International Journal of Biotechnology and Biomedical Sciences
p-ISSN 2454-4582, e-ISSN 2454-7808, Volume 2, Issue 3; July-December, 2016 pp. 211-215

http://www.krishisanskriti.org

Characterization of a Novel Human Protein Expected to be Involved in Signal Transduction

Namrata Pandey[1], P[r]atibha Sharma[2], Rashmi[3], Navodita[4], Poonam Sharma[5] and Jasvinder Kaur[6]

[1,2,3,4,5,6]Gargi College University of Delhi
E-mail: [1]namratapandey3897@gmail.com, [2]sharmapratibha562@gmail.com, [3]rashmijolly597@gmail.com, [4]navodita.sood@gmail.com, [5]poonamsharma.gargi@gmail.com, [6]jasvinder1511@gmail.com

Abstract—*To understand the functions of proteins at a molecular level, it is often necessary to determine their three-dimensional structure as structure dictates function. The sequence of a protein reveals much about its evolutionary history. Protein structures are more conserved than the amino acid sequences. Hence, it would be easier to determine function by comparing structures that are conserved and to determine phylogenetic relationships between diverse species. Functions of the vast majority of proteins - found in humans and elsewhere - remain entirely unknown. Understanding the proteome is acquired by investigating, characterizing, and cataloging proteins. Computational predictive models and mathematical algorithms helps in deciphering structure and functions of uncharacterized proteins. C16orf70 protein from human (UniProt ID: Q9BSU1) belongs to UPF0183 family of proteins. This protein family is yet to be characterized completely. The protein is widespread in occurrence, localized in kidney, respiratory, digestive & reproductive system. The sequence of the protein was procured from public databases and the protein was structurally characterized using bioinformatics tools and a series of prediction models. The protein was found to be conserved from Caenorhabditis elegans to Arabidopsis thaliana. Nine N-Myristoylation sites and various casein kinase and tyrosine kinase phosphorylation sites in the protein point towards its likely role in signal transduction.*

1. INTRODUCTION

The evolution of life has diversified the life forms from single organism to an everlasting list of organisms, with one macromolecule playing a vital role throughout. Protein, the harbinger of ever increasing complexity and functionality of cell, has the crucial role of maintaining the cell's existence. To understand the functions of proteins at a molecular level, it is often necessary to determine their three-dimensional structure as structure dictates function. Polypeptide chains with similar amino acid sequence but different structural conformation can have a different function altogether. This is quite evident from the fact that misfolded or improperly folded proteins are a cause of many disorders including Alzheimer disease, Parkinson's disease, Huntington disease, and transmissible spongiform encephalopathies (prion disease). Also, information on protein structure is a key to drug discovery. Proteins resemble one another in amino acid sequence only if they have a common ancestor. The sequence of a protein reveals much about its evolutionary history. Protein structures are more conserved than the amino acid sequences. Hence, it would easier to determine function by comparing structures that are conserved and to determine phylogenetic relationships between diverse species.

Humans have genes encoding about 20,500 proteins. Functions of the vast majority of proteins - found in humans and elsewhere - remain entirely unknown. Only a small fraction of these proteins have been studied intensely enough to be well understood. Understanding the proteome is acquired by investigating, characterizing, and cataloging proteins. Bioinformatics is very instrumental in organizing and analyzing the large amount of proteomics data collected with the help of high throughput technologies. Computational predictive models and mathematical algorithms helps in deciphering structure and functions of uncharacterized proteins.

2. METHODOLOGY

The sequence of the uncharacterized protein under study was derived from UniProt (UniProt ID Q9BSU1). Basic protein features were procured using ProtParam and Prot Scale tools of ExPasy server [8]. Protein post-translational modifications were predicted using the protein motif scanner ScanProsite and Motif Scan [9]. Motif scan results were obtained using Pfam HMMs that predicted local models using pfam_fs. Kyte-Doolitle hydropathy plot was generated using the Kyte-Doolitle program of UVafasta server [10]. Sequence analysis; structure and function prediction was done using PredictProtein server [11]. Protein three dimensional structure was obtained from PyMol molecular visualization tool. Potential cleavage sites of proteases were predicted using Peptide Cutter tool of ExPasy server [8]. Transmembrane helix prediction was done using the TMpred tool of EMBnet-Server [13]. Functional domain prediction was done using ProFunc tool of EMBL-EBI [14]. ProBis Server [15] was used to detect structurally similar binding sites. Local alignment was performed using cutoff Z score of 1.0. The PTMs were identified by a variety of database of motifs ROSITE patterns

(frequent match producers) [freq_pat], HAMAP profiles [hamap], PeroxiBase profiles [perox], More profiles [pre], PROSITE profiles [prf] [8].

