Analysis of Obligate and Non-obligate Complexes using Desolvation Energies in Domain-domain Interactions

Mina Maleki , Md. Mominul Aziz , and Luis Rueda
School of Computer Science
University of Windsor
401 Sunset Avenue, Windsor, Ontario N9B 3P4, Canada
{maleki,azizc,lrueda}@uwindsor.ca

ABSTRACT

Protein-protein interactions (PPI) are important in most biological processes and their study is crucial in many applications. Identification of types of protein complexes is a particular problem that has drawn the attention of the research community in the past few years. We focus on obligate and non-obligate complexes, their prediction and analysis. We propose a prediction model to distinguish between these two types of complexes, which uses desolvation energies of domain-domain interactions (DDI), pairs of atoms and amino acids present in the interfaces of such complexes. Principal components of the data were found and then the prediction is performed via linear dimensionality reduction (LDR) and support vector machines (SVM). Our results on a newly compiled dataset, namely binary-PPID, which is a joint and modified version of two well-known datasets consisting of 146 obligate and 169 non-obligate complexes, show that the best prediction is achieved with SVM (77.78%) when using desolvation energies of atom type features. Furthermore, a detailed analysis shows that different DDIs are present in obligate and non-obligate complexes, and that homo-DDIs are more likely to be present in obligate interactions.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation, Feature evaluation and selection*

General Terms

Algorithms, Performance, Experimentation

Keywords

protein-protein interaction; domain-domain interaction; complex type prediction

1. INTRODUCTION

Protein interactions are important in many essential biological processes in living cells, including signal transduction, transport, cellular motion and gene regulation. As a consequence of this, the identification of protein-protein interactions (PPIs) is a key topic in life science research. Prediction of PPIs has been studied mostly using computational approaches and from many different perspectives. Prediction of interfaces (interactions between subunits) in different molecules includes analysis of patches, sites, amino acids, or even specific atoms. The physicochemical and geometric arrangement of subunits in protein complexes is best known as docking. An important aspect that has recently drawn the attention of the research community is to predict "when" the interactions will occur – this is mostly studied at the protein interaction network level. Another important aspect in studying PPIs is the identification of different types of complexes, including similarities between subunits (homo/hetero-oligomers), number of subunits involved in the interaction (dimers, trimers, etc.), duration of the interaction (transient vs. permanent), stability of the interaction (non-obligate vs. obligate), among others; we focus on the latter problem.

Obligate interactions are usually considered as permanent, while non-obligate interaction can be either permanent or transient [1]. Non-obligate and transient interactions are more difficult to study and understand due to their instability and short life, while obligate and permanent interactions last for a longer period of time, and hence are more stable [2]. For these reasons, an important problem is to distinguish between obligate and non-obligate complexes. To study the behavior of obligate and non-obligate interactions, in [3], it was shown that non-obligate complexes are rich in aromatic residues and arginine, while depleted in other charged residues. The study of [4] suggested that mobility differences of amino acids are more significant for obligate and large interface complexes than for transient and medium-sized ones.

Some studies in PPI consider the analysis of a wide range of parameters, including desolvation energies, amino acid composition, conservation, electrostatic energies, and hydrophobicity for predicting obligate and non-obligate complexes. In [1], a classification of obligate and non-obligate interactions was proposed where interactions are classified based on the lifetime of the complex. In [5], three different types of interactions were studied, namely crystal packing, obligate and non-obligate interactions. That study was based on using solvent accessible surface area, conservation scores, and the shapes of the interfaces. After classifying obligate and transient protein interactions based on 300 different interface attributes in [6], the difference in molecular weight between interacting chains was reported as the best single feature to distinguish transient from obligate interactions. Based on their results, interactions with the same molecular weight or large interfaces are obligate.

