Classification of protein binding ligands using their structural dispersion

G.A.I.C. Premarathna[1] and Leif Ellingson[2]

[1]Department of Mathematics and Statistics, Minnesota State University, Mankato, Minnesota, United States of America
[2]Department of Mathematics and Statistics, Texas Tech University, Lubbock, Texas, United States of America

Abstract

It is known that a protein's biological function is in some way related to its physical structure. Many researchers have studied this relationship both for the entire backbone structures of proteins as well as their binding sites, which are where binding activity occurs. However, despite this research, it remains an open challenge to predict a protein's function from its structure. The main purpose of this research is to gain a better understanding of how structure relates to binding activity and to classify proteins according to function via structural information. First, we performed the classification of binding sites for the data set arising from Ellingson and Zhang (2012) through the use of logistic regression. Then we approach the problem from the data set compiled by Kahraman et al. (2007). We calculated the covariance matrices of the binding sites' coordinates, which use the distance of each atom to the center of mass, and calculated the distance from an atom to the 1st, 2nd and 3rd principal axes. Then we obtained covariance matrices of these distances to serve as our data objects. Finally, we performed classification on these matrices using a variety of techniques, including nearest neighbor.

1 Introduction

Proteins are molecules consisting of chains of amino acids that fold into a 3-dimensional structure. They perform various functions by binding to various chemicals. Binding occurs at binding sites near surface of the protein. In protein-ligand binding, the ligand is usually a signal-triggering molecule binding to a site on a target protein.

The Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) is the largest data bank that provides information about the 3D structures of proteins and nucleic acid. As of April 18, 2017, there are $129,184$ biological macromolecular structural information files available in PDB and about 92.9% of them are proteins. X-ray crystallography and Nuclear Magnetic Resonance (NMR) are a few common methods used to obtain the protein structure. As of 2003 and 2010, respectively, Berman et al. (2003) and Chruszcz et al. (2010) showed that as many as 26% of the entries in the PDB have either unknown or putative function. Because much work has been done in this area, those figures have changed due to the discovering of functions and the additions of new structures to the database. The most up to date figures that are readily accessible are from the inventory done by Nadzirin and Firdaus-Raih (2012), which shows that there are about 42.53%

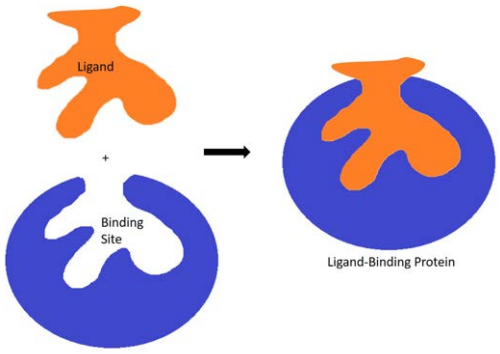


Figure 1: Protein Ligand Binding

of PDB entries that were categorized as proteins of unknown functions. By seeing how these figures have changed through past few years, we can understand the amount of research activity that has been going on over these years. As a result, the development of different context-based and structure based method is expanding drastically for prediction of unknown protein function.

1.1 General Problem

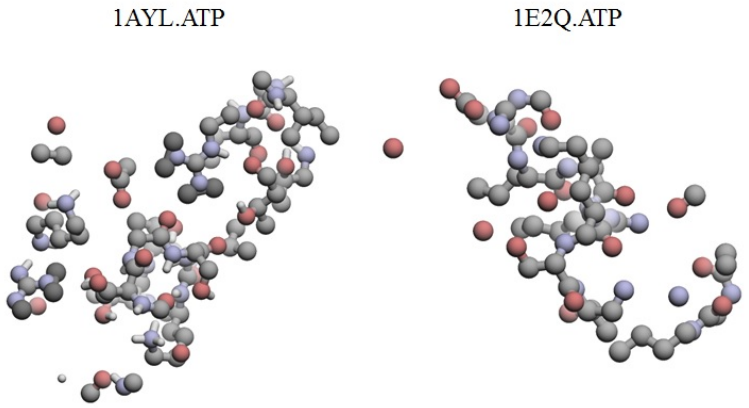
A common hypothesis in the literature is that proteins with similar functions should have binding sites with similar shapes and chemical properties. Some of the literature that suggest the above hypothesis are Ellingson and Zhang (2012), Hoffmann et al. (2010), and Kahraman et al. (2007). Therefore, analyzing the structure of proteins has become a very popular method in predicting unknown protein functions. One of the most useful applications of protein function prediction is in discovery of effective drug design.

The main goal of this research is to gain a better understanding of how structure relates to binding activity and then classify proteins according to function via structural information.

