Abstract

FERNANDES, ANDREW DELLANO. Quantifying Phylogenetic Conservation in Protein Molecular Evolution. (Under the direction of William R. Atchley.)

This dissertation examines the problem of quantifying amino acid conservation in proteins molecular evolution. Ideally, this conservation is quantified by inferring the rate of evolution at each amino acid site of a multiple-alignment. However, current rate-inference methods have three problematic assumptions. The methods assume that (a) the rates of all sites are independent, (b) the rates are drawn from a known prior distribution, and (c) the mean rate across sites is approximately one. The problems are two-fold. First, the assumptions of site-rate independence and known mean rate are contradictory. To see the contradiction, consider a two-site alignment with known rate of ~0.5 at site one. The rate at site two is unknown under the independent-sites assumption, but is ~1.5 by the assumption of known mean rate. Second, if the rates are drawn from a known prior distribution, the assumption of known distribution implies the question "which distribution?". Previous work has focused only on selecting better families of rate distributions, often at the expense of additionally parameterizing the evolutionary model. Herein, I develop a method of inferring rates requiring only the assumption of known mean rate, and not requiring additional parameterization. Thus a model of evolution based on our method is a more general framework for inferring rates than previous work. Since a known mean rate is required to distinguish evolutionary rate from time, our method is arguably the most general possible that allows rate and time to be fully and independently identified. The method is assessed by investigating conservation in the Myc, Max, and p53 transcription-factor families.

Quantifying Phylogenetic Conservation in Protein Molecular Evolution

by

Andrew Dellano Fernandes

A dissertation submitted to the Graduate Faculty of

North Carolina State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Biomathematics

Raleigh, North Carolina

2006

Approved by:

William R. Atchley
Chair of Advisory Committee

Charles E. Smith

Eric A. Stone

Steffen Heber

Dedication

This dissertation is dedicated to my wife, Bonnie Deroo, for all her love, kindness, and friendship, and to our dog, Keemun, who doesn't understand genetics, statistics, or mathematics but who nevertheless contributed to the successful completion of this work.

Biography

Andrew Dellano Fernandes graduated in 1994 from the University of Waterloo, Ontario, with an Honors Co-Operative Bachelor of Science in both Biochemistry and Mathematics. His Master of Science degree in Mathematics was completed in 1998 under Michael C. Mackey of McGill University, Québec, at the Center for Nonlinear Dynamics in Physiology and Medicine. This dissertation represents one of the final requirements of his Doctor of Philosophy degree under the direction of William R. Atchley at the North Carolina State University. His major is Biomathematics and minor is Genetics.

Acknowledgements

Heartfelt thanks and appreciation go to my advisor, Bill Atchley, for the personal, professional, and financial support he provided.

My spouse, Bonnie Deroo, supported me in untold ways. Words fail me.

Thanks go to advisory committee, including Jeff Thorne, Eric Stone, Steffen Heber, and Charlie Smith, for their thoughts, comments, and time.

Fellow lab-member Kevin Scott kept me statistically honest.

Special thanks are in order to both faculty and staff of the Biomathematics Program and Department of Genetics at the North Carolina State University for providing much-needed support and camaraderie.

Financial support was gratefully received from the National Institutes of Health (GM45344), and the North Carolina State University.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation examines the problem of quantifying amino acid conservation in proteins molecular evolution. Ideally, this conservation is quantified by inferring the rate of evolution at each amino acid site of a multiple-alignment. However, current rate-inference methods have three problematic assumptions. The methods assume that (a) the rates of all sites are independent, (b) the rates are drawn from a known prior distribution, and (c) the mean rate across sites is approximately one. The problems are two-fold. First, the assumptions of site-rate independence and known mean rate are contradictory. To see the contradiction, consider a two-site alignment with known rate of ~0.5 at site one. The rate at site two is unknown under the independent-sites assumption, but is ~1.5 by the assumption of known mean rate. Second, if the rates are drawn from a known prior distribution, the assumption of known distribution implies the question "which distribution?". Previous work has focused only on selecting better families of rate distributions, often at the expense of additionally parameterizing the evolutionary model.

Herein, I develop a method of inferring rates requiring only the assumption of known mean rate, and not requiring additional parameterization. Thus a model of evolution based on our method is a more general framework for inferring rates than previous work. Since a known mean rate is required to distinguish evolutionary rate from time, our method is arguably the most general possible that allows rate and time to be fully and independently identified. The method is assessed by investigating conservation in the Myc, Max, and p53 transcription-factor families.

The work is presented in a somewhat counterintuitive order. Originally, this dissertation began with Chapter 3, an investigation of site conservation within the p53 family of tumor suppressor proteins. I determined that most methods of quantifying conservation were deficient in at least two areas. First, phylogenetic correlation among sequences was often unaccounted for. This is a particular problem for p53 because most known sequences are from the closely related chordate clade. Without correcting for phylogenetic correlation, excess weight is given to the chordates, overwhelming information from non-chordate

clades. Second, some conservation measures did not correct for amino acid physiochemical similarity and evolutionary exchangeability. Thus the small change of leucine to isoleucine was given the same weight as the large change of phenylalanine to proline. The only measure of conservation incorporating phylogenetic and amino acid similarity data was evolutionary rate, as given in standard stochastic models of protein evolution. Unfortunately, estimating rates through these models used the three assumptions of site independence, known distribution, and fixed mean rate, as described above. Chapter 2 describes the methods I developed to avoid these assumptions and the problems that they create.

The process of developing this method required the many computations of likelihood integrals. Calculation of these integrals consumes the majority of computer time in many molecular evolution studies, including ours. Evolutionary biologist Joseph Felsenstein noted in 2001 that the numerical integration technique of Gaussian quadrature could be used to significantly speed up evaluation of these integrals. Unfortunately, integration schemes for only a few specific forms of the integral were known. Chapter 1 presents work extending the evaluation of likelihood integrals by Gaussian quadrature to arbitrary integrands. It has been reviewed by Felsenstein and has been accepted for publication in *Evolutionary Bioinformatics Online*.

Therefore this dissertation is presented in reverse-telescopic order: investigations of the p53 family begat investigations of site-specific rate inference. Rate inference in turn begat questions regarding the evaluation of likelihood integrals. First, I solved the problem of integration. Next, I completed the work on rate inference. Finally, our results were applied to the p53 family.

I hope that the tools developed herein will be valuable to the field of molecular evolution in general, and look forward to building on them in the future.

Chapter 1

Gaussian Quadrature Formulae for Arbitrary Positive Measures

(Arbitrary Gaussian Quadrature Formulae)

Andrew D. Fernandes[1,2,4] & William R. Atchley[1,2,3]

[1]Graduate Program in Biomathematics
[2]Center for Computational Biology
[3]Department of Genetics
North Carolina State University
Raleigh, NC 27695-7614

[4]Corresponding Author
andrew@fernandes.org

Abstract

We present computational methods and subroutines to compute Gaussian quadrature integration formulas for arbitrary positive measures. For expensive integrands that can be factored into well-known forms, Gaussian quadrature schemes allow for efficient evaluation of high-accuracy and -precision numerical integrals, especially compared to general *ad hoc* schemes. In addition, for certain well-known density measures (the normal, gamma, log-normal, Student's *t*, inverse-gamma, beta, and Fisher's *F*) we present exact formulae for computing the respective quadrature scheme.

Availability: Source code is freely available online as a C-linkable ISO C++ library under a BSD-style license from ⟨http://www.fernandes.org/gaussqr⟩. The library may be built using single, double, or extended precision arithmetic.

Contact: Andrew D. Fernandes ⟨andrew@fernandes.org⟩

Motivation

This paper is concerned with the efficient and accurate calculation of likelihood integrals of the form

$$\Pr(H|D) \propto \int_{h \in H} \Pr(D|h) \cdot \Pr(h)\, dh, \tag{1}$$

through the construction of a Gaussian-type quadrature scheme that is optimized specifically for the known prior distribution $\Pr(h)$. Our specific motivation stems from studies in the molecular evolution of protein sequences where it is important to take variation of evolutionary rates among sites into account when inferring phylogenies. In the context of this specific problem, both Felsenstein (2001; 2004) and Mayrose et al. (2005) pointed out that Gaussian quadrature formulae can be used to provide more accurate and more rapidly convergent numerical integration methods than the more common "equal percentile" method of Yang (1994). Unfortunately, Gaussian-type quadrature formulae have only been derived for a relatively small number of prior distributions. In the context of molecular evolution, the two most common priors are the gamma and log-normal distributions. Gaussian quadrature formulae for the gamma distribution are already known as "Generalized Gauss-Laguerre" quadrature (Felsenstein 2001), although admittedly the mathematical similarity between these schemes is not obvious with the usual formulation of Gauss-Laguerre quadrature. Thus their equivalence is generally not appreciated. Unfortunately, until now explicit Gaussian quadrature formulae were not available for log-normal (or other) priors commonly used in computational biology.

The purpose of this paper is to provide an efficient and rapid algorithm with accompanying computer library that permits computation of Gaussian quadrature rules for *arbitrary* prior distributions. In some cases, we derive analytic formulae for specific common distributions. Although motivated by a specific application to integrals found in the field of molecular evolution, we stress that our methods (and computer code) are applicable to the solution of numerical integration problems in general.

Problem Statement

We wish to find a set $i = 0,1,2,\ldots,(n-1)$ of weights w_i and abscissae x_i such that the approximation

$$\int_a^b w(x)\cdot f(x)\,dx \approx \sum_{i=0}^{n-1} w_i \cdot f(x_i) \tag{2}$$

is exact whenever f is a polynomial of degree $2n-1$ or less, and $w(x)$ is a known "weight function". In our case $w(x)$ represents the positive density measure of our prior likelihood. A good and complete modern reference covering the theory of Gaussian (and related) types of quadrature rules can be found in Gautschi (2004). If f is expanded as a polynomial series, inspection suggests that any quadrature scheme will depend on the raw moments of $w(x)$. Indeed, defining the (real) inner product

$$\langle f\,|\,g\rangle = \int f(x)\,g(x)\cdot w(x)\,dx, \tag{3}$$

it is well known that there always exists a set of polynomials, orthogonal with respect to this inner product, such that

$$\begin{gathered} p_{-1} = 0, \quad p_0 = 1 \\ p_{i+1}(x) = (x - a_i)\cdot p_i(x) - b_i \cdot p_{i-1}(x), \quad i = 0,1,2,\ldots \end{gathered} \tag{4}$$

and where the recurrence coefficients a_i and b_i can be calculated explicitly from

$$\begin{aligned} a_i &= \frac{\langle x\cdot p_i\,|\,p_i\rangle}{\langle p_i\,|\,p_i\rangle}, \quad i = 0,1,2,\ldots \\ b_i &= \frac{\langle p_i\,|\,p_i\rangle}{\langle p_{i-1}\,|\,p_{i-1}\rangle}, \quad i = 1,2,\ldots \end{aligned} \tag{5}$$

with the coefficient b_0 being arbitrary and set by convention such that $b_0 = \int w(x)\,dx$. Therefore the first n recursion coefficient pairs are uniquely determined by the first $2n$ moments of the measure w. Once the coefficients a_i and b_i are known, they can be assembled into the tridiagonal Jacobi matrix

$$J = \begin{bmatrix} a_0 & \sqrt{b_1} & & & & \\ \sqrt{b_1} & a_1 & \sqrt{b_2} & & & \\ & \sqrt{b_2} & \ddots & \ddots & & \\ & & \ddots & \ddots & \sqrt{b_{n-2}} & \\ & & & \sqrt{b_{n-2}} & a_{n-2} & \sqrt{b_{n-1}} \\ & & & & \sqrt{b_{n-1}} & a_{n-1} \end{bmatrix}. \quad (6)$$

The desired abscissae x_i are then equal to the eigenvalues of J, and the desired weights are given by the relationship

$$w_i = b_0 \cdot q_{i,0}^2, \quad (7)$$

where $q_{i,0}$ is the first component of the normalized eigenvector q_i of matrix J.

Solution Methods

Formulae are known that explicitly express the recursion coefficients a_j and b_j in terms of the raw moments of w. Unfortunately, these explicit representations are extremely ill-conditioned and thus are not usable even for "well behaved" weight functions or quadrature schemes of fairly low order n. If the integrals of Equation (5) can be calculated efficiently and accurately, Stieltjes' Procedure calculates the recursion coefficients via iterative application of Equations (4) and (5) forming the sequence $\{p_{-1}, p_0\} \to \{a_0, b_1\} \to \{p_1\} \to \{a_1, b_2\} \to \{p_2\} \to \cdots$. Although better behaved than explicit computation, Stieltjes' Procedure also tends to be moderately ill-conditioned (Press, Teukolsky et al. 1997) and therefore of limited value. Alternatively, the Sack-Donovan-Wheeler algorithm (Press, Teukolsky et al. 1997) has been suggested as a way to overcome the numerical instabilities inherent in Stieltjes' Procedure by utilizing modified moments rather than raw moments in Equations (4) and (5). The downside of the Sack-Donovan-Wheeler algorithm is its reliance on the *a priori* selection of a "good" polynomial basis for the moments of $w(x)$, in itself a fairly difficult and subjective procedure that is dependent on

heuristic approximation of the moments of $w(x)$. Forming such an approximation may be as or more difficulty than solving the original problem.

Recently, a general-purpose and unconditionally stable algorithm to calculate Gaussian weights and abscissae for any positive measure has been proposed (Gander and Karp 2001). The method is based on the observation (Boley and Golub 1987) that the *discrete* measure

$$\omega_m(x) = \sum_{i=0}^{m-1} \omega_i \cdot \delta(x - \xi_i) \tag{8}$$

can have its weights and abscissae assembled into a sparse matrix

$$W_m = \begin{bmatrix} 1 & \sqrt{\omega_0} & \sqrt{\omega_1} & \cdots & \sqrt{\omega_{m-2}} & \sqrt{\omega_{m-1}} \\ \sqrt{\omega_0} & \xi_1 & & & & \\ \sqrt{\omega_1} & & \xi_2 & & & \\ \vdots & & & \ddots & & \\ \sqrt{\omega_{m-2}} & & & & \xi_{m-2} & \\ \sqrt{\omega_{m-1}} & & & & & \xi_{m-1} \end{bmatrix} \tag{9}$$

that is orthogonally similar to the Jacobi matrix

$$J_m = \begin{bmatrix} 1 & \sqrt{b_0} & & & & & \\ \sqrt{b_0} & a_0 & \sqrt{b_1} & & & & \\ & \sqrt{b_1} & a_1 & \sqrt{b_2} & & & \\ & & \sqrt{b_2} & \ddots & \ddots & & \\ & & & \ddots & \ddots & \sqrt{b_{m-2}} & \\ & & & & \sqrt{b_{m-2}} & a_{m-2} & \sqrt{b_{m-1}} \\ & & & & & \sqrt{b_{m-1}} & a_{m-1} \end{bmatrix}, \tag{10}$$

where b_0 is the ω_m-measure of the entire domain of ω_m. Gander and Karp showed that if a sequence of discrete measures given by Equation (8) converged to a given continuous measure such that $\lim_{m\to\infty} \omega_m(x) = w(x)$, then J_m would similarly converge to the recurrence coefficients of the continuous measure. Such convergence had already been noted and

exploited several years before in the ORTHPOL software package (Gautschi 1994). A re-implementation, modernization, and modification of some of Gautschi's algorithms form the core of our work. To continue, given J_m, standard eigen-decomposition algorithms for symmetric tridiagonal matrices can be used to compute the Gaussian quadrature weights and abscissae for the given weight function. In summary, the weights and abscissae of an arbitrary positive measure $w(x)$ can be as determined by first finding a discrete $\omega_m(x)$ that approximates $w(x)$ "well enough", using the Lanczos reduction algorithm to transform $W_m \to J_m$, concomitantly obtaining the recursion coefficients $\{a_i, b_i\}$, and then eigen-decomposing J_m to determine the final weights and abscissae $\{x_i, w_i\}$ via Equation (7).

Algorithmic Details

The implementation details for the overall process, starting from a given weight function and ending with a set of Gaussian quadrature weights and abscissae, are best elucidated by a worked example. Assume we are given the weight function $w(x) \propto e^{-x}$, $x \geq 0$, where we do not know the normalization constant $1 \Big/ \int w(x)\, dx$ and do not recognize e^{-x} as the weight function for the well-known Gauss-Laguerre quadrature scheme. Our first step is to select a sequence of measures, as per Equation (8), that converges to the measure $e^{-x}\, dx$. Following Gautschi (1994), we use a classical numerical integration scheme to approximate $\int w(x)\, dx$, namely the Fejér Type-2 integration rule (Gautschi originally used the Fejér Type-1 rule). Fejér integration rules are very similar to the well-known Clenshaw-Curtis integration rules over the domain $z \in [-1,1]$. However, the Fejér rules are open-ended, do not require evaluation at the domain endpoints, and are therefore more suitable for measures with non-compact support. Fejér Type-2 rules also have an efficiency advantage over the Type-1 rules in the fact that the n-point Type-2 abscissae are an interleaved subset the $(2n+1)$-point Type-2 abscissae. Therefore, the Type-2 rules allow us to reuse all previously calculated ordinates when the number of integration points is doubled. Lastly, Fejér Type-2 integration weights can be calculated very rapidly via real inverse Fast Fourier Transform (Waldvogel

2006), allowing a large number of points to be efficiently utilized in approximating $\int w(x)\,dx$. The supplied subroutine fejer2_abscissae calculates the required abscissae and integration weights $\{z_i, q_i\}$ for a given number of abscissae $i = 0,1,2,\ldots,(m-1)$. The transformation $g(z) = \dfrac{(1+z)}{(1-z)}$ is used via the subroutine map_fejer2_domain to map $z \in (-1,1) \rightarrow x \in (0,\infty)$ and change the variable of integration such that $\int_0^{\infty} e^{-x}\,dx = \int_{-1}^{+1} e^{-g(z)} \cdot g'(z)\,dz$, giving the final abscissae and weights $\{\xi_i, \omega_i\}$ for Equation (9), where $\xi_i = g(z_i)$ and $\omega_i = q_i \cdot w(g(z_i)) \cdot g'(z_i)$. Note that the subroutine map_fejer2_domain is capable of mapping the Fejér interval to other arbitrary finite and non-finite domain intervals in addition to the particular transformation $g(z)$ utilized here.

The tridiagonalization of W_m in Equation (9) to J_m in Equation (10) can be accomplished by using the subroutine lanczos_tridiagonalize, a subroutine that exploits the sparsity structure of Equation (9) via the Lanczos algorithm (Golub and Van Loan 1996) for efficient tridiagonalization. Lastly, the eigen-decomposition of J_m in Equation (10) and subsequent calculation of the final Gaussian quadrature rule for $w(x)$ via Equation (7) is accomplished by use of the subroutine gaussqr_from_rcoeffs, where the eigen-decomposition is performed using a modified implicit-shift QL algorithm. Note that the coefficient b_0 returned from lanczos_tridiagonalize estimates $\int w(x)\,dx$ for the given m. Thus, we can set $b_0 = 1$ prior to calling gaussqr_from_rcoeffs to normalize $w(x)$ without explicitly knowing or calculating the actual normalization coefficient. In many cases, this can significantly speed up the calculation of $w(x)$. For common distributions such as the normal, gamma, log-normal, and others, the utility function standard_distribution_rcoeffs is supplied to compute recursion coefficients directly.

Lastly, we must ensure that m is large enough so that $\omega_m(x)$ approximates $w(x)$ sufficiently closely to further ensure that the $i = 0,1,2,\ldots(n-1) < m$ computed quadrature

points $\{x_i, w_i\}$ converge. The subroutine relative_error computes the maximum relative error between its two vector arguments. Since w_i is guaranteed to be positive for all non-negative measures $w(x)$, it suffices (and simplifies matters) to verify convergence of w_i without explicit regard to the convergence of x_i.

Implementation Details

In using the subroutines presented, there are a few subtleties in the overall procedure that can be exploited in order to address non-standard situations or increase computational efficiency. First, we note that the discrete measure denoted by Equation (8) can be used to approximate *any* finite union of disjoint intervals. For instance, if we wished to use the (admittedly contrived) implicit weight function

$$w(x) \propto \begin{cases} e^{-x}, & 0 \le x < 1 \\ 1/x^2, & 1 \le x \end{cases} \tag{11}$$

over support $0 \le x$. Our subroutines could be applied twice, once for each continuous interval, yielding two discrete-measure approximations, each with approximate normalization consonant. The two discrete measures could then be combined into a set of abscissae and weights $\{\xi_i, \omega_i\}$ that would then be subject to the Lanczos tridiagonalization procedure in order to determine the recursion coefficients of Equation (11). Note that the normalization of Equation (11) is computed "on the fly" and therefore allows great flexibility in choosing the weight function $w(x)$. Furthermore, note that the example weight function of Equation (11) is not even continuous at $x = 1$.

Second, we note that computing an m-node Fejér Type-2 integration scheme is done by performing a real inverse fast Fourier transform of size $(m+1)$. Although the subroutine supplied is capable of computing inverse Fourier transforms of almost arbitrary size, the transform is efficient *only* if $(m+1)$ has divisors from the set $\{2,3,4,5\}$. To further increase efficiency, we note that the Fejér Type-2 nodes are simple to compute via

$$z_i = \cos\left(\frac{(i+1)\cdot\pi}{m+1}\right), \quad i = 0,1,\ldots,(m-1), \tag{12}$$

implying that an m_1-point and m_2-point integration scheme will share common abscissae if (m_1+1) and (m_2+1) have a common divisor. Having common abscissae imply that previously computed values of $w(g(z_i))$ could be reused as m increases, thus increasing the efficiency of approximating $w(x)$. Therefore the recommended sequence of m for fejer2_abscissae follows $\{3,7,15,31,63,\ldots\}$. For very simple, well-behaved weight functions, it may be preferable to simply use m of a few hundred or few thousand, and not worry excessively about convergence when m is small. Such an approach may be indicated when pre-computing quadrature schemes for a parameterized family of weight functions; the shape parameter of the unit-mean gamma distribution, for example. Rather than determining quadrature points for every desired shape parameter, it may make more sense to pre-compute weights and abscissae as functions of the shape parameter at particular parameter values, and then interpolate a quadrature scheme for all "in-between" parameter values. Obviously, Fejér nodes and weights can be pre-computed as well.

There may be situations where it is useful to know the analytic form of a particular weight function's recursion coefficients. In particular, well-known density functions can often have their recurrence relationships determined by Stieltjes' Procedure, and a representative sample of such is shown in Table 1. Recursion coefficients computed from this table can be supplied directly to subroutine gaussqr_from_rcoeffs, although better numeric stability may be achieved by approximating these densities via standard_distribution_rcoeffs. Note that Gaussian quadrature schemes may not exist for all distributions at all parameter values. In these cases, non-existence of the quadrature scheme is due to the non-existence of the distribution's relevant higher-order moments. In any case, caution should be exercised in utilizing Table 1 for these distributions lest numerical truncation error inadvertently become too great. Lastly, as Table 1 shows, it is often possible to extract a common factor λ from the recursion coefficients. Such a common factor merely scales the eigenvalues of J_m while leaving the eigenvectors alone, and thus may be safely ignored prior to eigen-decomposition.

We conclude with a reminder that our choice of the Fejér Type-2 integration points for computing the approximation $\lim_{m\to\infty} \omega_m(x) = w(x)$ is quite arbitrary, and other integration schemes may be more appropriate given a different family of weight functions. For instance, a simple $1/m$ "equal-percentile" approach, reminiscent of Yang (1994), may be more efficient than a Fejér-like scheme for weight functions with numerous sharp peaks. Further, rational-quadrature schemes may be a better choice for measures with poles near the measure's support (Gautschi 1999; Weideman and Laurie 2000; Van Deun, Bultheel et al. 2006). In any case, the Fejér Type-2 scheme utilized here should prove adequate for most common weight functions utilized in likelihood calculations today.

Usage Guidelines

Two approximations must be made to construct a set of quadrature abscissae and weights. First, the number of discrete points that will be used to approximate the weight function must be chosen. Second, the number of quadrature points to compute the final likelihood integral must be chosen. In this section, we provide guidance on how to select the appropriate number of points in each case.

First, when approximating $w(x)$ by a discrete measure, we exploit efficiencies inherent in the FFT and sparsity structure of matrices W_m and J_m to quickly and efficiently approximate $w(x)$ with thousands (1023, 2047, or more) points. For example, using 1023 points to approximate a standard $N(0,1)$ distribution results in quadrature coefficients, correct to within one part in 2×10^{-15} (the limit of machine precision), to be calculated in negligible time compared to all but the most trivial phylogenetic likelihood calculations.

Guidance for the second case, the number of quadrature points to use, is more difficult to give because of the main convergence property of Gaussian quadrature: the rate of convergence depends critically on how well the integrand can be approximated by a polynomial. The better the approximation, the more rapid the convergence. Unfortunately, the converse is also true; functions that are poorly approximated by polynomials may have far *worse* convergence characteristics than other numerical integration schemes. The best

guidance on picking the number of quadrature points for a particular integrand may come from trial and error: keep increasing the number of points until numerical convergence seems to be achieved. This empirical "try it and see" approach has been utilized by Yang (1994), Mayrose et al. (2004), among others and is commonly advised.

In an effort to provide a more concrete example of how Gaussian quadrature fares in a sample integrand from molecular evolution studies, consider one site of a four sequence alignment where every nucleotide is different (one each of A, C, G, and T), and we know *a priori* that all four sequences share an unknown common ancestor one time unit in the past. Assuming a normalized Jukes-Cantor (1969) model of evolution yields a likelihood function of

$$f(r) \propto \left(1 + 3 \cdot e^{-\frac{4}{3}r}\right)\left(1 - e^{-\frac{4}{3}r}\right)^3 \tag{13}$$

for a given evolutionary rate r. We assume unit-proportionality for convenience. Further assuming that rates are distributed according to a unit-mean Gamma distribution with coefficient of variation $\sqrt{2}$ results in a weight function of

$$w(r) = 4 \cdot r \cdot e^{-2r}. \tag{14}$$

The likelihood of our data given our model can then be calculated analytically, resulting in

$$\int_0^\infty h(r)\, dr = \frac{30080}{53361} \approx 0.5637076, \tag{15}$$

where

$$h(r) = w(r) \cdot f(r). \tag{16}$$

A graph depicting the relative shapes of f, g, and h is shown in Figure 1. A plot of the relationship between the number of quadrature points and the relative error of the integral in Equation (15) is shown in Figure 2. Seven quadrature points result in a relative error of about $0.15\,\%$, and twenty points result in a relative error of about $1.1 \times 10^{-6}\,\%$. Note that seven or more quadrature points demarks the asymptotic domain for numerical convergence where the error decreases polynomially with the number of quadrature points.

A detailed examination of the twenty-quadrature point case shows an interesting optimization that applies to likelihood functions such as Equation (13), where the likelihood approaches a constant value as its argument approaches infinity. Recall that Gaussian quadrature schemes are designed to optimally integrate polynomials $p(x)$, and that complex analysis tells us that for polynomials, $|p(x)| \to \infty$ as $|x| \to \infty$. For $w(x) \cdot p(x)$ to be integrable, $|w(x)| \to 0$ relatively rapidly as $|x| \to \infty$. Therefore we expect the quadrature weight w_i to rapidly become very small as the magnitude of its respective abscissa x_i increases. An illustration of the magnitudes of $\{x_i, w_i\}$ for a twenty-point quadrature scheme for our $h(r)$ example, above, is shown in Figure 3. Note that after the first ten to twelve abscissae have been summed, the contribution of the remaining eight to ten points will be negligible; the integration scheme assumes that $f(r)$ will be polynomially large when in fact it is almost constant. Thus we can gain the accuracy benefits of using a twenty-point integrator while incurring the cost of only ten evaluations of $f(r)$.

Acknowledgements

The authors wish to thank Jeff Thorne and Joseph Felsenstein for helpful comments and suggestions during manuscript review. Financial support was provided by the National Institutes of Health (GM45344) and the North Carolina State University.

Table 1

Exact recursion coefficients for selected probability distributions. For Equations (6) and (7), we scale the recursion coefficients such that $a_i = \lambda \cdot a_i'$ and $b_i = \lambda^2 \cdot b_i'$. Note that for n recursion coefficients, at least the first $2n$ moments must exist. There is also a special case for the Beta distribution: $a_0 = 1/2$ when $\alpha = \beta = 1$ (the uniform distribution).

