Biological Protein-Protein Interaction Prediction Using Binding Free Energies and Linear Dimensionality Reduction

L. Rueda[1], Carolina Garate[2], Sridip Banerjee[1], and Md. Mominul Aziz[1]

[1] School of Computer Science, University of Windsor, 401 Sunset Ave., Windsor, ON, N9B 3P4, Canada
{lrueda,banerje1,azizc}@cs.uwindsor.ca
[2] Department of Computer Science, University of Concepcion, Concepcion, Chile
cgarate@udec.cl

Abstract. An important issue in understanding and classifying protein-protein interactions (PPI) is to characterize their interfaces in order to discriminate between transient and obligate complexes. We propose a classification approach to discriminate between these two types of complexes. Our approach uses contact and binding free energies of the residues present in the interaction, which are the input features for the classifiers. A total of 282 features are extracted for each complex, and the classification is performed via recently proposed dimensionality reduction (LDR) methods, including the well-know Fisher's discriminant analysis and two heteroscedastic approaches. The results on a standard benchmark of transient and obligate protein complexes show that LDR approaches achieve a very high classification accuracy (over 78%), outperforming various support vector machines and nearest-neighbor classifiers. An additional insight on the proposed approach and experiments on different subsets of features shows that solvation energies can be used in the classification, leading to a performance comparable to using the full binding free energies of the interaction.

Keywords: protein-protein interaction, classification, binding free energy, linear dimensionality reduction.

1 Introduction

Protein-protein interaction (PPI) is involved in multiple cellular processes such as signal transduction, immune response, regulation of gene expression, and different processes where the oligomerization is a requirement to achieve a biologically active state. In this context, interactions can be attractive or repulsive, which may result in the formation of intermolecular clusters or aggregates. Although PPI depends on the protein surfaces and on the environmental conditions, many efforts have been made to understand the factors responsible for interactions between proteins at the atomic level [1,2,3]. PPI has been studied from many different perspectives and for different purposes. According to [4], prediction of protein interactions can be focused on three main goals: (i) predicting

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 383–394, 2010.

the interfaces involved in the interaction, (ii) predicting the spatial arrangement of the interacting chains or molecules, and (iii) predicting the identity of the molecules involved in the interaction. One typical case of the latter main goal is to differentiate between specific types of PPI, namely obligate versus transient interactions, i.e., interactions that can be identified by its duration. Characterizing PPI in terms of specific goals including prediction of different types can be carried out in many different ways and using many different descriptors or features [5], including solvent accessibility, residual vicinity, shape of the structure of the interface, secondary structure, planarity, conservation scores, physicochemical features, hydrophobicity electrostatic and solvation energies, just to mention a few. In this work, we focus on using energetic features.

Some of the studies in PPI consider the characterization of the geometry [6], physicochemical properties [7], the preference of residues to appear on the surface [8], and the role of hydrogen bridges, saline bridges and hydrophobic and polar interactions on the proteins surfaces [9]. Other studies include the analysis of the loss of surface accessible to solvent [10] as a result of the interaction and the analysis of the conservation of residues in the interaction surface [11]. In an upper level, amino acids composition of protein-protein interfaces have been studied to infer the composition of the residues at the interface, which is generally different from the rest of the surface. A comprehensive study was conducted by the authors of [12], who studied six types of interfaces: intra and inter domains, homo and hetero-oligomers, and obligate and transient complexes. That study concluded that the amino acid composition of these surfaces are different, as there is only 1.5% of similarity between the internal and external surfaces, and 0.2% similarity between hetero surfaces belonging to obligate homo complex and transient homo complexes. They found, on the other hand, a 16.3% similarity between homo and hetero complexes.

