BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i126–i135
doi:10.1093/bioinformatics/bti1032

Detecting coevolving amino acid sites using Bayesian mutational mapping

Matthew W. Dimmic[1,*], Melissa J. Hubisz[1],
Carlos D. Bustamante[1] and Rasmus Nielsen[1,2]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 13101, USA and [2]Center for Bioinformatics, University of Copenhagen, Copenhagen 2100 Kbh Ø, Denmark

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: The evolution of protein sequences is constrained by complex interactions between amino acid residues. Because harmful substitutions may be compensated for by other substitutions at neighboring sites, residues can coevolve. We describe a Bayesian phylogenetic approach to the detection of coevolving residues in protein families. This method, Bayesian mutational mapping (BMM), assigns mutations to the branches of the evolutionary tree stochastically, and then test statistics are calculated to determine whether a coevolutionary signal exists in the mapping. Posterior predictive P-values provide an estimate of significance, and specificity is maintained by integrating over uncertainty in the estimation of the tree topology, branch lengths and substitution rates. A coevolutionary Markov model for codon substitution is also described, and this model is used as the basis of several test statistics.

Results: Results on simulated coevolutionary data indicate that the BMM method can successfully detect nearly all coevolving sites when the model has been correctly specified, and that non-parametric statistics such as mutual information are generally less powerful than parametric statistics. On a dataset of eukaryotic proteins from the phosphoglycerate kinase (PGK) family, interdomain site contacts yield a significantly greater coevolutionary signal than interdomain non-contacts, an indication that the method provides information about interacting sites. Failure to account for the heterogeneity in rates across sites in PGK resulted in a less discriminating test, yielding a marked increase in the number of reported positives at both contact and non-contact sites.

Contact: matt@dimmic.net

Supplementary information: http://www.dimmic.net/supplement/

INTRODUCTION

The primary unit of phenotypic expression for a protein-coding gene is the amino acid site, and each site's evolution is constrained by a myriad of factors which contribute to the protein's function. For example, the residue at the site must pack correctly against other residues in the folded protein, it may catalyze a reaction in the active site, and it may be involved in binding or recognition of amino acid sites on other proteins. Because each amino acid's constraints are dependent on interactions with other residues, a mutation at nearby sites can change these constraints. In analogy with classical genetics, if each site is considered to be a single locus with 20 possible alleles, and the fitness of each amino acid 'allele' depends on the amino acids with which it interacts, then a substitution will alter the fitness landscape at the interacting sites. Changes to this landscape can in turn change the rate of evolution at the affected sites (Fitch and Markowitz, 1970), leading to concerted evolution and correlated substitutions.

Detection of coevolving sites has the potential to greatly aid fields such as protein threading, structure recognition and binding site detection; and there is a keen interest in developing methods for correlated mutational analysis (Gobel *et al.*, 1994; Shindyalov *et al.*, 1994; Pazos *et al.*, 1997; Pollock *et al.*, 1999; Atchley *et al.*, 2000; Pritchard *et al.*, 2001; Hamilton *et al.*, 2004). One general conclusion of these studies is that, although some coevolution does occur among neighboring residues, the signal from extant sequences is weak. One plausible explanation is that the detection methods are powerful enough but that coevolution is rare: many mutations are likely to be too deleterious, and thus there will be no opportunity for a further compensatory mutation (Govindarajan *et al.*, 2003). Proteins also have other compensatory mechanisms which would confound simple pairwise relationships; for example, a mutation at a variety of distant sites in proteins could subtly shift whole secondary structures, relieving the steric strain of an unfavorable mutation (Pollock *et al.*, 1999). Even if this is the case, where coevolution does occur it is likely that current methods have not yet reached the limit of detection. Previous studies have found that alignment-based methods can be biased because they do not account for spurious correlations due to the evolutionary history of the sequences (Pollock and Taylor, 1997; Atchley *et al.*, 2000; Tillier and Lui, 2003). The

[*]To whom correspondence should be addressed at Divergence, Inc., St. Louis, MO 63141, USA.

false positive (FP) rate can be reduced by explicitly accounting for the phylogenetic tree topology (Pollock *et al*., 1999; Fukami-Kobayashi *et al*., 2002) and power can be increased by modeling 'nuisance parameters' such as branch length, tree topology, and evolutionary rate (Tufféry and Darlu, 2000).

Bayesian phylogenetic methods can deal effectively with these types of concerns by integrating over nuisance parameters to focus on parameters of interest (Huelsenbeck *et al*., 2001). Such methods have been applied to a variety of evolutionary problems, for example to detect sites undergoing adaptive evolution (Huelsenbeck and Dyer, 2004), to infer particular branches of the evolutionary tree where adaptive evolution may have occurred (Guindon *et al*., 2004), and to determine the rooting of evolutionary trees (Huelsenbeck *et al*., 2002).

Here we apply the method of Bayesian mutational mapping (BMM) to the detection of coevolving sites in proteins, via a novel parametric Markov model to describe coevolving site pairs. In spirit, BMM resembles a method described by Fukami-Kobayashi *et al*. (2002), where the mutations at a pair of sites are mapped onto the branches of the tree, and coevolution is inferred when the mutations at the site pair tend to co-occur in evolutionary time. However, BMM differs from this method (and others like it) in several important respects. First, the mutational maps are informed by a model of evolution, allowing explicit assumptions to be tested using well-developed likelihood hypothesis testing techniques. Second, BMM does not require specification of a single tree topology or a small set of mutational maps, but instead uses Markov Chain Monte Carlo (MCMC) integration to account for uncertainty in the phylogeny and branch lengths, as well as variance in the estimates of the mutation times, model parameters, and evolutionary rates. In this paper a model of coevolution is developed, and the performance of the model is compared to other test statistics on simulated datasets and on the phosphoglycerate kinase (PGK) protein family.

