Assessing the accuracy of direct-coupling analysis for RNA contact prediction

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

[1] Scuola Internazionale Superiore di Studi Avanzati, International School for Advanced Studies, via Bonomea 265, 34136 Trieste, Italy
[2] Center for Complexity and Biosystems and Department of Physics, Università degli Studi di Milano and INFN, via Celoria 16, 20133 Milano, Italy

Abstract

Many non-coding RNAs are known to play a role in the cell directly linked to their structure. Structure prediction based on the sole sequence is however a challenging task. On the other hand, thanks to the low cost of sequencing technologies, a very large number of homologous sequences are becoming available for many RNA families. In the protein community, it has emerged in the last decade the idea of exploiting the covariance of mutations within a family to predict the protein structure using the direct-coupling-analysis (DCA) method. The application of DCA to RNA systems has been limited so far. We here perform an assessment of the DCA method on 17 riboswitch families, comparing it with the commonly used mutual information analysis and with state-of-the-art R-scape covariance method. We also compare different flavors of DCA, including mean-field, pseudo-likelihood, and a proposed stochastic procedure (Boltzmann learning) for solving exactly the DCA inverse problem. Boltzmann learning outperforms the other methods in predicting contacts observed in high resolution crystal structures.

1 Introduction

The number of non-coding RNAs with a known functional role has steadily increased in the last years [Morris and Mattick, 2014, Hon *et al.*, 2017]. For a large fraction of them, their function has been suggested to be directly related to their structure [Smith *et al.*, 2013]. For paradigmatic cases such as ribozymes [Doherty and Doudna, 2000], that catalyze chemical reactions, and riboswitches [Serganov and Nudler, 2013], whose aptamer domain has evolved in order to specifically bind physiological metabolites, a well-defined three-dimensional structure is required for function. Secondary structure can be

---

*To whom correspondence should be addressed. Email: bussi@sissa.it

inferred using thermodynamic models [Mathews *et al.*, 2016], often used in combination with chemical probing data [Weeks, 2010]. Tertiary structure is usually determined using more complex techniques based on nuclear magnetic resonance [Rinnenthal *et al.*, 2011] or X-ray diffraction [Westhof, 2015]. Predicting RNA tertiary structure from sequence alone is still very difficult, as it can be seen by the relatively poor predictive performances of molecular dynamics simulations [Šponer *et al.*, 2018] and knowledge-based potentials [Miao *et al.*, 2017]. The low cost of sequencing techniques, however, lead to the accumulation of a vast number of sequence data for many homologous RNA families [Nawrocki *et al.*, 2014]. Covariance of aligned homologous sequences has been traditionally used to help or validate three-dimensional structural modeling (see, *e.g.*, [Michel and Westhof, 1990] and [Costa and Michel, 1997] for early examples). Systematic approaches based on mutual information analysis [Eddy and Durbin, 1994] and related methods [Pang *et al.*, 2005] are now routinely used to construct covariance models and score putative contacts. Recently, a G-test-based statistical procedure called R-scape has been shown to be more robust than plain mutual information analysis for RNA systems [Rivas *et al.*, 2017]. In the last years, in the protein community it has emerged the idea of using so-called direct coupling analysis (DCA) in order to construct a probabilistic model capable to generate the correlations observed in the analyzed sequences [Morcos *et al.*, 2011, Marks *et al.*, 2011, Nguyen *et al.*, 2017, Cocco *et al.*, 2018]: strong direct couplings in the model indicate spatial proximity. The solution of the corresponding inverse model is usually found in the so-called mean-field approximation [Morcos *et al.*, 2011], that is strongly correlated with the sparse inverse covariance approach [Jones *et al.*, 2011]. A further improvement in the level of approximation of the inferred solution is reached when maximizing the conditional likelihood (or *pseudo-likelihood*), which is a consistent estimator of the full likelihood but involves a tractable maximization [Ekeberg *et al.*, 2013] and is considered as the state-of-the-art method for protein sequences.

Whereas covariance methods have been applied to RNA systems since a long time, the application of DCA to RNA structure prediction has so far been limited. The coevolution of bases in RNA fragments with known structure has been investigated [Dutheil *et al.*, 2010], observing strong correlations in Watson-Crick (WC) pairs and much weaker correlations in non-WC pairs. DCA has been first applied to RNA in two pioneering works, using either the mean-field approximation [De Leonardis *et al.*, 2015] or a pseudo-likelihood maximization [Weinreb *et al.*, 2016]. A later work also used the mean-field approximation to infer contacts [Wang *et al.*, 2017]. The mentioned applications of DCA to RNA structure prediction focused on the prediction of RNA three-dimensional structure based on the combination of DCA with some underlying coarse-grain model [De Leonardis *et al.*, 2015, Weinreb *et al.*, 2016, Wang *et al.*, 2017]. However, the performance of the DCA alone is difficult to assess from these works, since the reported results largely depend on the accuracy of the utilized coarse-grain models. In addition, within the DCA procedure there are a number of subtle arbitrary choices that might significantly affect the result, including the choice

of a suitable sequence-alignment algorithm and the identification of the correct threshold for contact prediction. Due to the relatively weak degree of coevolution in RNA [Dutheil *et al.*, 2010] as compared, for instance, to proteins, a careful analysis of the different methods that can be used to quantify it is particularly urgent.

In this paper, we report a systematic analysis of the performance of DCA methods for 17 riboswitch families chosen among those for which at least one high-resolution crystallographic structure is available. Riboswitches are ubiquitous in bacteria and thus show a significant degree of sequence heterogeneity within each family. A stochastic procedure based on Boltzmann learning for solving exactly the DCA inverse problem is introduced and compared with the mean-field solution and the pseudo-likelihood maximization approach, as well as with mutual information and R-scape method. A rigorous cross-validation procedure that allows to find a portable threshold to identify predicted contacts is also introduced. Whereas Boltzmann learning is usually considered as a numerically unfeasible procedure in DCA, we here show that it can be effectively used to infer parameters that reproduce correctly the statistical properties of the analyzed alignments and that correlate with experimental contacts better than those predicted using alternative approximations.

2 Materials and Methods

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* [Thompson *et al.*, 1994] and *Infernal* [Nawrocki and Eddy, 2013].

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 [De Leonardis *et al.*, 2015] 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).

The idea of DCA is to construct a probability distribution that can generate an ensemble of sequences compatible with the available ones. The probability distribution that maximizes the entropy, thus minimizing the amount of information, among those compatible with the empirical frequencies has the following form:

$$P(\{\sigma\}) = \frac{1}{Z} \exp\left(\sum_i h_i(\sigma_i) + \sum_{ij} J_{ij}(\sigma_i, \sigma_j)\right) \tag{3}$$

The frequencies of nucleotides and co-occurence of nucleotides corresponding to the model, $f_i(\sigma)$ and $f_{ij}(\sigma, \tau)$, should coincide with the frequencies observed in the MSA, $F_i(\sigma)$ and $F_{ij}(\sigma, \tau)$. The coupling matrix J only contains direct interactions and is free of indirect correlations, hence the name direct couplings. In the following we discuss the technical details associated to the Boltzmann learning procedure introduced here. For a more general introduction to DCA and to the other methods to perform DCA (mean field and pseudolikelihood) see Supplementary Methods.

