8-2015

COMPUTATIONAL MODELING OF RNA-SMALL MOLECULE AND RNA-PROTEIN INTERACTIONS

Lu Chen

COMPUTATIONAL MODELING OF RNA-SMALL MOLECULE AND RNA-PROTEIN INTERACTIONS

by

Lu Chen

APPROVED:

Advisory Professor, Shuxing Zhang, Ph.D.

George A. Calin, Ph.D.

Xiongbin Lu, Ph.D.

Edward Nikonowicz, Ph.D.

Wenyi Wang, Ph.D.

APPROVED:

Dean, The University of Texas

Graduate School of Biomedical Sciences at Houston

COMPUTATIONAL MODELING OF RNA-SMALL MOLECULE AND RNA-PROTEIN INTERACTIONS

A

DISSERTATION

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

M.D. Anderson Cancer Center

Graduate School of Biomedical Sciences

In Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Lu Chen, B.S.

Houston, Texas

August 2015

Dedication

To my darling wife, Xiaofei Xiong and son, Ryan X. Chen who have loved, inspired, encouraged, motivated and supported me on whatever decision I have made. My parents, Jianjun Chen and Guoyun Zhang, who are always supportive in my scientific career. Dr. Shuxing Zhang, greatest mentor in my life.

Acknowledgements

I acknowledge all the hardworking scientists in Dr. Shuxing Zhang's lab, including John Morrow, Micheal Cato, Zhi Tan, Srinivas Alla, Hoang Tran, Lei Du-cuny, Longzhang Tian, Sharangdhar Phatak, Nathan Ihle, Ryan Watkins. I also thank Matri and Paloma in George Calin's lab, and great researchers in Dr. Edward Nikonowicz's lab for their dedicate contributions to experimental validations.

I owe great acknowledgement to my advisory committee, Shuxing Zhang, George Calin, Xiongbin Lu, Edward Nikonowicz, Wenyi Wang and John Ladbury for your valuable inputs and insights into this thesis. I could not have done it without your support.

Special thanks to University of Texas M.D. Anderson and University of Austin for providing state-of-the-art HPC resources. Thank TACC for free academic license for Gaussian09. Thank Dr. Jinbo Xu for providing the source code and assisting me on deploying RaptorX for RNA-protein interface threading.

COMPUTATIONAL MODELING OF RNA-SMALL MOLECULE AND RNA-PROTEIN INTERACTIONS

By Lu Chen, B.S.

Advisor: Shuxing Zhang, Ph.D.

The past decade has witnessed an era of RNA biology; despite the considerable discoveries nowadays, challenges still remain when one aims to screen RNA-interacting small molecule or RNA-interacting protein. These challenges imply an immediate need for cost-efficient while predictive computational tools capable of generating insightful hypotheses to discover novel RNA-interacting small molecule or RNA-interacting protein. Thus, we implemented novel computational models in this dissertation to predict RNA-ligand interactions (Chapter 1) and RNA-protein interactions (Chapter 2).

Targeting RNA has not garnered comparable interest as protein, and is restricted by lack of computational tools for structure-based drug design. To test the potential of translating molecular docking tools designed for protein to RNA-ligand docking and virtual screening, we benchmarked 5 docking software and 11 scoring functions to assess their performances in pose reproduction, pose ranking, score-RMSD correlation and virtual screening. From this benchmark, we proposed a three-step docking pipelines optimized for virtual screening against RNAs with different flexibility properties. Using this pipeline, we have successfully

identified a selective compound binding to GA:UU motif. Both NMR and the subsequent MD simulation proved its selective binding to GA:UU motif flanked by two tandem flexible base pairs next to GA. Consistent to the 3D model, SAR analysis revealed that any *R*-group substitution would abolish the binding.

Current computational methods for RNA-protein interaction prediction (sequence-based or structure-based) are either short of interpretability or robustness. Aware of these pitfalls, we implemented RNA-Protein interaction prediction through Interface Threading (*RPIT*), which identifies and references a known RNA-protein interface as the template to infer the region where the interaction occurs and predict the interacting propensity based on the interface profiles. To estimate the propensity more accurately, we implemented five statistical scoring functions based our unique collection of non-redundant protein-RNA interaction database. Our benchmark using leave-protein-out cross validation and two external validation sets resulted in overall 70%-80% accuracy of *RPIT*. Compared with other methods, *RPIT* offers an inexpensive but robust method for *in silico* prediction of RNA-protein interaction networks, and for prioritizing putative RNA-protein pairs using virtual screening.

Table of Contents

List of Illustrations

Chapter 1 (no illustrations)

Chapter 2

Chapter 3

Chapter 4 (no illustrations)

List of Tables

Chapter 1 (no tables)

Chapter 2

Chapter 3

Chapter 4 (no tables)

Abbreviations

μL: microliter

μM: micromolar

PDB: Protein Data Bank

NMR: Nuclear magnetic resonance

RDC: Residual dipolar coupling

ROC: Receiver operating characteristics

AUC: Area under the curve

VUS: Volume under the surface

RMSD: Root mean square deviation

RMSE: Root mean square error

PPI: Protein-protein interaction

RPI: RNA-protein interaction

H-bond: Hydrogen bond

vdW: van der Waals

PCA: Principle component analysis

ANOVA: Analysis of variance

RF: Random forest

SVM: Support vector machine

KNN: K-nearest neighbor

Chapter 1: Introduction

1.1 Targeting RNA with small molecules

1.1.1 RNA as therapeutic target

Recent advancements in RNA biology refresh our understandings of life and potentiate the strategy of targeting RNA for a large multitude of diseases. DNAs and proteins have received much attention as therapeutic targets of small molecules, but RNAs have not garnered comparable interest for a variety of reasons including relatively few and ill-defined structures, the intrinsic dynamics of RNAs, and sometimes less appreciated link between RNA molecules and biological functions. Historically, targeting RNA for therapeutic development has been envisaged by many to be a cost-expensive strategy. However, several pioneer studies have provided proof-of-principles that targeting RNA is a feasible strategy for treatment infectious diseases and cancers. Targets that are mostly investigated includes prokaryotic rRNA A-site [1-3], HIV-1 TAR RNA [4-6] and riboswitches [7-9]. Furthermore, researchers are exploring new-generation, drug-like compounds for disease-related RNAs including CUG- or CCUG-repeated mRNA [10-12], miRNA [13, 14] and internal ribosome entry site (IRES) [15, 16]. All these efforts represent a paradigm-shift strategy to target a more upstream biomolecule, that is, hub RNA, which regulates multiple disease-related proteins.

1.1.2 Hit identification via molecular docking

A number of strategies have been used for lead identification targeting RNA, including high-throughput screening, rational design by NMR or computational modeling. Conventional high-throughput small molecule screening methods are well-suited to catalysis-based assays,

but are limited in screening compounds for RNA binding by detection assays that generally rely on binding-coupled conformational changes which compete with intrinsic RNA dynamics. Therefore, virtual high-throughput screening (vHTS) using molecular docking has become one of the core lead discovery technologies in the pharmaceutical industry [17], which provides a practical route to identify more selective RNA-binding compounds in a more efficient fashion.

Molecular docking is one of the key strategies for computational structure-based drug design [18]. The goal of molecular docking is to predict the favored binding mode of a small molecule (ligand) in a macromolecule pocket (e.g., protein or nucleic acid) with respect to the 3D structure [19]. Docking has become a popular structure-based approach to prioritize active compounds from a large chemical database prior to expensive and time-consuming experimental validation. In general, molecular docking procedure can be divided into two steps: conformational sampling and scoring. During the conformational sampling phase, a large amount of ligand conformations and coordinates will be numerated and submit a few to the second phase based upon a fast, but less accurate scoring function which roughly evaluates the fitness of binding. In the second phase, a more accurate but more complicated scoring function will be applied to differentiate the “good” (energetically-favored) poses against the “bad” (energetically-prohibited) poses. Although ranking compounds according to relative binding affinity still remain challenging, docking-based virtual screening has been employed for lead identification and optimization for a number of protein targets, which has been reviewed by Chen et al. [18].

1.1.3 Current in silico methods of targeting RNA

Like protein, RNA can fold into well-defined tertiary structures (such as helix, hairpin, bulge and pseudoknot), providing the structural basis for structure-based rational design. There have been several studies which aim to translate the docking/scoring functions that have led to great successes for protein targets, but are parameterized exclusively using protein-ligand complex, to RNA target. For example, GOLD and Glide [20] and AutoDock4 [21, 22] have been benchmarked for their usage in docking small molecules to RNA receptor. Others were seeking to implement RNA-specific scoring functions, e.g., force field-based scoring functions based implicit solvent models [23], empirical scoring function [2, 24, 25] and knowledge-based scoring function [26]. The tools that model a flexible RNA receptor, such as MORDOR (molecular recognition with a driven dynamics optimizer) [27], may give more accurate predictions, yet not feasible to screen a large chemical database. None of these computational tools have been benchmarked using publicly available dataset, and thus the predictive capability of these models still remains ambiguous. Actually, we have found that the docking parameters widely used in proteins may not be well translated to RNA systems. For instance, electrostatic attraction between RNA backbone and positively charge group (such as piperazine) can be overestimated [23, 28, 29], and desolvation term need improvement [21]. Hence, we believe that a mature structure-based modeling technique designed specifically for RNAs, e.g., docking-based virtual screening, is still lacking, despite the efforts mentioned above.

1.2 Discovering novel RNA-protein interaction

1.2.1 Emerging RNA-protein interactions (RPI)

The past decade has witnessed an era of RNA biology: new RNA, new functionalities, and new interactions. RNA-protein interaction (RPI) takes a major proportion in these exciting discoveries, owing to its critical roles in cellular processes, such as transcription, translation and regulation [30]. Ribosome and spliceosome are the two well-known examples of large bio-machineries involving complex RPI. Various non-coding RNAs, such as microRNA (miRNA), long non-coding RNA (lncRNA) and Piwi-interacting RNA (piRNA), interplay with a large number of proteins via indirect mechanism or direct binding [31]. For example, a vast majority of lncRNA reported in the literature is able to form machinery with multiple proteins. lncRNA that folds into complex tertiary structure has been shown to modulates the transcriptional factors that regulate the gene-specific transcription, basal transcription machinery, splicing and translation [32]. Recent discoveries of new functionalities of miRNA, e.g., direct binding to hnRNP-E2 [33], ELAVL1 [34], or being the native ligand of Toll-like receptors (TLR) [35, 36], have updated the dogmatic understanding of microRNA. On the other hand, more studies focused on the biogenesis of miRNAs, which is regulated at posttranscriptional level via various RNA-binding proteins (e.g., hnRNPA1 [37-39], PTBP1 [39], KSRP [40-42], Lin28 [43]). piRNA is another representative protein-binding non-coding RNA that form RNA-protein complexes through interacting with piwi proteins [44]. This RPI mediates the epigenetic and posttranscriptional gene regulations, especially in germline cells [45].

1.2.2 RNA-protein interface

Current understandings of RNA-protein binding interface primarily come from the analysis of high resolution structures. For example, several analyses based upon small datasets from PDB (81 complexes [46], 54 crystal structures [47], 77 complexes [48], 41 complexes [49], 89 complexes [50], 152 complexes [51]) have provided insightful knowledge of the physicochemical patterns that are essential to form a RPI. Despite the trivial differences between studies, most of them did reach a consensus. From a structural perspective, Huang et al. summarized four features of RPI interfaces that are significantly different from PPI interface: (1) The atomic packing of RPI interfaces is looser than that of PPI interfaces; (2) There is a strong residue preference at RPI interface-positively charged residues are significantly favored (Arg and Lys) whereas negatively charged residues (Asp and Glu) are disfavored; (3) Stacking interaction plays a more critical role in RPI than PPI, especially the π-π stacking between aromatic amino acids (His, Tyr and Trp) and nucleotide base; (4) Secondary structure states of amino acids and nucleotides are important at RPI interface [52]. All these RPI-specific features should be considered when one designs statistical scoring functions to assess the fitness of RNA-protein binding. These signatures, however, bring both insights and challenges. With respect to feature (1), macromolecular docking, which determine the fitness of binding based on structural complementarity between RNA and protein, is historically optimized to result a compact interface. As to feature (2), despite the preference of positively-charged protein residue at the interface, the contributions of such electrostatic attraction to RNA-protein binding affinity can be easily overestimated, compared with other more sequence-specific type of interaction. Regarding feature (3), to the best of our

knowledge, there is no grounded mathematical model to quantitatively evaluate the propensity of stacking. Finally, unlike secondary structure states of protein residues, which have 3 major clusters (helix, sheet and coil), the base pairing states of nucleic acid is more complicated. Other than well-defined Watson-Crick and G-U wobble base pairing, there are still hundreds of noncanonical base pair types, triplex or quadruplex [53]. Other than the challenges from the modeling perspective, the statistical significance of these conclusions still remain elusive due to the paucity of 3D structure of protein-RNA complexes. Thus, it is crucial to perform more comprehensive structural analyses using a larger dataset to achieve greater statistical power and make more accurate inferences on the protein-RNA binding patterns when designing scoring functions in RPI prediction.

1.2.3 Current in silico methods of predicting RPI

In sharp contrast of advancements in RNA biology, there are only 1,585 protein-RNA complex structures deposited in PDB as of April 2014, which only represents a tiny island (<1.5%) compared with all macromolecular structure repository in PDB. Due to the technical issue in solving crystal/NMR structure of protein-RNA complex, high-throughput experiments to identify RPI are being developed to provide better understanding of the complex RPI networks, but they are usually expensive and time-consuming. As a consequence, there are immediate needs of developing computational tools for RPI prediction that help generate valuable hypotheses and prioritize insightful RPIs for experimental validation.

From the best of our knowledge, current computational methods of predicting RPI fall into two categories: sequence-based and structure-based methods. *RPISeq* [54] and *catRAPID* [55] are sequence-based methods. *RPISeq* utilizes machine learning classifiers to predict protein-RNA interaction propensity purely from sequence information, whereas *catRAPID* calculates the protein-RNA interaction propensity through combining various physiochemical properties, such as H-bond, vdW, secondary structure. Structure-based methods take advantage of 3D structures of protein and RNA, and employ molecular docking strategy to evaluate the structural complementarity based on RNA-protein statistical scoring function. For example, Péres-Cano et al. developed a new protein-RNA docking scheme in which FTDock was used to generate rigid-body binding modes and rescored by an in-house derived statistical amino acid-nucleotide potential [56]. Similarly, 3dRPC applied a novel protocol including two modules, RPDock and DECK-RP [52]. RPDock is a new docking procedure that discretizes molecules and charges, and considers geometric and electrostatic complementarities as well as stacking interactions. DECK-RP is a coarse-grained, knowledge-based statistical potential to evaluate the predicted RNA-protein complex, which takes into account the secondary structure and interface preferences of protein/RNA residues [52]. Other efforts on the development of protein-RNA statistical potentials, such as DARS-RNP, QUASI-RNP[57] and Li et al.[58], have resulted in comparable performances according to their benchmarks. However, either sequence-based or structure-based methods have its merits and pitfalls. Sequence-based method is based on simple assumption and thereby more robust, for example, using conjoint triad descriptors [59]; however, it could be sensitive to noise as it fails to discriminate the interface with other part of the molecule.

Structure-based method, on the other hand, restricts its application only for the protein / RNA targets that have 3D structure. Therefore, a method that balances the robustness and accuracy of RPI prediction is urgently needed.

Chapter 2: Computational modeling of RNA-small molecule interaction

2.1 Introduction

Due to the challenges we have described in Chapter 1.1, we think there is an immediate need of exploring current computational tools and implementing new ones to model RNA-ligand interaction more accurately, and prioritize compounds via virtual screening more effectively. Herein, we have benchmarked 5 popular docking programs, including GOLD 5.0.1 [60], Glide 5.6 [61], Surflex 2.415 [62], AutoDock 4.1 [63, 64] and rDock 2006.2 [24], and 11 scoring functions to explore their capability in RNA-small molecule docking. **Fig. 2.1** shows an overview of structure-based virtual screening pipeline. A typical structure-based virtual high-throughput screening (vHTS) can be divided in to three steps: sample ligand conformations (step 1), score and rank the poses for each molecule based on a scoring function (step 2), score and rank the molecules and estimate the relative binding affinity for the optimal pose provided by step 2 based on a second scoring function (step 3). The rescoring scheme is believed to improve the results when two scoring functions have complementary strengths: one is better at ranking poses and the other ranking actives [65]. Based on this "complementary" hypothesis, we comprehensively evaluated the docking

performances at these three levels, and explored exhaustively for the best docking-scoring-rescoring strategies using various statistical metrics. As a result, we proposed a rational workflow for structure-based modeling for RNA-targeted drug discovery for RNA, which has demonstrated a significant improvement of virtual screening enrichment in two independent benchmarks [66].

In a follow-up case study, we validated the effectiveness of our pipeline in which we have successfully identified small-molecule inhibitor that binds selectively to RNAs containing GA:UU internal loop motif. NMR validated the binding site specificity and the essential context adjacent to the motif. This tandem mismatch internal loop, $\begin{matrix} 5'GUGA3' \\ 3'CUAU5' \end{matrix}$ (or called GA:UU RNA), is a highly conserved motif in prokaryotic large ribosomal subunit (LSU) as a part of a conserved 58-nt fragment. It is the binding domain of ribosomal protein L11, and this thermodynamically destabilizing internal loop is crucial for binding of L11 [67]. The discovery of small-molecule binder targeting this rRNA motif has the potential to destabilize the L11 binding. From the druggability perspective, selective small molecule inhibitor targeting prokaryotic rRNA internal loop, such as A-site, has been proved an effective strategy of designing antibiotic drugs. However, the most thoroughly studied RNA-binding antibiotics, notably aminoglycosides, have very low bioavailability. Development of non-aminoglycoside antibiotics targeting bacteria rRNA will improve the pharmacokinetics profiles and provide possible solution to overcome drug resistance.

Here we hypothesize that RNA-small molecule docking composed of three independent steps, each of which needs a fine-tuned docking/scoring combination to maximize the predictive ability in a virtual screening scenario. In order to validate our hypothesis, we proposed several specific aims:

1. Benchmark open-source and commercially available docking/scoring method to identify best strategy for pose reproduction, pose ranking and active ranking.
2. Knowing the challenge in active ranking, optimize the scoring function so that the docking score has a better representation of the experimentally determined binding affinity for RNA-small molecule interaction.
3. Apply the derived structure-based drug discovery pipeline to a real-world problem: to identify novel inhibitors that bind selectively to GA:UU RNA motif.

2.2 Materials and Methods: Benchmarking, Development and Application

2.2.1 Benchmark datasets

Most of the currently published datasets are either too small or lack target diversity [20, 21, 23, 24, 26, 27]. Based on these datasets, we compiled our own dataset of high-resolution RNA-ligand complex structures by removing those low-resolution, redundant structures as well as those structures with critical structural defects. This resulted in a unique collection of 56 RNA-ligand complex structures with 36 high-resolution (<3.0Å) crystal and 20 NMR structures. Another issue of the published datasets was that over 65% of the ligands were aminoglycosides or low-affinity binders (e.g. spermine) [20]. To avoid the potential problems of overweighting any type of RNA ligand, we reduced the number of aminoglycosides and

low-affinity binders, but increased the number of high-affinity small molecules. Our curation encompassed a large variety of RNA targets including: RNA aptamers, prokaryotic and eukaryotic rRNA A-sites, ribozymes, riboswitches, and viral RNAs (TAR RNA, HCV IRES domain, etc.). These RNA-small molecule complexes are listed in **Table 2.1**.

We also compiled a second dataset which contains 45 RNA-ligand binding affinity values for benchmark currently available scoring functions and to derive RNA-specific docking scoring function (**Table 2.2**). Briefly, dissociation constant (K_d) or binding free energy values were carefully collected from literature, and we compared them with PDBBind database (http://www.pdbbind-cn.org) [69] and other reports/databases to ensure the consistency. If the variance between K_d values is within 10-fold difference, we calculated the average values; otherwise, data will be discarded. Notably, we used 2 μM as the K_d of gentamicin C1a-rRNA A-site complex (1BYJ) because this is the K_d under room temperature, instead of 0.01μM (K_d under 4°C) [70]. In addition, K_d for neomycinB-HIV-1 TAR RNA complex (1QD3) should be 5.9±4μM. The K_d values used in previous studies were for U24C TAR RNA mutant [21, 24, 71]. The binding free energy were converted from K_d using $\Delta G = RT\ln(K_d)$ under room temperature (300K).

Fig. 2.1. **An overview of structure-based virtual screening pipeline**. A typical virtual screening can be divided into three steps: for each candidate molecule, docking program should do conformational sampling (step 1) and select an optimal pose based on a scoring function (step 2). An additional scoring of the optimal pose for each molecule might be performed after pose selection to estimate the relative binding affinity (step 3). Finally, the molecules that have good predicted binding affinity will be prioritized for experimental validation.

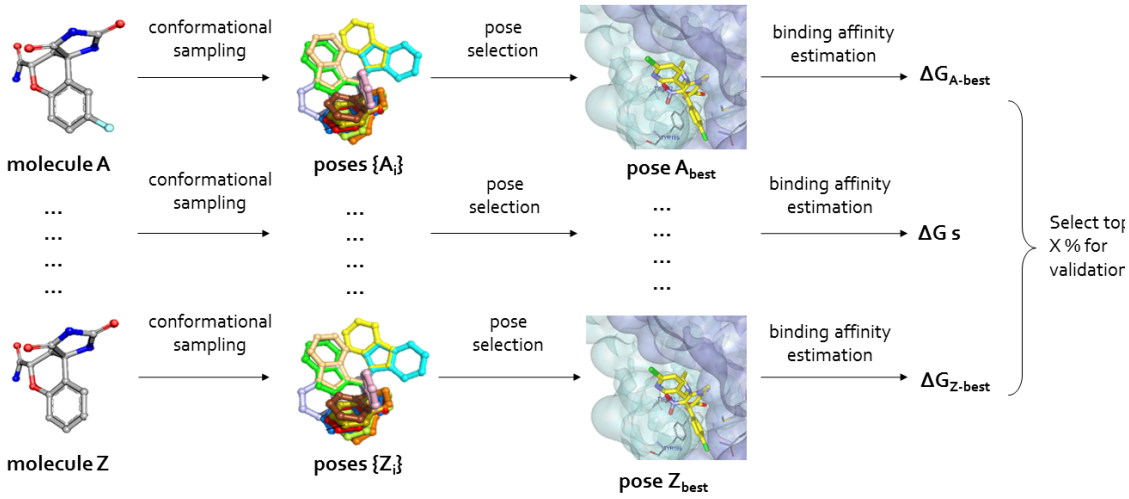


Table 2.1 List of 56 PDBs used in binding mode reproduction study

PDB ID	Res (Å)	Ligand	RNA
1F1T	2.8	N,N'-tetramethyl-rosamine	Malachite green aptamer RNA
1F27	1.3	Biotin	Biotin-binding aptamer RNA
1J7T	2.5	Paromomycin	Bacterial rRNA A-site
1NTB	2.9	Streptomycin	Streptomycin RNA aptamer
1YRJ	2.7	Apramycin	Bacterial rRNA A-site
2F4T	3	Designed antibiotics	Bacterial rRNA A-site
2FCZ	2.01	Ribostamycin	HIV-1 DIS Kissing loop
2ET8	2.5	Neamine	Bacterial rRNA A-site
2O3V	2.8	Paromamine derivative NB33	Human rRNA A-site
2OE8	1.8	Apramycin	Human rRNA A-site
1LC4	2.54	Tobramycin	Bacterial rRNA A-site
1MWL	2.4	Geneticin	Bacterial rRNA A-site
1U8D	1.95	Hypoxanthine	xpt-pbuX B. subtilis guanine riboswitch
1Y26	2.1	Adenine	Vibrio vulnificus adenosine riboswitch
2BE0	2.63	Paromomycin Derivative JS5-39	Bacterial rRNA A-site
1YKV	3.3	DAI	Diels-Alder ribozyme
2G5Q	2.7	Amikacin containing L-haba	Bacterial rRNA A-site
2GDI	2.05	Thiamine Diphosphate	Thiamine pyrophosphate-sensing riboswitch.
2GIS	2.9	S-Adenosylmethionine	S-adenosylmethionine riboswitch (T. tengcongensis)
3LA5	1.7	Azacytosine	Engineered A-riboswitch
3F2Q	2.95	Flavin mononucleotide	Flavin mononucleotide riboswitch
3DIL	1.9	Lysine	Thermotoga maritima Lysine riboswitch
2Z74	2.2	Alpha-D-glucose-6-phosphate	T. tengcongensis glmS ribozyme
2Z75	1.7	glucosamine 6-phosphate	T. tengcongensis glmS ribozyme
1ZZ5	3	Neomycin Derivative	rRNA A-site
3Q3Z	2.51	C-di-GMP	Clostridium acetobutylicum c-di-GMP-binding riboswitches
2ESI	3	Kanamycin A	Bacterial rRNA A-site
2FD0	1.8	Lividomycin	HIV-1 DIS Kissing loop
3NPQ	2.18	S-adenosylhomocysteine	Ralstonia solanacearum S-adenosyl-(L)-homocysteine (SAH) riboswitches
2PWT	1.8	L-HABA containing aminoglycoside	Bacterial rRNA A-site
3DVV	2	Ribostamycin	HIV-1 F DIS extended duplex
3GX2	2.9	Sinefungin	T. tengcongensis SAM-I riboswitch (variant)
1Y27	2.4	Guanine	Bacillus subtilis G-riboswitch xpt
3GX3	2.7	SAH	T. tengcongensis SAM-I riboswitch (variant)
3GX5	2.4	SAM	T. tengcongensis SAM-I riboswitch (variant)
3GX7	2.95	SAM	T. tengcongensis SAM-I riboswitch (double mutated variant)
1FMN	NMR	Flavin mononucleotide	FMN aptamer
1UUD	NMR	P14	HIV-1 TAR RNA

2KU0	NMR	ISI	HCV IRES domain IIa RNA
1AM0	NMR	AMP	AMP aptamer
1LVJ	NMR	PMZ	HIV-1 TAR RNA
1TOB	NMR	Tobramycin	antibiotic-RNA aptamer
1EHT	NMR	Theophylline	Theophylline-binding RNA
1BYJ	NMR	Gentamicin C1	Bacterial rRNA A-site
1PBR	NMR	Paromomycin	Bacterial rRNA A-site
1AKX	NMR	Arginine	HIV-2 TAR RNA
1FYP	NMR	Paromomycin	Human rRNA A-site
2KGP	NMR	Novantrone	tau pre-mRNA splicing regulatory element
1EI2	NMR	Neomycin	RNA major groove in Tau Exon 10 splicing regulatory element
1KOD	NMR	Citrulline (arginine derivative)	Citrulline aptamer
1QD3	NMR	Neomycin B in the minor groove	HIV-1 TAR RNA
1KOC	NMR	Arginine	arginine aptamer
1NEM	NMR	Neomycin B in the major groove	Neomycin B RNA aptamer
2TOB	NMR	Tobramycin	tobramycin-RNA aptamer
2KTZ	NMR	ISH	HCV IRES domain IIa RNA
1Q8N	NMR	Malachite green	Malachite green aptamer RNA

Table 2.2 Experimental binding free energy values used for benchmarking and optimizing score functions

PDB ID	Type	Binding free energy (kJ/mol)	Dissociation constant
1F1T [1]	Crystal	-42.23	$K_D \approx 0.04\mu M$
1F27 [1]	Crystal	-29.8	$K_D \approx 6.0\mu M$
1J7T [2]	Crystal	-38.47	$K_d = 0.2\pm0.042\mu M$
1NTB [1,2]	Crystal	-34.46	$K_d \approx 1\mu M$
1YRJ [1,2]	Crystal	-30.91	$K_d = 2\pm0.20\mu M/6.3\mu M$
2F4T [1,2]	Crystal	-32.49	$K_d = 2.2\pm0.1\mu M$
2FCZ [1,2]	Crystal	-28.62	$K_d = 10.4\pm1.4\mu M$
2ET8 [1,2]	Crystal	-27.99	$K_d = 7.8\mu M/19\pm1\mu M$
2O3V [1,2]	Crystal	-30.21	$K_a = 1.8\pm0.1\times10^5\ \mu M^{-1}$
2OE8 [1]	Crystal	-36.19	$K_d = 0.5\mu M$
1LC4 [1,2]	Crystal	-33.06	$K_d = 1.5\mu M/2\pm0.22\mu M$
1U8D [1]	Crystal	-35.24	$K_d = 0.732\mu M$
1YKV [1]	Crystal	-28.72	$K_d \approx 10\mu M$
3LA5 [1]	Crystal	-34.46	$K_d = 1\pm0.016\mu M$
3DIL	Crystal	-40.2	$K_d = 0.10\pm0.03\mu M$ (with K^+ and Mg^{2+})
3Q3Z [1]	Crystal	-49.72	$K_d = 0.0022\pm0.0002\mu M$
2ESI	Crystal	-27.25	$K_d = 18\mu M$
2FD0 [2]	Crystal	-43.04	$K_d = 0.032\pm0.007\mu M$
3GX3	Crystal	-23.83	$K_d = 71\pm2\mu M$
3GX5	Crystal	-39.55	$K_d = 0.13\pm0.01\mu M$
3GX7	Crystal	-25.89	$K_d = 31\pm1\mu M$
1FMN [1]	NMR	-35.9	
2KU0 [1,2]	NMR	-32.08	$K_D = 2.6\mu M$
1AM0	NMR	-28.5	
1LVJ	NMR	-39.97	
1TOB [2]	NMR	-52.2	
1EHT [1]	NMR	-36.5	
1BYJ [1,2]	NMR	-32.73	$K_d = 2.0\mu M$ (room temperature) / $0.01\mu M$ (4°C)
1PBR [1,2]	NMR	-38.2	
1EI2 [1,2]	NMR	-34.23	
1KOD [1,2]	NMR	-23.8	
1QD3	NMR	-30.03	$K_D = 5.9\pm4\mu M$; $K_D = 0.92\mu M$ (U24C mutant)
1KOC [1,2]	NMR	-24.1	
1NEM [1]	NMR	-39.9	

2TOB [2]	NMR	-51.2	
2KTZ [1,2]	NMR	-28.98	K_d = 9µM
1Q8N [1]	NMR	-35.02	K_D = 0.8µM
3SD1 [3]	Crystal	-27.25	$K_{D,app}$ = 18±1µM
2YGH [3]	Crystal	-37.38	K_d = 0.31±0.06µM (G2na mutation)
3SKI [3]	Crystal	-40.20	K_D = 0.1±0.01µM (20mM Mg^{2+})
2L94 [3]	NMR	-19.78	$K_{d,app}$ = 360±26µM
3GER [3]	Crystal	-34.75	K_D = 0.89±0.06µM
2G5K [3]	Crystal	-36.19	K_d = 0.5µM
2BEE [3]	Crystal	-40.20	K_d = 0.1µM
2BE0 [3]	Crystal	-39.55	K_d = 0.13µM

2.2.2 Molecular docking and decoy generation

Throughout Chapter 1, we denoted "A:B" as the method that "docking using A program and scoring with B scoring function". RNA molecules and ligands were prepared using Protein Preparation Wizard (Maestro). For NMR structures, we used the average structure and energy minimized. All of RNA phosphates were manually deprotonated in case of software errors. The ligands were protonated/deprotonated using Epik (Schrödinger) at PH 7.0 [72]. If RNA has symmetric binding sites and identical ligands, the region with the lowest B-factors was retained. The ligands were minimized, and molecular docking and rescoring were performed using the similar approaches as previously described [66]. Briefly, we benchmarked five docking programs (GOLD 5.0.1, Glide 5.6, Surflex v2.415, AutoDock 4.1 and rDock 2006.2) combined with their native scoring functions to generate 10 poses using the parameters in **Appendix 1**. In order to ensure the high diversity and quality of the conformational decoys, we employed GOLD:GOLD Fitness to generate 100 conformational decoys for each RNA-compound complex using the tuned parameters for genetic algorithm.

