2014 3rd International Conference on Environment Energy and Biotechnology
IPCBEE vol.70 (2014) © (2014) IACSIT Press, Singapore
DOI: 10.7763/IPCBEE. 2014. V70. 7

Improved Prediction of Protein-Small Organic Ligand Binding Sites Via Consensus-Based Ranking with Linear Regression

Ibrahim Hijazi [1] and Lukasz Kurgan [1+]

[1] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada

Abstract. Prediction of binding of small organic ligands to proteins based on the knowledge of protein structures finds applications in rational drug discovery and elucidation of various cellular-level processes. Recent work shows that predictive quality of computational predictors of these binding events can be improved with the use of a consensus-based approach that combines predictions from several base predictors. We designed a novel type of a consensus, called ConSitePred, which uses a regression-based meta-predictor to (re)rank predictions from four well-performing base methods. The regression uses a vector of six custom-designed and empirically selected features that quantify atomic composition of the protein nearby the predicted binding site and presence and quality of other binding site predictions that are close to the predicted site. We empirically show that ConSitePred's predictions improve over the predictions of a comprehensive set of ten existing predictors, including its four base methods. Our method provides an alternative to other consensuses-based approaches that are based on clustering predictions from the base methods.

Keywords: protein-ligand binding, protein-ligand interactions, binding sites, binding pockets, prediction

1. Introduction

The past two decades observed a substantial interest in computational studies of interactions between proteins and other molecules. These include investigations that have analysed and characterized protein-protein [1]-[3], protein-DNA [4], protein-RNA [4]-[6], and protein-small ligand [7]-[9] interactions. Numerous computational methods that predict binding (interaction) sites on the protein surface were also developed [10]-[13].

In this work we focus on computational prediction of interactions with small (less than 100 non-hydrogen atoms) organic molecules. These molecules constitute nearly 90% of the drugs approved by the U.S. Food and Drug Administration [14]. Moreover, they are of particular interest because they are involved in numerous cellular activities, such as cellular signalling, growth of neurons, and regulation of cell cycles [15]-[17]. Knowledge of binding sites of these small molecules is important for rational drug discovery [18], [19]. Here, we use "ligand" and "binding site" terms to refer to the small organic compounds and the sites on the protein structure where they bind, respectively.

The field of structure-based prediction of ligand binding sites was recently reviewed in [11]. There are three types of predictive approaches that are based on geometrical analysis, calculation of binding energy, and threading using structural templates. The geometry-based predictors include SURFNET [20], PocketFinder [21], PASS [22], LIGSITE^{csc} [23], PocketPicker [24], ConCavity [25], and Fpocket [26]. The representative energy-based and threading-based approaches are Q-SiteFinder [27] and Findsite [28], respectively. The review [11] included a large-scale comparative evaluation of several publicly available predictors, and suggested that improved predictive quality can be obtained by building consensus-based methods, i.e., meta-methods that combine predictions from several base methods. MetaPocket [29] is a consensus approach that combines four base methods: LIGSITE^{csc}, PASS, Q-SiteFinder, and SURFNET. MetaPocket perform predictions in three steps: (1) it collects top three predictions (represented by predicted

[+] Corresponding author. Tel.: + 1 780 494-5488; fax: + 1 780 492-1811
E-mail address: lkurgan@ece.ualberta.ca

positions of the center of ligand) from each of the four methods; (2) it clusters these 12 predictions using hierarchical clustering according to their spatial similarity (distance); and (3) it ranks all clusters based on the sum of the (normalized) z-scores of the predicted pockets included in the cluster. The result is a ranked list of (new) predictions that are computed based on center of mass of each cluster.