2.1 Maximum likelihood and Boltzmann learning

Given a set of independent equilibrium configurations $\{\sigma^b\}_{b=1}^B$ of the model (Eq. 3) such that $P(\sigma) = \prod_{b=1}^B P(\sigma^b)$, a statistical approach to infer parameters $\{h, J\}$ is to let them maximize the likelihood, i.e. the probability of generating the data set for a given set of parameters [Ekeberg *et al.*, 2013]. This can be equivalently done minimizing the negative log likelihood divided by the effective number of sequences:

$$l = -\frac{1}{B} \sum_{b=1}^B \log P(\sigma^b) \tag{4}$$

Minimizing l with respect to local fields h_i gives

$$\begin{aligned}
\frac{\partial l}{\partial h_i(\sigma)} &= \frac{1}{B} \sum_{b=1}^B \left(\frac{\partial \log Z}{\partial h_i(\sigma)} - \delta(\sigma_i^b, \sigma)\right) = \\
&= \frac{1}{B} \sum_{b=1}^B \left(f_i(\sigma) - \delta(\sigma_i^b, \sigma)\right) \\
&= f_i(\sigma) - F_i(\sigma) = 0
\end{aligned} \tag{5}$$

Similarly, minimizing l with respect to the couplings gives

$$\frac{\partial l}{\partial J_{ij}(\sigma, \tau)} = f_{ij}(\sigma, \tau) - F_{ij}(\sigma, \tau) = 0 \tag{6}$$

4

These equations show that the model with the maximum likelihood to reproduce the sequences observed in the MSA is the one with frequencies identical to those observed in the MSA.

A possible strategy to minimize l is *gradient descent*, that is an iterative algorithm in which parameters are adjusted by forcing them to follow the opposite direction of the function gradient [Ackley *et al.*, 1987, Sutto *et al.*, 2015, Barrat-Charlaix *et al.*, 2016, Haldane *et al.*, 2016, Figliuzzi *et al.*, 2018]. The value of the parameters $\boldsymbol{\theta}$ at iteration $k+1$ can be obtained from the value of $\boldsymbol{\theta}$ at the iteration k as

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta l(\theta) = \theta_t - \eta_t(f(\theta) - F) \tag{7}$$

where η_t is the *learning rate* and t is the fictitious time, corresponding to the iteration number. Calculation of the gradient requires evaluation of an average over all the possible sequences. This average can be computed with a Metropolis-Hastings algorithm in sequence space, but might be very expensive due to the large size of the sequence space. In addition, the average should be recomputed at every iteration. We here propose to use the instantaneous value of $\delta(\sigma_i, \sigma)$, where σ_i is the identity of the nucleotide at position i in the simulated sequence, as an unbiased estimator of $f_i(\sigma)$ in order to update the parameters more frequently, resulting in a *stochastic gradient descent* procedure that forces the system to sample the posterior distribution. The procedure can be easily parallelized, so that at each iteration the new set $\boldsymbol{\theta}$ is an average of the updated parameters over all processes. We here used 20 simultaneous simulations initialized from 20 random sequences chosen in the MSA. Once parameters are stably fluctuating around a given value, their optimal value can be estimated by taking a time average of $\boldsymbol{\theta}$ over a suitable time window [Cesari *et al.*, 2018]. At that point, a new simulation could be performed using the time-averaged parameters. Such a simulation can be used to rigorously validate the obtained parameters.

We here choose a learning rate η_t in the class *search then converge* [Darken and Moody, 1990]:

$$\eta_t = \frac{\alpha}{1 + \frac{t}{\tau_S}} \tag{8}$$

This function is close to α for small t ("search phase"). For $t \gg \tau_S$ the function decreases as $1/t$ ("converge phase"). Since it is based on Boltzmann sampling of the sequence space, we refer to this procedure as *Boltzmann learning*. The exact algorithm is described in Supporting Information and the employed C code is available at https://github.com/bussilab/bl-dca. We notice that in the algorithm implemented here, at variance with others proposed before [Sutto *et al.*, 2015, Figliuzzi *et al.*, 2018], the Lagrangian multipliers are evolved every few Monte Carlo iterations using istantaneous values rather than averages obtained from converged trajectories. In Ref. [Figliuzzi *et al.*, 2018] a change of variables of the model parameters was proposed to make the minimization easier. This idea might be beneficial also in our algorithm.

2.2 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:

$$RMSD = \sqrt{< (f_{ij}(\sigma_i, \tau_j) - F_{ij}(\sigma_i, \tau_j))^2 >_{\{ij\},\{\sigma_i, \tau_j\}}} \tag{9}$$

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.

2.3 Validation of the predicted contacts

We perform this analysis on sequences of a number of riboswitches families classified in the Rfam database [Nawrocki *et al.*, 2014]. 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 Supporting Information (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 [Lu *et al.*, 2015], that takes into account all hydrogen bonds and classify base pairs according to the Westhof-Leontis nomenclature [Leontis and Westhof, 2001]. 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 [Leontis *et al.*, 2002, Stombaugh *et al.*, 2009]. 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) [Dunn *et al.*, 2007].

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:

$$sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$precision = \frac{TP}{TP + FP} \tag{11}$$

The Matthews correlation coefficient (MCC) can be defined as the geometric average of sensitivity and precision [Matthews, 1975, Gorodkin *et al.*, 2001]

$$MCC = \sqrt{sensitivity \cdot precision} \tag{12}$$

and is equivalent to the interaction network fidelity [Parisien *et al.*, 2009]. 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:

$$\bar{S} = \arg\max_S \left(\prod_\mu^{N_s} MCC_\mu(S) \right)^{\frac{1}{N_s}} \tag{13}$$

3   Results

We here report an extensive assessment of the capability of covariance-based methods to infer contacts in RNA systems. In particular, we focus on direct-coupling-analysis (DCA) methods, which require the coupling constants of a Potts model that reproduces empirical covariations to be estimated. We thus first assess the capability of different methods to infer correct couplings. We then compare the high-score contacts with those observed in high resolution crystallographic structures in order to assess the capability of these methods to enhance RNA structure prediction.

The majority of the results presented in the main text are obtained using the Infernal MSA, and equivalent results obtained using ClustalW alignments are presented in Supporting Information (Figures 6 to 11). Similarly, the effect of not applying the average-product correction (APC) is reported in Supporting Information (Table 4).

7

3.1 Capability of the inferred couplings to reproduce frequencies

As a first step, we compared the absolute capability of the discussed methods to infer a Potts model compatible with the frequencies observed in the MSA. As shown in Figure 1, the Boltzmann learning procedure is capable to infer a Potts model that generates sequences with the correct frequencies. The two displayed families are those where the model frequencies agree best (PDB: 3F2Q) or worst (PDB: 3IRW) with the empirical ones. For 3IRW there are still visible mismatches, whereas for 3F2Q the modeled and empirical frequencies are virtually identical. On the other hand, the couplings inferred using the pseudo-likelihood or the mean-field approximation do not reproduce correctly the empirical frequencies. This is expected, since the mean-field approximation is not meant to be precise but rather a quick method to compute an approximation to the real couplings. Particularly striking is the case of the pseudo-likelihood for 3IRW, where there is no apparent correlation between the modeled and the empirical frequencies.