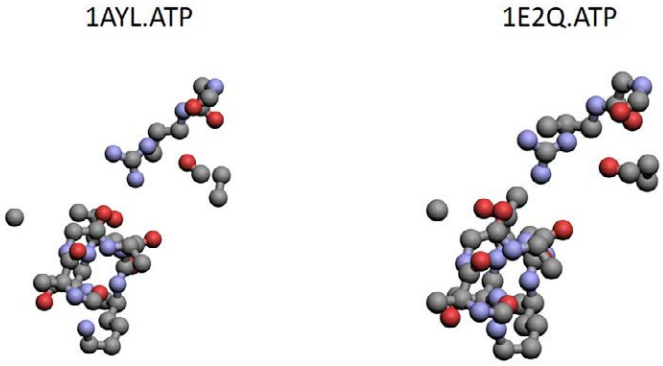
1.2 Background

Many researchers have conducted ligand-binding protein prediction studies by taking structural information into consideration as an initial step towards protein function prediction. The different types of such structure-based approaches are shape based methods, alignment base methods, graph-theoretic approaches, machine learning methods and, model-based methods. Some of the examples of these types of approaches can be found in Bertolazzi et al. (2014), Nussinov and Wolfson (1991), Fischer et al. (1994), Wallace et al. (1997), Kinoshita et al. (2001), Najmanovich et al. (2008), Hertz and Yanover (2006), Zhang et al. (2006), Ellingson and Zhang (2012), and Hoffmann et al. (2010).

Out of all these methods, more focus will be given to the alignment based method as I will be comparing the implemented model-based method results. In the alignment based method, binding sites are compared after finding spatially equivalent positions by superimposing them pairwise according to some criteria. These ideas have been used by different researchers. Shulman-Peleg et al. (2005) talked about two web servers and software packages named SiteEngine and Interface-to-Interface (I2I)-SiteEngine, for the recognition of the similarity of binding sites and interfaces. Gold and Jackson (2006) talks about



(a) Binding sites 1AYL and 1E2Q from ligand ATP



(b) Binding sites 1AYL and 1E2Q from ligand ATP after alignment (Ellingson and Zhang (2012))

Figure 2: Aligning binding sites

a database named SiteBase, which holds information about structural similarity between known ligand-binding sites. For the comparison of these binding sites, geometric hashing was used and the equivalent atom constellations between pairs of binding sites were identified. Hoffmann et al. (2010) talks about assessing similarity between pockets in protein binding sites by aligning them in 3D space and comparing the results with a convolution kernel. Then Ellingson and Zhang (2012) talks about a new algorithm named TIPSA (Triangulation-based Iterative-closest-point for Protein Surface Alignment) based on the iterative closest point (ICP) algorithm (Besl and McKay (1992)). Figure 2 describes an example of two binding sites that bind to the same ligand group after aligning and superimposing them by rotation and translation. It also supports the common hypothesis that binding sites with similar shapes bind to similar chemicals.

In the model-based method, features of binding sites for a given group were used and then implement a model for categorization. Ellingson and Zhang (2012) talked about radius of gyration as a model based approach. For a single binding site, the radius of gyration can be calculated by simply calculating the standard deviation of distance of the atom to the center of mass, as shown below:

$$Rg = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|x_i - \bar{x}\|^2}. \quad (1)$$

However, this does not contain much useful information, as it restrict all the information to one dimension.

2 Methodology

In this research, the ligand-binding protein prediction problem is approached by taking a higher level, objected oriented approach that summarizes the description of the binding site, so that it reduces the amount of information lost compared to most of the other approaches. Thereby, a model-based method is considered, where the nonparametric model is implemented by using the features of the binding sites for a given ligand group for understanding and classification purposes.

A model-based method that we named Covariances of Distances to Principal Axis (CDPA) was implemented by developing a new representation to each binding site using the entire covariance matrix. Since the coordinates of each binding site were based upon X-ray crystallography, each binding site has an arbitrary x, y and z coordinate system. As such, using distances from each atom to the x, y and z axes will not give a clear picture of the variability within each binding site. Therefore, in this method, the distance from each atom to the three principal axes that provide the three orthogonal directions of maximal variation in atom coordinates were found. For each binding site, atom coordinates were already known. Therefore, to find the three principal axes, the covariance of the coordinates were calculated and then we found the three eiganvectors that give the maximal variation. Once the principal axes were found for each binding site, the orthogonal projection of all atoms to each principal axis were calculated for every binding sites. Figure 3 shows an example of how to obtain the three principal axes.

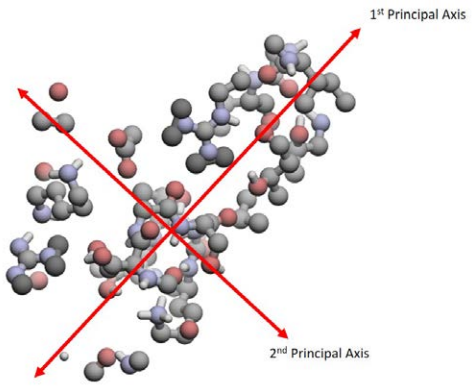


Figure 3: Orthogonal principal axes of 1AYL-ATP (note: The third principal axis is orthogonal to the page.)

