Thirteen dubious ways to detect conserved structural RNAs

William Gao\textsuperscript{1}, Ann Yang\textsuperscript{2}, Elena Rivas\textsuperscript{2*}
\textsuperscript{1} Department of Genetics
University of Pennsylvania, Philadelphia, PA 19104, USA
\textsuperscript{2} Department of Molecular and Cellular Biology,
Harvard University, Cambridge, MA 02138, USA
\textsuperscript{*} Corresponding author: elenarivas@fas.harvard.edu

Abstract

Covariation induced by compensatory base substitutions in RNA alignments is a great way to deduce conserved RNA structure, in principle. In practice, success depends on many factors, importantly the quality and depth of the alignment and the choice of covariation statistic. Measuring covariation between pairs of aligned positions is easy. However, using covariation to infer evolutionarily conserved RNA structure is complicated by other extraneous sources of covariation such as that resulting from homologous sequences having evolved from a common ancestor. In order to provide evidence of evolutionarily conserved RNA structure, a method to distinguish covariation due to sources other than RNA structure is necessary. Moreover, there are several sorts of artifactually generated covariation signals that can further confound the analysis. Additionally, some covariation signal is difficult to detect due to incomplete comparative data. Here, we investigate and critically discuss the practice of inferring conserved RNA structure by comparative sequence analysis. We provide new methods on how to approach and decide which of the numerous long non-coding RNAs (lncRNAs) have biologically relevant structures.

Is it a conserved RNA structure, or not? ...or can we even tell?

In this article we discuss how to identify conserved RNA structure, that is, how to determine if a genomic region is under selective pressure to maintain some particular RNA secondary structure.

Identifying evolutionarily conserved structural RNAs requires the use of comparative data. A conserved RNA structure results in a pattern of compensatory base pair substitutions that tend to preserve the base pairs [1]. As a quantitative measure of the amount of compensatory base pair substitutions, covariation has been used very successfully to infer the structure of many fundamental structural RNAs such as tRNAs [2], ribosomal RNA [3, 4], group I introns [5], and ribozymes [6], to name a few. Numerous computational methods have been designed to determine RNA structures using covariation [7, 8, 9, 10, 11]. Many others have taken advantage of covariation in related tasks such as RNA homology searches [12, 13, 14], RNA structural alignments [15, 16], and RNA genefinding [17, 18, 19, 20]. However, identifying novel conserved RNA structures is a question fundamentally different from RNA structure prediction, RNA homology searches or structural alignment.

Unlike structure prediction where we assume there is a conserved structure, and unlike homology search or alignment where we are happy to use primary sequence conservation alone, asserting the
presence of a conserved RNA structure requires knowing if the covariation we see is more than what we expect from other biological explanations. The question is not what are the base pairs in this structure, but instead, does this RNA have a conserved RNA structure to start with?

Identifying conserved RNA structure is necessarily a computational analysis question. Essentially, any RNA has structure; SHAPE and other chemical probing methods [21, 22, 23, 24, 25] will still give a signal in random sequences, and will not be able to distinguish a random sequence from one under selective pressure to conserve one or several particular structures. The experiment one would like to do to demonstrate the importance of a particular RNA structure is compensatory mutation analysis. That is, to show that single mutations that disrupt either side of an inferred base pair disrupt function, but a double mutant that restores base pairing also restores the function. Evolution has already done that experiment on a massive scale. Computationally, we can analyze the sequences that have successfully mutated and maintained structure and function in vivo.

But why care specifically about conserved RNA structure? As stated above and as reasoned by Vicens & Kieft [26], RNA sequences of any kind, including mRNAs or even random ones, will produce some folding under biological conditions. It is also likely that for a given transcript many different structures coexist, based on the many different theoretical folding possibilities that present comparable thermodynamic stabilities. The ensemble of structures, likely transient and dynamic, may be related to an overall functional state of the RNA. For instance, mRNA global structuredness seems to increase RNA stability, which has implications in translation efficiency [27]. However, when RNA is involved in some specific function, often it relies on conserved RNA structures involved in processes conserved across species. There are many well known classes of functional conserved structural RNAs involved in fundamental cellular functions. An excellent compilation of conserved RNA sequence and structure can be found in the database of RNA families Rfam [28]. We hypothesize that other RNAs with conserved structures remain to be discovered which are likely to lead us to specific and perhaps novel cellular mechanisms.

Identifying novel conserved RNA structures is really two computational tasks: (1) To determine whether the amount of observed pairwise covariation is more than we expect if there is not a conserved structure. For this we need a null hypothesis to compare to [29, 30]. (2) To determine how much pairwise covariation we expect to see if there were a conserved structure, given how much sequence variation there is available in the alignment because sometimes, with highly conserved regions, there will not be enough variation to know if there is an evolutionarily conserved structure. The first problem is about statistical significance, the second is about statistical power.

We have developed a computational tool (R-scape) that addresses both questions [31, 32], which is succinctly described in Figure 1. Here, we will use R-scape as an specific example, but the main considerations apply broadly.

For the first task (statistical significance), the main source of confounding pairwise covariation comes from the phylogenetic relatedness of the sequences. The covariation observed between pairs of columns in an input multiple sequence alignment is compared against that of related synthetic alignments generated under the null hypothesis, from which we can assign an E-value (that is, the expected number of null pairs of columns with a similar covariation score or bigger) for all pairs in the input alignment. We say that a base pair significantly covaries if it has an R-scape E-value < 0.05, by default.

For the second task (statistical power), we use a source of true conserved structures such as Rfam [28]. Statistical power is how much significant covariation to expect if there were a conserved structure given the sequence variation observed in the alignment. The power of a base pair is a
Figure 1: The comparative signals utilized by R-scape. We use a toy alignment of 5 sequences (S) and 15 nts in length (L) to show the principles behind R-scape. Given the alignment, a phylogenetic tree is produced. The alignment and the tree are used both to generate null alignments (Figure 3a), as well as to estimate the number of substitutions per position. The null alignments are used to generate a null distribution of covariation scores due strictly to the phylogenetic relationships amongst the sequences. The substitutions are used to estimate covariation power, that is, the probability of the pair been called significant if it were a base pair. R-scape uses both covariation and power to decide on whether the alignment supports a conserved structure, rejects the hypothesis of a conserved structure or cannot make any inference about whether there is a conserved RNA structure or not. For this toy alignment and using default setting, R-scape identifies five significantly covarying base pairs with E-values smaller than 0.05.

probability between 0 and 1 that the base pair is expected to be detected as significantly covarying based on the variability that is present.

Using covariation and power we can distinguish between three different scenarios of RNA conservation: significant covariation indicates the presence of a conserved base pair. In the absence of covariation, power allows us to distinguish two other possible scenarios: lack of power together with lack of covariation means that the alignment is too conserved at the primary sequence level to give any information about whether there is a conserved structure or not. On the other hand, power in the absence of covariation is evidence against a conserved base pair.

