Amino Acids (2012) 43:583–594
DOI 10.1007/s00726-011-1106-9

ORIGINAL ARTICLE

Fuzzy clustering of physicochemical and biochemical properties of amino Acids

Indrajit Saha · Ujjwal Maulik · Sanghamitra Bandyopadhyay · Dariusz Plewczynski

Received: 25 May 2011 / Accepted: 23 September 2011 / Published online: 13 October 2011

Abstract In this article, we categorize presently available experimental and theoretical knowledge of various physicochemical and biochemical features of amino acids, as collected in the AAindex database of known 544 amino acid (AA) indices. Previously reported 402 indices were categorized into six groups using hierarchical clustering technique and 142 were left unclustered. However, due to the increasing diversity of the database these indices are overlapping, therefore crisp clustering method may not provide optimal results. Moreover, in various large-scale bioinformatics analyses of whole proteomes, the proper selection of amino acid indices representing their biological significance is crucial for efficient and error-prone encoding of the short functional sequence motifs. In most cases, researchers perform exhaustive manual selection of the most informative indices. These two facts motivated us to analyse the widely used AA indices. The main goal of this article is twofold. First, we present a novel method of partitioning the bioinformatics data using consensus fuzzy clustering, where the recently proposed fuzzy clustering techniques are exploited. Second, we prepare three high quality subsets of all available indices. Superiority of the consensus fuzzy clustering method is demonstrated quantitatively, visually and statistically by comparing it with the previously proposed hierarchical clustered results. The processed AAindex1 database, supplementary material and the software are available at http://sysbio.icm.edu.pl/aaindex/.

Keywords Amino acids · AAindex database · Consensus fuzzy clustering · High-quality indices · Validity measures · Physico-chemical features

Introduction

Amino acids are chemical entities containing an amine group and a carboxylic acid group and a side chain that varies between different amino acids. They form proteins, which are critical to life, and have many important functions in living cells. Twenty naturally occurring amino acids with different physicochemical and biochemical properties are the structural building blocks of proteins. A wide diversity of properties of amino acids have been investigated through a large number of experiments and theoretical studies. Each of these amino acid properties can be represented by a vector of 20 numerical values, and we refer to it as an amino acid index. Nakai et al. (1988) came up with 222 amino acid indices from published literature and investigated the relationships among them using hierarchical clustering analysis. Subsequently, Tomii and Kanehisa (1996) enriched the AAindex database with 42 amino acid mutation matrices and released as the

I. Saha · D. Plewczynski (✉)
Interdisciplinary Centre for Mathematical
and Computational Modelling,
University of Warsaw, 02-106 Warsaw, Poland
e-mail: darman@icm.edu.pl

I. Saha
e-mail: indra@icm.edu.pl

U. Maulik
Department of Computer Science and Engineering,
Jadavpur University, Kolkata 700032, West Bengal, India
e-mail: umaulik@cse.jdvu.ac.in

S. Bandyopadhyay
Machine Intelligence Unit, Indian Statistical Institute,
Kolkata 700108, West Bengal, India
e-mail: sanghami@isical.ac.in

AAindex2. Recently, 47 amino acid contact potential matrices have been reported as AAindex3.[1] The database is continuously updated by Kawashima et al. (1999, 2008) and Kawashima and Kanehisa (2000). Currently, 544 amino acid indices are released in AAindex1 database (see footnote 1). Selection of the minimal/optimal set of amino acid indices for different bioinformatics applications is a difficult task and often involves adhoc/sub-optimal choices. It is, therefore, necessary to group similar indices in clusters and label representative cluster-indices. Moreover, apart from AAindex database, Chou's pseudo amino acid composition (PseAA) (Chou 2001) and its several variants, described in Nanni et al. (2010), have been used as an alternative tool to deal with sequence-related systems and protein-related problems (Chou 2009; Nanni et al. 2010; Georgiou et al. 2009, 2010; Pape et al. 2010; Wang et al. 2010). Clustering (Hartigan 1975; Jain and Dubes 1988; Oliveira and Pedrycz 2007) is a widely used technique in data mining application for discovering patterns in underlying data, which partitions the input space into K regions depending on some similarity/dissimilarity metric, where the value of K may or may not be known a priori. Clustering can be performed in two different modes: (1) crisp and (2) fuzzy. In crisp clustering, the clusters are disjoint and nonoverlapping in nature. Any pattern may belong to one and only one class in this case. In fuzzy clustering, a pattern may belong to all the classes with a certain fuzzy membership grade. Due to the overlapping nature of the AAindex1 database, we decided to work on the field of evolutionary partitional fuzzy clustering methods. Moreover, it has been observed by our recent experimental study that no single method outperforms all others over a wide array of different applications (Plewczynski et al. 2010b). Thus, the consensus of all methods is typically applied to provide the best solution. Therefore, we propose a consensus fuzzy clustering (CFC) technique, which analyzes the AAindex1 database for known and unknown number of clusters by exploiting the capability of recently developed fuzzy clustering techniques. It has also been observed that the index encoding scheme of cluster medoids, used in the fuzzy c-medoids (FCMdd) (Krishnapuram et al. 1999) algorithm, provides better results over real valued encoding scheme of cluster centres as used in fuzzy c-means (FCM) (Bezdek 1981). Thus, the different advanced hybridization forms of FCMdd, like differential evolution-based fuzzy c-medoids (DEFCMdd) (Maulik et al. 2010; Maulik and Saha 2009) clustering, and genetic algorithm-based fuzzy c-medoids (GAFCMdd) (Maulik et al. 2010; Maulik and Saha 2009; Maulik and Bandyopadhyay 2000) clustering algorithms are tested. In the case of finding the optimal number of clusters, automatic differential evolution-based fuzzy clustering (ADEFC) (Maulik and Saha 2010) and variable length genetic algorithm (Bandyopadhyay and Pal 2001)-based fuzzy clustering (VGAFC) (Maulik and Bandyopadhyay 2003) are used, which measure the Xie–Beni (XB) (Xie and Beni 1991) index in fitness computation. Thereafter, the consensus result of all methods is taken by a majority voting procedure. Effectiveness of the proposed method is demonstrated quantitatively and visually. Also Wilcoxon rank sum test (Hollander and Wolfe 1999) is conducted to judge the statistical significance and statbility of clusters found by the proposed method. In bioinformatics research on protein sequences, the AAindex1 database has been used in wide range applications, e.g., prediction of post-translational modification (PTM) sites of proteins (Plewczynski et al. 2008; Basu and Plewczynski 2010), protein subcellular localization (Huanga et al. 2007; Tantoso and Li 2008; Liao et al. 2010; Laurila and Vihinen 2010), immunogenicity of MHC class I binding peptides (Tung and Ho 2007; Tian et al. 2009), protein SUMO modification site (Liu et al. 2007; Lu et al. 2010), coordinated substitutions in multiple alignments of protein sequences (Afonnikov and Kolchanov 2004), HIV protease cleavage site prediction (Ogul 2009; Nanni and Lumini 2009), and many more (Jiang et al. 2009; Liang et al. 2009; Soga et al. 2010; Chen et al. 2010; Pugalenthi et al. 2010). In all these cases, selection of proper amino acid indices is crucial, where this paper also attempts to make a humble contribution. The notable work, available in the literature so far, on clustering of amino acid indices is by Tomii and Kanehisa (1996) and Kawashima et al. (2008). They categorized 402 indices into six groups using hierarchical clustering technique. Those clusters/groups represent alpha and turn propensities, beta propensity, composition, hydrophobicity, physicochemical properties and other properties. However, there is no work available on clustering of the latest AAindex database, consisting of 544 amino acid indices. To address this issue, three standardized sets of high-quality indices (HQIs) are generated in our current work by analyzing the AAindex1 database using a CFC technique.

