# Protein classification using texture descriptors extracted from the protein backbone image

Loris Nanni, Jian-Yu Shi, Sheryl Brahnam, Alessandra Lumini




# Protein classification using texture descriptors extracted from the protein backbone image

Loris Nanni, Jian-Yu Shi, Sheryl Brahnam, Alessandra Lumini

PII: S0022-5193(10)00143-8
DOI: doi:10.1016/j.jtbi.2010.03.020
Reference: YJTBI 5921

To appear in: *Journal of Theoretical Biology*

Received date: 7 December 2009
Revised date: 28 January 2010
Accepted date: 11 March 2010

Cite this article as: Loris Nanni, Jian-Yu Shi, Sheryl Brahnam and Alessandra Lumini, Protein classification using texture descriptors extracted from the protein backbone image, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.03.020

# Protein classification using texture descriptors extracted from the protein backbone image

Loris Nanni[1] Jian-Yu Shi[2] Sheryl Brahnam[3] Alessandra Lumini[1]

[1]DEIS, IEIIT—CNR, Università di Bologna, Viale Risorgimento 2, 40136 Bologna, Italy

loris.nanni@unibo.it; alessandra.lumini@unibo.it

[2] School of Life Science , School of Computer Science and Technology, Northwestern Polytechnical University, Xi'An, China jianyushi@nwpu.edu.cn

[3]Computer Information Systems, Missouri State University, 901 S. National, Springfield, MO 65804, USA sbrahnam@missouristate.edu



**Abstract**

In this work we propose a method for protein classification that combines different texture descriptors extracted from the 2-D distance matrix obtained from the 3-D tertiary structure of a given protein. Instead of considering all atoms in the protein, the distance matrix is calculated by considering only those atoms that belong to the protein backbone. The positive results reported in this paper offer further experimental confirmation that the distance matrix contains sufficient information for describing a protein. Moreover, we show that combining features extracted from the primary structure with features extracted from the distance matrix increases the performance of our classification system. We demonstrate this finding by comparing the performance of an ensemble of classifiers that uses the combined features. The classifiers used in our experiments are support vector machines and random subspace of support vector machines. The experimental results, validated using three different datasets (protein fold recognition, DNA-binding proteins recognition, biological processes and molecular functions recognition) along with different texture feature extraction methods (variants of local binary patterns, radon feature transform based approaches, and Haralick descriptors) demonstrate the effectiveness of the proposed approach. Particularly

interesting are the results in the classification of 27 types of structural properties: our proposed approach achieves significant improvement compared with other reported methods.

**Key Words:** protein classification; texture descriptors; primary structure**;** local binary patterns; Radon transform; Haralick features; support vector machines.

# 1 Introduction

An important area of research involves finding good methods for extracting a set of features from a protein (Chou & Zhang, 1995). There are two general views on how extraction should be accomplished. Based on the wide assumption that structural features are closely related to sequence composition (Krissinel, 2007; Bastolla et al., 2008), one popular approach, called the *indirect representation of protein spatial structure*, extracts features from a sequence to perform classification. Indirect representation can be organized mainly into two types: one based on the statistical analysis of amino acid residues (Shi et al., 2008) and the other based on amino acid indices (Cai et al., 2002; Shi et al., 2008). Probably the most famous method for extracting a set of features from the amino-acid sequence is Chou's pseudo amino acid (PseAA) composition approach, and several variants of this method have been proposed in the literature, including hydropathy scales (Chou, 2005), physicochemical distance (Chou, 2000), digital code (Gao et al., 2005), complexity factor (Xiao et al., 2005; Xiao et al., 2006b), digital signal (Xiao & Chou, 2007), Fourier low-frequency spectrum (Liu et al., 2005), cellular automata (Xiao et al., 2006a), and genetic programming (Nanni & Lumini, 2008b). As summarized in recent comprehensive reviews, (Chou & Shen, 2007; Chou, 2009), since the concept of pseudo amino acid composition (PseAAC) was proposed (Chou, 2001), it has provided a very flexible mathematical frame for investigators to incorporate their desired information into the representation of protein samples. According to its original definition, the PseAAC is actually formulated by a set of discrete numbers (Chou, 2001) as

long as it is different from the classical amino acid composition (AAC) and it is derived from a protein sequence that is able to harbor some sort of its sequence order and pattern information, or able to reflect some physicochemical and biochemical properties of the constituent amino acids. The PseAAC approach has also been widely used to deal with many protein-related problems and sequence-related systems (see, e.g., (Chen et al., 2009; Ding et al., 2009; Ding and Zhang, 2008; Fang et al., 2008; Georgiou et al., 2009; Gonzalez-Diaz et al., 2008; Jiang et al., 2008; Li and Li, 2008; Lin, 2008; Lin et al., 2008; Lin et al., 2009; Qiu et al., 2009; Zeng et al., 2009; Zhang and Fang, 2008; Zhang et al., 2008; Zhou et al., 2007) and a long list of PseAAC-related references cited in a recent review (Chou, 2009)). As summarized in (Chou, 2009), until now 16 different PseAAC modes have been used to represent the samples of proteins for predicting their attributes. Each of these modes has its own advantage and disadvantage. Other new developments for predicting various protein features based on sequence information are: (Lin et al., 2009; Xiao et al., 2008a; Xiao et al., 2008b; Xiao et al., 2009a; Xiao et al., 2009b; Xiao et al. 2009c).

In contrast to the indirect approach is the view that features should be extracted directly from an analysis of the protein's spatial structure. This direct approach to feature representation can be grouped into three general types: one based on the spatial atom distribution (Daras et al., 2006), a second on its topological structure (Anne, 2004), and a third on its geometrical shape (Sayre & Singh, 2008).

Generally, the indirect representation is lower in computational cost but provides a higher dimensional feature set whereas the direct representation is higher in computational cost but provides a lower dimensional feature set. While the lower computational cost involved in the indirect approach is desirable, the higher dimensional representation requires the application of the most advanced techniques in pattern recognition, for example, building ensembles of classifiers for improving the performance of stand-alone methods (Chou & Cai, 2006; Sarda et al, 2005; Nanni & Lumini, 2006; Nanni & Lumini, 2008).

In this paper we apply new pattern recognition techniques to the indirect representation of protein features by examining texture descriptors extracted from the 2-D distance matrix of different protein classification datasets. We validate the use of texture descriptors by testing many variants of state-of-the-art texture descriptors. In particular, we examine variants of the well-known local binary patterns. Our experiments show that the best performance is obtained when the idea of dominant local binary patterns (DLB) (Liao et al., 2009) is combined with local ternary patterns (LTP) (Tan & Triggs, 2007). With DLP, the most frequent rotation invariant patterns are selected. LTP is proposed for obtaining a noise robust texture descriptor. It is based on encoding the gray level difference *d* between a pixel **x** and its neighborhood **u** by 3 values.

A further improvement of performance is obtained by employing a "supervised" random subspace of classifiers where each bin of the histogram has a probability of belonging to a given subspace according to its occurrence frequencies in the training data (Nanni et al., 2009). Another approach for improving classification performance is to extract the texture descriptors not only from the whole distance matrix but also from some selected sub-windows.

