Journal of Universal Computer Science, vol. 24, no. 6 (2018), 711-724
submitted: 30/6/17, accepted: 15/12/17, appeared: 28/6/18 © J.UCS

Identifying Cleavage Sites of Gelatinases A and B by Integrating Feature Computing Models

Quan Zou
(College of Optoelectronic Engineering, Shenzhen University
Shenzhen, P.R.China
zouquan@szu.edu.cn)

Chi-Wei Chen
(Department of Computer Science and Engineering, Institute of Genomics and Bioinformatics
National Chung Hsing University, Taichung, Taiwan
d103056006@mail.nchu.edu.tw)

Hao-Chen Chang
(Institute of Genomics and Bioinformatics, National Chung Hsing University
Taichung, Taiwan
caty.e.e@hotmail.com)

Yen-Wei Chu
(Institute of Genomics and Bioinformatics, Biotechnology Center, Agricultural Biotechnology
Center, Institute of Molecular Biology, Graduate Institute of Biotechnology
National Chung Hsing University, Taichung, Taiwan
ywchu@nchu.edu.tw, corresponding author)

Abstract: Gelatinases proteases with the ability to cleave the extracellular matrix (ECM). Two types of gelatinases exist: Gelatinase A, also referred to as matrix metalloproteinase-2 (MMP-2), and gelatinase B, also referred to as matrix metalloproteinase-9 (MMP-9). MMP-2 and MMP-9 degrade ECM, which is highly expressed during tumor metastasis. The poor therapeutic effects of inhibitors can be attributed to the high structural homology shared by members of the matrix metalloproteinase family. The highly similar structures of these proteases preclude the specific binding of inhibitor drugs. Moreover, the regulatory pathways of MMP-2 and MMP-9 remain poorly understood. An accurate model for the prediction of substrates and the cleavage sites of gelatinases should be developed to enable screening and exploring the physiological and pathological mechanisms of these enzymes. Prediction is based on various types of information on binary integration, physical–chemical properties, protein stability, solvent accessibility, and protein secondary structure. In this study, the first level of the prediction model was constructed on the basis of intergroup differences and support vector machine. Predictive probability was then taken as the characteristic of the second level of the prediction model, which was constructed using different machine-learning methods. The Mathews correlation coefficients of the MMP-2 and MMP-9 prediction models were 89.4% and 64.4%, respectively. The physical–chemical properties of the active sites of MMP-2 and MMP-4 were selected for analysis. The completion of this prediction system will aid the discovery of regulatory paths and novel applications of MMP-2 and MMP-9, as well as provide references for drug design.

Keywords: Gelatinase, MMP-2, MMP-9, Machine Learning, Support Vector Machine
Categories: I.2.1, I. 2.4, I.2.6, I.2.8, J.3

1 Introduction

After undergoing translation, proteins modify and regulate various cellular metabolic processes. Modification through translation is reversible. By contrast, hydrolysis, in which the peptide bond that connects an amino acid with an amide is cleaved by a protease, is irreversible. Proteases have extensive and important influences on various proteins. Approximately 2% of the human genome encode for proteases, and approximately 5%–10% of proteases have known drug targets [Puente et al. 2003, Overall and Blobel 2007]. Cleavage studies are performed to identify the specific function and regulatory pathway of a given protease in a specific organism. In addition, the specificity of the positive substrate site is used as a reference for drug design. However, proteins are highly complex and dynamic. Proteins have unique compositions and undergo different modification processes in different tissues and cells, as well as in different diseases and disease stages. Therefore, proteases have different cleavage sites [Doucet et al. 2008, López-Otín and Overall 2002] in every single protein.

Extracellular matrix (EM) degradation by matrix metalloproteinases (MMPs) is highly associated with tumorigenesis [Kessenbrock et al. 2010]. Immunosuppressant drugs were first investigated twenty years ago and are currently undergoing human clinical trials. However, clinical trials have shown that immunosuppressants have poor therapeutic effects. Although they can suppress tumor metastasis, immunosuppressants have failed to improve the survival rate of patients with cancer and have instead caused several side effects, such as muscle pain and joint disease. The poor therapeutic effects of immunosuppresants may be attributed to the following two points: 1) All 23 MMPs are structurally similar and contain a zinc peptide restriction enzyme. Immunosuppressants cannot suppress specific MMPs and may suppress nontargeted MMPs. Nonspecific targeting by immunosuppressants may thus affect other physical functions. 2) In addition to the EM, MMPs act on cytohormones, cell membrane receptors, and growth factors. Thus, MMPs affect cell growth, differentiation, movement, and other mechanisms. Furthermore, knowledge of the regulatory pathways and substrates of MMPs remains incomplete; therefore, the complete scope of immunosuppressant influence cannot be predicted [Coussens et al. 2002, Turk 2006, Drag and Salvesen 2010].