In Figure 2 we report the RMSD between the empirical and model frequencies for all the investigated families. The learning parameters for the Boltzmann learning simulation were chosen in order to minimize the RMSD value reported here ($\alpha = 0.01$, $\tau_S = 1000$). A negative control is performed comparing empirical frequencies with the ones calculated on random sequences ($f_{ij} = 1/25$), and a positive control computing the statistical error due to finite size of the alignment, in order to set a reference for RMSD values. In addition, we compare empirical frequencies against the ones calculated on the 20 MSA sequences initializing the parallelized Boltzmann learning simulation, so to ensure that frequencies are not reproduced thanks to the statistics resulting from the initial sequences but rather thanks to a correct choice of the coupling parameters. For all families, the resulting RMSD obtained with the Boltzmann learning couplings is lower than the one obtained using the 20 sequences from the MSA, indicating that the chosen couplings are shifting the distribution towards the empirical one. In some cases the RMSD reaches the statistical error expected with a finite number of sequences (positive control). Whereas this is expected since the Boltzmann learning procedure is exaclty trained to reproduce these frequencies, it is not obvious that this result can be achieved in a feasible computational time scale. On the contrary, both the pseudo-likelihood and mean-field approximation present an RMSD systematically larger than the one obtained from 20 sequences from the MSA. This indicates that the couplings inferred using these approximated methods are not leading to a Potts model that reproduces the experimental frequencies.

We notice that the adopted pseudo-likelihood implementation employs a regularization term in order to improve predictions when the number of sequences is low. This term is usually tuned in order to improve the rank of true contacts and not the frequencies reported here. We thus tested parameters obtained using a lower regularization term obtaining similar results (Supporting Information, Figure 12). Given that pseudo-likelihood is known to con-
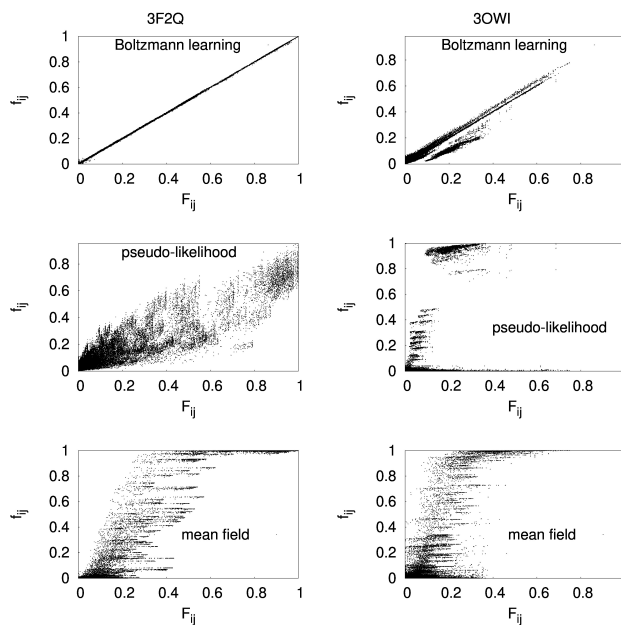
Figure 1: FMN riboswitch (PDB code 3F2Q) (1) and c-di-GMP-I (PDB code 3IRW) (1). Comparison between modeled $f_{ij}(\sigma, \tau)$ and empirical $F_{ij}(\sigma, \tau)$ frequencies $\forall i, j, \sigma, \tau$, obtained from DCA via Boltzmann learning, mean-field approximation and pseudo-likelihood maximization.

verge to the exact value in the limit of an infinite number of sequences (see, e.g., [Arnold and Strauss, 1991] and [Ravikumar *et al.*, 2010]), this discrepancy should be attributed to the typical size of the used alignments. We also notice that in multiple cases the frequencies obtained using couplings inferred with pseudo-likelihood tend to be larger than the empirical ones. Since the RMSD is highly sensitive to large deviations, this can cause some of the systems to be in less agreement with natural sequences than the employed negative control, which instead consists by construction of homogeneous frequencies. Qualitatively, the deviation observed here is similar to the one reported for protein systems in [Figliuzzi *et al.*, 2018].

3.2 Validation of contact prediction

As we have seen so far, Boltzmann learning is the only procedure capable to infer correct couplings. However, this does not necessarily imply that it is also the method capable of most correct contact predictions. Indeed, one cannot give for granted that the exact parameters of the Potts model are correlated with structural contacts. We here validate the predictions against a set of crystallographic structures by computing the MCC between the predicted and
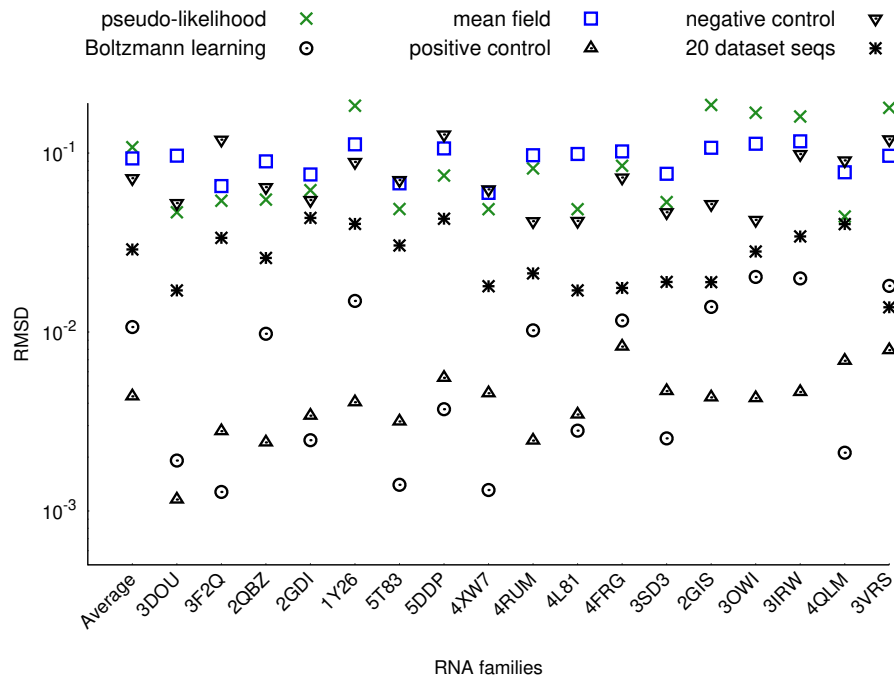
9

Figure 2: Capability of the inferred couplings to reproduce frequencies using different methods (Boltzmann learning, pseudo-likelihood and mean-field DCA). 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. We report a positive control (statistical error due to the finite number of sequence), a negative control (RMSD between empirical sequences and a random sequence) and the RMSD from the ensemble of the 20 sequences used as a starting point of Boltzmann learning simulations. Families are labeled using the PDB code of the representative crystallographic structure. Average RMSD is also reported.

empirical contacts. The general approach used to predict contacts from DCA is to extract the residue pairs with the highest couplings. Similarly, contacts can be predicted choosing pairs with the highest mutual information or the lowest E-value provided by R-scape. In order to fairly choose the threshold we adopted a cross-validation procedure: the $\overline{MCC}$ of each system is the one corresponding to a score cutoff $\overline{S}$ maximizing the average MCC (Eq. 13), calculated excluding that system. The choice of the threshold for covariance scores of the different models can be generalized to an independent data set, since the optimal threshold has a similar value for all systems (Supporting Information, Table 2 and 3). We also tested the more standard procedure of choosing as predictions a given fraction of the length N (Supporting Information, Table 11). For R-scape we used the recommended threshold corresponding to an E-value equal to 0.05.

