An approach for prediction of the protein's amino acid residues that are part of binding region

Georgina Mirceva[1], Andreja Naumoski[1] and Andrea Kulakov[1]

[1] Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering, Rugjer Boskovikj 16, 1000 Skopje, Macedonia

georgina.mirceva@finki.ukim.mk, andreja.naumoski@finki.ukim.mk, andrea.kulakov@finki.ukim.mk

Abstract. Proteomics is a research area that analyze protein molecules in order to determine their sequence, structure and functions. It is very popular because proteins play an important part in the processes in the cells of organisms. There are plethora of methods for protein annotation, which try to found out the functions of the proteins in the interactions between the proteins. However, they cannot follow the speed of determination of protein structures, thus the gap between the known protein structures and those that are functionally annotated continually rises. Therefore, there is a noticeable necessity for developing fast and accurate computational methods for determination of protein functions. In this paper, we present an approach for prediction of the amino acid residues that are part of binding region where an interaction occurs with other protein structure. These predictions could be used for protein annotation. We present some results of the evaluation of our approach, and we compare it with some approaches found in the literature.

Keywords. Protein molecule, protein function, protein binding site.

1 Introduction

The examination of protein molecules is very attractive nowadays. Proteomics is the research area that study proteins and tries to find their sequence and structure that could be further used to determine their functions. Proteins are one of the essential constituents of the cells, thus they influence many processes in the organisms. The knowledge about the functions of protein molecules could be used in drug design, thus providing a way to regulate these processes where they are involved. That is why proteomics is very popular, since it could improve humans' life. As novel technologies are introduced, the protein structures are determined with faster rate. On the other hand, the methods for protein function determination is not performed that fast, so the number of proteins molecules with known structure and unknown functions continually rises. Therefore, there is a necessity for fast methods for functionally annotating protein structures. Different methods from the literature consider different information in order to determine the functions of proteins. Proteins with same predecessor

are more likely to share their functions, therefore some methods [1] try to identify the homologous proteins of the inspected protein structure in order to determine its functions. Because the proteins with common predecessor share same functions, therefore it is thought that those parts of the protein structures that remain the same during evolution are those that determine the functions. Therefore, the researchers have developed another group of methods [2] that aim to find the parts of protein structures that do not change during evolution, which are called conserved parts, and then to analyze their features in order to find out the protein functions. Based on the knowledge gathered about the proteins that interact, protein-protein interactions networks are built. There are many methods [3] that perform functional annotation of protein structures by analyzing these protein-protein interaction networks. In this research we focus on the methods [4] that identify the binding sites of the examined protein where interactions with other protein structures occur and then determine the protein functions by analyzing the features of the binding sites.

In this paper, we aim to find out which of the amino acid residues of a given protein structures could be part of binding region. The determination of the binding sites is based on the features of the amino acid residues. For that purpose, in this research we take into consideration the following features of the amino acid residues: Accessible Surface Area (ASA) [5], Relative ASA (RASA) [6], depth index (DPX) [7], protrusion index (CX) [8] and hydrophobicity [9]. These features are most frequently used for identification of the binding sites of protein structures. Based on these features, a classification model could be generated. Because amino acid residues have several atoms, we can extract various features by taking into consideration different sets of atoms. In this way we obtain a larger number of features that could provide more accurate estimation. Since the dimensionality of the dataset would be huge because the number of samples is high, therefore we can apply some technique for selecting the most significant features. By using a suitable technique for feature selection, besides reduction of the dataset we also can advance the predictions and reduce the training and testing times. Also, by using lower number of features, generally less complex models are obtained, which could be read easily. In our previous research [10], we used various classification methods for model induction. In the same research, we also applied different feature selection techniques in order to obtain a dataset with the most significant features. In [10], we made evaluation on two datasets. Later, in [11] we performed protein binding sites identification on another dataset by using several classification methods. In this paper, we use the dataset used in [11] and we make analysis of the prediction power of the models obtained by using the approach that we used in [10].