In addition to investigating the performance of various texture descriptors, we build on the idea that descriptors based on different extraction notions give complementary information. We do this by combining a direct (Chou's amino acid) descriptor with an indirect representation (protein spatial structure features extracted from the distance matrix) using the sum rule. Our investigation shows that the classifiers trained with features extracted from the amino-acid sequence are partially independent to the classifiers trained with features extracted from the distance matrix. The experimental results show that the proposed ensemble of classifiers combining direct and indirect descriptors outperforms stand-alone approaches.

The fusion approaches are a well known technique for improving the performance of the stand-alone methods, remarkable successes of using fusion approaches are: (Chou and Shen, 2008b) protein subcellular location prediction, (Chou and Shen, 2008a) protease type prediction, (Chou and

Shen, 2009b; Shen et al., 2009) protein folding rate prediction, (Chou and Shen, 2007b) signal peptide prediction, (Chou and Shen, 2007a) membrane protein type prediction.

The remainder of this paper is organized as follows. In section 2, we introduce our proposed approach and the feature extraction methods investigated in this paper. In section 3, we report experimental results obtained on three benchmark databases. Finally, in section 4, we summarize results and draw a few conclusions.

## 2 Proposed approach

One objective in this work is to explore new methods for extracting features from the distance matrix that works well on several benchmark datasets. In addition, we want to investigate fusion using a standard method for extracting features from the primary sequence of the protein. The protein descriptor used in our experiments is Chou's well-known pseudo amino acid descriptor (Chou, 2005).

In (Shi & Zhang, 2009), the authors show that Haralick features and the Radon transform produce a good texture descriptor for the distance matrix. In this paper, we compare several variants of the LBP, which we use as a new texture descriptor. We demonstrate that it is well suited for the distance matrix.

To improve performance further, we use a random subspace of support vector machines as the classifier, and we extract the features not only from the whole distance matrix but also from the II, III, and IV quadrant of the whole image (see figure 1). For each image a different classifier is trained and the results are combined using the sum rule.

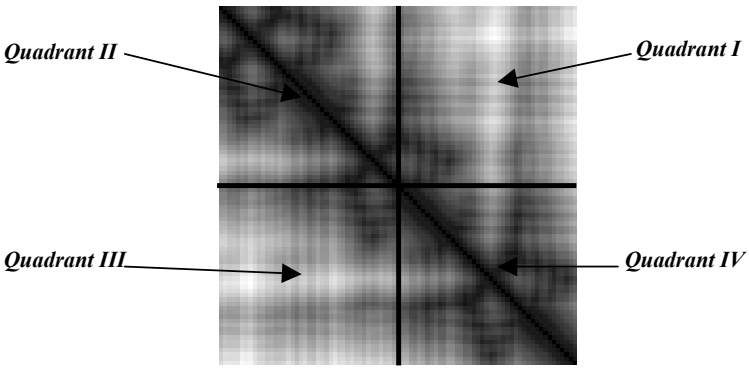



**Figure 1**. Subwindows of the whole distance matrix.

In addition we explore using a random subspace of support vector machine where the features are the following: dominant local ternary patterns (LTP), Haralick features, Radon transform, and discrete cosine transform (Radon+DCT), and Chou's amino acid sequence descriptor.

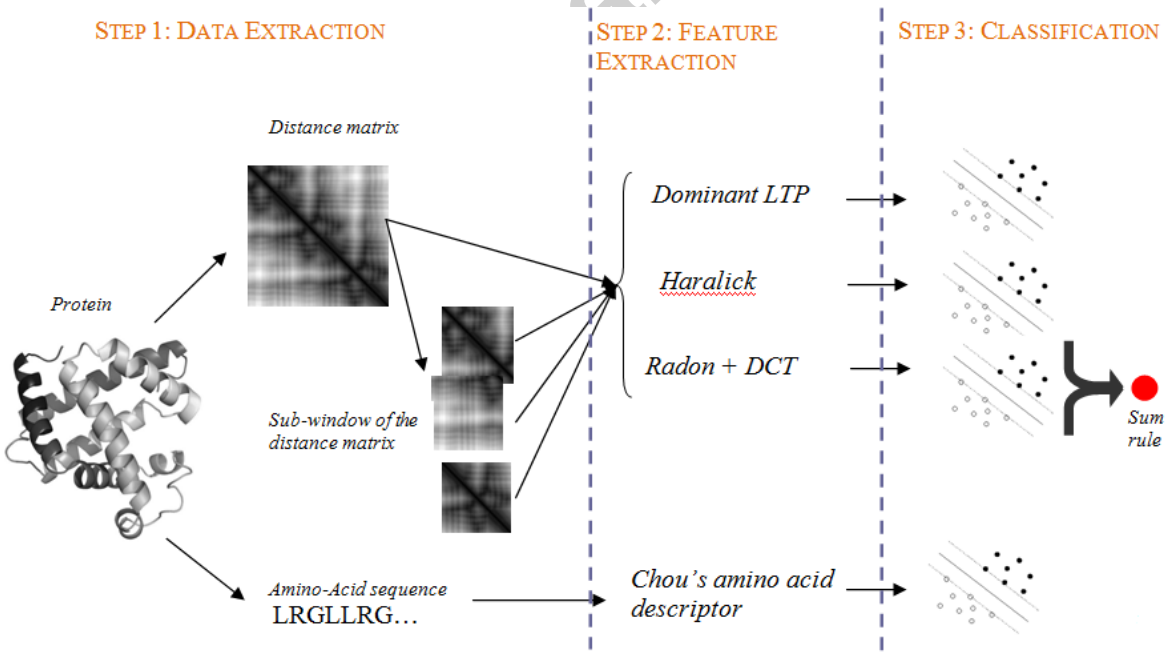



**Figure 2**. Proposed system for protein classification

The architecture of our best performing system (random subspace of support vector machines) is presented in figure 2.

A general description of each step in our classification experiments is provided below.

### 2.1 Extracting features from the distance matrix of the protein backbone

Protein is a composite of 20 types of amino acid residues. Various physicochemical properties and different counts and sequenced orders of these residues are the keys for deciding and producing the diversity of protein spatial structures. Unfortunately, how these keys work together is not fully understood. This fact brings out the difficulty of describing, analyzing, and characterizing protein structure. Instead of considering all atoms, many researchers use the $C_\alpha$ atoms of protein to characterize the whole protein structure. This *protein backbone* reflects the topology and the folding of protein (Taylor & Orengo, 1989). An effective representation of the backbone is the distance matrix (DM). It contains sufficient information of the proteins structure as the original 3D backbone structure can be reconstructed from DM using distance geometry methods (Timothy et al., 1983).

