INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

S.I.S.S.A.

DOCTORAL THESIS

Covariance models for RNA structure prediction

Author:

Francesca CUTURELLO

Supervisor:

Prof. Giovanni BUSSI

A thesis submitted in fulfillment of the requirements

for the degree of Doctor of Physics

in the

Molecular and Statistical Biophysics

October 11, 2019

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many different types of RNA molecules do not encode proteins, but rather play important roles in a wide range of cellular processes, including protein synthesis, gene regulation, protein transport and splicing (Morris and Mattick, 2014; Hon et al., 2017). The majority of RNAs conserve a particular three dimensional structure that is energetically favorable and inherent to their function (Smith et al., 2013), which is often conserved across evolutionary timescales. An RNA's structure is determined by intramolecular interactions between different elements in the polynucleotide chain, as well as by intermolecular interactions with other RNAs or proteins. Many of these interactions are hydrogen bonds formed by the base-pairing of two RNA nucleotides. Base-pairs commonly occur in groups, or stems, that form helices because they allow thermodynamically favorable stacking of the π bonds of the bases' aromatic rings. The set of stems in the RNA defines its secondary structure and can be inferred using thermodynamic models (Mathews, Turner, and Watson, 2016), often used in combination with chemical probing data (Weeks, 2010). There are other sets of base-pairs besides Watson-Cricks, called non-canonical (Stombaugh et al., 2009; Leontis and Westhof, 2001). The hierarchy of RNA structure is schematically shown in figure 1.1.

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 three dimensional structure of RNAs can be determined using x-ray crystallography (Westhof, 2015). Nuclear magnetic resonance (NMR) (Rinnenthal et al., 2011) has been used to solve the structure of short RNA motifs, but it's hard to use for

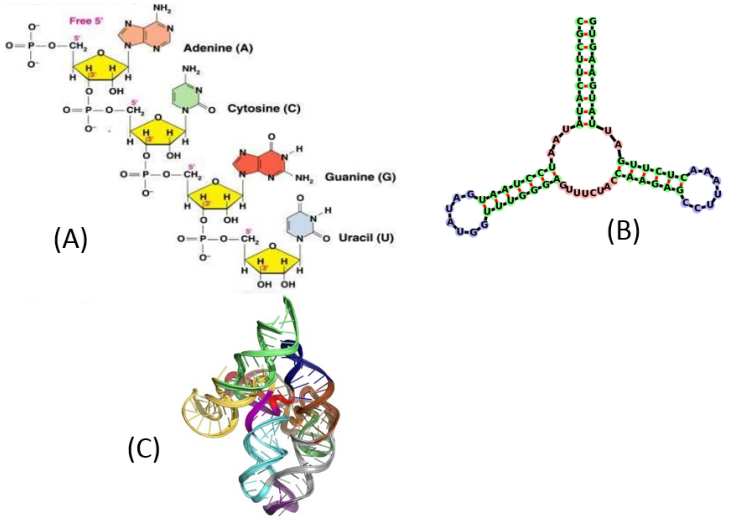


FIGURE 1.1: Hierarchy of RNA structure. Primary structure (A) is the nucleotides sequence. Secondary structure (B) is formed by consecutive canonical hydrogen bonded pairs (Watson-Crick pairing). Tertiary structure (C) is the three-dimensional shape to the molecule and includes non canonical pairs.

large RNAs. These techniques are expensive and time-consuming. Given the lower cost of sequencing techniques, an attractive alternative method for inferring RNA structure is based exclusively on sequence analysis. Since genomes of all organisms are evolutionarily related, many genes can be classified into families, composed of homologous elements. Families of homologous RNAs can be related by a phylogenetic tree rooted at the oldest ancestral sequence of the family: along each branch of the tree the sequences have evolved independently accumulating mutations, but, importantly, the function of the molecule has acted as an evolutionary constraint. Organisms whose sequence undergoes mutations that negatively affect function are less likely to survive to the next generation. Structural inference exploits the fact that nucleotides that form base-pairs in RNA structure tend to covary throughout evolution, in order to preserve the function of the molecule in cell processes (Nawrocki, 2009). A necessary first step for identifying covariation is alignment of homologous sequences. The goal of the alignment is to arrange sequences so that homologous nucleotides in each sequence occur in the same column of the alignment. The compensatory changes create patterns in multiple sequence alignments that are sometimes even recognizable by eye. The coevolution of bases in RNA fragments with

known structure has been investigated (Dutheil, Jossinet, and Westhof, 2010), observing strong correlations in Watson-Crick (WC) pairs and much weaker correlations in non-WC pairs.

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, Zecchina, and Berg, 2017; Cocco et al., 2018): strong direct couplings in the model indicate spatial proximity. The solution of the corresponding inverse model has been often obtained through 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.

The application of DCA to RNA structure prediction has so far been limited. 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 approach to infer contacts (Wang et al., 2017a). 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., 2017a). 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.

In this thesis, I 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. 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 (Eddy and Durbin, 1994) and R-scape method (Rivas, Clements, and Eddy, 2017). 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, it is shown 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.

In inverse problems, parameters of the model are inferred based on observations maximizing a likelihood function. This maximization is usually performed using regularization terms in order to avoid overfitting (Ekeberg et al., 2013; Marruzzo et al., 2017; Tyagi et al., 2016; Ravikumar, Wainwright, Lafferty, et al., 2010; Figliuzzi, Barrat-Charlaix, and Weigt, 2018) , especially when a limited number of training examples is available. In a Bayesian framework, the regularization term can be interpreted as a prior information on the parameters that is encoded in the process (Zhu, Chen, and Xing, 2014; Baldassi et al., 2014). In principle, any prior information about the parameters can be included in order to make their estimation more reliable and thus decrease both systematic errors, due to approximations in the model, and statistical errors, due to finite sample sizes. In the second part of the thesis, I show how to use a maximum posterior estimation procedure including a general prior in order to improve the solution of inverse Ising-like models. The procedure is first illustrated on a simple 10×10 Ising model. Then, the capability of DCA in predicting RNA contacts is enhanced by including in the parameters Boltzmann learning external information obtained from a secondary structure prediction algorithm. In DCA literature, l_2 regularization is usually adopted to avoid overfitting (Ekeberg et al., 2013; Figliuzzi, Barrat-Charlaix, and Weigt, 2018) and systematic error is tackled by post-processing in some advanced way the resulting couplings (Schug et al., 2009; Chen et al., 2011; Ma et al., 2015; Wang et al., 2017a; Wang et al., 2017b). The idea of helping the inference including external knowledge is found to significantly improve the accuracy of contact prediction.

In Chapter 2 of this Thesis I introduce the DCA formalism as well as the Boltzmann learning procedure that I developed and implemented. In Chapter 3 I show

the results obtained for RNA contact prediction on a database of 17 riboswitches. Finally, in Chapter 4 I illustrate the developed formalism to use informative priors in inferring couplings and I show how it can be adopted to significantly improve DCA predictions also when based on alignments that are suboptimal or of very limited size.

The material included in this Thesis has been partly adapted from two manuscripts (Cuturello, Tiana, and Bussi, 2019; Cuturello, Tiana, and Bussi, "Encoding prior information in inverse Ising-like models", in preparation). In addition, I contributed to another manuscript (Calonaci, Cuturello, Jones, Sattler and Bussi, in preparation), where data obtained through the models that I developed are used for structure prediction in combination with chemical probing experiments. The results of this manuscript are not included in this Thesis.

Chapter 2

Covariance models

The low cost of sequencing techniques lead to the accumulation of a vast number of sequence data for many homologous RNA families (Nawrocki et al., 2014). 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, Clements, and Eddy, 2017). In the protein community it has been widely and successfully employed direct coupling analysis (DCA) for residues contact prediction, which is a probabilistic model capable of generating the correlations observed in the analyzed alignments of homologous sequences (Morcos et al., 2011; Marks et al., 2011; Nguyen, Zecchina, and Berg, 2017; Cocco et al., 2018). It arises from maximum entropy principle and it consists of an inverse Potts problem that can be solved in first approximation through the mean-field approach (Morcos et al., 2011). Another possibility to further improve the level of approximation is to infer the solution through maximization of the conditional likelihood (or *pseudo-likelihood*), which is a consistent estimator of the full likelihood (Ekeberg et al., 2013). We propose a stochastic procedure (*Boltzmann learning*) for solving exactly the DCA inverse problem.

All mentioned covariance methods are described, including a number of subtle arbitrary choices that might significantly affect the result of the analysis, such as alignment methods, similarity based re-weighting of sequences, average product correction and Gauge choice. A description of the adopted measures for validation of results are also reported.

2.1 Alignment methods

The analysis of nucleotide co-evolution requires homologous RNA sequences to be aligned through a process named multiple sequence alignment (MSA). The results of any co-evolutionary analysis depends on this initial step. We here tested two commonly used MSA algorithms, namely those implemented in *ClustalW* (Thompson, Higgins, and Gibson, 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:

$$F_i(\sigma) = \frac{1}{B} \sum_{b=1}^{B} \delta(\sigma_i^b, \sigma) \tag{2.1}$$

$$F_{ij}(\sigma, \tau) = \frac{1}{B} \sum_{b=1}^{B} \delta(\sigma_i^b, \sigma)\delta(\sigma_j^b, \tau) \tag{2.2}$$

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.

2.1.1 *Infernal* alignment

Infernal is a tool for building consensus RNA secondary structure profiles called covariance models (CMs), and uses them to search nucleic acid sequence databases for homologous RNAs, or to create structure-based multiple sequence alignments. CMs are a special case of stochastic context-free grammars providing a statistical framework for combining sequence and secondary structure conservation information in a single consistent scoring system, designed for modeling RNA consensus sequence and structure. It relies on the typical conservation of secondary structure in RNA families, which are easier to asses than the full tertiary structure and thus are often available. We consider the use of secondary structure information in the alignment procedure as a strong bias in the input of covariance analysis driven by a previous

knowledge of the molecular structure. Even though the predictions from covariance models on these kind of alignments are very accurate, this tool can't be employed for future blind contact predictions on families for which no structure information is available.

2.1.2 *ClustalW* alignment

ClustalW is a necessary tool for contact prediction on RNA families lacking an assessed experimental structure. It uses a progressive alignment method: it aligns the most similar sequences first, then it progressively aligns more distant groups of sequences until a global alignment is created. *ClustalW* is a matrix-based algorithm, since the first step consists in computing a distance matrix between each pair of sequences, known as pairwise sequence alignment. Then, a neighbor-joining method is adopted to build the tree. *ClustalW* performs well when the data set contains sequences with varied degrees of divergence because the guide tree is less sensitive to noise in this case.

2.2 Re-weighting

The inference of the Potts parameters relies on the assumption that samples of sequences are independently generated from the model, which is not true for biological sequences. In databases there are many RNA sequences from related species which did not have enough time to reach statistical independence while evolving. Moreover, the selection of species to be sequenced is dictated by human interest. In order to reduce the effect of possible biases on the sampling of RNA sequences in databases, a common heuristic approach is sequence re-weighting. Two sequences are considered similar if the fraction of positions with coincident nucleotides (*similarity*) is larger than a given similarity threshold x:

$$n_b = |s \in \{1, ..., B\} : similarity(\sigma^s, \sigma^b) > x| \tag{2.3}$$

The inverse of n_b, $\omega_b = \frac{1}{n_b}$, gives a weight for the sequence contribution to frequencies and the effective number of sequences in the alignment is given by:

$$B_{eff} = \sum_{b=1}^{B} \omega_b \tag{2.4}$$

The effect of re-weighting is to make the sequence density homogeneous in sequences space.

2.3 Mutual information

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

$$MI_{ij} = \sum_{\sigma_i,\tau_j} F_{ij}(\sigma_i,\tau_j) \ln \frac{F_{ij}(\sigma_i,\tau_j)}{F_i(\sigma_i)F_j(\tau_j)} \equiv S_{ij} \tag{2.5}$$

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 (Eddy and Durbin, 1994) and its capability to predict contacts in RNA has been reported to be surpassed by DCA-based methods (De Leonardis et al., 2015).

2.4 R-scape

R-scape (RNA Significant Covariation Above Phylogenetic Expectation) is a software associating E-values to the pairs showing a significant covariation pattern in a MSA. E-values are determined using a null model of covariation entirely due to phylogeny. Covariation scores are calculated using the default G-test measure implemented in the package rscape_v1.2.3 at `http://eddylab.org/software/rscape`.

$$GT(i,j) = 2B \sum_{\sigma_i,\tau_j} F_{ij}(\sigma_i,\tau_j) log \frac{F_{ij}(\sigma_i,\tau_j)}{F_i(\sigma_i)F_j(\tau_j)} \equiv 2B \cdot M_{ij} \tag{2.6}$$

Significantly co-varying pairs are those with a low associated E-value.

2.5 Average product correction

The average product correction (APC) was introduced with the aim of removing the contribution of shared ancestry and phylogenetic history from pair covariance scores (Dunn, Wahl, and Gloor, 2007):

$$S_{ij}^{APC} = S_{ij} - \frac{\sum_i S_{ij} \sum_j S_{ij}}{\sum_{i,j} S_{ij}} \tag{2.7}$$

This is empirically found to remove the influence of positions entropy from the scores, even though the reason why this particular functional form works so well on DCA is still unknown (Ekeberg, Hartonen, and Aurell, 2014).

2.6 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, namely single-site and two-sites frequency counts (Morcos et al., 2011), such that

$$F_i(\sigma_i) = \sum_{\{\sigma_k | k \neq i\}} P(\sigma_1, ..., \sigma_N) \equiv f_i(\sigma_i) \tag{2.8}$$

$$F_{ij}(\sigma_i, \tau_j) = \sum_{\{\sigma_k | k \neq i,j\}} P(\sigma_1, ..., \sigma_N) \equiv f_{ij}(\sigma_i, \tau_j) \tag{2.9}$$

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

$$P(\sigma) = \frac{1}{Z} \exp\left(\sum_i h_i(\sigma_i) + \sum_{ij} J_{ij}(\sigma_i, \sigma_j)\right) \tag{2.10}$$

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

$$Z = \sum_{\{\sigma\}} \exp\left(\sum_i h_i(\sigma_i) + \sum_{ij} J_{ij}(\sigma_i, \sigma_j)\right) \tag{2.11}$$

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. Once parameters $h_i(\sigma)$ and $J_{ij}(\sigma, \tau)$ have been determined, the Frobenius norm of the coupling matrices can be used to obtain a scalar value for each pair of positions (De Leonardis et al., 2015; Cocco, Monasson, and Weigt, 2013; Ekeberg, Hartonen, and Aurell, 2014):

$$S_{ij} = \sqrt{\sum_{\{\sigma,\tau\}} J_{ij}(\sigma, \tau)^2} \tag{2.12}$$

We will discuss three different approaches that can be used to determine the parameters of the model: the mean-field approximation (Morcos et al., 2011), the pseudo-likelihood maximization (Ekeberg, Hartonen, and Aurell, 2014), and a Boltzmann-learning approach proposed here.

2.6.1 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 Morcos et al., 2011, by which coupling matrices are estimated as the inverse of the connected correlation matrices:

$$J_{ij}(\sigma_i, \sigma_j) \simeq -C_{ij}^{-1}(\sigma_i, \sigma_j) \tag{2.13}$$

and the local fields are estimated as:

$$h_i(\sigma_i) \simeq \ln \frac{F_i(\sigma_i)}{F_i(\bar{\sigma}_i)} - \sum_{j, j \neq i} \sum_{\substack{\sigma_i, \\ \sigma_i \neq \bar{\sigma}_i}} J_{ij}(\sigma_i, \sigma_j) F_j(\sigma_j) \tag{2.14}$$

where $C_{ij}(\sigma_i, \sigma_j) = F_{ij}(\sigma_i, \sigma_j) - F_i(\sigma_i) F_j(\sigma_j)$ is the correlation matrix and $\bar{\sigma}$ is an arbitrarily chosen letter of the alphabet, usually the one representing gaps. To make the matrix $C_{ij}(\sigma_i, \sigma_j)$ invertible and alleviate finite sample effects it is common to add

pseudo-counts. We adopt the same approach as in De Leonardis et al., 2015:

$$\hat{F}_i = (1-\lambda)F_i + \frac{\lambda}{5} \tag{2.15}$$

$$\hat{F}_{ij} = (1-\lambda)F_{ij} + \frac{\lambda}{25}(1-\delta_{ij}) + \frac{\lambda}{5}\delta_{ij}\delta_{\sigma_i\sigma_j} \tag{2.16}$$

where $\lambda = 0.5$.

2.6.2 Pseudo-likelihood maximization

An alternative approach to estimate the DCA inverse problem solution is the maximization of the conditional likelihood (or *pseudo-likelihood*) (Ekeberg et al., 2013), which is a consistent estimator of the full likelihood, but involves a tractable maximization and is considered as the state-of-the-art method for protein sequences. This is equivalent to minimizing the negative pseudo-log likelihood function:

$$l_{pseudo} = -\frac{1}{B}\sum_{r}\sum_{b=1}^{B}\log P(\sigma_r^b|\boldsymbol{\sigma}_{\backslash r}^b) \tag{2.17}$$

Here $\boldsymbol{\sigma}_{\backslash r}^b$ denotes the identity of all the nucleotides *except* the one at position r, and thus $P(\sigma_r^b|\boldsymbol{\sigma}_{\backslash r}^b)$ is the conditional probability of observing one variable σ_r given all the other variables. When data is abundant, the conditional likelihood tends to the full likelihood function (see, e.g., Arnold and Strauss, 1991). 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. We exploit the asymmetric pseudo-likelihood maximization (Ekeberg, Hartonen, and Aurell, 2014) as implemented at `https://github.com/magnusekeberg/plmDCA`.

