Supporting Information

Francesca Cuturello [1], Guido Tiana [2], Giovanni Bussi [1]

February 18, 2020

1 Supplementary Methods

1.1 Alignments

In order to conduct a proper analysis of nucleotide co-evolution, homologous RNA sequences need to be aligned through a process named multiple sequence alignment (MSA). A number of different algorithms have been proposed to this aim. The results of any co-evolutionary analysis will depend on this initial step. We here tested two commonly used MSA algorithms, namely those implemented in *ClustalW* [1] and *Infernal* [2].

MSAs are matrices $\{\sigma^b\}_{b=1}^{B}$ of B homologous RNA sequences that have been aligned through insertion of gaps to have a common length N, so that each sequence can be represented as $\sigma^b = \{\sigma_1^b, ..., \sigma_N^b\}$. Vector σ has entries from a $q = 5$ letters alphabet $\{A, U, C, G, -\}$ coding for nucleotide type, where $-$ represents a gap. $F_i(\sigma)$ denotes the empirical frequency of nucleotide σ at position i and $F_{ij}(\sigma, \tau)$ the frequency of co-occurence of nucleotides σ and τ at positions i and j, respectively:

$$F_i(\sigma) = \frac{1}{B}\sum_{b=1}^{B}\delta(\sigma_i^b, \sigma) \tag{1}$$

$$F_{ij}(\sigma, \tau) = \frac{1}{B}\sum_{b=1}^{B}\delta(\sigma_i^b, \sigma)\delta(\sigma_j^b, \tau) \tag{2}$$

Here δ is the Kronecker symbol (which equals one if the two arguments coincide and zero elsewhere) and σ_k^b is the nucleotide located at position k in the b-th sequence of the MSA. In order to reduce the effect of possible sampling biases in the MSA we adopt the reweighting scheme as in [3] with sequences similarity threshold 0.9. However, we did not find significant difference in test cases where the reweighting scheme was omitted (Supporting Information, Table 5).

1.2 Direct coupling analysis

The idea of direct coupling analysis is to infer a global statistical model $P(\sigma)$ that is able to generate the empirical data (single-site and two-sites frequency

counts) [4], such that

$$F_i(\sigma_i) = \sum_{\{\sigma_k | k \neq i\}} P(\sigma_1, ..., \sigma_N) \equiv f_i(\sigma_i) \tag{3}$$

$$F_{ij}(\sigma_i, \tau_j) = \sum_{\{\sigma_k | k \neq i,j\}} P(\sigma_1, ..., \sigma_N) \equiv f_{ij}(\sigma_i, \tau_j) \tag{4}$$

Introducing a set of Lagrange multipliers $\boldsymbol{\theta} \equiv \{h_i(\sigma), J_{ij}(\sigma, \tau)\}$ to constrain the model averages $\boldsymbol{f} \equiv \{f_i(\sigma), f_{ij}(\sigma, \tau)\}$ to the observed frequencies $\mathbf{F}$, the maximum entropy distribution over the sequences takes the form

$$P(\{\sigma\}) = \frac{1}{Z} \exp\left(\sum_i h_i(\sigma_i) + \sum_{ij} J_{ij}(\sigma_i, \sigma_j) \right) \tag{5}$$

corresponding to a five-states fully connected Potts model, where

$$Z = \sum_{\{\sigma\}} \exp\left(\sum_i h_i(\sigma_i) + \sum_{ij} J_{ij}(\sigma_i, \sigma_j) \right) \tag{6}$$

is the partition function, $h_i(\sigma_i)$ are called *local fields*, while $J_{ij}(\sigma, \tau)$ are called *direct couplings* and can be interpreted as the direct interaction between nucleotides σ and τ at positions i and j, after disentangling them from the interaction with nucleotides sited at other positions. The partition function requires a sum to be done over all the possible sequences of a given length. For a multiple sequence alignment of length N nucleotides, this would amount to 5^N different sequences, where 5 includes the 4 nucleobases and the gap. Once parameters $h_i(\sigma)$ and $J_{ij}(\sigma, \tau)$ have been determined, the Frobenius norm [3, 5, 6] of the coupling matrices can be used to obtain a scalar value for each pair of positions:

$$S_{ij} = \sqrt{\sum_{\{\sigma,\tau\}} J_{ij}(\sigma, \tau)^2} \tag{7}$$

We will discuss three different approaches that can be used to determine the parameters of the model: the mean-field approximation [4], the pseudo-likelihood maximization [6], and a Boltzmann-learning approach proposed here.

1.3 Mean field approximation

In the mean-field approximation, the effect of all nucleotides on any given one is approximated by a single averaged effect, reducing a many-body problem to a one-body problem. The mean-field approach is the one adopted in [4], by which coupling matrices are estimated as the inverse of the connected correlation matrices: $J_{ij}(\sigma_i, \sigma_j) \simeq -C_{ij}^{-1}(\sigma_i, \sigma_j)$, where $C_{ij}(\sigma_i, \sigma_j) = F_{ij}(\sigma_i, \sigma_j) - F_i(\sigma_i)F_j(\sigma_j)$, and the local fields are estimated as $h_i(\sigma_i) \simeq$

$\ln \frac{F_i(\sigma_i)}{F_i(\bar{\sigma_i})} - \sum_{j,j\neq i} \sum_{\substack{\sigma_i,\\ \sigma_i\neq\bar{\sigma}_i}} J_{ij}(\sigma_i,\sigma_j)F_j(\sigma_j)$, where $\bar{\sigma}$ is an arbitrarily chosen letter of the alphabet, usually the one representing gaps. To make the matrix invertible and alleviate finite sample effects it is common to add pseudo-counts as $\hat{F}_i = (1-\lambda)F_i + \frac{\lambda}{5}$ and $\hat{F}_{ij} = (1-\lambda)F_{ij} + \frac{\lambda}{25}(1-\delta_{ij}) + \frac{\lambda}{5}\delta_{ij}\delta_{\sigma_i\sigma_j}$, where $\lambda = 0.5$ [3].

1.4 Pseudo-likelihood maximization

An alternative approach to estimate the DCA inverse problem solution can be that of minimizing the negative pseudo-log likelihood function $l_{pseudo} = -\frac{1}{B}\sum_r \sum_{b=1}^{B} \log P(\sigma_r^b|\boldsymbol{\sigma}_{\setminus r}^b)$. Here $\boldsymbol{\sigma}_{\setminus r}^b$ denotes the identity of all the nucleotides *except* the one at position r, and thus $P(\sigma_r^b|\boldsymbol{\sigma}_{\setminus r}^b)$ is the conditional probability of observing one variable σ_r given all the other variables. When data is abundant, maximizing the conditional likelihood tends to maximizing the full likelihood (see, e.g., [7, 8]). Pseudo-likelihood maximization allows to overcome the intractable evaluation of the full partition function, since calculating the normalization of the conditional probability only requires an empirical average over the dataset. In this paper we will exploit the asymmetric pseudo-likelihood maximization [6] as implemented at `https://github.com/magnusekeberg/plmDCA`.

1.5 Gauge invariance and regularization

The number of model parameters in Eq. 5 is $\frac{N(N-1)}{2}q^2 + Nq$ but the model is over-parametrized, in the sense that distinct parameter sets can describe the same probability distribution. This is because the consistency conditions (Eq. 3) are not independent, single-site marginals being implied by the two-sites marginals and all distributions being normalized; thus the number of independent parameters turns out to be $\frac{N(N-1)}{2}(q-1)^2 + N(q-1)$ [9]. In order to remove the degeneracy of the mean-field solution so to obtain a unique and reproducible result, a possible *gauge* choice for the Potts model [10, 3] is the one minimizing the norm of couplings matrices (Eq. 7), namely $\sum_{\{\tau\}} J_{ij}(\sigma,\tau) = \sum_{\{\tau\}} J_{ij}(\tau,\sigma) = \sum_{\{\tau\}} h_i(\tau) = 0$, $\forall\ i,j,\sigma,\tau$. Another possible gauge is the one in which parameters relative to a specific letter of the alphabet $\bar{\sigma}$ (usually the one representing the gaps) are set to zero: $J_{ij}(\bar{\sigma},\tau) = J_{ij}(\tau,\bar{\sigma}) = h_i(\bar{\sigma}) = 0$, $\forall\ i,j,\tau$ In the Boltzmann learning and pseudo-likelihood maximization frameworks, the degeneracy can alternatively be removed by minimizing a function obtained by the addition of an l_2-regularization term to $l(\theta)$ [10], such that:

$$\theta = \arg\min_{\theta}\{l(\theta) + R(\theta)\} \tag{8}$$

where $R(\theta) = \frac{k}{2}\sum_p \theta_p^2$ and $p = \{1, ..., \frac{N(N-1)}{2}q^2 + Nq\}$. For the Boltzmann learning approach we heuristically observed that a regularization is not necessary and that results are not sensitive to the choice of k, and we thus decided not to use any regularization term. For pseudo-likelihood we used a value of k

depending on the alignment size, using the default options supplied by the employed software. Different prefactors where also tested (see Figure 12).