This paper is structured in the following way. In Section 2, we present our approach for identification of the amino acid residues that are part of binding region. Section 3 provides some results of the evaluation of our approach and also contains results from the comparison with some approaches for protein binding sites detection found in the literature. Finally, Section 4 gives conclusions and directions for future work.

2 Our Approach for Prediction of Protein Binding Sites

The prediction of protein binding sites is made by analyzing the features of the amino acid residues that constitute the inspected protein structure. In order to do that, we apply an approach that has three steps for training the classification model, see Fig. 1. In the first step, the features of the residues are extracted. Next, in the second step, we perform selection of the most significant features. Finally, in the third step, prediction model is generated by applying some classification method. In the following subsections, we describe these steps in more details.

2.1 First Step: Extraction of the Features of Amino Acid Residues

In this paper, we consider the following features of the amino acid residues: Accessible Surface Area (ASA) [5], Relative Accessible Surface Area (RASA) [6], depth index (DPX) [7], protrusion index (CX) [8] and hydrophobicity [9].

ASA [5] is a feature that shows how much of the area of a given atom could be touched by another protein structure. It is estimated by using a sphere with a predefined radius, which is traversed over the protein structure. By using the approximation described in [5], the surface area that is accessible to this sphere is assessed. Because the number of constituting atoms is different for different amino acid, therefore it is better to use a feature that would provide equity for each amino acid. Therefore, the Relative ASA (RASA) [6] is also widely used for protein binding sites prediction. Namely, the calculated value for ASA for a given amino acid residue is divided by the standard value for ASA for that particular amino acid, and thus the RASA feature is obtained. In this research, the NACCESS program [6] is used for extracting the ASA and RASA features.

The depth index (DPX) [7] is a feature of an atom that shows how far it is from the surface of the protein structure. It is calculated as Euclidean distance between the observed atom and its nearby atom that has ASA larger than zero.

The protrusion index (CX) [8] is also very significant feature, which gives useful evidence that could be used for binding sites prediction. It shows if the surrounding of an atom is dense region filled with many other atoms or not. For each atom, the surrounding within a given radius is observed, and the depth index is calculated by dividing the non-occupied volume and occupied volume within the observed sphere. This is done by using the technique presented in [8].

The amino acids are not equally found in each part of the protein. Specifically, some amino acids are regularly found near the protein surface, while others have a tendency to be placed close to the center of molecule. This is known as hydrophobic effect, and the researchers have proposed different scales that define the preferences of the amino acids to be located in a particular region. In this paper we use the hydrophobicity scale proposed by Kyte and Doolittle [9].

