PNAS
www.pnas.org

Supplementary Information for

Simultaneous Bayesian inference of phylogeny and molecular coevolution

Meyer X., Dib L., Silvestro D. and Salamin N.

Xavier Meyer.
E-mail: xavier.meyer@berkeley.edu

This PDF file includes:

Supporting Information Text

Efficiently exploring the pairs space ($k > 0$).

Drawing pairs and profiles. Sampling the parameters ρ_k representing the sites being coevolving and their profile is extremely challenging because there is an extremely high number of possible combinations. Indeed, up to $\frac{N(N-1)}{2}$ different pairs can be derived from a sequence of N nucleotides (quadratic growth). Additionally, the profile of coevolution must be sampled for each of the k pairs among 192 possible existing profiles. Using uniform proposal to blindly propose parameter values for ρ_k would thus result in a very inefficient sampler with a high rejection rate for such moves due to the size and complexity of this parameter space.

Instead of using uniform proposal, we construct proposals based on scores representing how likely it is for a pairs of sites to coevolve. These scores are computed once prior to a MCMC analysis using inexpensive statistical computations. These computations require that we identify the number of profiles that can be expressed for any given pair of sites (i, j). For each pair and observable profile, we then count the number of observed data in the alignment at sites (i, j) that are not included in the profile. For instance given a profile defined by AA and CC, any pairs of nucleotides observed at positions i and j of the alignment other than AA and CC would be considered as conflicting with the profile and thus counted.

Using the number of profiles per pair of positions and the number of conflicts per profile, we defined a score $S(i, j)$ favoring pairs having few profiles, few conflicts for their best profile and a significant variance in the pairs found that are within the profile (SI Appendix, Algorithm S1). The first component ($S1$) of the score penalizes pairs of position having many possible profiles, the second, ($S2$), penalizes pairs showing many conflicts and the last, ($S3$), penalizes co-conserved pairs. All the components are then scaled such as to fit within the $[0, 1]$ interval and then weighted for importance.

The second score $R(i, j, \phi)$ rates each observed profile in function of its conflicts count: the less being the best. This score is normalized and can be used to obtain a discrete probability distribution $p(\phi|i, j)$ from which profiles ϕ can be independently drawn. Similarly, the score $S(\cdot, \cdot)$ is employed to define the probability of a position i, or pair of positions (i, j), given the set of available independent sites S_{indep}. We define the probability of drawing a position i as

$$p(i|S_{indep}) = \frac{max_j(S(i,j))}{\sum_{l \in S_{indep}} max_j(S(l,j))}. \quad [1]$$

Given a position A and the set of independent sites, we define the conditional probability of drawing a pair containing i as

$$p((i,j)|S_{indep}, i) = \frac{S(i,j)}{\sum_{j \in S_{indep}} S(i,j)}. \quad [2]$$

Frequency of proposals. In CoevRJ, proposals are randomly selected at each iteration according to a weighted probability distribution. Transdimensional moves are proposed with frequencies defined as a function of k. These changes in proposal probabilities must therefore be accounted for in the acceptance ratio. Frequencies affected by k are $f_+(k)$ for the creation of pairs, $f_-(k)$ for the deletion, f_{swap1} for the swap of a pair with an independent position and f_{swap2} for the swap between two coevolving pairs.

Swapping moves are applied with a constant frequency f_{cst} with respect to k. Proposals moving from M_k to M_{k+1} have a frequency decreasing as k increases and given by $f_+(k) = 0.99 - 0.98 \cdot k/K_{max}$. The reciprocal proposals moving from M_{k+1} to M_k have a frequency increasing with the number of pairs k, and is given by $f_-(k) = 0.1 + 0.98 \cdot k/K_{max}$. Finally, these frequencies must be adapted for the boundary cases and are given by

$$\begin{cases} k = 0 & f_+(k), f_-(k) = f_{swap1} = f_{swap2} = 0 \\ k = 1 & f_+(k), f_-(k), f_{swap1} = f_{cst}, f_{swap2} = 0 \\ k = K_{max} & f_+(k) = 0, f_-(k), f_{swap1} = 0, f_{swap2} = f_{cst} \\ else & f_+(k), f_-(k), f_{swap1} = f_{cst}, f_{swap2} = f_{cst} \end{cases}$$

The probability of applying a proposal x having frequency f_x is then defined as

$$Pr(x|k) = \frac{(1 - f_{cst}) \cdot f_x}{f_+(k) + f_-(k) + f_{swap1} + f_{swap2}}$$

These frequencies are changing with k and are not symmetrical when moving through the space of models. Therefore, equation 3 (main article) must account for these probabilities by conditioning the probability of a move defined as $q(M)$ and $q(M')$ with its probability $Pr(x|k)$.

1. Supplementary figures

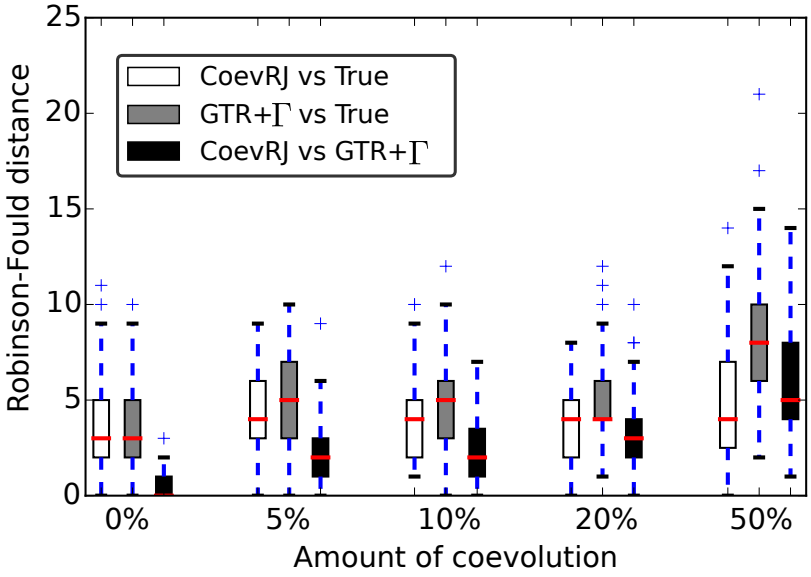
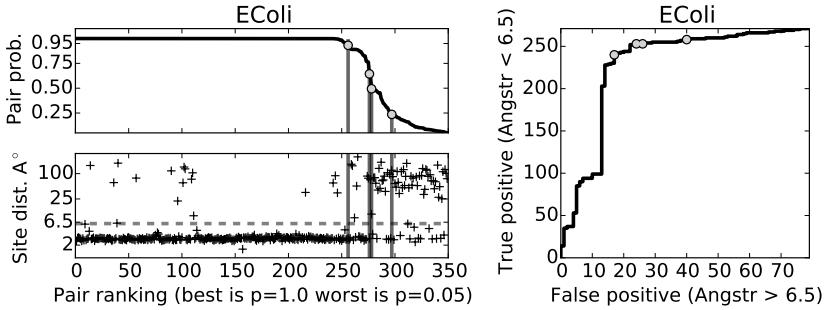
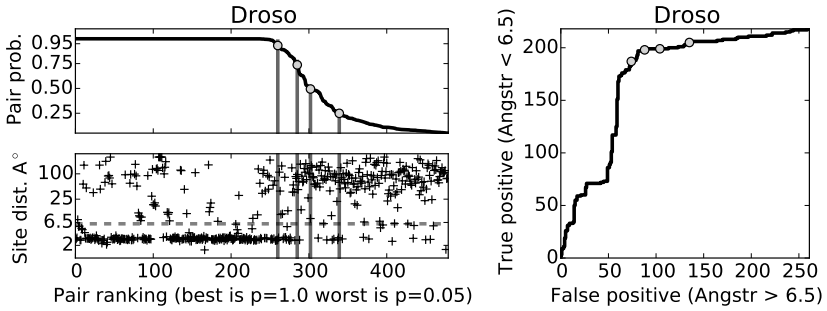


