Towards the Sequence and Structural Prediction of Proteases - An *in-silico* Study

M. Bhat and S. A. M. Rizvi

***Abstract*— Proteases cut long sequences of amino acids into fragments and regulate many physiological processes, so, it is called by many as "biology's version of swiss army knives". Different types of proteases have different action mechanisms, biological processes and therefore differ in their structures too. Sequence comparison is considered as backbone of bioinformatics and bioinformatics is the computing response to the molecular revolution in biology. 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 functions of an organism or entire species as a whole.**

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 identify their types and also can predict their secondary or tertiary structures. To address these problems, a 3-layer predictor, is proposed to develop by fusing the functional domain and sequential evolution information: the first layer is for identifying the query protein as protease or non protease; if it is a protease, the process will automatically go to the second layer to further identify it amongst the six types of proteases, and the third layer will be for structural analysis. Besides, analysis based on phylogenetic relation of these proteases by constructing their phylogenetic trees in the light of evolution can be done. Storing all the information extracted from these sequences in a new database is another perspective of the present study.

Index Terms—**Motif finding, Phylogeny, Proteases, Sequence Alignments, Secondary or Tertiary Structure Prediction.**

I. INTRODUCTION

Proteases are vitally important for life cycles and have become a main target in drug development. According to their action mechanisms, proteases are classified into six types:

(1) *Serine proteases*, (2) *Threonine proteases*, (3) *Cysteine proteases*, (4) *Aspartic acid proteases*, (5) *Metalloproteases*, (6) *Glutamic acid proteases.*

Traditionally, the main applications of in-silico sequence alignments have included motif finding, secondary or tertiary structure prediction, function prediction, phylogenetic tree reconstruction, and much minor but useful application such as data validation. With the avalanche of protein sequences generation during the post genomic age, it is highly desirable for both basic research and drug designers to develop a fast and reliable method for identifying the types of proteases and their structures according to their sequences or even just for whether they are proteases or not.

II. METHODOLOGY

In-silico analysis of these protein sequences includes following steps:-

First, the sequences of these proteins are collected from the databases such as, MEROPS, PDB, NCBI, GenBank, DDBJ, EMBL, Prosite, KEGG, pfam, EC Enzymes.

Second, the sequence comparison and analysis is done using the methods as shown in flow chart. (Fig. 1)

Third, after sequence collection and comparison, structure prediction of these sequences is done. (Fig. 2)

Last, sequence and structural analysis are followed by the phylogenetic analysis. (Fig. 3)

Manuscript received July 6, 2010. ("Towards the Sequence and Structural Prediction of Proteases"- An *in-silico* Study)
M. Bhat (Research scholar) is with the Department of Computer Science, Jamia Millia Islamia, New Delhi-25, India (phone: 009891372524; e-mail: menaxi.jmi@gmail.com). S. A. M. Rizvi (Associate Professor) is with the Department of Computer Science, Jamia Millia Islamia, New Delhi-25, India (phone: 009891820606; e-mail: samsam_rizvi@yahoo.com).

