Abbasi *et al. BioData Mining* (2020) 13:20
https://doi.org/10.1186/s13040-020-00231-w

BioData Mining

RESEARCH Open Access

ISLAND: in-silico proteins binding affinity prediction using sequence information

Wajid Arshad Abbasi[1,2*], Adiba Yaseen[2], Fahad Ul Hassan[2], Saiqa Andleeb[3] and Fayyaz Ul Amir Afsar Minhas[4*]

* Correspondence: wajidarshad@gmail.com; fayyaz.minhas@warwick.ac.uk; fayyazafsar@gmail.com
[1]Computational Biology and Data Analysis Laboratory, Department of Computer Science and Information Technology, King Abdullah Campus, University of Azad Jammu & Kashmir, Muzaffarabad, Pakistan
[4]Department of Computer Science and the PathLAKE Consortium, University of Warwick, Coventry, UK
Full list of author information is available at the end of the article

Abstract

Background: Determining binding affinity in protein-protein interactions is important in the discovery and design of novel therapeutics and mutagenesis studies. Determination of binding affinity of proteins in the formation of protein complexes requires sophisticated, expensive and time-consuming experimentation which can be replaced with computational methods. Most computational prediction techniques require protein structures that limit their applicability to protein complexes with known structures. In this work, we explore sequence-based protein binding affinity prediction using machine learning.

Method: We have used protein sequence information instead of protein structures along with machine learning techniques to accurately predict the protein binding affinity.

Results: We present our findings that the true generalization performance of even the state-of-the-art sequence-only predictor is far from satisfactory and that the development of machine learning methods for binding affinity prediction with improved generalization performance is still an open problem. We have also proposed a sequence-based novel protein binding affinity predictor called ISLAND which gives better accuracy than existing methods over the same validation set as well as on external independent test dataset. A cloud-based webserver implementation of ISLAND and its python code are available at https://sites.google.com/view/wajidarshad/software.

Conclusion: This paper highlights the fact that the true generalization performance of even the state-of-the-art sequence-only predictor of binding affinity is far from satisfactory and that the development of effective and practical methods in this domain is still an open problem.

Keywords: Protein sequence analysis, Protein-protein interaction, Support vector machines, Web services, Binding affinity

Background

Protein binding affinity is a key factor in enabling protein interactions and defining structure-function relationships that drive biological processes [1]. Accurate measurement of binding affinity is crucial in understanding complex biochemical pathways and to uncover protein interaction networks. It is also measured as part of drug discovery

and design to improve drug specificity [2]. It can be measured in terms of the disassociation constant (K_d) through different experimental methods such as Nuclear magnetic resonance spectroscopy, gel-shift and pull-down assays, analytical ultracentrifugation, Surface Plasmon Resonance (SPR), spectroscopic assays, etc [3, 4]. However, the accuracy of these methods depends on dissociation rates and these methods cannot be applied at a large scale due to cost and time constraints [3, 5]. Therefore, accurate computational techniques can play an important role in the affinity determination of protein complexes.

Various computational methods for binding affinity prediction have been proposed based on free energy perturbation, empirical scoring, and force-field potentials [6–12]. These scoring function based methods are typically trained and evaluated on limited datasets. These methods fail to accurately predict binding affinities for diverse datasets [13].

Among computational binding affinity prediction methods, machine learning is preferred because of its implicit treatment of any relevant factors involved in protein-protein interactions (PPIs) and the flexibility of using empirical data instead of a fixed or predetermined function form [14]. A representation of the design and use of machine learning models for binding affinity prediction is given in Fig. 1. Machine learning based affinity prediction models require a dataset of diverse protein complexes with experimentally determined affinity values for training. By extracting the feature representation of protein complexes, a regression model is trained which can be used for affinity prediction of a novel complex (Fig. 1). A number of machine learning based studies for protein binding affinity prediction have been proposed in the literature [5, 15–19]. Most of these studies are based on a protein binding affinity benchmark dataset with 3-D structures of 144 protein complexes [20]. The affinity prediction models proposed by Moal et al., Tian et al., and Vangone and Bonvin in their studies are based on 3-D protein structures [5, 15, 16]. However, protein structures are not available for most protein complexes. Consequently, the sequence-based prediction of binding affinity is an important research problem.

