Computers in Biology and Medicine 43 (2013) 889–899

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Keratin protein property based classification of mammals and non-mammals using machine learning techniques

Amit Kumar Banerjee [a], Vadlamani Ravi [b,*], U.S.N. Murty [a], Anirudh P. Shanbhag [a], V. Lakshmi Prasanna [a]

[a] Bioinformatics Group, Biology Division, CSIR-Indian Institute of Chemical Technology, Tarnaka, Uppal Road, Hyderabad 500607, Andhra Pradesh, India
[b] Institute for Development and Research in Banking Technology (IDBRT), Castle Hills Road No 1, Masab Tank, Hyderabad 500057, Andhra Pradesh, India

ARTICLE INFO

Article history:
Received 22 December 2011
Accepted 9 April 2013

Keywords:
Biological classification
Data mining
Support Vector Machines (SVM)
Machine learning
Artificial Intelligence (AI)
Artificial Neural Networks (ANN)
Keratin
Logistic regression
Meta-modeling
Tree induction
Rule induction
Discriminant analysis

ABSTRACT

Keratin protein is ubiquitous in most vertebrates and invertebrates, and has several important cellular and extracellular functions that are related to survival and protection. Keratin function has played a significant role in the natural selection of an organism. Hence, it acts as a marker of evolution. Much information about an organism and its evolution can therefore be obtained by investigating this important protein. In the present study, Keratin sequences were extracted from public data repositories and various important sequential, structural and physicochemical properties were computed and used for preparing the dataset. The dataset containing two classes, namely mammals (Class-1) and non-mammals (Class-0), was prepared, and rigorous classification analysis was performed. To reduce the complexity of the dataset containing 56 parameters and to achieve improved accuracy, feature selection was done using the *t*-statistic. The 20 best features (parameters) were selected for further classification analysis using computational algorithms which included SVM, KNN, Neural Network, Logistic regression, Meta-modeling, Tree Induction, Rule Induction, Discriminant analysis and Bayesian Modeling. Statistical methods were used to evaluate the output. Logistic regression was found to be the most effective algorithm for classification, with greater than 96% accuracy using a 10-fold cross validation analysis. KNN, SVM and Rule Induction algorithms also were found to be efficacious for classification.

1. Introduction

Among different biological paradoxes existing presently, exact and efficient classification of organisms remains top priority. It is of paramount importance due to its pressing need in basic and applied bioscience research. Immense morphological, anatomical and genetic complexity of individual organism has made this problem almost unsolvable since time immemorial. Increase in the interdisciplinary approaches for tackling difficult problems in science, availability of heap of molecular data in the public data repositories and revolution in the existing machine learning methodologies provides us an opportunity to explore this classical issue of biological sciences.

According to Mayr and Bock, biological classification refers to the categorization of the entities in a hierarchical manner where every hierarchy consists of closely related classes [1]. In simple terms, a class is known as a cluster of similar entities, when presence of common traits or attributes in a collection is considered as similar [1]. Linnaeus introduced the concept of biological classification based on common physical features as a means for grouping species [2]. Continuous methodological revisions for grouping species have been performed by the experts including modern molecular phylogenetic techniques to meet the criteria of the Darwinian principles. Till we receive the exact answer, changes in scientific approaches to achieve the same are likely to continue. Understanding each organism based on a complete set of criteria is almost impossible, therefore, an effort is made here to classify mammals and non-mammals taking into account various properties (56) of a single important protein molecule, i.e., Keratin. This particular protein was selected due to its structural and functional intricacies and importance.

Among different structural protein families, Keratin is a significant one. It is fibrous in nature and acts as a structural part of nails, hairs and outer skin layer. Assembled units of Keratin monomers form filament bundles to develop unmineralized tissues in different species. Keratinocytes are rich in filaments of Keratin especially in cornified epidermal layer. Basically, two types of Keratins are found, namely, α- and β-Keratins. Though it is beyond proof at present, it is speculated that different body parts of dinosaurs were composed of various types of Keratins [3]. Based on the intermediate filament, Keratins are of different types, among which polymers of type I and type II intermediate

* Corresponding author. Tel.: +91 40 23534981.
E-mail address: rav_padma@yahoo.com (V. Ravi).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.04.007

filaments are found in some of the chordates. Various non-chordate organisms including nematodes have exclusively type V intermediate filaments [3]. Keratinization also plays a critical role during the programmed cell death process [4]. Replacement and shedding of keratinized epidermal cells is another interesting phenomenon which is thoroughly being studied with relevance to the understanding of regeneration [5]. Structurally, Keratin molecules show interesting diversity which may be due to the involvement of multiple protein coding genes as identified for β-Keratins in feathers and this is probably characteristic of all Keratins [5].

Classification of such a diverse and essential protein is of immense importance which tempted us to select this protein for our study. Probably due to its complexity in structural integrity and diversity in the genes, not much information is available on classification of Keratin, especially, using computational approach.

An attempt was made to understand the chemical relation of the basic amino acids of this protein earlier [6]. Relationship of different types of this protein molecule was studied with respect to the acidic and basic amino acids extensively [7]. Understanding and comparing different types of Keratins and their sequential and structural features requires more sophisticated approaches. Comparative proteomics is increasingly being applied for enriching knowledge in this aspect [8]. This protein is also being used as marker for keratinocyte differentiation [9]. Experimental evidences proved that Keratin is associated with several human diseases, cancer in particular [10–12].

Grouping proteins manually based on their parameters is not only cumbersome but also confusing due to variation and overlapping nature of values of properties. Classification of this protein based on several parameters derived experimentally such as immunoreactivity, isoelectric point and mode of expression has been attempted in the past [13]. Application of monoclonal antibodies for cataloging and characterization of epithelial Keratins in mammals was found to be promising [14]. Strategies have been devised to classify mammalian Keratins based on the presence of high sulfur content [15]. Keeping these isolated efforts aside, the protein under discussion has not been classified extensively either experimentally or theoretically till date. However, Keratin Associated Protein (KRTAP), present in wide group of mammalian species, was subjected for categorization studies in the recent past. The KRTAP family is unique for mammals and several mammalian KRTAP genes had been characterized so far along with gene repertoire in some rodents [16]. Interestingly, humans contain equal number of KRTAP genes as found in different primates besides prominent Keratin related phenotypical differences.

We have adopted advanced bioinformatics and computational classification strategy. Several examples were reported where extensive high-end computational approaches were employed to understand, identify and classify multifaceted biological data including proteins [17–19]. Complex classification exercises were successfully performed on gene, protein, spacer sequences, microarray and disease related data [20–24]. Similarly, different unique and novel approaches were also adopted to understand and classify the datasets. Out of different advanced approaches, Artificial Neural Network (ANN) [23], Radial Basis Function Network (RBFN), Support Vector Machines (SVM) [25,26], Decision rule based approaches [27], Self Organizing Maps (SOM) [28–30], Genetic Programming and GATree [31] have been used meticulously. With time, such studies have also become convoluted owing to availability of mammoth data generated through high throughput experiments. In parallel, advances in computing methodologies have proved helpful in computing numerous parameters theoretically, thus increasing the secondary data pool. Large number of attributes is being considered to obtain accurate and robust output for various types of data categorization. Selection of proper attributes is another major factor. Feature selection techniques are of great help in this regard [32]. Significant variables were identified and sorted out based on their statistical importance to reduce the computational involvedness. Identifying exact and effective algorithm for a particular classification problem is also a tedious process. Therefore, reasonably applying various methodologies and reporting the most efficient ones may be helpful [33].

We have followed a comparative approach available in the RapidMiner platform [34] to understand and classify the Keratin dataset based on the mammalian and non-mammalian origin. Fifty six computed parameters were made part of the analysis to attain the goal, as classifying with less number of parameters may yield less satisfying results with low confidence.

