Identification of Protein-Ligand Binding Site Using Multi-Clustering and Support Vector Machine

Ginny Y. Wong
Centre for Signal Processing
Dept of EIE, PolyU
Hung Hom, Hong Kong
ginnyyk.wong@connect.polyu.hk

Frank H.F. Leung
Centre for Signal Processing
Dept of EIE, PolyU
Hung Hom, Hong Kong
frank-h-f.leung@polyu.edu.hk

Steve S.H. Ling
Centre for Health Technologies
Faculty of Engg & IT, UTS
NSW, Australia
steve.ling@uts.edu.au

***Abstract*— Multi-clustering has been widely used. It acts as a pre-training process for the identification of protein-ligand binding site in this paper. Then, the Support Vector Machine (SVM) is employed to classify the pockets that are most likely to bind ligands with the attributes of geometric characteristics, interaction potential, offset from protein, conservation score and properties surrounding the pockets. Our approach is compared to LIGSITE^{CSC}, SURFNET, Fpocket, Q-SiteFinder, ConCavity, and MetaPocket on the 198 drug-target protein complexes. The results show that our approach improves the success rate from 82% to 86%.**

Keywords—SVM, multi-clustering, protein-ligand binding site

I. INTRODUCTION

The drug discovery process starts with target identification and validation. This operation searches the causes of the phenotype of the disease. Protein plays a critical role in causing the symptoms of a human disease. Activating or inhibiting its function can have a positive effect on the disease [1]. After the relationship between the target and disease has been found, the next operation of drug discovery is to find a method to modify that target. This consists of protein-protein and protein-ligand (small chemical molecule) interactions.

Taking advantage of the three-dimensional (3D) structure of a protein, structure-based drug design (SBDD) attempts to contribute to drug discovery [2]. The 3D structure of a protein can be obtained experimentally with x-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy. Another method is to construct the protein based on its amino acid sequence and a similar protein with a known 3D structure. All this information can be found from the Protein Data Bank (PDB) [3] or Protein Quaternary Structure file server (PQS) [4], which show the atomic coordinates and the quaternary structure of proteins respectively. This has made the SBDD more and more feasible because the 3D atoms' arrangements of proteins allow the prediction of protein and ligand binding sites, which is an important pre-requisite of SBDD [5]. When the protein's structure is known, different approaches can be applied to find the ligand, such as virtual screening, docking and de novo drug design [6].

The protein-ligand binding sites are located in the pockets (clefts) on the surface of proteins. The prediction of pockets has been examined with the information regarding the proteins' sequence or structure. The sequence conservation was analysed to predict the residues involved in ligand binding [7]–[8]. The structural information includes the studies of geometry and interaction energy of proteins. In POCKET [9], LIGSITE [10], and SURFNET [11], the studies only use the geometric characteristics and assume that the binding site is usually located in the largest pocket. On the other hand, some methods like PocketFinder [12] and Q-SiteFinder [13] focused on the energetic criteria by calculating the van der Waals interaction potential. However, the structure-based methods are not so capable of tackling the multi-chain problems of proteins. The methods may treat the gaps among the chains of proteins as pockets incorrectly. Therefore, LIGSITE^{CSC} [14] and ConCavity [15] suggested that the sequence conservation should be integrated with the structural pocket identification to get more accurate binding sites of proteins, particularly the multi-chain proteins. MetaPocket [16-17] was a combination of eight predictors, including LIGSITE^{CSC} [14], PASS [18], Q-SiteFinder [13], SURFNET [11], GHECOM [19], ConCavity [15], Fpocket [20], and POCASA [21]. It ranked the predicted binding sites of the eight methods and found the potential binding sites according to their spatial similarity.

The prediction of binding sites is practically a binary classification problem to classify whether some grid points are likely to bind with ligands or not. The above methods calculate a score for each grid point based on the corresponding protein characteristics, and predict the binding sites based on these scores. However, the methods to determine these scores are not easy to decide. Therefore, the Support Vector Machine (SVM) was applied in our previous work [22] to predict the binding sites by using 29 proteins' attributes, including the geometric characteristics, interaction energy, sequence conservation, distance from protein, and the properties of the surrounding grid points. Like most of the datasets in bioinformatics, the data of the binding sites have the problems of being imbalanced and in large data scales [23]. Therefore, random under-sampling and filtering are also applied to reduce the data size.

