Machine Learning Model for Domain Based Classification of MMP's

K Pardasani, B Pant, K Pant

Citation

K Pardasani, B Pant, K Pant. *Machine Learning Model for Domain Based Classification of MMP's*. The Internet Journal of Genomics and Proteomics. 2009 Volume 5 Number 2.

Abstract

Purpose: Classification of human matrix metalloproteinase's (MMP's or mmp's) using various machine learning techniques so as to reduce the time and economic constraint which protein classification poses on the existing wet lab techniques. Methods: The domain information of various MMP's was obtained from UniProtKB/Swiss-Prot of ExPASy Proteomics Server. Various machine learning tools like Naïve Byes, Random Forest and Decision tree were used. The domain data for all the eight classes was incorporated into the three classifiers. Results: Out of the three the Naïve Bayes and Random forest performed best and gave accuracy of 78.26% and 73.91% where out of 23 MMP's taken for cross validation 18 were correctly classified by Naïve Bayes classifier.Discussion and Conclusion: The above classifier takes into account the domain structure of all the known 23 MMP's as the Naïve Bayes classifier performs the best among all the three classifiers, it qualifies as most suitable choice for classification. Information about substrate specificity and tissue inhibitor of matrix metalloproteinase's (TIMP) was not included since information about these is not completely available. No such algorithm for classification of MMP's has been reported so far in the literature. Such classification can be extended to other proteins as well where adequate domain information is available.

INTRODUCTION

MMP's have major role to play in skin aging. Besides other causes of aging exposure to ultraviolet (referred to as UVA or UVB) radiation emanating from sunlight accounts for about 90% of the symptoms of premature skin ageing, and most of these effects occur by 20 years of age. Both UVA and UVB rays cause damage leading to wrinkles, lower immunity against infection, ageing skin disorders, and cancer. Even small amounts of UV radiation damage collagen fibers (the major structural protein in the skin) and cause accumulation of abnormal elastin (the protein that causes tissue to stretch). During the process, large amounts of enzymes called metalloproteinase's are produced. The normal function of these enzymes is to remodel the sun-injured tissue by synthesizing and reforming collagen. This is an imperfect process, however, and to achieve it, some of these enzymes actually degrade collagen. The result is an uneven formation (matrix) of disorganized collagen fibers called solar scars. If this process of imperfect skin rebuilding occurs over and over, wrinkles result. Sunlight damages collagen fibers (the major structural protein in the skin) and causes accumulation of abnormal elastin (the protein that causes tissue to stretch) which leads to production of metalloproteinase's [1].

MMP's have been found to play important role in various aspects of cancer. These clinical trials have led to the recognition that specific mmps are used in conjunction with cytotoxic chemotherapy in early stage of cancer. In case of cancer MMP-2 and MMP-9 play an important role in degradation of type IV collagen which is a major protein component of basement membrane [11], [12].

Many reports have indicated the role of MMP1 and MMP9 in rheumatoid and osteoarthritis. The enzyme aggrecanase, a member of the ADAM family of metalloproteinases is thought to play an important role in articular damage [13].

Various studies have shown an increased expression of MMP-9 at the sites of atherosclerosis and aneurysm formation [2]. Secretion and activation of MMP's by macrophages induces degradation of E.C.M. in the atherosclerosis plaque and plaque rupture. Hence MMPs are proposed to represent sensitive markers of inflammation in patients with coronary heart disease.

Increased level of MMP's has been found in various lung diseases including acute respiratory distress syndrome, asthama, bronchiectasis, and cystic fibrosis. More studies regarding MMP inhibitors are required to be done as potential therapy [3].

MMP's have been found to be a key player in neurological disorders like Multiple sclerosis and Guillain-Barre's syndrome etc [2], [4].

MMP8 and MMP9 which are stored in the granules of polymorphonuclear leukocytes are found to be involved in inflammatory and infectious processes. It was proposed that specific MMP9 inhibition constitutes a potential approach for the treatment of septic shock syndromes [5].