1.6 Mutual information

The mutual information between two positions i and j is defined as

$$MI_{ij} = \sum_{\sigma_i,\tau_j} F_{ij}(\sigma_i,\tau_j) \ln \frac{F_{ij}(\sigma_i,\tau_j)}{F_i(\sigma_i)F_j(\tau_j)} \equiv S_{ij} \tag{9}$$

and is a local measure of the mutual dependence between two random variables, quantifying how much the uncertainty about one of the two variables is reduced by knowing the other. It is the simplest possible way to assess covariance [11] and its capability to predict contacts in RNA has been reported to be surpassed by DCA-based methods [3].

1.7 Capability of the inferred couplings to reproduce frequencies

The capability of various DCA methods to infer correct parameters for the Potts model can be quantified by computing the root-mean-square deviation (RMSD) between model and observed pair frequencies:

$$RMSD = \sqrt{< (f_{ij}(\sigma_i,\tau_j) - F_{ij}(\sigma_i,\tau_j))^2 >_{\{ij\},\{\sigma_i,\tau_j\}}} \tag{10}$$

For Boltzmann-learning DCA, the model frequencies are calculated in the validation phase of simulations, and the RMSD can be used to assess the convergence of the simulation. For other DCA methods one can simply use the estimated couplings to run a simulation in sequence space.

1.8 Validation of the predicted contacts

We perform this analysis on sequences of a number of riboswitches families classified in the Rfam database [12]. Columns with more than 90% of gaps were removed from the alignments in order to make the maximization faster and to avoid overfitting the model on positions of the alignment that are not relevant. The 17 RNA families have been chosen among those for which at least one high-resolution crystallographic structure have been reported, ruling out from the analysis the structures annotated as interacting double chains. A full list is reported in Table 1. The number of nucleotides in each chain ranges between 52 and 161, and the effective number of sequences between 25 and 1078 (all details are reported in Supporting Information, Table 1). The lowest quality structure in the data set has been solved with resolution 2.95Å. Contacts in the reference PDB structures are annotated with DSSR [13], that takes into account all hydrogen bonds and classify base pairs according to the Westhof-Leontis nomenclature [14]. This is different from other works where

the geometric distance between heavy atoms belonging to each nucleotide, thus including also backbone atoms, is used, and is expected to better report on the direct base-base contacts that are supposed to be associated to covariation. We decided to ignore stacking interactions since coevolution in RNA is mostly related to isostericity [15, 16]. All the used MSAs as well as files containing the annotation of each base pair are available at https://github.com/bussilab/bl-dca.

Before computing the one-site and two-sites frequencies, the columns of the MSA where the sequence corresponding to the reference crystallographic structure had a gap were eliminated by the alignment. Whereas this step should not be in principle required, preliminary calculations showed that this pruning improves the quality of the results for all the tested DCA methods (data not shown). In addition, we applied to the score of each contact (Eq. SI5) the so-called average-product correction (APC) [17].

Evaluation of the performance of RNA contact prediction methods requires the number of correct predictions (true positives, TP), the number of contacts predicted but absent in the native structure (false positives, FP), and the number of contacts present in the native structure but not predicted (false negatives, FN). Two common measures are sensitivity and precision, where *sensitivity* is the fraction of correctly predicted base pairs of all true base pairs, while *precision* is the fraction of true base pairs of all predicted base pairs:

$$sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$precision = \frac{TP}{TP + FP} \tag{12}$$

The Matthews correlation coefficient (MCC) can be defined as the geometric average of sensitivity and precision [18, 19]

$$MCC = \sqrt{sensitivity \cdot precision} \tag{13}$$

and is equivalent to the interaction network fidelity [20]. To turn contact scores S_{ij} (either Eq. SI7, for mutual information, or Eq. SI5, for DCA, or E-values, for R-scape) into predictions it is necessary to assume a threshold $\bar{S}$. The predicted contacts will be those scored by a value above (below, for R-scape) $\bar{S}$. For R-scape, we used the recommended threshold $\bar{S} = 0.05$. For the other methods, we chose the threshold score maximizing the MCC, corresponding to the optimal compromise between precision and sensitivity. For each covariance method, the MCC as a function of the threshold score S shows a similar behavior for all the N_s=17 systems, their peaks falling at very similar positions. This suggests the possibility to set a unique threshold for each covariance method that maximizes the MCC geometric average over all systems:

$$\bar{S} = \arg\max_{S} \left(\prod_{\mu}^{N_s} MCC_\mu(S) \right)^{\frac{1}{N_s}} \tag{14}$$

Table 1: PDB, RFAMcode molecule name, alignment length and size, effective alignment size after reweighting of the data set.

PDB	**RFAM**	**molecule name**	**length**	**size**	**size$_{\mathbf{eff}}$**
4L81	RF01725	SAM-I/IV variant riboswitch	97	693	128
2GDI	RF00059	TPP riboswitch	80	10858	1054
3F2Q	RF00050	FMN riboswitch	109	3144	1078
2GIS	RF00162	SAM riboswitch	93	4903	910
1Y26	RF00167	Purine riboswitch	71	2589	508
3DOU	RF00168	Lysine riboswitch	161	1870	832
4QLM	RF00379	ydaO/yuaA leader	108	2723	1067
2QBZ	RF00380	ykoK leader	153	850	240
5T83	RF00442	ykkC-yxkD leader	89	687	138
3OWI	RF00504	Glycine riboswitch	88	4602	985
3IRW	RF01051	Cyclic di-GMP-I riboswitch	91	2231	578
4FRG	RF01689	AdoCbl variant RNA	84	189	25
3VRS	RF01734	Fluoride riboswitch	52	1426	312
5DDP	RF01739	Glutamine riboswitch	61	1138	179
4XW7	RF01750	ZMP/ZTP riboswitch	64	1197	432
3SD3	RF01831	THF riboswitch	89	547	205
4RUM	RF02683	NiCo riboswitch	92	207	42

Table 2: $\overline{MCC}$ with optimal covariance score threshold $\overline{S}$ for Boltzmann learning DCA, pseudo-likelihood DCA, mean field DCA, mutual information for each of 17 RNA families, obtained through cross-validation procedure. Alignments are performed with *Infernal*.

	Boltzmann learning DCA		**Pseudo-likelihood DCA**		**mean field DCA**		**mutual information**	
PDB	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$
3DOU	0.68	1.09	0.59	0.65	0.67	1.0	0.68	0.22
3F2Q	0.58	1.09	0.58	0.65	0.56	1.0	0.55	0.22
2QBZ	0.55	1.09	0.50	0.78	0.52	1.0	0.53	0.22
2GDI	0.55	1.09	0.51	0.65	0.57	1.0	0.48	0.22
1Y26	0.69	1.09	0.67	0.65	0.63	0.99	0.63	0.22
5T83	0.58	1.09	0.58	0.65	0.58	1.0	0.53	0.22
5DDP	0.65	1.09	0.63	0.65	0.66	1.0	0.65	0.22
4XW7	0.59	1.24	0.63	0.65	0.59	1.0	0.55	0.22
4RUM	0.60	1.19	0.39	0.78	0.54	1.06	0.55	0.22
4L81	0.46	1.09	0.45	0.78	0.43	1.0	0.35	0.22
4FRG	0.63	1.09	0.49	0.78	0.50	0.99	0.64	0.22
3SD3	0.67	1.05	0.69	0.65	0.67	1.0	0.63	0.22
2GIS	0.67	1.14	0.74	0.65	0.44	1.03	0.37	0.22
3OWI	0.73	1.11	0.73	0.65	0.67	1.0	0.29	0.24
3IRW	0.58	1.09	0.56	0.65	0.50	1.0	0.35	0.22
4QLM	0.56	1.05	0.58	0.65	0.49	1.0	0.43	0.22
3VRS	0.64	1.11	0.71	0.65	0.71	1.0	0.67	0.22

Table 3: Clustal alignment. $\overline{MCC}$ with optimal covariance score threshold $\overline{S}$ for Boltzmann learning DCA, pseudo-likelihood DCA, mean field DCA, mutual information for each of 17 RNA families, obtained through cross-validation procedure.

