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 identifying protein-ligand binding in structure-based drug design. Then, the Support Vector Machine (SVM) is employed to classify the sites most likely for binding ligands. Three types of attributes are used, namely geometry-based, energy-based, and sequence conservation. Comparison is made on 198 drug-target protein complexes with LIGSITE^{CSC}, SURFNET, Fpocket, Q-SiteFinder, ConCavity, and MetaPocket. The results show an improved success rate of up to 86%.**

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

I. INTRODUCTION

The process of structure-based drug design (SBDD) [2] requires the identification and validation of the target protein for searching the causes of the disease phenotype. Protein has a critical role for the symptoms of diseases. On functioning, a drug activates or inhibits the protein by modifying it for a curing effect [1]. After the relation between the disease and target protein has been found, the next step is to find the method of modifying the target. It is generally referred to as protein-protein or protein-ligand (small chemical molecule) interactions.

SBDD considers a protein's three-dimensional (3D) structure, which can be found experimentally by Nuclear Magnetic Resonance (NMR) spectroscopy or x-ray crystallography. The protein structure can also be constructed based on its amino acid sequence and a similar protein of which the 3D structure is known. Public domains that contain this information include the Protein Data Bank (PDB) [3] (that tells the atomic coordinates) and Protein Quaternary Structure file server (PQS) [4]. They enhance the feasibility of SBDD as the knowledge of some protein's 3D structures enables the prediction of binding sites for protein and ligand, which is a pre-requisite of SBDD [5]. When a protein's structure is known and its binding sites are predicted, finding a suitable ligand (the drug) can be realised by docking, de novo drug design, or virtual screening [6].

Protein-ligand binding sites usually are found in the pockets of the protein surfaces. The prediction of pockets as binding sites has been studied based on the proteins' structure or sequence. The geometric properties are considered in POCKET [9], LIGSITE [10], and SURFNET [11]. It is often assumed that the binding site is found in the largest pocket. By analysing the sequence conservation, the residues involved in ligand binding were predicted [7]–[8]. Other methods like PocketFinder [12] and Q-SiteFinder [13] consider the energy of the van der Waals interaction potential. Nevertheless, the above methods are not so capable of handling the multi-chain proteins, where the gaps among the protein chains might be predicted as pockets incorrectly. LIGSITE^{CSC} [14] and ConCavity [15] suggested that the sequence conservation should be integrated with the structural pocket identification to predict the binding sites of proteins, especially the multi-chain ones, more accurately. MetaPocket [16-17] is a combination of eight predictors, namely LIGSITE^{CSC} [14], PASS [18], Q-SiteFinder [13], SURFNET [11], GHECOM [19], ConCavity [15], Fpocket [20], and POCASA [21]. It ranks the predicted binding sites of the eight methods and finds the potential binding sites based on their spatial similarity.

Predicting protein-ligand binding sites can be formulated as a binary classification problem to evaluate how likely some grid points around the protein surface will bind with ligands. A score for each grid point is obtained based on the protein properties. However, the many methods to determine these scores increase the complexity of the problem. Thus, we applied the Support Vector Machine (SVM) in our previous work [22] to do the prediction. A total of 29 attributes of proteins were used, which cover all the aforementioned protein information. Nevertheless, like many bioinformatics datasets, the data of binding sites have the problems of being large-sized and imbalanced [23]. Thus, the data size in the previous work was reduced by random under-sampling and filtering.

In this paper, the unsupervised pre-training process of multi-clustering is proposed to further improve our previous prediction method [22]. Multi-clustering has been widely used in different areas, including big data [24], feature selection [25], data reduction [26], and deep 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 to generate eight classification models. By applying multi-clustering in this way, we can focus on a particular type of attribute to design the SVM, which makes the SVM in each group to be simpler and have better performance when compared with an SVM designed for data in all groups for single clustering. SVM is used as the classifier

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

because it can offer more robust performance than other conventional classifiers based on our experimental results. The 198 drug-target dataset developed in MetaPocket [17] is used to evaluate our method. Only the largest three binding sites are predicted. Each site is denoted by a centre point. The proposed method is compared with six other published methods, namely LIGSITECSC, SURFNET, Fpocket [20], Q-SiteFinder, Con-Cavity, and MetaPocket. A new evaluation method, which is different from that in our previous work, is applied. It is closer to the method in [17] so that the comparison results are more meaningful.