Description of fuzzy clustering algorithms and validity measures

Algorithms for known number of clusters

Fuzzy c-means clustering

The FCM algorithm (Bezdek 1981) is a widely used technique that uses the principles of fuzzy sets to evolve a partition matrix $U(X)$ while minimizing the measure

[1] http://www.genome.jp/aaindex/.

$$J_m = \sum_{j=1}^{n}\sum_{k=1}^{K} u_{k,j}^{m} D^2(z_k, x_j), \quad 1 \leq m \leq \infty \tag{1}$$

where n is the number of data objects, K represents number of clusters, $u_{k,j}$ is cluster membership of jth point in the kth cluster and m denotes the fuzzy exponent. $D(z_k, x_j)$ denotes the distance of point x_j from the kth cluster centre z_k. FCM algorithm starts with random initial K cluster centres, and then at every iteration, it finds the fuzzy membership of each data points using the following equation:

$$u_{k,i} = \frac{\left(\frac{1}{D(z_k,x_i)}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{K}\left(\frac{1}{D(z_j,x_i)}\right)^{\frac{1}{m-1}}}, \quad \text{for} \quad 1 \leq k \leq K, \quad 1 \leq i \leq n \tag{2}$$

The cluster centres are recomputed using the following equation:

$$z_k = \frac{\sum_{i=1}^{n} u_{k,i}^{m} x_i}{\sum_{i=1}^{n} u_{k,i}^{m}} \quad 1 \leq k \leq K \tag{3}$$

The algorithm terminates when there is no further change in the cluster centres. Finally, each data point is assigned to the cluster to which it has maximum membership.

Fuzzy c-medoids clustering

The FCMdd (Krishnapuram et al. 1999) algorithm is the extension of FCM (Bezdek 1981) algorithm replacing cluster means with cluster medoids. A medoid is defined as follows: Let $V = \{v_1, v_2, \ldots, v_i\}$ be a set of z objects. The medoid of V is an object $O \in V$ such that the sum of distances from O to other objects in V is minimum. The aim of FCMdd algorithm is to cluster the dataset $X = \{x_1, x_2, \ldots, x_n\}$ into K partitions so that the J_m (Eq. 1) is minimized. The FCMdd algorithm also iteratively estimates the partition matrix $U(X)$ followed by computation of new cluster medoids. It starts with random initial C medoids, and then at every iteration it finds the fuzzy membership of each object to every cluster using the Eq. 2. Based on the membership values, the cluster medoids are recomputed as follows:

$$p_k = \operatorname{argmin}_{1 \leq j \leq n} \sum_{i=1}^{n} u_{k,i}^{m} D(x_j, x_i), \quad 1 \leq k \leq K \tag{4}$$

and

$$z_k = x_{pk}, 1 \leq k \leq K \tag{5}$$

The algorithm terminates when there is no significant improvement in J_m value. Finally, assignment of each data point is performed in a manner identical to that of FCM.

Differential evolution-based fuzzy c-medoids clustering

In DEFCMdd (Maulik et al. 2010; Maulik and Saha 2009) clustering, the medoids of the clusters are encoded in the vector. For initializing a vector, C medoids are randomly selected from n data points. The fitness of a vector indicates the degree of goodness of the solution, which is defined by J_m. The objective is, therefore, to minimize the J_m index for achieving proper clustering. Subsequently, the medoids encoded in a vector are updated using Eqs. 3 and 5. The process of mutation and crossover follows Eqs. 6 and 9.

$$\vartheta_k(t+1) = \vartheta_m(t) + F(\vartheta_r(t) - \vartheta_j(t)) \tag{6}$$

Here $\vartheta_m(t)$, $\vartheta_r(t)$ and $\vartheta_j(t)$ are randomly taken vectors from the current population (indicated by t time stamp) with the d dimensions for the mutant vector $\vartheta_k(t+1)$. F is the scaling factor usually $\in$ [0,1]. If the index value of $\vartheta_k(t+1)$ lies beyond the permissible range of $\{1, \ldots, n\}$ then it is scaled using one of the following two operations:

$$\vartheta_k(t+1) - n \tag{7}$$

and

$$\vartheta_k(t+1) + n \tag{8}$$

To increase the diversity of the perturbed parameter vectors, crossover is introduced.

$$U_{jk}(t+1) = \begin{cases} \vartheta_{jk}(t+1) \\ \quad \text{if } \text{rand}_j(0,1) \leq \text{CR} \quad \text{or} \quad j = \text{rand}(k) \\ \vartheta_{jk}(t) \\ \quad \text{if } \text{rand}_j(0,1) > \text{CR} \quad \text{and} \quad j \neq \text{rand}(k) \end{cases} \tag{9}$$

In Eq. (9), rand_j (0,1) is the jth evaluation of a uniform random number generator with outcome $\in$ [0, 1]. CR is the crossover rate $\in$ [0, 1], which has to be determined by the user. rand(k) is a randomly chosen index $\in \{1, 2, \ldots, d\}$, which ensures that $U_k(t+1)$ gets at least one parameter from $\vartheta_k(t+1)$. To make the population for the next generation, the trial vector $U_k(t+1)$ is compared with the target vector $\vartheta_k(t)$ using the greedy criterion. If vector $U_k(t+1)$ yields a better fitness value than $\vartheta_k(t)$, then $U_k(t+1)$ is set to $\vartheta_k(t)$; otherwise, the old value $\vartheta_k(t)$ is retained. The algorithm is terminated after a fixed number of generations. The algorithm is outlined in Fig. 1.

```
Initialization of Population.
(Here each vector encodes index corresponding to K medoids)
Evaluation of Fitness using Eqn. (1).
Repeat
 Mutation
 Crossover
 Evaluation of Fitness using Eqn. (1).
 Selection.
Until (termination criteria are met)
```

