Base Pairing Constraints Drive Structural Epistasis in Ribosomal RNA Sequences

Julien Y. Dutheil,∗1,2 Fabrice Jossinet,3 and Eric Westhof3

1Bioinformatics Research Center (BiRC), Aarhus University, Århus, Denmark
2Institut des Sciences de l’Évolution—Montpellier, Université Montpellier 2, Montpellier, France
3Architecture et Réactivité de l’ARN, Institut de Biologie Moléculaire et Cellulaire du CNRS, Université de Strasbourg, Strasbourg, France

*Corresponding author: E-mail: julien.dutheil@univ-montp2.fr.

Associate editor: Manolo Gouy

Abstract

It has long been accepted that the structural constraints stemming from the 3D structure of ribosomal RNA (rRNA) lead to coevolution through compensating mutations between interacting sites. State-of-the-art methods for detecting coevolving sites, however, while reaching high levels of specificity and sensitivity for Watson–Crick (WC) pairs of the helices defining the secondary structure, only scarcely reveal tertiary interactions occurring at the level of the 3D structure. In order to understand the relative failure of coevolutionary methods to detect such interactions, we analyze 2,682 interacting sites derived from high-resolution structures, which include a comprehensive data set of rRNA sequences from Archaea and Bacteria. We report a striking difference in the amount of coevolution between WC and non-WC pairs. In order to understand this pattern, we derive fitness landscapes from the geometry of base pairing interactions and construct neutral networks of substitutions for each type of interaction. These networks show that coevolution is a property of WC pairs because, unlike non-WC pairs, their landscapes exhibit fitness valleys, a single mutation in a WC pair resulting in a fitness drop. Second, we used the inferred neutral networks to estimate the level of constraint acting on each type of base pair and showed that it correlates negatively with the observed rate of substitutions for all non-WC pairs. WC pairs appear as outliers, fixing more substitutions than expected according to their level of constraint. We here propose that the rate of substitution in WC pairs is due to coevolution resulting from constraints acting at intermediate levels of organization, namely the one of the helical stem with its forming WC pairs. In agreement with this hypothesis, we report a significant excess of intrahelical, inter-WC pairs coevolution compared with interhelices pairs. Altogether, these results show that detailed biochemical knowledge is required and has to be incorporated into evolutionary reasoning in order to understand the fine patterns of variation at the molecular level. They also demonstrate that coevolutionary analysis provides almost exclusively 2D information and only little 3D signal.

Key words: ribosomal RNA, coevolution, fitness landscape, base pair geometry, structural constraint, epistasis.

Introduction

The 3D structures of ribosomal RNAs (rRNAs) are shaped by numerous interactions between residues of the nucleotide sequence (Ban et al. 2000; Wimberly et al. 2000; Harms et al. 2001; Schuwirth et al. 2005). The secondary structure of RNA (double-stranded regions, namely helices or stems) relies on the Watson–Crick (WC) complementarity between nucleotides. This complementarity forms the basis of ab initio secondary structure prediction methods from sequence alignments (Fox and Woese 1975; Noller and Woese 1981; James et al. 1988; Gutell et al. 1994). RNA molecules further adopt a 3D (tertiary) structure stabilized by a variety of interactions between the nucleotides, as revealed by the recent availability of high-resolution crystal structures (Leontis and Westhof 2001). A surprising result, however, obtained using detailed evolutionary modeling is that even state-of-the-art methods only predict a tiny fraction of these tertiary interactions while they recover most of the secondary structure (Dutheil et al. 2005; Yeang et al. 2007). Several explanations can be put forward to explain this discrepancy. First, as the tertiary structure of ribosomes has been resolved much longer after the secondary structure, researchers might have considered non-WC pairs as false positives of their methods and focused on the detection of stems. This hypothesis, however, does not hold for recent works for which 3D structures of ribosomes were available. Alternatively, the patterns of substitutions of sites involved in tertiary structure significantly differ from those involved in secondary structure interactions. For instance, such sites could be highly constrained and appear as totally conserved in an alignment, hence depriving any evolutionary analysis of signal. The predictive signal of WC pairs results from their strong coevolutionary pattern, which has been studied in great details in several types of RNA molecules and organisms. This coevolution is due to the occurrence of compensating mutations that restore the fitness drop after a mutation has occurred at the interacting site on the opposite strand. From a phylogenetic perspective, this implies that these sites will undergo simultaneous substitutions.
Whether non-WC pairs can undergo cosubstitutions has never been assessed. High-resolution crystal structures have provided the field with a detailed inventory of all base-pairing interactions (Leontis et al. 2002; Stombaugh et al. 2009), leading to an unique opportunity to understand how patterns of sequence variation are shaped by structural constraints. Base pairing interactions can be classified according to their interacting edges and the relative orientation of the riboses. Each nucleotide has three edges, Watson–Crick (W), Hoogsteen (H), and sugar (S), leading to six possible combinations (WW, WH, WS, HS, HH, and SS), which can each be either in the cis (C) or trans (T) orientations (Leontis and Westhof 2001). Essential features of RNA structure result from base pairing, base stacking, and backbone conformation, notably the distance between two interacting sites. The term isostericity was coined to describe pairs that have a similar backbone distance for a given interaction type, and isostericity matrices are used to summarize the information for all pairs (Leontis and Westhof 2001). From an evolutionary point of view, this implies that substitutions between isosteric pairs are neutral, whereas substitutions altering the isostericity are deleterious (fig. 1). Isostericity matrices can hence be used to elucidate the underlying fitness landscape and predict evolutionary features like the occurrence of coevolution (defined as nonindependent site evolution or intragenic epistasis) or the rate of substitutions.