In this paper, multi-clustering acts as an unsupervised pre-training process to improve our prediction method. Multi-clustering is widely used in different areas, such as big data [24], feature selection [25], data reduction [26], and deep

The work described in this paper was substantially supported by a grant from The Hong Kong Polytechnic University (Project Account Code. G-YN19).

learning [27]–[28]. After the training dataset is generated, it is clustered into eight groups depending on the type of attributes. SVM is then applied on each group of data. Therefore, eight classification models are generated.

The 198 drug-target dataset, which is developed in MetaPocket [17], is used to evaluate our method in this paper. Only the top three largest binding sites are predicted and each site is represented as a centre point in this experiment. Our approach is compared with six other published methods. They are LIGSITECSC, SURFNET, Fpocket [20], Q-SiteFinder, Con- Cavity, and MetaPocket. A new evaluation method, which is different from that in our previous study, is applied. It is more similar to the method in [17] in order to show comparison results more properly.

This paper is organized as follows. In Section II, the prediction methods for binding sites with consideration of the proteins' sequence and geometrical structure are described. Section III describes the attributes considered for each grid point. In Section IV, the overall process and the selected training data are introduced. The adopted evaluation method is discussed in Section V. Section VI shows the results of our proposed method. A conclusion will be drawn in Section VII.

II. Prediction Of Protein-Ligand Binding Site

This section describes the most common approaches of binding site prediction.

A. POCKET and LIGSITE

POCKET [9] is one of the geometry-based methods to define the binding sites. Firstly, a 3D grid is generated. Secondly, a distance check is applied on the grid to make sure the atoms of protein do not overlap with the grid point. All the grid points, which do not overlap with the atoms of protein, are labelled as solvent. If the grid points outside the protein are enclosed by the protein surface in opposite directions of the same axis (i.e. the grid points are enclosed by pairs of atoms within the protein), it is called a protein-solvent-protein (PSP) event (Fig. 1).

LIGSITE [10] is an extension to POCKET [9] with the scanning directions being increased. LIGSITE scans for the pockets along three axes and four cubic diagonals while POCKET only scans three axes. Both of them considered the identification of PSP events on the basis of atom coordinates. Some value will be assigned to each grid point, which is actually the number of PSP events occurred in the scanning directions. That means, the higher the value of a grid point, the more likely the grid point will be a pocket. Fig. 1 shows the PSP events of two enclosed grid points. This method only focuses on the geometric characteristics and does not consider any other properties of the protein.

B. SURFNET

SURFNET [11] is another geometry-based method to define the binding sites. Like LIGSITE, a 3D grid is generated first. The grid values of SURFNET are calculated by counting the number of constructed spheres. Firstly, pairs of relevant atoms are taken within the protein. Then, testing spheres are formed between the pairs. If the sphere overlaps with other atoms, the radius decreases until no overlapping occurs (Fig. 2). Only the distance between two atoms within 10 Å is considered. The sphere of radius smaller than 1.5 Å is also ignored. If the grid points are out of the pockets, the distances between pairs of atoms are very large or cannot be found. On the contrary, if the grid points are inside the pockets, more than

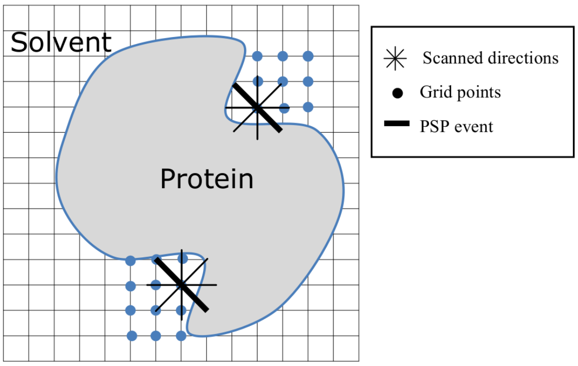

Fig. 1. PSP event used to describe the geometric feature of a grid point. It counts the number of scanning directions that the pair of protein atoms can enclose the grid point. For the POCKET method, the maximum number of PSP event is three while it is seven for the LIGSITE method.

