BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts

Peng Xiong, Chengxin Zhang, Wei Zheng and Yang Zhang

Department of Computational Medicine and Bioinformatics, *University of Michigan, Ann Arbor, MI 48109, USA*

Correspondence to Yang Zhang: *zhng@umich.edu*
http://dx.doi.org/10.1016/j.jmb.2016.11.022
Edited by Michael Sternberg

Abstract

Understanding how gene-level mutations affect the binding affinity of protein–protein interactions is a key issue of protein engineering. Due to the complexity of the problem, using physical force field to predict the mutation-induced binding free-energy change remains challenging. In this work, we present a renewed approach to calculate the impact of gene mutations on the binding affinity through the structure-based profiling of protein–protein interfaces, where the binding free-energy change ($\Delta\Delta G$) is counted as the logarithm of relative probability of mutant amino acids over wild-type ones in the interface alignment matrix; three pseudo-counts are introduced to alleviate the limit of the current interface library. Compared with a previous profile score that was based on the log-odds likelihood calculation, the correlation between predicted and experimental $\Delta\Delta G$ of single-site mutations is increased in this approach from 0.33 to 0.68. The structure-based profile score is found complementary to the physical potentials, where a linear combination of the profile score with the FoldX potential could increase the $\Delta\Delta G$ correlation from 0.46 to 0.74. It is also shown that the profile score is robust for counting the coupling effect of multiple individual mutations. For the mutations involving more than two mutation sites where the correlation between FoldX and experimental data vanishes, the profile-based calculation retains a strong correlation with the experimental measurements.

Introduction

Protein–protein interaction plays an essential role in many biological processes, ranging from immune defense to cellular communication [1]. The ability to rationally design protein mutants with improved binding affinity is important in developing protein inhibitor as therapeutic agents. However, predicting the binding affinity change upon amino acid mutations remains a challenge for physical force field due to the lack of accurate methods describing the interactions [2]. Another major obstacle for physics-based approach is the need of building full-atomic model for the mutant complex, where both backbone and side-chain conformation can change due to the mutations in the interface region. Since physical energy calculation is sensitive to the subtle changes of atomistic structure models, this greatly restricts its performance, especially when multiple mutant sites are involved, where the complex coupling effect from individual residues could further complicate the problem [3].

A useful approach to alleviate this limitation is the utilization of evolutionary interface structure profiles built from the multiple sequence alignments of analogous protein–protein interactions collected from known protein–protein interface databases [4]. The assumption behind the idea is simple, that is, amino acids with a higher degree of conservation in the structural and evolutionary analogies tend to have a higher binding affinity. Due to the fundamentally different principles that they are built on, the information from the structure profile is complementary to the physics-based energy terms. The profile score has also an advantage compared with physics-based energy terms in that the calculation is not sensitive to the accuracy of the complex structures of target proteins, which enables the utilization of low-resolution models from threading and docking

Please cite this article as: P. Xiong, et al., BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts, J. Mol. Biol. (2016), http://dx.doi.org/10.1016/j.jmb.2016.11.022

approaches [5,6] for mutation calculations. Meanwhile, the coupling of different mutation sites can make it challenging for the physical potential to predict the effect from multiple-point mutations, where the correlation may still hold in the profiling score of interfaces given the completeness of the statistics.

How to build a connection between the interface profiles and the protein binding affinity change upon mutations is, however, a question, especially given the limit number of known interactions in the current structure library [7,8]. One of the mostly used methods is the log-odd likelihood score [9] that is equivalent to the average of the substitution scales from the Dayhoff [10] or BLOSUM [11] matrix between target amino acid and all amino acids at the corresponding position in the multiple interface alignments; this approach was taken in the former version of BindProf [4]. Another approach is to calculate the binding free-energy change according to the statistical energy derived from the Boltzmann distribution. Since the number of collected sequences is often much lower than that required for stable Boltzmann statistics, here it is necessary to make corrections to the amino acid probability in the interface multiple structural alignment (iMSA) according to amino acid substitution matrix [12].

In this work, we tested both methods to predict the binding affinity change in an experimental database of mutant protein interactions, SKEMPI [13]. We found that the statistical energy has a significantly improved correlation with the experimental $\Delta\Delta G$ than the log-odds–based profiling approach. It also shows advantage compared to the state of the art physical potentials [14] in both nsSNP and multiple-point mutations. An online server and the standalone open-source program of the approach, called BindProfX, are freely available at http://zhanglab.ccmb.med.umich.edu/BindProfX/.

