BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 9 2014, pages 1241–1249
doi:10.1093/bioinformatics/btu012

Phylogenetics

Advance Access publication January 11, 2014

Evolutionary footprint of coevolving positions in genes

Linda Dib[1,2], Daniele Silvestro[1,2,3] and Nicolas Salamin[1,2,*]

[1]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland, [2]Swiss Institute of Bioinformatics, Quartier Sorge, 1015 Lausanne, Switzerland and [3]Department of Plant and Environmental Sciences, University of Gothenburg, Carl Skottsbergs gata 22B, 413 19 Gothenburg, Sweden

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The analysis of molecular coevolution provides information on the potential functional and structural implication of positions along DNA sequences, and several methods are available to identify coevolving positions using probabilistic or combinatorial approaches. The specific nucleotide or amino acid profile associated with the coevolution process is, however, not estimated, but only known profiles, such as the Watson–Crick constraint, are usually considered *a priori* in current measures of coevolution.

Results: Here, we propose a new probabilistic model, Coev, to identify coevolving positions and their associated profile in DNA sequences while incorporating the underlying phylogenetic relationships. The process of coevolution is modeled by a 16×16 instantaneous rate matrix that includes rates of transition as well as a profile of coevolution. We used simulated, empirical and illustrative data to evaluate our model and to compare it with a model of 'independent' evolution using Akaike Information Criterion. We showed that the Coev model is able to discriminate between coevolving and non-coevolving positions and provides better specificity and specificity than other available approaches. We further demonstrate that the identification of the profile of coevolution can shed new light on the process of dependent substitution during lineage evolution.

Availability: http://www2.unil.ch/phylo/bioinformatics/coev

Contact: nicolas.salamin@unil.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 28, 2013; revised on December 28, 2013; accepted on January 3, 2014

1 INTRODUCTION

Coevolution is defined as 'the modification of a biological object triggered by the change of a related object' (Yip *et al.*, 2008). This process of dependent evolution has been described in various biological systems and can be an essential process behind changes occurring at both morphological level, e.g. coevolution of female and male genital morphology (Fitzpatrick *et al.*, 2012), and molecular level (Gobel *et al.*, 2004), e.g. ligand–receptor interactions (Chockalingam *et al.*, 2005).

At the molecular level, studies of dependent substitutions have gained importance in evolutionary biology because of the potential functional and structural interpretations that can be made on the positions identified as coevolving. For example, coevolving positions in proteins are known to be involved in allosteric communication, to create physically connected networks that link distant functional positions in the tertiary structure, and to modify the structure of the protein (Baussand and Carbone, 2009; Lockless and Ranganathan, 1999). Further, it has been shown recently that coevolving fragments within protein sequences can also be involved in binding specificity and folding constraints (Dib and Carbone, 2012). They are thus good indicators to explain folding intermediates, peptide assembly and key mutations with known roles in genetic diseases (Dib and Carbone, 2012). Coevolution within RNA sequences also revealed structurally and functionally important positions (Dutheil *et al.*, 2010), which are often located on helices and are subject to Watson–Crick constraint (i.e. guanine–cytosine and adenine–thymine complementarity). For example, the positions 245 and 283 of the 16S ribosomal RNA in *Thermus thermophilus* coevolved under the specific Watson–Crick profile {CC, UU}, respectively, to maintain the specific hydrogen bonds necessary for the stability of the helix hairpin-like structure (Cannone *et al.*, 2002).

Current methods designed to detect coevolution focus primarily on the identification of pair of positions along the aligned sequence by evaluating a score of coevolution. Some methods consider solely the set of aligned sequences to estimate these coevolving positions (Carbone and Dib, 2011; Codoñer and Fares, 2008; Fares and Travers, 2006; Gloor *et al.*, 2005; Lockless and Ranganathan, 1999; Yip *et al.*, 2008), while others take into account the evolutionary history of the observed sequences (Baussand and Carbone, 2009; Corbi *et al.*, 2012; Dib and Carbone, 2012; Dutheil *et al.*, 2005; Yeang *et al.*, 2007).

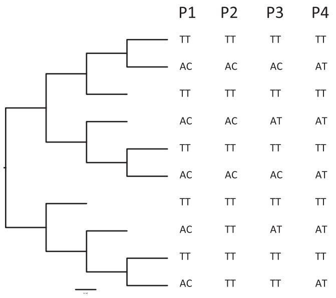
The score of coevolution usually represents the correlation of the nucleotide or amino acid patterns found at two different positions along a multiple alignment. This correlation can be considered as the outcome of the dependent evolutionary process that is depicted by coevolution and the correlated pairs constitute the set of coevolving nucleotides or amino acids. This set is defined here as the profile of coevolution and is similar to the profile of site classes present in phylogenetic mixture models (e.g. Lartillot and Philippe, 2004). While current methods designed to detect coevolving positions have proven very useful, they do not estimate the coevolving profile that corresponds to the structural or functional constraints involved in the evolutionary process (Dutheil *et al.*, 2010). This lack of a mechanistic component induces the use of arbitrary rules to identify the profile of coevolution for protein-coding sequences, such as using the most frequent pattern within a pair of coevolving positions (Lockless and Ranganathan, 1999). Alternatively, the profile of

[*] To whom correspondence should be addressed.

coevolution can be set *a priori*, like in the case of RNA sequences where Watson–Crick profiles are known to be important. Setting a profile *a priori* was done in a Bayesian context (Dutheil and Galtier, 2007) by constraining the mapping of substitutions that occurred independently at each site on the branches of the phylogenetic tree using, as weights, the biochemical properties of the nucleotides defining the Watson–Crick constraint. Similarly, this constraint was used as modifiers of the rate parameters of a dependent model describing the process of coevolution (Yeang *et al.*, 2007). State transitions that establish or maintain Watson–Crick base pairs were favored and coevolution under this specific constraint was assessed through likelihood ratio tests. These two approaches confirmed that a well-known coevolving pair was constrained by a Watson–Crick profile in 16S RNA sequences (Dutheil *et al.*, 2005; Yeang *et al.*, 2007). However, other known types of constraints, such as the Hoogsteen base pairing (Westhof and Fritsch, 2000), were not explicitly incorporated in the model and thus not identified in these analyses. Furthermore, the integration of particular constraints in the models may be difficult in the case of amino acid or protein-coding DNA sequences, for which little *a priori* information is available regarding the coevolving profiles. Thus, the coevolving profile should be considered as a parameter to be estimated from the data, rather than being defined on the basis of the frequencies of coevolving patterns or of known constraints.