2.6.3 Maximum likelihood and Boltzmann learning

Given a set of independent equilibrium configurations $\{\sigma^b\}_{b=1}^B$ of the model (Eq.2.10) 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 number of sequences:

$$l = -\frac{1}{B}\sum_{b=1}^{B}\log P(\sigma^b) \tag{2.18}$$

Minimizing l with respect to local fields h_i gives

$$\begin{aligned}\frac{\partial l}{\partial h_i(\sigma)} &= \frac{1}{B}\sum_{b=1}^{B}\left(\frac{\partial \log Z}{\partial h_i(\sigma)} - \delta(\sigma_i^b,\sigma)\right) = \\ &= \frac{1}{B}\sum_{b=1}^{B}\Big(f_i(\sigma) - \delta(\sigma_i^b,\sigma)\Big) \\ &= f_i(\sigma) - F_i(\sigma) = 0\end{aligned} \tag{2.19}$$

Similarly, minimizing l with respect to the couplings gives

$$\frac{\partial l}{\partial J_{ij}(\sigma,\tau)} = f_{ij}(\sigma,\tau) - F_{ij}(\sigma,\tau) = 0 \tag{2.20}$$

These equations show that the model maximizing the likelihood of parameters 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, Hinton, and Sejnowski, 1987; Sutto et al., 2015; Barrat-Charlaix, Figliuzzi, and Weigt, 2016; Haldane et al., 2016; Figliuzzi, Barrat-Charlaix, and Weigt, 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

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta l(\theta) = \theta_t - \eta_t(f(\theta) - F) \tag{2.21}$$

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 such 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 $f_i(\sigma)$; this procedure allows to update the parameters more frequently, resulting in a *stochastic gradient descent* that forces the system to sample the posterior distribution. The algorithm 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 use 20 simultaneous simulations initialized from 20 sequences randomly 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, Reißer, and Bussi, 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. The learning rate η_t belongs to the class *search then converge* (Darken and Moody, 1990):

$$\eta_t = \frac{\alpha}{1 + \frac{t}{\tau_S}} \tag{2.22}$$

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, this procedure is named *Boltzmann learning*. The exact algorithm is described in ***Algorithm* 1** and the employed C code is available at `https://github.com/bussilab/bl-dca`. In the algorithm implemented here, at variance with others proposed before (Sutto et al., 2015; Figliuzzi, Barrat-Charlaix, and Weigt, 2018), the Lagrangian multipliers are evolved every few Monte Carlo iterations using istantaneous values rather than averages obtained from converged trajectories. A change of variables of the model parameters was proposed to make the minimization easier (Figliuzzi, Barrat-Charlaix, and Weigt, 2018). This idea might be beneficial also in our algorithm.

2.6.4 Gauge invariance and regularization

The number of model parameters in Eq. 2.10 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. 2.8, 2.9) 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

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 $\alpha = \left(1, \frac{P_{new}}{P_{old}}\right)$, 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 frequencies.
3. **Validation:** Repeat step 2 using parameters $\{h, J\}$ computed as averages over the last 5000 Monte Carlo sweeps of step 2.

out to be $\frac{N(N-1)}{2}(q-1)^2 + N(q-1)$ (Weigt et al., 2009). 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 (Ekeberg et al., 2013; De Leonardis et al., 2015) is the one minimizing the norm of couplings matrices (Eq. 2.12):

$$\sum_{\{\tau\}} J_{ij}(\sigma, \tau) = \sum_{\{\sigma\}} J_{ij}(\tau, \sigma) = \sum_{\{\tau\}} h_i(\tau) = 0 \quad \forall\ i, j \tag{2.23}$$

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:

$$J_{ij}(\bar{\sigma}, \tau) = J_{ij}(\tau, \bar{\sigma}) = h_i(\bar{\sigma}) = 0 \quad \forall\ i, j, \tau \tag{2.24}$$

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 $l(\theta)$ (Ekeberg et al., 2013), such that:

$$\theta = \arg\min_{\theta}\{l(\theta) + R(\theta)\} \tag{2.25}$$

where

$$R(\theta) = \frac{k}{2}\sum_{p}\theta_p^2\ , \quad p = \{1, ..., \frac{N(N-1)}{2}q^2 + Nq\} \tag{2.26}$$

For pseudo-likelihood we use a value of k depending on the alignment size, adopting the default options supplied by the employed software. 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. We thus decided not to use any regularization term, since the chosen length of simulation is indeed playing the role of an early stopping, a form of regularization commonly used in gradient descent algorithms (Yao, Rosasco, and Caponnetto, 2007). The regularization term can also be interpreted in a Bayesian framework as a prior knowledge on the system. The possibility to use informative priors in DCA is discussed in Chapter 4.

2.6.5 Validation of the inferred couplings

The capability of the discussed DCA methods to infer a Potts model compatible with the frequencies observed in the MSA can be quantified by computing the root-mean-square deviation (RMSD) between model and observed pair frequencies:

$$RMSD = \sqrt{< (f_{ij}(\sigma_i, \tau_j) - F_{ij}(\sigma_i, \tau_j))^2 >_{\{ij\},\{\sigma_i,\tau_j\}}} \tag{2.27}$$

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

2.7 Validation of predicted contacts

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:

$$sensitivity = \frac{TP}{TP + FN} \tag{2.28}$$

$$precision = \frac{TP}{TP + FP} \tag{2.29}$$

The Matthews correlation coefficient (MCC) can be defined as the geometric average of sensitivity and precision (Matthews, 1975; Gorodkin, Stricklin, and Stormo, 2001):

$$MCC = \sqrt{sensitivity \cdot precision} \tag{2.30}$$

and is equivalent to the interaction network fidelity (Parisien et al., 2009). To turn contact scores S_{ij} 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) $\bar{S}$. In order to allow for a fair comparison between different covariance methods, we choose the threshold score maximizing the MCC, corresponding to the optimal compromise between precision and sensitivity as illustrated in Figure 2.1 for a test system (PDB: LY26). For each covariance method, the MCC as a function of the threshold score S shows a similar behavior for all the N_s 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:

$$\bar{S} = \arg\max_{S} \left(\prod_{\mu}^{N_s} MCC_{\mu}(S) \right)^{\frac{1}{N_s}} \tag{2.31}$$

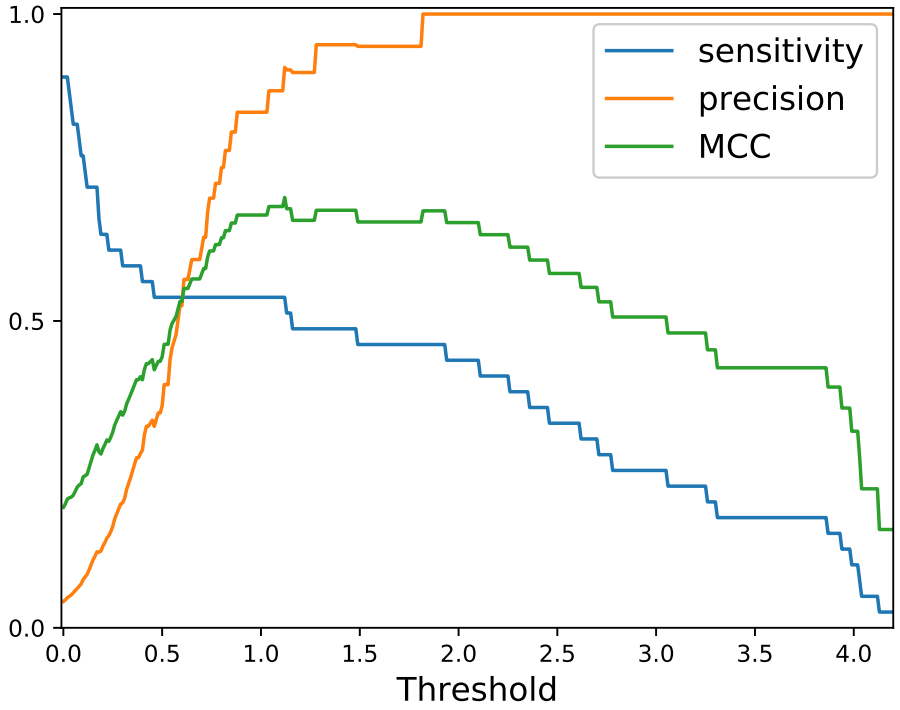


FIGURE 2.1: Typical behavior of sensitivity, precision and MCC illustrated for a test Ising system (PDB: LY26, Boltzmann learning DCA on *Infernal* alignment). Notice that the number of false positives grows as the score threshold decreases. Sensitivity is not exactly 1 at zero threshold because of APC correction introducing some negative values among pair scores.

Chapter 3

RNA contact prediction

An extensive assessment of the capability of covariance-based methods to infer contacts on a data set of 17 RNA families is reported in this chapter. In particular, the focus is on direct coupling analysis methods, which require the coupling constants of a Potts model to be estimated. We first assess the capability of different DCA solutions 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 all covariance methods to enhance RNA structure prediction. We also analyze the robustness of the different models with respect to the choice of the threshold on covariance scores and discuss the influence of some possible arbitrary choice on the accuracy of predictions. The performance of the methods in terms of a number of measures is reported, such as precision, sensitivity, number of tertiary contacts and false positives, sensitivity to contacts in secondary structures and number of stacked base pairs among false positives. Moreover, the predicted contacts of the two most accurate methods are explicitly shown, comparing directly the best and worst performing systems.

3.1 Data set

The analysis is performed on sequences of 17 riboswitches families classified in the Rfam database (Nawrocki et al., 2014). Riboswitches are ubiquitous in bacteria and thus show a significant degree of sequence heterogeneity within each family. The RNA families have been chosen among those for which at least one high-resolution

crystallographic structure has been reported, ruling out from the analysis the structures annotated as interacting double chains. The full dataset is listed in Table 3.1. The number of nucleotides in each chain ranges between 52 and 161, and the number of sequences between 189 and 10858. The lowest quality structure in the data set has been solved with resolution 2.95 Å. Contacts in the reference PDB structures are annotated with DSSR (Lu, Bussemaker, and Olson, 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, using the geometric distance between heavy atoms belonging to each nucleotide (thus including also backbone atoms), 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 co-evolution in RNA is mostly related to isostericity (Leontis, Stombaugh, and Westhof, 2002; Stombaugh et al., 2009), which is the property of some of the base pairs of a given family (e.g. Watson-Crick family) to assume very similar positions and distances between the C1' carbon atoms. All the used MSAs as well as files containing the annotation of each base pair are available at `https://github.com/bussilab/bl-dca`. Columns with more than 90% of gaps were removed from the alignments in order to make the maximization faster and to avoid overfitting on positions of the alignment that are not relevant. 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.

The required times for all the tested covariance methods scale roughly as the number of nucleotides squared and are listed in Table 3.2 for the largest and smallest molecules in the data set. (Dunn, Wahl, and Gloor, 2007).

3.2 Validation of the inferred couplings

As shown in Figure 3.1, the Boltzmann learning procedure is capable to infer a Potts model that generates sequences with the correct frequencies. The two displayed

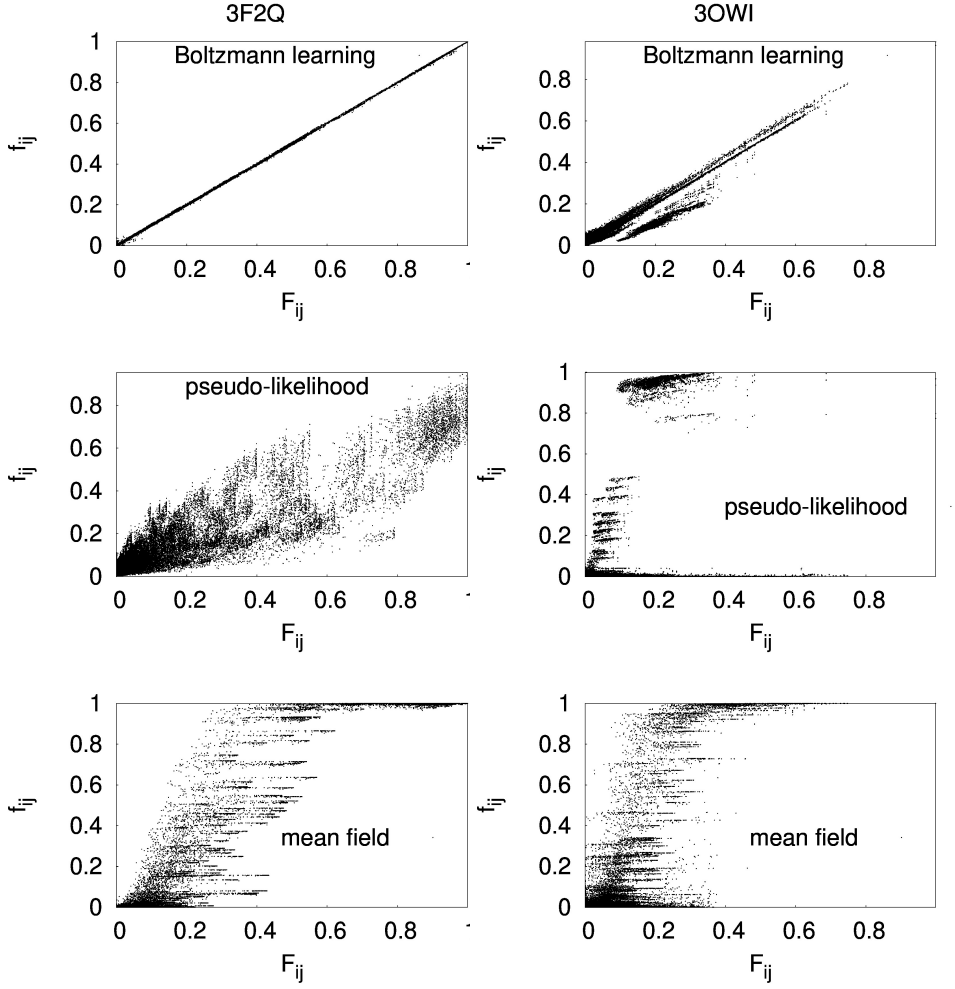
TABLE 3.1: PDB, RFAMcode molecule name, alignment length and size, effective alignment size after reweighting (similarity treshold x=0.9 on *Infernal* alignments).

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 3.2: Computational time for the smallest and largest investigated systems. 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 predictions 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

families are those where the model frequencies agree best (PDB: 3F2Q, Figure 3.1a) or worst (PDB: 3IRW, Figure 3.1b) 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 of 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 3.2 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 ($f_{ij} = 1/25$). The positive control sets a reference for RMSD values and corresponds to the statistical error on the frequencies calculated via the bootstrap procedure, by uniformly sampling with repetitions a number of sequences corresponding to the MSA size and calculating the RMSD between the so obtained averages and the empirical frequencies. In addition, we compare empirical



(A) (B)

FIGURE 3.1: FMN riboswitch (PDB code 3F2Q, figure 3.1a) and c-di-GMP-I (PDB code 3OWI, figure 3.1b). 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.

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 (Figliuzzi, Barrat-Charlaix, and Weigt, 2018; Gao, Zhou, and Aurell, 2018).

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 (Figure 3.3). Given that pseudo-likelihood is known to converge to the exact value in the limit of an infinite number of sequences (see, e.g., Arnold and Strauss, 1991), this discrepancy should be attributed to the typical size of the used alignments. 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 (Figliuzzi, Barrat-Charlaix, and Weigt, 2018).

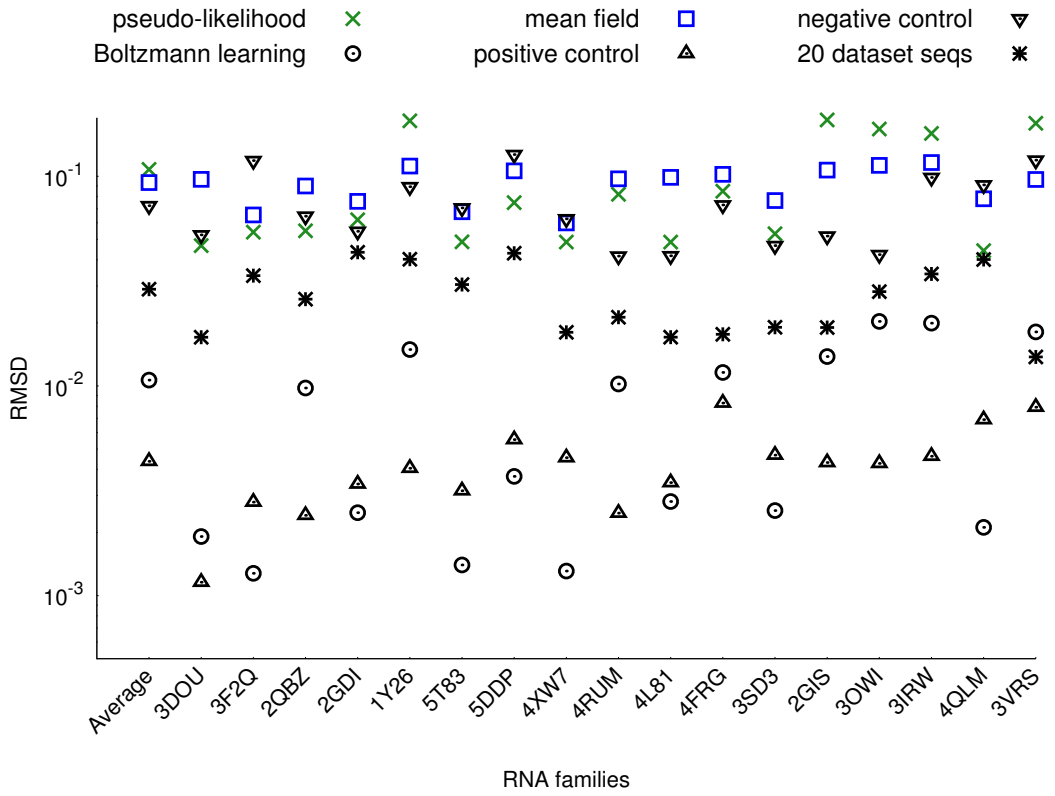


FIGURE 3.2: Validation of the coupling parameters inferred using different methods (Boltzmann learning, pseudo-likelihood and mean-field DCA). The validation is done running a parallel MC simulation on 20 sequences and calculating the root-mean-square deviation (RMSD) between the obtained frequencies and the empirical ones. We report a positive control (statistical error due to 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 the Boltzmann learning simulations. Families are labeled using the PDB code of the representative crystallographic structure. Average RMSD is reported in first column. *Infernal* alignments.