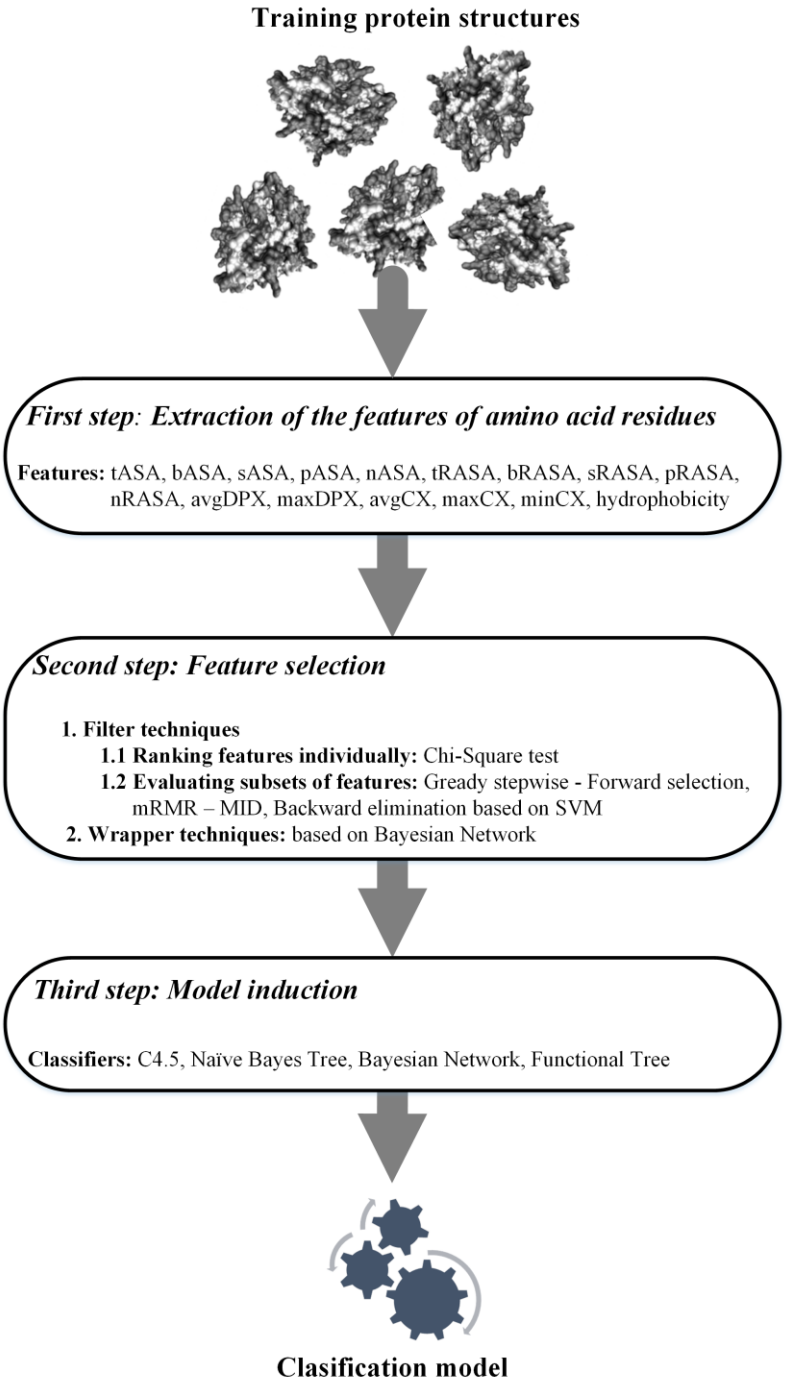

Fig. 1. Description of the three steps performed in the training phase of our approach.

ASA, RASA, DPX and CX are calculated for each atom of the protein structure that is analyzed. However, in the prediction of binding sites, the samples correspond to the amino acid residues, which are composed of a number of atoms that are folded in a particular way in the three-dimensional space. In order to calculate the features for a particular amino acid residue, we accumulate the values for the corresponding feature by taking into account different sets of atoms. Regarding ASA and RASA, we calculate the totalASA (tASA) and totalRASA (tRASA) by summing the values for all atoms. Based on the location of the atoms, we differentiate the atoms that are on the backbone from those that are on the side-chain, thus providing backboneASA (bASA), side-chainASA (sASA), backboneRASA (bRASA) and side-chainRASA (sRASA). By considering only the polar or non-polar atoms, we obtain polarASA (pASA), non-polarASA (nASA), polarRASA (pRASA) and non-polarRASA (nRASA). Regarding DPX and CX, we combine the features of all atoms that are part of the residue by considering their average, maximal and minimal value, thus obtaining avgDPX, maxDPX, minDPX, avgCX, maxCX and minCX. The hydrophobicity feature is a feature of the amino acids, so since in our case the amino acid residues itself are samples in the dataset, therefore there in no need to aggregate over its constituting atoms.

2.2 Second Step: Feature Selection

In this research, we apply the feature selection techniques that showed best performance in our previous research made in [10]. We can differentiate filter and wrapper [12] techniques for feature selection. In wrapper techniques [12], some classifier is used for generating classification model and the final objective function is optimized, while in filter techniques some other objective function is optimized.

From the filter techniques, we can distinguish two categories, which are techniques where the features are assessed separately, and techniques where subsets of features are inspected. In the first category, the dependency between a given feature and the class feature is studied, and the features with highest dependency are designated. Nevertheless, with this techniques, the redundancy is not considered. Therefore, the techniques from the second category look over a given subset of features by computing the dependency with the class feature and also consider the redundancy among the features from the observed subset.