2.2.3 Evaluation of pose reproduction

Both RMSD between experimental structures and predicted docking poses and pose ranking were considered. To simplify the expression, we defined *C(x, y)* as the criterion that "at least one pose (RMSD < yÅ) was predicted within the top x poses". To evaluate the overall ability of docking/scoring programs to reproduce experimentally determined binding mode, we implemented volume under the surface (VUS) metric to describe overall performance of pose reproduction. VUS was calculated as the sum of the volume of all triangular prisms under this

surface. Briefly, a series of coordinates were obtained based on their RMSD cutoff (X dimension), ranking cutoff (Y dimension), and the number (Z dimension) of successfully reproduced structures satisfying *C(x, y)*. RMSD cutoff had interval of 0.5Å, and that for ranking was 1. The surface was made by connecting any two adjacent points and then partitioned into a series of triangles. Any of these triangles and their projections on the *XY* plane was used to define the triangular prism unit. Detailed calculation of the volume of each triangular prism unit and VUS were demonstrated in **Appendix 2**. The ideal VUS was calculated as 10(RMSD cutoff)×9(rank cutoff)×56(number of targets).

2.2.4 Evaluation of pose ranking

For each RNA-compound complex, we generated 100 decoys to the corresponding RNA as we described in 1.2.2. Together with the native pose, we obtained 101 RMSD-docking score data points for each RNA-ligand pair. For native pose ranking study, we scored these 101 poses using different scoring functions as aforementioned. The ranking of native poses for 56 targets were calculated, and we calculated the recovery curves as the ranking cutoffs (X axis) against the cumulative number of targets (Y axis) in which the ranking of native pose was smaller than the ranking cutoff. Meanwhile, spearman's rank correlation coefficient was used to evaluate the ranking capability. To make the docking scores positively correlated with RMSD (the higher the scores, the higher the RMSD), we used the negative value of GOLD Fitness, ChemScore, ASP and Surflex-dock scores. If a pose was assigned a score with the absolute value more than 1000 (outliers), this RMSD-score pair will be excluded. The Spearman's rank correlation coefficient (ρ) was computed using

$\rho = \frac{\sum_i (r_{RMSD,i} - r_{RMSD}^{avg})(r_{score,i} - r_{score}^{avg})}{\sqrt{\sum_i (r_{RMSD,i} - r_{RMSD}^{avg})^2 \sum_i (r_{score,i} - r_{score}^{avg})^2}}$, where $r_{RMSD,i}$ and $r_{score,i}$ are the rankings of the RMSD and score for the pose *i*, and we took the average of the ranks for tied values. r_{RMSD}^{avg} and r_{score}^{avg} are the average ranks of RMSD and score for 101 poses. We classified the resulted 56 ρ values (calculated from 56 RNA-ligand complexes) for each scoring function into three groups based on the widely-used criteria: weak correlation: $\rho<0.3$, moderate correlation: $0.3\leq\rho<0.5$, strong correlation: $\rho\geq0.5$.

2.2.5 Evaluation of virtual screening

Two different targets were assessed, bacterial 16S rRNA A-site (representing open and flexible binding site, PDB ID: 1J7T [73]) and lysine riboswitch (representing closed and rigid binding site, PDB ID: 3DIL [74]). We collected 75 known rRNA inhibitors including 34 drug-like small molecules from the Foloppe dataset [2] and 31 aminoglycoside mimetics from the Zhou dataset [3]. Additionally, we obtained 11 aminoglycoside inhibitors which have the crystal structures in complex with the bacterial rRNA A-site (1J7T, 1YRJ, 2F4T, 2ET8, 1LC4, 1MWL, 2BE0, 2G5Q, 2ESI, 2PWT and 1BYJ). For virtual screening against lysine riboswitch, we collected 14 compounds including 7 known inhibitors and 7 experimentally validated inactives [7]. In order to avoid artificial enrichment [75], a focused library containing 942 drug-like and positively charged decoys was generated from MayBridge database. We assumed this randomly constructed decoy library does not include or include very few active compounds as previous studies did. The area under the curve (AUC) for the receiver operating characteristic (ROC) curve was used to assess the virtual screening enrichment.

2.2.6 Evaluation of docking score-binding affinity correlation

The Pearson correlation coefficients (R^2) between these docking scores and their corresponding binding affinities were calculated. Three common outliers, 1LVJ, 1TOB and 2TOB, were excluded during analysis, as they contained many unfavorable steric clashes in the NMR structures.

2.2.7 RNA-specific scoring function optimization

The weak correlation between docking score and binding affinity might be because most of the current scoring functions were derived from protein-ligand complexes. To implement RNA-specific scoring function, we optimized the energetic coefficients in AutoDock4.1 scoring function using dataset provided in **Table 2.2**. This empirical scoring function was shown as Equation (2). The parameters (A, B, C, D, S, V) were obtained from default AutoDock4 scoring function [64]. We optimized the coefficients, W_{vdW}, W_{hbond}, W_{elec}, W_{sol} and W_{tors} using multiple linear regression.

$$\Delta G_{bind} = W_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}\right)+W_{hbond}\sum_{i,j}\xi(t)\left(\frac{C_{ij}}{r_{ij}^{12}}-\frac{D_{ij}}{r_{ij}^{10}}\right)+W_{elec}\sum_{i\ j}\frac{q_i q_j}{\varepsilon(r_{ij})r_{ij}}+W_{sol}\sum_{i,j}\left(S_i V_j + S_j V_i\right)e^{\frac{-r_{ij}^2}{2\sigma^2}}+W_{tors}N_{tors}$$

Besides R^2, we calculated leave-one-out (LOO) cross-validation correlation coefficients (Q^2) and validated against an external test set consisting of eight complexes to evaluate the predictive power of our new scoring function.

2.2.8 MD simulations of GA:UU RNA-inhibitor complex

All simulation systems were set up using GROMACS 5.05 [76], using a similar protocol published previously [77]. The topology and charges of inhibitors were prepared using Gaussian09 at B3LYP/6-311++G(d,p) level of theory on Texas Advanced Computing Center (TACC). We used ff99bsc0 force field [78, 79] for RNA and general AMBER force field (GAFF) for inhibitor, prepared by ACPYPE [80]. The RNA-inhibitor complex was solvated in TIP3P water and neutralized with sodium ions. The simulation boxes were prepared so that the no RNA or inhibitor atom was within 14Å away from the edge. The system was minimized and equilibrated for 2 ns before production runs. The production simulations were performed for 660 ns, with constant pressure maintained by Berendsen barsostat (1 bar), constant temperature maintained by Berendsen thermostats (300K), LINCS, smooth particle mesh Ewald, 10 Å cutoff for short-range interactions, and 2-fs time step for bonded, van der Waals and short-range Coulomb interactions. Snapshots were taken every 20ps for further analysis.

2.2.9 Preparation of RNA samples

A total of five RNA constructs were prepared in order to evaluate the binding specificity (canonical base pairs are italic characters):

RNA1 (wildtype): $\begin{matrix} 5' \text{GGGCUGUGAUGCUU} \\ 3' \text{CCCGACUAUACGGC} \end{matrix}$

RNA2 (miR-328): $\begin{matrix} 5' \text{GGGUGGUGGAUUUU} \\ 3' \text{CCCACUUACUAAGC} \end{matrix}$

RNA3 (mutU5A): $\begin{matrix} 5' \text{GGGCAGUGAUGCUU} \\ 3' \text{CCCGACUAUACGGC} \end{matrix}$

RNA4 (mutU5A-ΔAU): $\begin{matrix}5'\text{GGGCAGUGAGCUU}\\3'\text{CCCGACUAUCGGC}\end{matrix}$

RNA5 (miR-10b): $\begin{matrix}5'\text{GGAUACCCUGUACUU}\\3'\text{CCUAAGGGG-AUGGC}\end{matrix}$

These RNAs were prepared by *in vitro* transcription with T7 RNA polymerase either unlabeled or $^{13}C/^{15}N$ -labeled 5'-NTPs (nucleoside triphosphates), and purified using the standard protocol described previously [81]. The integrity of the RNA molecules was evaluated using denaturing PAGE.

2.2.10 Nuclear magnetic resonance (NMR)

Spectrums of the RNA and the DMSO (solvent) were used as controls. All NMR spectra were acquired on Varian Inova 600 and 800 MHz spectrometers equipped with cryogenically cooled ^{1}H-[^{13}C, ^{15}N] probes and solvent suppression was achieved using binomial read pulses, as previously described [81]. 2D ^{13}C-^{1}H HSQC (Heteronuclear Single Quantum Coherence) spectra were collected to identify ^{13}C-^{1}H chemical shifts. NMR spectra were processed and analyzed by Felix 2007 (Felix NMR Inc., San Diego, CA). Peaks in the samples with the RNA and small molecules were compared to the control spectra to predict RNA-compound interactions.

2.3 Results: Benchmarking and optimizing docking method for RNA target

2.3.1 GOLD:GOLD Fitness and rDock:rDock_solv are the best pose generators

We first benchmarked the docking and scoring combinations for their ability to reproduce the ligand binding pose similar to the experimentally determined binding mode. An ideal RNA

docking method should be able to perform a thorough conformational sampling and identify at least one near-native pose. **Table 2.3** showed that, if we arbitrarily employed *C(5, 3.0)* (the top 5 pose includes at least one near-native pose with RMSD<3.0Å) to define a successful docking case, GOLD:GOLD Fitness and rDock:rDock_solv outperformed other methods, both with 73.21% success rate. Additionally, GOLD:ChemScore, GOLD:ASP, Glide:GlideScore(SP), Glide:Emodel(SP) and rDock:rDock obtained more than 50% docking success rate. In contrast, the success rates for Glide:GlideScore(XP), Glide:Emodel(XP), Surflex and AutoDock4.1 (default) were low, ranging from 30.36% to 44.64%. All programs, especially AutoDock4.1 and Surflex, had weak performance (<60%) on flexible and extensively-charged aminoglycosides. When more stringent criteria = *C(3, 1.5)* was used, the accuracy decreased but GOLD:GOLD Fitness and rDock:rDock_solv remained as the best methods (>40%). When compared with rDock:rDock_solv, the GOLD:GOLD Fitness achieved better performance for the pose reproduction on aminoglycosides-RNA complexes such as 1J7T, 2FCZ, 2BE0, 1NEM and 2TOB, whereas rDock:rDock_solv produced more accurate binding modes for drug-like ligand such as 2Z74, 2Z75, 1EHT and 1AKX. The detailed results (scores, RMSD and statistics) are available from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3869234.

To better demonstrate the relationship between the pose reproduction accuracy and RMSD or ranking, we illustrated our results with **Fig. 2.2A**, in which the heavy-atom RMSD and the ranking of pose were considered simultaneously. VUS represents the overall ability of reproducing near-native binding modes. It showed that GOLD:GOLD Fitness achieved the

best VUS (78.11%), while rDock:rDock_solv was the second best (**Table 2.3** and **Fig. 2.2A**). We proposed to employ the contour of 50% success rate to guide the pose selection in RNA docking: if one aims to cover at least one near-native pose (RMSD<3.0Å) with 50% probablity, at least top five poses should be kept when using GOLD:GOLD Fitness. In contrast, we should keep at least top 20 poses to achieve 50% success for Surflex and AutoDock 4.1 (**Fig. 2.2B**). From these assessments, we suggest that GOLD:GOLD Fitness and rDock:rDock_solv be the best methods for pose reproduction in RNA small molecule docking.

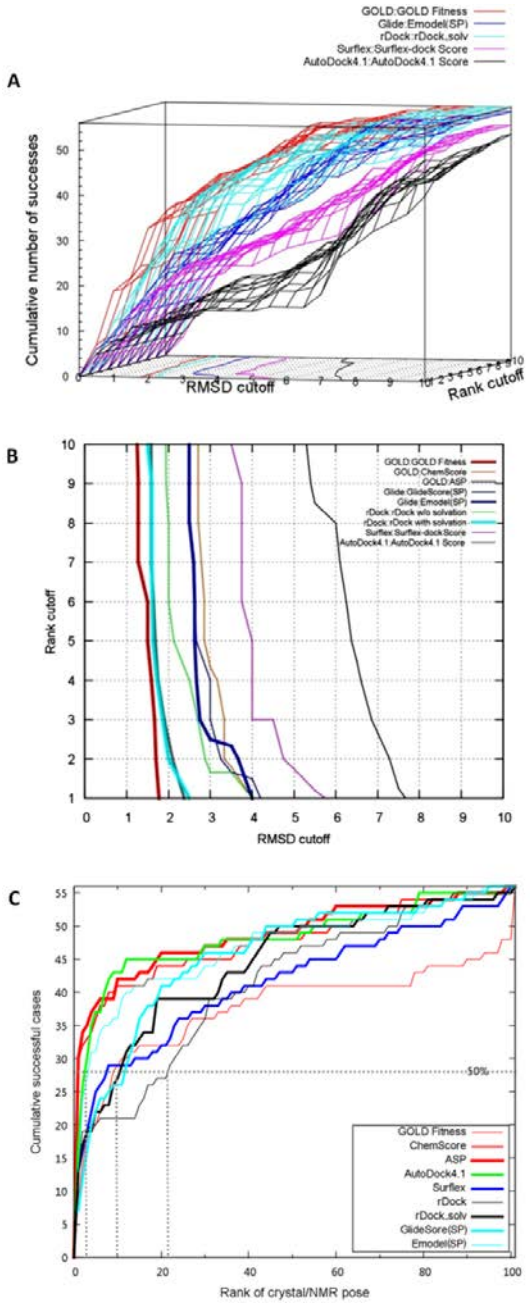
As expected, we observed that the average docking accuracy on crystal structures was higher than that on NMR structures for all of 11 current docking/scoring combinations (58.84% versus 42.27%, $p = 0.06$). Not surprisingly, the pose reproduction performance on small-molecule RNA ligands was remarkably better than that on flexible aminoglycosides (64.55% versus 39.51%, $p<0.01$). Among the failed cases (defined as two or less docking programs are able to satisfy *C(5, 3.0)*), five are crystal structures (2O3V, 2BE0, 2FD0, 2PWT and 2Z75) and seven are NMR structures (1UUD, 1LVJ, 1TOB, 1AKX, 1EI2, 1KOD and 1QD3). We found that the current methods were usually less accurate on RNA complexes containing large aminoglycosides (e.g. lividomycin, paromomycin, etc.), weak RNA binders (e.g. arginine and citrulline), or phosphate-containing hydrophilic ligands (glucosamine 6-phosphate). As negatively-charged moieties can form specific interactions with RNA phosphates in the presence of metal ions acting as the “metal bridge”, such as 2GDI and 2Z74, we tried docking with consideration of metal ions. As expected, we could significantly

improve the pose prediction of the diphosphate tail of thiamine diphosphate in 2GDI when the Mg^{2+} ion was taken into account as part of RNA targets.

Table 2.3 Performances of binding mode reproduction. 56 RNA-ligand complexes list in Table 2.1 were benchmarked using different docking/scoring combinations. The values in the brackets indicated the total number of structure complexes in the category. The values before the parentheses were the results satisfying *C(5, 3.0)*, and the values in the parentheses were for *C(3, 1.5)*.

Docking program	**GOLD 5.0.1**			**AutoDock 4.1**	**Surflex 2.415**	
Scoring function	GOLD Fitness	ChemScore	ASP	Autodock4.1 Score	Surflex-dock Score	
Aminoglycoside [26]	18 (9)	13 (3)	15 (9)	1 (1)	4 (2)	
Small Molecule [30]	23 (15)	17 (10)	22 (15)	16 (9)	21 (13)	
X-ray crystal [36]	29 (19)	26 (13)	29 (21)	13 (8)	17 (11)	
NMR [20]	12 (5)	4 (0)	8 (3)	4 (2)	8 (4)	
Total [56]	41 (24)	30 (13)	37 (24)	17 (10)	25 (15)	
Overall Success Rate %	73.21 (42.86)	53.57 (23.21)	66.07 (42.86)	30.36 (17.86)	44.64 (26.79)	
VUS %	78.11	65.48	70.17	43.3	55.22	
Score-binding affinity correlation R^2	0.25	0.03	0.29	0.22	0.05	
Docking program	**Glide 5.6**				**rDock 2006.2**	
Scoring function	GlideScore (SP)	Emodel (SP)	GlideScore (XP)	Emodel (XP)	rDock	rDock_solv
Aminoglycoside [26]	12 (3)	13 (3)	4 (2)	4 (2)	13 (6)	16 (8)
Small Molecule [30]	18 (13)	18 (13)	16 (11)	15 (11)	21 (13)	25 (15)
X-ray crystal [36]	20 (9)	21 (9)	12 (8)	12 (8)	24 (16)	29 (17)
NMR [20]	10 (7)	10 (7)	8 (5)	7 (5)	10 (3)	12 (6)
Total [56]	30 (16)	31 (16)	20 (13)	19 (13)	34 (19)	41 (23)
Overall Success Rate %	53.57 (28.57)	55.36 (28.57)	35.71 (23.21)	33.93 (23.21)	60.71 (33.93)	73.21 (41.07)
VUS %	65.41	66.01	NA	NA	63.09	73.13
Score-binding affinity correlation R^2	0.1	0.14	NA	NA	0.15	0.18

Fig. 2.2. Analysis of the binding mode reproduction performance (A). The cumulative success rate in 3D representation. Only the scoring functions which obtained the highest VUS for each docking method were selected for illustration. The contour on the XY (RMSD-Rank) plane represented the 50% (Z=28) success rate (the binding mode can be reproduced for 50% of RNA-ligand complexes); (B). The 50% success contour (Z=28) for all available scoring functions (GlideScore (XP) and Emodel (XP) were not included due to the unavailability of VUS values). (C). The cumulative success rate for 56 RNA-ligand complexes based on the ranking of X-ray/NMR determined poses against 100 decoys. The 50% success line and the corresponding rankings to achieve 50% success were shown as dots.



2.3.2 ASP: best pose selector

Native pose ranking evaluates the ability to differentiate the experimental pose from the decoy poses for different scoring functions. This was assessed by investigating two metrics: the ranking of native poses, and the Spearman's correlation between scores and RMSDs. Since GOLD:GOLD Fitness outperformed other docking programs on the coverage of the near-native poses as aforementioned, it was utilized to generate 100 decoys for each complex hoping to obtain a decoy set with a smooth transition from near-native binding mode to unfavorable one. We investigated whether a given scoring function could obtain the highest rankings for experimentally determined poses. Analogous to IC50 (in assessing biological activity), we used 50% success rate to evaluate the performance of different docking/scoring methods. As demonstrated in **Fig. 2.2C**, the 50% success rate line (dashed) clustered scoring functions into three groups: ASP, ChemScore, AutoDock4.1 Score and Emodel (SP) were the first group; the second group included other scoring functions, except rDock which ranked the lowest as the 3rd group. **Fig. 2.2C** indicated that GOLD Fitness has 50% of possibility to rank the native ligand conformation within top 10% of the predicted poses, whereas for ASP, ChemScore, AutoDock4.1 Score and Emodel (SP), this value reduced to top 5%. The native pose ranking performance for different docking/scoring schemes varied with different types of RNA structure. For example, most programs performed significantly better for crystal structures than NMR structures (69.14% versus 38.89%, $p < 0.01$) with the top 10 as the cutoff to define a successful ranking case. Surprisingly, ASP was remarkably better in crystal structure ranking, in which only two targets (2O3V and 3DIL) failed, while AutoDock4.1 outperformed others on ranking NMR structures. Taken together, these data suggested that

RNA targets with different structural resolutions should be rescored with respective appropriate scoring functions (e.g., ASP or AutoDock4.1) after the initial step of docking with GOLD:GOLD Fitness or rDock:rDock_solv.

For score-RMSD correlation study, we grouped the performances for 56 cases based on the strength of correlation for each scoring function. Consequently, ASP, GlideScore (SP) and Emodel (SP) were the best three scoring functions which had most cases with moderate or strong correlations (**Table 2.4**). rDock, rDock_solv and Surflex-dock scores obtained fair performance, which could derive weak or strong correlations for more than 1/3 of cases. Surprisingly, GOLD Fitness could not achieve satisfactory performance to enrich the near-native ligand conformations (44 cases obtained the weak correlations) (**Table 2.4**). Combined with the native pose ranking analysis, these results demonstrated that other scoring functions such as ASP could enrich the near-native poses when applied to decoy poses generated by GOLD:GOLD Fitness.

Table 2.4. Score-RMSD Spearman's rank correlations. The values indicated the number of RNA-ligand complexes fit in each correlation category (Weak: $\rho<0.3$, Moderate: $0.3\leq\rho<0.5$, Strong: $\rho\geq0.5$). Top 3 scoring functions are in bold.

	Weak	Moderate	Strong
GOLD Fitness	44	5	7
ChemScore	41	7	8
ASP	33	15	8
GlideScore (SP)	31	15	10
Emodel (SP)	29	14	13
Surflex-dock Score	38	12	6
AutoDock4.1 Score	40	5	11
rDock	35	12	9
rDock_solv	36	12	8

2.3.3 ASP rescoring improves the pose generation

As we have identified ASP as the most robust scoring function for pose ranking, we validate whether rescoring with ASP is able to improve the identification of near-native binding poses generated by GOLD:GOLD Fitness or rDock:rDock_solv without artificial parameters designed for decoy generation. **Fig. 2.3** showed the average RMSD-ranking relationship of ASP rescoring based on GOLD:GOLD fitness or rDock:rDock_solv predicted poses. Obviously, after ASP rescoring, low-RMSD poses were more likely to appear in top tiers (top 5) compared with using either GOLD:GOLD fitness or rDock:rDock_solv alone. For GOLD:GOLD fitness, the number of complexes satisfying *C(5, 3.0)* increased from 41 to 44, while this number for *C(3, 1.5)* increased from 24 to 30, compared to original GOLD:GOLD Fitness performance. Specifically, we observed that the best RMSD in top 5-scored docking conformations of 2GDI, 2Z74, 2PWT and 1ZZ5 was significantly reduced (below 3.0Å) after ASP rescoring. In contrast, GOLD:GOLD Fitness alone failed to identified the near-native conformation for these complexes (**Fig. 2.4**). Furthermore, VUS increased from 78.11% to 79.18%. Compared with the docking accuracy using GOLD:GOLD Fitness alone, the average RMSD for the top-scored conformations was further reduced to 2.61±0.38Å (**Fig. 2.3 (up)**). Similarly, ASP rescoring improved VUS from 73.13% to 75.24% and the average RMSD of top-scored poses was reduced to 2.92±0.49 Å (**Fig. 2.3 (down)**). Combined with native pose ranking and RMSD-score correlation results, our results confirmed that ASP has the best ability for pose ranking, and ASP rescoring can significantly enrich the near-native poses generated by GOLD:GOLD Fitness or rDock:rDock_solv for pose reproduction purpose in RNA-ligand docking.

Fig. 2.3. ASP rescoring improves the ranking of poses (overall statistics). (Up) ASP rescoring based on the poses generated by GOLD:GOLD_Fitness (Down) ASP rescoring based on the poses generated by rDock:rDock_solv.

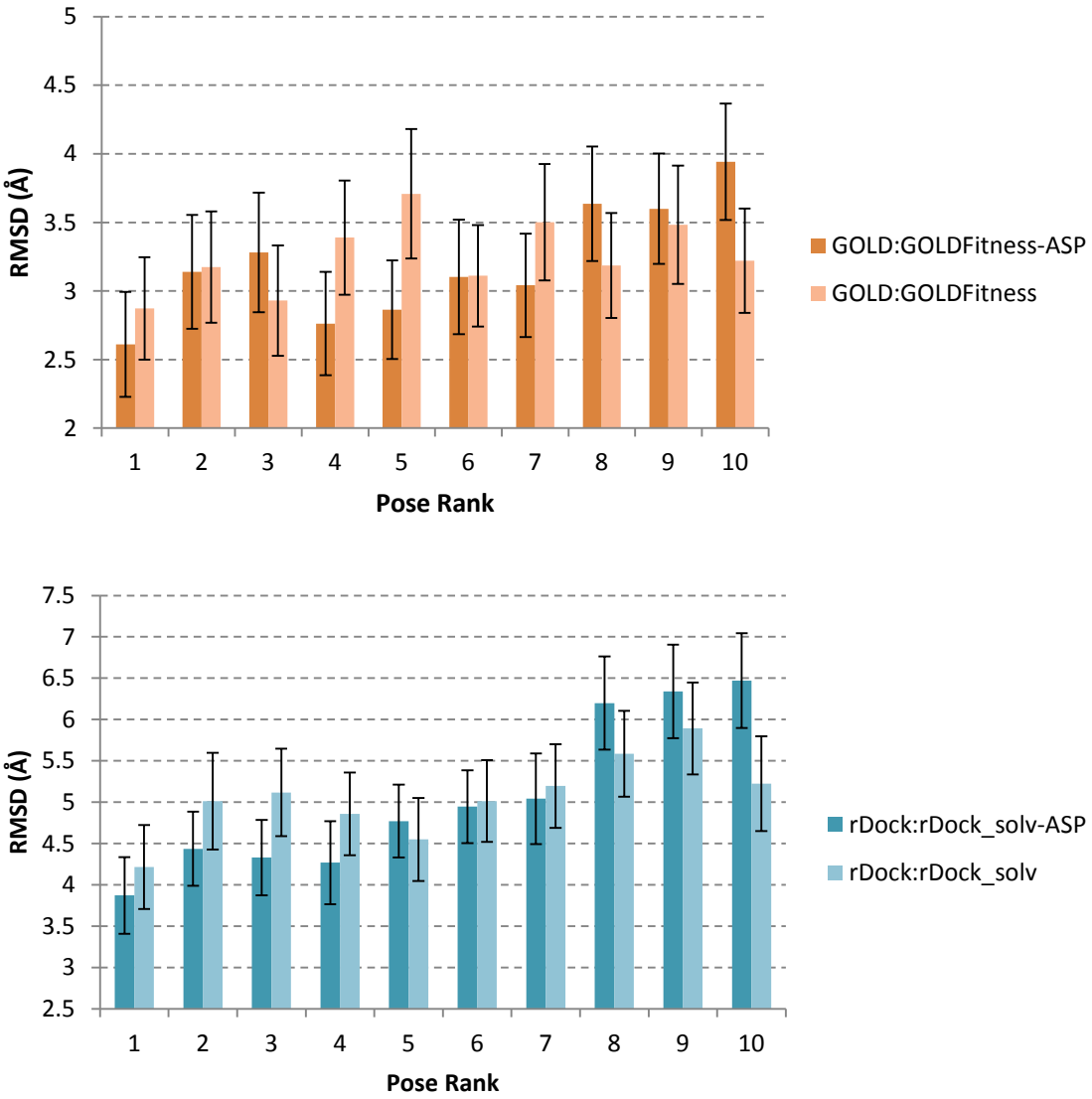
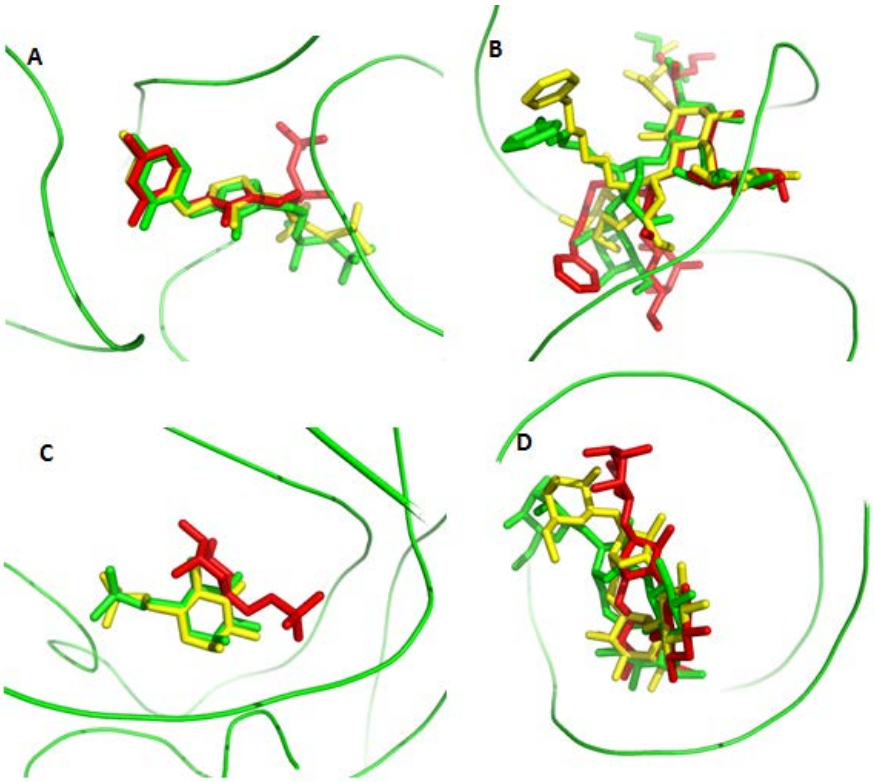


Fig. 2.4. **ASP rescoring improves the ranking of poses (molecular view).** Experimental structures were in green (RNAs in ribbons, ligands in sticks). Only the docking conformation with the lowest RMSD selected from the top five-scored poses were shown. GOLD:GOLD Fitness poses were colored red, while ASP rescored poses are colored yellow. (A) 2GDI; (B) 2PWT; (C) 2Z74; (D) 1ZZ5.



2.3.3 Improved score-binding affinity correlation by iMDLScores

Here we benchmarked the score-binding affinity correlation to assess the ability of scoring function to differentiate the binder against the non-binder. Surprisingly, we found all existing scoring functions received poor correlations ($R^2 < 0.3$) (**Table 2.3** and **Fig. 2.5**). To improve the correlation with the experimentally measured binding affinity, we developed new scoring functions, iMDLScore1 and iMDLScore2, using our RNA-ligand binding free energy datasets (**Table 2.2**). This was done by optimizing AutoDock4.1 scoring terms, W_{vdw}, W_{hbond}, W_{elec}, W_{sol}, W_{tors}, using multi-linear regression (MLR). We derived iMDLScore1 using the full dataset, in which the contributions of those scoring terms are 0.1460 for vdW, 0.0745 for hbond, 0.0559 for electrostatic, and 0.3073 for torsions (**Table 2.5**). iMDLScore1 achieved a significantly better correlation ($R^2 = 0.70$) between the docking scores and binding affinities. When iMDLScore1 was further validated against an external test set consisting of eight complexes, the $R^2 = 0.82$, and the root-mean-square error (RMSE) of prediction = 4.09kJ/mol (**Fig. 2.6A**).

