Propensity Vectors of Low-ASA Residue Pairs in the Distinction of Protein Interactions

Qian Liu and Jinyan Li*

Bioinformatics Research Center & School of Computer Engineering

Nanyang Technological University, Singapore 639798

Short title: Distinguishable low-ASA residue pairs

Keywords: Low-ASA Residue Pair, O-Ring-Surrounded Region, Propensity Vector, OringPV

Correspondence Author*:

Professor Jinyan Li
School of Computer Engineering
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
Tel: (65) 67906253 (office)
Fax: (65) 67926559
Email: jyli@ntu.edu.sg

Abstract

We introduce low-ASA residue pairs as classification features for distinguishing the different types of protein interactions. A low-ASA residue pair is defined as two contact residues each from one chain that have a small solvent accessible surface area (ASA). This notion of residue pairs is novel as it first combines residue pairs with the O-ring theory, an influential proposition stating that the binding hot spots at the interface are often surrounded by a ring of energetically less important residues. As binding hot spots lie in the core of the stability for protein interactions, we believe that low-ASA residue pairs can sharpen the distinction of protein interactions. The main part of our feature vector is 210-dimensional, consisting of all possible low-ASA residue pairs; the value of every feature is determined by a propensity measure. Our classification method is called OringPV which uses propensity vectors of protein interactions for support vector machine. OringPV is tested on three benchmark datasets for a variety of classification tasks such as the distinction between crystal packing and biological interactions, the distinction between two different types of biological interactions, etc. The evaluation frameworks include within-dataset, cross-dataset comparison, and leave-one-out cross-validation. The results show that low-ASA residue pairs and the propensity vector description of protein interactions are truly strong in the distinction. In particular, many cross-dataset generalization capability tests have achieved excellent recalls and overall accuracies, much outperforming existing benchmark methods.

1 Introduction

Close interactions are necessary and indispensable for proteins to fulfill molecular functions and biological processes. Their binding behavior and the associated physicochemical properties are complicated and amazing. One of the fundamental problems is to characterize the types of protein interactions by using their structure information at the residue or atom level. Protein complexes that are determined by the popular and prolific technique X-ray crystallography can be broadly classified into crystal packing or biological interaction according to the biological reality of the contact. Crystal packing interactions/contacts are enforced by the crystallographic packing environment and formed during the crystallization process, but they do not occur in solution or in their physiological states[1]. Biological interactions have been carefully studied and categorized into sub-groups according to various criteria, such as permanent versus transient complexes on the basis of the lifetime of the complexes[2], and homo-oligomers versus hetero-oligomers according to whether the interactions occur between identical chains or not[2]. Depending on whether the protomers of interactions can be found or not as stable structures on their own *in vivo*[2], biological interactions can also be grouped into obligate or non-obligate interactions[2]. Focusing on the different transition processes in protein folding and binding[3], two-state folding complexes and three-state complexes were also used to describe obligate and non-obligate interactions. Further, Ofran and Rost's work[4] categorized biological interactions in a more complicate way with six subtypes. The different types of biological interactions possess their unique binding behaviors. For example, obligate interactions are stable, and their protein chains function only in the complex form[2]; however, the protomers in transient interactions associate to accomplish a particular function upon a molecular stimulus and dissociate after that[2]. Therefore, deep understanding of these binding behaviors can be particularly useful for reliable predictions on the types of interactions in new protein complexes[5–7], and it is also helpful for docking algorithms to construct the protein quaternary structures[8] and to identify protein binding sites[9,10].

Outstanding chemical, physical and geometric properties[9,11] have been extensively explored in literature to describe and characterize protein interactions and their binding interfaces. These properties include hydrophobicity and polarity[12–15] as chemical features, interface size and contact area as physical features, and planarity, shape complementarity, circularity[13,16] and secondary structure[9] as geometric features. Other features such as residue conservation[6,11], residue composition[6,17–19] and propensity[15], residue pairs[7,20] and atomic pairs[5,21–23] have been also proposed. Many of them have been involved in important findings. For example, biological interactions are found to be significantly different in residue composition from the rest of protein surfaces[15,16,24,25], while crystal packing possess similar composition to the rest of protein surfaces[26]. The physical feature, interface area of biological interactions, is found to be much larger than that in non-biological interactions[6,11,14,15,17,26–28].

These properties have also been integrated by many classification algorithms to discriminate biological interactions from crystal packing[15,29], and to distinguish different types of biological interactions[5,6,30]. Bernauer *et al.* developed structure-based scoring functions[7,31] and later the DiMoVo method[32] for identifying biological interactions. Their key idea is on a Voronoi tessellation which nicely describe the geometric and physicochemical complementarities of protein interfaces. Zhu *et al.*[6] proposed several descriptors of interfaces, such as interface area, amino acid composition and gap volume, to distinguish obligate, non-obligate interactions and crystal packing. More recently, atom/residue pairs were conceptualized and used in the distinction of protein interactions[5,8,21,22,30], e.g., by the ACV (atom contact vector) method[5].

The common approach adopted by the above classification methods is that the classification properties are all taken from interface residues whose surface accessibility change is >0.1 Å^2 [14,15] or >1.0 Å^2 [6,16] upon the formation of complexes. In this work, we further narrow down the scope of interfacial residues to concentrate on those of low solvent exposure. Our new notion is called *low-ASA residue pairs.* A low-ASA residue pair is defined as two contact residues whose ASA (solvent accessibility surface area) is very small. The notion of low-ASA residue pairs is in agreement with the influential O-ring hypothesis[33–36], and one of its successors, the insightful "coupling" proposition[37]. The O-ring theory states that the binding 'hot spot' residues are usually clustered and located at the center of the interfaces, and they are often surrounded by energetically less important residues shaped like an O-ring for occluding water molecules. The "coupling" theory[37] highlights that the hot spot residues are always coupled to each other with a short distance between the two sides of the interface. Therefore, given a low-ASA residue pair, the two residues are both buried by O-ring residues (i.e., the residues on the O-ring), and the spatial compactness between them is very tight. Thus, low-ASA residue pairs probably form a special area that is richer of hot spot residues than the other areas of the interface. As binding hot spots lie in the core of the stability for protein interactions[33], we believe that low-ASA residue pairs can sharpen the difference between different types of protein interactions. With low-ASA residue pairs, an immediate ease is to accurately identify crystal packing, because the interfaces of crystal packing contain few fully buried atoms[15].

A residue in a low-ASA residue pair may have multiple partners. Assume (A_1, B_1) is a low-ASA residue pair in an interacting chain pair, it is often the case that (A_1, B_2), (A_1, B_3), or even (A_1, B_4) is also a low-ASA residue pair. With this regard, we introduce O-ring-surrounded regions. Given an interacting chain pair, its O-ring-surrounded region is the union of all low-ASA residue pairs of this interacting chain pair.

We propose to construct a propensity vector to characterize the interaction behavior of an interacting chain pair by using all of its low-ASA residue pairs as features. The propensity vector of the low-ASA residue pairs consists of 213 feature elements: three elements are used for describing the summary information of its O-ring-surrounded region, and the remaining 210 elements are reserved for the propensity values of all possible low-ASA residue pairs contained in the O-ring-surrounded region against the rest of protein surfaces. Assume we are given n number of protein interactions in a classification problem, then n propensity vectors will be constructed accordingly. Each of these propensity vectors will be labeled with the types of biological interactions, or crystal packing in the training data.

Propensity vectors of our low-ASA residue pairs are related to but different from the residue-pair method proposed in [7,20]. Firstly, our residue pairs are low-ASA residue pairs satisfying the O-ring theory and the coupling proposition, while residue pairs of [7,20] are just interface residue pairs which include residues outside the O-ring surrounded region. Secondly, we calculate propensity values of the low-ASA residue pairs which are totally different from frequency values of residue pairs as used in [7,20]. Propensity values of residues are more competitive to frequency values to improve classification performance as early observed by Bahadur *et al.*[15]. Finally, propensity vectors do not require the assumptions of additivity as [7,20] required. Our propensity vector is also different from the frequency vector of atom pairs proposed in ACV[5,30]. Compared with ACV, our vectors take into account more biologically useful properties, such as residue composition and propensity of residue pairs, for signifying the physicochemical properties of interfaces.

The discriminating power of propensity vectors of protein interactions are tested on three benchmark datasets[6,15,21]. In the experiments, we consider a variety of classification tasks, such as distinction between biological interactions and crystal packing, distinction within biological interactions (e.g., obligate vs non-obligate interactions) and the 3-class classification problem. The performance evaluation is also measured under a variety of frameworks, including within-dataset comparison, cross-dataset generalization capability test, and LOOCV (leave-one-out cross-validation). Our comprehensive comparison results have shown that low-ASA residue pairs and the propensity vector description of protein interactions are truly strong in the prediction of protein interaction types. In particular, many cross-dataset generalization capability tests have achieved excellent recalls and overall accuracies, much outperforming existing benchmark methods.