Given a protein $P_i$ as first step we need to extract its backbone: it is described as a vector $B_i = \left\{\mathbf{Coor}_{\alpha,1}^i, \mathbf{Coor}_{\alpha,2}^i, \cdots, \mathbf{Coor}_{\alpha,N}^i\right\}$, where $\mathbf{Coor}_{\alpha,n}^i$ is coordinates vector of the n$^{\text{th}}$ $C_\alpha$ atom. The distance matrix is defined as the matrix $DM = \left\{dm_i(p,q) = dist\left(\mathbf{Coor}_{\alpha,p}^i, \mathbf{Coor}_{\alpha,q}^i\right)\right\}$ where $dist(\cdot)$ is simply the Euclidean distance between the two set of coordinates (considered as a vector) and $1 \le p,q \le N$.[1]

Since DM maintains sufficient 3-D structural information, similar protein backbones are expected to have such distance matrices with similar properties. In our model, DM is regarded as a grayscale image. It is interesting to note that features extracted from DM are invariant to rotation and translation.

[1] The matlab code for extracting the distance matrix is available at http://bias.csr.unibo.it/nanni/DM.zip

### 2.2 Dominant Local ternary patterns

The basic idea of LBP is to examine the joint distribution of gray scale values of a circularly symmetric neighbor set of $P$ pixels around a pixel $\mathbf{x}$ on a circle of radius $R$. The LBP histogram of dimension $N$ (usually $N=P+2$) is obtained considering all the pixels of a given image. The difference $d$ between $\mathbf{x}$ and its neighborhood $\mathbf{u}$ is encoded by 2 values:

$$d = \begin{cases} 1 & \mathbf{u} \geq \mathbf{x} \\ 0 & \text{otherwise} \end{cases}$$

A pattern that contains at most two bitwise 0 to 1 or 1 to 0 transitions (circular binary code) is called uniform patterns (e.g., 11111111, 00000110 or 10000111 are all uniform patterns). In uniform LBP only these distributions are considered in the histogram (where a single bin contains all the non-uniform patterns).

In (Liao et al., 2009), better performance (compared with using the uniform patterns) is obtained when the patterns that represent $K$% ($K$=90% in their work) of the whole pattern occurrences in the training data are selected. To make LBP more robust to noise, (Tan & Triggs, 2007) used Local Ternary Patterns (LTP). In LTP the difference $d$ between $\mathbf{x}$ and its neighborhood $\mathbf{u}$ is encoded by 3 values according to a threshold $\tau$ (here $\tau$=3): 1 if $\mathbf{u} \geq \mathbf{x} + \tau$; -1 if $\mathbf{u} \leq \mathbf{x} - \tau$; else 0. The ternary pattern is then split into two binary patterns by considering its positive and negative components. The histograms that are computed from the binary patterns which are concatenated to form the feature vector.

In our experiments, we obtain the best performance by combining the idea of dominant LBP with LTP. For this purpose, we modify the original LBP code found at http://www.ee.oulu.fi/mvg/page/lbp_matlab[2].

[2] we have extracted the uniform patterns with *getmapping011*(16,'riu2') while for dominant LBP/LTP we use *getmapping011(16,'u2')*

Furthermore, when we use dominant LTP as feature extractor, we do not use the standard random subspace method but rather a "supervised" approach. The standard random subspace method (Ho, 1998) modifies the training dataset by generating $K$ (where $K$=100 in this paper) new training sets containing only a random subset of 60% of the features. Classifiers are then trained on these modified training sets. In "supervised" random subspace, each feature is not randomly selected but has a probability of being chosen that is based on the occurrence frequencies in the training data. Given $\mathbf{x}_i$, the sum of the occurrence frequency of the $i$-th bin of the histogram in the training data, the probability of the $i$-th bin to be chosen is given by $\mathbf{x}_i/\sum_i \mathbf{x}_i$. In "supervised" random subspace, the scores of the set of classifiers are combined using the sum rule.

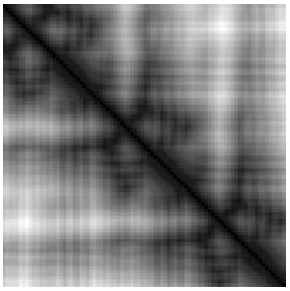
It should be noted that the "supervised" random subspace method is only used with the dominant LBP/LTP features. When random subspace is coupled with other texture descriptors, the standard random subspace is used.

**2.3 Radon transform**

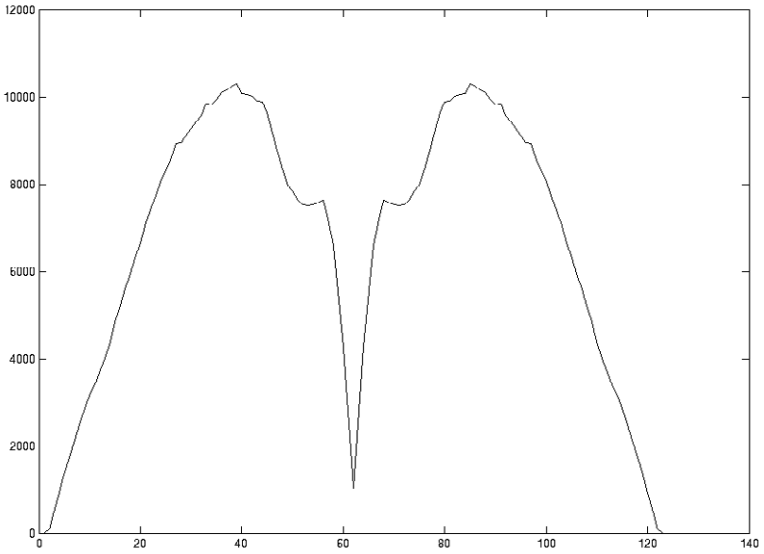
This feature extraction method is performed by selecting the first discrete cosine transform (duda et al., 2001) coefficients from the Radon Transform of the image (Kourosh & Hamid, 2005). The Radon transform is the projection of the image intensity along a radial line oriented at a specific angle (in this paper we use an angle of 45°). We have also tested other methods for selecting discrete cosine coefficients[3] (DCT), e.g., the coefficients with higher variance in the training set and the "discriminative power" method proposed in (Dabbaghchian & Ghaemmaghami, 2009). But in our tests, the best performance is obtained by selecting the first discrete cosine coefficients. In figure 3 an example of this feature extraction is reported.