2. Materials and methods

2.1. Sequence retrieval

Complete sequences of Keratin protein were extracted from different public domain databases such as NCBI, Swiss Prot, UniProt, PIR and EMBL. To avoid ambiguity in sequence length, literature was referred and the sequences with length ranging between 301 and 699 amino acids were collected. All partial and other associated sequences such as Keratin associated proteins etc. were eliminated from the initial dataset extracted from individual database. Removal of the repetitive sequences present in different databases was another issue which was handled using standalone protein–protein BLAST (BLASTp). After initial filtering, sequences obtained from one database were checked through local BLAST against the sequences of another database. Output received with 100% identity values, i.e., exact same protein present in the other database with similar or dissimilar annotation was removed and rest of the sequences were added to the main dataset. This process was repeated until all the datasets belonging to different databases were cross validated and all repetitive sequences were removed (Fig. 1). Once the initial dataset was prepared, all the sequences were sorted based on the source organisms and subjected for further analysis.

2.2. Computation of protein features

Significance of protein properties in determining its structure and function is unanimously accepted and vastly reported in the literature. For understanding the classification, we have considered numerous protein physicochemical properties and computed their numerical values. To obtain information from the considered sequences, PROTPARAM and PROTSCALE servers were utilized [35].

All the parameters in the PROTPARAM server were computed which include number of amino acids, molecular weight, theoretical PI, amino acid composition (Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val), total negatively charged amino acids, total positively charged amino acids, atomic composition (carbon, hydrogen, nitrogen, oxygen and sulfur), total atoms, extinction co-efficient, aliphatic index, Grand Average Hydropathicity (GRAVY) and instability index.

In a similar fashion, selected important parameters (sequential and structural) were computed using PROTSCALE. The extracted parameters include number of codon(s), bulkiness, polarity (Zimmerman), refractivity, recognition factors, hydrophobicity (Kyte & Doolittle), transmembrane tendency, buried residues percentage, accessible residues, ratio of hetero end/side, average area buried, average flexibility, alpha-helix (Chou & Fasman), beta-sheet (Chou & Fasman), beta-turn (Chou & Fasman), coil (Deleage & Roux), total

beta-strand, antiparallel beta-strand, parallel beta-strand and relative mutability. As PROTSCALE provides minimum and maximum scale values, therefore, the average scale values were computed (Eq. 1).

$$A_p = (M + m)/2 \quad (1)$$

where A_p is the average parametric value, M and m are the maximum and minimum parameter values respectively.

2.3. Classification analyses

Classification of mammals and non-mammals sequences depending on the Keratin protein sequence derived physicochemical properties was performed before and after feature selection. After performing the feature selection and selecting the top 20 features out of 56 total parameters (variables), a rigorous classification exercise was adopted to identify the best classification methodology for the developed dataset in this study.

2.3.1. Data normalization

All the parameters computed by the tools have different scales, units and dimensions. The ranges of the values were also having vast differences from each other. The final dataset obtained after combining all the generated data was highly heterogeneous in nature. Therefore, transformation of the data to a homogeneous dataset was essential and normalization (Eq. 2) of the final dataset was performed applying the following equation:

$$(N) = \frac{O-m}{M-m} \quad (2)$$

where (N) denotes normalized output, O represents the original data value, M and m refers to the maximum and minimum data values of a parameter respectively. All the normalized values of the

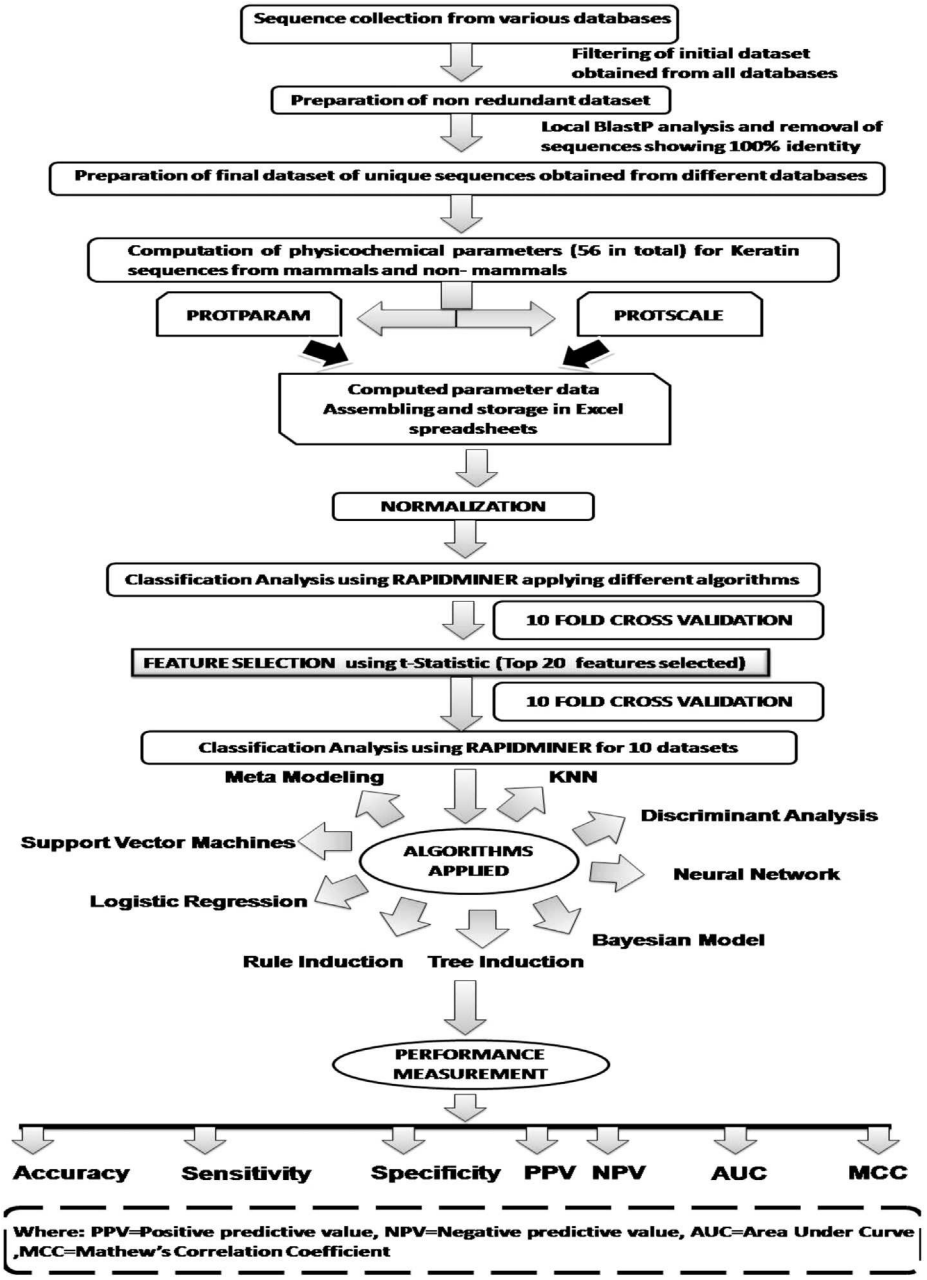

Fig. 1. Workflow of the study adopted for the analysis.

transformed dataset ranged between 0 and 1. Classification exercises were performed on this transformed normalized dataset.

2.3.2. Tool and approaches for classification

RapidMiner was selected as the tool for analysis, the variation in methodologies made this software an interesting platform to extract the best methodology and obtain the classification results. Different techniques available in RapidMiner which include Lazy Modeling (KNN, Default modeling), Tree Induction (Random forest, Decision tree, Random tree and CHAID), Rule Induction, Logistic Regression, Support Vector Machine, Bayesian Modeling, Neural Network, Discriminant analysis and Meta-modeling has been used to fulfill the purpose. The adopted classification strategy is represented in Fig. 1.

