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:

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

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

2.2.2. Position specific scoring matrix

Position specific scoring matrix (PSSM) is an evolutionary information technique which is commonly employed for patterns (motifs) representation in biological sequences [50]. In the recent few years, various studies demonstrated that the evolutionary information reflected by PSSM have provided power features for solving different bioinformatics problems [51]. Indeed, PSSM is the most adopted feature extraction strategy by the researchers for various classification problems. For this reason, PSSM was taken as feature extraction method in the current study. A protein sequence *P* with *L* residues of amino acid is formulated as follows;

$$P_{PSSM} = \begin{bmatrix} E_{1\to 1} & E_{1\to 2} & \cdots & E_{1\to j} & \cdots & E_{1\to 20} \\ E_{2\to 1} & E_{2\to 2} & \cdots & E_{2\to j} & \cdots & E_{2\to 20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{i\to 1} & E_{i\to 2} & \cdots & E_{i\to j} & \cdots & E_{i\to 20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{L\to 1} & E_{L\to 2} & \cdots & E_{L\to j} & \cdots & E_{L\to 20} \end{bmatrix} \tag{3}$$

Where $E_{i\to j}$ indicates the amino acid residue score in position *ith* of protein sequence being substituted to amino acid type *j* during the evolution computing processes. The values of amino acids are represented by ($j = 1, 2, \ldots., 20$) in their alphabetical order. The original values of PSSM for the ECMP sequences were obtained by executing its sequences against PSI-BLAST to search the Swiss-Prot database through three iterations with 0.001 as the E-value cut-off against the sequence of protein *P*, which has *L*20* scores as described in Eq. (3). Further, the P_{PSSM} is normalized by a standard conversion as follows;

$$E_{i\to j} = \frac{E^0_{i\to j} - \overline{E^0_i}}{SD\left(\overline{E^0_i}\right)} \quad (i = 1, 2, \ldots, L; \quad j = 1, 2, \ldots, 20) \tag{4}$$

Where $E^0_{i\to j}$represents original values of evolutionary information generated by PSI-BLAST, which can be both positive and negative integers. $\overline{E}^0_i$Indicates the average of $E^0_{i\to j}$ over ($j = 1, 2, \ldots, 20$) and $SD(\overline{E}^0_i)$represents the standard deviation. In Eq. (3), *L* shows the length of protein sequences. As we know the length of protein sequence is different in the benchmark dataset, and it is not possible to develop predictor which is capable of handling sequences of different lengths. As the sequences provided have different length. This problem can be solved by expressing the protein sequence in a fixed length feature vector as described below;

$$\overline{P_{20}} = \left[\overline{E_1} \quad \overline{E_2} \quad ,\ldots., \quad \overline{E_{20}}\right]^T \tag{5}$$

$$\overline{E}_j = \frac{1}{L}\sum_{i=1}^{L} E_{i\to j} \quad (j = 1, 2, \ldots, 20) \tag{6}$$

where $\overline{E}_j$ indicates the average score of the amino acid residues in the protein *P* being substituted to amino acid type *j* during the biological evolutionary processes [52].

2.2.3. GreyPSSM approach

Grey system theory was introduced by Deng Julong in 1989 [53]. According to this theory, if the information of the investigating system is fully known, the system is called as “white system', if the information of the investigating system is partially known, such system is called “grey system”; and if the information of the investigating system is completely unknown, the system is called a “black system'. Therefore, the model developed on the basis of grey system theory is known as “grey model”. The grey model is particularly useful for solving complicated problems that are lack of sufficient information, or need to process uncertain information and to reduce random effects of acquired data. More recently, Xiao et al. used grey model based model for identification of antifreeze proteins [42], Qiu et al. for prediction of lysine ubiquitination sites in proteins [54], Xiao et al., for identification of Enzymes Catalytic Sites [55] and Min et al., for Interaction between Enzymes and Drugs in Cellular Networking [56]. In the current study, we have used grey system to incorporate PSSM matrix to formulate grey model with sequence evolutionary information. In Eq. (5), defined PSSM based feature vector of 20-D which is based on linear probabilities of amino acids and does not contain any sequential information. In order to compute sequential information, grey system can be utilized to further defined 60-D components;

$$\overline{P}_{j+20} = \delta_j \quad (j = 1, 2, 3\ldots..60) \tag{7}$$

where

$$\delta_j = \begin{cases} a^1_1 & when \quad u = 1 \\ a^1_2 & when \quad u = 2 \\ b^1 & when \quad u = 3 \\ \vdots & \quad \vdots \qquad \quad \vdots \\ a^{20}_1 & when \quad u = 58 \\ a^{20}_2 & when \quad u = 59 \\ b^{20} & when \quad u = 60 \end{cases} \tag{8}$$

where

$$\begin{bmatrix} a^j_1 \\ a^j_2 \\ b^j \end{bmatrix} = \left(B^T_j B_j\right)^{-1} B^T_j U_j \quad (j = 1, 2, \ldots., 20) \tag{9}$$

In the above equation

$$B_j = \begin{bmatrix} -E_{2,j} & \left(-E_{1,j} - 0.5E_{2,j}\right) & 1 \\ -E_{3,j} & \left(-\sum_{i=1}^{2} E_{i,j} - 0.5E_{3,j}\right) & 1 \\ \vdots & \vdots & \vdots \\ -E_{L,j} & \left(-\sum_{i=1}^{L-1} E_{i,j} - 0.5E_{L,j}\right) & 1 \end{bmatrix} \tag{10}$$

and

$$U_j = \begin{bmatrix} E_{2,j} & E_{1,j} \\ E_{3,j} & E_{2,j} \\ \vdots & \vdots \\ E_{L,j} & E_{L-1,j} \end{bmatrix} \tag{11}$$

Now, each of these protein sequences in the dataset can be formulated by a 80-D vector via Equation (7), which is obtained by incorporating sequence evolutionary information using the grey system model (Equation (8)–(11)).