Acute and chronic wounds are found to have high levels of MMP2 and MMP9. It has been suggested that ulcers generate a local environment of activated MMPs which delay the process of healing [2]. MMP9 has been found to play a part in blistering skin diseases and contact hypersensitivity. MMP's have long been implicated in periodontal disease and more recently, in inflammatory bowel diseases [7].

Thus there is a need for better understanding of the role played by these MMPs in pathophysiological conditions of the body for the molecular biologists and scientists for devising new vaccine candidates and therapeutic agents for prevention and cure of disease. Thus there arises a need for identification and classification of MMP's and mapping them with functional and structural aspects. MMP's have been classified by using substrate specificity and cellular localization [16]. Some wet lab experiments have been performed for identification and classification of MMP's but they are highly expensive and time consuming. Traditionally MMP's have been classified on evolutionary and functional basis. But no such attempt has been made to classify MMPs on the basis of domain knowledge using computational techniques.

In view of the above a computational model has been developed for classification of MMP's using the domain structure. The results are cross validated with existing classes in UniProtKB/Swiss-Prot of ExPASy Proteomics Server.

MMP's are made up of the following homologous domains:

1) Predomain- It contains a signal peptide or a leader sequence which targets MMPs to the secretory or plasma membrane insertion pathway.

2) Prodomain– Contains a conserved cysteine switch motif of PRCXXPD for making the proMMP latent by occupying the active site Zinc and making the enzyme inaccessible to substrate.

3) Zinc containing catalytic domain--The structure of this domain is quiet similar in all MMPs. It has a motif, HEF/LGHS/ALGLXHS, which coordinates a zinc atom at the active site. Besides catalytic zinc active site contains structural zinc and two to three calcium ions. A sub-site- or S1'-pocket- or channel-like structure is a binding site for a substrate or inhibitor molecule within the active site, which is quiet different in size and shape among various MMPs.

4) Hemopexin domain which mediates interactions with substrates and confers specificity of the enzymes; and

5) Hinge region which links the catalytic and the hemopexin domain [14], [15].

Besides this other domains are found in MMPs which give additional properties to them.

The smallest MMP in size, MMP-7 does not contain hemopexin domain yet displays substrate specificity. The membrane type MMPs (MMP-14, MMP-15, MMP-16,MMP-24) contain a domain which is trans membranous and about 20 amino acid in length with a small cytoplasmic domain included in the structure as well. The other membrane type MMPs contain a glycosylphosphatidyl inositol linkage for attaching them to cell surface hence classified as glycosyl-phosphatidyl inositol (GPI)-linked MMPs. The gelatin binding MMPs contain three internal repeats called fibronectin domain for binding to their substrate. Furin-activated secreted MMPs (MMP-11 and MMP-28) contain recognition motif for furin-like serine proteinases present in their catalytic domain for intracellular activation. This motif is also found in the vitronectin-like insert MMPs (MMP-21), and the MT-MMPs [16]. MMP-23 has cysteine array and immunoglobulin (Ig)-like domains but no conserved hemopexin-like domain [17]. It is also classified as type II transmembrane MMP, since it has an amino-terminal signal anchor (CA) targeting it to the cell membrane [17].

MATERIAL AND METHODS

All the above said domains were incorporated into the classifier on the basis of which MMPs were divided into eight classes. For classification purpose various classes were abbreviated as follows

Class1) Minimal domain MMP's (MD)

Class2) Simple Hemopexin domain containing MMP's (SHDC)

Class3) Gelatin binding MMP's (GB)

Class4) Furin activated secreted MMP's (FAS)

Class5) Transmembrane MMP's (Type1)(TMT)

Class6) Glycosyl-phosphatidyl inositol linked MMP's(GLM)

Class7) Vitronectin-like insert MMP's (VLI)

Class8) Cysteine/Proline-rich IL-1Receptor-like domain MMP's (CPR)