We investigate an alternative design of a consensus. Similarly as in MetaPocket, we empirically select four base methods but instead of using clustering and generating new predictions, we use a meta-predictor to (re)rank predictions collected from the four base methods. Our meta-predictor, named ConSitePred, represents each prediction from each base method using a vector of numerical descriptors (features). We considered features that quantify certain structural properties of the input protein nearby the predicted binding site, atomic and amino acid (AA) composition of the predicted binding site, and certain geometric properties of the input protein and the predicted binding site in relation to predictions from other base methods. An empirically selected (well-performing) subset of these features is inputted into a prediction model that provides a score, which in turn is used to rank the predictions from the base methods. We carefully designed our ConSitePred including feature selection (from over 200 considered features) and selection of predictive model (out of two possible choices).

2. Materials and Methods

2.1. Datasets and evaluation protocols

We use the high-quality benchmark dataset proposed in [11]. Three proteins in that dataset could not be processed by DSSP (Kabsch and Sander, 1983), which we need to generate the features, and thus were excluded. Some of the considered predictors could not perform predictions for another 32 proteins, which were also excluded. We randomly divided the remaining 216 proteins into two similarly-sized subsets, one that was used for training the meta-predictor (TRAINING dataset with 110 proteins) and the other that was used to perform out-of-sample testing (TEST dataset with 106 proteins). All design steps of the meta-predictor were performed based on five-fold cross validation on the TRAINING dataset; the final design was tested and compared with other predictors on the TEST dataset.

As proposed in [11]-[28], we use the center-to-center distance (Dcc) between predicted and native (true) positions of the ligand, to assess predictive quality. For a given protein with *n* ligand binding sites we assess the top *n* predictions (based on a ranking generated by a given method, including our ConSitePred), and we aggregate this assessment over the entire dataset. Next, we compute a success rate for a given distance cutoff, i.e., predictions for which Dcc is smaller than cutoff are assumed correct and we compute ratio of these correct predictions among all native ligand binding sites in the entire dataset. We use cutoff ranging from 1Å to 20Å, with step of 1Å. The success rates are plotted against the cutoff values forming the success rate curve (Fig. 1). Finally, we compute normalized area under the success rate curve (AUS) when considering distances between 1Å and 10Å (AUS_{10}) and between 1Å and 20Å (AUS_{20}). We normalize the area under the success rate curve by the highest attainable value, which is 10 and 20 for AUS_{10} and AUS_{20}, respectively. Higher AUS values correspond to more accurate predictions. AUS_{10} focuses on predictions that are closer to the native binding site; AUS_{20} gives a more comprehensive evaluation of a larger set of predictions.

2.2. Design of the meta-predictor

The score generated by the meta-predictor is a predicted distance to the native binding site. The true distance (i.e., Euclidian distance between a given prediction and the closest native binding site) is transformed with the help of logistic function: $-1+2/(1+e^{-0.5*\text{distance}})$. This transformation forces the predictor to focus on minimizing the errors for small distances, i.e., to obtain higher quality predictions closer to the native binding site, rather than to minimize errors irrespective of the distance. Based on empirical comparison of predictive quality of ten predictors (SURFNET, PocketFinder, PASS, $\text{LIGSITE}^{\text{csc}}$, PocketPicker, ConCavity, Fpocket, Q-SiteFinder, Findsite, and MetaPocket) on the TRAINING dataset, we selected the top four performing methods to include in our meta-predictors. They include Findsite, Concavity, Q-SiteFinder, and MetaPocket. This result agrees with the results in [11].

We generated total of 242 features:

- features based on the input protein structure and sequence

    - solvent accessibility (generated with DSSP) of AA closest the predicted binding site, grouped by AA types; sum of solvent accessibilities of AAs that are within a radius of 4, 6, 8, and 10Å from the predicted site; count of AAs with solvent accessibility > 0.25 (solvent exposed AAs) that are within the radius of 4, 6, 8, and 10Å from the predicted site, including their type; (105 features)
    - secondary structure (helix/strand/coil; generated with DSSP) of AAs that are within the radius of 4, 6, 8, and 10Å from the predicted site; min. distance between the predicted site and the closest helix/strand/coil; composition of helix/strand/coil conformations in the entire protein; (18 features)
    - AA composition of the entire protein; (20 features)
    - composition of the predicted binding site including min. distance between a given binding site prediction and a particular AA type; atom type (H, C, N, O, and S) closest to the predicted site with the cutoff of 4 and 6Å; min. distance between the predicted site and closest H, C, N, O, and S atoms of the protein; count of H, C, N, O, and S atoms within a radius of 4, 6, 8, and 10Å from the predicted site; (56 features)
    - shape of the input protein fold, expressed by its radius of gyration; (1 feature)
- features based on the predicted binding site including average and min. distances between the considered predicted site and all sites predicted by the other base methods; count of other predictions within a radius of 4, 6, 8, and 10Å from the predicted site, also aggregated by the base method name; name of the method that generated the considered predicted site; and scores generated by the base methods for a given predicted site, e.g., from Findsite we used fraction of templates that shared this pocket, number of templates used to evaluate this pocket, and max., min. and average TM-scores and RMSDs. (42 features)

Next, we performed empirical feature selection to select a subset of features that are relevant to our objective. First, we removed low quality features that do not correlate with the output of the prediction (the transformed distance to the native binding site) using Pearson correlation coefficient (PCC). We calculated PCC for every considered feature for each training fold from the 5-fold cross validation on the TRAINING dataset. We only consider features with the average (over the five folds) PCC > 0.2; other features were removed. Next, we performed greedy, wrapper-based feature selection to remove redundant features. We sorted the remaining features by their average absolute PCC in the descending order and used two predictors: linear regression (LR) and the support vector regression (SVR). We performed two greedy search types:

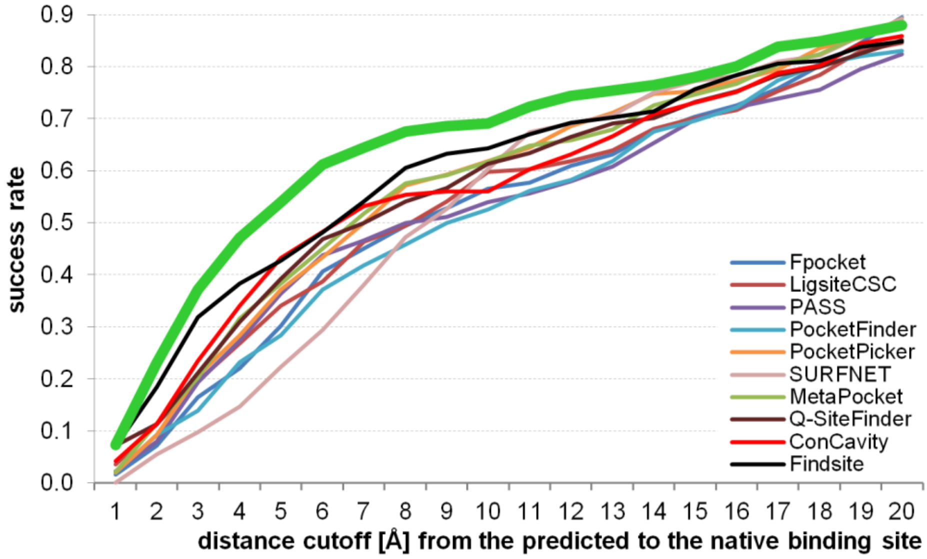
- Forward (FFS) where we start with the top-ranked feature and we add the next-ranked feature if it improves the prediction quality based on the 5-fold cross validation on the TRAINING dataset.
- Backward (BFS) where we start with all features and we remove the next-lower-ranked feature (starting with lowest-ranked feature) if the prediction quality (based on the 5-fold cross validation on the TRAINING dataset) does not deteriorate due to the removal.

We parameterized the SVR model before and after the second step of the feature selection. We considered Gaussian kernel with complexity parameter = 2^{-5}, 2^{-4},… 2^{5} and gamma = 2^{-10}, 2^{-8}, … 2^{4} and selected the parameter values that provided the highest predictive quality based on the 5-fold cross validation on the TRAINING dataset.