A known challenge in RNA virtual screening is to enrich the actives from a focused library with positively-charged molecules because most RNA binders are potentially positively charged. To overcome this problem, we derived a second scoring function, iMDLScore2, with a dataset containing 18 complexes with only positively charged ligands. For iMDLScore2, the contribution are 0.1634 (vdW), 0.2436 (hbond), 0.2311 (electrostatic), and 0.2212 (torsion) (**Table 2.5**). Interestingly, R^2 and Q^2 (leave-one-out cross validation R^2) for the training set reached 0.79 and 0.62, and R^2 (test set) = 0.76. RMSE of prediction (4.35kJ/mol) was

comparable to that of iMDLScore1 (**Fig. 2.6B**). Q^2, R^2 and RMSE of prediction indicated the better predictive power of RNA-ligand binding affinities by both iMDLScore1 and iMDLScore2, compared with any other existing scoring functions.

Fig. 2.5. **Binding free energies-score correlation for ASP, GOLD_Fitness, AutoDock4.1 Score (default)**. Three outliers, 1TOB, 2TOB and 1LVJ, were highlighted in rectangles.

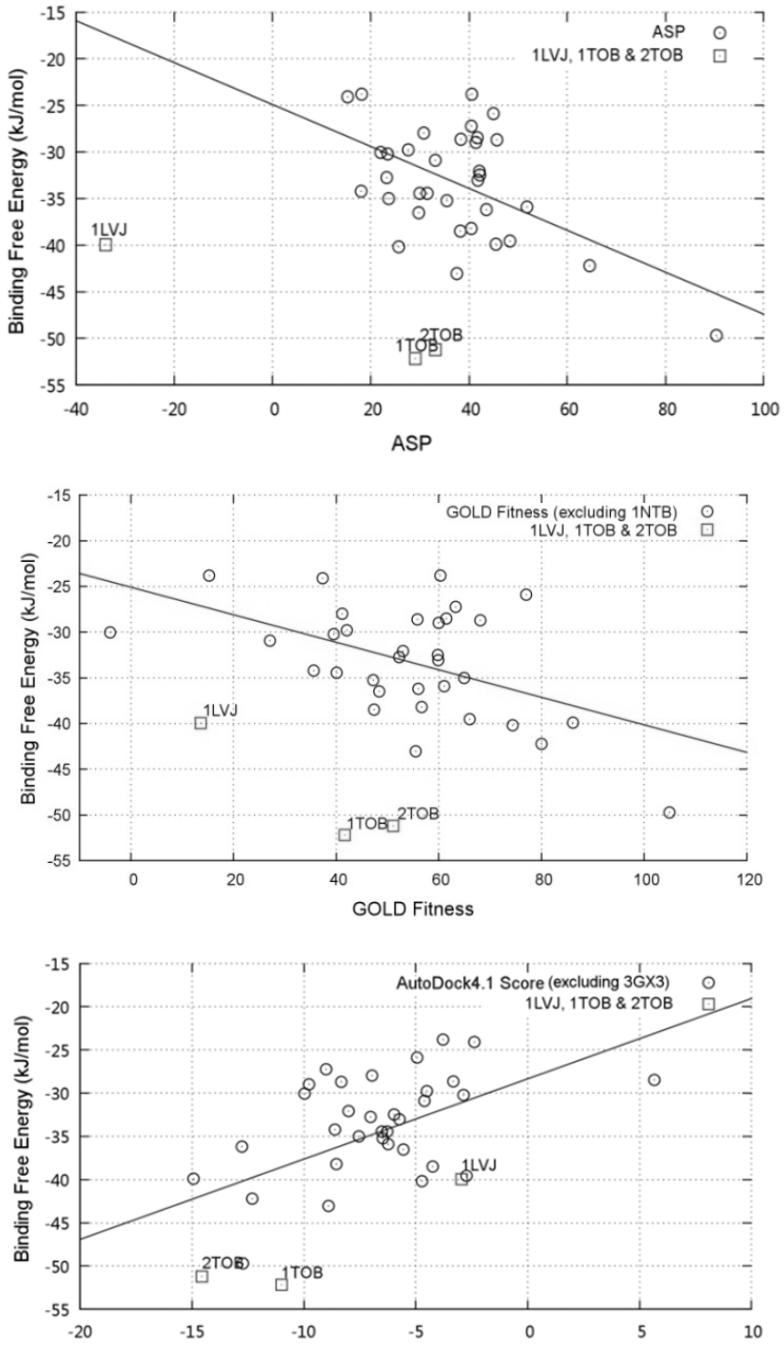


Fig. 2.6. Score-binding affinity correlation for iMDLScores. (A) iMDLScore1. (B) iMDLScore2.

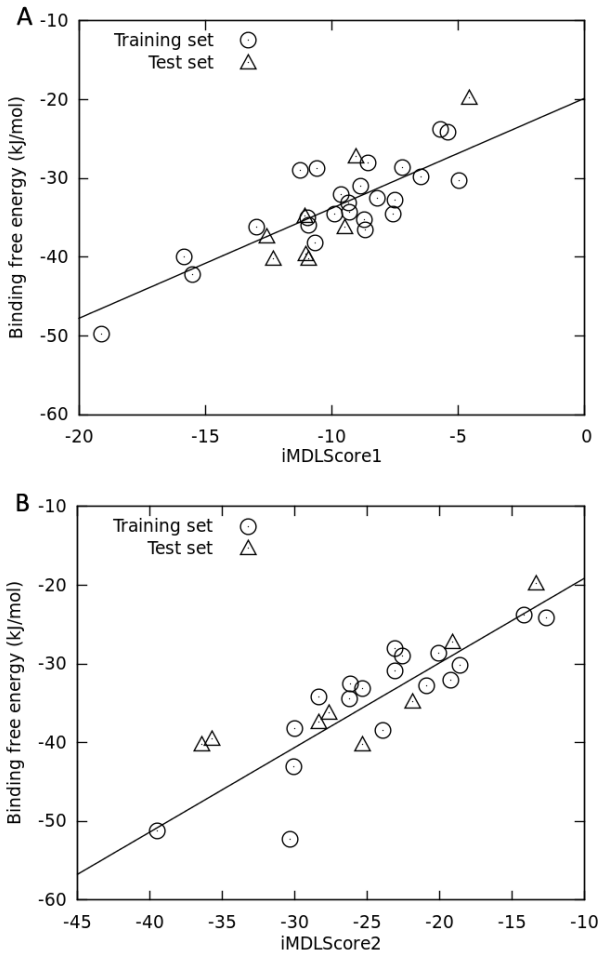


Table 2.5. Contributions of AutoDock energetic terms and associated performances in binding affinity correlation study.

Parameter	Default	iMDLScore1	iMDLScore2
vdW	0.1662	0.146	0.1634
hbond	0.1209	0.07451	0.2436
electrostatic	0.1406	0.05593	0.2311
desolvation	0.1322	0	0
torsion	0.2983	0.3073	0.2212
No. of complexes as training set	NA	25	18
R^2 (training set)	0.22	0.70	0.79
LOO Q^2 (training set)	NA	0.44	0.62
R^2 (test set)	NA	0.82	0.76
RMSE of prediction (kJ/mol, test set)	NA	4.09	4.35

2.3.4 Novel three-step virtual screening scheme improves the enrichment

Our ultimate goal is to identify an optimal pipeline for vHTS against RNA targets. In our benchmark, the ROC AUCs for Foloppe dataset are around 0.6 for both GOLD:GOLD fitness and rDock:rDock_solv, whereas the ROC AUCs for lysine riboswitch decoys are 0.82 and 0.86 for GOLD:GOLD fitness and rDock:rDock_solv, respectively (**Table 2.6**). As expected, three-step virtual screening, namely docking – rescoring (poses) – rescoring (compounds), could significantly improve the virtual screening enrichment in both cases. For Foloppe dataset, the enrichment was significantly increased by rescoring either rDock:rDock_solv or GOLD:GOLD Fitness generated poses using iMDLScore2 (AUC=0.74 and 0.69, compared with 0.61 and 0.58 without rescoring) (**Fig. 2.7**). For lysine riboswitch, however, all AutoDock-related could not obtain as good AUC (AUC <0.85) as other rescoring schemes (AUC>0.95) (**Table 2.6**). Additionally, we investigated whether any rescoring scheme could improve the differentiation of the seven known lysine riboswitch inhibitors from the seven experimentally validated lysine-analog decoys (more challenging due to the chemical similarity between actives and inactives). We found that GOLD:GOLD_Fitness combined with rDock_solv rescoring achieved the best enrichment (AUC=0.86) (**Fig. 2.7**) and ranked all seven active compounds within top eight.

We are surprised to find that the optimal combination of the methods for these two targets is different. We hypothesize that it was due to distinctive flexibility of the binding site. B-factors analyses of active site of 16S rRNA A-site were statistically higher than other part of the RNA (p=0.002) (**Fig 1.8A**), indicating that rRNA A-site is a flexible target. Furthermore,

normal mode analysis using oGNM [82] confirmed this local flexibility (**Fig. 2.8B**), because significant fluctuation of the A-site residues could be observed within five lowest-frequency modes (low-frequency motions are expected to have larger contribution to the conformational changes [83]). In contrast, based on the crystal structure of lysine riboswitch, the ligand (lysine) is completely enveloped in the rigid binding pocket of lysine riboswitch, and only the small molecules which can sterically fit the pocket can be accommodated. B-factor analysis demonstrated that lysine-binding pocket in this riboswitch was statistically more rigid than other residues (**Fig. 2.8A**). Normal mode analysis further confirmed the rigidity of this pocket (**Fig. 2.8C**).

Fig. 2.7. ROC curves of the virtual screening experiments. (A). Virtual screening against the 16S rRNA A-site using the Foloppe dataset. (B). Virtual screening against the lysine riboswitch using 7 known active compounds. (C-D). ROC comparison of the virtual screening performances of AutoDock4.1 and iMDLScore1/iMDLScore2 scoring functions with rRNA A-site (C). and lysine riboswitch (D). GOLD:GOLD Fitness dockings were in thin lines, while rDock:rDock_solv dockings were in thick lines. AutoDock4.1 default scoring function, iMDLScore1 and iMDLScore2 were colored red, blue and black, respectively.

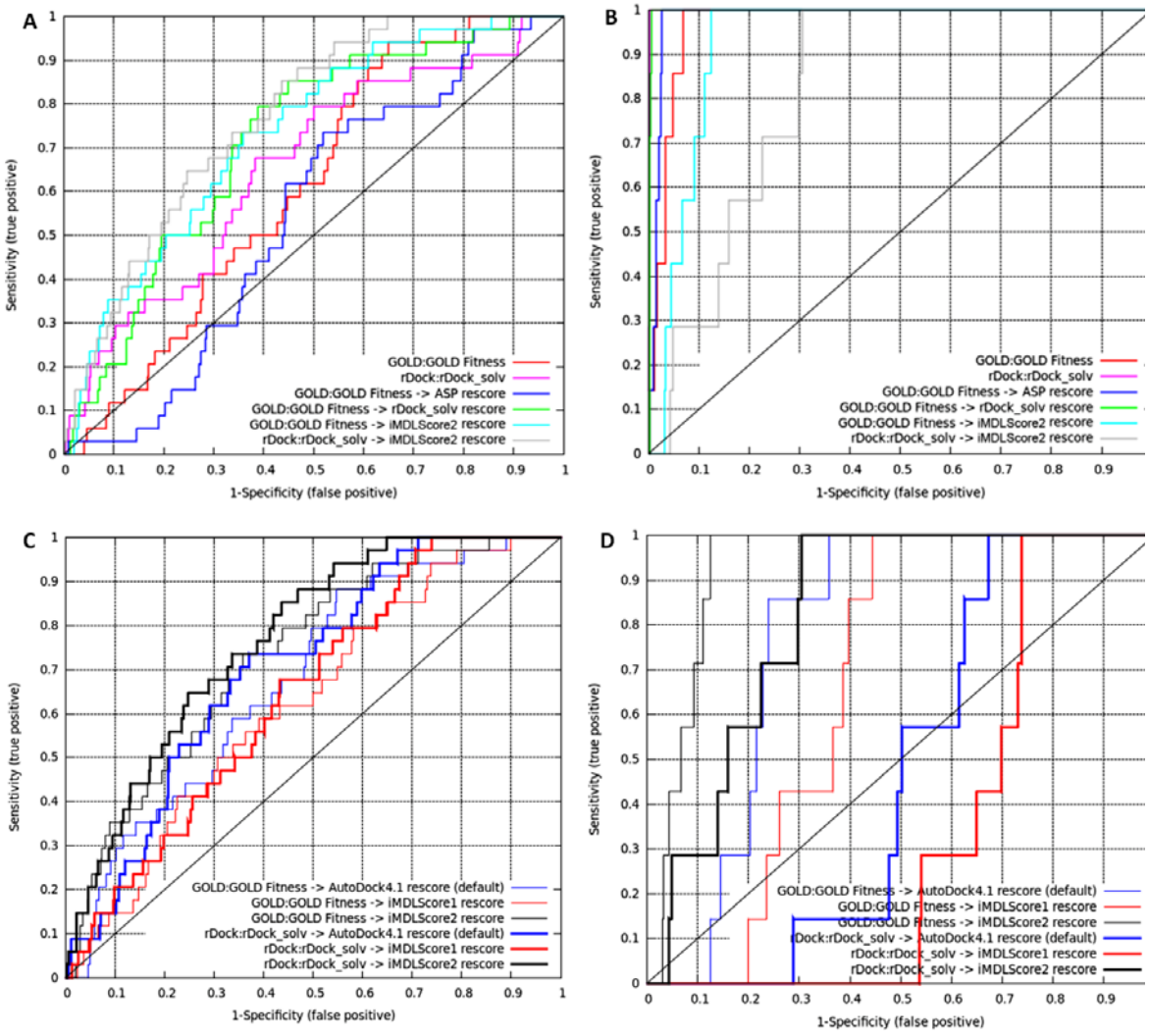


Table 2.6. ROC AUC for various docking and scoring combinations in virtual screening.

Lysine riboswitch (3DIL)		Bacterial rRNA A-site (1J7T)				
7 known inhibitors [2]	7 known inhibitors [1]	Foloppe dataset [1]	Zhou dataset [1]	Aminoglycosides [1]	Rescoring function	Initial docking & scoring function
0.82	0.97	0.58	1	1	None	GOLD:GOLD Fitness
0.86	0.999	0.61	1	1	None	rDock:rDock_solv
0.51	0.98	0.5	NA	NA	ASP	GOLD:GOLD Fitness
0.86	0.998	0.68	NA	NA	rDock_solv	GOLD:GOLD Fitness
NA	0.77	0.64	NA	NA	AutoDock4.1 Score	GOLD:GOLD Fitness
NA	0.66	0.58	NA	NA	iMDLScore1	GOLD:GOLD Fitness
0.51	0.92	0.69	NA	NA	iMDLScore2	GOLD:GOLD Fitness
NA	0.46	0.67	NA	NA	AutoDock4.1 Score	rDock:rDock_solv
NA	0.33	0.61	NA	NA	iMDLScore1	rDock:rDock_solv
0.51	0.81	0.74	NA	NA	iMDLScore2	rDock:rDock_solv

[1] The decoy set was MayBridge dataset

[2] The decoy set was seven known lysine analogs inactive to lysine riboswitch

Fig. 2.8. Difference between flexible and rigid RNA targets. (A) B-factor distribution. (B) Predicted flexibility of 16S rRNA A-site based on normal mode analysis. (C) Predicted flexibility profile of lysine riboswitch based on normal model analysis.

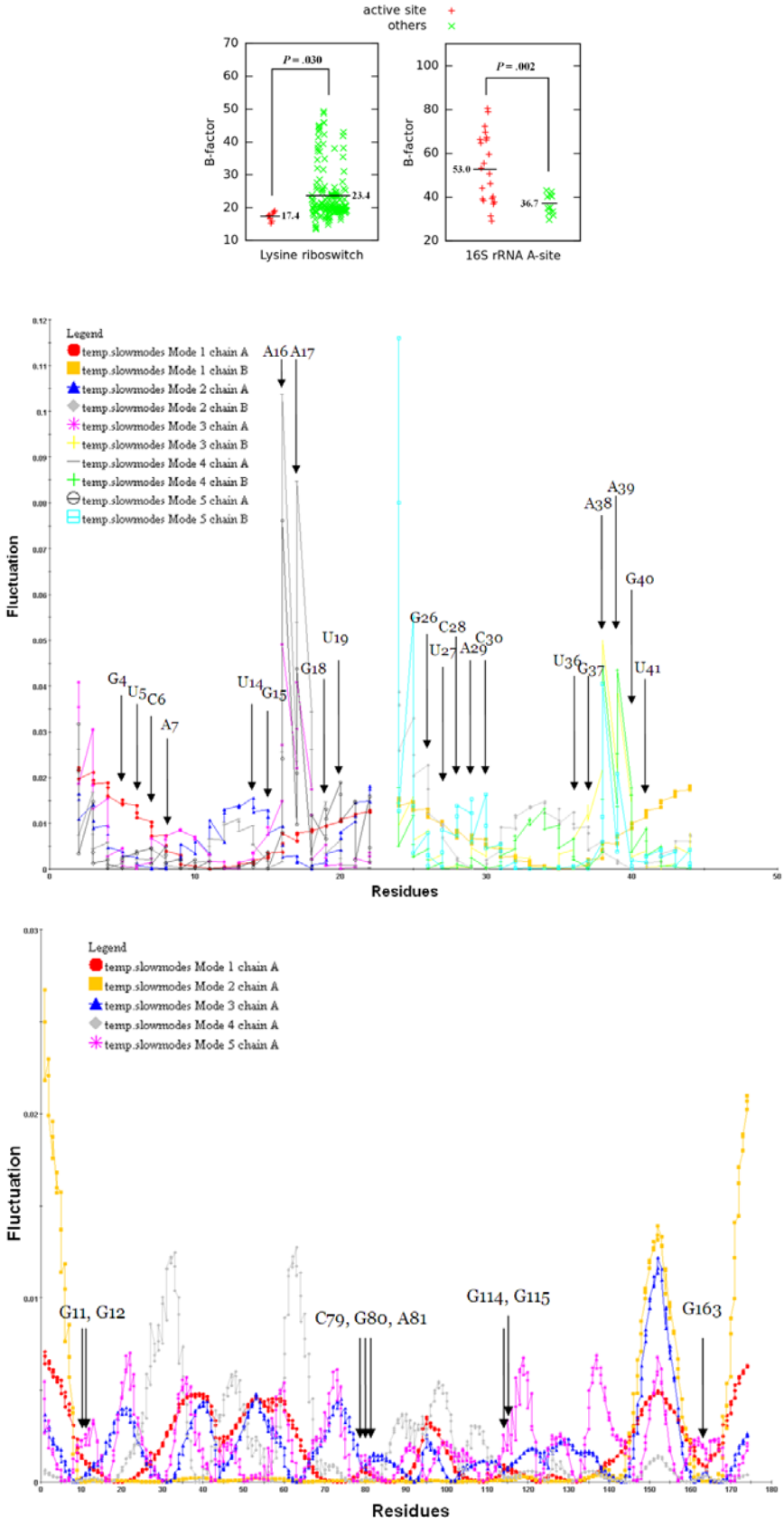
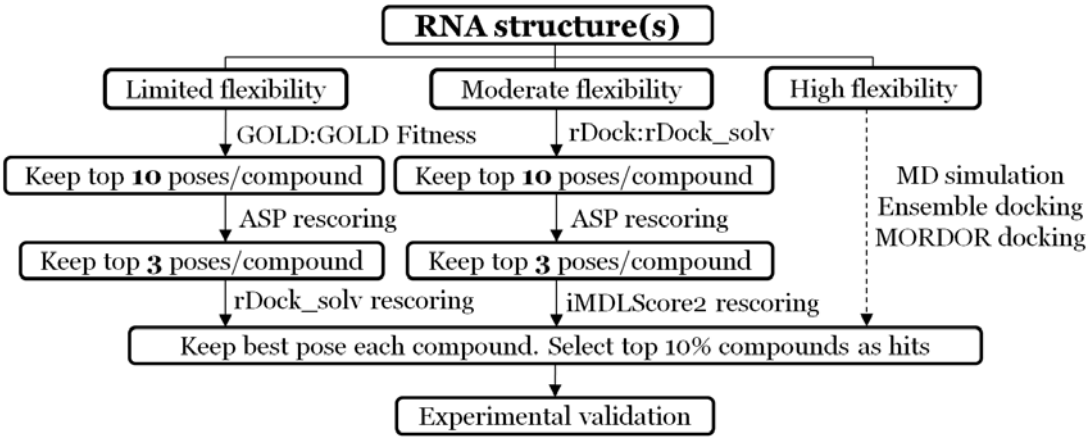


Fig. 2.9. The suggested workflow for structure-based virtual screening for RNA-targeted inhibitor discovery.



2.4 Results: Application of three-step docking scheme to identify novel RNA-small molecule interaction

2.4.1 Identify small molecules that binds GA:UU RNA internal loop

To identify the potential binders for GA:UU internal loop, we performed *in silico* high-throughput virtual screening using the protocol we derived in Chapter 1.3. Specifically, we screened ChemBridge diverse set (containing 100,000 compounds) and MayBridge (containing 14,400 compounds) using "GOLD:GOLD_Fitness-ASP-rDock_solv" pipeline for an NMR ensemble containing GA:UU motif provided by Dr. Nikonowicz group. Upon clustering analysis and visualization of the molecular interaction, we selected 15 compounds for experimental evaluations using 1D and 2D NMR. 1D NMR confirmed that two out of 15 compounds (compound **423** and **449**) are able to affect the chemical shifts for GA:UU motif. Compound **423** contains an amino-benzothiazole scaffold, whereas compound **449** contains a 1,4-dihydroquinoxaline-2,3-dione moiety. The chemical structure of 449 is shown in **Fig. 2.10**.

2.4.2 Experimental validation by NMR

The 1D imino proton resonances reflect the presence or absence of the RNA base pairing. Using 1D NH NMR, we identified 2-amino-1,3-benzothiazole-6-carboxamide (compound **423**) as the most potent and selective compound. The imino spectra of U22 and U7, which form noncanonical base pair in the unbound structure, exhibit weaker chemical shift and selective broadening by the addition of compound **423**. Meanwhile, some chemical shifts from a G-C base pair (~12.7ppm) and A-U base pair (~13.2ppm) were displaced (**Fig. 2.10A-**

B). However, the addition of compound **423** did not perturb the NH spectrum of a control RNA (RNA5: pre-miR-10b) that does not contain GA:UU motif (**Fig. 2.10D-E**). This result demonstrates that compound **423** does discriminate bulge / mismatch nucleotide identity. Although this does not indicate **423** is specific for the GA:UU motif, these results demonstrate this compound does discriminate bulge/mismatch nucleotide identity. In comparison, compound **449** shows a weaker effect on the 1D NH spectra (**Fig. 2.10C**).

According to 2D ^{13}C-^{1}H HSQC spectrum, the addition of compound **423** abolishes the chemical shifts from U7H6, G8H1', A9H8, A9H2, and the chemical shift from G8H8 becomes weak (**Fig. 2.11**). Chemical shift from U7H1' is also altered (**Fig. 2.11**). The peak perturbations in the NH and base spectra indicate that the binding of compound **423** should occur in the GA:UU tandem mismatch motif, but most of the effects are caused by the binding of G8 and A9. To further explore the binding context, 2D NMR was performed on 3 variants. The first one has GU wobble base pair at UU side and GC base pair at GA side (RNA2: $\begin{matrix} 5'GUGG3' \\ 3'UUAC5' \end{matrix}$), whereas the 2nd and 3rd variants extends the UU side with a AA:AU motif (RNA3: $\begin{matrix} 5'AGUGAU3' \\ 3'ACUAUA5' \end{matrix}$, RNA4: $\begin{matrix} 5'AGUGAG3' \\ 3'ACUAUC5' \end{matrix}$). Compared with the original RNA molecule ($\begin{matrix} 5'UGUGAU3' \\ 3'ACUAUA5' \end{matrix}$), RNA2 is much less stable at UU side (flanked by GU base pair) and more stable at GA side (flanked by GC base pair). RNA3 is less stable at UU side (flanked by GC+AA base pair), whereas RNA4 is more stable at GA side (flanked by AU+GC base pair). Consequently, we observed changes on RNA3, but not RNA4 and RNA5 using ^{13}C-^{1}H HSQC spectra (**Fig. 2.11B-C**). In fact, the effects on NMR spectra demonstrated

inverse correlation with the rigidity at GA side (rigidity at GA side: RNA2 > RNA4 > RNA1 & RNA3). This suggested that the flexibility adjacent to the GA base pair be another attribute that determines the selectivity of **423**.

Fig. 2.10. 1D NH spectra. of the GA:UU mismatch (A-C) and pre-miR-10b RNA hairpins (D,E). (A,D) RNA (0.015 mM) in 5% DMSO, (B,E) with 2-amino-1,3-benzothiazole-6-carboxamide (1), and (C) 5,7-dimethyl-1,4-dihydro-2,3-quinoxalinedione (2). Peaks altered by the compounds are labeled (*). **423** interacts with GA:UU mismatches but not A-A or bulged G. The interaction of compound **449** is weaker. Compound concentrations are 0.1 mM. NH spectrum is unaffected by 5% DMSO.

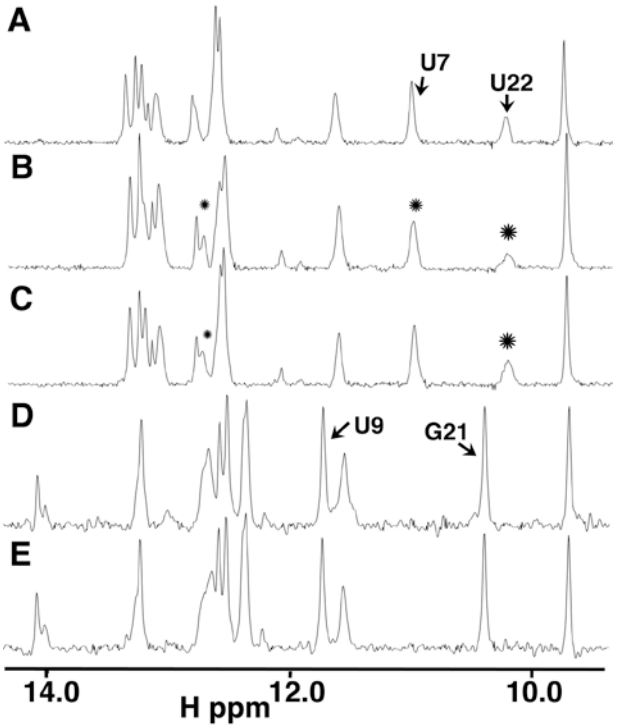
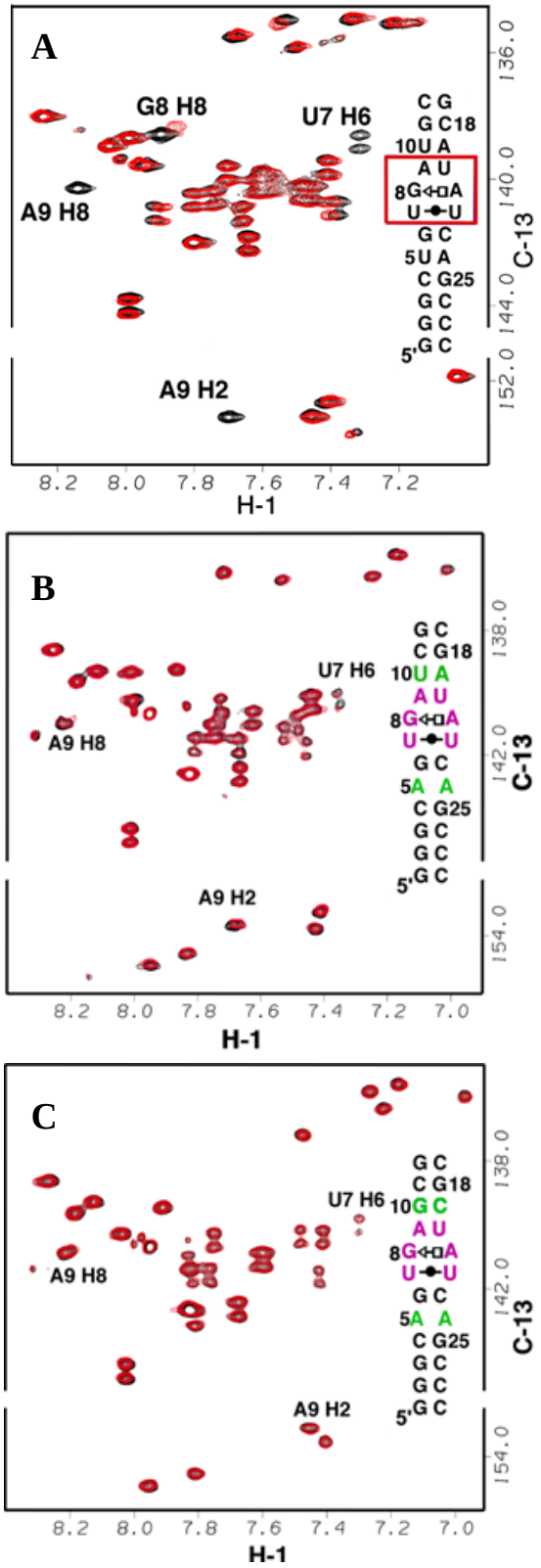


Fig. 2.11. **2D ^{1}H-^{13}C spectrum**. Base region of 2D ^{1}H-^{13}C spectrum of GA:UU containing RNA molecule (black). The spectrum after addition of compound **423** (red) leads to exchange broadening of labeled peaks. (A) RNA1(wildtype). (B) RNA3(mutU5A). (C) RNA4 (mutU5A-ΔAU).



2.4.3 Molecular dynamics study

To further investigate the selective binding mechanisms of **423** to GA:UU RNA internal loop, intensive MD simulations (~660 ns) were performed. **Fig. 2.12A** showed the smoothed trajectory of compound **423** over a total of 660 ns simulation in the explicit solvent. Intriguingly, we observed a complete binding circle of **423**'s associating or dissociating with GA:UU motif (**Fig. 2.12A-B**), indicating that the MD simulation we performed has sampled adequate configurational spaces and could be employed as a structural ensemble for further analysis. Based on MD simulation, we observed six periods in which compound 423 formed stable binding to RNA (**Fig. 2.12C**). Starting from the docking conformation, the **423** quickly associated with the minor groove of GA:UU motif (center of mass (COM) distance between GA:UU and **423** = 10Å) and formed specific and stable binding for ~100 ns (molecular details will be discussed later). Then **423** disassociated from GA:UU motif to the bulk solvent, and form nonspecific stacking with RNA terminal nucleotides periodically (at 150ns, 230ns, 300ns respectively). Compound **423** also interacted with the major groove formed by GA:UU motif in a nonspecific and transient manner (at ~450ns). Finally, it traveled back to the minor groove of GA:UU motif, in which the binding mode is almost identical to that in Stage I (COM distance = 8.2Å). After binding specifically to GA:UU for ~80ns, **423** once again disassociated from the RNA molecule and began a second binding circle (**Fig. 2.12C**).

