International Journal of Computer Applications (0975 – 8887)
Volume 43– No.7, April 2012

Computational Analysis of Proteases Domains using Hidden Markov Model

Meenakshi Bhat
Senior Research Fellow
Department of Computer Science, Jamia Millia Islamia University, New Delhi, India

S. A. M. Rizvi
Associate Professor
Department of Computer Science, Jamia Millia Islamia University, New Delhi, India

ABSTRACT
In this paper, we present a three-layered predictor, Profinder, for identification and analysis of protein enzyme 'Protease'. This predictor is shaped by collecting the protease family domains represented by multiple sequence alignments and hidden morkov modeling techniques. Present study here is an attempt to develop a specific algorithm for searching particular domains in the genome sequences of these protein enzymes. Therefore, it is important for both basic research and drug discovery to consider the following two problems. Given the sequence of a protein, determine whether the protein is a protease or not? And if so, then which class of proteases? It is only on the basis of their sequence analysis, one can identify their types and also can predict their secondary or tertiary structures. User can test their sequences in fasta format for identification of proteases domain and therefore can get some insights on their fuctions and secondary structures.

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 this present *in-silico* study.

General Terms
Bioinformatics, *In-Silico* Analysis

Keywords
Proteases, Motifs, Sequence Alignment, Protein Domains, Hidden Morkov Model.

1. INTRODUCTION
Peptidases and their substrates are of great relevance to biology, medicine and biotechnology. Proteases, known as proteinases or peptidases are protein enzymes, which cut long sequences of amino acids and therefore called by many as biology's version of "Swiss Army Knives" [2]. Six catalytic types of peptidases can now be recognized as serine, threonine, cysteine, glutamic, aspartic or metallo groups play primary roles in catalysis[1]. The serine, threonine and cysteine peptidases are catalytically very different from the aspartic and metallopeptidases as in former group; the nucleophile of the catalytic site is part of an amino acid, whereas it is an activated water molecule in the other groups [16].

2. HIDDEN MORKOV MODELS
Molecular biologists use hidden Markov models (HMMs) as a popular tool to statistically describe biological sequence families [13]. This statistical description can then be used for sensitive and selective database scanning, e.g., new protein sequences are compared with a set of HMMs to detect functional similarities [14].

3. MATERIALS AND METHODS

3.1 Data Collection and Sequence Analysis
To develop a powerful statistical predictor, the first important thing is to construct a high quality and accurate dataset. After intensive literature survey, protease databases were searched and to begin with, MEROPS (http://merops.sanger.ac.uk) [8] is a comprehensive and excellent database of proteases which relies on large scale experiments and automatic annotation [3]. In all 3,051 proteases sequences (dataset) [4] from MEROPS classification system were compiled and studies, distribution of which is, Aspartic acid (A1, A2, A3….) 258, Cysteine (C1….) 589, Glutamic (G1…..)39, Metallo (M1….) 1040, Serine (S1….) 1063, Threonine (T1….) 62, multiple sequence alignment and phylogenetic reconstruction of the sequences (dataset) were carried out using clustalW 2.0[9]. Profile HMMs were built using HMMER 3.0 [7].

3.2 Tool development
The Dataset.

Multiple Sequence Alignment of six various types of Proteases from dataset using ClustalW 2.0 is done.

Hidden Markov modeling technique is used to make profiles of these types of protein enzymes (*hmmbuild*) - Profile hidden markov modeling analysis [14] was carried out by HMMER 3.0 package. Profile HMMs are statistical models of multiple sequence alignments [5], or even of single sequences. They capture position-specific information about how conserved each column of the alignment is, and which residues are likely. HMMER 3.0 [6] is a widely used open source implementation of profile HMM algorithms with protein databases written in the C programming language [9]. We have measured the performance of the *hmmbuild* algorithm, which is part of the HMMER 3.0 package. *hmmbuild* also aligns a query profile HMM to all protein sequences of a given database. A variety of different multiple sequence alignment formats are permitted (Clustal, Msf, Selex, Stockholm and aligned Fasta format). Once, the multiple sequence alignment is converted to a profile HMM using *hmmbuild* (in its default mode). Concisely, HMM technique is used to make profiles of these types of protein enzymes

(*hmmbuild*). A sequence profile may also be represented by a hidden Markov model, referred to as a profile HMM. Combination of profiles of these six types of profiles leads to a protein identifier (*catdata*) - Profinder.

