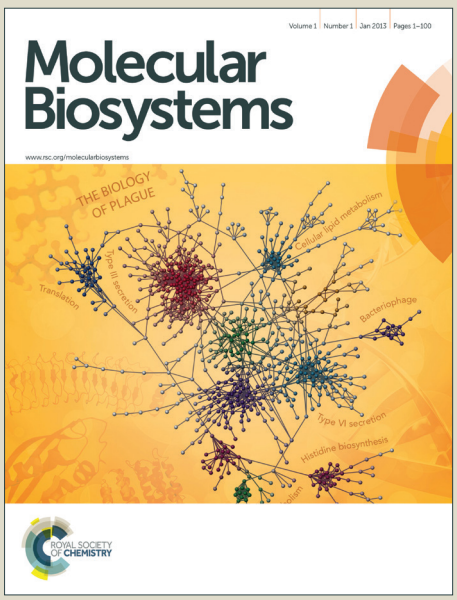
Molecular BioSystems

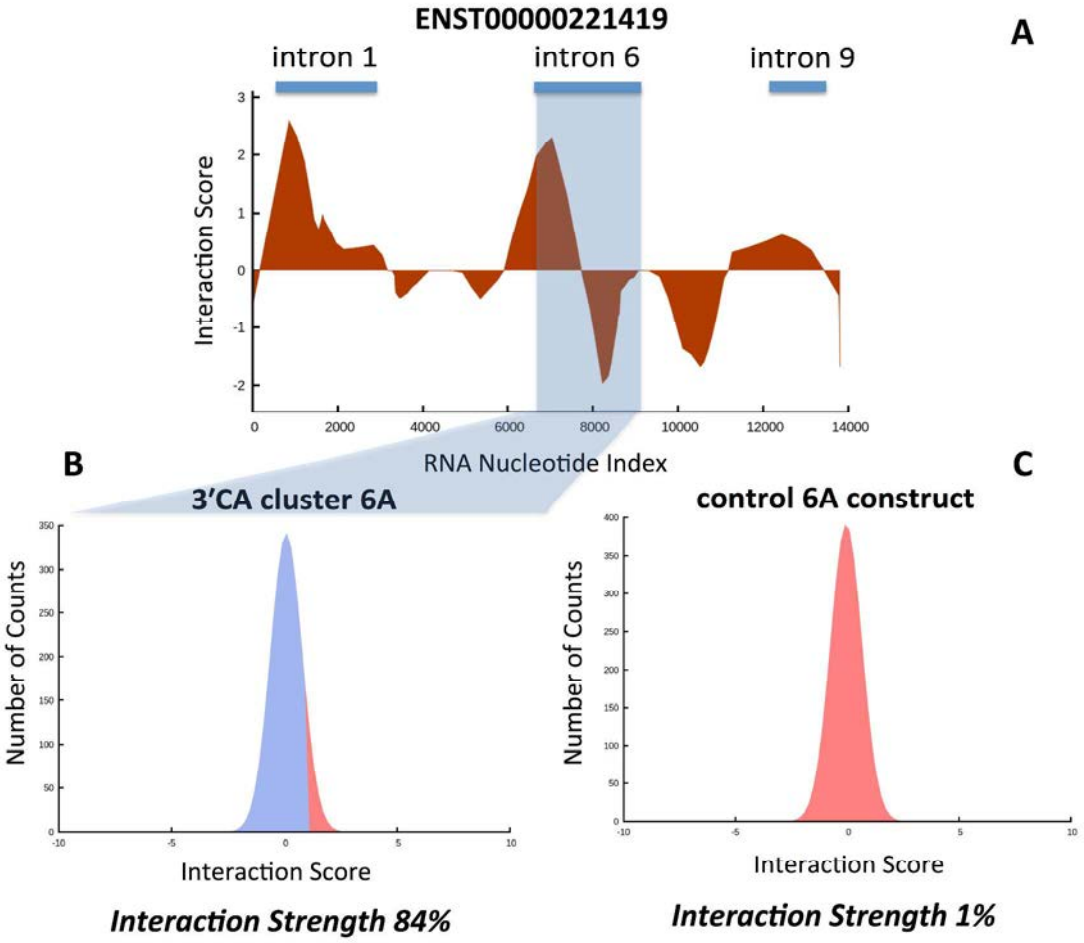
Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Table of contents entry



We review latest advances and future challenges in experimental and computational investigation of protein-RNA networks.

Discovery of Protein-RNA Networks

Davide Cirillo[1,2] , Carmen Maria Livi[1,2], Federico Agostini[1,2] and Gian Gaetano Tartaglia[1,2]

[1]Gene Function and Evolution, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.

Corresponding author: Gian Gaetano Tartaglia.

Telephone: +34 933160116.

Email: gian.tartaglia@crg.eu.

Abstract

Coding and non-coding RNAs associate with proteins to perform important functions in the cell. Protein-RNA complexes are essential components of the ribosomal and spliceosomal machinery, are involved in epigenetic regulation and form non-membrane-bound aggregates known as granules. Despite the functional importance of ribonucleoprotein interactions, the precise mechanisms of macromolecular recognition are still poorly understood.

Here, we present the latest developments for experimental and computational investigation of protein-RNA interactions. We compare performances of different algorithms and discuss how predictive models allow the large-scale investigation of ribonucleoprotein associations. Specifically, we focus on approaches to decipher mechanisms regulating the activity of transcripts in protein networks. Finally, the *cat*RAPID *omics express* method is introduced for the analysis of protein-RNA expression networks.

Introduction

Recent approaches based on nucleotide-enhanced UV crosslinking and immunoprecipitation (CLIP) identified a number of previously unknown proteins with an RNA-binding activity [1,2]. As RNA-binding proteins (RBPs) orchestrate many post-transcriptional events and influence gene expression by acting at various steps of RNA metabolism [3], protein-RNA associations could be important players in regulatory networks [4] . Intriguingly, only a fraction of the genome (i.e. about 1.4% in humans) is translated into proteins, while > 50% of the mammalian genome is predicted to be transcribed, which suggests that a large number of RNAs might contribute to biological processes by associations with RBPs [5–7].

Despite the increasing amount of high-throughput data, basic questions regarding remain to be addressed: How do protein and RNA recognize each other? What are the mechanisms leading to formation of assemblies such as ribonucleoprotein aggregates? Is it possible to build models to predict protein-RNA associations and exploit theoretical frameworks to investigate functional and dysfunctional complexes?

Here we present state-of-the-art experimental and computational approaches to investigate protein-RNA associations. We describe predictive models for the characterization of ribonucleoprotein complexes and introduce the latest developments in the field including cat*RAPID omics express*. Finally, we discuss future challenges for the prediction of RNA structure and propensity to form ribonucleoprotein aggregates.

Quantitative approaches to detect protein-RNA interactions

Detection of RNA targets and identification of binding sites is usually based on *in vitro* and *in vivo* experiments such as systematic evolution of ligands by exponential enrichment (SELEX) [8] and immunoprecipitation (IP) [9,10]. Although accurate, these approaches require considerable amount of work for the optimization of experimental conditions [11,12]:

- RNA immunoprecipitation (RIP) is the most common approach to reveal interaction between proteins and ribonucleic acids. To perform RIP, it is necessary to use an antibody directed against the RNA-binding protein of interest to pull down associated RNAs from cellular extracts. RNA sequences are identified using

qPCR, microarrays and next-generation sequencing [13]. Two relevant issues limit the application of the method: i) the low resolution (i.e., the binding sites cannot be identified) and high propensity to include indirect interactions; ii) the propensity of protein-RNA complexes to re-assemble after cell lysis, which might introduce artifacts [14]. A RIP variant is being developed to detect RNA interactions with nuclear chromatin. In this case, the approach exploits a formaldehyde fixation step to lock RNA-chromatin interactions. The crosslinking method allows identification of indirect protein-RNA interactions as well as detection of higher molecular weight macromolecular complexes.

