ELSEVIER
Journal of
Theoretical
Biology

Journal of Theoretical Biology 317 (2013) 377–383

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

EcmPred: Prediction of extracellular matrix proteins based on random forest with maximum relevance minimum redundancy feature selection

Krishna Kumar Kandaswamy [a,b,c,*], Ganesan Pugalenthi [d], Kai-Uwe Kalies [e], Enno Hartmann [e], Thomas Martinetz [a]

[a] *Institute for Neuro- and Bioinformatics, University of Luebeck, Germany*
[b] *Graduate School for Computing in Medicine and Life Sciences, University of Luebeck, Germany*
[c] *Max Planck Institute for Biology of Ageing, Germany*
[d] *Bioinformatics Group, Bioscience Core Lab, King Abdullah University of Science and Technology (KAUST), Saudi Arabia*
[e] *Centre for Structural and Cell Biology in Medicine, Institute of Biology, University of Luebeck, Germany*

HIGHLIGHTS

- ► The random forest algorithm has been used to predict extracellular matrix proteins.
- ► We could extract the best features using mRMR feature selection.
- ► We have identified novel extracellular matrix proteins in the human proteome.
- ► The results are compared with previous work, and our algorithm is more advanced.

ARTICLE INFO

Article history:
Received 18 January 2012
Received in revised form
8 October 2012
Accepted 9 October 2012
Available online 1 November 2012

Keywords:
Extracellular proteins
Sequence properties
Maximum relevance minimum redundancy (mRMR)
Random forest
Human proteome

ABSTRACT

The extracellular matrix (ECM) is a major component of tissues of multicellular organisms. It consists of secreted macromolecules, mainly polysaccharides and glycoproteins. Malfunctions of ECM proteins lead to severe disorders such as marfan syndrome, osteogenesis imperfecta, numerous chondrodysplasias, and skin diseases. In this work, we report a random forest approach, EcmPred, for the prediction of ECM proteins from protein sequences. EcmPred was trained on a dataset containing 300 ECM and 300 non-ECM and tested on a dataset containing 145 ECM and 4187 non-ECM proteins. EcmPred achieved 83% accuracy on the training and 77% on the test dataset. EcmPred predicted 15 out of 20 experimentally verified ECM proteins. By scanning the entire human proteome, we predicted novel ECM proteins validated with gene ontology and InterPro. The dataset and standalone version of the EcmPred software is available at http://www.inb.uni-luebeck.de/tools-demos/Extracellular_matrix_proteins/EcmPred.

1. Introduction

The tissues of multicellular organisms are formed by cells and a network of macromolecules secreted by them, which is called extracellular matrix (ECM) (Lewin et al., 2007). It consists of glycosaminoglycans, proteoglycans, fibrous proteins like collagenes, adhesive glycoproteins, enzymes involved in formation and remodeling of the ECM, like metalloproteases, and other factors (Lewin et al., 2007). In the tissues the ECM integrates the cells and provides structural support. In addition, it also influences the fate of cells during differentiation, morphogenesis, aging or pathogenesis (Schwartz et al., 1995; Burridge and Chrzanowska-Wodnicka, 1996; Wary et al., 1996). The ECM can coordinate cell functions by transducing signals across the plasma membrane. This can be achieved either directly by ECM molecules or indirectly by signal molecules, like growth factors, cytokines, chemokines, and hormones, which are sequestered in local depots within the ECM (Kim et al., 2011; Nelson and Bissell, 2006). At first glance, the extracellular matrix seems to be a static structure with a slow turnover. However, it turned out that the ECM can easily adapt to changing conditions by a dynamic remodeling of its compounds (Green and Lund, 2005).

Malfunctions of ECM proteins lead to severe disorders that are linked to the structural functions of ECM molecules, such as the marfan syndrome, osteogenesis imperfecta, numerous chondrodysplasias, and skin diseases (Green and Lund, 2005; Aszodi,

[*] Corresponding author at: Max Planck Institute for Biology of Ageing, Germany. Tel.: +49 22147889664; fax: +49 22147897402.
E-mail address: Krishna.Kandaswamy@age.mpg.de (K.K. Kandaswamy).

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.10.015

2006; Bateman et al., 2009; Bruckner-Tuderman and Bruckner, 1998). Moreover, tumor growth, metastasis, inflammation, and other disorders can occur as a consequence of ECM malfunctions (Nelson and Bissell, 2006; Campbell et al., 2010; Sorokin, 2010). Thus, extracellular matrix proteins promise great possibilities as therapeutic targets or diagnostic markers (Grønborg et al., 2006).