Two types of gelatainases exist: Gelatinase A, also referred to as matrix metalloproteinase-2 (MMP-2) and gelatinase B, also referred to as matrix metalloproteinase-9 (MMP-9). MMP-2 and MMP-9 are potential targets and biomarkers in cancer treatment. These proteinases can be distinguished from other MMPs by the presence of three fiber chain protein domains in their catalytic regions. MMP-2 and MMP-9 have similar structures and can be differentiated from each other on the basis of the length of the zone connecting their catalytic regions and heme-binding domains [Zou et al 2016]. Although they share common substrates, they independently catalyze different substrates and affect different messaging pathways [Bauvois 2012]. In contrast to other proteases, like caspase, with specific positive amino acid sites and that cleave the peptide bond after aspartate [Timmer et al. 2009], MMP-9 and MMP-2 lack fixed positive amino acid sites and instead have different amino acids in their positive sites [Prudova et al. 2010]. Hence, predicting the positive sites and substrates of these MMPs is challenging. The substrate cleavage site can be identified through mass spectrometry analysis. Nevertheless, protein expression and translation vary across different types of cells at different stages of the cell cycle, thus complicating the prediction of positive sites and substrates. Comprehensive substrate identification through experiments is expensive and time consuming. Therefore, the in silico prediction of MMP-2 and MMP-9 positive sites has been proposed for the mass annotation of substrate cleavage sites. In silico prediction methods can be applied in biotic experiments to identify candidates with high accuracy rate [Wang et al. 2017].

Numerous systems for the prediction of protease-substrate positions exist [Song et al. 2011] and commonly adopt fractional calculation or machine learning [Wei et al. 2017]. Systems that are based on fractional calculation include GPS-CCD, CaSPreditor, PoPS, and SitePrediction. GPS-CCD makes use of BLOSUM 62 to transfer amino acids through an algorithm to predict the substrate positive site of Calpain. CaSPreditor uses BLOSUM 62 for transfer and adds a PEST-like sequence to calculate fractions [Liu et al. 2011]. PoPS allows users to define the physical–chemical fraction and weight of a specific substrate to calculate fractions [Boyd et al. 2004]. SitePrediction applies the appearance rate of each kind of amino acid and calculates the fraction of amino acid substitution matrix on every position to predict the substrate positive site [Verspurten et al. 2009]. PoPS and SitePrediction can provide substrate positive site predictions, secondary structure prediction, and other extra information for all currently known proteases. Methods based on machine learning include Pripper, Cascleave, and PCSS. Pripper has been used to predict the substrate positive site of caspase through the adoption of binary code in support vector machine (SVM), random forest and J48 [Piippo et al. 2010]. Cascleave, another system for the prediction of the caspase substrate positive site, added structure information and Bi-profile Bayesian signature code [Jia et al. 2017], but not binary code, to support vector regression and provide predictions [Song et al. 2010]. PCSS make the users input training data sets themselves and takes advantage of sequence and structure information to support the vector machine in constructing prediction system. Given that the users input training data sets themselves, the substrates of all kinds of protease can be predicted [Barkan et al. 2010].

This research built a high-precision positive-site prediction system for MMP-2 and MMP-9. This prediction system has two levels. The first level is constructed on the basis of binary information, physical–chemical properties, and structural information. Protein characteristics, such as the fold-change of amino acid number, were adopted to illustrate positive and negative sites in databases and to coordinate with SVM to build a model with four characteristics. The secondary level of this system integrates the prediction confidence index of every feature model and compares various machine learning methods to build models. The feature models on the first level all test seven different groups of negative sites to identify the most suitable dataset that can represent the characteristic of feature codes with high accuracy. Consequently, this whole prediction system can learn additional negative site information. To identify the reason for the failure of MMP-2 and MMP-9 as cancer-targeting drugs, physical–chemical property selection was performed in further steps. The accuracy rate of the prediction system was improved by exploiting fold change and amino acid composition information

2 Research methods

2.1 Data Collection

MEROPS [Rawlings et al. 2016] was used to screen out human MMP-2 and MMP-9 substrates, remove duplicates, and redundant data. CD-HIT removed 70% of similar proteins [Li and Godzik 2006]. The experimental data for MMP-2 included 1269 substrates and cleavage sites in 630 proteins, whereas those for MMP-9 data included 269 cleavage sites in 42 proteins. Cleavage sites that were not annotated as MMP-2 or MMP-9 cleavage sites were defined as negative sites. The number of negative sites was high (330457 and 30558) because MMP-2 and MMP-9 do not possess specific amino acids at their cleavage sites. In this experiment, seven negative sets (N1–N7) were randomly selected with a positive set number of 1:1, and each negative set with a positive set formed seven training sets trained in different coding methods. Thereafter, a negative set (N8) other than N1 to N7 was randomly selected as the training set for the second layer.

2.2 Construction of the Prediction System

In this study, binary, physical–chemical property, structural information, and fold-change features were encoded. The second-layer model was constructed to improve prediction accuracy. To build the second layer, four feature models of the first layer were constructed by using SVM, and the probability confidence scores of prediction results from four feature models were merged. To find the optimal method for building the two-layer prediction system, the prediction model of the second layer was built with LibSVM, J48, Random Forest, and IBK classifiers [Frank et al. 2004]. The performance of each classifier was tested through 10-fold cross-validation.