To improve the accuracy of the classifier further the localization of MMP,s are included as one of the characteristics. The domain and motif knowledge of all MMPs was collected from Swissprot/Uniprot server of Expasy and the model was built using various modules of Weka. The three machine learning classifiers namely Random Forest, Naïve Bayes and Decision Tree have been developed using above data. The details of three approaches employed for classification are given below

RANDOM FOREST

Random Forest is a class of ensemble method specially designed for decision tree classifiers .It combines the prediction made by multiple decision trees where each tree is generated based on the value of an independent set of random vectors .The random vectors are generated from a fixed probability distribution .Bagging using decision trees is a special case of random forests ,where randomness is injected into the model building process by randomly choosing N samples with replacement ,from the original training set. It has been theoretically proved that the upper bound for generalization error of random forests converges to the following expression when the number of trees is sufficiently large [8].

Figure 1

$$\text{Generalization error} \le \frac{\rho\left(1-s^2\right)}{s^2}$$

Where is the average correlation among the trees and s is a quantity that measures the strength of the tree classifier. The strength of a set of classifier refers to the average performance of the classifier where performance is measured probabilistically in terms of the classifier margin.

Figure 2

$$\text{margin}, M(X,Y) = P(Y_\theta = Y) - \max_{Z \neq Y} P(Y_\theta = Z)$$

Where $Y_{\square}$ is the predicted class of X according to a classifier built from some random vector . The higher the margin is, the more likely it is that the classifier correctly predicts a given example X.

NAÏVE BAYES

A naïve bayes classifier estimates the class conditional probability by assuming that the attributes are conditionally independent, given the class label y. The conditional independence assumption can be formally stated as follows [8].

The main advantage of Bayesian classifiers is that they are probabilistic models, robust to real data noise and missing values. The Naive Bayes classifier assumes independence of the attributes used in classification but it has been tested on several artificial and real data sets, showing good performances even when strong attribute dependences are present. In addition, the Naive Bayes classifier can outperform other powerful classifiers when the sample size is small [9]. Since it also has advantages in terms of simplicity, learning speed, classification speed, storage space and incrementality its use is preferred more often.

Figure 3

$$P(X|Y=y) = \prod_{i=1}^{a} P(X_i|Y=y)$$

DECISION TREE

A decision tree (or tree diagram) is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities. In data mining and machine learning, a decision tree is a predictive model; that is, a mapping from observations about an item to conclusions about its target value. More descriptive names for such tree models are classification tree (discrete outcome) or regression tree (continuous outcome). In these tree structures, leaves represent classifications and branches

represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a decision tree from data is called decision tree learning, or (colloquially) decision trees [8].

The above three classifiers are simulated to obtain the results and their comparative analysis has been performed.

RESULT AND DISCUSSION

The three domains viz pre, pro and catalytic were discretized as low, medium and high with the values obtained from Uniprot/Tremble database of Expasy server. Since rest of the domains was present in few MMPs only, hence they were indicated as either present or absent.

The confusion matrix generated from the above is given as under

Figure 4

```
=== Confusion Matrix ===

 a b c d e f g h   <-- classified as
 9 0 0 0 0 0 0 0 | a = SHDC
 1 1 0 0 0 0 0 0 | b = GB
 0 0 2 0 0 0 0 0 | c = MD
 2 0 0 0 0 0 0 0 | d = FAS
 0 0 0 0 4 0 0 0 | e = TMT
 0 0 0 1 0 0 0 0 | f = VLI
 0 0 1 0 0 0 0 0 | g = CPR
 0 0 0 0 0 0 0 2 | h = GLM

=== Detailed Accuracy By Class ===
```

Figure 5

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.214	0.75	1	0.857	1	SHDC
0.5	0	1	0.5	0.667	1	GB
1	0.048	0.667	1	0.8	1	MD
0	0.048	0	0	0	0.929	FAS
1	0	1	1	1	1	TMT
0	0	0	0	0	0.045	VLI
0	0	0	0	0	1	CPR
1	0	1	1	1	1	GLM
0.783	0.092	0.699	0.783	0.724	0.952	Weighted Avg.