Fig. 1 DEFCMdd algorithm

```
Initial Population.
(Here each chromosome encodes index corresponding to K medoids)
Fitness computation using $J_m$.  (see Eqn. (1))
Repeat
  Roulette Wheel Selection.
  Crossover.
  Mutation.
  Fitness Computation.
Until (termination criteria are met)
```

Fig. 2 GAFCMdd algorithm

Genetic algorithm-based fuzzy c-medoids clustering

GA-based fuzzy c-medoids (GAFCMdd) (Maulik et al. 2010; Maulik and Saha 2009; Maulik and Bandyopadhyay 2000) clustering algorithm also uses the same encoding policy as DEFCMdd to represent the vectors. The fitness of each chromosome is computed using Eq. 1. Subsequently, the medoids encoded in a chromosome are also updated using Eqs. 3 and 5. Conventional proportional selection has been implemented on the population. The standard single point crossover is applied stochastically with probability μ_c. Each chromosome also undergoes mutation with a fixed probability μ_m. Termination condition is the same as the other algorithm. The elitism model of GAs has been used, where the best chromosome seen till the current generation is stored in a location within the population. The best chromosome of the last generation provides the solution to the clustering problem. Figure 2 demonstrates the GAFCMdd algorithm.

Algorithms for unknown number of clusters

Automatic differential evolution-based fuzzy clustering

Automatic differential evolution-based fuzzy clustering (Maulik and Saha 2010) has been developed on the framework of differential evolution (DE). The technique uses a masker along with the initial population of DE, which contains 0's and 1's. The value 1 in the masker cell indicates that the encoded medoids in the same position of the vector is valid, otherwise not. Fitness of the each vector is computed by XB index (Xie and Beni 1991). Let $\{z_1, z_2, \ldots, z_K\}$ be the set of K cluster medoids encoded in a vector. The XB index is defined as a function of the ratio of the total variation σ to the minimum separation sep of the clusters. Here σ and sep can be written as

$$\sigma(U, Z; X) = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{k,i}^2 D^2(z_k, x_i), \tag{10}$$

and

$$\text{sep}(Z) = min_{i \neq j} \| z_i - z_j \|^2, \tag{11}$$

where $\|.\|$ is the Euclidean norm, and $D(z_k, x_i)$, as mentioned earlier, is the distance between the pattern x_i and the cluster medoid z_k. The XB index is then define as

$$XB(U, Z; X) = \frac{\sigma(U, Z; X)}{n \times \text{sep}(Z)} \tag{12}$$

Note that when the partitioning is compact and good, value of σ should be low while sep should be high, thereby yielding lower values of the XB index. The objective is, therefore, to minimize the XB index for achieving proper clustering. Moreover, the process of mutation, crossover and selection are the same as in DE, and it terminates after a fixed number of generations.

Variable length genetic algorithm-based fuzzy clustering

A variable string length GA (VGA)-based clustering technique has been developed by Maulik and Bandyopadhyay (2003) on the framework of genetic algorithm (GA), where real valued encoding of cluster centres is used. However, index-encoding of cluster medoids is implemented due to context of this article. The algorithm automatically evolves the number of clusters as well as the partitioning and minimizes the XB cluster validity index. Since the number of clusters is considered to be variable, the string lengths of different chromosomes in the same population are allowed to vary. The selection, crossover and mutation operations are also performed in each generation. Elitism is also incorporated to keep track of the best chromosome obtained so far.

Cluster validity indices

Minkowski Score (Jardine and Sibson 1971), Kappa Index (Cohen 1960) (these indices are applied when true cluster is known), and Silhouette Index ($S(C)$) (Rousseeuw 1987) are used for evaluating the performance of the clustering algorithms.

Minkowski Score

The performances of the clustering algorithms are evaluated in terms of the *Minkowski Score* (MS) (Jardine and Sibson 1971). This is a measure of the quality of a solution given the true clustering. Let T be the "true" solution and S the solution we wish to measure. Denote by n_{11} the number of pairs of elements that are in the same cluster in both S and T. Denote by n_{01} the number of pairs that are in the same cluster only in S, and by n_{10} the number of pairs that are in the same cluster in T. Minkowski Score (MS) is then defined as:

$$\mathrm{MS} = \sqrt{\frac{n_{01} + n_{10}}{n_{11} + n_{10}}} \tag{13}$$

For MS, the optimum score is 0, with lower scores being "better".

Kappa index

The kappa index was developed by Cohen (1960) and used by the medical community as a useful measure of classification accuracy. The score of the method is derived from the contingency table called confusion matrix, where each element C_{ij} is the number of records pertaining to cluster i that have been automatically classified in cluster j. So the diagonal elements correspond to the records that have been correctly classified. Overall accuracy (% correct) and kappa index are computed by Eqs. 14 and 15, respectively.

$$\% \text{ correct} = \frac{\sum_k C_{kk}}{n} \times 100 \tag{14}$$

$$\text{kappa} = \frac{n \sum_k C_{kk} - \sum_k C_{k+} C_{+k}}{n^2 - \sum_k C_{k+} C_{+k}} \tag{15}$$

where $C_{k+} = \sum_j C_{kj}$, $C_{+k} = \sum_i C_{ik}$ and n is the number of data points. Kappa values range from 0 to 1. Higher value of kappa (close to 1) indicates better accuracy.

Silhouette index

Silhouette index (Rousseeuw 1987) reflects the compactness and separation of the clusters. Given a set of n samples $S = \{s_1, s_2, \ldots, s_n\}$ and a clustering of the samples $C = \{C_1, C_2, \ldots, C_K\}$, the silhouette width $S(s_i)$ for each sample s_i belonging to cluster C_j denotes a confidence measure of belongingness, and it is defined as follows:

$$S(s_i) = \frac{b(s_i) - a(s_i)}{\max\{a(s_i), b(s_i)\}} \tag{16}$$

Here $a(s_i)$ denotes the average distance of the sample s_i from the other samples of the cluster to which sample s_i is assigned, and $b(s_i)$ represents the minimum of the average distances of sample s_i from the samples of the clusters $C_l, l = 1, 2, \ldots, K$, and $l \neq j$. The value of $S(s_i)$ lies between -1 and 1. Large value of $S(s_i)$ (approaching 1) indicates that the sample s_i is well clustered. Overall silhouette index $S(C)$ of a clustering C is defined as

$$S(C) = \frac{1}{n} \sum_{i=1}^{n} S(s_i) \tag{17}$$

Greater value of $S(C)$ (approaching 1) indicates that most of the samples are correctly clustered and this, in turn, reflects better clustering solution.

