Predicting functions of disordered proteins with MoRFpred

Christopher J. Oldfield[1], Vladimir N. Uversky[2,3,*] and Lukasz Kurgan[1,*]

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA
[2]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute,, Morsani College of Medicine, University of South Florida, Tampa, FL, 33612, USA
[3]Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, 142292, Russian Federation

*Corresponding authors: vuversky@health.usf.edu; lkurgan@vcu.edu.

Summary
Intrinsically disordered proteins and regions are involved in a wide range of cellular functions and they often facilitate protein-protein interactions. Molecular recognition features (MoRFs) are segments of intrinsically disordered regions that bind to partner proteins, where binding is concomitant with a transition to a structured conformation. MoRFs facilitate translation, transport, signaling and regulatory processes and are found across all domains of life. A popular computational tool, MoRFpred, accurately predicts MoRFs in protein sequences. MoRFpred is implemented as a user-friendly webserver that is freely available at http://biomine.cs.vcu.edu/servers/MoRFpred/. We describe this predictor, explain how to run the webserver, and show how to interpret the results it generates. We also demonstrate the utility of this webserver based on two case studies, focusing on the relevance of evolutionary conservation of MoRF regions.

Key words
Intrinsic disorder; prediction; Molecular recognition features; MoRFs; protein-protein interactions; MoRFpred.

1 Introduction
Intrinsically disordered proteins (IDPs) and protein regions (IDRs) are incompetent in forming stable three dimensional structure, yet perform varied and vital biological functions [1-4]. The lack of the prerequisite of a stable structure for function creates several challenges in the study of IDPs and IDRs, both experimental and computational [3]. The crux of these challenges on the computational side is the lack of conservation in many IDRs relative to structured proteins. Without the need to maintain rigid structures many IDRs diverge drastically, even in closely related species [5]. Lack of conservation confounds established methods for function annotation that rely on sequence similarity to transfer functional annotations. Lack of conservation is not universal in IDRs; many IDRs may be conserved, or more commonly conserved in portions of their sequences [5].

One mechanism of IDP function is short functional elements within IDRs. The evolutionary origin of these functional elements is seemingly idiosyncratic, where some examples have been found to be

evolutionarily conserved [6-8], and others have been proposed to be emergent sequence features [9,10]. A common function of these short functional elements is binding to molecular partners, often other proteins [11,12]. These types of features are likely common across many biological processes [13], such as cell cycle regulation, modulation of cellular structure, and apoptosis.

One model of these functional elements is known as molecular recognition features (MoRFs) [14]. It models functional elements within IDRs as short regions of increased structural propensity within longer regions of intrinsic disorder [13]. Examples of these types of functional regions can readily be inferred from protein structures and sequence properties [11]. Several predictors of MoRFs have been developed [13-17]. Initial predictors relied on direct interpretation of the MoRF model, by scanning for patterns in prediction of intrinsic disorder and employing a second level of prediction over patterns of interest [13,15]. The most recent MoRF predictors, including MoRFpred, relax the strict reliance on disorder prediction patterns while still directly considering disorder predictions [17]. Several other methods of MoRF prediction have been independently developed [13,14,18,16,19,20,15,21]. In addition to MoRFs, several other related models of functional elements with IDRs have been proposed. Eukaryotic linear motifs (ELMs) model these elements as short sequence motifs which can be predicted by pattern matching and filtering spurious matches [22]. Though they are very different models, MoRF and ELM predictions are frequently coincident [23]. Further, several generalized models of binding regions within IDRs have been developed [24-27]. Relative to other methods, MoRFpred was developed on a well-defined dataset with short functional elements that bind to other proteins within larger regions of intrinsic disorder. Like all methods of this type, the specificity is difficult to assess exactly, but this predictor features a good estimated sensitivity [17]. MoRFpred is useful for gaining insight into the function of novel IDPs.

MoRFpred is available as a user-friendly webserver at http://biomine.cs.vcu.edu/servers/MoRFpred/. This server was been extensively used by the community since it was released in early 2012. Usage data collected with the Google Analytics platform reveals that MoRFpred was utilized close to 9000 times by over 2700 unique users from 711 cities and 71 countries. The article that introduces this computational tool was already cited 175 times (source: Google Scholar on June 29, 2018).

