GLOBAL JOURNAL OF BIO-SCIENCE AND BIOTECHNOLOGY

www.scienceandnature.org

INSILICO STUDIES OF CELLULASE OF *ASPERGILLUS TERREUS*

S. Maragathavalli, S.V. Megha, S. Brindha, V. Karthikeyan & B. Annadurai
Research and Development Centre, Bharathiyar University, Coimbatore-641 041
*Corresponding author: E-mail: mailme_emerald@yahoo.com

ABSTRACT
Cellulases refer to a class of enzymes produced majorly by fungi, bacteria and protozoans that catalyze cellulolysis. Cellulase enzyme is used extensively in various industries, especially in textile, food and in the bioconversion of lignocellulosic wastes to alcohol. The extensive use of cellulase in industries depends on the cost of the enzyme and hence considerable research is being carried out to isolate better microbial strains and also to develop new fermentation processes with the aim to reduce the product cost. Cellulases from different strains of Pseudomonas species were analyzed using computational tools. The physicochemical properties of the selected cellulases were analyzed by using ExPASy's Prot Param tool and it was found that the molecular weight (M.Wt) ranges between 40927.4-100058.7 Da. Isoelectric Points (pI) of all the organisms were found to be acidic in nature. The aliphatic index infers that all the cellulases are stable. The negative value of GRAVY indicates that there will be better interaction with water. The secondary structure prediction was done by SOPMA which showed that random coils dominated all the other conformations. Multiple sequence analysis and evolutionary analysis of cellulases were carried out by CLC workbench. The Phylogenetic analysis was done using Neighbour joining method. The 3D structures of cellualses were obtained by ESyPred 3D server.

KEY WORDS: *ExPasy, Pi, M.Wt., GRAVY, EMBL-EBI, DDBJ, NCBI, OMIM, PubMed, Cellulase.*

INTRODUCTION

Cellulase I.U.B.:3.2.1.41, 4-(1, 3; 1, 4)-β-D-Glucan-4-glucanohydrolase, Cellulase refers to a group of enzymes which, acting together, hydrolyze cellulose. It has been reviewed by Emert *et al*. (1974) and Whitaker (1971). Cellulose is a linear polysaccharide of glucose residues connected by β-1,4 linkages. Like chitin it is not cross-linked. Native crystalline cellulose is insoluble and occurs as fibers of densely packed, hydrogen bonded, anhydro glucose chains of 15 to 10,000 glucose units. Its density and complexity make it very resistant to hydrolysis without preliminary chemical or mechanical degradation or swelling. In nature cellulose is usually associated with other polysaccharides such as xylan or lignin. It is the skeletal basis of plant cell walls. According to Spano *et al*. (1975) cellulose is the most abundant organic source of food, fuel and chemicals. However, its usefulness is dependent upon its hydrolysis to glucose. Acid and high temperature degradation is unsatisfactory in that the resulting sugars are decomposed; also, waste cellulose contains impurities that generate unwanted by products under these harsh conditions. Cellulase is a group of enzymes that catalyses cellulolysis. It is mainly produced by fungi, bacteria and some protozoans. The active research of cellulases was started in 1950. After knowing its potentiality to convert lignocellulases. It is studies extensively due to their applications in the hydrolysis of cellulose, the most abundant biopolymer and potential source of utilizable sugar, which serves as a raw material in the production of chemicals and fuel (Ali et al 2011, Pradeep *et al*., 2012). Since, Cellulases is used mostly in textiles, food and the bioconversion lignocellulosic waste to alcohol, it becomes industrially important. Because largely is used in the industries, large scale of production (Microbial strains). Isolation and purification, Procedures are required. In addition to that the computational tools and *insilico* studies are required to preserve and reduce the cost of cellulase. Bioinformatics revolutionized the field of molecular biology. The raw sequences information of proteins and nucleic acid can convert to analytical and relative information with the help of soft computing tools. Prediction of protein function is important application of bioinformatics (Prasanth *et al*., 2010). In this chapter Bioinformatics analysis characterization of cellulases from *Aspergillus terreus* species were carried out. Hence, the *insilico* studies by using ExPASy, Protparam tool are used for determining molecular weight. The secondary structure prediction has to be done by SOPMA to show the random coily multiple sequence analysis evolutionary analysis of cellulases has to be carried out by CLC work bench. The phylogenic analysis has to be done using neighbor joining method. The 3D structure of cellulases has to be done by ESyPREDE Spred 3D server. These parameters will help.The biochemist and physiologies in extraction, purification, separation and industrial application of the enzymes.

SYSTEM & METHODS

Description of databases used

1. NCBI

The international Nucleotide Sequence Database Collaboration (referred as "GenBank") is a joint production of the nucleotide sequence database by the DDBJ (DNA Data Bank of Japan), EBI (European Bioinformatics Institute) and NCBI (National Center for Biotechnology Information).The nucleotide sequence

databases are data repositories, accepting nucleic acid sequence data from the community and making it freely available. The databases strive for completeness, with the aim of recording ever publicity known nucleic acid sequence.

2. EBI

The EMBL-EBI lies in the 55 acres of landscaped parkland in rural Cambridge shire that make up the Welcome Trust Genome Campus. The Campus also houses the Welcome Trust Sanger Institute, making it one of the world's largest concentrations of expertise in genomics and bioinformatics. The EMBL-EBI grew out of EMBL's pioneering work to provide public biological databases to the research community. Although we are geographically separated from EMBL's main headquarters in Heidelberg and its other sites in Grenoble, Hamburg and Monterotondo, the EMBL-EBI is an integral part of EMBL. We play a vital role in achieving EMBL's mission of providing a top-quality research environment that also develops new technologies, and provides services and training to Europe's molecular life scientists. Like the other EMBL sites, we have an extremely cosmopolitan staff base, and alumni who have moved on to successful careers all over the world.

