Published in final edited form as:

Mol Biosyst. 2013 October ; 9(10): . doi:10.1039/c3mb70167k.

Prediction of RNA binding proteins comes of age from low resolution to high resolution†

Huiying Zhao^{a,b}, Yuedong Yang^{a,b}, and Yaoqi Zhou^{a,b,‡}

Yaoqi Zhou: yqzhou@iupui.edu

^aSchool of Informatics, Indiana University Purdue University, Indianapolis, Indiana 46202, USA

^bCenter for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA

Abstract

Networks of protein–RNA interactions is likely to be larger than protein–protein and protein–DNA interaction networks because RNA transcripts are encoded tens of times more than proteins (*e.g.* only 3% of human genome coded for proteins), have diverse function and localization, and are controlled by proteins from birth (transcription) to death (degradation). This massive network is evidenced by several recent experimental discoveries of large numbers of previously unknown RNA-binding proteins (RBPs). Meanwhile, more than 400 non-redundant protein–RNA complex structures (at 25% sequence identity or less) have been deposited into the protein databank. These sequences and structural resources for RBPs provide ample data for the development of computational techniques dedicated to RBP prediction, as experimentally determining RNA-binding functions is time-consuming and expensive. This review compares traditional machine-learning based approaches with emerging template-based methods at several levels of prediction resolution ranging from two-state binding/non-binding prediction, to binding residue prediction and protein–RNA complex structure prediction. The analysis indicates that the two approaches are complementary and their combinations may lead to further improvements.

Introduction

RNA is directly involved in a wide variety of functions ranging from protein synthesis, post-transcriptional modification, to post-transcriptional regulation. Unlike DNA, located mostly in the cell nucleus, RNA is transcribed in the nucleus and transported to the cytoplasm as non-coding RNA or for translation. The diverse localization and different functionalities of RNA transcripts^{1–3} along with only 3% of the human genome being coded for proteins⁴ suggest that the network of protein–RNA interactions is likely to be larger and more complex than those of protein–DNA and protein–protein interactions.⁵ These RNA-binding proteins (RBPs) are challenging to locate experimentally although some progress in high-throughput biochemical approaches has been made^{5–8} and hundreds of novel unconventional or moonlighting RBPs have been discovered.^{9–11} This, however, has only scratched the surface of RBPs and their associated post-transcriptional networks.

†Electronic supplementary information (ESI) available. See DOI: 10.1039/c3mb70167k

© The Royal Society of Chemistry 2013

Correspondence to: Yaoqi Zhou, yqzhou@iupui.edu.

‡Current address: Institute for Glycomics and School of Information and Communication Technology, Griffith University, Parklands Dr. QLD 4215, Australia. yaoqi.zhou@griffith.edu.au

A complete understanding of the protein–RNA interaction between a specific protein and an RNA requires the determination their complex structure. In spite of difficulties in solving protein–RNA complex structures,^{12–14} the number of non-redundant complex structures deposited in PDB has quadrupled from 45 per year in 2001 to 181 in 2012 (at 90% sequence identity cutoff), as shown in Fig. 1. By comparison, the number of deposited structures has only tripled from 2831 in 2001 to 8944 in 2012 (<http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=90>). The growing number of protein–RNA complex structures provides an increasingly larger dataset for analyzing the principles of protein–RNA recognition.^{15–20} However, not all members in the same structural folds have RNA-binding activities. For example, the Structural Classification Of Proteins (SCOP)²¹ has 44 folds shared by both RNA and non-RNA binding proteins.²²

The challenges and expense to experimentally determine RBPs necessitate the development of accurate and efficient computational techniques. In this review article, we will classify different computational methods according to the resolution of their prediction from low, medium, high to the highest. A low-resolution prediction is a simple two-state prediction of whether a protein is RNA binding or non-RNA binding. A medium-resolution prediction locates the amino-acid region of an RBP that binds to RNA (RNA binding site/motif prediction). A high-resolution prediction indicates the types of RNA binding to an RBP. The highest resolution prediction will predict the three-dimensional structure of protein–RNA complexes with a predicted RNA binding sequence. The highest resolution prediction can simultaneously make all lower resolution predictions including the RNA type, RNA-binding site, and the two-state RBP/non-RBP classification, but not vice versa. Most computational methods developed so far have focused on low to medium resolution prediction.^{23,24}

Another way to classify computational methods is based on the information employed in the prediction. A structure-based method attempts to predict RNA-binding based on a known protein structure. Structure-based methods are important because of the small number of known protein–RNA complex structures relative to all structures solved. As of June 18th, 2013, there are 91 550 structures in the PDB but of these only 1541 entries (1.7%) are RNA–protein complexes. Some of those other 98.3% of structures could have unknown hidden RNA-binding activities as a large number of RBPs are yet to be discovered. Moreover, many structures generated from structure genomics projects have unknown biological functions.²⁵ A sequence-based method, on the other hand, makes a prediction with the protein sequence as the only input. Sequence-based methods can be coupled with predicted or actual structural information as well as predicted binding information to improve the overall accuracy of prediction.

Here, we will provide a brief review based on the resolution as well as the information (*i.e.* sequence *versus* structure-based) employed in prediction.

