Predicting Protein-Ligand Binding Site using Support Vector Machine with Protein Properties

Ginny Y. Wong, *Student Member, IEEE*, Frank H.F. Leung, *Senior Member, IEEE*, and S.H. Ling, *Senior Member, IEEE*

Abstract

Identification of protein-ligand binding site is an important task in structure-based drug design and docking algorithms. In the past two decades, different approaches have been developed to predict the binding site, such as the geometric, energetic and sequence-based methods. When scores are calculated from these methods, the algorithm for doing classification becomes very important and can affect the prediction results greatly. In this paper, the Support Vector Machine (SVM) is used to cluster 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, LIGSITE^{csc}, SURFNET, Fpocket, PocketFinder, Q-SiteFinder, ConCavity, and MetaPocket on the dataset LigASite and 198 drug-target protein complexes. The results show that our approach improves the success rate from 60% to 80% at AUC measure and from 61% to 66% at top-1 prediction. Our method also provides more comprehensive results than the others.

Index Terms

Ginny Y. Wong and Frank H.F. Leung are with the Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong. (e-mail: ginnyyk.wong@connect.polyu.hk; frank-h-f.leung@polyu.edu.hk).

S.H. Ling is with the Centre for Health Technologies, Faculty of Engineering and Information Technology, University of Technology Sydney, NSW, Australia. (e-mail: steve.ling@uts.edu.au).

bioinformatics, protein-ligand binding sites, binding sites predication, structure-based drug design, support vector machine (SVM).

I. Introduction

Bioinformatics has been an active research area in the last three decades. It concerns the application of computational methods to biological problems. Recently, prediction has become a very popular issue in bioinformatics. Example problems include gene's structure prediction [1] , function prediction [2], protein functional site identification [3]–[4], and disease classification [5]. Computational approaches have started to take an important role in bioinformatics for the past two decades because of their ability to handle a huge data size at a relatively low cost, and the rapid increase of computational power. The identification of protein-ligand binding site using computational approaches is the focus of this paper.

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 [6]. After the relationship between the target and disease has been found, the next operation of drug discovery is found 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 the protein, structure-based drug design (SBDD) attempts to contribute to drug discovery [7]. The 3D structure of 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) [8] or Protein Quaternary Structure file server (PQS) [9], which show the atomic coordinates and the quaternary structure of protein 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 prerequisite of SBDD [10]. One famous example use of SBDD is the inhibition of the HIV protease. The drug is highly effective against HIV [11]. Different approaches can be applied to find the ligand, such as virtual screening, docking and de novo drug design, when the protein's

structure is known [12].

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 analyzed to predict the residues involved in ligand binding [13]–[14]. The structural information includes the studies of geometry and interaction energy of proteins. In POCKET[15], LIGSITE[16], and SURFNET[17], the studies only use the geometric characteristics and believe that the binding site is usually located in the largest pocket. On the other hand, some methods like PocketFinder[18] and Q-SiteFinder[19] 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 protein. The methods may treat the gaps among the chains of protein as pockets incorrectly. Therefore, LIGSITE^{csc}[20] and ConCavity[21] suggeste that the sequence conservation should be integrated with the structural pocket identification to get the more accurate binding sites of proteins, particularly the multi-chain proteins.

There is a drawback when binding sites are identified from the above approaches. The binding sites predictions of each method were based on different scores, which were calculated from the corresponding protein characteristics. The simplest method was setting a threshold to help determining the binding sites [16]. If the score of a point was greater than the threshold, that point would be identified as the binding site. In [18], mean and standard deviation of the scores were considered on finding the threshold. The results of these approaches are easily affected by the grid format and the threshold needed to be set carefully; otherwise the results would not be satisfactory. Machine learning techniques have been widely applied in bioinformatics and have shown satisfactory performance in the binding site prediction [22]–[25]. In this paper, support vector machine (SVM) is proposed for handling this problem. Moreover, SVM [26]–[29] has shown its high applicability and advantage on classifying high-dimensional and large datasets in [30]–[31].

The prediction of the binding site can be formulated as a problem of binary classification: discriminating whether a location is likely to bind the ligand or not. SVM is one of the tools that use supervised learning for doing classification. It mainly applies two techniques to solve the classification problem: the formulation of a

large-margin hyperplane and the use of a kernel function. SVM can construct an $(n-1)$-dimensional hyperplane in an n-dimensional space to separate the data, where each datum is represented by an n-dimensional vector.

We train the SVM to generate the hyperplane by using 29 proteins' attributes, including the geometric characteristics, interaction energy, sequence conservation, distance from protein, and the properties of the surrounding grid points. A radial basis function (RBF) is used as the SVM kernel since a non-linear classification model is needed and RBF is a common kernel to handle this problem. Like most of the datasets in bioinformatics, the data of the binding sites have the problem of being imbalanced and in large data scales [32]. Therefore, down sampling and filtering are also be applied to reduce the data size.

Two experiments are used to evaluate our approach. The first one uses LigASite [33] as the dataset which is suggested in ConCavity. The predicted binding sites are represented as grid points in this experiment. Our approach is compared with four other methods. They are LIGSITE, SURFNET, PocketFinder and ConCavity. The other experiment uses 198 drug-target dataset which is developed in MetaPocket [34]. Only the top three largest binding sites are predicted and represented as one center point of each site in this experiment. Our approach is compared with six other methods. They are LIGSITE^{csc}, SURFNET, Fpocket [35], Q-SiteFinder, ConCavity, and MetaPocket. Two different measurements are applied since the representations of the binding sites are different in these experiments.

This paper is organized as follows. In Section II, the prediction methods for binding sites with consideration of the proteins' sequence and geometrical structure, and the problem of the evaluation methods are described. In Section III, the details of SVM and the selected attributes are introduced. The adopted evaluation method is discussed in Section IV. Section V shows the result of our proposed method and a conclusion will be drawn in Section VI.

II. Prediction of Protein-Ligand Binding Site

This section describes the three most common approaches of binding site prediction. Then, the problems of their evaluation methods, which are tackled by our proposed method, are discussed.

A. POCKET and LIGSITE

POCKET [15] 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 labeled 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).

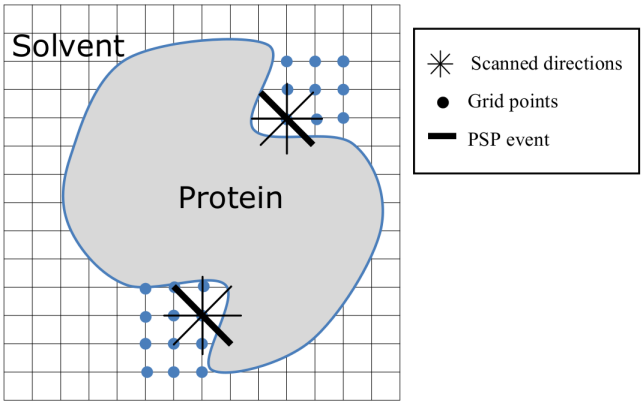


Fig. 1. PSP event used to describe the geometric feature of a grid point. It counts the number scanning directions that the pair of protein atoms can enclose the grid point. For POCKET method, the maximum number of PSP event is three while it is seven for LIGSITE method.