3. PUBMED

PubMed, available via the NCBI Entrez retrieval system, was developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), located at the U.S. National Institutes of Health (NIH). Entrez is the text-based search and retrieval system used at NCBI for services including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, and many others. PubMed provides access to citations from biomedical literature. Link Out provides access to full text articles at journal Web sites and other related Web resources. PubMed also provides access and links to the other Entrez molecular biology resources.

4. MEDLINE

Medline Plus will direct you to information to help answer health questions.Medline Plus brings together authoritative information from NLM, the National Institutes of Health (NIH), and other government agencies and health-related organizations. Pre formulated MEDLINE searches are included in Medline Plus and give easy access to medical journal articles. Medline Plus also has extensive information about drugs, an illustrated medical encyclopedia, interactive patient tutorials, and latest health news

5. SWISS PROT

Swiss Prot is a curate biological database of protein sequences created in 1986 by Amos Bairoch during his PhD and developed by the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. It strives to provide a high level of annotation (such as the description of the function of protein, its domains structure, post translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other bimolecular database as well as extensive external documentation. The protein sequence databases are the most comprehensive source of information on proteins. It is necessary to distinguish between universal databases covering proteins from all species and specialized data collections storing information about specific groups of proteins, or about the proteins of a specific organism.

6. PDB

The PDB is the single worldwide repository for the processing and distribution of 3-D structure data of large molecules of proteins and nucleic acids, as determined by X-ray crystallography or nuclear magnetic resonance (NMR) imaging. The molecules described by the files are usually viewed locally by dedicated software, or can be visualized on the World Wide Web. The number of known protein structures is increasing very rapidly and these are available on the protein Data Bank. There is also a database of structures of 'small' molecules, of interest to biologists concerned with protein-ligand interactions, from the Cambridge Crystallographic Data Centre.

7. RCSB DATABASE

The World Wide Web site of the protein data bank at the RCSB offers a number of services for submitting and retrieving three dimensional structure data. The homepage of the RCSB site provides links to services for depositing three-dimensional structures, information on how to obtain the status of structures undergoing processing for submission. Ways to download the PDB database and links to other relevant sites and software.

DESCRIPTION OF TOOLS USED

Protparam

ProtParam is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss Prot or TrEMBL or for a user entered sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half life, instability index, aliphatic index and grand average of hydropathicity (GRAVY)

Compute PI/MW

Compute pI/Mw is a tool which allows the computation of the theoretical pI (isoelectric point) and Mw (molecular weight) for a list of Prot or entries or for user entered sequences

Reverse translate

Reverse Translate accepts a protein sequence as input and uses a codon usage table to generate a DNA sequence representing the most likely non degenerate coding sequence. A consensus sequence derived from all the possible codons for each amino acid is also returned. Use Reverse Translate when designing PCR primers to anneal to an un sequenced coding sequence from a related species.

Profile scan

Profile, Scan uses a method called pfscan to find similarities between a protein or nucleic acid query sequence and a profile library. In this case, three profile libraries are available for searching. First is PROSITE an ExPASy database that catalogs biologically significant sites through the use of motif and sequence profiles and patterns. Second is Pfam. Which is a collection of protein domain families that differ from most such collections in one important aspect the initial alignment of the protein domains is families that differ from most such collections in one important aspect the initial alignment of the protein domains is done by hand? Rather than by depending on

automated methods. As such Pfam contains slightly over 500 entries but the entries are potentially of higher quality. The third profile set is referred to as the Gribskov collection.

SOPMA

The protein Sequence Analysis server at the Centre National de la Recherché Scientifique in Lyons. France takes a unique approach in making secondary Structure predictions: rather than using a single method, it uses five, the predictions from which are subsequently used to come up with a "consensus predictions." The methods used are the Garnier Gibrat Robson method the Levin homolog method the double-prediction method the PhD method described above as part of Predict Protein, and the method of CNRS itself, called SOPMA as Briefly, this self optimized prediction method builds sub databases is quences with known secondary structure prediction based on sequence similarity. The information from the sub databases is then used to generate a prediction on the query sequence.

SIGNALP

SignalP predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models.

TARGETP

TargetP predicts the subcellular location of eukaryotic proteins. The location assignment is based on the predicted presence of any of the N terminal presequences: chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP) or secretory pathway signal peptide (SP).

CHOLOROP

The ChloroP server predicts the presence of chloroplast transit peptides (cTP) in protein sequences and the location of potential cTP cleavage sites. A related service TargetP predicts the subcellular location of proteins by integrating predictions of chloroplast transit peptides, signal peptides and mitochondrial targeting peptides.