2 Methods

In this section, we give an overview to the three test datasets. Then we present a formal definition for low-ASA residue pairs and O-ring-surrounded regions, followed by a description of how to compute a propensity vector of the low-ASA residue pairs for a protein interaction. We also introduce our classification method, OringPV, and describe performance evaluation measurements.

2.1 Three Benchmark Datasets

The first benchmark dataset is the BNCP-CS dataset[6] which comprises 75 obligate interactions, 62 non-obligate interactions and 106 crystal packing. Zhu *et al.*[6] tested their NOXClass method on this dataset to predict the three types of protein interactions. Here, the obligate and non-obligate interactions in the BNCP-CS dataset are specially denoted by BNCP-CS^{bio}, and the set of obligate interactions and crystal packing are denoted by BNCP-CS^{hocp}.

The second dataset is the Ponstingl dataset[21]. It consists of 95 monomers and 76 homodimers. On this datset, Ponstingl *et al.* tested their score schemes[21] for the distinction between homodimeric proteins and monomeric proteins, and Mintseris and Weng tested their ACV method[5].

The third dataset is the Bahadur dataset[15] which contains 70 heterodimeric complexes (also termed as protein-protein complexes in [24]), 122 homodimeric proteins and 188 crystal packing. The homodimeric proteins and crystal packing of this dataset were used previously in [7,15]. In this work, the homodimeric proteins and crystal packing are specially denoted by Bahadur^{hocp}, and the heterodimeric complexes and homodimeric proteins in the Bahadur dataset are denoted by Bahadur^{bio} dataset.

2.2 Low-ASA Residue Pairs and Their Propensity Vector

Let C_1 and C_2 be a pair of interacting polypeptide chains, two residues $r_1 \in C_1$ and $r_2 \in C_2$ are defined as a **low-ASA residue pair**, if (i) the ASA of r_1 and of r_2 are both small, ideally with a value close to zero, namely no much contact with water solvent; (ii) the minimum of the atom distances between r_1 and r_2 is less than a threshold plus the van der Waals radii of the corresponding atoms. Here, we set this threshold value, denoted by d_{tw}, as a real number less than the van der Waals diameter (2.75 Å) of water molecules. The first criteria of this definition captures the idea of the influential O-ring theory, indicating that the two residues r_1 and r_2 should satisfy a proposed condition[33, 36] for them to be in a hot spot. The second criteria best follows the spirit of the coupling proposition, emphasizing the importance of a water-free distance between two contact residues. The notion of low-ASA residue pairs also more-or-less shares a light with our recent "double water exclusion" hypothesis[38] which was proposed to refine the O-ring theory for the binding hot spots at protein interfaces.

The **O-ring-surrounded region** of an interacting chain pair is the union of all low-ASA residue pairs of this interacting chain pair. Usually, such a region is covered by one O-ring, i.e., a mono-island region. However, sometimes, two subsets of low-ASA residue pairs in the O-ring-surrounded region of an interacting chain pair may not share any common residues with each other. Such a region may be covered by two or more O-rings, i.e., a multi-island region. Nevertheless, we treat the union of all low-ASA residue pairs as the O-ring-surrounded region of the chain pair. Figure 1 shows an example of O-ring-surrounded region which is located at the binding site between chain F and G of PDB entry 1GLA. This O-ring-surrounded region consists of 24 low-ASA residue pairs involving 14 residues at chain F and 8 residues at chain G. Of the 24 low-ASA residue pairs, some are duplicates. For example, residue pair (THR, VAL) occurs four times, residue pair (PHE, THR) occurs three times, and both residue pair (ILE, PHE) and (ILE, SER) occur twice. All the rest occur only once in this O-ring-surrounded region. Actually, the size of O-ring-surrounded regions varies greatly among protein interfaces, especially among different types of protein interfaces. This can be seen from Table I which shows the size information of O-ring-surrounded regions for the interactions in the BNCP-CS dataset. We note that a size of zero means the O-ring-surrounded region is an empty set of residue.

Given an interacting polypeptide chain pair C_1 and C_2, computationally, we use two steps to locate low-ASA residue pairs:

(1) we take the NACCESS software[39] to remove those residues that have a relative accessible surface area in the complexed form bigger than a threshold. In this work, this threshold is

TABLE I: The size of O-ring-surrounded regions for the BNCP-CS dataset.

Interaction	number of residues (and residue pairs)		
Type	minimum	average	maximum
Crystal packing	0 (0)	13 ±*7.4*	40 (46)
Transient	12 (13)	45 ±*20.6*	146 (214)
Obligate	23 (28)	83 ±*53.2*	288 (393)

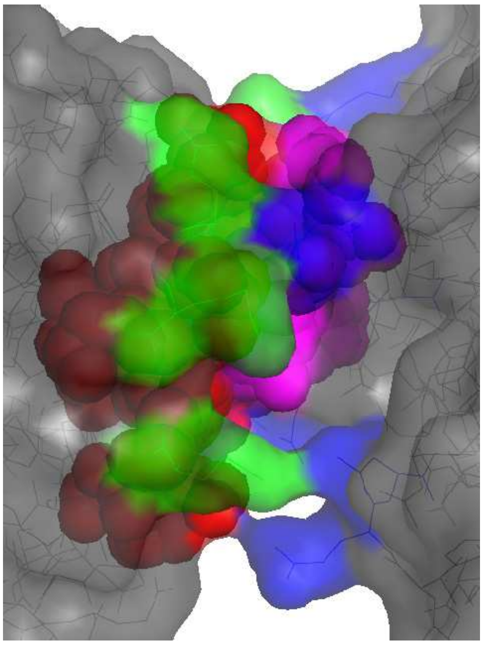


Fig. 1: The O-ring-surrounded region between chain F and G of 1GLA in the 'spheres' view in PyMOL with the color red and magenta. The residues in green and blue are interfacial residues of large ASA which are filtered and are not used to form low-ASA residue pairs.

set as 36% following the one recommended by [38]. Theoretically, this threshold should be close to zero, but in real case, it is too strict, leaving too small number of residues for statistical analysis. On the other hand, some hot-spot residues have relatively large ASA[34]. How to determine the optimal threshold is still a difficult problem.

(2) Let C'_1 and C'_2 be the residue set of C_1 and of C_2, respectively, after the residue removal by step 1. Let $r_i \in C'_1$ and $r_j \in C'_2$, we calculate the distance between all possible atom pairs of r_i and r_j. If the minimal distance is less than their van der Waals radii plus d_{tw}, then r_i and r_j is a low-ASA residue pair.

We would like to also point out that there is no gold standard about how to determine an optimal d_{tw}. We have tried to set d_{tw} as every value from 0.5 Å to 2.75 Å with step 0.25 Å. We found that the classification performance had only small variation across most of these situations when d_{tw} changed. In this work, we only report the results when d_{tw} was set as 1.5 Å. Note that under this setting (d_{tw}=1.5 Å), the performance was not always the best on all of the datasets.

Construction of our Propensity Vectors: Given an interacting protein chain pair, C_1 and C_2, we construct a propensity vector based on the low-ASA residue pairs of this interaction. This propensity vector consists of two parts. (i) At the *summary part*, there are three feature elements for describing the summary information of the O-ring-surrounded region—two numbers (each for one chain) of the contact residues from the O-ring-surrounded region, and the total number of low-ASA residue pairs; (ii) at the *propensity part*, there are 210 elements ($C_{20}^2 + 20$=210) each for describing the propensity value of one of the all possible residue pairs (r_i, r_j), $i \leq j, i, j = 1, ..., 20$, between the O-ring-surrounded region and the surface area of this protein chain pair. The propensity value of a residue pair (r_i, r_j) is calculated by

$$p_{(r_i,r_j)}= \log(\frac{f_{(r_i,r_j)}}{(f_{r_i^1} f_{r_j^2}+f_{r_i^2} f_{r_j^1})/2} + 1) \tag{1}$$

where $f_{(r_i,r_j)}$ is the frequency fraction of the residue pair (r_i, r_j) in the O-ring-surrounded region; $f_{r_k^c}$, k=i, j, is the frequency fraction of the residue r_k in the surface residues of the protein chain C_c (c=$1, 2$); the number 2 is used to compensate the double expected count of the pairs for surface residues. A residue is considered as a surface residue if its relative accessibility is greater than 25%[40]. Note that our $f_{r_k^c}$ is based on the protein surface residues instead of interface residues[20]. Reasons why we calculate propensity values of residue pairs by using $f_{r_k^c}$ in the protein surface residues include: (i) The interface between two proteins in crystal packing occurs by chance, and the residue composition in the interface is similar to that in the rest of protein surfaces[26]; (ii) However, at the interfaces of biological interactions, the residue composition is statistically different from protein surfaces[15,16,24,25]. Thus, the interfaces of biological interactions and crystal packing have different propensities to be compared with protein surfaces.

