# *In-Silico* approach to phylogentic analysis of differentially expressed protein Matrix Metalloproteinase proteinase -12 associated with Lung Cancer

Bhagavathi S. [1], Gulshan Wadhwa [2] Anil Prakash[3]

[1][3] *Department of Biotechnology & Microbiology, Barkatullah University, Bhopal*
[2] *Dept. of Biotechnology, Apex Bioinformatics Centre, Ministry of Science and Technology, Government of India, Block-2, 7th Floor, C.G.O. Complex Lodhi Road, New Delhi-110003*
*Correspondence: S.BHAGAVATHI*

**Abstract**

Phylogenies ; The evolutionary histories of groups of species are one of the most widely used tools throughout the life sciences, as well as objects of research with in systematic, evolutionary biology. In every phylogenetic analysis reconstruction produces trees. These trees represent the evolutionary histories of many groups of organisms, bacteria due to horizontal gene transfer and plants due to process of hybridization. In this paper a model has been employed to reconstruct phylogenetic network in Matrix metalloproteinase-12 gene (MMP12). Through comparison with several species of healthy organism, one may determine where the defective mutation is located, and then determine how to treat the disease. For this purpose, we have taken up the strategy of phylogenetics of genes causable for lung cancer with bioinformatics approach. This strategy may help us to identify the mutations that had occurred in evolutionary conserved residues. We feel this method can be useful for understanding evolutionary rate variation, and for understanding selection variation on different characters. To better understand the roles that MMPs play today in development and disease, it is helpful to understand their historical functions and their evolution.

Keywords- Phylogeny, mutations, species, MMP12

## I. INTRODUCTION

PHYLOGENIES are the main tool for representing evolutionary relationships among biological entities. The biologists, mathematicians, and computer scientists are working to design a variety of methods for their reconstruction. Almost all such methods, however construct trees; yet scientists have long recognized that trees oversimplify our view of evolution science. Comparisons of related protein and nucleotide sequences have facilitated many recent advances in understanding the information content and function of genetic sequence. Statistical phylogenetics is computationally intensive, resulting in considerable attention meted on techniques. The startling recent advances in sequencing technology or fuelling a concomitant increase in the scale and ambition of phylogenetic analysis, however this enthusiasm belies and fundamental limitation in statistical phylogenetics as the number of sequence increases, the size of parameter space specifically the number of possible phytogenetic histories explodes. Many biological analyses involve the construction of a phylogenetic tree for some set of sequence data, and a variety of methods for inferring phylogenies are available. However, the choice of phylogenetic method can have a strong influence on the tree obtained for a given set of sequence data, both in terms of its topology and branch lengths. In addition, different gene trees can be obtained for some fixed set of species, where each gene tree is based on a different set of orthologous sequences chosen for the analysis. Comparison between gene trees and species trees can reveal concensus patterns of evolution as well as genes that diverge from this pattern. Recent large-scale studies of individuals within a population have demonstrated that there is widespread variation in copy number in many gene families. In addition, there is increasing evidence that the variation in gene copy number can give rise to substantial phenotypic effects. In some cases, these variations have been shown to be adaptive. These observations show that a full understanding of the evolution of biological function requires an

understanding of gene gain and gene loss. Accurate, robust evolutionary models of gain and loss events are, therefore, required. **[1]**.

**Disease importance of Metallo proteinase domain 12**

**Metallo proteinase domain 12** is a family of proteolytic enzymes that break down proteins in the extracellular matrix. Matrixmetalloproteinases (MMP's) facilitate cellular invasion by degrading the extracellular matrix, and their regulation is partially dependent on transcription **[2]**. MMPs are regulated by specific inhibitors known as the tissue inhibitors of metallo proteinases (TIMPs). These are highly expressed in many different cancer types including lung cancer. The importance of MMP12 in human lung has however been shown by its role in the development of emphysema caused by cigarette smoke.

There is considerable evidence that MMPs in the lung play a role in host defense **[3] [4].** In this article we present evidence from the literature that supports a direct or an indirect role for several of the MMP family members in various lung diseases.

