

# Unspecific binding but specific disruption of the group I intron by the StpA chaperone

Vladimir Reinharz and Tsvi Tlusty\*

*Center for Soft and Living Matter,  
Institute for Basic Science,  
Ulsan, Republic of Korea*

(Dated: October 10, 2019)

Chaperone proteins — the most disordered among all protein groups — help RNAs fold into their functional structure by destabilizing misfolded configurations or stabilizing the functional ones. But disentangling the mechanism underlying RNA chaperoning is challenging, mostly due to inherent disorder of the chaperones and the transient nature of their interactions with RNA. In particular, it is unclear how specific the interactions are and what role is played by amino acid charge and polarity patterns. Here, we address these questions in the RNA chaperone StpA. By adapting direct coupling analysis (DCA) to treat in tandem sequences written in two alphabets, nucleotides and amino acids, we could analyze StpA-RNA interactions and identify a two-pronged mechanism: StpA disrupts *specific* positions in the group I intron while *globally* and loosely binding to the entire structure. Moreover, the interaction is governed by the charge pattern: negatively charged regions in the destabilizing StpA N-terminal affect a few specific positions in the RNA, located in stems and in the pseudoknot. In contrast, positive regions in the C-terminal contain strongly coupled amino acids that promote non-specific or weakly-specific binding to the RNA. The present study opens new avenues to examine the functions of disordered proteins and to design disruptive proteins based on their charge patterns.

arXiv:1911.03046v1 [q-bio.BM] 8 Nov 2019

---

\* tsvitlusty@gmail.com

## I. INTRODUCTION

There is mounting evidence for the existence of Intrinsically Disordered Proteins (IDPs) that lack specific structures (Babu *et al.*, 2012). These proteins do not fold into a well-defined conformation (Wright and Dyson, 2015), although some may acquire a specific structure given the right context. IDPs are at the core of key biological assemblies and processes, such as membrane-less organelles (Nott *et al.*, 2015), cell signaling (Wright and Dyson, 2015), and cell division (Buske and Levin, 2013). Disordered regions may exert entropic forces on the proteins they bind and thereby shift the ensemble of protein structures towards one with higher binding affinity (Keul *et al.*, 2018). While our repertoire of IDPs is steadily growing (Piovesan *et al.*, 2017; Schad *et al.*, 2017; Varadi and Tompa, 2015), the function of most is yet to be discovered (Papaleo *et al.*, 2016; Van Der Lee *et al.*, 2014). Nevertheless, analysis suggests that a crucial determinant of the global shape and function of IDPs is their charge pattern (Das *et al.*, 2015).

A prominent class of IDPs is that of chaperones whose fraction of disordered residues, 54% on average, is the highest among all functional classes of proteins (Tompa and Csermely, 2004). A particularly important subclass are those that chaperone RNA folding: To perform their functions, non-coding RNAs rely on well-conserved structures, which have been used for sequence alignment and putative RNA prediction (Nawrocki and Eddy, 2013). While some non-coding RNAs are able to attain those structures by themselves, chaperone proteins are essential in stabilizing correct conformations or in destabilizing, and thus rescuing, misfolded RNAs (Bhaskaran and Russell, 2007; Papasaikas and Valcárcel, 2016; Woodson, 2010).

A prime example of chaperone-dependent RNA is the Group I Intron (GII), which has an elaborate functional structure (Michel and Westhof, 1990). Two chaperones take part in the folding of this RNA. One is the Cyt-18 protein that stabilizes the active structure (Guo and Lambowitz, 1992; Mohr *et al.*, 1992). The second chaperone is the StpA protein, which is known to destabilize misfolded GII structure (Mayer *et al.*, 2007; Waldsich, 2002). The structures of Cyt-18 and its complex with the GII are well-determined (Paukstelis *et al.*, 2008). In contrast, most of the StpA protein, 73% of the residues, is known to be disordered. StpA consists of two domains, the N-terminal and C-terminal. Excising the C-terminal from the sequence increases the efficacy of the chaperone, while mutations in the C-terminal hinder its binding capacity (Mayer *et al.*, 2007). An entropy transfer model has been proposed, where rapid and transient binding disturbs the structure, thus allowing it to refold (Tompa and Csermely, 2004). But many questions regarding the specifics of the destabilization function remain open. An inherent obstacle in understanding the mechanisms of disordered proteins, such as StpA, is the lack of functional structure. The StpA-GII problem is even more challenging since the other partner in the interaction, the GII RNA, is misfolded and therefore lacks a specific structure as well.

To overcome the lack of structures, one may leverage the accelerated growth in the number of known sequences and use them for multiple sequence alignments (MSA). As of August 2018, GenBank (Sayers *et al.*, 2019) had sequences totaling over  $3.7 \times 10^{12}$  nucleotides from 420 000 species, an increase of 40% from the previous year. Techniques such as Direct Coupling Analysis (DCA) extract from the MSAs amino acid contacts and 3D structures (Burger and Van Nimwegen, 2010; Marks *et al.*, 2011; Ovchinnikov *et al.*, 2015), protein-protein interaction sites (Morcos *et al.*, 2011; Ovchinnikov *et al.*, 2014), RNA ligand binding pockets (Reinharz *et al.*, 2016), RNA tertiary contacts (De Leonardis *et al.*, 2015), and RNA-protein interaction sites (Weinreb *et al.*, 2016). These studies have also demonstrated that many IDPs have strong correlations, hinting at context dependant structures (Toth-Petroczy *et al.*, 2016), though in the last study StpA did not exhibit any particular structure. So far, however, IDP-RNA interactions — which are essential in many molecular systems, in particular chaperones — have not been examined, perhaps due to the difficulty of analyzing the interaction of two objects that lack defined structures and whose sequences are written in different alphabets.

All this motivates the present study in which we adapt the DCA method to concurrently process proteins and RNAs, which not only differ in the size of their alphabets but, on top of that, have high variability in

sequence conservation. The adaptive method, termed  $\alpha\beta$ DCA, produces the first analysis of the interaction of a disordered protein, StpA, with a non-coding RNA, the group I intron. Our method identifies 90 strongly coupled pairs between StpA and GII. We find that the charge pattern governs the type of interactions: The N-terminal of StpA destabilizes the RNA by a few specific interactions among negatively charged regions of the protein and regions of the GII, which are critically misfolded in the structures ensemble or impede functional loops from forming. In the C-terminal, strongly coupled amino acids are mostly in positively charged regions, and their interaction of these amino acids with the RNA is weakly specific and almost uniformly distributed over the entire GII sequence. Moreover, while both terminals are of roughly the same length, only 21% of the top DCA scores are in the N-terminal. These findings propose a charge-dependent two-pronged mechanism of unspecific binding but specific disruption by chaperone IDPs.