3. RESULTS AND DISCUSSION

The protein under study (C16orf70) is widespread in occurrence, it has been reported to be localized in hepatocytes, gallbladder cells, pancreatic exocrine glandular cells, skin, skeletal muscle, esophagus, duodenum, glandular cells of small intestine, colon and rectum, glandular cells of female reproductive system, thyroid, parathyroid and adrenal glands, respiratory epithelial cells in nasopharynx and bronchus and also myocytes in heart muscle (human protein atlas).

C16orf70 protein from human (UniProt ID: Q9BSU1) belongs to UPF0183 family of proteins. This protein family is yet to be characterized completely and contains proteins from diverse species like Caenorhabditis elegans to Arabidopsis thaliana. The protein has two isoforms 1 and 2. It is the Isoform-1 that been chosen as 'canonical form'. A canonical form of a protein is a sequence of amino acids that reflects the most common choice of base or amino acid at each position i.e., the most prevalent and most similar to orthologous sequences found in other species. Hence, it allows the clearest description of domains, isoforms, polymorphisms, post-translational modifications, etc. It has a length of 422 amino acids and a molecular weight of 47.5 kDa.

The locus is human chromosome 16: 67143861 – 67182442.

1. Characterization of the protein based on sequence

Theoretical pI is 7.64. A total of 39 negatively (Asp + Glu) and 40 positively charged residues (Arg + Lys) are present. Extinction coefficient (M-1 cm-1 at 280 nm measured in water) of the protein is 40965 assuming all pairs of Cys residues form cystines; and 40340 assuming all Cys residues are reduced. Further, the instability index (II) was computed to be 48.29. This classified the protein as unstable.

The protein has an aliphatic index of 82.46. The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins.

2. Possible cleavage sites

The number of cleavage sites were in the order maximum for the enzyme Proteinase K (198) followed by pepsin (139); thermolysin (124); chymotrypsin (109). Moderate number of sites was obtained for clostripain, endopeptidases, formic acid, glutamyl endopeptidase, staphylococcal peptidase I, CNBr and trypsin. Very few sites were available for action for Iodosobenzoic acid, BNPS-Skatole and caspase 1.

3. Motif scan results

i. N-Myristoylation and N-Glycosylation sites

ScanProsite predicted nine N-myristoylation sites (22-27; 93-98; 118-123; 133-138; 194-199; 207-212; 221-226; 340-345; & 381-386) and two N-Glycosylation sites (59-62; 375-378) in the protein. N-Myristoylation is found in proteins involved in signal transduction pathway; few integral membrane proteins, and few disease related proteins. Many N-myristoylated proteins are membrane bound and can be found in the plasma membrane or other intracellular membranes in eukaryotic cells. Myristate hydrophobically inserts into the lipid bilayer, and approximately 10 of the 14 carbons penetrate the hydrocarbon core of bilayer. Additional protein-protein interactions may occur at the plasma membrane that serve to preferentially enhance the binding of certain N-myristoylated proteins to the plasma membrane [1].

Whereas, N-Glycosylation has role in trafficking of apical membrane proteins in epithelia or can be required in apical sorting of glycoproteins [3]. N-linked carbohydrates play roles in diverse biological processes such as protein folding and conformation, stability, and targeting to subcellular and extracellular sites, as well as cell-matrix and cell-cell interactions. Most membrane proteins targeted to the plasma membrane possess N-linked glycans. Glycosylation serves as an important apical membrane trafficking signal that has been well acknowledged. The developmental importance of N-glycosylation is reflected in such functions as morphogenesis, proliferation, differentiation, and apoptosis [2].

ii. Phosphorylation

Four casein kinase II phosphorylation sites (142-145; 178-181; 244-247; & 349-352); two protein kinase C phosphorylation sites (167-169; 301-303;) and one tyrosine kinase phosphorylation site at position 233-240.

iii. NHL repeats and Phosphatidylinositol-specific phospholipase X-box domain profile

PI-PLC X-Box (138-156) and NHL-repeat with a low confidence level (298-310) was predicted. The X-box is an important region for catalytic activity of PI-PLCs. Many residues are highly conserved in these domains of all eukaryotic PI-PLCs.