From the first category, we apply the Chi-Square test [13]. Because this technique works with discrete attributes, therefore, first we discretize the features as in [10].

From the techniques of the second category, we apply greedy stepwise forward selection, minimum-Redundancy-Maximum-Relevance (mRMR) [14] technique and recursive backward elimination technique based on support vector machines (SVM) [15].

Greedy stepwise is a sequential technique because the features are injected or removed sequentially. In forward selection, we begin with empty set of features, and then the features are injected one by one based on that which feature provides highest growth in the Pearson's correlation coefficient. There is also reverse option called backward elimination where we begin with a set that holds all the features, and then

the features are removed one after another. In this paper, we use the greedy stepwise forward selection technique.

In the latest literature, the mRMR technique is very popular, and it is applied in many research areas. It attempts to maximize the relevance and to minimize the redundancy at the same time. In this paper, we use its Mutual Information Difference (MID) scheme [16] where the difference among the relevance and redundancy is maximized. This technique could be used only for discrete attributes, for that reason we make discretization of the features in ten intervals with same length.

The last filter technique applied in this research is the technique presented in [15], where recursive backward elimination of the features is made. The significance of the features is assessed based on the values of the weights obtained with the SVM classification models that are generated for the observed sets of features.

In wrapper techniques [12], some classifier is applied for generating models and the finial objective function is optimized. In this paper, we apply the Bayesian Network classifier [17] in the wrapper technique, because it was among the best feature selection techniques in the analyses made in [10]. The significance of a given set of features is assessed by carrying out 2-fold cross validation on the training set. In this study, we use the same settings for the feature selection techniques as in [10].

2.3 Third Step: Model Induction

In order to generate prediction models, we apply the classification methods that attained best results in the analyses that we made in [10]. We use the following classification methods: C4.5 Tree [18], Naïve Bayes Tree [19], Bayesian Network [17] and Functional Trees [20].

3 Evaluation

For evaluation of our approach, we use the same dataset that we used in [11]. The dataset is obtained from the LigASite v7.0 database [21] that holds redundant and non-redundant datasets. In the non-redundant dataset the protein chains have less than 25% similarity in their sequences. The set used for testing has samples that relate to the amino acid residues of protein chains from the non-redundant set. With the residues from the other chains that are not present in the non-redundant set we form the set used for training the models.

The amino acid residues that are deeply in the protein structure could not be touched from another protein, thus they cannot be elements of a binding region. In order to shrink the size of the dataset, we throw away the samples that relate to the residues that have less than 5% accessible area [22]. After cleaning the surface amino acid residues, the number of samples in the training and test sets is 132773 and 105408 correspondingly. Since we take into consideration only the surface residues, therefore minDPX equals zero for all samples, so we discard this feature.

The training dataset that is obtained in this way is a non-balanced set because for the class feature we have significantly larger number of samples that relate to the

residues that are not part of binding region versus the samples that represents the residues that constitute a binding region. To evade creating models that are predisposed for the class that is foremost, we sample the training set to 20% of its original size by keeping uniform distribution of the class feature. Then, we normalize the descriptive features in the interval [0;1].

First, we made experiments by using different feature selection techniques and applying different classification methods for creating model. The results achieved for AUC-ROC are presented in Table 1. The best results for each classification method are shown as bolded. By using C4.5 Tree and Naïve Bayes Tree, the best results are obtained by using the Chi-Square test for choosing features. On the other hand, when Bayesian Network is used for model induction, the most accurate model is generated when we use the corresponding wrapper technique based on the same classifier. Functional Tree classifier achieved best results in combination with the same feature selection technique (wrapper based on Bayesian Network). Generally, Naïve Bayes Tree and Functional Tree showed as better classification methods for generating models, and Chi-Squared test and the wrapper based on Bayesian Network showed best performance among the feature selection techniques.

Table 1. The results for AUC-ROC obtained by using different feature selection techniques and different classification methods.

