Chemometrics and Intelligent Laboratory Systems 174 (2018) 22–32

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Improving prediction of extracellular matrix proteins using evolutionary information via a grey system model and asymmetric under-sampling technique

Muhammad Kabir [a], Saeed Ahmad [a], Muhammad Iqbal [b], Zar Nawab Khan Swati [a,c], Zi Liu [a], Dong-Jun Yu [a,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] Department of Computer Science, Abdul Wali Khan University Mardan, Khyber Pakhtunkhwa 23200, Pakistan
[c] Department of Computer Science, Karakoram International University, Gilgit-Baltistan 15100, Pakistan

ARTICLE INFO

Keywords:
Extracellular matrix proteins
Evolutionary information
Grey system model
GreyPSSM
Asymmetric under-sampling
Support vector machine

ABSTRACT

Extracellular Matrix proteins (ECMP) play vigorous part in performing various biological functions including cell migration, adhesion, proliferation, differentiation. Furthermore, embryonic development, angiogenesis, gene expression, and tumor growth are also regulated by ECMP. In view of this incredible significance, precise and reliable identification of ECMP through computational techniques is highly requisite. Although, previous works made substantial improvement, however, accurately predicting ECMP from primary protein sequence is still at the infant stage due to the rapid growth of proteins samples in online databases. In the current study, a novel sequence-based prediction method called TargetECMP has been proposed, which is based on the evolutionary information extracted via a grey system model. It utilizes asymmetric under-sampling approach for splitting the benchmark dataset into eleven subsets in order to avoid class imbalance problem. Jackknife cross-validation test is performed with support vector machine (SVM) on each subset of data and then ensemble majority voting is utilized to integrate outputs of SVM against each subset. The experimental results achieved by TargetECMP outperformed the existing predictor on both benchmark dataset and independent dataset. Owning to best prediction results provided by TargetECMP, it is demonstrated that the analysis will provide novel insights into basic research, drug discovery and academia in general and function of extracellular matrix proteins in particular.

1. Introduction

Extracellular matrix proteins (ECMP) make a class of secreted proteins, and get together as a broad network on the surface of the cell [1]. It assembles a complex structure of proteins secreted by cells that provide physical and chemical support to neighbor cells [2]. The composition of ECMP varies among multicellular structures; however, cell adhesion, cell-to-cell communication, homeostasis, tissue morphogenesis, differentiation, regulation of embryonic development, angiogenesis, gene expression, and tumor growth are certain common functions of the ECMP [3,4]. ECMP are classified from a broad spectrum into two categories: (i) collagens; (ii) proteoglycans [3]. First class, collagens, are synthesized by fibroblast cells [5]. It is most abundant protein found in mammals and almost 90% parts of the bones matrix proteins contains collagens [6,7]. Second class, proteoglycans, is proved to be in playing important role in migration, cell adhesion and proliferation. It also imparts a framework to other human body parts like, cartilage, blood vessels and bones. This class is further sub-divided into chondroitin sulfate, Heparan sulfate, and keratin sulfate. Chondroitin is the principal component of ligaments tendons and aorta [8]. Heparan sulfate also accomplish important activities, for example embryonic development, angiogenesis, and blood clothing etc. [9]. Similarly, keratin sulfate is also vital part of animal's horns [10,11]. Elastin is among one of the principal part of ECMP that provide mechanical and structural support to body organs of various mammals, like contraction and extraction of muscular tissues, that can help in the spinal card and neck movement [10]. Furthermore, ECMP are of great significant component of bones engineering wound healing, body growth and inflammation processes. Metal abnormalities,

* Corresponding author.
E-mail addresses: kabiricp@gmail.com, mdkabir@njust.edu.cn (M. Kabir), saeed.ahmad075@gmail.com (S. Ahmad), mdiqbalpk@gmail.com (M. Iqbal), zarnawab@kiu.edu.pk (Z.N. Khan Swati), liuzi189836@163.com (Z. Liu), njyudj@njust.edu.cn (D.-J. Yu).

https://doi.org/10.1016/j.chemolab.2018.01.004
Received 30 July 2017; Received in revised form 4 December 2017; Accepted 12 January 2018
Available online 4 February 2018

epidermolysis, bullosa, Ehlers Danlos Syndrome and cancer are several fetal diseases which are caused by dis-ordering and deregulations in collagen coding genes [12,13]. Deficiencies in some ECMP cause Williams syndrome and cutislaxa [12].

Owing the substantial potential of ECMP in different biological processes, events and aspects, long sequence of efforts were noted till now to develop computational models for its prediction. In this regard, Juan et al. developed ECMPP predictor for its identification [2]. Likewise, Position specific scoring matrix (PSSM) in combination with support vector machine (SVM) technique was developed by Anitha et al. [14]. A web server ECMPRED (ECM PREDiction) has also been established in this area [15]. PECM model, which comprises Pseudo amino acid composition (PseAAC) in conjunction with SVM was developed by Zhang et al. [16]. Various models consisting of hybrid features spaces were also proposed [17–19]. More recently, a hybrid model was proposed for ECMP prediction [20]. In this model amino acid composition (AAC) [21,22], PseAAC [21–24] and Dipeptide composition [25–27] were used to extract features and then hybrid those spaces and passed into various classification algorithms like, k-nearest neighbor, SVM, random forest, Naïve Bayes and AdaBoost.M [20].

