LRT derived by parametric bootstrap (y-axis), limited to the meaningful test p-value range of <0.01. The χ^2_1 distribution is too liberal (lying below the $x=y$ line), the χ^2_2 is too conservative, while the mixture is approximately correct.
(PDF)

Figure S2 Simulation parameters for generating datasets for evaluating the empirical Bayes inference of branch-site combinations under selection. Branches are colored according the the value of ω used to evolve sequences along them; branches simulated under positive selection are also labeled with ω values.
(PDF)

Figure S3 Summary of empirical Bayes inference of branches under selection on data simulated using the selective parameters from Figure S2. Each branch is colored according to the proportion of times it was found to have an empirical Bayes factor of 20 or greater at sites with MEME p-value of 0.05 or less. Branches with $>5\%$ detection rates are also labeled with the values of the rates.
(PDF)

Table S1 False positive rates for data sets simulated under strict neutrality using empirical trees from TreeBase. The entries are sorted in order of increasing mean false positive rate derived from simulated data (10 replicates per tree). Mean divergence between any pair of leaves in a given tree is reported in expected nucleotide substitutions per site. False positive range reports the minimum and maximum values for false positive rates for an individual replicate. 95% confidence intervals are derived from the binomial distribution with the probability of success $p = 0.05$, and the number of trials N equal to the number of codons. This range provides the expected spread of per replicate false positive rates for a test that has the probability of making a false positive error of exactly 0.05 over N tests.
(PDF)

Table S2 False positive rates for three empirical trees from TreeBase when the parameters of the null model are varied: 20% of the branches are simulated with the foreground ω, and the remainder under the background ω. 10 replicates with 300 codons each per tree-ω pair were simulated. The synonymous rate was set to 0.52 for the first 150 codons, 0.9 for the next 100 codons, and 1.58 for the last 50 codons.
(PDF)

Table S3 Comparative performance of FEL and MEME on simulated data where ω does not vary among tree branches. The rate of false positives (FP) and power are reported for a fixed nominal test p-value of 0.05. Power is also shown for the p-value that achieves FP of 0.05, estimated empirically from the distribution of p-values on the subset of sites evolving neutrally.
(PDF)

Table S4 Positively selected sites in abalone sperm lysin. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S5 Positively selected sites in camelid VHH. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S6 Positively selected sites in Diatom silicon transporters found by MEME at $p \leq 0.05$. The FEL result column summarizes the classification obtained by FEL. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S7 Positively selected sites in Drosophila *adh* found by MEME at $p \leq 0.05$. The FEL result column summarizes the classification obtained by FEL. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S8 Positively selected sites in Echinoderm histone H3. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S9 Positively selected sites in Flavivirus NS5. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S10 Positively selected sites in Hepatitis D virus Ag. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S11 Positively selected sites in HIV-1 reverse transcriptase (*rt*). + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S12 Positively selected sites in HIV-1 viral infectivity factor (*vif*). + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S13 Positively selected sites in Influenza A virus hemagglutinin (H3N2 serotype). Superscript letters after the site indicate the epitope in which substitutions can affect phenotype. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S14 Positively selected sites in Japanese encephalitis virus *env*. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S15 Positively selected sites in mammalian β-globin. The FEL result column summarizes the classification obtained by FEL. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S16 Positively selected sites in primate cytochrome c oxidase subunit 1 (*COX1*). + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S17 Positively selected sites in Salmonella *recA*. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites

(FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S18 Positively selected sites in vertebrate rhodopsin. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S19 Positively selected sites in West Nile virus NS3. + stands for a positively selected site and − stands for a negatively selected site (FEL $p > 0.1$). ++ and −− reflect borderline significant sites (FEL p between 0.05 and 0.1). +++ and −−− denote significant sites (FEL $p \leq 0.05$).
(PDF)

Table S20 Test p-values for positively selected sites found by MEME in a set of 38 vertebrate rhodopsin sequences analyzed with REL methods in Yokoyama2008fk. Sites with $p \leq 0.05$ are shown in bold. The partial ordering of subsets is as follows: Squirrelfish ⊂ Fish ⊂ All, Coelacanth and tetrapods ⊂ All. Sites found to be under positive selection with posterior probability of >95% (M8 model) in Yokoyama2008fk in at least one of the subsets are marked with $\star$.
(PDF)

Table S21 Test p-values for positively selected sites found by MEME in a set of 86 influenza A virus hemagglutinin sequences (Set 3) and its various subsets, analyzed with REL methods in Chen2011fk. Sites with $p \leq 0.05$ are shown in bold. The partial ordering of subsets is as follows: Set 4 ⊂ Set 1 ⊂ Set 3, Set 5 ⊂ Set 2 ⊂ Set 3, Set 6 ⊂ Set 3, Set 7 ⊂ Set 3. Sites found to be under positive selection with posterior probability of >95% (M3 model) in Chen2011fk in at least one of the subsets are marked with $\star$.
(PDF)

Text S1 Supplementary methods, results, and discussion.
(PDF)

Author Contributions

Conceived and designed the experiments: BM KS SLKP. Performed the experiments: BM JOW SM TW SLKP. Analyzed the data: BM JOW SM TW SLKP. Contributed reagents/materials/analysis tools: BM SLKP. Wrote the paper: BM JOW KS SLKP.

References

1. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715–724.
2. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–36.
3. Delport W, Scheffler K, Seoighe C (2009) Models of coding sequence evolution. Brief Bioinform 10: 97–109.
4. Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26: 255–271.
5. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335: 167–70.
6. Bonhoeffer S, Holmes EC, Nowak MA (1995) Causes of HIV diversity. Nature 376: 125.
7. Messier W, Stewart CB (1997) Episodic adaptive evolution of primate lysozymes. Nature 385: 151–4.
8. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–6.
9. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3: e170. doi:10.1371/journal.pbio.0030170
10. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–36.
11. Sawyer SL, Wu LI, Emerman M, Malik HS (2005) Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A 102: 2832–7.
12. Brault AC, Huang CYH, Langevin SA, Kinney RM, Bowen RA, et al. (2007) A single positively selected West Nile viral mutation confers increased virogenesis in American crows. Nat Genet 39: 1162–6.
13. Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP (2004) Modeling the site-specific variation of selection patterns along lineages. Proceedings of the National Academy of Sciences of the United States of America 101: 12957–12962.
14. Delport W, Scheffler K, Seoighe C (2008) Frequent toggling between alternative amino acids is driven by selection in HIV-1. PLoS Pathog 4: e1000242. doi:10.1371/journal.ppat.1000242
15. Seoighe C, Ketwaroo F, Pillay V, Scheffler K, Wood N, et al. (2007) A model of directional selection applied to the evolution of drug resistance in HIV-1. Mol Biol Evol 24: 1025–31.
16. Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to inuenza A virus. Mol Biol Evol 25: 1809–24.
17. Yang Z, dos Reis M (2011) Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28: 1217–28.
18. Yokoyama S, Tada T, Zhang H, Britt L (2008) Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proc Natl Acad Sci U S A 105: 13480–5.
19. Chen J, Sun Y (2011) Variation in the analysis of positively selected sites using nonsynonymous/synonymous rate ratios: an example using inuenza virus. PLoS ONE 6: e19996. doi:10.1371/journal.pone.0019996
20. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, et al. (2011) A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28: 3033–3043.
21. Kosakovsky Pond S, Delport W, Muse SV, Scheffler K (2010) Correcting the bias of empirical frequency parameter estimators in codon models. PLoS ONE 5: e11230. doi:10.1371/journal.pone.0011230
22. Felsenstein J (1981) Evolutionary trees from DNA-sequences – a maximum-likelihood approach. J Mol Evol 17: 368–376.
23. Yang Z (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human inuenza virus A. Journal of Molecular Evolution 51: 423–432.
24. Kosakovsky Pond SL, Frost SDW (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208–1222.
25. Wertheim JO, Kosakovsky Pond SL (2011) Purifying selection can obscure the ancient age of viral lineages. Mol Biol Evol.
26. Pond SK, Muse SV (2005) Site-to-site variation of synonymous substitution rates. Mol Biol Evol 22: 2375–85.
27. Self SG, Liang KY (1987) Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. J Am Stat Assoc 82: 605–310.
28. Yang ZH, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics 155: 431–449.
29. Nozawa M, Suzuki Y, Nei M (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci U S A 106: 6700–5.
30. Pond SLK, Scheffler K, Gravenor MB, Poon AFY, Frost SDW (2010) Evolutionary fingerprinting of genes. Mol Biol Evol 27: 520–36.
31. Wu KP, Wu CW, Tsao YP, Kuo TW, Lou YC, et al. (2003) Structural basis of a avivirus recognized by its neutralizing antibody: solution structure of the domain III of the Japanese encephalitis virus envelope protein. J Biol Chem 278: 46007–13.
32. Gangwar RS, Shil P, Cherian SS, Gore MM (2011) Delineation of an epitope on domain I of Japanese encephalitis virus Envelope glycoprotein using monoclonal antibodies. Virus Res 158: 179–87.
33. Tanaka T, Nei M (1989) Positive darwinian selection observed at the variable-region genes of immunoglobulins. Mol Biol Evol 6: 447–59.
34. Harmsen M, Ruuls R, Nijman I, Niewold T, Frenken L, et al. (2000) Llama heavy-chain V regions consist of at least four distinct subfamilies revealing novel sequence features. Molecular Immunology 37: 579–590.
35. Su C, Nguyen VK, Nei M (2002) Adaptive evolution of variable region genes encoding an unusual type of immunoglobulin in camelids. Mol Biol Evol 19: 205–15.
36. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57: pp. 289–300.
37. Tuffley C, Steel M (1998) Modeling the covarion hypothesis of nucleotide substitution. Mathematical biosciences 147: 63–91.
38. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–4.

Modeling HIV-1 Drug Resistance as Episodic Directional Selection

Ben Murrell[1,2], Tulio de Oliveira[3,4], Chris Seebregts[1,5], Sergei L. Kosakovsky Pond[6], Konrad Scheffler[2,6]* on behalf of the Southern African Treatment and Resistance Network (SATuRN) Consortium

[1] Biomedical Informatics Research Division, eHealth Research and Innovation Platform, Medical Research Council, Tygerberg, South Africa, [2] Computer Science Division, Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa, [3] Africa Centre for Health and Population Studies, Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa, [4] Research Department of Infection and Population Health, University College London, London, United Kingdom, [5] School of Computer Science, University of KwaZulu-Natal, Durban, South Africa, [6] Department of Medicine, University of California San Diego, San Diego, California, United States of America

Abstract

The evolution of substitutions conferring drug resistance to HIV-1 is both episodic, occurring when patients are on antiretroviral therapy, and strongly directional, with site-specific resistant residues increasing in frequency over time. While methods exist to detect episodic diversifying selection and continuous directional selection, no evolutionary model combining these two properties has been proposed. We present two models of episodic directional selection (MEDS and EDEPS) which allow the *a priori* specification of lineages expected to have undergone directional selection. The models infer the sites and target residues that were likely subject to directional selection, using either codon or protein sequences. Compared to its null model of episodic diversifying selection, MEDS provides a superior fit to most sites known to be involved in drug resistance, and neither one test for episodic diversifying selection nor another for constant directional selection are able to detect as many true positives as MEDS and EDEPS while maintaining acceptable levels of false positives. This suggests that episodic directional selection is a better description of the process driving the evolution of drug resistance.

Citation: Murrell B, de Oliveira T, Seebregts C, Kosakovsky Pond SL, Scheffler K, et al. (2012) Modeling HIV-1 Drug Resistance as Episodic Directional Selection. PLoS Comput Biol 8(5): e1002507. doi:10.1371/journal.pcbi.1002507

Editor: Christophe Fraser, Imperial College London, United Kingdom

Received September 22, 2011; **Accepted** March 16, 2012; **Published** May 10, 2012

Funding: This research was supported by Europeaid Grant number SANTE/2007/147-790 from the European Commission; BM is supported by the same Europeaid Grant. TdO's work on this paper was funded by the same Europeaid grant, by the Africa Centre for Health and Population Studies Wellcome Trust Core Grant 082384/Z/07/Z and the grant entitled Swiss-Prot/South Africa: Protein Bioinformatics Resource Development for Important Health-related Pathogens" under the Switzerland-South Africa Collaborative Research Program. Funding for the UCSD computing cluster has been provided by the Joint DMS/NIGMS Mathematical Biology Initiative through Grant NSF-0714991 and the National Institutes of Health grants AI47745 and AI74621. HyPhy custom script development was supported by the National Institute Of General Medical Sciences (grant GM093939). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: kscheffler@ucsd.edu

Introduction

Among positively selected evolutionary changes, a distinction can be made between *diversifying selection*, where any nucleotide substitutions that change the amino acid are favored, and *directional selection*, where only substitutions towards a small number of target amino acids are selected for. Detection of genes or sites evolving under positive selection [1–6] has been dominated by methods which explicitly or implicitly assume *diversifying* positive selection. This assumption allows evolution to be modeled as a continuous-time Markov process without assuming that any particular residue is the preferred target of substitutions at any sites. For most models of diversifying selection, apart from a single rate governing amino acid change, the process is no different from one site to the next. By contrast, models have been proposed in which specific residues do have special status at specific sites. In models of toggling selection [7], substitutions away from a site-specific "wild type" amino acid are likely to be followed by reversions to that amino acid. Models of directional selection [8,9] allow substitution rates towards a site-specific "target" amino acid to be accelerated. By making a distinction among all possible targets of a substitution, such models allow the detection of positive selection favoring mutations towards one amino acid, even at sites where the overall rate of amino acid change is decreased by purifying selection. For a review of codon models of selection, see [10].

A second distinction is that between selective pressure that is constant over time, and selective pressure that changes over time, possibly instantaneously – we shall refer to the latter as *episodic selection*. Several authors have studied models that allow evolutionary rates to change over time, including models in which the selective pressure is different on different branches [11–14] as well as various models [15–17] in which the rate of evolution at any site may change at any point in time. We are specifically interested in the former type of model, under which rate changes occur simultaneously at a particular set of sites - as would be expected under an external change in selective pressure, *i.e.* episodic selection. This type of selection is applicable to countless real world scenarios that have been studied extensively: examples include the evolution of lysozyme in response to diet changes [18], the adaptation of HIV to different host populations [14], the evolution of the rhodopsin pigment following changes in habitat [19], and the adaptation of HIV-1 [20,21] and Influenza A Virus (IAV) [22]

Author Summary

When exposed to treatment, HIV-1 and other rapidly evolving viruses have the capacity to acquire drug resistance mutations (DRAMs), which limit the efficacy of antivirals. There are a number of experimentally well characterized HIV-1 DRAMs, but many mutations whose roles are not fully understood have also been reported. In this manuscript we construct evolutionary models that identify the locations and targets of mutations conferring resistance to antiretrovirals from viral sequences sampled from treated and untreated individuals. While the evolution of drug resistance is a classic example of natural selection, existing analyses fail to detect the majority of DRAMs. We show that, in order to identify resistance mutations from sequence data, it is necessary to recognize that in this case natural selection is both episodic (it only operates when the virus is exposed to the drugs) and directional (only mutations to a particular amino-acid confer resistance while allowing the virus to continue replicating). The new class of models that allow for the episodic and directional nature of adaptive evolution performs very well at recovering known DRAMs, can be useful at identifying unknown resistance-associated mutations, and is generally applicable to a variety of biological scenarios where similar selective forces are at play.

genes following zoonosis events. For a review on the evidence for episodic selection in large numbers of protein sequences, see [23].

Here, we consider the evolution of drug resistance in HIV-1 following the treatment of a subset of the host population. We expect that selective pressure will be both episodic, with drug-induced adaptive amino acid changes occurring only in patients receiving therapy, and directional, with site-specific target residues increasing in frequency over time in the treated subset. HIV-1 experiences a variety of other selective pressures, most prominently due to host immune response (e.g. [14,24]), but because such response is nearly unique in each host, we expect that the majority of concerted selective changes in subjects on treatment will be drug-induced.

Previous approaches to detect positive selection driving treatment resistance have had variable success. Crandall *et al.* [25] showed that normalized ratios of non-synonymous to synonymous substitution counts (d_N/d_S) obtained by the counting method of Nei and Gojobori [1] failed to show consistent evidence of selection, despite obvious resistance associated substitutions occurring in parallel in many patients. Chen *et al.* [26] used a contingency-table counting method to characterize positive selection towards specific amino acids in a sample of approximately 40000 sequences. However, their approach ignored the phylogenetic relationships between samples which can cause selection to be conflated with founder effects [22,27]. Lemey *et al.* [28] used the branch-site model of Yang and Nielsen [12] – a model of episodic diversifying selection – to analyze the evolution of drug resistance over a transmission chain. A number of sites were inferred to be under positive selection, of which some were associated with drug resistance. Seoighe *et al.* [8] modeled the evolution of reverse transcriptase between pre- and post-treatment samples from 300 patients. They successfully detected some of the major drug resistance mutations with few false positives.

In this paper we aim to demonstrate that explicitly modeling the directional and episodic character of the evolution of drug resistance increases the power and accuracy to detect drug resistance sites. We introduce a codon-based Model of Episodic Directional Selection (MEDS) and a model of protein evolution called Episodic Directional Evolution of Protein Sequences (EDEPS), and show that both models outperform models that lack either the episodic or directional components.

Models

MEDS

Our codon model of episodic directional selection assumes that branches on the phylogenetic tree can be partitioned into foreground (F) and background (B) subsets *a priori*. Evolution along background branches is described by a standard codon model (Q^B, see below). In the model for foreground branches (Q^F), directional selection is incorporated via an elevated rate of substitutions towards a target amino acid.

MEDS extends two previously proposed models of coding sequence evolution: 1) the episodic component of MEDS is structurally identical to the Internal Fixed Effects Likelihood (IFEL) model proposed in [14], although IFEL is used to detect diversifying selection along internal branches only, and, 2) the directional component is introduced in a manner similar to that in the model of directional selection proposed by Seoighe *et al.* [8]. We used $MG94 \times REV$ [29] as our baseline codon model: it combines a general time-reversible (GTR) model of nucleotide substitution with separate synonymous (α) and non-synonymous (β) rates. To facilitate reading, table 1 summarizes the properties of each model.

Following Seoighe *et al.* [8] we add a directional selection parameter ω_T to modulate the rate of substitutions to the target residue T in the model assigned to foreground branches. If $AA(x)$ represents the amino-acid encoded by codon x, then the instantaneous rates of change between codons i and j ($i \neq j$) are given by:

Table 1. Summary of models described in this manuscript.

Model	Data	Baseline model	Site variation	Lineage variation	Selection test	Citation
MEDS	Codon	MG94 × REV[a]	Fixed effects	Episodic	Directional	This paper
FEEDS	Codon	MG94 × REV	Fixed effects	Episodic	Diversifying	[14][b]
DEPS	Protein	HIV-Between[c]	Random effects	Constant	Directional	[9]
EDEPS	Protein	HIV-Between	Random effects	Episodic	Directional	This paper

[a] [29].

[b] FEEDS has the same structure as a model called IFEL in that paper, but the use here is novel.