2 Materials and Methods

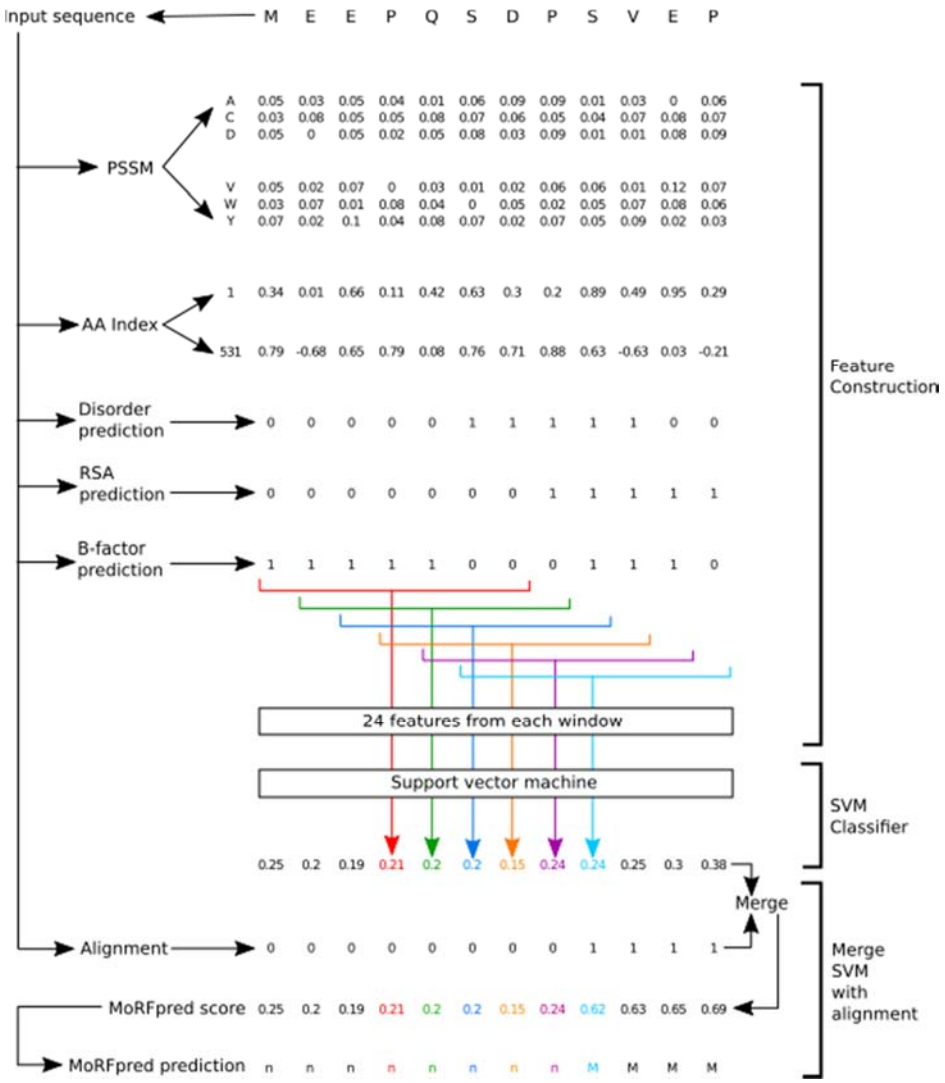
2.1 Datasets

For training of MoRFpred, a set of MoRFs was constructed beginning with known binding regions from Protein Data Bank (PDB) [28]. Bound peptides from PDB were carefully filtered for clear binding to a longer protein chain and mapped back to their source proteins. This procedure resulted in a dataset of 842 MoRFs. To avoid training and testing on similar proteins, these MoRFs were grouped into 427 clusters and divided into testing and training sets. This gave a training and testing sets with 421 and 419 MoRFs, respectively, with no protein more than 30% identical between the two sets (*see* **Note 1**).

A set of negative examples that do not contain MoRFs with near certainty was constructed from protein chains that have been completely structurally characterized by X-ray crystallography at a high-resolution. The chance of intrinsic disorder in the negative set was minimized by only selecting

monomeric proteins without large co-factors that contained no missing residues due to lack of electron density. Further, any protein with a significant amount of predicted intrinsic disorder, >30% of residues, was discarded. Filtering for proteins with less than 30% identity resulted in a set of 28 proteins.

Figure 1. Architecture of MoRFpred. The input sequence is used to generate sequence properties, from which input features are derived by windowed averaging. A support vector machine predicts MoRFs based on these input features. The SVM prediction is merged with similarity based predictions to produce the final MoRFpred score, where scores above 0.5 are predicted MoRFs (M) and those less than 0.5 are predicted non-MoRFs (n).



2.2 Architecture

MoRFpred is a support vector machine (SVM) over a rich feature space merged with a sequence similarity-based prediction (Fig. 1). Features considered for the linear kernel SVM predictor included 5 disorder prediction methods [29-32], relative solvent accessible surface prediction [33], B-factor prediction [34], PSI-BLAST generated position specific scoring matrices (PSSMs), and amino acid propensity scales from AAindex [35]. Two broad sets of features were used from each of these methods:

(1) per-residue over a window of 25 residues and (2) values aggregated over a window. Aggregation methods included taking the difference over a window of 25 residues and a smaller window, which captures the features found to be useful for previous MoRF predictors. For example, previous MoRF predictors relied on elevated predicted disorder surrounding a predicted MoRF, but depressed values for the MoRF region itself. Indeed, the corresponding difference-based aggregation was found to be one of the strongest MoRF features.

Feature selection for the SVM predictor was based on a best-first iterative addition of ranked features. Features were ranked based on a combination of biserial correlations [36] and single feature predictive performance, where poorly correlated or performing features were removed from consideration. Iterative addition of features was based on a modified 5-fold cross validation procedure, where a feature was only added if it improved prediction performance by at least 1%.

Similarity-based predictions were done using a PSI-blast search against MoRF containing proteins in the training set. PSI-blast matches were selected based on an *e*-value threshold. An *e*-value = 0.5 was selected based on optimization of performance of the merged predictor. MoRF annotations from the training set are transferred to the query protein based on the PSI-BLAST alignment. These transferred annotations are merged with SVM-based prediction by adding one to the SVM-prediction result and dividing by two, which ensures that the merged prediction for transferred annotations will be over the 0.5 threshold.

2.3 Predictive Quality

Prediction performance was assessed in the MoRFpred publication [17], using the true positive and false positive rates, overall accuracy (ACC), area under the ROC curve (AUC), and the success rate. Success rate is a per-sequence measure of performance, where a sequence is considered successfully predicted if the MoRF residues have a higher average prediction score than the non-MoRF residues.

MoRFpred performance was assessed in comparison to ANCHOR and previously developed MoRF predictors. MoRFpred had ACC = 94.7% and the highest performance by success rate and AUC evaluations, with values of 71.8% and 67.3%, respectively. The original MoRF predictor had a very low false positive rate, which artificially inflated its ACC value due to the large proportion of non-MoRF residues in the data set. Adjusting the MoRFpred threshold to an equally low false positive rate results in nearly double the true positive rate of the original MoRF predictor.

