Efficient Design of Bio-Basis Function to Predict Protein Functional Sites Using Kernel-Based Classifiers

Pradipta Maji* and Chandra Das

***Abstract*—In order to apply the powerful kernel-based pattern recognition algorithms such as support vector machines to predict functional sites in proteins, amino acids need encoding prior to input. In this regard, a new string kernel function, termed as the *modified bio-basis function*, is proposed that maps a nonnumerical sequence space to a numerical feature space. The proposed string kernel function is developed based on the conventional bio-basis function and needs a bio-basis string as a support like conventional kernel function. The concept of *zone of influence* of a bio-basis string is introduced in the proposed kernel function to take into account the influence of each bio-basis string in nonnumerical sequence space. An efficient method is described to select a set of bio-basis strings for the proposed kernel function, integrating the Fisher ratio and a novel concept of *degree of resemblance*. The integration enables the method to select a reduced set of relevant and nonredundant bio-basis strings.**

***Index Terms*—Bioinformatics, functional site prediction, pattern recognition, sequence analysis, support vector machines.**

I. Introduction

RECENT advancement and wide use of high-throughput technology for biological research are producing an enormous amount of biological data. The successful analysis of biological data has become critical. Although laboratory experiment is the most effective method to analyze the biological data, it is very financially expensive and labor intensive. Pattern recognition techniques and machine learning methods provide useful tools for analyzing the biological data [1]–[4].

The prediction of functional sites in proteins is an important issue in protein function studies and hence drug design [5], [6]. The problem of protein functional sites prediction deals with the protein subsequences. The subsequences are obtained from a whole protein sequence through moving a fixed length sliding window residue by residue. The residues within a scan form a subsequence. A functional site is said to present within a subsequence, if there exists a match between the subsequence and a consensus pattern of a specific function; and the subsequence is labeled as functional, otherwise it is labeled as nonfunctional. Therefore, the goal of this problem is to classify a subsequence whether it is functional or nonfunctional [6].

Manuscript received June 24, 2009; revised August 26, 2010; accepted September 16, 2010. Date of publication September 30, 2010; date of current version February 02, 2011. *Asterisk indicates corresponding author.*

*P. Maji is with the Machine Intelligence Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata, 700 108, India (e-mail: pmaji@isical.ac.in).

C. Das is with the Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, 700 152, India.

In classification, the main objective is to train a model based on the labeled data. The trained model is then used for classifying novel data. Classification analysis requires two descriptions of an object: one is the set of features that are used as inputs to train the model and the other is the class label. Classification analysis aims to find a mapping function from the features to the class label.

To analyze protein sequences or subsequences, BLAST [7], suffix-tree based algorithms [8], regular expression matching representations [9], and finite state machines [10], [11] are a few of the many pattern recognition algorithms that use characters or strings as their primitive type. However, some other pattern recognition algorithms such as artificial neural networks trained with back-propagation [3], [12], [13], Kohonen's self-organizing map [1], feed-forward and recurrent neural networks [4], [5], bio-basis function neural networks [14]–[19], and support vector machines [6], [20], [21] work with numerical features to predict different functional sites in proteins such as protease cleavage sites of the human immunodeficiency virus (HIV) and the hepatitis C virus [13], linkage sites of glycoprotein [3], [19], enzyme active sites [22], posttranslational phosphorylation sites [15], immunological domains [23], Trypsin cleavage sites [16], protein–protein interaction sites [21], and so forth. Hence, in order to apply the powerful kernel-based pattern recognition algorithms such as support vector machines to predict functional sites in proteins, the protein subsequences therefore have to be encoded prior to input. The objective of coding biological information in subsequences is to provide a method for converting nonnumerical attributes in subsequences to numerical features.

There are two main methods for coding a subsequence: distributed encoding [12] and the bio-basis function method [14]–[16]. In the distributed encoding method, each of 20 amino acids is encoded using a 20-bit binary vector [12]. In effect, a subsequence with m amino acids or residues is converted into a binary string of length $20m$. Hence, in this method, the input space for modeling is expanded unnecessarily [13]. Moreover, the use of the Euclidean distance may not be able to encode biological content in sequences efficiently [13].

In this background, the concept of bio-basis function has been proposed in [14]–[16] for encoding subsequences. The bio-basis function is a string kernel function that takes an input subsequence and a reference string as its two arguments; the reference string is termed as the bio-basis string. A set of bio-basis functions transforms a nonnumerical subsequence to a numerical feature vector. Transformation of an input subsequence to a numerical feature vector is performed based on the similarity of the input subsequence and a set of bio-basis strings. The similarity

TABLE I
DAYHOFF MATRIX: 1 POINT MUTATION IS ACCEPTED PER 100 RESIDUES

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	40	24	32	32	16	36	28	28	28	24	28	32	36	32	24	36	36	32	8	20
C	24	88	12	12	16	20	20	24	12	8	12	16	20	12	16	32	24	24	0	32
D	32	12	48	44	8	36	36	24	32	16	20	40	28	40	28	32	32	24	4	16
E	32	12	44	48	12	32	36	24	32	20	24	36	28	40	28	32	32	24	4	16
F	16	16	8	12	68	12	24	36	12	40	32	16	12	12	16	20	20	28	32	60
G	36	20	36	32	12	52	24	20	24	16	20	32	28	28	20	36	32	28	4	12
H	28	20	36	36	24	24	56	24	32	24	24	40	32	44	40	28	28	24	20	32
I	28	24	24	24	36	20	24	52	24	40	40	24	24	24	24	28	32	48	12	28
K	28	12	32	32	12	24	32	24	52	20	32	36	28	36	44	32	32	24	20	16
L	24	8	16	20	40	16	24	40	20	56	48	20	20	24	20	20	24	40	24	28
M	28	12	20	24	32	20	24	40	32	48	56	24	24	28	32	24	28	40	16	24
N	32	16	40	36	16	32	40	24	36	20	24	40	28	36	32	36	32	24	16	24
P	36	20	28	28	12	28	32	24	28	20	24	28	56	32	32	36	32	28	8	12
Q	32	12	40	40	12	28	44	24	36	24	28	36	32	48	36	28	28	24	12	16
R	24	16	28	28	16	20	40	24	44	20	32	32	32	36	56	32	28	24	40	16
S	36	32	32	32	20	36	28	28	32	20	24	36	36	28	32	40	36	28	24	20
T	36	24	32	32	20	32	28	32	32	24	28	32	32	28	28	36	44	32	12	20
V	32	24	24	24	28	28	24	48	24	40	40	24	28	24	24	28	32	48	8	24
W	8	0	4	4	32	4	20	12	20	24	16	16	8	12	40	24	12	8	100	32
Y	20	32	16	16	60	12	32	28	16	28	24	24	12	16	16	20	20	24	32	72

is calculated using an amino acid mutation matrix. The bio-basis strings are also the subsequences of the protein sequence that are used to transform all subsequences into numerical feature vectors with dimension equal to the number of bio-basis strings. The bio-basis function has been successfully applied to predict different functional sites in proteins [14]–[19].