Different studies have claimed that only a few highly conserved residues are crucial for protein interactions [7, 8]. Moreover, it has been shown that physical interactions between proteins are mostly

BIOKDD 2011, August 2011, San Diego, CA, USA
Copyright 2011 ACM 978-1-4503-0839-7 ...$10.00.

controlled by their domains, as a domain is often the minimal and fundamental module corresponding to a biochemical function [7, 8]. Thus, in previous studies, the physical interaction between proteins is analyzed in terms of the interaction between residues of their structural domains. For example, in [7], interactions between residues were used for finding obligate and non-obligate residue contacts of PPIs. That study concluded that non-obligate interfaces occupy less than 2% of the area of the domain surfaces, while the number of obligatory interfaces is between 0–6%. In [8], the interface of 750 transient DDIs, interactions between domains that are part of different proteins, and 2,000 obligate interactions were studied. The interactions between domains of one amino acid chain were analyzed to obtain a better understanding of molecular recognition and identify frequent amino acids in the interfaces and on the surfaces of PPIs. Also, in [9], the domain information from protein complexes was used to predict four different types of PPIs including transient enzyme inhibitor/non enzyme inhibitor and permanent homo/hetero obligate complexes.

In a recent work [10], an approach to distinguish between obligate and non-obligate complexes has been proposed in which desolvation energies of amino acids and atoms present in the interfaces of PPIs are considered as the input features of the classifiers. The results of that classifier show that desolvation energies are better discriminant than solvent accessibility and conservation properties. In this paper, we present an analysis of PPIs that uses properties of DDIs present in the interface to predict obligate and non-obligate protein-protein interactions. Desolvation energies of atom and amino acid pairs present in the interface of DDIs as well as desolvation energies of all atom and amino acid pairs present in the interface of interacting complexes are used in the prediction. We have also performed an analysis on the DDIs present in the two types of interactions. A visual analysis shows that that unique pairs can be identified for both types of interactions, and highlight the presence of homo-DDIs in obligate interactions. The prediction approach resorts on two state-of-the-art classification techniques of linear dimensionality reduction (LDR) and support vector machines (SVM). Ten-fold cross validation of the proposed scheme on our binary-PPID dataset, which is an extended dataset that we compiled from two well-known datasets of [5] and [11], demonstrates that (a) using desolvation energies of atom type features are better than the features used in [5] for predicting obligate and non-obligate complexes, achieving 77.78% classification accuracy in comparison to 71.80% (b) atom type features are better than amino acid type features for prediction of these two types of complexes (c) although the prediction accuracies by considering atom and amino acid pairs present in the interacting domains instead of all interacting atom and amino acid pairs of two chains are low, they are still acceptable and provide additional information about the specific domains.

2. MATERIALS AND METHODS

2.1 Dataset

We have compiled a new dataset by merging two existing, pre-classified datasets of protein complexes obtained from the studies of Zhu et al. [5], and Mintseris and Weng [11]. The former dataset contains 75 obligate and 62 non-obligate interactions while the latter contains 115 obligate and 212 transient interactions. There are 39 common interactions between these two datasets and hence the redundant complexes were removed. In addition, we carefully examined all the interactions and removed complexes with contradicting class labels. For example "*1eg9,A:B*" is classified as both obligate and non-obligate in [5] and [11]. In total, seven complexes: *1eg9, 1hsa, 1i1a, 1raf, 1d09, 1jkj* and *1cqi*, showed this contradiction and were then removed from the new dataset. After this pre-processing stage, the new dataset resulted in 417 complexes from which 182 were obligate and 235 were non-obligate. In this study, each complex is considered as the interaction of two chains (two single sub-units). Since the dataset of [11] considers the interaction of two units in which each may contain more than one chain, e.g., "*1qfu,AB:HL*", all these complexes were converted to interactions between two single chains (binary interactions). For this, all binary interactions of each of the 93 multiple-chain complexes were identified, obtaining 289 interactions, and each of these was converted into a separate complex in the new dataset. For example, the multiple-chain of *1qfu* was transformed to four binary chains as follows: *A:H*, *A:L*, *B:H* and *B:L*. Another step involves taking the whole dataset of binary complexes and filtering non-interacting pairs. Using the interface definition of [12], complexes with interacting chains with less than five interface residues were removed. Two residues (from different chains) are considered to be interacting if at least one pair of atoms from these residues is 5Å or less apart from each other. This resulted in a dataset that contains 516 complexes, from which 303 are non-obligate and 213 are obligate binary interactions. In a final step, we collected the domains contained in each interacting chain from the Pfam database [13]. The complexes that do not have any domain in at least one of their subunits were discarded in the analysis. This resulted in our final dataset of 315 complexes, from which 146 are obligate complexes and 169 are non-obligate complexes - we call this dataset binary protein-protein interactions by considering domain definitions (binary-PPID). The PDB IDs of these complexes and the interacting chains are shown in Table 1.