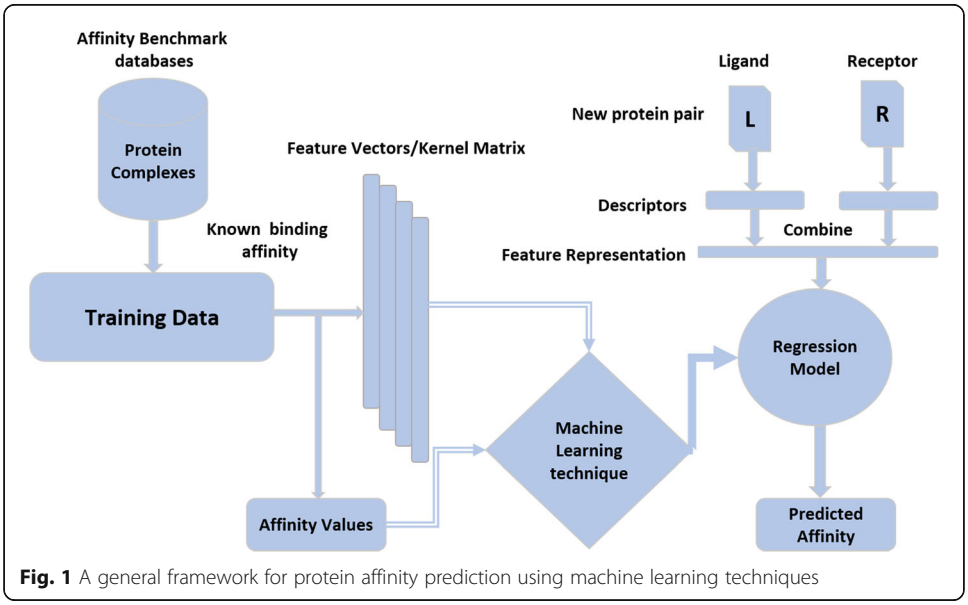

Fig. 1 A general framework for protein affinity prediction using machine learning techniques

Sequence-based binding affinity prediction is challenging because proteins interaction and binding affinity are dependent upon protein structures and functions.

Among sequence-based protein binding affinity prediction models using protein binding affinity benchmark dataset, the model proposed by Yugandhar and Gromiha (PPA-Pred2) is the state of the art absolute binding affinity predictor [17]. PPA-Pred2 claims high accuracy with a high correlation score between true and predicted binding affinity values [21]. However, their proposed model performed poorly on an external validation dataset [22]. Furthermore, their prediction errors are, surprisingly, lower than the reported deviation in experimental measurements of binding affinity values and the error rates of structure-based prediction techniques [20, 22]. Yugandhar and Gromiha have attributed this issue to the difference in experimental conditions and computational platforms [21]. In this work, we have replicated the validation of PPA-Prep2 on an external independent test dataset as performed by Moal *et. al.* [22]. Moreover, protein binding affinity prediction models proposed by Chen M, et al. and Srinivasulu YS, et al., had not been evaluated using any external validation datasets, and also these studies did not provide an interface to perform such a validation [18, 19]. These simple researches have highlighted the need to revisit sequence-based binding affinity prediction and develop novel predictors that can be used in a practical setting. To address this, we have proposed a new binding affinity prediction model called ISLAND (In SiLico protein AffiNity preDictor). Our proposed model uses sequence features alone and gives higher prediction accuracy than the PPA-Pred2 web server.

Methods

In this section, we have discussed in detail the methodology adopted to develop and evaluate the performance of sequence-based protein binding affinity predictors.

Datasets and preprocessing

We have used protein binding affinity benchmark dataset 2.0 for evaluation of PPA-Pred2 webserver and development of the proposed method ISLAND [20]. This dataset contains 144 non-redundant complexes of proteins for which both bound and unbound structures of the ligand and receptor proteins are available. Protein binding affinities are given in terms of binding free energy (ΔG) and disassociation constant (K_d). The binding free energy (ΔG) ranges from − 18.58 to − 4.29. Following the same data curation and preprocessing technique used by Yugandhar and Gromiha, we have selected 135 complexes from this dataset [17]. This allows us to have a direct comparison of our method with the one proposed by Yugandhar and Gromiha [17].

We have also used an external independent test dataset of 39 protein-protein complexes with known binding free energy (ΔG) to perform a stringent test of performance comparison between PPA-Pred2 and ISLAND. This dataset is derived from Chen et al. by removing complexes having more than two chains, involving chains of size less than 50 residues, and having an overlap with training data [23]. This dataset has also been used by Moal *et. al.* in their evaluation of binding affinity prediction techniques [22].

Evaluation of the PPA-Pred2 webserver

In order to investigate the accuracy of PPA-Pred2, we evaluated its performance on the selected dataset. For this purpose, we accessed PPA-Pred2 [17] through its webserver (URL: http://www.iitm.ac.in/bioinfo/PPA_Pred/) on 03-02-2017. This webserver takes amino acid sequences of ligand and receptor of a protein complex and returns predicted values of change in binding free energy (ΔG) and disassociation constant (K_d) [17]. The results obtained through this evaluation will also serve as a baseline in this study.

Sequence homology as affinity predictor

In order to confirm whether simple homology is enough to predict protein binding affinity accurately or not, we have developed a sequence homology-based protein binding affinity predictor as a baseline. For this purpose, we predicted the affinity value of a query protein complex based on the affinity value of its closest homolog in our dataset of protein complexes with known binding affinity values. We performed the Smith-Waterman alignment to determine the degree of homology between two protein complexes using BLOSUM-62 substitution matrix with gap opening and extension penalties of – 11 and – 1, respectively [24, 25].

Proposed methodology

We have developed a sequence-only regression model called ISLAND (In SiLico protein AffiNity preDictor), to predict absolute protein binding affinity values rather calssifying protein complexes into low and high affinity as in case of LUPI [26]. To develop ISLAND, we have used different regression methods, evaluation protocols, and sequence-based feature extraction techniques. The methodology adopted for the development of the ISLAND is detailed below.