The most important issue for bio-basis function is how to select a reduced set of most relevant and nonredundant bio-basis strings. Berry *et al.* [15] used the Fisher ratio for selection of bio-basis strings. Yang and Thomson [16] proposed a method to select bio-basis strings using mutual information. In principle, the bio-basis strings in nonnumerical sequence space should be such that the degree of resemblance between pairs of bio-basis strings would be as minimum as possible. Each of them would then represent a unique feature in numerical feature space. However, the methods proposed in [15] and [16] have not adequately addressed this problem. Also, it has not been paid much attention earlier. Moreover, the bio-basis function proposed in [14]–[16] does not take into account the impact or influence of each bio-basis string in nonnumerical sequence space.

In this paper, a new string kernel function, termed as the *modified bio-basis function*, is proposed that modifies existing bio-basis function. The concept of *zone of influence* of the bio-basis string is incorporated in the proposed kernel function to take into account the influence or impact of each bio-basis string in nonnumerical sequence space. An efficient method is presented, integrating the Fisher ratio and the novel concept of *degree of resemblance*, to select most relevant and distinct bio-basis strings for the proposed string kernel functions. Instead of using symmetric similarity measure as in [15], the asymmetric biological dissimilarity is used to calculate the Fisher ratio, which is shown to be more effective for selection of most relevant bio-basis strings. The *degree of resemblance* enables efficient selection of a set of distinct bio-basis strings. In effect, it reduces the redundant features in numerical feature space.

The structure of the rest of this paper is as follows: Section II briefly introduces necessary notions of existing bio-basis function, and related bio-basis string selection methods proposed by Berry *et al.* [15] and Yang and Thomson [16]. In Section III, the *modified bio-basis function* is presented, while an efficient bio-basis string selection method is proposed in Section IV. Concluding remarks are given in Section V.

II. BIO-BASIS FUNCTION AND SELECTION METHODS

In this section, the basic notion in the theory of bio-basis function and the bio-basis string selection methods of Yang and Thomson [16] and Berry *et al.* [15] are reported.

A widely used method in sequence analysis is the sequence alignment [7], [24]. In this method, the function of a sequence is annotated through aligning a novel sequence with known sequences. If the alignment between a novel sequence and a known sequence gives a very high similarity or homology score, the novel sequence is believed to have the same or similar function as the known sequence. In this method, an amino acid mutation matrix is commonly used. Each mutation matrix has 20 columns and 20 rows. A value at the nth row and mth column is a probability or a likelihood value that the nth amino acid mutates to the mth amino acid after a particular evolutionary time [25], [26]. The mutation probabilities as similarities among amino acids are therefore metrics. The Dayhoff matrix (Table I) was the first mutation matrix developed in 1978 [27] and many new mutation matrices were developed later on, for instance, the Blosum62 matrix [25]. However, the above method may not be useful directly for subsequence analysis. Because a subsequence may not contain enough information for conventional alignment.

To alleviate this problem, the concept of bio-basis function is introduced in [14]–[16] for subsequence analysis. It is based on the principle of conventional alignment technique. Using a table look-up technique, a homology score as a similarity value can be obtained for a pair of subsequences. The nongapped pairwise alignment technique is used to calculate this similarity value, where no deletion or insertion is used to align two subsequences [14]–[16]. For ease of subsequent discussions, the following terminology is used in rest of the paper.

- $\Omega = \{A, C, \ldots, W, Y\}$ be the set of 20 amino acids.
- n represents the total number of subsequences with m amino acids or residues obtained from a whole amino acid sequence of length L through moving a sliding window of size m residue by residue. Hence, $n = L - m + 1$.

- $X = \{x_1, \ldots, x_j, \ldots, x_n\}$ be the set of n subsequences with m residues, $\forall j, x_j \subset \mathcal{A}^m$.
- c represents the total number of bio-basis strings.
- $V = \{v_1, \ldots, v_i, \ldots, v_c\}$ be the set of c bio-basis strings and $\forall i, v_i \in X$.
- $x_j[k] \in \mathcal{A}, v_i[k] \in \mathcal{A}, \forall_{k=1}^{m}$.

A. Bio-Basis Function

The definition of bio-basis function is as follows [14], [16]:

$$f(x_j, v_i) = \exp\left\{\gamma_b \frac{h(x_j, v_i) \quad h(v_i, v_i)}{h(v_i, v_i)}\right\} \tag{1}$$

- $h(x_j, v_i)$ is the homology score between a subsequence x_j and a bio-basis string v_i;
- $h(v_i, v_i)$ denotes the maximum homology score of the ith bio-basis string v_i; and
- γ_b is a constant and typically chosen to be 1 [14], [15].

Suppose both x_j and v_i have m residues, the homology score between x_j and v_i is then defined as

$$h(x_j, v_i) = \sum_{k=1}^{m} \mathbb{M}(x_j[k], v_i[k]) \tag{2}$$

where $\mathbb{M}(x_j[k], v_i[k])$ can be obtained from an amino acid mutation matrix through a table look-up method. Note that $x_j[k], v_i[k] \in \mathcal{A}$ and $\mathcal{A}$ is a set of 20 amino acids.

