Cell Reports
Resource

Discovering RNA-Protein Interactome by Using Chemical Context Profiling of the RNA-Protein Interface

Marc Parisien,[1] Xiaoyun Wang,[1] George Perdrizet II,[1] Corissa Lamphear,[4] Carol A. Fierke,[4] Ketan C. Maheshwari,[2] Michael J. Wilde,[2] Tobin R. Sosnick,[1,2,3,*] and Tao Pan[1,3,*]
[1]Department of Biochemistry and Molecular Biology
[2]The Computation Institute
[3]Institute for Biophysical Dynamics
University of Chicago, Chicago, IL 60637, USA
[4]Department of Chemistry, University of Michigan, Ann Arbor, MI 48109, USA
*Correspondence: trsosnic@uchicago.edu (T.R.S.), taopan@uchicago.edu (T.P.)
http://dx.doi.org/10.1016/j.celrep.2013.04.010

SUMMARY

RNA-protein (RNP) interactions generally are required for RNA function. At least 5% of human genes code for RNA-binding proteins. Whereas many approaches can identify the RNA partners for a specific protein, finding the protein partners for a specific RNA is difficult. We present a machine-learning method that scores a protein's binding potential for an RNA structure by utilizing the chemical context profiles of the interface from known RNP structures. Our approach is applicable even when only a single RNP structure is available. We examined 801 mammalian proteins and find that 37 (4.6%) potentially bind transfer RNA (tRNA). Most are enzymes involved in cellular processes unrelated to translation and were not known to interact with RNA. We experimentally tested six positive and three negative predictions for tRNA binding in vivo, and all nine predictions were correct. Our computational approach provides a powerful complement to experiments in discovering new RNPs.

INTRODUCTION

Over 10^5 RNAs are present in a mammalian cell and essentially all function through interaction with proteins. Recent studies indicate that a human cell contains more than 10^3 mRNA-binding proteins; over 35% of these were previously not known to interact with any RNA (Baltz et al., 2012; Castello et al., 2012), suggesting that many RNA-protein (RNP) complexes remain to be identified. Experimental approaches such as CLIP-seq and PAR-clip apply high-throughput sequencing techniques and readily identify RNA partners for a given protein on the genomic scale in cells (Hafner et al., 2010; Scheibe et al., 2012; Zhang and Darnell, 2011; Zhang et al., 2010).

Despite these advances, identification of protein partners that bind to a specific RNA remains challenging. The low abundance of many cellular RNAs often prohibits identification of bound proteins at the genomic scale. RNP interactions in cells can be transient because a given RNA can exchange protein partners during its maturation, function, and degradation. Differential expression of proteins and RNAs in distinct cell types and physiological states also contribute in making the determination of RNP interactome difficult on the basis of experimental approaches alone.

To complement experimental methods, computational approaches are highly desirable for predicting the RNP interactome (Bellucci et al., 2011; Li et al., 2012a; Pons et al., 2010; Puton et al., 2012; Setny and Zacharias, 2011; Tuszynska and Bujnicki, 2011; Zhao et al., 2011; Zheng et al., 2007). No current method, however, can provide accurate genome-wide predictions of RNPs without a priori geometrical restraints or assumptions on the protein motifs that contact the RNA. Nevertheless, recent progress including an expanded structure database (Chruszcz et al., 2010), large-scale computational resources (Wilde et al., 2011), and structural prediction for proteins (Moult et al., 2011) and RNAs (Cruz et al., 2012) have led to the development of docking and scoring methods as the first stage to uncover RNPs on the genomic scale. Numerous studies indicate that RNP interactions involve electrostatics (Bahadur et al., 2008; Chen and Lim, 2008; Polozov et al., 2006; Shazman and Mandel-Gutfreund, 2008; Tworowski et al., 2005), specific amino acid-nucleotide partners such as overrepresentation of arginine (Kim et al., 2006; Pérez-Cano and Fernández-Recio, 2010), and other factors (Draper, 1999; Fulle and Gohlke, 2010; Lunde et al., 2007). Almost all RNP docking predictions utilize statistical potentials incorporating either contact- or distance-based statistics found in solved RNPs and used thereafter as scoring function.

Here, we introduce a machine-learning-based approach to predict unknown RNP complexes followed by experimental validation (Figure 1A). We first identify transfer RNA (tRNA)-binding proteins because tRNA may have an extensive yet uncharacterized protein interaction network, and we can train using many solved transfer RNP (tRNP) structures. We also demonstrate the feasibility for non-tRNA motifs where only a single RNP structure containing the motif is available. We computationally

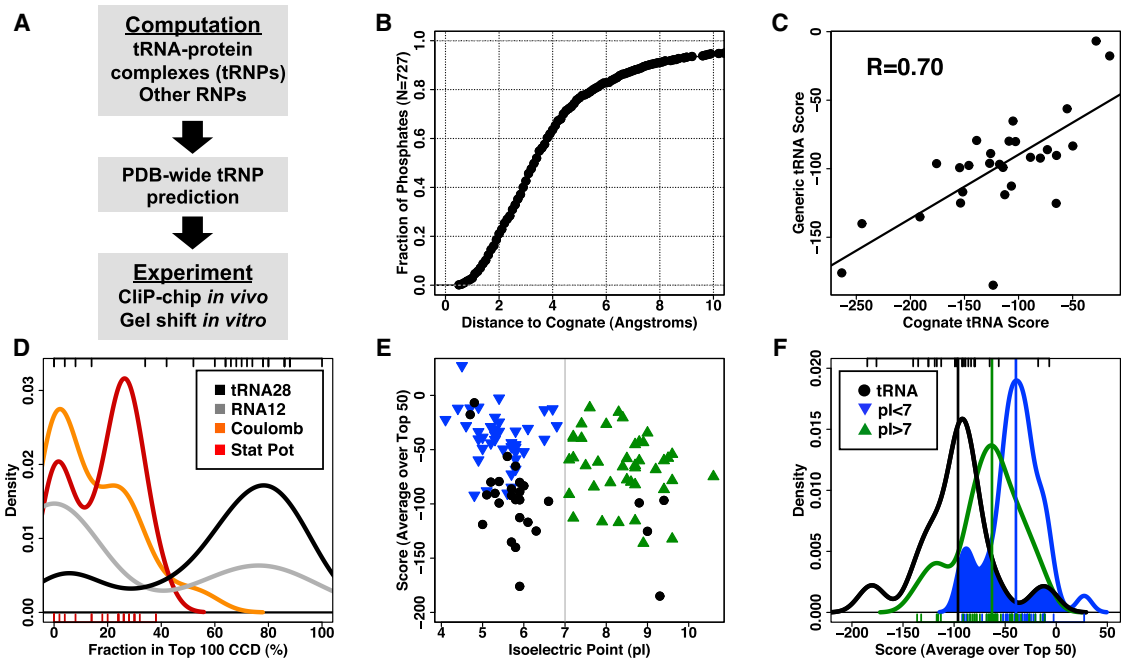

Figure 1. Performance of Scoring Functions

(A) Flow diagram of this work. We first analyze RNPs using a machine-learning, CCP-based approach. We then perform PDB-wide prediction of tRNP complexes followed by experimental validations in vivo and in vitro.

(B) A prototypical $tRNA^{Phe}$ can mimic many cognate tRNAs in the 28 known tRNPs. After superposition, 80% of the phosphate groups of $tRNA^{Phe}$ is within 6 Å of the corresponding phosphate groups of the cognate RNA.

(C) Docking scores for cognate tRNAs and $tRNA^{Phe}$ correlate with a Pearson's R value of 0.70.

(D) Four scoring functions are evaluated for their ability to identify native-like interfaces, quantified as the fraction of the top 50 scoring poses also having one of the 100 most native-like interfaces as characterized using CCD: tRNA specific trained on 28 tRNPs (tRNA28; black), general RNP trained on 12 non-tRNPs (tRNA12; gray), Coulombic terms only (orange), and an all-atom statistical potential (Stat Pot; red).

(E) Distribution of scores with respect to pI of the protein. The 28 tRNPs used in the training are shown as black circles. Randomly selected proteins from the PDB are shown as blue diamonds for those with pI <7 and as green triangles with pI >7.

