A Model to Predict and Analyze Protein-protein Interaction Types Using Electrostatic Energies

Gokul Vasudev and Luis Rueda
School of Computer Science, University of Windsor
401 Sunset Avenue, Windsor, ON N9B 3P4, Canada
E-mail:{vasudevg,lrueda}@uwindsor.ca

***Abstract*—Identification and analysis of types of protein-protein interactions (PPI) is an important problem in molecular biology because of their key role in many biological processes in living cells. We propose a model to predict and analyze protein interaction types using electrostatic energies as properties to distinguish between obligate and non-obligate interactions. Our prediction approach uses electrostatic energies for pairs of atoms and amino acids present in interfaces where the interaction occurs. Our results confirm that electrostatic energy is an important property to predict obligate and non obligate protein interaction types achieving accuracy of over 96% on two well known datasets. The classifiers used are support vector machines and linear dimensionality reduction.**

***Index Terms*—protein-protein interaction; electrostatic energy; complex type prediction**

I. Introduction

Proteins control almost all biological systems in a cell including nutrient uptake, gene expression, cell growth, proliferation, inter-cellular communication, among others. Prediction of interaction types in protein-protein interactions (PPI) and analyzing relevant properties for prediction have been studied from various perspectives. Proteins bind to each other through a combination of hydrophobic bonding, van der Waals forces and salt bridges at specific binding domains on each protein. The strength of the binding depends on the size of the binding domains. These domains can be large surfaces, small binding clefts, a few peptides, or several amino acids.

Prediction of PPI has gained much interest in recent years with over 20 different proposed methods [1]. Characterizing the properties of protein interaction types can be done by studying their structural information. Thus, structure-based prediction methods including computational approaches, homology modelling, threading-based methods and protein-protein docking are more accurate than those which do not employ structural data [1]. These studies have been carried out mostly by relying on biological knowledge about the atoms or molecules, which normally, are selected manually by observing groups of complexes or on prediction results. Another important aspect in studying PPI is to predict 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) [2], stability of the interaction (non-obligate vs. obligate) [3], among others; we focus on the latter problem.

Obligate interactions are usually considered as permanent, while non-obligate interactions can be either permanent or transient [4]. 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 [5]. The study of [6] suggested that mobility differences of amino acids are more significant for obligate and large interface complexes than for transient and medium-sized ones.

In this context, the features are the observed properties of each complex that are used for prediction. Physiochemical properties of proteins are very powerful and the use of them in prediction has been extensively reported in literature. Interacting regions can be characterized by diverse sets of physiochemical properties [1] and other sequence features. In addition, other properties have been used for PPI prediction, such as analysis of solvent accessibility [3], geometry, hydrophobicity, sequence-based features and desolvation energy [7]. Based on interface properties such as interface area and ratio of area [3], Zhu *et al.* predicted biological and crystal packing interactions using a support vector machine (SVM). A model that uses solvent accessible surface area and other interface properties for prediction of types (obligate and non-obligate) was reported in [8]. A very recent work presented in [7] shows the use of desolvation energies to predict obligate and non-obligate complexes using SVM and linear dimensionality reduction (LDR).

In this study, we use electrostatic energies as properties to predict obligate and non obligate PPIs. Non-covalent interactions are very common between macromolecules (including proteins). There are three types of non-covalent interactions [9]: (i) Electrostatic interactions - they occur between electrically charged atoms having both positive and negative interactions, (ii) Vander Waal interactions - they occur between any pair of charged atoms that are close to each other, (iii) non-polar interactions - these are attractive interactions occurring between atoms that do not have any charge. We focus on electrostatic energies.

In this paper, we propose a model that uses electrostatic energies as properties for both atom and amino acid pairs present in the interface to predict obligate and non-obligate interaction types. We use LDR and SVM as the classifiers to predict these types. Our prediction results for well-known datasets, namely the Zhu [3] and Mintseris [2] datasets shows

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

an impressive accuracy in prediction of more than 96%, which implies an increase of at least 5% from previous approaches.

II. Materials and Methods

The method that we propose for prediction of types of PPIs uses electrostatic energies as properties, and classification is performed using LDR and SVM. We call this method prediction of protein interaction types using electrostatic energies (PPIEE).

