Local phase quantization texture descriptor for protein classification

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

[1]Computer Information Systems, Missouri State University, 901 S. National, Springfield, MO 65804, USA
sbrahnam@missouristate.edu
[2]DEIS, IEIIT—CNR, Università di Bologna, Viale Risorgimento 2, 40136 Bologna, Italy
loris.nanni@unibo.it; alessandra.lumini@unibo.it
[3] School of Life Science , School of Computer Science and Technology, Northwestern Polytechnical University, Xi'An, China jianyushi@nwpu.edu.cn

Abstract

In this work we propose a method for protein classification based on a texture descriptor, called local phase quantization that utilizes phase information computed from the image extracted from the 3-D tertiary structure of a given protein. To build this texture, the Euclidean distance is calculated between all the atoms that belong to the protein backbone. Moreover, we study classification fusion with a state-of-the-art method for describing the proteins: the Chou's pseudo amino acid descriptor. Our experiments show that the fusion between the two approaches improves the performance of Chou's pseudo amino acid descriptor. We use support vector machines as our base classifier. The effectiveness of our approach is demonstrated using four benchmark datasets (protein fold recognition, DNA-binding proteins recognition, biological processes and molecular functions recognition/enzyme classification).

Keywords: protein classification; texture descriptors; primary structure**;** local phase quantization; support vector machines.

1 Introduction

Finding effective feature extraction methods is still one of most important ongoing issues in protein classification[4]. There are two general views on how extraction should be accomplished: the indirect and direct methods. *Indirect representation of protein spatial structure*, is based on the widely held assumption that structural features are closely related to sequence composition [7, 8]. Thus this method extracts features from a sequence. Perhaps the most famous indirect representation is pseudo amino acid (PseAA) composition [10], with its many variants, see, for instance, [11-14]. In the direct approach feature extraction is accomplished via an analysis of the protein's spatial structure. The direct method of representation can be grouped into three general types: one based on the spatial atom distribution [15], a second on its topological structure [16], and a third on its geometrical shape [17].

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, see, e.g., [3, 18-20].

In this paper we apply a new pattern recognition techniques that combines an indirect (Chou's amino acid) descriptor with a direct representation (namely, protein spatial structure features extracted from the distance matrix). The experimental results show that combining direct and indirect descriptors using an ensemble of classifiers outperforms previous standalone approaches.

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

2 Proposed approach

In [9] the authors show that Haralick features and the Radon transform produce a good texture descriptor for the distance matrix of the protein backbone. The main aim of this work is to propose a single set of texture features that works well in this problem. The protein descriptor used in our experiments is Chou's well-known pseudo amino acid descriptor [11]. The architecture of our best performing system is presented in figure 2. A general description of each step in our approach is provided below.

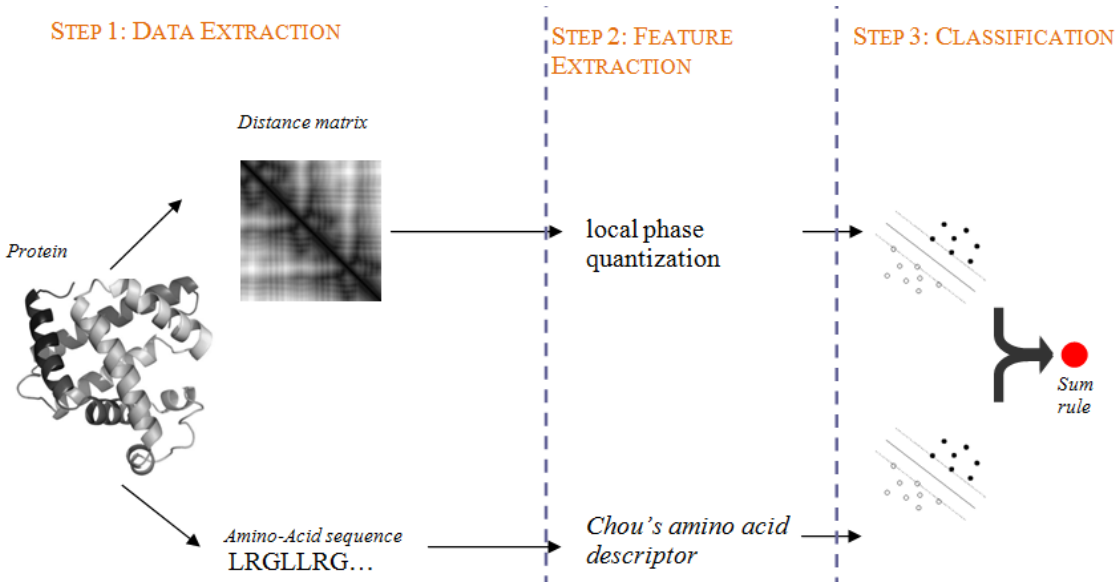

Figure 1. Proposed system for protein classification.