We take an example to show how the propensity value of a feature is obtained. Residue pair (THR, VAL) is a low-ASA residue pair in the O-ring-surrounded region shown in Figure 1. This residue pair occurs four times in this region, so $f_{(THR,VAL)}$ is 0.167 (4/24). And then $f^F_{(THR)}$=0.076, $f^G_{(THR)}$=0.0286, $f^F_{(VAL)}$=0.038 and $f^G_{(VAL)}$=0.0514. Therefore, the propensity value of this residue pair, $p_{(THR,VAL)}$=$log(\frac{0.167}{(0.076*0.0514+0.038*0.0286)/2} + 1)$ =4.22.

Construction of binary vectors and frequency vectors: In parallel, a binary vector of low-ASA residue pairs and a frequency vector of low-ASA residue pairs are also proposed in comparison to the above propensity vector. The binary vector and the frequency vector both have the same 213 feature elements as the propensity vector does. The only difference is on the feature values. The values of the 210 feature elements (i.e., the second part of the vector) of a binary vector indicate whether the residue pairs occur in the O-ring-surrounded region or not; similarly, the 210 feature values of the frequency vector are the frequency values of the residue pairs in the O-ring-surrounded region. These binary vectors and frequency vectors are called low-ASA binary vectors and low-ASA frequency vectors respectively in this work. They are different from those defined in [7,20] which determine the feature values by using the entire interface residues (without considering the water accessibility restriction). However, we use residues in O-ring-surrounded region which are the energetically most important subset of residues in the binding. In this work, the binary vectors and the frequency vectors based on the entire interfaces (without considering the water accessibility restriction) are termed *traditional* binary vectors and *traditional* frequency vectors respectively.

2.3 OringPV: Our Classification Method

Given an interaction classification task, we first construct a propensity vector for every interaction in the training data and also in the test data. Then, we take these training propensity vectors as input and feed to Support Vector Machine (SVM)[41] to build a classifier. As low-ASA residue pairs are heavily involved in the construction of the propensity vectors, we name this learning process OringPV (short for learning by Propensity Vectors of low-ASA residue pairs in O-ring-surrounded regions).

In this work, all classification tasks are performed by running the *libsvm* software package[42] which contains an implementation of the SVM learning method. A Radial Basis (RBF) kernel function was chosen in the training. To determine optimized C and γ for the RBF kernel functions, a grid search heuristics[43] was imposed on training data with 10-fold cross-validation.

For the test datasets above that have three types of protein interactions, such as the BNCP-CS dataset and the Bahadur dataset, our OringPV method employs a two-stage SVM: the first-stage SVM is used to discriminate crystal packing and biological interactions, and the second-stage SVM is subsequently to differentiate different types of biological interactions.

2.4 Performance Measures

To quantify sensitivity performance of classification methods, we denote biological interactions as positive set and crystal packing as negative set when identifying biological interactions from non-biological interactions, and one type of biological interactions (e.g., obligate interactions in the BNCP-CS dataset) as positive set and the other type of biological interactions (e.g., non-obligate interactions in the BNCP-CS dataset) as negative set in the process of distinguishing two types of biological interactions. In addition to *sensitivity* (the fraction of correctly predicted positive interactions over all positive interactions), the performance is evaluated also based on *precision* (the percentage of correctly predicted positive interactions over all predicted interactions), *specificity* (the fraction of correctly predicted negative interactions over all negative interactions), *accuracy* (the number of correctly predicted positive and negative interactions divided by the number of all interactions), as well as Receiver Operating Characteristics (ROC) curves and their Area Under the ROC curves (AUC).

3 Classification Results

Our experiments are conducted under the following four aspects of considerations:

- Leave-one-out cross-validation (LOOCV) within the datasets for showing the outstanding capability of our OringPV method to distinguish different types of protein interactions.

- The comparison of the propensity vectors with low-ASA binary vectors, low-ASA frequency vectors, and the vectors without considering the water accessibility restriction to show the subtle and deep discriminating power of the propensity vectors.

- Within-dataset comparison with benchmark classification methods such as the NOXclass method[6], the DiMoVo method[7], and other methods[5, 15, 21]. The evaluation frameworks for each dataset strictly follow those set by these literature methods.

- Cross-dataset test for comparison between our OringPV method and the literature methods. We use one dataset for training and other datasets for performance testing. It is similar to the so called independent or blind data testing scheme. This is a more reliable approach to testing a classifier's generalization capability.

TABLE II: The overall performance of our OringPV method under LOOCV procedure. The numbers in parentheses are for classification performance by the traditional frequency vectors.

Dataset and		Sensitivity	Specificity	Precision	Accuracy(%)		
Interaction Types		(%)	(%)	(%)	SVM Stage 1	SVM Stage 2	SVM Overall
BNCP-CS	OB	90.7(88.2)	97.0(92.9)	93.2(84.8)	96.7(95.1)	87.9(82.9)	92.2(88.1)
	NO	90.3(79.0)	93.4(93.4)	82.4(80.3)			
	CP	94.3(93.4)	98.5(96.4)	98.0(95.2)			
Bahadur	HO	83.6(81.1)	96.9(88.0)	92.7(76.2)	87.9(78.7)	87.4(73.8)	85.8(75.5)
	NO	71.4(51.4)	95.5(94.5)	78.1(67.9)			
	CP	92.6(80.9)	83.3(76.6)	84.5(77.2)			
Ponstingl	HO	88.2(81.6)	-	93.1(88.6)	-	-	91.8(87.1)
	CP	94.7(91.6)	-	90.9(86.1)			
Bahadur^{hocp}	HO	85.2(80.3)	-	96.3(84.5)	-	-	92.9(86.5)
	CP	97.9(90.4)	-	91.1(87.6)			
BNCP-CS^{bio}	OB	89.3(86.7)	-	94.4(86.7)	-	-	91.2(85.4)
	NO	93.5(83.9)	-	87.9(83.9)			
Bahadur^{bio}	HO	95.9(90.2)	-	96.7(92.4)	-	-	95.3(89.1)
	NO	94.3(87.1)	-	93.0(83.6)			

'OB', 'NO', 'CP' and 'HO' represent obligate interactions, non-obligate interactions (heterodimeric complexes), crystal packing and homodimeric interactions, respectively; '-' means the values are not applicable.

3.1 LOOCV Performance by OringPV within Datasets

The OringPV's sensitivity, precision, specificity and accuracy are presented in Table II, while the corresponding confusion matrix results are reported in Supplementary Table I and Supplementary Table II.

The results in Table II together with some results from Supplementary Table I demonstrate a strong capability of distinguishing biological interactions and crystal packing by our OringPV method. For example, it can effectively identify homodimers from monomers with an accuracy of 91.8% on the Ponstingl dataset and 92.9% on the Bahadur^{hocp} dataset. For differentiating biological interactions from crystal packing, the accuracy of OringPV can reach a level as high as 96.7% on the BNCP-CS dataset and 87.9% on the Bahadur dataset. Therefore, we can conjecture that low-ASA residue pairs and the propensity vectors can capture the signature patterns of biological interactions, and our OringPV method can identify them from crystal packing with high accuracy.

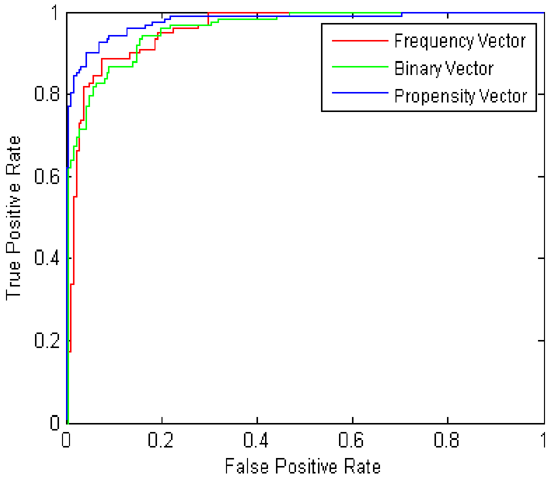


Fig. 2: The ROC curves of the classification between homodimeric proteins and monomers on the Bahadur^{hocp} dataset for frequency vectors, binary vectors and propensity vectors.

Outstanding classification performance by OringPV to distinguish between two types of biological interactions is also shown in Table II, and the corresponding confusion matrix results are reported in Supplementary Table II. The point we want to make here is: when an interaction is confirmed as biological interaction, OringPV can exactly tell which type of biological interaction it belongs to with high accuracy. For example, only 12 interactions were misclassified out of 137 interactions in the BNCP-CS^{bio} dataset, and only 9 misclassified in the 192 interactions of the Bahadur^{bio} dataset, which means 91.2% and 95.3% accuracy respectively.