To study the behavior of transient and obligate interactions, in [13], a classification of these two types of interactions was proposed, where interactions are classified based on the lifetime of the complex. Obligate interactions are usually more stable, while transient interactions are less stable and, hence, more difficult to discriminate and understand, due to their short life [14]. Protomers from obligate complexes do not exist as stable structures in vivo, whereas protomers of non-obligate complexes may dissociate from each other and stay as stable and functional units. For these reasons, it is one of the prime importance of proteomics to distinguish between obligate and transient complexes. Additionally, in [15], it was proposed that interfaces in obligate complexes are inherently hydrophobic. Another work that deserves attention is that of Zhu et al. [16], in which three different types of interaction are studied, namely crystal packing, obligate and non-obligate interactions. Their study is based on using solvent accessible surface area, conservation scores, and shapes of the interfaces.

The interfaces of some transient complexes were also found to be with clusters of hydrophobic residues [17]. Moreover, they are rich in aromatic residues and arginine but depleted in other charged residues [18]. However, hydrophobicity at the interfaces of transient complexes is not as distinguishable from the remainder

of the surface as hydrophobicity at the interfaces of the obligate complexes [18]. As a result, it is difficult to make an accurate prediction of the interfaces of transient complexes using a single parameter of residue interface propensity.

In [19], a research on protein-protein interactions was conducted in which each interaction is analyzed in physical interaction, co-complex relationship and co-member of the pathway (i.e. enzymes are involved in enzyme or metabolic ways). This study attempted to determine the accuracy of predictions of interactions, applying six different classification methods, namely random forest (RF), RF-based k-NN, Bayes, decision trees, logistic regression, and support vector machines (SVM). RF was shown to be the most robust and efficient method among the six aforementioned approaches for predicting protein-protein interactions. While this study concluded that the co-complex relationship is the easiest to predict, the situation could change when larger datasets are available.

Although interfaces have been the main subject of study to predict protein-protein interactions, an accuracy of 70% has been independently achieved by several different groups [20,21,22,23]. These approaches have been carried out by analyzing a wide range of parameters, including solvation energies, amino acid composition, conservation, electrostatic energies, and hydrophobicity. In a recent work, prediction of four different PPI types has been performed, including transient Enzyme inhibitor/Non Enzyme inhibitor and permanent homo/hetero obligate complexes [24]. That work uses association rules to understand and characterize the diverse kinds of interactions, and carry out experiments on 147 pre-classified complexes (a smaller set than the one used in [25], and which is used here).

In this paper, a classification approach to predict transient and obligate protein-protein interactions is proposed. We use heteroscedastic linear discriminant analysis as the primary classification method, which is discussed in Section 2. For each protein complex, its three-dimensional structure, obtained from the Protein Data Bank (PDB) [26], is processed to extract binding free energies, namely solvation and electrostatic, producing as many as 282 features. The details of this process are discussed in Section 4. Other two classifiers, namely the k-nearest neighbor and a support vector machine, are also used for experimental comparison (briefly discussed in Section 3). Experiments on more than 400 transient and obligate complexes on two different datasets show a high accuracy in classification, more than 78% – the discussions of these experiments are in Section 5. Further analysis on the results demonstrate that solvation energies are crucial in distinguishing transient and obligate complexes, and using these features solo leads to a performance comparable to, if not better than, using the full binding free energies of the interaction.

2 Linear Dimensionality Reduction

In this section, we discuss the homoscedastic and heteroscedastic classifiers used in our approach. Linear dimensionality reduction (LDR) is a well-studied topic in the field of pattern recognition. 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 lineal transformation. The advantage of using a linear transformation is that, although the derivation of the underlying transformation may be slower, the classification is extremely fast as it performs linear-time operations to reduce to dimensions, typically, much lower than the original one.

We consider two classes, ω_1 y 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 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. Various criteria have been proposed to measure this separability [27]. We consider three LDR methods: (a) the well-know Fisher's discriminant analysis (FDA) [28,29], a recently proposed heteroscedastic discriminant analysis (HDA) approach [30], and the even more recent Chernoff discriminant analysis (CDA) approach [27] – a brief discussion of these three follows.