[c] [37].

doi:10.1371/journal.pcbi.1002507.t001

$$Q^F_{i,j}=\begin{cases}0,\text{when } i \text{ and } j \text{ differ at more than one position, otherwise :}\\ \theta(i,j)\times\alpha\times\pi_{ij} \text{ when } AA(i)=AA(j)\\ \theta(i,j)\times\beta_F\times\pi_{ij} \text{ when } AA(i)\neq T\&AA(i)\neq T\&AA(i)\neq AA(j)\\ \theta(i,j)\times\beta_F\times\omega_T\times\pi_{ij} \text{ when } AA(i)\neq T\&AA(j)=T\\ \theta(i,j)\times\beta_F\times 1/\omega_T\times\pi_{ij} \text{ when } AA(i)=T\&AA(i)\neq T\end{cases}\quad(1)$$

for the foreground and

$$Q^B_{i,j}=\begin{cases}0,\text{when } i \text{ and } j \text{ differ at more than one position, otherwise :}\\ \theta(i,j)\times\alpha\times\pi_{ij} \text{ when } AA(i)=AA(j)\\ \theta(i,j)\times\beta_B\times\pi_{ij} \text{ when } AA(i)\neq AA(j)\end{cases}\quad(2)$$

for the background branches. We assume that α does not change significantly between foreground and background branches. Indeed, available evidence (e.g. [30–32]) suggests that synonymous rate variation among sites is due to biological processes which change slowly, e.g. RNA secondary structure, transcriptional or translational efficiency, relative to the nearly instant change in the selective environment due to the presence of ARV. In principle, the model can readily handle such variation. β^F and β^B can be inferred independently. $\theta(i,j)$ is the GTR-based rate of the underlying nucleotide substitution from codon i to j, shared between Q^F and Q^B. Equilibrium frequency parameters π_{ij} are derived with the corrected $F3\times4$ estimator [33]. While the same π_{ij} values are used for background and foreground models, when $\omega_T\neq1$ the equilibrium frequencies of Q^F will depart from those dictated by π_{ij}, although we do not need to calculate these new equilibrium frequencies explicitly. This feature is essential because directional evolution changes the character frequencies at a site. We also experimented with a version of the model where the factor $\frac{1}{\omega_T}$ in the last line of Equation 1 was omitted – this had essentially no impact on the results. To ensure that Q defines a valid Markov process generator, along the diagonal of Qwe set:

$$Q_{i,i}=-\sum_{j,j\neq i}Q_{i,j}.\quad(3)$$

Model fitting proceeds in two stages: (a) estimating the parameters shared across sites, and (b) site-wise analysis [6,34]. The branch lengths and Q^F and Q^B, without the directional component (*i.e.* $\omega_T=1$), are first optimized over the entire alignment to obtain gene-wide parameter estimates in the presence of potentially ubiquitous purifying or diversifying selection. The nucleotide rate parameters ($\theta(i,j)$) and relative branch lengths are then fixed for subsequent analyses. From then, the analysis proceeds site by site. We define the null model by setting $\omega_T=1$, a special case of the alternative directional model (ω_T is free to vary), and equivalent to IFEL [14]. The null model has 3 free parameters per site: α (synonymous substitution rate), β^F (non-synonymous substitution rate along foreground lineages) and β^B (non-synonymous substitution rate along backfround lineages). The alternative model has a single additional parameter, ω_T, biasing substitutions towards T. To test for selection towards amino acid T at a specific site, we obtain maximum likelihood scores for the null and alternative models and perform a likelihood-ratio test (LRT) with one degree of freedom based on the asymptotic χ^2 distribution of the likelihood-ratio statistic.

The above test treats nucleotide substitution rates and branch length parameters at a single site as known, even though these are estimated across sites under a simpler model. It is possible that this could affect inference if these estimates were substantially biased. Our simulations suggest that the test performs well in spite of this computational shortcut, and using different models to infer these parameters does not substantially affect the test results on the empirical data we analyze here. Additionally, the χ^2 asymptotic approximation implicit in MEDS relies on the intuition that when the number of sequences increases, the number of branches in the tree will increase, so that substitutions on those branches will constitute different (although dependent) realizations of the process. We note that the asymptotic approximation for our test requires not only many branches but also many foreground branches. While theoretical results justifying our use of the χ^2 approximation are currently lacking, our simulations (see below) suggest that the use of the χ^2 appears to lead to a conservative test for the conditions we have examined.

Scanning a site for selection towards any possible amino acid (T) involves testing 20 hypotheses, and we employ Bonferroni correction [35] to control the site-wise Type I error rate. For computational efficiency, we skip invariant sites and restrict potential values of T to those observed at a given site. Because these reductions are informed by the data, we still employ the 20-test Bonferroni correction at each site.

FEEDS

To assess the importance of the directional component of MEDS, we adapt IFEL to test for episodic diversifying selection along foreground branches and use it as a benchmark for MEDS. As the branches of interest are mostly terminal, the name, IFEL, is no longer appropriate, and we rename the model FEEDS, for 'Fixed Effects Episodic Diversifying Selection'. The alternative model for FEEDS is identical to the null model for MEDS, allowing α, β^F and β^B to vary for each site. To test for non-neutral selection along foreground branches, we set up a null model with $\beta^F=\alpha$, and use an LRT (one degree of freedom) to determine whether the alternative model fits better than the null model. If $\beta^F>\alpha$ results in a significant likelihood improvement, we have evidence for diversifying selection along foreground branches. This test for episodic diversifying selection has three features that distinguish it from the popular branch-site model of Yang and Nielsen [12] and Zhang, Nielsen and Yang [36]: 1) it uses a sitewise likelihood-ratio test [5], otherwise known as a fixed effects likelihood [6] approach, 2) it allows site-to-site synonymous rate variation, which has been shown to be ubiquitous and can cause spurious detection of diversifying selection if ignored [29] and 3) it allows diversifying selection on the background branches in both the null and alternative models. MEDS shares these properties, allowing us to attribute any performance differences specifically to the directional component of MEDS.

DEPS

Throughout the analyses we also compare our results against DEPS (full results in tables S1 to S3), a method for detecting non-episodic directional selection proposed by Kosakovsky Pond *et al.* [9]. DEPS identifies sites with increased substitution rates towards specific amino acids, but it differs from MEDS in three ways: 1) DEPS models directional selection at the amino acid level rather than the codon level, 2) DEPS uses a Random Effects Likelihood (REL) framework to bias selection towards target amino acids across all sites, relying on an empirical Bayes analysis to identify sites of interest and 3) in DEPS, directional selection affects all branches of the phylogeny.

Episodic DEPS

It is a straightforward exercise to modify DEPS to incorporate the episodic nature of MEDS – namely, we restrict accelerated substitutions towards a target residue T (and retard substitutions away from it) to foreground branches, while background branches always evolve according to the baseline protein substitution model specific to HIV-1 [37]. The entire testing framework of DEPS, as described in Kosakovsky Pond *et al.* [9], applies without change. It is well known that amino acid substitution rates depend on the residues involved (e.g. see [38]), and specifying a baseline model which includes unequal substitution rates provides a qualitative advance over MEDS. Conversely, because DEPS works with protein sequences, the natural proxy of approximately neutral evolution (the rate of synonymous substitutions) is not available.

All models and their accompanying LRTs are implemented in a HyPhy Batch Language script [39], and all code and test datasets are available on the MEDS section of the HyPhy wiki (www.hyphy.org) and included in the latest HyPhy distribution (version 2.0020101225 or later).

Datasets

We analyzed three HIV-1 datasets obtained from the South African mirror of the Stanford HIV Drug Resistance Database (HIVdb) [40,41]. Synthetic datasets were generated by simulation to investigate the power and false positive rate of MEDS. The primary goal of this paper is to show that MEDS and EDEPS perform well on medium-sized datasets constructed under a variety of conditions. Every empirical dataset includes sequences sampled from both treated and untreated patients, but we varied the inclusion criteria from one dataset to the next. An ideal dataset for detecting drug resistance would include pre- and post-treatment samples from the same patients (as in our reverse transcriptase dataset), but often such data are not available, e.g. when sequences are obtained from patients experiencing regimen failure. To evaluate the performance of MEDS and EDEPS when pre- and post-treatment sequence pairs were not available (our protease and integrate datasets), we selected pre-treatment sequences using heuristic measures of proximity to the post-treatment samples, as one would be forced to do under such circumstances. Exactly which factors are responsible for performance variation is left as a topic for future research. The objective of each analysis was to detect sites (and corresponding amino acids) that are involved in drug resistance. For validation, we used the curated list of drug resistance associated mutations (DRAMs) which is available from the Stanford HIVdb (http://hivdb.stanford.edu). This list is produced every year and approved by the International AIDS Society (http://www.iasusa.org/resistance_mutations/). These mutations have been rigorously validated with genotype-phenotype and genotype-clinical data and are known to confer varying levels of resistance to one or more antiretroviral agents – they can therefore be used as a ground truth for evaluating the performance of our methods.

We screened each sequence for evidence of recombination (known to have a biasing effect on selection detection, e.g. [42]) using SCUEAL [43] and excluded any sequences showing $>90\%$ support for either inter- or intra-subtype recombination, and using the Rega HIV-1 Subtyping tool Version 2.0 [44], excluding any sequences with clear inter-subtype recombination.

Reverse transcriptase. The first dataset comprises pairs of reverse transcriptase (RT) isolates obtained before and after the initiation of highly active anti-retroviral therapy (HAART) from 241 patients (482 sequences). The data were obtained from the Stanford HIVdb using a query that retrieved paired samples from the same patient, filtered on the earlier sample being Reverse Transcriptase Inhibitor (RTI) naive, and the later sample taken during therapy with at least one Non-Nucleoside RTI (NNRTI) *and* at least one Nucleoside RTI (NRTI). The topology of the phylogeny was estimated using PhyML [45] (settings for all datasets: REV model with tree search by Nearest Neighbor Interchange and Subtree Pruning and Regrafting), and all terminal branches leading to post-treatment sequences were selected as foreground (see Figure S1). As an artifact of older sequencing assays [14], a large number of sequences were missing data at the beginning and end of RT, hence we analyzed the region from codon 40 to 250. Six sequences were excluded from our analyses because they displayed evidence of recombination.

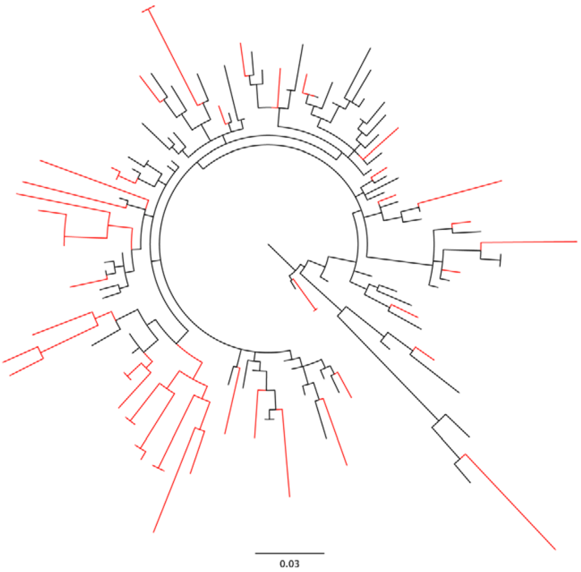

Figure 1. The maximum-likelihood phylogeny for the protease dataset. Foreground branches are marked in red. All terminal foreground branches lead to sequences obtained from patients who had been receiving antiretroviral therapy. See text for details of how we determined which internal branches were assigned to foreground. MEDS and EDEPS allow the presence of a directional component along the foreground branches where antiretroviral therapy exerts selective pressure.
doi:10.1371/journal.pcbi.1002507.g001

Protease. A dataset consisting of 49 protease isolates (from 37 patients), sampled post-Protease Inhibitor (PI) treatment was retrieved from HIVdb (query: Number of PIs = 3, Subtype = C). Additionally, the entire collection of treatment naive protease isolates was obtained, and all full length sequences were searched for two sequences nearest (under the Hamming distance) to each of the 49 post-treatment sequences. The final dataset was constructed by combining the post-treatment and closely related naive sequences: a total of 122 sequences, as some naive sequences were closely related to more than one post-PI sequence. Since protease is only 297 nucleotides long, we were concerned that convergent evolution due to drug resistance might inflate the apparent relatedness between some of the treatment resistant sequences [46], hence we excluded the major resistance sites before reconstructing the phylogeny, using PhyML. As there are many instances where a number of post-treatment sequences were sampled from a single patient, we adopted a recursive branch labeling strategy for the internal branches. All terminal branches leading to post-PI and PI-naive isolates were labeled as foreground and background respectively, and internal branches were labeled

Table 2. Sites under episodic directional and episodic diversifying selection in reverse transcriptase.

Site	Target	MEDS p-value[a]	ω_T (Lower 99% CI)[b]	FEEDS p-value[c]	EDEPS Bayes Factor[d]	Resistance
41	L	0.00259	1937 (280.06)	-[e]	-	NRTI[f]
62	V	-	-	-	313	NRTI
64	K	0.00244	11.99 (5.58)	0.0067	-	
77	L	-	-	-	211	NRTI
98	S	0.00488	>1000 (>1000)	-	-	
100	I	<0.0001	>1000 (524.49)	-	$>10^5$	NNRTI[g]
102	⋆[h]	-	-	0.0025	-	
103	N	<0.0001	629.9 (466.73)	<0.0001	$>10^5$	NNRTI
104	Y	0.00244	>1000 (90.81)	-	-	
115	F	-	-	-	3110	NRTI
116	Y	0.00319	>1000 (179.80)	-	-	NRTI
151	M	<0.0001	>1000 (186.13)	-	$>10^5$	NRTI
151	Q	0.00023	13.96 (7.04)	-	-	
162	S	-	-	-	1772	
165	L	<0.0001	>1000 (>1000)	-	2245	
174	R	-	-	-	105	
181	I	<0.0001	>1000 (118.72)	-	$>10^5$	NNRTI
184	V	<0.0001	25.82 (16.68)	-	$>10^5$	NRTI
188	L	<0.0001	377.93 (32.42)	0.0002	$>10^5$	NNRTI
188	Y	<0.0001	17.61 (11.15)	-	-	
190	S	<0.0001	75.85 (26.09)	-	$>10^5$	NNRTI
200	⋆	-	-	<0.0001	-	
215	F	0.00282	160.65 (10.36)	-	2727	NRTI
215	T	0.00035	15.19 (6.69)	-	-	
228	R	0.00029	72.2 (14.09)	-	1401	NRTI accessory
230	L	0.00297	>1000 (44.6)	-	$>10^5$	NNRTI
245	⋆	-	-	0.0006	-	
286	A	0.00085	>1000 (>1000)	-	-	

[a] MEDS versus FEEDS LRT, testing for directional selection.
[b] the lower bound of the approximate 99% confidence interval calculated from profile likelihood.
[c] $\beta_F > \alpha$ LRT, testing for diversifying selection.
[d] Empirical Bayes analysis, testing for directional selection on protein data.
[e] '-': not significant.
[f] Nucleoside reverse-transcriptase inhibitor.
[g] Non-nucleoside reverse-transcriptase inhibitor.
[h] ⋆: detected only by FEEDS which does not identify a target AA.
doi:10.1371/journal.pcbi.1002507.t002

as foreground if both child branches were foreground, and background otherwise (See figure 1). This labeling ensures that drug resistance selection occurs only on foreground branches. Because there may be portions of foreground branches not under drug selection, the effect of potential mislabeling is to dilute the signal along foreground branches and reduce the power of the test. No sequences showed evidence of recombination.

Integrase. The post-treatment sequences for the final empirical dataset were 83 integrase isolates sampled from 40 patients after Integrase Inhibitor (II, Raltegravir) therapy. 1237 II-naive isolates were obtained from the Stanford HIVdb, and the final Raltegravir dataset was made up of 315 sequences: the 83 post-II isolates, plus the union of the 25 II-naive isolates nearest to each of the 83 post-II isolates under the HKY85 distance [47]. The topology of the phylogeny was again estimated using PhyML, and the foreground region was labeled in the same fashion as the protease dataset (see Figure S2). 20 sequences were excluded for showing evidence of recombination.

Power simulations. We investigated the power of MEDS by simulating alignments over a balanced 64-taxon phylogeny (see Figure S3 for an example). All parameters were varied (see Text S1 for complete details). Of particular interest, we simulated under 4, 8, 16 or 32 foreground branches and, selecting a random target amino acid T for each site, the directional selection parameter ω_T took values of 2, 5, 10, 100 and 1000. These ω_T values are in a reasonable range: in our three empirical datasets, the 25%, 50% and 75% percentiles of the maximum-likelihood estimates of ω_T values for detected substitutions are 32.2, 629.9 and 5544.3. 400 sites were simulated for each ω_T value, for each number of foreground branches, yielding 8000 simulated sites. To assist in understanding the effects of ω_T and the size of the foreground subset, we also recorded the number of

Table 3. Sites under episodic directional and episodic diversifying selection in protease.

Site	Target	MEDS p-value[a]	ω_T (Lower 99% CI)[b]	FEEDS p-value[c]	EDEPS Bayes Factor[d]	Resistance
10	⋆[e]	-[f]	-	0.0005	-	PI[g] accessory
12	T	<0.0001	28.88 (8.58)	-	-	
13	V	0.0059	490.2 (138)	-	145	PI accessory
35	D	0.0035	8.56 (1.99)	-	-	
54	⋆	-	-	0.0026	-	PI
60	E	<0.0001	>1000 (>1000)	-	-	PI accessory
61	E	<0.0001	>1000 (>1000)	-	-	
71	V	-	-	0.0011	257	PI accessory
74	S	0.0007	19.93 (4.08)	0.0013	-	PI accessory
82	A	-	-	<0.0001	$>10^5$	PI
84	V	0.00798	890.3 (248.19)	-	$>10^5$	PI
90	M	<0.0001	>1000 (986.17)	<0.0001	$>10^5$	PI
93	L	0.0078	>1000 (6.36)	-	-	PI accessory

[a] MEDS versus FEEDS LRT, testing for directional selection.
[b] 99% lower confidence interval calculated from the likelihood profile.
[c] $\beta_F > \alpha$ LRT, testing for diversifying selection.
[d] Empirical Bayes analysis, testing for directional selection on protein data.
[e] ⋆: detected only by FEEDS which does not identify a target AA.
[f] '-': not significant.
[g] Protease inhibitor.

doi:10.1371/journal.pcbi.1002507.t003

substitutions towards the target amino acid that occurred along foreground branches.

In real evolving systems, the modeling assumption of selection towards a single target amino acid could be violated. We investigated how such deviations would impact the power of the model by simulating directional selection towards two target amino acids, with substitutions towards one target accelerated on 8 foreground branches, and substitutions towards another accelerated on a different 8 foreground branches. The parameters were varied in the same manner as the single-target power simulation, and 1600 sites were simulated for each ω_T value, again yielding 8000 sites in total.

False positive simulations. We used exactly the same simulation configuration and parameters to asses the rates of false positives under the null model ($\omega_T = 1$). We simulated 400 sites for each of 4, 8, 16 or 32 foreground branches.

In evolving proteins, each site could have its own site-specific selective constraints governing amino acid distributions. MEDS assumes that background equilibrium frequencies are the same for all sites, and a potential concern is that deviations from this

Table 4. Sites under episodic directional and episodic diversifying selection in integrase.

Site	Target	MEDS p-value[a]	ω_T (Lower 99% CI)[b]	FEEDS p-value[c]	EDEPS Bayes Factor[d]	Resistance
72	I	0.0095	>1000 (533.76)	-[e]	-	INI[f] accessory
97	A	0.0028	337 (105.52)	<0.0001	$>10^5$	INI accessory
140	S	<0.0001	>1000 (>1000)	0.0003	$>10^5$	INI
143	R	0.0015	23.5 (3.83)	<0.0001	$>10^5$	INI
148	H	<0.0001	35.5 (14.53)	<0.0001	$>10^5$	INI
155	H	<0.0001	>1000 (>1000)	0.0006	$>10^5$	INI
163	R	-	-	-	1143	INI accessory
221	Q	-	-	-	107	
227	⋆[g]	-	-	0.0064	-	
230	⋆	-	-	0.0048	-	INI accessory

[a] MEDS versus FEEDS LRT, testing for directional selection.
[b] 99% lower confidence interval calculated from the likelihood profile.
[c] $\beta_F > \alpha$ LRT, testing for diversifying selection.
[d] Empirical Bayes analysis, testing for directional selection on protein data.
[e] '-': not significant.
[f] Integrase inhibitor.
[g] ⋆: detected only by FEEDS which does not identify a target AA.

doi:10.1371/journal.pcbi.1002507.t004

modeling assumption could lead to excessive false positives. To investigate this, we simulated data under a version of the null model where each site's amino acid equilibrium frequencies were sampled from a symmetric Dirichlet distribution with density

$$f(q_p;\alpha)\sim \prod_{p=1}^{20} q_p^{\alpha-1} \tag{4}$$

The concentration parameter α took values 0.005, 0.05, 0.5 and 5, varying the equilibrium frequency distributions from extremely peaked to relatively flat. Each sampled amino-acid frequency q_p was evenly distributed among all codons encoding p and a version of Q^B with the Goldman-Yang parameterization of equilibrium frequencies [4] was employed to simulate codon sequence data.