Using covariation and power, we can re-shape the question of RNA structure inference. The combination of covariation and power lets us predict negative pairs. A negative pair has plenty of variation, enough that if the two aligned columns form a conserved base pair we would expect to see significant covariation, but there is not. Having negative pairs provides a novel constraint for RNA structure inference, and we recently described a method CaCoFold (Cascade variation/covariation Constrained Folding algorithm), that uses both positive and negative pair information to infer conserved structures. CaCoFold can propose any folding topology provided that it shows significant
covariation. The amount of support from (significantly covarying) positive pairs provides a direct measure of the reliability of the proposed structure [33].

Using R-scape (or conceptually related approaches) enables genome-wide screens for novel conserved RNA structures. In particular, it enables screens of long non-coding RNAs (lncRNAs) for conserved RNA structures. LncRNAs are mRNA-like transcripts of more than 200 nucleotides characterized by the sole property that they do not appear to encode a protein. In humans alone, between ten to hundred thousand lncRNAs have been reported [34, 35, 36]. These transcripts tend to be expressed at low levels [37], are rapidly degraded by the RNA exosome [38, 39], and have on average longer transcriptional burst periods relative to mRNAs, resulting in larger cell-to-cell transcriptional variability [40].

While only a handful of lncRNAs have been functionally characterized, it is often assumed that lncRNAs exert their function by means of RNA structure. While it is expected that any RNA may arrange in a compact folded state [26], it cannot be a priori assumed that any of these lncRNAs has any particular conserved structure that may serve a specific function.

Because of the large genome size and number of RNAs in an organism, especially vertebrates, an important issue affecting most previous methods’ ability to reliably detect structural RNAs is specificity, i.e. false positive rate [17, 18]. The tally of structural lncRNAs has been affected by many difficulties that result from combining a signal difficult to characterize [41] with the large size of vertebrate genomes and transcriptomes. Existing predictions of human structural RNAs are large and highly variable in numbers, with little overlap amongst different methods, and large numbers of estimated false positives [42, 43, 44, 45].

The general framework described in Figure 1 is an approach for interrogating transcripts or genomic regions and selecting the subset that are likely to have a conserved RNA structure. This method has controllable specificity as we showed in a recent analysis of novel ncRNAs in fungi, where we identified 21 new structural RNAs, 17 of which are novel, with an estimated specificity of 0.003 total false positives in a search over 134,000 alignments of conserved noncoding genomic regions, and 78% (62/80) sensitivity to detect known structural RNAs in S. cerevisiae [46].

Controversy still remains over lncRNAs and how to decide which have biologically relevant structures. In our experience so far using R-scape at genome scale, we have observed a variety of important sources of confounding covariation. Here, with our title inspired by a famous numerical analysis paper in the same spirit,¹ we aim to investigate the main practical issues in inferring the presence of an evolutionarily conserved RNA structure. Some issues have to do with the choice of covariation statistic and its compatibility with how the null hypothesis is chosen. Other issues have to do with the quality of the alignment from which the covariation signal is extracted. For each of these thirteen issues, we provide examples, a way to identify similar artifacts, and a way forward.

---

¹“Nineteen dubious ways to compute the exponential of a matrix” [47]
Conserved RNA Structure

thirteen issues

Figure 2: Thirteen issues to be aware of in the detection of conserved RNA structure. Given an RNA sequence of length L, such that the sequence is conserved forming an alignment of S homologs, several issues need to be considered in the analysis of whether the RNA has a conserved structure. Issue [1] refers to the need of using an alignment. Two issues [2, 3] have to do with correcting for covariation not due to RNA structure. Issue [4] addresses the assumptions of the null hypothesis. Two issues [5, 6] have to do with using a good covariation measure compatible with the null hypothesis. Issue [7] refers to other sources of covariation that are not phylogeny or RNA structure. Issue [8] refers to when a conserved RNA structure may not indicate a functional RNA, for instance with the nuclear mitochondrial RNAs (NUMTs). Finally, four issues [9, 10, 11, 12, 13] have to do with the quality of the alignment and the evolutionary divergence of the comparative data.

[1] Beware of single-sequence analyses

Single-sequence analyses cannot inform about conserved RNA structure. An alignment is needed.

Single-sequence analyses typically use a computational RNA folding method [48, 49, 50, 51, 52, 53, 54, 55, 56, 57] constrained by chemical probing data obtained from high-throughput experiments either in vitro or in vivo [23, 58]. The computational method explores the most stable structures based on the sequence alone. The chemical probing provides a measure of base reactivities informative of the folded state of the RNA molecule. The combination of those two sources of information result in predicted structures for one single RNA sequence. By sampling from the distribution of possible structures, these single-sequence methods can report one or more as possible alternative structural conformations of the RNA molecule [59, 24, 25].

Regardless of the quality of these single sequence structural predictions, these methods do not provide any information regarding whether the structures are conserved or biologically relevant. Since all RNA molecules are bound to adopt some folded conformation [26], one should expect predictions of stable RNA structures in essentially any sequence. However, a stable structure is not equivalent to a conserved structure. Only comparative data can infer whether a structure is
### Thirteen Issues to beware of

1. **Beware of single-sequence analyses**
   Single-sequence methods (experimental and computational combined) are modestly informative of possible structures, but cannot determine if the structures are conserved.

2. **Beware of phylogenetic covariation**
   Phylogenetic covariation is pervasive in all alignments and confounds structural covariation.

3. **Beware of what the null hypothesis destroys**
   Any property of the input alignment that the null alignments break (by some shuffling procedure, for example) can lead to spurious significance if the covariation measure depends on it.

4. **Beware of what the covariation statistic actually measures**
   A statistic (such as RAF or RAFTS) that combines covariation with other signals cannot be used to assess the significance of the structure; the statistic can be driven by the other signals (such as sequence conservation).

5. **Beware of machine-learned covariation measures**
   They do not automatically provide assurance of structural covariation.

6. **Beware of what covariance models measure**
   Significance with respect to a covariance model does not guarantee covariation, good alignments and homology E-values can occur just due to sequence conservation and consistency with an RNA fold.

7. **Beware of covariation due to sources others than RNA structure or phylogeny**
   Rejecting the null hypothesis of phylogenetic covariation does not automatically validate the structural RNA hypothesis. Alignments of protein-coding exons have covariation within codon positions induced both by synonymous and non-synonymous substitutions.

8. **Beware of structural fragments resulting from horizontal transfer**
   A conserved RNA structure may not indicate a functional RNA, for instance the nuclear mitochondrial RNAs (NUMTs) derived from mitochondrial tRNAs or rRNA.

9. **Beware of pseudogenes**
   Pseudogenes can reduce the covariation in an alignment.

10. **Beware of spurious covariation induced by misalignments**
    Alignment methods that favor covariation over conservation can create spurious covariation.

11. **Beware of spurious covariation induced by non-homologous sequences**
    Structural alignments that include non-homologous sequences can create spurious covariation.