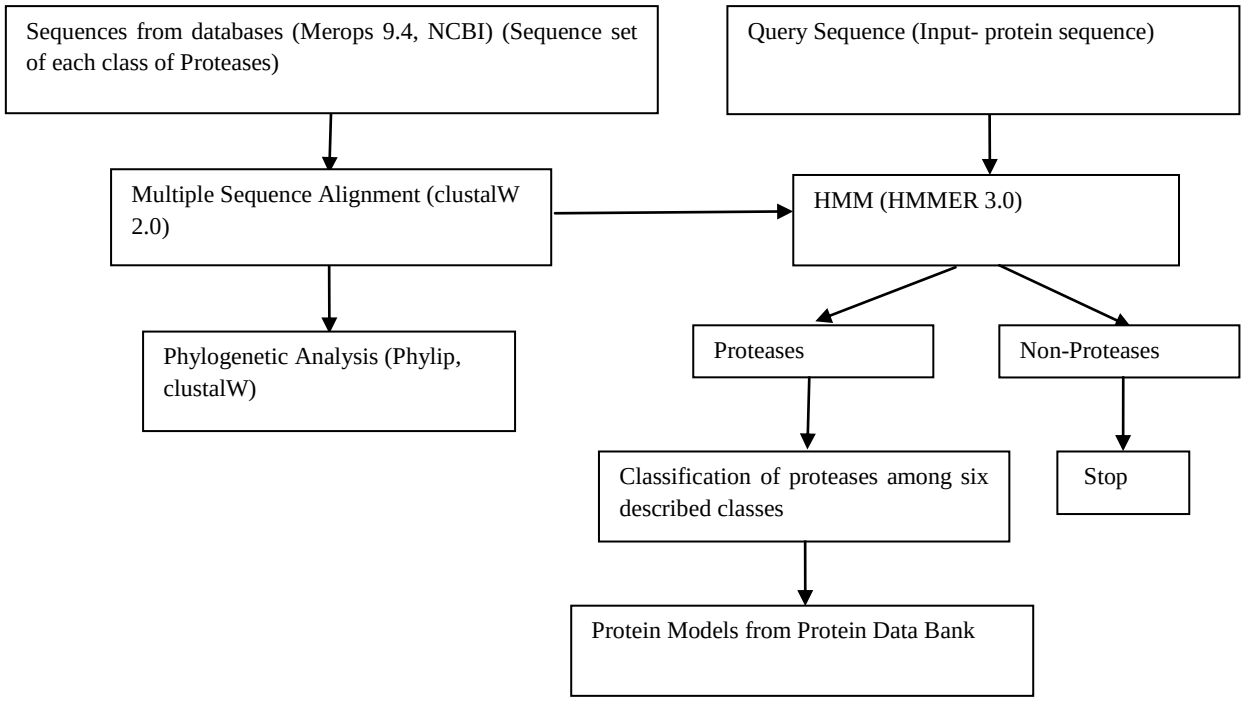


Figure 1: Frame Work of Profinder

3.3 Case Study

Sequences have been taken from the dataset to cross validate the results of our tool. Results from Profinder has been compared to pfam[10] (Wellcome Trust Sanger Institute, EXPASY) commonly used HMM based tool for profile searches and profinder is performing equally well and the advantage of profinder over existing tool for domain classification [11] is that it can go to upto three layers prediction. Results can be seen below (in Table 1 and Table 2) as a graphical representation of scores, alignments and HMM alignments.