As a negative control we show the MCC obtained assuming randomly chosen scores. In this case, the precision is equal to the number of native contacts (N_{native}) over the total number of possible contacts ($\frac{N(N-1)}{2}$) irrespectively of the chosen threshold, whereas the sensitivity is maximized when the threshold is chosen such that all the possible contacts are predicted and is equal to 1. The corresponding MCC is thus $\sqrt{\frac{2N_{native}}{N(N-1)}}$. Finally, we also computed the cross-validated MCC obtained with a thermodynamic model applied to the sequence associated to each crystallographic structure, by using as scoring the pairing probabilities computed with ViennaRNA [Mathews $et\ al.$, 2004, Lorenz $et\ al.$, 2011]. These results do not exploit the covariance information and are thus instructive to assess its importance.

Results of the cross-validation procedure for each system (Figure 3) indicate that direct coupling analysis outperforms mutual information and R-scape, and in particular Boltzmann learning performs the most accurate prediction. In addition, the results on individual families show that the choice of threshold covariance score is more consistent for Boltzmann learning when compared to pseudo-likelihood DCA. In order to quantify this effect we introduce a transferability index $\phi = \frac{1}{N_s} \sum_{\mu}^{N_s} \frac{\overline{MCC_\mu}}{MCC_\mu^{max}}$, which is the ratio between the cross-validated MCC for system μ ($\overline{MCC_\mu}$) described above and the maximum MCC that can be obtained by choosing the optimal threshold for each system MCC_μ^{max}, averaged over all systems. This value amounts to $\phi = 0.96$ for BL and to $\phi = 0.91$ for pseudo-likelihood DCA, suggesting that for the latter case the accuracy of contact prediction is more sensible to the choice of the cutoff, which is less easily transferable between different systems. Results for mean field DCA and mutual information are $\phi = 0.95$ and $\phi = 0.92$, respectively. Finally, we notice that all the DCA methods perform better than thermodynamic models alone (Supporting Information, Table 8).

3.3   Influence of alignment method

We then used the two most accurate covariance methods (Boltzmann learning and pseudo-likelihood DCA) to assess the influence of the alignment method.
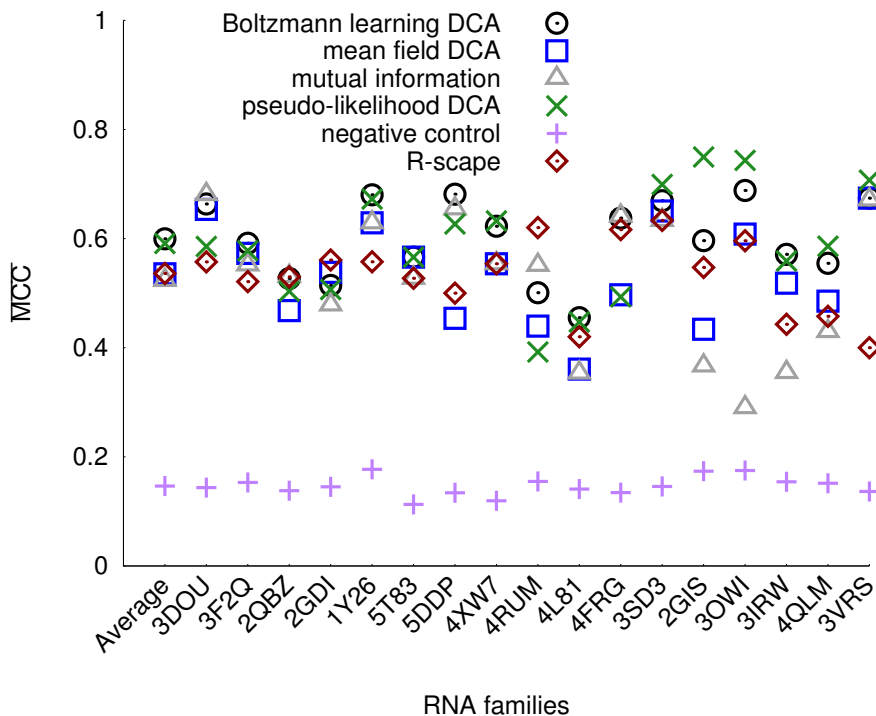
Figure 3: $\overline{MCC}$ of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information, and R-scape for 17 RNA families at the threshold obtained through cross-validation procedure. The recommended threshold 0.05 was used for R-scape. Families are labeled using the PDB code of the representative crystallographic structure. Average $\overline{MCC}$ is also reported. Alignments are performed with *Infernal*.

In particular, we considered the MSA methods implemented in *ClustalW* and *Infernal* packages. The average MCC over all RNA families when varying threshold S is systematically higher if sequences are aligned with *Infernal* rather than *ClustalW* (Figure 4). We attribute this improvement in the quality of prediction performance to the use of consensus secondary structure in *Infernal* [Nawrocki and Eddy, 2013]. The discrepancy between the accuracies of contact prediction using two different alignment methods enlightens the necessity of efficient tools to improve covariance analysis input quality. Interestingly, the threshold score $\overline{S}$ maximizing the MCC is the same for the Boltzmann learning performed on the two different MSAs. This suggests the robustness of the adopted procedure to assess the optimal threshold score (Eq. 13), again enlightening a greater consistency in its choice for the Boltzmann learning with respect to pseudo-likelihood maximization framework. Given its better performance, the *Infernal* MSA method is used in the rest of the main text.
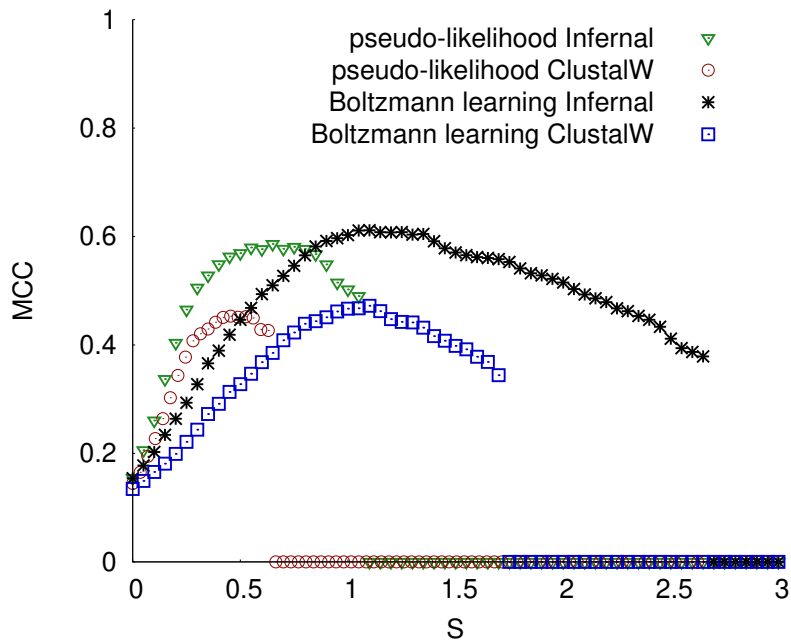
12