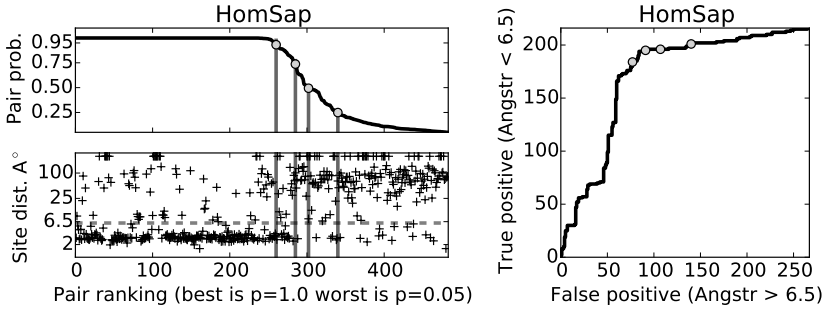
Fig. S1. Robinson-Foulds distances (Robinson & Foulds 1981) between the trees topologies inferred with CoevRJ, GTR+Γ and the one employed for the simulations (True).



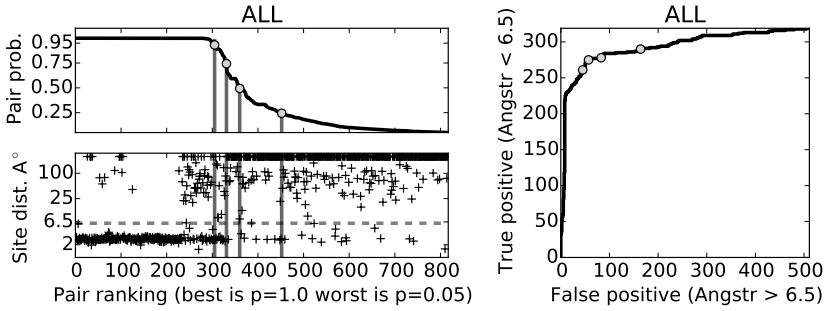
(a) *E. Coli*



(b) *D. Melanogaster*



(c) *H. Sapiens*



(d) All three structures: *E. Coli*, *D. Melanogaster* and *H. Sapiens*

Fig. S2. Validation of the pairs predicted as coevolving by CoevRJ on various structures. Only pairs that map on the structure are reported in figure (a), (b) and (c). The left figures reports the pairs probability against distance in the RNA structure. The right figures reports amount of pairs predicted as coevolving but unconfirmed by the structure against the ones confirmed and predicted as coevolving.

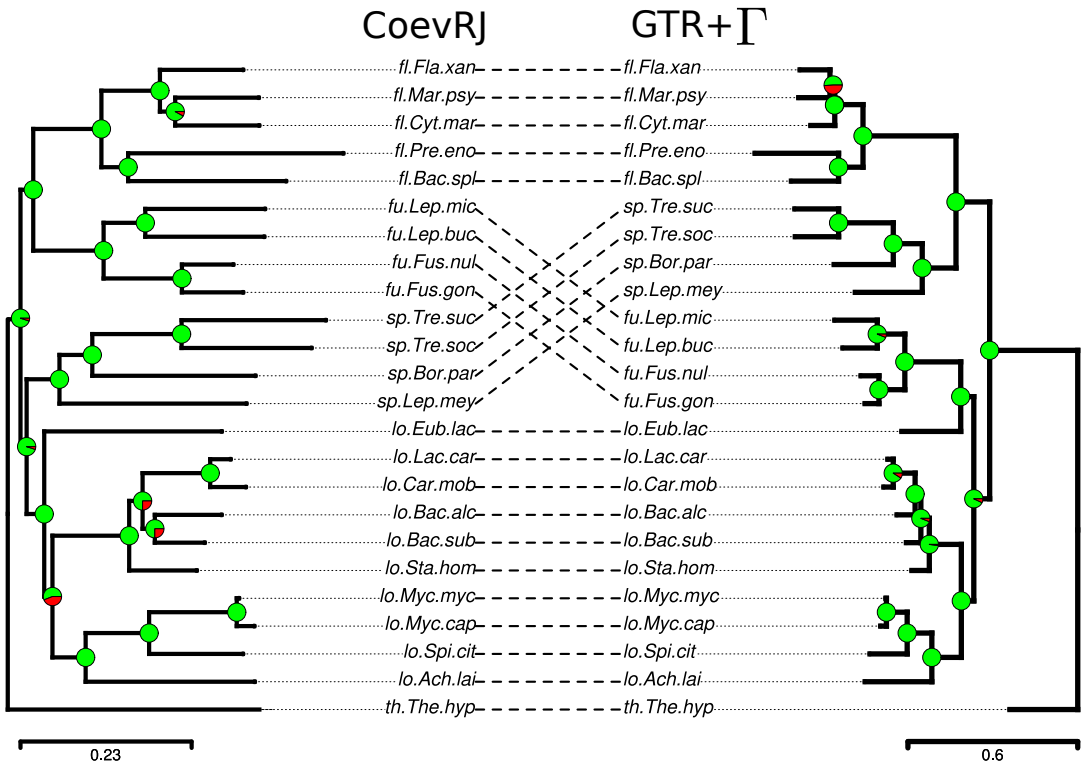


Fig. S3. Comparison of the phylogenetic trees inferred with CoevRJ and GTR+Γ for the Flaviobactera, Fusobacteria, Spyrochaetes and Fermeticutes clades. The pie charts represent the inferred probability for each node to be present in the summarized phylogeny: a fully green pie chart identifies a node supported by a probability of 1.