Distribution	Non-normalized Density Function	Domain, if not $\mathbf{R}$	a_i' $i = 0, 1, 2, 3, \ldots, (n-1)$	b_i' $i = 1, 2, 3, \ldots, (n-1)$	λ
Normal	$\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$	$\sigma^2 > 0$	μ	$i \cdot \sigma^2$	1
Gamma	$(x)^{\alpha-1} \cdot \exp(-x/\beta)$	$x > 0$ $\alpha, \beta > 0$	$\alpha + 2i$	$i \cdot (\alpha + i - 1)$	β
Log-Normal	$\left(\frac{1}{x}\right) \cdot \exp\left(\frac{-(\ln(x)-\mu)^2}{2\sigma^2}\right)$	$x, \sigma^2 > 0$ $\varsigma = e^{\sigma^2}$	$\varsigma^{(2i-1)/2} \cdot (\varsigma^i \cdot (\varsigma + 1) - 1)$	$\varsigma^{(3i-2)} \cdot (\varsigma^i - 1)$	e^{μ}
Student's t	$\left(1 + \frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$	$\nu > 0$	0	$\frac{i \cdot \nu \cdot (\nu - i + 1)}{(\nu - 2i) \cdot (\nu - 2i + 2)}$	1
Inverse Gamma	$(x)^{-\alpha-1} \cdot \exp(-\beta/x)$	$x > 0$ $\alpha, \beta > 0$	$\frac{(\alpha + 1)}{(\alpha - 2i + 1) \cdot (\alpha - 2i - 1)}$	$\frac{i \cdot (\alpha - i + 1)}{(\alpha - 2i) \cdot (\alpha - 2i + 1)^2 \cdot (\alpha - 2i + 2)}$	β
Beta	$(x)^{\alpha-1} \cdot (1-x)^{\beta-1}$	$0 < x < 1$ $\alpha, \beta > 0$ $\gamma = \alpha + \beta$	$\frac{\alpha \cdot \gamma + (2i-2) \cdot \alpha + 2i \cdot \beta + i \cdot (2i-2)}{(\gamma + 2i) \cdot (\gamma + 2i - 2)}$	$\frac{i \cdot (\gamma + i - 2) \cdot (\alpha + i - 1) \cdot (\beta + i - 1)}{(\gamma + 2i - 1) \cdot (\gamma + 2i - 2)^2 \cdot (\gamma + 2i - 3)}$	1
Fisher's F	$\left(\frac{1}{x}\right) \cdot \sqrt{\frac{x^{\nu_1}}{(\nu_1 \cdot x + \nu_2)^{(\nu_1 + \nu_2)}}}$	$x > 0$ $\nu_1, \nu_2 > 0$	$\frac{(\nu_1 \cdot \nu_2 + 2 \cdot \nu_1 + 4i \cdot \nu_2 - 8i^2)}{(\nu_2 - 4i - 2) \cdot (\nu_2 - 4i + 2)}$	$\frac{2i \cdot (\nu_1 + 2i - 2) \cdot (\nu_2 - 2i + 2) \cdot (\nu_1 + \nu_2 - 2i)}{(\nu_2 - 4i) \cdot (\nu_2 - 4i + 2)^2 \cdot (\nu_2 - 4i + 4)}$	$\frac{\nu_2}{\nu_1}$

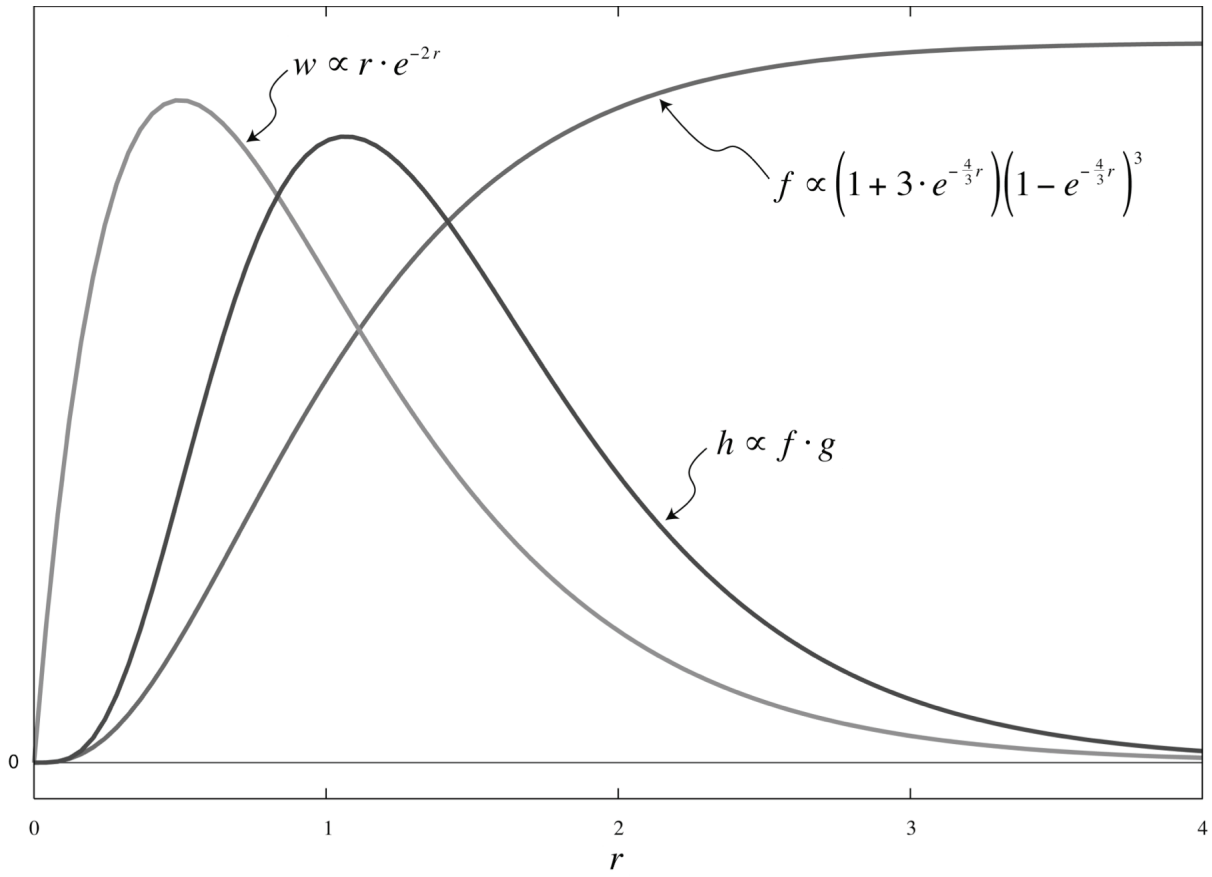

Figure 1

A graphical depiction of the relative shapes of Equations (13), (14), and (16).

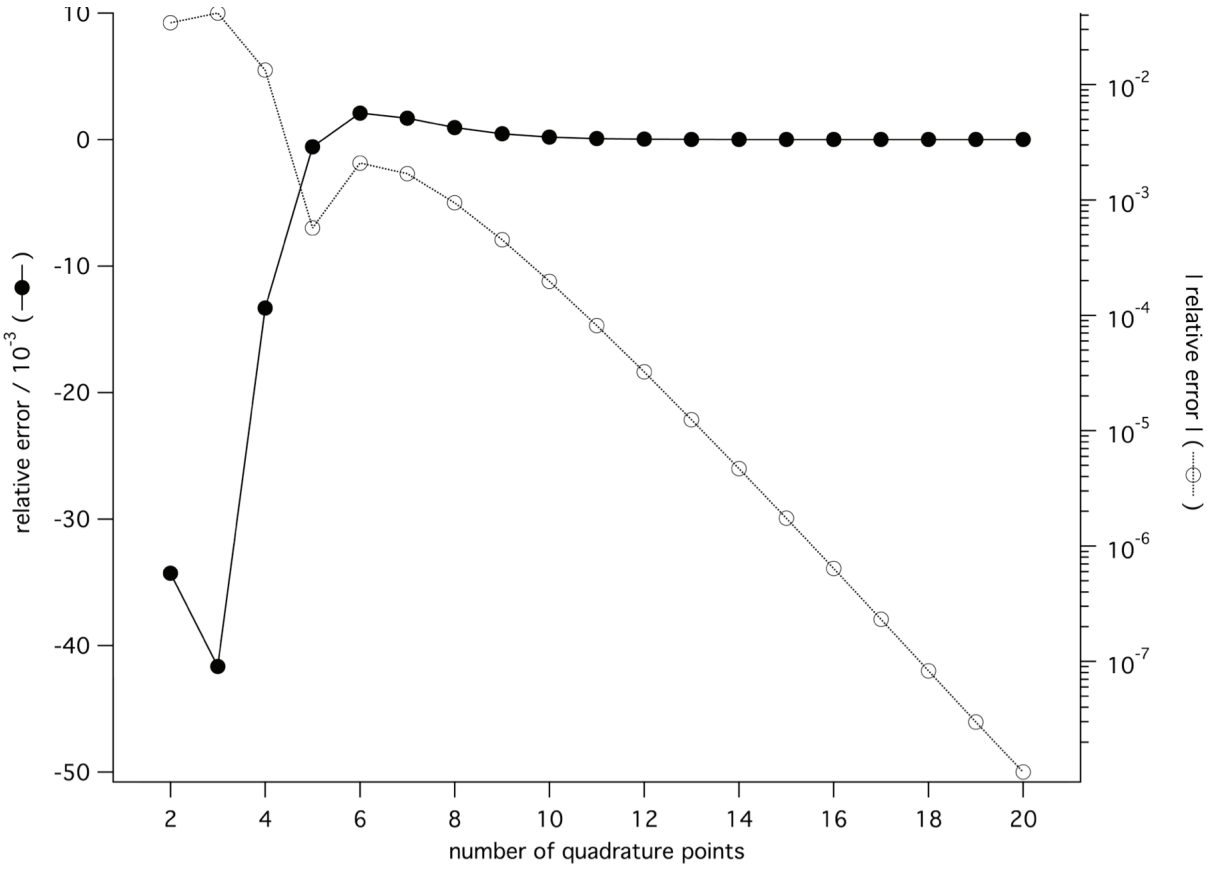

Figure 2

The number of quadrature points versus the relative error in the sample molecular evolution integration problem.

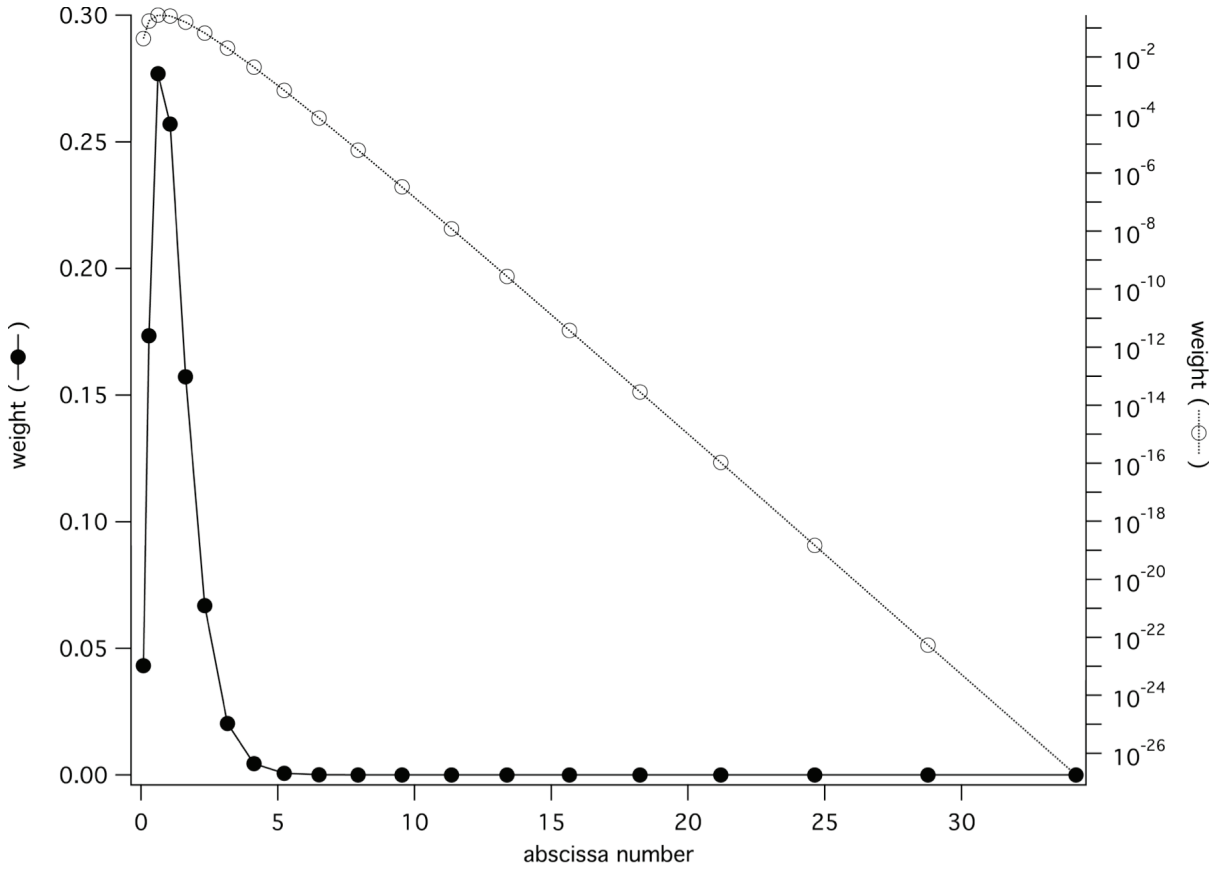

Figure 3

Gaussian quadrature weights and abscissae for $n = 20$ points for the sample molecular evolution integration problem.

References

Boley, D. and G. H. Golub (1987). "A survey of matrix inverse eigenvalue problems." Inverse Problems 3(4): 595-622.

Felsenstein, J. (2001). "Taking Variation of Evolutionary Rates Between Sites into Account in Inferring Phylogenies." Journal of Molecular Evolution 53(4 - 5): 447.

Felsenstein, J. (2004). Inferring Phylogenies. Sunderland, MA, USA, Sinauer Associates.

Gander, M. J. and A. H. Karp (2001). "Stable computation of high order Gauss quadrature rules using discretization for measures in radiation transfer." Journal of Quantitative Spectroscopy & Radiative Transfer 68(2): 213-223.

Gautschi, W. (1994). "Algorithm 726: ORTHPOL - A Package of Routines for Generating Orthogonal Polynomials and Gauss-Type Quadrature Rules." ACM Transactions on Mathematical Software 20(1): 21-62.

Gautschi, W. (1999). "Algorithm 793: GQRAT - Gauss quadrature for rational functions." ACM Transactions on Mathematical Software 25(2): 213-239.

Gautschi, W. (2004). Orthogonal polynomials: computation and approximation. New York, Oxford University Press.

Golub, G. H. and C. F. Van Loan (1996). Matrix computations. Baltimore, Johns Hopkins University Press.

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. Mammalian Protein Metabolism. H. N. Munro. New York, Academic Press. 3: 21-132.

Mayrose, I., N. Friedman, et al. (2005). "A Gamma mixture model better accounts for among site rate heterogeneity." Bioinformatics 21: 151-158.

Mayrose, I., D. Graur, et al. (2004). "Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior." Molecular Biology and Evolution 21(9): 1781-1791.

Press, W. H., S. A. Teukolsky, et al. (1997). Numerical recipes in C: The art of scientific computing. Cambridge, Cambridge University Press.

Van Deun, J., A. Bultheel, et al. (2006). "On computing rational Gauss-Chebyshev quadrature formulas." Mathematics of Computation 75: 307-326.

Waldvogel, J. (2006). "Fast Construction of the Fejér and Clenshaw-Curtis Quadrature Rules." BIT Numerical Mathematics 46(1): 195-202.

Weideman, J. A. C. and D. P. Laurie (2000). "Quadrature rules based on partial fraction expansions." Numerical Algorithms 24(1 - 2): 159.

Yang, Z. (1994). "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods." Journal of Molecular Evolution 39(3): 306-14.

Chapter 2

Detecting Conserved Motifs using Site-Specific Rate Inference with Objective Non-Informative Priors

(Detecting Conserved Motifs using Site-Specific Rate Inference)

Andrew D. Fernandes[1,2,4] & William R. Atchley[1,2,3]

[1]Graduate Program in Biomathematics
[2]Center for Computational Biology
[3]Department of Genetics
North Carolina State University
Raleigh, NC 27695-7614

[4]Corresponding Author
andrew@fernandes.org

Abstract

Motivation: The detection of conserved protein sequence motifs depends critically on the precise definition of "conservation" used. Current measures of conservation suffer from one or more of three flaws: they may not account for similarities or differences among amino acids, they may not correct for the phylogenetic correlation among the given sequences, or they may depend to an unknown extent on a parameterized *ad hoc* prior distribution. We propose a novel method of detecting conserved motifs that does not suffer from these flaws. The method first infers site-specific rates of evolution using a class of objectively non-informative rate priors, and then associates low rates of evolution with conserved, and hence functionally or structurally important, sites. An extensive analysis of conservation within proteins from the Myc-Max transcription factor family is given, with detailed comparison to previous methods of detecting conservation.

Results: Given a sequence alignment, an estimate of the alignment's phylogeny, and a model of amino acid substitution, our method efficiently estimates the rate of evolution at each site using a prior chosen from a class of objectively non-informative rate priors. While generally superior to other methods of detecting conservation, the results are critically dependent on the quality of the sequence alignment analyzed.

Contact: Andrew D. Fernandes ⟨andrew@fernandes.org⟩

William R. Atchley ⟨bill@atchleylab.org⟩

Keywords: site-specific rate variation; protein evolution; likelihood; conservation; Bayesian estimation; prior distribution; Markov-chain Monte Carlo

Introduction

In proteins, the rate of evolution at each residue site is expected to vary according to the specific selective constraints at that site. Sites that are under strong selective constraints should be relatively highly conserved, while sites under lesser selective pressure should be more variable, with the observed rate corresponding to the level of purifying selection at that site (Kimura 1983). Therefore conserved sites are often strongly suggestive of structural or functional importance. Such conserved residues may be involved ligand binding, secondary or tertiary structure conformation, DNA binding, protein folding, or protein-protein interactions. A relatively small, contiguous set of relatively conserved sites that accurately identify a specific set of homologous proteins, such as the basic Helix-Loop-Helix (bHLH) transcriptional regulators, is called a *sequence signature* or *sequence motif* (Atchley, Terhalle et al. 1999; Atchley and Fernandes 2005). This paper is concerned with the problem of determining which sites, given an aligned set of protein sequences and their associated phylogeny, should be denoted as being conserved and are therefore candidate sequence motifs.

Previous Work

Although many different site-specific conservation scores for multiple alignments have been proposed in the past (reviewed in Valdar 2002), none of these methods make full use of the information contained in the phylogenetic tree that relates the sequences, or the evolutionary exchangeability of different amino acids. A conservation score based on entropy that incorporated phylogenetic information was introduced by Wollenberg and Atchley (2000), but this score did not account for the effective similarity (such as isoleucine and leucine) or dissimilarity (such as proline and tryptophan) of individual amino acids. More recently still, at least two independent frameworks, namely MrBayes (Ronquist and Huelsenbeck 2003) and rate4site (Pupko, Bell et al. 2002; Mayrose, Graur et al. 2004) have been proposed to measure site conservation for amino acid data by inferring the evolutionary rate at each alignment site using a detailed probabilistic model of sequence evolution. Since evolutionary rates measure the expected number of amino acid replacement events per unit time, lower evolutionary rates indicate fewer replacements, and hence greater conservation;

conversely, higher rates indicate proportionally less conservation at that site. Both of these programs infer evolutionary rates using a Bayesian framework in essentially similar manners. For brevity, we will describe the general procedure utilized by these programs to infer rates, ignoring minor details where they differ.

Assume an alignment A and phylogenetic tree T such that site i of the alignment, denoted by A_i, as shown in Figure 1. Branch lengths of the phylogenetic tree represent the average number of amino acid replacements between nodes of each branch, averaged across all sites. The site-specific rate r_i indicates how quickly site i evolves relative to the average. Therefore, following standard likelihood calculations, we can write

$$P\left(A_i|T,r_i\right)=\sum_x P(x)\left(\sum_y P\left(y|x,r_i t_6\right)P\left(R|y,r_i t_1\right)P\left(K|y,r_i t_2\right)\right) \times\left(\sum_z\left(P\left(z|x,r_i t_8\right)P\left(N|z,r_i t_3\right)\sum_w P\left(w|z,r_i t_7\right)P\left(I|w,r_i t_4\right)P\left(L|w,r_i t_5\right)\right)\right), \quad (17)$$

where $P(x)$ denotes the background (prior) probability of amino acid x, and $P\left(y|x,r_i t_j\right)$ denotes the probability that amino acid x will be replaced by amino acid y along a branch length of t_j if the site is evolving at rate r_i. In practice, Felsenstein's (1981) post-order tree traversal algorithm is used to calculate this likelihood efficiently. If we use a time-reversible model of amino acid substitution, as is usually the case, the root node could be placed anywhere on the tree without changing the computed likelihood. To infer rates given the data, Bayes' Theorem can be used to set

$$P\left(r_i|T,A_i\right)\propto P\left(A_i|T,r_i\right)\cdot P\left(r_i\right). \quad (18)$$

Thus we are left with the necessity of specifying the prior density $P\left(r_i\right)$ for each of the rates.

Almost without exception, the Gamma distribution is used for the rate prior (Yang and Kumar 1996), simply because it is a "nice" unimodal distribution supported over the non-negative real numbers. Occasionally, the log-Normal distribution is used instead of the Gamma distribution, but its use has largely fallen out of favor (Felsenstein 2001). The

Gamma distribution with shape parameter $\alpha > 0$ and inverse-scale parameter β has mean α/β and variance α/β^2. Setting $\alpha = \beta$ yields a distribution of mean one and variance $1/\alpha^2$, with the shape of the density being determined by α alone. When $\alpha > 1$ the density is bell-shaped and unimodal; as α gets larger there is progressively less rate heterogeneity and rates tend to be progressively closer to one. When $\alpha \leq 1$ the density is L-shaped in the manner of the Exponential distribution; as α shrinks the density becomes progressively more skewed and obtains a wider range of rate heterogeneity.

Once a rate prior has been selected, the mean and variance of the posterior rate distribution can now be calculated via

$$\begin{aligned} \mathrm{E}\left[r_i \middle| A,T\right] &= \int_0^\infty r_i \cdot P\left(r_i \middle| A_i,T\right) dr_i = \bar{r}_i \\ &\propto \int_0^\infty r_i \cdot P\left(A_i \middle| r_i,T\right) \cdot P\left(dr_i\right) \end{aligned} \tag{19}$$

and

$$\begin{aligned} \mathrm{Var}\left[r_i \middle| A,T\right] &= \int_0^\infty \left(r_i - \bar{r}_i\right)^2 \cdot P\left(r_i \middle| A_i,T\right) dr_i \\ &\propto \int_0^\infty \left(r_i - \bar{r}_i\right)^2 \cdot P\left(A_i \middle| r_i,T\right) \cdot P\left(dr_i\right) \end{aligned} . \tag{20}$$

The integrals given in Equations (19) and (20) are analytically intractable and must be approximated by an appropriate numerical scheme. The most popular is the "equal probability discretization" quadrature scheme introduced by Yang (1994) that approximates the integral

$$\int_{x \in X} f\left(x\right) dx \approx \sum_{i=1}^{m} w_i \cdot f\left(x_i\right) \tag{21}$$

for some discrete set of *knot* points $x_i \in X$ and a given set of *weights* w_i. Specifically, Yang's scheme partitions X into m equal-percentile sets, assigns x_i to be the mean of percentile-set i, and assigns all $w_i = 1/m$. Felsenstein (2001) has argued that more advanced

integration technique of *Gaussian quadrature* yields a higher accuracy result for comparatively fewer evaluations of $f(x_i)$. Gaussian quadrature uses a somewhat involved procedure to find a set of knots and weights for Equation (21) that yields numerically exact results for a large class of functions. Further details of constructing Gaussian quadrature schemes for arbitrary functions f can be found in Fernandes and Atchley (2006).

Finally, if alignment A consists of n sites, the likelihood of a given set of rates $\{r_1, r_2, \ldots, r_n\}$ can be computed by the relationship

$$\begin{aligned} P(r_1, r_2, \ldots, r_n | A, T) &= \prod_{i=1}^{n} P(r_i | A_i, T) \\ &\propto \prod_{i=1}^{n} P(A_i | r_i, T) \cdot P(r_i) \end{aligned}, \tag{22}$$

from which summary statistics can be computed. This is termed the Empirical Bayesian (EB) approach for inferring site-specific rates. The term "empirical" is used because the shape parameter α of the Gamma distribution rate prior is generally unknown and hence must be estimated from the data (Robbins 1956; Leonard and Hsu 2001). Because of the difficulty of choosing a prior distribution that is optimal from everyone's differing points of view, it has been argued that no prior for the rates should be assumed, the equivalent of assuming $P(r_i) = 1$. This choice of prior is termed the *Laplace prior* and implies that Equation (22) can now be interpreted in a Maximum Likelihood (ML) framework if desired. In the ML interpretation, the best estimates $\hat{r}_i$ for the rate parameters r_i are

$$\hat{r}_i = \arg\max_{r_i} \left[P(A_i | r_i, T) \right], \tag{23}$$

with related significance and variance estimates coming from classical ML theory. In fact, the original formulation of rate4site (see Pupko, Bell et al. 2002) used an ML framework to infer evolutionary rates. They found that the lack of a rate prior could, in rare circumstance, lead to the case where $\hat{r}_i \to \infty$ if an alignment site had all different amino acids at that site. However, later work on rate4site (see Mayrose, Graur et al. 2004) showed through extensive simulation tests that the EB approach was superior to ML methods for site-specific rate

inference; so much so, in fact, that the EB approach was recommended for all general use. In some sense, the superiority of Bayesian over ML methods in the context of rate inference was anticipated by Felsenstein (2001) who pointed out that a model with a separate rate for each site would result in the number of parameters increasing at the same pace as the increase in alignment length, causing the likelihood methods to lose their property of consistency. Thus many of the asymptotic convergence properties that ML methods rely on would no longer hold, implying that much of the standard ML machinery would be invalidated. Note that in a Bayesian context, estimating too many parameters from the data results in (potentially drastic) variance inflation, but does not invalidate the analytical context on which the estimation is based.

Criticisms of EB Rate Inference

Having described the EB methodology for rate inference, we turn now to criticisms of this methodology. Our goal is to formulate a method of rate inference that addresses these criticisms and ameliorates the difficulties that arise from them. The criticisms we present below are *not* intended to imply that EB rate inference is somehow incorrect. Rather, we hope to merely highlight the major assumptions and approximations inherent in EB rate inference, pointing out why the criticism is significant. After all, there is a large body of established literature that uses EB methods very similar to what we have described, and uses them with notable success. Any proposed new methodology would be suspect if it produced results that were not consistent with this body of previous work.

The Gamma Distribution

The first criticism involves the *ad hoc* selection of the Gamma (or Gamma-like) distribution as an appropriate data-independent prior distribution for the rates. The Gamma distribution is traditionally selected because its density can assume a wide variety of shapes over its parameter support and not because of any *a priori* biological relevance (Felsenstein 2001). In cases where inspection reveals that the Gamma distribution alone does not provide a sufficiently realistic model of rate heterogeneity, the distribution has been supplemented in various ways. For instance, provision for a fraction of invariant sites can be added (Gu, Fu et

al. 1995), a mixture of Gamma distributions can be utilized (Mayrose, Friedman et al. 2005), or other more general parameterized distributions (Kosakovsky Pond and Frost 2005) can be used. Generally speaking, the more flexible the parameterized family of distributions utilized for a prior, the more likely that the prior will capture the "true" distribution of rates that are present, but unobserved, in the data.

Unfortunately, it is difficult to determine when the rate prior has been made flexible enough to "correctly" model the data. For instance, we could utilize Bayes factors (Robert 2001) or Akaike's Information Criterion (AIC, 1974) to select between two different hypothetical rate priors: the first consisting of a Gamma distribution, the second consisting of a Gamma distribution augmented with a proportion of invariant sites. The latter model is clearly more flexible than the former and may well better model the existing rate heterogeneity. However, for a shape parameter greater than one, for example, both models assume that large rates are roughly exponentially less likely than lower rates. If this assumption is not true and a relatively large fraction of sites are evolving quickly, then the posterior rate distribution may be significantly yet undetectably biased due to the fact that neither prior adequately captures the true behavior of the rate distribution.

Distribution Parameters

Utilizing Gamma (or Gamma-like) distributions as rate priors requires estimation of a shape parameter from the data. As we make our prior more flexible, most schemes previously suggested in the literature require a greater number of parameters to be estimated from the data. Augmenting the Gamma distribution with an invariant proportion of sites requires estimation of that proportion. Using mixtures of Gamma distributions require estimating the shape of each density as well as their relative proportions. Unfortunately, adding too many parameters to our model for the distribution of rates can result in overfitting, and hence undesired inflation of each rate parameter's variance. Ideally we would prefer a prior that yields maximum flexibility with a minimum number of parameters, preferably none. If, however, sites were always treated as completely independent as per Equation (22), there would appear to be no non-parameterized distribution adequate to the task of modeling every possible distribution of rates.

Influence of the Prior

If an empirical prior is used, how can we know whether the data or the prior dominates the posterior? One simple idea is to use two alternative priors and check the resultant posteriors for agreement. But this procedure is rather indirect and, moreover, may be more informative about the two alternative priors than about the data. In some ways this restates the previous two sections that criticized both the use of the Gamma distribution and requirement for parameter estimation. Our point here is deeper, however. Given any set of alternative priors, how can we determine which one, if any, is the most appropriate?

The field of objective Bayesian analysis attempts to answer this question by turning the question on its head. Rather than verifying what comes out of an analysis, what if we are more careful about what goes into it? If the prior is designed to satisfy a rigorous and objective set of criteria that quantify our prior ignorance about the parameter in question, it is logical to assume that we have a better understanding of how the prior has influenced the posterior. An analogous situation is the observation that, by using better ingredients, we will likely bake a better cake. A prior that is designed to satisfy such criteria is termed an objective prior (Berger 2006). An excellent review of the selection of prior distributions by formal rules has been written by Kass and Wasserman (1996). Note that objective priors should not be automatically interpreted as being automatically "better" than either subjective or empirical priors. Rather, a prior is deemed "objective" when it has been chosen to satisfy objective criteria that capture, in a rigorous way, the precise nature of our ignorance. One may thus object to the criteria, but not the resultant prior.

Since rates are by definition non-negative any reasonable rate prior should have support of the non-negative real numbers and be integrable. Although improper (non-integrable) priors are often used in Bayesian analysis, their use in likelihood calculations as per Equation (17) results in improper posteriors. Given this restriction, relatively few distributions can model rates appropriately. Furthermore, most of these distributions tend to subjectively be quite similar. For example, for shape parameter greater than one, both Gamma and log-Normal distributions are unimodal and bell-shaped. Simply changing from one prior to the other would likely be insufficient to judge the influence of the prior over the

posterior. Unfortunately, it is unclear how to construct an objective prior that could be utilized within in the current framework for rate inference.

Reparameterization Invariance

One criterion that many feel is important for a Bayesian prior is re-parameterization invariance (Robert 2001). Reparameterization invariance means that no matter how our parameter of interest is measured, the prior always expresses the same equivalent belief. Stated mathematically, a prior density h is termed invariant if

$$h(\phi) = h\big(g(\phi)\big) \cdot \left| \frac{\partial g(\phi)}{\partial \phi} \right| \tag{24}$$

whenever g is a one-to-one differentiable transformation satisfying

$$f\big(x|\phi\big) = f\Big(g(x)\big|\, g(\phi)\Big) \cdot \left| \frac{\partial g(x)}{\partial x} \right| \tag{25}$$

for all x and ϕ, where $f\big(x|\phi\big)$ is the likelihood density. That is, it should make no difference whether evolutionary change is measured as a rate parameter ("average amino acid replacements per unit time") or its reciprocal, a time parameter ("average time per amino acid replacement"). Both quantities represent a measure of the same physical quantity, and therefore an analysis performed on one parameter should yield identical inferences to an analysis of the other. Unfortunately, the Gamma-family of distributions does not satisfy the invariance principle. Furthermore, given the described rate inference framework, it is not clear how to construct an invariant prior for site-specific rate parameters.

Confounding with Time

Perhaps the most serious criticism against the current EB rate inference method is the mathematical confounding of evolutionary rates with divergence times between the nodes of the phylogenetic tree. Thus, there is no way to distinguish between (a) doubling all of the inferred rates and (b) halving all of the tree's branch lengths. Unfortunately, lack of identifiability results in numerous practical problems, the most important of which is the fact that each rate is not an absolute measure. Instead, each rate is a measure only in relation to all

other rates. Generally, rates are relative to a mean rate of one. Thus, a rate of two indicates amino acid replacement is twice as fast as the average, a rate of half indicates a rate half the average. To interpret rates as relative measures, the rate prior distribution is usually constrained to have a unit mean to de-confound rate and time.

