Rough-Fuzzy *C*-Medoids Algorithm and Selection of Bio-Basis for Amino Acid Sequence Analysis

Pradipta Maji and Sankar K. Pal, *Fellow*, *IEEE*

Abstract—In most pattern recognition algorithms, amino acids cannot be used directly as inputs since they are nonnumerical variables. They, therefore, need encoding prior to input. In this regard, bio-basis function maps a nonnumerical sequence space to a numerical feature space. It is designed using an amino acid mutation matrix. One of the important issues for the bio-basis function is how to select the minimum set of bio-bases with maximum information. In this paper, we describe an algorithm, termed as rough-fuzzy c-medoids (RFCMdd) algorithm, to select the most informative bio-bases. It is comprised of a judicious integration of the principles of rough sets, fuzzy sets, the c-medoids algorithm, and the amino acid mutation matrix. While the membership function of fuzzy sets enables efficient handling of overlapping partitions, the concept of lower and upper bounds of rough sets deals with uncertainty, vagueness, and incompleteness in class definition. The concept of crisp lower bound and fuzzy boundary of a class, introduced in RFCMdd, enables efficient selection of the minimum set of the most informative bio-bases. Some new indices are introduced for evaluating quantitatively the quality of selected bio-bases. The effectiveness of the proposed algorithm, along with a comparison with other algorithms, has been demonstrated on different types of protein data sets.

Index Terms—Pattern recognition, data mining, c-medoids algorithm, fuzzy sets, rough sets, bioinformatics.

1 Introduction

Recent advancement and wide use of high-throughput technology for biological research are producing an enormous size of biological data. Data mining techniques and machine learning methods provide useful tools for analyzing these biological data. The successful analysis of biological sequences relies on the efficient coding of the biological information contained in sequences/subsequences. For example, to recognize functional sites within a biological sequence, the subsequences obtained through moving a fixed length sliding window are generally analyzed. The problem with using most pattern recognition algorithms to analyze these biological subsequences is that they cannot recognize nonnumerical features such as the biochemical codes of amino acids. Investigating a proper encoding process prior to modeling the amino acids is then critical.

The most commonly used method for coding a subsequence is distributed encoding, which encodes each of 20 amino acids using a 20-bit binary vector [1]. However, in this method, the input space is expanded unnecessarily. Also, this method may not be able to encode biological content in sequences efficiently. On the other hand, different distances for different amino acid pairs have been defined, by various mutation matrices, and validated [2], [3], [4]. However, they cannot be used directly for encoding an amino acid to a unique numerical value.

In this background, Thomson et al. [5], Berry et al. [6], and Yang and Thomson [7] proposed the concept of a bio-basis function for analyzing biological sequences. It uses a kernel function to transform biological sequences to feature vectors directly. Bio-bases consist of sections of biological sequences that code for a feature of interest in the study and are responsible for the transformation of biological data to high-dimensional feature space. Transformation of input data to high-dimensional feature space is performed based on the similarity of an input sequence to a bio-basis with reference to a biological similarity matrix. Thus, the biological content in the sequences can be maximally utilized for accurate modeling. The use of similarity matrices to map features allows the bio-basis function to analyze biological sequences without the need for encoding.

The most important issue for bio-basis function is how to select the minimum set of bio-bases with maximum information. Berry et al. [6] used genetic algorithms for bio-bases selection considering the Fisher ratio as the fitness function. Yang and Thomson [7] proposed a method to select bio-bases using mutual information (MI). In principle, the bio-bases in nonnumerical sequence space should be such that the degree of resemblance (DOR) between pairs of bio-bases would be as minimum as possible. Each of them would then represent a unique feature in numerical feature space. As this is a feature selection problem, the clustering method can be used, which partitions the given biological sequences into subgroups around each bio-basis, each of which should be as homogeneous/informative as possible. However, the methods proposed in [6] and [7] have not adequately addressed this problem. Also, not much attention has been paid to it earlier.

• *The authors are with the Center for Soft Computing Research, Indian Statistical Institute, Kolkata, 203 B. T. Road, 700108, India. E-mail: {pmaji, sankar}@isical.ac.in.*

Manuscript received 20 June 2006; revised 20 Nov. 2006; accepted 29 Jan. 2007; published online 6 Feb. 2007.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0307-0606.
Digital Object Identifier no. 10.1109/TKDE.2007.1027.

In biological sequences, the only available information is the numerical values that represent the degrees to which pairs of sequences in the data set are related. Algorithms that generate partitions of that type of relational data are usually referred to as relational or pairwise clustering algorithms. A well-known relational clustering algorithm is c-medoids due to Kaufman and Rousseeuw [8]. The c-medoids algorithm is applicable to situations where the objects to be clustered cannot be represented by numerical features, rather, only represented with similarities/dissimilarities between pairs of objects. Therefore, the relational clustering algorithms can be used to cluster biological subsequences if one can come up with a similarity measure to quantify the DOR between the pairs of subsequences. The pairwise similarities are usually stored in the form of a matrix called the similarity matrix.

One of the main problems in biological subsequence analysis is uncertainty. Some of the sources of this uncertainty include incompleteness and vagueness in class definitions. In this background, the possibility concept introduced by the fuzzy-sets theory [9] and rough-sets theory [10] have gained popularity in modeling and propagating uncertainty. Both fuzzy sets and rough sets provide a mathematical framework to capture uncertainties associated with the data [11], [12], [13], [14]. Two of the early rough-clustering algorithms are those due to Hirano and Tsumoto [15] and De [16]. Other notable algorithms include rough c-means [17], rough self organizing map [18], rough support vector clustering [19], and so forth. In [20], the indiscernibility relation of rough sets has been used to initialize the expectation-maximization algorithm. The most notable fuzzy relational algorithm is fuzzy c-medoids (FCMdd) due to Krishnapuram et al. [21]. Recently, combining rough sets and fuzzy sets, Mitra et al. proposed rough-fuzzy collaborative clustering [22].

In this paper, we propose an algorithm, termed as RFCMdd algorithm, based on rough sets and fuzzy sets to select the most informative bio-bases. Although the membership function of fuzzy sets enables efficient handling of overlapping partitions, the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in class definition. Each partition is represented by a medoid (bio-basis), a crisp lower approximation, and a fuzzy boundary. The lower approximation influences the fuzziness of the final partition. The medoid (bio-basis) depends on the weighting average of the crisp lower approximation and fuzzy boundary. The concept of "DOR," based on nongapped pairwise homology alignment score, circumvents the initialization and local minima problems of c-medoids and enables efficient selection of the minimum set of the most informative bio-bases. Some quantitative measures are introduced based on MI and nongapped pairwise homology alignment scores to evaluate the quality of selected bio-bases. The effectiveness of the proposed algorithm, along with a comparison with hard c-medoids (HCMdd) [8], rough c-medoids (RCMDdd), FCMdd [21], Berry et al.'s method [6], and Yang and Thomson's method [7], has been demonstrated on different protein data sets.

The structure of the rest of this paper is given as follows: Section 2 briefly introduces necessary notions of bio-basis function, rough sets, and fuzzy sets. In Section 3, a new c-medoids algorithm is proposed based on rough sets and fuzzy sets for bio-bases selection. Some quantitative measures are presented in Section 4 to select the most informative bio-bases. A few case studies and a comparison with other methods are presented in Section 5. Concluding remarks are given in Section 6.

2 Bio-Basis Function, Rough Set, and Fuzzy Set

In this section, the basic notions in the theories of bio-basis function, rough sets, and fuzzy sets are reported.

2.1 Bio-Basis Function

The most successful method of sequence analysis is homology alignment [23], [24]. In this method, the function of a sequence is annotated through aligning a novel sequence with known sequences. If the homology alignment between a novel sequence and a known sequence gives a very high similarity score, the novel sequence is believed to have the same or similar function as the known sequence. In homology alignment, 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 [3], [4].

However, the principle of homology alignment cannot be used directly for subsequence analysis because a subsequence may not contain enough information for conventional homology alignment. A high homology alignment score between a novel subsequence and a known subsequence cannot assert that two subsequences have the same function. However, it can be assumed that they may have the same function statistically.

The design of bio-basis function is based on the principle of conventional homology alignment used in biology. Using a table lookup technique, a homology alignment score as a similarity value can be obtained for a pair of subsequences. The nongapped homology alignment method is used to calculate this similarity value, where no deletion or insertion is used to align two subsequences. The definition of bio-basis function is given as follows [5], [7]:

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

where $h(x_j, v_i)$ is the nongapped pairwise homology alignment score between a subsequence x_j and a bio-basis v_i calculated using an amino acid mutation matrix [3], [4], $h(v_i, v_i)$ denotes the maximum homology alignment score of the ith bio-basis v_i, and γ is a constant. Let $\mathbb{A}$ be the set of 20 amino acids, $X = \{x_1, \cdots, x_j, \cdots, x_n\}$ be the set of n subsequences with m residues, and $V = \{v_1, \cdots, v_i, \cdots, v_c\} \subset X$ be the set of c bio-bases such that $v_{ik},\ x_{jk} \in \mathbb{A},\ \forall_{i=1}^{c},\ \forall_{j=1}^{n},\ \forall_{k=1}^{m}$. The nongapped pairwise homology alignment score between x_j and v_i is then defined as

$$h(x_j, v_i) = \sum_{k=1}^{m} M(x_{jk}, v_{ik}), \quad (2)$$

where $M(x_{jk}, v_{ik})$ can be obtained from an amino acid mutation matrix through a table lookup method. The function value is high if two subsequences are similar or close to each other and one for two identical subsequences. The value is small if two subsequences are distinct.

The bio-basis function transforms various homology alignment scores to a real number as a similarity within the interval [0, 1]. Each bio-basis is a feature dimension in a numerical feature space. It needs a subsequence as a support. A collection of c bio-bases formulates a numerical feature space $\mathbb{R}^c$. After the mapping using bio-bases, a nonnumerical subsequence space $\mathbb{A}^m$ will be mapped to a c-dimensional numerical feature space $\mathbb{R}^c$, that is, $\mathbb{A}^m \rightarrow \mathbb{R}^c$.

The most important assumption about bio-basis function is that the distribution of the amino acids in sequences depends on the specificity of the sequences. If the distribution of amino acids is random, the selection of bio-basis will be very difficult. Fortunately, the biological experiments have shown that the distribution of amino acids at the specific subsites in sequences does depend on the specificity of the sequences.