- CLIP [15] exploits crosslinking and nuclease digestion, enabling stringent purification of RNA-protein complexes through size separation by gel electrophoresis to reveal which RNAs are bound and where on the sequence the interaction occurs. A variant of this technique, called individual-nucleotide resolution CLIP (iCLIP) allows detection of RNA-protein interactions with single-base precision [16]. Two key differences between CLIP and RIP are the crosslinking and gel-purification steps. The RNA molecules in the RNA-protein complexes are radioactively end-labeled, resolved by SDS-PAGE and transferred to a membrane, which enables visualization of the complex and ensures that no non-specific RNA is co-purified.

- ChIRP (chromatin isolation by RNA purification), CHART (capture hybridization analysis of RNA targets) and RAP (RNA antisense purification) exploit biotinylated oligonucleotides complementary to the RNA of interest as a way to pull down associated proteins [17,18]. Mass spectrometry and next-generation sequencing are employed to identify proteins associated with RNA and genomic locations at which those interactions occur.

The field of protein-RNA interaction is evolving rapidly thanks to high-throughput technologies [16] and the basic principles regulating the formation of ribonucleoprotein complexes are starting to be elucidated. Nevertheless, a number of crucial questions are emerging from experimental works [19,20]: How many proteins have RNA binding abilities [2] ? Do non-canonical RNA-binding regions occur more often than previously thought [1] ? What is the role of RNA structure in macromolecular recognition [21,22] ? Are there special RNA-mediated mechanisms regulating cell homeostasis [23,24] ?

Computational methods for prediction of protein-RNA interactions

Physico-chemical properties are particularly useful to identify binding regions in protein and RNA molecules. A number of algorithms, such as RNABindR [25], SCRPRED [26] and the *clever*Suite [27], have been trained to predict the RNA-binding propensity of proteins using primary structure information. Recent computational methods focus on the simultaneous predictions of contact regions for both protein and RNA, which is essential to capture the specificity of ribonucleoprotein complexes.

In 2011 the *cat*RAPID algorithm was released to predict protein associations with coding and non-coding transcripts [28]. The method was trained on 858 not redundant protein-RNA complexes available in the Protein Data Bank (http://www.rcsb.org) to discriminate interacting and non-interacting molecules using the information contained in primary structure. *cat*RAPID was tested on the non-nucleic-acid-binding proteins (NNBP) dataset (area under ROC curve of 0.92) [29], the non-coding RNA and protein interactions (NPInter) database (area under the ROC curve of 0.88) [30], and a number of interactions validated by RIP and CLIP approaches (RNase P and MRP complexes, XIST network and RBP-associated transcriptomes) [23,24,31,32].

At the same time *cat*RAPID was published, Pancaldi and Baehler introduced an approach based on Support Vector Machine (SVM) and Random Forest (RF) to predict RBP targets in yeast [33]. To rationalize the factors contributing to the formation of ribonucleoprotein complexes, the authors studied untranslated region (UTR) properties, RNA structures, expression levels, gene ontology (GO) associations and physico-chemical features. A subset of 40 RBPs along with the corresponding experimental targets for a total of 12000 interactions were used to validate the method. The findings of this analysis can be summarized as follows:

- High nitrogen content and high isoelectric point discriminate RBPs from other proteins;
- A significant correlation between RNA length and relative amount of Glycine, Isoleucine and Valine has been reported;
- Proteins with high-isoelectric points tend to bind to long mRNAs containing a large number of stem-loops;
- RBPs sharing common targets often interact with each other and bind to the mRNAs of their interaction partners, building an auto-regulatory system.

To test the predictive power of the method, the authors performed cross-validation and reported accuracy of 0.69, an Area Under the ROC Curve of 0.77 and sensitivity and specificity around 0.7. SVM performed better than RF, but only 14 out of 76 RBP targets could be well discriminated. The approach presented in this study is not available in form of web-server / source-code, which limits its use.

Always in 2011, Muppirala *et al.* developed *RPIseq* to predict protein-RNA associations using SVM and RF approaches [34]. In contrast to Pancaldi and Baehler, *RPIseq* predictions are based on primary structure. In *RPIseq,* RNA sequences are encoded with the normalized frequency of nucleotide tetrads (total of 256 characteristics), while protein sequences are represented using conjoint triad (total of 343 characteristics):

- The nucleotide tetrads are 4-mer combinations of [A,C,G,U];
- The protein triad divides the 20 amino acids in 7 classes: [A,G,V], [I,L,F,P], [Y,M,T,S] , [H,N,Q,W], [R,K], [D,E] and [C].

RPIseq [34] training has been performed on two different datasets obtained from Protein-RNA Interface Database (PRIDB)[35]: a larger set containing ribosomal complexes and a smaller set without ribosomal proteins-RNA associations. On both sets, RF outperforms SVM in both accuracy and true positive rate. Both methods show good performances on the dataset containing ribosomal information (SVM: accuracy=0.87; RF: accuracy= 0.89). The algorithms have been additionally applied to predict protein interactions with non-coding RNAs downloaded from NPInter [30]. When trained on the larger dataset, RF correctly predicted 80% of NPInter interactions, while SVM only 66%.

In 2012, Wang *et al.* [36] developed a sequence-based Naïve Bayes classifier to predict interactions between RBPs and non-coding RNAs. Three different datasets were used to validate the method: PRIDB[35] with and without ribosomal complexes and NPInter[30]. The following features are used as input:

- RNA sequences are analyzed using 3-mer occurrence of [A,C,G,U];
- Four classes [D,E], [H,R,K], [C,G,N,Q,S,T,Y] and [A,F,I,L,M,P,V,W] are employed for amino acid frequencies.

In a 10-fold cross validation, Naïve Bayes and extended Naïve Bayes classifiers obtained similar results with accuracies around 0.7, specificities of 0.9 and sensitivities of 0.3-0.4 on all the datasets.

A major advantage of *cat*RAPID [28] and *RPIseq* [34] is their online availability, whereas the algorithms by Pancaldi and Baehler [33] and Wang *et al.* [36] are not publicly available.

The cat*RAPID* modules

In the last years, a number of algorithms have been implemented to investigate mechanisms associated with protein-RNA interactions. We focused on large-scale predictions and comparison with experimental data technologies such as CLIP. The *cat*RAPID modules to compute protein-RNA interactions are available at our group webpage http://service.tartaglialab.com/page/catrapid_group. At present, 4 algorithms are available: *cat*RAPID *graphic*, *cat*RAPID *fragments*, *cat*RAPID *strength*, *cat*RAPID *omics*. Here, an overview is provided of the different modules with related examples (Table 1).

