A Model-Based Approach for Detecting Coevolving Positions in a Molecule

Julien Dutheil, Tal Pupko,† Alain Jean-Marie,‡ and Nicolas Galtier**

*CNRS UMR 5171 Laboratoire "Génome, Populations, Interactions, Adaptation," Université Montpellier II, Montpellier Cedex, France; †The Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel; and ‡CNRS UMR 5506 Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, Université Montpellier II, Montpellier Cedex, France

We present a new method for detecting coevolving sites in molecules. The method relies on a set of aligned sequences (nucleic acid or protein) and uses Markov models of evolution to map the substitutions that occurred at each site onto the branches of the underlying phylogenetic tree. This mapping takes into account the uncertainty over ancestral states and among-site rate variation. We then build, for each site, a "substitution vector" containing the posterior estimates of the number of substitutions in each branch. The amount of coevolution for a pair of sites is then measured as the Pearson correlation coefficient between the two corresponding substitution vectors and compared to the expectation under the null hypothesis of independence. We applied the method to a 79-species bacterial ribosomal RNA data set, for which extensive structural characterization has been done over the last 30 years. More than 95% of the intramolecular predicted pairs of sites correspond to known interacting site pairs.

Introduction

Distinct positions in proteins and RNA molecules might not evolve independently because of shared structural or functional constraints. In proteins, two amino acid sites that are in close proximity in the three-dimensional (3D) structure might coevolve in a complementary manner. This is, for instance, the case of the "small-to-large" mutation Ala129 to Met in bacteriophage T4 lysozyme, which is compensated by the "large-to-small" mutation Leu121 to Ala (Baldwin et al. 1996). RNA molecules—mainly transfer RNA and ribosomal RNA (rRNA)—also evolve under structural constraints. Their particular folding leads to the formation of characteristic secondary motifs required for the function of the molecule, namely, loops (single-stranded regions) and stems (double-stranded regions with a DNA-like double-helix structure). Stem positions are known to evolve in a compensatory way (Woese 1987; Rousset, Pélandakis, and Solignac 1991).

Detecting sites that do not evolve independently is important for the understanding of the various structural and functional constraints acting on a molecule at the level of specific sites. Such understanding is important for elucidating the mechanisms of molecular evolution and might also have practical applications in structure prediction and drug design.

Because many biochemical mechanisms may lead to nonindependent evolution, it is difficult to take nonindependence into account in evolutionary models. Some attempts have been made in the particular case of rRNA (Tillier and Collins 1998; Savill, Hoyle, and Higgs 2001) and proteins (Pollock, Taylor, and Goldman 1999). Tillier and Collins (1995) assessed the impact of the independence hypothesis on tree reconstruction methods. They showed that nonindependence among sites may be seen as a redundancy of the phylogenetic signal within the data and hence may lead to a reduced tree reconstruction efficiency. Following this idea, Galtier (2004) showed that signal redundancy may lead to overestimated bootstrap support values because coevolving sites will tend to support the same topology, either correct or incorrect. The detection of coevolving sites could therefore help improve phylogenetic reconstructions.

The earliest methods for detecting coevolving sites were proposed by structural biologists. They used comparative sequence analysis to detect excessively frequent co-occurrences of states at pairs of sites (Altschuh et al. 1987; Gutell et al. 1992; Neher 1994). Such methods succeeded in detecting coevolving sites but led to many false positive predictions. The main reason is that biological sequence data sets depart from the independence assumption not only because of possible functional interactions but also because of shared evolutionary history. Thus, a method for detecting coevolving sites should distinguish the desired structural/functional correlation signal from the phylogenetic one.

A method suggested by Tillier and Lui (2003) aims to remove the phylogenetic component from the computation of coevolving positions. However, this method does not rely on an evolutionary substitution model and hence cannot take into account factors such as substitution probabilities among states or the among-site rate variation. Model-based methods rely on standard Markov models of sequence evolution, considering the hypothesis of independence as the null hypothesis. Two approaches for model-based inference have been suggested. In the first approach, a likelihood ratio test is used to compare the independence (single site) model to a joint model allowing coevolution between a given number of coevolving sites (Pollock, Taylor, and Goldman 1999; Akmaev, Kelley, and Stormo 2000). Alternatively, Tufféry and Darlu (2000), following Shindyalov, Kolchanov, and Sander (1994), proposed a method based on the (model-based) mapping of cosubstitutions onto the tree. A cosubstitution was defined as two different sites undergoing a substitution on the same branch of the tree. For each site, substitutions were mapped onto the (presumably known) phylogeny by reconstructing ancestral states at each node and assigning one substitution to a given branch if the states at the top and bottom nodes of that branch differ. Then, the number of cosubstitutions for a pair of sites

Key words: coevolution, RNA structure prediction, correlated substitutions, intramolecular and intermolecular interaction, Markov models.

E-mail: julien.dutheil@univ-montp2.fr.

Mol. Biol. Evol. 22(9):1919–1928. 2005
doi:10.1093/molbev/msi183
Advance Access publication June 8, 2005

was calculated and compared to the expected number under the null hypothesis of independence (Tufféry and Darlu 2000). This method does not account for multiple substitutions or take into account the uncertainty in the reconstruction of ancestral states.

Here, we introduce a new empirical Bayesian method for the detection of coevolving positions, taking into account the uncertainty in substitution mappings, multiple substitutions, and among-site rate variation. We apply it to a bacterial rRNA data set to test whether the method could detect site pairs involved in documented structural motifs. We succeeded in retrieving a large number of significant coevolving site pairs, almost 90% of which match already known structural pairs involved in stems. Among the remaining 10%, we show by using 3D structure information and previous studies that at least 26 pairs out of 42 are true coevolving site pairs. Hence, for this rRNA data set, more than 95% of the predicted interactions are true "coevolving pairs."

Methods

The method analyzes a set of aligned sequences (D) using a phylogeny (assumed to be known), a Markovian substitution model, and a discrete rate distribution across sites. The set of parameters Θ, including branch lengths, the entries in the substitution matrix, and the rate distribution parameters, is estimated using the maximum likelihood (ML) method prior to the cosubstitution analysis. The HKY85 + Γ (Hasegawa, Kishino, and Yano 1985) substitution model was used, with a four-class discretized gamma rate distribution (Yang 1994).

The method relies on two points: the mapping of substitutions along the tree for each site and its uncertainty and the estimation of the degree to which such substitutions coevolve for a pair of sites.

Substitution Vectors

Let D_i be the ith site of the data set, i.e., a column of the alignment. Let $V_i=(v_{i,1}, \ldots, v_{i,b}, \ldots, v_{i,m})$ be a vector of dimension m, the number of branches in the tree, where $v_{i,b}$, is the posterior estimate of the number of substitutions that occurred on branch b for site i. V_i is called the substitution vector for site i and is estimated as follows.