[3] For both Radon and DCT we have used the official matlab function (radon.m and dct.m)

*Distance matrix*



*Radon transform for 45 degrees.*




*DCT transform of the Radon coefficients*

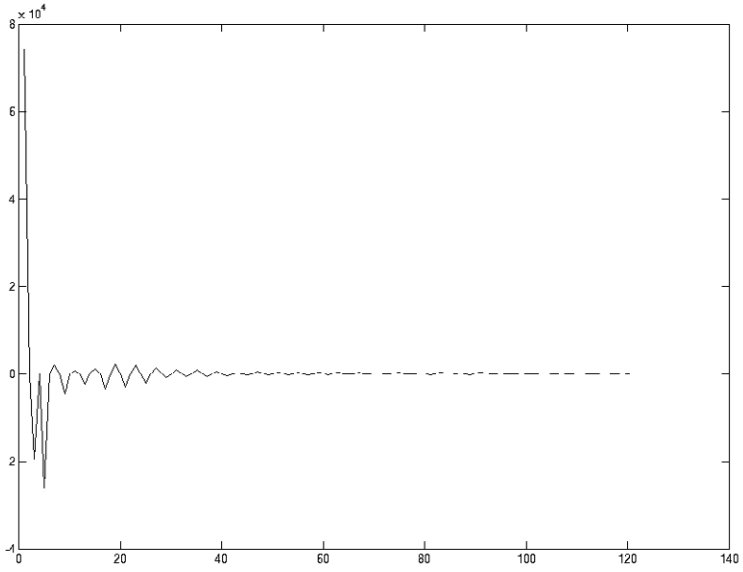



**Figure 3**. Radon+DCT feature extraction.

### 2.4 Haralick texture features

The Haralick texture features descriptor was proposed 30 years ago by (Haralick, 1979). It is based on the spatial gray level dependence matrices (SGLD), or co-occurrence matrix. Given an image with $N$ gray levels, the SGLD matrix at angle $\theta$ is a matrix of size $N \times N$. Each element in the matrix is a count of the total number of pairs of gray levels $i$ and $j$ at a distance $d$ along the direction $\theta$.

Thirteen features are calculated[4] from a SGLD matrix at a fixed angle $\theta$: energy, correlation, inertia, entropy, inverse difference moment, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy, and two information measures of correlation.

In this work we test two different feature sets extracted using the Haralick's method:

- *HARA1*, concatenation of the Haralick features extracted by considering two angles ( 0° and 90°), with $d$=1;
- *HARA3,* concatenation of the features extracted by considering four angles (0°, 45°, 135° and 90°[5]), with $d$=1;

### 2.5 Chou's pseudo amino acid composition (PseAA)

In (Zeng et al., 2009) a sequence-based algorithm combining the augmented Chou's pseudo amino acid composition based on auto covariance is presented. A set of pseudo amino acid based features are extracted from a given protein as the concatenation of the 20 standard amino acid composition values and $m$ (where $m$=20) values, reflecting the effect of sequence order (where $m$ is a parameter denoting the maximum distance between two considered amino acids $i$,$j$):

$$C_{20+l}^{d} = \sum_{k=1}^{Len-l} \frac{(Index(A(k),d) - M_d) \cdot (Index(A(k+l),d) - M_d)}{V_d \cdot (Len - l)} \quad l \epsilon [1..m]$$

[4] Implemented as in Haralick Texture Features Matlab Toolbox v0.1b shalinig@ece.utexas.edu - www.bme.utexas.edu/reasearch/informatics

[5] Notice that it is the same as 0 degree to subwindow II and IV

where $A(k)$ denotes the index of the amino acid in the $k^{th}$ position of the protein, *Len* is the length of the protein, $d$ denotes the selected physicochemical property, and the function $index(i,d)$ returns the value of the property $d$ for the amino acid $i$.

$M_d$ and $V_d$ are normalization factors denoting the average and the variance of the physicochemical property $d$ on the 20 amino acids:

$$M_d = \frac{1}{20}\sum_{i=1}^{20} index(i,d)$$

$$V_d = \frac{1}{20}\sum_{i=1}^{20} (index(i,d) - M_d)^2$$

We create 50 different Chou's pseudo amino acid feature vectors using 50 different psychochemical properties extracted from the AAindex (Kawashima & Kanehisa, 2000). For each Chou's pseudo amino acid feature vector a different support vector machine is trained. These 50 classifiers are then combined using the sum rule.



### 2.6 Classification system

For the classifier, we have used the well-known support vector machine (SVM) (Cristianini & Shawe-Taylor, 2000). It is a two class classifier which aims at finding the hyperplane that separates the training patterns of two classes by maximizing the distance between the hyperplane and the two classes. When it is not possible to find a linear decision boundary, a kernel function can be used to project the data onto a higher-dimensional feature space where a hyperplane separating the two classes can be found. Some typical kernels used in SVM include polynomial kernels and radial basis function kernels. It should be noted that the features used for training SVM are linearly normalized to [0 1]. In our work we use the OSU matlab toolbox (http://sourceforge.net/projects/svm/).

SVM-based machine learning algorithm was used in predicting protein subcellular location (Chou and Cai, 2002), membrane protein type (Cai et al., 2003a; Cai et al., 2004a), protein structural class (Cai et al., 2002a), specificity of GalNAc-transferase (Cai et al., 2002b), HIV protease cleavage sites in protein (Cai et al., 2002c), beta-turn types (Cai et al., 2002d), protein signal sequences and their cleavage sites (Cai et al., 2003b), alpha-turn types (Cai et al., 2003c), catalytic triads of serine hydrolases (Cai et al., 2004b), B-cell epitope prediction (Chen et al., 2007).

# 3 Datasets

All experiments were performed using the following datasets: Protein fold recognition, DNA-binding proteins, and the GO dataset. Each of these datasets and the testing protocols are briefly described in this section.

### 3.1 Protein fold recognition (FOLD)

The fold database used in our experiments is derived from the work of (Ding and Dubchak, 2001). It contains a training set and a testing set that contain 313 and 385 proteins. The sequence similarities are less 35% and 40% respectively, and the class numbers are both 27. The training set is used to build the classifier models, and we independently used the testing set to evaluate performance as this testing protocol is widely used in the literature for this dataset. The whole database can be downloaded from http://ranger.uta.edu/~chqding/protein/. Some sample of distance matrices extracted from the proteins of this dataset are reported in figure 4.

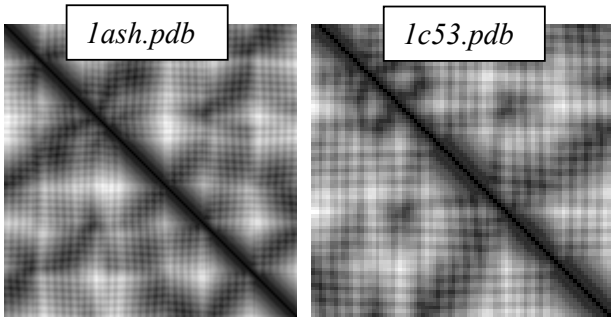
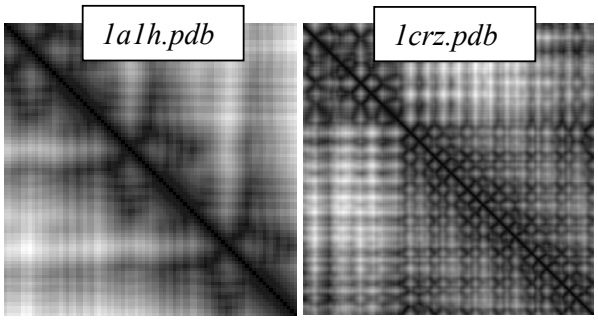



**Figure 4**. Examples of different classes of the FOLD dataset.

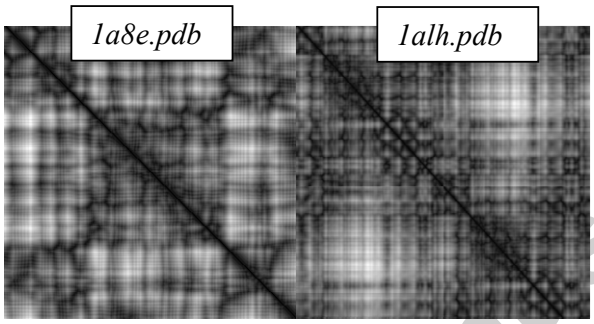
### 3.2 DNA-binding proteins (DNA)

This is the dataset is reported in (Fang et al., 2008) and contains 118 DNA-binding Proteins and 231 Non-DNA-binding proteins. These proteins have less than 35% sequence identity between each pair. DNA-binding proteins are proteins that are composed of DNA-binding domains and thus have a specific or general affinity for either single or double stranded DNA. Sequence-specific DNA-binding proteins generally interact with the major groove of B-DNA.

