bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

PremPRI: Predicting the Effects of Single Mutations on Protein-RNA Interactions

Ning Zhang[1], Haoyu Lu[1], Yuting Chen[1], Zefeng Zhu[1], Qing Yang[1], Shuqin Wang[1] and Minghui Li[1,*]

[1]Center for Systems Biology, Department of Bioinformatics, School of Biology and Basic Medical Sciences, Soochow University, Suzhou 215123, China

*corresponding author, minghui.li@suda.edu.cn

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

Protein-RNA interactions are crucial for many cellular processes, such as protein synthesis and regulation of gene expression. Missense mutations that alter protein-RNA interaction may contribute to the pathogenesis of many diseases. Here we introduce a new computational method PremPRI, which predicts the effects of single mutations occurring in RNA binding proteins on the protein-RNA interaction by calculating the binding affinity changes quantitatively. The multiple linear regression scoring function of PremPRI is composed of 11 sequence- and structure-based features, and is parameterized on 248 mutations from 50 protein-RNA complexes. Our model shows a good agreement between calculated and experimental values with Pearson correlation coefficient of 0.72 and the corresponding root-mean-square error of 0.76 kcal mol^{-1}, and outperforms three other available methods. PremPRI can be used for finding functionally important variants, understanding the molecular mechanisms underlying the effects, and designing new protein-RNA interaction inhibitors. PremPRI is freely available at http://lilab.jysw.suda.edu.cn/research/PremPRI/.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

The interactions between protein and RNA are crucial for many cellular processes, such as protein synthesis and regulation of gene expression (1-6). Missense mutations occurring in these RNA binding proteins that alter protein-RNA interactions may cause significant deviation from the normal function of these proteins, potentially leading to various disorders including cancer (7,8). Indeed, several comprehensive studies of the structural nature of mutations in cancer and rare Mendelian diseases have shown that mutations located on binding interfaces may induce macromolecular interaction perturbations and play "driver" or "damaging" roles in many cancers and Mendelian diseases (9-16). Quantifying the effects of missense mutations on specific protein-RNA interactions requires assessing the binding affinity changes upon introducing mutations that can be accurately measured by traditional mutagenesis technologies. However, methods such as surface plasmon resonance (17) and isothermal titration calorimetry (18) used to measure binding affinity are costly and time-consuming. Therefore, developing reliable computational approaches provides an alternative way to investigate the effects of mutations on proteins and their complexes on a large scale, which will facilitate the identification of functionally important missense mutations and the discovery of the molecular mechanisms that cause diseases.

Many efforts to computationally predict and model the effects of missense mutations on protein stability and protein-protein interactions have been made that help uncover the relationships between genotype and phenotype (19-25). However, predicting the impacts of mutations on protein–nucleic acid interactions has been more intractable and very few methods have been proposed that can do this (26-30). One of the reasons that has hindered the development of methods is due to the complexity of nucleic acid chemistry and binding that has limited the availability of high-quality data. We have previously used molecular mechanics force fields and statistical potentials to model the impacts of single mutations on protein-protein (22) and protein-DNA interactions (27). Peng et al. combined modified MM/PBSA based energy terms with additional knowledge-based terms to predict the protein-DNA binding affinity changes upon single mutations (28). Pires et al. used graph-based signatures to estimate the effects of single mutations on protein-nucleic acid binding (26). Moreover, several computational approaches have been proposed for predicting binding hot spots at protein-RNA binding interfaces (29-31). Although these advancements have been made, the issue of estimating the effects of mutations on protein-RNA interactions is still at the initial stage.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To address this need, we developed a computational method, PremPRI, for characterizing the effects of single mutations on protein-RNA interactions by calculating the binding affinity change quantitatively. PremPRI uses a novel multiple linear regression scoring function composed of 11 sequence- and structure-based features, and achieves significantly better performance than other predictors. PremPRI can be applied to many tasks, such as finding potential disease-causing and cancer driver missense mutations and understanding their molecular mechanisms, and designing inhibitors of protein-RNA interactions.