(F) Overlap between the scores of the three sets from (E). The blue region highlights the overlap between the known tRNPs with randomly selected proteins with pI <7. The tick marks at the top of the plots indicate the scores of the known tRNPs, whereas those at the bottom are for the two randomly chosen protein sets. See also Tables S1, S2, S3, S4, S9, and S12 and Figures S1, S2, S3, S5, and S6.

test ~800 mammalian proteins and identify dozens of novel hits for tRNA binding. We experimentally test six positive and three negative predictions in mammalian cells and find that all nine predictions are correct.

RESULTS

Docking Decoys and Representation of RNA and Protein in the Complex

We selected tRNA to begin our study for a variety of computational and biological reasons. It is relatively small (~76 nt) and has a well-defined structure that increases the likelihood of success with rigid body-docking methods. Mammalian cells contain up to 10^8 tRNA molecules distributed among ~50 isoacceptor families. tRNA is ancient and may have evolved to have an intricate network of protein partners. Finally, the presence of many tRNP complexes in the PDB (Berman et al., 2007) combined with contemporary machine-learning techniques enable the derivation of a scoring function well suited for finding novel tRNPs.

The computational approach has three steps: docking, scoring, and ranking of a known protein structure against the canonical structure of yeast $tRNA^{Phe}$ (Figure S1A). The computation starts without any a priori geometrical information, sequence conservation, or knowledge of RNP-binding motifs. We only use the coordinates of 28 known tRNPs (Table S1). The program FTdock (Gabb et al., 1997; Sternberg et al., 2000) is used to generate up to 10^5 initial decoys for each tRNP pair in the learning set. The large number of decoys adds to the difficulty in selecting native-like poses, thus sharpens the specificity of our scoring function. Examples of docked complex involve the protein surface and many tRNA orientations and generate many near-native docking poses (Figure S1B). We modified the FTdock program to work on large computer networks by sharing the loads among many processors. This level of resources is essential for the present study.

To derive a scoring function, a coarse-grained representation is used to drastically speed up computations while retaining the primary determinants of tRNP interaction. The Cβ carbon is selected as the interacting center for each amino acid, whereas

a heavy atom in the major groove, the minor groove, and the phosphate group is selected as the interaction points for each nucleotide types (Figure S1C). An example for 1 of the 300 interacting pairs (20 amino acids × 15 nt interacting points) is shown between the alanine and phosphate groups of adenosine (Figure S1D). The interaction strength is assumed to be constant within the average extent < e > past the Cβ atom for each amino acid type; beyond this extent, the energy decays as 1/*r* such that $f(r) = 1/\max(3.5\text{Å}, r - \langle e \rangle)$.

Chemical Context Profile and Chemical Context Discrepancy

Our goal is to predict a realistic docking geometry with a chemically reasonable set of contacts. We developed chemical context profile (CCP), which is a representation of a docking pose designed to capture the preferences of the contacting chemical contexts at the docking interface. CCP is a 300-dimensional vector where each dimension is the interaction strength associated with each of the 300 RNP pairwise interaction types. The strength is chosen to be inversely proportional to the distance separating the pair.

We compare the CCP of a predicted complex with that of the native complex by comparing the magnitudes of their corresponding CCP entries. The vector representation enables this comparison by calculating the angle between the predicted and native CCP vectors. This angle is the chemical context discrepancy (CCD). CCD properly identifies native-like docking poses for multimeric proteins and is strongly preferred over the conventional rmsd to guide machine learning in the derivation of a scoring function. The CCP/CCD paradigm has been shown to be useful in the studies of DNA-binding potential and DNA-protein docking sites (Parisien et al., 2012), although substantial modification is needed to investigate RNP complexes.

CCD quantifies the differences among chemical complementarity of docked complexes. For instance, among the 10^5 decoys of the symmetric dimer 1ASY, the CCD plot features two minima: one at low rmsd, and another at rmsd values around 60 Å (Figure S2). In this particular example, the use of rmsd as a measure of goodness of fit would fool machine learning by providing for discordant inputs: one is good (low rmsd), the other is bad (rmsd near 60 Å), although both have exactly the same interaction interface. Because CCD properly identifies native-like docking poses for multimeric proteins, it is strongly preferred over the conventional rmsd to guide machine learning in the derivation of a scoring function.

Cognate versus Generic tRNA

A large number of binding surfaces are provided by the 2,000+ tRNAs (Chan and Lowe, 2009), which are further enhanced by the presence of modified nucleotides (Motorin and Helm, 2010). This diversity is currently intractable computationally. To reduce the search, we utilize a single prototypical, free-form tRNA, yeast $tRNA^{Phe}$ (*4tra*). The use of the free-form $tRNA^{Phe}$ scaffold is warranted because ~80% of the phosphate groups of $tRNA^{Phe}$ are within 6 Å of the cognate tRNAs among the 28 known tRNPs (Figure 1B). Bound tRNAs undergo conformational changes with respect to their free-form states, but the extent of the conformational changes has an upper-bound limit within one-half of the width on an RNA helix (~6 Å). Furthermore, the score of the cognate tRNP interaction is proportional to that for $tRNA^{Phe}$ with a Pearson's correlation coefficient of R = 0.7 (Figure 1C). Hence, the free-form $tRNA^{Phe}$ is a credible surrogate for many tRNAs. This simplification greatly reduces the computational requirements at the expense of rendering our method partially insensitive to tRNA sequence and modification. We will not identify all tRNA-binding proteins or be able to predict which specific tRNA binds to which protein. This deficiency is compensated for here because of the potential of our method to identify new tRNA-binding proteins at the genomic scale.

Scoring Function

The CCP captures the interaction of the protein and RNA moieties at the tRNP interface and serves as the basis of our scoring function. However, the high dimensionality of the CCP vector makes it difficult to weigh the magnitude and the sign of the RNP interaction for all 300 entries. To reduce the number of required entries and to identify key RNP interactions, we utilize a forward version of the sequential feature selection (SFS) approach (Romero and Sopena, 2008). SFS enables the identification of the most-important interacting pairs among all possible ones. We activate, one at a time, an interacting pair and evaluate its performance at identifying native-like docked conformations. The activated pair that gives rise to the best performance is permanently activated, and the process is repeated. Hence, at each step, we identify an important interaction term, although we do not know if it is used to directly identify native poses or to discriminate against the large decoy sets.

As more key interaction terms are picked, the identification performance plateaus and even starts to decrease when too many terms are used (Figure S3A). This decrease is due to more interaction terms being better able to discriminate the native poses during training, but it starts to "learn-by-heart" these docking poses at the expense of poorer identification performance on unseen cases. After activating up to 32 terms, the optimal number of interacting pairs is found to be only 12 before overtraining sets in Figures S3B and S3C. This low number of parameters makes it likely that the scoring function will be robust through a range of various tRNA shapes and sequences. The final scoring function has the form

$$S = Coulomb + \overrightarrow{\omega_{ccp}} \bullet \overrightarrow{ccp},$$

where $\overrightarrow{\omega_{ccp}}$ has just 12 nonzero dimensions and therefore is exceedingly fast computationally, enabling scoring of very large decoy sets at the genomic scale.

The attractive and repulsive interplay between specific pairs of moieties reflects the specificity of tRNP docking (Table S2). The docking poses are too imprecise to incorporate hydrogen-bonding interactions and may explain why the Coulomb term is inadequate by itself. To bind a tRNA, a protein has to pierce the negatively charged envelope of the nucleic acid. The largest electropositive patch, however, may not be the actual binding surface. For example, arginine specifically interacts with most nucleotides in addition to the electrostatic term. Although Arg residue can bind to the major groove or the Hoogsteen edge of

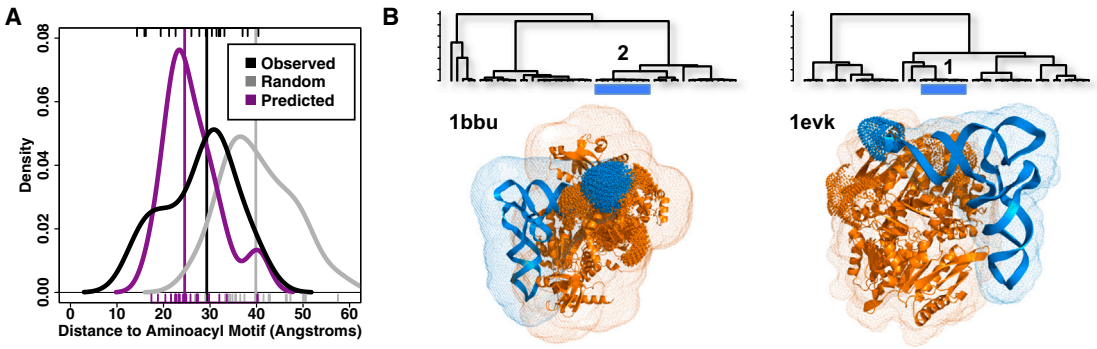

Figure 2. Computational Test of Known tRNA-Binding Proteins

(A) Distance distribution of aminoacyl motif residues to tRNA phosphate 76 for 24 aaRSs for which only the apo structure has been solved. Our predicted distribution of average distances (magenta) is compared to the observed distribution of 19 solved aaRS-tRNA complexes (black). A poorly performing scoring function for tRNP such as RNA12 is unable to position the tRNA phosphate 76 close to the active site of the aaRS (gray).

