DeepECMP: Predicting Extracellular Matrix Proteins using Deep Learning

Mohamed Ghafoor
University of Liverpool
Liverpool, L69 3BX
sgmghafo@liverpool.ac.uk

Anh Nguyen
University of Liverpool
Liverpool, L69 3BX
anh.nguyen@liverpool.ac.uk

Abstract

Introduction: The extracellular matrix (ECM) is a network of proteins and carbohydrates that has a structural and biochemical function. The ECM plays an important role in differentiation, migration and signaling. Several studies have predicted ECM proteins using machine learning algorithms such as Random Forests, K-nearest neighbours and support vector machines but is yet to be explored using deep learning.

Method: DeepECMP was developed using several previously used ECM datasets, asymmetric undersampling and an ensemble of 11 feed-forward neural networks.

Results: The performance of DeepECMP was 83.6% balanced accuracy which outperformed several algorithms. In addition, the pipeline of DeepECMP has been shown to be highly efficient.

Discussion: This paper is the first to focus on utilizing deep learning for ECM prediction. Several limitations are overcome by DeepECMP such as computational expense, availability to the public and usability outside of the human species.

1. Introduction

1.1. Proteins and the Extracellular Matrix

Proteins play an essential role in biological systems. Each protein consists of a variable sequence of 20 amino acids. The protein sequence in turn determines its 3D structure, physio-chemical properties and function. Information about proteins, including their sequences, are available on databases such as UniProtKB [1].

The extracellular matrix (ECM) is a biochemically active scaffold structure within biological tissues (Figure 1). Structurally, it is composed of various combinations of proteins and carbohydrates, such as collagen, proteoglycans and glycoproteins. Functionally, it provides mechanical support to tissues and interacts with several biochemical pathways, such as ECM components interacting with receptors and signaling molecules binding to the ECM [4].

Abnormalities in ECM can be associated with pathologies such as excess deposition in cancers and excess degradation in osteoarthritis [4]. Therefore, an improved understanding of extracellular matrix proteins could lead to novel therapeutic targets for a range of disorders.

1.2. Deep Learning

Machine learning (ML) utilizes algorithms that learn patterns from data and are able to make future predictions. Within ML are neural networks, adaptable functions inspired by biological neurons in the brain, that map given inputs to outputs. A specific class of neural networks with multiple layers, or depth in terms of layers, are referred to as deep neural networks or deep learning (DL). DL, unlike classical ML, does not require feature extraction and hence
domain specific knowledge [7, 5]. Moreover, DL has various architectures suitable for various tasks, such as convoluted neural networks for image recognition [11] and recurrent neural networks for text processing [12].

1.3. Predicting Extracellular Matrix Proteins Using Machine Learning

Several papers have utilized ML to predict ECM proteins. The first ECM predictor was ECMPP [8] which used 109 ECM proteins and 1,430 non-ECM proteins, 96 features were generated using sequential information (amino acid composition and protein domains) and evolutionary information. The feature selection method used was mean decrease accuracy. From these, 13 optimal features were selected and classified using random forest (RF) and support vector machine (SVM) algorithms [8]. The performance of ECMPP was 95.6% accuracy, 56.3% sensitivity, 99.2% specificity and 77.8% balanced accuracy.

ECMPred provided a larger dataset that was utilized in subsequent studies, which included 445 metazoan ECM and 4464 metazoan non-ECM proteins. In addition, 20 experimentally verified ECM proteins were introduced. Protein sequence information (functional groups at C and N terminals) as well as physio-chemical properties were used to derive 68 features. The algorithm used was RF and feature selection used was maximum relevance minimum redundancy. The performance was, 77.0% accuracy, 65.0% sensitivity, 77.0% specificity and 71.0% balanced accuracy. The authors scanned the whole human genome in search for novel ECM proteins. In addition, the software was made available at the time of publication, however it is no longer available [10].