Results

Reverse transcriptase

MEDS detected twenty substitutions at seventeen sites under significant directional selection at $p \leq 0.01$, after correcting for multiple tests (see tables 2 and S4). Of these, five are known NRTI drug resistance associated mutations (DRAMs) (41L, 116Y, 151M, 184V and 215F) and six are known NNRTI DRAMs (100I, 103N, 181I, 188L, 190S and 230L). Additionally, 228R is listed as an accessory NRTI mutation. The eight detected substitutions that have not been experimentally or clinically associated with drug resistance are 64K, 98S, 104Y, 151Q, 165L, 188Y, 215T and 286A. Interestingly, at three of these sites (151, 188 and 215) selection was detected both towards the wildtype and towards resistant residues. EDEPS agreed with MEDS on eleven sites, detected additional DRAMs 62V, 77L and 115F, missed two MEDS-reported DRAMs (41L and 116Y), and found episodic selection at 162S and 174R which are not known to confer drug resistance.

Remarkably, FEEDS detected only six sites under diversifying selection (table S5), two of which are known resistance mutations, strongly supporting the inclusion of a directional component in the model. A continuous directional selection model (DEPS) detected 46 sites under directional selection with Bayes factors >100 (see table S1), only ten of which are on the HIVdb list. This indicates that focusing our attention on branches where the evolutionary environment shifts is advantageous for finding evidence of adaptive response to such shifts.

Protease

MEDS detected nine substitutions under directional selection at $p \leq 0.01$ (tables 3 and S6). Of these, two are major DRAMs (90M and 84V). Three are accessory polymorphic mutations (13V, 60E and 93L) under selective pressure from the drugs. 74S is a non-polymorphic accessory mutation. EDEPS agreed with MEDS on three (13V, 84V and 90M), detected one more major mutation, 82A, and an accessory mutation at 71V. Interestingly, 60E and 61E found by MEDS involve substitutions ($D \rightarrow E$ and $Q \rightarrow E$) which, in HIV, are much more frequent than the mean substitution rate [37]. Because MEDS sets the background rate of non-synonymous substitutions to the same value for all pairs of residues, it could use ω_T to compensate for the *overall* underestimation of rates that are much greater than the mean rate.

FEEDS identified six sites involved in diversifying selection (table S7), with all six listed on HIVdb. In addition to two sites already detected by MEDS (74 and 90), sites 10 and 71 are listed as accessory mutations, while 54 and 82 are major resistance mutations. DEPS appeared to be much more conservative on this dataset, detecting four sites under directional selection, two of which are listed on HIVdb (see table S2).

Integrase

MEDS detected six substitutions under significant directional selection at the 1% level (see tables 4 and S8). Four (140S, 143R, 148H and 155H) appear on the HIVdb list of mutations associated with a $>5-10$ fold decrease in Raltegravir susceptibility. Two are listed as mutations selected by Raltegravir (72I and 97A). EDEPS confirmed five DRAMs (97A, 140S, 143R, 148H and 155H), together with a 163R accessory substitution and a 221Q mutation which is not a known DRAM.

FEEDS found seven sites under diversifying selection (table S9), six of which are known resistance mutations. 230 is the only correctly identified resistance site in the integrase dataset that is detected as being under diversifying selection by FEEDS, but not directional selection by MEDS. 230 R and N are listed as selected by Raltegravir. DEPS detected 39 substitutions under directional selection (see table S3), nine of which appear on the HIVdb list.

Comparing methods

Comparing the fit of FEEDS and MEDS on *known* resistance sites in all three datasets, LRTs reject a null model of FEEDS in favor of MEDS on 24 sites, with FEEDS being favored on five (four from protease and one from integrase). Note that FEEDS might still be useful for detecting these sites, but the LRT demonstrates that MEDS is a better model of the process. This suggests that episodic directional selection is, in most cases, a better characterization of the evolution of drug resistance. Overall, FEEDS detects fourteen true positives, while MEDS and EDEPS detect 24 each (although not the same 24). Where FEEDS appears to have a reasonably low rate of false positives but misses a large number of true positives, DEPS detects a large number of true positives but with a very high false positive rate. This is expected as DEPS will detect substitutions under selection along background branches that are not related to drug resistance.

Power simulations

The power of MEDS, like that of other codon methods, strongly depends on the information content of the sequences, specifically on the number of times that substitutions toward the target occur along the foreground lineages. For example, even when ω_T is 1000, no substitutions towards T occur on half the sites simulated on the phylogeny with sixteen foreground branches. The primary reason for this is that ω_T affects only the instantaneous substitution rate from a codon to its direct neighbors; if none of the direct

Table 5. Single target power simulations: power as a function of ω_T.

# FG branches	ω_T				
	2	5	10	100	1000
4	0 (8)[a]	0 (16)	0 (37)	0.31 (110)	0.79 (155)
8	0 (11)	0 (18)	0.04 (62)	0.51 (129)	0.73 (170)
16	0 (31)	0.018 (54)	0.036 (83)	0.59 (177)	0.71 (201)
32	0.02 (62)	0.03 (71)	0.16 (116)	0.68 (223)	0.80 (282)

[a] Numbers in brackets are the number of times at least one substitution towards the target occurred along foreground branches: *i.e.* the denominator for the proportion of detections.

doi:10.1371/journal.pcbi.1002507.t005

Table 6. Single target power simulations: power as a function of number of substitutions to target AA along foreground branches, pooling over ω_T.

# FG branches	# substitutions to target AA					
	0	1	2	3	4	≥5
4	0 (1674)[a]	0 (119)	0.2 (58)	0.77 (48)	0.99 (111)	N/A
8	0 (1610)	0 (146)	0.23 (53)	0.69 (26)	1 (21)	0.99 (144)
16	0 (1454)	0 (200)	0.34 (92)	0.49 (39)	0.79 (34)	0.97 (181)
32	0 (1246)	0.03 (234)	0.4 (107)	0.41 (70)	0.70 (46)	0.97 (297)

[a]Numbers in brackets are the number of times that many substitutions towards the target occurred along foreground branches: *i.e. the denominator for the proportion of detections*.

doi:10.1371/journal.pcbi.1002507.t006

neighbors of T are visited along a foreground branch, a change in ω_T will not affect the process.

Hence, we tabulate MEDS results only for sites with at least one substitution towards the target on any foreground branch. Table 5 shows that the power is positively correlated with ω_T. MEDS appears to be quite powerful, even when the number of foreground branches is small, achieving, for example, 51% power with $\omega_T = 100$ with only eight foreground branches. Table 6 displays the power of MEDS conditioned on the number of substitutions towards the target on foreground branches. With only one substitution there is almost no power, but moderate power ($\approx 30\%$) occurs with two substitutions towards T, and with five or more substitutions towards T, the power is almost 100%.

For data simulated with two target residues, each on eight foreground branches, the occurrence of at least one substitution towards *both* targets is infrequent. From 4800 sites simulated with ω_T values of 2, 5 and 10, this occurs only 58 times, and is never detected. From 1600 sites simulated with $\omega_T = 100$ for both targets, substitutions to both targets occur 174 times. MEDS detects substitutions to at least one target in 47% of such sites, but only detects substitutions to both targets in 5% of such sites. With $\omega_T = 1000$, we see 306 of 1600 sites with substitutions to both targets, and MEDS detects substitutions to at least one target in 86% of these sites, and to both targets in 31%.

Table 7 shows how the power increases with the number of substitutions towards both targets on the foreground branches. Since there too many possible combinations and too few observations, we display power in a cumulative manner (*i.e.* $\geq N$ substitutions towards both targets).

False positive simulations

MEDS behaves conservatively. With data simulated under the null model, far fewer sites are identified as under episodic directional selection than would be expected from the nominal p-value thresholds. Across all four foreground configurations, only one false positive detection ($p < 0.01$, with Bonferroni correction) occurs on the 32 foreground branch phylogeny, and none on the others. With $p < 0.05$, with 4, 8, 16 and 32 branches, we have false positive rates of 0, 0.0025, 0.0075 and 0.01; with $p < 0.1$, we have 0.005, 0.005, 0.0125 and 0.02, respectively. This is most likely due to FEL methods being generally conservative [6] as well as the conservative nature of Bonferroni correction. The effect of the correction is compounded because increasing the frequency of one amino acid reduces the frequency of the others, and thus the twenty tests are not independent. Table 8 shows the false positive rate for alignments simulated under site specific equilibrium frequencies. MEDS is still conservative under this scenario, and the false positive rates do not appear to be influenced by the concentration parameter.

Table 8. False positives with site specific equilibrium frequencies as a function of the concentration parameter α and the nominal p-value of the test.

α parameter:	0.005	0.05	0.5	5
$p = 0.01$	0.005	0.0025	0.0025	0.0075
$p = 0.05$	0.02	0.0175	0.02	0.015
$p = 0.1$	0.0325	0.0325	0.035	0.0375

doi:10.1371/journal.pcbi.1002507.t008

Discussion

We have proposed a codon (MEDS) and a protein (EDEPS) model of episodic directional selection, and demonstrated their performance on three HIV-1 datasets, where drug-induced directional episodic selection is expected to operate. We have also proposed a model of episodic diversifying selection (FEEDS), to rigorously evaluate the importance of modeling the directional component of natural selection. As expected, on all datasets, our episodic directional tests strongly outperform a test for continuous directional selection (DEPS) for detecting drug resistance sites. The assumptions of DEPS are inappropriate for the analysis of episodic selection, where selection is limited to specific regions of the phylogeny, because DEPS assumes uniform selection over the whole phylogeny. This serves as a caution against using suboptimal models, rather than a criticism of DEPS.

We tested MEDS with extensive simulations. MEDS is a conservative test, even when strong constraints on the amino acid state space are introduced in the form of site-specific equilibrium frequencies. Under the alternative model, good power is achieved even when relatively few substitutions towards target amino acids take place along foreground branches. When we deviate from the alternative model and elevate the substitution rate towards several target residues, the power to detect both targets is lower than it would be assuming independence. This reduction in power is

Table 7. Dual target power simulations: power as a function of number of substitutions to two target AAs.

Substitutions to both targets[a]:	≥1	≥2	≥3	≥4	≥5	≥6	≥7	=8
MEDS detects at least one target:	0.64	0.81	0.89	0.92	0.95	0.98	1	1
MEDS detects both targets:	0.19	0.36	0.48	0.52	0.63	0.76	0.78	0.81
Total sites:	538	288	214	179	132	99	69	32

[a]Substitutions along foreground branches. Each target has 8 foreground branches along which changes towards it were accelerated.

doi:10.1371/journal.pcbi.1002507.t007

expected: as the number of targets along foreground branches increases, the directional nature of the process is lost.

Hughes [48] argues that diversifying selection is only appropriate for modeling pathogen-host co-evolution, and that the constantly shifting environment is required for the standard diversifying selection model to be appropriate. Our results highlight that models of diversifying selection also serve as reasonable approximations in instances where selective constraints allow escape to many different residues, such as codon 54 in protease, which has V, T, A, L and M as major drug resistant residues. However, at most sites conferring drug resistance, directional models better approximate reality – positive selection acts only on one or a few specific mutations, while the rest are suppressed by purifying selection. The simulations presented in Table 7 illustrate how much power MEDS can be expected to have in cases such as site 54 in protease. This example also suggests a future extension of MEDS, where instead of considering one target residue at a time, substitution rates could be elevated towards *classes* of target residues.

Another interesting property of directional models is exemplified by a substitution in the protease dataset. 93L is a polymorphic mutation selected for by protease inhibitors. Despite L already being the most common residue in subtype C, the model detects selective pressure towards it – the proportion of L residues is indeed lower in nave sequences. At the population level this appears as purifying selection: the most common amino acid increases in frequency. This is nevertheless detected by our test. Far from being problematic, such information could be useful for directing treatment, if it turns out that patients with I at position 93 are more susceptible to PI therapy. Such observations should, of course, be directly verified with clinical data.

There are clear differences in organism-wide amino acid exchangeabilities in HIV-1 [37], yet the null model of MEDS (and the vast majority of other codon-models) posit that the non-synonymous substitution rate does not depend on the residues. We evaluated the effect of this assumption by comparing MEDS with an episodic version of DEPS – a test that specifically incorporates a heterogeneous exchangeability matrix in the evolutionary model. With a few exceptions, MEDS and EDEPS return overlapping sets of directionally evolving residues and identify the same targets. There are several sites in protease and integrase, where MEDS may be misclassifying non-uniform exchangeabilities as directional selection, hence another extension of MEDS would be to incorporate multiple non-synonymous substitution rates [38].

MEDS and EDEPS were designed with HIV-1 drug resistance in mind, but should be applicable wherever episodic directional selection occurs along multiple lineages. To use the models, two specific conditions must be met: 1) Lineages expected to be under directional selection must be known *a priori*, at least approximately. This is necessary to partition the phylogeny into foreground and background regions. 2) A rich collection of background sequences are needed. With HIV-1, this translates to requiring treatment naive sequences. Variety in these sequences is also important. If all the background sequences were so closely related that the foreground and background regions were separated by a single branch, if would be difficult to separate directional selection from founder effects, which would result in a loss of power. If the background sequences are spread about the phylogeny, however, founder effects are rendered unlikely and the test for directional selection should be well powered.

With HIV-1 drug resistance datasets, the foreground labeling strategy might prove important. On the RT dataset, branch-labeling was straightforward, as we had access to pre-treatment sequences for each patient. This is not the case for most real-world datasets, and other approximate labeling schemes, as well as the robustness of the results to these differences, should be investigated.

Another consideration is the rooting of the tree. With directional models, the expected amino acid frequencies change across the phylogeny, and the position of the root becomes important [9]. With MEDS and EDEPS, the directional component only affects foreground branches. Consequently, the tree can be rooted on any background branch and the likelihood will be unaffected [49].

Amidst growing concerns about an epidemic of circulating drug resistant HIV-1, the WHO and SATuRN are recommending a scale-up of drug resistance surveillance [41,50]. This is to ensure the long-term success of the world's largest antiretroviral treatment programs, located in Africa. We see improved models of the sequence evolution playing a role in characterizing local differences in treatment resistance patterns, perhaps driven by different treatment regimens, adherence and transmission dynamics, and possibly identifying new resistance mutations.

Supporting Information

Figure S1 The maximum-likelihood phylogeny for the reverse transcriptase dataset. Foreground branches are marked in red. All terminal foreground branches lead to sequences obtained from patients who had been receiving antiretroviral therapy.
(PDF)

Figure S2 The maximum-likelihood phylogeny for the integrase dataset. Foreground branches are marked in red. All terminal foreground branches lead to sequences obtained from patients who had been receiving antiretroviral therapy.
(PDF)

Figure S3 A balanced phylogeny used for simulations. Foreground branches are marked in red. See Text S1 for further simulation details.
(PDF)

Table S1 Reverse transcriptase results - DEPS.
(PDF)

Table S2 Protease results - DEPS.
(PDF)

Table S3 Integrase results - DEPS.
(PDF)

Table S4 Reverse Transcriptase - MEDS: Maximum likelihood parameter values for the test for episodic directional selection.
(PDF)

Table S5 Reverse Transcriptase - FEEDS: Maximum likelihood parameter values for the test for episodic diversifying selection.
(PDF)

Table S6 Protease - MEDS: Maximum likelihood parameter values for the test for episodic directional selection.
(PDF)

Table S7 Protease - FEEDS: Maximum likelihood parameter values for the test for episodic diversifying selection.
(PDF)

Table S8 Integrase - MEDS: Maximum likelihood parameter values for the test for episodic directional selection.
(PDF)

Table S9 Integrase - FEEDS: Maximum likelihood parameter values for the test for episodic diversifying selection.
(PDF)

Text S1 Simulation details. The variation in nuisance parameters used for our simulations.
(PDF)

References

1. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.
2. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the adh locus in drosophila. Nature 351: 652–654.
3. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715–724.
4. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding dna sequences. Mol Biol Evol 11: 725–736.
5. Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. Genetics 169: 1753–1762.
6. Kosakovsky Pond SL, Frost SDW (2005) Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208–1222.
7. Delport W, Scheffler K, Seoighe C (2009) Models of coding sequence evolution. Brief Bioinform 10: 97–109.
8. Seoighe C, Ketwaroo F, Pillay V, Scheffler K, Wood N, et al. (2007) A model of directional selection applied to the evolution of drug resistance in hiv-1. Mol Biol Evol 24: 1025–1031.
9. Kosakovsky Pond SL, Poon AFY, Leigh Brown AJ, Frost SDW (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza a virus. Mol Biol Evol 25: 1809–1824.
10. Anisimova M, Kosiol C (2009) Investigating Protein-Coding Sequence Evolution with Probabilistic Codon Substitution Models. Mol Biol Evol 26: 255–271.
11. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568–573.
12. Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908–917.
13. Kosakovsky Pond SL, Frost SDW (2005) A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol 22: 478–485.
14. Kosakovsky Pond SL, Frost SDW, Grossman Z, Gravenor MB, Richman DD, et al. (2006) Adaptation to different human populations by hiv-1 revealed by codon-based analyses. PLoS Comput Biol 2: e62+.
15. Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15: 1647–57.
16. Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18: 866–873.
17. Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP (2004) Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci U S A 101: 12957–12962.
18. Messier W, Stewart CB (1997) Episodic adaptive evolution of primate lysozymes. Nature 385: 151–154.
19. Yokoyama S, Tada T, Zhang H, Britt L (2008) Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proc Natl Acad Sci U S A 105: 13480–13485.
20. Wain LV, Bailes E, Bibollet-Ruche F, Decker JM, Keele BF, et al. (2007) Adaptation of HIV-1 to its human host. Mol Biol Evol 24: 1853–60.
21. Ngandu N, Seoighe C, Scheffler K (2009) Evidence of hiv-1 adaptation to host hla alleles following chimp-to-human transmission. Virol J 6: 164+.
22. Tamuri AU, dos Reis M, Hay AJ, Goldstein RA (2009) Identifying changes in selective constraints: Host shifts in influenza. PLoS Comput Biol 5: e1000564+.
23. Studer RA, Robinson-Rechavi M (2009) Evidence for an episodic model of protein sequence evolution. Biochem Soc T 37: 783–786.
24. Frost SDW, Wrin T, Smith DM, Kosakovsky Pond SL, Liu Y, et al. (2005) Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection. Proc Natl Acad Sci U S A 102: 18514–9.
25. Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP (1999) Parallel evolution of drug resistance in hiv: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol Biol Evol 16: 372–382.
26. Chen L, Perlina A, Lee CJ (2004) Positive selection detection in 40,000 human immunodeficiency virus (hiv) type 1 sequences automatically identifies drug resistance and positive fitness mutations in hiv protease and reverse transcriptase. J Virol 78: 3722–3732.
27. Felsenstein JJ (1985) Phylogenies and the comparative method. Am Nat 125: 1–15.
28. Lemey P, Derdelinckx I, Rambaut A, Van Laethem K, Dumont S, et al. (2005) Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. J Virol 79: 11981–11989.
29. Kosakovsky Pond SL, Muse SV (2005) Site-to-site variation of synonymous substitution rates. Mol Biol Evol 22: 2375–2385.
30. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7: 98–108.
31. Zhou T, Gu W, Wilke CO (2010) Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol 27: 1912–22.
32. Sanjuán R, Bordería AV (2011) Interplay between RNA structure and protein evolution in HIV-1. Mol Biol Evol 28: 1333–8.
33. Kosakovsky Pond S, Delport W, Muse SV, Scheffler K (2010) Correcting the Bias of Empirical Frequency Parameter Estimators in Codon Models. PLoS ONE 5: e11230+.
34. Yang Z (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus a. J Mol Evol 51: 423–32.
35. Rice WR (1989) Analyzing tables of statistical tests. Evolution 43: 223–225.
36. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472–2479.
37. Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, et al. (2007) Hiv-specific probabilistic models of protein evolution. PLoS One 2: e503.
38. Delport W, Scheffler K, Botha G, Gravenor MB, Muse SV, et al. (2010) Codontest: modeling amino acid substitution preferences in coding sequences. PLoS Comput Biol 6: e1000885.
39. Kosakovsky Pond SL, Frost SDW, Muse SV (2005) Hyphy: hypothesis testing using phylogenies. Bioinformatics 21: 676–679.
40. Rhee SYY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, et al. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res 31: 298–303.
41. de Oliveira T, Shafer RW, Seebregts C (2010) Public database for HIV drug resistance in southern Africa. Nature 464: 673.
42. Scheffler K, Martin DP, Seoighe C (2006) Robust inference of positive selection from recombining coding sequences. Bioinformatics 22: 2493–9.
43. Kosakovsky Pond SL, Posada D, Stawiski E, Chappey C, Poon AFY, et al. (2009) An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in hiv-1. PLoS Comput Biol 5: e1000581.
44. Alcantara LCJC, Cassol S, Libin P, Deforche K, Pybus OG, et al. (2009) A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. Nucleic Acids Res 37: W634–42.
45. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol 59: 307–321.
46. Doolittle RF (1994) Convergent evolution: the need to be explicit. Trends Biochem Sci 19: 15–18.
47. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174.
48. Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity 99: 364–373.
49. Lacerda M, Scheffler K, Seoighe C (2010) Epitope discovery with phylogenetic hidden markov models. Mol Biol Evol 27: 1212–1220.
50. Jordan MR, Bennett DE, Bertagnolio S, Gilks CF, Sutherland D (2008) World health organization surveys to monitor hiv drug resistance prevention and associated factors in sentinel antiretroviral treatment sites. Antivir Ther 13 Suppl 2: 15–23.