2.2 Rough Sets

The theory of rough sets begins with the notion of an approximation space, which is a pair $< U, R >$, where U is a nonempty set (the universe of discourse), and R is an equivalence relation on U, that is, R is reflexive, symmetric, and transitive. The relation R decomposes the set U into disjoint classes in such a way that two elements x and y are in the same class if and only if $(x, y) \in R$. Let us denote by U/R the quotient set of U by the relation R, and

$$U/R = \{X_1, \cdots, X_i, \cdots, X_p\},$$

where X_i is an equivalence class of R, $i = 1, 2, \cdots, p$. If two elements x and y in U belong to the same equivalence class $X_i \in U/R$, we say that x and y are indistinguishable. The equivalence classes of R and the empty set $\emptyset$ are the elementary sets in the approximation space $< U, R >$. Given an arbitrary set $X \in 2^U$, in general, it may not be possible to describe X precisely in $< U, R >$. One may characterize X by a pair of lower and upper approximations defined as [10]

$$\underline{R}(X) = \bigcup_{X_i \subseteq X} X_i; \quad \overline{R}(X) = \bigcup_{X_i \cap X \neq \emptyset} X_i.$$

That is, the lower approximation $\underline{R}(X)$ is the union of all the elementary sets, which are subsets of X, and the upper approximation $\overline{R}(X)$ is the union of all the elementary sets, which have a nonempty intersection with X. The interval $[\underline{R}(X), \overline{R}(X)]$ is the representation of an ordinary set X in the approximation space $< U, R >$ or simply called the rough set of X. The lower (respectively, upper) approximation $\underline{R}(X)$ (respectively, $\overline{R}(X)$) is interpreted as the collection of those elements of U that definitely (respectively, possibly) belong to X. Further,

- a set $X \in 2^U$ is said to be definable (or exact) in $< U, R >$ if and only if $\underline{R}(X) = \overline{R}(X)$;
- for any $X, Y \in 2^U$, X is said to be roughly included in Y, denoted by $X \tilde{\subset} Y$, if and only if $\underline{R}(X) \subseteq \underline{R}(Y)$ and $\overline{R}(X) \subseteq \overline{R}(Y)$; and
- X and Y are said to be roughly equal, denoted by $X \simeq_R Y$, in $< U, R >$ if and only if $\underline{R}(X) = \underline{R}(Y)$ and $\overline{R}(X) = \overline{R}(Y)$.

In [10], Pawlak discusses two numerical characterizations of imprecision of a subset X in the approximation space $< U, R >$: accuracy and roughness. The accuracy of X, denoted by $\alpha_R(X)$, is the ratio of the number of objects in its lower approximation to that in its upper approximation, namely,

$$\alpha_R(X) = \frac{|\underline{R}(X)|}{|\overline{R}(X)|}.$$

The roughness of X, denoted by $\rho_R(X)$, is defined by subtracting the accuracy from 1:

$$\rho_R(X) = 1 - \alpha_R(X) = 1\frac{|\underline{R}(X)|}{|\overline{R}(X)|}.$$

Note that the lower the roughness of a subset, the better is its approximation. Further, the following observations are easily obtained:

1. As $\underline{R}(X) \subseteq X \subseteq \overline{R}(X)$, $0 \leq \rho_R(X) \leq 1$.
2. By convention, when $X = \emptyset$, $\underline{R}(X) = \overline{R}(X) = \emptyset$ and $\rho_R(X) = 0$.
3. $\rho_R(X) = 0$ if and only if X is definable in $< U, R >$.

2.3 Fuzzy Set

Let U be a finite and nonempty set called universe. A fuzzy set F of U is a mapping from U into the unit interval [0, 1]:

$$\mu_F : U \rightarrow [0, 1],$$

where, for each $x \in U$, we call $\mu_F(x)$ the membership degree of x in F. Practically, we may consider U as a set of objects of concern, and a crisp subset of U represents a nonvague concept imposed on objects in U. Then, a fuzzy set F of U is thought of as a mathematical representation of a vague concept described linguistically. The support of fuzzy set F is the crisp set that contains all the elements of U that have a nonzero membership value in F [9].

A function mapping all the elements in a crisp set into real numbers in [0, 1] is called a membership function. The larger value of the membership function represents the higher degree of the membership. It means how closely an element resembles an ideal element. Membership functions can represent the uncertainty using some particular functions. These functions transform the linguistic variables into numerical calculations by setting some parameters. The fuzzy decisions can then be made.

3 Rough Fuzzy C-Medoids Algorithm

In this section, we first describe two existing relational clustering algorithms—HCMdd [8] and FCMdd [21], for selection of bio-bases. Next, we propose two relational algorithms—RCMdd and RFCMdd, incorporating the concept of lower and upper approximations of rough sets into HCMdd and FCMdd, respectively. Some quantitative measures are introduced to select the minimum set of the most informative bio-bases.

3.1 HCMdd

The HCMdd algorithm [8] uses the most centrally located object in a cluster, which is termed as the medoid. A medoid is essentially an existing data from the cluster, which is closest to the mean of the cluster.

The objective of the HCMdd algorithm for selection of bio-bases is to assign n subsequences to c clusters. Each of the clusters β_i is represented by a bio-basis v_i, which is the medoid for that cluster. The process begins by randomly choosing c subsequences as the bio-bases. The subsequences are assigned to one of the c clusters based on the maximum value of the nongapped pairwise homology alignment score $h(x_j, v_i)$ between the subsequence x_j and the bio-basis v_i. After the assignment of all the subsequences to various clusters, the new bio-bases are calculated as follows:

$$v_i = x_q, \tag{3}$$

where q is given by

$$q = \arg\max\{h(x_k, x_j)\}; \quad x_j \in \beta_i; \quad x_k \in \beta_i,$$

and $h(x_k, x_j)$ can be calculated as per (2). The basic steps are outlined as follows:

1. Arbitrarily choose c subsequences as the initial bio-bases v_i, $i = 1, 2, \cdots, c$.
2. Assign each remaining subsequence to the cluster for the closest bio-basis.
3. Compute the new bio-basis as per (3).
4. Repeat Steps 2 and 3 until no more new assignments can be made.

3.2 FCMdd

This provides a fuzzification of the HCMdd algorithm [21]. For bio-bases selection, it maximizes

$$J = \sum_{j=1}^{n}\sum_{i=1}^{c}(\mu_{ij})^{\acute{m}}\{h(x_j, v_i)\}, \tag{4}$$

where $1 \leq \acute{m} < \infty$ is the fuzzifier, and $\mu_{ij} \in [0, 1]$ is the fuzzy membership of the subsequence x_j in cluster β_i, such that

$$\mu_{ij} = \sum_{l=1}^{c}\left\{\frac{h(x_j, v_i)}{h(x_j, v_l)}\right\}^{\frac{1}{\acute{m}-1}} \tag{5}$$

subject to

$$\sum_{i=1}^{c}\mu_{ij} = 1, \forall j, \text{ and } 0 < \sum_{j=1}^{n}\mu_{ij} < n, \forall i.$$

The new bio-bases are calculated as

$$v_i = x_q, \tag{6}$$

where q is given by

$$q = \arg\max\sum_{k=1}^{n}(\mu_{ik})^{\acute{m}}\{h(x_k, x_j)\}; \quad 1 \leq j \leq n.$$

The algorithm proceeds as follows:

1. Assign initial bio-bases v_i, $i = 1, 2, \cdots, c$. Choose values for fuzzifier $\acute{m}$ and threshold ϵ_1. Set iteration counter $t = 1$.

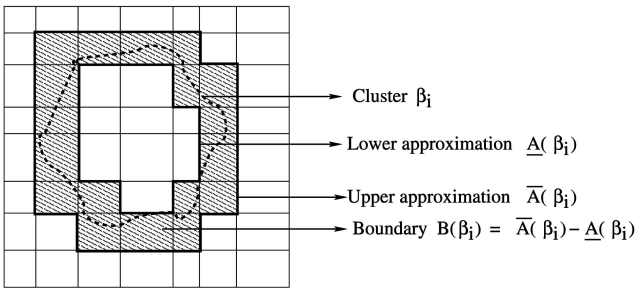


Fig. 1. RCMdd: cluster β_i is represented by lower and upper bounds $[\underline{A}(\beta_i), \overline{A}(\beta_i)]$.

2. Compute membership μ_{ij} by (5) for c clusters and n subsequences.
3. Update bio-basis v_i by (6).
4. Repeat steps 2 to 4, by incrementing t, until $|\mu_{ij}(t) - \mu_{ij}(t-1)| > \epsilon_1$.

3.3 RCMdd

Let $\underline{A}(\beta_i)$ and $\overline{A}(\beta_i)$ be the lower and upper approximations of cluster β_i, and $B(\beta_i) = \overline{A}(\beta_i) - \underline{A}(\beta_i)$ denotes the boundary region of cluster β_i (Fig. 1). In the RCMdd algorithm, the concept of c-medoids algorithm is extended by viewing each cluster β_i as an interval or rough set. However, it is possible to define a pair of lower and upper bounds $[\underline{A}(\beta_i), \overline{A}(\beta_i)]$ or a rough set for every set $\beta_i \subseteq U$, U is the set of objects of concern [10]. The family of upper and lower bounds are required to follow some of the basic rough set properties such as

1. an object x_j can be part of at most one lower bound;
2. $x_j \in \underline{A}(\beta_i) \Rightarrow x_j \in \overline{A}(\beta_i)$; and
3. an object x_j is not part of any lower bound $\Rightarrow x_j$, belongs to two or more upper bounds.

Incorporating rough sets into c-medoids algorithm, we propose RCMdd for generating bio-bases. It adds the concept of lower and upper bounds of rough sets into HCMdd algorithm. It classifies the subsequence space into two parts—lower approximation and boundary region. The bio-basis (medoid) is calculated based on the weighting average of the lower bound and boundary region. All the subsequences in lower approximation take the same weight w, whereas all the subsequences in boundary take another weighting index $\tilde{w}$ uniformly. Calculation of the bio-bases is modified to include the effects of lower, as well as upper, bounds. The modified bio-bases calculation for RCMdd is given by

$$v_i = x_q, \tag{7}$$

where q is given by

$$q = \arg\max\begin{cases} w \times \mathcal{A} + \tilde{w} \times \mathcal{B} & \text{if } \underline{A}(\beta_i) \neq \emptyset, B(\beta_i) \neq \emptyset \\ \mathcal{A} & \text{if } \underline{A}(\beta_i) \neq \emptyset, B(\beta_i) = \emptyset \\ \mathcal{B} & \text{if } \underline{A}(\beta_i) = \emptyset, B(\beta_i) \neq \emptyset \end{cases}$$

$$\mathcal{A} = \sum_{x_k \in \underline{A}(\beta_i)} h(x_k, x_j); \quad \mathcal{B} = \sum_{x_k \in B(\beta_i)} h(x_k, x_j).$$

The parameters w and $\tilde{w}$ $(= 1 - w)$ correspond to the relative importance of lower bound and boundary region. Since the subsequences lying in lower approximation definitely

belong to a cluster, they are assigned a higher weight w compared to $\tilde{w}$ of the subsequences lying in the boundary region. That is, $0 < \tilde{w} < w < 1$. The main steps of RCMdd algorithm are described as follows:

1. Assign initial bio-bases v_i, $i = 1, 2, \cdots, c$. Choose a value for threshold ϵ_2.
2. For each subsequence x_j, calculate the homology alignment score $h(x_j, v_i)$ between itself and the bio-basis v_i of cluster β_i.
3. If $h(x_j, v_i)$ is maximum for $1 \leq i \leq c$ and
$$h(x_j, v_i) - h(x_j, v_k) \leq \epsilon_2,$$
then $x_j \in \overline{A}(\beta_i)$ and $x_j \in \overline{A}(\beta_k)$. Furthermore, x_j is not part of any lower bound.
4. Otherwise, $x_j \in \underline{A}(\beta_i)$ such that $h(x_j, v_i)$ is the maximum for $1 \leq i \leq c$. In addition, by properties of rough sets, $x_j \in \overline{A}(\beta_i)$.
5. Compute the new bio-basis as per (7).
6. Repeat Steps 2 to 5 until no more new assignments can be made.

3.4 RFCMdd

Incorporating both fuzzy sets and rough sets, next, we propose another version of c-medoids algorithm, termed as RFCMdd. The proposed RFCMdd algorithm adds the concept of fuzzy membership of fuzzy sets and the lower and upper approximations of rough sets into the c-medoids algorithm. Although the lower and upper bounds of rough sets deal with uncertainty, vagueness, and incompleteness in class definition, the membership of fuzzy sets enables efficient handling of overlapping partitions.

