Similarity of Protein-RNA Interfaces Based on Motif Analysis

Brian T. Sutch, Eric J. Chambers, Melina Z. Bayramyan, Timothy K. Gallaher, and Ian S. Haworth*

Department of Pharmacology & Pharmaceutical Sciences, University of Southern California, Los Angeles, California 90089-9121

Received May 1, 2009

We have developed a method for determination of the similarity of pairs of protein-RNA complexes, which we refer to as SIMA (Similarity by Identity and Motif Alignment). The key element in the SIMA method is the description of the protein-RNA interface in terms of motifs (salt bridges, aromatic stacking interactions, nonaromatic stacks, hydrophobic interactions, and hydrogen-bonded motifs), in addition to single hydrogen bonds and van der Waals contacts. Based on a pairwise scoring function combining motif alignment with identity of the protein and RNA sequences, we define a SIMA score for any pair of protein-RNA complexes. A positive score indicates similarity between the complexes. We used the SIMA method to identify 284 nonredundant binary protein-RNA complexes out of 776 such complexes in 382 nonribosomal protein-RNA structure files obtained from the RCSB database. SIMA allows rapid and quantitative comparison of protein-RNA interfaces and may be useful for interface classification with potential functional and evolutionary implications.

BACKGROUND

Protein-RNA complexes play a central role in many biological processes, including translation,[1] mRNA stabilization,[2] ribosome formation,[3] and viral packaging.[4] Over the last ten years, the structures of many of these complexes have been solved by X-ray crystallography or NMR spectroscopy. Biophysical approaches have also suggested basic principles of protein-RNA recognition: complexes often form with large conformational changes in the protein,[5,6] and the binding region of the RNA molecule is often single-stranded[7] and may be unstructured before binding. Kinetically, protein-RNA association is commonly driven by long-range electrostatic forces[8] followed by formation of hydrogen bonds and hydrophobic interactions, and complexes with biological function typically have binding constants in the nanomolar range.[9]

Statistical analyses of protein-RNA interactions have been reported by several groups.[10−18] In an early analysis of 32 protein-RNA complexes, Jones et al.[10] found a roughly even distribution of hydrogen bond contacts of the protein with the base and backbone of RNA, more van der Waals contacts than hydrogen bonds at the interface. a high frequency of protein contacts with guanine and uracil, and frequent RNA contacts with arginine and aromatic amino acids. Allers and Shamoo[11] developed and used the ENTANGLE algorithm in an analysis of a larger set of 45 protein-RNA structures, with findings of frequent hydrogen bonds to the protein backbone, propensities for RNA base stacking with aromatic amino acids, and evidence that different categories of protein-RNA complexes (particularly ribosomal and nonribosomal) may be characterized by different trends in atomic level interactions. Using a similar sized data set, Treger and Westhof[12] found evidence for an increased number of Ser-A, Gly-G, and Asn-U interactions and also pointed out the key role of water at the protein-RNA interface. Ser-A and Asn-U interactions were also commonly found in the data set examined by Jeong et al.[13]

Lejeune et al.[14] derived an order of frequency of amino acid-nucleotide interactions of all types based on an analysis of 49 protein-RNA complexes, which were selected using a computational approach to eliminate homologous complexes based on the protein sequence. Morozova et al.[15] utilized a technique of superposition of bases with the corresponding binding pocket to show that protein-RNA recognition exploits a variety of potential conformations of ssRNA to achieve specificity through glove-like binding pockets capable of discriminating between different bases with as few as two contacts. The increasing number of available structures has allowed characterization of recognition sites for RNA classes in more recent studies.[16,17] Ellis et al.[16] found a preference of van der Waals interactions over hydrogen bonding and a prevalence of interactions through the RNA backbone in a data set of 89 complexes, with the key observation that the distribution of contacts is based on the functional class of the bound RNA. Bahadur et al.[17] proposed an average protein-RNA interface containing 20 hydrogen bonds (including 7 to phosphate, 5 to the sugar 2′-OH, and 6 to the bases) and 32 water molecules. Electrostatic and hydrogen bonding interactions are also found in incidental contacts based on an analysis of crystal packing of protein-RNA complexes.[18]

There are currently more than 700 entries for protein-RNA complexes in the RCSB protein data bank.[19,20] This growing number of structures increases the difficulty of extraction of a set of representative complexes for structural and geometrical analysis with avoidance of duplication through inclusion of similar protein-RNA interfaces. Therefore, assessment of the structural similarity of two protein-RNA

* Corresponding author e-mail: ihaworth@usc.edu.

10.1021/ci900154a CCC: $40.75
Published on Web 09/04/2009

Table 1. Definition of Protein-RNA Contact Motifs

motif	protein component	RNA component	required contacts[a]
aromatic stack	F,W,Y: ring atoms in side chain	bases	≥12
nonaromatic stack	D, E, H, N, Q, R: sp^2 side chain atoms	bases	≥12
hydrophobic	A, I, L, M, V: all side chain atoms	bases	≥3[b]
salt bridge	K, R: amine or guanidinium atoms	phosphodiester: P, O1P, and O2P	1
double H-bond	oxygen and nitrogen atoms	bases and backbones	2

[a] A contact is defined as an atom−atom distance of <4.0 Å. [b] Also $d_{P\text{-}R}$ < 4.5 Å, where $d_{P\text{-}R}$ is the distance between a defined atom in each amino acid atom and the center of the base.

interfaces remains as an important challenge. Zhou et al.[21] have recently described an approach to this problem based on alignment of RNA nucleotides with interactions with the protein partner in combination with SCOP protein classification. Importantly, this approach provided the first derivation of a numerical measure of protein-RNA interface similarity.[21] In the current work, we have developed a complementary approach, which we refer to as SIMA (Similarity by Identity and Motif Alignment), for quantitative comparison of protein-RNA interfaces based on structural motifs that form the interface.

