Helix Vol.5: 735-738 (2015)

Structural Analysis of hypothetical protein Accession No CAI56786.1 of *Homo sapiens*

[*1]Shahin Ruby Qureshi, [2]Jyothsna Desu,
[1]Sikkim Manipal University, Mnipal, [2] BioAxis DNA Research Centre, Hyderabad
*Email: ansarruby22@gmail.com

Received: 21st July 2015, Accepted: 5th Aug 2015, Published: 1st Sep 2015

Abstract
Bioinformatics has been serving to identify many unknown sequences and bring their importance into lime light. The methodology used here is the comparative study of the unknown sequence with the known sequences in the database thereby annotating the unknown sequence. The aim of the current was to charecterize and annotates a Hypothetical human protein with accession no **CAI56786.1** obtained from NCBI database using various insilico tools and software. Several approaches have been made for predicting protein structure and function using the information derived from sequence similarity, phylogenetic profiles, protein domains etc. The classical way to infer function is based on sequence similarity using sequence database searching programs such as FASTA and PSI-BLAST, CLUSTAL W, to find out SNPs and close similarities with other species. SMART tool has been used to predict domain identification, Protparam for physical and chemical properties of protein. Secondary structure prediction was done using SOPMA, 3D structure was visualized using Rasmol software.

Keywords:
Annotation, Accession Number, PSI BLAST, SOPMA, RASMOL, NCBI

Introduction

A hypothetical protein is a protein that is predicted to be expressed from an open reading frame, but for which there is no experimental evidence of translation. Hypothetical proteins constitute a substantial fraction of proteomes of human as well as of other eukaryotes. With the general belief that the majority of hypothetical proteins are the product of pseudogenes, it is essential to have a tool with the ability of pinpointing the minority of hypothetical proteins with a high probability of being expressed.

Here, we present an *in silico* selection strategy where eukaryotic hypothetical proteins are sorted according to two criteria that can be reliably identified *in silico*: the presence of subcellular targeting signals and presence of characterized protein domains. To validate the selection strategy we applied it on a database of human hypothetical proteins dating to 2006 and compared the proteins predicted to be expressed by our selecting strategy, with their status in 2008. For the comparison we focused on mitochondrial proteins, since considerable amounts of research have focused on this field in between 2006 and 2008. Therefore, many proteins, defined as hypothetical in 2006, have later been characterized as mitochondrial.

Materials and Methods:

Sequence Retrieval:

SRS a Sequence Retrieval System is a step by step method of collecting the user specific data from the enormous data base. The Mater Data Base NCBI has been used here for the collection of the required Hypothetical sequence.

Insilico Analysis:

Various Bioinformatics tools and softwares have been used sequentially to annotate the protein based on comparative analysis.

BLAST Similarity Search:

BLAST is a basic tool from NCBI server for the pair wise alignment and comparison. It performs an Iterative BLAST using local alignment system and searches for a similar sequence in the data base specified. The parameters of BLAST include Score, E value, Query coverage, Identity and Gap.

SMART Domain Identification:

To identify the functional regions in the protein sequence termed Domains, SMRT Simple Modular Architectural Research Tool can be used. It not only provides the regions of functionality but also infers regarding the specific function and phylogeny of the sequence.

Fingerprint Scan:

Fingerprint is a conserved domain of a protein that decodes its protein family relation and evolution. An unknown protein can be identified or traced its origin using the Fingerprint Scan. It depicts the family to which a protein belongs and the patterns that categorise the protein to the specific family.

CLUSTAL W Multiple Sequence Alignment:

Clustal W tool a product of EMBL is used to compare two or more sequences at once using the Global alignment methodology. It compares the sequences for the identification of conserved regions among them. Based on the result of Clustal W a phylogenetic tree can be developed that would

enable the production of a distance tree called the Phylogenetic tree, Dendrogram or the Cladogram depicting the evolutionary relation among the studied sequences.
Here the Clustal W is not only used for the idenitification of the conservation but also for the identification of those unconserved regions or SNP's that would constitute for the evolution.

