In Silico Prediction of an Unknown Function of a Protein Using Bioinformatics Tools

[1]Dr.T. V. Sai Krishna, Dr. [2]A. Yesubabu, [3]Dr. Deepak Nedunuri, [4]Ch. Madhava Rao
[1]*Professor & Head, Department of CSE, QIS Institute of Technology, Ongole, Andhra Pradesh.*
[2]*Professor & Head, Department of CSE, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh.*
[3]*Associate Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh.*
[4]*Assistant Professor, Department of ECE, Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh.*

Abstract:-Understanding the function of genes is an important aspect of genomics. An unknown functional protein was selected from PDB database. X-ray crystal structure of protein *Escherichia coli*, 1U5W having unknown function was randomly chosen to study the functional aspects. *Escherichia coli* is one of the best-studied prokaryotic model organisms. FASTA sequence of the organism was taken from the pdb file and is analyzed using BLAST. Data bases such as non-redundant protein (nr), reference proteins (refseq protein), swiss protein sequences (Swissprot), patented protein sequences (pat), protein data bank proteins (pdb) and environmental samples (env nr) are used, where, identities and similarities were observed from non-redundant protein sequences (98%-36%) and reference proteins (98%-34%). Other databases were not considered as they reported low similarities (48%-34%). The BLAST analysis showed maximum (that is above 50%) similarity with NTPase protein sequences in each database. Hence, it can be stated that the 1U5W protein belongs to NTPase family.

KeyWords: Domain of Unknown Function, NTPase, PDB, BLAST

I. INTRODUCTION

The sequence of a genome contains the plans of the possible life of an organism, but implementation of genetic information depends on the functions of the proteins and nucleic acids (1). Many individual proteins of known sequence and structure present challenges to the understanding of their function. Whole-genome sequencing projects are a major source of proteins of unknown function. 3D structure can aid the assignment of function; motivating the challenge of structural genomics projects to make structural information available for novel uncharacterized proteins (2). Nevertheless, prediction of protein function from sequence and structure is a difficult problem, because homologous proteins often have different functions. Many methods of function prediction rely on identifying similarity in sequence and/or structure between a protein of unknown function and one or more well-understood proteins. Alternative methods include inferring conservation patterns in members of a functionally uncharacterized family for which many sequences and structures are known (3).

Genome sequencing projects are producing linear amino acid sequences, but full understanding of the biological role of these proteins will require knowledge of their structure and function (4). Although experimental structure determination methods are providing high-resolution structure information about a subset of the proteins, computational structure prediction methods will provide valuable information for the large fraction of sequences whose structures will not be determined experimentally. There are now plenty of proteins, which have a totally unknown function. Most often, only the sequence of the protein is known, but there are also hundreds of protein structures of unknown function, which are provided by the structural genomics centers (5). Sometimes the proteins come from prokaryotes where the operons make it possible to infer the function of a protein from its genomic context, but this is more complicated in eukaryotes. Bioinformatics analyses of whole genome sequences highlight the problem of identifying the biochemical and cellular functions of many gene products that are at present uncharacterized (6). Determination of their three-dimensional structures, either experimentally or by prediction, provides a powerful tool to address function, since it is at this level that biological activity is expressed (7).

Many of the protein domains have unknown function. But these protein domains of unknown function participate in metabolic pathways of an organism and cause adverse effects. Sometimes the function of the protein may change due to mutations like insertions, deletions and substitutions. The main objective of the paper is to predict the protein domain of unknown function and its classification using Bioinformatics tools.

II. MATERIALS AND METHODS

Selection of Hypothetical Protein

Hypothetical protein is a protein whose existence has been predicted, but for which there is no experimental evidence that it is expressed in vivo. These are searched in a protein structure database, Protein Data Bank (8).

Similarity Search

To predict the function of the query protein, similarity search was carried out and by this search the proteins that may exhibit sequence or structural similarity with the hypothetical protein were performed using NCBI BLASTP similarity program (9). BLAST (Basic Local Alignment Search Tool) is a program for sequence similarity searching developed at NCBI and is instrumental in identifying genes and genetic features.

The FASTA sequence of the 1U5W protein was given as the query sequence and was searched for the similar proteins in different databases using BLASTP program. The FASTA sequence must start with greater than (>) symbol with pdb id. The databases used for similarity search were:

- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- Swiss prot protein sequences (swissprot)
- Patented protein sequences (pat)
- Protein databank proteins (pdb)
- Environmental samples (env_nr)

Sequence Analysis

Pair wise alignments followed by multiple sequence alignments were carried out between the hypothetical protein and various proteins from database. For this purpose the CLUSTALW2 from EBI (10) was used.