The AUS_{10} and AUS_{20} values of the resulting 4 setups based on the 5-fold cross validation on the TRAINING dataset are: 0.43 and 0.63 for LR model and FFS (6 features); 0.38 and 0.57 for LR and BFS (30 features); 0.40 and 0.59 for SVR and FFS (7 features); 0.41 and 0.59 for SVR and BFF (33 features). To compare, AUS_{10} and AUS_{20} values of the four base methods on the TRAINING datasets are: 0.39 and 0.59 for Findsite; 0.33 and 0.55 for ConCavity; 0.29 and 0.51 for Q-SiteFinder; and 0.27 and 0.52 for MetaPocket. These empirical results reveal that the LR with FFS provides the highest predictive performance and that this meta-predictor outperforms the base methods. This setup was used to implement our ConSitePred method.

3. Results and Discussion

The proposed ConSitePred is empirically compared with a comprehensive set of ten existing predictors, including SURFNET, PocketFinder, PASS, LIGSITE^{csc}, PocketPicker, ConCavity, Fpocket, Q-SiteFinder, Findsite, and MetaPocket, on the TEST dataset (Fig. 1). The results show that the proposed consensus-based approach generates promising results that improve over the results of the other predictors, including the four base methods. The success rates of ConSitePred are higher across the entire range of the distance cutoff values. This means that ConSitePred provides good predictive performance when the user is interested in predictions that are both very close to the native site and possibly farther away. Using the cutoff of 4Å, which was suggested in [28] since this value is similar to the radius of gyration of considered ligands, ConSitePred correctly predicts 42% of binding sites compared to the 35% obtained by the second best Findsite. The areas under the success rate curve AUS_{10} and AUS_{20} of ConSitePred are 0.46 and 0.63, respectively. These values are larger by 100%*(0.46-0.39)/0.39 = 18% and 100%*(0.63-0.57)/0.57 = 11% than the corresponding area values of the second best Findsite.



Method	AUS_{10}	AUS_{20}
ConSitePred	0.46	0.63
Findsite	0.39	0.57
ConCavity	0.35	0.54
MetaPocket	0.35	0.55
Q-SiteFinder	0.34	0.54
PocketPicker	0.34	0.55
$Ligsite^{CSC}$	0.31	0.51
PASS	0.31	0.49
Fpocket	0.29	0.50
PocketFinder	0.28	0.48
SURFNET	0.25	0.51

Fig. 1: Predictive quality of ConSitePred and other considered predictors on the TEST dataset measures using the success rate curves and the corresponding AUS_{10} and AUS_{20} values. The methods are sorted by the AUS_{10} value.