Feature selection technique	C4.5 Tree	Naïve Bayes Tree	Bayesian Network	Functional Tree
Chi-Squared	**0.623**	**0.624**	0.609	0.619
Forward selection	0.616	0.621	0.609	0.618
mRMR (MID)	0.607	0.620	0.594	0.613
Backward elimination based on SVM	0.615	0.623	0.606	0.620
Wrapper based on Bayesian Network	0.616	0.619	**0.615**	**0.623**

We also made experiments for comparison with some other approaches for identification of the amino acid residues that form a binding region. In the comparison, we use distance-based [23], [24], conservation based [25], [26], [27], and pocket finding [28], [29], [30] approaches. Also, the ConCavity approach [31] is used in the comparison, which combines the sequence conservation calculated with the Jensen-Shannon divergence (JSD) approach [27] with the pocket finding approaches mentioned before. The results for AUC-ROC obtained in this comparison analysis are given in Table 2. The second column of Table 2 shows the type of the examined approach, i.e.: D (distance-based), C (conservation-based) and P (pocket finding).

The results show that on this dataset our approach for binding sites prediction is better than the distance-based and conservation-based approaches used in the comparison, while the approaches that use some pocket finding algorithm outperform our approach. However, in [10] we made analysis by using two different datasets that holds various types of proteins. The results presented in [10] showed that for the proteins from the LigASite database [21], the approaches based on pocket finding are

best, but on the other dataset used in [10], these approaches showed significantly lower performance than the other approaches. The conclusions remain the same as in [10] that the existing approaches attain good results only for a given group of proteins, while our approach is more general and by applying feature selection it tries to adapt towards the characteristics of the studied group of proteins.

Table 2. The results for AUC-ROC by using different approaches.

Approach	Type	Reference	AUC-ROC
Our approach		[10]	0.624
Atom nucleus distance	D	[23]	0.522
PIADA	D	[24]	0.523
ASA change	C	[25]	0.530
Van der Waals distance	C	[26]	0.517
JSD	C	[27]	0.609
LigSite	P	[28]	0.792
PocketFinder	P	[29]	0.804
Surfnet	P	[30]	0.785
Concavity LigSite	C+P	[31]	0.835
Concavity PocketFinder	C+P	[31]	0.836
Concavity Surfnet	C+P	[31]	0.819

4 Conclusions and Future Work

In this paper we applied an approach for identification of the proteins' amino acid residues that are part of binding region. First, the features of the residues are extracted, and then by using a feature selection techniques we try to find out which features should be considered in the dataset. Finally, we generate a model which identifies which of the samples corresponds to residues that could form a binding region.

To get a better picture about the performance of our approach, we made comparison with several approaches from the literature. For that purpose we considered distance-based, conservation-based and pocket finding approaches. On the dataset used in this research, our approach has better predictions than the distance-based and conservation-based approaches, while the approaches based on pocket finding attains highest performances. However, in [10], we showed that these approaches used for comparison have different performances on different groups of proteins, that is not a case with our approach which shows as very stable since it is universal and with the application of feature selection technique it attempts to adjust for the observed group of proteins.

We identified several directions for enhancement of this approach. First, and probably the most important is to extend the number of features by finding out some other features that are significant and that could be beneficial for making good predictions.

Regarding feature selection and model induction, we will carry on our hunt for finding the most suitable feature selection technique and classification method. As it was described in the introduction section, the protein binding sites prediction is necessary in order to make functional annotation of the proteins. So, besides identification of the residues that form a binding region, we will also explore for methods for determination of the functions of proteins.

Acknowledgments. This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University in Skopje", Skopje, R. Macedonia.

