Communications in Information and Systems
Volume 20, Number 1, 1–22, 2020

Scoring functions for protein-RNA complex structure prediction: advances, applications, and future directions

Liming Qiu and Xiaoqin Zou[*]

Protein-RNA interaction is among the most essential of biological events in living cells, being involved in protein synthesizing, RNA processing and transport, DNA transcription, and regulation of gene expression, and many other critical bio-molecular activities. A thorough understanding of this interaction is of paramount importance in fundamental study of a variety of vital cellular processes and therapeutic application for remedy of a broad range of diseases. Experimental high-resolution 3D structure determination is the primary source of knowledge for protein-RNA complexes. However, due to technical limitations, the existing techniques for experimental structure determination couldn't match the demand from fast growing interest in academia and industry. This problem necessitates the alternative high-throughput computational method for protein-RNA complex structure prediction. Similar to the in silico methods used for protein-protein and protein-DNA interactions, a reliable prediction of protein-RNA complex structure requires a scoring function with commensurate discriminatory power. Derived from determined structures and purposed to predict the to-be-determined structures, the scoring function is not only a predictive tool but also a gauge of our knowledge of protein-RNA interaction. In this review, we present an overview of the status of existing scoring functions and the scientific principle behind their constructions as well as their strengths and limitations. Finally, we will discuss about future directions of the scoring function development for protein-RNA structure prediction.

[*] Correspondence to: X. Zou, zoux@missouri.edu (email), 573-882-6045 (tel.), 573-884-4232 (fax).

1. Introduction

It has been of great interest to understand protein-RNA interactions at the molecular level due to their widespread involvement in vital cellular processes. Among these processes is the translation of a genetic code to a protein, in which specific recognition of tRNA by aminoacyl-tRNA synthetase is essential to correct gene expression. Despite the general similarity of tRNAs in their three-dimensional (3D) structures, synthetase is able to discriminate and aminoacylate their cognate tRNA based on a number of identity elements [1, 2, 3]. More surprisingly, discrimination based on a similar set of identity elements can involve completely different interaction modes [1, 4, 5, 6]. Ribosome, the molecular machine for protein synthesis in all living organisms, also highlights the significance of protein-RNA interactions. Structural and functional characterization of ribosome has been of long-standing scientific interest since its discovery in the mid-50s [7]. This highly complex structure is composed of ribosomal proteins and rRNAs. The interactions of ribosomal proteins and rRNAs are a key to the understanding of the working principle of ribosome. It had not been clear whether the ribosomal proteins function as the scaffold to facilitate the catalytic reaction by rRNAs, or vice versa, until the atomic structure of the large ribosomal subunit was solved and thereby confirmed the former conjecture [8, 9]. Besides these two prominent examples, RNA also participates in a broad range of important biological processes, such as RNA splicing [10, 11], signal transduction [12, 13], and immune response [14], through complex formation with partner proteins into ribonucleoprotein particles (RNP). Anomalies in protein-RNA interactions are implicated in numerous human diseases, including cancers, AIDS, and neurodegenerative disorders [15]. Without doubt, advances in medical treatment of these diseases would benefit from insightful knowledge on protein-RNA interactions.

Being one of the primary sources of knowledge on protein-RNA interactions, detailed structural information of protein-RNA complexes is relatively rare, considering the enormous amount of RNAs in a cell and their potential binding proteins. As of November 1, 2014, there existed only 1746 protein-RNA crystal structures in the Protein Data Bank (PDB) [16]. The number of high-resolution 3D structures of protein-RNA complexes increases slowly, because it is time consuming and laborious to resolve macromolecular structures by X-ray crystallography or NMR in general. In particular for protein-RNA complexes, the highly flexible and complicated interaction patterns of RNA further complicate the structure determination by these experimental measures. As an alternative tool, computational structure prediction has received substantial attention in recent years. A reliable computational method can effectively bridge the gap between the scarcity of resolved protein-RNA complex structures and the abundance of biological processes bearing significant scientific interests. The value of the computational methods is even more accentuated in the cases where RNA-protein interactions are intrinsically transient or exhibit a broad binding selectivity [17, 18].

Despite the paucity in the structural data, studies on statistical characterization of the available protein-RNA complex structures have uncovered consistent features. Overrepresentation of a group of residues was identified in several investigations [19, 20, 22, 23, 24], indicating the dominant role of electrostatics in protein-RNA interactions. Hydrogen-bonds were found to be important for specific interactions at the protein-RNA interface [19, 20, 25]. Base-dependent glove-like binding pockets in proteins enabling hydrogen-bonding, van der Waals and non-polar interactions were shown to be fundamental to the specific recognition of nucleotide bases [26]. Local geometry at protein-RNA interfaces was reported to be conserved over known structures [27]. That is, the joint distribution of the location and orientation of contacting residues with respect to their interface nucleotide bases peaks in a restricted region. Knowledges from these characterization studies lay the groundwork for protein-RNA structure prediction.

One of the key bottlenecks for protein-RNA structure prediction is a reliable scoring function. The purpose of a scoring function is to ascertain the likelihood of a model structure(or a pose) being the native structure. In the paradigm of computational structure prediction, first, a set of model structures is generated by a conformation search algorithm. Then, the scoring function is applied to identify the native-like structure as the structure with the lowest energy score. The performance of a scoring function is evaluated from its discriminatory power in differentiating the native structure from the decoys. As expected, no scoring function can achieve the ideal performance due to our limited knowledge on protein-RNA interactions. In fact, existing scoring functions are constructed based on various assumptions using different approaches, and have respective advantages and disadvantages. In this review, we will give an overview of the scoring functions that have been used for protein-RNA structure prediction. For the purpose of clarity, we categorize scoring functions into groups according to their commonality in methodology, and discuss one or more representatives for each group in detail about their strengths and limitations. We also present the benchmarks for training and assessing the scoring functions, and the applications of protein-RNA structure prediction. Finally, we discuss about future directions for the scoring function development in this area.