12. **Beware of diverse evolutionary rates in structural RNAs**
    It may be a challenge to identify sequenced genomes at the right evolutionary distance for the detection of some structural RNAs.

13. **Beware of structural alignments with inconsistent base pair power**
    When the power of a base pair is not consistent whether calculated by double substitutions or by the sum of single substitutions, that suggests evidence against a conserved base pair.

As an example, the SARS-CoV-2 virus genome includes several known RNA structures, such as the 3' and 5' UTRs, the s2m RNA, a frameshift structure, and the packing signal. Recently, other structural elements have been predicted by the hundreds [60], but only a small fraction seem to be evolutionarily conserved [61]. These evolutionarily conserved structures have a higher likelihood to be involved in some functional role.

**[2] Beware of phylogenetic covariation**

*Phylogenetic covariation needs to be taken into account because it creates a confounding covariation signal.*

It has long been established that a conserved RNA structure induces covariation in an alignment. Different mathematical measures of covariation exist, such as mutual information (MI) [63], G-test [64], and others [65, 66]. Using covariation information, the consensus structures of both
subunits of ribosomal RNA were deduced, a huge success for two of the largest and most conserved structural RNA molecules known [4, 67, 68]. However, the reciprocal is not true; covariation does not necessarily imply the presence of a conserved RNA structure. After all, most pairs of positions in an alignment have some amount of covariation. How much covariation is enough to decide there is evidence of a conserved base pair?

When we remember that the reason why the sequences are alignable is because they share a common ancestor then we find another source of covariation not due to RNA structure. Independent substitutions can occur in two different sites in an ancestral sequence on one branch of the phylogeny. Because all descendants will tend to inherit both substitutions, the contrast between sequences in the clade with the substitutions versus other sequences without them appears to be a pairwise covariation signal when compared to simple measures that assume that sequences are independent [29]. Phylogenetic covariation has to be taken into account in order to detect the evolutionary signal left behind by a conserved RNA structure.

We first introduced R-scape to empirically determine the distribution of covariation scores for null alignments such that covariation generated by the phylogeny (by independent substitutions, sometimes co-occurring in the same branch) are preserved while randomizing the site of those substitutions to destroy the sort of recurrent pairwise compensatory substitutions observed for structural covariation [31]. The null alignments are constructed using the phylogenetic tree inferred for the input alignment. Each null alignment includes the same total number of substitutions per branch as the input alignment. However for the null alignments, the sites of substitutions per branch are selected at random amongst all possible sites, thus destroying any existing site correlation in the input alignment but retaining the possible phylogenetic ones. For each input alignment, a collection of null alignments are synthesized and a distribution of expected covariation due to phylogeny is calculated. This distribution is used to estimate for each base pair in the input alignment an expected number (E-value) of just phylogenetically related pairs that could have similar covariation scores (Figure 3a).

[3] Beware of what the null hypothesis destroys

The properties of the original alignment that are randomized away in the null alignments can appear as significant if measured—but may not be structural covariation.

Here, we show an example of an alignment property that may interfere with structural RNA significance.

The evolutionary signal of residue substitutions used in covariation analysis is almost always accompanied by insertions and deletions. Conserved RNA structures can have variable number of base pairs and even different numbers of helices. Those events introduce gaps in alignments. Covariation measures such as mutual information (MI) do not take gaps into account. MI is defined as the ratio of the joint $P_{ij}^{ab}$ to marginal $P_i^a$ residue probabilities for any pair of positions $(i, j)$, with $a, b = \{A, C, G, U\}$,

$$\text{MI}(i, j) = \sum_{a,b} P_{ij}^{ab} \log \frac{P_{ij}^{ab}}{P_i^a P_j^b}.$$  

On the other hand, G-test, another measure of pairwise covariation, uses a similar expression but
measures observed counts instead,

\[ \text{G-test}(i, j) = 2 \sum_{a,b} \text{Obs}_{ij}^{ab} \log \frac{\text{Obs}_{ij}^{ab}}{\text{Obs}_i^{a} \text{Obs}_j^{b}}, \]

thus introducing an additional dependence on how many symbols are observed in the columns, which is a function on how gappy they are. That is, while two sets of pairs with the same residue probabilities have the same MI regardless of number of gaps, their G-test will be higher for the pair with the fewer gaps.

Thus when the process of generating null alignments does not respect gap structure, and the input alignment has a non-uniform gap structure (as in Figure 3b top) a covariation measure such as G-test can wrongly assign higher scores to pairs because of their higher occupancy, relative to randomized null alignments (Figure 3b bottom left). Null alignments that respect the input alignment gap structure (Figure 3b bottom right) will not fool covariation signals like G-test with extraneous signal from inhomogeneous gap distributions. On the other hand, MI which is agnostic to occupancy, performs similarly in both types of null alignments, and finds no significant covariation in either one of them, but MI is noisier in low occupancy columns, which is also undesirable.

R-scape uses G-test as its default covariation measure. Since version v2.0.0.g and later, R-scape creates null alignments that preserve the position of gaps in the input alignment.

[4] Beware of what the covariation statistic actually measures

A ‘covariation’ statistic that measures more than just covariation may detect significance when there is no covariation.

There are covariation statistics that measure a combination of covariation, conservation and compatibility with an RNA structure, such as the RNAalifold (RAF) measure [69] or RNAalifold with stacking (RAFS) [65]. For every pairwise sequence comparison, the RAF measure has a positive score (+2,+1) for a double and a half compensatory change respectively. In addition, it has a negative term that penalizes (with score -1 per sequence occurrence) base pairs that are non-Watson-Crick and also base pairs that include at least one gap. By penalizing gaps, they effectively favor conserved pairs. By penalizing non-Watson-Crick base pairs, the statistic favors consistency with an RNA structure.

RAF(S) measures have been used successfully for making RNA alignments, as in the method RNAalifold [16], where conservation and consistency are both useful information, since the method assumes that the sequences to be aligned are both homologs and structural.

However RAF(S) measures create artifactual ‘significant covariation’ because of how they interact with the method to estimate null alignments to assess a conserved RNA structure. The random-site sampling method used to place the substitutions of a given evolutionary history does not preserve position-specific sequence conservation in the original alignment. In Figure 3a, we can see that while the original alignment has four completely conserved positions, only two of those have survived in the left null alignment, and none in the right null alignment. Thus, because a RAF(S) statistic rewards a pair of highly conserved positions compatible with a base pair (even with no compensatory substitutions), and R-scape’s null alignments do not preserve position-specific conservation patterns, any statistic that mixes covariation with conservation (such as RAFTS [65]) can wrongly assign pairs with zero covariation and no compensatory substitutions at all as “significant” [70].
Figure 4 shows two published examples of this artifact [71, 72]. For the lncRNA HOT AIR domain 1, [71] proposes a structure with several base pairs deemed to covary according to R-scape using the RAfS measure (a variant of RAfS that includes an average product correction [66]). A close analysis of a five base pair helix where three of those pairs are deemed significant shows that in all three cases, the right hand side of the base pair consists of a completely conserved column, thus lacking any covariation (Figure 4a). A standard analysis with R-scape default covariation measure results in no significant covariation for any pair in the whole HOT AIR D1 alignment. On the other hand, the alignment has sufficient power to expect 21 covarying base pairs, overall suggesting evidence against the presence of a conserved RNA structure.