References

1. Todd, A.E., Orengo, C.A., Thornton, J.M.: Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. 307(4), 1113–1143 (2001).
2. Panchenko, A.R., Kondrashov, F., Bryant, S.: Prediction of functional sites by analysis of sequence and structure conservation. Protein Science 13(4), 884–892 (2004).
3. Kirac, M., Ozsoyoglul, G., Yang, J.: Annotating proteins by mining protein interaction networks. Bioinformatics 22(14), e260–e270 (2006).
4. Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., Nussinov, R.: A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. Briefings in Bioinformatics 10(3), 217–232 (2009).
5. Shrake, A., Rupley, J.A.: Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol. 79(2), 351–371 (1973).
6. Hubbard, S.J., Thornton, J.M.: NACCESS, Computer Program. Department of Biochemistry and Molecular Biology, University College London, London, UK (1993).
7. Pintar, A., Carugo, O., Pongor, S.: DPX: for the analysis of the protein core. Bioinformatics 19(2), 313–314 (2003).
8. Pintar, A., Carugo, O., Pongor, S.: CX, an algorithm that identifies protruding atoms in proteins. Bioinformatics 18(7), 980–984 (2002).
9. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157(1), 105–132 (1982).
10. Mirceva, G., Kulakov, A.: Improvement of protein binding sites prediction by selecting amino acid residues' features. J. Struct. Biol. 189(1), 9–19 (2015).
11. Mirceva, G., Naumoski, A., Kulakov, A.: Comparative analysis of methods for determination of protein binding sites. In: 14th International Conference on Informatics and Information Technologies. Mavrovo, Macedonia (2017).
12. Kohavi, R., John, G.H.: Wrappers for Feature Subset Selection. Artif. Intell. 97(1), 273–324 (1997).
13. Liu, H., Setiono, R.: Chi2: Feature Selection and Discretization of Numeric Attributes. In: Vassilopoulos, J.F., (ed.) Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence, pp. 388–391. IEEE Computer Society, USA (1995).
14. Peng, H.C., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE T. Pattern. Anal. 27(8), 1226–1238 (2005).
15. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46(1–3), 389–422 (2002).
16. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. 3(2), 185–205 (2005).
17. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Mach. Learn. 29(2–3), 131–163 (1997).

18. Quinlan, R.: C4.5: Programs for Machine Learning. 1st edn. Morgan Kaufmann Publishers, San Mateo, CA, USA (1993).
19. Kohavi, R.: Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In: Simoudis, E., Han, J., Fayyad, U., (eds.) Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD 1996), pp. 202–207. AAAI Press, Menlo Park, CA, USA (1996).
20. Gama, J.: Functional Trees. Mach. Learn. 55(3), 219–250 (2004).
21. Dessailly, B.H., Lensink, M.F., Orengo, C.A., Wodak, S.J.: LigASite a database of biologically relevant binding sites in proteins with known apo-structures. Nucleic Acids Res. 36 (Database issue), D667–D673 (2008).
22. Chothia, C.: The Nature of the Accessible and Buried Surfaces in Proteins. J. Mol. Biol. 105(1), 1–12 (1976).
23. Ofran, Y., Rost, B.: Predicted protein-protein interaction sites from local sequence information. FEBS Lett. 544(1–3), 236–239 (2003).
24. Mihel, J., Šikić, M., Tomić, S., Jeren, B., Vlahoviček, K.: PSAIA – Protein Structure and Interaction Analyzer. BMC Struct. Biol. 8, 21 (2008).
25. Jones, S., Thornton, J.M.: Analysis of protein-protein interaction sites using surface patches. J. Mol. Biol. 272(1), 121–132 (1997).
26. Aytuna, A.S., Gursoy, A., Keskin, O.: Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics 21(12), 2850–2855 (2005).
27. Capra, J.A., Singh, M.: Predicting functionally important residues from sequence conservation. Bioinformatics 23(15), 1875–1882 (2007).
28. Hendlich, M., Rippmann, F., Barnickel, G.: LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J. Mol. Graph. Model. 15(6), 359–363 (1997).
29. An, J., Totrov, M., Abagyan, R.: Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol. Cell Proteomics 4(6), 752–761 (2005).
30. Laskowski, R.: SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. J. Mol. Graph. 13(5), 323–330 (1995).
31. Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M., Funkhouser, T.A.: Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. PLoS Comput. Biol. 5(12), e1000585 (2009).