Suppose d_{kj} is the distance from k^{th} atom to j^{th} principal axis, where, k=1, 2, ..., n_i ; j=1, 2, 3 ; i=1, 2, ..., 100 or i=1, 2, ..., 972 and n_i is the number of atoms for the i^{th} binding site. Then the distance matrix d_i can be represented as (2).

$$d_i = \left(\vec{d_{.1}} \quad \vec{d_{.2}} \quad \vec{d_{.3}}\right). \tag{2}$$

Once the distance matrices (d_i) were calculated for all binding sites, the covariance matrice (S_i) of the distance matrix was calculated for each binding site *i* and can be represented as (3).

$$S_i = Cov(d_i) = \begin{pmatrix} Var(d_{.1}) & Cov(d_{.1}, d_{.2}) & Cov(d_{.1}, d_{.3}) \\ Cov(d_{.2}, d_{.1}) & Var(d_{.2}) & Cov(d_{.2}, d_{.3}) \\ Cov(d_{.3}, d_{.1}) & Cov(d_{.3}, d_{.2}) & Var(d_{.3}) \end{pmatrix} \tag{3}$$

where, *i*=1, 2, . . . , 100 or *i*=1, 2, . . . , 972

The new representation of the binding sites are now in the form of covariance matrices. By doing this, we will not only consider the amount of variability, but also the shape of the variability of the binding sites. These new covariance matrices are now 3×3 Symmetric Positive Definite (SPD) matrices.

For example, from the extended Kahraman dataset, the binding site 1cbq, which binds with the ligand PO4, has only 8 atoms within $5.3\mathring{A}$ distance. This is the 582^{nd} observation of the sorted extended Kahraman dataset. The distance of each atom to each principal axis for the binding site 1cbq that binds with ligand PO4 can be shown using a 8×3 matrix as shown in equation (4). Then, the covariance matrix for ligand-binding site, 1cbq.PO4 can be shown as equation (5).

$$d_{582} = \begin{array}{c} \\ Atom1 \\ Atom2 \\ Atom3 \\ Atom4 \\ Atom5 \\ Atom6 \\ Atom7 \\ Atom8 \end{array} \begin{array}{c} \begin{array}{ccc} d_1 & d_2 & d_3 \end{array} \\ \begin{pmatrix} 1.7037 & 4.8011 & 5.0779 \\ 1.6471 & 2.3297 & 2.7237 \\ 0.7226 & 1.0903 & 1.3059 \\ 0.6659 & 0.9572 & 0.8708 \\ 1.5013 & 1.0586 & 1.2798 \\ 0.5727 & 3.1320 & 3.1330 \\ 1.5059 & 2.7635 & 3.1433 \\ 0.6002 & 3.5621 & 3.5920 \end{pmatrix} \end{array} \tag{4}$$

$$S_{582} = \begin{pmatrix} 0.2638 & 0.1667 & 0.2445 \\ 0.1667 & 1.9077 & 1.9459 \\ 0.2445 & 1.9459 & 2.0161 \end{pmatrix} \tag{5}$$

Once the covariance matrices of the distances for all binding sites were calculated, the primary model is the Covariance of Distances to Principal Axis (CDPA). In order to perform the classification of protein ligand-binding sites using CDPA, a non-parametric model based upon Mahalanobis distance of CDPA was used as the similarity measure. In order to calculate the Mahalanobis distances, means and covariances of S_i are required. Instead of using the covariance matrix S_i, to calculate Mahalanobis distance, the vectorized form of S_i and vectorized form of mean of S_i for each ligand is considered . Equation (6) represents the vectorized form of S_i.

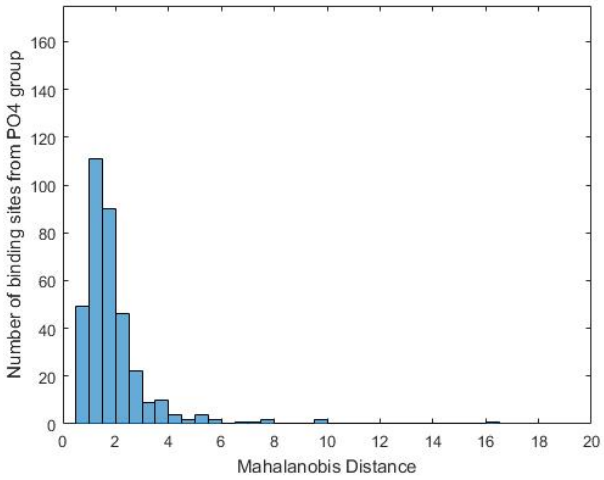


Figure 4: Mahalanobis Distance from each binding sites in PO4 group to the mean of PO4 group

$$Svec_i = \begin{pmatrix} Var(d_{.1}) \\ Var(d_{.2}) \\ Var(d_{.3}) \\ \sqrt{2} * Cov(d_{.1}, d_{.2}) \\ \sqrt{2} * Cov(d_{.1}, d_{.3}) \\ \sqrt{2} * Cov(d_{.2}, d_{.3}) \end{pmatrix} \tag{6}$$

The reason why the last three entires of the equation (6) are multiplied by the $\sqrt{2}$ is so that the Euclidean distance between any two observations remains the same whether in matrix or vector form. In other words, the Frobenius norm of the matrix will be equal to the norm of the vectorized form of the matrix.