Further examination of the average 3D model from Stage I and VI revealed that compound **423** primarily bind to the minor groove formed by G8, A9, U20, A21 (**Fig. 2.13A**) The benzothiazole moiety formed aromatic stacking on A21, and the amine group interacted with

U20 through intermolecular H-bond (**Fig. 2.13A** and **Fig. 2.13C**). Moreover, the sheared GA base pair exhibited an out-of-plane conformation (propeller twist) by 40° after the addition of **423** (**Fig. 2.13A**). Compared with the unbound structure (**Fig. 2.13B** and **Fig. 2.13D**), we observed that (1) A9H2 and G8H1' directly interact with **423** which explains the missing chemical shifts seen in the unbound RNA; (2) Due to the conformational changes caused by GA base pair propeller-twist, G8H8 and A9H8 changed their chemical environment; (3) The chemical shift changes of U7H6 and U7H1' is likely due to the change of sugar puckering of G6 (arrow highlighted in **Fig. 2.13C-D**). These findings are all consistent to 2D ^{13}C-^{1}H HSQC spectra. Moreover, **Fig. 2.14** demonstrated that destabilizing the two base pairs next to GA side is an essential step before compound **423** binds to GA:UU motif. In comparison, the base pair stability at +3 position (**Fig. 2.14 (bottom)**) does not correlate with the compound binding event. Therefore, MD simulation confirms the RNA specificity we observed in Chapter 1.4.2 that the flexibility at GA side may enhance the binding affinity.

2.4.4 Structure-activity relationship (SAR) analysis

Based on this 3D structural model, we further validated our hypothesis using rational designed structure-activity relationship (SAR) study. SAR demonstrated that any substitution on the R_1, R_2 or R_3 group failed to show any changes in the NMR spectra at concentration as high as 0.2mM (**Table 2.7**). This SAR result was consistent to the 3D model predicted by MD simulation, in which the amine group forms H-bond with U20, and the R_3-carboxamide forms polar contacts with G8 ribose ring such that any hydrophobic substitution is likely to abolish the binding. NMR proved that moving R_3-carboxamide to R_2 position is also detrimental, as

indicated by the 3D model that R_2 is exposed to the solvent and has minimal contribution to the binding (**Fig. 2.13A**).

Fig. 2.12. MD simulations of compound 423 binding to GA:UU motif. (A) Smoothed trajectory of GA:UU RNA and compound **423** over 660ns simulation. The color ranges from red to blue, denoting the time-dependent evolution of the complex structure from 0ns to 660ns. (B) Representative structures of through MD simulation. (C) Center of mass distance between GA:UU and compound **423**. Each stable state is assigned an ID, whose 3D structure has been illustrated in (B).

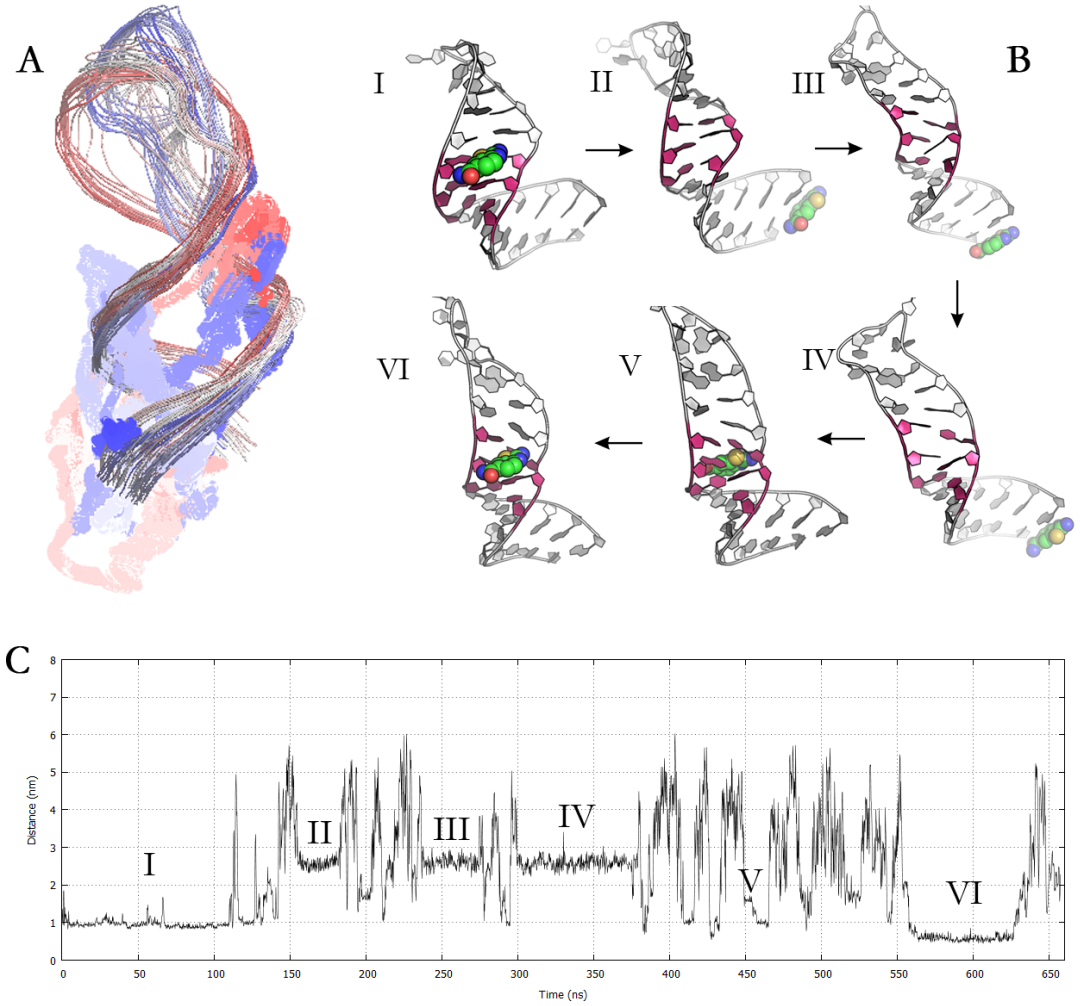


Fig. 2.13. 3D model of compound 423 binding to UU:GA motif. Compound **423** are shown in red sticks, and the atoms altered by addition of **423** are shown in sphere. The changes of sugar puckering are highlighted with arrows. (A) Minor groove view of 423-bound structure. (B) Minor groove view of unbound structure. (C) Major groove view of 423-bound structure. (D) Major groove view of unbound structure.

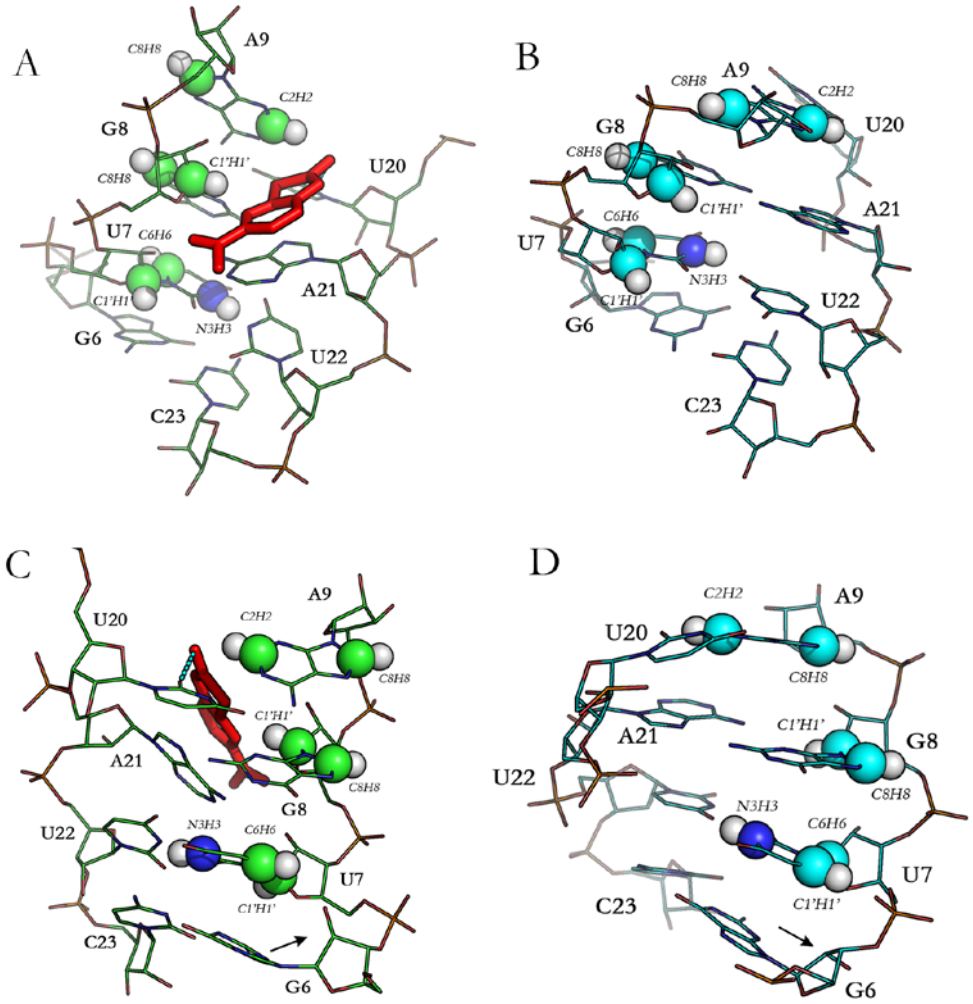


Fig. 2.14. Base pair flexibility of the context of GA:UU motif. Time-dependent distance between A9N1-U20N3, U10N3-A19N1, and G9N1-C18N3, which denote the three base pair from the GA base pair. The arrows highlighted the time point that **423** start to associate with GA:UU motif.

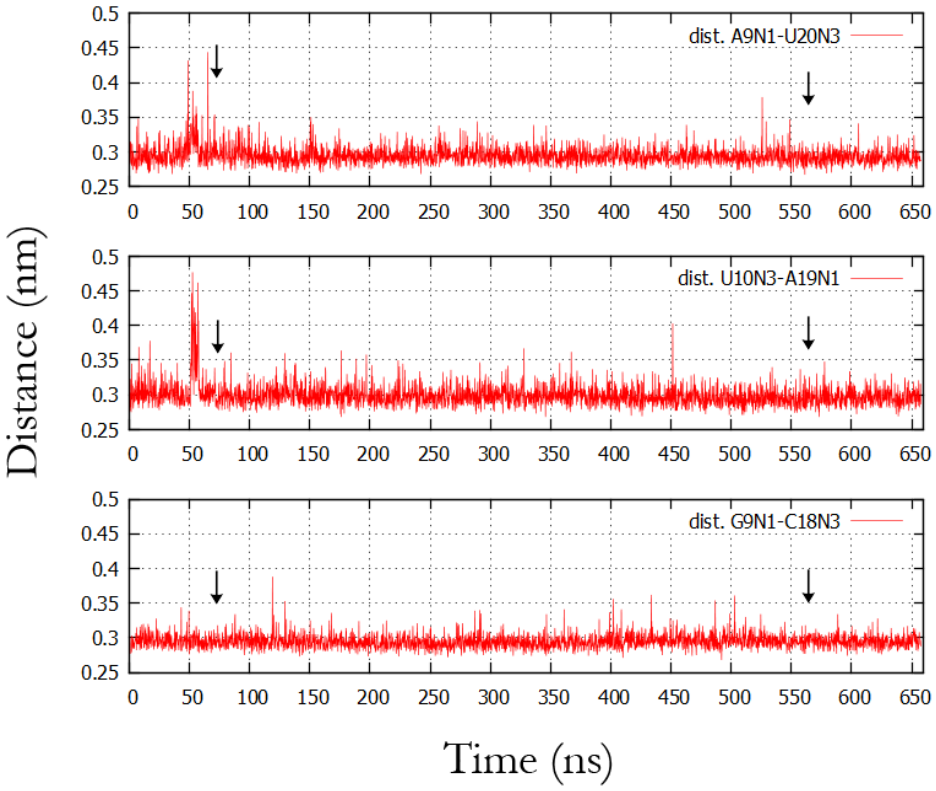


Table 2.7. Structure-activity relationship of 423 series compounds.

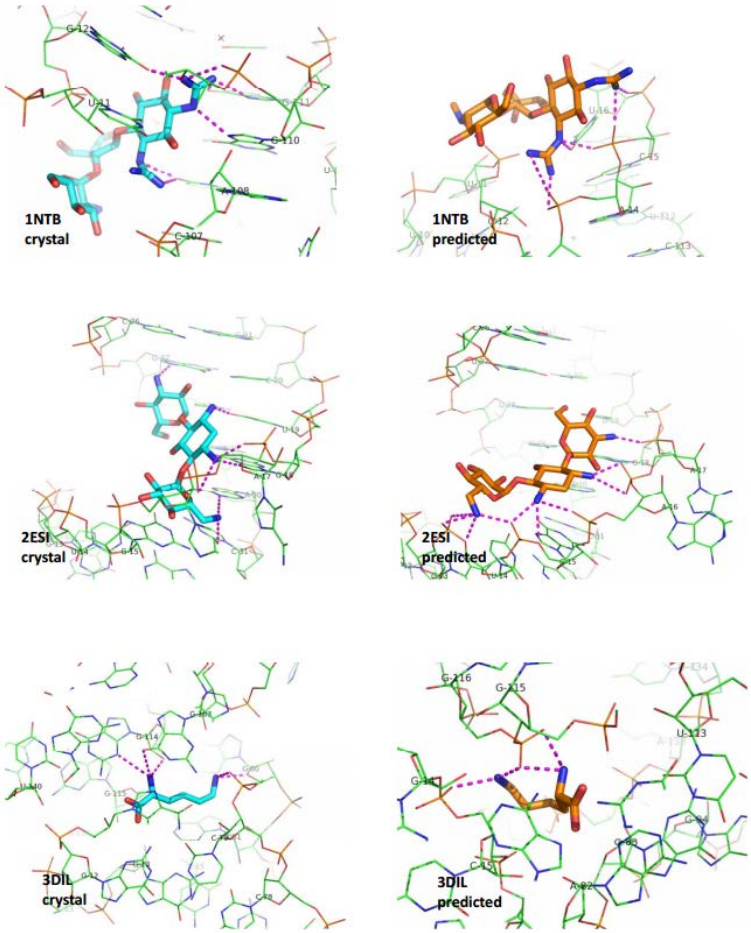
R_1	R_2	R_3	Activity
-H	-H	$-CONH_2$	active
-H	-H	$-NHCOCH_3$	inactive
-H	-H	$-CH(CH_3)CH_2$-**423**	inactive
-H	$-CONHNH_2$	-H	inactive
$-NH_2$	-H	$-CONHNH_2$	inactive

2.5 Discussion

RNA represents a historically important, but less systematically investigated therapeutic target. The discovery of various aminoglycosides inhibitors target prokaryotic 16S rRNA is a proof-of-principle that RNA can be specifically targeted by small molecules. On the other hand, rational design of small-molecule inhibitor targeting specific RNA motif less appreciated due to lack of reliable *in silico* tools for structure-based drug design. To address this issue, we have benchmarked and identified an optimal docking / scoring pipeline for RNA-ligand modeling and virtual screening through a comprehensive evaluation in three different aspects. First, we have identified GOLD:GOLD Fitness and rDock:rDock_solv as the best pose predictors and are most appropriate for the initial binding modes generation. Nevertheless, we proved that rescoring of the predicted binding modes is a necessary step to enhance the enrichment of true positives in virtual screening exercises. To this end, we discovered that ASP, rather than GOLD Fitness or rDock_solv scoring function, achieved the best performance in pose ranking evaluation. Second, scoring functions can be generally classified as soft-core potentials (e.g. AutoDock scores and iMDLScores) and hard-core potentials (rDock_solv and ASP). Based on the structural resolution and flexibility of the binding sites, hard- or soft-core potentials may behave distinctively, as we summarized in **Fig. 2.9**. Hard-core scoring functions, for example, usually result in better ROC AUC than the soft-core ones when the target is an RNA ensemble. Third, implementation of RNA-specific scoring function (e.g. iMDLScore2) improved the virtual screening enrichment as well as the accuracy of RNA-ligand binding affinity prediction.

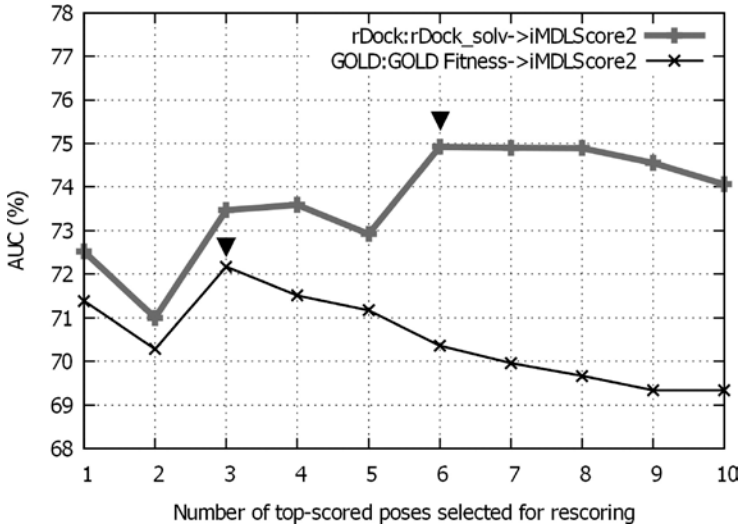
In agreement with previous docking benchmarks [84], we report that the docking/scoring combination that is specialized in binding mode reproduction does not correlate its corresponding performance in virtual screening. Hence, we exhaustively explored the best strategy independently for pose generation, pose ranking and hit ranking, so called three-step RNA virtual screening. Consistent to our hypothesis, some scoring function such as iMDLScore2 which is good at hit identification (the final step) performs poorly in initial pose generation (the first step). We found that if iMDLScore2 was biased to the electrostatic interactions with RNA backbone if it was selected for pose generation (**Fig. 2.15**). Indeed, we concluded that so far no existing docking/scoring combination can have satisfactory performances on all three steps. Our three-step pipeline circumvented the pitfalls of traditional one-step docking-scoring by separating the conformation-wise pose generation/selection and ligand-wise hit selection, and it outperformed other methods in our virtual screening benchmarks (**Fig. 2.7**).

Fig. 2.15. **Comparisons of AutoDock4.1:iMDLScore2 predicted binding modes with experimental structures**. 1NTB, 2ESI and 3DIL were used as the examples to demonstrate the overestimation of polar interactions with RNA phosphate for initial pose generation purpose. RNA receptors were shown in green lines, while experimentally determined binding modes are shown in cyan sticks. AutoDock4.1 generated pose with the best RMSD were in orange sticks. The interactions between basic guanidinium/amine groups with RNA atoms were labeled with magenta dashes. We could observe that these basic groups were predicted dominantly to form interaction with the backbone phosphates; actually, the H-bonds with RNA base atoms and cation-π interactions were more favorable.



Compared with one-step docking process, one of the limitations of our three-step pipeline is that current implementation cannot derive a statistical model of determining how many candidates we should retain for the next step. Here we conceptualized a successful virtual screening exercise as a process not only to cover the near-native ligand conformation, but also to enrich the true positive from these predicted conformations. Hence, the hit rate is a net effect of pose selection and compound selection, and we observed that forwarding too many poses for each ligand to hit selection stage may harm the virtual screening performance. **Fig. 2.16** showed that keeping top three poses in GOLD:GOLD Fitness phase achieved the best ROC AUC in iMDLScore2 rescoring phase. ROC AUC is declining when no. retained poses is increasing. Similar trend is also found when rDock:rDock_solv is used as the first phase, but the maximum performance occurs only when we retains top 6 poses. Our data suggest that even though arbitrarily keeping top 10 poses can only compromise <3% ROC AUC, the number of poses that can achieve the best performance is still hard to estimate *a priori*. Thus, we will continue to develop statistical model to predict the *a priori* no. of poses based on the features of target and screening library.

Fig. 2.16. **ROC AUC against number of candidate poses selected for iMDLScore2 rescoring for 16S rRNA A-site.** The downward-pointing triangle (▼) represents the number of picked poses corresponding to the best ROC AUC (turning point).



Chapter 3: Computational modeling of novel RNA-protein interaction

3.1 Introduction

All aforementioned pitfalls described in Chapter 1.2 motivated the implementation of interface threading for RPI prediction. The idea of interface threading is inspired by *iWRAP* [85] which predicts protein-protein interaction (PPI) by referencing to template PPI(s). Our implementation, RPI prediction through Interface Threading (or *RPIT*), identifies and references an RNA-protein interface as a template to estimate the interface region where the interaction occurs. To estimate the interaction propensity between different types of amino acids and nucleotides more precisely, we have implemented a set of statistical scoring functions based on our unique collection of non-redundant protein-RNA interaction database. Compared with sequence-based methods, interface threading model not only predicts accurately the probability of the RNA-protein direct binding, but also infers the sequence elements that are most attributable to binding. This is significantly appealing when the size of RNA (or protein) is huge so that trivial mutagenesis study is prohibitive when validating the computational model. Compared with structure-based method, *RPIT* offers an inexpensive but robust method for *in silico* prediction of RNA-protein interaction networks, and for prioritizing putative RNA-protein pairs for experimental validation.

Here we hypothesized that the interaction propensity between protein and RNA is dominated by interface regions, and mutations on a distal region contribute less than those close to the interface. In order to validate this hypothesis, there are four specific aims (**Fig. 3.1**):

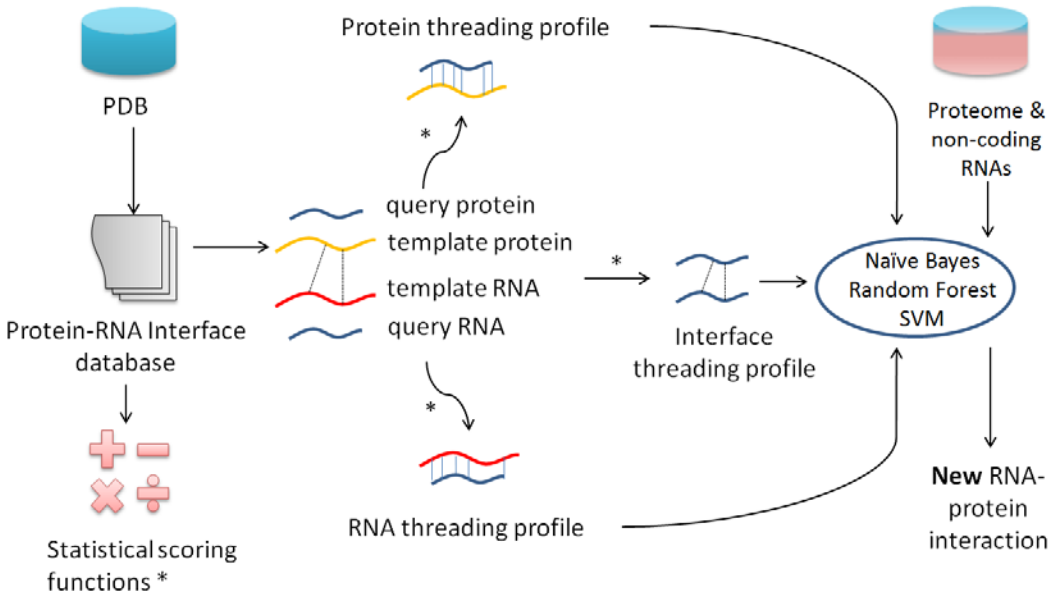
1. Develop an interface template database to which query protein-pair can thread.

2. Develop statistical scoring functions to evaluate the threading performance.

3. Design novel alignment (threading) paradigms to incorporate 2nd structural and interface importance information.

4. Implement classifier to predict the probability of interaction based on alignment and scoring functions.

Fig. 3.1. **An overview of protein-RNA interface threading pipeline.** There are four milestones of protein-RNA interface threading. 1. Develop an interface template database. 2. Develop statistical scoring functions. 3. Design alignment (threading) algorithms. 4. Implement functions to predict the probability of interaction based on alignment and scoring.



3.2 Materials and Methods: Development, Validation and Application

3.3.1 Non-redundant protein-RNA interfaces database (nrPR)

As of April 26th 2014, there are totally 1,585 protein-RNA complex structures deposited in PDB. From this collection, we have curated 20,111 protein-RNA interaction pairs (termed as *totPR* dataset), each of which contains at least five residue-based intermolecular interactions. We further removed the redundants from *totPR* (using 0.8 similarity cutoff for protein and 0.6 for RNA, considering sequence, secondary structure and types of interaction simultaneously), ultimately resulting in 5,471 non-redundant interaction pair (termed as *nrPR* database). This *nrPR* database will be used as the threading templates as well as the training set to derive statistical scoring functions. Notably, we keep the non-standard amino acids and nucleotide intact in the 3D structure, but may ignore them when deriving statistical scoring functions. For NMR structures, we select the best representative model according to 'selection_criteria' tag in the mmCIF file. The diversity of the protein or RNA sequences in *nrPR* is analyzed by principle component analysis (PCA) using conjoint triad descriptors [59]. There are 343 features for protein sequence and 256 features for RNA sequence. In this chapter, we may use "query" and "target" interchangeably.

3.3.2 Statistical Scoring Functions

We designed five knowledge-based statistical scoring functions to determine the fitness of interface threading. Generally, ***PInter*** (or ***RInter***) estimates the interaction propensity that evaluates how favorable an interfacial protein (or RNA) residue to form a specific interaction. ***PDist*** (or ***RDist***) estimates the distance propensity that evaluates how favorable a protein (or

RNA) residue is found on/close to the interface. Here we defined 12 types of RPI which empirically summarize the fundamental amino acid-nucleotide contacts at atomic level (**Table 3.1**). **Fig. 3.2** illustrates the schematic view of these 12 types of interaction.

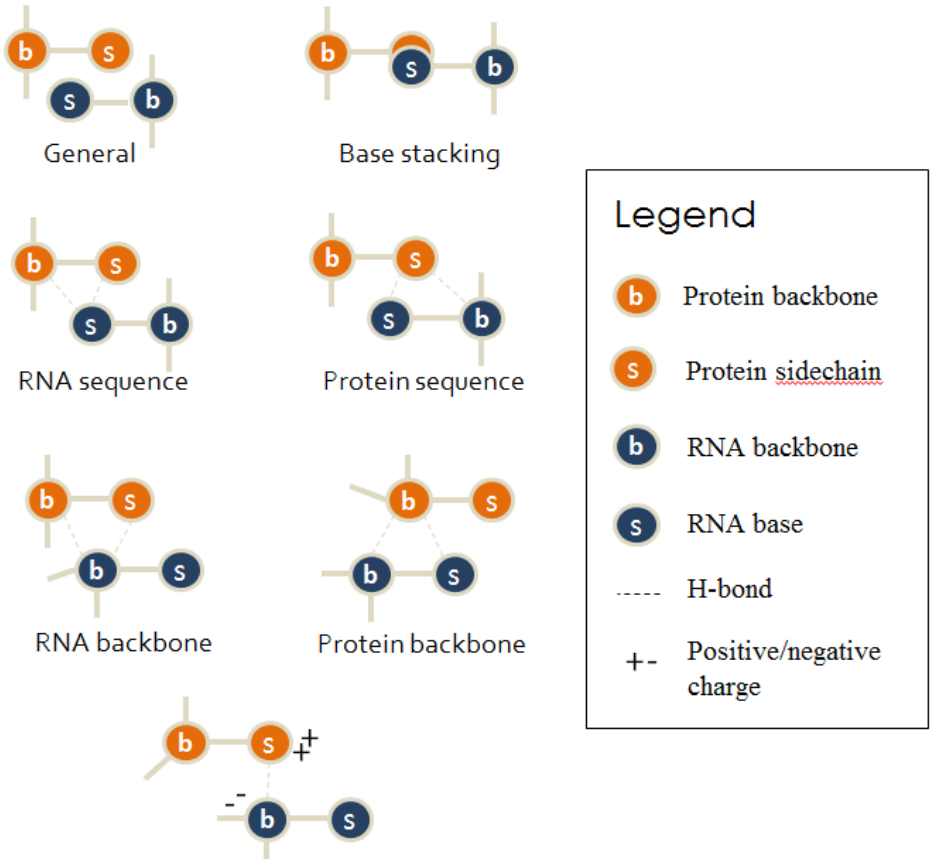
The secondary structure of protein residues were analyzed by Stride [86]. As Stride predicts 7 types of protein secondary structures (H=α-helix, G=3_{10}-helix, I=π-helix, E=extended conformation, B(or b)=isolated bridge, T=turn), we clustered these 7 secondary structure codes into helix, sheet and coil as following: $helix \in \{\mathrm{H}, G, I\}$, $sheet \in \{E, B(b)\}$, $coil \in \{T, C\}$. RNA secondary structures were analyzed using DSSR v1.0.2, a new component of 3DNA suite of software programs [87]. We define a paired state to be Watson-Crick base pairing (19-XIX or 20-XX) or G-U wobble base pairing (28-XXVIII), and an unpaired state to be any other noncanonical base pair or unpaired bulge. vdW interaction denotes any atom pair with distance shorter than 4.0Å. Salt bridge (or electrostatic attraction) denotes the interaction involving a phosphate atom of nucleotide and an atom of basic amino acid whose distance is shorter than 4.5Å. All H-bonds, aromatic stacking, cation-π interactions, and electrostatic attractions were computed by Molecular Operating Environment (MOE). Then we confirmed the MOE assignments and assign aromatic-like stacking and hydrophobic stacking using the criteria described in **Table 3.1,** based on the number of atomic contacts.

Table 3.1 Summary of 12 types of RPI

Applicable amino acids [b]	Definition	Comment	Type	Category
all	*dist* < 4.0Å [a]	van der Waals interaction	vdw	General
H, F, Y, W	*N*(vdw between protein sidechain and RNA base) ≥ 12 [a]	aromatic stacking on RNA base	arom	Base stacking
D, N, E, Q, R	*N*(vdw between protein sidechain and RNA base) ≥ 12	aromatic-like stacking on RNA base	arom_l	
A, V, P, I, L, M	*N*(vdw between protein sidechain and RNA base) ≥ 3	hydrophobic stacking with RNA base	hy	
K, R	*N*(vdw between positively charged protein sidechain and RNA base) ≥ 3 [a]	cation-π interaction with RNA base	cpi	
all	*N*(H-bond with RNA base) ≥ 2 [a]	at least two h-bond interactions with RNA base	hbr_2	RNA sequence recognition
all	*N*(H-bond with RNA base) ≥ 1 [a]	one h-bond interaction with RNA base	hbr_1	
D, E, H, K, N, Q, R, S, T, W, Y	*N*(H-bond with protein sidechain) ≥ 2 [a]	at least two h-bond interactions with protein side chain	hbp_2	Protein sequence recognition
D, E, H, K, N, Q, R, S, T, W, Y	*N*(H-bond with protein sidechain) ≥ 1 [a]	one h-bond interaction with protein side chain	hbp_1	
all	*N*(H-bond with RNA backbone) ≥ 1 [a]	at least one H-bond interaction with RNA backbone	bb_r	RNA backbone recognition
all	*N*(H-bond with protein backbone) ≥ 1 [a]	at least one H-bond interaction with protein backbone	bb_p	Protein backbone recognition
K, R	*dist*(positively charged protein sidechain, RNA phosphate) < 4.5 Å [a]	basic amino acid interacting with RNA phosphate (<4.5Å)	salt	Long-range non-specific

[a] Interaction that are confirmed by MOE
[b] 12 interaction types are applicable to all nucleotides

Fig. 3.2. Schematic view of 7 major categories of RPI types.