Let $\{x_1, \ldots, x_a\}$ be a particular joint reconstruction of ancestral states for site i and b a branch of length t, a being the number of inner nodes in the tree. Let x_p and x_q be the states at the top and bottom nodes, respectively, of this branch for this reconstruction. $v_{i,b}$, is the expected number of substitutions on a branch of length t knowing its initial state x_p and final state x_q (named n_{x_p,x_q}) times the probability of the joint reconstruction, summed over all joint reconstructions:

$$v_{i,b} = \sum_{\{x_1,\ldots,x_p,x_q,\ldots,x_a\}} P(x_1, \ldots, x_a \mid D_i, \Theta) \times n_{x_p,x_q}(t). \quad (1)$$

We can rewrite this equation by grouping all reconstructions with identical x_p and x_q:

$$v_{i,b} = \sum_{x_p}\sum_{x_q}\left(n_{x_p,x_q}(t) \times \sum_{\{x_1,\ldots,x_a\}\setminus\{x_p,x_q\}} P(x_1, \ldots, x_a \mid D_i, \Theta)\right). \quad (2)$$

The rightmost summation is equivalent to the "two-states joint" probability, $P(xp, xq|Di, \Theta)$:

$$v_{i,b} = \sum_{x_p}\sum_{x_q} P(x_p, x_q \mid D_i, \Theta) \times n_{x_p,x_q}(t). \quad (3)$$

Estimating the Number of Substitutions According to Initial and Final States

The usual way for mapping substitutions is to count zero substitution when the two states at a branch are identical and one substitution if they are different (Tufféry and Darlu 2000; Nielsen 2002). Because this does not account for multiple substitutions, we chose to estimate the conditional number of substitutions, $n_{x_p,x_q}(t)$. Computations are made as shown in Jean-Marie et al. (unpublished data). Let $N(t)$ be the number of jumps of the Markov chain on a given branch of length t and note that

$$n_{x_p,x_q}(t) = E(N(t) \mid x_p, x_q). \quad (4)$$

We introduce $m_{x_p,x_q}(t)$ defined as the joint expectation

$$m_{x_p,x_q}(t) = E(N(t), x_q \mid x_p). \quad (5)$$

We have

$$n_{x_p,x_q}(t) = \frac{m_{x_p,x_q}(t)}{p_{x_p,x_q}(t)}, \quad (6)$$

where $p_{x_p,x_q}(t)=P(x_q|x_p)$. Let $M(t)=\{m_{x_p,x_q}(t)\}$, Jean-Marie et al. (unpublished data) showed that

$$M(t) = \sum_{n=1}^{\infty} \frac{t^n}{n!} \sum_{p=0}^{n-1} Q^p (Q + \Lambda) Q^{n-p-1}, \quad (7)$$

where $Q=\{q_{x_p,x_q}\}$ is the generator of the Markov chain and $\Lambda=\{\lambda_{x_p,x_q}\}$ the diagonal matrix with all substitution rates ($\lambda_{x_p,x_q} = -q_{x_p,x_q}$, if $x_p=x_q$, else 0). All $n_{x_p,x_q}(t)$ can be computed approximatively by truncating the series (n = 10 gives a good approximation).

Taking Among-Site Rate Variations into Account

Equation (3) may be rewritten to account for among-site rate variation (assuming a discrete distribution of rates) by summing over all rate classes:

$$v_{i,b} = \sum_{c}\sum_{x_p}\sum_{x_q} P(x_p, x_q, r_c \mid D_i, \Theta) \times n_{x_p,x_q,r_c}(t), \quad (8)$$

where r_c is the rate of class c and $n_{x_p,x_q,r_c}(t)$ is the conditional number of substitutions expected on a branch of length t knowing states x_p and x_q and the rate r_c. This number is equal to the expected number of substitutions on a branch of length $t \times r_c$:

$$v_{i,b} = \sum_c \sum_{x_p} \sum_{x_q} P(x_p, x_q, r_c \mid D_i, \Theta) \times n_{x_p,x_q}(t \times r_c). \quad (9)$$

The first term in the above summation is the joint probability of having state x_p at the bottom node, state x_q at the top node, and rate class c given the data and parameters. It can be computed as follows:

$$P(x_p, x_q, r_c \mid D_i, \Theta) = \frac{P(x_p, x_q, r_c, D_i \mid \Theta)}{P(D_i \mid \Theta)} = \frac{P(x_p, x_q, D_i \mid \Theta, r_c) \times P(r_c)}{P(D_i \mid \Theta)}, \quad (10)$$

where the first term of the numerator is the likelihood for site i conditional on states x_p and x_q at top and bottom nodes and rate equal to r_c. This likelihood is computed as described by Felsenstein (1981), after having multiplied all branch lengths by r_c (Yang 1993) and summing over all possible ancestral states at each node except for the top and bottom nodes of branch b, for which states x_p and x_q are fixed. $P(r_c)$ is the prior probability for site i of being in rate class c, and $P(D_i|\Theta)$ is the likelihood for site i. This calculation extends the empirical Bayesian estimation of ancestral states probability of Yang, Kumar, and Nei (1995) to the two-states joint case, as previously proposed by Galtier and Boursot (2000) and Pupko et al. (2003).

Replacing equations (10) and (6) into (9) allows one to calculate estimated substitutions vectors V_i in time proportional to the number of sites and to the square of the number of sequences.

Coevolution Statistic for a Pair of Sites

The amount of coevolution for a pair of sites is measured by taking the Pearson correlation coefficient of the two corresponding substitution vectors:

$$\rho_{i,j} = \frac{cov(V_i, V_j)}{sd(V_i) \times sd(V_j)}. \quad (11)$$

Positive values of ρ mean that the two sites tend to undergo substitutions on the same branches, whereas ρ values close to 0 are expected if the two sites evolve independently. One can look for coevolution by measuring correlation coefficients for pairs of sites within a single molecule (intramolecular coevolution) or by taking each site in a distinct data set (intermolecular coevolution).

Mean Posterior Rates

We computed the mean posterior rates $\hat{r}_i$ for each site i by averaging upon each rate class:

$$\hat{r}_i = \sum_c \left(\frac{P(D_i \mid \Theta, r_c) \times P(r_c)}{P(D_i \mid \Theta)} \times r_c \right). \quad (12)$$

This is the mean of all r_c weighted by the posterior probabilities of being in each class (Mayrose et al. 2004).

rRNA Sequence Data

Two data sets of bacterial large subunit (LSU) and small subunit (SSU) rRNA were built, each with 79 sequences from the same 79 species. Aligned sequences were retrieved from the rRNA database (Wuyts, Perrière, and Van De Peer 2004). Alignments were inspected by eye and slightly modified. All ambiguously aligned and gap-containing sites were discarded from the analysis. The total number of analyzed sites was 2,312 (LSU) and 1,300 (SSU). Data are available in *Supplementary Material*.

