Journal of Theoretical Biology 403 (2016) 30–37

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Machine learning approaches for discrimination of Extracellular Matrix proteins using hybrid feature space

Farman Ali, Maqsood Hayat *

Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan

HIGHLIGHTS

- Computational model is developed for prediction of Extracellular Matrix (ECM) proteins.
- PseAAC and Dipeptide is used as feature extraction schemes.
- Various classification algorithms are utilized.
- KNN achieved quite promising results.

ARTICLE INFO

Article history:
Received 25 December 2015
Received in revised form
2 May 2016
Accepted 3 May 2016
Available online 11 May 2016

Keywords:
Extracellular Matrix
Di-peptide composition
K-Nearest Neighbor
PseAAC

ABSTRACT

Extracellular Matrix (ECM) proteins are the vital type of proteins that are secreted by resident cells. ECM proteins perform several significant functions including adhesion, differentiation, cell migration and proliferation. In addition, ECM proteins regulate angiogenesis process, embryonic development, tumor growth and gene expression. Due to tremendous biological significance of the ECM proteins and rapidly increases of protein sequences in databases, it is indispensable to introduce a new high throughput computation model that can accurately identify ECM proteins. Various traditional models have been developed, but they are laborious and tedious. In this work, an effective and high throughput computational classification model is proposed for discrimination of ECM proteins. In this model, protein sequences are formulated using amino acid composition, pseudo amino acid composition (PseAAC) and dipeptide composition (DPC) techniques. Further, various combination of feature extraction techniques are fused to form hybrid feature spaces. Several classifiers were employed. Among these classifiers, K-Nearest Neighbor obtained outstanding performance in combination with the hybrid feature space of PseAAC and DPC. The obtained accuracy of our proposed model is 96.76%, which the highest success rate has been reported in the literature so far.

1. Introduction

Extracellular Matrix (ECM) is a complex structure of proteins released by cells that furnish structural support to the neighbor cells. ECM proteins are categorized into two main classes: (i) collagens; (ii) proteoglycans (Yang et al., 2015). The first class collagens are synthesized by fibroblast cells (Chan et al., 2015). Approximately, 90% parts of the bones matrix protein are consisting of collagens and also found the most abundant protein in mammals (Di-Lullo et al., 2002; Karsenty and Park, 1995; Kern et al., 2001). Furthermore, collagens also perform a vital role in cell adhesion, migration and proliferation. It provides framework to other parts of human body such as blood vessels, cartilage, bones and corneas. The second class proteoglycans are further classified into Heparan sulfate, chondroitin sulfate and keratin sulfate. Heparan sulfate performs various activities like angiogenesis, blood clotting and embryonic development while chondroitin is a major constituent of aorta, ligaments and tendons (Hensch, 2005). Similarly, keratin sulfate is also important part of horns of animals. Elastin is one of the crucial part of ECM proteins providing structural and mechanical supports to various mammals body organs such as extension and contraction of muscular tissues, which help in the spinal cord and neck movement (Li et al., 1998; Rosenbloom et al., 1993). Besides these, EMC proteins are also essential part of bones engineering, body growth, wound healing and inflammation processes (Peach et al., 1993). Disorders and deregulations in collagen encoding genes cause several diseases such as developmental abnormalities, epidermolysis bullosa, Ehlers Danlos Syndrome and cancer (Gurvan et al., 2010). Deficiencies in some ECM

* Corresponding author.
E-mail addresses: Maqsood.hayat@gmail.com, m.hayat@awkum.edu.pk (M. Hayat).

http://dx.doi.org/10.1016/j.jtbi.2016.05.011
0022-5193/

proteins cause Williams syndrome and cutis laxa (Provenzano et al., 2009).