Methods that estimate coevolving profiles in DNA, RNA or protein sequences are currently not available. However, some methods have been developed for binary encoded data where coevolving profiles represent presence or absence of characters such as genes, restriction sites, introns, indels and methylation sites (Cohen *et al.*, 2013; Franceschini *et al.*, 2013). Identifying the profiles of binary character states involved in correlated evolution is also an important part of the analyses of phenotypic data. These approaches have progressed from using maximum parsimony criteria (Boussau *et al.*, 2004; Mirkin *et al.*, 2003) to a full probabilistic framework (Csuros, 2005; Hao and Golding, 2006; Pagel, 1994) in which the dynamics of presence and absence of events are assumed to follow a continuous-time Markov process along the phylogenetic tree. In this context, it is possible to define a dependent substitution model, which is expressed as a 4×4 matrix of instantaneous transition rates, to explain the correlated evolution of two binary characters (Pagel, 1994). Each row and column of this matrix specifies a pair of character states built on an alphabet of size 2 (i.e. the character states of binary traits), which defines two possible profiles of coevolution {00, 11} or {01, 10}. This model can be compared to an independent model of evolution, which assumes a single 2×2 substitution matrix, through likelihood ratio tests (Pagel, 1994).

It is possible to apply the ideas used for binary data to identify profiles in molecular data. The nucleic (or proteic) alphabet can be reduced to an alphabet of size two using the physico-chemical properties behind the nucleotides or amino acids (Pollock *et al.*, 1999). This reduction is, however, arbitrary and could omit important factors that play a role in the evolution of the molecular sequences. For this reason, we propose in this study a new Markov model that describes the evolutionary process of coevolving positions along DNA sequences. Because correlated positions within nucleotide and protein sequences are the result of an evolutionary process, a better understanding of the coevolving profile should be obtained by incorporating the underlying evolutionary history of the sequences (i.e. their phylogenetic relationships and the associated profile; Dutheil *et al.*, 2010). We thus developed a dependent model of nucleotide substitutions based on a 16×16 instantaneous rate matrix that includes four substitution rates s, d, $r1$, $r2$ and a fifth parameter ϕ representing the profile of coevolution. We implemented the model in maximum likelihood and Bayesian frameworks. The performance of the model was tested simulated datasets and we compared the model's predictions with existing methodologies on empirical data.



P1	P2	P3	P4
TT	TT	TT	TT
AC	AC	AC	AT
TT	TT	TT	TT
AC	AC	AT	AT
TT	TT	TT	TT
AC	AC	AC	AT
TT	TT	TT	TT
AC	TT	AT	AT
TT	TT	TT	TT
AC	TT	TT	AT

Fig. 1. Illustrative example tree. Tree with 10 leaves and a fixed branch length (0.1). On the right-hand side of the tree we give a replicate for each of the four pairs P1–P4 described below in Table 1

2 MATERIALS AND METHODS

2.1 Definitions and notations

Let us consider the set S of aligned orthologous DNA sequences and a phylogenetic tree τ. We define a combination $\mathcal{C}$ as the association of two letters $l_1(\mathcal{C})$, $l_2(\mathcal{C})$ from the same aligned sequence seq $\in S$, where $l_1(\mathcal{C})$ is the letter at the first position in seq and $l_2(\mathcal{C})$ is the letter at the second position in seq. The maximal number of possible combinations depends on the size of the alphabet and is equal to 16 in the case of nucleotide sequences. For a pair of positions, the two combinations $\mathcal{C}_1$ and $\mathcal{C}_2$ are called 'conflicting combinations' if either $l_1(\mathcal{C}_1)$ is equal to $l_1(\mathcal{C}_2)$ or $l_2(\mathcal{C}_1)$ is equal to $l_2(\mathcal{C}_2)$, that is, if two words share a common letter at any one of the two positions.

For a pair of positions, every combination $\mathcal{C}_i$ has a number of occurrences $o(\mathcal{C}_i)$ within the associated alignment S and a combination frequency $f(\mathcal{C}_i)$, which is $o(\mathcal{C}_i)$ divided by the total number of aligned sequences. In the pair $P1$ (Fig. 1), for example, combinations AC and TT occur at the same frequencies, whereas in pair $P2$, the number of occurrences o(TT) is higher than the other and, thus, f(TT) > f(AC) (Table 1).

We define as Comb the set of combinations observed in a pair of sites, which represents a subset of the 16 possible combinations. For example, pair $P3$ (Fig. 1) has Comb = {AC, TT, AT}.

The 'proportion of conflicts' expresses the proportion of conflictual combinations at two positions p_1 and p_2 as

$$\frac{\sum\limits_{C_i \in \text{Comb}} \sum\limits_{C_j \in \text{Comb},\, C_i \neq C_j,\, C_i \text{ conflict } C_j} o(C_i) \times o(C_j)}{\sum\limits_{C_i \in \text{Comb}} \sum\limits_{C_j \in \text{Comb},\, C_i \neq C_j} o(C_i) \times o(C_j)}. \qquad (1)$$

A coevolving profile $\phi \in \mathcal{P}$, is a subset of Comb that does not include any pairs of conflicting combinations. In the case of nucleotide sequences, the size of a profile varies from 2 to 4 and the total number of profiles is 192 ($|\mathcal{P}| = 192$; Supplementary Figure S1; Supplementary Material 1). The frequency of a profile is further calculated as the sum of the combination frequencies composing it. For instance, in the pair *P*3 (Fig. 1), the profile {AC, TT} is of size 2 and has a frequency of 0.8.

Table 1. Illustrative example: properties

	P1	P2	P3	P4
f(TT)	50%	75%	40%	50%
f(AC)	50%	25%	40%	0%
f(AT)	0%	0%	20%	50%
f(TA)	0%	0%	0%	0%
$S_{\text{comp}}(p_1, p_2)$	1	1	0.7	1
$S_{\text{comp}}(p_2, p_1)$	1	1	0.7	0.5
Property of conflicts	0	0	57.14%	100%

Combinations frequency (first four rows), S_{comp} scores (rows 5 and 6) and 'proportion of conflict' (row 7) of the four pairs (P1, P2, P3 and P4). The profile for all of those pairs is {AC, TT}. The proportion of conflicting combinations (Equation (1)) increases as S_{comp} score decreases. Notice that considering two positions p_1 and p_2 of a pair, the score S_{comp} is not symmetric.