Unfortunately, constraining the prior alone is insufficient because it does not equivalently constrain the posterior. The posterior rate distribution remains with rate and time confounded. For example, suppose we have data x consisting of m samples of a Normal distribution that has known variance σ^2 and unknown mean μ. To estimate $\mathrm{E}\left[\mu | x, \sigma^2\right]$ in a Bayesian framework requires a prior distribution for μ. We assume the prior to be Gaussian such that $\mu \sim N\left(\mu_0, \sigma_0^2\right)$, where the mean μ_0 and variance σ_0^2 of the prior are arbitrary hyperparameters. Setting

$$P\left(\mu | x, \sigma^2\right) \propto P\left(x | \mu, \sigma^2\right) \cdot P(\mu) \tag{26}$$

yields

$$\mu | x, \sigma^2 \sim N\left(\mu_*, \sigma_*^2\right), \tag{27}$$

where

$$\mu_* = \sigma_*^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{m\bar{x}}{\sigma^2} \right) \quad \text{and} \quad \sigma_*^2 = \left(\frac{1}{\sigma_0^2} + \frac{m}{\sigma^2} \right)^{-1}. \tag{28}$$

Notice that no matter how tight the prior's variance σ_0^2 nor the location of the prior's mean μ_0, given enough samples eventually $\sigma_*^2 \to \sigma^2 / m \to 0$ and $\mu_* \to \bar{x}$. Restricting the prior of the mean does not restrict the mean of the posterior; the information inherent in the data eventually overwhelms the information provided by the prior. Returning to the case of Gamma distributed rate priors, the implication of our example is that restricting the rate prior to have unit mean does not restrict the mean of the rate posterior in any meaningful way. Without somehow restricting the mean of the posterior, rates and times remain confounded.

Why is confounding a problem? Consider the case where we hold the alignment A and tree topology T fixed but we wish to estimate not only the site-specific rates r, but the tree's branch lengths t as well. Then

$$\begin{aligned} P(r,t|A,T) &\propto P(A,T|r,t)\cdot P(r,t) \\ &\overset{?}{=} P(A,T|r,t)\cdot P(r)\cdot P(t) \end{aligned} \tag{29}$$

in the current framework. However, if rates and times are confounded, it is impossible to construct a meaningful prior for t, as $P(r,t)$ cannot be separated into $P(r)\cdot P(t)$. Of course, there is no *a priori* reason that rates and times be separable, but if they are not interpretation becomes very difficult. For instance, consider the simple problem of computing the conditional probability $P(r|t)$. Unless the probabilities of r and t are separable, such a seemingly simple statement is meaningless and cannot be computed. This is a very practical problem that has been tackled in a variety of ways.

Different researchers have attempted to solve the problems that emerge from rate-time confounding in different ways. MrBayes (version 3.1.2) does not attempt to adjust its branch lengths to ensure the posterior mean rate is equal to one. Simultaneous interpretation of the posterior rate and branch length distributions is therefore problematic since $P(r,t)$ is not integrable and is therefore not a density, as $P(r\eta,t/\eta)$ is constant for $\eta\in[0,\infty)$. In contrast, rate4site continually and forcibly adjusts its posterior rate distribution to have unit mean (Mayrose, Graur et al. 2004), and an iterative procedure as suggested by Meyer and von Haeseler (2003) is used to simultaneously infer rates and divergence times. In order to estimate branch lengths, first rates are estimated, keeping divergence times constant. The rates are normalized, and then the divergence times are inferred keeping the rates constant. The process is iterated until convergence. The question then is how to interpret the inferred parameters: neither the ML or EB frameworks fully apply to the result. Again, formulating a logically consistent interpretation is difficult because the assumption that site-specific rates are fully independent of each other necessarily implies that rates and times must be confounded.

Methods

Here we describe a model of rate heterogeneity that *explicitly* models rates as relative quantities; n rates are described by $n-1$ parameters, effectively decoupling rates and times. Although the site rates are no longer independent, the corresponding rate prior is much more general than the gamma distribution or related mixture models allow. We describe our method in three steps: how the rates are modeled, how an appropriate prior can be chosen, and how the posterior can be sampled.

Modeling Rates

Assume a multiple sequence alignment A and phylogenetic tree T. Let $r=\{r_1,r_2,\ldots,r_n\}$ be the set of evolutionary rates for the n alignment sites. Bayes' Theorem gives the posterior probability

$$P(r_1,r_2,\ldots,r_n|A,T)\propto P(A|r_1,r_2,\ldots,r_n,T)\cdot P(r_1,r_2,\ldots,r_n|T). \quad (30)$$

Since rate and time are confounded, a constraint is placed on $\{r_1,r_2,\ldots,r_n\}$ to separate rates from divergence times. Restricting the rates to almost any $(n-1)$-dimensional smooth manifold will suffice; the one chosen here is that the mean rate is equal to one:

$$\frac{r_1+r_2+\cdots+r_n}{n}=1. \quad (31)$$

Such a constraint is desirable for three reasons. First, this assumption is commonly used in phylogenetic studies involving site-specific rate heterogeneity. Second, Equation (31) sets a natural scale by which rates can be compared and then classified. Roughly speaking, conserved sites should have rate less than one and variable sites a rate greater than one. Third, if the rates are re-scaled and denoted by $\theta_i=r_i/n$, then each relative rate proportion $\theta_i\in[0,1]$ and $\sum\theta_i=1$. An instance of the complete set of rate proportions $\theta=\{\theta_1,\theta_2,\ldots,\theta_n\}$ can be interpreted as a coordinate of the n-dimensional unit simplex $\mathbb{S}^n\equiv\{\theta_1,\theta_2,\ldots,\theta_n|0\le\theta_i\le 1\ \forall i,\sum\theta_i=1\}$. Simplexes are the natural supports for both the

Multinomial distribution and the Dirichlet distribution, a fact used later. Graphical depictions of $\mathbb{S}^2$, $\mathbb{S}^3$, and $\mathbb{S}^4$ are shown in Figure 2. Note that the arithmetic mean of Equation (31) is not the only constraint that could be chosen. For instance, a constraint using the harmonic or geometric mean could be used instead. However, none of these alternative constraints have the justification or simplicity of interpretation of Equation (31). Also note that the rates of sites $1,2,\ldots,n$ are often considered as a sample from a larger (infinite) population, implying that their mean need not be exactly one. We consider including such sampling variation later.

An additional benefit of the arithmetic mean constraint of Equation (31) is that the rates no longer range over the entire positive real line. Instead, the constraint $0 \le r_i \le n$ results in a finite and bounded parameter space, implying that an infinite rate is impossible.

Objective Priors for Proportions

As we have already discussed, and as further elucidated by Kass and Wasserman, there are two related but different interpretations of what an objective prior is (1996). First is that objective priors are formal representations of ignorance. Second is that there is no objective, unique prior that represents ignorance, mainly because there is no unique notion of ignorance. Instead, a particular objective prior should be selected for use as a default by public agreement, much like units of weight or length. We use this default when there is either insufficient information or insurmountable difficulty exists to define the prior. We define three different objective priors on the simplex support for a set of relative proportions.

The Uniform Prior

The uniform prior has been a widely used for its conceptual and mathematical simplicity. The uniform prior is identical to the maximum-entropy prior if the support of the prior is finite. Jaynes (1968) argued that if p is the density of the prior, then the entropy

$$\mathcal{H}(p) = -\int p \cdot \ln p \, d\mu(p) \tag{32}$$

represents the amount of uncertainty implied by p with respect to the base density μ. For finite supports, μ is implicitly chosen to be Lebesgue (uniform) measure. By choosing a

density p that maximizes the entropy, we maximize the uncertainty and hence minimize the information present in the prior.

Although maximum entropy priors have been used successfully in a broad range of problems, it is not necessarily ideal for all circumstances. For instance, they are not invariant to reparameterization, and they are subject to numerous inferential paradoxes. For a review and critique of maximum-entropy priors, please refer to Seidenfeld (1987).

The Multinomial Prior

Invariance to reparameterization is a highly desirable property for a prior. Unfortunately, developing such a prior satisfying Equations (24) and (25) is difficult for a likelihood given by Equation (17). Consider that the estimated quantities θ denote a set of relative proportions. With the *ansatz* that when nothing else is known, any set of experiments that measure a set of relative proportions should have the same prior for them. Thus, all experiments should share the same prior. Now consider the following experiment: a large set of colored balls, with n different colors, is placed in an urn. Balls are sampled with replacement with the goal of inferring the relative proportions of colors. This experiment describes a classic inferential problem based on the multinomial distribution, and has been extensively studied. Associating the relative fraction of each color with the relative rate fraction θ_i, we draw an analogy between the inference of multinomial parameters and rate fraction parameters.

For inference about multinomial parameters, Jeffreys (1946; 1961) proposed using a reparameterization-invariant prior that is proportional to the square root of the determinant of the Fisher information matrix

$$P(d\theta) \propto \sqrt{\det\left[\mathcal{I}\left(\theta|A,T\right)\right]}\, d\theta\,, \tag{33}$$

where $\mathcal{I}\left(\theta|A,T\right)$ denotes the Fisher information matrix of the likelihood. Applied to a set of unknown proportions with multinomial likelihood, Jeffreys' procedure results in the prior

$$P(d\theta) \propto \frac{d\theta}{\sqrt{\theta_1 \cdot \theta_2 \cdot \cdots \cdot \theta_n}}\,. \tag{34}$$

This so-called the *Jeffreys prior* works well for univariate priors. In the multivariate case it is known to sometimes give results incommensurate with conventional statistical theory (Syversveen 1998). Therefore, the Jeffreys prior should not be accepted as the multivariate "prior of choice" without additional justification.

To ameliorate problems with the multivariate Jeffreys prior, the reference prior construction was created by Bernardo (1974), further refined by Berger and Bernardo (1989), and fully developed Bernardo and Smith (1994). This class of priors is identical to the Jeffreys prior in the univariate case, but is often quite advantageous to the Jeffreys prior in the multivariate case. A "reference prior" is a prior that maximizes the expected Kullback-Leibler divergence of the posterior distribution relative to the prior. This maximizes the expected posterior information about θ when the prior density is $P(d\theta)$. The case of inferring multinomial proportions has been analyzed in detail by Berger (1992) and Bernardo (1998). Their analyses yield results identical to the Jeffreys prior, a coincidence giving strong evidence for the correctness and consistency of the Jeffreys prior for the multinomial distribution.

Goyal (2005) showed that the Jeffreys prior can also be obtained from, and is logically equivalent to, an intuitively reasonable information-theoretical invariance principle in the limit as $n \to \infty$. This result is comforting because the greatest criticism of the ansatz equating balls in an urn with evolutionary rates is the fact that the analogy only holds if it is given that there are an infinite number of balls.

The Group-Invariant Prior

Group-invariant priors are constructed to be invariant to specific types of group operation, rather than arbitrary reparameterization. For instance, given a three-site alignment, the prior information we have about rates $\{r_1, r_2, r_3\}$ is identical the prior information we have of the permuted rates $\{r_3, r_1, r_2\}$, thus the prior for these rates should be invariant to permutation. Both the uniform and Jeffreys priors for proportions are invariant with respect to the permutation group of order n on the space of ordered parameter indices.

Another type of desirable group-invariance is based on a more subtle argument. Given three unknown rate proportions $\{\theta_1,\theta_2,\theta_3\}$, suppose we have three known positive constants $\{a,b,c\}$. Our ignorance of $\{\theta_1,\theta_2,\theta_3\}$, is equal to our ignorance of the scaled relative proportions $\{a\theta_1,b\theta_2,c\theta_3\}$, after normalization. The process of scaling and normalization defines a Lie group on $\mathbb{S}^n$, and this group is isomorphic to the standard Euclidian group of $\mathbb{R}^{n-1}$ under addition. Details, including proof of the wider result that $\mathbb{S}^n$ is isomorphic to the vector space $\mathbb{R}^{n-1}$, are in Egozcue et al. (2003).

The appropriate prior for Lie group invariance to hold is given by the Jacobian of the transformation $T:\mathbb{S}^n \mapsto \mathbb{R}^{n-1}$, given by

$$P(d\theta) \propto \frac{d\theta}{\theta_1 \cdot \theta_2 \cdot \cdots \cdot \theta_n}. \tag{35}$$

It formally encapsulates the idea that if $x>0$ is unknown and $y>0$ is known, our ignorance of x the same as our ignorance of $x \cdot y$. The fact that this prior is improper is an unfortunate but not insurmountable difficulty.

Relating the Priors

The most notable connection among the priors discussed is that, despite having diverse philosophical and logical underpinnings, all three priors are members of the same distribution family. Specifically,

$$\theta \sim \text{Dirichlet}(\delta,\delta,\ldots,\delta) \tag{36}$$

for some $0<\delta\le 1$, with the group-invariant prior of Equation (35) being a limiting case. It is possible to choose $\delta>1$, but we find no compelling reason to do so. A graphical comparison of the $\delta=\{0,\frac{1}{2},1\}$ cases, where $n=2$, is depicted in Figure 3. As δ decreases, increasingly more weight is given to the prior probability that θ is either near zero or near one, corresponding to cases of high or low conservation, respectively.

Although any positive value of δ is technically valid, as δ decreases the distribution becomes singular until $\delta \to 0$ and the distribution becomes improper. Although improper

priors are frequently used in Bayesian analysis, they cannot be used here since they will always result in an improper posterior. Worse, as the prior becomes more singular, numerical methods often ill-conditioned. Therefore δ should not be too small.

Computing the Posterior

Given a likelihood of the form of Equation (17) and a model of amino acid replacement, the posterior rate distribution can be sampled the Metropolis-Hastings (MH) Markov Chain Monte Carlo (MCMC) technique (Robert and Casella 2004). Amino acid transition probabilities were specified by the wag evolutionary Markov process of Whelan and Goldman (2001) such that the probability of transition from amino acid k to amino acid j is given by

$$P\left(j|k,rt\right)=\left[e^{rtQ}\right]_{j.k}, \tag{37}$$

where Q is the wag rate matrix. The prior probability of observing amino acid x, denoted by $P\left(x\right)$ in Equation (17), was taken as the wag default.

Sampling the Posterior

The MH-MCMC algorithm is a three-step process. Given an initial parameter estimate θ_{old}, a new estimate θ_{new} is drawn from density $P\left(\theta_{\text{new}}|\theta_{\text{old}}\right)$. The new value is then accepted if

$$u<\frac{P\left(\theta_{\text{old}}|\theta_{\text{new}}\right)\cdot P\left(\theta_{\text{new}}|A,T\right)}{P\left(\theta_{\text{new}}|\theta_{\text{old}}\right)\cdot P\left(\theta_{\text{old}}|A,T\right)}, \tag{38}$$

where u is drawn from a uniform $\left[0,1\right]$-distribution, and rejected otherwise. The procedure is iterated indefinitely, resulting in consecutive parameter values being samples from the posterior.

Correct specification of the proposal density $P\left(\theta_{\text{new}}|\theta_{\text{old}}\right)$ is important. Since the Dirichlet distribution is one of only well-known distributions supported on the simplex, it would seem natural to base the proposal density on it. Implementation experience has shown that doing so

leads to unacceptable MCMC convergence, likely due to the highly asymmetrical nature of the Dirichlet density when its parameters are very small or very large, resulting in a low acceptance rate. We were unable to find any variation of Dirichlet-based proposal densities that had displayed acceptable convergence properties.

Instead, a variation of a simplex point-picking algorithm was used to indirectly derive a proposal density such that repeated sampling from the proposal density was equivalent to performing a uniform random walk over the simplex. Traditionally, simplex point-picking is accomplished via well-known properties connecting the Dirichlet and Gamma distributions. Specifically, if θ is uniformly distributed over $\mathbb{S}^n$, then $\theta \sim \text{Dirichlet}(1,1,\ldots,1)$ and draws of θ can be made by summing and normalizing n unit-exponential deviates (Devroye 1986). For our purposes, an alternative point picking algorithm based on order statistics (Balakrishnan and Cohen 1991), as proposed by Kraemer (1999) and refined by Smith and Tromble (2004), is more constructive. They recommend the following steps to generate θ:

Set $x_0 \leftarrow 0$ and $x_n \leftarrow 1$.

Generate $n-1$ uniform random values from the open interval $x_i \in (0,1)$.

Sort the set of points $\{x_0, x_1, \ldots, x_n\}$ into increasing order.

The n final coordinates $\{\theta_1, \theta_2, \ldots, \theta_n\}$ are given by $\theta_i \leftarrow x_i - x_{i-1}$.

If any $\theta_i = 0$, rerun the algorithm to generate a new set of θ.

Viewing the $n-1$ internal coordinates x_i as particles within a unit-interval "box" yields an immediate procedure for perturbing θ in a manner guaranteed to uniformly sample the simplex. Given an initial $n-1$ points x_i, as above, and a maximum displacement $0 < \varepsilon_{\max} \ll 1$, iterate the following:

1. Draw a set of random $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{n-1}\}$, such that $-\varepsilon_{\max} < \varepsilon_i \le \varepsilon_{\max}$ uniformly.
2. Set $x_i \leftarrow x_i + \varepsilon_i$ for each $i \in \{1,2,\ldots,n-1\}$.

3. For any $x_i < 0$, "reflect" the point back into the unit interval via $x_i \leftarrow |x_i|$.

4. For any $x_i > 1$, "reflect" the point back into the unit interval via $x_i \leftarrow 2 - x_i$.

5. Sort the set of points $\{x_0, x_1, \ldots, x_n\}$ into increasing order.

6. The n final coordinates $\{\theta_1, \theta_2, \ldots, \theta_n\}$ are given by $\theta_i \leftarrow x_i - x_{i-1}$.

7. If any $\theta_i = 0$, rerun the algorithm using the original points x_i.

This procedure is based on the formal equivalence between points uniformly distributed over $\mathbb{S}^n$ and a diffusion process of $n-1$ particles in a one-dimensional support with perfectly reflecting boundary conditions (Feller 1967; Strang 1986). This "diffusion" process is reversible and symmetric such that $P(\theta_{\text{new}}|\theta_{\text{old}}) = P(\theta_{\text{old}}|\theta_{\text{new}})$, simplifying numerical implementation.

The optimal choice of $\varepsilon_{\max}$ varies with simplex dimension (number of alignment sites) n. To compare step sizes for different data sets, a geometric argument (and practical experience) suggests that ε_i be divided by $\sqrt{n}$ to yield commensurate step sizes.

Sampling the Mean Rate

If we assume that the mean rate has sampling variability and is not exactly one, we set

$$\sum_{i=1}^{n} \theta_i = \mu \approx 1, \tag{39}$$

and a prior for μ is required. Under the mild assumption that each rate has a unit-Exponential distribution, a reasonable prior for μ is the unit-mean Gamma distribution

$$P(d\mu) = \frac{n^n}{\Gamma(n)} \cdot \mu^{(n-1)} e^{(-n\mu)} d\mu, \tag{40}$$

with n being the alignment length. Representative densities are shown in Figure 4. The distribution variance decreases in proportion to $\sqrt{n}$, implying that there is no significant change in shape for alignments of broadly similar length.

Results

Two related protein families were selected for study: the Myc transcription factor and its heterodimer binding partner Max. Site specific evolutionary rates were inferred for each protein family. Before discussing specific results, however, we make note of some general observations regarding the performance of our rate inferring algorithm.

Convergence of the MCMC algorithm to the posterior was relatively rapid for short $(\approx 100 \text{ site})$ alignments and appeared to be insensitive to starting values of the relative rate proportions. Therefore initial values of $\theta_i = 1/n$ were used for all studies, corresponding to the null hypothesis of equal rates across all sites (rate homogeneity). The posterior rate distribution did not noticeably deviate from unimodality in any example studied.

Changing the δ hyperparameter had little influence on the posterior, with smaller values of δ accentuating extreme values of θ_i. The magnitude of this effect was not significant when taken in context of the rates' variance, unless δ was small enough to induce convergence difficulties. After preliminary experimentation, a median value of $\delta = 1/2$, corresponding to the multinomial prior, was selected to act as a default. No convergence difficulties were observed at this value. Trial rate computations incorporating Equations (39) and (40) indicated that even for short alignments, the effects of modeling mean rate fluctuation was to slightly inflate the variance of estimated rates, and appeared to be essentially negligible.

We note that sampling from the posterior can be performed quickly. For constant branch lengths, likelihoods can be pre-computed at each site i for a given range of rates-times-time. Rather than re-computing the likelihood at every change of r_i, a cached or interpolated value can be used to speed up overall calculation. When branch lengths are not constant, intermediate caching of constant branch-specific values can be employed to further reduce computational effort.

The Myc Motif

The Myc-Max-Mad transcription network of bHLH proteins is essential for control of cell growth, proliferation, differentiation, and apoptosis. *Myc* is a well-established oncogene whose deregulated expression is responsible for a wide range of human cancers. Approximately 70,000 cancer deaths in the United States each year arise from misregulation of *Myc* (Grandori, Cowley et al. 2000; Luscher 2001; Nasi, Ciarapica et al. 2001; Zhou and Hurlin 2001; Levens 2003). A comprehensive analysis of conservation in the bHLH-leucine-zipper (bHLHz) domain in a diverse set of Myc homologs performed by Atchley and Fernandes (2005) used site-specific entropy profiles to quantitatively score conservation and use that score to generate a Myc-family predictive motif. Specifically, the Boltzmann-Shannon entropy $\mathcal{H}_i$ for alignment site i was calculated as

$$\mathcal{H}_i = \mathcal{H}(A_i) = -\sum_{j=1}^{20} p_{i,j} \log_{20}\left(p_{i,j}\right), \tag{41}$$

where $p_{i,j}$ is the probability of having amino acid j at site i and $0 \le \mathcal{H}_i \le 1$. Smaller values of $\mathcal{H}_i$ indicate lower variability and hence greater evolutionary conservation at that site. The method was remarkably successful despite of a lack of explicit correction for phylogenetic or amino acid similarity correlation effects, and seemed to give results that are were superior than other conservation scores. A large number of these conservation scores were reviewed by Valdar (2002). The resultant entropy profile for the bHLHz region of Myc is shown in Figure 5. The alignment used herein had the hypervariable loop region removed, resulting in an alignment that was almost gap-free and where most residues display some degree of conservation. The data consists of 45 sequences with an aligned length of 79 sites. The alignment and phylogenetic tree, well as a graphic depiction of the alignment, are available as online as Supplementary Material. The alignments are included as Supplements A and B.

For comparison, Figure 6 displays a box-plot showing the rates (and their approximate marginal distributions) as inferred using our method, where rates are forced to always have a mean of one. Qualitatively, Figures 5 and 6 display similar relative features. The entropic conservation scores of Figure 5 are highly correlated $(r \approx 0.77)$ with the evolutionary rates

of Figure 6. Quantitatively, differences are seen mostly in the relative magnitudes of corresponding sites in each figure. The entropy of each site tends to be either quite high or quite low, especially in relative comparison with its corresponding inferred rate. Note that Figures 5 and 6 are significantly less similar to the conservation, quality, and consensus scores (their inverses) as computed by the jalview multiple alignment viewer (Clamp, Cuff et al. 2004) and shown in the Figure Supplements A and B. These examples highlight the importance of choosing an appropriate measure of evolutionary change when scoring relative conservation.

More importantly, examining the differences between Figures 5 and 6 highlight the importance of taking phylogeny and amino acid exchangeability into account. For example, site 3 consists of about $2/3$ lysine and $1/3$ arginine, both of which are basic and hydrophilic amino acids. The evolutionary exchangeability of these amino acids, coupled with the phylogenetic relationship of their sequences, causes the entropy method to significantly underestimate the amount of conservation at that site. Another example can be seen in site 6, featuring a predominance of threonine and asparagine. The entropic profile here indicates very little conservation. However, after accounting for phylogeny, the evolutionary rate of site 6 appears likely to be lower than average, and hence somewhat conserved. Further inspection reveals numerous other sites where phylogeny or exchangeability are large factors in site conservation.

Comparing our results with those from rate4site (Mayrose, Graur et al. 2004) is instructive as both methods infer rates with a process that incorporates, in identical ways, both phylogenetic and amino acid mutation information. The only difference between them is the computation of and choice of prior distribution for the rates. Figure 7 depicts the inferred rates and associated variances of the EB rate4site method using an optimized Gamma shape parameter of $\alpha = 1.35$, and 25-75% Confidence Intervals (CI) about the mean estimated rate. The tree topology and branch lengths were held constant and were identical for both algorithms.

Both Figures 6 and 7 are qualitatively and quantitatively similar, indicating that both methods yield consistent and similar results. Quantitative differences can be seen, however,

at the extremities of the rates; rate4site tends to shift both low and high rates towards a rate of one, as compared with Figure 6. The origin of this rate shift can be understood as an artifact of the Gamma rate prior with shape parameter $\alpha = 1.35$. With this shape, rates near to zero must be shifted upward in accordance with the prior. Furthermore, the roughly exponential-like tail of the Gamma prior tends to downshift rates of variable sites, thus underestimating them. These underestimations are of particular importance in determining the variability of the estimated rate, as seen in Figure 6, alignment sites 1, 9, 24, and 31 in particular.

Figure 8 depicts a frequency histogram of our Myc rate means versus a plot of two unit-mean Gamma distribution, one with shape parameter $\alpha = 1.35$ representing the optimal prior found by rate4site, and one with $\alpha = 1.54$ representing the Gamma distribution's best fit to the histogram data. A quartile plot is also shown to assess the fit of the observed distributions to the theoretical. Although the exponential-like decay of larger rates appears well modeled, rates of the order of 0 to 1.5 appear to be poorly represented. Accordingly, the Cramer-von Mises and Anderson-Darling tests both indicate that there is marginal evidence $(p < 0.11)$ that the best-fit Gamma distribution did not adequately model the data. We consider this p-value marginal since it referrers to the non-optimal Gamma shape. Comparing the two rate distributions must be done carefully, however, as rate4site assumes independent sites, and hence the joint probability of all the rates is equal to the product of the marginal rate distributions. With our method, however, rates are not independent and the joint rate distribution is *not* equal to the product of the marginal distributions. Figure 8 effectively compares only the means of the marginal rate distributions, and may be misleading as to how the rate distributions actually compare.

Lastly, another significant difference between the two methods is that the rate4site confidence intervals of Figure 7 appear to be roughly discretized, possibly an artifact of the discretized Gamma approximation that the algorithm uses. It is not clear whether the confidence intervals as displayed are accurate, or if they have been disproportionately influenced by the discretization.

The Max Protein

Although inter-quartile dispersion of the estimated rates are roughly equal between Figures 6 and 7, stability of our algorithm was important to investigate when an alignment consists of a large number of non-conserved sites. Also important was our algorithm's ability to discriminate conserved from non-conserved sites. Thus, we computed the posterior distribution of rates for Max, a binding partner of Myc.

Protein-protein interactions between Myc and Max are an essential element in proper functioning of the Myc-Max-Mad transcription factor network. Mad-Max heterodimers repress the expression of Myc and initiate differentiation. Although capable of weak homodimerization, proper Myc function requires heterodimerization with Max (Grandori, Cowley et al. 2000; Zhou and Hurlin 2001; Nair and Burley 2003). We aligned entire protein sequences of 23 Max homologs using the local alignment method of DIALIGN-T (Subramanian, Weyer-Menkhoff et al. 2005), resulting in a total alignment length of 380. Local alignment methods tend to produce better alignments than global alignment methods when sequences are substantially divergent, as is the case with Max. As a result, Max alignments are heavily gapped. This is significant since likelihood calculations treat gaps as the logical equivalent of "any amino acid" (Felsenstein 2004), directly affecting the inferred rates.

A box-plot showing all 380 inferred rates for Max is shown in Figure 9, with a detail enlargement in Figure 10. Despite the use of relatively few alignment sequences, the estimated rates have tight confidence intervals, as confirmed via multiple long-run MCMC samplings. Conserved sites can be discerned by inspection: sites with rate less than one are conserved, with the degree of conservation given directly by the rate. Anomalies are also notable with respect to gaps and autapomorphies.

Sites 123-127 show an insertion autapomorphy common to *Cyprinus* (carps) and *Danio* (zebrafish), resulting in 20 of 23 sequences being gapped. There is also an arguable misalignment of two valine residues within the gapped sequences. Nontheless, rates inferred for these sites tend to be very low, indicating putative conservation. This implied

conservation is correct, but only within the relevant species group; normally one would not consider a set of gaps to be "conserved".

An interesting autapomorphy depicted at site 102 consists of 22 sequences with aspartic acid (acidic and hydrophilic), while one sequence contains alanine (hydrophobic). Again, though arguably a misalignment, this single non-conservative substitution causes the inferred rate to be roughly 3.5, one of the largest rates present.

Figure 11 depicts a frequency histogram of our Max rate means versus a plot of two unit-mean Gamma distribution, one with shape parameter $\alpha = 0.935$ representing the optimal prior found by rate4site, and one with $\alpha = 1.17$ representing the Gamma distribution's best fit to the histogram data. A quartile plot is also shown to assess the fit of the observed distributions to the theoretical. Due to the large number of rates rate samples present, the deviation of the observed distribution to the best-fit Gamma is notably higher. The quartile plot indicates systematic and widespread deviation from the Gamma distribution, especially at lower rates. This deviation is evidenced by the Cramer-von Mises and Anderson-Darling tests both indicating strong evidence $(p < 0.001)$ that the best-fit Gamma distribution does not adequately model the data. Again, we warn about the difficulty comparing the two distributions, as per the case with Myc.

Discussion

In this study we showed how site-specific rates could be inferred for a given multiple alignment and phylogeny, and from those rates, conserved resides detected. Unlike previous work involving site-rate heterogeneity, our work does not assume any particular form for the prior distribution of rates. Instead, by estimating n relative rates through $n-1$ parameters, our effective rate prior is a parameterless distribution that can be thought of as a superset containing all rate distributions of unit mean. Such a rate prior is arguably more general than other types of parameterized families, such as unit-mean Gamma or unit-mean Gamma mixtures. Thus our method estimates site-specific rates using a prior that has the least-possible prior information, and hence incorporates the least-possible bias into the results.

Numerical performance of any rate inference method is of crucial importance. Our method appears to have robust convergence characteristics and for short $(\approx 100 \text{ site})$ alignments generally attains rapid convergence. Longer $(\approx 1000 \text{ site})$ alignments can take prohibitively long to converge, however, and we plan future work on enhancing the MCMC transition kernel to improve convergence and sampling efficiency.

An important observation was that variances of the inferred rates did not increase with the number of alignment sites considered. This observation seems to imply that inferring one parameter per site does not automatically inflate variance estimates, as is the case when models are over-parameterized. In fact, our observations seem to indicate that the widths of rate confidence intervals are far more dependent on the number of sequences and their phylogenetic relationship than the number of alignment sites. For Myc, Max, and other examples tested during development, convergence was generally rapid and insensitive to starting values of θ. It would appear then that the posterior rate distribution is essentially unimodal and fairly sharp, implying that MH-MCMC convergence should rarely be problematic.

In terms of numerical implementation, our method represents a significant departure from traditional likelihood calculations that incorporate rate heterogeneity. Currently, likelihood calculations are generally computed along the lines of

$$P(A,T) = \prod_{i=1}^{n} \int_{0}^{\infty} P(A_i, T | r_i)\, dP(r_i), \tag{42}$$

where $P(r_i)$ is the rate prior (generally the Gamma distribution) and integral is approximated by numerical quadrature. To actually calculate $P(A|T)$ requires iterating over two nested loops: the inner loop computing the integral and the outer loop counting sites. In contrast, our method entails picking a set of $n-1$ relative rates r and computing $P(A|T,r)$. The posterior density is sampled via the MCMC algorithm; no *a priori* selection of number and magnitude of "rate categories" are required. These two approaches are quite different from an implementation point of view, and therefore it may not be entirely straightforward to

implement our scheme within existing phylogenetics programs such as MrBayes (Ronquist and Huelsenbeck 2003) or beast (Drummond and Rambaut 2003).