We here report the first attempt to build neutral networks (Gavrilets 2004) of interacting pairs based on fitness landscapes derived from structural constraints of the folded molecule. We gather a large data set of nonredundant Bacteria and Archaea sequences for which high-resolution structures are available. We infer the substitution histories of known interacting sites and study their patterns in the light of the reconstructed neutral networks.

**Material and Methods**

**Structural Alignments**

Raw sequences were retrieved from the European rRNA database (Wuyts et al. 2004). We used the S2S application to construct the structural alignments (Jossinet and Westhof 2005). S2S allows the user to align the RNA sequences against a reference molecule for which a structural mask has been calculated. From a crystal structure, S2S identifies and classifies all the base pairs formed in the nucleic acid structures according to the Leontis–Westhof classification (Leontis et al. 2002), using the RNAVIEW algorithm (Jossinet and Westhof 2005). S2S then automatically recovers this set of base pairs to generate a structural mask. During the manual editing of the multiple alignment, this mask is used to display the conservation of base pairs,
Evolutionary Analyses
Performing evolutionary analysis on all the sequences requires considerable computer resources. Moreover, all the sequences are not informative due to redundancy. We submitted the structural alignments to the following preprocessing pipeline:

1. Remove all unidentified sequences from the data set.
2. Build a BioNJ tree from the resulting data set using a Jukes–Cantor substitution model and equal substitution rates across sites.
3. Submit the resulting tree to phylogenetic sampling (program BppPhySamp from the Bio++ programs suite package; Dutheil and Boussau 2008): if two sequences have less than 0.005 substitution per aligned position, only the longest one is kept.

The resulting alignments were used to build a maximum likelihood phylogenetic tree, using the PhyML program, with a general time reversible substitution model, with gamma (four classes) + invariant rates across sites distribution. From this step and through all the following analysis, only positions without gaps were considered. Table 1 provides a summary of the final data sets used in the analysis.

Estimation of Substitution Rates
We used the empirical Bayes approach to estimate site-specific evolutionary rates (Mayrose et al. 2004), using the General Time Reversible + gamma + invariant model estimated by PhyML. This rate is a relative adimensional rate because a calibration point, for instance, using the fossil record, is required to obtain absolute rates in numbers of substitutions per million years.

Measuring Coevolution
The degree of “nonindependent evolution” for a pair of sites was assessed by 1) measuring the cooccurrence of substitutions for the pair of sites and 2) computing the probability that such a cooccurrence might be observed by chance under the null hypothesis of independent evolution. Step 1 is achieved through substitution mapping: The probabilistic substitution model estimated during the phylogenetic analysis is used to compute the average number of substitutions that occurred on a particular branch of the tree for a particular site (Dutheil et al. 2005). The set of all numbers of substitutions for all branches for a particular site is called the “substitution vector” of the site. We then measured the amount of coevolution for a pair of sites by taking the Pearson correlation coefficient of the corresponding substitution vectors. The null distribution of such a statistic is not analytically tractable because the observations—that is, the sequences—are not independent due to their shared evolutionary history. To assess the null distribution, we relied on a parametric bootstrap approach. We simulated several pairs of independent sites using the model estimated from the data, but assuming independent evolution, and recorded for each simulated pair the obtained correlation coefficient. By comparing the correlation obtained from real data and the distribution obtained from simulations, we can compute the P value of the observed correlation. In this work, we introduced a modification of the algorithm from Dutheil and Galtier (2007). As we tested candidate positions and did not perform a search over all possible groups, we could relax a few approximations in the simulations. For each candidate pair, we simulated two sites with substitution rates close to the posterior estimates of the corresponding sites in the data and recorded the resulting coevolution measure. We used the norm of the substitution vector as an indication of the substitution rate. We performed 1,000 simulations per tested pair and computed the P value for a group as \((N + 1)/(1001)\), where \(N\) is the number of simulated pairs with a coevolution measure greater or equal to the observed one in the data. A false discovery rate method was then used to correct for multiple testing (Benjamini and Hochberg 1995). The new algorithm introduced here is available in the CoMap program, version 1.3.0, distributed along with the source code at http://home.gna.org/comap/.

