JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 1, 2010

Pp. 97–105
DOI: 10.1089/cmb.2008.0236

Prediction of Extracellular Matrix Proteins Based on Distinctive Sequence and Domain Characteristics

*JUHYUN JUNG,[1] *TAEWOO RYU,[2] YONGDEUK HWANG,[1]
EUNJUNG LEE,[1] AND DOHEON LEE[1]

ABSTRACT

Extracellular matrix (ECM) proteins are secreted to the exterior of the cell, and function as mediators between resident cells and the external environment. These proteins not only support cellular structure but also participate in diverse processes, including growth, hormonal response, homeostasis, and disease progression. Despite their importance, current knowledge of the number and functions of ECM proteins is limited. Here, we propose a computational method to predict ECM proteins. Specific features, such as ECM domain score and repetitive residues, were utilized for prediction. Based on previously employed and newly generated features, discriminatory characteristics for ECM protein categorization were determined, which significantly improved the performance of Random Forest and support vector machine (SVM) classification. We additionally predicted novel ECM proteins from non-annotated human proteins, validated with gene ontology and earlier literature. Our novel prediction method is available at http://biosoft.kaist.ac.kr/ecm.

Key words: ECM, extracellular matrix proteins, protein localization, Random Forest, support vector machine.

1. INTRODUCTION

PROTEIN LOCALIZATION provides a crucial indication of biological function. For instance, transcription factors are largely localized in the nucleus, and proteins related to respiration are present in mitochondria. On the other hand, secreted proteins are exported to the exterior of the cell membrane, and play a role in communication with neighboring cells (Jacobs et al., 2008).

Extracellular matrix (ECM) proteins constitute a class of secreted proteins, and assemble as a massive network on the cell surface. These proteins support the cell surface microenvironment, and additionally influence critical cell behavior, such as proliferation, survival, and differentiation. Consequently, dysregulation of ECM proteins may cause diseases, including developmental abnormalities and cancer (Pupa et al., 2002). Recent studies report the presence of some of these proteins in body fluid, suggesting they may have practical applications as disease-specific markers (Gronborg et al., 2006). For example, KLK6, one of the secreted serine proteases, is a suggested serum marker for the diagnosis of early-stage ovarian cancer

[1]Department of Bio and Brain Engineering, KAIST, Daejeon, Korea.
[2]Bioinformatics Research Center, KRIBB, Daejeon, Korea.
*These two authors contributed equally to this work.

(Diamandis et al., 2000). Therefore, identification of ECM proteins is a significant step in understanding cancer progression and providing effective therapeutic targets or diagnostic markers.

A large proportion of eukaryotic protein localization has not been annotated experimentally (Nair and Rost, 2005). Because experimental verification is labor-intensive and time-consuming, several computational programs have been developed to predict protein localization (Chou, 2000; Lee et al., 2006; Xie et al., 2005). Most prediction methods have solved this problem in one of two ways. One procedure is based on specific organelle-targeting sequences, while the other utilizes statistical differences in general sequence characteristics, such as amino acid composition. For secreted proteins, sorting signal peptides on the N-termini are incorporated for classification and prediction in SignalP (Bendtsen et al., 2004b) and WoLF PSORT (Nakai and Horton, 1999). Klee and Sosa (2007) gave a comprehensive review on multiple techniques for protein subcellular localization with more detail. Despite the successful development of localization-prediction programs, no method has been developed specifically for predicting ECM proteins, which have distinct features from other secreted proteins.

Here, we propose a computational method to distinguish ECM proteins among the secreted protein groups. Thirteen distinctive features, including characteristic repeats, functional domains, and evolutionary amino acid composition, were identified as discriminatory for ECM proteins, which significantly improved the performance of classification in conjunction with Random Forest (Leo, 2001) and support vector machine (SVM) (Xie et al., 2005). Furthermore, 20 candidate ECM proteins were predicted from non-annotated human proteins using these distinctive features.