**II. Materials & Methods**

**A. Sequence alignment**

Most of the sampled gene regions were protein-coding, making alignment relatively straightforward. DNA sequences were translated to amino acids using MacClade version 4.0 **[5] [6]** and any gaps were placed so as to maintain the integrity of codon triplets and the alignment of amino acids. Alignment of MMP12 genes was more challenging. Most programs in the PHYLIP package require a set of aligned sequences and for our target gene , sequence similarity search of nucleotide sequences were performed by WU-BLAST(WEBSITE: [**7]** WU-BLAST 2 stands for Washington University Basic local alignment search tool version 2.0. The emphasis of this tool is to find regions of sequence similarity quickly with minimum loss of sensitivity. This will yield functional and evolutionary clues about the structure and function of the novel query sequence.

**Methods:**

**B. Selecting the Appropriate sequences**

A basic assumption of all phylogenetic analyses is that orthologous genes are being compared. This may seem obvious but genes subject to horizontal transfer or orphan genes, for e.g. will produce spurious results they are subjected to different evolutionary constraints from the ancestral genes.

Out groups provide a reference used to measure distances and help determine the root of a tree when a actual ancestral sequence is not available. An out group is the closest relative that does not belong to the group under study. For example, to build a tree of mammalian sequences, a bird sequence may provide a suitable out group. In this case a plant sequence would be a poor choice, because plants are very distant relatives and would degrade the alignments and the distance estimations.

***DNA or Protein:*** When too many mutations accumulate sequences become saturated with mutations. Consider a position in State A mutating to state B. As more mutations occur the chances increase that it mutates to a third state C or back to state A, making us underestimate the number of mutations.

Apart from the redundancy in the genetic code, the protein is usually the functional product of the gene and the preservation of the protein function is a driving force for sequence conservation. Protein sequence therefore change much more slowly than DNA sequences and are first choice for studying different relationship of genes that changes very rapidly.

In some cases, when a gene changes very slowly or when very close relationships are being examined, when the genes are very small, the peptide sequence may not contain enough information to resolve trees and DNA sequences may be a better choice. So we chose only DNA sequences.

**III. RESULTS**

**A. PHYLOGENETIC ANALYSES**

Our primary approach for integrating data from multiple species was to combine these data into a single matrix, in various combinations . Combined analysis for MMP12 genes is uncontroversial, because they are linked and therefore share a single phylogenetic history. The phylogenetic history of the genetically linked MMP12 genes may differ from that of the species phylogeny for these same

reasons. However, we expect that problems of discordance between gene and species trees will generally be phylogenetically localized and involve a limited number of species when they do occur, in which case combined analysis of multiple species should yield the correct answer. Nevertheless under some circumstances many genes may converge on an incorrect answer and mislead a combined analysis and attention may be common on short branches . To deal with such circumstances, we applied species-tree methods (i.e., BEST) to the nuclear data. Fortunately, our results showed much congruence between the trees from combined species , and BEST analysis of nucDNA data, which bolsters our confidence in the idea that all of these trees generally reflect the species phylogeny.

Data were analyzed using parsimony, Bayesian, and likelihood methods. Data from different species were generally combined (concatenated).Nucleotide sequences had been accessed from NCBI'S website [**8]** and sequence similarity search sequences were used for phylogenetic analyses. Phylogenetic analysis was performed with PHYLIP package **[9]** and the result was further verified by MEGA4 software [**10].** The top 11 sequences with high sequence similarity were chosen for further steps. The list of the top 11 sequences are Humans (NM_002426.4; *Homosapiens*), Rodents Mouse (NM_008605.3; Mus musculus), Cow (NM_001206640.1; Bos Taurus), Rat ( NM_053963.2 ; *Rattus Norvegicus*) Rabbit (NM_00108277.1; *Oryctolagus cuniculus*), Pig(NM_001099938.1; *Susscrofa),* Gorilla (DQ482230.1; Troglodytes) , Cat (JP013446.1; *Mustela putorius*), Dog(XM_849501.2 ; *Canis familiaris*), Chimpanzee ( XM_508724.3*; Pantroglodytes*), Orangutan ( DQ482231.1;*Pongo abelii*). The details are given in the table.