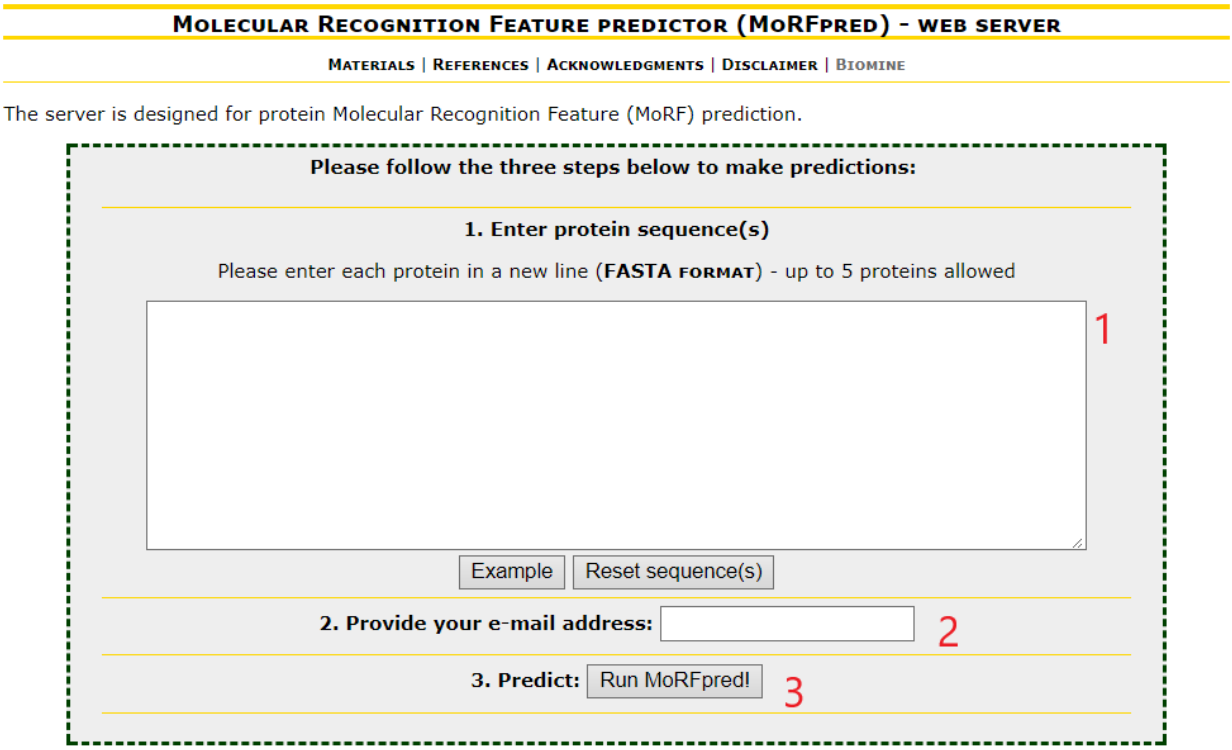
2.4 Webserver

The MoRFpred webserver is freely available at http://biomine.cs.vcu.edu/servers/MoRFpred/. The server can be accessed with an internet connection and any modern web browser. All computations that are needed to complete predictions are performed on the server side.

On our webserver, sequences submitted for prediction will be returned within 20-minutes of submission (*see* **Note 2**). The runtime of MoRFpred is dominated by the PSI-BLAST prediction, whose runtime varies with protein length and database similarity.

The main server page is where proteins are submitted for prediction. The webserver only requires FASTA sequences of the proteins of interest to preform MoRFpred predictions. Up to 5 FASTA formatted protein sequences may be entered into the large text entry field per submission. An email address is required for each submission. All required programs for generating prediction features, including PSI-BLAST, and disorder, RSA, and B-factor predictions, are run automatically by scripts on the server. Upon completion of predictions for each submission, the server will send an email notification with links to the prediction results.

Figure 2. Primary MoRFpred page, for submission of sequences for prediction. Red numbers indicate the sequence of steps required to submit a prediction.



2.5 Running MoRFpred

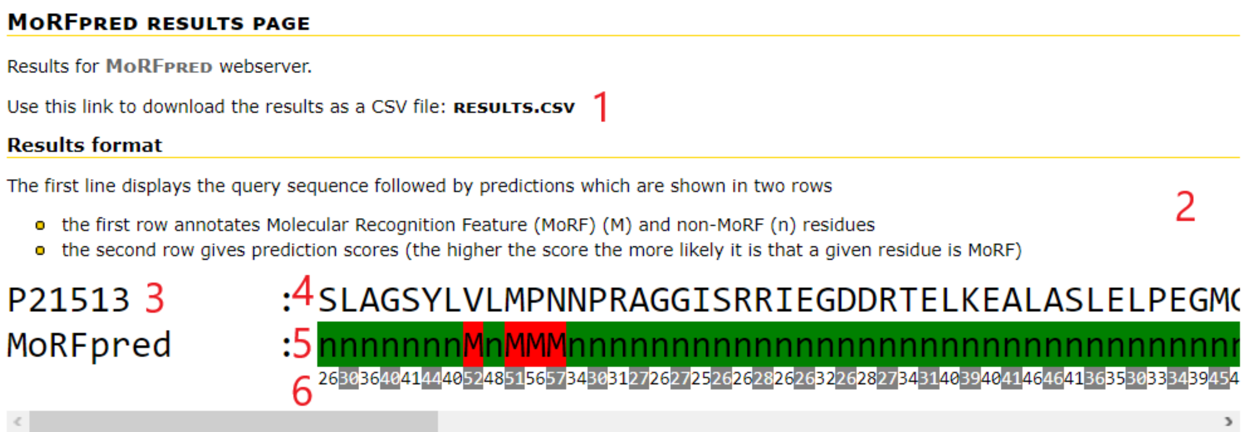
From the main server page, three steps are required to submit sequences to obtain the MoRFpred's predictions (Fig. 2, steps are highlighted with red numbers corresponding to the step):

1. Copy your FASTA formatted sequence (*see* **Note 3**) from its source file or webpage and paste it into the text box (*see* **Notes 4 and 5**).
2. Enter an email address. This is the address to which links to the prediction results will be sent.
3. Click "Run MoRFpred!". This submits the sequences to our server for MoRFpred predictions.

Once sequences are submitted for prediction, the browser is redirected to a status page that gives the current position of the submission in the server queue. This page will be automatically redirected to the

results page when predictions are completed. The queue on the server is first come first serve, and if there is a large number of submission, predictions may be delayed. Even if the web page is closed at this point, links to predictions will still be received through email (*see* **Note 6**).

Figure 3. MoRFpred prediction results page. Red numbers correspond to the primary features of the results page.