2.2 Model of coevolving substitutions

A dependent model of evolution for binary characters can be derived from the standard models of substitutions by extending the dimensionality of the instantaneous rate matrix *Q* and evaluating the likelihood of a pair of positions simultaneously (Pagel, 1994). Using a similar approach, a dependent model for DNA sequences will be based on a *Q* matrix of size 16 × 16 to account for the four states ($\mathcal{A} = \{A, C, G, T\}$) representing the nucleotide alphabet. For amino acid sequences, the instantaneous rate matrix would be of size 400 × 400. Coevolution further posits that a substitution should trigger the change at another position during a coevolution event and that these two events should not happen simultaneously. Consequently, all double substitutions have a rate of 0 (Pagel, 1994). The *Q* matrix of a generalized model of dependent evolution for DNA sequences will thus be composed of 96 non-zero different instantaneous rates (Equation (2)), which will clearly lead to overparameterization of the model. We propose a new way to restrict the number of parameters by incorporating into the dependent model the discrete parameter ϕ representing the coevolving profile. This will reduce the number of parameters to estimate from 96 to 4. There are at most 192 distinct profiles ϕ that can be formed by combining two nucleotide positions. For each of these profiles, a unique 16 × 16 *Q* matrix is constructed by distributing differently the four parameters depending on the combination involved in each profile (Supplementary Figure S4 and Table S1; Supplementary Material 1). Given a profile, ϕ, the instantaneous rate matrix *Q* of our model, Coev, is modeled as follows:

$$Q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by two nucleotide positions,} \\ r1 & \text{if } \{i,j\} \notin \phi \text{ and if } i \text{ differs from } j \text{ at position 1,} \\ r2 & \text{if } \{i,j\} \notin \phi \text{ and if } i \text{ differs from } j \text{ at position 2,} \\ s & \text{if } i \in \phi \text{ and } j \notin \phi, \\ d & \text{if } i \notin \phi \text{ and } j \in \phi \end{cases} \tag{2}$$

The parameter *s* is the rate of transition from a coevolving combination present in the profile to a non-coevolving combination. Conversely, *d* is the rate of transition from one non-coevolving to a coevolving combination (Equation (2); Supplementary Figure S3A; Supplementary Material 1). The additional parameters *r*1 and *r*2 are the rates of transitions between two non-coevolving combinations at positions 1 and 2, respectively (Supplementary Figure S3B; Supplementary Material 1). For simplicity, we considered in the following a Jukes–Cantor (JC) model of substitutions (Jukes and Cantor, 1969), where position 1 evolves under a single rate *r*1 and position 2 evolves under another rate *r*2. This can easily be modified to use any existing substitution models. The rate parameters of the Coev model represent continuous variables potentially taking any positive value (i.e. $s, d, r1, r2 \in [0, \infty]$).

2.3 Maximum likelihood estimation

In a Maximum Likelihood (ML) framework, we want to estimate the probability of a pair of positions *X*, which represents one combination of characters for each species in *S*,

$$\text{Prob}(X|\phi, s, d, r1, r2, \tau, \nu) \tag{3}$$

coevolving under the model Coev along a phylogenetic tree with topology τ and branch lengths ν. For simplicity, we assume that these τ and ν parameters are known and are not estimated during the ML optimization. We use Felsenstein's pruning algorithm (Felsenstein, 1981) to evaluate the likelihood of the model. This is done by calculating, for each branch of a phylogenetic tree, the transition probability matrix $P(t) = e^{Qt}$, where the branch length *t* is a finite time interval. The amount of data in *X* is not sufficient to estimate the frequencies at equilibrium and the model assumes that the frequencies of all combinations are equal (Pagel, 1994).

The parameter ϕ is a discrete parameter and an exhaustive search through the 192 possible profiles of $\mathcal{P}$ is performed to find the profile that best fit the data *X*. For each profile ϕ, the continuous parameters are estimated by numerical integration to obtain the values that maximized the probability of *X* (Equation (3)) using the Nelder–Mead algorithm (Nelder and Mead, 1965).

The Coev model can be tested against an 'independent' model that assumes that the two positions are evolving independently from each other. The fit of the Coev model with respect to the 'independent' model was estimated using Δ Akaike information criterion ($\Delta\text{AIC} = \text{AIC}_{\text{independent}} - \text{AIC}_{\text{Coev}}$). This procedure can be used to assess whether a pair of positions *X* is coevolving by directly indicating if the Coev model better fits the data than the 'independent' one. For example, the pairs 1, 2 and 3 in Figure 1 were coevolving with the profile {AC,TT} (Supplementary Table S7; Supplementary Material 1). In contrast, the Comb set of pair 4 was composed of two conflicting combinations AC and AT and $\mathcal{P} = \emptyset$, which means that the positions forming pair 4 are not coevolving. The examples illustrated by pairs 1, 2 and 3 (Fig. 1) also showed that the ΔAIC scores for Coev are providing similar results than the non-parametric scores of coevolution (here S_{comp}; Table 1; Dip and Carbone, 2012) and that the ΔAIC decreases under weaker coevolution (Supplementary Table S7; Supplementary Material 1). To delimit with confidence the list of coevolving pairs for a given dataset, a distribution of expected ΔAIC was obtained for each dataset by simulating alignments based on the same phylogenetic tree as the original data but evolving the nucleotides under the 'independent' model. The 95th percentile of this expected ΔAIC distribution provided a threshold to consider the observed ΔAIC for Coev to be large enough to be accepted as evidence for coevolution (see Supplementary Material 1 for more details).

2.4 Bayesian implementation

As is commonly done in other evolutionary models, we first tested all combination of pairs *X* in an alignment *S* for coevolution using the ML implementation and then used the Bayesian approach to fully estimate the uncertainty in the parameter estimates of the Coev model for pairs that did coevolve based on the ΔAIC described above (e.g. FitzJohn, 2012). The Bayesian framework used a Markov Chain Monte Carlo algorithm (MCMC) to sample the rate parameters ($s, d, r1, r2$) and estimated the profile ϕ from their posterior distribution. We updated the rate parameters by applying a uniform sliding window on the log scale of the parameters (Ronquist *et al.*, 2009), while we randomly drew the profile based on a uniform prior distribution. It should be noted that while a change of profile forces a reassignment of potentially all rates within the 16×16 instantaneous rate matrix, the number of parameters, and thus the dimension of the model, remains unchanged. We implemented a Metropolis-coupled MCMC (MC^3) algorithm (Altekar *et al.*, 2004) to move across the discrete parameter space defined by the finite number of profiles and to improve the mixing of the chains. Swaps between chains were randomly proposed every fifth generations, and posterior estimates of the parameters were obtained from 1 000 000 MC^3 generations (after the burnin phase), sampling parameters every 1000 generations. We assessed the efficiency of the chain mixing by measuring the effective sample size of the different parameters and by examining the MCMC log files in Tracer (Drummond and Rambaut, 2007). The sampling frequencies of each profile were calculated from the MCMC samples and used as approximations for the respective posterior probabilities. The rate parameters were summarized as mean values and 95% credibility intervals, calculated as highest posterior densities.