We calculated the ROC curves and their AUC values of OringPV in every above individual classification experiment with the optimal C and γ. Again, it is confirmed that OringPV can well characterize the distinction of different protein interactions. For example, in Figure 2, the AUC values of classification performance for low-ASA frequency vectors, low-ASA binary vectors and our propensity vectors on the Bahadur^{hocp} dataset are as high as 0.9571, 0.9573 and 0.9789 respectively.

3.2 Propensity vector in comparison to propensity vector without the ASA restriction, and in comparison to traditional frequency vector

Recall that a low-ASA residue pair is a contact residue pair that satisfies the ASA restriction, i.e. its ASA is required to be small. This ASA restriction is sometimes referred to as residue filtering. In most cases, the classification performance of OringPV can be improved a lot if we take low-ASA residue pairs instead of using the residue pairs that do not apply the ASA filtering, as shown in

TABLE III: The accuracy performance comparison among propensity vector (PV), binary vector (BV) and frequency vector (FV) with/without ASA restriction.

Dataset/SVM-stage		PV^{OR}	BV^{OR}	FV^{OR}	PV^{NonOR}	BV^{NonOR}	FV^{NonOR}
Ponstingl		91.8	86.5	83.6	89.5	88.3	87.1
Bahadur	SVM 1	87.9	82.6	82.6	85.2	81.6	78.7
	SVM 2	87.4	81.7	79.0	84.5	82.0	73.7
	overall SVM	85.8	81.1	79.7	83.2	81.0	75.5
Bahadur^{hocp}		92.9	87.4	88.4	86.8	85.8	86.5
Bahadur^{bio}		95.3	94.8	89.5	96.8	95.3	89.0
BNCP-CS		92.2	89.7	90.9	93.8	89.3	88.1

'SVM 1', 'SVM 2' and 'overall SVM' represent the first stage, the second stage of a SVM classifier and the whole classifier. $^{OR}/^{NonOR}$ means that the vectors are with/without ASA restriction.

column 2 and 5 in Table III. In particular, on the Bahadur^{hocp} dataset, the accuracy of OringPV is improved from 86.8% to 92.9% when the ASA filtering is applied to concentrate on low-ASA residue pairs. It seems that the ASA filtering is very sensitive to the performance improvement for identifying biological interactions from crystal packing. However, for distinguishing the two types of biological interactions, the filtering cannot improve the performance as it does when classifying biological and non-biological interactions. This difference is probably attributed to the degree of the ASA filtering. The ASA filtering removes about 17.9% residue pairs from the contact residue pairs of the biological interactions, and 31.8% residue pairs from crystal packing. In other words, the binding interfaces of biological interactions much more satisfy the requirement of O-ring-surrounded regions than non-biological interactions. Thus, the ASA filtering can achieve more improvement for identifying biological interactions from crystal packing and has lesser impact on classification of two types of biological interactions.

The numbers in the parentheses of Table II, Supplementary Table I and Supplementary Table II are the classification performance achieved by the traditional frequency vectors on the three benchmark datasets. (See the definition of the traditional frequency vectors at Section 2.2.) This performance comparison suggests that the propensity vectors can much outperform the traditional frequency vectors and improve the classification accuracy by 3% to 10% in most cases. For example, on the Bahadur dataset, the accuracy is improved from 75.5% to 85.8% by the propensity vectors.

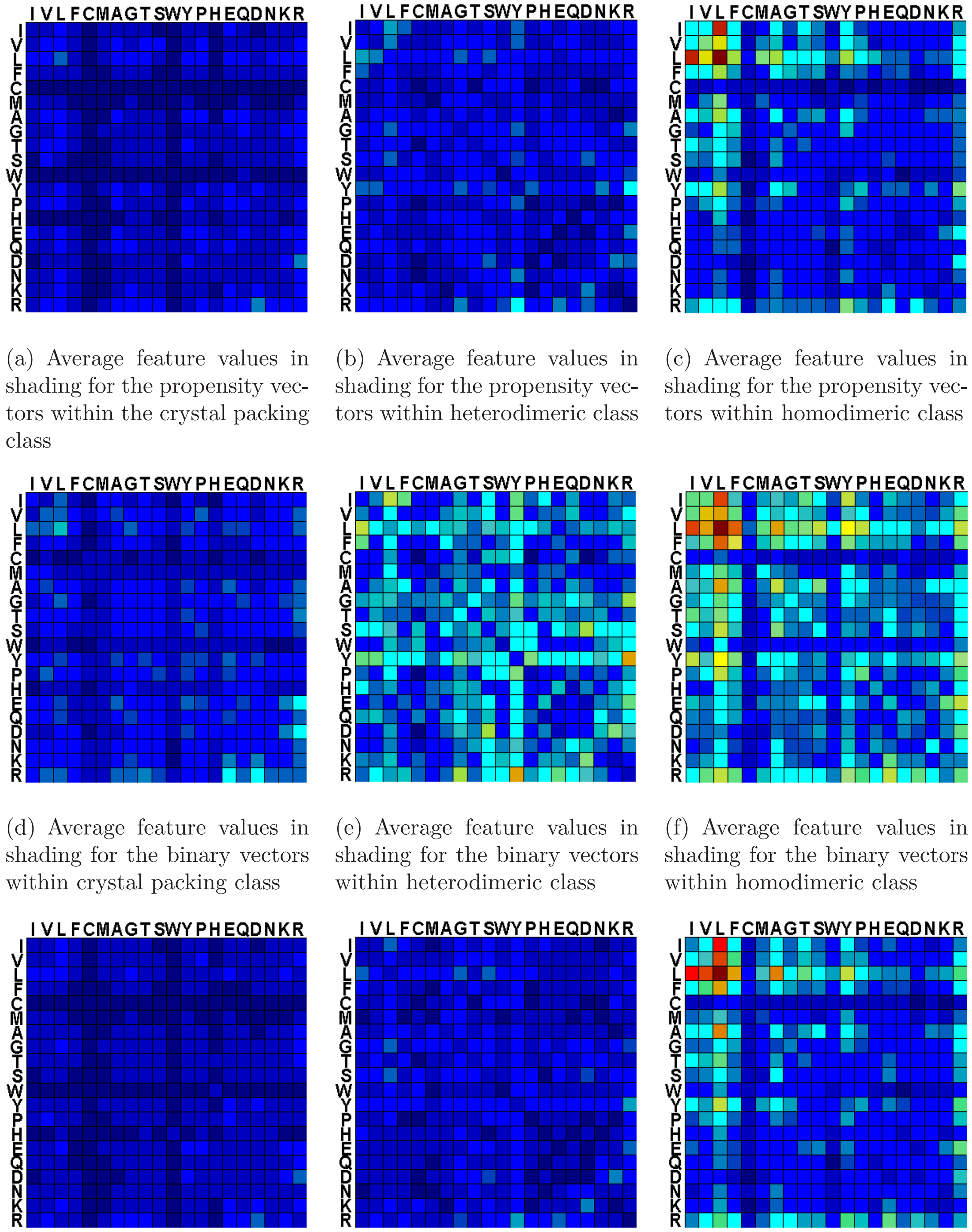
3.3 Propensity Vector in Comparison to Low-ASA Binary Vector and Low-ASA Frequency Vector

The performance of the propensity vectors is also compared to those achieved by low-ASA binary vectors and low-ASA frequency vectors, as shown in column 3 vs column 6, and in column 4 vs column 7 in Table III. First of all, the performance between the binary and frequency vector is very similar to each other, though in most cases the binary vector can achieve a bit better performance than the frequency vector. In fact, the accuracy difference between these two kinds of vectors is in the range from -2% to 6%. Take the performance on the Bahadur^{bio} dataset as an example, the binary vector method has an accuracy of 94.8% and outperforms the frequency vectors (89.5%). But on the Bahadur^{hocp} dataset, the frequency vectors have an 88.4% accuracy, slightly better than the accuracy 87.4% of the binary vector method.

OringPV outperforms both the binary vector and the frequency vector for almost every classification task involved in this work. For example, on the Bahadur dataset, our OringPV method achieved an accuracy of 85.8%, much higher than 81.1% by the binary vectors and 79.7% by the frequency vectors. The superior performance by OringPV is possibly attributed to two biological observations: (i) the residue composition in biological interaction interfaces is different from that in the rest of protein surfaces, but those in non-biological interactions are similar to each other[15,16,24,25]; (ii) the reside-residue pairing preference in obligate and transient interactions was also found to be different[19]. Thus, it is only propensity vectors rather than binary vectors or frequency vectors that can translate the ideas behind these facts into sharp discrimination power of a classifier, while binary vectors and frequency vectors are concentrated on the binding hot spots (O-ring-surrounded regions) only and unable to capture the relative difference between the binding interfaces and the rest of protein surfaces.

