IACSIT International Journal of Engineering and Technology, Vol.2, No.3, June 2010
ISSN: 1793-8236

# Multi Class Classification Approach for Classification of ADAMs, MMPs and Their Subclasses

Kumud pant, Neeru Adlakha and Alok Mittal

*Abstract*— **The MMPs and ADAMs are cell surface proteases which belong to metalloprotease family. They play an important role in skin aging, skin disorders, anticancer therapy and other physiological disorders. Thus there arises the need to understand the relationships among various parameters of these proteins for prediction of their classes, structures and functionality. The computational approaches for prediction of their classes are fast and economical therefore can be used to complement the existing wet lab techniques. Realizing their importance, in this paper an attempt has been made to correlate them with their amino acid composition and predict them with fair accuracy. This is a novel method where ADAMs and MMPs have been classified on the basis of amino acid composition using Support Vector Machine. The SVM has been implemented using Lib SVM package. The method discriminates MMP subfamily from ADAM proteases with Matthew's correlation coefficient of 0.98 using amino acid composition. The method is further able to predict three major subclasses or subfamilies of MMPs with an overall Matthew's correlation coefficient [MCC] and accuracy of 0.782 and 89.01% respectively using amino acid composition. The performance of the method was evaluated using 5-fold cross-validation where accuracy of 98% was obtained.**

*Index Terms*— **Metalloproteinases; Support vector machine; Amino Acid composition; Kernel functions..**

## I. Introduction

The MMPs and ADAMs belong to a family of metalloproteinase's [MPs]. Both are zinc proteases responsible for degradation of extracellular matrix proteins and receptors. Zinc proteases are subdivided according to the primary structure of their catalytic sites and include gluzincin, metzincin, inuzincin, carboxypeptidase, and DD carboxypeptidase subgroups [13]. The metzincin subgroup [to which the ADAMs belongs] is further divided into serralysins, astacins, matrixins, and ADAMalysins [13], [15]. The matrixins comprise the matrix metalloproteases, or MMPs and ADAMalysins comprise the ADAMs [16].

MMPs [Matrix Metalloproteinases] are divided into eight subgroups on the basis of their domain organization. They contain predomain, prodomain and catalytic domain which are also shared by others besides hemopexin, transmembrane and cytoplasmic domain [15]. Till now 28 members have been discovered out of which 23 are found in human.

ADAMs are transmembrane MPs that are distinct from the MMPs in that they also have an extracellular disintegrin domain and a cytoplasmic domain that can associate with intracellular proteins. It also is made of metalloproteinase and cysteine-rich, EGF-like domains, followed [in most] by transmembrane and cytoplasmic regions. It is the presence of these two domains that give the ADAMs their name [a disintegrin and metalloprotease] also called as MDC [Metalloprotease, Disintegrin, Cysteine-rich] proteins. Till now 34 different types of ADAMs have been discovered in various organisms. The ADAMalysins subfamily also contains the class III snake venom metalloproteases and the ADAM-TS proteases, which although similar to the ADAMs, can be distinguished structurally [15].

Studies on MMPs began with the discovery of collagenase in the tadpole tail in 1962, and now it has become a vast research field. This family of enzymes plays a central role in the pathological degradation and physiological turnover of the extracellular matrix [ECM] and other proteins. The ADAM family is a recently developed gene family, and the field is growing rapidly. These proteins are involved in important pathobiological events such as processing of growth factors and cytokines, ECM degradation and shedding of membrane proteins. The MMPs and ADAMs share the substrate specificity and their expression pattern under certain conditions, but their mutual roles are not well understood [17].

Despite emerging implications for ADAMs and matrix metalloproteinase's [MMPs] in disease progression, the mechanisms that lead to activation of specific ADAMs and MMPs and their actions in various diseases is still incompletely understood. These enzymes are the principle agents responsible for extracellular matrix degradation and remodeling, and play important roles in development, wound healing, and in the pathology of diseases such as arthritis and cancer [9]. Alzheimer's disease, as well as most of other neurodegenerative disorders is characterized by the deposition of insoluble proteinaceous aggregates. Hence ADAM-dependent proteolytic attack could represent a valuable therapeutic target [3].