Statistical Analysis
All analyses have been conducted using the R software (R Development Core Team 2008). In order to test for intrahelical coevolution involving noninteracting sites, we contrasted two types of site pairs. A first set was computed by sampling sites within helices, excluding any already documented direct interaction. If a helix is less than 10 base pairs long, an exhaustive sampling was performed. Otherwise, ten pairs were drawn randomly. Another set was obtained by sampling 1,000 pairs of sites, which belong to distinct helices and which do not participate in any known documented interaction. This sampling procedure ensures that the evolutionary rate distribution of the sites of the two sets is the same and allows testing specifically for extra WC intrahelices coevolution, whereas the interhelices set being a control for which no strong coevolution signal is expected. The coevolution signal was evaluated using the same procedure as presented before, and the distributions of \(P\) values for all pairs for the two sets were compared using a Wilcoxon unilateral test. Data sets and detailed results are available as supplementary material, Supplementary Material online.

### Table 1. Description of the Structure and Sequence Data.

<table>
<thead>
<tr>
<th>Species</th>
<th>SU</th>
<th>Number of nonredundant sequences</th>
<th>Number of sites</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em></td>
<td>LSU</td>
<td>2A8B</td>
<td>951</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>SSU</td>
<td>2AVY</td>
<td>567</td>
</tr>
<tr>
<td><em>Halocarcula marismortui</em></td>
<td>LSU</td>
<td>IJ2</td>
<td>1,164</td>
</tr>
</tbody>
</table>

SU, subunit; LSU, large subunit; SSU, small subunit.
Results and Discussion
In order to assess the importance of coevolutionary patterns in real data, we manually aligned a set of 16S and 23S rRNA sequences extracted from fully sequenced genomes against the available crystal structures of rRNAs. We aligned the Archaea 23S rRNA sequences against the resolved crystal structure of *Haloarcula marismortui* ribosome (Protein Data Bank [PDB] ID: 1JJ2; Klein et al. 2001), and the bacterial 16S and 23S rRNA sequences against the structures of the *Escherichia coli* ribosome (PDB ID: 2AVY and 2AWB, respectively; Schuwirth et al. 2005). We further edited the alignments to remove redundant sequences (see Material and Methods). A summary of the data is shown in Table 1. We reconstructed a maximum likelihood phylogenetic tree for each of the three data sets and used the estimated parameters, including phylogeny, to infer 1) the substitution map with number of substitutions that occurred on each branch of the tree of each column in

<table>
<thead>
<tr>
<th>Interaction type</th>
<th>Cis orientation</th>
<th>Trans orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHH</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>CHS</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>CSS</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>CWH</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>CWS</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>THH</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>THS</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>TSS</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>TWH</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>TWS</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>TWW</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

![Figure 2](image-url) Proportion of coevolving pairs as a function of interaction type. The actual numbers of significant pairs over the total number of pairs are shown on top of the bars. Interactions are classified according to their interacting edges (W, Watson–Crick; H, Hoogsteen; and S, sugar) and their orientation (C, cis and T, trans).

![Figure 3](image-url) Site-specific evolutionary rate and interaction geometry. (a) Logarithm of substitution rates plotted as a function of interaction type. The widths of the boxes are proportional to the square root of the number of observations. Nonoverlapping notches show significant difference in median. The “No struct.” category contains sites that are not involved in any interaction. The median of the “No struct.” category is plotted as a dotted horizontal line on other plots for comparison. H, Hoogsteen edge; S, sugar edge; and W, Watson–Crick edge. (b) Proportion of significantly coevolving interactions as a function of the rate of substitution of the sites. Rate categories have been designed in order to represent the same amount of data (25% of interactions each).
the alignment (Dutheil et al. 2005) and 2) the site-specific substitution rates. For each documented interaction, we then quantified the amount of coevolution that occurred between two positions by calculating the Pearson correlation coefficient of the corresponding substitution maps. This coevolution measure hence assesses the cooccurrence of substitutions in the branches of the phylogenetic tree (cosubstitutions; Tuffery and Darlu 2000). We evaluated the significance of the measured correlations while accounting for phylogenetic relationships by a parametric bootstrap approach.

