Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 625403, 8 pages
http://dx.doi.org/10.1155/2013/625403

*Research Article*

# Application of Improved Three-Dimensional Kernel Approach to Prediction of Protein Structural Class

**Xu Liu,[1] Yuchao Zhang,[2,3] Hua Yang,[1] Lisheng Wang,[1] and Shuaibing Liu[1,4]**

[1] *School of Chemistry & Chemical Engineering, Guangxi University, Guangxi Province, Nanning 530004, China*
[2] *State Key Laboratory of Medical Genomics, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai Jiao Tong University School of Medicine, Shanghai 200240, China*
[3] *Graduate School of the Chinese Academy of Sciences, Beijing 100049, China*
[4] *College of Pharmacy, Guangxi University of Chinese Medicine, Nanning 530001, China*

Correspondence should be addressed to Xu Liu; wendaoliuxu@hotmail.com

Received 25 March 2013; Revised 4 May 2013; Accepted 10 May 2013

Academic Editor: Bing Niu



Kernel methods, such as kernel PCA, kernel PLS, and support vector machines, are widely known machine learning techniques in biology, medicine, chemistry, and material science. Based on nonlinear mapping and Coulomb function, two 3D kernel approaches were improved and applied to predictions of the four protein tertiary structural classes of domains (all-$\alpha$, all-$\beta$, $\alpha/\beta$, and $\alpha+\beta$) and five membrane protein types with satisfactory results. In a benchmark test, the performances of improved 3D kernel approach were compared with those of neural networks, support vector machines, and ensemble algorithm. Demonstration through leave-one-out cross-validation on working datasets constructed by investigators indicated that new kernel approaches outperformed other predictors. It has not escaped our notice that 3D kernel approaches may hold a high potential for improving the quality in predicting the other protein features as well. Or at the very least, it will play a complementary role to many of the existing algorithms in this regard.

## 1. Introduction

Due to the rapid development of genome and protein science, the biological information has expanded dramatically. Therefore, it is very important and highly desirable for computers to manage, organize, and interpret the information. As a part of biochemistry, study of protein structure classes has become a hot topic, because of experimental and theoretical purposes. Artificial neural networks, support vector machines, kernel methods, and ensemble algorithms are widely known machine learning techniques in biology, medicine, chemistry, and material science [1–10]. In this work, two classification problems, protein's tertiary structure classes of domains and membrane protein types, were researched with some machine learning techniques.

Several motifs pack together to form compact, local, and semiindependent units called domains. The details of proteins domains structures are extremely complicated and irregular. But their overall structural frames are simple, regular, and truly elegant [11–13]. Many protein domains often have similar or identical folding patterns even if they are quite different according to their sequences [14–16]. The overall 3D structure of the polypeptide chain is referred to as the protein's tertiary structure. Levitt and Chothia proposed to classify protein tertiary structures into the following four structural classes based on the secondary structural content of the domains. (1) All-$\alpha$: it is formed essentially by $\alpha$-helices. This class is dominated by small folds, many of which form a simple bundle with helices running up and down. (2) All-$\beta$: this class has a core composed of antiparallel $\beta$-sheets, usually two sheets pack against each other. (3) $\alpha/\beta$: this class contains both $\alpha$-helices and $\beta$-strands that are largely interspersed in forming mainly parallel $\beta$-sheet; (4) $\alpha+\beta$: this class also contains both of the two secondary structure elements that, however, are largely segregated in forming mainly antiparallel $\beta$-sheets.

This concept of structural class has ever since been widely used as an important attribute for characterizing the overall folding type of proteins domains. Lots of methods have been

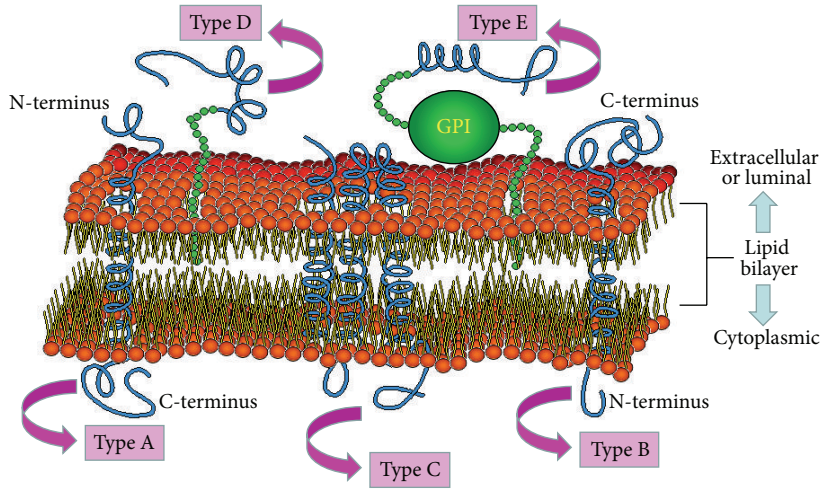


Figure 1: Five types of membrane proteins.

made to predict the structural classes based on the knowledge of protein sequences [17].

The research of membrane protein type is also important because of the special biological functions. The biomembrane usually contains some specific proteins and lipid components that enable it to perform its unique roles in the cell and organelle.