2.5 Empirical simulated and illustrative datasets

We analyzed (i) two datasets to compare model predictions on real biological sequences (empirical datasets), (ii) datasets of three different sizes with each 180 simulated positions (simulated datasets) to evaluate the models performance and (iii) two additional datasets to study the intrinsic properties of Coev model (illustrative datasets).

The first empirical dataset was a DNA alignment of the large subunit of the ribulose-bisphosphate carboxylase gene (*rbcL*). The dataset contained 422 species of Poaceae obtained from GPWG2 (2012). We used a larger but taxonomically more focused set of sequences than Wang *et al.* (2011) who looked for co-evolving positions in 142 *rbcL* angiosperms, gymnosperms, ferns and mosses sequences. The second dataset was obtained from Yeang *et al.* (2007) and included 146 sequences of 16S RNA spanning several kingdoms of life (animals, plants, fungi, archea and bacteria). The 16S ribosomal RNA sequence is well known for its coevolving pair 245, 283 constituted of 68 CC combinations and 65 UU combinations across different lineages (Dutheil *et al.*, 2005; Yeang *et al.*, 2007).

For both empirical datasets, we filtered the alignment to remove highly conserved positions (i.e. >90% conservation) as well as all positions with at least one insertion or deletion. We estimated the presence of coevolution on all pairs of positions of the filtered alignments of the two empirical datasets using the ML implementation of Coev. The coevolving positions identified by ML were subsequently analyzed with the Bayesian implementation to evaluate the posterior probabilities of the coevolving profiles. For comparison, we also ran two non-parametric methods (score of comparison, S_{comp}; Dib and Carbone, 2012) and Mutual Information (MI; Gloor *et al.*, 2005) and two parametric models: CO (Yeang *et al.*, 2007) and CoMap.

To further evaluate the performances of our new model, we simulated three datasets of variable size (33, 67 and 110 species). Tree topologies were randomly created and the branches of the phylogenetic trees were randomly drawn form an exponential distribution with $\lambda = 0.5$. The coevolving positions were created by simulating convergent codons in different lineages differing by a single nucleotide (e.g. Methionine ATG to Lysine AAG) in the coevolving lineages following the approach used by Christin *et al.* (2012). The second position of each codon was then kept. This has the advantage to use a model of evolution (here codon model of substitution) that is different from Coev, while creating a pattern of nucleotides that mimics coevolution. It will thus not favor our model over any alternative approaches to measure coevolution. For each dataset, we simulated 20 coevolving positions and concatenated these positions with 160 independently evolving positions that were simulated without forcing coevolution. Each of these datasets resulted in the simulation of 190 pairs of coevolving positions ($(20{*}20 - 20)/2$) and 15 920 non-coevolving pairs and allowed us to evaluate each method using standard performance measures. Specifically, we estimated the number of positions correctly predicted as coevolving (true positives, TP), the number of residues correctly predicted as non-coevolving (true negatives, TN), the number of non-coevolving positions predicted as coevolving (false positives, FP) and the number of coevolving residues predicted as non-coevolving (false negatives, FN). From these measures, we calculated the sensitivity TP/(TP + FN), specificity TN/(TN + FP), accuracy (TP + TN)/(TP + FN + TN + FP) and positive predictive value TP/(TP + FP) for each method tested. We compared the performance of Coev to CoMap, S_{comp} and CO methods. Coev performance was evaluated by considering pairs with ΔAIC values higher than the 95th percentile value of the ΔAIC distribution issued from independently evolving positions. The performance of CoMap was evaluated by considering coevolving pairs with a stat score >0.75 and a *P*-value <0.05, while the performance of S_{comp} was evaluated by considering pairs with either a score equal to 1 (SComp I) or the 1% top scores (SComp II) as coevolving. For the CO parametric model, we considered pairs with a likelihood ratio >6 log units as coevolving Yeang *et al.* (2007).

In the Supplementary Material 1, we also described two illustrative datasets (S1 and S2) specifically designed to highlight the intrinsic properties of the model. Those datasets were designed to explore the properties of the Coev model and in particular to assess the effect of conflicting combinations on the model.

The dataset S1 was composed of five pairs of positions. The first pair, $P1(S1)$, had a $\text{Comb}_{P1(S1)}$ size equal to two ($\text{Comb}_{P1(S1)} = \{AA, TT\}$). The number of occurrences of each combination is reported in Supplementary Table S3 (Supplementary Material 1). The four other pairs were created by adding each time one new conflictual combination to the previous pair. The dataset S2 was built with the same rationale but starting with a pair of positions with a complex profile of coevolution ($\text{Comb}_{P1(S2)} = \{AA, CC, GG, TT\}$).

3 RESULTS

3.1 Ribulose-bisphosphate carboxylase gene

We ran our ML implementation on the 74 filtered positions of *rbcL* and found, out of the 2701 possible pairs tested, that Coev was preferred over the ‘independent’ model for a total of 103 pairs of positions. The three coevolving nucleotide pairs (positions 401, 950 and 1058 in the alignment) that obtained the best ΔAIC belong to codon 133 of the N-terminal domain, and codons 326 and 362 of the C-terminal domain. These positions are not in close proximity in the DNA sequence. However, they form a triplet that is in direct contact in the 3D crystallized structure of *Oryza sativa* RuBisCO protein (pdb 1WDD; Supplementary Figure S12; Supplementary Material 1). The three amino acids are linked to the binding site of the CAP A 1001 substrate, suggesting their involvement in the allosteric movements of the RuBisCO (Lockless and Ranganathan, 1999). The ΔAIC ranged from 6.65 to 171.53 log units. The nucleotide pair 950–1058 had the highest ΔAIC value for the coevolving profile {AA, GG} (Supplementary Table S5; Supplementary Material 1). The Bayesian analysis over this pair of positions confirmed that {AA, GG} is the profile with the highest posterior probability (100%).

The highest MI score obtained over the 2701 possible pairs of the *rbcL* gene was 0.9 (Gloor *et al.*, 2005). The MI score could, in theory, vary between 0 and 2 and scores below 0.9 indicates that none of the pairs are coevolving (Supplementary Material 2). High MI score are obtained when the frequencies of the non-conflicting combinations are homogeneous, which is not the case for the 103 pairs of the *rbcL* gene identified by our Coev model. For instance, the AA combination frequency for the pair 950–1058 was 363 out of 422, which lead to a MI score of 0. Further, none of the 103 pairs had a S_{comp} score >0.86 and the S_{comp} score associated with the pair 950–1058 was intermediate (0.54 and 0.51). This is certainly due to the fact that the pair was partially composed of conflicting combinations (Comb = {AA, AG, GA, GG}). The pair 656–818 had the highest S_{comp} value (0.86) and the ΔAIC of the Coev model was not the highest (35.98). When we looked closely at the combinations associated with this pair, we observed that the AC combination had the highest frequency (70% or 297 species out of 422) and the combination with the second highest frequency was TC (17%). The two most frequent combinations were conflicting combinations and could not be the profile of coevolution. However, this pair of positions had seven possible profiles:{TC, AA}; {TC, AT}; {AC, TA}; {AA, CC}; {CC, TA}; {CC, AT}; {AT, TA}. Among those, {TC, AA} had the highest ΔAIC score for the Coev model (35.98) and {AT, TA} had a comparable ΔAIC score (34.71). The two profiles had no common combination and the *s* or *d* estimated values for each were similar ($s = 0.06$, $d = 1.66$ for {TC, AA} and $s = 0.07$, $d = 1.62$ for {AT, TA}).