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. The well-known FDA criterion consists of maximizing the Mahalanobis distance between the transformed distributions by finding $\mathbf{A}$ that maximizes the following function [28]:

$$J_{FDA}(\mathbf{A}) = tr\left\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_E\mathbf{A}^t)\right\} . \tag{1}$$

The matrix $\mathbf{A}$ that maximizes (1) is obtained by finding the eigenvalue decomposition of the matrix:

$$\mathbf{S}_{FDA} = \mathbf{S}_W^{-1}\mathbf{S}_E , \tag{2}$$

and taking the d eigenvectors whose eigenvalues are the largest ones. Since $\mathbf{S}_E$ is of rank one, $\mathbf{S}_W^{-1}\mathbf{S}_E$ is also of rank one. Thus, the eigenvalue decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_E$ leads to only one non-zero eigenvalue, and hence FDA can only reduce to dimension $d = 1$.

HDA has been recently proposed as a new LDR technique for normally distributed classes [30], which takes the Chernoff distance in the original space into consideration to minimize the error rate in the transformed space. It can be seen as a generalization of FDA to consider heteroscedastic classes, and the aim is to obtain the matrix $\mathbf{A}$ that maximizes the function:

$$\begin{aligned} J_{HDA}(\mathbf{A}) = tr\Big\{&(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\left[\mathbf{A}\mathbf{S}_E\mathbf{A}^t\right. \\ &\left.-\mathbf{A}\mathbf{S}_W^{\frac{1}{2}}\frac{p_1\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}})+p_2\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1p_2}\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t\right]\Big\} \end{aligned} \tag{3}$$

where the logarithm of a matrix $\mathbf{M}$, $\log(\mathbf{M})$, is defined as:

$$\log(\mathbf{M}) \triangleq \boldsymbol{\Phi}\log(\boldsymbol{\Lambda})\boldsymbol{\Phi}^{-1} . \tag{4}$$

with $\boldsymbol{\Phi}$ and $\boldsymbol{\Lambda}$ representing the eigenvectors and eigenvalues of $\mathbf{M}$, respectively.

The solution to this criterion is given by computing the eigenvalue decomposition of:

$$\mathbf{S}_{HDA} = \mathbf{S}_W^{-1} \left[\mathbf{S}_E - \mathbf{S}_W^{\frac{1}{2}} \frac{p_1 \log(\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_1 \mathbf{S}_W^{-\frac{1}{2}}) + p_2 \log(\mathbf{S}_W^{-\frac{1}{2}} \mathbf{S}_2 \mathbf{S}_W^{-\frac{1}{2}})}{p_1 p_2} \mathbf{S}_W^{\frac{1}{2}}\right] \tag{5}$$

and choosing the d eigenvectors whose corresponding eigenvalues are the largest ones.

CDA is an LDR method that has been recently proposed, and its aim is to maximize the separability of the distributions in the transformed space measured by the Chernoff distance between the two classes. CDA assumes that the classes are normally distributed (in the original and transformed spaces), maximizing the following function [27]:

$$\begin{aligned} J_{CDA}(\mathbf{A}) = tr\{&p_1 p_2 \mathbf{A}\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} \\ &+ \log(\mathbf{A}\mathbf{S}_W\mathbf{A}^t) - p_1 \log(\mathbf{A}\mathbf{S}_1\mathbf{A}^t) - p_2 \log(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)\} \end{aligned} \tag{6}$$

where $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$, $\mathbf{S}_E = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$.

It has been shown in [27] that for any normally distributed random vectors, $\mathbf{x_1}$ and $\mathbf{x_2}$, there always exists an orthogonal matrix $\mathbf{Q}$, where $\mathbf{Q}\mathbf{Q}^t = \mathbf{I}$, such that $J_{CDA}(\mathbf{A}) = J_{CDA}(\mathbf{Q})$ for any $\mathbf{A}$ or rank d. Thus, without loss of generality, here, we assume that $\mathbf{A}$ is an orthogonal matrix. In [27], a gradient-based algorithm was proposed, which maximizes the function (6) in an iterative way. The algorithm starts with an arbitrary orthogonal matrix $\mathbf{A}^{(1)}$, and at step $k+1$, $\mathbf{A}^{(k+1)}$ is computed as follows:

$$\mathbf{A}^{(k+1)} = \mathbf{A}^{(k)} + \alpha_k \nabla J_{CDA}(\mathbf{A}^{(k)}) \tag{7}$$