2. METHODS

2.1. Data set

Metazoan secreted protein information were obtained from Swiss-Prot release 51 (Boeckmann et al., 2003). To increase the confidence of the classification results, proteins with experimentally confirmed annotations were used. To avoid data redundancy, proteins displaying over 40% sequence similarity were additionally excluded using Cd-hit (Li and Godzik, 2006). Consequently, 109 ECM and 1,430 secreted but not localized in ECM (hereinafter referred to simply as ''non-ECM'') proteins were obtained.

2.2. Feature generation

We generated five novel features specific to ECM proteins in addition to 91 features which were conventionally used by localization-prediction methods (all features are listed in Table S1; see online Supplementary Material at www.liebertonline.com). To describe characteristics of ECM proteins, we initially defined the repetitive residue (F2 in Table 1). The number of repeat patterns ($N_{i,j}$) was counted for each gap size (i), i.e., the number of amino acids between any two repeats, and amino acid (j) in the protein sequence. The most frequently occurring repeat patterns for each gap size were selected.

$$M_i = \max\{N_{i,0}, N_{i,1}, \ldots N_{i,20}\}$$

All M_i in the sequence were summed:

$$S_{sum} = \sum_{i=0}^{L} M_i$$

where L represents the sequence length. Because longer sequences presumably have more repeats, S_{sum} was normalized:

$$S_{final} = S_{sum}/L'$$

The scaled sequence length (L') was calculated as follows:

$$L' = L/\sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}}$$

where x_j, $j = 1 \ldots n$ is the sequence length, and n represents the number of sequences in the dataset.

Table 1. Distinctive Features of ECM Proteins

	Feature name	*MDA*	μ_{ecm}	$\mu_{non\text{-}ecm}$	*t score*
F1	ECM domain score	0.974	0.62	0.00	8.759
F2	**Repetitive residue**	0.952	294.33	212.07	4.480
F3	**Molecular weight**	0.760	113,470.86	35,069.44	7.610
F4	**Sequence length**	0.705	1,033.39	316.03	7.610
F5	**Repeated domains**	0.692	5.46	0.50	6.145
F6	Tyr composition	0.450	0.03	0.03	0.877
F7	Aliphatic residue composition	0.361	0.18	0.20	4.726
F8	Leu evolutionary composition	0.352	0.08	0.09	2.780
F9	Cys evolutionary composition	0.347	0.04	0.03	0.952
F10	**Glycine-x-y repeats**	0.279	12.34	2.37	4.774
F11	Cys composition	0.273	0.04	0.04	0.738
F12	Gln evolutionary composition	0.268	0.05	0.04	2.595
F13	Arg evolutionary composition	0.246	0.05	0.05	0.556

Novel features distinguishing ECM proteins are emphasized in bold.

ECM, extracellular matrix; mean decrease accuracy (MDA), the discriminatory power of each feature provided by Random Forest; μ_{ecm}, mean value of ECM proteins; $\mu_{non\text{-}ecm}$, mean value of non-ECM proteins; *t* score, discriminatory power of features between ECM and non-ECM proteins evaluated using the Student's *t*-test.

Next, we employed the molecular weight (F3 in Table 1) and sequence length (F4) from the Swiss-Prot database. The numbers of domain duplication (F5) and Gly-x-y repeat (F10) occurrences were also counted.

2.3. Feature selection and classification

We utilized Random Forest, an ensemble classifier that generates multiple decision trees by bootstrapping samples with replacement, and aggregates the results (Leo, 2001). Random Forest starts by sampling *n* proteins with replacement from all training samples, and also randomly selects a subset of features among the 96 features. Given a subset of selected samples and features, the best splitting feature values are selected for each node of a decision tree. Each decision tree grows as large as possible in this way, and the best model is obtained from the average of trees in the forest.

During bootstrap training, about one-third of the samples are left out, designated "out-of-bag" (OOB). Because these samples are applied only in the test step, an unbiased error rate is achieved by running each OOB sample down all trees. To measure the feature importance, the mean decrease accuracy (MDA) was calculated as follows: first, the number of correctly classified samples was calculated for each tree using the original values of the k^{th} feature (n1), and also using randomly permuted values of the k^{th} feature (n2). Next, the differences between the values obtained (n1-n2) were averaged across all trees, and normalized by assuming the normal distribution. Then, the OOB error rate was estimated by adding features in decreasing order of feature importance. Thirteen features were found to be sufficient for the OOB error rate to converge to its minimum value (Fig. S1; see online Supplementary Material at www.liebertonline.com).

