PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

PremPDI estimates and interprets the effects of missense mutations on protein-DNA interactions

Ning Zhang[1☯], **Yuting Chen**[1☯], **Feiyang Zhao**[1], **Qing Yang**[1], **Franco L. Simonetti**[2], **Minghui Li**[1]*

1 School of Biology and Basic Medical Sciences, Soochow University, Suzhou, China, **2** Institute Leloir Foundation, Buenos Aires, Argentina

☯ These authors contributed equally to this work.
* minghui.li@suda.edu.cn

Abstract

Protein-DNA interactions play important roles in regulations of many vital cellular processes, including transcription, translation, DNA replication and recombination. Sequence variants occurring in these DNA binding proteins that alter protein-DNA interactions may cause significant perturbations or complete abolishment of function, potentially leading to diseases. Developing a mechanistic understanding of impacts of variants on protein-DNA interactions becomes a persistent need. To address this need we introduce a new computational method PremPDI that predicts the effect of single missense mutation in the protein on the protein-DNA interaction and calculates the quantitative binding affinity change. The PremPDI method is based on molecular mechanics force fields and fast side-chain optimization algorithms with parameters optimized on experimental sets of 219 mutations from 49 protein-DNA complexes. PremPDI yields a very good agreement between predicted and experimental values with Pearson correlation coefficient of 0.71 and root-mean-square error of 0.86 kcal mol^{-1}. The PremPDI server could map mutations on a structural protein-DNA complex, calculate the associated changes in binding affinity, determine the deleterious effect of a mutation, and produce a mutant structural model for download. PremPDI can be applied to many tasks, such as determination of potential damaging mutations in cancer and other diseases. PremPDI is available at http://lilab.jysw.suda.edu.cn/research/PremPDI/.

Author summary

Developing methods for accurate prediction of effects of amino acid substitutions on protein-DNA interactions is important for a wide range of biomedical applications such as understanding disease-causing mechanism of missense mutations and guiding protein engineering. Very few methods have been developed for predicting the effects of mutations on protein-DNA binding affinity. Here we report a new computational method, PRedicts the Effects of single Mutations on Protein-DNA Interactions (PremPDI). The core of the PremPDI method is based on molecular mechanics force fields and fast side-

OPEN ACCESS

Citation: Zhang N, Chen Y, Zhao F, Yang Q, Simonetti FL, Li M (2018) PremPDI estimates and interprets the effects of missense mutations on protein-DNA interactions. PLoS Comput Biol 14(12): e1006615. https://doi.org/10.1371/journal.pcbi.1006615

Editor: Emil Alexov, Clemson University, UNITED STATES

Received: September 11, 2018

Accepted: November 1, 2018

Published: December 11, 2018

Data Availability Statement: The dataset is available for download from http://lilab.jysw.suda.edu.cn/research/PremPDI/download.

Funding: 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. FLS was supported by the Argentinian National Research Council (CONICET). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

chain optimization algorithms that makes the PremPDI algorithm efficient and being fast enough to handle large number of cases. The performance of the PremPDI protocol was tested against experimentally determined binding free energy changes of 219 mutations from 49 protein-DNA complexes and yields very good correlation coefficient. The PremPDI webserver is available to the community at http://lilab.jysw.suda.edu.cn/research/PremPDI/.

Introduction

There has been a rapid development of genome-wide techniques in the last decade along with significant lowering of the cost of gene sequencing, which generated widely available genomic data. However, the interpretation of genomic data and prediction of the association of genetic variations with diseases and phenotypes still require significant improvement [1]. Crucial prerequisite for proper biological function is a protein's ability to establish highly selective interactions with macromolecular partners. Protein-DNA interactions play important roles in regulations of many vital cellular processes, including transcription, translation, DNA replication, repair and recombination. Sequence variants occurring in these DNA binding proteins that alter protein-DNA interactions may cause significant perturbations or complete abolishment of function, potentially leading to many diseases, such as cancer and heart diseases [2–4]. One possible way to assess the effect of a mutation on protein-DNA interaction is to experimentally measure the binding affinity change. However, while site-directed mutagenesis methods are inexpensive and fast, surface plasmon resonance [5], isothermal titration calorimetry [6], FRET [7] and other methods used to measure binding affinity can be time-consuming and costly. Therefore, the development of reliable computational approaches to predict the effects of missense mutations on proteins and their complexes would give us important clues for identifying functionally important missense mutations, understanding the molecular mechanisms of diseases and facilitating their treatment and prevention.

With recent rapid advances in computational biology, many approaches have been developed to offer a phenotypic classification of mutations into damaging and neutral categories [8–10], to calculate the impact of mutations on protein stability [11–13] and protein-protein interactions [14–18]. Previously, we developed two methods for predicting the effect of single mutation on protein-protein binding affinity change. One used modified MM/PBSA, statistical scoring energy functions and structure minimization protocol with explicit solvent model [17]. The other updated method of MutaBind [14], which combined additional features and used a 100-step energy minimization in the gas phase that considerably increases the prediction accuracy and calculation speed. Our method was applied to predict the effects of cancer mutations on the binding between CBL ubiquitin ligase and E2 conjugating enzyme, where predicted binding affinity changes were successfully compared with the experiments using cancer and non-cancer cell lines [19]. However, very few methods can predict the effects of mutations on protein-DNA binding affinity [20]. Very recently, two prediction methods with servers, mCSM-NA [21] and SAMPDI [22], were proposed for performing this task. mCSM-NA relies on graph-based signatures and can predict the effect of single mutation on protein-DNA and protein-RNA binding, while SAMPDI combines modified MM/PBSA based energy terms with additional knowledge-based terms for predicting the protein-DNA binding affinity change upon single mutation. As we know, machine learning methods that use different features and training sets may produce different performances on diverse mutations and complexes[23]. Therefore, more fast and accurate computational methods need to be developed for increasing the range of applications on different kinds of complexes and mutations and explaining the mechanisms, such as the molecular mechanisms of disease progression caused by mutations.

To address this need we present a new computational method and webserver, PremPDI (http://lilab.jysw.suda.edu.cn/research/PremPDI/) which is based on molecular mechanics force fields and fast side-chain optimization algorithms. PremPDI can evaluate the effects of sequence variants and disease mutations (both interfacial and non-interfacial mutations) on protein-DNA interactions; calculate the quantitative change in binding affinity upon single mutation; assess deleterious effects and produce models of mutant complexes. PremPDI is validated using different types of cross-validation and is compared with two other methods using a variety of training and test sets. PremPDI can be applied to many tasks, including finding potential driver missense mutations in cancer, investigating the effects of sequence variations on protein fitness in evolution and protein design.