The two data sets (LSU and SSU) were first concatenated to estimate a common phylogeny with the ML method using the PhyML software (Guindon and Gascuel 2003). As for the coevolution analysis, we used the HKY85 + Γ model with a four-class discretized gamma rate distribution. Other parameters were considered specific and reestimated separately from each data set. Substitution vectors were estimated for every site, and correlation coefficient ρ was calculated for every pair of sites within and between data sets (2,673,828 pairs for the LSU, 845,650 pairs for the SSU, and 3,005,600 pairs for the interaction).

Simulations

We used a parametric bootstrap approach to evaluate the null distribution of ρ. All simulations were performed under a HKY85 + Γ model. Specifically, the null distribution was estimated by simulating 100,000 independent pairs and computing ρ for each pair. Three distributions of ρ were built, using each data set separately (intramolecular study) and then together (intermolecular study).

The same approach was used to evaluate if the number of site pairs in the data set with ρ higher than a specific threshold is greater than the chance expectation. Fifty simulated data sets were generated for the LSU and SSU data sets, with the same numbers of species and sites. For each data set, we counted the number of site pairs with ρ greater than the considered threshold. Parameters used for simulation were the ones estimated from each data set. These parameter values were also used for the estimation of substitution vectors and were not reestimated for each simulated data set. This is hence an approximated parametric bootstrap, similar to the resampling of estimated log-likelihoods (RELL) method of Kishino, Miyata, and Hasegawa (1990).

Structural Data

The secondary structures of *Escherichia coli*'s rRNA sequences (accession number U18997) were used as a reference for the determination of structural pairs. The structures used are the ones given in the Dedicated Comparative Sequence Editor (DCSE) alignment from the rRNA database. The RNAViz2 software (De Rijk, Wuyts, and De Wachter 2003) was used for RNA representation and site

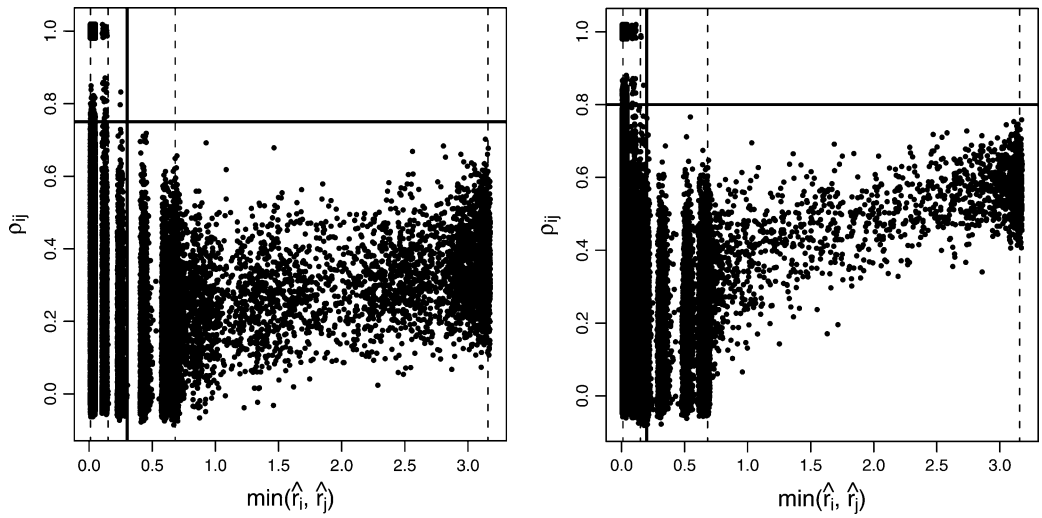

FIG. 1.—Correlation coefficient ρ plotted against the minimal posterior rate $\hat{r}_{min}$, for 100,000 simulated independent pairs. Dashed lines are drawn at prior rates of the four classes of the gamma distribution. Continuous lines represent the $\hat{r}_m$ and ρ_m thresholds used in the analysis (see *Results*).

visualization. 3D analysis was performed using the *Thermus thermophilus* 5.5 Å structure (Protein Data Bank [PDB] accession number 1GIX and 1GIY for LSU and SSU, respectively; Yusupov et al. 2001), *T. thermophilus* 3.0 Å SSU structure 1J5E (Wimberly et al. 2000), and *Deinococcus radiodurans* 3.1 Å LSU structure 1NKW (Harms et al. 2001). We used the MolScript (Kraulis 1991) and the Raster3d (Merritt and Bacon 1997) softwares for 3D representation of molecules.

Results

The Expected Distribution of the Correlation Coefficient

We developed a new measure of the amount of coevolution for a pair of sites, defined as the Pearson correlation coefficient ρ between two corresponding substitution vectors. Positive ρ means that the two sites tend to substitute on the same branch (cosubstitute, see *Methods*). Because positive values of ρ may be obtained by chance, we determined the null distribution of ρ, assuming that sites evolve independently. This was achieved by conducting simulations (parametric bootstrapping) using parameter values estimated from either the LSU or the SSU data set. The distribution of ρ depends on the evolutionary rate of the two sites tested. This is seen in figure 1, where the correlation coefficient ρ is plotted as a function of the minimal posterior rate $\hat{r}_{\min}$. The relationship between ρ and $\hat{r}_{\min}$ appears complex. Slow-evolving sites confer a high variance to ρ estimates, so that high values of ρ can be reached just by chance. Two invariable sites, for example, have identical substitution maps and a correlation coefficient equal to one. Slow sites, therefore, will be ignored when trying to detect coevolving pairs. For fast-evolving pairs, ρ tends to be positive and positively correlated with $\hat{r}_{\min}$. This is due to the phylogenetic correlation: sites tend to undergo more substitutions in the long branches and fewer in the short ones. This effect is more obvious when sites undergo a large number of substitutions. Phylogenetic correlation appears to be stronger in the SSU data set. This is consistent with the fact that the SSU tree is longer than the LSU tree.

From these simulations, we want to determine a correlation threshold (ρ_m) above which a given pair of sites from real rRNA data should be considered as significantly departing from the independence hypothesis. According to the above discussion, this threshold would ideally depend on the evolutionary rates of the sites in the considered pair. Based on the results shown in figure 1, we decided to consider a site pair as coevolving if (1) its $\hat{r}_{\min}$ is greater than $\hat{r}_m = 0.3$ (LSU) or 0.2 (SSU) and (2) its ρ is greater than $\rho_m = 0.75$ (LSU) or 0.8 (SSU). Virtually, no values of ρ greater than these thresholds are expected under the null hypothesis of independence. We call sites with $\hat{r} > \hat{r}_m$ "fast-evolving sites." Pairs with $\hat{r}_{\min} > \hat{r}_m$ and $\rho > \rho_m$ are defined as coevolving pairs.