In order to calculate the mean covariance ($\bar{S}_j$) of the j^{th} ligand, the covariance matrices of the binding sites that corresponds to each ligands are considered and their mean is calculated. For all nine groups, 3×3 mean covariances were found. Then, the vectorized form of each mean covariances $\overline{Svec}_j$ is considered. Next, the covariance of $Svec_i$ of j^{th} ligand ($Cov(Svec_{ij})$) is calculated and we named that 6×6 matrix as Σ_j, where $j = 1, 2, \ldots, 9$. Then the Mahalanobis distance D_i from each binding site i to the mean of each ligand group j is calculated and the equation is given in (7).

$$D_i = \sqrt{(Svec_i - \overline{Svec}_j)' \Sigma_j^{-1} (Svec_i - \overline{Svec}_j)} \tag{7}$$

where, *i*=1, 2, ..., 100 or *i*=1, 2, ..., 972 and *j*=1, 2, ..., 9.

Figures 4 and 5 represent an example of the reasons why this is considered a model-based approach. Figure 4 represents the frequency histogram of Mahalanobis distances from each binding site in PO4 to the mean of PO4. Figure 5 represents the frequency histogram of Mahalanobis distances from each binding site that are not coming from PO4 to the mean of PO4. Figure 4 depicts that almost all binding sites from PO4 are closer to their mean, whereas Figure 5 shows that most of the binding sites that are coming from other ligand groups except PO4, have higher Mahalanobis distances to the mean of PO4.

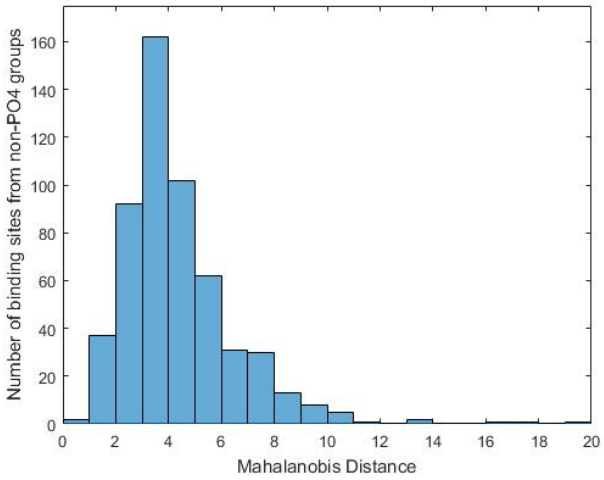


Figure 5: Mahalanobis Distance from each binding sites in non-PO4 groups to the mean of PO4 group

Once D_i is calculated for each binding site, two classification procedures were used. The 1-nearest mean classification and predicted probabilities of Polytomous (Multinomial) Logistic Regression model were used to classify these binding sites into the ligand group. For each, the classification error is calculated as the proportion of misclassified binding sites.

Then the results obtained using the model-based approach are compared to the alignment-based method used by Ellingson and Zhang (2012) and Hoffmann et al. (2010), where they superimpose binding sites pairwise, based on certain criteria. We compare our approach to just two methods because many methods are not readily accessible for use with new data and only very few methods use a common benchmark dataset. Therefore, we used those methods that used a common benchmark data set to compare our results with the same dataset.

3 Data

Kahraman et al. (2007), which is known in the literature as the Kahraman dataset. It consists of 100 protein binding sites which bind to one of 10 ligands (AMP, ATP, FAD, FMN, GLC, HEM, NAD, PO4, EST, AND). The second dataset is the one that used in Hoffmann et al. (2010), which they named the extended Kahraman Dataset. It consists of 972 protein binding sites, of which the Kahraman dataset is a subset. These sites also bind to one of the above same 10 ligands. There are only 2 and 4 binding sites that forms with ligand AND and EST, respectively. Therefore, both ligand groups are considered as one ligand group named "Steroid". These ligands vary in size and flexibility. PO4 is the smallest ligand in size and the most rigid molecule. FAD is the largest in size and is the highest in flexibility.