Proposed consensus fuzzy clustering algorithm

We have analyzed the AAindex1 database in two different phases. Phase I explores the capability of fuzzy clustering methods for known number of clusters. Thereafter, a consensus result is obtained through majority vote over all fuzzy clustering methods implemented. These phases are described in greater detail below.

Phase I

In this phase, the results of hierarchical clustering for 402 AA indices have been verified by fuzzy clustering algorithms. For this purpose, different fuzzy clustering techniques, like DEFCMdd clustering, GAFCMdd clustering, FCMdd and FCM have been used. Each of these algorithms is executed separately for producing the fuzzy partitions and then different cluster validity measures have been used for the evaluation of the algorithms. Different steps of this phase are described below as well as highlighted in Fig. 3 by red color.

Step 1: Input the AAindex1 database of 402 indices with the known number of clusters.

Step 2: Execute N number of fuzzy clustering algorithms.

Step 3: Evaluate the performance of N number of fuzzy clustering algorithms using different validity measure to ensure that the number of clusters used in Step1 are stable.

Phase II

This phase is more advanced than the earlier phase. The enhanced AAindex1 database of 544 indices has been used. The database is tested by ADEFC and variable length genetic algorithm-based fuzzy clustering (VGAFC) techniques for finding optimal number of clusters automatically. After that, the earlier fuzzy clustering techniques are used to fix the optimal number of clusters as stable clusters. Finally, the results of all six fuzzy clustering methods are used to create a consensus using majority voting procedure. The phase II of the proposed method is described below and its block diagram is shown in Fig. 3.

Step 1: Use enhanced database of AAindex1 for predicting the number of clusters.

Step 2: Execute M number of automatic fuzzy clustering methods to determine the number of clusters.

Step 3: Evaluate the performance of M number of fuzzy clustering algorithms using different validity measures.

Step 4: Execute N number of fuzzy clustering methods with the predicted number of clusters found in

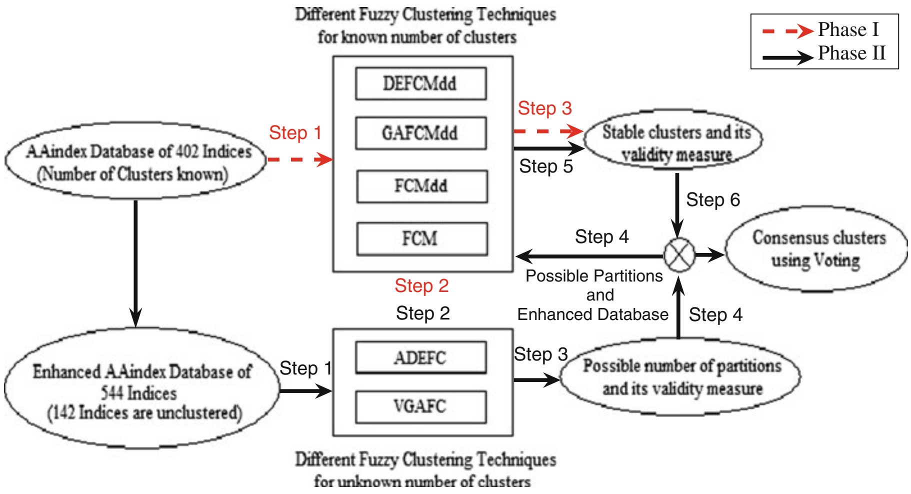

Different Fuzzy Clustering Techniques
for known number of clusters
Phase I
Phase II
DEFCMdd
GAFCMdd
FCMdd
FCM
Step 1
Step 3
Step 5
Stable clusters and its validity measure
AAindex Database of 402 Indices
(Number of Clusters known)
Step 6
Step 4
Consensus clusters using Voting
Possible Partitions
and
Enhanced Database
Step 2
Step 2
Step 4
Enhanced AAindex Database of 544 Indices
(142 Indices are unclustered)
Step 1
ADEFC
VGAFC
Step 3
Possible number of partitions and its validity measure
Different Fuzzy Clustering Techniques
for unknown number of clusters

Fig. 3 Block diagram of the proposed method

Step 2 to ensure that the number of clusters are stable.

Step 5: Repeat the Step 3 of *Phase I* to evaluate the performance of N number of fuzzy clustering algorithms.

Step 6: It is divided into two sub-steps. One is for creating equivalence among all different solutions, and the other for a consensus result among those solutions.

(a) Before creating the consensus clustering result among $M + N$ number of methods, reorganization of data points is required to make them consistent with each other. Thus, the cluster j in the first solution should be equivalent to cluster j in all the other solutions. For example, the solution string $\{aabbccc\}$ is equivalent to $\{bbccaaa\}$. The reorganization is done in such a way that each δ_i, where $i = 2, 3, \ldots, M + N$ and δ_i is a solution string, becomes consistent with δ_1.

(b) Apply consensus method on the label vectors δ_i, $i = 1, 2, \ldots, M + N$ to obtain the final clustering label vector δ. The majority voting is used to create the consensus clustering result, and it is performed as follows: assign each point $k = 1, \ldots, n$ to the cluster j where the label j appears the maximum number of times among all the labels for the point k in all the δ_i.

Experimental results

Description of AAindex1 database

The AAIndex1 currently contains 544 amino acid indices. Each entry consists of an accession number, a short description of the index, the reference information and the numerical values for the properties of 20 amino acids.

Distance measures

The Pearson correlation-based distance measure has been used as this is the commonly used distance metric for clustering AAindex1 database (Tomii and Kanehisa 1996; Kawashima et al. 2008). Given two sample vectors, s_i and s_j, Pearson correlation coefficient Cor(s_i, s_j) between them is computed as:

$$\mathrm{Cor}(s_i, s_j) = \frac{\sum_{l=1}^{p}(s_{il} - \mu_{s_i})(s_{jl} - \mu_{s_j})}{\sqrt{\sum_{l=1}^{p}(s_{il} - \mu_{si})^2}\sqrt{\sum_{l=1}^{p}(s_{jl} - \mu_{sj})^2}}. \qquad (18)$$

Here μ_{s_i} and μ_{s_j} represent the arithmetic means of the components of the sample vectors s_i and s_j, respectively. Pearson correlation coefficient defined in Eq. 18 is a measure of similarity between two samples in the feature space. The distance between two samples s_i and s_j is computed as 1 − mod(Cor(s_i, s_j)), which represents the dissimilarity between those two samples.