Methods

Experimental datasets used for training

The ProNIT database (32) includes experimentally measured values of changes in binding affinity ($\Delta\Delta G$) between proteins and nucleic acids upon single amino acid substitutions along with the experimentally determined 3D structures of protein-nucleic acid complexes. A recently proposed database, dbAMEPNI (33), collected more experimentally measured binding affinity changes between protein and nucleic acid induced by alanine-scanning mutations. In order to construct a accurate training dataset, we removed the complexes with modified residues/nucleotides located at the protein-RNA binding interface, the complexes in which the protein length is less than 20 amino acids or RNA length is less than five nucleotides, and the mutations that occurred at metal coordination sites from the above two data resources. Furthermore, to avoid the inconsistency between RNA sequences used for measuring binding affinities and those presented in the 3D complex structures used for developing prediction model, we compared the sequence similarity between the RNA sequences at the protein-RNA binding sites and the corresponding sequences used to determine binding affinity. The RNA sequences in binding affinity measurements were either obtained from the ProNIT database or manually collected from the corresponding references. The entries with high sequence similarity (80%) were retained. In addition, there are three mutations with multiple experimental measurements of $\Delta\Delta G$ and the difference between the maximal and minimal $\Delta\Delta G$ for each of them is less than 1.0 kcal mol^{-1}, thus the average value was used for these cases. As a result, 108 single mutations from 30 protein-RNA complexes from the ProNIT and dbAMEPNI databases were retained.

Moreover, 140 additional mutations obtained from the PrabHot benchmark and independent datasets were added into our training dataset, and they satisfy all of above criteria

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(29). PrabHot is a computational method for identifying hot spots at the protein-RNA binding interfaces. Therefore, the final experimental dataset used to parameterize the PremPRI model includes 248 single mutations from 50 protein-RNA complexes (named as S248) (Table S1). The number of mutations for each protein-RNA complex is shown in Figure S1. In the S248 dataset, only 16 mutations have experimental pH values. Thus, the neutral pH was chosen at which the default charged states are assigned to the ionizable residues. We also compared our training dataset of S248 with them used for developing mCSM-NA (26) and PrabHot methods (29), and the details are shown in the Table S1.

Structural optimization protocol

Three dimensional structures of protein-RNA complexes were obtained from the Protein Data Bank (PDB) (34), and the biological assembly 1 of crystal structure or the first model of NMR was used as the initial wild-type structure. Then we used the BuildModel module of FoldX software package (20,35) to produce mutant structures. Next we applied the VMD program (36) to add missing heavy side-chain and hydrogen atoms to wild-type and mutant structures using the topology file of the CHARMM36 force field (37). After that we carried out a 1000-step energy minimization in the gas phase for each protein-RNA complex during which the harmonic restraints with the force constant of 5 kcal mol^{-1} $Å^{-2}$ were applied on the backbone atoms of all residues. The NAMD program v2.12 (38) and the CHARMM36 force field (37) were used to perform the energy minimization. A 12 Å cutoff distance for non-bonded interactions was applied to the molecular systems and lengths of hydrogen-containing bonds were constrained by the SHAKE algorithm (39). The flowchart of structure optimization protocol is shown in Figure S2. The minimized structures of wild-type and mutant protein-RNA complexes were used for the following calculations of energy features.

The PremPRI model

The multiple linear regression scoring function of PremPRI is composed of 11 features and all of them have significant contribution to the quality of the model (p-value < 0.01, t-test). The p-value and the importance of each feature are presented in Table S2 and the description is shown below:

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- $\Delta\Delta E_{vdw}$ is the difference of Van der Waals interaction energies between mutant and wild type ($\Delta\Delta E_{vdw} = \Delta E_{vdw}^{mut} - \Delta E_{vdw}^{wt}$). ΔE_{vdw} is the difference of Van der Waals energies between a protein-RNA complex and each binding partner (partner1: protein; partner2: RNA), which is calculated using the ENERGY module of CHARMM program (40).
- $\Delta\Delta E_{vdw.re}$ is the difference of Van der Waals repulsive energies between mutant and wild type. Here, the Van der Waals repulsive energy only calculates the repulsion between the residue at the mutated site and the nucleic acids.
- $\Delta\Delta E_{elec}$ is the difference of electrostatic interaction energies between mutant and wild type ($\Delta\Delta E_{elec} = \Delta E_{elec}^{mut} - \Delta E_{elec}^{wt}$). ΔE_{elec} is the electrostatic interaction energies between the residue at the mutated site and its contact residues/nucleic acids. If any side-chain atom of a residue or nucleic acid is located within 10 Å from any side-chain atom of the mutated site, we defined it as a contact residue or nucleic acid. The calculation is carried out using the ENERGY module of CHARMM program.
- N_{inter} is the number of amino acids at the protein-RNA binding interface. If the solvent accessible surface area of a residue in the protein is more than that in the complex, we define it as the interface residue. The SASA module of CHARMM is used to calculate the solvent accessible surface area.
- $R_{L/SA}$ is the ratio of protein length and its surface area. $R_{L/SA} = \frac{\text{Length}}{\text{SASA}}$, Length is the total number of residues in the protein-RNA complex, and SASA is the solvent accessible surface area of the unbound protein that is extracted from the minimized wild-type complex structure.
- Closeness is the inverse of total shortest-path distance between the node of mutated site and other nodes in the remaining residue interactive network. It is defined as:

$$C(u) = (n-1)/\sum_{v \in V, v \neq u} d(u,v)$$

where $d(u,v)$ is the shortest-path distance between the node u of mutated site and any node v. V is the set of all nodes and n is the number of nodes in the network. The shortest-path distance between two nodes refers to the minimum number of nodes that reach from one node to the other (41). Cα atom of a residue is considered as a node and if the distance