Sequence-based features

In machine learning based prediction models, we require a feature representation of each example for training and testing (Fig. 1). Therefore, we have represented each complex in our dataset through a feature representation obtained from individual chains in the ligand (l) and receptor (r) of each complex. We used several explicit features and various kernel representations to model sequence-based attributes of protein complexes. We discuss the sequence-based features used in this study below.

Explicit features

Amino acid composition features (AAC) These features capture the occurrences of different amino acids in a protein sequence. It gives a 20-dimensional feature vector $\phi_{AAC}(s)$ of a given sequence s such that the $\phi_{AAC}(s)_k$ contains the number of times amino acid k occurs in s [27]. This feature representation has successfully been used to predict protein interactions, binding sites, and prion activity [27–29].

Average BLOSUM-62 features (Blosum) In contrast to AAC, this feature representation models the substitutions of physiochemically similar amino acids in a protein. In

this feature representation, protein sequence $\boldsymbol{s}$ is converted into a 20-dimensional feature vector by simply averaging the columns from the BLOSUM-62 substitution matrix corresponding to the amino acids in the given sequence. Mathematically, $\boldsymbol{\phi_{Blosum}(s)}=\frac{1}{|s|}\sum_{i=1}^{|s|}\boldsymbol{B_i}$, where $\boldsymbol{B_i}$ is the column of the BLOSUM-62 substitution matrix [24] corresponding to the ith residue in $\boldsymbol{s}$.

Propy features (propy) In order to capture the biophysical properties of amino acids and sequence-derived structural features of a given protein sequence, we used a feature extraction package called propy [30]. It gives a 1537-dimensional feature representation $\boldsymbol{\phi_{propy}(s)}$ of a given sequence $\boldsymbol{s}$. This representation includes pseudo-amino acid compositions (PseAAC), autocorrelation descriptors, sequence-order-coupling number, quasi-sequence-order descriptors, amino acid composition, transition and the distribution of various structural and physicochemical properties [31, 32].

Position specific scoring matrix features (PSSM) This feature representation models the evolutionary relationships between proteins. To get this representation, we used the Position Specific Scoring Matrix (PSSM) of a given protein sequence [33]. We obtained the PSSM for each protein chain in a complex by using PSI-BLAST for three iterations against the non-redundant (nr) protein database with an e-value threshold of $\mathbf{10^{-3}}$ [33, 34]. In this feature representation, we represent the protein sequence $\boldsymbol{s}$ by the average of columns in its PSSM. This results in a 20-dimensional feature vector $\boldsymbol{\phi_{PSSM}(s)}=\frac{1}{|s|}\sum_{i=1}^{|s|}\boldsymbol{F_i^s}$, where $\boldsymbol{F_i^s}$ is the column in the PSSM corresponding to the ith residue in s.

ProtParam features (ProtParam) In order to capture different physiochemical properties of a protein such as the molecular weight of the protein, aromaticity, instability index, isoelectric point, and secondary structure fractions, we have used ProParam ExPASy tools to get ProtParam representation [35–37]. This leads to a 7-dimensional feature representation $\boldsymbol{\phi_{ProtParam}(s)}$ of a given sequence $\boldsymbol{s}$.

Kernel representations

In addition to using explicit protein sequence features in our machine learning models for binding affinity prediction, we have also experimented with different sequence-based kernel [38, 39]. Kernel methods present an alternate way of sequence representation by modeling the degree of similarity between protein sequences instead of an explicit feature representation [38]. Kernel-based methods such as support vector machines and support vector regression can make use of these kernel function scores in their training and testing [40]. Different sequence kernels used in this work are described below. Each of these kernels $\boldsymbol{k(a,b)}$ can be interpreted as a function that measures the degree of similarity between sequences $\boldsymbol{a}$ and $\boldsymbol{b}$.

Smith-Waterman alignment kernel (SW kernel) We have used the Smith-Waterman alignment algorithm for determining the degree of similarity between two protein sequences [25]. The Smith-Waterman kernel $\boldsymbol{k_{SW}(a,b)}$ is simply the alignment score obtained from the Smith-Waterman local alignment algorithm using BLOSUM-62

substitution matrix with gap opening and extension penalties of − 11 and − 1, respectively. It is important to note that this kernel may not satisfy the Mercer's conditions as the eigen values of the kernel matrix may be negative [41]. We addressed this issue by subtracting the most negative eigenvalue of the original kernel matrix from its diagonal elements [42]. From a theoretical point of view, this kernel can be interpreted as the optimal local alignment score of the two sequences [42]. Mathematically, the Smith-Waterman alignment score $\boldsymbol{k_{SW}(a,b)}$ between sequences, $\boldsymbol{a}$ and $\boldsymbol{b}$ can be written as follows [42].

$$\boldsymbol{k_{SW}(a,b) = max_{\pi\in\Pi(l,r)} p(a,b,\pi)} \tag{1}$$

Here, $\boldsymbol{\Pi(a,b)}$ denote the set of all possible local alignments between $\boldsymbol{a}$ and $\boldsymbol{b}$, and $\boldsymbol{p(a,b,\pi)}$ represents the score of the local alignment $\boldsymbol{\pi\epsilon\Pi(a,b)}$ between $\boldsymbol{a}$ and $\boldsymbol{b}$.