2. Propensity-based scoring functions

The outcome of several structural characterization studies [19, 20, 22, 23, 24] indicated a higher frequency for a subset of residues and bases than random occurring at protein-RNA interfaces and motivated the development of propensity-based scoring functions. Propensity is a measure of the tendency of certain chemical structures to occur at the interface in the context of protein-RNA interactions. In general, the propensity for an amino acid residue or nucleotide base of type k is defined as a ratio of the frequencies:

$$P_k = \frac{N_k^I / \sum_k N_k^I}{N_k^A / \sum_k N_k^A} \tag{1}$$

where N_k^I is the number of residues/bases of type k involved in the interface, and $\sum_k N_k^I$ is the total number of interface residues/bases. Likewise, N_k^A is the total number of residues/bases of type k, and $\sum_k N_k^A$ is the total number of residues/bases. It is worth pointing out that the definition of interfacial residues/bases is research group specific; it could be based on a universal distance range such as between 1 and 5 Å [20], or a set of interaction-type and atom-group dependent distance criteria [23]. Despite the difference, a number of studies [19, 20, 28, 23] on the propensities of protein residues and nucleotide bases found that positively charged residues, Arg and Lys, have the highest propensity at the protein-RNA interface. In contrast, the results regarding the propensities of aromatic residues and nucleotide bases do not converge, probably due to the paucity of high-resolution protein-RNA complex structures and the differences in the methods for analysis.

With a different definition for propensity, Fernández-Recio and colleagues [29] investigated residue and base propensities on a larger set of resolved protein-RNA complex. The authors put more emphasis on the surface residues or bases by defining the propensity as

$$P_k = \frac{N_k^I / \sum_k N_k^I}{N_k^S / \sum_k N_k^S} \tag{2}$$

where N_k^I and $\sum_k N_k^I$ have the same definitions given in equation (1). N_k^S and $\sum_k N_k^S$ are the number of residues/bases of type k and total number of residues/bases on the surface, respectively. The results confirmed the highest propensity for Arg and Lys at the interface. The results also showed a high propensity for His, which had not been reported in previous studies. Hydrophobic residues were found to be the least favored at the interface. Aromatic residues, with a low propensity from statistics, were found to be important in pairing with unpaired RNA bases. On the other hand, no statistically significant difference in propensity was found among RNA bases. Moreover, protein-RNA interaction, similar to that of protein-DNA interaction, was mainly through electrostatic forces other than the hydrophobic and desolvation effects that dominate protein-protein binding.

Meanwhile, a set of propensity-based statistical potentials was developed by the same group [30], in which a pairwise residue-nucleotide interface propensity is defined as follows:

$$P^I_{pq} = \frac{N^I_{pq}/\sum_{pq} N^I_{pq}}{N^S_p/\sum_p N^S_p \times N^S_q/\sum_q N^S_q} \tag{3}$$

where N^I_{pq} is the number of pairs between amino acid type p and nucleotide base type q at the interface, $\sum_{pq} N^I_{pq}$ is the total number of residue-base pairs at the interface. N^S_p and N^S_q are the number of residues of type p and number of bases of type q on the surface, respectively. $\sum_p N^S_p$ and $\sum_q N^S_q$ are the total number of residues and total number of bases on the surface, respectively. The pairing of a residue and a base is based on a cutoff distance between nearest atoms. From the inverse Boltzmann formula, a free-energy-like statistical potential is calculated as

$$\Delta G^{stat}_{pq} = -RT\ln(P^I_{pq}) \tag{4}$$

In equation (4) ΔG^{stat}_{pa} represents the energy for the pair formed by a residue of type p and a base of type q at the interface in accordance with the propensity data, R is the gas constant, and T is the absolute temperature. The sum of the energies for all pairs of residues and bases at the interface, as evaluated by equation (5), amounts to the score for a given protein-RNA complex structure:

$$\Delta G^{stat} = \sum_{pq} \Delta G^{stat}_{pq} \tag{5}$$

The dependence of this energy scoring function on the distance is considered only through the cutoff distance for the assessment of residue-base pairs in contact. As a result, the contact-based scoring function has the disadvantage of being insensitive to model structures (e.g., with identical contact pairs), but in the meanwhile has the advantage of being tolerant to minor conformational change. In respect of differential propensities for residues and bases in different secondary structures at the interface [19, 20, 28, 31], Li *et al.* incorporated secondary-structure information into their scoring function, by a modified definition of propensity for residue base pairs which separate the same residue-base pairs of different secondary structures into statistically independent terms [32].