(B) Two examples of positioning of the tRNA-3′CCA tail in the aaRS's catalytic site. Shown on top is the rmsd structural clustering of the best 50 scores. Also shown is the centroid of the cluster with the least average distance of the tRNA 3′CCA tail (blue dots) to the protein's catalytic site residues (orange dots).

See also Tables S3 and S5.

guanosine (Kondo and Westhof, 2011), this interaction type is not emphasized as strongly in our scoring function compared to the other arginine-related interactions.

The magnitude and sign of these 12 weights in $\overrightarrow{\omega_{ccp}}$ are set in a subsequent learning step. The scoring function is chosen to identify native-like docking poses with low CCD values, while correlating the total scores computed for yeast $tRNA^{Phe}$ with those computed for the cognate tRNA (Figure 1C). Despite the simplicity of the scoring function, it can identify many native-like docking poses in very large decoy sets.

Scoring tRNP Complexes

We apply two filters to define a successful docking trial. The first filter consists of retaining the best 50 scores among the 10^5 docking poses. The next filter consists of retaining those docking poses that are native-like as defined by having a CCD rank lower than 100. A successful trial is defined as when more than ten poses pass both filters. Nativeness is determined by the CCD value obtained from the vector dot product of the candidate and native CCP vectors. The method is successful for 23 of the 28 tRNPs (82%) and outperforms methods using Coulomb terms or an all-atom statistical potential (Figure 1D; Table S1).

Our scoring function is able to identify authentic tRNA-binding proteins (Figures 1E and 1F; Table S3). The scores of the proteins in the known 28 tRNPs are well separated from those of 81 randomly selected proteins from the PDB (Table S4) that are presumed not to bind tRNA. Among the 28 tRNPs used in the machine learning, 24 have an isoelectric point (pI) below 7 as do half of the known RNA-binding proteins (Castello et al., 2012). For these acidic proteins, binding affinity is achieved through detailed interactions rather than generic charge-charge interactions. This property likely increases binding specificity at the expense of affinity. In fact, for the 28 known tRNA binders, the mean score of the 24 acidic proteins is 1.5 SDs better than for the four basic proteins with only a 29% overlap (Extended Discussion, 1). This result indicates that our scoring function for tRNA binding is largely trained for acidic proteins.

To further investigate the methodology, we examine the CCP contribution derived from the two major functional hot spots in tRNA: the acceptor stem (AA, nt 1–4); and the anticodon stem loop (AN, nt 33–37). One contacting nucleotide contributes approximately a value of 40 to the AA or AN score. For the 28 tRNPs, the contributions for these two regions are 140 ± 109 and 68 ± 67, respectively (Table S3). Predicted and cognate AA or AN values highly correlate (R^2 = 0.8 with a slope of 1.05 for AA, and R^2 = 0.5 with a slope of 1.06 for AN; Table S3) when the cognate score is greater than 40.

We use both CCD and AA/AN scores to predict tRNA binding for a protein. On the basis of the 28 tRNP training set, we set the CCD score to be lower than −50, which is 1.3× SD above the average of the known tRNA-binding proteins (−98 ± 38; Figure 1F), and the AA or AN score to be greater than 40.

We tested our ability to identify known tRNA-binding proteins using their solved apo structure (Figure 2; Table S5). This test serves to quantify our ability to identify tRNA-binding proteins, the correctness of the docking pose, and as a realistic control for our PDB-wide screen for tRNA binding where only the unbound protein structure is available. We choose 24 aminoacyl-tRNA synthetases (aaRSs), and 20 (83%) are identified as tRNA binding (Table S5), indicating that our scoring function has a very good chance of identifying potential new tRNA-binding proteins even when only the apo structure is known.

We compared the distance between aaRS's active site and the amino acid attachment site on tRNA's 3′ terminal residue (A76) in our predictions with the distribution in known structures (TRNASYNTH motif in the PRINTS database [Attwood, 2002] via the InterPro web site [Hunter et al., 2009]). For the 19 solved aaRS-tRNA complexes, the average distance of active site residues to A76 is less than 30 Å with a SD of 10 Å (Table S3). This distribution is comparable for our predicted docked tRNA conformations with the 24 aaRS apo-proteins (Figure 2A; Table S5). Our scoring function is therefore able to position the 3′CCA tail of tRNA within the catalytic sites of aaRSs (examples shown in Figure 2B). This result indicates that our approach can

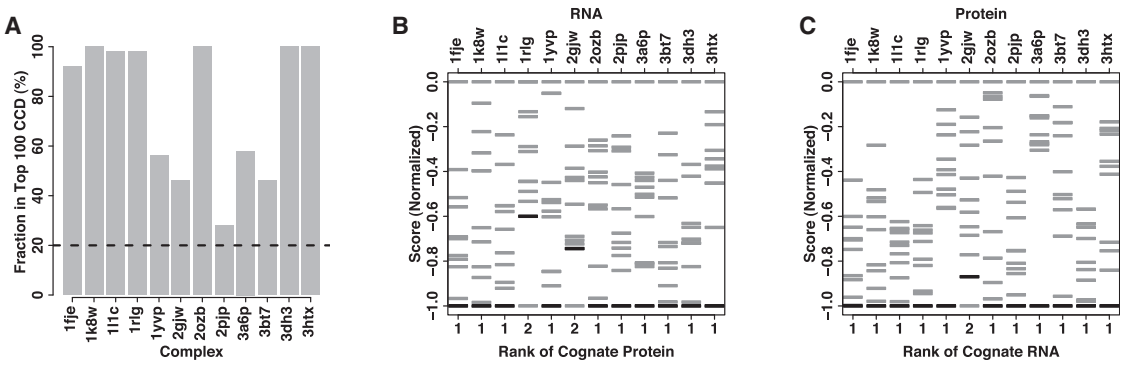

Figure 3. Performance of Individual Scoring Functions for 12 Non-tRNA-RNPs with Distinct RNAs

(A) Fraction of the top 50 predictions in the top 100 poses for each protein. A threshold of 20% generates at least ten native-like poses among the top 50 scores among 10^5 decoys and is considered a success (dashed line). The same scoring function is also applied to simultaneously select (B) the cognate protein for each RNA and (C) the cognate RNA for each protein. For ease of comparison, the scores of the various complexes are normalized such that the lowest (best) score is −1, and the highest (worst) score is 0. Cognate partners are indicated with the thick black bar; the x axis shows the ranking of the cognate complex for each RNP. See also Tables S6 and S10.

be successful in the identification of docked conformations on apo structures, although not yet at atomic resolution.

Scoring for Non-tRNA Structural Motifs

We next examined whether our CCP-based scoring also works for other RNPs. This task is far more challenging because the number of solved RNP structures is far fewer for non-tRNA structural motifs. We selected 12 distinct RNP structures to carry out this analysis (Table S6). Our approach is inaccurate when all structures are used to generate a single scoring function, presumably due to the large diversity of interactions among distinct RNPs (Extended Discussion, 2).