*cat*RAPID *graphic.* The contributions of secondary structure, hydrogen bonding and van der Waals' are combined together into the *interaction profile*:

$$\vec{\Phi}_x = \alpha_H \vec{\mathrm{H}}_x + \alpha_W \vec{\mathrm{W}}_x + \alpha_S \vec{\mathrm{S}}_x \quad (1)$$

where the variable x indicates RNA ($x = r$) or protein ($x = p$). The $\vec{\mathrm{S}}$ term designates the profile associated with secondary structure occupancy of each nucleotide (or amino acid) in RNA (protein) sequence:

$$\vec{S} = S_1, S_2, \ldots, S_{length} \quad (2)$$

The RNAplot algorithm is employed to generate secondary structure coordinates of a number of models [37]. Using the nucleotide coordinates, we define *secondary structure occupancy* by counting the number of contacts made by each nucleotide within the different regions of the chain (Figure 1). High values of *secondary structure occupancy* indicate that base pairing occurs in regions with high propensity to form hairpin-loops, while low values are associated with junctions or multi-loops. Similarly, $\vec{\mathrm{H}}$ represents the hydrogen-bonding and $\vec{\mathrm{W}}$ the van der Waals' profile [38]. The *interaction propensity* π is defined as the inner product between the protein propensity profile $\vec{\Psi}_p$ and the RNA propensity profile $\vec{\Psi}_r$ weighted by the *interaction matrix* I:

$$\pi = \vec{\Psi}_p \mathrm{I} \vec{\Psi}_r \quad (3)$$

The matrix I has been derived using a *Montecarlo* approach to guarantee optimal space sampling in the parameters space. The algorithm predicts the interaction propensity of a protein-RNA pair reporting the *discriminative power* DP, which is a measure of the interaction potential with respect to the training sets [29]. DP ranges from 0% (the case of interest is predicted to be negative) to 100% (the case of interest is predicted to be positive). In general, DP values above 50% indicate that the interaction is likely to take place, whereas DPs above 75% represent high-confidence predictions. The *cat*RAPID *graphic* module predicts the interaction propensity of a protein-RNA pair reporting the DP and a heatmap of the interaction scores along the sequences. The module accepts protein sequences with a length ranging between 50 and 750 amino acids and RNA sequences between 50 and 1200 nucleotides and is more accurate on small transcripts [32].

catRAPID strength. This module calculates the interaction of a protein-RNA pair with respect to a reference set [32]. Random associations between polypeptide and nucleotide sequences are used for the reference set. Reference sequences have the same lengths as the pair of interest to guarantee that the interaction strength is independent of protein and RNA lengths [32]. The interaction strength ranges from 0% (no interaction) to 100% (strong interaction). Interaction strengths above 50% indicate a high propensity to bind (Figure 2). In a previous study, it has been observed that the strength correlate with chemical affinities [32], which suggests that the interactions propensity can be used to estimate the strength of association. It is important to mention that the interaction strength provides a better estimate of the binding than the discriminative power, as it is evaluated on a larger set of interactions and excludes potential biases arising from protein/RNA sequences lengths.

*cat*RAPID *fragments.* Due to the conformational space of nucleotide chains, prediction of RNA secondary structures is difficult when RNA sequences are > 1200 nucleotides and simulations cannot be completed on standard processors (2.5 GHz; 4 to 8 GB memory). To overcome this limitation, a procedure called *fragmentation* was introduced. This involves the division of polypeptide and nucleotide sequences into fragments followed by prediction of the interaction propensities [31,32]. Two types of fragmentation are possible:

- *Protein and RNA* uniform fragmentation (for transcripts smaller than 10000 nucleotides) [31]: The fragmentation approach is based on the division of protein and RNA sequences into overlapping segments. This analysis of fragments is particularly useful to identify protein and RNA regions involved in the binding [23,31].

- *Long RNA* weighted fragmentation (for transcripts larger than 10000 nucleotides) [32]: The use of RNA fragments is introduced to identify RNA regions involved in protein binding (Figure 3). The RNALfold algorithm from the Vienna package is employed to select RNA fragments in the range 100-200 nucleotides with predicted stable secondary structure [32].

*cat*RAPID *omics.* We have recently developed an algorithm to allow fast calculation of ribonucleoprotein associations in *Caenorhabditis elegans, Danio rerio, Drosophila melanogaster, Homo sapiens, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae and Xenopus tropicalis* [39]. The algorithm computes the interaction between a molecule (protein or transcript) and the pre-compiled reference library (transcriptome or proteome) for each model organism. In addition to the *interaction propensities, discriminative power* and *interaction strength,* the approach allows detection of RNA-binding regions in proteins and recognition motifs in RNA molecules. The method has been validated on Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) data and predicts associations with high significance (p-values < 0.05).

Examples of predictions and comparison between predictive methods

In a recent study, the *cat*RAPID approach has been employed to investigate the occurrence of ribonucleoprotein associations in biological pathways [23]. In this analysis, the interaction potential was computed for 295 × 10^6 protein-RNA pairs reported in Reactome [40] and 65 × 10^6 associations available from NCI-Nature Pathway Interaction Database (NCI-PID) [41]. One of the main results of this study is that around 1000 genes encoding aggregation-prone and structurally disordered proteins have a high propensity to interact with their own mRNAs (autogenous interactions). Here, experimental evidence available in literature is used to compare *cat*RAPID performances with other computational methods (Table 2)[42–48] on autogenous interactions:

- Heterogeneous nuclear ribonucleoprotein L hnRNP-L is able to induce non-sense mediated decay by binding to its own mRNA [48]. Our predictions, carried out with *cat*RAPID fragments ("Long RNA" fragmentation option [32]; see Methods*, catRAPID fragments*) indicate that hnRNP-L interacts with its own transcript within three different intronic regions located between exons 1-2, 6-7 and 9-10, which is in agreement with experimental evidence [48]. More specifically, hnRNP-L is predicted to bind with strong propensity to the 3' CA cluster 6A (interaction strength = 84 %;

Figure 3; Table 2) of the *hnRNP-L* RNA (intron 6 of transcript ENST00000221419 corresponding to nucleotides 39332858-39332174 of NC_000019.9) and not to sequence 6A (position 39332443-39332174; interaction strength = 1 %; Figure 3; Table 2), which is in agreement with the *in vitro* assays performed by Rossbach *et al.* [48]. Similarly to our calculations, *RPIseq* predicts region 39332858-39332174 to be interacting with hnRNP-L (RF score = 0.75 and SVM score = 0.88), while fragment 39332443-39332174 has RF score = 0.85 and SVM score = 0.77.

As reported in Table 2, *RPIseq* shows excellent true positive rate and high false positive rate. It is likely that, due to the heterogeneous composition of training datasets, algorithms show different predictive power. Nevertheless, it is advisable to use all the available methods, as comparative analyses provide precious information for the designing of new experiments.

The examples used here (Table 2) refer to interactions occurring between transcriptional and translational products of the same gene. *cat*RAPID predictions indicate that a large number of proteins undergo autogenous associations in intronic/UTR regions [23]. As the maximum levels of mRNA expression are intrinsically correlated with the aggregation rates of encoded proteins [49,50], autogenous interactions could represent a homeostatic mechanism to regulate expression via feedback loops, thus limiting protein production and the tendency of proteins to aggregate [51,52]. In this regards, it is likely that autogenous interactions play a major role in regulation the expression of dosage-sensitive genes [53,54]. At present, we do not know if self-regulatory mechanisms represent a way of avoiding production of highly concentrated and potentially toxic protein products [23] or derive from a primordial and ribosomal-independent mechanism of translation [55].

*cat*RAPID *omics express*

*cat*RAPID *omics express* (http://service.tartaglialab.com/page/catrapid_express_omics_group) is a recent implementation of our *cat*RAPID *omics* [39] algorithm to investigate the connection between expression networks and interaction propensities of protein-RNA pairs [24] (Table 1). Our algorithm allows the calculation of both interaction propensities and expression patterns for a given protein with respect to the human transcriptome (or given RNA with respect to the human nucleic-acid binding proteome). Using this approach, we found that interaction between RBPs and mRNAs is with high statistical significance related to the probability that the two molecules have linked patterns of expression in a number of human tissues [24]. More specifically, it has been observed a strong enrichment in functions related to cell-cycle control for positively correlated patterns and survival, growth and differentiation for negatively correlated patterns. Intriguingly, about 90% of genes in both categories are listed in the gene index of the National Institutes of Health's Cancer Genome Anatomy Project, with a large number of tumor suppressors featuring in the former category and many transcription regulators appearing in the latter. Our analysis reveals that modifications in the expression network could trigger aberrant interactions that lead to pathogenic events, including cancer [24].