Consider two bio-basis strings $v_1 = \mathrm{KPRT}$ and $v_2 = \mathrm{YKAE}$, and a subsequence $x_1 = \mathrm{IPRS}$ having $m = 4$ residues. The nongapped pairwise homology score is calculated between the subsequence x_1 and each bio-basis string considering the mutation probabilities as in Table I. For first bio-basis string v_1, four mutation probabilities are

$$\begin{aligned}
&\mathbb{M}(x_1[1], v_1[1]) = \mathbb{M}(\mathrm{I}, \mathrm{K}) = 24;\\
&\mathbb{M}(x_1[2], v_1[2]) = \mathbb{M}(\mathrm{P}, \mathrm{P}) = 56;\\
&\mathbb{M}(x_1[3], v_1[3]) = \mathbb{M}(\mathrm{R}, \mathrm{R}) = 56;\\
&\mathbb{M}(x_1[4], v_1[4]) = \mathbb{M}(\mathrm{S}, \mathrm{T}) = 36.
\end{aligned}$$

Hence, the homology score between the subsequence x_1 and the bio-basis string v_1 is given by

$$h(x_1, v_1) = \sum_{k=1}^{4} \mathbb{M}(x_1[k], v_1[k]) = 172.$$

Similarly, the value of $h(x_1, v_2)$ between the subsequence x_1 and the bio-basis string v_2 is as follows:

$$h(x_1, v_2) = \sum_{k=1}^{4} \mathbb{M}(x_1[k], v_2[k]) = 112.$$

The maximum homology scores of two bio-basis strings v_1 and v_2 are given by

$$h(v_1, v_1) = 208 \quad \text{and} \quad h(v_2, v_2) = 212.$$

Considering the value of $\gamma_b = 1$

$$f(x_1, v_1) \quad \exp\left\{\gamma_b \frac{h(x_1, v_1) - h(v_1, v_1)}{h(v_1, v_1)}\right\} \quad 0.633334;$$

$$f(x_1, v_2) = \exp\left\{\gamma_b \frac{h(x_1, v_2) - h(v_2, v_2)}{h(v_2, v_2)}\right\} = 0.287988;$$

$$f(v_2, v_2) \quad \exp\left\{\gamma_b \frac{h(v_2, v_2) - h(v_2, v_2)}{h(v_2, v_2)}\right\} \quad 1.000000.$$

Hence, the value of bio-basis function $f(x_i, x_j)$ is high if two subsequences x_i and x_j are similar or close to each other. The function value is one if two subsequences are identical, and small if they are distinct. The function needs a subsequence as a support, which is termed as the bio-basis string. Each bio-basis string is a feature dimension in a numerical feature space. If $\mathcal{A}$ is used to denote a collection of 20 amino acids, an input space of all potential subsequences with m residues is $\mathcal{A}^m$. Then, a collection of c bio-basis strings formulates a numerical feature space $\Re^c$, to which a nonnumerical sequence space $\mathcal{A}^m$ is mapped for analysis, that is, $\mathcal{A}^m \to \Re^c$. The bio-basis function can transform various homology scores to a real number as a similarity within the interval $[0, 1]$, that is,

$$0 \le f(x_j, v_i) \le 1. \tag{3}$$

B. Bio-Basis String Selection Using Mutual Information

In [16], Yang and Thomson proposed a method for bio-basis string selection using mutual information [28]. The necessity for a bio-basis string to be an independent and informative feature can be determined by the shared information between the bio-basis string and the rest as well as the shared information between the bio-basis string and class label [16].

The mutual information is quantified as the difference between the initial uncertainty and the conditional uncertainty. Let $\Phi = \{v_i\}$ be a set of selected bio-basis strings, $\Theta = \{v_k\}$ a set of candidate bio-basis strings. $\Phi = \phi$ (empty) at the beginning. A prior probability of a bio-basis string v_k is referred as $\mathrm{p}(v_k)$. The initial uncertainty of v_k is defined as

$$H(v_k) = -\mathrm{p}(v_k) \ln \mathrm{p}(v_k). \tag{4}$$

Similarly, the joint entropy of two bio-basis strings v_k and v_i is given by

$$H(v_k, v_i) = -\mathrm{p}(v_k, v_i) \ln \mathrm{p}(v_k, v_i) \tag{5}$$

where $v_i \in \Phi$ and $v_k \in \Theta$. The mutual information between v_k and v_i is therefore given by

$$\begin{aligned}
I(v_k, v_i) &= H(v_k) + H(v_i) - H(v_k, v_i)\\
&\{-\mathrm{p}(v_k) \ln \mathrm{p}(v_k) - \mathrm{p}(v_i) \ln \mathrm{p}(v_i)\\
&- \mathrm{p}(v_k, v_i) \ln \mathrm{p}(v_k, v_i)\}.
\end{aligned} \tag{6}$$

However, the mutual information of v_k with respect to all the bio-basis strings in Φ is

$$I(v_k, \Phi) = \sum_{v_i \in \Phi} I(v_k, v_i). \tag{7}$$

Combining (6) and (7), we get [16]

$$I(v_k, \Phi) = \sum_{v_i \in \Phi} \mathrm{p}(v_k, v_i) \ln \left\{\frac{\mathrm{p}(v_k, v_i)}{\mathrm{p}(v_k)\mathrm{p}(v_i)}\right\}. \tag{8}$$

Replacing Φ with the class label $\Omega = \{\Omega_1, \ldots, \Omega_j, \ldots, \Omega_M\}$, the mutual information

$$I(v_k, \Omega) \quad \sum_{\Omega_j \in \Omega} \mathrm{p}(v_k, \Omega_j) \ln \left\{ \frac{\mathrm{p}(v_k, \Omega_j)}{\mathrm{p}(v_k)\mathrm{p}(\Omega_j)} \right\} \tag{9}$$

measures the mutual relationship between v_k and Ω. A bio-basis string whose $I(v_k, \Omega)$ value is the largest will be selected as v_k and will make the largest contribution to modeling (discrimination using Ω) among all the remaining bio-basis strings in Θ. Therefore, there are two mutual information measurements for v_k, the mutual information between v_k and $\Omega(I(v_k, \Omega))$ and the mutual information between v_k and $\Phi(I(v_k, \Phi))$. In this method, the following criterion is used for the selection of bio-basis strings [16], [29]

$$J(v_k) = \alpha_{\mathrm{YT}} I(v_k, \Omega) \quad (1 \quad \alpha_{\mathrm{YT}}) I(v_k, \Phi) \tag{10}$$

where α_{YT} is a constant. In the current study, the value of α_{YT} is set at 0.7 to give more weightage in discrimination [16], [29]. The major drawback of the method proposed by Yang and Thomson [16] is that a huge number of prior and joint probabilities are to be calculated, which makes the method computationally expensive.