Due to advances in sequencing technologies, tremendous amounts of DNA and protein sequences have accumulated in databases. Most of these sequences have unknown functions. It is very important to extract relevant biological information from sequences for functional annotation. Since the function of a protein is closely associated with its subcellular localization, the ability to predict the protein's subcellular localization will be useful in the characterization of the expressed sequences of unknown functions (Horton et al., 2007; Chou and Shen, 2010a).

Various machine learning methods are available for predicting protein subcellular localizations (Chou and Shen, 2010b; Shen and Chou, 2009; Chou and Shen, 2007a; Chou and Shen, 2007b; Chou et al., 2011). Protein subcellular localization prediction for eukaryotes (Chou and Shen, 2007b; Chou et al., 2011), humans (Chou and Shen, 2006a; Chou et al., 2012), plants (Chou and Shen, 2006b; Wu et al., 2011), viruses (Chou and Shen, 2006c; Xiao et al., 2011), gram negative bacteria (Chou and Shen, 2006d) and gram positive bacteria (Wu et al., 2012) have also been carried out. Several methods have been proposed for the identification of secretory proteins that follow the classical secretory pathway (Bendtsen et al., 2004) and non-classical secretory pathway (Kandaswamy et al., 2010). Even though there are various tools available for predicting subcellular localizations and protein secretion, there is no method with sufficient accuracy to predict ECM proteins among the secreted protein groups.

Recently, an in-silico model (ECMPP) has been developed to predict ECM proteins (Jung et al., 2010). It uses the Support Vector Machine (SVM) and Random Forest (RF) to distinguish ECM proteins based on 13 distinctive features. However, the performance of this method mainly depends on the position specific scoring matrix (PSSM) profile, which needs sufficiently many sequence homologs to derive a sequence alignment. In this work, we present a random forest method, EcmPred, to identify extracellular matrix (ECM) proteins from sequence derived properties such as frequency of amino acid/amino acid groups and physicochemical properties. EcmPred achieves 83% and 77% accuracy on training and test data, respectively.

2. Materials and methods

2.1. Data set

We performed an extensive database and literature curation to collect sequences pertaining to extracellular matrix proteins. The dataset containing 17233 Metazoan secreted protein sequences was obtained from Swiss-Prot release 67 (Boeckmann et al., 2003). Out of these 17233 sequences, 1103 sequences are extracellular matrix proteins (positive dataset), and the remaining 16130 proteins are secreted proteins without extracellular matrix annotation (negative dataset). The positive and negative datasets were made completely non-redundant by allowing a sequence identity between any two proteins of not more than 70% (Li et al., 2001). Finally, the training dataset consisted of 445 extracellular proteins that form the positive dataset and 4187 non-ECM proteins that form the negative dataset.

2.1.1. Training set

300 ECM proteins were randomly selected from the 445 ECM proteins for the positive training dataset. Similarly, 300 non-ECM proteins were randomly taken from the 4187 non-ECM proteins for the negative training dataset.

2.1.2. Test set

The remaining 145 ECM proteins served as positive dataset for testing. The remaining 3887 non-ECM proteins (after excluding 300 non-ECM proteins for training) were used as a negative dataset for testing.

2.1.3. Human proteome screening

A human proteome database containing 86845 protein sequences was downloaded from the IPI database release 3.66 (http://www.ebi.ac.uk/IPI/) (Kersey et al., 2004). Transmembrane proteins were removed using TMHMM (Krogh et al., 2001). Finally, we obtained 65508 protein sequences for the computational screening and identification of novel ECM proteins.

2.2. Features

Amino acid composition is one of the most basic characteristics of the proteome and is extensively used in sequence based prediction studies (Horton et al., 2007). Instead of using the conventional 20-D amino acid composition, another new concept called "pseudo amino acid composition" has been reported in order to include the sequence-order information which leads to a higher success rate in sequence based prediction studies (Chou, 2001; Chou, 2005; Shen and Chou, 2006; Chou and Cai, 2005). PseAAC, a server based on the concept of pseudo amino acid composition, provides a flexible way to generate various kinds of pseudo amino acid compositions for a given protein sequence (Chou, 2001; Chou, 2005). The general form of PseAAC for a protein P can be expressed as (see Eq.(6) of Chou (2011)).

$$P = [\psi_1 \quad \psi_2 \quad \psi_3 \ldots \psi_u \ldots \psi_\Omega]^T \tag{1}$$

The subscript omega (Ω) is 68 and ψ_u (u=1, 2,., 68) is corresponding to each of the 68 sequence derived features (Table 1).

It has been reported that signal peptides play a vital role in protein secretion (Walter et al., 1984). Generally, signal peptides occur within the first 30 residues from N-terminal. In order to use signal peptide information, each sequence is split into two segments. For a sequence with L residues length, the first 30 residues from the N terminal (residues 1–30) form segment 1 and the remaining residues (residues 31–L) form segment 2.