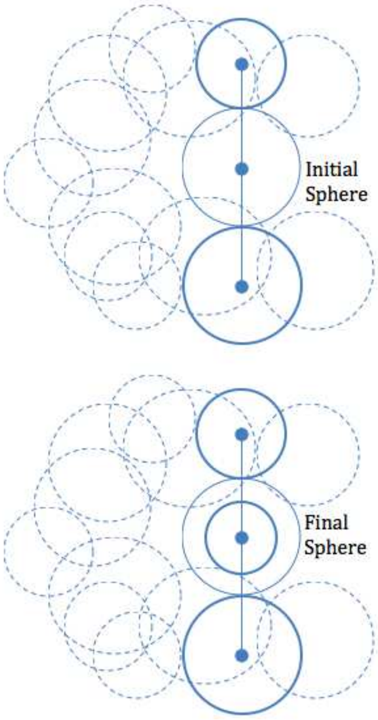

Fig. 2. SURFNET. There are three solid line circles and several dotted line circles in each graph. The top and bottom solid line circles represent the pair of relevant atoms and the middle one shows the constructed sphere of a grid point. The dotted line circles represent the other atoms that surround the testing grid point. The initial sphere in the upper graph overlaps with other atoms. Therefore, its radius decreases until no overlapping occurs to form the final sphere in the lower graph.

one sphere can be formed.

C. PocketFinder

PocketFinder [12] is an energy-based method of ligand binding site prediction. It uses the van der Waals interaction energy between the protein and a simple atomic probe to locate the binding sites with high energy. A 3D grid potential map is generated first. The potential at grid point p is calculated by the Lennard-Jones formula:

$$V(p)=\sum_{i=1}^{N}\left(\frac{C_{12}^{i}}{r_{pi}^{12}}-\frac{C_{6}^{i}}{r_{pi}^{6}}\right) \qquad (1)$$

where C_{12}^{i} and C_{6}^{i} are constants, which are the typical 12-6 Lennard-Jones parameters used to model the van der Waals interaction energy between a carbon atom placed at the grid point *p* and the protein atom *i*; *N* is the total number of protein atoms. r_{pi}^{12} and r_{pi}^{6} are the powers 12 and 6 of r_{pi} respectively, where r_{pi} is the distance between the grid point *p* and the protein atom *i*. The first term describes the repulsion between atoms when they are very close to each other. The second term describes the attraction between atoms at long distance.

D. Sequence Conservation

As not all residues in protein are equally important, conservation analysis is a very useful method to predict those functionally important residues in the protein sequence [29]-[31]. Sequence conservation has also been shown to be strongly correlated with ligand binding sites [7]-[8]. Therefore, [15] suggested combining the sequence conservation and the structure of protein to predict the protein ligand binding sites by weighting every pair of protein atoms.

III. Protein Attributes Used

This section describes the 29 protein attributes which are introduced in [22] by us. These attributes are also used in this paper for the training and testing for the identification of the protein-ligand binding sites.

1) Grid values: These are the two values of each grid point that are calculated by LIGSITE and SURFNET. They can represent the binding site preference based on geometric characteristics.

2) Interaction potential: This energy is the same as the van der Waals interaction potential of an atomic probe with the protein [12]. The calculation is done by the PocketFinder method, which is mentioned in Section II. The Lennard-Jones formula (1) is used to estimate the interaction potential between the protein and a carbon atom placed at the grid point.

3) Conservation score: Conservation score is obtained from a residue-level analysis to identify which residues in a protein are responsible for its function. The score of each grid point is the conservation score of the nearest residue. Jensen-Shannon divergence (JSD) method is used to calculate the score since it has been shown to provide an outstanding performance in identifying residues near bound ligands [31]. It is an open source program which is freely available in its webpage [31].

4) Distance from protein: The squared distance from each grid point to the closest point on the van der Waals surface of the protein is calculated. When the grid points are too far from the atoms, they are not likely to be a pocket. In the experiment, almost 90% of ligand atoms are located within 5Å of the protein's van der Waals surface. Hence, the grid points with the squared distance larger that 5Å are filtered out in order to reduce the huge data size.

5) Properties of surrounding grid points: All the binding sites are formed by many grid points (the distance between two grid points is 1Å [15]), so the properties of the grid points nearby are also relevant features to the prediction. The six connected points (as shown in Fig. 3) are selected and their properties of grid values of LIGSITE and SURFNET, interaction potential, and conservation score are used as the attributes. Together with the distance of the selected grid point from protein, there are totally 29 features assigned as the attributes of each selected grid point.