For a visual display of the subtle and deep discriminating power provided by OringPV, we draw a picture according to the average feature value of every residue pair within different interaction types. Let (r_i, r_j) be a residue pair and N be the total number of interactions in a class, for example, in the homodimeric class, then this average is calculated by $V_{(r_i,r_j)} = \frac{\sum_{k=1}^{N} V^k_{(r_i,r_j)}}{N}$, where $V^k_{(r_i,r_j)}$ is the propensity value, or binary value, or frequency value of the residue pair (r_i, r_j) of the kth interaction, depending on what kind of vectors is used in the classification. Similarly, we calculate such averages for the heterodimeric class, and for the crystal packing class. The visualization for the Bahadur dataset is shown in Figure 3. We also employ Wilcoxon signed rank test[44] to calculate the statistical significance of the difference for each pairs of vectors with the same kind. To show the process of the signed rank test, let's take for example propensity vectors of homodimeric interactions and crystal packing. Assume their propensity vector representations are $P^1 = \{p^1_{(r_i,r_j)}\}$ and $P^2 = \{p^2_{(r_i,r_j)}\}$ respectively where (r_i, r_j), $i \leq j, i, j = 1, ..., 20$, are residue pairs of residue r_i and residue r_j, and $p_{(r_i,r_j)}$ is propensity value. Then the significance level of their difference, p-value, is produced by (i) calculating the value difference of each feature $p^1_{(r_i,r_j)}$ and $p^2_{(r_i,r_j)}$, and ranking every feature based on the absolute value of the difference; (ii) restoring the signs of the differences to the ranks for obtaining the signed ranks; (iii) summing those ranks with positive signs to W_+; (iv) calculating the normalized test statistics by $T_{value} = \frac{W_+ - E(W_+)}{Var(W_+)}$, where $E(W_+) = \frac{n(n+1)}{4}$, $Var(W_+) = \frac{n(n+1)(2n+1)}{24}$, and n is the number of features with nonzero difference values; (v) obtaining p-value according to t-distribution with degree of freedom n, and value T_{value}; theoretically, smaller a p-value is, more likely a pair of vectors have significant difference. The similar way above can also be used to calculate p-value for the other two pairs of propensity vectors, and for pairs of binary vectors or of frequency vectors. Finally, the p-values for three pairs of three types of interactions based on propensity vectors, binary vectors and frequency vectors are shown in Table IV where 3 p-values in each row are for the pair of crystal packing and heterodimeric complexes, the pair of crystal packing and homodimeric proteins, and the pair of heterodimeric complexes and homodimeric proteins, respectively. It can be noted from Figure 3 and Table IV that: (i) the propensity vectors make clear distinction between the three types of protein interactions, and the smallest difference among three pairs of propensity vectors has p-value 1.11E-16 (see the three subfigures at the first row of Figure 3); (ii) the binary vectors show the difference between biological and non-biological interactions, but they sometimes confuse homodimeric proteins with heterodimeric complexes with p-value 4.61E-5 (see the middle three subfigures); (iii) the frequency vectors can clearly differentiate homodimeric proteins from others, but they maybe mislead the discrimination between crystal packing and heterodimeric complexes with p-value 2.0E-15 (see the three subfigures at the third row).

Overall in summary, the vectors with the ASA filtering, especially our OringPV method, outperform the vectors without the ASA restriction in most cases. In fact, both the pair propensity idea and the ASA filtering can sharpen the difference for the distinction of different types of interactions.



(a) Average feature values in shading for the propensity vectors within the crystal packing class

(b) Average feature values in shading for the propensity vectors within heterodimeric class

(c) Average feature values in shading for the propensity vectors within homodimeric class

(d) Average feature values in shading for the binary vectors within crystal packing class

(e) Average feature values in shading for the binary vectors within heterodimeric class

(f) Average feature values in shading for the binary vectors within homodimeric class

(g) Average feature values in shading for the frequency vectors within crystal packing class

(h) Average feature values in shading for the frequency vectors within heterodimeric class

(i) Average feature values in shading for the frequency vectors within homodimeric class

Fig. 3: A visual display of the discriminative power carried by the three types of vectors (propensity vector, low-ASA binary vector, and low-ASA frequency vector) for the distinction of protein interactions in the Bahadur dataset. In above figures, each row and column represents a different residue, and the residues are ordered according to their hydrophobicity with I as the most hydrophobic and R as the least hydrophobic. The colors from blue to red indicate the magnitude of values from the smallest to the largest.

TABLE IV: The p-values of Wilcoxon signed rank test for three pairs of the three types of interactions based on propensity vectors, binary vectors or frequency vectors.

Vectors	p-values		
	CP-PX	CP-HO	HO-PX
Propensity Vectors	1.11E-16	<1E-324	<1E-324
Binary Vectors	<1E-324	<1E-324	4.61E-5
Frequency vectors	2.0E-15	<1E-324	<1E-324

'CP', 'PX' and 'HO' represent crystal packing, heterodimeric complexes and homodimeric proteins respectively in the Bahadur dataset. Each column is one pair of the three different types of interactions.

3.4 Performance Comparison of OringPV with NOXclass, DiMoVo, and Other Methods

We have taken two comparison approaches: one is within-dataset comparison, and the second is cross-dataset comparison. In these experiments, the thresholds for relative ASA and for atomic distance (d_{tw}) are always fixed at 36% and 1.5 Å respectively. Parameters C and γ in the RBF kernel function of SVM are optimized in the training process of OringPV by cross-validation as what exactly done by the existing methods.

3.4.1 Within-Dataset Comparison with NOXclass, DiMoVo, and Other Methods

NOXclass[6] is a highly accurate algorithm trained on the BNCP-CS dataset to differentiate obligate from non-obligate interactions, and it also identifies crystal packing interactions. It is a multi-stage SVM prediction method, using interface properties such as interface area, ratio of interface area to protein surface area, and amino acid composition of the interface as input. For a fair comparison between the performance of OringPV and NOXclass on the BNCP-CS dataset, we take the two evaluation frameworks (EF) originally set by the NOXclass method in [6].

Under the first evaluation framework (EF1), the whole BNCP-CS dataset was used in choosing optimal parameters of SVM; then the performance of the classifier is measured also on the same data. (This approach is called maximized training dataset method[6].) The main objective of this kind of learning is to see whether a classifier can have an optimistic learning on all existing data despite of a possible overfitting problem. The result under this evaluation framework is shown in Table V. Overall in the distinction of the three types of protein interactions, OringPV achieved an accuracy of

TABLE V: The within-dataset performance comparison between OringPV and NOXclass.

Evaluation Framework	Method	Accuracy(%)		
		SVM 1	SVM 2	overall SVM
EF1	NOXclass	97.9	86.4	91.8
	OringPV	97.5	91.5	94.7
EF2	NOXclass	94.5	75.2	83.1
	OringPV	97.0	86.0	91.4

'SVM 1' and 'SVM 2' represent the first stage and the second stage of a SVM classifier. The performance of NOXclass is taken from [6].

94.7%, higher than NOXclass' 91.8%. This accuracy improvement is attributed to the outstanding performance of OringPV for distinguishing the two types of biological interactions (91.5% versus NOXclass' 86.4%).

The second evaluation framework (EF2) takes into consideration of testing performance. It was set as a 5-time 3-fold cross-validation integrated by a 10-fold cross-validation for training parameter selection[6]. Under this framework, the BNCP-CS dataset is randomly divided into three parts: iteratively each of the three parts for testing, and the other two parts for training in which the selection of parameter values is optimized by 10-fold cross-validation. Then this procedure is repeated five times to get an average performance of the classifier. The result is shown in Table V. It can be seen that our OringPV method has improved NOXclass' performance significantly from 83.1% to 91.4%. Again, this significant improvement comes from the sharper distinction capability of OringPV for distinguishing the two types of biological interactions.

On the Ponstingl dataset, a score cut-off method[21] misclassified 12 interactions, achieving an accuracy of 93.0%. The ACV method[5] (without symmetric consideration) also achieved the same level 93.0% accuracy. Our OringPV method achieved an accuracy of 94.7% (under the same evaluation framework). This is a slightly better performance than the two existing benchmark methods. Furthermore, the generalized error rate of 200 bootstrap samples for the atom-pair scoring schemes is 12.5%[21], which is worse than our OringPV method's 8.2% error rate under the LOOCV procedure.

On the Bahadur^{hocp} dataset, Bahadur *et al.* tested the performance of a new score cut-off method[15], which is a method based on interface area, shape and atomic packing density, residue propensity, etc. This score cut-off method[15] misclassified 17 of the total 310 interactions, while our OringPV method misclassified 18 interactions. More recently, Bernauer *et al.*[7] proposed a new method called DiMoVo. DiMoVo is a binary classifier developed specifically to discriminate crystal packing and biological interactions. It was evaluated on the Bahadur^{hocp} dataset and achieved a 95% accuracy under the LOOCV procedure with 0.5 as the cut-off score. Our OringPV method achieved a comparable performance (an accuracy of 94.2%) with optimal parameters.