More success is obtained when each RNP is used to derive its own scoring function, in particular, in the correct identification of RNP pairs and docking sites (Figure 3). A robust scoring function can be derived from a single RNP by requiring that it simultaneously satisfies three criteria: (1) it identifies the native docking pose in large decoy sets; (2) the protein recognizes the cognate RNA scaffold among noncognate RNAs; and (3) the RNA recognizes the cognate protein among noncognate proteins. We examined the entire 12 protein × 12 RNA docking space by generating 144 decoy sets containing a total of 1.44 × 10^7 decoys. Because only one cognate RNP is used in machine learning, the learning capacity is decreased, and the task of learning is made more stringent by using only 6 randomly chosen weights instead of the 12 used for tRNPs (Extended Experimental Procedures). We applied the same criteria used for tRNPs to define a successful docking trial. We found the cognate docking site for all 12 RNPs (Figure 3A). For the identification of native RNP pairings, 10 of 12 proteins have the best score, and the 2 remaining proteins have the second-best score for docking with cognate RNA (Figure 3B); 11 out of 12 RNAs have the best score, and the 1 remaining RNA has the second-best score for docking with cognate protein (Figure 3C). This result shows that the CCP approach is capable of finding RNPs as long as each RNA structural motif is trained separately. To discover unknown RNPs for a non-tRNA structural motif, the scoring function should be made more specific by requiring it to have maximal score separation between the cognate and noncognate RNPs, as we have done for tRNPs (Figure 1F). Nevertheless, our result demonstrates that a single-solved RNP structure is sufficient to generate a viable, CCP-based scoring function for RNP.

PDB-wide Computational Screen

In order to discover tRNPs, we computationally screened 801 unique mammalian proteins for their propensity to bind tRNA (Table S7). We screened only acidic proteins (pI <7) because our tRNP scoring is largely trained for acidic proteins. Using both the CCD and the AA/AN scores, 37 proteins or ∼4.6% of the screened proteins are identified as tRNA binding (Table 1). Three proteins that are known to interact with tRNA have scores worse than −50, representing a false-negative rate of less than 0.4% (Table S8). Among the 24 known tRNA-binding proteins screened separately, our positive rate was >83% (Figure 2).

Protein binding to a tRNA may be functionally important as illustrated with several predicted complexes (Figure 4, shown according to the overall scores). These predicted tRNA-binding proteins are involved in a wide range of cellular processes, including protein modification, histone modification, cell-cycle control, gluconeogenesis, glutathione synthesis, and membrane trafficking. These proteins are all cellular enzymes that are previously not known to interact with any RNA; tRNA binding may help regulate the activity of these enzymes.

It is commonly assumed that a large positive electrostatic patch on the protein surface is a prerequisite for nucleic acid binding (Ahmad and Sarai, 2011; Bahadur et al., 2008; Chen and Lim, 2008; Polozov et al., 2006; Shazman and Mandel-Gutfreund, 2008; Tworowski et al., 2005; Tworowski and Safro, 2003). The proteins we analyzed here for potential tRNA binding all have pI below 7, and their ability to dock tRNA relies on the spatial organization of their positively charged amino acids. The total binding score is not entirely of Coulombic nature (Table 1; total score versus Coulomb score). The location of the largest positive patch does not always coincide with the predicted tRNA-binding site. Other precedents for unfavorable Coulomb interactions between proteins and their bound tRNAs can be found in known tRNPs (Figure 1D; Table S3), and the Coulomb

Table 1. Predicted Mammalian tRNA-Binding Proteins

PDB	Score	Coul	AA	AN	pI	Description
2D39-1	−91.8	−16.8	45.8	20.5	5.6	Ficolin-1
2IAG-1	−89.2	−55.4	54.3	7.0	6.7	Prostacyclin synthase
1KHB-1[a]	−84.7	−25.8	71.6	12.7	5.7	Phosphoenolpyruvate carboxykinase, cytosolic (gtp)
1R42-1	−82.5	+12.6	66.3	13.2	4.8	Angiotensin i/collectrin homology domain
3IFQ-1	−79.5	−34.1	44.8	26.8	6.2	Plakoglobin/e-cadherin
1B41-1,-2	−78.2	−35.2	52.2	7.8	6.4	Acetylcholinesterase/fasciculin-2
2BYD-1	−76.9	−34.1	41.2	15.6	6.4	hspc223
3I2B-4	−74.2	−54.1	52.1	28.8	6.5	6-Pyruvoyl tetrahydrobiopterin synthase
1ND7-1	−74.2	−26.4	50.0	5.7	5.9	ww domain-containing protein 1
1SIQ-1	−73.7	−25.0	44.0	20.3	6.0	Glutaryl-coa dehydrogenase
2P0A-1	−73.7	−46.3	73.5	12.5	6.4	Synapsin-3
2Q32-1	−71.8	−10.4	61.0	3.7	5.5	Heme oxygenase 2
2FMM-1	−70.5	−21.0	46.2	20.8	6.2	Protein emsy/chromobox protein homolog 1
1D8D-1[a]	−70.3	−33.2	34.3	43.1	5.9	FT (α)/FT (β)
2CJW-1	−69.5	−35.1	42.2	19.3	6.0	gtp-binding protein gem/gtp-binding protein gem
3I2B-3	−69.2	−49.9	45.6	17.1	6.2	6-Pyruvoyl tetrahydrobiopterin synthase
2GAO-1[a]	−69.1	−31.7	52.4	13.4	6.6	gtp-binding protein sar1a
2VGQ-1	−68.6	+1.0	59.5	21.8	4.8	Maltose-binding periplasmic protein
1CJL-1	−67.2	−7.7	64.7	5.8	5.3	Procathepsin l
3ISQ-1	−66.9	−48.4	50.1	16.1	6.7	4-Hydroxyphenylpyruvate dioxygenase
3EHT-1	−65.3	−7.3	54.5	12.4	5.0	crfr1 extracellular domain and mbp/corticoliberin
2IGQ-1[a]	−63.0	−31.6	131.1	0.2	5.8	Euchromatic histone methyltransferase 1
2GL7-1	−62.8	−44.6	52.0	20.5	6.6	β-catenin/transcription factor 7-like 2/b-cell lymphoma 9
2B7A-1	−62.0	−22.6	48.8	5.0	6.8	Tyrosine-protein kinase jak2
3D8B-1	−61.2	−10.1	48.6	8.8	5.3	Fidgetin-like protein 1
2PET-1	−60.7	−33.7	49.4	3.7	6.3	Lutheran blood group glycoprotein
1S9I-1[a]	−59.5	−36.1	104.8	1.5	6.3	Dual-specificity mitogen-activated protein kinase kinase 2
2HGS-1[a]	−58.2	−28.3	42.7	3.6	5.6	Protein (glutathione synthetase)
2G01-2	−55.1	−26.3	91.2	6.6	6.4	Protein kinase 8/c-jun-amino-terminal kinase-interacting 1
3FQW-1	−54.1	−39.7	54.7	13.6	6.1	hla class I histocompatibility antigen, a-2 a/b-2-microglobulin
1FCH-3,-4	−53.4	+19.8	43.9	21.9	4.7	Peroxisomal targeting signal 1 receptor
1EZF-2	−51.5	−12.7	62.4	4.6	5.5	Farnesyl-diphosphate FT
1JUO-1	−51.1	−7.3	43.7	12.3	4.9	Sorcin
1DHS-1	−50.7	−3.4	69.6	1.6	5.3	Deoxyhypusine synthase
2VUX-1	−50.4	−8.6	55.2	2.3	5.2	Ribonucleoside-diphosphate reductase subunit m2 b
3N1G-1	−50.3	−11.9	49.1	3.2	5.2	Desert hedgehog protein/brother of cdo
3E7O-1,-2	−50.2	−26.5	59.0	7.3	6.2	Mitogen-activated protein kinase 9

Each protein is indicated by its PDB entry. Scores shown are the average over the top 50 best scores. Coulomb (Coul) is the electrostatic contribution. AA is the AA contact, and AN is the AN contact.
See also Tables S7 and S8.
[a]Proteins tested for tRNA binding in vivo.

field alone does not perform well at coordinating docking configurations (Figure 1D).