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

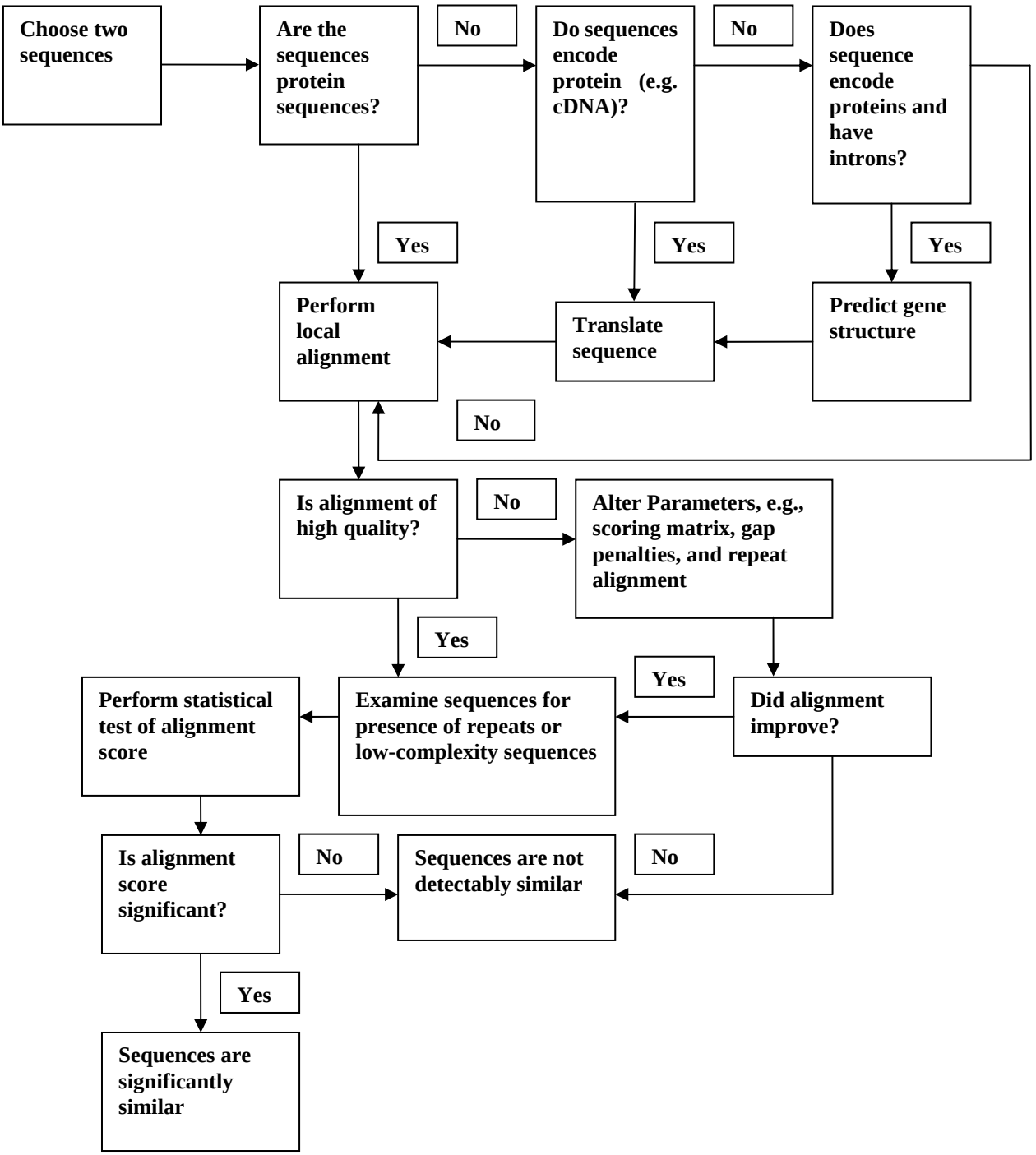

Fig. 1: Sequence Analysis

A. Sequence Analysis

Explanation:-Sequences can be either DNA or protein sequences. If one is a DNA sequence, then it is translated into protein sequence and thereafter alignment is done. As far as alignment is concerned, protein sequence alignment should be of high quality. Some of the characteristics of which are; relatively few gaps confined to particular regions of the alignment, notable identities, good conservation substitutions (structurally and chemically similar amino acids) etc. All these parameters are altered to achieve a good test of signifance, which is done to analyze the degree of similarity between the sequences.