NHL Repeats is an amino acid sequence found in protein. These are found in Serine/ Threonine protein Kinases which are transmembrane receptor with an intra-cellular N-terminal Kinase domain and an extra-cellular C-terminal sensor domain. However, no transmembrane domains were found in the protein. The occurrence of the NHL domain in a variety of proteins – regulatory and nonregulatory, cytoplasmic and nuclear suggests a general function for this domain – for example, in protein–protein interactions.

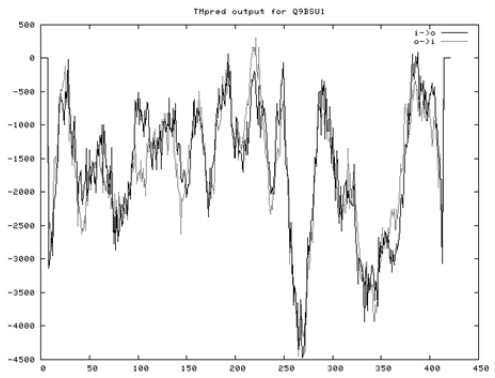

Fig. 1: No transmembrane domains were found.

The protein C16Orf70 or Lin10 shares 98% and 97% sequence similarity to the homologs from *Rattus norvegicus* and *Mus musculus*. 54% identity to *Drosophila melanogaster* with 98% query coverage. 37% identity to *Caenorhabditis elegans* with 95% query coverage. 34% similarity with 92% query coverage to *Arabidopsis thaliana*. Thus, the protein was found to be relatively conserved across various species.

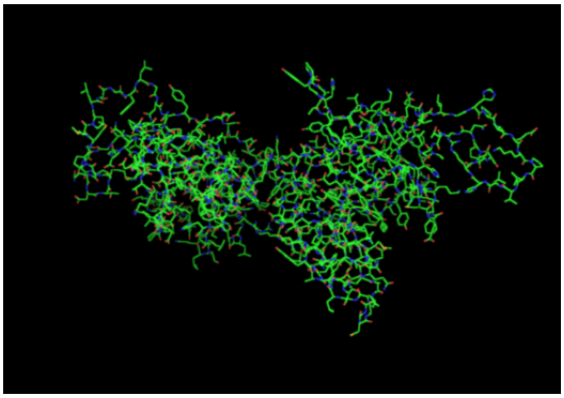

Fig. 2a: 3-D structure of the protein C16orf70. The blue bars depict the amino acid side chains.

A Ramachandran plot was constructed for the 3-D structure of the protein [12]. It showed the darkest areas (here shown in red) as "core" regions representing the most favorable combinations of phi-psi values.

Ideally, one would hope to have over 90% of the residues in these "core" regions. The percentage of residues in the "core" regions is one of the better guides to stereochemical quality. The different regions on the Ramachandran plot are as described in [7]. The two most favoured regions are the "core" and "allowed" regions which correspond to 10° x 10° pixels having more than 100 and 8 residues in them, respectively.

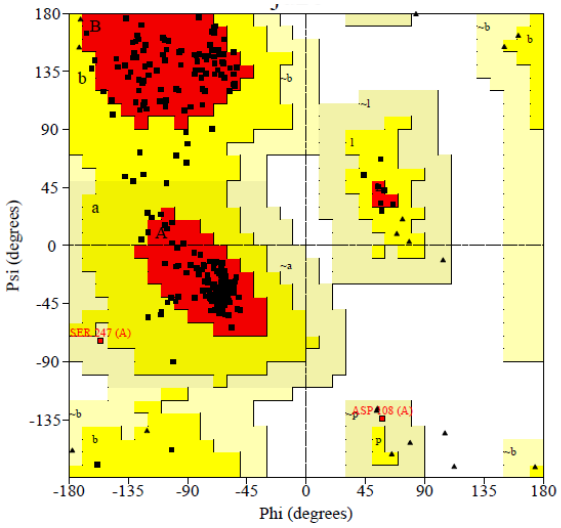

Fig. 2b: Ramachandran plot of the predicted 3-D structure. Glycine residues are separately identified by triangles. The regions are labelled as follows: A-Core alpha L-Core left-handed alpha; a-Allowed alpha; l- Allowed left-handed alpha; ~a-Generous alpha; ~l- Generous left-handed alpha; B-Core beta; p- Allowed epsilon; b-Allowed beta; ~p-Generous epsilon; ~b- Generous beta