Furthermore, several studies show that many membrane proteins are also the key targets of drug discovery, particularly membrane channel proteins [18–20]. Membrane proteins can be further classified into the five types [21–23]: (a) type A membrane protein is single-pass transmembrane protein which has an extracellular (or luminal) N-terminus and cytoplasmic C-terminus for a cell (or organelle) membrane; (b) type B membrane protein is single-pass transmembrane protein which has an extracellular (or luminal) C-terminus and cytoplasmic N-terminus for a cell (or organelle) membrane; (c) type C is multipass transmembrane protein: the polypeptide crosses the lipid bilayer multiple times; (d) type D membrane proteins are lipid chain-anchored membrane proteins: they are bound to the membrane by one or more covalently attached fatty acid chains or other types of lipid chains called prenyl groups; (e) type E is GPI-anchored membrane protein which is bound to the membrane by a glycosylphosphatidylinositol (GPI) anchor.

Researchers have applied classification algorithm to predict the types of membrane proteins based on their amino acid composition [24, 25]. Figure 1 shows the forms and the locations of different membrane proteins.

The first goal of this paper is to illustrate the application of 3D kernel approach as a relatively new tool in proteins domains field for classification purposes. And the second goal is to show that the new approach can be applied to analysis of membrane protein types.

## 2. Materials and Methods

*2.1. Kernel Function.* Kernel function was originally a kind of functions used in integral operator research. However, Vapnik implemented this function in his newly invented SVMs method [26]. The use of kernel function makes SVMs able to treat nonlinear data processing problems by using linear algorithms. The basic idea of kernel function is to map the data $\mathbf{X}$ into a higher-dimensional feature space $\mathbf{F}$ via a nonlinear mapping $\Phi$ and then to do classification and regression in this space. There are four commonly used kernel functions:

linear kernel

$$K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x} \cdot \mathbf{y} \rangle + \theta. \tag{1}$$

polynomial kernel

$$K(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x} \cdot \mathbf{y} \rangle + \theta)^d. \tag{2}$$

Gaussian (RBF) kernel

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right). \tag{3}$$

sigmoid kernel

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\nu \langle \mathbf{x} \cdot \mathbf{y} \rangle + r). \tag{4}$$

The elegance of using kernel function lied in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(\mathbf{x})$. Any function that satisfies Mercer's condition can be used as kernel function.

*2.2. Kernel PCA.* Principal component analysis (PCA) is a versatile and easy-to-use multivariate mathematical-statistical method in multivariate data analysis and the extraction of maximal information [27, 28]. It is a linear transformation approach that compresses high-dimensional data with minimum loss of data information. PCA is performed in the original sample space, whereas kernel PCA (KPCA) applies

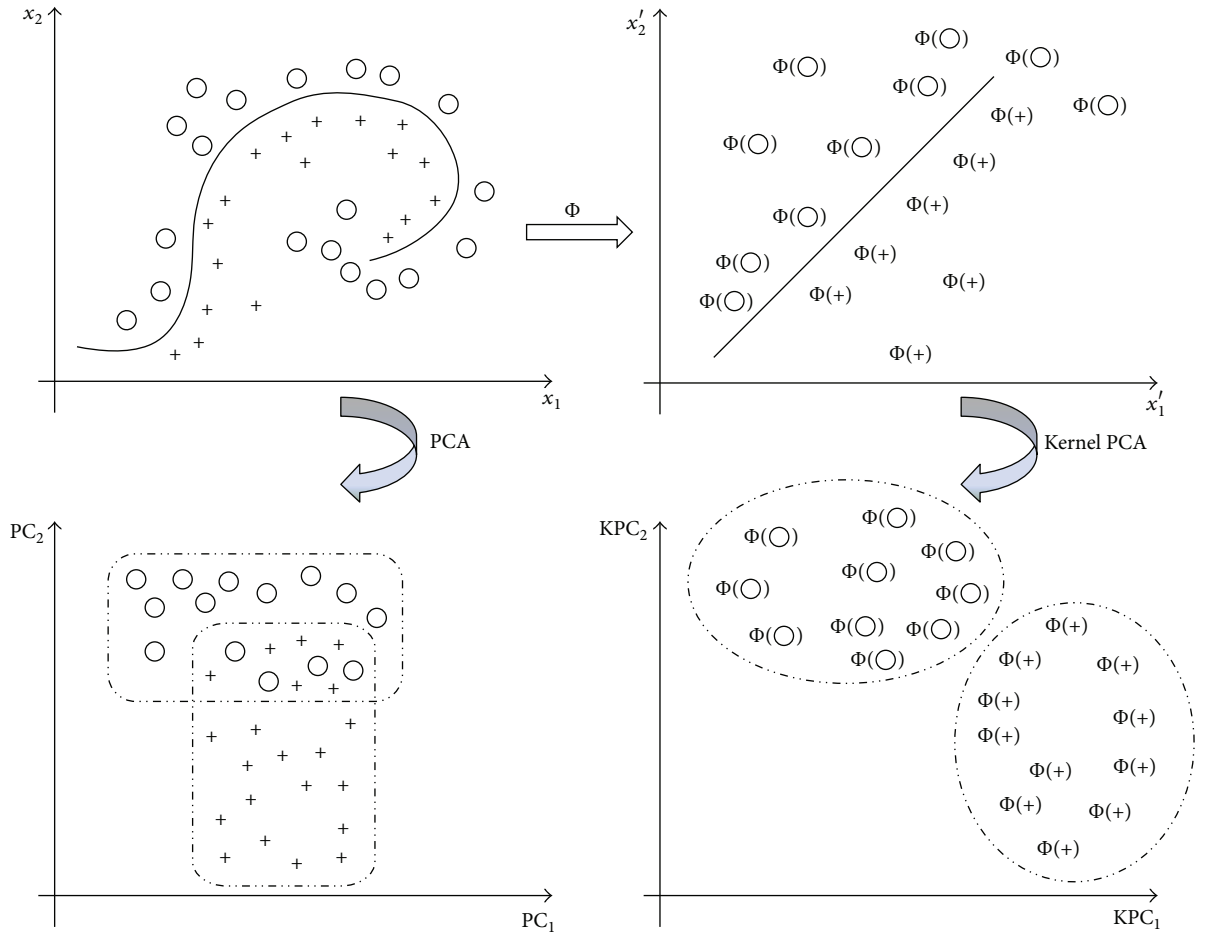


Figure 2: The mapping Φ embeds the data points in a feature space.