Local alignment kernel (LA kernel) Local alignment kernel is useful for comparing sequences of different lengths that share common parts [40, 42]. In contrast to the Smith-Waterman alignment kernel which considers only the optimal alignment, this kernel sums up contributions of all the possible local alignments of input sequences. Mathematically, the local alignment score $k_{LA}(a,b)$ between sequences, a and b can be written as follows [42].

$$k^{\beta}_{LA}(a,b) = \sum\nolimits_{\pi\in\Pi(a,b)} \exp(\beta p(a,b,\pi)) \tag{2}$$

Here in Eq. (2), $\beta \geq 0$ is a parameter that controls the sensitivity of the LA kernel. For larger values of β score of LA kernel approaches SW kernel score [42]. We have used $\beta = 0.1$ based on empirical performance.

Mismatch kernel (MM kernel) The mismatch kernel captures the degree of overlap between subsequences of the two sequences while allowing mismatches [43]. MM kernel $k^{k,m}_{MM}(a,b)$ gives the number of subsequences of length k that are present in both the input sequences a and b with a maximum of m mismatches. Ranges for the values of k and m are 3 – 9 and 0 – 5, respectively. We have used $k = 5$ and $m = 3$ based on empirical performance.

Complex level features representation

We need to predict protein binding affinity at the complex level. Since we have extracted features at the chain level, therefore, we require a mechanism to obtain a complex level feature representation from individual chains. The basic mechanism of combining individual chain level feature representation from each ligand and receptor to form a complex level representation is shown in Fig. 2. Complex level representation is obtained for explicit features by concatenation of chain level features and for kernels by adding kernels over the constituent chains of a complex.

Feature concatenation

In our machine learning model, a complex c is represented by the tuple $c \equiv ((l,r),y)$, where (l,r) is the pair of ligand and receptor proteins in the complex and y is the corresponding affinity value. To generate the complex level feature representation $\boldsymbol{\psi}(c)$, we

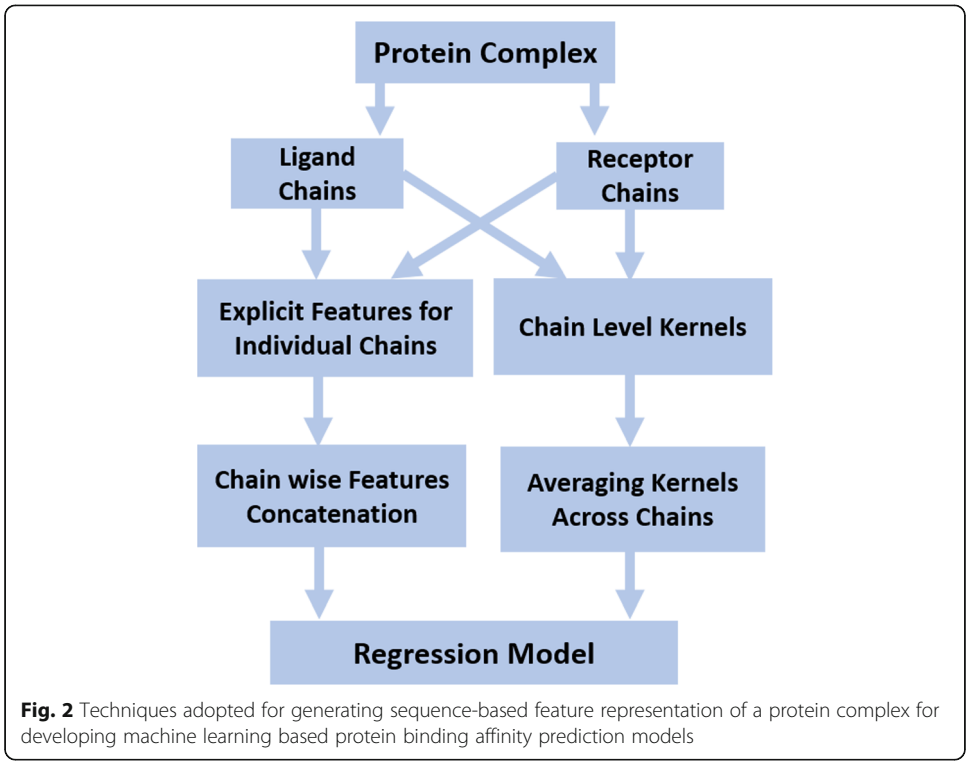

Fig. 2 Techniques adopted for generating sequence-based feature representation of a protein complex for developing machine learning based protein binding affinity prediction models

simple concatenate the feature representations of respective ligand and receptor as $\boldsymbol{\psi}(c) = \begin{bmatrix} \boldsymbol{\psi}_{Avg}(l) \\ \boldsymbol{\psi}_{Avg}(r) \end{bmatrix}$. Here, $\boldsymbol{\psi}_{Avg}(l) = \frac{1}{|l|}\sum_{q\in l}\boldsymbol{\phi}(q)$ and $\boldsymbol{\psi}_{Avg}(r) = \frac{1}{|r|}\sum_{q\in r}\boldsymbol{\phi}(q)$ are the explicit feature representations averaged across all the chains present in the ligand and receptor proteins, respectively. This method of feature representation generation has already been used for protein interacting residues predictor [44].