A. Datasets used

Two pre-classified datasets of protein complexes were obtained from the studies of Zhu *et al.* [3], referred to as the ZH dataset and Mintseris *et al.* [2] referred to as the MW dataset. The ZH dataset contains 75 obligate and 62 non-obligate complexes and the MW dataset contains 209 non-obligate and 115 obligate complexes. In this study, some complexes were excluded from the ZH dataset, namely 1cc0 A:E, 1qbk B:C, 1b8a A:B, 1cli A:B, 1qav A:B, 1bkd R:S and 1nse A:B, since APBS could not compute their electrostatic energies for all the atoms in the interface. Similarly, for the MW dataset, 24 complexes were also left out: 1b7y A:B, 1be3 CDEGK:A, 1jb0 AB:C, 1jb0 AB:D, 1jb0 AB:E, 1jro A:BD, 1jv2 A:B, 1k28 A:D, 1kqf A:B, 1ldj A:B, 1m2v A:B, 1mjg AB:M, 1nbw AC:B, 1prc C:HLM, 1bgx HL:T, 1de4 CF:A, 1ezv E:XY, 1is8 ABEJCIDHGF:KLOMN, 1m2o AC:B, 1o94 AB:CD, 1qfu AB:HL, 2hmi AB:CD, 4cpa I:0 and 2q33 A:B.

B. Electrostatic Energies

Electrostatic interactions are important in understanding intermolecular interactions, since they are long ranged interactions and because of their influence in charged molecules [10]. This motivated us to focus on electrostatic energies and hence use them as properties for predicting interaction types. In proteins, these interactions can occur between charged atoms belonging to different molecules, between charged atoms on the protein surface, charges in the environment, and also, between charged atoms of the proteins with those in the ligand. Also, hydrogen bonds are very common in proteins and play a significant role such as imparting specificity to proteins.

In order to compute electrostatic energies we use software packages PDB2PQR [11] and APBS [12]. For each complex in our datasets, structural data from the Protein Data Bank (PDB) [13] were collected. PDB2PQR is a tool that automates common tasks for preparing structures for electrostatic calculations. Its purpose is to add missing heavy atoms, place missing hydrogen atoms and assign charge and radii to PDB files. The output of this package is a PQR structure file, which is the input for APBS. There is an option to customize the parameters of PDB2PQR according to users needs. The parameters which we used for our experiment are as follows:

1) Forcefield: We use the AMBER forcefield for our experiments.
2) APBS input: We specify "apbs-input" as one of the parameters so that new files are created (with .in extension) which act as input for APBS.
3) Chain name: We also specify "- -chain" so that the chain name appears in the PQR file formats.

APBS is a software package used for calculating electrostatic energies for interactions between solutes in salty and aqueous media. It solves the Poisson-Boltzmann equation numerically and evaluates electrostatic calculations ranging from tens to millions of atoms. We ran APBS for the interacting chains for all the complexes present in the ZH and MW datasets. For this, the parameters were set as follows:

1) Calcenergy: We change this parameter to "Calcenergy comps" to calculate and return the total electrostatic energy for the entire molecule as well as electrostatic energy components for each atom.
2) cglen: It specifies the length of course grid for multi-grid calculations.

To compute the features for classification, we consider the interaction between the i^{th} atom of the first interacting chain and the j^{th} atom of the second interacting chain. Then, we calculate distance between all possible atom pairs. If an atom pair is separated by a distance less than or equal to 10 Å, then that pair is considered to be in the interface of that complex [9]. In most studies, the inter-atom distances are no more than 7 Å. However, due to the long ranged nature for electrostatic interactions 10 Å or even more, is a suitable distance for this case; even atoms that are under the surface of proteins pose electrostatic forces towards the stability of the protein complex.

Figure 1 depicts the quaternary structure of an obligate complex, PDB ID 1b8j, along with its interacting chains A and B (represented in different colors: red and blue). Atoms that are under a certain threshold distance act as interface atoms and are represented in yellow and purple colors, respectively. Here, we consider a pair to be in the interface if the atoms are 10Å or less apart from each other. Since electrostatic interactions are long ranged, we observe a large interface area.

Two interacting chains were extracted from each PDB file. We consider 18 different atom types as in [14]. Thus, for each protein complex a feature vector with ${}^{18}_{2}C$+18 = 171 values were obtained, where each feature contains the cumulative sum of electrostatic energies for all pairs of atoms of the same type. Since the order of interacting atom pairs is not important, the final length of the feature vector for each complex is 171, which corresponds to the number of unique pairs. Since APBS outputs electrostatic energies for both atoms in the pair, we compute the average of the electrostatic energies and then use these values in the cumulative sum. We found that these averages are a good representation for the affinity of the proteins and hence for the stability of the complex. Average values close to zero mean high affinity, while values far away from zero mean low affinity. This implies more powerful prediction of obligate and non-obligate interactions. We also consider pairs of amino acids, and for this, we