Methods

Compilation of experimental datasets of mutations

ProNIT database [24] includes experimentally measured values of changes in binding free energies upon single and multiple amino acid substitutions (called “mutations” hereafter) derived from the scientific literatures for protein-nucleic acid complexes with experimentally determined structures. dbAMEPNI database [25], being developed recently, focuses on the effects of single alanine-scanning mutations on the experimentally measured binding affinities between protein and nucleic acid. It comprises a total of 577 mutations with quantitatively characterized thermodynamic effects, among of them 345 were taken from ProNIT database. Both databases were used for compiling the dataset for parameterization of PremPDI. The following criteria were applied in constructing our dataset: removal complexes without wild-type protein structures or with modified residues or nucleotides at the binding interface of protein-DNA; removal mutations for their mutated sites with missing coordinates in the corresponding wild-type complex structures; eliminating ProNIT entries with multiple mutations restricting our set to single mutations. Furthermore, to avoid the inconsistency between nucleic acids used for measuring binding affinity and those for developing prediction model based on complex 3D structures, we carried out the comparison of sequence similarity between the nucleic acids of binding sites observed in the protein-DNA structures and the sequences used in the corresponding experiments. Then the entries with high sequence similarity (80%) for the nucleic acids in the binding interface were kept. ProNIT database includes the sequences of DNA used for measuring binding affinity, while dbAMEPNI database does not. So, we manually compiled them from the corresponding references. There are some entries where several experimental values are available for the same mutation. For these cases that are not drastically different from each other, we used an average value of experimental changes in binding free energy. In addition, 20 mutations from five protein-DNA complexes abstracted from SAMPDI training set [22] were also included in our dataset. As a result, the experimental set used in this study includes 219 single mutations from 49 wild-type protein-DNA complexes (it will be referred to as “Prempdi”) (S1 Table). Only 105 mutations obtained from ProNIT database have the information of experimental pH. Thus, we chose the experimental pH to be neutral assuming that at neutral pH the ionizable residues have default charged states. The number of mutations for each protein-DNA complex is shown in S1 Fig We also compared our dataset with the training datasets used for developing SAMPDI and mCSM methods, and the details are shown in S1 Table.

Structure optimization protocol

Crystal or NMR structures of wild-type protein-DNA complexes were obtained from the Protein Data Bank (PDB) [26], and biological assembly 1 of crystal structure or the first model of NMR was used as the initial structure. First we introduced a single mutation on the wild-type Protein-DNA complex structure using BuildModel module from FoldX [27] software package. Missing heavy side chain atoms and hydrogen atoms were added for the wild type and mutant using VMD program [28] based on the topology file from the CHARMM36 force field [29]. Then a 100-step energy minimization in the gas phase was carried out for both wild type and mutant using harmonic restraints (with the force constant of 5 kcal mol^{-1} Å^{-2}) applied on the backbone atoms of all residues. Minimization was done only for protein-DNA complexes, and protein or nucleic acid structures of binding partners were retained assuming the rigid-body binding. The energy minimization was carried out with NAMD program version 2.12 [30] using the CHARMM36 force field [29]. A 12 Å cutoff distance for nonbonded interactions was applied to the systems. Lengths of hydrogen-containing bonds were constrained by the SHAKE algorithm [31]. The current structure optimization protocol was chosen based on its highest accuracy and speed. The performances for other structure optimization protocols that have been tried are shown in S2 Table. The minimized structures of wild-type and mutant complexes were used for the calculation of energy terms.

Calculation of binding energy terms

Our goal is to design a method to assess the effects of mutations on protein-DNA binding. Mutations can affect binding in different ways [32]. They may change the components of protein-DNA interaction energies, may affect the solvation of a complex, may change the hydrogen-bond network and may directly disrupt binding hotspot sites [33]. Besides, the interactions between protein and the two types of nucleic acids (DNA and RNA) are also different, which was validated by a detailed computational comparison at the atomic contact level [34]. Here, through analysis of different kinds of protein sequence and structural features (S3 Table shows all features considered in our model selection), we found that nine features contributed significantly to the quality of multiple linear regression model (MLR) for the calculation of $\Delta\Delta G$ value (change in binding affinity upon mutation) affecting protein-DNA interactions (Table 1). The features that contribute significantly to the quality of PremPDI model are described below.

Table 1. The p-value and importance of each feature in energy function for binding affinity change determined by multiple linear regression (MLR).

Feature	P-value	Importance
$SA^{wt}_{com/p2}$	5.74e-09	0.47
$\Delta\Delta G_{solv}$	6.34e-08	0.41
ΔN^{p1-p2}_{Hbond}	5.37e-06	0.33
$\Delta E^{mut.(p1-p2)}_{elec}$	1.98e-07	0.28
$N^{wt.(site-all)}_{Hbond}$	2.33e-06	0.27
L_{mut}	9.42e-04	0.26
$\Delta\Delta E^{site-all}_{vdw}$	4.30e-04	0.18
ΔE_{fold}	3.03e-03	0.17
$\Delta^{mut}_{location}$	2.28e-03	0.17

All features have significant contribution to the quality of the model with p-value < 0.01 (t-test). Standardized coefficients are used for describing the importance for MLR.

https://doi.org/10.1371/journal.pcbi.1006615.t001

- $\Delta\Delta G_{solv}$ is the difference between polar solvation energies of mutant and wild-type protein-DNA complexes ($\Delta\Delta G_{solv} = \Delta G_{solv}^{mut} - \Delta G_{solv}^{wt}$). ΔG_{solv}^{wt} and ΔG_{solv}^{mut} are the differences between polar solvation energies of a complex and each interacting partner ($\Delta G = G_{com} - G_{p1} - G_{p2}$) (p1: partner1, proteins; p2: partner2, DNA) in water for wild-type and mutant complexes respectively. These terms are calculated from solving the Poisson-Boltzmann equation with PBEQ module [35] of CHARMM program [36]. For the PB calculation, dielectric constants, $\varepsilon = 2, 6, 10, 14, 18$ and 20, were tested using the optimized minimization protocol and energy function. As a result, $\varepsilon = 2$ for the protein interior and $\varepsilon = 80$ for the exterior aqueous environment were used for polar solvation energy calculations in our energy model with the best performance (the testing results using different dielectric constants are shown in S4 Table). The ion concentration of zero was used for energy calculation[17].

- $\Delta\Delta E_{vdw}^{site-all}$ is the difference between Van der Waals interaction energies of mutant and wild type ($\Delta\Delta E_{vdw}^{site-all} = \Delta E_{vdw}^{mut.(site-all)} - \Delta E_{vdw}^{wt.(site-all)}$). $\Delta E_{vdw}^{wt.(site-all)}$ and $\Delta E_{vdw}^{mut.(site-all)}$ are Van der Waals interaction energies between residue in the mutated site and the rest of protein-DNA complex located within 10 Å from it for wild-type and mutant complexes respectively. They are calculated using ENERGY module of CHARMM program [36].

