BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 10(8)

Domain analyses of Usher syndrome causing Clarin-1 and GPR98 protein models

Sehrish Haider Khan[1], Muhammad Rizwan Javed[1]*, Muhammad Qasim[1], Samar Shahzadi[1], Asma Jalil[1] & Shahid ur Rehman[2]

[1]Department of Bioinformatics and Biotechnology, Government College University Faisalabad (GCUF), Allama Iqbal Road 38000, Faisalabad, Pakistan; [2]Department of Poultry Husbandry, University of Agriculture, Faisalabad, Pakistan; Muhammad Rizwan Javed - Email: rizwan@gcuf.edu.pk; Phone: +92-(0)301-6012931; *Corresponding author

Received June 22, 2014; Revised July 06, 2014; Accepted July 07, 2014; Published August 30, 2014

Abstract:
Usher syndrome is an autosomal recessive disorder that causes hearing loss, Retinitis Pigmentosa (RP) and vestibular dysfunction. It is clinically and genetically heterogeneous disorder which is clinically divided into three types i.e. type I, type II and type III. To date, there are about twelve loci and ten identified genes which are associated with Usher syndrome. A mutation in any of these genes e.g. CDH23, CLRN1, GPR98, MYO7A, PCDH15, USH1C, USH1G, USH2A and DFNB31 can result in Usher syndrome or non-syndromic deafness. These genes provide instructions for making proteins that play important roles in normal hearing, balance and vision. Studies have shown that protein structures of only seven genes have been determined experimentally and there are still three genes whose structures are unavailable. These genes are Clarin-1, GPR98 and Usherin. In the absence of an experimentally determined structure, homology modeling and threading often provide a useful 3D model of a protein. Therefore in the current study Clarin-1 and GPR98 proteins have been analyzed for signal peptide, domains and motifs. Clarin-1 protein was found to be without any signal peptide and consists of prokar lipoprotein domain. Clarin-1 is classified within claudin 2 super family and consists of twelve motifs. Whereas, GPR98 has a 29 amino acids long signal peptide and classified within GPCR family 2 having Concanavalin A-like lectin/glucanase superfamily. It was found to be consists of GPS and G protein receptor F2 domains and twenty nine motifs. Their 3D structures have been predicted using I-TASSER server. The model of Clarin-1 showed only *a-helix* but no beta sheets while model of GPR98 showed both *a-helix* and β sheets. The predicted structures were then evaluated and validated by MolProbity and Ramachandran plot. The evaluation of the predicted structures showed 78.9% residues of Clarin-1 and 78.9% residues of GPR98 within favored regions. The findings of present study has resulted in the three dimensional structure prediction and conserved domain analysis which will be quite beneficial in better understanding of molecular components, protein-protein interaction, clinical heterogeneity and pathophysiology of Usher syndrome.

Background:
Usher syndrome (USH) is an autosomal recessive disorder that causes deafness (hearing loss), blindness (retinis pigmentosa) and vestibular areflexia (dysfunction). It was discovered in 1858 by Von Graefe. He found hearing and vision loss and vestibular impairment in the persons suffering from Usher syndrome. Later on a British ophthalmologist found that this syndrome is inheritable **[1]**. Mutations in different genes or different mutations in the same genes produce same or clinically heterogeneous features in affected individuals. On the basis of retinal degeneration level, hearing impairment and presence or absence of vestibular dysfunction; severity of Usher syndrome can be determined. Chales Usher had initially put Usher Syndrome in two categories **[2]**. Currently, this syndrome is clinically divided into three types i.e. type I, type II and type III. Type I is most severe form in which children are deaf at birth with severe balance and vision problems, usually loss of peripheral vision followed by continuous retinal degeneration resulting in complete blindness in the second or third decade of life. Usher syndrome type I has genetically six subtypes i.e.

USHIB-USHIG. MYO7A is the one of the most common genes of Usher syndrome type I that causes USHIB. MYO7A is responsible for organelle transport and clearance of optin from cilium **[3].** Diagnosis of the disease can be confirmed by mutational analysis of the reported genes and electroretinography (ERG) of affected individuals **[3-5]**.

By the analysis of USA and UK population, 27 different mutations have been identified where 19 mutations are new and remaining are missense mutations. 35-39% is found in USHIB & USHID, 11% in USHIF, while 7% in USHIC and USHIG **[6]**. In Pakistani population, data collection over more than 400 families shows that 10% of the children are suffering from USH1. Half of them are linked with the subtype USHIB and approximately 30% are linked with USHID, while remaining is equally divided with USHIC and USHIF. Relative abundance in the sampling of Pakistani deaf children shows 42.9% of the children are affected with USH1 and 4.1% with USH2 **[7]**.