C. Selection of Bio-Basis Strings Using Fisher Ratio

In [15], Berry *et al.* proposed a method to select a set $V = \{v_1, \ldots, v_i, \ldots, v_c\}$ of c bio-basis strings from the whole set $X = \{x_1, \ldots, x_j, \ldots, x_n\}$ of n subsequences based on their discriminant capability. The discriminant capability of each subsequence x_i is calculated using the Fisher ratio [30] that follows next:

$$\mathrm{F}(x_i) = \frac{|\mu_{A_i} - \mu_{B_i}}{\sqrt{\sigma^2_{A_i} + \sigma^2_{B_i}}} \tag{11}$$

where μ_{A_i} and σ_{A_i} denote the mean and standard deviation of similarity values of subsequences presented in group X_A with respect to the subsequence x_i.

The n subsequences of X would have different compositions of amino acids. Hence, they should have different pairwise alignment scores with the other subsequences of X. As the class labels of these training subsequences are known, these subsequences can be partitioned into two groups or classes (functional and nonfunctional), which are denoted as $X_A \subset X$ and $X_B \subset X$, respectively. Denoting the similarity between two subsequences x_i and x_j as $h(x_j, x_i)$, the mean and standard deviation values for these two groups with respect to the subsequence x_i can be written as

$$\mu_{A_i} = E_A[h(x_j, x_i)] = \frac{1}{n_A} \sum h(x_j, x_i) \quad \forall x_j \in X_A \tag{12}$$

$$\mu_{B_i} \quad E_B\, h(x_k, x_i) \quad \frac{1}{n_B} \sum h(x_k, x_i) \quad \forall x_k \in X_B \tag{13}$$

$$\begin{aligned} \sigma^2_{A_i} &= E_A[h^2(x_j, x_i)] \quad [E_A[h(x_j, x_i)]]^2 \\ &= \frac{1}{n_A} \sum \{h(x_j, x_i) - \mu_{A_i}\}^2; \quad \forall x_j \in X_A \end{aligned} \tag{14}$$

$$\begin{aligned} \sigma^2_{B_i} &= E_B[h^2(x_k, x_i)] \quad [E_B[h(x_k, x_i)]]^2 \\ &= \frac{1}{n_B} \sum \{h(x_k, x_i) - \mu_{B_i}\}^2; \quad \forall x_k \in X_B \end{aligned} \tag{15}$$

where n_A, n_B are the number of subsequences in X_A and X_B, respectively. $E[h(x_j, x_i)]$ and $E[h^2(x_k, x_i)]$ represent the zero-mean, first-, and second-order moment of similarity, that is, expectation of $h(x_j, x_i)$ and $h^2(x_j, x_i)$, respectively. Hence, the numerator and denominator of (11) can be written as

$$|\mu_{A_i} - \mu_{B_i}| \quad |E_A[h(x_j, x_i)] - E_B\, h(x_k, x_i)|; \tag{16}$$

$$\begin{aligned} \sigma^2_{A_i} + \sigma^2_{B_i} = &\{E_A[h^2(x_j, x_i)] + E_B[h^2(x_k, x_i)]\} \\ &- \{[E_A[h(x_j, x_i)]]^2 + [E_B[h(x_k, x_i)]]^2\}. \end{aligned} \tag{17}$$

The Fisher ratio is used to maximize the discriminant capability of a subsequence in terms of interclass separation (as large as possible) and intraclass spread between subsequences (as small as possible). The larger the Fisher ratio value, the larger the discriminant capability of the subsequence. Based on the values of the Fisher ratio, n subsequences of X can be ranked from the strongest discriminant capability to the weakest one. The method yields a set V of c subsequences from X as the bio-basis strings which possess good discriminant capability between two classes, having evolved from original data set. The basic steps of this method follows next.

1) Calculate the discriminant capabilities of all subsequences of X using the Fisher ratio as in (11).
2) Rank all subsequences of X based on the values of Fisher ratio in descending order.
3) Select first c subsequences from X as the set V of bio-basis strings.

Step 3 of this algorithm is not necessarily optimal. It selects c subsequences as the bio-basis strings based on their discriminant capabilities without considering similarity among them. However, the bio-basis strings in nonnumerical sequence space should be such that the similarity between pairs of bio-basis strings would be as minimum as possible. Each of them would then represent a unique feature in numerical feature space. The methods proposed in [15] and [16] have not adequately addressed this problem. Also, not much attention has been paid to it earlier.

III. Proposed String Kernel Function

In this section, a new string kernel function is presented based on the concepts of biological dissimilarity and *zone of influence* of the bio-basis string.

A. Asymmetricity of Biological Dissimilarity

Here, two asymmetric dissimilarities are defined between two subsequences x_i and x_j as follows:

$$\begin{aligned} d_{x_i \to x_j} &= d(x_j, x_i) = \{h(x_i, x_i) - h(x_j, x_i)\} \\ d_{x_j \to x_i} &= d(x_i, x_j) = \{h(x_j, x_j) - h(x_i, x_j)\} \end{aligned} \tag{18}$$

where $d_{x_j \to x_i}$ denotes the dissimilarity of subsequence x_j from the subsequence x_i and $h(x_i, x_j) = h(x_j, x_i)$ is the nongapped pairwise homology score between x_i and x_j.

Consider two subsequences $x_i = \mathrm{KPRT}$ and $x_j = \mathrm{YKAE}$ with four residues. According to the Dayhoff mutation matrix (Table I), the nongapped pairwise homology score between two subsequences x_i and x_j is therefore $h(x_i, x_j) = h(x_j, x_i) = 100$, while the maximum homology scores of two subsequences x_i and x_j are given by $h(x_i, x_i) = 208$ and $h(x_j, x_j) = 212$,

respectively. Hence, the dissimilarity of subsequence x_j from the subsequence x_i is given by

$$d(x_j, x_i) \quad \{h(x_i, x_i) - h(x_j, x_i)\} \quad 208 - 100 \quad 108$$

whereas the dissimilarity of x_i from x_j is as follows:

$$d(x_i, x_j) \quad \{h(x_j, x_j) - h(x_i, x_j)\} \quad 212 - 100 \quad 112.$$

Thus, the dissimilarity is asymmetric in nature, that is,

$$d(x_j, x_i) \neq d(x_i, x_j). \tag{19}$$

The asymmetricity reflects domain organizations of two subsequences x_i and x_j. When two subsequences x_i and x_j consist of the same single domain, $d(x_j, x_i)$ and $d(x_i, x_j)$ will be similar small values. However, suppose that x_i has one extra domain, then $d(x_j, x_i)$ becomes large even if $d(x_i, x_j)$ is small. These dissimilarities may be used for clustering of protein sequences or subsequences so that domain organizations are well reflected. The asymmetric property of the biological dissimilarity was also observed by Stojmirovic [31] and Itoh *et al.* [32]. The asymmetric dissimilarity might be a powerful tool to cluster sequences or subsequences and to explore the gene or protein universe.