- $\Delta E_{elec}^{mut.(p1-p2)}$ is electrostatic interaction energy between protein and DNA within 10 Å from each other in mutant. They are calculated using ENERGY module of CHARMM program [36].

- $\Delta N_{Hbond}^{p1-p2}\, and\, N_{Hbond}^{wt.(site-all)}$: ΔN_{Hbond}^{p1-p2} is the difference between the number of hydrogen bonds formed in mutant and wild-type protein-DNA complexes ($\Delta N_{Hbond}^{p1-p2} = N_{Hbond}^{mut.(p1-p2)} - N_{Hbond}^{wt.(p1-p2)}$). $N_{Hbond}^{wt.(p1-p2)}$ and $N_{Hbond}^{mut.(p1-p2)}$ terms account for the number of hydrogen bonds formed between protein and DNA for wild-type and mutant complexes respectively; $N_{Hbond}^{wt.(site-all)}$ is the number of hydrogen bonds formed between residue in the mutated site and the rest of wild-type protein-DNA complex. Hydrogen bonds are identified with the CORMAN command of CHARMM program using the following criteria: the maximum distance between acceptor and hydrogen is 2.5 Å and the minimum angle of donor–hydrogen–acceptor is 90˚.

- $SA_{com/p2}^{wt}$ is the ratio of SA_{com}^{wt} and SA_{p2}^{wt}. SA_{com}^{wt} and SA_{p2}^{wt} are the solvent accessible surface areas of complex and DNA respectively for wild type. Solvent accessible surface area is calculated using SASA module of CHARMM program.

- $\Delta_{location}^{mut}$ is equal to 1 if the mutation occurs on protein-DNA interface, otherwise it is 0. We define a residue to be located on a protein-DNA interface if residue's solvent accessibility in the complex is lower than in the corresponding unbound partners.

- ΔE_{fold} is a pairwise statistical potential for protein folding which was obtained from an optimization procedure that maximizes thermodynamic stability for all proteins simultaneously [37]. It is obtained from Amino Acid Index Database with identifier of MIRL960101 (AAindex, http://www.genome.jp/aaindex/).

- L_{mut} *is* the length of mutated protein chain.

Results and discussion

Model training through multiple linear regression

The p-value and contribution of each term to the PremPDI model are shown in Table 1, and all terms contribute significantly to the energy model with p-values less than 0.01. If we train and test our model on the 'Prempdi' set, the Pearson correlation coefficient between

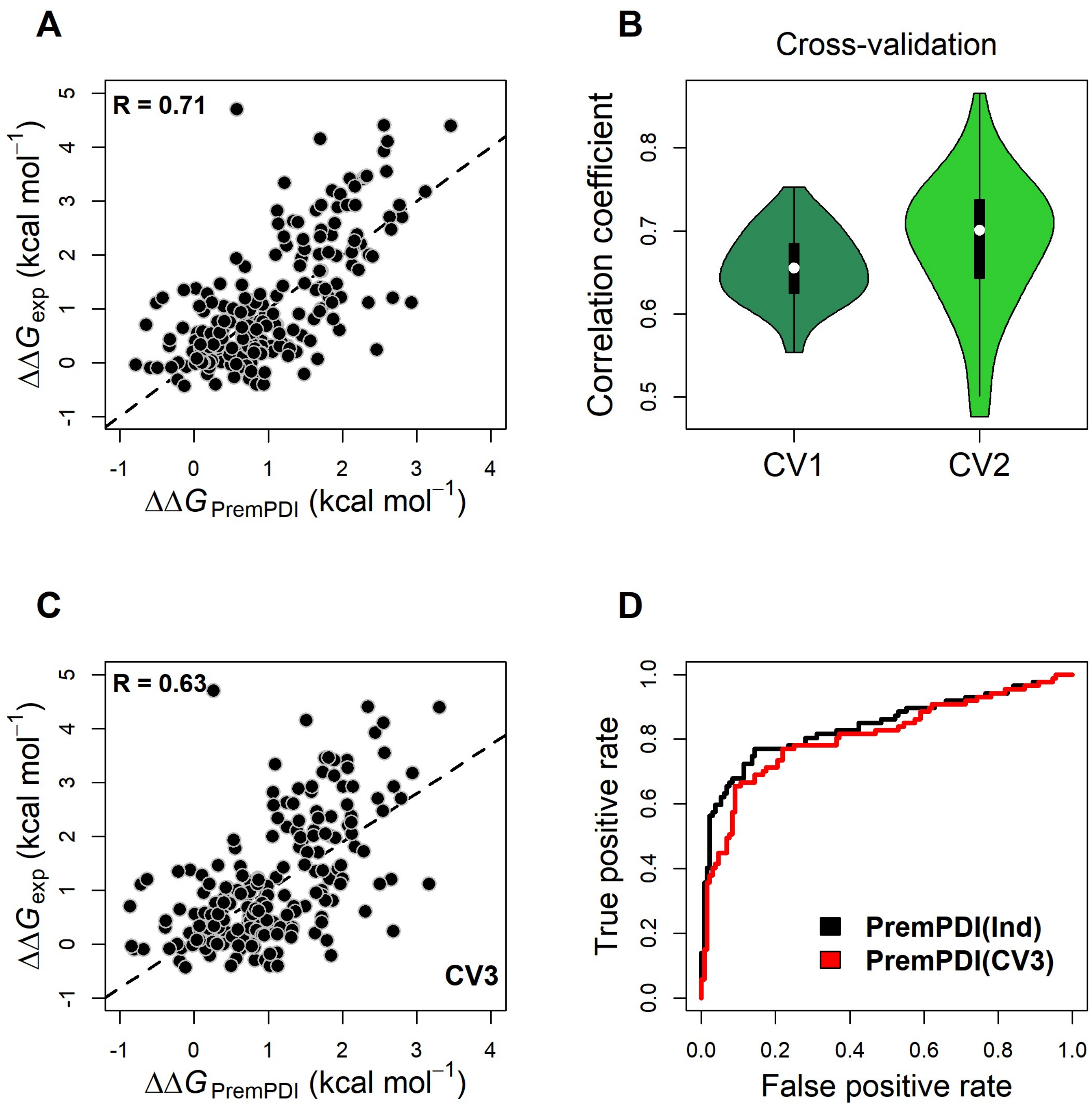


Fig 1. PremPDI performance. Pearson correlation coefficients between experimental and calculated changes in binding free energies (ΔΔG) for "Prempdi" training/test set (a), for two types of cross-validation (CV1 and CV2) (b) and for "leave-one-complex-out" cross-validation (CV3) (c). ROC curves for predictions of deleterious mutations applied on "Prempdi" set (d).

https://doi.org/10.1371/journal.pcbi.1006615.g001

experimental and calculated changes in binding free energies is R = 0.71 (Fig 1a and Table 2) and the corresponding root-mean-square error (RMSE) is 0.86 (Table 2). Among 219 mutations in "Prempdi" dataset, 179 ones belong to alanine-scanning single mutations defined as substitutions of residues into alanine and 134 ones located on the interfaces of protein-DNA complexes according to our definition (see Method' section). The results show that our model

Table 2. PremPDI performance.

