Cloud Publications
International Journal of Advanced Bioinformatics and Computational Biology
2013, Volume 1, Issue 1, pp. 22-30, Article ID Med-84

Research Article | Open Access

In Silico Structure Modeling and Characterization of Hypothetical Proteins Present in Human Fetal Brain

Parakh Sharma, Komal Patil, Devangi Sarang and Pramod Shinde

Department of Bioinformatics, Guru Nanak Khalsa College, Mumbai, Maharashtra, India

Correspondence should be addressed to Parakh Sharma, parakh.sehgal09@gmail.com

Publication Date: 14 June 2013

Article Link: http://medical.cloud-journals.com/index.php/IJABCB/article/view/Med-84

Abstract Human genome project provided infrastructure of biological information and essential information of gene sequences make it possible to predict proteins for which there is no experimental evidence. Characterization of hypothetical proteins is provided by sequence similarity with proteins of known function. Aim of this study was to describe Hypothetical proteins [Q9Y4F4, Q14166, Q9NUQ6, Q9H0W9 and Q9H0R4] in the perireticular nucleus, a key structure in human brain development. The functional domain prediction was carried out using InterPro, BLASTcds, COGs and CDART. I-TASSER server was used for tertiary structure prediction and structure comparison studies performed using VAST and DALI. This study revealed some probable functions of hypothetical proteins. These proteins were predicted to serve the functional roles in various activities like DNA - protein binding, protein modification process, ligase activity, hydrolase activity and are also involved in the metabolic processes.

Keywords *Hypothetical Protein, Perireticular Nucleus, Bioinformatics, Characterization, Domain*

1. Introduction

The thalamic perireticular nucleus (PN) and reticular nucleus (RN) play the key role in the human brain development and are scattered within the internal brain capsule [1, 2, 3]. These nuclei lie directly in the path of corticofugal and corticopetal axons during development [1]. A common feature of these nuclei in different species is the immunoreactivity for some calcium binding proteins with a developmental pattern of expression [4]. Evidence suggests that the perireticular neurons in various species decrease in number with increasing gestation, but in humans this finding has not been supported by quantitative data [5].

In the development process from fetus to adult human brain undergo in continuous change. During this period, brain proteins undergo lots of structural changes resulting in the difficulties in structural data of these proteins. Hence, these limitations have provided the vast scope in the system biology and structural bioinformatics fields [6]. Hypothetical proteins (HP) are the proteins predicted from

nucleic acid sequences only and protein sequences with unknown function. HP's are awaiting experiments to show their existence at the protein level and subsequent bioinformatics handling in order to assign proteins a tentative function is mandatory [7]. The focus of the study was on the protein sequence data to predict protein structure and perform the annotations.

The prediction of experimental proteins structure and associate functions is very expensive and tedious process. The process of protein annotation involves two phases viz. first is sequence characterization i.e. identification of motifs and domains; second is function prediction. On the basis of sequence one can go for its function prediction [8]. Protein function is a prediction based on identification of short consensus sequences with known functions. These consensus sequence patterns are termed motifs and domains. Motifs are stable arrangements of several elements of secondary structure and the connections between them [9]. These discrete structural units are assumed to fold independently of the rest of the protein and to have its own function, hence are having significant role in the unknown protein characterisation [10].

For the functional characterization of HPs, protein sequences are examined for the presence of functional domains. Structures of the respective proteins can be predicted on the basis of its sequence using various structure modeling servers can be used based on threading and *ab initio* approach. The structure comparison study was performed for structural homologs finding.

2. Methodology

HP sequences were extracted from UniProtKB (http://www.uniprot.org/) with Q9Y4F4, Q14166, Q9NUQ6, Q9H0W9, and Q9H0R4 UniProt identifiers.

2.1. Structural Modeling

The structures for Q9H0W9 and Q9H0R4 proteins are available in PDB. For Q9Y4F4 structure modeling was not possible due to sequence length and homology limitation. Hence the focus for the study was on structure modelling for Q14166 and Q9NUQ6 proteins using I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) [11, 12] server.

2.2. Functional Characterization

The identification of motifs and domains in proteins is an important aspect of the classification of protein sequences and functional annotation.