B. Modified Bio-Basis Function

The design of new string kernel function is based on the principle of asymmetric biological dissimilarity. Using a table look-up technique, a biological dissimilarity is calculated for a pair of subsequences based on an amino acid mutation matrix. The nongapped pairwise alignment method is used to calculate this dissimilarity, where no deletion or insertion is used to align two subsequences. The definition of the *modified bio-basis function* is as follows:

$$f_{\text{modified}}(x_j, v_i) = \exp\left\{\gamma_{mb}\frac{h(x_j, v_i) - h(v_i, v_i)}{\eta_i}\right\}$$

$$\text{that is, } f_{\text{modified}}(x_j, v_i) = \exp\left\{\gamma_{mb}\frac{d(x_j, v_i)}{\eta_i}\right\} \tag{20}$$

where γ_{mb} is a constant and typically chosen to be 1 similar as γ_b in (1). The parameter η_i in (20) represents the *zone of influence* of the ith bio-basis string v_i. Combining (1) and (20), the relation between the existing bio-basis function and *modified bio-basis function* is

$$\ln|f_{\text{modified}}(x_j, v_i)| \quad \left[\frac{\gamma_{mb}}{\gamma_b}\right]\left[\frac{h(v_i, v_i)}{\eta_i}\right]\ln|f(x_j, v_i)|. \tag{21}$$

If all the subsequences are partitioned into a set of disjoint clusters considering each bio-basis string as a cluster prototype, then the *zone of influence* of each bio-basis string represents the variance of that string with respect to the subsequences presented in that cluster. That is, the parameter η_i represents the radius of the cluster associated with the bio-basis string v_i. The value of η_i could be the same for all bio-basis strings if all of them are expected to form similar clusters in nonnumerical sequence space. In general, it is desirable that η_i should relate to the overall size and shape of the cluster associated with the bio-basis string v_i. In the present research work, the following definition is used:

$$\eta_i = \frac{1}{n_i}\sum_{x_j} d(x_j, v_i) = \frac{1}{n_i}\sum_{x_j}\{h(v_i, v_i) \quad h(x_j, v_i)\} \tag{22}$$

where n_i is the number of subsequences presented in the cluster associated with the ith bio-basis string v_i (that is, the number of subsequences having minimum dissimilarity with the ith bio-basis string v_i among all the bio-basis strings) and $\{h(v_i, v_i) \quad h(x_j, v_i)\}$ is the dissimilarity of the subsequence x_j from the bio-basis string v_i. In other words, the value of η_i represents the average dissimilarity of input subsequences from their corresponding bio-basis string v_i.

Hence, the main difference between the proposed and existing string kernel functions is that the former normalizes the asymmetric dissimilarity value by the *zone of influence* or variance of the bio-basis string considering the impact or influence of the bio-basis string in nonnumerical sequence space, while the later, as in (1), does not take into account this, rather it normalizes the dissimilarity value using maximum homology score of that bio-basis string.

IV. Proposed Bio-Basis String Selection Method

In real biological data analysis, the data set may contain a number of similar or redundant subsequences with low discriminant capability or relevance to the classes. The selection of such similar and nonrelevant subsequences as the bio-basis strings may lead to a reduction in the useful information in numerical feature space. Ideally, the selected bio-basis strings should have high discriminant capability with the classes while the similarity among them would be as low as possible. The subsequences with high discriminant capability are expected to be able to predict the classes of the subsequences. However, the prediction capability may be reduced if many similar subsequences are selected as the bio-basis strings. In contrast, a data set that contains subsequences not only with high relevance with respect to the classes but with low mutual redundancy is more effective in its prediction capability. Hence, to assess the effectiveness of the subsequences as the bio-basis strings, both relevance and redundancy (similarity) need to be measured quantitatively. The proposed bio-basis string selection method addresses the above issues through following three phases:

1) computation of the discriminant capability or relevance of each subsequence;
2) determination of the nonrelevant subsequences; and
3) computation of the similarity or redundancy among subsequences.

An asymmetric biological dissimilarity based Fisher ratio is chosen here to compute the discriminant capability or relevance of each subsequence to the classes, while a novel concept of the *degree of resemblance* is used to calculate the mutual redundancy (similarity) among subsequences. The nonrelevant subsequences are discarded using a nearest mean classifier [30]. Next the calculation of the Fisher ratio using asymmetric biological dissimilarity is provided, along with the concept of *degree of resemblance* and the principle of nearest mean classifier.

A. Fisher Ratio Using Biological Dissimilarity

In the proposed method, the Fisher ratio [30] is used to measure the discriminant capability or relevance of each subsequence $x_i \in X$. The Fisher ratio is calculated based on the asymmetric biological dissimilarity. As the class labels of all training subsequences are known, the set X can be partitioned into two groups or classes X_A (functional) and X_B (nonfunctional), where

$$X_A \cap X_B = \emptyset; \quad X_A \cup X_B = X; \tag{23}$$

$$|X_A| = n_A; \quad |X_B| = n_B; \quad n_A + n_B = n. \tag{24}$$

Hence, each subsequence $x_i \in X$ should have n_A and n_B dissimilarity values with the subsequences of X_A and X_B, respectively. Denoting the dissimilarity of the subsequence x_j from the subsequence x_i as $d(x_j, x_i)$, the mean and standard deviation values for the two classes X_A and X_B with respect to the subsequence x_i are as follows:

$$\bar{\mu}_{A_i} = \frac{1}{n_A} \sum d^2(x_j, x_i); \qquad \forall x_j \in X_A \tag{25}$$

$$\bar{\mu}_{B_i} = \frac{1}{n_B} \sum d^2(x_k, x_i); \qquad \forall x_k \in X_B \tag{26}$$

$$\bar{\sigma}^2_{A_i} = \frac{1}{n_A} \sum \{d^2(x_j, x_i) - \bar{\mu}_{A_i}\}^2; \quad \forall x_j \in X_A \tag{27}$$

$$\bar{\sigma}^2_{B_i} = \frac{1}{n_B} \sum \{d^2(x_k, x_i) - \bar{\mu}_{B_i}\}^2; \quad \forall x_k \in X_B \tag{28}$$

where $\bar{\mu}_{A_i}$, $\bar{\mu}_{B_i}$, $\bar{\sigma}_{A_i}$, and $\bar{\sigma}_{B_i}$ represent the mean and standard deviation values of the subsequence x_i for two groups X_A and X_B, respectively. These four quantities are calculated based on the square of biological dissimilarity with respect to the subsequence x_i. Based on these four quantities, the discriminant capability of each subsequence x_i is computed using the Fisher ratio that is as follows:

$$\mathbb{F}(x_i) = \frac{|\bar{\mu}_{A_i} - \bar{\mu}_{B_i}|}{\sqrt{\bar{\sigma}^2_{A_i} + \bar{\sigma}^2_{B_i}}}. \tag{29}$$