The same RAfS analysis [71] has also been used on a proposed structure for the lncRNA MEG3 [72]. Figure 4b shows MEG3 putative helix H11 for which 7 potentially covarying pairs are presented, two of them claimed significant. A closer analysis to the alignment shows that 4 of the 7 pairs are completely conserved (including one of the two significantly covarying according to RAfS), and the other 3 have minimal mutual information. A standard analysis with R-scape also results in no significant covariation, but the alignment has sufficient power to expect 35 covarying base pairs.

Since version 1.4.0, R-scape disallows the use of any RAF-related covariation measure to assess statistical significance. We strongly discourage using earlier versions of R-scape in order to apply a RAF(S) measure to the statistical test.

[5] Beware of machine-learned covariation measures

Models that infer pairwise coupling terms describing the alignment do not automatically provide evidence for a conserved RNA structure.

Models from machine learning and statistical physics with pairwise coupling parameters have been applied to RNA alignments. Amongst those, direct coupling analyses (DCA) based on statistical Potts models [73] have attracted a lot of attention in recent years, both for the prediction of protein 3D structures [74, 75, 76, 77, 78], as well as for RNA structure [79, 80, 81]. The advantage of these methods for prediction accuracy of RNA base pairs seems to be limited to non-Watson-Crick pairs [82]. Importantly, DCA does not separate phylogenetic covariation, and thus does not reliably detect only RNA structural covariation. The R-scape statistical significance test could be applied to DCA covariation scores but the time needed to estimate the parameters makes the approach impractical.

Recent developments in 3D structure prediction using deep learning methods show that adding evolutionary information in the form of an alignment and covariation scores and incorporating both jointly into attention methods [83] is very good for improving 3D structure determination both for proteins [84] as well as for RNA [85]. However, the question of whether there is or is not a conserved RNA structure at all cannot be addressed directly by those methods. Methods for 3D RNA structure prediction are tested on RNA-Puzzles competitions [86, 87], where it is assumed to begin with that the mystery RNA is structural and conserved, and the emphasis is on the problem of getting the molecule’s spatial coordinates correctly determined.


Significance with respect to an RNA covariance model does not guarantee covariation.
Table 1: Evolutionary analysis of the 40,078 conserved RNA structures from Seemann et al., (2022). This subset was selected by Seemann et al. because their sequence is conserved in at least three species (human, rhesus macaque and mouse) with an average sequence length ≥ 80 nts and the human sequence has ≥ 100 nts, and because their 100-way vertebrate UCSC Genome Browser alignment [88] has a good CMfinder score [19]. We provided results for three different types of alignments: the original alignments, and two realignments of the sequences to an Infernal model created for the human sequence using either the program cmalign or cmsearch with E-value 1×10⁻³ which sometimes does not include some of the sequences in the alignment [14]. The no Rfam hit candidates are obtained after removing all hits to Rfam using the Infernal program cmscan. The NUMTs candidates are identified as having homology to the human mitochondrial genome using nhmmer. The final set of five (cmf.M0062940, cmf.M0914129, cmf.M0950988, cmf.M2247904, cmf.M2299369) results after removing one candidate (cmf.M2304242) that appears to be a protein-coding exon. Definitions: CANNOT TELL = at least one of three alignments: (expected cov ≤ 5) / no cov. NEGATIVES = at least one of three alignments: (expected cov ≥ 8) / no cov. STRUCTURAL = three alignments have at least three covariations. Note that the sum of the three categories CANNOT TELL, NEGATIVES and STRUCTURAL does not need to add up to the total.

An RNA covariance model (CM) is a probabilistic model (a stochastic context free grammar) that describes a particular conserved RNA structure. The method Infernal [14], given a structural alignment, builds a CM model that represents the structure and all its expected sequence and structural variability. Infernal CM models can be used to search sequence databases for structural homologs and to align them. CMs measure a mix of covariation, conservation and compatibility with an RNA structure, which are all informative properties for the purpose of homology search and alignment. However, much as the issue with the RAF(S) statistics, a CM score can create artifactual ‘significant covariation’.

Several publications have used high CM scores as evidence that a conserved RNA structure exists [45, 89, 90]. As an example, we concentrate on a filtered subset of 40,078 RNA selected from a screen that predicted over 516,000 conserved vertebrate RNAs structures predicted using CMfinder [45] believed to be a trusted subset of de novo conserved RNA structures [90].

We have investigated the subset of 40,078 CMfinder-derived candidates for the presence of a conserved RNA structure. First, because the candidates were selected based on the human-centered 100-way vertebrate UCSC Genome Browser alignment [45] which is not aware of any structural constraints, we tried to enhance the possible structural signal by making realignments producing Infernal models and creating two additional alignments per candidate, one using the Infernal tool cmalign which simply realigns all the sequences according to the CM, and another one using the Infernal tool cmsearch which only aligns sequences below an E-value threshold to assess homology. Our investigations are based on the combined information provided by the three types of alignments.

A reanalysis using R-scape of the subset of 40,078 CMfinder-derived candidates indicates that only a small fraction of the candidates (2.5%, 1,021/40,078) have evidence of a conserved RNA
structure (Table 1). Why do many candidates have good scores, but so few of them have evidence of a conserved RNA structure according to the R-scape approach?

A critical component in a CM is the base pair emission probabilities that assign a probability to the 16 possible pairwise nucleotides [12]. Each base pair in a given CM gets assigned a different pair emission probability distribution. For instance, the first base pair of the Rfam CM model for the SECIS_1 element (RF00031) is a G:C consensus pair with the following probability distribution (calculated before adding pseudocounts, which are used to avoid the case of zero probabilities),

\[
P_{\text{RF00031 first bp (G:C)}} = \begin{pmatrix}
A & 0.000 & 0.000 & 0.000 & 0.047 \\
C & 0.000 & 0.035 & 0.201 & 0.000 \\
G & 0.000 & 0.348 & 0.000 & 0.123 \\
U & 0.231 & 0.015 & 0.000 & 0.000 \\
\end{pmatrix}, \text{ such that } \sum_{a,b=1}^{4} P_{\text{RF00031 first bp (G:C)}}(a,b) = 1.
\]

For this consensus G:C pair, while compensatory double-substitutions such as C:G and A:U (in green) and half-compensatory G:U (in blue) have high probability, still the highest probability correspond to the actual consensus G:C pair (bold). Thus, the CM is modeling both base pair conservation and well as covariation associated to RNA secondary structure constraints.