3. RESULTS

The overall workflow of the proposed method is depicted in Figure 1. Training data were curated from Swiss-Prot version 51 (Boeckmann et al., 2003). In total, 96 features, including five novel characteristics, were generated to represent ECM protein-specific characteristics (Table S1; see Supplementary Material at www.liebertonline.com). Feature values were calculated in two ways: (1) amino acid composition and repeat residues were directly obtained from sequences; and (2) meta-information, including evolutionary composition and domain-related features were obtained using other programs (i.e., Pfam/HMMER and PSI-BLAST). The top 13 features, which displayed the minimum error rate, were selected according to the MDA, a feature importance value from the Random Forest algorithm (Fig. S1; see online Supplementary Material at www.liebertonline.com). Classification performance was significantly improved with the aid of these distinctive features and two machine learning techniques (Random Forest and SVM). Finally, novel ECM proteins were predicted from non-annotated proteins using our method.

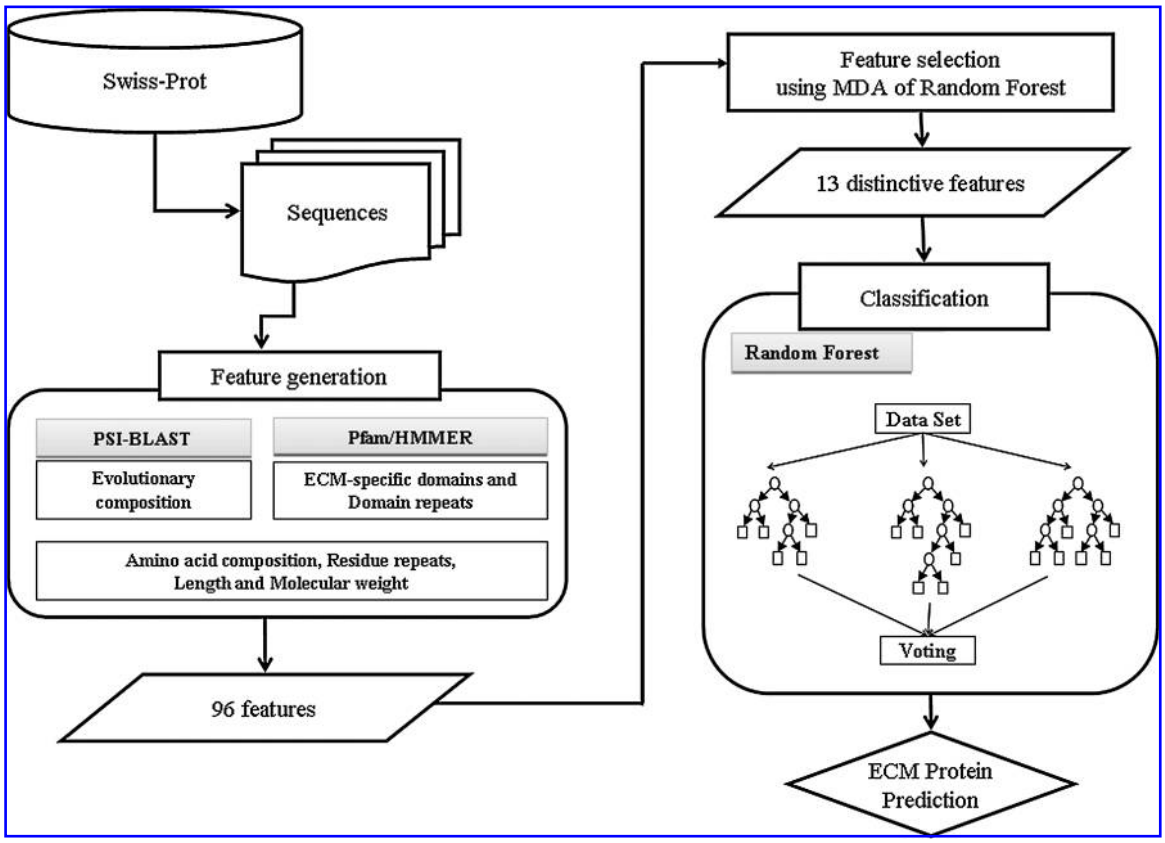

FIG. 1. A schematic diagram of the proposed method. Various sequence and domain features were generated from metazoan protein sequences downloaded from Swiss-Prot release 51. Among them, the top 13 discriminatory features were selected by the mean decrease accuracy of Random Forest.

3.1. Distinctive features of ECM proteins

Initially, 96 features were generated. However, use of a large number of features often triggers practical problems in machine learning, such as the curse of dimensionality and prevention of model interpretation. To reduce noisy and irrelevant features, we used MDA provided by Random Forest, which reflects the influence of a specific feature on the classification error rate (Leo, 2001). Features were ranked based on MDA scores, and added sequentially to the feature set starting from the feature with the highest score. The overall error rate was calculated at each iteration. We found that use of the top 13 features gave the best classification performance (Fig. S1; see online Supplementary Material at www.liebertonline.com). These features, including five ECM protein-specific characteristics (F2, F3, F4, F5, and F10) and MDA scores, are presented in Table 1.