where the gradient for J_{CDA} is:

$$\begin{aligned} \frac{\partial J_{CDA}}{\partial \mathbf{A}} = \nabla J_{CDA}(\mathbf{A}) = 2p_1p_2 \big[&\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} \\ &-\mathbf{S}_W\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_E\mathbf{A}^t)(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\big]^t \\ +2\big[&\mathbf{S}_W\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} - p_1\mathbf{S}_1\mathbf{A}^t(\mathbf{A}\mathbf{S}_1\mathbf{A}^t)^{-1} \\ &-p_2\mathbf{S}_2\mathbf{A}^t(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)^{-1}\big]^t \end{aligned}$$

For this gradient algorithm, a learning rate, α_k needs to be computed. In order to ensure that the gradient algorithm converges, α_k needs to be maximized. In [27], the secant method is proposed for this, and the aim is to maximize the function:

$$\phi_k(\alpha) = J_{CDA}(\mathbf{A}^{(k)} + \alpha \nabla J_{CDA}(\mathbf{A}^{(k)})) \tag{8}$$

Starting with two initial values $\alpha^{(0)}$ and $\alpha^{(1)}$, the value of $\alpha^{(j+1)}$ at time $j+1$ is iteratively found as follows:

$$\alpha^{(j+1)} = \alpha^{(j)} + \frac{\alpha^{(j)} - \alpha^{(j-1)}}{\frac{d\phi_k}{d\alpha}(\alpha^{(j)}) - \frac{d\phi_k}{d\alpha}(\alpha^{(j-1)})} \frac{d\phi_k}{d\alpha}(\alpha^{(j)}) \tag{9}$$

where

$$\frac{d\phi_k}{d\alpha}(\alpha) = [\nabla J_{CDA}(\mathbf{A}^{(k)} + \alpha \nabla J_{CDA}(\mathbf{A}^{(k)}))] \cdot \nabla J_{CDA}(\mathbf{A}^{(k)}) \tag{10}$$

The operator "$\cdot$" is the dot product between two matrices, and is computed, for any two matrices $\mathbf{C}$ and $\mathbf{D}$, as $\mathbf{C} \cdot \mathbf{D} = tr\{\mathbf{C}\ \mathbf{D}\}$. The value of $\nabla J_{CDA}(\mathbf{A}^{(k)} + \alpha \nabla J_{CDA}(\mathbf{A}^{(k)}))$ is computed by replacing $\mathbf{A}$ for $\mathbf{A}+\alpha \nabla J_{CDA}(\mathbf{A})$ in the equation (8).

Finally, with the definition of $\frac{d\phi_k}{d\alpha}(\alpha)$, Equation (9) can be solved, and the gradient algorithm continues with the next iteration. The complete algorithm can be found in [27]. One of the keys in this algorithm is the initialization of the matrix $\mathbf{A}$, and in this work, we have performed ten different initializations and then chosen the solution for $\mathbf{A}$ that gives the maximum Chernoff distance.

3 Other Classifiers

In order to compare the LDR methods with other benchmarks, a classification was performed with two other state-of-the-art classifiers: k-nearest-neighbor (k-NN) and an SVM. For the k-NN classifier, six different distance functions were implemented, namely angle, Chebychev, Euclidean, Manhattan, Minkowski and Pearson correlation. For each distance function, different values of $k = 1, ..., 20$ were evaluated, where the maximum value of $k = 20$ was taken roughly from $\sqrt{N}$ with N being the total number of complexes. The resulting accuracies were evaluated to observe the best overall performance of each distance, and hence we chose the Euclidean distance. The resulting accuracies of k-NN with the Euclidean distance, and the best value of k from 1 to 20 are reported in Section 5.

For the SVM classifier, different kernels were implemented and evaluated using the OSU-SVM toolbox in Matlab [31]. Three different types of kernels were implemented, namely polynomial, radial basis function (RBF), and sigmoid. For the polynomial kernel, polynomials of degree $p = 2, 3, ..., 8$ were considered. For the RBF, the parameters C and γ were optimized using grid search. As in k-NN, these different classifiers were evaluated and the maximum accuracy for all datasets resulted from the RBF kernel, with the parameters C and γ optimized. These results are reported on the fifth column of both Tables.