2.2 Features

We use desolvation energies as the predicting properties, which are shown to be very efficient for prediction of obligate and non-obligate complexes [10]. Knowledge-based contact potential that accounts for hydrophobic interactions, self-energy change upon desolvation of charged and polar atom groups, and side-chain entropy loss compose the so-called binding-free energy. In [14], the total desolvation energy is defined as follows:

$$\Delta G_{des} = g(r)\Sigma\Sigma e_{ij}. \tag{1}$$

If we are considering the interaction between the i^{th} atom of a ligand and the j^{th} atom of a receptor then e_{ij} is the atomic contact potential (ACP) [15] between them, and $g(r)$ is a smooth function based on their distance. The value of $g(r)$ is 1 for atoms that are less than 5 Å apart [14]. For simplicity, we consider the smooth function to be linear. Within the range of 5 and 7 Å, the value of $g(r)$ is $(7 - r)/2$.

We collected the structural data from the Protein Data Bank (PDB) [16] for each complex in our dataset. After adding domain information obtained from Pfam to each atom present in the chain, each PDB file was divided into two different ligand and receptor files based on its side chains. From [15], we know that there are 18 atom types. Thus, for each protein complex a feature vector with 18^2 values was obtained, where each feature contains the desolvation energy of a pair of atom types. As the order of interacting atom pairs is not important, the final length of feature vector for each complex was 171 that correspond to unique pairs. We also considered pairs of amino acids, and for this, we computed desolvation energy values for each pair of atoms using Eq. (1) and accumulated the values for each pair of amino acids. Avoiding repeated pairs resulted in 210 different features (unique pair of amino acids).

Table 1: binary-PPID dataset (146 obligate and 169 non-obligate binary complexes).

Obligate Complexes

1a0f , A:B	1byk , A:B	1eex , A:B	1hcn , A:B	1jk0 , A:B	1li1 , A:C	1qbi , A:B	2hdh , A:B
1a6d , A:B	1c3o , A:B	1eex , A:G	1hfe , L:S	1jk8 , A:B	1li1 , B:C	1qdl , A:B	2hhm , A:B
1ahj , A:B	1c7n , A:B	1efv , A:B	1hgx , A:B	1jkm , A:B	1lti , C:G	1qfe , A:B	2kau , A:C
1aj8 , A:B	1ccw , A:B	1ep3 , A:B	1hjr , A:C	1jmx , A:G	1lti , C:H	1qfh , A:B	2kau , B:C
1ajs , A:B	1cmb , A:B	1ezv , D:H	1hr6 , A:B	1jnr , A:B	1lti , C:D	1qu7 , A:B	2min , A:B
1aq6 , A:B	1cnz , A:B	1ezv , C:F	1hss , A:B	1jro , A:B	1lti , C:F	1sgf , A:B	2mta , A:H
1b34 , A:B	1coz , A:B	1f6y , A:B	1ihf , A:B	1jwh , A:C	1lti , C:E	1sgf , A:Y	2nac , A:B
1b3a , A:B	1cpc , A:B	1ffu , A:C	1jb0 , B:E	1jwh , A:D	1luc , A:B	1spp , A:B	2pfl , A:B
1b4u , A:B	1dce , A:B	1ffv , A:B	1jb0 , B:E	1k3u , A:B	1mro , A:B	1spu , A:B	2utg , A:B
1b5e , A:B	1dii , A:C	1fm0 , D:E	1jb0 , B:D	1k8k , A:B	1mro , B:C	1trk , A:B	3gtu , A:B
1b7b , A:C	1dj7 , A:B	1g8k , A:B	1jb0 , B:D	1k8k , B:F	1mro , A:C	1vcb , A:B	3pce , A:M
1b7y , A:B	1dkf , A:B	1gka , A:B	1jb0 , A:E	1k8k , A:E	1msp , A:B	1vlt , A:B	3tmk , A:B
1b8j , A:B	1dm0 , A:D	1go3 , E:F	1jb0 , A:E	1k8k , C:F	1poi , A:B	1wgj , A:B	4rub , A:T
1b8m , A:B	1dm0 , A:E	1gpe , A:B	1jb0 , A:C	1k8k , D:F	1pp2 , L:R	1xso , A:B	
1b9m , A:B	1dor , A:B	1gpw , A:B	1jb0 , C:E	1kpe , A:B	1prc , C:H	1ypi , A:B	
1be3 , G:A	1dtw , A:B	1gux , A:B	1jb0 , B:C	1kqf , B:C	1prc , C:L	1ytf , C:D	
1bjn , A:B	1dxt , A:B	1h2a , L:S	1jb0 , A:D	1ktd , A:B	1prc , C:M	2aai , A:B	
1brm , A:B	1e8o , A:B	1h2r , L:S	1jb0 , A:D	1l7v , A:C	1qae , A:B	2ae2 , A:B	
1byf , A:B	1e9z , A:B	1h8e , A:D	1jb0 , C:D	1ld8 , A:B	1qax , A:B	2ahj , A:B	