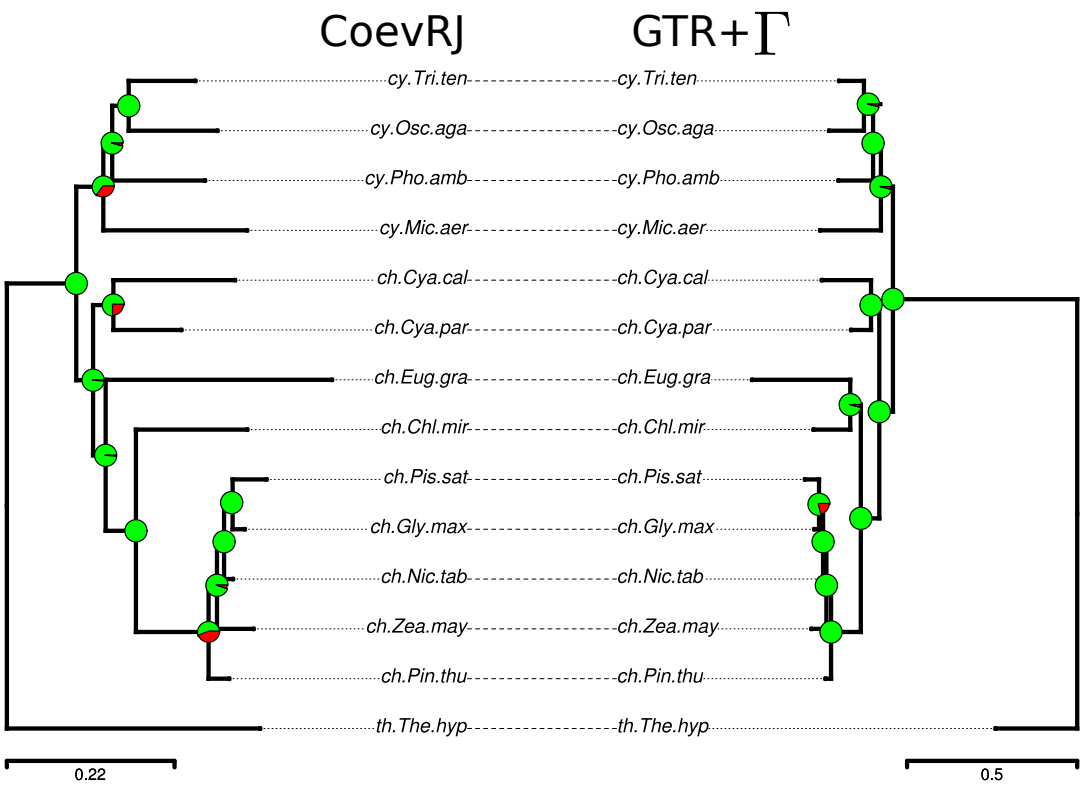


Fig. S4. Comparison of the phylogenetic trees inferred with CoevRJ and GTR+Γ for the Cytobaceteria and Chloroplast clades. The pie charts represent the inferred probability for each node to be present in the summarized phylogeny: a fully green pie chart identifies a node supported by a probability of 1.

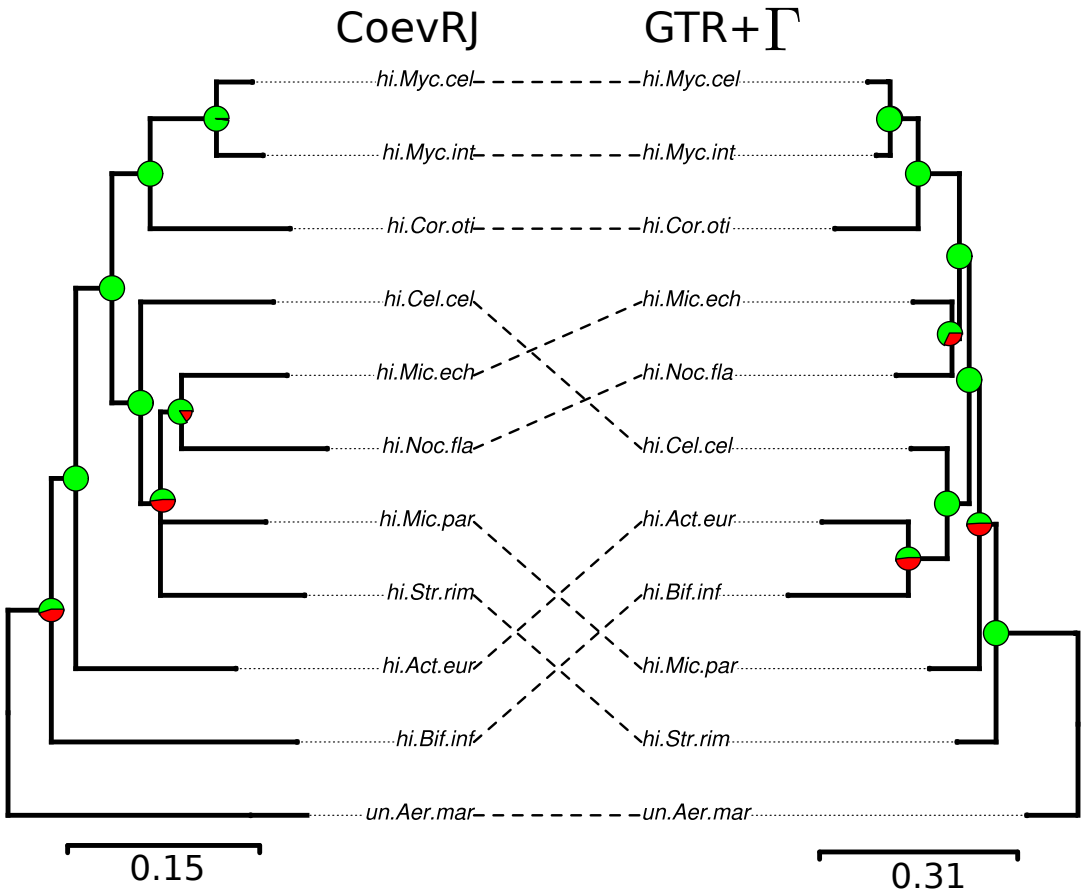


Fig. S5. Comparison of the phylogenetic trees inferred with CoevRJ and GTR+Γ for the Actinobaceteria clade. The pie charts represent the inferred probability for each node to be present in the summarized phylogeny: a fully green pie chart identifies a node supported by a probability of 1.