The CoMap model predicted 164 pairs of co-evolving positions with a *P*-value <0.05. However, none of the predicted pairs had a stat score that exceeded 0.75 (the 164 stat scores vary from 0.10 to 0.56) and 12 out of the 164 co-evolving pairs are also predicted by Coev. Wang *et al.* (2011) originally reported that about half of the sites of the *rbcL* gene are co-evolving whereas the new analysis (using the 422 sequences filtered dataset) showed that ∼40% (30 out of 74) of the sites are coevolving when using CoMap. This difference is likely due to the dependency of CoMap to the number of species (Fig. 2).

Finally, the CO model predicted 15 pairs of coevolving positions. The three coevolving pairs with the highest ΔAIC were not identified by CO. However, the pairs 61–407 identified by CO with the highest score (19 log units) was predicted by Coev among the 103 list of pair of positions preferred over the 'independent' model. This pair was mainly composed of two conflicting combinations CT and CG and its estimated profile was {CT, TA}.

3.2 16S ribosomal RNA

The 16S ribosomal RNA empirical dataset was first used to analyze the well-known coevolving pair of positions 245–283, which have been found as coevolving under the Watson–Crick profile {CC,UU} by both structural and experimental analyses (Cannone *et al.*, 2002). ML analyses showed that the Coev model for this pair had a ΔAIC = 53.71 (Supplementary Table S6; Supplementary Material 1), which confirmed that {CC,UU} was the best profile of coevolution for this pair.

Over all the 23 005 possible pairs that can be tested with the 16S ribosomal RNA dataset, we found, however, that the best coevolving pair was not the positions 245–283, but rather the positions 1950–2017 (ΔAIC of 164.72). Using the 95th percentile of the expected ΔAIC distribution as the threshold of coevolution (ΔAIC = 1.77; see Supplementary Material 1), 1008 pairs were selected as coevolving. Among those, the top 3% of the pairs with the highest ΔAIC displayed profiles of different complexity: 12 pairs had a profile of size 2, 12 pairs had a profile of size 3, five pairs had a profile of size 4. Among those 29 pairs, 25 showed a typical Watson–Crick profile. However, the pair with the highest ΔAIC had a complex Watson–Crick profile {AT, CG, GC, TA} with repeated occurrences of each combination across lineages of the phylogenetic tree (Supplementary Table S7; Supplementary Material 1).

Several of the 29 pairs identified by Coev were also identified by S_{comp} (pairs 1970–2011, 1950–2017, 4410–4420, 4429–4439, 1948–2020) or MI (pairs 1950–2017, 2065–3569, 1556–1592, 1951–2016, 3384–3407, 4411–4419). Nevertheless, no correlation was observed between the ΔAIC values and S_{comp} or MI scores (Pearson correlation = 0.19 and 0.1, respectively). We looked more closely at the pair 1948–2020, whose proportion of conflict was 0 and S_{comp} score was maximal, but whose ΔAIC for Coev was not the highest (52.05 log units). Its ΔAIC was not the highest, it was still larger than the threshold of coevolution defined for this empirical dataset and its estimated profile was {CG, GC, TA}. The GC combination occurred in all bacterial, archae and protist lineages, as well as the two eukaryote organelles. In contrast, all multicellular eukaryotes had the TA combination, except two fungi (*Fellomyces ogasawarenisis* and *Bullera huiaensis*) and one animal (*Strongylocentrotus intermedius*), which retained the ancestral CG combination (Supplementary Figure S10; Supplementary Material 1). The double substitution between TA and CG was thus specific to the multicellular eukaryotic lineages and, although it is not possible to date precisely the acquisition of the TA combination, it suggests a local coevolution within the phylogenetic tree in the early evolution of the multicellular eukaryotes. The Coev model is thus sensitive to local coevolution patterns and, in contrast to other methods, can distinguish between global and local coevolving positions.

We further validated our predictions by localizing the coevolving positions on the 2D and 3D structure available for the 16S subunit (pdb 2AVY) of *Escherichia coli*. We found that 24 of the 29 pairs identified by our model were connected in the 2D structure (Supplementary Figure S13 (left); Supplementary Material 1) and 26 of the 29 pairs were connected in the 3D structure (Supplementary Figure S13 (right); Supplementary Material 1). In contrast, only five and six of the 15 pairs of positions predicted by S_{comp} demonstrate direct contact of the nucleic acid pairs in the 2D and 3D structure, respectively. For CO, Yeang *et al.* (2007) reported that 15 of the 41 predicted pairs are in direct contact in the 3D structure.

3.3 Simulated and illustrative datasets

The Coev model outperformed other parametric and nonparametric methods for the three simulated datasets of size 110, 67 and 33 sequences (Fig. 2; Supplementary Table S9; Supplementary Material 1). In particular, the sensitivity was found to be consistently higher in the simulations suggesting that the Coev model can reliably detect true coevolving positions (i.e. all of the 190 co-evolving pairs simulated for each of datasets). The sensitivity of Coev did not appear to be affected by the size of the tree whereas a substantial decrease was observed in S_{comp} and CoMap (9- and 4-fold, respectively) with increasing tree size (Fig. 2; Supplementary Table S9;