Frequency of 10 functional groups: We categorized 20 amino acids into 10 functional groups based on the presence of side chain chemical groups such as phenyl (F/W/Y), carboxyl (D/E), imidazole (H), primary amine (K), guanidino (R), thiol (C), sulfur (M), amido (Q/N), hydroxyl (S/T), and non-polar (A/G/I/L/V/P) (Kandaswamy et al., 2010). The frequencies of these 10 functional groups were calculated for segment 1 and 2.

Table 1
List of 68 features.

Name of the feature	Number of features
Frequency of 10 functional groups in segment 1 (first 30 residues from N-terminal)	10
Frequency of 24 physicochemical properties in segment 1 (first 30 residues from N-terminal)	24
Frequency of 10 functional groups in segment 2	10
Frequency of 24 physicochemical properties in segment 2	24
Total	68

Physicochemical properties: We took 24 physicochemical properties from the UMBC AAIndex database (Kawashima et al., 2008). These physicochemical properties include molecular weight, hydrophobicity, hydrophilicity, refractivity, average accessible surface area, flexibility, melting point, side chain volume, side chain hydrophobicity, normalized frequency of beta-sheet and alpha helix, refractivity, membrane buriability, retention coefficient, steric hindrance, optical activity, polarity, heat capacity, and isoelectric point. For each sequence, 24 physicochemical property values were calculated by taking the sum of each physicochemical property value over all residues of the sequence and divide it by the length of the sequence.

2.3. Classification protocol

Random forests (RF) (Breiman, 2001) have been used for a large number of classification as well as regression tasks (Kandaswamy et al., 2010; Kumar et al., 2009; Dudoit et al., 2002; Lee et al., 2005; Jia and Hu, 2011; Kandaswamy et al., 2011; Lin et al., 2011; Pugalenthi et al., 2012; Qiu and Wang, 2011; Shameer et al., 2011). A typical random forest consists of a set of binary decision trees (Ho et al., 1994). Random forests generate multiple decision trees for a given training set and use a weighted average of the trees for the final decision (Breiman, 2001).

Random forest is a very popular ensemble method that is robust to noise, robust against overfitting, fast, and offers possibilities for an explanation and visualization of its outputs. Random forest ‘‘grows’’ and combines a large number of classification trees (Ho et al., 1994; Ho, 1998; Ho, 2002). Two random elements serve to obtain a random forest, bagging and random split selection. Bagging is done here by sampling multiple times with replacement from the original training data set. Thus in the resulting samples, a certain event may appear several times, and other events not at all (Breiman, 2001).

Random forests are trained in a supervised way. Training involves a tree construction as well as assigning to each leaf node the information about the training samples reaching this leaf node, e.g. the class distribution in the case of classification tasks. At runtime, a test sample is passed down all the trees of the forest, and the output is computed by averaging the distributions recorded at the reached leaf nodes. The RF algorithm was implemented with the random forest R package (Liaw and Wiener, 2002).

2.4. Maximum Relevance Minimum Redundancy (mRMR)

Feature selection is important in many pattern recognition problems for excluding irrelevant and redundant features. It allows to reduce system complexity and processing time and often improves the recognition accuracy. The minimal-redundancy-maximal-relevance (mRMR) algorithm is a sequential forward selection algorithm and was first developed by Peng et al. (2005) to analyze the importance of different features. mRMR uses mutual information to select M features that best fulfill the minimal redundancy and maximal relevance criterion. A detailed description of the mRMR method can be found in Peng et al. (2005).

The relevance and redundancy are both measured by the mutual information (MI) defined as

$$I(x,y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \qquad (2)$$

x and y are two random variables. $p(x,y)$ is their joint probability density, and $p(x)$ and $p(y)$ are their marginal probability densities, respectively.

Let F represent the whole feature set, while F_s denotes the already-selected feature set which contains m features, and F_t denotes the yet-to-be-screened feature set which contains n features. Relevance D of the feature f in F_t with the target c can be calculated by

$$D = I(f,c) \qquad (3)$$

The redundancy R of the feature f in F_t with all the features in F_s can be calculated by

$$R = \frac{1}{m}\sum_{\tilde{f}\in F_s} I\left(f,\tilde{f}\right) \qquad (4)$$

To obtain the feature F_t with maximum relevance and minimum redundancy, Eqs. (3) and (4) are combined with the mRMR function

$$\max_{f\in F_t}\left[I(f,c) - \frac{1}{m}\sum_{\tilde{f}\in F_s} I\left(f,\tilde{f}\right)\right] \qquad (5)$$

For a feature set with M features, the feature evaluation will continue M rounds. After these evaluations, we will get a feature set s by the mRMR method

$$s = \{f_1, f_2, f_3 \ldots, f_h, \ldots f_M\} \qquad (6)$$

The feature index h indicates the importance of the respective feature. Better features will be extracted earlier with a smaller index h.