Detecting the Coevolving Sites

Having set the two ρ_m and $\hat{r}_m$ thresholds, we checked if there are coevolving pairs in the rRNA data sets. We found 258 coevolving pairs for the LSU data set and 126 for the SSU data set. Figure 2 (graphs 1 and 3) shows the right-tail distributions of the ρ statistic measured on real LSU and SSU data sets. We then tested whether these numbers are significant by simulating each data set under the independence hypothesis and counting the number of pairs for which $\rho > \rho_m$. Figure 2 (graphs 2 and 4) shows the mean distribution over 50 simulated data sets. There is a striking difference between real and simulated data sets: only 5.9 and 0.94 site pairs, respectively, reach the ρ_m value in the simulated data sets. All detected pairs with $\rho > \rho_m$ and $\hat{r}_{\min} > \hat{r}_m$ were drawn on the *E. coli* structure (*Supplementary Material*).

Comparison with Secondary Structure

The number of detected coevolving sites was significantly greater than the chance expectation. In order to characterize the biological relevance of these sites, we analyzed the secondary structure of rRNA given in the rRNA database (Wuyts, Perrière, and Van De Peer 2004). Out of the 258 coevolving sites in the LSU data, 225 (87%) were

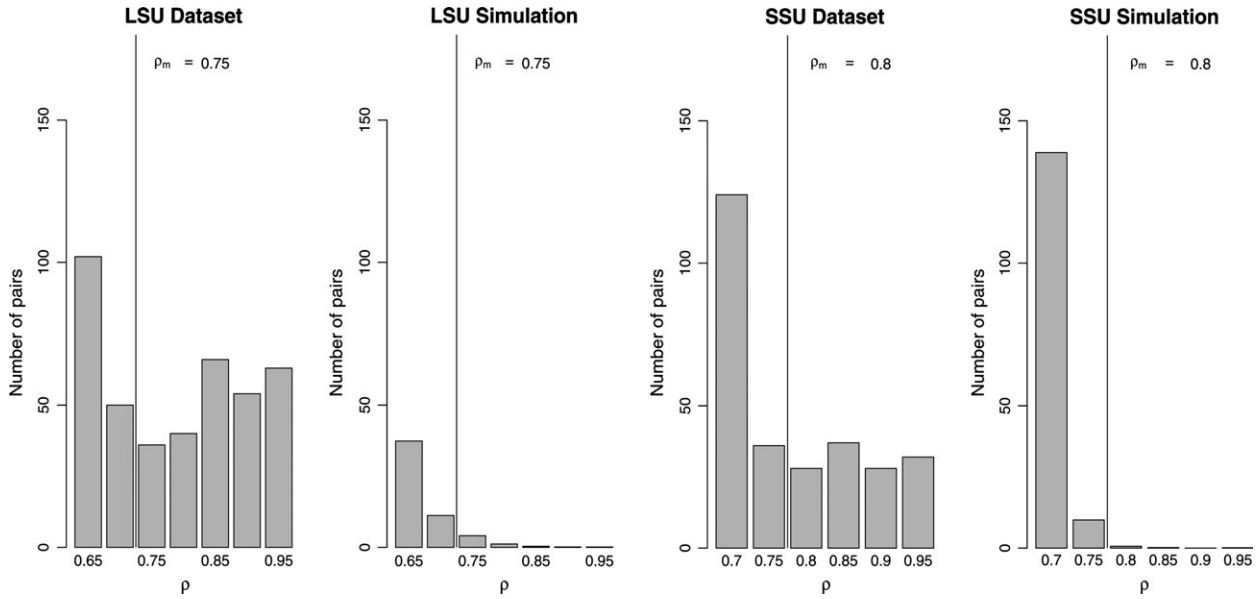

FIG. 2.—Right-tail distribution of ρ for the LSU and SSU data sets. The number of pairs is plotted against the correlation coefficient ρ, for rRNA data set (graphs 1 and 3) and corresponding simulated data sets (graphs 2 and 4). In the latter case, the average number over 50 simulations was used. Only fast-evolving site pairs, i.e., when both sites have a posterior rate greater than 0.3 (LSU) or 0.2 (SSU), were used. Vertical lines correspond to the correlation threshold ρ_m used in each case.

found to match already known stem pairs. For the SSU data, the ratio was even greater: out of the 126 coevolving pairs, 117 (93%) are known stem pairs.

Comparison with Tertiary Structure

Forty-two detected pairs did not match any known stem pair in the secondary structure. These pairs were further examined using the 3D structure of both *T. thermophilus* (LSU and SSU) and *D. radiodurans* (for the LSU). Results are shown in tables 2 and 3. We classified these pairs into four categories (see table 1 and fig. 3 for examples), ranging from nondocumented Watson-Crick (WC) pairs to structurally distant sites. Only sites falling into category 4 cannot be confirmed as coevolving and may be false positives. It is noteworthy that this concerns 10 pairs out of the 258 + 126 detected ones. Pairs in category 3b correspond to ambiguously predicted stems and/or probable frameshifts. This is the case of the E25 stem of the LSU for instance. The DCSE structure of *E. coli* shows two unpaired loops, whereas the Comparative RNA Web site (CRW, Cannone et al. 2002) predicts it as a stem, with several mismatches. Shifting down the right strand from one nucleotide, as suggested by our results, leads to a similar number of mismatches because the left strand is made of four consecutive G's (fig. 4). Out of 384 detected pairs, 342 correspond to unambiguous stem pairs and 26 to confirmed tertiary interactions (categories 1–3a). This leads to a score of 95.8% successful predictions

Table 1
Classification of Detected Coevolving Sites That Are not Stem Pairs

Type	Description	Example
1	Canonical WC base pair. Pairs are of type AU or GC and have two or three hydrogen bonds. Rings are in the same plane.	Figure 3(*A* and *B*)
2	Bases are close enough to allow hydrogen bonds but are not canonical WC pairs.	
2a	Pair belongs to a stem (local mismatch)	
2b	Pair does not belong to a stem	Figure 3*C*
3	Bases are close enough to be interacting, but residues do not allow hydrogen bonds. Coevolution may hence be due to the following:	
3a	Hydrophobic or van der Waals interaction	Figure 3*D*
3b	Frameshift	Figure 5
4	Bases are not close enough to be directly interacting.	