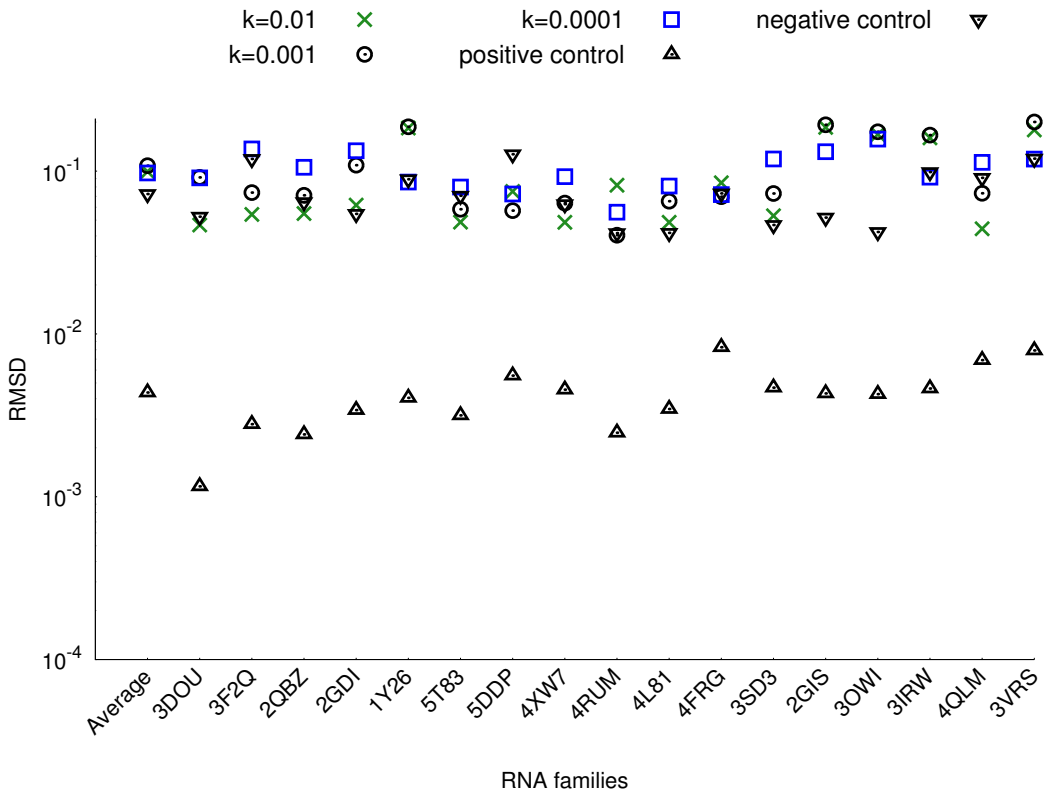


FIGURE 3.3: Validation of the coupling parameters inferred via the l_2-regularized pseudo-likelihood maximization method implemented at *https://github.com/magnusekeberg/plmDCA*, adopting different regularization strengths k. The validation is done running a parallel MC simulation on 20 sequences and calculating the root-mean-square deviation (RMSD) between the obtained frequencies and the empirical ones. The positive control is the statistical error due to finite number of sequence, and the negative control is the RMSD between empirical sequences and a random sequence. *Infernal* alignment.

3.3 Validation of predicted contacts

From the previous results it is clear how 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, this does not necessarily imply that the exact parameters of the Potts model are correlated with structural contacts. The predictions is validated against a set of crystallographic structures by computing the Matthews correlation coefficient (MCC) between the predicted and 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. 2.31), calculated excluding that system. 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 $\sqrt{\frac{2N_{native}}{N(N-1)}}$.

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 (Table 3.3 and 3.4). 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. This seems to be due to the dependency of the $ł_2$-regularization strength on the family size in the adopted plmDCA code, which induces lower optimal score thresholds for the less numerous families. In order to quantify this effect we introduce a transferability index $\phi = \frac{1}{N_s}\sum_{\mu}^{N_s}\frac{\overline{MCC_\mu}}{\overline{MCC_\mu^{max}}}$, which is the ratio between the cross-validated MCC for system μ ($\overline{MCC}_\mu$) and the maximum MCC that can be obtained by choosing the optimal threshold for each system MCC_μ^{max}, averaged over all systems. For the *Infernal* alignments, this value amounts to $\phi = 0.96$ for BL and to $\phi = 0.91$ for pseudo-likelihood DCA, suggesting

TABLE 3.3: *Infernal* alignments. $\overline{MCC}$ with optimal covariance score threshold $\overline{S}$ for Boltzmann learning DCA, pseudo-likelihood DCA, mean field DCA, mutual information, R-scape for each of 17 RNA families, obtained through cross-validation procedure.

	Boltzmann learning DCA		**Pseudo-likelihood DCA**		**mean field DCA**		**mutual information**		**R-scape**	
PDB	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$
3DOU	0.68	1.09	0.59	0.65	0.67	1.0	0.68	0.22	0.53	0.5
3F2Q	0.58	1.09	0.58	0.65	0.56	1.0	0.55	0.22	0.53	0.5
2QBZ	0.55	1.09	0.50	0.78	0.52	1.0	0.53	0.22	0.55	0.5
2GDI	0.55	1.09	0.51	0.65	0.57	1.0	0.48	0.22	0.53	0.5
1Y26	0.69	1.09	0.67	0.65	0.63	0.99	0.63	0.22	0.46	0.5
5T83	0.58	1.09	0.58	0.65	0.58	1.0	0.53	0.22	0.57	0.5
5DDP	0.65	1.09	0.63	0.65	0.66	1.0	0.65	0.22	0.52	0.5
4XW7	0.59	1.24	0.63	0.65	0.59	1.0	0.55	0.22	0.53	0.5
4RUM	0.60	1.19	0.39	0.78	0.54	1.06	0.55	0.22	0.65	0.5
4L81	0.46	1.09	0.45	0.78	0.43	1.0	0.35	0.22	0.43	0.5
4FRG	0.63	1.09	0.49	0.78	0.50	0.99	0.64	0.22	0.65	0.5
3SD3	0.67	1.05	0.69	0.65	0.67	1.0	0.63	0.22	0.63	0.5
2GIS	0.67	1.14	0.74	0.65	0.44	1.03	0.37	0.22	0.59	0.5
3OWI	0.73	1.11	0.73	0.65	0.67	1.0	0.29	0.24	0.62	0.5
3IRW	0.58	1.09	0.56	0.65	0.50	1.0	0.35	0.22	0.42	0.3
4QLM	0.56	1.05	0.58	0.65	0.49	1.0	0.43	0.22	0.43	0.5
3VRS	0.64	1.11	0.71	0.65	0.71	1.0	0.67	0.22	0.42	0.1

TABLE 3.4: *ClustalW* alignment. $\overline{MCC}$ with optimal covariance score threshold $\overline{S}$ for Boltzmann learning DCA, pseudo-likelihood DCA, mean field DCA, mutual information, R-scape for each of 17 RNA families, obtained through cross-validation procedure.

	Boltzmann learning DCA		**Pseudo-likelihood DCA**		**mean field DCA**		**mutual information**		**R-scape**	
PDB	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$	$\overline{MCC}$	$\overline{S}$
3DOU	0.47	1.07	0.45	0.43	0.42	0.82	0.47	0.20	0.50	1.3
3F2Q	0.48	0.99	0.45	0.43	0.32	0.80	0.31	0.20	0.43	1.3
2QBZ	0.49	1.07	0.46	0.51	0.45	0.80	0.39	0.20	0.46	1.3
2GDI	0.44	1.07	0.35	0.47	0.35	0.82	0.29	0.20	0.45	1.3
1Y26	0.57	1.07	0.50	0.43	0.51	0.82	0.32	0.20	0.38	1.3
5T83	0.41	1.07	0.38	0.43	0.32	0.82	0.44	0.20	0.44	1.1
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	0.40	1.3
4RUM	0.46	1.07	0.32	0.51	0.24	0.80	0.37	0.20	0.51	1.3
4L81	0.27	1.07	0.29	0.45	0.18	0.80	0.16	0.20	0.26	1.3
4FRG	0.59	1.07	0.44	0.57	0.34	0.82	0.40	0.20	0.57	1.3
3SD3	0.71	1.07	0.72	0.45	0.58	0.8	0.50	0.20	0.60	1.3
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	0.32	2.8
3IRW	0.55	1.07	0.37	0.44	0.39	0.80	0.25	0.20	0.28	1.1
4QLM	0.38	1.07	0.45	0.51	0.30	0.80	0.10	0.23	0.20	1.3
3VRS	0.55	1.08	0.42	0.43	0.42	0.82	0.34	0.20	0.30	1.3

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.

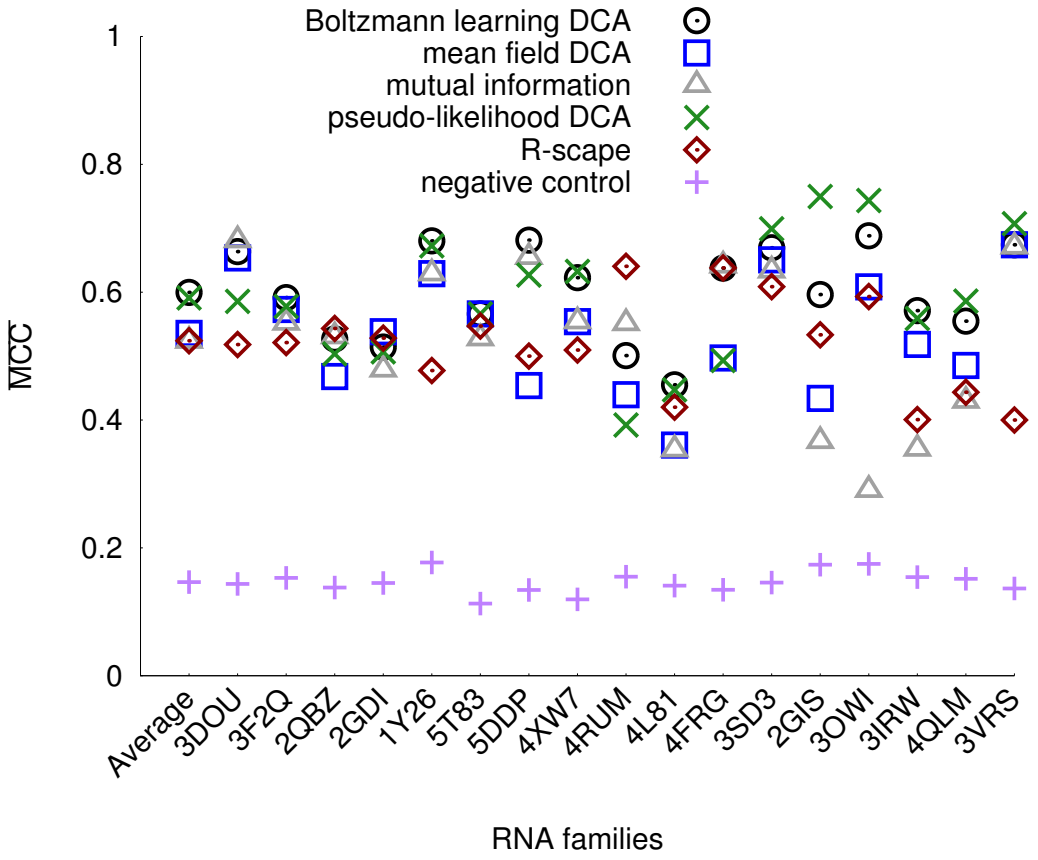


FIGURE 3.4: *Infernal* alignment. $\overline{MCC}$ of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information, R-scape and negative control for 17 RNA families at optimal threshold score obtained through cross-validation procedure. Families are labeled using the PDB code of the representative crystallographic structure. Average $\overline{MCC}$ reported in first column.

Results of the cross-validation procedure for each system (Figure 3.4 for *Infernal* alignments) indicate that direct coupling analysis outperforms mutual information and R-scape, and in particular Boltzmann learning performs the most accurate prediction. We considered the MSA methods implemented in both *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 3.5). We attribute this improvement in the quality of prediction performance to the use of consensus secondary structure in *Infernal* (Nawrocki and Eddy, 2013). The

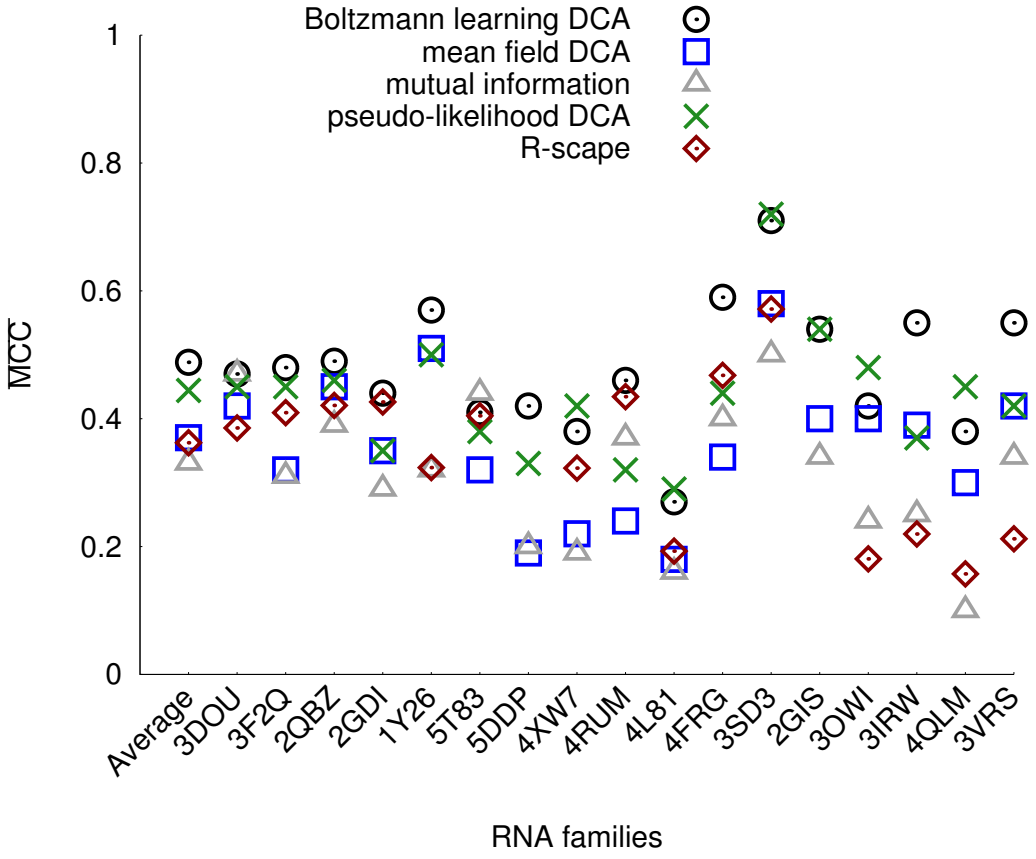


FIGURE 3.5: *ClustalW* alignments. $\overline{MCC}$ of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information, and R-scape for 17 RNA families at optimal threshold score obtained through cross-validation procedure. Families are labeled using the PDB code of the representative crystallographic structure. Average $\overline{MCC}$ is reported in the first column.

average MCC curves of two most accurate covariance methods (Boltzmann learning and pseudo-likelihood DCA) are shown to compare the two alignment methods directly (Figure 3.6). 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. 2.31), again enlightening a greater consistency in its choice for the Boltzmann learning with respect to pseudo-likelihood maximization framework.

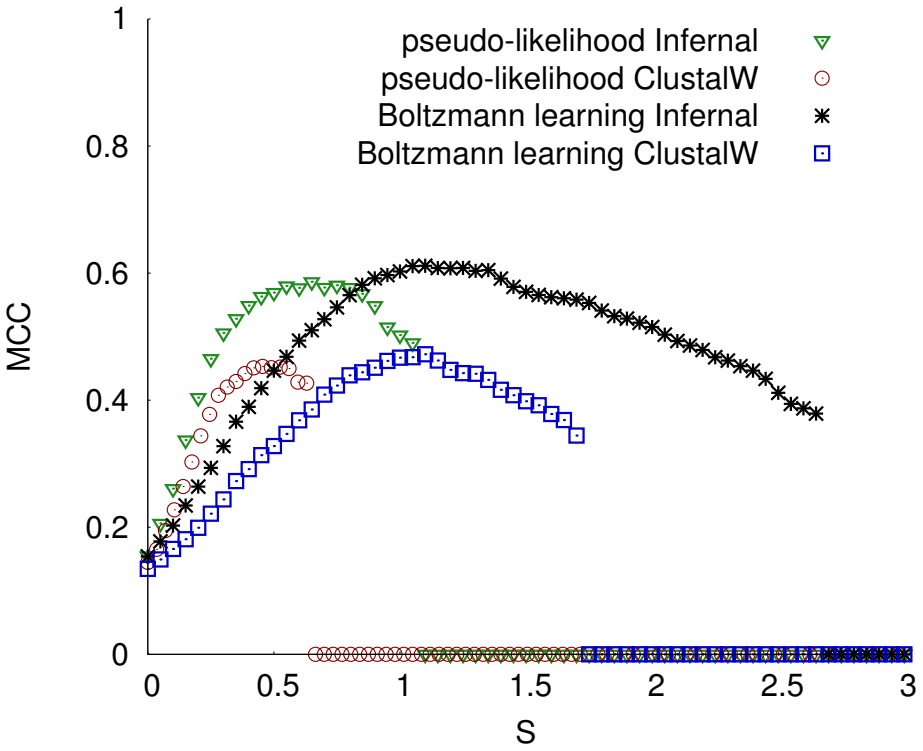


FIGURE 3.6: MCC geometric average over the 17 systems as a function of threshold scores S for Boltzmann learning and pseudo-likelihood DCA. MSAs are performed with *ClustalW* and *Infernal*, as indicated. The sharp decrease after some method-dependent value of S is due to the fact that when the threshold 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

Sensitivity and precision are independently monitored for each RNA family at cross-validated thresholds, in order to better quantify the capability of the investigated methods to provide useful information about contacts. 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 (Figures 3.7, 3.8). In particular, we notice that the two least populous families (PDB: 4FRG and 4RUM) show the lowest sensitivities for pseudo-likelihood DCA due to a lower optimal score thresholds for these systems induced by stronger adopted regularizations. It is interesting to notice how on *ClustalW* alignments the sensitivity drops significantly for mutual information, R-scape and mean field DCA, while it is slightly lower for pseudo-likelihood and Boltzmann learning DCA compared to results obtained on *Infernal* alignments (consistently with the level of approximation of the models).