Non-obligate Complexes

1a14 , L:N	1bml , A:C	1eai , A:C	1fq1 , A:B	1i4d , B:D	1jsu , B:C	1n2c , B:E	2btc , E:I
1a14 , H:N	1buh , A:B	1eay , A:C	1fqj , A:C	1i4d , A:D	1jsu , A:C	1n2c , A:E	2btf , A:P
1a2k , B:C	1c1y , A:B	1ebd , A:C	1frv , A:B	1i7w , A:B	1jtg , A:B	1n2c , B:F	2mta , A:L
1a4y , A:B	1c4z , A:D	1ebd , B:C	1fss , A:B	1i85 , B:D	1jw9 , B:D	1nbf , A:D	2mta , A:C
1acb , E:I	1cc0 , A:E	1eer , A:B	1gaq , A:B	1i8l , A:C	1k5d , A:B	1nf5 , A:B	2mta , H:L
1agr , E:A	1cgi , E:I	1efu , A:B	1gcq , B:C	1ib1 , B:E	1kcg , A:C	1noc , A:B	2pcb , A:B
1akj , B:D	1cmx , A:B	1efx , A:D	1gh6 , A:B	1ib1 , A:E	1kcg , B:C	1pdk , A:B	2pcc , A:B
1akj , A:D	1cs4 , A:C	1eja , A:B	1gl1 , A:I	1icf , B:I	1kkl , A:H	1qbk , B:C	2prg , B:C
1ar1 , B:D	1cs4 , B:C	1es7 , C:B	1gla , F:G	1ijk , A:B	1kkl , C:H	1qgw , A:C	2sic , E:I
1avg , H:I	1cse , I:E	1es7 , A:B	1gp2 , A:B	1ijk , A:C	1kmi , Y:Z	1rlb , A:E	2tec , E:I
1avw , A:B	1cvs , A:C	1eth , A:B	1grn , A:B	1is8 , C:M	1kxp , A:D	1rlb , C:E	3hhr , A:B
1avx , A:B	1d4x , A:G	1euv , A:B	1gvn , A:B	1is8 , B:L	1kyo , O:W	1rlb , B:E	3sgb , E:I
1avz , B:C	1d5x , A:C	1evt , A:C	1gzs , A:B	1is8 , E:O	1lb1 , A:B	1rrp , A:B	3ygs , C:P
1awc , A:B	1de4 , C:A	1f02 , I:T	1h2k , A:S	1is8 , D:N	1lpb , A:B	1stf , E:I	4htc , H:I
1ay7 , A:B	1dev , A:B	1f34 , A:B	1h59 , A:B	1is8 , A:K	1m10 , A:B	1t7p , A:B	4sgb , E:I
1azz , A:D	1df9 , B:C	1f3v , A:B	1hlu , A:P	1is8 , D:O	1m1e , A:B	1tab , E:I	
1azz , A:D	1dfj , E:I	1f80 , A:E	1hwg , A:C	1is8 , A:L	1m4u , A:L	1tgs , I:Z	
1b6c , A:B	1doa , A:B	1fak , H:T	1hwg , A:B	1is8 , E:K	1mah , A:F	1toc , B:R	
1b9y , A:C	1du3 , A:D	1fg9 , B:C	1hzz , B:C	1is8 , C:N	1mbu , A:C	1uea , A:B	
1bdj , A:B	1du3 , A:F	1fg9 , A:C	1i2m , A:B	1is8 , B:M	1ml0 , A:D	1wq1 , G:R	
1bi8 , A:B	1dx5 , M:I	1fin , A:B	1i3o , D:E	1itb , A:B	1mr1 , A:D	1ycs , A:B	
1bkd , R:S	1e6e , A:B	1fle , E:I	1i3o , B:E	1jch , A:B	1n2c , A:F	1zbd , A:B	

Table 2: Description of the subsets of features used in this study.