RESULTS

1. Prot Param Result:

User-provided sequence:

10 20 30 40 50 60
MNCRKYLLSG LAVFGLAATS AVAALSTDDY VEAAWMTTRF FGAQRSGQGP NWILDGTSNP
70 80 90 100 110 120
TSFTKDSYNG KDVSGGWFDC GDHVMYGQSQ GYASYVLALA YAEFTEVSTT FILVTTPTTR
130 140 150 160 170 180
KPTTTPMKSG KPNKVRDLLE ELRYEADFWV KAAIDGNNFV TVKGDGNADH QKWVTAGAMS
190 200 210 220 230 240
KLGSGEGGEP RCITGNANDG FTSGLAAAML AVMARVDPDT ANQAKYLKAA KTAYSYAKSH
250 260 270 280 290 300
KGVTNSQGFY ESSWWDGRWE DGPFLAELEL YRTTGENSYK TAAIDRYDNL KFSLGEGTHF
310 320 330 340 350 360
MYSNVVPLSA VMAEAVFEET PHGMRKEAIG VLDLIYEEKA KDKIFQNPNG MGSGKFPVRV
370 380 390 400 410 420
PSGGAFLYAL SDKFNNTNEH MEMIEKNVSY LLGDNGSKKS YVVGFSKNGA NAPSRPHHRG
430 440 450
YYANEKRWRR SRRCSESSRK EQALGRYDCW RLY .

Number of amino acids: 453

Molecular weight: 50042.0

Theoretical pI: 8.35

Amino acid composition:

Ala (A)	45 9.9%
Arg (R)	23 5.1%
Asn (N)	25 5.5%
Asp (D)	24 5.3%
Cys (C)	5 1.1%
Gln (Q)	9 2.0%
Glu (E)	26 5.7%
Gly (G)	44 9.7%
His (H)	8 1.8%
Ile (I)	9 2.0%
Leu (L)	30 6.6%
Lys (K)	30 6.6%
Met (M)	13 2.9%
Phe (F)	20 4.4%
Pro (P)	16 3.5%
Ser (S)	36 7.9%
Thr (T)	30 6.6%

Trp (W) 10	2.2%
Tyr (Y) 24	5.3%
Val (V) 26	5.7%

Total number of negatively charged residues (Asp + Glu): 50
Total number of positively charged residues (Arg + Lys): 53
Atomic composition:

Carbon	C	2218
Hydrogen	H	3378
Nitrogen	N	612
Oxygen	O	678
Sulfur	S	18

Formula: $C_{2218}H_{3378}N_{612}O_{678}S_{18}$
Total number of atoms: 6904
Extinction coefficients
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.
Ext. coefficient 91010
Abs 0.1% (=1 g/l) 1.819, assuming ALL Cys residues appear as half cystines
Ext. coefficient 90760
Abs 0.1% (=1 g/l) 1.814, assuming NO Cys residues appear as half cystines
Estimated half-life
The N-terminal of the sequence considered is M (Met).
The estimated half-life is: 30 hours (mammalian reticulocytes, *in vitro*).
>20 hours (yeast, *in vivo*).
>10 hours (Escherichia coli, in vivo).
Instability index
The instability index (II) is computed to be 29.52
This classifies the protein as stable.

Result:
The given protein is calculated with various parameters to give good results. There are number of parameters are used to analyse the given protein sequence.

1. Calculated number of amino acid present in the given protein sequence is 453.
2. Calculated the molecular weight of protein sequence is 50042.0
3. Calculated the theoretical isoelectric point of protein sequence is 8.35
4. Calculated each amino acid sequence molecular weight represents in a molecular percent.
5. Calculated the number of positive charges amino acids is 53
6. Calculated the number of negative charge amino acids is 50
7. Calculated the number of atoms of amino acid of protein is 6904
8. Give the molecular formula of given protein sequence is C2218 H3378 N612 O678 S18
9. The extinction co-efficient of protein sequence is calculated.
10. Computed the half-life of given protein is 30 (hours)
11. Computed the instability of given protein sequence is 29.52

Inference
From this result I have got protein molecular weight, isoelectric point, each aminoacids composition with molar percent values, positive charge residues and negative charge residues, number of atoms of amino acids, formula of proteins, total number of atoms, stability of proteins, half of proteins and extinction co-efficient in the protein sequence.

2. Compute pI/Mw Result
Theoretical pI/Mw (average) for the user-entered sequence:

10	20	30	40	50	60
MNCRKYLLSG	LAVFGLAATS	AVAALSTDDY	VEAAWMTTRF	FGAQRSGQGP	NWILDGTSNP
70	80	90	100	110	120
TSFTKDSYNG	KDVSGGWFDC	GDHVMYGQSQ	GYASYVLALA	YAEFTEVSTT	FILVTTPTTR
130	140	150	160	170	180
KPTTTPMKSG	KPNKVRDLLE	ELRYEADFWV	KAAIDGNNFV	TVKGDGNADH	QKWVTAGAMS
190	200	210	220	230	240
KLGSGEGGEP	RCITGNANDG	FTSGLAAAML	AVMARVDPDT	ANQAKYLKAA	KTAYSYAKSH
250	260	270	280	290	300
KGVTNSQGFY	ESSWWDGRWE	DGPFLAELEL	YRTTGENSYK	TAAIDRYDNL	KFSLGEGTHF
310	320	330	340	350	360
MYSNVVPLSA	VMAEAVFEET	PHGMRKEAIG	VLDLIYEEKA	KDKIFQNPNG	MGSGKFPVRV
370	380	390	400	410	420
PSGGAFLYAL	SDKFNNTNEH	MEMIEKNVSY	LLGDNGSKKS	YVVGFSKNGA	NAPSRPHHRG
430		440			450
YYANEKRWRR SRRCSESSRK EQALGRYDCW RLY					

Theoretical pI/Mw: 8.35 /50042.0

Result

The given protein is calculated with various parameters to give good results. There are number of parameters are used to analyse the given protein sequence.

Inference

In this site calculate the molecular weight of the 1ut9A sequence is 50042.0 and isoelectric point is 8.35.