4 Protein-Protein Interaction Classification

To begin the classification process two dataset of transient and obligate complexes were obtained from previous works of [25] and [16]. Two types of complexes were classified as one of two classes: transient or obligate. Each complex is listed in the form of one or more chains for ligand and receptor respectively. The relevant data about the structure of the complex was obtained from the Protein Data Bank (PDB) [26]. When more than one chain are present on either

ligand or receptor, they are merged into a single one, producing a complex with two interacting chains, one for the ligand and another for the receptor.

Obtaining binding free energies, even for a single complex, may take a considerable amount of time. Thus, for this purpose feature extraction is performed using FastContact [32], an approach that obtains a fast estimate of the binding free energy based on a statistically determined solvation contact potential and Coulomb electrostatics with a distance-dependent dielectric constant. The interaction between two chains is estimated as the sum of the standard intermolecular Coulombic electrostatic potential ($4r$ used as the distance-dependent dielectric constant), plus the most essential features of solvation free energy that includes hydrophobic interactions. For each complex, FastContact delivers the electrostatic energy, solvation free energy, and the top 20 maximum and minimum values (along with the corresponding residue number and amino acid) for: (i) residues contributing to the binding free energy, (ii) ligand residues contributing to the solvation free energy, (iii) ligand residues contributing to the electrostatic energy, (iv) receptor residues contributing to the solvation free energy, (v) receptor residues contributing to the electrostatic energy, (vi) receptor-ligand residue solvation constants, and (vii) receptor-ligand residue electrostatic constants.

For each complex, all energy values (minimum and maximum) were obtained as indicated in (i)-(vii). Thus, all these values (with the residue numbers) and the total solvation and electrostatic energy values compose a total of 282 features.

Due to the large number of features present in most datasets, compared to the number of samples, problems of dimensionality arise. More precisely, ill-conditioned matrices would be present when applying LDR methods, and hence principal component analysis is applied to each dataset by removing all components which are less than 10^{-5} times the largest eigenvalue of the within-class scatter of the dataset.

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 either of the LDR methods discussed in Section 2. The resulting vector $\mathbf{y}$ is then passed through a quadratic Bayesian (QB) classifier [28], which is the optimal classifier for normal distributions.

5 Experimental Results

To create the datasets for classification, two pre-classified datasets of protein complexes ware obtained from the studies of [25] and [16]. The first set of proteins, Mintseris et al. dataset, contains complexes of two classes: 209 transient complexes and 115 obligate complexes. The second dataset, Zhu et al. dataset, contains 62 transient complexes and 75 obligate complexes as two different classes for classification. The main datasets were created by retrieving each complex from PDB, and then obtaining the 282 features by invoking FastContact, as discussed in Section 4.

To study the effects of the different types of energies and ligand/receptor, we created a total of 13 different subsets of features for each dataset including:

Table 1. Results of classification accuracy for the 13 PPI subsets extracted from Mintseris et al. dataset [25], using different LDR methods and a comparison with k-NN and SVM

				QB					
Subset	n	k-NN	SVM	FDA	d^*	HDA	d^*	CDA	d^*
All Energetic	282	76.38	77.30	70.38	1	77.50	4	76.87	9
Binding Free Energy	40	69.94	72.09	71.86	1	75.20	6	73.33	5
Ligand Energy	80	72.70	74.54	66.58	1	76.42	8	76.44	6
Ligand Solvation	40	77.91	75.46	69.65	1	75.81	2	74.86	8
Ligand Electrostatic	40	69.33	70.86	72.09	1	72.14	7	71.84	4
Receptor Energies	80	74.54	74.23	67.17	1	76.42	6	76.74	11
Receptor Solvation	40	75.46	75.46	68.73	1	75.50	1	74.60	3
Receptor Electrostatic	40	72.09	70.55	68.47	1	69.71	3	70.31	12
Ligand-Receptor Energies	80	71.78	71.78	67.91	1	75.94	7	75.32	7
Ligand-Receptor Solv.	40	72.09	70.55	65.64	1	71.84	9	72.76	4
Ligand-Receptor Elect.	40	73.62	74.85	72.78	1	75.48	20	75.50	13
Solvation	120	78.53	76.07	65.72	1	76.70	14	76.41	11
Electrostatic	120	71.78	71.17	65.72	1	76.70	14	76.41	11