After correction for multiple testing (Benjamini and Hochberg 1995) assuming a false discovery rate of 5%, we found that 42% of the tested interactions are significantly coevolving. The actual proportions range from 54% for \textit{E. coli} large subunit (LSU) to 29% for \textit{H. marismortui} LSU (this is due to a smaller number of sequences available for the Archaea data set). The results show a significant discrepancy in proportion of coevolving pairs between the interaction types (fig. 2; \( P < 0.1\%), chi-squared test). The WC interactions, cis-WW and, to some extent, trans-WW, are the most frequently significant, with 57% of interactions being significantly coevolving, whereas this proportion is only 7% for non-WC interactions (a value which is slightly—yet significantly—more than expected under our false discovery rate of 5%; binomial test, \( P = 1.1\%\)). If the WC interactions are excluded, the differences in the proportion of coevolving pairs between the remaining interaction types are no longer significant. We report a significant difference in substitution rates according to the geometry of interactions, demonstrating distinct evolutionary constraints (fig. 3a). Positions involved in cis interaction with their WC edge evolve at a rate comparable

**FIG. 4.** Neutral networks deduced from isostery matrices for all interaction types. Lines connect states with equivalent fitness that can be reached by single neutral mutation events at one of the two interacting sites.

**FIG. 5.** Substitution rate as a function of IDI. The bars show a 95% confidence interval of the means. The size of points is proportional to the square root of the number of observations. The biggest point, with IDI = 4.1 corresponds to the CWW pairs. The regression line was computed without the CWW pairs.
with the sites not involved in any documented interaction. On the other hand, positions involved in trans interactions and in interactions using the Hoogsteen edge underwent fewer substitutions. As slow sites have less signal for detecting coevolution, we compared the amount of coevolution between the geometrical types for sites with equivalent rates. In this way, we showed that WC interactions are still those exhibiting the highest proportion of coevolving pairs, demonstrating that the scarcity of significant interactions is not due to a lack of power of the method but reflects a difference in structural constraints (fig. 3b).

The lack of coevolution signal in tertiary interactions can be interpreted in terms of neutral networks. Neutral networks connect different sequence states, distant by only one mutation, that possess the same fitness (Gavrilets 2004). Coevolution—defined as nonindependent evolution or intragenic epistasis—occurs when the change from one state to another involves simultaneous substitution events at different positions. If the two states belong to the same network, the coevolution is dispensable as the new state can also be reached by several single substitution events. When the two states are on two unconnected networks, the coevolution is mandatory, as a mutation at one of the two interacting sites breaks functionality, whereas substitutions at both sites conserve it. The system hence has to go through a fitness valley, and the second mutation is therefore said to compensate the first one (Stephan 1996; Gavrilets 2004). We used the concept of base pair isostericity as a proxy for structural functionality and derived all neutral networks from the available isostericity matrices presented in Stombaugh et al. (2009; fig. 4). Such an analysis reveals epistatic changes with the possibility of compensating mutations in all but the cis-SS and trans-SS interaction types. A base pair in a given site is said to be epistatic when its presence affects the probability of fixation of a mutation at another site in contact. For the SS types, it appears that all interactions have the same fitness, resulting in a full independence of the evolution of individual sites. The WC pairs (cis- and trans-WW), however, are the only types for which coevolution is mandatory, with the exception of the GU wobble pair, which can be a stable intermediate when transitioning between GC and AU. We hence hypothesize that in non-WC interactions, evolution proceeds mostly along neutral networks, resulting in a decoupling of the substitution processes at interacting sites and an apparent independent evolution.