To show the performance of *cat*RAPID *omics express*, which is here released with a web service interface, we collected recent CLIP experiments [56–60] and assessed the ability of the algorithm to predict interactions between RBPs and their targets with available expression data (Figure 4). *cat*RAPID *omics express* predictions achieves significant performances (p-values < 0.05; Fisher's exact test) in remarkable agreement with genome-wide experimental data.

In these calculations, expression profiles are derived from RNA sequencing data in 14 human tissues (ArrayExpress: E-MTAB-513) [61]. The normalized relative abundances are assigned respectively to proteins and RNAs using a homology-based criterion [24]. Pearson's coefficient calculated across expression levels for all the tissues represents the correlation of the constitutive expression levels associated with every protein-RNA pair. The absolute value of expression correlation is added to the sum of interaction propensity values to rank the results [39]. Quantitative predictions on the binding propensities of full-length proteins (alternatively, nucleic acid binding regions) and transcripts (alternatively, predicted stable secondary structure fragments) are provided as output.

Concluding remarks

The field of protein-RNA interactions is moving fast and a number of fascinating hypotheses have been recently formulated on the evolution of ribonucleoprotein complexes [1,62]. Computational models represent an important source of information that can be exploited to identify trends, understand the principles of molecular recognition and design new experiments. Indeed, improvement of theoretical models and subsequent validation of predictions is crucial to achieve a better description of the role of coding and non-coding RNAs in protein networks, especially in human disease [63]. As shown for *cat*RAPID *omics express*, computational methods greatly benefit from integration with experimental data coming from different sources, including lncRNAdb (repository for long noncoding RNAs) [64], ☐NRED (database long noncoding RNA expression) [65], ☐NONCODE (integrated knowledge dataset of non-coding RNAs) [66], ☐HMDD (human microRNA disease database) [67], ☐OMIM (list of human genes and genetic disorders) [68] and ☐GAD (Genetic Association Database) [69].

Synergy between computational and experimental approaches is expected to improve our understanding of ribonucleoprotein networks. At present, two important challenges can be identified for future research: i) development of methods to accurately predict RNA structure; ii) integration of existing tools to elucidate mechanisms leading to formation of complexes such as ribonucleoprotein aggregates.

Structural models. *cat*RAPID calculations rely on the *Vienna* algorithm to generate accurate predictions of secondary structure ensembles [70]. In the future, it will be crucial to improve performances of computational approaches to achieve a more accurate characterization of RNA regions involved in protein binding. At present, classical experimental methods for RNA structure determination include X-ray crystallography, NMR, cryo-electron microscopy and chemical and enzymatic probing. However, these methods are only applicable to analyze a single RNA per experiment and constrained by the length of probed transcripts.

A relatively new and promising large-scale technique for structure determination is Parallel Analysis of RNA Structure (PARS), which is based on deep sequencing of precise RNA fragments generated by single strand specific enzyme S1 and double-strand specific enzyme V1 [71]. A similar approach exploits high-throughput sequencing of fragments generated by single-strand specific nuclease P1 and has been applied to non-coding RNAs

in different cells [72]. In this case, the Selective 2'-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) chemistry, combined with multiplexed bar coding and next generation sequencing, was able to measure the structures of a complex pool of RNAs [73].

Methods based on technologies such as PARS and SHAPE could be very useful for investigation of RNA structure and will provide new data to train predictive algorithms. Nevertheless, it is important to mention that the structure measured with PARSE and SHAPE could be significantly different from that observed *in vivo* [74], as proteins are known to influence RNA folding.

Ribonucleoprotein aggregates. Using *cat*RAPID to investigate protein-RNA associations, it has been observed that several proteins including Muscle-blind-like MBNL1 and the heterogeneous nuclear ribonucleoproteins hnRNP-A1, hnRNP-A2/B1, hnRNP-C, hnRNP-D, hnRNP-E, and hnRNP-G, bind to CGG repetitions in the 5' UTR of *FMR1* [31]. These ribonucleoprotein associations are particularly relevant because they occur in a neurodegenerative disorder called Fragile X-associated tremor/ataxia syndrome [75,76].

How often do RNA molecules promote sequestration of proteins in the cell? Previous studies have reported cases of phase separation in cytoplasm and nucleoplasm, which, similarly to lipid-raft formation in membranes, results in the formation of droplets [77]. These droplets define specific, non-membrane-bound accumulations rich in proteins and RNA (examples include nucleoli, stress granules and Cajal bodies), and are in many cases known to be the sites of mRNA storage, processing, and decay [78,79]. Intriguingly, it has been proposed that the packaging of cytoplasmic mRNA into discrete ribonucleoprotein granules regulates gene expression by delaying the translation of specific transcripts [80]. At present, it is not possible to state if ribonucleoprotein granules are functional assemblies or pathological transitions to amyloid structures [79]. As a matter of fact, recent experiments showed that several disease-related mutations of TDP-43 and FUS promote granule formation [81].

What are the molecular features underlying the formation of ribonucleoprotein aggregates? Theoretical approaches for prediction of protein aggregation could provide insights into this mechanism [76–78]. Indeed, aggregation can be predicted with high accuracy using physico-chemical features such as hydrophobicity, secondary structure propensity and solvent accessibility [85]. According to our calculations, structural disorder

regions of proteins interact with RNA [24] and this could have a strong impact on aggregation [86] and toxicity [87]. It is possible that stable RNA secondary structures, especially those enriched in GC content, contribute to the spatial rearrangement of disordered regions of proteins [23]. We envisage that the simultaneous investigation of RNA-binding ability and aggregation propensity of proteins will be key to understand pathogenesis of several disorders, including neurodegeneration and cancer [62].

In conclusion, the methods and ideas discussed here have been developed in an exciting moment of the post-genomic era [61]. For the very first time, experimental and computational approaches have started to unveil the complexity of our genomes and protein-RNA emerged as key players in a large number of regulatory processes [88]. It is our hope that the works presented hereby will inspire other researchers to validate the large-scale models and generate new hypotheses.

Acknowledgements

The authors would like to thank Domenica "Mimma" Marchese, Andreas Zanzoni, Benedetta Bolognesi, Giovanni Bussotti and Roderic Guigo' for stimulating discussions.

Our work was supported by the Ministerio de Economia y Competividad (SAF2011-26211 to G.G.T.) and the European Research Council (ERC Starting Grant to G.G.T.).