The most powerful feature in ECM protein classification was the ECM domain score (F1 in Table 1). It was calculated by counting the occurrence of 34 ECM-specific domains that were identified to occur only in ECM proteins using Pfam-HMMer (Finn et al. 2006) (Table S2; see online Supplementary Material at www.liebertonline.com). About half of the known ECM proteins (53/109) in Swiss-Prot contained at least one specific domain, as reflected by a high MDA score.

Repeat patterns (i.e., F2, F5, and F10) were other discriminatory features. ECM proteins have been known to contain several residue repeats, such as coiled-coil structures (Hohenester and Engel 2002). The feature "Repetitive residue," (F2) representing the occurrence of residues at regular intervals, captured this characteristic. In addition, ECM proteins often have repeat domains (F5) replicated from a common ancestor domain, which alter their structure and function independently (Hohenester and Engel 2002). Collagenous proteins constituting the largest family of ECM proteins display a particular type of repeat pattern, i.e., Glycine-x-y repeats (Hohenester and Engel, 2002), represented by F10. The mean values of these repeats in ECM proteins were significantly greater than those in non-ECM proteins, as reflected by high t scores (Table 1). Figure 2 shows examples of these repeat patterns in the protein sequences.

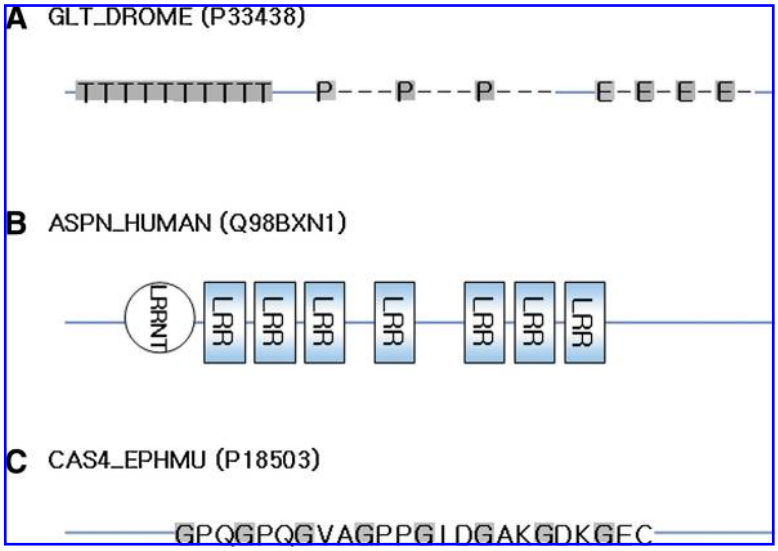

FIG. 2. Examples of (**A**) repetitive residues, (**B**) domain repeats and (**C**) glycine-x-y repeats. (A) Fruitfly Glutactin, a basement membrane-related glycoprotein, contains various types of repetitive residues. (B) ASPN associated with cartilage matrix contains seven replicate leucine-rich repeat (LRR) domains. (C) Glycines are frequently repeated with two gaps in collagenous proteins.