2.3. Support vector machine as classifier

Support vector machine (SVM) is a statistical based classification engine which was first introduced by Cortes and Vapnik in 1995. It has been extensively applied, and widely adopted classifier by the researchers in the area of pattern recognition, machine learning, data mining and bioinformatics [26,57–60]. At first, SVM was introduced for dealing with binary classification problems, but later on it was modified in 1999 to handle multi class problems. Currently, SVM has the capability to deal with multi-label problems as well. “One verses one” (OVO) and “one verses rest” (OVR) techniques are applied to the traditional SVM for classification of multi-class problems. The advantage of SVM over the other classification engines is that it transforms all the training data into a high dimensional feature space and seeking a hyperplane which separates positive instances of data from the negative instances of data [61]. Different kernel functions like linear, polynomial, Radial Base Function (RBF) and sigmoid are usually employed for the optimization of model in

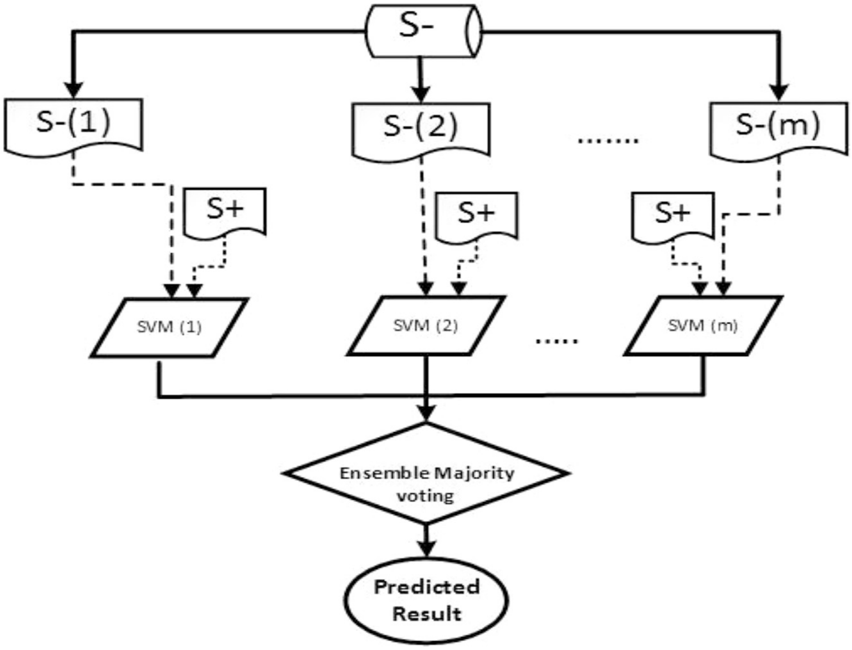


Fig. 1. A flowchart to describe the formulation of subsets from the benchmark dataset.

SVM. In the current study RBF was utilized as a kernel function with kernel width parameter γ and regularization parameter C. The values of C and γ were determined via an optimization procedure using a grid search approach.

$$K(x_i, x_j) = \exp(-\gamma|x_i - x_j|^2) \tag{12}$$

In the above equation the width of the Gaussian function shows through parameter γ. For the ECMP prediction, LIBSVM package was exploited which is available at http://www.csie.ntu.edu.tw/cjlin/libsvm and is free for downloading.

2.4. Learning from imbalance data with asymmetric under-sampling

As described in the dataset section, the benchmark dataset is imbalance i.e. the number of negative samples is very large as compared to positive samples. The positive samples are 410 and negative samples are 4464. In the development of computational predictor, an important step is the classification of the data samples to certain classes. It has been noted that if the positive and negative samples are not equal in size i.e. imbalance dataset, the classification engines cannot classify them precisely as they tends towards the majority class [25,62,63]. This biasness of classification algorithms is one of the main challenges in machine learning based approaches. Different researchers have proposed their own techniques to deal with imbalanced problems as discussed in the introduction section.

In this study, we have utilized the asymmetric under-sampling technique which employs the divide and conquer rule to deal with class imbalance problem [42]. In this method, the imbalance benchmark dataset is divided into different subsets keeping the number of positive samples constant and therein varying the number of negative samples. In this way, 11 subsets were constructed from the original benchmark dataset where each subset is treated as a new dataset as shown in Fig. 1. Further, to incorporate all the sequences for the demonstration of the quality assessment of the proposed predictor, the output of each subset is fused into and ensemble SVM strategy using majority voting ensemble technique [64,65]. By using this way, all the sequences in the dataset can be utilized simultaneously removing the imbalance problem. The process of how ensemble majority voting technique can be described as below; Let suppose SVM be the base classifier, then the ensemble classification can be expressed as below;

$$EnsC = SVM_1 \oplus SVM_2 \oplus SVM_3 \oplus SVM_4 \oplus .... \oplus SVM_{11} \tag{13}$$

Where *EnsC* represents the ensemble classification *SVM₁*, *SVM₂*, *SVM₃*, *SVM₄*, up to *SVM₁₁* indicates the outputs of SVM classifier against each subset of benchmark dataset. Suppose the predicted result of individual feature space for the protein query *P* is;

$$\{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}\} \in \{L_1, L_2\} \tag{14}$$

where $C_1, C_2, C_3, ..., C_{11}$are the outputs of SVM classifier for feature spaces generated from 11 subsets, $L_1$and L_2 represents the labels of ECMP and non-ECMP class respectively.

$$Y_j = \sum_{i=1}^{11} \delta(C_i, L_j), (j = 1, 2) \tag{15}$$

where

$$\delta(C_i, L_j) = \begin{pmatrix} 1, & if \quad C_i \in L_j \\ 0, & otherwise \end{pmatrix} \tag{16}$$

Finally, the output of ensemble SVM classifier combined through majority voting is obtained as;

$$MV_{Ens} = MAX(Y_1, Y_2, Y_3, ..., Y_{11}) \tag{17}$$

where MV_{Ens} is the predicted result of the ensemble majority voting technique, *MAX* represents choosing the maximum class number between ECMP and non-ECMP.

3. Performance assessment metrics

In machine learning, several statistical metrics are utilized to facilitate the quantitative analysis of a computational predictor [37]. In this study, we have employed the following five performance evaluation parameters to evaluate the prediction capability of our developed approach.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{18}$$

Table 1
Performance analysis of various feature extraction methods with jackknife test on the benchmark datasets with different negative datasets.