III. RESULTS AND DISCUSSION

A general search with 'hypothetical protein' as keyword in PDB resulted 1698 structure hits. Further, filtration was carried by considering the following criteria: Structural genomics projects; Experimental methods (X-ray diffraction), where 174 hits are obtained. Later, a sequence length between 180 to 220 amino acid residues and x-ray resolution between 2.0 to 2.5 A0 resulted in 13 structure hits. Among these 13 proteins, 1U5W protein was randomly selected (Figure 1).

The sequence of 1U5W protein with a chain length of 184 residues has 38% helical and 20% beta sheet composition. Fasta sequence is given below.

>1U5W:A|PDBID|CHAIN|SEQUENCE
MLIMHQVVCATTNPAKIQAILQAFHEIFGEGSCHIASVA
VESGVPEQPFGSEETRAGARNRVANARRLLPEADFWV
AIEAGIDGDSTFSWVVIENASQRGEARSATLPLPAVILE
KVREGEALGPVMSRYTGIDEIGRKEGAIGVFTAGKLTR
ASVYHQAVILALSPFHNAVYSGRVEHHHHHH

The above sequence was subjected to BLAST and the scores and identities of biological sequences from each data base with 1U5W protein are tabulated. The proteins from different databases that have specific function which has similarity with those of the hypothetical protein were selected. More similarities were observed from non-redundant protein sequences and reference protein sequences. They are given in Table 1.

NCBI was used to obtain entire details of resulted biologically similar proteins with their FASTA sequences when their protein id was given for search. From the above table it is evidenced that NTPase displayed more similarity with 1U5W and hence all the NTPase fasta format sequences are selected to perform multiple alignments. The output is shown in Figure 2.

From the above analysis of protein 1U5W NTPase protein appeared more number of times from non redundant and reference protein databases. Hence from this analysis it can be observed that 1U5W belongs to NTPase protein as it showed more identity.

Figures and Tables

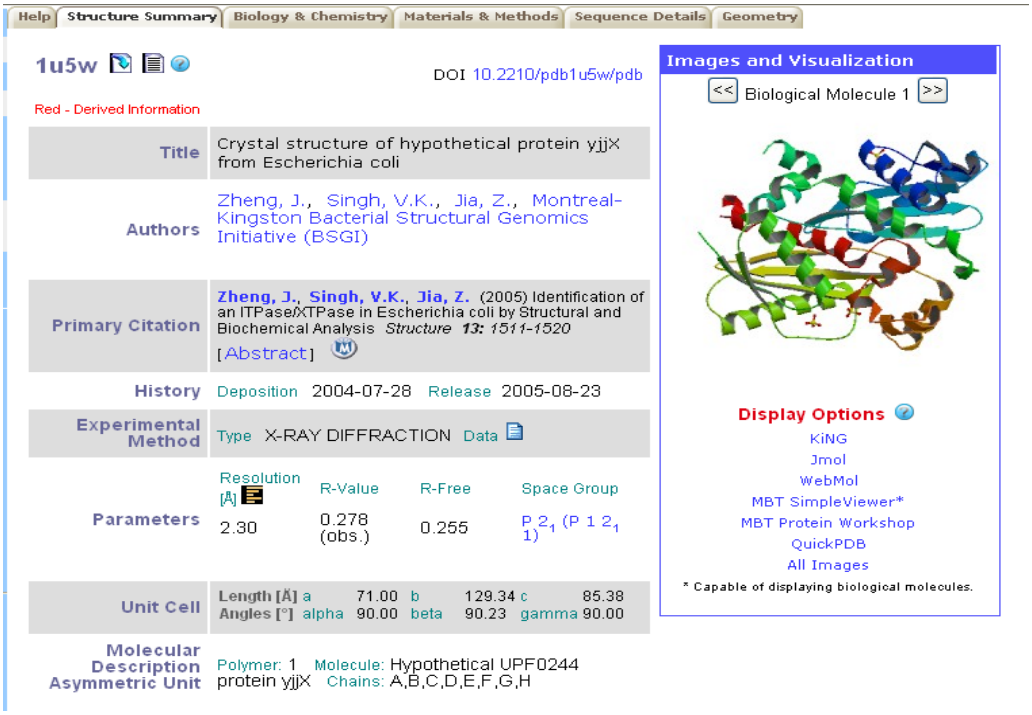

Figure 1: 1U5W protein with unknown function.