METHODS

The application of BMM to the detection of correlated substitutions is motivated by this coevolutionary hypothesis: if two amino acid residues interact, their evolutionary fitness landscapes will depend on the amino acid at the interacting site. Therefore, a substitution at one site will potentially affect the rate of substitution at the other site, and mutations at the sites will tend to cluster together in evolutionary history. By mapping the probable pattern of mutations onto the evolutionary tree, we seek to detect the sites where these clusters have occurred more often than they would by chance. Such a technique requires:

(1) A method for mapping the mutations onto the evolutionary tree of the protein family (in this case, BMM),

(2) Test statistics to identify site pairs where the mutations support the coevolutionary hypothesis and

(3) A method for assessing the significance of each test statistic relative to the null hypothesis of no coevolution.

Bayesian Mutational Mapping (BMM)

The posterior distribution of mutational mappings is sampled on the coding sequences of the protein, using the method described in Nielsen (2002) and Huelsenbeck *et al*. (2003), with an addition to utilize rates across sites. Briefly, given an alignment of protein-coding nucleotide sequences and a set of paired columns A, B in the alignment to compare:

(1) Using the MCMC technique, draw N_{gen} samples G_n from the posterior distribution of trees and branch lengths $\{T, l\}$, nucleotide model parameters $\{R, \pi\}$, and rate parameter α, so that $G_n = \{T, l, R, \pi, \alpha\}_n$. The rate parameter controls the shape of the Γ-distribution, which is discretized into N_r possible rate categories (Yang, 1994). This step is performed using the program MrBayes (Ronquist and Huelsenbeck, 2003).

(2) For each iteration n, where the posterior sample is G_n:

(a) Sample a site-specific rate for each site from the posterior distribution of rates across sites. In each codon site s, each nucleotide position s_x is assigned to a rate category $r_n^{(sz)}$ stochastically based on the posterior probability of category r at that site. This yields a vector of rates at each codon, $r_n^{(s)} = \left\{ r_n^{(s1)}, r_n^{(s2)}, r_n^{(s3)} \right\}$.

(b) Sample mutational maps $M_n^{(A)}, M_n^{(B)}$ from the posterior distribution of mappings for each site in the set of paired columns to be tested. The details of sampling each map are identical to Nielsen (2002), with the addition that the branch lengths at that site are all scaled by the factor $r_n^{(sz)}$ calculated in the previous step.

(c) Calculate the value of each test statistic T_n for each site pair $\{A, B\}$:

$$T_n^{(A,B)} \mid G_n, \{M_n, r_n, D\}^{(A)}, \{M_n, r_n, D\}^{(B)} \tag{1}$$

This notation demonstrates that each calculation of the test statistic T_n is dependent upon the posterior sample of the substitution parameters and tree topology G_n, the posterior sample of mutational maps M_n, the posterior sample of rates at nucleotide positions r_n, and the data D at that site. These dependencies will be implied hereafter.

(3) Once Step 2 has been completed, calculate the expected value of each test statistic $\langle T \rangle$ by summing over all

samples from the posterior distribution:

$$\langle T \rangle = \sum_{n}^{N_{\text{gen}}} T_n \tag{2}$$

In all cases in this study, MrBayes was allowed to proceed for 1 100 000 iterations, with the first 100 000 iterations discarded as burn-in and every 1000 iterations sampled thereafter. Four rate categories were used in the approximation to the Γ-distribution (one of them invariant), and the GTR model was used to obtain the nucleotide substitution rates and stationary frequencies, with the default priors specified.

Markov model of coevolving protein sites

To test the power of the method and the various test statistics, sequence alignments were simulated using a novel Markov model of coevolving protein sites. The model is similar to the codon model of Nielsen and Yang (1998) (the NY model), but it instead describes the rate of a codon substitution $u \to v$ at a protein site A, which is correlated with site B. If the codon change is non-synonymous then the amino acid substitution is an $a \to j$ substitution. The current amino acid at site B is b, and simultaneous nucleotide substitutions are disallowed. The rate of codon substitution is:

$$Q_{uv}^{(A)} = \begin{cases} 0 & \text{if } u, v \text{ differ at} > 1 \text{ position} \\ \mu R_{uv} & \text{if } u \to v \text{ is synonymous} \\ \mu R_{uv} \omega_A & \text{if } \psi(j,b) = \psi(a,b) \\ \mu R_{uv} \omega_A Z_{AB} & \text{if } \psi(j,b) > \psi(a,b) \\ \mu R_{uv} \omega_A Z_{AB}^{-1} & \text{if } \psi(j,b) < \psi(a,b) \end{cases} \tag{3}$$

Here μ is a scaling factor that determines the overall rate of substitution, and R_{uv} is the base rate of the single-nucleotide mutation that yields a $u \to v$ codon mutation, which can be specified using a model such as the GTR or HKY (Hasegawa *et al.*, 1985) models. The next term, ω_A, is the rate scalar for all non-synonymous changes at site A, independent of the type of substitution. It is typically <1, indicating purifying selection. These elements of the model are exactly equivalent to the NY model.

The coevolutionary rate parameter in the model is Z_{AB}, the strength of the coevolutionary effect. It is based on the 20×20 interaction matrix, Ψ. If the substitution $a \to j$ is favorable in the context of the amino acid b at the coevolving site, $\psi(j,b) - \psi(a,b) > 0$ and the substitution rate increases by an amount Z_{AB}. Unfavorable substitutions, those where $\psi(j,b) - \psi(a,b) < 0$, have a decrease in rate by a multiple of Z_{AB}^{-1}. Note, Z is always positive.

