Chapter 2

A Computational Approach for the Discovery of Protein–RNA Networks

Domenica Marchese, Carmen Maria Livi, and Gian Gaetano Tartaglia

Abstract

Protein–RNA interactions play important roles in a wide variety of cellular processes, ranging from transcriptional and posttranscriptional regulation of genes to host defense against pathogens. In this chapter we present the computational approach *cat*RAPID to predict protein–RNA interactions and discuss how it could be used to find trends in ribonucleoprotein networks. We envisage that the combination of computational and experimental approaches will be crucial to unravel the role of coding and noncoding RNAs in protein networks.

Key words Protein–RNA interactions, Interaction prediction, Ribonucleoprotein networks, Messenger RNA, Noncoding RNA, *cat*RAPID

1 Introduction

The human genome harbors >1500 genes encoding proteins containing at least one RNA-binding domain (RBD) [1]. The number of proteins with identified RNA-binding ability (RBP), either possessing canonical or noncanonical RBDs [2, 3], is increasing. The fact that some proteins bind to transcripts through domains or regions that are not specifically evolved to this precise purpose [3, 4] is particularly intriguing. Indeed, recent manuscripts suggest a scenario where unexpected players can exert crucial functions in processes that were previously thought of as exclusively regulated by selected RBD-containing proteins [5].

Computational models represent an important source of information that can be exploited to identify hidden trends and understand the basics of molecular recognition. As a matter of fact, bioinformatics tools can perform exhaustive analyses and extract distinctive features, hence facilitating the design of new experiments. For example, it has been shown in several studies that the

Erik Dassi (ed.), *Post-Transcriptional Gene Regulation*, Methods in Molecular Biology, vol. 1358,
DOI 10.1007/978-1-4939-3067-8_2, © Springer Science+Business Media New York 2016

composition of primary protein structure, and the physicochemical properties associated with it, can be used to describe the amino acid regions that are more likely to be involved in binding to RNA molecules [6, 7]. Due to the limitations of current experimental approaches, it remains difficult to simultaneously investigate the plethora of RBPs bound to a single transcript and RNA regions that are likely to be involved in the binding. This has resulted in experimentalists having to rely on protein analysis to investigate specific signatures.

We developed an algorithm, *cat*RAPID, to investigate protein–RNA associations involved in regulatory mechanisms [8]. We trained *cat*RAPID on a large set of protein–RNA pairs available in the Protein Data Bank [9] to discriminate interacting and noninteracting molecules using the information contained in primary structures. *cat*RAPID relies on the ViennaRNA package [10], which has an accuracy of ~76 % [11], to generate predictions of secondary structure ensembles. These structures are then analyzed to extract information on the pairing profile of each nucleotide. By means of this procedure, the probability of *cat*RAPID predicting a protein–RNA interaction has a 72 % correlation with secondary structure information. However, a higher correlation factor is consistently expected with the enhancement of secondary structure prediction accuracies. As the predictive power of global RNA structure becomes less accurate as the length of the RNA increases [12], we developed the *cat*RAPID *fragments* module that exploits the RNALfold algorithm [11] to determine interactions for the most stable local structure.

2 *cat*RAPID Modules

The *cat*RAPID approach (http://s.tartaglialab.com/page/catrapid_group) [8, 13] has been developed to predict protein associations with coding and noncoding RNAs [14, 15] (Table 1). In our method, the contributions of secondary structure, hydrogen bonding, and van der Waals are combined together into the *interaction profile*:

$$\vec{\Phi}_x = \alpha_{\mathrm{H}} \vec{\mathrm{H}}_x + \alpha_{\mathrm{W}} \vec{\mathrm{W}}_x + \alpha_{\mathrm{S}} \vec{\mathrm{S}}_x$$

where the variable x indicates RNA ($x = \underline{\mathrm{r}}$) or protein ($x = \mathrm{p}$). The hydrogen bonding profile, denoted by $\vec{\mathrm{H}}$, is the hydrogen bonding ability of each amino acid (or nucleotide) in a protein (or RNA) sequence:

$$\vec{\mathrm{H}} = \mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_{\mathrm{length}}$$

Table 1
Algorithms of the *cat*RAPID suite. Computational models, their applications and examples