LIGSITE [16] is an extension to POCKET [15] with the scanning directions being different. Both of them considered the identification of PSP events on the basis of atom coordinates. LIGSITE scans for the pockets along three axes and four cubic diagonals while POCKET only scans three axes. 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 [17] 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, the 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 one sphere can be formed.

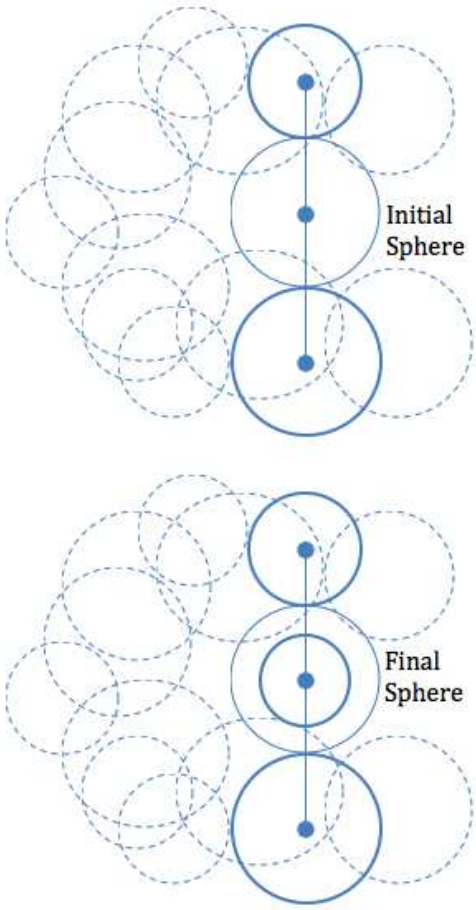


Fig. 2. SURFNET. There are three solid line circles and several dotted line circles in each graph. The top and bottom solid line circles represent the pair of relevant atoms and the middle one shows the constructed sphere of a grid point. The dotted line circles represent the other atoms that surround the testing gris 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.

C. PocketFinder

PocketFinder [18] 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:

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

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 the 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 [36]–[38]. Sequence conservation has also been shown to be strongly correlated with ligand binding sites [13]-[14]. Therefore, [21] suggested combining the sequence conservation and the structure of protein to predict the protein ligand binding sites by weighting every pair of protein atoms.

E. Problems of evaluation methods

There are several evaluation methods to determine the binding site after the corresponding values are calculated by the above approaches. The simplest one is to apply a threshold to the grid point value to determine if the grid points belong to a pocket [16]. This threshold is set to all proteins and does not consider the difference among them. A poor scenario may cluster most of the grid points as pockets if the threshold is too low, or the number of pockets is much smaller than that of binding sites if the threshold is too large.

Another method calculates the mean and standard deviation of the grid points' values to determine the threshold for each protein [18]. Although this approach calculates the threshold for different proteins, the threshold depends

on the grid points' values. If the grids embedded in the protein vary, the mean and standard deviation of the grid points' values will be different. That means, the threshold and the number of pockets could be varying for a particular protein used.

In [21], a binary search for the grid threshold is performed. The binary search produces a culled set of pockets, which have specified properties based on the sizes and shapes of the pockets. When the method iterates, the grid points are adjusted until the set of pockets meet all the properties. Although this approach can consider the sizes and shapes of the pockets, all the grid thresholds are set by the users, and we do not know which values of thresholds are suitable for a given protein.

We make use of the characteristics of the above three approaches, which can describe the properties of the protein-ligand binding sites. To overcome the weakness of the different evaluation methods, we employ the SVM to achieve the goal. The process of SVM is discussed in the following section.

III. Methodology

In order to alleviate the problems mentioned above, the SVM is applied. This section explains the collection of datasets and attributes used in this paper first. The details of the SVM classifier follow. Then the overall flow of our method is described.

A. Datasets

We have used two sets of proteins to evaluate our method. The first one is the non-redundant LigASite (v9.4) dataset [33], which is suggested in [21]. The other one is the 198 drug-target complexes, which are developed in [34]. For the dataset of LigASite, only six main classes of enzyme (categorized for 272 protein complexes) from the dataset are selected. They are transferase, hydrolase, oxidoreductase, lyase, ligase and isomerase, which occupied around 70% of LigASite. Fig. 3 shows the percentages of the number of proteins distributed among these six enzyme classes. Fig. 4 shows the number of chains distributed in the selected proteins of LigASite.

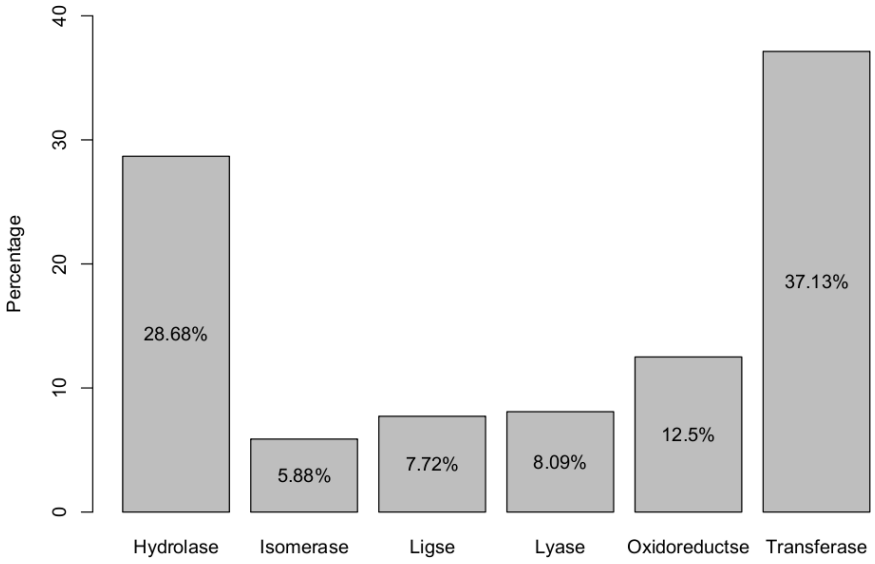


Fig. 3. Percentages of distribution among six enzyme classes. Most of the proteins in LigASite belong to transferase type. The second is hydrolase. The contribution of the other classes are almost the same.