Benchmark Datasets	GreyPSSM				PSSM				SAAC			
	Acc (%)	*Sen* (%)	*Spe* (%)	*MCC*	*Acc* (%)	*Sen* (%)	*Spe* (%)	*MCC*	*Acc* (%)	*Sen* (%)	*Spe* (%)	*MCC*
$S(1) = S^+ \cup S^-(1)$	93.04	92.62	93.47	0.861	91.95	91.58	92.31	0.839	89.39	90.42	88.35	0.788
$S(2) = S^+ \cup S^-(2)$	95.24	94.44	96.02	0.905	94.87	95.00	94.75	0.897	92.56	92.37	92.74	0.851
$S(3) = S^+ \cup S^-(3)$	92.92	92.68	93.17	0.858	92.68	92.80	92.56	0.853	89.87	91.28	88.47	0.798
$S(4) = S^+ \cup S^-(4)$	94.63	94.14	95.12	0.892	93.04	92.50	93.59	0.861	89.51	90.36	88.65	0.790
$S(5) = S^+ \cup S^-(5)$	91.95	91.21	92.68	0.839	90.73	90.85	90.60	0.814	87.92	88.84	87.01	0.758
$S(6) = S^+ \cup S^-(6)$	94.75	95.06	94.45	0.895	93.65	94.02	93.29	0.873	92.43	92.92	91.95	0.848
$S(7) = S^+ \cup S^-(7)$	92.92	93.04	92.80	0.858	91.70	91.46	91.95	0.834	88.41	90.30	86.52	0.769
$S(8) = S^+ \cup S^-(8)$	93.65	92.07	95.24	0.874	92.19	92.43	91.95	0.843	89.87	90.54	89.20	0.797
$S(9) = S^+ \cup S^-(9)$	91.09	89.69	92.50	0.822	90.24	90.36	90.12	0.804	88.29	89.75	86.82	0.766
$S(10) = S^+ \cup S^-(10)$	92.68	92.43	92.92	0.853	91.21	89.39	93.04	0.825	88.53	89.26	87.80	0.770
$S(11) = S^+ \cup S^-(11)$	91.46	91.09	91.82	0.829	89.75	89.63	89.87	0.795	88.78	90.24	87.31	0.776
Average	93.12	92.58	93.65	0.862	92.00	91.82	92.18	0.839	89.59	90.57	88.62	0.791
Integrated	94.13	93.16	94.22	0.798	92.43	91.83	92.50	0.755	88.12	89.79	87.97	0.663

$$Spe = \frac{TN}{TN + FP} \tag{19}$$

$$Sen = \frac{TP}{TP + FN} \tag{20}$$

$$BAcc = \frac{1}{2}(Spe + Sen) \tag{21}$$

$$MCC = \frac{(TP.TN) - (FP.FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{22}$$

In the above equations (18)–(22), where TP is True Positive, TN is False Negative, TN is True Negative and FP is False Positive.

4. Results and discussion

In designing sequence-based computational predictors, one of the important step is to objectively and properly evaluate its prediction quality [22]. For this purpose, two important aspects should be kept in mind. First, to select some metrices which effectively measure the prediction accuracy and second, to utlized some statistical testing techinques to derive the metrics values [66]. In the below sections, we have addressed these problems.

4.1. Cross validation

Cross-validation is an essential in analyzing the predictive performance of computational methods. Three statistical tests including jackknife, k-subsampling and independent dataset tests are utilized in this regard. Among these tests, jackknife is deemed to be the widely adopted by the researchers due to its effectiveness. It has the ability to return unique result for a given benchmark dataset. Accordingly, in this study, jackknife cross-validation test is performed on both benchmark dataset and independent dataset [67]. The aim of this study is to predict whether a given protein query is ECMP or not.

4.2. Contribution of various features

The performance of different feature extraction techniques using the SVM ensemble strategy is given in Table 1. The ensemble SVM was formed by combining the predictive output of SVM against each subspace through majority voting technique. Fig. 1 shows the process of how to combined prediction of each SVM for the final output. As illustrated above that, 11 subspaces were formed using asymmetric under-sampling technique from the benchmark dataset in order to minimize the effect of class imbalance. The predictive values given in Table 1 demonstrated that GreyPSSM based integrated SVM produced higher results; 94.13% accuracy, 93.16% sensitivity, 94.22% specificity and 0.798 MCC. Likewise PSSM based integrated method achieved 92.43%, 91.83%, 92.50% and 0.755 of accuracy, sensitivity, specificity and MCC respectively.

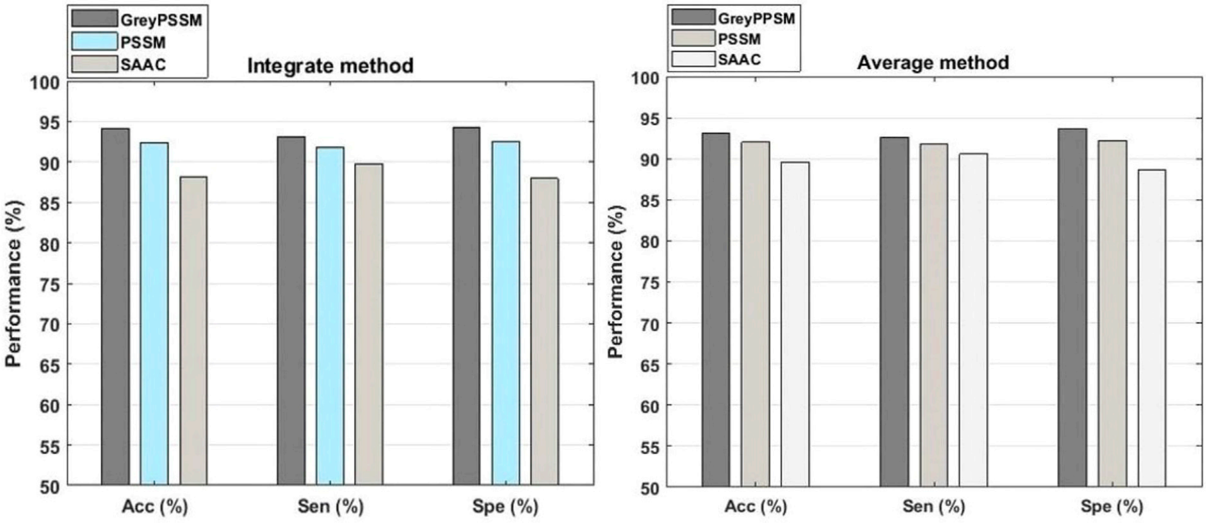


Fig. 2. Performance comparison integrated and average method on GreyPSSM, PSSM and SAAC.

Table 2
Analysis of imbalance dataset using different ratios of positive and negative data samples.