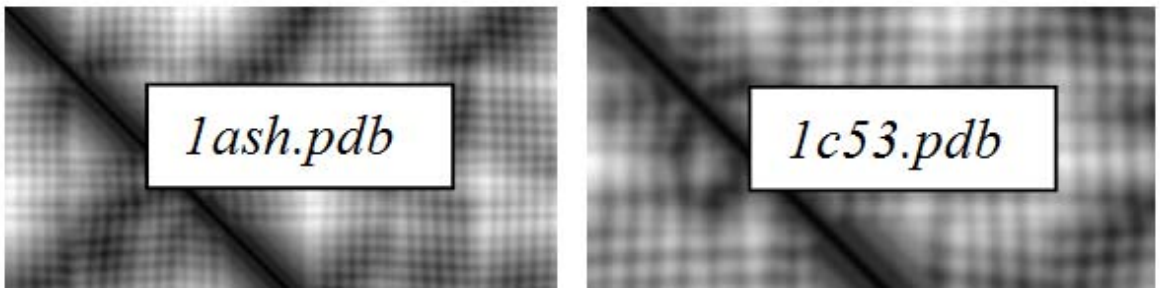

Figure 2. Examples of different classes of the BINDING dataset.

2.1 Data extraction

In step 1, we extract features from the distance matrix of the protein backbone. Diverse protein molecules differ in the number, type, physicochemical properties of amino acid residues, and their distribution along the polypeptide chain. These distinctions produce the diversity of protein spatial structures. Unfortunately, how these distinctions work together is not fully understood. This fact brings out the difficulty of describing, analyzing, and characterizing protein conformation.

Instead of considering all atoms, many researchers use the protein backbone to characterize the whole protein structure. Protein backbone is composed sequentially only by C_α atoms and reflects the topology and the folding of protein [21]. An effective representation of the backbone is the distance matrix (DM) which contains sufficient information of the proteins structure because the original 3D backbone structure can be reconstructed from DM using distance geometry methods [22].

Given a protein P_i, its backbone can be 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 nth C_α atom. Then its DM 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 from which the

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

extracted features are invariant to rotation and translation. We show an example of DM of two proteins in Figure 2.

2.2 Feature extraction

In step 2 we extract features using local phase quantization and Chou's pseudo amino acid composition (PseAA). Local phase quantization is applied to the DM. Chou's pseudo amino acid composition (PseAA) is created from the AAindex [23] as described below.

Local phase quantization

The Local Phase Quantization (LPQ) operator was originally proposed by Ojansivu and Heikkila as a texture descriptor [24]. LPQ is based on the blur invariance property of the Fourier phase spectrum. It uses the local phase information extracted using the 2-D short-term Fourier transform (STFT) computed over a rectangular neighborhood at each pixel position of the image. In LPQ only four complex coefficients are considered, corresponding to 2-D frequencies. For more mathematical details, refer to [24]. In our experiments, we use the original code shared by the inventors of LPQ. We use the images resized to 100×100 before the feature extraction step.

Chou's pseudo amino acid composition (PseAA)

In [25] a sequence-based algorithm is presented that combines the augmented Chou's pseudo amino acid composition based on auto covariance. 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 that reflect the effect of the sequence order: m is a parameter denoting the maximum distance between two considered amino acids i, j:

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

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:

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

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

In our experiments, we create 50 different Chou's pseudo amino acid feature vectors using 50 different psychochemical properties extracted from the AAindex [23]. A different support vector machine is trained for each Chou's pseudo amino acid feature vector. The 50 classifiers are then combined using the sum rule.

2.3 Classification

The classifier we use in our experiments is the well-known support vector machine (SVM) [26]. SVM finds the hyperplane that separates the training patterns of two classes by maximizing the distance between the hyperplane and the two classes. Where it is not possible to find a linear decision boundary, a kernel function can be used that projects the data onto a higher-dimensional feature space. Commonly used kernels include polynomial kernels and radial basis function kernels. All 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/).