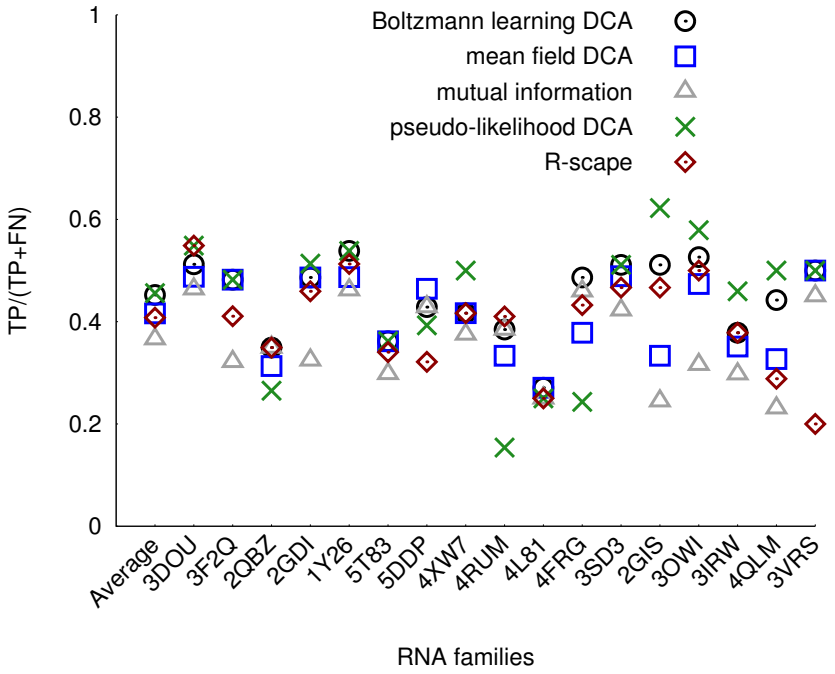


FIGURE 3.7: *Infernal* alignment. Sensitivity of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average sensitivity is reported in the first column.

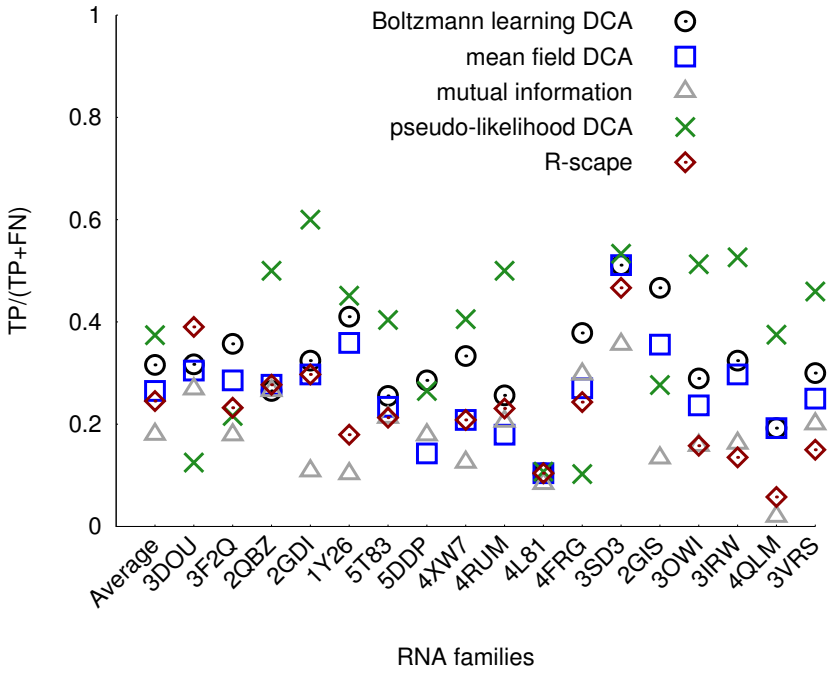


FIGURE 3.8: *ClustalW* alignment. Sensitivity of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average sensitivity is reported in the first column.

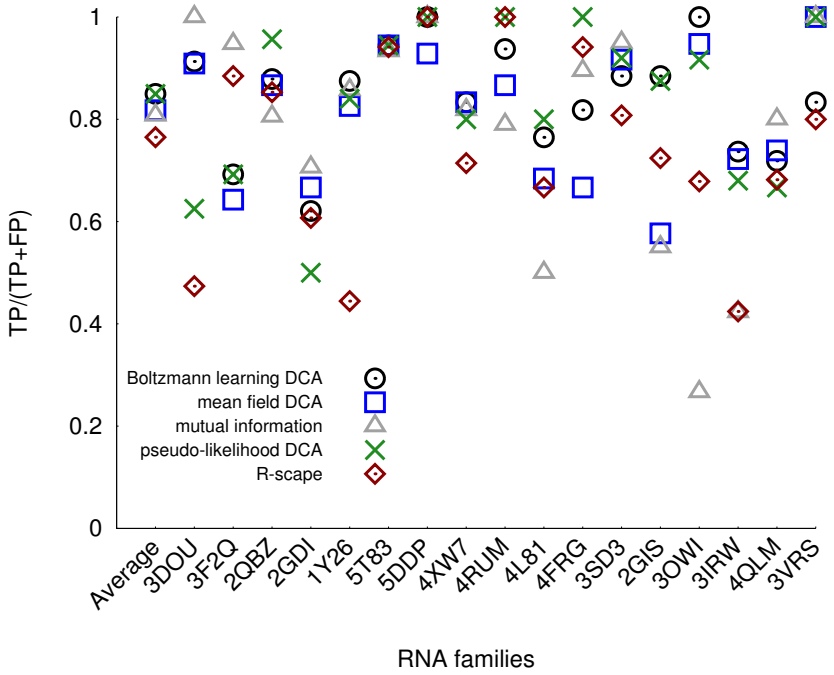


FIGURE 3.9: *Infernal* alignment. Precision of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average precision is reported in the first column.

The average precision instead ranges between 0.7 and 0.9 on *Infernal* alignments, indicating that the number of falsely predicted contacts is rather small (Figure 3.9). On the other hand, on *ClustalW* alignments there is a visible drop in precision (ranging between 0.5 and 0.8) concerning all methods but Boltzmann learning DCA (Figure 3.10). 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 an E-value threshold chosen to maximize the MCC in a training set. By using the recommended threshold (E-value$<$0.05) we would have obtained a higher precision, a lower sensitivity, and a lower MCC.

In order to assess the capability of these methods to probe RNA tertiary structure it is useful to look at the sensitivity value restricted to secondary contacts, obtained considering only base pairs contained in stems (in this analysis also pseudo-knots are included as stems). The sensitivity to contacts in stem is quite high for all methods on *Infernal* alignment, their average ranging between 0.6 and 0.7 with a systematic benefit in adopting pseudo-likelihood or Boltzmann learning DCA (Figure 3.11).

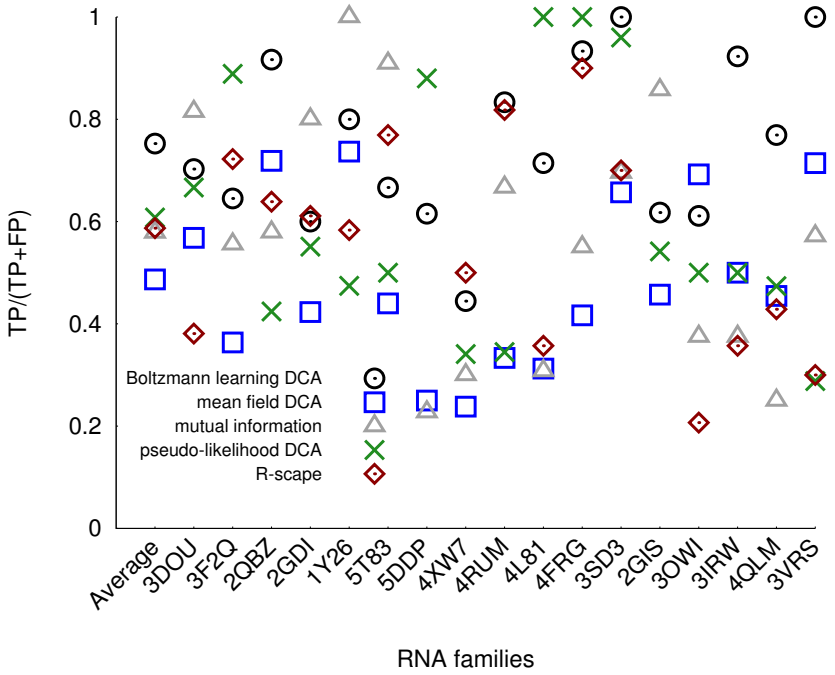


FIGURE 3.10: *ClustalW* alignment. Precision of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average precision is reported in the first column.

All methods show a visible loss in secondary structure sensitivity when adopting *ClustalW* alignments (Figure 3.12).

The number of true positive tertiary contacts are reported in figures 3.13 and 3.14. A contact is here considered as tertiary irrespectively of which edges are shared between nucleobases, and might even be an isolated WC pair. In other words, tertiary contacts are base-pairs that don't belong to any stem. In general, DCA is able to identify not only cWW pairs (Leontis and Westhof, 2001), where covariance is mostly associated to canonical pairs (GC, AU, and GU), but also a number of non-canonical pairs (see Table 3.5 as an example, where we also report the total number of tertiary contacts present in each of the 17 folded RNA structures). When looking at the absolute number of incorrect predictions the Boltzmann learning DCA provides the smallest average number (Figures 3.15 and 3.16). In particular, R-scape and pseudo-likelihood DCA report 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 these methods. A more careful eye on incorrect predictions reveals that couplings in consecutive nucleotides might be affected by a bias in the dinucleotide

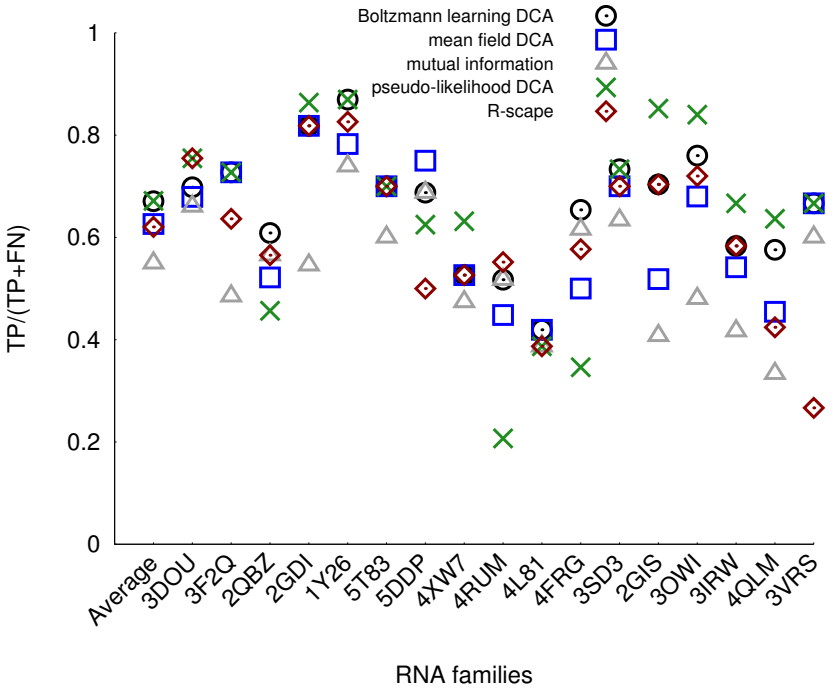


FIGURE 3.11: *Infernal* alignment. Sensitivity to contacts in stems (RNA secondary structure) of Boltzmann learning DCA, mean field DCA, mutual information and R-scape for all families. Families are labeled using the PDB code of the representative crystallographic structure. Average reported in first column.

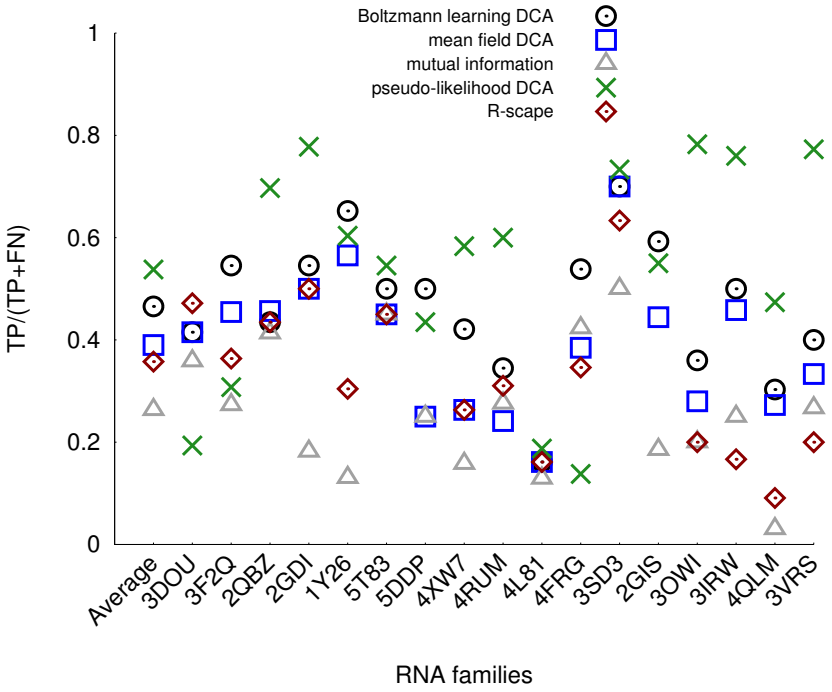


FIGURE 3.12: *ClustalW* alignment. Sensitivity to contacts in stems (RNA secondary structure) of Boltzmann learning DCA, mean field DCA, mutual information and R-scape for all families. Families are labeled using the PDB code of the representative crystallographic structure. Average reported in first column.

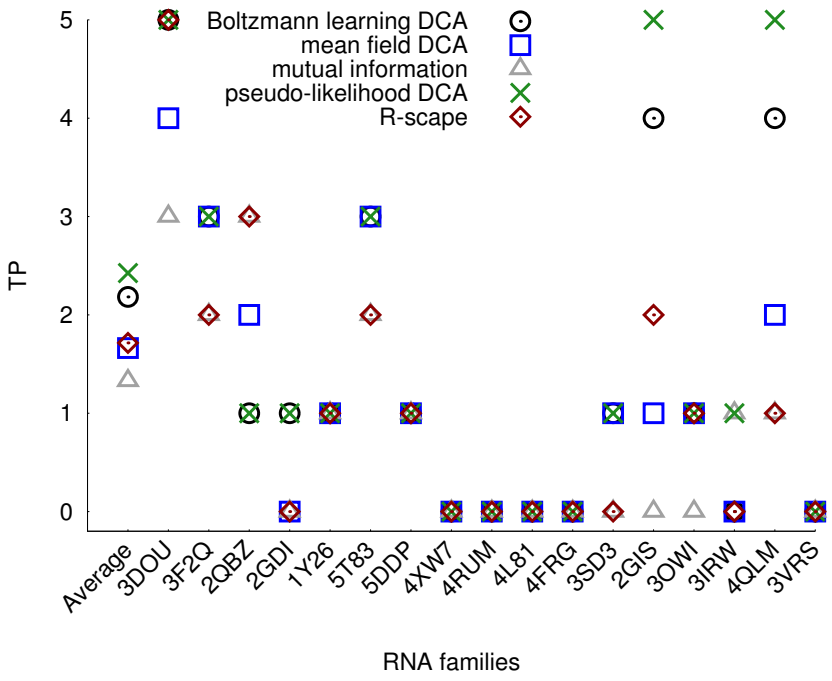


FIGURE 3.13: *Infernal* alignment. Number of correctly predicted (True Positives) tertiary contacts of Boltzmann learning DCA, mean field DCA, mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average reported in first column.