Different ratios of Positive Vs. Negative Samples	GreyPSSM				PSSM				SAAC			
	Acc (%)	*Sen* (%)	*Spe* (%)	*BAcc* (%)	*Acc* (%)	*Sen* (%)	*Spe* (%)	*BAcc* (%)	*Acc* (%)	*Sen* (%)	*Spe* (%)	*BAcc* (%)
1:1	93.04	92.62	93.47	93.04	91.95	91.58	92.31	91.94	89.39	90.42	88.35	89.39
1:2	94.79	92.15	96.11	94.13	94.06	90.81	95.69	93.25	87.56	88.53	87.07	87.80
1:3	94.81	89.72	96.51	93.11	94.39	89.75	95.93	92.84	84.57	67.89	90.13	79.01
1:4	95.51	89.34	97.05	93.19	95.41	90.51	96.64	93.57	86.43	59.92	93.06	76.49
1:5	95.65	87.82	97.21	92.52	95.60	88.17	97.09	92.63	87.88	58.08	93.84	75.96
1:6	96.30	88.51	97.60	93.06	96.20	87.97	97.57	92.77	89.40	58.11	94.62	76.37
1:7	96.46	88.35	97.62	92.98	96.03	88.38	97.12	92.75	90.57	57.97	95.23	76.60
1:8	96.88	87.95	97.99	92.97	96.42	87.47	97.54	92.51	91.49	57.81	95.69	76.75
1:9	97.14	87.59	98.20	92.90	96.78	88.26	97.72	92.99	92.29	57.97	96.10	77.04
1:10	97.29	87.54	98.26	92.90	96.85	85.98	97.93	91.96	92.90	57.79	96.41	77.10
1:11	97.37	86.97	98.32	92.65	97.06	87.43	97.97	92.70	93.37	57.78	96.64	77.21

SAAC based ensemble SVM achieved accuracy of 88.12%, sensitivity of 87.79%, specificity of 87.97% and MCC of 0.663. It was noted that GreyPSSM based integrated method obtained higher performance as compared to PSSM and SAAC based integrated method as well as average method. These results demonstrate that GreyPSSM integrated method can yield very reliable and better results. A comparative analysis has been presented in Fig. 2, to demonstrate the reliability and effectiveness of the proposed integrated method with average method. It can be observed from Fig. 2, that integrated method produced higher predictive performance on all feature extraction methods as compared to average method for all performance measure parameters.

4.3. Analyzing the effect of imbalance data on classification

A comparative analysis among GreyPSSM, PSSM and SAAC feature spaces are drawn for imbalance dataset in this section. We analyze the effect of imbalance dataset on the prediction performance of computational predictors. In this regard, different ratios of positive and negative data samples were considered for evaluation of computational predictor. It can be seen from the predictive results given in Table 2 that, although the overall accuracy of the predictor is getting high as we increase the number of negative samples, but the balance accuracy decrease at the same time. This shows that, if the dataset is imbalanced, the classifier bends towards the majority class, thus providing higher overall accuracy. This is not a good approach indeed we want to have a predictor which is can precisely classify both positive and negative samples of data. In order to provide better analysis, we have drawn the predictive results of these methods in Fig. 3 (a) (b) an (c) for GreyPSSM, PSSM and SAAC respectively. The results perceived that Grey model based PSSM, GreyPSSM, is better as compared to PSSM and SAAC based features.

4.4. Comparison between balance and imbalance data

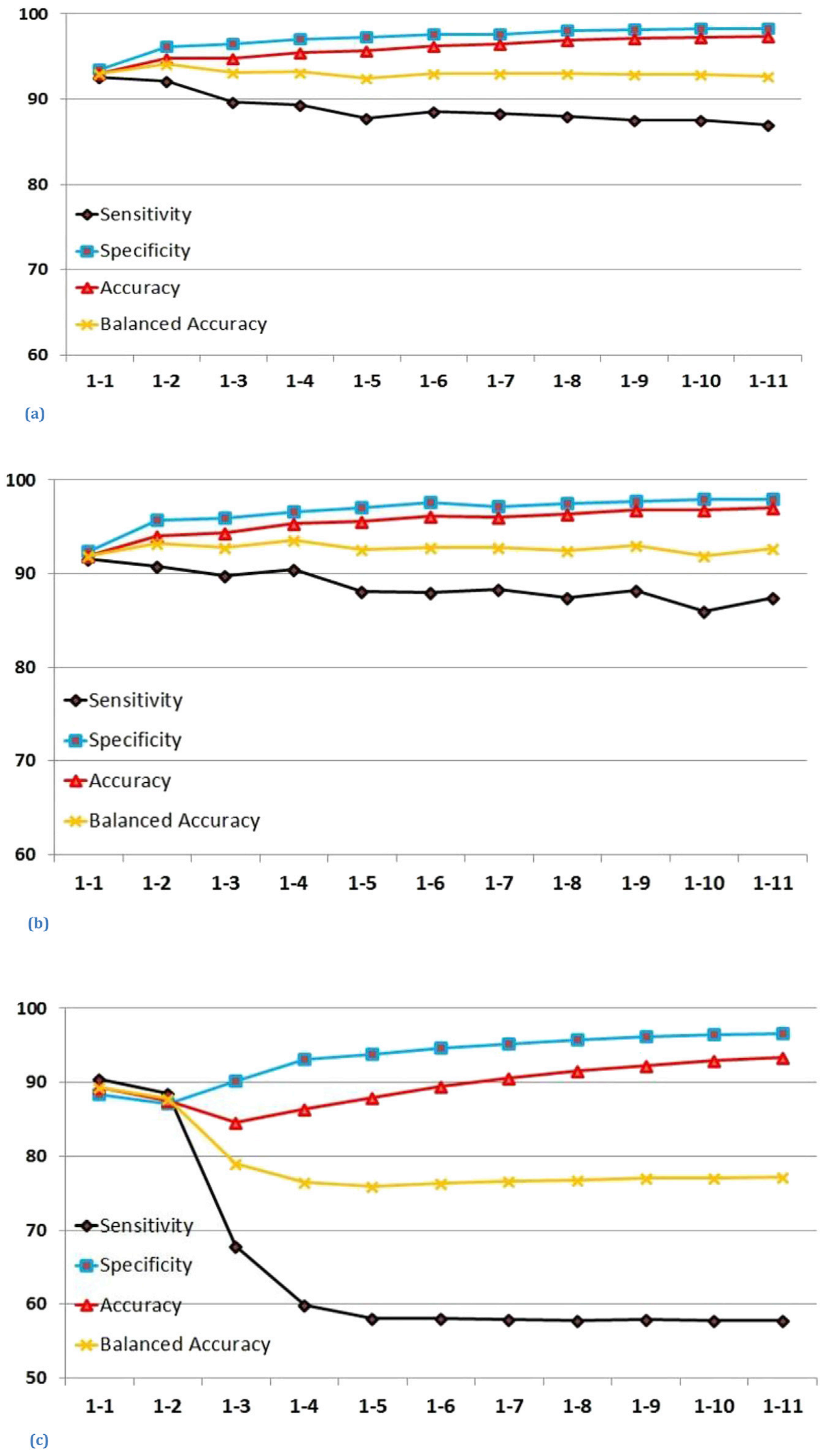
In this section, a comparative analysis has been drawn between balance dataset and imbalance dataset. The empirical results achieved by SVM with balance dataset using asymmetric under-sampling technique and imbalance dataset are illustrated in Table 3. Three different feature extraction methods are applied in order to investigate the strength of proposed model. In case of imbalance dataset, among the feature extraction methods, GreyPSSM based feature space yielded better results. In contrast, the performance of PSSM and SAAC based feature space is quite encouraging but worse as compared to GreyPSSM based feature space. It can be analyzed from Table 3 that the accuracy of GreyPSSM, PSSM and SAAC based features are very high as compared to balanced accuracy. This is because of class imbalance problem; in which classifier tends towards the majority class thus providing higher specificity but lower sensitivity or vice versa. In order to enhance the generalization power of classifier and minimize the effect of biasness, the benchmark dataset has been balanced using asymmetric under-sampling technique. The experimental results of balanced dataset named as integrated method for GreyPSSM, PSSM and SAAC feature spaces are reported in Table 3. The empirical results demonstrated that, the performance of GreyPSSM, PSSM and SAAC based integrated feature spaces are consistent and better than imbalance method. GreyPSSM based feature space in combination with integrated method provided higher ranks for all the performance metrics. Likewise, PSSM and SAAC based feature space in combination with integrated method provided encouraging results but worse than GreyPSSM based integrated method. By revisiting Table 3, it can be analyzed that GreyPSSM based features yielded better prediction performance on both imbalance method and integrated method as compared to PSSM and SAAC based methods.