Some sample of distance matrices extracted from the proteins of this dataset are reported in figure 5. For this database we used the ten-fold cross validation protocol.




*a)* Binding class



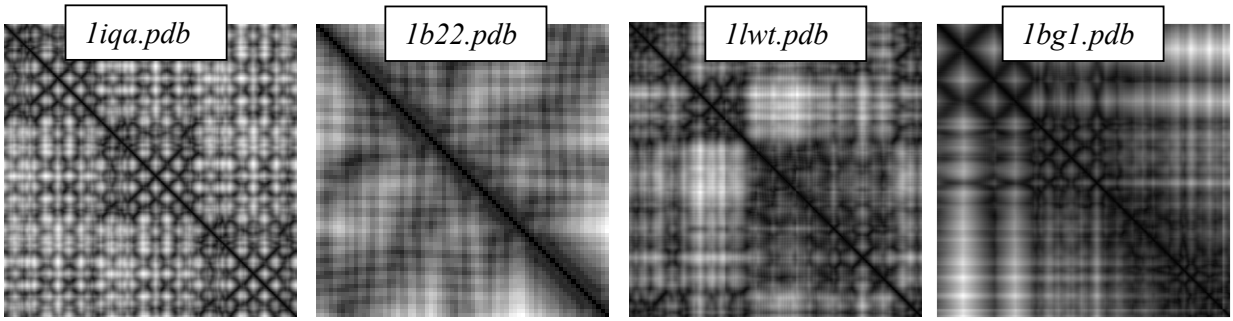

*b)* Non-binding

**Figure 5**. Examples of different classes of the BINDING dataset.

### 3.3 GO dataset (GO)

This dataset was reported in (Nanni et al., 2009b). It was created by collecting proteins according to GO annotations, distinguishing between the biological processes "immune response" (33 proteins), "DNA repair" (43 proteins), and between the molecular functions "substrate specific transporter activity" (39 proteins) and "signal transducer activity" (53 proteins). The presence of highly similar proteins in the same class was avoided by removing sequences having more than 30% identity. We randomly extracted 20% of the proteins for building the testing set, and this procedure was repeated 50 times. The results were then are averaged.

Some sample of distance matrices extracted from the proteins of this dataset are reported in figure 6.




*a*) immune *b*) DNA repair *c*) substrate *d*) signal

**Figure 6**. Examples of different classes of the GO dataset.



## 4 Experimental results

Performance is evaluated using the area under the Receiver Operating Characteristic (ROC) curve. The area under the ROC curve (AUC[6]) (Fawcett, 2004) is a scalar measure to evaluate performance which can be interpreted as the probability that the classifier will assign a higher score to a randomly picked positive sample than to a randomly picked negative sample. In the **GO** dataset, which is a four class problem, the AUC is calculated using the one versus all approach (a given class is considered as "positive" and all the other classes are considered as "negative") and the average AUC is reported.

In the following tests, Table 1, the AUC is reported as obtained using the following methods:

- (Shi & Zhang, 2009), the method reported in (Shi & Zhang, 2009), where few selected Radon features and Haralick feature are extracted from the distance matrix;
- PSEAA, the Chou's pseudo amino acid composition method explained in section 2.5;
- LBP, standard LBP with $P$=16 and $R$=2, the entire DM is used;
- LTP, standard LTP with $P$=16 and $R$=2, the entire DM is used;
- DLBP, dominant LBP with $P$=16, $R$=2 and $K$=90%, the entire DM is used;

[6] EUC is implemented as in dd_tools 0.95 davidt@ph.tn.tudelft.nl

- DLTP, dominant LTP with $P$=16, $R$=2 and $K$=90%, the entire DM is used;
- RS-DLTP, a random subspace of DLTP, the entire DM is used;
- MI DLTP, fusion of four DLTPs, one extracted from the entire DM and the others extracted from three sub-windows of the DM;
- MI(RS-DLTP), fusion of four RS-DLTPs, one extracted from the entire DM and the others extracted from three sub-windows of the DM;
- RADON+DCT, the descriptor described in section 3.3, the entire DM is used and the first 20 DCT coefficients are retained;
- RS-RADON+DCT, the random subspace version of RADON+DCT;
- MI(RS-RADON+DCT), fusion of four RS-RADON+DCT, one extracted from the entire DM and the others from three sub-windows of the DM. The first 20 DCT coefficients are retained when the DM is used but with the sub-windows the first 10 coefficients are retained;
- HARA1, the descriptor described in section 3.4, only the entire DM is used;
- HARA3, the descriptor described in section 3.4, only the entire DM is used;
- MI(HARA3), fusion of four HARA3, one extracted from the entire DM and the others from three sub-windows of the DM.

- In each cell of table 1 when the texture descriptors are used, there are two values: the first is the AUC obtained when the original distance matrix is used, and second, a value between parenthesis, is the AUC obtained when the matrix is resized to 100×100 before the feature extraction step. Also, for MI(RS-DLTP) and MI HARA3, we use the images resized to 100×100 before the feature extraction step, while for MI(RS-RADON+DCT) we use the original images.

| | | *DATASETS* | | |
|---|---|---|---|---|
| | | FOLD | BINDING | GO |
| *FEATURE EXTRACTION* | PseAA | 51.95 | 90.96 | 69.53 |
| | (Shi & Zhang, 2009) | 72.99 | 88.86 | 55.34 |
| | LBP | 56.88 (50.91) | 74.15 (73.82) | 56.9 (55.7) |
| | LTP | 61.56 (56.62) | 75.57 (72.20) | 60.2 (56.5) |
| | DLBP | 58.96 (55.84) | 69.47 (83.67) | 59.1 (59.3) |
| | DLTP | 68.57 (67.01) | 78.23 (83.90) | 61.7 (60.2) |
| | RS-DLTP | 69.87 (65.19) | 76.82 (76.91) | 60.5 (61.0) |
| | MI DLTP | 70.91 (69.87) | 79.53 (**84.38**) | 61.1 (61.1) |
| | **MI(RS-DLTP)** | 72.99 (**75.58**) | 75.72 (82.00) | **62.7** (62.2) |
| | RADON+DCT | 62.60 (52.21) | 73.79 (73.76) | 51.1 (51.1) |
| | RS-RADON+DCT | 65.45 (52.99) | 74.48 (78.79) | 53.5 (54.1) |
| | **MI(RS-RADON+DCT)** | **71.17** (65.97) | **79.29** (78.11) | **55.21** (53.9) |
| | HARA1 | 48.57 (50.39) | **88.88** (85.09) | 52.02 (**62.99**) |
| | HARA3 | 57.14 (58.70) | 88.21 (85.44) | 55.85 (61.05) |
| | **MI HARA3** | 70.91 (**71.69**) | 88.41 (87.32) | 61.73 (61.10) |