## II. RESULTS

We extended the classic mean-field approximation DCA (mfDCA) method for treating paired sequences that are written in different alphabets and have different levels of sequence conservation (for details see Sec. IV). First, we tested this simple DCA variant – which we call  $\alpha\beta$ DCA (for treating varying alphabets) – against two other DCA implementations: Gremlin, an implementation of Markov random-field DCA (Ovchinnikov *et al.*, 2014), and EVcouplings (Hopf *et al.*, 2018), an implementation of pseudo-likelihood DCA (plmDCA). For the benchmark of the 5S–RL18 ribosomal complex, the adaptive  $\alpha\beta$ DCA method predicts more contacts in its top scores (see Sec. IV.E). Additionally, we observe that the mfDCA method outperforms Gremlin in the GII alignment, most probably owing to the correct pseudo-count for a 5-letter alphabet, rather than that of the 21-letter alphabet of proteins used in Gremlin. In the following, we apply the  $\alpha\beta$ DCA method to analyze the StpA–GII alignment (the code and alignment are available at: <https://github.com/vreinharz/ProtRNADCA>).

### A. $\alpha\beta$ DCA exhibits significant scores for strongly-coupled StpA–RNA contacts

The DCA method identifies strong couplings, indicating significant physical interactions. These significant scores emerge as outliers departing from the bulk distribution of the DCA scores. Sequence conservation is a critical factor, as too high conservation level prohibits co-evolution analysis. Fig. 1a shows the secondary structure of the GII RNA together with its long range interactions and sequence conservation values (the overall maximal conservation is shown in Fig. S1). As a test for how predictive the DCA is, we compare the distribution of  $\alpha\beta$ DCA scores from the StpA–GII alignment with those obtained from the same alignment but with shuffled sequences. The scores of the original alignment spread over a much wider range than that of the shuffled alignment, thus confirming that the DCA analysis is predictive (Fig. S2).

The StpA–GII amino acid–nucleotide pairs with the strongest DCA couplings are shown in Fig. 1b. The significant pairs are defined as those deviating by more than  $4\sigma$  from the average of the bulk distribution. The  $\alpha\beta$ DCA identifies 90 significant pairs, 15% less than those extracted by the standard DCA, which disregards the difference in alphabet and sequence similarity between the RNAs and the proteins.

### B. Inferred protein–RNA interactions are selective in the N-terminal and global in the C-terminal of StpA

The N- and C-terminals of StpA are known to interact differently with the RNA (Waldsich, 2002). This motivates us to characterize the number and distribution of high  $\alpha\beta$ DCA scores, which indicate strong physical couplings, in each of these two regions. Since RNA structures fluctuate within a dynamic ensemble (McCaskill, 1990), we examine the interactions in light of the two main structure ensembles and the functional structure, and in particular link the distribution of strong couplings along the RNA.

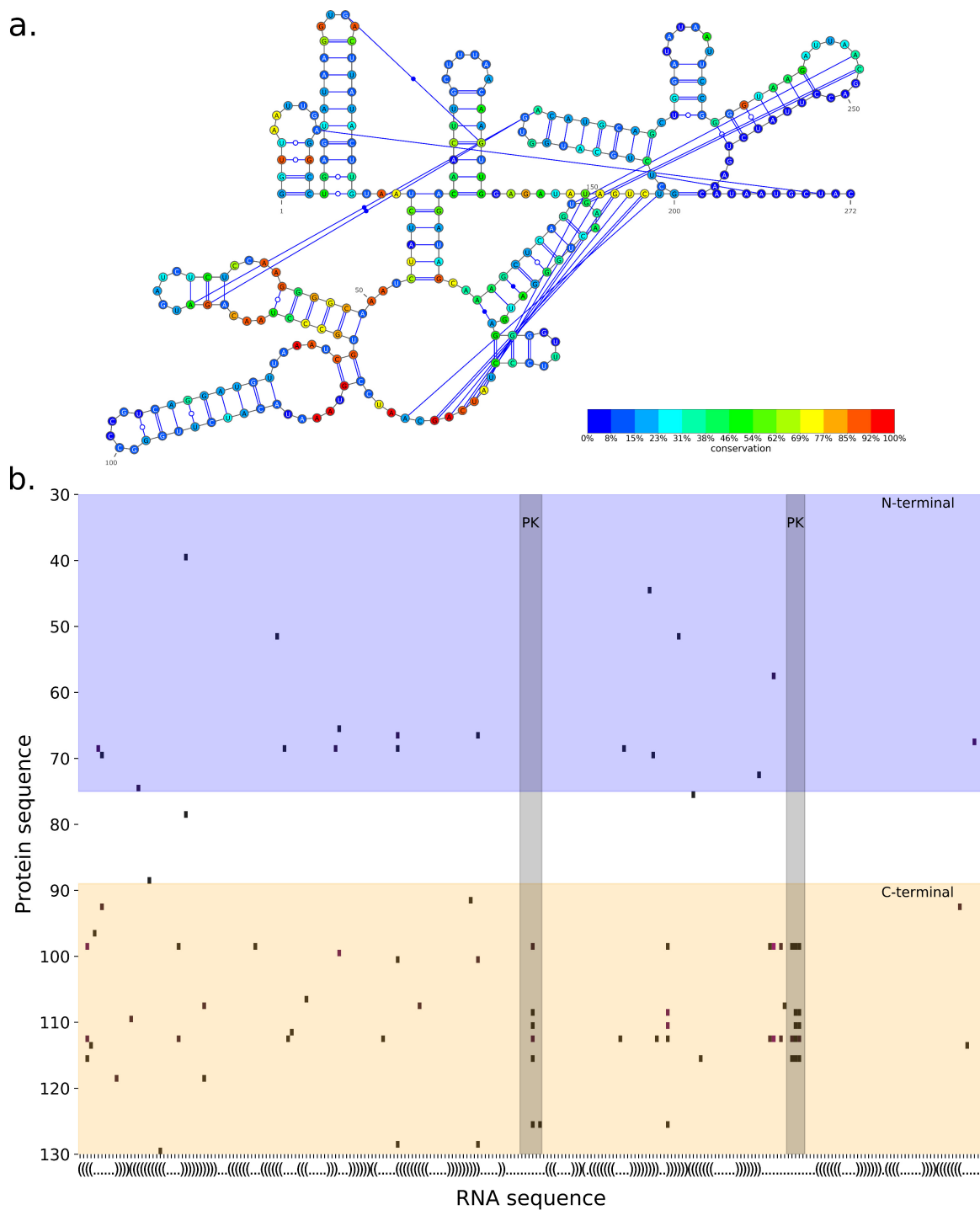


FIG. 1: **a.** GII secondary structure and its sequence conservation. The last 25 positions have no conservation levels since they are excluded from the alignment (see Sec. IV.C). **b.** Positions of significant  $\alpha\beta$ DCA scores ( $\geq 4\sigma$  above average) between the StpA protein (vertical axis) and the GII RNA (horizontal axis). The RNA axis is labeled with the secondary structure in parentheses notation. The blue region is the N-terminal of StpA and the orange region its C-terminal. The pale grey regions are the pseudoknot (PK) of GII. The last 25 positions of GII are omitted (See Sec. IV.C).

To this end, we plot in Fig. 2 the net charge distribution along the StpA protein (averaged over a window of 5 amino acids), above the two main clusters of the GII RNA structure ensemble as predicted by **RNAstructure** (Reuter and Mathews, 2010). The arcs in the upper part depict bonds in the main cluster, whose probability is 68.2%, and the arcs in the lower part show bonds in the second main cluster, of probability 31.8% (Aalberts and Jannen, 2013). The red discs represent stems in the functional structure that are absent from both ensembles, in particular the pseudoknot. Note that while pseudoknots cannot be predicted with **RNAstructure**, they could not be inferred even with **RNAPKplex**, which was designed for this purpose (Lorenz *et al.*, 2011).