One of the more interesting and unexpected capabilities of our algorithm has been the detection of hypervariable sites – sites whose inferred rate of evolution is much greater than one. Such cases may indicate a significant evolutionary fact about a site, or, as in the case of Max, a potential misalignment. In particular, by not assuming *a priori* that large rates are roughly exponentially less likely than small rates, our method may be suitable for detecting both conserved sites and sites evolving significantly *more* rapidly than others, and may prove a valuable avenue of future research.

Finally, an honest elucidation of our method begs the following question. In general, the only observed data we are given is the set of sequences. Their alignment is inferred, usually via multiple-alignment algorithm, and the phylogenetic tree is then further inferred from the alignment. Our method currently assumes that the tree is known and fixed. However, rate heterogeneity is important to consider when deriving evolutionary relationships and divergence times. Thus we are left with the dichotomy of requiring rate heterogeneity to infer the phylogeny, and requiring the phylogeny to infer rate heterogeneity.

For the time being, it is fortunate that the site-independent Gamma distribution often provides a reasonable starting point because that is how we currently estimate the phylogeny. Of particular interest to our group is the construction of a reasonable and objective prior for branch lengths and tree topology, thus allowing fully Bayesian inference of rates, times, and topology. Such a treatment would provide a comprehensive and consistent framework for molecular phylogenetics, and is actively being pursued.

Acknowledgements

The authors wish to thank Jeff Thorne, Charlie Smith, Steffen Heber, Kevin Scott, Spencer Muse, Eric Stone, Benjamin Redelings, and Bonnie Deroo for helpful comments and suggestions during manuscript preparations. Data analysis was done in large part with the R system (2006). Financial support was provided by the National Institutes of Health (GM45344), and the North Carolina State University.

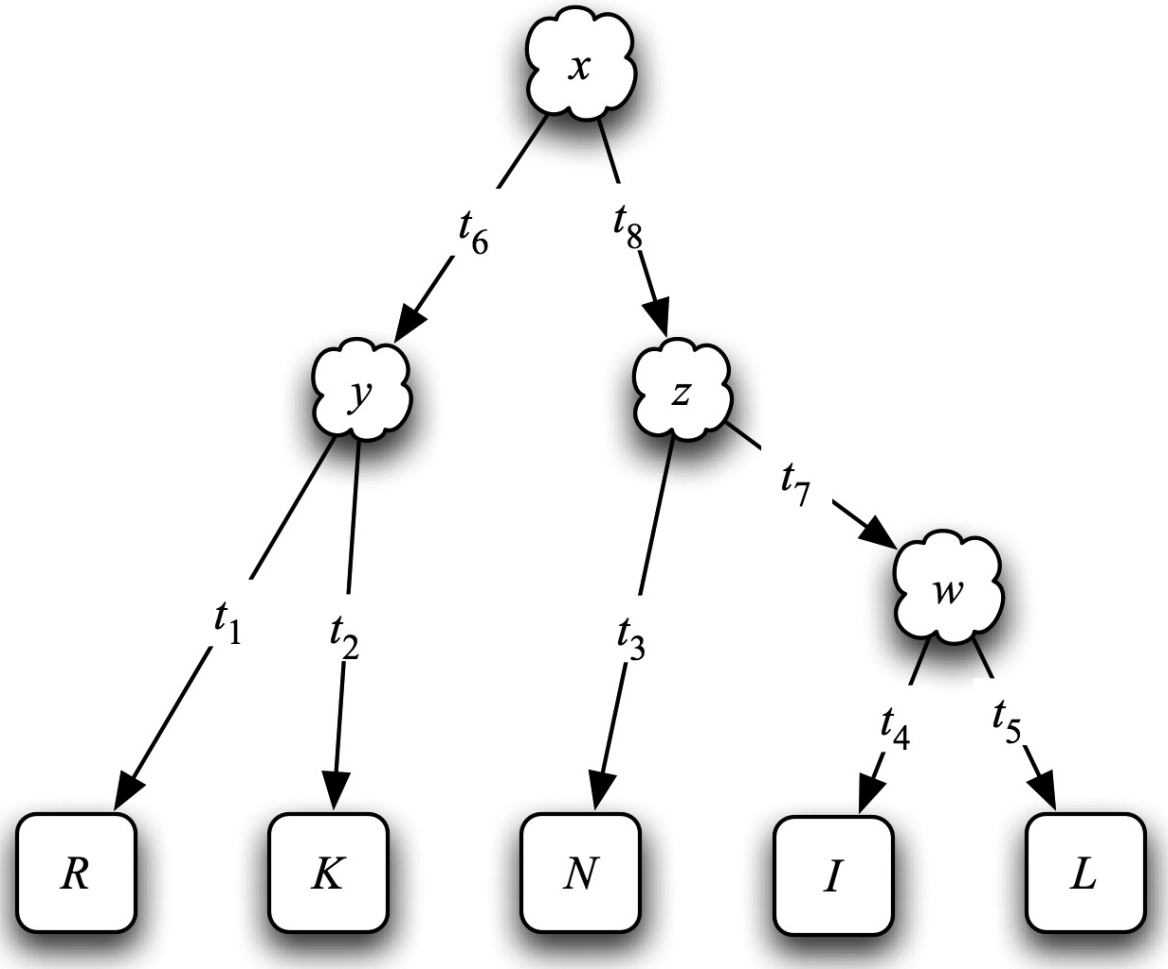

Figure 1

A sample phylogenetic tree showing individual branch lengths and alignment data for a single site. Observed data are shown in boxes while inferred ancestral states are shown in clouds.

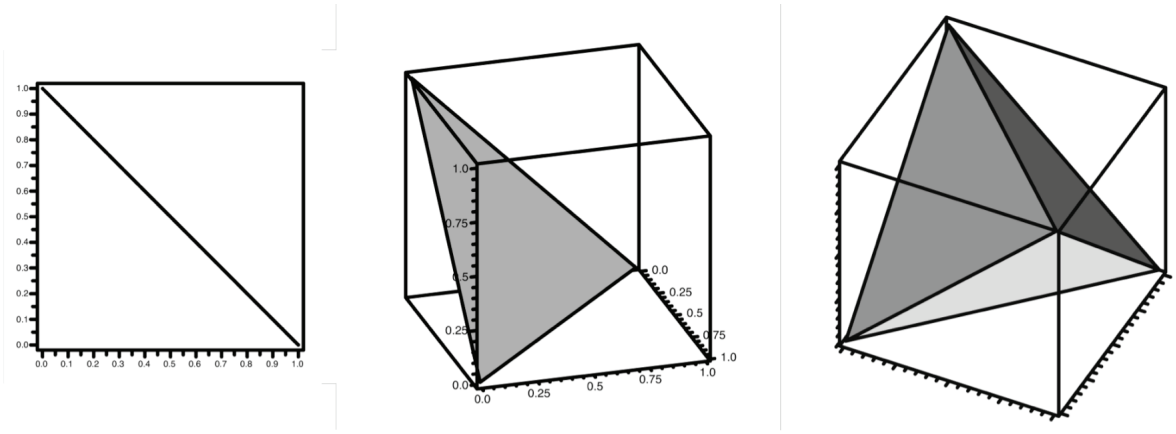

Figure 2

The first three standard simplexes with enclosing coordinate axes. From left to right: $\mathbb{S}^2$ (a line segment), $\mathbb{S}^3$ (an equilateral triangle), and $\mathbb{S}^4$ (a solid, regular tetrahedron). For clarity, the fourth dimension of $\mathbb{S}^4$ is not shown.

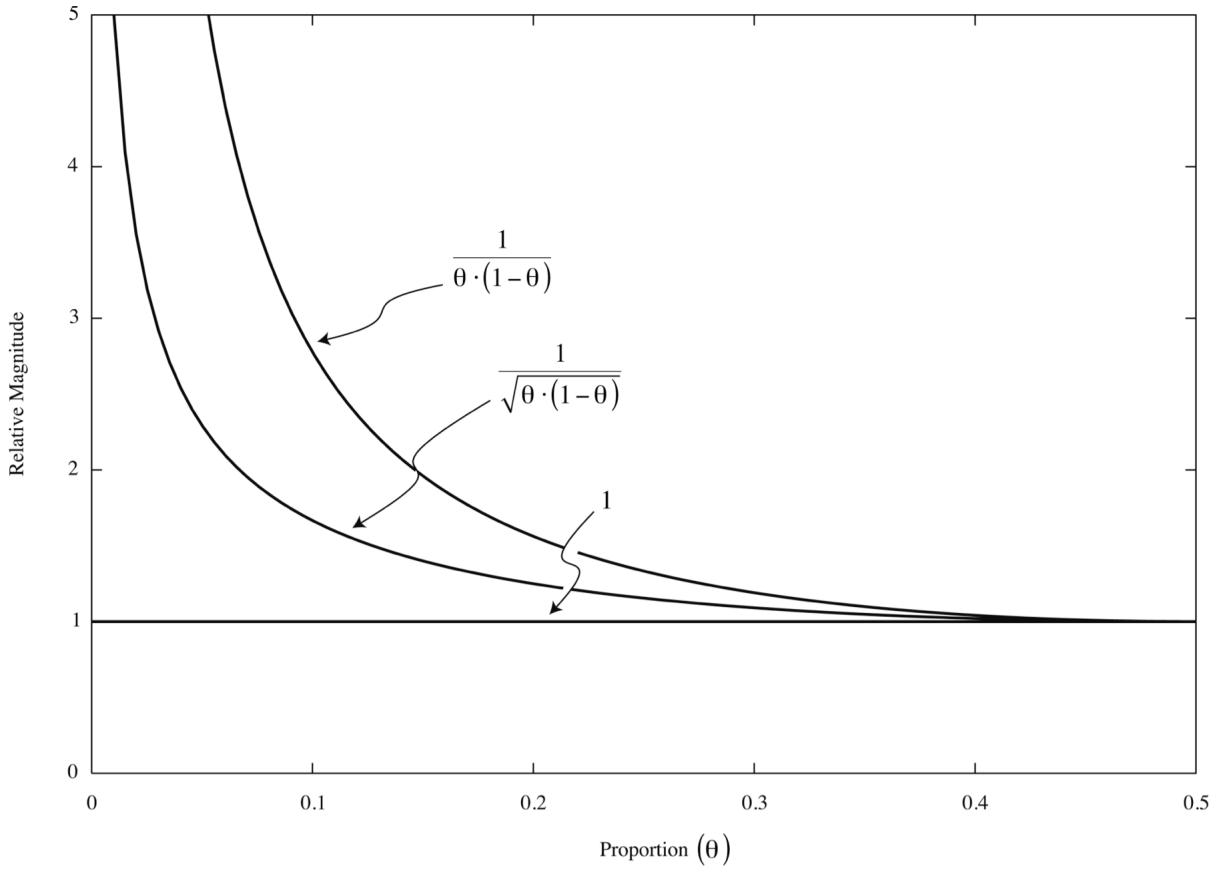

Figure 3

A graphic comparison showing the relative magnitudes of the three prior distributions discussed for relative proportions, assuming $n = 2$. The functions are symmetric around $\theta = 0.5$.

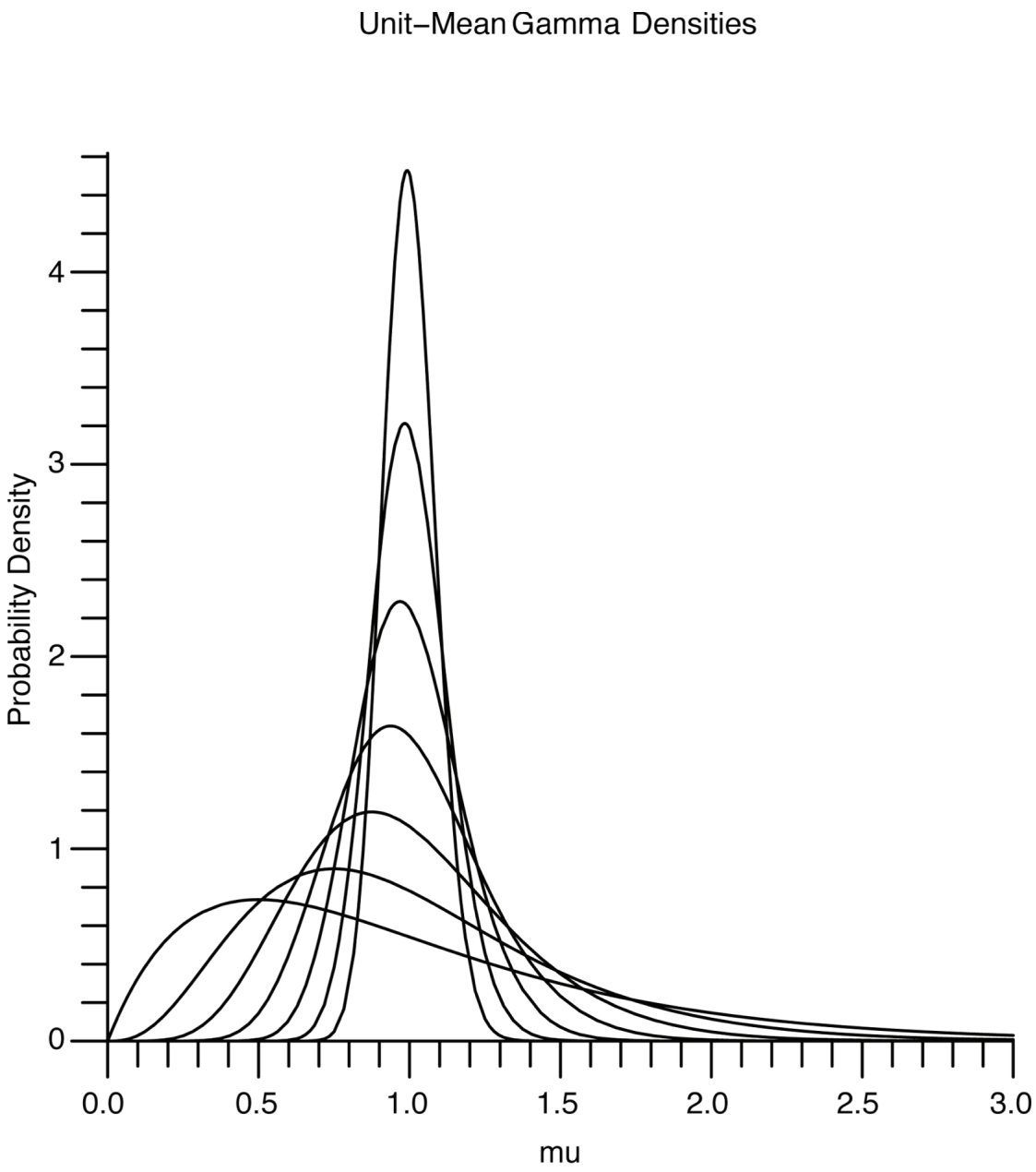

Figure 4

Unit-mean gamma distribution densities as per Equation (40) for $n = \{2, 4, 8, 16, 32, 64, 128\}$. Larger values of n result in distributions with progressively sharper peaks.

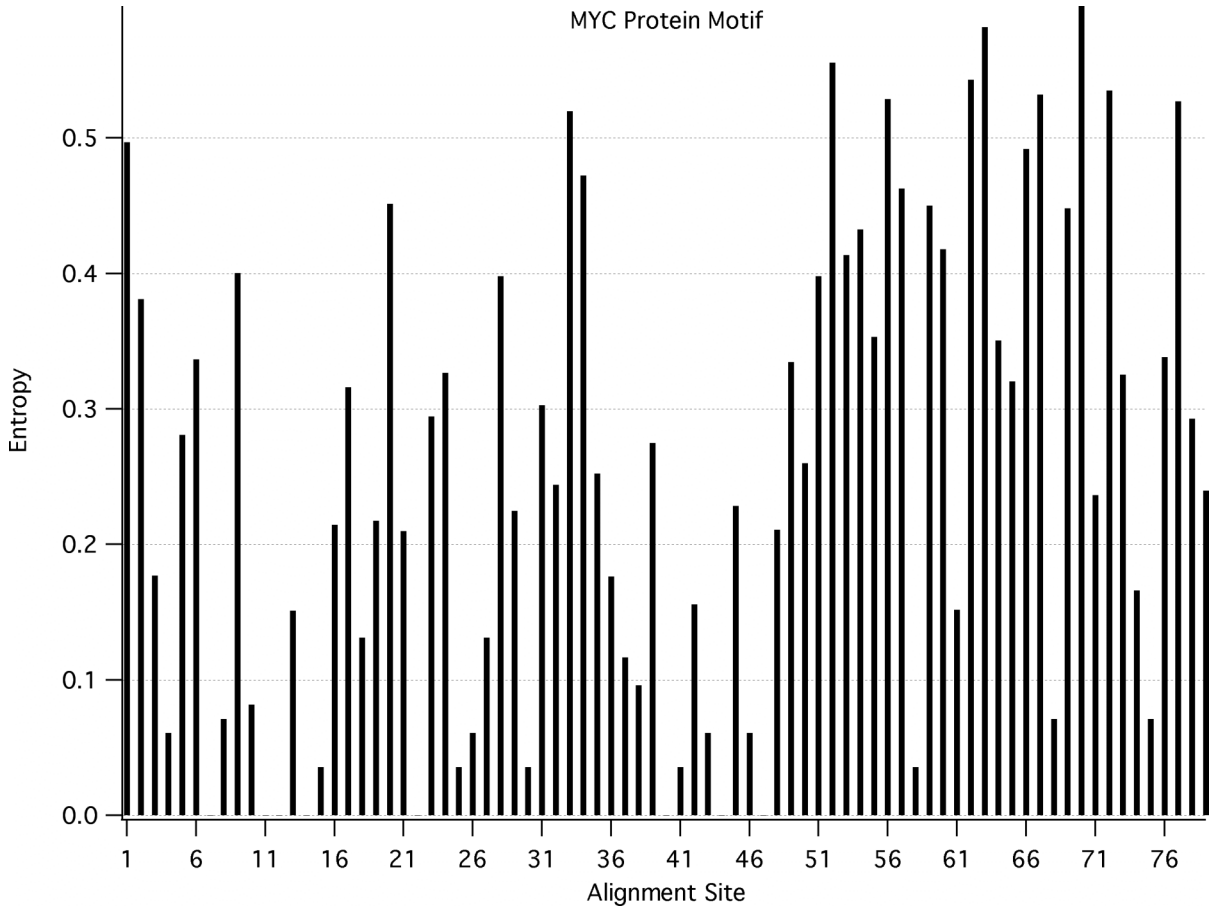

Figure 5

Entropy profile for the conserved residues of Myc, as per Atchley and Fernandes (2005). The alignment sites have been renumbered for consistency with other figures.

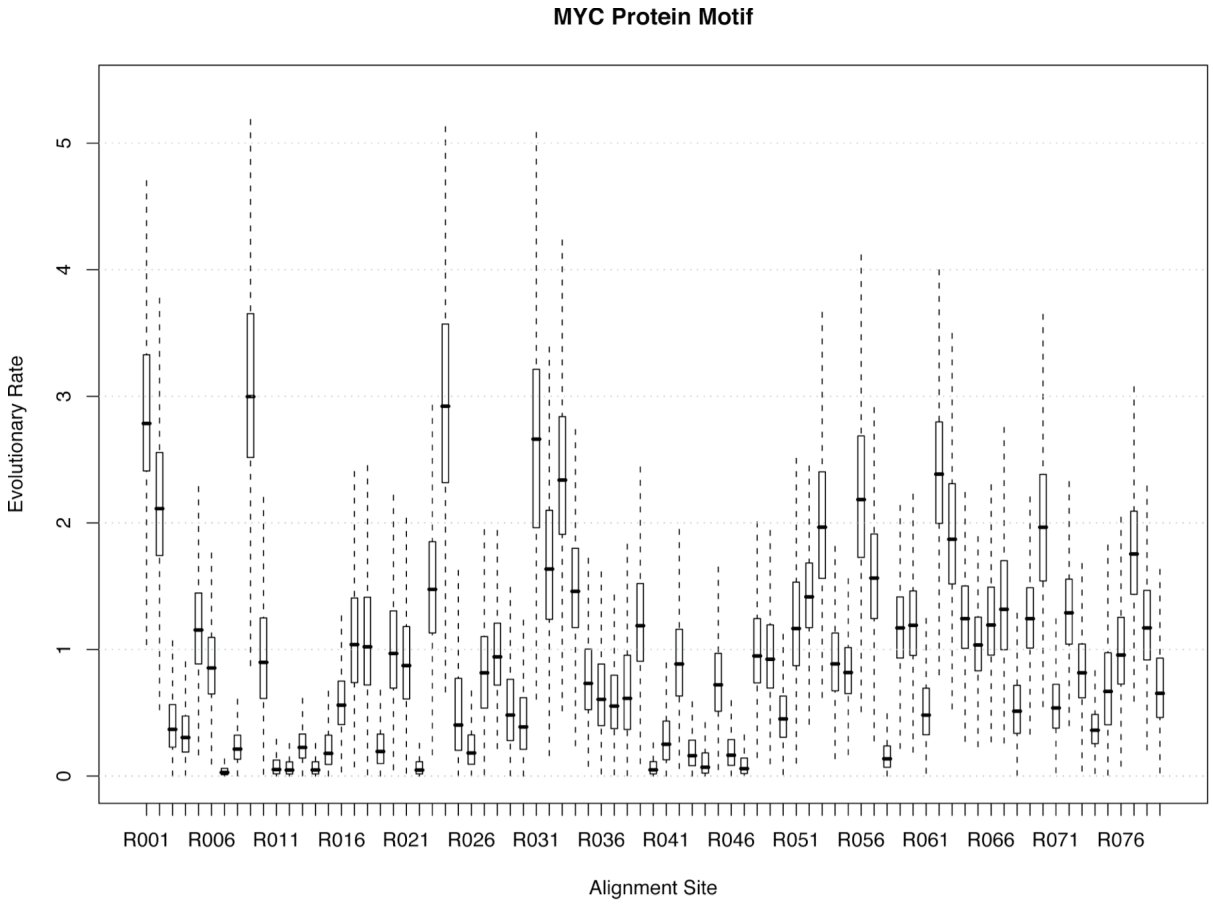

Figure 6

Distribution of rates for the conserved residues of Myc, as given by the methods presented herein.

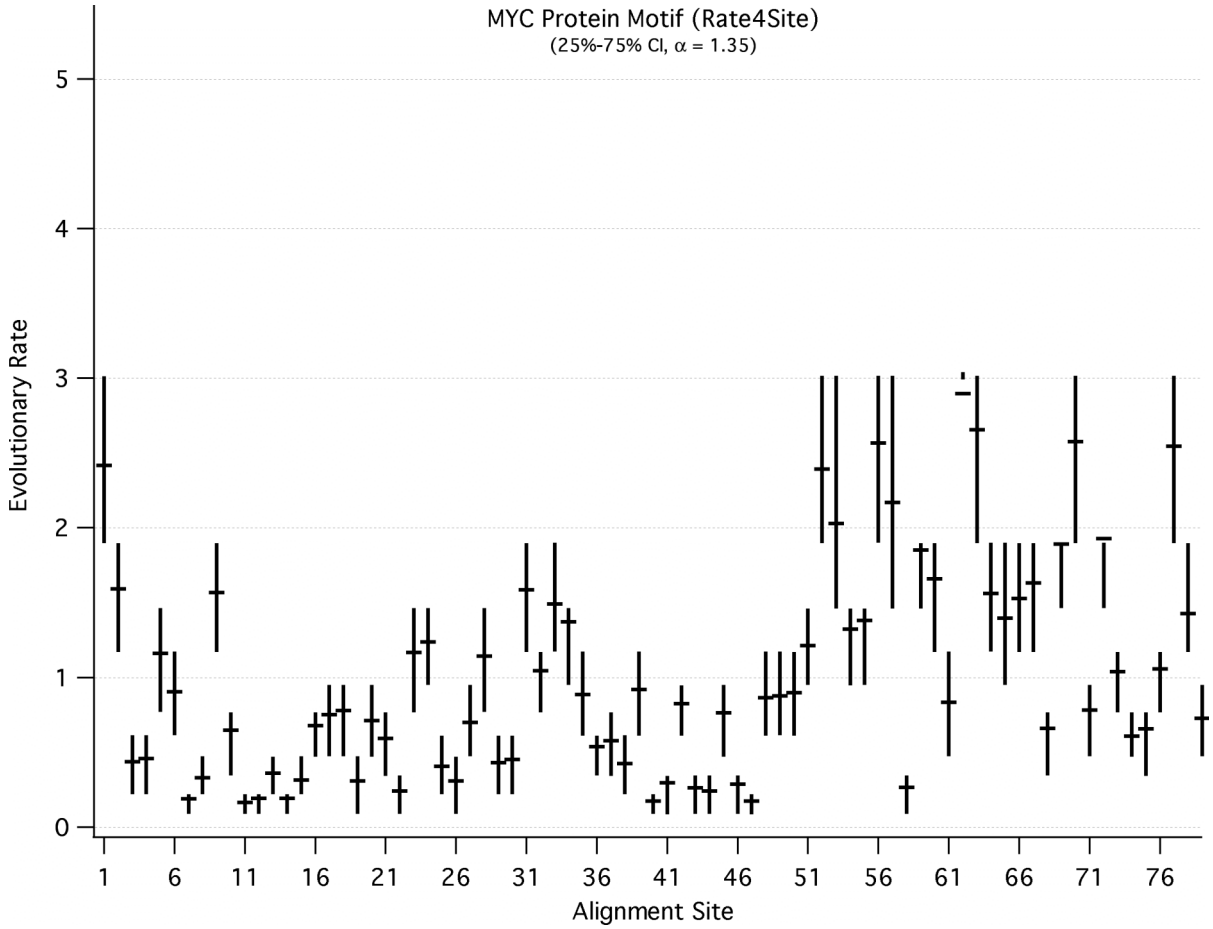

Figure 7

Distribution of rates for the conserved residues of Myc, as given by the EB rate4site method.

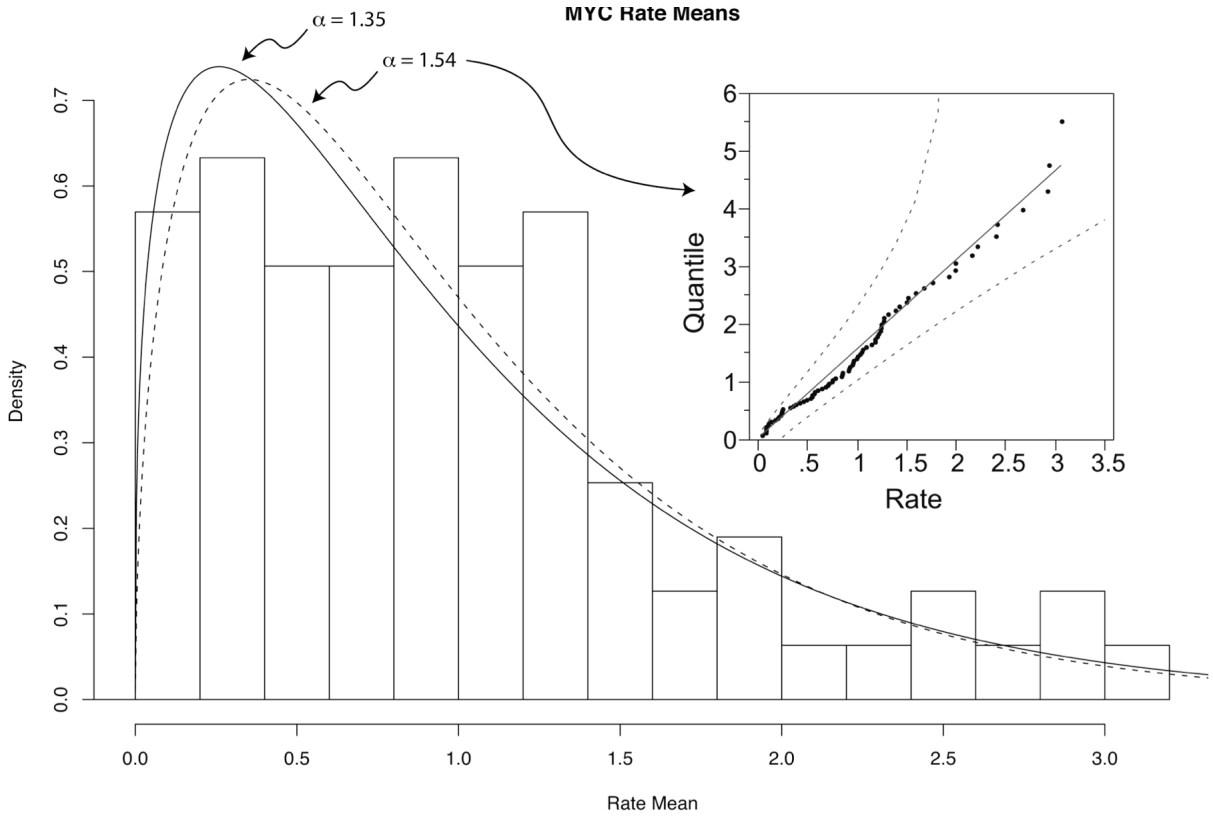

Figure 8

The distribution of rate means for Myc as calculated by our method (histogram) overlaid with two unit-mean Gamma densities, based on 79 sites: the one used by rate4site (solid, optimized shape parameter $\alpha = 1.35$), and the best Gamma density fit to our data (dashed, shape parameter $\alpha = 1.54$). The quantile plot shows the deviation of the histogram data from the best-fit Gamma density, with marginal evidence for lack-of-fit $(p < 0.11)$.