Figure 4: Geometric-average MCC as a function of threshold scores S for Boltzmann learning and pseudo-likelihood DCA. MSAs are performed with *ClustalW* and *Infernal*, as indicated. The sharp decrease after some method-dependent value of S is due to the fact that when the threshold is too large the number of correctly predicted contacts in at least one of the 17 investigated systems drops to zero.

3.4 Precision and sensitivity

In order to better quantify the capability of the investigated methods to provide useful information about contacts, we independently monitor sensitivity and precision for each RNA family at cross-validation thresholds. The average sensitivity values are around 0.3–0.4, indicating that approximately one third of the contacts present in the native structure can be predicted with these procedures (Supporting Information, Figure 1). Qualitatively, it appears that correctly predicted contacts are scattered along the sequences. The average precision instead ranges between 0.7 and 0.9, indicating that the number of falsely predicted contacts is rather small (Supporting Information, Figure 2). The Boltzmann learning and pseudo-likelihood DCA report higher sensitivity and precision than the other methods. R-scape presents a higher sensitivity when compared with mutual information and a similar precision. We notice that R-scape results reported here are obtained using the recommended threshold (E-value<0.05). Results obtained choosing the E-value that maximizes the MCC are reported in Supporting Information (Tables 9 and 10). In order to assess the capability of these methods to probe RNA tertiary structure we also report the sensitivity value restricted to secondary contacts, obtained considering only base pairs contained in stems, and the number of true positive tertiary contacts, with results similar to those reported above (Supporting Information, Figures 3 and 4). A contact is thus here considered as tertiary irrespectively of which edges are shared between nucleobases, and might even be an isolated WC pair. In general, DCA is able to identify not only cWW [Leontis and Westhof, 2001] pairs, where covariance is mostly associated to canonical pairs (GC, AU, and GU), but also a number of non-canonical pairs (see Table 6). When looking at the absolute number of incorrect predictions the Boltzmann learning DCA provides the smallest average number (Supporting Information, Figure 5). In particular, pseudo-likelihood DCA reports a very large number of false positives for a few systems. Also in this case, this is a consequence of the poor transferability of the cutoff for contact prediction in pseudo-likelihood DCA. A more careful eye on incorrect predictions reveals that $\approx 50\%$ of false positives predicted by all DCA methods are actually stacking interactions not included in the true-positive list since we only considered base pairings in reference native structures (Supporting Information, Table 7). In addition, couplings in consecutive nucleotides might be affected by a bias in the dinucleotide distribution.

3.5 Typical contact predictions

It is instructive to visualize which specific contacts are correctly predicted and which ones are not for individual systems. We first discuss the predictions on the systems where Boltzmann learning and pseudo-likelihood DCA result in the highest MCC (glycine riboswitch, PDB 3OWI, and SAM riboswitch, PDB 2GIS, respectively). In the glycine riboswitch, Figure 5, we see that the two methods give comparable results. All the four native stems are predicted, although pseudo-likelihood DCA predicts a slightly larger number of correct pairs.

Also a non-stem WC contact is identified. In the SAM riboswitch, Figure 6, we see that the pseudo-likelihood DCA predicts a significantly larger number of correct contacts. Notably, both methods are capable to identify contacts in a pseudoknotted helix between residues 25–28 and residues 68–65. These examples show that in the best cases these methods allow full helices to be identified accompanied by a small number of critical tertiary contacts. It is also useful to consider the cases resulting in the lowest MCC (SAM-I/IV riboswitch, PDB 4L81, for Boltzmann learning and NiCo riboswitch, PDB 4RUM, for pseudo-likelihood DCA). In the SAM-I/IV riboswitch the two methods give comparable results, and only a limited number of secondary contacts are correctly predicted (Figure 7). The stem between position 10 and position 20 shows a number of false positives. In this case, a helix with a register shifted by one nucleotide is suggested by the both DCA predictions. In more detail, we do not expect the alternative register to have a significant population in solution, since it would be capped by a AGAC tetraloop, whereas the reference crystal structure displays a common GAGA tetraloop. We interpret both sets of false positives as errors in the MSA. Indeed, especially with sequences consisting of consecutive identical nucleotides, one cannot assume the alignment procedure to correctly place gaps in the MSA. As a consequence, the reference structure for which the PDB is available might be misaligned with the majority of the homologous sequences in the MSA, resulting in predicted contacts shifted by one position upstream or downstream. Remarkably, many WC pairs close to the binding site of the riboswitch are predicted (G10/C21, G22/U50 and G23/C49; ligand directly interacts with nucleotides C7, A25 and U47). In the NiCo riboswitch, Figure 8, pseudo-likelihood DCA only predicts 6 correct helical contacts, whereas Boltzmann learning DCA is capable to predict a number of contacts in the helices, even though resulting in several false positives.

3.6 Validation on non-riboswitch systems

We further validated the whole procedure by considering 4 additional families including ribosomial RNA subunits, transfer RNA (tRNA), and a purely eukaryotic spliceosomal RNA. All the parameters of the Boltzmann learning simulations were chosen identical to those used for the riboswitch families. The threshold used to convert scores into predictions was taken as 1.06, which is the one that maximizes the MCC on the 17 riboswitch families. Results are reported in Supporting Information, Table 12 and are slightly worse than those obtained for riboswitch families, with the exception of tRNA.