Usher syndrome type II (USHIIA-USHIIC) is less severe than type I. Children born with type II have moderate to severe hearing loss and normal balance. Usher syndrome type III (USHIIIA) is less severe than type I and II. Persons with type III may develop progressive hearing loss and visual problem but have normal vestibular function. So far, 12 loci and 10 genes have been identified out of which six (MYO7A, CDH23, PCDH15, USH1E, USHIC and USHIG) are responsible for Usher type I, three (USH2A, GPR98 and DFNB31) for type II and only one (USH3A) for Usher type III **[3, 7-10]**.

Mutations in MYO7A cause USHIB; USH1C & USH1D are caused by mutations in USH1C and CDH23, respectively; while USHIF and non-syndromic deafness are caused by mutations in PCDH15. USHIB encodes Myosin VIIA that acts as motor protein, USHID encodes cadherin 23 and USHIF encodes photocadherin15, these are two cell-cell adhesion proteins. USHIC encodes harmonin and USHIG encodes SANS, both are scaffold proteins. These proteins create a network with each other by binding to PZD domain in USHIC harmonin protein and its scaffold function is carried out with SANS protein. In USH type 2, the USH2A encodes a protein named as Usherin and USH2C encodes G-protein, both are trans-membrane proteins. The proteins of USH type II are also a part of protein network and involved in protein interactions. They also serve as a candidate for USH2B. USH type I and type II proteins form interactions in the protein network that shows the pathophysiological pathway in USH.

Studies have shown that protein structures of only seven USH genes have been determined experimentally and there are still three genes whose protein structures are unavailable. These genes are Clarin-1 (USH3A), GPR98 (USH2C) and Usherin (USH2A). In the absence of an experimentally determined structure, comparative or homology modeling and threading often provide a useful 3D model for a protein that is related to at least one protein structure. Therefore in the current study Clarin-1 and GPR98 proteins have been selected for *in-silico* analysis and 3D structure prediction. The current findings will help in the determination of protein-protein interactions which will be beneficial in understanding the molecular basis of both auditory and photo-transduction.

Methodology:

Sequence Retrieval

Amino acid sequences of the Usher genes Clarin-1 (GeneBank Accession # NG_009168) and GPR98 (GeneBank Accession # NG_007083) were obtained from Universal Protein Resource **[11]**.

Prediction and Removal of Signal Peptide

Usher syndrome proteins may contain signal peptide in their sequence. To find the signal peptide, SignalP-4.1 Server was used. The SignalP-4.1 server determines the presence and area of signal-peptide cleavage-sites in amino acid sequence of proteins **[11]**.

Domain Analysis

A protein sequence with a precise 3D structure and a precise function always contains a domain. Domains of the Usher proteins under study were analyzed and identified by various databases; Prosite database using ScanProsite at Expasy, InterPro database using InterProScan at EBI, Pfam database using Search Pfam at Sanger Institute **[11-12]** and with the help of Conserved Domain Database using CDD search at NCBI **[13].**

Motif Search

The existence of a specific motif reveals particular functions of the proteins. The conserved functional motifs in the protein sequence of Usher proteins under study were found by Motif Search Server using Prosite pattern, Prosite profile, PRODOM and PRINTS databases **[11]**.

Structure Prediction

I-TASSER server was used for protein-structure predictions. For protein structure and function prediction by I-TASSER, the server was accessed through the link http://zhanglab.ccmb.med.umich.edu/I-TASSER **[14]**. The protein sequences were submitted in the query box or the files were uploaded through local computer in FASTA format. The results received via email were analyzed further.

Structure Visualization

To view the structures in atomic coordinate files with pdb extensions, to enable the manipulation of the image and to visualize the molecules with different perspectives, molecular visualization tool is needed. The PDB files obtained from I-TASSER server were visualized with UCSC CHIMERA **[15]**.

Model Evaluation and Validation

After the prediction of 3D structures, the predicted models were evaluated and validated through MolProbity and Ramachandran Plot **[16]** for their sterio-chemical properties. MolProbity server computed Ramachandarn values for dihedral angles, poor rotameric conformations, Cβ deviation, bad angles and bond lengths of all the residues.