Author Contributions

Conceived and designed the experiments: BM SLKP KS. Performed the experiments: BM SLKP. Analyzed the data: BM SLKP KS. Wrote the paper: BM TdO CS SLKP KS.

FUBAR : An efficient analysis of the molecular footprint of natural selection

Ben Murrell, Sasha Moola, Amandla Mabona, Thomas Weighill, Daniel Sheward
Sergei L. Kosakovsky Pond and Konrad Scheffler*

August 1, 2012

Abstract

Model-based selection analyses (such as those performed by PAML and HyPhy) often model the site-to-site variation in selection parameters as a random effect. Due to computational limitations, these methods are restricted to using a relatively small number of discrete rate categories, placing unrealistic constraints on the distribution of selection parameters over sites. Such methods are also prohibitively slow for large alignments. We present an approximate hierarchical Bayesian method that allows rich, flexible site-to-site variation, which improves the statistical performance of the method, while still detecting selection much faster than current methods.

By exploiting some commonly used approximations, our Fast Unconstrained Bayesian AppRoximation (FUBAR) can accurately identify positive and purifying selection orders of magnitude faster than existing random effects methods and 3 to 20 times faster than fixed effects methods (with the disparity increasing for larger alignments). We introduce a fast Markov chain Monte Carlo (MCMC) routine that allows a flexible distribution over the selection parameters to be specified, with no parametric constraints on the shape of this distribution. This flexible distribution allows information to be shared between sites, yielding greater power to detect positive selection than that of fixed effects methods, but without the potential bias introduced by the overly restrictive distributions used by current random effects models. We demonstrate the utility of these computational speedups by analyzing selection on a large influenza haemagglutinin dataset (3142 sequences).

FUBAR is available as a batch file within the latest HyPhy distribution, as well as on the Datamonkey web server (`http://www.datamonkey.org/`)

1 Introduction

When natural selection has driven adaptive change throughout the evolutionary history of a group of related organisms, a detectable trace of this adaptation may be left upon their protein coding sequences, courtesy of the structure of the genetic code: the relative rate of non-synonymous substitutions at some sites may be inflated beyond that of the synonymous rate. This is called positive selection. Conversely, purifying selection - where mutations that modify the protein are less likely to go to fixation - causes the non-synonymous rate to be smaller than the synonymous rate. While purifying selection is pervasive (conserving useful protein structures), positive selection is usually more interesting, pointing to host-pathogen co-evolution (Hughes and Nei, 1988) or adaptation to environmental changes (Messier and Stewart, 1997).

This paper revisits the problem of site-to-site variation in selection intensity. Two codon model-based (Muse and Gaut, 1994; Goldman and Yang, 1994) techniques have dominated the literature: fixed

[*] to whom correspondence should be addressed

and random effects. Fixed effects approaches (Kosakovsky Pond and Frost, 2005) to detecting sites evolving under selection make no assumptions about how the selection coefficients α (the synonymous rate) and β (the non-synonymous rate) are distributed over sites, having two free parameters per site. Computationally efficient estimation of site-specific α and β parameters requires that the topology and branch proportions of the tree be fixed in advance, usually estimated under a simpler model.

Random effects models (Nielsen and Yang, 1998; Pond and Muse, 2005), on the other hand, explicitly model a distribution over α and β, under the assumption that each site is an independent and identically distributed draw from this distribution. The parameters that control the distribution are shared across all sites, so the model uses a smaller number of parameters than fixed effects models. This comes at the cost of flexibility, since a form for this distribution must be specified. For computational tractability, discrete distributions are used, where α and β can take a small number of values. The computational complexity of the likelihood calculation increases linearly with the number of categories, so the discretization is often quite coarse. As we show in section 3.3, overly coarse distributions can mislead inference.

On the one hand, fixed effects models make no assumptions about the distribution of the selection parameters over sites, so α and β can take any value, but evidence from one site cannot inform our expectations regarding another. On the other hand, random effects models - which do allow such sharing of information between sites - are forced, by computational considerations, to make overly restrictive assumptions and overly coarse discretizations of the parameter space, so α and β are restricted to be one of a small number of categories. Where the restrictive assumptions are relaxed, a substantial computational cost is incurred (Huelsenbeck et al., 2006).

We propose FUBAR (a Fast Unconstrained Bayesian AppRoximation), which exploits a collection of computational shortcuts to speed up the detection of positive (or purifying) selection. Like the fixed effects method of Kosakovsky Pond and Frost (2005), FUBAR estimates the branch proportion parameters and nucleotide substitution rates using a nucleotide model, and fixes these parameters for the subsequent selection analysis. The key contribution of FUBAR is that, with these parameters estimated in advance, one can efficiently precompute a dense “grid” of conditional likelihoods - the probability of the data at each site given particular values for α and β. Approximate site specific inference - whether frequentist likelihood ratio tests, random effects empirical Bayes, or fully hierarchical Bayes - can then be performed efficiently on this grid of precomputed conditional likelihoods.

With the substantial computational saving afforded by a precomputed grid, far more complex models of site-to-site variation may be considered. The default settings for FUBAR, for example, recommend 400 categories for α, β, compared to 9 for the random effects approach in Pond and Muse (2005). To handle the statistical complexity required by a models with a large number of categories, we employ a hierarchical Bayesian approach with a Dirichlet hyperprior over $p(\alpha, \beta)$, and we integrate over the uncertainty in $p(\alpha, \beta)$ using Markov Chain Monte Carlo (MCMC). This hierarchical Bayesian approach should be far less vulnerable to model misspecification than approaches that assume a small number of categories or a parametric form for the distribution of α, β, while at the same time avoiding over-parameterization, since the uncertainty in the parameters is marginalized out.

Despite this model complexity, the precomputed conditional likelihood grid allows FUBAR to run orders of magnitude faster than current fixed and random effects approaches. With such a large number of α, β categories, FUBAR combines the site-to-site flexibility enjoyed by fixed effects methods with the ability to share information across sites enjoyed by random effects methods, at a fraction of the computational cost of either. This allows the analysis of very large datasets that would otherwise be intractable using model based methods. See table 1 for a summary of the trade-offs and gains made by FUBAR compared to other methods: fixed effects, exemplified by FEL, HyPhy’s two rate fixed effects method (Kosakovsky Pond and Frost, 2005), and random effects, exemplified by REL, HyPhy’s two rate random effects method (Pond and Muse, 2005).

Table 1: Trade-offs made by different approaches to accommodate variation from site to site.

	Random effects	Fixed effects	FUBAR
Flexible variation of α, β	No	Yes	Yes
Shares information across sites	Yes	No	Yes
$\mathcal{T}$ and $\mathcal{N}$ from full model	Yes	No	No
Computational efficiency	Poor	Moderate	Excellent

2 Methods

Following Muse and Gaut (1994), we model the process for a particular branch at a site as an instantaneous rate matrix, $Q = \{q_{ij}\}$, with elements that describe the rate of substitution of codon i with codon j:

$$q_{ij}(\alpha, \beta, \Pi, \mathcal{N}) = \begin{cases} \alpha \pi_{ij} n_{ij}, & \delta(i,j) = 1,\ AA(i) = AA(j), \\ \beta \pi_{ij} n_{ij}, & \delta(i,j) = 1,\ AA(i) \neq AA(j), \\ 0, & \delta(i,j) > 1, \\ -\sum_{k \neq i} q_{ik}, & i = j. \end{cases} \tag{1}$$

$\delta(i,j)$ counts the number of nucleotide differences between codons i and j. α and β parameterize the rates of synonymous and non-synonymous substitutions respectively. n_{ij} (comprising $\mathcal{N}$) are the nucleotide mutational biases, which we model using the 5-parameter general time reversible nucleotide model (GTR; Tavaré, 1986). π_{ij} (comprising Π) denote the equilibrium frequency parameters.

We denote a phylogenetic tree $\mathcal{T}$, specifying both the tree topology and, for every branch b, a branch length parameter, t_b. The probability of changing from codon i to j at a site along branch b in time t_b, is recorded in the corresponding element of the transition matrix e^{Qt_b}. The likelihood of observing the site given the model parameters is calculated using the Felsenstein's pruning algorithm (Felsenstein, 1981). The goal of a selection analysis is to infer values for α and β for each site, and provide a measure of evidence for the hypotheses that $\alpha > \beta$ or $\alpha < \beta$.

2.1 Random Effects models

2.1.1 Calculating the likelihood.

The model used by FUBAR requires that the synonymous and non-synonymous rates vary across sites. To achieve this, we follow Pond and Muse (2005) and treat α and β as random effects, specifying a distribution from which they are drawn, and we integrate over that distribution to calculate the marginal likelihoods. For computational tractability these distributions are discrete. Furthermore, the sites are assumed to evolve independently, with the overall likelihood being the product of the site likelihoods. Thus, if x_i denotes the i^{th} site of the alignment X, then the overall likelihood can be calculated as

$$p(X|\mathcal{T}, \Pi, \mathcal{N}, \theta) = \prod_i \sum_{\alpha,\beta} p(x_i|\alpha, \beta, \mathcal{T}, \Pi, \mathcal{N}) p(\alpha, \beta|\theta) \tag{2}$$

where $p(\alpha, \beta|\theta)$ specifies the probability of each (α, β) combination, and θ is a set of parameters governing this distribution.

2.1.2 Site-specific inference.

If we had a fixed phylogeny with branch lengths, $\hat{\mathcal{T}}$, fixed equilibrium frequencies, $\hat{\Pi}$, fixed nucleotide rates, $\hat{\mathcal{N}}$, as well as a fixed prior distribution, $p(\alpha, \beta|\hat{\theta})$, then we could calculate the site-specific posteriors using Bayes theorem:

$$p(\alpha, \beta | x_i, \hat{\mathcal{T}}, \hat{\Pi}, \hat{\mathcal{N}}, \hat{\theta}) = \frac{p(x_i | \alpha, \beta, \hat{\mathcal{T}}, \hat{\Pi}, \hat{\mathcal{N}}) p(\alpha, \beta | \hat{\theta})}{\sum_{\alpha,\beta} p(x_i | \alpha, \beta, \hat{\mathcal{T}}, \hat{\Pi}, \hat{\mathcal{N}}) p(\alpha, \beta | \hat{\theta})} \tag{3}$$

The posterior probability that positive selection occurred at a site is the total probability where $\beta > \alpha$:

$$p(\beta > \alpha | x_i, \hat{\mathcal{T}}, \hat{\Pi}, \hat{\mathcal{N}}, \hat{\theta}) = \sum_{\beta > \alpha} p(\alpha, \beta | x_i, \hat{\mathcal{T}}, \hat{\Pi}, \hat{\mathcal{N}}, \hat{\theta}) \tag{4}$$

and Bayes factors can be calculated straightforwardly:

$$BF(i) = \frac{\frac{p(\beta > \alpha | x_i, \hat{\theta})}{1 - p(\beta > \alpha | x_i, \hat{\theta})}}{\frac{p(\beta > \alpha | \hat{\theta})}{1 - p(\beta > \alpha | \hat{\theta})}} \tag{5}$$

The empirical Bayes approach would optimize the likelihood function, obtaining maximum likelihood estimates of all parameter values. Evidence for selection at each site is then calculated using the Bayesian inference machinery, ignoring uncertainty in the parameter estimates. FUBAR seeks to estimate models with a large number of α and β categories, and with no constraints on the shape of the distribution over these categories, which renders empirical Bayes both computationally intractable and statistically unsound. FUBAR thus turns to approximate hierarchical Bayes, exploiting heuristics to make the estimation of such complex models not just computationally tractable and statistically robust, but fast as well.

2.2 Recycling conditional likelihoods

To prevent having to recalculate the conditional likelihoods, $p(x_i | \alpha, \beta, \mathcal{T}, \Pi, \mathcal{N})$, we fix all parameters that would affect them in advance. The equilibrium frequency parameters, $\hat{\Pi}$, are estimated directly from the sequence data, using a counting approach (Kosakovsky Pond et al., 2010). The nucleotide substitution rates, $\hat{\mathcal{N}}$, and the tree topology and branch proportion parameters, $\hat{\mathcal{T}}$, are fixed at the maximum likelihood estimates under a nucleotide model.

To construct a distribution over the selection parameters, both the values of α and β and their associated probabilities, $p(\alpha, \beta | \theta)$, must be specified. Random effects models typically parameterize the values of α and β as a function of θ, so their 'locations' move during optimization. We avoid this by fixing the locations of α and β, and modify the distribution by varying the weights, $p(\alpha, \beta | \theta)$. With a coarse grid this would be problematic, but as the grid resolution increases, the ability to move the category locations becomes irrelevant. The appropriateness of these computational simplifications is dealt with further in the Discussion section.

The value of each conditional likelihood $p(x_i | \alpha, \beta, \mathcal{T}, \Pi, \mathcal{N})$ depends on the values of α and β, the branch lengths, the equilibrium frequencies, and the nucleotide model. Every time existing random effects methods compute the marginal likelihood for a specific set of parameter values, the conditional likelihoods need to be calculated for every value of α and β for each site, often ruling out the use of random effects models on large alignments. Our approach exploits the fact that, if we use all three shortcuts described above to estimate the branch proportions $\hat{\mathcal{T}}$, nucleotide substitution rates $\hat{\mathcal{N}}$, and equilibrium frequencies $\hat{\Pi}$ in advance, *and* we use fixed positions of α and β for modeling our selection parameters, then the conditional likelihoods $p(x_i | \alpha, \beta, \hat{\mathcal{T}}, \hat{\Pi}, \hat{\mathcal{N}})$ only need to be computed once, rather than every time we compute the marginal likelihood. As can be seen in equation 2, with these precomputed conditional likelihoods, the marginal likelihood is now dependent only upon the probability masses of each α and β pair, $p(\alpha, \beta | \theta)$, and can thus be computed very quickly, reduced to matrix algebra. This is much faster than recalculating all the conditional likelihoods each time, and allows us to consider models with a much finer discretization of the selection parameter space than existing random effects methods, in a fraction of their runtimes.

2.3 Dirichlet hyperprior

The distribution of the selection parameters α and β will differ from one gene to another, and needs to be informed by the data, so we cannot fix it in advance. Nor do we want to assume some parametric form for the distribution, since that limits the flexibility of the method, and could bias inference. Further, estimating 400 category weights as free parameters by maximum likelihood would be overparameterized, so the naive empirical Bayes approach is unhelpful here. Instead, we adopt an approximate hierarchical Bayesian approach (also called Bayes empirical Bayes; Yang, Wong and Nielsen, 2005) and integrate over the uncertainty in the category weights.

To achieve this, we introduce a distribution over our α, β grid that does not assume any parametric form, letting each of the α, β categories have a separate weight (for a $k \times k$ grid, we can represent each distribution over α and β by the k^2-vector $\Psi = (\psi_{1,1}, \psi_{1,2}, ..., \psi_{k,k-1}, \psi_{k,k})$ where $\psi_{i,j}$ is the weight of the i-th α, j-th β category). In general Ψ will have a very large number of dimensions, so we cannot straightforwardly evaluate the integral. Instead we assume a symmetric Dirichlet hyperprior distribution over Ψ, and use MCMC to drawn samples from an approximate the posterior distribution.

2.4 MCMC sampling

To perform the MCMC sampling, we implemented the Metropolis algorithm in HyPhy. We begin in a random initial state, where all of the $\psi_{i,j}$ are drawn from the Dirichlet hyperprior. To propose a change to Ψ, we first randomly pick two elements (uniformly). We sample our perturbation from a uniform distribution between 0 and 0.01, and add this to the first element and subtract it from the second.

To decide whether to accept or reject a proposed state, we draw a from a uniform (0,1) distribution, and accept the state when:

$$a < \frac{p(X|\Psi^{t+1})p(\Psi^{t+1})}{p(X|\Psi^{t})p(\Psi^{t})} \tag{6}$$

where the $p(\Psi)$ values are obtained from the Dirichlet distribution, and $p(X|\Psi)$ is just the likelihood (equation 2), but calculated from our precomputed conditional likelihoods using matrix multiplication. The resulting MCMC chain can be computed extremely efficiently, completing millions of iterations in a few minutes, which is sufficient to produce almost identical site posteriors on separate runs.

The full hierarchical Bayesian calculation of site-specific posterior distributions over α, β would require running a separate MCMC chain for each site, leaving out the contribution of the data at that site when estimating the site-specific prior over α, β. That site-specific prior estimated from all the other sites would then be used to compute the site posterior distribution over α, β. However, as the number of sites increases, the effect of leaving out a single site diminishes, and so, as a further approximation, we run a single chain for all sites.

We asses MCMC convergence using potential scale reduction factors (PSRFs) and effective sample size (Gelman et al., 2003) computed for the posterior probabilities of positive selection for each site. For all datasets tested, an MCMC chain length of 2×10^6 with the first half discarded as burn-in yields good convergence (assessed by running 10 MCMC chains in parallel from random starting positions) with PSRFs close to 1 and effective sample sizes over 100 for most sites.

Table 2: Comparative performance of FEL and FUBAR on simulated data. The rate of false positives (FP) and power are reported for a fixed nominal test p-value of 0.05 for FEL, and a posterior threshold of 0.9 for FUBAR. To achieve a fair comparison between tests with different measures of evidence, power is also shown for the p-value or posterior threshold that achieves FP of 0.05, estimated empirically from the distribution of p-values or posteriors on the subset of sites evolving neutrally.