For all protein structures, information about the 3D structures of full binding sites were downloaded from the PDB (Protein Data Bank) that were determined by X-ray crystallography (Henrick and Thornton (1998)). Then, all the binding sites that consist of those atoms within radius of $5.3\mathring{A}$ (Angstroms) to the specified ligand were extracted. Figure 6 shows an example to a full binding site of 1AYL and an extracted binding site within $5.3\mathring{A}$ to the

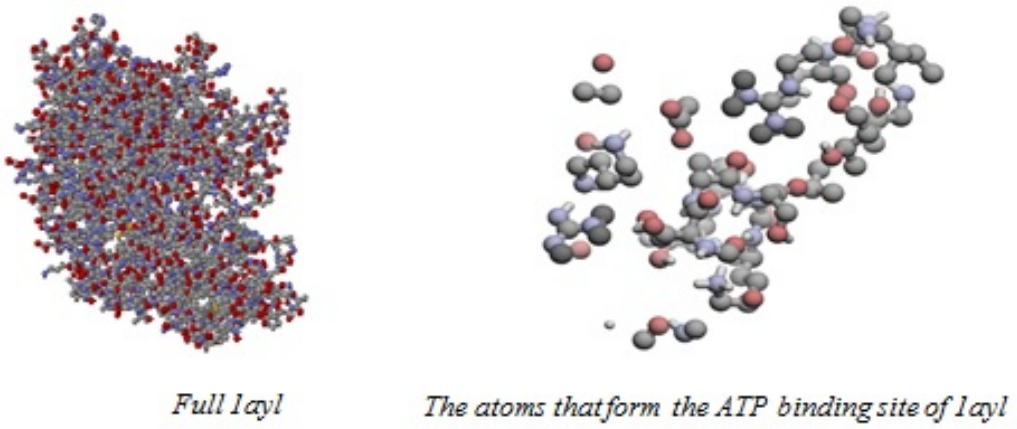


Figure 6: Full binding site 1AYL and the extracted binding site 1AYL within the radius of $5.3\mathring{A}$ to the ligand ATP.

ligand group ATP. We selected this radius of $5.3\mathring{A}$, because Hoffmann et al. (2010) found experimentally that it corresponds to a good default value. It also facilitates comparisons to Ellingson and Zhang (2012) and Hoffmann et al. (2010). Tables 1 and 2 show the number of binding sites that belong to each ligand group of the Kahraman dataset and the extended Kahraman dataset, respectively.

Table 1: Kahraman dataset

	AMP	ATP	FAD	FMN	GLC	HEM	NAD	PO4	STEROID
Number of Sites	9	14	10	6	5	16	15	20	5

Table 2: extended Kahraman dataset

	AMP	ATP	FAD	FMN	GLC	HEM	NAD	PO4	STEROID
Number of Sites	63	78	79	58	88	113	91	389	6

4 Results

4.1 Ligand Classification for Kahraman dataset

The results obtained for the Kahraman dataset were considered first. Multi Dimensional Scaling (MDS) is used to plot a visual representation of the covarince matrix representations. Figure 7 is the MDS plot for the Kahraman dataset. It shows a clear separation between the groups. This gives us a motivation that this method could work well.

Then, the summary of classification error found using CDPA with the nearest mean classification is provided in Table 3.

To validate the model, a testing set is simulated by adding different amounts of noise to each coordinate. We used 100 different noise levels from 0.01 to 1 with 0.01 increments.