kernel functions in the input space to achieve the same effect of the expensive nonlinear mapping.

From Figure 2, it is found that the basic idea of KPCA is to map the original dataset into some higher dimensional feature space. In this complex space, PCA can be applied to establish a linear relationship which is nonlinear in the original input space [29, 30]. For the special case in which $\Phi(\mathbf{x}) = \mathbf{x}$, KPCA is equivalent to linear PCA. From this viewpoint, KPCA can be regarded as a generalized version of linear PCA.

For PCA, with data $\mathbf{X} = [x_1, x_2, \ldots, x_n]^T \in R^p$, one can first compute the covariance matrix $\mathbf{C}$:

$$\mathbf{C} = \frac{1}{n}\mathbf{X}^{\mathbf{T}}\mathbf{X}. \tag{5}$$

A principal component $\mathbf{v}$ is computed by solving the following eigenvalue problem:

$$\boldsymbol{\lambda}\mathbf{v} = \mathbf{C}\mathbf{v} = \frac{1}{n}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{v}. \tag{6}$$

Thus, the eigenvectors can be written as ($\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$)

$$\mathbf{v} = \sum_{i=1}^{n} \alpha_i x_i = \mathbf{X}^{\mathbf{T}}\boldsymbol{\alpha}. \tag{7}$$

Then, the eigen value problem can be represented by the following simple form:

$$\boldsymbol{\lambda}\boldsymbol{\alpha} = \frac{1}{n}\mathbf{K}\boldsymbol{\alpha}, \tag{8}$$

where $\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathbf{T}} \in R^{n\times n}$ is a linear kernel matrix. To derive KPCA, one firstly needs to map the data $\mathbf{X}$ into a feature space $\mathbf{F}$ (i.e., $\mathbf{M} = [\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_n)]^T$). Hence, a nonlinear kernel matrix $\mathbf{K}$ ($\mathbf{K} = \mathbf{M}\mathbf{M}^{\mathbf{T}} \in R^{n\times n}$) can be directly generated by means of specific kernel function ((1), (2), (3), and (4)). For extracting features of a new sample $x$ with KPCA, one simply projects the mapped sample $\Phi(\mathbf{x})$ onto the first $k$ projections $\mathbf{V}_k$,

$$\mathbf{V}_k \cdot \Phi(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^k \left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \right\rangle. \tag{9}$$

KPCA is to map the original data (in the input space) with nonlinear features into kernel feature space in which the linear PCA algorithm is then performed. Therefore, KPCA, being suitable to describe the nonlinear structure of data set, can be regarded as a generalized version of linear PCA.

*2.3. GDA.* Generalized discriminant analysis (GDA) is a method designed for nonlinear classification [31–33]. It is a nonlinear extension of linear discriminant analysis (LDA) based on a kernel function $\Phi$ which transforms the original space $\mathbf{X}$ to a new high dimensional feature space $\mathbf{F}$. The within-class (or total) scatter ($\mathbf{W}^{\Phi}$) and between-class scatter ($\mathbf{B}^{\Phi}$) matrixes of the nonlinearly mapped data are as follows:

$$\mathbf{W}^{\Phi} = \sum_{c=1}^{C} \sum_{\mathbf{x} \in \mathbf{X}_c} \Phi(\mathbf{x}) \Phi(\mathbf{x})^T, \tag{10}$$

$$\mathbf{B}^{\Phi} = \sum_{c=1}^{C} M_c \mathbf{m}_c^{\Phi} \left(\mathbf{m}_c^{\Phi}\right)^T. \tag{11}$$

In (11), $\mathbf{m}_c$ is the mean of class $\mathbf{X}_c$ and $M_c$ is the number of samples belonging to $\mathbf{X}_c$. The aim of the GDA is to find such projection matrix $\mathbf{U}^{\Phi}$ that maximizes the following Fisher criterion:

$$\mathbf{U}_{\text{opt}}^{\Phi} = \arg\max \frac{\left|\left(\mathbf{U}^{\Phi}\right)^T \mathbf{B}^{\Phi} \mathbf{U}^{\Phi}\right|}{\left|\left(\mathbf{U}^{\Phi}\right)^T \mathbf{W}^{\Phi} \mathbf{U}^{\Phi}\right|} = \left[\mathbf{u}_1^{\Phi}, \ldots, \mathbf{u}_N^{\Phi}\right]. \tag{12}$$