In addition, the ROC curve has been plotted along with their AUC values as shown in Fig. 4 (a), (b) and (c) for GreyPSSM, PSSM and SAAC feature extraction methods respectively. It can be observed from that proposed integrated method achieved best performance regarding AUC values from imbalanced method on all feature extraction methods.

The following points may be the reason why balance dataset can provide sustainable and persistent results as compared to imbalance problem. (1) First, class imbalance problem can significantly degrade the performance of classifiers. (2) Second, balancing dataset can minimize the effect of biasness (3) third, integrating individual subsets of benchmark dataset using majority voting can incorporate all the sequences therein avoiding class imbalance problem (4) fourth, evolutionary information encoded by grey model system are more reliable as compared to PSSM and SAAC based feature information.

4.5. Comparison of TargetECMP and existing methods on benchmark dataset

In this section, to show the effectiveness and discriminative power, we have drawn a comparative analysis of our method with the existing predictors i.e. ECMPP [2], PECMP [16], ECMPRED [15] IECMP [3] and ECMP-HybKNN [20]. Table 4, demonstrated the values of different predictors for various performance assessment parameters using benchmark dataset. First we compare the anticipated method with other state-of-the-art methods using jackknife cross-validation tests. PECMP method [16] proposed by Zhang et al. provided accuracy of 93.1%, specificity of 97.1%, sensitivity of 49% and balanced accuracy of 73.1%. Later, Jung et al. developed ECMPP method [2] and provided 56.3%, 99.2%, 95.6% and 77.8% of sensitivity, specificity, accuracy and balanced accuracy respectively. Kandaswamy et al. proposed ECMPRED predictor [15] which resulted in 65% of sensitivity, 77% of specificity, 83% of accuracy and 71% of balanced accuracy. Yang et al. proposed an ensemble method and provided encouraging results of 87.8%, 84.9%, 85.1% and 86.4% of sensitivity, specificity, accuracy and balanced accuracy respectively. Recently, Farman et al. developed a hybrid approach with KNN as classification engine. Their predictive results are; 84.1% of sensitivity, 97.8% of specificity, 96.7% of accuracy and 90.9% of



(a)

(b)

(c)

Fig. 3. (a): Effect of imbalance dataset using with different ratios on GreyPSSM. (b): Effect of imbalance dataset using with different ratios on PSSM. (c): Effect of imbalance dataset using with different ratios on SAAC.

Table 3
Comparison imbalance method and integrated method on benchmark.

Prediction Method	GreyPSSM				PSSM				SAAC			
	Acc (%)	*Sen* (%)	*Spe* (%)	*BAcc* (%)	*Acc* (%)	*Sen* (%)	*Spe* (%)	*BAcc* (%)	*Acc* (%)	*Sen* (%)	*Spe* (%)	*BAcc* (%)
Imbalance Method	97.37	86.97	98.32	92.65	97.07	87.43	97.97	92.70	93.37	57.78	96.64	77.21
Integrated Method	94.13	93.16	94.22	93.69	92.44	91.83	92.50	92.16	88.13	89.79	87.97	88.88