all 282 values, binding free energies, ligand/receptor solvation/electrostatic energies, ligand-receptor solvation and electrostatic energies, and solvation and electrostatic energies. The 13 datasets along with a short description in column one are listed in Tables 1 and 2. The second column lists the number of features of each dataset. As discussed earlier, PCA was applied to some datasets to avoid ill-conditioned matrices.

To study the performance of the classifiers, a 10-fold cross validation procedure was carried out, and then the average accuracy was computed, where accuracy for each individual fold was computed as follows: $acc = (TP + TN)/N_f$, where TN and TP are the true positive (obligate) and true negative (transient) counters, and N_f is the total number of complexes in the test set of the corresponding fold.

For the LDR schemes, three different classifiers were implemented and evaluated, namely the combinations of three LDR criteria discussed in Section 2, FDA, HDA and CDA, combined with a quadratic Bayesian (QB), implemented as discussed in Section 4. Note that we have also tested the classification with a linear Bayesian classifier, which yielded much lower classification accuracies than the QB. Then, only the results for QB are reported. For each of these classifiers reduction to dimensions $d = 1, ..., 20$ were performed, followed by QB. The dimensions that resulted in the best average accuracy for the 10-fold cross validation for each classifier are listed in the tables in the subsequent columns. Each column reports the highest average accuracy among all possible reduced dimensions, as well as the dimension in which the best accuracy is obtained, namely d^*. Since the classification problem is two-class, FDA always leads to

Table 2. Results of classification accuracy for the 13 PPI subsets extracted from Zhu et al. dataset [16], using different LDR methods and a comparison with k-NN and SVM

				QB					
Subset	n	k-NN	SVM	FDA	d^*	HDA	d^*	CDA	d^*
All Energetic	282	67.15	65.69	58.62	1	65.08	1	<u>69.12</u>	15
Binding Free Energy	40	64.96	59.85	55.59	1	<u>65.74</u>	9	63.23	7
Ligand Energy	80	68.61	69.34	60.05	1	70.60	15	<u>72.08</u>	5
Ligand Solvation	40	<u>70.80</u>	<u>70.80</u>	62.35	1	70.64	18	69.26	6
Ligand Electrostatic	40	<u>64.23</u>	62.77	49.55	1	60.51	4	59.58	3
Receptor Energies	80	65.69	67.88	52.54	1	68.86	13	<u>72.79</u>	19
Receptor Solvation	40	<u>76.64</u>	64.96	66.05	1	74.03	11	73.97	13
Receptor Electrostatic	40	61.31	64.96	54.95	1	65.48	6	<u>67.48</u>	5
Ligand-Receptor Energies	80	67.15	67.88	67.22	1	69.16	17	<u>70.97</u>	5
Ligand-Receptor Solv.	40	70.8	70.07	70.71	1	72.08	10	<u>72.18</u>	18
Ligand-Receptor Elect.	40	61.31	55.47	60.27	1	66.72	16	<u>67.54</u>	17
Solvation	120	73.72	71.53	51.41	1	65.33	6	<u>75.41</u>	7
Electrostatic	120	69.34	<u>72.99</u>	53.61	1	63.53	14	64.10	1

reducing to dimension one. The best accuracy for each method for each dataset is underlined to indicate the classifier that performed best of all for that dataset.