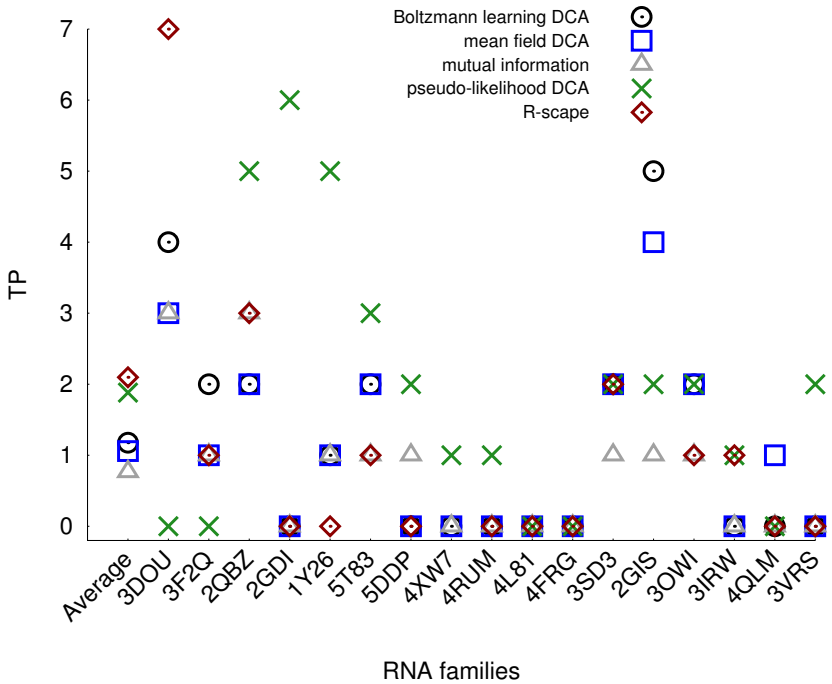


FIGURE 3.14: *ClustalW* alignment. Number of correctly predicted (True Positives) tertiary contacts of Boltzmann learning DCA, mean field DCA, mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average reported in the first column.

TABLE 3.5: 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

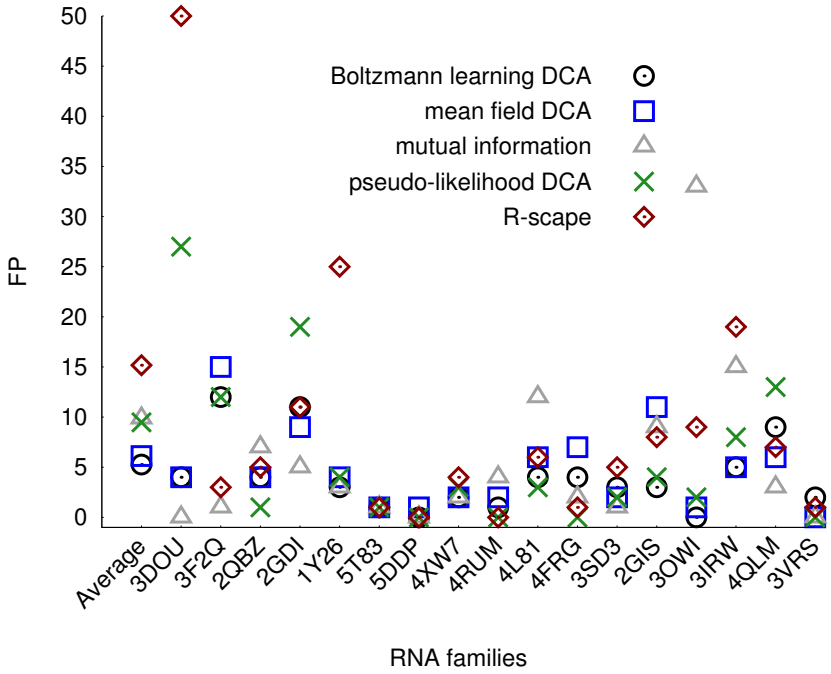


FIGURE 3.15: *Infernal* alignment. Number of incorrect predictions (False Positives) of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average reported in the first column.

distribution, due to mutational pressure produced by distinct probability of different substitution types (Belalov and Lukashev, 2013).

3.5 Typical contact predictions

It is instructive to visualize specific contact predictions on individual systems. First, we 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 3.17, 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 3.18, 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

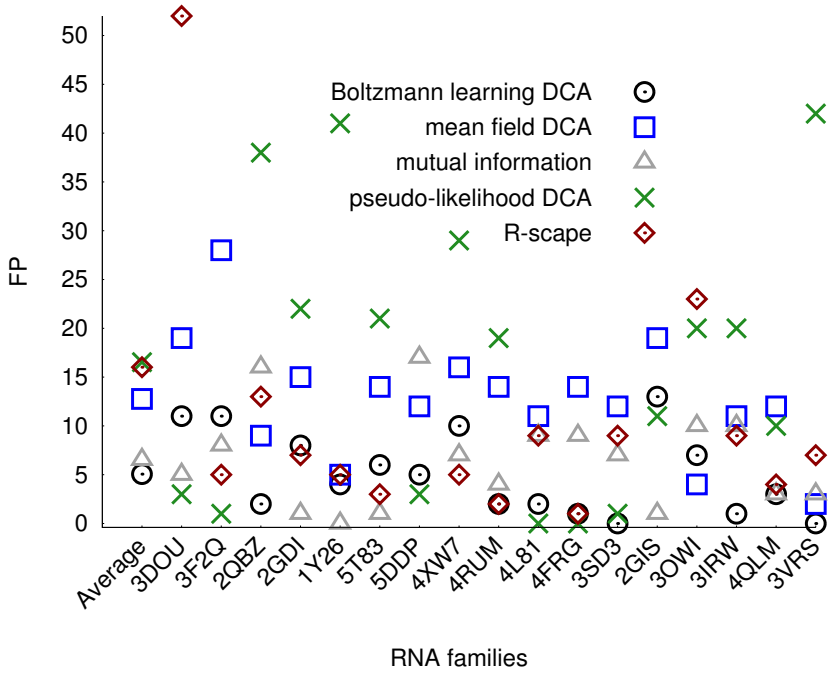
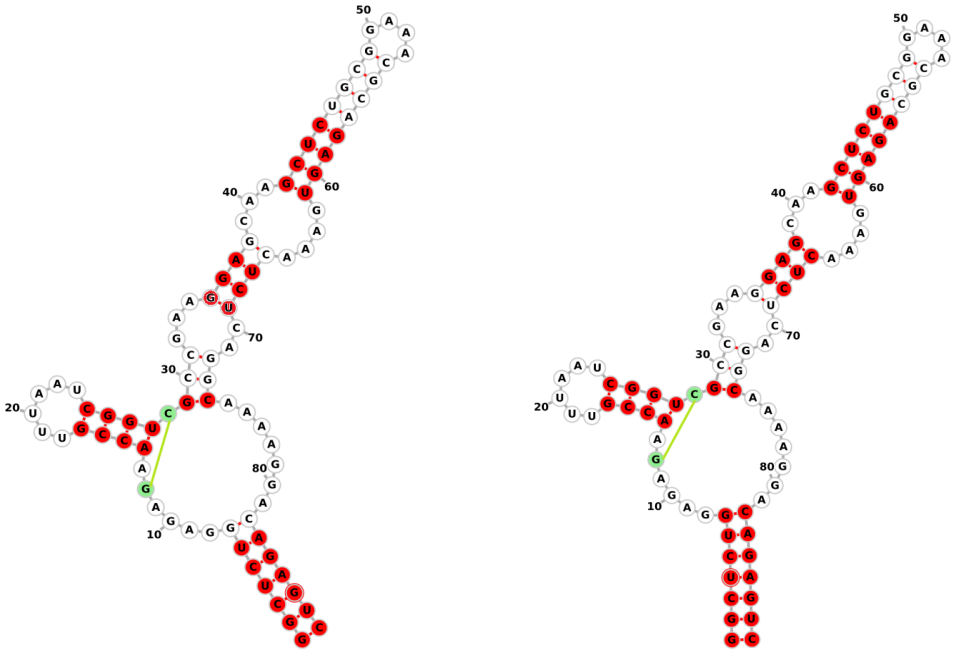


FIGURE 3.16: *ClustalW* alignment. Number of incorrect predictions (False Positives) of Boltzmann learning DCA, pseudo-likelihood DCA, mean-field DCA, mutual information and R-scape for all RNA families. Families are labeled using the PDB code of the representative crystallographic structure. Average reported in the first column.

that in the best cases these methods allow full helices to be identified accompanied by a small number of critical tertiary contacts.

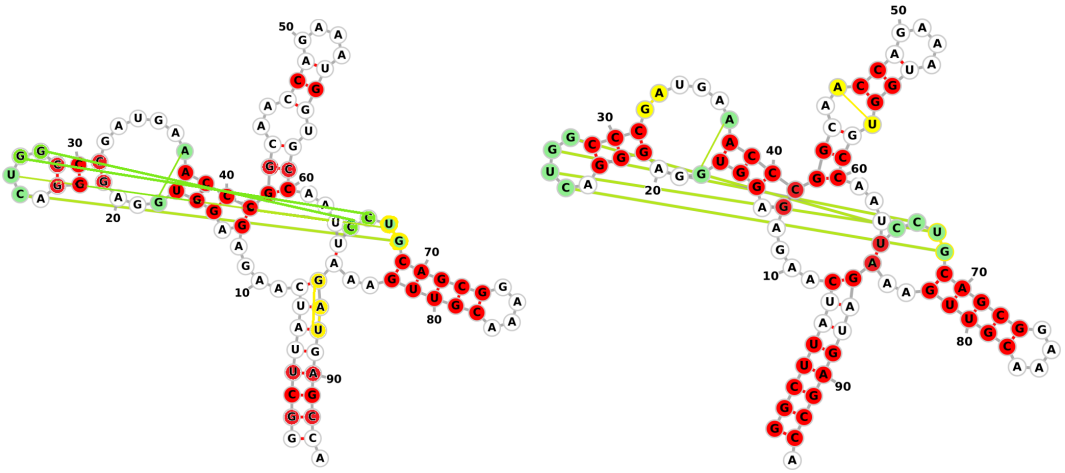
It is also useful to consider the cases resulting in the lowest MCC (SAM-I/IV riboswitch, PDB 4L81, for Boltzmann learning and NiCo riboswitch, PDB 4RUM, for pseudo-likelihood DCA). In the SAM-I/IV riboswitch the two methods give comparable results, and only a limited number of secondary contacts are correctly predicted (Figure 3.19). 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



(A) (B)

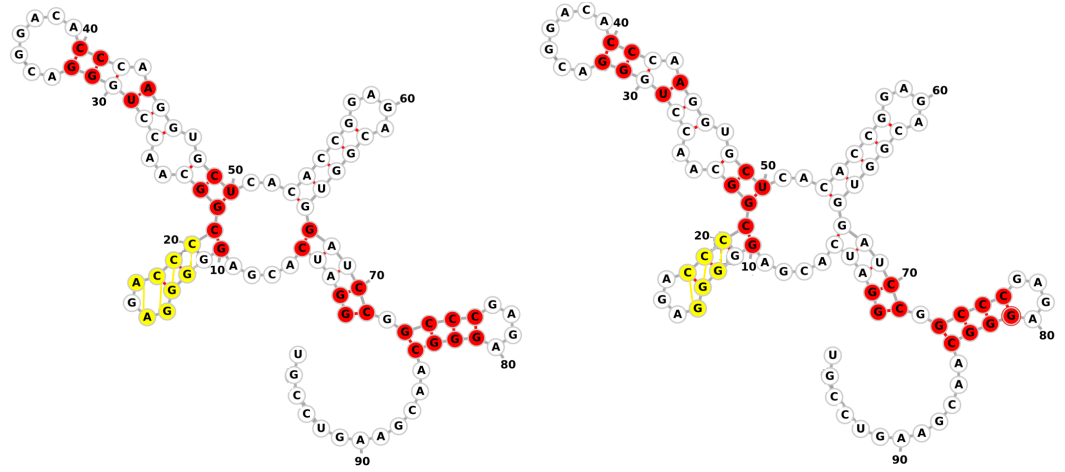
FIGURE 3.17: Glycine riboswitch (PDB code 3OWI) most accurate Boltzmann learning prediction (A) and respective pseudo-likelihood prediction (B). Correctly predicted contacts in secondary structure are shown in red. Correctly predicted tertiary contacts are shown in green. False positives are shown in yellow. We notice that G12/C28 pair is here labeled as tertiary since it corresponds to a isolated Watson-Crick pair in the reference structure.

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 3.20, 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.



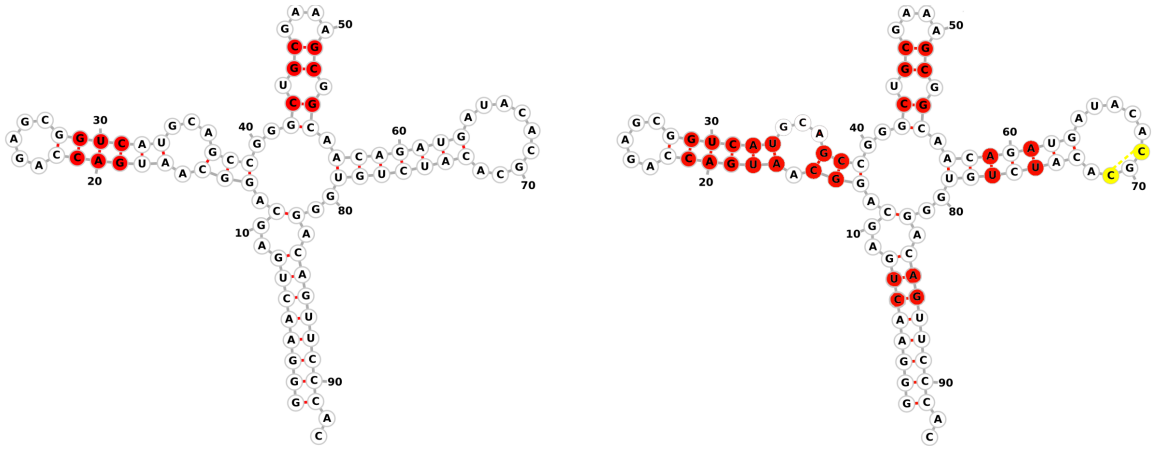
(A) (B)

FIGURE 3.18: SAM riboswitch (PDB code 2GIS), most accurate pseudo-likelihood prediction (A) and respective Boltzmann learning prediction (B). Correctly predicted contacts in secondary structure are shown in red. Correctly predicted tertiary contacts are shown in green. False positives are shown in yellow.



(A) (B)

FIGURE 3.19: SAM-I/IV riboswitch (PDB code 4L81), least accurate Boltzmann learning prediction (A) and respective pseudo-likelihood prediction (B). Correctly predicted contacts in secondary structure are shown in red. Correctly predicted tertiary contacts are shown in green. False positives are shown in yellow.



(A) (B)

FIGURE 3.20: NiCo riboswitch (PDB code 4RUM), least accurate pseudo-likelihood prediction (A) and respective Boltzmann learning prediction (B). Correctly predicted contacts in secondary structure are shown in red. Correctly predicted tertiary contacts are shown in green. False positives are shown in yellow.

3.6 Influence of MSA columns removal

The effect of removing a priori the MSA columns corresponding to gaps in the target sequence is to reduce the computational cost required for inferring DCA couplings for all methods. The Boltzmann learning DCA predictions are unaltered by this step while simulations on the full alignments are much more time demanding when compared with the reduced ones. On the other hand, results obtained through the pseudo-likelihood maximization approach are sensitive to the removal of columns, as it can be seen from the MCC geometric average over all systems as a function of the score threshold (Fig 3.21). The worse performance of plm-DCA on the full MSA could be due to a stronger effect of the regularization term when more columns are present, leading to a higher discrepancy among ranges of coupling values for families with heterogeneous numbers of sequences. In particular, scores obtained from the least numerous families (PDB: 4FRG and 4RUM) show lower values with respect to the other systems, causing the significant drop in the average MCC.

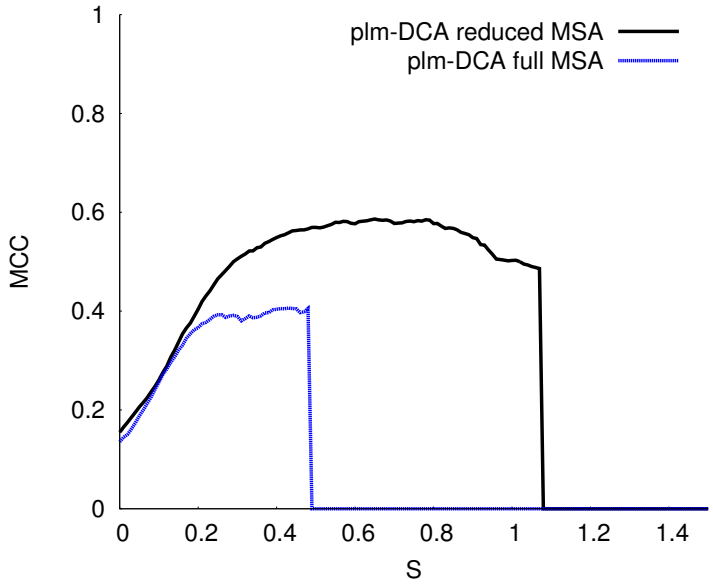


FIGURE 3.21: Average MCC at various score thresholds resulting from pseudo-likelihood DCA on the entire alignments (full MSA) and on alignments without columns corresponding to gaps in the target sequence. *Infernal* alignments.

3.7 Re-weighting

The accuracy of prediction (maximum average MCC) as a function of different similarity thresholds for sequence re-weighting obtained with Boltzmann learning DCA (Figure 3.22) reveals that results from this method are poorly affected by the reweighting procedure. In particular, there is a modest improvement if frequency counts from sequences with very high similarity in the MSA (x=0.9) are under-weighted.

3.8 APC correction

The empirical average product correction is found to improve the accuracy of prediction of DCA for this dataset. Table 3.6 shows the comparison between the maximum average MCC obtained through all DCA methods performed with and without the adoption of the APC correction.