	Boltzmann learning DCA		**Pseudo-likelihood DCA**		**mean field DCA**		**mutual information**	
PDB	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$
3DOU	0.47	1.07	0.45	0.43	0.42	0.82	0.47	0.20
3F2Q	0.48	0.99	0.45	0.43	0.32	0.80	0.31	0.20
2QBZ	0.49	1.07	0.46	0.51	0.45	0.80	0.39	0.20
2GDI	0.44	1.07	0.35	0.47	0.35	0.82	0.29	0.20
1Y26	0.57	1.07	0.50	0.43	0.51	0.82	0.32	0.20
5T83	0.41	1.07	0.38	0.43	0.32	0.82	0.44	0.20
5DDP	0.42	1.10	0.33	0.51	0.19	0.82	0.20	0.20
4XW7	0.38	1.07	0.42	0.43	0.22	0.80	0.19	0.20
4RUM	0.46	1.07	0.32	0.51	0.24	0.80	0.37	0.20
4L81	0.27	1.07	0.29	0.45	0.18	0.80	0.16	0.20
4FRG	0.59	1.07	0.44	0.57	0.34	0.82	0.40	0.20
3SD3	0.71	1.07	0.72	0.45	0.58	0.8	0.50	0.20
2GIS	0.54	0.99	0.54	0.43	0.40	0.82	0.34	0.20
3OWI	0.42	1.07	0.48	0.47	0.40	0.82	0.24	0.20
3IRW	0.55	1.07	0.37	0.44	0.39	0.80	0.25	0.20
4QLM	0.38	1.07	0.45	0.51	0.30	0.80	0.10	0.23
3VRS	0.55	1.08	0.42	0.43	0.42	0.82	0.34	0.20

Table 4: Average $\overline{MCC}$ at optimal covariance score threshold for DCA methods with and without APC correction. Alignments are performed with *Infernal*.

	Boltzmann learning DCA		**Pseudo-likelihood DCA**		**mean field DCA**	
	APC	no APC	APC	no APC	APC	no APC
average $\overline{MCC}$	0.61	0.59	0.59	0.56	0.57	0.54

Table 5: Reweighting scheme: two sequences are considered similar if the fraction of positions with coincident nucleotides (*similarity*) is larger than a given similarity threshold x: $n_b = |s \in \{1, ..., B\} : similarity(\sigma^s, \sigma^b) > x|$. The inverse of n_b, $\omega_b = \frac{1}{n_b}$, gives a weight for the sequence contribution to frequencies ($B_{eff} = \sum_{b=1}^{B} \omega_b$ is then the effective number of sequences in the alignment). In this table we report the average $\overline{MCC}$ at optimal covariance score threshold for pseudo-likelihood DCA in the reweighting scheme adopting different similarity thresholds x. Alignments are performed with *Infernal*.

	x=0.7	**x=0.8**	**x=0.9**	**x=1.0**
average $\overline{MCC}$	0.59	0.59	0.59	0.59

Table 6: Non-canonical tertiary contacts predicted via Boltzmann learning DCA on Infernal alignments.

PDB	**Total non-canonical contacts**	**Predicted non-canonical contacts**	**Type of base pairing**
1Y26	12	1	cSS
2GDI	14	1	tSS
2GIS	14	4	cSS,tSH,c.H,tWS
2QBZ	30	1	tSH
3DOU	21	5	t.H,tSS,tSH,tHS,tHS
3F2Q	17	3	tHS,cSS,cHW
3IRW	11	0	-
3OWI	11	1	tHS
3VRS	5	0	-
5SD3	10	1	tHS
4FRG	11	0	-
4L81	15	0	-
4RUM	7	0	-
4QLM	15	4	tSH,tSH,cSS,tHS
4XW7	5	0	-
5DDP	11	1	...
5T83	21	3	t.H, tSH,tHW

Table 7: Total stacked false positives (base atomes distance < 3.5 Å in the pdb reference structure) over total false positives for all methods. (*Infernal* alignment).

	Boltzmann learning	**Pseudo-likelihood DCA**	**Mean Field**	**Mutual Information**
stacked FP / FP	0.43	0.46	0.39	0.39

Table 8: $\overline{MCC}$ for each of 17 RNA families obtained through cross-validation procedure with optimal probability threshold $\overline{S}$. Base pairing probabilities are calculated from the RNAfold program available in the ViennaRNA package. We notice that for PDB 5T83 the MCC is zero for thresholds larger than ≈0.5, leading to a very low $\overline{S}$ whenever that system is included in the training set.

PDB	$\overline{MCC}$	$\overline{S}$
3DOU	0.51	0.25
3F2Q	0.52	0.25
2QBZ	0.52	0.25
2GDI	0.52	0.25
1Y26	0.51	0.25
5T83	0.58	0.72
5DDP	0.51	0.25
4XW7	0.51	0.25
4RUM	0.50	0.25
4L81	0.51	0.25
4FRG	0.53	0.26
3SD3	0.52	0.25
2GIS	0.51	0.25
3OWI	0.50	0.25
3IRW	0.53	0.25
4QLM	0.52	0.25
3VRS	0.54	0.25

Table 9: *Infernal* alignment. $\overline{MCC}$ for each of 17 RNA families obtained through cross-validation procedure with optimal E-value threshold $\overline{S}$ and MCC at default R-scape threshold (E-value=0.05). E-values for base pairs are calculated from the R-scape program.

PDB	$\overline{S}$	$\overline{MCC}$ at $\overline{S}$	MCC at $S = 0.05$
3DOU	0.5	0.51	0.56
2GIS	0.5	0.59	0.55
1Y26	0.5	0.48	0.56
3VRS	0.1	0.4	0.4
2GDI	0.5	0.53	0.56
5T83	0.5	0.57	0.53
4FRG	0.5	0.64	0.62
4L81	0.5	0.41	0.42
2QBZ	0.5	0.54	0.53
3F2Q	0.5	0.60	0.52
4QLM	0.5	0.44	0.46
3IRW	0.5	0.40	0.44
5DDP	0.3	0.53	0.5
4RUM	0.5	0.64	0.62
3SD3	0.5	0.61	0.63
3OWI	0.5	0.58	0.60
4XW7	0.5	0.51	0.55

Table 10: *ClustalW* alignment. $\overline{MCC}$ for each of 17 RNA families obtained through cross-validation procedure with optimal E-value threshold $\overline{S}$ and MCC at default R-scape threshold (E-value=0.05). E-values for base pairs are calculated from the R-scape program. We notice that for PDB 5DDP and 2GIS the MCC is zero for E-value thresholds smaller than 10.

PDB	$\overline{S}$	$\overline{MCC}$ at $\overline{S}$	MCC at $S = 0.05$
3DOU	1.3	0.40	0.45
2GIS	-	0.0	0.0
1Y26	1.3	0.35	0.16
3VRS	1.3	0.27	0.22
2GDI	1.3	0.44	0.44
5T83	1.3	0.42	0.42
4FRG	1.3	0.47	0.44
4L81	1.3	0.18	0.19
2QBZ	1.3	0.44	0.46
3F2Q	1.3	0.43	0.40
4QLM	1.3	0.16	0.20
3IRW	1.1	0.23	0.28
5DDP	-	0.0	0.0
4RUM	1.1	0.45	0.28
3SD3	1.3	0.58	0.58
3OWI	2.8	0.18	0.21
4XW7	1.3	0.32	0.25

Table 11: $\overline{MCC}$ for each of 17 RNA families obtained considering as contact predictions the top N/2 coupling scores (where N is the sequence length). Average MCC is 0.53. Infernal alignments.

PDB	$\overline{MCC}$
4L81	0.58
2GDI	0.50
3F2Q	0.56
2GIS	0.49
1Y26	0.56
3DOU	0.43
4QLM	0.51
2QBZ	0.50
5T83	0.31
3OWI	0.53
3IRW	0.69
4FRG	0.56
3VRS	0.41
5DDP	0.49
4XW7	0.57
3SD3	0.61
4RUM	0.51

Table 12: Contact prediction via Boltzmann learning DCA on ribosomial RNA subunits 58S and 5S (PDB 1FFK and 2WW9), tRNA (PDB 1ASY) and U4 spliceosomal RNA (PDB 2N7M). MCC obtained at optimal score threshold 1.06. Infernal alignments.

PDB	**RFAM**	**molecule name**	**length**	**size**	**MCC**
1FFK	RF00001	5S ribosomal RNA	122	139785	0.49
2WW9	RF00002	58S ribosomal RNA	63	4727	0.35
1ASY	RF00005	tRNA	75	100000	0.74
2N7M	RF00015	U4 spliceosomal RNA	92	7670	0.38

Algorithm 1 Boltzmann learning direct coupling analysis

1. **Initialization:**
   - Choose randomly 20 sequences from the MSA.
   - Initialize model parameters $\{h, J\}$ to zero.
2. **Learning:** Loop over 100000 Monte Carlo sweeps. For each sweep:
   - Loop over the 20 sequences. For each sequence k:
     - Loop over nucleotide of each sequence. For each nucleotide i:
       - Propose a new random nucleotide at position i
       - Compute the acceptance as $\alpha = \left(1, \frac{P_{new}}{P_{old}}\right)$, where P_{new} and P_{old} are the probabilities of old and new nucleotides at position i according to model parameters $\{h, J\}$.
       - Accept/reject comparing α with a uniform random number in $[0, 1)$.
     - Compute frequencies on the 20 sequences.
   - Update parameters $\{h, J\}$ estimating likelihood gradient based on current frequences.
3. **Validation:** Repeat step 2 using parameters $\{h, J\}$ computed as averages over the last 5000 Monte Carlo sweeps of step 2.