2.6 MoRFpred Results

The results page includes a link to the raw results (Fig. 3, red 1) as well as a color-coded text display of MoRFpred results (Fig. 3, red 2). The raw results (results.csv) file gives results for each submitted sequence, each in three lines, which are comma delimited:

- The input sequence: the FASTA header followed by each residue of the input sequence.
- Binary MoRFpred predictions: the string ‘MoRFpred’ followed by one character for each residue in the input sequence – ‘O’ for non-MoRF residues and ‘D’ for MoRF residues.
- The raw prediction output: the string ‘prob’ followed by one floating point number between 0 and 1 for each residue of the input sequence. Predicted MoRF residues correspond to values greater than 0.5.

The color-coded text display of MoRFpred predictions includes the FASTA header of each input sequence (Fig. 3, red 3), and several aligned rows. The rows are aligned by residue from N-terminus to C-terminus. The rows are, from top to bottom, are the input sequence (Fig. 3, red 4), binary MoRFpred results (Fig. 3, red 5) indicating non-MoRF residues (green ‘n’) and MoRF residues (red ‘M’) and the raw prediction value (Fig. 3, red 6) multiplied by 10 and rounded, with alternating residues in black and white.

The notification email contains links both to the results page (Fig. 4, red 1) and to the raw results file (Fig. 4, red 2). This email can be saved to access results at a later time.

Figure 4. Notification email. Red numbers correspond to links to prediction results.

Predictions for MoRFpred job id: 20171009045915 are ready.

Upon the usage the users are requested to use the following citations:

Miri Disfani F, Hsu WL, Mizianty MJ, Oldfield C, Xue B, Dunker AK, Uversky VN, Kurgan LA, 2012. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics, 28(12): i75-i83.

You can find the results for this job at: http://biomine.cs.vcu.edu/webresults/MoRFpred/20171009045915/results.html 1

The CSV file can be found here: http://biomine.cs.vcu.edu/webresults/MoRFpred/20171009045915/results.csv 2

The webserver can be found here: http://biomine.cs.vcu.edu/servers/MoRFpred/

Thank you for using our webserver,
Biomine group

3 Case Studies

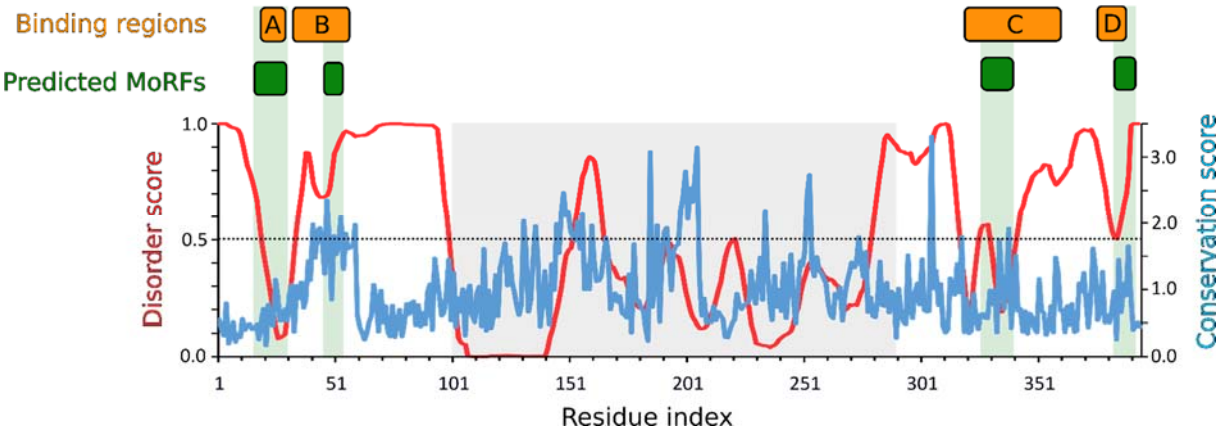
As subjects of the case studies we selected two proteins of different origin, human p53 (a 393 residue-long protein) and RNase E from *E.coli* (a 1,061 residue-long protein). These two proteins have very different biological functions, are characterized by different levels of intrinsic disorder, and possess different numbers of MoRFs.

3.1 Case Study: p53

Because of its crucial biological roles in regulation of apoptosis, genomic stability, and inhibition of angiogenesis, as well as many mechanisms of anticancer activity, cellular tumor antigen p53 is one of the most studied proteins. The p53 signaling pathway is activated in response to a variety of stress signals. Activated p53 is accumulated in the nucleus, where it's binding to specific DNA results in the induction or inhibition of a realm of different genes [37,38], many of which are involved in apoptosis, growth arrest, or senescence [39-42]. In the unstressed mammalian cells, continuous ubiquitination of the non-phosphorylated p53 by double minute-2 ubiquitin ligase (MDM2) [43] and subsequent proteasomal degradation ensure short life-time and low levels of p53. There is also a negative feedback between the p53 and Akt pathways [44], where Akt is activated in cells exposed to various stimuli ranging from hormones to growth factors, and to extracellular matrix components [45] and controls the MDM2-mediated targeting of p53 for degradation [46]. Loss of p53 function due to mutations in this protein or some other alterations in the pathways leading to its activation and regulation, is a common feature in the majority of human cancers [47]. Such mutations account for ~90% of cancer-related mutations in the *TP53* gene and are found in 50% of human cancers [48]. For example, up to 50% of advanced-stage prostate cancers contain mutations in p53 [49], and progression of prostate cancer to metastatic disease is characterized by the loss of p53 [50]. Furthermore, p53 levels may have prognostic value in urological oncology [51].

There are three major functional domains in human p53, the intrinsically disordered N-terminal regulatory domain (residues 1-92), the ordered central DNA binding domain (DBD, residues 94-292) [52-54], and the intrinsically disordered C-terminal oligomerization and regulatory domain (residues 293-393) [55]. The regulatory domains can be further subdivided into functional subdomains/regions, such as transactivation domain 1 (TAD1) (residues 1-40), TAD2 (residues 40-60), and a proline-rich region, PR (residues 64-92), in the N-terminal regulatory domain, and tetramerization or oligomerization domain (OD; residues 325-356), and a regulatory C-terminal domain (CTD; residues 356-393) in the C-terminal regulatory domain [55,56]. The N-terminal and C-terminal regulatory domains show exceptional binding promiscuity. Some of the illustrative examples of proteins interacting with the N-terminal transactivation region of p53 include CBP/p300, CSN5/Jab1, MDM2, RPA, TFIIH, and TFIID [57], whereas the CTD of p53 is engaged in interaction with 14-3-3, GSK3β, hGcn5, PARP-1, $S100B_{\beta\beta}$, TAF, TAF1, TRRAP, to name a few [57]. Importantly, despite their crucial role in biological activities of p53, the regulatory regions of this protein are characterized by relatively poor evolutionary conservation, whereas the central DBD domain is highly conserved among different species. Irrespective of the general lack of conservation, there are four MoRFs in human p53 that overlap with or are included into the known binding sites of this protein (see Figure 5).