>MER017342

IATESFLASRAGYEVYKKGGNAADAVVAASAVLTYTLPHLGGVGGDFLAVVHKGDRAESVLGLGWAPRKIPDRPPRRGIQS
AVVPGYIAGLVEFHKRYGALPWGEVIGTALDFMRKATVHPSLAAAIERHRDLLASDPGGRLYLALPLEPGAPYKIEPLLKLW
GVLRDNPLLFYDEIARDLEQYGYFELDDFLKYRPEVKPPVFINYDGWVIYEAPPPSLGFVVLLTVKLSQPVKSPFSYARIKNTV
AALRKAHWARDAYLHDGEVPVYDILSGGVQLGEADKPEPTPGTTYLAAGDKELTVSAIQSLYYPFGSGFTDLKWGITFNNR
ASDFTTGLNRPAPRKRPSHTLSAVVMVKEGEAWALGASAGHYRPAIYAQLIQNVIWYGMSPREAILAPRFIWTGGWEVQAE
EGWEAGLGVTIVKYPSRMGVAAALRKRNNYIAAVADIRGDGL

>MER022928

SGTSSSTEEPFPVLAVGKDGRGNYYVNSTFGTPGQRQRLLVDIIQPYINLVSGTSESHNEYSGVYHKHPSYLMNDSTSSVPVSP
GQIYEISFIDGRAVNCTLVTDDMNFTNVSSENSSTALITDLMVTRDNVQFNSGSLSISNVSFFDIQSSNFKTSGLLGLSGKVTNP
GNAIDSSQYTEQSYFLSLLKDADIIESSSYSLWLAGDTSTYKTYRDPISNCGKLLLGGVDPSLFTGTLGKFDLIPYVDPVSNAVS
VGYPIVPLGPIYIVSNSGQSLNMTSKDFLSPALLDSTSSVSYLPTSTIIQIAVQIAATYVESLDRWLVQCSIADMGVSLGFRLREL
TIEIPLRDLLSSTYDTSTNSSMFFSSGQEACFLTLYANTNTGVNILGEAFIKNIYMAMDLEDNTIAIAQAKKVEDDAVTE

>MER069322

LVLGGPMGIRDLGTSTYPWLIEEVGLIKEALNQDVGIIGVCLGAQLLAYAAGGDVERLQDELYHQPLVEIGWDNIFPQSLDN
NCKLTTLLNHSFPVLHWHGDRILLPNTAELIASSYRCKEQLFSISSLAYGIQFH

4. PREDICTION ACCURACY

I) Query: MER017342 [L=451]

Scores for complete sequence (score includes all domains):

--- full sequence --- --- best 1 domain --- -#dom-

E-value score bias E-value score bias exp N Model Description

------- ------ ----- ------- ------ ----- ---- -- -------- -----------

3e-57 183.4 0.0 3.4e-57 183.2 0.0 1.0 1 thrfasta

Threonine Proteases

II) Query: MER022928 [L=420]

Scores for complete sequence (score includes all domains):

--- full sequence --- --- best 1 domain --- -#dom-

E-value score bias E-value score bias exp N Model Description

------- ------ ----- ------- ------ ----- ---- -- -------- -----------

5.3e-24 74.2 20.1 6.5e-24 73.9 13.9 1.1 1 aspfasta

Aspartic Acid Proteases

III) Query: MER069322 [L=138]

Scores for complete sequence (score includes all domains):

--- full sequence --- --- best 1 domain --- -#dom-

E-value score bias E-value score bias exp N Model Description

------- ------ ----- ------- ------ ----- ---- -- -------- -----------

[No hits detected that satisfy reporting thresholds]

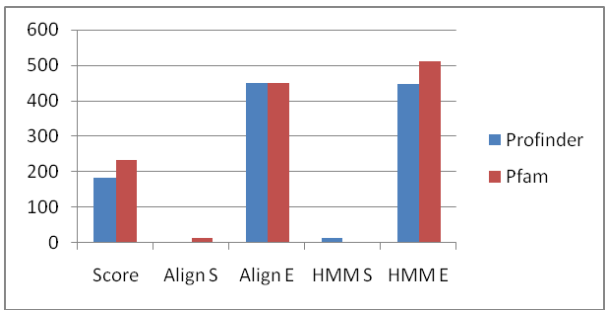
Domain annotation for each model (and alignments):

[No targets detected that satisfy reporting thresholds]