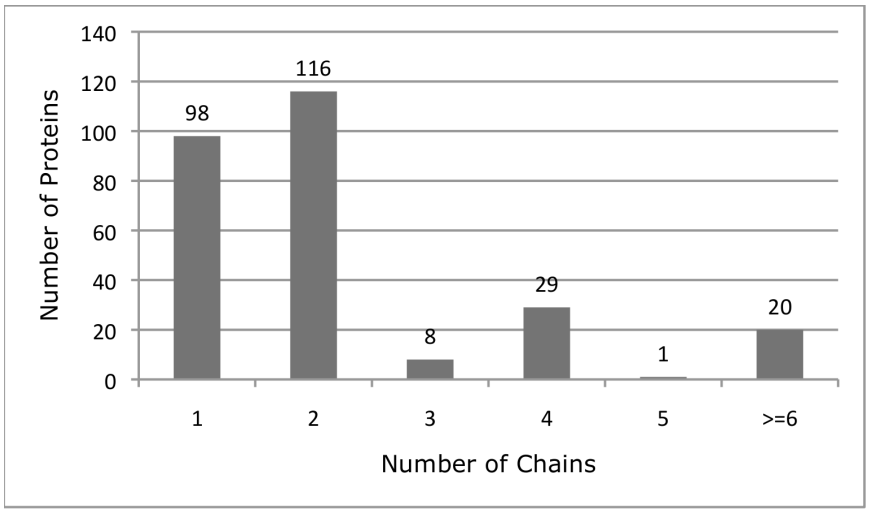


Fig. 4. Distribution of the number of chains in the selected proteins of LigASite. Most of the proteins have less than three chains.

B. Protein Properties Used for Training and Testing

The structure of proteins with bound ligands are obtained from the Protein Data Bank (PDB) [8], which is a collection of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule.

After the structure of each protein is retrieved, the 3D grid is generated by covering the free-space surrounding

the proteins. The program is based on the source of ConCavity, which is available on its website. The attributes of each grid point used in SVM are calculated based on the protein properties in the following:

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.

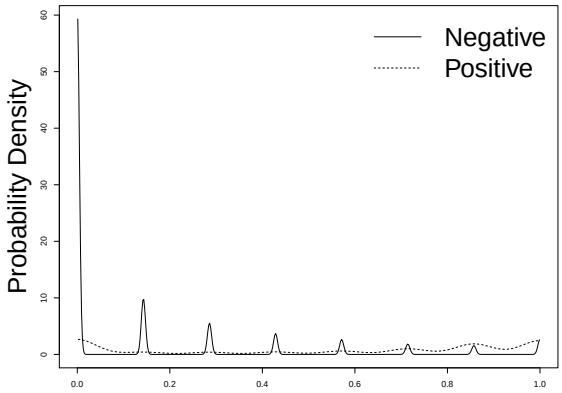
2) Interaction potential: This energy is the same as the van der Waals interaction potential of an atomic probe with the protein [18]. The calculation is done by the PocketFinder method, which is mentioned in Section II. The Lennard-Jones formula (1) is used to estimate the interaction potential between 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 chosen to calculate the score since it has been shown to provide an outstanding performance in identifying residues near bound ligands in [38]. It is an open source program which is freely available in its webpage [38].

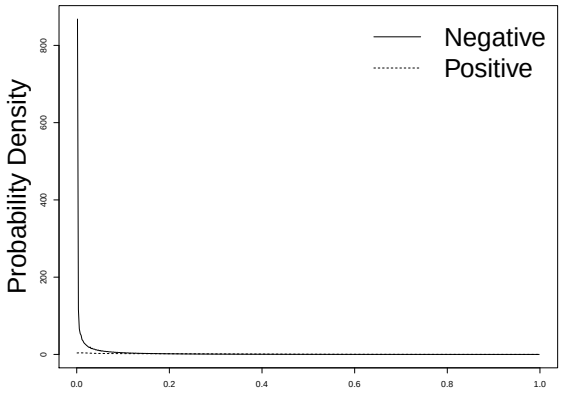
4) Distance from protein: The squared distance from each grid point to the closest point on the van der Waals surface of the protein is calculated. When the grid points are too far from the atoms, they are not likely to be a pocket. In the experiment, almost 90% of ligand atoms are located within 5Å 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.

To explain the relationship between the binding sites and the selected attributes, graphs of probability density for the normalized attribute values are shown in Fig. 5. The solid line represents the corresponding probability density for non-binding sites (negative class) and the dotted line represents the corresponding probability density for binding sites (positive class). Some of the attributes, such as LIGSITE values, SURFNET values, and interaction potential, show a very high density of small values when the grid points are located at non-binding sites. On the other hand, these attributes show a small difference on the density when the grid points are located at binding sites. This difference is shown more clearly in Fig. 6. We can see that the values of these attributes are relevant to the location of the binding sites. However, it is difficult to use only one property to classify the

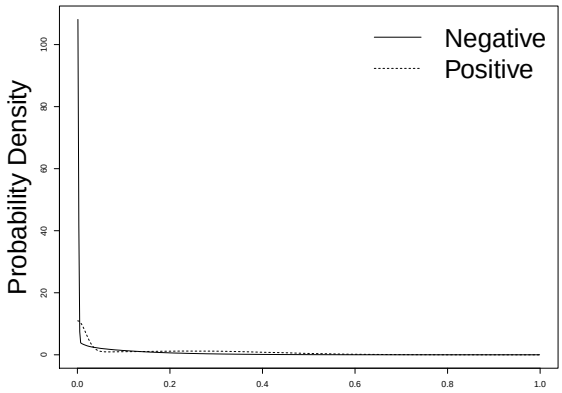
binding sites. Therefore, we propose to use all of them as the features of the training set for an SVM.



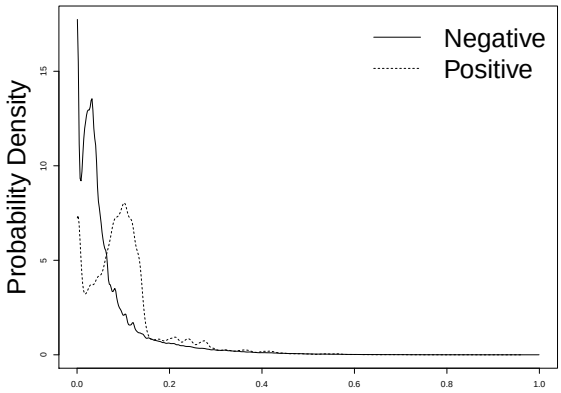
(a) Normalized grid values calculated by LIGSITE.



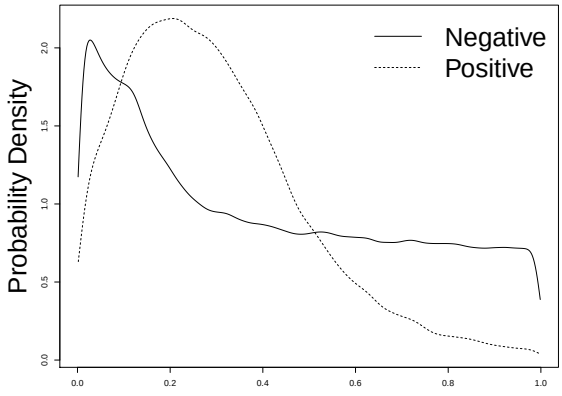
(b) Normalized grid values calculated by SURFNET.