Calculation was performed for 10 datasets which were again divided into training and testing sets. For all algorithms applied, the average accuracy values were noted. The data division for test and training was performed using the 10 fold cross validation method.

2.3.3. Algorithms selected for classification

The datasets were analyzed using nine algorithms for classification purposes viz KNN (K Nearest neighbor), Meta-modeling, Neural Networks, SVM (Support Vector Machine), Tree Induction, Rule Induction, Bayesian inference, Logistic regression and Discriminant analysis. Details of the methodologies adopted for the analysis with relevance to this classification exercises are briefly discussed in the following section.

2.3.3.1. KNN (K-Nearest Neighbor method). This method performs classification depending upon the training examples existing in the vicinity of the considered feature space. In this method, each pattern sample is corresponded as a point in a large dimensional feature space. The K-Nearest Neighbor algorithm helps in classifying an object by considering the majority vote contribution of its neighbors. The object under discussion is then assigned to the most common class out of K nearest neighbors where k is a positive integer. When the value of k equals 1, then the neighbor class in proximity is assigned to the object [36].

2.3.3.2. Neural Networks. Basic component of a feed forward neural network is multilayer perceptron where an input and output are mapped through a non-linear function. The prediction method is always divided into training and testing phases which further depend on three major states, namely, input data computation, hidden layer calculation followed by output prediction. The activation state (S_{pj}) (Eq. 3) is achieved by the summed up product of each input vectors (X_{pi}) along with a bias node vector multiplied with the initial random weight values (W_{ij}) in an n-dimensional space.

$$(S_{pj}) = \sum_{i=1}^{N} (W_{ij}X_{pi}) + (W_N + 1, j) \tag{3}$$

where, (S_{pj}) is the activation state, (X_{pi}) is input vectors and (W_{ij}) is initial random weight values, and rest is bias node details.

In the next stage of computation, output for the hidden layers is calculated using sigmoid, Gaussian or hyperbolic functions. Computation of the output for hidden layer neuron (O_{pj}) (Eq. 4) using sigmoid function obeys the following equation where f is known as a non-decreasing function.

$$O_{pj} = f(S_{pj}) = \frac{1}{1 + e^{-S_{pj}}} \tag{4}$$

where, O_{pj} is the output for hidden layer neuron,

During the computation of output layer, similar calculation is performed as above but the input vectors are switched by hidden layer output vectors along with the altered new weights.

To estimate the performance and learning progress, neural network in this case commonly reduces the cost which remains as sum of squared error or mean-squared error that eventually tries to bring down the average squared error between the network output, $f(x)$, and the intended value y for all the example pairs. The cost is minimized using the descent gradient method which is a part of the well-known backpropagation algorithm for training neural networks [23].

2.3.3.3. Support Vector Machines (SVM). A support vector machine develops hyperplane in a large dimensional space which can be used for several important data mining and statistical analysis related to classification and other problems. Logically, the best separation is achieved by the hyperplane containing the largest functional margin distance from the nearest training data points of any class. Statistically, with large margin, low generalization error can be achieved for the classifier [25].

SVM searches for optimal hyperplane for classification of the two-class problem. Support vectors are nothing but combined input points which play role in finding solution for obtaining an optimal hyperplane. Given a set of points $b \in \Re$ with $I = 1 \ldots N$. Each point x_i belongs to either of two classes with the label $y_i \in \{-1, +1\}$,

The set S is linearly separable if there exist $w \in \Re^n$ and $b \in \Re$ such that,

$$\begin{cases} w^T\varphi(X_i) + b \geq +1, & y_i = +1, \\ w^T\varphi(X_i) + b \leq -1 & y_i = -1, \end{cases} \tag{5}$$

which is equivalent to

$$y_i[w^T\varphi(X_i) + b] \geq 1, \quad i = 1, \ldots, N. \tag{6}$$

The input space is mapped into a large dimensional feature space by a non-linear function $\varphi(X)$. A hyperplane ($w^T\varphi(X) + b = 0$) is formed in the developed feature space due to the inequalities which discriminates both the classes for a typical two-dimensional case. By minimizing $w^T w$, the margin between two classes is maximized. Different types of kernel functions $K(,.,)$can be used for any calculation which are linear kernel ($K(X, X_i) = X_i^T X,$), polynomial kernel of d degree ($K(X, X_i) = (1 + X_i^T X/C)^d$), RBF kernel $K(X, X_i) = \exp{-(\|X - X_i\|^2/2\sigma^2)}$ and MLP kernel ($K(X, X_i) = \tanh(\kappa X_i^T X + \theta)$) where $d, c, \sigma, \kappa$and θ are constants.

2.3.3.4. Tree induction. A tree-like model consisting of various decisions along with their all possible consequences containing random event outcomes, different resource costs and plausible utility is considered as tree induction. Major applications of decision trees are observed in operations research and also in different bioinformatics analysis. The decision tree performs recursive partitioning on a given dataset. It involves a chain of branching operations depending on the comparative quantitative information of the given dataset. The overall representation of the algorithm is like computing ordered and dependent (binary) query series and arriving to a final decision.

2.3.3.5. Rule induction. Rule induction is an observation based method where formal rules are generated following a set of observations. Extracted rules represent the overall or partial model of the considered dataset along with respective global or local pattern representation.

2.3.3.6. Logistic regression. Logit function based logistic curve is used to fit the data in consideration for a particular event where

occurrence probability is computed using this methodology. Actually, logistic regression follows the binomial regression and develops a generalized linear model for discrete variables (Eq. 7).

$$\text{Logit}(P) = \ln\frac{P}{(1-P)} = m\alpha + c \tag{7}$$

where P is the probability of occurrence of an event and $(1-P)$ denotes the probability of the same event for not occurring. Therefore, Logit function of the P event becomes equal to the logarithm value of the odds of the event occurring which in turn is linearly proportional to the predictor α and c is a constant. A logistic function is determined by a sigmoid curve proposed by Verhulst.

2.3.3.7. Bayesian inference. Determination of likelihood for a specific hypothesis by applying prior probability over hypotheses where observed evidences are available is known as the Bayesian inference. In other words, the chance of a specific hypothesis of being correct provided some observed evidence arises from the combination of the inherent chances of the hypothesis and its compatibility to the observed evidence in consideration related to the hypothesis (Eq. 8). In this manner, the posterior probability is computed maintaining Baye's rule.

$$P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)} \tag{8}$$

where, probability of any selected, hypothesis A is computed when observed evidence B is given, provided that B is not a part of prior probability computation. This $P(A|B)$ is the posterior probability when the hypothesis A has been chosen after the occurrence of B observed evidences. Therefore, the probability of the hypothesis A given the B observed evidences equals to the likelihood of B given A multiplied by the prior probability, i.e., occurrence of the hypothesis before B is observed ($P(A)$) upon the marginal likelihood ($P(B)$).

2.3.3.8. Discriminant analysis. Classifying a group of observations into some predesignated classes is denoted as discriminant analysis. A group of predictors or input variables is used to decide and designate the class for an observation. For known classes, models are developed based on some set of observations or training datasets. A group of linear functions is developed from the available predictors depending on the kind of training data. These linear functions are designated as discriminant functions which are employed in finding the linear combination of features for isolating two or more objects.

2.3.3.9. Meta-modeling. For a particular dataset having predefined class or classes, different combinatorial approaches are adopted such as framing, rule and constraint generation, model development etc. This logically ordered concept development for modeling purpose under the restricted domain condition to attain a better classification is known as Meta-modeling. In other words, a Meta-model is a sub-abstraction of a developed model which extracts and represents the important properties of a model.