Results & Discussion:

In the current study the amino acid sequence of Clarin-1 (GenBank Accession # NG_009168) and GPR98 (GenBank

Accession # NG_007083) were retrieved from UniProt (Accession no. P58418 and Q8WXG9, respectively).

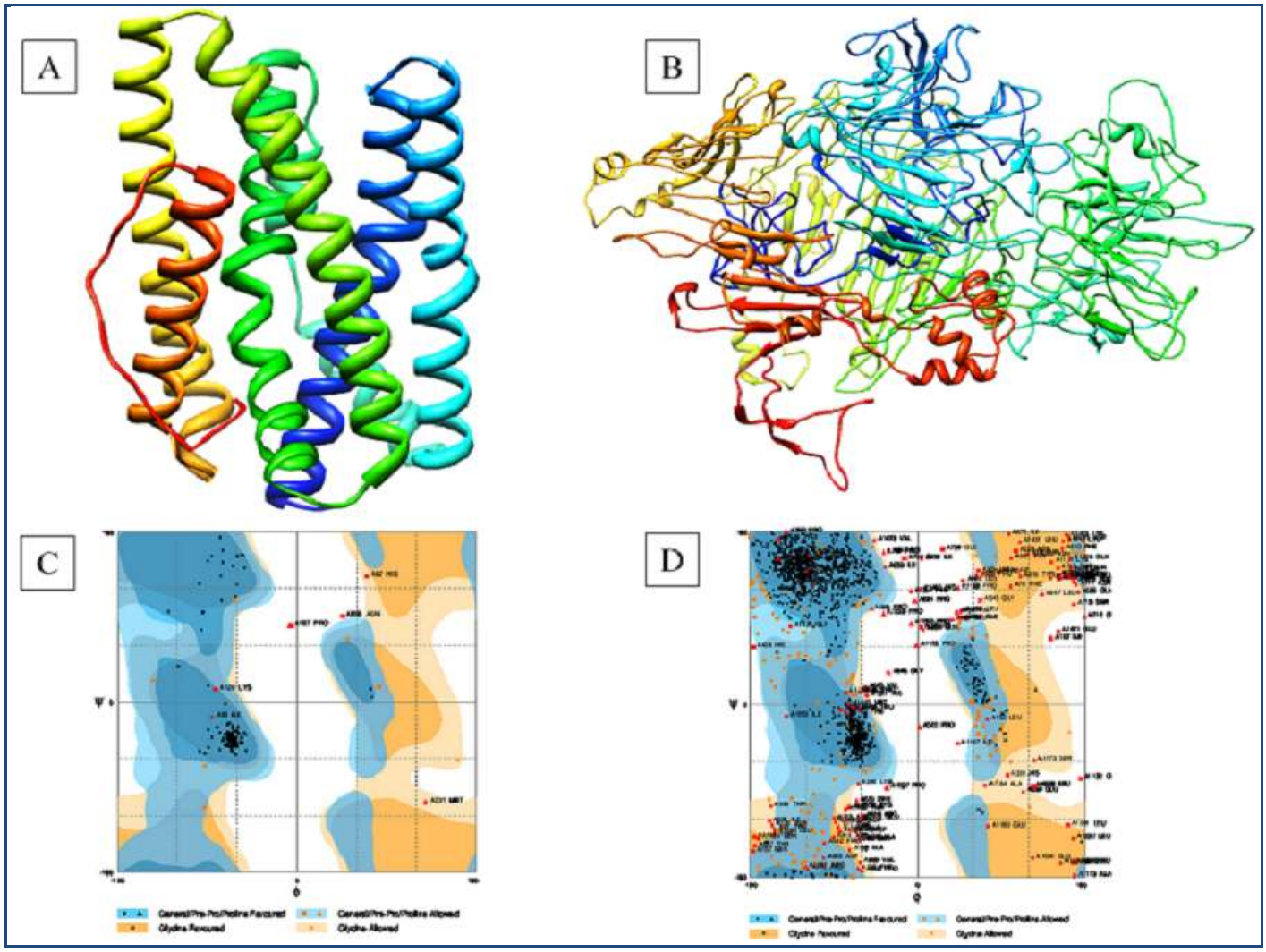

Figure 1: I-TASSER predicted structures of **(A)** Clarin-1; **(B)** GPR98 using templates PDB ID: 4DR0A (having 27% homology with the Clarin-1) & PDB ID: 4IGlA (having 22% homology with the GPR98), respectively. The model of Clarin-1 showed only *a-helix* but no beta sheets while model of GPR98 showed both *a-helix* (spiral) and *β* sheets (broad arrows). Ramachandran plot analysis of **(C)** Clarin-1 and **(D)** GPR98 protein structures to visualize dihedral angles; φ against ψ.