Two medium-resolution propensity-based scoring functions, termed DARS-RNP and QUASI-RNP, in conjunction with a reduced representation of the protein and the RNA were introduced by Tuszynska and Bujnicki [33]. In this reduced representation, amino residues are represented by one to three united atoms [33, 34, 35], and for RNA two united atoms are used for the backbone and one/two atoms for pyrimidines/purines. The reduced representation is more amenable to the above two propensity-based scoring functions which take into account scoring dependencies on distances, angles, and interaction sites. The two scoring functions share a common mathematical form:

$$E = E_d + E_a + E_s + E_p \tag{6}$$

where E is the total score, E_d, E_a, E_s and E_p are the distance-dependent energy, angle-dependent energy, site-dependent energy, and steric clash penalty term, respectively, and they are equally weighted. Among these energy terms, E_d, E_a, and E_s are both calculated by the same formula:

$$\epsilon(i,j,d) = -RT \ln \frac{N_{obs}(i,j,d)}{N_{exp}(i,j,d)} \tag{7}$$

The ϵ in equation (7) is either E_d, E_a, or E_s. In the case of E_d and E_a, $N_{obs}(i,j,d)$ is the number of contacts between united atoms of type i and type j lodged in distance and angle bin d, respectively; with the implicit understanding that atom i and atom j belong to different components forming a complex. It is important to note that for E_a, the angle-dependent energy, the angle for type i and j necessarily and implicitly involve a common united atom type to form an angle, but for simplicity in presenting the theory, the vertex atom is suppressed. For the case of E_s, the parameter d represents one of the three types of edge of RNA bases that is able to form hydrogen bonds with another base, i.e., the Watson-Crick edge, the Hoogsteen edge, and the sugar edge [21]. The last energy term E_p is the penalty for steric clashes between united atoms, which disfavors pairs of united atoms closer to each other than a cutoff distance. The decoupling of the distance- and angle-dependent energy terms in equation (6) is employed because sampling for a joint distribution of the distances and angles would be difficult since samples for certain bins would be exceedingly rare. This scarce data problem is common in building knowledge-based scoring functions especially when structural information is limited. The general remedy is to adjust the bin size and/or decouple binning in a parametric space into binnings in lower parametric spaces. There is no standard bin size, and the specific choice of bin size is justified by testing. For the DARS-RNP and QUASI-RNP, the bin size is 1 Å for the distance-dependent energy and 20 degrees for the angle-dependent energy. While the calculation of $N_{obs}(i, j, d)$ from a training set is straightforward, the calculation of $N_{exp}(i, j, d)$, the expected value for atoms of type i and type j in bin d for the reference state, is complicated. The DARS-RNP and QUASI-RNP scoring functions employ different ways to calculate $N_{exp}(i, j, d)$. For QUASI-RNP, $N_{exp}(i, j, d) = X_i * X_j * N_{obs}(d)$, where X_i and X_j are the mole fractions of atom type i and type j in the training set, respectively. $N_{obs}(d)$ is the number of contacts in bin d irrespective of atom types. For DARS-RNP, $N_{exp}(i, j, d)$ is the normalized number of contacts for atom types i and j in bin d, calculated from a set of 1000 decoys generated for each complex in the training set by the GRAMM [36]. The energy terms E_s and E_p, identical in both scoring functions, take care of the probability of residues interacting with edges of nucleotides, and steric clashes, respectively. Based on the results for two bound docking set [33], the DARS-RNP showed a slightly better performance than QUASI-RNP. The finding is not unexpected, because for DARS-RNP the statistics of amino acid and nucleotide contacts were counted from a much larger training set.

The employment of reduced representation makes medium-resolution scoring functions less susceptible to conformational changes; as a result, the scoring functions are expected to have more discriminatory power in situations where complex formation induces minor conformation changes in its components. Meanwhile, as compared to the Fernández scoring functions, these scoring functions have a better spatial resolution and a better treatment of the reference state, hence a higher accuracy in differentiating near-native structure from decoys. Xiao and co-workers also developed a coarse-grained distance-dependent scoring function DECK-RP [37] that takes into account the secondary structures of protein and RNA. For DECK-RP, amino acids are classified into 7 types of interaction centers based on the dipoles and volumes of side chains, but the 4 nucleotide types are unchanged. The secondary structures of protein and RNA are also considered. Following the DSSP [38] notations, only three classes of secondary structures are included for proteins, which are denoted as X (for I, G and S), Y (for E, B, T and random coils), and Z (for H). For RNA, only paired and unpaired bases are counted. The DECK-RP virtually integrates the features of Li *et al*'s scoring function and DARS-RNP.

3. Atomic-level statistical scoring functions

Scoring functions at the atomic level provide higher discriminatory power than scoring functions at the residue level, when low-RMSD, near-native models exist among other decoys, benefitting from a higher spatial resolution of the embedded energy function. A knowledge-based scoring function at the atomic level originally constructed for protein-DNA interaction [39] was adapted by Varani and co-workers [40] for the purpose of protein-RNA complex structure prediction [40]. Recently, two other pairwise, distance-dependent knowledge-based scoring functions also at the atomic level, dRNA [47] and ITScore-PR [41], were developed with careful handling of the reference state problem. Since the introduction of statistical potentials [72, 73, 42] more than three decades ago, the issue of the reference state has been central to the construction of pairwise, distance-dependent statistical potentials [75, 76]. For macromolecules like proteins and RNAs, the connectivity between residues or bases and finite atomic volumes forbid a trivial simplification of the joint distribution over pairwise distances without loss of statistical rigorousness. However, because of the strong inter-correlation between pairwise distances, the exact functional form of the joint distribution remains a puzzle. Without an explicit justification, it is assumed distributions for individual pairs over distance are independent, giving rise to statistical potentials with a tractable mathematical form. Nevertheless, it has been demonstrated that the incorrect pairwise decomposition can be effectively patched by a suitable definition of the reference state [45, 46, 80] to discern native binding modes from nonnative modes.

The scoring function dRNA developed by Zhou *et al.* [47] employed the distance-scaled, finite ideal-gas reference (DFIRE) technique for the construction of reference state. First used in protein-folding prediction [48], DFIRE essentially applies a weight function of r^α (with $\alpha < 2$) to the atom-pair density over separation r such that this weighted density is equal to the atom-pair density for ideal-gas. When adapted for protein-RNA interaction [47], a volume-fraction factor is needed to account for the fact that protein and RNA atoms with residue/base-specific types do not mix with each other.

