METHODS AND APPLICATIONS

The dataset for protein–RNA binding affinity

Xiufeng Yang, Haotian Li, Yangyu Huang, and Shiyong Liu*

Biomolecular Physics and Modeling Group, Department of Physics, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

Received 16 July 2013; Accepted 7 October 2013
DOI: 10.1002/pro.2383
Published online 11 October 2013 proteinscience.org

Abstract: We have developed a non-redundant protein–RNA binding benchmark dataset derived from the available protein–RNA structures in the Protein Database Bank. It consists of 73 complexes with measured binding affinity. The experimental conditions (pH and temperature) for binding affinity measurements are also listed in our dataset. This binding affinity dataset can be used to compare and develop protein–RNA scoring functions. The predicted binding free energy of the 73 complexes from three available scoring functions for protein–RNA docking has a low correlation with the binding Gibbs free energy calculated from K_d. © 2013 The Protein Society

Keywords: scoring function; binding free energy; protein-RNA docking; binding affinity benchmark

Introduction

RNA–protein interactions play key roles in all kinds of biological processes. High resolution structures of protein–RNA complexes are necessary for understanding mechanisms of protein–RNA interactions. Unfortunately, it is difficult and slow to determine the 3D structure of protein–RNA complexes by X-ray crystallography and nuclear magnetic resonance spectroscopy. Alternatively, computational protein–RNA docking provides another way to build the 3D structure of a protein–RNA complex if the unbound protein and RNA structures are available.

In the past decade, a number of methods have been developed to identify protein–RNA binding sites experimentally[1] and computationally.[2–5] That binding site information could be used to improve protein–RNA docking. However, there are still very few methods for protein–RNA docking and scoring. In 2011, Setny *et al.*[6] developed a coarse-grained force field for protein–RNA docking, which only predicted one case in the top 100 from seven unbound protein–RNA cases. Also in 2011, Tuszynska *et al.*[7] published two knowledge-based scoring functions, which were tested on eight unbound-unbound protein–RNA docking decoys made by the GRAMM program. The results showed that these potentials recognized near-native structures for four out of eight cases. At the same time, Li *et al.*[8] proposed a residue-nucleotide propensity potential, in which they found the secondary structure information for RNA is a key factor to the predictive power of their pair potentials. It is expected that more and more reliable docking and scoring methods will be developed in the near future.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Natural Science Foundation of China; Grant number: 31100522; Grant sponsor: National High Technology Research and Development Program of China; Grant number: 2012AA020402; Grant sponsor: Specialized Research Fund for the Doctoral Program of Higher Education; Grant number: 20110142120038; Grant sponsor: National Natural Science Foundation of China; Grant number: 31100522.

*Correspondence to: Shiyong Liu, Biomolecular Physics and Modeling Group, Department of Physics, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China.
E-mail: liushiyong@gmail.com

In order to measure and compare the performance of different methods for predicting protein–RNA complex structures, two protein–RNA docking benchmarks were released recently.[9,10] Although a series of binding affinity benchmarks[11,12] are used for the development of the scoring functions in protein–ligand and protein–protein docking simulation, binding affinity benchmark for protein–RNA scoring functions is still lacking. Based on protein–protein binding dataset, protein–protein kinetic rate constants were well studied and a number of correlations with $\log k_{on}$ were identified.[13] They found that the most important correlated factor was the energy difference between the unbound and bound conformational state, which was calculated by some either coarse grained or atomistic pair potentials. Since lack of a binding affinity dataset of protein–RNA has become a bottleneck for developing more accurate scoring functions, we have decided to collect the experimentally measured binding affinity data from the scientific literature. Only protein–RNA complexes with experimentally determined structures were considered in this work. This RNA–protein binding benchmark will benefit the research related to protein–RNA docking and binding mechanism.