ECM proteins usually have high molecular weights and large number of amino acids as shown in Table 1. For instance, LAMB1, an ECM glycoprotein, contains 1,786 amino acids with a molecular weight of about 198 kilodaltons. The molecular weights (F3) and sequence lengths (F4) of ECM proteins were more than three times higher than those of non-ECM proteins on average, thus leading to high MDA and *t* scores.

Amino acid composition is strongly correlated with localization, possibly because the protein surface should contain appropriate residues for the local environment (Andrade et al., 1998). The amino acid composition was obtained by counting the numbers of each residue and dividing by the protein length. Because evolutionary information is effective in localization prediction (Xie et al., 2005), the evolutionary composition was also adopted in our study. The position-specific scoring matrix (PSSM) of a protein was obtained by running PSI-BLAST, and all values for each amino acid in the matrix were summed and divided by the sequence length. The evolutionary composition calculated in this way was highly ranked, together with amino acid and categorical composition (F6–9 and F11–13).

3.2. Classification using Random Forest and support vector machine

The ultimate goal of this study was to predict novel ECM proteins. To achieve this, we built classifiers and validated their accuracies using Random Forest, which has several advantages for classification. The program runs efficiently on a large dataset, provides importance values of features, supports unbiased error estimates, and balances the prediction error when the class sizes are different. We utilized Random Forest as a default classifier with three different feature sets: (1) 91 previous features, (2) 91 previous plus five novel features, and (3) top 13 distinctive features. Fivefold cross-validation was conducted, and receiver operating characteristic (ROC) curves were drawn to evaluate the classification performance over the entire range of specificity and sensitivity values (Fig. 3A). Precision, accuracy, and Matthew's correlation coefficient (MCC) (Matthews, 1975) were additionally measured (Table S3; see online Supplementary Material at www.liebertonline.com). The 13 selected features improved classification performance over other feature sets, as shown in Figure 3A and Table S3 (see online Supplementary Material at www.liebertonline.com). However, it was questionable whether the improved performance relied on the specific machine learning, i.e., Random Forest. Accordingly, SVM, a widely used classification method in many bioinformatics fields (Lo et al., 2005), was additionally applied to the same dataset and feature groups. The reduced feature set repeatedly led to the best performance in SVM classification. However, the performance of the 2nd feature set was not comparable to the best performance of the 3rd set, but was slightly better than that of the 1st set (Fig. 3B). This discrepancy between Random Forest and SVM may arise because Random Forest selected features during training with the 2nd set, similar to the process of generating the 3rd feature set.