2.4. Feature selection

It is a known fact that augmentation in the number of features increases the complication in classification. Reduction in feature space dimension while keeping the contribution of individual feature intact statistically will aid in better, robust and improved result. Feature selection techniques aid enormously in such case. In this study, initially we have considered total 56 parameters (35 from PROTPARAM and 21 from PROTSCALE). Feature selection was performed for the considered dataset using t-statistic (Eqs. 9 and 10) for unequal sample size and unequal variance.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{{}^{S}\overline{X}_1 - \overline{X}_2} \tag{9}$$

where

$${}^{S}\overline{X}_1 - \overline{X}_2 = \sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}} \tag{10}$$

In this case, s^2 is the estimator of the variance which is unbiased in nature and computed for the two samples where n represents the participants number for group one (1) and group two (2) respectively. Depending on the t-statistic based feature selection exercise, top 20 features were considered for further analysis i.e number of codon(s), total number of atoms, average area buried, molecular weight, theoretical PI, composition of amino acids like Alanine, Aspergine, Cystine, Glutamine, Glutamate, Leucine, Lysine, Methionine, Phenylalanine, Serine, Valine and elements like Hydrogen, Nitrogen, Oxygen and Sulfur.

2.5. Statistical analysis of output

Results of basic statistical analysis provide interesting insight for understanding the behavior of the features along with different stochastic measures which cumulatively aided in reaching the conclusion. Applied measures are discussed below.

2.5.1. Calculation of the accuracy

Analysis of the obtained accuracy value is essential to understand the performance of the algorithms. Therefore, percentage accuracy (Eq. 11) was computed for each case and compared to find out the best algorithm suitable for the considered dataset.

Formula used for accuracy computation is as follows:

$$\text{Acc} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{11}$$

where Acc denotes the accuracy values and TP, TN, FP and FN represents true positive, true negative, false positive and false negative values respectively.

2.5.2. Computations of sensitivity and specificity

After feature selection and performing the classification exercises, the sensitivity and specificity (Eqs. 12 and 13) values computed from the generated confusion matrix for the obtained results.

Sensitivity and specificity were computed employing the following formulae:

$$S_n = (\text{TP})/(\text{TP} + \text{FN}) \tag{12}$$

$$S_p = (\text{TN})/(\text{TN} + \text{FP}) \tag{13}$$

where, S_n and S_p denotes sensitivity and specificity and TP, FN, TN, FP refer true positive, false negative, true negative and false positive respectively.

2.5.3. Computation of positive predictive value and negative predictive value

To obtain an insight about the accuracy of the output, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) (Eqs. 14 and 15) were computed for each output from the confusion matrix. The following formulae were used to obtain the result:

$$\text{PPV} = (\text{TP})/(\text{TP} + \text{FP}) \tag{14}$$

$$\text{NPV} = (\text{TN})/(\text{TN} + \text{FN}) \tag{15}$$

where TP,FP,TN and FN represents true positive, false positive, true negative and false negative values of the confusion matrix during calculation.

Area Under Curve (AUC) (Eq. 16) reveals the importance of a particular methodology by assessing its sensitivity and specificity together. More the area under the curve, better the result. We have computed the AUC values directly from the sensitivity and specificity output applying the following formula:

$$\mathrm{AUC} = 50 \times (\mathrm{Sensitivity} + \mathrm{Specificity}) \tag{16}$$

where, AUC is the Area Under the Curve which is equal to the summed up value of sensitivity and specificity multiplied by a constant.

2.5.4. Computation of Matthew's correlation coefficient

Application of different algorithms, analyzing and comparing their performances is a quite cumbersome task. Matthew's Correlation Coefficient (Eq. 17) is an important measure which represents the single correlation value of the whole confusion matrix. It is exhaustively used for the performance measurement of the unbalanced observations, such as, the one used in this study where the mammals and non-mammals observations vary in terms of number of sequences. The measure is performed using the following equation:

$$\mathrm{MCC} = \frac{(\mathrm{TP}.\mathrm{TN}) - (\mathrm{FP}.\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{17}$$

In each step of filtering, all the prepared feature sets were compiled in Excel spreadsheets till normalization. The final input for RapidMiner was provided as comma separated data file (.csv) format. All the output files such as confusion matrix obtained from RapidMiner were exported to excel spreadsheets for detailed analysis of performance evaluation. All required computations were performed in Quad Core Windows workstations with sufficient storage memory and RAM.

3. Results and discussion

Biological data analysis probably demands maximum attention in the present era due to its intricacy, diversity and basic importance. Extracting novel information in terms of relations, correlations and influences are now possible through data mining techniques. Classification of heterogeneous dataset containing multiple vital attributes with higher statistical values and simultaneously maintaining the biological importance is interesting. The obtained output for this study is described stepwise in the following section in detail.

3.1. Collected sequence information

394 Keratin sequences having the length range of 301–699 were selected. Some classes such as *Protozoa, Annelida*, *Arthropoda*, etc. were eliminated due to less sample numbers and difference in sequence lengths. The total non-mammalian sequences used in the study were 142 while mammalian sequences were 252 in number. The distribution of the sequences is shown in Fig. 2.

3.2. Analysis of amino acid

Calculation of the percentage of amino acids in the considered Keratin protein sequences in both mammals and non-mammals revealed that the mammals contain more polar amino acids except Asparagine which is found to be slightly higher (about 0.113%) in non-mammals (Figs. 3 and 4).

The analysis showed that non-polar amino acids are more in the Keratin sequences of the non-mammals except Alanine and Leucine which were 0.297% and 1.97% higher than mammals. The average distribution of protein length for each organism and other vital parameters is shown in Fig. 3 and in Supplementary materials (Fig. 1). The amino acid composition was observed to be constant amongst primates and non-primates however in Prototherians, it was found to be varying significantly.

3.3. Classification analysis

Following the aim of the project, classification of the mammals and non-mammals group of organisms based on the computed protein physicochemical property values was performed using most of the possible methodologies available for classification. All the sub-options in every algorithm were tested such as different kernel types like dot, radial, polynomial, neural, anova, epachnenikov, Gaussian combination, multiquadric, different measures such as mixed, nominal and numeric measures, iteration numbers and kernel numbers. Output obtained after feature selection for 10 datasets is provided in supplementary Table 1 where accuracy of the best obtained methodology with their most efficient sub-option is reported. The testing was performed for dataset which is out of the training set considered for better generalization. To find out the best algorithm for such classification study, the strategy adopted (shown in Fig. 1) yielded better results after feature selection.

3.3.1. Feature selection

As mentioned in the methodology section, we have adopted a *t*-statistic based computation for selecting out the important features out of the computed 56 features considered for the initial study. The selected features based on the computations shown in Eqs. (9) and (10) are shown in Table 1. Best 20 features were selected for further classification analysis.

Interestingly, feature selection analysis resulted in some important parameters which are very essential for structural and functional integrity of a protein molecule.

3.3.2. Obtained classification results

The algorithms considered in RapidMiner were KNN, Bayesian Modeling, Tree Induction, Rule Induction, Neural Network, Logistic Regression, Support Vector Machines, Discriminant analysis and Meta-modeling.

The classification exercises were performed applying all possible options available under the individual method. For instance, during KNN modeling, *K* values were altered and the best obtained results were reported. In a similar fashion, quadratic discriminant analyses were tried under discriminant analysis. Naive based kernel with greedy approaches was adopted for Bayesian modeling. Gain ratios, different size of splits were also tried in ID3 algorithm for Tree Induction methodologies. Different types of rule inductions were adopted for Rule Induction methodologies. Various perceptrons were experimented for Neural Network approach. Altered experiments were performed with different kernel types for Logistic regression analysis and Support Vector Machine analysis. In continuation, various available options were considered for Meta-modeling. The best obtained outputs and their statistics are reported in the following section.