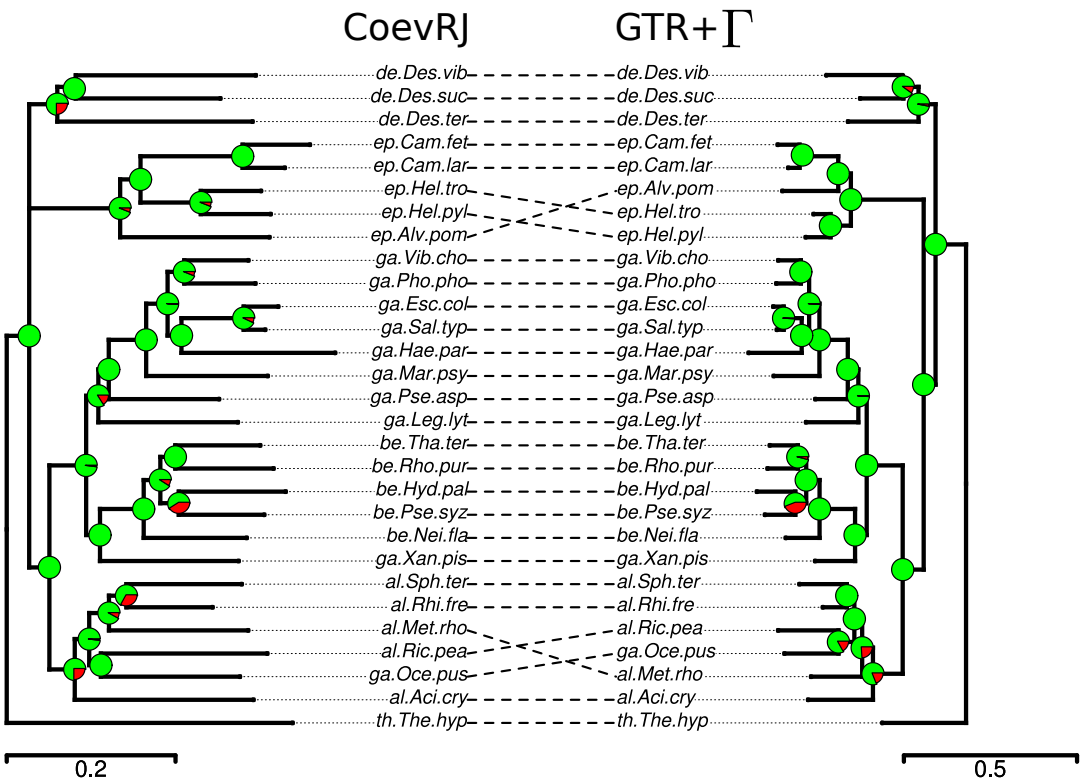


Fig. S6. Comparison of the phylogenetic trees inferred with CoevRJ and GTR+Γ for the Proteobacteria clade. The pie charts represent the inferred probability for each node to be present in the summarized phylogeny: a fully green pie chart identifies a node supported by a probability of 1.

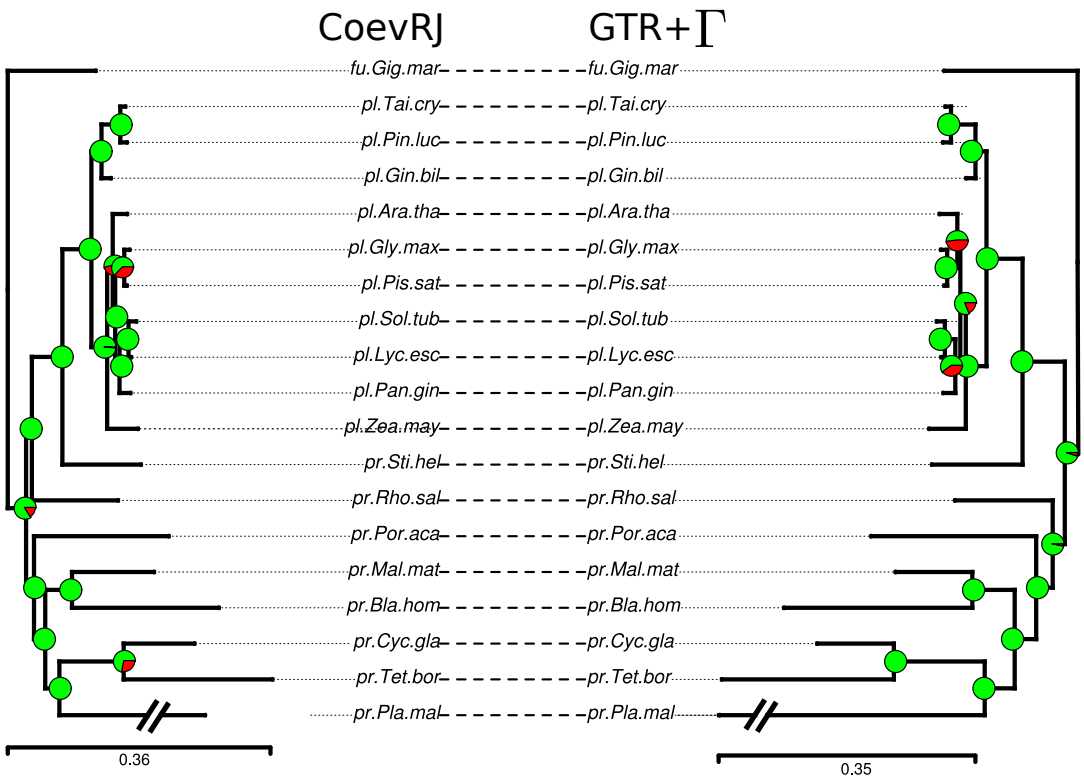


Fig. S7. Comparison of the phylogenetic trees inferred with CoevRJ and GTR+Γ for the plants and protists clade. The pie charts represent the inferred probability for each node to be present in the summarized phylogeny: a fully green pie chart identifies a node supported by a probability of 1.

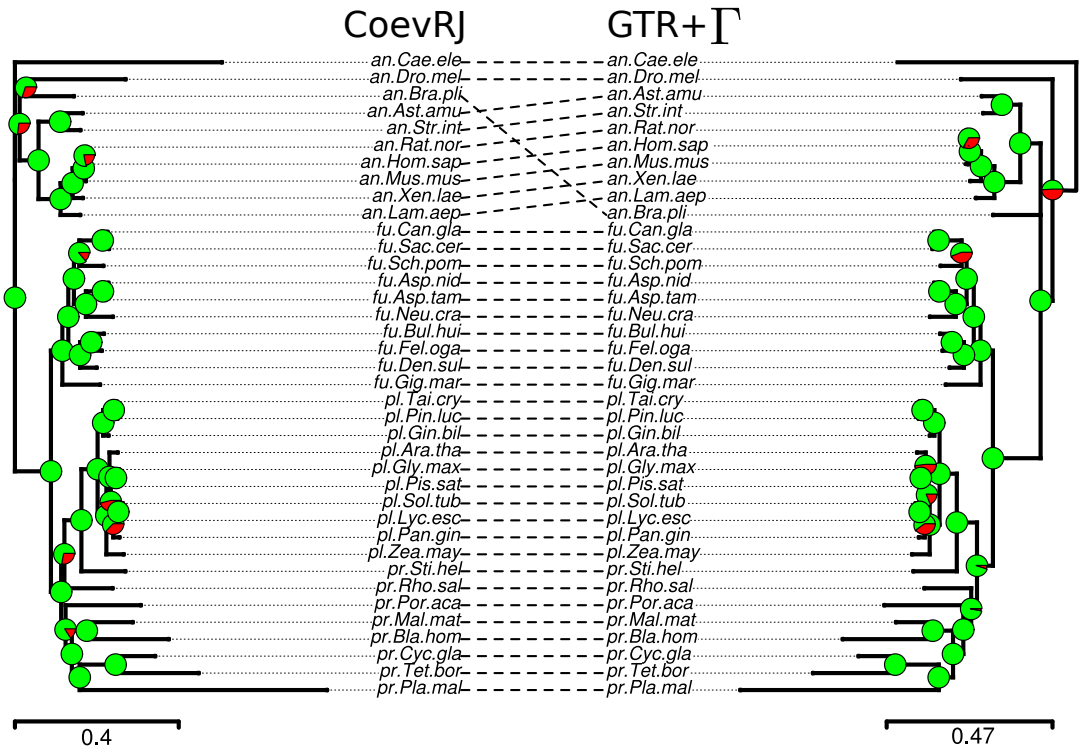


Fig. S8. Comparison of the phylogenetic trees inferred with CoevRJ and GTR+Γ for the Eukaryotes clade. The pie charts represent the inferred probability for each node to be present in the summarized phylogeny: a fully green pie chart identifies a node supported by a probability of 1.