Results and Discussion

Binding affinities dataset and statistical potentials

We have assembled a protein–RNA binding affinity dataset that includes the experimentally characterized equilibrium dissociation constants (K_d) of 73 protein–RNA complexes along with the methods used to determine them. It also includes the experimental conditions (pH and temperature) at which the K_d values were measured. The binding Gibbs free energy calculated with the K_d. can provide us a factor to access the scoring functions in protein-RNA docking. Two medium-resolution knowledge-based potentials (QUASI-RNP and DARS-RNP) for scoring protein–RNA models have been proposed.[7] Both statistical potentials comprise four terms: a distance-dependent energy term, an angular-dependent energy term, a site-dependent energy term, and a penalty for steric clashes. Equal weights to all four terms are assigned. We have tested the 73 protein–RNA complexes native structures with DARS-RNP, QUASI-RNP potentials and our group's newly developed scoring function DECK-RP[14] (Fig. 1). All correlation coefficients between the score and observed binding free energy (calculated for scatter plots from Fig. 1) are low (DARS-RNP: $R = 0.20$, QUASI: $R = 0.21$, DECK-RP: $R = -0.12$). The result shows the scores were not directly correlated with observed binding free energies; hence, we need to develop better scoring function for protein–RNA docking. Based on our protein–RNA binding dataset, these weights may be optimized to improve the accuracy of the scoring functions.

Bind or not bind?

Given a protein–RNA binding dataset, it is possible to develop a model to predict protein–RNA binding affinity by using a large set of molecular descriptors. Similar approach has been successfully implemented on protein–protein binding affinity prediction.[15] If protein–RNA binding free energy model is constructed, we would assess whether the protein could bind RNA. It would open a new way to do structure-based design of protein–RNA interactions by screening PDB database. Recently, two groups reported that they successfully designed protein–RNA interaction[16,17] for PUF proteins by using a yeast three-hybrid system. Due to many potential biological and medical applications, the design of protein–RNA interaction by a computational method will be explored in the future.

Binding affinities dataset and docking benchmark

Three protein–RNA docking benchmarks were released recently.[9,10,18] Benchmark(I)[9] includes 36 unbound-bound cases and 9 unbound-unbound cases. Benchmark(II)[10] is composed of 71 cases, which includes an additional set of 35 cases by homology modeling. These two benchmarks would contribute to the better understanding and prediction of protein–RNA interactions. The third dataset[18] of 72 targets consists of 52 unbound–unbound test complexes, and 20 unbound–bound test complexes. The dataset constructed by us is a binding affinity dataset. Compared with benchmark (II), there are 10 cases in our dataset that are contained in the experiment set, eight cases contained in the homology modeling set. With the development of technology, more binding affinity values for protein–RNA complexes will become available. It will help improve the veracity and reliability of protein–RNA docking assessing.

Methods

Dataset of protein–RNA complexes

There were 1495 protein–RNA complex structures that have been deposited into the Protein Database Bank (PDB) on September 16, 2013. First, we built a protein–RNA complex structure set. Those cases that do not meet the following two conditions were filtered out: (1) The RNA sequence has at least five nucleotides, the protein sequence has at least twenty amino acids; (2) The structure of protein–RNA complexes are determined by X-ray crystallography or NMR. And the three-dimensional structures solved from X-ray crystallography should have a resolution better than 3 Å. Larger Ribosome complex and virus