Each site-pair therefore exists in one of two evolutionary regimes. When the a, b amino acid interaction is favorable, there are two possible substitution rates: $Q_{uv} \propto \omega$ (the basal NY rate) for any change to another favorable pairing, and $Q_{uv} \propto \omega\, Z^{-1}$ for a change to an unfavorable pairing. As Z increases, the rate of an unfavorable change will decrease. Once an unfavorable mutation is accepted, the site-pair enters a new regime: $Q_{uv} \propto \omega$ for any change to another unfavorable pairing, but now $Q_{uv} \propto \omega\, Z$ for a change to a favorable state. This results in a transitory increase in the mean substitution rate at both sites until the favorable pairing is restored, and substitutions at the sites will be correlated in time. Practically speaking, this means that a site in a favorable pairing will tend to remain in that state longer than a site in an unfavorable pairing, and the stationary frequency of each codon pair can be calculated analytically (not shown).

Test statistics

Once the distribution of mutational maps has been determined, a test statistic is required to evaluate the hypothesis of correlated substitution. Several test statistics will be evaluated here, divided into two broad categories: parametric and non-parametric tests. A parametric test is defined here as a test which involves a parameterized Markov model of evolution. The non-parametric test statistics do not use an explicit evolutionary model for their calculation, but instead rely on entropic measures or descriptive measures of correlation.[1]

With the exception of the mutual information-based test statistic MI (see below), all the test statistics require the specification of a 20×20 interaction matrix, Ψ. If a matrix entry $\psi(a,b) > 0$, an interaction between those amino acids at two sites in the protein is considered to be favorable, while if $\psi(a,b) < 0$ the interaction is unfavorable. In this study, the Ψ values are set with the assumption that the following interactions are favorable: hydrophobic residue pairs, polar pairs, and charged amino acid pairs (as long as the charge is of opposite sign). All other interactions such as polar–hydrophobic are unfavorable.

Each test statistic $\langle T \rangle$ is calculated as an expectation over the samples from the posterior distribution G_n and mutational mappings M_n as given in Equation (2). Each of the calculations shown below is for a single iteration n, but the subscript-n's are removed for ease of reading.

LR *and* $\hat{Z}$ (*likelihood ratio test and correlation parameter*) The likelihood ratio tests the strength of support for an alternative hypothesis H_1 against the null hypothesis H_0. In this case, H_1 is the hypothesis that the substitution rate at one site is dependent upon the current amino acid at its partner site, while the null hypothesis H_0 is that the sites evolve independently. The likelihood ratio is calculated using the model introduced above, where $H_0 : Z = 1$ (fixed) and $H_1 : Z \geq 1$. By making the simplifying assumption that the interaction matrix Ψ has only two possible values $\psi_{ab} = \pm \frac{1}{2}$, the log-likelihood ratio for a single mutational mapping can be

[1] While the non-parametric tests do not model coevolution as a parameterized process, all tests utilize the same site-independent Markov model to sample mutational mappings and to calculate P-values.

calculated analytically as:

$$\mathrm{LR} = N_{\mathrm{non}} \log \left(\frac{\widehat{\mu\omega}_{\mathrm{corr}}}{\widehat{\mu\omega}_o} \right) + N_d \log \left(\hat{Z} \right) \tag{4}$$

where N_{non} is the total number of non-synonymous mutations at the pair of sites, $N_d = N_+ - N_-$, the difference between the number of favorable and unfavorable mutations at the pair of sites given Ψ, and

$$\hat{Z} = \frac{c_o N_d + \sqrt{c_o^2 N_d^2 + 4c_1 c_2 \left(N_{\mathrm{non}}^2 - N_d^2 \right)}}{2c_1 \left(N_{\mathrm{non}} - N_d \right)} \tag{5}$$

$$\widehat{\mu\omega}_{\mathrm{corr}} = \frac{c_o N_{\mathrm{non}} - \sqrt{c_o^2 N_d^2 + 4c_1 c_2 \left(N_{\mathrm{non}}^2 - N_d^2 \right)}}{c_o^2 - 4c_1 c_2} \tag{6}$$

$$\widehat{\mu\omega}_o = \frac{N_{\mathrm{non}}}{c_0 + c_1 + c_2} \tag{7}$$

The c_i terms use draws from the posterior distribution of the GTR model and mutational partitions on the tree. These terms are related to the amount of time spent in each amino acid state, weighted by the basal transition rate out of that state; due to space constraints their derivation cannot be shown here but is provided in the Supplementary Material at the authors' web site. $\hat{Z}$ is the maximum likelihood estimate of the strength of correlated evolution between the sites, and can itself be used as a test statistic. The LR was set to 0 (no support for the alternative hypothesis) in the following limiting cases: $N_{\mathrm{non}} = 0$ at either site A or B, $N_d = N_{\mathrm{non}}$, and $\hat{Z} < 1$.

The LR test here differs from the common approach in phylogenetic likelihood calculations, which is to use a pruning algorithm to sum over all possible ancestral states (Felsenstein, 1981; Pollock *et al.*, 1999). Instead, the ML estimates of the model parameters in Equations (5)–(7) are conditional on a single sample from the posterior distribution of ancestral states, and therefore so is the LR in Equation (4). The test statistic's value, $\langle \mathrm{LR} \rangle$, is then obtained by averaging over many such samples. The advantage to using the BMM approach over a pruning algorithm is that the maximum likelihood estimates of the model parameters can be calculated analytically, without requiring numerical optimization.

W_+ (*weighted difference in escape times*) The coevolutionary hypothesis implies that an amino acid in an unfavorable paired state will be more likely to be substituted with a favorable paired state, and the waiting time in unfavorable states is expected to be low. Because this waiting time can also be affected by the structure and degeneracy of the genetic code, it may be important to correct for this on the codon level. The W_+ statistic utilizes the waiting times spent in each codon, weighted by the expected rate of the observed substitutions under a model of no coevolution:

$$W_+ = \frac{c_2 - c_1}{c_0 + c_1 + c_2 + c_3} \tag{8}$$

The c_i terms are identical to the those in the calculation of LR. Briefly, if an amino acid pair is in a favorable state of interaction, then it will spend a long time 'waiting' to mutate to unfavorable interaction states, measured as a high value of c_2. Conversely, when the residue pair is not physicochemically favorable, the rate of mutation to favorable states will be higher than in the null case and the waiting time spent in those states will be short, measured as a low value of c_1. This test is parametric in the sense that it relies on values drawn directly from the Bayesian posterior distributions, but it does not rely on a particular model of coevolution (other than the specification of the interaction matrix).