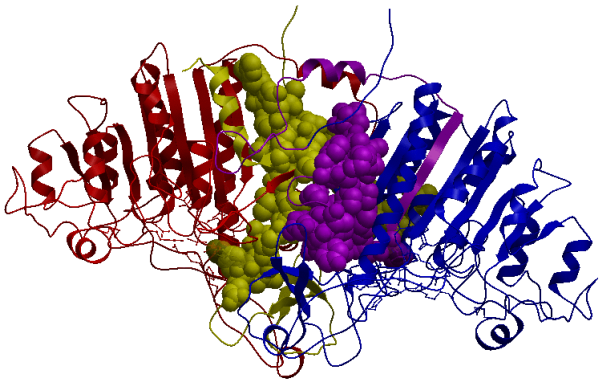

Fig. 1. Quaternary structure of complex 1b8j, visualized with ICM browser showing interface atoms for two interacting chains A and B.

compute $\binom{20}{2}C$+20 = 210 electrostatic energy values for all pairs of atoms, by accumulating the values of the corresponding atoms for each pair of amino acids.

By using electrostatic energies for different types of features for the ZH and MW datasets, four subsets of features for prediction and evaluation were generated. These are listed in Table I. The conventions used in this table are ZH-AT which stands for the ZH dataset, atom type, while ZH-AA stands for the ZH dataset, amino acid type. Similarly, MW-AT stands for the MW dataset, atom type, while MW-AA stands for the MW dataset, amino acid type.

TABLE I
DESCRIPTION OF DATASETS USED IN THIS STUDY.

Authors	Reference	Atom type	Amino acid type
Mintseris *et al.*	[2]	MW-AT	MW-AA
Zhu *et al.*	[3]	ZH-AT	ZH-AA

III. THE PREDICTION METHODS

One of the techniques used for prediction of protein interaction types is LDR. The basic idea of LDR is to represent an object of dimension n onto a lower-dimensional vector of dimension d, achieving this by performing a linear transformation. Each class, obligate or non-obligate is represented by 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 as *a priori* probabilities. 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. We consider three LDR methods as follows [15]:

- Fisher's discriminant analysis (FDA)
- Heteroscedastic discriminant analysis (HDA)
- Chernoff discriminant analysis (CDA)

The resulting vectors $\mathbf{y}_i$ are then input to a quadratic Bayesian (QB) classifier and a linear Bayesian (LB) classifier which is obtained by deriving a Bayesian classifier with a common covariance matrix. More details about the LDR methods and classification can be found in [15].

We also use SVM, which takes the set of input vectors and predicts the possible classes of output based on support vectors [16]. The kernel trick allows optimization in a much higher dimensional space that hopefully makes class separation easier. For all datasets, ZH-AT, ZH-AA, MW-AA and MW-AT, we use linear and radial basis function (RBF) kernels for the SVM. We have optimized the values for C and Gamma on a grid search in order to obtain better accuracies for each dataset.

IV. RESULTS AND DISCUSSIONS

For the LDR schemes, six different classifiers were implemented and evaluated, namely the combinations of FDA, HDA and CDA with QB and LB classifiers. In a 10-fold cross validation process, reductions to dimensions $d = 1, 2, \ldots, 20$ were performed, followed by QB and LB. The maximum average classification accuracy was taken into account for each classifier, which is the one that is reported for each dataset, accuracy was computed as follows: $acc = (TP + TN)/N$, where TP and TN are the total numbers of true positive (obligate) and true negative (non-obligate) counters over the 10 folds, respectively, and N is the total number of complexes in the dataset.

After running PPIEE on the ZH-AT dataset, we observe that the accuracy of SVM is 96.18%. As mentioned earlier, the RBF kernel was also applied with optimized values of C and Gamma obtaining a better accuracy of 97.17%. Considering the ZH dataset for amino acid type using LDR, accuracy is 96.06%. Using SVM and after optimizing the values of C and Gamma, increases accuracy to 96.18%. Similarly, the RBF kernel was also applied to MW datasets with the optimized values of C and Gamma. For the MW-AT dataset, we obtained a better accuracy of 96.74% while for MW-AA type datasets, we obtained accuracy of 95.29%.

In Table II, we also compare our results with other prediction models using different properties. We compare our approach with prediction of obligate and non-obligate using desolvation energies as properties [7]. Their work was based on the ZH and MW datasets. The best accuracy they obtained was 83.21%. In our work, the best accuracy obtained is 97.17% using the same dataset with electrostatic energies rather that desolvation energies. For MW datasets, their best accuracy is 78.83%, while using electrostatic energies, our best accuracy is 97.36%.