Visualization

In this article, for visualization of the datasets, well-known visual assessment of clustering tendency (VAT) representation (Bezdek and Hathaway 2002) is used. To visualize a clustering solution, first the points are reordered according to the class labels given by the solution. Thereafter, the distance matrix is computed on this reordered data matrix. In the graphical plot of the distance matrix, the boxes lying on the main diagonal represent the clustering structure.

Input parameters

The population size and number of generation used for DEFCMdd, GAFCMdd, ADEFC and VGAFC algorithms

Table 1 Average values of cluster validity indices for AAindex1 database of 402 indices in Phase I

Algorithms	No. of clusters known	MS	Kappa	% Correct	S(C)
DEFCMdd		0.5251	0.8102	78.0362	0.3602
GAFCMdd		0.5604	0.7725	74.3663	0.3371
FCMdd	6	0.6507	0.6807	66.7386	0.2462
FCM		0.7263	0.6326	61.2208	0.1883

Table 2 Number of clusters predicted and average values of cluster validity indices for AAindex1 database of 544 indices in Phase II

Algorithms	Predicted no. of clusters	XB	S(C)
ADEFC	8	0.2033	0.5295
VGAFC	8	0.2104	0.5201

are 20 and 100, respectively. The crossover probability (CR) and mutation factors (F) for DEFCMdd and ADEFC are set to 0.8 and 1, respectively. For GAFCMdd and VGAFC, the crossover and mutation probabilities are taken to be 0.8 and 0.3, respectively. The FCMdd and FCM algorithms are executed till it converges to the final solution. Also for the probabilistic/stochastic nature, each algorithm has run for 50 times to show consistency in producing the better results. Note that the input parameters used here are fixed either following the literature or experimentally (Maulik and Bandyopadhyay 2000, 2003; Maulik and Saha 2009; Maulik et al. 2010). The performance of the clustering methods is evaluated by measuring Minkowski Score (MS) (Jardine and Sibson 1971), Kappa index (Cohen 1960), and Silhouette Index (Rousseeuw 1987).

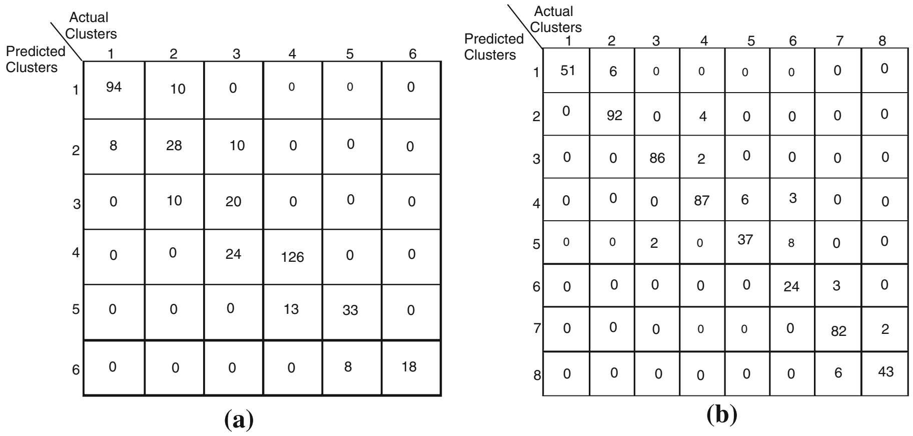
Results and discussion

To analyze the AAindex1 database, different fuzzy clustering algorithms are used in two phases and the average results of 50 consecutive runs of those algorithms are reported in Tables 1, 2, and 3. Here, phase 1 is conducted for the known number of clusters of 402 AA indices. The results are reported in Table 1, which shows the quality of different fuzzy clustering algorithms in terms of cluster validity measures. It is also observed from Table 1 that the DEFCMdd provides better results. However, in phase II, CFC outperforms the others. Hence, Tables 2 and 3 have been designed to show the effectiveness of different fuzzy clustering algorithms. At the beginning of phase II, the enhanced AAindex1 database of 544 indices is examined by ADEFC and VGAFC techniques. The number of clusters found by these two methods is mentioned in Table 2. Table 2 also shows that ADEFC provides better results over VGAFC in terms of validity measures. However, the number of clusters found by both of these algorithms is similar. Thereafter, different fuzzy clustering algorithms (for known number of clusters) are then evaluated by comparing the clustering results of ADEFC and reported in Table 3. Effectiveness of the results is demonstrated by confusion matrix and boxplot in Figs. 4 and 5, respectively. Moreover, for the enhanced database of AAindex1, it has also been observed that the optimal number of clusters is ‘8’ whereas, earlier it was ‘6’ for reduced database of AAindex1. The true clusters plot are shown in Fig. 6 for ADEFC and hierarchical clustered result. It also very clear from Fig. 6 that the ADEFC performs better for producing the optimal number of clusters.

Tables 4, 5 and 6 represent the in-depth analysis of each cluster produced by ADEFC and CFC algorithms, respectively. Table 4 shows that the earlier clustering results have been fragmented into different clusters for ADEFC algorithm, and this observation is also supported by other algorithms in Table 5. For example, in Table 5, the number of AAindex1 indices belonging to cluster 4 are 96, 91, 93, 96, 92, 98 based on seven different algorithms. Moreover, the mapping of the clusters found by CFC is given in Table 6. The name of the clusters is provided based on the mapping of known clusters and predicted clusters, which gives us three new clusters, named as electric properties, residue propensity and intrinsic propensities. These names are given by in-depth study of each AA index. For electric

Table 3 Average values of cluster validity indices for AAindex1 database of 544 indices in Phase II

Algorithms	No. of clusters known (taken from Table 2)	MS	Kappa	% Correct	S(C)
CFC		0.3026	0.9502	94.2551	0.5707
DEFCMdd		0.3307	0.9373	91.4462	0.5386
GAFCMdd	8	0.3844	0.9142	89.0473	0.5011
FCMdd		0.4682	0.8461	81.5506	0.3204
FCM		0.5448	0.7582	73.2731	0.2517



(a)

Predicted Clusters \ Actual Clusters	1	2	3	4	5	6
1	94	10	0	0	0	0
2	8	28	10	0	0	0
3	0	10	20	0	0	0
4	0	0	24	126	0	0
5	0	0	0	13	33	0
6	0	0	0	0	8	18