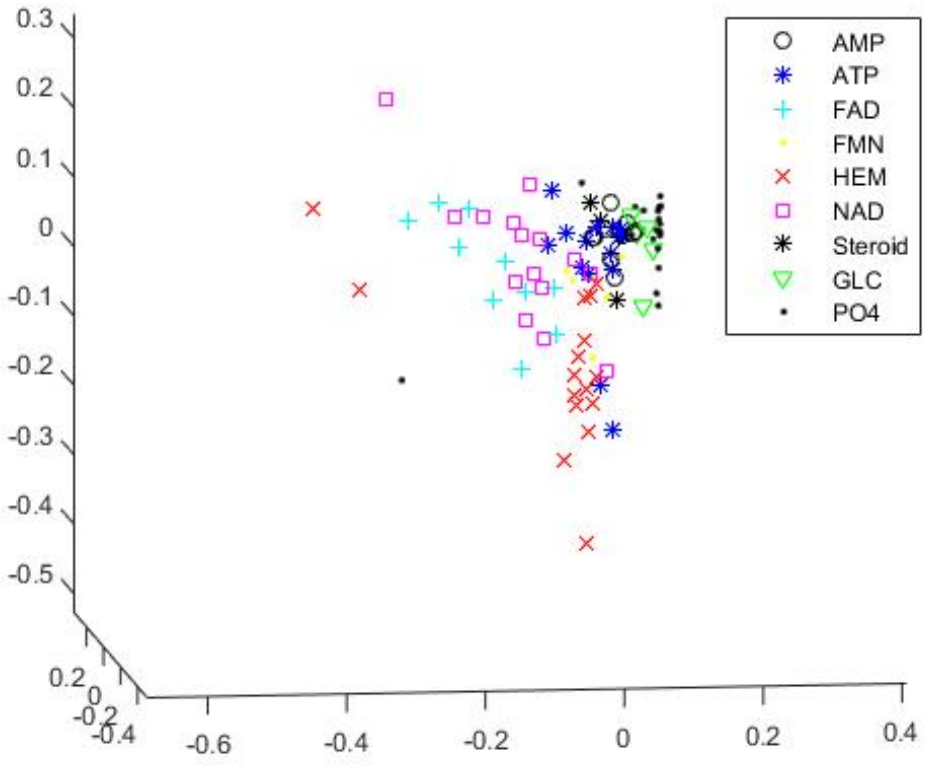


Figure 7: MDS plot for Kahraman (5.3 Å) dataset.

Then each such noise added to the data was replicated 100 times and we found their classification errors. Figure 8 represents the classification error with respect to the different noise level for the Kahraman dataset.

Figure 8 explains that, for each increasing noise, there is a change in the model with a consistency. That is, even though the classification error increases when the noise is increasing, there is a trend between the variables. As a result, it validates the model performance. Here, the noise is added to each coordinates and adding a noise of 1 means a lot of noise. That is because, when selecting a binding site, we considered the atoms that are within a $5.3\mathring{A}$. Therefore the entire diameter of it in any dimension should be no more than roughly about $10\mathring{A}$. Therefore standard deviation of the noise been $1\mathring{A}$ to each coordinate would dramatically impact the position of each atom and as a result the classification error will increase significantly. But with a smaller amount of noise it does not change very much and the classification error is also closer to the CE of original data. Also, it can be observed that the amount of thickness of the confidence band is low during lower level of noises and it then being some what constant at higher level of noises. It can also be noticed that up until about noise of 0.25, we are still not very far from what we originally have with the data and its CEs are still lower than other methods.

Table 3: Results for nearest mean classification for the Kahraman (5.3 Å) dataset.

Method	Classification Error
TYPSA:TI	0.43
Gyr	0.54
TYPSA:TI+Gyr	0.29
$Sup - CK_L$	0.27
CDPA with nearest mean classification	**0.15**

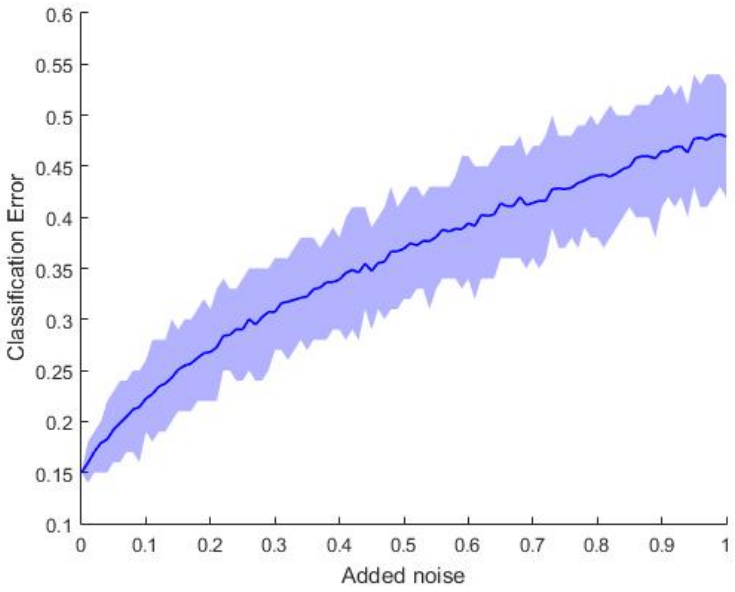


Figure 8: Classification error with respect to different noise level for Kahraman dataset

4.2 Ligand Classification for Extended Kahraman dataset

The extended Kahraman dataset, the summary of classification error found using CDPA is provided in Table 4. Ellingson and Zhang (2012) unable to perform their method to the extended Kahraman dataset introduced by Hoffmann et al. (2010), due to the extreme computational cost of that method.

Here, CDPA shows a higher CE than Hoffmann et al. (2010)'s method. This is mainly because Hoffmann et al. (2010) considered a pairwise alignment method. For instance, if there are two binding sites that are closer to each other but far away from their actual ligand group, but still according to the pairwise alignment, they will still be grouped in to the actual ligand group and using the model based method those binding sites will be misclassified into a ligand group that is more closer to it.

Figure 9 represents a visual comparison of classification performed using nearest mean and predicted probabilities with added noise for the simulated dataset. The classification performed using nearest mean shows that the mean line is closer to a line and it has a increasing linear trend between CE and noise level. However, the classification performed using predicted probabilities of polytomous logistic regression is not linear and it starts off really slow and then its starts increasing with the more noise. That is, it is less sensitive to smaller noise than with nearest mean classification.