Thus, a CM pair emission probability distribution reflects both the conservation and the covariation observed for that base pair, which is exactly what you want for the purpose of homology detection, but a confounded measure for detecting structural covariation. A CM will assign high scores simply due to sequence conservation and not to structure conservation.

Indeed, going back to our example, we observe that for more than half of the candidates (53%, 21,204/40,078) the sequences are too conserved to be able to assess whether there is an RNA conserved structure or not (no covariation and no power).

Power is defined as the expected covariation that we should observed if there were a conserved structure, given the amount of variability we observe in the columns. We calculate power by estimating for a collection of structural RNAs the observed covariation as a function of the number of substitutions in the base pairs. No covariation in the absence of power indicates that the alignment has no information to decide on whether a structure is conserved or not. No covariation in the presence of power shows evidence against a conserved RNA structure.

For the collection of 40,078 candidates, almost a third of the candidates (29%, 11,534/40,078) do not have covariation but they have sufficient power (Table 1). We leave a follow up discussion of the 1,021 candidates with covariation support for Issue 8.

[7] Beware of covariation due to sources others than RNA structure or phylogeny

Rejecting a null hypothesis (phylogenetic covariation) does not mean that your favorite hypothesis (RNA structural covariation) is true.

We have encountered an unexpected source of pairwise covariation that arises neither from phylogeny nor from RNA structure. In alignments of protein-coding exons, we detect significant covariation between the positions in a given codon (within-codon covariation). Covariation between adjacent positions cannot be due to Watson-Crick RNA base pairs as those are required to be separated by at least three unpaired nucleotides. (Incidentally, there are non-Watson-Crick adjacent pairs, but those are few and do not show much covariation.)
Using alignments for all intronless protein-coding genes in S. cerevisiae, we observe that overall 38.5% of all the significant covariations are within-codon covariation. That is in contrast with RNA structural covariations which are predominantly distant (95% of covariation in S. cerevisiae structural ncRNAs are more than three nucleotides apart) [91]. Other studies have also reported coding covariations [92]. Significant covariation between adjacent or next to adjacent positions is a signal indicative of protein-coding exons, not yet exploited, that could be incorporated into methods to detect protein coding exons from alignments.

Within-codon covariation is the result of both synonymous and non-synonymous codon substitutions.

Here we show that the within-codon covariation signal in protein-coding exons arises from constraints of the genetic code resulting from a combination of codon preferences and amino acid evolution.

Figure 5 describes a simple model of protein-codon evolution in which the probability of observing a codon c1c2c3 at time t resulting from an ancestral amino acid a can be computed by combining a standard amino acid substitution matrix with standard codon biases as,

\[
P(c_1c_2c_3 \mid a, t) = \sum_{b=1}^{20} P(c_1c_2c_3 \mid b) P(b \mid a, t).
\]

In this evolutionary model, the first term describing codon biases accounts for one source of variation. Biologically, we use them to represent synonymous substitutions, which are substitutions that do not alter the amino acid. Most synonymous substitutions only involve variation at the third codon position (referred to as wobble) as in the case of cysteine, which can be encoded with either U or C at the third position so long as the first and second positions are fixed as U and G. However, the three 6-box amino acids (leucine, arginine, and serine) can undergo synonymous mutations that also alter the first and/or second codon positions.

The second term describes amino acid substitutions, the second mechanism of protein evolution in our model. In contrast to synonymous substitutions, nonsynonymous substitutions can greatly change the codon’s nucleotide composition as it is no longer restrained to encode the same amino acid. This allows it to theoretically become any of the other 63 codons. Biologically, however, it has been observed that some amino acid substitutions are much more common than others. This is because changing the amino acid may result in deleterious changes to protein structure and function. Thus, it has been observed that amino acid substitutions tend to favor changes that maintain similar chemical properties such as polarity, charge, and shape. To describe the likelihood of an amino acid either staying the same (synonymous substitution) or mutating into another amino acid (nonsynonymous substitution), matrices have been computed using real protein alignments. These matrices can be adjusted with respect to a parameter t that represents a unit of evolutionary time.

While our model involves both synonymous and nonsynonymous substitutions, this simplified model ignores insertions and deletions. Nor does it consider nonsense mutations, which would change a sense codon to a stop codon probably worth including, but the amino acid substitution models like BLOSUM do not consider sense-to-stop substitutions (or vice versa), which is why our model does not either. It also assumes that all sequences in the alignment are independently derived from the ancestral sequence (a star topology).
Figure 6a describes the amount of within-codon covariation (c1-c2, c1-c3 or c2-c3, where c1, c2, c3 are the codon positions) observed for the different encoded amino acids. We observe that a small set of amino acids drives each of the three different cases.

Using the evolutionary model in Figure 5, we can calculate the expected within-codon covariation as the mutual information of the corresponding marginal probabilities, and we can compare those to the within-codon covariation observed in the yeast exon alignments. We present those correlations in Figure 6b where the expected covariation have been calculated at \( t = 0 \) which corresponds to the average pairwise identity (58%) found in the yeast intronless aligneds. We observe that there is a good correlation between observed and expected covariation given the evolutionary model of Figure 5.

Figure 6c describes more generally the predicted evolution of within-codon covariation as a function of evolutionary divergence given the model. We observe that for short evolutionary divergence, within-codon covariation is dominated by the three amino acids coded by 6 different codons. As those are the only amino acids that support synonymous changes in all codon positions, thus they are the only cases that in the absence of non-synonymous substitutions can result in within-codon covariation. As divergence increases, non-synonymous substitutions become more prevalent and within-codon covariations are spread more regularly amongst the different amino acids. In particular, amino acids that are restricted in genetic code usage (methionine, tryptophan) or have strong evolutionary constraints per the BLOSUM matrix (cysteine, tryptophan) have the highest within-codon covariation at moderate evolutionary divergences.

\textit{mRNA-induced covariation can coexist with RNA structural covariation.}

That is the case of the transfer-messenger RNA (tmRNA), a bacterial structural RNA that also includes a protein-coding sequence. When a mRNA lacking a stop codon gets stalled at the ribosome during translation, tmRNAs provide a mRNA template ending with a stop codon that frees the ribosome from the defective mRNA. The mRNA template sequence is also part of the structure making half of a RNA base paired helix. In alignments of tmRNAs from different bacteria, R-scape can find non-phylogenetic covariation that can be easily associated either to the hairpin elements or to coding structure [33].

\textbf{[8] Beware of structural fragments}

\textit{Sometimes a structural RNA may not be functional.}

One such case are the nuclear mitochondrial DNA sequences (NUMTs) which are inserted in nuclear genomes in multiple copies include mitochondrial rRNA and tRNA fragments and show evolutionary conservation in the nuclear genome.

Within the 40,078 proposed conserved human RNA structures of [90], we identify a subset of 1,021 with covariation support (at least 3 covarying base pairs in all three alignments). Further analysis indicates that only 71 of them do not show any similarity to an already known Rfam family. Those similarities are mostly fragments of rRNA (137 LSU, 117 SSU), tRNA(371), snoRNAs (65) and miRNAs (82) (Table 1).