In FCMdd, the bio-basis (medoid) depends on the fuzzy membership values of different subsequences. Whereas in RFCMdd, after computing the memberships for c clusters and n subsequences, the membership values of each subsequence are sorted, and the difference of the two highest memberships is compared with a threshold value ϵ_2. Let μ_{ij} and μ_{kj} be the highest and second highest memberships of subsequence x_j. If $(\mu_{ij} - \mu_{kj}) > \epsilon_2$, then $x_j \in \underline{A}(\beta_i)$, as well as $x_j \in \overline{A}(\beta_i)$ and $x_j \notin \underline{A}(\beta_k)$; otherwise, $x_j \in B(\beta_i)$ and $x_j \in B(\beta_k)$. That is, the proposed algorithm first separates the "core" and overlapping portions of each cluster β_i based on the threshold value ϵ_2. The "core" portion of the cluster β_i is represented by its lower approximation $\underline{A}(\beta_i)$, whereas the boundary region $B(\beta_i)$ represents the overlapping portion. In effect, it minimizes the vagueness and incompleteness in cluster definition.

According to the definitions of lower approximations and boundary of rough sets, if a subsequence $x_j \in \underline{A}(\beta_i)$, then $x_j \notin \underline{A}(\beta_k)$, $\forall k \neq i$, and $x_j \notin B(\beta_i)$, $\forall i$. That is, the subsequence x_j is contained in β_i definitely. Thus, the weights of the subsequences in the lower approximation of a cluster should be independent of other bio-bases and clusters and should not be coupled with their similarity with respect to other bio-bases. Also, the subsequences in lower approximation of a cluster should have similar influence on the corresponding bio-basis and cluster. Although if $x_j \in B(\beta_i)$, then the subsequence x_j possibly belongs to β_i and potentially belongs to another cluster.

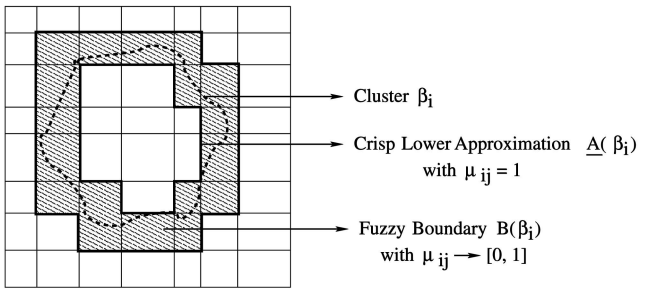


Fig. 2. RFCMdd: cluster β_i is represented by a crisp lower bound and fuzzy boundary.

Hence, the subsequences in boundary regions should have different influence on the bio-bases and clusters.

Therefore, in RFCMdd, after assigning each subsequence in the lower approximations and boundary regions of different clusters based on ϵ_2, the memberships μ_{ij} of the subsequences are modified. The membership values of the subsequences in lower approximation are set to 1, whereas those in the boundary regions remain unchanged. In other words, the proposed c-medoids first partitions the data into two classes—lower approximation and boundary. The concept of fuzzy memberships is applied only to the subsequences of boundary region, which enables the algorithm to handle overlapping clusters. Thus, in RFCMdd, each cluster is represented by a bio-basis (medoid), a crisp lower approximation, and a fuzzy boundary (Fig. 2). The lower approximation influences the fuzziness of final partition. The FCMdd can be reduced from RFCMdd when $\underline{A}(\beta_i) = \emptyset$, $\forall i$. Thus, the proposed algorithm is the generalization of existing FCMdd algorithm.

The new bio-bases are calculated based on the weighting average of the crisp lower approximation and fuzzy boundary. Computation of the bio-bases is modified to include the effects of both fuzzy membership and lower and upper bounds. Since the subsequences lying in lower approximation definitely belong to a cluster, they are assigned a higher weight compared to that of the subsequences lying in the boundary region. The modified bio-bases calculation for RFCMdds is therefore given by

$$v_i = x_q, \tag{8}$$

where q is given by

$$q = \arg\max \begin{cases} w \times \mathcal{A} + \tilde{w} \times \mathcal{B} & \text{if } \underline{A}(\beta_i) \neq \emptyset, B(\beta_i) \neq \emptyset \\ \mathcal{A} & \text{if } \underline{A}(\beta_i) \neq \emptyset, B(\beta_i) = \emptyset \\ \mathcal{B} & \text{if } \underline{A}(\beta_i) = \emptyset, B(\beta_i) \neq \emptyset \end{cases}$$

$$\mathcal{A} = \sum_{x_k \in \underline{A}(\beta_i)} h(x_k, x_j); \; \mathcal{B} = \sum_{x_k \in B(\beta_i)} (\mu_{ik})^{\acute{m}} h(x_k, x_j).$$

The main steps of this algorithm proceeds as follows:

1. Assign initial bio-bases v_i, $i = 1, 2, \cdots, c$. Choose values for fuzzifier $\acute{m}$ and thresholds ϵ_1 and ϵ_2. Set the iteration counter $t = 1$.
2. Compute the membership μ_{ij} by (5) for c clusters and n subsequences.

3. If μ_{ij} is maximum for $1 \leq i \leq c$ and $(\mu_{ij} - \mu_{kj}) \leq \epsilon_2$, then $x_j \in \overline{A}(\beta_i)$ and $x_j \in \overline{A}(\beta_k)$. Furthermore, x_j is not part of any lower bound.
4. Otherwise, $x_j \in \underline{A}(\beta_i)$ such that μ_{ij} is the maximum for $1 \leq i \leq c$. In addition, by properties of rough sets, $x_j \in \overline{A}(\beta_i)$.
5. Compute the new bio-basis as per (8).
6. Repeat Steps 2 to 6, by incrementing t, until $|\mu_{ij}(t) - \mu_{ij}(t-1)| > \epsilon_1$.

3.5 Selection of Initial Bio-Basis

A limitation of the c-medoids algorithm is that it can only achieve a local optimum solution that depends on the initial choice of the bio-bases. Consequently, computing resources may be wasted in that some initial bio-bases get stuck in regions of the input space with a scarcity of data points and may therefore never have the chance to move to new locations where they are needed. To overcome this limitation of the c-medoids algorithm, next, we propose a method to select initial bio-bases, which is based on a similarity measure using amino acid mutation matrix. It enables the algorithm to converge to an optimum or near optimum solutions (bio-bases).

Prior to describing the proposed method for selecting initial bio-bases, next, we provide a quantitative measure to evaluate the similarity between two subsequences in terms of nongapped pairwise homology alignment score:

- DOR. The DOR between two subsequences x_i and x_j is defined as

$$\mathrm{DOR}(x_j, x_i) = \frac{h(x_j, x_i)}{h(x_i, x_i)}. \tag{9}$$

  It is the ratio between the nongapped pairwise homology alignment scores of two input subsequences x_i and x_j based on an amino acid mutation matrix to the maximum homology alignment score of the subsequence x_i. It is used to quantify the similarity in terms of homology alignment score between pairs of subsequences. If the functions of two subsequences are different, the DOR between them is small. A high value of $\mathrm{DOR}(x_i, x_j)$ between two subsequences x_i and x_j asserts that they may have the same function statistically. If two subsequences are same, the DOR between them is maximum, that is, $\mathrm{DOR}(x_i, x_i) = 1$. Thus, $0 < \mathrm{DOR}(x_i, x_j) \leq 1$. Also, $\mathrm{DOR}(x_i, x_j) \neq \mathrm{DOR}(x_j, x_i)$.

Based on the concept of DOR, next, we describe the method for selecting initial bio-bases. The main steps of this method proceeds as follows:

1. For each subsequence x_i, calculate $\mathrm{DOR}(x_j, x_i)$ between itself and the subsequence x_j, $\forall_{j=1}^{n}$.
2. Calculate the similarity score between subsequences x_i and x_j:

$$\mathrm{S}(x_j, x_i) = \begin{cases} 1 & \text{if } \mathrm{DOR}(x_j, x_i) > \epsilon_3 \\ 0 & \text{otherwise.} \end{cases}$$

3. For each x_i, calculate total number of similar subsequences of x_i as

$$\mathrm{N}(x_i) = \sum_{j=1}^{n} \mathrm{S}(x_j, x_i).$$

4. Sort n subsequences according to their values of $\mathrm{N}(x_i)$ such that $\mathrm{N}(x_1) > \mathrm{N}(x_2) > \cdots > \mathrm{N}(x_n)$.
5. If $\mathrm{N}(x_i) > \mathrm{N}(x_j)$ and $\mathrm{DOR}(x_j, x_i) > \epsilon_3$, then x_j cannot be considered as a bio-basis, resulting in a reduced set of subsequences to be considered for initial bio-bases.
6. Let there be $\acute{n}$ subsequences in the reduced set having $\mathrm{N}(x_i)$ values such that $\mathrm{N}(x_1) > \mathrm{N}(x_2) > \cdots > \mathrm{N}(x_{\acute{n}})$. A heuristic threshold function can be defined as [12]

$$\mathrm{Tr} = \frac{\mathrm{R}}{\epsilon_4};\ \text{where } \mathrm{R} = \sum_{i=1}^{\acute{n}} \frac{1}{\mathrm{N}(x_i) - \mathrm{N}(x_{i+1})},$$

   where ϵ_4 is a constant (say, $= 0.5$), so that all subsequences in a reduced set having $\mathrm{N}(x_i)$ value higher than it are regarded as the initial bio-bases.

The value of Tr is high if most of the $\mathrm{N}(x_i)$s are large and close to each other. The above condition occurs when a small number of large clusters are present. On the other hand, if the $\mathrm{N}(x_i)$s have wide variation among them, then the number of clusters with smaller size increases. Accordingly, Tr attains a lower value automatically.

Note that the main motive of introducing this threshold function lies in reducing the number of bio-bases. We attempt to eliminate noisy bio-bases (subsequence representatives having lower values of $\mathrm{N}(x_i)$) from the whole subsequences. The whole approach is therefore data dependent.

4 Quantitative Measure

In this section, we propose some quantitative indices to evaluate the quality of selected bio-bases incorporating the concepts of nongapped pairwise homology alignment scores and MI.

4.1 Using Homology Alignment Score

Based on the nongapped pairwise homology alignment scores, next, we introduce two indices—β index and γ index for evaluating quantitatively the quality of selected bio-bases:

- β index. It is defined as

$$\beta = \frac{1}{c} \sum_{i=1}^{c} \frac{1}{n_i} \sum_{x_j \in \beta_i} \frac{h(x_j, v_i)}{h(v_i, v_i)}, \tag{10}$$

  i.e.,

$$\beta = \frac{1}{c} \sum_{i=1}^{c} \frac{1}{n_i} \sum_{x_j \in \beta_i} \mathrm{DOR}(x_j, v_i),$$

  where n_i is the number of subsequences in the ith cluster β_i, and $h(x_j, v_i)$ is the nongapped pairwise homology alignment scores, obtained using an amino acid mutation matrix, between subsequence x_j and

bio-basis v_i. The β index is the average normalized homology alignment scores of input subsequences with respect to their corresponding bio-bases. A good clustering procedure for bio-bases selection should make all input subsequences as similar to their bio-bases as possible. The β index increases with the increase in homology alignment scores within a cluster. Therefore, for a given data set and c value, the higher the homology alignment scores within the clusters, the higher would be the β value. The value of β also increases with c. In an extreme case, when the number of clusters is maximum, that is, $c = n$, the total number of subsequences in the data set, we have $\beta = 1$. Thus, $0 < \beta \leq 1$.