Test set	Method	R	RMSE (kcal mol^{-1})	Slope
Prempdi	PremPDI	0.71	0.86	1
	PremPDI (CV1)	0.68	0.90	0.94
	PremPDI (CV2)	0.68	0.90	0.95
	PremPDI (CV3)	0.63	0.95	0.90
Alanine-scanning mutations	PremPDI	0.68	0.87	0.97
	PremPDI (CV3)	0.58	0.96	0.87
Non-Alanine- scanning mutations	PremPDI	0.64	0.81	0.88
	PremPDI (CV3)	0.58	0.88	0.72
Interfacial mutations	PremPDI	0.71	0.86	1.01
	PremPDI (CV3)	0.64	0.95	0.89
Non-interfacial mutations	PremPDI	0.69	0.85	0.98
	PremPDI (CV3)	0.59	0.95	0.91

R: Pearson correlation coefficient between experimental and predicted ΔΔG values. RMSE: root-mean square error. The last column shows 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). CV1 and CV2 results for "Alanine-scanning mutations", "Non-Alanine-scanning mutations", "Interfacial mutations" and "Non-interfacial mutations" test sets are shown in S5 Table.

https://doi.org/10.1371/journal.pcbi.1006615.t002

does not present bias to alanine-scanning mutations and yields good performance for non-alanine-scanning mutations with R = 0.64 and RMSE = 0.81 (Table 2). As was shown previously [14,17], mutations located on the interface region present average larger effects on protein-protein interactions and are better predicted compared to non-interface mutations. In this study, PremPDI yields statistically significant correlation (p-value < 0.01) in predicting non-interfacial mutations and the correlation reaches value as high as 0.69 and RMSE is 0.85. We also tried several other machine learning methods such as random forest, support vector machine and neural network to build our model using these nine features. Cross-validation and leave one complex validation that will be discussed in the next section show that multiple linear regression represents the best performance.

In addition, we performed multicollinearity analysis to investigate the linear association across each feature. Pearson correlation matrixes and variance inflation factors (VIF) for the energy features in PremPDI are shown in S6 Table. The results show that $\Delta\Delta G_{solv}$ has relatively strong correlation with ΔN_{Hbond}^{p1-p2} (R = -0.71), $SA_{com/p2}^{wt}$ has relatively strong correlation with L_{mut} with R of 0.74, and the rest of the correlations are either small or are not significantly different from zero. The VIFs of all features are less than three representing relatively low multicollinearity. We removed highly correlated features from our energy function that results in decrease of prediction accuracy. For instance, removal ΔN_{Hbond}^{p1-p2} from PremPDI MLR model leads to the decrease of correlation from 0.71 to 0.68. Thus, all nine features were kept in our final model to achieve the optimal performance.

PremPDI takes about five minutes to perform calculations for a single mutation in a protein-DNA complex with 300 residues and 30 nucleotides running on a single processor core, and it requires additional two-to-three minutes for each additional mutation per complex.

Evaluating the performance of PremPDI using cross-validation and leave one complex validation

Our goal is to construct a computational method that can achieve a high prediction accuracy for large and diverse sets of single mutations. In many cases, overfitting may occur when the parameters of computational methods are tuned to minimize the mean square deviations of predicted from experimental values in the training set, thus leading to the decreased generalized performance [38]. At the same time the training set should be as comprehensive as possible, while in our study the data set used for training and testing is relatively small. To address this issue, we performed three types of cross-validation. In case of "CV1" cross-validation (Fig 1b), 50% mutations selected randomly from "Prempdi" set were used for training and the remaining mutations for testing, the procedure was repeated 50 times. In "CV2" cross-validation we randomly chose 80% of all mutations as training and used the remaining 20% mutations for testing, also repeated 50 times. The average Pearson correlation coefficient is R = 0.68 for both "CV1" and "CV2" with small standard error of 0.06 (Fig 1b). The RMSE is 0.9 kcal mol^{-1} for both cross validations (Table 2).

Since the prediction accuracy of mutational effects largely depends on sequence and structure of a complex, we performed a "leave-one-complex-out" procedure ("CV3" cross-validation). Namely, we trained the parameters on experimental $\Delta\Delta G$ values of mutations from 48 protein-DNA complexes and then applied the model to mutations from the remaining one complex. This procedure was repeated for each complex. The Pearson correlation coefficient between experimental and computed $\Delta\Delta G$ values using this procedure is R = 0.63 with RMSE of 0.95 kcal mol^{-1} (Fig 1c and Table 2). In addition, for alanine-scanning, non-alanine-scanning, interfacial and non-interfacial mutations, they also present relatively high correlation coefficients and low RMSEs in "CV3" cross-validation, especially for interfacial mutations (Table 2).

We also analyzed the variation of the weighting coefficient for each feature in "CV1", "CV2" and "CV3" cross-validation respectively. The results are shown in S7 Table. The standard deviations of the weighting coefficients are relatively small even for "CV1" cross-validation, 50% mutations from "Prempdi" set were used for training and the remaining mutations for testing, which indicates the variation is not significant across each fold. In addition, the average weighting coefficients in each cross-validation were compared with the weighting coefficients of the final PremPDI model and the results show that the differences for all energy features are very small. All the validations indicate that our PremPDI model does not overfit on its training set and all features have significant contribution to the energy function.