3. Reverse Translate Results

Results for 394 residues sequence "P23665| GUNA_ FIBSU Endoglucanase A - Fibrobacter MNCRKYLLSGLAVFGLAATSAVAALSTDDYVEAA WMTTRFFGAQRSGQGPNWILDGTSNP"starting "TSFTKDSYNG"

>reverse translation of P23665|GUNA_FIBSU Endoglucanase A - Fibrobacter MNCRKYLLSGLAVFGLAATSAVAALSTDDYVEAA WMTTRFFGAQRSGQGPNWILDGTSNP to a 1182 base sequence of most likely codons.

accagctttaccaaagatagctataacggcaaagatgtgagcggcggctggtttg attgcggcgatcatgtgatgtatggccagagccagggctatgcgagctatgtgct ggcgctggcgtatgcggaatttaccgaagtgagcaccacctttattctggtgacca ccccgaccacccgcaaaccgaccaccaccccgatgaaaagcggcaaaccgaa caaagtgcgcgatctgctggaagaactgcgctatgaagcggatttttgggtgaaa gcggcgattgatggcaacaactttgtgaccgtgaaaggcgatggcaacgcggat catcagaaatgggtgaccgcgggcgcgatgagcaaactgggcagcggcgaag gcggcgaaccgcgctgcattaccggcaacgcgaacgatggctttaccagcggc ctggcggcggcgatgctggcggtgatggcgcgcgtggatccggataccgcgaa ccaggcgaaatatctgaaagcggcgaaaaccgcgtatagctatgcgaaaagcca taaaggcgtgaccaacagccagggcttttatgaaagcagctggtgggatggccg ctgggaagatggcccgtttctggcggaactggaactgtatcgcaccaccggcga aaacagctataaaaccgcggcgattgatcgctatgataacctgaaatttagcctgg gcgaaggcacccatttttatgtatagcaacgtggtgccgctgagcgcggtgatggc ggaagcggtgtttgaagaaaccccgcatggcatgcgcaaagaagcgattggcgt gctggatctgatttatgaagaaaaagcgaaagataaaatttttcagaacccgaacg gcatgggcagcggcaaatttccggtgcgcgtgccgagcggcggcgcgtttctgt atgcgctgagcgataaatttaacaacaccaacgaacatatggaaatgattgaaaaa aacgtgagctatctgctgggcgataacggcagcaaaaaaagctatgtggtgggct ttagcaaaaacggcgcgaacgcgccgagccgcccgcatcatcgcggctattatg cgaacgaaaaacgctggcgccgcagccgccgctgcagcgaaagcagccgca aagaacaggcgctgggccgctatgattgctggcgcctgtattaa

Result

The given protein sequence is converted to DNA by using Reverse Translate.

4. ScanProsite Results Viewer

This view shows ScanProsite results together with ProRule-based predicted intra-domain features

Found: 1 hit in 1 sequence

MNCRKYLLSGLAVFGLAATSAVAALSTDDYVEAA WMTTRFFGAQRSGQGPNWILDGTSNPTSFTKDSYN GKDVSGGWFDCGDHVMYGQSQGYASYVLALAYA EFTEVSTTFILVTTPTTRKPTTTPMKSGKPNKVRDLL EELRYEADFWVKAAIDGNNFVTVKGDGNADHQK WVTAGAMSKLGSGEGGEPRCITGNANDGFTSGLAA AMLAVMARVDPDTANQAKYLKAAKTAYSYAKSH KGVTNSQGFYESSWWDGRWEDGPFLAELELYRTT GENSYKTAAIDRYDNLKFSLGEGTHFMYSNVVPLS AVMAEAVFEETPHGMRKEAIGVLDLIYEEKAKDKIF QNPNGMGSGKFPVRVPSGGAFLYALSDKFNNTNEH MEMIEKNVSYLLGDNGSKKSYVVGFSKNGANAPSR PHHRGYYANEKRWRRSRRCSESSRKEQALGRYDC WRLY

Inference

Scan prosite search a given protein against prosite database to occurrence of pattern and profile.GLYCOSYL_HYDROL_F9_1 active site is found between 403 - 419 in the sequence.

5. Motif Scan Results:

	Motif Scan Results	user: anonymous
Query Protein	temporarily stored here.	
Database of motifs	PROSITE patterns, PROSITE patterns (frequent match producers), PROSITE profiles, HAMAP profiles, Pfam HMMs (local models), Pfam HMMs (global models).	
Reference	Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K &Bairoch A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.***30**:235-238	

SUMMARY

The matched sequences reported by search programs can be classified as true positives andfalse positives (the sequences missed by the program are the negatives). A true positive is a sequence that shares similarity with the query because both have evolved (diverged) from a common ancestral sequence, and is thus a true homolog. Similarity can also be sometimes attributed to evaluative convergence. A sequence is considered a false positive if the observed similarity is attributable to chance. It must be stressed that only biological arguments can enable one to decide whether a sequence should be regarded as a true or false positive. Nevertheless, a statistical analysis based on sound principles can help in the decision, because some matches are more likely to have been produced by chance than others (see statistical interpretation below). In

addition we provide match status codes to help the biologist in interpreting the matches. The matched sequences reported by search programs can be classified as true positives and false positives (the sequences missed by the program are the negatives). A true positive is a sequence that shares similarity with the query because both have evolved (diverged) from a common ancestral sequence, and is thus a true homolog. Similarity can also be sometimes attributed to evolutive convergence. A sequence is considered a false positive if the observed similarity is attributable to chance. It must be stressed that only biological arguments can enable one to decide

whether a sequence should be regarded as a true or false positive. Nevertheless, a statistical analysis based on sound principles can help in the decision, because some matches are more likely to have been produced by chance than others (see statistical interpretation below). In addition we provide match status codes to help the biologist in interpreting the matches.