(c) Normalized interaction potential.



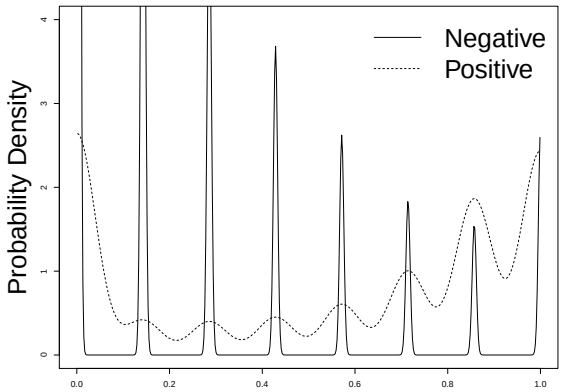
(d) Normalized conservation score.



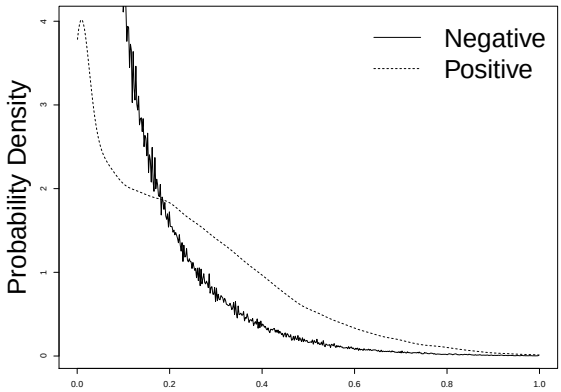
(e) Normalized distance from protein.

Fig. 5. Probability density functions of protein properties.

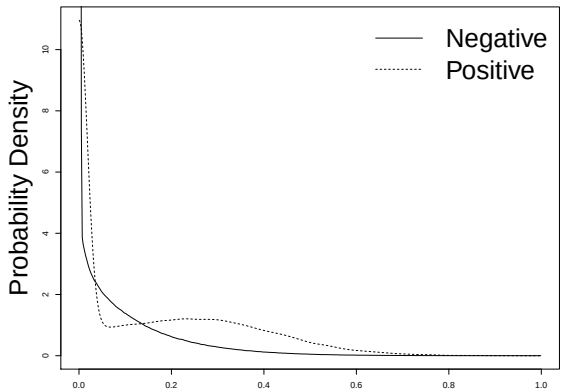
5) Properties of surrounding grid points: All the binding sites are formed by many grid points (the distance between two grid points is 1Å [21]), so the properties of the grid points nearby are also relevant features to the prediction. The six connected points (as shown in Fig. 7) are selected and their properties (1) to (3) as described



(a) Normalized grid values calculated by LIGSITE.



(b) Normalized grid values calculated by SURFNET.



(c) Normalized interaction potential.

Fig. 6. A zoom-in version of Fig. 5 of properties: (a) LIGSITE values, (b) SURFNET values, and (c) Interaction potential.

above are used as the attributes. The point in the middle of the cube in Fig. 7 is the selected grid point to be classified. There are totally 29 features assigned as the SVM attributes.

C. Classification with Support Vector Machine

Machine learning methods have been applied to predict catalytic sites [22]-[39]. In this paper, one of the machine learning tools, the support vector machine (SVM), is employed to predict the protein-ligand binding sites.

To avoid the drawbacks mentioned in Section II, SVM is employed to classify which grid points are most likely to bind the ligands based on the properties of grid values, interaction potential, sequence conservation score, distance from protein, and the surrounding grid points. A common kernel, the radial basis function, is used to construct a non-linear hyperplane. The program called SVM^{light} is used, which is available from its

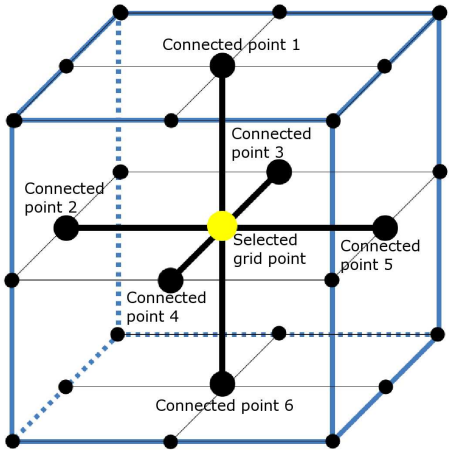


Fig. 7. Six connected grid points of a selected grid point. All the black spots in the graph represent the grid points. The middle one is the selected gris point to be classified and the larger black spots are the connected grids points: their properties are used as the attributes of the classification.

website [40].

SVM basically is a binary classifier. Let a vector $\mathbf{x}$ be denoted by $[x_j]$, $j = 1, \ldots, m$, where m is the number of attributes and $[x_j]$ is a point in an m-dimensional vector space. The notation $\mathbf{x}_i$ is the i-th vector in a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $y_i \in \{-1, 1\}$ is the condition label for a binary classification problem and n is the number of examples (grid points). To construct the SVM, all training samples are first mapped to a feature space by a non-linear function $\phi(\mathbf{x}_i)$. A separating hyperplane in the feature space can be expressed as

$$\begin{aligned} f(\mathbf{x}) &= \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b \\ &= \sum_{j=1}^{m} w_j \phi(x_j) + b \end{aligned} \tag{2}$$

where $\mathbf{w}$ is the weight vector and b is the bias.

The optimal separating hyperplane is defined as a linear classifier which can separate the two classes of training samples with the largest marginal width, and the solution $\boldsymbol{\alpha} = [\alpha_i]$ is obtained by maximizing the following function:

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \tag{3}$$

subject to:

$$\sum_{i=1}^{n} y_i \alpha_i = 0, \quad i = 1, \ldots, n, \tag{4}$$

where $0 \leq \alpha_i \leq C * C_{factor}$ (for positive samples) and $0 \leq \alpha_i \leq C$ (for negative samples). C is the regularization parameter controlling the tradeoff between training error and margin. The larger the value of C, the larger penalty is assigned to errors. C_{factor} is a cost-factor, which makes the training errors on positive samples outweighing the errors on negative sampled [41].

In the above optimization problem, only those items with $\alpha_i > 0$ can remain. The samples $\mathbf{x}_i$ that lie along or within the margins of the decision boundary (by Kuhn-Tucker theorem) are called the support vectors. The weight vector in (2) can be expressed in terms of $\mathbf{x}_i$ and the solutions α_i of the optimization function (3):

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(\mathbf{x}_i) \tag{5}$$

where $\alpha_i \geq 0$.

Then, the separating hyperplane in (2) becomes

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle + b \tag{6}$$

To avoid the computation of the inner product $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle$ in the high dimensional space during the optimization of (3), the kernel function that can satisfy the Mercer's condition is introduced:

$$K(\mathbf{x}_i, \mathbf{x}) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle \tag{7}$$

The kernel function can be computed efficiently and solve the problem of mapping the samples to the potentially high dimensional feature space.