For the Mintseris et al. dataset (Table 1), it is clearly observable that the best performance was achieved by LDR methods combined with the QB classifier. Of these, the LDR criterion that achieves the best performance is HDA in as many as 6 out of 13 cases. Also, the classification of all LDR methods achieves the best performance in most of the cases, 10 out of 13 cases. This demonstrates that LDR methods perform better than k-NN and SVM. On the other hand, k-NN achieves better performance in more cases than the SVM, even though the results of these two are comparable in most of the cases. Regarding individual subsets, we observe that the best overall classification performance, 78.53%, was achieved by k-NN on Solvation energies. A comparison with other subsets, such as All Energetic, suggests that using a subset of features, such as Solvation energies, achieves an even better classification performance. In terms of energetic values, solvation leads to better performance than electrostatic values. This suggests that solvation is more important in classifying transient and obligate complexes. Additionally, using Solvation energies from the ligand only (just 40 features) leads to a classification accuracy of 77.91%, achieved by k-NN, which is no less than 1% below the best overall accuracy, obtained from all solvation values.

For the Zhu et al. dataset (Table 2), we observe that the best overall performance is delivered, again, using Solvation energies only, leading to an accuracy of 75.41%, which is achieved by CDA. Moreover, using the Solvation energies of the receptor only leads to an accuracy of 74.03%, slightly below that of using all Solvation energies. For this dataset, CDA is the best performer, yielding the

highest accuracy in 8 out of 13 subsets. Again, as in Mintseris et al. dataset, the LDR methods perform much better than k-NN and SVM, and the Solvation energies by themselves can differentiate between the two types of complexes.

A final analysis of the results is done on the power of dimensionality reduction of the schemes. We observe that the best overall classification accuracy was obtained by HDA and CDA, while reducing from dimensions 120 to 14 and 11. In Zhu et al. dataset, the best classification accuracy achieved by CDA is 75.41%, while reducing from dimension 120 to 7. This thus implies not only a gain in classification accuracy but also in terms of classification speed. Similar results can be observed in the other cases, and hence demonstrating the power and simplicity of LDR schemes in this classification problem. To conclude, we emphasize that using a subset of features tends to be more productive than using all features, and hence demonstrating that the approach of considering different subsets of features leads to feature selection methods, even though more sophisticated approaches for feature selection could be used [33], a problem that is currently being investigated.

6 Conclusion

We have proposed a classification approach for transient and obligate protein-protein complexes. We have used linear dimensionality reduction (LDR) that involve homoscedastic and heteroscedastic criteria coupled with a quadratic Bayesian classifier. The results on two datasets of pre-classified complexes show that the LDR schemes coupled with QB achieves the best overall classification performance, even better than k-NN and an SVM with an RBF kernel. Comprehensive tests have been carried out in as many as 13 subsets of different features and for each dataset, showing that the best classification performance is achieved by using a smaller subset of features, solvation energies for the ligand or receptor. The results suggest that the proposed approach also performs feature selection, a problem that is currently being investigated. Other interesting problems that deserve investigation are the use of this approach in different protein-protein interaction classification problems, including intra and inter domains, homo and hetero-oligomers, and the use of other features, such as solvent accessibility, residual vicinity, shape of the structure of the interface, secondary structure, planarity, conservation scores, physicochemical features, hydrophobicity and others.

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 grant.