In contrast, the scoring function of ITScore-PR developed by Huang and Zou [41] circumvents the reference state problem by using an iterative approach. This approach considers the entire energy landscape (embedded in nonnative structures) rather than only the global minimum of energy (corresponding to the native structures). Specifically, ITScore-PR is defined as the sum of statistical potentials $u_{ij}(r)$ over all atom pairs of $(i,\, j)$; the structure with the lowest score is the predicted binding mode. The separations r between atom i and atom j are divided into bins of a size of 0.2 Å. The potential $u_{ij}(r)$ reflecting the effective total interactions including electrostatic contribution is determined via an iterative formula:

$$u_{ij}^{(n+1)}(r) = u_{ij}^{(n)}(r) + \Delta u_{ij}^{(n)}(r) \tag{8}$$

$$\Delta u_{ij}^{(n)}(r) = \frac{1}{2} k_B T [g_{ij}^{(n)}(r) - g_{ij}^{obs}(r)] \tag{9}$$

Here n denotes the iteration step, $g_{ij}^{(n)}(r)$ and $g_{ij}^{obs}(r)$ stand for the radial distribution functions calculated according to $u_{ij}^{(n)}(r)$ and calculated from the native crystal structures in the training set, respectively. $g_{ij}^{(n)}(r)$ is calculated by

$$g_{ij}^{(n)}(r) = \frac{1}{K} \sum_{k=1}^{K} \sum_{l=0}^{L} P_k^l g_{ij}^{kl}(r) \tag{10}$$

where the radial distribution function $g_{ij}^{kl}(r)$ for atom pair $(i,\, j)$ observed in the l-th binding mode of the k-th complex is weighted by the score-dependent Boltzmann probability P_k^l that is obtained using the potentials $u_{ij}^{(n)}(r)$, and K is the total number of complexes in the training set used to produce $g_{ij}^{(n)}(r)$. The L binding modes for each complex can be generated by Monte Carlo or docking programs. In a similar way,

$$g_{ij}^{obs}(r) = \frac{1}{K} \sum_{k=1}^{K} g_{ij}^{k*}(r) \tag{11}$$

is the average of $g_{ij}^{k*}(r)$, the radial distribution function determined from the k-th complex in the training set, over all of the K complexes. The iteration continues through equation (8)–(11) until all native structures in the training set can be discriminated from decoys. ITScore-PR has been systematically tested and showed a consistently better performance than other scoring functions, particularly for rigid docking [41].

The validity of knowledge-based pairwise distance-dependent scoring function is conditioned on the extracted structural features, in this case the pairwise distance, sampled from the native structures obeying Boltzmann distribution at temperature T, which is restricted by the scarcity of solved protein-RNA complex structures. The published ITScore-PR was derived from a set of 110 nonredundant protein-RNA complexes, but can be easily improved upon the availability of a larger training set. ITScore-PR can also be extended to include modified RNA bases in a straightforward manner.

4. Chemical context profile based scoring function

A novel scoring function based on a weighted chemical context profile (CCP) was introduced by Parisien *et al.* [49]. This scoring function also uses a reduced representation for the protein and RNA. For the protein, the C_β carbon atom is chosen to be the interaction center for each residue. For the RNA, a heavy atom in the major groove (M), minor groove (m), and phosphate group (P) is selected, respectively, as interaction centers for each type of nucleotide (i.e., A, T, U, and G). In total, there are $20 \times 12 = 240$ possible interaction pairs between the protein and the RNA. A universal, distance-dependent interaction strength is assigned to each interaction pair:

$$f(r) = \frac{1}{\max(3.5\text{Å}, r - \hat{e})} \tag{12}$$

where r is the distance between the interaction centers, and $\hat{e}$ is the average distance between C_β and its partner interaction center. The CCP is defined as a 240-dimensional vector:

$$\overrightarrow{CCP} = \left(\sum_{C_\beta}^{Ala} \sum_{M}^{A} f(r), \sum_{C_\beta}^{Ala} \sum_{m}^{A} f(r), \sum_{C_\beta}^{Ala} \sum_{P}^{A} f(r), \cdots, \sum_{C_\beta}^{Val} \sum_{P}^{T} f(r) \right) \tag{13}$$

Each component of CCP corresponds to the summation of the interaction strengths for all the interaction pairs of the same kind. For example, the first component is for Ala and the major groove center. The CCP vector essentially captures the chemical context at the protein-RNA interface. The definition of an angle made by two vectors in the real space is generalized to the CCP space to define the chemical context discrepancy (CCD), whose defining relation in terms of two arbitrary vectors $\overrightarrow{CCP_1}$ and $\overrightarrow{CCP_2}$ is

$$\cos(CCD) = \frac{\left(\overrightarrow{CCP_1} \cdot \overrightarrow{CCP_2}\right)}{\left(\left|\overrightarrow{CCP_1}\right| \times \left|\overrightarrow{CCP_2}\right|\right)}. \tag{14}$$