Due to the remarkable significance of ECM proteins in various biological events, a series of computational prediction models have been carried out. In this regard, Juan et al., developed ECMPP predictor for identification of ECM proteins (Jung et al., 2010). Similarly, Anitha et al., utilized Position Specific Scoring Matrix (PSSM) in combination with Support Vector Machine (SVM) (Anitha et al., 2012), while Kandaswamy et al., established ECMPRED (ECM PREDiction) web server (Kandaswamy et al., 2013). Zhang et al., developed PECM model which comprises of PseAAC in conjunction with SVM (Zhang et al., 2014). Several researchers employed the concept of ensemble classifiers and hybrid feature space (Cai, 2003; Huang et al., 2011; Xiao and Wang, 2009; Xiao and Wang, 2011). Recently, Yang et al., developed a model employing an ensemble method in conjunction with hybrid features (Yang et al., 2015). In their model, features were extracted using frequency of physicochemical properties and functional groups and provided as input to the Random Forest classifier. Although these approaches have achieved some reasonable results but still there exists some room of improvement. The important parameter for measuring the performance of computational method is accuracy but sometimes accuracy does not reflect the desire results. In such circumstances, sensitivity and specificity are computing to show the true reflection of all the desire classes. In existing methods, some of them have less accuracy while some have reasonable accuracy but worse sensitivity. In order to overcome the deficiencies of exiting methods there needs such a computational model which target both the sensitivity and specificity. In this regards, an effective and high throughput computational model is proposed. In this model, salient features of protein were extracted using amino acid composition (AAC), di-peptide composition (DPC) and pseudo amino acid composition (PseAAC). Further, various combination of feature spaces were merged to build hybrid feature spaces. The performance of classifiers is assessed using 10-fold cross validation test. The framework of the proposed model is showed in Fig. 1.

The remaining paper is structured as following: materials and methods are expressed in Section 2; evaluating metrics are presented in Section 3; results and discussion are described in Section 4 and finally conclusion is provided in Section 5.

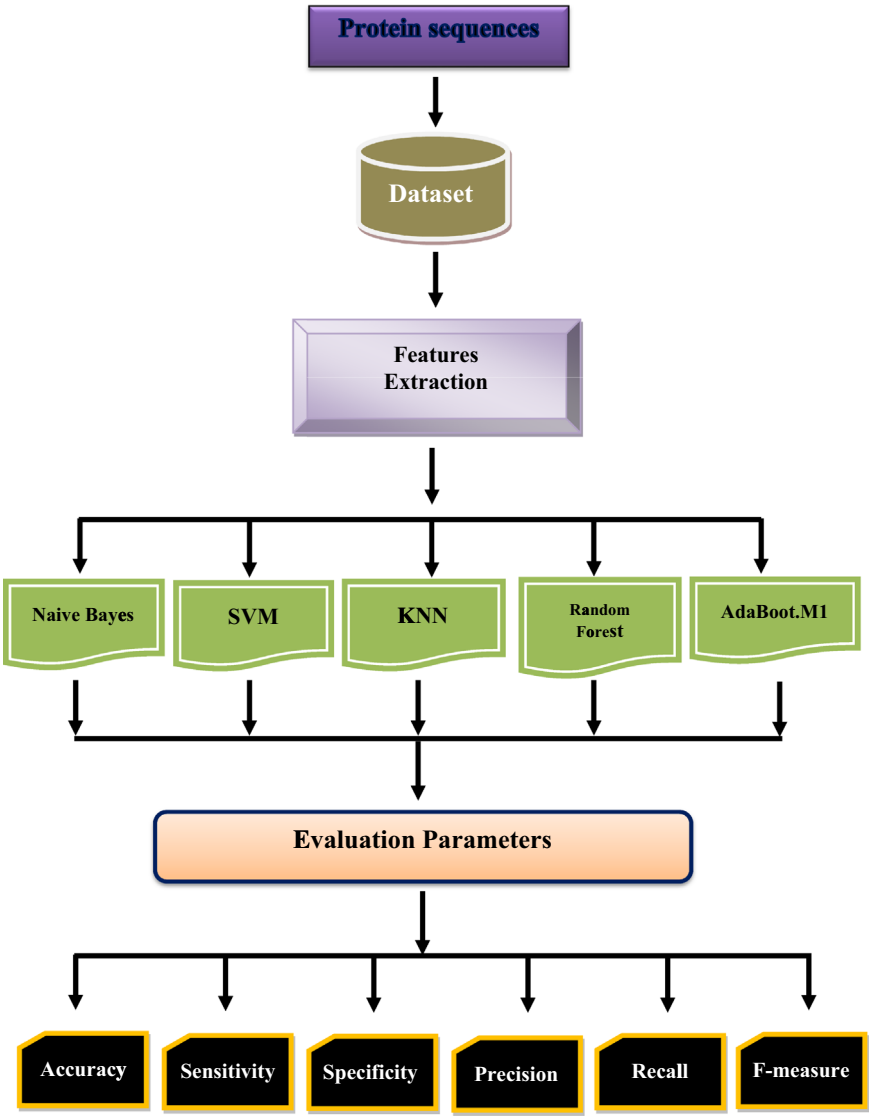

Fig. 1. Framework of Proposed model.