References

1. Janin, J.: Kinetics and thermodynamics of protein-protein interactions from a structural perpective. In: Protein-Protein Recognition, p. 344. Oxford University Press, Oxford (2000)
2. Jones, S., Thornton, J.M.: Analysis and classification of protein-protein interactions from a structural perspective. In: Protein-Protein Recognition. Oxford University Press, Oxford (2000)
3. Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M., Sali, A.: A structural perspective on protein-protein interactions. Curr. Opin. Struct. Biol. 14(3), 313–324 (2004)
4. Kurareva, I., Abagyan, R.: Predicting molecular interactions in structural proteomics. In: Nussinov, R., Shreiber, G. (eds.) Computational Protein-Protein Interactions, pp. 185–209. CRC Press, Boca Raton (2009)
5. Ofran, Y.: Prediction of protein interaction sites. In: Nussinov, R., Shreiber, G. (eds.) Computational Protein-Protein Interactions, pp. 167–184. CRC Press, Boca Raton (2009)
6. Lawrence, M.C., Colman, P.M.: Shape complementarity at protein/protein interfaces. J. Mol. Biol. 234(4), 946–950 (1993)
7. Chakrabarti, P., Janin, J.: Dissecting protein-protein recognition sites. Proteins 47(3), 334–343 (2002)
8. Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A., Kornberg, R.D.: Structural basis of transcription: an RNA polymerase II elongation complex at 3. 3 A resolution. Science 292(5523), 1876–1882 (2001)
9. Xu, D., Tsai, C., Nussinov, R.: Hydrogen bonds and salt bridges accross protein-protein interfaces. Protein Eng. 10(9), 999–1012 (1997)
10. Shanahan, H., Thornton, J.: Amino acid architecture and the distribution of polar atoms on the surfaces of proteins. Biopolymers 78(6), 318–328 (2005)
11. Ma, B., Elkayam, T., Wolfson, H.: Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc. Natl. Acad. Sci., USA 100(10), 5772–5777 (2003)
12. Ofran, Y., Rost, B.: Analysing six types of protein-protein interfaces. J. Mol. Biol. 325(2), 377–387 (2003)
13. Nooren, I., Thornton, J.: Diversity of protein-protein interactions. EMBO Journal 22(14), 3846–3892 (2003)
14. Jones, S., Thornton, J.M.: Principles of protein-protein interactions. Proc. Natl Acad. Sci, USA 93(1), 13–20 (1996)
15. Glaser, F., Steinberg, D.M., Vakser, I.A., Ben-Tal, N.: Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 43(2), 89–102 (2001)
16. Zhu, H., Domingues, F., Sommer, I., Lengauer, T.: Noxclass: Prediction of protein-protein interaction types. BMC Bioinformatics 7(27) (2006) doi:10.1186/1471–2105–7–27
17. Young, J.: A role for surface hydrophobicity in protein protein recognition. Protein Sci. 3, 717–729 (1994)
18. LoConte, L., Chothia, C., Janin, J.: The atomic structure of protein-protein recognition sites. J. Mol. Biol. 285(5), 2177–2198 (1999)
19. Qi, Y., Bar-Joseph, Z., Klein-Seetharaman, J.: Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins 63(3), 490–500 (2006)

20. Bordner, A.J., Abagyan, R.: Statistical analysis and prediction of protein-protein interfaces. Proteins 60(3), 353–366 (2005)
21. Caffrey, H.J., Somaroo, S.: Are protein protein interfaces more conserved in sequence than the rest of the protein surface? Protein Science 13, 190–202 (2004)
22. Neuvirth, S., Raz, R.: ProMate. a structure based prediction program to identify the location of protein protein binding sites. J. Mol. Biol. 338, 181–199 (2004)
23. Zhou, H., Shan, Y.: Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44(3), 336–343 (2001)
24. Park, S.H., Reyes, J., Gilbert, D., Kim, J.W., Kim, S.: Prediction of protein-protein interaction types using association rule based classification. BMC Bioinformatics 10(36) (2009) doi:10.1186/1471-2105-10-36
25. Mintseris, J., Weng, Z.: Structure, function, and evolution of transient and obligate protein-protein interactions. Proc. Natl. Acad. Sci., USA 102(31), 10930–10935 (2005)
26. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P.: The Protein Data Bank. Nucleic Acids Research 28, 235–242 (2000)
27. Rueda, L., Herrera, M.: Linear Dimensionality Reduction by Maximizing the Chernoff Distance in the Transformed Space. Pattern Recognition 41(10), 3138–3152 (2008)
28. Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd edn. John Wiley and Sons, Inc., New York (2000)
29. Fisher, R.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7, 179–188 (1936)
30. Loog, M., Duin, P.: Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(6), 732–739 (2004)
31. Ivanciuc, O.: Applications of Support Vector Machines in Chemistry. In: Reviews in Computational Chemistry, pp. 291–400. Wiley, Chichester (2007)
32. Camacho, C., Zhang, C.: FastContact: rapid estimate of contact and binding free energies. Bioinformatics 21(10), 2534–2536 (2005)
33. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 3rd edn. Elsevier Academic Press, Amsterdam (2006)