References

1. A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, N. E. Davey, D. T. Humphreys, T. Preiss, L. M. Steinmetz, J. Krijgsveld, and M. W. Hentze, *Cell*, 2012, **149**, 1393–1406.
2. A. G. Baltz, M. Munschauer, B. Schwanhäusser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, and M. Landthaler, *Mol. Cell*, 2012, **46**, 674–690.
3. H. Siomi and G. Dreyfuss, *Curr. Opin. Genet. Dev.*, 1997, **7**, 345–353.
4. D. D. Licatalosi and R. B. Darnell, *Nat Rev Genet*, 2010, **11**, 75–87.
5. M.-C. Tsai, O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, and H. Y. Chang, *Science*, 2010, **329**, 689–693.
6. A. M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn, *Proc. Natl. Acad. Sci. U.S.A.*, 2009, **106**, 11667–11672.
7. I. Iglesias-Platas, A. Martin-Trujillo, D. Cirillo, F. Court, A. Guillaumet-Adkins, C. Camprubi, D. Bourc'his, K. Hata, R. Feil, G. Tartaglia, P. Arnaud, and D. Monk, *PLoS ONE*, 2012, **7**, e38907.
8. C. Tuerk and L. Gold, *Science*, 1990, **249**, 505–510.
9. M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A.-C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl, *J Vis Exp*, 2010.
10. S. Kishore, L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, and M. Zavolan, *Nat. Methods*, 2011, **8**, 559–564.
11. M.-L. Ankö and K. M. Neugebauer, *Trends Biochem. Sci.*, 2012, **37**, 255–262.
12. T. Puton, L. Kozlowski, I. Tuszynska, K. Rother, and J. M. Bujnicki, *Journal of structural biology*, 2011.
13. J. D. Keene, J. M. Komisarow, and M. B. Friedersdorf, *Nat. Protocols*, 2006, **1**, 302–307.
14. S. Mili and J. A. Steitz, *RNA*, 2004, **10**, 1692–1694.
15. J. König, K. Zarnack, N. M. Luscombe, and J. Ule, *Nat. Rev. Genet.*, 2011, **13**, 77–83.
16. I. Huppertz, J. Attig, A. D'Ambrogio, L. E. Easton, C. R. Sibley, Y. Sugimoto, M. Tajnik, J. König, and J. Ule, *Methods*, 2013.
17. C. Chu, K. Qu, F. L. Zhong, S. E. Artandi, and H. Y. Chang, *Mol. Cell*, 2011, **44**, 667–678.
18. M. D. Simon, C. I. Wang, P. V. Kharchenko, J. A. West, B. A. Chapman, A. A. Alekseyenko, M. L. Borowsky, M. I. Kuroda, and R. E. Kingston, *Proc. Natl. Acad. Sci. U.S.A.*, 2011, **108**, 20497–20502.
19. A. M. Khalil and J. L. Rinn, *Semin. Cell Dev. Biol.*, 2011, **22**, 359–365.
20. M. Guttman and J. L. Rinn, *Nature*, 2012, **482**, 339–346.
21. M. Parisien, X. Wang, G. Perdrizet II, C. Lamphear, C. A. Fierke, K. C. Maheshwari, M. J. Wilde, T. R. Sosnick, and T. Pan, *Cell Reports*, 2013, **3**, 1703–1713.
22. F. Di Palma, F. Colizzi, and G. Bussi, *RNA*, 2013, **19**, 1517–1524.
23. A. Zanzoni, D. Marchese, F. Agostini, B. Bolognesi, D. Cirillo, M. Botta-Orfila, C. M. Livi, S. Rodriguez-Mulero, and G. G. Tartaglia, *Nucl. Acids Res.*, 2013, gkt794.

24. D. Cirillo, D. Marchese, F. Agostini, C. M. Livi, T. Botta-Orfila, and G. G. Tartaglia, *Genome Biol.*, 2014, **15**, R13.
25. M. Terribilini, J. D. Sander, J.-H. Lee, P. Zaback, R. L. Jernigan, V. Honavar, and D. Dobbs, *Nucleic Acids Res.*, 2007, **35**, W578–584.
26. M. Fernandez, Y. Kumagai, D. M. Standley, A. Sarai, K. Mizuguchi, and S. Ahmad, *BMC Bioinformatics*, 2011, **12 Suppl 13**, S5.
27. P. Klus, B. Bolognesi, F. Agostini, D. Marchese, A. Zanzoni, and G. G. Tartaglia, *Bioinformatics*, 2014.
28. M. Bellucci, F. Agostini, M. Masin, and G. G. Tartaglia, *Nat. Methods*, 2011, **8**, 444–445.
29. E. W. Stawiski, L. M. Gregoret, and Y. Mandel-Gutfreund, *J. Mol. Biol*, 2003, **326**, 1065–1079.
30. T. Wu, J. Wang, C. Liu, Y. Zhang, B. Shi, X. Zhu, Z. Zhang, G. Skogerbø, L. Chen, H. Lu, Y. Zhao, and R. Chen, *Nucleic Acids Res*, 2006, **34**, D150–152.
31. D. Cirillo, F. Agostini, P. Klus, D. Marchese, S. Rodriguez, B. Bolognesi, and G. G. Tartaglia, *RNA*, 2013, **19**, 129–140.
32. F. Agostini, D. Cirillo, B. Bolognesi, and G. G. Tartaglia, *Nucleic Acids Res.*, 2013, **41**, e31.
33. V. Pancaldi and J. Bähler, *Nucleic Acids Res.*, 2011, **39**, 5826–5836.
34. U. K. Muppirala, V. G. Honavar, and D. Dobbs, *BMC Bioinformatics*, 2011, **12**, 489.
35. B. A. Lewis, R. R. Walia, M. Terribilini, J. Ferguson, C. Zheng, V. Honavar, and D. Dobbs, *Nucleic Acids Res*, 2011, **39**, D277–D282.
36. Y. Wang, X. Chen, Z.-P. Liu, Q. Huang, Y. Wang, D. Xu, X.-S. Zhang, R. Chen, and L. Chen, *Mol. BioSyst.*, 2012, **9**, 133–142.
37. A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neubock, and I. L. Hofacker, *Nucleic Acids Research*, 2008, **36**, W70–W74.
38. D. Cirillo, F. Agostini, and G. G. Tartaglia, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2013, **3**, 161–175.
39. F. Agostini, A. Zanzoni, P. Klus, D. Marchese, D. Cirillo, and G. G. Tartaglia, *Bioinformatics*, 2013, **29**, 2928–2930.
40. D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein, *Nucleic Acids Res.*, 2011, **39**, D691–697.
41. C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, *Nucl. Acids Res.*, 2009, **37**, D674–D679.
42. C. Schaeffer, B. Bardoni, J.-L. Mandel, B. Ehresmann, C. Ehresmann, and H. Moine, *EMBO J*, 2001, **20**, 4803–4813.
43. A. Sureau, R. Gattoni, Y. Dooghe, J. Stévenin, and J. Soret, *EMBO J.*, 2001, **20**, 1785–1796.
44. Y. M. Ayala, L. De Conti, S. E. Avendaño-Vázquez, A. Dhir, M. Romano, A. D'Ambrogio, J. Tollervey, J. Ule, M. Baralle, E. Buratti, and F. E. Baralle, *EMBO J*, 2011, **30**, 277–288.
45. E. Chu, D. M. Koeller, J. L. Casey, J. C. Drake, B. A. Chabner, P. C. Elwood, S. Zinn, and C. J. Allegra, *Proc. Natl. Acad. Sci. U.S.A.*, 1991, **88**, 8977–8981.
46. E. Chu, T. Takechi, K. L. Jones, D. M. Voeller, S. M. Copur, G. F. Maley, F. Maley, S. Segal, and C. J. Allegra, *Mol. Cell. Biol.*, 1995, **15**, 179–185.