Low resolution function prediction: two-state prediction (RBP prediction)

Structure-based inference of RBPs

Negatively charged RNA preferentially binds to positively charged proteins. Electrostatic interactions obviously are an important feature for the detection of RBPs. Shazman and Mandel-Gutfreund²⁶ employed Support Vector Machines (SVM) to combine electrostatic patches, solvent accessibility, cleft sizes and other global protein features for RBP prediction. This method trained on 76 RNA binding proteins and 246 non-nucleic acid binding proteins and achieved a Matthews correlation coefficient (MCC) of 0.72 based on the leave-one-out test. MCC is a correlation coefficient between actual and predicted binary classifications. It is 1 for perfect prediction and 0 for random prediction. Despite its high MCC value, the method is unable to distinguish RBPs from DNA-binding proteins. Ahmad

and Sarai²⁷ employed neural networks that are based on charge, dipole moment and three eigenvalues of quadrupole moments generated from the structure. It was trained on 160 RBPs and 2441 non-RBPs and achieved 0.79 for an area under the ROC curve based on the leave-one-out test (1 for perfect prediction and 0.5 for random prediction). Table 1 provides a list of features for the two methods described above. More recently, a template-based approach was developed by utilizing known protein–RNA complex structures.^{28,29} In this method, a target structure is aligned to the templates in the template library and an RBP is predicted if the structural similarity between the target and a template is higher than a certain threshold. Several structural alignment programs were tested. Among them, SPalign²⁸ was found to give the highest MCC value of 0.37 based on a dataset of 212 RNA binding domains and 6761 non-RNA binding domains with 250 RNA-binding domains as templates. When the query structure is compared to template structures, the templates with sequence identity >30% to the query sequence are excluded in order to test the ability of the method to detect remote structural homologs. SPOT-struc (RNA)²⁹ is an improved method based on a structural similarity score only by using a relative structural similarity between RBPs and non-RBPs and by predicting the binding affinity between the query protein and the template RNA with a statistical energy function based on distance-scaled finite ideal gas reference state (DFIRE).³⁰ It achieves an MCC value of 0.57 for the same dataset above. SPOT-struc (RNA) has the ability to separate RNA- from DNA-binding proteins because it yields zero false positives after excluding proteins known to bind both DNA and RNA when applied to a dataset of 331 DNA binding domains.

Sequence-based inference of RBPs

The main limitation of a structure-based technique is that the structures for most proteins are not yet known. One common technique is homology-based prediction assuming that proteins with similar sequences are likely to perform the same function. Enzymes,^{31–33} for example, tend to have a conserved function, if they share more than 40% to 50% sequence identity. However, such prediction will produce false negatives by failure to detect functionally identical remote homologs³⁴ and false positives by ignoring possible functional divergence for highly homologous sequences.³¹ Thus, there is a need to go beyond simple homology-based search. Several SVM-based tools^{35–41} were developed. Different methods mainly differ in features employed as summarized in Table 1. Commonly used features are the composition of amino-acid residues, hydrophobicity, amino acid composition, charge, hydrophobicity and accessible surface area. Early studies^{35,36} did not remove homologous sequences in training and testing. Due to limitation of SVMs, most methods were trained with nearly equal number of RBPs and non-RBPs.^{37–41} In a real-world situation, RBPs are only a fraction of all proteins. The reported MCC values are 0.53 for a dataset of 134 RBPs and 134 non-RBPs,³⁹ 0.51 for a dataset of 69 RBPs and 100 non-RBPs by RNAPred,⁴⁰ 0.65 for a dataset of 687 RBPs and 687 non-RBPs.⁴¹ The structure-based method SPOT-struc²⁹ was extended to sequence-based (SPOT-seq²²) by employing predicted structures from a template-based technique called SPARKS-X.⁴² More specifically, SPARKS-X attempts to match the query sequence to the templates of known protein–RNA complex structures. If a match is found (based on a Z-score), a binding affinity is predicted based on a knowledge-based energy function. The query sequence is an RBP if the binding affinity is higher than an optimized threshold. This coupled structure and binding prediction leads to an MCC value of 0.62 for independent test on 215 RBPs and 5765 non-RBPs with a template library of 1164 RNA-binding domains and RNA-binding chains. This MCC value is even higher than 0.56 given by SPOT-struc for the same dataset despite of using predicted structures in SPOT-seq, rather than actual structures in SPOT-struc, suggesting possible cancellation of errors in structure and binding prediction.

Method comparison

There is a lack of comparison between different methods for RBP prediction. Most methods described above do not have web-servers or their web-servers are no longer functional. We found only two available servers (RNApred,⁴⁰ <http://www.imtech.res.in/raghava/rnapred/> and SVMprot³⁶). Both of these are sequence-based methods. They are compared to SPOT-seq along with structure-based techniques SPOT-struc and SPalign with a dataset of 257 RBPs and 5765 non-RBPs in Table 2. This dataset is an independent test set for SPOT-seq. Here, we define sensitivity as $TP/(TP + FN)$, precision as $TP/(TP + FP)$, and $MCC = (TP \times TN - FP \times FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$ where TP, FP, TN, and FN are numbers of true and false positives and true and false negatives, respectively. RNApred predicted 203 out of 257 RBPs and 2415 out of 5765 non-RBPs as RBPs. As shown in Table 2, RNApred achieved an MCC value of 0.15, sensitivity of 79%, and precision of 8%. SVMprot yields the MCC of 0.19, sensitivity of 50% and precision of 13%. By comparison, SPOT-seq has a MCC value of 0.60, sensitivity of 44%, precision of 84% for the same dataset. Thus SPOT-seq is significantly more powerful in separating RNA from non-RNA binding proteins. It is even more powerful than structure-based techniques that achieved MCC values of 0.46 (SPalign) and 0.50 (SPOT-struc). Fig. 2 displays the Receiver Operating Characteristic (ROC) curves for these sequence and structure-based methods. It is clear that SPOT-seq, the template-based technique, is substantially more accurate than other sequence-based, machine-learning techniques (RNApred and SVMprot) or structure-based techniques (SPalign and SPOT-struc). For structure-based techniques, SPalign, although is less accurate than SPOT-struc at low false positive rates, has higher sensitivity at high false positive rates. This suggests that replacing TMalign employed in SPOT-struc by SPalign for pairwise structure alignment will likely further increase the power of SPOT-struc.