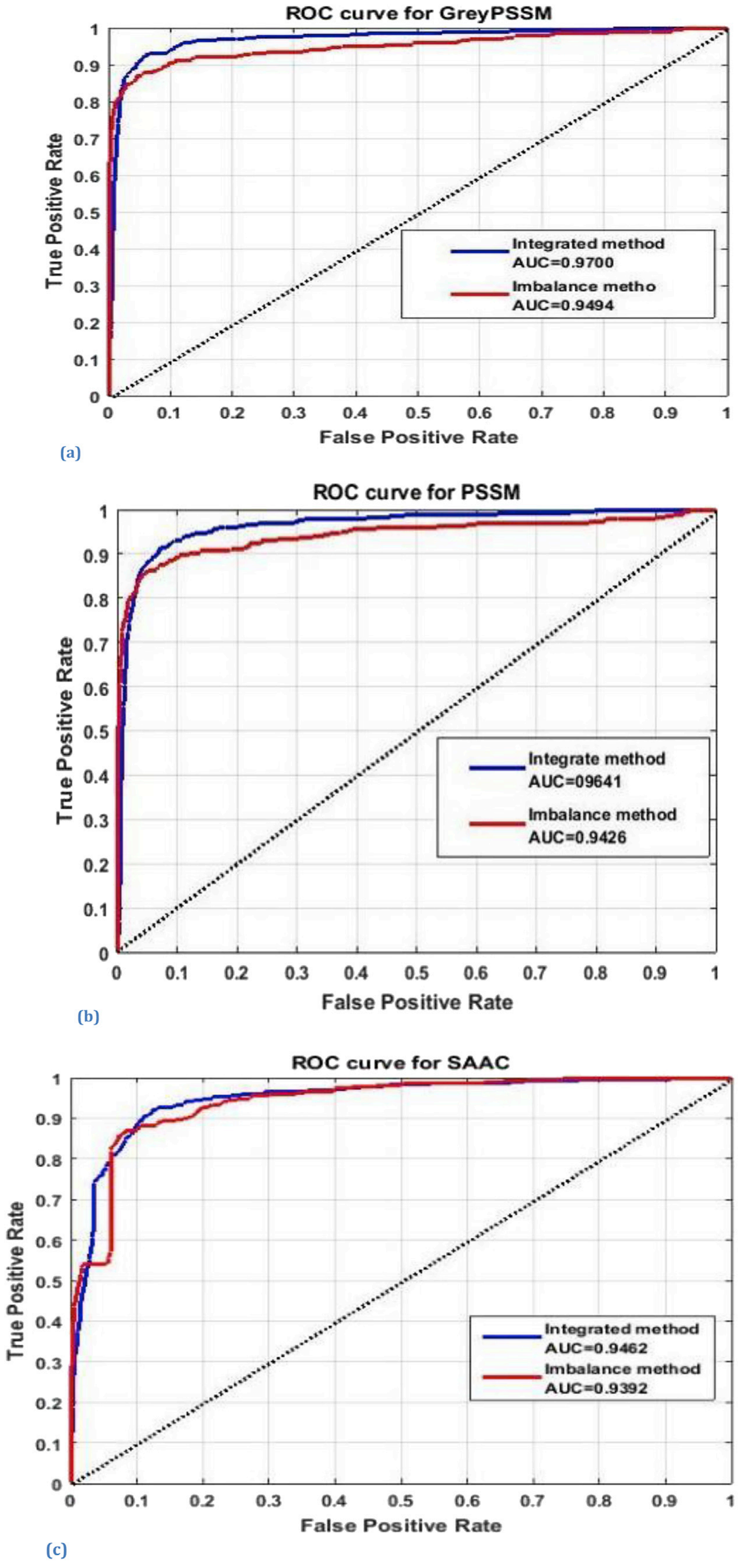


Fig. 4. (a): The ROC curve of integrate method and imbalance method for GreyPSSM. 4(b): The ROC curve of integrate method and imbalance method for PSSM. (c): The ROC curve of integrate method and imbalance method for SAAC.

Table 4
Comparison of proposed method with existing methods on benchmark dataset.

Predictor	*Acc* (%)	*Sen* (%)	*Spe* (%)	*BAcc* (%)
ECMPP [2]	95.6	56.3	99.2	77.8
PECMP [16]	93.1	49.0	97.1	73.1
ECMPRED [15]	83.0	65.0	77.0	71.0
IECMP [3]	85.1	87.8	84.9	86.4
ECMP-HybKNN [20]	96.7	84.1	97.8	90.9
TargetECMP	94.1	93.1	94.2	93.7

balanced accuracy. On the other hand, our proposed method, TargetECMP, enhanced the predictive performance and yielded better results. Our method achieved sensitivity of 93.1%, specificity of 94.2%, and accuracy of 94.1% and the most important metric balanced accuracy of 93.7%. Although, accuracy and specificity of our method is lower as compared to some other methods, but the sensitivity and balanced accuracy of our method is promising. This indicates that our method effectively handled class imbalanced problem with asymmetric under-sampling technique by minimizing the difference between sensitivity and specificity. A graphical representation to compare all the methods for ECMP is illustrated in Fig. 5. Based on the above observation and comparison, we concluded that our method is better is much better than exiting models in the literature so far. All these promising achievements are possible because of asymmetric under-sampling technique with grey model based features.

4.6. Comparison of TargetECMP and existing methods on independent dataset

The performance of our predictor was examined using independent dataset. The predicted outputs for different parameters are illustrated in Table 5. The predicted results indicate that proposed method achieved remarkable results as compared to previously published methods regarding ECM proteins. For the comparison, we selected four state-of-the-art predictors known as ECMPP, PECMP, ECMPRED and IECMP. ECMP-HybKNN [20] method didn't use independent dataset for in their studies that's why we have not mentioned their method in the comparison of independent dataset. The sensitivity produced by our model is 82.3% which is high than that of 29.4%, 43.5%, 65.5% and 76.5% for ECMPP, PECMP, ECMPRED and IECMP respectively. The specificity of our anticipated method is 90.7% which is higher than the specificity of ECMPRED and IECMP but lower than that of ECMPP and PECMP. Likewise, our predictor achieved an accuracy of 87.4% which are higher than other predictors. By looking into the most important parameters i.e. BAcc, proposed predictor outperform other methods on this performance parameter. Proposed method attained 86.5%, ECMPP achieved 64%, PECMP obtained 68.7%, and ECMPRED acquired 55% while the most recent predictor IECMP accomplished 77.5% of balanced accuracy. It can see from the predictive outputs of that proposed method outperform other existing predictor using independent dataset. To provide clearer snapshot of the values listed in Table 5, graphical representation has been drawn in Fig. 6.

Table 5
Comparison of proposed predictor with existing methods on independent dataset.

Predictor	*Acc* (%)	*Sen* (%)	*Spe* (%)	*BAcc* (%)
ECMPP [2]	71.2	29.4	98.5	64.0
PECMP [16]	74.0	43.5	93.8	68.7
ECMPRED [15]	53.5	65.5	47.8	55.0
IECMP [3]	77.7	76.5	78.5	77.5
TargetECMP	87.4	82.3	90.7	86.5

5. Conclusion

In this study, we developed a promising computational predictor for improving the prediction performance of extracellular matrix proteins (ECMP). The benchmark dataset provided for ECMP are imbalance which creates the biasness during the classification. To avoid the biasness and construct effective predictor with higher prediction quality as well as keeping all the data for consideration, we have divided the benchmark dataset into different subsets. During this formulation of subsets, the positive data samples are retained constant and the negative are changed keeping evenly the ratio of negative and positive samples. Sequence based method SAAC, the evolutionary information in the form of PSSM and grey model based on PSSM are utilized for encoding of ECMP sequences. Based on the above criteria we implemented a new computational method called TargetECMP. The proposed method was compared with the existing approaches by performing stringent jackknife tests and independent validation tests. The experimental results achieved by our proposed method outperformed the existing predictor on both datasets. It is anticipated that the outcomes of this study will enrich our understanding of sequence-based ECMP prediction and can potentially be applied to other protein sequence-related prediction and imbalanced datasets problems.

Since user friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods [60,68–70]. In future, we shall make efforts to provide a web-server for the proposed method in this paper.

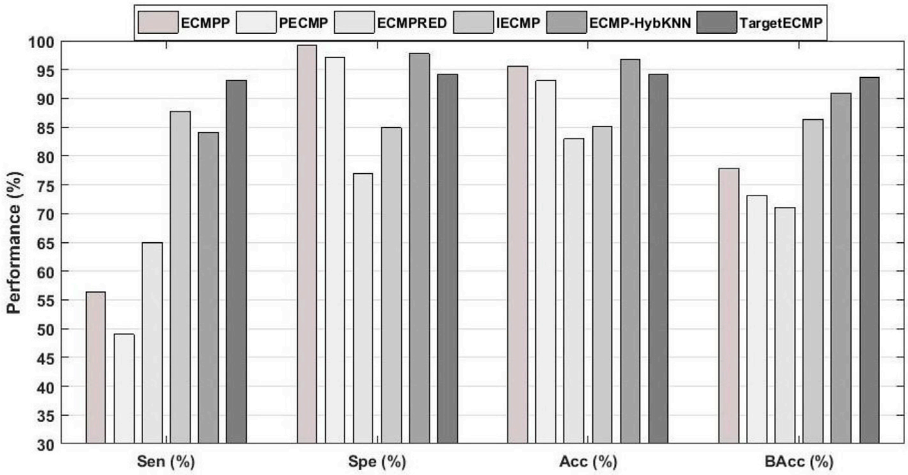


Fig. 5. Performance comparison of various predictors with the TargetECMP on benchmark dataset.