METHODS

Definition of Protein-RNA Motifs and Contacts. To define a protein-RNA interface, we first determine all atom−atom contacts of <4.0 Å and compute the number of residue−residue (amino acid-nucleotide) pairs based on the presence of at least one atom−atom contact for each pair. This analysis is performed in our PRORNA algorithm (Sutch et al., to be submitted). All contacts are based on distances between non-hydrogen atoms, since most available structures do not contain protons. The contact(s) for each pair are then examined to determine if the pair forms a motif, based on the criteria in Table 1. Stacking interactions involving bases with aromatic amino acid side chains (phenylalanine, tryptophan, and tyrosine) and nonaromatic amino acid side chains (aspartic acid, glutamic acid, histidine, asparagine, glutamine, and arginine) are detected based on a count of >12 contacts between the designated atoms (the criterion was developed using preliminary trials to determine the appropriate cutoff). Hydrophobic contacts are defined by >3 contacts between bases and hydrophobic amino acid side chains, with the additional requirement of a distance of <4.5 Å between a defined atom in each amino acid and the center of the base. Salt bridges are defined by a single contact between the lysine NZ atom or any atom of the arginine guanidinium group (NE, CZ, NH1, and NH2) with any atom in the phosphodiester linkage (P, OP1, and OP2). Hydrogen bonds are defined as a contact between defined oxygen and nitrogen atoms where at least one atom can carry a proton (i.e., contacts between two carbonyl oxygens, for example, would be excluded). Two such hydrogen bonds within an amino acid-nucleotide pair result in the definition of a double H-bond motif.

Each pair that meets the criteria for a given motif is designated as such, and all the contacts involved in the motif are eliminated from consideration as individual contacts. The remaining contacts (which may include a hydrogen bond or a double H-bond motif) are considered to be secondary contacts that arise due to formation of the motif but are not directly involved in the motif. The remaining, nonmotif pairs are then examined to determine if a single hydrogen bond is formed between the pair, and, if so, the pair is defined as a H-bonded pair. The remaining contacts in such pairs are then considered to be secondary contacts that arise due to formation of the hydrogen bond. Finally, all pairs that have neither motifs nor hydrogen bonds are defined as vdW pairs.

Similarity by Identity and Motif Alignment (SIMA). To determine a nonredundant set of protein-RNA complexes for an unbiased analysis of the frequency of motif contacts and geometrical data, we developed an alignment method using a scoring system that simultaneously reflects RNA sequence identity and protein-RNA contacts: Similarity by Identity and Motif Alignment (SIMA). This method is implemented in PRORNA, with the program running in the alignment mode. The program accepts input from a FASTA format file of sequences of protein-RNA complexes downloaded from the RCSB protein data bank.[19] Using the FASTA file and the corresponding PDB structure file, PRORNA was used to perform a motif-based analysis of every binary combination of protein and RNA chains within the structure file. PDB structures with multiple protein and RNA chains give many protein-RNA binary combinations in which the chains have no interaction and these were eliminated. Protein-RNA interfaces with <5 amino acid-nucleotide pairwise interactions were also eliminated at this stage since these were judged to be incidental contacts arising from crystal packing.

The binary protein-RNA complexes that remained after the initial processing step were then compared using the SIMA procedure. For each complex, the "protein-interacting region" of the RNA (the interface) was defined from the first nucleotide to the last nucleotide with an interaction with the protein, with addition of up to 5 nucleotides at each end. Interface pairs were then compared, with the limitation that the comparison was made over a length ≥75% of the longer interface in each pair. This length requirement was determined in preliminary runs with variation of this criterion.

For pairs of interfaces meeting this criterion, the primary motif contact or other interaction for each nucleotide in the two chains was compared at each RNA position, using simultaneous evaluation of the sequence identity and motif/interaction contact. For a position with a matching nucleotide, matching motifs or matching interactions receive a positive score (Table 2), a matching noncontact receives a score of zero, and any mismatched contact (for example, a cytosine with a single hydrogen bond to an amino acid versus a cytosine at the same position in the other interface that interacts with the protein through a double hydrogen bond motif) receives a score of −3. These scores are then modulated based on a match (+3) or mismatch (−3) between the interacting amino acid at the particular position in the two interfaces (Table 2). A position in the interface with a mismatched nucleotide is scored as −5 points. Salt bridges are scored independently of the sequence, with a matching salt bridge scored as +10 and the presence of a salt bridge

Table 2. SIMA Scoring System for Assessment of Similarity of Protein-RNA Interfaces

motif/interaction	code[a]	SIMA score
aromatic stack	S	10
nonaromatic stack	N	10
hydrophobic	H	10
double H-bond	2	10
hydrogen bond	1	5
vdW contact	c	2
clash	x	2
no contact		0
different motifs or interactions		−3
same amino acids		3
different amino acids		−3
different nucleotides		−5
salt bridge agreement[b]		10
salt bridge disagreement[b]		−3

[a] Single-letter code used to define the motif in the SIMA procedure (see Figures 1 and 2). [b] Salt bridge scores are independent of the nucleotide identity.

in one complex and no salt bridge in the second complex is scored as −3.

The component scores are summed for each possible comparison between two interfaces and the highest score is regarded as the SIMA score for the two interfaces. A positive SIMA score indicates that the complexes are "similar". The parameters for the SIMA analysis were determined in preliminary runs with variation of the scoring function. We note that we use the sum of the SIMA scores, rather than the average over the number of interface nucleotides, as the total score for each pairwise interface comparison. Currently, we view the algorithm as capable only of giving a binary result: i.e., interfaces are similar (positive SIMA score) or not similar (negative SIMA score). Therefore, only the sign of the SIMA score is important. The absolute score may indicate greater or lesser similarity, but calibration of the degree of similarity will require further exploration of the scoring system. Examples of the results of the SIMA procedure are given below.

RESULTS

Determination of a Nonredundant Set of Protein-RNA Complexes. Determination of a nonredundant set of protein-RNA interfaces is important for calculation of statistically unbiased geometrical and frequency data for protein-RNA contacts. A search of the RCSB database using the advanced search facility[20] with "Molecule/chain ID" set to "Protein = Yes, RNA = Yes, DNA = No" returned a total of 709 protein-RNA PDB structure files (search date: 2009-2-1). Using "Advanced/Keyword" set to "ribosom*" and "Full Text", the files were separated into 314 ribosomal and 395 nonribosomal files. Of the nonribosomal files, 13 were eliminated due to the presence of many short RNA chains, leaving 382 nonribosomal files for further analysis.