The overall strategy for classification analysis was performed following the 10 fold cross validation method of data division and statistical measures were performed for the obtained results from

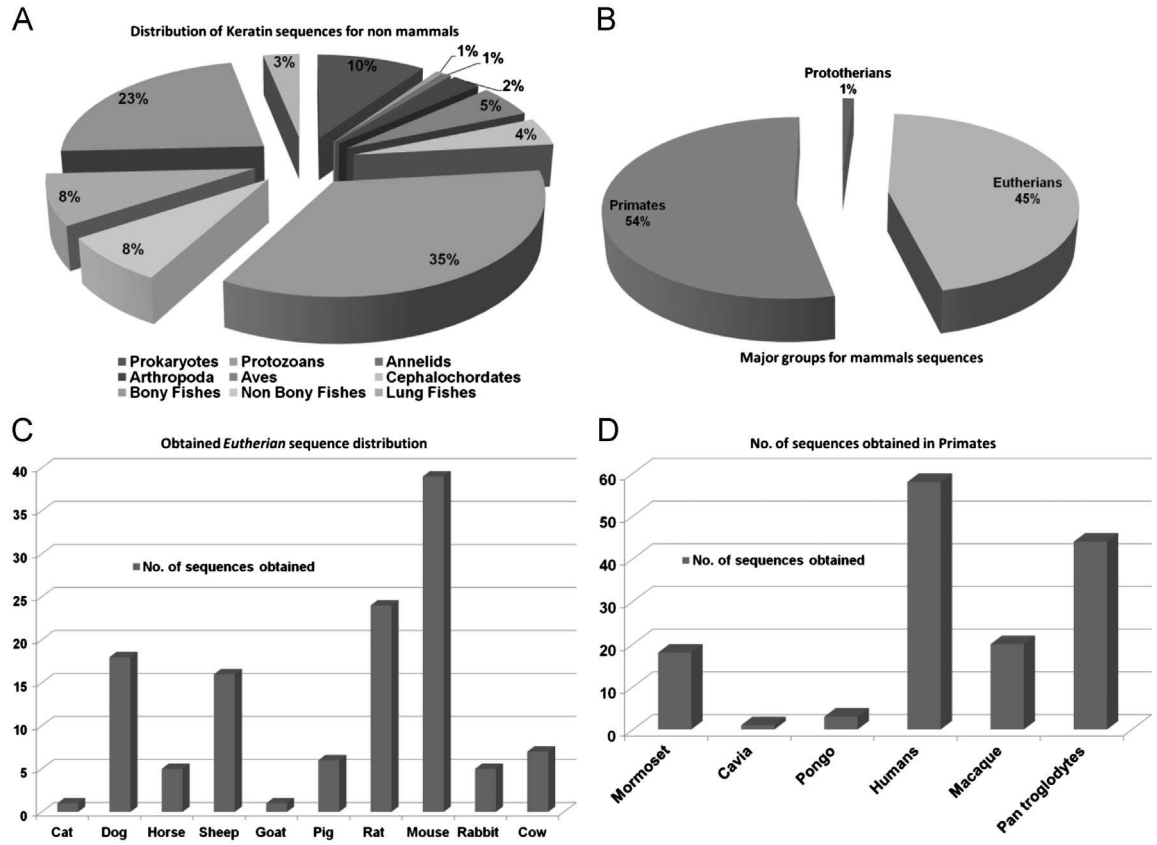

Fig. 2. Graphs depicting the sequences collected. (A) Distribution of non-mammals sequence collection. (B) Mammals sequence collection distribution. (C) Obtained Eutherion sequence distribution. (D) Primate sequences considered.

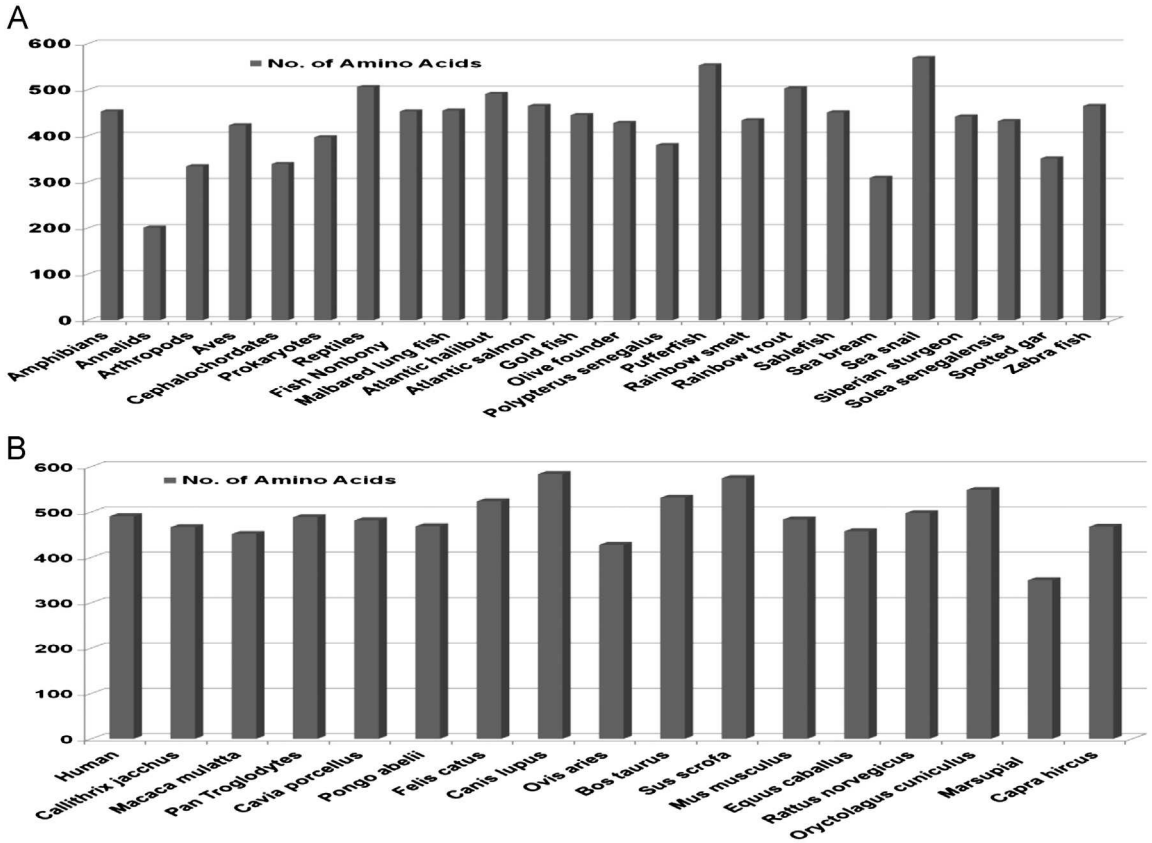

Fig. 3. Obtained amino acid length distribution. Distribution of average amino acid sequence length in (A) non-mammals and (B) mammals group.