Signal Peptide Analysis

Before using the retrieved protein sequences for homology modeling, the sequences were analyzed for the presence of signal peptide through SignalP-4.1 server. A signal peptide is a short length peptide chain that helps in the transportation of protein inside and outside of the cell. This short chain is eliminated during maturation of the protein. Therefore, the signal amino acid sequence should be eliminated from primary sequence. The elimination of signal peptide confirms the structure prediction according to the original structure of the protein. According to Signal-HMM result of Clarin-1, the maximum cleavage (max. C) site probability was 0.500 at 28th amino acid and there was no signal peptide in the protein. Whereas, in GPR98 sequence analysis Signal-HMM results showed that the maximum cleavage (max. C) site probability of the GPR98 signal peptide was 0.594 at 30th amino acid and a signal peptide of 29 amino acids was present (the sequence of which was removed before proceeding further).

Domain Analysis

With the help of Prosite, Pfam, InterProScan and CDD search, a number of domain hits were found within the proteins being analyzed. Prosite results showed the existence of prokar lipoprotein domain within Clarin-1 and presence of two domains, GPS and G protein receptor F2 within GPR98. Pfam and InterProScan results showed that Clarin-1 is classified within claudin 2 super family that play role in sterocilia and photoreceptor cell synapses. There were four transmembrane domains in Clarin-1 which play role in photoreceptor cell synapses, pathophysiological pathway and hair cell development. While, in GPR98 the InterProScan results also showed the presence of GPS domain. The protein is classified within GPCR family 2 and Concanavalin A-like lectin/glucanase superfamily. The Pfam analysis showed the presence of three domains Calx-Beta, Laminin G3 and EPTP families within GPR98 protein. Additionally in GPR98, seven PZD binding domain hits were found that help in the formation

of hair cell stereocilia **[17].** The study of Clarin-1 protein by means of Conserved Domain Database (CDD) Search tool indicated that there is no conserved domain within this protein while twenty five Calx-Beta domains within GPR98 were observed.

Motif Analysis

Motifs are amino acid residues within a domain that occur consistently and are responsible to carry out the specific function of the particular protein. Multiple short conserved motifs drawn from sequence alignments may carry the Fingerprints. Fingerprints are very helpful in evolutionary studies and in assigning a recently sequenced protein to a particular family. In PRINTS and PROSITE, fingerprints are created from gapped alignments, but in case of BLOCKS from un-gapped alignments. From PROSITE and PRINTS, fingerprints may consist of numerous motifs, that's why these are more flexible and strong as compared to a single PROSITE motif. It is helpful to verify similar data from more than one database and then compare their results. Therefore, fingerprints in the query sequences were found by MOTIF search server using PROSITE, PRODOM and PRINTS. Results of Clarin-1 showed 12 motifs with prokar lipoprotein being the most prominent in PRODOM and Rhodopsin-like GPCR superfamily signature in PRINTS; while in GPR98, twenty nine motifs were observed. GPR98 also contains several repeated motifs including some calcium binding Calx-Beta repeats and seven copies of an epitempin repeats.

Structure Prediction, Model Evaluation and Validation

The tertiary structures of Clarin-1 and GPR98 proteins were predicted using I-TASSER server. In case of Clarin-1, homology modeling through I-TASSER was done using template (PDB ID: 4DR0A) which has 27% homology with the Clarin-1. The model of Clarin-1 showed only *a-helix* but no beta sheets in it **(Figure1).** Similarly the same procedure was repeated with GPR98 by using a template (PDB ID: 4IGlA; having 22% similarity) and predicted model of GPR98 showed both *a-helix* and *β* sheets **(Figure1).**

Both predicted structures were then statistically evaluated and validated as shown in **Table 1 (see supplementary material)**. Validation and evaluation of 3D structure of proteins, complexes and nucleic acid can be done automatically and efficiently by MolProbity **[18]**. When φ and ψ angles of the amino acids of a particular protein are plotted against each other, the resulting diagram is called a Ramachandran plot **(Figure1).** It is basically a technique to visualize dihedral angles φ against ψ of all residues, which is the back bone of protein structure. Clarin-1 and GPR98 showed same percentage of residues in favored region when evaluated by Ramachandran plot. Clarin-1 showed 78.9% of its residues in favored region. Similarly, the predicted model of GPR98 also had 78.9% residues in favored region as shown in **Table 1 (see supplementary material)**.