Figure 5. Case study: p53. The correspondence between intrinsic disorder predictions (red line), sequence conservation (blue line), binding regions (orange boxes), and predicted MoRF regions (green boxes) is show. Binding regions are discussed in the text. Sequence conservation is calculated from a set of p53 orthologs (OrthoDB) as the relative profile entropy over maximum entropy weighted sequence (large values indicate greater conservation).



The first MoRF (see Figure 5, box A) coincides with the MDM2 binding site of p53. MDM2 is the E3 ubiquitin-protein ligase that is known as an important oncogene due to its overexpression in many human cancers, such as breast, colon, and prostate cancers, as well as hematologic malignancies and sarcomas [58]. MDM2 is most famous for its vital role in the p53 regulation via binding to a short stretch (residues 13-29) of the p53 TAD1 that prevents p53-driven activation or inhibition of various genes, via the MDM2-mediated p53 ubiquitination that targets this protein for the proteasomal degradation, and via active p53 transport out of the nucleus due to the presence of a nuclear export signal in MDM2 [59,60]. Therefore, alteration of the p53-MDM2 interaction pathway considered as a promising target

for cancer therapy [58]. X-ray crystallographic studies of the p53-MDM2 complex revealed that the MDM2 binding region of p53 forms an α-helical structure bound to a deep groove on the surface of the N-terminal domain of MDM2 (residues 17-125) [61].

The second MoRF (see Figure 5, box B) is included into the p53N fragment (residues 33-60) responsible for the p53 interaction with the N-terminal domain of the single-stranded DNA (ssDNA)-binding protein, replication protein A (RPA) [62]. This RPA70N domain is characterized by an oligonucleotide/ oligosaccharide-binding fold typical for the ssDNA-binding domains, whereas the p53N fragment, which is disordered in isolation, forms two amphipathic helices, H1 and H2, following RPA70N binding [62]. Also, unlike other MoRFs in this protein, this MoRF displays a large amount of sequence conservation (see Figure 5, conservation score).

The third MoRF (see Figure 5, box C) is a part of the p53 tetramerization domain (325-356), structure of which represents a short β-strand (residues 326–333) followed by an α-helix (residues 335–355). These two structural elements are connected by a sharp turn facilitated by a conserved glycine residue (Gly334). Two monomers of the p53 tetramerization domain associate to form an antiparallel double-stranded sheet and the antiparallel association of their helices forms a two-helical bundle. Four chains form a tetramer that can be described as a dimer of primary dimers [63].

The fourth MoRF (see Figure 5, box D) is a part of the highly promiscuous C-terminal binding region of p53 (residues 374-388) that can bind to cyclin A [64], sirtuin [65], CBP [66], or S100ββ [67]. It was pointed out that upon interaction with different partners, this binding region of p53 displays all three major secondary structure types in the four complexes [68], where its core fragment becomes an α-helix when bound to S100ββ [67], a β-strand when bound to sirtuin [65], and a coil with two distinct backbone trajectories when bound to CBP [66] and cyclin A2[64].

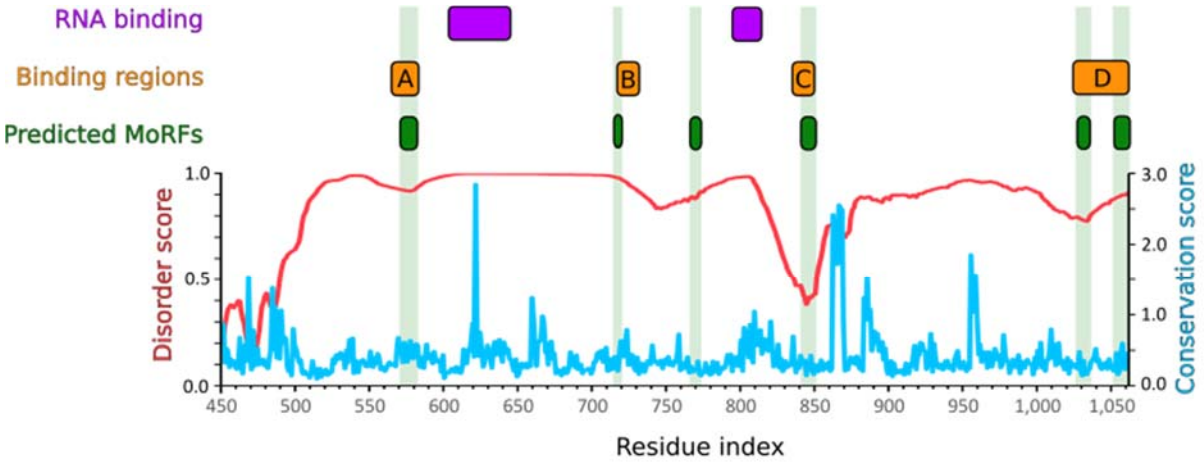
MoRFpred correctly identifies the four MoRF regions in p53 (see Figure 5, green boxes), in spite of the significantly different conservation profiles of the four MoRF regions (see Figure 5, blue lines). We note the relatively low conservation of the first, third and fourth MoRF region and much higher conservation values for the second region. Interestingly, the red lines that identify the putative propensities for disorder, which were generated with VSL2B [69], correctly identify both termini of p53 as intrinsically disordered. However, they also register dips where the MoRFs are located. These dips are a by-product of the fact that MoRF regions become structured upon interacting with the protein partner, reducing the inherent propensity of these amino acids to be intrinsically disordered.