The significant scores between the RNA and the protein are denoted by lines. There are 90 significant scores ( $\geq 4\sigma$ ) between the protein and the RNA: 19 in negative regions of the N-terminal (dark lines), 69 in mostly positive regions of the C-terminal (light gray lines), and 2 in the linker between the N and C-terminals (dashed lines).

### C. High scores correspond to close nucleotides in the 3D structure of GII

To check whether the RNA alignment is informative by itself, we examine the DCA scores among all pairs of RNA positions. To validate the quality of the RNA alignment, we compared the physical contacts predicted by DCA to the 3D structure of the *td* GII RNA (available at <http://www-ibmc.u-strasbg.fr/spip-arn/spip.php?rubrique136>). We computed DCA scores using two methods, the mean-field approximation (mfDCA) and Gremlin (Ovchinnikov *et al.*, 2014). We note that  $\alpha\beta$ DCA is identical to mfDCA when treating a single alphabet. We consider as a good prediction a pair of nucleotides closer than 8Å in the 3D structure. Fig. 3 shows the number of these true positives (distance  $< 8\text{Å}$ ) for the hundred top scores. While the first 40 top scores are well predicted by both methods, the Gremlin method is outperformed by mfDCA in the next 60 scores.

## III. DISCUSSION

The StpA protein destabilizes the misfolded GII RNA, allowing it to achieve its functional structure. Experiments have shown that the binding is transient and weak, with little specificity (Doetsch *et al.*, 2011; Waldsich, 2002). Mutation studies provide evidence for GII-StpA interactions: Mutations in the StpA C-terminal reduce the binding affinity between StpA and the group I intron, while complete deletion of the C-terminal increases the efficiency of StpA as a chaperone (Waldsich, 2002). A C-terminal mutation, glycine 126 changed to valine, weakens the binding and increases the efficiency of StpA (Mayer *et al.*, 2007). In the following, we further expand the understanding and provide a detailed picture of the GII-StpA mechanism, based on the  $\alpha\beta$ DCA results, which identify the coupled amino acids and nucleotides responsible for both binding and destabilizing interactions.

### A. Binding is mediated by positively charged regions of StpA

Binding of StpA to GII is driven by electrostatic forces mediated by positively charged amino acids (Mayer *et al.*, 2007). This is confirmed by the  $\alpha\beta$ DCA showing that the vast majority of high scores in the C-terminal are in positively charged regions (Fig. 2). This also implies that most of the binding energy comes from amino acids in the C-terminal. It was conjectured that binding is only weakly specific and prefers unstructured RNAs (Mayer *et al.*, 2007), and our analysis confirms this as well, showing spread of top  $\alpha\beta$ DCA scores all over the RNA. Fig. 4 shows the cumulative number of top scores of with the C-terminal along the RNA, demonstrating the roughly uniform spread (with gaps excluded), with notable enrichment before position 200.

In a fine-grained examination, one notices several interactions of special interest. The glycine at position 126 of the protein, which is known to strongly reduce binding affinity when mutated, takes part in three

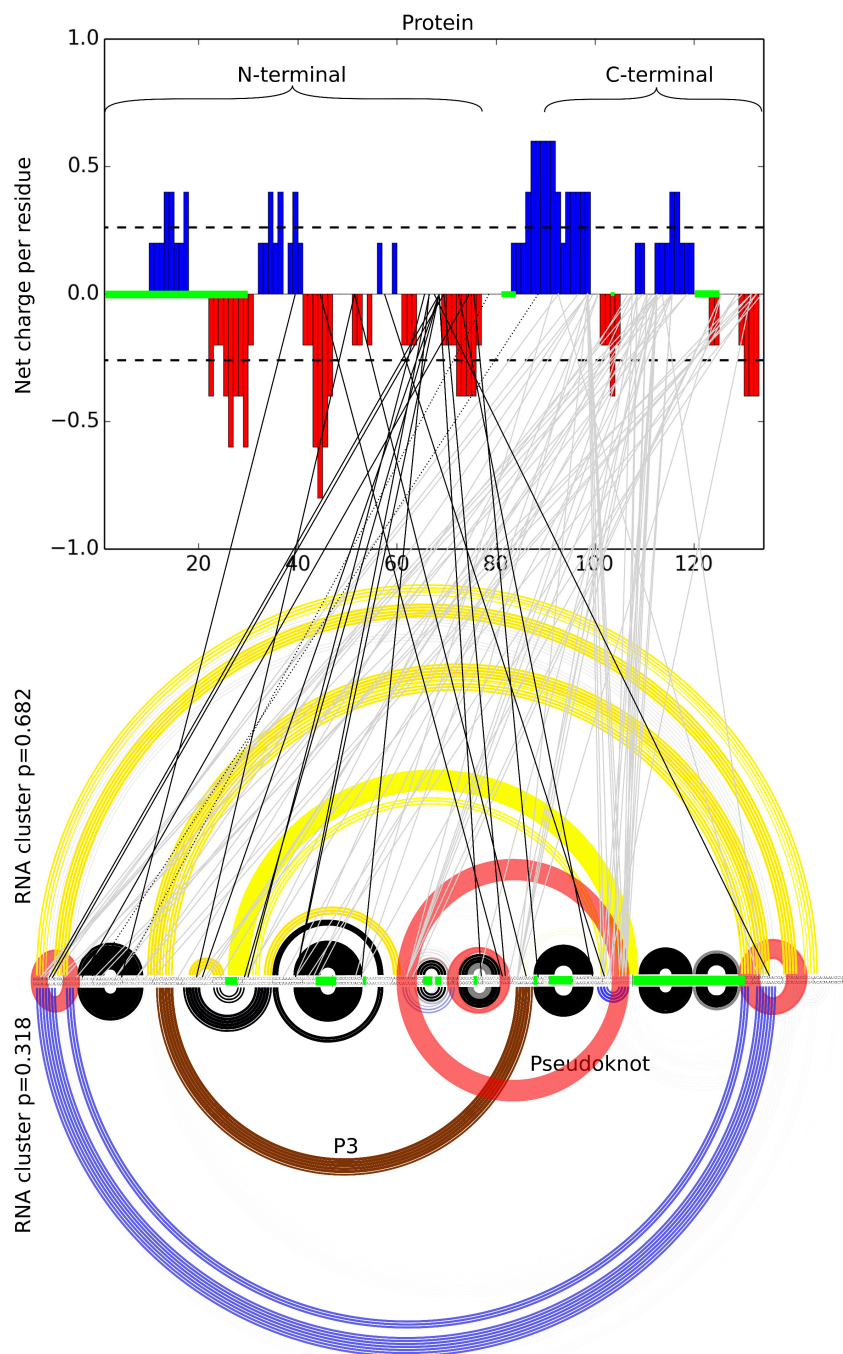


FIG. 2: Significant scores between the protein and the RNA. Top: the protein charge distribution. Bottom: the RNA sequence with the two main structure clusters. Arcs represent base pairs: yellow only in the main, most probable cluster, blue only in the secondary, least probable cluster, and black in the functional structure. Red discs are base pairs in the functional structure absent from both clusters. Brown arcs are the P3 stem. Positions highlighted in green in the protein and RNA had more than 50% of gaps in the alignment and were therefore omitted from the analysis. Significant DCA scores are denoted by lines: dark between the N-terminal and the RNA, light gray lines between the C-terminal and the RNA.



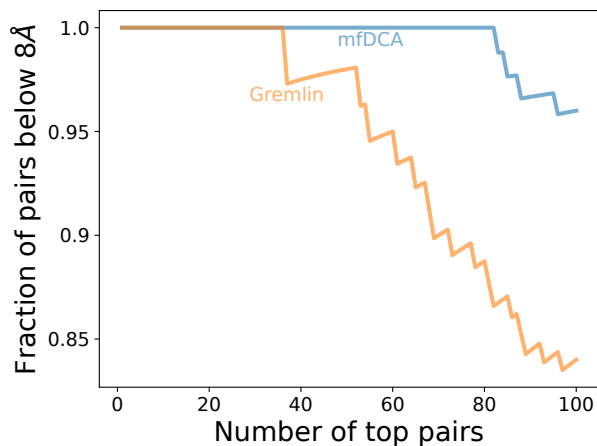


FIG. 3: Evaluating the predictive power of the group I intron RNA sequence alignment. Fraction of nucleotide pairs closer than  $8\text{\AA}$  for the top DCA values using mfDCA and Gremlin.

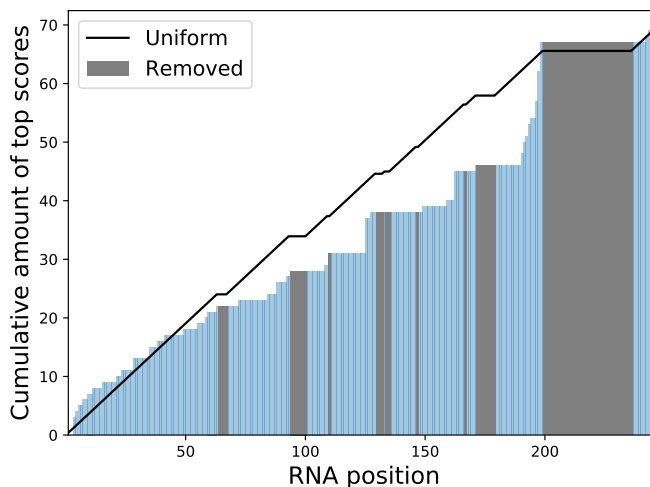


FIG. 4: Cumulative distribution of top  $\alpha\beta$ DCA scores of amino acids in the C-terminal of StpA coupled with nucleotides along the RNA. Positions with over 50% of gaps omitted from the analysis are in grey. The black curve is the cumulative uniform distribution with the same gaps.

different pairs. Position 113 of the protein — which participates in 14 different pairs, more than any other amino acid — is strongly coupled to positions 125 and 162 in the RNA, which themselves are also coupled with glycine 126. Position 125 of the RNA resides in the 5'-end of the pseudoknot, and position 162 in the 3'-end of the P3 stem. The two RNA regions with the strongest coupling to the C-terminal are both ends of the pseudoknot, which are involved in erroneous base pairs in the two dominant structures. This may explain why the isolated C-terminal is a much inferior chaperone than the whole protein. While important misfolded regions are disrupted, strong electrostatic binding slows the release of StpA, thereby impeding the correct folding of the RNA.

## B. Destabilization is mediated by negatively charged regions targeting specific RNA positions

Removing the linker and C-terminal increases by 50% the efficiency of StpA, implying that the N-terminal drives the destabilization (Mayer *et al.*, 2007). While the N-terminal composes 48.5% of StpA, it contains

only 21% of the strong couplings, 19 out of 90. The black lines in Fig. 2 show the coupled pairs between StpA and GII. Of those pairs, 7 are coupled with regions that determine the functional RNA conformation in both structure ensemble, in particular the one position paired with the positively charged amino acid at position 39 in the N-terminal. The other 12 pairs destabilize four different regions.

In both ensembles, the functional short stems at the beginning and end of the GII sequence are blocked by a stem linking those two parts together. We find three interactions that target this region: First, the 3'-end of the pseudoknot, in both ensembles, is blocked by misfolded stems which are strongly correlated with a position of the N-terminal. Second, following the 5'-end of the P3 stem, the region between positions 60 and 85 of the RNA has the right conformation in the less probable ensemble and is targeted by three couples. Finally, 3'-end of the P3 stem, present only in the least probable ensemble, is involved in one coupling. The last 3 coupled positions are in a hairpin stem preceding the P3 3'-end. This stem is missing functional base pairs in both ensembles, two of the coupled pairs are in positions lacking a base pair, the third in the unpaired region of the hairpin.

Without StpA, around 55% of the RNA is able to fold into its functional self-splicing form, and this folding fraction rises to roughly 80% in the presence of the chaperone (Mayer *et al.*, 2007). The strong correlations we observe manifest an interplay between the two main structure ensembles of the RNA, with the less probable one presenting most of the correct base pairs. Regions that contain functional stems in the least probable ensemble are all targeted by destabilizing couplings with the N-terminal. In both ensembles, the functional but energetically unfavorable pseudoknot has stems in its 3'-end impeding its formation. These stems are also destabilized by the N-terminal.

### C. Conclusion

DCA methods have been applied to infer protein structure, and protein-protein or protein-RNA interactions (Weinreb *et al.*, 2016). DCA demonstrated high correlations among amino acids in IDPs, suggesting that many IDPs do exhibit structure in a particular context (Toth-Petroczy *et al.*, 2016). In the present study, we expanded DCA to account for the different alphabets and different levels of sequence diversity in the concatenated sequences of protein and RNA used for the alignment. We used this adapted  $\alpha\beta$ DCA method to infer the strong couplings between a non-coding RNA, GII, and its disordered protein chaperone, StpA. Understanding the StpA-GII is particularly challenging, since on top of the inherent disorder of the protein, the misfolded RNA also lacks a well-defined structure.

The present  $\alpha\beta$ DCA method produces 15% less significant contacts than the traditional mfDCA. In cases where the structure is unknown, a rather arbitrary significance threshold must be chosen. Having less scores departing from the distribution indicates better discrimination of important co-evolving pairs. We observe that the binding, mediated by electrostatic forces of positively charged amino acids, is *non-specific* or only weakly specific. These strong couplings, observed in the positively charged regions in the C-terminal of StpA, are paired with evenly distributed nucleotides along the RNA sequence. In contrast, the  $\alpha\beta$ DCA shows that the structural disruption driven by the N-terminal is mediated by negatively amino acids that target *specific* regions of the RNA sequence. In particular, regions in the two main structure ensembles of the RNA impeding the formation of the first and last stem, as the pseudoknot, are strongly coupled with the N-terminal. Stems in the more probable structure ensemble — which are conflicting with the functional stems present in the lower probability ensemble — are also targeted.

The present study is the first direct coupling analysis of the coupling between a disordered chaperone and its RNA target. Charge patterns have been known to be crucial for the global structure of disordered proteins, and here we shed some light on how they can affect destabilization mechanisms involved in RNA chaperoning. The analysis suggest several concrete experimental tests, for example mutations at positions 99 and 113 in the C-terminal are expected to significantly decrease binding affinity. The  $\alpha\beta$ DCA variant used in the study is simple and general enough to be easily applied for investigating other IDP-RNA mechanisms. An interesting application of the present analysis is the identification of chaperone IDPs from their charge



pattern. Those patterns could also be used to design novel destabilizing proteins.

## IV. METHOD

We first present a modified DCA algorithm, termed  $\alpha\beta$ DCA, adapted for treating paired sequences that are written in different alphabets and have different sequence conservation levels. The different nature of the paired sequences influences the normalization factors that are crucial to predict the disentangled covariations. To illustrate the method, we show how the data for the StpA protein and the group I intron RNA was gathered, and how the alignment was built. The code is freely available at: <https://github.com/vreinharz/ProtRNADCA>.

### A. $\alpha\beta$ DCA: direct coupling analysis for varying alphabets and sequence conservation

Direct coupling analysis (DCA) has proved extremely useful for disentangling covariations between non-interacting residues in multiple sequence alignment (MSA) (Morcos *et al.*, 2011; Weigt *et al.*, 2009). It aims to find the Potts model that maximizes the entropy, in order to infer the most likely probability having the given dinucleotide marginals without any additional constraints (Weigt *et al.*, 2009). The original method was constructed to treat alignments of sequences written in the same alphabet, namely the protein amino acids written in the language of the genetic code. We modify this method to treat in tandem two alphabets, of sizes  $r$  and  $s$ . Given a sequence of  $n$  characters, we assume that the first  $\zeta$  elements are from the alphabet of size  $r$ , and the last  $n - \zeta$  from the alphabet of size  $s$ . In this study, the first alphabet is of the protein amino acids and a gap, hence  $r = 21$ , and the second is of the RNA nucleotides and a gap, i.e.  $s = 5$ .

The MSA of  $M$  sequences of length  $n$  is recorded as its sequence of columns  $\{C_1^p, \dots, C_n^p\}$  where  $p \in [1, \dots, M]$  are the  $M$  sequences and  $1, \dots, n$  are the columns. Since the proteins and RNAs have different sequence similarity and alphabets, we define two values for calibrating the pseudo-count:

$$\begin{aligned} m_p^{\text{prot}} &= \sum_{q=1}^M \left[ 1 \text{ if } \text{similarity}(C_{1,\dots,\zeta}^p, C_{1,\dots,\zeta}^q) > 80\% \right], \\ m_p^{\text{rna}} &= \sum_{q=1}^M \left[ 1 \text{ if } \text{similarity}(C_{\zeta+1,\dots,n}^p, C_{\zeta+1,\dots,n}^q) > 80\% \right], \end{aligned} \quad (1)$$

where  $\text{similarity}(C_{a,\dots,b}^p, C_{a,\dots,b}^q) > 80\%$  is true if sequences  $C^p$  and  $C^q$  are identical in over 80% of the positions between  $a$  and  $b$ . We note that the values of  $m_p^{\text{prot}}$  and  $m_p^{\text{rna}}$  are at least 1 since each sequence is identical to itself. We additionally define  $M_{\text{eff}}^{\text{prot}} = \sum_{p=1}^M 1/m_p^{\text{prot}}$  and  $M_{\text{eff}}^{\text{rna}} = \sum_{p=1}^M 1/m_p^{\text{rna}}$ . The parameter  $\lambda$  is a pseudocount set to the appropriate value of  $M_{\text{eff}}^{\text{prot}}$  or  $M_{\text{eff}}^{\text{rna}}$ , as in previous studies.

The frequencies of each letter in each column, and of each pair of letters for each pair of positions, need to be re-weighted as following. We define the frequency count of a letter  $\alpha$  at column  $i$ , given the indicator function  $\mathbb{1}$ , as:

$$f_i(\alpha) = \begin{cases} \frac{1}{M_{\text{eff}}^{\text{prot}} + \lambda} \left( \frac{\lambda}{r} + \sum_{p=1}^M \frac{1}{m_p^{\text{prot}}} \mathbb{1}_{\alpha, C_i^p} \right) & : i \leq \zeta \\ \frac{1}{M_{\text{eff}}^{\text{rna}} + \lambda} \left( \frac{\lambda}{s} + \sum_{p=1}^M \frac{1}{m_p^{\text{rna}}} \mathbb{1}_{\alpha, C_i^p} \right) & : \zeta < i \end{cases} \quad (2)$$

Similarly, the frequency count of a pair of letters  $(\alpha, \beta)$  at positions  $(i, j)$  is defined as:

$$f_{ij}(\alpha, \beta) = \begin{cases} \frac{1}{M_{\text{eff}}^{\text{prot}} + \lambda} \left( \frac{\lambda}{r^2} + \sum_{p=1}^M \frac{1}{m_p^{\text{prot}}} \mathbb{1}_{\alpha, C_i^p} \mathbb{1}_{\beta, C_j^p} \right) & : i < j \leq \zeta \\ \frac{1}{\frac{1}{2}(M_{\text{eff}}^{\text{prot}} + M_{\text{eff}}^{\text{rna}}) + \lambda} \left( \frac{\lambda}{rs} + \sum_{p=1}^M \frac{1}{\frac{1}{2}(m_p^{\text{prot}} + m_p^{\text{rna}})} \mathbb{1}_{\alpha, C_i^p} \mathbb{1}_{\beta, C_j^p} \right) & : i \leq \zeta < j \\ \frac{1}{M_{\text{eff}}^{\text{rna}} + \lambda} \left( \frac{\lambda}{s^2} + \sum_{p=1}^M \frac{1}{m_p^{\text{rna}}} \mathbb{1}_{\alpha, C_i^p} \mathbb{1}_{\beta, C_j^p} \right) & : \zeta < i < j \end{cases} \quad (3)$$

The rest of the equations follow closely the formulation in (Morcos *et al.*, 2011). The coupling value  $e_{ij}(\alpha, \beta)$ , between two letters  $(\alpha, \beta)$  at positions  $(i, j)$ , is calculated through the set of  $n(n-1)/2$  matrices  $\epsilon$ ,

the connected correlation matrix. For each pair of positions  $i, j$ , one defines a matrix  $\vartheta_{ij}$ , whose dimension is:  $(r-1)^2$  if  $i < j \leq \zeta$ ,  $(r-1)(s-1)$  if  $i \leq \zeta < j$ , and  $(s-1)^2$  if  $\zeta < i < j$ . For all  $i \in [1, \dots, n]$ ,  $j \in [1, \dots, n]$  the entries of  $\vartheta_{ij}$  are:

$$\vartheta_{ij}(\alpha, \beta) = f_{ij}(\alpha, \beta) - f_i(\alpha)f_j(\beta) \quad (4)$$

where  $\alpha$  and  $\beta$  take all possible  $r-1$  or  $s-1$  values, depending on the index  $i$  and  $j$ . Finally, the coupling between positions  $i, j$  is obtained by inverting  $\vartheta$

$$e_{ij} = -(\vartheta_{ij}^{-1}), \quad (5)$$

where that block matrix is extended with 0s so that the dimension of  $e_{ij}$  is  $r^2$  if  $i < j \leq \zeta$ ,  $rs$  if  $i \leq \zeta < j$ , and  $s^2$  if  $\zeta < i < j$ . The inverse of the connected correlation matrix return the negative coupling term, we correct it by taking minus its value (Morcos *et al.*, 2011).

We can now define a pseudo-probability,  $P_{ij}(\alpha, \beta)$ , of observing  $(\alpha, \beta)$  at positions  $(i, j)$ , given auxiliary residue fields  $\tilde{h}$  for each position:

$$P_{ij}(\alpha, \beta) = \frac{1}{\mathcal{Z}} \exp \left[ e_{ij}(\alpha, \beta) + \tilde{h}_i(\alpha) + \tilde{h}_j(\beta) \right], \quad (6)$$

where  $\mathcal{Z}$  is the normalization factor. The values of the fields  $\tilde{h}$  are determined by the observed single residue count, and must satisfy the system of equations:

$$f_i(\alpha) = \sum_{\gamma} P_{ij}(\alpha, \gamma), \quad f_j(\beta) = \sum_{\gamma} P_{i,j}(\gamma, \beta), \quad (7)$$

noting we must assume that if  $i \leq \zeta : \tilde{h}_i(r) = 0$  (resp. if  $\zeta < i : \tilde{h}_i(s) = 0$ ).

At this point, we can compute the directed information between two positions,  $D_{ij}$ , as:

$$D_{ij} = \sum_{\alpha, \beta} P_{ij}(\alpha, \beta) \ln \frac{P_{ij}(\alpha, \beta)}{f_i(\alpha)f_j(\beta)}. \quad (8)$$

Finally, the distortion of the scores due to the undersampling effect is corrected using an average product correction (APC) method (Dunn *et al.*, 2007).

## B. StpA homologues

The StpA protein from *Escherichia coli* (strain K12) sequence is

```
MSVMLQSLNNIRTLMARAREFSIDVLEEMLEKFRVVTKERREEEQQQRELAERQEKIST
WLELMKADGINPEELLGNSSAAAPRAGKKRQPRPAKYKFTDVNGETKKTWTGQGRTPKPIA
QALAEGKSLDDFLI.
```

The distribution of charges along the sequence is a known indicator of the global conformation of disordered proteins (Holehouse *et al.*, 2017). The Das-Pappu phase diagram shows that the StpA protein belongs to the ensemble of “Janus sequences”. Those are collapsed or expanded depending of context, and most functional disordered proteins belong to that group. This region of Janus sequences contains 40% of known disordered proteins (Das *et al.*, 2015), whereas another 25% reside in the strong polyampholyte region, and 30% are classified as weak polyampholyte.

The jackhmmmer method (Potter *et al.*, 2018) was run iteratively 13 times, until the number of sequences added to the matches was less than 1% of the already identified ones. We identified 21593 matches, 5749 of them unique. jackhmmmer provides a sequence alignment of all the hits, which belong to 7539 different taxa. Every sequence in GenBank (Sayers *et al.*, 2019) associated with those taxa was downloaded, a total of 633 GB of data.

### C. Group I intron

The *td* group I intron (GII) sequence from *phage T4 thymidylate-synthase* is:

```
gguUAAUUGAGGCCUGAGUAAUAGGUGACUUAUACUUGUAAUCUAUCAAACGGGGAACC
UCUCUAGUAGACAAUCCCGUGCUAAAUUGUAGGACUGCCCGGGUUCUACAUAAAUGCCUA
ACGACUAUCCCUUUGGGGAGUAGGGUCAAGUGACUCGAAACGAUAGACAACUUGCUUUA
CAAGUUGGAGAUUAGUCUGCUCUGCAUGGUGACAUGCAGCUGGAUAAUUCGGGGUA
AGAUUAACGACCUUAUCUGAACAUAAUGcuac
```

and its functional secondary structure is

```
(((((.....))))((((((((((.....))))))))))...((((((...((((((.....
(((.....)))...))))))((.....((((((((((.....)))))))))).....))...
[. [[[[[.(((.....))(.((((((((((.....))))))))..))))))((((((.....
)))))).....]]]]]...((((((((((.....)))))))).(((.....))))((((
((.....)))))).....
```

where the pseudoknot is indicated with square brackets, '[' and ']'.

The GII has 14 different subgroups, which have been cataloged in the GISSD database (Zhou *et al.*, 2008). Identification and alignment of GII sequences are highly dependent of the subgroup they belong to (Nawrocki *et al.*, 2018). Therefore, for each subgroup, we generated a covariance model using **Infernal** (Nawrocki and Eddy, 2013). The IA2 subgroup is the most compatible with GII. With GII, **Infernal** reports an E-value of  $1.7 \times 10^{-36}$  and 63% of the base pairs are well predicted. In particular, the complete P3 stem (brown in Fig. 2) is perfectly aligned with the consensus structure. We note that while the sequence has 273 nucleotides, only the first 248 were matched. The rest of the analysis is performed on those 248 nucleotides.

A search of matches to the IA2 subgroup was then computed with the **cmsearch** routine of **Infernal**, on all sequences from the 7359 taxa gathered previously. A total of 7542 sequences were identified as significant — e-value < 0.01 — with default parameters, 471 of them unique. The **cmsearch** tool returns an alignment of those sequences.

### D. Protein–RNA alignment

Duplicate proteins and RNAs were removed from each taxon. Every possible protein–RNA pair inside a taxon was concatenated together. This yielded a total of 13 230 couples, 10 013 of which unique.

Only columns where StpA and the GII have less than 50% of gaps were kept. In total, 39 positions of the proteins were removed, the N-terminal’s first 30 positions, 6 in the C-terminal and 2 in the linker. In the RNA, 64 positions were removed. The resulting protein alignment is composed of 95 columns and the RNA alignment of 184.

### E. 5S RNA–RL18 protein interactions

We compare four DCA methods for the benchmark of inferring the interactions between the 5S RNA and the RL18 protein. The four methods are: (i) standard mfDCA, where the pseudocount is kept at 11 for every position in our alignment, (ii) Our  $\alpha\beta$ DCA implementation of mfDCA with adaptive pseudocount, (iii) The implementation EVcouplings (Hopf *et al.*, 2018) of pseudo-likelihood DCA (plmDCA), and (vi) The Markov-random field DCA as implemented in Gremlin (Ovchinnikov *et al.*, 2014).

We used the protein alignment of RL18 provided in (Weinreb *et al.*, 2016). The RNA sequences were recovered from the Rfam family RF00001 (Kalvari *et al.*, 2017). We followed the protocol of Sec. IV.D. Due to the large amount of sequences, we selected randomly one pair of protein–RNA per taxonomic family,

as in (Weinreb *et al.*, 2016). The alignment before removing columns with over 50% of gaps is available at <https://github.com/vreinharz/ProtRNADCA>. We computed amino acid–nucleotide distances in the 4V4Q protein structure (Schuwirth *et al.*, 2005). Pairs with distance shorter than 10Å are considered to be in contact.

We show in Fig. S3 the results of the first top 100 scores for each method. Only mfDCA and  $\alpha\beta$ DCA (mfDCA adaptive) have their highest scores correctly predicting a contact. While mfDCA’s fourth hit is correct but not the one in  $\alpha\beta$ DCA method, the opposite occurs at their sixth top score. Both methods outperform Gremlin and EVcouplings on the top 20 scores. While mfDCA and  $\alpha\beta$ DCA true positives steadily declines as more top scores are taken into account, Gremlin sees an increase to up to 50% at its 30<sup>th</sup> score. All methods then converge to roughly 22% true positive when the first 100 scores are taken into account.

The overlap of scores over  $4\sigma$  from each bulk distribution is shown in Fig. S4 (Heberle *et al.*, 2015). While 95% of those overlap between mfDCA and  $\alpha\beta$ DCA, they are almost completely exclusive from Gremlin and EVcouplings top results. None of the top pairs is identified by all of the four methods and only 2 are shared by mfDCA,  $\alpha\beta$ DCA and Gremlin. This is the only overlap between any three methods.

## ACKNOWLEDGMENTS

We thank Thomas Hopf and Debora S. Marks for help with the EVcoupling suite, Sergey Ovchinnikov for providing Gremlin, Eric Westhof for providing the structure of the td-Intron, and Eric Nawrocki for help with infernal and understanding how to generate the subgroup GII alignments. This work was supported by the taxpayers of South Korea through the Institute for Basic Science, project code IBS-R020.

## REFERENCES

- Aalberts, D. P., and W. K. Jannen, 2013, *Rna* **19**(4), 475.
- Babu, M. M., R. W. Kriwacki, and R. V. Pappu, 2012, *Science* **337**(6101), 1460.
- Bhaskaran, H., and R. Russell, 2007, *Nature* **449**(7165), 1014.
- Burger, L., and E. Van Nimwegen, 2010, *PLoS computational biology* **6**(1), e1000633.
- Buske, P., and P. A. Levin, 2013, *Molecular microbiology* **89**(2), 249.
- Das, R. K., K. M. Ruff, and R. V. Pappu, 2015, *Current opinion in structural biology* **32**, 102.
- De Leonardis, E., B. Lutz, S. Ratz, S. Cocco, R. Monasson, A. Schug, and M. Weigt, 2015, *Nucleic acids research* **43**(21), 10444.
- Doetsch, M., R. Schroeder, and B. Fürtig, 2011, *The FEBS journal* **278**(10), 1634.
- Dunn, S. D., L. M. Wahl, and G. B. Gloor, 2007, *Bioinformatics* **24**(3), 333.
- Guo, Q., and A. M. Lambowitz, 1992, *Genes & development* **6**(8), 1357.
- Heberle, H., G. V. Meirelles, F. R. da Silva, G. P. Telles, and R. Minghim, 2015, *BMC bioinformatics* **16**(1), 169.
- Holehouse, A. S., R. K. Das, J. N. Ahad, M. O. Richardson, and R. V. Pappu, 2017, *Biophysical journal* **112**(1), 16.
- Hopf, T. A., A. G. Green, B. Schubert, S. Mersmann, C. P. Schärfe, J. B. Ingraham, A. Toth-Petroczy, K. Brock, A. J. Riesselman, P. Palmedo, *et al.*, 2018, *Bioinformatics* **35**(9), 1582.
- Kalvari, I., J. Argasinska, N. Quinones-Olvera, E. P. Nawrocki, E. Rivas, S. R. Eddy, A. Bateman, R. D. Finn, and A. I. Petrov, 2017, *Nucleic acids research* **46**(D1), D335.

- Keul, N. D., K. Oruganty, E. T. S. Bergman, N. R. Beattie, W. E. McDonald, R. Kadirvelraj, M. L. Gross, R. S. Phillips, S. C. Harvey, and Z. A. Wood, 2018, *Nature* **563**(7732), 584.
- Lorenz, R., S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, 2011, *Algorithms for Molecular Biology* **6**(1), 26.
- Marks, D. S., L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, 2011, *PloS one* **6**(12), e28766.
- Mayer, O., L. Rajkowitzsch, C. Lorenz, R. Konrat, and R. Schroeder, 2007, *Nucleic Acids Research* **35**(4), 1257, ISSN 0305-1048, 1362-4962.
- McCaskill, J. S., 1990, *Biopolymers: Original Research on Biomolecules* **29**(6-7), 1105.
- Michel, F., and E. Westhof, 1990, *Journal of molecular biology* **216**(3), 585.
- Mohr, G., A. Zhang, J. A. Gianelos, M. Belfort, and A. M. Lambowitz, 1992, *Cell* **69**(3), 483.
- Morcos, F., A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, 2011, *Proceedings of the National Academy of Sciences* **108**(49), E1293.
- Nawrocki, E. P., and S. R. Eddy, 2013, *Bioinformatics* **29**(22), 2933.
- Nawrocki, E. P., T. A. Jones, and S. R. Eddy, 2018, *Nucleic Acids Research* **46**(15), 7970.
- Nott, T. J., E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowitz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, *et al.*, 2015, *Molecular cell* **57**(5), 936.
- Ovchinnikov, S., H. Kamisetty, and D. Baker, 2014, *Elife* **3**, e02030.
- Ovchinnikov, S., L. Kinch, H. Park, Y. Liao, J. Pei, D. E. Kim, H. Kamisetty, N. V. Grishin, and D. Baker, 2015, *Elife* **4**, e09248.
- Papaleo, E., G. Saladino, M. Lambrugh, K. Lindorff-Larsen, F. L. Gervasio, and R. Nussinov, 2016, *Chemical reviews* **116**(11), 6391.
- Papasaikas, P., and J. Valcárcel, 2016, *Trends in biochemical sciences* **41**(1), 33.
- Paukstelis, P. J., J.-H. Chen, E. Chase, A. M. Lambowitz, and B. L. Golden, 2008, *Nature* **451**(7174), 94.
- Piovesan, D., F. Tabaro, L. Paladin, M. Necci, I. Mičetić, C. Camilloni, N. Davey, Z. Dosztányi, B. Mészáros, A. M. Monzon, *et al.*, 2017, *Nucleic acids research* **46**(D1), D471.
- Potter, S. C., A. Luciani, S. R. Eddy, Y. Park, R. Lopez, and R. D. Finn, 2018, *Nucleic Acids Research* **46**(W1), W200.
- Reinharz, V., Y. Ponty, and J. Waldispühl, 2016, *Nucleic Acids Research* **44**(11), e104.
- Reuter, J. S., and D. H. Mathews, 2010, *BMC bioinformatics* **11**(1), 129.
- Sayers, E. W., M. Cavanaugh, K. Clark, J. Ostell, K. D. Pruitt, and I. Karsch-Mizrachi, 2019, *Nucleic Acids Research* **47**(D1), D94.
- Schad, E., E. Fichó, R. Pancsa, I. Simon, Z. Dosztányi, and B. Mészáros, 2017, *Bioinformatics* **34**(3), 535.
- Schuwirth, B. S., M. A. Borovinskaya, C. W. Hau, W. Zhang, A. Vila-Sanjurjo, J. M. Holton, and J. H. D. Cate, 2005, *Science* **310**(5749), 827.
- Tompa, P., and P. Csermely, 2004, *The FASEB Journal* **18**(11), 1169.
- Toth-Petroczy, A., P. Palmedo, J. Ingraham, T. A. Hopf, B. Berger, C. Sander, and D. S. Marks, 2016, *Cell* **167**(1), 158.



- Van Der Lee, R., M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, *et al.*, 2014, *Chemical reviews* **114**(13), 6589.
- Varadi, M., and P. Tompa, 2015, in *Intrinsically disordered proteins studied by NMR spectroscopy* (Springer), pp. 335–349.
- Waldsich, C., 2002, *Genes & Development* **16**(17), 2300, ISSN 08909369.
- Weigt, M., R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, 2009, *Proceedings of the National Academy of Sciences* **106**(1), 67.
- Weinreb, C., A. J. Riesselman, J. B. Ingraham, T. Gross, C. Sander, and D. S. Marks, 2016, *Cell* **165**(4), 963.
- Woodson, S. A., 2010, *RNA biology* **7**(6), 677.
- Wright, P. E., and H. J. Dyson, 2015, *Nature reviews Molecular cell biology* **16**(1), 18.
- Zhou, Y., C. Lu, Q.-J. Wu, Y. Wang, Z.-T. Sun, J.-C. Deng, and Y. Zhang, 2008, *Nucleic Acids Research* **36**(suppl\_1), D31.

## SUPPLEMENTARY MATERIAL

### 1. Group I intron sequence diversity

Fig. S1 shows the conservation of the most frequent nucleotide in the GII alignment. Necessarily, each position must be at least as conserved as the ones in the GII sequence conservation figure. Each nucleotide shown is the most frequent one. If only A or G are present in that position, an R is shown for purine. If only C or U are present in that position, a Y is shown for pyrimidine.

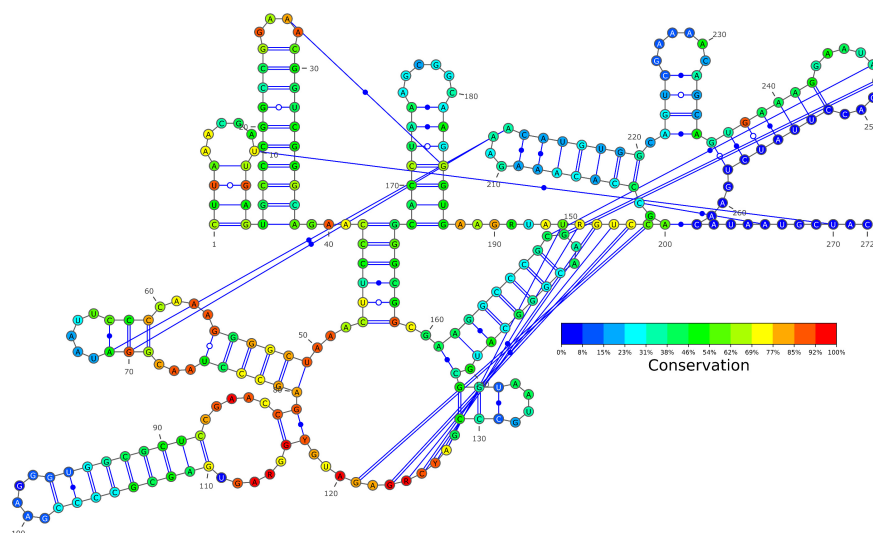


FIG. S1: The most conserved nucleotide for each position, with its percentage of conservation.

### 2. APC values vs. random shuffling of the aligned sequences

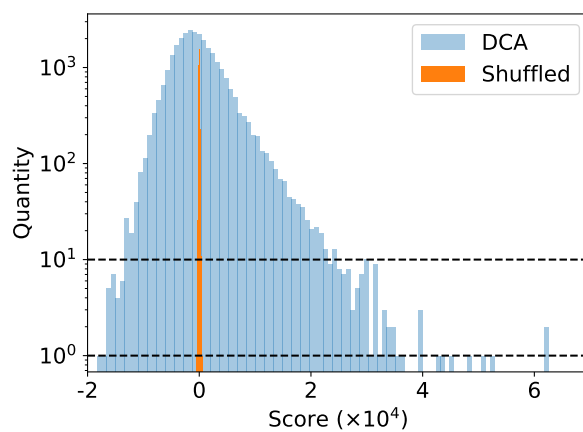


FIG. S2: The distribution of APC values obtained by our method on the StpA-GII alignment compared to the same values after shuffling the sequences. There are 100 blue and 100 orange bins. The orange bins are therefore narrower.

### 3. Comparing four DCA methods on the ribosomal 5S–RL18 complex

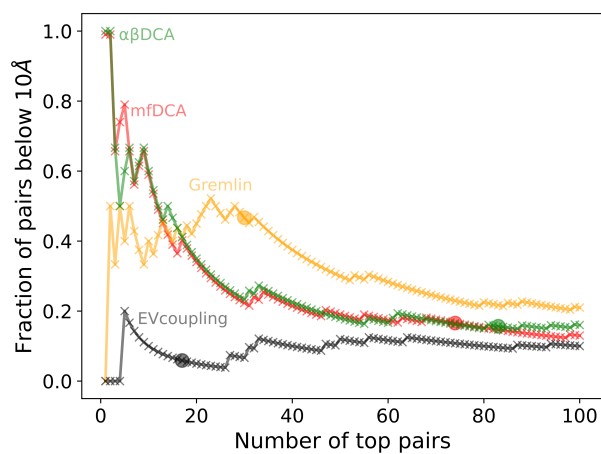


FIG. S3: Comparing four DCA methods for the benchmark of inferring the 5S–RL18 complex from PDB 4V4Q. The graphs show fraction of pairs with a distance below 10Å for the top 100 DCA values for each method. The circles indicate the last score over  $4\sigma$  from the bulk distribution.

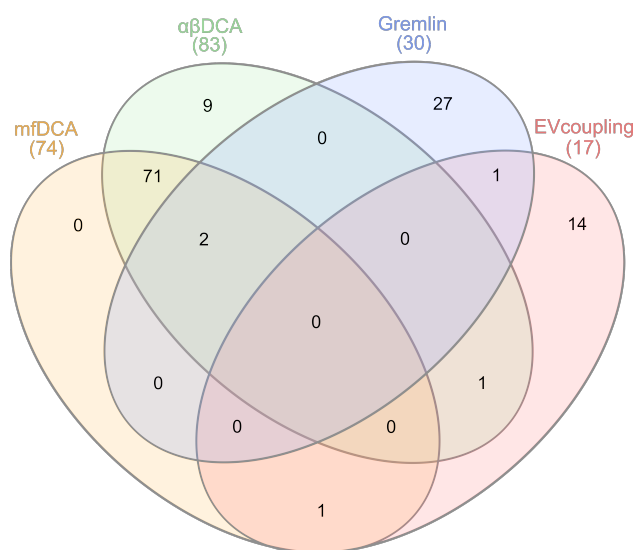


FIG. S4: Intersection of the scores over  $4\sigma$  for each of the four DCA methods on the 5s–RL18 complex.