4 Discussion

We here report a systematic assessment of RNA contact prediction based on aligned homologous sequences using mutual information analysis, R-scape, and DCA. When compared to previous works [De Leonardis *et al.*, 2015, Weinreb *et al.*, 2016, Wang *et al.*, 2017], our analysis focuses on the DCA cal-
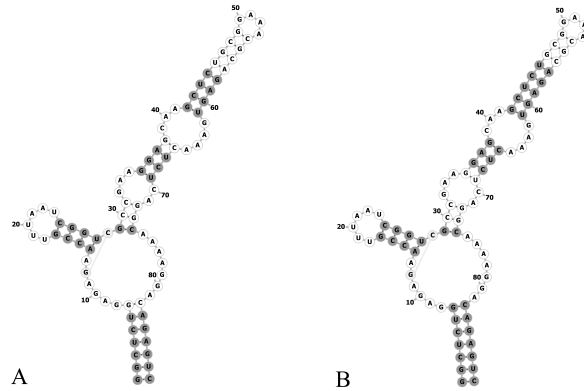
Figure 5: Glycine riboswitch (PDB code 3OWI) most accurate Boltzmann learning prediction (5A) and respective pseudo-likelihood prediction (5B). Correctly predicted contacts in secondary structure are shown in dark-grey. Correctly predicted tertiary contacts are shown in light-grey. We notice that G12/C28 pair is here labeled as tertiary since it corresponds to a isolated Watson-Crick pair in the reference structure.
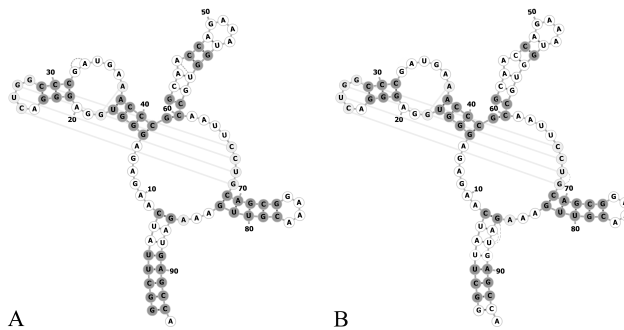

Figure 6: SAM riboswitch (PDB code 2GIS), best accurate pseudo-likelihood prediction (6A) and respective Boltzmann learning prediction (6B). Correctly predicted contacts in secondary structure are shown in dark-grey. Correctly predicted tertiary contacts are shown in light-grey. False positives are shown with dashed lines.
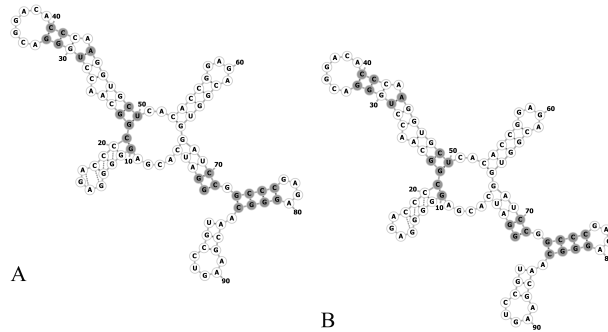
Figure 7: SAM-I/IV riboswitch (PDB code 4L81), least accurate Boltzmann learning prediction (7A) and respective pseudo-likelihood prediction (7B). Correctly predicted contacts in secondary structure are shown in dark-grey. Correctly predicted tertiary contacts are shown in light-grey. False positives are shown with dashed lines.
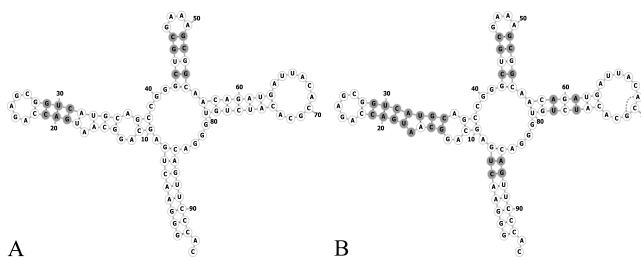

Figure 8: NiCo riboswitch (PDB code 4RUM), least accurate pseudo-likelihood prediction (8A) and respective Boltzmann learning prediction (8B). Correctly predicted contacts in secondary structure are shown in dark-grey. Correctly predicted tertiary contacts are shown in light-grey. False positives are shown with dashed lines.

culation and does not convert the resulting couplings into a structural model. The capability of various DCA-based methods to reproduce empirical frequencies from the MSA is evaluated. Native contacts in a set of reference structures are carefully annotated and compared with the predicted ones, in order to quantify the fraction of correctly predicted contacts (precision) and the fraction of predicted native contacts (sensitivity). In particular, since coevolution in RNA is expected to be related to isostericity [Leontis *et al.*, 2002, Stombaugh *et al.*, 2009], we only considered base pairing and excluded other base-backbone or backbone-backbone contacts.

Our results show that approximately 40% of the total native contacts can be predicted by this procedure. A large fraction of the predicted contacts are secondary structure contacts or pseudoknotted helices. However, in most of the analyzed structures, at least one tertiary contact is correctly predicted. In addition, the number of false positives is very small ($\approx 10\%$ of the predicted contacts). In many cases, false positives are just labeled so by our decision to exclude stacking interactions from the true contacts. In other cases, false positives are a consequence of an erroneous alignment of some of the sequences. Some false positives are genuinely caused by numerical noises or by the assumptions behind the Potts model. We notice that in principle the detrimental effect of false positives on the accuracy of structure prediction might be mitigated by using approaches where contacts that are not compatible with the predicted structure are discarded iteratively [Weinreb *et al.*, 2016]. As a general consideration, it must be kept in mind that strong couplings as predicted by DCA are a signature of co-evolutionary pressure but not necessarily of spatial proximity. For instance, functionally related elements that are far from each other in space might exhibit coevolution. Relationships of this kind could in principle decrease the precision of the method in predicting contacts. In principle, highly conserved residues carry a limited amount of information and could thus reduce the sensitivity of the method, although in practice we never observed a very high conservation in the analyzed bacterial sequences. Eukariotic sequences might be more sensible to this issue, as it can be seen by the lower performance of the method when applied to spliceosomal RNA.

Importantly, we developed a rigorous manner to establish a threshold for contact prediction. In particular, once a figure of merit capable to take into account both the method precision and sensitivity has been defined, an optimal threshold can be found on a specific training set. We here used the Mathews correlation coefficient, that corresponds to the interaction network fidelity [Parisien *et al.*, 2009] widely used in the RNA structure-prediction community [Miao *et al.*, 2017]. The resulting thresholds are different depending on the used method (mutual information vs the tested DCA methods), but are transferable across different RNA families as illustrated by our cross-validation analysis.

It is important to observe that RNA molecules often display dynamics (i.e. coexistence of multiple structures) related to function, and that perhaps riboswitches are the paradigmatic example where multiple structures are required for function. For instance, some of the false positives might correspond to true contacts in an alternative, biologically functional structure (e.g., on and off state

of the riboswitch). This fact might affect the results of the comparison reported here. Nevertheless, we believe that high resolution X-ray structures still represent the best proxy for the correct solution structure and as such they should be used for a critical assessment. Without having an experimentally determined ensemble, it appears difficult to assume that the observed false positives are, by chance, important contacts in alternative structures.

A crucial finding is that the here introduced stochastic solution of the inverse problem (Boltzmann learning) is feasible on these systems and outperforms the other DCA approaches. The resulting Potts models were shown to reproduce correctly the empirical frequencies from the MSA. Whereas the fact that the mean-field approach provides an approximate solution is well-known [Nguyen *et al.*, 2017, Cocco *et al.*, 2018], no such comparison has been reported on RNA DCA yet. In addition, we show that, although it is supposed to be capable to infer correct couplings at least in the limit of a large number of sequences, also the pseudo-likelihood approximation is not capable to reproduce the correct frequencies with the employed datasets. This fact was recently observed for protein systems [Figliuzzi *et al.*, 2018], where it was also observed that in spite of this disagreement the contact predictions obtained with the pseudo-likelihood approximation are of quality similar to those obtained with Boltzmann learning DCA.