S_+ (*difference in time spent in each state*) This non-parametric statistic, identical to the expected frequency of association used by Huelsenbeck *et al*. (2003), measures the difference in the time spent in favorable states versus the expected time spent in favorable states:

$$S_+ = \sum_{a,b} t_{ab} - t_a t_b \quad \text{for all } a, b, \text{ where } \psi_{ab} > 0 \tag{9}$$

where t_{ab} is the total time spent in a favorable amino acid state pair, and t_a and t_b are the total time spent in those states at sites A and B, independent of the other site. It is similar in spirit to the W_+ statistic, although it does not adjust for the differential expectation of mutations due to the genetic code and the nucleotide substitution rates, as W_+ does.

MI *and* MI_+ (*mutual information in alignment*) MI is an alignment-based measure; it does not use any information from the tree (Atchley *et al*., 2000). This statistic calculates the mutual information contained in correlated substitutions which appear in the alignment:

$$\mathrm{MI} = \sum_{a,b} p_{ab} \log_2 \left(\frac{p_{ab}}{p_a q_b} \right) \tag{10}$$

where p_a and q_b are the fraction of sequences with amino acids a and b at sites A and B, respectively, while p_{ab} is the fraction of sequences with both a at A and b at B. MI_+ only counts the mutual information contained in site pairs where the amino acid pair interaction is favorable.

Posterior predictive P-values

Posterior predictive P-values are used to determine whether the null hypothesis of no coevolution at a pair of sites can be reliably rejected for each test statistic. They represent the probability of a test statistic's value, given data simulated under the null hypothesis of no coevolution, and are similar in spirit to the parametric bootstrap typically used in likelihood calculations (Pollock *et al*., 1999). The difference is that the bootstrap evaluates the expected distribution of the statistics at a single estimate of model and tree parameters, whereas the posterior predictive P-values average this expected distribution over the posterior distribution of unknown parameters.

These P-values are calculated using the method described in Nielsen (2002), with an addition to utilize rates across sites. Iterating over posterior samples G_i:

(1) For each sample G_i, record the set of ordered rates $r_i^{(s)} = \left\{r^{(s1)}, r^{(s2)}, r^{(s3)}\right\}_i$ assigned to each codon pair to be tested in Step 2a. For example, if there are 4 rate categories, the rate categories at nucleotide positions 1, 2, and 3 of the codon at site A for iteration i might be $r_i^{(A)} = \{1, 1, 3\}$, and those at site B might be $r_i^{(B)} = \{2, 1, 4\}$.

(2) For each $r_i^{(A)}$ and $r_i^{(B)}$, simulate a replicate alignment column $D_i^{(\mathrm{sim})}(r)$ under the hypothesis that the nucleotide sites are uncorrelated. This is done by simulating the evolution of three nucleotide sites (one for each codon position) using the model, branch lengths, and tree in the sample G_i using the techniques from Nielsen (2002), with the branches at each nucleotide site scaled by the appropriate factor in r. Therefore, if there are N_r rate categories, there will potentially be $2N_r^3$ replicate alignment columns simulated per iteration,[2] although in practice it may be fewer if some rate category combinations are not found among the sites to be tested.

(3) Once a replicate dataset $D_i^{(\mathrm{sim})}$ has been simulated for each iteration i, proceed as if that dataset were the true data, just as in Step 2 of the procedure described in section . In other words, for each replicate pair $D_i^{(A)}$ and $D_i^{(B)}$, iterate over the draws from the posterior G_n, where this inner loop is denoted by the index n:

   (a) Generate a 'null' mutational map $M_{in}^{(A)}$ and $M_{in}^{(B)}$, using the posterior sample G_n and the posterior rate sample for that pair of sites, $r_i^{(A)}$ and $r_i^{(B)}$.

   (b) Calculate the value of each test statistic under the null hypothesis, $T_{in}^{(\mathrm{null})}$, for that site pair $\{A, B\}$ using the mutational maps, rates, tree topology, and substitution parameters drawn from the posterior.

(4) Calculate the mean null value of each test statistic $T_i^{(\mathrm{null})}(A, B)$ for each site pair in each simulated dataset i:

$$\langle T\rangle_i^{(\mathrm{null})} = \sum_n T_{in}^{(\mathrm{null})} \tag{11}$$

(5) Once $\langle T\rangle_i^{(\mathrm{null})}(A, B)$ has been calculated for each test statistic on each simulated dataset, rank them against the value of the test statistic on the actual data for the site pair $\langle T\rangle(A, B)$ to determine the posterior predictive P-value.

It is important to note that, by using the rate information to generate data replicates, each site pair $\{A, B\}$ is compared only with replicate sites that have evolved at the same rate. For example, if a pair of sites has many mutations, there will be many co-occurring mutations by chance. This would cause the value of $\langle T\rangle$ on the real data to be high even when the sites were not truly coevolving, potentially yielding a high FP rate. By generating the distribution of $\langle T\rangle^{(\mathrm{null})}$ using the observed rate distribution for the site pair, these FPs are minimized, because fast-evolving sites will also yield a higher $\langle T\rangle^{(\mathrm{null})}$ leading to higher (less significant) P-values.