We checked the CRW to compare our results with those of other methods. It was possible to retrieve from this knowledge database a list of site pairs previously predicted as being involved in tertiary interaction by a battery of approaches, including comparative analysis and secondary structure prediction methods. We found that all sites from category 1 to 3a were actually known to be interacting (see tables 2 and 3). Sixty-one (LSU + SSU) pairs of sites in the CRW were not detected by our method. Three explanations for the nonidentification of these reported pairs are as follows: (1) 7 pairs were considered as ambiguously aligned in our analysis, (2) 6 pairs include a site with a gap in our alignment, and (3) 39 pairs include a site evolving too slowly (posterior rate <0.3 for LSU, 0.2 for SSU). All these sites were hence excluded from our analysis. Thus, as far as 3D interactions are concerned, only nine CRW-documented site pairs potentially detectable by our method were missed.

Intermolecular Interaction

We also applied the method to search for interactions between the two subunits. This has been achieved by measuring all possible correlation coefficients between

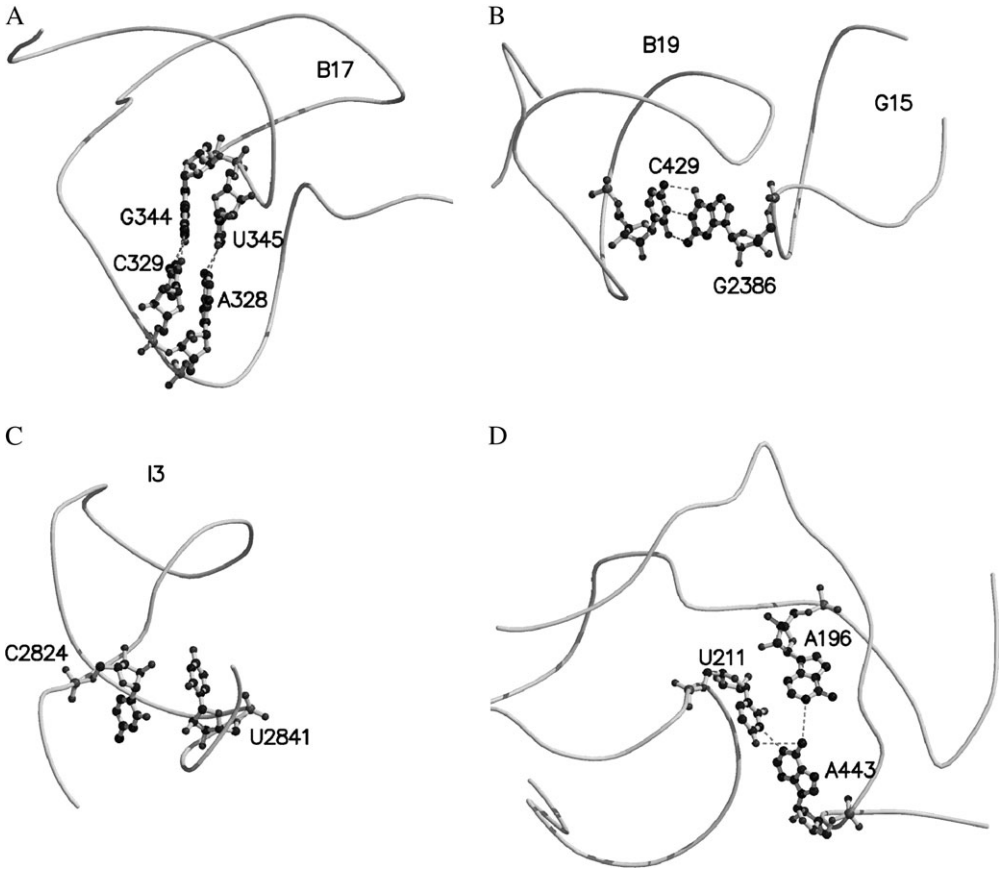

FIG. 3.—Example of detected coevolving sites that do not belong to a stem in *Escherichia coli*. (*A*) Small stem near B17, (*B*) long-range WC interaction, (*C*) triple interaction between B11 and B13, and (*D*) probable size interaction near I3. See figures in Supplementary Material online for two-dimensional pairs representation and table 1 for a classification of nonstem detected sites.

one site from LSU and one site from SSU. The null distribution is different from the intra-subunit ones: fewer site pairs with high ρ values are observed. This may be because the estimated branch lengths are different between the two subunits. No high value of ρ is observed from the real data too. We checked the 10 pairs with greatest ρ and did not find already documented interacting pairs, mostly because sites known to be involved in interactions between LSU and SSU are highly conserved through evolution. Yusupov et al. (2001) give a list of these sites.

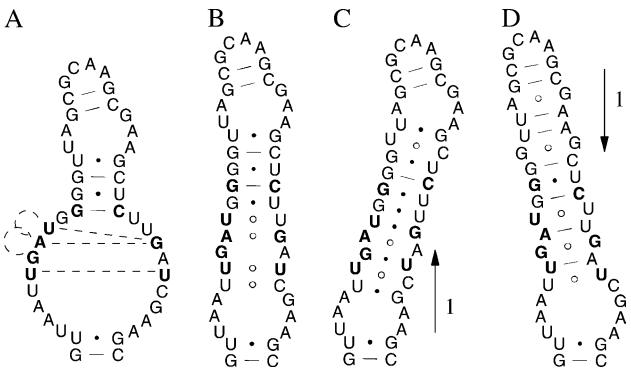

FIG. 4.—Probable example of detected coevolution due to frameshifts during bacterial evolution. E25 stem in *Escherichia coli*: (*A*) structure in the DCSE database, (*B*) structure as proposed by the CRW, (*C*) alternative model with the right strand shifted down by 1 nt, and (*D*) right strand shifted up by 1 nt.

Impact of the Model, Rate Distribution, and Tree Topology on the Results

We examined the robustness of our method with respect to the substitution model used in the analysis. This was done by repeating our analysis with several other models (all substitutions equal [Jukes and Cantor 1969], distinct transitions and transversions rates [Kimura 1980], and transitions and transversions + distinct GC and AU proportions [Tamura 1992]). None of the models significantly changed the results. We also tested a model with a constant distribution of rates among sites. The method performed poorly because slow-evolving sites that are not removed from the analysis lead to false-positive detected pairs.

We assessed the importance of the tree topology—assumed to be known in the analysis—by conducting the analysis on several randomly generated trees. These topologies were obtained by randomly pruning/regrafting subtrees from a neighbor-joining input tree, with the constraint of keeping nodes with bootstrap support >90% unaltered. Here again, results were essentially unchanged.

Finally, we estimated the minimum number of sequences required for detecting nonindependent sites.

