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