In analogous to RMSD, CCD is indicative of similarity between two complex structures in the context of the interfacial chemical context, with a small value of CCD implying high similarity. The CCP-based scoring function is designed to identify native-like structures with low CCD value with respect to the native structure. The CCP-based score S is evaluated as

$$S = E_{coulomb} + \overrightarrow{\omega_{CCP}} \cdot \overrightarrow{CCP}, \tag{15}$$

where $E_{coulomb}$ is the generic electrostatic energy term, and $\overrightarrow{\omega_{CCP}}$ is a dual vector of the space that enables the weighted sum of CCP components. Only a subset of the components of the dual vector $\overrightarrow{\omega_{CCP}}$ is different from zero; hence, not all components of a CCP vector contribute to the score. A machine-learning approach [49, 50] is used to extract the CCP components having the most significant contributions to the discriminatory power, and at the same time, to parameterize the components of $\overrightarrow{\omega_{CCP}}$ accordingly. It should be pointed out that $\overrightarrow{\omega_{CCP}}$ and the contributing components of the CCP vector depend on the training set. If a new training set is used, a new training cycle is required. The discriminatory power of the scoring function was fully illustrated by a 100% accuracy in predicting six tRNA binding proteins that are not included in the training set [49]. Additionally, the CCP-based scoring function exhibits an interesting feature that only few, even one, resolved protein-RNA structures are sufficient to produce a scoring function that is able to discriminate native-like structures for other protein-RNA complexes sharing the same structural motifs. Using an ensemble of complex structures with disparate structural motifs actually compromise the performance of the resulting scoring function. On the contrary, scoring functions using statistical potential as described in the previous sections rely on a large set of known complex structures to achieve their discriminatory power for diverse complexes. The CCP-based scoring function is very useful in screening proteins that bind to a specific RNA in which similar structural motifs are involved. Moreover, the fact that only a small subset of components of the CCP vector are used for scoring implies that specific interactions dominate protein-RNA interactions, in consistent with previous structural characterization studies.

5. Benchmarks for assessment of scoring functions

Performance of various scoring functions can be compared on a docking benchmark. For this purpose, several benchmarks have been constructed and used. The benchmark assembled by Perez-Cano *et al.* [51] contains a total of 106 test cases that cover all major protein-RNA functional classes. Among these test cases, 71 cases have both bound and unbound protein and

RNA structures that were experimentally determined, for some cases RNA structures that bound to nonhomologous proteins are treated as the unbound structures. If there are no unbound structures, unbound structures were modeled based on homologous templates. Huang and Zou [52] also constructed a benchmark composed of 72 protein-RNA complex structures that covers diverse types of interactions and demonstrates various degrees of conformational change between apo- and holo-structures. Based on the extent of conformational change of unbound structures upon binding, the 72 structures can be divided into 49 easy, 16 medium and 7 difficult targets. Also, Xiao *et al.* [37] merged parts of the Perez-Cano and Huang and Zou sets into an extended benchmark that is referred to as RPDOCK set here. The performance of several scoring functions representative of propensity-based and iterative statistical potential principles are summarized in Table 1 in terms of rate of success in identifying the native structure as the top 1 model and in including the native structures in the top 10 models [52]. No benchmarking for the CCP-based scoring function has been found. Overall, ITScore-PR has the consistently best performance across the three benchmark sets. It is important to emphasize the dependence of performance of a scoring function on the generated decoys and the result assessment method, as illustrated by results on the RPDOCK set. Since the RPDOCK set is a mix of the other two benchmark sets, the performance of a scoring function on RPDOCK set is expected to be close to that of Huang and Zou or Perez-Cano set. However, there was a substantial discrepancy between the two, for example, the success rate of top 1 model for ITScore-PR was 48% on the RPDOCK set versus 24% on the Huang and Zou set. This unexpected difference is caused by the different docking algorithms used in generating decoys, i.e., ZDOCK 2.1 for Huang and Zou set as opposed to RPDOCK for the RPDOCK set, as well as a different criterion for success assessment between the RPDOCK and the Huang and Zou sets.

6. Discussion

We have described and discussed three kinds of scoring functions designed for protein-RNA structure prediction. Each of these scoring functions holds a unique aspect with respect to each other. They have different spatial resolution and are constructed with different methods. Nonetheless they share one common feature which is the absence of underpinning physical arguments. For the case of propensity based scoring functions, despite the resemblance in mathematical form, the relation between the score and free energy in statistical mechanics has not been formally established. For the case of the

Table 1: Success rates of scoring functions on different benchmark test sets

scoring function	Huang and Zou success rate (%)		Perez-Cano success rate (%)		RPDOCK success rate (%)	
	top 1	top 10	top 1	top 10	top 1	top 10
ITScore-PR	24	46	22.2	∼ 40	25.6 (48)	46.5 (62)
dRNA	24	44	13.9	∼ 40	–	–
DARS-RNP	16	38	15.3	26.4	13.6 (38)	36.4 (54)
QUASI-RNP	14	32	11.1	22.2	–	–
DECK-RP	–	–	–	–	22.7 (32)	45.5 (52)
Li	–	–	–	–	15.9 (10)	27.3 (32)

The RPDOCK benchmark is a combination of the Huang and Zou as well as Perez-Cano sets. The number in parentheses corresponds to the Huang and Zou set.

iterative statistical scoring function, although the concept of radial distribution function that was used in construction of the scoring function is closely related to theory for liquid, justification for its application to protein and RNA was not discussed explicitly. Actually an analystic treatment of the connectivity issue of atoms is still intractable. Moreover, the CCP-based scoring function is entirely based on a machine-learning approach. The validity of these scoring functions comes from their usefulness in applications. Ideally a scoring function calculates the free energy of the target complex structure, accounting for both enthalpy and entropy of the interaction. However, entropy as a measurement of the disorder of a system, can not be determined from the system's mechanical variables. Perhaps molecular dynamics (MD) simulation is a potential theoretical solution. Indeed, several MD simulation studies [65, 66, 67, 68] were devoted to investigation of protein RNA interactions. In theory, with a long enough MD simulation trajectory, the problem of free energy can be transformed into a counting problem. As a matter of fact, numerous computational techniques have been proposed to shorten the time it takes for a free energy calculation to converge. However, employing MD simulation for macromolecule structure prediction in general is still not feasible. In a typical docking program, the scoring function needs to evaluate tens of thousands of decoys in a short period of time. This exceeds the capability of MD simulations at present.