Evaluating the performance of PremPDI to predict deleterious effects of mutations

Predicting the quantitative values of binding affinity changes is quite challenging. A much easier task, attempted by many studies, is to classify mutations based on their effects into deleterious or neutral. Several thresholds of experimentally determined $\Delta\Delta G$, 1, 1.5, 2.0 and 2.5 kcal mol^{-1}, were tested for defining mutations with deleterious (highly destabilizing) effects (see S2 Fig). The number of mutations in each category is shown in S2a Fig Threshold of 1 kcal mol^{-1} has the most balanced dataset. To quantify the performance of PremPDI scores, we performed Receiver Operating Characteristics (ROC) and precision-recall analyses. Sensitivity or true positive rate was defined as TPR = TP/(TP + FN) and specificity or true negative rate was defined as TNR = 1-FPR = TN/(FP+TN). Additionally, in order to account for imbalances in the labeled dataset, the quality of the predictions was described by Matthews correlation coefficient (MCC), a performance measure which is known to be more robust on unbalanced datasets:

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

S2b–S2e Fig show the ROC and precision-recall curves by applying PremPDI on the "Prempdi" training/test set using different thresholds. S2f Fig depicts the basic summary of performance metrics, including AUC for ROC and precision-recall curves and MCC. The results show that threshold of 1.5 kcal mol^{-1} has the highest AUC-ROC of 0.91 and MCC of 0.61 in distinguishing deleterious and neutral mutations (S2b and S2f Fig). Threshold of 1 kcal mol^{-1} has the highest AUC-PR of 0.83 and its AUC-ROC and MCC is 0.84 and 0.58 respectively (S2d and S2f Fig). S2c and S2e Fig show that threshold of 1 kcal mol^{-1} classification has the best performance in the deleterious mutation prediction with less than 10% false positive rate and more than 50% precision. Here, we choose $\Delta\Delta G_{exp}$ = 1 kcal mol^{-1} as the threshold to define deleterious effect, and it is also in agreement with SAMPDI method for classifying large and small effects [22]. Fig 1d shows the ROC curves for PremPDI and PremPDI (CV3) to distinguish deleterious and neutral effects using threshold of 1 kcal mol^{-1}. Therefore, PremPDI classifies a mutation as deleterious if its predicted $\Delta\Delta G$ is higher or equal to 1.10 kcal mol^{-1} (S3 Fig). This threshold corresponds to 14% FPR and 77% TPR which minimizes the value of error $\mathrm{ER} = \sqrt{(1 - TPR)^2 + FPR^2}$ to compensate retrieval sensitivity and specificity.

Comparison of PremPDI with other methods

We compared our method with the other two available machine learning methods, mCSM-NA [21] and SAMPDI [22]. mCSM-NA uses graph-based signatures to calculate the changes in protein-nucleic acid binding affinity upon single mutations. SAMPDI uses a combination of modified MM/PBSA based energy terms with additional knowledge-based terms to predict the $\Delta\Delta G$ values of interfacial mutations for protein-DNA complexes. The training sets for parameterizing PremPDI method and the other two have some differences, which is shown in S1 Table. Among 219 mutations from 49 complexes in PremPDI training set ("Prempdi"), 105 mutations from 16 complexes overlap with mCSM-NA training set of "Mcsm" (the overlapped set is named as "P.O.M") and 77 mutations from 11 complexes overlap with SAMPDI training set of "Sampdi" (the overlapped set is named as "P.O.S"). 114 mutations from 33 complexes in "Prempdi" are not included in the "Mcsm" (named as "P.D.M") and 142 mutations from 43 complexes in "Prempdi" are not in the "Sampdi" (named as "P.D.S"). Since SAMPDI is used in particular for interfacial mutations, we created a subset of "P.D.S" and named it as "P.D.S.I" that includes 77 interfacial mutations from 32 complexes.

We performed several types of comparisons between our method and the other two using four different test sets. "P.O.M" or "P.O.S" is the test set of overlapped mutations used for developing PremPDI and mCSM or SAMPDI respectively. So, we compared PremPDI with them using the model that built on the whole 'Prempdi' dataset. "P.D.M" or "P.D.S.I" test set represents the mutations that are included in the 'Prempdi' but not in the 'Mcsm' or 'Sampdi'. So, to be fair, we used both "leave-one-complex-out" (CV3) results and the model built on the independent 'P.O.M' or 'Prempdi-P.D.S.I' dataset (named as PremPDI(Ind)) to compare with the other methods respectively. Pearson correlation coefficients and RMSE between experimental measurements ($\Delta\Delta G_{exp}$) and predictions show that PremPDI presents a similar performance with mCSM-NA method and performs better than SAMPDI in predicting quantitative values of $\Delta\Delta$G (Table 3). ROC curves shown in Fig 2 and AUC-ROC, AUC-PR and MCC values presented in Table 3 (The number of mutations in each category is shown in S4 Fig) demonstrate that the performance of PremPDI is notable in estimating deleterious effects (highly destabilizing) for all test sets and better than mCSM-NA and SAMPDI methods.

Webserver input

The main requirement of the webserver is the 3D structure of a protein-DNA complex. The users can either input PDB code of the complex, then structures of either biological assemblies

Table 3. Comparison of methods' performances on different test sets.

Test set	Training set	Method	R	RMSE (kcal mol^{-1})	AUC-ROC	AUC-PR	MCC
P.O.M	Prempdi	PremPDI	0.80	0.81	0.88	0.87	0.54
	Mcsm	mCSM	0.76	0.95	0.82	0.79	0.50
P.O.S	Prempdi	PremPDI	0.68	0.63	0.88	0.81	0.52
	Sampdi	SAMPDI	0.39	0.80	0.66	0.53	0.27
P.D.M	Prempdi	PremPDI(CV3)	0.51	0.97	0.78	0.72	0.54
	P.O.M	PremPDI(Ind)	0.51	1	0.77	0.72	0.41
	Mcsm	mCSM	0.54	1.17	0.69	0.65	0.28
P.D.S.I	Prempdi	PremPDI(CV3)	0.70	1.10	0.85	0.82	0.62
	Prempdi- P.D.S.I	PremPDI(Ind)	0.74	1.08	0.85	0.83	0.68
	Sampdi	SAMPDI	0.53	1.32	0.79	0.71	0.35

R: Pearson correlation coefficient between experimental and predicted ΔΔG values. RMSE: root-mean square error. AUC-ROC: the AUC values of ROC curves. AUC-PR: the AUC values of Precision-recall curves. MCC: Matthews correlation. All correlation coefficients are statistically significantly different from zero (p-value < 0.01). The descriptions of training and test set are shown in S1 Table. Nine mutations do not have SAMPDI scores in the P.D.S.I test set, so they were excluded in the comparison.

https://doi.org/10.1371/journal.pcbi.1006615.t003

or asymmetric unit will be retrieved from the Protein Data Bank, or they can upload their own file with atomic coordinates. In either case, the structure file should contain at least two chains.