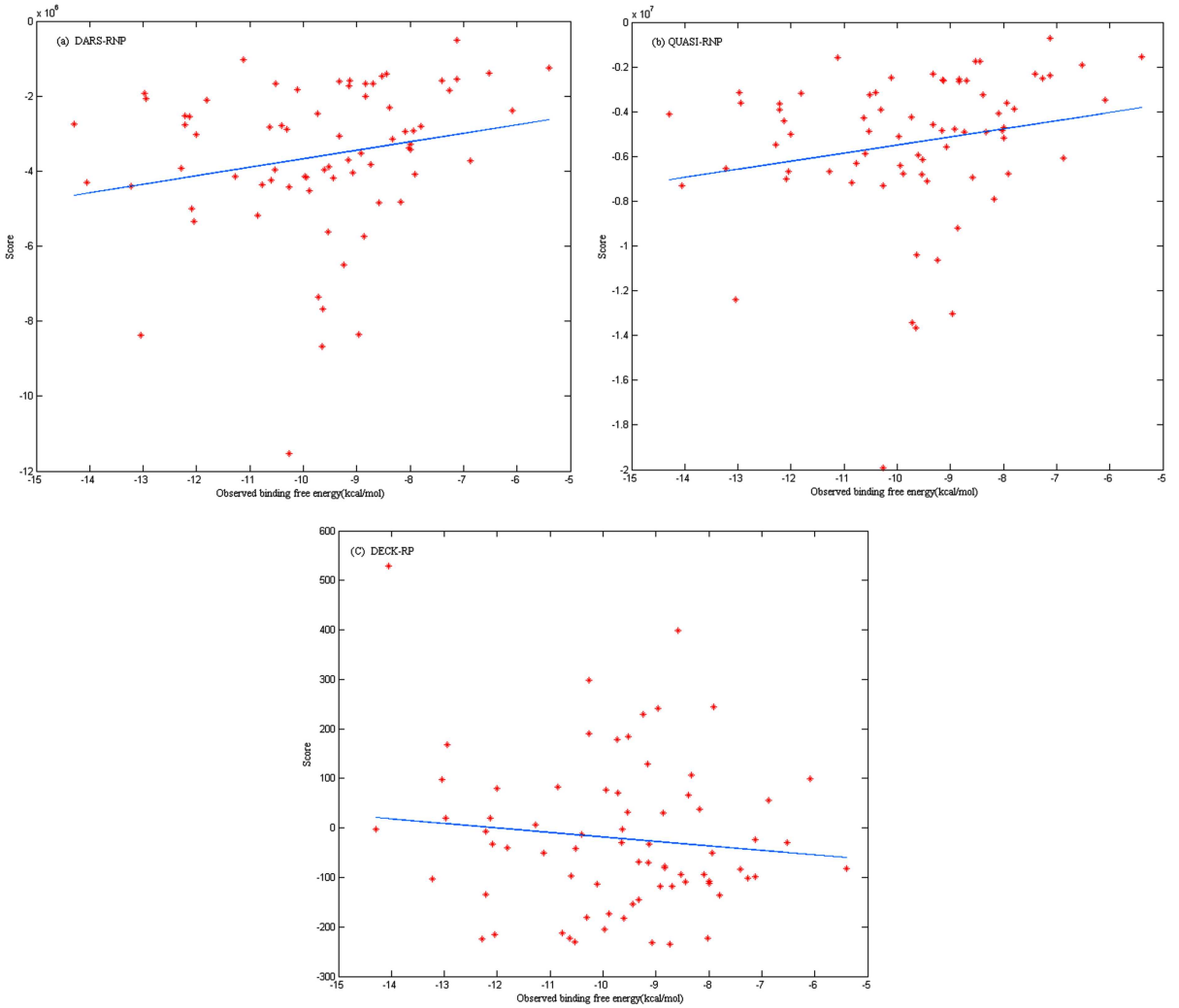

Figure 1. The correlation between score (respectively tested in three different scoring functions: DARS-RNP, QUASI-RNP, DECK-RP) and observed binding free energy. Their correlation coefficients obtained with MATLAB are low (DARS-RNP: $R = 0.20$, QUASI: $R = 0.21$, DECK-RP: $R = -0.12$) [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