I Mutant, for analyzing the effect of SNP or unconserved regions, on the stability of the protein:
I Mutant tool enable the user to identify the effect of a single amino acid substitution on the stability of the protein. Here the tool has been employed to check for the stability changes that occur due to the presence of an SNP or unconserved region in the MSA.

Testing the efficacy of two missing fragments in hypothetical protein:
As per the results of MSA there were few unconserved domains and some missing fragments in the Hypothetical protein when compared to the other sequences under comparison. The effect of these to missing fragments in the functionality of the protein has been analyzed using comparative study of SMART and the Clustal W.

Physico chemical Characterization of the Protein using Protparam:
Protparam is an ExPasy tool used for the characterization of the protein. This tool uses the theoretical calculations and algorithenms for the characterization of the proteins.

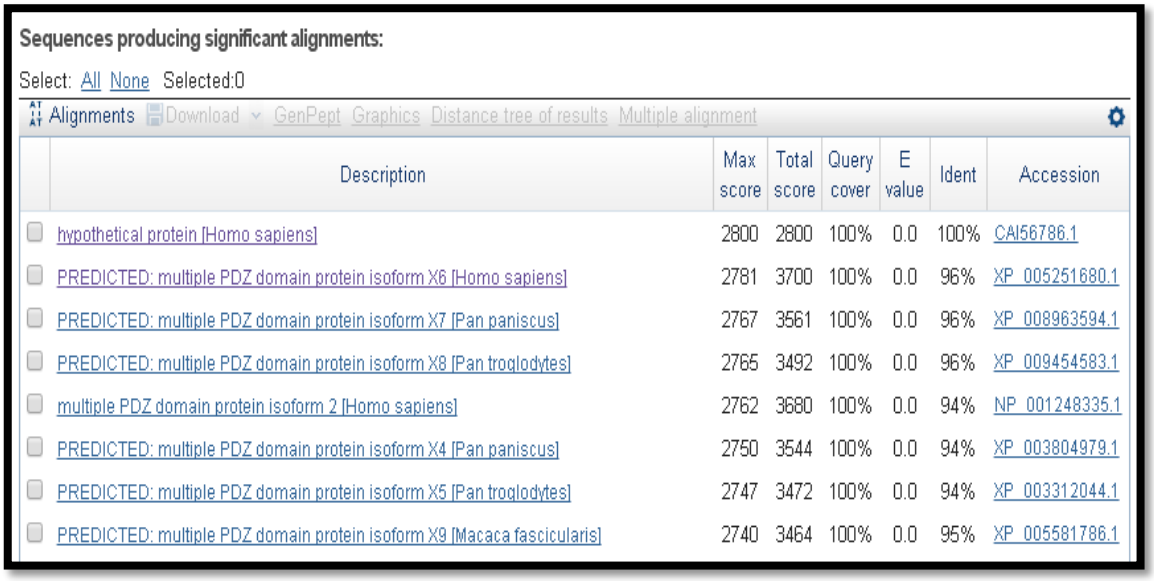
SOPMA for the secondary structural Characterization of the protein:
The tool helps to detect the secondary structural conformations in the protein sequences. Based on this data a proteins stability and solubility can be predicted. The tool would provide an insight to the secondary structural conformations present at each amino acid point in the sequence.

3D Structure Prediction of the Protein using Phyre:
This tool would perform an internal BLAST with the PDB Database so as to receive the PDB id's of those structures sharing sequence similarity of the query protein.

Results and Discussion:
Sequence Retrieval:
The Hypothetical protein CAI156786 was retrieved from NCBI data base using AND operator and a regular Sequence retrieval system (SRS). The length of the sequence was found to be 1378 aa with various domains. There were multiple PDZ domains found in the protein sequence.

BLAST Analysis for Similarity search:
To estimate the similarity between the Hypothetical sequence and the Database sequence BLAST tool has been used. The results of the BLAST have been depicted in the Fig: 1 shown below.