2.2.1. On the Basis of Sequence

Functional characterisation was performed using InterPro (www.ebi.ac.uk/interpro/) [13], BLAST cds (www.ncbi.nlm.nih.gov/Structure/cdd/) [14, 15, 16], COGs (www.ncbi.nlm.nih.gov/COG/) [17] and CDART (www.ncbi.nlm.nih.gov/Structure/) [18] tools which showed the defined conserved regions in the sequences and guided the path in the classification of conserved protein families. The results of all four tools were compared and common results used for the analysis.

2.2.2. On the Basis of Structure

Structures characterization was performed using VAST (www.ncbi.nlm.nih.gov/Structure/VAST/) [19] and DALI (www.ekhidna.biocenter.helsinki.fi/dali_server/) [20] server to obtain structural homologous proteins. Structural domains are indentified and search is made to identify the functions of domains present in the structures.

3. Results and Discussion

The result of functional characterization using protein sequences were tabulated in Tables 1, 2 and 3 with detailed signature matches. Q9Y4F4 protein has Armadillo-type fold-IPR016024 and Armadillo-like helical-leucine-rich repeat variant domains [21]. The sequence from 351 - 1,715 assigned to PTHR21567 family which belongs to the structural protein class and play a role in the DNA binding. It is also involved in cellular process like cell cycle-mitosis. Q14166 protein has Tubulin-tyrosine ligase domain [22]. The sequence from 22 – 634 amino acids assigned to PTHR12241 TTL-Related family which belongs to the ligase protein class. It is involved in the cellular protein modification process and also has tubulin-tyrosine ligase activity [22, 23].

Table 1: *Probable Conserved Domains in HP's showing the Protein Family Membership within Panther and Pfam Databases Predicted Using Interpro Server*

Uniprot ID	Protein Family Membership		Detailed Signature Matches
Q9Y4F4	None predicted	Armadillo-type fold-IPR016024	128 - 333,superfamily - ssf48371 (arm-type_fold) 349 - 1,714, superfamily - ssf48371 (arm-type_fold)
		Armadillo-like helical-IPR011989 Leucine-rich Repeat Variant	140 - 296, 359 - 541, 1,250 - 1,380, 1,435 - 1,641 (g3dsa:1.25.10.10)
		Unintegrated signatures-no IPR	351 - 1,715, panther -pthr21567 (pthr21567), 351 - 1,715, panther - pthr21567:sf5 (pthr21567:sf5)
Q14166	Tubulin-tyrosine ligase (IPR004344)	Tubulin-tyrosine ligase	348 - 639,pfam-pf03133 (ttl) 300 - 644,prosite profiles -ps51221 (ttl)
		Unintegrated signatures	22 - 634,panther -pthr12241 (pthr12241) 22 - 634,panther-pthr12241:sf13 (pthr12241:sf13)
Q9NUQ6	Protein of unknown function DUF1387 (IPR009816)	Protein of unknown function DUF1387	59 - 368,pfam-pf07139 (duf1387)
		UBA-like	9-72,superfamily - ssf46934 (uba_like)
		Unintegrated signatures	1 - 558, panther-pthr32353 (pthr32353) 1 - 558,panther - pthr32353:sf1 (pthr32353:sf1)
Q9H0W9	None predicted	Domain of unknown function DUF1907	19 - 303, pfam- pf08925 (duf1907)
		Unintegrated signatures	1 - 315, panther - pthr13204 (pthr13204) 1 - 315, panther - pthr13204:sf0 (pthr13204:sf0) 3 - 315, superfamily - ssf117856 (ssf117856)
Q9H0R4	HAD-superfamily hydrolase, subfamily IIA (IPR006357) HAD-superfamily hydrolase, subfamily IIA, hypothetical 2 (IPR006355)	HAD-superfamily hydrolase, subfamily IIA	10-88,pfam,pf13344 (hydrolase_6) 10 - 226,tigrfams - tigr01460 (had-sf-iia)
		HAD-superfamily hydrolase, subfamily IIA, hypothetical 2	7 - 257,tigrfams - tigr01458 (had-sf-iia-hyp3)
		HAD-like domain	5 - 79, 176 – 257- (g3dsa:3.40.50.1000) 7 - 253, superfamily - ssf56784 (had-like_dom)
		Nitrophenylphosphatase-like domain	80 - 175- (g3dsa:3.40.50.10410)
		Unintegrated signatures	3 - 115,pfamb - pb010872 (pfam-b_10872) 177 - 246, pfam - pf13242 (hydrolase_like) 5 - 258,panther - pthr19288 (pthr19288) 5 - 258, panther - pthr19288:sf1 (pthr19288:sf1)