Simulated datasets

The test statistics were evaluated on sequence datasets which were generated using the Markov model of coevolution described above. Sequence evolution was simulated on a 32-taxon symmetric balanced tree with equal branch lengths. 800 codons were simulated for each dataset. The first 600 sites represent 300 codon pairs which were coevolving, with the coevolutionary rate parameter Z_{sim} set to $Z_{\mathrm{sim}} = \mathrm{e}^{1/2}$ for moderate coevolution and $Z_{\mathrm{sim}} = \mathrm{e}^1$ for strong coevolution. On the null (non-coevolving) datasets, the first 300 codon pairs were simulated with $Z_{\mathrm{sim}} = 1$. ω_{sim} was set to 0.6 at these sites, so on the moderate coevolutionary dataset $\omega Z_{\mathrm{sim}} \approx 1$ (which means a compensatory non-synonymous mutation occurs at the rate of synonymous change) and on the strong coevolutionary dataset $\omega Z_{\mathrm{sim}} \approx 1.6$, which puts the site-pair in an adaptational regime for compensatory changes.

The remaining 200 codons in each dataset were included to represent the fact that in real datasets, the coevolving sites will be mixed in among independent sites. In these codon positions $Z_{\mathrm{sim}} = 1$ (no coevolution) and ω was mixed: $\omega_{\mathrm{sim}} = 0, 0.3, 0.6, 1, 1.3$ for 20, 60, 70, 30 and 20 codons, respectively. At all codon sites, the rate of each nucleotide transition was set to twice the rate of transversion, all nucleotide frequencies were equal, and the initial frequency of each codon at the root was calculated from the model parameters.

PGK dataset

Thirty four eukaryotic nucleotide sequences of PGK were assembled as follows: SWISSPROT identifiers were taken from Pfam entry PF00162 (Bateman *et al.*, 2004), the GenBank sequence identifiers were then taken from each SWISSPROT entry, and the nucleotide sequences were downloaded from GenBank and assembled. For alignment, the codon sequences were translated into amino acids and aligned to the structural alignment of the closed-form crystal structure of bacterial PGK (1VPE) and the open-form yeast structure 1QPG (Berman *et al.*, 2000). Supporting scripts were written in Python with the aid of the Biopython package (www.biopython.org).

[2] Since sites A and B are generated with an identical model, it might seem that only N_r^3 columns need to be simulated. This causes problems when two sites for comparison have the same r, as their $D_m^{(\mathrm{sim})}$ columns would be identical, leading to a false inference of coevolution between the pair.

Table 1. Summary of test statistics using BMM

Test	Mean $\langle T \rangle$			% Pairs detected		
statistics	Strong	Moderate	None	Strong	Moderate	None
LR	6.93	2.94	0.32	99.0	59.7	3.3
$\hat{Z}$	2.05	1.52	1.03	97.7	54.3	2.7
W_+	0.16	0.10	0.00	84.3	31.7	1.0
S_+	0.14	0.10	0.00	50.7	28.3	2.0
MI_+	1.63	1.41	0.99	58.7	30.3	1.7
MI	1.88	1.95	1.97	1.3	1.0	2.3

All results are based on simulated datasets as described under the Simulation section, with 1 expected nucleotide mutation per codon per branch and $Z_{\text{sim}} = 2.72$, 1.65 and 1 for strong, moderate and no correlation between interacting site pairs respectively.

Sites were chosen for testing based on their proximity in the closed-form crystal structure 1VPE. Site-pairs were considered in contact if the minimum inter-residue distance of all pairs of heavy atoms was ≤ 8 Å. This is a liberal cutoff (not all residues will be assured of making contact), but it accommodates transient contacts due to flexibility in the protein structure. Site-pairs in the non-contact set had a minimum distance of ≥ 16 Å. To account for any systematic bias due to non-coevolutionary reasons, each non-contact pair consisted of one site from the set of contact pairs and another site chosen randomly that was at least 16 Å away. Only interdomain contact and non-contact pairs were considered, where the domains were defined as positions 1–187 and 188–399 in the reference sequence. Once gapped and invariant sites were removed from the dataset, there were 116 site pairs in the contact set and 670 site pairs in the non-contact set.

RESULTS

Comparison of test statistics

Simulated datasets were used to compare the power of the various test statistics (Table 1), which are divided into parametric and non-parametric measures. The positive correlation between LR and Z_{sim} indicates that, as the strength of the coevolution increases, the data's support for the coevolutionary model also increases. The mean value of the $\hat{Z}$ parameter estimate on the null dataset is close to the true value of $Z_{\text{sim}} = 1$, but the estimates of $\hat{Z}$ are biased toward lower values relative to the values used for simulation on the coevolving datasets ($Z_{\text{sim}} = 2.72$ and 1.65 for strong and moderate coevolution respectively). The S_+ test indicates that the time spent in favorable site pairings is 10–14% greater than expected under the null, with similar results found when the waiting times are weighted by the expected nucleotide substitution rates using the W_+ statistic. Because the measured values of the test statistics can be dependent on variables such as branch length and tree topology, the posterior predictive P-values provide a more reliable estimate of the power of each statistic. Among the 300 site-pairs simulated with strong coevolution, the LR test could detect 297 of them at the 0.01 level, a 99% true positive (TP) rate. In contrast, the best non-parametric test, MI_+, could detect 59% of the strongly coevolving sites. When simulated with more modest coevolution, LR could detect nearly 60% of the sites at the 0.01 level, and the TP rate for MI_+ dropped to 30%.

Table 1 also indicates that the FP rate is slightly elevated from the expectation of 1%. In some situations a lower Type I error rate is desired; Figure 1 illustrates the power of the most sensitive parametric and non-parametric test statistics LR and MI_+ at critical values between <0.001 and 0.05. In general, the parametric tests outperform the non-parametric tests, especially as the degree of coevolution becomes stronger. The high sensitivity of the LR and $\hat{Z}$ tests could be attributable to the fact that they are based on the same Markov model used for simulation (although the values of the parameters are not fixed at the true values but are maximum likelihood estimates). Nevertheless, it is gratifying to find that, when the model of coevolution resembles the true coevolutionary process, the BMM method can accurately detect coevolving sites. It is also an indication that the method is robust to its approximation of the true null model, the use of a nucleotide site-independent model to estimate the posterior predictive P-values. Finally, it is worth noting that another parametric test, W_+, also generally performs better than the non-parametric tests, and it is not based on the coevolutionary simulation model.