Combining kernels

To make predictions at the complex level from sequence-based kernels, we have developed a complex-level kernel by simply averaging the kernel function values of individual chains from the two complexes [38]. Mathematically, the kernel over complexes c and $c^{'}$ is given by $K(c, c^{'}) = \frac{1}{|c|\times|c^{'}|}\sum_{q\in c, q^{'}\in c^{'}} k(q, q^{'})$, where $k(q, q^{'})$ is the chain level kernel over two chains from the two complexes.

Regression models

Here, we begin by presenting the binding affinity prediction problem as a regression problem. In machine learning based affinity prediction, a dataset consisting of N examples (c_i, y_i), where $i = 1...N$. In this representation, c_i is a complex with known binding affinity y_i. The feature representation of c_i is $\boldsymbol{\psi}(c_i)$. Our objective in machine learning based regression is to train a model $f(c)$ that can predict the binding affinity of the complex c. The learned regression function $f(\cdot)$ should generalize well over previously unseen complexes. We used the following regression techniques through Scikit-learn to

get different regression models [45]. It is also important to note that the feature representations are normalized to have unit norm and standardized to zero mean and unit standard deviation before using them in the regression model.

Ordinary least-squares regression (OLSR)

Ordinary least squares (OLS) estimates the regression function $f(c) = \boldsymbol{w}^T\boldsymbol{\psi}(c) + b$ by minimizing the sum of squared error between the actual and predicted affinity values $\min_{\boldsymbol{w},b} \sum_{i}^{N} (y_i - f(\boldsymbol{c}_i))^2$ [46]. Here, $\boldsymbol{w}$ and b are parameters to be learned. This technique has been used previously for protein binding affinity prediction [17]. We have used this technique as a baseline in our study.

Support vector regression (SVR)

Support Vector Machines have been effectively used to solve different computational problems in bioinformatics [47]. Support Vector Regression (SVR) performs regression using ε-insensitive loss and, by controlling model complexity [48]. Training a SVR for protein binding affinity prediction involves optimizing the objective function given in Eq. (3) to learn a regression function $f(c) = \boldsymbol{w}^T\boldsymbol{\psi}(c) + b$.

$$min_{w,b}\frac{1}{2}\|\boldsymbol{w}\|^2 + \mathrm{C}\sum_{i=1}^{N}\left(\xi_i^+ + \xi_i^-\right)$$

$$Such\ that\ for\ all\ i: \begin{cases} y_i - f(\boldsymbol{c}_i) \le \varepsilon + \xi_i^+ \\ f(\boldsymbol{c}_i) - y_i \le \varepsilon + \xi_i^- \\ \xi_i^+, \xi_i^- \ge 0 \end{cases} \tag{3}$$

Here, $\frac{1}{2}\|\boldsymbol{w}\|^2$ controls the margin, ξ_i^+ and ξ_i^- capture the extent of margin violation for a given training example and C is the penalty of such violations [47]. We used both linear and radial basis function (rbf) SVR in this study. The values of C, gamma, and epsilon were optimized during model selection. SVR has already been used for the same purpose in previous studies [17].

Random Forest regression (RFR)

Random Forest regression (RFR) is an ensemble of regression trees used for nonlinear regression [49]. Each regression tree in the RF is based on randomly sampled subsets of input features. We optimized RF with respect to the number of decision trees and a minimum number of samples required to split in this study using grid search. This regression technique has been used in many related studies [15, 50, 51].

Model evaluation and performance assessment

To evaluate the performance of all the trained regression models, we have used Leave One Complex Out (LOCO) cross-validation (CV) [52]. In LOCO, a regression model is developed with $(N-1)$ complexes and tested on the left out complex. This process is repeated for all the N complexes present in the dataset. We used Root Mean Squared Error $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(y_i - f(c_i))^2}$ and Pearson correlation coefficient (P_r) between the predicted $f(c_i)$ and actual y_i, as performance measures for model evaluation and

performance assessment. To check the statistical significance of the results, we have also estimated the *P*-value of the correlation coefficient scores. We used grid search over training data to find the optimal values of hyper-parameters of different regression models.

Webserver

We have deployed ISLAND as a webserver that takes a pair of protein sequences in plain text and predicts their binding affinity. After the successful submission of protein sequences, the result page shows predicted binding affinity along with disassociation constant (K_d). A Python implementation of the proposed method together with a webserver is available at http://faculty.pieas.edu.pk/fayyaz/software.html# island.

Results and discussion

In this section, we discuss the results and give details of the major outcomes of our study.

Binding affinity prediction through sequence homology

As a baseline, we have obtained the predicted affinity values of all 135 complexes in our dataset using a sequence homology-based affinity prediction method. The Pearson correlation coefficient (P_r) between predicted and experimental values of ΔG is 0.29 with a Root Mean Squared Error (*RMSE*) of 3.20. These results with poor correlation and high RMSE value show that the sequence homology only cannot be effectively used to predict the binding affinity of the protein complexes. As discussed in the next section, our machine learning based method performs significantly better than homology-based predictions.