The Local Pairwise Alignment of Two Sequences

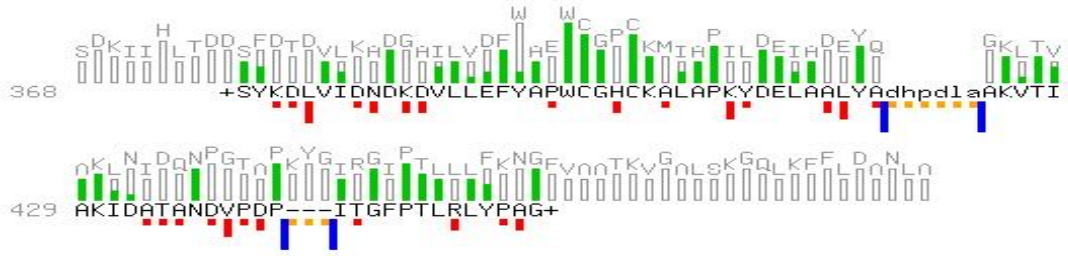
Here below, the classical text representation of a pairwise alignment of two sequences (THIO_ECOLI and PDI_ASPNG). This alignment was obtained with the Smith-Waterman algorithm, a BLOSUM62 similarity matrix and (-11/-1) for gap opening and extension penalties, respectively.

```
37.5% identity in 80 aa overlap; score: 122

                  20        30        40        50              60
THIO_ECOLI           SFDIDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQ------
GKLTVAKLNIDQNP
           :.   :.  :  .:..:.: ::: :: .::  ::.:  :       .:.:.::..
:
PDI_ASPNG
SYKDLVIDNDKDVLLEFYAPWCGHCKALAPKYDELAALYADHPDLAAKVTIAKIDATAND
           370       380       390       400       410       420

               70        80
THIO_ECOLI GTAPKYGIRGIPTLLLFKNG
             :   : :.::: :.  :
PDI_ASPNG  VPDP---ITGFPTLRLYPAG
           430          440
```

An alternative graphical representation of the same local alignment is proposed in the figure below.

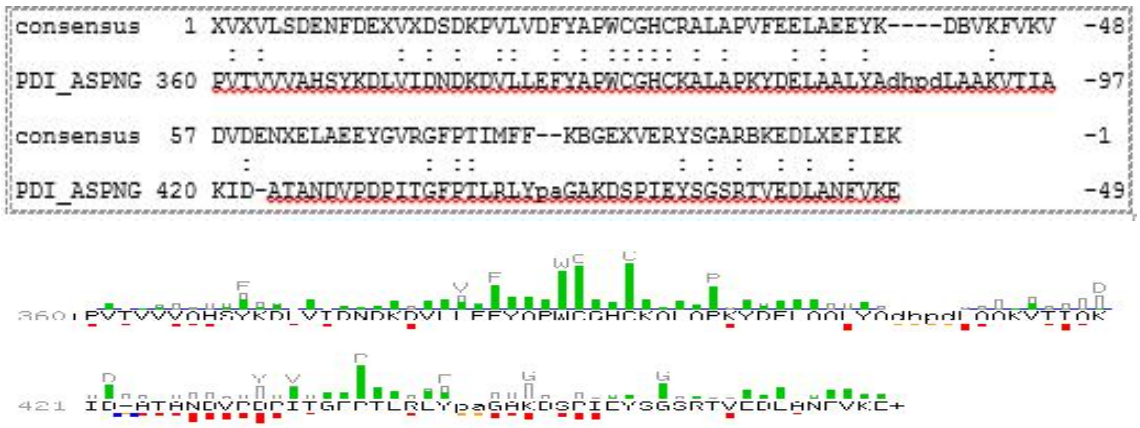


- The amino acids of the query sequence (THIO_ECOLI) are represented using the grayed residues at the top of the grayed background histogram. Hence the full length of the query sequence is shown.
- The local alignment of PDI_ASPNG on the query is represented by the sequence in black. The "+" signs at both ends of the aligned sub-sequence indicate that the alignment is local on PDI_ASPNG (the symbols "<" and ">" can be used to tag sequence extremities).
- The Smith-Waterman score (122) is proportional to the sum of the areas of the red, blue and orange rectangles. The areas of the rectangles located below the aligned sequence are negative.
- The area of every red rectangle corresponds to the score attributed by the similarity matrix to an observed pair of amino acids. The underlying gray rectangles represent the maximal score possible at every position of the query, which correspond to the diagonal elements of the similarity matrix in this example.
- Two gaps are present in this example. The first one is an insertion (relative to the query) and is represented with lowercase letters. The second one is a deletion (relative to the query) and is represented with "-".

The cost of a gap is proportional to the sum of the areas of the adjacent blue and yellow rectangles. The area of the two blue rectangles represents the "gap existence" cost which is equally divided into an opening and a closing penalty. The orange rectangles represent the costs for extending the gap.

Alignment of a Sequence on a Profile

The pairwise alignment below corresponds to the one obtained when the PDI_ASPNG sequence is searched with the THIOREDOXIN_2 profile. For the sake of the textual representation, the profile positions were symbolized by the residues of the "consensus" sequence of the multiple sequence alignment from which the profile was derived. This alignment is not fundamentally different from the one considered before but the textual representation does not reveal the additional information carried on by the profile scoring system, that eventually makes the identification by the profile so "informative". The alternative graphical representation of this alignment reveals much of this extra information.