structure were removed. This results in 554 protein–RNA biological complexes in our structure dataset. The protein sequences in these complexes with at least 70% sequence identity were assigned into 261 clusters according to the weekly clustering of protein chains in the PDB by BLASTClust (ftp://resources.rcsb.org/sequence/clusters/bc-70.out). One complex in each cluster is kept. Second, we search the scientific literature for binding affinity data for those protein–RNA complexes selected above. There is a primary reference, which can be retrieved from the corresponding PDB file, associated with every structure deposited in PDB. Therefore, we can scan the reference to search for the complex's binding affinity. If its authors have measured the binding affinity for the complex, it is expected that the binding affinity would be released in the original article. The value appears in publications in the form of either equilibrium constants (K_d or $K_a = 1/K_d$), or as the ratio $K_d = k_d/k_a$ of rate constants measured from surface plasmon resonance and other kinetic measurements.[12] We can also obtain the binding affinity information through citations in the primary reference. If the binding affinity was not available in these publications mentioned above, we search for the values in Google Scholar with some key words such as the component molecules of the complex, "binding affinity/K_d" and one of the main methods used to measure the protein–RNA binding affinity. If one of the 261 complexes does not have available binding affinity, in order to expand the dataset, we would continue to search for binding affinity for another complex which has at least 70% protein sequence identity to replace this one. Finally, we collected K_d values for 73 complexes (Supporting Information Dataset) and compiled a binding affinity dataset along with reference citations, molecules description, as well as the experimental conditions

(pH and temperature). The dissociation Gibbs free energy was calculated (some cases do not have a stated temperature, we adopt room temperature in the calculation) by the equation[12]:

$$\Delta G^{\circ} = -RT\ln K_d$$

To maximize their reliability, we have double checked the 73 values. The reference citations were presented for the convenience of rechecking and obtaining further details. With the development of technology, more binding affinity values for protein–RNA complex will become available. We are committed to update the database yearly so as to enhance its usefulness to both improving and developing the docking scoring functions.

References

1. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141:129–141.
2. Kim OT, Yura K, Go N (2006) Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. Nucleic Acids Res 34:6450–6460.
3. Zhao H, Yang Y, Zhou Y (2011) Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucleic Acids Res 39:3017–3025.
4. Fernandez M, Kumagai Y, Standley DM, Sarai A, Mizuguchi K, Ahmad S (2011) Prediction of dinucleotide-specific RNA-binding sites in proteins. BMC Bioinform 12:S5.
5. Dror I, Shazman S, Mukherjee S, Zhang Y, Glaser F, Mandel-Gutfreund Y (2012) Predicting nucleic acid binding interfaces from structural models of proteins. Proteins: Structure, Function, and Bioinformatics 80, 482–489.
6. Setny P, Zacharias M (2011) A coarse-grained force field for protein-RNA docking. Nucleic Acids Res 39: 9118–9129.
7. Tuszynska I, Bujnicki JM (2011) DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. BMC Bioinform 12:348.
8. Li CH, Cao LB, Su JG, Yang YX, Wang CX (2012) A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. Proteins 80:14–24.
9. Barik A, Nithin C, Manasa P, Bahadur RP (2012) A protein-RNA docking benchmark (I): Non-redundant cases. Proteins 80:1866–1871.
10. Perez-Cano L, Jimenez-Garcia B, Fernandez-Recio J (2012) A protein-RNA docking benchmark (II): Extended set from experimental and homology modeling data. Proteins 80:1872–1882.
11. Wang R, Fang X, Lu Y, Wang S (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem 47:2977–2980.
12. Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AM, Janin J (2011) A structure-based benchmark for protein-protein binding affinity. Protein Sci 20:482–491.
13. Moal IH, Bates PA (2012) Kinetic rate constant prediction supports the conformational selection mechanism of protein binding. PLoS Comp Biol 8:e1002351.
14. Huang Y, Liu S, Guo D, Li L, Xiao Y (2013) A novel protocol for three-dimensional structure prediction of RNA-protein complexes. Scientific Rep 3:1887.
15. Moal IH, Agius R, Bates PA (2011) Protein-protein binding affinity prediction on a diverse set of structures. Bioinformatics 27:3002–3009.
16. Filipovska A, Razif MF, Nygard KK, Rackham O (2011) A universal code for RNA recognition by PUF proteins. Nat Chem Biol 7:425–427.
17. Dong S, Wang Y, Cassidy-Amstutz C, Lu G, Bigler R, Jezyk MR, Li C, Hall TM, Wang Z (2011) Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. J Biol Chem 286: 26732–26742.
18. Huang S-Y, Zou X (2012) A nonredundant structure dataset for benchmarking protein-RNA computational docking. J Comp Chem 34:311–318.