3.3.2.1 *PInter* and *PDist:* RNA-binding ability for amino acids

PInter scoring function quantifies the probability of a given amino acid (*i*) holding the secondary structure (ss_k) to form a specific interaction (t_m) to any nucleotide. $i \in 20$ standard amino acids, $k \in \{helix, sheet, coil\}$, $m \in 12$ interaction types. The statistical potential of ***PInter*** is written as following:

$$P_{Inter}\left(p_i^{ss_k}, t_m\right) = \ln\left(\frac{N_{obs}(p_i^{ss_k}, t_m)}{N_{\exp}(p_i^{ss_k}, t_m)}\right) = \ln\left(\frac{N_{obs}(p_i^{ss_k}, t_m) + B_c/60}{X_{i,ss_k} \cdot X_{t_m} \cdot \left[N_{obs}(p_\cdot^{ss\cdot}, t_\cdot) + B_c\right]}\right)$$

$$X_{i,ss_k} = \frac{N_{obs}(p_i^{ss_k})}{N_{obs}(p_\cdot^{ss\cdot} \in T_m)}$$

$$X_{t_m} = \frac{N_{obs}(p_\cdot^{ss\cdot}, t_m)}{N_{obs}(p_\cdot^{ss\cdot}, t_\cdot)}$$

$$B_c = \sqrt{N_{obs}(p_\cdot^{ss\cdot}, t_m)}$$

$N_{obs}(p_i^{ss_k}, t_m)$ is the observed number of interaction of a protein residue, $p_i^{ss_k}$, with any nucleotide using interaction type, t_m, whereas $N_{\exp}(p_i^{ss_k}, t_m)$ is the expected number of interaction of a protein residue, $p_i^{ss_k}$, with any nucleotide using interaction type, t_m. The expected value is calculated in a similar manner as in χ^2-test. Here, we estimates the unbiased fraction of residue $p_i^{ss_k}$ (X_{i,ss_k}) and type t_m (X_{t_m}) from all observed interactions ($N_{obs}(p_\cdot^{ss\cdot}, t_\cdot)$), in which dot '.' denotes any residue, secondary structure or interaction type. To take zero observation $N_{obs}(p_i^{ss_k}, t_m)$ into consideration, a pseudocount B_c is added to both denominator and numerator. B_c is calculated as the square root of all observations $N_{obs}(p_\cdot^{ss\cdot}, t_\cdot)$. Of note, when computing the unbiased fraction of protein residue X_{i,ss_k}, interaction propensity of some interaction type (e.g., aromatic or aromatic-like stacking,

cation-π) can be inflated, due to the physicochemical nature of different amino acid. For example, Ala, Val, Leu and Ile by nature are incapable of forming H-bond with its sidechain, and thus hbp_2 and hbp_1 are not applicable to these residues. Therefore, we apply an "interaction type correction" when computing the denominator for X_{i,ss_k}. $N_{obs}(p_{\cdot}^{ss\cdot} \in T_m)$ denotes the number of observed amino acid in which the interaction category that t_m is belonged to is applicable. For instance, when calculating interaction propensity of (Phe, α-helix, aromatic stacking (arom)), $N_{obs}(p_{\cdot}^{ss\cdot} \in T_m)$ will only counts the residues that are able to form aromatic stacking, aromatic-like stacking, hydrophobic stacking or cation-π interaction. This can significantly reduce the bias toward amino acid-specific interaction type.

PDist scoring function quantifies the probability of a given amino acid (*i*) holding the secondary structure (ss_k) to reside at most d_θ Å from any interfacial amino acid. $i \in 20$ standard amino acids, $k \in \{helix, sheet, coil\}, \theta \in \{0 \ldots 20\}$. When the given amino acid forms direct interaction with RNA (interfacial residue), the distance is zero. The statistical potential of ***PDist*** is written as following:

$$P_{Dist}\left(p_i^{ss_k}, d_\theta\right) = \ln\left(\frac{N_{obs}(p_i^{ss_k}, d \le d_\theta)}{N_{\exp}(p_i^{ss_k}, d \le d_\theta)}\right) = \ln\left(\frac{N_{obs}(p_i^{ss_k}, d \le d_\theta)}{X_{i,ss_k} \cdot X_{d_\theta} \cdot N_{obs}(p_{\cdot}^{ss\cdot}, d_{\cdot})}\right)$$

$$X_{i,ss_k} = \frac{N_{obs}(p_i^{ss_k}, d_{\cdot})}{N_{obs}(p_{\cdot}^{ss\cdot}, d_{\cdot})}$$

$$X_{d_\theta} = \frac{N_{obs}(p_{\cdot}^{ss\cdot}, d \le d_\theta)}{N_{obs}(p_{\cdot}^{ss\cdot}, d_{\cdot})}$$

Basically, $N_{obs}(p_i^{ss_k}, d \le d_\theta)$ is the number of observed cases for a specific protein residue $p_i^{ss_k}$ to be found at a distance equal or less than d_θ. $N_{\exp}(p_i^{ss_k}, d \le d_\theta)$ is the number of expected cases for a specific protein residue $p_i^{ss_k}$ to be found at a distance equal or less than d_θ.To calculate the expected value, we estimate the unbiased fraction as we did in ***PInter***. $N_{\exp}(p_i^{ss_k}, d \le d_\theta)$ is computed as the multiplication of unbiased fraction of $p_i^{ss_k}$ (X_{i,ss_k}), unbiased fraction of all residues with maximal distance to interfacial atom $d \le d_\theta$ (X_{d_θ}) and total number of observations for any residue at any distance ($N_{obs}(p_\cdot^{ss\cdot}, d_\cdot)$). Different from interaction propensity, no pseudocount was applied here, as theoretically the occurrence of protein residue at some distance is assumed to be random enough to prohibit zero observation.

3.3.2.2 *RInter* and *RDist*: Protein-binding ability for nucleotides

RInter scoring function quantifies the probability of a given nucleotide (*j*) holding the secondary structure (ss_l) to form a specific interaction (t_m) to any amino acid. $i \in 4$ standard nucleotide, $k \in \{WC/GU, others\}$, $m \in 12$ interaction types. The statistical potential of ***RInter*** is derived as following:

$$R_{Inter}\left(r_j^{ss_l}, t_m\right) = \ln\left(\frac{N_{obs}(r_j^{ss_l}, t_m)}{N_{\exp}(r_j^{ss_l}, t_m)}\right) = \ln\left(\frac{N_{obs}(r_j^{ss_l}, t_m) + B_c/8}{X_{j,ss_l} \cdot X_{t_m} \cdot \left[N_{obs}(r_\cdot^{ss\cdot}, t_\cdot) + B_c\right]}\right)$$

$$X_{j,ss_l} = \frac{N_{obs}(r_j^{ss_l})}{N_{obs}(r_\cdot^{ss\cdot})}$$

$$X_{t_m} = \frac{N_{obs}(r_\cdot^{ss\cdot}, t_m)}{N_{obs}(r_\cdot^{ss\cdot}, t_\cdot)}$$

$$B_c = \sqrt{N_{obs}(r_\cdot^{ss\cdot}, t_m)}$$

Here all parameters used in ***RInter*** scoring function are similar to those in ***PInter***, except that 1) the pseudocount are divided by 8, because there are only 8 combinations of nucleotide type and 2 secondary structure type for RNA; 2) there is no interaction type correction when computing unbiased fraction of nucleotide, X_{j,ss_l}, because all 12 interaction types are applicable to all nucleotides (A, U, G, C).

Similar to protein, we define ***RDist*** as following, in which X_{j,ss_l} and X_{d_θ} are the unbiased fractions for RNA nucleotide ($r_j^{ss_l}$) and the maximal distance from any interface nucleotide ($d \le d_\theta$), respectively:

$$R_{Dist}\left(r_j^{ss_l}, d_\theta\right) = \ln\left(\frac{N_{obs}(r_j^{ss_l}, d \le d_\theta)}{N_{\exp}(r_j^{ss_l}, d \le d_\theta)}\right) = \ln\left(\frac{N_{obs}(r_j^{ss_l}, d \le d_\theta)}{X_{j,ss_l} \cdot X_{d_\theta} \cdot N_{obs}(r_\cdot^{ss_\cdot}, d_\cdot)}\right)$$

$$X_{j,ss_l} = \frac{N_{obs}(r_j^{ss_l}, d_\cdot)}{N_{obs}(r_\cdot^{ss_\cdot}, d_\cdot)}$$

$$X_{d_\theta} = \frac{N_{obs}(r_\cdot^{ss_\cdot}, d \le d_\theta)}{N_{obs}(r_\cdot^{ss_\cdot}, d_\cdot)}$$

3.3.2.3 Protein-RNA interface fitness: *PRInter*

PRInter scoring function quantifies the probability of a given amino acid (*i*) holding the secondary structure (ss_k) to form a specific interaction (t_m) to a given nucleotide (*j*) holding the secondary structure (ss_l). The definitions of *i*, *j*, ss_k, ss_l, t_m have been described above. The statistical potential of ***PRInter*** is given as following:

$$PR_{Inter}\left(p_i^{ss_k}, r_j^{ss_l}, t_m\right) = \ln\left(\frac{N_{obs}(p_i^{ss_k}, r_j^{ss_l}, t_m)}{N_{\exp}(p_i^{ss_k}, r_j^{ss_l}, t_m)}\right)$$

$$= \ln\left(\frac{N_{obs}(p_i^{ss_k}, r_j^{ss_l}, t_m) + B_c/480}{X_{i,ss_k} \cdot X_{j,ss_l} \cdot X_{t_m} \cdot \left[N_{obs}(p_\cdot^{ss\cdot}, r_\cdot^{ss\cdot}, t_\cdot) + B_c\right]}\right)$$

$$X_{i,ss_k} = \frac{N_{obs}(p_i^{ss_k})}{N_{obs}(p_\cdot^{ss\cdot} \in T_m)}$$

$$X_{j,ss_l} = \frac{N_{obs}(r_j^{ss_l})}{N_{obs}(r_\cdot^{ss\cdot})}$$

$$X_{t_m} = \frac{N_{obs}(p_\cdot^{ss\cdot}, r_\cdot^{ss\cdot}, t_m)}{N_{obs}(p_\cdot^{ss\cdot}, r_\cdot^{ss\cdot}, t_\cdot)}$$

$$B_c = \sqrt{N_{obs}(p_\cdot^{ss\cdot}, r_\cdot^{ss\cdot}, t_m)}$$

Similar to other interaction propensity scoring functions, it is computed from the logarithm of observed cases ($N_{obs}(p_i^{ss_k}, r_j^{ss_l}, t_m)$) over expected cases ($N_{\exp}(p_i^{ss_k}, r_j^{ss_l}, t_m)$). ***PRInter*** applied three unbiased fractions X_{i,ss_k}, X_{j,ss_l}, X_{t_m} and the pseudocount will be divided by 480 as there are in total 480 combinations (20 amino acids, 3 amino acid secondary structure states, 4 nucleotides, 2 nucleotides secondary structure states). As we discussed in ***PInter***, Interaction type correction term is applied to ***PRInter*** as we have discussed in ***PInter***.

3.3.3 Develop protein-RNA threading and scoring scheme

3.3.3.1 Protein threading and scoring

RaptorX, the best template-based modeling method in Critical Assessment of Protein Structure Prediction 9 (CASP9), is used for protein interface threading problem. *RaptorX* is equipped an integer linear programming (ILP) scheme so that when searching and aligning to template(s), *RaptorX* optimizes the objective function which involves sequence profile similarity, statistical potential-based sequence similarity, secondary structure profile

similarity, solvent accessibility, contact capacity, environment fitness, sequence identity, alignment length and gaps simultaneously [77,78]. Therefore, this tool is ideal for fold recognition using low-homology template(s). Here, non-redundant protein structures in *nrPR* database will be treated as the templates to thread the target protein sequence. Based on the alignment provided by *RaptorX*, we calculated the interface threading score as following:

$$E_{pThread} = \left(E^{q}_{pInter} + E^{q}_{pDist} + E^{q}_{pGap_{penalty}}\right) - \left(E^{t}_{pInter} + E^{t}_{pDist}\right)$$

$$E^{norm}_{pThread} = E_{pThread} \Big/ \sum_{t} \psi\left(d^{t}_{\theta}\right)$$

Where:

$$E^{q}_{pInter} = \sum_{t,q\in aligned} \sum_{m\in t} \begin{cases} \sum_{k\in\{H,E,C\}} \left[p_{ss_k} \cdot P_{Inter}\left(p_i^{ss_k}, t_m^{t}\right)\right] \\ \arg\max_{q'\in[q-3,q+3]} \sum_{k\in\{H,E,C\}} \left[p_{ss_k} \cdot P_{Inter}\left(p_i^{ss_k}, t_m^{t}\right)\right], \ if\ m \notin T_{m,i}\ and\ m = salt \\ \sum_{m'\in T_{m,i}} \sum_{k\in\{H,E,C\}} \left[p_{ss_k} \cdot P_{Inter}\left(p_i^{ss_k}, t_{m'}^{t}\right)\right], \ if\ m \notin T_{m,i}\ and\ m \in base\ stacking \end{cases}$$

$$E^{q}_{pDist} = \sum_{t,q\in aligned} \sum_{k\in\{H,E,C\}} P_{Dist}\left(p_i^{ss_k}, d^{t}_{\theta}\right)$$

$$E^{q}_{pGap_{penalty}} = -\sum_{t\neq -,q=-} \left[P_{Dist}\left(p_i^{ss_k}, d^{t}_{\theta}\right) + \alpha \cdot \psi\left(d^{t}_{\theta}\right)\right]$$

$$E^{t}_{pInter} = \sum_{t} \sum_{m\in t} P_{Inter}\left(p_i^{ss_k}, t_m^{t}\right)$$

$$E^{t}_{pDist} = \sum_{t} P_{Dist}\left(p_i^{ss_k}, d^{t}_{\theta}\right)$$

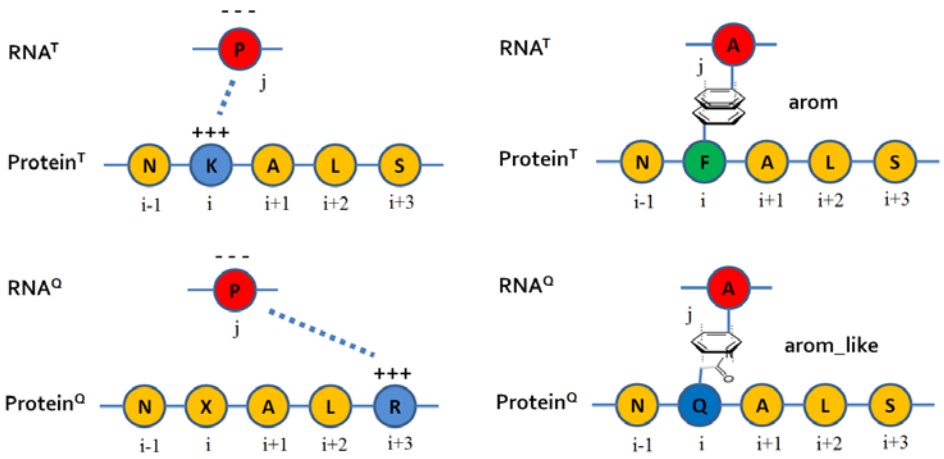
The secondary structure of the target protein will be predicted by *RaptorX* internally using PSIPRED method [88]. The probability of each residue being α-helix, β-sheet or loop, namely p_{ss_k}, will be incorporated into the function. The overscript (e.g., $\overset{t}{X}$ or $\overset{q}{X}$) indicates whether the profile (*X*) is retrieved from template or query. To compare the protein interface threading scores for proteins with different length, we normalize the final score (calculated from query threading scores minus template threading scores) by "effective length", in which greater weight will be placed on the region that is closer to interface, we transform the distance ($\overset{t}{d}_\theta$) using a sigmoid function (ψ):

$$\psi\left(\overset{t}{d}_\theta\right)=\frac{1}{1+e^{-A\left(\overset{t}{d}_\theta - s\right)}}$$

Then effective length is the summation of $\psi\left(\overset{t}{d}_\theta\right)$ for all the positions in template sequence. Here constant c defines the minimum weight after transformation, A controls the overall steepness of sigmoid curve and *s* affects descending rate for small $\overset{t}{d}_\theta$ values. We use, s=8.0, A= -0.8 for maximum performance. Considering the nature of RPI, we take into account two nonspecific interaction schemes in protein threading scoring or protein-RNA interface threading scoring (**Fig. 3.3**). First, salt bridge (electrostatic attraction) is considered as sequence-independent interaction. If a query alignment position (*q*) fails to form salt bridge interaction with RNA (as indicated by the template), we search the surrounding *q* ±3 positions for Arg or Lys, and use the best score when calculating the contribution of this position to E^q_{pInter} (**Fig. 3.3 left**). Second, base stacking interaction is considered as type-independent interaction. In this exception, if a query alignment position (*q*) is unable to form a specific

stacking interaction (t_m) (as indicated by the template), we calculate the interaction propensity using other t'_m s which are classified in the same category of t_m (**Fig. 3.3 right**). The gap penalty coefficient α is -2 when there presents a gap in the query sequence.

Fig. 3.3. **Scheme of the nonspecific interactions in *PRInter* scoring**. (A) salt bridge (electrostatic attraction) is considered as sequence-independent interaction. (B) base stacking interaction is considered as type-independent interaction.



3.3.3.2 RNA threading and scoring

LocARNA utilizes dynamic programming (DP) for RNA alignment with dedicate consideration of secondary structure of nucleotides [89, 90]. The secondary structure and the base pair probability of the query RNA is predicted by CentroidFold [91], one of the most robust RNA secondary structure prediction tools benchmarked by *CompaRNA* [92]. The base pairing probabilities for each nucleotide in the query sequence are predicted using CentroidFold, and will be used as inputs for *LocARNA*. Similar to that for protein threading, we have RNA threading score as following:

$$E_{pThread} = \left(E^{q}_{pInter} + E^{q}_{pDist} + E^{q}_{pGap_{penalty}}\right) - \left(E^{t}_{pInter} + E^{t}_{pDist}\right)$$

$$E^{norm}_{pThread} = E_{pThread} \Big/ \sum_{t} \psi\left(\overset{t}{d}_{\theta}\right)$$

Where:

$$E^{q}_{pInter} = \sum_{t,q\in aligned} \sum_{m\in t} \sum_{k\in\{\mathrm{H,E,C}\}} \left[p_{ss_k} \cdot P_{Inter}\left(\overset{q}{p}_{i}^{ss_k}, \overset{t}{t}_{m}\right)\right]$$

$$E^{q}_{pDist} = \sum_{t,q\in aligned} \sum_{k\in\{\mathrm{H,E,C}\}} P_{Dist}\left(\overset{q}{p}_{i}^{ss_k}, \overset{t}{d}_{\theta}\right)$$

$$E^{q}_{pGap_{penalty}} = -\sum_{t\neq -,q=-} \left[P_{Dist}\left(\overset{t}{p}_{i}^{ss_k}, \overset{t}{d}_{\theta}\right) + \alpha \cdot \psi\left(\overset{t}{d}_{\theta}\right)\right]$$

$$E^{t}_{pInter} = \sum_{t} \sum_{m\in t} P_{Inter}\left(\overset{t}{p}_{i}^{ss_k}, \overset{t}{t}_{m}\right)$$

$$E^{t}_{pDist} = \sum_{t} P_{Dist}\left(\overset{t}{p}_{i}^{ss_k}, \overset{t}{d}_{\theta}\right)$$

Besides R_{Inter} and R_{Dist}, no nonspecific threading term (e.g., salt or base stacking terms) are applied. Since it is infeasible to estimate the importance of RNA secondary structure *a priori*, we perform greedy searches for the optimal values of "structweight" parameter at an interval of 50, and "indel" and "indel-opening" at an interval of 0.5, and the RNA alignment that obtains the best score will be retained. Other parameters are identical to those in protein threading.

3.3.3.3 Protein-RNA interface threading and scoring

Based on the protein alignment provided by *RaptorX* and RNA alignment provided by *LocARNA*, we are able to predict the query interface and align to the template interface. Here we hypothesized that 1) the interaction type of each residue/nucleotide at interface can be inferred from its homologous interface; 2) the missing residues, namely gaps, at interface alignment are detrimental to the binding. The performance of protein-RNA interface threading is scored by $E^{norm}_{prThread}$, as following:

$$E_{prThread} = \left(E^{q}_{prInter} + E^{q}_{prGap_{penalty}}\right) - \left(E^{t}_{Inter}\right)$$

$$E^{norm}_{prThread} = E_{prThread} / N_{t^p \otimes t^r}$$

Where:

$$E^{q}_{prInter} = \sum_{\substack{t^p \otimes t^r \\ t^p, q^p \in protein \\ t^r, q^r \in RNA}} \sum_{m \in t} \begin{cases} \sum_{k \in \{H,E,C\}} \sum_{l \in \{WC, nP\}} \left[p_{ss_k} p_{ss_l} \cdot PR_{Inter} \left(p_i^{ss_k, q^p}, r_j^{ss_l, q^r}, t_m^{t} \right) \right] \\ \arg\max_{q'^p \in [q^p - 3, q^p + 3]} \sum_{k \in \{H,E,C\}} \sum_{l \in \{WC, nP\}} \left[p_{ss_k} p_{ss_l} \cdot PR_{Inter} \left(p_i^{ss_k, q'^p}, r_j^{ss_l, q^r}, t_m^{t} \right) \right], \ if\ m \notin T_{m,i}\ and\ m = salt \\ \sum_{m' \in T_{m,i}} \sum_{k \in \{H,E,C\}} \sum_{l \in \{WC, nP\}} \left[p_{ss_k} p_{ss_l} \cdot PR_{Inter} \left(p_i^{ss_k, q^p}, r_j^{ss_l, q^r}, t_{m'}^{t} \right) \right], \ if\ m \notin T_{m,i}\ and\ m \in base\ stackin \end{cases}$$

$$E^{q}_{prGap_{penalty}} = -\left(\sum_{t \neq -, q = -} \omega_{half} + \sum_{t = -, q \neq -} \omega_{half} + \sum_{t = -, q = -} \omega_{full} \right)$$

$$E^{t}_{prInter} = \sum_{\substack{t^p \otimes t^r \\ t^p, q^p \in protein \\ t^r, q^r \in RNA}} \sum_{m \in t} PR_{Inter} \left(p_i^{ss_k, t^p}, r_j^{ss_l, t^r}, t_m^{t} \right)$$

Similar to the calculation for protein threading score, we applied the salt and stacking corrections for nonspecific interaction schemes in interface threading. Here $t^p \otimes t^r$ denotes the interface element in the template or the corresponding query interface, where a template protein residue in the template (t^p) interacts with a template nucleotide (t^r). Therefore, $N_{t^p \otimes t^r}$ denotes the number of direct contacts in the template, which will be used as a normalization factor when computing $E^{norm}_{prThread}$. As the secondary structure states of amino acids and nucleotides are all from predictions, the interaction propensity are computed as the sum of dot product of the residues with all secondary structure states (for amino acids: H, E, C and for nucleotides: WC, nP).

3.3.4 Develop Random Forest classification models

3.3.4.1 Collect interface profiles to train classification models

Unlike protein-protein interaction, the publicly available resources for protein-RNA interaction are greatly limited. Furthermore, it is generally more dangerous to scramble the positive dataset to derive the non-interacting negative controls for protein-RNA interaction, as RNA only contains four types of residue (variables), where the probability of chance binding is significantly higher than that of protein when data shuffling is performed. Here we train the machine learning classifier with three resources: *NPInter* [93], *RBPDB* [94] and *NNBP* [95]. Briefly, we have collected 14,623 protein-RNA positive pairs from *NPInter*, and 3,649 negative pairs from *RBPDB* using PSSM motif scanning searching for RNAs in *NPInter* that are less likely to bind (<5%). In addition, for each protein in *NNBP* that is confirmed not to interact with any nucleosides, we randomly selected 50 RNAs from *NPInter*, and formed 12,500 negative pairs.

Two independent datasets were collected for external validation. First dataset contains 11,709 protein-mRNA interaction pairs from *Saccharomyces* genome database (SGD) [96]. The 4,706 negative pairs generated by random shuffling were obtained from *RPIseq* validation set, which was retrieved from [54]. This independent dataset was used in the previous method benchmarks, such as *RPISeq* [54] and Pancaldi and Bähler et al. [97]. The second dataset was compiled from 42 most recent discoveries of miRNA-protein interactions (**Table 3.2**), and we will compare the performances of interface threading method with those by *RPISeq* using these two validation sets.

Table 3.2. External validation dataset (II). In the binding column, 0-no binding detected, 1-binding is observed. CoIP: coimmunoprecipitation. MS: mass spectrum. WB: western blot. RIP: RNA immunoprecipitation.

Gene	Uniprot ID	miRNA	Experiment	PubMed_ID	Bind?
HNRNPA1	P09651	pre-mir-18a	CoIP	17558416	1
HNRNPA1	P09651	pre-let-7a-1	CoIP	20639884	1
HNRNPA1	P09651	pre-mir-101-1	RNA chromatography MS; WB	18995836	1
HNRNPL	P14866	pre-let-7a-1	RNA chromatography MS; WB	18995836	1
PCBP2	Q15366	mir-181b-5p	RIP	20211135	0
PCBP2	Q15366	mir-330-5p	UV crosslinking; RIP	20211135	0
PCBP2	Q15366	mir-328	UV crosslinking; RIP; EMSA	20211135	1
PTBP1	P26599	pre-mir-101-1	RNA chromatography MS; WB	18995836	1
HNRNPK	P61978	mir-328	Preliminary data	20211135	1
ELAVL1	Q15717	mir-29b-3p(mut)	CoIP	23901138	0
ELAVL1	Q15717	mir-29b-3p	CoIP	23901138	1
KHSRP	Q92945	pre-mir-21	CoIP; NMR; UV crosslinking	19458619	1
KHSRP	Q92945	pre-mir-1-2	crosslinking; CoIP	23221640	1
KHSRP	Q92945	pre-let-7a-1	CoIP; NMR; UV crosslinking	20639884; 19458619	1
HNRNPD	Q14103	pre-mir-155	Preliminary data	19423639	0
KHSRP	Q92945	pre-mir-23b	CoIP	19423639	0
ZFP36	P26651	pre-mir-155	CoIP	19423639	0
KHSRP	Q92945	pre-mir-155	CoIP	19423639	1
LIN28b	Q6ZN17	pre-let-7f1	Crystalized	22078496	1
LIN28b	Q6ZN17	pre-let-7d	Crystalized	22078496	1
LIN28b	Q6ZN17	pre-let-7g	Crystalized	22078496	1
TLR7	Q9NYK1	let-7b-5p	Indirect assay with exogenous miRNA	22610069	1
TLR7	Q9NYK1	let-7a-5p	Indirect assay with exogenous miRNA	22610069	1
TLR7	Q9NYK1	let-7c	Indirect assay with exogenous miRNA	22610069	1
TLR7	Q9NYK1	let-7g-5p	Indirect assay with exogenous miRNA	22610069	1
TLR7	Q9NYK1	mir-599	Indirect assay with exogenous miRNA	22610069	1
TLR7	Q9NYK1	mir-124-3p	Indirect assay with exogenous miRNA	22610069	0
TLR8	Q9NR97	mir-21-5p	CoIP; Colocalization	22753494	1
TLR8	Q9NR97	mir-29a-3p	CoIP; Colocalization	22753494	1
TLR8	Q9NR97	mir-16-5p	CoIP	22753494	0
TLR8	Q9NR97	mir-147a	Indirect assay with exogenous miRNA	22753494	1

QKI6	Q9QYS9	mir-20a-3p	CoIP	22751500	1
QKI6	Q9QYS9	pre-mir-20a	CoIP	22751500	0
QKI6	Q9QYS9	mir-18a-5p	CoIP	22751500	0
QKI6	Q9QYS9	mir-20a-5p	CoIP	22751500	0
QKI6	Q9QYS9	pre-mir-7-1	CoIP	23319046	1
SND1	Q7KZF4	pre-miR-92a-2	RIP	23770094	1
SND1	Q7KZF4	mir-17-5p	RIP	23770094	1
SND1	Q7KZF4	mir-18a-5p	RIP	23770094	1
SND1	Q7KZF4	mir-19a-5p	RIP	23770094	1
SND1	Q7KZF4	mir-20a-5p	RIP	23770094	1
SND1	Q7KZF4	mir-92a-5p	RIP	23770094	1

3.3.4.2 *RPIT*-RF model

The ultimate goal is to determine whether the query protein interacts with the query RNA based on the interface score profiles computed previously. Since only a few protein-RNA pairs interact *in vivo*, the main challenge is to discriminate the true interactions from the false ones. Here, we extract a vector of interface profile, $X_{Interface}$, and feed this profile to various classifiers, which compute the probability of interacting:

$$p = \zeta\left(X_{Interface}\right)$$
$$X_{Interface} = \left\{\left\langle pThread\right\rangle, \left\langle rThread\right\rangle, \left\langle prThread\right\rangle\right\}$$

Where:

$$\left\langle pThread\right\rangle = \left\{\begin{matrix} tP, qP, \varepsilon P, iP, rap_{score}, rap_{pval}, rap_{NEFF}, E^{q}_{pInter}, N^{q}_{pInter}, E^{t}_{pInter}, N^{t}_{pInter}, \\ E^{q}_{pDist}, E^{t}_{pDist}, E^{q}_{pGap_{penalty}}, N^{q}_{pGap_{penalty}}, E^{q}_{pGap_{affinity}}, N^{q}_{pGap_{affinity}}, E_{pThread}, E^{norm}_{pThread} \end{matrix}\right\}$$

$$\left\langle rThread\right\rangle = \left\{\begin{matrix} tR, qR, \varepsilon R, iR, loc_{score}, loc_{ssweight}, loc_{gap}, E^{q}_{rInter}, N^{q}_{rInter}, E^{t}_{rInter}, N^{t}_{rInter}, \\ E^{q}_{rDist}, E^{t}_{rDist}, E^{q}_{rGap_{penalty}}, N^{q}_{rGap_{penalty}}, E^{q}_{rGap_{affinity}}, N^{q}_{rGap_{affinity}}, E_{rThread}, E^{norm}_{rThread} \end{matrix}\right\}$$

$$\left\langle prThread\right\rangle = \left\{E^{q}_{prInter}, N^{q}_{prInter}, E^{t}_{prInter}, N^{t}_{prInter}, E^{q}_{prGap_{penalty}}, N^{q}_{prGap_{half}}, N^{q}_{prGap_{full}}, E_{prThread}, E^{norm}_{prThread}\right\}$$

tP and tR are the length of template protein or RNA. qP and qR are the length of query protein (or RNA). εP and εR are the effective length of the protein (or RNA), which are employed to normalize the interaction score. iP and iR are the sequence identity between the template and query protein (or RNA). $rap_{score}, rap_{pval}, rap_{NEFF}$ are the threading score, p-value, and NEFF value calculated by *RaptorX*. $loc_{score}, loc_{ssweight}, loc_{gap}$ are the best threading score, secondary structure weight, and gap penalty score from *LocARNA*. Feature importance showed that features other than $rap_{score}, rap_{pval}, rap_{NEFF}$, $E_{pThread}, E^{norm}_{pThread}$, $E_{rThread}, E^{norm}_{rThread}$

$E_{prThread}, E_{prThread}^{norm}$ are less informative (data not shown). Thus, the final classifier will only have the above nine features as the attributes.