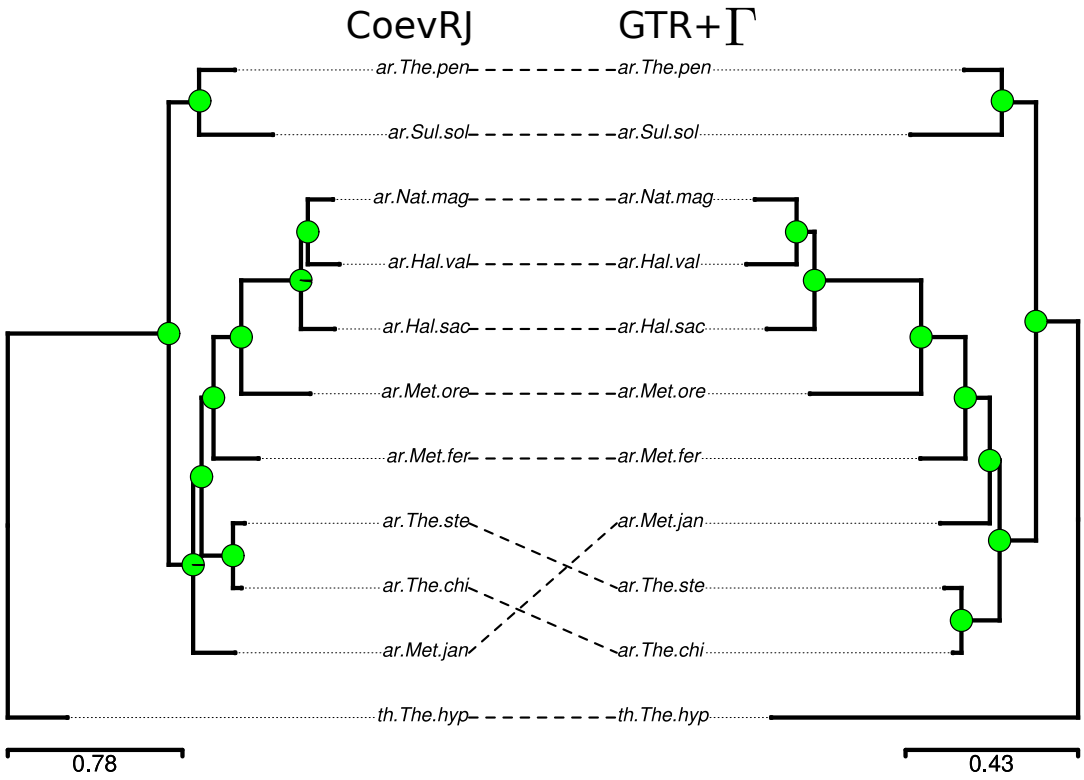


Fig. S9. Comparison of the phylogenetic trees inferred with CoevRJ and GTR+Γ for the Archea clade. The pie charts represent the inferred probability for each node to be present in the summarized phylogeny: a fully green pie chart identifies a node supported by a probability of 1.

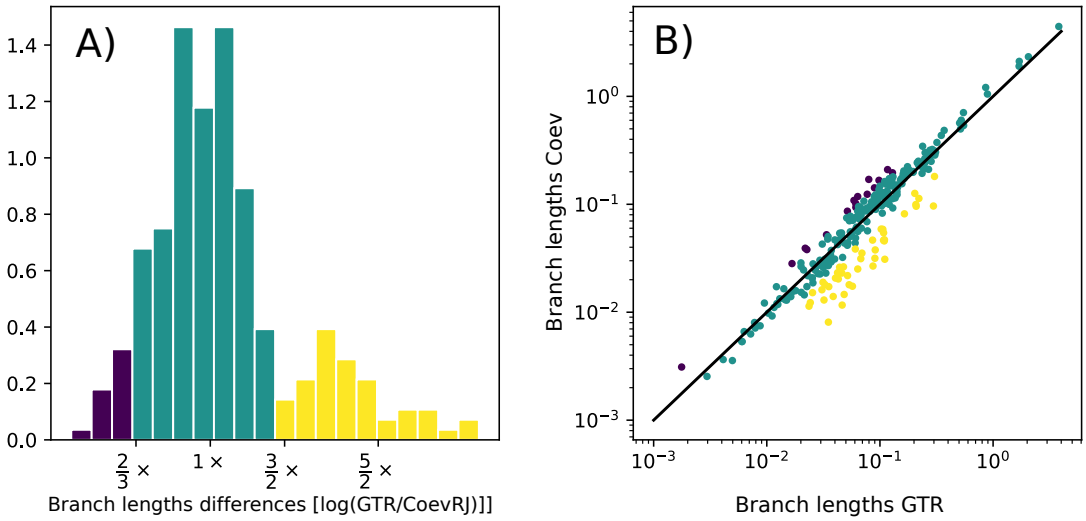


Fig. S10. Distribution of the branch lengths ratio between the consensus tree inferred under the GTR+Γ model and the CoevRJ model on the 16S rRNA dataset. Figure A) shows the histogram of the ratios where bins are colored according to the magnitude of these ratios. Violet bins represent branches that were inferred to be at least 2/3 smaller with the GTR+Γ model. Green bins are branches that that are within a 2/3 to 3/2 ratios. Yellow bins represent branches that were inferred as at least 3/2 longer with the GTR+Γ model. Figure B) shows values of the branch lengths inferred under both models. Each point represents a pair of branch lengths that is colored according to Figure A) colors.

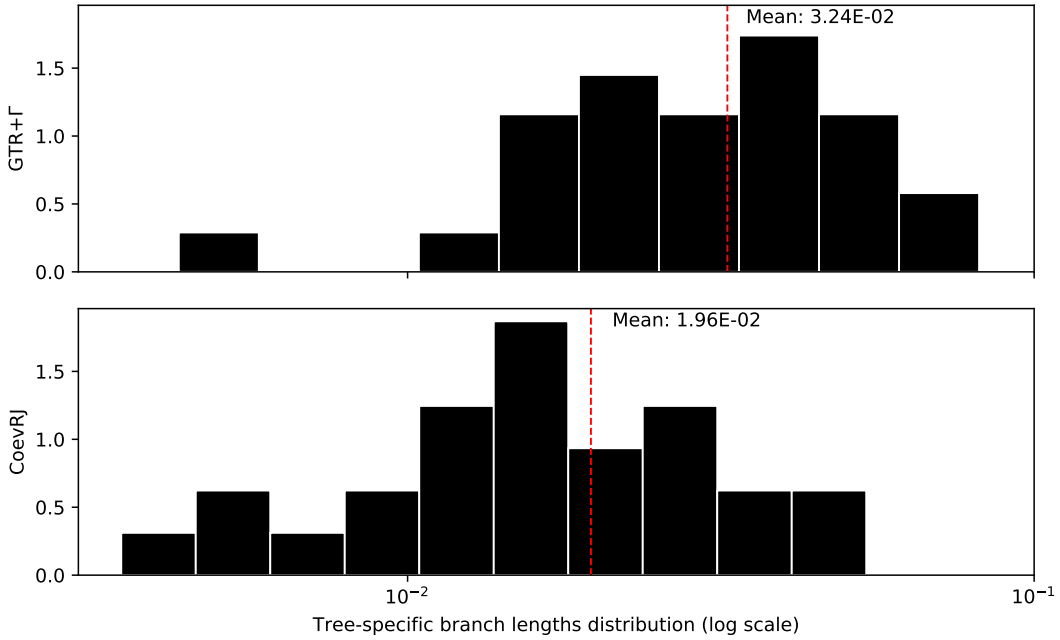


Fig. S11. Distributions of the inferred tree-specific branch lengths for the GTR+Γ and the CoevRJ models on the 16S rRNA dataset. Tree-specific branches are branches (or bipartitions) specific to one of the consensus trees.