Table 2
Nonstem Pairs Detected for the LSU Data Set

Site Pair										
Alignment	Escherichia coli	1GIY	1NKW	$\hat{r}_{min}$	ρ	States (Ec)	%WC	Type	Location	CRW
908, 926	338, 355	317, 334	328, 345	3.157	0.953	GC	0.975	1	B16–B17 (see fig. 3*A*)	Yes
909, 925	339, 354	318, 333	329, 344	1.293	0.858	CG	0.975	1		Yes
1160, 7936	437, 2429	416, 2407	429, 2386	0.662	0.999	UA	1.000	1	G15–B19 (see fig. 3*B*)	Yes
1318, 6551	552, 2040	531, 2018	541, 2001	0.691	0.853	CG	0.987	1	Center of the 2D representation	Yes
4032, 6550	1283, 2039	1262, 2017	1275, 2000	0.681	0.960	AU	1.000	1		Yes
2925, 3090	883, 936	862, 915	875, 927	0.663	0.992	GC	0.962	1	Between D13 and D14	Yes
4183, 4442	1365, 1424	1344, 1403	1357, 1416	0.777	0.967	UA	1.000	1	E7–E10	Yes
6151, 6161	1773, 1777	1752, 1756	1743, 1747	0.679	0.762	CG	0.949	1	Loop between E20 and E21	Yes
6192, 8146	1803, 2608	1782, 2586	1773, 2565	0.684	1.000	UU	0.000	1	E22 and loop near G20	Yes
7705, 7822	2351, 2408	2329, 2386	2308, 2365	2.302	0.902	UA	1.000	1	G12–G17	Yes
7706, 7821	2352, 2407	2330, 2385	2309, 2364	0.447	0.888	GC	1.000	1		Yes
8964, 9126	2834, 2908	2813, 2887	2788, 2862	3.157	0.914	AA	0.911	1	I1	Yes
462, 482	240, 255	219, 234	196, 211	0.415	0.825	AU	0.987	2b	Triplet (see fig. 3*C*)	Yes
462, 1176	240, 451	219, 430	196, 443	0.415	0.824	AA	0.000	2b		Yes
482, 1176	255, 451	234, 430	211, 443	0.588	0.998	UA	1.000	2b		Yes
6349, 6410	1877, 1907	1856, 1886	1848, 1869	3.122	0.836	UU	0.747	2a	E25	Yes
6351, 6408	1879, 1905	1858, 1884	1850, 1867	0.654	0.854	AG	0.013	2a		Yes
6350, 6351	1878, 1879	1857, 1858	1849, 1850	0.606	0.787	GA	0.000	3b		
6351, 6352	1879, 1880	1858, 1859	1850, 1851	0.683	0.794	AU	0.215	3b		
6352, 6408	1880, 1905	1859, 1884	1851, 1867	0.654	0.800	UG	0.013	3b		
6709, 6766	2133, 2166	2111, 2144	2094, NF	0.492	0.893	UG	0.000	2b	G4. Alpha-carbons of sites are close in *Thermus thermophilus* but sites are not in the *Deinococcus radiodurans* file	Yes
6709, 6769	2133, 2169	2111, 2147	2094, 2121	0.504	0.997	UA	1.000	1		Yes
6710, 6806	2134, 2191	2112, 2169	2095, 2165	0.396	0.954	GA	0.000	2b		Yes
6711, 6807	2135, 2192	2113, 2170	NF, 2166	0.681	0.876	UA	0.975	1		Yes
6766, 6769	2166, 2169	2144, 2147	NF, 2121	0.492	0.858	GA	0.000	1		Yes
9033, 9099	2870, 2887	2849, 2866	2824, 2841	3.150	0.771	UU	0.000	3a	Loops within I3	Yes
178, 6512	123, 2002	102, 1980	100, 1963	0.458	0.780	UG	0.000	4	B6–E21	
1147, 2352	425, 599	404, 578	417, 587	0.407	0.781	AG	0.000	4	Near B19–near D1	
1320, 4133	554, 1338	533, 1317	543, 1330	0.603	0.763	GG	0.051	4	Center and E6	
2328, 4133	581, 1338	560, 1317	569, 1330	0.609	0.790	CG	0.861	4	C1–E6	
2347, 3837	594, 1218	573, 1197	582, 1211	0.448	0.784	UG	0.013	4	Center–near D21	
2831, 6350	829, 1878	808, 1857	821, 1849	0.606	0.849	GG	0.000	4	D6 and loop near E25	
5344, 8443	1712, 2701	1691, 2679	1708, 2658	0.625	0.771	CA	0.051	4	E19–H3	

NOTE.—Positions are given for the database alignment, the *E. coli* sequence, and the *T. thermophilus* (1GIY) and *D. radiodurans* (1NKW) structures. $\hat{r}_{min}$: minimum posterior rate; ρ: estimated correlation coefficient; States: site states in *E. coli*; %WC: percentage of WC pairs among species; Type: interaction type (see table 1); CRW: is pair documented on the CRW; 2D: two-dimensional; NF: not found in PDB file.

We generated several sub–data sets by randomly selecting sequences in our data set and reperformed the analysis. Small sub–data sets (<30 sequences) lead to a large number of detected pairs, including many false positives, because high ρ values were more likely to occur under the hypothesis of independence. We used the percentage of stem pairs among detected pairs as an indicator of how well the method performs. Plotting this indicator as a function of the number of sequences used leads to a sigmoidal curve with a plateau reached at 60 sequences (results not shown).

Methodological Improvements

Our method is based on an improved substitution mapping. Improvements concern uncertainty over ancestral states and multiple substitution events. As a consequence

Table 3
Nonstem Pairs Detected for the SSU Data Set

Site Pair								
Alignment	Escherichia coli, 1GIX, 1J5E	$\hat{r}_{min}$	ρ	State (Ec)	%WC	Type	Location	CRW
513, 539	152, 169	0.683	0.885	AC	0.000	2a	Near stem 9	Yes
1432, 1498	245, 283	0.661	1.000	UU	0.000	2a	Loop with helix 12	Yes
3841, 3860	722, 733	0.681	0.932	GG	0.000	2a	Loop near helix 26	Yes
6080, 6082	1415, 1416	0.279	0.806	GG	0.000	3b	Stem 49	
6082, 6315	1416, 1485	0.268	0.807	GU	0.329	3b		
6360, 6361	1520, 1521	0.686	0.836	CC	0.025	3b	3′ end	
3793, 5636	690, 1233	0.249	0.806	GG	0.000	4	Helices 25–34	
4827, 5593	1075, 1202	0.246	0.816	UU	0.076	4	Helices 38–40	
4837, 5593	1082, 1202	0.246	0.854	AU	0.861	4		

NOTE.—See legend in table 2.

of this new mapping procedure, we use the Pearson correlation coefficient not the number of cosubstitution events (Tufféry and Darlu 2000) as a measure of the amount of coevolution for a pair of sites. In order to test which of these improvements matter, we tested them separately on the LSU data set. Our method detected 258 pairs, among which 246 were already documented as coevolving. For each method assessed, we sorted site pairs according to their correlation statistic and recorded the best 258 pairs. Among these, we counted the number of pairs documented as coevolving, thus characterizing the efficiency of various methods in a comparable way. Results are shown in table 4. It appears that removing the correction for multiple substitutions (i.e., counting one substitution when states at the bottom and top nodes on a branch are different, zero substitution otherwise) detected four additional pairs. Actually, this uncorrected method detected eight pairs not detected by our method, and our method detected four pairs not detected by the uncorrected one. Our unbiased estimate of site-specific, branch-specific number of changes does not, therefore, appear to improve the cosubstitution analysis. Averaging over all mappings slightly improves the efficiency, whereas the use of the correlation coefficient instead of the number of cosubstitutions more than doubles the number of interacting sites detected.