Name	Feature Type	Interacting Chains	DDIs
PPID-AT	atom type	✓	-
PPID-AA	amino acid	✓	-
PPID-ATD	atom type	-	✓
PPID-AAD	amino acid	-	✓

A posterior step was to identify the 317 unique domains present in the interface of at least one complex in the dataset. Considering all pairs of domains, the desolvation energies for all atoms and amino acids present in each interacting domains were calculated using Eq. (1) and finally each complex had 171 atom type and 210 amino acid type features. By using desolvation energies for different types of features, four subsets of features for prediction and evaluation were generated (Table 2). The names of the subsets are PPID-X where X is AT for atom type, AA for amino acid pairs, ATD for atoms in interacting domains (DDIs) or AAD for amino acid pairs in interacting domains.

2.3 Prediction Methods

2.3.1 Linear Dimensionality Reduction

One of the approaches we have used for prediction is LDR. The basic idea of LDR is to represent an object of dimension n as a lower-dimensional vector of dimension d, achieving this by performing a linear transformation. We consider two classes, ω_1 and ω_2, represented by two normally distributed random vectors $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{S}_1)$ and $\mathbf{x}_2 \sim N(\mathbf{m}_2, \mathbf{S}_2)$, respectively, with p_1 and p_2 the *a priori* probabilities. After the LDR is applied, two new random vectors $\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1$ and $\mathbf{y}_2 = \mathbf{A}\mathbf{x}_2$, where $\mathbf{y}_1 \sim N(\mathbf{A}\mathbf{m}_1; \mathbf{A}\mathbf{S}_1\mathbf{A}^t)$ and $\mathbf{y}_2 \sim N(\mathbf{A}\mathbf{m}_2; \mathbf{A}\mathbf{S}_2\mathbf{A}^t)$ with $\mathbf{m}_i$ and $\mathbf{S}_i$ being the mean vectors and covariance matrices in the original space, respectively. The aim of LDR is to find a linear transformation matrix $\mathbf{A}$ in such a way that the new classes ($\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$) are as separable as possible. Let $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$ and $\mathbf{S}_E = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ be the within-class and between-class scatter matrices respectively. Various criteria have been proposed to measure this separability [17]. We consider the following two LDR methods:

(a) The heteroscedastic discriminant analysis (HDA) approach [17], which aims to obtain the matrix $\mathbf{A}$ that maximizes the following function, which is optimized via eigenvalue decomposition:

$$J_{HDA}(\mathbf{A}) = tr\left\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\left[\mathbf{A}\mathbf{S}_E\mathbf{A}^t \right.\right.$$
$$\left.\left. -\mathbf{A}\mathbf{S}_W^{\frac{1}{2}} \frac{p_1 \log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}}) + p_2 \log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1 p_2}\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t\right]\right\} . \quad (2)$$

(b) The Chernoff discriminant analysis (CDA) approach [17], which aims to maximize the following function, which is maximized via a gradient-based algorithm:

$$J_{CDA}(\mathbf{A}) = tr\{p_1p_2\mathbf{A}\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}$$
$$+ \log(\mathbf{A}\mathbf{S}_W\mathbf{A}^t) - p_1\log(\mathbf{A}\mathbf{S}_1\mathbf{A}^t) - p_2\log(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)\}. \quad (3)$$

In order to classify each complex, first a linear algebraic operation $\mathbf{y} = \mathbf{A}\mathbf{x}$ is applied to the n-dimensional vector, obtaining $\mathbf{y}$, a d-dimensional vector, where d is ideally much smaller than n. The linear transformation matrix $\mathbf{A}$ corresponds to the one obtained by one of the LDR methods, namely HDA or CDA. The resulting vector $\mathbf{y}$ is then passed through a Quadratic Bayesian (QB) classifier [17], which is the optimal classifier for normal distributions. For additional tests, a linear Bayesian (LB) classifiers is considered, by deriving a Bayesian classifier with a common covariance matrix: $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$.