Simulation	FP : Power		Power at FP= 0.05	
	FEL	FUBAR	FEL	FUBAR
Encephalitis virus env				
$\omega^+ = 1.25$	0.01:0.03	0.00:0.01	0.04	0.10
$\omega^+ = 1.5$	0.00:0.03	0.00:0.02	0.09	0.14
$\omega^+ = 1.75$	0.00:0.03	0.00:0.04	0.08	0.17
$\omega^+ = 2$	0.00:0.05	0.00:0.07	0.13	0.24
$\omega^+ = 3$	0.00:0.09	0.00:0.20	0.19	0.38
$\omega^+ = 5$	0.00:0.19	0.00:0.44	0.34	0.60
$\omega^+ = 8$	0.00:0.28	0.00:0.60	0.50	0.74
$\omega^+ = 12$	0.00:0.34	0.00:0.67	0.54	0.82
$\omega^+ = 16$	0.00:0.38	0.00:0.77	0.63	0.85
Vertebrate Rhodopsin				
$\omega^+ = 1.25$	0.01:0.07	0.00:0.04	0.07	0.12
$\omega^+ = 1.5$	0.01:0.08	0.00:0.08	0.08	0.18
$\omega^+ = 1.75$	0.01:0.13	0.01:0.15	0.14	0.26
$\omega^+ = 2$	0.01:0.19	0.01:0.27	0.13	0.37
$\omega^+ = 3$	0.01:0.32	0.01:0.57	0.34	0.59
$\omega^+ = 5$	0.01:0.48	0.01:0.80	0.51	0.88
$\omega^+ = 8$	0.01:0.67	0.01:0.96	0.74	0.98
$\omega^+ = 12$	0.00:0.71	0.00:0.99	0.80	1.00
$\omega^+ = 16$	0.00:0.76	0.00:0.99	0.88	1.00
Camelid VHH				
$\omega^+ = 1.25$	0.01:0.11	0.01:0.09	0.06	0.09
$\omega^+ = 1.5$	0.02:0.19	0.01:0.20	0.14	0.21
$\omega^+ = 1.75$	0.01:0.34	0.01:0.42	0.26	0.53
$\omega^+ = 2$	0.01:0.51	0.01:0.60	0.48	0.62
$\omega^+ = 3$	0.01:0.74	0.01:0.74	0.64	0.78
$\omega^+ = 5$	0.01:0.93	0.01:0.95	0.93	0.97
$\omega^+ = 8$	0.01:0.98	0.01:0.99	0.98	0.99
$\omega^+ = 12$	0.01:0.97	0.01:1.00	0.97	1.00
$\omega^+ = 16$	0.02:0.99	0.03:1.00	0.99	1.00

3 Results

3.1 Power and false positive rates

To assess the statistical properties of FUBAR, we compared power and false positive rates between FUBAR and FEL, using a collection of simulated alignments where the values for α and β varied from one site to another. The data were simulated over phylogenies estimated from 3 empirical datasets of varying size: 23 Encephalitis virus *env* sequences, 38 vertebrate rhodopsin sequences, and 212 camelid

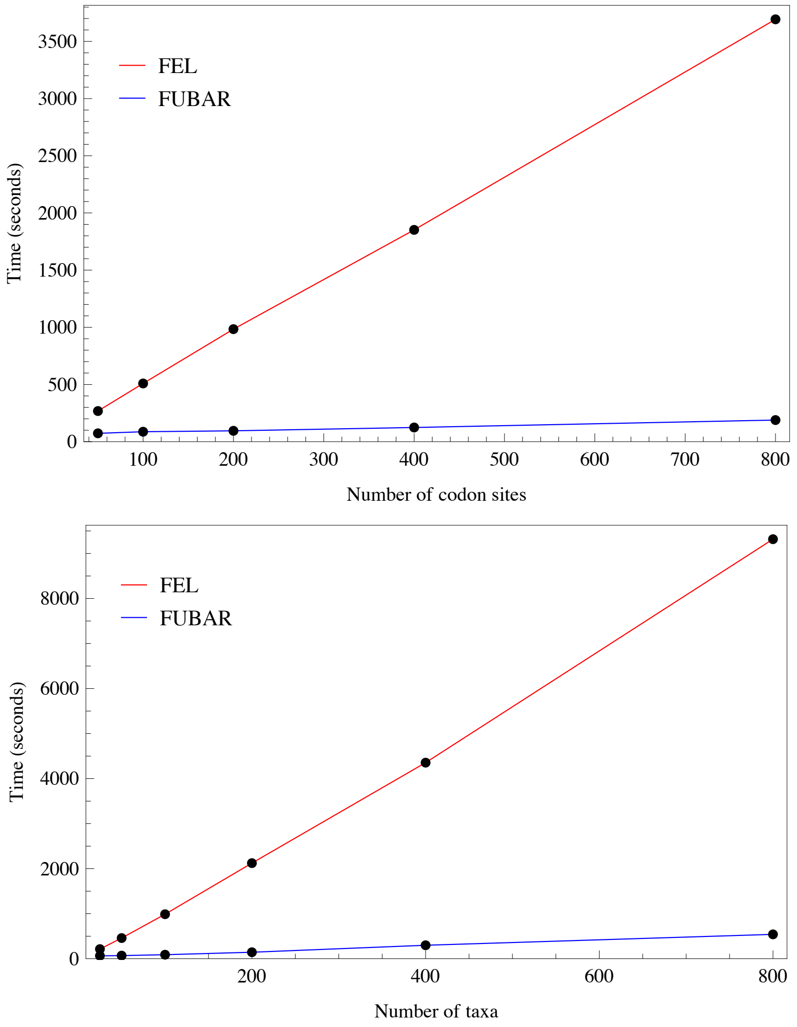

Figure 1: These plots depict the execution times for FEL and FUBAR while varying the number of codon sites (top) and number of taxa (bottom).

VHH sequences (see Murrell et al., 2012b, for details).

Table 2 demonstrates the superiority of FUBAR over FEL. At a posterior threshold of 0.9, FUBAR achieves very low false positive rates on data that was simulated under neutrality, and has superior power in 22/27 configurations. To achieve a fair comparison between tests with different measures of evidence – p-values vs. posteriors – the thresholds of both were adjusted so that FEL and FUBAR both achieve false positive rates of 0.05 on neutral data. This makes the superiority of FUBAR even clearer. FUBAR has greater power in every case, and the difference is sometimes substantial, with FUBAR having over twice the power of FEL in some configurations.

3.2 Speed comparisons

Random effects likelihood methods are typically very computationally intensive and very difficult to parallelize, precluding their use on very large alignments with many sequences. Fixed effects methods are faster and typically parallelized, so FEL was used as a comparator to get an estimate of the efficiency gains enjoyed by FUBAR. A very large HIV-1 *env* alignment was obtained from LANL, stripped of gaps and subsampled to create alignments of varying size. To investigate how computation time increases with the number of sites, we sampled 100 taxa randomly from the *env* alignment, and created 5 alignments with 50, 100, 200, 400 and 800 randomly sampled codon sites. To investigate how computation time increases with taxa, we fixed the number of sites to 200 and sampled alignments with 25, 50, 100, 200, 400 and 800 taxa. All phylogenies were estimated with FastTree 2 (Price, Dehal and Arkin, 2010) using the GTR nucleotide model. FEL and FUBAR were compared on a computing cluster, with the analyses running in parallel on 10 nodes each. FUBAR was consistently faster than FEL across all tested alignments. As can be seen in figure 1, FEL took from 3.3 times longer (214 seconds for FEL vs 65 seconds for FUBAR) for the smallest alignment, to 19.5 times longer (1 hours 2 minutes for FEL vs 3 minutes for FUBAR) for the largest alignment, with the relative disparity increasing uniformly with alignment size. REL, using only 3 categories each for α and β, which we did not run on all the alignments, took 22 minutes 25 seconds for the smallest alignment and 35 hours 29 minutes for the largest.

3.3 Robustness to model misspecification

Before FUBAR, random effects models typically used a small number of categories to capture variation from one site to another. We wanted to investigate how empirical Bayesian inference behaves when the model is misspecified, and, in particular, when the model is too simple to accommodate the data, since this is almost universally true of most models for real datasets. An extreme example of this is PAML's M2a model (Wong et al., 2004), which allows 3 categories for β/α. To investigate the impact of this kind of model misspecification upon inference, we simulated 1000 sites using a constant $\alpha = 1$ but with β taking values of 0.2 (50%), 1 (30%), 3 (10%) and 11 (10%). This models a situation with no synonymous rate variation, where most sites were under purifying or neutral selection, and a smaller proportion of sites were under either weak or strong selection. This is a small violation of PAML's M2a model, whose alternative model allows 3 categories: 1 purifying, 1 neutral and 1 positive selection category. We were particularly interested in the weak selection category, where the true ω ($= \beta/\alpha$) for all sites was 3.

PAML's ω^+ (its only category with $\omega > 1$) must attempt to accommodate both the $\omega = 3$ and the $\omega = 11$ sites, and the resulting maximum likelihood estimate (MLE) is 7.6. For any given prior, the posterior $P(M_{\omega^1})/P(M_{\omega^+})$ is proportional to $P(D_i|M_{\omega^1})/P(D_i|M_{\omega^+})$, so, when posteriors for either neutral or positive selection are calculated under this model, the ratio of the likelihood evaluated at $\omega = 1$ and the likelihood evaluated at $\omega = 7.6$ is the relevant contribution from the data at that site. The true peak of the likelihoods for most sites of interest is between these values, declining to either side. For some sites, the likelihood at $\omega = 1$ is higher than at $\omega = 7.6$, but the reverse is true for other sites.

The true value is between PAML's two available options. The model restrictions preclude detecting all such sites with confidence (which would be optimal, since all have a true $\omega = 3$). Under these conditions we would hope, from an inferential procedure, that the posteriors would reflect substantial uncertainty. Instead, what we observe is 42% of sites show either strong evidence for selection, with posteriors > 0.90, or, critically, strong evidence *against* selection, with 41% of sites showing posteriors < 0.1. Very few sites show moderate posteriors - the uncertainty we would hope for from a slight model misspecification.

FUBAR, with its dense conditional likelihood grid, is capable of learning the presence of both $\omega > 1$ categories in the data, evaluating the likelihood much closer to the expected peak near $\omega = 3$. The results are much more sensible. Of sites simulated under a true ω of 3, 81% were detected with posteriors > 0.90. The mean posterior across all sites with true $\omega = 3$ was 0.93, vs M2a's 0.52. Under FUBAR, no sites show evidence against positive selection. Figure 2 shows the distribution of the posteriors for both M2a

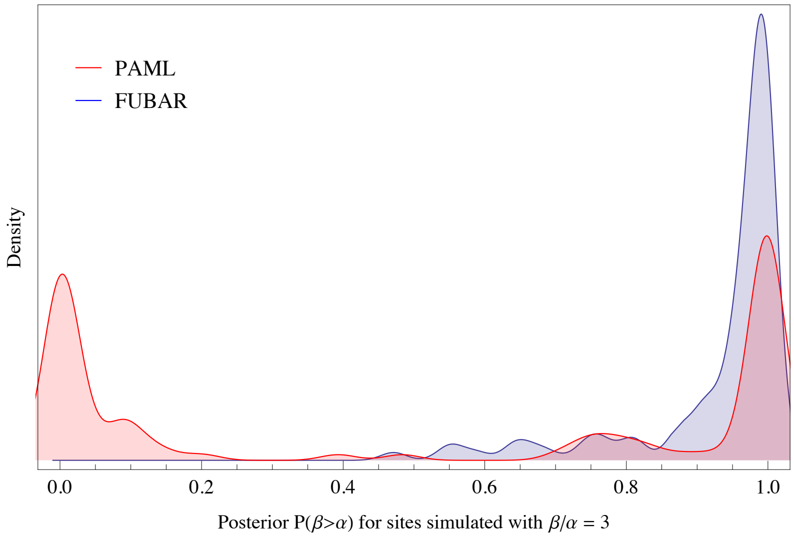

Figure 2: Inference using Bayes' rule under mispecified models. The smoothed histogram depicts the density of posteriors for sites simulated under a true $\omega = 3$ (see text for description of the simulation). PAML's M2a (the red curve) confidently identifies positive selection in nearly half of these sites, but also declares strong evidence *against* positive selection in half. FUBAR, on the other hand, detects most of the sites, and doesn't appear to suffer the same overconfidence problem.

and FUBAR.

3.4 Empirical example - Influenza A H

To demonstrate the use of FUBAR, we analyzed a collection of global human IAV haemagglutinin subtype 3 (H3) sequences from the NCBI Influenza Virus Database (`http://www.ncbi.nlm.nih.gov/genomes/FLU/`). The influenza haemagglutinin glycoprotein (HA) mediates the entry of the virus into cells, and is the target of neutralizing antibodies. We reconstructed the phylogeny (see figure 3) for 3142 sequences using FastTree 2 (Price, Dehal and Arkin, 2010). The FUBAR selection analysis (which was restricted to using 10 nodes, just as for the timing comparisons) took one and a half hours. Figure 3 shows the distribution of $\beta - \alpha$ across HA, with the mode at mild purifying selection ($\beta < \alpha$), and with a minority of sites under positive selection ($\beta > \alpha$). We use $\beta - \alpha$ rather than the posterior $P(\beta > \alpha)$ because, with so many sequences, the posteriors can confidently report positive selection even when it is very weak, and so we examine the estimated *magnitude* of positive selection instead. As a measure of the magnitude of selection, β/α is very skew, but $\beta - \alpha$, with neutrality at 0, is more amenable to visualization. All sites described below are codon sites, as opposed to sites in each subunit, unless stated otherwise.

Codon sites under positive selection are almost exclusively localized to the globular head. Using $\beta - \alpha > 1$ as a working definition of strong positive selection, 11 codons were identified. Of these, 7 sites (154, 161, 173, 210, 241, 242, and 245) are clustered in and around the receptor binding site and fall broadly within 2 of the classical, major antigenic regions (regions A and B; Caton et al., 1982) - these "regions" are called "sites" in the influenza literature, but we use "regions" here to prevent confusion. Sites 66 and 69 fall broadly within region C (with site 61 located in close proximity). The remaining site under strong positive selection (19; HA1 site 3), is not part of the crystallized structure, but would likely lie near the base of the membrane-proximal stem. In contrast to regions A, B and C, region D is not under similarly strong positive selection, suggesting that it may represent a more sub-dominant target for neutralizing antibodies.

The majority of sites under strong purifying selection are located within the membrane-proximal stem. Antibodies to the HA stem are less common, but have nevertheless been shown to be able to mediate neutralization by inhibiting viral fusion with the host cell (Okuno et al., 1993; Varecková et al., 2003). This is consistent with the identification of broadly cross-reactive antibodies that target this region (Sui et al., 2009; Ekiert et al., 2009; Wang et al., 2010; Corti et al., 2011), and reinforces the haemagglutinin stem as an attractive target for influenza vaccines.

Interestingly, site 553 (HA2 site 208) is under extremely strong purifying selection ($\beta - \alpha = -23.5022$), although its function is not clear.

Of sites in the globular head under strong purifying selection, sites 181, 203, 234, and 238 are clustered together in the quaternary structure at the protomer interface of the globular head, potentially representing a more accessible target for cross-neutralizing antibodies. While site 181 represents an N-linked glycosylation site which could potentially shield this region from antibody binding, it is also conceivable that the glycan may contribute to epitope formation. Several potent and broadly cross-neutralizing HIV antibodies (PG9/PG16-like and PGT-128-like antibodies) are dependent on both a peptide and a glycan component for binding (McLellan et al., 2011; Pejchal et al., 2011), providing a precedent for this mode of recognition.

4 Discussion and Conclusion

Traditionally, phylogenetic models of evolution have employed computational shortcuts to speed up likelihood optimization. One widespread example involves the equilibrium frequencies: an estimate of the equilibrium frequency parameters, $\hat{\Pi}$, is often counted directly off the sequence data, invoking a stationarity assumption to reduce the number of parameters that need to be optimized (Kosakovsky Pond et al., 2010). This works because inference under the model is seldom very sensitive to the typical magnitude of the deviations of these quick-and-dirty calculations from the actual maximum likelihood estimates. Another example is that estimates of the nucleotide substitution rates, $\hat{\mathcal{N}}$, may be calculated using a simpler model – such as a codon model that does not allow site-to-site variability in selection intensity – and then fixed for the optimization of the more complicated model (Kosakovsky Pond and Frost, 2005). This works for the same reason as the shortcut estimate $\hat{\Pi}$: inference is not affected.

A less common shortcut estimates the branch proportions under a simple model and fixes them, although the overall tree length is still allowed to vary. This is adopted in the fixed effects models of Kosakovsky Pond and Frost (2005). However, the acceptability of this approximation is implicit in the standard Bayes empirical Bayes approach of Yang, Wong and Nielsen (2005): Bayes empirical Bayes acknowledges that uncertainty exists about the MLEs for parameters, but that only some parameters affect site-specific empirical Bayesian inference of positive selection. For parameters that matter, distributions over these parameters are specified, and the uncertainty about these parameters is integrated out. Parameters deemed not to affect inference are left at their MLEs. Branch lengths, as well as nucleotide substitution rates and equilibrium frequencies, count among the latter. If a parameter matters little enough for uncertainty around the MLE to be ignored, then using a shortcut estimate from a simpler model should also be admissible.

FUBAR uses a precomputed grid of conditional likelihoods, which relies on the same set of computational shortcuts as the fixed effects models: the branch lengths, nucleotide substitution rates and equilibrium frequencies must all being fixed in advance, using estimates from simpler models. Standard REL models do not need to take these shortcuts (although they sometimes do for computational reasons), but are computationally constrained to use a different shortcut: a much coarser discretization over the α, β space. FUBAR takes shortcuts when estimating unimportant parameters to avoid shortcuts when estimating important ones.

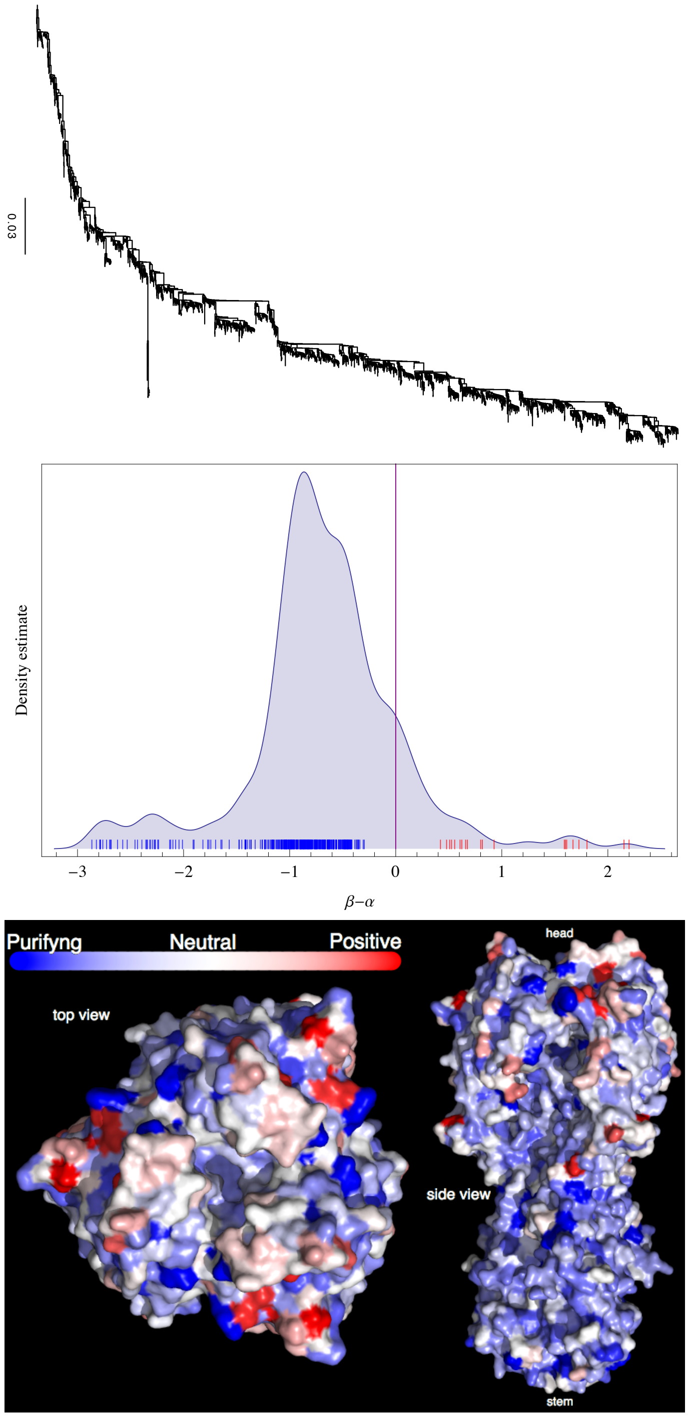

Figure 3: Top: The H3 phylogeny with 3142 coding sequences. Middle: The smoothed histogram of $\beta - \alpha$ across H3, with the greatest density at mild purifying selection ($\beta < \alpha$), and fewer sites under positive selection ($\beta > \alpha$). The notches depict sites with posteriors greater than 0.9 for positive (red) or purifying (blue) selection. Bottom: The inferred $\beta - \alpha$ values mapped to the HA protein (PDB 3ZTJ; Corti et al., 2011), displayed from two viewpoints. Red regions with stronger diversifying selection are likely involved in immune escape. These primarily occur on the “head” of the protein, with mostly purifying selection on the membrane proximal stem. See text for further detail.