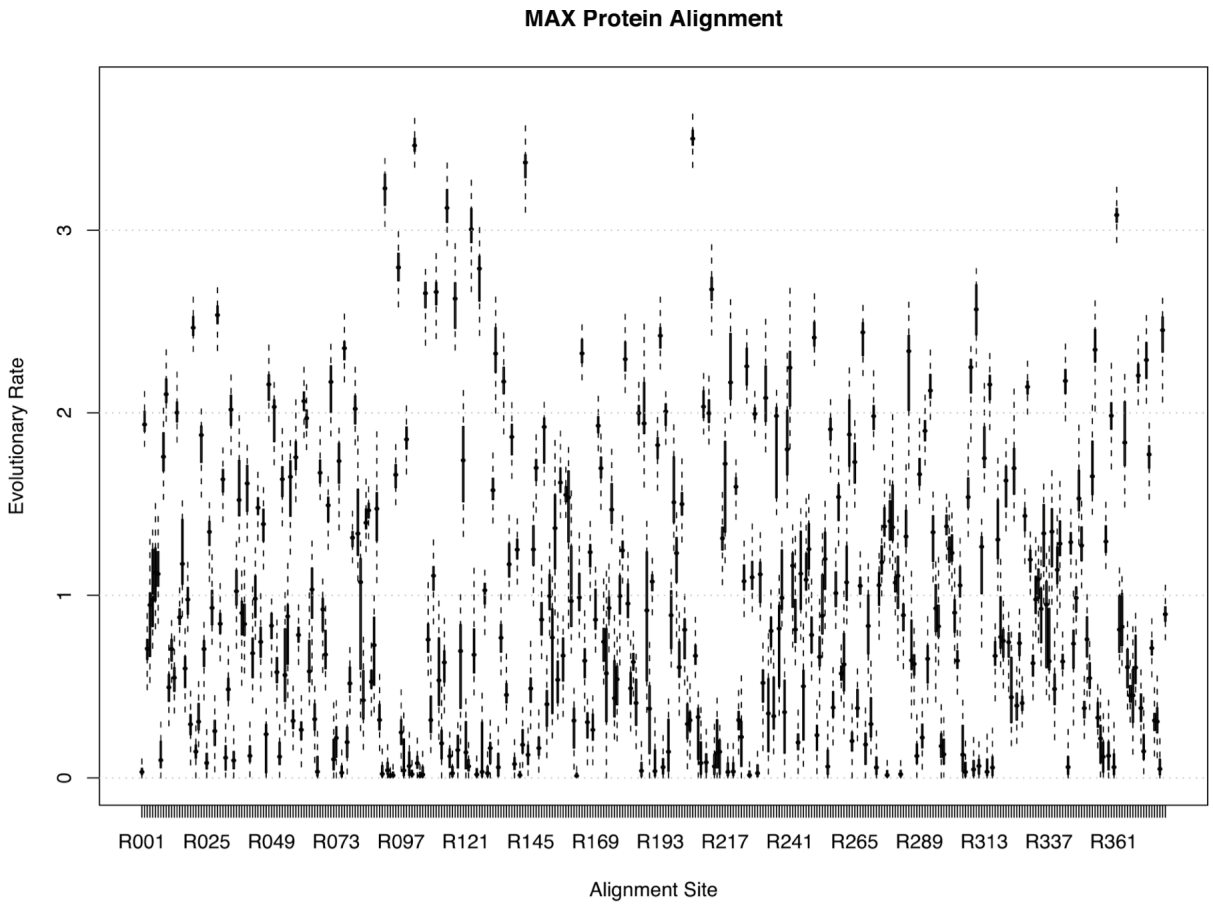

Figure 9

Distribution of rates for the conserved residues of Max, as given by the methods presented herein.

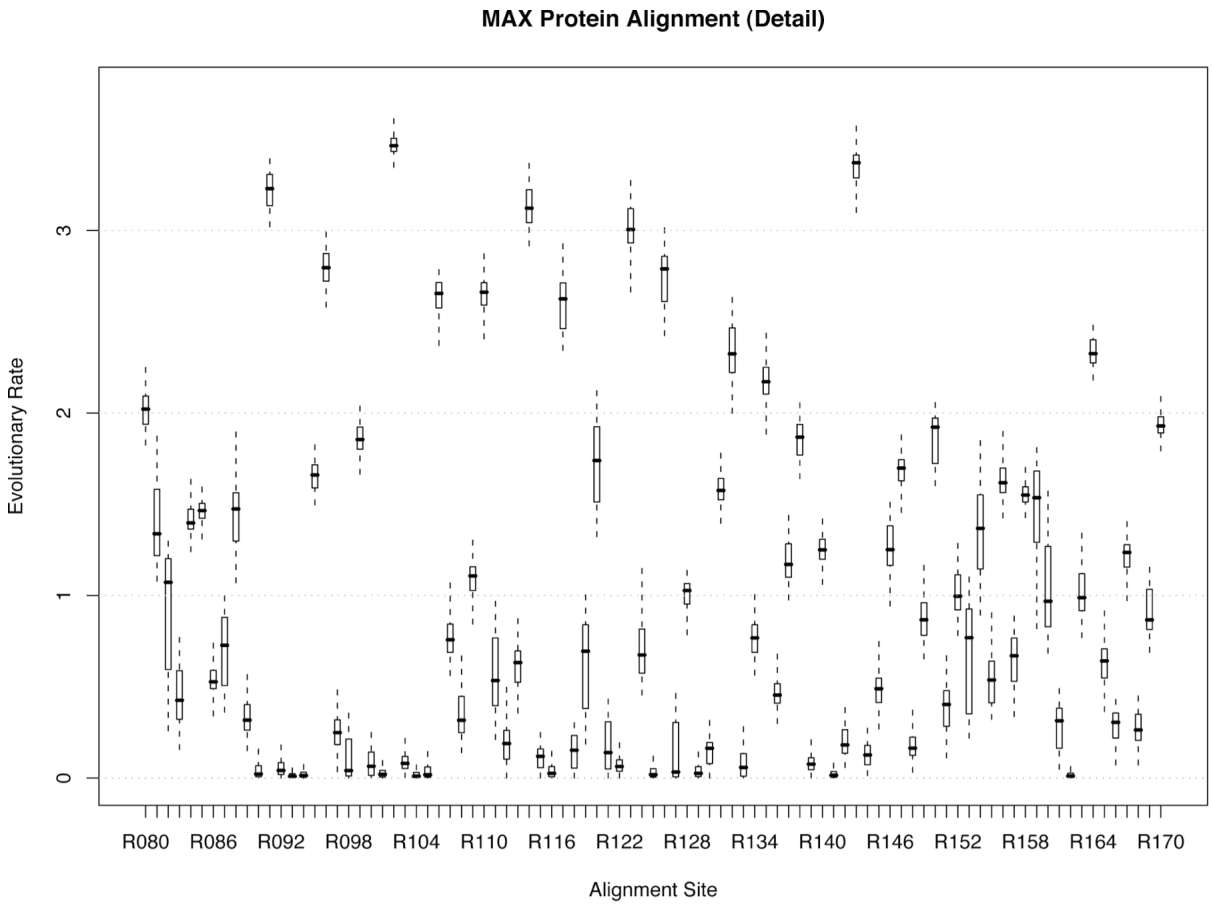

Figure 10

Detail region of Figure 9.

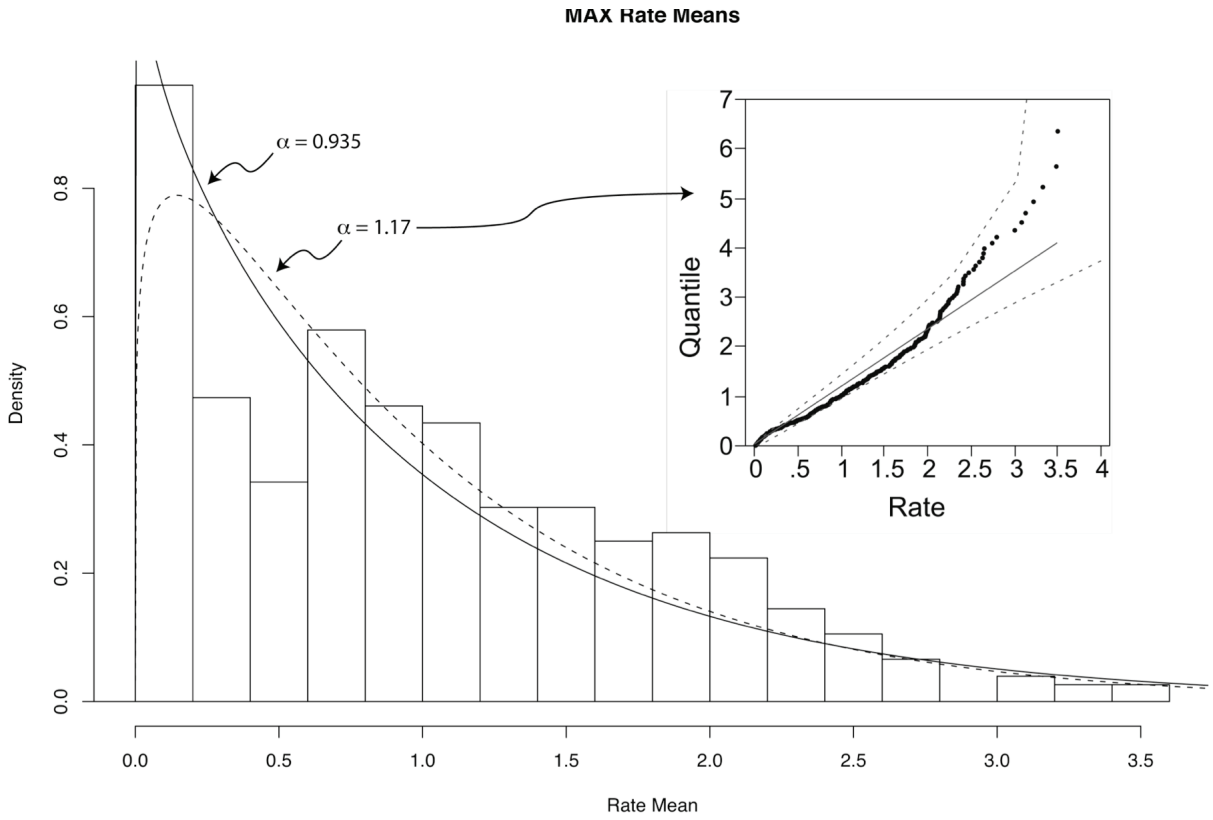

Figure 11

The distribution of rate means for Max as calculated by our method (histogram) overlaid with two unit-mean Gamma densities, based on 380 sites: the one used by rate4site (solid, optimized shape parameter $\alpha = 0.935$) and the best Gamma density fit to our data (dashed, shape parameter $\alpha = 1.17$). The quantile plot shows the deviation of the histogram data from the best-fit Gamma density, with strong evidence for lack-of-fit $(p < 0.001)$.

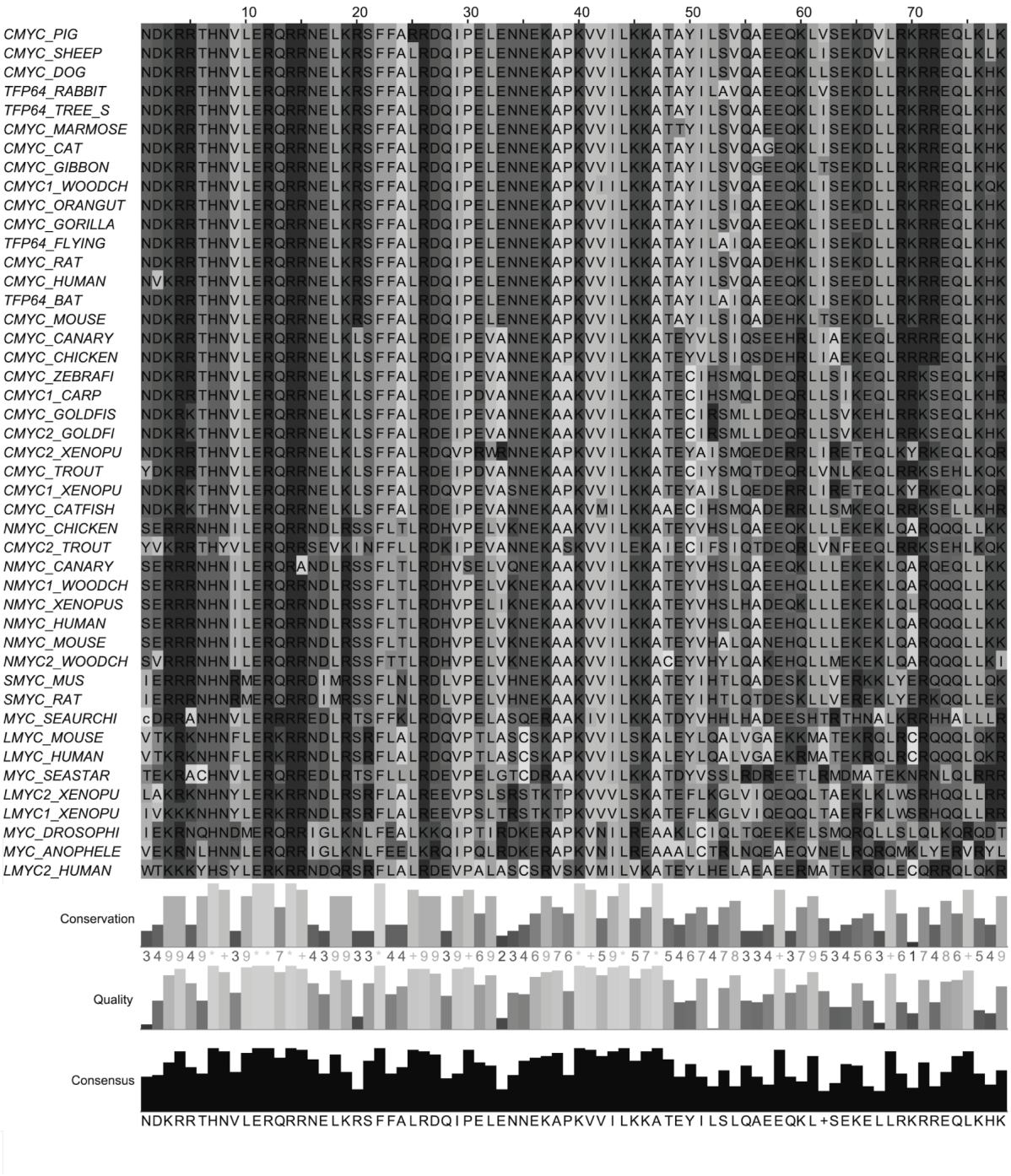

Figure Supplement A

The aligned Myc bHLHz region.

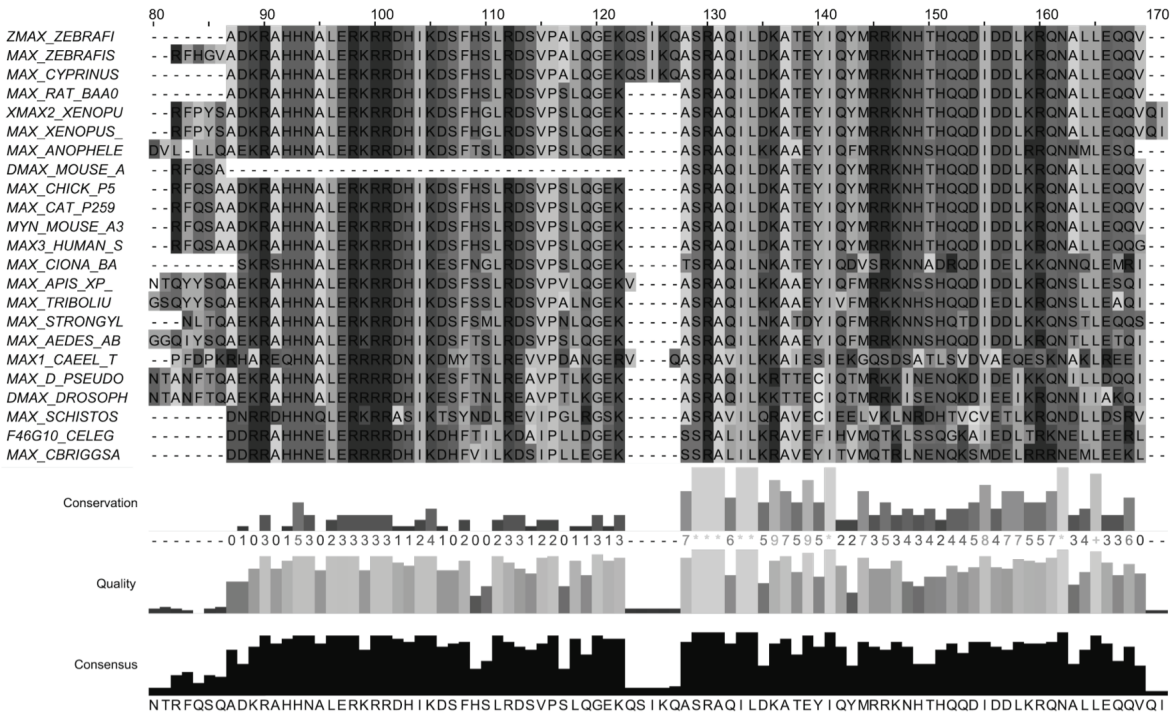

Figure Supplement B

Detail region of the Max homolog alignment.

References

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

Atchley, W. R. and A. D. Fernandes (2005). "Sequence signatures and the probabilistic identification of proteins in the Myc-Max-Mad network." Proceedings of the National Academy of Sciences of the United States of America 102(18): 6401-6406.

Atchley, W. R., W. Terhalle, et al. (1999). "Positional dependence, cliques, and predictive motifs in the bHLH protein domain." Journal of Molecular Evolution 48(5): 501-516.

Balakrishnan, N. and A. C. Cohen (1991). Order statistics and Inference: Estimation Methods. San Diego, CA, Academic Press.

Berger, J. (2006). "The Case for Objective Bayesian Analysis." Bayesian Analysis 1(3): 385-402.

Berger, J. O. and J. M. Bernardo (1989). "Estimating a Product of Means - Bayesian-Analysis with Reference Priors." Journal of the American Statistical Association 84(405): 200-207.

Berger, J. O. and J. M. Bernardo (1992). "Ordered Group Reference Priors with Application to the Multinomial Problem." Biometrika 79(1): 25-37.

Bernardo, J. M. (1974). "Reference Posterior Distributions for Bayesian Inference." Proceedings of the Royal Society of London Series B 41: 113-147.

Bernardo, J. M. and J. M. Ramon (1998). "An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters." Journal of the Royal Statistical Society Series D 47(1): 101-135.

Bernardo, J. M. and A. Smith (1994). Bayesian Theory. New York, John Wiley and Sons.

Clamp, M., J. Cuff, et al. (2004). "The Jalview Java alignment editor." Bioinformatics 20(3): 426-427.

Devroye, L. (1986). Non-uniform random variate generation. New York, Springer-Verlag: http://cg.scs.carleton.ca/~luc/rnbookindex.html.

Drummond, A. and A. Rambaut. (2003). "BEAST." Revision 1.3. Retrieved 26 October 2005, from http://evolve.zoo.ox.ac.uk/beast/.

Egozcue, J. J., V. Pawlowsky-Glahn, et al. (2003). "Isometric Logratio Transformations for Compositional Data Analysis." Mathematical Geology 35(3): 279-300.

Feller, W. (1967). An introduction to probability theory and its application. New York,, Wiley.

Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." Journal of Molecular Evolution 17(6): 368-76.

Felsenstein, J. (2001). "Taking Variation of Evolutionary Rates Between Sites into Account in Inferring Phylogenies." Journal of Molecular Evolution 53(4 - 5): 447.

Felsenstein, J. (2004). Inferring Phylogenies. Sunderland, Massachusetts, Sinauer Associates.

Fernandes, A. D. and W. R. Atchley (2006). "Gaussian Quadrature Formulae for Arbitrary Positive Measures." Evolutionary Bioinformatics Online 2: 261-269.

Goyal, P. (2005). Prior Probabilities: An Information-Theoretic Approach. Bayesian inference and maximum entropy methods in science and engineering. K. H. Knuth. Melville, N.Y., American Institute of Physics: 366-373.

Grandori, C., S. M. Cowley, et al. (2000). "The Myc/Max/Mad network and the transcriptional control of cell behavior." Annual Review of Cell And Developmental Biology 16: 653-699.

Gu, X., Y. X. Fu, et al. (1995). "Maximum-Likelihood-Estimation of the Heterogeneity of Substitution Rate among Nucleotide Sites." Molecular Biology and Evolution 12(4): 546-557.

Jaynes, E. T. (1968). "Prior Probabilities." IEEE Transactions on Systems Science and Cybernetics 4(3): 227-241.

Jeffreys, H. (1946). "An Invariant Form for the Prior Probability in Estimation Problems." Proceedings of the Royal Society of London Series A 186: 453-461.

Jeffreys, H. (1961). Theory of Probability. Oxford, Clarendon Press.

Kass, R. E. and L. Wasserman (1996). "The Selection of Prior Distributions by Formal Rules." Journal of the American Statistical Association 91(435): 1343-1370.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. Evolution of Genes and Proteins. K. M. Nei and R. Sunderland, MA, Sinauer Associates: 208-233.

Kosakovsky Pond, S. L. and S. D. W. Frost (2005). "A Simple Hierarchical Approach to Modeling Distributions of Substitution Rates." Molecular Biology and Evolution 22(2): 223-234.

Kraemer, H. (1999). "Sampling Uniformly from the N-Simplex." from http://groups.google.com/group/sci.stat.math/browse_thread/thread/35257b2947e00915/eae0fc182e3f220e.

Leonard, T. and J. S. J. Hsu (2001). Bayesian methods : an analysis for statisticians and interdisciplinary researchers. Cambridge ; New York, Cambridge University Press.

Levens, D. L. (2003). "Reconstructing MYC." Genes and Development 17(9): 1071-1077.

Luscher, B. (2001). "Function and regulation of the transcription factors of the Mye/Max/Mad network." Gene 277(1-2): 1-14.

Mayrose, I., N. Friedman, et al. (2005). "A Gamma mixture model better accounts for among site rate heterogeneity." Bioinformatics 21: 151-158.

Mayrose, I., D. Graur, et al. (2004). "Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior." Molecular Biology and Evolution 21(9): 1781-1791.

Meyer, S. and A. von Haeseler (2003). "Identifying site-specific substitution rates." Molecular Biology and Evolution 20(2): 182-189.

Nair, S. K. and S. K. Burley (2003). "X-ray structures of Myc-Max and Mad-Max recognizing DNA: Molecular bases of regulation by proto-oncogenic transcription factors." Cell 112(2): 193-205.

Nasi, S., R. Ciarapica, et al. (2001). "Making decisions through Myc." FEBS Letters 490(3): 153-162.

Pupko, T., R. E. Bell, et al. (2002). "Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues." Bioinformatics 18(Suppl 1): S71-S77.

R Development Core Team (2006). R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing, http://www.R-project.org.

Robbins, H. (1956). An Empirical Bayes Approach to Statistics. Proceeding of the Third Berkeley Symposium on Mathematical Statistics, Berkeley, University of California Press.

Robert, C. P. (2001). The Bayesian choice : from decision-theoretic foundations to computational implementation. New York, Springer.

Robert, C. P. and G. Casella (2004). Monte Carlo Statistical Methods. New York, Springer-Verlag.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics 19(12): 1572-4.

Seidenfeld, T. (1987). Entropy and Uncertainty. Foundations of statistical inference. I. B. MacNeill and G. J. Umphrey. Norwell, MA, Reidel: 259-287.

Smith, N. A. and R. W. Tromble. (2004). "Sampling Uniformly from the Unit Simplex." Technical Report, Johns Hopkins University, available at http://nlp.cs.jhu.edu/~nasmith/sampling.pdf.

Strang, G. (1986). Introduction to applied mathematics. Wellesley, Mass., Wellesley-Cambridge Press.

Subramanian, A. R., J. Weyer-Menkhoff, et al. (2005). "DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment." BMC Bioinformatics 6: 66.

Syversveen, A. R. (1998). "Noninformative Bayesian Priors. Interpretation and Problems With Construction and Applications." from http://www.math.ntnu.no/preprint/statistics/1998/S3-1998.ps.

Valdar, W. S. J. (2002). "Scoring residue conservation." Proteins: Structure, Function, and Genetics 48(2): 227-241.

Whelan, S. and N. Goldman (2001). "A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach." Molecular Biology and Evolution 18(5): 691-699.

Wollenberg, K. R. and W. R. Atchley (2000). "Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap." Proceedings of the National Academy of Sciences of the United States of America 97(7): 3288-3291.

Yang, Z. (1994). "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods." Journal of Molecular Evolution 39(3): 306-14.

Yang, Z. H. and S. Kumar (1996). "Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites." Molecular Biology and Evolution 13(5): 650-659.

Zhou, Z. Q. and P. J. Hurlin (2001). "The interplay between Mad and Myc in proliferation and differentiation." Trends In Cell Biology 11(11): S10-S14.

Chapter 3

Molecular Evolution and Conservation in the p53 Family of Tumor-Suppressor Proteins

(Molecular Evolution of p53)

Andrew D. Fernandes[1,2,4] & William R. Atchley[1,2,3]

[1]Graduate Program in Biomathematics
[2]Center for Computational Biology
[3]Department of Genetics
North Carolina State University
Raleigh, NC 27695-7614

[4]Corresponding Author
andrew@fernandes.org

Abstract

p53 is a widely recognized tumor-suppressor protein. It has been estimated that p53 is either mutated or lost in ~50% of all human cancer cases worldwide. However, a detailed evolutionary history of p53 has yet to be inferred, partly because it is often unclear whether relevant sequences displaying low similarity share a true evolutionary relationship. In this paper, we examine both statistical and biochemical evidence supporting putative phylogenetic relationships among various non-chordate species. This is particularly true with regard to the postulated β-sandwich super-family of transcription factors to which p53 may belong. Both conserved and putatively hypervariable sites are examined, especially with respect to known p53 structure and function. Lastly, we examine the consistency and implications of evidence pointing both toward and away from the currently accepted p53 phylogeny.

Background

Genes are typically classified into gene families, the members of which are readily identifiable via shared functional or structural characteristics, or inferred evolutionary history. Gene families commonly arise through gene-duplication and subsequent divergence (Ohno 1970; Lynch and Conery 2000). However, the tumor suppressor p53 is an intriguing exception to this established duplicate-diverge paradigm. Currently, the only known p53 paralogs are the p63 and p73 transcription factor families (Ichimiya, Nakagawara et al. 2000; Moll, Erster et al. 2001; Strano, Rossi et al. 2001; Westfall and Pietenpol 2004), with p63 and p73 being more closely related to each other than to p53. Note that p63 is often denoted as p73L for "p73-like".

As a widely recognized tumor-suppressor gene, it has been estimated that p53 is either mutated or lost in ~50% of all human cancer cases (Hollstein, Sidransky et al. 1991; Levine, Momand et al. 1991). The tumor-suppressing nature of p53 appears to stem from its ability, upon recognition of DNA damage, to (a) activate DNA repair mechanisms, (b) arrest the cell cycle at the G_1/S regulation point, and (c) initiate apoptosis if the damage is irreparable (Levine, Momand et al. 1991; Hofseth, Hussain et al. 2004). Expression levels of p53 are regulated by a complex interaction network, exhibiting a nonlinear "pulsed" dose-response to DNA damage (Casci 2004; Lahav, Rosenfeld et al. 2004). Additional functions of p53 have also been reported, such as a putative role in mitochondrial respiration (Matoba, Kang et al. 2006).

With its implied health importance, it is surprising that the p53 family is evolutionarily constrained to such a few members. Laboratory investigation of putative p53 homologs within protstomes, echinoderms, and protists, would suggest that p53 is highly conserved only within *chordate* taxa (Strano, Rossi et al. 2001), but may be highly divergent in other phyla. The scarcity of sequence data for p53 outside chordates, coupled with the low observed sequence variability of chordate p53 family members, results in surprisingly low statistical power for the identification of p53 motifs in non-chordates. There is little quantitative information about site-specific evolutionary rates in conserved domains among p53 orthologs and p63 and p73 paralogs. Inter-species multiple-alignments, for example from

Soussi and May (1996, Figure 2) and Glazko, et al. (2004, Figure 3), generally show strong conservation especially in the DNA-binding and tetramerization domains (Soussi, Defromentel et al. 1990).

Without a reasonable estimate of phylogeny, multiple alignments can only provide limited information about how conserved a particular motif actually is, since single mutation events can affect multiple alignment rows. Furthermore, since unambiguous p53-family sequences are mostly known from chordates, it would be incorrect to estimate relative site-specific substitution rates without incorporating both phylogenetic corrections and non-chordate sequences.

In this paper we examine a diverse collection of protein sequences including chordate p53, p63, and p73 as well as putative arthropod, mollusc, nematode, echinoderm, and protist p53 family members. The goal is to answer several important evolutionary questions. First, what non-chordate proteins are true homologs to the p53 family? Second, when does the p53 family arise in evolution? Third, do conserved sites also occur outside of the DNA binding domain in the p53 family?

First, we describe the chordate p53 family, then next describe non-chordate p53 sequences, examining the experimental evidence identifying them as p53 homologs.

Chordate p53 Biology

Some aspects of the molecular biology of the human p53 family are summarized in Figure 1. Human p53 is translated from a single mRNA with a single open reading frame. The human *p53* gene is ~20 kb long and contains eleven exons (Soussi and May 1996). It is 393 amino acids long and includes three functional domains: an N-terminal transactivation domain (TAD), a central DNA binding domain (DBD), and a C-terminal oligomerization domain (OLD). Experimental and clinical data suggest that mutation in any of these domains may significantly impair both the binding of p53 to recognition sites of target genes and subsequent transcription (Levine, Momand et al. 1991; Strano, Rossi et al. 2001). C-terminal alternatively spliced products in humans appear rarely, are expressed at low levels, and seem to be confined to quiescent lymphocytes (Flaman, Waridel et al. 1996), although recent

evidence suggests that these isoforms have functional importance and have been evolutionarily conserved (Courtois, de Fromentel et al. 2004; Bourdon, Fernandes et al. 2005). In contrast, C-terminal alternate splicing occurs more readily in rodents and may have species-specific implications (Almog, Goldfinger et al. 2000; Laverdiere, Beaudoin et al. 2000).

Complexes of the p53 DBD (residues 102-292) bound to DNA have been crystallized and structurally resolved (Cho, Gorina et al. 1994), although whole p53 tetramers bound to DNA have yet to be produced. The core structure of the DBD appears as a pair of anti-parallel β-sheets sandwiched together. The sandwich seems to have two functions: (a) it forms a scaffold for two large loops and a loop-sheet-helix structure and (b) it interacts with the major groove of DNA during binding. The two loops are held together in tetrahedral coordination with a zinc ion and they interact with the minor groove of DNA during binding. Both the loops and loop-sheet-helix motif appear to be the most conserved elements of the DBD (and p53 in general), an observation that supports the hypothesis that DNA binding is a critical function of p53.

The canonical human p53 target binding sequence consists of tandem repeats of the palindromic sequence RRRCWWGYYY where R is a purine, C is cytosine, W is adenine or thymidine, G is guanine, and Y is a pyrimidine (El-Deiry, Kern et al. 1992; Funk, Pak et al. 1992). Variations of this motif are known to exhibit a wide range of binding specificities. Furthermore, the motif is commonly repeated as a pair of inverted repeats, separated by between zero and fourteen spacer nucleotides. Unfortunately, motif plasticity and variation in binding specificity make it difficult to identify true p53 binding sites in the genome.