- γ Index. It can be defined as

$$\gamma = \max_{i,j} \frac{1}{2} \left\{ \frac{h(v_j, v_i)}{h(v_i, v_i)} + \frac{h(v_i, v_j)}{h(v_j, v_j)} \right\}, \tag{11}$$

i.e.,

$$\gamma = \max_{i,j} \frac{1}{2} \{ \mathrm{DOR}(v_j, v_i) + \mathrm{DOR}(v_i, v_j) \}$$

$0 < \gamma < 1$. The γ index calculates the maximum normalized homology alignment score between bio-bases. A good clustering procedure for bio-bases selection should make the homology alignment score between all bio-bases as low as possible. The γ index minimizes the between-cluster homology alignment score.

4.2 Using MI

Using the concept of MI, one can measure the within-cluster and between-cluster shared information. In principle, MI is regarded as a nonlinear correlation function and can be used to measure the mutual relation between a bio-basis and the subsequences, as well as the mutual relation, between each pair of bio-bases. It is used to quantify the information shared by two objects. If two independent objects do not share much information, the MI value between them is small. Although two highly nonlinearly correlated objects will demonstrate a high MI value. In the present case, the objects can be the bio-bases and the subsequences.

Based on the MI, the β index would be as follows:

$$\overline{\beta} = \frac{1}{c} \sum_{i=1}^{c} \frac{1}{n_i} \sum_{x_j \in \beta_i} \frac{\mathrm{MI}(x_j, v_i)}{\mathrm{MI}(v_i, v_i)}. \tag{12}$$

$\mathrm{MI}(x_i, x_j)$ is the MI between subsequences x_i and x_j. The mutual information $\mathrm{MI}(x_i, x_j)$ is defined as

$$\mathrm{MI}(x_i, x_j) = \mathrm{H}(x_i) + \mathrm{H}(x_j) - \mathrm{H}(x_i, x_j) \tag{13}$$

with $\mathrm{H}(x_i)$ and $\mathrm{H}(x_j)$ being the entropy of subsequences x_i and x_j, respectively, and $\mathrm{H}(x_i, x_j)$ as their joint entropy. $\mathrm{H}(x_i)$ and $\mathrm{H}(x_i, x_j)$ are defined as

$$\mathrm{H}(x_i) = -\mathrm{p}(x_i) \mathrm{lnp}(x_i) \tag{14}$$

$$\mathrm{H}(x_i, x_j) = -\mathrm{p}(x_i, x_j) \mathrm{lnp}(x_i, x_j). \tag{15}$$

$\mathrm{p}(x_i)$ and $\mathrm{p}(x_i, x_j)$ are the a priori probability of x_i and joint probability of x_i and x_j, respectively. The $\overline{\beta}$ index is the average normalized MI of input subsequences with respect to their corresponding bio-bases. A bio-bases selection procedure should make the shared information between all input subsequences and their bio-bases as high as possible. The $\overline{\beta}$ index increases with the increase in MI within a cluster. Therefore, for a given data set and c value, the higher the MI within the clusters, the higher the $\overline{\beta}$ value would be. The value of $\overline{\beta}$ also increases with c. When $c = n$, $\overline{\beta} = 1$. Thus, $0 < \overline{\beta} \leq 1$.

Similarly, γ index would be

$$\overline{\gamma} = \max_{i,j} \frac{1}{2} \left\{ \frac{\mathrm{MI}(v_i, v_j)}{\mathrm{MI}(v_i, v_i)} + \frac{\mathrm{MI}(v_i, v_j)}{\mathrm{MI}(v_j, v_j)} \right\}. \tag{16}$$

The $\overline{\gamma}$ index calculates the maximum normalized MI between bio-bases. A good clustering procedure for bio-bases selection should make the shared information between all bio-bases as low as possible. The $\overline{\gamma}$ index minimizes the between-cluster MI.

5 Experimental Results

The performance of RFCMdd is compared extensively with that of various other related ones. These involve different combinations of the individual components of the hybrid scheme, as well as other related schemes. The algorithms compared are

1. HCMdd [8],
2. RCMdd,
3. FCMdd [21],
4. the method proposed by Yang and Thomson [7] using MI, and
5. the method proposed by Berry et al. [6] using genetic algorithms and Fisher ratio (GAFR).

All the experiments are implemented in C language and run in a LINUX environment having a machine configuration of Pentium IV, 3.2 GHz, 1 Mbyte cache, and 1 Gbyte of RAM.

5.1 Description of a Data Set

To analyze the performance of proposed method, we have used real data sets of five whole human immunodeficiency virus (HIV) protein sequences, Cai-Chou HIV data set [25], and caspase cleavage protein sequences. The initial bio-bases for different c-medoids algorithms, which represent crude clusters in the nonnumerical sequence space, have been generated by the methodology described in Section 3.5. The Dayhoff amino acid mutation matrix [2], [3], [4] is used to calculate the nongapped pairwise homology alignment score between two subsequences.

5.1.1 Five Whole HIV Protein Sequences

HIV protease belongs to the family of aspartyl proteases, which have been well characterized as proteolytic enzymes. The catalytic component is composed of carboxyl groups from two aspartyl residues located in both NH_2-terminal and COOH-terminal halves of the enzyme molecule in HIV protease [26]. They are strongly substrate selective and cleavage specific, demonstrating their capability of cleaving large virus-specific polypeptides called polyproteins between a specific pair of amino acids. Miller et al. showed

TABLE 1
Five Whole HIV Protein Sequences from NCBI

Accession No	Length	Cleavage Sites at P_1
AAC82593	500	132(MA/CA), 363(CA/p2), 377(p2/NC), 432(NC/p1), 448(p1/p6)
AAG42635	498	132(MA/CA), 363(CA/p2), 376(p2/NC), 430(NC/p1), 446(p1/p6)
AAO40777	500	132(MA/CA), 363(CA/p2), 377(p2/NC), 432(NC/p1), 448(p1/p6)
NP_057849	1435	488(TF/PR), 587(PR/RT), 1027(RT/RH), 1147(RH/IN)
NP_057850	500	132(MA/CA), 363(CA/p2), 377(p2/NC), 432(NC/p1), 448(p1/p6)

TABLE 2
Thirteen Caspase Cleavage Proteins from NCBI

Proteins	Gene	Length	Cleavage sites
O00273	DFFA	331	117(C3), 224(C3)
Q07817	BCL2L1	233	61(C1)
P11862	GAS2	314	279(C1)
P08592	APP	770	672(C6)
P05067	APP	770	672(C6), 739(C3/C6/C8/C9)
Q9JJV8	BCL2	236	64(C3 and C9)
P10415	BCL2	239	34(C3)
O43903	GAS2	313	278(C)
Q12772	SREBF2	1141	468(C3 and C7)
Q13546	RIPK1	671	324(C8)
Q08378	GOLGA3	1498	59(C2), 139(C3), 311(C7)
O60216	RAD21	631	279(C3/C7)
O95155	UBE4B	1302	109(C3/C7), 123(C6)

that the cleavage sites in HIV polyprotein can extend to an octapeptide region [27]. The amino acid residues within this octapeptide region are represented by

$$P_4\text{-}P_3\text{-}P_2\text{-}P_1\text{-}P_{1'}\text{-}P_{2'}\text{-}P_{3'}\text{-}P_{4'},$$

where P_4-P_3-P_2-P_1 is the NH_2-terminal half and $P_{1'}$-$P_{2'}$-$P_{3'}$-$P_{4'}$ is the COOH-terminal half. Their counterparts in the HIV protease are represented by S_4-S_3-S_2-S_1-$S_{1'}$-$S_{2'}$-$S_{3'}$-$S_{4'}$ [28]. The HIV protease cleavage site is exactly between P_1 and $P_{1'}$.

The five whole HIV protein sequences have been downloaded from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov). The accession numbers are AAC82593, AAG42635, AAO40777, NP_057849, and NP_057850. Details of these five sequences are included in Table 1. Note that MA, CA, NC, TF, PR, RT, RH, and IN are matrix protein, capsid protein, nucleocapsid core protein, transframe peptide, protease, reverse transcriptase, RNAse, and integrase, respectively. They are all cleavage products of HIV protease. p1, p2, and p6 are also cleavage products [29]. For instance, 132 (MA/CA) means that the cleavage site is between the residues 132 (P_1) and 133 $(P_{1'})$, and the cleavage split the polyprotein producing two functional proteins: the matrix protein and the capsid protein. The subsequences from each of the five whole protein sequences are obtained through moving a sliding window with eight residues. Once a subsequence is produced, it is considered as functional if there is a cleavage site between P_1-$P_{1'}$; otherwise, it is labeled as nonfunctional. The total number of subsequences with eight residues in AAC82593, AAG42635, AAO40777, NP_057849, and NP_057850 are 493, 491, 493, 1428, and 493, respectively.

5.1.2 Cai-Chou HIV Data Set

In [25], Cai and Chou have described a benchmark data set of HIV. It consists of 114 positive oligopeptides and 248 negative oligopeptides, in total, 362 subsequences with eight residues. The data set has been collected from the University of Exeter, UK.

5.1.3 Caspase Cleavage Data Set

The programmed cell death, also known as apoptosis, is a gene-directed mechanism, which serves to regulate and control both cell death and tissue homeostasis during the development and the maturation of cells. The importance of apoptosis study is that many diseases such as cancer, ischemic damage, and so forth, result from apoptosis malfunction. A family of cysteine proteases called caspases, which are expressed initially in the cell as proenzymes, is the key to apoptosis [30]. As caspase cleavage is the key to programmed cell death, the study of caspase inhibitors could represent effective drugs against some disease where blocking apoptosis is desirable. Without a careful study of caspase cleavage specificity, effective drug design could be difficult.

The 13 protein sequences containing various experimentally determined caspase cleavage sites have been downloaded from NCBI (http://www.ncbi.nih.gov). Table 2 represents the information of these sequences. Ci depicts the $i\mathrm{th}$ caspase. The total number of noncleaved subsequences is about 8,340, whereas the number of cleaved subsequences is only 18. In total, there are 8,358 subsequences with eight residues.

5.2 Example

Consider the data set NP_057849 with sequence length 1,435. The number of subsequences obtained through moving a sliding window with eight residues is 1,428. The parameters generated in the DOR-based initialization method for bio-bases selection are shown in Table 3 only for NP_057849 data, as an example. The values of input parameters used are also presented in Table 3.