Discriminating RBPs from DNA-binding proteins

DNA-binding proteins are important controls for examining the accuracy of RBP prediction because DNA-binding interfaces are also positively charged as RNA-binding interfaces. Most methods are either unable to separate RNA from DNA binding proteins or not tested in this aspect. Table 2 confirms the high false positive rates given by RNApred (~69%) and SVM-prot (22%) when tested on 245 DNA-binding proteins, compared to zero-false positive rates given by SPOT-seq. These 245 DNA-binding proteins are a subset of DB250 which are modified by excluding 5 RNA and DNA-binding proteins.²²

Medium resolution function prediction: binding residues prediction

Locating functional residues is an important first step in understanding the mechanism of function. Thus, there are a significant number of studies predicting RNA-binding residues. Most studies are machine-learning techniques trained from sequences or structures of known RBPs.

Structure-based prediction

Finding discriminative structural features is key for accurate prediction of RNA-binding residues from a given structure. Table 3 lists structural features employed by several structure-based techniques.^{43–50} The methods range from docking, random forest classifier, neural network, SVM, naïve Bayes classifier to linear regression. The notable features are sequence conservation, secondary structures, types of amino acid residues, solvent accessible surface area and interface propensity. There are some overlaps between the features employed for RBP prediction and binding residue prediction, except that one focuses on the whole protein level and the other is on the residue level. Unlike machine-

learning based methods, template-based methods can predict RBPs and their binding residues at the same time because template-based methods predict complex structures between target structure and template RNA that can be employed for annotating binding residues. The accuracy of binding residue prediction depends on strongly on structural alignment programs. SPOT-struc (RNA)²⁹ is based on an alignment program called TM-align.⁵¹ Another method SPalign that permits a size-independent alignment was developed to further improve the accuracy of alignment and identification of binding regions.²⁸

Sequence-based prediction

For sequence-based predictions, the prominent feature is sequence similarity and evolution information.^{39,52–59} Additional features as shown in Table 3 include properties of amino acid residues, predicted secondary structures and solvent accessibility. Most methods are based on SVM. These features are typical of those utilized in secondary structure prediction and ASA prediction as well (*e.g.* ref. 60). All above methods are machine-learning based tools. Similar to SPOT-struc, SPOT-seq²² can infer RNA-binding residues according to predicted RNA–protein complexes between the model structure of the target protein and the structure of template RNA.

Method comparison

The published performance for various methods in structure and sequence-based RNA-binding residues can be found in Table S1 (ESI[†]). One conclusion is that structure-based techniques do not have any advantage over sequence-based techniques. The second conclusion is that all methods have MCC below 0.6. However, different datasets make comparisons between different methods impossible. To compare different methods, we built a dataset of 106 RNA-binding domains (RB106) that were released in 2011 and 2012. RB106 is a non-redundant dataset with pairwise sequence identity lower than 35%. However, only 67 domains in 106 domains were predicted as RBPs by SPOT-seq because of a lack of templates or low binding affinity. Thus, we also show results for the RB67 set. In addition, we further removed the domains that have more than 45% sequence identity with RNA-binding domains released before 2011. This leads to a small dataset of 20 RBPs (RB20). We employed 45% sequence identity cutoff here because a lower cutoff will lead to fewer new RNA-binding complex structures.

Table 4 lists the performance of various structure and sequence-based techniques for three datasets (RB106, RB67 and RB20). In structure-based techniques, SPalign is consistently the top performing of three structure-based techniques (SPalign, SPOT-struc and KYG). In both SPalign and SPOT-struc, all templates more than 30% sequence identity to the target are removed. In sequence-based methods, BINDN+ has the best performance in the MCC value for RB106 (MCC = 0.59), followed by PBRpred (MCC = 0.57). For RB20, PBRpred gives the highest MCC value (0.39), followed by BINDN+ (0.38) and RNABindR (0.37). SPOT-seq, on the other hand, yields the highest MCC value for those proteins predicted as RBPs (0.63 for RB67). SPOT-seq achieved an MCC value of 0.33 for RB20 by using the templates that have no sequence identity higher than 30% to target. It seems that machine-learning-based techniques are as accurate as or more accurate than template-based techniques in predicting RNA binding residues. All methods, however, decrease their accuracy significantly for RB20, suggesting that homologous sequences are the reasons for improved prediction for RB106. The performance of various methods is also compared by the ROC curves in Fig. 3. Regardless of datasets, two best performing methods are RBPpred and BindN+.