3.3.4.3 Metrics for model quality assessment

We evaluated the predictive ability of classifiers by sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 measurement, Matthews correlation coefficient (MCC), overall accuracy, and ROC AUC. These metrics are calculated as following, in which TP = true positive, FP = false positive, TN = true negative, FN = false negative:

Sensitivity = $TP / (TP + FN)$

Specificity = $TN / (TN + FP)$

PPV = $TP / (TP + FP)$

NPV = $TN / (TN + FN)$

F1 = $2 \times sensitivity \times PPV / (sensitivity + PPV)$

MCC = $(TP \times TN - FP \times FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$

Accuracy = $(TP + TN) / (TP + TN + FP + FN)$

The F-measurement is a harmonic mean of precision and recall. F-measurement, e.g., F1 value, ranges from 0 to 1, and value close to 1 indicates perfect classifier. MCC value, often known as φ-coefficient, is essentially a correlation coefficient between the observed and predicted classification. Its value ranges from -1 to 1 where 1 indicates prefect classification, 0 means random, and -1 indicates a total disagreement. The ROC AUC evaluates the

performance of binary classifier with varied discrimination threshold. Its value ranges from 0 to 1 and AUC = 1 indicates a perfect classifier and AUC = 0.5 indicates random classifier.

3.3 Results: Interface threading approach to predict RNA-protein binding

3.3.1 nrPR database

The quality and diversity of *nrPR* template database may directly affect the accuracy and applicability of interface threading implementation. *nrPR* database consists of 5,471 non-redundant protein-RNA pairs, including 69% crystal structures, 2% NMR structures and 29% electron microscopy structures (**Fig. 3.4A**). A majority of crystal structures (76.5%) acquire acceptable resolutions (resolution < 3.5Å), with most around 3.0Å (**Fig. 3.4B**). The median resolution is 3.1Å. Although the quality of 3D structures are not ideal compared with other collections, we think it is acceptable if considering the tradeoff between database coverage and quality. Indeed, the statistical scoring functions in this implementation can tolerate the trivial errors in structural models, because they only consider a binary response (interface or non-interface residue) for interaction propensity calculation and the distance range (d<cutoff) for distance propensity calculation.

Analysis of the lengths of interacting protein and RNA pairs has shown three major clusters in *nrPR* (**Fig. 3.4C**). First cluster involves small to large size protein (75-500 aa) interacting with small RNA (<500 nt). Second and third clusters include medium size protein (100-250 aa) interacting with medium (1000-2000 nt) or large size RNA (2500-3500 nt), respectively. Regardless the wide variation in macromolecular lengths, the number of interfacial amino

acids and nucleotides show a clear correlation (R^2 = 0.84) (**Fig. 3.4D**), in which amino acid on average interacts with 0.68 nucleotide at interface region (amino acid to nucleotide ratio (ANR) = 1.48 ± 0.99). This ANR is in contrast to that of protein-DNA interface, whose ANR is about 2 (24 ± 6 aa vs. 12 ± 3 nt [98] and 52 ± 25 aa vs. 23 ± 9 nt [51]). Our data is also distinct from any previous statistical analyses using small dataset (< 200 samples), which usually reported ANR > 2.5 [51]. We observed a considerable variation of ANR values (ranging from 0.33 to 12.5), which indicates that diverse protein/RNA families have been collected. Interesting, we find length of protein (or RNA) non-informative to predict of number of interface residue, as there are no correlations between length of protein (RNA) and number of interfacial residues (nucleotides) (**Fig. 3.4E-F**). PCA using triad conjoint descriptors demonstrates that the *nrPR* database is absent of significant clusters (**Fig. 3.5A**), in which first two principle components (PCs) only accounts for <10% variance amongst all protein sequences. Pairwise-sequence/secondary structure/interaction similarity distribution could be fitted to a normal distribution, $p \sim \mathcal{N}(mean = 42.15\%, \sigma^2 = 66.77)$ with low RMSE = 3.02 (**Fig. 3.5B**). All these data suggest that *nrPR* database represents an unbiased collection of RNA-protein interfaces, and the diversity in sequence/structure/interaction should be sufficient to achieve statistical power for scoring functions implementations.

Fig. 3.4. **Statistics of nrPR database.** (A) composition of 3D structures. (B) Distribution of resolution for all cystal structures. (C) Distribution of the length of protein vs. length of RNA for protein-RNA interacting pairs in *nrPR*. (D) Distribution of number of interfacial amino acid and interfacial nucleotides for the interfaces in *nrPR*. (E) Distribution of the number of interfaical residues vs. length of protein. (F) Distribution of the number of interfaical nucleotides vs. length of RNA.

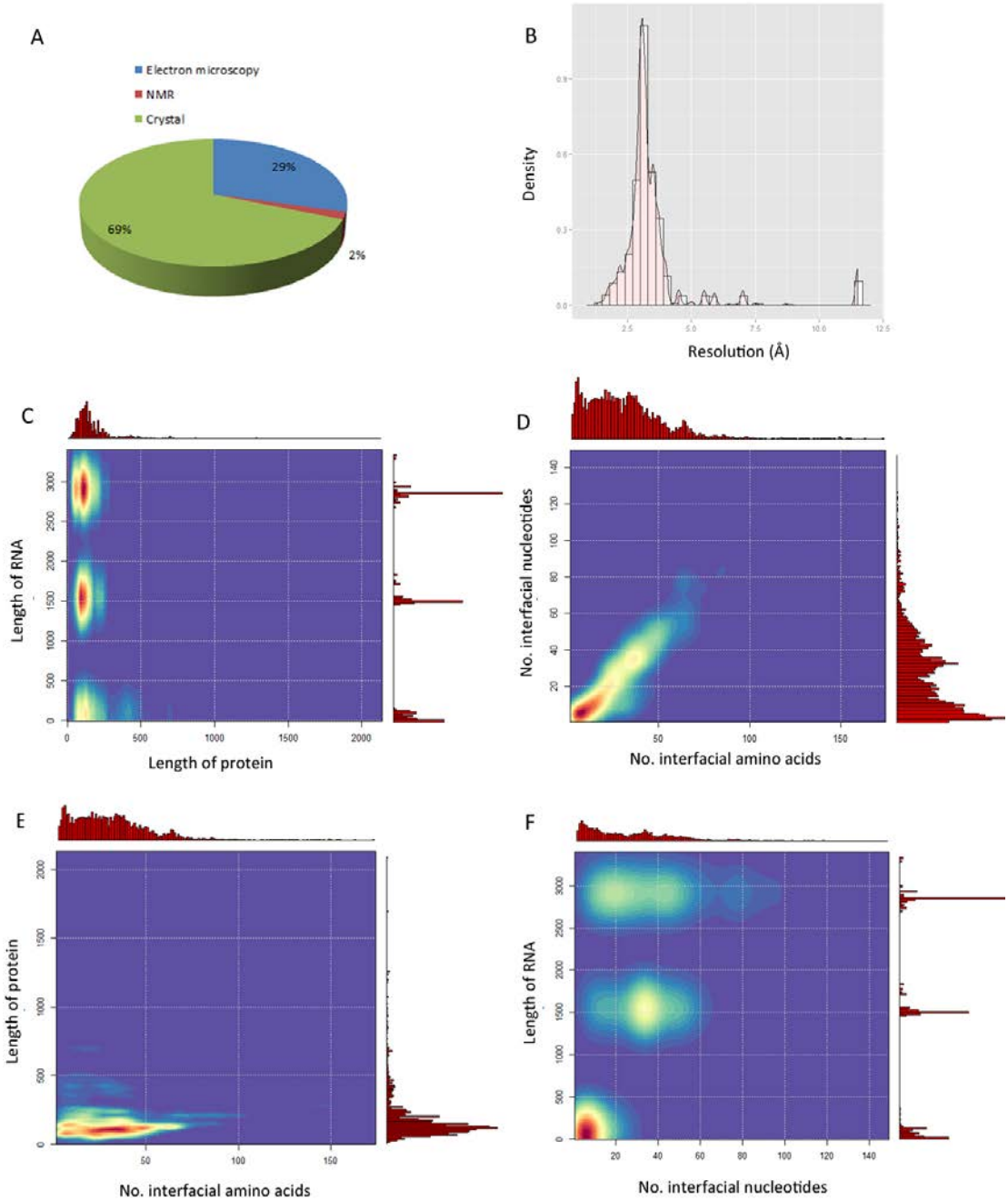
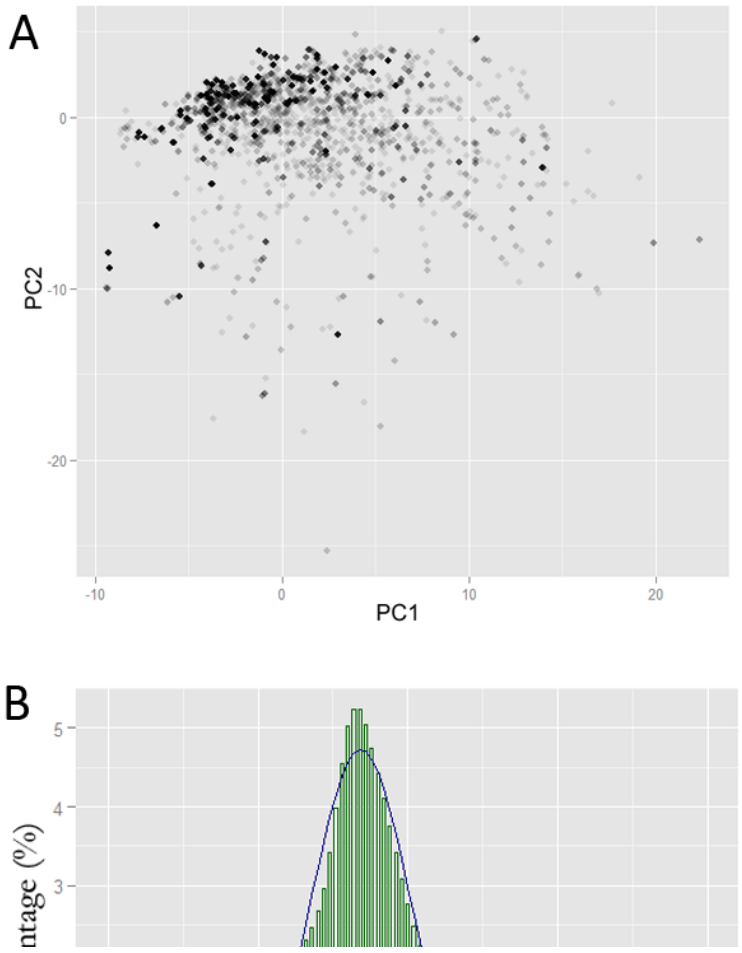


Fig. 3.5. **Sequence and structural diversity of nrPR database.** (A) Scatter plot of first two principle components of nrPR database using conjoint triad descriptors. (B) Distribution of pairwise-RNA-protein pair similarity in nrPR database. The blue line indicates the normal distribution of the pairwise similarity.



3.3.2 Statistical scoring functions

Table 3.3 and **Table 3.4** summarized some basic statistics for amino acid or nucleotide at protein-RNA interface. We find that amino acids in coil state are statistically more favorable at protein-RNA interface than other secondary structure states ($p < 0.01$) (**Fig. 3.6**). This agrees with previous finding that protein-RNA interfaces are packed less tightly than those of protein-DNA and protein-protein interface.

Fig. 3.7A shows the heat map for ***PInter*** scoring function. Consistent to the amino acid preference [52] and secondary structure preference we have observed, general vdW interaction potentials exhibit significant variances ($p = 5.15e^{-15}$ for amino acid factor and $p = 1e^{-11}$ for secondary structure factor, two-way ANOVA). Coiled amino acid and positively charged residues are more favorable to protein-RNA interface. Other type-specific potentials, such as arom, arom_l, cpi, hy and salt, hbp_1, hbp_2 obtain expected preferences to specific amino acids. T, S, R, K, H, Q, N are more likely to be recognized by RNA nucleotides by forming more than two H-bonds with their sidechain atoms ($p < 1e^{-5}$). Intriguingly, negatively charged residue (D, E) as well as these seven residues (T, S, R, K, H, Q, N) are more likely to recognize specific nucleotide judged by high hbr_2 propensities ($p < 1e^{-5}$). Later we will show in **Fig. 3.8** that even if D and E are generally disfavored on protein-RNA interface, they preferably form H-bonds with specific nucleotides if they happen to be on the interface.

In contrast, the potential of RNA nucleotides to interaction with protein depend more on secondary structure states (**Fig. 3.7B**). Arom, arom_l, cpi, hbr_2, and hy interactions have

greater propensity to occur on unpaired/noncanonical base paired nucleotides than Watson-Crick/G-U wobble base paired ones. In terms of sequence-specific interaction, guanine and cytosine are most likely to be "recognized" by forming more than two H-bonds with protein residues. However, nucleotide type and secondary structure states cannot determine the interaction propensity to form backbone recognition ($p_{bb\_p,\ ss}$ = 0.61, $p_{bb\_p,\ na}$ = 0.94, $p_{bb\_r,\ ss}$ = 0.08, $p_{bb\_r,\ na}$ = 0.42), protein sequence recognition (hbp_1, hbp_2), salt and vdW interactions according to two-way ANOVA test.

Fig. 3.8 summarizes the interaction propensities between interfacial protein residues and RNA nucleotides. The nonspecific interactions (arom, arom_l, bb_p, bb_r, cpi, hy, salt and vdw) show similar patterns with those of ***PInter***. With respect to hbr_2 and hbp_2 propensities, some favorable amino acid-nucleotide specific interactions have been detected (superscripts indicate the secondary structure state):

1. A^{nP} is favorably recognized by V^{E} / T / S / Q / N^{E} / K^{E} / I^{E} / H^{E} / D^{E} / C^{C} / A^{E}, whereas A^{nP} : T, A^{nP} : S, A^{nP} : N^{E}, A^{nP} : K^{E} shows more specific bilateral recognition judged by both high hbp_2 and hbr_2 propensities, and other interactions are unilateral recognition only to nucleotide. Compared to the bilateral recognitions involved with other nucleotides^{nP}, there is a statistically significant weaker propensities for A^{nP} : R (p = 3.5e^{-5}). **Fig. 3.9A-C** shows some typical interactions found for nucleotide^{nP} : R.
2. C^{nP} is preferably recognized by T / S / R / M^{C} / K, in which C^{nP} : T, C^{nP} : S, C^{nP} : R and C^{nP} : S are bilateral recognitions, whereas C^{nP} : M^{C} is unilateral to nucleotide.

Compared to the bilateral recognitions involved with other nucleotides^{nP}, there is a statistically significant weaker propensities for C^{nP} : Q (p = 0.0007).

3. G^{nP} is preferably recognized by T / S / R / Q / N^{H} / N^{C} / M^{C} / K / E / D. All except C^{nP} : M^{C} recognize both amino acid and nucleotide types bilaterally. G^{nP} : D and G^{nP} : E interactions are uniquely enriched for G^{nP} (p = 0.015 for G^{nP} : D and p = 0.0006 for G^{nP} : E). **Fig. 3.9D-E** illustrates typical sequence-specific interactions of G^{nP} : D and G^{nP} : E.
4. U^{nP} is preferably recognized by T / S / R / Q / N / M^{C} / K / E / D. All except C^{nP} : M^{C} recognize both amino acid and nucleotide types bilaterally. U^{nP} : N, in particular, is enriched compared with other nucleotide-Gln sequence-specific interactions (p = 0.028). **Fig. 3.9F** demonstrates a representative interaction pattern of U^{nP} : N.
5. Surprisingly, G^{WC} is the only paired nucleotide^{WC} that have significantly greater propensity to be recognized by T, S, R, Q, N, K ($p < 10^{-5}$). Examples of G^{WC}-recognition interaction are showed in **Fig. 3.9G-H**.

Amino acids demonstrate distinctive propensities to be on or close to protein-RNA interface (**Fig. 3.10**). We can classify 20 amino acids into several groups based on their respective distance propensity profile. (1) Non-aromatic, hydrophobic residues (Ala, Ile, Leu, Val), especially in helix or sheet forms, are strongly unfavorable to protein-RNA interface until 5Å. (2) Negatively charged residues (Asp and Glu) are unfavorable in any secondary structure state, even at 10Å. (3) Sulfur-containing residues (Cys and Met) slightly prefer the interface regions when they are in coil state, but disfavor when in helix or sheet state. (4) Hydrophobic

residues with carbon-only ring (Phe, Pro) have neutral preference to the interface when in coil state and slightly unfavorable in other states. (5) Gly has neutral preference at all states. (6) Neutral hydrophilic residues (Asn, Gln, Ser, Thr) in coil form slightly favor the interface, but neutral when in other forms. (7) Aromatic residues whose sidechains can be H-bond donor/acceptor (Tyr, Trp) strongly favor the interface when in coil state, but neutral in other states. (8) Positively-charged residues (His, Lys and Arg) significantly favor the protein-RNA interface in any states. In comparison, the distance potential for RNA nucleotides fail to show any significant difference between A, U, G, C and the potentials are always neutral at any distance (**Fig. 3.11**), indicating that the distance propensity for RNA nucleotides might be non-informative for RPI prediction.

Table 3.3. Statistics of protein amino acids in *nrPR* database. Secondary structure states were considered: H=helix, E=sheet, C=coil. The vdw interaction statistics for each residue type were not shown as it equals to the total number of interfacial residue.

AA-ss	Interface / total	% Interface	Mean dist.	arom	arom_l	hy	cpi	hbr_2	hbr_1	hbp_2	hbp_1	bb_p	bb_r	salt
A	9989 / 73833	13.53%	6.04 ± 5.48											
A-H	2873 / 31334	9.17%	6.42 ± 5.49	0	0	117	0	0	42	0	0	326	284	0
A-E	812 / 9203	8.82%	5.97 ± 5.11	0	0	26	0	5	30	0	0	87	62	0
A-C	6304 / 33296	18.93%	5.72 ± 5.55	0	0	343	0	7	214	0	0	1151	986	0
C	779 / 6927	11.25%	5.99 ± 5.43											
C-H	144 / 1892	7.61%	7.30 ± 5.41	0	0	0	0	0	13	1	15	11	13	0
C-E	117 / 1682	6.96%	6.19 ± 5.23	0	0	0	0	0	3	0	14	6	17	0
C-C	518 / 3353	15.45%	5.17 ± 5.40	0	0	0	0	4	32	11	91	44	110	0
D	4353 / 38901	11.19%	7.06 ± 5.69											
D-H	890 / 10020	8.88%	7.59 ± 5.67	0	10	0	0	7	40	23	203	36	210	0
D-E	615 / 4918	12.51%	6.68 ± 5.39	0	0	0	0	10	81	9	177	20	115	0
D-C	2848 / 23963	11.88%	6.92 ± 5.75	0	45	0	0	27	233	59	491	271	558	0
E	5454 / 58839	9.27%	7.39 ± 5.65											
E-H	1670 / 23926	6.98%	7.78 ± 5.63	0	13	0	0	78	116	84	274	91	241	0
E-E	889 / 8339	10.66%	6.94 ± 5.41	0	17	0	0	31	80	40	214	88	225	0
E-C	2895 / 26574	10.89%	7.21 ± 5.73	0	23	0	0	21	160	48	420	288	561	0
F	4421 / 28797	15.35%	6.17 ± 5.47											
F-H	716 / 8617	8.31%	6.92 ± 5.48	102	0	0	0	0	5	0	0	27	22	0
F-E	1147 / 7380	15.54%	5.90 ± 5.30	155	0	0	0	8	9	0	0	91	75	0
F-C	2558 / 12800	19.98%	5.84 ± 5.52	210	0	0	0	8	24	0	0	165	133	0
G	13582 / 63588	21.36%	5.63 ± 5.73											
G-H	1438 / 7849	18.32%	5.88 ± 5.82	0	0	0	0	0	55	0	0	281	228	0
G-E	1006 / 6958	14.46%	6.06 ± 5.50	0	0	0	0	2	8	0	0	139	129	0
G-C	11138 / 48781	22.83%	5.53 ± 5.74	0	0	0	0	9	465	0	0	2554	2147	0
H	6454 / 18531	34.83%	4.59 ± 5.60											
H-H	1427 / 4884	29.22%	5.38 ± 5.89	82	0	0	0	0	43	5	357	49	356	0
H-E	1125 / 3670	30.65%	4.82 ± 5.46	45	0	0	0	4	23	4	225	35	220	0
H-C	3902 / 9977	39.11%	4.12 ± 5.45	240	0	0	0	2	192	27	807	337	934	0
I	5946 / 51161	11.62%	6.15 ± 5.40											
I-H	1415 / 15768	8.97%	6.47 ± 5.54	0	0	273	0	0	33	0	0	54	21	0
I-E	1560 / 16466	9.47%	6.14 ± 5.23	0	0	215	0	12	37	0	0	145	111	0
I-C	2971 / 18927	15.70%	5.91 ± 5.41	0	0	456	0	3	58	0	0	265	213	0
K	25285 / 73269	34.51%	4.86 ± 5.73											
K-H	5910 / 22622	26.13%	5.61 ± 5.89	0	0	0	29	107	558	672	2986	367	3188	2859
K-E	3728 / 10347	36.03%	4.72 ± 5.45	0	0	0	17	43	296	340	1872	223	1962	1802
K-C	15647 / 40300	38.83%	4.50 ± 5.68	0	0	0	67	111	1128	1665	7112	1484	8068	6966
L	7856 / 73356	10.71%	6.25 ± 5.36											
L-H	1963 / 30793	6.37%	6.68 ± 5.37	0	0	180	0	0	21	0	0	146	125	0
L-E	1295 / 14028	9.23%	6.26 ± 5.10	0	0	203	0	4	12	0	0	83	67	0
L-C	4598 / 28535	16.11%	5.81 ± 5.43	0	0	615	0	5	62	0	0	401	336	0
M	3110 / 17551	17.72%	6.10 ± 5.64											
M-H	737 / 6807	10.83%	6.82 ± 5.69	0	0	136	0	0	35	0	94	76	128	0
M-E	380 / 2874	13.22%	5.90 ± 5.17	0	0	53	0	0	2	0	22	35	54	0
M-C	1993 / 7870	25.32%	5.57 ± 5.70	0	0	478	0	24	172	0	204	436	480	0
N	7356 / 30214	24.35%	5.88 ± 5.89											
N-H	1860 / 8224	22.62%	6.06 ± 5.83	0	28	0	0	53	149	129	616	52	609	0
N-E	643 / 3307	19.44%	6.48 ± 5.94	0	9	0	0	8	59	27	225	34	207	0
N-C	4853 / 18683	25.98%	5.70 ± 5.90	0	64	0	0	33	344	191	1404	482	1586	0
P	6565 / 36190	18.14%	6.28 ± 5.82											
P-H	656 / 5018	13.07%	7.14 ± 5.86	0	0	59	0	0	3	0	0	30	27	0
P-E	524 / 3728	14.06%	6.22 ± 5.30	0	0	144	0	0	5	0	0	82	77	0
P-C	5385 / 27444	19.62%	6.14 ± 5.87	0	0	663	0	3	71	0	0	417	352	0
Q	6814 / 29162	23.37%	5.64 ± 5.70											
Q-H	1897 / 11007	17.23%	6.32 ± 5.71	0	15	0	0	45	253	135	693	84	662	0
Q-E	1054 / 4282	24.61%	5.52 ± 5.52	0	16	0	0	41	108	87	302	67	335	0
Q-C	3863 / 13873	27.85%	5.17 ± 5.70	0	91	0	0	36	277	175	1112	454	1314	0
R	30332 / 72257	41.98%	4.16 ± 5.58											
R-H	8508 / 23446	36.29%	4.74 ± 5.81	0	241	0	267	119	334	1554	3508	357	4345	4741

R-E	4678 / 11723	39.90%	4.09 ± 5.29	0	95	0	91	53	206	743	1713	231	2164	2293
R-C	17146 / 37088	46.23%	3.83 ± 5.50	0	821	0	613	225	1013	2842	6502	1615	8271	8625
S	9482 / 40149	23.62%	5.94 ± 6.00											
S-H	1948 / 10576	18.42%	6.71 ± 6.08	0	0	0	0	21	217	136	826	229	837	0
S-E	1053 / 5872	17.93%	6.36 ± 5.73	0	0	0	0	19	113	67	424	50	403	0
S-C	6481 / 23701	27.34%	5.51 ± 5.99	0	0	0	0	142	676	443	2436	1032	2633	0
T	9265 / 41581	22.28%	5.65 ± 5.69											
T-H	1356 / 8832	15.35%	6.65 ± 5.88	0	0	0	0	6	115	64	558	146	566	0
T-E	1993 / 10204	19.53%	5.62 ± 5.35	0	0	0	0	61	114	114	573	240	691	0
T-C	5916 / 22545	26.24%	5.30 ± 5.73	0	0	0	0	89	478	349	2019	802	2268	0
V	7860 / 70838	11.10%	5.99 ± 5.22											
V-H	1179 / 16774	7.03%	6.42 ± 5.34	0	0	133	0	0	12	0	0	69	57	0
V-E	2053 / 25822	7.95%	5.99 ± 4.93	0	0	109	0	13	26	0	0	168	131	0
V-C	4628 / 28242	16.39%	5.74 ± 5.40	0	0	492	0	2	37	0	0	474	440	0
W	1602 / 6505	24.63%	5.54 ± 5.61											
W-H	362 / 2282	15.86%	6.41 ± 5.74	54	0	0	0	0	5	3	46	12	51	0
W-E	277 / 1417	19.55%	5.88 ± 5.49	26	0	0	0	0	13	5	57	11	62	0
W-C	963 / 2806	34.32%	4.71 ± 5.45	141	0	0	0	0	15	2	134	136	231	0
Y	6220 / 24749	25.13%	5.44 ± 5.70											
Y-H	1305 / 7706	16.93%	6.55 ± 5.87	134	0	0	0	2	50	36	444	35	470	0
Y-E	1590 / 6648	23.92%	5.45 ± 5.49	159	0	0	0	4	105	69	557	57	599	0
Y-C	3325 / 10395	31.99%	4.67 ± 5.58	242	0	0	0	3	110	86	890	175	988	0

Table 3.4. Statistics of RNA nucleotides in *nrPR* database. Secondary structure states were considered: WC=Watson-Crick or GU wobble base pair, nP=other noncanonical base pair or no base pair. The vdw interaction statistics for each residue type were not shown as it equals to the total number of interfacial residue.

NA-ss	Interface / total	% Interface	Mean dist.	arom	arom_l	hy	cpi	hbr_2	hbr_1	hbp_2	hbp_1	bb_p	bb_r	salt
A	38140 / 2070177	1.84%	9.27 ±6.54											
A-WC	8437 / 442783	1.91%	9.05 ±6.54	14	33	131	26	7	320	536	2270	933	3157	1752
A-nP	29703 / 1627394	1.83%	9.32 ±6.54	694	485	1603	414	183	1362	1755	7185	2828	9452	5911
U	27533 / 1570531	1.75%	9.16 ±6.50											
U-WC	11438 / 678220	1.69%	9.14 ±6.51	11	51	78	46	12	479	751	3187	1356	4407	2663
U-nP	16095 / 892311	1.80%	9.18 ±6.50	308	262	936	181	169	1437	1020	4655	1840	5480	3435
G	48756 / 2810137	1.74%	9.11 ±6.47											
G-WC	30060 / 1744408	1.72%	9.07 ±6.45	94	138	496	49	311	2034	1975	8429	3661	10876	6643
G-nP	18696 / 1065729	1.75%	9.19 ±6.50	326	392	919	258	538	1861	1548	5534	1808	6201	4194
C	36535 / 2082378	1.75%	9.12 ±6.46											
C-WC	25852 / 1509379	1.71%	9.13 ±6.49	33	69	175	37	6	673	1786	7624	2674	10315	6659
C-nP	10683 / 572999	1.86%	9.08 ±6.40	217	124	620	97	272	1000	909	3184	1204	3724	2356

Fig. 3.6. Percentage of interfacial protein residue with different secondary structure states. The boxplot was generated by ggplot2 library in R statistical package.

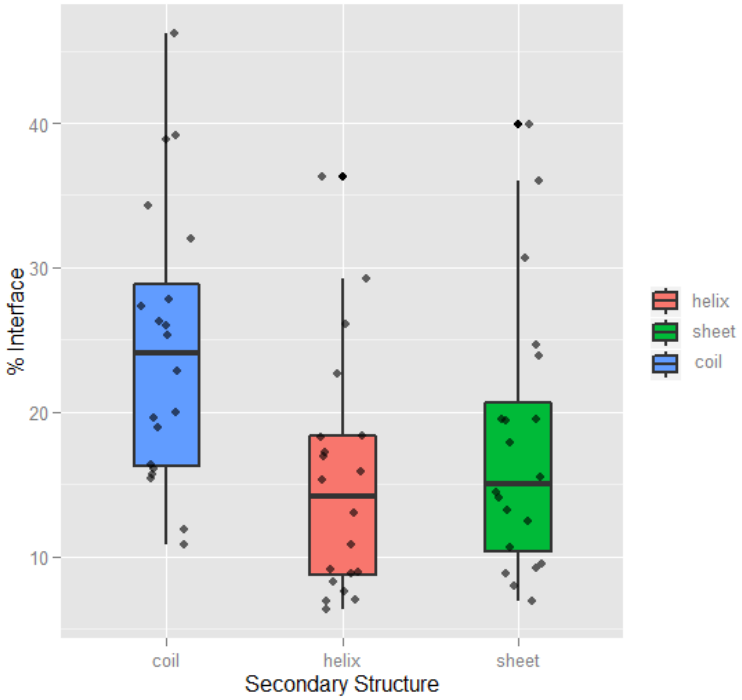


Fig. 3.7. Heat map of interaction potentials for protein or RNA residues. (A) amino acids in different secondary structure states. (B) nucleotides in different secondary structure states.

Fig. 3.8. Heat map of interaction potentials between protein-RNA residues.