After the structure was retrieved correctly, the server will display a 3D view of the complex colored by chains or partners using the GLmol software. Each chain is listed with the corresponding protein or nucleic acid name. At the second step, two interacting partners should be defined. The user can assign one or multiple chains to either Partner 1 or Partner 2, but both partners should include at least one chain. Here, we restrict Partner 1 to proteins and Partner 2 to DNA and the selected protein/DNA chain will be put into the box of Partner1/Partner2 automatically. Only the selected chains of two partners will be taken into account during the calculation. If the interface size between two partners is more than 100 Å^2, we define them interacting with each other and then perform the calculation. Interface size is calculated as the difference between the solvent accessible surface areas of complex and unbound partners.

The third step is to select mutations (Fig 3). Each mutation will be treated independently and up to 16 single mutations can be selected for one submission. After the chain and the mutated residue are selected, they can be visualized in the wild-type complex using the 3D viewer.

Webserver output

For each mutation of a protein-DNA complex, PremPDI server provides the following results:

- ΔΔG (kcal mol^{-1}), predicted binding affinity change induced by single mutation. Positive and negative signs correspond to destabilizing and stabilizing mutations predicted to decrease and increase binding affinity respectively.
- Interface (yes/no), PremPDI defines a residue to be located on the protein-DNA interface if residue's solvent accessibility in the complex is lower than in the corresponding unbound partners.
- Deleterious (yes/no), PremPDI classifies a mutation as deleterious if ΔΔG is higher or equal to 1.10 kcal mol^{-1}. This threshold corresponds to a minimum value of ER to compensate retrieval sensitivity and specificity.
- Coordinates of the minimized mutant structure are provided for download.

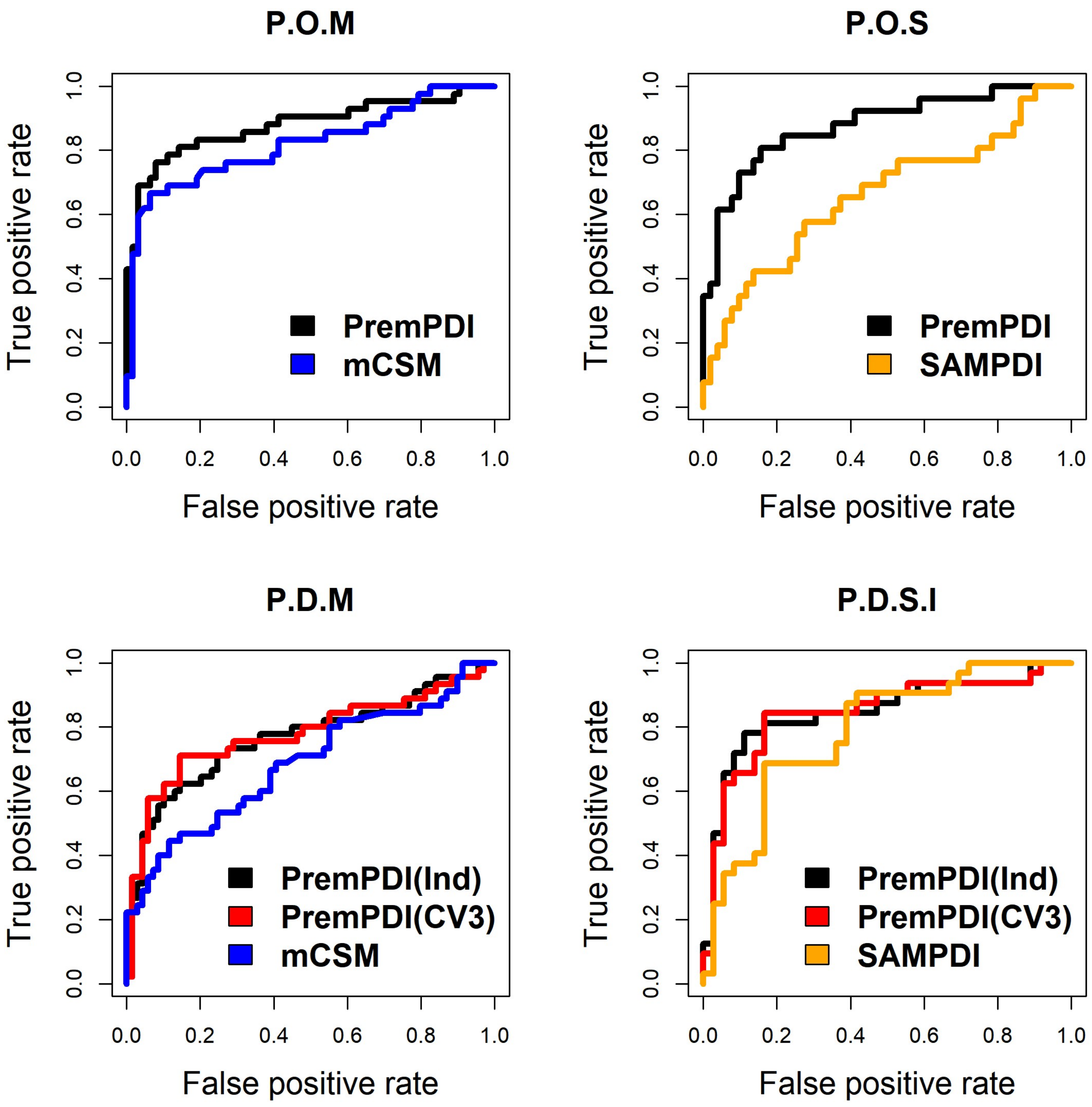
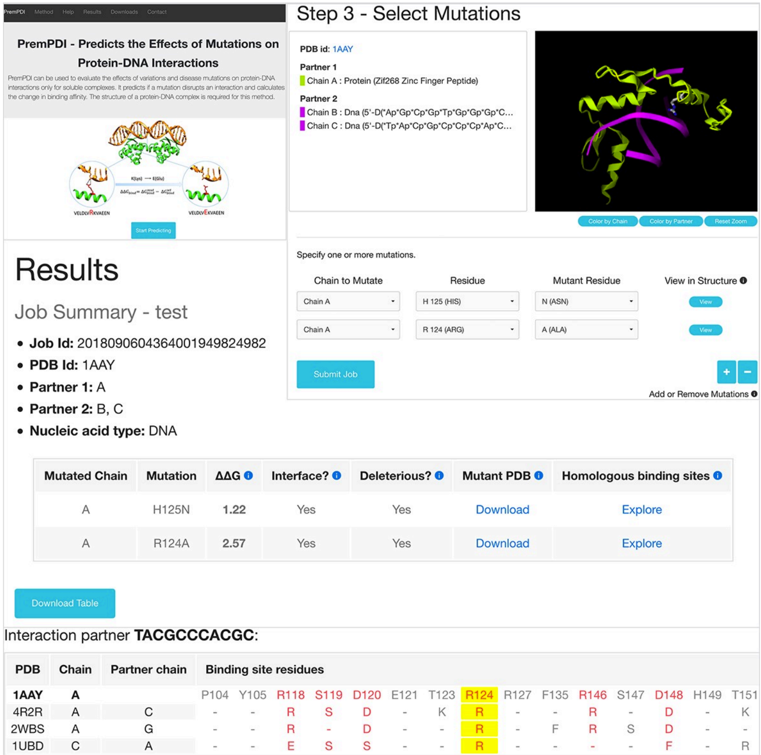


Fig 2. Assessment of classification performance between deleterious and neutral mutations. ROC curves for PremPDI, mCSM-NA and SAMPDI methods applied on different training and test set. More information is shown in Table 3.

https://doi.org/10.1371/journal.pcbi.1006615.g002

- Protein binding sites in protein-DNA complexes homologous to the query are identified using Inferred Biomolecular Interactions Server at NCBI (IBIS) server [39]. It allows testing mutations of aligned binding site residues in homologous protein-DNA in PremPDI.