TABLE II
COMPARISON (IN TERMS OF ACCURACY) OF OUR APPROACH WITH DESOLVATION ENERGY AS PROPERTIES FOR ZH AND MW DATASETS.

		Properties (energy)	
Dataset type	Features	Desolvation	Electrostatic
ZH-AA	210	78.39%	96.18%
ZH-AT	171	83.21%	97.17%
MW-AA	210	78.83%	95.38%
MW-AT	171	78.53%	97.36%

Table III, compares PPIEE with different approaches already reported in literature. We compare our method with the one proposed by Zhu *et al.* [3] namely NOXClass, which uses four or six interface properties including interface area, interface area ratio, conservation score and gap volume index. Rueda *et al.* [8] extended these properties by considering each amino acid individually as a property, for interface area and frequency and hence adding 40 more features, i.e., a total of 46 features. Note that we do not use these features (we rather use electrostatic energies) but we include them for comparison.

As shown in the table, Zhu *et al.* predicted obligate and non-obligate interactions using interface properties and obtained an accuracy of 75.2%. Also, Rueda *et al.* predicted obligate and non-obligate interactions using interface properties with solvent accessible surface area and obtained an accuracy of 81.83%. With our approach, we predict obligate and non-obligate interactions using 210 features for electrostatic energies as properties and obtained an accuracy of 96.18%.

We also compare our results with the approach of [17] on the BNCP-CS dataset, which consists of 75 obligate interactions and 62 non obligate interactions. In their approach, they calculate solvent accessible surface area and apply SVM to predict the type of interaction. In their experiments they obtained an accuracy of 92.2%, while in our experiments using electrostatic energies as properties, the best accuracy obtained by PPIEE is **97.17%**, an increase of about 5% with respect to [17].

TABLE III
PREDICTION RESULTS AND COMPARISON WITH OTHER APPROACHES AND PROPERTIES ON THE ZH-AA DATASET.

No. of Features	Classifier	Accuracy	Properties used	Method
6	SVM	75.2%	Interface	Zhu *et al.* [3]
26	LDR	78.27%	Interface	Rueda *et al.* [8]
46	LDR	81.83%	Solvent accessible area and interface area	Rueda *et al.* [8]
210	SVM	92.20%	Solvent accessible area	Liu *et al.* [17]
210	SVM	96.18%	Electrostatic energies	PPIEE

Also, in Table IV, we compare our results for the MW-AA dataset. The comparison with Zhu *et al.* [3] and Rueda *et al.* [8] is made on 4, 24 and 44 features (that is, excluding conservation scores and gap volume index) since these are the features that yield the highest accuracy. As shown in the table, Rueda *et al.* predicted obligate and non-obligate interactions for different numbers of features using interface properties and obtained an accuracy of 77.54%. Also, Rueda *et al.* predicted interactions using interface properties combined with solvent accessible surface area and obtained an accuracy of 77.25%. With our approach, we predict obligate and non-obligate interactions using 210 features for electrostatic energies as properties and obtained an accuracy of 95.38%, implying an increase of more than 15% with respect to [8].

TABLE IV
PREDICTION RESULTS AND COMPARISON WITH OTHER APPROACHES AND PROPERTIES ON THE MW-AA DATASET.

No. of Features	Classifier	Accuracy	Properties used	Method
4	LDR	77.96%	Interface	Rueda *et al.* [8]
24	LDR	77.54%	Interface	Rueda *et al.* [8]
44	LDR	77.25%	Solvent accessible area and interface area	Rueda *et al.* [8]
210	SVM	95.38%	Electrostatic energies	PPIEE

Figure 2 shows an obligate complex along with the electrostatic potential for three different cases: Figure 2(a) shows one subunit (Chain A) of PDB ID 2min, Figure 2(b) depicts another subunit (Chain B) and Figure 2(c) shows Chains A and B combined. To visualize the effect of electrostatic energies for prediction, we show these proteins plotted over solvent accessible surface area, generated with the help of Jmol embedded in APBS. Observing Figure 2(a) carefully, the highlighted yellow portion has positive electrostatic potential (shown in blue), while in Figure 2(b), the highlighted yellow portion has negative electrostatic potential (shown in red). The interaction between the two chains takes place at these regions as shown in Figure 2(c). The positive and negative potentials on the corresponding areas of the interface of A and B yield very high affinity, and hence a favourable scenario for the obligate complex. This is the main feature that we exploit to predict the stability of the protein complex and it is corroborated in our experimental results.