From the theory of reproducing kernels, any solution $\mathbf{u}^{\Phi} \in \mathbf{F}$ must lie in the span of all training samples in $\mathbf{F}$:

$$\mathbf{u}^{\Phi} = \sum_{c=1}^{C} \sum_{i=1}^{M_c} \alpha_{ci} \Phi(\mathbf{x}_{ci}), \tag{13}$$

where $\alpha_{ci}$ are some real weights and $x_{ci}$ is the $i$th sample of the class $c$. The solution is obtained by solving ($\boldsymbol{\alpha} = [\boldsymbol{\alpha}_c]$, $c = 1, 2, \ldots, C$; $\boldsymbol{\alpha}_c = [\boldsymbol{\alpha}_{ci}]$, $i = 1, 2, \ldots, M_c$):

$$\boldsymbol{\lambda} = \frac{\boldsymbol{\alpha}^T \mathbf{KDK} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{KK} \boldsymbol{\alpha}}. \tag{14}$$

$\mathbf{K}$ is the $n \times n$ kernel matrix composed of the dot products of nonlinearly mapped data. And $\mathbf{D} = \text{diag}(\mathbf{D}_1, \ldots, \mathbf{D}_c)$, where $\mathbf{D}_i$ is a $n_i \times n_i$ matrix with entries all equal to $1/n_i$.

*2.4. New Improved 3D Kernel Approach: 3D KPCA and 3D GDA.* Traditional KPCA and GDA are typical multivariate two-dimension statistical methods. In this work, KPCA and GDA are improved with three-dimensional projection and the concept of electric field intensity.

Firstly, the data of training samples are projected onto three-dimensional space by KPCA or GDA algorithm with satisfactory classification effect. The three-dimensional coordinate axes are, respectively, the first kernel principal component, second kernel principal component, and third kernel principal component or the direction vectors of generalized discriminant analysis.

Secondly, we need to estimate the class (unknown) of new projection points, such as membrane protein types of test sample data. There are two estimation methods in this work: K-Nearest Neighbor algorithm (KNN) [34] and class intensity model.

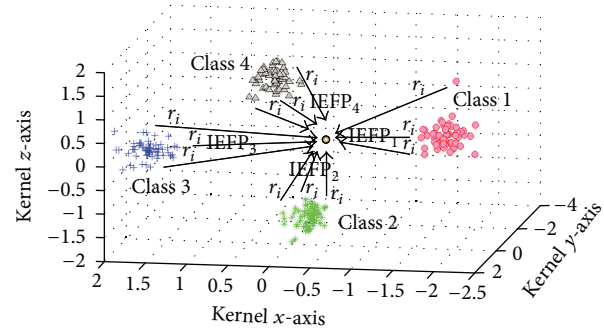


Figure 3: IEFP with different classes in 3D kernel space.

KNN algorithm estimation: new projection point (test sample) is classified by a majority vote of its neighbors (training samples in kernel three-dimensional space).

Class intensity model estimation: the projection point of one training data can be considered as point charge. The species of charge is related to the class of sample. And the Electric Quantity of Point Charge (EQPC) is related to the number of samples ($n_C$) which belongs to some class:

$$\text{EQPC}_C = \frac{1}{n_C}. \tag{15}$$

The value of EQPC is negative related with the sample amount of same class. Based on the Coulomb law and formula of intensity of electric field, the Intensity of Electric Field of one Point (IEFP) in 3D space is

$$\text{IEFP}_C = \sum_{i=1}^{n_C} \frac{\text{EQPC}_C}{r_i^2}, \tag{16}$$

where $r$ is distance between point charge and the space point.

Therefore, in class intensity model, IEFP is a criterion of classification. For example, there are four classes in training data: class 1, class 2, class 3, and class 4 in Figure 3. After projecting with kernel methods, all projection class charge points of training data can form a space electric field. The test sample can be projected onto this space with the same kernel methods. Figure 3 illustrates the relationship between point charge of different class and corresponding IEFP. To project position of test sample, if there exist $\text{IEFP}_1 > \text{IEFP}_2$, $\text{IEFP}_1 > \text{IEFP}_3$ and $\text{IEFP}_1 > \text{IEFP}_4$, test sample should belong to class 1.

## 3. Results and Discussion

*3.1. System and Software Used for Data Analysis.* The calculations were carried out using the Intel(R) Core(TM) Duo CPU T5870 GHz computer running Windows XP operating system. All the learning input data were range-scaled to [0~1] in this work. The improved 3D kernel approach software package including 3D kernel PCA and 3D GDA was programmed in our laboratory referring to the literature [29, 31] based on statistical pattern recognition toolbox for MATLAB [35].

*3.2. Application of Improved 3D Kernel Approach to Protein's Tertiary Structure Classes of Domains.* The protein datasets

Table 1: LOOCV success rates by component-coupled, neural network, SVMs, AdaBoost, and improved 3D kernel approach.

| Dataset | Algorithm | All-$\alpha$ | All-$\beta$ | $\alpha/\beta$ | $\alpha+\beta$ | Overall |
|---|---|---|---|---|---|---|
| Dataset A (277 domains) | Component-coupled | 84.3% | 82.0% | 81.5% | 67.7% | 79.1% |
| | Neural networks | 68.6% | 85.2% | 86.4% | 56.9% | 74.7% |
| | SVMs | 74.3% | 82.0% | 87.7% | 72.3% | 79.4% |
| | AdaBoost | 87.1% | 95.1% | 98.7% | 81.5% | 90.9% |
| | 3D kernel | 88.6% | 85.3% | 93.8% | 77.0% | 86.6% |
| Dataset B (498 domains) | Component-coupled | 93.5% | 88.9% | 90.4% | 84.5% | 89.2% |
| | Neural networks | 86.0% | 96.0% | 88.2% | 86.0% | 89.2% |
| | SVMs | 88.8% | 95.2% | 96.3% | 91.5% | 93.2% |
| | AdaBoost | 96.2% | 92.1% | 98.5% | 89.9% | 94.2% |
| | 3D kernel | 91.6% | 95.3% | 99.3% | 92.3% | 95.0% |