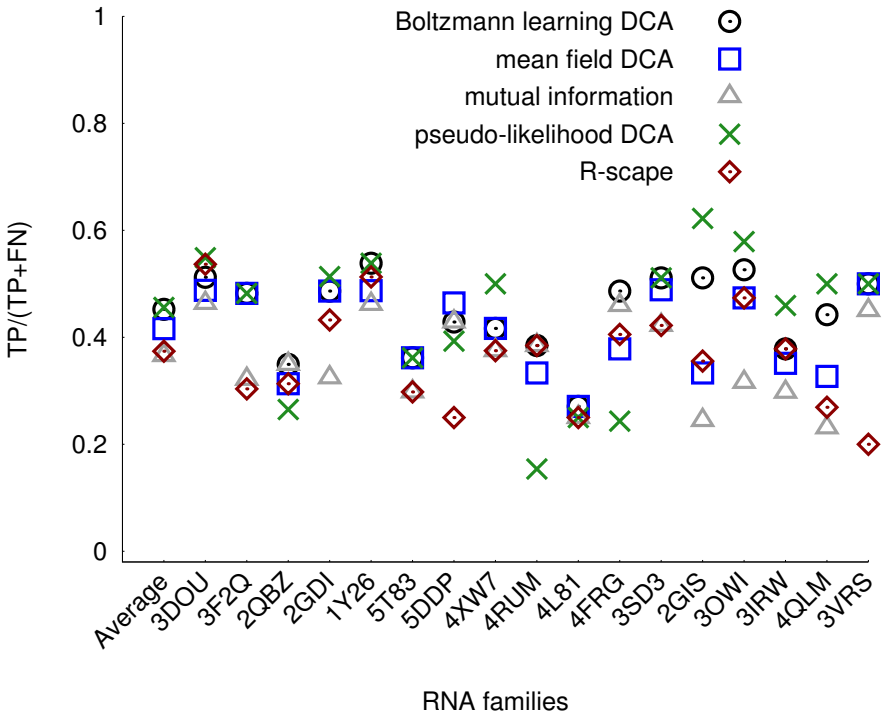

Figure 1: Sensitivity of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

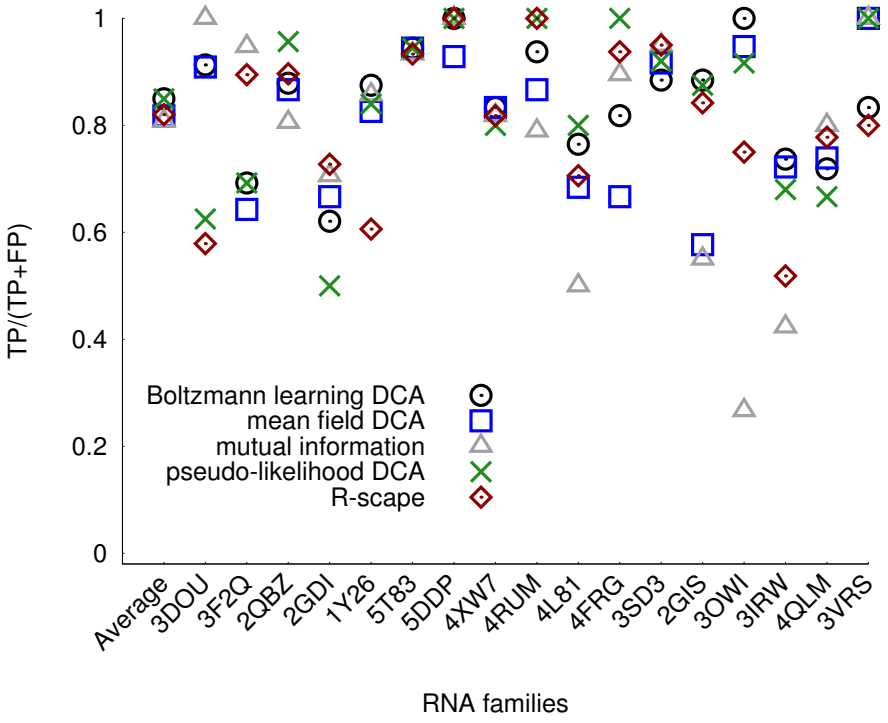

Figure 2: Precision of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

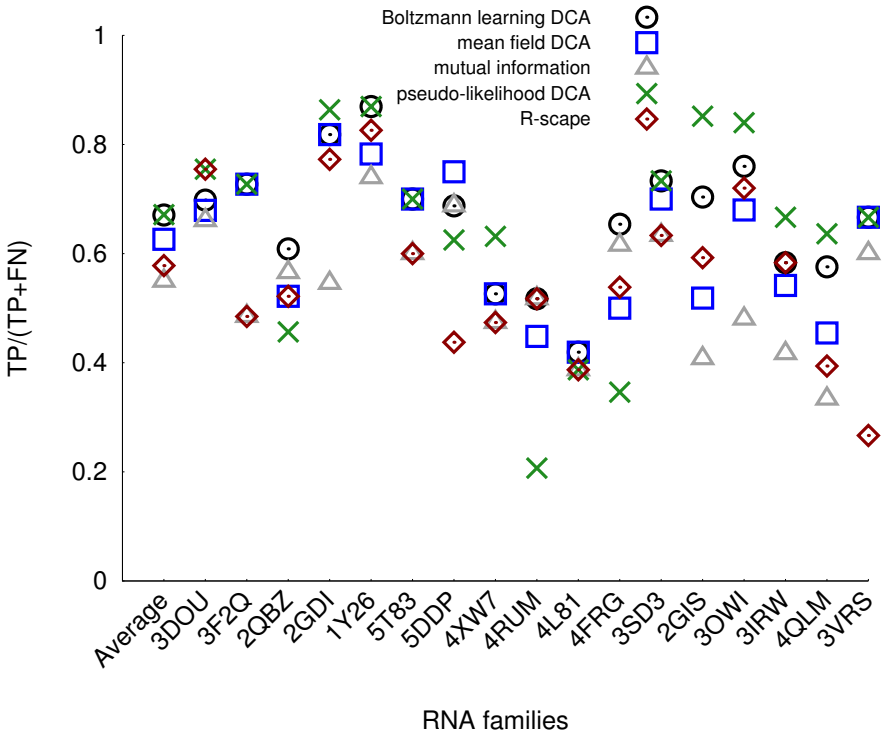

Figure 3: Sensitivity to contacts in stems (RNA secondary structure) of Boltzmann learning DCA, mean field DCA, mutual information and R-scape for all families. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

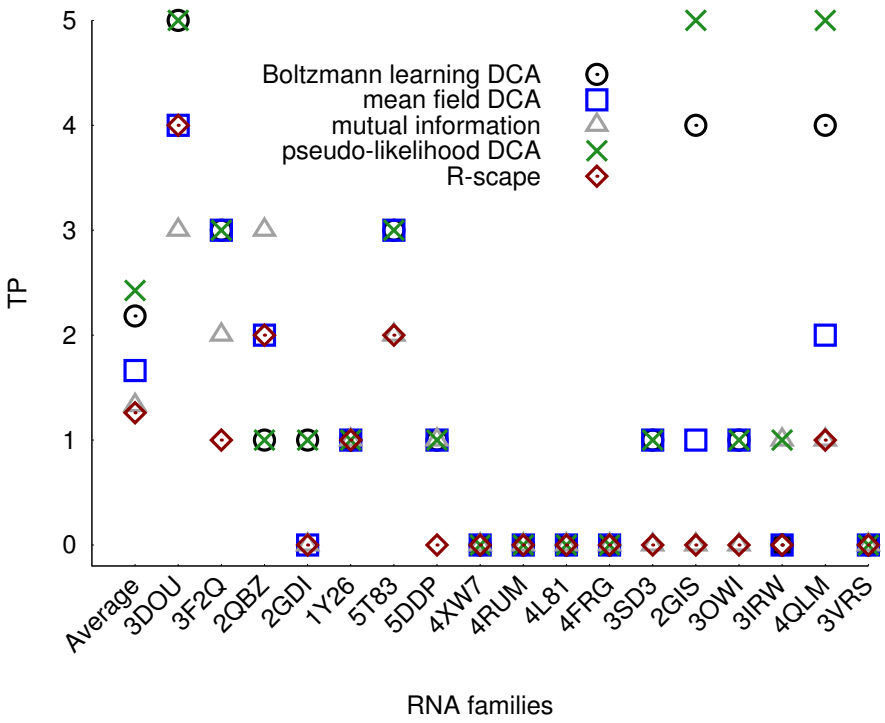

Figure 4: Number of correctly predicted (True Positives) tertiary contacts of Boltzmann learning DCA, mean field DCA, mutual information and R-scape for all RNA families. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