PECM [17] utilized the dataset created in ECMPred as the training dataset with proteins of length over 3000 or below 50 removed. In addition, proteins with ambiguous amino acids (B,J,O,U,X and Z) were removed. This resulted in 410 metazoan ECM proteins and 4464 metazoan non-ECM proteins as the training dataset. The study also created an independent human test set, which included 85 ECM proteins and 130 non-ECM proteins. The features extracted included sequence information (77 features), physio-chemical properties (82 features), evolutionary information (80 features) and structural information (76 features). The feature selection method was Fisher-Markov selector and the ML algorithm SVM. The performance of the algorithm was 86.5% accuracy, 75.9% sensitivity, 86.9% specificity and 81.4% balanced accuracy. PECM was also the first to introduce position specific scoring matrix, an important albeit computationally expensive feature.

Identify ECM Proteins (IECMP) [15] used the dataset developed by ECMPred for training and testing. In addition, an independent dataset containing 159 human non-ECM proteins and 3681 human non-ECM proteins were used for testing. The features extracted were sequence information (77 features), physiochemical features (82 features), evolutionary information (80 features) and structural information (76 features). Importantly, IECMP was the first to introduce under-sampling and an ensemble as a solution to class imbalance. A webserver was originally made available however it is no longer so. Features were ranked using information gain ratio and incremental feature selection with an ensemble of 11 RF classifiers. The performance of the algorithm was 85.1% accuracy, 87.8% sensitivity, 84.9% specificity and 86.6% balanced accuracy.

ECMP-HybKNN [2] was the first algorithm to use K-nearest Neighbour (KNN) to predict ECM proteins. The dataset used was again the same developed for ECM-Pred. Sequence information was used to extract features such as amino acid composition, di-peptide composition and pseudo amino acid composition. The feature selection method used was maximum relevance minimum redundancy. The accuracy was 96.7%, sensitivity 84.2%, specificity 97.8% and balanced accuracy 91.0%. HybKNN produced significant improvements to the previous algorithms, however the algorithm was not made available to the public.

BAMORF used the ECMPred dataset and used sequence information and physio-chemical proprieties as the main features. Feature selection used was binary animal migration and the algorithm of choice a RF. The performance of the algorithm was 85.0% sensitivity, 86.5% specificity, and 85.7% balanced accuracy. The tool was not publically available [6].

TargetECMP [9] only used grey position specific scoring matrix (PSSM) as the main feature and also utilized the ECMPred dataset. The performance was significantly improved to 93.1% sensitivity, 94.2% specificity and 93.7% balanced accuracy. The algorithm used was SVM.

ECMPride [13] was tested on both the ECMPred dataset and a novel human ECM dataset, which included 521 human ECM proteins and 11,336 human non-ECM proteins. A total of 167 features were extracted using ECM-domains, physio-chemical properties and position-specific scoring matrices. Feature selection method was maximum relevance minimum redundancy. Undersampling was utilized with an ensemble of 99 RFs. The performance of ECM-Pride on the ECMPred dataset was 86.4% accuracy, 87.8% sensitivity, 86.2% specificity and 87.0% balanced accuracy. On the new, human ECM dataset, balanced accuracy increased to 91.4%.

The current state of the art, ECMP-RF [14] which utilized the ECMPred dataset. Synthetic Minority Oversampling was utilized with elastic net and an RF to produce significant improvements. A sensitivity of 98.8%, specificity of 95.8% and balanced accuracy of 97.3%. The source code was made available, however a webserver or a trained algorithm to be utilized by the public was not.
All studies focused on classical ML algorithms, such as RF, SVM and KNN while DL has yet to be explored. A limitation of current algorithms is accessibility to the general public. Of the above algorithms, only ECMPride is available to the public in a relatively user friendly manner. Moreover, ECMPride can only be utilized to make predictions about human proteins.

1.4. Our Contribution

This paper has four contributions. Firstly, we are the first to focus on DL for the prediction of ECM proteins. Secondly, we are the first to utilize ProtVec in the prediction of ECM proteins, creating a more efficient pipeline. Thirdly, DeepECMP is the only publicly available algorithm that can be used to predict ECM proteins outside of the human species. Finally, all our source code and data has been shared to encourage further optimization.

2. Methods

2.1. Datasets

Several datasets have been developed for the prediction of ECM proteins. To enable comparability, DeepECMP was tested on all the available datasets. These include a benchmark (ECMPred), ECMPride and an independent dataset. The benchmark dataset used in several studies is the Kandaswamy dataset, which includes 445 metazoan ECM proteins and 4,187 metazoan non-ECM proteins. The dataset is imbalanced as there are significantly more negative samples compared to positives [10].