Table 2: *Probable PANTHER Protein Families and Class Present in HP's With Their Corresponding GO Terms Including Cellular Component, Molecular Function and Involvement in Biological Processes*

Protein	PANTHER Family	PANTHER Protein Class	GO Cellular Component	GO Molecular Function	GO Biological Process
Q9y4f4	Clasp (Pthr21567)	Structural Protein	-	Structural Molecule Activity Go:0005488 Binding	Cellular Process-Cell Cycle-Mitosis
Q14166	Ttl-Related (Pthr12241)	Ligase	Intracellular -Cytoskeleton - Microtubule	Structural Molecule Activity Go:0004835 Ttl Activity	Go:0006464 Cellular Protein Modification Process
Q9nuq6	Family Not Named (Pthr32353)	-	None Predicted.	Go:0005515 Protein Binding	-
Q9h0w9	Ptd012 Protein (Pthr13204)	-	Go:0005634 Nucleus	-	-
Q9h0r4	4-Nitrophenylphosphatase-Related (Pthr19288)	Phosphatase	-	Go:0016787 Hydrolase Activity	Phosphate Metabolic Process

Key: None predicted

Table 3: *Probable Conserved Domains in HP's predicted by BLASTcds, COG's and CDART*

Uniprot Id	BLASTcds		COGs		CDART
	Domain	Multidomain			
Q9Y4F4	No	No	KOG2933	Uncharacterized conserved protein	No
Q14166	pfam03133, Tubulin Tyrosine Ligase (TTL) - family	No	KOG2155	TTL-related protein	TTL cl18406
Q9NUQ6	pfam07139, Protein of unknown function (DUF1387)	Yes	No	No	No
Q9H0W9	pfam08925, Domain of Unknown Function (DUF1907)	No	KOG4048	Uncharacterized conserved protein	DUF1907 cl07499
Q9H0R4	1. Haloacid dehalogenase-like hydrolases (HAD)_like(cd01427) 2. HAD_like(cd01427) 3.HAD-SF-IIA-hyp3(TIGR01458)	Yes Yes Yes	KOG3040	Predicted sugar phosphatase (HAD superfamily)	HAD-like superfamily cl17915

Q9NUQ6 protein has Domain of Unknown Function DUF1387 is a multi domain and it belongs to UBA-like superfamily [24]. The sequence from 1 – 558 amino acids is assigned to PTHR32353 family having the protein binding activity. Q9H0W9 protein has Domain of Unknown Function DUF1907 present in the nucleus. The sequence from 1 – 315 amino acids is assigned to PTHR13204 family is PTD012 Protein. Q9H0R4 protein belongs to HAD-superfamily hydrolase, subfamily IIA and nitrophenyl phosphatase-like domain which is a multidomain. The sequence from 5 – 258 amino acids is assigned to PTHR19288 having hydrolase activity [24, 25].

I-TASSER model for Q14166 and Q9NUQ6 proteins were predicted to be with -0.42 and -2.01 appropriate C-score suggesting modelled structures are of good quality (Figures 1 and 2). C-score is a confidence score for estimating the quality of predicted models by I-TASSER. It is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations [11, 12].