The results discussed above were measured on simulated data, where one nucleotide substitution is expected per codon per branch in the evolutionary tree (because $\omega_{\text{sim}} = 0.6$, the amino acid substitution rate is lower). Figure 2 examines the effect of increasing sequence divergence on the P-values for several statistics. Because the BMM method uses mutational information as data, more mutations on the tree provide more potential evidence for (or against) coevolution. For this reason, even though the value of the estimate may change with increasing evolutionary distance (not shown), power tends to increase and is retained even when sequence divergence is high. For example, when the branch length is 4, there are on average 4 codon substitutions per site per branch (8 between nearest sequence neighbors), yet LR can still detect nearly every co-evolving site with a low FP rate (Fig. 2).

With one exception, each of the statistics shown in Table 1 requires the specification of an interaction matrix, the set of amino acid interactions which are favorable and unfavorable. The exception is MI, which calculates the mutual information in all amino acid substitutions at both sites, not only those which are compensatory. When compared with MI_+ (which utilizes only compensatory substitutions), MI shows an increase in the mutual information but a dramatic loss in power in the posterior predictive P-values. Other tests showed a similar loss of power when no interaction matrix is specified (not shown). This reinforces the conclusions of previous studies (Tufféry and Darlu, 2000; Fukami-Kobayashi *et al.*, 2002): for a test statistic to maintain power, some hypothesis of the

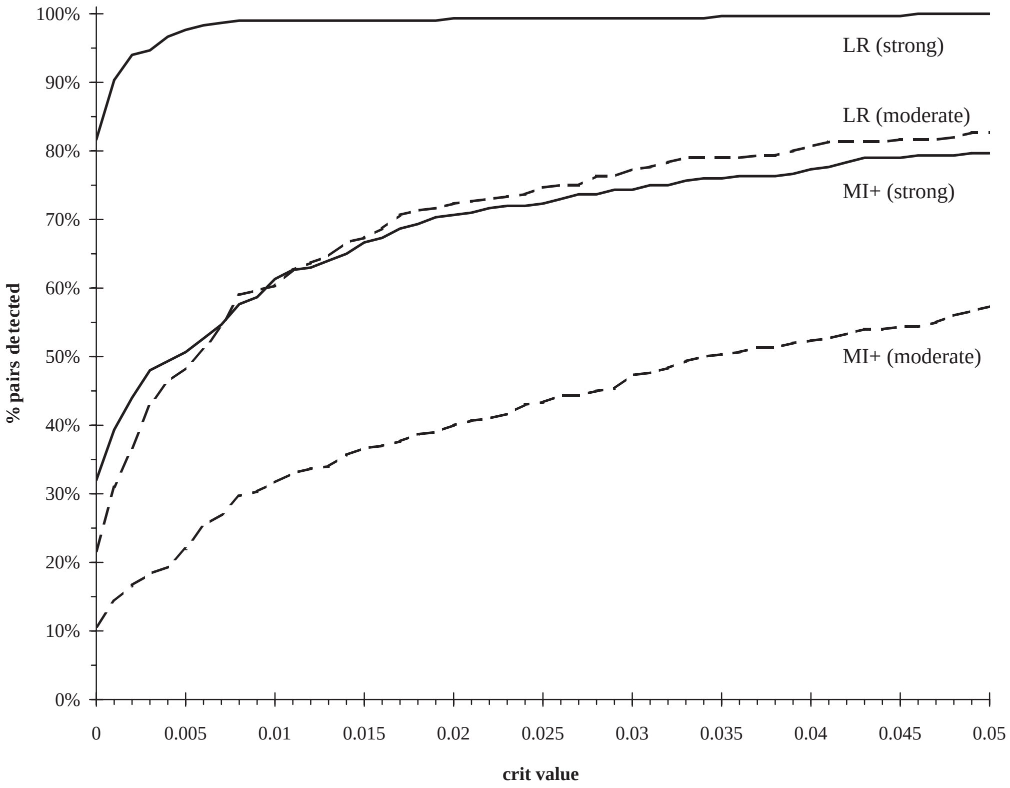

Fig. 1. Sensitivity (TP rate) of LR and MI_+ at varying critical values on a simulated dataset. Simulation details are given in the Methods, and $Z_{\mathrm{sim}} = 2.72$ and 1.65 for strongly and moderately coevolving sites respectively.

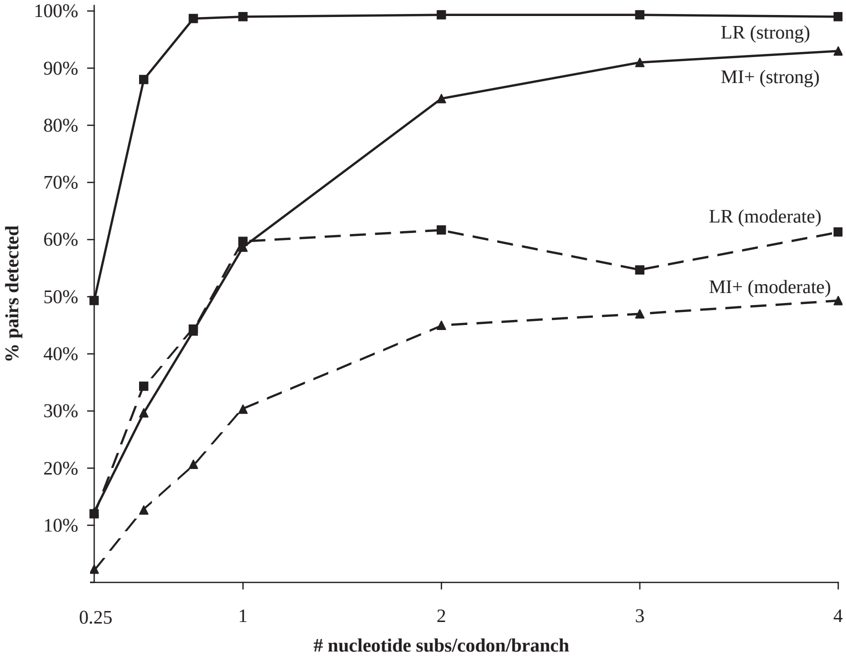

Fig. 2. Power of selected test statistics for different levels of divergence, represented as the percentage of significant site pairs detected at a critical P-value of 0.01. Simulation details are given in the Methods.