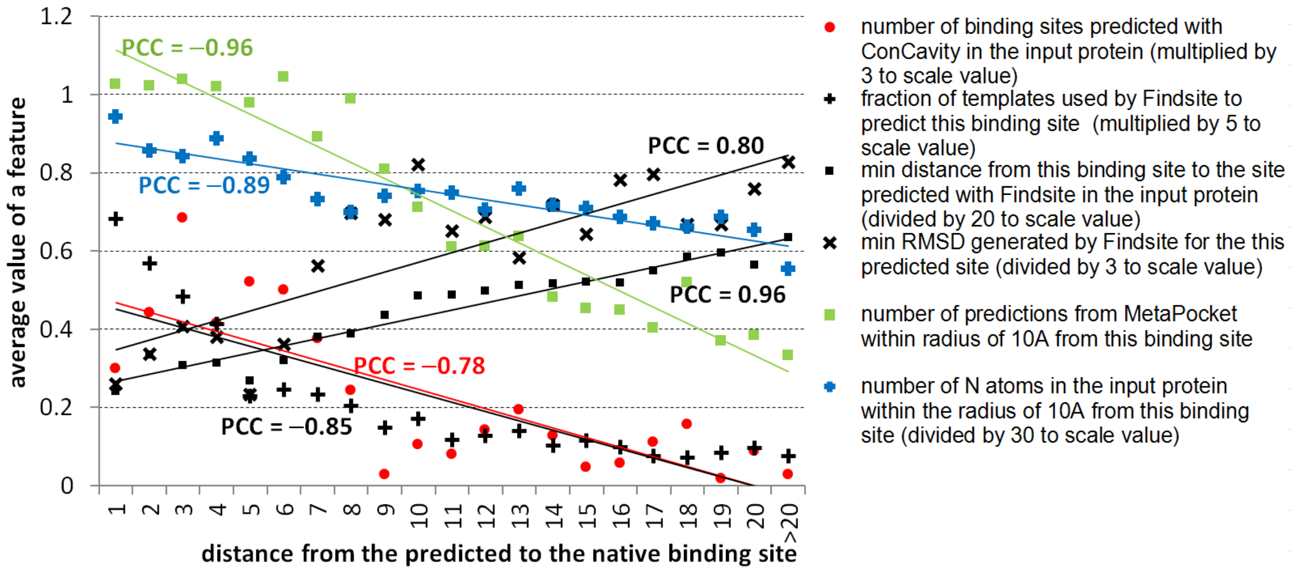


Fig. 2: Features used by ConSitePred (listed on the right). The *y*-axis shows average values of a given feature computed over all predicted sites in the TRAINING dataset for a given range of values of the distance given in *x*-axis; some value were scaled, as explained in the legend. Markers show the average values; lines show linear fit into the relation between the average values and the distance with the corresponding values of Pearson correlation coefficients (PCCs).

We attribute these improvements to the use of a well-performing set of novel features that were empirically selected from the considered comprehensive feature set. Instead of combining the predictions of the four base methods together through clustering, like in the other meta-method Meta-Pocket [29], we rank the predictions (from Findsite, Concavity, Q-SiteFinder, and MetaPocket) using linear regression with six features that were empirically selected using the TRAINING dataset. The selected features quantify several different aspects of each predicted binding site and are highly correlated with the distance from the predicted to the native binding site (|PCC| ≥ 0.78), see Figure 2; the latter suggests that they provide useful predictive input. Three features (shown using black markers in Fig. 2) consider the quality of the predicted and nearby predictions from Findsite, e.g., black + markers denote feature that quantifies fraction of structural template used by Findsite's threading that is higher for the predictions that are closer to the native site (that have lower distance). This is a credible relationship since higher number of available templates usually results in higher quality of threading. Two other features consider location and number of predictions from ConCavity (red markers) and MetaPocket (green markers) that are nearby the predicted binding site. These features reveal that when more of these predictions are closer to the predicted site then this site is more likely to be correct (distance is lower). Finally, the last feature (blue markers) shows that distance is lower (i.e., prediction is more accurate) when the number of nitrogen atoms in the input protein that are close to the predicted site (<10Å away) is higher. This is likely related to the fact that nitrogen atoms may form covalent bonds with the considered ligands [7]. Moreover, besides the use of novel features, another reason for the favorable success rates of ConSitePred is the careful, empirical design that included selection of well-performing base methods and prediction model.

To conclude, our results demonstrate that the quality of the prediction of binding sites of small organic compounds in protein structures can be improved with a consensus-based approach that (re)ranks predictions generated by well-performing predictors using a meta-predictor.

4. References

[1] M. Baaden and S. J. Marrink. Coarse-grain modelling of protein-protein interactions. *Curr Opin Struct Biol*. 2013, **23**(6):878-886.

[2] G. Brady, P. Stouten. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des*. 2000, **14**:383-401.

[3] N. Brooijmans, and I.D. Kuntz. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct*. 2003, **32**:335-373.

[4] J. A. Capra, R. A. Laskowski, J. M. Thornton, M. Singh, T. A. Funkhouser. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol*. 2009, **5**:e1000585.