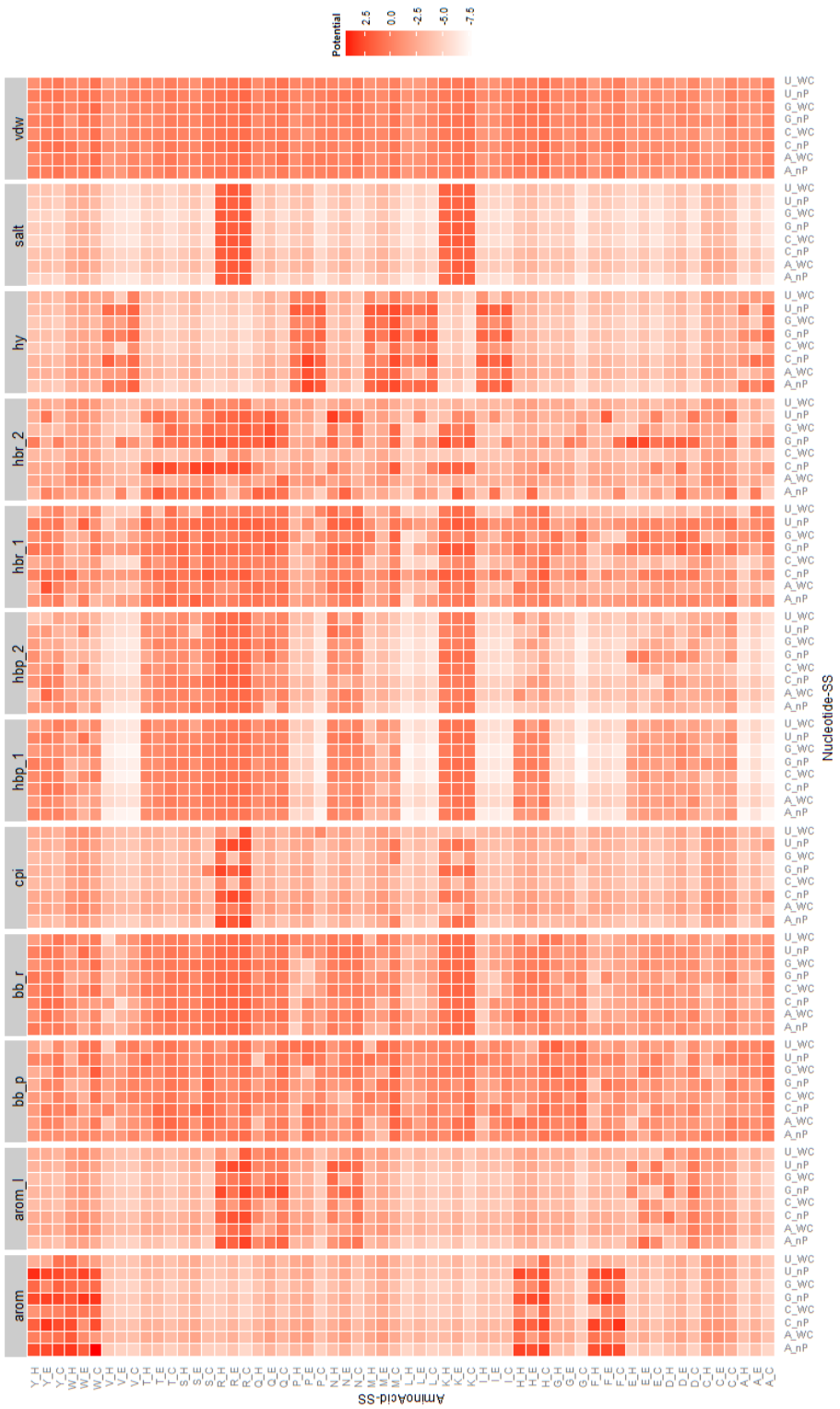


Fig. 3.9. **Representative bilateral sequence-recognition interaction on protein-RNA interface.** Intermolecular H-bonds are displayed as yellow dashes.

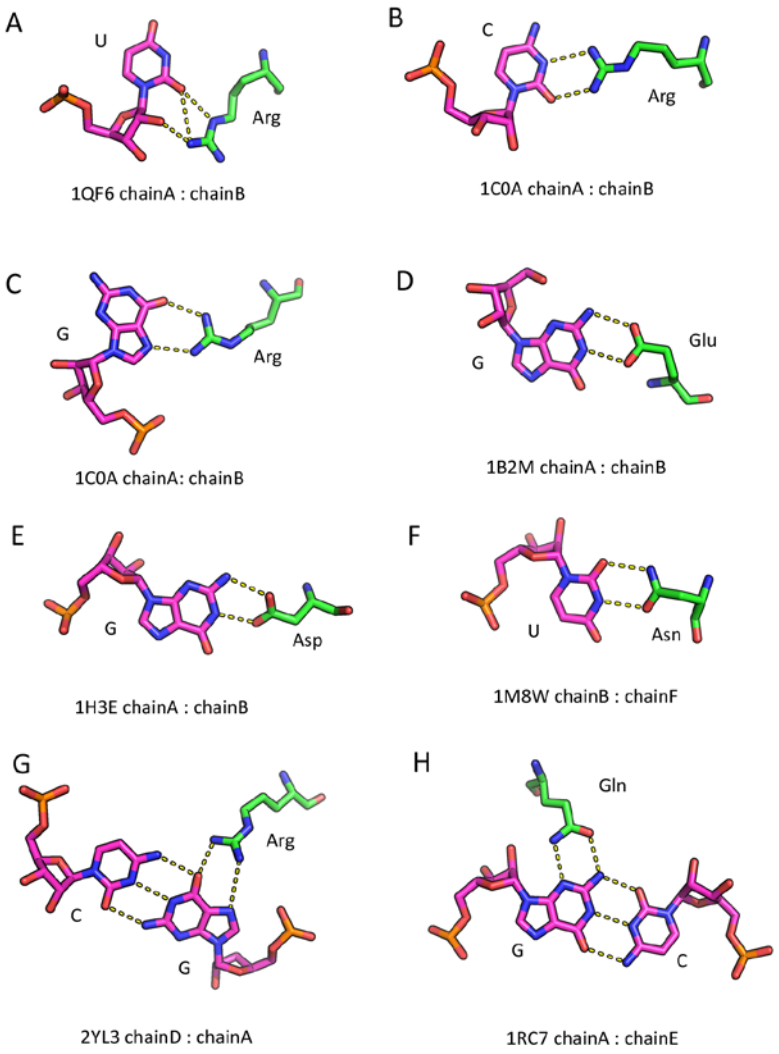


Fig. 3.10. **Distance potentials for protein residues.** The amino acids are sorted in a rough ascending order according to overall preferences to protein-RNA interface (most disfavored to most favored).

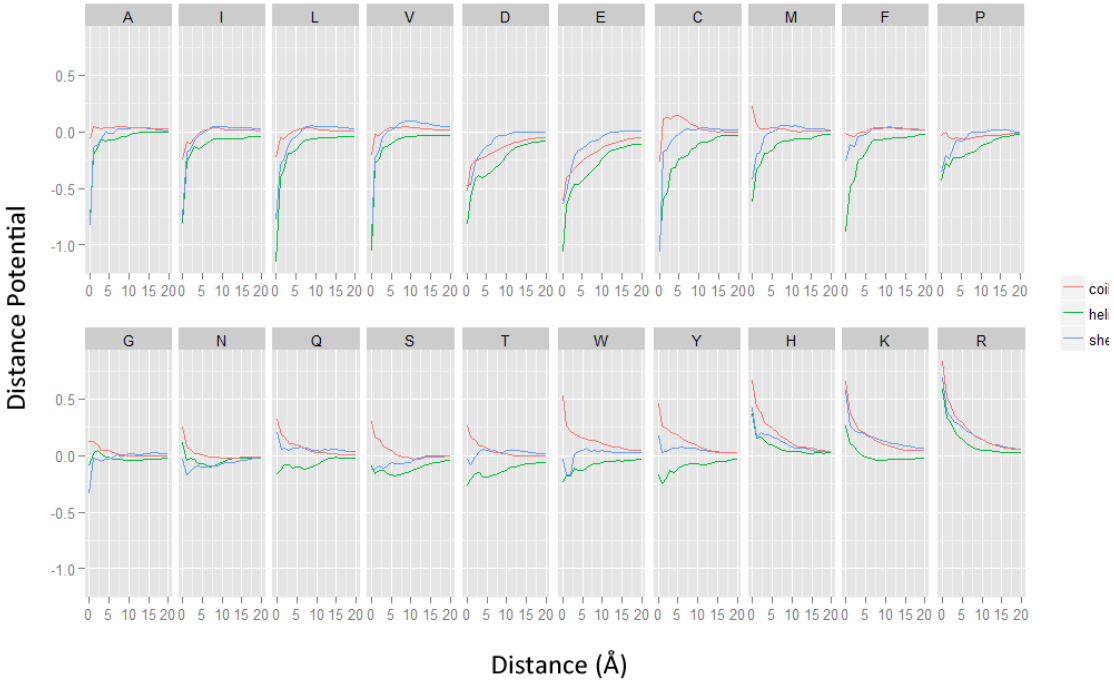
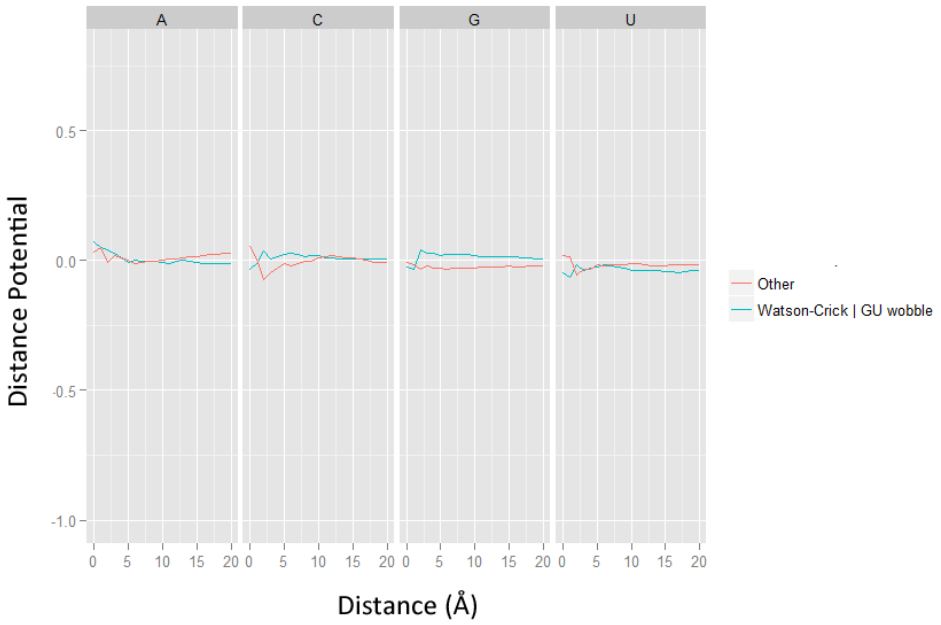


Fig. 3.11. **Distance potentials for RNA nucleotides.** Unlike those for amino acids, nucleotides are not sorted due to insignificant difference among groups.



3.3.3 Performance evaluation of RPIT

The conventional way of evaluating the robustness of a classifier is through leave-one-out or leave-group-out cross validation. However, due to the sparseness of protein profile in the training set (protein profile is heavily clustered due to the lack of mutagenesis data), these cross validation strategies could be unreasonably biased. Actually, we found that the classification metrics (e.g., sensitivity, specificity, PPV, NPV, ROC AUC, etc.) were way above 95% if 10-fold or 20-fold stratified cross validation was performed, even using different classifiers (data not shown). Instead, we assessed the robustness of model by leave-protein-out cross-validation (LPOCV). As a result, the random forest classifier (*RPIT*-RF) and quadratic discriminant analysis classifier (*RPIT*-QDA) outperformed other methods in terms of their outstanding ROC AUCs ($AUC^{RPIT\text{-}RF}$ = 0.93, $AUC^{RPIT\text{-}QDA}$ = 0.93) (**Table 3.5** and **Fig. 3.12A**). In particular, LPOCV of *RPIT*-RF resulted in outstanding predictive metrics (sensitivity = 0.89, specificity = 0.84, MCC = 0.69). Removal of any attributes from *RPIT*-RF (either protein, RNA or interface term) compromised the LPOCV performance (**Fig. 3.12B**), indicating that the interface threading score carry relevant information to make reasonable RPI predictions. Furthermore, random forest classifier using only protein and RNA sequence identity information predict significantly worse than the *RPIT*-RF (**Fig. 3.12B**), indicating that the naïve assumption that interfaces with similar protein/RNA sequences have similar binding response is not applicable here. Y-randomization abolished the predictive ability, which implied that the model was not generated by chance (**Fig. 3.12B**). Similar trends could also be observed for *RPIT*-QDA in LPOCV (**Fig. 3.12C**).

Then we validated our *RPIT*-RF and *RPIT*-QDA classifiers with an independent external test set, comprising of 11709 known yeast mRNA-protein interaction pairs and 4709 negative pairs generated with data shuffling. Similar to LPOCV results, *PRIT*-RF achieved the best performance with respect to ROC AUC (AUC = 0.71) (**Table 3.5**). In comparison, other method, except AdaBoost (another ensemble classifier), received ROC AUC close to 0.5 (**Fig. 3.13A** and **Table 3.5**), which indicates the low predictive capabilities for QDA, LDA, KNN and Naïve Bayes. Notably, we observed that most classifiers, including *RPIT*-RF, obtained much better performances in sensitivity than specificity. This is probably due to the fact that the negative set we used for external validation is originated from data shuffling, and the probability of have false negative pair can be significant. In addition, removal of any attributes (protein/RNA/interface) or using only sequence identity information as features could dramatically compromise ROC AUCs ($AUC^{\Delta Protein}$ = 0.45, $AUC^{\Delta RNA}$ = 0.51, $AUC^{\Delta Interface}$ = 0.59), as shown in **Fig. 3.13B**. Consistent to the observation from LPOCV, random forest classifier fed with protein and RNA sequence identity information predicted RPI no better than Y-randomization in this validation ($AUC^{SeqIden}$ = 0.51) (**Fig. 3.13B**).

To further validate our model in a more unbiased manner, we evaluated *RPIT*-RF on 42 most recent discoveries of miRNA-protein interactions (**Table 3.2**). Here the ROC AUC metric may not achieve enough statistical power due to the small size of this dataset. We herein only reported the overall accuracy. The predictive accuracy of *RPIT*-RF is 71.5%, which is superior to that of *RPISeq*-RF (accuracy = 56.1%) and *RPISeq*-SVM (accuracy = 63.4%). Although more aggressive validation is indeed needed, this validation set showed a proof-of-

principle that *RPIT*-RF can be used for prioritizing novel miRNA-protein interaction by virtual screening.

Table 3.5 Performance of different classifiers in protein-RNA interface threading. The training set was validated by leave-protein-out cross-validation, and we validated classifiers based on an external test set.

		RF	QDA	AdaBoost	LDA	kNN	NaïveBayes
Training Set	Sensitivity	0.89	0.89	0.77	0.68	0.68	0.91
	Specificity	0.84	0.67	0.86	0.86	0.86	0.64
	PPV	0.72	0.56	0.72	0.70	0.69	0.54
	NPV	0.94	0.93	0.89	0.85	0.85	0.94
	F1	0.79	0.69	0.74	0.69	0.68	0.67
	MCC	0.69	0.52	0.62	0.55	0.54	0.51
	Accuracy	0.86	0.74	0.83	0.81	0.80	0.72
	ROC AUC	0.93	0.86	0.89	0.85	0.83	0.83
Test Set	Sensitivity	0.79	0.82	0.75	0.33	0.41	0.84
	Specificity	0.49	0.18	0.35	0.61	0.53	0.14
	PPV	0.79	0.71	0.74	0.68	0.68	0.71
	NPV	0.48	0.29	0.36	0.27	0.27	0.26
	F1	0.79	0.76	0.75	0.45	0.51	0.77
	MCC	0.28	0.00	0.10	-0.05	-0.05	-0.03
	Accuracy	0.70	0.63	0.64	0.41	0.44	0.64
	ROC AUC	0.71	0.48	0.61	0.47	0.46	0.49

Fig. 3.12. **ROCs in LPOCV.** (A) Comparison of different classifiers. The grey dashed lines indicates the random prediction. (B) Comparison of random forest classifiers with/without critical interface threading attributes (protein/RNA/interface), using only sequence identities, and Y-randomization. (C) Comparison of QDA classifiers with/without critical interface threading attributes (protein/RNA/interface), using only sequence identities, and from Y-randomization.

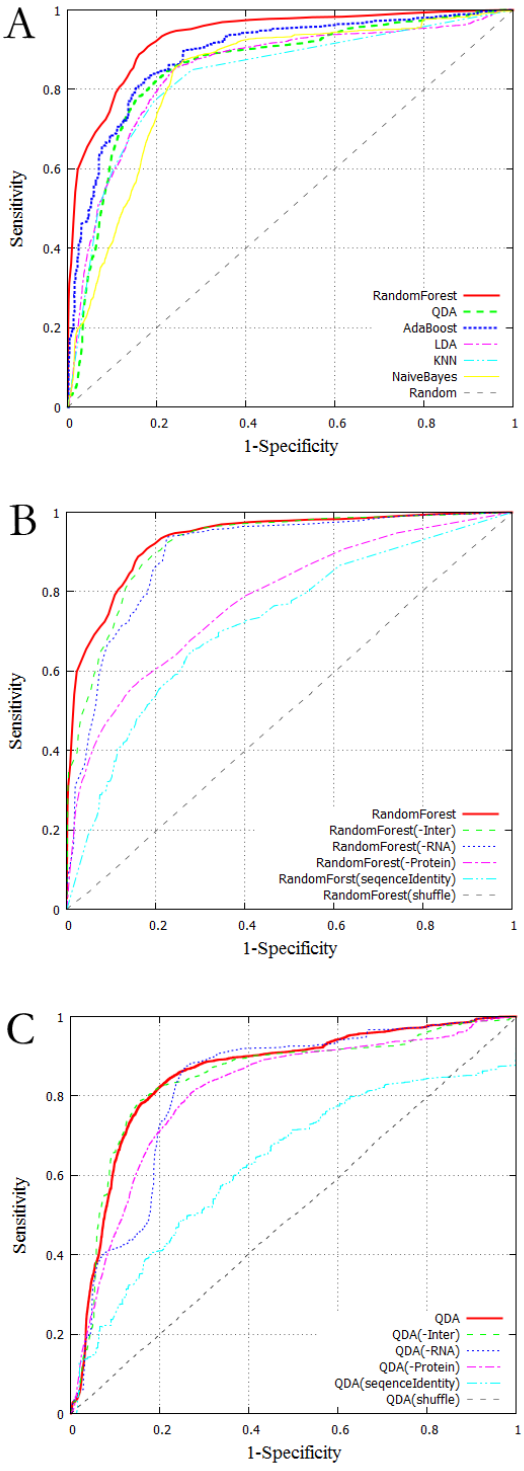
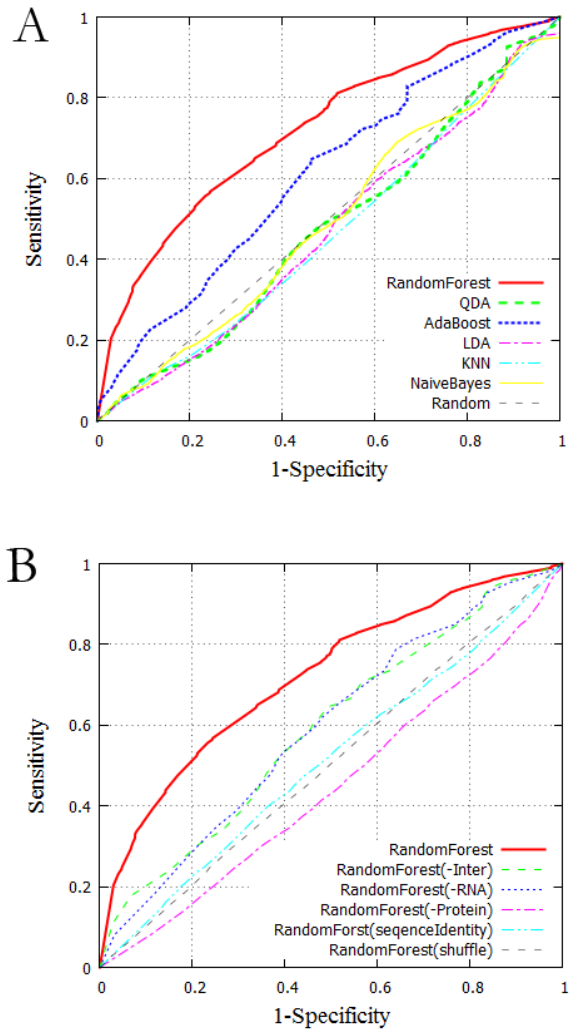


Fig. 3.13. **ROCs in external validation.** (A) Comparison of different classifiers. The grey dashed lines indicates the random prediction. (B) Comparison of random forest classifiers with/without critical interface threading attributes (protein/RNA/interface), using only sequence identities, and from Y-randomization.



3.5 Discussion

Computational modeling of RPI primarily concentrated on the identification of interface residue (or nucleotide) that is likely to bind nucleotide (or protein residue) (see [99] for complete review). In contrast, the "interaction pair prediction problem" is largely overlooked. Despite the recent advancement of experimental high-throughput screening technology (e.g., PAR-CLIP [100], RIP-Chip [101], RNAcompete [102], HITS-CLIP [103]) which shed new light on RPI network, computational method that predict RPI network is still in the "budding stage". As we have discussed in the introduction, either sequence-based or structure-based method has its respective advantages and limitations. Taking the challenge to balance the model robustness (drawback of structure-based method) and noise tolerance capacity (drawback of sequence-based method), in this chapter we implemented an interface threading pipeline, called *RPIT*, for *in silico* prediction of RNA-protein interactions (RPI) using a reference RNA-protein interface as template and in-house developed statistical scoring functions. Compared with template-free, sequence-based method, interface threading restrains the alignment and scoring to only those residues which are most likely to be involved in the RPI. Compared with structure-based method, *RPIT* is independent of the 3D structure information and is more robust when the sequence homology is so low that hampers the prediction of tertiary structure. On the whole, our results showed encouraging accuracy (70%-80%), which is comparable to that of *RPISeq* (78% for *RPISeq*-RF and 65% for *RPISeq*-SVM) and Pancaldi and Bähler et al.'s, accuracy = 70%. Furthermore, *RPIT*-RF is more robust due to a significantly reduction of features (9-feature vector) compared to *RPISeq* which used 599 conjoint triad features, and Pancaldi and Bähler et al.'s which utilized 100 different features

(including mRNA half-life, predicted secondary structure, GO annotation, amino acid composition, codon bias, etc.) which is often unavailable in many cases. Third, two independent validations have suggested that *RPIT*-RF is valuable for predicting and analyzing regulatory RPI networks.

Chapter 4: Summary and future directions

4.1 Summary of three-step virtual screening and its application

In Chapter 2.3, we have benchmarked and compared the possibility of translating 5 docking software and 11 scoring functions to RNA-ligand docking and virtual screening using the largest-ever RNA-ligand complex structure dataset and RNA-ligand binding affinity dataset. Comprehensive statistical analyses have been applied to assess the performance in various aspects: pose reproduction, pose ranking, score-RMSD correlation, and virtual screening enrichment. From this benchmark, we have successfully identified the best combinations for RNA virtual screening: rDock:rDock_solv – ASP rescoring – iMDLScore2 second rescoring for flat, open and flexible binding sites of RNAs, while GOLD:GOLD Fitness – ASP rescoring – rDock_solv second rescoring could be more appropriate for solvent inaccessible and rigid RNA targets, as demonstrated by **Fig. 2.9**.

GA:UU tandem mismatch is a conserved RNA motif frequently found in bacteria rRNA. Using the three-step docking/scoring scheme for structure-based drug design that we have developed in Chapter 2.3, we have successfully identified compound **423** that demonstrates specific binding to RNAs at GA:UU internal loop. Both 1D and 2D NMR spectra proved that compound **423** interacts with G-A sheared base pair meanwhile disrupts the hydrogen bonding between U-U. As expected, the base pairs flexibility, especially from the GA side, contributes to the binding specificity. Ultimately, SAR analysis shows that any *R*-group substitution will abolish its binding to GA:UU motif.

4.2 Summary of RPIT implementation

In Chapter 3, we have implemented an interface threading scheme, called *RPIT*, for accurate prediction of RNA-protein interaction partner using sequences as input. Interface threading circumvents the pitfalls of pure sequence-based or structure-based methods, but identifies and references a known RNA-protein interface as template to make inferences on the region where the interaction occurs, and predict the interacting propensity based on the interface profiles. Briefly, we generated the template database and five statistical scoring functions from our unique collection of 5,471 non-redundant protein-RNA pairs (*nrPR*) from PDB database. The statistical scoring functions evaluate the protein-binding propensity, RNA-binding propensity and RNA-protein binding complementarity as a function of residue type or distance to interface. The interface threading algorithm takes into consideration the residue types, secondary structure state, distance to interface residue, interaction types, and statistical correction for nonspecific interaction while performing alignment. Upon evaluation, *RPIT* random forest classifier (*RPIT*-RF) achieved the best performance in leave-protein-out cross-validation ($AUC^{RPIT\text{-}RF}$ = 0.93, MCC = 0.69) and independent external validation using RPI from yeast ($AUC^{RPIT\text{-}RF}$ = 0.71, MCC = 0.28). These predictions were significantly better than that baseline model generated with Y-randomization or sequence identity attributes. The attributes of the classifier (protein, RNA or interface profile), moreover, showed reasonable contributions and removal any of them significantly impair the predictive ability. Compared with *RPISeq* method, *RPIT*-RF achieved comparable accuracy in yeast validation set (~70% accuracy) and superior accuracy in miRNA-protein interaction validation set (71% accuracy).

4.3 Future directions in modeling RNA-small molecule interactions

Accurate scoring remains a great challenge in vHTS, even for protein target. When optimizing RNA-ligand scoring function, we observed that distal mutation may significantly affect the binding affinity in RNA system. For instance, the mutations on three distal base pairs on SAM-I riboswitch (PDB ID: 3GX7) decrease the binding affinity by 300 fold but cause minimal changes in the binding modes of SAM (RMSD < 0.5Å). Current scoring functions are incapable of estimating free energy due to the thermodynamic changes of RNA structure. Therefore, future works are still needed to derive RNA-specific atom typing [24], intermolecular potential [104, 105] and nucleotide rotamer library for flexible docking. Docking small molecule to flexible RNA, in particular, is considered more challenging due to the lack of rotamer libraries. Current docking methods model the flexible RNA by soft potentials [106], structural ensembles [6], or doing post-docking local optimization [27]. The generation of RNA ensemble has led to a great success in developing specific inhibitor target HIV-1 TAR RNA; however, the performance of RNA ensemble docking varies on targets, scoring functions and other factors, and virtual screening performance may not be improved when flexibility is introduced [107]. Thus, exploration of flexible RNA docking and scoring will be one of the future directions of our research, which may be realized by incorporating NMR RDC restraints into the scoring function.

In this thesis, three-step virtual screening pipeline has been successfully applied to a disease-related RNA motif, GA:UU tandem mismatch. To this end, we have identified a small-molecule **423** that specifically recognizes GA:UU motif, validated by 1D and 2D NMR

spectra. According to NMR structures we determined, the unbound-GA:UU motif flanked by GC and AU base pairs shows sheared GA noncanonical base pair and UU is paired thru two internucleotide H-bonds. In another NMR model (PDB ID: 2JSE [67]), however, the H-bonds between UU are absent if GA:UU motif is flanked by two GC Watson-Crick base pairs. This indicates that the GA:UU motif is intrinsically thermodynamically unstable, and can be easily perturbed by surrounding nucleotides, as well as small molecules. Based on 1D and 2D NMR spectra, compound **423** is able to perturb the UU base pair and but binds primarily to U7 and G8 region. Surprisingly enough, the base pair stability of UU or next to UU side failed to infer any variation of binding, but the base pair stability at GA side inversely correlates the binding (tandem AU base pair > AU+GC > GC+AU). This specificity of RNA context fits the 3D model generated by MD simulation, in which the benzothiazole ring stacks on A21 and the amine group form interaction with the AU base pair adjacent to G8:A21 base pair, not UU base pair. However, this model failed to provide direct evidence to explain the destabilizing of UU base pair. We speculated that the weaker peak from UU base pair is because of the enhancement of the exchange rate of uridine imino hydrogen atom with the solvent since GA base pair is propelled. As the mechanism of **423** being selective to GA:UA:AU context still remains unclear, lead optimization and more SAR studies are currently undergoing and more 2D and 3D structural information are being collected to determine more molecular mechanisms of its specificity. Meanwhile, we are designing more RNA vairants (e.g., GA:CC motif) to further investigate binding motif more thoroughly. If necessary, compound **423** or its derivatives can be designed as a molecular probe to quantify the GA:UU RNA expression

in a cellular system, or to study the thermodynamic stability of a new RNA motif tagged with GA:UU motif.

4.4 Future directions in modeling RNA-protein interactions

As a prototyping implementation, one of the central assumptions is that the interface and the interaction type are generally inheritable from its homologous template. This has inevitably simplified the interface threading problem because (i) RPI interface can be assembled by discontinued fragments such that they might have different order in the target and the template; (2) the confidence of interface threading can be greatly compromised if no threading template could be identified; (3) homologous interfaces may not have the identical interaction profile.

The limitation (1) could be partially mitigated by using the across-family templates, which flattens the scoring function by only considering the convergently evolved interface motifs, as did by iWARP [85]. As *RPIT* by nature is a template-based method, the limitation (2) is so far infeasible to address without a consensus strategy, that is, to combine the template-free, *de novo* scoring scheme or classifiers for consensus prediction. In fact, the conventional definition of "template" (<30% sequence identity) need expansion under this circumstance. *RPIT* has employed the best algorithms to date that greatly overcome the "twilight zone" of sequence identities in template-based homology modeling (30%). Based on our evaluation, the sequence identity demonstrated minimal contribution to RPI prediction compared with random guessing (**Fig. 3.13** and **Fig. 3.14**), which further indicates that sequence homology

may be less informative than fold homology. In fact, if we could segment the entire RPI interface into modular motifs or networks (as we proposed when discussing limitation (1)), the common issue of "lack of homologous template" in template-based modeling may be significantly mitigated. Finally, we may address the 3rd limitation by across-family interaction network analysis, which aims to derive a feasible probabilistic model to make inferences on preferable interaction types and interactive residues in that sub-chemical environment.

The second pitfall of *RPIT* is that, as we discussed in the Chapter 4.3, current implementation is unable to account for the thermodynamic changes by distal mutations. Actually, the underlying assumption of interface threading is that distal residue is generally less informative than the residue that is closer to the interface when one predicts the binding. Making such assumption simplifies the model, however, sacrificed the situation that the mutations that destabilize the integrity of macromolecule (especially protein) may significantly affect the binding affinity. More specifically, in the current *RPIT* implementation, solvent accessible surface area (SASA) is not yet considered due to lack of biophysical model to estimate the SASA of target interface, and when gap is presented. Based on the previous study of RNA-protein interface, SASA is indeed a unique characteristic in RNA-protein interface compared with that for protein-protein interface or protein-DNA interface [51]. More theoretical models which confers the template SASA to RNA-protein interface whose SASA information is absent is aggressively needed for further optimization. Another aspect of future work is to account for the thermodynamic contributions, which may involve the incorporation of pre-calculated residue flexibility profile or NMR restraints.

