αβDCA method identifies unspecific binding but specific disruption of the group I intron by the StpA chaperone

VLADIMIR REINHARZ[1,2] and TSVI TLUSTY[1,3]

[1]Center for Soft and Living Matter, Institute for Basic Science, Ulsan 44919, Republic of Korea
[2]Department of Computer Science, Université du Québec à Montréal, Montréal, H2X 3Y7, Canada
[3]Department of Physics, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea

ABSTRACT

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 because of 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. We adapted direct coupling analysis (DCA) into the αβDCA method that can treat in tandem sequences written in two alphabets, nucleotides and amino acids. With αβDCA, we could analyze StpA–RNA interactions and show consistency with a previously proposed two-pronged mechanism: StpA disrupts *specific* positions in the group I intron while *globally* and loosely binding to the entire structure. Moreover, the interactions are strongly associated with the charge pattern: Negatively charged regions in the destabilizing StpA amino-terminal affect a few specific positions in the RNA, located in stems and in the pseudoknot. In contrast, positive regions in the carboxy-terminal contain strongly coupled amino acids that promote nonspecific 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.

Keywords: group I intron; StpA chaperone; direct coupling analysis; RNA structure; disordered protein

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 toward one with higher binding affinity (Keul et al. 2018). Although our repertoire of IDPs is steadily growing (Varadi and Tompa 2015; Piovesan et al. 2017; Schad et al. 2017), the function of most is yet to be discovered (Van Der Lee et al. 2014; Papaleo et al. 2016). 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 is those that chaperone RNA folding: To perform their functions, noncoding RNAs rely on well-conserved structures, which have been used for sequence alignment and putative RNA prediction (Nawrocki and Eddy 2013). Although some noncoding 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; Woodson 2010; Papasaikas and Valcárcel 2016).

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

Corresponding author: tsvi@unist.ac.kr
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.074336.119.

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 (Waldsich et al. 2002; Mayer et al. 2007). 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 amino-terminal and carboxy-terminal. Excising the carboxy-terminal from the sequence increases the efficacy of the chaperone, whereas mutations in the carboxy-terminal hinder its binding capacity (Mayer et al. 2007). An entropy transfer model has been proposed, in which 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 because 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 (MSAs). As of August 2018, GenBank (Sayers et al. 2019) had sequences totaling more than 3.7×10^{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-dependent structures (Toth-Petroczy et al. 2016), although 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 because of 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 αβDCA, produces the first analysis of the interaction of a disordered protein, StpA, with a noncoding RNA, the group I intron. Our method identifies 90 strongly coupled pairs between StpA and GII. The inferred locations of those pairs are consistent with the results of Mayer et al. (2007).

We find that the charge pattern is strongly associated with the type of interactions: The amino terminal of StpA, which is known to destabilize the RNA, exhibits a few specific interactions among negatively charged regions of the protein and regions of the GII, which are critically misfolded in the structure's ensemble or impede functional loops from forming. In the carboxyl 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, although both terminals are of roughly the same length, only 21% of the top DCA scores are in the amino terminal. These findings propose a charge-dependent two-pronged mechanism of unspecific binding but specific disruption by chaperone IDPs.

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 Materials and Methods). First, we tested this simple DCA variant—which we call αβDCA (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 pseudolikelihood DCA (plmDCA). For the benchmark of the 5S–RL18 ribosomal complex, the adaptive αβDCA method predicts more contacts in its top scores (see section "5S RNA–RL18 protein interactions" in the Materials and Methods). Additionally, we observe that the mfDCA method outperforms Gremlin in the GII alignment, most probably owing to the correct pseudocount for a five-letter alphabet, rather than that of the 21-letter alphabet of proteins used in Gremlin. In the following, we apply the αβDCA method to analyze the StpA–GII alignment (code and alignment are available at https://gitlab.info.uqam.ca/cbe/abDCA).

αβDCA 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 a conservation level prohibits coevolution analysis. Figure 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. 2). To test whether the alignment contains more information than an ensemble of random sequences, we compare the distribution of αβDCA scores