The organization of this paper is as follows. The methods of predicting protein-ligand binding sites are described in Section II. Section III explains the attributes we consider for each grid point. In Section IV, the overall process and the selected training data are detailed. The proposed evaluation method is explained in Section V. Section VI presents the results from the proposed method. The paper's conclusion is given in Section VII.

II. Predicting Sites for Protein-Ligand Binding

A. POCKET and LIGSITE

POCKET [9] is a geometry-based method for predicting the binding sites, which involves a 3D grid as shown in Fig. 1. A distance check assures that the grid points will not overlap with the protein atoms. Those grid points not overlapping with the protein atoms are labelled under solvent. A protein-solvent-protein (PSP) event occurs if a grid point in solvent is enclosed by pairs of protein atoms in opposite directions of the same axis.

As an extension to POCKET [9], LIGSITE [10] has increased directions of scanning. The number of PSP events of a grid-point in the scanning directions is obtained (Fig 1). A larger value of it represents a higher chance of that grid point being a pocket. This method considers only the geometric properties of the target protein.

B. SURFNET

SURFNET [11] is another geometry-based method to find the binding sites. Its grid values are based on the number of constructed spheres within pairs of relevant protein atoms that has no overlapping occurred (Fig. 2). It can be seen that for grid points outside some pockets, the distances between pairs of atoms are very long. On the other hand, for grid points inside pockets, more than one sphere can be formed.

C. PocketFinder

PocketFinder [12] is an energy-based method to predict ligand binding sites. The van der Waals interaction potential at a grid point p between the protein and a simple atomic probe is given by the Lennard-Jones formula:

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

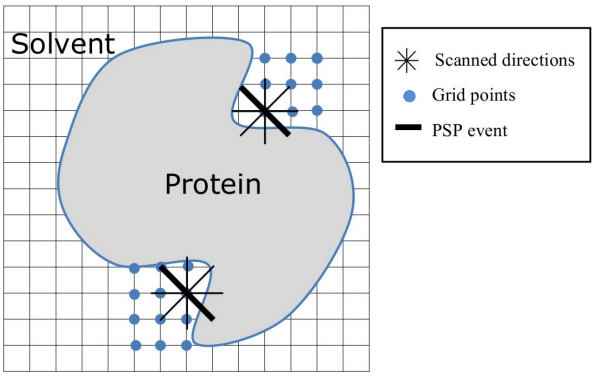

Fig. 1. Geometric feature of PSP event of a grid point.

where C_{12}^{i} and C_{6}^{i} are the typical 12-6 Lennard-Jones parameters that model the van der Waals interaction energy between protein atom i and a carbon atom placed at the grid point p at a distance of r_{pi}; 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.

D. Sequence Conservation

As residues in protein are not equally important, conservation analysis can be used to predict those residues in the protein sequence that are functionally important [29]-[31].

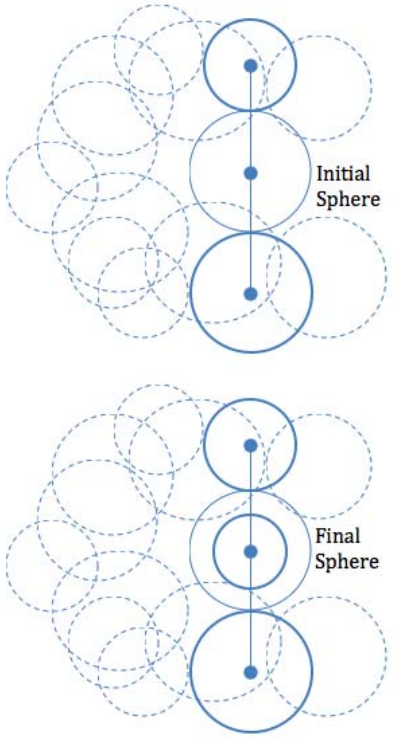

Fig. 2. SURFNET. Solid line circles are the concerned protein atoms with the constructed sphere of a grid point between them. The constructed sphere's radius decreases until no overlapping occurs.

It has been shown that sequence conservation has a strong correlation with ligand binding sites [7]-[8]. In [15], the sequence conservation is combined with the information of the protein structure to predict the binding sites.

III. Protein Attributes Used

The protein attributes considered by this paper are based on the common approaches of binding site prediction described in the previous section. A total of 29 attributes [22] are used for the training of the SVM, which is then tested for identifying the sites of protein-ligand binding. For a given grid-point, these attributes (features) include:

1) Grid values: The geometric property values according to LIGSITE and SURFNET.