Table 4
Efficiency of Methodological Improvements

Method	Ancestral States	Probability	Mult. Subst.	Statistic	N. Det.
TD2000	Estimated	Marginal	No	N. of cosub.	89
	Estimated	Marginal	No	Corr. coefficient	233
	Averaged	Joint	No	Corr. coefficient	250
This article	Averaged	Joint	Yes	Corr. coefficient	246

NOTE.—Mult. Subst, correction for multiple substitutions; N. Det, number of known interacting pairs among the first 258 detected pairs; TD2000, Tufféry and Darlu (2000); N. of cosub., number of cosubstitution; Corr. coefficient, correlation coefficient.

Discussion

A model-based method for detecting coevolving site pairs is proposed. It is based on the comparison of substitution maps of the candidate site pairs. The method was applied to bacterial rRNA data, a benchmark data set for which a tremendous effort of structural characterization has been made over the last 30 years (Gutell et al. 1992).

Biochemical Constraints Underlying the Cosubstitution Process

The major part of detected pairs belongs to stems, where coevolution is due to the selection for WC pairs. The canonical WC pairs (AU and GC) are more stable in stems, although the GU pair is sometimes allowed (Tillier and Collins 1998). A striking result is the importance of these interactions in nonstem paired detected sites (interactions of type 1, 15 pairs out of 26 confirmed pairs).

More generally, hydrogen bond interactions appear to be the main source of chemical interaction in coevolving rRNA. This is the case for stem pairs, types 1 and 2. Another probable source of coevolution is strand shifting, leading one site to interact with one base on the other strand in one species and with an adjacent site in another related species. This phenomenon may explain type 3b interactions. Finally, we found one case where van der Waals interactions are a likely cause for coevolution (see fig. 3*D*).

Genetics Mechanisms Underlying the Cosubstitution Process

The near exclusivity of WC pairs in stem pairs raises the question of the underlying substitution process. Indeed, substituting a WC pair to another involves two simultaneous substitution events. Authors have hence hypothesized that the GU pair could be a less deleterious intermediate (Rousset, Pélandakis, and Solignac 1991), but data analysis shows that this may be the case only in highly variable regions (Tillier and Collins 1998). The GU pair is indeed frequently observed in several but not all stem pairs. Moreover, the GU intermediate does not explain all observed substitutions (in our data set, no stem pair with high rate is compatible with a pure GC ↔ GU ↔ AU or CG ↔ UG ↔ UA scenario). GU intermediates hence exist and have a sufficiently high lifetime to be observed in present day sequences, but other intermediates must be invoked to explain the evolution of interacting sites (see Higgs 1998 for theoretical development).

Our interpretation is that the main factor responsible for ribosome structure evolution lies in the variations of constraints across space and time. For instance, a given pair of interacting sites could temporarily allow a non-WC state provided that another site pair in the neighborhood is WC. For intermediate levels of constraint intensity, the GU pairs, but not other non-WC pairs, could be acceptable.

Performance of the Method

Our method succeeded in retrieving most of the coevolution information in the RNA data. Indeed, there are around 700 stem pairs in the LSU and 400 in the SSU of *E. coli*, which represent $\simeq 40\%$ of the sites. Among these pairs, about 510 (LSU) and 330 (SSU) are homologous stem pairs, i.e., pairs made of sites that are correctly aligned, and about 320 and 190 have a sufficiently high rate of evolution to be included in the analysis. The method succeeded in detecting 227 (LSU) and 117 (SSU) of these pairs, i.e., 67% of detectable pairs in both cases (cf. *Comparison with Secondary Structure*). Additionally, the method detected interactions that probably result from frameshifts and are not documented on the CRW.

Using a lower ρ threshold increased the number of recovered stem pairs but led to more false positives. For instance, using a ρ threshold of 0.7 instead of 0.75 for the LSU data set added 115 detected pairs including only 35 additional stem pairs and 0 documented tertiary interaction. Using a ρ threshold of 0.8, however, removed six out of seven type 4 interactions.

The power of the method relies on the ability to take among-site rate variation into account, which enabled us to remove slow-evolving sites from the analysis. With our method, the coevolutionary signal of these sites—if there is any—does not clearly rise above the noise, as shown

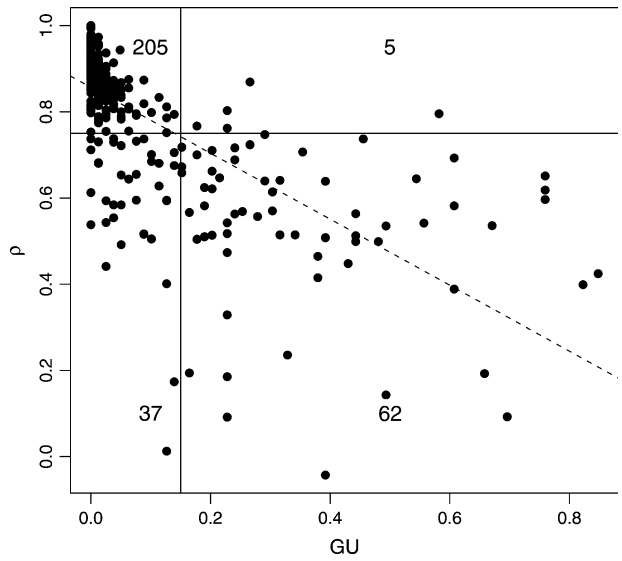

FIG. 5.—Correlation between ρ and GU content. Each dot is for an LSU stem pair with posterior rates greater than 0.3. The GU content of a site pair is the proportion of sequences in the data set showing states GU or UG for a pair. The horizontal line shows the detection level (0.75), and the vertical line was arbitrarily set to 15% of GU. Numbers correspond to frequencies in each quadrant.

by the null distribution of the ρ statistic. Considering only fast-evolving sites allowed us to choose a statistic threshold leading to solid predictions with very few false positives, which is a strength of the method. One may also wish to be more exhaustive and choose a lower threshold to retrieve a larger number of coevolving pairs but with potentially more false positives.