MARGIN CURVE

Confidence is measured by the difference between the estimated probabilities of the true class and that of most likely predicted class other than the true class, a quantity known as the margin. The larger the margin, the more confident the classifier is in predicting the true class. It turns out that boosting can increase the margin long after the training error has dropped to zero. The effect can be visualized by plotting the cumulative distribution of the margin values of all the training instances for different numbers of boosting iterations, giving a graph known as the margin curve .With the above data the margin curve was generated with high values.

On X-axis margin number was plotted with number of instances on Y-axis. Out of the 23 instances 5 instances have values lying between 0 and -1 indicating incorrect classification and the values lying above 0 and 1 numbered to 18 indicating the correctly classified instances in the training dataset.

Figure 6

Figure1---Margin curve for decision tree

Figure 7

Figure 2---Margin curve for Random Forest

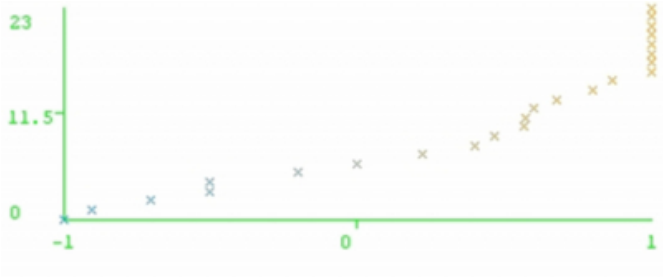


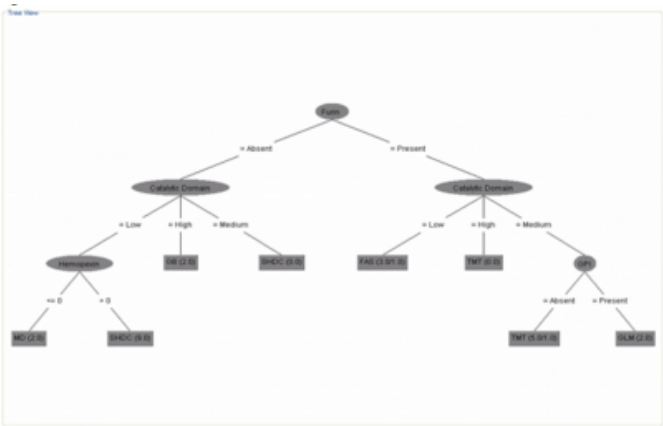
Figure 8

Figure 3---Margin curve for Naïve Bayes

The J48 algorithm of Weka is used to obtain the tree view of classes as shown in fig4

Figure 9

Figure4---Decision Tree classification of MMPs



The accuracy of results obtained by different algorithms is presented in Table -1

Figure 10

Table 1: Accuracy of classifiers

S.no	Name of Algorithm	Accuracy
1	Random Forest	73.913%
2	Decision Tree	56.52%
3	Naïve Bayes	78.26%

Thus we observe that out of the 23 MMPs taken for cross validation 18 were classified correctly whereas 5 were classified incorrectly by naïve bayes classifier. This accounts to 78.2609% accuracy which was the highest among all the three classifier used here so far. Thus the above classifier is able to classify MMPs into eight classes for which no algorithm has been reported in the literature so far. We can increase the instances by adding domain data of other organisms like mouse, rat, pig and others but it does not give any significant change. This implies that the human instances are alone sufficient to develop the classifier. The reason is that similarity is 75-85% for amino acid composition and domain identity is 100% among human and other organism. Hence inclusion of domain data of other organisms will not only increase the instances but also increase the redundancy. The same model can be applied for organism like mouse, rat etc. for which domain information is available in UniProtKB/Swiss-Prot of ExPASy Proteomics Server.