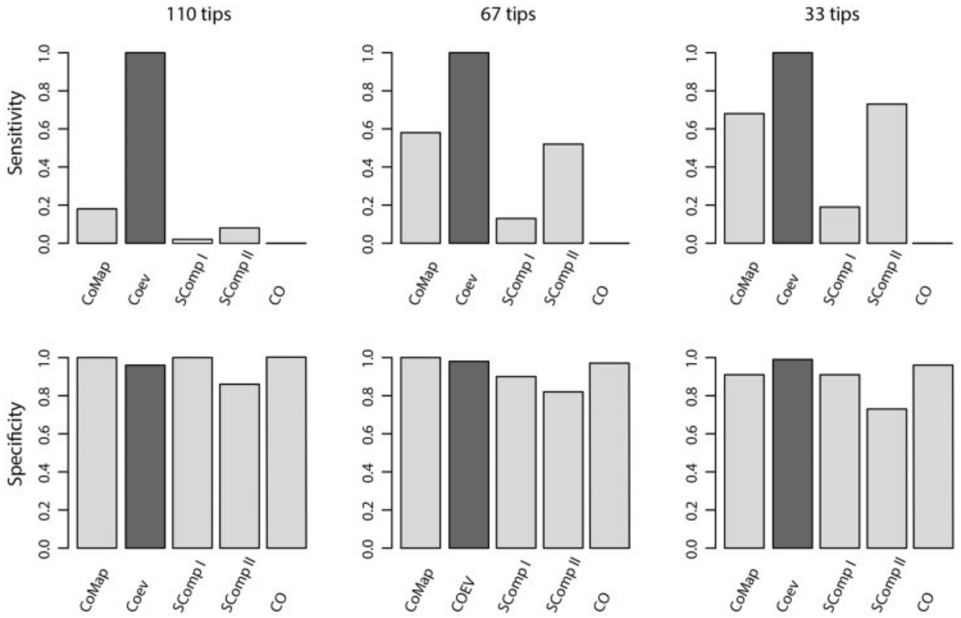


Fig. 2. Performance: sensitivity and specificity plots. Sensitivity and specificity plots comparing the performance of the Coev model to CoMap, S_{comp} and CO. For each dataset (110, 67 and 33 tips), we simulated an alignment of 180 positions (20 coevolving positions and 160 independently evolving). SComp I is the performance of S_{comp} method when only maximal scores of 1 are considered, whereas SComp II considered the 1% top scores. CoMap's performance where evaluated when considering coevolving pairs with a stat score >0.75 and a P-value <0.05. While the specificity does not strikingly change across datasets and methods, the sensitivity is higher for the Coev when compared to the other methods

Supplementary Material 1). Noticeably, the CO model failed to recover the true co-evolving positions in all datasets ($TP = 0$). These observations were previously shown by Dib and Carbone (2012).

Moreover, two illustrative datasets (S1 and S2) were used to describe intrinsic properties of the Coev model (Supplementary Material 1). The analysis of S1 and S2 showed a strong negative correlation of -0.994 between the proportion of conflicts and the ΔAIC values. Furthermore, the simulations revealed that the s/d ratio can be seen as a measure of the coevolution strength (Supplementary Figure S7; Supplementary Material 1) since s/d ratio decreased as the signal of coevolution became stronger. In the perfectly coevolving case, s/d tended towards zero (Supplementary Table S8; Supplementary Material 1), whereas this ratio increased until the two parameters became equal when the pairs of positions evolved independently. Further, the $r1$, $r2$ parameters do not incorporate information about coevolution since their values can fluctuate from zero to infinity without any correlation with the ΔAIC of Coev (Supplementary Table S8; Supplementary Material 1).

We also designed an experiment (Supplementary Material 1) to show that, contrary to the S_{comp} score, Coev can distinguish between coevolving and co-inherited pairs of positions (Supplementary Figure S9; Supplementary Material 1). A co-inherited pair is defined as a combination acquired only once during the evolution and further inherited by all the descendants. Our results showed that the ΔAIC difference increased with the number of occurrences of the combination in the phylogenetic tree (Supplementary Figure S9; Supplementary Material 1).

4 DISCUSSION

We presented a new dependent model of evolution, Coev, based on the broad class of Markov models. Coev can reliably identify the coevolving pairs of nucleotide positions and their molecular footprint, expressed by the profile of coevolution. The Coev model incorporates free parameters that are easily interpreted from an evolutionary perspective and that govern the change of one combination into another within pairs of coevolving positions. Our model is mechanistic and describes the process of evolution that is responsible for the simultaneous evolution of dependent sites along DNA or amino acid sequences. The profile of coevolution is explicitly incorporated in the model as a discrete parameter and thus can be estimated from the data. It is an essential parameter in our model since it shapes the instantaneous rate matrix at the base of Coev (Supplementary Figure S4; Supplementary Material 1) and defines the number of occurrences of the rate parameters s, d, $r1$ and $r2$ (Supplementary Table S1; Supplementary Material 1). The parameter s is the rate of transition from a coevolving combination to a non-coevolving one. This parameter thus represents the rate of a necessary step (yielding conflict) that precedes the shift to a new coevolving combination. A new substitution is then required to re-establish the coevolving profile, and this event is modeled by the parameter d. This parameter d therefore represents the rate of transition from a non-coevolving combination to a coevolving one. Parameters s and d describe the two temporally successive substitutions that are the minimum requirement to shift from one coevolving combination to another in the profile. The additional parameters $r1$ and $r2$ represent the rates of independent substitution between non-coevolving combinations.

Conflictual substitutions in each site are modeled by a single independent rate regardless of the type of substitution, e.g. transition/transversion, thus resembling a JC model. The model can however be extended to incorporate more complex nucleotide substitution models such as the GTR (Tavare, 1986).

4.1 Coev profile and influence of the phylogenetic tree

Our approach can provide a better understanding of the evolutionary forces shaping pairs of coevolving positions by estimating the profile that best fits the coevolving pair using a probabilistic model of dependent nucleotide evolution. We implemented our model in a Bayesian framework that further provides the posterior probability associated with each profile for a given pair of coevolving positions. The data-driven estimation of the profile of coevolution is novel and expands the potential of coevolution analyses by allowing us to identify patterns of coevolution without restricting itself to known profiles, such as Watson–Crick constraint for RNA sequences.

Based on illustrative and empirical datasets, we found that the best fitting profile is not necessarily the most frequently observed one, especially in the presence of conflicts (e.g. $P5(S1)$ in Supplementary Table S8; Supplementary Material 1). Further, the selection of the profile that best describes the coevolution at two positions depends on the associated phylogenetic tree. Using the Coev model, we showed that the coevolving profiles in ribosomal RNA are not necessarily Watson–Crick profiles but can rather involve complex patterns with, for example, three or four combinations. Additionally, the $\{$CC, UU$\}$ profile prediction for the pair 245, 283 in the 16S ribosomal RNA sequence identified by other methods (Dutheil *et al.*, 2005; Yeang *et al.*, 2007), was estimated by Coev without assuming any *a priori* weighting based on physico-chemical properties of the nucleotides. This is certainly a strength of our approach as it allows the estimation of coevolution due to other less known constraints or to extend these analyses to protein coding genes that are not affected by Watson–Crick constraint.