IV. Methodology

A. Overall Process

In this paper, multi-clustering acts as a unsupervised pre-training process to improve the prediction result. The protein attributes are first divided into three types, including geometry-based, energy-based, and sequence conservation. For the geometry-based type, the attributes consist of the grid values of

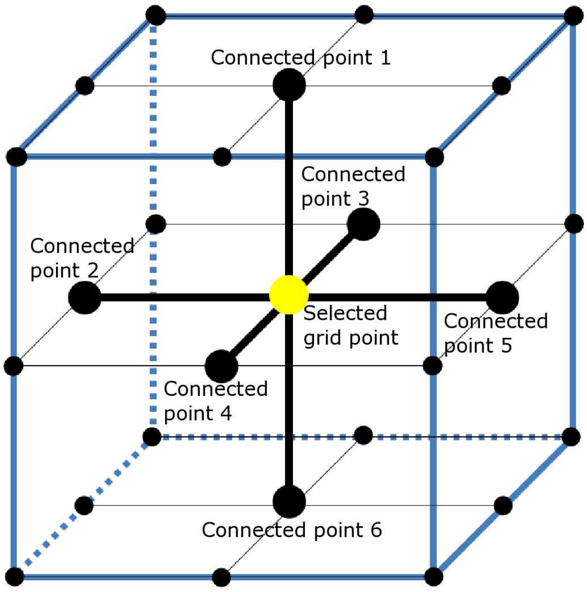

Fig. 3. Six connected grid points of a selected grid point. All the spots in the graph represent the grid points. The middle one is the selected grid point and the larger black spots are the connected grids points: their properties are used as the attributes of the classification.

LIGSITE and SURFNET, distance from protein, and the grid values of the six connected grid points. For the energy-based type, the attributes consist of the interaction potential and that of the six connected grid points. For the sequence conservation type, the attributes consist of the conservation score and that of the six connected grid points.

K-means clustering [32] is then applied to cluster the training data into two regions for each type of attributes, and each region contains half of the number of data. (Therefore, K=2 in this case.) Only one type of attributes is used for each clustering, while the other types of attributes are set to zero for simplicity. As we have three types of attributes, a 3-bit binary code can be assigned and totally eight regions of clustered data are formed. The centroid of each region is calculated. Fig. 4 shows an example of the multi-clustering. SVM is then applied to the training data of each region to form eight classification models of binding sites.

Both the learning and classifying process of SVM are used in the SVM^{light} program. For the testing datasets, each testing protein is also built with the 29 attributes. The grid points of each testing protein are clustered into 8 regions based on the centroids calculated in the training set. The grid points are classified by the corresponding models to identify whether they are potential binding sites. The potential binding sites are then clustered into different groups by K-means clustering, where the initial value of K depends on the number of potential binding sites. The value of K will decrease if empty clusters are formed during the clustering process. After clustering, each group is represented by a centroid that corresponds to an identified binding site. Fig. 5 shows the overall process of the proposed prediction method.

B. Datasets

In this paper, the training set is the same as the one in our previous study [22], which is 15% of the LigASite (v9.4) dataset (40 proteins) as shown in Table I.

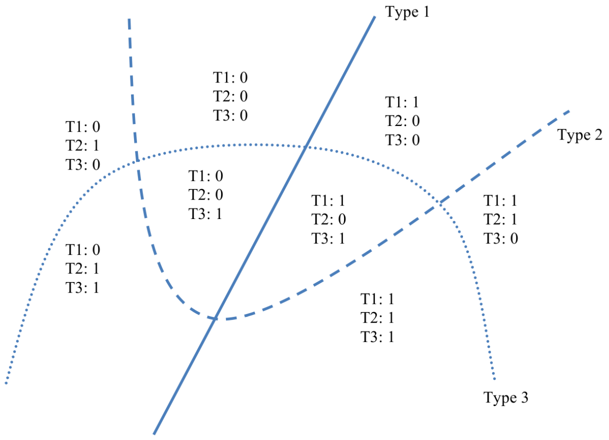

Fig. 4. Example of multi-clustering. The three lines represent the clustering border of different types of attributes and separate the data into eight regions.

TABLE I. TRAINING DATA SET.