Let $\kappa_i = h(x_i, x_i)$ represents the maximum homology score of the subsequence x_i. The above four quantities can, then, be written using κ_i as

$$\bar{\mu}_{A_i} = \left\{\kappa_i^2 + E_A[h^2(x_j, x_i)] - 2\kappa_i E_A[h(x_j, x_i)]\right\} \tag{30}$$

$$\bar{\mu}_{B_i} = \left\{\kappa_i^2 + E_B[h^2(x_k, x_i)] - 2\kappa_i E_B[h(x_k, x_i)]\right\} \tag{31}$$

$$\begin{aligned}\bar{\sigma}^2_{A_i} = \{&4\kappa_i^2(E_A[h^2(x_j, x_i)] \\ &[E_A[h(x_j, x_i)]]^2) \\ &- 4\kappa_i(E_A[h^3(x_j, x_i)] - E_A[h(x_j, x_i)] \\ &\times E_A[h^2(x_j, x_i)]) \\ &- [E_A[h^2(x_j, x_i)]]^2 + E_A[h^4(x_j, x_i)]\}\end{aligned} \tag{32}$$

$$\begin{aligned}\text{and } \bar{\sigma}^2_{B_i} = \{&4\kappa_i^2(E_B[h^2(x_k, x_i)] \\ &- [E_B[h(x_k, x_i)]]^2) \\ &- 4\kappa_i(E_B[h^3(x_k, x_i)] \\ &E_B[h(x_k, x_i)]E_B[h^2(x_k, x_i)]) \\ &- [E_B[h^2(x_k, x_i)]]^2 + E_B[h^4(x_k, x_i)]\}\end{aligned} \tag{33}$$

where $E[h^r(x_j, x_i)]$ represents the zero-mean, rth order moment of similarity $h(x_j, x_i)$ between two subsequences x_i and x_j. Now, the numerator of (29) is given by

$$\begin{aligned}|\bar{\mu}_{A_i} - \bar{\mu}_{B_i}| = |\{&E_A[h^2(x_j, x_i)] - E_B[h^2(x_k, x_i)]\} \\ &-2\kappa_i\{E_A[h(x_j, x_i)] - E_B[h(x_k, x_i)]\}|.\end{aligned} \tag{34}$$

Hence, the numerator of (29) not only depends on the difference of zero-mean, first-order moment of similarity of two groups as in (11), it also takes into account the difference of zero-mean, second order moment of similarity as well as the maximum homology score of the subsequence x_i. That is, the numerator of (29) depends on following three factors:

- difference of zero-mean, first-order moment of similarity of two groups, $\{E_A[h(x_j, x_i)] - E_B[h(x_k, x_i)]\}$;
- difference of zero-mean, second order moment of similarity of two groups, $\{E_A[h^2(x_j, x_i)] - E_B[h^2(x_k, x_i)]\}$;
- maximum homology score of the subsequence x_i, that is, $\kappa_i = h(x_i, x_i)$.

Similarly, the denominator of (29) contains the following terms:

$$\begin{aligned}\sigma^2_{A_i} + \sigma^2_{B_i} = [&4\kappa_i^2\{E_A[h^2(x_j, x_i)] + E_B[h^2(x_k, x_i)]\} \\ &- 4\kappa_i^2\{E_A[h(x_j, x_i)]^2 + [E_B[h(x_k, x_i)]]^2\} \\ &- 4\kappa_i\{E_A[h^3(x_j, x_i)] + E_B[h^3(x_k, x_i)]\} \\ &+ 4\kappa_i\{E_A[h(x_j, x_i)]E_A[h^2(x_j, x_i)] \\ &+ E_B[h(x_k, x_i)]E_B[h^2(x_k, x_i)]\} \\ &\{[E_A[h^2(x_j, x_i)]]^2 + [E_B[h^2(x_k, x_i)]]^2\} \\ &+ \{E_A[h^4(x_j, x_i)] + E_B[h^4(x_k, x_i)]\}].\end{aligned} \tag{35}$$

Hence, the denominator of (29) considers the zero-mean, higher order (up to fourth order) moment of similarity of two groups as well as the maximum homology score of the subsequence x_i, while that of (11) only takes into account the zero-mean, first-, and second-order moment of similarity of two groups and does not consider the maximum homology score of the subsequence x_i. In effect, (29) calculates the discriminant capability of each subsequence x_i more accurately.

B. Nearest Mean Classifier

After computing the discriminant capability or relevance $\mathbb{F}(x_i)$ of each subsequence $x_i \in X$ using the Fisher ratio according to (29), the nonrelevant subsequences are discarded based on a threshold value δ. The subsequences those have the Fisher ratio values larger than or equal to the threshold value δ are considered as the candidate bio-basis strings. The value of δ is obtained using the concept of nearest mean classifier.

The proposed technique assumes at least one bio-basis string in the set X. If the Fisher ratio value $\mathbb{F}(x_s)$ of the subsequence x_s is the maximum, then x_s is declared to be the first bio-basis string. In order to find other candidate bio-basis strings, the threshold value δ is calculated using the nearest mean classifier. To obtain the reliable arithmetic mean, the subsequence x_s and those have the Fisher ratio values less than $\mathbb{F}(x_s)/10$ are removed [33], [34]. The mean $\mathcal{M}$ of the Fisher ratio values of the remaining subsequences is then calculated. Finally, the minimum mean distance is calculated as follows:

$$\mathcal{D}(x_s) = \min|\mathbb{F}(x_i) - \mathcal{M}| \quad 1 < i < n \tag{36}$$

where the Fisher ratio value $\mathbf{F}(x_s)$ of the subsequence x_s has the minimum distance with $\mathcal{M}$. To make the threshold value noise-insensitive, the Fisher ratio value $\mathbf{F}(x_s)$ that is closest to the mean $\mathcal{M}$ is set as δ, rather than the mean itself, that is,

$$\delta = \bar{\mathbf{\Gamma}}(x_s). \tag{37}$$

The basic steps of this approach follow next.