The independent dataset has 85 human ECM proteins and 130 human non-ECM proteins [16].

The ECMPride dataset has 521 human ECM proteins and 11,336 human intracellular proteins [13].

2.2. Protein Embedding

ProtVec is a pre-trained skip-gram neural network which embeds proteins of variable length into 100-d vectors. It was created for general use in bioinformatics and has been used for protein family classification [3]. A limitation of using ProtVec is sequences with ambiguous are required to be removed, although they represent a small subset of datasets.

2.3. Feed Forward Neural Network Architectures

The pipeline of DeepECMP involves the use of 11 feed-forward neural networks (FNNs). Each FNN architecture has 5 layers with 512, 256, 128, 128 and 1 units respectively and l2 regularization set to 0.0001. The activation functions used were rectified linear unit (ReLU) for the hidden layers and sigmoid for the output layer. The optimizer was set to adaptive learning rate (Adam). (Figure 2).

2.4. Performance Evaluation

For evaluation on the benchmark dataset and on the ECMPride dataset, 10-fold stratified cross validation was used. To create each fold, 1/10th of the positive and negative samples were placed into each fold. Subsequently, 1 fold is used for testing while the remaining 9 folds are used for training. The process is repeated 10 times where each fold acts as the test fold.

For evaluation on the independent dataset, the whole benchmark dataset was used for training.

The metrics used in this study were sensitivity, specificity, accuracy and balanced accuracy for DeepECMP:

- **Accuracy**(Acc) = \( \frac{TP + TN}{TP + TN + FP + FN} \)
- **Sensitivity**(SN) = \( \frac{TP}{TP + FN} \)
- **Specificity**(SP) = \( \frac{TN}{TN + FP} \)
- **BalancedAccuracy**(BAcc) = \( \frac{SN + SP}{2} \)

2.5. Under-sampling and Deep Ensemble

In the DeepECMP pipeline, asymmetric under-sampling was used to create 11 subsets of the original dataset. Asymmetric under-sampling was only utilized on the benchmark dataset, it has significantly more negative samples compared to positive samples.

All of the positive samples are used and are the same in each of the new subsets (e.g. if 400 positive examples, each of the 11 new datasets will contain 400 positive samples). However, 11 subsets equal in size to the positive samples are selected from the negative samples (e.g. if 4000 negative samples, 400 are selected randomly). This results in 11
new datasets that contain the same number of positive and negative examples. Each of the 11 new datasets are used to train 11 FNNs.

3. Results and Discussion

3.1. DeepECMP

3.1.1 The Effect of Imbalanced Dataset

To explore the effect of using a larger imbalanced training dataset, different ratios of positive samples to negative were utilized (1:1, 1:2, 1:3, 1:4 and 1:5). A balanced dataset resulted in high sensitivity but low specificity. However, as the dataset become more imbalanced, there was a reduction in sensitivity and an increase in specificity (Figure 3). Simply using more data does not improve performance in an imbalanced dataset.

![Figure 3. Imbalanced Dataset. Different ratios of positive to negative samples were created. A 1:1 ratio resulted in low specificity and high sensitivity. However, as the ratio of negative samples to positives increased, there was a reduction in sensitivity and an increase in specificity.](image)

3.1.2 DeepECMP Pipeline

DeepECMP is available as a Google Co-labs notebook which has been set up in a user friendly manner for simple prediction of ECM proteins along with a user manual. In addition, the code and datasets used in this study are available on the following link: GitHub.

The DeepECMP pipeline begins with creating 11 new subsets from the original dataset using under sampling. Subsequently, all sequences are represented as a 100-d vector using ProtVec. Each of the 11 newly created datasets are then used to train an ensemble of 11 FNNs. To make a prediction, the input enters each FNN which each output a prediction, the prediction with the highest frequency becomes the final prediction of the system (Figure 4).

| Method      | Acc  | Sen  | Spe  | BAce |
|-------------|------|------|------|------|
| Single      | 78.54| 78.69| 53.17| 65.93|
| Ensemble    | 79.99| 88.01| 79.13| 83.57|

3.1.3 Asymmetric Under-sampling and A Deep Learning Ensemble Improved Performance

Asymmetric under-sampling has been used to improve algorithmic performance while using an imbalanced dataset \[16, 13\]. In this study asymmetric under-sampling was used to create 11 subsets that were utilized to train 11 separate FNNs. For a prediction to be made, the input is run through each model and the most common prediction is selected as the final prediction.

The ensemble of FNNs significantly improved performance compared to a single FNN trained on a balanced dataset (Table 1).

3.1.4 Comparison to Other Predictors

DeepECMP reached 83.6% balanced accuracy on the benchmark dataset (Table 2), 71.0% balanced accuracy on the independent dataset (Table 3) and 77.5% balanced accuracy on the ECMPride dataset (Table 4). By using balanced accuracy as the metric to rank performance, on the benchmark dataset DeepECMP ranked 4/7, on the independent dataset 3/7 and 3/3 on the ECMPride dataset.

3.1.5 DeepECMP is highly efficient and publically available

To test the efficiency of PSSM compared to ProtVec, five proteins from the benchmark dataset were selected randomly. The time taken to generate a PSSM using the standard configuration of NCBI PSI-BLAST was monitored. In addition, the standard ProtVec method utilized in DeepECMP was performed for comparison. The results highlight the computational expense associated with PSSM (Table 5).

Although the performance of DeepECMP is not the best available, it is among the most efficient along with ECMP, ECMpred and ECMP-HybKNN which do not use PSSM. Among these, based on balanced accuracy, only ECMP-HybKNN outperforms DeepECMP. ECMPride used PSSM and is publically available however the authors pre-calculated the PSSM features for the entire human proteome which limits ECMPride to prediction of human proteins. A limitation of ECMP-HybKNN and other algorithms that outperform DeepECMP is they are not available to the public.

DeepECMP has been made publically available as a Google Co-labs Notebook. DeepECMP and a detailed user
Asymmetric under-sampling is first used to create 11 balanced subsets of the original unbalanced dataset. Protein sequences are then converted into a 100-d vector using ProtVec. An ensemble of 11 feed-forward neural networks are trained using the balanced datasets. When making a prediction, each trained feed-forward neural networks make a prediction and the most common becomes the final prediction. The metrics used in this study include specificity, sensitivity and balanced accuracy.

**Table 2. Comparison of methods on benchmark dataset**

| Reference | Predictor | Method | Acc  | Sen  | Spe  | BAnc |
|-----------|-----------|--------|------|------|------|------|
| [8]       | ECMPP     | RF     | 95.6 | 56.3 | 99.2 | 77.8 |
| [17]      | PECMP     | SVM    | 93.1 | 49.0 | 97.1 | 73.1 |
| [10]      | ECMPRED   | RF     | 83.0 | 65.0 | 77.0 | 71.0 |
| [15]      | IECMP     | RF     | 85.1 | 84.9 | 87.0 | 86.4 |
| [2]       | ECMP-HybKNN | KNN   | 96.7 | 84.1 | 97.2 | 90.9 |
| [9]       | TargetECMP | SVM   | 94.1 | 93.1 | 94.2 | 93.7 |
| Our Method | DeepECMP  | FNN    | 80.0 | 88.0 | 79.1 | 83.6 |

**Table 3. Comparison of methods on independent dataset**

| Predictor   | Sen  | Spe  | Acc  | BAnc |
|-------------|------|------|------|------|
| ECMPP       | 29.4 | 98.5 | 71.2 | 64.0 |
| PECMP       | 43.5 | 93.8 | 74.0 | 68.7 |
| EcmPred     | 62.2 | 47.8 | 53.5 | 55.0 |
| IECMP       | 76.5 | 78.5 | 77.7 | 77.5 |
| ECMPride    | 87.8 | 86.2 | 86.4 | 87.0 |
| TargetECMP  | 87.4 | 82.3 | 90.7 | 86.5 |
| DeepECMP    | 68.2 | 73.9 | 71.6 | 71.0 |

**Table 4. ECMPride dataset comparison**

| Predictor   | Sen  | Spe  | Acc  | BAnc |
|-------------|------|------|------|------|
| ECMPride    | 89.3 | 93.6 | 92.4 | 91.4 |
| EcmPred     | 84.6 | 91.6 | 91.5 | 88.1 |
| DeepECMP    | 83.7 | 71.3 | 71.8 | 77.5 |