High-resolution function prediction: binding RNA type prediction

Predicting the type of RNA binding with a given RBP provides more detailed information on the function of RNA-binding proteins. Yue *et al.*³⁷ developed a sequence-based predictor for separating rRNA-binding from RNA-binding proteins. They found that rRNA-binding proteins can be more accurately predicted than other types of RNA-binding proteins. Shazman and Mandel-Gutfreund²⁶ employed a multi-class SVM to classify rRNA, tRNA, and mRNA-binding proteins based on electrostatic properties derived from protein structures. It has the highest success rate for tRNA-binding proteins (13/13) but a lower success rate for rRNA (32/46) and mRNA (17/23) binding proteins. This method, however, cannot separate RNA from DNA binding proteins. SPOT-seq can predict the RNA types by assuming that the query protein and its matching template RBP bind to the same type of RNA.²² SPOT-seq achieved success rate of 69% (33/48) for tRNA, 56% (15/27) for rRNA and 96% (54/56) for mRNA for an independent test set of 215 RNA-binding proteins, compared to 62%, 73% and 91% for the training set of 216 RBPs. It should be noted, however, that the RNA structural motif, rather than the RNA functional type, is the key to the RBP function as many proteins can bind with different types of RNAs.

Highest resolution function prediction: protein–RNA complex structure prediction

To understand the mechanism of protein–RNA binding, atomic resolution of protein–RNA complex structures is required. One method to predict protein–RNA complex structures is protein–RNA docking which relies on known protein and RNA structures. Such docking techniques for protein–RNA interactions can be modified from many docking software tools for protein–protein and protein–ligand docking after equipping with a scoring/energy function for protein–RNA interaction. For example, Zheng *et al.* utilized the RosettaDocking⁶¹ program to generate protein–RNA complex decoys and evaluate the ability of a knowledge-based energy function based on a conditional-probability function to discriminate docking decoys.⁶² Perez-Cano *et al.* employed the FTDOCK⁶³ program plus propensity-based statistical potentials.⁴⁹ Tuszynska and Bujnicki employed the GRAMM⁶⁴ docking program and two separate statistical potentials (QUASI-RNP based a quasi-chemical reference state and DARS-RNA based on the reference state from decoys) for scoring.⁶⁵ Setny and Zacharias employed the protein-docking program ATTRACT⁶⁶ and a knowledge-based energy function employing a quasi-chemical approximation.⁶⁷ These studies demonstrated the usefulness of knowledge-based energy functions for decoy discrimination and selection of near-native docking decoys. A DFIRE-based statistical potential for protein–RNA interaction²⁹ is found to be useful for increasing true positive rates while decreasing false positive rates of predicting RNA-binding proteins. Protein–RNA docking, however, is more challenging than protein–protein docking because RNA structures are more flexible than protein structures. This is demonstrated by critical assessment of predicted interaction (CAPRI, 2009). CAPRI, which typically assessed protein–protein docking models, included a protein–RNA complex structure in a recent round.⁶⁸ All docking predictions failed for this protein–RNA complex target because of inaccurate model RNA structures.

Another approach to predict protein–RNA complex structures is to use known protein–RNA complex structures as templates. SPOT-seq²² and SPOT-struc²⁹ are sequence and structure-based techniques for predicting protein–RNA binding complex structures based on template-based structure prediction program SPARKS X and structural alignment program TM-align,⁵¹ respectively. Both methods can provide quite accurate prediction of binding residues and complex structures if a significantly matching template is found. For example, the average structural similarity between predicted and actual structures is about 0.66

according to the TM-score, regardless of training or testing. About 77% of structures predicted are less than 4 Å root-mean-squared distance from the respective native structures.²² One example is shown in Fig. 4. In this figure the target protein is 1m8yB (human Puf protein, Pumilio1), the SPOT-seq selected template is 3k5qA (Caenorhabditis elegans fem-3 binding factor 2). The sequence identity between these two proteins is 24.9%. The advantage of SPOT-seq or SPOT-struc is their computational efficiency that allows large genome-scale prediction.

Summary and outlook

A constant increase in the number of protein–RNA complex structures makes it possible to develop various techniques to predict RNA-binding proteins with different levels of functional details. Unlike sequence-based approaches using templates, sequence-based techniques using machine-learning methods are ineffective in distinguishing RNA-binding proteins from DNA-binding proteins. On the other hand, for a known RNA-binding protein, the best machine-learning techniques are often more accurate in locating RNA-binding residues than a template-based approach. This is true particularly for those proteins that are not predicted as RBPs by the template-based approach. Thus, combining these two approaches may further improve the accuracy of RNA-binding function prediction.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Notes and references

- Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schonbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, Brusica V, Chothia C, Corbani LE, Cousins S, Dalla E, Dragani TA, Fletcher CF, Forrest A, Frazer KS, Gaasterland T, Gariboldi M, Gissi C, Godzik A, Gough J, Grimmond S, Gustincich S, Hirokawa N, Jackson JJ, Jarvis ED, Kanai A, Kawaji H, Kawasaki Y, Kedzierski RM, King BL, Konagaya A, Kurochkin IV, Lee Y, Lenhard B, Lyons PA, Maglott DR, Maltais L, Marchionni L, McKenzie L, Miki H, Nagashima T, Numata K, Okido T, Pavan WJ, Pertea G, Pesole G, Petrovsky N, Pillai R, Pontius JU, Qi D, Ramachandran S, Ravasi T, Reed JC, Reed DJ, Reid J, Ring BZ, Ringwald M, Sandelin A, Schneider C, Semple CAM, Setou M, Shimada K, Sultana R, Takenaka Y, Taylor MS, Teasdale RD, Tomita M, Verardo R, Wagner L, Wahlestedt C, Wang Y, Watanabe Y, Wells C, Wilming LG, Wynshaw-Boris A, Yanagisawa M, Yang I, Yang L, Yuan Z, Zavolan M, Zhu Y, Zimmer A, Carninci P, Hayatsu N, Hirozane-Kishikawa T, Konno H, Nakamura M, Sakazume N, Sato K, Shiraki T, Waki K, Kawai J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Imotani K, Ishii Y, Itoh M, Kagawa I, Miyazaki A, Sakai K, Sasaki D, Shibata K, Shinagawa A, Yasunishi A, Yoshino M, Waterston R, Lander ES, Rogers J, Birney E, Hayashizaki Y, Consortium F, Grp RGER. *Nature*. 2002; 420:563–573. [PubMed: 12466851]
- Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, Adachi J, Fukuda S, Aizawa K, Izawa M, Nishi K, Kiyosawa H, Kondo S, Yamanaka I, Saito T, Okazaki Y, Gojobori T, Bono H, Kasukawa T, Saito R, Kadota K, Matsuda H, Ashburner M, Batalov S, Casavant T, Fleischmann W, Gaasterland T, Gissi C, King B, Kochiwa H, Kuehl P, Lewis S, Matsuo Y, Nikaido I, Pesole G, Quackenbush J, Schriml ML, Staubli F, Suzuki R, Tomita M, Wagner L, Washio T, Sakai K, Okido T, Furuno M, Aono H, Baldarelli R, Barsh G, Blake J, Boffelli D, Bojunga N, Carninci P, de Bonaldo MF, Brownstein MJ, Bult C, Fletcher C, Fujita M, Gariboldi M, Gustincich S, Hill D, Hofmann M, Hume DA, Kamiya M, Lee NH, Lyons P, Marchionni L, Mashima J, Mazzairelli J, Mombaerts P, Nordone P, Ring B, Ringwald M, Rodriguez I, Sakamoto N, Sasaki H, Sato K, Schonbach C, Seya T, Shibata Y, Storch KF, Suzuki H, Toyooka K, Wang KH, Weitz C, Whittaker C, Wilming L, Wynshaw-Boris A, Yoshida K, Hasegawa Y,

Kawaji H, Kohtsuki S, Hayashizaki Y, Grp RGER. *Nature*. 2001; 409:685–690. [PubMed: 11217851]

3. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng ZP, Snyder M, Dermitzakis ET, Stamatoyannopoulos JA, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SCJ, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Dutta A, Guigo R, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng DY, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Flicek P, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu JQ, Lian Z, Lian J, Newburger P, Zhang XQ, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Dermitzakis ET, Margulies EH, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan YJ, Snyder M, Birney E, Struhl K, Gerstein M, Antonarakis SE, Gingeras TR, Brown JB, Flicek P, Fu YT, Keefe D, Birney E, Denoeud F, Gerstein M, Green ED, Kapranov P, Karaoz U, Myers RM, Noble WS, Reymond A, Rozowsky J, Struhl K, Siepel A, Stamatoyannopoulos JA, Taylor CM, Taylor J, Thurman RE, Tullius TD, Washietl S, Zheng DY, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Collins FS, Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Hou MM, Taylor J, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Brown JB, Huang HY, Zhang NR, Bickel P, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Sringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Gerstein M, Antonarakis SE, Batzoglu S, Goldman N, Hardison RC, Haussler D, Miller W, Pachter L, Green ED, Sidow A, Weng ZP, Trinklein ND, Fu YT, Zhang ZDD, Karaoz U, Barrera L, Stuart R, Zheng DY, Ghosh S, Flicek P, King DC, Taylor J, Ameer A, Enroth S, Bieda MC, Koch CM, Hirsch HA, Wei CL, Cheng J, Kim J, Bhinge AA, Giresi PG, Jiang N, Liu J, Yao F, Sung WK, Chiu KP, Vega VB, Lee CWH, Ng P, Shahab A, Sekinger EA, Yang A, Moqtaderi Z, Zhu Z, Xu XQ, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Clelland GK, Wilcox S, Dillon SC, Andrews RM, Fowler JC, Couttet P, James KD, Lefebvre GC, Bruce AW, Dovey OM, Ellis PD, Dhami P, Langford CF, Carter NP, Vetrie D, Kapranov P, Nix DA, Bell I, Patel S, Rozowsky J, Euskirchen G, Hartman S, Lian J, Wu JQ, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu CX, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang XL, Xu MS, Haidar JNS, Yu Y, Birney E, Weissman S, Ruan YJ, Lieb JD, Iyer VR, Green RD, Gingeras TR, Wadelius C, Dunham I, Struhl K, Hardison RC, Gerstein M, Farnham PJ, Myers RM, Ren B, Snyder M, Thomas DJ, Rosenbloom K, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakapallayil A, Barber G, Kuhn RM, Karolchik D, Haussler D, Kent WJ, Dermitzakis ET, Armengol L, Bird CP, Clark TG, Cooper GM, de Bakker PIW, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Thomas DJ, Woodroffe A, Batzoglu S, Davydov E, Dimas A, Eyraes E, Hallgrimsdottir IB, Hardison RC, Huppert J, Sidow A, Taylor J, Trumbower H, Zody MC, Guigo R, Mullikin JC, Abecasis GR, Estivill X, Birney E, Bouffard GG, Guan XB, Hansen NF, Idol JR, Maduro VVB, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang HY, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu BL, de Jong PJ, Consortium EP. *Nature*. 2007; 447:799–816. [PubMed: 17571346]
4. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. *Nat Rev Genet*. 2010; 11:559–571. [PubMed: 20628352]
5. Iio H, Loiselle D, Haystead TA, Macara IG. *Nucleic Acids Res*. 2011; 39:E53. [PubMed: 21300640]