Results can be viewed directly on the browser (Fig 3) or downloaded as a plain text file.



PremPDI - Predicts the Effects of Mutations on Protein-DNA Interactions

PremPDI can be used to evaluate the effects of variations and disease mutations on protein-DNA interactions only for soluble complexes. It predicts if a mutation disrupts an interaction and calculates the change in binding affinity. The structure of a protein-DNA complex is required for this method.

Step 3 - Select Mutations

PDB id: 1AAY

Partner 1

🟩 Chain A : Protein (Zif268 Zinc Finger Peptide)

Partner 2

🟪 Chain B : Dna (5'-D(*Ap*Gp*Cp*Gp*Tp*Gp*Gp*Gp*C...

🟪 Chain C : Dna (5'-D(*Tp*Ap*Cp*Gp*Cp*Cp*Cp*Ap*C...

Specify one or more mutations.

Chain to Mutate	Residue	Mutant Residue	View in Structure
Chain A	H 125 (HIS)	N (ASN)	View
Chain A	R 124 (ARG)	A (ALA)	View

Add or Remove Mutations

Results

Job Summary - test

- **Job Id:** 20180906043640019498249 82
- **PDB Id:** 1AAY
- **Partner 1:** A
- **Partner 2:** B, C
- **Nucleic acid type:** DNA

Mutated Chain	Mutation	ΔΔG	Interface?	Deleterious?	Mutant PDB	Homologous binding sites
A	H125N	1.22	Yes	Yes	Download	Explore
A	R124A	2.57	Yes	Yes	Download	Explore

Interaction partner **TACGCCCACGC**:

PDB	Chain	Partner chain	Binding site residues													
1AAY	**A**		P104	Y105	R118	S119	D120	E121	T123	R124	R127	F135	R146	S147	D148	H149	T151
4R2R	A	C	-	-	R	S	D	-	K	R	-	-	R	-	D	-	K
2WBS	A	G	-	-	R	-	D	-	-	R	-	F	R	S	D	-	-
1UBD	C	A	-	-	E	S	S	-	-	R	-	-	-	-	F	-	R

Fig 3. Left corner: The entry page of PremPDI server; right corner: The third step for selecting mutations, wild-type residue (R124) in the mutated site is shown in the 3D viewer; and bottom: Final results table and alignment of homologous binding sites.

https://doi.org/10.1371/journal.pcbi.1006615.g003

Supporting information

S1 Fig. The number of mutations for each protein-DNA complex.
(DOCX)

S2 Fig. Assessment of classification performance between deleterious and neutral mutations by applying PremPDI on "Prempdi" dataset using different thresholds. (a) The definition and the number of deleterious, neutral and stabilizing mutations for four thresholds. (b) ROC curves. (c) shows the ROC curves corresponding to FPR less than 10%. (d) Precision-recall curves. (e) shows the precision-recall curves corresponding to precision over 50%. (f) The AUC values of ROC curves and Precision-recall curves, and Matthews correlation (MCC) for four thresholds. The best performance is shown in bold font.
(DOCX)

S3 Fig. ROC curve for predicting deleterious mutations by applying PremPDI on the training set of "Prempdi". Red point corresponds to the minimization of the value of error ER $= \sqrt{(1 - TPR)^2 + FPR^2}$.
(DOCX)

S4 Fig. The number of deleterious, neutral and stabilizing mutations for four datasets of P.O.M, P.O.S, P.D.M and P.D.S.I. Nine mutations do not have SAMPDI scores in the P.D.S.I test set, so they were excluded in the comparison.
(DOCX)

S1 Table. The number of mutations in different data sets.
(DOCX)

S2 Table. Correlation between predicted and experimental values of ΔΔG for different structure optimization protocols. All calculations were performed by PremPDI energy function. "Prempdi-dbAMEPNI" includes 126 mutations and the mutations from dbAMEPNI database were not included in it. R: Pearson correlation coefficient between experimental and predicted ΔΔG values, and RMSE: root-mean squared error.
(DOCX)

S3 Table. Features considered in model selection.
(DOCX)

S4 Table. PremPDI performance using different dielectric constants for protein interior in the PB calculation.
(DOCX)

S5 Table. PremPDI performance.
(DOCX)

S6 Table. Correlation matrixes and variance inflation factors (VIF) for the energy features in PremPDI. Correlation coefficients that are greater than 0.5 are underlined. Only correlation coefficients that are statistically significantly different from zero ($P\text{-value} < 0.01$) are shown.
(DOCX)

S7 Table. Average weighting coefficients and corresponding standard deviation (in brackets) for all energy features in "CV1", "CV2" and "CV3" cross-validation respectively. The weighting coefficients from the final PremPDI model were also shown for the comparison.
(DOCX)

Acknowledgments

We thank Dr. Anna R. Panchenko for providing useful suggestions.

Author Contributions

Conceptualization: Minghui Li.

Data curation: Ning Zhang, Yuting Chen, Minghui Li.

Formal analysis: Ning Zhang, Yuting Chen, Minghui Li.

Funding acquisition: Minghui Li.

Investigation: Ning Zhang, Minghui Li.

Methodology: Ning Zhang, Yuting Chen, Qing Yang, Minghui Li.

Project administration: Minghui Li.

Resources: Feiyang Zhao, Franco L. Simonetti, Minghui Li.

Software: Ning Zhang, Yuting Chen, Minghui Li.

Supervision: Minghui Li.

Validation: Ning Zhang, Yuting Chen, Minghui Li.

Visualization: Ning Zhang, Yuting Chen, Minghui Li.

Writing – original draft: Minghui Li.

Writing – review & editing: Minghui Li.