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between two Cα atoms is less than 6 Å, we defined them as having a direct interaction. Closeness is calculated using the Python package NetworkX (42).

- ΔSA is the difference of solvent accessible surface areas between mutant and wild type ($\Delta SA = SA^{mut}_{site} - SA^{wt}_{site}$). SA_{site} is the solvent accessible surface area of the residue at the mutated site in the unbound protein that is extracted from the minimized complex structure.
- $P_{coil} = \frac{N_{e.c}}{N_{All}}$, $N_{e.c}$ and N_{All} are the number of exposed residues in the coil conformation and all residues in the mutated chain, respectively. Secondary structure elements other than α-helices and β-strands are defined as coil, and are assigned by DSSP program (43). If the ratio of the solvent accessible surface area of a residue in the complex and in solvent is more than 0.25 (44), we defined this residue as the exposed residue.
- ΔOMH is the difference of hydrophobicity scale between mutant and wild-type residue type. The hydrophobicity scale (OMH) for each type of amino acid residue was derived by considering the observed frequency of amino acid replacements among thousands of related structures, which was taken from the study of (45) directly.
- ΔP_{FWY} and ΔP_{KR-DE}: $\Delta P_{FWY} = P^{mut}_{FWY} - P^{wt}_{FWY}$, $P_{FWY} = \frac{N_{FWY}}{N_{All}}$, N_{FWY} and N_{All} are the number of aromatic residues (F, W and Y) and all residues in the mutated protein chain respectively; $\Delta P_{KR-DE} = P^{mut}_{KR-DE} - P^{wt}_{KR-DE}$, $P_{KR-DE} = \frac{N_{KR}-N_{DE}}{N_{All}}$, N_{KR} and N_{DE} are the number of positively charged residues (K and R) and negatively charged residues (D and E) in the mutated chain respectively.

We also tested the performance of other machine learning algorithms, including Random Forest (RF), Back Propagation Neural Network (BPNN), Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost), however, the multiple linear regression algorithm shows the best performance (Table S3).

Statistical analysis and evaluation of performance

Pearson correlation coefficient (R) and Root Mean Square Error (RMSE) are used to measure the agreement between experimental and predicted values of binding affinity change. All correlation coefficients presented in the paper are significantly different from zero (p-value < 0.01, *t*-test). RMSE (kcal mol^{-1}) is the standard deviation of the residuals (prediction errors). For

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

checking whether the difference in performance between PremPRI and other methods is significant, Hittner2003 test (46,47) is used for comparing whether the difference in correlation coefficients between PremPRI and the other methods is significant. The changes in the area under the receiver operating characteristics curve (AUC of ROC) are tested by Delong test (48). All tests are implemented in R.

Receiver Operating Characteristics and precision-recall analyses were performed to quantify the performance of different methods in distinguishing mutations highly decreasing binding affinity ($\Delta\Delta G_{exp} \geq 1$ kcal mol^{-1}) from others. True positive rate (TPR) and false positive rate (FPR) is defined as TPR=TP/(TP + FN) and FPR=FP/(FP+TN) (TP: true positive; TN: true negative; FP: false positive; FN: false negative) respectively. In addition, the maximal Matthews correlation coefficient (MCC) value is reported for each method by calculating the MCC across a range of thresholds:

$$\mathrm{MCC} = \frac{TP * TN - \mathrm{FP} * \mathrm{FN}}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

RESULTS and DISCUSSION

Multiple linear regression model of PremPRI

The PremPRI model is built using multiple linear regression algorithm and composed of ten structure- and one sequence-based features. The p-value and importance for each feature are shown in Table S2 that indicate all of the features contribute significantly to the model. To investigate the linear associations among different features, we performed multicollinearity analysis. Table S4 presents that the variance inflation factor (VIF) of each feature is less than three, indicating low collinear relationships among 11 independent variables in our multiple linear regression model.