Random effects inference allows one to learn the distribution of selection parameters from the alignment, and lets this distribution influence site-wise inference. It is interesting to contrast the performance of M2a and FUBAR with a FEL analysis of the 4 category simulation in section 3.3. FEL, which does not learn a prior distribution, has poor performance compared to FUBAR, detecting only 33 of 100 sites with $\beta/\alpha = 3$ at the $p < 0.05$ level (which also gives 8 false positives on the neutral or purifying sites, compared to FUBAR's 2 false positives). But FEL, lacking an α, β prior, reflects the uncertainty appropriately, and, while it fails to detect many sites, it never becomes overconfident in the incorrect result, unlike PAML's M2a. So, where site specific inference is not constrained by any kind of shared prior, the uncertainty is large and the power is low. But, when it is constrained incorrectly, inference can be both confident and wrong at the same time, which is positively misleading. When the distribution over α and β is learned correctly, however, the uncertainty around the α, β estimate for each site is substantially reduced, since the prior distribution imposes appropriate constraints over which α, β values are probable. This allows FUBAR a substantial increase in power over both FEL and M2a for this example. While the simulating distribution used in this example might be extreme, it serves to illustrate that a correctly learned informative distribution substantially improves inference.

Methods proposed up until now all have unrealistic constraints: some assume no synonymous rate variation, others confine sites to a small number of classes, and most assume independence from one site to another. These restrictive assumptions almost guarantee that they will be unable to capture the true data generating distribution. Some restrictive assumptions are more harmful than others, so systematic examination of their individual and collective effects is recommended. Here, we have shown with the example in 3.3 that an overly coarse grid can be very misleading when violated. Our Bayesian method approximates the α, β distribution as flexibly as the grid resolution allows, and is computationally cheap due to shortcuts that are mild when compared with the coarse discretizations assumed by existing random effects methods.

The methodology behind FUBAR can be applied to other models. DEPS (Kosakovsky Pond et al., 2008) and EDEPS (Murrell et al., 2012a) model amino acid evolution, and detect directional selection - where the substitution rate towards a particular amino acid is elevated - using a random effects approach with 1 neutral and 1 positive selection category. This could benefit from a grid-based implementation, allowing a large number of non-neutral categories, which should improve the statistical performance of the method, while achieving speedups similar to those observed in FUBAR. Methods that accommodate site-to-site variability in many parameters, such as MEME (Murrell et al., 2012b) and MEDS (Murrell et al., 2012a), may not be amenable to grid-based approaches, however, because the number of grid points grows exponentially in the number of dimensions that vary from site to site.

References

Caton, A. J., G. G. Brownlee, J. W. Yewdell, and W. Gerhard. 1982. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell **31**(2):417–427.

Corti, D., J. Voss, S. J. Gamblin, G. Codoni, A. Macagno, D. Jarrossay, S. G. Vachieri, D. Pinna, A. Minola, F. Vanzetta, C. Silacci, B. M. Fernandez-Rodriguez, G. Agatic, S. Bianchi, I. Giacchetto-Sasselli, L. Calder, F. Sallusto, P. Collins, L. F. Haire, N. Temperton, J. P. M. Langedijk, J. J. Skehel, and A. Lanzavecchia. 2011. A Neutralizing Antibody Selected from Plasma Cells That Binds to Group 1 and Group 2 Influenza A Hemagglutinins. Science **333**(6044):850–856.

Ekiert, D. C., G. Bhabha, M.-A. Elsliger, R. H. E. Friesen, M. Jongeneelen, M. Throsby, J. Goudsmit, and I. A. Wilson. 2009. Antibody Recognition of a Highly Conserved Influenza Virus Epitope. Science **324**(5924):246–251.

Felsenstein, J. 1981. Evolutionary trees from DNA-sequences – a maximum-likelihood approach. J Mol Evol **17**:368–376.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science). Chapman and Hall/CRC, 2nd ed.

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

Huelsenbeck, J. P., S. Jain, S. W. Frost, and S. K. L. Pond. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proceedings of the National Academy of Sciences of the United States of America **103**(16):6263–6268.

Hughes, A. L., and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class i loci reveals overdominant selection. Nature **335**(6186):167–70.

Kosakovsky Pond, S., W. Delport, S. V. Muse, and K. Scheffler. 2010. Correcting the bias of empirical frequency parameter estimators in codon models. PLoS One **30**:e11230.

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

Kosakovsky Pond, S. L., A. F. Y. Poon, A. J. Leigh Brown, and S. D. W. Frost. 2008. A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza a virus. Mol Biol Evol **25**(9):1809–24.

McLellan, J. S., M. Pancera, C. Carrico, J. Gorman, J.-P. P. Julien, R. Khayat, R. Louder, R. Pejchal, M. Sastry, K. Dai, S. O'Dell, N. Patel, S. Shahzad-ul Hussan, Y. Yang, B. Zhang, T. Zhou, J. Zhu, J. C. Boyington, G.-Y. Y. Chuang, D. Diwanji, I. Georgiev, Y. D. D. Kwon, D. Lee, M. K. Louder, S. Moquin, S. D. Schmidt, Z.-Y. Y. Yang, M. Bonsignori, J. A. Crump, S. H. Kapiga, N. E. Sam, B. F. Haynes, D. R. Burton, W. C. Koff, L. M. Walker, S. Phogat, R. Wyatt, J. Orwenyo, L.-X. X. Wang, J. Arthos, C. A. Bewley, J. R. Mascola, G. J. Nabel, W. R. Schief, A. B. Ward, I. A. Wilson, and P. D. Kwong. 2011. Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. Nature **480**(7377):336–343.

Messier, W., and C. B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**(6612):151–4.

Murrell, B., T. de Oliveira, C. Seebregts, S. L. Kosakovsky Pond, K. Scheffler, on behalf of the Southern African Treatment, and R. N. S. Consortium. 2012a. Modeling HIV-1 Drug Resistance as Episodic Directional Selection. PLoS Comput Biol **8**(5):e1002507+.

Murrell, B., J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler, and S. L. Kosakovsky Pond. 2012b. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genet **8**(7):e1002764+.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol **11**(5):715–24.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**(3):929–36.

Okuno, Y., Y. Isegawa, F. Sasao, and S. Ueda. 1993. A common neutralizing epitope conserved between the hemagglutinins of influenza A virus H1 and H2 strains. Journal of virology **67**(5):2552–2558.

Pejchal, R., K. J. Doores, L. M. Walker, R. Khayat, P.-S. Huang, S.-K. Wang, R. L. Stanfield, J.-P. Julien, A. Ramos, M. Crispin, R. Depetris, U. Katpally, A. Marozsan, A. Cupo, S. Maloveste, Y. Liu, R. McBride, Y. Ito, R. W. Sanders, C. Ogohara, J. C. Paulson, T. Feizi, C. N. Scanlan, C.-H. Wong, J. P. Moore, W. C. Olson, A. B. Ward, P. Poignard, W. R. Schief, D. R. Burton, and I. A. Wilson. 2011. A Potent and Broad Neutralizing Antibody Recognizes and Penetrates the HIV Glycan Shield. Science **334**(6059):1097–1103.

Pond, S. K., and S. V. Muse. 2005. Site-to-site variation of synonymous substitution rates. Mol Biol Evol **22**(12):2375–85.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2 Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE **5**(3):e9490+.

Sui, J., W. C. Hwang, S. Perez, G. Wei, D. Aird, L.-m. M. Chen, E. Santelli, B. Stec, G. Cadwell, M. Ali, H. Wan, A. Murakami, A. Yammanuru, T. Han, N. J. Cox, L. A. Bankston, R. O. Donis, R. C. Liddington, and W. A. Marasco. 2009. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. Nature structural & molecular biology **16**(3):265–273.

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

Vareckovà, E., V. Mucha, S. A. Wharton, and F. Kostolanský. 2003. Inhibition of fusion activity of influenza A haemagglutinin mediated by HA2-specific monoclonal antibodies. Archives of virology **148**(3):469–486.

Wang, T. T., G. S. Tan, R. Hai, N. Pica, E. Petersen, T. M. Moran, and P. Palese. 2010. Broadly Protective Monoclonal Antibodies against H3 Influenza Viruses following Sequential Immunization with Different Hemagglutinins. PLoS Pathog **6**(2):e1000796+.

Wong, W. S. W., Z. Yang, N. Goldman, and R. Nielsen. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics **168**(2):1041–51.

Yang, Z., W. S. W. Wong, and R. Nielsen. 2005. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Molecular Biology and Evolution **22**(4):1107–1118.

Non-Negative Matrix Factorization for Learning Alignment-Specific Models of Protein Evolution

Ben Murrell[1,2], Thomas Weighill[2], Jan Buys[2], Robert Ketteringham[2], Sasha Moola[3], Gerdus Benade[2], Lise du Buisson[2], Daniel Kaliski[3], Tristan Hands[3], Konrad Scheffler[2]*

1 Biomedical Informatics Research Division, eHealth Research and Innovation Platform, Medical Research Council, Cape Town, Western Cape, South Africa, **2** Stellenbosch University, Stellenbosch, Western Cape, South Africa, **3** University of Cape Town, Cape Town, Western Cape, South Africa

Abstract

Models of protein evolution currently come in two flavors: generalist and specialist. Generalist models (e.g. PAM, JTT, WAG) adopt a one-size-fits-all approach, where a single model is estimated from a number of different protein alignments. Specialist models (e.g. mtREV, rtREV, HIVbetween) can be estimated when a large quantity of data are available for a single organism or gene, and are intended for use on that organism or gene only. Unsurprisingly, specialist models outperform generalist models, but in most instances there simply are not enough data available to estimate them. We propose a method for estimating alignment-specific models of protein evolution in which the complexity of the model is adapted to suit the richness of the data. Our method uses non-negative matrix factorization (NNMF) to learn a set of basis matrices from a general dataset containing a large number of alignments of different proteins, thus capturing the dimensions of important variation. It then learns a set of weights that are specific to the organism or gene of interest and for which only a smaller dataset is available. Thus the alignment-specific model is obtained as a weighted sum of the basis matrices. Having been constrained to vary along only as many dimensions as the data justify, the model has far fewer parameters than would be required to estimate a specialist model. We show that our NNMF procedure produces models that outperform existing methods on all but one of 50 test alignments. The basis matrices we obtain confirm the expectation that amino acid properties tend to be conserved, and allow us to quantify, on specific alignments, how the strength of conservation varies across different properties. We also apply our new models to phylogeny inference and show that the resulting phylogenies are different from, and have improved likelihood over, those inferred under standard models.

Citation: Murrell B, Weighill T, Buys J, Ketteringham R, Moola S, et al. (2011) Non-Negative Matrix Factorization for Learning Alignment-Specific Models of Protein Evolution. PLoS ONE 6(12): e28898. doi:10.1371/journal.pone.0028898

Editor: Thomas Mailund, Aarhus University, Denmark

Received September 22, 2011; **Accepted** November 16, 2011; **Published** December 22, 2011

Funding: This research was funded by Europeaid grant number SANTE/2007/174-790 from the European Commission. Funding for the UCSD computing cluster was provided by the Joint DMS/NIGMS Mathematical Biology Initiative through Grant NSF-0714991 and the National Institutes of Health grant AI47745. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: kscheffler@cs.sun.ac.za

Introduction

Empirical models of protein evolution, as pioneered by Dayhoff and colleagues [1,2], have found wide use across varied domains: sequence alignment [3], phylogenetics [4], and as baseline models against which positive selection is detected [5]. These models describe molecular evolution at the amino acid level by quantifying the relative substitution rates between different amino acids. Such rates are an aggregation over multiple distinct phenomena: the structure of the genetic code, which renders some mutations less likely to occur; and differences in the physicochemical properties of the amino acids themselves, which, along with the environment of the organism, will determine which substitutions are deleterious, tolerated or adaptive.

The original approach by Dayhoff *et al.* used a maximum parsimony procedure to reconstruct the ancestral sequences and phylogeny for a collection of protein families and counted the amino acid substitutions across this phylogeny. Their PAM (point accepted mutation) matrices were derived from rates of amino acid exchange estimated from these counts. Jones et al. [6] automated a similar procedure which ran on a much larger dataset, producing the JTT amino acid rate matrix. A further refinement to these "counting" methods was contributed by Kosiol and Goldman [7]. Whelan and Goldman [8] made use of a maximum likelihood approach which, unlike the counting methods mentioned above, finds the amino acid substitution matrix while simultaneously optimizing the branch lengths of the phylogeny, thus incorporating the possibility of multiple substitutions taking place along any given branch. In constructing their WAG matrix, they applied an approximation of this technique to a large dataset.

The above models are generalist in that they use the same set of relative amino acid exchangeabilities for all genes and all organisms. However, since these exchangeabilities can vary considerably between genes and/or organisms, researchers have also constructed specialist models. Such models are estimated from – and intended for use on – a specific gene, organism or genetic code. Adachi and Hasegawa [9] estimated an empirical amino acid substitution rate matrix for mitochondrial DNA-encoded proteins, using the maximum likelihood method on a dataset consisting of mtDNA-encoded sequences from vertebrate species. Yang et al. [10] used a similar technique to derive a substitution rate matrix from the mtDNA mammalian dataset of Cao et al. [11]. Both of these are intended for use only on mitochondrial sequences. Dimmic et al. [12] optimized an amino acid substitution rate matrix via maximum likelihood, using a set of retroviral pol protein sequences. Nickle et al. [13] derived two substitution rate matrices with maximum likelihood, each using different HIV protein sequence datasets. The first matrix (HIVwithin) was derived by applying maximum likelihood to pairs of within-individual protein sequences, while the second (HIVbetween) made use of a set of consensus sequences obtained from a population of individuals. In all cases, specialist models fit alignments for their particular system better than generalist models.

Specialist models are better than generalist ones, but specialist models simply don't exist for most alignments. If the alignment is very large, one can estimate a fully parameterized general reversible model (often referred to as REV), which involves estimating 190 parameters. With most alignments, however, this will be severely over-parameterized. Computational biologists who want to analyze a single alignment for which a specialist model has not been constructed are therefore forced to resort to using a generalist model. This is the problem we seek to address: constructing alignment-specific models of protein evolution without over-fitting, allowing the model to be just as complex as the data justify.

We investigate a compromise between generalist and specialist models by first extracting, from a large dataset, the important dimensions of variation in amino acid substitution rates, and then using these to constrain our models. We propose the following three step approach: First, we estimate a separate REV amino acid rate matrix for each of a number of reasonably large alignments. These provide a library of specialist models, each with 190 rate parameters. Second, we apply non-negative matrix factorization – a dimensionality reduction technique – to find a smaller set of 'basis' rate matrices, whose non-negative weighted combinations best approximate the original REV estimates. Finally, for a new alignment (which is not contained in the original dataset and may be relatively small), we model the amino acid rate matrix as a weighted combination of our set of basis matrices. During this final step, we optimize over both the number of combination weights and their values. NNMF is thus used to approximate the space of useful models, reducing the number of parameters required to explore it. Rate matrices for specific alignments are estimated by searching within this lower-dimensional parameter space.

The basis matrices obtained by our NNMF procedure are interesting in that they reveal a set of components from which the eventual rate matrices are comprised – each alignment-specific rate matrix is the sum of positive multiples of the basis matrices. By measuring, for each basis matrix, the correlation between the amino acid exchangeabilities and the strength of the different physicochemical properties of the amino acids being exchanged, we obtain an indication of how the degree of conservation of the different properties varies between different alignments.

Using a separate test dataset, we show that models estimated through our procedure outperform existing models in terms of Akaike's information criterion (AIC) on all but one of 50 alignments tested. Finally, we use our models to infer phylogenies and show that this leads to phylogenetic trees that are structurally different and have higher likelihood than maximum likelihood trees obtained using standard methods.

Methods

We start by briefly reviewing phylogenetic models of protein evolution. Substitutions along every branch of a phylogenetic tree are described by a continuous time Markov process, defined by an instantaneous rate matrix, Q. The elements q_{ij} are the rates of substituting amino acid i with amino acid j. From the rate matrix Q and the length of a branch in the phylogeny, t, a transition probability matrix for that branch can be calculated using the matrix exponential:

$$P(t) = e^{Qt}. \tag{1}$$

The constraint $q_{ii} = -\sum_{\forall j \neq i} q_{ij}$ is required for Q to be a valid Markov process generator. The (ij) elements of $P(t)$ describe the probabilities of substituting amino acid i with amino acid j after time t. With these transition probabilities along the branches of a phylogeny, the likelihood of an alignment can be calculated using Felsenstein's pruning algorithm [4].

We assume the Markov process is reversible: that is, Q can be decomposed into the product of a symmetric matrix S and a diagonal matrix Π, where the elements of the diagonal of Π, π_j, are the equilibrium frequencies for the j^{th} amino acid in the Markov process defined by $Q = S\Pi$, with $\sum_j \pi_j = 1$. Throughout this paper we adopt a common approximation by estimating the equilibrium frequencies π_j as the empirical amino acid frequencies counted across all sites in the alignment.

S is the 20×20 symmetric amino acid exchangeability matrix. Given the symmetry and the constraints on the diagonal elements, this leaves 190 parameters that need to be specified to define the model of protein evolution over a given phylogeny. Our focus in this study is the estimation of these parameters.

Estimating reversible protein models

To characterize the important dimensions of relative substitution rate variation, we first estimate a general reversible (REV) model – where the 190 parameters of S are estimated by maximum likelihood – from each of a large number m of large alignments. We use the procedure described in [13] to estimate a REV model for each alignment. For computational reasons we use a single rate class, ignoring site-to-site amino acid rate variation (although we show that this can be added at a later stage of our procedure).

Non-negative matrix factorization

Non-negative matrix factorization (NNMF) is a tool for dimensionality reduction [14,15] of datasets in which the values, like the rates in the rate matrix S, are constrained to be non-negative. Instead of applying it to data, we use it to reduce the dimensionality of our models. We start by arranging the parameters of each specialist REV model into a vector of dimension $n = 190$. The set of m such vectors combine to form a $n \times m$ matrix V (Figure 1, Table 1) representing the full set of specialist rate matrices. For a given factorization rank $r \ll n$, the NNMF procedure now finds an $n \times r$ matrix W and an $r \times m$ matrix H such that $WH \approx V$. This is done by minimizing an objective function: we chose to minimize the sum of squared differences between WH and V.