3.9 Influence of stacking

It is worthy to notice that in many cases false positives are just labeled so by our decision to exclude stacking interactions from the true contacts. The fraction of stacked

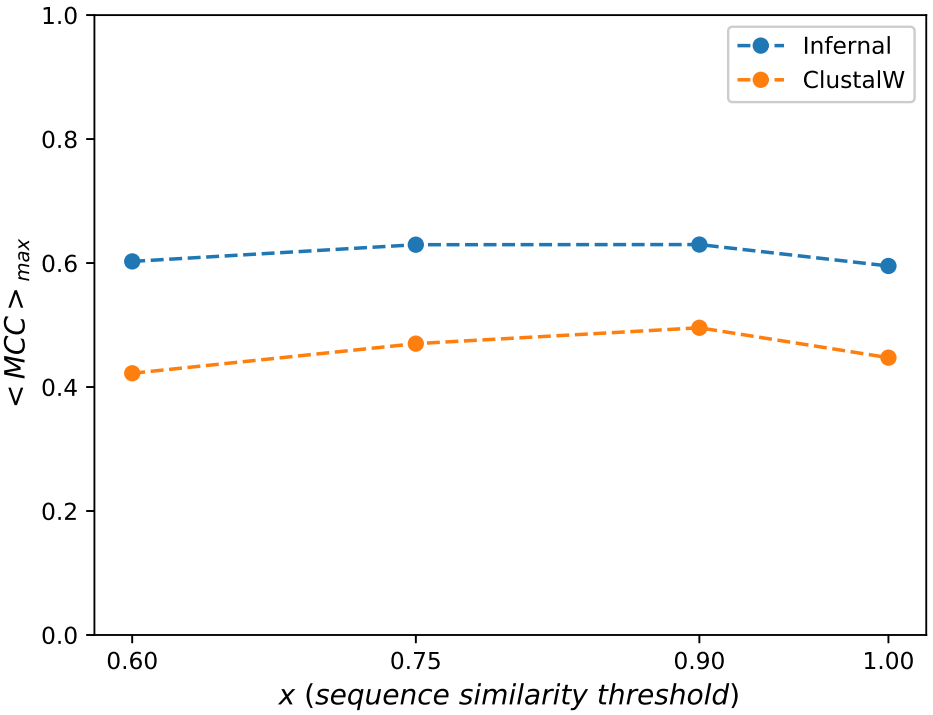


FIGURE 3.22: Maximum average MCC at various similarity thresholds for sequence re-weighting.

pairs that are reported as false positives over the total number of false positives is reported in Table 3.7 for all different DCA methods as an average over all the 17 RNA molecules. Among the few false positives, $\approx 50\%$ are truly stacked pairs in the pdb reference structure (base atomes distance < 3.5 Å).

TABLE 3.6: Maximum average $\overline{MCC}$ 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 3.7: Average fraction of stacked false positives (base atoms distance < 3.5 Å in the pdb reference structure) for all methods at optimal score threshold. *Infernal* alignment.

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

3.10 Validation on non-riboswitch systems

The DCA Boltzmann learning procedure is further validated 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 (leave-one-out procedure on the new systems). Results are reported in Table 3.8 and are slightly worse than those obtained for riboswitch families, with the exception of tRNA.

TABLE 3.8: 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

3.11 Discussion

A systematic assessment of RNA contact prediction based on the co-evolution analysis of nucleotides in aligned homologous sequences is carried out comparing mutual information analysis, R-scape, and DCA. Differently from other previous works (De Leonardis et al., 2015; Weinreb et al., 2016; Wang et al., 2017a), our analysis does not convert the resulting couplings into a structural model and focuses on the DCA calculation. 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 particular, we only considered base pairing and excluded other base-backbone or backbone-backbone contacts.

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. 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 worse 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, 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 stochastic solution of the inverse problem here introduced (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, Zecchina, and Berg, 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, Barrat-Charlaix, and Weigt, 2018).

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. 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.

The impact on contact prediction of other sometime overlooked choices (re-weighting 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. Moreover, we consider the discrepancy between the two alignments methods particularly remarkable since future blind contact predictions, on families for which no structure information is available, require the adoption of the *ClustalW* procedure.

Chapter 4

Encoding prior information in inverse Ising-like models

Inverse problems in statistical physics arise from the need to create models capable to interpret large amounts of data (Nguyen, Zecchina, and Berg, 2017). They consist in using the result of some observations to infer the values of the parameters characterizing the system under investigation. The solution to such problem can be difficult to assess because different values of the model parameters may be consistent with the data, or their discovery may require the exploration of a huge parameter space. Inverse Ising or Potts models are among the simplest physical models used in this context and have been applied in a number of fields, ranging from reconstruction of gene regulatory network (Lezon et al., 2006) to solution of diluted Sherrington-Kirkpatrick models (Aurell and Ekeberg, 2012), to biomolecular contact predictions starting from co-evolutionary information (Morcos et al., 2011). In the latest case, that is direct coupling analysis discussed in Chapter 2, the applicability of these method is intrinsically limited by two types of error: (a) statistical, that is the size of the available dataset (for DCA the number of sequences); and (b) systematic, that is the intrinsic error in the interpretation of the obtained parameters (for DCA the assumption that large couplings correspond to physical contacts).

In inverse problems, parameters of the model are inferred based on observations maximizing a likelihood function or a suitable approximation to it. This maximization is usually performed using regularization terms in order to avoid overfitting (Ekeberg et al., 2013; Marruzzo et al., 2017; Tyagi et al., 2016; Ravikumar, Wainwright, Lafferty, et al., 2010; Figliuzzi, Barrat-Charlaix, and Weigt, 2018) , especially

when a limited number of training examples is available, thus limiting the impact of the statistical error. In a Bayesian framework, the regularization term can be interpreted as a prior information on the parameters that is encoded in the process (Zhu, Chen, and Xing, 2014; Baldassi et al., 2014). For instance, a l_2 regularization is equivalent to a Gaussian prior on the parameters of the model. In principle, any prior information about the parameters can be included in order to make their estimation more reliable and thus decrease both systematic and statistical errors.

In what follows, we show how to use an informative prior to improve the solution of inverse Ising-like models. We first illustrate the procedure on a simple 10 spins Ising model, that can be solved by complete enumeration, where we use synthetic data to emulate the *a priori* knowledge on the parameters. Statistical and systematic errors are artificially introduced to test the capability of an informative prior to cure for both types of error. We then show how the introduced technique can be used in a real-life application of direct coupling analysis, namely to the prediction of contacts in RNA systems. In particular, we perform DCA through the Boltzmann learning technique and include in the parameters learning process information obtained from a secondary structure prediction algorithm. The idea of helping the inference including external knowledge is new in DCA literature, where instead l_2 regularization is usually adopted to avoid overfitting (Ekeberg et al., 2013; Figliuzzi, Barrat-Charlaix, and Weigt, 2018) and systematic error is only tackled by post-processing in some advanced way the resulting couplings (Schug et al., 2009; Chen et al., 2011; Ma et al., 2015; Wang et al., 2017a; Wang et al., 2017b).

4.1 Ising model

We first consider a system of 10 spins interacting through an Ising Hamiltonian, with possible states $\sigma = \pm\frac{1}{2}$. The couplings $\bar{J}$ are chosen randomly from a Gaussian distribution with zero average and variance Var(J) =0.5. The equilibrium distribution reads

$$P(\{\sigma\}) \propto \exp\left(\sum_{ij} \bar{J}_{ij}\sigma_i\sigma_j\right) \tag{4.1}$$

Since the number of possible state is only $2^{10} = 1024$, the partition function, as well as the average of any possible observable, can be computed by explicitly enumerating them. We then generate a limited number N_s of states, σ, drawn from the distribution in Eq. (4.1), and use the information contained in these states to infer the couplings J. Inference is done maximizing the posterior probability of the couplings given the observed states:

$$P(\mathbf{J}|\sigma) = P(\sigma|\mathbf{J})P(\mathbf{J}) = \left(\prod_{i=s}^{N_s} P(\{\sigma\}_s|\mathbf{J})\right) P(\mathbf{J}) \tag{4.2}$$

or, equivalently, minimizing the negative log-probability divided by the number of observations:

$$\mathcal{L} = -\frac{1}{N_s}\sum_s \log P(\{\sigma\}_i|\mathbf{J}) - \frac{1}{N_s}\log P(\mathbf{J}) \tag{4.3}$$

Here $P(\mathbf{J})$ encodes our prior knowledge on the model parameters. Assuming that we have an independent manner to estimate the parameters of the model, we here choose a prior in the form

$$-\log P(\mathbf{J}) = -\frac{\lambda}{2}\sum_{ij}(J_{ij} - J_{ij}^{prior})^2 \tag{4.4}$$

J^{prior} are generated as $J^{prior} = \bar{J} + \epsilon_{prior}\mathcal{N}(0,1)$, where ϵ_{prior} represents the error in our a priori estimate of the couplings and is here set to 1. Here λ is a hyper-parameter that must be properly chosen to maximize the performance of the inference procedure. $\lambda = 0$ corresponds to ignoring the prior knowledge, whereas $\lambda \to \infty$ corresponds to only using the prior knowledge ignoring any information arising from the observed states. Since in DCA applications one is interested in the ranking of the

couplings rather than on their precise values, we judge the quality of the inference procedure by computing the Pearson correlation coefficient $\rho_{J,J}$ between the reference couplings and the inferred ones. Alternative metrics to evaluate the ranking, such as the Kendall coefficient of concordance, can be used as well.

4.1.1 Statistical error

We first mimic the presence of statistical error by inferring couplings on finite samples of states N_s. The Pearson correlation coefficient $\rho_{J,J}$ is shown as a function of the number of states in Fig. 4.1, as computed for different choices of λ. Results using the Kendall coefficient of concordance are analogous, as shown in Fig 4.2.

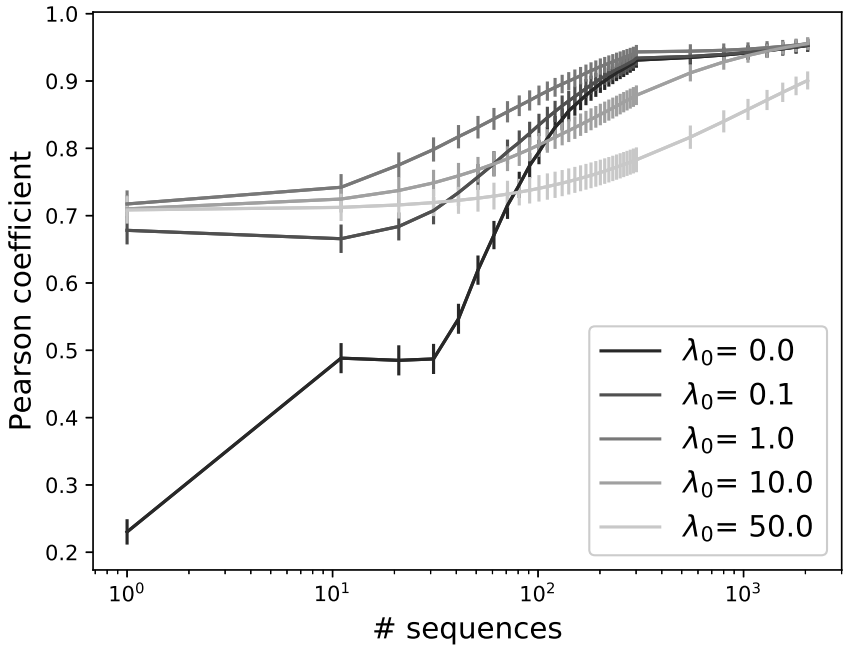


FIGURE 4.1: Ising system. Pearson correlation coefficient between ground truth and inferred couplings for various numbers of sequences and different values of λ_0. Average and standard deviation over 500 realizations of the system.

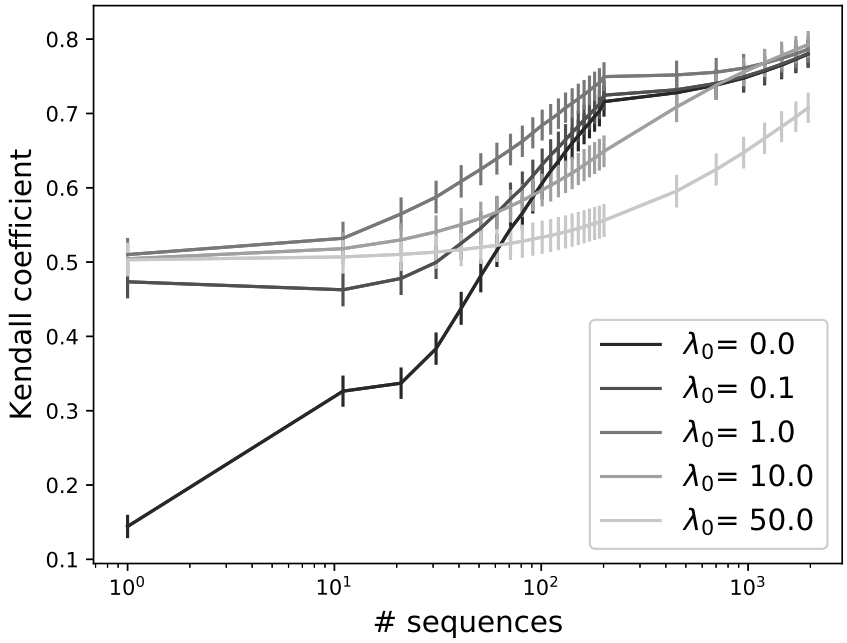


FIGURE 4.2: Ising system. Kendall coefficient between ground truth and inferred couplings for various numbers of sequences and different values of λ_0. Average and standard deviation over 500 realizations of the system.

Here it can be seen that there is a unique value of λ that leads to the highest correlation between the exact and the inferred couplings independently of the number of observed states. This specific value depends on the exact value of the error in the a priori estimate ϵ_{prior}. In this example, since the prior estimate of the couplings was chosen by adding normalized Gaussian numbers to the exact couplings, its value is exactly $\lambda = 1$. However, in a real application, where we are not sure about the error in the *a priori* estimate, this parameter should be chosen via a cross-validation procedure. In any case, it is crucial to underline that the optimal lambda does not depend on the finite size sampling errors but only on the error in the prior.

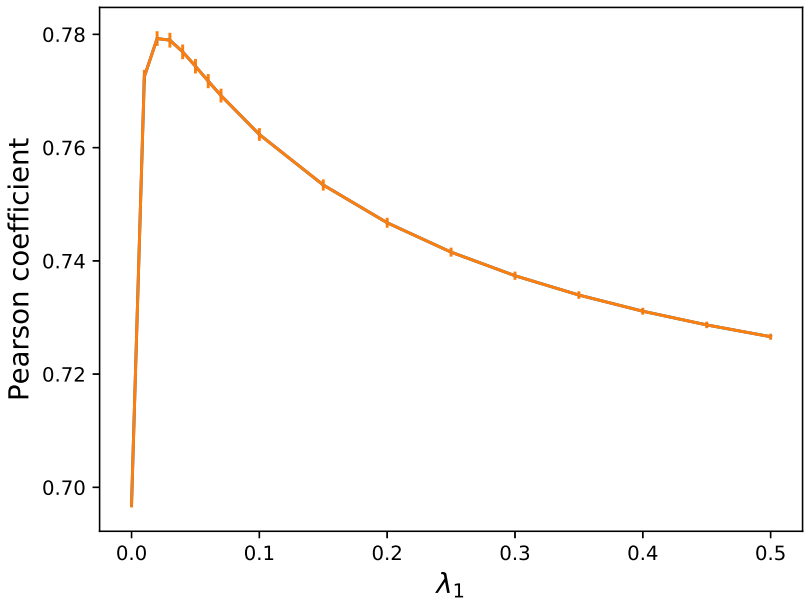
4.1.2 Systematic error

We can then artificially include a systematic error by performing inference on a set of states generated from a distribution *different* from Eq. (4.1). Since the choice of such distribution is arbitrary, we build it to obtain the same correlation with the ground truth solution as the one between the latter and the prior distribution. In

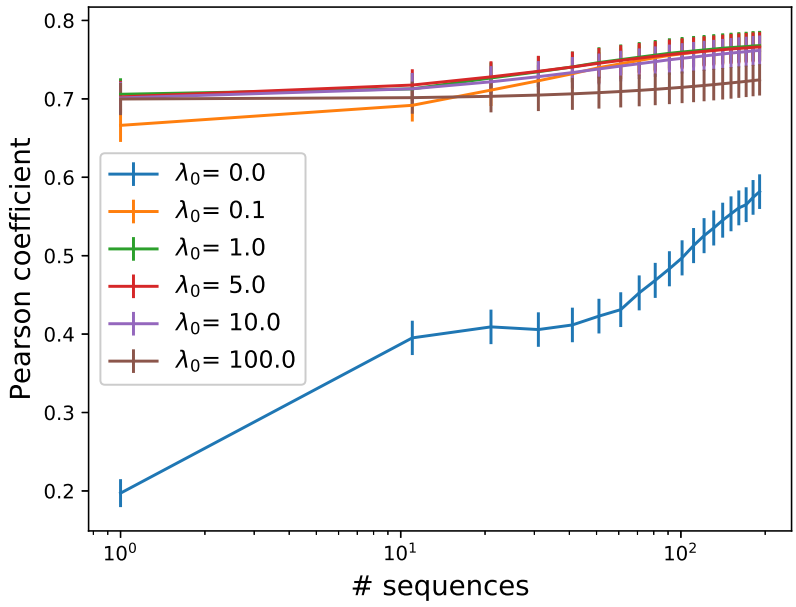
this way, the induced error in the estimate of parameters and the error in the prior contribute similarly to the inference. To this aim, we generate a new set of couplings $J_{err} = \bar{J} + \alpha\mathcal{N}(0,1)$, where α was iteratively adjusted until the Pearson correlation coefficient between the exact couplings $\bar{J}$ and those with the systematic error J_{err} is $\rho_{\bar{J},J_{err}} = \rho_{J,J_{prior}}$.

We then notice that, whereas in presence of purely statistical error, the contribution of the prior should disappear for a large number of observations, in presence of both statistical and systematic error, the contribution of the prior should be retained also for a large number of observations. We thus choose the prior hyper-parameter in the form $\lambda = \lambda_0 + N_s\lambda_1$, where λ_0 accounts for the error in the prior and λ_1 accounts for the systematic error in the observed states.