Results and Discussion

Experimental mutation datasets

Experimental protein–protein binding affinity data were derived from the SKEMPI database [13], which contains the experimentally measured binding affinity change of protein complexes upon amino acid mutations. To construct the structure profile, we collected a non-redundant set of the interface mutations from SKEPMPI, where the residues with the nearest atomic distance <5 Å to the opposite chain are identified as interface residues. Average $\Delta\Delta$G value is used when there are multiple entries for the same mutation. The final non-redundant interface mutation set contains entries for 114 protein complexes, where 1131 entries are single-point mutations, 195 are double-points mutations, and 76 are three or higher-order mutations. A list of the non-redundant mutations is available at http://zhanglab.ccmb.med.umich.edu/BindProfX/download/.

Statistical protein binding energy from interface analogy alignments

For a query protein–protein complex, its interface is structurally aligned to the interfaces in the non-redundant interface library (NIL) collected from the PIFACE library [7] (see Methods). The interface comparison is performed by the I-align program [15], where all interfaces with a high interface similarity score (IS score, see Methods) are used to construct an iMSA matrix. The binding free-energy change upon mutation is calculated by

$$\Delta\Delta G_{\text{evo}}(i) = -\lambda \ln \frac{P(A_{\text{Mut}}, i)}{P(A_{WT}, i)} = -\lambda \ln \frac{N_{\text{obs}}(A_{\text{Mut}}, i) + N_{\text{pesudo}}(A_{\text{Mut}}, i)}{N_{\text{obs}}(A_{\text{WT}}, i) + N_{\text{pseudo}}(A_{\text{WT}}, i)} \quad (1)$$

where $P(A_{\text{Mut}}, i)$ and $P(A_{\text{WT}}, i)$ are the possibility of mutant and wild-type amino acids, respectively, appearing at the ith position of iMSA. $N_{\text{obs}}(\text{Mut}, i)$ and $N_{\text{obs}}(\text{WT}, i)$ are the number of the corresponding amino acids observed in the iMSA matrix, where $N_{\text{pseudo}}(A, i)$ is the corresponding pseudo-count number introduced to offset the limit of statistics that will be discussed in detail in next section.

In Fig. 1a, we listed the correlation coefficient of the predicted and experimental binding free-energy changes ($\Delta\Delta G$) upon the single-point mutations in SKEPMPI. The experiments were performed at different IS score cutoffs (from 0.2 to 0.98). It was shown that the profile score has an inverted U-shaped curve versus the IS score cutoff. This is understandable because the iMSA with a high IS score cutoff may contain too few interface samples that can reduce the efficiency of the statistical counting, while a too low IS score cutoff should introduce false-positive interfaces into the matrix. The method achieves a reasonable score with correlation coefficient above 0.5 for all cutoffs in [0.45, 0.65]. The best correlation coefficient is achieved at IS score cutoff = 0.55 which has a $\Delta\Delta G$ correlation coefficient = 0.68.

We also listed in the figure the correlation coefficient data when we use the Gribskov log-odds profile score [9], that is, using the log-odds difference between the wild-type and mutant amino acid to calculate the free-energy changes, which has been adopted in the first version of BindProf [4]. The data show a similar inverted U-shaped curve versus different IS score cutoffs, with a platform in [0.3, 0.8]. However, the best performance, with a $\Delta\Delta G$ correlation coefficient of 0.33, is much lower than the pseudo-count-assisted Boltzmann probability calculations. The result is largely consistent with the results obtained by the

Please cite this article as: P. Xiong, et al., BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts, J. Mol. Biol. (2016), http://dx.doi.org/10.1016/j.jmb.2016.11.022