Fig 1: The below figure shows the BLAST output with all the Data base sequences sharing the similarity to query



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
hypothetical protein [Homo sapiens]	2800	2800	100%	0.0	100%	CAI56786.1
PREDICTED: multiple PDZ domain protein isoform X6 [Homo sapiens]	2781	3700	100%	0.0	96%	XP_005251680.1
PREDICTED: multiple PDZ domain protein isoform X7 [Pan paniscus]	2767	3561	100%	0.0	96%	XP_008963594.1
PREDICTED: multiple PDZ domain protein isoform X8 [Pan troglodytes]	2765	3492	100%	0.0	96%	XP_009454583.1
multiple PDZ domain protein isoform 2 [Homo sapiens]	2762	3680	100%	0.0	94%	NP_001248335.1
PREDICTED: multiple PDZ domain protein isoform X4 [Pan paniscus]	2750	3544	100%	0.0	94%	XP_003804979.1
PREDICTED: multiple PDZ domain protein isoform X5 [Pan troglodytes]	2747	3472	100%	0.0	94%	XP_003312044.1
PREDICTED: multiple PDZ domain protein isoform X9 [Macaca fascicularis]	2740	3464	100%	0.0	95%	XP_005581786.1

As per the above result of BLAST it can be seen that there is only 1 protein named Multiple PDZ domain Isoform OF Homo sapiens that shares good similarity with the query sequence, whereas all the other similar sequences are Predicted, Thus cannot be considered for the annotation.The Multiple PDZ domain shares the Identity of 94% with 100% query coverage and score of 2762 bits.

SMART based Domain Analysis:
SMART has been used to identify the functional domains present in the Hypothetical protein. The results show that the protein has multiple copies of the same domain PDZ, which is also called DHR (Dlg homologous region) or GLGF (relatively well conserved tetrapeptide in these domains). Some of the PDZ regions are known to bind C-terminal polypeptides; others appear to bind internal (non-C-terminal) polypeptides. Different PDZs possess different binding specificities.

Multiple sequence alignment using SDSC Biology workbench and identification of the SNPs:
The best results of BLAST showing maximum sequence similarity with the Query sequence were collected. All these sequences were further submitted for MSA using Clutsal W, for the identification of the conserved domains and the SNP's in the sequences. The SNP'S and the Missing fragments in the query sequence with respect to the collected Database sequences

The Sequences used for Comparison are:
XP_005251680.1, XP_008963594.1, XP_009454583.1, NP_001248335.1 and XP_003804979.1

From the results of ClustalW following conclusions can be made:

1) The query sequence starts from the pattern "VGHHFIR" and it does not contain a peptide from MLEAIDK to GISLEAT" compared to the other sequences in the alignment.
2) There is an amino acid mismatch in the pattern SVLPEGPV and the amino acid change includes between E and Q. The query and all the others contain E except 2 sequences.
3) There is a peptide missing from SLDLCD to PLAMW in the query and all the other sequences contain the same.
4) There is a fragment missing in the query and few other proteins considered from VDGMD to IINRPRAP

Table 1: Showing the SNP Role and their effect using I Mutant 2.0:
I-mutant tool has been used to predict protein stability and also to find alternate SNPs in comparison with the wild type.

SNP	WT A.Acid	NEW	I Mutant Stability
S/C 1270	**S**	**C**	**Decreases**
V/M 1258	**V**	**M**	**Decreases**
M/I 1173	**M**	**I**	**Decreases**
Deletion	**SLDLC—LAMW**		
T/A 496	**T**	**A**	**Decreases**
E/D 474	**E**	**D**	**Decreases**
C/S	**C**	**S**	**Decreases**
Deletion	**VDGDMDL—NRPRAP**		
E/Q	**E**	**Q**	**Decreases**

The above table shows all the variations obtained in the MSA among the selected sequences with the hypothetical protein.