(b)

Predicted Clusters \ Actual Clusters	1	2	3	4	5	6	7	8
1	51	6	0	0	0	0	0	0
2	0	92	0	4	0	0	0	0
3	0	0	86	2	0	0	0	0
4	0	0	0	87	6	3	0	0
5	0	0	2	0	37	8	0	0
6	0	0	0	0	0	24	3	0
7	0	0	0	0	0	0	82	2
8	0	0	0	0	0	0	6	43

Fig. 4 The best Confusion matrix produced by **a** DEFCMdd for 402 indices, **b** consensus fuzzy clustering for 544 indices, out of 50 runs

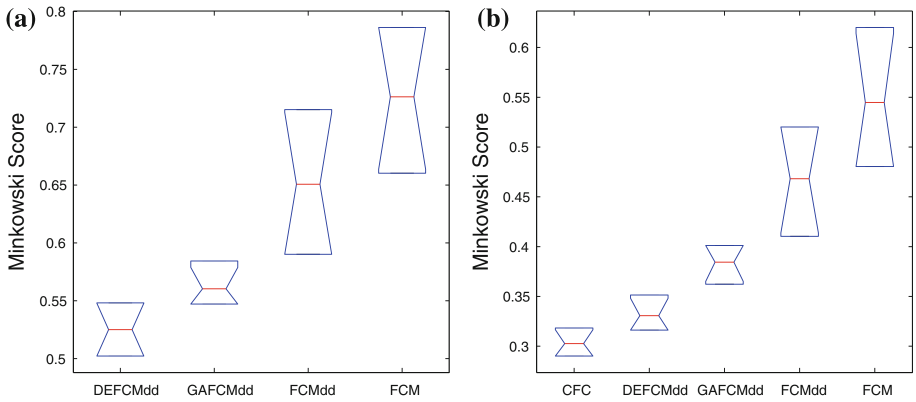


Fig. 5 Boxplot of different clustering algorithms. **a** '6' clusters for 402 indices, **b** '8' clusters for 544 indices, out of 50 runs

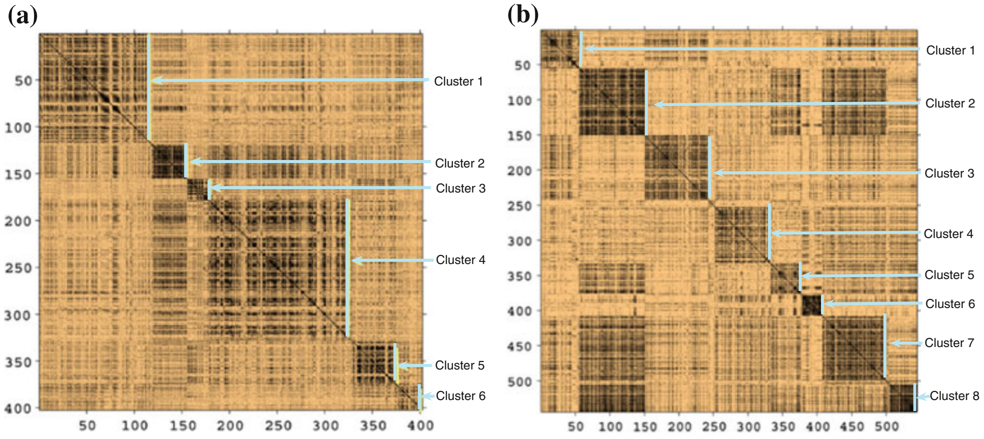


Fig. 6 True clusters plot of AAindex1 database using VAT representation. **a** '6' clusters for 402 indices, **b** '8' clusters for 544 indices found by ADEFC

properties and residue propensity, most of the indices came from original clusters called alpha and turn propensities and hydrophobicity, respectively. The electric properties describe isoelectric point and polarity of amino acid indices, whereas molecular weight, average accessible surface area and mutability are described by residue propensity. However, intrinsic propensities are formed mostly by the unclustered AA indices, and it describes hydration potential, refractivity, optical activity and flexibility. It is also observed that original clusters are fragmented into other clusters to some extent. The cluster called other properties has now been resolved by assigning them in alpha and turn propensities and physicochemical properties. Moreover, names of the current eight clusters are electric properties, hydrophobicity, alpha and turn propensities, physicochemical properties, residue propensity, composition, beta propensity, and intrinsic propensities.

Table 4 Name of the clusters of AAindex1 database

For Fig. 6a			For Fig. 6b		
Clusters	No. of indices	Name of the cluster	Clusters	No. of indices	Name of the cluster
Cluster 1	118	Alpha and turn propensities	Cluster 1	55	Electric properties
Cluster 2	37	Beta propensity	Cluster 2	95	Hydrophobicity
Cluster 3	24	Composition	Cluster 3	92	Alpha and turn propensities
Cluster 4	149	Hydrophobicity	Cluster 4	91	Physicochemical properties
Cluster 5	46	Physicochemical properties	Cluster 5	41	Residue propensity
Cluster 6	28	Other properties	Cluster 6	33	Composition
			Cluster 7	92	Beta propensity
			Cluster 8	45	Intrinsic propensities

Table 5 Number of AAindex1 indices in each cluster for different algorithms

Algorithms	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
CFC	57	96	88	96	47	27	84	49
ADEFC	55	95	92	91	41	33	92	45
VGAFC	52	104	88	93	45	29	88	45
DEFCMdd	57	96	88	96	43	31	84	49
GAFCMdd	56	96	89	92	49	29	82	51
FCMdd	56	105	82	98	43	30	81	49
FCM	54	109	82	82	49	35	79	54

High-quality indices

To provide different subsets of HQIs from the consensus clusters, three different approaches are used. For computing the HQIs 8 (HQI8), medoids (centres) of each cluster are considered and these become AA indices called BLAM930101, BIOV880101, MAXF760101, TSAJ990101, NAKH920108, CEDJ970104, LIFS790101, MIYS990104. Similarly for HQI24 and HQI40, three and five AA indices are considered from each cluster, respectively. For computing HQI24, including the cluster medoids, two other indices farthest from the medoids are taken from each cluster. These two farthest indices are less significant for that cluster, which gives more diversable properties of amino acid to that subset. Similarly for HQI40, including the indices computed in HQI24 for each cluster, other two nearest indices of the medoids are considered, that gives strength to the property of medoids indices. All of these HQIs, HQI8, HQI24 and HQI40 are separately given in the supplementary with their amino acid values. Computational process of HQIs is illustrated by Fig. 7.