2.3 Feature Model of the First Level

The first layer has four feature models: binary, physical-chemical property, structural information and fold change. The window size indicates that the first amino acid at the N-terminus of the cleaved peptide bond is P1, and the first amino acid at the C-terminus is P1 '. The largest window size is the twentieth amino acid P20, P19, P18 ... P1, P1 '... P18', P19', P20'.

Binary: Amino acid fragments of various lengths from P20 to P20' are encoding in vector format. The 20 amino acids including Gap are encoded in a vector of 21 dimensions, with the amino acid occupying a dimension of one. E.g:
Alanine(A) 000000000000000000001
Cysteine(C) 000000000000000000010

Fold change: P20-P20 'fragments of the positive and negative sites of MMP-2 and MMP-9 in training set were calculated by Icelogo software [Colaert et al. 2009]. The frequency of occurrence of each amino acid on each of the positions is calculated according to the following formula and divided to give the value of the fold change. The difference in frequency of occurrence of each amino acid between the positive site and the negative site to show the differential nature of positive and negative data for different dataset. For example, if P+N1 is used to build the model, the positive set will be divided into 10 parts for the 10-fold cross-validation, and the Fold Change substitution table will be calculated respectively with the negative set of the sum N2-N7.

$$Frequency+ = \frac{AA1}{all\ AA} \text{ in positive set} \quad (1)$$

$$Frequency- = \frac{AA1}{all\ AA} \text{ in negative set} \quad (2)$$

$$\text{Fold chang} = \frac{Frequency+}{Frequency-} \quad (3)$$

Structural information: The protein sequence was sent to DISOPRED [Ward et al. 2004] and NetSurfP [Petersen et al. 2009] for prediction and encoding with P20-P20 'fragment. The encoded information contained the probability of disorder; relative surface accessibility; absolute surface accessibility; z-fit score; and probability of alpha-helix, beta-strand, and coil formation.

Physical–chemical property: The physical–chemical properties of amino acids were retrieved from the AAindex [Kawashima et al. 2008] database. 544 data to remove null values and Pearson correlation coefficient greater than 0.8 with 371 data. Venkatarajanet al. incorporated 237 physical–chemical properties of amino acids into a five-vector feature [Venkatarajan and Braun 2001]. The gap value was defined as the average of 20 amino acid attribute values. Therefore, 376 physical–chemical features were used. The cleavage sites and noncleavage site sequences of P20–P20, with positive (+1) and negative ('1), were encoded with 376 physical–chemical properties, each of which will be tested through Pearson correlation. When the correlation coefficient |R| exceeded 0.05–0.3 and the P-value was less than 0.001, the relative features of this site were selected, and the remaining features were removed. Therefore, not every site was encoded, and different numbers of sites were encoded.

2.4 Integration Model for the Secondary Level

The predictive model was built with the training set with the best predictive performance in each feature encoding. The negative site of the seven negative sets that do not contain the first layer was randomly selected, and the ratio of the positive site to the negative site was 1:1, which becomes the eighth negative set (N8). Then, the subset was input in the four feature models of first layer to predict the output of the four models and the probability of positive sites and negative sites. This probability was used as the feature of the second layer. LibSVM, Multilayer Perceptron, J48, Random Forest, and IBK were tested to find the best classifiers for the second layer.

2.5 Similarity Analysis of Data Sets

BLOSUM62 Matrix was used to calculate the similarity between two amino acids on the basis of the positive set and seven negative sites (N1–N7). The alignment scores for each amino acid in each of the positions to the other amino acids in the same position were averaged. For example, in 100 pieces of data, the first site of the first amino acid will be the other 99 amino acids as inferred from BLOSUM62 matrix conversion scores. The alignment scores, specifically, the first sites of the similarity score, were then averaged. P20–P20' would have an average score.

2.6 Evaluation of Model Prediction Ability

Accuracy (Acc), sensitivity (Sn), specificity (Sp), and the Matthews correlation coefficient (MCC) were used to evaluate the predictive ability of each system. Four measures were defined:

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{4}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$Sp = \frac{TN}{TN + FP} \tag{6}$$

$$Sn = \frac{TP}{TP + FN} \tag{7}$$

where TP, FP, FN and TN are true positives, false positives, false negatives, and true negatives, respectively. Sn and Sp represent the rate of true positives and true negatives respectively. Acc is the overall accuracy of prediction. Additionally, MCC is a measure of the quality of the classifications, and the value may range between -1 (an inverse prediction) and +1 (a perfect prediction), with 0 denoting a random prediction.