The similarity score of each subsequence in the original set and reduced set are shown in Fig. 3. The initial bio-bases for c-$\mathrm{medoids}$ algorithms have been obtained from reduced set using the threshold value of Tr. The initial bio-bases

TABLE 3
DOR-Based Initialization of Bio-Bases for NP_057849

Sequence length = 1435
Number of subsequences $n = 1428$
Value of $\epsilon_3 = 0.75$
Number of subsequences in reduced set $\acute{n} = 223$
Value of $\epsilon_4 = 0.5$; Value of Tr = 35.32
Number of bio-bases $c = 36$
Value of fuzzifier $\acute{m} = 2.0$
Values of $w = 0.7$ and $\tilde{w} = 0.3$
Parameters: $\epsilon_1 = 0.001$ and $\epsilon_2 = 0.2$
Quantitative Measures:
HCMdd: $\beta = 0.643$, $\gamma = 0.751$, $\overline{\beta} = 0.807$, $\overline{\gamma} = 1.000$
RCMdd: $\beta = 0.651$, $\gamma = 0.751$, $\overline{\beta} = 0.822$, $\overline{\gamma} = 1.000$
FCMdd: $\beta = 0.767$, $\gamma = 0.701$, $\overline{\beta} = 0.823$, $\overline{\gamma} = 0.956$
RFCMdd: $\beta = 0.836$, $\gamma = 0.681$, $\overline{\beta} = 0.866$, $\overline{\gamma} = 0.913$

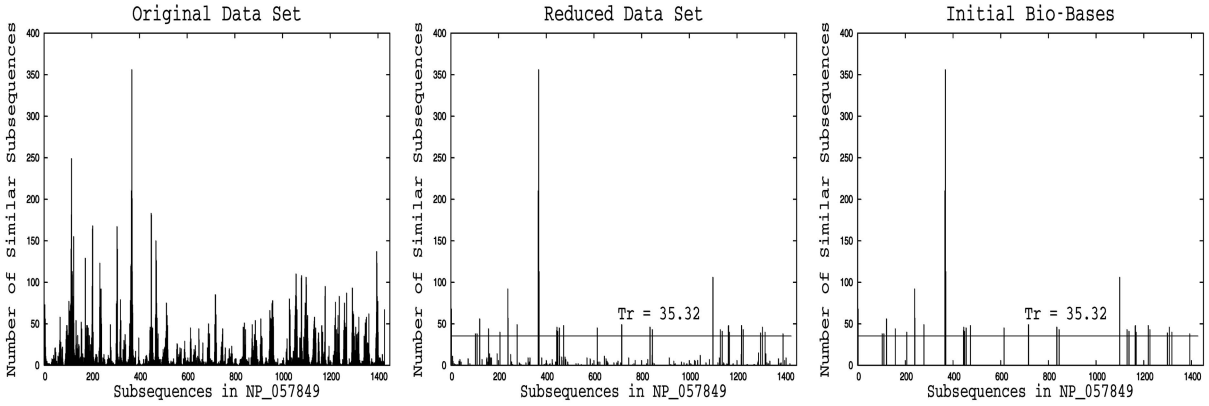


Fig. 3. Similarity scores of subsequences of HIV protein NP_057849 considering $\epsilon_3 = 0.75$ and $\epsilon_4 = 0.50$. (a) Similarity scores in original data set. (b) Similarity scores in reduced data set. (c) Similarity scores of initial bio-bases.

with similarity scores are also shown in Fig. 3. Each c-medoids algorithm has been evolved using these initial bio-bases. The performance obtained by the c-medoids algorithms are shown in Table 3.

5.3 Performance Analysis

The experimental results on the data sets, reported in Section 5.1, are presented in Tables 4, 5, 6, 7, 8, 9, and 10. Subsequent discussions analyze the results presented in these tables with respect to β, γ, $\overline{\beta}$, $\overline{\gamma}$, and execution time.

5.3.1 Optimum Value of Parameter ϵ_3

Table 4 reports the values of β, γ, $\overline{\beta}$, and $\overline{\gamma}$ of different algorithms for the data set NP_057849. Results are presented for different values of ϵ_3. Fig. 4 shows the similarity scores of initial bio-bases as a function of ϵ_3. The parameters generated from the data set NP_057849 are shown in Table 4. The value of c is computed using the method described in Section 3.5. It may be noted that the optimal choice of c is a function of the value ϵ_3. In Fig. 4, it is seen that as the value of ϵ_3 increases, the initial bio-bases,

TABLE 4
Performance of Different Algorithms on NP_057849

Parameters	Algorithms	β	γ	$\overline{\beta}$	$\overline{\gamma}$
	RFCMdd	0.736	0.914	0.817	1.000
$\epsilon_3 = 0.60$	FCMdd	0.719	0.914	0.805	1.000
$\acute{n} = 15$	RCMdd	0.612	0.938	0.805	1.000
Tr = 4.05	HCMdd	0.607	0.938	0.801	1.000
$c = 13$	MI	0.611	0.944	0.813	1.000
	GAFR	0.609	0.962	0.804	1.000
	RFCMdd	0.801	0.821	0.822	1.000
$\epsilon_3 = 0.65$	FCMdd	0.746	0.837	0.811	1.000
$\acute{n} = 34$	RCMdd	0.632	0.836	0.807	1.000
Tr = 6.02	HCMdd	0.618	0.844	0.800	1.000
$c = 26$	MI	0.624	0.913	0.801	1.000
	GAFR	0.616	0.902	0.811	1.000
	RFCMdd	0.801	0.819	0.822	0.982
$\epsilon_3 = 0.70$	FCMdd	0.746	0.828	0.811	0.996
$\acute{n} = 84$	RCMdd	0.635	0.829	0.812	1.000
Tr = 16.58	HCMdd	0.621	0.827	0.803	1.000
$c = 27$	MI	0.625	0.913	0.801	1.000
	GAFR	0.618	0.902	0.810	1.000
	RFCMdd	0.836	0.681	0.866	0.913
$\epsilon_3 = 0.75$	FCMdd	0.767	0.701	0.823	0.956
$\acute{n} = 223$	RCMdd	0.651	0.751	0.822	1.000
Tr = 35.32	HCMdd	0.643	0.751	0.807	1.000
$c = 36$	MI	0.637	0.854	0.802	1.000
	GAFR	0.646	0.872	0.811	1.000
	RFCMdd	0.682	0.937	0.809	1.000
$\epsilon_3 = 0.80$	FCMdd	0.667	0.941	0.805	1.000
$\acute{n} = 594$	RCMdd	0.604	0.941	0.805	1.000
Tr = 28.05	HCMdd	0.605	0.938	0.807	1.000
$c = 6$	MI	0.611	0.938	0.811	1.000
	GAFR	0.608	0.957	0.803	1.000

TABLE 5
Performance of Different c-Medoids Algorithms

Data	Algorithms	Bio-bases	β	γ	$\overline{\beta}$	$\overline{\gamma}$
A	HCMdd	Random	0.615	0.817	0.809	1.000
A		Proposed	0.719	0.702	0.852	1.000
C	FCMdd	Random	0.655	0.791	0.821	1.000
8		Proposed	0.814	0.680	0.901	0.956
2	RCMdd	Random	0.674	0.813	0.825	1.000
5		Proposed	0.815	0.677	0.872	0.983
9	RFCMdd	Random	0.713	0.728	0.847	0.987
3		Proposed	0.874	0.633	0.913	0.916
A	HCMdd	Random	0.657	0.799	0.803	1.000
A		Proposed	0.714	0.664	0.853	1.000
G	FCMdd	Random	0.698	0.706	0.818	1.000
4		Proposed	0.807	0.674	0.892	0.924
2	RCMdd	Random	0.685	0.709	0.812	1.000
6		Proposed	0.768	0.681	0.882	1.000
3	RFCMdd	Random	0.717	0.719	0.847	1.000
5		Proposed	0.831	0.611	0.912	0.957
A	HCMdd	Random	0.651	0.864	0.837	1.000
A		Proposed	0.794	0.723	0.881	1.000
O	FCMdd	Random	0.718	0.804	0.842	1.000
4		Proposed	0.817	0.634	0.912	0.977
0	RCMdd	Random	0.717	0.791	0.847	1.000
7		Proposed	0.809	0.633	0.879	0.977
7	RFCMdd	Random	0.759	0.793	0.890	1.000
7		Proposed	0.856	0.613	0.930	0.947
N	HCMdd	Random	0.601	0.882	0.801	1.000
P		Proposed	0.643	0.751	0.807	1.000
0	FCMdd	Random	0.606	0.802	0.811	1.000
5		Proposed	0.767	0.701	0.823	0.956
7	RCMdd	Random	0.600	0.811	0.801	1.000
8		Proposed	0.651	0.751	0.822	1.000
4	RFCMdd	Random	0.698	0.798	0.804	1.000
9		Proposed	0.836	0.681	0.866	0.913
N	HCMdd	Random	0.611	0.913	0.792	1.000
P		Proposed	0.714	0.719	0.801	1.000
0	FCMdd	Random	0.648	0.881	0.796	1.000
5		Proposed	0.784	0.692	0.886	0.983
7	RCMdd	Random	0.639	0.895	0.794	1.000
8		Proposed	0.758	0.702	0.826	0.993
5	RFCMdd	Random	0.702	0.824	0.803	1.000
0		Proposed	0.851	0.629	0.911	0.928

TABLE 6
Execution Time (ms) of Different c-Medoids Algorithms

Methods	Bio-bases	AAC8 2593	AAG4 2635	AAO4 0777	NP_05 7849	NP_05 7850
RFCMdd	Random	10326	17553	16218	316764	18038
	Proposed	8981	12510	13698	251058	11749
FCMdd	Random	7349	16342	11079	293264	13217
	Proposed	5898	11998	9131	240834	9174
RCMdd	Random	6108	13816	8053	268199	10318
	Proposed	5691	8015	5880	160563	5895
HCMdd	Random	2359	2574	2418	8728	2164
	Proposed	535	534	532	4397	529

which represent the crude clusters, are becoming more prominent. The best result is achieved at $\epsilon_3 = 0.75$. The subsequences selected as initial bio-bases at $\epsilon_3 = 0.75$ have higher values of $\mathrm{N}(x_i)$. For the purpose of comparison, c bio-bases are generated using the methods proposed by Berry et al. (GAFR) and Yang and Thomson (MI).

It is seen from the results in Table 4 that the RFCMdd achieves consistently better performance than other algorithms with respect to the values of β, γ, $\overline{\beta}$, and $\overline{\gamma}$ for different values of ϵ_3. Also, the results reported in Table 4 establish the fact that as the value of ϵ_3 increases, the performance of RFCMdd also increases. The best performance with respect to the values of β, γ, $\overline{\beta}$, and $\overline{\gamma}$ is achieved with $\epsilon_3 = 0.75$. At $\epsilon_3 = 0.75$, the values of $\mathrm{N}(x_i)$ for most of the subsequences in the reduced data set are large and close to each other. Therefore, the threshold Tr attains a higher value compared to that of other values of ϵ_3. In effect, the subsequences selected as initial bio-bases with $\epsilon_3 = 0.75$ have higher values of $\mathrm{N}(x_i)$. Hence, the quality of generated clusters using different c-medoids algorithms is better compared to other values of ϵ_3.

5.3.2 Random versus DOR-Based Initialization

Tables 5 and 6 provide comparative results of different c-medoids algorithms with random initialization of bio-bases and the DOR-based initialization method described in Section 3.5 considering $\epsilon_3 = 0.75$. The DOR-based initialization is found to improve the performance in terms of β, γ, $\overline{\beta}$, and $\overline{\gamma}$ and reduce the time requirement of all c-medoids algorithms. It is also observed that HCMdd with DOR-based initialization performs similar to RFCMdd with random initialization, although it is expected that RFCMdd is superior to HCMdd in partitioning subsequences. Although, in random initialization, the c-medoids algorithms get stuck in local optimums, the DOR-based scheme enables the algorithms to converge to an optimum or near optimum solutions. In effect, the execution time required for different c-medoids is lesser in DOR-based initialization compared to random initialization.

5.3.3 Optimum Values of Parameters $\acute{m}$, w, and ϵ_2

The fuzzifier $\acute{m}$ has an influence on the clustering performance of both RFCMdd and FCMdd. Similarly, the performance of RFCMdd and RCMdd depends on the parameter w and the threshold ϵ_2. Tables 7, 8, and 9 report the performance of different c-medoids algorithms for different values of $\acute{m}$, w, and ϵ_2, respectively. The results and subsequent discussions are presented in these tables with respect to β, γ, $\overline{\beta}$, and $\overline{\gamma}$.