6. Zhang CL, Darnell RB. *Nat Biotechnol.* 2011; 29:U607–U686.
7. Galante PAF, Sandhu D, Abreu RD, Gradassi M, Slager N, Vogel C, de Souza SJ, Penalva LOF. *RNA Biol.* 2009; 6:426–433. [PubMed: 19458496]
8. Konig J, Zarnack K, Luscombe NM, Ule J. *Nat Rev Genet.* 2011; 13:77–83. [PubMed: 22251872]
9. Tsvetanova NG, Klass DM, Salzman J, Brown PO. *PLoS One.* 2010; 5:e12671. [PubMed: 20844764]
10. Scherrer T, Mittal N, Janga SC, Gerber AP. *PLoS One.* 2010; 5:e15499. [PubMed: 21124907]
11. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsvelde J, Hentze MW. *Cell.* 2012; 149:1393–1406. [PubMed: 22658674]
12. Price SR, Ito N, Oubridge C, Avis JM, Nagai K. *J Mol Biol.* 1995; 249:398–408. [PubMed: 7540213]
13. Ke A, Doudna JA. *Methods.* 2004; 34:408–414. [PubMed: 15325657]
14. Perederina A, Krasilnikov AS. *Methods Mol Biol.* 2012; 905:123–143. [PubMed: 22736002]
15. Chen Y, Varani G. *FEBS J.* 2005; 272:2088–2097. [PubMed: 15853794]
16. Morozova N, Allers J, Myers J, Shamoo Y. *Bioinformatics.* 2006; 22:2746–2752. [PubMed: 16966360]
17. Ellis JJ, Broom M, Jones S. *Proteins.* 2007; 66:903–911. [PubMed: 17186525]
18. Treger M, Westhof E. *J Mol Recognit.* 2001; 14:199–214. [PubMed: 11500966]
19. Clery A, Blatter M, Allain FH. *Curr Opin Struct Biol.* 2008; 18:290–298. [PubMed: 18515081]
20. Shulman-Peleg A, Nussinov R, Wolfson HJ. *Nucleic Acids Res.* 2009; 37:D369–D373. [PubMed: 18953028]
21. Murzin AG, Brenner SE, Hubbard T, Chothia C. *J Mol Biol.* 1995; 247:536–540. [PubMed: 7723011]
22. Zhao H, Yang Y, Zhou Y. *RNA Biol.* 2011; 8:988–996. [PubMed: 21955494]
23. Puton T, Kozlowski L, Tuszyńska I, Rother K, Bujnicki JM. *J Struct Biol.* 2012; 179:261–268. [PubMed: 22019768]
24. Walia RR, Caragea C, Lewis BA, Towfic FG, Terribilini M, El-Manzalawy Y, Dobbs D, Honavar V. *BMC Bioinf.* 2012; 13:89.
25. Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA, Godzik A. *PLoS Biol.* 2009; 7:e1000205. [PubMed: 19787035]
26. Shazman S, Mandel-Gutfreund Y. *PLOS Comput Bio.* 2008; 4:E1000146. [PubMed: 18716674]
27. Ahmad S, Sarai A. *BMC Struct Biol.* 2011; 11:8. [PubMed: 21284850]
28. Yang Y, Zhan J, Zhao H, Zhou Y. *Proteins.* 2012; 80:2080–2088. [PubMed: 22522696]
29. Zhao H, Yang Y, Zhou Y. *Nucleic Acids Res.* 2011; 39:3017–3025. [PubMed: 21183467]
30. Zhou HY, Zhou YQ. *Protein Sci.* 2002; 11:2714–2726. [PubMed: 12381853]
31. Todd AE, Orenge CA, Thornton JM. *J Mol Biol.* 2001; 307:1113–1143. [PubMed: 11286560]
32. Tian WD, Skolnick J. *J Mol Biol.* 2003; 333:863–882. [PubMed: 14568541]
33. Rost B. *J Mol Biol.* 2002; 318:595–608. [PubMed: 12051862]
34. Weinhold N, Sander O, Domingues FS, Lengauer T, Sommer I. *PLoS Comput Biol.* 2008; 4:e1000105. [PubMed: 18604264]
35. Cai YD, Lin SL. *Biochim Biophys Acta, Proteins Proteomics.* 2003; 1648:127–133.
36. Han LY, Cai CZ, Lo SL, Chung MCM, Chen YZ. *RNA.* 2004; 10:355–368. [PubMed: 14970381]
37. Yu XJ, Cao JP, Cai YD, Shi TL, Li YX. *J Theor Biol.* 2006; 240:175–184. [PubMed: 16274699]
38. Shao XJ, Tian YJ, Wu LY, Wang Y, Jing L, Deng NY. *J Theor Biol.* 2009; 258:289–293. [PubMed: 19490865]
39. Spriggs RV, Murakami Y, Nakamura H, Jones S. *Bioinformatics.* 2009; 25:1492–1497. [PubMed: 19389733]
40. Kumar M, Gromiha MM, Raghava GPS. *J Mol Recognit.* 2011; 24:303–313. [PubMed: 20677174]
41. Peng CR, Liu L, Niu B, Lv YL, Li MJ, Yuan YL, Zhu YB, Lu WC, Cai YD. *J Biomed Biotechnol.* 2011; 2011:506205. [PubMed: 21826121]