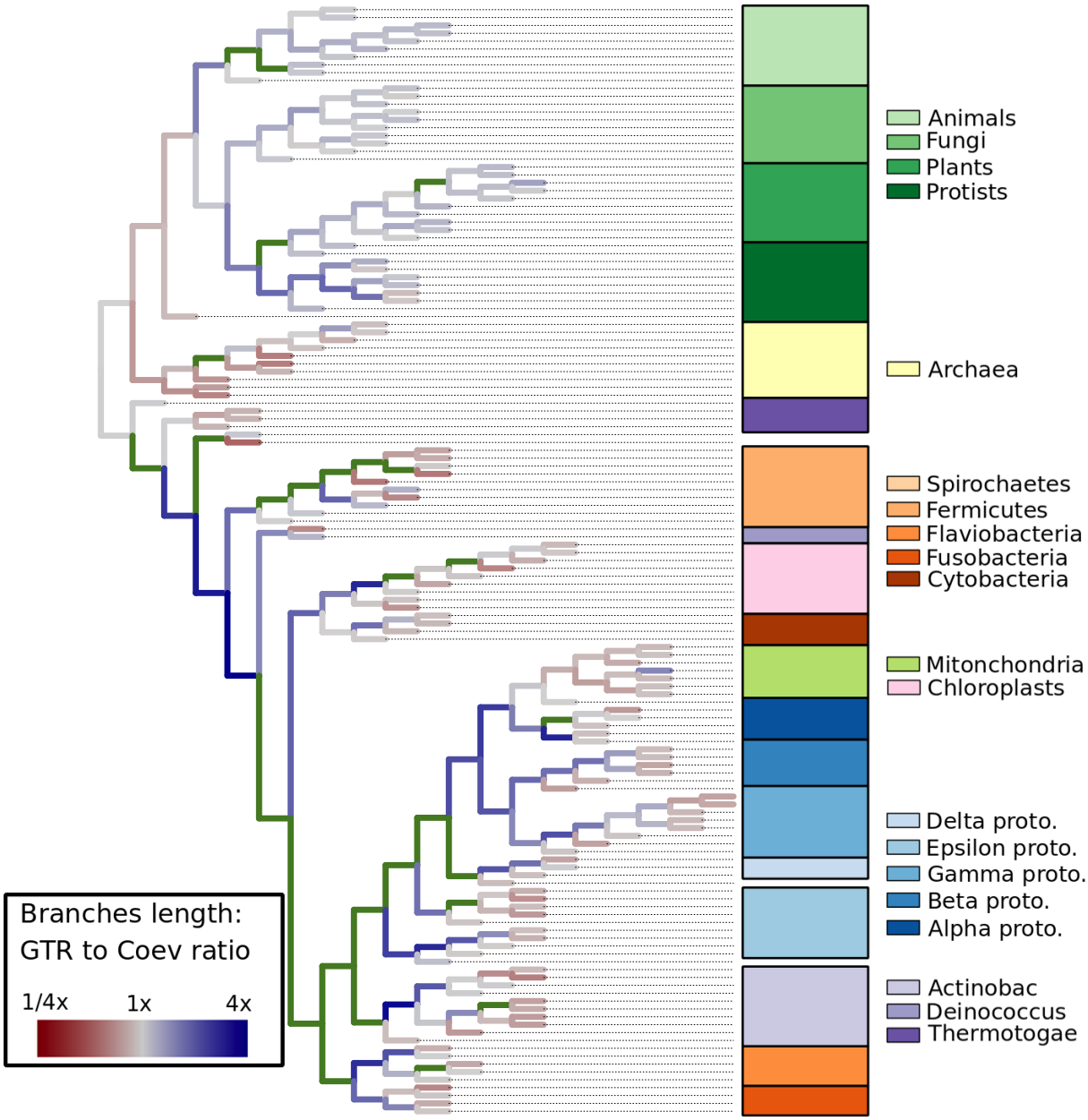


Fig. S12. Branch lengths ratios between the consensus tree inferred under the GTR+Γ model and the CoevRJ model on the 16S rRNA dataset. The phylogenetic tree represented is the consensus tree obtained under the CoevRJ model. Branches that are specific to this consensus tree, and therefore not directly comparable, are colored in green. Branches in shades of blue are inferred as smaller under the CoevRJ model (longer under the GTR+Γ model) and branches in shades of red are inferred as longer under the CoevRJ model (smaller under the GTR+Γ model).