3.2 Case Study: RNase E

Endoribonucleases are hydrolytic enzymes that catalyze the endonucleolytic cleavage of RNA, have various specificities, are universally present in all organisms, and typically operate under tight cellular regulation. Endoribonucleases are involved in the maturation, modification, and degradation of different RNAs [70]. There are at least five endoribonucleases in *E. coli* (RNases I*, III, E, G, P). Among various activities attributed to RNase E are processing of transfer RNA, 9S ribosomal RNA, the catalytic RNA of

RNase P, the transfer/messenger RNA (t/mRNA) that rescues stalled ribosomes [71-73], and general mRNA decay [74].

Figure 6. Case study: RNase E. The correspondence between intrinsic disorder predictions (red line), sequence conservation (blue line), binding regions (orange boxes), and predicted MoRF regions (green boxes) is show. Binding regions are discussed in the text. Sequence conservation is calculated from a set of RNase E orthologs (OrthoDB) as the relative profile entropy over maximum entropy weighted sequence (large values indicate greater conservation).



Being one of the larger *E. coli* proteins, RNase E consists of 1,061 amino acid residues [75,76]. There are two functionally different domains in this protein, the catalytic N-terminal domain (NTD; residues 1–498) and the regulatory C-terminal domain (CTD; residues 499–1,061) [77-79]. Although the NTD is relatively conserved and has numerous homologues [80], there is little sequence conservation in the CTD [81], which is also characterized by low sequence complexity. The purified CTD was shown to be mostly disordered by a set of biophysical techniques, such as limited proteolysis, SDS–PAGE, SAXS, and far-UV CD [82]. Despite being highly disordered, the CTD was shown to interact with other degradosome components and with structured RNA [82]. In agreement with these experimental data, computational analysis clearly indicated that the NTD of RNase E was expected to be mostly structured, whereas the CTD had characteristics of a highly disordered protein [82].

The CTD is highly disordered, which is in agreement with the high values of the putative propensities for disorder generated for this protein with VSL2B [69] (see Figure 6, red line). CTD is also characterized by the presence of four regions of increased structural propensity (labeled as segments A, B, C, and D, respectively), which correspond to MoRFs. The four MoRF were correctly identified by the MoRFpred method (green boxes). Importantly, all these segments are related to various biological activities of RNase E, such as membrane targeting and CTD self-association (segment A corresponding to residues 565–585) or interactions with the components of the RNA degradosome, helicase (segment B, which is a portion of the arginine-rich domain (residues 628–843)) [79,83], enolase (segment C (residues 833-850),) [82], and polynucleotide phosphorylase PNPase (segment D, RNase E residues 1021–1061) [82]. Like in the case of p53, some of the MoRF regions (see Figure 6, segments C and D) are concomitant

with a substantial decrease in the putative propensity for disorder (red line), but the remaining two regions do not register these dips. However, MoRFpred is still capable of identifying these MoRF regions, in spite of their high propensity for disorder and lack of conservation (blue line).

4 Notes

1. The datasets can be downloaded from http://biomine.cs.vcu.edu/servers/MoRFpred/
2. To partially compensate for the long runtime of the algorithm, up to five sequences can be submitted simultaneously to the webserver. As soon as the results for one batch of up to five sequences are returned, another set of sequences can be submitted.
3. In FASTA format, each sequence is prefixed by a line beginning with '>' followed by some identifying text. The sequence should begin on the following line. For example: …
4. Up to 5 sequences can be submitted at a time. Ensure that each sequence has its own 'FASTA header', which is a separate line beginning with '>'.
5. The maximum length of each submitted sequence is 1000 residues.
6. It is advised to store or bookmark the link at this point. Predictions are stored on the server for at least three months, and keeping the link will allow return to the results pages. It also protects against lost predictions, in the case that an incorrect notification email address was entered.

References

1. Wang C, Uversky VN, Kurgan L (2016) Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. Proteomics 16 (10):1486-1498.
2. Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, Hu G, Uversky VN, Kurgan L (2015) Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci 72 (1):137-151.
3. Habchi J, Tompa P, Longhi S, Uversky VN (2014) Introducing protein intrinsic disorder. Chem Rev 114 (13):6561-6588.
4. Dunker AK, Babu MM, Barbar E, Blackledge M, Bondos SE, Dosztányi Z, Dyson HJ, Forman-Kay J, Fuxreiter M, Gsponer J, Han K-H, Jones DT, Longhi S, Metallo SJ, Nishikawa K, Nussinov R, Obradovic Z, Pappu RV, Rost B, Selenko P, Subramaniam V, Sussman JL, Tompa P, Uversky VN (2013) What's in a name? Why these proteins are intrinsically disordered. Intrinsically Disordered Proteins 1 (1):e24157.
5. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ (2002) Evolutionary rate heterogeneity in proteins with long disordered regions. Journal of molecular evolution 55.
6. Meszaros B, Tompa P, Simon I, Dosztanyi Z (2007) Molecular principles of the interactions of disordered proteins. J Mol Biol 372 (2):549-561.
7. Trudeau T, Nassar R, Cumberworth A, Wong ET, Woollard G, Gsponer J (2013) Structure and intrinsic disorder in protein autoinhibition. Structure (London, England : 1993) 21 (3):332-341.
8. Varadi M, Guharoy M, Zsolyomi F, Tompa P (2015) DisCons: a novel tool to quantify and classify evolutionary conservation of intrinsic protein disorder. BMC Bioinformatics 16 (1).
9. Ait-Bara S, Carpousis AJ, Quentin Y (2015) RNase E in the gamma-Proteobacteria: conservation of intrinsically disordered noncatalytic region and molecular evolution of microdomains. Molecular genetics and genomics : MGG 290 (3):847-862.
10. Davey NE, Cyert MS, Moses AM (2015) Short linear motifs – ex nihilo evolution of protein regulation. Cell Communication and Signaling 13 (1):43.

11. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN (2006) Analysis of molecular recognition features (MoRFs). Journal of molecular biology 362 (5):1043-1059.
12. Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, Uversky VN, Dunker AK (2007) Characterization of molecular recognition features, MoRFs, and their binding partners. Journal of proteome research 6 (6):2351-2366.
13. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry 44 (37):12454-12470.
14. Yan J, Dunker AK, Uversky VN, Kurgan L (2016) Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst 12 (3):697-710.
15. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK (2007) Mining α-Helix-Forming Molecular Recognition Features with Cross Species Sequence Alignments†. Biochemistry 46 (47):13468-13477.
16. Malhis N, Gsponer J (2015) Computational identification of MoRFs in protein sequences. Bioinformatics 31 (11):1738-1744.
17. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28 (12):i75-83.
18. Malhis N, Jacobson M, Gsponer J (2016) MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. Nucleic Acids Res.
19. Jones DT, Cozzetto D (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31 (6):857-863.
20. Fang C, Noguchi T, Tominaga D, Yamana H (2013) MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinformatics 14:300.
21. Xue B, Dunker AK, Uversky VN (2010) Retro-MoRFs: Identifying Protein Binding Sites by Normal and Reverse Alignment and Intrinsic Disorder Prediction. International Journal of Molecular Sciences 11 (10):3725-3747.
22. Puntervoll P, Linding R, Gemünd C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DMA, Ausiello G, Brannetti B, Costantini A, Ferrè F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Küster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31 (13):3625-3630.
23. Meszaros B, Dosztanyi Z, Simon I (2012) Disordered binding regions and linear motifs--bridging the gap between two models of molecular recognition. PLoS One 7 (10):e46829.
24. Peng Z, Wang C, Uversky VN, Kurgan L (2017) Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. Methods Mol Biol 1484:187-203.
25. Meszaros B, Simon I, Dosztanyi Z (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 5 (5):e1000376.
26. Dosztanyi Z, Meszaros B, Simon I (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics 25 (20):2745-2746.
27. Khan W, Duffy F, Pollastri G, Shields DC, Mooney C (2013) Predicting Binding within Disordered Protein Regions to Structurally Characterised Peptide-Binding Domains. PLoS ONE 8 (9):e72838.
28. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28 (1):235-242.
29. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21 (16):3433-3434.

30. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT (2004) The DISOPRED server for the prediction of protein disorder. Bioinformatics 20 (13):2138-2139.
31. McGuffin LJ (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics 24 (16):1798-1804.
32. Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Disfani FM, Kurgan L (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26 (18):i489-496.
33. Faraggi E, Xue B, Zhou Y (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins 74 (4):847-856.
34. Schlessinger A, Yachdav G, Rost B (2006) PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 22 (7):891-893.
35. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36 (Database issue):D202-205.
36. Tate RF (1954) Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation. The Annals of Mathematical Statistics 25 (3):603-607.
37. Zhao R, Gish K, Murphy M, Yin Y, Notterman D, Hoffman WH, Tom E, Mack DH, Levine AJ (2000) Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. Genes Dev 14 (8):981-993.
38. Balint EE, Vousden KH (2001) Activation and activities of the p53 tumour suppressor protein. Br J Cancer 85 (12):1813-1823.
39. el-Deiry WS (1998) Regulation of p53 downstream genes. Semin Cancer Biol 8 (5):345-357.
40. Yu J, Zhang L, Hwang PM, Rago C, Kinzler KW, Vogelstein B (1999) Identification and classification of p53-regulated genes. Proc Natl Acad Sci U S A 96 (25):14517-14522.
41. Sax JK, El-Deiry WS (2003) p53-induced gene expression analysis. Methods Mol Biol 234:65-71.
42. Fridman JS, Lowe SW (2003) Control of apoptosis by p53. Oncogene 22 (56):9030-9040.
43. Anderson CW, Appella E (2004) Signaling to the p53 tumor suppressor through pathways activated by genotoxic and nongenotoxic stress. In: Bradshaw RA, Dennis EA (eds) Handbook of Cell Signaling. Academic Press, New York, pp 237-247.
44. Gottlieb TM, Leal JF, Seger R, Taya Y, Oren M (2002) Cross-talk between Akt, p53 and Mdm2: possible implications for the regulation of apoptosis. Oncogene 21 (8):1299-1303.
45. Nicholson KM, Anderson NG (2002) The protein kinase B/Akt signalling pathway in human malignancy. Cell Signal 14 (5):381-395.
46. Abraham AG, O'Neill E (2014) PI3K/Akt-mediated regulation of p53 in cancer. Biochem Soc Trans 42 (4):798-803.
47. Muller PA, Vousden KH (2013) p53 mutations in cancer. Nat Cell Biol 15 (1):2-8.
48. Soussi T, Beroud C (2001) Assessing TP53 status in human tumours to evaluate clinical outcome. Nat Rev Cancer 1 (3):233-240.
49. Bookstein R (1994) Tumor suppressor genes in prostatic oncogenesis. J Cell Biochem Suppl 19:217-223.
50. Pencik J, Wiebringhaus R, Susani M, Culig Z, Kenner L (2015) IL-6/STAT3/ARF: the guardians of senescence, cancer progression and metastasis in prostate cancer. Swiss Med Wkly 145:w14215.
51. Wolff JM, Stephenson RN, Jakse G, Habib FK (1994) Retinoblastoma and p53 genes as prognostic indicators in urological oncology. Urol Int 53 (1):1-5.
52. Joerger AC, Ang HC, Veprintsev DB, Blair CM, Fersht AR (2005) Structures of p53 cancer mutants and mechanism of rescue by second-site suppressor mutations. J Biol Chem 280 (16):16030-16037.
53. Canadillas JM, Tidow H, Freund SM, Rutherford TJ, Ang HC, Fersht AR (2006) Solution structure of p53 core domain: structural basis for its instability. Proc Natl Acad Sci U S A 103 (7):2109-2114.