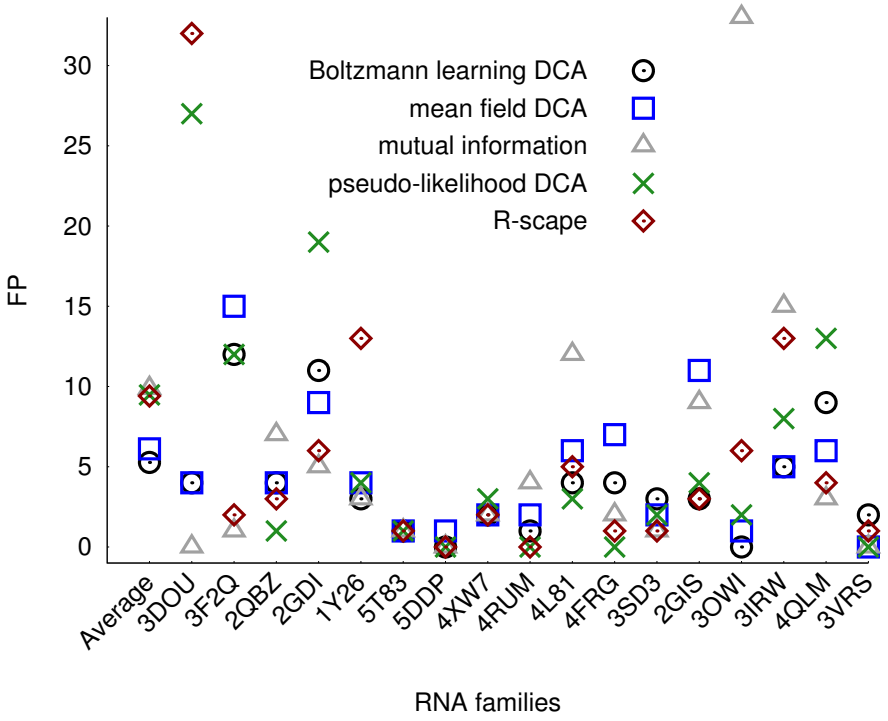

Figure 5: Number of incorrect predictions (False Positives) of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information and R-scape for all RNA families.Families are labeled using the PDB code of the representative crystallographic structure. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

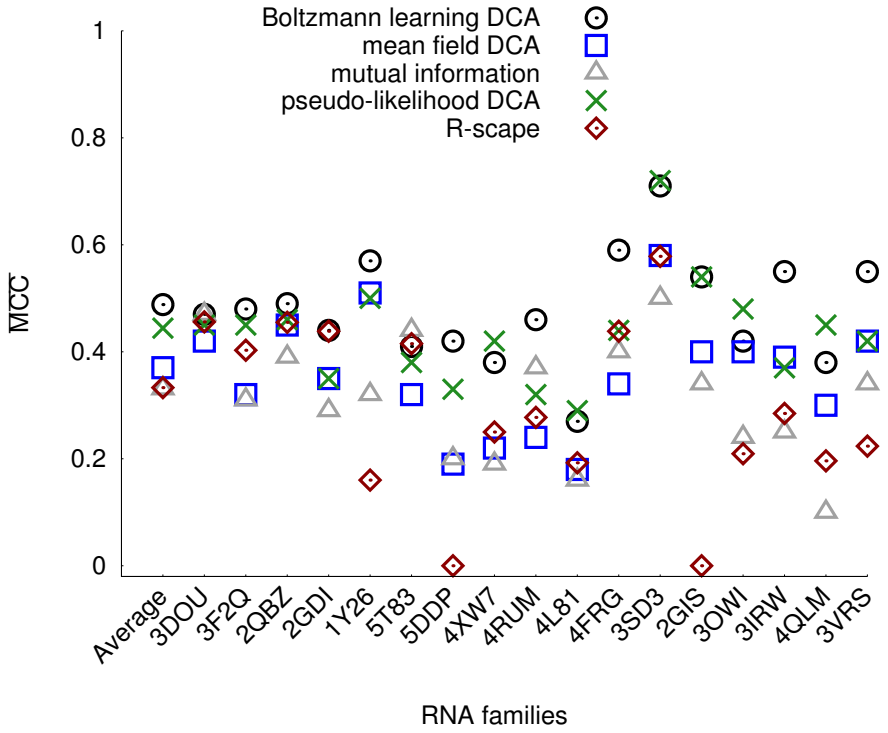

Figure 6: Clustal alignment. $\overline{MCC}$ of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information and R-scape for 17 RNA families at the threshold obtained through cross-validation procedure. Families are labeled using the PDB code of the representative crystallographic structure. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

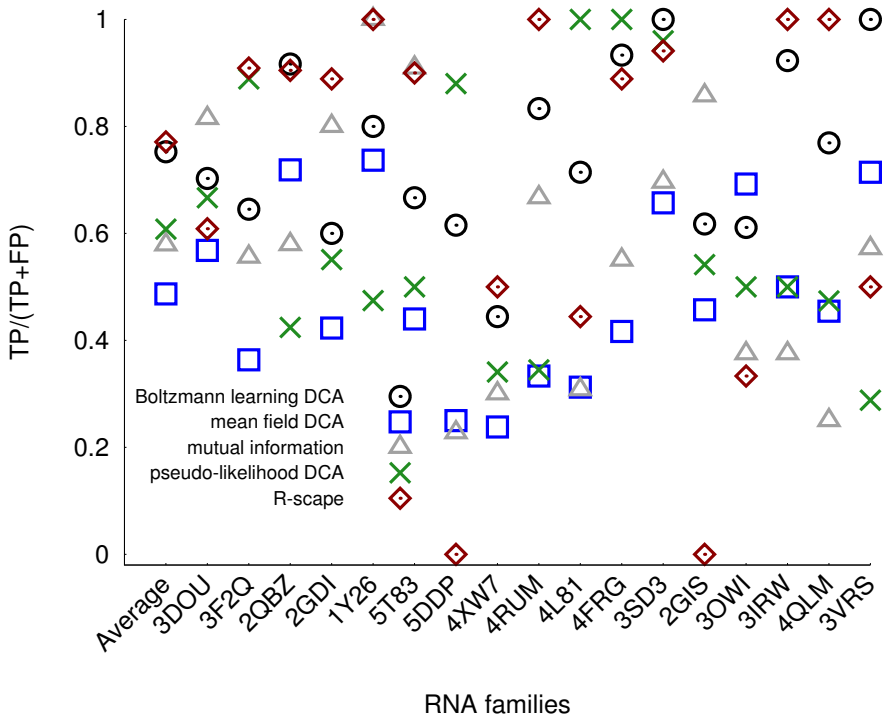

Figure 7: Clustal alignment. Precision of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

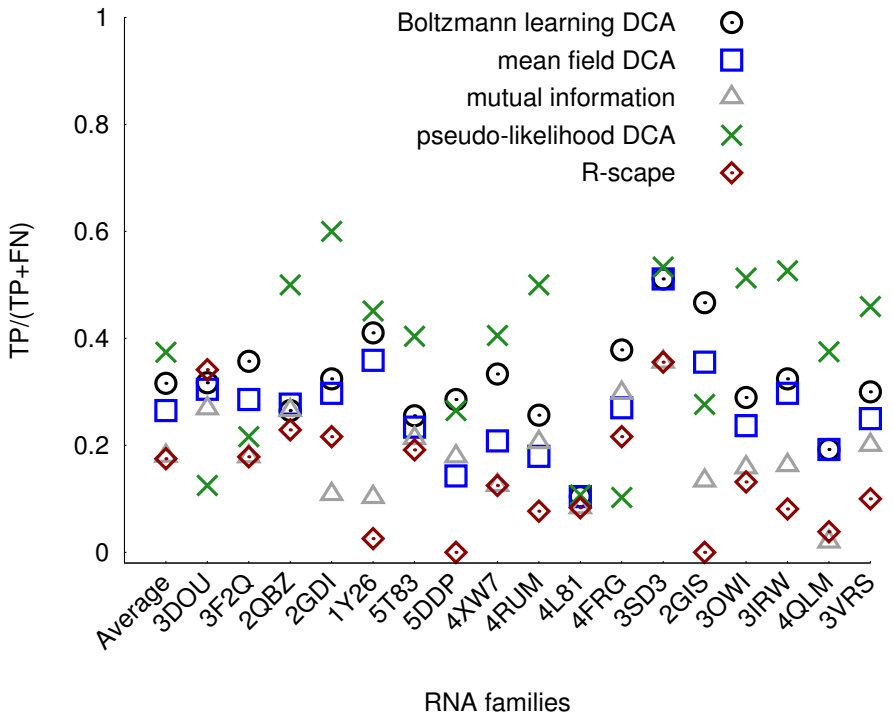

Figure 8: Clustal alignment. Sensitivity of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