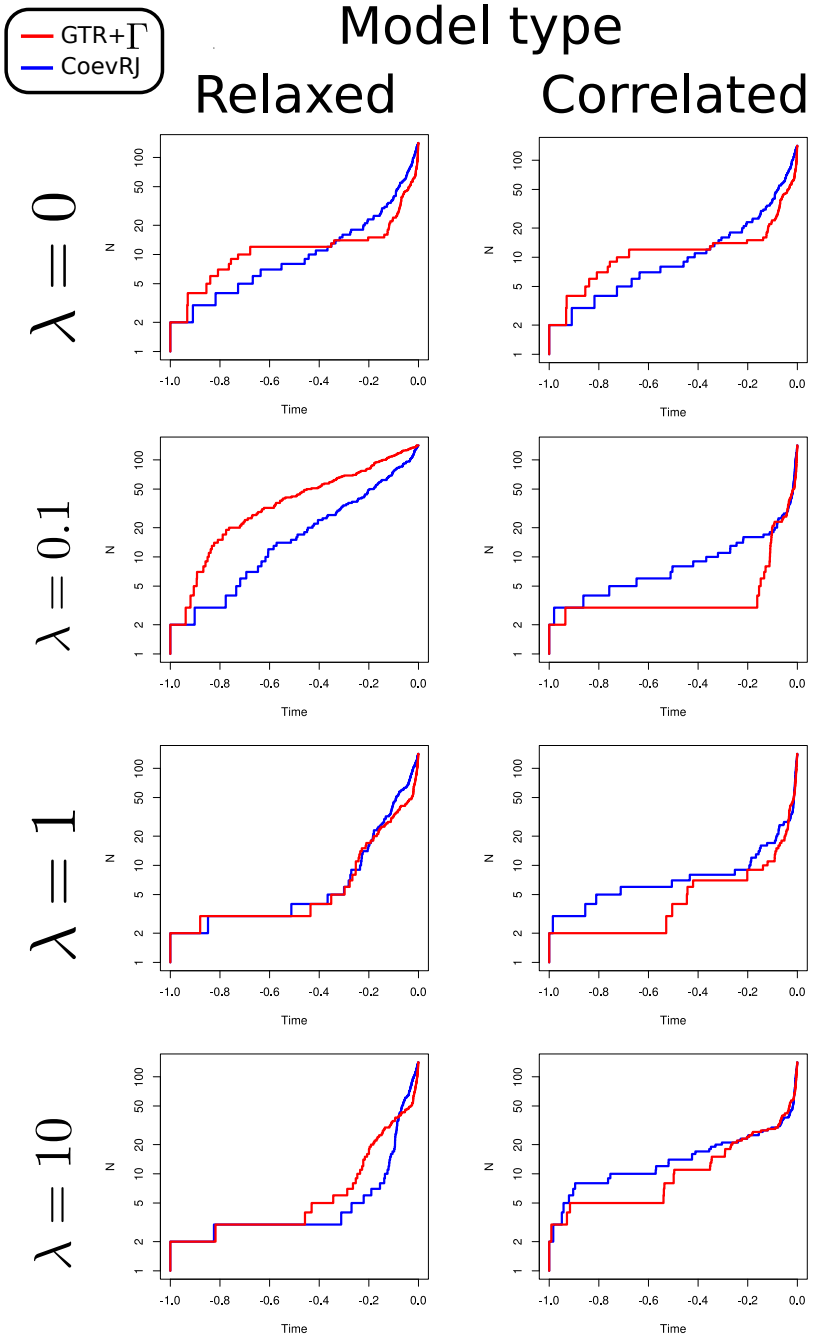


Fig. S13. Molecular dating of the phylogenetic tree inferred for the 16S ribosomal RNA. Lineages through time are reported for the *Relaxed* and *Correlated* penalized likelihood models with various parametrization for the λ parameter informing the strength with which the rate are constrained along branches.

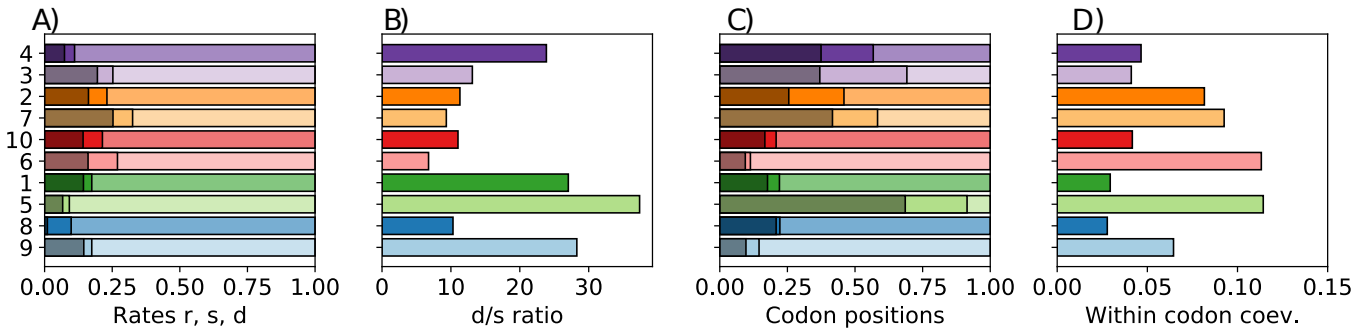


Fig. S14. Supplementary analysis for the 10 protein-coding genes identified by their assigned number on the left side of the figures. Figure A) shows the inferred rates (r, s, d) with color shading identifying the rates : dark tint for rate r, normal tint for rate s and light tint for rate d. Figure B) reports the inferred d/s ratio. Figure C) reports the overall frequencies of the codon positions for sites inferred as coevolving with probability > 0.5. Darker, normal and lighter tints of the colors identifies the frequency of the first, second and third positions respectively (e.g. dataset 4 has $40\%, 20\%, 40\%$ for positions 1,2,3 respectively). Figure D) shows the percentage of coevolving pairs whose sites were within the same codon.

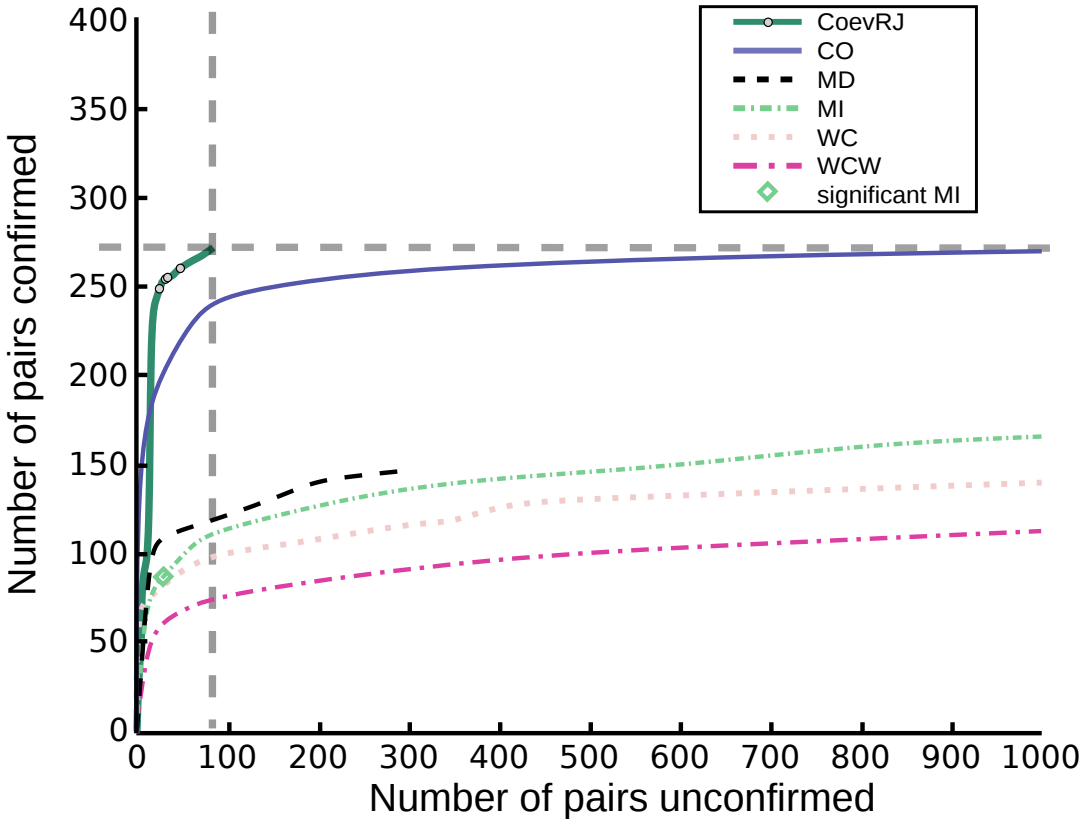


Fig. S15. Comparison of coevolution predictions on the 16S ribosomal RNA dataset for different methods and CoevRJ using the *E. Coli* structure. This figures reports the pairs unconfirmed by the structure against the ones confirmed (closely located on the structure). Pairs are ranked in function of the score returned by the method (posterior probability for CoevRJ). The threshold on the significance of the predicted pairs enables CoevRJ to report mostly confirmed pairs. In absence of such a threshold for the other methods, pairs predicted as coevolving are reported until the 1000th unconfirmed pairs is reached. This figure is based on results from Yeang et al. 2007 (1) in addition of the one from CoevRJ.