2.3.2 Support Vector Machines

SVMs are well known machine learning techniques used for classification, regression and other tasks. The aim of SVM is to find the support vectors (most difficult vectors to be classified), and derive a linear classifier, which ideally separates the space into two regions. Classification is normally inefficient when using a linear classifier, because the data is not linearly separable, and so the use of kernels is crucial in mapping the data onto a higher dimensional space in which the classification is much more efficient. There are number of kernels that can be used in SVM models. In our model, we use polynomial, radial basis function (RBF) and sigmoid.

3. RESULTS AND DISCUSSIONS

3.1 Experimental Settings

For the LDR schemes, four different classifiers were implemented and evaluated, namely the combinations of HDA and CDA, and QB and LB classifiers. In a 10-fold cross validation setup, reductions to dimensions $d = 1, \ldots, 20$ were performed, followed by QB and LB, and the maximum average classification accuracy was recorded for each classifier. The best accuracy for each method for each dataset is bolded to indicate the classifier that performed the best for that dataset. Principal component analysis (PCA) was used as a pre-processing step to eliminate ill-conditioned matrices present in the LDR step. To select the principal components, we used different threshold values (from $\lambda_{max}10^{-2}$ to $\lambda_{max}10^{-7}$), where λ_{max} is the largest eigenvalue of the scatter matrix. The results for the threshold that achieves the highest accuracy are reported.

The SVM was also trained in a 10-fold cross validation setup with three kernels: RBF, polynomial and sigmoid. The training was carried out with the LIBSVM package [18]. A grid search was performed on the parameters gamma and C, choosing the ones that gives the maximum average accuracy for all kernels. For the polynomial kernel, the degree of the polynomial was set to 3.

The subsets of features shown in Table 2 were used for prediction. To analyze the power of desolvation energy in discriminating obligate and non-obligate complexes, NOXclass [5] was also applied to our binary-PPID dataset. The following four interface properties were analyzed, since in [5], these properties were recognized as the best ones for prediction of different types of protein protein interactions:

- Interface area
- Interface area ratio
- Amino acid composition of the interface
- Correlation between amino acid compositions of interface and protein surface

We used NACCESS [19] to calculate solvent accessible surface area (SASA). After running the classifiers in a 10-fold cross validation procedure for all subsets of features, the average accuracies were computed. The accuracy for each individual fold was computed as follows: $acc = (TP + TN)/N_f$, where TP and TN are the true positive (obligate) and true negative (non-obligate) counters respectively, and N_f is the total number of complexes in the test set of the corresponding fold.

Table 3: Prediction results for SVM and LDR classifiers on binary-PPID dataset.

	SVM			LDR			
	RBF	Polynomial	Sigmoid	Linear		Quadratic	
				HDA	CDA	HDA	CDA
PPID-AT	**77.78**	76.83	72.70	71.76	74.08	72.73	**74.55**
PPID-AA	**75.56**	71.43	71.11	71.46	**71.81**	71.46	65.07
PPID-ATD	**70.30**	67.62	67.43	68.66	68.06	**70.25**	68.97
PPID-AAD	**69.84**	67.62	66.35	67.34	66.12	**68.32**	62.80
PPID-NOXclass	**72.38**	69.84	69.52	68.89	**71.80**	67.71	68.97

3.2 Analysis of Prediction

The results of SVM and LDR classifiers with different subsets of features are depicted in Table 3. For SVM, it is clearly seen that the RBF kernel performs better that polynomial and sigmoid kernels for all subsets of features. The atom type features present in interacting chains (PPID-AT) are best classified with SVM and a RBF kernel, achieving an average accuracy of 77.78%, while accuracy for atom type features present in interacting domains (PPID-ATD) is 70.30%. Similarly, the subset of amino acid type features present in interacting chains (PPID-AA) with 75.56% classification accuracy yields more efficient predictions than using the subset of amino acid type features present in DDIs (PPID-AAD) with 69.84% classification accuracy. Furthermore, the subset based on NOXclass features (with best accuracy of 72.38%) perform worse than the best subset based on desolvation energy properties (PPID-AT) on a SVM classifier.

For LDR, the best accuracy, 74.55%, is achieved by CDA with the quadratic classifier, which is still lower than the best accuracy achieved by SVM. Note that both of them are on the PPID-AT subset. Additionally, as in SVM, subsets of atom and amino acid type features present in interacting chains perform better than those in DDIs. Also, the NOXclass subset of features (PPID-NOXclass) yields lower accuracy (71.80%) than PPID-AT, which is based on calculation of desolvation energies only, and also DDI subsets.