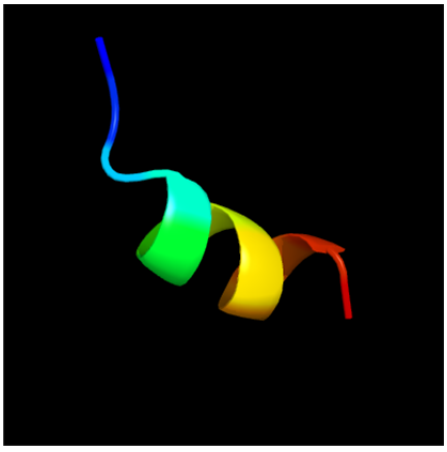

Fig. 3: N → C terminus domain model dimensions (Å): X: 12.810 Y: 20.741 Z: 12.977 predicted with 8 % confidence and 3 % coverage generated by profunc.

Fold: DNA/RNA-binding 3-helical bundle

Superfamily: RNA polymerase subunit RPB10

Family: RNA polymerase subunit RPB10

A Zn-binding site was predicted near the N-terminus of the protein. Structural determination of this subunit reveals a zinc-bundle topology, consisting of three alpha-helices stabilized by a zinc ion. Zinc finger (Znf) domains are relatively small protein motifs which contain multiple finger-like protrusions that make tandem contacts with their target molecule. Some of

these domains bind zinc. These domains are recognized to bind DNA, RNA, protein and/or lipid substrates. Those that bind to phosphatidylinositol 3-phosphate are often found in proteins targeted to lipid membranes that are involved in regulating membrane traffic. Some zinc finger domains target proteins to endosomes by binding specifically to phosphatidylinositol-3 phosphate at the membrane surface. This might be related to the presence of caspase I cleavage sites on the protein.

The *C. elegans* Lin-10 protein was originally found to be a critical component of the polarized trafficking machinery in roundworm epithelium cells [4]. Originally Lin-10 was thought to be unrelated to previously identified proteins. However recent work has reassigned the product of the lin-10 gene as a homologue of the X11 family of proteins. X11a, mLin-2, and mLin-7 interact both biochemically and genetically and likely control protein targeting in an evolutionarily conserved fashion [5]. All X11 family members have conserved PTB and PDZ domains but divergent amino termini. The PDZ domain is comprised of roughly 100 amino acids, which form a modular pocket that interacts with a ligand—usually the C-terminal tail of a protein. The finding that PTB domains can bind to their target peptides in a phosphor-tyrosine-independent fashion indicates that these domains can be involved in diverse cellular functions, not just signaling downstream of tyrosine kinases. This suggests a possible divergence between worm and mammalian epithelia because, in worm epithelia, the X11a homologue, Lin-10, is crucial for basolateral targeting. The heterotrimeric complex contains several protein-protein interaction domains that would be useful to contact a large number of different proteins. In addition, the presence of the PTB domain, a domain that can bind to beta turn motifs, might make X11 proteins particularly suitable for detecting trafficking signals with tyrosine-based motifs. PDZ domain proteins have been demonstrated to play a role in receptor and channel clustering at synaptic junctions. PDZ domains have also been implicated as being important in protein targeting to specific membrane surfaces. A PDZ point mutation enhances surface delivery of exogenous glutamate receptors in transfected neurons, suggesting that mLin-10 may regulate AMPA receptor trafficking in vivo [6].