The fuzzifier $\acute{m}$ controls the extent of membership sharing between fuzzy clusters. In Table 7, it is seen that as the value of $\acute{m}$ increases, the values of β and $\overline{\beta}$ increase, whereas γ and $\overline{\gamma}$ decrease. The RFCMdd and FCMdd achieve their best performance with $\acute{m} = 2.0$ for HIV protein NP_057849, $\acute{m} = 1.9$ and 2.0 for Cai-Chou HIV data set, and $\acute{m} = 2.0$ for caspase cleavage protein sequences, respectively. However, for $\acute{m} > 2.0$, the performance of both

TABLE 7
Performance of RFCMdd and FCMdd for Different Values of Fuzzifier $\acute{m}$

Value of $\acute{m}$	Algorithms	HIV Protein NP_057849				Cai-Chou HIV Data Set				Caspase Cleavage Proteins			
		β	γ	$\overline{\beta}$	$\overline{\gamma}$	β	γ	$\overline{\beta}$	$\overline{\gamma}$	β	γ	$\overline{\beta}$	$\overline{\gamma}$
1.5	RFCMdd	0.759	0.744	0.832	0.997	0.755	0.714	0.867	0.992	0.748	0.662	0.882	0.989
	FCMdd	0.699	0.733	0.811	1.000	0.732	0.753	0.824	1.000	0.734	0.678	0.871	1.000
1.6	RFCMdd	0.762	0.717	0.839	0.966	0.781	0.692	0.878	0.979	0.773	0.658	0.899	0.985
	FCMdd	0.716	0.726	0.814	1.000	0.739	0.749	0.833	0.994	0.761	0.677	0.882	1.000
1.7	RFCMdd	0.799	0.702	0.843	0.956	0.794	0.677	0.895	0.950	0.785	0.647	0.907	0.977
	FCMdd	0.725	0.746	0.817	1.000	0.750	0.728	0.868	0.973	0.772	0.671	0.883	0.978
1.8	RFCMdd	0.814	0.695	0.852	0.947	0.818	0.639	0.907	0.932	0.803	0.628	0.923	0.972
	FCMdd	0.738	0.729	0.818	0.985	0.764	0.695	0.890	0.954	0.795	0.671	0.890	0.978
1.9	RFCMdd	0.831	0.681	0.858	0.913	0.829	0.618	0.911	0.927	0.814	0.611	0.937	0.965
	FCMdd	0.755	0.702	0.821	0.972	0.809	0.656	0.903	0.941	0.808	0.668	0.898	0.962
2.0	RFCMdd	0.836	0.681	0.866	0.913	0.829	0.618	0.911	0.927	0.839	0.608	0.942	0.944
	FCMdd	0.767	0.701	0.823	0.956	0.809	0.656	0.903	0.941	0.816	0.662	0.901	0.953
2.1	RFCMdd	0.835	0.684	0.861	0.927	0.811	0.622	0.908	0.945	0.826	0.617	0.935	0.949
	FCMdd	0.754	0.701	0.820	0.956	0.802	0.671	0.901	0.948	0.801	0.665	0.899	0.973
2.2	RFCMdd	0.817	0.699	0.847	0.931	0.802	0.640	0.903	0.958	0.817	0.639	0.928	0.954
	FCMdd	0.751	0.722	0.813	0.978	0.767	0.692	0.892	0.977	0.798	0.665	0.895	0.973
2.3	RFCMdd	0.816	0.712	0.847	0.959	0.791	0.658	0.882	0.961	0.801	0.641	0.901	0.961
	FCMdd	0.734	0.759	0.809	0.991	0.760	0.703	0.877	0.982	0.784	0.668	0.886	0.979
2.4	RFCMdd	0.802	0.712	0.835	0.959	0.774	0.699	0.878	0.967	0.792	0.642	0.894	0.961
	FCMdd	0.712	0.771	0.808	1.000	0.752	0.726	0.876	0.983	0.763	0.672	0.885	0.981
2.5	RFCMdd	0.795	0.751	0.829	0.984	0.767	0.711	0.863	0.967	0.785	0.657	0.887	0.979
	FCMdd	0.701	0.771	0.801	1.000	0.751	0.744	0.854	0.989	0.755	0.673	0.874	0.996

TABLE 8
Performance of RFCMdd and RCMdd for Different Values of Parameter $w\ (= 1 - \tilde{w})$

Value of w	Algorithms	HIV Protein NP_057849				Cai-Chou HIV Data Set				Caspase Cleavage Proteins			
		β	γ	$\overline{\beta}$	$\overline{\gamma}$	β	γ	$\overline{\beta}$	$\overline{\gamma}$	β	γ	$\overline{\beta}$	$\overline{\gamma}$
0.51	RFCMdd	0.632	0.941	0.742	1.000	0.684	0.827	0.806	1.000	0.683	0.714	0.808	1.000
	RCMdd	0.604	0.952	0.719	1.000	0.649	0.856	0.782	1.000	0.668	0.733	0.781	1.000
0.55	RFCMdd	0.637	0.839	0.756	1.000	0.739	0.743	0.837	1.000	0.727	0.688	0.833	1.000
	RCMdd	0.604	0.844	0.729	1.000	0.718	0.807	0.796	1.000	0.714	0.709	0.803	1.000
0.60	RFCMdd	0.648	0.750	0.795	0.983	0.788	0.708	0.883	0.991	0.779	0.649	0.883	0.983
	RCMdd	0.617	0.766	0.746	1.000	0.731	0.733	0.807	1.000	0.762	0.697	0.837	1.000
0.65	RFCMdd	0.817	0.708	0.839	0.934	0.811	0.633	0.902	0.958	0.821	0.622	0.927	0.957
	RCMdd	0.644	0.761	0.807	1.000	0.763	0.694	0.866	0.998	0.774	0.680	0.852	1.000
0.70	RFCMdd	0.836	0.681	0.866	0.913	0.829	0.618	0.911	0.927	0.839	0.608	0.942	0.944
	RCMdd	0.651	0.751	0.822	1.000	0.771	0.677	0.897	0.993	0.782	0.673	0.887	1.000
0.75	RFCMdd	0.819	0.713	0.844	0.940	0.807	0.627	0.899	0.951	0.838	0.611	0.939	0.953
	RCMdd	0.647	0.758	0.806	1.000	0.764	0.698	0.871	1.000	0.771	0.684	0.858	1.000
0.80	RFCMdd	0.766	0.784	0.813	0.992	0.793	0.651	0.874	0.978	0.817	0.622	0.914	0.964
	RCMdd	0.645	0.821	0.792	1.000	0.753	0.704	0.864	1.000	0.753	0.702	0.851	1.000
0.85	RFCMdd	0.713	0.839	0.802	1.000	0.781	0.692	0.853	0.991	0.804	0.647	0.887	1.000
	RCMdd	0.642	0.861	0.785	1.000	0.747	0.718	0.847	1.000	0.736	0.719	0.843	1.000
0.90	RFCMdd	0.648	0.841	0.788	1.000	0.748	0.711	0.829	1.000	0.761	0.682	0.825	1.000
	RCMdd	0.641	0.862	0.781	1.000	0.736	0.727	0.821	1.000	0.708	0.732	0.816	1.000
0.95	RFCMdd	0.639	0.862	0.759	1.000	0.702	0.774	0.818	1.000	0.728	0.711	0.814	1.000
	RCMdd	0.635	0.865	0.761	1.000	0.698	0.781	0.815	1.000	0.681	0.753	0.803	1.000
0.99	RFCMdd	0.602	0.968	0.736	1.000	0.671	0.813	0.802	1.000	0.675	0.762	0.798	1.000
	RCMdd	0.601	0.968	0.736	1.000	0.671	0.813	0.802	1.000	0.674	0.762	0.794	1.000

TABLE 9
Performance of RFCMdd and RCMdd for Different Values of Parameter ϵ_2

Value of ϵ_2	Algorithms	HIV Protein NP_057849				Cai-Chou HIV Data Set				Caspase Cleavage Proteins			
		β	γ	$\overline{\beta}$	$\overline{\gamma}$	β	γ	$\overline{\beta}$	$\overline{\gamma}$	β	γ	$\overline{\beta}$	$\overline{\gamma}$
0.00	RFCMdd	0.643	0.751	0.807	1.000	0.713	0.782	0.817	1.000	0.707	0.698	0.862	1.000
	RCMdd	0.643	0.751	0.807	1.000	0.713	0.782	0.817	1.000	0.707	0.698	0.862	1.000
0.05	RFCMdd	0.704	0.723	0.812	1.000	0.753	0.707	0.868	1.000	0.766	0.683	0.881	1.000
	RCMdd	0.644	0.751	0.810	1.000	0.716	0.754	0.839	1.000	0.723	0.687	0.863	1.000
0.10	RFCMdd	0.793	0.709	0.837	0.981	0.794	0.683	0.882	0.991	0.801	0.641	0.907	0.995
	RCMdd	0.647	0.751	0.814	1.000	0.738	0.726	0.841	1.000	0.738	0.681	0.872	1.000
0.15	RFCMdd	0.811	0.702	0.855	0.946	0.806	0.629	0.902	0.964	0.819	0.622	0.928	0.973
	RCMdd	0.651	0.751	0.819	1.000	0.744	0.694	0.856	1.000	0.764	0.678	0.879	1.000
0.20	RFCMdd	0.836	0.681	0.866	0.913	0.829	0.618	0.911	0.927	0.839	0.608	0.942	0.944
	RCMdd	0.651	0.751	0.822	1.000	0.771	0.677	0.897	0.993	0.782	0.673	0.887	1.000
0.25	RFCMdd	0.836	0.707	0.852	0.936	0.811	0.638	0.907	0.952	0.814	0.631	0.932	0.980
	RCMdd	0.651	0.792	0.819	1.000	0.759	0.698	0.881	1.000	0.767	0.692	0.855	1.000
0.30	RFCMdd	0.817	0.718	0.844	0.990	0.805	0.681	0.894	0.988	0.791	0.667	0.908	0.995
	RCMdd	0.648	0.828	0.801	1.000	0.738	0.731	0.878	1.000	0.741	0.723	0.839	1.000
0.35	RFCMdd	0.801	0.739	0.831	1.000	0.784	0.704	0.875	1.000	0.772	0.671	0.881	1.000
	RCMdd	0.631	0.857	0.779	1.000	0.706	0.757	0.849	1.000	0.728	0.756	0.814	1.000
0.40	RFCMdd	0.792	0.784	0.804	1.000	0.757	0.762	0.872	1.000	0.759	0.699	0.863	1.000
	RCMdd	0.629	0.914	0.758	1.000	0.681	0.796	0.826	1.000	0.719	0.779	0.794	1.000
0.45	RFCMdd	0.716	0.833	0.796	1.000	0.732	0.783	0.850	1.000	0.706	0.737	0.827	1.000
	RCMdd	0.617	0.957	0.792	1.000	0.667	0.817	0.803	1.000	0.678	0.793	0.779	1.000
0.50	RFCMdd	0.708	0.864	0.781	1.000	0.713	0.805	0.813	1.000	0.684	0.798	0.769	1.000
	RCMdd	0.617	0.962	0.781	1.000	0.659	0.836	0.793	1.000	0.659	0.822	0.746	1.000

algorithms decreases with the increase in $\acute{m}$. That is, the best performance of RFCMdd and FCMdd is achieved when the fuzzy membership value of a subsequence in a cluster is equal to its normalized homology alignment score with respect to all the bio-bases.