4.2 Coev model compared to other models

Some of the most widely used methods to estimate coevolution, such as MI (Gloor *et al.*, 2005), SCA (Lockless and Ranganathan, 1999) and ELSC (Yip *et al.*, 2008), do not incorporate the evolutionary history of the gene under consideration and any random assignment of sequences along the underlying tree produces the same score of coevolution. Other methods do use the topology of the phylogenetic tree to estimate local scores of coevolution for each node before merging them hierarchically to obtain a single score for the whole tree (Dib and Carbone, 2012). We compared the performance of Coev to available methods and showed that it outperforms existing approaches especially when the number of sequences in the alignment increases (Fig. 2). Moreover, available parametric and non-parametric methods do not attempt to capture the long-term evolutionary process shaping DNA sequences, but rather focus on the product of this process only. This misses important information, such as selective pressure and evolutionary constraints, which are key elements to explain how and why a pair of positions can coevolve. Few attempts have been made to model coevolution directly, but they all differ in several aspects from our Coev model.

First, the model used in the CoMap method was an attempt to capture the process of coevolution in DNA sequences (Dutheil and Galtier, 2007). It used a Bayesian approach to map the substitutions that occurred at each site independently onto the branches of the underlying phylogenetic tree. Mutations occurring at two sites are thus not correlated, which violates the specific assumption of coevolution (Bollback, 2005; Huelsenbeck *et al.*, 2003). This could explain the lower sensitivity observed in our simulations for CoMap and highlight the necessity to model the precise process of coevolution to be able to correctly predict correlated pairs of positions. Several studies used CoMap to analyse the co-evolving positions within and among genes (Corbi *et al.*, 2012; Wang *et al.*, 2011). For instance, Wang *et al.* (2011) post-processed CoMap predictions to learn about the combinations properties found in the co-evolving amino-acid positions of the *rbcL* dataset. They looked for the most frequent combination in the co-evolving positions and described the bio-chemical properties of the amino acids composing the co-evolving positions. However, we showed using Coev that the most frequent combination is not necessarily part of the co-evolutionary profile and Coev is now able to estimate the profile of co-evolution along a phylogentic tree. This profile provides the biochemical properties and evolutionary constraint associated with a pair of positions. Second, a Markov model that, like Coev model, used an instantaneous rate matrix to model the coevolution of pairs of positions has also been proposed (Yeang *et al.*, 2007, CO model). We compared the CO model with Coev on the *rbcL* empirical data and showed that CO was not able to capture the same coevolving pairs of positions and more specifically the pair 950–1058 where AA and GG combinations appear in several lineages along the tree. The CO model accounts for non-independent evolution of the pair of positions by specifying rate parameters for double substitutions and down-weighting the single substitution rates. This has the effect of penalizing single changes, but the model still allows simultaneous substitutions within a small unit of time. The latter assumption is in contradiction with a coevolution process, where a substitution should trigger the change at another position (Pagel, 1994). This contradiction could explain the lack of sensitivity of CO in our simulations, which confirmed previous studies (Dib and Carbone, 2012). Further, the different assumptions made by the CO and Coev models, might explain the discrepancies in the prediction of coevolving sites for the *rbcL* dataset.

Finally, the dependent model proposed by Pagel (1994) for phenotypic data was modified and adapted for protein data (Pollock *et al.*, 1999), but this has not been extended to the full alphabet of the sequences at hand. Instead, a categorization of the physico-chemical properties of the amino acid was used to create an alphabet of size 2 (e.g. positively and negatively charged residues; large and small residues), and the dependent model of Pagel (1994) was applied directly. The reduction of the original amino acid alphabet into two categories has the advantage of assessing potential important characteristics of the protein, but multiple categorization will be necessary to test which physico-chemical property is the most pertinent for the coevolution. It is also possible that different lineages could present coevolution between different physico-chemical properties for the same pair of positions depending on the DNA-sequence analyses. In addition, the dismissal of the original alphabet will not allow the characterization of the evolutionary process that has lead to the binary coevolving pattern.

The Coev model is conceptually different from the approaches that develop dependent models of evolution. Our model can describe the process of evolution of a coevolving pair, estimate the associated profile, reconstruct the ancestral states of dependent positions and provide the probability vector for several ancestors (see Supplementary Figure S8; Supplementary Material 1). We have shown as well that the phylogenetic tree is an essential aspect of the coevolution process and that the estimation of the profile, but also the prediction of coevolving positions, can be affected by the underlying phylogenetic tree.

4.3 Global versus lineage specific coevolution

The Coev model assumes that a pair of positions coevolve under the same evolutionary process along the whole phylogenetic tree. It is however likely that during the acquisition of a new function, a gene will be under different selective constraints in different lineages. This will create branch-specific coevolving positions in the phylogenetic tree that none of current methods are able to account for. For example, the evolution of genotypic convergence has been documented recently (Castoe *et al.*, 2009; Christin *et al.*, 2007) and it illustrates that different processes can take place in specific lineages. For example, different codons are used in the different lineages to create the convergent functional protein in the C4 grasses. Thus, it will be important in future development to define an evolutionary framework able to detect these lineage-specific constraints, especially as gene-tree estimation can be biased due to constraints or selective pressures (Christin *et al.*, 2012).

Additionally, the Coev model can help to distinguish coevolving from co-inherited pairs of positions, which are defined as combinations acquired once in the evolution of a lineage and inherited by all its descendants. The pair of positions 1948–2020 of the 16S ribosomal RNA family is an example of a co-inherited position acquired at the origin of the multicellular organisms and further lost only in few species. None of the available methods consider the number of times a combination is acquired along the tree when assessing a score of coevolution (Dutheil, 2012), whereas in the Coev model we observed higher ΔAIC for pairs whose combinations have been acquired multiple times in different lineages. This is the case even though our model does not explicitly count the number of acquisitions (Supplementary Material 1). However, establishing whether the pair 1948–2020 is coevolving or co-inherited is still an open question. One difficulty comes from the current definition for co-inheritance, which is simplistic and assumes that a coevolving pair is necessarily coevolving under a simple profile (i.e. of size 2; Dutheil, 2012). However, the pair 1948–2020 coevolved under a complex profile that presents both co-inherited and coevolving combinations. The substitutions of GC to TA combination were co-inherited by the ancestor of the multicellular organisms. However, the CG combination was later acquired again several times and can be considered as coevolution (Supplementary Figure S10; Supplementary Material 1). This example clearly shows the difficulty to assess if two positions are coevolving or co-inherited and the definition of co-inheritance given by Dutheil (2012) should be revisited in order to take into account complex profiles.

Lineage-specific constraint and co-inherited pairs affected our model since their associated ΔAIC was weaker than coevolving pairs with combinations acquired several times in the phylogenetic tree. One possible move forward to assess locally coevolving pairs would be to extend the Coev model and allow the rate parameters and coevolving profiles to vary across the branches of a phylogenetic tree. There is also the need to better assess the links existing between coevolution and the well-known selective pressures affecting molecular sequences. This is an important area to develop and it can help understand better protein signatures that are used to identify the function and the role of newly sequenced proteins. For instance, one could look for Hoogsteen pair of positions that interact in the tertiary structures of RNA sequences but that do not necessarily evolve simultaneously and present conflict in the observed data.