47. N. M. Parakhnevitch, A. V. Ivanov, A. A. Malygin, and G. G. Karpova, *Mol Biol*, 2007, **41**, 44–51.
48. O. Rossbach, L.-H. Hung, S. Schreiner, I. Grishina, M. Heiner, J. Hui, and A. Bindereif, *Mol. Cell. Biol.*, 2009, **29**, 1442–1451.
49. G. G. Tartaglia, S. Pechmann, C. M. Dobson, and M. Vendruscolo, *Trends Biochem Sci*, 2007, **32**, 204–6.
50. A. J. Baldwin, T. P. J. Knowles, G. G. Tartaglia, A. W. Fitzpatrick, G. L. Devlin, S. L. Shammas, C. A. Waudby, M. F. Mossuto, S. Meehan, S. L. Gras, J. Christodoulou, S. J. Anthony-Cahill, P. D. Barker, M. Vendruscolo, and C. M. Dobson, *J. Am. Chem. Soc.*, 2011, **133**, 14160–14163.
51. G. G. Tartaglia and M. Vendruscolo, *Mol. BioSyst.*, 2009, **5**, 1873–1876.
52. P. Ciryam, G. G. Tartaglia, R. I. Morimoto, C. M. Dobson, and M. Vendruscolo, *Cell Rep*, 2013, **5**, 781–790.
53. H. Moriya, K. Makanae, K. Watanabe, A. Chino, and Y. Shimizu-Yoshida, *Mol Biosyst*, 2012, **8**, 2513–2522.
54. R. Sopko, D. Huang, N. Preston, G. Chua, B. Papp, K. Kafadar, M. Snyder, S. G. Oliver, M. Cyert, T. R. Hughes, C. Boone, and B. Andrews, *Molecular Cell*, 2006, **21**, 319–330.
55. C. R. Woese, D. H. Dugre, W. C. Saxinger, and S. A. Dugre, *Proc Natl Acad Sci U S A*, 1966, **55**, 966–974.
56. M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano Jr, A.-C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl, *Cell*, 2010, **141**, 129–141.
57. Z. Wang, M. Kayikci, M. Briese, K. Zarnack, N. M. Luscombe, G. Rot, B. Zupan, T. Curk, and J. Ule, *PLoS Biol.*, 2010, **8**, e1000530.
58. J. I. Hoell, E. Larsson, S. Runge, J. D. Nusbaum, S. Duggimpudi, T. A. Farazi, M. Hafner, A. Borkhardt, C. Sander, and T. Tuschl, *Nat. Struct. Mol. Biol.*, 2011, **18**, 1428–1431.
59. D. T. Vo, D. Subramaniam, M. Remke, T. L. Burton, P. J. Uren, J. A. Gelfond, R. de Sousa Abreu, S. C. Burns, M. Qiao, U. Suresh, A. Korshunov, A. M. Dubuc, P. A. Northcott, A. D. Smith, S. M. Pfister, M. D. Taylor, S. C. Janga, S. Anant, C. Vogel, and L. O. F. Penalva, *Am. J. Pathol.*, 2012, **181**, 1762–1772.
60. Y. Xue, Y. Zhou, T. Wu, T. Zhu, X. Ji, Y.-S. Kwon, C. Zhang, G. Yeo, D. L. Black, H. Sun, X.-D. Fu, and Y. Zhang, *Mol. Cell*, 2009, **36**, 996–1006.
61. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard, *Genome Res.*, 2012, **22**, 1760–1774.
62. B. Wolozin, *Molecular Neurodegeneration*, 2012, **7**, 56.
63. G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan, and Q. Cui, *Nucleic Acids Res.*, 2013, **41**, D983–986.
64. P. P. Amaral, M. B. Clark, D. K. Gascoigne, M. E. Dinger, and J. S. Mattick, *Nucleic Acids Res.*, 2011, **39**, D146–151.
65. M. E. Dinger, K. C. Pang, T. R. Mercer, M. L. Crowe, S. M. Grimmond, and J. S. Mattick, *Nucleic Acids Res.*, 2009, **37**, D122–126.

66. D. Bu, K. Yu, S. Sun, C. Xie, G. Skogerbø, R. Miao, H. Xiao, Q. Liao, H. Luo, G. Zhao, H. Zhao, Z. Liu, C. Liu, R. Chen, and Y. Zhao, *Nucleic Acids Res.*, 2012, **40**, D210–215.
67. Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, *Nucleic Acids Res.*, 2014, **42**, D1070–1074.
68. J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, *Nucleic Acids Res*, 2009, **37**, D793–D796.
69. K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, *Nat. Genet.*, 2004, **36**, 431–432.
70. I. L. Hofacker, *Nucleic Acids Research*, 2003, **31**, 3429 –3431.
71. M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal, *Nature*, 2010, **467**, 103–107.
72. Y. Wan, M. Kertesz, R. C. Spitale, E. Segal, and H. Chang, *Nat Rev Genet*, 2011, **12**.
73. J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin, *Proc Natl Acad Sci U S A*, 2011, **108**, 11063–11068.
74. M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal, *Nature*, 2010, **467**, 103–107.
75. P. J. Hagerman and R. J. Hagerman, *Am. J. Hum. Genet.*, 2004, **74**, 805–816.
76. C. Sellier, F. Rau, Y. Liu, F. Tassone, R. K. Hukema, R. Gattoni, A. Schneider, S. Richard, R. Willemsen, D. J. Elliott, P. J. Hagerman, and N. Charlet-Berguerand, *EMBO J.*, 2010, **29**, 1248–1261.
77. R. Narayanaswamy, M. Levy, M. Tsechansky, G. M. Stovall, J. D. O'Connell, J. Mirrielees, A. D. Ellington, and E. M. Marcotte, *PNAS*, 2009, pnas.0812771106.
78. L. Malinovska, S. Kroschwald, and S. Alberti, *Biochim. Biophys. Acta*, 2013, **1834**, 918–931.
79. M. Ramaswami, J. P. Taylor, and R. Parker, *Cell*, 2013, **154**, 727–736.
80. N. Kedersha and P. Anderson, *Meth. Enzymol.*, 2007, **431**, 61–81.
81. T. Murakami, S.-P. Yang, L. Xie, T. Kawano, D. Fu, A. Mukai, C. Bohm, F. Chen, J. Robertson, H. Suzuki, G. G. Tartaglia, M. Vendruscolo, G. S. K. Schierle, F. T. S. Chan, A. Moloney, D. Crowther, C. F. Kaminski, M. Zhen, and P. St George-Hyslop, *Human Molecular Genetics*, 2011.
82. G. G. Tartaglia, A. Cavalli, R. Pellarin, and A. Caflisch, *Protein Sci*, 2005, **14**, 2723–34.
83. G. G. Tartaglia, A. P. Pawar, S. Campioni, C. M. Dobson, F. Chiti, and M. Vendruscolo, *J Mol Biol*, 2008, **380**, 425–36.
84. F. Agostini, M. Vendruscolo, and G. G. Tartaglia, *J. Mol. Biol.*, 2012, **421**, 237–241.
85. G. G. Tartaglia and M. Vendruscolo, *Chem Soc Rev*, 2008, **37**, 1395–401.
86. H. Olzscha, S. M. Schermann, A. C. Woerner, S. Pinkert, M. H. Hecht, G. G. Tartaglia, M. Vendruscolo, M. Hayer-Hartl, F. U. Hartl, and R. M. Vabulas, *Cell*, 2011, **144**, 67–78.
87. T. Vavouri, J. I. Semple, R. Garcia-Verdugo, and B. Lehner, *Cell*, 2009, **138**, 198–208.
88. S. Altman, *RNA*, 2013, **19**, 589–590.
89. A. Zanzoni, D. Marchese, F. Agostini, B. Bolognesi, D. Cirillo, M. Botta-Orfila, C. M. Livi, S. Rodriguez-Mulero, and G. G. Tartaglia, *Nucleic Acids Res.*, 2013, **41**, 9987–9998.
90. F. Zalfa, M. Giorgi, B. Primerano, A. Moro, A. Di Penta, S. Reis, B. Oostra, and C. Bagni, *Cell*, 2003, **112**, 317–327.
91. F. Zalfa, S. Adinolfi, I. Napoli, E. Kühn-Hölsken, H. Urlaub, T. Achsel, A. Pastore, and C. Bagni, *J. Biol. Chem.*, 2005, **280**, 33403–33410.

92. C. Lacoux, D. D. Marino, P. P. Boyl, F. Zalfa, B. Yan, M. T. Ciotti, M. Falconi, H. Urlaub, T. Achsel, A. Mougin, M. Caizergues-Ferrer, and C. Bagni, *Nucl. Acids Res.*, 2012, gkr1254.

Table 1. *cat*RAPID *modules.* Synopsis of *cat*RAPID algorithms, their use and related examples [23,24,31,32].