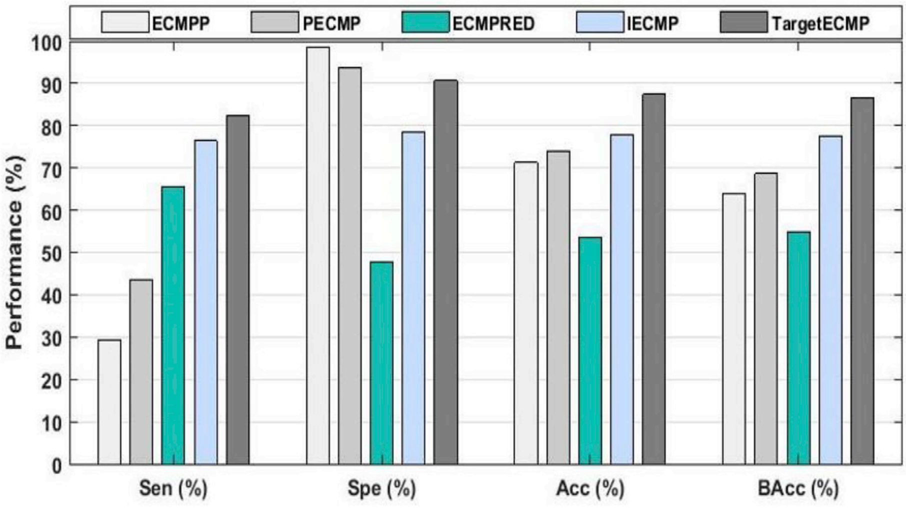


Fig. 6. Comparison of TargetECMP with the existing methods on independent dataset.

Conflicts of interest

The authors declare that they have no interests of conflict.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61772273, 61373062) and the Fundamental Research Funds for the Central Universities (No. 30916011327).

D. J. Yu is the corresponding authors for this paper.

References

[1] J.M. Jacobs, et al., The mammary epithelial cell secretome and its regulation by signal transduction pathways, J. Proteome Res. (2008) 558–569 [cited 7 2].

[2] J. Jung, et al., Prediction of extracellular matrix proteins based on distinctive sequence and domain characteristics, J. Comput. Biol. 17 (1) (2010) 97–105.

[3] R. Yang, et al., An ensemble method with hybrid features to identify extracellular matrix proteins, PLoS One 10 (2) (2015), e0117804.

[4] M.A. Karsdal, et al., Extracellular matrix remodeling: the common denominator in connective tissue diseases possibilities for evaluation and current understanding of the matrix as more than a passive architecture, but a key player in tissue failure, Assay Drug Dev. Technol. 11 (2) (2013) 70–92.

[5] J.F. Chan, et al., Middle East respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease, Clin. Microbiol. Rev. 28 (2) (2015) 465–522.

[6] G.A. Di Lullo, et al., Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen, J. Biol. Chem. 277 (6) (2002) 4223–4231.

[7] B. Kern, et al., Cbfa1 contributes to the osteoblast-specific expression of type I collagen genes, J. Biol. Chem. 276 (10) (2001) 7101–7107.

[8] T.K. Hensch, Critical period mechanisms in developing visual cortex, Curr. Top. Dev. Biol. 69 (2005) 215–237.

[9] C. Chagnot, et al., Bacterial adhesion to animal tissues: protein determinants for recognition of extracellular matrix components, Cell Microbiol. 14 (11) (2012) 1687–1696.

[10] D.Y. Li, et al., Elastin is an essential determinant of arterial morphogenesis, Nature 393 (6682) (1998) 276–280.

[11] J. Rosenbloom, W. Abrams, R. Mecham, Extracellular matrix 4: the elastic fiber, FASEB J. 7 (13) (1993) 1208–1218.

[12] R.J. Peach, et al., Identification of hyaluronic acid binding sites in the extracellular domain of CD44, J. Cell Biol. 122 (1) (1993) 257–264.

[13] P.P. Provenzano, et al., Matrix density-induced mechanoregulation of breast cell phenotype, signaling and gene expression through a FAK–ERK linkage, Oncogene 28 (49) (2009) 4326–4343.

[14] J. Anitha, et al., Prediction of extracellular matrix proteins using SVMhmm classifier, IJCA special issue on advanced computing and communication technologies for HPC applications 1 (2012) 7–11.

[15] K.K. Kandaswamy, et al., EcmPred: prediction of extracellular matrix proteins based on random forest with maximum relevance minimum redundancy feature selection, J. Theor. Biol. 317 (2013) 377–383.

[16] J. Zhang, et al., PECM: prediction of extracellular matrix proteins using the concept of Chou's pseudo amino acid composition, J. Theor. Biol. 363 (2014) 412–418.

[17] Y.-D. Cai, G.-P. Zhou, K.-C. Chou, Support vector machines for predicting membrane protein types by using functional domain composition, Biophys. J. 84 (5) (2003) 3257–3263.

[18] T. Huang, et al., Predicting transcriptional activity of multiple site p53 mutants based on hybrid properties, PLoS One 6 (8) (2011), e22940.

[19] X. Xiao, P. Wang, K.-C. Chou, Predicting the quaternary structure attribute of a protein by hybridizing functional domain composition and pseudo amino acid composition, J. Appl. Crystallogr. 42 (2) (2009) 169–173.

[20] F. Ali, M. Hayat, Machine learning approaches for discrimination of Extracellular Matrix proteins using hybrid feature space, J. Theor. Biol. 403 (2016) 30–37.

[21] H. Lin, The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition, J. Theor. Biol. 252 (2) (2008) 350–356.

[22] H. Lin, W. Chen, Prediction of thermophilic proteins using feature selection technique, J. Microbiol. Meth. 84 (1) (2011) 67–70.

[23] H. Ding, L. Luo, H. Lin, Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition, Protein Pept. Lett. 16 (4) (2009) 351–355.

[24] H. Lin, et al., Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition, Protein Pept. Lett. 15 (7) (2008) 739–744.

