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, Pêlandakis, 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 non-independence among sites 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 biologist. 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, Tuffery 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.

*Corresponding author.

References


doi:10.1093/molbev/msi183

Advance Access publication June 8, 2005

© The Author 2005. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.

For permissions, please e-mail: journals.permissions@oupjournals.org
was calculated and compared to the expected number under
the null hypothesis of independence (Tuffery and Darlu
2000). This method does not account for multiple substitu-
tions or take into account the uncertainty in the recon-
struction of ancestral states.

Here, we introduce a new empirical Bayesian method
for the detection of coevolving positions, taking into ac-
count 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 sig-
ificant 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 infor-
mation 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 \(\Theta\), including branch lengths,
the entries in the substitution matrix, and the rate distribu-
tion parameters, is estimated using the maximum likeli-
hood (ML) method prior to the co-substitution analysis. The
HKY85 + \(\Gamma\) (Hasegawa, Kishino, and Yano 1985) substi-
tution model was used, with a four-class discretized gamma
rate distribution (Yang 1994).

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

Substitution Vectors

Let \(D_i\) be the \(i\)th site of the data set, i.e., a column of
the alignment. Let \(V_i = (v_{i1}, \ldots, v_{i, s_1}, \ldots, v_{i, s_m})\) 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 \(n_{x_p,x_q}(t)\)) times the proba-
bility of the joint reconstruction, summed over all joint
reconstructions:

\[
v_{i,b} = \sum_{\{x_1, \ldots, x_a, x_p, x_q\}} P(x_1, \ldots, x_a | D_i, \Theta) \times n_{x_p,x_q}(t).
\]  \hspace{1cm} (1)

We can rewrite this equation by grouping all reconstruc-
tions with identical \(x_p\) and \(x_q\):

\[
v_{i,b} = \sum_{x_p} \sum_{x_q} P(x_p, x_q | D_i, \Theta) \times n_{x_p,x_q}(t).
\]  \hspace{1cm} (2)

The rightmost summation is equivalent to the “two-states
joint” probability, \(P(x_p, x_q | D_i, \Theta)\):

\[
v_{i,b} = \sum_{x_p} \sum_{x_q} P(x_p, x_q | D_i, \Theta) \times n_{x_p,x_q}(t).
\]  \hspace{1cm} (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 (Tuffery
and Darlu 2000; Nielsen 2002). Because this does not account
for multiple substitutions, we chose to estimate the condi-
tional 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

\[
n_{x_p,x_q}(t) = E(N(t) | x_p, x_q).
\]  \hspace{1cm} (4)

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

\[
m_{x_p,x_q}(t) = E(N(t), x_q | x_p).
\]  \hspace{1cm} (5)

We have

\[
n_{x_p,x_q}(t) = \frac{m_{x_p,x_q}(t)}{p_{x_p,x_q}(t)},
\]  \hspace{1cm} (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

\[
M(t) = \sum_{a=1}^{m} \sum_{b=0}^{n-1} \sum_{p=0}^{\infty} O^p \cdot (Q + \Lambda)^{n-p-1},
\]  \hspace{1cm} (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 subsition 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:

\[
v_{i,b} = \sum_{r} \sum_{x_q} P(x_p, x_q, r | D_i, \Theta) \times n_{x_p,x_q,r}(t),
\]  \hspace{1cm} (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 \( t \times r_c \):

\[
v_{c,b} = \sum c \sum x_p \sum x_q n_{x_p,x_q}(t \times r_c).
\]

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:

\[
P(x_p,x_q,r_c \mid D_i, \Theta) = \frac{P(x_p,x_q,D_i \mid \Theta)}{P(D_i \mid \Theta)} = \frac{P(x_p,x_q,D_i \mid \Theta, r_c) \times P(r_c)}{P(D_i \mid \Theta)},
\]

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 \( P(D_i \mid \Theta) \) 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:

\[
\rho_{i,j} = \frac{\text{cov}(V_i, V_j)}{sd(V_i) \times sd(V_j)}.
\]

Positive values of \( \rho \) mean that the two sites tend to undergo substitutions on the same branches, whereas \( \rho \) 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 (intramolecular coevolution).