**Table 1**. Comparison among different methods.

From the results reported in table 1, the following conclusions can be drawn:

- Using the three sub-windows of the DM improves performance;
- The dominant LTP with the LBP/LTP texture descriptor works well in this particular application--the standard LBP/LTP obtains the worse performance;
- The features used in (Shi & Zhang, 2009), a set of selected Haralick features and moments extracted from the Radon coefficients, works particularly well in the BINDING dataset and well enough in the FOLD dataset. In these datasets the structural configuration of the proteins is very important. We want to stress the low performance of PSEAA in the FOLD dataset. The GO dataset the features of (Shi & Zhang, 2009) works poorly probably because of the low performance of the Radon based features in this difficult dataset.

Notice that in our work we do not perform any feature selection. We prefer to use longer feature vectors since the feature selection could be dataset dependent. In our work we want to propose a general method that works well on several datasets. If a training set is used to select a set of features for a particular dataset, the performance would likely be improved.

In table 2, we report some tests performed for optimizing the parameters of the texture descriptors, as follows:

- Different values for the parameters $\tau$ and $K$ of dominant LTP;
- MULTIRES, a combination between a dominant LTP with $P$=16 and $R$=2 and $P$=8 and $R$=1, both with $\tau$=3;
- R X-Y, is the MI(RS-RADON+DCT) method where the first X coefficients are extracted from the entire DM, and the first Y coefficients are extracted from the sub-windows of DM.
- MI HARA3 $d$=X, is the method MI HARA3 where we change the value of the parameter $d$.

| | | *DATASETS* | | |
|---|---|---|---|---|
| | | FOLD | BINDING | GO |
| *PARAMETERS* | τ=3 *K*=90% | 75.58 | 82.00 | 62.20 |
| | τ=1 *K*=90% | 70.17 | 79.67 | 59.24 |
| | τ=5 *K*=90% | 74.03 | 81.56 | 60.37 |
| | τ=3 *K*=80% | 72.99 | 77.80 | 60.52 |
| | τ=3 *K*=85% | 74.55 | 80.28 | 60.62 |
| | τ=3 *K*=95% | 74.81 | 83.75 | 62.07 |
| | MULTIRES | 73.25 | 80.78 | 61.31 |
| | R 25-10 | 71.17 | 79.29 | 55.21 |
| | R 40-20 | 70.65 | 81.06 | 59.95 |
| | R 15-5 | 66.23 | 80.27 | 51.60 |
| | MI HARA3 *d*=1 | 71.69 | 87.32 | 61.10 |
| | MI HARA3 *d*=2 | 72.47 | 85.25 | 63.64 |
| | MI HARA3 *d*=3 | 74.03 | 82.98 | 62.76 |

**Table 2**. Parameters optimization.

From the results reported in table 2, it is clear that the best configurations are R 40-20 for RADON+DCT, *d*=1 for MI HARA3, and τ=3 and *K*=90% for dominant LTP.

Finally, we report the results of the following fusions by sum rule (notice that before the fusion the scores of each classifier are normalized to a mean of 0 and a standard deviation of 1):

- FUS1, fusion by sum rule among MI(RS-DLTP), MI(RS-RADON+DCT) and MI HARA3;
- FUS2, fusion by weighted sum rule among MI(RS-DLTP), MI(RS-RADON+DCT) and MI HARA3. The weights of MI(RS-DLTP) and MI(RS-RADON+DCT) are 0.25, while the weight of MI HARA3 is 1;
- FUS3, fusion by sum rule among PSEAA, MI(RS-DLTP), MI(RS-RADON+DCT) and MI HARA3;
- FUS4, fusion by weighted sum rule among PSEAA, MI(RS-DLTP), MI(RS-Radon+Dct) and MI HARA3. The weight of PSEAA is 5, while the weights of the other methods are 1.

| Method | FOLD | BINDING | GO |
|---|---|---|---|
| FUS1 | 80.78 | 87.97 | 64.44 |
| FUS2 | 76.62 | 89.41 | 63.59 |
| FUS3 | **81.04** | 91.06 | 69.16 |
| **FUS4** | 64.68 | **92.16** | **71.53** |

**Table 3**. Fusion approaches.

Looking at table 3, it is clear that the best approach is different for the FOLD dataset and for the BINDING/GO dataset. This is due to the different performance of PSEAA. It works better than a texture descriptor in the general classification problems in the BINDING or GO dataset but works worst in the structural classification problem (where the features extracted from the distance matrix work very well). Obviously, in this particular problem, the structural classification using the distance matrix brings more information than the amino-acid sequence.

FUS3 obtains the best performance in the FOLD dataset, while in the BINDING/GO dataset the best performance is obtained by FUS4 (where PSEAA has an higher weight).

The results also suggest that some ad-hoc fusion methods should be studied independently in each dataset. For example, in the BINDING dataset fusion by sum rule between MI HARA3 and PSEAA (the two best descriptors in that dataset) obtains an AUC of 92.52. If in the BINDING dataset, we combine using sum rule MI HARA3, PSEAA and (Shi & Zhang, 2009), we obtain an an AUC of 93.58.

In order to validate the effectiveness of the presented method, we compare it with several methods in different literatures all of which used the same benchmark data sets and the same testing protocol of original paper (Ding & Dubchak, 2001), it is the dataset named FOLD in this work. The comparison is listed in Table 4. (Ding & Dubchak, 2001) proposed six kinds of features denoted by

C,S,H,P,V and Z respectively. Letter C is just the popular amino acid composition, while the left five letters indicate the features of Polarity, Polarizability, Normalized Van Der Waals volume, Hydrophobicity and Predicted secondary structure respectively. (Chinnasamy et al., 2005; Shi et al., 2006) are based on the same features CSHPVZ but different classifier systems. (Huang et al., 2003) combines CSHPVZ with bigram-coded feature (B) and spaced bigram-coded feature(SB), (Lin et al., 2007) does the same work as (Huang et al., 2003) but improves the classifier system by the technique of data fusion.

| *Method* | *Accuracy (%)* |
|---|---|
| (Ding & Dubchak, 2001) | 56.50 |
| (Chinnasamy et al., 2005) | 58.18 |
| (Shi et al., 2006) | 61.04 |
| (Huang et al., 2003) | 65.50 |
| (Lin et al., 2007) | 69.60 |
| (Shi & Zhang, 2009) | 72.99 |
| FUS3 | **81.04** |

**Table 4**. Other comparisons in the FOLD dataset.

The results reported in table 4 demonstrate that our proposed system (FUS3) outperforms other methods with the highest accuracy of classification.