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

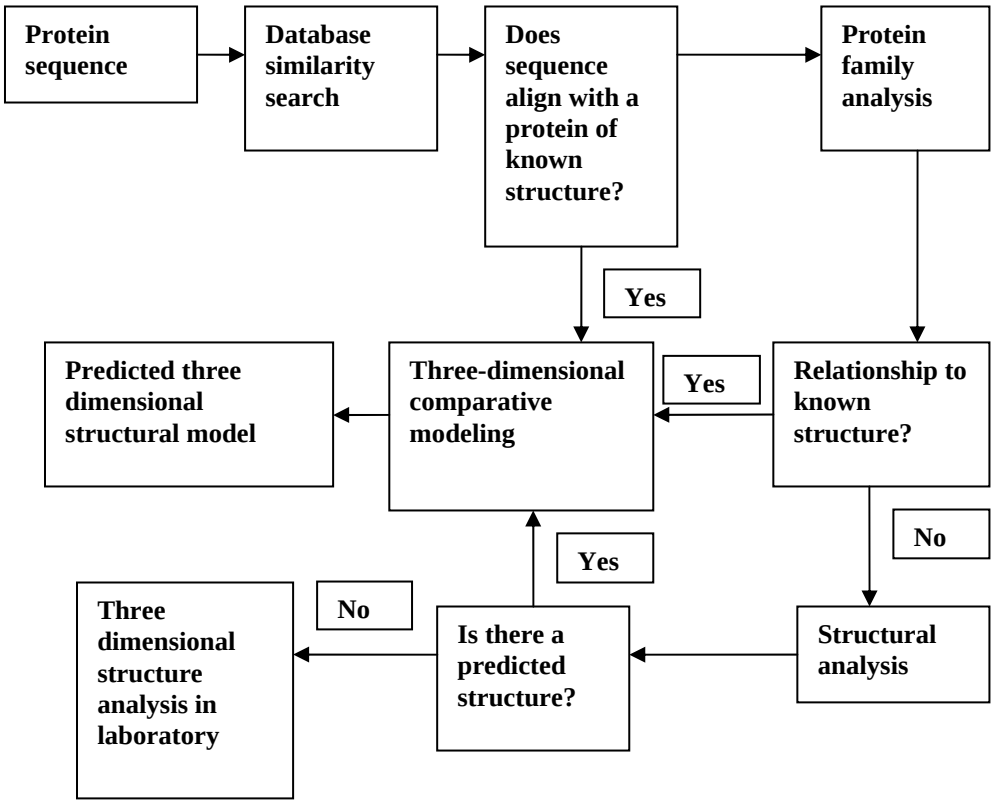

Fig. 2: Structure Prediction

B. Structure Prediction

Explanation:-Database similarity search for a protein sequence is done to know, to which protein family it belongs to? and if it is having the same structural folds as that of a particular fold family or not. If the given protein sequence belongs to a particular protein of a known family, three-dimensional modeling is done and structure of given protein sequence can be predicted. If otherwise, protein is having a known three-dimensional structure; it can be classified into fold families based on the common arrangements of secondary structures.