1) Compute the mean $\mathcal{M}$ of the Fisher ratio values of the subsequences without considering the best and below one tenth best Fisher ratio values.
2) Find out the Fisher ratio $\bar{\mathbf{\Gamma}}(x_s)$ of the subsequence x_s that has the minimum distance with $\mathcal{M}$ and set the threshold value $\delta = \bar{\mathbf{\Gamma}}(x_s)$.
3) Remove those subsequences with Fisher ratio values below the threshold δ.

After eliminating nonrelevant subsequences using the principle of nearest mean classifier, the redundancy among existing subsequences (candidate bio-basis strings) is calculated in terms of nongapped homology score. A quantitative measure is introduced next to compute the similarity or redundancy between two subsequences.

C. Degree of Resemblance

The *degree of resemblance* of the subsequence x_j with respect to the subsequence x_i is defined as

$$\mathrm{DOR}(x_j, x_i) = \frac{h(x_j, x_i)}{h(x_i, x_i)}. \tag{38}$$

It is the ratio between the nongapped pairwise homology score of two input subsequences x_i and x_j to the maximum homology score of the subsequence x_i. It is used to quantify the similarity in terms of homology score between pairs of subsequences. Combining (18) and (38), the relation between the *degree of resemblance* and the asymmetric dissimilarity of the subsequence x_j with respect to the subsequence x_i is

$$d(x_j, x_i) = h(x_i, x_i)[1 - \mathrm{DOR}(x_j, x_i)]. \tag{39}$$

The *degree of resemblance* is asymmetric in nature, that is,

$$\mathrm{DOR}(x_i, x_j) \neq \mathrm{DOR}(x_j, x_i). \tag{40}$$

This asymmetric property makes a reference subsequence different from the subsequence under study. It helps to find out redundant subsequences with respect to a selected bio-basis string. If two subsequences are different, the *degree of resemblance* between them is small. A high value of $\mathrm{DOR}(x_i, x_j)$ between two subsequences x_i and x_j asserts that the similarity between them is high. If two subsequences are same, the *degree of resemblance* between them is maximum, that is, 1. Thus,

$$0 < \mathrm{DOR}(x_i, x_j) \leq 1.$$

D. Details of the Algorithm

While the Fisher ratio is used to calculate the discriminant capability or relevance of each subsequence, the *degree of resemblance* takes into account the similarity or redundancy between two subsequences. Based on the concept of *degree of resemblance* as in (38) and the Fisher ratio as in (29), the method for selecting a reduced set of most relevant bio-basis strings is described next. The algorithm proceeds as follows:

Selection of Bio-Basis Strings

- Input: $X = \{x_1, \ldots, x_j, \ldots, x_n\}$ be the set of n subsequences with m residues, where $x_j \subset \mathcal{A}^m$ and $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \ldots, \mathrm{W}, \mathrm{Y}\}$ be the set of 20 amino acids.
- Output: $V = \{v_1, \ldots, v_i, \ldots, v_c\}$ be the set of c bio-basis strings with m residues, where $v_i \in X$ and $v_{ik} \in \mathcal{A}$.

begin:

1) Initialize $\bar{V} \leftarrow X$ and $V \leftarrow \emptyset$.
2) Calculate the discriminant capabilities of all subsequences of $\bar{V}$ using the Fisher ratio as in (29).
3) Compute the threshold value δ using (37).
4) Remove the subsequences from $\bar{V}$ those have Fisher ratio values below the threshold δ.
5) Repeat steps a) and b) for all the remaining subsequences of $\bar{V}$.
   a) Select a subsequence from $\bar{V}$ as the candidate bio-basis string of V that has the highest Fisher ratio value (maximum discriminant capability).
   b) Remove the subsequences from $\bar{V}$ those have the DOR values with respect to the selected bio-basis string of step a) above the threshold ξ.
6) Stop.

Note that the main motive of introducing the concepts of *degree of resemblance* and nearest mean classifier lies in reducing the number of bio-basis strings. That is, both attempt to eliminate nonrelevant and redundant bio-basis strings from the whole subsequences. The whole approach is therefore data dependent.

V. Conclusion

The main contribution of this paper is twofold, namely:

1) the development of a new string kernel function based on the principle of biological dissimilarity and the concept of *zone of influence* of bio-basis string; and
2) the development of a method for selection of a reduced set of most relevant and nonredundant bio-basis strings.

The concept of *zone of influence* introduced in the proposed *modified bio-basis function* normalizes the biological dissimilarity. As the *modified bio-basis function* takes into account the influence of each bio-basis string in nonnumerical sequence space, it can transform nonnumerical sequence space to numerical feature space more accurately than the existing bio-basis function. Moreover, the proposed bio-basis string selection method overcomes the limitations of existing string selection methods and is expected to be more effective to select a reduced set of most relevant and nonredundant bio-basis strings. However, the effectiveness of the proposed kernel function and proposed bio-basis string selection method, along with a comparison with existing bio-basis function and related bio-basis string selection methods, is demonstrated in [35] using different protein data sets.

ACKNOWLEDGMENT

The authors would like to thank Dr. Kuntal Ghosh of Indian Statistical Institute, Kolkata, India, and anonymous referees for providing helpful comments and valuable criticisms on the original version of the manuscript.

REFERENCES

[1] P. Arrigo, F. Giuliano, and G. Damiani, "Identification of a new motif on nucleic acid sequence data using Kohonen's self-organising map," *CABIOS*, vol. 7, pp. 353–357, 1991.

[2] E. A. Ferran and P. Ferrara, "Topological maps of protein sequences," *Biol. Cybern.*, vol. 65, pp. 451–458, 1991.

[3] Y. D. Cai and K. C. Chou, "Artificial neural network model for predicting HIV protease cleavage sites in protein," *Adv. Eng. Softw.*, vol. 29, no. 2, pp. 119–128, 1998.

[4] P. Baldi, G. Pollastri, C. A. Anderson, and S. Brunak, "Matching protein beta-sheet partners by feedforward and recurrent neural networks," in *Proc. Int. Conf. Intell. Syst. Mol. Biol. (ISMB)*, 1995, vol. 8, pp. 25–36.

[5] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*. Cambridge, MA: MIT Press, 1998.

[6] Z. R. Yang, "Biological application of support vector machines," *Briefings Bioinf.*, vol. 5, no. 4, pp. 328–338, 2004.