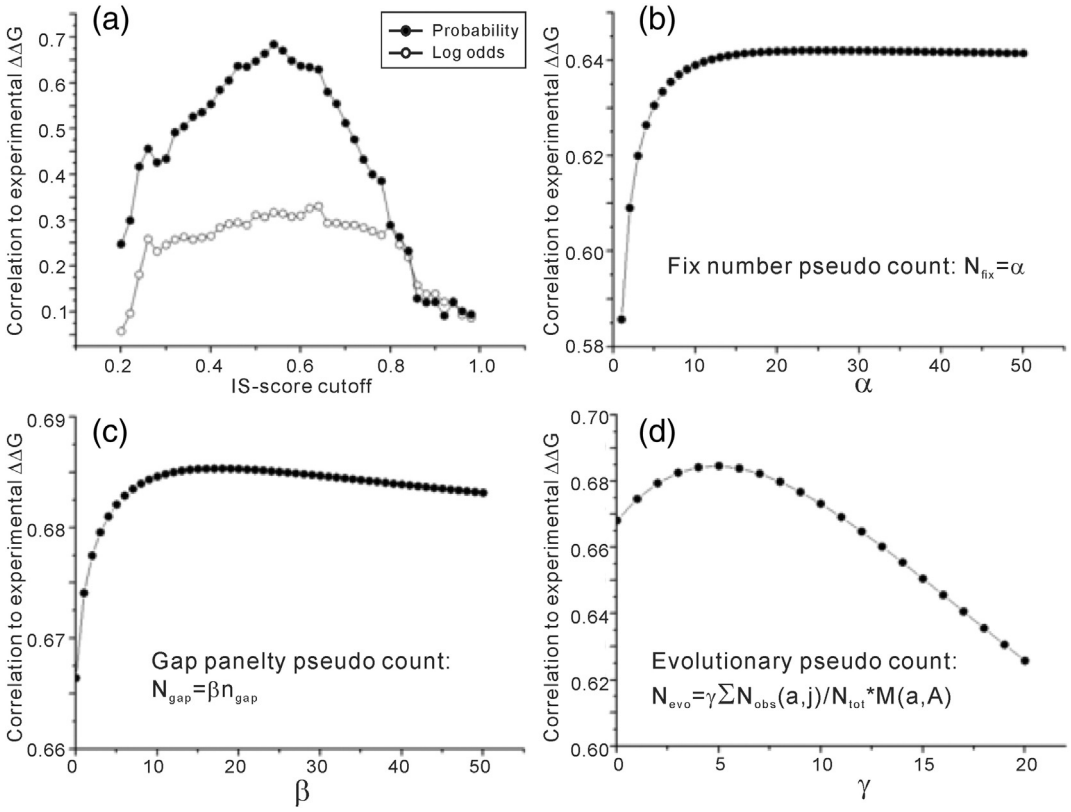

Fig. 1. Correlation coefficient between the predicted and experimental $\Delta\Delta G$ using difference scoring functions and parameters. (a) Profile score by BindProf and BindProfX at different IS score cutoffs. (b) Dependence of correlation coefficient on the fix number pseudo-count, where $\beta=\gamma=0$. (c) Dependence of correlation coefficient on the gap penalty pseudo-count, where $a=25, \gamma=0$. (d) Dependence of correlation coefficient on the evolutionary pseudo-count, where $a=25, \beta=15$.

BindProf program, although BindProf used a different interface template structure library from multiple-chain threading approach that consists of ~55,000 protein–protein complexes [16]. These data suggest that the use of Boltzmann probability with pseudo-counts could better associate the interface structural profiles with the binding free-energy change calculations.

Improving binding free-energy calculation with pseudo-counts

The accurate calculation of the mutation probability requests for an infinite number of structural analogies in the interface MSA. However, we could only find on average five structurally similar interfaces to the target with an IS score cutoff of 0.55. To construct an optimal binding free-energy model from structurally derived sequence information, we take the strategy of introducing pseudo-counts to Eq. (1) to partly alleviate the limit of the current structure library. The pseudo-counts consist of three parts, that is, $N_{pesudo}=N_{fix}+N_{gap}+N_{evo}$, where N_{fix}, N_{gap} and N_{evo} are the fix-number pseudo-count, gap penalty count and evolutional pseudo-count, respectively.

Fix number pseudo-count

The fix-number pseudo-count is a constant parameter, that is, $N_{fix}=\alpha$, added to the observation of both mutant and wild-type frequency. To test how this parameter affects the binding correlation results, we change the fixed pseudo-count number from 1 to 50, with the resultant correlation curve shown in Fig. 1b. When α increases from 1 to 15, the correlation coefficient rapidly increases from 0.586 to 0.641, and then keeps almost unchanged afterward. In our calculation, we take $\alpha=25$. Given the low number of interface analogies identified on average (~5), the data suggest that N_{fix} should be much larger than the number of observed amino acids in the interface alignments.

Pseudo-count of gap penalty

The second part in pseudo-count is used to decrease the absolute value of predicted $\Delta\Delta G$ for the positions with gap in the alignment, where $N_{gap}=\beta n_{gap}$ is proportional to the number of gaps (n_{gap}) at each position of iMSA. This count is introduced based on the observation that the prediction

Please cite this article as: P. Xiong, et al., BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts, J. Mol. Biol. (2016), http://dx.doi.org/10.1016/j.jmb.2016.11.022

accuracy of mutants with gapped alignment is much worse than mutants without gap; that is, the average correlation coefficient of predicted and experimental $\Delta\Delta G$ values is only 0.25 at the positions with gaps in the alignment, where it increases to 0.65 at the positions without gap. Physically, such dependence may be attributed to the fact that the structure of proteins at the poorly aligned positions has a higher variation and the structure profiles derived at these positions are therefore less reliable.

Although the introduction of pseudo-count could not increase the accuracy of $\Delta\Delta G$ prediction of mutants with gaps in alignment, it helps balance the contribution of the unreliably alignment regions to the total score calculation when combined with other binding score. As it is shown in Fig. 1c, this pseudo-count indeed slightly increases the overall correlation coefficient between predicted and experimental $\Delta\Delta G$ values, where an optimal parameter $\beta = 15$ was selected.