W now represents a set of r basis matrices: each column contains the 190 parameters of a single basis matrix, and the S matrix for any of the training alignments can be reconstructed (approximately) by forming a weighted sum over these basis matrices. The weights in this sum are stored in the column of H corresponding to the training alignment in question. One way of interpreting the factorization is that the set of basis matrices in W captures the dimensions of important variation between different rate matrices representing the training alignments, so that they form a set of components out of which any of the rate matrices can

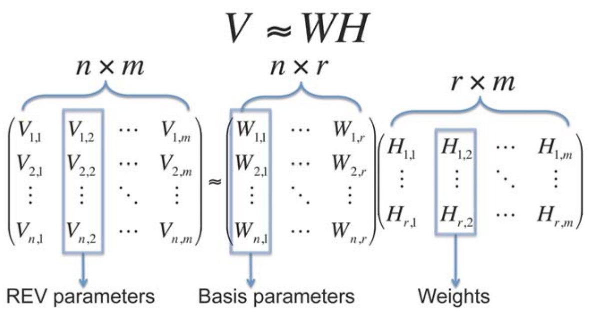

Figure 1. Non-negative matrix factorization.
doi:10.1371/journal.pone.0028898.g001

be built up. Our key assumption is that this will also be the case for alignments not in the training dataset: after paying the fixed cost of learning the $190 \times r$ parameters in W from the training dataset, we propose to represent any alignment using only r weight parameters instead of 190 independent rate parameters (Figure 2).

NNMF proceeds by an iterative algorithm, converging on a local minimum of the sum of squared error. It is thus potentially sensitive to initial conditions. To ensure decent performance, we began with 20 different random initial conditions and optimized the factorization for 2000 iterations each. The best resulting factorization was then further refined for an additional 5000 iterations.

Fitting basis models to new data: optimizing over combination weights

Given a collection of r basis exchangeability matrices, B_i (the columns of W arranged as a reversible rate matrix), their associated weights, w_i, where i goes from 1 to r, a combined exchangeability matrix S is parameterized by:

$$S = \sum_{i=1}^{r} w_i \times B_i \tag{2}$$

We add the constraint that $\sum_i w_i = 1$: since rate and time are confounded, and since the branch lengths are free parameters, this does not entail loss of generality. With a new test alignment (that was not included in the original factorization over the training data) and a collection of basis rate matrices, we can now optimize the weights w_i (and branch lengths) to obtain the maximum likelihood combined model for the alignment. This is in contrast to model selection approaches such as ProtTest [16] which select a single model from a collection of existing models. Importantly, the combined model can itself be represented as a single numeric rate matrix, and can thus be used by any application that allows for custom amino acid rate matrices, such as HyPhy [17], PAML [18] or PhyML [19].

The flagship method presented in this paper applies this approach to our NNMF-estimated basis matrices (we refer to this method as "NNMF"). We also introduce a method that uses the same mixture approach, but differs from NNMF, in that it uses a collection of existing numeric rate matrices for its basis matrices , and we name the resulting model the 'Mixture of Existing Rates' (MOER) model. For any given test alignment, both models use mixture components that are fixed in advance, but NNMF obtains these by factorizing a large dataset, while MOER uses existing "average" model estimates. The models we chose to combine in MOER are those available by default in the HyPhy software package: Dayhoff, JTT, WAG, rtREV, mtMAM, mtREV, HIVwithin and HIVbetween. For both NNMF and MOER, the equilibrium frequencies used when modeling the test alignments are estimated from the amino acid counts.

These are also the fixed rate models we use as a comparison for NNMF and MOER to asses the performance of our methods, since they are standardly used in the literature. Under a fixed rate model, the branch lengths are optimized to maximize the likelihood, but the exchangeability matrix itself has no flexibility. Each fixed rate model is a special case of MOER, when the weights for all but a single matrix go to 0. MOER will thus always obtain better likelihoods than any single fixed-rate model, but our model comparison measure will penalize against the extra parameters if they prove unnecessary.

Selecting the optimal factorization rank for a given alignment

The NNMF decomposition requires the specification of a factorization rank: the number of basis matrices to be estimated. Since the optimal number of basis matrices for a new alignment depends on the details of that alignment – larger alignments can justify more parameters – no single factorization will suffice. Instead, we obtain factorizations for a range of different ranks. To select the best NNMF model for each new alignment, we maximize the likelihood function for every rank, and select the model with the best (minimum) AICc(Akaike's information criterion with a small sample correction [20]) score, which prevents over-fitting by penalizing the inclusion of additional parameters:

$$AICc = -2L + 2p + \frac{2p(p+1)}{n-p-1} \tag{3}$$

Table 1. Interpretation of the matrix factorization in Figure 1.

m	Number of training alignments
n	Number of parameters per rate matrix (190)
r	Number of basis matrices
Column of V	Specialist REV model corresponding to one training alignment
V	Library of specialist REV models
Column of W	One basis matrix
W	Set of r basis matrices
Column of H	Set of weights with which to combine basis matrices to obtain model for one training alignment
H	Set of weights for training dataset

doi:10.1371/journal.pone.0028898.t001

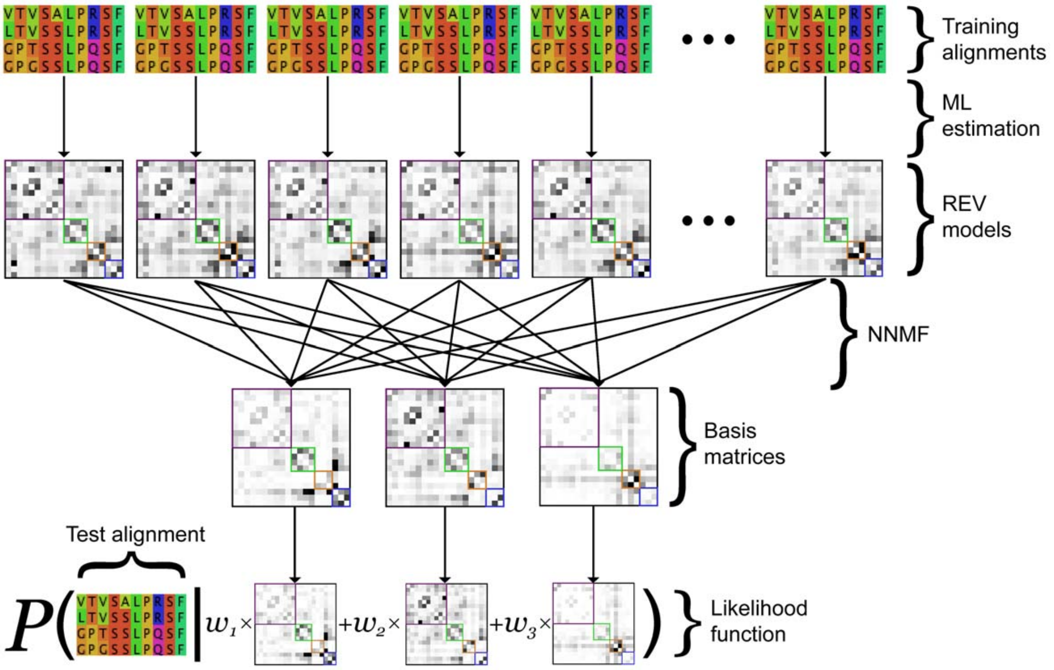

Figure 2. Learning models of protein evolution with NNMF. A schematic overview of the procedure.
doi:10.1371/journal.pone.0028898.g002

where L is the log-likelihood, p is the number of parameters and n is the number of observations. Counting the number of observations is not straightforward: taking the total number of characters in the alignment is problematic because amino acids at the same site are extremely correlated. (If one were to do this, one could add duplicate sequences which would increase the number of observations without being at all informative.) Instead, we use the number of sites as the number of observations. This can lead to problems when branch lengths are included as parameters, because as the number of branches approaches the number of sites (specifically, when $p = n - 1$), the second order term becomes undefined. This is not just a theoretical concern: it actually occurs for one of our test alignments. To remedy this, we exclude branch lengths from our model parameter count. Excluding branch lengths as parameters when extra taxa are not counted as extra observations makes intuitive sense: adding taxa increases the number of branch length parameters to be estimated while providing the required information to estimate those parameters, but is not correspondingly informative for estimation of the other model parameters. For further discussion of these issues, see [21].

Phylogeny comparison

To determine whether improvements in model fit would make a difference to the topology of the inferred phylogeny, we compared the best NNMF model to WAG, the existing amino acid model with the best overall fit on our 50 test alignments. We constructed 50 phylogenies using WAG, and 50 using the best NNMF model. Topology search was performed in PhyML [19] with nearest-neighbor interchange plus subtree pruning and regrafting, and we disallowed rate variation due to computational restrictions. We compared topologies under the Robinson-Foulds symmetric difference [22] using PHYLIP [23].

Data

Training and test alignments were selected from the Pandit database [24], with the selection based on the size of the alignments (Figure 3). For our training dataset (293 alignments in total) we used all alignments with number of sequences >50, alignment length >200 and number of sequences $\times$ alignment length >15000, with the exception of one very large alignment (number of sequences $\times$ alignment length $=989720$) that exceeded our computational resources. The number of sequences per alignment ranged from 51 to 797, with a median of 95 and an inter-quartile range (IQR) of 77. The alignment length ranged from 201 to 1767, with a median of 339 and an IQR of 230.75. All trees used to train the models were also obtained from the Pandit database.

We then adjusted our size criteria to yield a test dataset containing the 50 "next largest" alignments: number of sequences >45, alignment length >195, number of sequences $\times$ alignment length >11800, but excluding all training alignments. The number of sequences per alignment ranged from 46 to 182, with a median of 51 and an IQR of 12. The alignment length ranged from 196 to 926, with a median of 249 and an IQR of 207. Trees were again obtained from the Pandit database.

Implementation

HyPhy [17] was used to estimating the original 293 REV models from the Pandit alignments, using code from [13]. The non-negative matrix factorization was performed in Matlab. Optimizing over basis matrix combination weights for all factorization ranks was performed in HyPhy, as was the comparison of protein models. HyPhy Batch Language (HBL) code for optimizing over combination weights is available online (www.cs.sun.ac.za/ bmurrell/nnmf/), along with the basis matri-

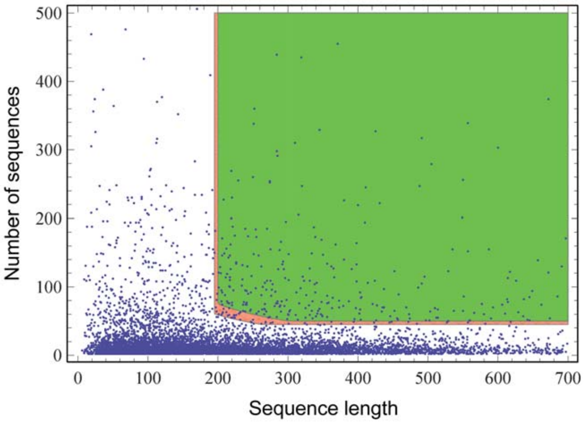

Figure 3. Selecting the larger Pandit alignments. Each blue dot represents an alignment in the Pandit database. The green region covers the alignments used in the training set, and the thin red region covers those in the test set.
doi:10.1371/journal.pone.0028898.g003

ces. A web script for converting from this output to a rate matrix that is usable by PAML and PhyML is also available at the same url.

Results

The basis matrices

We first consider the set of basis matrices obtained on the training alignments. Figure 4 shows that, as expected, the sum of squared errors decreases as the number of basis matrices increases. To investigate the first few sets of basis matrices, we use the Stanfel classification [25] of amino acids according to their physicochemical properties. Figure 5 shows the basis matrices obtained for the first 5 ranks, with the amino acid ordering chosen so as to group amino acids with similar properties together. We observe that, when one or two rate classes are used, the larger rates (darker squares) occur more frequently within the same class than between classes. Thus these rate matrices capture the fact that, on average, physicochemical properties tend to be conserved.

As more rate matrices are added, the variation between different alignments becomes better resolved. By the third factorization ($r=3$), a basis matrix occurs with larger rates (involving Cysteine) occurring between classes. This reflects that, in some alignments, these rates are accelerated while in other alignments they are not: the NNMF analysis indicates that whether these rates are high or low is an important dimension of variation across the training alignments. We also notice that the exchangeabilities of Cysteine with other amino acids are not elevated independently: in alignments where the Cysteine↔Histidine exchangeability is elevated, the Cysteine↔Leucine and Cysteine↔Arginine exchangeabilities also tend to be elevated. This may reflect that the properties under conservation in these alignments, along with the relative importances of these properties, differ from those used to define the Stanfel classification; rather than speculating about the underlying biochemistry, we restrict ourselves to pointing out that the set of basis matrices provides a far richer description of amino acid exchangeability, and how this varies between alignments, than can be achieved by classifying the amino acids into a predefined set of non-overlapping categories.

With $r=5$ we see that Tryptophan has increased exchangeability with most other amino acids in a subset of alignments. It would be interesting to establish the underlying causes of such effects; for now we merely note that they are easily observable. Inspection of the basis matrices for larger values of r would lead to many similar observations.

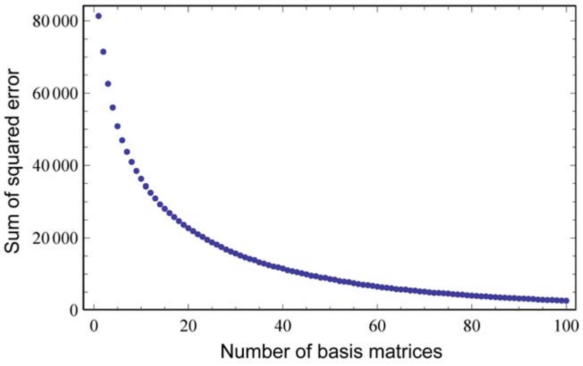

Figure 4. Convergence of NNMF. The sum of squared error decreases as more basis matrices are included.
doi:10.1371/journal.pone.0028898.g004

Figure 6 displays the correlations of the rates in the basis matrices for the first 5 factorizations with 5 amino acid properties (chemical composition, polarity, volume, isoelectric point and hydropathy). The values for these properties were obtained from [26]. Here we are correlating the rate of substitution between two amino acids with the difference between their values of the relevant property. As expected, negative correlations predominate: amino acids with larger differences are less frequently exchanged. The horizontal black line (at -0.169) indicates the threshold for significant negative correlation ($p<0.01$, one-tailed correlation test, $n=190$). The relationships between the chemical properties and the basis matrices clearly vary across the factorizations. For instance, the fifth basis matrix for $r=5$ (which as we saw corresponds to an elevation of the overall exchangeability of Tryptophan) corresponds with significant conservation of polarity, isoelectric point and hydropathy (evidently, exchanging Tryptophan for another amino acid does not affect these properties very much on average), but no conservation of chemical composition or volume (Tryptophan substitutions do affect these properties).

NNMF consistently yields better models than other approaches

For each of the 50 Pandit test alignments, we optimized the weight vectors and computed the AICc scores for the first 40 factorizations (from 1 to 40 basis matrices; we stopped at 40 because finding weights by maximum likelihood is computationally intensive, taking, for example, 2 to 3 hours to get up to 40 with datasets of around 600 codons and 50 sequences, but taking substantially longer as larger numbers of basis matrices are considered). The number of basis matrices that minimized the AICc was dependent on the alignment. This optimal number ranged from 11 to 40, with a median of 30.5 and an interquartile range (IQR) of 11. Figure 7 shows the distribution of the optimal number of basis matrices for the best NNMF model across all 50 test datasets.

From the 50 test datasets, we also computed AICc scores for the MOER model, as well as for each named amino acid model implemented in HyPhy, the REV model and the REV 1-step model (which fixes to 0 the rates of all amino acid substitutions that require more than one nucleotide change). Following Burnham and Anderson [27], we compute ΔAICc scores, which are the

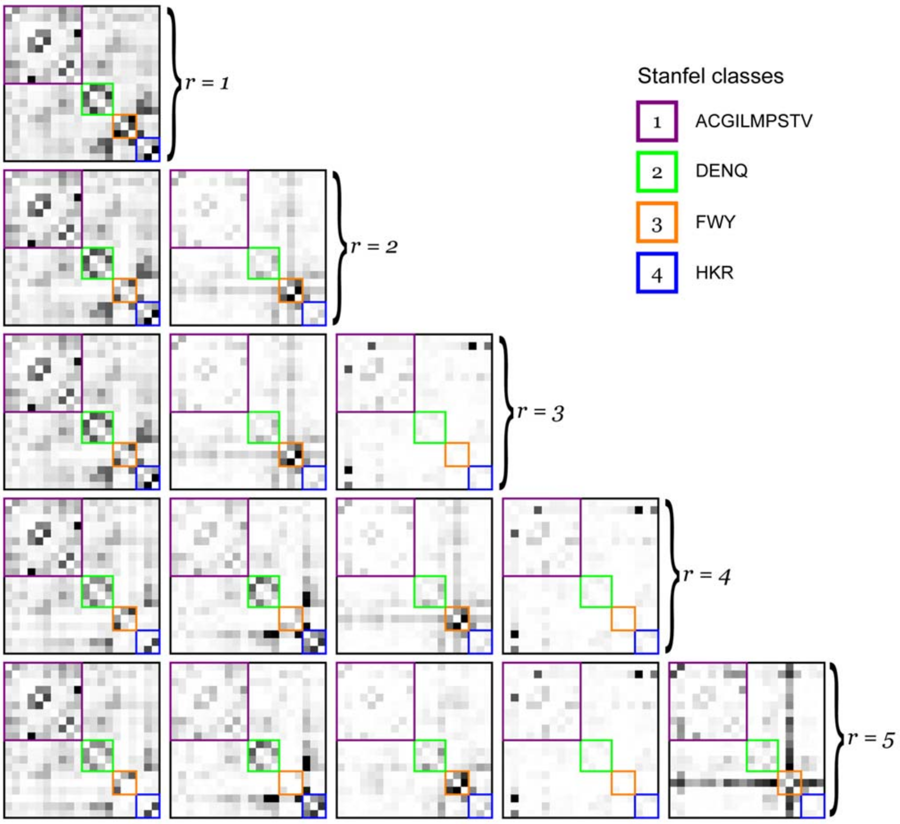

Figure 5. NNMF basis matrices. The set of NNMF basis matrices obtained for ranks ranging from 1 to 5. Amino acids are ordered according to their Stanfel classification [25]. Rates are indicated in grayscale, with pure white being a rate of zero and pure black being the maximum rate in the matrix.
doi:10.1371/journal.pone.0028898.g005

AICc scores for each model minus the best AICc for that dataset. The best model will thus have $\Delta\text{AICc}=0$. Models with $\Delta\text{AICc}>10$ have "essentially no support" [27]. Table 2 summarizes the frequency of each model's ΔAICc scores. The NNMF procedure for finding models appears to consistently outperform the others, obtaining the best AICc on 49 of 50 datasets. REV won on a single alignment, which, unsurprisingly, was the largest alignment and thus able to justify the full 190 rate parameters. The best NNMF model on this dataset had a ΔAICc of 0.34, which indicates that it has only slightly less support than REV.

Our approach of selecting the factorization rank using AICc is equivalent to selecting the best of the 40 NNMF models under consideration. Such a model selection step arguably gives NNMF an unfair advantage over the other models; although it is not standard procedure in the AIC literature, it may be more correct to add a penalty to the AICc scores of NNMF. Though not strictly appropriate for this context, a Bayesian argument can be used to estimate the appropriate size of this penalty: if we are comparing NNMF as a whole procedure against a single other model and we distribute the prior probability for NNMF uniformly over the 40 NNMF candidate models, we would introduce a penalty of at most $-\log\frac{1}{40}\approx 3.7$ to the resulting marginal likelihood for the NNMF procedure. This would amount to a maximum AICc penalty of approximately 7.4 to the scores for NNMF. Applying this penalty in Table 2 does not substantially affect the results. Furthermore, if we fix the number of basis matrices used (we picked 20) for all alignments, we still outperform WAG (the best overall fixed model) on all alignments with a median AICc improvement of 225 points. This is despite removing the model's ability to adapt its complexity to suit the data. That the improvement remains is not surprising: even a fixed amount of flexibility is better than none, as long as it does not require too many parameters for any particular alignment.