The sequences of the protein and RNA chains of the 382 PDB files were downloaded from the RCSB in FASTA format and used directly as input into PRORNA with the program running in the SIMA alignment mode. The 382 PDB files were analyzed as 1746 binary protein-RNA combinations; for example, a PDB file with 2 protein chains and 2 RNA chains gives 4 binary protein-RNA combinations. Of the 1746 combinations, exclusion criteria of no contact between the protein and RNA chains and <5 pairwise (amino acid-nucleotide) interactions between the chains led to elimination of 876 and 94 combinations, respectively (Table 3).

The remaining 776 combinations (binary protein-RNA complexes) were subjected to SIMA analysis. This resulted in identification of 284 unique protein-RNA interfaces and elimination of 492 redundant interfaces due to similarity with one of the 284 interfaces (Supporting Information Table 1). SIMA examines interfaces in the order in which it receives them, via an alphabetically sorted (FASTA format) input file. Interfaces are either eliminated because they are similar to a previously examined interface or are accepted as unique. The 284 nonredundant interfaces originated from 188 PDB files. For example, PDBID 1DI2 appears twice in Supporting Information Table 1 as a complex of protein chain B with RNA chains D and E. In the SIMA analysis, the binary protein-RNA combinations 1DI2:BD and 1DI2:BE were found to be similar to other protein-RNA combinations within the 1DI2 structure file (AD and AE, respectively) and also with combinations in a different structure file, 1RC7. Hence, 1DI2:AD, 1DI2:AE, and all 1RC7 combinations are listed as eliminated in Supporting Information Table 1 due to similarity with the 1DI2:BD and 1DI2:BE entries.

Examples of Complex Similarity. To illustrate the power of the SIMA method to identify protein-RNA interface similarity we describe two examples in detail. In the first (Figure 1), the interface of protein chain A and RNA chain B within 1CX0.pdb (1CX0:AB) is compared to the interface of 1M5V:FE. SIMA detects the two interfaces along the RNA chains and then scores these interfaces. A visual comparison of the complexes does not immediately indicate similarity (Figure 1). However, the SIMA score is 25.0, which indicates that the interfaces are similar. The interfaces are defined for positions 38−63 (26 nucleotides) in the RNA chain of 1CX0:AB and 27−51 (25 nucleotides) in 1M5V:FE. The SIMA algorithm finds the highest similarity score for a 20-nucleotide regions spanning 3−22 of the 26-nucleotide interface of 1CX0:AB and 2−21 of the 25-nucleotide interface of 1M5V:FE. The compared regions of the interfaces are delineated by the dashed lines in Figure 1. A 20-nucleotide comparison is the minimum length that meets the length parameter of ≥75% of the longer interface. For this reason, the comparison begins with G3 in the 1CX0:AB interface and A2 in the 1M5V:FE interface. Since this is a sequence mismatch (Table 2) it receives a score of −5 points (first number in the row of scores in Figure 1). The scores for the remaining 19 positions can be understood with reference to Table 2. For example, position 4 is an A in both interfaces, but in 1CX0:AB this A forms a hydrogen bond (code 1) to a lysine (K), whereas the A at position 4 in 1M5V:FE does not interact with the protein; therefore, there is a difference in the interaction (hydrogen bond vs nothing, −3 points) and in the interacting amino acid (lysine vs nothing, also −3 points), giving a total score for position 4 of −6 points.

The second example (Figure 2) demonstrates the ability of SIMA to compare large interfaces with complicated multiple protein-RNA contacts. In this case the two com-