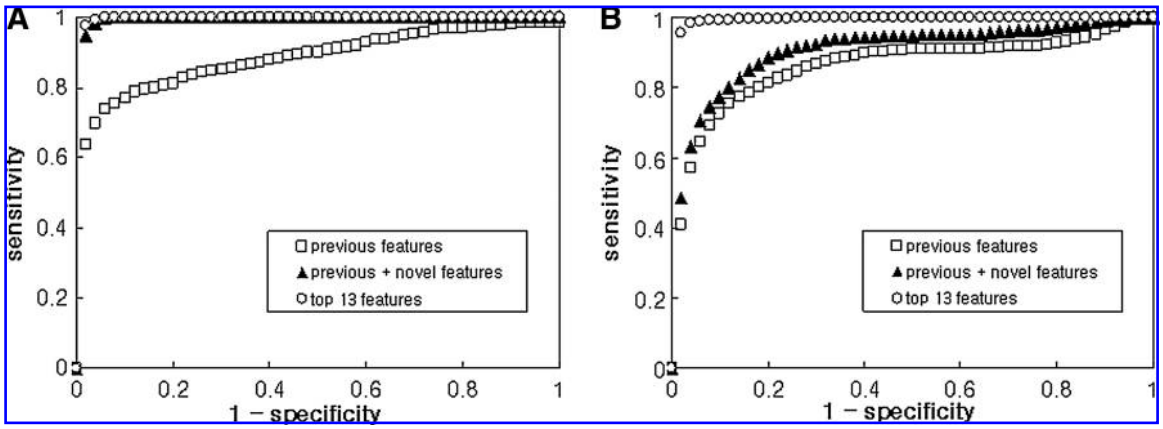

FIG. 3. ROC curve of (**A**) Random Forest and (**B**) SVM built with conventional 91 features, conventional 91 features plus 5 novel features, and the top 13 discriminatory features selected among 96 features. The top 13 features displayed the best performance in both classification methods.

3.3. Prediction of candidate ECM proteins

We applied the proposed method to predict novel ECM proteins in humans. Firstly, 4,163 non-annotated human proteins were retrieved from Swiss-Prot. Secreted protein candidates were then sought for the input set of our method because our method was specialized to identify ECM proteins among secreted groups. We employed SignalP, which ensured high predictive accuracy, to filter out non-secretory proteins under the assumption that most ECM proteins contain signal peptides in order to be secreted (Bendtsen et al., 2004b). Even though the secretory pathway without an N-terminal signal peptide has been reported (Bendtsen et al., 2004a), we did not consider it in this study since the number of proteins entering the pathway is still very limited and its mechanism is not yet clear. From the initial set, 454 proteins were predicted to contain at least one extracellular secretory signal peptide, and hence were regarded as candidate secreted proteins. Using Random Forest and 13 selected features, 20 proteins with a prediction score of 0.6 (an arbitrary cut-off) or higher were classified as putative ECM proteins, and their functions investigated using gene ontology (GO) (Gene Ontology Project, 2008) (Table 2). By the continuous update of GO, half of the predicted candidates were annotated with the component term ''extracellular region.'' This high ratio confirms the predictive power of the proposed method.

While no ontology evidence was obtained for the remaining 10 proteins, a literature survey provided further information. For example, P98066 (official symbol: TNFAIP6) was reported to be a tumor necrosis factor—alpha (TNFα)–induced secretory protein involved in ECM stability and cell migration (Kohda et al. 1996), and O43261 (DLEU1) was used as a surface marker in a leukemia study (Moretta et al. 1987). P01619 and P04434 are immunoglobulin kappa chain proteins with unknown localization. Because antibodies are secreted into the body fluid or anchored on the surface of immune cells, it is postulated that these proteins are localized in the ECM. Other immunoglobulin proteins specified in Table 2 (P01601, P01764, P04433, P01602, and P04211) were also annotated with the term ''extracellular region,'' supporting this possibility.

Highly ranked features in Table 1 generally dominated the prediction results (Tables 2 and S4; see online Supplementary Material at www.liebertonline.com). For example, proteins with ECM-specific domains (F1), large molecular weights (F3), and sequence lengths (F4) displayed high prediction scores. On the other hand, less important features did not influence the score to a significant extent. While limited information is available on the localization and biological roles of these candidates, the high feature values support the possibility that they are true ECM proteins.

4. DISCUSSION

ECM maintains cell shape and controls the communication with the environment. While most cellular proteins are concealed by the lipid bilayer membrane, ECM proteins are displayed on the surface, and often