Mean Posterior Rates

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

\[
\hat{r}_i = \sum c \left( \frac{P(D_i \mid \Theta, r_c) \times P(r_c) \times r_c}{P(D_i \mid \Theta)} \right).
\]

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 + \( \Gamma \) 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 \( \rho \) 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 \( \rho \). All simulations were performed under a HKY85 + \( \Gamma \) model. Specifically, the null distribution was estimated by simulating 100,000 independent pairs and computing \( \rho \) for each pair. Three distributions of \( \rho \) 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 \( \rho \) 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 \( \rho \) 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
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 $\hat{r}$ between two corresponding substitution vectors. Positive $\hat{r}$ means that the two sites tend to substitute on the same branch (cosubstitute, see Methods). Because positive values of $\hat{r}$ may be obtained by chance, we determined the null distribution of $\hat{r}$, 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 $\hat{r}$ depends on the evolutionary rate of the two sites tested. This is seen in figure 1, where the correlation coefficient $\hat{r}$ is plotted as a function of the minimal posterior rate $\hat{r}_{\text{min}}$. The relationship between $\hat{r}$ and $\hat{r}_{\text{min}}$ appears complex. Slow-evolving sites confer a high variance to $\hat{r}$ estimates, so that high values of $\hat{r}$ 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, $\hat{r}$ tends to be positive and positively correlated with $\hat{r}_{\text{min}}$. 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 ($\hat{r}_{\text{m}}$) 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 $\hat{r}_{\text{min}}$ is greater than $\hat{r}_m = 0.3$ (LSU) or 0.2 (SSU) and (2) its $\hat{r}$ is greater than $\hat{r}_{\text{m}} = 0.75$ (LSU) or 0.8 (SSU). Virtually, no values of $\hat{r}$ 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 $\hat{r}_{\text{min}} > \hat{r}_m$ and $\hat{r} > \hat{r}_{\text{m}}$ are defined as coevolving pairs.

Detecting the Coevolving Sites

Having set the two $\hat{r}_m$ and $\hat{r}_{\text{m}}$ 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 $\hat{r}$ 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 $\hat{r} > \hat{r}_m$. 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 $\hat{r}_m$ value in the simulated data sets. All detected pairs with $\hat{r} > \hat{r}_m$ and $\hat{r}_{\text{min}} > \hat{r}_m$ 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 tRNA given in the tRNA database (Wuyts, Perrière, and Van De Peer 2004). Out of the 258 coevolving sites in the LSU data, 225 (87%) were
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. thermophillus* (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 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.

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.

Table 1

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

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...
one site from LSU and one site from SSU. The null distribution is different from the intra-subunit ones: fewer site pairs with high $q$ values are observed. This may be because the estimated branch lengths are different between the two subunits. No high value of $q$ is observed from the real data too. We checked the 10 pairs with greatest $q$ 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.

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.
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 \( q \) 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 of these improvements, we have been able to detect coevolving positions in a molecular sequence with higher accuracy.

### Table 3

**Nonstem Pairs Detected for the LSU Data Set**

<table>
<thead>
<tr>
<th>Site Pair</th>
<th>1GIY</th>
<th>INKW</th>
<th>( \hat{\rho} )</th>
<th>States (Ec)</th>
<th>%WC</th>
<th>Type</th>
<th>Location</th>
<th>CRW</th>
</tr>
</thead>
<tbody>
<tr>
<td>908, 926</td>
<td>338, 355</td>
<td>317, 334</td>
<td>0.683</td>
<td>A</td>
<td>0.000</td>
<td>3a</td>
<td>Loops within I3</td>
<td>Yes</td>
</tr>
<tr>
<td>178, 6512</td>
<td>123, 2002</td>
<td>102, 1980</td>
<td>0.661</td>
<td>UU</td>
<td>0.000</td>
<td>3a</td>
<td>Loop with helix 12</td>
<td>Yes</td>
</tr>
<tr>
<td>1147, 2352</td>
<td>425, 599</td>
<td>404, 578</td>
<td>0.458</td>
<td>GG</td>
<td>0.000</td>
<td>3b</td>
<td>Near helix 26</td>
<td>Yes</td>
</tr>
<tr>
<td>2328, 4133</td>
<td>581, 1338</td>
<td>560, 1317</td>
<td>0.609</td>
<td>GC</td>
<td>0.000</td>
<td>3a</td>
<td>Loops near helix 26</td>
<td>Yes</td>
</tr>
<tr>
<td>2347, 3837</td>
<td>594, 1218</td>
<td>573, 1197</td>
<td>0.790</td>
<td>UG</td>
<td>0.000</td>
<td>3b</td>
<td>Center–near D21</td>
<td>Yes</td>
</tr>
<tr>
<td>2831, 6350</td>
<td>829, 1878</td>
<td>808, 1857</td>
<td>0.894</td>
<td>GG</td>
<td>0.000</td>
<td>3b</td>
<td>D6 and loop near E25</td>
<td>Yes</td>
</tr>
<tr>
<td>5344, 8443</td>
<td>1712, 2701</td>
<td>1691, 2679</td>
<td>0.625</td>
<td>UA</td>
<td>0.000</td>
<td>3b</td>
<td>E9–H3</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**NOTE.**—Positions are given for the database alignment, the *E. coli* sequence, and the *T. thermophilus* (1GIY) and *D. radiodurans* (1NKW) structures. \( \hat{\rho} \): minimum posterior rate; \( \hat{\rho} \): 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.