Evolutionary pseudo-count

The third part of pseudo-count is a sum of the counts of the amino acids evolutionally related to the observed amino acid in the iMSA, which is designed to offset the missing of the amino acid variations due to the limited number of interface analogies, i.e.

$$N_{\text{evo}}(A,i) = \gamma \sum_{a=1}^{20} \frac{N_{\text{obs}}(a,i)}{N_{\text{tot}}} M(a,A) \qquad (2)$$

where $N_{\text{obs}}(a,i)/N_{\text{tot}}$ is the relative frequency of amino acid a appearing at the ith position of iMSA. $M(a,A)$ is the interface probability transition matrix (iPTM) derived from the PIFACE homologous interface structures (see Methods and Table S1 in Supplementary Materials).

It is of interest to compare interface-based iPTM and the widely used BLOSUM-based PTM that was derived from the homologous sequence blocks [11] (Table S2). One of the most significant differences is the frequency of substitution between the polar and hydrophobic amino acids, where the frequency in the iPTM is higher than that in the BLOSUM PTM. For example, $T(\text{Glu}, \text{Leu})$ is 0.093 in the iPTM and 0.026 in the BLOSUM PTM, and $T(\text{Leu}, \text{Glu})$ is 0.042 in the iPTM and 0.022 in BLOSUM PTM. This probably reflects the enhanced interactions between the polar and non-polar residues in the interface regions.

Fig. 1d shows the dependence of the $\Delta\Delta G$ correlation on the evolutionary pseudo-counts, where adding a small number of evolutional pseudo-count could increase the correlation coefficient from 0.667 to 0.686 at $\gamma = 5$. Although the overall enhancement is relatively low, this pseudo-count is more helpful if the number of structural neighbors is low. For those targets with only one or two structurally similar interfaces, for example, the correlation coefficient increased from 0.207 to 0.323.

Result of cross-validations

The results in Fig. 1 were obtained using four parameters (IS score cutoff, α, β, γ) trained on the global samples, which may have a danger of over-fitting. To examine this issue, we made a protein-level cross validation on the SKEMPI dataset, in which the mutation samples were randomly divided into five groups according to the proteins that the mutations are associated with, where sequence identity between the proteins in any two groups is lower than 30% (a list of proteins in the five groups is displayed in Table S3). Next, we randomly select three groups as the training set to optimize the parameters and test the BindProfX potential on the rest two groups. Out of the 10 test experiments, the average correlation coefficient between the predicted and experimental $\Delta\Delta G$ is 0.663 for the training set and 0.625 for the testing set. The data show a slight dependence of the correlation on the training process because the correlation coefficient on the testing is slightly lower than the training set. However, the difference is statistically insignificantly, as the p value in Student's t test is 0.29.

As a control, we also performed a mutation-level cross-validation test based on random split of the mutation samples without applying the 30% sequence identity cutoff. The average correlation coefficient between the predicted and experimental $\Delta\Delta G$ is 0.662 for training and 0.657 for testing sets. As expected, the difference between the training and test becomes smaller in the mutation-level cross-validation compared to the protein-level ones, probably due to the effect of homologous correlation between test and training samples, which is consistent with the observation made on the cross-validation of monomer-chain mutations [17,18].

Overall, the magnitudes of the correlation coefficients in the cross-validations are largely comparable to that of Fig. 1, suggesting the robustness of the BindProfX performance. Although the global parameter training can slightly improve the correlation, it does not count for the major contribution of the BindProfX performance.

Combination of profile score with physics based potentials

FoldX [14] is an empirical potential that combines multiple physics-based energy terms and has been widely used to calculate protein–protein binding energy. In the recent study [4], it was shown that FoldX outperforms all other physics-based potentials

Please cite this article as: P. Xiong, et al., BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts, J. Mol. Biol. (2016), http://dx.doi.org/10.1016/j.jmb.2016.11.022

in generating the $\Delta\Delta G$ prediction with the highest correlation with the experimental data. When using this potential to predict $\Delta\Delta G$, the energy terms and parameters were optimized to reproduce experimental $\Delta\Delta G$ upon single mutation. One assumption here is that the backbone structure will be kept unchanged after single amino acid substitution, with the mutant structure often built by optimizing the side chain conformation. To calculate the FoldX score, the crystal structures of experimental proteins are downloaded from SKEMPI Web site (http://life.bsc.es/pid/mutation_database/database.html), with water and ligand molecules removed from the structural files. RepairPDB function within FoldX is used to perform a quick optimization in native structures, and BuildModel function is then used to generate the structures of mutant complex. Finally, AnalyseComplex function is used to determine the interaction energy of all wild-type and mutant complexes.