The main disadvantage of DeepECMP is the lower performance. However, DeepECMP includes the whole pipeline required to make a prediction, is significantly faster and has no limitations on the sequence to be predicted (such as only human proteins).
Table 5. Time comparison of PSSM and ProtVec

| Protein   | PSSM Time (s) | ProtVec Time (s) |
|-----------|---------------|------------------|
| Q9P0Z3    | 61.845        | 0.017            |
| P54356    | 62.561        | 0.01             |
| A9QWP9    | 61.838        | 0.003            |
| P18917    | 60.921        | 0.006            |

4. Conclusion

A novel DL algorithm was developed in this study that could efficiently predict ECM proteins. Several studies have focused on predicting ECM proteins using ML and have reached high performance. However, there exists several limitation of current algorithms: 1) high computational expense, 2) lack of availability to the public and 3) usable only on human proteins. ECMPrildr overcome the computational expense limitation by pre-calculating the extraction of features such as PSSM/domains for the entire human genome, however this limits the algorithm to only predicting human proteins. Although not as high performing as other algorithms which are not publicly available, ECMPrildr is a better option than DeepECMP if the proteins to be predicted are part of the human genome. If a member of the public would like to predict whether a non human protein is part of the ECM in a binary fashion, DeepECMP is the only option available. Furthermore, the datasets and architecture of DeepECMP has been made available for further development.

References

[1] Uniprot: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49(D1):D480–D489, 2021.
[2] F. Ali and M. Hayat. Machine learning approaches for discrimination of extracellular matrix proteins using hybrid feature space. Journal of theoretical biology, 403:30–37, 2016.
[3] E. Asgari and M. R. Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. PloS one, 10(11):e0141287, 2015.
[4] C. Bonnans, J. Chou, and Z. Werb. Remodelling the extracellular matrix in development and disease. Nature reviews Molecular cell biology, 15(12):786–801, 2014.
[5] T.-T. Do, A. Nguyen, and I. Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In ICRA, 2018.
[6] L. Guan, S. Zhang, and H. Xu. Bamorf: a novel computational method for predicting the extracellular matrix proteins. IEEE Access, 5:18498–18505, 2017.
[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
[8] J. Jung, T. Ryu, Y. Hwang, E. Lee, and D. Lee. Prediction of extracellular matrix proteins based on distinctive sequence and domain characteristics. Journal of Computational Biology, 17(1):97–105, 2010.
[9] M. Kabir, S. Ahmad, M. Iqbal, Z. N. K. Swati, Z. Liu, and D.-J. Yu. Improving prediction of extracellular matrix proteins using evolutionary information via a grey system model and asymmetric under-sampling technique. Chemometrics and Intelligent Laboratory Systems, 174:22–32, 2018.
[10] K. K. Kandaswamy, G. Pugalenthii, K.-U. Kalies, E. Hartmann, and T. Martinetz. Ecmpred: Prediction of extracellular matrix proteins based on random forest with maximum relevance minimum redundancy feature selection. Journal of theoretical biology, 317:377–383, 2013.
[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. NIPS, 2012.
[12] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. nature, 521(7553):436–444, 2015.
[13] B. Liu, L. Leng, X. Sun, Y. Wang, J. Ma, and Y. Zhu. Ecmpred: prediction of human extracellular matrix proteins based on the ideal dataset using hybrid features with domain evidence. PeerJ, 8:e9066, 2020.
[14] M. Wang, L. Yue, X. Cui, C. Chen, H. Zhou, Q. Ma, and B. Yu. Prediction of extracellular matrix proteins by fusing multiple feature information, elastic net, and random forest algorithm. Mathematics, 8(2):169, 2020.
[15] R. Yang, C. Zhang, R. Gao, and L. Zhang. An ensemble method with hybrid features to identify extracellular matrix proteins. PloS one, 10(2):e0117804, 2015.
[16] R. Yang, C. Zhang, R. Gao, and L. Zhang. An ensemble method with hybrid features to identify extracellular matrix proteins. PloS one, 10(2):e0117804, 2015.
[17] J. Zhang, P. Sun, X. Zhao, and Z. Ma. Pecm: Prediction of extracellular matrix proteins using the concept of chou’s pseudo amino acid composition. Journal of Theoretical Biology, 363:412–418, 2014.