## 4. Conclusion and Discussion

This paper focused on the study of texture descriptors for training an ensemble of machine learning algorithms for protein classification. The texture descriptors are extracted from the 2-D distance matrix obtained from the 3-D tertiary structure of a given protein.

Based on an analysis of prior research, we propose a new method based on a fusion of three texture descriptors and on the pseudo Chou's amino acid descriptor. Moreover, to improve the

performance, we extract the texture descriptors not only from the entire distance matrix but also from some selected sub-windows.

The ensemble proposed in this work has been tested on three datasets. The experimental results show that the proposed ensemble of classifiers outperforms stand-alone approaches. Particularly interesting are the results in the classification of 27 types of structural properties.

The best practical finding revealed in this work is that texture descriptors extracted from the 2-D distance matrix and amino acid descriptors should be combined to obtain a very reliable method of classification.

As further future work we want to test more texture descriptors, some preliminary tests have shown interesting results. We have tested: an holistic method based on the neighborhood preserving embedding method (NPE)[7] (He et al., 2005), which is a subspace learning algorithm aimed at preserving the global Euclidean structure of the space, obtaining a 40% accuracy in the FOLD dataset; a recent Gabor based descriptors (Guo et al., 2009) obtaining a 66% accuracy in the FOLD dataset. Both these results are obtained considering also the sub-windows of the distance matrix and the random subspace as classifier.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors (Chou and Shen, 2009a), we shall make efforts in our future work to provide a web-server for the method presented in this paper.



## References

Anne P. (2004): Voronoi and Voronoi-Related Tessellations in Studies of Protein Structure and Interaction. Current Opinion in Structural Biology, 14 (2),233 - 241

[7] The matlab code is available at http://www.cs.uiuc.edu/homes/dengcai2/Data/data.html (Accessed 15 July 2009)

Bastolla U., Ortíz A. R., Porto M., Teichert F. (2008): Effective Connectivity Profile: A Structural Representation That Evidences the Relationship between Protein Structures and Sequences. Proteins: Structure, Function, and Bioinformatics, 73(4),872-888

Cai Y. D., Liu X. J., Xu X. B., Chou K. C. (2002): Support Vector Machines for Prediction of Protein Subcellular Location by Incorporating Quasi-Sequence-Order Effect. Journal of Cellular Biochemistry, 84(2),343-348

Cai, Y.D., Zhou, G.P., and Chou, K.C., 2003a. Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84, 3257-3263.

Cai, Y.D., Lin, S., and Chou, K.C., 2003b. Support vector machines for prediction of protein signal sequences and their cleavage sites. Peptides 24, 159-161.

Cai, Y.D., Liu, X.J., Xu, X.B., and Chou, K.C., 2002a. Prediction of protein structural classes by support vector machines. Comput Chem 26, 293-296.

Cai, Y.D., Liu, X.J., Xu, X.B., and Chou, K.C., 2002b. Support vector machines for predicting the specificity of GalNAc-transferase. Peptides 23, 205-208.

Cai, Y.D., Liu, X.J., Xu, X.B., and Chou, K.C., 2002c. Support Vector Machines for predicting HIV protease cleavage sites in protein. J Comput Chem 23, 267-274.

Cai, Y.D., Liu, X.J., Xu, X.B., and Chou, K.C., 2002d. Support vector machines for the classification and prediction of beta-turn types. J Pept Sci 8, 297-301.

Cai, Y.D., Feng, K.Y., Li, Y.X., and Chou, K.C., 2003c. Support vector machine for predicting alpha-turn types. Peptides 24, 629-30.

Cai, Y.D., Pong-Wong, R., Feng, K., Jen, J.C.H., and Chou, K.C., 2004a. Application of SVM to predict membrane protein types. J Theor Biol 226, 373-376.

Cai, Y.D., Zhou, G.P., Jen, C.H., Lin, S.L., and Chou, K.C., 2004b. Identify catalytic triads of serine hydrolases by support vector machines. J Theor Biol 228, 551-557.

Chen, C., Chen, L., Zou, X., and Cai, P., 2009. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein & Peptide Letters 16, 27-31.

Chinnasamy, W. K. Sung and A. Mittal, "Protein structure and fold prediction using tree-augmented naive bayesian classifier," Journal of Bioinformatics and Computational Biology, vol. 3, pp.803–820, 2005.

Chou KC and Zhang CT (1995). Review: Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30: 275-349.

Chou KC (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10-19.

Chou KC (2000). Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochemical & Biophysical Research Communications 278: 477-483.

Chou, K.C., and Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277, 45765-45769.

Chou KC and Cai YD. (2006). Predicting protein-protein interactions from sequences in a hybridization space. Journal of Proteome Research, 5, 316-322.

Chou KC and Shen HB (2007). Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry 370: 1-16.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol.44, 60) 43, 246-255.

Chou, K.C., 2009. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics 6, 262-274.

Chou, K.C., and Shen, H.B., 2007a. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Comm 360, 339-345.

Chou, K.C., and Shen, H.B., 2007b. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm 357, 633-640.

Chou, K.C., and Shen, H.B., 2008a. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem. Biophys. Res. Comm. 376, 321-325.

Chou, K.C., and Shen, H.B., 2008b. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3, 153-162.

Chou, K.C., and Shen, H.B., 2009a. Review: recent advances in developing web-servers for predicting protein attributes. Natural Science 2, 63-92 (openly accessible at http://www.scirp.org/journal/NS/).

Chou, K.C., and Shen, H.B., 2009b. FoldRate: A web-server for predicting protein folding rates from primary sequence. The Open Bioinformatics Journal 3, 31-50 (openly accessible athttp://www.bentham.org/open/tobioij/).

Cristianini N, Shawe-Taylor J (2000). An introduction to Support vector machines and other kernel-based learning methods, Cambridge University Press.

Dabbaghchian S and Ghaemmaghami MP. (2009), Feature ex- traction using discrete cosine transform and discrimination power analysis with a face recognition technology, PatternRecognition, doi:10.1016/j.patcog.2009.11.001

Daras P., Zarpalas D., Axenopoulos A., Tzovaras D., Strintzis M. G. (2006): Three-Dimensional Shape-Structure Comparison Method for Protein Classification. IEEE Trans Comput Biol Bioinformatics, 3(3),193-207.

Ding CHQ and Dubchak I (2001), Multi-class protein fold recognition using support vector machines and neural networks, Bioinformatics, vol. 17, pp.349–358,.

Ding, H., Luo, L., and Lin, H., 2009. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein & Peptide Letters 16, 351-355.

Ding, Y.S., and Zhang, T.L., 2008. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recognition Letters 29, 1887-1892.

Duda R, Hart P, Stork D, (2001) Pattern Classification, Wiley, New York.

Fang Y, Guo Y, Feng Y, Li M (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34(1):103–109. doi:10.1007/s00726-007-0568-2

Fawcett T (2004). ROC Graphs: Notes and Practical Considerations for Researchers, Technical report, Palo Alto, USA: HP Laboratories.

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005). Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373-376.

Georgiou, D.N., Karakasidis, T.E., Nieto, J.J., and Torres, A., 2009. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J Theor Biol 257, 17-26.