The performance of PremPRI trained and tested on the S248 set is shown in Figure 1a and Table 1. The Pearson correlation coefficient between experimental and calculated binding affinity change is 0.72 and the corresponding RMSE and slope is 0.76 and 1.00 respectively. Moreover, PremPRI does not present bias to the majority of alanine-scanning mutations and yields a R of 0.71 for alanine-scanning and 0.78 non-alanine-scanning mutations respectively (Table S5). In addition, a better performance is observed for mutations from protein-single stranded RNA

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

complexes (Protein-ssRNA) than those from protein-double stranded RNA complexes (Protein-dsRNA) (Table S5).

PremPRI takes about four minutes to perform the calculation for a single mutation in a protein-RNA complex with ~ 330 residues and ~ 70 nucleotides, and requires 30 seconds for each additional mutation introduced in the same complex. Therefore, the running time for calculating one thousand mutations is about 14 hours.

Performance on three types of cross-validation

Overfitting is one of the major problems in machine learning. In order to check whether our method has this issue, we performed three types of cross-validation (CV1, CV2 and CV3). In the CV1 and CV2, we randomly chose 50% and 80% of all mutations from the S248 dataset to train the model, respectively, and used the remaining mutations for blind testing; both procedures were repeated 100 times. In CV3, the model was trained and tested using non-overlapping sets of protein-RNA complexes. Namely, we left one complex and its mutations out as the test set and trained the model using the rest of the complexes/mutations (leave-one-complex-out validation); this process was repeated for each complex.

The average Pearson correlation coefficient for both CV1 and CV2 is 0.68, and the R for each round out of 100 is greater than 0.42 (Table 1 and Fig. 1b). In the leave-one-complex-out validation (CV3), the correlation coefficient reaches to 0.61 (Fig. 1c) and RMSE = 0.87 kcal mol^{-1} (Table 1). In addition, the performances for the alanine-scanning and non-alanine-scanning mutations, and the mutations from Protein-ssRNA and Protein-dsRNA complexes remain relatively high in the CV3 with R of ~ 0.60 for each category of mutations (Table S5). Moreover, the analysis of the variation of the weighting coefficient for each feature in the three types of cross-validation also illustrates that our PremPRI model does not overfit on its training dataset and all features have significant contribution to the energy function (Table S6).

Comparison with other methods

We compared our method with three other available computational methods, mCSM-NA (26), PrabHot (29) and FoldX5.0 (35), for predicting the effects of mutations on protein-RNA interactions. mCSM-NA calculates the changes in protein-nucleic acid binding affinity induced by single mutations using graph-based signatures. PrabHot uses a combination of sequence-,

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

structure- and residue-interaction-network-based features to identify hot spots at the protein-RNA binding interfaces without calculating the values of binding affinity change. FoldX5.0 calculates the binding affinity changes using an empirical energy function that is parametrized on experimentally determined unfolding free energy changes. All of them are machine learning approaches and the training datasets for parameterizing mCSM-NA and PrabHot are shown in the Table S1. The number of common mutations between S248 and the training datasets of mCSM-NA and PrabHot is 16 and 92 respectively, and there is no common mutation between S248 and the training dataset of FoldX5.0 since it is parameterized on unfolding free energy changes. The performances presented in Table 2 illustrate that PremPRI performs better than mCSM-NA and FoldX methods with the highest R and lowest RMSE.

Moreover, we performed ROC and precision-recall analyses in order to estimate the performance of PremPRI to predict the highly decreasing mutations ($\Delta\Delta G_{exp} \geq 1$ kcal mol^{-1}). Figure 2a and Table 2 demonstrate the high performance of PremPRI in predicting mutations highly decreasing protein-RNA binding affinity, outperforming mCSM-NA, FoldX and PrabHot methods. In addition, our method performs well for both interfacial and non-interfacial mutations (Fig. 2b).

Furthermore, we applied all methods to a test case that includes three single mutations occurring in the highly conserved triad Thr-Met-Gly of Ribosomal protein L1 and their experimental values of binding affinity changes (49). All three mutations lower the protein-RNA binding largely (The experimental values of $\Delta\Delta G_{exp}$ are presented in the Table S7). Crystal structure of ribosomal protein L1 from Thermus thermophilus (TthL1) in complex with a specific 80 nt fragment of 23S rRNA (PDB ID: 3U4M) is used to perform the calculations. The predictions shown in Table S7 indicate that the PremPRI has the best performance and predicts all three mutations as highly decreasing mutations.