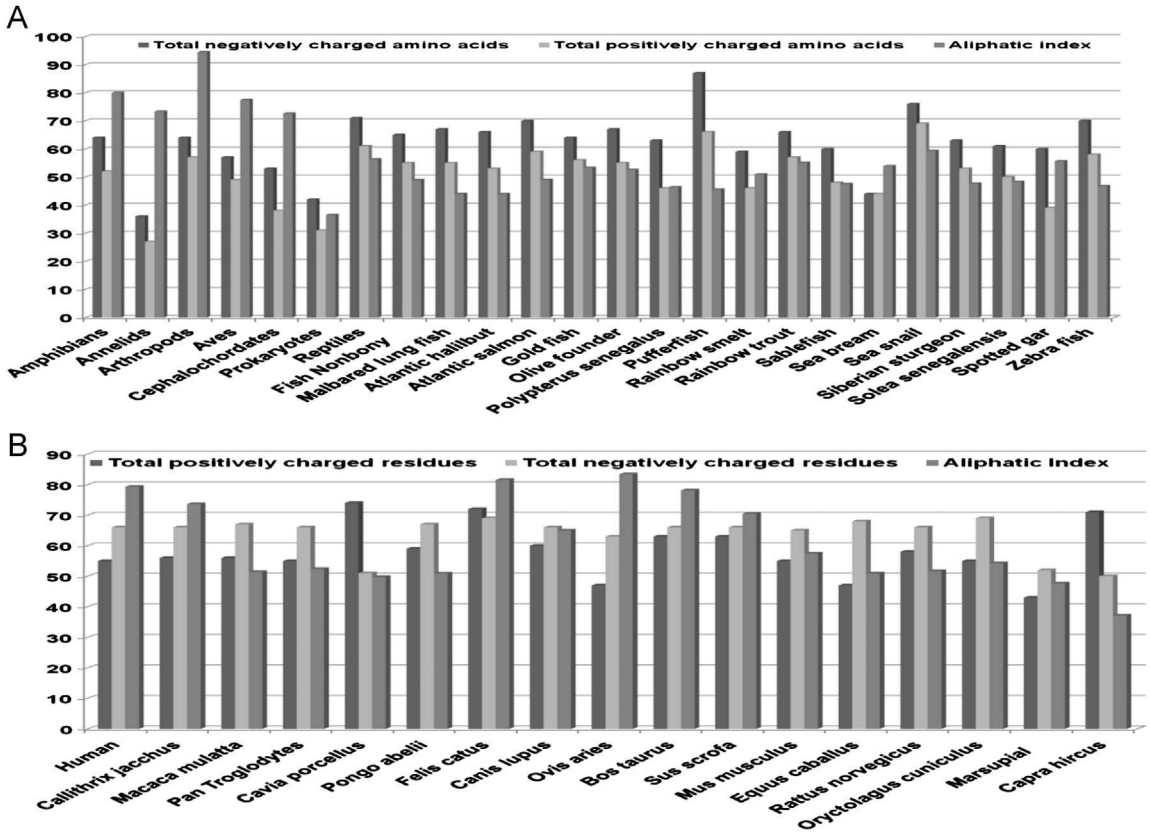

Fig. 4. Computation of the vital parameters such as average values of total negatively charged amino acids, total positively charged amino acids and Aliphatic Index in non-mammals (A) and (B) mammals.

Table 1
Top 20 features obtained after performing *t*-statistic.

Obtained parameter (variables)	*t*-statistic value	Obtained parameter (variables)	*t*-statistic value
Average area buried	11.58	Val	5.23
Gln	11.53	Nitrogen percentage	4.91
Phe	9.42	No. of amino acids	4.89
Cys	8.13	Leu	4.76
Asn	6.39	Hydrogen percentage	4.62
Total atoms	5.94	Sulfur percentage	4.62
Met	5.65	Ser	4.60
Molecular weight	5.49	Ala	4.55
Oxygen percentage	5.38	Lys	4.29
Theoretical PI	5.29	Glu	3.77

Table 2
Obtained average percent accuracy for different methodologies adopted for the 10 Fold cross validation approach.

Method	Percent accuracy
KNN	95.52
Discriminant analysis	62.54
Bayesian model	81.70
Tree induction	68.27
Rule induction	94.11
Neural network	72.16
Logistic regression	96.47
SVM	95.05
Meta-modeling	70.23

the analyses. A total of 10 subsets of the data were tested and best results are reported (Suppl.Table 1).

Deviation was observed in the performance of Discriminant analysis, Bayesian modeling, Meta-modeling and Tree Induction. Stability and improvement in outputs were observed for KNN, Neural Network, Rule Induction, Logistic Regression and Support Vector Machine approach (Table 2).

Best result was obtained for the logistic regression method with a percent accuracy value of 96.47%. The other comparable better results were observed for KNN (95.52%), SVM (95.05%) and Rule Induction (94.11%).

To obtain further insight on true and false prediction tendency, sensitivity and specificity were computed from the generated confusion matrix. Comparative understanding of sensitivity and specificity is essential to estimate the performance of a specific algorithm or methodology. It is known that high sensitivity and low specificity signifies over-prediction tendency of a particular method while low sensitivity and high specificity specifies too

Table 3
Comparative sensitivity and specificity for the considered methodologies during 10 Fold Cross validation analyses.

Method	Average sensitivity	Average specificity
KNN	0.95	0.94
Discriminant analysis	0.94	0.5
Bayesian model	0.91	0.63
Tree induction	0.55	0.91
Rule induction	0.93	0.94
Neural network	0.97	0.27
Logistic regression	0.97	0.95
SVM	0.95	0.93
Meta-modeling	0.56	0.94

Table 4
Comparative Positive Predictive Value (PPV) and Negative Predictive Value (NPV) for the considered methodologies during 10 Fold Cross validation analyses.

Method adopted	Average PPV	Average NPV
KNN	0.97	0.92
Discriminant analysis	0.63	0.33
Bayesian model	0.81	0.81
Tree induction	0.92	0.53
Rule induction	0.96	0.89
Neural network	0.7	0.87
Logistic regression	0.97	0.95
SVM	0.96	0.92
Meta-Modeling	0.95	0.55

Table 5
Comparative Average AUC value calculated from sensitivity and specificity results for different methods.

Method	Average AUC
KNN	95.38
Discriminant analysis	49.98
Bayesian model	77.77
Tree induction	73.4
Rule induction	94.22
Neural network	62.32
Logistic regression	96.15
SVM	94.8
Meta-modeling	75.61

Table 6
Comparative Average Matthew's Correlation Coefficient value calculated from confusion matrix for different methods.

Method	Matthew's Correlation Coefficient
KNN	0.90400184
Discriminant analysis	0.00692
Bayesian model	0.593782416
Tree induction	0.464085566
Rule induction	0.875638356
Neural network	0.375468868
Logistic regression	0.924091169
SVM	0.893740742
Meta-modeling	0.507881539

much conservation in the prediction. Therefore, to understand the comparable performance and efficiency of the approaches considered for the study, sensitivity (S_n) and specificity (S_p) were computed for all the methods (Table 3).

High specificity and sensitivity signifies the stability and better performance of an approach adopted for the study. In this study, KNN, Rule Induction, Logistic Regression and SVM approach turned out to be the best suited methodologies for the dataset considered. The KNN method showed 0.95 and 0.94 for sensitivity and specificity respectively. In a similar fashion, Rule Induction showed 0.93 and 0.94 sensitivity and specificity respectively. Logistic regression resulted in 0.97 and 0.95 whereas SVM yielded 0.95 and 0.93 for sensitivity and specificity respectively. The other approaches showed lot of variation in sensitivity and specificity which signifies the instability of these methods in classifying the considered dataset.

Similar information was obtained from the PPV and NPV calculations using the Eqs. (14) and (15) in the methodology section. Better average PPV and NPV were observed for KNN, Rule Induction, Logistic Regression and SVM (Table 4). This again confirms the suitability of these methodologies for this study.

Applying the equation (Eq. 16) mentioned in the methodology section, area under curve was computed for all the 10 datasets used, for each methodology adopted for this study. Few methodologies seem to be promising with an average value of more than 90. Logistic regression, KNN provided best results (Table 5) followed by the SVM and Rule Induction method. This output suggests that these methodologies are suitable to obtain better and robust classification for similar kind of data analysis.