The ADAM are a fascinating family of transmembrane and secreted proteins with important roles in regulating cell phenotype via their effects on cell adhesion, migration, proteolysis and signaling. They are involved in diverse processes such as development, cell-cell interactions and protein ectodomain shedding [13]. Recently they have been

Manuscript received December 16, 2009.

K. Pant is with Bioinformatics Department, MANIT, Bhopal, India [e-mail: pant.kumud@gmail.com].

N. Adlakha is with Applied Mathematics Department, SVNIT, Surat, India [e-mail: nad@ashd.svnit.ac.in].

A. Mittal is with Department of Chemistry, MANIT, Bhopal, India [email: aljymittal@yahoo.co.in].

found to play an important role in various types of cancer [2].

Similarly MMPs are found to be a key player in skin aging and many physiological disorders. In animal body they have role in cancer, neurological disorders, rheumatoid, osteoarthritis etc. to name a few. Synthetic inhibitors have recently been developed for MMPs, so as to control cartilage matrix degradation by them in arthritis and osteoporosis [11].

Some attempts for the understanding of structure and function of the ADAM and MMP family of proteins have been made within the past five years. These results have demonstrated the importance of these proteins in diverse biological processes. Studies have also raised many interesting questions that remain to be answered. Questions concerning substrate specificities of these ADAMs and MMPs, the physiological regulators that activate or inhibit these proteases the regulation of the protease, adhesion and signaling activities of the ADAMs and MMPs in response to developmental, physiological and pathological stimuli etc. remain unanswered [17].

The identification of novel type of cell surface proteases and their cognate substrate is the major focus of pharmaceutical companies. Hence, highly accurate identification of protease types will solve the problem of efficacy and side effects of various drugs. Currently, efforts are underway to develop new therapeutic agents and elucidation of metabolic pathway associated with diseases. Moreover, the mere understanding of different types of ADAM and MMPs and their substrate-binding properties will assist in finding novel drug target with minimum side effects. The experimental attempts are reported in the literature for functional classification of MMPs and ADAMs using the structure of their catalytic sites [13]. But no computational technique is available in the literature for classification of MMPs and ADAMs based parameter like amino acid composition. Since the experimental identifications of them are labor and cost-intensive task, the computational biology can provide a better alternative to develop a method for classifying different enzymes of each.

In view of the above an attempt has been made in this paper to develop a computational approach for predicting and classifying two types of cell surface proteases, ADAMs and MMPs. The classification of MMPs has been extended further where they have been divided into 4 subclasses. In the first step, binary classification method has been adopted where the metalloproteinases can be discriminated as MMPs or ADAMs. In the second step a multiclass classification approach has been used where MMPs have been divided further into 4 subfamilies.

It has been shown in past that SVM is an elegant technique for the classification of biological data [1], [10], [4]-[7]. Here the same SVM model has been developed for amino acid composition based prediction identification and classification of cell surface proteases MMPs and ADAMs.

This paper is a step in the direction where machine learning and computational biology techniques can be used to compliment existing wet lab techniques.

## II. Materials And Methods

### A. Recognition of MMPs from ADAMs

Initially, we developed an SVM module for identifying MMPs from ADAM proteins. The dataset was derived from uniprotKB of Expasy server [12]. The final dataset consisted of 187 proteins belonging to both MMPs and ADAMs subclass of metalloproteinase's family [MPs].

To validate our methodology 178 globular proteins belonging to various enzyme classes other than hydrolase [E.C. EC 3.4.24] to which metalloproteinase's belong were taken into consideration. They were treated as negative instances.

The performance of the module was evaluated using a 5-fold cross-validation test. The SVM was trained with a fixed-dimensions vector [20] obtained on the basis of the amino acid composition of protein sequences

### B. Recognition of subfamilies of MMPs

Prediction of subfamilies of matrix metalloproteinase's is a multi-class classification problem. In this case, the three dataset belonging to three subfamilies and one dataset including members of rest of the matrix metalloproteinase's were taken for classification. To handle this multi-class situation, we designed a series of binary SVMs. For N class classification, N SVMs were constructed. The ith SVM was trained with all samples of the ith subfamily being labeled as positive, and the samples of all other subfamilies being labeled as negative. The SVMs trained in this way were referred to as 1-v-r SVMs. In this classification approach, each of the unknown proteins will achieve four scores. An unknown protein will be classified into the subfamily that corresponds to the 1-v-r SVM with the highest output score

The dataset consisted of all MMPs proteins family members extracted from swissprot/uniprot data bank of expasy server [12]. All the entries marked as fragments were not included in the dataset. Since the number of sequences belonging to many subclasses of MMPs are too less to be of any statistical significance therefore they were combined as belonging to one. The various subclasses of MMPs on the basis of domain organization are depicted in figure1 below.