Propensity-based and iterative statistical scoring functions can be classified together as a single group of statistical potential based scoring function. These score functions evaluate the score for protein-RNA interaction as a weighted sum of the pairwise statistical potentials. For those scoring functions described in this review, a uniform weight was applied. However, the success of the CCP-based scoring function in using a small subset of intercomponent interactions for prediction suggests that specificity is achieved by only a few critical interactions. Therefore, the choice of uniform weight may not be optimum. On the other hand, the CCP-based scoring function has a strong dependence on its training set in the sense that while a scoring function has remarkable discriminatory power for complex sharing a common chemical context for interaction, it may not be as powerful for other complexes. The need for reparameterization of weights in a case dependent fashion raise a speculation that nonlinear many-body interaction among atoms or residues has a significant contribution to molecular specificity. The results of structural characterization of protein-RNA complex structures, such as binding pocket for bases [26] and hydrogen-bond network [19, 20, 25], already hint at the many-body effect. Without a theoretical framework, it is difficult to incorporate the potential nonlinear effect into existing scoring functions. Yet, we can use these knowledge from structural characterization as constraints to filter out inconsistent decoys before the application of scoring functions. It is reasonable to expect a native-like structure would satisfy these structural characteristics provided they are correct.

A major challenge for interaction prediction in general is conformational change of components upon binding. Flexible docking with the ability to handle conformational change is still an open question. For molecules like protein and RNA, the possible number of conformations increases exponentially with the length of their primary sequence. Conformation enumeration by brute force is out of scope of consideration. The FFT-based algorithm [36] is highly efficient only for rigid-body docking. The issue of conformational change can be partly compensated for by the characteristic low spatial sensitivity of some of the scoring functions. The introduced scoring functions span a narrow range of spatial resolution. Being the one with lowest resolution, the residue level propensity-based scoring function by Fernández-Recio *et al.* [29] has tolerance for minor conformational changes because its distance dependence is effected by a maximum intermolecular atom-to-atom cutoff distance, typical of 4 Å, used as the pairing criterion. On the other end of the range is the iterative statistical scoring function by Zou *et al.* [41], the distance dependence of the function is at atomic level. The high spatial resolution gives the function outstanding discriminatory power if a low RMSD native-like structure is present. Studies [69, 70, 71] have shown that although binding induced conformational change is common for RNP, major conformational changes are rare events. Therefore, before the advent of a ultimate solution to conformational change issue, we may rely on scoring functions with lower than atomic resolution to tackle problems with minor conformational changes.

The significance of a computational structure prediction method in biological studies resides in its ability to generate trustworthy 3D structures and to motivate interpretations of experimental results from a structural perspective. Tao *et al.* [49] demonstrated the usefulness of their prototypical CCP-based scoring function in identifying RNA binding proteins for a tRNA molecule. This finding as the first step in resolving a biological network of protein RNA interaction demonstrates the potential of the scoring function. Although the accuracy of existing scoring functions is always less than 100%, and there is no error estimation for the predicted structure, scoring function can still be valuable in experimental result interpretation. In fact, besides the high-resolution structure information from crystallography and NMR, there exist a large amount of low-resolution structural data from electron microscopy and other experimental methods. With the help of a scoring function, ambiguities in these low-resolution data can be resolved.

Acknowledgement

This work was supported by NIH R01GM109980 (PI: XZ) to XZ.

References

[1] Draper DE. Protein-RNA Recognition. *Annu Rev Biochem* 1995, 64:593–620.

[2] McClain WH. Rules that govern tRNA identity in protein synthesis. *J Mol Biol* 1993, 234(2):257–280.

[3] Giegé R, Sissler M, Florentz C. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res* 1998, 26(22):5017–5035.

[4] Rould MA, Perona JJ, Söll D, Steitz TA. Structure of E. coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 A resolution. *Science* 1989, 246:1135–1142.

[5] Rould MA, Perona JJ, Steitz TA. Structural basis of anticodon loop recognition by glutaminyl-tRNA synthetase. *Nature* 1991, 352:213–218.

[6] Ruff M, Krishnaswamy S, Boeglin M, Poterszman A, Mitschler A, Podjarny A, Rees B, Thierry JC, Moras D. Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp). *Science* 1991, 252(5013):1682–1689.

[7] Palade GE. A small particular component of the cytoplasm. *J Biophys Biochem Cytol* 1955, 1:59–68.

[8] Cech TR. The Ribosome Is a Ribozyme. *Science* 2000, 289:878–879.

[9] Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science* 2000, 289:905–920.

[10] Lerner MR, Steitz JA. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc Natl Acad Sci U S A* 1979, 76(11):5495–5499.

[11] Lerner MR, Boyle JA, Mount SM, Wolin SL, Steitz JA. Are snRNPs involved in splicing? *Nature* 1980, 283(5743):220–224.

[12] Walter P, Blobel G. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* 1982, 299(5885):691–698.

[13] Politz JC, Yarovoi S, Kilroy SM, Gowda K, Zwieb C, Pederson T. Signal recognition particle components in the nucleolus. *Proc Natl Acad Sci U S A* 2000, 97(1):55–60.