Since for an infinite number of states $\lambda \approx \lambda_1$, we first optimize the value of λ_1 by performing inference using the pairwise correlations computed directly from the probability in Eq. (4.1), thus removing the contribution of the statistical error. The result is shown in Fig 4.3a where it can be seen that for the chosen parameters the optimal choice is $\lambda_1 = 0.03$. This specific value depends on the magnitude of the artificially introduced systematic error. We then choose the value of λ_0 that maximizes ρ for a finite number of states (Fig 4.3b). We underline that also in this case the optimal choice of the parameters optimizes ρ over the whole range of values of N_s. We also notice that the result shows little dependence on the exact choices of the parameters λ_0 and λ_1, thus indicating that the described procedure is very robust against the choice of these parameters.



(A)



(B)

FIGURE 4.3: Ising system. Pearson correlation coefficient in presence of systematic error. The optimal λ_1 at full sample size (4.3a) is used to find optimal λ_0 in presence of statistical error due to finite sample size (4.3b). Average and standard deviation over 500 realizations of the system.

4.2 DCA including informative prior

In order to show how prior information can be used in a DCA context, we apply it to the prediction of contacts in RNA molecules. The analysis is performed on sequences of 17 riboswitches families classified in the Rfam database (Table 3.1). High-score contacts are compared with those observed in high resolution crystallographic structures. The measure for the accuracy of prediction is the Matthews correlation coefficient (MCC).

In many previous works the *Infernal* algorithm is used to produce MSA (De Leonardis et al., 2015; Weinreb et al., 2016; Wang et al., 2017a; Cuturello, Tiana, and Bussi, 2019). In order to quantify the performance in a blind prediction, we choose to use the *ClustalW* method that doesn't involve any RNA family structural knowledge in the procedure. To this aim, we propose to exploit the maximum a posteriori estimation procedure (Eq. 4.2) to include structural information, provided by a secondary structure prediction algorithm based on thermodynamic parameters, in the DCA couplings learning process.

Other sources of prior knowledge were tested, even though results are not shown in this Thesis. In particular, couplings corresponding to Watson-Crick pairs can be incremented based on the values of the adjacent Watson-Crick couplings:

$$-\log P(\mathbf{J}) \propto \lambda \sum_{ij} \left(\sum_{\{\sigma,\tau\}} J_{ij}(\sigma,\tau)^2 - \frac{1}{N_{WC}} \sum_{\{\sigma,\tau\}\in WC} J_{i-1,j+1}(\sigma,\tau) J_{ij}(\sigma,\tau) \right) \tag{4.5}$$

where N_{WC} is the number of possible Watson-Crick pairs (N_{WC}=6, including wobble GU pairs). Such prior is found to increase the accuracy of secondary structure prediction, but in what follows we decided to focus on the inclusion in the model of external structural information since this last strategy allows to better improve the prediction performance. Moreover, a prior penalizing pairs with a low isostericity score was tested with no success. This could be attributed to the arbitrary choice of such score function, which we constructed as a sum of frequencies restricted to pairs of nucleotides annotated as isosteric in a given interaction family (Leontis, Stombaugh, and Westhof, 2002). For all pairs of positions in the MSA, the isostericity score is eventually the maximum value of such sum among those corresponding to

each type of base pairing.

4.2.1 ViennaRNA

The ViennaRNA package (Hofacker et al., 1994; Mathews et al., 2004; Lorenz et al., 2011) implements an algorithm devoted to secondary structure prediction, taking as an input the single RNA molecule sequence. It is a thermodynamic model (nearest neighbors energy parametrization) solving the RNA folding problem by means of dynamic programming (Nussinov and Jacobson, 1980). The RNAfold program in the package calculates minimum free energy secondary structures and allows for option $-\boldsymbol{p}$ to compute the partition function and base pairing probability matrix p_{ij}. These canonical base pair probabilities can be used to score putative secondary contacts:

$$S_{ij} \equiv p_{ij}^{RNAfold} \tag{4.6}$$

The average MCC is computed after ranking the RNAfold scores (Eq. 4.6) and predicting as interacting pairs those scored above threshold S (Figure 4.4). This is an unusual procedure in ViennaRNA contact prediction context, but it is necessary for consistency with chapter 3. We can see that the accuracy of the method is overall comparable with that of Boltzmann learning DCA on *ClustalW* (Figure 3.5).

It is to notice that the average here is computed on 15 systems instead of 17, two RNA systems being excluded here. This is because for systems PDB:5T83 and PDB:4QLM ViennaRNA gives low accuracy results and RNAfold base pairing probabilities are much lower than for the other systems. This is a sign of lack of portability for the MCC-based score threshold in the ViennaRNA contact prediction context (Table 4.1).

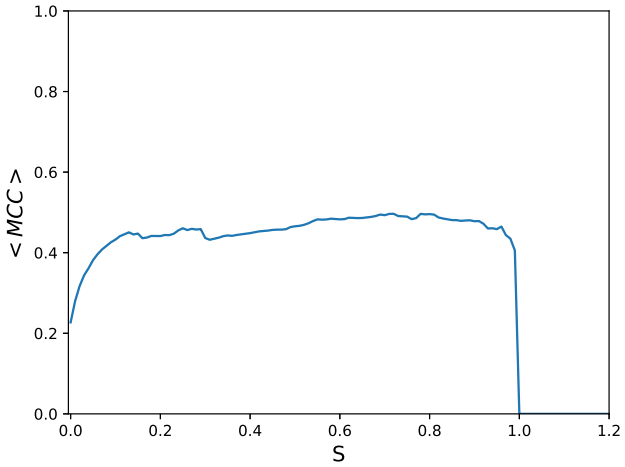


FIGURE 4.4: ViennaRNA package. Average MCC curve over 15 RNA families. Covariance scores S are base pair probabilities obtained via the RNAfold program.

TABLE 4.1: $\overline{MCC}$ with optimal probability threshold $\overline{S}$ for each of 17 RNA families, obtained through cross-validation procedure. Base pairing probabilities are calculated from the RNAfold program available in the ViennaRNA package. For systems PDB:5T83 and PDB:4QLM the MCC is zero for thresholds larger than ≈0.5.

PDB	$\overline{MCC}$	$\overline{S}$
3DOU	0.48	0.72
3F2Q	0.51	0.72
2QBZ	0.51	0.72
2GDI	0.48	0.72
1Y26	0.51	0.71
5T83	-	-
5DDP	0.48	0.78
4XW7	0.52	0.72
4RUM	0.48	0.72
4L81	0.48	0.8
4FRG	0.49	0.8
3SD3	0.49	0.78
2GIS	0.50	0.72
3OWI	0.53	0.71
3IRW	0.48	0.72
4QLM	-	-
3VRS	0.50	0.78

4.2.2 Prior distribution and hyper parameters

In order to enhance the capability of DCA on ClustalW in predicting RNA contacts, we propose the inclusion of structural information provided by ViennaRNA in the parameters learning process (the Boltzmann learning technique discussed in chapter 2). The canonical base pair probabilities p_{ij} can be used to build the following prior distribution:

$$-\log P(\mathbf{J}) \propto \frac{\lambda}{2} \sum_{ij} \sum_{\{\sigma,\tau\}} \frac{J_{ij}(\sigma,\tau)^2}{(p_{ij}^{RNAfold})^2 + \epsilon} \tag{4.7}$$

This function penalizes couplings corresponding to low RNAfold probability pairs and thus it differs from the usually adopted l_2-regularization term, which doesn't carry any external information in the Gaussian variance. Pseudo-count ϵ in the denominator (ϵ=0.05) is used to avoid infinite penalization on couplings corresponding to a null base pair probability in ViennaRNA.

As for the Ising model, hyper parameter $\lambda = \lambda_0 + N_S\lambda_1$ accounts for both the statistical error due to finite number of sequences (λ_0) and the systematic error due to mistakes in the alignments (λ_1). Such parameters are searched through an iterative procedure. In the first step, we look for the value of λ_1 maximizing the average MCC using all the available sequences and under the assumption of a negligible statistical error (Figure 4.5). Adopting the optimized λ_1, a sample of possible values of λ_0 is scanned including the dependence of λ on N_s, so to find the optimal value (Figure 4.6). In this second round, the regularization term lambda has thus a different weight on families of different size. The obtained λ_0 is in turn used to search for possible values of λ_1 further improving the accuracy of predictions. The procedure stops after this iteration since the optimal λ_1 equals that found in the first step of the optimization. Values of hyper parameters maximizing the accuracy are $\lambda_0 = 10$ and $\lambda_1 = 0.02$, validated for all system through leave-one-out procedure.

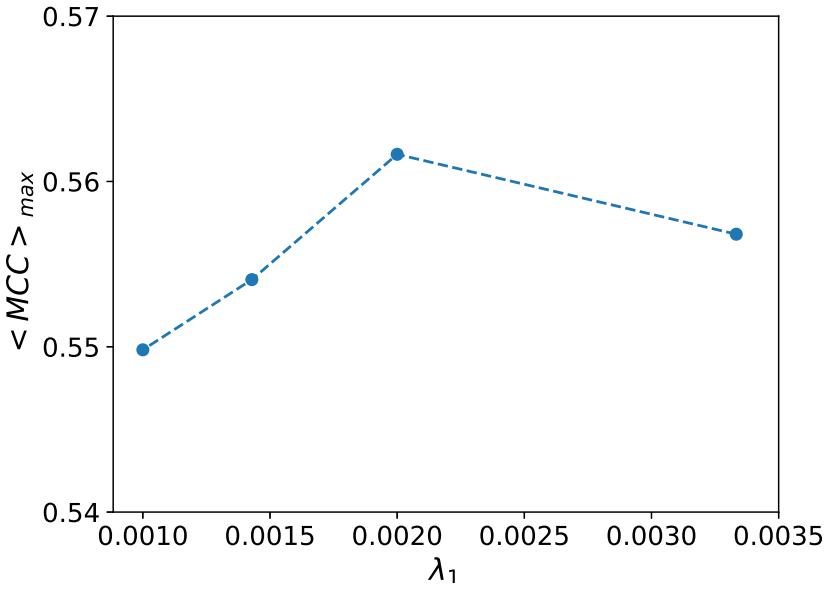


FIGURE 4.5: First step of iterative grid searching of prior hyper parameter λ_1: maximum average MCC for different λ_1 values under the approximation of null statistical error (λ_0=0). Optimal value is λ_1=0.002. *ClustalW* alignments.

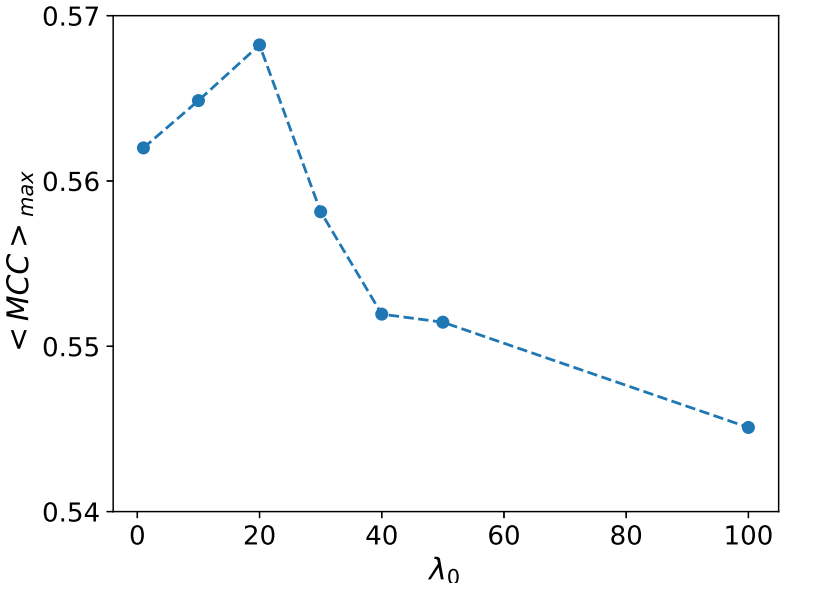


FIGURE 4.6: Second step of iterative grid searching of prior hyper parameter λ_0: maximum average MCC for different λ_0 values at fixed $\lambda_1 = 0.002$. Optimal value is λ_0=20. *ClustalW* alignments.

4.2.3 RNA contact prediction

The posterior distribution obtained regularizing the DCA likelihood through the ViennaRNA Gaussian prior can be maximized via the Boltzmann learning algorithm, and its capability to infer the correct contacts can be compared with both pure DCA and standard ViennaRNA. Prior hyperparameters are set to their optimal values derived in 4.2.2 (λ_0=10 and λ_1=0.02). It is interesting to look at the accuracy of predictions when artificially reducing the sequence sample size. In Figure 4.7 the maximum average MCC is shown at various percentages of sequences randomly extracted from the full MSAs.

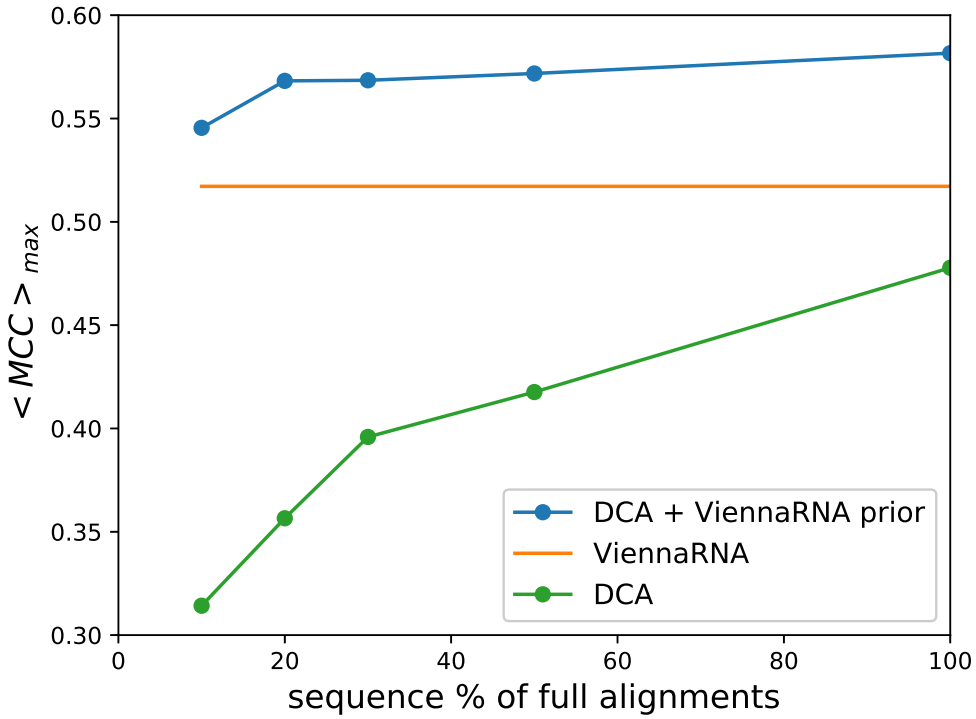


FIGURE 4.7: Maximum average MCC. Comparison of contact prediction accuracies resulting from DCA, ViennaRNA and DCA including ViennaRNA prior for various percentages of sequences randomly extracted from the full MSAs. *ClustalW* alignments.

The accuracy resulting from the inclusion of the Vienna prior in the DCA inference process outperforms on average both DCA and ViennaRNA methods. This is not obvious, since the former consists of a combination of results obtained via the other two methods. Even if the sequence sample is very small, the posterior maximization method is able to correct for the error in DCA induced by undersampling

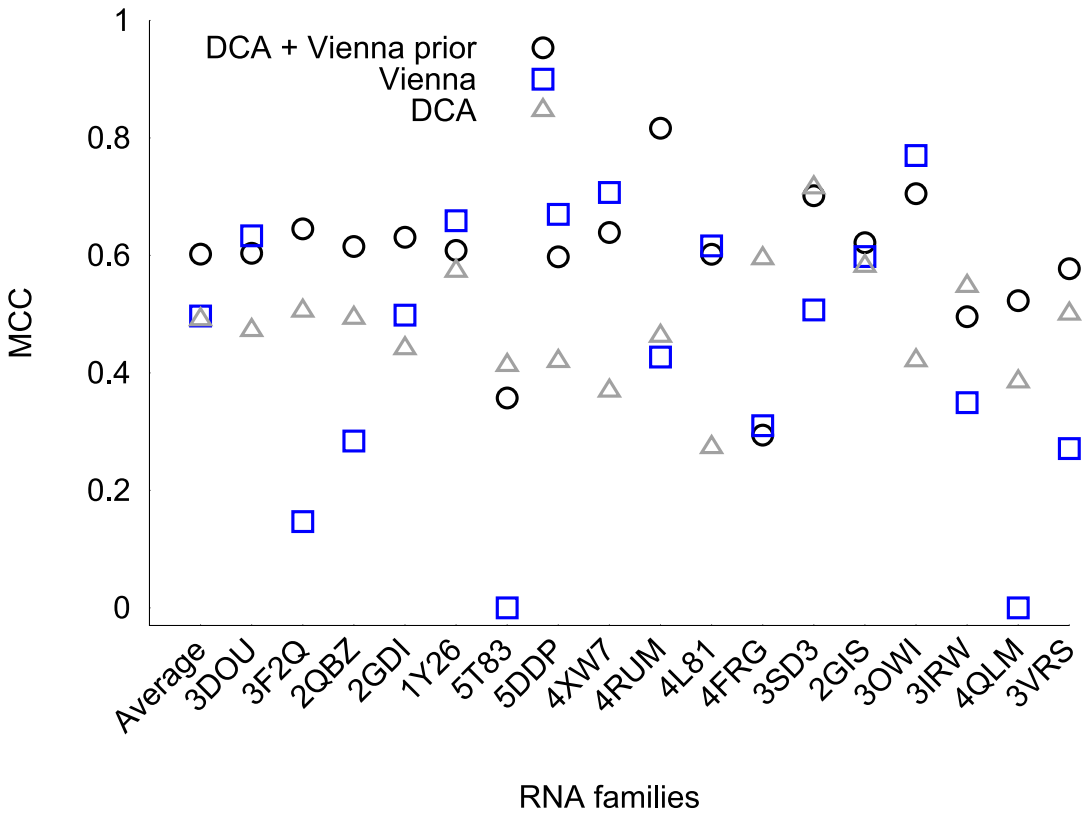


FIGURE 4.8: Average MCC obtained via DCA, ViennaRNA, and DCA including the ViennaRNA prior for 17 RNA families at optimal threshold score obtained through cross-validation procedure. Families are labeled using the PDB code of the representative crystallographic structure. Average $\overline{MCC}$ is reported in the first column. *ClustalW* alignments.

of sequences, while still taking the effect of co-evolution into account. At small sample sizes the performance of DCA alone drops dramatically due to not sufficient statistics, while the adoption of the prior in DCA is able to cure for the introduced statistical error relying on ViennaRNA predictions, but still outperforming also this latter method.