3 Results

3.1 Prediction Efficiency of the First Level

Binary model (B)- We used binary code to obtain the primary protein structure. We screened different window sizes which are from the cleavage site to the amino-terminal (N) end to the carboxyl-terminal (C) end for each 20 amino acids to optimize the accuracy. As fig.1 shows, both MMP-2 and MMP-9 model were significantly increased in MCC when the fragment from the beginning of P3. In MMP-2 model, the fragments from P3-15 of N-terminal to P3'-20' of C-terminal have above 0.5 in MCC. The best performance is obtained in P3-13' fragment. The fragment length of MMP-9 from P3-15 with C-terminal P2 'to P20' have more than 0.4 of MCC, the best MCC convergence is at P3 to P9'. The convergence region shows the determined length of amino acid specificity around the cleavage site. In MMP-2 model, the Sn of every training set was more than 0.9, Sp was more than 0.85, and ACC was up to 0.9. The 7th training set (P4-P5') has the highest MCC, which is 0.808. The diversity of MCC in MMP-2 model training sets is 0.034 (table 1). In MMP-9 model, every training set got more than 0.7 in Sn, 0.75 in Sp, and around 0.75 in ACC. The 7th training set (P3-P5') has the best MCC at 0.570. The difference between the best and the worst training set is 0.108 (Table 2).

MMP-2 first layer models	Acc (model)	Highest MCC (model)	Lowest MCC (model)
Binary	0.903 (7)	0.808 (7)	0.774 (2)
Fold change	0.897 (3)	0.796 (3)	0.687 (2)
Structure	0.824 (7)	0.649(7)	0.609 (2)
Physical-chemical property (|R| > 0.05)	0.907 (1)	0.814 (1)	0.778 (6)

Table 1: Acc, Highest and lowest MCC of training sets for the MPP-2 first-layer model

MMP-9 first layer models	Acc (model)	Highest MCC (model)	Lowest MCC (model)
Binary	0.782 (7)	0.570 (7)	0.462 (5)
Fold change	0.690 (2)	0.386 (2)	0.305 (4)
Structure	0.699 (2)	0.400 (2)	0.309 (1)
Physical-chemical property (|R| > 0.2)	0.805 (7)	0.613 (7)	0.446 (5)

Table 2: Acc, Highest and lowest MCC of training sets for the MPP-9 first-layer model

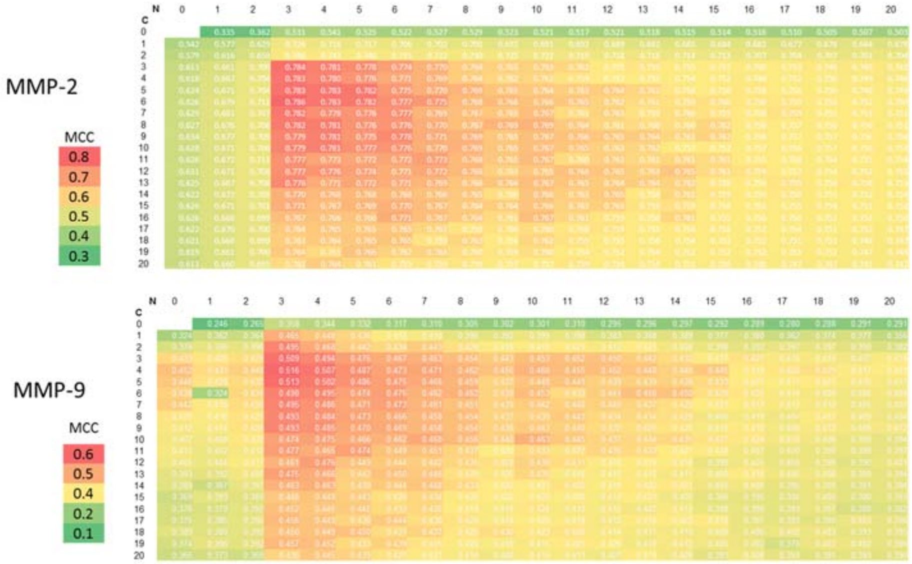


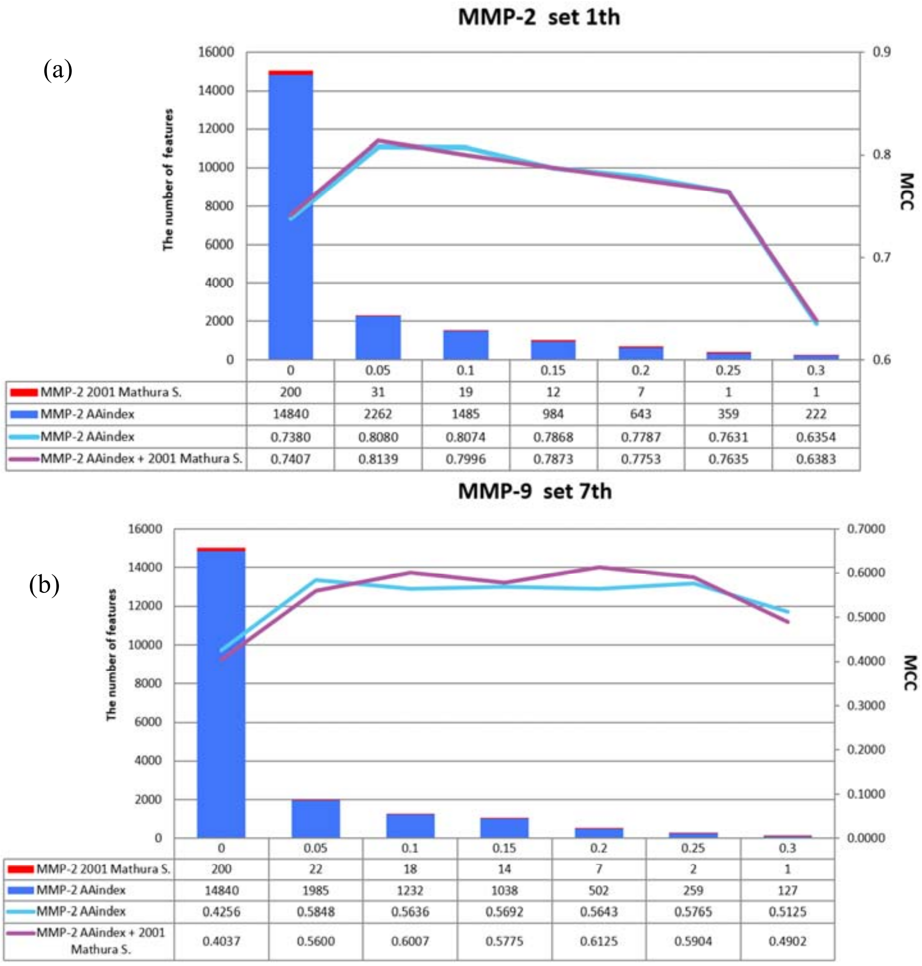
Figure 1: MCC of various window sizes obtained through binary coding (note: X and Y axis represent the length of amino acid from the N to C terminals) (a) Heat map of MMP-2 (b) Heat map of MMP-9