Expression of the confusion matrix in a single value for better resolution and understanding for an unbalanced data was performed very well using Matthew's Correlation Coefficient (MCC). The obtained values range between −1 and +1 where values close to the positive results depict better performance for an algorithm in consideration. MCC values were computed applying equation number 17 and provided in Table 6. It is clearly evident that performance of the logistic regression method is best followed by the KNN, SVM and Rule induction techniques.

4. Conclusion

Proper classification of organisms remains as one of the classical problems of biological sciences. So far, different approaches have been adopted to search a definite solution but nothing has come out as unique strategy for resolving this issue. Moreover, there are ongoing debates on considering the data-types for particular organisms. It is quite difficult and tedious to distinguish sibling species solely based on morphological and anatomical data and this warrants the need for novel and out of the box approaches. Keeping these facts in mind, we have tried to frame the problem in a multi-dimensional manner. Instead of relying on a single parameter, we have taken into account 56 parameters followed by extensive statistical analyses to extract the most informative variables those are helpful for classification with high accuracy values. The heterogeneity of the parameters under discussion is enormous. Therefore, comparative strategy was implemented to reduce the ineffective attributes using feature selection. In other aspect, this exercise also helped to find out the most valuable features as discussed in the result section. Employment of advanced data mining methods, no doubt, extends the scope of the analysis but selection of a specific algorithm and the data analysis method again raises the debate. To seek an answer to this issue, we have adopted a comparative approach where all available methods were tried to select the best method such as logistic regression, KNN etc. in this case. This whole approach seems to be most apt way which leaves tremendous scope for extension of dataset with more organisms as well as choice of methodology. Increment in the efficiency and hybridization of the methodologies may aid in finding some definite answer for the classification and phylogenetic analysis of organisms where multiple complex parameters could be incorporated in the near future.

Conflict of interest statement

We, the authors, have no conflict of interest related to the manuscript entitled “Keratin protein property based classification of mammals and non-mammals using intelligent classification techniques”.

Acknowledgments

AKB thanks Council of Scientific and Industrial Research (CSIR) for Senior Research Fellowship (SRF).

Appendix A. Supporting Information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compbiomed.2013.04.007.

References

[1] E. Mayr, W.J. Bock, Classifications and other ordering systems, J. Zool. Syst. Evol. Res. 40 (2002) 169–194.
[2] M.J. Anderson, Carl Linnaeus: Father of Classification, Enslow Publishers, United States, ISBN 978-0-89490-786-9, 1997.
[3] C.P Hickman, L.S Roberts, A.L. Larson, Integrated Principles of Zoology, McGraw-Hill, Dubuque, IA538.
[4] E.A. Kogan, D.A. Ugriumov, G. Jaques, Morphologic and molecular-genetic characteristics of keratinization and apoptosis in squamous cell lung carcinoma, Arkh. Patol. 62 (3) (2000) 16–20.
[5] L. Kreplak, J. Doucet, P. Dumas, F. Briki, New aspects of the alpha-helix to beta-sheet transition in stretched hard alpha-keratin fibers, Biophys. J. 87 (1) (2004) 640–647.
[6] R.J. Block, H.B. Vickery, The basic amino acids of proteins. A chemical relationship between various keratins, J. Biol. Chem. 93 (1931) 113–117.
[7] M.H. Lynch, W.M. O'Guin, C. Hardy, L. Mak, T. Sun, Acidic and basic hair/nail (hard) keratins: their colocalization in upper cortical and cuticle cells of the human hair follicle and their relationship to soft keratins, J. Cell Biol. 103 (6) (1986) 2593–2606.
[8] J.E. Plowman, The proteomics of keratin proteins, J. Chromatograph. B 849 (2007) 181–189.
[9] A. Schermer, J.V. Jester, C. Hardy, D. Milano, T. Sun, Transient synthesis of K6 and K16 keratins in regenerating rabbit corneal epithelium: keratin markers for an alternative pathway of keratinocyte differentiation, Differentiation 42 (1989) 103–110.
[10] A.D. Irvine, W.H.I. Mclean, Human keratin diseases: the increasing spectrum of disease and subtlety of the phenotype–genotype correlation, Br. J. Dermatol. 140 (1999) 815–828.
[11] J.W. Said, G. Nash, G. Tepper, S.P. Banks-Schlegel, Keratin proteins and carcinoembryonic antigen in lung carcinoma: an immunoperoxidase study of fifty-four cases, with ultrastructural correlations, Hum. Pathol. 14 (1) (1999) 70–76.
[12] S.P. Banks-Schlegel, E.M. Mcdowell, T.S. Wilson, B.F. Trump, C.C. Harris, Keratin proteins in human lung carcinomas. combined use of morphology, keratin immunocytochemistry, and keratin immunoprecipitation, Am. J. Physiol. 114 (2) (1984) 273–286.
[13] R. Eichner, P. Bonitz, T. Sun, Classification of epidermal keratins according to their immunoreactivity, isoelectric point, and mode of expression, J. Cell Biol. 98 (1984) 1388–1396.
[14] D. Cooper, A. Schemer, R. Pruss, T. Sun, The use of aIF, AEl, and AE3 monoclonal antibodies for the identification and classification of mammalian epithelial keratins, Differentiation 28 (1984) 30–35.
[15] R.C. Marshall, M.J. Frenkel, J.M. Gillespie, High-sulfur proteins in mammalian keratins: a possible aid in classification, Aus. J. Zool. 25 (1) (1977) 121–132.
[16] D. Wu, D.M. Irwin, Y. Zhang, Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair, BMC Evol. Biol. 8 (2008) 241.
[17] P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, G. Valiente, Compression-based classification of biological sequences and structures via the universal similarity metric: experimental assessment, BMC Bioinf. 8 (2007) 252.
[18] A.K. Banerjee, N. Arora, U.S.N. Murty, Classification and regression tree (CART) analysis for deriving variable importance of parameters influencing average flexibility of CaMK kinase family, Electron. J. Biol. 4 (1) (2008) 27–33.
[19] A.K. Banerjee, N. Arora, V. Pranitha, U.S.N. Murty, Exploring the interplay of sequence and structural features in determining the flexibility of AGC kinase protein family: a bioinformatics approach, J. Proteom. Bioinf. 1 (2008) 77–89.
[20] M. Zhu, S. Zhao, Candidate gene identification approach: progress and challenges, Int. J. Biol. Sci., 3, 420–427.
[21] Y. Qi, Z. Bar-Joseph, J. Klein-Seetharaman, Evaluation of different biological data and computational classification methods for use in protein interaction prediction, Proteins Struct. Funct. Bioinf. 63 (2006) 490–500.
[22] Q. Liu, A.H. Sung, M. Qiao, Z. Chen, J.Y. Yang, M.Q. Yang, X. Huang, Y. Deng, Comparison of feature selection and classification for MALDI-MS data, BMC Genomics (2009), (Suppl 1):S3.
[23] A.K. Banerjee, K. Kiran, U.S.N. Murty, Ch. Venkateswarlu, Classification and identification of mosquito species using artificial neural networks, Comput. Biol. Chem. 32 (6) (2008) 442–447.
[24] J. Nahar, S. Ali, Y.P. Chen, Microarray data classification using automatic SVM kernel selection, DNA Cell Biol. 26 (10) (2007) 707–712.
[25] W.S. Noble, What is a support vector machine? Nat. Biotechnol. 24 (12) (2006) 1565–1567.
[26] H. Sahbi, D. Geman, A hierarchy of support vector machines for pattern detection, J. Mach. Learn. Res. 7 (2006) 2087–2123.
[27] M.R. Guarracino, A. Chinchuluun, P.M. Pardalos, Decision rules for efficient classification of biological data, Optim. Lett. 3 (2009) 357–366.
[28] U.S.N. Murty, A.K. Banerjee, N. Arora, Application of Kohonen maps for solving the classification puzzle in AGC kinase protein sequences, Interdiscip. Sci. Comput. Life Sci. 1 (2009) 173–178.
[29] A.K. Banerjee, M. Sunita, M. Naveen, U.S.N. Murty, Classification and clustering analysis of pyruvate dehydrogenase enzyme based on their physicochemical properties, Bioinformation 4 (10) (2010) 456–462.
[30] U.S.N. Murty, A.K. Banerjee, N. Arora, An in silico approach to cluster CAM kinase protein sequences, J. Proteom. Bioinf. 2 (2009) 97–107.
[31] A. Kulkarni, B.S.C.N. Kumar, V. Ravi, Colon cancer prediction with genetics profiles using evolutionary techniques, Expert Syst. Appl. 38 (3) (2011) 2752–2757.
[32] B. Jin, A. Strasburger, S.J. Laken, F.A. Kozel, K.A. Johnson, M.S. George, X. Lu, Feature selection for fMRI-based deception detection, BMC Bioinform. (2009), (Suppl 9):S15.
[33] A.K. Banerjee, N. Harikrishna, J.V. Kumar, U.S.N. Murty, Towards classifying organisms based on their protein physicochemical properties using comparative intelligent techniques, Appl. Artif. Intell. 25 (5) (2011) 426–439.
[34] S.D. Abdelmessih, F. Shafait, M. Reif, M. Goldstein, Landmarking for Meta-learning using RapidMiner, in: Proceedings of the RapidMiner Community Meeting and Conference, RCOMM'10. Dortmund, Germany, 2010.
[35] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, A. Bairoch, in: John M. Walker (Ed.), The Proteomics Protocols Handbook, Humana Press 571–607.
[36] Y. Wang, Y. Li, Prediction of protein subcellular locations using fuzzy k-NN method, Bioinformatics 20 (1) (2003) 21–28.