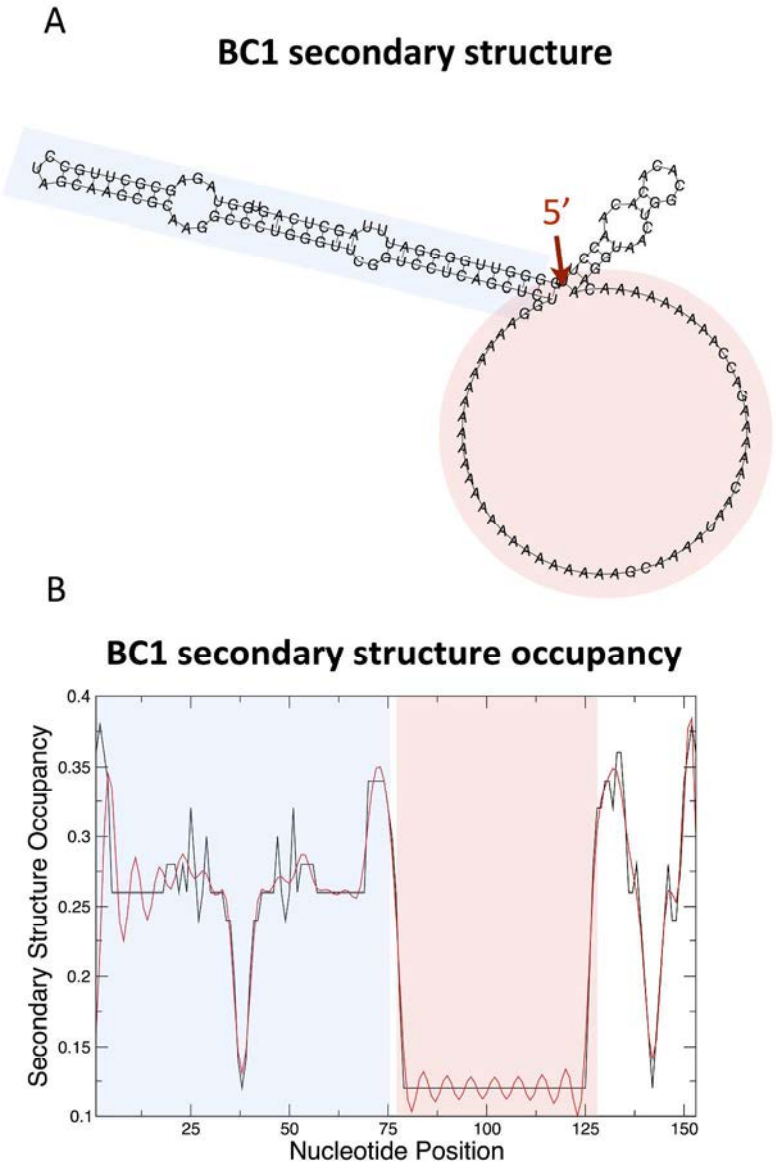
Table 2*. Predictions and comparison between predictive methods.* Interaction scores of known associations (bold characters) and negative controls. The *cat*RAPID [32] and RPIseq [34] performances are compared on autogenous interactions [89].

Figure 1. *Secondary structure occupancy.* A) Example of secondary structure prediction for the non-coding RNA BC1 as predicted by Vienna RNAfold (centroid model) [90]. B) High values of the *secondary structure occupancy* profile [28] indicate that base pairing occurs in regions with high propensity to form stem loops (blue box), while low values are associated with loops or junctions (pink region).

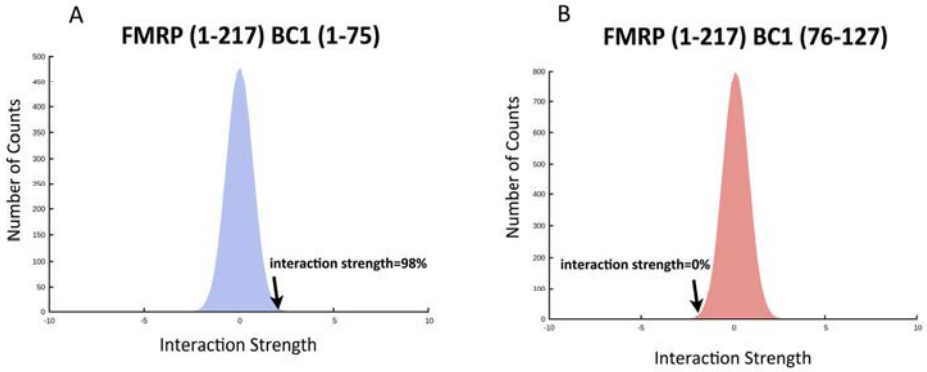
Figure 2. *Interaction strength.* In agreement with experimental evidence [91,92], we predict that the N-terminus of fragile X mental retardation protein FMRP (amino acids 1-217) (A) binds to the 5' stem loop of BC1 transcript (nucleotides 1-75), (B) does not interact with the loop region of BC1 transcript (nucleotides 76-127). Here, the interaction strength algorithm is used to estimate the interaction propensity of the protein-RNA pair [31].

Figure 3. *Long RNA fragmention*. A) Using the *catRAPID fragments* algorithm [23,28], we are able to reproduce experimental evidence on the interaction of hnRNP-L with its own transcript [48]. Our predictions indicate that the binding occurs in three different intronic regions located between exons 1-2, 6-7 and 9-10, in agreement with experimental evidence [48]; B) We predict that hnRNP-L protein binds with high affinity (interaction strength = 84%) to the 3' CA cluster 6A of the hnRNP-L gene and not to C) the control 6A (interaction strength = 1%), as shown by *in vitro* splicing assays performed by Rossbach *et al.* [48].

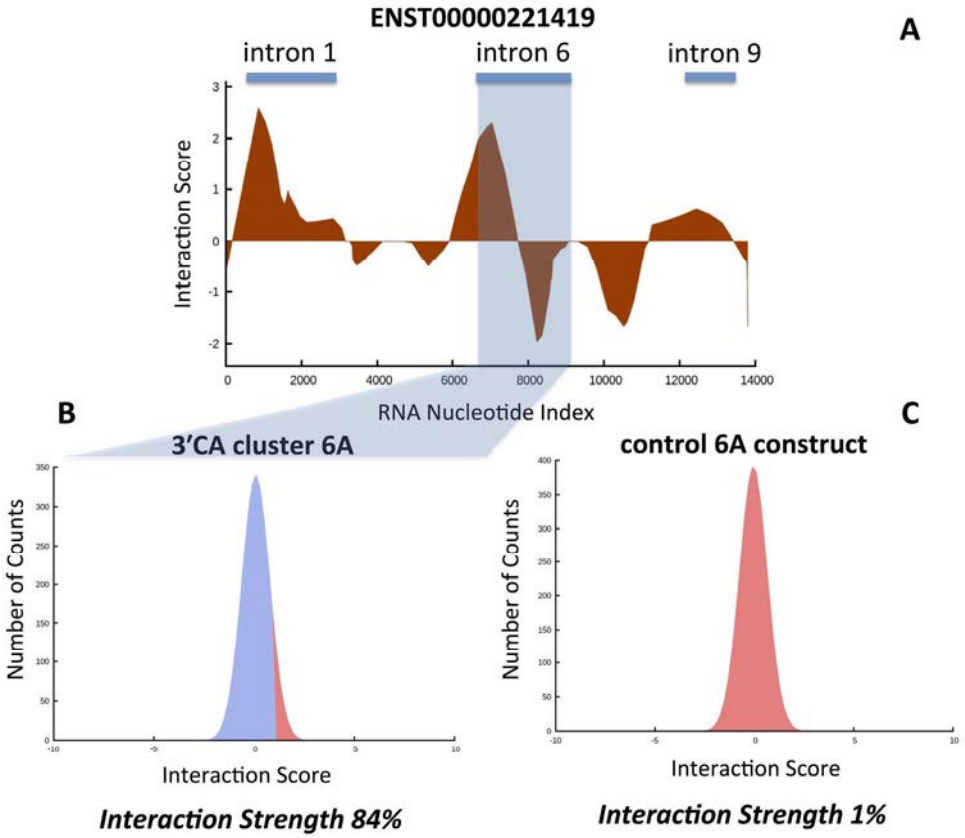
Figure 4. *cat*RAPID *omics express.* We show performances of our new algorithm *cat*RAPID *omics express* [24] on the interactomes of IGF2B1 (Insulin-like growth factor 2 mRNA-binding protein 1), TIA1 (T-cell-restricted intracellular antigen-1), FUS (Translocated in liposarcoma protein), MSI (RNA-binding protein Musashi homolog 1) and PTBP1 (Polypyrimidine tract-binding protein 1 PTB1) [9,57–60]. The significance of our predictions was assessed using Fisher's exact test (dashed line corresponds to *p-value* = 0.1) and 0.9-quantile of rank score distribution as performance measure. (FUS: 1030 interactions; MSI: 352 interactions; PTBP1: 1567 interactions; TIA1: 1237 interactions; IGF2BP1-3: 3299 interactions).