Fold-change model- Icelogo provided the fold change value of P20–P20' for 40 amino acids. The value provided the fold difference in the ratio of amino acid type on every position in the positive and negative sets. The Sn and Sp of the MMP-2 model exceeded 0.83, and ACC was approximately 0.85. The optimal training set for the MMP-2 model was the third set, which had a MCC of 0.796. In this coding, the diversity of MCC in training sets is 0.796. The Sn and Sp of the MMP-9 model exceeded 0.65, and ACC was approximately 0.65. The best training set for this model was the second set, which had a MCC of 0.4. The diversity of MMP-9 training sets is 0.081 in MCC.

Structure model- The output of the structure prediction system was used as a feature. The output included relative surface accessibility, absolute surface accessibility, z-fit score, probability for alpha-helix, probability for beta-strand, and probability for coil information. Each training set for the MMP-2 model yielded Sn and Sp values of more than 0.8 and ACC values of approximately 0.8. The seventh set was the best predictor training set and exhibited an MCC of 0.649. Each training set for the MMP-9 model yielded Sn and Sp values of more than 0.65 and ACC of approximately 0.65. The best predictor training set was the second set, which had an MCC of 0.4. The difference between the best and the worst MCC in the training sets of MMP-2 and 9 is 0.04 and 0.091.

Physical–chemical property model- The physical–chemical properties of the amino acid sequences of P20 to P20' were encoded and then subjected to the correlation coefficient test to perform feature selection. It would present the importance of specific position in sequence then the final coding was a discontinuity fragment. Different features for encoding were obtained after 0 to 0.3 threshold selection. In MMP-2 data, the MCC of the selected feature-built set was approximately 0.8 at |R|> 0.05. As |R| increased, the MCC gradually decreased to approximately 0.6 (Figure 2 (a)). The first training set had the highest MCC of 0.814 with |R|> 0.05. At the same |R|> 0.05, the worst MCC was 0.778 for the sixth training set. The MCC of each dataset for MMP-9 fluctuated with increasing |R| (Fig.2(B)). The seventh training set of MMP-9 had the best MCC of 0.613 as |R| increased to 0.2. At the same |R|> 0.2, the fifth training set had the the worst MCC of 0.446.

The addition of five-dimensional amino acid properties (Venkatarajan et al., 2001) improved accuracy. The correlation coefficient increased as the number of selected features decreased. The five-dimensional amino acid profile accounted for approximately 1% of the overall profile, with AAindex accounting for a large number of features. The five-dimensional feature of MMP-2 increased the MCC by 0.01 when |R|> 0.1 (Fig.2 (a)). In the MMP-9 model when |R|> 0.1 to 0.2, the addition of the five-dimensional protein features caused MCC to increase. In particular, when R = 0.2, MCC increased by 0.04 (Fig. 2 (b)).



	0	0.05	0.1	0.15	0.2	0.25	0.3
MMP-2 2001 Mathura S.	200	31	19	12	7	1	1
MMP-2 AAindex	14840	2262	1485	984	643	359	222
MMP-2 AAindex	0.7380	0.8080	0.8074	0.7868	0.7787	0.7631	0.6354
MMP-2 AAindex + 2001 Mathura S.	0.7407	0.8139	0.7996	0.7873	0.7753	0.7635	0.6383

	0	0.05	0.1	0.15	0.2	0.25	0.3
MMP-2 2001 Mathura S.	200	22	18	14	7	2	1
MMP-2 AAindex	14840	1985	1232	1038	502	259	127
MMP-2 AAindex	0.4256	0.5848	0.5636	0.5692	0.5643	0.5765	0.5125
MMP-2 AAindex + 2001 Mathura S.	0.4037	0.5600	0.6007	0.5775	0.6125	0.5904	0.4902