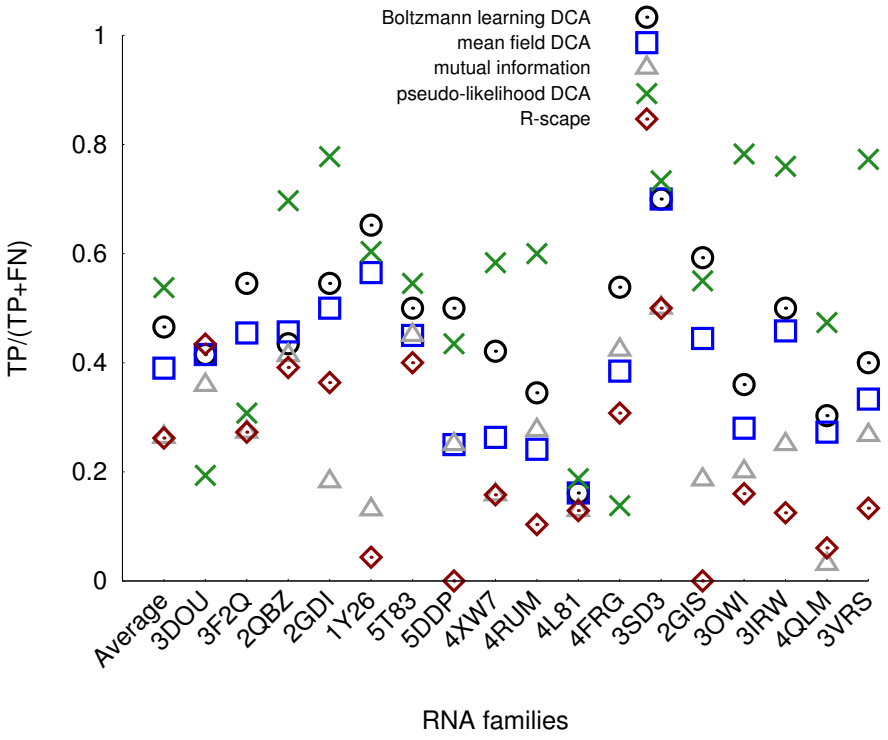

Figure 9: Clustal alignment. Sensitivity to contacts in stems (RNA secondary structure) of Boltzmann learning DCA, mean field DCA, mutual information and R-scape for all families. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

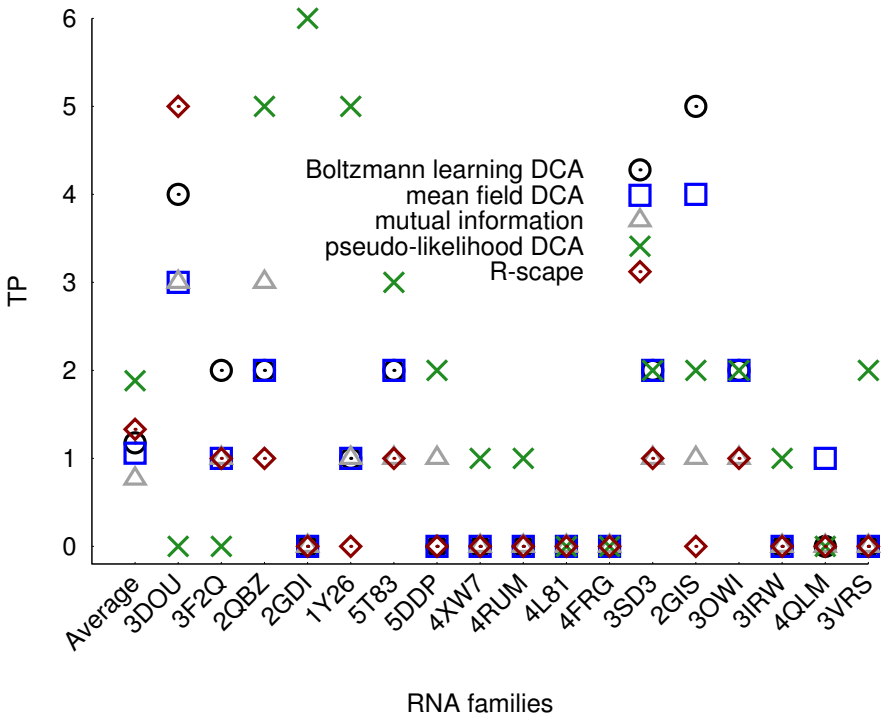

Figure 10: Clustal alignment. Number of correctly predicted (True Positives) tertiary contacts of Boltzmann learning DCA, mean field DCA, mutual information and R-scape for all RNA families. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

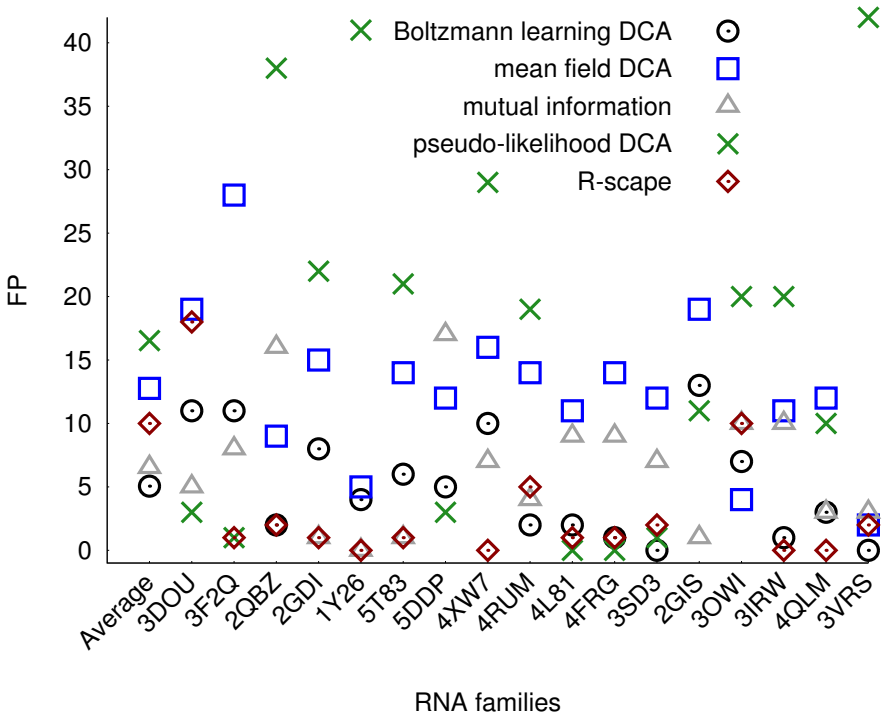

Figure 11: Clustal alignment. Number of incorrect predictions (False Positives) of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information and R-scape for all RNA families.Families are labeled using the PDB code of the representative crystallographic structure. Average is reported in first column. Score threshold is obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape.