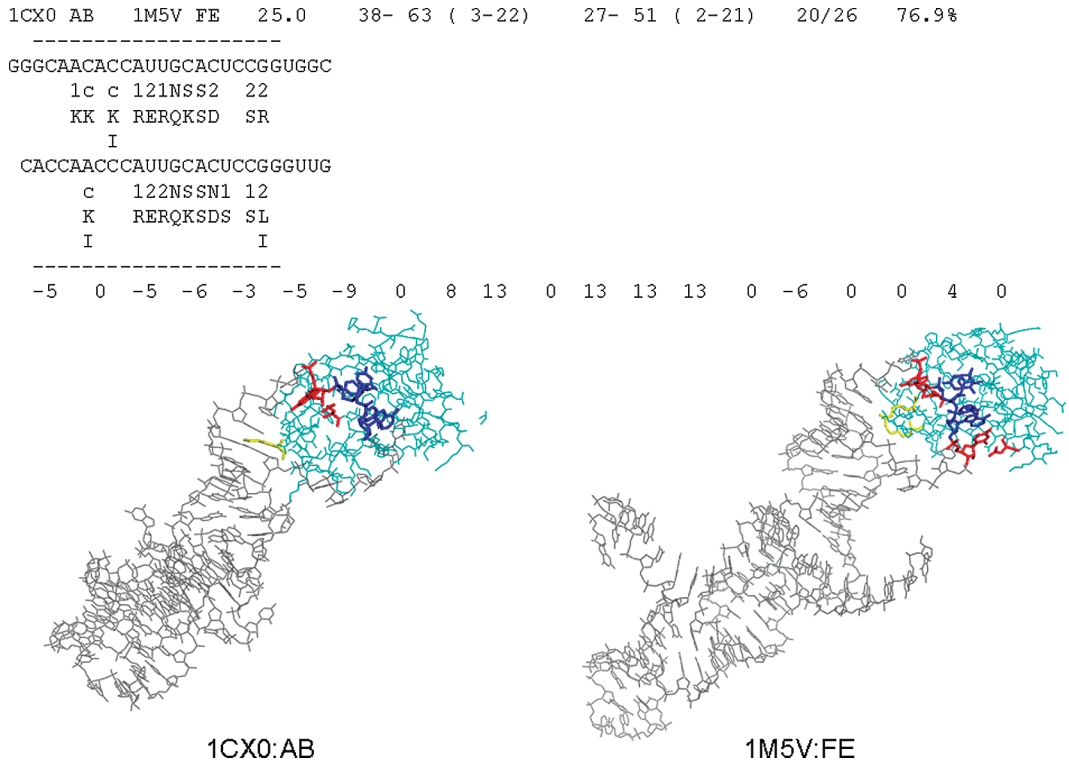

Figure 1. SIMA comparison of the interfaces between chains A (protein) and B (RNA) of 1CX0 (1CX0:AB) and chains F and E of 1M5V (1M5V:FE). The 1CX0:AB interface spans RNA residues 38−63 of chain B and the 1M5V:FE interface spans residues 27−51 of chain E. In the SIMA analysis the two interfaces are aligned, and each position in a particular alignment is evaluated based on scoring the system in Table 2. The highest SIMA score (25.0) was obtained for a comparison of positions 3−22 in the 1CX0:AB interface (bases 41−60) and 2−21 in the 1M5V:FE interface (bases 28−47). The comparison is performed over 20 bases out of a total of 26 in the longer interface (76.9%, which meets the length parameter requirement of >75%; see text). The dashed lines delineate the region of comparison and the interaction (motifs, see codes in Table 3) and interacting nucleotide and amino acid are given in single-letter codes for each interface. The scores for each position in the favored alignment are given below the alignment. In the images of the two complexes, the protein is shown in light blue and the RNA in gray. Aromatic stacking interactions are shown in blue, nonaromatic stacking interactions in red, and salt bridges in yellow. The images are from PYMOL based on output from the SIMA analysis.

plexes are from the same PDB file, 1J2B, which has two protein chains (A and B) and two RNA chains (C and D). The RNA chains differ in length and in sequence at the 5′ end. The 1J2B:AC and 1J2B:BD protein-RNA interfaces are very different at the 5′ end of the RNA strands and show subtle differences along the length of the interface. SIMA is able to align the interfaces and identify the common interactions in the alignment over a 57-nucleotide region using the scoring system described above. The similarity of these interfaces is particularly dependent on common salt bridge interactions and less on RNA sequence-specific motifs. The 1J2B file contains two other protein-RNA interfaces, 1J2B:AD and 1J2B:BC, which are not found to be similar to each other or to the 1J2B:AC and 1J2B:BD interfaces, and hence three of the four interfaces within the 1J2B file are retained as nonredundant (Supporting Information Table 1).

Overview of Similarity for Multiple Interfaces. The pairwise comparisons illustrated above can be used to view similarity across multiple interfaces. The SIMA output includes a simple text file that can be used as input into the GCLUTO program,[21] which allows visualization and clustering of the SIMA data. In Supporting Information Figure 1 we show the 776 × 776 comparison for all interfaces examined in this work. These data are more easily viewed electronically, and the GCLUTO interface permits zooming in on particular areas of interest. To illustrate this, we focus on two regions of the full data set in Figures 3 and 4, in which darker shades of red and green indicate greater similarity and greater nonsimilarity, respectively, between each interface pair.

Figure 3 shows a comparison of 33 × 33 interfaces corresponding to square A in Supporting Information Figure 1. The structures are listed in alphabetical order of PDB IDs. In the general naming convention used for RCSB files, similar structures often have similar IDs. In the SIMA comparison, this results in a greater intensity and frequency of red squares close to the diagonal. The data representation readily shows relationships within and between files: for example, the 1Q2R:AE interface is shown to be similar to the 1Q2R:CF, 1Q2S:AE, and 1Q2S:CF interfaces. "Off-diagonal" similarity is also apparent between 1O0B:AB and 1QRS:AB, which are visually similar interfaces.

Figure 4 shows an off-diagonal region comparing 87 × 74 interfaces from 1QFQ:BA through 1WWD:AB on the vertical axis and 2NUG:AC through 2V3C:DN on the horizontal axis. The presence of off-diagonal positive scores are the more important feature showing that frequently structures which are not named similarly are similar to one or several other otherwise unrelated interfaces. For example, interface 2NZ4:AP, which includes the U1A protein bound to an engineered ribozyme, is found to be similar to the original U1A/RNA complex in 1URN:BQ and to two other