Figure 2: Variation in MCC after the selection of physical–chemical property models based on correlation coefficient

3.2 Prediction Efficiency of the Second Level

The secondary level integrated the outputs of every coding model in the primary level. The MMP-2 model combined the best performance of the seventh set for binary coding, the third set for fold change, the seventh set for structural information, and the first set for physical–chemical property. The MMP-9 model also combined the best performance of the seventh set for binary coding, the second set for fold change, the second set for structural information, and the seventh set for physical–chemical property (Table 2). We compared the prediction performance of the binary, physical–chemical, structural, and fold change feature models with LibSVM, J48, Random Forest, and IBK. The prediction results for the second layer of MMP-2 are as follows: The MCC of three or four feature models exceeded 0.8. The best result was obtained by a prediction system constructed by LibSVM with four feature models. This system had an MCC of 0.894. The second layer of MMP-9 with the optimal MCC of 0.644 was constructed using four feature models and IBK (K = 23). Prediction performance decreased in the absence of the FG feature model. The addition of the FG model to the MMP-2 and MMP-9 models increased prediction performance by 5.52% and 10.62%, respectively (Table 3). The final complete prediction system was built on the first layer of the best MCC of each feature model and the second layer of the machine-learning method.

MMP-2 second layer	J48	Random Forest	IBK	LibSVM
3 models	0.825	0.821	0.829	0.834
4 models	0.883	0.888	0.893	0.894
Increase	5.8%	6.7%	6.4%	6%

Table 3: Comparison of the MCC of the MPP-2 first-layer model based on different machine learning methods and with or without FG

MMP-9 second layer	J48	Random Forest	IBK	LibSVM
3 models	0.522	0.447	0.503	0.512
4 models	0.566	0.600	0.644	0.625
Increase	4.4%	15.3%	14.1%	11.3%

Table 4: Comparison of the MCC of the MPP-9 first-layer model based on different machine learning methods and with or without FG

4 Discussion

The gap between the highest and lowest MCC in the four codes of MMP-2 was approximately 0.03 and was different by approximately 0.1 from the code of MMP-9. This result can be attributed to the five-fold difference between the sizes of MMP-2 and MMP-9 data. In every kind of code, the difference between the MCC of MMP-2 and that of MMP-9 exceeded 0.2 because massive amounts of data were available for MMP-2 (1269 positive sites) and could facilitate the construction of a prediction system on the basis of negative sites. By contrast, less material was available for MMP-9 and concentrated negative sites were preferentially selected, thus introducing prediction efficiency bias and influencing the accuracy of the prediction system. The difference in groups with the highest and lowest MCC can be attributed to the random selection of negative sites with diverse compositions. Calculating the similarity of positive and negative sites with BLOSU62 revealed that the trends of every negative set were not in complete accordance with those of negative sets. (Table 3) The difference between each negative set and positive set resulted in the difference in prediction system training. Hence, the accuracy rate changed in accordance with the negative set. Notably, the similarity curve of positive and negative sites in amino acids were distinct if MMP-2 was at the period from P4 to P4′ (Table 6(a) and

MMP-9 was at P13, P3, P1–P3′, P12′, P15′, or P18′ (Table 6(b)). These positions corresponded to the code positions selected from physical–chemical property codes on the basis of correlation coefficients. The application of the relevant coefficient to select characteristics not improved MCC but also conformed to the relationship of sequence similarity.