Table 2. Candidate ECM Proteins Predicted Using Our Method

Swiss-Prot accession number	*Official symbol*	*ECM score*[a]	*Short description (Gene Ontology: Cellular Component)*
P78539	SRPX	0.92	Cell surface and membrane
P09486	SPARC	0.918	Basement membrane, extracellular region
P10915	HAPLN1	0.917	Extracellular region
P98066	TNFAIP6	0.91	N/A
P59022	DSCR10	0.731	N/A
O43261	DLEU1	0.72	N/A
O75556	SCGB2A1	0.716	Extracellular region
Q13296	SCGB2A2	0.698	Extracellular region
P01619	N/A	0.698	N/A
P01601	IGKC	0.686	Extracellular region
Q16517	NNAT	0.682	N/A
P01764	IGHV@	0.68	Extracellular region
P04433	IGKV3D-11	0.656	Extracellular region
P58511	C21orf51	0.656	N/A
P01124	N/A	0.655	N/A
P01602	IGKV1-5	0.641	Extracellular region
Q7Z4B0	C18orf20	0.623	N/A
P04434	N/A	0.62	N/A
P04211	N/A	0.613	Extracellular region
Q9BXH5	TTY6	0.604	N/A

[a]Score indicating whether a protein belongs to the ECM protein class.

Human proteins with unknown localization in Swiss-Prot were analyzed with our method. Using Random Forest with the set of 13 selected features, 20 ECM protein candidates were obtained with a cut-off value of 0.6.

ECM, extracellular matrix.

serve as specific markers for cell status or therapeutic targets. However, the number of annotated ECM proteins is too limited at present for further analysis. In this study, we attempted to predict ECM proteins in extracellular space. Firstly, we proposed five novel characteristics of ECM proteins, which proved remarkably useful for the classification process. Secondly, we built a highly accurate classifier for ECM proteins using Random Forest and a reduced feature set. Consequently, 20 candidate human ECM proteins were predicted using our classification engine.

It should be noted that gene ontology and protein-protein interactions, frequently used in earlier classification studies (Chou and Cai, 2005; Lee et al., 2006), were not employed in order to avoid biased learning, since such prior knowledge might be helpful only for well-annotated proteins. As the prediction of localization for unknown proteins is more important, our classification was solely based on sequence-based information, such as domain and sequence length.

One major difficulty in this task involved data imbalance, i.e., only 109 metazoan ECM proteins were annotated, compared to 1,430 non-ECM proteins. This low number and ratio represent an obvious obstacle in classification. Evidently, while Random Forest balances the prediction error by bootstrapping samples in such cases, better prediction will be possible when more ECM proteins are available.

ECM proteins are strongly advocated as effective diagnosis markers and therapeutic targets (Table 2). For instance, O75556 and Q13296 are utilized as potential serum protein markers for breast cancer (Bernstein et al., 2005; Sasaki et al., 2007), while evidence of P78539, P09486, P98066, Q16517, and P01764 as markers or drug targets has been reported (Davi et al., 2008; Garcia et al., 2006; Kim et al., 2003; Krstulja et al., 2008; Uchihara et al., 2007). The identification of ECM proteins should be helpful in the analysis of ECM-related function and disease.

ACKNOWLEDGMENTS

We would like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics for providing research facilities. This work was supported by the National Research Laboratory Program (grant R0A-

2005-000-10094-0) and the Korean Systems Biology Program (grant M10309020000-03B5002-00000) from the Ministry of Education, Science and Technology through the National Research Foundation.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Andrade, M.A., O'Donoghue, S.I., and Rost, B. 1998. Adaptation of protein surfaces to subcellular location. *J. Mol. Biol.* 276, 517–525.

Bendtsen, J.D., Jensen, L.J., Blom, N., et al. 2004a. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* 17, 349–356.

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

Bernstein, J.L., Godbold, J.H., Raptis, G., et al. 2005. Identification of mammaglobin as a novel serum marker for breast cancer. *Clin. Cancer Res.* 11, 6528–6535.

Boeckmann, B., Bairoch, A., Apweiler, R., et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365–370.

Chou, K.C. 2000. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.* 278, 477–483.

Chou, K.C., and Cai, Y.D. 2005. Predicting protein localization in budding yeast. *Bioinformatics* 21, 944–950.

Davi, F., Rosenquist, R., Ghia, P., et al. 2008. Determination of IGHV gene mutational status in chronic lymphocytic leukemia: bioinformatics advances meet clinical needs. *Leukemia* 22, 212–214.

Diamandis, E.P., Yousef, G.M., Soosaipillai, A.R., et al. 2000. Human kallikrein 6 (zyme/protease M/neurosin): a new serum biomarker of ovarian carcinoma. *Clin. Biochem.* 33, 579–583.