2.5. Evaluation parameter

The performance of various models developed in this study was computed by using threshold-dependent as well as threshold-independent parameters. As threshold-dependent parameters, we used sensitivity, specificity, overall accuracy, and Matthew's correlation coefficient (MCC). These measurements are expressed in terms of true positive (TP), false negative (FN), true negative (TN), and false positive (FP).

2.5.1. Sensitivity

Percentage of correctly predicted ECM proteins within the positive classifications:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad (7)$$

2.5.2. Specificity

Percentage of correctly predicted non-ECM proteins within the negative classifications:

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (8)$$

2.5.3. Accuracy

Percentage of correctly predicted ECM and non-ECM proteins:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \qquad (9)$$

2.5.4. Matthews's Correlation Coefficient (MCC)

It is the statistical parameter to assess the quality of prediction and to take care of the unbalancing in data. Matthew's correlation coefficient ranges from $-1 \le \text{MCC} \le 1$. A value of $\text{MCC}=1$ indicates the best possible prediction while $\text{MCC}=-1$ indicates the worst possible prediction (or anti-correlation). Finally, $\text{MCC}=0$

would be expected for a random prediction scheme.

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \tag{10}$$

2.5.5. Area under the Curve (AUC)

Most of the above measures have the common drawback that their value depends on the selected threshold. The so-called Receiver Operating Curve (ROC) provides a threshold independent measure. The ROC is a plot between sensitivity (TP/TP+FN) and specificity (FP/FP+TN).

3. Results and discussion

3.1. Classification by EcmPred

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical applications: independent dataset test, subsampling (or K-fold crossover) test, and jackknife test (Chou, 2011). However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as demonstrated by Eqs. 28–32 of Chou (2011). Therefore, the jackknife test has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors (Hayat and Khan, 2012; Hayat and Khan, 2011; Harihar and Selvaraj, 2011; Nanni et al., 2012; Esmaeili et al., 2010; Georgiou et al., 2009; Mohabatkar, 2010). However, to reduce the computational time, we adopted the independent test dataset cross-validation in this study, as it is done by many investigators with the SVM as the prediction engine.

We trained our random forest model on the training dataset containing 300 ECM proteins and 300 non-ECM proteins. Our model achieved 82% training accuracy using all the features (68 features). To identify the most prominent features, we carried out feature selection with mRMR. We selected six different feature subsets by decreasing the number of features, and the performance of each feature subset was evaluated. Using 40 features, we obtained 83% training accuracy, which is comparable to the accuracy obtained using 68 features. A similar performance was observed using 10, 20, 30, 50 and 60 features.

In order to examine the performance of the newly developed model, we tested our training model on a test dataset containing 145 ECM proteins and 3887 non-ECM proteins. As shown in Table 2, we obtained 75% accuracy using all the features with a sensitivity of 63%, a specificity of 76%, and a MCC of 0.1702. Using 40 features, our model obtained 77% accuracy with 65% sensitivity, 77% specificity, and a MCC of 0.1906. The performance values were shown for single runs. This result suggests that our feature reduction approach selected useful features by eliminating correlated and noisy features. The list of 40 features is available at http://www.inb.uni-luebeck.de/tools-demos/Extracellular_matrix_proteins/EcmPred.

We also investigated the influence of the feature reduction by plotting Receiver Operating Characteristic (ROC) curves (Fig. 1) derived from the sensitivity (true positive rate) and specificity (false positive rate) values for the classifiers using all the features and the top 40 features, respectively. The area under curve for all features was 0.76 and for the top 40 features was 0.79, respectively.

3.2. Prediction result for known ECM proteins

We collected 20 experimentally verified extracellular matrix proteins from human. Criteria for selection were clear experimental evidence within the literature for the given sequence entry. We tested the efficiency of EcmPred and ECMPP (Jung et al., 2010) using these 20 proteins (Table 3). As shown in Table 3, EcmPred (top 40 features) correctly predicts 15 proteins as extracellular matrix proteins, whereas ECMPP predicts only 6 proteins.