Binding affinity prediction through ISLAND

We have evaluated the performance of three different regression models (OLSR, RFR, and SVR) along with eight different types of sequence descriptors with LOCO cross-validation over the docking benchmark dataset. The results of this analysis are shown in Table 1 in the form of Root Mean Squared Error (*RMSE*) and Pearson correlation coefficient (P_r) along with statistical significance (*P*-value). With explicit features, we obtained a maximum correlation of 0.41 with RMSE = 2.60 between predicted and experimental values of ΔG using propy through SVR (Table 1). While using kernel descriptors, we obtained a maximum correlation of 0.44 with an RMSE = 2.56 between predicted and experimental ΔG values using the local alignment kernel (see in Table 1). We have achieved the best performance through local kernel across all sequence descriptors used in this study as shown in Table 1. Moreover, LA kernel performs better than SW kernel because of considering the effect of all the local alignments rather taking the best alignment as in the case of SW kernel. The RMSE value of ISLAND predictions is quite close to the range of experimental uncertainties (1–2 kcal/mol) as reported by Kastritis et al. [20]. Our proposed method outperforms the previous sequence-based method proposed by Srinivasulu YS, et al., with a reported correlation coefficient of 0.34 through Jackknife cross validation [19]. Another protein sequence-based method involving deep learning proposed by Chen M, et al., reported a higher

Table 1 Performance of regression models trained on the range of protein sequence descriptors using loco cross validation

Feature Descriptors	Regression Models								
	OLSR			RFR			SVR		
	P_r	*P*-value	RMSE	P_r	*P*-value	RMSE	P_r	*P*-value	RMSE
AAC	0.20	1.5×10^{-2}	3.19	0.40	6.4×10^{-7}	2.66	0.40	1.0×10^{-6}	2.69
Blosum	0.20	1.4×10^{-2}	3.10	0.37	2.8×10^{-7}	2.71	0.39	1.5×10^{-5}	2.67
propy	0.14	1.3×10^{-1}	3.67	0.39	3.0×10^{-3}	2.64	0.41	1.1×10^{-6}	2.60
PSSM	0.19	7.2×10^{-1}	3.68	0.38	1.1×10^{-5}	2.67	0.37	1.5×10^{-5}	2.66
ProtParam	0.25	3.0×10^{-3}	2.82	0.34	4.7×10^{-5}	2.72	0.37	9.4×10^{-6}	2.64
SW kernel	results not applicable						0.37	2.1×10^{-6}	2.63
LA kernel							**0.44**	$\mathbf{1.2 \times 10^{-8}}$	**2.56**
MM kernel							0.38	7.1×10^{-6}	2.66

accuracy with a correlation coefficient of 0.873 using 10-fold cross-validation and SKEMPI dataset [18, 53]. However, the cross-validation scheme adopted by Chen M, et al., may not conform to the underlying problem as SKEMPI dataset involves more than one mutant proteins of a single protein complex [18, 52, 53].

The performance of the ISLAND is also comparable with the methods based on 3-D protein structures such as DFIRE ($P_r = 0.35$), PMF ($P_r = 0.37$), RBF ($P_r = 0.44$), M5' ($P_r = 0.45$) and RF ($P_r = 0.48$) as reported by Moal et al. [15]. Despite getting the comparable performance of ISLAND with structure-based methods, there is still a lot of room for improvement in affinity prediction from sequence information alone.

Comparison using external independent test dataset

We obtained the predicted binding affinity values for all the complexes in our external validation dataset using both PPA-Pred2 and ISLAND. We have seen a significant performance improvement of the ISLAND in terms of RMSE between predicted and experimental ΔG values. We obtained an RMSE of 1.98 with ISLA ND whereas PPA_Pred2 gives us an RMSE of 4.78. We have also seen a significant performance improvement of both the methods in terms of Pearson correlation coefficient and absolute error with values 0.35, 1.52 and 0.05, 2.63 through ISLAND and PPA_Pred2, respectively. We have also shown a comparison between ISLAND and PPA-Pred2 in terms of absolute error between predicted and actual binding affinity values of all the complexes in our validation set in Fig. 3. The binding affinity of > 60% complexes were predicted within an absolute error of 1.5 kcal/mol using ISLAND, whereas, through PPA-Pred2 absolute error for these complexes is above 2.5 kcal/mol (see in Fig. 3). These results show better performance of our proposed method for binding affinity prediction of proteins in a complex in comparison to PPA-Pred2. These performance improvements of ISLAND over PPA-Pred2 are based on a proper model selection with parameters tuned using grid search and better feature engineering by using different kernels. Moreover, these results also support the criticism of Moal *et. at.*, on PPA-Pred2 and suggest a need