Radial basis function is used as the kernel in this paper, which is defined by

$$K(\mathbf{x}_i, \mathbf{x}) = exp(-\frac{1}{\sigma} \|\mathbf{x}_i - \mathbf{x}\|^2) \tag{8}$$

where $\sigma > 0$ is the parameter to determine the width of the radial basis function. It controls the flexibility of the classifier. When σ decreases, the flexibility of the resulting classifier in fitting the training data increases, and this might easily lead to over-fitting.

Around 15% of the proteins in LigASite (40 proteins) are selected as the training set of SVM since we found that the results are only slightly different with more proteins as training data. Moreover, since the number of grid points of each protein is very large, more proteins will cause the training time to increase greatly. Based on the consideration of interpretability, only 15% (40 proteins) are selected as the training set. The training data are selected randomly with the same distribution of enzyme type as that of the whole dataset (shown in Fig. 3). The protein used in the training set are shown in Table I.

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

Like most of the datasets in bioinformatics, the dataset used in this paper also encounters the problem of being imbalanced, i.e. the number of positive samples (the grid points of binding site) is much less than the negative samples (the other grid points). Under-sampling is applied to reduce this problem. After several experiments, the one proper proportion between the negative samples and the positive samples is 2:1. Therefore, the negative samples are selected randomly to get this ratio in the training set.

The flowchart for the prediction of protein-ligand binding site is shown in Fig. 8. The training dataset is built with the 29 attributes of each grid point by using ConCavity program and the 3D grid space is set as 1 Å. The training set undergoes random under-sampling, so that the ratio 2:1 between the negative and positive samples can be obtained. SVM is applied on the re-sampled training set to form the classification model. This model

is used later to classify the grid points of the testing proteins. The prediction datasets of each testing protein are also built with the 29 attributes by using ConCavity program. Both the learning and classifying process of SVM are used in the SVM^{light} program. From previous studies, the cavities with volume small than 100 Å^3 are ignored since ligands are not likely to be bound in small cavities.

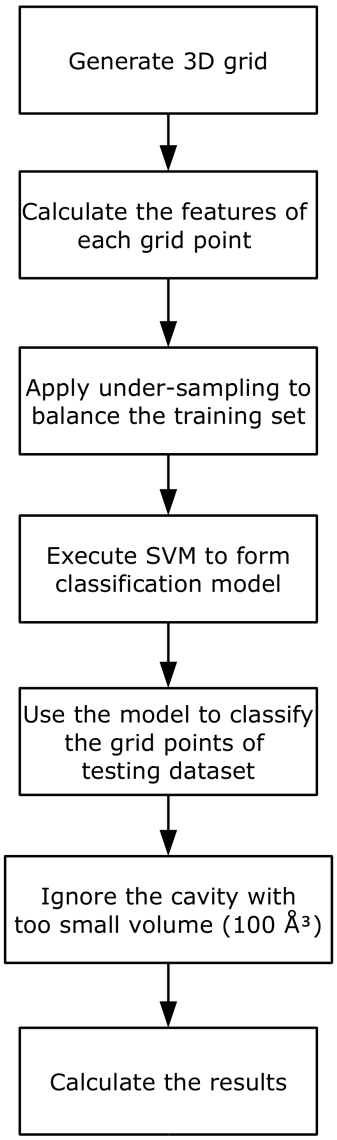


Fig. 8. The flowchart for prediction of protein-ligand binding site.

IV. EVALUATION

To evaluate and compare our method to the others, the same performance measurement should be used. We have applied two different measurements on different methods and datasets.

A. Dataset of LigAsite

For this dataset, grid points are used to represent the potential binding sites. If a grid point is clustered as not suitable for binding ligands, a zero value will be assigned to it. Therefore, the prediction of ligand binding sites can be represented by non-zero values of the grid point, which represent the potential of binding sites. The prediction can be validated by computing the difference with the grid points of known ligands. We define the grid points of the ligand atoms calculated from PDB as the positive samples and the other grid points as the negative samples.

The terms of precision and recall are introduced [42] to measure the performance of the classification of imbalanced testing data. The definitions of precision and recall are given as follows:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. The high value of precision indicates that the predicted positive samples are most likely relevant. The high value of recall indicates that most of the positive samples can be predicted correctly.

Another term called $F - measure$ [42], which is a function of precision and recall, is introduced. It is a popular evaluation metric for imbalanced problems. In principle, $F - measure$ represents a harmonic mean between precision and recall. It is defined as follows:

$$F - measure = \frac{2 * precision * recall}{precision + recall} \tag{11}$$

The area under the receiver operating characteristic (ROC) curve (AUC) is also commonly used to measure the performance of classification. The AUC metric [43] is the probability of correctly identifying a random sample

and can be defined as:

$$AUC = \frac{1 + Recall - FP_{rate}}{2} \tag{12}$$

where $Recall$ is defined in (10) and $FP_{rate} = \frac{FP}{FP+TN}$, FP is the number of false positives and TN is the number of true negatives. FP_{rate} defines the percentage of true negatives cases misclassified as positives.

B. 198 Drug-target Complexes for Testing

After the grid points of potential binding sites are predicted by SVM, the top three largest sites [34] are selected and each site is represented by a grid point in the center of it. [34] has also proved that most of ligands bind to large pockets. Therefore, they suggested an evaluation method for comparing the top three largest sites only.

First, the real binding sites are defined from PDB and each site is represented by a grid point in the center of it. These grid points of real binding sites are compared with the top three largest predicted sites. There are sometimes more than one binding site within a protein. A prediction is counted as a hit if at least one binding stie in the given protein can be located correctly. Using the same approach of [34], 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:

$$success\_rate = \frac{N_{HIT}}{N_P} \tag{13}$$

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.

V. RESULTS

In this paper, the value of σ in (8) is set to the usually chosen value of 1, the value of C_{factor} of α_i in (4) is set to 1, and the value of C in (4) is equal to $n/(\sum_{i=1}^{n} \mathbf{x}_i \cdot \mathbf{x}_i) = 0.7635$, where $\mathbf{x}_i$ is the i-th vector in the training dataset and n is the number of samples in the dataset. The reason of choosing these SVM parameter values is as follows. Table II shows the validation results for different parameters of the SVM classifier. Six random proteins from different enzyme classes are chosen to generate the validation dataset. They are 2cwh, 1g6c, 3p0x, 1wxg, 3kco, and 1k54. In the experiment, the values of σ and C_{factor} differ from 0.5 to 2. The default value of C is

TABLE II

PERFORMANCE OF THE PARAMETERS OF SVM CLASSIFIER.