42. Yang Y, Faraggi E, Zhao H, Zhou Y. *Bioinformatics*. 2011; 27:2076–2082. [PubMed: 21666270]
43. Jeong E, Chung IF, Miyano S. *Genome informatics. International Conference on Genome Informatics*. 2004; 15:105–116.
44. Kim OT, Yura K, Go N. *Nucleic Acids Res*. 2006; 34:6450–6460. [PubMed: 17130160]
45. Chen YC, Lim C. *Nucleic Acids Res*. 2008; 36:e29. [PubMed: 18276647]
46. Maetschke SR, Yuan Z. *BMC Bioinf*. 2009; 10:341.
47. Towfic F, Caragea C, Gemperline DC, Dobbs D, Honavar V. *Int J Data Min Bioinf*. 2010; 4:21–43.
48. Li Q, Cao ZX, Liu HY. *Protein Pept Lett*. 2010; 17:287–296. [PubMed: 19508202]
49. Perez-Cano L, Solernou A, Pons C, Fernandez-Recio J. *Pac Symp Biocomput*. 2010; 15:269–280.
50. Liu ZP, Wu LY, Wang Y, Zhang XS, Chen L. *Bioinformatics*. 2010; 26:1616–1622. [PubMed: 20483814]
51. Zhang Y, Skolnick J. *Nucleic Acids Res*. 2005; 33:2302–2309. [PubMed: 15849316]
52. Wang L, Brown SJ. *Nucleic Acids Res*. 2006; 34:W243–248. [PubMed: 16845003]
53. Kumar M, Gromiha AM, Raghava GPS. *Proteins*. 2008; 71:189–194. [PubMed: 17932917]
54. Terribilini M, Sander JD, Lee JH, Zaback P, Jernigan RL, Honavar V, Dobbs D. *Nucleic Acids Res*. 2007; 35:W578–W584. [PubMed: 17483510]
55. Murakami Y, Spriggs RV, Nakamura H, Jones S. *Nucleic Acids Res*. 2010; 38:W412–W416. [PubMed: 20507911]
56. Cheng CW, Su EY, Hwang JK, Sung TY, Hsu WL. *BMC Bioinf*. 2008; 9:S6.
57. Carson MB, Langlois R, Lu H. *Nucleic Acids Res*. 2010; 38:W431–W435. [PubMed: 20478832]
58. Zhang T, Zhang H, Chen K, Ruan JS, Shen SY, Kurgan L. *Curr Protein Pept Sci*. 2010; 11:609–628. [PubMed: 20887256]
59. Ma X, Guo J, Wu JS, Liu HD, Yu JF, Xie JM, Sun XA. *Proteins*. 2011; 79:1230–1239. [PubMed: 21268114]
60. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. *J Comput Chem*. 2011; 33:259–263. [PubMed: 22045506]
61. Chen Y, Kortemme T, Robertson T, Baker D, Varani G. *Nucleic Acids Res*. 2004; 32:5147–5162. [PubMed: 15459285]
62. Zheng S, Robertson TA, Varani G. *FEBS J*. 2007; 274:6378–6391. [PubMed: 18005254]
63. Gabb HA, Jackson RM, Sternberg MJE. *J Mol Biol*. 1997; 272:106–120. [PubMed: 9299341]
64. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. *Proc Natl Acad Sci U S A*. 1992; 89:2195–2199. [PubMed: 1549581]
65. Tuszyńska I, Bujnicki JM. *BMC Bioinf*. 2011; 12:348.
66. Zacharias M. *Proteins*. 2005; 60:252–256. [PubMed: 15981270]
67. Setny P, Zacharias M. *Nucleic Acids Res*. 2011; 39:9118–9129. [PubMed: 21846771]
68. Lensink MF, Wodak SJ. *Proteins*. 2010; 78:3073–3084. [PubMed: 20806235]
69. Fujishima K, Komasa M, Kitamura S, Suzuki H, Tomita M, Kanai A. *DNA Res*. 2007; 14:91–102. [PubMed: 17573465]
70. Pintar A, Carugo O, Pongor S. *Bioinformatics*. 2002; 18:980–984. [PubMed: 12117796]
71. Wang L, Huang C, Yang MQ, Yang JY. *BMC Syst Biol*. 2010; 4(suppl 1):S3. [PubMed: 20522253]

Fig. 1. Number of protein–RNA complex structures deposited in the PDB since 2001. The statistics for 2013 includes data up to June 18, 2013.

Fig. 2.
The ROC curves for several RBP predictors. SPOT-seq, RNA-pred and SVMprot are sequence-based methods while SPalign and SPOT-struc are structure-based.

Fig. 3.
Performance of RNA-binding predictions for the RB20 dataset by several sequence and structure-based techniques.

Fig. 4. Comparison between the predicted (red) and actual (green) structure and predicted (yellow) and actual (blue) binding residues. The actual RNA structure is cyan and that of the predicted structure is orange. The target is 1m8yB and the template is 3k5qA.