4. CONCLUSION

C16orf70 protein from human (UniProt ID: Q9BSU1) belongs to UPF0183 family of proteins. This protein family is yet to be characterized completely and contains proteins from diverse species like *Caenorhabditis elegans* to *Arabidopsis thaliana*. The protein is widespread in occurrence, localized from kidney, respiratory, digestive to reproductive system. The instability index (II) was computed to be 48.29**.** This classified the protein as unstable. The protein has an aliphatic index of 82.46. It may be regarded as a positive factor for the increase of thermostability of the protein. Nine N-Myristoylation sites and various casein kinase and tyrosine kinase phosphorylation sites in the protein point towards its possible role as a signal transduction protein. No transmembrane domains were however found. The protein was found to be relatively conserved across various species starting from *C. elegans* to *Homo sapiens* to *Arabidopsis thaliana*. The predicted tertiary structure saw 98 % of the residues falling in favored regions in the Ramachandran plot and only about 3 % in outlier regions. A DNA/RNA-binding 3-helical bundle was predicted in the protein with just 8 % confidence and 3 % coverage. Zn-binding site is near the N-terminus. Structure determination of this subunit reveals a zinc-bundle topology, consisting of three alpha-helices stabilized by a zinc ion. Since the protein is a homologue of the X11 family of proteins, it may be involved in protein targeting or signal transduction.

5. ACKNOWLEDGEMENTS

The work was supported by Gargi College, University of Delhi and DBT STAR college grant.

REFERENCES

[1] Resh, M. D., “Fatty acylation of proteins: new insights into membrane targeting of myristoylated and palmitoylated proteins”, *Biochimica et Biophysica Acta*, 1451, 1, September 1999, pp. 1-16

[2] Vagin, O., Kraut, J. A., Sachs, G., “Role of N-Glycosylation in trafficking of apical membrane proteins in epithelia”, *Am J Physio/ Renal Physio*, 296, 3, March 2009, pp. F459-F469

[3] Cooper, G. M., Hausman, R. E. “Normal N-Glycosylation is required for apical sorting of glycoproteins”, *Cell- a molecular approach*, 4th edition, ASM Press

[4] Kim, S. K., Horvitz, H. R., “The Caenorhabditis elegans gene lin-lO is broadly expressed while required specifically for the cell fates”, 4, 1990, *Genes & Development,* pp. 357-371.

[5] Borg, J-P., Straight, S. W., Kaech, S. M., Borg, M. T., Kroon, D. E., Karnak, D., Turner, R. S., Kim, S. K., Margolis, B., “Identification of an Evolutionarily Conserved Heterotrimeric Protein Complex Involved in Protein Targeting”, *J. Biol. Chem.,* 273, 48, November 1998, pp. 31633-31636.

[6] Stricker, N. L., Huganir, R. L., “The PDZ domains of mLin-10 regulate its trans-Golgi network targeting and the surface expression of AMPA receptors”, *Neuropharmacology*, 45, 6, November 2003, pp. 837–848.

[7] Morris, A. L.; MacArthur, M. W.; Hutchinson, E, G.; Thornton, J. M. "Stereochemical quality of protein structure coordinates". *Proteins: Structure, Function, and Genetics*, 12, 4, April 1992, pp. 345–364.

[8] Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., Bairoch, A. “Protein Identification and Analysis Tools on the ExPASy Server” *(In) John M. Walker (ed): The Proteomics Protocols Handbook,* Humana Press, 2005, pp. 571-607.

[9] De Castro, E., Sigrist, C. J. A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., Hulo, N. “ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins”. *Nucleic Acids Res*., 34, July 2006, pp. W362-W365.

[10] Pearson, W. R., Lipman, D. J. "Improved tools for biological sequence comparison". *Proc. Natl. Acad. Sci. USA*., 85, 8, April 1988, pp. 2444-2448.

[11] Rost, B., Yachdav, G., Liu, J. "The PredictProtein server". *Nucleic Acids Res*., 32, 2004, (Web Server issue), W321-W326.

[12] The PyMOL Molecular Graphics System, Version 1.1 Schrödinger, LLC.

[13] Hofmann, K., Stoffel, W. "TMbase - A database of membrane spanning proteins segments". *Biol. Chem. Hoppe-Seyler,* 374, 166, 1993.

[14] Laskowski, R. A., Watson, J. D., Thornton, J. M. "ProFunc: a server for predicting protein function from 3D structure". *Nucleic Acids Res*. 33, July 2005 (Web Server issue), pp. W89–W93.

[15] Konc, J., Janezic, D. "ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment". *Bioinformatics*, 26, 9, March 2010, pp. 1160-1168.