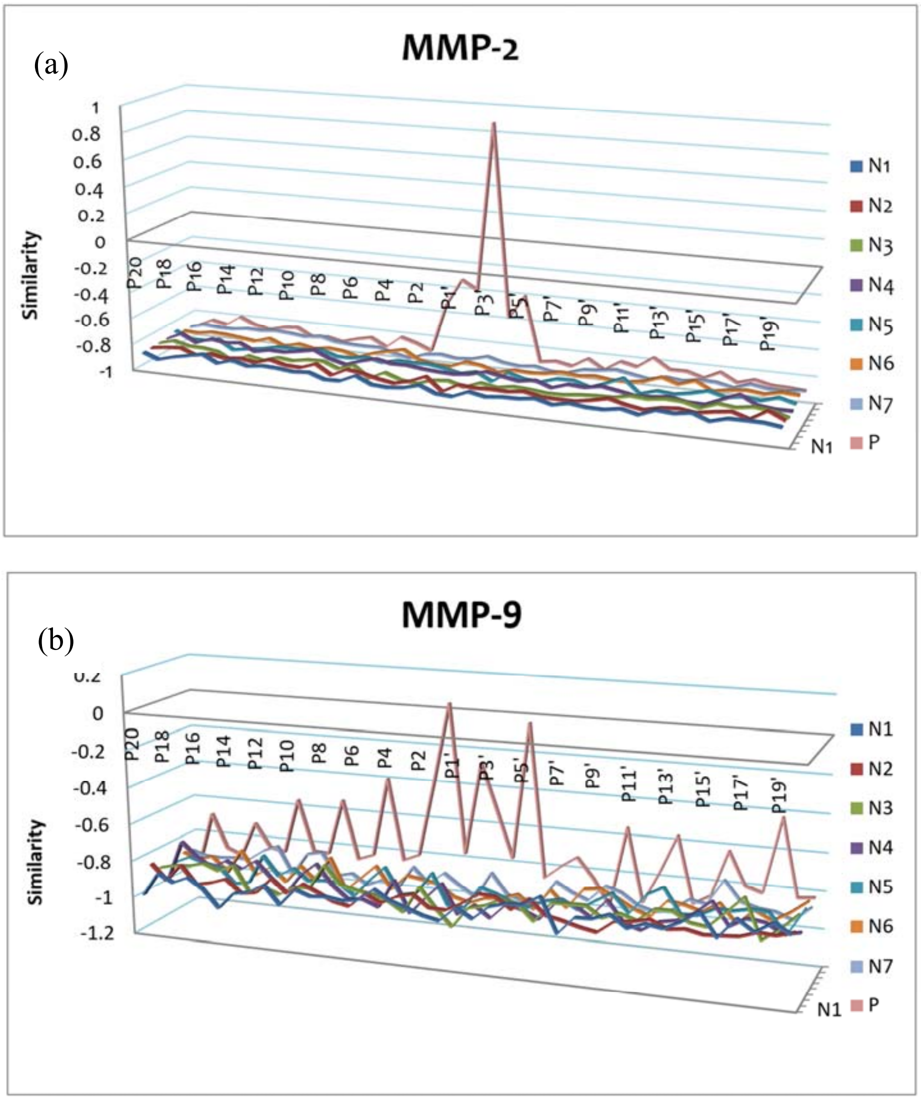


Figure 3: Similarity ratio of positive- and negative-site amino acid sequence. P is a positive site base. N1 is the negative site base of the first group, and so on. (a) Similarity tendency graph of MMP-2 data. (b) Similarity tendency graph of MMP-9 data.

The flow design of this experiment is to allow machine learning to obtain additional negative sets and allow every code to identify the suitable training set. Then, the code integration of different meanings on the secondary level yields high numbers of negative messages that solve the problem of overly large negative site and provide references for future prediction system frameworks. This experiment is the first to adopt the fold-change value provided by Icelog to identify multiplier difference between positive and negative sets. Integrating this characteristic in the secondary level helped improve the prediction efficiency. This code can resolve the problem of excessive negative data as well. If a suitable negative site group can be selected to delegate the whole negative site materials, the appropriate fold-change value can be achieved and used to describe the difference in comprehensiveness.

This system is more user-friendly and precise than current prediction systems. Furthermore, this system is the only full-time prediction system that predicts the tangent point of MMP-2 and MMP-9 and is designed to accept massive inputs of protein sequences at a time. MMP-2 and MMP-9 are closely related with tumor metastasis. In this study, a cleavage point prediction system for MMP-2 and MMP-9 was constructed to predict new possible substrates and then estimate other regulatory pathways for these metalloproteinases. This system will aid provide references for drug design.

Acknowledgements

This work was supported in part by the Ministry of Science and Technology, Taiwan, Republic of China (under grants 106-2221-E-005-007-MY2 and 107-2634-F-005-002) and National Chung Hsing University and Chung Shan Medical University (under grants NCHU-CSMU10612).

References

[Bauvois 2012] Bauvois, B.: “New facets of matrix metalloproteinases MMP-2 and MMP-9 as cell surface transducers: Outside-in signaling and relationship to tumor progression.” Biochimica et Biophysica Acta (BBA)-Reviews on Cancer 1825.1 (2012): 29-36.

[Barkan et al. 2010] Barkan, D. T., Hostetter, D. R., Mahrus, S., Pieper, U., Wells, J. A., Craik, C. S., Sali, A.: “Prediction of protease substrates using sequence and structure features.” Bioinformatics 26.14 (2010): 1714-1722.

[Boyd et al. 2004] Boyd, S. E., de la Banda, M. G., Pike, R. N., Whisstock, J. C., Rudy, G. B.: “PoPS: a computational tool for modeling and predicting protease specificity.” Computational Systems Bioinformatics Conference (2004): 372-381.

[Colaert et al. 2009] Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J., Gevaert, K.: “Improved visualization of protein consensus sequences by iceLogo.” Nature methods 6.11 (2009): 786-787.

[Coussens et al. 2002] Coussens, L. M., Fingleton, B., Matrisian, L. M.: “Matrix metalloproteinase inhibitors and cancer—trials and tribulations.” Science 295.5564 (2002): 2387-2392.

[Doucet et al. 2008] Doucet, A., Butler, G. S., Rodríguez, D., Prudova, A., Overall, C. M.: “Metadegradomics.” Molecular & Cellular Proteomics 7.10 (2008): 1925-1951.

[Drag and Salvesen 2010] Drag, M., Salvesen, G. S.: "Emerging principles in protease-based drug discovery." Nature reviews Drug discovery 9.9 (2010): 690-701.