[25] M. Khan, et al., Unb-DPC: identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC, J. Theor. Biol. 415 (2017) 13–19.

[26] S. Ahmad, M. Kabir, M. Hayat, Identification of Heat Shock Protein families and J-protein types by incorporating Dipeptide Composition into Chou's general PseAAC, Comput. Meth. Progr. Biomed. 122 (2) (2015) 165–174.

[27] H. Lin, et al., Predicting cancerlectins by the optimal g-gap dipeptides, Sci. Rep. (2015) 5.

[28] J. Hu, et al., A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction, PloS One 9 (9) (2014), e107676.

[29] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, Comput. Intell. 20 (1) (2004) 18–36.

[30] J. Laurikkala, Improving identification of difficult small classes by balancing class distribution, in: Conference on Artificial Intelligence in Medicine in Europe, Springer, 2001.

[31] K.M. Ting, An instance-weighting method to induce cost-sensitive trees, IEEE Trans. Knowl. Data Eng. 14 (3) (2002) 659–665.

[32] S. Ertekin, et al., Learning on the border: active learning in imbalanced data classification, in: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, ACM, 2007.

[33] S. Ertekin, J. Huang, C.L. Giles, Active learning for class imbalance problem, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2007.

[34] G. Wu, E.Y. Chang, KBA: kernel boundary alignment considering imbalanced data distribution, IEEE Trans. Knowl. Data Eng. 17 (6) (2005) 786–795.

[35] B.X. Wang, N. Japkowicz, Boosting support vector machines for imbalanced data sets, Knowl. Inf. Syst. 25 (1) (2010) 1–20.

[36] P. Kang, S. Cho, EUS SVMs: ensemble of under-sampled SVMs for data imbalance problems, in: Neural Information Processing, Springer, 2006.

[37] M. Waris, et al., Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix, Neurocomputing 199 (2016) 154–162.

[38] M. Kabir, D.-J. Yu, Predicting DNase I hypersensitive sites via un-biased pseudo trinucleotide composition, Chemometr. Intell. Lab. Syst. (2017).

[39] L. Ma, S. Fan, CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests, BMC Bioinf. 18 (1) (2017) 169.

[40] W. Lin, D. Xu, Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types, Bioinformatics 32 (24) (2016) 3745–3752.

[41] C. Jia, Y. Zuo, S-SulfPred: a sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique, J. Theor. Biol. 422 (2017) 84–89.

[42] X. Xiao, M. Hui, Z. Liu, iAFP-ense: an ensemble classifier for identifying antifreeze protein by incorporating grey model and PSSM into PseAAC, J. Membr. Biol. 249 (6) (2016) 845–854.

[43] M. Kabir, M. Hayat, iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples, Mol. Genet. Genom. 291 (1) (2016) 285–296.

[44] Z. Liu, et al., pRNAm-PC: predicting N 6-methyladenosine sites in RNA sequences via physical–chemical properties, Anal. Biochem. 497 (2016) 60–67.

[45] X. Cheng, X. Xiao, K.-C. Chou, pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC, Mol. Biosyst. 13 (9) (2017) 1722–1727.

[46] X. Cheng, X. Xiao, K.-C. Chou, pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, Gene 628 (2017) 315–321.

[47] S. Ali, A. Majid, A. Khan, IDM-PhyChm-Ens: intelligent decision-making ensemble methodology for classification of human breast cancer using physicochemical properties of amino acids, Amino Acids 46 (4) (2014) 977–993.

[48] M. Hayat, A. Khan, MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM, J. Theor. Biol. 292 (2012) 93–102.

[49] M. Hayat, A. Khan, M. Yeasin, Prediction of membrane proteins using split amino acid and ensemble classification, Amino Acids 42 (6) (2012) 2447–2460.

[50] X. He, et al., TargetFreeze: identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition, J. Membr. Biol. 248 (6) (2015) 1005–1014.

[51] A. Dehzangi, et al., PSSM-Suc: accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction, J. Theor. Biol. 425 (2017) 97–102.

[52] J. Wang, et al., POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles, Bioinformatics 33 (17) (2017) 2756–2758.

[53] J.L. Deng, Introduction to Grey system theory, J. Grey Syst. 1 (1) (1989) 1–24.

[54] W.-R. Qiu, et al., iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model, J. Biomol. Struct. Dyn. 33 (8) (2015) 1731–1742.

[55] X. Xiao, et al., iCataly-PseAAC: identification of enzymes catalytic sites using sequence evolution information with grey model GM (2, 1), J. Membr. Biol. 248 (6) (2015) 1033–1041.

[56] J.-L. Min, X. Xiao, K.-C. Chou, iEzy-Drug: a web server for identifying the interaction between enzymes and drugs in cellular networking, BioMed Res. Int. (2013) 2013.

[57] J. Hu, et al., TargetCrys: protein crystallization prediction by fusing multi-view features with two-layered SVM, Amino Acids 48 (11) (2016) 2533–2547.

[58] W. Chen, et al., iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties, Bioinformatics 33 (22) (2017) 3518–3523.

[59] P. Feng, et al., iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Mol. Ther. Nucleic Acids 7 (2017) 155–163.

[60] W.-C. Li, et al., iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition, Chemometr. Intell. Lab. Syst. 141 (2015) 100–106.

[61] K. Ahmad, M. Waris, M. Hayat, Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition, J. Membr. Biol. 249 (3) (2016) 293–304.

[62] M. Khan, et al., Bi-PSSM: position specific scoring matrix based intelligent computational model for identification of mycobacterial membrane proteins, J. Theor. Biol. 435 (2017) 116–124.

[63] J. Ahmad, F. Javed, M. Hayat, Intelligent computational model for classification of sub-Golgi protein using oversampling and Fisher feature selection methods, Artif. Intell. Med. 78 (2017) 14–22.

[64] J. Jia, et al., iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC, J. Theor. Biol. 377 (2015) 47–56.

[65] D.-J. Yu, et al., Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble, BMC Bioinf. 15 (1) (2014) 297.

[66] J. Hu, et al., Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs, in: IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2016.

[67] Z.U. Khan, M. Hayat, M.A. Khan, Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model, J. Theor. Biol. 365 (2015) 197–203.

[68] Z.-Y. Liang, et al., Pro54DB: a database for experimentally verified sigma-54 promoters, Bioinformatics 33 (3) (2017) 467–469.

[69] W.-X. Liu, et al., Identifying the subfamilies of voltage-gated potassium channels using feature selection technique, Int. J. Mol. Sci. 15 (7) (2014) 12940–12951.

[70] H. Ding, et al., Prediction of Golgi-resident protein types by using feature selection technique, Chemometr. Intell. Lab. Syst. 124 (2013) 9–13.