With all these within-dataset comparison results, we can note that OringPV is much more accurate than NOXclass, and it is comparable to the score cut-off methods and DiMoVo. Our superior performance over these score cut-off methods and DiMoVo is presented in the following subsection.

3.4.2 Cross-Dataset Test for Performance Comparison

As introduced, cross-dataset test refers to an evaluation framework for a classifier where two datasets (usually from different authors) are given: one is used for training the classifier, and the other is for testing. This is a less-biased assessment to demonstrate the high reliability and generalization capability of a classifier. In this subsection, we compare the performance of OringPV under this evaluation framework with NOXclass and DiMoVo. The reason for choosing only NOXclass and DiMoVo is because their executable codes are available from the authors. However, both NOXclass and DiMoVo are final, user-end software programs built-in with the whole BNCP-CS and Bahadur^{hocp} dataset respectively in the training. To our best knowledge, we are unable to train them again on a different dataset, but they can be used to get a test accuracy for new datasets.

For a fair comparison to DiMoVo, we trained OringPV on the whole Bahadur^{hocp} dataset as well. Table VI shows the test performance of both OringPV and DiMoVo on the Ponstingl dataset and on the BNCP-CS^{hocp} dataset for distinguishing between obligate (homodimeric) interactions and crystal packing (monomers). We can see that on the Ponstingl dataset, our OringPV method and the DiMoVo method achieved a comparable accuracy and recall. However, on the BNCP-CS^{hocp} dataset, our OringPV method had a much better overall performance than the DiMoVo method (97.2% versus 89.0%), and especially on the homodimeric interactions, a significant recall improvement is from DiMoVo's 76% to 94.7%. This poor performance by DiMoVo is no surprise, and it is in agreement with the limitations of DiMoVo as discussed by the authors previously[7].

NOXclass is capable of conducting two kinds of classification tasks: (i) binary distinction between obligate and non-obligate interactions, and (ii) binary distinction between crystal packing (monomers) and biological interactions. For task (i), we used the whole BNCP-CS^{bio} dataset as OringPV's training data, and the Bahadur^{bio} dataset was used as the independent testing dataset

TABLE VI: The cross-dataset performance comparison between OringPV and DiMoVo.

Tested Dataset	Method	Recall CP	Recall HO	Accuracy(%)
BNCP-CS^{hocp}	DiMoVo	98.1	76.0	89.0
	OringPV	99.1	94.7	97.2
Ponstingl	DiMoVo	97.9	92.1	95.3
	OringPV	96.8	96.1	96.5

'Recall CP' and 'Recall HO' represent the recall of crystal packing and of homodimeric interactions respectively. The performance of DiMoVo is obtained from the website "http://cgal.inria.fr/DiMoVo/".

to assess the prediction performance of both NOXclass and OringPV. Based on the result shown in Table VII, we can see that the performance of OringPV is tremendously better than NOXclass. In particular, the recall rate on the non-obligate interactions is 20 points higher, and the accuracy is 13 points higher. Such an excellent performance is almost maintained by OringPV when it was trained on the Bahadur^{bio} dataset and tested on the BNCP-CS^{bio} dataset (shown in third row in Table VII). We were unable to report NOXclass' performance for this case, as its trained model is fixed on the BNCP-CS dataset only.

For task (ii), we trained OringPV on the BNCP-CS dataset and tested on the Ponstingl dataset. The accuracies by NOXclass and by OringPV are 86% and 87.7% respectively. However, when these two classifiers were used to predict whether the interactions in the Bahadur dataset are or not crystal packing, both performances were not good[6]. It may be due to that the non-biological interactions in the Bahadur dataset have large interfaces[6] similar to the size of biological interactions. This causes difficulties for the classifiers to learn necessary information to clearly classify biological and non-biological interactions when they are trained on the BNCP-CS dataset whose crystal packing generally have smaller interfaces than biological interactions.

We have further examined the capability of distinguishing homodimers and monomers by OringPV. We merged the two datasets of Bahadur^{hocp} and BNCP-CS^{hocp} as training data for OringPV. Then OringPV was tested on the Ponstingl dataset. The accuracy reached to 98.8%, an almost perfect accuracy (namely, only 1 homodimer and 1 monomer misclassified).

We note that the interpretation of the cross-dataset test results should be taken with some caution. The concern is the inter-dataset redundancy. Actually in this work, the redundancy among these three datasets does not play an important impact in the performance evaluation. This point can be verified as follows. OringPV's LOOCV accuracy (91.8%) on the Ponstingl dataset is less

TABLE VII: The cross-dataset performance comparison bewteen OringPV and NOXclass to distinguish the two types of biological interactions.

Tested Dataset	Method	Recall OB	Recall NO	Accuracy(%)
Bahadur^{bio}	NOXclass	84.4	78.6	82.3
	OringPV	93.4	98.6	95.3
BNCP-CS^{bio}	OringPV	97.3	79.0	89.1

'Recall OB' and 'Recall NO' represent the recall of obligate (homodimeric) interactions and non-obligate interactions (heterodimeric complexes) respectively. The performance of NOXclass is taken from [6].

TABLE VIII: The accuracy of OringPV in comparison to NOXclass and DiMoVo.

Tested Dataset	OringPV vs NOXclass (%)		OringPV vs DiMoVo (%)		Tested Dataset
(Training on BNCP-CS)	OringPV	NOXclass	OringPV	DiMoVo	(Training on Bahadur^{hocp})
BNCP-CS	*94.7^a (91.4^b)*	*91.8^a (83.1^b)*	97.2	89.0	BNCP-CS^{hocp}
Bahadur^{bio}	95.3	82.3	*94.2*	*95.0*	Bahadur^{hocp}
Ponstingl	87.7	86	96.5	95.3	Ponstingl

The *italic numbers* are for within-dataset comparison and the others for cross-dataset comparison. ^{a} and ^{b} stand for EF1 and EF2 respectively.

than the cross-dataset accuracy (97.7%) on the Ponstingl dataset when OringPV is trained on the Bahadur^{hocp} dataset. This may speculate some redundancy concern over the Ponstingl dataset and the Bahadur^{hocp} dataset. However, OringPV's LOOCV accuracy (92.9%) on the Bahadur^{hocp} dataset is higher than the cross-dataset accuracy on the Bahadur^{hocp} dataset when OringPV is trained on the Ponstingl dataset or on the BNCP-CS^{hocp} dataset. (The corresponding accuracy is 90% or 85.8%.)

To conclude Section 3.4, we use Table VIII to summarize the various and critical comparison results.

4 Insights into Low-ASA Residue Pairs and Propensity Vectors: A Discussion Based on Misclassifications

The comprehensive comparison results have already shown that OringPV is a highly accurate classifier to distinguish different types of protein interactions no matter the performance evaluation is by within-dataset, cross-dataset, or LOOCV. However, the focus on this section is different: we discuss

TABLE IX: Performance trend of a classifier when training datasets change.

Tested Dataset	Training Dataset (Accuracy %)		
	BNCP-CS^{hocp}	Ponstingl	Bahadur^{hocp}
Bahadur^{hocp}	85.8	90	*92.9*
Ponstingl	88.3	*91.8*	97.7
BNCP-CS^{hocp}	*98.3*	98.3	98.3

The *italic numbers* are for within-dataset comparison and others for cross-dataset comparison. In this table, the accuracies are a little different from Table VIII due to that the training datasets and training frameworks in Table VIII are under the same ones as what exactly done by the existing methods, such as leave-one-out learning with a 5-fold cross-validation procedure when OringPV is compared with DiMoVo in Table VIII. But OringPV is trained with leave-one-out cross-validation here.

why some interactions are still misclassified and give insights into the reasons.

We start with an interesting observation on the performance change when OringPV turned back to make predictions on the BNCP-CS^{bio} dataset. Recall from Table VII that OringPV achieved a testing accuracy of 95.3% on the Bahadur^{bio} dataset when it was trained on the BNCP-CS^{bio} dataset. However, its testing accuracy reduced to 89.1% on the BNCP-CS^{bio} dataset when it was trained on the Bahadur^{bio} dataset. To find out the reason behind this discrepancy, we examined the size of the O-ring-surrounded regions of the interactions in these two datasets. We obtained that on average, both the O-ring-surrounded region size of the interactions and the size variance in the BNCP-CS^{bio} dataset are much bigger than those in the Bahadur^{bio} dataset. This indicates that the diversity of the O-ring-surrounded regions in the BNCP-CS^{bio} dataset seems to overwhelm the cases in the Bahadur^{bio} dataset. Therefore, the testing performance on the BNCP-CS^{bio} dataset can be sacrificed as OringPV may not learn enough from the less-diversified Bahadur^{bio} dataset.