ONLINE WEBSERVER

Input

The PremPRI webserver requires the 3D structure of a protein-RNA complex that can be retrieved from the Protein Data Bank with the PDB code input or provided by a file with the atomic coordinates uploaded by the user (Fig. S3a). In either case, the structure file must include at least two chains, one is protein and the other is RNA. After the structure is retrieved correctly, the server

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

will display a 3D view of the complex and list the corresponding protein or RNA name for each chain (Fig. S3b). At the second step, two interaction partners should be defined, and one or multiple chains can be assigned to each partner. Only assigned chains will be considered during the prediction. At the third step of selecting mutations, three options are provided allowing users to do large-scale mutational scanning (see Fig. S3c). In the option of "Specify One or More Mutations Manually", the user can submit the specified mutations and visualize each mutated site in the protein-RNA complex structure. "Alanine Scanning for Each Protein Chain" option is used to perform alanine scanning for each protein chain. The "Upload Mutation List" option allows users to submit a list of mutations specified in the uploaded file.

Output

For each single mutation in a protein-RNA complex, the PremPRI server provides: ΔΔG (kcal mol^{-1}), predicted binding affinity change, and positive and negative signs correspond to the mutations decreasing and increasing binding affinity respectively; Interface (yes/no), shows whether the mutation occurred at the protein-RNA binding interface. When a residue's solvent accessible surface area in the complex is lower than in the unbound partner, it is defined as located at the interface. In addition, the minimized 3D structure of the mutant complex is provided for the user to download. Furthermore, for each mutation, PremPRI provides an interactive 3D viewer which shows the non-covalent interactions between the residue in the mutated site and its adjacent residues/nucleic acids, generated by Arpeggio (50). An example is provided in Figure S4.

ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China (Grant No. 31701136), Natural Science Foundation of Jiangsu Province, China (Grant No. BK20170335), and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Competing interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