Experimental Validation

We tested six predicted hits for tRNA binding in cells along with three proteins that we predicted not to bind tRNA (Figure 5). All nine proteins are known to be of moderate abundance in HEK293T (Uhlen et al., 2010). We applied UV-crosslinking immunoprecipitation followed by tRNA microarray (CLIP-chip) for the experimental validation (Ule et al., 2003, 2005; Zhang and Darnell, 2011). In CLIP-chip (Figure 5A), HEK293T cells were UV crosslinked, lysed, and the crosslinked RNP complex was purified at high stringency using antibodies specific for the protein of interest. The purified complex was then treated with Proteinase K to completely degrade the protein, and all RNAs present were visualized by 3′ ^{32}P labeling. Radiolabeled RNAs of tRNA sizes were excised from the denaturing gel, and the tRNAs were identified by

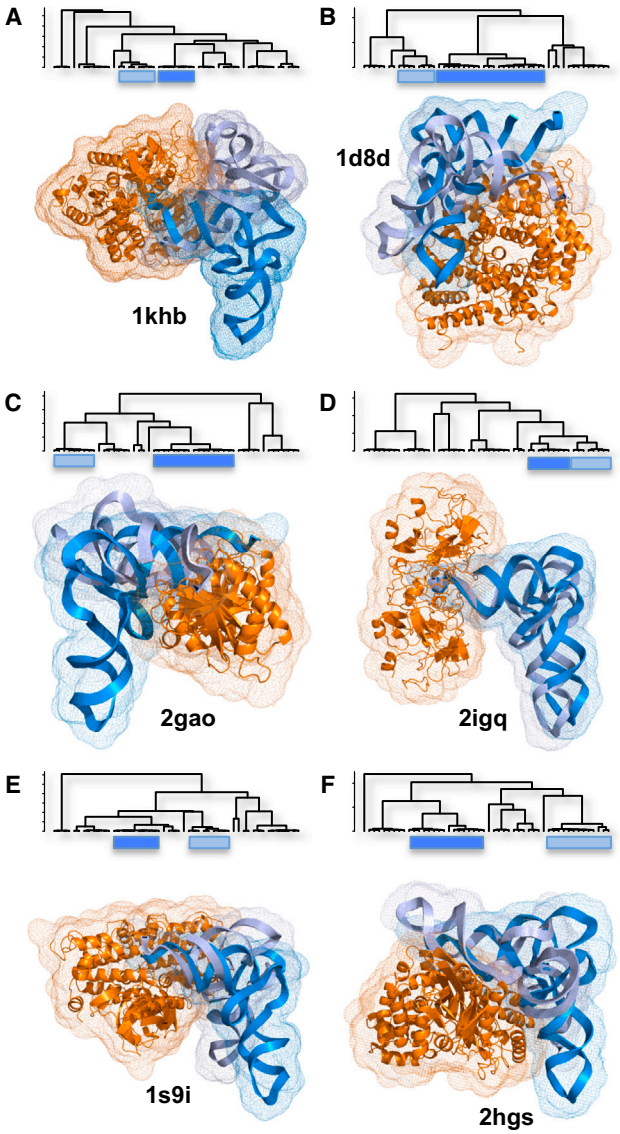

Figure 4. Six Predicted tRNPs that Are Experimentally Tested

Rmsd clustering of tRNAs (blue) after optimal superposition of proteins (orange) for the 50 top-scoring poses. Our method does not yet predict a single specific structure at atomic resolution; rather, many top-scoring docked configurations form structural clusters. The centroid centers of the largest (1; marine) and second-largest (2; teal) clusters are shown on top. These two clusters are shown as ribbons inside shells in the 3D representations.

(A) Phosphoenolpyruvate carboxykinase (1KHB) is a metabolic enzyme that converts oxaloacetate to phosphoenolpyruvate.

(B) FT (1d8D) is a protein modification enzyme that adds a farnesyl group to proteins with a CaaX motif near C terminus.

(C) GTP-binding protein SAR1a (2GAO) is a GTPase involved in membrane trafficking of other proteins.

(D) Euchromatic histone methyltransferase 1 (2IGQ) is a histone modification enzyme.

(E) Dual-specificity mitogen-activated protein kinase kinase 2 (1S9I) is a protein kinase involved in cell-cycle regulation.

(F) Glutathione synthetase (2HGS) is an enzyme that synthesizes a metabolite to maintain cellular redox state.

tRNA microarrays. We applied one positive control using the antibody against EF-1α, which is known to bind all elongator tRNAs, and two negative controls using nonimmunized serum (IgG) and GFP antibody (Figures 5B–5D). We find that our predictions are 100% accurate. There were no unsuccessful trials.

We further tested whether tRNA binding occurs in vitro for one predicted protein, farnesyltransferase (FT; Figure 6). Recombinant rat FT was purified from *E. coli*, and tRNA binding was examined by native gel shift followed by tRNA microarrays. Using total human tRNAs, two gel-shifted complexes were identified, and many, but not all, tRNAs were bound by FT, similar to the binding observed in vivo (Figures 6A and 6B). The half-saturation point was ~0.3 μM for complex 1 and ~2 μM for complex 2 (Figure 6C). The cellular concentration of total tRNA in a mammalian cell is on the order of ~30–100 μM, and an average concentration of a tRNA isoacceptor is in the micromolar range (Dittmar et al., 2006). The gel shift result indicates that FT binds tRNA at affinities relevant to physiological conditions.

DISCUSSION

We have described a computational approach to enable PDB-wide screening of potential RNA-binding proteins. Our scoring function is devised with CCP, a compact representation of a docking configuration that reflects the stereochemical features of the binding interface. The quality of a docked pose with respect to the cognate RNP structure is measured using the CCD, instead of the more conventional rmsd. CCD is better suited here in part due to its robustness to chemically equivalent docking sites. For tRNP, our scoring function requires only 12 parameters obtained using machine-learning methods. This small parameter space contrasts with the several orders of magnitude larger number of parameters required to encode statistical potentials, a standard approach for addressing this problem (Tuszynska and Bujnicki, 2011). We further show that our approach of CCP coupled with machine learning can be applied to non-tRNP complexes.

Several studies have previously proposed RNP scoring functions mainly of the contact-only or distance-dependent types; they perform poorly or feature many thousands of parameters. These scoring functions are knowledge based because they are derived from statistics extracted from solved RNP structures. Because structures for relatively few RNPs have been solved so far, the validity of scoring functions based on low-statistical counts and high-dimensional feature spaces is difficult to assess. Our scoring function, however, is tolerant to the atomic-level faults at the atomic scale but precise enough to assess the potential RNA-binding ability of a protein. Furthermore, our scoring function can identify the native docking pose among many thousand alternatives for a given RNP pair. We still make use of solved RNPs, not to provide for statistical counts but to guide the machine-learning process at identifying native-like docked complexes.

One disadvantage of ours and many other approaches is that potential conformational changes upon binding are not explicitly considered. Induced fit and conformational selection are common in RNP formation (Fulle and Gohlke, 2010; Shajani et al., 2011), although major changes are rare (Ellis and Jones, 2008).