Conclusion:

Advances in molecular genetics techniques have revolutionized the identification of new mutations in various genes but usually the effect of these new mutations upon mutated protein structure and function, protein-protein interactions, onset and severity of disease is not focused. The present manuscript answers this fundamental question. The signal peptide analysis showed the presence of 29 amino acid long signal peptide in GPR98, whereas no signal peptide was observed in Clarin-1. Both proteins consist of a number of conserved domains which play important role in sterocilia development and photoreceptor cell synapses. Clarin-1 showed twelve motifs with prokar lipoprotein being the most prominent while in GPR98 twenty nine motifs were observed including calcium binding Calx-Beta repeats. The Clarin-1 predicted structure showed only *a-helix* while in GPR98 structure both *a-helix* and *β* sheets were present. The study will help to better understand the pathophysiology of disease, impact of allele variants on protein and to devise new tools for therapeutic intervention.

References:

[1] Millan J *et al. J Ophthalmol.* 2011 doi:10.1155/2011/417217:1 [PMID: 21234346]
[2] Wu YW & Chiu CC, *Psychiatry Clin Neurosci.* 2006 **60:** 626 [PMID: 16958948]
[3] William DS, *Vision Res.* 2008 **48:** 433 [PMID: 17936325]
[4] Mets MB *et al. Trans Am Ophthalmol Soc.* 2000 **98:** 237 [PMID: 11190026]
[5] Hashimoto T *et al. Gene Ther.* 2007 **14:** 584 [PMID: 17268537]
[6] Ouyang XM *et al. Hum Genet.* 2005 **116:** 292 [PMID: 15660226]
[7] Aparisi MJ *et al. Mol Vis.* 2010 **16:** 2948 [PMID: 21203349]
[8] Astuto LM *et al. Am J Hum Genet.* 2000 **67:** 1569 [PMID: 11060213]
[9] Reiners J *et al. Exp Eye Res.* 2006 **83:** 97 [PMID: 16545802]
[10] Yan D & Liu XZ, *J Hum Genet.* 2010 **55:** 327 [PMID: 20379205]
[11] Sehar U *et al. Bioinformation* 2013 **9:** 725 [PMID: 23976829] (www.cbs.dtu.dk/services/SignalP/).
[12] Punta M *et al. Nucleic Acids Res.* 2012 **40:** D290 [PMID: 22127870]
[13] Marchler BA *et al. Nucleic Acids Res.* 2011 **39:** D225 [PMID: 21109532]
[14] Ahmad R *et al. Bioinformation*. 2013 **9:** 802 [PMID: 24143049]
[15] Pettersen EF *et al. J Comput Chem.* 2004 **25:** 1605 [PMID: 15264254] (http://www.cgl.ucsf.edu/chimera/).
[16] Lovell SC *et al. Proteins* 2003 **50:** 437 [PMID: 12557186] (http://mobprobity.biochem.duke.edu/).
[17] Grati M *et al. J Neurosci.* 2012 **32:** 14288 [PMID: 23055499]
[18] Chen VB *et al. Acta Crystallogr D Biol Crystallogr.* 2010 **66:** 12 [PMID: 20057044]

Edited by P Kangueane
Citation: **Khan *et al.*** Bioinformation 10(8): 491-495 (2014)

BIOINFORMATION

open access

Supplementary material:

Table 1: Statistical results of I-TASSER predicted structure evaluation by MolProbity and Ramachandran plot analysis.

Determination of Protein Geometry Parameters by MolProbity	**Observed**	
	Clarin-1	**GPR98**
Poor Rotamers	5.70%	5.53%
Ramachandran outliers	2.17%	5.53%
Ramachandran favored	93.91%	84.90%
Cβ deviations >0.25A	16	105
Residues with bad bonds	0.00%	0.00%
Residues with bad angles	0.1%	4.77%
Ramachandran Plot Statistics		
Number of residues in favored region (%age)	217 (78.9%)	1155 (78.9%)
Number of residues in allowed region (%age)	7 (3.0%)	176 (12.0%)
Number of residues in outlier region (%age)	6 (2.6%)	133 (9.1%)