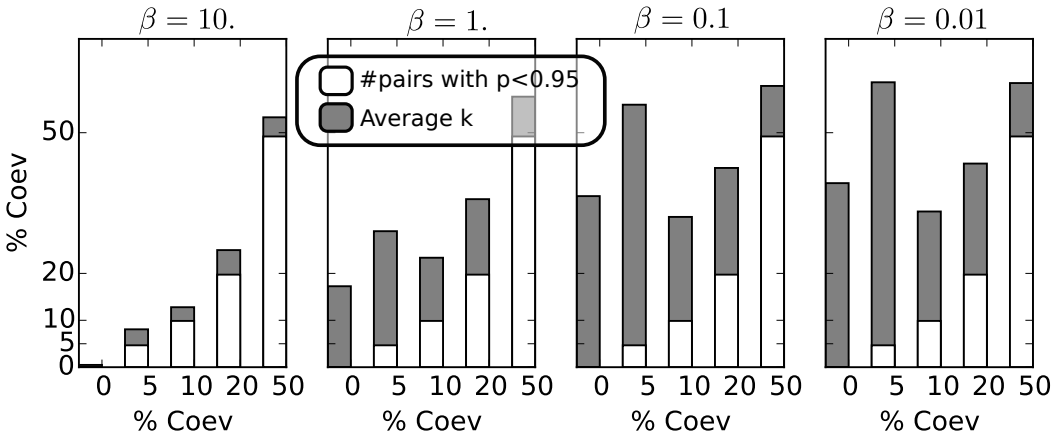


Fig. S16. Effects resulting from different parametrization of the hyperprior on k. Each grey bar represent the average of three k values inferred on simulated datasets. White bars represent the average number of correctly predicted pairs inferred with $p > 0.95$ on these datasets.

2. Supplementary Tables

Table S1. CoevRJ performance at predicting coevolving pairs on the simulated datasets. Only pairs inferred with probability > 0.95 are reported.

% .	Sensitivity	Specificity	Accuracy	PPV
Coev	TP/(TP+FN)	TN/(TN+FP)	(TP+TN)/All	TP/(TP+FP)
0%	-	1	0.99	-
5%	0.98	1	1	1
10%	0.98	1	1	1
25%	0.98	1	1	1
50%	0.99	1	1	1

Table S2. Coevolving pairs inferred with $p > 0.75$ having a distance greater than 6.5 $\mathring{A}$ on the *E.Coli* structure. Positions in bold are reported to have contact with proteins in the small ribosomal subunit (2). The name and the paired position in the protein are reported in the parentheses. Profiles annotated with a $^+$ have a pure Watson-Crick profile, pairs with a * have a partial Watson-Crick profile and the reminder are fully not overlapping with Watson-Crick pairings.

E. Coli		Alignment		Pair Prob	Dist	Profile
P1	P2	P1	P2	Prob	Ang	
204	1336	564	3467	1	177.2	*AA,CT,GC
100	1260	243	3314	1	155.8	$^\times$AT,TA
608 (S16-18)	**1130** (S9-3)	1196	2564	1	143.4	AA,TG,GT
609 (S16-18)	**1125** (S10-5)	1197	2559	1	129.6	$^\times$AT,CG,TA,GC
588 (S8)	**1289** (S7)	1172	3362	1	116.3	TT,GA
618 (S16)	1166	1249	3128	1	104.2	*AA,CG,TT
330	424	721	842	1	78.3	*AC,CG,TT
617 (S16-44)	1448	1205	3593	1	73.6	$^\times$AT,TA,GC
195 (S20-68)	622	552	1253	1	61.2	$^\times$AT,CG,GC
605	1449	1191	3594	1	60.5	AA,CC,GT
1322 (S19-78)	**1371** (S9)	3451	3507	1	35.7	AC,CG,GT
597 (S8-94)	651	1182	1332	1	22.3	*AA,GC
619 (S4-134)	623	1250	1292	1	10.0	AG,CA,TC,GT
202	464	560	981	1	6.6	*AA,CG,TC,GT
292	**1525** (S11-120)	681	3749	0.999	60.0	AA,GG
428 (S4-10)	1261	848	3315	0.997	89.2	TT,GA
937	**1248** (S16-1229)	2288	3230	0.995	34.8	AA,CT
80	845	116	2173	0.891	166.6	TA,GG
462	**705** (S11)	933	1853	0.891	153.0	AC,CA,TT,GG
1025	1036	2417	2449	0.891	8.9	* AC,CG,TA,GT
621	**642** (S8-113)	1252	1322	0.889	50.7	AA,CC,TT,GG
541 (S4-42)	**1148** (S9-16)	1118	3094	0.768	87.3	AA,TC,GT
173	**587** (S8-92)	351	1165	0.763	77.5	AA,TG,GT

Table S3. Eukaryotes datasets from Ensembl/Selectome.

Id	Name	Annotation	# Seq	Length	Filtered Length
1	ENSGT00390000004753.1	Ribosomal Protein L7a	88	805	796
2	ENSGT00550000074423.1	Olfactory receptor 507	256	1111	709
3	ENSGT00600000084009.1	Olfactory receptor family 8 subfamily D	336	1126	883
4	ENSGT00600000084235.1	Protocadherin beta 4	195	2674	2116
5	ENSGT00620000087601.2	Bone Morphogenetic Protein	157	1954	907
6	ENSGT00660000095099.1	Actg2 (actin, gamma 2)	160	1132	1120
7	ENSGT00660000095126.1	Olfactory receptor family 6 subfamily C	315	997	928
8	ENSGT00660000095289.1	Tubulin Beta 3 Class III	111	1336	1321
9	ENSGT00670000097815.3	Elongation factor1-alpha1	86	1387	1387
10	ENSGT00680000099543.3	Tubulin, alpha 1A	192	1720	976

Table S4. Values employed for the priors on the rates defining the Coev Q matrices.

Default prior			Empirical prior		
i	μ_{ψ_i}	$\sigma^2_{\psi_i}$	i	μ_{ψ_i}	$\sigma^2_{\psi_i}$
1	0	0.25	1	0.26	0.37
3	3.5	0.25	3	3.13	0.66

3. Supplementary Algorithm

Algorithm 1 Score algorithm for the proposal of pairs.

for all Possible unique pairs $(i, j) \in N \times N$ **do**
  $\Phi \leftarrow$ Possible profiles given X_i and X_j
  $\chi \leftarrow$ Conflict count for each profile $\phi \in \Phi$
  $\phi_{best} \leftarrow argmin_{\phi \in \Phi}(\chi(\phi))$
  $\sigma^2_{best} \leftarrow$ variance of observed pairs count within profile ϕ_{best}
  $S1 \leftarrow |\Phi|$
  $S2 \leftarrow |\chi(\phi_{best})|$
  $S3 \leftarrow \sigma^2_{best}$
  $S1', S2', S3' \leftarrow \text{normalize}(S1, S2, S3)$
  $S(i, j) \leftarrow \prod_{i=1}^{3} (S'_i)^{w_i}$
  $R(A, B, \phi) \leftarrow \dfrac{\chi(\phi)}{\sum_{\phi \in \Phi} \chi(\phi)}$
return $S(\cdot, \cdot), R(\cdot, \cdot, \cdot)$

References

1. Yeang CH, Darot JFJ, Noller HF, Haussler D (2007) Detecting the Coevolution of Biosequences—An Example of RNA Interaction Prediction. *Molecular Biology and Evolution* 24(9):2119–2131.
2. Cannone JJ, 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 Bioinformatics* 3:2.