The second observation is about the big change of the testing performance on the same dataset when OringPV's training data is shift from one dataset to another. This can be seen from Table IX that: (i) the testing performance on Bahadur^{hocp} changes from 85.8% to 90% to 92.9% if OringPV's training data is switched from BNCP-CS^{hocp} to the Ponstingl dataset and to Bahadur^{hocp} (LOOCV is used if training data is the same as test data); (ii) the testing performance on the Ponstingl dataset changes from 88.3% to 91.8% to 97.7% if OringPV's training data is switched from BNCP-CS^{hocp} to the Ponstingl dataset to Bahadur^{hocp}; (iii) however, the OringPV's testing performance on BNCP-CS^{hocp} has no significant change and maintains at the high level of a 98.3% accuracy when the training dataset is switched from BNCP-CS^{hocp} to the Ponstingl dataset and to Bahadur^{hocp}. It seems that

Bahadur^{hocp} is the most accountable training dataset, and the Ponstingl dataset is the second most accountable, while BNCP-CS^{hocp} is less reliable, in this special classification task of distinguishing between obligate (homodimeric) interactions and crystal packing.

To understand the deep reasons, we examine two factors: the average size of O-ring-surrounded regions in each dataset, and the propensity vector values of O-ring-surrounded regions with various sizes. Our examination shows that the average size for O-ring-surrounded regions of the crystal packing in the Bahadur^{hocp} dataset and the size variance are both bigger than those in the Ponstingl dataset, which are both bigger than those in BNCP-CS^{hocp}. On the contrary, the average size for O-ring-surrounded regions of the homodimeric interactions in the Bahadur^{hocp} dataset and the size variance are both similar to those in the Ponstingl dataset, which are both smaller than those in BNCP-CS^{hocp}. This means that crystal packing and homodimeric interactions within the Bahadur^{hocp} dataset are more complicated though separable than those within BNCP-CS^{hocp} and the Ponstingl dataset are. Therefore, OringPV, especially propensity part of propensity vectors, can learn better if its training data is switched to the Bahadur^{hocp} dataset. On the other hand, smaller O-ring-surrounded regions are easier to produce random sharp propensity values in a vector due to that the propensity vectors have 210 dimensions of residue pairs together describing the whole interfaces. Further, the dataset with smaller O-ring-surrounded regions and also with smaller size variance might make the propensity part of propensity vectors less helpful in the classification process. It is why the testing performance is better when OringPV is trained on the datasets whose crystal packing have larger O-ring-surrounded regions and larger size variance, such as the Bahadur^{hocp} dataset, than when OringPV is trained on the datasets whose crystal packing have smaller O-ring-surrounded regions and smaller size variance, such as the BNCP-CS^{hocp} dataset.

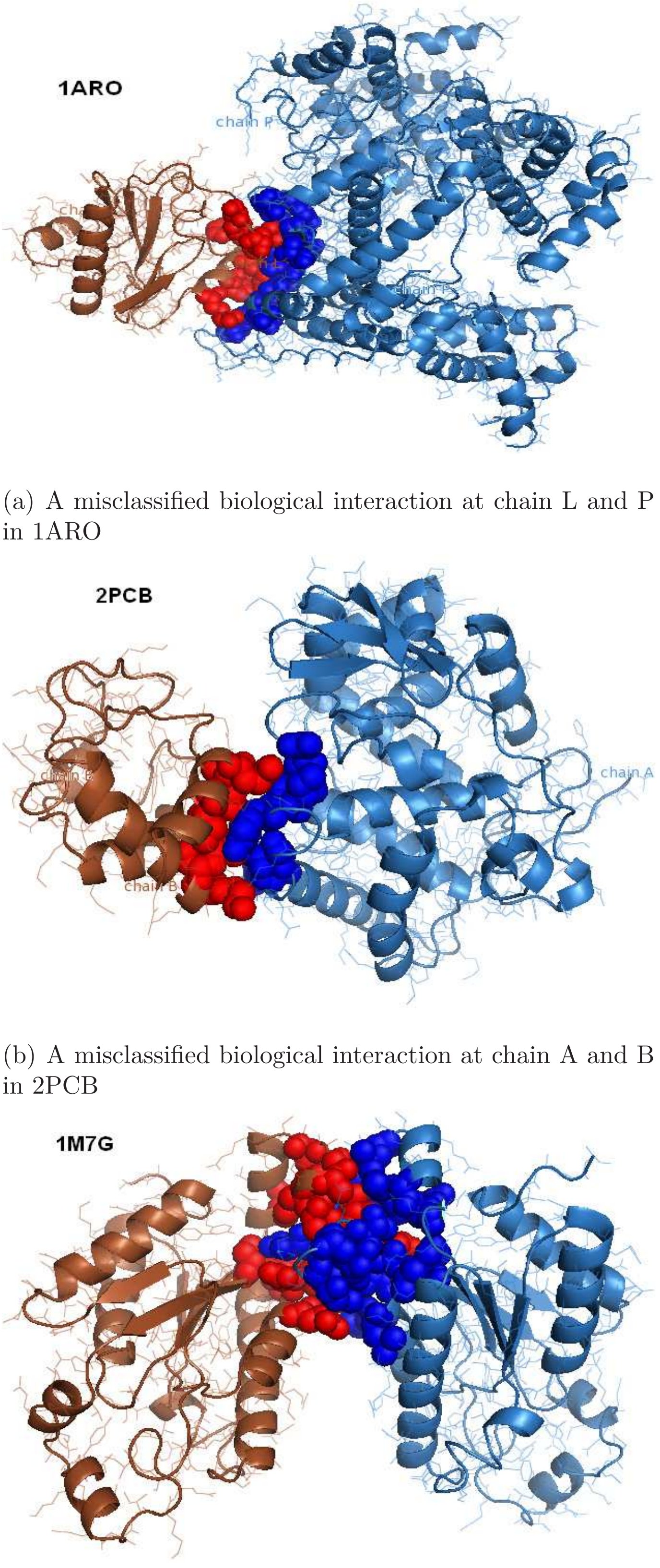
Our third observation is as follows. A biological interaction is easy to be wrongly predicted as crystal packing if its O-ring-surrounded region is small, and a crystal packing is also easy to be wrongly predicted as biological interaction if its O-ring-surrounded region is large. We show three examples taken from the BNCP-CS dataset in the process of identifying biological interactions from crystal packing. Under LOOCV, OringPV misclassifies 8 interactions (2 non-obligate interactions are wrongly predicted as non-biological interactions, and 1 crystal packing as obligate, and 5 crystal packing as non-obligate interactions). If OringPV is trained on the Bahadur dataset, it wrongly classifies 12 interactions (2 obligate interactions and 7 non-obligate interactions are grouped into non-biological interactions, and 1 and 2 crystal packing into obligate and non-obligate interactions respectively). Under these two prediction approaches, there are two common biological interactions misclassified as non-biological interactions, and one common crystal packing that is wrongly predicted as a biological interaction. Interestingly, DiMoVo also wrongly classifies these three interactions.

Of the two wrongly predicted biological interactions, one is in the PDB entry 1ARO between chains L (T7 lysozyme) and P (T7 RNA polymerase), and the other is in 2PCB (A complex between electron transfer partners) between chains A (Cytochrome C peroxidase) and B (Cytochrome C). Figure 4(a)(b) show their structures. Recall that the propensity vectors are based on the number of O-ring-surrounded residues and the number of surface residues. We examine the number of surface residues, N_S, of 1ARO and of 2PCB, in comparison to the number of residues, N_O, in their O-ring-surrounded regions. The investigation indicates that N_O/N_Ss of 1ARO and of 2PCB are the minimum in all biological interactions. A smaller N_O/N_S means that residue pairs in the O-ring-surrounded region are too fewer to make propensity values outstanding. We also observed that when N_O/N_S is less than 0.1, 3 out of 4 biological interactions are wrongly classified. The misclassified prediction is likely due to that those O-ring-surrounded regions of misclassified interactions are so small that the propensity part of propensity vectors has many random sharp propensity values, thus misleading the true propensity properties of those residue pairs.

The PDB entry of the one commonly misclassified crystal packing is 1M7G (APS kinase from Penicillium Chrysogenum: ternary structure with ADP and APS) whose structure is shown in Figure 4(c). Similar to the above data analysis, N_O/N_S of 1M7G is found with the maximum N_O/N_S in all crystal packing. Furthermore, there are 7 misclassified non-biological interactions *(by the LOOCV procedure or by OringPV trained on the Bahadur dataset)* among the top 11 maximum of N_O/N_S descending ranking of crystal packing. The N_O/N_S values of those 7 misclassified crystal packing are all larger than 0.13. A possible reason is that crystal packing can occasionally form 'abnormal' large O-ring-surrounded regions, which computationally results in propensity values similar to those of biological interactions.