Statistical significance test

A non-parametric statistical significance test called Wilcoxon's rank sum test (Hollander and Wolfe 1999) for independent samples has been conducted at the 5% significance level to show that the statistical significance and clusters found by CFC did not arise by chance. For this purpose, results are obtained by comparing pairs of algorithms, in particular, CFC is compared to each four methods (for phase II). For phase I, there are actually only three pairs of comparisons (one less than in phase II), since DEFCMdd is compared to three other methods. Each group consists of the Minkowski Score (MS) produced by 50 consecutive runs of the corresponding algorithm.

To establish that this goodness is statistically significant, Table 7 reports the p values produced by Wilcoxons rank sum test for comparison of two groups (one group corresponding to DEFCMdd and another group corresponding to some other algorithm in phase I and in phase II, one group corresponding to CFC and another group corresponding to some other algorithm) at a time. As a null hypothesis, it is assumed that there is no significant difference between the median values of two groups. Whereas, according to the alternative hypothesis there is a significant difference in the median values of the two groups. The test reflects the stability and reliability of the algorithm. All the p values reported in the table are less than 0.05 (5% significance level). For example, the rank sum test between the algorithms CFC and DEFCMdd in phase II produced a p value of 0.0012, which is very small. This is strong evidence against the null hypothesis, indicating that the better median values of the performance metrics produced by

Table 6 Mapping of consensus clusters

Predicted clusters	Known clusters							Name of the clusters
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7 (unclustered New Indices)	
Cluster 1	30	0	0	3	3	1	20	Electric properties
Cluster 2	0	0	0	72	0	0	24	Hydrophobicity
Cluster 3	62	0	0	0	2	11	13	Alpha and turn propensities
Cluster 4	2	5	4	18	36	12	19	Physicochemical properties
Cluster 5	5	2	8	19	2	0	11	Residue propensity
Cluster 6	4	0	10	0	1	2	10	Composition
Cluster 7	15	30	2	33	2	2	0	Beta propensity
Cluster 8	0	0	0	4	0	0	45	Intrinsic propensities

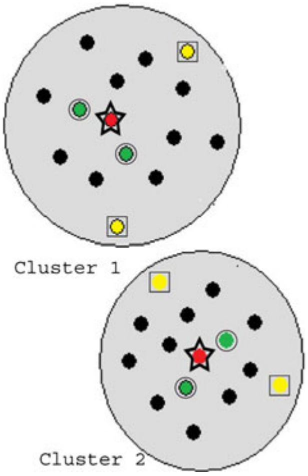


Fig. 7 Illustrated the computational process of HQIs for two clusters, '*star*' points are considered for HQI2, '*star* +*square*' points are considered for HQI6, and '*star* + *square* + *circle*' points are considered for HQI10. In our case, number of clusters is 8, hence, we got HQI8, HQI24 and HQI40

CFC are statistically significant and have not occurred by chance. Similar result is obtained for other case and for all other algorithms compared to CFC, establishing the significant superiority of the proposed method, which also gives that the clusters formed were not the result of random chance.

Conclusion

Summarizing, this article poses two different issues. First, we propose a novel classification method based on fuzzy clustering and second, to provide three subsets of HQIs to the research community from large AAindex1 database. For the first purpose, several recently developed fuzzy clustering techniques are used to analyze the currently released AAindex1 database. We found novel clusters that divide the AAindex1 database on more clear and biologically meaningful way. The novel clusters describe some of the properties of amino acids like isoelectric point, polarity, molecular weight, average accessible surface area, mutability, hydration potential, refractivity, optical activity and flexibility. We also resolved the problem of unknown amino acid indices by assigning them to clusters that have defined biological meaning. Thereafter, majority voting among the all fuzzy clustering methods are taken to create a consensus clusters. After applying the above procedure, we prepared three datasets of HQIs. The first dataset of HQI8 contains eight HQIs, which belongs at the medoids (centres) of each cluster. Similarly, HQI24 and HQI40 contain 24 and 40 indices, respectively. For HQI24, two most less significant indices are taken from each cluster to provide the versatility of subset. However, HQI40 gives the more strength to the medoids indices. These three datasets of HQIs are very effective for machine learning applications of protein sequences, where the short fragments of chains of amino acids can be encoded very easily and effectively. As a scope of further research, developed code of CFC can be used for other bioinformatics applications that utilize amino acids physico-chemical features for machine learning or data mining classification tasks. Representation of amino acids using vector of real numbers can be further explored in physical chemistry, e.g., in computational studies of polymers (Plewczynski et al. 2007; Rodriguez-Soca et al. 2010;

Table 7 *p* values produced by Wilcoxon's rank sum test comparing DEFCMdd and CFC with other algorithms in Phase I and Phase II

Algorithms	Phase I	Phase II
DEFCMdd	–	0.0012
GAFCMdd	0.0019	0.0018
FCMdd	1.4851e−004	1.6083e−004
FCM	1.4851e−005	1.1051e−005

Lu et al. 2007), of selecting inhibitors for a given protein target (Plewczynski et al. 2006, 2010a, b), for PTMs prediction (Plewczynski et al. 2008; Basu and Plewczynski 2010) as well as for finding co-expressed genes (Liu et al. 2008; Kim et al. 2006) in large-scale microarray experiments. The authors are currently working in these directions.

Acknowledgments This work was supported by the Polish Ministry of Education and Science (grants: N301 159735, N518 409238) and Department of Science and Technology, India (Grant No. DST/INT/MEX/RPO-04/2008(ii)). We would also like to give thanks to the reviewer for their valuable comments.

Conflict of interest The authors declare that they have no conflict of interest.