Type of analysis	Algorithm	Result	Examples
The protein–RNA pair of interest are <750 aa and 1200 nt in length	*cat*RAPID *graphic* and *strength* modules	The score will provide the *propensity* to interact as well as an estimate of the *strength* of interaction	CSR system [13] FMRP [16]
Protein (or RNA) is larger than 750 aa (1200 nt)	*cat*RAPID *fragments* (*protein and RNA* option)	The *binding sites* of both molecules are visualized	SNCA [28] UNR (this work)
RNA is >10,000 nt and protein <750 aa	*cat*RAPID *fragments* (*long RNA* option)	The *binding sites* of the protein on the RNA sequence are identified	hnRNP-L [18] Xist [14]
Protein (transcript) partners of an RNA (protein) of interest	*cat*RAPID *omics*	*Propensity, strengths, binding motifs* are ranked in a table	HuR [19] LIN28B [19]
Interacting protein (transcript) partners co-expressed in human tissues	*cat*RAPID *omics express*	*Propensity, strengths, binding motifs* and *expression patterns* are characterized	TIA1 [18] MSI [18]

Similarly, $\vec{S}$ represents the secondary structure occupancy profile and $\overrightarrow{W}$ the van der Waals' profile. The *interaction propensity* π is defined as the inner product between the protein propensity profile $\overrightarrow{\Psi}_p$ and the RNA propensity profile $\overrightarrow{\Psi}_r$ weighted by the *interaction matrix I*:

$$\pi = \overrightarrow{\Psi}_p I \overrightarrow{\Psi}_r$$

The algorithms to compute protein–RNA interactions are available at our group webpage http://service.tartaglialab.com/page/catrapid_group

2.1 catRAPID Graphic

Our original algorithm predicts the interaction propensity of a protein–RNA pair reporting the discriminative power DP, which is a measure of interaction strength with respect to the training sets [8]. The DP ranges from 0 % (the case of interest is predicted to be negative) to 100 % (the case of interest is a positive). DP values above 50 % indicate that the interaction is likely to take place, whereas DPs above 75 % represent high-confidence predictions. Due to computational requirements (intense CPU usage), the *cat*RAPID graphic algorithm accepts protein sequences with a length between 50 and 750 aa and RNA sequences between 50 and 1200 nt [15].

2.2 catRAPID Fragments

When input sequences exceed the length compatible with our computational requirements (i.e.: protein length > 750aa or RNA length > 1200 nt), the *cat*RAPID *graphic* cannot be used to calculate the interaction propensity [14, 15]. To overcome this limitation, we developed a procedure called *fragmentation*, which cuts polypeptide and nucleotide sequences into fragments followed by the prediction of the interaction propensities. Two types of fragmentation are possible:

- *Protein and RNA uniform fragmentation* (transcripts < 10,000 nt) [15]: The fragmentation approach is based on the division of protein and RNA sequences into 104 overlapping segments. This analysis is particularly useful to identify regions involved in the binding.
- *Long RNA* weighted fragmentation (for transcripts > 10,000 nt) [14]: The use of RNA fragments is introduced to identify RNA regions involved in protein binding. The RNALfold algorithm from Vienna package is employed to select RNA fragments in the range between 100 and 200 nt with predicted stable secondary structure.

2.3 catRAPID Strength

We previously observed that the strength correlates with chemical affinities [14], which suggests that the interaction propensity can be used to estimate the strength of association [16]. *cat*RAPID *strength* algorithm calculates the strength of a protein–RNA pair with respect to a reference set [13]. Random associations between polypeptide and nucleotide sequences are used to build the reference set. Since little interaction propensities are expected from random associations, the reference set is considered a negative control. Reference sequences have the same length as the pair of interest to guarantee that the interaction strength is independent of protein and RNA length. The interaction strength ranges from 0 % (noninteracting) to 100 % (interacting). Interaction strengths above 50 % indicate propensity to bind.

2.4 catRAPID Omics

The method is based on *cat*RAPID [8] algorithm and performs high-throughput predictions of protein–RNA interactions. *cat*RAPID *omics* enables: (1) the calculation of protein–RNA interactions on a large scale (up to 105 associations) in a reasonable time; (2) the submission of protein and RNA sequences without any length restriction; and (3) to focus on specific protein regions able to bind nucleic acid molecules [17] (Table 2).

- The time required by the original *cat*RAPID algorithm for predicting a single RNA–protein interaction strictly depends on the features of the input molecules, which are computed on the fly for each submission (using parallel calculation, <10 min are required for proteomic interactions of one RNA molecule).