As a subsequent validation, we would like to examine the effectiveness of *RPIT*-RF for discovery of novel miRNA-protein interaction. We are interested in discovering novel RPI from human kinome and miRNAome. The hypothesis underlying this discovery-oriented study is based on the finding of an RNA aptamer in complex with *Bos taurus* G protein-coupled receptor kinase 2 (GRK2) [108], and we speculated that the precursor miRNAs is likely to function as endogenous inhibitor of protein kinase. To benchmark the speed of *RPIT*, we have prioritized several promising miRNA-kinase interaction from randomly generated 825,600 pairs within 48 hours, and the follow-up experimental validation is ongoing. We are hoping to identify paradigm-shift function of miRNA for better understanding of disease related miRNA regulatory network in the near future.

Appendix

1. Docking parameters

GOLD 5.0.1 (CCDC): For docking and virtual screening, default parameters were set for GOLD Fitness, ChemScore and ASP scores. "Allow early termination" and soft potentials were turned off, and 200% search efficiency was employed to allow maximal exploration of ligand conformation. We used 20 genetic algorithm (GA) runs with internal energy offset. For pose reproduction analysis, the radius of the binding pocket was set as the maximal atomic distance from the geometrical center of the ligand plus 3Å. The top 10 ranked docking poses were retained for the 3D cumulative success rate analysis, cross-docking, and virtual screening studies. To perform the native pose ranking and RMSD-score correlation study, we found that the GOLD:GOLD Fitness combination with 100 population and 1000 maxops could help us obtain high diversity and quality of the conformational decoys. Therefore, GOLD:GOLD Fitness was employed to generate 100 conformational decoys for each target. Rescoring was conducted with the GOLD rescore option, in which poses would be optimized by the program.

Glide 5.6 (Schrödinger): Default parameters were employed for both Glide standard precision (SP) and extra precision (XP) docking. Both GlideScore and Emodel score were evaluated.

Multiple starting conformations were prepared with LigPrep2.0. The binding site was defined as a box centered on the geometrical center of the bound ligand with each length equivalent to the maximal atomic distance from the center of the ligand plus 3Å. Flexible hydroxyl groups involved in the ligand binding were selected. The ligand internal energy offset option was turned on. The top 10 ranked poses were minimized and retained. Rescoring was performed by choosing "Refine (do not dock)" option. The decoys with no valid poses after minimization were excluded in RMSD-score correlation analysis, but included in other evaluations as bad poses (GlideScore or Emodel=10000).

Surflex 2.415 (Tripos): The binding pockets were defined by the area around the experimentally determined ligand structure. The protomol_bloat=5 was set for pocket identification. We used 4 additional starting poses and explored the best spin density parameter using 3, 5 and 10. Self_scoring option was turned on. We kept 10 final poses for analysis, and rescoring was performed by "-opt" flag.

rDock 2006.2: Radius of binding pocket was maximal atomic distance from the geometrical center of the ligand plus 3Å, and site searching scoring function was RbtCavityGridSF. Default parameters from “dock.prm” (standard scoring function) and “dock_solv.prm” (scoring function with solvation term) were used for scoring. We performed 200 separate runs for each docking exercise in order to cover enough conformational space. Top 10 ranked poses were retained. Rescoring was performed using the parameter in “minimise.prm” and “minimize_solv.prm” for rescoring with and without the solvation term, respectively.

AutoDock 4.1: The definition of grid box was the same as that of Glide with 0.2Å grid spacing. Lamarckian Genetic Algorithm (LGA) was used to perform 100 GA runs. Other parameters, such as 200 individuals in populations, 500,000 maximum energy evaluations, and 30,000 maximum generations were employed for LGA. The top 10 clusters were retained for analysis. Rescoring was performed using AutoDockTools4 using optimized parameters.

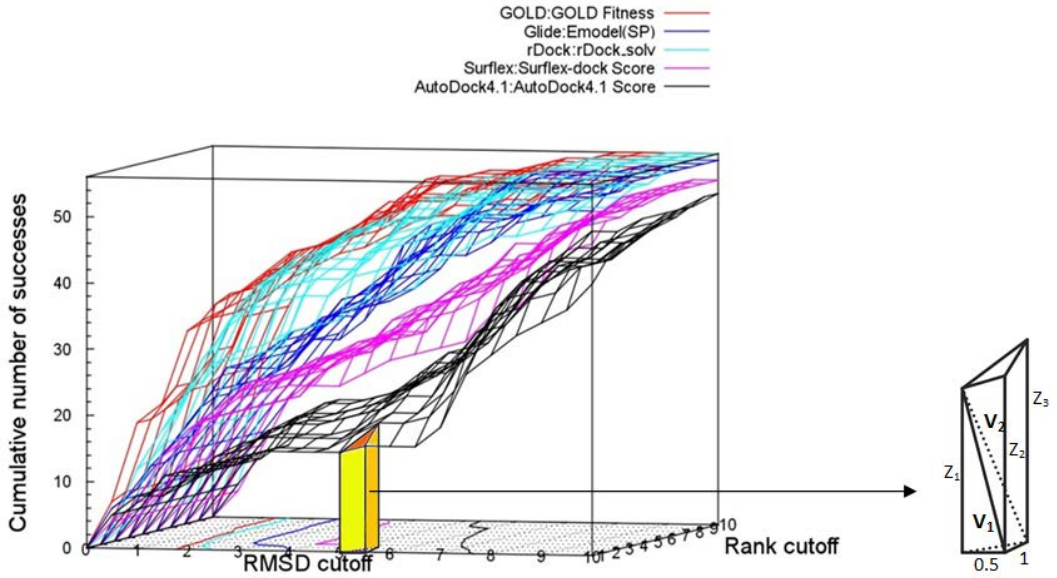
2. Volume under the surface (VUS) calculation

VUS were estimated as the sum of the volume of all triangular prism units under the surface, therefore

$$VUS = \sum (V_{triangular\_prism})$$

The volume of each triangular prism unit ($V_{triangular\ prism}$) was calculated by the following equation. Each triangular prism unit was broken down into a tetrahedron (V_1) and a tetragonal pyramid (V_2), as illustrated below. Z_1, Z_2 and Z_3 were the Z coordinates of triangle vertices on the 3D cumulative success rate surface, and we assume $Z_1 \leq Z_2 \leq Z_3$. Thus,

$$\begin{aligned} V_{triangular\_prism} &= V_1 + V_2 \\ &= (\frac{1}{6} \times 1 \times 0.5 \times Z_1) + (\frac{1}{3} \times \frac{Z_2 + Z_3}{2} \times 1 \times 0.5) \\ &= \frac{1}{12}(Z_1 + Z_2 + Z_3) \end{aligned}$$


GOLD:GOLD Fitness
Glide:Emodel(SP)
rDock:rDock_solv
Surflex:Surflex-dock Score
AutoDock4.1:AutoDock4.1 Score
Cumulative number of successes
RMSD cutoff
Rank cutoff
Z3
Z2
Z1
V2
V1
0.5
1

Bibliography

[1] Q. Chen, R.H. Shafer, I.D. Kuntz, Structure-based discovery of ligands targeted to the RNA double helix, Biochemistry, 36 (1997) 11402-11407.

[2] N. Foloppe, I.J. Chen, B. Davis, A. Hold, D. Morley, R. Howes, A structure-based strategy to identify new molecular scaffolds targeting the bacterial ribosomal A-site, Bioorg Med Chem, 12 (2004) 935-947.

[3] Y. Zhou, V.E. Gregor, B.K. Ayida, G.C. Winters, Z. Sun, D. Murphy, G. Haley, D. Bailey, J.M. Froelich, S. Fish, S.E. Webber, T. Hermann, D. Wall, Synthesis and SAR of 3,5-diamino-piperidine derivatives: novel antibacterial translation inhibitors as aminoglycoside mimetics, Bioorg Med Chem Lett, 17 (2007) 1206-1210.

[4] Z. Du, K.E. Lind, T.L. James, Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening, Chem Biol, 9 (2002) 707-712.

[5] A.V. Filikov, V. Mohan, T.A. Vickers, R.H. Griffey, P.D. Cook, R.A. Abagyan, T.L. James, Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR, J Comput Aided Mol Des, 14 (2000) 593-610.

[6] A.C. Stelzer, A.T. Frank, J.D. Kratz, M.D. Swanson, M.J. Gonzalez-Hernandez, J. Lee, I. Andricioaei, D.M. Markovitz, H.M. Al-Hashimi, Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble, Nat Chem Biol, 7 (2011) 553-559.

[7] K.F. Blount, J.X. Wang, J. Lim, N. Sudarsan, R.R. Breaker, Antibacterial lysine analogs that target lysine riboswitches, Nat Chem Biol, 3 (2007) 44-49.

[8] P. Daldrop, F.E. Reyes, D.A. Robinson, C.M. Hammond, D.M. Lilley, R.T. Batey, R. Brenk, Novel ligands for a purine riboswitch discovered by RNA-ligand docking, Chem Biol, 18 (2011) 324-335.

[9] J. Mulhbacher, E. Brouillette, M. Allard, L.C. Fortier, F. Malouin, D.A. Lafontaine, Novel riboswitch ligand analogs as selective inhibitors of guanine-related metabolic pathways, PLoS Pathog, 6 (2010) e1000865.

[10] A. Pushechnikov, M.M. Lee, J.L. Childs-Disney, K. Sobczak, J.M. French, C.A. Thornton, M.D. Disney, Rational design of ligands targeting triplet repeating transcripts that cause RNA dominant disease: application to myotonic muscular dystrophy type 1 and spinocerebellar ataxia type 3, J Am Chem Soc, 131 (2009) 9767-9779.

[11] J.F. Arambula, S.R. Ramisetty, A.M. Baranger, S.C. Zimmerman, A simple ligand that selectively targets CUG trinucleotide repeats and inhibits MBNL protein binding, Proc Natl Acad Sci U S A, 106 (2009) 16068-16073.

[12] C.H. Wong, Y. Fu, S.R. Ramisetty, A.M. Baranger, S.C. Zimmerman, Selective inhibition of MBNL1-CCUG interaction by small molecules toward potential therapeutic agents for myotonic dystrophy type 2 (DM2), Nucleic Acids Res, 39 (2011) 8881-8890.

[13] D.D. Young, C.M. Connelly, C. Grohmann, A. Deiters, Small molecule modifiers of microRNA miR-122 function for the treatment of hepatitis C virus infection and hepatocellular carcinoma, J Am Chem Soc, 132 (2010) 7976-7981.

[14] K. Gumireddy, D.D. Young, X. Xiong, J.B. Hogenesch, Q. Huang, A. Deiters, Small-molecule inhibitors of microrna miR-21 function, Angew Chem Int Ed Engl, 47 (2008) 7482-7484.

[15] J. Parsons, M.P. Castaldi, S. Dutta, S.M. Dibrov, D.L. Wyles, T. Hermann, Conformational inhibition of the hepatitis C virus internal ribosome entry site RNA, Nat Chem Biol, 5 (2009) 823-825.

[16] P.P. Seth, A. Miyaji, E.A. Jefferson, K.A. Sannes-Lowery, S.A. Osgood, S.S. Propp, R. Ranken, C. Massire, R. Sampath, D.J. Ecker, E.E. Swayze, R.H. Griffey, SAR by MS: discovery of a new class of RNA-binding small molecules for the hepatitis C virus: internal ribosome entry site IIA subdomain, J Med Chem, 48 (2005) 7099-7102.

[17] A.C. Good, S.R. Krystek, J.S. Mason, High-throughput and virtual screening: core lead discovery technologies move towards integration, Drug discovery today, 5 (2000) 61-69.

[18] L. Chen, J.K. Morrow, H.T. Tran, S.S. Phatak, L. Du-Cuny, S. Zhang, From laptop to benchtop to bedside: structure-based drug design on protein targets, Current pharmaceutical design, 18 (2012) 1217-1239.

[19] G.M. Morris, M. Lim-Wilby, Molecular docking, Methods in molecular biology, 443 (2008) 365-382.

[20] Y. Li, J. Shen, X. Sun, W. Li, G. Liu, Y. Tang, Accuracy assessment of protein-based docking programs against RNA targets, J Chem Inf Model, 50 (2010) 1134-1146.

[21] C. Detering, G. Varani, Validation of automated docking programs for docking and database screening against RNA drug targets, Journal of medicinal chemistry, 47 (2004) 4188-4201.

[22] N. Moitessier, E. Westhof, S. Hanessian, Docking of aminoglycosides to hydrated and flexible RNA, Journal of medicinal chemistry, 49 (2006) 1023-1033.

[23] P.T. Lang, S.R. Brozell, S. Mukherjee, E.F. Pettersen, E.C. Meng, V. Thomas, R.C. Rizzo, D.A. Case, T.L. James, I.D. Kuntz, DOCK 6: combining techniques to model RNA-small molecule complexes, Rna, 15 (2009) 1219-1230.

[24] S.D. Morley, M. Afshar, Validation of an empirical RNA-ligand scoring function for fast flexible docking using Ribodock, Journal of computer-aided molecular design, 18 (2004) 189-208.

[25] I.G. Pinto, C. Guilbert, N.B. Ulyanov, J. Stearns, T.L. James, Discovery of ligands for a novel target, the human telomerase RNA, based on flexible-target virtual screening and NMR, J Med Chem, 51 (2008) 7205-7215.

[26] P. Pfeffer, H. Gohlke, DrugScoreRNA--knowledge-based scoring function to predict RNA-ligand interactions, J Chem Inf Model, 47 (2007) 1868-1876.

[27] C. Guilbert, T.L. James, Docking to RNA via root-mean-square-deviation-driven energy minimization with flexible ligands and flexible targets, J Chem Inf Model, 48 (2008) 1257-1268.

[28] K.E. Lind, Z. Du, K. Fujinaga, B.M. Peterlin, T.L. James, Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA, Chem Biol, 9 (2002) 185-193.

[29] N. Foloppe, N. Matassova, F. Aboul-Ela, Towards the discovery of drug-like RNA ligands?, Drug discovery today, 11 (2006) 1019-1027.

[30] T. Glisovic, J.L. Bachorik, J. Yong, G. Dreyfuss, RNA-binding proteins and post-transcriptional gene regulation, FEBS Lett, 582 (2008) 1977-1986.

[31] J.T. Kung, D. Colognori, J.T. Lee, Long noncoding RNAs: past, present, and future, Genetics, 193 (2013) 651-669.

[32] J.E. Wilusz, H. Sunwoo, D.L. Spector, Long noncoding RNAs: functional surprises from the RNA world, Genes & development, 23 (2009) 1494-1504.

[33] A.M. Eiring, J.G. Harb, P. Neviani, C. Garton, J.J. Oaks, R. Spizzo, S. Liu, S. Schwind, R. Santhanam, C.J. Hickey, H. Becker, J.C. Chandler, R. Andino, J. Cortes, P. Hokland, C.S. Huettner, R. Bhatia, D.C. Roy, S.A. Liebhaber, M.A. Caligiuri, G. Marcucci, R. Garzon, C.M. Croce, G.A. Calin, D. Perrotti, miR-328 functions as an RNA decoy to modulate hnRNP E2 regulation of mRNA translation in leukemic blasts, Cell, 140 (2010) 652-665.

[34] M.Y. Balkhi, O.H. Iwenofu, N. Bakkar, K.J. Ladner, D.S. Chandler, P.J. Houghton, C.A. London, W. Kraybill, D. Perrotti, C.M. Croce, C. Keller, D.C. Guttridge, miR-29 acts as a decoy in sarcomas to protect the tumor suppressor A20 mRNA from degradation by HuR, Science signaling, 6 (2013) ra63.

[35] X. Chen, H. Liang, J. Zhang, K. Zen, C.Y. Zhang, microRNAs are ligands of Toll-like receptors, Rna, 19 (2013) 737-739.

[36] M. Fabbri, A. Paone, F. Calore, R. Galli, C.M. Croce, A new role for microRNAs, as ligands of Toll-like receptors, RNA biology, 10 (2013) 169-174.

[37] S. Guil, J.F. Caceres, The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a, Nat Struct Mol Biol, 14 (2007) 591-596.

[38] G. Michlewski, J.F. Caceres, Antagonistic role of hnRNP A1 and KSRP in the regulation of let-7a biogenesis, Nat Struct Mol Biol, 17 (2010) 1011-1018.

[39] G. Michlewski, S. Guil, C.A. Semple, J.F. Caceres, Posttranscriptional regulation of miRNAs harboring conserved terminal loops, Mol Cell, 32 (2008) 383-393.

[40] H. Towbin, P. Wenter, B. Guennewig, J. Imig, J.A. Zagalak, A.P. Gerber, J. Hall, Systematic screens of proteins binding to synthetic microRNA precursors, Nucleic Acids Res, 41 (2013) e47.

[41] M. Trabucchi, P. Briata, M. Garcia-Mayoral, A.D. Haase, W. Filipowicz, A. Ramos, R. Gherzi, M.G. Rosenfeld, The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs, Nature, 459 (2009) 1010-1014.

[42] T. Ruggiero, M. Trabucchi, F. De Santa, S. Zupo, B.D. Harfe, M.T. McManus, M.G. Rosenfeld, P. Briata, R. Gherzi, LPS induces KH-type splicing regulatory protein-dependent processing of microRNA-155 precursors in macrophages, FASEB journal : official publication of the Federation of American Societies for Experimental Biology, 23 (2009) 2898-2908.

[43] Y. Nam, C. Chen, R.I. Gregory, J.J. Chou, P. Sliz, Molecular basis for interaction of let-7 microRNAs with Lin28, Cell, 147 (2011) 1080-1091.

[44] A.G. Seto, R.E. Kingston, N.C. Lau, The coming of age for Piwi proteins, Molecular Cell, 26 (2007) 603-609.

[45] M.C. Siomi, K. Sato, D. Pezic, A.A. Aravin, PIWI-interacting small RNAs: the vanguard of genome defence, Nat Rev Mol Cell Bio, 12 (2011) 246-258.

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

[47] M. Treger, E. Westhof, Statistical analysis of atomic contacts at RNA–protein interfaces, Journal of Molecular Recognition, 14 (2001) 199-214.

[48] A. Gupta, M. Gribskov, The role of RNA sequence and structure in RNA--protein interactions, Journal of molecular biology, 409 (2011) 574-587.

[49] N. Morozova, J. Allers, J. Myers, Y. Shamoo, Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures, Bioinformatics, 22 (2006) 2746-2752.

[50] J.J. Ellis, M. Broom, S. Jones, Protein-RNA interactions: structural analysis and functional classes, Proteins, 66 (2007) 903-911.

[51] A. Barik, N. C, S.P. Pilla, R.P. Bahadur, Molecular architecture of protein-RNA recognition sites, Journal of biomolecular structure & dynamics, (2015) 1-14.

[52] Y. Huang, S. Liu, D. Guo, L. Li, Y. Xiao, A novel protocol for three-dimensional structure prediction of RNA-protein complexes, Scientific reports, 3 (2013) 1887.

[53] U. Nagaswamy, N. Voss, Z. Zhang, G.E. Fox, Database of non-canonical base pairs found in known RNA structures, Nucleic Acids Res, 28 (2000) 375-376.

[54] U. Muppirala, V. Honavar, D. Dobbs, Predicting RNA-Protein Interactions Using Only Sequence Information, BMC bioinformatics, 12 (2011) 489.

[55] M. Bellucci, F. Agostini, M. Masin, G.G. Tartaglia, Predicting protein associations with long noncoding RNAs, Nature methods, 8 (2011) 444-445.

[56] L. Perez-Cano, A. Solernou, C. Pons, J. Fernandez-Recio, Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical

potentials, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, (2010) 293-301.

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

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

[59] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein-protein interactions based only on sequences information, Proc Natl Acad Sci U S A, 104 (2007) 4337-4341.

[60] M.L. Verdonk, J.C. Cole, M.J. Hartshorn, C.W. Murray, R.D. Taylor, Improved protein–ligand docking using GOLD, Proteins: Structure, Function, and Bioinformatics, 52 (2003) 609-623.

[61] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, J Med Chem, 47 (2004) 1739-1749.

[62] A.N. Jain, Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search, J Comput Aided Mol Des, 21 (2007) 281-306.

[63] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, Journal of Computational Chemistry, 19 (1998) 1639-1662.

[64] R. Huey, G.M. Morris, A.J. Olson, D.S. Goodsell, A semiempirical free energy force field with charge-based desolvation, J Comput Chem, 28 (2007) 1145-1152.

[65] N.M. O'Boyle, J.W. Liebeschuetz, J.C. Cole, Testing assumptions and hypotheses for rescoring success in protein-ligand docking, J Chem Inf Model, 49 (2009) 1871-1878.

[66] L. Chen, G.A. Calin, S. Zhang, Novel insights of structure-based modeling for RNA-targeted drug discovery, J Chem Inf Model, 52 (2012) 2741-2753.

[67] N. Shankar, T. Xia, S.D. Kennedy, T.R. Krugh, D.H. Mathews, D.H. Turner, NMR reveals the absence of hydrogen bonding in adjacent UU and AG mismatches in an isolated internal loop from ribosomal RNA, Biochemistry, 46 (2007) 12665-12678.

[68] M. Davlieva, J. Donarski, J. Wang, Y. Shamoo, E.P. Nikonowicz, Structure analysis of free and bound states of an RNA aptamer against ribosomal protein S8 from Bacillus anthracis, Nucleic Acids Res, 42 (2014) 10795-10808.

[69] Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu, R. Wang, PDB-wide collection of binding data: current status of the PDBbind database, Bioinformatics, 31 (2015) 405-412.

[70] S. Yoshizawa, D. Fourmy, J.D. Puglisi, Structural origins of gentamicin antibiotic action, EMBO J, 17 (1998) 6437-6448.

[71] F. Barbault, L. Zhang, L. Zhang, B.T. Fan, Parametrization of a specific free energy function for automated docking against RNA targets using neural networks, Chemometrics and Intelligent Laboratory Systems, 82 (2006) 269-275.

[72] J.C. Shelley, A. Cholleti, L.L. Frye, J.R. Greenwood, M.R. Timlin, M. Uchimaya, Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules, J Comput Aided Mol Des, 21 (2007) 681-691.

[73] Q. Vicens, E. Westhof, Crystal structure of paromomycin docked into the eubacterial ribosomal decoding A site, Structure, 9 (2001) 647-658.

[74] A. Serganov, L. Huang, D.J. Patel, Structural insights into amino acid binding and gene control by a lysine riboswitch, Nature, 455 (2008) 1263-1267.

[75] M.L. Verdonk, V. Berdini, M.J. Hartshorn, W.T. Mooij, C.W. Murray, R.D. Taylor, P. Watson, Virtual screening using protein-ligand docking: avoiding artificial enrichment, J Chem Inf Comput Sci, 44 (2004) 793-806.

[76] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J. Berendsen, GROMACS: fast, flexible, and free, Journal of computational chemistry, 26 (2005) 1701-1718.

[77] L. Chen, L. Du-Cuny, S. Moses, S. Dumas, Z. Song, A.H. Rezaeian, H.K. Lin, E.J. Meuillet, S. Zhang, Novel inhibitors induce large conformational changes of GAB1 pleckstrin homology domain and kill breast cancer cells, PLoS computational biology, 11 (2015) e1004021.

[78] A. Perez, I. Marchan, D. Svozil, J. Sponer, T.E. Cheatham, 3rd, C.A. Laughton, M. Orozco, Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers, Biophysical journal, 92 (2007) 3817-3829.

[79] D. Svozil, J.E. Sponer, I. Marchan, A. Perez, T.E. Cheatham, 3rd, F. Forti, F.J. Luque, M. Orozco, J. Sponer, Geometrical and electronic structure variability of the sugar-phosphate backbone in nucleic acids, The journal of physical chemistry. B, 112 (2008) 8188-8197.

[80] A.W. Sousa da Silva, W.F. Vranken, ACPYPE - AnteChamber PYthon Parser interfacE, BMC research notes, 5 (2012) 367.

[81] A.T. Chang, E.P. Nikonowicz, Solution NMR determination of hydrogen bonding and base pairing between the glyQS T box riboswitch Specifier domain and the anticodon loop of tRNA(Gly), FEBS Lett, 587 (2013) 3495-3499.

[82] L.W. Yang, A.J. Rader, X. Liu, C.J. Jursa, S.C. Chen, H.A. Karimi, I. Bahar, oGNM: online computation of structural dynamics using the Gaussian Network Model, Nucleic acids research, 34 (2006) W24-31.

[83] S.E. Dobbins, V.I. Lesk, M.J. Sternberg, Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking, Proceedings of the National Academy of Sciences of the United States of America, 105 (2008) 10390-10395.

[84] G.L. Warren, C.W. Andrews, A.M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus, S. Senger, G. Tedesco, I.D. Wall, J.M. Woolven, C.E. Peishoff, M.S. Head, A critical assessment of docking programs and scoring functions, J Med Chem, 49 (2006) 5912-5931.

[85] R. Hosur, J. Xu, J. Bienkowska, B. Berger, iWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions, Journal of molecular biology, 405 (2011) 1295-1310.

[86] D. Frishman, P. Argos, Knowledge-based protein secondary structure assignment, Proteins, 23 (1995) 566-579.

[87] X.J. Lu, W.K. Olson, H.J. Bussemaker, The RNA backbone plays a crucial role in mediating the intrinsic stability of the GpU dinucleotide platform and the GpUpA/GpA miniduplex, Nucleic Acids Res, 38 (2010) 4868-4876.

[88] L.J. McGuffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server, Bioinformatics, 16 (2000) 404-405.

[89] S. Will, T. Joshi, I.L. Hofacker, P.F. Stadler, R. Backofen, LocARNA-P: accurate boundary prediction and improved detection of structural RNAs, Rna, 18 (2012) 900-914.

[90] C. Smith, S. Heyne, A.S. Richter, S. Will, R. Backofen, Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA, Nucleic Acids Res, 38 (2010) W373-377.

[91] K. Sato, M. Hamada, K. Asai, T. Mituyama, CENTROIDFOLD: a web server for RNA secondary structure prediction, Nucleic Acids Res, 37 (2009) W277-280.

[92] T. Puton, L.P. Kozlowski, K.M. Rother, J.M. Bujnicki, CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction, Nucleic Acids Res, 42 (2014) 5403-5406.

[93] J. Yuan, W. Wu, C. Xie, G. Zhao, Y. Zhao, R. Chen, NPInter v2.0: an updated database of ncRNA interactions, Nucleic Acids Res, 42 (2014) D104-108.

[94] K.B. Cook, H. Kazan, K. Zuberi, Q. Morris, T.R. Hughes, RBPDB: a database of RNA-binding specificities, Nucleic Acids Res, 39 (2011) D301-308.

[95] E.W. Stawiski, L.M. Gregoret, Y. Mandel-Gutfreund, Annotating nucleic acid-binding function based on protein structure, Journal of molecular biology, 326 (2003) 1065-1079.

[96] K.R. Christie, S. Weng, R. Balakrishnan, M.C. Costanzo, K. Dolinski, S.S. Dwight, S.R. Engel, B. Feierbach, D.G. Fisk, J.E. Hirschman, E.L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C.L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein, J.M. Cherry, Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms, Nucleic Acids Res, 32 (2004) D311-314.

[97] V. Pancaldi, J. Bahler, In silico characterization and prediction of global protein-mRNA interactions in yeast, Nucleic Acids Res, 39 (2011) 5826-5836.

[98] K. Nadassy, S.J. Wodak, J. Janin, Structural features of protein-nucleic acid recognition sites, Biochemistry, 38 (1999) 1999-2017.

[99] D. Cirillo, F. Agostini, G.G. Tartaglia, Predictions of protein–RNA interactions, Wiley Interdisciplinary Reviews: Computational Molecular Science, 3 (2013) 161-175.

[100] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, Jr., A.C. Jungkamp, M. Munschauer, A. Ulrich, G.S. Wardle, S. Dewell, M. Zavolan, T. Tuschl, Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP, Cell, 141 (2010) 129-141.

[101] J.D. Keene, J.M. Komisarow, M.B. Friedersdorf, RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts, Nature protocols, 1 (2006) 302-307.

[102] D. Ray, H. Kazan, E.T. Chan, L. Pena Castillo, S. Chaudhry, S. Talukder, B.J. Blencowe, Q. Morris, T.R. Hughes, Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins, Nature biotechnology, 27 (2009) 667-670.

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

[104] J. Wang, W. Wang, S. Huo, M. Lee, P.A. Kollman, Solvation Model Based on Weighted Solvent Accessible Surface Area, The Journal of Physical Chemistry B, 105 (2001) 5055-5067.

[105] L. Wesson, D. Eisenberg, Atomic solvation parameters applied to molecular dynamics of proteins in solution, Protein Sci, 1 (1992) 227-235.

[106] D.M. Krüger, J. Bergs, S. Kazemi, H. Gohlke, Target Flexibility in RNA−Ligand Docking Modeled by Elastic Potential Grids, ACS Medicinal Chemistry Letters, 2 (2011) 489-493.

[107] O. Korb, T.S. Olsson, S.J. Bowden, R.J. Hall, M.L. Verdonk, J.W. Liebeschuetz, J.C. Cole, Potential and limitations of ensemble docking, J Chem Inf Model, 52 (2012) 1262-1274.

[108] V.M. Tesmer, S. Lennarz, G. Mayer, J.J. Tesmer, Molecular mechanism for inhibition of g protein-coupled receptor kinase 2 by a selective RNA aptamer, Structure, 20 (2012) 1300-1309.

Vita

Lu Chen was born in Shanghai, China on December 20, 1987, the son of Guoyun Zhang and Jianjun Chen. After completing his work at Shanghai Experimental School, Shanghai, China in 2005, he entered Fudan University in Shanghai, China. He received the degree of Bachelor of Sciences with a major in biological science from Fudan University in June, 2009. In September of 2009, he entered The University of Texas Graduate School of Biomedical Sciences at Houston.

Publications

Chen L, Moses S, Du-Cuny L, Dumas S, et al. Novel inhibitors induce large conformational changes of GAB1 pleckstrin homology domain and kill breast cancer cells. *PLoS Comput Biol*. 2015; 11(1): e1004021.

Chen L, Calin GA, Zhang S. Novel insights of structure-based modeling for RNA-targeted drug discovery. J *Chem Inf Model*. 2012; 52(10):2741-53.

Chen L, Morrow JK, Tran HT, Phatak SS, Du-Cuny L, Zhang S. From laptop to benchtop to bedside: Structure-based drug design on protein targets. *Curr Pharm Des*. 2012; 18(9):1217-39.

Du-Cuny L*, **Chen L***, Zhang S. A critical assessment of combined ligand-based and structure-based approaches to hERG channel blocker modeling. *J Chem Inf Model*. 2011; 51(11):2948-60. (*Co-first author)

Permanent address:

No.1, Lane 58, De Ping Road,

Room 803

Shanghai, 200136 China