Figure 4.8 shows the MCC of the 17 RNA families corresponding to predictions at optimal threshold from the three methods on the full MSAs. The values are more scattered for ViennaRNA method, ranging between 0 and 0.8, due to a less portable score threshold for this method. The MCC of ViennaRNA is the highest in 5 cases, while pure DCA outperforms the other methods in 3 cases; the posterior maximization procedure reports the highest MCC on 9 systems.

It is interesting to investigate how the two components of the MCC, precision and sensitivity, contribute to its value. We compare the average sensitivity, precision

and sensitivity to contacts in secondary structures at optimal threshold of pure DCA, DCA combined with the ViennaRNA prior and ViennaRNA (Figure 4.9). While precision is high and very similar for all methods (almost 0.8), including the prior causes a significant improvement in sensitivity, and in particular in the sensitivity to contacts in stems when compared to DCA. The reason is that ViennaRNA is a tool designed to detect contacts that are formed in RNA secondary structures and does not give any information about base pairs in tertiary structures.

Overall, via the Boltzmann learning algorithm we are able to incorporate in the direct coupling analysis external information about the possible base pairs in the structured molecule, while still retaining information arising from co-evolutionary signals. The performance obtained through maximization of the posterior are comparable with results of DCA on the high quality but structure based *Infernal* alignments. Even though *Infernal* alignments can serve as a reliable benchmark for assessing the capability of predictions from co-variance models, we are aware that the inclusion in the alignment procedure of structure knowledge (when available) poses prohibitive limits for prediction on RNA families for which no experimental structure is available. For this reason, we consider the improvement of the DCA performance on *ClustalW* obtained via the maximum posterior estimation procedure worthy of future investigations.

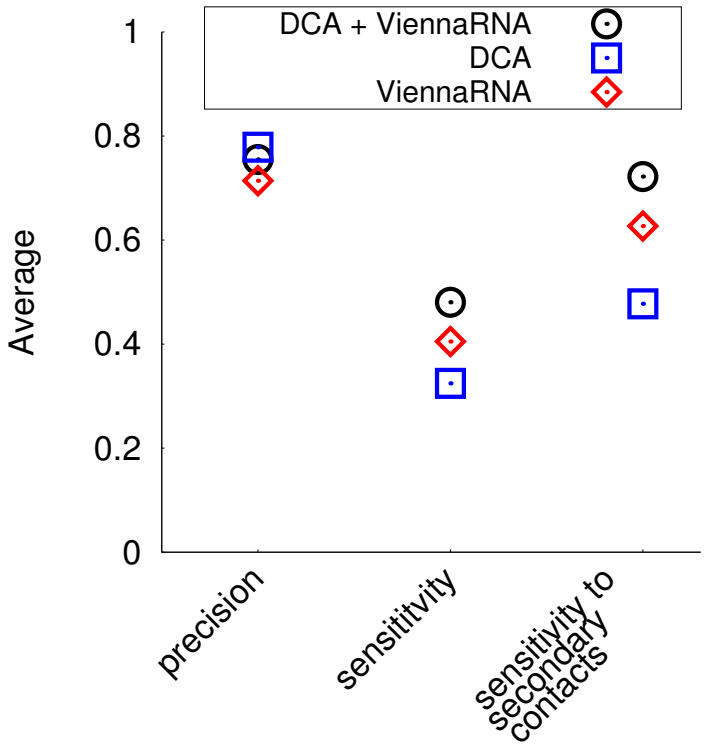


FIGURE 4.9: Average sensitivity, precision and sensitivity to contacts in stems at optimal threshold for DCA and DCA+ViennaRNA prior (average over 17 systems) and ViennaRNA (average over 15 systems). *ClustalW* alignments.

Chapter 5

Conclusion

In this thesis I discussed several methodological issues related to the prediction of contacts in RNA molecules based on co-evolutionary information. In particular, in Chapter 2 I introduced the theory of covariance models, with a particular focus on direct coupling analysis and the Boltzmann learning algorithm that I developed. Results of the application of these methods to RNA contact prediction are shown in Chapter 3 for a set of 17 riboswitches families. Among the tested methods, the Boltzmann learning approach is the one that allows to simultaneously maximize accuracy and precision on the considered data set. If one also includes the cost of a later 3D structure prediction and refinement, the extra computational time required by such algorithm can be considered as absolutely worth. Moreover, the fast Boltzmann learning procedure introduced here, based on a stochastic gradient descent devoted to the minimization of the exact negative log-likelihood, could be productively used in protein systems as well.

As shown in Chapter 3, the procedure employed to align homologous sequences has a significant impact on the accuracy of the prediction. Results show that *Infernal* algorithm, based on experimental secondary structures or on a previous prediction, performs significantly better than the *ClustalW* algorithm. This observation suggests that *ClustalW* deserves a particular attention, since it is an essential tool to quantify the performance in a blind prediction. A possible perspective is to use the couplings obtained with the DCA approach to further refine the multiple sequence alignments. In particular, once a putative Potts Hamiltonian has been found, one might try to reposition the gaps in order to minimize the total energy of the sequence. A possible

strategy is to shift gaps along the sequences through Monte Carlo moves. Parallelizing the procedure, different starting point of the process can be generated and the alignment in which sequences take the highest average probability can be selected. This procedure is however very expensive, since it requires an exhaustive exploration of the space of possible moves, and preliminary attempts were not found to improve the accuracy of contact prediction.

In Chapter 4 I introduced a method based on maximum a posteriori estimation procedure with an informative prior. I used this method to exploit structural information in the DCA parameters learning process, with the aim of improving the prediction performance on *ClustalW* alignments. The Boltzmann learning technique is adopted to include prior knowledge about the couplings, provided by ViennaRNA secondary structure prediction algorithm. The prior hyper-parameters are easy to choose and their optimal values are portable across different RNA families. The method allows to improve the accuracy of contact prediction decreasing both systematic error, possibly due to alignment mistakes, and statistical error, due to finite alignments size. The introduced formalism opens the way to the possibility of including prior information about isostericity matrices in RNA contact prediction, other secondary structure prediction algorithms for proteins, or even three-dimensional modeling tools both at the coarse-grained or at the atomistic level.

Bibliography

Ackley, David H, Geoffrey E Hinton, and Terrence J Sejnowski (1987). "A learning algorithm for Boltzmann machines". In: *Readings in Computer Vision*. Elsevier, pp. 522–533.

Arnold, Barry C and David Strauss (1991). "Pseudolikelihood estimation: some examples". In: *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 233–243.

Aurell, Erik and Magnus Ekeberg (2012). "Inverse Ising inference using all the data". In: *Physical review letters* 108.9, p. 090201.

Baldassi, Carlo et al. (2014). "Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners". In: *PloS one* 9.3, e92721.

Barrat-Charlaix, Pierre, Matteo Figliuzzi, and Martin Weigt (2016). "Improving landscape inference by integrating heterogeneous data in the inverse Ising problem". In: *Sci. Rep.* 6, p. 37812.

Belalov, Ilya S and Alexander N Lukashev (2013). "Causes and implications of codon usage bias in RNA viruses". In: *PLoS One* 8.2, e56642.

Cesari, Andrea, Sabine Reißer, and Giovanni Bussi (2018). "Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments". In: *Computation* 6.1, p. 15.

Chen, Zhen et al. (2011). "Integrating molecular dynamics and co-evolutionary analysis for reliable target prediction and deregulation of the allosteric inhibition of aspartokinase for amino acid production". In: *Journal of biotechnology* 154.4, pp. 248–254.

Cocco, Simona, Remi Monasson, and Martin Weigt (2013). "From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction". In: *PLoS Comput. Biol.* 9.8, e1003176.

Cocco, Simona et al. (2018). "Inverse statistical physics of protein sequences: A key issues review". In: *Rep. Prog. Phys.* 81.3, p. 032601.

Cuturello, Francesca, Guido Tiana, and Giovanni Bussi (2019). "Assessing the accuracy of direct-coupling analysis for RNA contact prediction". In: *arXiv preprint arXiv:1812.07630v3*.

Darken, Christian and John Moody (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. Denver, Colorado, USA: Morgan Kaufmann Publishers Inc., pp. 832–838. ISBN: 1-55860-184-8. URL: http://dl.acm.org/citation.cfm?id=118850.119956.

De Leonardis, Eleonora et al. (2015). "Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction". In: *Nucleic Acids Res.* 43.21, pp. 10444–10455.

Dunn, Stanley D, Lindi M Wahl, and Gregory B Gloor (2007). "Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction". In: *Bioinformatics* 24.3, pp. 333–340.

Dutheil, Julien Y, Fabrice Jossinet, and Eric Westhof (2010). "Base pairing constraints drive structural epistasis in ribosomal RNA sequences". In: *Mol. Biol. Evol.* 27.8, pp. 1868–1876.

Eddy, Sean R and Richard Durbin (1994). "RNA sequence analysis using covariance models". In: *Nucleic Acids Res.* 22.11, pp. 2079–2088.

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

Ekeberg, Magnus et al. (2013). "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models". In: *Phys. Rev. E* 87.1, p. 012707.

Figliuzzi, Matteo, Pierre Barrat-Charlaix, and Martin Weigt (2018). "How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins?" In: *Mol. Biol. Evol.* 35.4, pp. 1018–1027.

Gao, Chen-Yi, Hai-Jun Zhou, and Erik Aurell (2018). "Correlation-compressed direct-coupling analysis". In: *Physical Review E* 98.3, p. 032407.

Gorodkin, Jan, Shawn Stricklin, and Gary Stormo (2001). "Discovering common stem-loop motifs in unaligned RNA sequences." In: *Nucleic Acids Res.* 29 10, pp. 2135–44.

Haldane, Allan et al. (2016). "Structural propensities of kinase family proteins from a Potts model of residue co-variation". In: *Protein Science* 25.8, pp. 1378–1384.

Hofacker, Ivo L et al. (1994). "Fast folding and comparison of RNA secondary structures". In: *Monatshefte fur Chemie/Chemical Monthly* 125.2, pp. 167–188.

Hon, Chung-Chau et al. (2017). "An atlas of human long non-coding RNAs with accurate 5′ ends". In: *Nature* 543.7644, p. 199.

Jones, David T et al. (2011). "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments". In: *Bioinformatics* 28.2, pp. 184–190.

Leontis, Neocles B, Jesse Stombaugh, and Eric Westhof (2002). "The non-Watson–Crick base pairs and their associated isostericity matrices". In: *Nucleic Acids Res.* 30.16, pp. 3497–3531.

Leontis, Neocles B and Eric Westhof (2001). "Geometric nomenclature and classification of RNA base pairs". In: *RNA* 7.4, pp. 499–512.

Lezon, Timothy R et al. (2006). "Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns". In: *Proceedings of the National Academy of Sciences* 103.50, pp. 19033–19038.

Lorenz, Ronny et al. (2011). "ViennaRNA Package 2.0". In: *Algorithms Mol. Biol.* 6.1, p. 26.

Lu, Xiang-Jun, Harmen J. Bussemaker, and Wilma K. Olson (2015). "DSSR: an integrated software tool for dissecting the spatial structure of RNA". In: *Nucleic Acids Res.* 43.21, e142.

Ma, Jianzhu et al. (2015). "Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning". In: *Bioinformatics* 31.21, pp. 3506–3513.

Marks, Debora S et al. (2011). "Protein 3D structure computed from evolutionary sequence variation". In: *PLoS ONE* 6.12, e28766.

Marruzzo, Alessia et al. (2017). "Improved pseudolikelihood regularization and decimation methods on non-linearly interacting systems with continuous variables". In: *arXiv preprint arXiv:1708.00787*.

Mathews, David H, Douglas H Turner, and Richard M Watson (2016). "RNA secondary structure prediction". In: *Curr. Protoc. Nucleic Acid Chem.* 67.1, pp. 11–2.

Mathews, David H et al. (2004). "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure". In: *Proc. Natl. Acad. Sci. USA* 101.19, pp. 7287–7292.

Matthews, Brian W (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *BBA-Prot. Struct.* 405.2, pp. 442–451.

Miao, Zhichao et al. (2017). "RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme". In: *RNA* 23.5, pp. 655–672.

Morcos, Faruck et al. (2011). "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". In: *Proc. Natl. Acad. Sci. U. S. A.* 108.49, E1293–E1301.

Morris, Kevin V and John S Mattick (2014). "The rise of regulatory RNA". In: *Nat. Rev. Genet.* 15.6, p. 423.

Nawrocki, Eric (2009). "Structural RNA homology search and alignment using covariance models". In:

Nawrocki, Eric P and Sean R Eddy (2013). "Infernal 1.1: 100-fold faster RNA homology searches". In: *Bioinformatics* 29.22, pp. 2933–2935.

Nawrocki, Eric P et al. (2014). "Rfam 12.0: updates to the RNA families database". In: *Nucleic Acids Res.* 43.D1, pp. D130–D137.

Nguyen, H Chau, Riccardo Zecchina, and Johannes Berg (2017). "Inverse statistical problems: from the inverse Ising problem to data science". In: *Advances in Physics* 66.3, pp. 197–261.

Nussinov, Ruth and Ann B Jacobson (1980). "Fast algorithm for predicting the secondary structure of single-stranded RNA". In: *Proceedings of the National Academy of Sciences* 77.11, pp. 6309–6313.

Pang, Phillip S et al. (2005). "Prediction of functional tertiary interactions and intermolecular interfaces from primary sequence data". In: *J. Exp. Zool. B Mol. Dev. Evol.* 304.1, pp. 50–63.

Parisien, Marc et al. (2009). "New metrics for comparing and assessing discrepancies between RNA 3D structures and models". In: *RNA* 15.10, pp. 1875–1885.

Ravikumar, Pradeep, Martin J Wainwright, John D Lafferty, et al. (2010). "High-dimensional Ising model selection using 1-regularized logistic regression". In: *The Annals of Statistics* 38.3, pp. 1287–1319.

Rinnenthal, Jorg et al. (2011). "Mapping the landscape of RNA dynamics with NMR spectroscopy". In: *Acc. Chem. Res.* 44.12, pp. 1292–1301.

Rivas, Elena, Jody Clements, and Sean R Eddy (2017). "A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs". In: *Nature Methods* 14.1, p. 45.

Schug, Alexander et al. (2009). "High-resolution protein complexes from integrating genomic information with molecular simulation". In: *Proceedings of the National Academy of Sciences* 106.52, pp. 22124–22129.

Smith, Martin A et al. (2013). "Widespread purifying selection on RNA structure in mammals". In: *Nucleic Acids Res.* 41.17, pp. 8220–8236.

Šponer, Jiri et al. (2018). "RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview". In: *Chem. Rev.* 118.8, pp. 4177–4338.

Stombaugh, Jesse et al. (2009). "Frequency and isostericity of RNA base pairs". In: *Nucleic Acids Res.* 37.7, pp. 2294–2312.

Sutto, Ludovico et al. (2015). "From residue coevolution to protein conformational ensembles and functional dynamics". In: *Proc. Natl. Acad. Sci. U. S. A.* 112.44, pp. 13567–13572.

Thompson, Julie D, Desmond G Higgins, and Toby J Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". In: *Nucleic Acids Res.* 22.22, pp. 4673–4680.

Tyagi, Payal et al. (2016). "Regularization and decimation pseudolikelihood approaches to statistical inference in X Y spin models". In: *Physical Review B* 94.2, p. 024203.

Wang, Jian et al. (2017a). "Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide–nucleotide interactions from direct coupling analysis". In: *Nucleic Acids Res.* 45.11, pp. 6299–6309.

Wang, Sheng et al. (2017b). "Accurate de novo prediction of protein contact map by ultra-deep learning model". In: *PLoS computational biology* 13.1, e1005324.

Weeks, Kevin M (2010). "Advances in RNA structure analysis by chemical probing". In: *Curr. Opin. Struct. Biol.* 20.3, pp. 295–304.

Weigt, Martin et al. (2009). "Identification of direct residue contacts in protein–protein interaction by message passing". In: *Proc. Natl. Acad. Sci. U. S. A.* 106.1, pp. 67–72.

Weinreb, Caleb et al. (2016). "3D RNA and functional interactions from evolutionary couplings". In: *Cell* 165.4, pp. 963–975.

Westhof, Eric (2015). "Twenty years of RNA crystallography". In: *RNA* 21.4, pp. 486–487.

Yao, Yuan, Lorenzo Rosasco, and Andrea Caponnetto (2007). "On early stopping in gradient descent learning". In: *Constructive Approximation* 26.2, pp. 289–315.

Zhu, Jun, Ning Chen, and Eric P Xing (2014). "Bayesian inference with posterior regularization and applications to infinite latent SVMs". In: *The Journal of Machine Learning Research* 15.1, pp. 1799–1847.