Using the same single-point mutation dataset from SKEMPI, the correlation coefficient between the FoldX score and the experimental binding energy change $\Delta\Delta G$ is 0.457, with a root mean square error (RMSE) being 2.31 kcal/mol, compared to the profile score that has the correlation coefficient 0.675 and RMSE 1.82 kcal/mol. These two score terms are complementary, which are built on different principles. If we combine these two scores by a simple linear combination: $\Delta\Delta G_{comb} = 0.9 \times \Delta\Delta G_{evo} + 0.4 \times \Delta\Delta G_{foldx}$, a correlation coefficient 0.738 can be achieved for the single-point mutations, with the RMSE reduced to 1.70 kcal/mol. The relative weight (0.9/0.4) was decided according to the optimized RMSE of the randomly selected half of the single-mutation database from SKEMPI. Fig. 2 presents the correlation between experimental $\Delta\Delta G$ data and that predicted by the combined evolutionary and FoldX scores, for the 1131 single-point mutation samples in SKEMPI. We note that we have tried both versions of FoldX.v3 and FoldX.v4 and found that the old version achieved a slightly higher correlation coefficient (0.457) than the newer version (0.430) for the single-point mutations. Therefore, all the data presented in this study were based on FoldX.v3 unless mentioned otherwise.

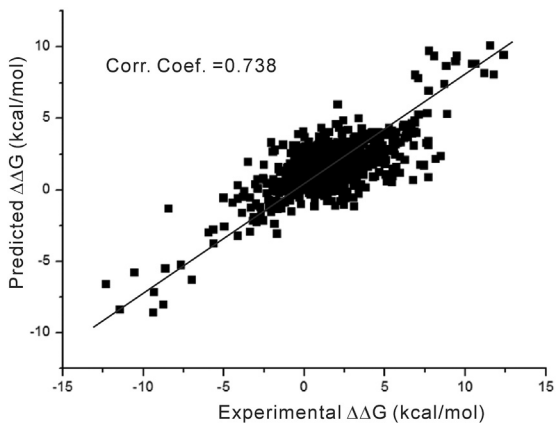

Fig. 2. Predicted versus experimental $\Delta\Delta G$ values using combined profile and FoldX score. Data contain all 1131 single-mutation samples taken from the SKEMPI database [13].

Free-energy changes on multiple mutations

The binding free-energy changes of mutations on different amino acids are usually non-additive due to the coupling effect between individual residues [3]; this makes the physics-based prediction of $\Delta\Delta G$ particularly difficult for the multiple-point mutations, because the coupling effect can be complicated, the strength of which depends on the separation of the mutated residues along the sequence and the relative spatial locations in the structure. Another difficulty to physics-based approaches is that the structural changes induced by multiple mutations are usually larger than that by single-point mutations, where the implicated fixed-backbone assumption is less likely held [14].

In Table 1, we list the correlation coefficient data of experimental and predicted $\Delta\Delta G$ on single and multiple mutations by different methods. Compared to the single-point mutations, the correlation by FoldX is reduced by 2.57 times on double mutations and almost vanishes for three and higher-order mutations (column 3). The correlation coefficient by the interface profile score is also reduced but the magnitude of reduction is much smaller; that is, the correlation decreases by 1.47 times on double mutation and 1.66 times on three and higher-order mutations, relative to the single mutations (column 6).

Interestingly, the correlation coefficient by the interface profile score is stronger for the multi-point mutations than that by the combined profile and FoldX scores (column 7). These data suggest that the profile score is more robust than the physics-based potentials in calculating the coupling effect of different amino acids in the multiple mutations. Accordingly, we use the combined score for single-point mutation, where only the profile score is utilized for treating multiple mutations in the BindProfX program (column 8).

Comparison of BindProfX with other statistical potentials

In addition to the physics-based approach of FoldX, we reported in Table 1 the results obtained from two statistical potentials: BeAtMuSiC [19] and Dcomplex [20]. In BeAtMuSiC, the binding free energy was calculated by a linear combination of multiple coarse-grained statistical terms with the weighting parameters determined by neural network training; the current BeAtMuSiC program allows only for

Please cite this article as: P. Xiong, et al., BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts, J. Mol. Biol. (2016), http://dx.doi.org/10.1016/j.jmb.2016.11.022

Table 1. Summary of correlation coefficients between predicted and experimental $\Delta\Delta G$ values by different scoring functions

N_{mut}[a]	N_{sam}[b]	C_{Fol}[c]	C_{BeA}[d]	C_{Dco}[e]	C_{Pro}[f]	C_{ProFol}[g]	C_{BinX}[h]
1	1131	0.457	0.272	0.056	0.675	0.738	0.738
2	195	0.178		−0.057	0.459	0.425	0.459
3 or more	76	0.046		0.015	0.406	0.319	0.406
All mixed	1402	0.445		0.183	0.651	0.658	0.691