V. CONCLUSION AND DISCUSSIONS

Our newly proposed model PPIEE works very well for distinguishing protein interaction types. Our prediction approach uses electrostatic energies as properties for pairs of atoms or amino acids present in the interfaces of such complexes. The classification is performed via different LDR methods and also SVM.

We observe that electrostatic energies turn out to be the best ones for prediction of interaction types on the basis of our experimental results. The reason for why electrostatic energies yield better prediction results is due to the fact that they are long ranged interactions which may go up to a 10 Å or more. As a result, it covers a broader (and deeper) area in the interface giving excellent results in classification. Also, they have more influence in polar and charged molecules. Thus, among various components of molecular interactions, electrostatic energies play a special role. The proposed features then exploit the high affinity of proteins to interact with each other (in terms of negative and positive potentials).

In the future, we plan to investigate domains and motifs

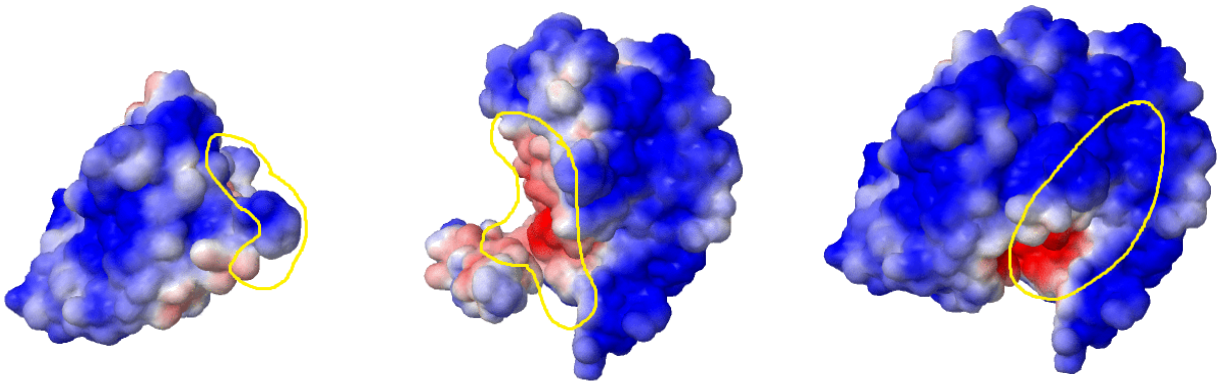

Fig. 2. Electrostatic potential of an obligate complex, PDB ID 2min, plotted over solvent accessible surface area before and after the interaction takes place. The plots were generated by Jmol embedded in APBS.

present in the interface in order to achieve a better insight on proteins, their interactions, and function. Also we intend to apply feature selection algorithms on the available datasets to obtain relevant pairs of atom types and amino acid types that are biologically meaningful.

REFERENCES

[1] S. Park, J. Reyes, D. Gilbert, J. Kim, and S. Kim, "Prediction of protein-protein interaction types using association rule based classification," *BMC Bioinformatics*, vol. 10, no. 1, p. 36, 2009.

[2] 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.

[3] 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.

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

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

[6] 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.

[7] M. M. Aziz, M. Maleki, L. Rueda, M. Raza, and S. Banerjee, "Prediction of biological protein-protein interactions using atom-type and amino acid properties," *Proteomics*, vol. 11, pp. 17–22, 2011.

[8] L. Rueda, S. Banerjee, M. Aziz, and M. Raza, "Protein-protein interaction prediction using desolvation energies and interface properties," in *Bioinformatics and Biomedicine (BIBM)*, 2010, pp. 17 –22.

[9] A. Kessel and N. Ben-Tal, *Introduction to Proteins: Structure, Function, and Motion*. CRC Press, 2010.

[10] N. Baker, "Continuum models for biomolecular solvation." 2008, pacific Northwest National Laboratory.

[11] T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, and N. A. Baker, "Pdb2pqr: expanding and upgrading automated preparation of biomolecular structures for molecular simulations," *Nucleic Acids Research*, vol. 35, no. suppl 2, pp. W522–W525, 2007.

[12] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Electrostatics of nanosystems: Application to microtubules and the ribosome," *Proceedings of the National Academy of Sciences*, vol. 98, no. 18, pp. 10 037–10 041, 2001.

[13] 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.

[14] 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.

[15] 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.

[16] O. Ivanciuc, "Applications of support vector machines in chemistry," *Reviews in computational chemistry*, pp. 291–400, 2007.

[17] Q. Liu and J. Li, "Propensity vectors of low-asa residue pairs in the distinction of protein interactions," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 3, pp. 589–602, 2010.