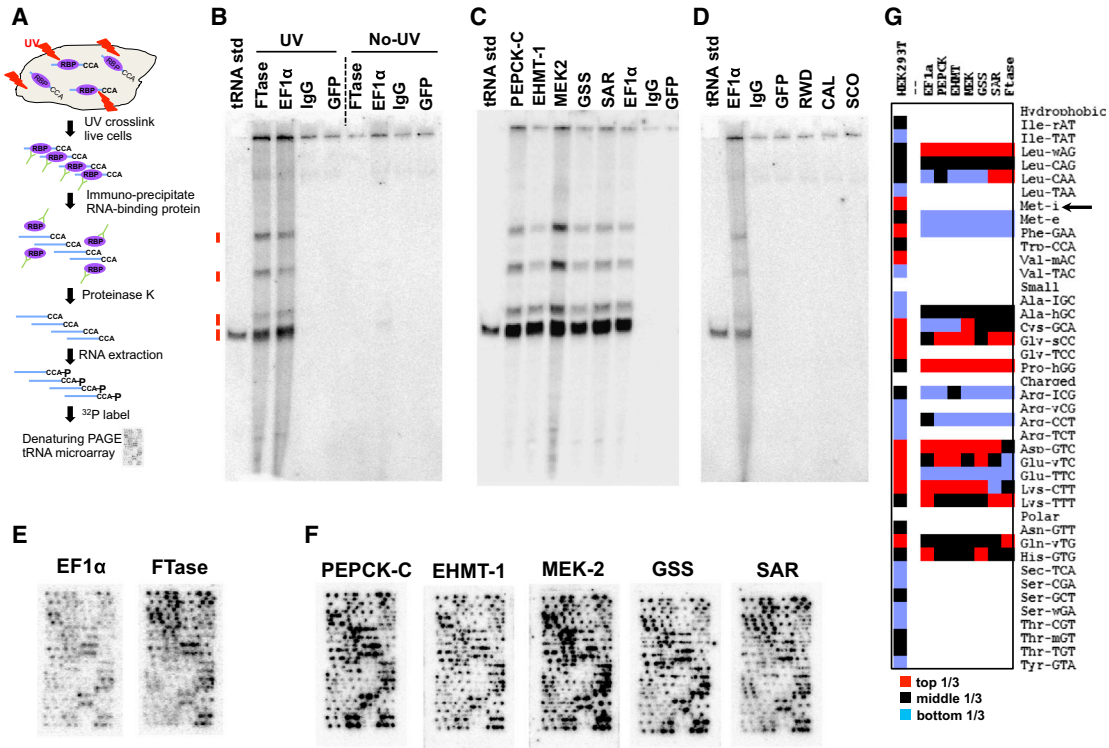

Figure 5. In Vivo Validation of Nine Predictions in HEK293T Cells

(A) CLIP-chip workflow. The final RNA products are ^{32}P labeled and analyzed on denaturing PAGE (B–D) and tRNA microarray (E–G). Among the nine proteins analyzed, six are predicted to bind tRNA (Table 1): FT (FTase), phosphoenolpyruvate carboxykinase (PEPCK-C), euchromatic histone methyltransferase 1 (EHMT-1), mitogen-activated protein kinase kinase 2 (MEK-2), glutathione synthetase (GSS), and GTP-binding protein SAR1a (SAR). Three are predicted not to bind tRNA (Table S8): RWD domain-containing protein 1 (RWD), calsequestrin-2 (CAL), and SCO1 protein homolog (SCO).

(B) PAGE analysis of FTase with or without UV crosslinking. EF1α binds all elongator tRNAs and is a positive control. Nonimmunized serum (IgG) and GFP antibody are negative controls. Yeast tRNA^{Phe} is a tRNA standard (std). Cells contain type I (75–78 nt) and type II (83–93 nt) tRNAs (indicated by red dashed lines).

(C) PAGE analysis of the five other predicted tRNA-binding proteins.

(D) PAGE analysis of three proteins that are predicted not to bind tRNA.

(E) tRNA microarray analysis of FTase and EF1α samples validates the identity of tRNAs derived from the bands shown in (B).

(F) tRNA microarray analysis of tRNA species for the other predicted tRNA-binding proteins from the bands shown in (C).

(G) Semiquantitative analysis of Clip-chip array results comparing the top 18 tRNA species crosslinked to the respective protein. The tRNA abundance is grouped in top, middle, and bottom thirds. As with EF-1α, all six proteins in various degrees prefer binding to elongator tRNAs, but not to initiator-tRNA$_i^{Met}$ even though tRNA$_i^{Met}$ is highly abundant in HEK293T cells (arrow). The tRNA abundance in HEK293T was determined previously using our standard, fluorescence-based array method (Novoa et al., 2012).

Folding upon binding represents an extreme class of conformational change, but it may be an option for high-confidence predictions. Another challenge in the computational analysis of RNP interactome is that many RNA-binding proteins recognize just RNA sequences that are in single-stranded regions (Extended Discussion, 3; Agostini et al., 2013; Goodarzi et al., 2012; Serganov and Patel, 2008; Shulman-Peleg et al., 2008).

An inherent limitation of using only solved structures is that many of these are protein domains, not full-length proteins. However, using mammalian protein domains is still valid because many mammalian proteins in the absence of their interacting partners are made of folded domains that are connected like beads on a string. Undoubtedly, our approach is incomplete and likely misses many tRNA-binding proteins in the cell. However, our approach is fundamentally useful: it predicts specific proteins that can be tested experimentally, and it is highly successful because we discovered six tRNP complexes in vivo.

The six tRNA-binding proteins have broad binding selectivity for tRNA, similar to many mRNA-binding proteins that recognize broad sequence/structural motifs in many mRNAs (Ascano et al., 2012; Licatalosi et al., 2008). Micromolar-binding affinity may be sufficient for tRNA-binding proteins to perform their function because it matches the cellular tRNA concentration. Noncanonical roles of tRNA are previously known for only two other mammalian proteins: the regulation of protein kinase GCN2 activity in stress response (Hinnebusch and Natarajan, 2002); and the prevention of HIV *gag* protein synthesis through binding to the innate immune protein Slfn11 (Li et al., 2012b). Our prediction and discovery of many tRNA-binding proteins suggest a widespread, noncanonical role for tRNP interaction in cellular communications between translation and other processes (Figure S4). In this model, when translation activity is high, most tRNAs are used by the ribosome, and only a small amount of tRNA is available to interact with other proteins. When translation

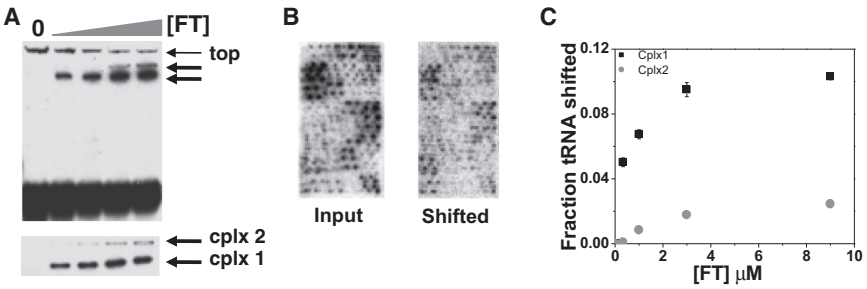

Figure 6. In Vitro Validation of a Predicted tRNP

(A) Native gel shift of recombinant rat FT with 5′ ^{32}P-labeled total human tRNA. Two shifted complexes (cplx) are present as shown in the full gel (top) and in the inset from a second gel shift experiment.

(B) tRNA microarray results comparing the input and the gel shift tRNAs. Many tRNAs were bound by FT, consistent with the CLIP-chip result in vivo.

(C) Quantitative analysis shows half-saturation in the range of ~0.3 and ~2 μM for complex 1 and 2, respectively.

activity is low, more tRNA becomes available to interact with other proteins, which may result in up- or downregulation of a wide array of diverse cellular processes.

In summary, we present computational RNP prediction at the genomic scale together with experimental validation of six RNPs. Our results raise important biological questions for the role of the discovered tRNPs. RNPs have become increasingly important in defining the functions of RNA and proteins. RNP formation occurs at different times and places and at different levels in cells; our computational approach can be an excellent complement to experiments by either suggesting new targets of investigation or providing independent validation and ranking of experimental results. Furthermore, our CCP-based scoring emphasizes binding specificity that is often more important than affinity in biological function (Shajani et al., 2011). Our approach should be useful to uncover and extend the network of RNA-binding proteins, thus achieving better understanding of RNA biology.

EXPERIMENTAL PROCEDURES

Sections described in Extended Experimental Procedures include tRNP benchmark, docking pose generation, machine learning, and validation of FT binding to tRNA.

CCP and CCD

Because docking prediction aims to reproduce the contacts observed in solved complexes, we employ a compact representation of any docking configuration that captures the nature and magnitude of the contacting moieties. This is done with the CCP, a 300-dimensional vector:

$$\overrightarrow{CCP} = \left(\sum_{C_\beta}^{ala} \sum_{M}^{A} f(r), \sum_{C_\beta}^{ala} \sum_{m}^{A} f(r), \sum_{C_\beta}^{ala} \sum_{P}^{A} f(r), \cdots, \sum_{C_\beta}^{val} \sum_{P}^{T} f(r) \right),$$

where the double sum is over a given pair of moieties (e.g., the first term is between all Cβ of alanine and major groove of adenosine). The 300 dimensions represent the product of the 20 amino acid types multiplied by 15 types for nucleic acids (three interacting centers, major groove [M], minor groove [m], and phosphate group [P], for the five nucleotides [A, C, G, U, and T] to cover both RNA and DNA). For RNA, all entries that pertain to thymines (T) have a CCP value of zero. The energy function $f(r)$ has a form similar to a Coulombic $1/r$ potential but with a separation distance that accounts for the average extent of the protein's side chain beyond the Cβ atom (Table S11): $f(r) = 1/\max(3.5\ \text{Å}, r - \langle \text{extent} \rangle)$. Subsets of the CCP vector can be used to estimate the extent of protein binding to particular stems loops of the tRNA. Because both the AA and the AN are two functional hot spots of tRNA, we use the summations of the CCP for nt 1–4 for AA and 33–37 for AN. The higher these values, the more the protein contacts these regions.