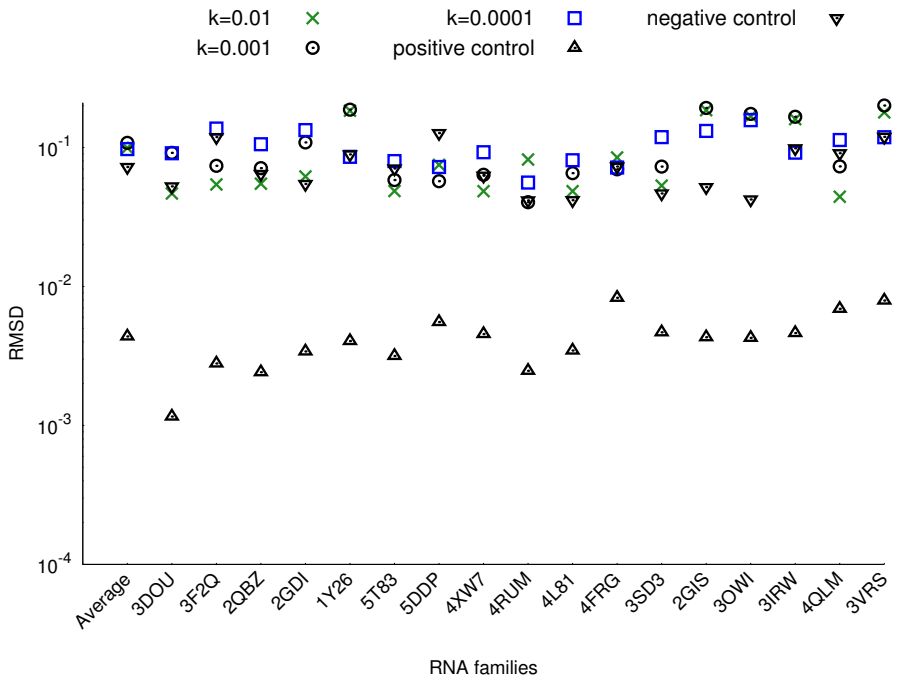

Figure 12: Validation of the coupling parameters inferred via the l_2-regularized pseudo-likelihood maximization method implemented at *https://github.com/magnusekeberg/plmDCA*, adopting different regularization strengths k. The validation is done running a parallel MC simulation on 20 sequences and calculating the root-mean-square deviation (RMSD) between the obtained frequencies and the empirical ones. The positive control is the statistical error due to the finite number of sequence, and the negative control is the RMSD between empirical sequences and a random sequence. Infernal alignment.