- In strong contrast to the previous example, the scoring system is heavily position-dependent: The area of every red rectangle corresponds to the score attributed by the profile for the presence of a particular residue at a particular position. The underlying gray rectangle represents the maximal score possible at that position. The amino acids of the profile consensus that might contribute the most to the profile score are represented in gray at the top of the background histogram.
- Three gaps are presented in this example. They score differently as the system of gap penalties is also position dependent in a profile.

Two cysteines are found among the highest scoring residues of the above example. Actually they form the active site of thioredoxins. A proline residue, which is quite distant on the sequence, also rewards a particularly high score. Actually, this proline is spatially located close to the active site as shown on the figure below. Obviously, this is a case where the alignment of a sequence on a profile can provides indication for the possible function of selected residues.

Description	N-glycosylation site.
myhits	query by motif
MyHits synonyms	ASN_GLYCOSYLATION , PS00001

```
ID   ASN_GLYCOSYLATION; PATTERN.
AC   PS00001;
DT   APR-1990 (CREATED); APR-1990 (DATA UPDATE); APR-
1990 (INFO UPDATE).
DE   N-glycosylation site.
PA   N-{P}-[ST]-{P}.
CC   /SITE=1,carbohydrate;
CC   /SKIP-FLAG=TRUE;
CC   /VERSION=1;
PR   PRU00498;
DO   PDOC00001;
//
```

Description	cAMP- and cGMP-dependent protein kinase phosphorylation site.
myhits	query by motif
MyHits synonyms	CAMP_PHOSPHO_SITE , PS00004

```
ID   CAMP_PHOSPHO_SITE; PATTERN.
AC   PS00004;
DT   APR-1990 (CREATED); APR-1990 (DATA UPDATE); APR-
1990 (INFO UPDATE).
DE        cAMP- and cGMP-dependent protein kinase
phosphorylation site.
PA   [RK](2)-x-[ST].
CC   /SITE=3,phosphorylation;
CC   /SKIP-FLAG=TRUE;
CC   /VERSION=1;
DO   PDOC00004;
//
```

Inference

Motifscan is a program for finding motifs in the given sequence. The above results significantly shows some the important motif, its functions and the family where the motifs belongs to which implies the protein sequence. The result also gave the post translational modification.

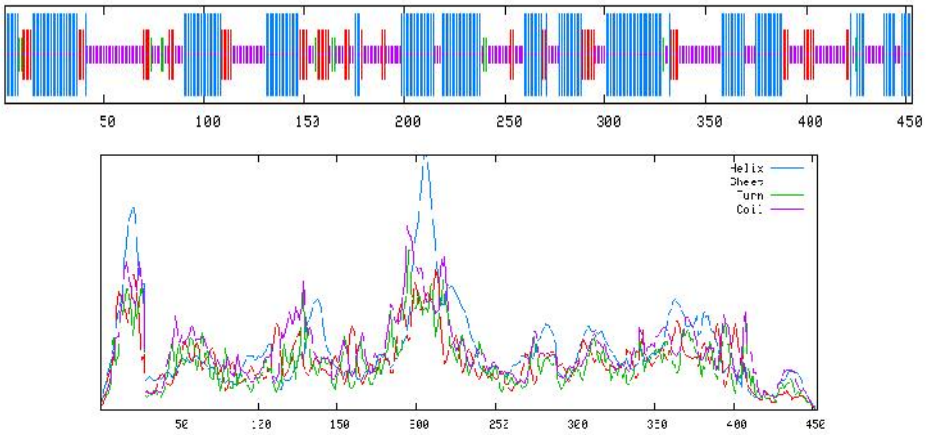
6. SOPMA result for: UNK_219630

Abstract Geourjon, C. & Deléage, G., SOPMA: Significant improvement in protein secondary structure prediction by consensus prediction from multiple Alignments. Cabios (1995) 11, 681-684

Sequence length: 453

SOPMA

Alpha helix	(Hh) :	201 is	44.37%
3_{10} helix	(Gg) :	0 is	0.00%
Pi helix	(Ii) :	0 is	0.00%
Beta bridge	(Bb) :	0 is	0.00%
Extended strand	(Ee) :	62 is	13.69%
Beta turn	(Tt) :	14 is	3.09%
Bend region	(Ss) :	0 is	0.00%
Random coil	(Cc) :	176 is	38.85%
Ambigous states	(?) :	0 is	0.00%
Other states	:	0 is	0.00%



Parameters

Window width : 17
Similarity threshold: 8
Number of states : 4

Result

1. The length of the given sequence is 453
2. The alpha helix of the given sequence in 44.30%
3. The beta strand of the given sequence in 13.69%
4. The beta turns of the given sequence in 3.09%
5. The coils of the given sequence in 38.85%
6. The output of above parameters values shows in a graphics display.

Inference

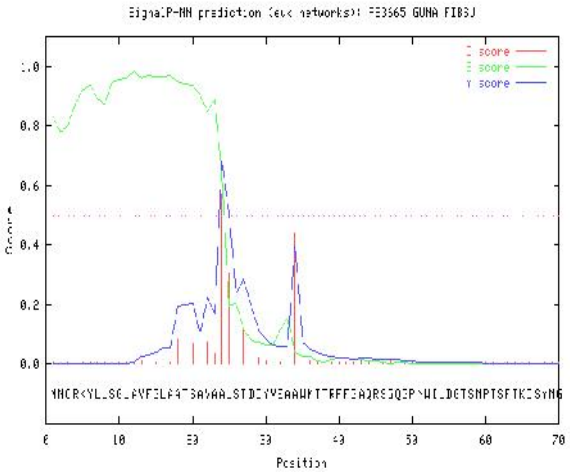
SOPMA predict secondary structure for 1ut9A and also it gives length, alpha helix, beta strand, beta turn, coils and the output of graphics display.