The overall improvement in the accuracy of the predictions, as measured by the MCC, when passing from state-of-the-art pseudo-likelihood DCA to Boltzmann-learning DCA is comparable to the one observed when passing from mean-field DCA to pseudo-likelihood DCA, which has been already shown to improve the quality of 3D structure prediction [De Leonardis *et al.*, 2015]. It is worth saying that the extra cost of the Boltzmann learning procedure is significant if one wants to characterize a large number of families. The required times for all the tested covariance methods scale roughly as the number of nucleotides squared and are listed in table 1 for the largest and smallest molecules in the data set. If we also include the cost of a later 3D structure prediction and refinement, we consider the extra cost of Boltzmann learning to be absolutely worth. We believe that the fast Boltzmann learning procedure introduced here based on a stochastic gradient descent could be fruitfully used in protein systems as well. We also notice that the stochastic procedure used here is closely related to similar techniques used in the molecular dynamics community in order to enforce preassigned distributions in the generation of molecular structures [Cesari *et al.*, 2018, Valsson and Parrinello, 2014, White and Voth, 2014, Cesari *et al.*, 2016]. We chose here to use the simplest possible optimization algorithm, but more advanced procedures might make the Boltzmann learning approach even faster.

We also tested the state-of-the-art pseudo-likelihood maximization approach, which is faster than the Boltzmann learning approach but, on the tested dataset, provides results of slightly inferior quality. Interestingly, the relatively good contact predictions obtained using pseudo-likelihood DCA are not paralleled by correct frequencies in the reconstructed Potts model. Similar results were obtained decreasing the regularization term usually employed in pseudo-likelihood

Table 1: Computational time for the smallest and largest system investigated. Machine hardware architecture: Intel E5-2620, 12 physical cores. Operating system: GNU/Linux. Mutual information, MF-DCA, and BL-DCA predictions were done using in house code. R-scape predictions were done using R-scape 1.2.3. PL-DCA were done using plmDCA_asymmetric_v2 code available on GitHub.

Method	3DOU (largest)	3VRS (smallest)
Boltzmann learning DCA	220 min	20 min
Pseudo-likelihood DCA	3 min	30 sec
R-scape	33 sec	9 sec
Mean field DCA	22 sec	4 sec
Mutual Information	15 sec	3 sec

DCA. This effect is likely due to the finite number of available sequences. A more important practical issue is that the optimal threshold used for contact predictions resulted less transferable across different families in pseudo-likelihood DCA when compared with Boltzmann learning DCA. This suggests that choosing a cutoff that can single out true contacts might be more difficult in this method.

The impact on contact prediction of other sometime overlooked choices (reweighting and APC correction) has also been assessed. Our results show that these choices lead to negligible or minor improvements to all the methods.

Finally, we show that the alignment procedure used to prepare the MSA has a significant impact on the accuracy of the prediction. Interestingly, the *Infernal* algorithm, that is based on a previous prediction of the secondary structure, performs significantly better than the *ClustalW* algorithm. Whereas this effect is somewhat expected, we are not aware of similar assessments done on DCA methods. We observe that the couplings obtained with the present approach might be used to further refine the multiple sequence alignments.

In conclusion, the direct coupling analysis method was assessed on a number of RNA families. We found that, in spite of the intrinsic approximations, this procedure is able to reliably predict a number of contacts in RNA molecules with known three-dimensional structure. Among the tested methods, the Boltzmann learning approach is the one that allows to simultaneously maximize accuracy and precision. In perspective, we foresee the possibility to explicitly use information about the isosteric RNA families [Leontis *et al.*, 2002, Stombaugh *et al.*, 2009] or include three-body terms [Schmidt and Hamacher, 2017] in order to further improve the accuracy of the predictions. Ultimately, we suggest the direct coupling analysis performed through the Boltzmann learning as the best available tool to enhance RNA

structure prediction for systems of up to a few hundreds nucleotides, taking advantage of only homologous sequences information.

5 Acknowledgements

References

[Ackley *et al.*, 1987] Ackley,D.H., Hinton,G.E. and Sejnowski,T.J. (1987) A learning algorithm for Boltzmann machines. In *Readings in Computer Vision*. Elsevier pp. 522–533.

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

[Barrat-Charlaix *et al.*, 2016] Barrat-Charlaix,P., Figliuzzi,M. and Weigt,M. (2016) Improving landscape inference by integrating heterogeneous data in the inverse ising problem. *Sci. Rep.,* **6**, 37812.

[Cesari *et al.*, 2016] Cesari,A., Gil-Ley,A. and Bussi,G. (2016) Combining simulations and solution experiments as a paradigm for RNA force field refinement. *J. Chem. Theory Comput.,* **12** (12), 6192–6200.

[Cesari *et al.*, 2018] Cesari,A., Reißer,S. and Bussi,G. (2018) Using the maximum entropy principle to combine simulations and solution experiments. *Computation,* **6** (1), 15.

[Cocco *et al.*, 2018] Cocco,S., Feinauer,C., Figliuzzi,M., Monasson,R. and Weigt,M. (2018) Inverse statistical physics of protein sequences: a key issues review. *Rep. Prog. Phys.,* **81** (3), 032601.

[Costa and Michel, 1997] Costa,M. and Michel,F. (1997) Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution. *EMBO J.,* **16** (11), 3289–3302.

[Darken and Moody, 1990] Darken,C. and Moody,J. (1990) Note on learning rate schedules for stochastic optimization. In *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3* NIPS-3 pp. 832–838 Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[De Leonardis *et al.*, 2015] De Leonardis,E., Lutz,B., Ratz,S., Cocco,S., Monasson,R., Schug,A. and Weigt,M. (2015) Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.,* **43** (21), 10444–10455.

[Doherty and Doudna, 2000] Doherty,E.A. and Doudna,J.A. (2000) Ribozyme structures and mechanisms. *Annu. Rev. Biochem.,* **69** (1), 597–615.

[Dunn *et al.*, 2007] Dunn,S.D., Wahl,L.M. and Gloor,G.B. (2007) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics,* **24** (3), 333–340.

[Dutheil *et al.*, 2010] Dutheil,J.Y., Jossinet,F. and Westhof,E. (2010) Base pairing constraints drive structural epistasis in ribosomal RNA sequences. *Mol. Biol. Evol.,* **27** (8), 1868–1876.

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

[Ekeberg *et al.*, 2013] Ekeberg,M., Lövkvist,C., Lan,Y., Weigt,M. and Aurell,E. (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys. Rev. E,* **87** (1), 012707.

[Figliuzzi *et al.*, 2018] Figliuzzi,M., Barrat-Charlaix,P. and Weigt,M. (2018) How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.,* **35** (4), 1018–1027.

[Gorodkin *et al.*, 2001] Gorodkin,J., Stricklin,S. and Stormo,G. (2001) Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.,* **29 10**, 2135–44.