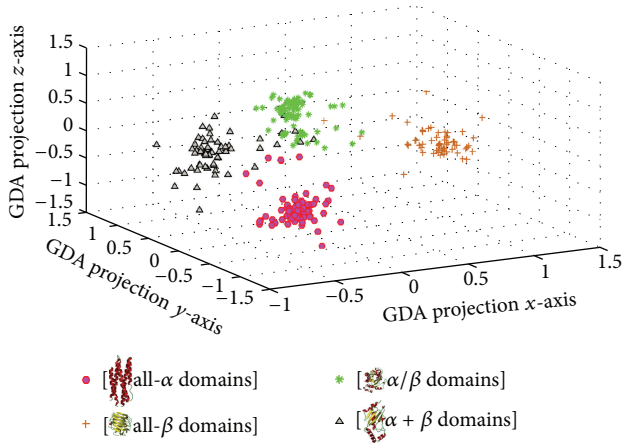


Figure 4: Distribution of different protein's tertiary structure classes data in 3D kernel space.

studied here were taken from Niu and his coworkers [17]. In dataset A, there are 277 protein domains, of which 70 are all-$\alpha$ domains, 61 all-$\beta$, 81 $\alpha/\beta$, and 65 $\alpha+\beta$. In dataset B, there are 498 protein domains, of which 107 are all-$\alpha$ domains, 126 all-$\beta$, 136 $\alpha/\beta$, and 129 $\alpha+\beta$. The amino acid composition was used to represent the sample of a protein domain.

To demonstrate the power of 3D kernel methods, computations were performed by the Leave-One-Out Cross-Validation (LOOCV), which are widely used by more and more investigators in testing the power of various predictors. As such, the data set of $n$ samples was divided into two disjoint subsets including a training data set ($n-1$ samples) and a test data set (only 1 sample). After developing each model based on the training set, the omitted data was predicted and the difference between experimental value and predicted value was calculated [36–38].

Based on dataset A, it was found that the projection with Gaussian (see (3), $\sigma = 0.5$) kernel function and KNN ($K = 3$) algorithm estimation was suitable for building 3D kernel PCA model with the better success rates.

Based on dataset B, it was found that the projection with polynomial (see (2), $d = 4$, $\theta = 1.5$) kernel function and class intensity model estimation was suitable for building 3D GDA model with the better success rates. Figure 4 illustrates the protein domains classes distribution of dataset B (498 samples) in 3D kernel space with GDA model. It can be seen that the data points, which belong to all-$\alpha$ domains, all-$\beta$ domains, $\alpha/\beta$ domains, and $\alpha+\beta$ domains respectively, are located in different regions with a correct classification result.

The success rates thus obtained are given in Table 1, where, for facilitating comparison, the corresponding rates obtained by component-coupled algorithm, neural networks, support vector machines (SVMs), and AdaBoost Learner [17] are also listed.

As it can be seen from Table 1, the performance of improved 3D kernel model outperforms those of component-coupled, neural networks, SVMs models but was a little worse than that of AdaBoost model for the dataset A (277 domains) available in LOOCV test. Based on dataset B (498 domains), improved 3D kernel learner is superior to all the other predictors in identifying the structural classification.

*3.3. Application of Improved 3D Kernel Approach to Classification of Membrane Proteins.* The membrane proteins dataset studied here was collected from the literature [25]. The dataset contains 2059 prokaryotic proteins (type A membrane proteins: 435; type B membrane proteins: 152; type C Multi-pass transmembrane proteins: 1311; type D lipid chain-anchored membrane proteins: 51; type E GPI-anchored membrane proteins: 110). The amino acid composition was selected as the input of the classification algorithm, and the computations were performed by LOOCV to test the power of various predictors. Based on dataset of membrane proteins, the classification flow chart (Figure 5) was obtained as follows.

From Figure 5, there are two steps in building classification model. Firstly, the 3D KPCA model with projection through polynomial (see (2), $d = 2$, $\theta = 0.1$) kernel function and KNN ($K = 5$) algorithm estimation was built to classify the multipass transmembrane proteins (type C) and the other membrane proteins (type A, type B, type D, and type E). Figure 6 illustrates the data distribution of type C and other membrane proteins in 3D kernel space with KPCA model.

Secondly, the 3D GDA model with Gaussian (see (3), $\sigma = 5$) kernel function and class intensity model estimation was built to classify type A, type B, type D, and type E membrane proteins.