The remaining 71 candidates would then appear to have high potential of being novel conserved structural RNAs. But further analysis indicates that many of them are homologous to fragments of mitochondrial tRNAs and rRNAs. These inserts named NUMTs do not seem to have any specific function. NUMTs are well known and present in all vertebrate species.

As an aside, we finally identify five of the 40,078 proposed structures with enough potential

[9] Beware of pseudogenes

*Pseudogenes increase apparent power but dilute covariation signal.*

There is an astounding proliferation of pseudogenes derived from ncRNAs [94]. Those are sequences related to a given ncRNA that show evidence of not performing the ncRNA function, either because of their fragmented or repetitive nature. Examples of human ncRNA pseudogenes of both kinds are the hundreds of 5S rRNA related sequences [95] or the SRP-derived Alu interspersed repeated elements, respectively.

Pseudogenes derived from ncRNAs retain both sequence and structural similarity, and they are often difficult to distinguish from the functional ncRNA. Indeed many pseudogenes show up in homology searches both structural or just sequence based. Because substitution in pseudogenes are not structurally constrained, they can appear at every position, and they will slowly degrade any evolutionary constraint of the functional ncRNA. Thus, alignments of structural ncRNAs that include pseudogenes will tend to show an overall increased power of covariation together with a decrease of covariation at base paired positions. We propose that in genomic searches for novel structural RNAs that result in many hits per genome, only should only retain the top scoring sequence per genome because that sequence is more likely to be the true ortholog rather than a confounding pseudogene.

As an example, we used the Rfam model for 5S rRNA (family RF00001) to search a database of invertebrate whole genomes from NCBI. We identified 97,352 homolog sequences (with E-value < $1 \times 10^{-10}$) in 386 genomes. *D. melanogaster* has 96 5S rRNA related sequences, and *Andricus curvator* (a gall wasp) the species with most homologs includes over 1,200 homologs. A comparison of the evolutionary support for the 5S rRNA structure alignments with and without pseudogenes is given in Figure 7. An alignment including only the best scoring sequence for each genome, likely to include true orthologs, results in 25 covarying base pairs (for 27 pairs expected to covary), out of the 38 base pairs in the 5S rRNA structure. However, an alignment with the same number of sequences but sampled at random from the whole alignment, likely including pseudogenes, reduces the number of covarying pairs to 15, while the expected number of covarying pairs rises to 32.

[10] Beware of spurious covariation induced by misalignments

*Misalignments of RNA helices with variable number of base pairs can induce spurious covariation when both base paired residues for one sequence are shifted simultaneously to wrong base paired positions within the helix.*

We have reported elsewhere [82] how misalignments can result in spurious covariation. Those result because base pair stems can have different number of base pairs in different species (Figure 8a).

An example is shown in Figure 8b. An alignment of the lncRNA COOLAIR appears to have covariation [96], but an automatic realignment shows a more conserved arrangement that increases sequence conservation while preserving all the proposed Watson-Crick base pairs, thus providing no evolutionary information as to whether there is an evolutionarily conserved structure for COOLAIR or not.

Paraphrasing Karlin & Altschul [97], high scoring structural alignments of non-homologous sequences will be pressured to resemble the proposed structure which can force spurious covariation.

The detection of homologs either by sequence alone (BLAST, nhmmer) or sequence and structure (Infernal) relies on the characterization of the distribution of scores of non-homologous sequences. High scoring non-homologous sequences will necessarily resemble the motif under consideration. In particular, for structural homology the inclusion of non-homologs will tend to force the alignment to introduce substitutions to fit the consensus structure and these will result in apparent covariation. This effect will be exacerbated when the sequence of interest has low complexity (such as including large runs of a repeated nucleotide, as a homopolymer of A nucleotides will appear to base pair with a homopolymer of Us).

Thus, when analyzing the evolutionary signal in structural alignments, it is important to avoid the inclusion of non homologous sequences. And it is safer to use sequence alignment alone, as opposed to risking the circularity inherent in performing an alignment that maximizes structural similarity, and then evaluating the support for that structure.

As an example, Figure 9 shows the 3′-UTR of the human CDKN1B gene for the p27 cyclin-dependent kinase inhibitor. The CDKN1B-3′-UTR is conserved in vertebrates. A structural alignment constructed using Infernal and using a proposed structure for the motif is presented in Figure 9a. This vertebrate alignment is likely to include real homologous sequences, but it does not provide much convincing evidence for the proposed structure to be conserved. Figure 9b, which uses a larger E-value cutoff (E < 1 versus E < 1 × 10^{-10}), shows an invertebrate structural alignment with similar number of sequences, all non-homologous sequences. This invertebrate alignment has very low % average pairwise sequence identity, and all proposed base pairs appear to covary! Figure 9c shows the highest scoring invertebrate sequence in this alignment (E-value 2 × 10^{-16}), and how by visual inspection it is obviously a non-homolog, because it is a repetitive sequence that even looks like a sequencing artifact (ordered runs of A, C, G, and Us). The alignment, by being forced to conform to the proposed structure, introduces a striking number of spurious covariations.

[12] Beware of diverse evolutionary rates in structural RNAs

Structural RNAs have a wide range of different evolutionary rates. The covariation evidence for evolutionary conservation of their consensus structure requires us to use comparative data at the optimal evolutionary distance.

Different known structural RNAs have many different evolutionary rates, and in order to identify them, we need to find alignments with enough sequence divergence to capture the covariation signal. This effect is exacerbated when searching for structural RNAs in vertebrates due to the much shorter evolutionary distances between the existing sequenced vertebrate genomes in comparison with those existing between bacterial or archaea genomes. For slowly evolving structural RNAs such as tRNAs or rRNAs, alignments restricted to vertebrates usually have too little variability to detect any signal of evolutionary conservation. On the other hand, structures such as the iron response elements (IRES) [98] or the selenocysteine insertion sequences (SECIS) [99] both found in vertebrate UTRs are short (less than 70 nts) and can go undetected if the homology search is not well tuned to their evolutionary divergence.

We have performed screens in the yeast S. cerevisiae and the pufferfish T. rubripes, in which regions in the genome of interest are compared using nhmmer to a database of whole genome
sequences, and the resulting alignment of homologs is evaluated for structural covariation using R-scape. Figure 10 shows controls for such screens using known structural RNAs either in yeast or pufferfish, using the same screen details [46], and different comparative data. We observe that while a large fraction (85%) of the fungal structural RNAs when compared to a database of fungi genomes would be detected (as having at least 3 significantly covarying base pairs), only a small fraction (8%) of the pufferfish structural RNAs would be detected searching in a database of vertebrate genomes. We also observe how a comparison to a database of invertebrate and fungi genomes increases performance in identifying known structural RNAs (24%). The reason for this is that most of the tested structural RNAs are slowly evolving and they have very high sequence similarity when restricted to vertebrate genomes alone.