Amit Kumar Banerjee is a Senior Research Fellow (C.S.I.R) working in Indian Institute of Chemical Technology, Hyderabad, India and pursuing his doctoral degree in the area of application of computational techniques in vector borne diseases. He holds a B.Sc. (honors) degree in Physiology followed by an M.Sc. and M.Phil. in Biotechnology, a PGDPL in Patent Laws, and Adv. Dip. in Bioinformatics and successfully topped the national level course on Adv. Course in Bioinformatics. He is having more than eight years of research experience in various directions of computational molecular biology and chemistry. His research interest encompasses protein structure prediction and docking, application of evolutionary algorithms and intelligent techniques in biological data classification, microRNA prediction and computational genomics in deciphering information with relation to vector borne diseases. So far, he has published 31 research articles in reputed peer reviewed journals, 11 in conference proceedings, 5 book chapters under eminent publishing houses. He is having 1 copyright related to the database development on mosquitoes.

Vadlamani Ravi is Associate Professor, in the Institute for Development and Research in Banking Technology, Hyderabad. He obtained his Ph.D. in the area of Soft Computing from Osmania University, Hyderabad and RWTH Aachen, Germany (2001); MS (Science and Technology) from BITS, Pilani (1991) and M. Sc. (Statistics & Operations Research) from IIT, Bombay (1987). At IDRBT, he spearheads the CRM lab and evangelizes it in a big way by conducting customized training programmes for bankers on CRM subsuming OCRM & ACRM; Data Warehousing & Data Mining and conducting POC for banks etc. He has 100 papers to his credit with the break-up of 51 papers in refereed International Journals, 4 papers in refereed National Journals, 33 papers in refereed International Conferences and 3 papers in refereed National Conferences and 9 invited book chapters. His papers appeared in Applied Soft Computing, Soft Computing, Asia–Pacific Journal of Operational Research, Decision Support Systems, European Journal of Operational Research, Expert Systems with Applications, Fuzzy Sets and Systems, IEEE Transactions on Fuzzy Systems, IEEE Transactions on Reliability, Information Sciences, Journal of Systems and Software, Neural Computing and Applications, Knowledge Based Systems, IJUFKS, IJCIA, IJAEC, IJDMMM, IJIDS, IJDATS, IJISSS, IJCIR, IJCISIM, IJBIC, Computers and Chemical Engineering, Canadian Geotechnical Journal, Biochemical Engineering Journal, Bioinformation, Journal of Services Research etc. He also edited a Book entitled "Advances in Banking Technology and Management: Impacts of ICT and CRM" (http://www.igi-global.com/reference/details.asp?id=6995), published by IGI Global, USA, 2007. Some of his research papers are listed in Top 25 Hottest Articles by Elsevier and World Scientific. He has an H-index of 13 and more than 500 citations for his papers (http://scholar.google.co.in/). Two Ph.D. students graduated under his supervision. So far, he advised 42 M.Tech./MCA/M.Sc projects and at least a dozen Summer Interns from various IITs. He currently supervises 5 M.Tech students. He is on the Expert Committee for some banks embarking on Data Warehouse, Data Mining and CRM. Prior to joining IDRBT as Assistant Professor in April 2005, he worked as a Faculty at the Institute of Systems Science (ISS), National University of Singapore (April 2002–March 2005). At ISS, he was involved in teaching M.Tech. (Knowledge Engineering) and research in the areas of Fuzzy Systems, Neural Networks, Soft Computing Systems and Data Mining & Machine Learning. Further, he consulted for Seagate Technologies, Singapore and Knowledge Dynamics Pte. Ltd., Singapore, on data mining projects. Before leaving for Singapore, he worked as Assistant Director (Scientist E1) from 1996 to 2002 and Scientist C from 1993 to 1996 at the Indian Institute of Chemical Technology (IICT), Hyderabad. He was deputed to RWTH Aachen (Aachen University of Technology) Germany under the DAAD Long Term Fellowship to carry out advanced research during

1997–1999. He earlier worked as Scientist B and Scientist C at the Central Building Research Institute, Roorkee (1988–1993) and was listed as an expert in Soft Computing by TIFAC, Government of India.

Upadhayula Suryanaryana Murty is presently working as Scientist 'G'/Director Grade Scientist, and Heading the Biology Division at Indian Institute of Chemical Technology (Council of Scientific and Industrial Research), Hyderabad. He has published over 100 research articles in peer reviewed national and international journals in the areas of vectors and vector borne diseases, microbiology, bioinformatics and sericulture. His research interests are in medical entomology, bioinformatics applications, data mining and integrated pest management. He has been presented with several recognitions, awards and travel fellowships for his contribution to the field of medical informatics. To mention a few WHO and Tropical Disease Research fellowship, German Research Foundation Fellowship by University of Heidelberg, best Project performance by Govt. of Mizoram, eNorthEast Award 2011 etc. He is having several copyrights for developing databases and software in the area of medical entomology.

Anirudh Pramod Shanbhag is having a master degree in life sciences in the area of Biochemistry and later on joined the Advanced Course in Bioinformatics jointly conducted by IICT, CDAC & JNTU through a nationwide competition. He scored more than 71% in the ACB course and ranked among the toppers. He was one of the most promising students of the batch and showed his interest and caliber during the course work. He worked in this project and performed excellent. Later on he joined the prestigious National Center for Biological Sciences, Bangalore as a project fellow and presently continuing there in a project dealing both experimental and theoretical areas.

V Lakshmi Prasanna possess a master degree in the area of Biotechnology from Gitam Institute of Science (GIS), Gitam University , Visakhapatnam, followed by which she joined the Advanced Course In Bioinformatics jointly conducted by IICT, CDAC & JNTU through a nationwide competition. She also possesses an Adv. Dip. in Tissue Culture Technology. Lakshmi pursued project works in the area of experimental immunology and microbiology from Indian Immunologicals Limited, Hyderabad during her graduation. She showed her interest and understanding during the course and later on continued working in some part of this project. Presently she is working in the corporate sector as a data miner.