Table 4: Results for nearest neighbor classification for the extended Kahraman (5.3 Å) dataset.

Method	Classification Error
TYPSA:TI	*
TYPSA:TI+Gyr	*
$Sup - CK_L$	0.19
CDPA with nearest neighbor classification	**0.32**
CDPA with predicted probabilities	**0.25**

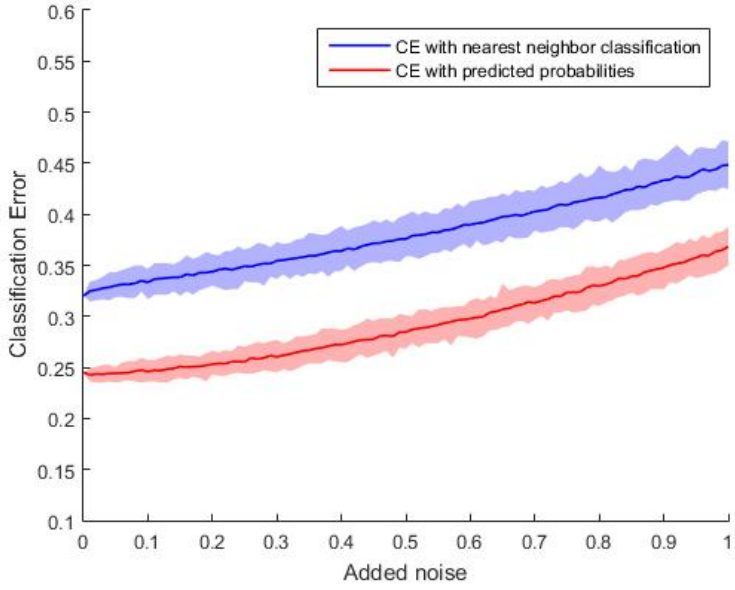


Figure 9: Classification Error with validation

Figure 10 shows the visual representation of all the binding sites together in a MDS plot. It clearly shows how each ligand groups disperse around, and how the model affect the classification directly.

5 Conclusion

To summarize, the performance of CDPA is better than that of the other methods using Kahraman dataset. The model-based approach used in this research is very fast compared to the alignment-based approaches used by Ellingson and Zhang (2012) and Hoffmann et al. (2010). For instance, in Ellingson and Zhang (2012), to perform pairwise alignment for the Kahraman dataset, it took between 2 to 6 seconds per alignment. Since there are 100 binding sites in this dataset, there will be 4950 pairs, requiring a minimum of 9900 seconds to compare all of the pairs. For Hoffmann et al. (2010), the algorithm running time per pockets pair varied between 0.2 and 1.3 seconds. This means that, at a minimum, it will take 990 seconds to compare all the pairs using that methodology. With CDPA, to calculate the similarity measure for all binding sites and classify them to their ligand group, for the Kahraman dataset, it took only 2.4 seconds. This illustrates the computational advantage of