The data set of MMPs consisted of A] Gelatin binding MMPs B] Simple hemopexin domain containing MMPs C] Transmembrane MMPs D] Others, which include members of rest of the classes which were too less in number to be examined individually. The number of sequences belonging to various subclasses used for training the classifier is mentioned in table 1. The test dataset included 11, 12, 17 and 15 instances belonging to A, B, C, and D subfamily other then those taken for training. Only non redundant sequences were taken for our study. Any two sequences with more then 90% sequence similarity were excluded from inclusion in the training data set since they can cause over training of the system

The number of instances for training was 85, whereas for testing other instances were used which also non fragmented non repeated entries also obtained from Expasy [12]. Entries marked as predicted or putative were also verified on testing as belonging to that class which explains the utility of the model and accuracy of amino acid composition in

differentiating proteins.

| Name of subclass | Member MMPs | Year of discovery |
|---|---|---|
| Minimal domain MMPs | MMP-7/ Matrilysin | 1980, 1988 |
| | MMP-26/ Endometase | 2001 |
| Simple hemopexin domain containing | MMP-1/ Collagenase-1 | 1962, 1986 |
| | MMP-8/ Collagenase-2 | 1968, 1990 |
| | MMP-13/ Collagenase-3 | 1994 |
| | MMP-18/ Collagenase-4 | 1996 |
| | MMP-3/ Stromelysin-1 | 1974,1985 |
| | MMP-10/ Stromelysin-2 | 1988 |
| | MMP-12/ Metalloelastase | 1981, 1992 |
| | MMP-19/ RASI-1 | 1996 |
| | MMP-20/ Enamelysin | 1997 |
| | MMP-22/ CMMP | 1998 |
| | MMP-27 | 2001 |
| Gelatin-binding MMPs | MMP2/ Gelatinase A | 1978, 1988 |
| | MMP9/ Gelatinase B | 1972, 1989 |
| Furin-activated secreted MMPs | MMP11/ Stromelysin-3 | 1990 |
| | MMP28/ Epilysin | 2001 |
| Transmembrane MMPs | MMP14/ MT1- MMP | 1994 |
| | MMP15/ MT2- MMP | 1995 |
| GPI-linked MMPs | MMP17/ MT4- MMP | 1996 |
| | MMP25/ MT6- MMP | 1999 |
| Vitronectin- like insect-less MMPs | MMP21/ XMMP | 1998 |
| Cysteine/ Proline-rich IL-1 receptor like domain MMPs | MMP23 | 1998 |

Figure 1. History and domain knowledge of MMPs with their history [20].

The data set of MMPs consisted of A] Gelatin binding MMPs B] Simple hemopexin domain containing MMPs C] Transmembrane MMPs D] Others, which include members of rest of the classes which were too less in number to be examined individually. The number of sequences belonging to various subclasses used for training the classifier is mentioned in table 1. The test dataset included 11, 12, 17 and 15 instances belonging to A, B, C, and D subfamily other then those taken for training. Only non redundant sequences were taken for our study. Any two sequences with more then 90% sequence similarity were excluded from inclusion in the training data set since they can cause over training of the system

TABLE I. THE NUMBER OF SEQUENCES BELONGING TO MATRIX METALLOPROTEINASE'S FAMILY USED FOR TRAINING THE CLASSIFIER

| Name of subfamily | Number of sequences |
|---|---|
| Gelatin binding domain MMPs | 12 |
| Simple hemopexin domain containing MMPs | 36 |
| Transmembrane MMPs | 15 |
| Others | 22 |

### C. Support vector machine [Binary classification]

Kernel-based techniques [such as support vector machines, Bayes point machines, kernel principal component analysis, and Gaussian processes] represent a major development in machine learning algorithms. Support vector machines [SVM] are a group of supervised learning methods that can be applied to classification or regression. In a short period of time, SVM found numerous applications in Bioinformatics.