Table 2. Mean and power of each test statistic on interdomain PGK contact pairs (<8 Å apart) versus non-contact pairs (>16 Å apart)

Test statistics	With rate heterogeneity				Without rate heterogeneity			
	Mean $\langle T \rangle$		% Pairs detected		Mean $\langle T \rangle$		% Pairs detected	
	Contact	Non-contact	Contact	Non-contact	Contact	Non-contact	Contact	Non-contact
LR	1.08	0.54	12.9	3.3	1.39	0.69	26.7	12.1
$\hat{Z}$	1.63	1.31	16.4	4.8	2.11	1.43	32.8	19.3
W_+	0.02	−0.01	2.6	1.2	0.03	0.00	22.4	15.7
S_+	0.00	0.00	0.0	1.2	0.00	0.00	0.0	0.9
MI_+	0.12	0.12	0.0	0.0	0.12	0.12	0.0	0.0
MI	0.24	0.24	0.9	0.1	0.24	0.24	0.0	0.0

Significance is measured at the 0.01 level, with 116 total pairs in the contact dataset and 670 total pairs in the non-contact dataset. On the left side, heterogeneous rates among sites were used to create the mutational maps and calculate the posterior predictive P-values.

physicochemical basis of coevolution is necessary. Also, the fact that an increase in mutual information can still yield a decrease in power demonstrates how the posterior predictive P-values correct for the 'noise' of substitutions under the null.

Using BMM to detect interacting sites: PGK

PGK is an enzyme involved in the glycolytic pathway, catalyzing the transfer of a phosphate from ATP to 3-phosphoglycerate. Catalysis requires closure of the N- and C-terminal domains of the protein, a hinge action which is thought to involve significant relative motion between the two domains (Chandra *et al*., 1998). Because interdomain interaction is required for enzymatic function, the evolution of the amino acids at this interface is likely to be constrained (Teichmann, 2002). This constraint is not expected to be absolute, however; many of the interacting residues are not completely conserved, indicating that there is some tolerance for amino acid substitutions at these positions. Since such substitutions may alter the fitness environment for nearby residues, this evolutionary regime of mild deleteriousness may provide fertile ground for compensatory mutations to occur. Coevolution between PGK domains has previously been observed using tree similarity methods (Goh *et al*., 2000).

To test whether the BMM method and coevolutionary model were capable of detecting coevolution in PGK, a dataset of eukaryotic PGK sequences was assembled as described in the Methods. The sites were paired into two sets according to their proximity in the closed-form crystal structure: interdomain contact pairs and interdomain non-contact pairs. Table 2 (left side) shows the difference in coevolutionary signal between these sets for each of the test statistics. The mean values of LR and $\hat{Z}$ are higher on the set of contact pairs, indicating greater support on average for the coevolutionary model and a stronger coevolutionary rate ratio at proximal residues. Several of the other test statistics also show a difference in their mean values, although the differences are less pronounced.

The differences in the posterior predictive P-values are more clear. At the 0.01 level of significance, 12.9% of the contact pairs returned a significant value of LR, versus 3.3% of the noncontact pairs. The difference for $\hat{Z}$ is even more dramatic (16.4% versus 4.8%), although the positive rate on the non-contact pairs is slightly elevated. None of the non-parametric tests find a significant signal on the interacting sites. When the significant contact pairs under the LR test are mapped onto the crystal structure of PGK, several of the pairs are located near the terminus of the hinge helix involved in domain closure. Since domain closure is thought to be effected by tightening or loosening of the winding in this helix (Chandra *et al*., 1998), the coevolutionary signal in these pairs may reveal functional importance. To test the importance of using site rates to inform the P-value calculations and mutational mapping steps, the analysis was repeated with homogeneous rates, and the results are shown on the right side of Table 2. (The Γ-distribution of rates was still used to inform the phylogeny during Step 1, when sampling trees and substitution parameters; it was not used in generating mutational maps or posterior predictive P-values.) Although, the differences are more pronounced in some tests and relatively more contact sites were found to be significant, in most cases more non-contact sites were significant as well. For example, although the number of significant pairs in the contact set with LR doubled to 27%, the number of significant pairs in the non-contact set also increased to 12%.

While it is possible that there is some true coevolution occurring in a few non-interacting site pairs, it is more likely that most of this increase is due to a spurious signal, which is dampened by the use of rate information in the posterior predictive P-values (Tufféry and Darlu, 2000). In addition, the number of truly coevolving site pairs among those in contact is unknown, but other studies have found that only a fraction of interacting sites in proteins seem to exhibit a coevolutionary signal. Nevertheless, it should be noted that the no-rates BMM method is computationally much faster than the method which uses rate information. Also, the significant sites detected using rate information are a complete subset of the sites detected without rates, indicating that the no-rates BMM method might be useful as a fast initial screening step in determining which sites in a protein to test with the more discriminate method.

DISCUSSION

In cases where correlated evolution has previously been examined, it has sometimes been observed that the coevolutionary signal is weak, if it exists at all. It is certainly possible that correlated substitutions are rare events in protein evolution, but these results indicate that we are not yet at the limit of detection. In most simulated regimes, the model-based tests introduced here have more power than the non-parametric tests. While the superior performance of the parametric tests on simulations is due in part to the fact that the simulations were conducted using the same model structure, it is important to note that the BMM method is capable of detecting nearly all the coevolving site pairs when the model has been correctly specified. In addition, the fact that the model-based tests detect more correlated substitutions in PGK's contact residues is evidence that the model does provide a practical approximation to the evolution of interacting sites.