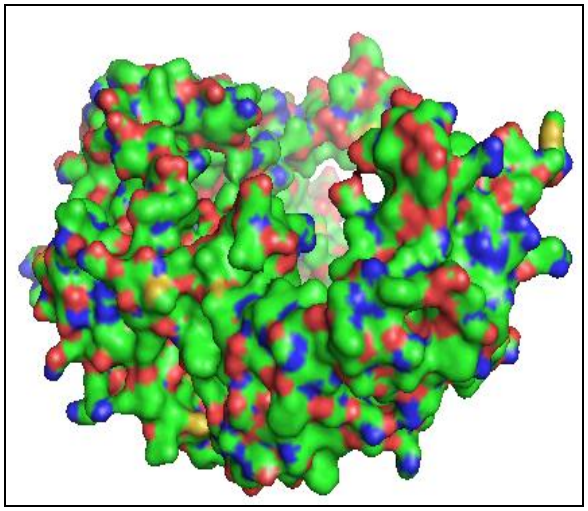

Figure 1: *Modeled Structure for HP Q14166*

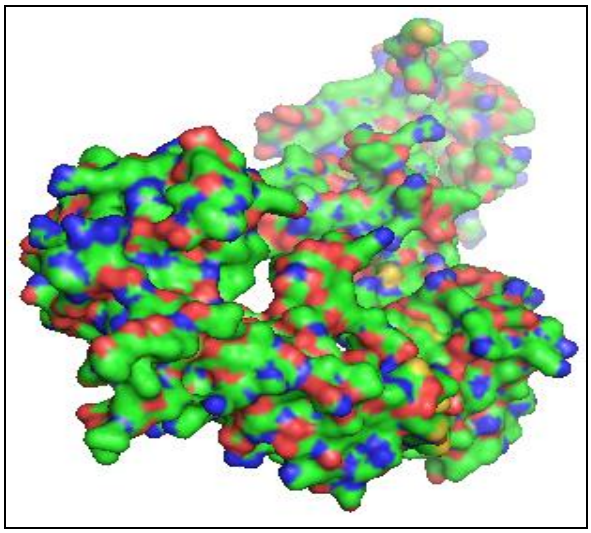

Figure 2: *Modeled Structure for HP Q9NUQ6*

Structural homologs for Q14166, Q9NUQ6, Q9H0W9 and Q9H0R4 proteins using VAST and DALI are tabulated in Tables 3, 4, 5, 6 and 7. Q14166 protein structure has Karyopherin Beta2 [26], Transportin-1 [27], Gtp-Binding Nuclear Protein Ran [28], Chromosome Region Maintenance,

Importin, Exportin-T and Snurportin-1 protein domains performing functional role in the cell nucleus [29]. Q9NUQ6 protein structure domains have found be related to Immunoglobulin protein having role in the antigen binding and Q9H0W9 protein structure has Putative DNA-Binding Protein domains [30]. Q9H0R4 protein structure has Haloacid Dehalogenase - Like Hydrolase Domain, Phospholysine Phosphohistidine Inorganic Pyrophosphate, Protein Nagd and P-Nitrophenyl Phosphatase (Pho2) domains [31, 32].

Table 4: *Structural Homologs for HP Q14166 Predicted Using DALI Server*

Sr. No.	Chain	Z-score	%id	Description
1	1qbk-B	20.7	8	Karyopherin Beta2
2	2ot8-A	16.9	7	Transportin-1
3	1wa5-C	13.0	4	Gtp-Binding Nuclear Protein Ran
4	4fgv-A	11.1	6	Chromosome Region Maintenance 1 (Crm1) Or Exporti
5	3nd2-A	11.0	8	Importin Subunit Beta-1
6	1z3h-B	10.9	3	Importin Alpha Re-Exporter
7	2jak-A	10.7	9	Serine/Threonine-Protein Phosphatase 2a 56 Kda Re
8	3icq-U	10.4	6	Exportin-T
9	3nby-D	10.2	4	Snurportin-1
10	2x19-B	10.1	9	IMPORTIN-13

Table 5: *Structural Homologs for HP Q9NUQ6 Predicted Using DALI Server*

Sr. No.	Chain	Z-score	%id	Description
1	1xed-E	12.0	5	Polymeric-Immunoglobulin Receptor
2	1zox-A	11.3	1	Clm-1
3	2nms-A	11.0	2	Cmrf35-Like-Molecule 1
4	3v4v-M	10.5	5	Integrin Alpha-4
5	2g60-H	10.5	10	Anti-Flag M2 Fab Light Chain
6	1psk-H	10.5	7	Antibody
7	1yee-H	10.5	6	Igg2a Fab Fragment (D2.5)
8	4ag4-H	10.5	6	Epithelial Discoidin Domain-Containing Receptor 1
9	3hc0-A	10.4	7	Immunoglobulin Igg1 Fab, Light Chain
10	1ahw-B	10.4	5	Immunoglobulin Fab 5g9 (Light Chain)