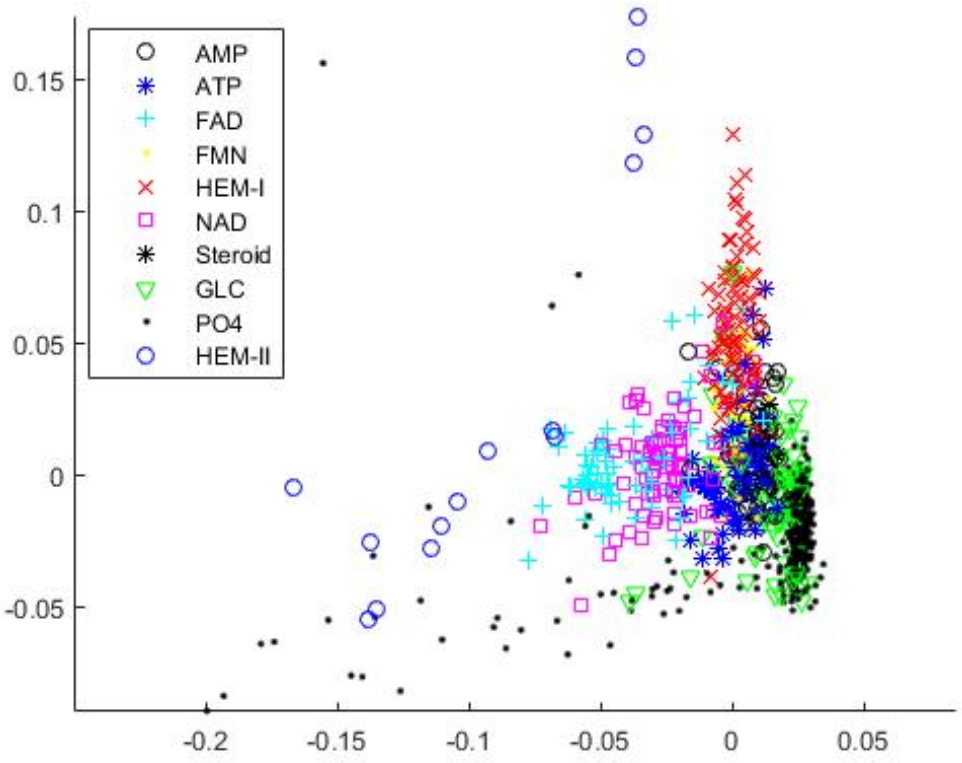


Figure 10: MDS plot of all ligands together for the extended Kahraman (5.3 Å) dataset

our model-based methodology since it is not restricted to pairwise comparisons. Furthermore, this computational efficiency is further aided by the method being alignment-free.

References

Berman, H. M., Hendrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. 10:980.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acid Research*, 28:235–242.

Bertolazzi, P., Guerra, C., and Liuzzi, G. (2014). Predicting protein-ligand and protein-peptide interfaces. *The European Physical Journal Plus*, pages 129–132.

Besl, P. and McKay, N. (1992). A method for registration of 3-d shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 14:239–256.

Bonnabel, S. and Sepulchre, R. (2009). Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31:1055–1070.

Burley, S. K. (2000). An overview of structural genomics. *Nature Structural & Molecular Biology*, 7:932–934.

Champe, P. C. and Harvey, R. A. (1994). *Biochemistry*. J. B. Lippincott Company, Philadelphia.

Chruszcz, M., Domagalski, M., Osinski, T., Wlodawer, A., and Minor, W. (2010). Unmet challenges of structural genomics. *Current Opinion in Structural Biology*, 20:587–597.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3:1102–1123.

Ellingson, L., Groisser, D., Osborne, D., Patrangenaru, V., , and Schwartzman, A. (2015). Nonparametric bootstrap of sample means of positive-definite matrices with an application to diffusion-tensor-imaging data. *Communications in Statistics - Simulation and Computation*.

Ellingson, L. and Zhang, J. (2012). Protein surface matching by combing local and global geometric information. *PLOS one*, 07.

Fischer, D., Wolfson, H., Lin, S. L., and Nussinov, R. (1994). Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: Potential implications to evolution and to protein folding. *Protein Science*, 3:769–778.

Gold, N. D. and Jackson, R. M. (2006). Sitesbase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Research*, 34:D231–D234.

Henrick, K. and Thornton, J. (1998). Psq: a protein quaternary structure file server. *Trends in Biochemical Sciences*, 23:358–361.

Hertz, T. and Yanover, C. (2006). Pepdist: A new framework for protein-peptide binding prediction based on learning peptide distance functions. *NIPS workshop on New Problems and Methods in Computational Biology*, 7:1–15.

Hoffmann, B., Zaslavskiy, M., Jean-Philippe, V., and Stoven, V. (2010). A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3d: application to ligand prediction. *BMC Bioinformatics*, 11(99).

Kahraman, A., Morris, R. J., Laskowski, R. A., and Thornton, J. M. (2007). Shape variation in protein binding pockets and their ligands. *Journal of Molecular Biology*, 368:283–301.

Kinoshita, K., Furui, J., and Nakamura, H. (2001). Identification of protein functions from a molecular surface database, ef-site. *Journal of Structural and Functional Genomics*, 2:9–22.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2:83–97.

Montelione, G. T. (2001). Structural genomics: An approach to the protein folding problem. *PNAS*, 98:13488–13489.

Nadzirin, N. and Firdaus-Raih, M. (2012). Proteins of unknown function in the protein data bank (pdb): An inventory of true uncharacterized proteins and computational tools for their analysis. *International Journal of Molecular Science*, 13:12761–12772.

Najmanovich, R., Kurbatova, N., and Thornton, J. (2008). Detection of 3d atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, 24:i105–i111.

Nussinov, R. and Wolfson, H. J. (1991). Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proceedings of the National Academy of Sciences of the United States of America(PNAS)*, 88:10495–10499.

Shulman-Peleg, A., Nussinov, R., and Wolfson, H. J. (2005). Siteengines: recognition and comparison of binding sites and proteinprotein interfaces. *Nucleic Acids Research*, 33:W337W341.

Wallace, A. C., Borkakoti, N., and Thornton, J. M. (1997). Tess: a geometric hashing algorithm for deriving 3d coordinate templates for searching structural databases. application to enzyme active sites. *Protein Science*, 6:2308–2323.

Zhang, L., Shao, C., Zheng, D., and Gao, Y. (2006). An integrated machine learning system to computationally screen protein databases for protein binding peptide ligands. *The American Society for Biochemistry and Molecular Biology, Inc*, 5:1224–1232.