C	C_{factor}	σ	F-measure	AUC
0.3817	0.5	0.5	0.3089	0.6891
0.3817	0.5	1	0.3053	0.6883
0.3817	0.5	2	0.2991	0.6866
0.3817	1	0.5	0.2756	0.7289
0.3817	1	1	0.2733	0.7295
0.3817	1	2	0.2687	0.7316
0.3817	2	0.5	0.2358	0.7420
0.3817	2	1	0.2359	0.7416
0.3817	2	2	0.2345	0.7409
0.7635	0.5	0.5	0.3118	0.6928
0.7635	0.5	1	0.3081	0.6934
0.7635	0.5	2	0.3009	0.6933
0.7635	1	0.5	0.2784	0.7275
0.7635	**1**	**1**	**0.2767**	**0.7269**
0.7635	1	2	0.2731	0.7277
0.7635	2	0.5	0.2397	0.7410
0.7635	2	1	0.2409	0.7397
0.7635	2	2	0.2413	0.7377
1.1452	0.5	0.5	0.3125	0.6943
1.1452	0.5	1	0.3075	0.6940
1.1452	0.5	2	0.3014	0.6955
1.1452	1	0.5	0.2797	0.7260
1.1452	1	1	0.2781	0.7251
1.1452	1	2	0.2749	0.7247
1.1452	2	0.5	0.2416	0.7398
1.1452	2	1	0.2430	0.7370
1.1452	2	2	0.2443	0.7343

0.7635 and it differs from a half to double of the default value. The results show that the increased F-measure may lead to the decreasing of AUC and the difference between the parameters is not significant. Therefore, the default values of each parameter are set to get a balance between F-measure and AUC.

A. Dataset of LigASite

In the first experiment, six enzyme classes are selected to compare our method with four other methods. They are LIGSITE, SURFNET, PocketFinder and ConCavity. Both LIGSITE and SURFNET used geometric charac-

teristics to predict the ligand binding site. PocketFinder used energy criteria and ConCavity used both geometric and sequence conservation properties to do the prediction. For the grid points determination, LIGSITE applied a threshold with the value of 5.5, SURFNET and PocketFinder determine the threshold value by considering the mean and standard deviation of the grid values. ConCavity applied a binary search to the grid points. The search was made by considering different specified properties based on the sizes and shapes of the pockets. Only the grid points, which met all the properties, were selected.

The success rate is calculated in terms of the $F-measure$ in (11) and AUC in (12). The $F-measure$ and AUC of the training data set are shown in Table III. Both results of sampled and non-sampled training data are given. The results of sampled training data are the classification outcome of the training set that is used to learn the classification model of SVM. As mentioned before, random under-sampling is applied before the SVM training to tackle the problem of imbalanced dataset. The results of non-sampled training data are the classification results of the training set provided by the trained SVM without applying any under-sampling. The other 85% of the selected proteins are then used as testing data to test the performance of our method.

TABLE III

SUCCESS RATE OF TRAINING DATA.

Dataset	F-measure	AUC
Sampled Training Data	0.8150	0.8585
Non-sampled Training Data	0.3360	0.8417

For testing data, the results in Table IV show that our method can classify the grid points correctly with a high value of AUC. The other methods always define the pockets with low AUC since the thresholds of the grid points are not always suitable to the proteins and only one property of protein is considered. The thresholds may be wrongly set by the user. On the contrary, we do not define any threshold for our method. We use SVM to train the system and cluster the grid points which are most likely to bind with ligands. The results also show that the success rate is not sensitive to the enzyme classes the proteins belong to. Both $F-measure$ and AUC show a small difference of values (around 10%) among the six enzyme classes.

Table V shows the $F-measure$ and AUC of testing datasets in different numbers of chains. The results can

TABLE IV

SUCCESS RATE OF TESTING DATA IN SIX ENZYME CLASSES.

Type	Method	F-measure	AUC
Transferase	Our Method	**0.3338**	**0.8162**
	LIGSITE	0.1622	0.6615
	SURFNET	0.2806	0.6516
	PocketFinder	0.08970	0.6353
	ConCavity	0.3195	0.6588
Hydrolase	Our Method	**0.3376**	**0.7548**
	LIGSITE	0.0982	0.6026
	SURFNET	0.2577	0.6332
	PocketFinder	0.07476	0.6132
	ConCavity	0.2963	0.6562
Oxidoreductase	Our Method	**0.3895**	**0.8208**
	LIGSITE	0.2044	0.6705
	SURFNET	0.3142	0.6467
	PocketFinder	0.1255	0.6396
	ConCavity	0.3314	0.6441
Lyase	Our Method	0.3025	**0.8464**
	LIGSITE	0.1507	0.7101
	SURFNET	0.2709	0.6698
	PocketFinder	0.06788	0.6349
	ConCavity	**0.3292**	0.6933
Ligase	Our Method	0.3453	**0.8407**
	LIGSITE	0.1540	0.6831
	SURFNET	0.2823	0.6612
	PocketFinder	0.07515	0.63915
	ConCavity	**0.3750**	0.6988
Isomerase	Our Method	**0.3442**	**0.7839**
	LIGSITE	0.1758	0.6685
	SURFNET	0.2497	0.6341
	PocketFinder	0.1205	0.6236
	ConCavity	0.2519	0.6177
Overall	Our Method	**0.3422**	**0.8105**
	LIGSITE	0.1576	0.7993
	SURFNET	0.2759	0.6494
	PocketFinder	0.07133	0.6310
	ConCavity	0.3172	0.6615

be interpreted by separating into groups. The first group is having 1 and 2 chains, which has the largest values of $F-measure$. The second groups is having 3 and 4 chains while the values of $F-measure$ are in between 0.28 and 0.293. The last groups is having 6 or more chains, which has the lowest values of $F-measure$. Generally, from the results of these three groups, F-measure decreases when the number of chains increases, expect when the number of chains is 5. Fig. 4 shows that there is only one protein with 5 chains. Therefore, the result of the case of 5 chains is not sufficient to reflect the trend. The values of AUC is insensitive to the number of chains. The reason is that more chains of the proteins means more complicated proteins' structure and the number of potential pockets on the proteins' surface increases. The method predicts some extra pockets which are not true binding sites.

TABLE V

SUCCESS RATE OF TESTING DATA IN DIFFERENT NUMBERS OF CHAINS.

No. of Chains	F-measure	AUC
1	0.3427	0.7950
2	0.3674	0.8057
3	0.2803	0.7976
4	0.2933	0.8105
5	0.4416	0.8989
$>= 6$	0.2575	0.7956

The grid points classified as binding sites are subject to further evaluation, which is carried out by computing the difference with the known bound ligands. Since each protein could be bound with more than one ligand, which might be unknown, both $F-measure$ and AUC may not reach at 1.0 and the results of all methods cannot reach a very high rate. Therefore, the comprehensive results are more important. After the binding sites are found, docking process and many medical experiments are needed to find a correct ligand to bind to the protein.

B. 198 Drug-target Complexes for Testing

In the second experiment, 198 drug-target protein complexes are used and our method is compared with six other approaches, based on the evaluation of top three largest binding sites. The six other approaches 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 is the extention 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 prediction. MetaPocket predictes the binding site by combining eight other approaches. Fig. 9 shows an example of binding sites prediction for the protein 1p5j. The real ligand is shown in red sticks at the center and the predicted pockets by all the seven approaches are shown in spheres with different color.