1pkj	3gd9	1lf3	3lem	1llo
1ybu	4tpi	3h72	2j4e	1rn8
2v8l	1x2b	1g97	2zhz	3a0t
1o26	1rzu	1znz	1ojz	1sqf
2gga	3gh6	3d1g	2jgv	1dy3
1jyl	2e1t	2ywm	1kwc	2g28
3d4p	2wyw	2dtt	1tjw	2za1
2art	1u7z	3gid	1i1h	2w1a

V. EVALUATION

To evaluate and compare our method to the others, the same performance measurement should be used. [17] proved that most of ligands bind to large pockets. Therefore, they suggested an evaluation method for comparing the top three largest sites only. After the grid points of binding sites are predicted by SVM, the top three largest sites are selected [17] and each site is represented by a grid point in the centre of it. Then, if the centre grid points of the three largest predicting sites are located inside the real binding sites (i.e. the distance between the centre grid points and any atoms of the ligand is within 4Å), the prediction will be counted as a hit, i.e. the binding site is identified correctly. There are sometimes more than one binding site within a protein. A prediction is counted as a hit if at least one binding site in the given protein can be located correctly. Using the same approach in [17], the top 1 to top 3 binding sites are evaluated separately. The success rate is calculated by the following equation to compare the performance of different methods:

$$success\_rate = \frac{N_{HIT}}{N_P} \qquad (2)$$

where N_{HIT} is the number of proteins that at least one binding sites can be located correctly and N_P is the total number of proteins in the dataset.

VI. RESULTS

This section shows the comparison of our method and the other prediction methods. Our method in this paper is named as MCSVMBs and our previous method in [22] is named as SVMBs. The 198 drug-target protein complexes [17] are used as the testing dataset. MCSVMBs is compared with other methods, based on the evaluation of the top three largest binding sites. They are LIGSITE^{CSC}, SURFNET, Fpocket, Q-SiteFinder, ConCavity, and MetaPocket. LIGSITE and PocketFinder are not applied in this experiment since LIGSITE^{CSC} and Q-SiteFinder are the extension of them respectively. All LIGSITE^{CSC}, SURFNET, and Fpocket use geometric characteristics to predict the ligand binding site. Q-SiteFinder uses energy criteria and ConCavity uses both geometric and sequence conservation properties to do the

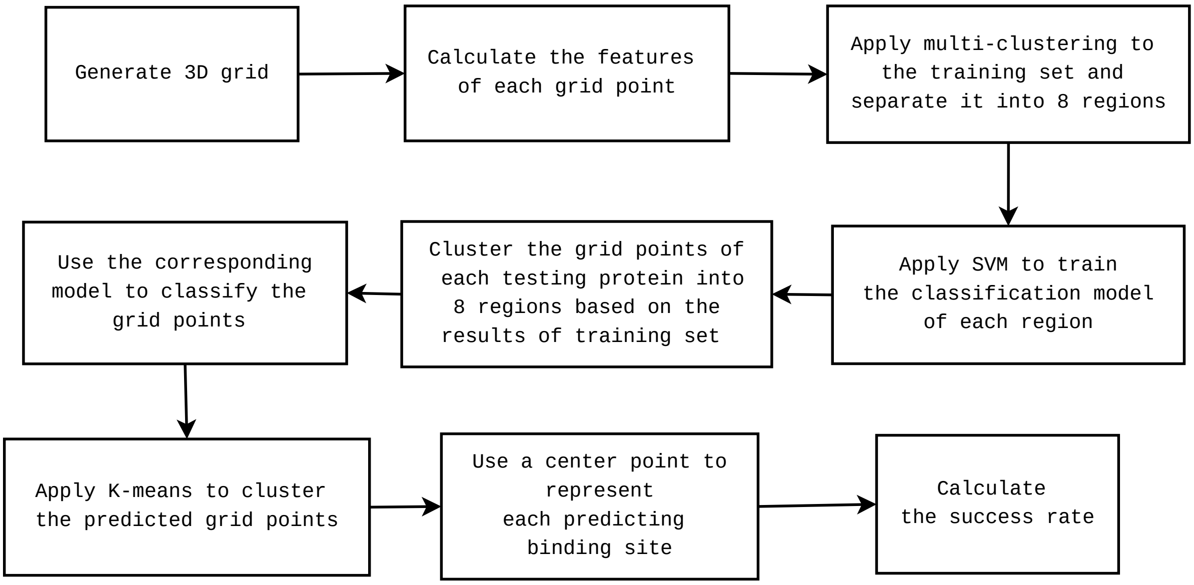
Generate 3D grid
Calculate the features of each grid point
Apply multi-clustering to the training set and separate it into 8 regions
Apply SVM to train the classification model of each region
Cluster the grid points of each testing protein into 8 regions based on the results of training set
Use the corresponding model to classify the grid points
Apply K-means to cluster the predicted grid points
Use a center point to represent each predicting binding site
Calculate the success rate

Fig. 5. Flowchart for prediction of protein-ligand binding site.

prediction. MetaPocket predicts the binding site by combining eight other approaches.