[Frank et a;. 2004] Frank, E., Hall, M., Trigg, L., Holmes, G., Witten, I. H.: "Data mining in bioinformatics using Weka." Bioinformatics 20.15 (2004): 2479-2481.

[Jia et al. 2017] Jia, C., He, W., Zou, Q.: "DephosSitePred: a high accuracy predictor for protein dephosphorylation sites." Combinatorial Chemistry & High Throughput Screening 20.2 (2017): 153-157

[Kawashima et al. 2008] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M.: "AAindex: amino acid index database, progress report 2008." Nucleic acids research 36.1 (2008): D202-D205.

[Kessenbrock et al. 2010] Kessenbrock, K., Plaks, V., Werb, Z.: "Matrix metalloproteinases: regulators of the tumor microenvironment." Cell 141.1 (2010): 52-67.

[Liu et al. 2011] Liu, Z., Cao, J., Gao, X., Ma, Q., Ren, J., Xue, Y.: "GPS-CCD: a novel computational program for the prediction of calpain cleavage sites." PLoS One 6.4 (2011): e19001.

[Li and Godzik 2006] Li, W., Godzik, A.: "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics 22.13 (2006): 1658-1659.

[López-Otín and Overall 2002] López-Otín, C., Overall, C. M.: "Protease degradomics: a new challenge for proteomics." Nature Reviews Molecular Cell Biology 3.7 (2002): 509-519.

[Overall and Blobel 2007] Overall, C. M., Blobel, C. P.: "In search of partners: linking extracellular proteases to substrates." Nature Reviews Molecular Cell Biology 8.3 (2007): 245-257.

[Petersen et al. 2009] Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M., Lundegaard, C.: "A generic method for assignment of reliability scores applied to solvent accessibility predictions." BMC Structural Biology 9.1 (2009): 51.

[Piippo et al. 2010] Piippo, M., Lietzén, N., Nevalainen, O. S., Salmi, J., Nyman, T. A.: "Pripper: prediction of caspase cleavage sites from whole proteomes." BMC Bioinformatics 11.1 (2010): 320.

[Puente et al. 2003] Puente, X. S., Sánchez, L. M., Overall, C. M., López-Otín, C.: "Human and mouse proteases: a comparative genomic approach." Nature Reviews Genetics 4.7 (2003): 544-558.

[Prudova et al. 2010] Prudova, A., Auf Dem Keller, U., Butler, G. S., Overall, C. M.: "Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics." Molecular & Cellular Proteomics 9.5 (2010): 894-911.

[Rawlings et al. 2016] Rawlings, N. D., Barrett, A. J., Finn, R. D.: "Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors." Nucleic acids research 44.D1 (2016): D343-D350.

[Song et al. 2011] Song, J., Tan, H., Boyd, S. E., Shen, H., Mahmood, K., Webb, G. I., Akutsu, T., Whisstock, J. C.: "Bioinformatic approaches for predicting substrates of proteases." Journal of bioinformatics and computational biology 9.1 (2011): 149-178.

[Song et al. 2010] Song, J., Tan, H., Shen, H., Mahmood, K., Boyd, S. E., Webb, G. I., Akutsu, T., Whisstock, J. C.: "Cascleave: towards more accurate prediction of caspase substrate cleavage sites." Bioinformatics 26.6 (2010): 752-760.

[Timmer et al. 2009] Timmer, J. C., Zhu, W., Pop, C., Regan, T., Snipas, S. J., Eroshkin, A. M., Riedl, S. J., Salvesen, G. S.: “Structural and kinetic determinants of protease substrates.” Nature structural & molecular biology 16.10 (2009): 1101-1108.

[Turk 2006] Turk, B.: “Targeting proteases: successes, failures and future prospects.” Nature reviews Drug discovery 5.9 (2006): 785-799.

[Venkatarajan and Braun 2001] Venkatarajan, M. S., Braun, W.: “New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties.” Journal of Molecular Modeling 7.12 (2001): 445-453.

[Verspurten et al. 2009] Verspurten, J., Gevaert, K., Declercq, W., Vandenabeele, P.: “SitePredicting the cleavage of proteinase substrates.” Trends in biochemical sciences 34.7 (2009): 319-323.

[Wang et al. 2017] Wang, Y., Song, J., Marquez-Lago, T. T., Leier, A., Li, C., Lithgow, T., Webb, G. I., Shen, H. B.: “Knowledge-transfer learning for prediction of matrix metalloprotease substrate-cleavage sites.” Scientific reports 7.1 (2017): 5755.

[Ward et al. 2004] Ward, J., Sodhi, J., McGuffin, L., Buxton, B., Jones, D.: “Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.” Journal of molecular biology 337.3 (2004): 635-645.

[Wei et al. 2017] Wei, L., Tang, J., Zou, Q.: “Local-DPP: An Improved DNA-binding Protein Prediction Method by Exploring Local Evolutionary Information.” Information Sciences 384.1 (2017):135-144.

[Zou et al 2016] Zou, Q., Wan, S., Ju, Y., Tang, J., Zeng, X. “Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy.” BMC System Biology 10.Suppl 4 (2016): 114.