Gonzalez-Diaz, H., Gonzalez-Diaz, Y., Santana, L., Ubeira, F.M., and Uriarte, E., 2008. Proteomics, networks, and connectivity indices. Proteomics 8, 750-778.

Yimo Guo, Guoying Zhao, Jie Chen, Matti Pietikäinen, Zhengguang Xu: A New Gabor Phase Difference Pattern for Face and Ear Recognition. CAIP 2009: 41-49

Haralick RM. (1979) Statistical and structural approaches to texture. Proceedings of the IEEE 67(5):768-804.

Huang C. D., Lin C. T. and Pal N. R., "Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification," IEEE Transactions on NanoBioscience, vol. 2, pp.221–232, 2003.

Ho TK (1998). The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell., 20(8) 832–844.

Jiang, X., Wei, R., Zhang, T.L., and Gu, Q., 2008. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein & Peptide Letters 15, 392-396.

Kawashima S, Kanehisa M (2000), AAindex: amino acid index database. Nucleic Acids Res., 28:374.

Krissinel E (2007) On the Relationship between Sequence and Structure Similarities in Proteomics Bioinformatics, 23(6),717-723

Kourosh JK and Hamid SZ (2005) Radon Transform Orientation Estimation for Rotation Invariant Texture Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, pp.1004-1008.

Li, F.M., and Li, Q.Z., 2008. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein & Peptide Letters 15, 612-616.

Liao, S., Law M.W.K., and Chung A.C.S. (2009) Dominant local binary patterns for texture classification. IEEE Transactions on Image Processing. 18(5): p. 1107 – 1118,.

Lin K. L., Lin C.Y., Huang C.D., Chang H. M., Yang C.Y., et al., "Feature selection and combination criteria for improving accuracy in protein structure prediction," IEEE Transactions on NanoBioscience, vol. 6, pp.186–196, 2007.

Lin, H., 2008. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252, 350-356.

Lin, H., Ding, H., Feng-Biao Guo, F.B., Zhang, A.Y., and Huang, J., 2008. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein & Peptide Letters 15, 739-744.

Lin, H., Wang, H., Ding, H., Chen, Y.L., and Li, Q.Z., 2009. Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition. Acta Biotheor 57, 321-330.

Liu H, Wang M and Chou KC (2005). Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336: 737-739.

Nanni, L; Lumini, A. (2006) MppS: an ensemble of Support Vector Machine based on multiple physicochemical properties of amino-acids, NeuroComputing, vol.69, no.13, pp.1688-1690, August 2006.

Nanni L, Lumini A (2008) A genetic approach for building different alphabets for peptide and protein classification, BMC Bioinformatics, 9:45.

Nanni L. and Lumini A. (2008b), Genetic programming for creating Chou's pseudoamino acid based features for submitochondria localization, Amino Acids, vol.34, no.4, pp.653-660.

Nanni L, Brahnam S, Lumini A (2009), Dominant Local Binary/Ternary Patterns and their weighted random subspace versions, submitted at CVPR 2010.

Nanni L, Mazzara S, Pattini L, Lumini A (2009b), Protein classification combining surface analysis and primary structure, Protein Engineering, Design and Selection, vol.22, no.4, pp.267-272, April 2009.

Qiu, J.D., Huang, J.H., Liang, R.P., and Lu, X.Q., 2009. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Analytical Biochemistry 390, 68-73.

Sayre T., Singh R.: Protein Structure Comparison and Alignment Using Residue Contexts. In: Advanced Information Networking and Applications - Workshops, 2008 AINAW 2008 22nd International Conference on 2008796-801. (2008)

Sarda D, Chua GH, Li K, Krishnan A (2005) , pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinformatics, 6:152

Shen, H.B., Song, J.N., and Chou, K.C., 2009. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. Journal of Biomedical Science and Engineering (JBiSE) 2, 136-143 (openly accessible at http://www.srpublishing.org/journal/jbise/). Biol 248, 546-551.

Shi J. Y., Pan Q., Zhang S. W., and Liang Y., "Protein fold recognition with support vector machines fusion network," Progress in Biochemistry and Biophysics, vol. 33, pp. 155–162, 2006

Shi J.-Y., Zhang S.-W., Pan Q., Zhou G.-P.: Using Pseudo Amino Acid Composition to Predict Protein Subcellular Location: Approached with Amino Acid Composition Distribution. Amino Acids2008, 35(2),321-327(2008)

Shi J-Y, Zhang Y-N (2009) Using Texture Descriptor and Radon Transform to Characterize Protein Structure and Build Fast Fold Recognition. International Association of Computer Science and Information Technology - Spring Conference, 2009. IACSITSC '09. 466-470.

Tan X, Triggs B (2007). Enhanced local texture feature sets for face recognition under difficult lighting conditions. In Analysis and Modelling of Faces and Gestures, volume 4778 of LNCS, pages 168-182. Springer, 2007.

Taylor WR and Orengo CA (1989), Protein structure alignment, Journal of Molecular Biology, vol. 208(1)pp.1– 22, 1989.

Timothy H, Irwin K and Gordon C (1983), The theory and practice of distance geometry," Bulletin of Mathematical Biology, vol. 45, pp.665–720.

Xiao X and Chou KC (2007). Digital coding of amino acids based on hydrophobic index. Protein & Peptide Letters 14: 871-875.

Xiao X, Shao S, Ding Y, Huang Z, Huang Y and Chou KC (2005). Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57-61.

Xiao X, Shao SH, Ding YS, Huang ZD and Chou KC (2006a). Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30: 49-54.

Xiao X, Shao SH, Huang ZD and Chou KC (2006b). Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. Journal of Computational Chemistry 27: 478-482.

Xiao, X., Lin, W.Z., and Chou, K.C., 2008a. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. J Comput Chem 29, 2018-2024.

Xiao, X., Wang, P., and Chou, K.C., 2008b. Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. J Theor Biol 254, 691-696.

Xiao, X., Wang, P., and Chou, K.C., 2009a. Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. J Appl Crystallogr 42, 169-173.

Xiao, X., Wang, P., and Chou, K.C., 2009b. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J Comput Chem 30, 1414-1423.

Xiao, X., Lin, W. Z., (2009c) Application of protein grey incidence degree measure to predict protein quaternary structural types. Amino Acids.37(4):741-749.

Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang (2005), Neighborhood Preserving Embedding, Tenth IEEE International Conference on Computer Vision (ICCV'2005), vol.2, pp.1208-1213.

Zeng, Y.H., Guo, Y.Z., Xiao, R.Q., Yang, L., Yu, L.Z., and Li, M.L., 2009. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J Theor Biol 259, 366-372.

Zhang, G.Y., and Fang, B.S., 2008. Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. J Theor Biol 253, 310-315.

Zhang, G.Y., Li, H.C., and Fang, B.S., 2008. Predicting lipase types by improved Chou's pseudo-amino acid composition. Protein & Peptide Letters 15, 1132-1137.

Zhou, X.B., Chen, C., Li, Z.C., and Zou, X.Y., 2007. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor