RILogo: Visualising RNA-RNA interactions

Peter Menzel, Stefan E. Seemann, and Jan Gorodkin

August 6, 2012

Definition of Mutual Information Measures

RILogo implements four mutual information like measures: MI and $\mathrm{MI}^{W\diamond P}$, two commonly used measures, as well as treeMI and $\mathrm{treeMI}^{W\diamond P}$, which extend the standard measures by a weighting scheme based on phylogenetic tree distances (new in *RILogo*).

The standard mutual information measure MI, based on the Kullback-Leibler divergence, for two base-paired alignment columns i and j is defined as

$$\mathrm{MI}_{ij} = O_{ij} \log_2 \frac{O_{ij}}{E_{ij}} + (1 - O_{ij}) \log_2 \frac{(1 - O_{ij})}{(1 - E_{ij})} \tag{1}$$

where O_{ij} is the frequency of observed base pairs and E_{ij} is the frequency of expected base pairs (Gorodkin *et al.* 1997). The observed and expected frequencies of canonical base pairs at columns i and j are generally counted as

$$O_{ij} = \frac{1}{N} \sum_{s \in S} C_s \theta(s_i, s_j); \quad E_{ij} = \frac{1}{N^2} \sum_{s \in S} C_s \theta(s_i) \cdot \sum_{s \in S} C_s \theta(s_j) \tag{2}$$

where S is the set of all sequences, $N = |S|$, and s_i is the i-th base in sequence s. The term C_s is the phylogenetic distance to the other sequences, see eq. (4) below. $\theta(s_i, s_j)$ is 1 if (s_i, s_j) is a canonical base pair and 0 otherwise. $\theta(s_i)$ is 1 if s_i is equal to any t_i that forms a canonical base pair (t_i, t_j) where $t \in S$ and 0 otherwise. Canonical base pairs are the pairs G:C, C:G, G:U, U:G, A:U, and U:A.

The mutual information based measure $\mathrm{MI}^{W\diamond P}$ only considers canonical base pairs and includes a gap penalty and is defined as

$$\mathrm{MI}^{W\diamond P}_{i,j} = O_{ij} \cdot \log_2 \frac{O_{ij}}{E_{ij}} - N^{\mathrm{G}}_{ij} \cdot \beta \tag{3}$$

with N^{G}_{ij} being the number of sequences that contain at least one gap in the two columns i and j and $\beta = \frac{1}{N}$ (Lindgreen *et al.* 2006). Note that $\mathrm{MI}^{W\diamond P}$ can become negative if many gaps are present in columns i and j, and is set to 0 in that case.

While $C_s = 1$ for all s in MI and $\mathrm{MI}^{W\diamond P}$, C_s is defined in treeMI and $\mathrm{treeMI}^{W\diamond P}$ as

$$C_s = 1 - \frac{d_{avg}(s)}{N_d} \tag{4}$$

where $d_{\text{avg}}(s)$ is the average pairwise distance from sequence s to all other sequences in the tree and $N_d = \sum_{s \in S} d_{\text{avg}}(s)$.

The term C_s now describes a weighting for each sequence s by its relative average evolutionary distance to all other sequences. Therefore, the information gained by a base pair is dependent from the conservation of its host sequence. Here, the sequence conservation is measured by the distance to all other sequences in the phylogenetic tree. The higher the distance $d_{\text{avg}}(s)$ is, the lower C_s becomes. The gained information becomes larger, the higher the sequence conservation is, with $C_s \rightarrow 1$. The information becomes lower if the sequence is very distant and, thus, compensatory base pair changes are expected by chance, *i.e.*, $C_s \rightarrow 0$.

With the normalisation factor $1/N$ in the calculation of the frequencies O_{ij} and E_{ij}, treeMI and treeMI$^{W \diamond P}$ can be greater than 0 even for completely conserved columns. To avoid this, the normalisation can be done by $1/N_d$. However, this will remove some information gained by the tree. Also columns with just one fully conserved column will get an MI of zero.

References

[Gorodkin *et al.*(1997)] Gorodkin, J., Heyer, L., Brunak, S., and Stormo, G. (1997). Displaying the information contents of structural RNA alignments: the structure logos. *Comput Appl Biosci*, **13**, 583–6.

[Lindgreen *et al.* (2006)] Lindgreen, S., Gardner, P., and Krogh, A. (2006). Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics*, **22**, 2988–95.