CONCLUSION

The above classifier takes into account the domain structure of all the known 23 MMPs as the naïve Bayes classifier performs the best among all the three classifiers, it qualifies as most suitable choice for classification, information about substrate specificity and tissue inhibitor of matrix metalloproteinases (TIMP) was not included since information about these is not completely available. The authors wish to incorporate it as soon as more information is available in the future. The above model is useful for generating information which can be of great use in prediction of structure and function of MMPs since they are key drug targets. The MMP's belonging to a particular class will have functional domains corresponding to that class which will ease in locating the active site(s) as well as the binding site(s) in the classified domain and hence it can be the potential active site or binding site for the drug. As more MMPs are discovered the above classifier can be trained to improve the accuracy of results.

References

1. www.Health-cares.net
2. PE Vanden Steen, B Dubois, I Nelissen, PM Rudd, RA Dwek, G Opdenakker, Biochemistry and molecular biology of gelatinase B or matrix metalloproteinase-9 (MMP-9). Crit Rev Biochem Mol Biol. 37 (2002) 376-536.
3. N Haseneen, G Vaday, S Zucker, HD Foda, Mechanical stretch induces MMP-2 release and activation in lung endothelium: role of EMMPRIN. Am J Physio Lung Cell Mol Physiol. 165 (2003) 541-L547.
4. GA Rosenberg, EY Estrada, JE Dencoff. Matrix metalloproteinases and TIMPs are associated
5. B Ubois, S Starckx, A Pagenstecher, J Oord, B Arnold, G Opdenakker. Gelatinase B deficiency protects against endotoxin shock. Eur J Immunol. 32 (2002) 2163-2171.
6. Z Liu, JM Shipley, X Zhou, LA Diaz, Z Werb, RM Senior. Gelatinase B-deficient mice are resistant to experimental bullous pemphigoid. J Exp Med. 188 (1998) 475-482.
7. TF Golub, Evans, Mc Namara TF, Lee HM, NS Ramamurthy. A non-antimicrobial tetracycline inhibits gingival matrix metalloproteinases and bone resorption. Ann New York Acad Sci. 732 (1994) 96-111.
8. Pang-Ning,Tan.M.Steinbach, V.Kumar. Introduction to Data Mining, 2008.
9. Luna De Ferrari, Aitken Stuart. Mining housekeeping genes with a Naive Bayes. BMC Genomics 7:277 (2006).
10. Amy Sang Qingxiang, Damon A. Douglas. Computational sequence analysis of matrix metalloproteinases. Journal of Protein Chemistry, Springer 15 (2005) 137-160.
11. S Zucker, D Pei, J Cao, Otin C Lopez, Eds. S. Zuker, W-T Chen. Membrane type-matrix metalloproteinases (MT-MMP) in Cell Surface Proteases. Academic Press (2003) 1-74
12. L Yan, S Zucker, B Tooe. Role of the multifunctional glycoprotein, EMMPRIN (Basigin; CD147) in tumor progression. Thrombosis & Haemostasis (2004) in press.
13. K Holmbeck, P Bianco, J Caterina. MT1-MMP deficient mice develop dwarfism, osteopenia, arthritis, and connective tissue disease due to inadequate collagen turnover. Cell. 99 (1999) 81-92.
14. H Nagase, F Woessner. Matrix metalloproteinases. J Biol Chem. 274(1999) 21491-21494.
15. H. Birkedal-Hansen. Proteolytic remodeling of extracellular matrix. Curr Opin Cell Biol. 7 (1995) 728-735.
16. M Egeblad, Z Werb. New functions for matrix metalloproteinases in cancer progression. Nat Rev Cancer 2(3)(2002) 161-74.
17. Pei et al. Cloning and Characterization of Human MMP-23, a New Matrix Metalloproteinases. Am J Physiol Lung Cell Mol Physiol (2000).

Author Information

K.R. Pardasani
Professor and Head, Dept. of Mathematics, MANIT

Bhasker Pant
Research Fellow, Dept. of Bioinformatics, MANIT

Kumud Pant
Research Fellow, Dept. of Bioinformatics, MANIT