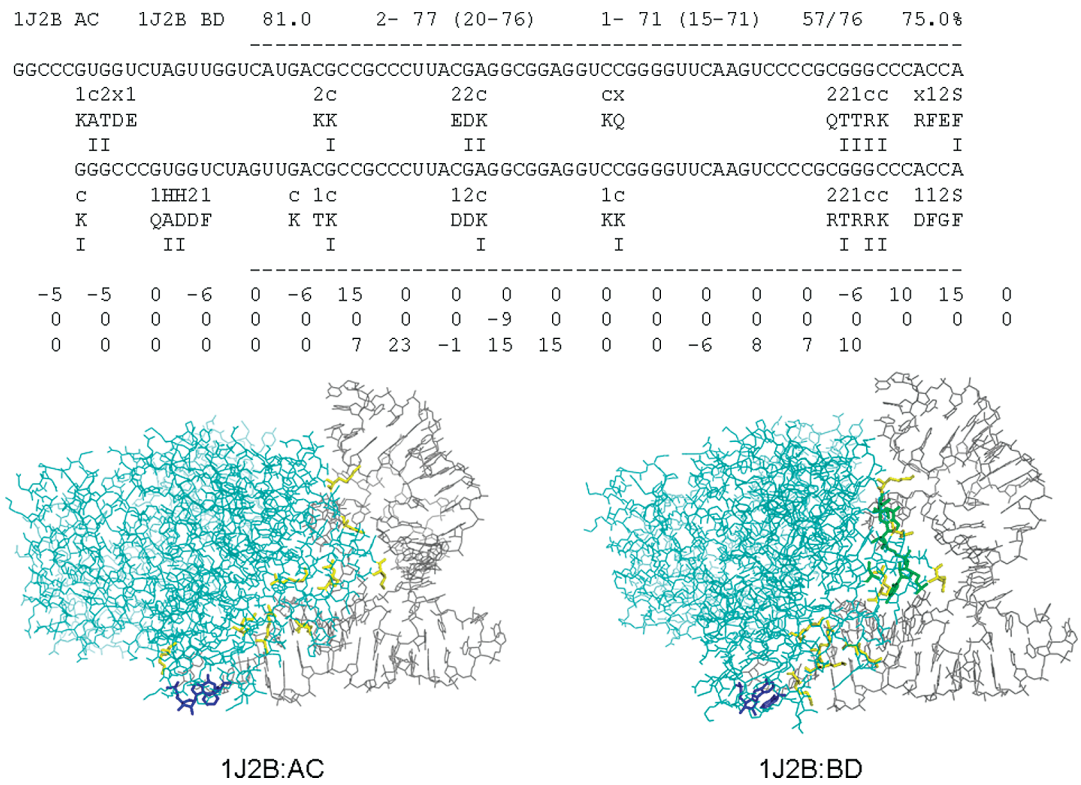

Figure 2. SIMA comparison of 1J2B:AC and 1J2B:BD interfaces. The details of the figure are as described in the legend to Figure 1. The score for the analysis is 112.0, which reflects the similarity between the two interfaces and the number of identical specific interactions. The main difference between the two interfaces is the presence of hydrophobic interactions in 1J2B:BD that are absent in 1J2B:AC. Coloring of the structures is as for Figure 1, with the addition of hydrophobic interactions shown in green.

Table 3. Summary of Protein-RNA Interface Similarity

item	N
total binary complexes	1746
eliminated: no contacts	876
eliminated: < 5 pairs	94
eliminated: similarity	492
retained	284

engineered U1A complexes with ribozymes (1SJ4:PR, 1VC0:AB). It is also apparent from Figure 4 (and Supporting Information Figure 1) that many other pairwise comparisons with similarity or near-similarity (a slightly negative SIMA score) are present in the data. Space precludes further analysis of these data in the current work, but we plan to examine these results in greater detail using the clustering methods in gCLUTO.

DISCUSSION

The basic characteristics of protein-RNA recognition have been established,[10−18] but a statistically robust geometrical analysis has been limited by the problem of defining a nonredundant set of protein-RNA complexes. This choice does not necessarily influence the broader conclusions drawn from earlier studies, but elimination of redundancy is important for compilation of a database that defines the probability of a particular interaction. Furthermore, the definition of similarity from study to study may vary, and a consistent and automated method for defining the similarity of pairs of protein-RNA interfaces is of value. The SIMA method addresses these issues using a motif-based description that provides a systematic breakdown of the interface beyond a relatively simple model based on hydrogen bonding and van der Waals contacts. This description is used in combination with the RNA and protein sequence identities to define the level of similarity (the SIMA score) between the two interfaces.

Methods based on a combination of geometrical information and sequence alignment for comparison of similarity have been described for protein-protein,[23] protein-DNA,[24] and, most recently, protein-RNA[21] interactions. Although differing in detail, these algorithms all represent interface 'shape' through a geometrical description of the interacting elements (amino acids, nucleotides). In SIMA, we also adopt this kind of approach, but we reduce the geometrical interaction into a 'motif', which can be represented as a single-digit code. We are then able to perform a sequence alignment based on both the nucleotide sequence and the 'motif' sequence. In this way, we can rapidly generate an optimal alignment of two interfaces, from which we can calculate the similarity score.

One difficulty in the SIMA method, which is common to many alignment methods, is establishment of an adequate comparison of two interfaces of different size. In the SIMA method, we define a "length parameter" as the minimum length of the shorter protein-RNA interface (based on the RNA length) expressed as a percentage of the length of the longer interface. We chose 75% as the length parameter based on preliminary runs (data not shown). A value of 25% resulted in overprediction of similarity based on comparison of relatively small regions of larger interfaces, whereas at 100% structures that were apparently similar on visual examination were not detected as such by SIMA. The choice

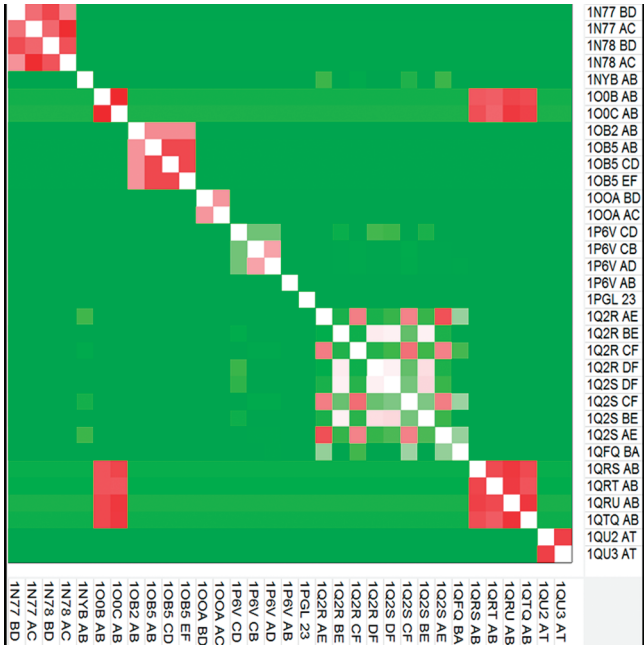

Figure 3. GCLUTO plot of a SIMA comparison of 33 × 33 interfaces from 1N77:BD through 1QU3:AT (square A in Supporting Information Figure 1). The diagonal is a zero score for comparison against the same structure. Positive scores are shown in red with increasing intensity corresponding to increasing similarity, and negative scores are shown in green with increasing intensity corresponding to increasing dissimilarity. The PDB IDs are listed in alphabetical order. The pairs of strand letters are shown as protein followed by RNA for a binary interface. The order of interfaces within a given PDB ID file is dependent on the order in the FASTA file downloaded from the RCSB.