The success rate of this experiment is calculated by (2). The prediction results of top 1 to top 3 binding sites for all approaches are evaluated separately. Table II shows the prediction results of MCSVMBs and the other seven approaches on the 198 drug-target dataset. MCSVMBs can achieve the highest success rate among all the methods. Table III shows the number of hit proteins among the seven methods on the drug-target dataset. The results of SVMBs are different from [22] since the evaluation method is different. There are 123 proteins that can have the binding sites correctly identified as the top 1 prediction. There are 33 and 14 proteins that can have the binding sites correctly identified as the top 2 and top 3 predictions respectively. There are 28 proteins that no associated binding sites can be identified correctly in the top 1-3 predictions. Our method can identify the highest number of binding sites among all methods.

VII. Conclusion

The identification of binding sites (pockets) is the pre-requisite for protein-ligand docking and an important step of structure-based drug design. The prediction of the protein-ligand binding site has been investigated in this paper. SVM is employed to identify the binding sites. It makes use of the attributes of geometric characteristics, interaction potential, distance from protein, conservation score and the grid points nearby to do the identification. Threshold assignment is no longer needed to determine the pockets. Distance filter and random under-sampling are also employed to reduce the effect of large data size and imbalanced data respectively.

Our approach is compared to LIGSITE^{CSC}, SURFNET, Fpocket, Q-SiteFinder, ConCavity, and MetaPocket on the 198 drug-target protein complexes. Only the top three largest binding sites are considered and each site is represented as one centre grid point. The results show that our approach performs better than the other approaches and predicts the binding sites correctly at 62.1% for top 1 prediction, 78.8% for top 1-2 prediction, and 85.9% at top 1-3 prediction. The binding sites identification can be treated as a preliminary step of the docking process. This study can be further developed in the application of ligands finding by virtual screening, docking or de novo drug design.

TABLE III. Number of Hit Proteins on 198 Drug-Target Dataset.

Method	***Top 1***	***Top 2***	***Top 3***	***None***
MCSVMBs	123	33	14	28
SVMBs	122	30	10	36
MetaPocket	121	17	9	51
LIGSITE^{CSC}	95	18	7	78
SURFNET	46	11	8	133
Fpocket	61	34	17	86
Q-SiteFinder	79	28	16	75
ConCavity	93	12	6	87

TABLE II. Success Rate (%) of Top 3 Binding Sites Predictions on 198 Drug-Target Dataset.

Method	***Top 1***	***Top 1-2***	***Top 1-3***
MCSVMBs	**62.1**	**78.8**	**85.9**
SVMBs	61.6	76.8	81.8
MetaPocket	61	70	74
LIGSITE^{CSC}	48	57	61
SURFNET	24	30	34
Fpocket	31	48	57
Q-SiteFinder	40	54	62
ConCavity	47	53	56

REFERENCES

[1] K. Qu and N. Brooijmans, “Structure-based drug design,” in *Computational Methods for Protein Structure Prediction and Modeling*, Y. Xu, D. Xu, and J. Liang, Eds. Springer New York, 2007, pp. 135–176.

[2] I. Kuntz, “Structure-based strategies for drug design and discovery,” *Science*, vol. 257, pp. 1078–1082, 1992.

[3] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, “The protein data bank,” *Necleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[4] K. Henrick and J. Thornton, “PQS: a protein quaternary structure file server,” *Trends in Biochemical Sciences*, vol. 23, no. 9, pp. 358–361, Sept. 1998.

[5] S. Kalyaanamoorthy and Y. Chen, “Structure-based drug design to augment hit discovery,” *Drug Discovery Today*, vol. 16, no. 17–18, pp. 831–839, 2011.

[6] A. Laurie and R. Jackson, “Methods for the prediction of proteinligand binding sites for structure-based drug design and virtual ligand screening,” *Current Protein and Peptide Science*, vol. 7, no. 5, pp. 395–406, Oct. 2006.

[7] S. Liang, C. Zhang, S. Liu, and Y. Zhou, “Protein binding site prediction using and empirical scoring function,” *Nucleic Acids Research*, vol. 34, pp. 3698–3707, 2006.

[8] T. Magliery and L. Regan, “Sequence variation in ligand binding sites in proteins,” *BMC Bioinformatics*, vol. 6, no. 1, p. 240, 2005.