Table 6: *Structural Homologs for HP Q9H0W9 Predicted Using DALI Server*

Sr. No.	Chain	Z-score	%identity	Description
1	1xcr-A	61.0	100	Hypothetical Protein Ptd012
2	1xcr-B	58.5	100	Hypothetical Protein Ptd012;
3	1xv2-C	15.9	9	Hypothetical Protein, Similar To Alpha
4	3hwu-A	6.4	9	Putative DNA-Binding Protein
5	2dt4-A	5.8	12	Hypothetical Protein Ph0802
6	2p6y-A	4.6	8	Hypothetical Protein Vca0587

Table 7: *Structural Homologs for HP Q9H0R4 Predicted Using DALI Server*

Sr. No.	Chain	Z-score	%identity	Description
1	3hlt-A	46.6	100	Hdhd2
2	2ho4-B	40.3	87	Haloacid Dehalogenase-Like Hydrolase Domain
3	2x4d-A	31.5	43	Phospholysine Phosphohistidine Inorganic Pyrophos
4	2c4n-A	30.7	24	Protein Nagd
5	3qgm-B	30.7	25	P-Nitrophenyl Phosphatase (Pho2)
6	1wvi-A	30.0	24	Putative Phosphatases Involved In N-Acetyl-Glucos
7	3epr-A	29.9	23	Hydrolase, Haloacid Dehalogenase-Like Family
8	1vjr-A	29.5	25	4-Nitrophenylphosphatase
9	1zjj-B	29.3	25	Hypothetical Protein Ph1952
10	1ys9-A	29.3	24	Protein Spy1043

4. Conclusion

HP's are of perireticular nucleus for human fetal brain, it is very much essential to know their function and role in the fetal brain development to understand neurological diseases and to develop drug remedy against brain disorders. Using sequence and structure comparison of HP's various structural and functional domains were predicted along with their associate functions. The study outcome with the important functions of studied HP's like DNA binding, cellular protein modification process, tubulin-tyrosine ligase activity, protein binding activity, hydrolase activity and is also involved in metabolic processes. Protein structure comparison studies predicted domains e.g. Karyopherin Beta2, Polymeric-Immunoglobulin Receptor, Putative DNA-Binding Protein, HDHD2 etc. Predicting protein function can help to identify new targets for known drugs, new functional analogues of known targets

in different organisms or cellular compartments, or to detect proteins homologous to the original target that are more suitable to experimental analysis and also support the study of protein role in the pathway analysis.

References

[1] Mitrofanis J. *Patterns of Antigenic Expression in the Thalamic Reticular Nucleus of Developing Rat.* J. Comp. Neurol. 1992a. 320; 161–181.

[2] Mitrofanis J. *Development of the Pathway from the Reticular and Perireticular Nuclei to the Thalamus in Ferrets: A Dil Study*. Eur. J. Neurosci. 1994a. 6; 1864-1882.

[3] Mitrofanis J. *Development of the Thalamic Reticular Nucleus in Ferrets with Special Reference to the Perigeniculate and Perireticular Cell Group.* Eur. J. Neurosci. (1994)b. 6; 253-263.

[4] Contreras-Rodríguez J. *Neurochemical Heterogeneity of the Thalamic Reticular and Perireticular Nuclei in Developing Rabbits: Patterns of Calbindin Expression, Brain Research.* Developmental brain research. 2003. 144 (2) 211-21.

[5] Cumhur Murat Tulay, et al. *Morphological Study of the Perireticular Nucleus in Human Fetal Brains.* J Anat. 2004. 205 (1) 57–63.

[6] Desler C., *et* al., *Genome-Wide Screens for Expressed Hypothetical Proteins.* Methods Mol Biol. 2012. 815; 25-38.

[7] Lubec G *et* al., *Searching For Hypothetical Proteins: Theory and Practice Based upon Original Data and Literature.* Prog Neurobiol. 2005. 77 (1-2) 90-127.

[8] Campbell ID. *Downing AK Building Protein Structure and Function from Modular Units.* Trends Biotechnol. 1994. 12; 168.172.