3 Datasets

All experiments are performed using the following datasets: Protein fold recognition (FOLD), DNA-binding proteins (DNA), GO dataset (GO), and Enzyme (ENZ). Below we briefly describe the main characteristics of each and the evaluation protocol we use in the classification stage of our experiments.

Protein fold recognition (FOLD)

The FOLD database used in our experiments is derived from the work of [1]. The trainin set contains 313 proteins and testing set contains 385. 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 use the testing set to evaluate performance. This testing protocol is widely used in the literature for this dataset. The entire database can be downloaded from http://ranger.uta.edu/~chqding/protein/.

DNA-binding proteins (DNA)

The DNA dataset is reported in [27] 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. For this database we use the ten-fold cross validation protocol.

GO dataset (GO)

This dataset was reported in [28]. 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 extract 20% of the proteins for building the testing set, and this procedure is repeated 50 times. The results are then averaged.

Enzyme (ENZ)

This dataset was reported in [28]. The PDB archive was used to retrieve this dataset. It includes proteins annotated as enzymes: 381 hydrolases and 713 different enzymes. For this database we use the ten-fold cross validation protocol.

4 Experimental results

In our experiments, performance is evaluated using the area under the Receiver Operating Characteristic (ROC) curve. The area under the ROC curve (AUC[2]) [29] is a scalar measure for evaluating performance. AUC can be characterized as the probability that the classifier will assign a higher score to a randomly picked positive sample as compared to a randomly picked negative sample. Because the **GO** dataset is a four class problem, the AUC is calculated using the one versus all approach, i.e., a given class is considered as "positive" and all the other classes are considered as "negative." The average AUC is then reported.

In Table 1, the experimental results are reported (AUC) using the following methods:

- (Shi & Zhang, 2009): the method reported in [9], 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 2.3;
- LBP: standard Local Binary Pattern descriptor with16 neighborhoods and radius of the operator=2 [30];
- LTP: standard Local Ternary Pattern descriptor with16 neighborhoods and radius of the operator=2 [30]
- DLBP: dominant Local Binary Pattern descriptor with16 neighborhoods and radius of the operator=2 and the 90% of the bins are selected [31];
- LPQ(*x*): the local phase quantization descriptor with the radius of the operator equal at *x*;
- LPQ(1+2+3+4): a combination by sum rule of LPQ(1), LPQ(2), LPQ(3) and LPQ(4);
- LPQ(1+2+3+4)+*K*×PseAA: fusion by weighted sum rule between PseAA and LPQ(1+2+3+4). The weight of PseAA is *K*, while the weight of LPQ(1+2+3+4) is 1. Before the fusion the scores of the two methods are normalized to mean 0 and standard deviation 1.

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

- The LPQ texture descriptor works well in this particular application;
  The multiresolution approach based on the combination of different LPQs with different radii of the operator outperforms the stand-alone LPQ;
- The features used in [9], a set of selected Haralick features and moments extracted from the Radon coefficients, works particularly well in the BINDING dataset, but in the other three datasets LPQ outperforms [9].
- In each dataset the fusion approach outperform PseAA. Notice that in the FOLD dataset, where the structural classification problem is addressed, the features extracted from the distance matrix work very well. Obviously, in this particular problem, the structural classification of the distance matrix brings more information than the amino-acid sequence (since the protein backbone is very important in this task).

Contrarily to what was reported in [32], we have shown that a stand-alone texture descriptor can be used as an efficient feature extraction from the DM. LPQ extracts a reliable set of features when using the entire DM, and its fusion with a standard protein descriptor as PseAA creates a high performance system.

To validate the effectiveness of our proposed method, we compare it with several other methods reported in the literature using the FOLD dataset. We also employ the same testing protocols reported in the original paper [1]. The comparison results are listed in Table 2. In [1], the authors propose using six kinds of features denoted by C,S,H,P,V and Z. The 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. In [2, 3] the authors use the same set of features as [1], but they experiment with different classifier systems. We report the best performing system. In [5] CSHPVZ is combined with bigram-coded feature (B) and spaced bigram-coded feature (SB). In [6], the features used by [5] are used but the classifier system uses fusion. The results reported in table 2 demonstrate that our proposed system, specifically LPQ (1+2+3+4), outperforms other reported methods with the highest accuracy of classification.