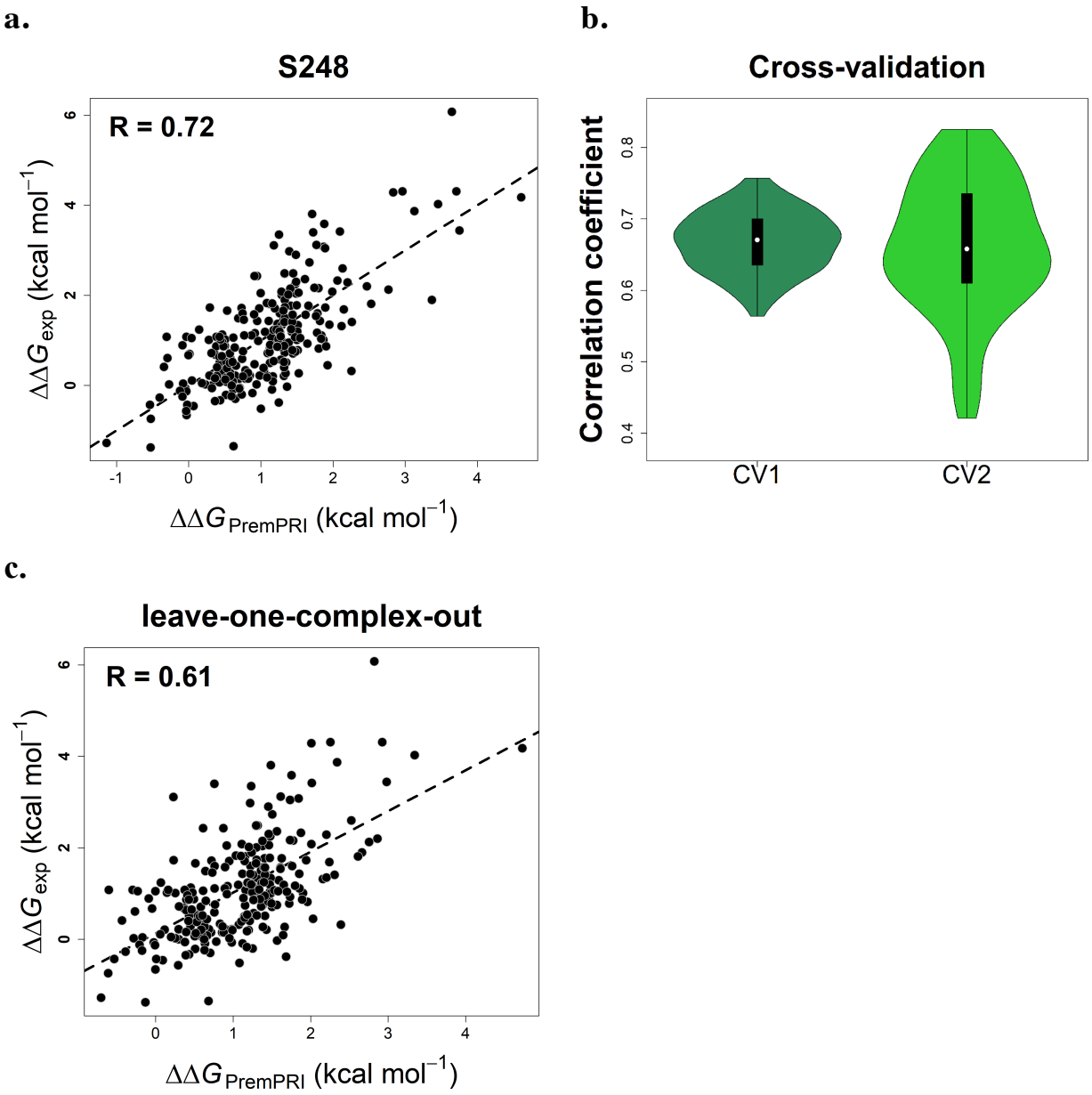

Figure 1. Pearson correlation coefficients between experimental and calculated changes in binding affinity for PremPRI trained and tested on S248 dataset (a), using two types of cross-validation (CV1 and CV2) (b) and performing leave-one-complex-out validation (CV3) (c), respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

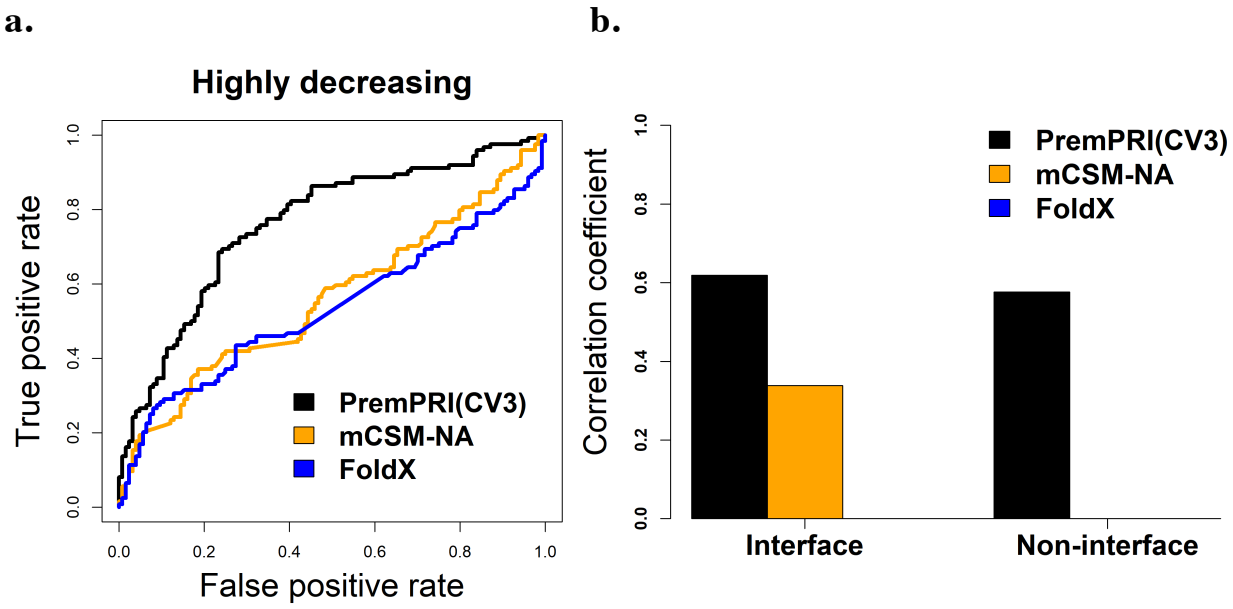

Figure 2. Performance for three methods of PremPRI, mCSM-NA and FoldX applied on the S248 dataset. The leave-one-complex-out validation (CV3) results of PremPRI are used. (a) ROC curves for predicting highly decreasing mutations. The number of highly decreasing mutations in S248 is 124. (b) Pearson correlation coefficients between predicted and experimental ΔΔG for mutations located at protein-RNA binding interface and non-interface, respectively. Only statistically significant correlation coefficients (p-value < 0.01, *t*-test) are shown in the figure.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. The performance for PremPRI trained and tested on S248 and performing three types of cross-validation.

Method	R	RMSE	Slope
PremPRI	0.72	0.76	1.00
PremPRI (CV1)	0.68	0.80	0.94
PremPRI (CV2)	0.68	0.80	0.95
PremPRI (CV3)	0.61	0.87	0.89

R: Pearson correlation coefficient between experimental and predicted ΔΔG values. RMSE (kcal mol^{-1}): root-mean square error. Slope: the slope of the regression line between experimental and predicted ΔΔG values. All correlation coefficients are statistically significantly different from zero (p-value < 0.01).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Comparison of methods' performances on the dataset of S248.

Method	R	RMSE	AUC-ROC	AUC-PR	MCC
PremPRI (CV3)	0.61	0.87	0.76	0.76	0.45
FoldX	0.20*	1.57	0.53*	0.59	0.24
mCSM-NA	0.24*	5.41	0.56*	0.60	0.22
PrabHot	-	-	0.58*	0.61	0.26

R: Pearson correlation coefficient. RMSE (kcal mol^{-1}): root-mean square error. AUC-ROC: the AUC values of ROC curves. AUC-PR: the AUC values of Precision-recall curves. MCC: maximal Matthews correlation coefficient. All correlation coefficients are statistically significantly different from zero (p-value < 0.01). *p-value < 0.01 compared to PremPRI (Hittner2003 and DeLong tests are used for comparing two correlation coefficients and AUC values, respectively).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Gerstberger, S., Hafner, M. and Tuschl, T. (2014) A census of human RNA-binding proteins. *Nature reviews. Genetics*, **15**, 829-845.
2. Mitchell, S.F. and Parker, R. (2014) Principles and properties of eukaryotic mRNPs. *Molecular cell*, **54**, 547-558.
3. Fabian, M.R., Sonenberg, N. and Filipowicz, W. (2010) Regulation of mRNA translation and stability by microRNAs. *Annual review of biochemistry*, **79**, 351-379.
4. Licatalosi, D.D. and Darnell, R.B. (2010) RNA processing and its regulation: global insights into biological networks. *Nature reviews. Genetics*, **11**, 75-87.
5. Mittal, N., Roy, N., Babu, M.M. and Janga, S.C. (2009) Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 20300-20305.
6. Hogan, D.J., Riordan, D.P., Gerber, A.P., Herschlag, D. and Brown, P.O. (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS biology*, **6**, e255.
7. Kechavarzi, B. and Janga, S.C. (2014) Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biol*, **15**, R14.
8. Lukong, K.E., Chang, K.W., Khandjian, E.W. and Richard, S. (2008) RNA-binding proteins in human genetic disease. *Trends in genetics : TIG*, **24**, 416-425.
9. Teng, S., Madej, T., Panchenko, A. and Alexov, E. (2009) Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. *Biophys J*, **96**, 2178-2188.
10. Nishi, H., Tyagi, M., Teng, S., Shoemaker, B.A., Hashimoto, K., Alexov, E., Wuchty, S. and Panchenko, A.R. (2013) Cancer missense mutations alter binding properties of proteins and their interaction networks. *Plos One*, **8**, e66273.
11. Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y. *et al.* (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, **161**, 647-660.
12. Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M. and Yu, H. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology*, **30**, 159-164.
13. Ozdemir, E.S., Gursoy, A. and Keskin, O. (2018) Analysis of single amino acid variations in singlet hot spots of protein-protein interfaces. *Bioinformatics (Oxford, England)*, **34**, i795-i801.
14. Jubb, H.C., Pandurangan, A.P., Turner, M.A., Ochoa-Montano, B., Blundell, T.L. and Ascher, D.B. (2017) Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in biophysics and molecular biology*, **128**, 3-13.
15. Kobren, S.N. and Singh, M. (2019) Systematic domain-based aggregation of protein structures highlights DNA-, RNA- and other ligand-binding positions. *Nucleic acids research*, **47**, 582-593.
16. Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J. and Godzik, A. (2015) A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS computational biology*, **11**, e1004518.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