| INDEX | ORGANISM | SCIENTIFIC NAME | GENERIC ID | REF NO. |
|---|---|---|---|---|
| 1. | Human | *Homo sapiens* | 261878521 | NM_002426.4 |
| 2. | Rodents-Mouse | *Mus musculus* | 115392137 | NM_008605.3 |
| 3. | Cow | *Bos Taurus* | 331284140 | NM_001206640.1 |
| 4. | Rat | *Rattus norvegicus* | 291327536 | NM_0536963.2 |
| 5. | Rabbit | *Oryctolagus cuniculus* | 130487231 | NM_00108277.1 |
| 6. | Pig | *Sus scrofa* | 153791307 | NM_ 001099938.1 |
| 7. | Gorilla | *Troglodytes* | 94961199 | DQ482230.1 |
| 8. | Cat | *Mustela putorius* | 355702773 | JP013446.1 |
| 9. | Dog | *Canis lupus familiaris* | 345799780 | XM_849501.2 |
| 10. | Chimpanzee | *Pan troglodytes* | 332837594 | XM_508724.3 |
| 11. | Orangutan | *Pongo abelii* | 94961200 | DQ482231.1 |

**B. PHYLOGENY CONSTRUCTION**

With the PHYLIP package, we had analyzed 11 sequences of highest similarity and inferred their phylogenetic relationship with respect to matrix metalloproteinase gene associated with lung cancer. With the protpars program unrooted tree was obtained. Fig.1 and rooted tree dendrogram tree was also obtained as shown in Fig.2.

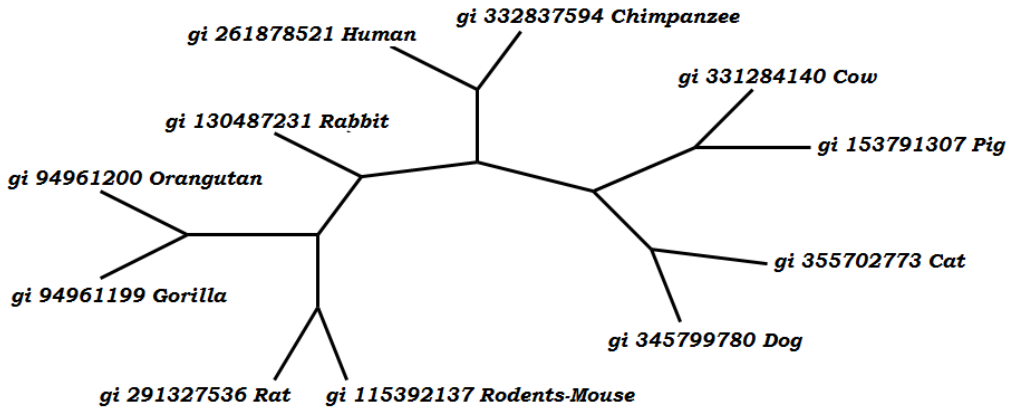


**Fig.1. UNROOTED TREE**

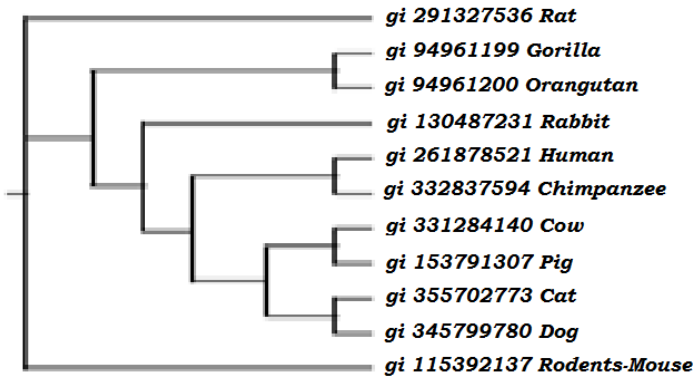


**Fig. 2. Rooted dendrogram**

The results produced by the phylip software reveals different distinct groups based on relative branch lengths as in Fig. 2. The comparison of nucleotides sequences strictly separated the rat, mouse which might be due to the deletion mutation at genetic level. The top organisms in the gene data base reveals five distinct groups based on their relative branch lengths. One group included the Mouse (*Mus musculus*), Second group include the Human ( *Homo sapiens* ), Third group includes the Rat (*Rattus norvegicus*), the fourth and fifth contained the even toed ungulates i.e., Cow ( *Bos Taurus*) and Pig (Sus scrofa) and other taxa which include primates.