As a fungi example, the yeast RPS2 5′-UTR structure that we re-identified in the fungi screen [100, 46] with only 40-50 nts, barely produced alignments in the first and second iteration (E < 1 × 10^{-10}), and only in the 3rd iteration (E < 1 × 10^{-5}) there was enough variation to detect covariation. And additional screens starting from four different fungi species (C. albicans, N. crassa, A. fumigatus, and S. pombe) expanding three different Ascomycota subphyla resulted in the discovery of different structural RNAs which were not found in the other subphyla [46].

As vertebrate examples, the IREs and SECIS structural RNAs both present in the pufferfish positive control did not report any homologs in an nhmmer search performed with E-values < 1 × 10^{-10}. In addition, using the UTR of the human GPX1 gene that includes a SECIS structure, an nhmmer search against a vertebrate database tuned to the evolutionary divergence of the structural RNaseP RNA, and using larger E-values < 1 results in an alignment of 156 vertebrate sequences with 61% identity, but it does not report any covarying base pairs (alignment is provided in the supplemental material).

Different sets of comparative genomes need to be used to optimally detect structural RNAs at a varied range of evolutionary rates.

[13] Beware of when the joint power of a base pair is not consistent with the sum of each position’s power

Inconsistent single versus double substitution power reflects lack of base pairing evidence.

In the calculation of statistical power, the default is to count the number of substitutions observed in both individual columns and sum them [32]. This default calculation assumes that the variation is shared more or less equally between the two columns. Alternative, we can calculate the number of substitutions that occur in both positions simultaneously. Alignments that do not share the variation between the two columns, or alignments that do not even share the presence of both residues in both position will lead to disparate results between these two different ways of estimating power. We show an example in Figure 11c.

Four alignment patterns, implications for inferring RNA structure.

We have discussed two computational measures for extracting the evolutionary information contained in an alignment with respect to the question of how to identify conserved RNA structure. One measure is the pairwise co-dependence above what would be expected if there were no conserved structure (significant covariation). We say that a base pair significantly covaries if it has an R-scape E-value < 0.05, by default. The other measure is how much significant covariation to expect if there were a conserved structure given the sequence variation observed in the alignment (power). The power of a base pair is the probability that the base pair is expected to be detected
as significantly covarying based only on the variability that is present in the alignment.

Figure 11 describes four possible scenarios based on significant covariation and power.

Significant covariation (covariation for short) necessarily should grow linearly with power (almost in a 1:1 ratio), but power can exist in the absence of covariation. Covariation indicates the presence of a conserved RNA structure (Figure 11a). But in the absence of covariation, three different scenarios can be expected based on power.

The variation observed for a pair of positions can be measured either by counting substitutions per position (single-subs power) or by measuring double substitutions (double-subs power). Double-subs power implies single-subs power, but the reciprocal is not true. Lack of covariation in the presence of double-subs power indicates evidence against an evolutionarily conserved structure (Figure 11b). Moreover, lack of covariation in the presence of single but not double subs power tends to indicate an alignment that does not support the proposed structure (Figure 11c). Finally lack of covariation in the absence of any power implies that the alignment cannot provided any information regarding the presence or not of a conserved RNA structure (Figure 11d).

**Future directions**

As in the field of matrix exponentials, we expect that twenty five years from now the field of computational methods for detecting structural RNAs will mature in important directions. Hopefully much earlier than in twenty five years, we will have a complete list of the structurally functional RNAs currently hiding in the extensive vertebrate transcriptional landscape. To achieve this goal, we need to develop more sensitive methods that without sacrificing specificity can identify the signals beyond covariation associated with conserved RNA structure. Importantly, the main source of artifacts is the quality of the alignment. Can we ever be sure that an alignment gives us the correct information as to whether the RNA has a conserved structure? More accurate models to perform structural alignments promise to be an important source of progress in the field.

Often thousands of lncRNAs are assumed to be fundamental regulators based on observations in a few specific ones. It seems prudent not to treat lncRNAs as a homogeneous class, and it seems fruitful to use stringent evidence to discern the wide variety of different things that different lncRNAs may be involved in. Those include cases where the act of transcription (rather than the RNA) seems to be the effector, cases where an RNA may work by binding proteins or other RNAs, cases of undetected protein-coding mRNAs, and cases of accidental and nonfunctional transcripts, none of which needs to involve conserved RNA structure. In addition, we expect a possibly small subset of lncRNAs to exert specific functions through conserved RNA structure. R-scape is a powerful statistical test to discern structural lncRNAs from other lncRNAs in exploratory comparative data analyses. It is our perspective that identifying vertebrate lncRNAs with genuinely conserved RNA structure would lead to the discovery of novel functions of RNA.

**Data availability**

All alignments and data created for this article are provided as part of the supplemental materials, except for the 4,826 intronless gene alignments produced for Figure 6a. Instead, we provide a one gene example with a tutorial on how to run the scripts. For the 40,078 Seemann et al. 2022

---

2 “Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later” [103]
proposed structures, we only provide the data for the subset of 1,021 proposed candidates with covariation support.

Acknowledgments

We would like to thank Sean R. Eddy for suggestions on the article, and Thomas A. Jones for never ending information about the sequences of ncRNAs. We thank David Haig for suggesting the CDKN1B 3'UTR example to us.