Although, all these discussed approaches behaved very well and strengthen research about ECMP, but in the current era we need fast, accurate and robust predictor. Also, the aforementioned predictors did not proposed any idea for dealing with class imbalance problem. Class imbalance is a problem which should be handled very carefully while considering the predictive performance of the computational predictor. In this regard, various methods have been proposed by different researchers to deal with imbalance learning which can be roughly grouped into three categories [28]; (1) sample rescaling-based methods [29,30], (2) learning based methods [31–34] and (3) hybrid methods [35,36], which is the combination of both sample rescaling-based methods and learning based methods. Among these solutions for imbalance problems, the sample rescaling method has been widely adopted by researchers. Sample rescaling-based method includes two strategies, i.e., over sampling and under-sampling, which attempts to balance the imbalance data class by changing the number and distributions within them. These methods have provided promising results in the last few decades dealing with different problems including; [25,37–41]. The aforementioned techniques have certain problems.

In the present study, to balance the benchmark dataset and improve the prediction quality of the proposed model, we have adopted the asymmetric under sampling technique [42]. Comparing to the traditional under-sampling strategy, where some data samples are removed, in asymmetric under-sampling technique we did not remove any samples from the original benchmark dataset. We constructed 11 subsets of the original dataset and then combined the prediction of each subset using ensemble approach. By using this strategy, we can incorporate all the available datasets to train and test the predictor effectively, but also avoid the class imbalance issue. The detail process of how the divide the benchmark dataset into various subsets will be discussed in the later sections.

Despite the progress, in this study we develop a computational model, TargetECMP, which can identify ECMP with reflection of the desired results. In this framework, during the first phase the dataset is divided into different subsets in order to overcome the imbalance problem. As feature extraction is the most essential step in developing computational model, in this study we considered three well-known feature extraction methods called split amino acid composition (SAAC) position specific scoring matrix (PSSM) and grey system model based position specific scoring matrix (GreyPSSM) to extract local and global features respectively. For better comparative analysis of our method with the state-of-the-art methods, the subsets of benchmark dataset is then combined using majority voting system. We also analyzed the effect of class imbalanced with our proposed method using jackknife cross-validation test with SVM as classification engine.

2. Materials and methods

2.1. Benchmark dataset

In order to effectively train and test the computational predictor, we need to have some valid benchmark dataset [43,44]. For this purpose, we have utilized the same datasets as previously used by different researchers in their studies [3,14–16]. The benchmark dataset and the independent dataset for the current study can be formulated as;

$$S_m = S_m^+ \cup S_m^- \tag{1}$$

where m = 1 represents the benchmark dataset and m = 2 indicates the independent dataset utilized in this study. Further, S_m^+ contains the positive sequences of ECMP and S_m^-comprises the negative samples of ECMP sequences and ∪ represents the symbol for "union" in the set theory. There are 410 a total of ECMP sequences (positive samples) and total numbers of non-ECMP sequences (negative samples) are 4464 in benchmark dataset. Likewise, there are 85 positive and 130 negative sequences in independent dataset respectively.

According to the report in some recent publications [45,46], to avoid homology bias and remove the redundant sequences from the benchmark dataset, a cutoff threshold of 25% was imposed in Refs. [45,46] to exclude those proteins from the benchmark datasets that have equal to or greater than 25% sequence identity to any other in a same subset. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the numbers of proteins for some subsets would be too few to have statistical significance.

2.2. Feature extraction

The primary structure of proteins is a polymer of amino acids which are formulated and folded according to the attributes of amino acids. These attribute are also known as features. Extracting nominal features from the biological sequences is considered to be the most important phase during the development of computational predictors [44]. The nominal features always have a positive impact on the predictive quality of computational models. Therefore, it is highly indispensable to use good feature extraction strategy. In view of this, we have utilized two feature abstraction methods [47]. These methods are split amino acid composition (SAAC), position specific scoring matrix (PSSM) and grey system model position specific scoring matrix (GreyPSSM). The former is used to capture local features while the latter two are utilized to incorporate evolutionary information of biological sequences.

2.2.1. Split amino acid composition (SAAC)

In traditional and typical amino acid composition (AAC), the relative frequency of each amino acid is calculated for construction of feature vector. The proteins have vital informative peptides at their N- or C terminus regions which are not considered while AAC feature formulation [48]. To exploit this complementary information from proteins, split amino acid composition (SAAC) was developed which decomposes the protein sequence into several fragments and then composition of each fragment is computed independently. In our SAAC model, each protein sequence is decomposed into three fragments; (i) 50-AA of N-terminus, (ii) 50-AA of C-terminus, and (iii) region between these two termini. The resultant feature vector is a 60D instead of 20D as in case of AAC [49]. The feature vector of SAAC is represented as:

$$P = \left[f_1^N, ...,f_{20}^N, f_1^{\text{int}}, ...,f_{20}^{\text{int}}, f_1^C, ...,f_{20}^C\right] \tag{2}$$

where *N*, *int* and *C* represents the N-terminus, integral segment and C-terminus respectively.