The MMPs belong to the metzincin superfamily that includes enzymes with similar metalloproteinase domains such as astacins, ADAMs, and ADAM-TS proteinases **[11]**.There is little consensus as to how the MMPs should be grouped , and different experts in the field classify them based on their structural similarities, substrate specificity, or tissue expression.

The comparison of amino acids in MMP12 sequences strictly separated the Rat from the mouse but that doesn't necessarily mean that the protein function is different from these two species **[12]** it might be due to the deletion mutations at the genetic level. A functional region might likely have a lower evolutionary constraint as compared to other region which could be argued on the basis, that most of the mutations are neutral. On the other hand two less divergent taxonomic group might have a fewer substitutions but these substitutions might occur in functionally important region. So the observed divergence between rodants and non-rodants probably reflected the differences in pattern of the expression between the two groups. The determination of the rodents branching point enables us to root the tree with non eutherian group which includes Gorilla (troglodytes).

So far from the phylogenetic point of view we can conclude that *Rattus norvegicus* and *Mus musculus* appears to be closer link to all other primates and non primates.

Next we assume to determine the unexpected Rabbit – Rodent divergence which is very similar to Human-Chimpanzee having a typical pseudo stratified surface epithelium **[13]** in which ciliated epithelial cells were the abundant cell types **[14]**.

From the phylogenetic point of view we can conclude that Rabbit appears to be closer to primates than rodents **[15]**

The comparison of Human and chimpanzee in the unrooted tree fig. 5 could be based upon the previous studies of Human chimpanzee transcription comparison which showed 39% of silent sites in Protein coding regions were under purifying selection **[16]** and that of humans the average substitution rates of silent sites was 30% lower in functional genes **[17]** than in pseudo genes, Pseudogenes are dysfunctional relatives of genes that have lost their protein-coding ability or are otherwise no longer expressed in the cell **[18]**. Also, the experimental evidence showed that the percentage of synonymous substitutions involved was comparable to that suggested to be under purifying selection in comparative inter and intra species studies **[19]**.

Further we had made an assumption that Primate Cercodactyla i.e., Pig & Cow CLADE with 97% which is grouped with a single branch.

Next we assume to determine the carnivores i.e., Cat & Dog are more closely related to the even toed ungulates i.e. Pig & Cow as shown in the fig.5 the unrooted tree.

So, therefore overall it can be concluded that there is a slight divergence of sequences in Primates & Non- Primate species. Thus it can be postulated that the remaining homologies were under a strong pressure and therefore had a critical role in the severity of disease development.

## IV. DISCUSSION

A phylogenetic tree is a leaf-labeled tree that models the evolution of a set of a taxa (species, genes, languages, placed at the leaves) from their most recent common ancestor (placed at the root). The internal nodes of the tree correspond to the speciation events. Many algorithms have been designed for the inference of phylogenetic trees, mainly from biomolecular (DNA, RNA, or amino-acid)sequences.

The matrix metalloproteinases (MMPs), a family of 25 secreted and cell surface-bound neutral proteinases, process a large array of extracellular and cell surface proteins under normal and pathological conditions. MMPs play critical roles in lung organogenesis, but their expression, for the most part, is down regulated after generation of the alveoli **[20].**

**Matrix metalloproteinase-12** (MMP-12) also known as **macrophage metalloelastase** (MME) or **macrophage elastase** (ME) is an enzyme that in humans is encoded by the *MMP12* gene. **[21]**. The matrix metalloproteinases (MMPs) consist of 24 known human zinc proteases with essential roles in breaking down components of the extracellular matrix (ECM). MMPs play critical roles in lung organogenesis, but their expression, for the most part, is down regulated after generation of the alveoli. Our knowledge about the resurgence of the MMPs that occurs in most inflammatory diseases of the lung is rapidly expanding. Although not all members of the MMP family are found within the lung tissue, many are upregulated during the acute and chronic phases of these diseases **[22]** .In consideration with phylogenetics we compare MMPs from lower phylogenetic ranks of species, e.g., invertebrates, with those of higher orders, e.g., mammals, a comparison that may allow lessons from a simple system to provide insights into the most complex biological functions of this family of enzymes. **[23]**. MMPs are widely distributed across phylogenies, but their structures are fairly conserved. Members of the MMP family have been found in soybean **[24] [25] [26]**.In invertebrates, MMP-like activity has been noted in *Balanus amphitrite* barnacle larvae **[27]**.To better understand the roles that MMPs play today in development and disease, it is helpful to understand their historical functions and their function in other organisms.