It is also interesting to look at the AICc scores excluding the NNMF models (Table 3). Here we see MOER finding the best model most often (21/50 times), with WAG a close second (15/50) and REV and REV 1-step next with 8/50 and 6/50 respectively. Predictably, most of the specialist models (mtMAM, mtREV 24, HIVwithin and HIVbetween) perform badly on datasets they were not intended for, with the exception of rtREV, which outperforms both JTT and Dayhoff (38, 10 and 2 wins respectively). Interestingly, in [13], rtREV was outperformed by generalist models WAG and JTT on HIV alignments containing the reverse transcriptase protein.

The use of constant rates across sites is an unrealistic assumption. It is possible to incorporate rate variation in a Random Effects Likelihood (REL) framework, where the rate at a site is modeled as a random draw from a discretized distribution. This incurs additional computational expense proportional to the number of rate categories used. To demonstrate that our results hold when rate variation is incorporated into all models, we

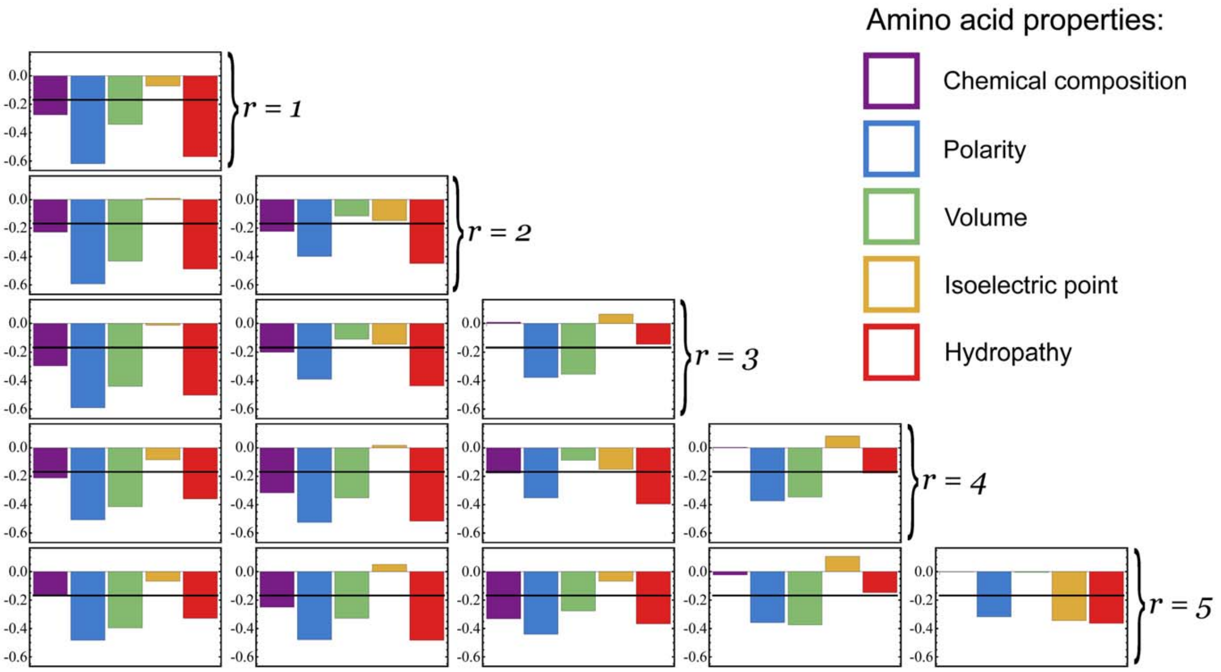

Figure 6. NNMF basis matrices correlate with amino acid properties. The correlations between amino acid properties and the basis matrices. The horizontal black line (at -0.16867) indicates the threshold for significant negative correlation ($p<0.01$, one tailed, $n=190$).
doi:10.1371/journal.pone.0028898.g006

randomly selected 10 test alignments and accounted for rate variation using a discretized gamma distribution with 4 rate categories. Table 4 displays the results for these 10 datasets. The conclusions are unchanged, and NNMF yields the best models for all 10 alignments.

NNMF models yield different phylogenies with better likelihoods

The Robinson-Foulds distance between the trees found using the WAG matrix and those found using the best NNMF model ranged from 0 to 98, with a median of 19 and an IQR of 24. This shows that the choice of model makes a difference to the estimated phylogeny. The NMMF phylogenies also have much higher likelihoods (and lower AICc scores) than the phylogenies estimated using WAG. When using maximum likelihood as a criterion for optimizing phylogenies, topologies and models that yield higher likelihoods should be preferred. This is not direct evidence that the NNMF procedure leads to more accurate trees (which would be difficult to demonstrate for a convincingly large sample), but it does suggest that we should expect such an improvement.

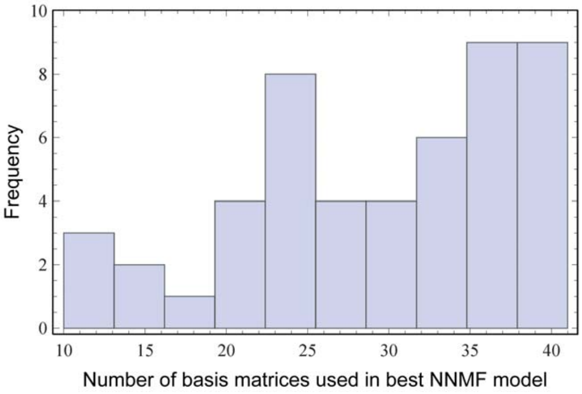

Figure 7. Distribution of the optimal number of basis matrices. The number of basis matrices that minimized the AICc across 50 test alignments.
doi:10.1371/journal.pone.0028898.g007

Bigger differences in likelihoods predict bigger differences in phylogenies. Figure 8 shows the relationship between the mean log-likelihood improvement per site for a given alignment and the Robinson-Foulds distance between the two resulting topologies. There is a strong positive correlation with $\rho=0.657$, $p=2\times10^{-6}$ (randomization test with 10^6 replicates). The slope of the best fitting line is 38.1, indicating a Robinson-Foulds distance increase of ≈ 38 for each log-likelihood per-site improvement.

Discussion

Model selection tools such as ModelTest [28] and its amino acid counterpart ProtTest [16] have been widely adopted for selecting the best fitting models for a given alignment. In this paper we show that, rather than simply selecting the best from a list of existing models, models of protein evolution can be tailored to specific alignments. Our NNMF framework has two primary strengths: 1) the model complexity adapts to fit the alignment, and 2) the dimensions along which the model can vary and the trajectory along which the complexity increases have been learnt, at least approximately, from a large collection of real alignments.

Since NNMF finds higher quality exchangeability matrices, we should expect it to benefit any application that uses such matrices. In this paper, we demonstrate an impact on phylogeny inference. Although we don't demonstrate it here, these rate matrices can also be used to construct scoring matrices for sequence alignments.

Table 2. ΔAICc scores for all models.

	0	≤2	≤4	≤8	≤16	≤32	≤64	≤128	≤256	≤512	≤1024	≤2048	≤4096	≤∞
NNMF	49	1												
MOER							1	7	18	20	3	1		
REV	1							2	4	7	4	6	5	21
REV-1 step									7	28	15			
Equal Input												14	27	9
Dayhoff										8	25	16	1	
JTT									2	11	24	12	1	
WAG								6	16	23	5			
rtREV									2	21	23	4		
mtMAM												11	30	9
mtREV 24											11	30	9	
HIVwithin												16	26	8
HIVbetween											6	29	14	1

Each table entry is the number of datasets with ΔAICc in that range. For any dataset, the best model has ΔAICc = 0. A model with ΔAICc > 10 has essentially no support.

doi:10.1371/journal.pone.0028898.t002

A procedure for doing this, along with software for generating the scoring matrices, is outlined in [13]. Given that an alignment is required before NNMF can be used, an iterative procedure, in which a guide alignment obtained from a standard scoring matrix is used to estimate an NNMF model, would have to be adopted. A scoring matrix based on this model can then be generated to refine the alignment.

Using more basis matrices

On our test alignments, we explored up to 40 basis matrices. This choice was motivated by computational considerations. The histogram of the optimal number of basis matrices for each dataset (Figure 7) suggests that using more basis matrices could lead to further improvement on some alignments. We provide basis matrices for the first 100 factorizations, so users can explore as many dimensions as their computational restrictions allow. It is worth pointing out that, when the number of basis matrices becomes 190, the NNMF model is equivalent to the REV model. This justifies the interpretation of the procedure as interpolating between a model with no flexibility and a fully flexible one.

Other approaches

CodonTest [26] is a recently proposed approach to solving a similar problem using a different approach, but at the codon rather than amino acid level. A genetic algorithm is used to find an optimal number of non-synonymous rate classes, as well as an assignment of particular non-synonymous substitution rates to these classes. The difference in the 'level' of modeling (codon *vs* protein) is superficial: applying our approach to codon models would be straightforward, though at some extra computational expense. The approach of CodonTest is different, in that it explores a much larger space of possible parameter clusters. While the difference in levels prevents direct comparison, we expect the NNMF approach to gain some additional leverage over that of CodonTest, because the set of subspaces it explores is learnt from a collection of training alignments, while CodonTest does not incorporate this prior information.

During the final preparation of this manuscript we became aware of recent work by Zoller and Schneider [29] in which a similar problem is tackled using an approach based on dimensionality reduction, again in the context of codon models rather than amino acid models. They used principal components analysis (PCA) to estimate a set of basis matrices, and, as in our approach, constructed their final model as a linear combination of these basis matrices. PCA has the advantage of being more computationally efficient than NNMF, but it lacks the non-negativity constraints. It is thus possible that certain linear combinations of PCA basis matrices will yield rates that are smaller than 0. Zoller and Schneider [29] circumvent this problem by explicitly resetting all negative rates to 0. That their model is applied to codon level data prevents a direct comparison, but future work will surely necessitate comparing different methods of dimensionality reduction for this task. We see their work as an encouraging sign that there is fertile ground for applying dimensionality reduction to phylogenetic models of evolution.

Practical recommendations

Our NNMF approach can be applied whenever a numeric model of amino acid evolution is required. The following procedure would appear sensible: First, estimate a guide tree using a fixed protein model. Then use the NNMF HBL program to find the best NNMF model. At this point, the model could be used to re-estimate the guide tree and iterate the NNMF procedure. Since each iteration should improve the model selection criterion (which is also bounded), this procedure should converge. Finally, the output can be converted to the form appropriate for the remaining analysis (phylogeny estimation, alignment etc). Some publicly available empirical rate matrices are provided with a fixed set of equilibrium frequencies. Importantly, our NNMF procedure used the empirical amino acid frequencies, and there are no such frequencies associated with any of our rate matrices, so any applications requiring equilibrium frequencies should use either the empirical frequencies, or estimate the equilibrium frequencies by maximum likelihood.

Rate variation may be introduced at any step. To save computation, one could use the NNMF HBL script without rate variation to obtain a rate matrix, and subsequently introduce rate variation. With more computational resources, rate variation can

Table 3. ΔAICc scores without NNMF.

	0	≤2	≤4	≤8	≤16	≤32	≤64	≤128	≤256	≤512	≤1024	≤2048	≤4096	≤∞
MOER	21			2	6	7	2	4	2	5		1		
REV	8						1	2	2	4	2	6	4	21
REV-1 step	6					1	4	4	18	13	4			
Equal Input												18	24	8
Dayhoff								1	2	13	24	9	1	
JTT								2	6	13	23	6		
WAG	15	2	3	3	3	7	4	2	5	6				
rtREV			1					3	18	18	9	1		
mtMAM												17	25	8
mtREV 24											24	19	7	
HIVwithin											1	17	24	8
HIVbetween											8	31	11	

Each table entry is the number of datasets with ΔAICc in that range. For any dataset, the best model has $\Delta\text{AICc}=0$. A model with $\Delta\text{AICc}>10$ has essentially no support.

doi:10.1371/journal.pone.0028898.t003

be included while optimizing over the combination weights. It is an open question whether including rate variation when estimating the original REV models (before the NNMF step) would significantly improve subsequent steps that also include rate variation. Results reported in [26] suggest that rate variation should be mostly orthogonal to estimating the relative substitution rates.

An approximate solution to a harder problem

Learning basis matrices by NNMF can be seen as an approximation to a more computationally challenging problem. It is possible to express the likelihood function for the factorization directly:

$$P(D|\theta)=\prod_{i=1}^{m} P(D_i|\sum_{j=1}^{r} w_{ij}\times B_j) \qquad (4)$$

where D_i is the i^{th} alignment in the training set, the likelihood within the sum is computed, as usual, using Felsenstein's pruning algorithm [4], and θ is the full collection of parameters, including weights and basis matrices. In this formulation, the rates in the basis matrices B_j and the combination weights w_{ij} could all be optimized numerically to maximize the overall likelihood on the training data. However, obtaining this optimal solution would be computationally challenging – our NNMF procedure approximates this by finding separate REV models that maximize the likelihood on each alignment, and then finding the factorization that most closely recovers these REV models in the mean square error sense. The implicit assumption is that this factorization will also yield good likelihoods. The computational saving relative to the full solution occurs in part because the REV models can be optimized separately for each training alignment.

Table 4. ΔAICc for all models with gamma rate variation (4 categories).

	0	≤2	≤4	≤8	≤16	≤32	≤64	≤128	≤256	≤512	≤1024	≤2048	≤4096	≤∞
NNMF	10													
MOER						1	3	4	2					
REV										1		1		8
REV-1 step										9	1			
Equal Input												7	2	1
Dayhoff									1	2	7			
JTT									2	4	4			
WAG							1	5	4					
rtREV								2	6	2				
mtMAM												8	2	
mtREV 24											8	2		
HIVwithin												6	3	1
HIVbetween											3	6	1	

Each table entry is the number of datasets with ΔAICc in that range. For any dataset, the best model has $\Delta\text{AICc}=0$. A model with $\Delta\text{AICc}>10$ has essentially no support.

doi:10.1371/journal.pone.0028898.t004

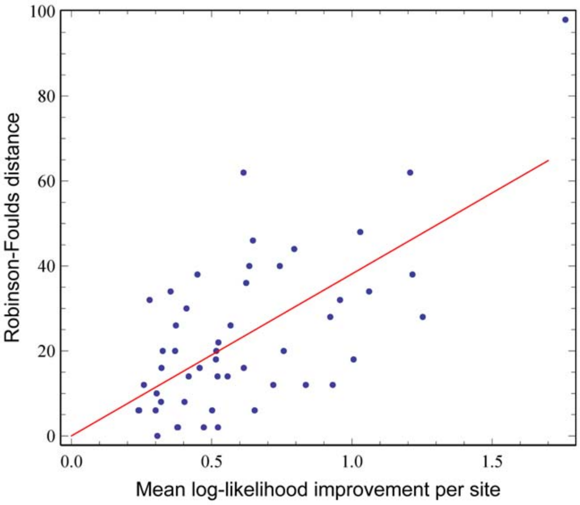

Figure 8. Likelihood improvement predicts phylogenetic difference. The difference between phylogenies increases as the mean likelihood difference per site between NNMF and WAG increases. $\rho = 0.657$, ($p = 2\times10^{-6}$, randomization test with 10^6 replicates). Assuming intercept of 0, slope = 38.1. Without this assumption, intercept = -0.31, slope = 38.5.
doi:10.1371/journal.pone.0028898.g008

Future avenues for research

Estimating a model of evolution that is specific to a single alignment clearly improves on the generalist approach. It is still, however, an incredibly coarse approximation to reality. The constraints and selective pressures on each site are most likely unique, but estimating a model for each site would be intractable, both computationally and statistically. Goldman *et al.* [30] took early steps in this direction, allowing the model of evolution to vary from site to site by using a Hidden Markov Model to capture the correlational structure across sites. Lartillot and Philippe [31] introduce a model that allows each site to belong to one of a number of classes, which differ in their equilibrium frequencies. A Dirichlet process prior is adopted to accommodate uncertainty about the number of classes, as well as the assignment of sites to classes. Le and Gascuel [32] also allow the substitution matrices to vary across sites. In their approach, they assume a small number (2 or 3) of distinct substitution processes, and their model treats each site as a random draw from one of these processes. This works well when clues about which process belongs to which site are available, but when the whole procedure is unsupervised the optimization appears to be difficult and sensitive to initial conditions [32,33]. Developing unsupervised approaches for estimating such models with larger numbers of distinct processes is an intriguing avenue for future research.

Acknowledgments

We thank Prof. Sergei Kosakovsky Pond for use of the UCSD computing cluster.

Author Contributions

Conceived and designed the experiments: BM KS. Performed the experiments: BM TW JB RK SM GB LdB DK TH. Analyzed the data: BM KS. Wrote the paper: BM KS.

References

1. Dayhoff MO, Eck RV, Park CM (1972) A model of evolutionary change in proteins. In: Dayhoff M, ed. Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington, D.C., volume 5. pp 89–99.
2. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff M, ed. Atlas of Protein Sequence and Structure, National Biomedical Research Foun- dation, Washington, D.C., volume 5, suppl. 3. pp 345–352.
3. Lipman DJ, Altschul SF, Kececioglu JD (1989) A tool for multiple sequence alignment. Proceedings of the National Academy of Sciences of the United States of America 86: 4412–4415.
4. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution 17: 368–376.
5. Kosakovsky Pond SL, Poon AF, Leigh Brown AJ, Frost SD (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to inuenza A virus. Mol Biol Evol 25: 1809–1824.
6. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–282.
7. Kosiol C, Goldman N (2005) Different Versions of the Dayhoff Rate Matrix. Molecular Biology and Evolution 22: 193–199.
8. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–699.
9. Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochon- drial DNA. Journal of molecular evolution 42: 459–468.
10. Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15: 1600–1611.
11. Cao Y, Waddell PJ, Okada N, Hasegawa M (1998) The complete mitochondrial DNA sequence of the shark Mustelus manazo: evaluating rooting contradictions to living bony vertebrates. Mol Biol Evol 15: 1637–1646.
12. Dimmic MW, Rest JS, Mindell DP, Goldstein RA (2002) rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. Journal of molecular evolution 55: 65–73.
13. Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, et al. (2007) HIV-Specific Probabilistic Models of Protein Evolution. PLoS ONE 2: e503+.
14. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788–791.
15. Devarajan K (2008) Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. PLoS Comput Biol 4: e1000029+.
16. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21: 2104–2105.
17. Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676–679.
18. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24: 1586–1591.
19. Guindon SA, Dufayard JFA, Lefort V, Anisimova M, Hordijk W, et al. (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59: 307–321.
20. Burnham KP, Anderson D (2002) Model Selection and Multi-Model Inference Springer, 2nd edition.
21. Posada D, Buckley TR (2004) Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. Systematic biology 53: 793–808.
22. Robinson D (1981) Comparison of phylogenetic trees. Mathematical Biosciences 53: 131–147.
23. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.
24. Whelan S, de Bakker PIW, Quevillon E, Rodriguez N, Goldman N (2006) PANDIT: an evolution- centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Research 34: D327–D331.
25. Stanfel L (1996) A New Approach to Clustering the Amino Acid. Journal of Theoretical Biology 183: 195–205.
26. Delport W, Scheffler K, Botha G, Gravenor MB, Muse SV, et al. (2010) CodonTest: Modeling Amino Acid Substitution Preferences in Coding Sequences. PLoS Comput Biol 6: e1000885+.
27. Burnham KP, Anderson DR (2004) Multimodel Inference. Sociological Methods & Research 33: 261–304.
28. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinfor- matics (Oxford, England) 14: 817–818.
29. Zoller S, Schneider A (2011) A new semi-empirical codon substitution model based on principal component analysis of Mammalian sequences. Mol Biol Evol;Advance access.
30. Goldman N, Thorne JL, Jones DT (1998) Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. Genetics 149: 445–458.

31. Lartillot N, Philippe H (2004) A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Molecular Biology and Evolution 21: 1095–1109.
32. Le SQ, Lartillot N, Gascuel O (2008) Phylogenetic mixture models for proteins. Philosophical Transactions of the Royal Society B: Biological Sciences 363: 3965–3976.
33. Le SQ, Gascuel O (2010) Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics Is Clearly Beneficial. Systematic Biology 59: 277–287.