Table 2
Composition of reference libraries used in *cat*RAPID *omics*

	Proteome					
	Full proteins		Domains		Transcriptome	
Model organisms	RNA	DNA	RNA	DNA	Coding	Noncoding
Caenorhabditis elegans	79	304	255	339	16613	8385
Danio rerio	82	323	311	391	21752	4589
Drosophila melanogaster	71	283	318	447	6307	1109
Homo sapiens	472	2152	1907	7432	105586	18553
Mus musculus	379	1518	1573	3073	42951	7243
Rattus norvegicus	168	592	689	902	13593	4823
Saccharomyces cerevisiae	261	389	508	431	3711	396
Xenopus tropicalis	70	184	279	253	2260	1278
Total	1582	5745	5840	13,268	98548	46376

Full-length (protein between 50 and 750 amino acids in length) and domains (derived from proteins >50 amino acids in length) are used as input of the method. Both sets are divided in additional groups, based on the ability of proteins to bind to RNA or DNA. Transcriptome searches use coding and noncoding RNAs, depending on the annotation in ENSEMBL version 68. The length of the transcripts in the datasets ranges from 50 to 1200 nucleotides, but longer RNAs can be added to the libraries

- To speed up the calculation of a far greater number of interactions, we introduce in *cat*RAPID *omics* a system of organism-specific feature libraries.

2.5 catRAPID Extensions

*cat*RAPID is interfaced with other methods to improve its predictive power [18]. Very recent implementations include the analysis of co-expression networks [19], the *clever*Suite approach to predict the RNA-binding ability of proteins [20] and the SeAMotE algorithm to identify regulatory elements coding/noncoding transcripts [21]:

- To train the *clever*Suite (http://s.tartaglialab.com/page/clever_suite), we focused on RNA-interacting proteins detected with UV cCL and PAR-CL protocols on proliferating HeLa cells followed by sequencing and compared them with the rest of cell lysate [3]. Analysis of physicochemical properties revealed a strong and consistent RNA binding property of the dataset (RNA-binding scales [3, 22, 23] discriminate 32–35 % of the entire database). The *clever*Suite selects the scales for nucleic acid binding [22, 23], membrane [24], burial [25] and aggregation [26] propensities, achieving a sensitivity of 0.72 and false positive rate of 0.24 on the entire dataset.

We applied the *clever*Suite to proteins that are classified as putative RNA-binding because they lack the canonical RNA-binding domains [3]. We observed correct classification associated with a sensitivity of 0.83 and false positive rate of 0.15, which indicates very high agreement with experimental data.

- Detection of regulatory motifs is a challenging task. For this reason, we developed the SeAMotE algorithm (http://s.tartaglialab.com/new_submission/seamote) [21], which provides an easy-to-use interface and allows the exhaustive analysis of large-scale datasets. Our approach offers unique features such as the discrimination based on the actual occurrences (i.e., pattern counts are not estimated) in the datasets, the choice of multiple reference backgrounds (shuffle, random, or custom) and the output of the most significant motifs in the whole span of tested motif widths, thus providing a wide range of solutions. In conclusion, our web-server is a powerful tool for the identification of enriched sequence patterns that characterize recognition process between proteins and nucleic acids. To evaluate SeAMotE performances on large-scale datasets, we collected recent CLIP experiments and assessed ability to identify significantly enriched motifs (Fisher's exact test). In each case analyzed, we compared RNAs bound to a specific protein (foreground set) with the same amount of non-interacting transcripts (background set). The DREME [27] algorithm was used as a reference to evaluate the performance of our system. Our method achieves both higher discrimination, which is the ability to separate the foreground from the background set, and significance, denoted by lower P-values associated with sequence motifs. In addition, SeAMoTe also shows very high sensitivity (~90 %) and accuracy (80 %).

3 *cat*RAPID Applications

3.1 Self-Regulatory Mechanisms Controlling Protein Production

We used the *cat*RAPID method to unravel self-regulatory pathways (autogenous interactions) controlling gene expression [15]. We discovered that aggregation-prone and structurally disordered proteins have a strong propensity to interact with their own mRNA [28]. Our results [15, 29] are in agreement with previous experimental work:

- It has been shown that the amyloidogenic TAR DNA binding protein 43 TDP-43 and *Fragile X* mental retardation protein FMRP interact with the 3′ UTR of their own mRNA to control protein production [15, 30, 31]. As overexpression leads to high protein concentration and enhanced amyloidogenicity [32, 33], it is possible that autogenous interactions prevent from generation of potentially toxic aggregates.