Finn, R.D., Mistry, J., Schuster-Bockler, B., et al. 2006. Pfam: clans, web tools and services. *Nucleic Acids Res.* 34, D247–D251.

Garcia, G.E., Wisniewski, H.G., Lucia, M.S., et al. 2006. 2-Methoxyestradiol inhibits prostate tumor development in transgenic adenocarcinoma of mouse prostate: role of tumor necrosis factor-alpha-stimulated gene 6. *Clin. Cancer Res.* 12, 980–988.

Gene Ontology Project. 2008. The Gene Ontology project in 2008. *Nucleic Acids Res.* 36, D440–D444.

Gronborg, M., Kristiansen, T.Z., Iwahori, A., et al. 2006. Biomarker discovery from pancreatic cancer secretome using a differential proteomic approach. *Mol. Cell Proteomics* 5, 157–171.

Hohenester, E., and Engel, J. 2002. Domain structure and organisation in extracellular matrix proteins. *Matrix Biol.* 21, 115–128.

Jacobs, J.M., Waters, K.M., Kathmann, L.E., et al. 2008. The mammary epithelial cell secretome and its regulation by signal transduction pathways. *J. Proteome Res.* 7, 558–569.

Kim, C.J., Shimakage, M., Kushima, R., et al. 2003. Down-regulation of drs mRNA in human prostate carcinomas. *Hum. Pathol.* 34, 654–657.

Klee, E.W., and Sosa, C.P. 2007. Computational classification of classically secreted proteins. *Drug Discov. Today* 12, 234–240.

Kohda, D., Morton, C.J., Parkar, A.A., et al. 1996. Solution structure of the link module: a hyaluronan-binding domain involved in extracellular matrix stability and cell migration. *Cell* 86, 767–775.

Krstulja, M., Car, A., Bonifacic, D., et al. 2008. Nasopharyngeal angiofibroma with intracellular accumulation of SPARC–a hypothesis (SPARC in nasopharyngeal angiofibroma). *Med. Hypoth.* 70, 600–604.

Lee, K., Kim, D.W., Na, D., et al. 2006. PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. *Nucleic Acids Res.* 34, 4655–4666.

Leo, B. 2001. Random Forests. *Mach. Learn.* 45, 5–32.

Li, W., and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659.

Lo, S.L., Cai, C.Z., Chen, Y.Z., et al. 2005. Effect of training datasets on support vector machine prediction of protein-protein interactions. *Proteomics* 5, 876–884.

Matthews, B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451.

Moretta, A., Poggi, A., Olive, D., et al. 1987. Selection and characterization of T-cell variants lacking molecules involved in T-cell activation (T3 T-cell receptor, T44, and T11): analysis of the functional relationship among different pathways of activation. *Proc. Natl. Acad. Sci. U.S.A.* 84, 1654–1658.

Nair, R., and Rost, B. 2005. Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.* 348, 85–100.

Nakai, K., and Horton, P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24, 34–36.

Pupa, S.M., Menard, S., Forti, S., et al. 2002. New insights into the role of extracellular matrix during tumor onset and progression. *J. Cell. Physiol.* 192, 259–267.

Sasaki, E., Tsunoda, N., Hatanaka, Y., et al. 2007. Breast-specific expression of MGB1/mammaglobin: an examination of 480 tumors from various organs and clinicopathological analysis of MGB1-positive breast cancers. *Mod. Pathol.* 20, 208–214.

Uchihara, T., Okubo, C., Tanaka, R., et al. 2007. Neuronatin expression and its clinicopathological significance in pulmonary non-small cell carcinoma. *J. Thorac. Oncol.* 2, 796–801.

Xie, D., Li, A., Wang, M., et al. 2005. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.* 33, W105–W110.

Address correspondence to:
Dr. Doheon Lee
Department of Bio and Brain Engineering
KAIST, 335 Gwahangno, Yuseong-gu
Daejeon, 305-701, Korea

E-mail: dhlee@kaist.ac.kr