54. Wang Y, Rosengarth A, Luecke H (2007) Structure of the human p53 core domain in the absence of DNA. Acta Crystallogr D Biol Crystallogr 63 (Pt 3):276-281.
55. Joerger AC, Fersht AR (2008) Structural biology of the tumor suppressor p53. Annu Rev Biochem 77:557-582.
56. Uversky VN, Oldfield CJ, Midic U, Xie H, Xue B, Vucetic S, Iakoucheva LM, Obradovic Z, Dunker AK (2009) Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. BMC Genomics 10 Suppl 1:S7.
57. Anderson CW, Appella E (2004) Signaling to the p53 tumor suppressor through pathways activated by genotoxic and nongenotoxic stress. In: Bradshaw RA, Dennis EA (eds) Handbook of Cell Signaling. Academic Press, New York, pp 237-247.
58. Bianco R, Ciardiello F, Tortora G (2005) Chemosensitization by antisense oligonucleotides targeting MDM2. Curr Cancer Drug Targets 5 (1):51-56.
59. Moll UM, Petrenko O (2003) The MDM2-p53 interaction. Mol Cancer Res 1 (14):1001-1008.
60. Nag S, Qin J, Srivenugopal KS, Wang M, Zhang R (2013) The MDM2-p53 pathway revisited. J Biomed Res 27 (4):254-271.
61. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, Levine AJ, Pavletich NP (1996) Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Science 274 (5289):948-953.
62. Bochkareva E, Kaustov L, Ayed A, Yi GS, Lu Y, Pineda-Lucena A, Liao JC, Okorokov AL, Milner J, Arrowsmith CH, Bochkarev A (2005) Single-stranded DNA mimicry in the p53 transactivation domain interaction with replication protein A. Proc Natl Acad Sci U S A 102 (43):15412-15417.
63. Mora P, Carbajo RJ, Pineda-Lucena A, Sanchez del Pino MM, Perez-Paya E (2008) Solvent-exposed residues located in the beta-sheet modulate the stability of the tetramerization domain of p53--a structural and combinatorial approach. Proteins 71 (4):1670-1685.
64. Lowe ED, Tews I, Cheng KY, Brown NR, Gul S, Noble ME, Gamblin SJ, Johnson LN (2002) Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin A. Biochemistry 41 (52):15625-15634.
65. Avalos JL, Celic I, Muhammad S, Cosgrove MS, Boeke JD, Wolberger C (2002) Structure of a Sir2 enzyme bound to an acetylated p53 peptide. Mol Cell 10 (3):523-535.
66. Mujtaba S, He Y, Zeng L, Yan S, Plotnikova O, Sachchidanand, Sanchez R, Zeleznik-Le NJ, Ronai Z, Zhou MM (2004) Structural mechanism of the bromodomain of the coactivator CBP in p53 transcriptional activation. Mol Cell 13 (2):251-263.
67. Rustandi RR, Baldisseri DM, Weber DJ (2000) Structure of the negative regulatory domain of p53 bound to S100B(betabeta). Nat Struct Biol 7 (7):570-574.
68. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics 9 Suppl 1:S1.
69. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 7:208.
70. Ehretsmann CP, Carpousis AJ, Krisch HM (1992) Specificity of Escherichia coli endoribonuclease RNase E: in vivo and in vitro analysis of mutants in a bacteriophage T4 mRNA processing site. Genes Dev 6 (1):149-159.
71. Huang H, Liao J, Cohen SN (1998) Poly(A)- and poly(U)-specific RNA 3' tail shortening by E. coli ribonuclease E. Nature 391 (6662):99-102.
72. Kushner SR (2002) mRNA decay in Escherichia coli comes of age. J Bacteriol 184 (17):4658-4665; discussion 4657.
73. Ow MC, Kushner SR (2002) Initiation of tRNA maturation by RNase E is essential for cell viability in E. coli. Genes Dev 16 (9):1102-1115.
74. Steege DA (2000) Emerging features of mRNA decay in bacteria. RNA 6 (8):1079-1090.

75. Casaregola S, Jacq A, Laoudj D, McGurk G, Margarson S, Tempete M, Norris V, Holland IB (1992) Cloning and analysis of the entire Escherichia coli ams gene. ams is identical to hmp1 and encodes a 114 kDa protein that migrates as a 180 kDa protein. Journal of molecular biology 228 (1):30-40.
76. Claverie-Martin F, Diaz-Torres MR, Yancey SD, Kushner SR (1991) Analysis of the altered mRNA stability (ams) gene from Escherichia coli. Nucleotide sequence, transcriptional analysis, and homology of its product to MRP3, a mitochondrial ribosomal protein from Neurospora crassa. J Biol Chem 266 (5):2843-2851.
77. Lopez PJ, Marchand I, Joyce SA, Dreyfus M (1999) The C-terminal half of RNase E, which organizes the Escherichia coli degradosome, participates in mRNA degradation but not rRNA processing in vivo. Mol Microbiol 33 (1):188-199.
78. Cohen SN, McDowall KJ (1997) RNase E: still a wonderfully mysterious enzyme. Mol Microbiol 23 (6):1099-1106.
79. McDowall KJ, Cohen SN (1996) The N-terminal domain of the rne gene product has RNase E activity and is non-overlapping with the arginine-rich RNA-binding site. Journal of molecular biology 255 (3):349-355.
80. Wachi M, Umitsuki G, Shimizu M, Takada A, Nagai K (1999) Escherichia coli cafA gene encodes a novel RNase, designated as RNase G, involved in processing of the 5' end of 16S rRNA. Biochem Biophys Res Commun 259 (2):483-488.
81. Kaberdin VR, Miczak A, Jakobsen JS, Lin-Chao S, McDowall KJ, von Gabain A (1998) The endoribonucleolytic N-terminal half of Escherichia coli RNase E is evolutionarily conserved in Synechocystis sp. and other bacteria but not the C-terminal half, which is sufficient for degradosome assembly. Proc Natl Acad Sci U S A 95 (20):11637-11642.
82. Callaghan AJ, Aurikko JP, Ilag LL, Gunter Grossmann J, Chandran V, Kuhnel K, Poljak L, Carpousis AJ, Robinson CV, Symmons MF, Luisi BF (2004) Studies of the RNA degradosome-organizing domain of the Escherichia coli ribonuclease RNase E. Journal of molecular biology 340 (5):965-979.
83. Taraseviciene L, Bjork GR, Uhlin BE (1995) Evidence for an RNA binding region in the Escherichia coli processing endoribonuclease RNase E. J Biol Chem 270 (44):26391-26398.