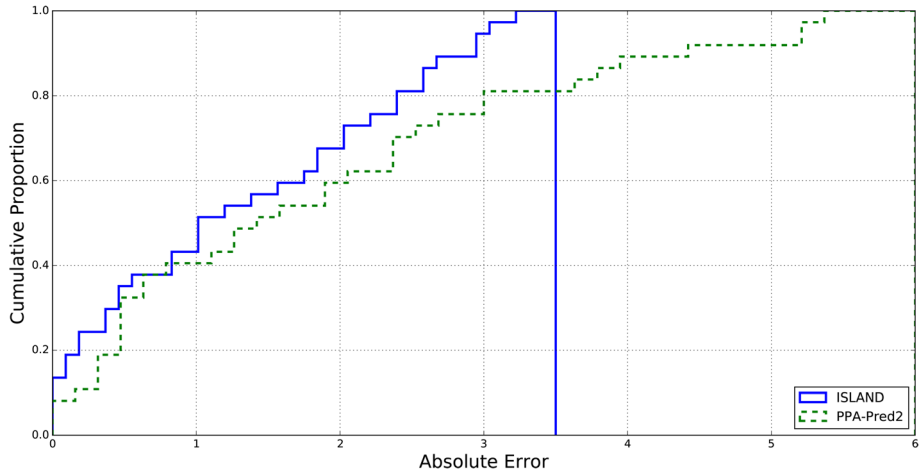

Fig. 3 Cumulative histogram of absolute error between actual and predicted binding affinity values through ISLAND and PPA-Pred2 on external independent validation dataset

for further work on methods of protein binding affinity prediction using sequence information [22].

Conclusions

This paper highlights the fact that the true generalization performance of even the state-of-the-art sequence-only predictor of binding affinity is far from satisfactory and that the development of effective and practical methods in this domain is still an open problem. As already suggested in recent studies by Dias & Kolaczkowski and Abbasi et al., to achieve better performance in this domain, we need either a significant increase in the amount of quality affinity data or methods of leveraging data from similar problems [26] [54]. We also propose a novel sequence-only predictor of binding affinity called ISLAND which gives better accuracy than PPA-Pred2 webserver and other existing methods over the same external independent test set.

Acknowledgments

The authors are thankful to K. Yugandhar and M. Michael Gromiha, Indian Institute of Technology Madras, India for providing relevant data for this study. We would also like to thank Dr. Hanif Durad and Dr. Javaid Khurshid, DCIS, PIEAS, Pakistan for helping us meet the computational requirements of the project. We also acknowledge the very fruitful discussions with Dr. Asa Ben-Hur, Colorado State University, Fort Collins, USA over the course of this project.

Authors' contributions

WAA developed the scientific workflow, performed the experiments, analyzed and interpreted the results, and was a major contributor in manuscript writing. FUH, AY and SA contributed to the analysis and interpretation of the results and writing of the manuscript. FuAAM conceived the idea, supervised the study and helped in manuscript writing. All authors have read and approved the final manuscript.

Funding

Wajid A. Abbasi is supported by a grant (PIN: 213–58990-2PS2–046) under the indigenous 5000 Ph.D. fellowship scheme from the Higher Education Commission (HEC) of Pakistan.
We would also like to acknowledge funding support from HEC under the National Research Program for Universities (NRPU) Project No. 6085.
We acknowledge the support from University of Warwick for open access publishing.

Availability of data and materials

All data generated or analyzed during this study are included in this paper or available at online repositories. A Python implementation of the proposed method together with a webserver is available at https://sites.google.com/view/wajidarshad/software and https://github.com/wajidarshad/ISLAND.

Ethics approval and consent to participate
This research does not involve human subjects, human material, or human data.

Consent for publication
This manuscript does not contain details, images, or videos relating to an individual person.

Competing interests
We have no conflict of interest to declare.

Author details
[1]Computational Biology and Data Analysis Laboratory, Department of Computer Science and Information Technology, King Abdullah Campus, University of Azad Jammu & Kashmir, Muzaffarabad, Pakistan. [2]Biomedical Informatics Research Laboratory, Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Nilore, Islamabad, Pakistan. [3]Biotechnology Laboratory, Department of Zoology, King Abdullah Campus, University of Azad Jammu & Kashmir, Muzaffarabad, Pakistan. [4]Department of Computer Science and the PathLAKE Consortium, University of Warwick, Coventry, UK.

Received: 3 April 2020 Accepted: 15 November 2020
Published online: 25 November 2020