5 Conclusion

In this work, we have introduced low-ASA residue pairs and O-ring-surrounded regions. The biological principle of this notion is based on the long standing O-ring theory and the coupling proposition. The water accessibility restriction and the contact distance are both considered in the definition of



(a) A misclassified biological interaction at chain L and P in 1ARO

(b) A misclassified biological interaction at chain A and B in 2PCB

(c) The misclassified crystal packing in 1M7G

Fig. 4: The interactions in (a) and (b) are biological interactions but wrongly predicted to crystal packing; the interaction in (c) is crystal packing but wrongly predicted to biological interaction. The O-ring-surrounded regions are in the 'spheres' view in PyMOL with red and blue colors.

low-ASA residue pairs. Thus, with this definition, the properties of binding hot spot residues are fully integrated into such residue pairs. We also introduced propensity vectors of low-ASA residue pairs and have suggested to use these propensity vectors to characterize the different types of protein interactions.

The OringPV method, our newly proposed learning scheme with propensity vectors as the input of SVM, has shown excellent performance in the prediction of the three types of protein interactions. The experiments are conducted on three benchmark datasets: the BNCP-CS dataset[6], the Ponstingl dataset[21], and the Bahadur dataset[15]. The performance is evaluated under the LOOCV procedure, and also under the comparison frameworks such as within/cross-dataset tests in comparison to widely accepted literature methods, including NOXClass[6] and DiMoVo[7]. The evaluation results demonstrate that the propensity vectors can signify important characteristics of protein interactions, and OringPV is highly accurate to identify biological interactions from non-biological interactions and to distinguish different types of biological interactions.

As a future work, OringPV perhaps can be used to determine and rank the fitness scores of all possible binding structures constructed by docking algorithms[45]. We also consider to apply low-ASA residue pairs and the propensity idea to deal with hot spot or interface prediction problems. In fact, those problems are similar to the current one, though they are beyond the scope of the current work. One consideration is that we construct propensity vectors for interacting chain pairs that contain a hot spot as the current work does. Second, we construct propensity vectors for interacting chain pairs where non-hot spots are identified. These two classes of propensity vectors, labeled with hot spot or non-hot spot, can be then used to train a classifier to predict whether a cluster of contact residue pairs is or not a hot spot. However, one difficulty of this future work is that we are lack of experimental data of non-hot spots. Perhaps, one-class learning algorithms[46] are useful. We leave all these details for readers who are interested in those problems.

Acknowledgement

This research work was funded by a Singapore MOE ARC Tier-2 grant (T208B2203).

References

[1] Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., and Nussinov, R. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. Brief Bioinform 2009; 10:217–232.

[2] Nooren, I. M. and Thornton, J. M. Diversity of protein-rotein interactions. EMBO J 2003; 22(14):3486–3492.

[3] Tsai, C. J., Xu, D., and Nussinov, R. Protein folding via binding and vice versa. Fold Des 1998; 3(4):R71–R80.

[4] Ofran, Y. and Rost, B. Analysing six types of protein-protein interfaces. J Mol Biol 2003; 325(2):377–387.

[5] Mintseris, J. and Weng, Z. Atomic contact vectors in protein-protein recognition. Proteins 2003; 53(3):629–639.

[6] Zhu, H., Domingues, F. S., Sommer, I., and Lengauer, T. NOXclass: prediction of protein-protein interaction types. BMC Bioinformatics 2006; 7.

[7] Bernauer, J., Bahadur, R. P. P., Rodier, F., Janin, J., and Poupon, A. DiMoVo: a voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. Bioinformatics 2008; 24:652–8.

[8] Moont, G., Gabb, H. A., and Sternberg, M. J. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 1999; 35(3):364–373.

[9] Neuvirth, H., Raz, R., and Schreiber, G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 2004; 338(1):181–199.

[10] Bradford, J. R. and Westhead, D. R. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics 2005; 21(8):1487–1494.

[11] Valdar, W. S. J. and Thornton, J. M. Conservation helps to identify biologically relevant crystal contacts. J Mol Biol 2001; 313(2):399–416.

[12] Young, L., Jernigan, R. L., and Covell, D. G. A role for surface hydrophobicity in protein-protein recognition. Protein Sci 1994; 3(5):717–729.

[13] Jones, S. and Thornton, J. M. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996; 93(1):13–20.

[14] Bahadur, R. P., Chakrabarti, P., Rodier, F., and Janin, J. Dissecting subunit interfaces in homodimeric proteins. Proteins 2003; 53(3):708–719.

[15] Bahadur, R. P., Chakrabarti, P., Rodier, F., and Janin, J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 2004; 336(4):943–955.

[16] Jones, S. and Thornton, J. M. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997; 272(1):121–132.

[17] De, S., Krishnadev, O., Srinivasan, N., and Rekha, N. Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. BMC Struct Biol 2005; 5.

[18] Saha, R. P., Bahadur, R. P., and Chakrabarti, P. Interresidue contacts in proteins and protein-protein interfaces and their use in characterizing the homodimeric interface. J Proteome Res August 2005; 4:1600–1609.

[19] Lukman, S., Sim, K., Li, J., and Chen, Y.-P. P. Interacting amino acid preferences of 3d pattern pairs at the binding sites of transient and obligate protein complexes. In *APBC* 2008; 69–78.

[20] Glaser, F., Steinberg, D. M., Vakser, I. A., and Ben-Tal, N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 2001; 43(2):89–102.

[21] Ponstingl, H., Henrick, K., and Thornton, J. M. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000; 41(1):47–57.

[22] Ponstingl, H., Kabir, T., and Thornton, J. M. Automatic inference of protein quaternary structure from crystals. Journal of Applied Crystallography 2003; 36(5):1116–1122.

[23] Zhang, C., Vasmatzis, G., Cornette, J., and DeLisi, C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997; 267:707–726(20).

[24] Lo Conte, L., Chothia, C., and Janin, J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999; 285(5):2177–2198.

[25] Janin, J., Miller, S., and Chothia, C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 1988; 204(1):155–164.

[26] Carugo, O. and Argos, P. Protein-protein crystal-packing contacts. Protein sci 1997; 6(10):2261–3.

[27] Janin, J. and Rodier, F. Protein-protein interaction at crystal contacts. Proteins 1995; 23(4):580–587.

[28] Janin, J. Specific versus non-specific contacts in protein crystals. Nature Structural Biology 1997; 4:973–974.

[29] Liu, S., Li, Q., and Lai, L. A combinatorial score to distinguish biological and nonbiological protein-protein interfaces. Proteins 2006; 64:68–78.

[30] Block, P., Paern, J., Hullermeier, E., Sanschagrin, P., Sotriffer, C. A., and Klebe, G. Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms. Proteins 2006; 65(3):607–622.

[31] Bernauer, J., Azé, J., Janin, J., and Poupon, A. A new protein-protein docking scoring function based on interface residue properties. Bioinformatics 2007; 5(23):555–62.

[32] Bernauer, J., Poupon, A., Azé, J., and Janin, J. A docking analysis of the statistical physics of protein-protein recognition. Phys Biol 2005; 2(1-2).

[33] Bogan, A. A. and Thorn, K. S. Anatomy of hot spots in protein interfaces. J Mol Biol 1998; 280(1):1–9.

[34] Thorn, K. S. and Bogan, A. A. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 2001; 17:284–285(3).

[35] DeLano, W. L. Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 2002; 12(1):14–20.

[36] Moreira, I. S. S., Fernandes, P. A. A., and Ramos, M. J. J. Hot spots-A review of the protein-protein interface determinant amino-acid residues. Proteins 2007; 68(4):803–812.

[37] Halperin, I., Wolfson, H., and Nussinov, R. Protein-protein interactions: Coupling of structurally conserved residues and of hot spots across interfaces. implications for docking. Structure 2004; 12(6):1027–1038.

[38] Li, J. and Liu, Q. 'Double water exclusion': a hypothesis refining the O-ring theory for the hot spots at protein interfaces. Bioinformatics 2009; 25(6):743–750.

[39] Hubbard, S. J. and Thornton, J. M. 'NACCESS', computer program. Technical report, Department of Biochemistry Molecular Biology, University College London, 1993.

[40] Pollastri, G., Baldi, P., Fariselli, P., and Casadio, R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002; 47(2):142–153.

[41] Cristianini, N. and Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press 2000.

[42] Chang, C. C. and Lin, C. J. LIBSVM: a library for support vector machines, 2001.

[43] Hsu, C., Chang, C., and Lin, C. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, 2003.

[44] Rice, J. A. Mathematical Statistics and Data Analysis. Duxbury Press 2001.

[45] Sobolev, V., Wade, R. C., Vriend, G., and Edelman, M. Molecular docking using surface complementarity. Proteins 1996; 25:120–129.

[46] Japkowicz, N. Concept learning in the absence of counterexamples: an autoassociation-based approach to classification. PhD thesis, New Brunswick, NJ, USA, 1999. Director-Hanson, Jose and Director-Kulikowski, Casimir.