2) Interaction potential: The PocketFinder method [12] that applies (1) to find the interaction potential.

3) Conservation score: Residue-level analysis was done to identify which residues in a protein are responsible for its function, which is reflected by a conservation score obtained by the Jensen-Shannon divergence (JSD) method [31]. Each grid point's score takes the conservation score of the nearest residue.

4) Distance to protein: It is the square of the distance between the nearest point on the van der Waals surface of the protein and the grid point. However, the grid points with the squared distance larger that 5Å are not considered as it has been found experimentally that nearly 90% of ligand atoms have a distance less than 5Å from the protein's van der Waals surface. By doing so, the data size can be reduced.

5) Attributes of nearby grid points: Binding sites practically involve many grid points (as the distance between two grid points is only 1Å [15]). The attributes of the nearby grid points are important for a good prediction. As shown in Fig. 3, the six connected points' properties of LIGSITE value, SURFNET value, interaction potential, and conservation score are used as the attributes too. Together with the distance of the selected grid point, totally 29 features are used as the attributes of each selected grid point.

IV. Methodology

A. Overall Process

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

The SVM^{light} program is used to realise the learning and classifying process. The radial basis function, which is commonly used to handle non-linear classification, is employed as the kernel of the SVM. On applying the trained SVM to do the testing, each protein in the dataset 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 used in our previous study [22], which contains 15% of the LigASite (v9.4) dataset (40 proteins) as shown in Table I.

V. Evaluation and Comparison

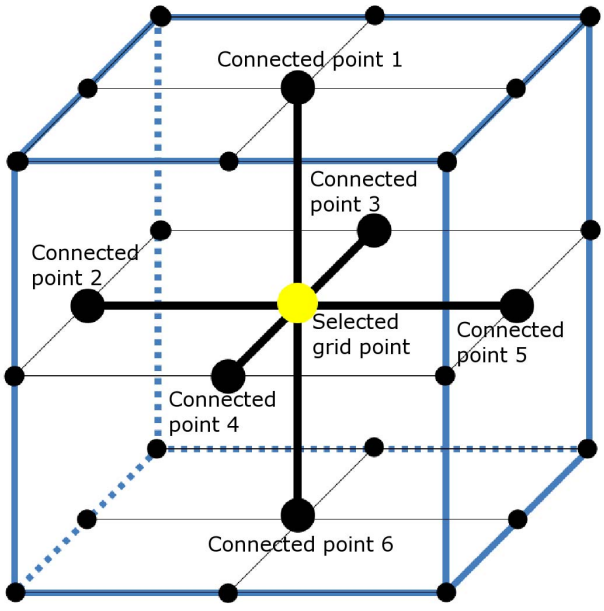

Fig. 3. Centre grid point and its six connected grid points.

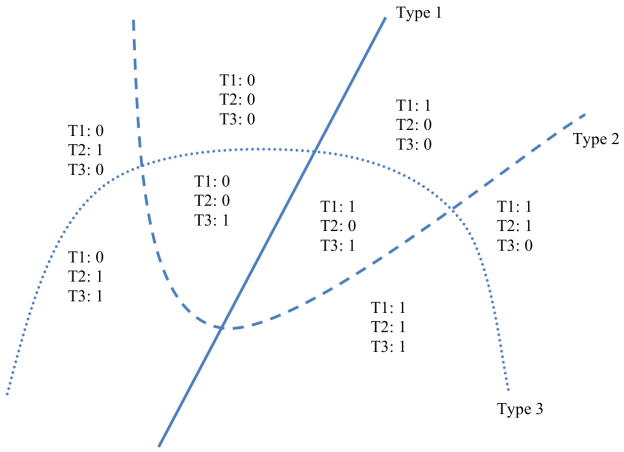

Fig. 5. Example of multi-clustering. The three lines represent the clustering border of different types of attributes and separate the data into eight regions.

The same performance measurement should be used to do the evaluation and comparison. [17] suggested an evaluation method for comparing only the largest three sites as most of the ligands would bind to large pockets only. After the SVM has predicted the binding sites grid points, the largest three sites are selected and represented by the grid points at the centre of them. Then, if the centre grid points of the three largest predicting sites are located inside the real binding sites (such that the distance between any atom of the ligand and the centre grid point is less than 4Å), the prediction will be counted as a hit (correct identification). A protein can have more than one binding site. If at least one binding site can be located correctly, the prediction is counted as a hit. Using the same evaluation method in [17], the top 1 to top 3 binding sites are obtained separately and the success rate is given by,

TABLE I. PROTEINS USED AS TRAINING DATASET.