Any RNP is represented with the CCP vector. The similarity of the model and the native complexes can be obtained by computing the angle between their CCP vectors. This angle, or CCD, is obtained from the vector dot product:

$$\cos(CCD) = \frac{\left(\overrightarrow{CCP_{nat}} \bullet \overrightarrow{CCP_{mdl}} \right)}{\left(\left| \overrightarrow{CCP_{nat}} \right| \times \left| \overrightarrow{CCP_{mdl}} \right| \right)}.$$

The more different the CCPs, which represent the chemical properties of the RNP interface, the greater the angle. One advantage of CCD is that each type of interacting center is grouped without consideration of where it lies on the interface. Hence, CCD can easily identify two native-like docking poses for a near-symmetric dimer, whereas only one of the models will have a low rmsd. We found that CCD improves the description on the chemical properties of the interface, whereas the use of rmsd to guide machine learning may lead to conflicting signals and inhibits the proper description of RNP interactions.

Scoring Function

By weighting the entries of a CCP, it can be used as a scoring function:

$$S = Coulomb + \overrightarrow{\omega_{ccp}} \bullet \overrightarrow{ccp},$$

where the total score is the sum of the Coulomb energy, (Charge1•Charge2)/r_{12}, plus a weighted CCP. Machine-learning methods are used to reduce the nonzero entries in ω_{ccp} to 12 from the original 300 dimensions. To evaluate CCD, we first use the full 300-dimensional CCP vector. However, a full CCP set contains too many dimensions for proper training and weighting to be used in a scoring function. We use an interaction matrix between protein (rows) and RNA (columns) moieties and group together columns to reduce the dimensionality. The CCP is brought down to 120 components by summing the interactions of both the major and minor grooves together, leading to 20 protein moieties interacting with only 6 nucleic acid moieties, the phosphate groups (regardless of nucleotide type), and the side chains (major and minor combined) of each five nucleotide types. The total number of pairs is 20× (1+5) = 120. After training, these 120 components are further reduced to 12 for the scoring tRNPs and to 6 for the scoring of other RNPs.

UV CLIP-Chip

To experimentally verify the predicted tRNPs, we performed the Clip-chip method in living mammalian cells (Figure 5A). CLIP is a widely used method that identifies the interaction between RNA and proteins through covalent bond formation upon UV irradiation and has been successfully applied to investigate RNP interactions in living cells (Ule et al., 2003; Zhang and Darnell, 2011). As described previously, we have developed tRNA microarray methods to identify and to determine the abundance of tRNAs in a cellular RNA mixture (Dittmar et al., 2006; Pavon-Eternod et al., 2009). In our study, CLIP was coupled with tRNA microarray analysis to identify the tRNAs that bind to their corresponding protein. To validate our predictions, we selected six predicted tRNA-binding proteins: FTase (1D8D), PEPCK (1KHB), EHMT (2IGQ), MEK (1S9I), GSS (2HGS), and SAR (2GAO). We also selected three proteins that are predicted not to bind tRNA: RWD (2EBM), CAL (2VAF), and SCO (1WP0).

In addition, we applied one positive control using antibody against EF1α, which is known to bind all elongator tRNAs, and two negative controls using normal IgG and GFP antibody to preclude nonspecific binding in our experiments.

The CLIP protocol was adapted from published studies with minor modifications (Ule et al., 2005). Typically, HEK293T cells were first grown in a 10 cm dish until ~80% confluency (~8×10^6 cells). Cells were placed in a Stratalinker on ice and irradiated once with 400 mJ/cm^2 at 254 nm and harvested with a cell scraper. Pellets of crosslinked cells were resuspended in 0.5 ml lysis buffer (1× PBS, 0.1% SDS, 1% Nonidet P-40, 0.5% sodium deoxycholate) with 400 U/ml RNase inhibitor (New England BioLabs) and freshly prepared protease inhibitor cocktail (Santa Cruz Biotechnology). Cell lysate was centrifuged at 17,000 × *g* for 30 min at 4°C after incubation on ice for 2 hr. The supernatant was precleared upon adding 20 μl Dynabeads protein A beads (Life Technologies) and incubation for 1 hr at 4°C. The supernatant was spun again at 17,000 × *g* for 10 min at 4°C and transferred to a fresh tube.

To prepare antibody-conjugated beads, 50 μl protein A beads in a fresh microtube were washed twice with 1 ml lysis buffer, then resuspended in 200 μl lysis buffer. A total of 4 μg of each antibody (FTase, PEPCK, EHMT, MEK, GSS, SAR, RWD, CAL, SCO, and GFP antibodies are from Santa Cruz Biotechnology; EF1α and rabbit IgG are from Cell Signaling Technology) was added to each bead batch. The mixture was rotated for 4 hr at room temperature and then washed three times with 1 ml lysis buffer. The lysis buffer was removed and the supernatant from above added to each of the antibody-conjugated beads. The mixture was rotated at 4°C overnight, and the supernatant was then discarded. Beads were washed 3× with 1 ml high-salt buffer (5× PBS, 0.1% SDS, 1% Nonidet P-40, 0.5% sodium deoxycholate) and 3× with 1 ml wash buffer (20 mM Tris-HCl [pH 7.4], 10 mM $MgCl_2$, 0.2% Tween 20).

After immunoprecipitation, beads were resuspended in 200 μl RNA elution buffer (100 mM Tris-HCl [pH 7.4], 10 mM EDTA, 1% SDS) containing 2 mg/ml Proteinase K (Ambion). Antibody-bound RNAs were released from the beads upon incubation at 50°C for 30 min. The Proteinase K was then removed upon extraction with 200 μl phenol/chloroform, and the RNA was recovered by ethanol precipitation. The RNA was 3′ ^{32}P labeled using [^{32}P] pCp and T4 RNA ligase (England et al., 1980). The ^{32}P-labeled mixture was directly analyzed on 10% denaturing PAGE containing 7 M urea using purified ^{32}P-labeled yeast tRNA^{Phe} as size control (Figures 5B–5D).

To analyze the ^{32}P-labeled RNA by tRNA microarray (Figures 5E and 5F), the corresponding tRNA sized bands were cut out of the gel and eluted with crush and soak buffer (50 mM KOAc/200 mM KCl [pH 7.0]) at 4°C overnight. The eluted RNA was recovered by ethanol precipitation and dissolved in water. tRNA microarray preparation, hybridization, and data analysis were performed according to methods described previously by Netzer et al. (2009) and Pavon-Eternod et al. (2009).

SUPPLEMENTAL INFORMATION

Supplemental Information includes an Extended Discussion, Extended Experimental Procedures, six figures, and twelve tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2013.04.010.

LICENSING INFORMATION

ACKNOWLEDGMENTS

This work was supported by an NIH grant (GM57880 to T.R.S. and T.P.), the NIH-supported computing resources of the "Beagle" Cray XE6 system (S10 RR029030-01), the NSF-supported ExTENCI project (OCI-1007115), and the computing resources of the Open Science Grid. M.P. was a Chicago Fellow of the University of Chicago and is a Natural Sciences and Engineering Research Council of Canada postdoctoral fellow. Computations were performed on the Godzilla, iBi, and the Beagle clusters at the University of Chicago. We thank B. Busby for computing assistance. We also thank Drs. Xiao-jing Yang and Karl Freed for stimulating discussions. A web server for predicting additional tRNA-protein complexes is available at http://godzilla.uchicago.edu/pages/duck-na/.

Received: June 12, 2012
Revised: March 4, 2013
Accepted: April 12, 2013
Published: May 9, 2013

REFERENCES

Agostini, F., Cirillo, D., Bolognesi, B., and Tartaglia, G.G. (2013). X-inactivation: quantitative predictions of protein interactions in the Xist network. Nucleic Acids Res. *41*, e31.

Ahmad, S., and Sarai, A. (2011). Analysis of electric moments of RNA-binding proteins: implications for mechanism and prediction. BMC Struct. Biol. *11*, 8.

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S., and Tuschl, T. (2012). Identification of RNA-protein interaction networks using PAR-CLIP. Wiley Interdiscip. Rev. RNA *3*, 159–177.

Attwood, T.K. (2002). The PRINTS database: a resource for identification of protein families. Brief. Bioinform. *3*, 252–263.