5 CONCLUSION

We presented a new, fully mechanistic, model that describes the processes governing the coevolution of a pair of positions. Conserved sites have been extracted from gene families for more than 40 years (Asthana *et al.*, 2007), but we propose that the evolutionary profile of coevolving sites should be added to these known signatures. This will help the community to classify highly divergent sequences and better interpret the function of new ones.

ACKNOWLEDGEMENT

We would like to thank W. Hordijk for the optimization of the C code. This work received support from the Vital-IT facilities and from the Swiss Institute of Bioinformatics.

Funding: University of Lausanne; Swiss National Science Foundation (grant 31003A-138282). DS received funding from the Wenner-Gren Foundation (Sweden).

Conflict of Interest: none declared.

REFERENCES

Altekar,G. *et al.* (2004) Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, **20**, 407–415.

Asthana,S. *et al.* (2007) Analysis of sequence conservation at nucleotide resolution. *Plos Comput. Biol.*, **3**, e254.

Baussand,J. and Carbone,A. (2009) A combinatorial approach to detect coevolved amino acid networks in protein families of variable divergence. *Plos Comput. Biol.*, **5**, e1000488.

Bollback,J.P. (2005) Posterior mapping and posterior predictive distributions. In: Nielsen,R. (ed.) *Statistical methods in molecular evolution*. Springer, New York, pp. 439–462.

Boussau,B. *et al.* (2004) Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc. Natl Acad. Sci. USA*, **101**, 9722–9727.

Cannone,J.J. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinform.*, **3**, 2.

Carbone,A. and Dib,L. (2011) Co-evolution and information signals in biological sequences. *Theor. Comput. Sci.*, **412**, 2486–2495.

Castoe,T.A. *et al.* (2009) Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl Acad. Sci. USA*, **106**, 8986–8991.

Chockalingam,K. *et al*. (2005) Directed evolution of specific receptor - ligand pairs for use in the creation of gene switches. *Proc. Natl Acad. Sci. USA*, **102**, 5691–5696.

Christin,P.A. *et al*. (2007) C4 Photosynthesis evolved in grasses via parallel adaptive genetic changes. *Curr. Biol.*, **17**, 1241–1247.

Christin,P.A. *et al*. (2012) Effect of genetic convergence on phylogenetic inference. *Mol. Phylogenet. Evol.*, **62**, 921–927.

Codoñer,F.M. and Fares,M.A. (2008) Why should we care about molecular coevolution? *Proc. Natl Acad. Sci. USA*, **102**, 5691–5696.

Cohen,O. *et al*. (2013) CoPAP: coevolution of presenceabsence patterns. *Nucleic Acids Res.*, **41**, W232–W237.

Corbi,J. *et al*. (2012) Accelerated evolution and coevolution drove the evolutionary history of AGPase sub-units during angiosperm radiation. *Ann. Bot-London.*, **109**, 693–708.

Csuros,M. (2005) Likely scenarios of intron evolution. In: McLysaght,A. and Huson,D. (eds) *Comparative Genomics*. Spring, Berlin, pp. 47–60.

Dib,L. and Carbone,A. (2012) Protein fragments: functional and structural roles of their coevolution networks. *Plos One*, **7**, e48124.

Drummond,A.J. and Rambaut,A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, **7**, 214.

Dutheil,J.Y. (2012) Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief. Bioinform.*, **13**, 228–243.

Dutheil,J. and Galtier,N. (2007) Detecting groups of coevolving positions in a molecule: a clustering approach. *BMC Evol. Biol.*, **7**, 242.

Dutheil,J. *et al*. (2005) A model-based approach for detecting coevolving positions in a molecule. *Mol. Phylogenet. Evol.*, **22**, 1919–1928.

Dutheil,J.Y. *et al*. (2010) Base pairing constraints drive structural epistasis in ribosomal RNA sequences. *Mol. Phylogenet. Evol.*, **27**, 1868–1876.

Fares,M.A. and Travers,S.A. (2006) A novel method to detect intra-molecular coevolution: adding a further dimension to selective constraints analyses. *Genetics*, **173**, 9–23.

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

FitzJohn,R. (2012) Diversitree: comparative phylogenetic analyses of diversification in R. *Meth. Ecol. Evol.*, **3**, 1084–1092.

Fitzpatrick,J.L. *et al*. (2012) Male contest competition and the coevolution of weaponry and testes in pinnipeds. *Evolution*, **66**, 3595–3604.

Franceschini,A. *et al*. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D80–D815.

Gloor,G.B. *et al*. (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, **44**, 7156–7165.

Gobel,U. *et al*. (2004) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.

GPWG2. (2012) New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *New Phytol.*, **193**, 304–312.

Hao,W. and Golding,G.B. (2006) The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res.*, **16**, 636–643.

Huelsenbeck,J.P. *et al*. (2003) Stochastic mapping of morphological characters. *Syst. Biol.*, **52**, 131–158.

Jukes,T. and Cantor,C. (1969) Evolution of protein molecules. In: Munro,H.H. (ed.) *Mammalian protein metabolism*. Academic Press, New York, pp. 21–132.

Lartillot,N. and Philippe,H. (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, **21**, 1095–1109.

Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.

Mirkin,B.G. *et al*. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.*, **3**, 2.

Nelder,J. and Mead,R. (1965) A simplex method for function minimization. *Computer J.*, **7**, 308–313.

Pagel,M. (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. B. Soc. B.*, **255**, 37–45.

Pollock,D.D. *et al*. (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.*, **287**, 187–198.

Ronquist,F. *et al*. (2009) A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. In: Lemey,P. and Vandamme,I.A-M. (eds) *The Phylogenetic Handbook*. 2nd edn. Cambridge University Press.

Tavare,S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura,R.M. (ed.) *Lectures on Mathematics in the Life Science*. American Mathematics Society, Providence, pp. 57–86.

Wang,M. *et al*. (2011) Coevolution of amino acid residues in the key photosynthetic enzyme Rubisco. *BMC Evol. Biol.*, **11**, 266.

Westhof,E. and Fritsch,V. (2000) RNA folding: beyond Watson–Crick pairs. *Structure*, **8**, R55–R65.

Yeang,C.H. *et al*. (2007) Detecting the coevolution of biosequences–an example of RNA interaction prediction. *Mol. Biol. Evol.*, **24**, 2119–2131.

Yip,K.Y. *et al*. (2008) An integrated system for studying residue coevolution in proteins. *Bioinformatics*, **24**, 290–292.