Members of the MMP family were originally identified by descriptive names that were assigned based on limited knowledge of their preferred substrate specificities, e.g., collagenases, gelatinases, stromelysins, and matrilysins **[28]**. A

sequential numbering system, loosely based on the order of discovery, was adopted when it became clear that more MMPs exist than was previously suspected and that the names based on substrate specificity type were inadequate.

The MMPs belong to the metzincin superfamily that includes enzymes with similar metalloproteinase domains such as astacins, ADAMs, and ADAM-TS proteinases . There is little consensus as to how the MMPs should be grouped, and different experts in the field classify them based on their structural similarities, substrate specificity, or tissue expression **[29]**.

Phylogenetic analyses were performed on the metalloprotease genes for each gene family. The accession numbers for protein sequences used in these studies are presented in Table 1. The genes identified were aligned with each gene family using CLUSTALX **[30]**

A preliminary bootstrapped Neighbour-joining tree was drawn using CLUSTALX and the sequences were then divided into sub-groups based on their position in the tree. For each sub-group, new multiple alignments were created, gap- containing sites were removed and four independent phylogenetic methods were performed. Neighbour joining trees and bootstrap replicates were generated using SEQBOOT, PROTDIST, NEIGHBOR and CONSENSE from the PHYLIP package using the default settings **[31]**. Maximum Parsimony trees and bootstrap replicates were obtained using SEQBOOT, PROTPARS and CONSENSE and Maximum Likelihood trees were inferred using PROML from the PHYLIP package using the default settings. Bayesian tree inference values were produced from the MrBayes programme **[32]** where Markov Chain Monte Carlo analysis was performed for all the 12 generations using all the programs.

Now that the presence of MMPs has been unequivocally established in lower phyla, it is time to focus on the structure-function relationship of these MMPs and utilize new model systems for investigating mechanisms for MMP action. To better understand the roles that MMPs play today in development and disease, it is helpful to understand their historical functions and their function in other organisms through phylogenetics.

## [6] Conclusion

The above work is an *in-silico* work; this work can serve as an predicted model and can be useful to understand their historical functions and their function in other organisms through phylogenetics against Lung Cancer. The *in-silico* approach helps researchers by giving them an in-hand idea so that they can happily advance towards the treatment of the disease.

## References

**[1] M . Ryan. Ames, Daniel Money, P. Vikramsinh Ghatge , Simon Whelan and C. Simon Lovell.** Determining the evolutionary history of gene families. *Bioinformatics* pp. 28(1): 48-55, 2012

**[2] L. Joni Rutter, I. Teresa, Mitchell, Giovanna Buttici', Jennifer Meyers, F . James, J .Gusella, Laurie, Ozelius and E. Constance, Brinckerhoff .** *Cancer research* pp. 58, 5321-5325, 1998.

**[3] W.C. Parks, S.D .Shapiro.** Matrix metalloproteinases in lung biology. *Respir Res* **pp**.2: 10–19, 2001.

**[4] J.J Atkinson, R.M. Senior.** Matrix metalloproteinase-9 in lung remodeling. *Am J Respir Cell Mol Biol* pp.28: 12–24, 2003.

**[5] D.R. Maddison, W.P**..**Maddison,** MacClade 4.0. Sinauer Associates,Sunderland, MA. 2000.

**[6] W.P . Maddison.**. Gene trees in species trees. *Syst. Biol*.pp. 46, 523–536.1997.

**[7]** http://www.ebi.ac.uk/Tools/blast2/index.html
**[8]** http://www.ncbi.nlm.nih.gov

**[9]** http://evolution.genetics.washington.edu/phylip.htm

**[10]** http://www.megasoftware.net

**[11] M.D . Sternlicht, Z.Werb** How matrix metalloproteinases regulate cell behavior. *Annu Rev Cell Dev Biol* pp.17: 463–516, 2001

[12] **Franco Pagani** *Genome.Res.* **pp 13, 831-837, 2005.**

**[13] C.G Plopper, A.T.Mariassy , D.N. Wilson, J.L .Alley, S.J. Nishio and P .Nerresheim..** Comparison of non-ciliated tracheal epithelial cells in six mammalian species : Ultra structure and population densities . *Exp.lung. Res.* pp 5, 281-294, 1983.