- The biosynthesis of tumor suppressor p53 is controlled by a translational autoregulatory feedback mechanism in which the p53 protein binds to its own mRNA in the 5′ terminal region, resulting in translational repression [34]. Indeed, it has been reported that naturally occurring mutations of p53 are associated with an increase of the aggregation potential [35]. In these regards, self-regulation of p53 can be seen as a way to control its aggregation potential.
- HSP70, the major stress-induced heat shock protein, regulates its own expression by interacting with its mRNA. Prolonged presence of HSP70 is detrimental for the cell, as it promotes aggregation. From ex vivo experiments, it has been shown that an increase in the degradation of HSP70 mRNA accompanies aggregation of HSP70 [36]. The interaction of HSP70 with its own mRNA (3′ UTR) suggests a self-limiting mechanism to reduce chaperone production and to avoid potential toxic effects in absence of stress [36].
- Moreover, the content of ribosomal proteins in eukaryotic cells is controlled by changes in the degradation rate of newly synthesized proteins. Such a high degree of coordination is achieved through the use of common regulatory elements in the genes and mRNAs of ribosomal proteins. In the majority of cases, regulation follows a feedback pattern, involving interactions of a ribosomal protein with its own pre-mRNA. This regulatory mechanism provides the required level of each individual ribosomal protein in the cell independently of other ribosomal proteins, which is crucial for extra-ribosomal functions. In the case of ribosomal proteins rpS26 and rpS13, high affinity for pre-mRNA fragments containing first introns has been found [37].

3.2 X-Chromosome Dosage Compensation

Dosage compensation of sex chromosomes equalizes expression of X-linked genes in organisms where males and females have a different number of X chromosomes. In mammals, Xist-mediated X chromosome inactivation (XCI) implies a complex network of macromolecular associations orchestrated by epigenetic modifiers as well as splicing and transcription factors.

- We used *cat*RAPID to investigate the interactions of the long noncoding Xist with Polycomb group proteins as well as YY1, SAF-A, ASF, and SATB1 proteins. In striking agreement with experimental evidence, we predicted protein binding sites and their affinities for Xist regions. We used our analysis to integrate the existing model of XCI into a new framework in which the transcriptional repressor YY1 tethers Xist to the X chromosome and nuclear matrix proteins SAF-A and SATB1 guide its translocation [14].

In *Drosophila melanogaster*, translational inhibition of male-specific *msl-2* messenger RNA by female-specific protein SXL is crucial for X-chromosome dosage compensation. Experimental studies identified an RNA-binding protein, UNR, as a fundamental co-repressor recruited by SXL to the 3′ UTR of *msl-2* mRNA for translation inhibition in females.

- RNA affinity chromatography and UV crosslinking assays show that UNR transcript and its 5′ UTR (nucleotides 1–261) efficiently bind to UNR protein, whereas 3′ UTR (nucleotides 261–447) does not [38]. Our calculations, carried out with *cat*RAPID fragments (“*Protein and RNA uniform fragmentation*” option), reproduce experimental results in great detail, identifying the cold shock domains (CSD; Fig. 1a) [39] involved