---

**Table 2**

**Nonstem Pairs Detected for the SSU Data Set**

<table>
<thead>
<tr>
<th>Site Pair</th>
<th>1GIY</th>
<th>INKW</th>
<th>( \hat{\rho} )</th>
<th>States (Ec)</th>
<th>%WC</th>
<th>Type</th>
<th>Location</th>
<th>CRW</th>
</tr>
</thead>
<tbody>
<tr>
<td>908, 926</td>
<td>338, 355</td>
<td>317, 334</td>
<td>0.683</td>
<td>A</td>
<td>0.000</td>
<td>3a</td>
<td>Loops within I3</td>
<td>Yes</td>
</tr>
<tr>
<td>178, 6512</td>
<td>123, 2002</td>
<td>102, 1980</td>
<td>0.661</td>
<td>UU</td>
<td>0.000</td>
<td>3a</td>
<td>Loop with helix 12</td>
<td>Yes</td>
</tr>
<tr>
<td>1147, 2352</td>
<td>425, 599</td>
<td>404, 578</td>
<td>0.458</td>
<td>GG</td>
<td>0.000</td>
<td>3b</td>
<td>Near helix 26</td>
<td>Yes</td>
</tr>
<tr>
<td>2328, 4133</td>
<td>581, 1338</td>
<td>560, 1317</td>
<td>0.609</td>
<td>GC</td>
<td>0.000</td>
<td>3a</td>
<td>Loops near helix 26</td>
<td>Yes</td>
</tr>
<tr>
<td>2347, 3837</td>
<td>594, 1218</td>
<td>573, 1197</td>
<td>0.790</td>
<td>UG</td>
<td>0.000</td>
<td>3b</td>
<td>Center–near D21</td>
<td>Yes</td>
</tr>
<tr>
<td>2831, 6350</td>
<td>829, 1878</td>
<td>808, 1857</td>
<td>0.894</td>
<td>GG</td>
<td>0.000</td>
<td>3b</td>
<td>D6 and loop near E25</td>
<td>Yes</td>
</tr>
<tr>
<td>5344, 8443</td>
<td>1712, 2701</td>
<td>1691, 2679</td>
<td>0.625</td>
<td>UA</td>
<td>0.000</td>
<td>3b</td>
<td>E9–H3</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**NOTE.**—See legend in table 2.

---

Detecting Coevolving Positions in a Molecule 1925
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.

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. 3D).

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 hypothe-

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TD2000</td>
<td>Estimated</td>
<td>Marginal</td>
<td>No</td>
<td>N. of cosub</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Estimated</td>
<td>Marginal</td>
<td>No</td>
<td>Corr. coefficient</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>Averaged</td>
<td>Joint</td>
<td>No</td>
<td>Corr. coefficient</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>Averaged</td>
<td>Joint</td>
<td>Yes</td>
<td>Corr. coefficient</td>
<td>246</td>
</tr>
</tbody>
</table>

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.

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 ≈ 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 $\rho$ threshold increased the number of recovered stem pairs but led to more false positives. For instance, using a $\rho$ 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 $\rho$ 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 sized 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.
numbers of a site pair is the proportion of sequences in the data set showing states GU or UG for a pair. The method may miss some coevolving pairs containing intermediate states such as the GU pair. 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 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 multivariate 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-Établissements 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 Yeshaya Horvitz Association.

Literature Cited


Neher, E. 1994. How frequent are correlated changes in families

William Martin, Associate Editor
Accepted May 27, 2005