Bahadur, R.P., Zacharias, M., and Janin, J. (2008). Dissecting protein-RNA recognition sites. Nucleic Acids Res. *36*, 2705–2716.

Baltz, A.G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., et al. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol. Cell *46*, 674–690.

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G.G. (2011). Predicting protein associations with long noncoding RNAs. Nat. Methods *8*, 444–445.

Berman, H., Henrick, K., Nakamura, H., and Markley, J.L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. *35*(Databaseissue), D301–D303.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., et al. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell *149*, 1393–1406.

Chan, P.P., and Lowe, T.M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res. *37*(Databaseissue), D93–D97.

Chen, Y.C., and Lim, C. (2008). Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. Nucleic Acids Res. *36*, e29.

Chruszcz, M., Domagalski, M., Osinski, T., Wlodawer, A., and Minor, W. (2010). Unmet challenges of structural genomics. Curr. Opin. Struct. Biol. *20*, 587–597.

Cruz, J.A., Blanchet, M.F., Boniecki, M., Bujnicki, J.M., Chen, S.J., Cao, S., Das, R., Ding, F., Dokholyan, N.V., Flores, S.C., et al. (2012). RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. RNA *18*, 610–625.

Dittmar, K.A., Goodenbour, J.M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. PLoS Genet. *2*, e221.

Draper, D.E. (1999). Themes in RNA-protein recognition. J. Mol. Biol. *293*, 255–270.

Ellis, J.J., and Jones, S. (2008). Evaluating conformational changes in protein structures binding RNA. Proteins *70*, 1518–1526.

England, T.E., Bruce, A.G., and Uhlenbeck, O.C. (1980). Specific labeling of 3′ termini of RNA with T4 RNA ligase. Methods Enzymol. *65*, 65–74.

Fulle, S., and Gohlke, H. (2010). Molecular recognition of RNA: challenges for modelling interactions and plasticity. J. Mol. Recognit. *23*, 220–231.

Gabb, H.A., Jackson, R.M., and Sternberg, M.J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. J. Mol. Biol. *272*, 106–120.

Goodarzi, H., Najafabadi, H.S., Oikonomou, P., Greco, T.M., Fish, L., Salavati, R., Cristea, I.M., and Tavazoie, S. (2012). Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature *485*, 264–268.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.

Hinnebusch, A.G., and Natarajan, K. (2002). Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. Eukaryot. Cell *1*, 22–32.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., et al. (2009). InterPro: the integrative protein signature database. Nucleic Acids Res. *37*(Databaseissue), D211–D215.

Kim, O.T., Yura, K., and Go, N. (2006). Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. Nucleic Acids Res. *34*, 6450–6460.

Kondo, J., and Westhof, E. (2011). Classification of pseudo pairs between nucleotide bases and amino acids by analysis of nucleotide-protein complexes. Nucleic Acids Res. *39*, 8628–8637.

Li, C.H., Cao, L.B., Su, J.G., Yang, Y.X., and Wang, C.X. (2012a). A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. Proteins *80*, 14–24.

Li, M., Kao, E., Gao, X., Sandig, H., Limmer, K., Pavon-Eternod, M., Jones, T.E., Landry, S., Pan, T., Weitzman, M.D., and David, M. (2012b). Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. Nature *491*, 125–128.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol. *8*, 479–490.

Motorin, Y., and Helm, M. (2010). tRNA stabilization by modified nucleotides. Biochemistry *49*, 4934–4944.

Moult, J., Fidelis, K., Kryshtafovych, A., and Tramontano, A. (2011). Critical assessment of methods of protein structure prediction (CASP)—round IX. Proteins *79*(*Suppl 10*), 1–5.

Netzer, N., Goodenbour, J.M., David, A., Dittmar, K.A., Jones, R.B., Schneider, J.R., Boone, D., Eves, E.M., Rosner, M.R., Gibbs, J.S., et al. (2009). Innate immune and chemically triggered oxidative stress modifies translational fidelity. Nature *462*, 522–526.

Novoa, E.M., Pavon-Eternod, M., Pan, T., and Ribas de Pouplana, L. (2012). A role for tRNA modifications in genome structure and codon usage. Cell *149*, 202–213.

Parisien, M., Freed, K.F., and Sosnick, T.R. (2012). On docking, scoring and assessing protein-DNA complexes in a rigid-body framework. PLoS One *7*, e32647.

Pavon-Eternod, M., Gomes, S., Geslain, R., Dai, Q., Rosner, M.R., and Pan, T. (2009). tRNA over-expression in breast cancer and functional consequences. Nucleic Acids Res. *37*, 7268–7280.

Pérez-Cano, L., and Fernández-Recio, J. (2010). Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. Proteins *78*, 25–35.

Polozov, R.V., Montrel, M., Ivanov, V.V., Melnikov, Y., and Sivozhelezov, V.S. (2006). Transfer RNAs: electrostatic patterns and an early stage of recognition by synthetases and elongation factor EF-Tu. Biochemistry *45*, 4481–4490.

Pons, C., Solernou, A., Perez-Cano, L., Grosdidier, S., and Fernandez-Recio, J. (2010). Optimization of pyDock for the new CAPRI challenges: docking of homology-based models, domain-domain assembly and protein-RNA binding. Proteins *78*, 3182–3188.

Puton, T., Kozlowski, L., Tuszynska, I., Rother, K., and Bujnicki, J.M. (2012). Computational methods for prediction of protein-RNA interactions. J. Struct. Biol. *179*, 261–268.

Romero, E., and Sopena, J.M. (2008). Performing feature selection with multilayer perceptrons. IEEE Trans. Neural Netw. *19*, 431–441.

Scheibe, M., Butter, F., Hafner, M., Tuschl, T., and Mann, M. (2012). Quantitative mass spectrometry and PAR-CLIP to identify RNA-protein interactions. Nucleic Acids Res. *40*, 9897–9902.

Serganov, A., and Patel, D.J. (2008). Towards deciphering the principles underlying an mRNA recognition code. Curr. Opin. Struct. Biol. *18*, 120–129.

Setny, P., and Zacharias, M. (2011). A coarse-grained force field for Protein-RNA docking. Nucleic Acids Res. *39*, 9118–9129.

Shajani, Z., Sykes, M.T., and Williamson, J.R. (2011). Assembly of bacterial ribosomes. Annu. Rev. Biochem. *80*, 501–526.

Shazman, S., and Mandel-Gutfreund, Y. (2008). Classifying RNA-binding proteins based on electrostatic properties. PLoS Comput. Biol. *4*, e1000146.

Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H.J. (2008). Prediction of interacting single-stranded RNA bases by protein-binding patterns. J. Mol. Biol. *379*, 299–316.

Sternberg, M.J., Gabb, H.A., Jackson, R.M., and Moont, G. (2000). Protein-protein docking. Generation and filtering of complexes. Methods Mol. Biol. *143*, 399–415.

Tuszynska, I., and Bujnicki, J.M. (2011). DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. BMC Bioinformatics *12*, 348.

Tworowski, D., and Safro, M. (2003). The long-range electrostatic interactions control tRNA-aminoacyl-tRNA synthetase complex formation. Protein Sci. *12*, 1247–1251.

Tworowski, D., Feldman, A.V., and Safro, M.G. (2005). Electrostatic potential of aminoacyl-tRNA synthetase navigates tRNA on its pathway to the binding site. J. Mol. Biol. *350*, 866–882.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. *28*, 1248–1250.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.

Ule, J., Jensen, K., Mele, A., and Darnell, R.B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods *37*, 376–386.

Wilde, M., Hategan, M., Wozniak, J.M., Clifford, B., Katz, D.S., and Foster, I. (2011). Swift: a language for distributed parallel scripting. Parallel Comput. *37*, 633–652.

Zhang, C., and Darnell, R.B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat. Biotechnol. *29*, 607–614.

Zhang, C., Frias, M.A., Mele, A., Ruggiu, M., Eom, T., Marney, C.B., Wang, H., Licatalosi, D.D., Fak, J.J., and Darnell, R.B. (2010). Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. Science *329*, 439–443.

Zhao, H., Yang, Y., and Zhou, Y. (2011). Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucleic Acids Res. *39*, 3017–3025.

Zheng, S., Robertson, T.A., and Varani, G. (2007). A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. FEBS J. *274*, 6378–6391.