7. SignalP 3.0 Server - prediction results

Using neural networks (NN) and hidden Markov models (HMM) trained on eukaryotes

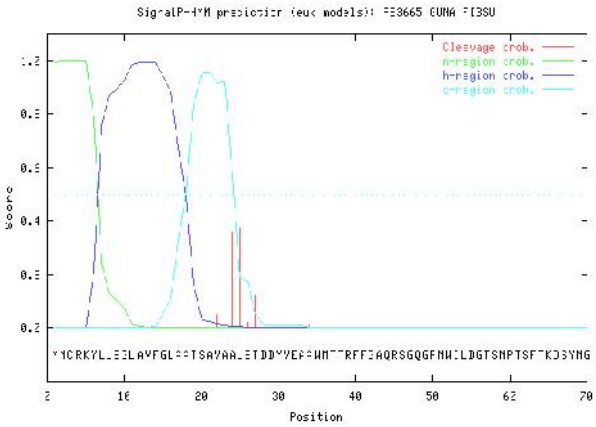
>P23665_GUNA_FIBSU Endoglucanase A - Fibrobacter succinogenes

SignalP-NN result:



data
>P23665_GUNA_FIBSU length = 70
Measure Position Value Cutoff signal peptide?
max. C 24 0.568 0.32 YES
max. Y 24 0.679 0.33 YES
max. S 12 0.985 0.87 YES
mean S 1-23 0.916 0.48 YES
D 1-23 0.798 0.43 YES
Most likely cleavage site between pos. 23 and 24: AVA-A

SignalP-HMM result



data
>P23665_GUNA_FIBSU
Prediction: Signal peptide
Signal peptide probability: 0.992
Signal anchor probability: 0.007
Max cleavage site probability: 0.375 between pos. 24 and 25

Inference

The result implies that the given protein sequence contain a signal sequence this gives a clue that the protein is both cytosolic protein by nn and by hmm. The sequence in the aminoterminal 24 or 25 amino acid from this results that our 1ut9A sequence contain any signal sequence which is in first 23 aminoacid by neural network.

TargetP 1.1 Server - prediction results

targetp v1.1 prediction results
####################################
Number of query sequences: 1
Cleavage site predictions included.
Using NON-PLANT networks.

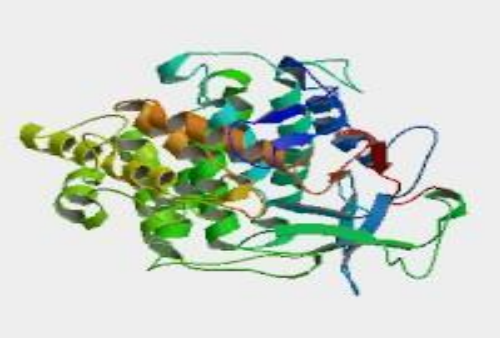
Name	Len	mTP	SP	other	Loc	RC	TPlen
P23665_GUNA_FIBSU	453	0.069	0.910	0.022	S	1	23
cutoff		0.000	0.000	0.000			

Inference

The result implies that the given protein contain a signal sequence having a probe of 0.069mTP. From this result is noted that protein contain signal sequence having destination mitochondria.

MOD BASE Result

Model Information



Sequence Model Coverage	
Sequence Identity	23.00%
E-Value	2e-58
Model Score	1.00
Target Region	6-422
Protein Length	453
Template PDB Code	1ut9A
Template Region	284-742
Dataset	SP/TR-2004
ModPipe Version	ModPipe1.0

All models for current sequence (Show filtered models only) Cross-references		
Template Structure		
PDB	1ut9	cellulose 1,4-beta-cellobiosidase: catalytic domain, residues 208-816
DBALI	1ut9A	
Jena Image Library	1ut9	
Target Sequence		
SwissProt	P23664	Endoglucanase A precursor (EC 3.2.1.4) (Endo-1,4-beta-glucanase) (Cellulase).
UniProt	P23664	
InterPro	P23664	
PFAM	P23664	
PRODOM	P23664	
SwissProt	P23665	Endoglucanase A precursor (EC 3.2.1.4) (Endo-1,4-beta-glucanase) (Cellulase).
UniProt	P23665	
InterPro	P23665	
PFAM	P23665	
PRODOM	P23665	
GenPept	121804	Guna_fibsu endoglucanase a precursor (endo-1,4-beta-glucanase) (cellulase)

10. PFAM Result:

This is the summary of UniProt entry GUNA_FIBSU (P23665).