[9] D. Levitt and L. Banaszak, “POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids,” *Journal of Molecular Graphics*, vol. 10, pp. 229–234, 1992.

[10] M. Hendlich, F. Rippmann, and G. Barnickel, “LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins,” *Journal of Molecular Graphics and Modelling*, vol. 15, no. 6, pp. 359–363, 1997.

[11] R. Laskowski, “SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions,” *Journal of Molecular Graphics*, vol. 13, pp. 323–330, 1995.

[12] J. An, M. Totrov, and R. Abagyan, “Pocketome via comprehensive identification and classification of ligand binding envelopes,” *Molecular and Cellular Proteomics*, vol. 4, pp. 752–761, 2005.

[13] A. Laurie and R. Jackson, “Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites,” *Bioinformatics*, vol. 21, pp. 1908–1916, 2005.

[14] B. Huang and M. Schroeder, “LIGSITEcsc: predicting ligand binding sites using the connolly surface and degree of conservation,” *BMC Structural Biology*, vol. 6, no. 1, p. 19, 2006.

[15] J. Capra, R. Laskowski, J. Thornton, M. Singh, and T. Funkhouser, “Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure,” *PLoS Computational Biology*, vol. 5, no. 12, 2009. [Online]. Available: http://compbio.cs.princeton.edu/concavity

[16] B. Huang, “Metapocket: a meta approach to improve protein ligand binding site prediction,” *Journal of Integrative Biology*, vol. 13, no. 4, pp. 325–330, 2009.

[17] Z. Zhang, Y. Li, B. Lin, M. Schroeder, and B. Huang, “Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction,” *Bioinformatics*, vol. 27, no. 15, pp. 2083–2088, 2011.

[18] G. Brady and P. Stouten, “Fast prediction and visualization of protein binding pockets with pass,” *Journal of Computer-Aided Molecular Design*, vol. 14, pp. 383–401, 2000.

[19] T. Kawabata, “Detection of multi-scale pockets on protein surfaces using mathematical morphology,” *Proteins*, vol. 78, pp. 1195–1121, 2010.

[20] V. Guilloux, P. Schmidtke, and P. Tuffery, “Fpocket: An open source platform for ligand pocket detection,” *BMC Bioinformatics*, vol. 10, no. 1, p. 168, 2009.

[21] J. Yu, Y. Zhou, I. Tanaka, and M. Yao, “Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere,” *Bioinformatics*, vol. 26, pp. 46–52, 2010.

[22] G. Wong, F. Leung, and S. Ling, “Predicting protein-ligand binding site using support vector machine with protein properties,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 6, pp. 1517–1529, 2013.

[23] A. Ben-Hur, C. Ong, S. Sonnenburg, B. Sch¨olkopf, and G. R¨atsch, “Support vector machines and kernels for computational biology,” *PLoS Computational Biology*, vol. 4, no. 10, 2008.

[24] X. Cai, F. Nie, and H. Huang, “Multi-view k-means clustering on big data,” in *Processdings of the 23rd International Joint Conference on Artificial Intelligence*, 2013, pp. 2598–2604.

[25] D. Cai, C. Zhang, and X. He, “Unsupervised feature selection for multicluster data,” in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 333–342.

[26] D. Meng, Y. Lee, and Z. Xu, “Passage method for nonlinear dimensionality reduction of data on multi-cluster manifolds,” *Pattern Recognition*, vol. 46, pp. 2175–2186, 2013.

[27] Y. Bengio, “Learning deep architectures for ai.” *Fundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[28] D. Erhan, Y. Bengio, A. Courville, P. Manzagol, and P. Vincent, “Why does unsupervised pre-training help deep learning?” *Journal of Machince Learning Research*, vol. 11, pp. 625–660, 2010.

[29] W. Valdar, “Scoring residue conservation,” *Proteins: Structure, Function, and Genetics*, vol. 48, no. 227–241, 2002.

[30] K. Wang and R. Samudrala, “Incorporating background frequency improves entropy-based residue conservation measures,” *BMC Bioinformatics*, vol. 7, no. 1, p. 385, 2006.

[31] J. Capra and M. Singh, “Predicting functionally important residues from sequence conservation,” *Bioinformatics*, vol. 23, pp. 1875–1882, 2007. [Online]. Available: http://compbio.cs.princeton.edu/conservation

[32] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.