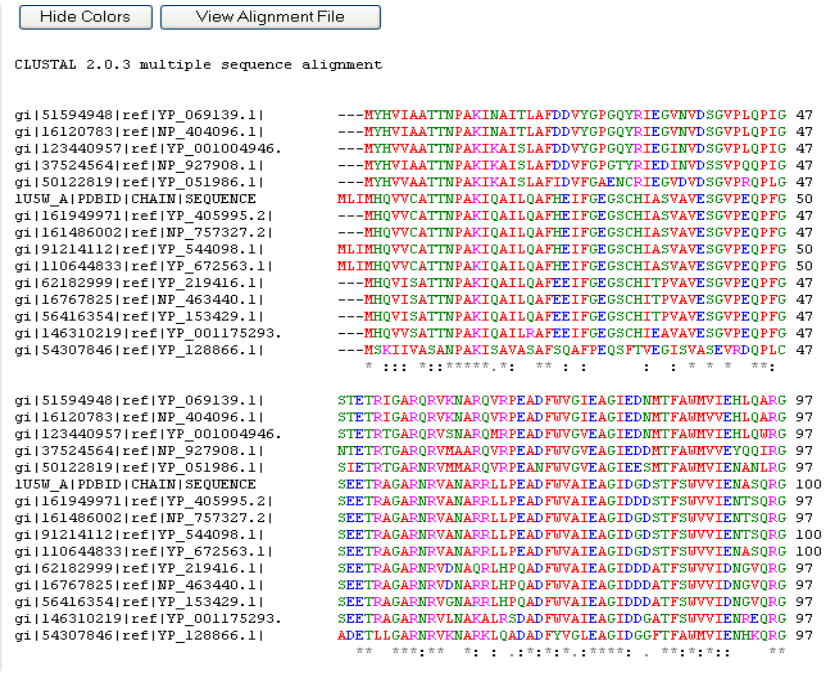

Figure 2: Multiple alignments of 1U5W with NTPase proteins.

Table 1: Results of BLASTp analysis showing biologically similar sequences with 1U5W.

PROTEIN	LENGTH	SCORE	E-value	IDENTITIES	POSITIVES	GAPS
NTPase protein	271	229 bits (583)	2e-58	124/270 (45%)	175/270 (64%)	2/270 (0%)
NTPase	171	313 bits (802)	3e-84	151/170 (88%)	161/170 (94%),	0/170 (0%)
NTPase	171	308 bits (790)	7e-83	149/170 (87%)	160/170 (94%),	0/170 (0%)
Phosphopantetheine adenylatetransferase protein	328	58.2 bits (139)	2e-08	53/182 (29%)	84/182 (46%)	20/182 (10%)
Hepatocyte growth factor receptor protein	1381	31.6 bits (70)	1.9	23/68 (33%)	33/68 (48%)	8/68 (11%)
Cysteine rich secretory protein	500	30.0 bits (66)	5.9	17/56 (30%)	27/56 (48%)	1/56 (1%)
3-methyl-2-oxybutanoate hydroxymethyl transferase protein	279	31.2 bits (69),	2.6	16/28(57%)	19/28 (67%)	2/28 (7%)

IV. CONCLUSION

Functional genomics is a field of molecular biology that attempts to make use of the vast wealth of data produced by genomic projects (such as genome sequencing projects) to describe gene and protein functions and interactions. Here, in this study an attempt was made to find the function of an unknown protein 1U5W. Blast P analysis of 1U5W sequence against non-redundant, reference protein, swiss prot protein sequences patented protein sequences, protein databank proteins, environmental samples, databases revealed 98% homology with NTPase protein, 57% with 3-methyl-2-oxobutanoate hydroxymethyl transferase, 33% with hepatocyte growth factor receptor and 29% with phosphopantethene adenyltransferase respectively. Based on this analysis multiple alignments conducted on all these proteins along with 1U5W sequence showed maximum number of residue conservations (reported as stars in alignment) with NTPase protein. Hence, this computational analysis and application of methodology in predicting the function of an unknown protein, 1U5W as NTPase protein justifies the utility of bioinformatics tools in recognizing functional aspects of a protein.

V. REFERENCES

[1]. Enault, F., Suhre, K., Abergel, C., Poirot, O., & Claverie, J. M. (2003). Annotation of bacterial genomes using improved phylogenomic profiles. Bioinformatics, 19(suppl 1), i105-i107.
[2]. Hawkins, T., Luban, S., & Kihara, D. (2006). Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Science, 15(6), 1550-1556.
[3]. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenomic profiles. Proc. Nat Acad. Sci. USA, 96, 4285–4288.
[4]. Jensen, L. J., Gupta, R., Staerfeldt, H. H., & Brunak, S. (2003). Prediction of human protein function according to Gene Ontology categories. Bioinformatics, 19(5), 635-642.
[5]. Minion, F.C., Lefkowitz, E.J., Madsen, M.L., Cleary, B.J., Swartzell, S.M., Mahairas, G.G., 2004. The genome sequence of Mycoplasma hyopneumoniae strain 232, the agent of swine mycoplasmosis. J. Bacteriol. 186, 7123–7133.
[6]. Chou, K.C., 2002. Prediction of protein signal sequences. Curr. Protein Pept. Sci. 3, 615–622.
[7]. Zheng,Y., Roberts,R.J. and Kasif,S. (2002) Genomic functional annotation using co-evolution profiles of gene clusters. Genome Biol., 3, 121–129.
[8]. www.rcsb.org/pdb
[9]. www.ncbi.nlm.nih.gov/blast
[10]. www.ebi.ac.uk/Tools/msa/clustalw2