[Haldane *et al.*, 2016] Haldane,A., Flynn,W.F., He,P., Vijayan,R. and Levy,R.M. (2016) Structural propensities of kinase family proteins from a potts model of residue co-variation. *Protein Science,* **25** (8), 1378–1384.

[Hon *et al.*, 2017] Hon,C.C., Ramilowski,J.A., Harshbarger,J., Bertin,N., Rackham,O.J., Gough,J., Denisenko,E., Schmeier,S., Poulsen,T.M., Severin,J. *et al.* (2017) An atlas of human long non-coding RNAs with accurate 5' ends. *Nature,* **543** (7644), 199.

[Jones *et al.*, 2011] Jones,D.T., Buchan,D.W., Cozzetto,D. and Pontil,M. (2011) Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics,* **28** (2), 184–190.

[Leontis *et al.*, 2002] Leontis,N.B., Stombaugh,J. and Westhof,E. (2002) The non-watson–crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.,* **30** (16), 3497–3531.

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

[Lorenz *et al.*, 2011] Lorenz,R., Bernhart,S.H., Zu Siederdissen,C.H., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA package 2.0. *Algorithms Mol. Biol.,* **6** (1), 26.

[Lu *et al.*, 2015] Lu,X.J., Bussemaker,H.J. and Olson,W.K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.,* **43** (21), e142.

[Marks *et al.*, 2011] Marks,D.S., Colwell,L.J., Sheridan,R., Hopf,T.A., Pagnani,A., Zecchina,R. and Sander,C. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE,* **6** (12), e28766.

[Mathews *et al.*, 2004] Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA,* **101** (19), 7287–7292.

[Mathews *et al.*, 2016] Mathews,D.H., Turner,D.H. and Watson,R.M. (2016) RNA secondary structure prediction. *Curr. Protoc. Nucleic Acid Chem.,* **67** (1), 11–2.

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

[Miao *et al.*, 2017] Miao,Z., Adamiak,R.W., Antczak,M., Batey,R.T., Becka,A.J., Biesiada,M., Boniecki,M.J., Bujnicki,J.M., Chen,S.J., Cheng,C.Y. *et al.* (2017) RNA-puzzles round iii: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA,* **23** (5), 655–672.

[Michel and Westhof, 1990] Michel,F. and Westhof,E. (1990) Modelling of the three-dimensional architecture of group i catalytic introns based on comparative sequence analysis. *J. Mol. Biol.,* **216** (3), 585–610.

[Morcos *et al.*, 2011] Morcos,F., Pagnani,A., Lunt,B., Bertolino,A., Marks,D.S., Sander,C., Zecchina,R., Onuchic,J.N., Hwa,T. and Weigt,M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.,* **108** (49), E1293–E1301.

[Morris and Mattick, 2014] Morris,K.V. and Mattick,J.S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.,* **15** (6), 423.

[Nawrocki *et al.*, 2014] Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2014) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.,* **43** (D1), D130–D137.

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

[Nguyen *et al.*, 2017] Nguyen,H.C., Zecchina,R. and Berg,J. (2017) Inverse statistical problems: from the inverse ising problem to data science. *Adv. Phys.,* **66** (3), 197–261.

[Pang *et al.*, 2005] Pang,P.S., Jankowsky,E., Wadley,L.M. and Pyle,A.M. (2005) Prediction of functional tertiary interactions and intermolecular interfaces from primary sequence data. *J. Exp. Zool. B Mol. Dev. Evol.,* **304** (1), 50–63.

[Parisien *et al.*, 2009] Parisien,M., Cruz,J.A., Westhof,É. and Major,F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA,* **15** (10), 1875–1885.

[Ravikumar *et al.*, 2010] Ravikumar,P., Wainwright,M.J. and Lafferty,J.D. (2010) High-dimensional ising model selection using l1-regularized logistic regression. *Ann. Stat.,* **38** (3), 1287–1319.

[Rinnenthal *et al.*, 2011] Rinnenthal,J., Buck,J., Ferner,J., Wacker,A., Fürtig,B. and Schwalbe,H. (2011) Mapping the landscape of RNA dynamics with nmr spectroscopy. *Acc. Chem. Res.,* **44** (12), 1292–1301.

[Rivas *et al.*, 2017] Rivas,E., Clements,J. and Eddy,S.R. (2017) A statistical test for conserved rna structure shows lack of evidence for structure in lncrnas. *Nature Methods,* **14** (1), 45.

[Schmidt and Hamacher, 2017] Schmidt,M. and Hamacher,K. (2017) Three-body interactions improve contact prediction within direct-coupling analysis. *Phys. Rev. E,* **96** (5), 052405.

[Serganov and Nudler, 2013] Serganov,A. and Nudler,E. (2013) A decade of riboswitches. *Cell,* **152** (1-2), 17–24.

[Smith *et al.*, 2013] Smith,M.A., Gesell,T., Stadler,P.F. and Mattick,J.S. (2013) Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.,* **41** (17), 8220–8236.

[Šponer *et al.*, 2018] Šponer,J., Bussi,G., Krepl,M., Banáš,P., Bottaro,S., Cunha,R.A., Gil-Ley,A., Pinamonti,G., Poblete,S., Jurečka,P. *et al.* (2018) RNA structural dynamics as captured by molecular simulations: a comprehensive overview. *Chem. Rev.,* **118** (8), 4177–4338.

[Stombaugh *et al.*, 2009] Stombaugh,J., Zirbel,C.L., Westhof,E. and Leontis,N.B. (2009) Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.,* **37** (7), 2294–2312.

[Sutto *et al.*, 2015] Sutto,L., Marsili,S., Valencia,A. and Gervasio,F.L. (2015) From residue coevolution to protein conformational ensembles and functional dynamics. *Proc. Natl. Acad. Sci. U. S. A.,* **112** (44), 13567–13572.

[Thompson *et al.*, 1994] Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) 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.

[Valsson and Parrinello, 2014] Valsson,O. and Parrinello,M. (2014) Variational approach to enhanced sampling and free energy calculations. *Phys. Rev. Lett.,* **113** (9), 090601.

[Wang *et al.*, 2017] Wang,J., Mao,K., Zhao,Y., Zeng,C., Xiang,J., Zhang,Y. and Xiao,Y. (2017) Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide–nucleotide interactions from direct coupling analysis. *Nucleic Acids Res.,* **45** (11), 6299–6309.

[Weeks, 2010] Weeks,K.M. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol.,* **20** (3), 295–304.

[Weinreb *et al.*, 2016] Weinreb,C., Riesselman,A.J., Ingraham,J.B., Gross,T., Sander,C. and Marks,D.S. (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell,* **165** (4), 963–975.

[Westhof, 2015] Westhof,E. (2015) Twenty years of RNA crystallography. *RNA,* **21** (4), 486–487.

[White and Voth, 2014] White,A.D. and Voth,G.A. (2014) Efficient and minimal method to bias molecular simulations with experimental data. *J. Chem. Theory Comput.,* **10** (8), 3023–3030.