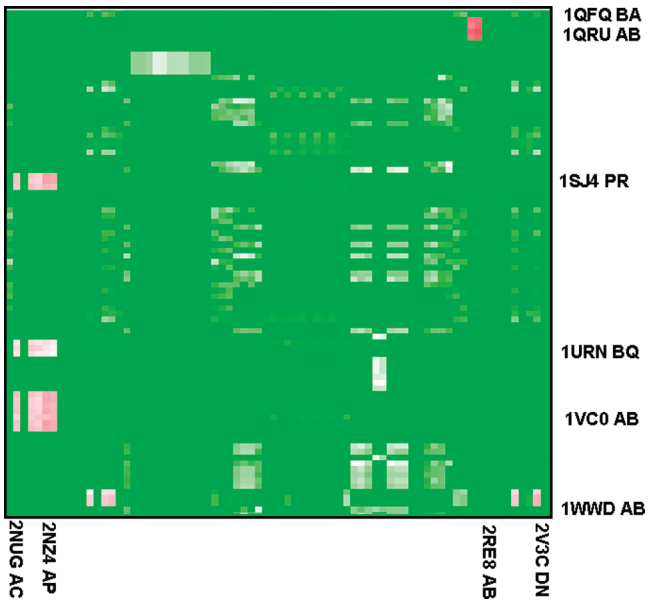

Figure 4. GCLUTO plot of a SIMA comparison of 87 × 74 interfaces from 1QFQ:BA to 1WWD:AB (vertical axis) against 2NUG:AC to 2V3C:DN (horizontal axis) (square B in Supporting Information Figure 1). The details of the figure are as described in the legend to Figure 3. This region is off-diagonal in the full 776 × 776 structure comparison, but several similar interfaces are identified. Interface 2NZ4:AP is similar to 1SJ4:PR, 1URN:BQ, and 1VC0:AB. Interface 2RE8:AB is similar to 1QRS, 1QRT, and 1QRU:AB.

of 75% seems acceptable for unrelated interfaces, but higher values might be appropriate for more closely related structures.

We envisage several applications of SIMA. Evolutionary analysis of protein-RNA complexes may become possible as the number of available structures increases. The coordinated evolution of interfaces and interface robustness could be examined via analysis of the relative similarity of interfaces. The potential to find structural similarity between distantly related interfaces or partial interfaces might open up new areas of bioinformatics of protein-RNA complexes as well as permitting homology modeling based on long-range evolutionary relationships. SIMA may also be useful in analysis of molecular dynamics simulations of protein-RNA complexes, where comparisons of changes over time within the same structure and annotation of the differences between structures are often difficult to visualize. The SIMA analysis can also be used to categorize structures into classes of protein-RNA interfaces.

A further application of SIMA will be in the detection of biologically important protein-RNA interactions. The recent findings reported by the ENCODE project[25] suggest that much is to be discovered regarding the role of RNA transcripts in the human genome. Methods for establishing the structural basis of protein-RNA association are of increasing importance, and several such algorithms have been described.[26−30] Improved descriptions of protein-RNA interfaces based on pairwise potential functions, thermodynamic analyses, and hydrogen bonding patterns[31−36] will improve the predictive accuracy, and the motif-based analysis within the SIMA approach can provide a significant contribution to this area by defining the nature of the protein and RNA surfaces that are likely to form complexes.

Finally, protein-RNA complexes also have potential from a therapeutic and diagnostic perspective. A particular area of interest is discovery of protein-targeted RNA aptamers: RNA molecules of typically a few tens of nucleotides that bind tightly to a specific protein.[37−39] Such aptamers are discovered using an evolutionary approach in the so-called SELEX method.[40,41] We are working toward a computational approach to RNA aptamer design using docking based on a database of protein-RNA interaction motifs. A specific goal in development of the SIMA approach was to establish a nonredundant set of protein-RNA interfaces for generation of a statistically robust database for use in rational *in silico* design of RNA aptamers. A geometrical and statistical analysis of the motifs in this database will be presented in a future paper.

Supporting Information Available: gCLUTO plot of the SIMA scores for an all-to-all comparison of 776 protein-RNA interfaces (Supporting Information Figure 1) and this same comparison in text format in Supporting Information Table 1, to allow identification of interfaces that were retained (left-hand column of the table) as 'unique' and those that were eliminated based on their SIMA similarity with a retained interface. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Amaral, P. P.; Dinger, M. E.; Mercer, T. R.; Mattick, J. S. The eukaryotic genome as an RNA machine. *Science* **2008**, *319* (5871), 1787–1789.