3.3. Screening for ECM in human proteome

To identify novel candidates in the human proteome as extracellular matrix proteins, we scanned the human proteome using SPRED (prediction of secretory proteins) (Kandaswamy et al., 2010) and EcmPred (Fig. 2). With SPRED, we classified these 65,508 protein sequences into 44,611 non-secreted proteins and 20,897 proteins located outside of the nucleo-cytoplasm. We predicted extracellular matrix proteins (6450) using EcmPred, leaving 14,447 proteins which do not belong to the class of extracellular matrix proteins. Subsequently, we removed putative proteins, isoform sequences, hypothetical proteins, fragmented proteins and false positives. The remaining 2201 protein

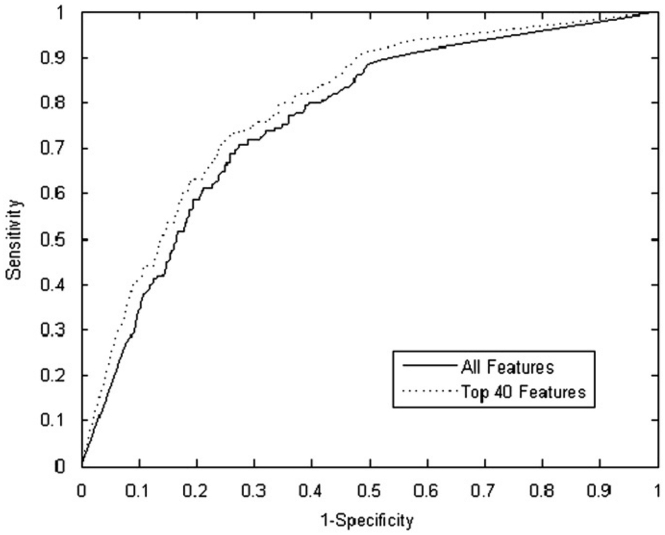

Fig. 1. ROC plot for Random Forest with all and the top 40 features.

Table 2
Performance of Random Forest using different feature subsets. Value inside the square brackets shows standard error of the mean from multiple runs.

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Test accuracy (%)	Training accuracy (%)
10	51 [0.62]	75 [0.31]	0.1123 [0.002]	74 [0.33]	73 [0.67]
20	48 [0.57]	77 [0.38]	0.1171 [0.004]	76 [0.30]	80 [0.60]
30	53 [0.51]	78 [0.30]	0.1378 [0.004]	77 [0.35]	81 [0.63]
40	**65 [0.54]**	**77 [0.34]**	**0.1906 [0.003]**	**77 [0.31]**	**83 [0.62]**
50	57 [0.51]	77 [0.33]	0.1493 [0.004]	76 [0.33]	82 [0.66]
60	60 [0.52]	77 [0.33]	0.1661 [0.003]	76 [0.32]	83 [0.62]
All features	63 [0.55]	76 [0.34]	0.1702 [0.003]	75 [0.33]	82 [0.64]

sequences were classified as extracellular matrix proteins. We investigated the top listed putative ECM proteins using InterPro (Hunter et al., 2009) and Gene ontology (GO) (Gene Ontology Consortium, 2010). Interpro annotation shows Collagen type XXI Alpha 1 and Adamts-like protein 2 as putative extracellular matrix proteins. Collagen, type V, alpha 1, Interphotoreceptor matrix proteoglycan 1, Protein Wnt, Galectin-1, and Galectin-7 were annotated with the Gene Ontology term "extracellular matrix." Thus, as could be expected by the composition of our training set we identified both, proteins forming the ECM network and more mobile proteins interacting transiently with the network. The complete list of predicted ECM proteins is provided at http://www.inb.uni-luebeck.de/tools-demos/Extracellular_matrix_proteins/EcmPred.

Table 3
Prediction result for 20 experimentally verified extracellular matrix proteins using EcmPred and ECMPP. "+" represents proteins correctly predicted as extracellular matrix proteins and "−" represents proteins not predicted as extracellular matrix proteins.

SwissProt ID	Protein annotation	ECMPRED	ECMPP
Q9BY76	Angiopoietin-related protein	+	–
P07355	Annexin A2	+	–
Q9BXN1	Asporin	+	+
P01137	Transforming growth factor beta-1	–	–
Q8N6G6	ADAMTS-like protein 1	+	–
P27797	Calreticulin	+	–
Q76M96	Coiled-coil domain-containing protein	+	+
Q07654	Trefoil factor 3	–	+
O75339	Cartilage intermediate layer protein 1	+	-
Q15063	Periostin	–	–
O43405	Cochlin	+	–
Q96P44	Collagen alpha-1(XXI) chain	+	+
P01009	Alpha-1-antitrypsin	–	–
Q14118	Dystroglycan	+	–
Q12805	EGF-containing fibulin-like extracellular matrix protein 1	+	–
Q75N90	Fibrillin-3	+	+
P09382	Galectin-1	+	+
Q8N2S1	Latent-transforming growth factor beta-binding protein 4	+	–
P27487	Dipeptidyl peptidase 4	–	–
P08253	72 kDa type IV collagenase	+	–

3.4. Comparison of EcmPred with other machine learning methods

The proposed EcmPred method was compared with several state-of-the-art classifiers such as J4.8, Support Vector Machine (SVM), Bayesnet, Logistic Regression, Decision Table, Multi-Layer-Perceptron and Adaboost (Bishop, 1995; Vapnik, 1998; Kohavi, 1995; Quinlan, 1993; Sumner et al., 2005). The data mining software WEKA is used to evaluate the performance of each classifier (Frank et al., 2004). The results based on 40 features are shown in Table 4. All models were tested on the test dataset containing 145 positive and 3887 negative sequences. The prediction accuracy of Random Forest is about 22% and 12% higher than Decision Table and Logistic Regression classifiers, respectively. The Specificity of the SVM is about 9% less than Random Forest. Although the performance of EcmPred and Bayesnet is comparable, the sensitivity is 8% less than with our model.

4. Conclusion

The extracellular matrix is the non-cellular component present within all tissues and organs. It provides physical scaffolding for the cellular constituents and initiates critical biochemical and biomechanical signals required for tissue morphogenesis, differentiation, and homeostasis. The extracellular matrix proteins

Table 4
Comparison of EcmPred with other machine learning methods.

Method	Sensitivity (%)	Specificity (%)	MCC	Test accuracy (%)
J4.8	57	66	0.0973	66
Bayesnet	57	76	0.1485	75
Adaboost	59	69	0.1107	59
Decision table	54	68	0.0893	55
Logistic	59	65	0.0971	65
SVM (polynomial)	56	68	0.1001	68
MLP	58	68	0.1039	59
EcmPred	**65**	**77**	**0.1906**	**77**

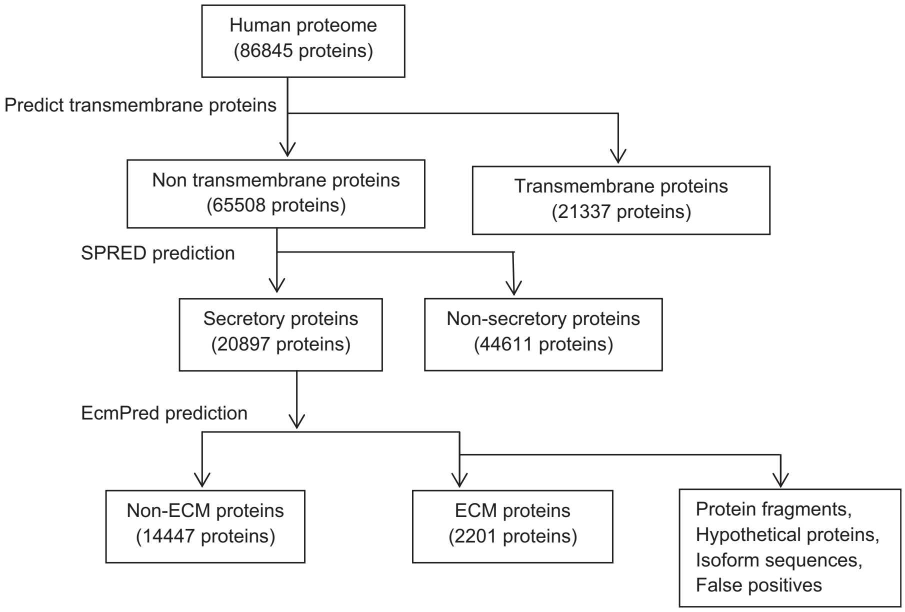

Fig. 2. Screening for ECM proteins in the human proteome.

promise great possibilities as therapeutic targets or diagnostic markers. Identification of ECM proteins is an essential and also difficult task. We implemented a Random Forest approach to predict ECM proteins based on sequence derived properties. High prediction accuracies on the training and test datasets show that EcmPred is a potentially useful tool for the prediction of extracellular matrix proteins from protein primary sequence. EcmPred performed better than ECMPP on experimentally verified ECM proteins. The identification of ECM proteins should be helpful for the analysis of ECM-related functions and diseases. Although our method performs better than the other methods, the prediction accuracy still can be improved by incorporating structural features.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors (Shen and Chou, 2009), we shall make efforts in our future work to provide a web-server for the method presented in this paper. The EcmPred program and dataset is available at http://www.inb.uni-luebeck.de/tools-demos/Extracellular_matrix_proteins/EcmPred.

Acknowledgments

This work was supported by the Graduate School for Computing in Medicine and Life Sciences funded by Germany's Excellence Initiative [DFG GSC 235/1]. KKK acknowledges Dr. Bianca Habermann, Max Planck Institute for Biology of Ageing, Germany for her support.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2012.10.015.

References

Aszodi, A., Legate, K.R., Nakchbandi, I., Fässler, R., 2006. What mouse mutants teach us about extracellular matrix function. Annu. Rev. Cell Dev. Biol. 22, 591–621.

Bateman, J.F., Boot-Handford, R.P., Lamande, S.R., 2009. Genetic diseases of connective tissues: cellular and extracellular effects of ECM mutations. Nat. Rev. Genet. 10, 173–183.

Bendtsen, J.D., Nielsen, H., von Heijne, G., Brunak, S., 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340, 783–795.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.

Bruckner-Tuderman, L., Bruckner, P., 1998. Genetic diseases of the extracellular matrix: more than just connective tissue disorders. J Mol. Med. (Berlin, Germany) 76, 226–237.

Burridge, K., Chrzanowska-Wodnicka, M., 1996. Focal adhesions, contractility, and signaling. Annu. Rev. Cell Dev. Biol. 12, 463–518.

Campbell, N.E., Kellenberger, L., Greenaway, J., Moorehead, R.A., Linnerth-Petrik, N.M., Petrik, J., 2010. Extracellular matrix proteins and tumor angiogenesis. J. Oncol., 586905.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct. Funct. Genet. (Erratum: ibid, 2001, vol. 44, 60) 43, 246–255.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J. Theor. Biol. 273, 236–247.

Chou, K.C., Cai, Y.D., 2005. Prediction of membrane protein types by incorporating amphipathic effects. J. Chem. Inf. Model. 45, 407–413.

Chou, K.C., Shen, H.B., 2006a. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. 347, 150–157.

Chou, K.C., Shen, H.B., 2006b. Large-scale plant protein subcellular location prediction. J. Cell. Biochem. 100, 665–678.

Chou, K.C., Shen, H.B., 2006c. Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85, 233–240.

Chou, K.C., Shen, H.B., 2006d. Large-scale predictions of Gram-negative bacterial protein subcellular locations. J. Proteome Res. 5, 3420–3428.

Chou, K.C., Shen, H.B., 2007a. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.

Chou, K.C., Shen, H.B., 2007b. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J. Proteome Res. 6, 1728–1734.

Chou, K.C., Shen, H.B., 2010a. Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Sci. 2, 1090–1103.

Chou, K.C., Shen, H.B., 2010b. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS One 5, e11335.

Chou, K.C., Wu, Z.C., Xiao, X., 2011. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6, e18258.

Chou, K.C., Wu, Z.C., Xiao, X., 2012. iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641.

Dudoit, S., Fridlyand, J., Speed, T.P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc. 97, 77–87.

Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J. Theor. Biol. 263, 203–209.

Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I.H., 2004. Data mining in bioinformatics using Weka. Bioinformatics 20, 2479–2481.

Gene Ontology Consortium, 2010. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res. 38, D331–D335. (Database issue).

Georgiou, D.N., Karakasidis, T.E., Nieto, J.J., Torres, A., 2009. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J. Theor. Biol. 257, 17–26.

Green, K.A., Lund, L.R., 2005. ECM degrading proteases and tissue remodelling in the mammary gland. Bioessays 27, 894–903.

Grønborg, M., Kristiansen, T.Z., Iwahori, A., Chang, R., Reddy, R., Sato, N., Molina, H., Jensen, O.N., Hruban, R.H., Goggins, M.G., Maitra, A., Pandey, A., 2006. Biomarker discovery from pancreatic cancer secretome using a differential proteomic approach. Mol. Cell Proteomics 5, 157–171.

Harihar, B., Selvaraj, S., 2011. Analysis of rate-limiting long-range contacts in the folding rate of three-state and two-state proteins. Protein Pept. Lett. 18, 1042–1052.

Hayat, M., Khan, A., 2011. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. J. Theor. Biol. 271, 10–17.

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

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844.

Ho, T.K., 2002. A data complexity analysis of comparative advantages of decision forest constructors. Pattern Anal. Appl. 5, 102–112.

Ho, T.K., Hull, J.J., Srihari, S.N., 1994. Decision combination in multiple classifier systems. IEEE Trans. Pattern Anal. Mach. Intell. 16, 66–75.

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., Nakai WoLF, K., 2007. PSORT: protein localization predictor. Nucleic Acids Res. 35, W585–W587.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H., Yeats, C., 2009. InterPro: the integrative protein signature database. Nucleic Acids Res. 37, D224–D228 (Database Issue).

Jia, S.C., Hu, X.Z., 2011. Using Random forest algorithm to predict beta-hairpin motifs. Protein Pept. Lett. 18, 609–617.

Jung, J., Ryu, T., Hwang, Y., Lee, E., Lee, D., 2010. Prediction of extracellular matrix proteins based on distinctive sequence and domain characteristics. J. Comput. Biol. 17, 97–105.

Kandaswamy, K.K., Pugalenthi, G., Hartmann, E., Kalies, K.U., Möller, S., Suganthan, P.N., Martinetz, T., 2010. SPRED: a machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. Biochem. Biophys. Res. Commun. 391, 1306–1311.

Kandaswamy, K.K., Chou, K.C., Martinetz, T., Möller, S., Suganthan, P.N., Sridharan, S., Pugalenthi, G., 2011. AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J. Theor. Biol. 270, 56–62.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M., 2008. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36, D202–D205.

Kersey, P.J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., Apweiler, R., 2004. The International Protein Index: an integrated database for proteomics experiments. Proteomics 4, 1985–1988.

Kim, S.H., Turnbull, J.E., Guimond, S.E., 2011. Extracellular matrix and cell signalling—the dynamic cooperation of integrin, proteoglycan and growth factor receptor. J. Endocrinol. 209, 139–151.

Kohavi, R., 1995. The Power of Decision Tables. In: Proceedings of the 8th European Conference on Machine Learning. pp. 174–189.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580.

Kumar, K.K., Pugalenthi, G., Suganthan, P.N., 2009. DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest. J. Biomol. Struct. Dyn. 26, 679–686.

Lee, J.W., Lee, J.B., Park, M., Song, S.M., 2005. An extensive comparison of recent classification tools applied to microarray data. Comput. Stat. Data Anal. 48, 869–885.

Lewin, B., Cassimeris, L., Lingappa, V., Plopper, G. (Eds.), 2007. Cells. Jones and Bartlett, Sudbury, MA.

Li, W., Jaroszewski, L., Godzik, A., 2001. Clustering of highly homologous sequences to reduce the size of large protein database. Bioinformatics 17, 282–283.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.

Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2011. iDNA-Prot: identification of DNA binding proteins using Random Forest with Grey Model. PLoS One 6, e24756.

Mohabatkar, H., 2010. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept. Lett. 17, 1207–1214.

Nanni, L., Lumini, A., Gupta, D., Garg, A., 2012. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans. Comput. Biol. Bioinform. 9, 467–475.

Nelson, C.M., Bissell, M.J., 2006. Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer. Annu. Rev. Cell Dev. Biol. 22, 287–309.

Peng, H.C., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27, 1226–1238.

Pugalenthi, G., Kandaswamy, K.K., Chou, K.C., Vivekanandan, S., Kolatkar, P., 2012. RSARF: prediction of residue solvent accessibility from protein sequence using Random Forest method. Protein Pept. Lett. 19, 50–56.

Qiu, Z., Wang, X., 2011. Improved prediction of protein ligand-binding sites using Random Forests. Protein Pept. Lett. 18, 1212–1218.

Quinlan, R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Schwartz, M.A., Schaller, M.D., Ginsberg, M.H., 1995. Integrins: emerging paradigms of signal transduction. Annu. Rev. Cell Dev. Biol. 11, 549–599.

Shameer, K., Pugalenthi, G., Kandaswamy, K.K., Sowdhamini, R., 2011. 3dswap-pred: prediction of 3D domain swapping from protein sequence using random forest approach. Protein Pept. Lett. 18, 1010–1020.

Shen, H.B., Chou, K.C., 2006. Ensemble classifier for protein fold pattern recognition. Bioinformatics 22, 1717–1722.

Shen, H.B., Chou, K.C., 2009. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. Anal. Biochem. 394, 269–274.

Sorokin, L., 2010. The impact of the extracellular matrix on inflammation. Nat. Rev. Immunol. 10, 712–723.

Sumner, M., Frank, E., Hall, M., 2005. Speeding up Logistic Model Tree Induction. In: Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases. pp. 675–683.

Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience, New York.

Walter, P., Gilmore, R., Blobel, G., 1984. Protein translocation across the endoplasmic reticulum. Cell 38, 5–8.

Wary, K.K., Mainiero, F., Isakoff, S.J., Marcantonio, E.E., Giancotti, F.G., 1996. The adaptor protein Shc couples a class of integrins to the control of cell cycle progression. Cell 87, 733–743.

Wu, Z.C., Xiao, X., Chou, K.C., 2011. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Mol. BioSyst. 7, 3287–3297.

Wu, Z.C., Xiao, X., Chou, K.C., 2012. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. Protein Pept. Lett. 19, 4–14.

Xiao, X., Wu, Z.C., Chou, K.C., 2011. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J. Theor. Biol. 284, 42–51.