17. Teh, H.F., Peh, W.Y., Su, X. and Thomsen, J.S. (2007) Characterization of protein--DNA interactions using surface plasmon resonance spectroscopy with various assay schemes. *Biochemistry-Us*, **46**, 2127-2135.
18. Velazquez-Campoy, A., Ohtaka, H., Nezami, A., Muzammil, S. and Freire, E. (2004) Isothermal titration calorimetry. *Current protocols in cell biology*, **Chapter 17**, Unit 17 18.
19. Getov, I., Petukh, M. and Alexov, E. (2016) SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach. *International journal of molecular sciences*, **17**, 512.
20. Guerois, R., Nielsen, J.E. and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, **320**, 369-387.
21. Dehouck, Y., Kwasigroch, J.M., Gilis, D. and Rooman, M. (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC bioinformatics*, **12**, 151.
22. Li, M., Simonetti, F.L., Goncearenco, A. and Panchenko, A.R. (2016) MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. *Nucleic acids research*, **44**, W494-501.
23. Dehouck, Y., Kwasigroch, J.M., Rooman, M. and Gilis, D. (2013) BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. *Nucleic acids research*, **41**, W333-339.
24. Rodrigues, C.H.M., Myung, Y., Pires, D.E.V. and Ascher, D.B. (2019) mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. *Nucleic acids research*, **47**, W338-w344.
25. Petukh, M., Dai, L. and Alexov, E. (2016) SAAMBE: Webserver to Predict the Charge of Binding Free Energy Caused by Amino Acids Mutations. *International journal of molecular sciences*, **17**, 547.
26. Pires, D.E.V. and Ascher, D.B. (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic acids research*, **45**, W241-w246.
27. Zhang, N., Chen, Y., Zhao, F., Yang, Q., Simonetti, F.L. and Li, M. (2018) PremPDI estimates and interprets the effects of missense mutations on protein-DNA interactions. *PLoS computational biology*, **14**, e1006615.
28. Peng, Y., Sun, L., Jia, Z., Li, L. and Alexov, E. (2018) Predicting protein-DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver. *Bioinformatics (Oxford, England)*, **34**, 779-786.
29. Pan, Y., Wang, Z., Zhan, W. and Deng, L. (2018) Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics (Oxford, England)*, **34**, 1473-1480.
30. Kruger, D.M., Neubacher, S. and Grossmann, T.N. (2018) Protein-RNA interactions: structural characteristics and hotspot amino acids. *RNA (New York, N.Y.)*, **24**, 1457-1465.
31. Barik, A., Nithin, C., Karampudi, N.B., Mukherjee, S. and Bahadur, R.P. (2016) Probing binding hot spots at protein-RNA recognition sites. *Nucleic acids research*, **44**, e9.
32. Kumar, M.D., Bava, K.A., Gromiha, M.M., Prabakaran, P., Kitajima, K., Uedaira, H. and Sarai, A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res*, **34**, D204-206.
33. Liu, L., Xiong, Y., Gao, H., Wei, D.Q., Mitchell, J.C. and Zhu, X. (2018) dbAMEPNI: a database of alanine mutagenic effects for protein-nucleic acid interactions. *Database*, **2018**.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.07.029520; this version posted April 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

34. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.
35. Delgado, J., Radusky, L.G., Cianferoni, D. and Serrano, L. (2019) FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics (Oxford, England)*, **35**, 4168-4169.
36. Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: visual molecular dynamics. *Journal of molecular graphics*, **14**, 33-38, 27-38.
37. MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S. *et al.* (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, **102**, 3586-3616.
38. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L. and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J Comput Chem*, **26**, 1781-1802.
39. Hoover, W.G. (1985) Canonical dynamics: Equilibrium phase-space distributions. *Physical review. A, General physics*, **31**, 1695-1697.
40. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry*, **4**, 187-217.
41. Chakrabarty, B. and Parekh, N. (2016) NAPS: Network Analysis of Protein Structures. *Nucleic acids research*, **44**, W375-382.
42. Aric Hagberg, P.S., Daniel Schult. (2008) Exploring network structure, dynamics, and function using NetworkX. *7th Python in Science Conference (SciPy2008)*, 11-15.
43. Joosten, R.P., te Beek, T.A., Krieger, E., Hekkelman, M.L., Hooft, R.W., Schneider, R., Sander, C. and Vriend, G. (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Res*, **39**, D411-419.
44. Brender, J.R. and Zhang, Y. (2015) Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLoS computational biology*, **11**, e1004494-e1004494.
45. Sweet, R.M. and Eisenberg, D. (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *Journal of Molecular Biology*, **171**, 479-488.
46. Hittner, J.B., May, K. and Silver, N.C. (2003) A Monte Carlo evaluation of tests for comparing dependent correlations. *J Gen Psychol*, **130**, 149-168.
47. Diedenhofen, B. and Musch, J. (2015) cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS One*, **10**, e0121945-e0121945.
48. DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837-845.
49. Tishchenko, S., Kostareva, O., Gabdulkhakov, A., Mikhaylina, A., Nikonova, E., Nevskaya, N., Sarskikh, A., Piendl, W., Garber, M. and Nikonov, S. (2015) Protein-RNA affinity of ribosomal protein L1 mutants does not correlate with the number of intermolecular interactions. *Acta crystallographica. Section D, Biological crystallography*, **71**, 376-386.
50. Jubb, H.C., Higueruelo, A.P., Ochoa-Monta?o, B., Pitt, W.R., Ascher, D.B. and Blundell, T.L. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J Mol Biol*, **429**, 365-371.