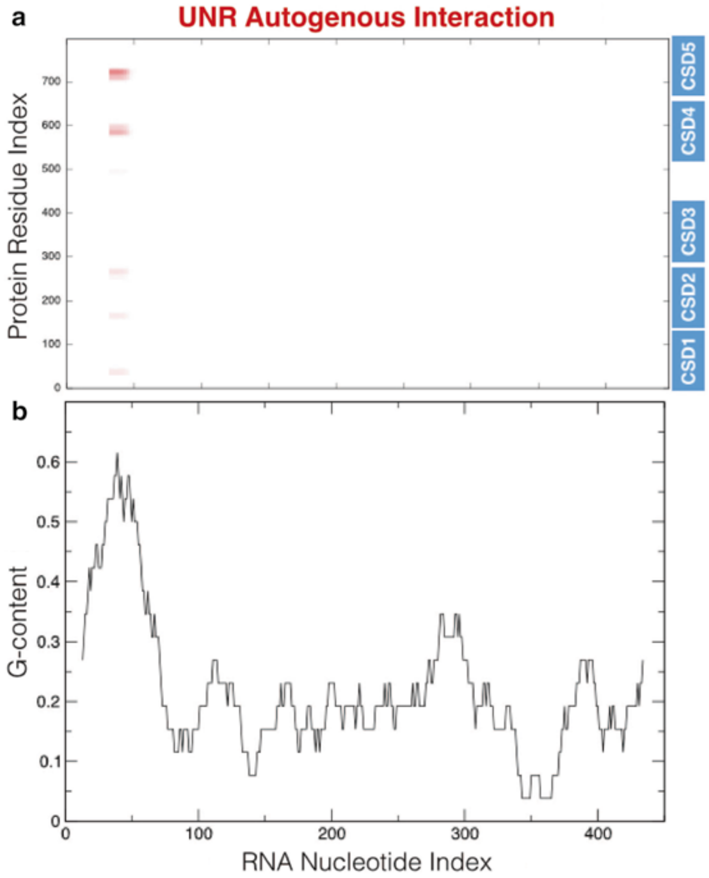

Fig. 1 UNR autogenous interactions. UNR transcript and its 5′ UTR (nucleotides 1–261) bind to UNR protein [38]. (**a**) Our calculations, carried out with *cat*RAPID fragments (“*Protein and RNA uniform fragmentation*” option), recapitulate experimental results in great detail, identifying cold shock domains (CSD) [39] involved in autogenous interaction; (**b**) In agreement with experimental evidence [38], UNR protein is predicted to bind to purine repeats, such as the guanine-rich region of UNR 5′ UTR

in autogenous interaction. In agreement with previous reports [38], UNR protein is predicted to bind to purine repeats, such as the guanine-rich region of its 5′ UTR (nucleotides 26–77; Fig. 1b).

3.3 Alternative Splicing

Recent studies indicate that nonsense-mediated decay (NMD) is an important element in alternative splicing regulation [40] and is associated with self-regulatory mechanisms:

- Polypyrimidine tract binding protein (PTB) regulates its own expression through a negative-feedback loop involving alternative splicing, which requires binding to mRNA and subsequent NMD triggered by exon skipping [41]. PTB autogenous interaction is particularly relevant because over-expression of the protein results in cell toxicity [42, 43].
- Similarly, heterogeneous nuclear ribonucleoprotein L hnRNP-L is able to induce NMD by associating with its mRNA [44]. Our predictions, carried out with *cat*RAPID fragments ("Long RNA" fragmentation option) indicate that hnRNP-L interacts with its own transcript in three different intronic regions located between exons 1–2, 6–7 and 9–10, which is in complete agreement with experimental evidence [44]. More specifically, we predict that hnRNP-L protein binds with high affinity to the 3′ CA cluster 6A of the hnRNP-L gene (intron 6) and not to sequence 6A (negative control), which is perfectly in agreement with the results of in vitro splicing assays performed by Rossbach et al. [44].

These and other results indicate that autogenous interactions occur in UTR/intronic regions and play a role in controlling protein production [28].

References

1. Ascano M, Gerstberger S, Tuschl T (2013) Multi-disciplinary methods to define RNA-protein interactions and regulatory networks. Curr Opin Genet Dev 23:20–28
2. Lunde BM, Moore C, Varani G (2007) RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol 8:479–490
3. Castello A et al (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell 149:1393–1406
4. Baltz AG et al (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol Cell 46:674–690
5. Kwon SC et al (2013) The RNA-binding protein repertoire of embryonic stem cells. Nat Struct Mol Biol 20:1122–1130
6. Terribilini M et al (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. Nucleic Acids Res 35: W578–W584
7. Fernandez M et al (2011) Prediction of dinucleotide-specific RNA-binding sites in proteins. BMC Bioinformatics 12(Suppl 13):S5
8. Bellucci M, Agostini F, Masin M, Tartaglia GG (2011) Predicting protein associations with long noncoding RNAs. Nat Methods 8: 444–445
9. Berman HM et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
10. Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31: 3429–3431