The key to the power of any coevolutionary detection method is that it accounts for the phylogenetic history of the proteins in some fashion, whether to inform a parametric model (Pollock *et al*., 1999), to reconstruct ancestral states (Fukami-Kobayashi *et al*., 2002), or to test for significance (Atchley *et al*., 2000). The BMM method and coevolutionary model described here utilize all these approaches, with the additional benefit that explicit knowledge of the true tree and its branch lengths are not required. This is less important when the true tree is well-known, such as in the case of the simulated datasets, which used a balanced tree that could be inferred without difficulty using neighbor-joining. But in the 34-sequence PGK dataset used here, there were 157 trees in the credible posterior set. While many of these trees involve small shifts in the topology, the Bayesian method accounts for this uncertainty by integrating over these trees. BMM also integrates over the substitution rates at each site, leading to a marked decrease in the rate of detection among non-contact residues.

When mapping the mutations and calculating P-values, one simplifying assumption is that of nucleotide independence. This is done both for convenience (current Bayesian software does not integrate over the coevolutionary model) and for speed (the computational cost of the simulation and mutational mapping steps increase dramatically with a large state space). As a result, the model used to simulate the sequences when calculating the P-values is not the true null model in the likelihood ratio, although it is still a site-independent model. The results on simulated datasets show that the method is relatively robust to these assumptions under the conditions tested, but it may be feasible to increase the method's power by incorporating the coevolutionary model into the integration and significance testing framework.

Future developments may also focus on improvements to the likelihood model or development of other test statistics. The significant improvement in the FP rate gained by including rate heterogeneity demonstrates how specificity can increase with a more accurate evolutionary model. The slight increase in the P-values for LR and $\hat{Z}$ on non-contact sites above the expectation of 1% indicates that there is still room for improvement. For example, the requirement for a binary interaction matrix is not an essential one for the model, only for the analytical solution to the maximum likelihood. It may be possible to derive an approximate maximum likelihood for the case of a continuum matrix, thereby allowing more elaborate interactions to be described. The model analyzed in this study assumes a constant 'background' mutational pressure—selection due to interaction with other sites not being tested—but the model can be extended to accommodate heterogeneity in this parameter. Once a new statistic is devised it should be relatively straightforward to test using BMM, since the BMM framework is quite flexible and can utilize a variety of tests which are informed by the changes on the tree.

ACKNOWLEDGEMENTS

The authors would like to thank Scott Williamson, John Huelsenbeck, Nick Goldman and Simon Whelan for helpful discussions. We thank an anonymous reviewer for their insightful comments which improved the manuscript. Computational resources were provided by the Cornell Theory Center, and we thank Daniel Sverdlik at the CTC for his assistance. Funding for M.W.D. was provided by a National Science Foundation Postdoctoral Fellowship in Biological Informatics, and M.J.H. was supported by a grant from the Human Frontiers in Science Program (HFSP).

REFERENCES

Atchley,W.R., Wollenberg,K.R., Fitch,W.M., Terhalle,W. and Dress,A.W. (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.*, **17**, 164–178.

Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, 138–141.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Chandra,N.R., Muirhead,H., Holbrook,J.J., Bernstein,B.E., Hol,W.G. and Sessions,R.B. (1998) A general method of domain closure is applied to phosphoglycerate kinase and the result compared with the crystal structure of a closed conformation of the enzyme. *Proteins*, **30**, 372–380.

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

Fitch,W.M. and Markowitz,E. (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.*, **4**, 579–593.

Fukami-Kobayashi,K., Schreiber,D.R. and Benner,S.A. (2002) Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *J. Mol. Biol.*, **319**, 729–743.

Gobel,U., Sander,C., Schneider,R. and Valencia,A. (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

Goh,C.S., Bogan,A.A., Joachimiak,M., Walther,D. and Cohen,F.E. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.

Govindarajan,S., Ness,J.E., Kim,S., Mundorff,E.C., Minshull,J. and Gustafsson,C. (2003) Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation. *J. Mol. Biol.*, **328**, 1061–1069.

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

Hamilton,N., Burrage,K., Ragan,M.A. and Huber,T. (2004) Protein contact prediction using patterns of correlation. *Proteins*, **56**, 679–684.

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

Huelsenbeck,J.P., Bollback,J.P. and Levine,A.M. (2002) Inferring the root of a phylogenetic tree. *Syst. Biol.*, **51**, 32–43.

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

Huelsenbeck,J.P., Nielsen,R. and Bollback,J.P. (2003) Stochastic mapping of morphological characters. *Syst. Biol.*, **52**, 131–158.

Huelsenbeck,J.P., Ronquist,F., Nielsen,R. and Bollback,J.P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

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

Nielsen,R. and Yang,Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **134**, 1271–1276.

Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol.*, **271**, 511–523.

Pollock,D.D. and Taylor,W.R. (1997) Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng.*, **10**, 647–657.

Pollock,D.D., Taylor,W.R. and Goldman,N. (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.*, **287**, 187–198.

Pritchard,L., Bladon,P., M O Mitchell,J. and J Dufton,M. (2001) Evaluation of a novel method for the identification of coevolving protein residues. *Protein Eng.*, **14**, 549–555.

Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Shindyalov,I.N., Kolchanov,N.A. and Sander,C. (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.*, **7**, 349–358.

Teichmann,S.A. (2002) The constraints protein–protein interactions place on sequence divergence. *J. Mol. Biol.*, **324**, 399–407.

Tillier,E.R.M. and Lui,T.W.H. (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, **19**, 750–755.

Tufféry,P. and Darlu,P. (2000) Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. *Mol. Biol. Evol.*, **17**, 1753–1759.

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