[14] Carpenter S, Aiello D, Atianand MK, Ricci EP, Gandhi P, Hall LL, Byron M, Monks B, Henry-Bezy M, Lawrence JB, O'Neill LAJ, Moore MJ, Caffrey DR, Fitzgerald KA. A Long Noncoding RNA Mediates Both Activation and Repression of Immune Response Genes. *Science* 2013, 341(6147):789–792.

[15] Lukong KE, Chang KW, Khandjian EW, Richard S. RNA-binding proteins in human genetic disease. *Trends Genet* 2008, 24:416–425.

[16] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000, 28:235–242.

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

[18] Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 2008, 456:464–469.

[19] Jones S, Daley DTA, Luscombe NM, Berman HM, Thornton JM. Protein-RNA interactions: a structural analysis. *Nucleic Acids Res* 2001, 29(4):943–954.

[20] Treger M, Westhof E. Statistical analysis of atomic contacts at RNA-protein interfaces. *J Mol Recognit* 2001, 14:199–214.

[21] Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. *Rna* 2001, 7(4):499–512.

[22] Jeong E, Kim H, Lee SW, Han K. Discovering the Interaction Propensities of Amino Acids and Nucleotides from Protein-RNA Complexes. *Mol Cells* 2003, 16(2):161–167.

[23] 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, 61:258–271.

[24] Kim OTP, 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.

[25] Allers J, Shamoo Y. Structure-based Analysis of Protein-RNA Interactions using the Program ENTANGLE. *J Mol Biol* 2001, 311:75–86.

[26] 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.

[27] Zhou P, Zou J, Tian F, Shang Z. Geometric Similarity Between Protein-RNA Interfaces. *J Comput Chem* 2009, 30:2738–2751.

[28] Kim H, Jeong E, Lee SW, Han K. Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. *FEBS Lett* 2003, 552:231–239.

[29] Pérez-Cano L, Fernández-Recio J. Optimal Protein-RNA Area, OPRA: A propensity-based method to identify RNA-binding sites on proteins. *Proteins* 2010, 78:25–35.

[30] Pérez-Cano L, Solernou A, Pons C, Fernández-Recio J. Structural Prediction of Protein-RNA Interaction by Computational Docking with Propensity-based Statistical Potentials. *Pac Symp Biocomput* 2010, pp. 293–301.

[31] Ellis JJ, Broom M, Jones S. Protein-RNA interactions: structural analysis and functional classes. *Proteins* 2007, 66:903–911.

[32] Li CH, Cao LB, Su JG, Yang YX, Wang CX. A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. *Proteins* 2012, 80:14–24.

[33] Tuszynska I, Bujnicki JM. DARS-RNP and QUASI-RNP: New statistical potentials for protein-RNA docking. *BMC Bioinformatics* 2011, 12:348.

[34] Boniecki M, Rotkiewic P, Skolnick J, Kolinski A. Protein fragment reconstruction using various modeling techniques. *J Comput Aided Mol Des* 2003, 17(11):725–738.

[35] Malolepsza E, Boniecki M, Kolinski A, Piela L. Theoretical model of prion propagation: a misfolded protein induces misfolding. *Proc Natl Acad Sci U S A* 2005, 102(22):7835–7840.

[36] Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A* 1992, 89(6):2195–2199.

[37] Huang Y, Liu S, Guo D, Li L, Xiao Y. A novel protocol for three-dimensional structure prediction of RNA-protein complexes. *Sci Rep* 2013, 3:1887.

[38] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983, 22(12):2577–2637.

[39] Robertson TA, Varani G. An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure. *Proteins* 2007, 66:359–374.

[40] Zheng S, Robertson TA, Varani G. A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. *Febs J* 2007, 274(24):6378–6391.

[41] Huang SY, Zou X. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic Acids Res* 2014, 42(7):e55.

[42] Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 1990, 213:859–883.

[43] Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci* 2006, 15:2507–2524.

[44] Huang SY, Zou X. An iterative knowledge-based scoring function for protein-protein recognition. *Proteins* 2008, 72:557–579.

[45] Huang SY, Zou X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J Comput Chem* 2006, 27:1865–1875.

[46] Huang SY, Zou X. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J Comput Chem* 2006, 27:1876–1882.

[47] Zhao H, Yang Y, Zhou Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res* 2010, 39:3017–3025.

[48] Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 2002, 11:2714–2726.

[49] Parisien M, Wang X, Perdrizet II G, Lamphear C, Fierke CA, Maheshwari KC, Wilde MJ, Sosnick TR, Pan T. Discovering RNA-Protein Interactome by Using Chemical Context Profiling of the RNA-Protein Interface. *Cell Reports* 2013, 3:1703–1713.

[50] Romero E, Sopena JM. Performing feature selection with multilayer perceptrons. *IEEE Trans Neural Netw* 2008, 19:431–441.

[51] Pérez-Cano L, Jiménez-García B, Fernánodez-Recio JFR. A protein-RNA docking benchmark (II): Extended set from experimental and homology modeling data. *Proteins* 2012, 80:1872–1882.

[52] Huang SY, Zou X. A nonredundant structure dataset for benchmarking protein-RNA computational docking. *J Comput Chem* 2013, 34:311–318.

[53] Thiel KW, Giangrande PH. Therapeutic applications of DNA and RNA aptamers. *Oligonucleotides* 2009, 19:209–222.

[54] Bagalkot V, Farokhzad OC, Langer R, Jon S. An aptamer-doxorubicin physical conjugate as a novel targeted drug-delivery platform. *Angew Chem Int Ed Engl* 2006, 45:8149–8152.

[55] Dhar S, Gu FX, Langer R, Farokhzad OC, Lippard SJ. Targeted delivery of cisplatin to prostate cancer cells by aptamer functionalized

Pt(IV) prodrug-PLGA-PEG nanoparticles. *Proc Natl Acad Sci U S A* 2008, 105:17356–17361.