The parameter w has an influence on the performance of RFCMdd and RCMdd. Since the subsequences lying in lower approximation definitely belong to a cluster, they are assigned a higher weight w compared to $\tilde{w}$ of the subsequences lying in boundary regions. Hence, for both RFCMdd and RCMdd, $0 < \tilde{w} < w < 1$. Table 8 presents the performance of RFCMdd and RCMdd for different values w considering $\acute{m} = 2.0$ and $\epsilon_2 = 0.20$. When the subsequences of both lower approximation and boundary regions are assigned approximately equal weights, the performance of RFCMdd and RCMdd is significantly poorer than HCMdd. As the value of w increases, the values of β and $\overline{\beta}$ increase, whereas the values of γ and $\overline{\gamma}$ decrease. The best performance of both algorithms is achieved with $w = 0.70$. The performance significantly reduces with $w \simeq 1.00$. In this case, since the clusters cannot see the subsequences of boundary regions, the mobility of the clusters and the bio-bases reduces. As a result, some bio-bases get stuck in the local optimum. On the other hand, when the value of

TABLE 10
Comparative Performance of Different Methods

Data Set	Algorithms	β	γ	$\overline{\beta}$	$\overline{\gamma}$	Time
AAC8 2593	RFCMdd	0.847	0.633	0.913	0.916	8981
	FCMdd	0.814	0.680	0.901	0.956	5898
	RCMdd	0.815	0.677	0.872	0.983	5691
	HCMdd	0.719	0.702	0.852	1.000	535
	MI	0.764	0.788	0.906	0.977	8617
	GAFR	0.736	0.814	0.826	1.000	12213
AAG4 2635	RFCMdd	0.831	0.611	0.912	0.957	12510
	FCMdd	0.807	0.674	0.892	0.924	11998
	RCMdd	0.768	0.681	0.882	1.000	8015
	HCMdd	0.714	0.664	0.853	1.000	534
	MI	0.732	0.637	0.829	0.989	13082
	GAFR	0.707	0.713	0.801	1.000	12694
AAO4 0777	RFCMdd	0.856	0.613	0.930	0.947	13698
	FCMdd	0.817	0.634	0.912	0.977	9131
	RCMdd	0.809	0.633	0.879	0.977	5880
	HCMdd	0.794	0.723	0.881	1.000	532
	MI	0.801	0.827	0.890	0.982	12974
	GAFR	0.773	0.912	0.863	1.000	11729
NP_05 7849	RFCMdd	0.836	0.681	0.866	0.913	251058
	FCMdd	0.767	0.701	0.823	0.956	240834
	RCMdd	0.651	0.751	0.822	1.000	160563
	HCMdd	0.643	0.751	0.807	1.000	4397
	MI	0.637	0.854	0.802	1.000	250138
	GAFR	0.646	0.872	0.811	1.000	291413
NP_05 7850	RFCMdd	0.851	0.629	0.911	0.928	11749
	FCMdd	0.784	0.692	0.886	0.983	9174
	RCMdd	0.758	0.702	0.826	0.993	5895
	HCMdd	0.714	0.719	0.801	1.000	529
	MI	0.736	0.829	0.833	1.000	9827
	GAFR	0.741	0.914	0.809	1.000	10873
Cai-Chou HIV Data	RFCMdd	0.829	0.618	0.911	0.927	6217
	FCMdd	0.809	0.656	0.903	0.941	4083
	RCMdd	0.771	0.677	0.897	0.993	3869
	HCMdd	0.713	0.782	0.817	1.000	718
	MI	0.764	0.774	0.890	1.000	6125
	GAFR	0.719	0.794	0.811	1.000	7016
Caspase Cleavage	RFCMdd	0.839	0.608	0.942	0.944	513704
	FCMdd	0.816	0.662	0.901	0.953	510961
	RCMdd	0.782	0.673	0.887	1.000	473380
	HCMdd	0.707	0.698	0.862	1.000	8326
	MI	0.732	0.728	0.869	1.000	511628
	GAFR	0.713	0.715	0.821	1.000	536571

$w = 0.70$, the subsequences of lower approximations are assigned a higher weight compared to that of boundary regions, as well as the clusters, and the bio-bases have a greater degree of freedom to move. In effect, the quality of generated clusters is better compared to other values of w.

The performance of RFCMdd and RCMdd also depends on the value of ϵ_2, which determines the class labels of all the subsequences. In other words, the RFCMdd and RCMdd partition the data set of a cluster into two classes—lower approximation and boundary, based on the value of ϵ_2. Table 9 presents the comparative performance of RFCMdd and RCMdd for different values of ϵ_2 considering $\acute{m} = 2.0$ and $w = 0.70$. For $\epsilon_2 = 0.0$, all the subsequences will be in lower approximations of different clusters and $B(\beta_i) = \emptyset$, $\forall i$. In effect, both RFCMdd and RCMdd reduce to conventional HCMdd. On the other hand, for $\epsilon_2 = 1.0$, $\underline{A}(\beta_i) = \emptyset$, $\forall i$, and all the subsequences will be in the boundary regions of different clusters. That is, the RFCMdd boils down to FCMdd. The best performance of RFCMdd and RCMdd with respect to β, $\overline{\beta}$, γ, and $\overline{\gamma}$ is achieved with $\epsilon_2 = 0.2$, which is approximately equal to the average difference of highest and second highest fuzzy membership values of all the subsequences. In practice, we find that both RFCMdd and RCMdd work well for $\epsilon_2 = \delta$, where

$$\delta = \frac{1}{n}\sum_{j=1}^{n}(\mu_{ij} - \mu_{kj}), \tag{17}$$

n is the total number of subsequences, and μ_{ij} and μ_{kj} are the highest and second highest fuzzy membership values of the subsequence x_j. The values of δ for HIV protein NP_057849, Cai-Chou HIV data set, and caspase cleavage proteins are 0.197, 0.201, and 0.198, respectively.

5.3.4 Comparative Performance of Different Algorithms

Finally, Table 10 provides the comparative results of different algorithms for the protein sequences reported in Section 5.1. It is seen that the RFCMdd with DOR-based initialization produces bio-bases having the highest β and $\overline{\beta}$ values and lowest γ and $\overline{\gamma}$ values for all the cases. Table 10 also provides execution time (in ms) of different algorithms for all protein data sets. The execution time required for RFCMdd is comparable to MI and GAFR. For the HCMdd, although the execution time is less, the performance is significantly poorer than that of RCMdd, FCMdd, and RFCMdd.

The following conclusions can be drawn from the results reported in Tables 4, 5, 6, 7, 8, 9, and 10:

1. It is observed that RFCMdd is superior to HCMdd both with random and DOR-based initialization. However, HCMdd requires considerably less time compared to RFCMdd. However, the performance of HCMdd is significantly poorer than RFCMdd. The performance of FCMdd and RCMdd are intermediate between RFCMdd and HCMdd.
2. The DOR-based initialization is found to improve the values of β, γ, $\overline{\beta}$, and $\overline{\gamma}$ and reduce the time requirement substantially for all c-medoids algorithms.
3. The use of rough sets and fuzzy sets adds a small computational load to the HCMdd algorithm; however, the corresponding integrated methods (FCMdd, RCMdd, and RFCMdd) show a definite increase in β and $\overline{\beta}$ values and decrease in γ and $\overline{\gamma}$ values.
4. Integration of three components—rough sets, fuzzy sets, and c-medoids, in the RFCMdd algorithm produces the minimum set of the most informative bio-bases in the least computation time compared to MI and GAFR.
5. It is observed that the RFCMdd algorithm requires significantly less time compared to MI and GAFR having comparable performance. Reduction in time is achieved due to DOR-based initialization. The DOR-based initialization reduces the convergence time of the RFCMdd algorithm considerably compared to random initialization.

The best performance of the proposed RFCMdd algorithm in terms of β, γ, $\overline{\beta}$, and $\overline{\gamma}$ is achieved due to the following reasons:

1. The DOR-based initialization of bio-bases enables the algorithm to converge to an optimum or near optimum solution.

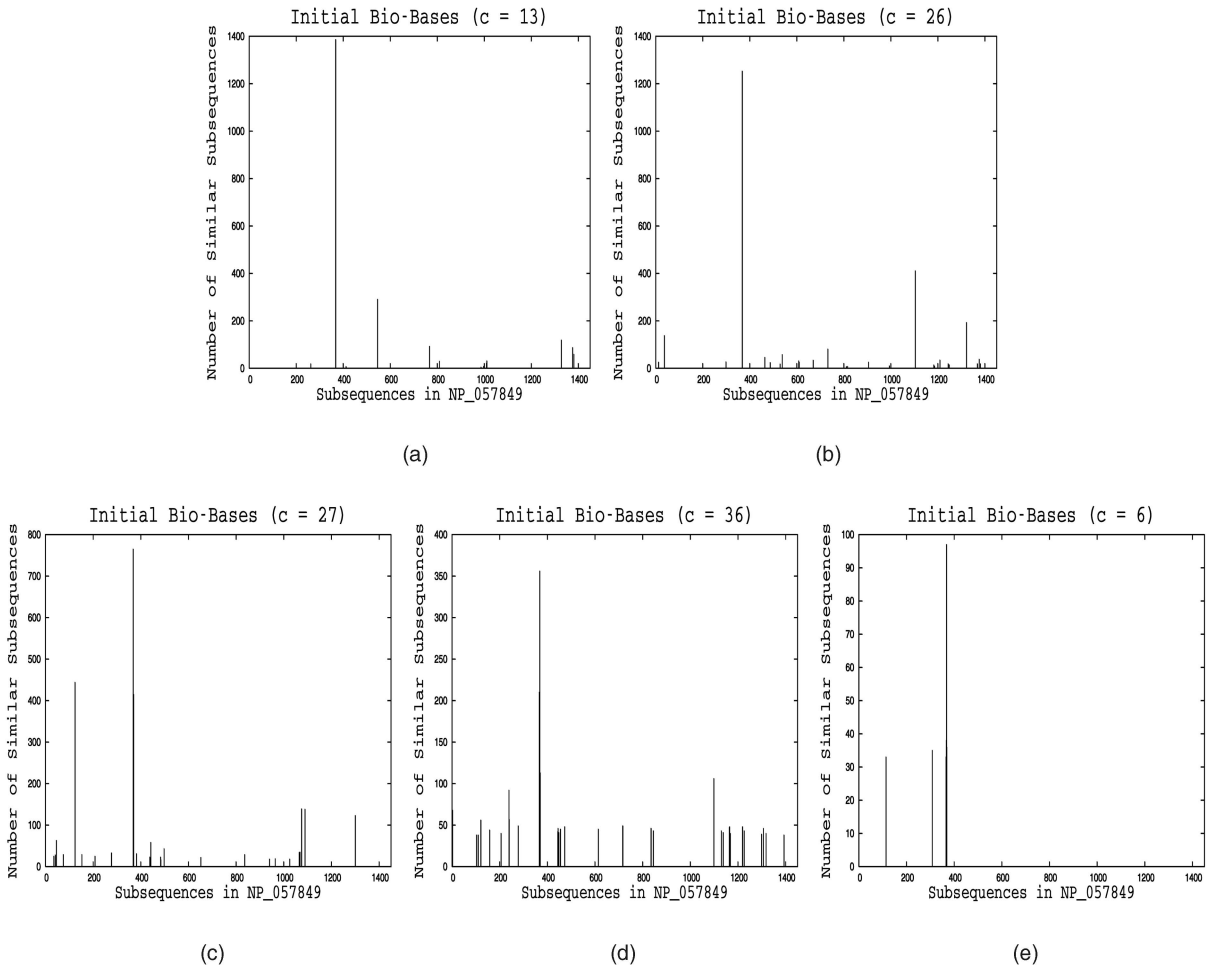


Fig. 4. Similarity scores of initial bio-bases of HIV protein NP_057849 for different values of ϵ_3 considering $\epsilon_4 = 0.50$. (a) $\epsilon_3 = 0.60$ and $\mathrm{Tr} = 4.05$. (b) $\epsilon_3 = 0.65$ and $\mathrm{Tr} = 6.02$. (c) $\epsilon_3 = 0.70$ and $\mathrm{Tr} = 16.58$. (d) $\epsilon_3 = 0.75$ and $\mathrm{Tr} = 35.32$. (e) $\epsilon_3 = 0.80$ and $\mathrm{Tr} = 28.05$.