Unlike p53, numerous splice-variants of both p63 and p73 occur. Human *p63* and *p73* genes are each ~65 kb in length and contain 15 and 14 exons, respectively (Ichimiya, Nakagawara et al. 2000; Westfall and Pietenpol 2004). They show high similarity to *p53* in their intron/exon structure (Moll, Erster et al. 2001; Strano, Rossi et al. 2001). A major difference between p53 and p63/p73 is that the latter have two additional N-terminal domains: a Sterile Alpha Motif (SAM) domain, implicated in protein-protein interactions, followed immediately by a post-SAM (PS) domain, implicated in transcriptional suppression

(Yang, Kaghad et al. 1998; Ichimiya, Nakagawara et al. 2000). Three major p63 isoforms (denoted α, β, and γ) can be translated from a single mRNA transcript resulting in isoforms with common 5′-termini and alternatively spliced 3′-termini. These isoforms have intact transactivation domains and are often termed TA-variants. Three additional N-terminal deleted isoforms (denoted ΔNα, ΔNβ, and ΔNγ) are generated by an internal promoter located upstream of exon three resulting in a total of six unique p63 isoforms (Levrero, De Laurenzi et al. 1999). Six TA-type isoforms of p73 (denoted α, β, γ, δ, ε, and ζ) have been reported. All of these differ by alternative splicing events at the C-terminus, which together with a murine N-terminal deleted variant (denoted ΔNα) give a total of seven isoforms (Strano, Rossi et al. 2001). It appears likely that all C-terminal p73 isoforms have endogenous ΔN-respective variants. Finally, all p63 isoforms appear to occur as variants, denoted "TA*", having 39 additional amino acids at the N-terminus (Yang, Kaghad et al. 1998; Yang, Kaghad et al. 2002).

Recent work by Ortt and Sinha (2006) used the Systematic Evolution of Ligands by Exponential Enrichment (SELEX) technique (Klug and Famulok 1994) to determine the optimal consensus binding sequence for human p63. They found it to be the non-palindromic ten base-pair motif WWACWTGTWT consisting of a CWTG core with adjacent W-rich 5′ and 3′ flanking regions. While similar to the optimal binding motif for p53, significant differences in both the core and flanking regions of each motif are evident. Unlike p53, no preference was found for binding to tandem repeats, although this absence could be an experimental artifact. Like p53, though, considerable variability in binding affinity was observed over various realizations of the motif.

In humans, p63 and p73 display higher similarity to each other than they do to p53. p63 and p73 exhibit 65% identity in the DBD including conservation of all DNA-contact and structural residues that are considered "hotspots" for tumorogenic mutation (Walker, Bond et al. 1999; Moll, Erster et al. 2001; Glazko, Koonin et al. 2004). Although often considered high in the p53 family literature, this level of similarity in a DNA-binding region is relatively

modest compared to other families of transcription factors such as AP-2 which may exhibit up to 90% similarity (Yang, Kaghad et al. 2002; Tummala, Romano et al. 2003).

These patterns of similarity suggest the evolutionary split between p63 and p73 was more recent than the split of these two from p53. Considerably less similarity exists among family member OLDs, suggesting that each protein evolved to support exclusive homo-oligomerization. Observations by Moll et al. (2001) and Strano et al. (2001) indicate that inter-family hetero-oligomerization does not appear occur at appreciable levels *in vivo*. Furthermore, *in vitro* two-hybrid binding and transcriptional activation data indicate that only homo-complexes are found in both normal and tumor cells.

Phenotypic studies of $p53^{-/-}$, $p63^{-/-}$, and $p73^{-/-}$ knockout mice indicate considerable functional divergence among these proteins. For instance, it is well known that p53-deficient mice are prone to a high incidence of spontaneous tumors, especially sarcomas and lymphomas (Donehower, Harvey et al. 1992). However, p63-deficient mice display severe ectoderm related developmental abnormalities, indicating a role in regenerative limb proliferation and craniofacial and epithelial development (Yang, Schweitzer et al. 1999). Other work shows a role for p63 in cellular senescence and aging (Keyes and Mills 2006), and it may be involved in the p53-response to UV irradiation (Yang, Kaghad et al. 2002). On the other hand, p73-deficient mice display neurological, pheromonal, and inflammatory defects, but lack spontaneous tumors (Yang, Walker et al. 2000). Other putative functions for p73 include involvement in gamma-radiation response, T-cell apoptosis, inflammatory response, and pheromone detection (Yang, Kaghad et al. 2002).

While p53 behaves as a model tumor-suppressor protein, it would appear that p63 and p73 may play an important role in ectodermal development and neurogenesis, respectively. Nonetheless, p63 at least is known to play an important role in the development of cancer (Westfall and Pietenpol 2004), possibly through an interaction with p53 (Yang, Kaghad et al. 2002). However, it is unclear whether p63 can be classified as an oncogene or tumor suppressor (Mills 2006).

Methods

Identification of both unambiguous and putative p53 family sequences was accomplished using several search strategies. First, we started with known chordate p53, p63, and p73 sequences as annotated within the UNIPROT 7.5 protein databank (Bairoch, Apweiler et al. 2005). Note that p63 is often aliased as p73L. These were augmented with entries from the PFAM 19.0 database (Finn, Mistry et al. 2006) using family entries p53 (PF00870), p53_tetramer (PF07710), and p53_TAD (PF08563). Both UNIPROT and GENBANK 154.0 (Benson, Karsch-Mizrachi et al. 2006) were scanned using BLAST and PSI-BLAST 2.2.14 (Altschul, Gish et al. 1990; Altschul, Madden et al. 1997) and HMMER 2.3.2 (Durbin 1998). Numerous literature and cross-reference searches were performed to help identify other putative homologs with sequence identity too low for purely computational detection. Results were manually curated to remove non-informative fragments, duplicate sequences, noninformative pseudogenes, and obvious sequencing errors. The resultant final database is presented as Table 1. Our sequence database was aligned using T-Coffee 4.45 (Notredame, Higgins et al. 2000). A Maximum Likelihood (ML) phylogenetic tree was estimated via PHYML 2.4.4 (Guindon and Gascuel 2003) using the WAG model of protein evolution (Whelan and Goldman 2001) and gamma-distributed rate heterogeneity. Topology, branch lengths, and gamma shape were optimized for tree reconstruction. Reliability of the tree was estimated by bootstrapping a consensus tree of 100 PHYML replicates. Conserved sites and corresponding implied rates of evolution were determined with MUTABLE 0.9.1 (Fernandes and Atchley 2006). The sequence database, alignment, a graphical view of the alignment, and phylogenetic tree are available online as Supplementary Material.

The statistical significance of conserved motifs was assessed by through motif profiles built using both PSI-BLAST and HMMER with the p53 alignment as input. For PSI-BLAST, three query profiles were built, each based on a human p53, p63, or p73 sequence. Iteration continued until convergence, using the entire p53-family database as the search target. For HMMER, profiles were built and calibrated using all four modes of HMM construction: default domain alignment, multi-hit local, global alignment, and local alignment. Standardized bit scores for each search were columnated and used as an indication of

"quality of fit" to the p53 family profile. A discussion of the interpretation of standardized bit scores can be found online at http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html, and will be briefly recapped, below.

Results & Discussion

Many of the sequences found have very low sequence similarity to human p53 family proteins. Nonetheless, they have been classified as p53 homologs in the literature based on biochemical, functional, and structural evidence. Below, we summarize evidence for classifying non-chordate sequences into the p53 family. With the exception of one echinoderm and one entamoeba sequence, all non-chordate sequences found were protostomes.

Non-chordate p53 Homologs

The close evolutionary relationship of chordate p53 paralogs and orthologs is exemplified by the high degree of similarity among them. Considerably less similarity exists between chordate and non-chordate p53 family sequences, making identification of non-chordate p53 family members difficult. As a consequence, such identification is rarely accomplished by purely computational sequence analyses. Low similarity is at least partially explained by the large evolutionary time scales separating chordate and non-chordate sequences. All non-chordate p53 sequences found by us belong to the protostomes with the exception of one echinoderm and one protist. Estimates of the time of protostome/deuterostome divergence vary from 670 (Ayala, Rzhetsky et al. 1998) to 730 (Peterson and Butterfield 2005) to 980 (Hedges, Blair et al. 2004) million years ago, with the echinoderm/chordate split occurring a mere 70 million years after the protostome/deuterostome split (Ayala, Rzhetsky et al. 1998). Further, many of the protostomes, particularly nematodes, are known to have relatively high rates of evolutionary change (Coghlan 2005).

Mollusc

Six mollusc sequences were found, including five bivalve (clam and mussel) and one cephalopod (squid) species. The mollusc sequences were among the most obvious p53 family

members, exhibiting 35% identity and 51% similarity with the human TA-p63α protein over its entirety, including the SAM domain. Studies show that considerable conservation also occurs at the gene level with extensive promoter, intron/exon structure, and 3′-UTR conservation both among mollusc sequences and between mollusc and chordate *p63* (Muttray, Cox et al. 2005). These mollusc homologs are relatively easy to detect computationally, due to their high similarity to *p53*. Thus similarity alone strong argues for true homology.

Specifically, mollusc *p53* homologs are reported for genera *Loligo* (squid) (Winge, Friend et al. 1996), *Mya* (soft-shelled clam) (Kelley, Winge et al. 2001), *Spisula* (surf-clam) (Cox, Stephens et al. 2003), and *Mytilus* (mussel) (Muttray, Cox et al. 2005). Unfortunately, reports include only data on the genetic architecture of these genes and have limited (if any) information on gene expression, protein function, or protein structure. Thus the mollusc data clearly indicate the presence of an ancestral *p53* gene, but its function is unclear.

Interestingly, Kelley et al. (2001) claim discovery of two distinct *p53* paralogs within *Mya*. One of these genes was proposed to be a homolog of p53, with the other being a homolog of p73. Both sequences are virtually identical over numerous functional domains. Similarly, Cox et al. (2003) also cloned two separate *p53*-like sequences from *Spisula*, but determined that their corresponding genes coded for identical protein sequences. Divergence between these genes was seen only in the 3′ UTR, making the clones either alternate transcripts of the same gene, or transcripts of two genes that are extremely recent duplicates. Finally, Muttray et al. (2005) found only one *p53* homolog in each *Mytilus* species examined. Given that *Mya* and *Spisula* are more similar to each other than either is to *Mytilus*, this may provide a clue as to when the putative gene duplication occurred within bivalves.

Nematode

Two putative p53 homologs were found in the nematodes *Caenorhabditis elegans* and *C. briggsae*. Statistically significant similarity to the human p53 family was detected only within a ~200 amino acid segment of the DNA binding domain, with 21% identity and 36% similarity for *C. elegans* and 14% identity and 34% similarity in *C. briggsae*. These

nematode orthologs display only 33% identity and 51% similarity to each other over their ~685 amino acid length. Such divergence is large considering that only 80 to 110 million years are estimated to separate *C. elegans* and *C. briggsae* (Gupta and Sternberg 2003). This relatively short time span is only slightly longer than the estimated 65 to 75 million years separating humans and mice whose p53 orthologs are 77% identical and 83% similar (Waterston, Lindblad-Toh et al. 2002).

Curiously, although homology was detected only within the DBD for *C. elegans* with respect to p53, p63, and p73, PSI-BLAST profile searching found a significant match $\left(E \approx 8 \times 10^{-67}\right)$ specifically between *C. briggsae* and human p73 over a much longer 424 amino acid region containing the DBD, with 15% identity and 30% similarity. It is not clear why a similar region could not be identified for *C. elegans*, even with exceedingly permissive search thresholds. This observation, coupled with observations from *Mollusca*, supports the hypothesis that p63/p73 represent the more ancestral state of the p53 family.

The *C. elegans* ortholog of *p53*, termed *cep-1* for "*C. elegans* p53-related 1", was discovered by Schumacher et al. (2001). Alignment profiles, generated from accepted *p53* sequences, were used to search the *C. elegans* genome. The resultant gene coded a 645 amino acid protein that, although displaying fairly low sequence identity with p53, nonetheless appears to conserve numerous residues required for both DNA and Zn^{2+} binding. Schumacher et al. further claim that a small set of conserved amino acids can also be detected in a putative C-terminal oligomerization domain, but the significance of such conservation is not supported by either our motif profiles or our conserved motif detection, discussed below. Perhaps the most significant observations of Schumacher et al. are that (a) the Cep-1 protein is a transcription factor, and (b) the *cep-1* gene is required for DNA damage-induced apoptosis; both these observations support the hypothesis that the function of *cep-1* is homologous to *p53*, even though the proteins that they code for have restricted similarity.

Given their low similarity, it is surprising that Cep-1 appears able to activate transcription from reporter plasmids containing human p53 binding sites (Derry, Putzke et al. 2001; Schumacher, Hofmann et al. 2001), and although it is not known if p53 and Cep-1 have

similar binding site recognition sequences, this observation still implies the hypothesis that p53 and Cep-1 have similar DNA binding specificities despite their low similarities. However, recent structural studies have shown that structural features of human p53 that are important for sequence-specific DNA binding are not conserved in Cep-1 (Huyen, Jeffrey et al. 2004), thus putting these proteins' functional homology to question.

An intriguing argument for functional homology comes from the iASPP (inhibitory member of the ASPP protein phosphatase family) protein which appears to be conserved from *C. elegans* to humans (Bergamaschi, Samuels et al. 2003). The iASPP protein is encoded by *PPP1R13L* in humans and *ape-1* in *C. elegans* and in both organisms appears to be a key inhibitor of p53 or Cep-1, respectively. Inhibition of iASPP by RNA-mediated interference induces p53-dependent apoptosis, while increased expression of iASPP confers resistance to both ultraviolet- and cisplatin-induced apoptosis, in both human and nematode cells. The conserved interaction between iASPP and p53/Cep-1 is unexpected since p53/Cep-1 conservation is only detectable in the DNA binding region, implying the conservation of protein-protein interactions without concomitant conservation of primary structure. Detection of additional putative p53 interaction partners has continued with somewhat limited success, for instance with the *C. elegans* GLD-1 germ line development regulatory protein, which is related to the mammalian Quaking family of proteins (Schumacher, Hanazawa et al. 2005).

Arthropod

Four distinct arthropod putative p53 homologs were found, all in the homometabolous insects *Drosophila melanogaster* (fruit fly), *Anopheles gambiae* (mosquito), *Leptinotarsa decemlineata* (potato beetle), and *Tribolium castaneum* (flour beetle). The *Drosophila* ortholog has been shown to display transcriptional activation activity, binding to the mammalian p53 recognition sequence (especially those of *p21* and *GADD45*), over-expression-dependent induced apoptosis but not G_1 arrest, and reduced inhibition-dependent DNA-damage-induced apoptosis (Jin, Martinek et al. 2000; Ollmann, Young et al. 2000). Structural threading provided evidence consistent with the hypothesis that the *Drosophila*-p53 DBD is similar to its chordate ortholog (Jin, Martinek et al. 2000). Statistically significant similarity to p53 can only be detected over a ~220 amino acid region in the DBD

where only 24% identity and 42% similarity is observed. Taken together, these observations support the hypothesis that these proteins are true orthologs within their DNA binding domains.

Interestingly, the region of similarity is expanded to ~303 amino acids (again including the DBD) specifically when *Drosophila* is compared to human p63 (not p53 or p73), whereas the *Anopheles* region of homology is increased by almost 100 amino acids if compared to p63 over p53. It is not clear if individual insect species similarities reflect a true closer relationship or merely statistical noise, as the homologous regions appear to be the effective threshold of detection. Hints that these additional regions of similarity may be statistical noise come from examination of the two beetle species: both beetles appear to have homologous region ~330-360 amino acids long; however the region of similarity is different for each of p53, p63, and p73. The only overlap among regions is within the DBD, where it would normally be expected.

Echinoderm

Statistical profile searches indicated a putative distant p53 homolog in the echinoderm *Strongylocentrotus purpuratus* (purple urchin) via conceptual translation of genomic sequences (see Table 1). This 665 amino acid protein has 35% identity and 49% similarity over a 291 amino acid region corresponding to the p53 DBD. However, a comparison with both human p63 and p73 shows 28% identity and 42% similarity over the entirety of each of these two proteins' lengths. This represents strong evidence for the hypothesis that echinoderm-p53 is much more similar to p63 or p73 than it is to p53.

A review of the literature revealed only one molecular study of urchin p53. Lesser et al. (2003) exposed embryos of the green urchin *Strongylocentrotus droebachiensis* to DNA-damaging doses of UVB radiation and used immunoblotting with polyclonal antibodies to human p53, p21, and CDC2 to assay cellular response. As expected, the cells arrested in the G_1/S transition with concomitant up-regulation of p53 and p21, and down-regulation of CDC2. Unfortunately, no details were provided on cross-reactivity or specificity of the antibodies used, with respect to the urchin. Previous similar work was performed with embryos of the chordate *Gadus morhua* (Atlantic cod fish), a species known to have a p53

ortholog very similar to human (Lesser, Farrell et al. 2001). Cross-reactivity with p63 or p73 would be expected to play a significant role in positively identifying an echinoderm p53 homolog via immunochemistry. Therefore, given the low sequence identity between human and urchin p53, these results should be interpreted with caution.

Entamoeba

The only non-metazoan, or indeed non-bilaterian, p53 homolog reported is the 430 amino acid protein Ehp53 from the protist *Entamoeba histolytica* (Mendoza, Orozco et al. 2003). Detectable homology is only found within a 50-residue motif within the p53 DBD, with 38% identity and 48% similarity to human p53, although this motif is reported to include seven of the eight DNA-binding residues and two of the four Zn^{2+}-binding sites described for human p53. No significant similarity was detected with respect to p63 or p73. Several lines of evidence were presented supporting the claim of homology. First, *E. histolytica* nuclear extracts were shown via gel-shift assay to contain a protein that binds the human p53 recognition consensus sequence. Next, monoclonal antibodies against human p53 recognized a single 53 kDa spot in two-dimensional gels, recognition that was inhibited by antibody binding. Lastly, expression levels of Ehp53 were increased after exposure to UV-radiation. Confocal microscopy indicated expression of Ehp53 in the nucleus, *E. histolytica* kinetoplastid organelles (EhKO) and cytoplasm.

Origin of the p53 Family

Structurally, p53 has been hypothesized to belong to the β-sandwich domain family of DNA-binding transcription factors (Berardi, Sun et al. 1999; Rudolph and Gergen 2001). Members of the β-sandwich domain family are structurally similar, but no comprehensive study has been done to assert actual homology. Family members include a diverse range of DNA-binding transcription factors including the NTD80/PhoG, RHD, Runt, T-Box, STAT, and p53 families of proteins. The β-sandwich domain is comprised mainly of β-strands arranged in an anti-parallel fashion to form a β-barrel with an s-type immunoglobulin (Ig) fold. DNA binding is usually mediated by loops that extend from the Ig-motif, often in coordination with a metal ion. The loops joining the β-strands are highly variable and are

sometimes important in additional protein-protein or protein-ligand interactions. At best, only a very low degree of sequence similarity is observed between β-sandwich family members.

If the *Entamoeba* sequence belongs to the β-sandwich transcription factor family, it would imply a family origin on the order of the 1.6 to 2.1 billion year old split between *Animalia* and *Fungi* (Knoll 1992; Hedges, Blair et al. 2004; Steenkamp, Wright et al. 2006). However, it should be noted that of all the β-sandwich transcription factors, homology searching within PFAM indicates that only members of the NTD80/PhoG are found outside *Metazoa*, with several examples detectable within *Alveolata* and *Mycetozoa* (non-metazoan eukaryotes), and numerous examples found within *Fungi*. Non-eukaryote family members are not detectable, nor are other representatives of the β-sandwich family.

Statistical Evidence

Statistical significance for significance of motifs generated via PSI-BLAST and HMMER is summarized in Figure 2, where the range of standardized bit-scores is depicted by a box-plot for each sequence in our database. The values are sorted by decreasing mean bit-score. Standardized bit-scores can be used to test the null hypothesis that two sequences will contain at least one pair of subsequences with high similarity score, purely by random chance. Given a standardized similarity bit score s, the probability p of this null hypothesis being falsely accepted is approximately $p = 1 - e^{-e^{-s}}$. Asymptotically, $s \to +\infty$ yields $p \approx e^{-s} + \mathcal{O}\left(e^{-s^2}\right)$, while $s \to -\infty$ yields $p \approx 1$. As a rough guide, $s \approx 2.97 \Rightarrow p = 0.05$ and $s \approx 4.60 \Rightarrow p = 0.01$, while $s \approx 13.8 \Rightarrow p = 1 \times 10^{-6}$. A strict interpretation of p requires independence of samples (an untrue assumption most sequence databases) and Bonferroni-correction for multiple tests. Similarity scores are based on parameterized computational models; therefore it is better to consider a computed similarity score s as an estimate of the true score, and not a simple point-value. Hence the boxes of Figure 2 provide an idea of the variability of "goodness of fit" between each database sequence and human p53, p63, or p73.

Family Membership

Strong support was found for all database sequences except for *Anopheles*, *Drosophila*, *Caenorhabditis*, and *Entamoeba* species (Figure 2). Although some homology searches found significant matching within all four of these genera, the mean bit-score was near zero for all of them, indicating statistically questionable homology. The greatest variation of bit-scores in these genera seemed to be caused by similarity of the target sequence to only one of p53, p63, or p73 with concomitant dissimilarity to the others. Nontheless, the range of bit-scores reported is indicative of overall statistical support for homology.

The inferred phylogenetic tree, with bootstrap support values for each branch, is shown in Figure 3. Within chordate p53, branch order and lengths mimic known evolutionary relationships. The six p63 (denoted p73L) and five p73 sequences were clearly resolved as distinct with 99% bootstrap support for their separation. Bootstrap support is low for the resolution of p53 from p63/p73 at 76%, likely due to the large C-terminal deletion of the SAM between these two families. This deletion creates many non-informative sites for possible selection during the bootstrap procedure. Restricting the bootstrap procedure to only sites present across both groups to specifically test the significance of this branch provides significantly more evidence for the given branching (data not shown).

A similar bootstrap support of 69% is found for the branching of the echinoderm from chordates. Taken together, the branching pattern presented supports the hypothesis that the gene duplication event separating the p53 from the p63/p73 lineage occurred not just after the protostome/deuterostome split, but after the echinoderm/chordate split as well. Furthermore, this evidence suggests that the ancestral sequence was significantly more similar to p63/p73 than p53, the former showing significant identity over the entire protein lengths including the SAM, and the latter showing significant identity only within the DBD of p53.

The next most distant branching is the protostome/deuterostome split as indicated via *Mollusca*, occurring with 96% bootstrap support. All molluscs clustered in the expected monophyletic relationship, with both branch lengths and sequence similarity supporting the hypothesis that the mollusc sequences are more similar to the deuterostome sequences than to

the insect, nematode, or protist sequences. Representation from an evolutionary intermediate such as *Annelida* would provide valuable insight into the rise of the p53 family.

The most important feature of the mollusc branch is the close evolutionary relationship between *Mya* paralogs. Bootstrap values of 100% support the hypothesis that these two genes are more similar to each other than any found in other bivalves. Thus the gene duplication event creating these paralogs appears to have occurred more recently than even the split between *Mytilus* and *Mya*. This observation does not support the hypothesis of Kelley et al. (2001) that the *Mya* gene pair are orthologous to the chordate p53 and p63/p73 pair.

The next three tree branches belong to insects, with the first branch toward the beetles *T. castaneum* and *L. decemlineata* being highly (84%) supported. Unsurprisingly, *D. melanogaster* and *A. gambiae* displayed fairly low similarity to each other, probably reflecting the estimated 250 million years (Yeates and Wiegmann 1999) and high rate of evolution (Bolshakov, Topalis et al. 2002) separating these organisms.

The subsequent branching of *C. briggsae* and *C. elegans* postulate these sequences are more primitive still. Lastly, rooting the tree on *E. histolytica* accentuates both its evolutionary distance and presumable similarity to the ancestral p53 family protein.

Relation to the β-Sandwich Family

The low inter-family similarity displayed by the β-sandwich transcription factors makes them particularly difficult to align. For our work, the PFAM clan CL007 provides the best quality inter-family alignment available. Rather than aligning the raw protein sequences, PFAM clans align the Hidden Markov Model (HMM) underlying each clan member, and then align individual sequences to each clan member HMM. A PFAM clan contains two or more PFAM families that are hypothesized to have arisen from a single evolutionary origin. Evidence of their evolutionary relationship is usually determined by similarity in tertiary structure or, when structures are not available, by common sequence motifs. Clan representation has been found to be suitable for protein families that are highly divergent and difficult to represent via HMM (Finn, Mistry et al. 2006). It is notable that the β-sandwich

clan members are unusually divergent from each other in comparison to other clans identified within PFAM.

The inferred phylogenetic tree for the β-sandwich clan is shown in Figure 4. Each of the NTD80/PhoG, RHD, Runt, T-Box, STAT, and p53 families were clearly resolved, even though support for some clades was somewhat low. Many of these low-support branches involve the p53 lineage. In particular, branches of low support that separate distinct lineages are noted as follows: (A) the main p53 super-family branch, (B) the known p53 family, (C) the non-chordate and non-mollusc branch, and (D) the Runt family. Using a gamma-distributed model of rate heterogeneity resulted in an optimized gamma shape-parameter of ~9-10, indicating both a very low degree of rate heterogeneity and very small set of conserved sites.

The branching pattern of non-chordate and non-mollusc sequences (Figure 4) implies these sequences are more similar to each other than to any of the chordate or mollusc p53 family members. Surprisingly, the echinoderm sequence is more similar to nematode and arthropod sequences than chordate, albeit with low support. Collapsing branches of low support, however, implies that the p53-questionable sequences of lineage (C) are almost as evolutionarily distant from the chordate p53 family as the p53 family is from the Runt family of transcription factors. Thus improvement in the classification of these sequences almost surely requires additional molecular evidence to augment the computational.

Evolutionary Rates

Rates of amino acid evolution with respect to human p53, p63, and p73 sequences are shown in Figures 5, 6, and 7. These rates are drawn from the posterior distribution of a probabilistic model of amino acid evolution, where the tree and branch-lengths are held constant (Fernandes and Atchley 2006). Phylogenetic correlation is fully accounted for within this model. Conserved residues have low rates while variable sites have high rates of evolution. All rates are relative to a mean rate of one. Note that very high rates signal hypervariable sites, and these sites may discriminate among evolutionary clades. More information on discriminating clades through diagnostic sites can be found in Atchley et al. (2006). To summarize the dense information within the three figures, Tables 2 and 3 detail

the 10% most and least conserved sites from our alignment, respectively. Biological functions are as noted in UNIPROT entries P53_HUMAN, P73L_ HUMAN, and P73_HUMAN.

All three figures and Table 2 give clear evidence for conservation within the p53 DBD. Surprisingly, strong conservation was not exhibited in the transactivation domain (TAD), except at two sites. Moderate to strong conservation of the TAD is often claimed, but our data restrict such conservation to within p53-proper. The last recognizable set of conserved sites appears within the SAM of p63 and p73. Such conservation is notable because the SAM is completely missing from the entire p53 clade, but appears to be conserved as strongly as the DBD. Note that Table 2 contains numerous conserved sites that often do not correspond to any p53, p63, or p73 site. The biological relevance of these sites is not clear, and should be subject to further investigation.

The hypervariable sites noted in Table 3 are interesting specifically because they also contain numerous transactivation, DNA-binding, and SAM sites. At least two sets of residues, each set involved in oligomerization and protein-protein interaction, are notably hypervariable. As with the conserved sites, Table 3 contains numerous conserved sites that often do not correspond to any p53, p63, or p73 site. Again, the biological relevance of these sites is not clear, and should be subject to further investigation.

Conclusions

Computational evidence for the origin of the p53 family is based on two criteria. First, similarity between p63/p73 and mollusc sequences implies an ancestral p53 extant prior to the protostome/deuterostome split. Second, the hypothesis that dual loss-of-function across species is more likely than a dual gain-of-function. That is, it is more likely that the ancestral p53 sequence contained a SAM-like domain that was subsequently lost or replaced in arthropods, nematodes, and chordates. It is correspondingly unlikely that ancestral-p53 resembled modern-p53, and a SAM-like domain was added independently in both protostome and deuterostome lines. An excellent visual summary of an evolutionary history of p53 consistent with our computational evidence was given by Yang (2002) and is presented as Figure 8. Note that since urchins appear to have preserved a recognizable SAM-

like domain, the ancestral gene-duplication leading to the modern p53 and p63/p73 lineages may have occurred after the rise of echinoderms.

Ancestral *p53* gene appears to have undergone duplication and divergence twice in independent lineages. One duplication event was recent and occurred within the bivalve molluscs. The second appears older and gave rise to the p53 and p63/p73 lineages of the chordates. Further duplication and divergence events are evident by two rat pseudogenes, UNIPROT entries Q6I9N3 and Q6QI11. Both entries appear to be nonfunctional, and were probably created by the movement of genomic transposable elements (Hulla 1992; Ciotta, Dogliotti et al. 1995). These events provide important clues as to how p53 duplication events have occurred over its history, especially since transposons are known to be active in some rodents (Weaver 2005).

Family Membership

Nematode, insect, and protist sequences are difficult to identify computationally. Their similarity to known p53 family sequences is so low that simple BLAST or PFAM database searches do not find them. Instead, their discovery requires much more sensitive search strategies. In database searching, *sensitivity* measures how many sequences are correctly identified, whereas, *selectivity* measures how many of those matches are correct. It is well known that higher sensitivity implies lower selectivity, and higher selectivity implies lower sensitivity. Therefore, a p53 search strategy that is sensitive enough to find nematode, insect, and protist sequences can be expected to erroneously find a large number of extraneous matches as well, due to decreased selectivity (Atchley, Terhalle et al. 1999; Atchley and Fernandes 2005).

Given a set of putative homologs, one may expect that their inferred multiple-alignment may help identify true family members. When sequence similarity is low, however, the alignment may depend more on the alignment-algorithm parameters than the input data. For instance, the T-COFFEE algorithm used aligned regions of the nematode, insect, and protist sequences with the p53 DBD. In contrast, the DIALIGN-T algorithm (Subramanian, Weyer-Menkhoff et al. 2005) rejects the hypothesis of homology and aligns these organisms in

independent blocks (data not shown). Both T-COFFEE and DIALIGN-T are suitable for aligning distant homologs such as those in BALiBASE (Thompson, Koehl et al. 2005). So which algorithm is correct? For p53, perhaps this question can be answered by simultaneous inference of phylogenetic tree and alignment, as done by Redelings et al. (2005) in the package BAli-Phy (Suchard and Redelings 2006). The fact remains that at low levels of similarity, very sequence data alone provides little information about homology.

It is inevitable that determining homology for distantly related proteins will require combined biological and statistical evidence. Numerous researchers have done precisely that, especially for the nematode, insect, and protist sequences. However, the biological evidence must be concordant with the statistical evidence that (a) putative homology for distant sequences is generally restricted to the DBD and (b) the ancestral p53 sequence is expected to be more p63/p73-like than p53-like. Close examination of their findings, however, reveals possible incompatibilities in these lines of evidence.

Biological Function

There are numerous examples of such possible incompatibility. For instance, the existence of putative *C. elegans* homologs for iASPP and GLD-1, two human p53 inhibitors, has been put forth as evidence of Cep-1/p53 homology. This observation appears to be incompatible with the observation that p63/p73 is closer to the ancestral parent of the sequences. For urchin, antibodies specific for human p53 are assumed to cross-react with the urchin p53 sequence. The cross-reactivity implies epitope, and hence structural, similarity. The antibodies used in the urchin study were shown to specifically and selectively bind to p53 from the chordate *Gadus morhua* (Atlantic cod fish), with no significant cross-reactivity to other proteins (Lesser, Farrell et al. 2001). Given the evidence that urchin p53 is really more p63/p73 in character, we would expect *not* to have significant binding, in contrast to what was reported.