Table 2: LOOCV success rates by covariant discriminant, neural network, SVM, bagging, and improved 3D kernel approach.

| Algorithm | Rate of correct prediction for each class | | | | | Overall rate of correct prediction |
|---|---|---|---|---|---|---|
| | Type A | Type B | Type C | Type D | Type E | |
| Covariant discriminant | 74.0% | 52.0% | 83.7% | 49% | 45.4% | 76.4% |
| Neural network | 75.63% | 30.92% | 88.86% | 50.98% | 30.91% | 77.76% |
| SVMs | 77.7% | 28.3% | 92.5% | 52.9% | 35.5% | 80.9% |
| Bagging | 79.80% | 48.68% | 93.21% | 49.02% | 60.91% | 84.18% |
| 3D kernel | 78.11% | 31.02% | 94.36% | 52.63% | 45.46% | 84.50% |

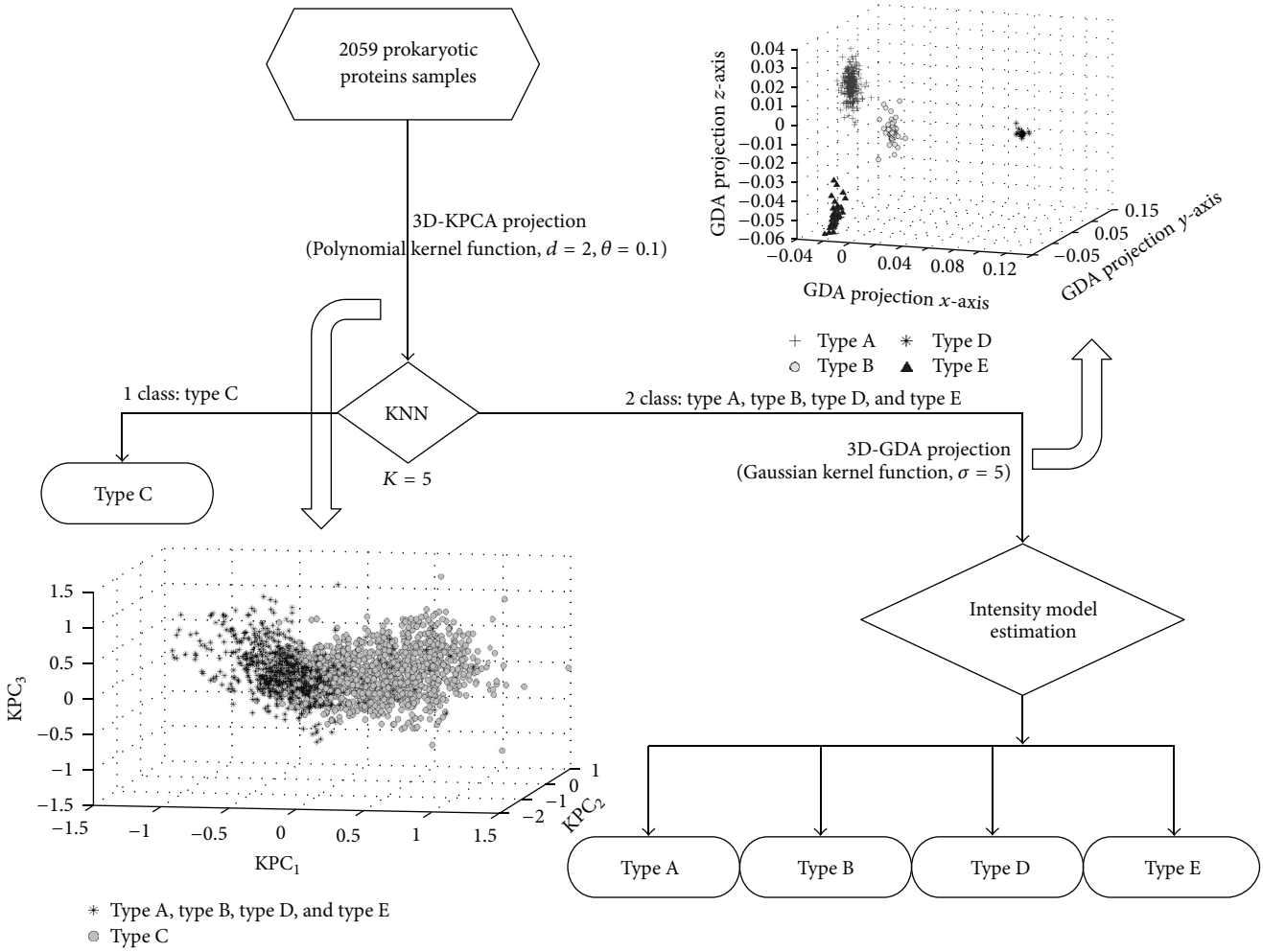


Figure 5: Classification flow chart of five type membrane proteins.

Figure 7 illustrates the data distribution of the type A, type B, type D, and type E membrane proteins in 3D kernel space with GDA model. 3D kernel method was compared with other machine learning classification methods: the covariant discriminant algorithm [23], neural networks, support vector machines, and Bagging [25], as is shown in Table 2.

As we can see from Table 2, correct classification rate of the LOOCV test applied 3D kernel algorithm outperformed other algorithms. It also means that 3D kernel method has learned very well through the membrane proteins training process.

## 4. Conclusions

The 3D kernel approach is very useful machine learning classifier. It has remarkably outperformed the powerful neural network, SVM classifiers, in predicting the protein domain structural classes for the two datasets constructed and membrane protein types for the same dataset constructed by previous investigators. It is thus anticipated that the 3D Kernel classifier can also be used to predict other protein attributes, such as sub-cellular localization [39–41], enzyme family and subfamily classes [42], and active sites of enzyme. The concepts of EQPC and IEFP can be easily extended to

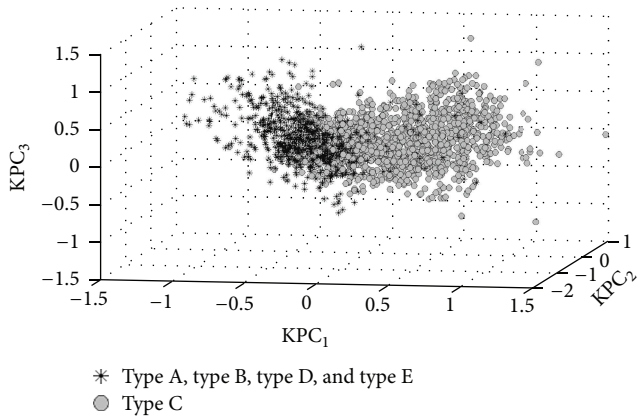


Figure 6: Distribution of the multipass transmembrane proteins (type C) and the other membrane proteins (type A, type B, type D and type E) data in 3D kernel space.

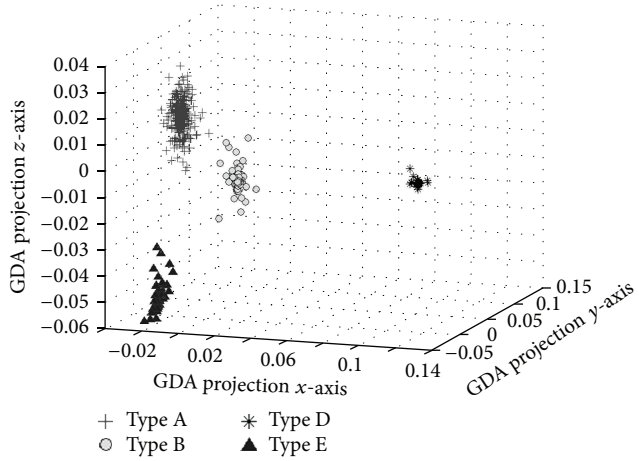


Figure 7: Distribution of the type A, type B, type D, and type E data in 3D kernel space.

many-dimensional space and could be improved to use four or more dimensions.

It could be concluded that 3D kernel approach is a robust and highly accurate classification technique that can be successfully applied to derive statistical models with statistical qualities and predictive capabilities for the protein location and function. The 3D kernel algorithm should be a complementary tool to the existing pattern recognition in chemometrics and bioinformatics.

## Authors' Contribution

Xu Liu and Yuchao Zhang contributed equally to this work.

## Acknowledgments

The project is financially supported by National Natural Science Foundation of China (nos. 20373040, 20973108, 20942005, and 21262005), Innovation Foundation of Guangxi University (nos. XBZ120947), and Innovation Foundation of Shanghai University (nos. A.10-0101-10-006). The work was supported by Guangxi Key Laboratory of Traditional Chinese Medicine Quality Standards (Guangxi Institute of Traditional Medical and Pharmaceutical Sciences) (guizhongzhongkai0802).

## References

[1] V. Brusic, G. Rudy, M. Honeyman, J. Hammer, and L. Harrison, "Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network," *Bioinformatics*, vol. 14, no. 2, pp. 121–130, 1998.

[2] L. Xu, L. Wencong, J. Shengli, L. Yawei, and C. Nianyi, "Support vector regression applied to materials optimization of sialon ceramics," *Chemometrics and Intelligent Laboratory Systems*, vol. 82, no. 1-2, pp. 8–14, 2006.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[4] B. M. Nicolaï, K. I. Theron, and J. Lammertyn, "Kernel PLS regression on wavelet transformed NIR spectra for prediction of sugar content of apple," *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 2, pp. 243–252, 2007.

[5] Y. Qu, B.-L. Adam, Y. Yasui et al., "Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients," *Clinical Chemistry*, vol. 48, no. 10, pp. 1835–1843, 2002.

[6] B. Niu, X.-C. Yuan, P. Roeper et al., "HIV-1 protease cleavage site prediction based on two-stage feature selection method," *Protein and Peptide Letters*, vol. 20, no. 3, pp. 290–298, 2013.

[7] B. Niu, Q. Su, X.-C. Yuan, W. Lu, and J. Ding, "QSAR study on 5-lipoxygenase inhibitors based on support vector machine," *Medicinal Chemistry*, vol. 8, no. 6, pp. 1108–1116, 2012.

[8] C.-R. Peng, W.-C. Lu, B. Niu, M.-J. Li, X.-Y. Yang, and M.-L. Wu, "Predicting the metabolic pathways of small molecules based on their physicochemical properties," *Protein & Peptide Letters*, vol. 19, pp. 1250–1256, 2012.

[9] Q. Su, W.-C. Lu, B. Niu, X. Liu, and T.-H. Gu, "Classification of the toxicity of some organic compounds to tadpoles (Rana Temporaria) through integrating multiple classifiers," *Molecular Informatics*, vol. 30, no. 8, pp. 672–675, 2011.

[10] B. Niu, W.-C. Lu, J. Ding et al., "Site of O-glycosylation prediction based on two stage feature selection," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 142–145, 2011.

[11] A. V. Finkelstein and O. B. Ptitsyn, "Why do globular proteins fit the limited set of foldin patterns?" *Progress in Biophysics and Molecular Biology*, vol. 50, no. 3, pp. 171–190, 1987.

[12] K.-C. Chou and L. Carlacci, "Energetic approach to the folding of $\alpha/\beta$ barrels," *Proteins: Structure, Function and Genetics*, vol. 9, no. 4, pp. 280–295, 1991.

[13] K.-C. Chou, "Progress in protein structural class prediction and its impact to bioinformatics and proteomics," *Current Protein & Peptide Science*, vol. 6, no. 5, pp. 423–436, 2005.

[14] K. Oxenoid and J. J. Chou, "The structure of phospholamban pentamer reveals a channel-like architecture in membranes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 31, pp. 10870–10875, 2005.

[15] J. S. Richardson, "$\beta$ sheet topology and the relatedness of proteins," *Nature*, vol. 268, no. 5620, pp. 495–500, 1977.

[16] O. B. Ptitsyn and A. V. Finkelstein, "Similarities of protein topologies: evolutionary divergence, functional convergence or

principles of folding?" *Quarterly Reviews of Biophysics*, vol. 13, no. 3, pp. 339–386, 1980.

[17] B. Niu, Y.-D. Cai, W.-C. Lu, G.-Z. Li, and K.-C. Chou, "Predicting protein structural class with AdaBoost Learner," *Protein and Peptide Letters*, vol. 13, no. 5, pp. 489–492, 2006.

[18] D. A. Doyle, J. M. Cabral, R. A. Pfuetzner et al., "The structure of the potassium channel: molecular basis of K+ conduction and selectivity," *Science*, vol. 280, no. 5360, pp. 69–77, 1998.

[19] J. R. Schnell and J. J. Chou, "Structure and mechanism of the M2 proton channel of influenza A virus," *Nature*, vol. 451, no. 7178, pp. 591–595, 2008.

[20] L. Stouffer Amanda, A. Rudresh, and S. David, "Structural basis for the function and inhibition of an influenza virus proton channel," *Nature*, vol. 451, pp. 596–599, 2008.

[21] M. D. Resh, "Myristylation and palmitylation of Src family members: the fats of the matter," *Cell*, vol. 76, no. 3, pp. 411–413, 1994.

[22] K.-C. Chou and D. W. Elrod, "Protein subcellular location prediction," *Protein Engineering*, vol. 12, no. 2, pp. 107–118, 1999.

[23] K.-C. Chou and D. W. Elrod, "Prediction of membrane protein types and subcellular locations," *Proteins*, vol. 34, pp. 137–153, 1999.

[24] K.-C. Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space," *Proteins: Structure, Function and Genetics*, vol. 21, no. 4, pp. 319–344, 1995.

[25] B. Niu, Y.-H. Jin, K.-Y. Feng et al., "Predicting membrane protein types with bagging learner," *Protein & Peptide Letters*, vol. 15, no. 6, pp. 590–594, 2008.

[26] V. Vapnik, *Statistical Learning Theory*, John Wiley & Johns, New York, NY, USA, 1998.

[27] D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte, and L. Kaufman, *Chemometrics: A Textbook*, Elsevier Science Publishers B.V., Amsterdam, The Netherlands, 1988.

[28] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no. 2, pp. 149–171, 1997.

[29] W. Wu, D. L. Massart, and S. de Jong, "The kernel PCA algorithms for wide data. Part I: theory and algorithms," *Chemometrics and Intelligent Laboratory Systems*, vol. 36, no. 2, pp. 165–172, 1997.

[30] D.-S. Cao, Y.-Z. Liang, Q.-S. Xu, Q.-N. Hu, L.-X. Zhang, and G.-H. Fu, "Exploring nonlinear relationships in chemical data using kernel-based methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 1, pp. 106–115, 2011.

[31] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[32] H. Yamamoto, H. Yamaji, Y. Abe et al., "Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables," *Chemometrics and Intelligent Laboratory Systems*, vol. 98, no. 2, pp. 136–142, 2009.

[33] H. Wang, Z. Hu, and Y. Zhao, "An efficient algorithm for generalized discriminant analysis using incomplete Cholesky decomposition," *Pattern Recognition Letters*, vol. 28, no. 2, pp. 254–259, 2007.

[34] B. S. Kim and S. B. Park, "A fast k nearest neighbor finding algorithm based on the ordered partition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 761–766, 1986.

[35] S. Sonnenburg, G. Rätsch, S. Henschel et al., "The Shogun machine learning toolbox," *The Journal of Machine Learning Research*, vol. 11, pp. 1799–1802, 2010.

[36] S. R. Amendolia, G. Cossu, M. L. Ganadu, B. Golosio, G. L. Masala, and G. M. Mura, "A comparative study of K-nearest neighbour, support vector machine and multi-layer perceptron for Thalassemia screening," *Chemometrics and Intelligent Laboratory Systems*, vol. 69, no. 1-2, pp. 13–20, 2003.

[37] M. Kearns and D. Ron, "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation," in *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pp. 152–162, ACM Press, July 1997.

[38] S. B. Holden, "PAC-like upper bounds for the sample complexity of leave-one-out cross-validation," in *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pp. 41–50, Desenzano del Garda, Italy, July 1996.

[39] G.-P. Zhou and K. Doctor, "Subcellular location prediction of apoptosis proteins," *Proteins: Structure, Function and Genetics*, vol. 50, no. 1, pp. 44–48, 2003.

[40] Y.-X. Pan, Z.-Z. Zhang, Z.-M. Guo, G.-Y. Feng, Z.-D. Huang, and L. He, "Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach," *Journal of Protein Chemistry*, vol. 22, no. 4, pp. 395–402, 2003.

[41] K.-C. Chou and Y.-D. Cai, "Predicting protein localization in budding yeast," *Bioinformatics*, vol. 21, no. 7, pp. 944–950, 2005.

[42] K.-C. Chou and Y.-D. Cai, "Predicting enzyme family class in a hybridization space," *Protein Science*, vol. 13, no. 11, pp. 2857–2863, 2004.