[a] N_{mut}: number of mutations involved in each sample.
[b] N_{sam}: number of mutation samples.
[c] C_{Fol}: correlation coefficient by FoldX.
[d] C_{BeA}: correlation coefficient by BeAtMuSic.
[e] C_{Dco}: correlation coefficient by Dcomplex.
[f] C_{Pro}: Correlation coefficient using Profile-score.
[g] C_{ProFol}: correlation coefficient by profile-score + FoldX.
[h] C_{BinX}: Correlation coefficient by BindProfX that uses combined score (profile-score + FoldX) for single-mutation and profile-score for multiple-mutations.

single-point mutation calculations. Dcomplex is a distance-specific contact potential trained on the monomer structures from the PDB [20]. The data in Table 1 show that both statistics-based potentials have a relatively lower correlation coefficient compared to the composite FoldX potential [14], which is consistent with the observation in the previous study [4]. The outperformance of BindProfX demonstrates again the advantage of the structure-profile based approach over both the statistics and physics-based potentials.

BindProfX on low-resolution complex structure models

The experiments tested above are all based on experimental complex structures. To examine the performance of BindProfX on low-resolution structures, we tested the profile score on an independent mutation set taken from the ZEMu dataset [21], where only complex reconstructed from the monomer structures solved in a unbound form used. Here, we only consider the mutations involved in dimer complexes with those involved in higher-order complexes excluded. We also found several mutations in the ZEMu dataset that have inconsistent amino acid type to that in the PDB structure (e.g., "Immunoglobulin FC/Fragment B of protein A" in 2jwd_3dz8_1fc2 has a single mutation "YC133W" in ZEMu, but residue 133 of chain C in the structure is actually W, not Y); these mutations have been excluded as well. A final set of 104 mutations from the ZEMu set is listed in Table S4.

To calculate the BindProfX score, we first constructed the complex structures by overlaying the unbound monomer structure to the template complex structure provided by ZEMu, using the TM-align structural alignment program [22]. The BindProfX program is then used to calculate $\Delta\Delta G$ values based on the interface analogous alignments searched from the NIL database. As shown in Table S4, BindProfX achieves an overall correlation of 0.454 to the experimental values, which is 3.8 times higher than that of ZEMu (0.118), which was based on FoldX potential [14] using the complex structure refined from MacroMoleculeBuilder (MMB) simulations after the structural-alignment overlay [21].

We note that the complex structures used by BindProfX were built by a simple structure-alignment overlay without the MMB refinement. Given that the structural profile construction is not sensitive to the atomic details of the interface structures, the skip of the MMB refinement should not affect much of the BindProfX performance. The ZEMu $\Delta\Delta G$ values were directly taken from the Supplementary Information of Ref. [21], where the correlation coefficient we obtained (0.118) is lower than that reported in the ZEMu paper (0.34); this is probably due to the fact that only a subset of the ZEMu dataset was used here, as the mutations from higher-order complex structures and those with inconsistent amino acid types have been excluded from our calculation.

Conclusion

We developed a renewed algorithm, BindProfX, to assess the binding free-energy changes of protein–protein interactions induced by single- or multiple-point mutations. Different from the statistics and physics-based approaches that are based on atomistic interactions between protein structures, BindProfX is built on the structural and evolutionary profile analyses that are derived from the interface analogies in the PDB structure databases. A large-scale benchmark test on mutation samples shows that the BindProfX generates $\Delta\Delta G$ with a stronger correlation with the experimental data than both the statistical and physical potentials.

Because the approach relies more on the global interface structure comparison rather than the subtle atomic-details of the complex structures, the BindProfX prediction has the potential to be used for the cases

Please cite this article as: P. Xiong, et al., BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts, J. Mol. Biol. (2016), http://dx.doi.org/10.1016/j.jmb.2016.11.022

with low-resolution complex structure built by protein docking [5,23] and multi-chain threading approaches [6,24]. On a test set of 104 mutations with only unbound monomer structures available [21], BindProfX uses the complex structure docked from structural overlay of monomers onto the homologous complexes and achieved a correlation 0.454 to the experiment that is 3.8 times higher than that calculated from the physics-based potentials. The study also demonstrated the robustness of BindProfX in predicting multiple-point mutations, in which the correlation from statistics and physics-based potentials almost vanishes but the profiling score retains a strong correlation coefficient (0.406) to the experimental $\Delta\Delta G$ data in the three and higher-order mutations.

Compared to an earlier version of BindProf [4], the major difference is in the derivation of the interface profile scoring function from the interface MSA. In BindProf, the profile was calculated based on the log-odds likelihood score [9], whereas the approach in this study calculated the binding free-energy as the logarithm of relative Boltzmann possibility of mutant and wild-type amino acids. Three pseudo-counts, including fix-number, gap penalty and evolutionary composition, were introduced to offset the limited counting of interface analogies in the current structure databases. These pseudo-counts were found to be particularly useful for the cases that have less than 3 homologous complexes identified.