References


Figure 3: Accounting for phylogenetic covariation. (a) We show the process of generating alignments under the null hypothesis of covariation due solely to phylogeny using the toy alignment introduced in Figure 1. For a given input alignment (in blue, top left), we build a phylogenetic tree, for which different sets of internal node sequences can be sampled using Fitch’s algorithm [62]. A particular sample of the three internal nodes for this toy alignment is given in blue (top right). At each node, we show the sampled sequence (in black) together with its ancestral sequence above (in gray), so that we can highlight the substitutions between the two (in brown). All null synthetic alignments follow the same phylogeny, but each null alignment is generated from a different sample of the input alignment’s internal nodes. An example null alignment (in yellow, bottom left) is generated introducing the same substitutions per branch than the given alignment, but the substitutions are applied at random positions in the corresponding null sequences (yellow, bottom right). As an example, we follow the base pair between positions 4:9 (one of the five significant base pairs indicated left with green lines). This 4:9 base pair includes five base-pair-preserving substitutions occurring at four branches. The same five substitutions appear at the same four branches in the null alignment, but at positions different from 4:9 and selected at random from other available positions. Both alignments have similar average percentage identity (41%), and the null sequences have the same base composition than their equivalent ones in the input alignment. Per-column conservation however is not preserved. (b) An example alignment with a non-uniform gap distribution that results in biased occupancy, with the right side of the alignment having higher occupancy than its left side. A null alignment sampling method that does not preserve the gap distribution of the input alignment (bottom left) may score as significant residues in the right side of the alignment when we use a covariation measure such as G-test (R-scape’s default) that is sensitive to occupancy. The problem is resolved by using a null alignment sampling algorithm that preserves the alignment occupancy (bottom right).
Figure 4: Tavares’ analysis [71] calls pairs of noncovarying conserved primary sequence positions “significantly covarying”. (a) Tavares et al. 37 sequence alignment of HOTAIR Domain 1 putative helix H11 where they show a 5 base pair helix and claim “covariation support” for three base pair (red arcs). The ACG in right hand side of the pairs is completely invariant, thus by definition there is no covariation. (Example from [70], as observed in Figure 5b of [71].) (b) Uroda et al. 42 sequence alignment of MEG3 putative helix H11 showing a nine base helix in which they claim covariation support for two base pairs (red arcs) and potential covariation support for five more (black arcs) using the same Tavares’ analysis. Four of the seven base pairs (GCCG-GGGG) show no covariation and no compensatory base pair substitutions, including one pair (C-G) claimed significant according to RAFS. The other pair claimed “significant” includes three double substitutions, but also four species where the base pair is not preserved, and also minimal Shannon mutual information (MI). Results obtained by running command line R-scape -s --RAFSp (from R-scape v0.8.2) on the provided alignments. Residues in red disrupt the proposed Watson-Crick pair, residues in blue show a half compensatory substitution, and residues in green show a double compensatory substitution. Shaded columns are 100% sequence conserved. Coordinates are relative to the given alignments (also provided with this Supplemental Materials).
Figure 5: Model of protein-codon evolution. Our model of protein evolution starts with an ancestral amino acid. The probability for amino acid $a$ being substituted by amino acid $b$ is calculated using a BLOSUM matrix at a specified value of evolutionary time $t$. In the case of nonsynonymous substitutions, $a \neq b$. In the case of synonymous substitutions, $a = b$. Finally, the codon is determined based on amino acid $b$'s codon biases. It is reasonable to assume that synonymous substitutions (represented by codon biases per amino acid) reach equilibrium quickly, while nonsynonymous substitutions keep evolving for much longer times. Thus, the model makes the approximation of using stationary (no time dependence) codon biases probability distributions.
**Figure 6: Within-codon protein-coding exon covariation.** For the 4,826 annotated intronless protein-coding genes in *S. cerevisiae*, we generated alignments using nhmmer [93] (E-value < 1.0 × 10^{-10}) searches of a database of 1,371 genomes of the Ascomycota fungal phylum [46]. (a) The observed fraction of codons for which the c1-c2 (or c1-c3 or c2-c3) codon positions significantly covary (R-scape E-value < 0.05) conditioned on the corresponding amino acid (as represented by the *S. cerevisiae* aligned codon). (b) Comparison of the observed within-codon covariation per amino acid to the expected covariation (measured by MI) given the model of evolution. For amino acid substitutions, we use a rate matrix derived from BLOSUM62 normalized to one substitution per site. We use the evolutionary divergence of \( t = 0.2 \) which produces on average alignments with average percentage identity of 58% similar to that observed on average for the intronless protein-coding alignments. Both the observed frequency of within-codon covariation and the expected mutual information are scaled to a total sum of one, just for better visualization. (c) Description of the expected (unscaled) MI for within-codon covariation as a function of divergence time. Trajectories that monotonically decrease with time correspond to the three amino acids (serine, arginine and leucine) decoded by 6 different codons. Arginine (CGN,AGR) and leucine (UUR,GUN) allow synonymous changes between the c1 and c3 codon positions, and the c2 position is fixed (middle panel), and serine (UCN,AGY) allows all three types (c1-c2, c1-c3 and c2-c3) of synonymous changes.
Figure 7: **Pseudogenes increase power but reduce covariation.** Invertebrate 5S rRNA sequences were selected and aligned using Infernal (cmsearch) and the 5S rRNA Rfam model (RF00001) against the NCBI database of invertebrate genomes. We identified 97,352 5S rRNA homologs with E-value < $1 \times 10^{-10}$ from 386 different invertebrate genomes. (a) Sub-alignment of 386 sequences using only the best hit per genome. (b) Sub-alignment of 386 5S rRNA related sequences selected at random amongst all significant hits. Evolutionary analysis of observed and expected covarying pairs performed using R-scape with option --s that evaluates how well a proposed consensus structure is supported by covariation.

Figure 8: **Base pair misalignments induce spurious covariation.** (a) Misalignment mechanism that results in spurious covariation at the expense of reducing sequence conservation while preserving the same structure. (b) Example of the artifact from a proposed structure for the lncRNA COOLAIR [96]. Sequences structurally realigned using Infernal show a completely conserved putative helix with no covariation support. Realigned groups of residues are equally colored in both alignments.
Figure 9: Structural alignments of non-homologous sequences can introduce spurious covariance. An Infernal model of the human CDKN1B-3' UTR is used to search a database of (a) vertebrates and (b) invertebrate genomes. We show the evolutionary analysis of the two alignments. (c) Alignment to the Infernal model of the best scoring invertebrate hit. The proposed structure is given in WUSS format (see Infernal documentation for a description of the WUSS structural notation).
Figure 10: Evolutionary signal for known conserved structural RNAs in eukaryotes; comparison between fungi and vertebrates. We show control screens in which a set of known structural RNAs in some genome (S. cerevisiae and T. rubripes) are used to search a genomic database with a homology method (nhmmer) to find homologs, build an alignment and evaluate the potential for a conserved RNA structure with R-scape. (a) Control screen of 80 known structural ncRNAs in the budding yeast S. cerevisiae using a database of 1371 complete fungal genomes [46]. Each RNA is represented by the observed number of significantly covarying base pairs versus the expected number of covarying pairs. We find that 69/80 RNAs have at least 3 covarying base pairs. (b) Control screen of 103 known structural ncRNAs in pufferfish (T. rubripes) compared to a database of 183 vertebrate genomes which yielded alignments for 88/103 RNAs. Only eight of the 88 RNAs have at least three covarying pairs. (c) Same 103 RNAs compared to a combined database of 969 invertebrate and 1371 fungal genomes which also yielded 88 alignments. 35/88 RNAs have at least 3 covarying base pairs. The vertebrate genomes were selected by clustering the collection of NCBI vertebrate genomes by the similarity of their RNaseP RNA, and taking a representative from each cluster (genomes in a cluster have at least 85% similarity to each other’s RNaseP RNAs) [101]. In all cases, we constructed multiple sequence alignments through three rounds of homology search (using nhmmer) at E-value cutoffs of $1 \times 10^{-4}$, $1 \times 10^{-10}$, and $1 \times 10^{-5}$, respectively.
Figure 11: Four different conservation patterns and their implications for inferring conserved RNA structure. (a) Detail of the vertebrate telomerase RNA Rfam seed alignment (RF00024). (b) From HOTAIR domain1, putative helix 3 from the alignment provided by Somarowthu et al. [102]. (c) From the same HOTAIR domain1 alignment, putative helix 6. (d) HOTAIR domain1, putative helix 11. R-scape report power by single substitutions by default. Option --doublesubs report power by double substitutions.