References

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular biology of the cell. 4th ed. New York: Garland Science; 2002. https://www.ncbi.nlm.nih.gov/books/NBK26911/. Accessed 15 Apr 2017.
2. Tomlinson IM. Next-generation protein drugs. Nat Biotechnol. 2004;22:521–2.
3. Wilkinson KD. Quantitative analysis of protein-protein interactions. Methods Mol Biol Clifton NJ. 2004;261:15–32.
4. Kastritis PL, Bonvin AMJJ. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. J R Soc Interface. 2013;10:20120835.
5. Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein–protein complexes. eLife. 2015;4: e07454.
6. Chothia C, Janin J. Principles of protein–protein recognition. Nature. 1975;256:705–8.
7. Horton N, Lewis M. Calculation of the free energy of association for protein complexes. Protein Sci Publ Protein Soc. 1992;1:169–81.
8. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol. 1998;280:1–9.
9. Qin S, Pang X, Zhou H-X. Automated prediction of protein association rate constants. Struct Lond Engl 1993. 2011;19: 1744–51.
10. Audie J, Scarlata S. A novel empirical free energy function that explains and predicts protein–protein binding affinities. Biophys Chem. 2007;129:198–211.
11. Ma XH, Wang CX, Li CH, Chen WZ. A fast empirical approach to binding free energy calculations based on protein interface information. Protein Eng. 2002;15:677–81.
12. Su Y, Zhou A, Xia X, Li W, Sun Z. Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. Protein Sci Publ Protein Soc. 2009;18:2550–8.
13. Kastritis PL, Bonvin AMJJ. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res. 2010;9:2216–25.
14. Ain QU, Aleksandrova A, Roessler FD, Ballester PJ. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdiscip Rev Comput Mol Sci. 2015;5:405–24.
15. Moal IH, Agius R, Bates PA. Protein-protein binding affinity prediction on a diverse set of structures. Bioinformatics (Oxford, England). 2011;27(21):3002–9. https://doi.org/10.1093/bioinformatics/btr513.
16. Tian F, Lv Y, Yang L. Structure-based prediction of protein-protein binding affinity with consideration of allosteric effect. Amino Acids. 2012;43:531–43.
17. Yugandhar K, Gromiha MM. Protein-protein binding affinity prediction from amino acid sequence. Bioinformatics. 2014; 30:3583–9.
18. Chen M, Ju CJ-T, Zhou G, Chen X, Zhang T, Chang K-W, et al. Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. Bioinformatics. 2019;35:i305–14.
19. Srinivasulu YS, Wang J-R, Hsu K-T, Tsai M-J, Charoenkwan P, Huang W-L, et al. Characterizing informative sequence descriptors and predicting binding affinities of heterodimeric protein complexes. BMC Bioinformatics. 2015;16:1–11.
20. Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AMJJ, et al. A structure-based benchmark for protein-protein binding affinity. Protein Sci Publ Protein Soc. 2011;20:482–91.
21. Yugandhar K, Gromiha MM. Response to the comment on 'protein-protein binding affinity prediction from amino acid sequence. Bioinformatics. 2015;31:978.
22. Yugandhar K, Gromiha MM. Protein–protein binding affinity prediction from amino acid sequence. Bioinformatics. 2014; 30(24):3583–9. https://doi.org/10.1093/bioinformatics/btu580.
23. Chen J, Sawyer N, Regan L. Protein–protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. Protein Sci Publ Protein Soc. 2013;22:510–5.
24. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol. 2004;22:1035–6.
25. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147:195–7.
26. Abbasi WA, Asif A, Ben-Hur A, Minhas FUAA. Learning protein binding affinity using privileged information. BMC Bioinformatics. 2018;19:425.
27. Leslie C, Eskin E, Noble WS. The spectrum kernel: a string kernel for SVM protein classification. Pac Symp Biocomput Pac Symp Biocomput. 2002;7:564–75.
28. Minhas FUAA, Ben-Hur A. Multiple instance learning of Calmodulin binding sites. Bioinformatics. 2012;28:i416–22.
29. Minhas FUAA, Ross ED, Ben-Hur A. Amino acid composition predicts prion activity. PLoS Comput Biol. 2017;13:e1005465.

30. Cao D-S, Xu Q-S, Liang Y-Z. Propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics. 2013;29:960–2.
31. Limongelli I, Marini S, Bellazzi R. PaPI: pseudo amino acid composition to score human protein-coding variants. BMC Bioinformatics. 2015;16:123.
32. Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 2006;34(suppl 2):W32–7.
33. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
34. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005;33(suppl 1):D501–4.
35. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker John M, editor. The Proteomics Protocols Handbook: Humana Press; 2005. p. 571–607. https://doi.org/10.1385/1-59259-890-0:571.
36. Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. 1994;22:3174–80.
37. Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. 1990;4:155–61.
38. Ben-Hur A, Noble WS. Kernel methods for predicting protein–protein interactions. Bioinformatics. 2005;21(suppl 1):i38–46.
39. Cortes C, Mohri M, Rostamizadeh A. Learning sequence kernels. In: 2008 IEEE Workshop on Machine Learning for Signal Processing; 2008. p. 2–8.
40. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. PLoS Comput Biol. 2008;4:e1000173.
41. Mercer J. Functions of positive and negative type, and their connection with the theory of integral equations. Philos Trans R Soc Lond Math Phys Eng Sci. 1909;209:415–46.
42. Saigo H, Vert J-P, Ueda N, Akutsu T. Protein homology detection using string alignment kernels. Bioinformatics. 2004;20:1682–9.
43. Leslie CS, Eskin E, Cohen A, Weston J, Noble WS. Mismatch string kernels for discriminative protein classification. Bioinformatics. 2004;20:467–76.
44. Ahmad S, Mizuguchi K. Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. PLoS One. 2011;6:e29104.
45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
46. Watson GS. Linear least squares regression. Ann Math Stat. 1967;38:1679–99.
47. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
48. Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput. 2004;14:199–222.
49. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
50. Li H, Leung K-S, Wong M-H, Ballester PJ. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study. BMC Bioinformatics. 2014;15:291.
51. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinforma Oxf Engl. 2010;26:1169–75.
52. Abbasi WA, Minhas FUAA. Issues in performance evaluation for host–pathogen protein interaction prediction. J Bioinforma Comput Biol. 2016;14:1650011.
53. Moal IH, Fernández-Recio J. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinforma Oxf Engl. 2012;28:2600–7.

Publisher's Note