[56] Gu F, Zhang L, Teply BA, Mann N, Wang A, Radovic-Moreno AF, Langer R, Farokhzad OC. Precise engineering of targeted nanoparticles by using self-assembled biointegrated block copolymers. *Proc Natl Acad Sci U S A* 2008, 105:2586–2591.

[57] Hicke BJ, Stephens AW, Gould T, Chang YF, Lynott CK, Heil J, Borkowski S, Hilger CS, Cook G, Warren S, Schmidt PG. Tumer targeting by an aptamer. *J Nucl Med* 2006, 47:668–678.

[58] Pastor F, Kolonias D, Giangrande PH, Gilboa E. Induction of tumour immunity by targeted inhibition of nonsense-mediated mRNA decay. *Nature* 2010, 465:227–230.

[59] Buff MC, Schafer F, Wulffen B, Muller J, Potzsch B, Heckel A, Mayer G. Dependence of aptamer activity on opposed terminal extensions: improvement of light-regulation efficiency. *Nucleic Acids Res* 2010, 38:2111–2118.

[60] Chakravarthy U, Adamis AP, Cunningham ETJ, Goldbaum M, Guyer DR, Katz B, Patel M. Year 2 efficacy results of 2 randomized controlled clinical trials of pegaptanib for neovascular age-related macular degeneration. *Ophthalmology* 2006, 113(1508):e1501–e1525.

[61] Ng EW, Adamis AP. Anti-VEGF aptamer (pegaptanib) therapy for ocular vascular diseases. *Ann N Y Acad Sci* 2006, 1082:151–171.

[62] Rockey WM, Hernandez FJ, Huang SY, Cao S, Howell CA, Thomas GS, Liu XY, Lapteva N, Spencer DM, McNamara II JO, Zou X, Chen SJ, Giangrande PH. Rational Truncation of an RNA Aptamer to Prostat-Specific Membrane Antigen Using Computational Structural Modeling. *Nucleic Acid Ther* 2011, 21(5):299–314.

[63] Huang SY, Zou X. MDockPP: A hierarchical approach for protein-protein docking and its application to CAPRI rounds 15–19. *Proteins* 2010, 78:3096–3103.

[64] Zhao H, Yang Y, Janga SC, Kao CC, Zhou Y. Prediction and validation of the unexplored RNA-binding protein atlas of the human proteome. *Proteins* 2014, 82:640–647.

[65] Reyes CM, Kollma PA. Structure and Thermodynamics of RNA-protein Binding: Using Molecular Dynamics and Free Energy Analyses to Cal-

culate the Free Energies of Binding and Conformational Change. *J Mol Biol* 2000, 297:1145–1158.

[66] MacKerell Jr AD, Nilsson L. Molecular dynamics simulations of nucleic acid-protein complexes. *Curr Opin Struct Biol* 2008, 18:194–199.

[67] Setny P, Zacharias M. A coarse-grained force field for Protein-RNA docking. *Nucleic Acids Res* 2011, 39(21):9118–9129.

[68] Henriksen NM, Roe DR, Cheatham III TE. Reliable oligonucleotide conformational ensemble generation in explicit solvent for force field assessment using reservoir replica exchange molecular dynamics simulations. *J Phys Chem B* 2013, 117(15):4014–4027.

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

[70] Ellis JJ, Jones S. Evaluating conformational changes in protein structures binding RNA. *Proteins* 2008, 70:1518–1526.

[71] Shajani Z, Sykes MT, Williamson JR. Assembly of bacterial ribosomes. *Annu Rev Biochem* 2011, 80:501–526.

[72] Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* 1976, 9:945–950.

[73] Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal-structures-Quasi-chemical approximation. *Macromolecules* 1985, 18:534–552.

[74] Zhao H, Yang Y, Zhou Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biology* 2011, 8:988–996.

[75] Thomas PD, Dill KA. Statistical potentials extracted from protein structures: How accurate are they? *J Mol Biol* 1996, 257:457–469.

[76] Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci USA* 1996, 93:11628–11633.

[77] Zhang L, Skolnick J. How do potentials derived from structural databases relate to “true” potentials? *Protein Sci* 1998, 7:112–122.

[78] Koppensteiner WA, Sippl MJ. Knowledge-based potentials – Back to the roots. *Biochemistry* (Moscow) 1998, 63:247–252.

[79] Zhang C, Liu S, Zhu Q, Zhou Y. A knowledge-based energy function for protein-ligand, protein-protein and protein-DNA complexes. *J Med Chem* 2005, 48:2325–2335. MR2238949

[80] Chuang, GY, Kozakov D, Brenke R, Comeau SR, Vajda S. DARS (Decoys As the Reference State) Potentials for Protein-Protein Docking. *Biophys J* 2008, 95:4217–4227.

[81] Perez C, Ortiz, AR. Evaluation of docking functions for protein-ligand interactions. *J Med Chem* 2001, 44:3768–3785.

[82] Huang SY, Zou X. Statistical mechanics-based method to extract atomic distance-dependent potentials from protein structures. *Proteins* 2011, 79:2648–2661.

Liming Qiu
Dalton Cardiovascular Research Center
University of Missouri
Columbia, Missouri 65211
USA
E-mail address: qiul@missouri.edu

Xiaoqin Zou
Dalton Cardiovascular Research Center
Department of Physics & Astronomy
Department of Biochemistry
Institute for Data Science and Informatics
University of Missouri
Columbia, Missouri 65211
USA
E-mail address: zoux@missouri.edu

Received October 4, 2019