The same argument applies to the immunochemical data from the protist *E. histolytica*, where antibodies specific for human p53 are shown cross-react with the protist p53 sequence. The case of *E. histolytica* is more incongruous than that of urchin, however, because not only

is sequence similarity with p53 is almost undetectably low, but the similarity is restricted to a short region tentatively ascribed similar to the DBD.

Studies of *E. histolytica*, *D. melanogaster*, and *C. elegans* homologs found the DNA recognition binding sequence preferentially bind the human p53 motif. However, given the differences in p53 and p63 binding domains, and the likelihood that p63 is closer to the ancestral sequence, we would expect the preferred recognition sequence to be closer to p63, not p53. Given the plasticity of binding specificity exhibited by p53 and p63, transcription factor binding specificity may not be good evidence for homology. In the future, binding specificities should be compared and contrasted over a broad range of binding motifs for p53, p63, and p73. Currently, the p73 recognition binding sequence is not known.

Lastly, possibly contradictory evidence exists with respect to the function of distant p53 relatives. Although p63 and p73 are suspected of playing a role in cellular response to DNA-damaging radiation, splice variants such as ΔNp63 are known to act as a p53-antagonist (Yang, Kaghad et al. 2002). Upon DNA damage, ΔNp63 is actively down-regulated as part of p53 induction. Thus even the function of ancestral-p53 is not clear, with evidence that it may have been either up- or down-regulated after DNA damage.

After reviewing the available data, we feel that the protist *E. histolytica* sequence to be the most incongruous. Other proteins, specifically from the insects and nematodes, seem to have stronger evidence for familial relation, but we stress that the nature of that relationship is actually quite unclear. Further work combining both statistical, phylogenetic, and biological information is clearly needed to draw an accurate picture of the rise of the p53 family of transcription factors. Clearly, functional studies of distantly related proteins should be interpreted with caution. It may not be correct to conclude that proteins that share as little as 15% identity have similar, or even related, biological function.

The β-Sandwich Clan

How much value should similarity of tertiary structure be given to questions of homology? Many dissimilar protein sequences fold into similar structures. A challenge facing phylogenetics is the discrimination between homology and analogy for structurally

similar domains that lack significant sequence similarity. For instance, Brendel et al. (1989) have argued statistically that the leucine zipper domain has arisen many times independently throughout history, making many of the leucine zipper transcription factors analogs, not homologs

Given the complexity of the β-sandwich domain, however, it is not clear if a similar statistical argument applies. Complexity and organization of the domain argue for homology. The extent of sequence variability permitted within the β-sheets and the linking loops argue that β-sandwich domains are not difficult to construct by random chance, implying analogy.

Theobald et al. (2005) investigated a set of divergent protein domains having similar tertiary structures and less than 40% sequence identity. They found that an all-against-all, profile-versus-profile analysis of these domains revealed many previously undetectable significant interrelationships among the domains. Unfortunately, the 40% identity used therein is much higher than the 15% identity displayed among many of our sequences, so their conclusion may not apply. In fact, we note that the NTD80/PhoG, RHD, Runt, T-Box, STAT, and p53 families comprising the PFAM β-sandwich clan display almost no inter-family similarity, in contrast to other PFAM clans. Other studies have concluded that use of structure information to increase alignment accuracy does *not* aid homologue detection, at least when using profile HMMs (Griffiths-Jones and Bateman 2002). Thus little evidence supports the clan's multiple-alignment and implied homology.

We believe that further investigation of the origin of the β-sandwich domain would provide valuable insight into the origin of these transcription factors. Such investigation would shed light not only on the origin of p53, but many other transcription factors that serve critical cellular functions.

Acknowledgements

The authors wish to thank Charlie Smith, Steffen Heber, Eric Stone, and Bonnie Deroo for helpful comments and suggestions during manuscript preparations. Data analysis was done in large part with the R system (2006). Financial support was provided by the National Institutes of Health (GM45344), and the North Carolina State University.

Table 1

Protein sequences, database accessions, taxonomy, and sequence classification for the sequences used in this study. Database identifiers are RefSeq (REF), UniProt (UP), and GenBank (GB). Conceptual translations of the two echinoderm DNA sequences were concatenated based on inspection of the genomic sequences.

ID	Database	Accession	Classification	Organism	Taxonomy
XP_784255_URCHIN_1	REF	XP_784255	echinoderm	*Strongylocentrotus purpuratus*	*Echinodermata*
XP_784325_URCHIN_2	REF	XP_784325	echinoderm	*Strongylocentrotus purpuratus*	*Echinodermata*
Q868M8_ENTHI	UP	Q868M8	entamoeba	*Entamoeba histolytica*	*Entamoeba*
Q9N6D8_DROME	UP	Q9N6D8	insect	*Drosophila melanogaster*	*Arthropoda*
Q7QAB9_ANOGA	UP	Q7QAB9	insect	*Anopheles gambiae*	*Arthropoda*
BD250011_LEPTDE	GB	BD250011	insect	*Leptinotarsa decemlineata*	*Arthropoda*
BD250012_TRICA	GB	BD250012	insect	*Tribolium castaneum*	*Arthropoda*
Q9NGC7_MYAAR	UP	Q9NGC7	mollusc	*Mya arenaria*	*Mollusca*
Q9NGC8_MYAAR	UP	Q9NGC8	mollusc	*Mya arenaria*	*Mollusca*
Q6WG20_SPISO	UP	Q6WG20	mollusc	*Spisula solidissima*	*Mollusca*
Q53CG6_MYTED	UP	Q53CG6	mollusc	*Mytilus edulis*	*Mollusca*
Q539B9_MYTTR	UP	Q539B9	mollusc	*Mytilus trossulus*	*Mollusca*
Q27937_LOLFO	UP	Q27937	mollusc	*Loligo forbesi*	*Mollusca*
Q61X87_CAEBR	UP	Q61X87	nematode	*Caenorhabditis briggsae*	*Nematoda*
Q20646_CAEEL	UP	Q20646	nematode	*Caenorhabditis elegans*	*Nematoda*
P53_ONCMY	UP	P25035	p53	*Oncorhynchus mykiss*	*Chordata*
P53_ORYLA	UP	P79820	p53	*Oryzias latipes*	*Chordata*
P53_XIPHE	UP	O57538	p53	*Xiphophorus helleri*	*Chordata*
P53_XIPMA	UP	Q92143	p53	*Xiphophorus maculatus*	*Chordata*
P53_PLAFE	UP	O12946	p53	*Platichthys flesus*	*Chordata*
P53_TETMU	UP	Q9W679	p53	*Tetraodon miurus*	*Chordata*
P53_BARBU	UP	Q9W678	p53	*Barbus barbus*	*Chordata*
P53_BRARE	UP	P79734	p53	*Brachydanio rerio*	*Chordata*
P53_ICTPU	UP	O93379	p53	*Ictalurus punctatus*	*Chordata*
P53_XENLA	UP	P07193	p53	*Xenopus laevis*	*Chordata*
P53_CHICK	UP	P10360	p53	*Gallus gallus*	*Chordata*
P53_RABIT	UP	Q95330	p53	*Oryctolagus cuniculus*	*Chordata*
P53_CAVPO	UP	Q9WUR6	p53	*Cavia porcellus*	*Chordata*
P53_CRIGR	UP	O09185	p53	*Cricetulus griseus*	*Chordata*
P53_MESAU	UP	Q00366	p53	*Mesocricetus auratus*	*Chordata*
P53_MOUSE	UP	P02340	p53	*Mus musculus*	*Chordata*
P53_RAT	UP	P10361	p53	*Rattus norvegicus*	*Chordata*
P53_MARMO	UP	O36006	p53	*Marmota monax*	*Chordata*
P53_SPEBE	UP	Q64662	p53	*Spermophilus beecheyi*	*Chordata*
P53_CERAE	UP	P13481	p53	*Cercopithecus aethiops*	*Chordata*
P53_MACFA	UP	P56423	p53	*Macaca fascicularis*	*Chordata*
P53_MACFU	UP	P61260	p53	*Macaca fuscata fuscata*	*Chordata*
P53_MACMU	UP	P56424	p53	*Macaca mulatta*	*Chordata*
P53_HUMAN	UP	P04637	p53	*Homo sapiens*	*Chordata*
P53_TUPGB	UP	Q9TTA1	p53	*Tupaia glis belangeri*	*Chordata*
P53_CANFA	UP	Q29537	p53	*Canis familiaris*	*Chordata*
P53_FELCA	UP	P41685	p53	*Felis silvestris catus*	*Chordata*
P53_DELLE	UP	Q8SPZ3	p53	*Delphinapterus leucas*	*Chordata*
P53_BOSIN	UP	P67938	p53	*Bos indicus*	*Chordata*
P53_BOVIN	UP	P67939	p53	*Bos taurus*	*Chordata*
P53_SHEEP	UP	P51664	p53	*Ovis aries*	*Chordata*
P53_PIG	UP	Q9TUB2	p53	*Sus scrofa*	*Chordata*

(Table 1 – Continued)

ID	Database	Accession	Classification	Organism	Taxonomy
P53_EQUAS	UP	Q29480	p53	*Equus asinus*	*Chordata*
P53_HORSE	UP	P79892	p53	*Equus caballus*	*Chordata*
Q8JHZ6_BRARE	UP	Q8JHZ6	p63	*Brachydanio rerio*	*Chordata*
Q98SW0_XENLA	UP	Q98SW0	p63	*Xenopus laevis*	*Chordata*
Q9DEC7_CHICK	UP	Q9DEC7	p63	*Gallus gallus*	*Chordata*
P73L_MOUSE	UP	O88898	p63	*Mus musculus*	*Chordata*
P73L_RAT	UP	Q9JJP6	p63	*Rattus norvegicus*	*Chordata*
P73L_HUMAN	UP	Q9H3D4	p63	*Homo sapiens*	*Chordata*
Q9W664_BARBU	UP	Q9W664	p73	*Barbus barbus*	*Chordata*
Q9JJP2_MOUSE	UP	Q9JJP2	p73	*Mus musculus*	*Chordata*
XP_342993_RAT	REF	XP_342993	p73	*Rattus norvegicus*	*Chordata*
P73_CERAE	UP	Q9XSK8	p73	*Cercopithecus aethiops*	*Chordata*
P73_HUMAN	UP	O15350	p73	*Homo sapiens*	*Chordata*

Table 2

Conserved sites found within the p53 family. The smallest 10% of mean rates were selected giving a range of mean rates from 0.00692 to 0.198.

Site	Mean Rate	p53	p63	p73	Function (with respect to p53)
4	0.007	M1	-	-	
7	0.094	-	E4	-	
12	0.102	-	A9	-	
13	0.160	-	T10	-	
36	0.164	-	E33	-	
38	0.197	-	Y35	-	
52	0.080	-	-	-	
62	0.193	L22	I58	L18	transactivation
63	0.142	W23	W59	W19	
74	0.162	-	-	-	
75	0.080	-	-	-	
89	0.187	-	-	-	
134	0.087	-	-	-	
136	0.139	-	-	-	
141	0.189	-	-	-	
149	0.121	-	-	-	
152	0.138	-	-	-	
155	0.068	-	-	-	
156	0.067	-	-	-	
157	0.091	-	-	-	
159	0.157	-	-	-	
171	0.198	-	-	-	
225	0.162	-	-	-	
226	0.193	-	-	-	
230	0.137	-	-	-	
240	0.121	-	-	-	
271	0.106	-	-	-	
278	0.099	-	-	-	
289	0.185	-	-	-	
299	0.081	-	-	-	
302	0.109	-	-	-	
311	0.104	-	-	-	
319	0.122	-	-	-	
320	0.147	Y103	Y171	Y121	DNA binding
327	0.132	F109	F177	F127	
344	0.158	C124	W192	W142	
346	0.123	Y126	Y194	Y144	
350	0.091	S127	S195	S145	
355	0.111	K132	K200	K150	
360	0.142	L137	I205	I155	
376	0.138	P151	P219	P169	
377	0.129	P152	P220	P170	
384	0.048	R158	R226	R176	
386	0.193	M160	M228	M178	
390	0.189	Y163	Y231	Y181	
391	0.160	K164	K232	K182	
405	0.090	-	-	-	
409	0.056	R175	R243	R193	
410	0.097	C176	C244	C194	
412	0.125	H178	N246	N196	
413	0.020	H179	H247	H197	
431	0.065	H193	H263	H213	
434	0.104	R196	R266	R216	
438	0.093	N200	N270	N220	
443	0.117	Y205	Y275	Y225	
457	0.056	R213	R283	R233	

(Table 2 – Continued)

Site	Mean Rate	p53	p63	p73	Function (with respect to p53)
459	0.161	S215	S285	S235	
463	0.029	P219	P289	P239	
467	0.174	P223	P293	P243	
468	0.108	E224	Q294	Q244	
485	0.136	M237	M307	M257	
486	0.044	C238	C308	C258	
487	0.155	N239	N309	N259	
489	0.113	S241	S311	S261	
490	0.014	C242	C312	C262	
491	0.194	M243	V313	V263	
495	0.116	M246	M316	M266	
496	0.075	N247	N317	N267	
497	0.035	R248	R318	R268	
498	0.065	R249	R319	R269	
502	0.192	I251	I321	I271	
508	0.147	T256	T326	T276	
509	0.074	L257	L327	L277	
510	0.134	E258	E328	E278	
514	0.109	G262	G332	G282	
520	0.180	R267	R337	R287	
526	0.159	R273	R343	R293	
528	0.042	C275	C345	C295	
530	0.070	C277	C347	C297	
531	0.089	P278	P348	P298	
533	0.115	-	-	-	
536	0.194	-	-	-	
542	0.034	R280	R350	R300	
543	0.048	D281	D351	D301	
544	0.096	R282	R352	R302	
550	0.115	E286	E356	E306	
658	0.153	-	-	-	
660	0.114	-	-	-	
661	0.153	-	-	-	
681	0.174	-	-	-	
689	0.172	-	-	-	
716	0.175	-	Q444	-	
725	0.194	-	-	-	
796	0.149	-	-	-	
855	0.089	-	F552	F496	Sterile Alpha Motif and transactivation inhibition (p63 & p73)
856	0.106	-	L553	L497	
869	0.123	-	F565	F509	
903	0.134	-	W598	W542	
904	0.173	-	K599	R543	
922	0.119	-	-	-	
987	0.198	-	W658	W609	
992	0.060	-	F663	F614	

Table 3

Hypervariable sites found within the p53 family. The largest 10% of mean rates were selected giving a range of mean rates from 2.11 to 7.25.

Site	Mean Rate	p53	p63	p73	Function (with respect to p53)
4	2.58	-	F3	-	
7	2.89	-	C8	-	
12	2.77	P4	P17	-	transactivation
13	2.46	-	H28	-	
36	3.00	-	S30	-	
38	3.13	S9	Q45	A6	
52	2.51	L14	L50	D10	
62	2.18	K24	D60	S20	
63	2.37	L25	F61	S21	
74	2.28	-	-	-	
75	3.50	-	-	-	
89	2.53	-	F76	Y28	
134	3.63	-	-	-	
136	3.10	-	-	-	
141	2.64	-	-	-	
149	2.91	-	-	-	
152	4.95	-	-	-	
155	2.18	E51	T112	A63	
156	3.02	-	-	-	
157	2.41	-	-	-	
159	2.70	-	-	-	
171	2.75	-	A158	T108	
225	2.30	S96	A164	V114	
226	2.17	-	-	-	
230	2.16	-		-	
240	3.56	-	-	-	
271	2.37	-	-	-	
278	2.51	-	-	-	
289	3.70	-	-	-	
299	2.70	-	-	-	
302	2.61	-	-	-	
311	2.46	-	-	-	
319	2.95	-	-	-	
320	2.34	G108	S176	H126	DNA binding
327	4.43	-	-	-	
344	4.20	-	-	-	
346	4.32	A129	E197	L147	
350	4.95	A138	A206	A156	
355	4.30	D148	M216	S166	
360	2.76	-	-	-	
376	5.14	-	-	-	
377	2.68	S166	A234	A184	
384	2.97	-	-	-	
386	4.80	-	-	-	
390	2.23	C182	S250	G200	
391	2.26	L188	I258	S208	
405	6.07	-	-	-	
409	2.64	L201	S271	N221	
410	2.40	-	-	-	
412	4.38	F212	G282	G232	
413	3.83	V217	L287	V237	
431	2.90	P222	P292	P242	
434	3.12	-	-	-	
438	2.46	Y236	F306	F256	
443	3.23	S240	S310	S260	
457	7.25	-	-	-	
459	3.17	-	-	-	

(Table 3 – Continued)

Site	Mean Rate	p53	p63	p73	Function (with respect to p53)
463	3.38	-	-	-	
467	4.21	-	-	-	
468	2.71	S260	R330	R280	
485	4.82	-	-	-	
486	4.02	A276	A346	A296	
487	2.87	-	-	-	
489	2.43	-	-	-	
490	3.25	-	-	-	
491	3.58	-	-	-	
495	2.55	L289	I359	Y309	
496	2.59	-	S367	S319	
497	2.29	-	-	-	
498	2.33	G302	D372	A324	Oligomerization, interaction with CARM1
502	2.48	N311	N381	S333	
508	3.08	-	-	-	
509	2.52	-	-	-	
510	3.06	-	-	-	
514	3.18	-	-	-	
520	2.22	-	-	-	
526	2.52	L344	I417	L371	Oligomerization, interaction with HIPK2
528	3.27	-	-	-	
530	2.26	-	-	-	
531	2.84	-	-	-	
533	2.30	G360	E433	D387	region between oligomerization and basic (repression of DNA-binding)
536	2.19	G361	T434	S388	domains
542	2.42	-	Q441	-	
543	2.15	-	I453	L402	
544	2.42	-	S471	G419	
550	2.59	-	-	-	
658	2.16	-	P521	S471	
660	2.39	-	P542	P486	Sterile Alpha Motif (p63 & p73)
661	2.31	-	-	-	
681	2.37	-	R555	G499	
689	2.32	-	S559	P503	
716	2.49	-	Y574	Y518	
725	2.13	-	-	-	
796	2.11	-	D582	E526	
855	3.16	-	A596	T540	
856	2.35	-	H604	L548	
869	3.48	-	S612	T556	Transactivation inhibition (p63 & p73)
903	2.67	-	-	G575	
904	2.27	-	A641	A585	
922	3.78	-	N662	G613	
987	3.23	-	-	-	
992	2.23	P390	-	T631	

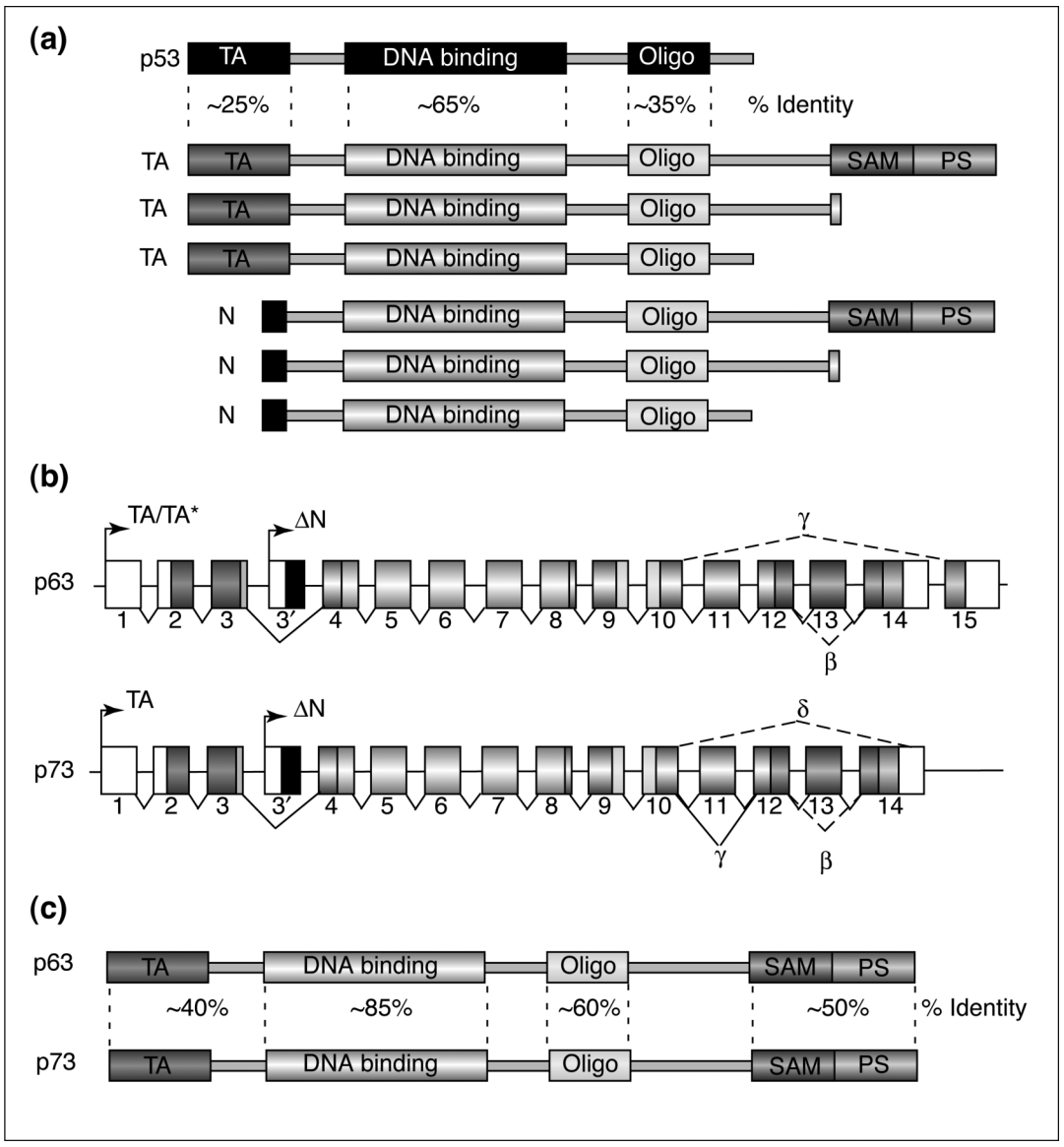

Figure 1

p53 family protein isoforms and genetic architecture, from Yang, et al. (2002) as described in the main text. (a) A comparison of p53 with the six major isoforms of each of p63 and p73. (b) Exon-intron arrangements of human p63 and p73, not drawn to scale. (c) Amino acid sequence identities between p63 and p73.

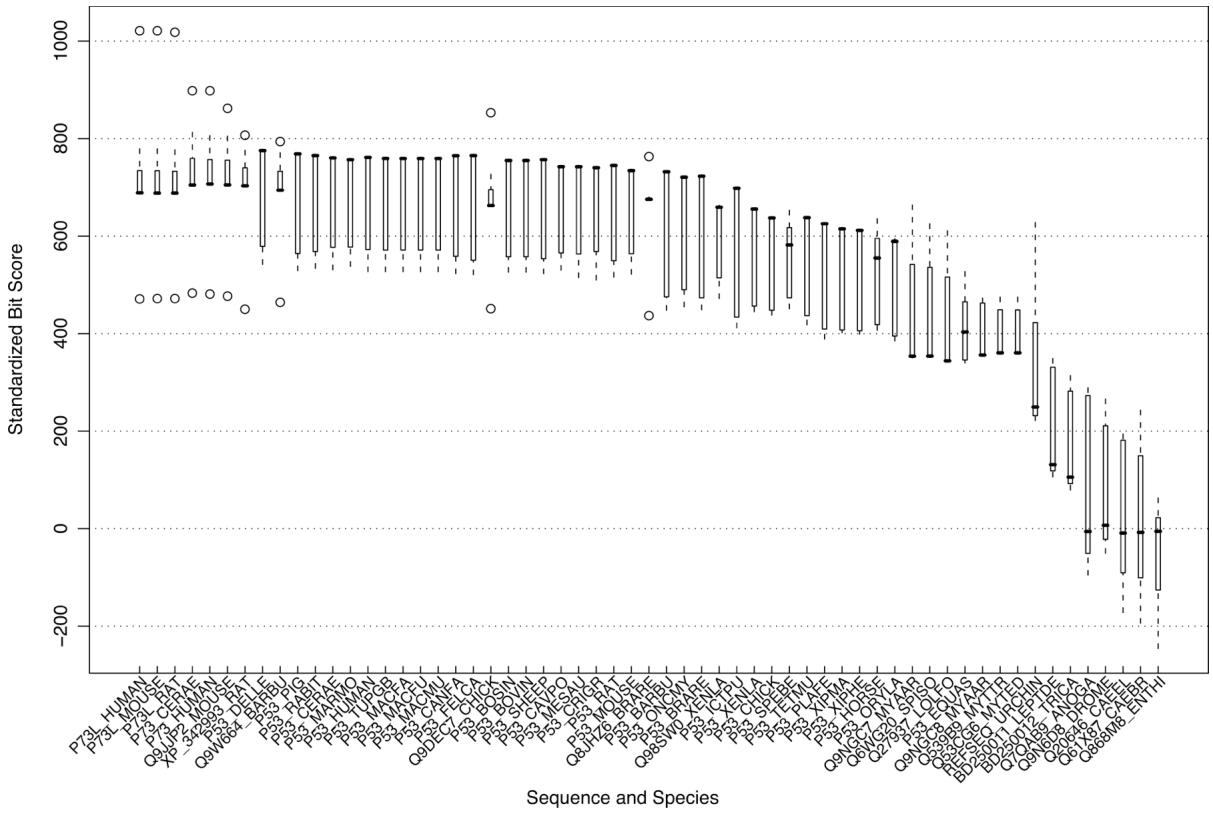

Figure 2

Significance of domain homology to p53 profile-based motifs, given by standardized bit-score. Details are given in the text. Organisms with the lowest bit-score are, from right to left: *E. histolytica*, *C. briggsae*, *C. elegans*, *D. melanogaster*, *A. gambiae*, *T. castaneum*, *L. decemlineata*, *S. purpuratus*, followed by mollusc species. The intervening P53_EQUAS sequence is a fragment from *Equus asinus* (donkey).

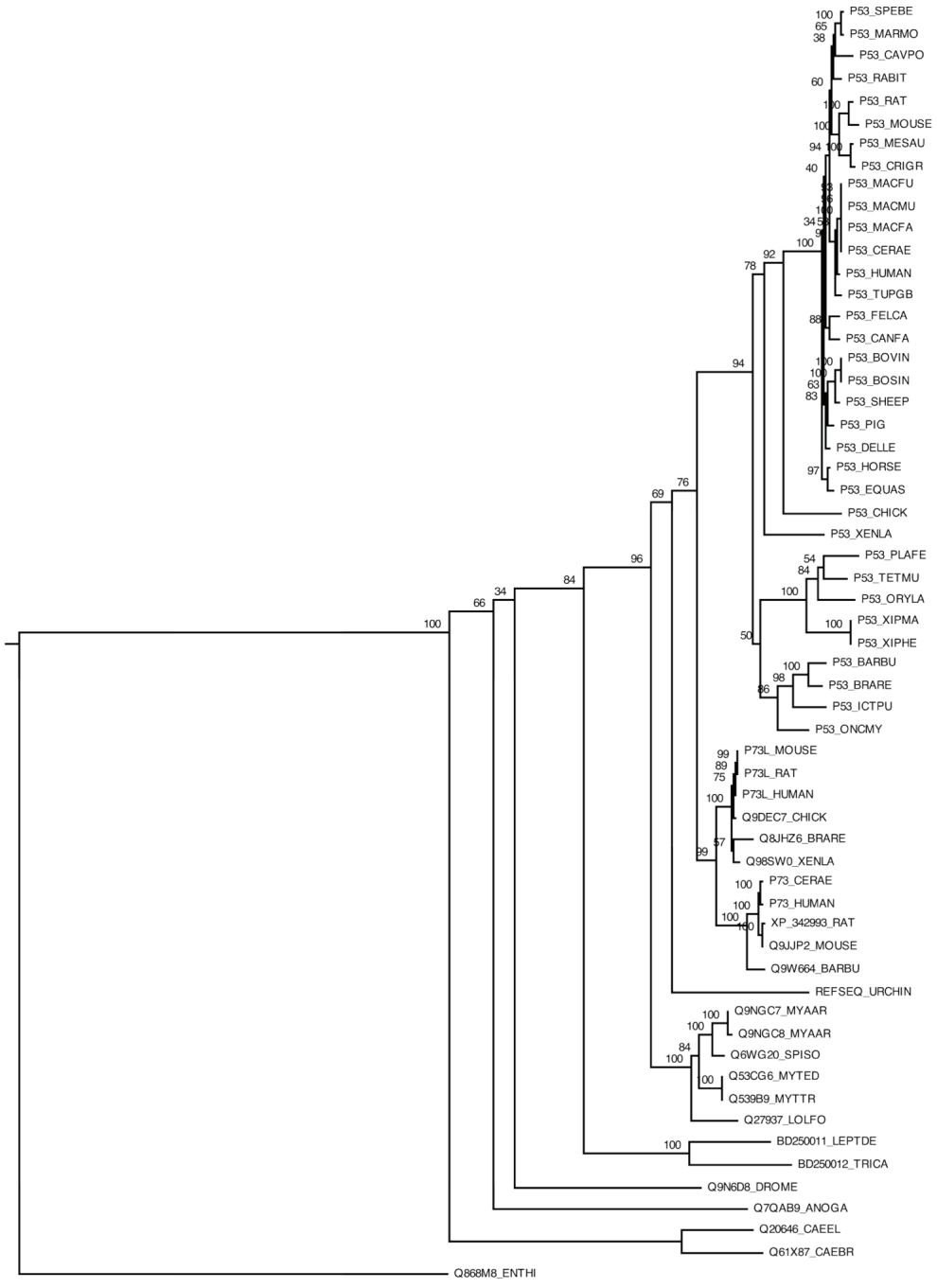

Figure 3

The inferred phylogenetic tree for the p53 family, showing bootstrap support values. The figure is discussed extensively in the text.

