In-silico Comprehensive Sequence And Structure Analysis of Proteases Family

Meenakshi Bhat and S. A. M. Rizvi
Department of Computer Science, Jamia Millia Islamia, New Delhi.
menaxi.jmi@gmail.com, samsam_rizvi@yahoo.com

ABSTRACT
Bioinformatics is the computing response to the molecular revolution in biology. This revolution has reshaped the life sciences and given us a deep understanding of DNA sequences, RNA synthesis, and the generation of proteins. An ever increasing number of biological modeling methods depend on the assembly of accurate Multiple Sequence Alignments (MSAs). Traditionally, the main applications of sequence alignments have included Motif finding, secondary or tertiary structure prediction, function prediction phylogenetic tree reconstruction, Hidden Markov Modeling (profiles), and much minor but useful application such as data validation [2]. A large majority of these applications are based on the analysis of protein sequences, with the notable exception of ribosomal RNA, possibly back-translated into nucleic acid sequences in the context of phylogenetic analysis. While this type of approaches still constitutes the vast majority of published applications for MSAs, recent biological discoveries coupled with the massive delivery of functional, structural and genomic data are rapidly expanding the potential scope of alignment methods. Sequence comparison is considered as backbone of Bioinformatics. Bioinformaticians and molecular biologists often need molecular sequences like DNA, RNA & proteins to compare them with each other in order to determine the degree of similarity on the basis of which various conclusions are derived regarding the features, structures, behavior and function of an organism or entire species as a whole [3]. Present Proposed study here is an attempt to develop a specific algorithm for searching particular pattern (motifs) in the genome sequences of the protein enzyme, proteases. On the basis of these sequence analysis, one can predict their secondary or tertiary structures. One can analyze the phylogenetic relation of these Proteases by constructing its phylogenetic trees in light of evolution. Storing all the information extracted from these sequences in new database is another perspective of the present study.

KEYWORDS
Multiple Sequence Alignments (MSAs), Hidden Markov Modeling (HMM), Motif finding, Secondary or Tertiary Structure Prediction,

1.0 INTRODUCTION
A protease is any enzyme that conducts proteolysis, that is, begins protein catabolism by hydrolysis of the peptide bonds that link amino acids together in the polypeptide chain.
pepsatin. Eukaryotic aspartic acid proteases include pepsins, cathepsins, and renins. They have a two-domain structure, probably arising from ancestral duplication. Retroviral and retrotransposon proteases (Pfam PF00077) are much smaller and appear to be homologous to a single domain of the eukaryotic aspartyl proteases.

D. Metalloprotease
Metalloproteases (or metalloproteases) constitute a family of enzymes from the group of proteases, classified by the nature of the most prominent functional group in their active site. There are two subgroups of metalloproteinases:
Exopeptidases: metalloexopeptidases.
Endopeptidases: metalloendopeptidases.

E. Threonine Proteases
Proteolytic enzyme with a threonine residue (Thr) in its active site is referred to as Threonine proteases. The prototype members of this class of enzymes are the proteasome catalytic subunits.

F. Glutamic acid Proteases
Glutamic acid proteases are a distinct, and recently re-classified, group of peptidases that are thought to be found only in fungi, found in all fungi except Saccharomyces class Glutamic proteases appear to be present in all other ascomycetes species examined. In addition to above describe mechanistic classes; there is a section of the Enzyme nomenclature which is allocated for proteases of unidentified catalytic mechanism [2]. This indicates that the catalytic mechanism has not been identified but the possibility remains that novel types of proteases do exist.[6][7]

2.0 MATERIALS AND METHODS
The analysis of these Protein sequences includes following steps:-
- Collection of information from PDB, NCBI, GenBank, DDBJ, EMBL, Prosite, BIND, MINT, KEGG, pfam, EC Enzymes.
- Flow charts given below explains the methods used for the analysis of sequences (As in figure 1 and in figure 2 )
- Phylogenetic Analysis consists of four main steps:- Constructing a multiple sequence alignment (As shown in figure 3)
- Determining the substitution model
- Tree building
- Tree evaluation
- Then comes the turn of structure prediction of these sequences (as is represented in figure 4)

CONCLUSION
Motif finding algorithmic approach is a significant approach in the study of DNA sequence analysis. One can find conserved regions (particular residues) after getting the DNA or Amino acid sequences of these particular Proteases, doing pair wise and multiple sequence alignment. Comprehensive in-silico analysis of these sequences will help in discovery of similar regulatory enzymes in other organisms which are yet to be experimentally characterized. Motif finding algorithmic approach. Phylogenetic analysis would in principle help in formulating predictive rules for detection of the line of diversion between proteases (detect the mutation). Storage of the collected information in the database for public use is another aspect-data making approach.[8]

FUTURE SCOPE
Reasons of their popularity include accumulated data on their enzymology, high throughput assays, biochemistry, physiology, pathology, 3D structures. Of the ~500 known human Proteases, ~15 are under investigation as potential drug target [7]. Although phylogenetic trees produced on the basis of sequenced genes or genomic data in different species can provide evolutionary insight, they have important role to represent the species evolutionary history, hybridization, convergent evolution, and conservation in sequences as well.

REFERENCES
Choose two sequences → Are the sequences protein sequences? → No → Do sequences encode protein (e.g. cDNA)? → No → Does sequence encode proteins and have introns? → No → Perform local alignment → Yes → Translate sequence → Yes → Predict gene structure → Yes → Alter Parameters, e.g., scoring matrix, gap penalties, and repeat alignment → No → Is alignment of high quality? → No → Examine sequences for presence of repeats or low-complexity sequences → Yes → Did alignment improve? → Yes → Perform statistical test of alignment score → Yes → Is alignment score significant? → Yes → Sequences are significantly similar → No → Sequences are not detectably similar → No → Sequences are not detectably similar → No → Sequences are not detectably similar → Figure 1
Choose three or more sequences

Are the sequences protein sequences?

Yes

Perform global alignment of sequences?

Is a convincing alignment produced?

Yes

Are the sequences cDNA sequences?

No

Translate into protein sequences

Are the sequences cDNA sequences?

Yes

Are the sequences genomic sequences that encode related proteins?

No

Predict gene structure

Are there a large number of sequences?

No

Make a profile or PSSM representation of the alignment

Yes

Analyze for patterns, repeats, etc.

Do the sequences encode RNA molecules?

No

Analyze promoter regions, intron-exon boundaries, etc.

Search for blocks

Yes

Analyze for secondary structure

Analyze for patterns, repeats, etc.

Analyze for secondary structure

Figure 2
In-silico Comprehensive Sequence and Structure Analysis of Proteases Family

Choose set of related sequence → Obtain multiple sequence alignment → Is there strong sequence similarity

- Yes → Maximum parsimony methods
- No → Is there clearly recognizable sequence similarity

- Yes → Distance method
- No → Maximum likelihood methods

Analyze how well data support prediction

Figure 3

Protein sequence → Database similarity search → Does sequence align with a protein of known structure?

- No → Protein family analysis
- Yes → Three-dimensional comparative modeling

- Yes → Relationship to known structure?
- No → Structural analysis
- Yes → Predicted three dimensional structural model

Three dimensional structure analysis in laboratory

Figure 4