The geometry of interactions, as measured by isostericity matrices, also enables the prediction of the relative amount of substitutions at each site. Interactions with fewer isosteric pairs are expected to undergo less
substitutions because of a greater number of mutations with negative effect. This is measured by the IsoDiscrepancy Index (IDI; Stombaugh et al. 2009). Our data show a weak but highly significant negative correlation between the mean evolutionary rate for each pair of interacting sites with the average IDI for the interaction family after cis-WW pairs have been removed (Kendall’s rank correlation tau = -0.09, P = 5.926 x 10^-05; see fig. 5). The cis-WW pairs have an abnormally high number of substitutions given the structural constraints imposed by their geometry. This implies that for a given pair of sites, nonisosteric mutations can be neutral or nearly neutral, which is apparently in contradiction with the strong coevolution signal and the high number of cosubstitutions observed. The occurrence of wobble pairs as intermediates does not account for this pattern because wobbling can only explain a certain subset of observed transitions. We propose a more general explanation that implies variation of local constraints in space and time, resulting from constraints acting at a higher level of structural organization. We hypothesize that the cost of nonisosteric mutations occurring in WC pairs is compensated by interactions between intrahelical pairs. If a minimal number of WC pairs is required for a helix to be stable, then the loss of a WC pair can be compensated by the gain of a WC pair at another position in the same helical stem, which predicts that sites, not in direct interaction, could coevolve. This was tested by comparing two sets of sites: one set of intrahelical sites and one set of interhelicals sites. Site pairs corresponding to a WC interaction were prohibited as we aimed at looking for coevolution between and not within pairs (see fig. 6).

The comparison between the two sets shows a significantly higher coevolution for the former than the latter in the Bacteria data sets (Wilcoxon test, P = 1.078 x 10^-10, 0.0068, 0.1842 for Bacteria large and small subunits and Archaea large subunit, respectively; see fig. 6). The Archaea data set also displays a smaller average P value but is not significant possibly because of the smaller number of sequences. In addition, we computed the “helical” distance for each pair of sites in the intrahelical data set. This distance is the number of nucleotides separating the two sites if they were on the same helix strand. We report a significant negative correlation between the coevolution signal and the distance between sites,
suggesting that inter-WC coevolution involves short-range interactions (Kendall’s tau, \( P = 5.34 \times 10^{-6}, 0.0012, 0.1641; \) see fig. 7). This hypothesis also predicts that between-pair coevolution is more likely to occur in long than in short stems, resulting in sites being more constrained and hence more slowly evolving in short stems. Our data show that these two variables are indeed positively and significantly correlated in the three data sets (Kendall’s tau = 0.22, \( P < 2.2 \times 10^{-16}; \) see fig 8). Such a correlation has previously been reported for bacterial RNase P RNA and transfer RNA (Parsch et al. 2000), vertebrate introns (Piskol and Stephan 2008), but not for rRNA. The generalization of these results to all types of RNA hence supports that the hierarchical organization of the molecular structure is responsible for the occurrence of cosubstitutions.

**Conclusions**

Although geometrically isosteric pairs may present different intrinsic energies or energetic stabilizations through stacking and interactions with the local environment, these results demonstrate that the geometry of RNA 3D interactions enables the prediction of patterns of epistasis at the molecular level. The substitution patterns resulting from tertiary interactions are distinct from the ones resulting from secondary interactions. These conclusions add evolutionary links to the known structural and dynamic differences between secondary and tertiary structures and interactions (Brion and Westhof 1997; Tinoco and Bustamante 1999). Isosteric (neutral) substitutions in WC interactions imply a coupling of single-site substitutions through compensating mutations. Conversely, the geometrical constraints on non-WC interactions predict the existence of neutral networks, with isosteric pairs being potentially substituted by single-site mutations, therefore decoupling the evolution of interacting sites. These results provide a basis for the understanding of molecular adaptive landscapes in RNA and corroborate the conclusions drawn from an analysis of tertiary motifs that display a high level of molecular interchangeability and neutrality (Lescoute and Westhof 2006; Cruz and Westhof 2009). Furthermore, they suggest that methods relying solely on alignment have little signal to predict tertiary structure. In order to deduce 3D contacts, such methods have therefore to be complemented by structural information stemming from sequence and structural databases accompanied by preliminary modeling.

**Supplementary Material**

Supplementary material is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

The authors would like to thank Thomas Bataillon, Bastien Boussau, Mikkel Schierup, Freddy Bugge Christiansen, and Eva Stukenbrock for critical reading and helpful comments on a previous version of this manuscript. This publication is contribution 2010-015 of the Institut des Sciences de l’Évolution de Montpellier (UMR 5554—CNRS). This work was supported by the French Agence Nationale de la Recherche “Domaines Emergents” (ANR-08-EMER-011 “PhylAriane”). Some of this work was made possible by a grant from the Human Frontier Science Program (RGP0032/2005-C to E.W.). J.Y.D. was funded by European Research Area in Plant Genomics (ERA-PG) A Relatives.

**References**