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

$$success\_rate = \frac{N_{HIT}}{N_P} \tag{2}$$

where N_P is the total number of proteins in the testing dataset, N_{HIT} is the number of proteins with at least one binding sites correctly located.

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 testing dataset is the 198 drug-target protein complexes [17]. Based on the evaluation of the largest three binding sites, MCSVMBs is compared with other methods of LIGSITE^{CSC}, SURFNET, Fpocket, Q-SiteFinder, ConCavity, and MetaPocket. LIGSITE and PocketFinder are not considered as LIGSITE^{CSC} and Q-SiteFinder have been the extension of them respectively. It should be noted that LIGSITE^{CSC}, SURFNET, and Fpocket use geometry-based type attributes, Q-SiteFinder uses energy-based type attributes, and ConCavity uses both geometry-based type and sequence

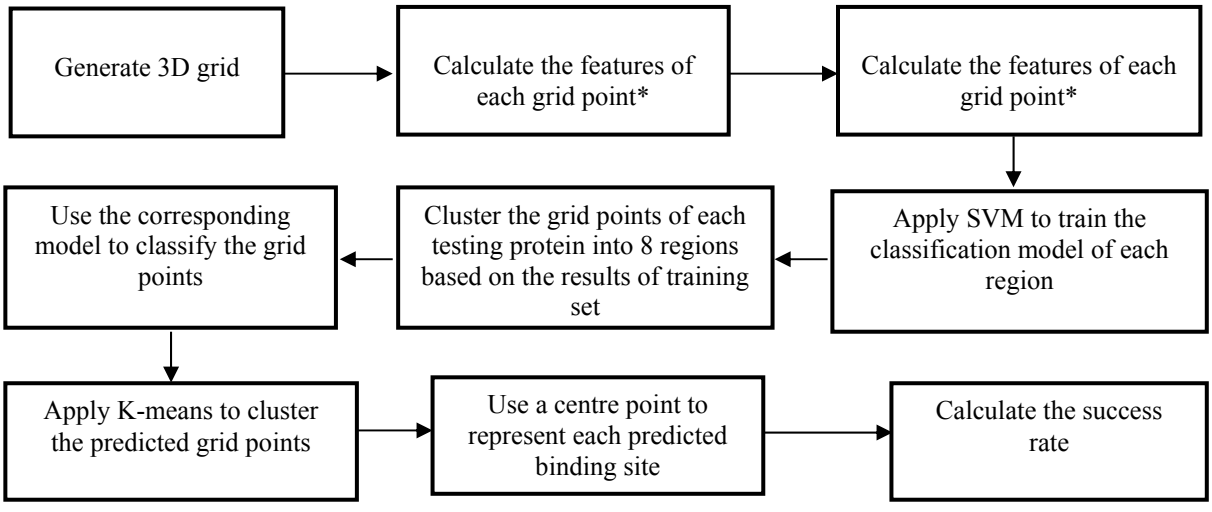

Fig. 4. Process of protein-ligand binding site prediction.

conservation type attributes to do the prediction. MetaPocket combine eight other methods to do the prediction.

The prediction results of top 1 to top 3 binding are evaluated separately, which are summarised in Table II. It can be seen that MCSVMBs gives the highest success rate. Table III lists the number of hit proteins. The results of SVMBs are different from [22] because the evaluation method is different. A total of 123, 33 and 14 proteins have the binding sites correctly identified as the top 1, top 2 and top 3 predictions respectively. Totally 28 proteins are not associated with any identified binding sites in the top 1-3 predictions. It can be seen that MCSVMBs identifies the largest number of binding sites.

VII. CONCLUSION

In structure-based drug design, identifying the binding sites is a pre-requisite for protein-ligand docking. In this paper, SVMs in eight clusters of data have been used to identify the binding sites. The geometry-based type, energy-based type and sequence conservation type attributes are considered on doing the multi-clustering identification. Assigning threshold value is no longer needed to determine the binding sites. Random under-sampling and distance filtering have been employed to reduce the effect of imbalanced data and large data size respectively.

MCSVMBs is compared to six other published methods using the 198 drug-target protein complexes. Only the largest three binding sites are considered. The results show that MCSVMBs performs better than the other methods.

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

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

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. 2, pp. 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.