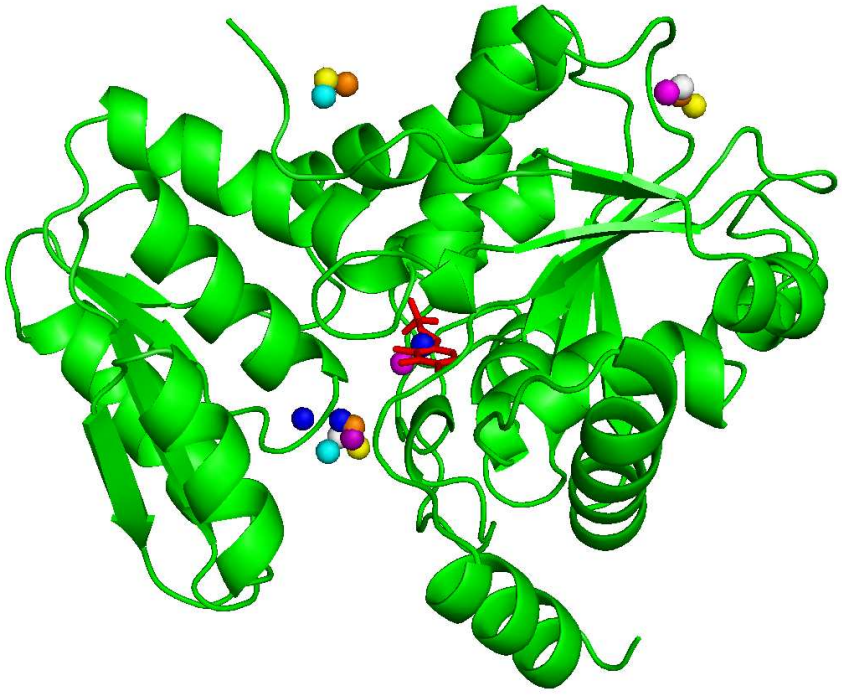


Fig. 9. The real ligand (red) binding site and the predicted pockets for protein 1p5j. The pockets sites of MetaPocket (orange), LIGSITE^{csc} (white), SURFNET (yellow), Fpocket (cyan), Q-SiteFinder (magenta), ConCavity (purple), and our method (blue) are shown in spheres.

The success rate of this experiment is calculated by (13). The prediction results of top 1 to top 3 binding sites for all approaches are evaluated separately. Table VI shows the prediction results of our method and the other six approaches on the 198 drug-target dataset. Our method can achieve the highest success rate among all the methods. Table VII shows the number of hit proteins among the seven methods on the drug-target dataset. There are 130 proteins that can have the binding sites correctly identified as the top 1 predictions. There are 37 and 7

proteins that can have the binding sites correctly identified as the top 2 and top 3 predictions respectively. There are 24 proteins that no associated binding sites can be identified correctly in the top 3 predictions. Our method can locate the highest number of binding sites among all methods.

TABLE VI

SUCCESS RATE (%) OF TOP 3 BINDING SITES PREDICTIONS ON 198 DRUG-TARGET DATASET.

Method	Top 1	Top 1-2	Top 1-3
Our Method	**66**	**84**	**88**
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 VII

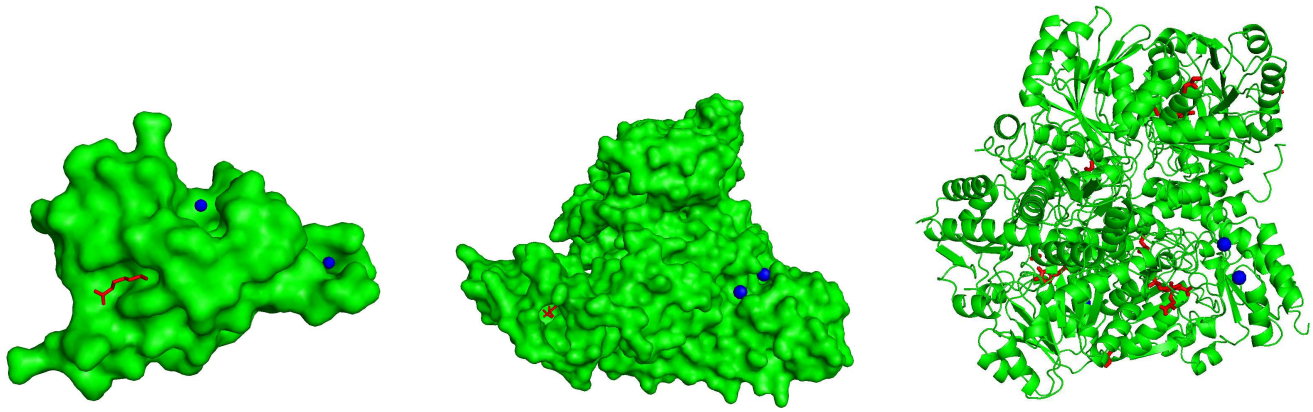
NUMBER OF HIT PROTEINS ON 198 DRUG-TARGET DATASET.

Method	Top 1	Top 2	Top 3	None
Our Method	130	37	7	24
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

The reason why our method can outperform the other methods is that no threshold is set to the grid points to identify the binding sites. Our method forms a training set with 29 different properties of some proteins first, and then applies an SVM to train a classification model. Finally, this model is used to predict the binding sites of other proteins. Besides, we have applied many different properties of protein, such as the geometric characteristics, interaction energy between protein and carbon probe, and sequence conservation score, to make the predictions; while some methods use only one property to locate the binding sites.

Our method still has some limitations. In some proteins, no binding sites can be predicted correctly. In the

drug-target dataset, 24 proteins cannot have the binding sites located correctly. There is also one case in the LigASite dataset. From these cases, we conclude with three limitations of our method. The first one is that ligands may bind to a flat region. Since our method tends to predict the binding sites inside a cavity or pocket, the sites in a flat region are difficult to locate. Ten cases in the drug-target dataset and the case of LigASite belong to this category. The second limitation is that ligands may bind to small cavities. Since only the top three largest binding sites are considered in the drug-target dataset, ligands in small cavities cannot be selected . There are eleven cases in the drug-target dataset belonging to this category. The third limitation is that the binding sites may be inside the proteins while only the pockets on the protein surface can be detected. There are three cases in the drug-target dataset belonged to this cateogry. Fig. 10 shows three examples of the difficult structures mentioned above. The real ligands are shown in red sticks. The predicted binding sites of our method are shown in blue spheres.



(a) 1pk2. (b) 1e7a. (c) 3cog.

Fig. 10. Examples of the three limitations of our method. (a) The ligand binds to a flat region. (b) The ligands bind to small cavities. (c) The binding sites are inside the protein.