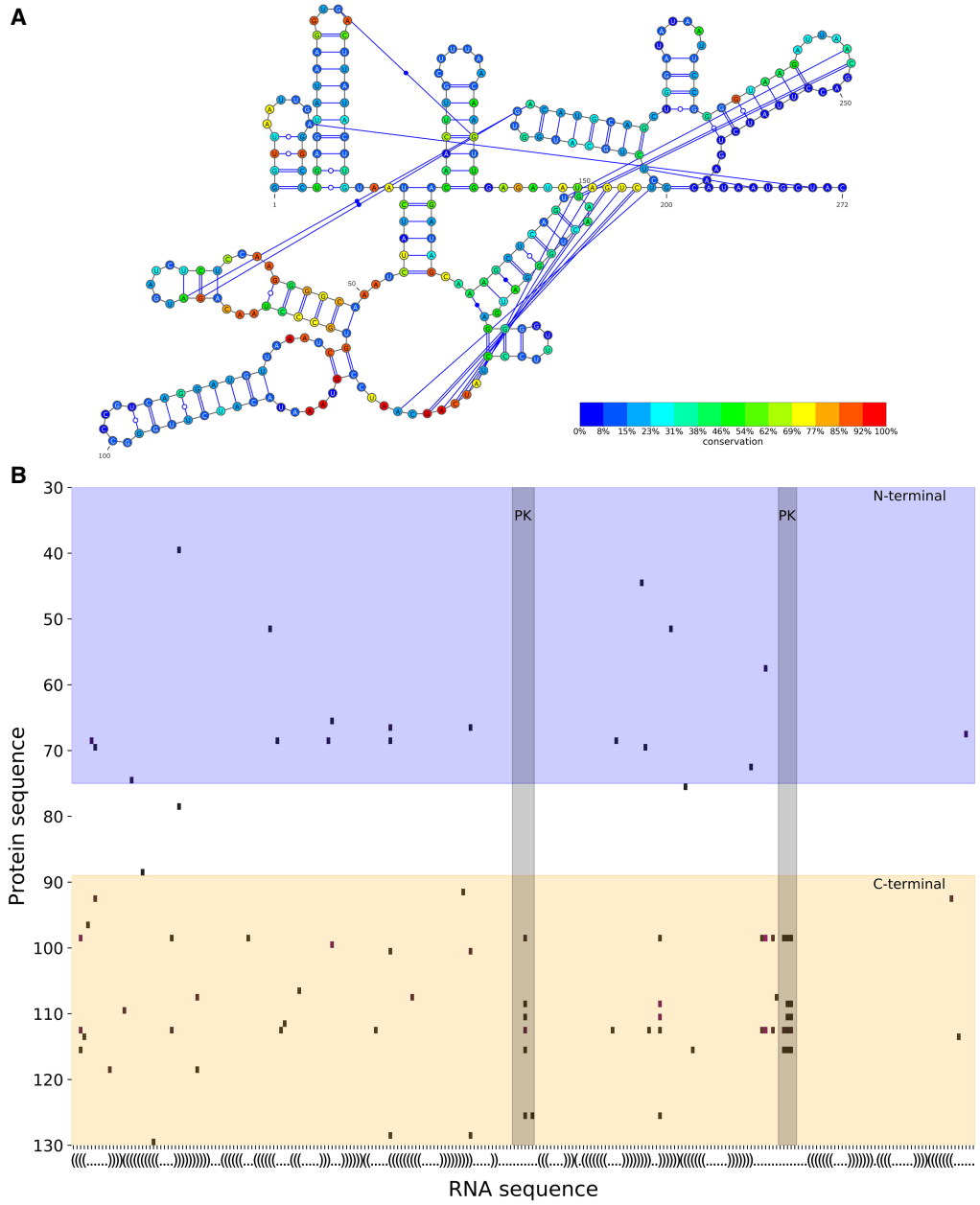

FIGURE 1. (*A*) GII secondary structure and its sequence conservation. The last 25 positions have no conservation levels because they are excluded from the alignment (see section "Group I intron" in Materials and Methods). (*B*) Positions of significant DCA scores (≥4σ above average) between the StpA protein (*y*-axis) and the GII RNA (*x*-axis). The RNA axis is labeled with the secondary structure in parentheses notation. The blue region is the amino terminal of StpA and the orange region its carboxyl terminal. The pale gray regions are the pseudoknot (PK) of GII. The last 25 positions of GII are omitted (see section "Group I intron" in Materials and Methods).

from the StpA–GII alignment with those obtained from the same alignment but with randomly shuffled sequences. The scores of the original alignment spread over a much wider range then that of the shuffled alignment, thus confirming that the DCA analysis extracts information from the alignment (Fig. 3).

The StpA–GII amino acid–nucleotide pairs with the strongest DCA couplings are shown in Figure 1B. The distribution of scores is assumed to be normal, with its average and standard deviation computed from the empirical data. Scores that are four standard deviations (4σ) above the average are deemed significant. The αβDCA 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. As shown below, in agreement with previous studies (Ovchinnikov et al. 2015; Toth-Petroczy et al. 2016), the number of false positives increases with the number of