References

Afonnikov DA, Kolchanov AN (2004) CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences. Nucleic Acids Res 32:W64–W68

Bandyopadhyay S, Pal SK (2001) Pixel classification using variable string genetic algorithms with chromosome differentiation. IEEE Trans Geosci Remote Sens 39(2):303–308

Basu S, Plewczynski D (2010) AMS 3.0: prediction of post-translational modifications. BMC Bioinform 11:210

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum, New York

Bezdek JC, Hathaway RJ (2002) VAT: a tool for visual assessment of (cluster) tendency. In: Proceedings of international joint conference on neural netwroks 3:2225–2230

Chen P, Liu C, Burge L, Li J, Mohammad M, Southerland W, Gloster C, Wang B (2010) DomSVR: domain boundary prediction with support vector regression from sequence information alone. Amino Acids 39(3):713–726

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins Struct Funct Genet 43:246–255

Chou KC (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteomics 6(4):262–274

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46

Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to chou's pseudo amino acid composition. J Theor Biol 257(1):17–26

Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2010) A study of entropy/clarity of genetic sequences using metric spaces and fuzzy sets. J Theor Biol 267(1):95–105

Hartigan JA (1975) Clustering algorithms. Wiley, New Jersey

Hollander M, Wolfe DA (1999) Nonparametric statistical methods. 2nd edn

Huanga WL, Tung CW, Huangc HL, Hwang SF, Hob SY (2007) ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. BioSystems 90:573–581

Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Englewood Cliffs

Jardine N, Sibson R (1971) Mathematical taxonomy. John Wiley and Sons, NY

Jiang Y, Iglinski P, Kurgan L (2009) Prediction of protein folding rates from primary sequences using hybrid sequence representation. J Comput Chem 30(5):772–783

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

Kawashima S, Ogata H, Kanehisa M (1999) AAindex: amino acid index database. Nucleic Acids Res 27:368–369

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202–D205

Kim SY, Lee JW, Bae JS (2006) Effect of data normalization on fuzzy clustering of DNA microarray data. BMC Bioinform 7:134

Krishnapuram R, Joshi A, Yi L (1999) A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. In: Proceedings of IEEE International Conference Fuzzy Systems—FUZZ-IEEE 99, pp 1281–1286

Laurila K, Vihinen M (2010) PROlocalizer: integrated web service for protein subcellular localization prediction. Amino Acids (2010, PMID:20811800)

Liang G, Yang L, Kang LY, Mei H, Li Z (2009) Using multidimensional patterns of amino acid attributes for QSAR analysis of peptides. Amino Acids 37(4):583–591

Liao B, Liao B, Sun X, Zeng Q (2010) A novel method for similarity analysis and protein subcellular localization prediction. Bioinformatics 26(21):2678–2683

Liu B, Li S, Wang Y, Lu L, Li Y, Cai Y (2007) Predicting the protein SUMO modification sites based on properties sequential forward selection (PSFS). Biochem Biophys Res Commun 358:136–139

Liu Q, Olman V, Liu H, Ye X, Qiu S, Xu Y (2008) RNACluster: an integrated tool for RNA secondary structure comparison and clustering. J Comput Chem 29(9):1517–1526

Lu L, Shi XH, Li SJ, Xie ZQ, Feng YL, Lu WC, Li YX, Li H, Cai YD (2010) Protein sumoylation sites prediction based on two-stage feature selection. Mol Divers 14:81–86

Lu Y, Bulka B, desJardins M, Freeland SJ (2007) Amino acid quantitative structure property relationship database: a web-based platform for quantitative investigations of amino acids. Protein Eng Des Sel 20:347–351

Maulik U, Bandyopadhyay S (2000) Genetic algorithm based clustering technique. Pattern Recogn 33:1455–1465

Maulik U, Bandyopadhyay S (2003) Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification. IEEE Trans Geosci Remote Sens 41(5):1075–1081

Maulik U, Bandyopadhyay S, Saha I (2010) Integrating clustering and supervised learning for categorical data analysis. IEEE Trans Syst Man Cybern Part A 40(4):664–675

Maulik U, Saha I (2009) Modified differential evolution based fuzzy clustering for pixel classification in remote sensing imagery. Pattern Recogn 42(9):2135–2149

Maulik U, Saha I (2010) Automatic fuzzy clustering using modified differential evolution for image classification. IEEE Trans Geosci Remote Sens 48(9):3503–3510

Nakai K, Kidera A, Kanehisa M (1988) Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng 2:93–100

Nanni L, Lumini A (2009) Using ensemble of classifiers for predicting hiv protease cleavage sites in proteins. Amino Acids 36(3):409–416

Nanni L, Shi JY, Brahnam S, Lumini A (2010) Protein classification using texture descriptors extracted from the protein backbone image. J Theor Biol 264(3):1024–1032

Ogul H (2009) Variable context markov chains for HIV protease cleavage site prediction. BioSystems 96:246–250

Oliveira JV, Pedrycz W (2007) Advances in fuzzy clustering and its applications. John Wiley & Sons, NY

Pape S, Hoffgaard F, Hamacher K (2010) Distance-dependent classification of amino acids by information theory. Proteins Struct Funct Bioform 78(10):2322–2328

Plewczynski D, Lazniewski M, Augustyniak R, Ginalski K (2010a) Can we trust docking results? evaluation of seven commonly used programs on pdbbind database. J Comput Chem 32(4):742–755

Plewczynski D, Lazniewski M, Grotthuss MV, Rychlewski L, Ginalski K (2010b) VoteDock: consensus docking method for prediction of protein-ligand interactions. J Comput Chem 32(4):568–581

Plewczynski D, Slabinski L, Tkacz A, Kajan L, Holm L, Ginalski K, Rychlewski L (2007) The RPSP: web server for prediction of signal peptides. Polymer 48(19):5493–5496

Plewczynski D, Spieser SAH, Koch U (2006) Assessing different classification methods for virtual screening. J Chem Inf Model 46:1098–1106

Plewczynski D, Tkacz A, Rychlewski LS, Ginalski K (2008) AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update. J Mol Model 14(1):69–76

Pugalenthi G, Kandaswamy KK, Suganthan PN, Archunan G, Sowdhamini R (2010) Identification of functionally diverse lipocalin proteins from sequence information using support vector machine. Amino Acids 39(3):777–783

Rodriguez-Soca Y, Munteanu CR, Dorado J, Rabunal J, Pazos A, Gonzalez-Diaz H (2010) Plasmod-PPI: a web-server predicting complex biopolymer targets in plasmodium with entropy measures of protein-protein interactions. Polymer 51(1):264–273

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

Soga S, Kuroda D, Shirai H, Kobori M, Hirayama N (2010) Use of amino acid composition to predict epitope residues of individual antibodies. Protein Eng Des Sel 23:441–448

Tantoso E, Li KB (2008) AAIndexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices. Amino Acids 35(2):345–353

Tian F, Yang L, Lv F, Yang Q, Zhou P (2009) In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure-activity relationship approach. Amino Acids 36(3):535–554

Tomii K, Kanehisa M (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 9:27–36

Tung WC, Ho YS (2007) POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. Bioinformatics 23:942–949

Wang S, Tian F, Qiu Y, Liu X (2010) Bilateral similarity function: a novel and universal method for similarity analysis of biological sequences. J Theor Biol 265(2):194–201

Xie XL, Beni G (1991) A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell 13:841–847