[5] K. Chen, and L. Kurgan. Investigation of atomic level patterns in protein-small ligand interactions. *PLoS ONE* 2009, **4**:e4473.

[6] K. Chen, M. J. Mizianty, J. Gao, L. Kurgan. A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure* 2011, **19**(5):613-621.

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

[8] T. Gallo Cassarino, L. Bordoli, T. Schwede. Assessment of ligand binding site predictions in CASP10. *Proteins* 2014, **82**(S2):154-163.

[9] M. Hendlich, F. Rippmann, G. Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J. *Mol. Graph. Model*. 1997, **15**:359-363.

[10] B. Huang. MetaPocket: a meta approach to improve protein ligand binding site prediction. *OMICS* 2009, **13**:325-330.

[11] B. Huang, M. Schroeder. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol*. 2006, **6**:19.

[12] S. Jones, and J. M. Thornton. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* 1996, **93**:13-20.

[13] S. Jones, and J. M. Thornton. Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.* 2004, **8**:3-7.

[14] W. Kabsch, C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983, **22**:2577-2637.

[15] K. Kasahara, M. Shirota, K. Kinoshita. Comprehensive classification and diversity assessment of atomic contacts

in protein-small ligand interactions. *J Chem Inf Model*. 2013, **53**(1):241-248.

[16] R. Laskowski. (1995). SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph*. **13**:323-330.

[17] A. T. Laurie and R. M. Jackson. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 2005, **21**:1908-1916.

[18] V. Le Guilloux, P. Schmidtke, P. Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* 2009, **10**:168.

[19] N. M. Luscombe, R. A. Laskowski, J. M. Thornton. Amino acidbase interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*. 2001, **29**:2860-2874.

[20] S. Mukherjee, B. R. Acharya, B. Bhattacharyya, G. Chakrabarti. Genistein arrests cell cycle progression of A549 cells at the G(2)/M phase and depolymerizes interphase microtubules through binding to a unique site of tubulin. *Biochemistry* 2010, **49**:1702-1712.

[21] Y. Murakami, K. Kinoshita, A.R. Kinjo, H. Nakamura. Exhaustive comparison and classification of ligand-binding surfaces in proteins. *Protein Sci*. 2013, **22**(10):1379-1391.

[22] A. G. Ngounou Wetie, I. Sokolowska, A. G. Woods, U. Roy, K. Deinhardt, C. C. Darie. Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cell Mol Life Sci*. 2014, **71**(2):205-228.

[23] Y. Popova, P. Thayumanavan, E. Lonati, M. Agrochao, J.M. Thevelein. Transport and signaling through the phosphate-binding site of the yeast Pho84 phosphate transceptor. *Proc. Natl. Acad. Sci. USA* 2010, **107**: 2890-2895.

[24] J. Skolnick and M. Brylinski. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. USA* 2008, **105**:129-134.

[25] A. Srinivas Reddy, L. Chen, S. Zhang. Structure-Based De Novo Drug Design, in *De novo Molecular Design* (Ed. G. Schneider), Wiley, 2014.

[26] M. Weisel, E. Proschak, G. Schneider. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* 2007, **1**:7.

[27] J. D. Whittard, T. Sakurai, M. R. Cassella, M. Gazdoiu, D. P. Felsenfeld. MAP kinase pathway-dependent phosphorylation of the L1-CAM ankyrin binding site regulates neuronal growth. *Mol. Biol Cell* 2006, **17**:2696-2706.

[28] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*. 2008, **36**:D901-906.

[29] T. Zhang, H. Zhang, K. Chen, J. Ruan, S. Shen, L. Kurgan. Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr. Protein Pept. Sci.* 2010, **11**:609-628.

[30] H. Zhu, I. Sommer, T. Lengauer, F.S. Domingues. Alignment of non-covalent interactions at protein-protein interfaces. *PLoS ONE* 2008, **3**:e1926.