The method is based on the cosubstitution mapping and does not make any hypothesis concerning the underlying mechanism and possible intermediates. This generality is a strength of the method, which can be applied to rRNA and protein data, but may also be a weakness. Indeed, the method may miss some coevolving pairs containing intermediate states such as the GU pair. Figure 5 shows the ρ statistic as a function of the GU content of each LSU stem pair with a minimum posterior rate of 0.3 (the GU content of a site pair is defined as the proportion of sequences in the data set showing states GU or UG for this pair). The method succeeded in retrieving 205/242 = 85% of the stem pairs with GU content $\leq$0.15, whereas it detected only 5/67 = 7% of the pairs with GU content $>$0.15. This probably comes from the fact that stem pairs for which GU is allowed can evolve according to the pathway GC → GU → AU, for instance, a pattern that does not involve any cosubstitution if the substitutions occur on different branches. This problem apparently explains a substantial fraction of the nondetected stem pairs (fig. 5).

Perspectives

An obvious perspective in this work is the extension to proteins. The method is alphabet independent and can be applied to protein data sets as is. In a preliminary analysis of a 100-species vertebrate myoglobin data set, however, the null distribution raised high values of ρ, and no significant coevolving pair could be retrieved. The low performance of the method on this data set can be explained in two ways. First, the method assumes that all substitutions are functionally equivalent. In proteins, however, the variety of amino acid properties makes this assumption unreliable. Second, amino acid sites probably do not coevolve in a pairwise fashion as in rRNA. The method therefore needs to be improved to deal with protein data sets, for instance, by incorporating chemical distances and/or using mutivariate analysis (e.g., Fleishman, Yifrach, and Ben-Tal 2004).

Finally, our method assumes that several parameters are known, namely, tree topology and branch lengths, substitution model, and rate distribution. These parameters were assumed to be equal to their ML estimate. One may account for the uncertainty of these values by encapsulating the method in a hierarchical Bayesian framework using Monte-Carlo Markov chains (e.g., Nielsen 2002).

Supplementary Material

The alignment of LSU and SSU sequences used are available in MASE format, with site selections. Two color pictures with all detected sites plotted on the *Escherichia coli* secondary structure are also given. Figures are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Nicolas Lartillot and Olivier Gascuel for helpful discussions and David Pollock for helpful comments on a previous version of this article. This work was supported by the French Programme Inter-Establissements Publics à Caractère Scientifique et Technique "Bioinformatique" and Action Concertée Incitative "Nouvelles Interfaces des Mathématiques." TP was supported by a grant in Complexity Science from the Yeshaia Horvitz Association.

Literature Cited

Akmaev, V. R., S. T. Kelley, and G. D. Stormo. 2000. Phylogenetically enhanced statistical tools for RNA structure prediction. Bioinformatics **16**:501–512.

Altschuh, D., A. M. Lesk, A. C. Bloomer, and A. Klug. 1987. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J. Mol. Biol. **193**:693–707.

Baldwin, E., J. Xu, O. Hajiseyedjavadi, W. A. Baase, and B. W. Matthews. 1996. Thermodynamic and structural compensation in "size-switch" core repacking variants of bacteriophage T4 lysozyme. J. Mol. Biol. **259**:542–559.

Cannone, J. J., S. Subramanian, M. N. Schnare et al. (14 coauthors). 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.

De Rijk, P., J. Wuyts, and R. De Wachter. 2003. RnaViz 2: an improved representation of RNA secondary structure. Bioinformatics **19**:299–300.

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

Fleishman, S. J., O. Yifrach, and N. Ben-Tal. 2004. An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. J. Mol. Biol. **340**:307–318.

Galtier, N. 2004. Sampling properties of the bootstrap support in molecular phylogeny: influence of nonindependence among sites. Syst. Biol. **53**:38–46.

Galtier, N., and P. Boursot. 2000. A new method for locating changes in a tree reveals distinct nucleotide polymorphism vs. divergence patterns in mouse mitochondrial control region. J. Mol. Evol. **50**:224–231.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**:696–704.

Gutell, R. R., A. Power, G. Z. Hertz, E. J. Putz, and G. D. Stormo. 1992. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acids Res. **20**:5785–5795.

Harms, J., F. Schluenzen, R. Zarivach, A. Bashan, S. Gat, I. Agmon, H. Bartels, F. Franceschi, and A. Yonath. 2001. High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. Cell **107**:679–688.

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

Higgs, P. G. 1998. Compensatory neutral mutations and the evolution of RNA. Genetica **103**:91–101.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of proteins molecules. Pp. 121–123 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum-likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

Kraulis, P. J. 1991. Molscript—a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. **24**:946–950.

Mayrose, I., D. Graur, N. Ben-Tal, and T. Pupko. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. **21**:1781–1791.

Merritt, E. A., and D. J. Bacon. 1997. Raster3d: photorealistic molecular graphics. Meth. Enzymol. **277**:505–524.

Neher, E. 1994. How frequent are correlated changes in families of protein sequences? Proc. Natl. Acad. Sci. USA **91**:98–102.

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

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

Pupko, T., R. Sharan, M. Hasegawa, R. Shamir, and D. Graur. 2003. Detecting excess radical replacements in phylogenetic trees. Gene **319**:127–135.

Rousset, F., M. Pélandakis, and M. Solignac. 1991. Evolution of compensatory substitutions through GU intermediate state in Drosophila rRNA. Proc. Natl. Acad. Sci. USA **88**:10032–10036.

Savill, N. J., D. C. Hoyle, and P. G. Higgs. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics **157**:399–411.

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

Tamura, K. 1992. The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA. Mol. Biol. Evol. **9**:814–825.

Tillier, E. R., and R. A. Collins. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics **148**:1993–2002.

Tillier, E. R. M., and R. A. Collins. 1995. Neighbor joining and maximum-likelihood with RNA sequences—addressing the interdependence of sites. Mol. Biol. Evol. **12**:7–15.

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

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

Wimberly, B. T., D. E. Brodersen, W. M. Clemons, R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. 2000. Structure of the 30S ribosomal subunit. Nature **407**:327–339.

Woese, C. R. 1987. Bacterial evolution. Microbiol. Rev. **51**:221–271.

Wuyts, J., G. Perrière, and Y. Van De Peer. 2004. The European ribosomal RNA database. Nucleic Acids Res. **32**(Database issue):D101–D103.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

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

Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

Yusupov, M. M., G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. Cate, and H. F. Noller. 2001. Crystal structure of the ribosome at 5.5 Å resolution. Science **292**:883–896.

William Martin, Associate Editor

Accepted May 27, 2005