11. Lorenz R et al (2011) ViennaRNA Package 2.0. Algorithms Mol Biol 6:26
12. Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics 5:105
13. Cirillo D, Agostini F, Tartaglia GG (2013) Predictions of protein–RNA interactions. WIREs Comput Mol Sci 3:161–175
14. Agostini F, Cirillo D, Bolognesi B, Tartaglia GG (2013) X-inactivation: quantitative predictions of protein interactions in the Xist network. Nucleic Acids Res 41:e31
15. Cirillo D et al (2013) Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. RNA 19:129–140
16. Johnson R, Noble W, Tartaglia GG, Buckley NJ (2012) Neurodegeneration as an RNA disorder. Prog Neurobiol 99:293–315
17. Finn RD et al (2009) The Pfam protein families database. Nucleic Acids Res 38(Database): D211–D222
18. Cirillo D, Livi CM, Agostini F, Tartaglia GG (2014) Discovery of protein-RNA networks. Mol Biosyst 10:1632–1642
19. Cirillo D et al (2014) Constitutive patterns of gene expression regulated by RNA-binding proteins. Genome Biol 15:R13
20. Klus P et al (2014) The cleverSuite approach for protein characterization: predictions of structural properties, solubility, chaperone requirements and RNA-binding abilities. Bioinformatics 30:1601–1608
21. Agostini F, Cirillo D, Ponti RD, Tartaglia GG (2014) SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences. BMC Genomics 15:925
22. Terribilini M et al (2006) Prediction of RNA binding sites in proteins from amino acid sequence. RNA 12:1450–1462
23. Lewis BA et al (2011) PRIDB: a protein–RNA interface database. Nucleic Acids Res 39:D277–D282
24. Nakashima H, Nishikawa K, Ooi T (1990) Distinct character in hydrophobicity of amino acid compositions of mitochondrial proteins. Proteins 8:173–178
25. Wertz DH, Scheraga HA (1978) Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. Macromolecules 11:9–15
26. Pawar AP et al (2005) Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. J Mol Biol 350:379–392
27. Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27:1653–1659
28. Zanzoni A et al (2013) Principles of self-organization in biological pathways: a hypothesis on the autogenous association of alpha-synuclein. Nucleic Acids Res 41(22): 9987–9998
29. Agostini F et al (2013) catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. Bioinformatics 29(22):2928–2930
30. Ayala YM et al (2011) TDP-43 regulates its mRNA levels through a negative feedback loop. EMBO J 30:277–288
31. Schaeffer C et al (2001) The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif. EMBO J 20:4803–4813
32. Tartaglia GG, Pechmann S, Dobson CM, Vendruscolo M (2007) Life on the edge: a link between gene expression levels and aggregation rates of human proteins. Trends Biochem Sci 32:204–206
33. Baldwin AJ et al (2011) Metastability of native proteins and the phenomenon of amyloid formation. J Am Chem Soc 133:14160–14163
34. Ewen ME, Miller SJ (1996) p53 and translational control. Biochim Biophys Acta 1242: 181–184
35. Xu J et al (2011) Gain of function of mutant p53 by coaggregation with multiple tumor suppressors. Nat Chem Biol 7:285–295
36. Balakrishnan K, De Maio A (2006) Heat shock protein 70 binds its own messenger ribonucleic acid as part of a gene expression self-limiting mechanism. Cell Stress Chaperones 11:44–50
37. Parakhnevitch NM, Ivanov AV, Malygin AA, Karpova GG (2007) Human ribosomal protein S13 inhibits splicing of its own pre-mRNA. Mol Biol 41:44–51
38. Schepens B et al (2007) A role for hnRNP C1/C2 and Unr in internal initiation of translation during mitosis. EMBO J 26:158–169
39. Triqueneaux G, Velten M, Franzon P, Dautry F, Jacquemin-Sablon H (1999) RNA binding specificity of Unr, a protein with five cold shock domains. Nucleic Acids Res 27:1926–1934
40. Kalyna M et al (2011) Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. Nucleic Acids Res. doi:10.1093/nar/gkr932
41. Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA, Smith CWJ (2004) Autoregulation

of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. Mol Cell 13:91–100

42. Lin S, Wang MJ, Tseng K-Y (2013) Polypyrimidine tract-binding protein induces p19(Ink4d) expression and inhibits the proliferation of H1299 cells. PLoS One 8:e58227

43. Izquierdo JM et al (2005) Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. Mol Cell 19:475–484

44. Rossbach O et al (2009) Auto- and cross-regulation of the hnRNP L proteins by alternative splicing. Mol Cell Biol 29:1442–1451