Introduction

This document contains definitions for a number of statistics used by Rfam for the presentation or RNA secondary structure and quality control. They are chiefly used to identify problematic families and basepairs by the curators of the database. These statistics are generally unpublished and are not optimised for tasks such as consensus structure prediction. So if you are tempted to write a paper about how useful or otherwise these are then we suggest you reconsider.

A tab-delimited file called **Rfam.qc** that summarizes many of these statistics should be available from the Rfam ftp site for the latest Rfam releases.

Secondary structure representations

Basic definitions: Consider a sequence 'a' annotated with a secondary structure S_{ij}, consisting of a set of doublets/basepairs i and j. Π^a_{ij} is as basepair indicator function; If a_i and a_j can form a canonical basepair {AU, CG or GU} then $\Pi^a_{ij} = 1$, otherwise $\Pi^a_{ij} = 0$. M is the number of sequences in the alignment.

W^a is a sequence weighting scheme for down- or up-weighting either an over- or under-represented sequence (a) respectively in an alignment. The sequence weights are computed using the program **weight** that ships with Sean Eddy's squid package. We use his implementation of the tree-based Gerstein, Sonnhammer and Chothia (GSC) weighting scheme[1].

Basepair conservation: This statistic was previously called "the fraction of canonical basepairs" (FCbp), which was a more accurate but less intuitive name. The basepair conservation for a given consensus pair is:

$$BPcons_{ij} := \frac{\sum_{a_{ij}, ij \in S_{ij}} (\Pi^a_{ij})}{M}$$

Covariation: If Ω^{ab}_{ij} is an indicator function such that if a_i and a_j and/or b_i and b_j cannot form a canonical basepair {AU, CG or GU} then $\Omega^{ab}_{ij} = 1$, otherwise $\Omega^{ab}_{ij} = 0$. $H(a_i a_j, b_i b_j)$ is the hamming distance between the two strings $a_i a_j$ and $b_i b_j$, this is the number of differences between the two strings (a.k.a. edit distance). E.g. $H(AU, AU) = 0$, $H(AU, GU) = 1$ & $H(AU, GC) = 2$. The per-basepair covariation statistic we use is defined as:

$$C_{ij} := \frac{\sum^{M}_{a=1,b=2,a<b,\ a_i:a_j \& b_i:b_j \notin \{-:-\},\ ij \in S_{ij}} \left(W^a + W^b\right) * \left(\Pi^{ab}_{ij} H(a_i a_j, b_i b_j) - \Omega^{ab}_{ij} H(a_i a_j, b_i b_j)\right)}{\tau_{ij}}$$

τ_{ij} is the total number of comparisons used to compute C_{ij}. In the absense of gap-gap cases $\tau_{ij} = \binom{M}{2}$. This measure is based upon the covariation measure used in RNAalifold [2].

Relative entropy: If $p_i(x)$ is the frequency of nucleotide x at position i in an alignment and $q(x)$ is the background frequency of nucleotide x (estimated from the total frequency of x in the alignment), then the relative entropy H_i is defined as:

$$H_i := \sum_{x \in \{A,C,G,U\}} p_i(x) \log_2 \frac{p_i(x)}{q(x)}$$

The **Sequence conservation** at a site in the alignment is defined as:

$$SC_i := \max_{x \in \{A,C,G,U\}} p_i(x)$$

Most informative sequence: any nucleotide that is over-represented relative to the background frequency at a certain site is projected into full IUPAC notation. Eg. if the frequencies of A and G at a site is 50% and background frequency for both is 25% then we use "R" to represent this site. Sites comprised of more than 50% gaps are in lowercase, otherwise uppercase is used. All the nucleotide IUPAC redundancy codes are given in table 1.

$$\text{I.e. If } \log_2 \frac{p_i(x)}{q(x)} > 0; x \rightarrow \{\text{IUPAC}\}_i$$

This can be re-written in a less complicated form as:

$$\text{I.e. If } p_i(x) > q(x); x \rightarrow \{\text{IUPAC}\}_i$$

1-letter codes	2-letter codes	3-letter codes	4-letter codes
A →A	A,G→R	C,G,U→B	A,C,G,U→N
C →C	U,C→Y	A,G,U→D	
G →G	A,C→M	A,C,U→H	
U,T→U	U,G→K	A,C,G→V	
	C,G→S		
	A,U→W		

Table 1: IUPAC nucleotide redundancy codes.

References

[1] M Gerstein, E L Sonnhammer, and C Chothia. Volume changes in protein evolution. *J Mol Biol*, 236(4):1067–78, Mar 1994.

[2] I L Hofacker, M Fekete, and P F Stadler. Secondary structure prediction for aligned rna sequences. *J Mol Biol*, 319(5):1059–1066, 2002.

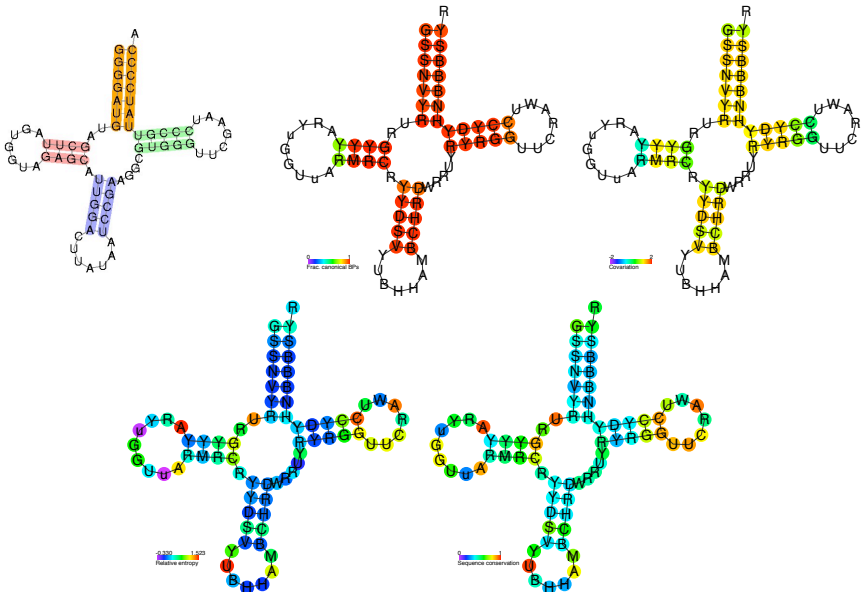

Figure 1: Example structures for RF00005 (tRNA). 1. Traditional Rfam graphic (these are good for comparison with HTML alignments). 2 & 3 Heatmaps: 2. shows the basepair conservation ($BPcons_{ij}$), 3. shows the covariation for each basepair (C_{ij}). 4. Positional entropy for each column in the alignment (H_i) 5. Sequence conservation for each column in the alignment (SC_i). The backbone is represented by the most informative sequence (MIS). The statistics and MIS are more formally defined below.