SVM is a supervised machine learning method which is based on the statistical learning theory [18], [19]. When used as a binary classifier, an SVM will construct a hyperplane, which acts as the decision surface between the two classes. This is achieved by maximizing the margin of separation between the hyperplane and those points nearest to it.

The SVMs were implemented using freely downloadable software, libSVM [8]. In this software there is a facility to define parameters and choose among various inbuilt kernals. They can be radial basis function [RBF] or a polynomial kernel [of given degree], linear, sigmoid.

### D. SVM software; LIBSVM

Simulations were preformed using LIBSVM version 2.89 [a freely available software package] [8]. For our study RBF Kernel was found to be the best. The SVM training was carried out by the optimization of value of regularization parameter and the value of RBF kernel parameter.

### E. Amino acid composition

Previously, this parameter has been used for predicting the subcellular localization of proteins [10]. The amino acid composition is the fraction of each amino acid type within a protein. The fractions of all 20 natural amino acids were calculated by using Equation 1,

Fraction of amino acid i [where i can be any amino acid]

$$= \frac{\text{Total Number of amino acid } i}{\text{Total number of amino acids in a protein}}$$

### F. Evaluation of Performance

The performance of our classifier was judged by 10 fold cross validation. The LibSVM provides a parameter selection tool using the RBF kernel: cross validation via grid search. A grid search was performed on C and Gamma using an inbuilt module of libSVM tools as shown in figure1. Here pairs of C

and Gamma are tried and the one with the best cross validation accuracy is picked. On using the values of C=2 and Gamma=0.125 obtained through grid search an accuracy of 98.87% was obtained.

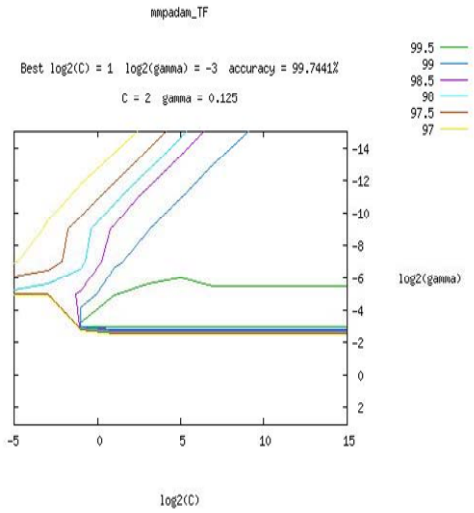


Figure 2. Coarse Grid Search on C = 2-5, 2-4 … 210 and Gamma = 25, 24 … 2 - 10

## III. Prdiction System Assessment

10-fold cross validation was performed to evaluate the accuracy of the classifier where, the data set was partitioned randomly to ten equally sized sets. The training and testing of each classifier was carried out ten times using one different set for testing and the other nine sets for training. Four threshold-dependent parameters, sensitivity, specificity, accuracy and Matthews's correlation coefficient [MCC] [14], were used to measure the performance of this module. Calculations of sensitivity, specificity, accuracy and MCC were carried out as follows:

Sensitivity = [TP/TP + FN] * 100 Eq. 1

Specificity = [T /TN + FP] * 100 Eq. 2

Accuracy = [TP+TN/TP+TN+FP+FN] * 100 Eq. 3

MCC = [TP * TN] – [FP * FN]/{[[TP + FP] * [TP + FN] * [TN + FP] * [TN+FN]]0.5 [TN + FP] * [TN+FN]]0.5 } Eq. 4

Where TP, TN, FP and FN represent true positive, true negative, false positive and false negative respectively

## IV. Results And Discussion

Both MMPs and ADAMs have been implicated in various diseases and a key player in many protein degradation processes. Realizing their implication in above we have chosen these two family of metalloproteases for our study.

The results of classifying MMPs from ADAMs using amino acid composition are given in Table2 and that of classifying subfamily of MMPs are given in Table3.

TABLE II. Performance Of Mmps And Adams Recognition

| S. No. | Protein sub family | MCC | Accuracy | Sensitivity | Recall | Precision |
|---|---|---|---|---|---|---|
| 1] | MMPs | 0.97 | 98.98% | 0.99 | 98.96 % | 98.96 % |
| 2] | ADAMs | 0.97 | 98.88% | 0.98 | 98.87 % | 98.87 % |

The results obtained here will be helpful in differentiating between different metalloproteinases i.e. whether MMPs or ADAMs. A new protein discovered can be shown to either belonging to ADAMs or MMP subfamily of proteins.