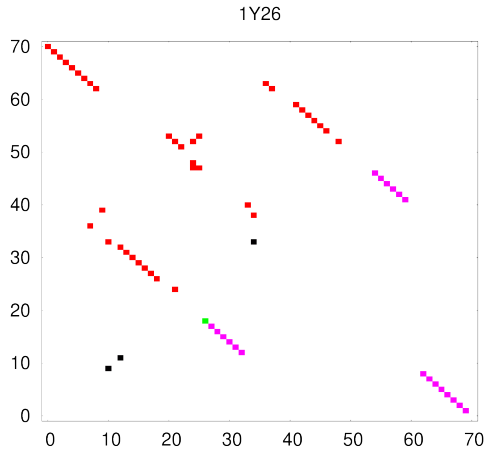

Figure 13: RF00167. Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

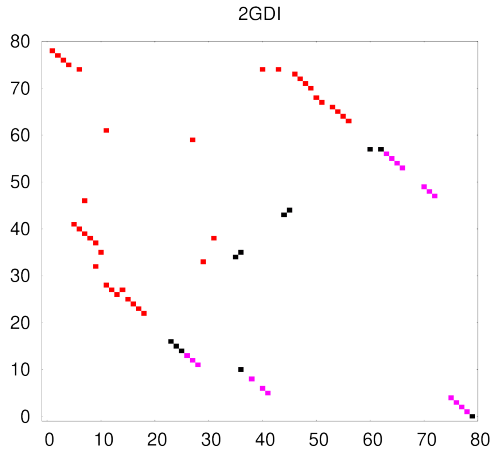

Figure 14: RF00059.Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

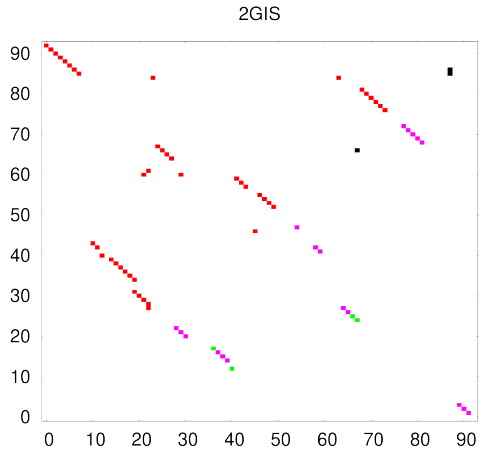

Figure 15: RF00162.Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

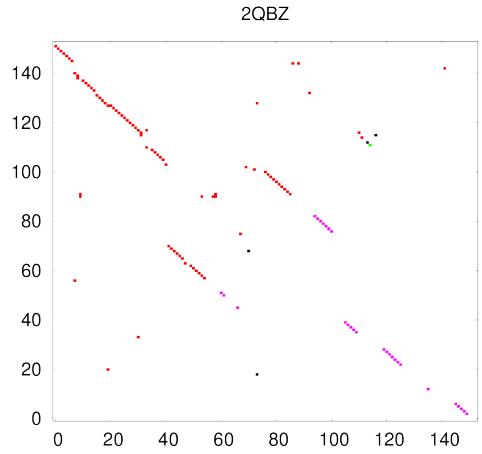

Figure 16: RF00380. Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

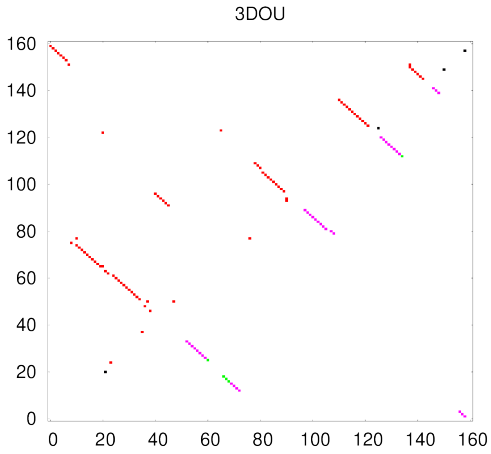

Figure 17: RF00168.Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

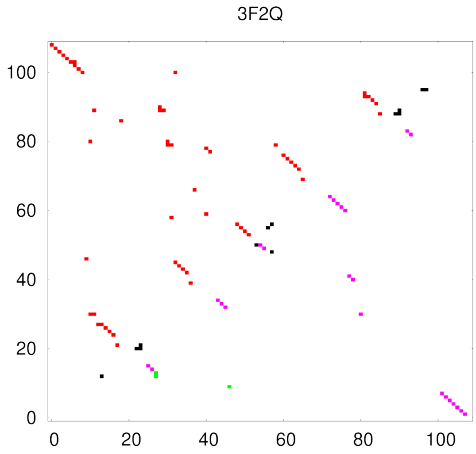

Figure 18: RF00050. Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

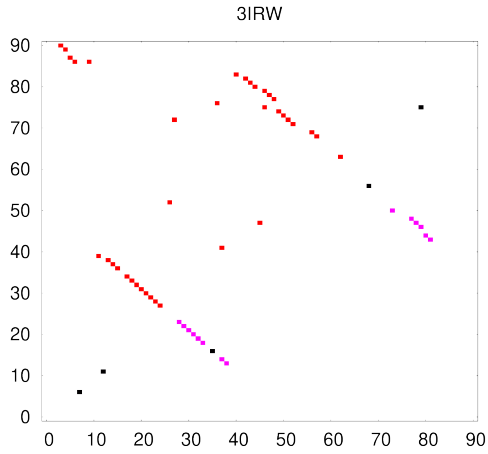

Figure 19: RF01051.Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

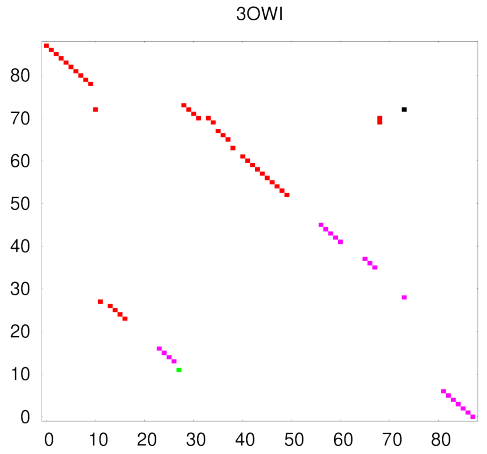

Figure 20: RF00504. Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

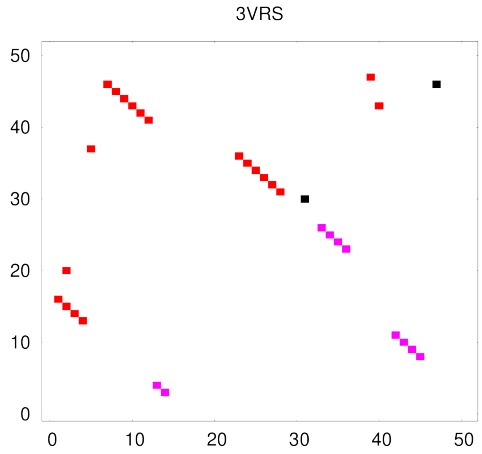

Figure 21: RF01734.Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

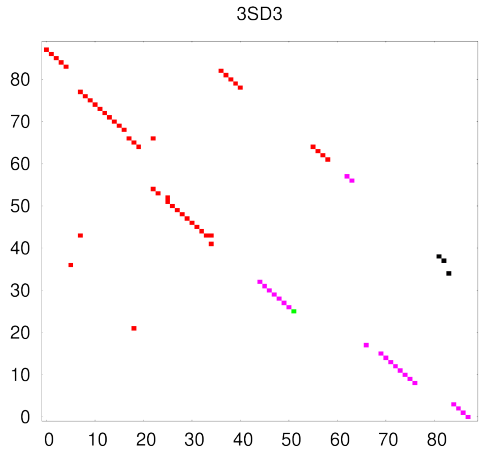

Figure 22: RF01831. Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

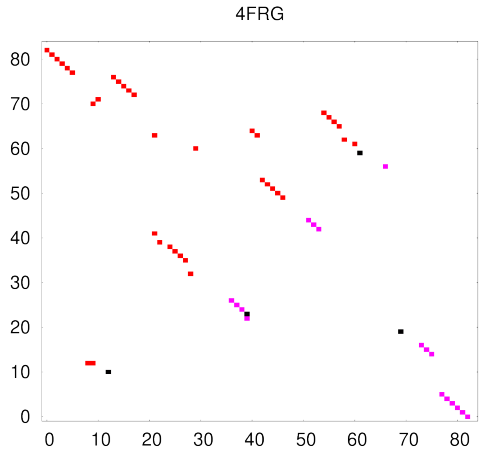

Figure 23: RF01689. Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

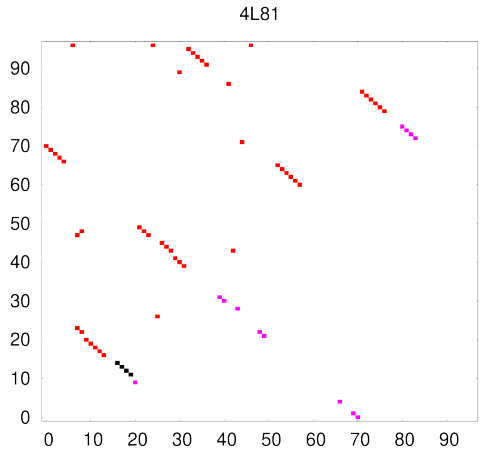

Figure 24: RF01725.Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

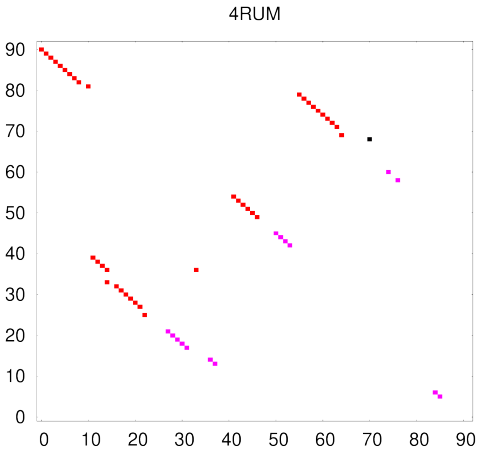

Figure 25: RF02683. Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

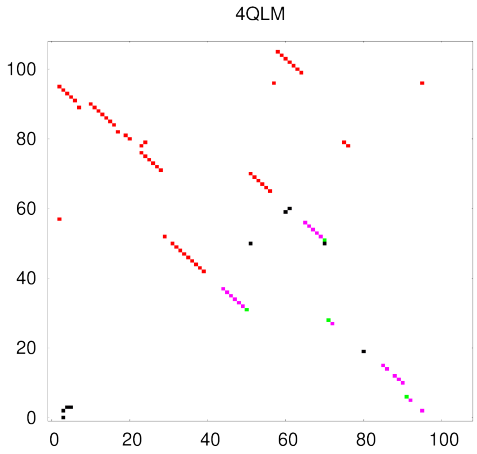

Figure 26: RF00379.Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

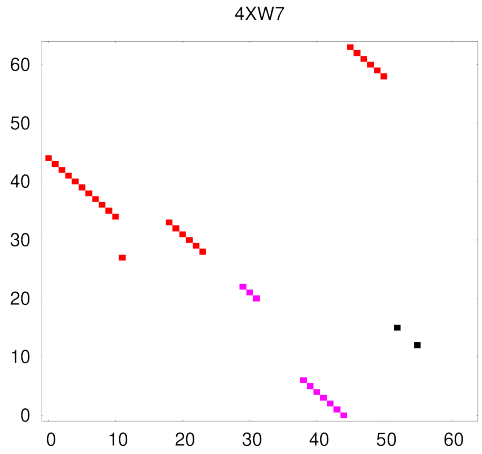

Figure 27: RF01750. Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

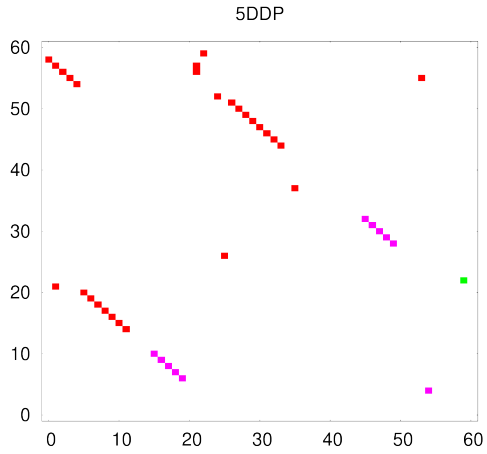

Figure 28: RF01739.Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

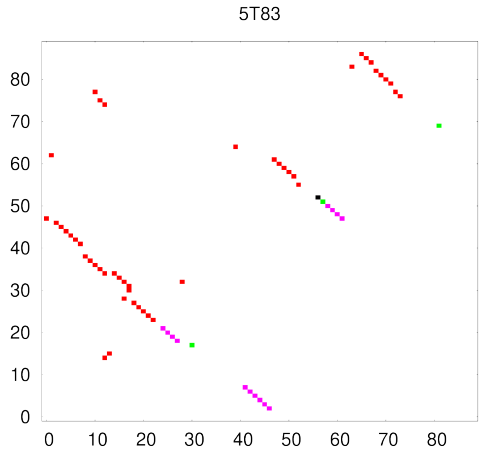

Figure 29: RF00442. Red: native structure base pairs in upper triangle. Magenta: correctly predicted secondary contacts. Green: correctly predicted tertiary contacts. Black: false positives.

References

[1] Julie D Thompson, Desmond G Higgins, and Toby J Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, 1994.

[2] Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

[3] Eleonora De Leonardis, Benjamin Lutz, Sebastian Ratz, Simona Cocco, Rémi Monasson, Alexander Schug, and Martin Weigt. Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.*, 43(21):10444–10455, 2015.

[4] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.*, 108(49):E1293–E1301, 2011.

[5] Simona Cocco, Remi Monasson, and Martin Weigt. From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. *PLoS Comput. Biol.*, 9(8):e1003176, 2013.

[6] Magnus Ekeberg, Tuomo Hartonen, and Erik Aurell. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.*, 276:341–356, 2014.

[7] Barry C Arnold and David Strauss. Pseudolikelihood estimation: some examples. *Sankhyā: The Indian Journal of Statistics, Series B*, 53:233–243, 1991.

[8] Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional ising model selection using l1-regularized logistic regression. *Ann. Stat.*, 38(3):1287–1319, 2010.

[9] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.*, 106(1):67–72, 2009.

[10] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys. Rev. E*, 87(1):012707, 2013.

[11] Sean R Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Res.*, 22(11):2079–2088, 1994.

[12] Eric P Nawrocki, Sarah W Burge, Alex Bateman, Jennifer Daub, Ruth Y Eberhardt, Sean R Eddy, Evan W Floden, Paul P Gardner, Thomas A Jones, John Tate, et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, 43(D1):D130–D137, 2014.

[13] Xiang-Jun Lu, Harmen J. Bussemaker, and Wilma K. Olson. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, 43(21):e142, 2015.

[14] Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499–512, 2001.

[15] Neocles B Leontis, Jesse Stombaugh, and Eric Westhof. The non-watson–crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, 30(16):3497–3531, 2002.

[16] Jesse Stombaugh, Craig L Zirbel, Eric Westhof, and Neocles B Leontis. Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.*, 37(7):2294–2312, 2009.

[17] Stanley D Dunn, Lindi M Wahl, and Gregory B Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, 2007.

[18] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *BBA-Prot. Struct.*, 405(2):442–451, 1975.

[19] Jan Gorodkin, Shawn Stricklin, and Gary Stormo. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.*, 29 10:2135–44, 2001.

[20] Marc Parisien, José Almeida Cruz, Éric Westhof, and François Major. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, 15(10):1875–1885, 2009.