2. The membership function of fuzzy sets handles efficiently overlapping partitions.
3. The concept of lower and upper bounds of rough sets deals with uncertainty, vagueness, and incompleteness in class definition.

In effect, the minimum set of bio-bases having maximum information is obtained using RFCMdd algorithm.

6 Conclusion

The contribution of the paper is threefold, namely,

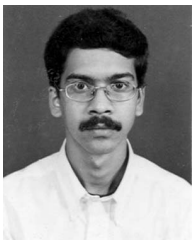
1. the development of a methodology integrating the merits of rough sets, fuzzy sets, c-medoids algorithm, and amino acid mutation matrix for bio-bases selection;
2. defining new measures based on MI and nongapped pairwise homology alignment score to evaluate the quality of selected bio-bases; and
3. demonstrating the effectiveness of the proposed algorithm, along with a comparison with other algorithms, on different types of protein data sets.

The concept of "DOR" is found to be successful in effectively circumventing the initialization and local minima problems of iterative refinement clustering algorithms like c-medoids. In addition, this concept enables efficient selection of the minimum set of the most informative bio-bases compared to existing methods. Although the methodology of integrating rough sets, fuzzy sets, and c-medoids algorithm has been efficiently demonstrated for biological sequence analysis, the concept can be applied to other relational unsupervised classification problems.

References

[1] N. Qian and T.J. Sejnowski, "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models," *J. Molecular Biology*, vol. 202, pp. 865-884, 1988.
[2] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt, "A Model of Evolutionary Change in Proteins. Matrices for Detecting Distant Relationships," *Atlas of Protein Sequence and Structure*, vol. 5, pp. 345-358, 1978.
[3] S. Henikoff and J.G. Henikoff, "Amino Acid Substitution Matrices from Protein Blocks," *Proc. Nat'l Academy of Sciences (PNAS '92)*, vol. 89, pp. 10915-10919, 1992.
[4] M.S. Johnson and J.P. Overington, "A Structural Basis for Sequence Comparisons: An Evaluation of Scoring Methodologies," *J. Molecular Biology*, vol. 233, pp. 716-738, 1993.
[5] 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, pp. 1741-1747, 2003.
[6] 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," *Computational Biology and Chemistry*, vol. 28, pp. 75-85, 2004.

[7] Z.R. Yang and R. Thomson, "Bio-Basis Function Neural Network for Prediction of Protease Cleavage Sites in Proteins," *IEEE Trans. Neural Networks,* vol. 16, no. 1, pp. 263-274, 2005.
[8] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data, An Introduction to Cluster Analysis.* John Wiley & Sons, 1990.
[9] L.A. Zadeh, "Fuzzy Sets," *Information and Control,* vol. 8, pp. 338-353, 1965.
[10] Z. Pawlak, *Rough Sets, Theoretical Aspects of Resoning About Data.* Kluwer Academic Publishers, 1991.
[11] D. Dubois and H. Prade, "Rough Fuzzy Sets and Fuzzy Rough Sets," *Int'l J. General Systems,* vol. 17, pp. 191-209, 1990.
[12] M. Banerjee, S. Mitra, and S.K. Pal, "Rough Fuzzy MLP: Knowledge Encoding and Classification," *IEEE Trans. Neural Networks,* vol. 9, no. 6, pp. 1203-1216, Nov. 1998.
[13] S.K. Pal, S. Mitra, and P. Mitra, "Rough-Fuzzy MLP: Modular Evolution, Rule Generation, and Evaluation," *IEEE Trans. Knowledge and Data Eng.,* vol. 15, no. 1, pp. 14-25, Jan./Feb. 2003.
[14] A. Skowron, R.W. Swiniarski, and P. Synak, "Approximation Spaces and Information Granulation," *Trans. Rough Sets,* vol. 3, pp. 175-189, 2005.
[15] S. Hirano and S. Tsumoto, "An Indiscernibility-Based Clustering Method with Iterative Refinement of Equivalence Relations: Rough Clustering," *J. Advanced Computational Intelligence and Intelligent Informatics,* vol. 7, no. 2, pp. 169-177, 2003.
[16] S.K. De, "A Rough Set Theoretic Approach to Clustering," *Fundamenta Informaticae,* vol. 62, nos. 3-4, pp. 409-417, 2004.
[17] P. Lingras and C. West, "Interval Set Clustering of Web Users with Rough K-Means," *J. Intelligent Information Systems,* vol. 23, no. 1, pp. 5-16, 2004.
[18] S.K. Pal, B.D. Gupta, and P. Mitra, "Rough Self Organizing Map," *Applied Intelligence,* vol. 21, no. 3, pp. 289-299, 2004.
[19] S. Asharaf, S.K. Shevade, and M.N. Murty, "Rough Support Vector Clustering," *Pattern Recognition,* vol. 38, pp. 1779-1783, 2005.
[20] S.K. Pal and P. Mitra, "Multispectral Image Segmentation Using the Rough Set-Initialized-EM Algorithm," *IEEE Trans. Geoscience and Remote Sensing,* vol. 40, no. 11, pp. 2495-2501, 2002.
[21] R. Krishnapuram, A. Joshi, O. Nasraoui, and L. Yi, "Low Complexity Fuzzy Relational Clustering Algorithms for Web Mining," *IEEE Trans. Fuzzy System,* vol. 9, pp. 595-607, 2001.
[22] S. Mitra, H. Banka, and W. Pedrycz, "Rough-Fuzzy Collaborative Clustering," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 36, no. 4, pp. 795-805, Aug. 2006.
[23] S.F. Altschul, W. Gish, W. Miller, E. Myers, and D.J. Lipman, "Basic Local Alignment Search Tool," *J. Molecular Biology,* vol. 215, pp. 403-410, 1990.
[24] S.F. Altschul, M.S. Boguski, W. Gish, and J.C. Wootton, "Issues in Searching Molecular Sequence Databases," *Nature Genetics,* vol. 6, pp. 119-129, 1994.
[25] Y.D. Cai and K.C. Chou, "Artificial Neural Network Model for Predicting HIV Protease Cleavage Sites in Protein," *Advances in Eng. Software,* vol. 29, no. 2, pp. 119-128, 1998.
[26] L.H. Pearl and W.R. Taylor, "A Structural Model for the Retroviral Proteases," *Nature,* vol. 329, pp. 351-354, 1987.
[27] M. Miller, J. Schneider, B.K. Sathayanarayana, M.V. Toth, G.R. Marshall, L. Clawson, L. Selk, S.B.H. Kent, and A. Wlodawer, "Structure of Complex of Synthetic HIV-1 Protease with Substrate-Based Inhibitor at 2.3 Resolution," *Science,* vol. 246, pp. 1149-1152, 1989.
[28] K.C. Chou, "A Vectorised Sequence-Coupling Model for Predicting HIV Protease Cleavage Sites in Proteins," *J. Biological Chemistry,* vol. 268, pp. 16938-16948, 1993.
[29] K.C. Chou, "Prediction of Human Immunodeficiency Virus Protease Cleavage Sites in Proteins," *Analytical Biochemistry,* vol. 233, pp. 1-14, 1996.
[30] T.T. Rohn, S.M. Cusack, S.R. Kessinger, and J.T. Oxford, "Caspase Activation Independent of Cell Death Is Required for Proper Cell Dispersal and Correct Morphology in PC12 Cells," *Experimental Cell Research,* vol. 293, pp. 215-225, 2004.



Pradipta Maji received the BSc (Hons) degree in physics in 1998, the MSc degree in electronics science in 2000, and the PhD degree in computer science in 2005, all from Jadavpur University, India. Currently, he is a lecturer in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. He is also associated with the Center for Soft Computing Research: A National Facility, Indian Statistical Institute, Kolkata, India. His research interests include pattern recognition, bioinformatics, medical image processing, cellular automata, neural networks, soft computing, and so forth. He has published more than 30 papers in international journals and conferences. He is also a reviewer of many international journals.

Sankar K. Pal received the PhD degree in radio physics and electronics from the University of Calcutta in 1979 and another PhD degree in electrical engineering along with a Diploma of Imperial College (DIC) from the University of London in 1982. He is the director and a distinguished scientist at the Indian Statistical Institute. He founded the Machine Intelligence Unit and the Center for Soft Computing Research: A National Facility in the Institute in Calcutta. He worked at the University of California, Berkeley, and the University of Maryland, College Park, in 1986-1987; the NASA Johnson Space Center, Houston, Texas, in 1990-1992 and 1994; and in the US Naval Research Laboratory, Washington, D.C., in 2004. Since 1997, he has been serving as a distinguished visitor of IEEE Computer Society for the Asia-Pacific Region and held several visiting positions at Hong Kong and Australian universities. He is a fellow of the IEEE, the Academy of Sciences for the Developing World, Italy, International Association for Pattern Recognition, USA, and all the four National Academies for Science/Engineering in India. He is the coauthor of 13 books and about 300 research publications in the areas of pattern recognition and machine learning, image processing, data mining and Web intelligence, soft computing, neural nets, genetic algorithms, fuzzy sets, rough sets, and bioinformatics. He has received the 1990 S.S. Bhatnagar Prize (which is the most coveted award for a scientist in India) and many prestigious awards in India and abroad including the 1999 G.D. Birla Award, the 1998 Om Bhasin Award, the 1993 Jawaharlal Nehru Fellowship, the 2000 Khwarizmi International Award from the Islamic Republic of Iran, the 2000-2001 FICCI Award, the 1993 Vikram Sarabhai Research Award, the 1993 NASA Tech Brief Award (USA), the 1994 I*EEE Transactions on Neural Networks* Outstanding Paper Award (USA), the 1995 NASA Patent Application Award (USA), the 1997 IETE-R.L. Wadhwa Gold Medal, the 2001 INSA-S.H. Zaheer Medal, and the 2005-2006 P.C. Mahalanobis Birth Centenary Award (Gold Medal) for Lifetime Achievement. He is an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Neural Networks* (1994-1998, 2003-2006), *Pattern Recognition Letters*, *Neurocomputing* (1995-2005), *Applied Intelligence*, *Information Sciences*, *Fuzzy Sets and Systems*, *Fundamenta Informaticae*, the *International Journal on Pattern Recognition and Artificial Intelligence*, the *International Journal on Computational Intelligence and Applications*, and Proceedings of the Indian National Science Academy (INSA-A); a member of the Executive Advisory Editorial Board of the *IEEE Transactions on Fuzzy Systems*, the *International Journal on Image and Graphics*, and the *International Journal of Approximate Reasoning*; and a guest editor of *Computer*.