Our results clearly highlight the importance of amino acid composition in differentiating between these families. This model can also be an important tool to understand the differences between MMP and ADAMs hence a step towards assisting various wet lab techniques in devising novel drugs and therapeutic agents against these two. The correlation of MMPs and ADAMs with their amino acid composition explored here can be useful to obtain better insight about these proteins. Their molecular and physiological roles along with the substrate affinity can also be correlated with amino acid composition.

The overall accuracy and MCC of the amino acid composition-based classifier for classifying the two subfamilies of metalloproteinases was 98% and 0.97, respectively. It proved that metalloproteinases can be correlated with amino acid composition and can be easily distinguished on this basis.

The receiver operating characteristics [ROC] score was usually used as the primary measure of the machine learning method performance and provided an overview of the possible cut-off levels in the test performance. The roc curves for both MMPs and ADAMs are depicted in Figure3 which shows that majority of instances fall in the true positive range

TABLE III. Performance of MMPs subfamily recognition using SVM

| Subfamily | Sensitivity | Specificity | Accuracy | MCC |
|---|---|---|---|---|
| Gelatin binding domain MMPs | 66.66% | 100% | 85% | 0.7237 |
| Simple hemopexin domain containing MMPs | 72.73% | 100% | 91.17% | 0.802 |
| Transmembrane MMPs | 71.42% | 90.90% | 83.87% | 0.681 |
| Others | 92.85% | 100% | 96% | 0.922 S |

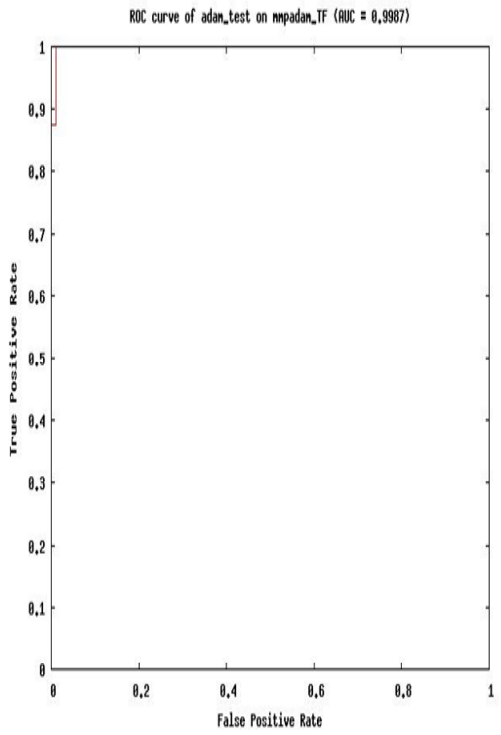


Figure 3. ROC curve for both MMP [above] and ADAM [below] proteases depicting FP and TP

## V. Conclusion

The SVM model developed here is computationally efficient and effective in predicting and classifying the MMPs and ADAMs. This is evident from the accuracy [98%] in the results. Further the amino acid composition contains very significant information for discriminating the classes of above proteins.

This model can be used to analyze other enzymes, such as entire proteomics data. Such type of prediction systems can be very useful for understanding the above proteases in a better way so as in conclusion, a novel method for classifying MMPs and ADAMs is presented.

This method will nicely complement the existing wet lab methods. It will assist in assigning the correct class to which these proteins belong or classifying them as either of the two subclasses. The prediction method presented here may be useful for the annotation of the piled-up proteomic data.

This model can also be an important tool to understand the differences between various subfamilies of MMPs and hence a step towards assisting various wet lab techniques in devising novel drugs and therapeutic agents against these two. Here we are attempting as well to correlate MMPs with their amino acid composition which can also help to better our understanding about these proteins. Their molecular and physiological roles along with the substrate affinity can also be correlated with amino acid composition. It is well established that machine learning methods require large number of examples for reliable prediction therefore due to less number of sequences we combined subclasses other then A, B, and C as one into D.

The author awaits discovery of more of these proteins in the future so that accuracy of the prediction model can be increased further and a server developed for public use.