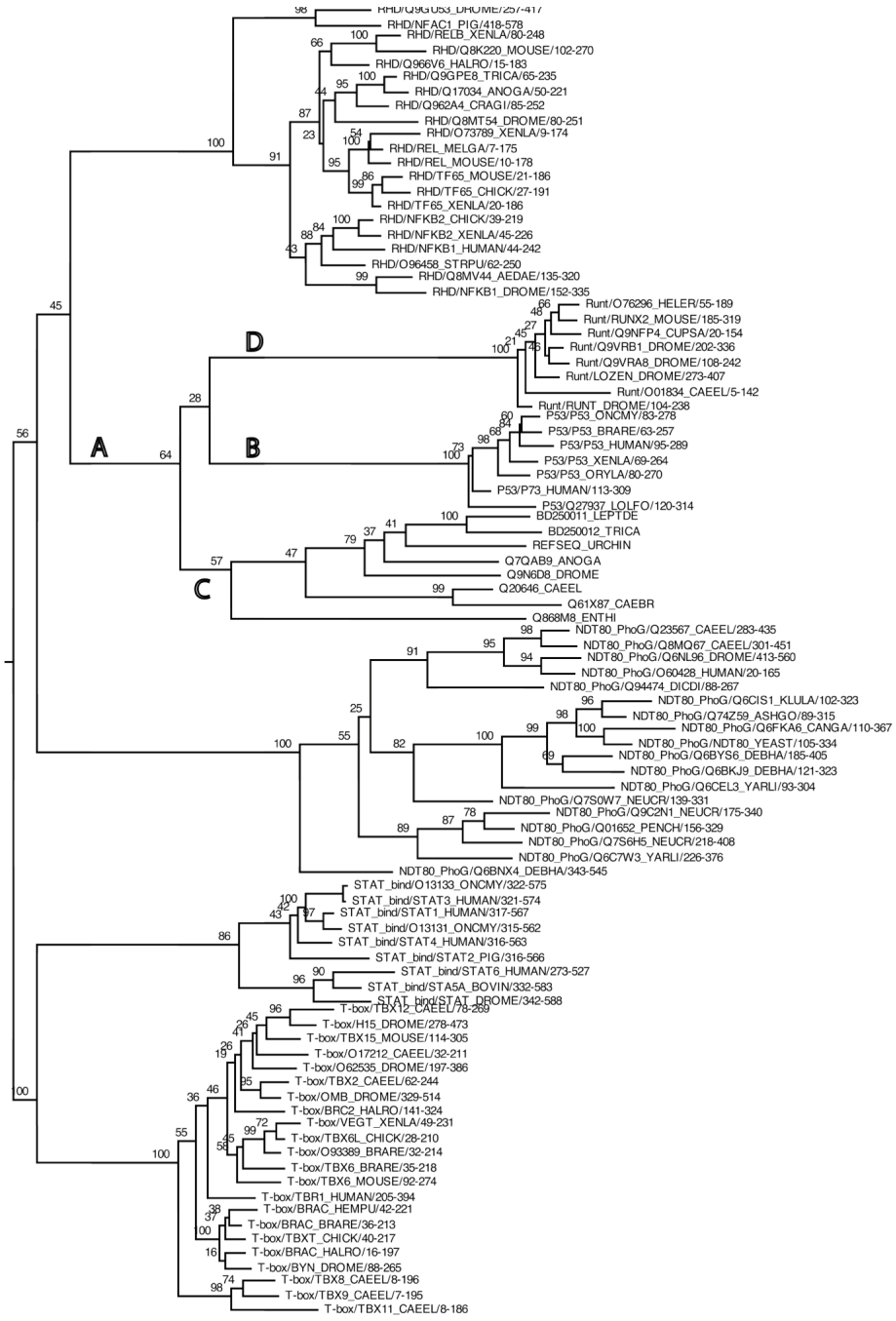

Figure 4

The inferred phylogenetic tree for the β-sandwich clan, with bootstrap values. Events: (A) main p53 super-family branch, (B) the known p53 family, (C) non-chordate and non-mollusc branch, (D) the Runt family.

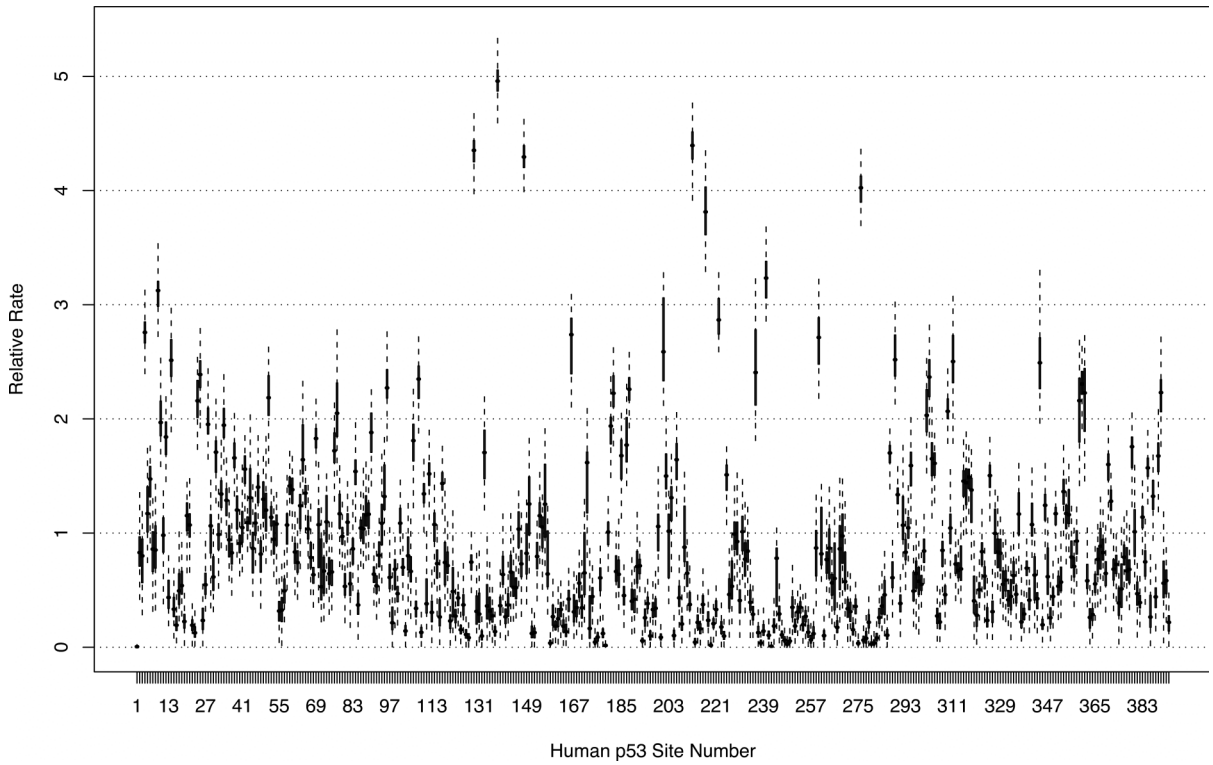

Figure 5

Site-specific relative rates of evolution with respect to human p53.

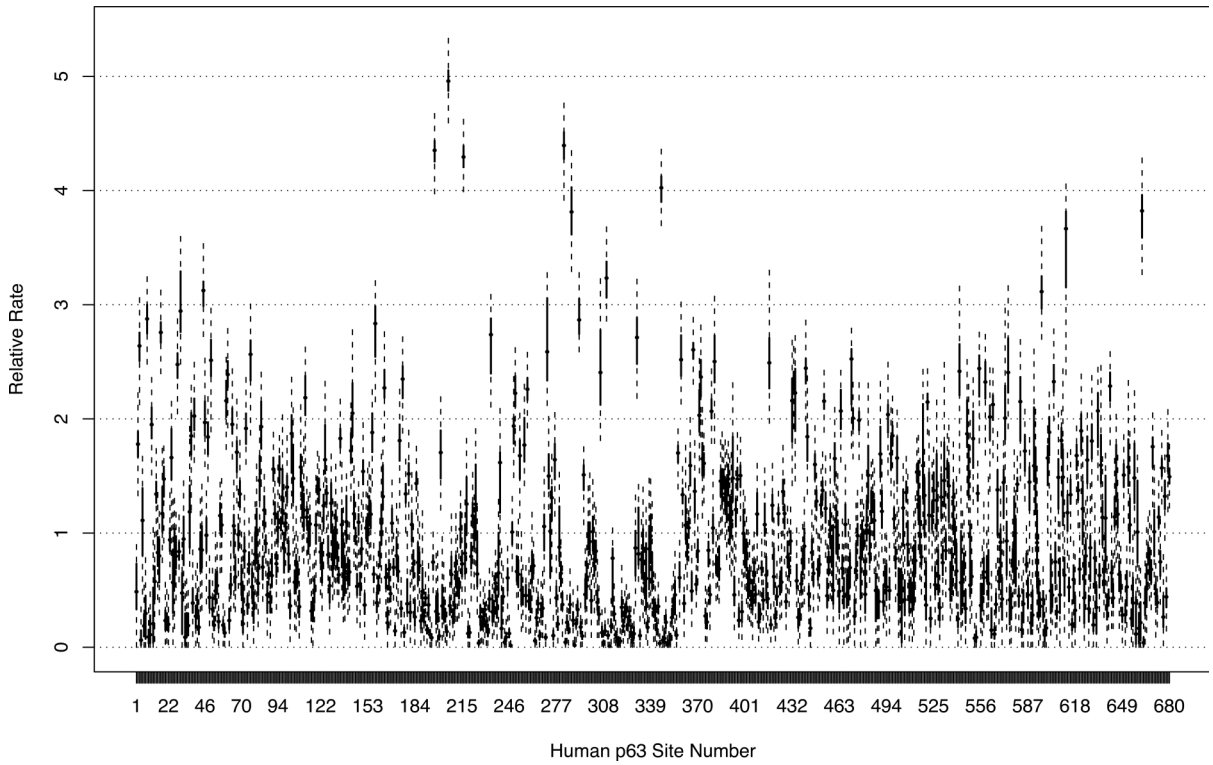

Figure 6

Site-specific relative rates of evolution with respect to human p63.

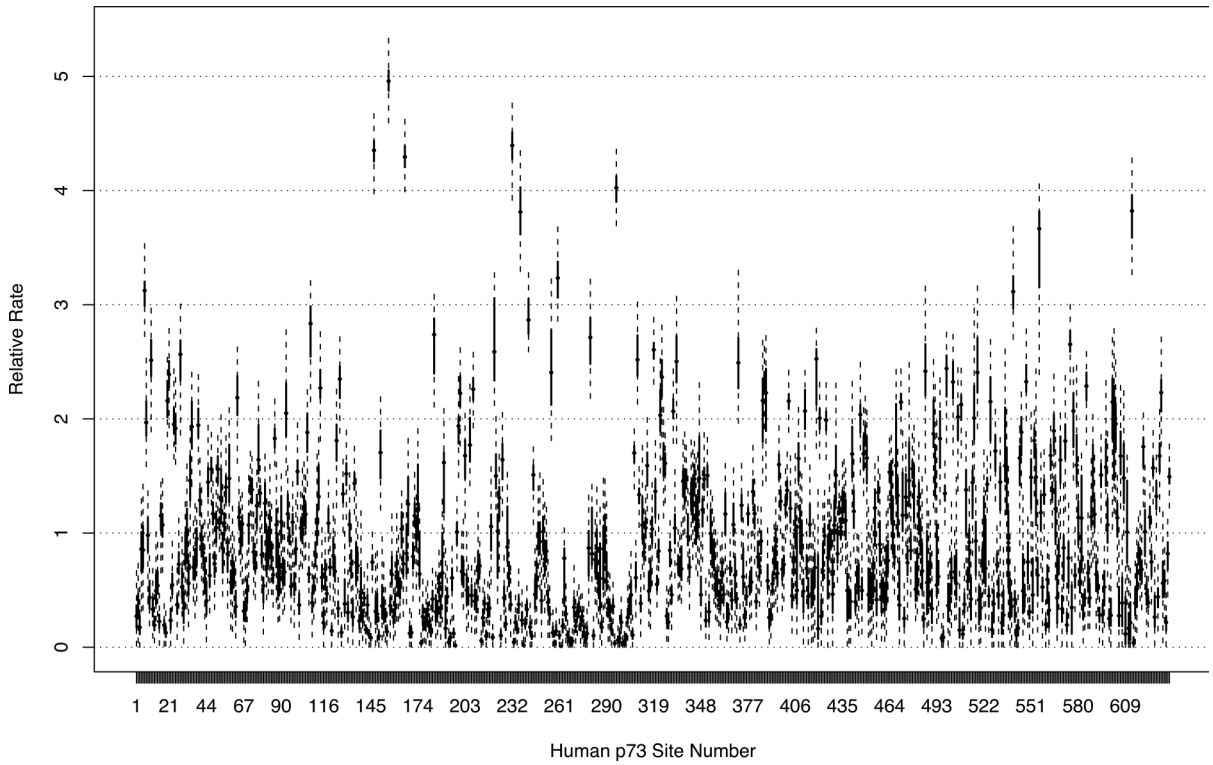

Figure 7

Site-specific relative rates of evolution with respect to human p73.

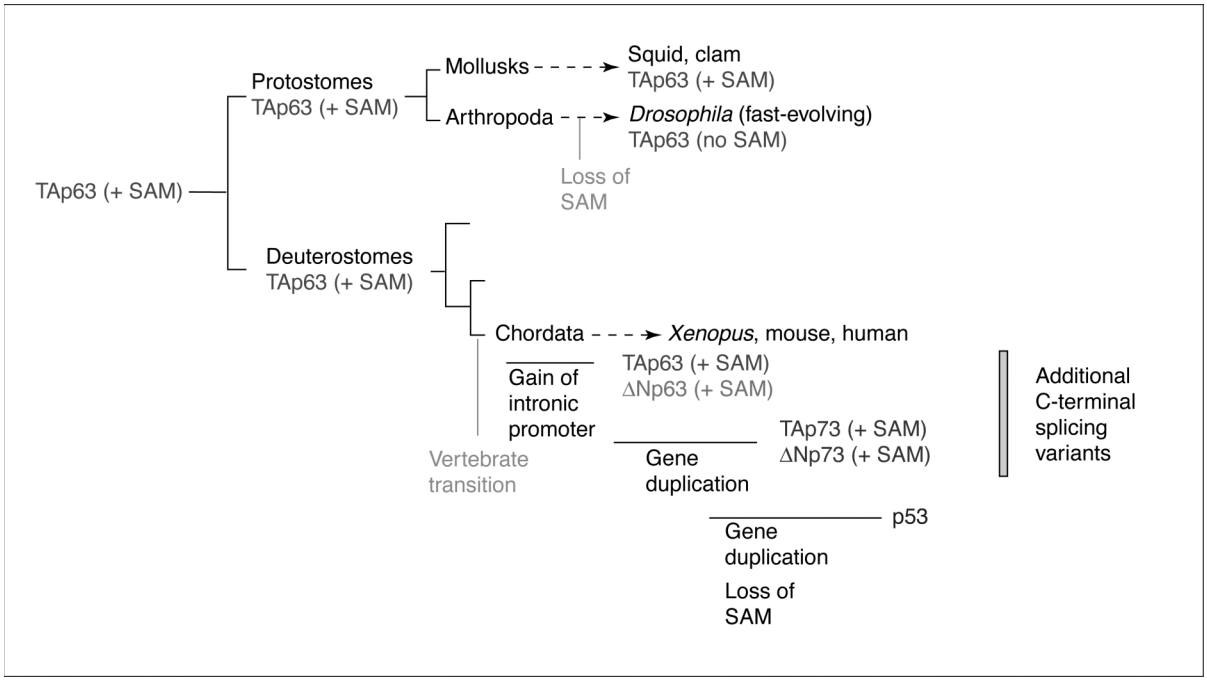

Figure 8

A hypothetical evolutionary history of the p53 family, from Yang, et al. (2002).

References

Almog, N., N. Goldfinger, et al. (2000). "p53-dependent apoptosis is regulated by a C-terminally alternatively spliced form of murine p53." Oncogene **19**(30): 3395-3403.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-10.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research **25**(17): 3389-402.

Atchley, W. R. and A. D. Fernandes (2005). "Sequence signatures and the probabilistic identification of proteins in the Myc-Max-Mad network." Proceedings of the National Academy of Sciences of the United States of America **102**(18): 6401-6406.

Atchley, W. R., W. Terhalle, et al. (1999). "Positional dependence, cliques, and predictive motifs in the bHLH protein domain." Journal of Molecular Evolution **48**(5): 501-516.

Atchley, W. R. and J. Zhao (2006). "Molecular Architecture of the DNA Binding Region and its Relationship to Classification of Basic Helix-Loop-Helix Proteins." Molecular Biology and Evolution **In Press**.

Ayala, F. J., A. Rzhetsky, et al. (1998). "Origin of the metazoan phyla: molecular clocks confirm paleontological estimates." Proceedings of the National Academy of Sciences of the United States of America **95**(2): 606-11.

Bairoch, A., R. Apweiler, et al. (2005). "The Universal Protein Resource (UniProt)." Nucleic Acids Research **33**(Database Issue): D154-9.

Benson, D. A., I. Karsch-Mizrachi, et al. (2006). "GenBank." Nucleic Acids Research **34**(Database Issue): D16-20.

Berardi, M. J., C. Sun, et al. (1999). "The Ig fold of the core binding factor α Runt domain is a member of a family of structurally and functionally related Ig-fold DNA-binding domains." Structure **7**(10): 1247-1256.

Bergamaschi, D., Y. Samuels, et al. (2003). "iASPP oncoprotein is a key inhibitor of p53 conserved from worm to human." Nature Genetics **33**(2): 162-167.

Bolshakov, V. N., P. Topalis, et al. (2002). "A comparative genomic analysis of two distant diptera, the fruit fly, Drosophila melanogaster, and the malaria mosquito, Anopheles gambiae." Genome Research **12**(1): 57-66.

Bourdon, J.-C., K. Fernandes, et al. (2005). "p53 isoforms can regulate p53 transcriptional activity." Genes and Development **19**(18): 2122-2137.

Brendel, V. and S. Karlin (1989). "Too many leucine zippers?" Nature **341**(6243): 574-5.

Casci, T. (2004). "p53, a protein with a pulse." Nature Reviews Genetics **5**(3): 162-163.

Cho, Y., S. Gorina, et al. (1994). "Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations." Science **265**(5170): 346-355.

Ciotta, C., E. Dogliotti, et al. (1995). "Mutation analysis in two newly identified rat p53 pseudogenes." Mutagenesis **10**(2): 123-128.

Coghlan, A. (2005). "Nematode genome evolution." WormBook, from http://www.wormbook.org.

Courtois, S., C. C. de Fromentel, et al. (2004). "p53 protein variants: structural and functional similarities with p63 and p73 isoforms." Oncogene **23**(3): 631-638.

Cox, R. L., R. E. Stephens, et al. (2003). "p63/73 homologues in surf clam: novel signaling motifs and implications for control of expression." Gene **320**: 49-58.

Derry, W. B., A. P. Putzke, et al. (2001). "Caenorhabditis elegans p53: role in apoptosis, meiosis, and stress resistance." Science **294**(5542): 591-5.

Donehower, L. A., M. Harvey, et al. (1992). "Mice Deficient for p53 Are Developmentally Normal but Susceptible to Spontaneous Tumors." Nature **356**(6366): 215-221.

Durbin, R. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, UK New York, Cambridge University Press.

El-Deiry, W. S., S. E. Kern, et al. (1992). "Definition of a consensus binding site for p53." Nature Genetics **1**(1): 45-49.

Fernandes, A. D. and W. R. Atchley (2006). Detecting Conserved Motifs using Site-Specific Rate Inference with Objective Non-informative Priors, North Carolina State University.

Finn, R. D., J. Mistry, et al. (2006). "Pfam: clans, web tools and services." Nucleic Acids Research **34**(S1): D247-251.

Flaman, J. M., F. Waridel, et al. (1996). "The human tumour suppressor gene p53 is alternatively spliced in normal cells." Oncogene **12**(4): 813-818.

Funk, W. D., D. T. Pak, et al. (1992). "A transcriptionally active DNA-binding site for human p53 protein complexes." Molecular and Cellular Biology **12**(6): 2866-2871.

Glazko, G. V., E. V. Koonin, et al. (2004). "Mutation hotspots in the p53 gene in tumors of different origin: correlation with evolutionary conservation and signs of positive selection." Biochimica et Biophysica Acta **1679**(2): 95-106.

Griffiths-Jones, S. and A. Bateman (2002). "The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs." Bioinformatics **18**(9): 1243-9.

Guindon, S. and O. Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Systematic Biology **52**(5): 696-704.

Gupta, B. P. and P. W. Sternberg (2003). "The draft genome sequence of the nematode Caenorhabditis briggsae, a companion to C. elegans." Genome Biology **4**(12): 238.

Hedges, S. B., J. E. Blair, et al. (2004). "A molecular timescale of eukaryote evolution and the rise of complex multicellular life." BMC Evolutionary Biology **4**: 2.

Hofseth, L. J., S. P. Hussain, et al. (2004). "p53: 25 years after its discovery." Trends in Pharmacological Sciences **25**(4): 177-181.

Hollstein, M., D. Sidransky, et al. (1991). "p53 Mutations In Human Cancers." Science **253**(5015): 49-53.

Hulla, J. E. (1992). "The rat genome contains a p53 pseudogene: detection of a processed pseudogene using PCR." PCR Methods and Applications **1**(4): 251-254.

Huyen, Y., P. D. Jeffrey, et al. (2004). "Structural Differences in the DNA Binding Domains of Human p53 and Its C. elegans Ortholog Cep-1." Structure **12**(7): 1237-1243.

Ichimiya, S., A. Nakagawara, et al. (2000). "p73: Structure and function." Pathology International **50**(8): 589-593.

Jin, S., S. Martinek, et al. (2000). "Identification and characterization of a p53 homologue in Drosophila melanogaster." Proceedings of the National Academy of Sciences of the United States of America **97**(13): 7301-6.

Kelley, M. L., P. Winge, et al. (2001). "Expression of homologues for p53 and p73 in the softshell clam (*Mya arenaria)*, a naturally-occurring model for human cancer." Oncogene **20**(6): 748-58.

Keyes, W. M. and A. A. Mills (2006). "p63: a new link between senescence and aging." Cell Cycle **5**(3): 260-5.

Klug, S. J. and M. Famulok (1994). "All You Wanted to Know About SELEX." Molecular Biology Reports **20**(2): 97-107.

Knoll, A. H. (1992). "The early evolution of eukaryotes: a geological perspective." Science **256**(5057): 622-7.

Lahav, G., N. Rosenfeld, et al. (2004). "Dynamics of the p53-Mdm2 feedback loop in individual cells." Nature Genetics **36**(2): 147-150.

Laverdiere, M., J. Beaudoin, et al. (2000). "Species-specific regulation of alternative splicing in the C-terminal region of the p53 tumor suppressor gene." Nucleic Acids Research **28**(6): 1489-97.

Lesser, M. P., J. H. Farrell, et al. (2001). "Oxidative stress, DNA damage and p53 expression in the larvae of atlantic cod (*Gadus morhua*) exposed to ultraviolet (290-400 nm) radiation." Journal of Experimental Biology **204**(Pt 1): 157-64.

Lesser, M. P., V. A. Kruse, et al. (2003). "Exposure to ultraviolet radiation causes apoptosis in developing sea urchin embryos." Journal of Experimental Biology **206**(Pt 22): 4097-103.

Levine, A. J., J. Momand, et al. (1991). "The p53 Tumor Suppressor Gene." Nature **351**(6326): 453-456.

Levrero, M., V. De Laurenzi, et al. (1999). "Structure, function and regulation of p63 and p73." Cell Death and Differentiation **6**(12): 1146-1153.

Lynch, M. and J. S. Conery (2000). "The Evolutionary Fate and Consequences of Duplicate Genes." Science **290**(5494): 1151-1155.

Matoba, S., J.-G. Kang, et al. (2006). "p53 Regulates Mitochondrial Respiration." Science **312**(5780): 1650-1653.

Mendoza, L., E. Orozco, et al. (2003). "Ehp53, an *Entamoeba histolytica* protein, ancestor of the mammalian tumour suppressor p53." Microbiology **149**(Pt 4): 885-93.

Mills, A. A. (2006). "p63: oncogene or tumor suppressor?" Current Opinion in Genetics and Development **16**(1): 38-44.

Moll, U. M., S. Erster, et al. (2001). "p53, p63 and p73 - solos, alliances and feuds among family members." Biochimica et Biophysica Acta **1552**(2): 47-59.

Muttray, A. F., R. L. Cox, et al. (2005). "Identification and phylogenetic comparison of p53 in two distinct mussel species (Mytilus)." Comparative Biochemistry and Physiology C **140**(2): 237-50.

Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." Journal of Molecular Biology **302**(1): 205-17.

Ohno, S. (1970). Evolution by gene duplication. New York, Springer-Verlag.

Ollmann, M., L. M. Young, et al. (2000). "Drosophila p53 is a structural and functional homolog of the tumor suppressor p53." Cell **101**(1): 91-101.

Ortt, K. and S. Sinha (2006). "Derivation of the consensus DNA-binding sequence for p63 reveals unique requirements that are distinct from p53." FEBS Letters **580**(18): 4544-4550.

Peterson, K. J. and N. J. Butterfield (2005). "Origin of the Eumetazoa: testing ecological predictions of molecular clocks against the Proterozoic fossil record." Proceedings of the National Academy of Sciences of the United States of America **102**(27): 9547-52.

R Development Core Team (2006). R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing, http://www.R-project.org.

Redelings, B. D. and M. A. Suchard (2005). "Joint Bayesian estimation of alignment and phylogeny." Systematic Biology **54**(3): 401-18.

Rudolph, M. J. and J. P. Gergen (2001). "DNA-binding by Ig-fold proteins." Nature Structural and Molecular Biology **8**(5): 384-386.

Schumacher, B., M. Hanazawa, et al. (2005). "Translational repression of C-elegans p53 by GLD-1 regulates DNA damage-induced apoptosis." Cell **120**(3): 357-368.

Schumacher, B., K. Hofmann, et al. (2001). "The *C. elegans* homolog of the p53 tumor suppressor is required for DNA damage-induced apoptosis." Current Biology **11**(21): 1722-1727.

Soussi, T., C. C. Defromentel, et al. (1990). "Structural Aspects of the p53 Protein In Relation To Gene Evolution." Oncogene **5**(7): 945-952.

Soussi, T. and P. May (1996). "Structural Aspects of the p53 Protein In Relation To Gene Evolution: A Second Look." Journal of Molecular Biology **260**(5): 623-637.

Steenkamp, E. T., J. Wright, et al. (2006). "The protistan origins of animals and fungi." Molecular Biology and Evolution **23**(1): 93-106.

Strano, S., M. Rossi, et al. (2001). "From p63 to p53 across p73." FEBS Letters **490**(3): 163-170.

Subramanian, A. R., J. Weyer-Menkhoff, et al. (2005). "DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment." BMC Bioinformatics **6**: 66.

Suchard, M. A. and B. D. Redelings (2006). "BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny." Bioinformatics **22**(16): 2047-8.

Theobald, D. L. and D. S. Wuttke (2005). "Divergent evolution within protein superfolds inferred from profile-based phylogenetics." Journal of Molecular Biology **354**(3): 722-37.

Thompson, J. D., P. Koehl, et al. (2005). "BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark." Proteins **61**(1): 127-36.

Tummala, R., R.-A. Romano, et al. (2003). "Molecular cloning and characterization of AP-2ε, a fifth member of the AP-2 family." Gene **321**: 93-102.

Walker, D. R., J. P. Bond, et al. (1999). "Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features." Oncogene **18**(1): 211-218.

Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Weaver, R. F. (2005). Molecular biology. Boston, McGraw-Hill.

Westfall, M. D. and J. A. Pietenpol (2004). "p63: molecular complexity in development and cancer." Carcinogenesis **25**(6): 857-864.

Whelan, S. and N. Goldman (2001). "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach." Molecular Biology and Evolution **18**(5): 691-9.

Winge, P., S. Friend, et al. (1996). "A p53 tumor suppressor homolog from *Loligo forbesi*." from http://www.uniprot.org/entry/Q27937_LOLFO.

Yang, A., M. Kaghad, et al. (2002). "On the shoulders of giants: p63, p73 and the rise of p53." Trends in Genetics **18**(2): 90-95.

Yang, A., M. Kaghad, et al. (1998). "p63, a p53 homolog at 3q27-29, encodes multiple products with transactivating, death-inducing, and dominant-negative activities." Molecular Cell **2**(3): 305-316.

Yang, A., R. Schweitzer, et al. (1999). "p63 is essential for regenerative proliferation in limb, craniofacial and epithelial development." Nature **398**(6729): 714-718.

Yang, A., N. Walker, et al. (2000). "p73-deficient mice have neurological, pheromonal and inflammatory defects but lack spontaneous tumours." Nature **404**(6773): 99-103.

Yeates, D. K. and B. M. Wiegmann (1999). "Congruence and controversy: toward a higher-level phylogeny of Diptera." Annual Review of Entomology **44**: 397-428.

Summary

Chapter 1 of this dissertation addressed the problem of how to efficiently evaluate arbitrary likelihood integrals through the use of Gaussian quadrature integration schemes. Such integrals are common in molecular evolution studies, and accelerating their evaluation allows studies of larger data sets than were practical before. I produced cross-platform computational libraries, the source code of which is freely available online, as a service to the research community.

Chapter 2 described a method of modeling protein evolution that allows site-specific rates to be identified independently of evolutionary time. The method is important for several reasons. First, specification of a prior rate distribution is not required. Since no rate prior is needed, there is no question of which prior to use, and no requirement to choose among different ones.

Second, since the method computationally separates rates from time, allowing each to be identified independently. In previous work, rate and time were not computationally identifiable. Without identifiability, simultaneous inference of rates and evolutionary divergence times cannot occur within a logically consistent framework. Although divergence times were treated as known and fixed in Chapter 2, the method can form the basis on which branch lengths and site-specific rates can be inferred simultaneously in a mathematically rigorous way.

Third, since the method does not assume an *a priori* distribution of rates, it is appears suitable for detecting both conserved and hypervariable sites. By assuming a prior rate distribution, previous work necessarily assumed that evolutionary large rates (much greater than one) are very unlikely. Since these large rates are indicative of hypervariable sites, prior rate distributions have a tendency to "discredit" evidence of significantly greater than average site variability. However, little investigation into the biological meaning of hypervariability has been done, and this is a lucrative area for future research.

In the same vein, using site-specific rates to detect conserved sites is not as straightforward as finding sites with low rates. To see why, consider that we are given that

the mean evolutionary rate is one. For each site, the posterior distribution of rates is inferred. Which sites should be denoted as conserved: sites with expected rate almost surely less than one, or some fraction of the slowest-evolving sites? The former answer implies that a site evolving at an estimated rate of 0.99 ± 0.0001 should be considered conserved, even though biologically speaking, it is evolving practically at the mean rate of all sites (one). The latter answer forces us to select a significance cut-off, implying that only five, ten, or other percentage of a protein's sites can ever be considered conserved, *a priori*. Both interpretations will likely need to be considered together. Furthermore, the correct use of rates to discover conserved sites will likely depend on the biological questions being asked.

Chapter 3 is interesting because, although a search query of "p53" in the PubMed journal database of the NCBI currently returns 40282 references, a comprehensive study of the molecular evolution of the p53 family has not been published. Most researchers seem to believe that p53 is so strongly conserved that little is gained from its phylogenetic study.

It was therefore surprising to discover the broad range of conflicting statistical and biological evidence classifying highly dissimilar sequences as p53 homologs. When biological, phylogenetic, and statistical evidence was combined, we discovered several unexplained inconsistencies in the data. These inconsistencies lead to the conclusion that some of the proteins reported in the literature may not be true p53 homologs, as purported.

Lastly, important questions regarding the rise of p53 involve its relationship to the β-sandwich family of transcription factors. Indeed, it is not clear even if there *is* such a family, as preliminary computational work provides little evidence for homology among family member groups. If sufficient evidence can be found supporting homology, it would provide an important foundation on which to study the rise of p53 during evolution.