Table 1

Structure and sequence-based features for RBP prediction

Authors/method name [ref.]	Technique	Features
Structure-based		
Shazman and Mandel-Gutfreund ²⁶	SVM	Electrostatic surface patches, molecular weight, solvent accessibility, dipole, quadrupole, patch size, size of the largest clefts, number of atoms in positive and negative patches, patch surface overlap
Ahmad and Sarai ²⁷	NN	Charge, dipole moment, quadrupole moment and functional property of protein chain
SPalign ²⁸	Template-based	Structural alignment
SPOT-struc ²⁹	Template-based	Structural alignment plus binding affinity estimation
Sequence-based		
Yu <i>et al.</i> ³⁷	SVM	Hydrophobicity, secondary structures, solvent accessibility, van der Waals volume, polarity, polarizability and amino acid composition
Kumar <i>et al.</i> ⁴¹	Voting	Hydrophobicity, predicted secondary structure, predicted solvent accessibility, normalized van der Waals volume, polarity, and polarizability
RNApred ⁴⁰	SVM	Residue composition, predicted RNA binding residues, and PSSM
Shao <i>et al.</i> ³⁸	SVM	Clustered amino acids according to dipoles and volumes of side chains
Cai <i>et al.</i> ³⁵	SVM	Pseudo-amino acid composition, charge, hydrophobicity, and accessible surface area
Fujishima <i>et al.</i> ⁶⁹	SVM	Amino acid composition and periodicities
SVMProt ³⁶	SVM	Amino acid composition, charge, polarity, and hydrophobicity
SPOT-seq ²²	Template-based	Sequence-to-structure match and binding affinity estimation

Table 2

Comparison of methods for low-resolution, two-state RBP prediction

Method	257 RBPs + 5765 non-RBPs						245 DBPs	
	MCC	SN	PR	TP	FN	TN	FP	FP
Structure-based								
SPalign ^{a 28}	0.46	33%	67%	85	172	5723	42	6
SPOT-Str ^{a 29}	0.50	32%	84%	83	174	5749	16	3
Sequence-based								
RNApred ⁴⁰	0.15	79%	8%	203	54	3350	2415	168
SVNprot ³⁶	0.19	50%	13%	129	128	4898	867	55
SPOT-seq ^{a 22}	0.60	44%	84%	114	143	5743	22	0

SN: sensitivity, PR: precision.

^aTemplate-based approach: templates with 30% or higher sequence identities to the query sequence are removed.

Table 3

Structure and sequence-based features for RNA-binding residue prediction

Author/method name [ref.]	Technique	Features
Structure-based		
Jeong <i>et al.</i> ⁴³	NN	Secondary structure and amino acid type
KYG ⁴⁴	Scoring	Residue doublet interface propensity and multiple sequence profiles
Chen and Lim ⁴⁵	Scoring	Surface binding pocket, electropositive atoms, and spatially evolution principle
Maetschke and Yuan ⁴⁶	SVM/Naïve Bayes	Residue contacts map, PSI-BLAST profile, and graph theory properties
Struct-NB ⁴⁷		Surface roughness, interface residue propensity and CX score ⁷⁰
Li <i>et al.</i> ⁴⁸	Linear reg.	PSSM, secondary structure and solvent accessibility
OPRA ⁴⁹	Docking	Pairwise residue-ribonucleotide interface propensities
Liu <i>et al.</i> ⁵⁰	Random forest	Random forest classifier & interaction propensities, physicochemical characteristics, hydrophobicity, rASA, secondary structure, and conservation score side-chain environment
Spalign ²⁸	Template-based	Structural alignment
SPOT-struc ²⁹	Template-based	Structural alignment plus binding affinity estimation
Sequence-based		
BindN ⁵²	SVM	Side chain pK _a value, hydrophobicity index, and molecular mass
RNABindR ⁵⁴	Naïve Bayes classifier	Hydrophobicity index, rASA, and predicted secondary structure
Pprint ⁵³	SVM	PSSM
RNAProB ⁵⁶	SVM	Smoothed PSSM
BindN+ ⁷¹	SVM	Side chain pK _a value, hydrophobicity index, molecular mass, and PSSM
NAPS ⁵⁷	Bootstrap aggregation and cost sensitivity learning	PSSM
PBRpred ⁵⁸	SVM	PSSM, predicted secondary structure and solvent accessibility
PiRaNhA ⁵⁵	SVM	PSSM, residue interface propensity, and predicted residue accessibility value
PRBR ⁵⁹	Random forest	Secondary structure, evolution information, conservation information of physicochemical properties of amino acids, polarity-charge, and hydrophobicity
SPOT-seq ²²	Template-based	Sequence-to-structure match and binding affinity estimation

Table 4

The performances of structure and sequence-based methods for predicting RNA-binding residues for three domain datasets (RB106, RB67, RB20)

Method name [ref.] Dataset	Sensitivity (%) RB106(RB67)	Precision (%) RB106(RB67)	MCC RB106(RB67)	MCC RB20
Structure-based				
KYG ⁴⁴	62(61)	61(65)	0.43(0.44)	0.26
SPalign ^{a28}	57(64)	61(67)	0.39(0.50)	0.43
SPOT-struc ^{a29}	55(61)	60(69)	0.36(0.49)	0.36
Sequence-based				
BindN ⁵²	57(56)	59(64)	0.39(0.40)	0.16
RNABindR ⁵⁴	69(77)	67(65)	0.52(0.53)	0.37
Pprint ⁵³	70(73)	64(65)	0.50(0.50)	0.24
BindN+ ⁷¹	70(73)	74(77)	0.59(0.62)	0.38
NAPS ⁵⁷	42(45)	55(56)	0.28(0.28)	0.18
PBRpred ⁵⁸	74(78)	69(70)	0.57(0.59)	0.39
PRBR ⁵⁹	55(56)	69(72)	0.46(0.47)	0.22
SPOT-seq ^{a 22}	81(68)	50(82)	0.39(0.63)	0.33

^aTemplate-based approach: templates with 30% or higher sequence identities to the query sequence are removed.