5. PERFORMANCE

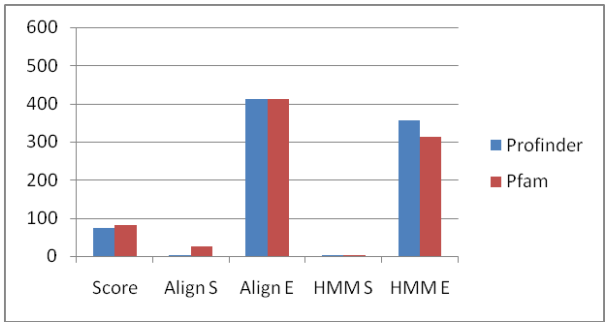
5.1.1 Sequence 1

Table 1:



5.1.2 Sequence 2

Table 2:



	E-value	Score	Align S	Align E	HMM S	HMM E
Profinder	3.4e-57	183.2	1	449	16	447
Pfam	3.8e-69	233.2	14	450	2	510

6. CONCLUSION

Biologists are in need of information in digital form for correct and meaningful interpretation of their biological data. The ongoing web interface is currently the only program available for detection and classification of these proteases, since no existing database or annotation program is able to identify, classify and represent 3d structures of these proteases one after another in a particular search algorithm. These features make this software a valuable resource which can aid in identification of proteases and its type found in newly sequenced genomes. This tool represents a collection of profile Hidden Markov Models (HMMs) based on a rigorous analysis of six distinct domain families of protein enzyme, Protease, namely, (i) Aspartic acid domains, (ii) Cystein domains, (iii) Glutamic (iv) Domain of Metalloproteases, (v) Serine, and (vi) Threonine domains. Focus here is on online automated, interactive and predictive search tool, the first of its kind dedicated to proteases domains.

To make our bioinformatics tool online available, proposed programme is written in Perl and uses the Bioperl toolkit. The website is developed using the PHP language and the Apache web server

7. ACKNOWLEDGMENTS

Funding from Union Grant Commission, Government of India is highly acknowledged.

8. REFERENCES

[1] A.J. Barrett, J.K. McDonald, Nomenclature: protease, proteinase and peptidase, Biochem. J. 237 (1986) 935.

[2] C. Seife, Blunting nature's Swiss army knife, Science 277 (1997) 1602–1603.

[3] Identification of proteases and their types by Hong-Bin Shen , Kuo-Chen Chou Published by Elsevier, Analytical Biochemistry 385 (2009) 153–160.

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

[5] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215:403-410.

[6] Pearson WR, Lipman DJ: Improved Tools for Biological Sequence Comparison. Proc Natl Acad Sci U S A 1988, 85:2444-2448.

[7] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, DJ Lipman: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25(17):3389-402.

[8] N.D. Rawlings, D.P. Tolle, A.J. Barrett, MEROPS: the peptidase database, Nucleic Acids Res. 32 (2004) D160–D164.

[9] Eddy, S.R. (1998) Profile hidden Markov models. Bioinformatics, 14, 755–763. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Research, 22(22), 4673-4680.

[10] Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. Nucleic Acids Res., 37, D229–D232.

[11] Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy, S.R., Sonnhammer, E.L. et al. (2008) The Pfam protein families database. Nucleic Acids Res., 36, D281–D288.

[12] Durbin R, Eddy S, Krogh A, Mitchison G: Biological Sequence Analysis: Probablistic Models of Protein and Nucleic Acids. Cambridge University Press, Cambridge, UK; 1998.

[13] HMMER user's guide: biological sequence analysis using profile hidden Markov models [ftp://selab.janelia.org/pub/software/hmmer/CURRENT/Userguide.pdf].

[14] Eddy SR: HMMER: Profile hidden Markov models for biological sequence analysis. 1998 [http://hmmer.janelia.org/].

[15] Puente XS, Sanchez LM, Overall CM, Lopez-Otin C: Human and mouse proteases: a comparative genomic approach. Nat Rev Genet 2003, 4:544-558.

[16] Barrett A.J., Rawlings ND, Woessner JF. The Handbook of Proteolytic Enzymes, 2nd ed. Academic Press, 2003. ISBN 0-12-079610-