(2) Yamazaki, S.; Takeshige, K. Protein synthesis inhibitors enhance the expression of mRNAs for early inducible inflammatory genes via mRNA stabilization. *Biochim. Biophys. Acta* **2008**, *1779* (2), 108–114.
(3) Erlacher, M. D.; Polacek, N. Ribosomal catalysis: the evolution of mechanistic concepts for peptide bond formation and peptidyl-tRNA hydrolysis. *RNA Biol.* **2008**, *5* (1), 5–12.
(4) Zhu, J.; Gopinath, K.; Murali, A.; Yi, G.; Hayward, S. D.; Zhu, H.; Kao, C. RNA-binding proteins that inhibit RNA virus infection. *Proc. Natl. Acad. Sci. U.S. A.* **2007**, *1049*, 3129–3134.
(5) Williamson, J. R. Induced fit in RNA-protein recognition. *Nat. Struct. Biol.* **2000**, *7* (10), 834–837.
(6) Ellis, J. J.; Jones, S. Evaluating conformational changes in protein structures binding RNA. *Proteins* **2008**, *70* (4), 1518–1526.
(7) Shulman-Peleg, A.; Shatsky, M.; Nussinov, R.; Wolfson, H. J. Prediction of interacting single-stranded RNA bases by protein-binding patterns. *J. Mol. Biol.* **2008**, *379* (2), 299–316.
(8) Treiber, D. K.; Williamson, J. R. Beyond kinetic traps in RNA folding. *Curr. Opin. Struct. Biol.* **2001**, *11* (3), 309–314.
(9) Ji, Z. L.; Chen, X.; Zhen, C. J.; Yao, L. X.; Han, L. Y.; Yeo, W. K.; Chung, P. C.; Puy, H. S.; Tay, Y. T.; Muhammad, A.; Chen, Y. Z. KDBI: Kinetic Data of Bio-molecular Interactions database. *Nucleic Acids Res.* **2003**, *31* (1), 255–257.
(10) Jones, S.; Daley, D. T.; Luscombe, N. M.; Berman, H. M.; Thornton, J. M. Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.* **2001**, *29* (4), 943–954.
(11) Allers, J.; Shamoo, Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.* **2001**, *311* (1), 75–86.
(12) Treger, M.; Westhof, E. Statistical analysis of atomic contacts at RNA-protein interfaces. *J. Mol. Recognit.* **2001**, *14* (4), 199–214.
(13) Jeong, E.; Kim, H.; Lee, S. W.; Han, K. Discovering the interaction propensities of amino acids and nucleotides from protein-RNA complexes. *Mol. Cells* **2003**, *16* (2), 161–167.
(14) Lejeune, D.; Delsaux, N.; Charloteaux, B.; Thomas, A.; Brasseur, R. Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins* **2005**, *61* (2), 258–271.
(15) Morozova, N.; Allers, J.; Myers, J.; Shamoo, Y. Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics* **2006**, *22* (22), 2746–2752.
(16) Ellis, J. J.; Broom, M.; Jones, S. Protein-RNA interactions: structural analysis and functional classes. *Proteins* **2007**, *66* (4), 903–911.
(17) Bahadur, R. P.; Zacharias, M.; Janin, J. Dissecting protein-RNA recognition sites. *Nucleic Acids Res.* **2008**, *36* (8), 2705–2716.
(18) Phipps, K. R.; Li, H. Protein-RNA contacts at crystal packing surfaces. *Proteins* **2007**, *67* (1), 121–127.
(19) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.
(20) Deshpande, N.; Addess, K. J.; Bluhm, W. F.; Merino-Ott, J. C.; Townsend-Merino, W.; Zhang, Q.; Knezevich, C.; Xie, L.; Chen, L.; Feng, Z.; Green, R. K.; Flippen-Anderson, J. L.; Westbrook, J.; Berman, H. M.; Bourne, P. E. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.* **2005**, *33* (Database issue), D233–D237.
(21) Zhou, P.; Zou, J.; Tian, F.; Shang, Z. Geometric similarity between protein-RNA interfaces. *J. Comput. Chem.* 2009Apr 27. [Epub ahead of print].
(22) Rasmussen, M.; Karypis, G. *gCLUTO: An Interactive Clustering, Visualization, and Analysis System*; UMN-CS TR-04-021; 2004.
(23) Kim, W. K.; Henschel, A.; Winter, C.; Schroeder, M. The many faces of protein-protein interactions: a compendium of interface geometry. *PLoS Comput. Biol.* **2006**, *2*, 1151–1164.
(24) Siggers, T. W.; Silkov, A.; Honig, B. Structural alignment of protein-DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.* **2005**, *345*, 1027–1045.
(25) Birney, E.; Stamatoyannopoulos, J. A.; Dutta, A.; Guigo, R.; Gingeras, T. R.; Margulies, E. H.; Weng, Z.; Snyder, M.; Dermitzakis, E. T.; Thurman, R. E.; Kuehn, M. S.; Taylor, C. M.; Neph, S.; Koch, C. M.; Asthana, S.; Malhotra, A.; Adzhubei, I.; Greenbaum, J. A.; Andrews, R. M.; Flicek, P.; Boyle, P. J.; Cao, H.; Carter, N. P.; Clelland, G. K.; Davis, S.; Day, N.; Dhami, P.; Dillon, S. C.; Dorschner, M. O.; Fiegler, H.; Giresi, P. G.; Goldy, J.; Hawrylycz, M.; Haydock, A.; Humbert, R.; James, K. D.; Johnson, B. E.; Johnson, E. M.; Frum, T. T.; Rosenzweig, E. R.; Karnani, N.; Lee, K.; Lefebvre, G. C.; Navas, P. A.; Neri, F.; Parker, S. C.; Sabo, P. J.; Sandstrom, R.; Shafer, A.; Vetrie, D.; Weaver, M.; Wilcox, S.; Yu, M.; Collins, F. S.; Dekker, J.; Lieb, J. D.; Tullius, T. D.; Crawford, G. E.; Sunyaev, S.; Noble, W. S.; Dunham, I.; Denoeud, F.; Reymond, A.; Kapranov, P.; Rozowsky, J.; Zheng, D.; Castelo, R.; Frankish, A.; Harrow, J.; Ghosh, S.; Sandelin, A.; Hofacker, I. L.; Baertsch, R.; Keefe, D.; Dike, S.; Cheng, J.; Hirsch, H. A.; Sekinger, E. A.; Lagarde, J.; Abril, J. F.; Shahab, A.; Flamm, C.; Fried, C.; Hackermuller, J.; Hertel, J.; Lindemeyer, M.; Missal, K.; Tanzer, A.; Washietl, S.; Korbel, J.; Emanuelsson, O.; Pedersen, J. S.; Holroyd, N.; Taylor, R.; Swarbreck, D.; Matthews, N.; Dickson, M. C.; Thomas, D. J.; Weirauch, M. T.; Gilbert, J.; Drenkow, J.; Bell, I.; Zhao, X.; Srinivasan, K. G.; Sung, W. K.; Ooi, H. S.; Chiu, K. P.; Foissac, S.; Alioto, T.; Brent, M.; Pachter, L.; Tress, M. L.; Valencia, A.; Choo, S. W.; Choo, C. Y.; Ucla, C.; Manzano, C.; Wyss, C.; Cheung, E.; Clark, T. G.; Brown, J. B.; Ganesh, M.; Patel, S.; Tammana, H.; Chrast, J.; Henrichsen, C. N.; Kai, C.; Kawai, J.; Nagalakshmi, U.; Wu, J.; Lian, Z.; Lian, J.; Newburger, P.; Zhang, X.; Bickel, P.; Mattick, J. S.; Carninci, P.; Hayashizaki, Y.; Weissman, S.; Hubbard, T.; Myers, R. M.; Rogers, J.; Stadler, P. F.; Lowe, T. M.; Wei, C. L.; Ruan, Y.; Struhl, K.; Gerstein, M.; Antonarakis, S. E.; Fu, Y.; Green, E. D.; Karaoz, U.; Siepel, A.; Taylor, J.; Liefer, L. A.; Wetterstrand, K. A.; Good, P. J.; Feingold, E. A.; Guyer, M. S.; Cooper, G. M.; Asimenos, G.; Dewey, C. N.; Hou, M.; Nikolaev, S.; Montoya-Burgos, J. I.; Loytynoja, A.; Whelan, S.; Pardi, F.; Massingham, T.; Huang, H.; Zhang, N. R.; Holmes, I.; Mullikin, J. C.; Ureta-Vidal, A.; Paten, B.; Seringhaus, M.; Church, D.; Rosenbloom, K.; Kent, W. J.; Stone, E. A.; Batzoglou, S.; Goldman, N.; Hardison, R. C.; Haussler, D.; Miller, W.; Sidow, A.; Trinklein, N. D.; Zhang, Z. D.; Barrera, L.; Stuart, R.; King, D. C.; Ameur, A.; Enroth, S.; Bieda, M. C.; Kim, J.; Bhinge, A. A.; Jiang, N.; Liu, J.; Yao, F.; Vega, V. B.; Lee, C. W.; Ng, P.; Yang, A.; Moqtaderi, Z.; Zhu, Z.; Xu, X.; Squazzo, S.; Oberley, M. J.; Inman, D.; Singer, M. A.; Richmond, T. A.; Munn, K. J.; Rada-Iglesias, A.; Wallerman, O.; Komorowski, J.; Fowler, J. C.; Couttet, P.; Bruce, A. W.; Dovey, O. M.; Ellis, P. D.; Langford, C. F.; Nix, D. A.; Euskirchen, G.; Hartman, S.; Urban, A. E.; Kraus, P.; Van Calcar, S.; Heintzman, N.; Kim, T. H.; Wang, K.; Qu, C.; Hon, G.; Luna, R.; Glass, C. K.; Rosenfeld, M. G.; Aldred, S. F.; Cooper, S. J.; Halees, A.; Lin, J. M.; Shulha, H. P.; Xu, M.; Haidar, J. N.; Yu, Y.; Iyer, V. R.; Green, R. D.; Wadelius, C.; Farnham, P. J.; Ren, B.; Harte, R. A.; Hinrichs, A. S.; Trumbower, H.; Clawson, H.; Hillman-Jackson, J.; Zweig, A. S.; Smith, K.; Thakkapallayil, A.; Barber, G.; Kuhn, R. M.; Karolchik, D.; Armengol, L.; Bird, C. P.; de Bakker, P. I.; Kern, A. D.; Lopez-Bigas, N.; Martin, J. D.; Stranger, B. E.; Woodroffe, A.; Davydov, E.; Dimas, A.; Eyras, E.; Hallgrimsdottir, I. B.; Huppert, J.; Zody, M. C.; Abecasis, G. R.; Estivill, X.; Bouffard, G. G.; Guan, X.; Hansen, N. F.; Idol, J. R.; Maduro, V. V.; Maskeri, B.; McDowell, J. C.; Park, M.; Thomas, P. J.; Young, A. C.; Blakesley, R. W.; Muzny, D. M.; Sodergren, E.; Wheeler, D. A.; Worley, K. C.; Jiang, H.; Weinstock, G. M.; Gibbs, R. A.; Graves, T.; Fulton, R.; Mardis, E. R.; Wilson, R. K.; Clamp, M.; Cuff, J.; Gnerre, S.; Jaffe, D. B.; Chang, J. L.; Lindblad-Toh, K.; Lander, E. S.; Koriabine, M.; Nefedov, M.; Osoegawa, K.; Yoshinaga, Y.; Zhu, B.; de Jong, P. J. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **2007**, *447* (7146), 799–816.
(26) Han, L. Y.; Cai, C. Z.; Lo, S. L.; Chung, M. C.; Chen, Y. Z. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA* **2004**, *10* (3), 355–368.
(27) Terribilini, M.; Sander, J. D.; Lee, J. H.; Zaback, P.; Jernigan, R. L.; Honavar, V.; Dobbs, D. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W578–W584.
(28) Tong, J.; Jiang, P.; Lu, Z. H. RISP: a web-based server for prediction of RNA-binding sites in proteins. *Comput. Methods Programs Biomed.* **2008**, *90* (2), 148–153.
(29) Wang, Y.; Xue, Z.; Shen, G.; Xu, J. PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids* **2008**, *35* (2), 295–302.
(30) Han, K.; Nepal, C. PRI-Modeler: extracting RNA structural elements from PDB files of protein-RNA complexes. *FEBS Lett.* **2007**, *581* (9), 1881–1890.
(31) Chen, Y.; Kortemme, T.; Robertson, T.; Baker, D.; Varani, G. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. *Nucleic Acids Res.* **2004**, *32* (17), 5147–1562.
(32) Kim, O. T.; Yura, K.; Go, N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.* **2006**, *34* (22), 6450–6460.
(33) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. *Nucleic Acids Res.* **2009**, *37* (Database issue), D369–D373.
(34) Stolarski, R. Thermodynamics of specific protein-RNA interactions. *Acta Biochim. Pol.* **2003**, *50* (2), 297–318.