Generally, it can be concluded that in our binary-PPID dataset:

(a) SVM with RBF kernel performs better than LDR methods in all subsets of features

(b) Amino acid type features (for both PPID-AA and PPID-AAT subsets) yeild lower accuracies than atom type features (PPID-AT and PPID-ATD) for both LDR and SVM classifiers

(c) Although the performance of both SVM and LDR classifiers are lower for subsets of DDI features (PPID-ATD and PPID-AAD) than subsets of interacting chain features (PPID-AT and PPID-AA), they are acceptable results.

(d) Desolvation energy properties are more powerful than four properties of NOXclass (interface area, interface area ratio, amino acid composition of the interface and correlation between amino acid compositions of interface and protein surface) in predicting obligate and non-obligate complexes.

3.3 Analysis of DDIs

As discussed earlier, the total number of DDIs among 317 existing domains of our binary-PPID dataset is $100,489$. After preprocessing and removing all zero-columns, we obtain only 256 DDI pairs of which 125 are obligate and 131 are non-obligate DDIs.

The most salient feature in our binary-PPID dataset is the fact that all DDIs are presented in either obligate or non-obligate complexes and there are no DDIs in both obligate and non-obligate. This clearly implies that the type of complex could just be predicted by the DDIs present in the interactions, achieving nearly perfect prediction rate of 100%. One could design a simple classifier that contains binary features and indicates the presence or absence of the DDI in the complex, and then a simple rule that checks those binary flags. However, this would not be the case when predicting new unknown complexes (not in this dataset). That is, when using the training data to test the classifier. When cross-validation is applied, as it is done in this paper, presence of a DDI in the training set may not imply its presence or absence in the test set. In addition, it is expected, though it would not be the case, that the DDI desolvation properties are much more informative than simply binary features indicating the presence or absence of the DDI in the complex.

We performed a visual analysis on our DDIs and discovered that from 317 existing domains in our binary-PPID dataset, 135 are present only in obligate DDIs, 158 are present only in non-obligate DDIs and 21 domains are in both obligate and non-obligate DDIs. We re-ordered the domain IDs based on their types (obligate, both and non-obligate). To provide a visual insight of the distribution of the DDIs in the different complexes, a schematic view of the DDIs in the dataset is shown in Figure. 1. It is clearly seen that the most homo-domain pairs are in obligate complexes (i.e. they lie on the diagonal line ($x = y$) of the plot). Only a small part of the domain IDs are common. This also implies we can achieve a reasonable prediction only by finding the domains of each unknown complex. This is an interesting issue that deserves a lot of attention, and that we are currently investigating.

4. CONCLUSION

We have proposed an approach for prediction and analysis of obligate and non-obligate protein complexes. We have investigated various interface properties of these interactions including atom and amino acid types present in interacting chains or domains. Various features are extracted from each complex, including the desolvation energies for atom and amino acid type pairs and also NOXclass properties. The classification is performed via different LDR methods that involve homoscedastic and heteroscedastic criteria and SVM with different kernels, namely RBF, polynomial and sigmoid.

The results on our binary-PPID dataset, which is a joint and modified version of two well-known datasets, show that the SVM classifier with 77.78% accuracy achieves much better classification performance, even better than LDR schemes coupled with quadratic and linear classifiers for all subset of features. The results also demonstrate that desolvation energy is better than interface area and composition for predicting obligate and non-obligate complexes.

Furthermore, visual and numerical analysis on DDIs show that (i) most homo-domain pairs are in obligate interactions and (ii) no common DDI is present in obligate and non-obligate complexes

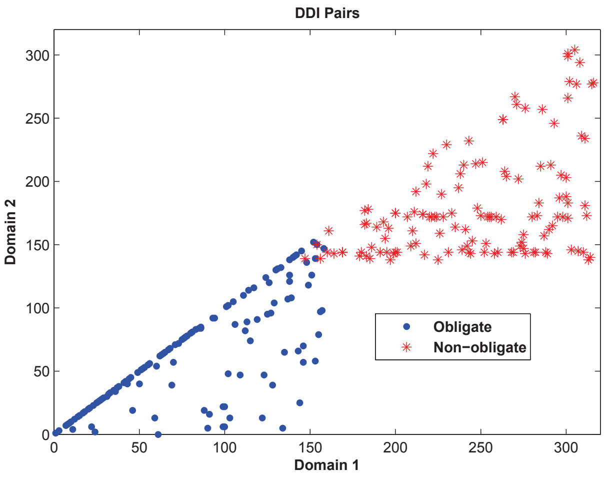

Figure 1: Schematic view of the DDI pairs in obligate and non-obligate interactions.

and all DDIs are present in either obligate or non-obligate complexes.