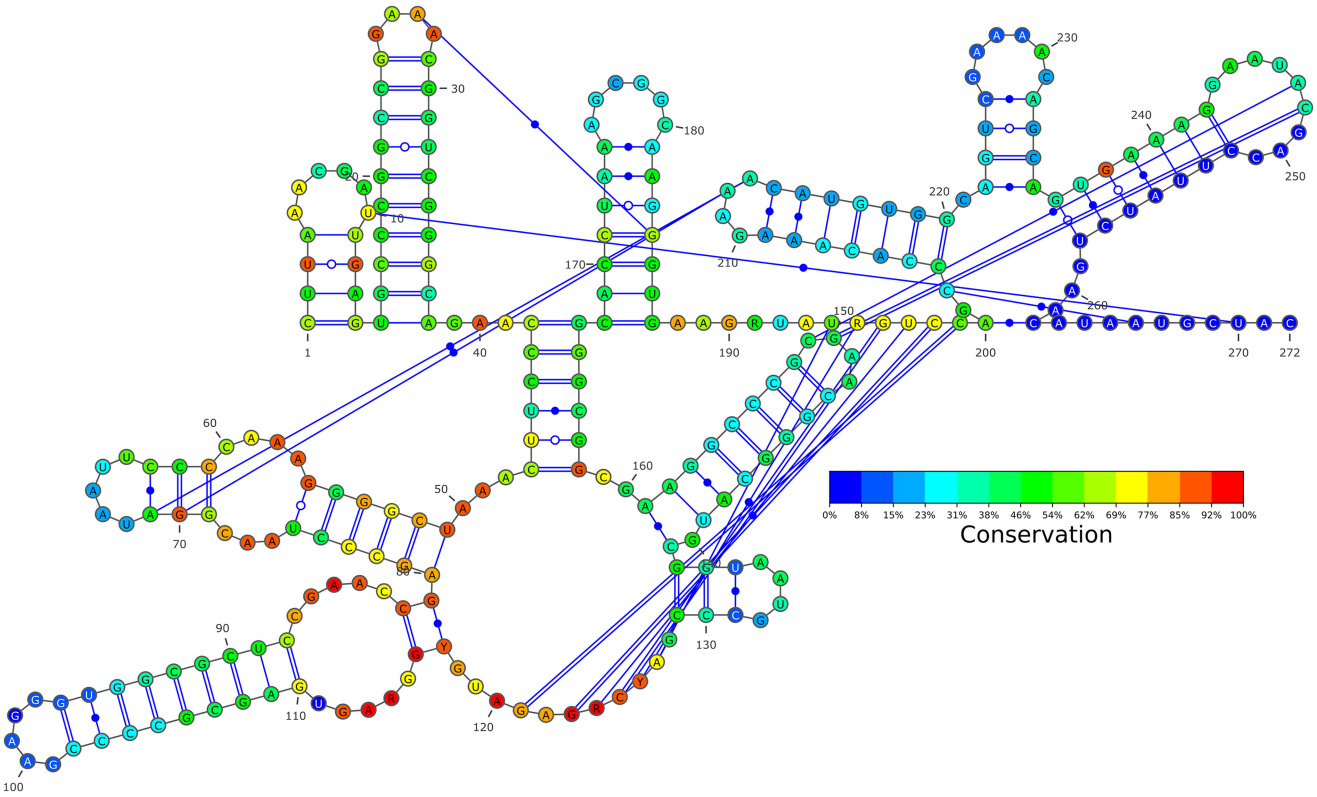

FIGURE 2. Global conservation. The most conserved nucleotide for each position, with its percentage of conservation. Each nucleotide shown is the most frequent one. If only A or G are present in that position, an R is shown for purine. If only C or U are present in that position, a Y is shown for pyrimidine.

selected pairs. One can therefore expect that the analysis that shows fewer significant scores will yield fewer errors.

Inferred protein–RNA interactions are selective in the amino terminal and global in the carboxyl terminal of StpA

The amino and carboxyl terminals of StpA are known to interact differently with the RNA (Waldsich et al. 2002). This motivates us to characterize the number and distribution of high αβDCA scores, which indicate strong physical couplings, in each of these two regions. Because 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.

To this end, from the RNA sequence, RNAstructure (Reuter and Mathews 2010) computes, in the McCaskill thermodynamic framework (McCaskill 1990), the probability of each possible base pair. Those pairing probabilities can be divided into two main structural ensembles to ease the visualization (Aalberts and Jannen 2013). We plot in Figure 4 the net charge distribution along the StpA protein (averaged over a window of five 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

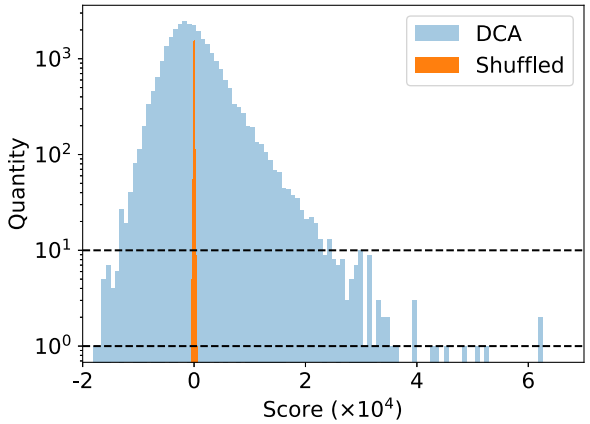

FIGURE 3. The distribution of APC values obtained by our method on the StpA–GII alignment compared to the same values after shuffling the sequences. There are 100 blue and 100 orange bins. The orange bins are therefore narrower.

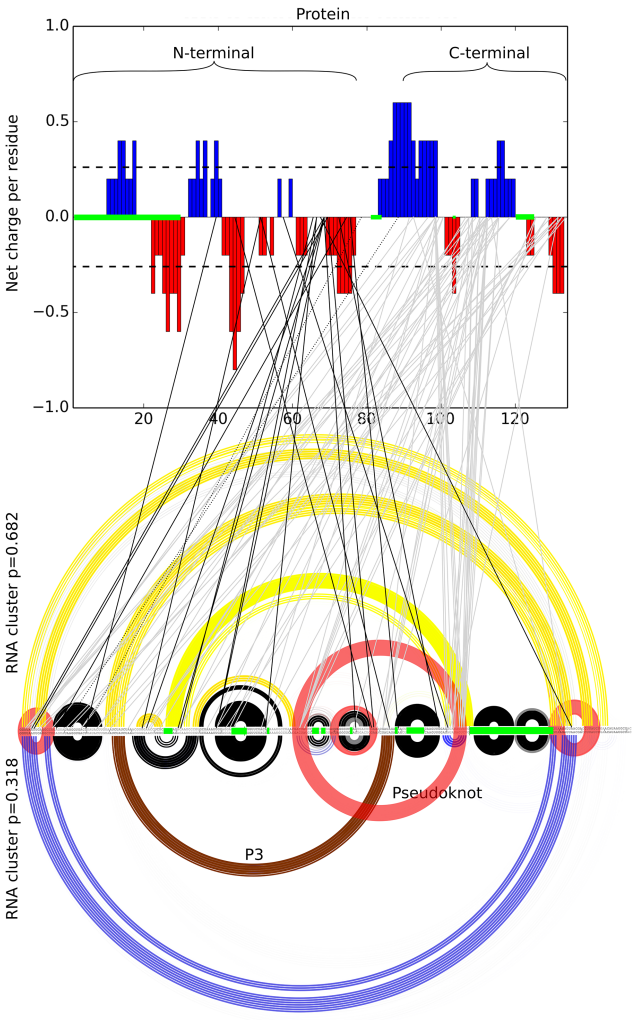

FIGURE 4. Significant scores between the protein and the RNA. (*Top*) The protein charge distribution. (*Bottom*) The RNA sequence with the two main structure clusters. Arcs represent base pairs: yellow only in the main, most probable cluster, blue only in the secondary, least probable cluster, and black 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 > 50% of gaps in the alignment and were therefore omitted from the analysis. Significant DCA scores are denoted by lines: dark between the amino-terminal and the RNA, light gray lines between the carboxy-terminal and the RNA.

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, as annotated by Waldsich et al. (2002). Note that although 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 (≥4σ) between the protein and the RNA: 19 in negative regions of the amino-terminal (dark lines), 69 in mostly positive regions of the carboxyl terminal (light gray lines), and two in the linker between the amino and carboxyl terminals (dashed lines).

More than 61% of carboxy-terminal significant amino acids have many globally distributed partners, on average 4.3 nt. In contrast, the N-amino acids show a more selective evolutionary signature with 67% of them exhibiting significant covariation with *only* 1 nt.

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 αβDCA is identical to mfDCA when treating a single alphabet. We consider as a good prediction a pair of nucleotides closer than 8 Å in the 3D structure. Figure 5 shows the number of these true positives (distance < 8 Å) for the hundred top scores. Although the first 40 top scores are well predicted by both methods, the Gremlin method is outperformed by mfDCA in the next 60 scores.

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 (Waldsich et al. 2002; Doetsch et al. 2011). Mutation studies provide evidence for GII–StpA

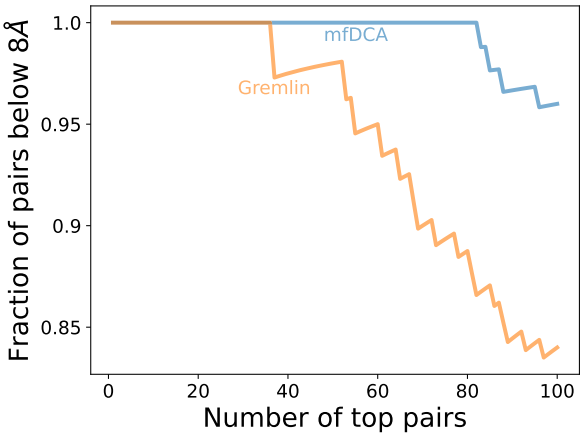

FIGURE 5. Evaluating the predictive power of the group I intron RNA sequence alignment. Fraction of nucleotide pairs closer than 8 Å for the top DCA values using mfDCA and Gremlin.

interactions: Mutations in the StpA carboxyl terminal reduce the binding affinity between StpA and the group I intron, whereas complete deletion of the carboxyl terminal increases the efficiency of StpA as a chaperone (Waldsich et al. 2002). A carboxy-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 of the GII–StpA mechanism, based on the αβDCA results. We show that the αβDCA results are consistent with previous experimental studies. Moreover, they put forward a detailed picture of coupled amino acids and nucleotides responsible for both binding and destabilizing interactions.

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 αβDCA showing that the vast majority of high scores in the carboxyl terminal are in positively charged regions (Fig. 4). Out of the 69 pairs, 44 (64%) are in positively charged regions, 18 (26%) in neutral regions, and seven (10%) in negatively charged regions. This also implies that most of the binding energy comes from amino acids in the carboxyl terminal. It was conjectured that binding is only weakly specific and prefers unstructured RNAs (Mayer et al. 2007). Our analysis is consistent with this conjecture, showing a spread of top αβDCA scores all over the RNA. Figure 6 shows the cumulative number of top scores of nucleotides with the carboxy-terminal along the RNA, demonstrating the roughly uniform spread (with gaps excluded), with notable enrichment before position 200.