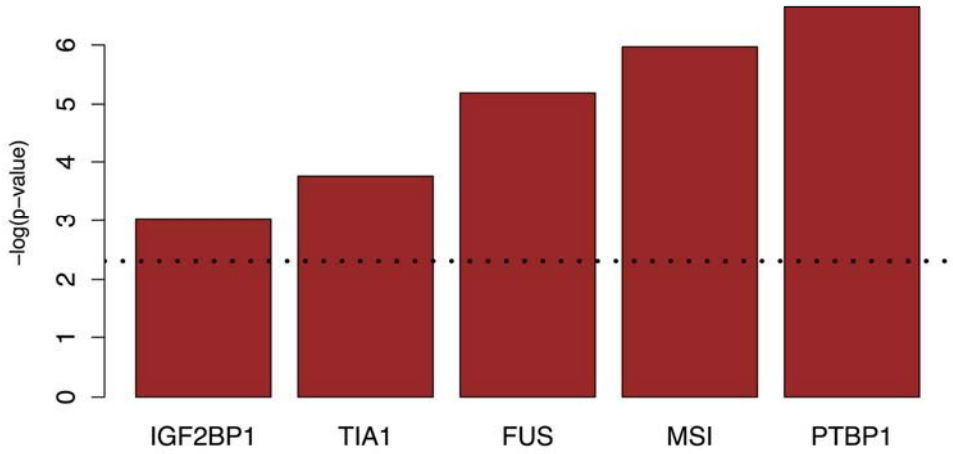
Secondary structure occupancy. A) Example of secondary structure prediction for the non-coding RNA BC1 as predicted by Vienna RNAfold (centroid model) 90. B) High values of the secondary structure occupancy profile 28 indicate that base pairing occurs in regions with high propensity to form stem loops (blue box), while low values are associated with loops or junctions (pink region).
259x389mm (300 x 300 DPI)



Interaction strength. In agreement with experimental evidence 91,92, we predict that the N-terminus of fragile X mental retardation protein FMRP (amino acids 1-217) (A) binds to the 5′ stem loop of BC1 transcript (nucleotides 1-75), (B) does not interact with the loop region of BC1 transcript (nucleotides 76-127). Here, the interaction strength algorithm is used to estimate the interaction propensity of the protein-RNA pair 31.
138x58mm (300 x 300 DPI)



Long RNA fragmention. A) Using the catRAPID fragments algorithm 23,28, we are able to reproduce experimental evidence on the interaction of hnRNP-L with its own transcript 48. Our predictions indicate that the binding occurs in three different intronic regions located between exons 1-2, 6-7 and 9-10, in agreement with experimental evidence 48; B) We predict that hnRNP-L protein binds with high affinity (interaction strength = 84%) to the 3' CA cluster 6A of the hnRNP-L gene and not to C) the control 6A (interaction strength = 1%), as shown by in vitro splicing assays performed by Rossbach et al. 48.
190x166mm (300 x 300 DPI)



catRAPID omics express. We show performances of the algorithm catRAPID omics express 24 on the interactomes of IGF2B1 (Insulin-like growth factor 2 mRNA-binding protein 1), TIA1 (T-cell-restricted intracellular antigen-1), FUS (Translocated in liposarcoma protein), MSI (RNA-binding protein Musashi homolog 1) and PTBP1 (Polypyrimidine tract-binding protein 1 PTB1) 9,56–59. The significance of our predictions was assessed using Fisher's exact test (dashed line corresponds to p-value = 0.1) and 0.9-quantile of rank score distribution as performance measure. (FUS: 1030 interactions; MSI: 352 interactions; PTBP1: 1567 interactions; TIA1: 1237 interactions; IGF2BP1-3: 3299 interactions).
127x79mm (300 x 300 DPI)

Type of Analysis	Algorithm	Features	Result	Examples
The protein-RNA pair of interest are < 750 aa and 1200 nt in length	*cat*RAPID *graphic* and *strength* modules	The *graphic* module calculates the interaction propensity of a protein-RNA pair. The *strength* module computes the interaction propensity with respect to a reference set.	The score will provide the *propensity* to interact as well as an estimate of the *strength* of interaction	RNAse P, HOTAIR [28]
The protein (or RNA) is larger than 750 aa (1200 nt)	*cat*RAPID fragments (*protein and RNA* option)	The algorithm automatically divides protein and RNA sequences into fragments and predicts interaction propensities.	The *binding sites* of both molecules are ranked and visualized	FMRP, TDP43 [31]
The RNA is > 10000 nt and the protein < 750 aa	Fragment module (*long RNA* option)	The algorithm divides the protein sequence into fragments. The entire protein is used to calculate the interaction propensity against the most stable local structures of the RNA. The interaction propensity is calculated between the protein and each RNA fragment.	The *binding sites* of the protein on the RNA sequence are provided	Xist [32], hnRNP-L
What are the protein (transcript) partners of an RNA (protein) of interest?	*cat*RAPID omics	The algorithm omputes the interaction between a protein (or transcript) and the transcriptome (or nucleotide-binding proteome) of a organism.	*Propensity, strengths, binding motifs* are ranked in a table	SRSF1, FUS [39]
What are the interacting protein (transcript) partners that are co-expressed in human tissues?	*cat*RAPID omics express	The algorithm allows identification of co-expressed protein and RNA pairs in human tissues.	*Propensity, strengths, binding motifs* and correlations of expression patterns are shown	TIA1, QKI [24]

protein	RNA		*cat*RAPID (interaction strength)	RPI*seq* (RF score)	RPI*seq* (SVM score)	Reference
FMRP	*FMR1* (XM_005262323.1)	**3`UTR (1744-1844)**	81%	0.60	0.43	[42]
		3`UTR (224-877)	1%	0.75	0.95	
SRSF2	*SRSF2* (NM_003016.4)	**region I/II of terminal exon (2521-2591)**	84%	0.15	0.16	[43]
		3`UTR (2592-2959)	0%	0.80	0.88	
TDP-43	*TARDBP* (XM_005263435.1)	**CDS (2271-2366)**	99%	0.60	0.90	[44]
		CDS (2838-3321)	21%	0.70	0.97	
TYMS	*TYMS* (XM_005258137.1)	**5` region (15-170)**	99%	0.55	0.52	[45] [46]
		3`UTR (994-1289)	18%	0.70	0.98	
RPS13	*RPS13* NC_000011.9	**intron1 (17099186-17098794)**	99%	0.65	0.84	[47]
		3`UTR (17095974-17095936)	4%	0.65	0.89	
hnRNP-L	*hnRNP-L* (NC_000019.9)	**intron 6 (39332858-39332174)**	84%	0.75	0.88	[48]
		intronic region 6A (39332443-39332174)	1%	0.85	0.77	