## Acknowledgment

The authors are highly thankful to the Department of Biotechnology, Delhi, India and M.P. Council of Science and Technology M.P., Bhopal, India for providing support in the form of Bioinfomatics infrastructure facility to carry out this work.

## References

[1] M. Bhasin and G.P. Raghava, “SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence,” Bioinformatics, vol. 20, pp. 421–423, 2004.
[2] C.P. Blobel, Cell, vol. 90, pp. 589, 1999.
[3] B. Vincent, “ADAM Proteases: Protective Role in Alzheimer's and Prion Diseases,” Current Alzheimer Research, vol. 10, pp. 165-174, 2006.
[4] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, “Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect,” J. Cell. Biochem, vol. 84, pp. 343–348, 2002.
[5] Y.D. Cai, R. Pong-Wong, K. Feng, J.C.H. Jen, K.C. Chou, “Application of SVM to predict membrane protein types,” J. Theor. Biol, vol. 226, pp. 373–376, 2004.
[6] Y.D. Cai, G.P. Zhou, K.C. Chou, “Support vector machines for predicting membrane protein types by using functional domain composition,” Biophys. J, vol. 84, pp. 3257–3263, 2003.
[7] Y.D. Cai, G.P. Zhou, C.H. Jen, S.L. Lin, K.C. Chou, “Identify catalytic triads of serine hydrolases by support vector machines,” J. Theor. Biol, vol. 228, pp. 551–557, 2004.
[8] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines. Software,” available at http://www.csie.ntu.edu.tw/~cjlin/libSVM, 2001.
[9] C. Chang and Z. Werb,” The many faces of metalloproteases: cell growth, invasion, angiogenesis and metastasis,” Am J Pathol, S37-S43, 2001.
[10] K.C. Chou, Y.D. Cai, “Using functional domain composition and support vector machines for prediction of protein subcellular location,” J. Biol. Chem, vol. 277, 45765–45769, 2002.
[11] S.1 Elliott, T. Cawston, “The Clinical Potential of Matrix Metalloproteinase Inhibitors in the Rheumatic Disorders authors, Drugs & Aging, vol. 18, pp. 87-99[13], 2001.
[12] ExPASy Proteomics Server, www.expasy.org/.
[13] N.M. Hooper, “Families of zinc metalloproteases,” FEBS Lett, vol. 354, pp. 1–6, 1994.
[14] S. Hua, and Z. Sun, Bioinformatics, vol. 17, pp. 721-728, 2001.
[15] Minireviews “ADAMs”, R&D Systems' 2001 Catalog.
[16] W. Stocker, F. Grams, U. Baumann, P. Reinemer, F.X. Gomis-Ruth, D.B. McKay, W. Bode, “The metzincins–topological and sequential relations between the astacins, ADAMalysins, serralysins, and matrixins [collagenases] define a superfamily of zinc-peptidases,” Protein Sci, vol. 4, pp. 823–840, 1995.
[17] Tenth Keio University International Symposium for Life Sciences and Medicine on October 17-19, 2001.
[18] V. Vapnik, “The nature of statistical learning theory,” Springer, 1995.
[19] M. Wang, J. Yang, G.P. Liu, Z.J. Xu, K.C. Chou, “ Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition,” Protein Eng. Des. Sel, vol. 17, pp. 509–516.
[20] J. Cao, M. D. and Stanley Zucker., “Biology and Chemistry of matrix metalloproteinases [MMPs],” Abcam: Resources, [2009].

**Biographical notes:**
**K. Pant** received her M. Sc. degree in Botechnology from Jiwaji University Gwalior, M. P., India. She is currently pursuing Ph.D degree in Department of Bioinformatics at Maulana Azad National Institute of Technology [MANIT], Bhopal, India.

**N. Adlakha i**s currently working as Associate Professor in Department of Applied Mathematics & Humanities, at Sardar Vallabhbhai National Institute of Technology [SVNIT], Surat, India. He has more then 30 research publications in both National and International Journals. A keen researcher and able administrator with multifaceted personality.

**A. Mittal** is currently working as Professor in Department of Chemistry, at Maulana Azad National Institute of Technology [MANIT], Bhopal, India. He has more then 50 research publications in both National and International Journals. A keen researcher and able administrator with multifaceted personality.