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

		DATASETS			
		FOLD	*BINDING*	*GO*	*ENZYME*
FEATURE EXTRACTION	*PseAA*	*51.95*	*90.96*	*69.53*	*69.82*
	(Shi & Zhang, 2009)	*72.99*	*88.86*	*55.34*	*62.97*
	LBP	*50.91*	*73.82*	*55.70*	*61.10*
	LTP	*56.62*	*72.20*	*56.50*	*62.10*
	DLBP	*55.84*	*83.67*	*59.30*	*63.00*
	LPQ (1)	*77.14*	*85.58*	*67.18*	*57.69*
	LPQ (2)	*85.45*	*82.42*	*66.44*	*60.75*
	LPQ (3)	*84.68*	*82.09*	*67.63*	*65.38*
	LPQ (4)	85.45	81.34	68.67	63.86
	LPQ (1+2+3+4)	**87.27**	85.48	69.43	65.56
	LPQ (1+2+3+4) + 1×PseAA	85.19	91.02	**74.58**	72.21
	LPQ (1+2+3+4) + 2×PseAA	78.70	91.76	73.46	**72.81**
	LPQ (1+2+3+4) + 3×PseAA	73.25	**91.98**	72.79	72.34

Table 1. Comparison of the different methods.

Method	***Accuracy (%)***
(Ding & Dubchak, 2001) [1]	56.50
(Chinnasamy et al., 2005) [2]	58.18
(Shi et al., 2006) }[3]	61.04
(Huang et al., 2003) [5]	65.50
(Lin et al., 2007) [6]	69.60
(Shi & Zhang, 2009) [9]	72.99
Our LPQ (1+2+3+4) method	**87.27**

Table 2. Best reported results in the literature using the FOLD dataset.

4. Conclusion and Discussion

This reported a study of texture descriptors for training an ensemble of classifiers 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. Our method combines direct and indirect feature extraction methods by fusing the texture descriptors and the pseudo Chou's amino acid descriptor. The ensemble system proposed in this paper was tested on four datasets, and the experimental results show that our proposed method outperforms stand-alone approaches.

The best practical finding revealed in this work is that a combination of texture descriptors extracted from the 2-D distance matrix and amino acid descriptors boost performance in classifier systems.

In the future, we plan on experimenting with more texture descriptors. Preliminary experiments have demonstrated promising results [32]. In particular we are looking at using a holistic method based on the neighborhood preserving embedding method (NPE)[3] [33]. This is a subspace learning algorithm aimed at preserving the global Euclidean structure of the space. Thus far we have obtained 40% accuracy using the FOLD dataset. We are also examining a recently proposed Gabor based descriptor [34]. Thus far we have obtained 66% accuracy on the FOLD dataset. Both these results consider extracting features from subwindows of the distance matrix using random subspace as the classifier.

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

References

[1] Ding, C. H. Q., and Dubchak, I., "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, pp. 349-358, 2001.

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

[3] Shi, J. Y., Pan, Q., Zhang, S. W., and Liang, Y., "Protein fold recognition with support vector machines fusion networ," *Progress in Biochemistry and Biophysics*, vol. 33, pp. 155-162, 2006.

[4] Chou, K. C., and Zhang, C. T., "Review: Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, pp. 275-349, 1995.

[5] 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.

[6] Lin, C. C., Tsai, Y. S., Lin, Y. S., Chiu, T. Y., Hsiung, C. C., Lee, M. I., Simpson, J. C., and Hsu, C. N., "Boosting multiclass learning with repeating codes and weak detectors for protein subcellular localization," *Bioinformatics*, vol. 23, no. 24, pp. 3374–3381, 2007.

[7] Krissinel, E., "On the relationship between sequence and structure similarities in proteomics," *Bioinformatics*, vol. 23, no. 6, pp. 717-723, 2007.

[8] Bastolla, U., Ortíz, A. R., Porto, M., and Teichert, F., "Effective connectivity profile: A structural representation that evidences the relationship between protein structures and sequences," *Proteins: Structure, Function, and Bioinformatics*, vol. 73, no. 4, pp. 872-888, 2008.

[9] Shi, J. Y., and Zhang, Y. N., "Using texture descriptor and radon transform to characterize protein structure and build fast fold recognition," in International Association of Computer Science and Information Technology (IACSITSC '09), 2009, pp. 466-470.

[10] Chou, K. C., "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology," *Current Proteomics*, vol. 6, pp. 262-274, 2009.

[11] Chou, K. C., "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, pp. 10-19, 2005.