[9] Chou KC. *Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition.* PROTEINS: Structure, Function, and Genetics. 2001. 43; 246.255.

[10] Ingolfsson H., *et* al. *Protein Domain Prediction.* Methods Mol Biol. 2008. 426; 117-143.

[11] Ambrish Roy., *et* al. *I-TASSER: A Unified Platform for Automated Protein Structure and Function Prediction.* Nature Protocols. 2010. 5; 725-738.

[12] Yang Zhang. *I-TASSER Server for Protein 3D Structure Prediction.* BMC Bioinformatics. 2008. 9; 40.

[13] Quevillon E., *et* al. *Interproscan: Protein Domains Identifier.* Nucleic Acids Res. 2005. 33; 116-120.

[14] Marchler-Bauer A, *Bryant SH,"CD-Search: Protein Domain Annotations on the Fly.* Nucleic Acids Res. 2004. 32; 327-331.

[15] Marchler-Bauer A., *et* al., *CDD: Specific Functional Annotation with the Conserved Domain Database.* Nucleic Acids Res. 2009. 37; 205-10.

[16] Marchler-Bauer A., *et* al., *CDD: A Conserved Domain Database for the Functional Annotation of Proteins.* Nucleic Acids Res. 2011. 39; 225-9.

[17] Tatusov R.L., et al. *The COG Database: An Updated Version Includes Eukaryotes.* BMC Bioinformatics. 2003. 4; 41.

[18] Geer L., et al., *CDART: Protein Homology by Domain Architecture.* Genome Res. 2001. 12 (10) 1619-23.

[19] Gibrat J.F., et al. *Surprising Similarities in Structure Comparison*, Curr Opin Struct Biol. 1996. 6 (3) 377-85.

[20] Holm L., et al. *Server: Conservation Mapping in 3D.* Nucl. Acids Res. 2010. 38; 545-549.

[21] Figueroa et al. *Biophysical Studies Support a Predicted Superhelical Structure with Armadillo Repeats For Ric-8.* Protein Science. 2009. 1139-1145.

[22] Ersfeld et al. *Characterization of the Tubulin-Tyrosine Ligase.* The Journal of Cell Biology.1993. 725-732.

[23] Trichet et al. *Characterization of the Human Tubulin Tyrosine Ligase-Like 1 Gene Mapping To 22q13. 1.* Gene 257. 2000; 109-117.

[24] Zhu et al. *SGNP: An Essential Stress Granule/Nucleolar Protein Potentially Involved In 5.8 S Rrna Processing/Transport*. PloS one. 2008.

[25] De Pierre et al. *Ecto-Enzymes of the Guinea Pig Polymorphonuclear Leukocyte I. Evidence for an Ecto-Adenosine Monophosphatase,-Adenosine Triphosphatase, and-P-Nitrophenyl Phosphatase.* Journal of Biological Chemistry.1974. 7111-7120.

[26] Chook et al. *Structure of the Nuclear Transport Complex Karyopherin-B2–Ran¨ Gppnhp.* Nature. 1999; 230-237.

[27] Imasaki et al. *Structural Basis for Substrate Recognition and Dissociation by Human Transportin 1.* Molecular Cell. 2007. 57-67.

[28] Moore et al. *The GTP-Binding Protein Ran/TC4 is required for Protein Import into the Nucleus.* 1993. 661-663.

[29] Lee et al. *A Novel Chromosome Region Maintenance 1-Independent Nuclear Export Signal of the Large Form of Hepatitis Delta Antigen That is required for the Viral Assembly.* Journal of Biological Chemistry. 2001. 8142-8148.

[30] Hoeffler et al. *Cyclic AMP-Responsive DNA-Binding Protein: Structure Based on a Cloned Placental Cdna.* Science. 1988. 1430-1433.

[31] Kuznetsova *et* al. *Genome-Wide Analysis of Substrate Specificities of the Escherichia Coli Haloacid Dehalogenase-Like Phosphatase Family.* Journal of Biological Chemistry. 2006. 36149-36161.

[32] Yokoi et al. *Molecular Cloning of A cDNA for The Human Phospholysine Phosphohistidine Inorganic Pyrophosphate Phosphatase.* Journal of Biochemistry. 2003. 607-614.