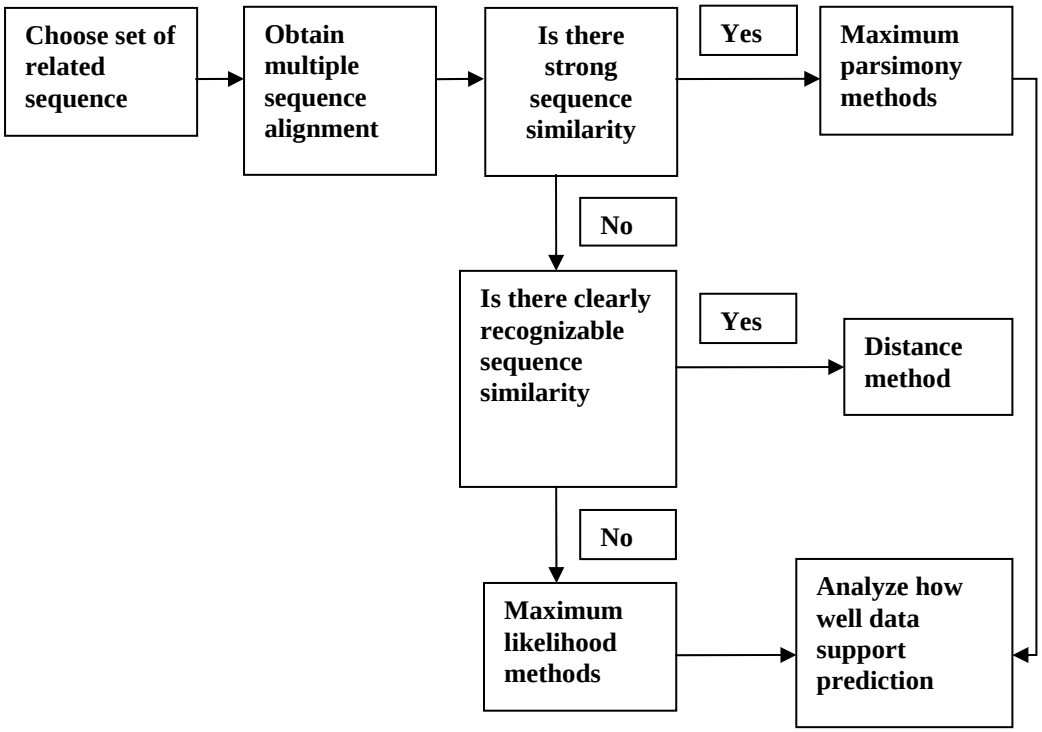

Fig. 3: Phylogenetic Analysis

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

C. Phylogenetic Analysis

Explanation:-The chosen sequences can be protein sequence or DNA depending upon the program one has decided to use for these sequences for their phylogenetic analysis. Second step refers to sequence alignment of these chosen sets of sequences and phylogenetic analysis should only be performed on those parts of sequences that can be aligned to a sufficient extent. In general phylogenetic methods analyze conserved regions that are represented in all the sequences. Three methods for phylogenetic analysis are defined till now depending upon the amount of similarity between the sequences of interest.

First, Maximum Parsimony Method; here the amount of variation among the sequences is negligible or in other words, sets of sequences show similarity to a great extent.

Second, Distance Method is used to predict an evolutionary tree when variation among these sets of sequences is present but amount is intermediate.

Third, Maximum Likelihood Method is used when among the set of sequences, variation is significantly more.

III. CONCLUSION

Motif finding algorithmic approach is a significant approach in the analysis of DNA and protein sequence. One can find conserved regions (particular residues) by performing sequence alignment of the amino acid sequences of these particular proteins (Proteases).

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. Given the sequence of an uncharacterized protein, one can identify whether it is a protease or non-protease? If it is, what type does it belong to? and its structural prediction also. Motif finding algorithmic approach, phylogenetic analysis would in principle help in formulating predictive rules for detection of the line of diversion between the proteases (detect the mutation).

IV. FUTURE SCOPE

Reasons of their popularity include accumulated data on their enzymologist, high throughput assays, biochemistry, physiology, pathology, 3D structures. Proteases account for about 2% of the human genome and 1–5% of genomes of infectious organisms. Of the ~500 known human Proteases, ~15 are under investigation as potential drug target.

REFERENCES

[1] Barrett, AJ; Rawlings, D.; Woessner, JF. Proteolytic enzymes. In: Oxford: Academic Press. , editor. *"Handbook of Proteolytic Enzymes"*. Barrett AJ, Rawlings D, Woessner JF; 1998. pp. 801–805.

[2] Hong-Bin Shen, Kuo-Chen Chou *"Identification of proteases and their types"*, published in Analytical Biochemistry 385 (2009) 153–160.

[3] David H Bos, Chris Mayfield and Dennis J Minchella *"Analysis of regulatory protease sequences identified through bioinformatic data mining of the Schistosoma mansoni genome"*, Published on 21 October 2009, BMC Genomics 2009.

[4] Kuo-Chen Chou, Hong-Bin Shen *"ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information"*, published in Biochemical and Biophysical Research Communications 376 (2008) 321–325.

[5] Kuo-Chen Chou, *"Structural Bioinformatics and its Impact to Biomedical Science"*, published in Science 12 September 1997.

[6] Meenakshi Bhat and S. A. M. Rizvi, *"In-silico Comprehensive Sequence Analysis of Proteases Family"*, published in the Proceedings of Ist IFIP International Conference on Bioinformatics in March 25-28, 2010 at Sardar Vallabhbhai National Institute of Technology, Surat, India.

[7] Meenakshi Bhat and S. A. M. Rizvi *"Comparative Phylogenetic Analysis of Protease"*, published in the Proceedings of IEEE International Advance computing Conference (IACC'09) held on March 6-7, 2009. ISBN No. - 978-98108-2465-5.

[8] David W. Mount, *"Bioinformatics- Sequence and Genome Analysis"*, University of Arizona, Tucson, Cold Spring Harbor, New York, USA.

ISBN: 978-988-18210-0-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)