**[14] P.K. Jeffrey,** Morphological features of airway surfaces epithelial cells and glands. Am. *Rev. Respir. Dis*. Pp. 128, S14-S20, 1983.

**[15] D.L Graur Duret and M. Gouy. Phylogenetic position of the order Lagomorpha (rabbits. Hares and allies***). Nature pp* 739:333-335, 1996.

**[16] I. Hellmann, S. Zollner, W. Enard, I. Ebersberger , B .Nickel and S. Paabo.** *Genome . Res.* pp .13, 831-837, 2003.

**[17] C.D Bustamante, R .Neilson, & D.L Harlt.** *Mol.Biol.Evol*.pp .19,110-117,2002.

**[18] E.F. Vanin (1985). "Processed pseudogenes:** characteristics and evolution". *Annu. Rev. Genet.* Pp. 19: 253–72.1985.

**[19] E. Buratti, T . Dork, E . Zuccato, F. Pagani , M. Romano and F.E .Baranello,**.*EMBO J*. pp .20,1774-1784, 2001.

**[20] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis,[a]Phylogenetic Inference,** *Molecular Systematics***,** *eds., pp. 407-514, Sinauer Assoc.*, 1996.

**[21] S.D. Shapiro , D.K .Kobayashi, T.J .Ley.** Cloning and characterization of a unique elastolytic metalloproteinase produced by human alveolar macrophages. *J. Biol. Chem.* Pp .**268** (32): 23824–9.1993.

**[22] J. Kendra , Greenlee, Zena werb and Farrah Kheradmand** *physiol rev pp*.87: 69–98, 2007.

**[23] M.D Sternlicht, Z . Werb.** ECM. Proteinases. In: *Guidebook to the Extracellular Matrix and Adhesion Proteins*, edited by T. Kreisand R. Vale. New York: Oxford Univ. Press, , p. 503–562. 1999.

**[24] G . McGeehan, W .Burkhart, R . Anderegg, J.D Becherer, J.W. Gillikin, J.S. Graham.** Sequencing and characterization of the soybean leaf metalloproteinase: structural and functional similarity to the matrix metalloproteinase family. *Plant Physiol pp*. 99: 1179–1183, 1992.

**[25] J.M .Maidment, D. Moore, G.P Murphy, G. Murphy, I.M .Clark.** Matrix metalloproteinase homologues from *Arabidopsis thaliana*: expression and activity. *J Biol Chem pp*.274: 34706–34710, 1999.plants and algae

**[26]** T.**Kinoshita, H.Fukuzawa , T.Shimada ,T.Saito , Y.Matsuda** Primary structure and expression of a gamete lytic enzyme in Chlamy.domonas-Reinhardtii: similarity of functional domains to matrix metalloproteases. *Proc Natl Acad Sci USA pp.* 89: 4693–4697, 1992

**[27]** F.**Mannello , L.Canesi ,M. Faimali , V. Piazza, G.Gallo , S. Geraci .** Characterization of metalloproteinase-like activities in barnacle (*Balanus amphitrite*) nauplii. *Comp Biochem Physiol B Biochem Mol Biol pp*.135: 17–24, 2003.

.

**[28] M.D.Sternlicht, Z.Werb .** How matrix metalloproteinases regulate cell behavior. *Annu Rev Cell Dev Biol* 17: 463–516, 2001

**[29]** W.C.**Parks , C.L.Wilson ,Y.S Lopez-Boado .** Matrix metalloproteinases as modulators of inflammation and innate immunity. *Nat Rev Immunol pp*.4: 617–629, 2004 .

**[30] J.D.Thompson,T.J. Gibson, F. Plewniak, ,F. Jeanmougin, D.G. Higgins, :** The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools **.** *Nucleic Acids Research*, pp.**24:**4877-4882. 1997

**[31] J.Felsenstein : PHYLIP.** Version 3.5c edition. Department of Genetics,University of Washington, Seattle.1993

**[32] J.P.Huelsenbeck, M.R.Bayes: Bayesian inference of phylogeny.** Department of Biology, University of Rochester , 2000.