Description:	Endoglucanase A precursor (EC 3.2.1.4) (Endo-1,4-beta-glucanase)(Cellulase).
Source organism:	Fibrobacter succinogenes (Bacteroides succinogenes). (NCBI taxonomy ID 833) ViewPfam genome data.
Length	453 amino acids

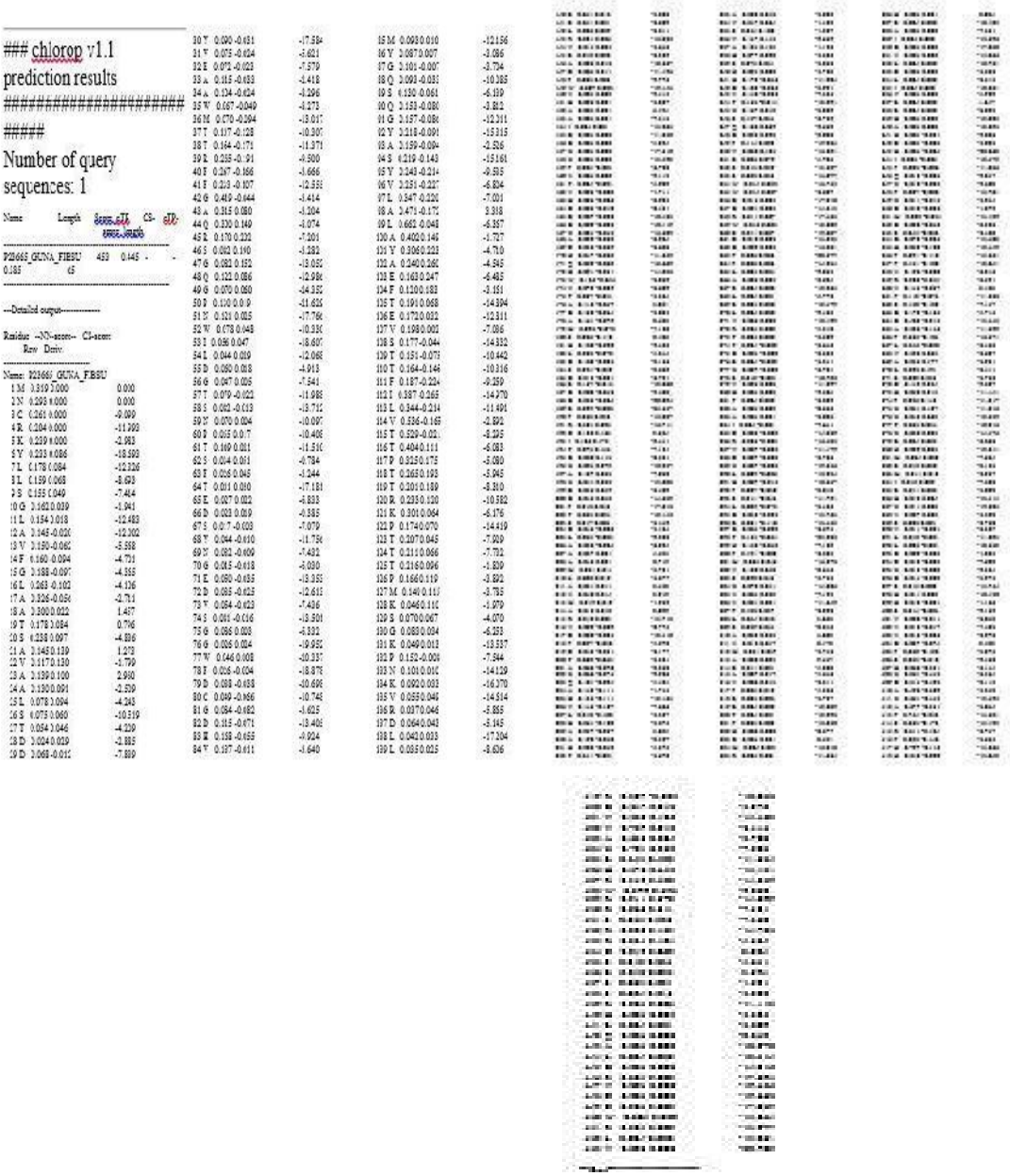
when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next*Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will

take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains. Note that some domains may be obscured by other, overlapping domains. This is noted in the table where applicable.

12. cholorop Result



INFERENCE

The result implies that the amino acid having both positive and negative CS-score. From this result that the protein sequence having the chloroP post translation modification.

REFERENCES

Ali, A.J., Ahmed, A.H., Zahraa, A. & Umar, A. (2011) Optimization of Cellulase Production by *Aspergillus niger* and *Tricoderma viride* using Sugar Cane Waste, *Journal of Yeast and Fungal Research*, 2(2), 19 – 23

Emert, G.H., Gum, E.K., Jr, Lang, J.**A.,** Liu, T.H. & Brown R.D. Jr (1974) in Food *related enzymes* (Whitaker, J. R., ed.) pp. 79-100, Amercian Chemical Society, Washington.

McAdam, Dale W. & Harry A. Whitaker (1971) "Language production: Electroencephalographic localization in the normal human brain." *Science* 172.3982 499-502.

Pradeep N V1*, Anupama1, Vidyashree K G2, Lakshmi P2 (2012)*In silico* Characterization of Industrial Important Cellulases using Computational Tools Advances in Life Science and Technology www.iiste.org ISSN 2224-7181 (Paper) ISSN 2225-062X
Vol 4, 8-15

Prashant, V.T., Uddhav, S.C., Madura, S. M., Vishal, P.D. & Renuka, R.K. (2010) Secondary Structure Prediction and Phylogenetic Analysis of Salt Tolerant Proteins, *Global Journal of Molecular Sciences*, 5 (1), 30-36.

Spano, L., Medeiros, J., Mandels, M. (1975) Enzymatic hydrolysis of cellulosic waste to glucose, pollution abatement DIV., Food SVCS. Labs. Us Army Natick, MA, USA, 7,

Whitaker, D.R. (1971) Cellulases. *The enzymes*, 5, 273-290.