Testing the efficacy of two missing fragments in hypothetical protein:
The SLDLC—LAMW region is not a domain
VDGDMDL—NRPRAP
Out of the two missing fragments the first one is not in the conserved region or the domain, thus can be neglected. The second fragments represent the Multiple PDZ domains with the accession no of NP_001248335

Physico Chemical characterization of the proteins:
To annotate the proteins physical and chemical properties Protparam has been used.

Number of amino acids: 1378
Molecular weight: 146693.1
Theoretical pI: 4.91

Ala (A) 91	6.6%
Arg (R) 56	4.1%
Asn (N) 51	3.7%
Asp (D) 74	5.4%
Cys (C) 19	1.4%
Gln (Q) 51	3.7%
Glu (E) 110	8.0%
Gly (G) 135	9.8%
His (H) 28	2.0%
Ile (I) 99	7.2%
Leu (L) 126	9.1%
Lys (K) 65	4.7%
Met (M) 26	1.9%
Phe (F) 28	2.0%

Pro (P) 77	5.6%
Ser (S) 150	10.9%
Thr (T) 73	5.3%
Trp (W) 5	0.4%
Tyr (Y) 22	1.6%
Val (V) 92	6.7%
Pyl (O) 0	0.0%
Sec (U) 0	0.0%

Total number of negatively charged residues (Asp + Glu): 184
Total number of positively charged residues (Arg + Lys): 121

As shown above the physicochemical properties of the protein state that it is an unstable protein with polar nature and the length of 1378 amino acids and the theoretical Iso electric point of 4.91. These properties would be helpful for designing the laboratory protocol for the extraction and purification of the protein.

Secondary Structural Confirmations of the Protein using SOPMA:
To analyze the secondary structural Confirmations of the hypothetical protein SOPMA has been used.
The results are depicted in the Fig 2
Fig 2: Secondary Structural Confirmations of the Protein using SOPMA:

```
SOPMA :
   Alpha helix      (Hh) :   311 is  22.57%
   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) :   350 is  25.40%
   Beta turn        (Tt) :   152 is  11.03%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :   565 is  41.00%
   Ambiguous states (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```

The above result shows that the Hypothetical protein contains 22.57% of Alpha helix confirmation, 25.40% Extended strand, 11.03% of Beta Turn and 41% of random coil confirmation. This shows that the protein has a maximum of random coil confirmation. The protein is hydrophilic.

Conclusion:
In the above work the structure analysis of hypothetical protein (homo sapiens) CAI56786 was performed, Structure of hypothetical protein may provide a hint for their biochemical or biophysical functions. 3D structure can aid the assignment of function for the characterization ofprotein.The multiple sequence alignmentwith the query sequence shows conserved as well as some SNPs regions which are in significant. The structure of protein shows multiple PDZ domains which can be predicted as a significant functional protein.

References

1. Oxford Journals, Nucleic Acid Research, 'Conserved hypothetical' proteins: prioritization of targets for experimental study', Michael Y. Galperin and Eugene V. Koonin*
2. International Journal Of Molecular Science, 'Function Prediction and Analysis of Mycobacterium tuberculosis Hypothetical Proteins', Gaston K. Mazandu and Nicola J. Mulder *
3. UniProtKB - O75970 (MPDZ_HUMAN), http://www.uniprot.org/uniprot/O75970
4. http://blast.ncbi.nlm.nih.gov/Blast.cgi
5. http://smart.embl-heidelberg.de/
6. Biochemistry and Molecular Biology Education. Volume 29, Issue 4, pages 165–166, July 2001. Additional Information. **How to Cite**. Jakobsson, E. (2001), http://seqtool.sdsc.edu/
7. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, and Stockinger H. ExPASy: SIB bioinformatics resource portal, Nucleic Acids Res, 40(W1):W597-W603, 2012
8. Bennett-Lovsey RM, Herbert AD, Sternberg MJE, Kelley LA. Proteins: Structure, Function, Bioinformatics, vol 70, 3, (2008).
9. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue): W306-10. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Capriotti E, Fariselli P, Casadio R.