Mathematically, the log-odds likelihood score counts only for the similarity between the target amino acids and a set of amino acids at the interested position by the average BLOSUM substitution scale, while the relative frequency of the target amino acids appearing at the position, which reflects the propensity of the amino acids among different binding sites, is missed. The BindProfX score measures both the probability of the target amino acids at the binding site and the effect of amino acid substitution with each other, with the latter counted by the evolutionary pseudo-count (N_{evo}) in Eq. (2). In fact, if we increase γ to make N_{evo} much larger than the observed amino acid number (N_{obs}), the result of BindProfX would become almost the same as that by the log-odds likelihood score. Similarly, if we reduced γ to 0, that is, ignoring the similarity between amino acids, the correlation will also decrease as shown in Fig. 1d. Thus, although the pseudo-counts were originally introduced for offsetting the limit of data samples, they provide a balanced count of different sources of information from frequency and mutation, the former of which has been missed in the log-odds likelihood scoring.

The large-scale benchmark data showed that the new approach could significantly enhance the correlation of calculated $\Delta\Delta G$ with the experimental data by almost doubling the correlation coefficient value provided by the profile score from BindProf. The second advantage of the approach is that there are only very few free parameters, including the weights combining physics-based potentials, whereas in the BindProf program a neural-network training was developed to combine different energy terms, which can be in the danger of over-training on the test and training samples [4].

Among many potential uses of the PPI mutation predictions, including, for example, function analyses of protein networks and disease diagnosis [2,25,26], the one particularly interesting to us is the possibility to switch and/or gain new protein–protein interactions through the redesign of nature protein sequences. The application of the BindProfX scoring approach to guide the evolutionary-based protein design [27] on protein–protein interactions is under progress.

Methods

NIL

Our NIL is derived from the PIFACE library [7] that consists of 130,209 protein dimer interfaces extracted from the PDB. These interfaces were first classified into 22,604 clusters according to interface structural similarity, where a set of non-redundant interfaces is then collected from each cluster with a sequence identity cutoff 50%. The final NIL contains 24,962 dimer interfaces from all the clusters, which can be downloaded at http://zhanglab.ccmb.med.umich.edu/BindProfX/download/.

Interface structure comparison and alignment

Interface alignment and structural analog search are performed by the I-align program [15], which is built on IS score:

$$\text{IS-score} = \frac{S + s_0}{1 + s_0} \tag{4}$$

where $S = \frac{1}{L_Q}\sum_{i=1}^{N_a} \frac{f_i}{1+(d_i/d_0)^2}$ is the raw interface similarity score, and $s_0 = 0.18 - \frac{0.35}{L_Q^{0.3}}$ is a scaling factor to normalize the interface size. In the raw score S, L_Q is the average number of interface residues, N_a is the number of aligned interface residues, f_i is the fraction of conserved interface contacts at ith aligned position, and d_i/d_0 is the normalized Cα distance at the ith aligned position. When running I-align, non-sequential alignment is allowed, that is, the alignment of interfacial residues does not need to follow their sequential order.

iPTM

A hierarchical procedure is developed to construct the pool of interface pairs that was used to create the iPTM. First, from the PIFACE library [7], 1083

Please cite this article as: P. Xiong, et al., BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts, J. Mol. Biol. (2016), http://dx.doi.org/10.1016/j.jmb.2016.11.022

representative dimer structures were selected from the 1083 dimer clusters that have >20 members. I-align is then used to compare the 1083 representative dimers with the 24,962 dimers in the NIL, resulting in 1083 new dimer clusters with a cutoff of IS score >0.4. An all-to-all sequence and interface alignment is then conducted on all the dimers in each of the new dimer clusters. Only those dimer pairs with an IS score >0.5 and sequence identity <0.7 were selected, which resulted in 40,299 aligned dimer pairs. Considering a Cα distance cutoff <1 Å, we obtained 1.16 million of aligned residue pairs from the 40,299 aligned dimer pairs.

The iPTM is derived by

$$M(A,B) = \frac{p(A,B)}{q(A)} \quad (5)$$

where $p(A,B) = [N(A,B) + N(B,A)]/2N_{\text{pair}}$ and $q(A) = N(A)/N_{\text{res}}$ • $N(A,B)$ is the number of residue pairs with amino acid type A and B, N_{pair} (=1.16 million) is the total number of residue pairs, $N(A)$ is the number of residues with type A, N_{res} ($=2^* N_{\text{pair}}$) is the total number of residues in the selected interface pool.

Acknowledgments

The work was supported in part by the National Institute of General Medical Sciences (GM083107, GM116960) and the National Science Foundation (DBI1564756).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2016.11.022.