References

1. Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E (2013) Molecular mechanisms of disease-causing missense mutations. J Mol Biol 425: 3919–3936. https://doi.org/10.1016/j.jmb.2013.07.014 PMID: 23871686
2. Muller PA, Vousden KH (2013) p53 mutations in cancer. Nat Cell Biol 15: 2–8. https://doi.org/10.1038/ncb2641 PMID: 23263379
3. Kechavarzi B, Janga SC (2014) Dissecting the expression landscape of RNA-binding proteins in human cancers. Genome Biol 15: R14. https://doi.org/10.1186/gb-2014-15-1-r14 PMID: 24410894
4. Sibanda BL, Chirgadze DY, Ascher DB, Blundell TL (2017) DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair. Science 355: 520–524. https://doi.org/10.1126/science.aak9654 PMID: 28154079
5. Teh HF, Peh WY, Su X, Thomsen JS (2007) Characterization of protein—DNA interactions using surface plasmon resonance spectroscopy with various assay schemes. Biochemistry 46: 2127–2135. https://doi.org/10.1021/bi061903t PMID: 17266332
6. Velazquez-Campoy A, Ohtaka H, Nezami A, Muzammil S, Freire E (2004) Isothermal titration calorimetry. Curr Protoc Cell Biol Chapter 17: Unit 17 18.
7. Hillisch A, Lorenz M, Diekmann S (2001) Recent advances in FRET: distance determination in protein-DNA complexes. Curr Opin Struct Biol 11: 201–207. PMID: 11297928
8. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812–3814. PMID: 12824425
9. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512
10. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS One 7: e46688. https://doi.org/10.1371/journal.pone.0046688 PMID: 23056405
11. Getov I, Petukh M, Alexov E (2016) SAAFEC: Predicting the Effect of Single Point Mutations on Protein Folding Free Energy Using a Knowledge-Modified MM/PBSA Approach. Int J Mol Sci 17: 512. https://doi.org/10.3390/ijms17040512 PMID: 27070572
12. Pires DE, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30: 335–342. https://doi.org/10.1093/bioinformatics/btt691 PMID: 24281696
13. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, et al. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25: 2537–2543. https://doi.org/10.1093/bioinformatics/btp445 PMID: 19654118
14. Li M, Simonetti FL, Goncearenco A, Panchenko AR (2016) MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. Nucleic Acids Res 44: W494–501. https://doi.org/10.1093/nar/gkw374 PMID: 27150810
15. Petukh M, Li M, Alexov E (2015) Predicting Binding Free Energy Change Caused by Point Mutations with Knowledge-Modified MM/PBSA Method. PLoS Comput Biol 11: e1004276. https://doi.org/10.1371/journal.pcbi.1004276 PMID: 26146996
16. Brender JR, Zhang Y (2015) Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. PLoS Comput Biol 11: e1004494. https://doi.org/10.1371/journal.pcbi.1004494 PMID: 26506533
17. Li M, Petukh M, Alexov E, Panchenko AR (2014) Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity. J Chem Theory Comput 10: 1770–1780. https://doi.org/10.1021/ct401022c PMID: 24803870
18. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D (2013) BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. Nucleic Acids Res 41: W333–339. https://doi.org/10.1093/nar/gkt450 PMID: 23723246
19. Li M, Kales SC, Ma K, Shoemaker BA, Crespo-Barreto J, et al. (2016) Balancing Protein Stability and Activity in Cancer: A New Approach for Identifying Driver Mutations Affecting CBL Ubiquitin Ligase Activation. Cancer Res 76: 561–571. https://doi.org/10.1158/0008-5472.CAN-14-3812 PMID: 26676746

20. Li M, Shoemaker BA, Thangudu RR, Ferraris JD, Burg MB, et al. (2013) Mutations in DNA-binding loop of NFAT5 transcription factor produce unique outcomes on protein-DNA binding and dynamics. J Phys Chem B 117: 13226–13234. https://doi.org/10.1021/jp403310a PMID: 23734591

21. Pires DEV, Ascher DB (2017) mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45: W241–W246. https://doi.org/10.1093/nar/gkx236 PMID: 28383703

22. Peng Y, Sun L, Jia Z, Li L, Alexov E (2017) Predicting protein-DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver. Bioinformatics https://doi.org/10.1093/bioinformatics/btx698 PMID: 29091991

23. Hassan MS, Shaalan AA, Dessouky MI, Abdelnaiem AE, ElHefnawi M (2019) A review study: Computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. Gene 680: 20–33. https://doi.org/10.1016/j.gene.2018.09.028 PMID: 30240882

24. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, et al. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 34: D204–206. https://doi.org/10.1093/nar/gkj103 PMID: 16381846

25. Liu L, Xiong Y, Gao H, Wei DQ, Mitchell JC, et al. (2018) dbAMEPNI: a database of alanine mutagenic effects for protein-nucleic acid interactions. Database 2018.

26. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242. PMID: 10592235

27. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320: 369–387. https://doi.org/10.1016/S0022-2836(02)00442-4 PMID: 12079393

28. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14: 33–38, 27–38. PMID: 8744570

29. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102: 3586–3616. https://doi.org/10.1021/jp973084f PMID: 24889800

30. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, et al. (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26: 1781–1802. https://doi.org/10.1002/jcc.20289 PMID: 16222654

31. Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. Phys Rev A Gen Phys 31: 1695–1697. PMID: 9895674

32. Luscombe NM, Laskowski RA, Thornton JM (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res 29: 2860–2874. PMID: 11433033

33. Cukuroglu E, Engin HB, Gursoy A, Keskin O (2014) Hot spots in protein-protein interfaces: towards drug discovery. Prog Biophys Mol Biol 116: 165–173. https://doi.org/10.1016/j.pbiomolbio.2014.06.003 PMID: 24997383

34. Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM (2001) Protein-RNA interactions: a structural analysis. Nucleic Acids Res 29: 943–954. PMID: 11160927

35. Im W, Beglov D, Roux B (1998) Continuum solvation model: Computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. Computer Physics Communications 111: 59–75.

36. Brooks BR, Brooks CL 3rd, Mackerell AD Jr., Nilsson L, Petrella RJ, et al. (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30: 1545–1614. https://doi.org/10.1002/jcc.21287 PMID: 19444816

37. Mirny LA, Shakhnovich EI (1996) How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 264: 1164–1179. https://doi.org/10.1006/jmbi.1996.0704 PMID: 9000638

38. Wei Q, Dunbrack RL Jr. (2013) The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One 8: e67863. https://doi.org/10.1371/journal.pone.0067863 PMID: 23874456

39. Shoemaker BA, Zhang D, Tyagi M, Thangudu RR, Fong JH, et al. (2012) IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. Nucleic Acids Res 40: D834–840. https://doi.org/10.1093/nar/gkr997 PMID: 22102591