Our future work involves the use of other features such as residual vicinity, shape of the structure of the interface, secondary structure, planarity, physicochemical features, hydrophobicity, structure of domains and many others in our dataset, and also identifying pseudo-domains and motifs present in interacting proteins.

Acknowledgments

This research work has been partially supported by NSERC, the Natural Sciences and Research Council of Canada, grant No. RGPIN 261360, and the University of Windsor, internal Start-up and Office of Research Services equipment grants.

5. REFERENCES

[1] I. Nooren and J. Thornton, "Diversity of protein-protein interactions," *EMBO Journal*, vol. 22, no. 14, pp. 3846–3892, 2003.

[2] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc. Natl Acad. Sci, USA*, vol. 93, no. 1, pp. 13–20, 1996.

[3] L. LoConte, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites," *J Mol Biol*, vol. 285, no. 5, pp. 2177–2198, 1999.

[4] O. K. A. Zen, C. Micheletti and R. Nussinov, "Comparing interfacial dynamics in protein-protein complexes: an elastic network approach," *BMC Structural Biology*, vol. 10, no. 26, 2010, doi: 10.1186/1472-6807-10-26.

[5] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer, "Noxclass: Prediction of protein-protein interaction types," *BMC Bioinformatics*, vol. 7, no. 27, 2006, doi:10.1186/1471-2105-7-27.

[6] M. S. S. Kottha, "Classifying permanent and transient protein interactions," *German Conference on Bioinformatics*, vol. 83GI, pp. 54–63, 2006.

[7] S. De, O. Krishnadev, N. Srinivasan, and N. Rekha, "Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different." *BMC Structural Biology*, vol. 5, no. 15, 2005.

[8] J. V. Eichborn, S. Günther, and R. Preissner, "Structural features and evolution of protein-protein interactions." *Intenational Conference of Genome Informatics.*, vol. 22, pp. 1–10, 2010.

[9] S. H. Park, J. Reyes, D. Gilbert, J. W. Kim, and S. Kim, "Prediction of protein-protein interaction types using association rule based classification," *BMC Bioinformatics*, vol. 10, no. 36, 2009, doi:10.1186/1471-2105-10-36.

[10] L. Rueda, S. Banerjee, M. M. Aziz, and M. Raza, "Protein-protein interaction prediction using desolvation energies and interface properties," proceedings of the 2nd. IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2010), pp. 17–22, 2010.

[11] J. Mintseris and Z. Weng, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proc Natl Acad Sci, USA*, vol. 102, no. 31, pp. 10 930–10 935, 2005.

[12] S. G. J. V. Eichborn and R. Preissner, "Structural features and evolution of protein-protein interactions," *Genome Inform*, vol. 22, pp. 1–10, 2010.

[13] [Online]. Available: http://pfam.sanger.ac.uk/

[14] C. Camacho and C. Zhang, "FastContact: rapid estimate of contact and binding free energies," *Bioinformatics*, vol. 21, no. 10, pp. 2534–2536, 2005.

[15] C. Zhang, G. Vasmatzis, J. L.Cornette, and C. DeLisi, "Determination of atomic desolvation energies from the structures of crystallized proteins," *J. Mol. Biol.*, vol. 267, pp. 707–726, 1997.

[16] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[17] L. Rueda and M. Herrera, "Linear Dimensionality Reduction by Maximizing the Chernoff Distance in the Transformed Space," *Pattern Recognition*, vol. 41, no. 10, pp. 3138–3152, 2008.

[18] C. L. C. Chang, "Libsvm: a library for support vector machines," last date accessed: May 31, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.pdf

[19] S. Hubbard and J. Thornton, "Naccess," last date accessed: May 31, 2011. [Online]. Available: www.bioinf.manchester.ac.uk/naccess/