Received 17 August 2016;
Received in revised form 22 November 2016;
Accepted 23 November 2016
Available online xxxx

Keywords:
protein–protein binding interaction;
non-synonymous single nucleotide polymorphisms;
multiple-point mutations;
interface structure alignment;
profile score

Abbreviations used:
ΔΔ*G*, binding free-energy change; NIL, non-redundant interface library; IS score, interface similarity score; iPTM, interface probability transition matrix; iMSA, interface multiple structural alignment; RMSE, root mean square error.

References

[1] H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41–42.
[2] C.M. Yates, M.J. Sternberg, The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein–protein interactions, J. Mol. Biol. 425 (2013) 3949–3963.
[3] A. Horovitz, Double-mutant cycles: a powerful tool for analyzing protein structure and function, Fold. Des. 1 (1996) R121–R126.
[4] J.R. Brender, Y. Zhang, Predicting the effect of mutations on protein–protein binding interactions through structure-based Interface profiles, PLoS Comput. Biol. 11 (2015), e1004494.
[5] M.F. Lensink, S.J. Wodak, Docking, scoring, and affinity prediction in CAPRI, Proteins 81 (2013) 2082–2095.
[6] A. Szilagyi, Y. Zhang, Template-based structure prediction of protein–protein interactions, Curr. Opin. Struct. Biol. 24 (2014) 10–23.
[7] E. Cukuroglu, A. Gursoy, R. Nussinov, O. Keskin, Non-redundant unique interface structures as templates for modeling protein interactions, PLoS One 9 (2014), e86738.
[8] L. Garma, S. Mukherjee, P. Mitra, Y. Zhang, How many protein–protein interactions types exist in nature? PLoS One 7 (2012), e38913.
[9] M. Gribskov, A.D. McLachlan, D. Eisenberg, Profile analysis: detection of distantly related proteins, Proc. Natl. Acad. Sci. U. S. A. 84 (1987) 4355–4358.
[10] M.O. Dayhoff, R.M. Schartz, B.C. Orcutt, in: M.O. Dayhoff (Ed.), A Model of Evolutionary Change in Proteins, Natl. Biomed. Res. Found., Washington, DC 1978, pp. 353–358.
[11] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 10915–10919.
[12] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[13] I.H. Moal, J. Fernandez-Recio, SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models, Bioinformatics 28 (2012) 2600–2607.
[14] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, L. Serrano, The FoldX web server: an online force field, Nucleic Acids Res. 33 (2005) W382–W388.
[15] M. Gao, J. Skolnick, iAlign: a method for the structural comparison of protein–protein interfaces, Bioinformatics 26 (2010) 2259–2265.
[16] S. Mukherjee, Y. Zhang, Protein–protein complex structure predictions by multimeric threading and template recombination, Structure 19 (2011) 955–966.
[17] L. Quan, Q. Lv, Y. Zhang, STRUM: structure-based prediction of protein stability changes upon single-point mutation, Bioinformatics 32 (2016) 2936–2946.
[18] D.E. Pires, D.B. Ascher, T.L. Blundell, mCSM: predicting the effects of mutations in proteins using graph-based signatures, Bioinformatics 30 (2014) 335–342.
[19] Y. Dehouck, J.M. Kwasigroch, M. Rooman, D. Gilis, BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations, Nucleic Acids Res. 41 (2013) W333–W339.
[20] S. Liu, C. Zhang, H.Y. Zhou, Y.Q. Zhou, A physical reference state unifies the structure-derived potential of mean force for protein folding and binding, Proteins Struct. Funct. Bioinf. 56 (2004) 93–101.

[21] D.F. Dourado, S.C. Flores, Modeling and fitting protein–protein complexes to predict change of binding energy, Sci. Rep. 6 (2016) 25406.
[22] Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Res. 33 (2005) 2302–2309.
[23] J. Janin, Assessing predictions of protein–protein interaction: the CAPRI experiment, Protein Sci. 14 (2005) 278–283.
[24] A. Guerler, B. Govindarajoo, Y. Zhang, Mapping monomeric threading to protein–protein structure prediction, J. Chem. Inf. Model. 53 (2013) 717–725.
[25] M. Gao, H. Zhou, J. Skolnick, Insights into disease-associated mutations in the human proteome through protein structural analysis, Structure 23 (2015) 1362–1369.
[26] R. Elber, From an SNP to a disease: a comprehensive statistical analysis, Structure 23 (2015) 1155.
[27] P. Mitra, D. Shultis, J.R. Brender, J. Czajka, D. Marsh, F. Gray, et al., An evolution-based approach to de novo protein design and case study on mycobacterium tuberculosis, PLoS Comput. Biol. 9 (2013), e1003298.

Please cite this article as: P. Xiong, et al., BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts, J. Mol. Biol. (2016), http://dx.doi.org/10.1016/j.jmb.2016.11.022