[7] S. F. Altschul, W. Gish, W. Miller, E. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.

[8] A. V. Aho and M. Corasick, "Efficient string matching: An aid to bibliographic search," *Commun. ACM*, vol. 18, no. 6, pp. 333–340, 1975.

[9] K. Thompson, "Regular expression search algorithm," *Commun. ACM*, vol. 11, no. 6, pp. 419–422, 1968.

[10] D. B. Searls and K. P. Murphy, "Automata-theoretic models of mutation and alignment," in *Proc. 3rd Int. Conf. Intell. Syst. Mol. Biol.*, 1995, pp. 341–349.

[11] D. B. Searls, "Sequence alignment through pictures," *Trends Genet.*, vol. 12, pp. 35–37, 1996.

[12] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *J. Mol. Biol.*, vol. 202, pp. 865–884, 1988.

[13] A. Narayanan, X. K. Wu, and Z. R. Yang, "Mining viral protease data to extract cleavage knowledge," *Bioinformatics*, vol. 18, pp. 5–13, 2002.

[14] R. Thomson, C. Hodgman, Z. R. Yang, and A. K. Doyle, "Characterising proteolytic cleavage site activity using bio-basis function neural network," *Bioinformatics*, vol. 19, no. 14, pp. 1741–1747, 2003.

[15] E. A. Berry, A. R. Dalby, and Z. R. Yang, "Reduced bio-basis function neural network for identification of protein phosphorylation sites: Comparison with pattern recognition algorithms," *Comput. Biol. Chem.*, vol. 28, pp. 75–85, 2004.

[16] Z. R. Yang and R. Thomson, "Bio-basis function neural network for prediction of protease cleavage sites in proteins," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 263–274, 2005.

[17] Z. R. Yang, "Prediction of caspase cleavage sites using Bayesian bio-basis function neural networks," *Bioinformatics*, vol. 21, no. 9, pp. 1831–1837, 2005.

[18] Z. R. Yang, R. Thomson, P. McNeil, and R. Esnouf, "RONN: Use of the bio-basis function neural network technique for the detection of natively disordered regions in proteins," *Bioinformatics*, vol. 21, no. 16, pp. 3369–3376, 2005.

[19] Z. R. Yang and K. C. Chou, "Predicting the linkage sites in glycoproteins using bio-basis function neural networks," *Bioinformatics*, vol. 20, no. 6, pp. 903–908, 2004.

[20] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou, "Support vector machines for predicting the specificity of GalNAc-transferase," *Peptides*, vol. 23, pp. 205–208, 2002.

[21] Y. Minakuchi, K. Satou, and A. Konagaya, "Prediction of protein-protein interaction sites using support vector machines," *Genome Inf.*, vol. 13, pp. 322–323, 2002.

[22] A. Gutteridge, G. J. Bartlett, and J. M. Thornton, "Using a neural network and spatial clustering to predict the location of active sites in enzymes," *J. Mol. Biol.*, vol. 330, pp. 719–734, 2003.

[23] Y. Bengio and Y. Pouliot, "Efficient recognition of immunological domains from amino acid sequences using a neural network," *CABIOS*, vol. 6, pp. 319–324, 1990.

[24] S. F. Altschul, M. S. Boguski, W. Gish, and J. C. Wootton, "Issues in searching molecular sequence databases," *Nature Genet.*, vol. 6, pp. 119–129, 1994.

[25] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *PNSA*, vol. 89, pp. 10 915–10 919, 1992.

[26] M. S. Johnson and J. P. Overington, "A structural basis for sequence comparisons: An evaluation of scoring methodologies," *J. Mol. Biol.*, vol. 233, pp. 716–738, 1993.

[27] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins. Matrices for detecting distant relationships," in *Atlas of Protein Sequence and Structure*. Washington, DC: Nat. Biomed. Res. Found., 1978, vol. 5, pp. 345–358.

[28] C. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: Univ. Illinois Press, 1964.

[29] Z. R. Yang, "Orthogonal kernel machine for the prediction of functional sites in proteins," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 1, pp. 100–106, 2005.

[30] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.

[31] A. Stojmirovic, "Quasi-metric spaces with measure," *Topol. Proc.*, vol. 28, no. 2, pp. 655–671, 2004.

[32] M. Itoh, S. Goto, T. Akutsu, and M. Kanehisa, "Fast and accurate database homology search using upper bounds of local alignment scores," *Bioinformatics*, vol. 21, no. 7, pp. 912–921, 2005.

[33] Y. M. Lui and H.-D. Cheng, "A new peak selection criterion based on minimizing the classification error," *Inf. Sci.*, vol. 94, pp. 213–233, 1996.

[34] P. Maji, M. K. Kundu, and B. Chanda, "Second order fuzzy measure and weighted co-occurrence matrix for segmentation of brain MR images," *Fundamenta Informaticae*, vol. 88, no. 1–2, pp. 161–176, 2008.

[35] P. Maji and C. Das, "Protein functional sites prediction using modified bio-basis function and quantitative indices," *IEEE Trans. NanoBiosci.*, accepted for publication.

Pradipta Maji received the B.Sc. degree in physics, the M.Sc. degree in electronics science, and the Ph.D. degree in the area of computer science from Jadavpur University, India, in 1998, 2000, and 2005, respectively.

Currently, he is an Assistant Professor in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. He has published around 60 papers in international journals and conferences. He is also a reviewer of many international journals. His research interests include pattern recognition, computational biology and bioinformatics, medical image processing, cellular automata, soft computing, and so forth.

Dr. Maji has received the 2006 Best Paper Award of the International Conference on Visual Information Engineering from The Institution of Engineering and Technology, U.K., the 2008 Microsoft Young Faculty Award from Microsoft Research Laboratory India Pvt., and the 2009 Young Scientist Award from the National Academy of Sciences, India, and has been selected as the 2009 Associate of the Indian Academy of Sciences, India.

Chandra Das received the B.Sc. degree in computer science, the M.Sc. degree in computer and information science, and the M.Tech. degree in computer science and engineering from the University of Calcutta, India, in 1999, 2001, and 2003, respectively, and has submitted the Ph.D. (Engg.) thesis at Jadavpur University, India, in 2010.

Currently, she is a Senior Lecturer in the Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India. She has a number of publications in international journals and conferences. Her research interests include pattern recognition, computational biology and bioinformatics, machine learning, soft computing, and so forth.