[12] Shi, J. Y., Zhang, S. W., Pan, Q., Cheng, Y. M., and Xie, J., "Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition," *Amino Acids*, vol. 33, no. 1, pp. 69-74, 2007.

[13] Shi, J. Y., Zhang, S. W., Pan, Q., and Zhou, G. P., "Using pseudo amino acid composition to predict protein subcellular location: Approached with amino acid composition distribution," *Amino Acids*, vol. 35, no. 2, pp. 321-327, 2008.

[14] Nanni, L., and Lumini, A., "Genetic programming for creating Chou's pseudoamino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653-660, 2008.

[15] Daras, P., Zarpalas, D., Axenopoulos, A., Tzovaras, D., and Strintzis, M. G., "Three-dimensional shape-structure comparison method for protein classification," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 3, no. 3, pp. 193-207, 2006.

[16] Anne, P., "Voronoi and voronoi-related tessellations in studies of protein structure and interaction," *Current Opinion in Structural Biology*, vol. 14, no. 2, pp. 233 - 241, 2004.

[17] Sayre, T., and Singh, R., "Protein structure comparison and alignment using residue contexts," in Advanced Information Networking and Applications - Workshops, AINAW 2008 22nd International Conference, 2008, pp. 796-801.

[18] Chou, K. C., and Cai, Y. D., "Predicting protein-protein interactions from sequences in a hybridization space," *Journal of Proteome Research*, vol. 5, pp. 316-322, 2006.

[19] Nanni, L., and Lumini, A., "MppS: An ensemble of support vector machines based on multiple physicochemical properties of amino-acids," *Neurocomputing*, vol. 69, no. 13, pp. 1688-1690, 2006.

[20] Nanni, L., and Lumini, A., "A genetic approach for building different alphabets for peptide and protein classification," *BMC Bioinformatics*, vol. 9, no. 45, 2008.

[21] Taylor, W. R., and Orengo, C. A., "Protein structure alignment," *Journal of Molecular Biology*, vol. 208, no. 1, pp. 1- 22, 1989.

[22] Timothy, H., Irwin, K., and Gordon, C., "The theory and practice of distance geometry," *Bulletin of Mathematical Biology*, vol. 45, pp. 665-720, 1983.

[23] Kawashima, S., and Kanehisa, M., "AAindex: amino acid index database," *Nucleic Acids Research*, vol. 20, no. 1, 374, 2000.

[24] Ojansivu, V., and Heikkila, J., "Blur insensitive texture classification using local phase quantization," in ICISP, 2008.

[25] Zeng, Y. H., Guo, Y. Z., Xiao, R. Q., Yang, L., Yu, L. Z., and Li, M. L., "Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach," *Journal of Theoretical Biology*, vol. 259, no. 2, pp. 366-72, 2009.

[26] Cristianini, N., and Shawe-Taylor, J., *An introduction to support vector machines and other kernel-based learning methods*, Cambridge, UK: Cambridge University Press, 2000.

[27] Fang, Y., Guo, Y., Feng, Y., and Li, M., "Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features," *Amino Acids*, vol. 34, no. 1, pp. 103-109, 2008.

[28] Nanni, L., Mazzara, S., Pattini, L., and Lumini, A., "Protein classification combining surface analysis and primary structure," *Protein Engineering, Design and Selection*, vol. 22, no. 4, pp. 267-272, 2009.

[29] Fawcett, T., *ROC graphs: Notes and practical considerations for researchers*, HP Laboratories, Palo Alto, USA, 2004.

[30] Tan, X., and Triggs, B., "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *Analysis and Modelling of Faces and Gestures*, vol. LNCS 4778, pp. 168-182, 2007.

[31] Liao, S., Law, M. W. K., and Chung, A. C. S., "Dominant local binary patterns for texture classification," *IEEE Transactions on Image Processing*, vol. 18, no. 5, pp. 1107 – 1118, 2009.

[32] Nanni, L., Shi, J.-Y., Brahnam, S., and Lumini, A., "Protein classification using texture descriptors extracted from the protein backbone image," *Journal of Theoretical Biology*, In Press.

[33] He, X., Cai, D., Yan, S., and Zhang, H.-J., "Neighborhood preserving embedding," in Tenth IEEE International Conference on Computer Vision (ICCV'2005), 2005.

[34] Guo, Y., Zhao, G., Chen, J., Pietikäinen, M., and Xu, Z., "A New Gabor Phase Difference Pattern For Face And Ear Recognition," in CAIP, 2009, pp. 41-49.