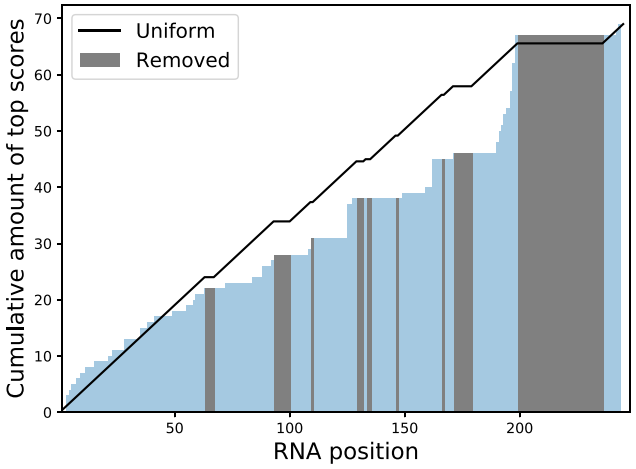

FIGURE 6. Cumulative distribution of top DCA scores of amino acids in the carboxyl terminal of StpA coupled with nucleotides along the RNA. Positions with >50% of gaps omitted from the analysis are in gray. The black curve is the cumulative uniform distribution with the same gaps.

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 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 carboxyl 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 carboxyl terminal is a much inferior chaperone than the whole protein. Our analysis is also consistent with the theory that although important misfolded regions are disrupted, strong electrostatic binding slows the release of StpA, thereby impeding the correct folding of the RNA.

Destabilization is mediated by negatively charged regions targeting specific RNA positions

Removing the linker and carboxyl terminal increases by 50% the efficiency of StpA, implying that the amino terminal drives the destabilization (Mayer et al. 2007). Although the amino-terminal composes 48.5% of StpA, it contains only 21% of the strong couplings, 19 out of 90. The black lines in Figure 4 show the coupled pairs between StpA and GII. Out of 19 significant pairs, one is in a positively charged region, five (26%) are in a neutral region, and the remaining 13 (68%) are in negatively charged regions.

Of those 19 pairs, seven are coupled with regions that determine the functional RNA conformation in both structure's ensembles, in particular the one position paired with the positively charged amino acid at position 39 in the amino terminal. The other 12 pairs correlate with four different regions that are expected to be destabilized as the probable structure conflicts with the functional one.

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 that are strongly correlated with a position of the amino 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, the 3′-end of the P3 stem, present only in the least probable ensemble, is involved in one coupling. The last three 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, ~55% of the RNA is able to fold into its functional self-splicing form, and this folding fraction rises to ~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 couplings with the destabilizing amino terminal. In both ensembles, the functional but energetically unfavorable pseudoknot has stems in its 3′-end impeding its formation. Our analysis proposes that the stems are also destabilized by the amino-terminal.

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 αβDCA method to infer the strong couplings between a noncoding RNA, GII, and its disordered protein chaperone, StpA. Understanding the StpA–GII is particularly challenging, because on top of the inherent disorder of the protein, the misfolded RNA also lacks a well-defined structure.

The present αβDCA method produces 15% less significant contacts than the traditional mfDCA. In cases in which the structure is unknown, a rather arbitrary significance threshold must be chosen. Having fewer scores departing from the distribution indicates better discrimination of important coevolving pairs. Our findings are consistent with experiments and a proposed mechanism in which the binding, mediated by electrostatic forces of positively charged amino acids, is *nonspecific* or only weakly specific. These strong couplings, observed in the positively charged regions in the carboxyl terminal of StpA, are paired with evenly distributed nucleotides along the RNA sequence. In contrast, the αβDCA suggests that the structural disruption driven by the amino terminal is mediated by negatively charged 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 amino 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 suggests several concrete experimental tests—for example, mutations at positions 99 and 113 in the carboxy-terminal are expected to significantly decrease binding affinity. The αβDCA 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.

MATERIALS AND METHODS

We first present a modified DCA algorithm, termed αβDCA, 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 were gathered, and how the alignment was built. The code is freely available at: https://gitlab.info.uqam.ca/cbe/abDCA.

αβDCA: direct coupling analysis for varying alphabets and sequence conservation

DCA has proved extremely useful for disentangling covariations between noninteracting residues in MSA (Weigt et al. 2009; Morcos et al. 2011). 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 ζ 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 $\{C_1^p, \ldots, C_n^p\}$, where $p \in [1, \ldots, M]$ are the M sequences and $1, \ldots, n$ are the columns. Because the proteins and RNAs have different sequence similarity and alphabets, we define two values for calibrating the pseudocount:

$$m_p^{\text{prot}} = \sum_{q=1}^{M} [1 \text{ if similarity}(C^p_{1,\ldots,\zeta}, C^q_{1,\ldots,\zeta}) > 80\%],$$

$$m_p^{\text{rna}} = \sum_{q=1}^{M} [1 \text{ if similarity}(C^p_{\zeta+1,\ldots,n}, C^q_{\zeta+1,\ldots,n}) > 80\%],$$

where similarity $(C^p_{a,\ldots,b}, C^q_{a,\ldots,b}) > 80\%$ is true if sequences C^p and C^q are identical in >80% of the positions between a and b. We

note that the values of m_p^{prot} and m_p^{rna} are at least 1 because each sequence is identical to itself. We additionally define

$$M_{\text{eff}}^{\text{prot}} = \sum_{p=1}^{M} 1/m_p^{\text{prot}} \quad \text{and} \quad M_{\text{eff}}^{\text{rna}} = \sum_{p=1}^{M} 1/m_p^{\text{rna}}$$

The parameter λ 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 reweighted as following. We define the frequency count of a letter α at column i, given the indicator function 1, as

$$f_i(\alpha) = \begin{cases} \dfrac{1}{M_{\text{eff}}^{prot} + \lambda}\left(\dfrac{\lambda}{r} + \displaystyle\sum_{p=1}^{M} \frac{1}{m_p^{prot}} 1_{\alpha, C_i^p}\right) : i \le \zeta \\ \dfrac{1}{M_{\text{eff}}^{rna} + \lambda}\left(\dfrac{\lambda}{s} + \displaystyle\sum_{p=1}^{M} \frac{1}{m_p^{rna}} 1_{\alpha, C_i^p}\right) : i > \zeta \end{cases}$$

Similarly, the frequency count of a pair of letters (α, β) at positions (i, j) is defined as

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

The rest of the equations follow closely the formulation in Morcos et al. (2011). The coupling value $e_{ij}(\alpha, \beta)$, between two letters (α, β) at positions (i, j), is calculated through the set of $n(n-1)/2$ matrices ∂, the connected correlation matrix. For each pair of positions i, j, one defines a matrix ∂_{ij}, whose dimension is $(r-1)^2$ if $i<j\le\zeta$, $(r-1)(s-1)$ if $i\le\zeta<j$, and $(s-1)^2$ if $\zeta<i<j$. For all $i\in[1, \ldots, n]$, $j\in[1, \ldots, n]$ the entries of ∂_{ij} are

$$\partial_{ij}(\alpha, \beta) = f_{ij}(\alpha, \beta) - f_i(\alpha) f_j(\beta),$$

where α and β 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 ∂:

$$e_{ij} = -(\partial_{ij}^{-1})$$

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

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

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

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:

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

Note that we must assume that if $i \le \zeta$:$\tilde{h}_i(r) = 0$ (respectively, if $\zeta < i$:$\tilde{h}_i(s) = 0$).

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

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

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

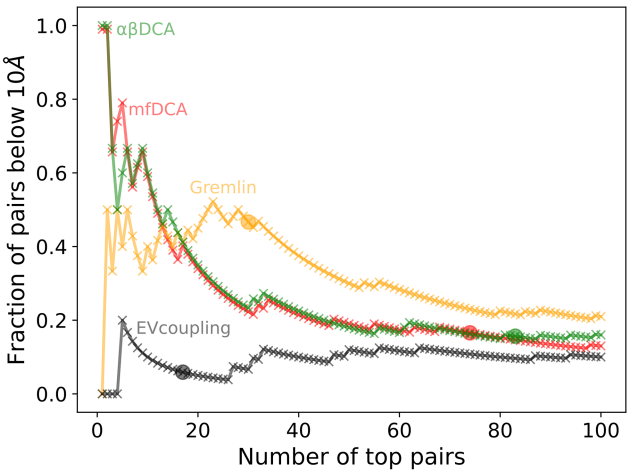

FIGURE 7. Comparing four DCA methods for the benchmark of inferring the 5S–RL18 complex from PDB 4V4Q. The graphs show fraction of pairs with a distance below 10 Å for the top 100 DCA values for each method. The circles indicate the last score over 4σ from the bulk distribution.

StpA homologs

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

MSVMLQSLNNIRTLRAMAREFSIDVLEEMLEKFRVVTKERREEE
EQQQRELAERQEKISTWLELMKADGINPEELLGNSSAAAPRAGKK
RQPRPAKYKFTDVNGETKTWTGQGRTPKPIAQALAEGKSLDDFLI.

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 on 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 jackhmmer method (Potter et al. 2018) was run iteratively 13 times, until the number of sequences added to the matches was <1% of the already identified ones. We identified 21,593 matches, 5749 of them unique. jackhmmer 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.

Group I intron

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

gguUAAUUGAGGCCUGAGUAUAAGGUGACUUAUACUUGU
AAUCUAUCUAAACGGGGAACCUCUCUAGUAGACAAUCCCG
UGCUAAAUUGUAGGACUGCCCGGGUUCUACAUAAAUGCCU
AACGACUAUCCCUUUGGGGAGUAGGGUCAAGUGACUCGA
AACGAUAGACAACUUGCUUUAACAAGUUGGAGAUAUAGUC
UGCUCUGCAUGGUGACAUGCAGCUGGAUAUAAUUCCGGGG
UAAGAUUAACGACCUUAUCUGAACAUAAUGcuac

and its functional secondary structure, from Waldsich et al. (2002), 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 on 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 1.7×10^{-36}, and 63% of the base pairs are well predicted. In particular, the complete P3 stem (brown in Fig. 4) is perfectly aligned with the consensus structure. We note that although the sequence has 273 nt, only the first 248 were matched. The rest of the analysis is performed on those 248 nt.

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.

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, of which 10,013 were unique.

Only columns in which StpA and the GII have <50% of gaps were kept. In total, 39 positions of the proteins were removed, the amino terminal's first 30 positions, six in the carboxyl terminal and two 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.

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, in which the pseudocount is kept at 21 for every position in our alignment, (ii) our αβDCA implementation of mfDCA with adaptive pseudocount, (iii) the implementation EVcouplings (Hopf et al. 2018) of pseudolikelihood DCA (plmDCA), and (vi) the Markov-random field DCA as implemented in Gremlin (Ovchinnikov et al. 2014).

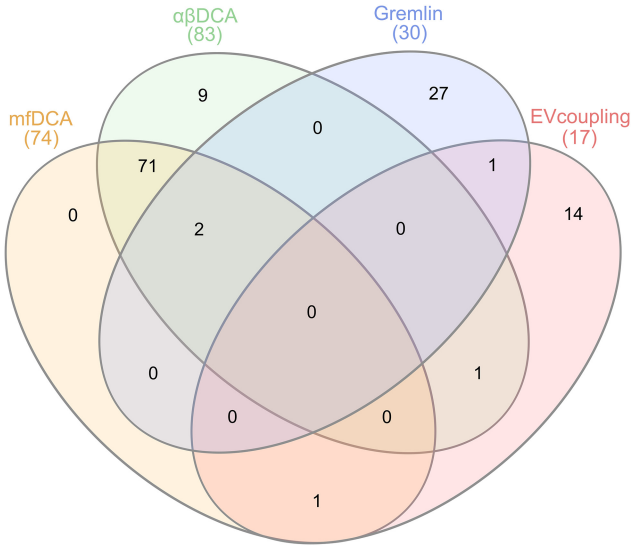

FIGURE 8. Intersection of the scores over 4σ for each of the four DCA methods on the 5s–RL18 complex.

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 the previous section. Because of 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 >50% of gaps is available at https://gitlab.info.uqam.ca/cbe/abDCA. We computed amino acid–nucleotide distances in the 4V4Q protein structure (Schuwirth et al. 2005). Pairs with distance shorter than 10 Å are considered to be in contact.

We show in Figure 7 the results of the first top 100 scores for each method. Only mfDCA and αβDCA (mfDCA adaptive) have their highest scores correctly predicting a contact. Although mfDCA's fourth hit is correct but not the one in the αβDCA method, the opposite occurs at their sixth top score. Both methods outperform Gremlin and EVcouplings on the top 20 scores. Although the true positives of mfDCA and αβDCA steadily decline as more top scores are taken into account, Gremlin sees an increase to up to 50% at its 30th score. All methods then converge to ~22% true positive when the first 100 scores are taken into account.

The overlap of scores over 4σ from each bulk distribution is shown in Figure 8 (Heberle et al. 2015). Although 95% of those overlap between mfDCA and αβDCA, 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 two are shared by mfDCA, αβDCA, 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, Ministry of Education, Science and Technology, project code IBS-R020. V.R. was also supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2020-05795).

Received December 12, 2019; accepted July 19, 2020.

REFERENCES

Aalberts DP, Jannen WK. 2013. Visualizing RNA base-pairing probabilities with RNAbow diagrams. *RNA* **19:** 475–478. doi:10.1261/rna.033365.112

Babu MM, Kriwacki RW, Pappu RV. 2012. Versatility from protein disorder. *Science* **337:** 1460–1461. doi:10.1126/science.1228775

Bhaskaran H, Russell R. 2007. Kinetic redistribution of native and misfolded RNAs by a DEAD-box chaperone. *Nature* **449:** 1014. doi:10.1038/nature06235

Burger L, Van Nimwegen E. 2010. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol* **6:** e1000633. doi:10.1371/journal.pcbi.1000633

Buske PJ, Levin PA. 2013. A flexible C-terminal linker is required for proper FtsZ assembly in vitro and cytokinetic ring formation in vivo. *Mol Microbiol* **89:** 249–263. doi:10.1111/mmi.12272

Das RK, Ruff KM, Pappu RV. 2015. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr Opin Struct Biol* **32:** 102–112. doi:10.1016/j.sbi.2015.03.008

De Leonardis E, Lutz B, Ratz S, Cocco S, Monasson R, Schug A, Weigt M. 2015. Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res* **43:** 10444–10455. doi:10.1093/nar/gkv932

Doetsch M, Schroeder R, Fürtig B. 2011. Transient RNA–protein interactions in RNA folding. *FEBS J* **278:** 1634–1642. doi:10.1111/j.1742-4658.2011.08094.x

Dunn SD, Wahl LM, Gloor GB. 2007. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24:** 333–340. doi:10.1093/bioinformatics/btm604

Guo Q, Lambowitz AM. 1992. A tyrosyl-TRNA synthetase binds specifically to the group I intron catalytic core. *Genes Dev* **6:** 1357–1372. doi:10.1101/gad.6.8.1357

Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R. 2015. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics* **16:** 169. doi:10.1186/s12859-015-0611-3

Holehouse AS, Das RK, Ahad JN, Richardson MO, Pappu RV. 2017. CIDER: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophys J* **112:** 16–21. doi:10.1016/j.bpj.2016.11.3200

Hopf TA, Green AG, Schubert B, Mersmann S, Schärfe CP, Ingraham JB, Toth-Petroczy A, Brock K , Riesselman AJ, Palmedo P, et al. 2018. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35:** 1582–1584. doi:10.1093/bioinformatics/bty862

Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD, Petrov AI. 2017. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46:** D335–D342. doi:10.1093/nar/gkx1038

Keul ND, Oruganty K, Schaper Bergman ET, Beattie NR, McDonald WE, Kadirvelraj R, Gross ML, Phillips RS, Harvey SC, Wood ZA. 2018. The entropic force generated by intrinsically disordered segments tunes protein function. *Nature* **563:** 584. doi:10.1038/s41586-018-0699-5

Lorenz R, Bernhart SH, Hoener Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6:** 26. doi:10.1186/1748-7188-6-26

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6:** e28766. doi:10.1371/journal.pone.0028766

Mayer O, Rajkowitsch L, Lorenz C, Konrat R, Schroeder R. 2007. RNA chaperone activity and RNA-binding properties of the *E. coli* protein StpA. *Nucleic Acids Res* **35:** 1257–1269. doi:10.1093/nar/gkl1143

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119. doi:10.1002/bip.360290621

Michel F, Westhof E. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol* **216:** 585–610. doi:10.1016/0022-2836(90)90386-Z

Mohr G, Zhang A, Gianelos JA, Belfort M, Lambowitz AM. 1992. The neurospora CYT-18 protein suppresses defects in the phage T4 *td* intron by stabilizing the catalytically active structure of the intron core. *Cell* **69:** 483–494. doi:10.1016/0092-8674(92)90449-M

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci* **108:** E1293–E1301. doi:10.1073/pnas.1111471108

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29:** 2933–2935. doi:10.1093/bioinformatics/btt509

Nawrocki EP, Jones TA, Eddy SR. 2018. Group I introns are widespread in archaea. *Nucleic Acids Res* **46:** 7970–7976. doi:10.1093/nar/gky414

Nott TJ, Petsalaki E, Farber P, Jervis D, Fussner E, Plochowietz A, Craggs TD, Bazett-Jones DP, Pawson T, Forman-Kay JD, et al. 2015. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol Cell* **57:** 936–947. doi:10.1016/j.molcel.2015.01.013

Ovchinnikov S, Kamisetty H, Baker D. 2014. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *Elife* **3:** e02030. doi:10.7554/eLife.02030

Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H, Grishin NV, Baker D. 2015. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* **4:** e09248. doi:10.7554/eLife.09248

Papaleo E, Saladino G, Lambrughi M, Lindorff-Larsen K, Gervasio FL, Nussinov R. 2016. The role of protein loops and linkers in conformational dynamics and allostery. *Chem Rev* **116:** 6391–6423. doi:10.1021/acs.chemrev.5b00623

Papasaikas P, Valcárcel J. 2016. The spliceosome: the ultimate RNA chaperone and sculptor. *Trends Biochem Sci* **41:** 33–45. doi:10.1016/j.tibs.2015.11.003

Paukstelis PJ, Chen J-H, Chase E, Lambowitz AM, Golden BL. 2008. Structure of a tyrosyl-tRNA synthetase splicing factor bound to a group I intron RNA. *Nature* **451:** 94. doi:10.1038/nature06413

Piovesan D, Tabaro F, Paladin L, Necci M, Mičetić I, Camilloni C, Davey N, Dosztányi Z, Mészáros B, Monzon AM, et al. 2017. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res* **46:** D471–D476. doi:10.1093/nar/gkx1071

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. *Nucleic Acids Res* **46:** W200–W204. doi:10.1093/nar/gky448

Reinharz V, Ponty Y, Waldispühl J. 2016. Combining structure probing data on RNA mutants with evolutionary information reveals RNA-binding interfaces. *Nucleic Acids Res* **44:** e104. doi:10.1093/nar/gkw217

Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11:** 129. doi:10.1186/1471-2105-11-129

Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2019. GenBank. *Nucleic Acids Res* **47:** D94–D99. doi:10.1093/nar/gky989

Schad E, Fichó E, Pancsa R, Simon I, Dosztányi Z, Mészáros B. 2017. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* **34:** 535–537. doi:10.1093/bioinformatics/btx640

Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurjo A, Holton JM, Doudna Cate JH. 2005. Structures of the bacterial ribosome at 3.5 Å resolution. *Science* **310:** 827–834. doi:10.1126/science.1117230

Tompa P, Csermely P. 2004. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J* **18:** 1169–1175. doi:10.1096/fj.04-1584rev

Toth-Petroczy A, Palmedo P, Ingraham J, Hopf TA, Berger B, Sander C, Marks DS. 2016. Structured states of disordered proteins from genomic sequences. *Cell* **167:** 158–170. doi:10.1016/j.cell.2016.09.010

Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Keith Dunker A, Fuxreiter M, Gough J, Gsponer J, Jones DT, et al. 2014. Classification of intrinsically disordered regions and proteins. *Chem Rev* **114:** 6589–6631. doi:10.1021/cr400525m

Varadi M, Tompa P. 2015. The protein ensemble database. *Adv Exp Med Biol* **870:** 335–349. doi:10.1007/978-3-319-20164-1_11

Waldsich C, Grossberger R, Schroeder R. 2002. RNA chaperone StpA loosens interactions of the tertiary structure in the *td* group I intron in vivo. *Genes Dev* **16:** 2300–2312. doi:10.1101/gad.231302

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci* **106:** 67–72. doi:10.1073/pnas.0805923106

Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS. 2016. 3D RNA and functional interactions from evolutionary couplings. *Cell* **165:** 963–975. doi:10.1016/j.cell.2016.03.030

Woodson SA. 2010. Taming free energy landscapes with RNA chaperones. *RNA Biol* **7:** 677–686. doi:10.4161/rna.7.6.13615

Wright PE, Dyson HJ. 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* **16:** 18. doi:10.1038/nrm3920

Zhou Y, Lu C, Wu Q-J, Wang Y, Sun Z-T, Deng J-C, Zhang Y. 2008. GISSD: group I intron sequence and structure database. *Nucleic Acids Res* **36:** D31–D37. doi:10.1093/nar/gkm766