(35) Walberer, B. J.; Cheng, A. C.; Frankel, A. D. Structural diversity and isomorphism of hydrogen-bonded base interactions in nucleic acids. *J. Mol. Biol.* **2003**, *327* (4), 767–780.
(36) Cheng, A. C.; Chen, W. W.; Fuhrmann, C. N.; Frankel, A. D. Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *J. Mol. Biol.* **2003**, *327* (4), 781–796.
(37) Levy-Nissenbaum, E.; Radovic-Moreno, A. F.; Wang, A. Z.; Langer, R.; Farokhzad, O. C. Nanotechnology and aptamers: applications in drug delivery. *Trends Biotechnol.* **2008**, *26* (8), 442–449.
(38) Stoltenburg, R.; Reinemann, C.; Strehlitz, B. SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol. Eng.* **2007**, *24* (4), 381–403.
(39) Mairal, T.; Ozalp, V. C.; Lozano Sanchez, P.; Mir, M.; Katakis, I.; O'Sullivan, C. K. Aptamers: molecular tools for analytical applications. *Anal. Bioanal. Chem.* **2008**, *390* (4), 989–1007.
(40) Tuerk, C.; Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **1990**, *249* (4968), 505–510.
(41) Ellington, A. D.; Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **1990**, *346* (6287), 818–822.

CI900154A