VI. Conclusion

The determination of binding sites (pockets) is the prerequisite 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 distinguish the binding sites. It makes use of the properties of geometric

characteristics, interaction potential, distance from protein, conservation score and the grid points nearby to identify the binding sites. 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, LIGSITE^{csc}, SURFNET, Fpocket, PocketFinder, Q-SiteFinder, ConCavity, and MetaPocket on the datasets of LigASite and 198 drug-target protein complexes. For the LigASite dataset, the binding sites are represented as grid points and our approach gets better results than the other approaches. The sites are predicted correctly in 35 % and 80 % of $F - measure$ and AUC respectively. The proposed method is shown to offer more comprehensive results than the others since more proteins fail to have the binding sites located when other approaches are used. For the 198 drug-target dataset, only the top three largest binding sites are considered and represented as one center point of each site. The results show that our approach performs better than the other approaches and predicts the binding sites correctly in 66% at top 1 prediction, 84% at top 1–2 prediction, and 88% 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.

ACKNOWLEDGMENT

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

REFERENCES

[1] I. Korf, P. Flicek, D. Duan, and M. Brent, “Integrating genomic homology into gene structure prediction,” *Bioinformatics*, vol. 17, pp. 140–148, 2001.

[2] O. Troyanskaya, K. Dolinski, A. Owen, R. Altman, and D. Botstein, “A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae),” in *Processdings of the National Academy of Sciences of the United States of America*, vol. 100, no. 14, 2003, pp. 8348–8353.

[3] A. Hatzigeorgiou, N. Mache, and M. Reczko, “Functional site prediction on the dna sequence by artificial neural networks,” in *Intelligence and Systems, 1996., IEEE International Joint Symposia on*, Nov. 1996, pp. 12–17.

[4] P. Aloy, E. Querol, F. Aviles, and M. Sternberg, “Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking,” *Journal of Molecular Biology*, vol. 211, pp. 395–408, 2001.

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine Learning*, vol. 46, pp. 389–422, 2001.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] A. Wlodawer and J. Vondrasek, “Inhibitors of HIV-1 protease: A major success of structure-assisted drug design,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 27, pp. 249–284, 1998.

[12] A. Laurie and R. Jackson, “Methods for the prediction of protein-ligand 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.

[13] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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

[22] A. Gutteridge, G. Bartlett, and J. Thornton, "Using a neural network and spatial clustering to predict the location of active sites in enzymes," *Journal of Molecular Biology*, vol. 330, pp. 719–734, 2003.

[23] A. Koike and T. Takagi, "Prediction of protein-protein interaction sites using support vector machines," *Protein Engineering, Design and Selection*, vol. 17, no. 2, pp. 165–173, 2004.

[24] M. Keil, T. Exner, and J. Brickmann, "Pattern recognition strategies for molecular surfaces: binding site prediction with a neural network," *Journal of Computational Chrmistry*, vol. 25, no. 6, pp. 779–789, 2004.

[25] J. Bradford and D. Westhead, "Improved prediction of protein-protein binding sites using a support vector machines approach," *Bioinformatics*, vol. 21, no. 8, pp. 1487–1494, 2005.

[26] B. Boser, I. Guyon, and V. Vapnik, "An training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, 1992, pp. 144–152.

[27] V. Vapnik, *The nature of statistical learning theory*, 2nd ed. Springer, 1999.

[28] B. Schölkopf and A. Smola, *Learning with kernels*. Cambridge (Massachusetts): MIT Press, 2002.

[29] K. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, pp. 181–201, 2001.

[30] B. Schölkopf, K. Tsuda, and J. Vert, *Kernel methods in computational biology*. Cambridge (Massachusetts): MIT Press, 2004.

[31] J. Vert, "Kernel methods in genomics and computational biology," in *Kernel Methods in Bioengineering, Signal and Image Processing*, G. Camps-Valls, J. Rojo-Álvarez, and M. M.-R. M, Eds. Idea Group, 2007, ch. 2, pp. 42–63.

[32] A. Ben-Hur, C. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support vector machines and kernels for computational biology," *PLoS Computational Biology*, vol. 4, no. 10, 2008.

[33] B. Dessailly, M. Lensink, C. Orengo, and S. Wodak, "LigASite: a database of biologically relevant binding sites in proteins with known apo-structures," *Nucleic Acids Research*, vol. 36, pp. 667–673, 2008.

[34] 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.

[35] 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.

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

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

[38] 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

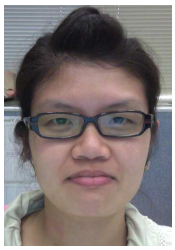
[39] N. Petrova and C. Wu, “Prediction of catalytic residues using support vector machines with selected protein sequence and structural properties,” *BMC Bioinformatics*, vol. 7, no. 1, p. 312, 2006.

[40] T. Joachims, “Making large-scale svm learning practical,” in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, 1999, ch. 11, pp. 169–184. [Online]. Available: http://svmlight.joachims.org/

[41] T. Joachims, *Learning to classify text using Support Vector Machines: methods, theory, and algorithms*. Kluwer Academic Publishers, 2002.

[42] Q. Gu, Z. Cai, L. Zhu, and B. Huang, “Data mining on imbalanced data sets,” in *International Conference on Advanced Computer Theory and Engineering*, 2008, pp. 1020–1024.

[43] S. García, J. Derrac, I. Triguero, C. Carmona, and F. Herrera, “Evolutionary-based selection of generalized instances for imbalanced classification,” *Knowledge-Based Systems*, vol. 25, no. 1, pp. 3–12, 2012.



Ginny Y. Wong (S’09) received the B.Eng. degree in Electronic and Information Engineering from the Hong Kong Polytechnic University, Hong Kong, in 2008. She is currently working towards her Ph.D. degree in the same department. Her current research interests include machine learning and bioinformatics.

Frank H.F. Leung was born in Hong Kong in 1964. He received the B.Eng. degree and the Ph.D. degree in Electronic Engineering from the Hong Kong Polytechnic University in 1988 and 1992 respectively.

He joined the Hong Kong Polytechnic University in 1992 and is now Associate Professor in the Department of Electronic and Information Engineering. He is active in research and has published over 180 research papers on Computational Intelligence, Control and Power Electronics. At present, he is involved in the R&D on Intelligent Systems. He has been serving as a reviewer for many international journals and helping the organization of many international conferences.

S.H. Ling (M06-SM12) received the B.Eng. degree in the Department of Electrical Engineering, M.Phil. and Ph.D. degrees from the Department of Electronic and Information Engineering in the Hong Kong Polytechnic University in 1999, 2002 and 2006 respectively. In 2008, he worked as Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Currently, he works in University of Technology, Sydney, Australia as Lecturer. He has authored and coauthored over 120 books, international journal and conference papers on computational Intelligence and its industrial applications. His current research interests include evolution computations, fuzzy logics, neural networks, hybrid systems and biomedical applications. Currently, he serves as Co-Editors-in-Chief for Journal of Intelligent Learning Systems and Applications.