Hypothetical Protein Predicted to be Tumor Suppressor: A protein Functional Analysis

Md. Abdul Kader
Mawlana Bhashani Science and Technology University
Md. Akash Ahmed
MBSTU: Mawlana Bhashani Science and Technology University
Md. Sharif Khan
MBSTU: Mawlana Bhashani Science and Technology University
Sheikh Abdullah Al Ashik
MBSTU: Mawlana Bhashani Science and Technology University
Md Shariful Islam (sharifbge@gmail.com)
University of Kentucky https://orcid.org/0000-0002-7631-882X
Mohammad Uzzal Hossain
National Institute of Biotechnology

Research Article

Keywords: Hypothetical protein, Functional annotation, Novel bacterium, VHL domain, Tumor suppressor

Posted Date: May 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-291087/v1

Version of Record: A version of this preprint was published at Genomics & Informatics on March 31st, 2022. See the published version at https://doi.org/10.5808/gi.21073.

Abstract

Background

Litorilituus sediminis is a Gram-negative, aerobic, novel bacterium under the family of Colwelliaceae, has a stunning hypothetical protein containing domain called von Hippel–Lindau (pVHL) that has significant tumor suppressor activity. Therefore, this study was designed to elucidate the structure and function of the biologically important hypothetical protein EMK97_00595 (QBG34344.1) using several bioinformatics tools.

Results

The functional annotation exposed that the hypothetical protein is an extracellular secretory soluble signal peptide and containing the VHL (VHL beta) domain that has a significant role in tumor suppression. This domain is conserved throughout evolution, as its homologs are available in various types of the organism like mammals, insects, and nematode. The gene product of VHL has a critical regulatory activity in the ubiquitous oxygen-sensing pathway. This domain has a significant role to inhibit cell proliferation, angiogenesis progression, kidney cancer, breast cancer, and colon cancer.

Conclusion

At last, the current study depicts that the annotated hypothetical protein is linked with tumor suppressor activity which might be of great interest to future research in the higher organism.

Background

Bacteria possess tremendous compatibility that can be used to the necessity of human welfare and *Litorilituus sediminis* can be one of them. *Litorilituus sediminis* is a Gram-negative, aerobic, curved-rod shaped, non-spore-forming, catalase, and oxidase-positive bacterium with the polar or sub-polar flagellum. It was isolated from a sediment sample that was collected from the coastal region of Qingdao, China [1]. This organism grew optimally at 37°C, pH 8–9. This type of bacterium was novel among the other genera under the family of Colwelliaceae. The characteristics like phenotypic, chemotaxonomic, and well-confirmed phylogenetic evidence of *Litorilitus* belongs to the family Colwelliaceae was distinctive that implied as a novel genus. This novel bacterium has a prominent concentration of cellular constituents comparing with other genera and these are C16:0 and C16:1 ω7c fatty acids, Phosphatidylethanolamine (PE), phosphatidylglycerol (PG), aminophospholipid (PN), and two amino lipids (AL1, AL2) as well as isoprenoid quinone 8 [1]. Along with bacterial cellular components, a profuse number of proteins exist where approximately 2% of the genes code for proteins as well as the remaining are non-coding or still functionally unknown [2].

The number of genes having unknown functions referred to as hypothetical proteins is present in each organism's genome [3] and these are a category of the protein whose existence is not confirmed by any

experimental evidence but can be predicted to be expressed from an open reading frame (ORF) [4]. The hypothetical proteins can be classified as uncharacterized protein families (UPF) which are experimentally verified to exist but have not been identified or linked to a known gene, and the other type is the domain of unknown functions (DUF) [5] that is experimentally characterized proteins in the absences of known functional or structural domains [6][7]. Despite the lack of functional characterization, they play a significant role in understanding biochemical and physiological pathways like to explore new structures and functions [8], pharmacological targets and markers [9], and early detection and benefits for proteomic and genomic research [10]. With the advancement of Computational Biology, it has become easier to analyze hypothetical proteins using bioinformatics tools that provide various advantages like the determination of 3D structural conformation, identification of new domains and motifs, assessment of new cascades and pathways, phylogenetic profiling, and functional annotation [11].

However, due to novel genera under the family of Colwelliaceae, this study intended to characterize the protein EMK97_00595 [*Litorilituus sediminis*], a family of von Hippel–Lindau (VHL) that have an overwhelming function as a tumor suppressor in higher organisms. The main feature of VHL is that it is a critical regulator of the ubiquitous oxygen-sensing pathway and can act as a substrate recognition component of an E3 ubiquitin ligase complex [12], also promote the degradation of epidermal growth factor receptor, pro-angiogenesis factors, remodeling of the extracellular matrix, and helps in apoptosis resulting tumor suppression[13].

In the higher organism during cellular normoxia when oxygen is available, the cellular HIFα is hydroxylated by prolyl hydroxylase and works as a felicitous substrate for pVHL which is a constitutive active site of E3 ubiquitin ligase. The hydroxyproline of hydroxylated HIFα provides a binding signal for pVHL, which leads to efficient ubiquitylation and proteasomal degradation of HIFα protein. On the other hand, in hypoxia condition HIFα is not prolyl hydroxylated and may escape pVHL recognition, resulting in accumulation of HIFα and formation of a complex with HIF1β, goes into the nucleus and activates a transcriptional program to cope with the short-term, long-term effects of oxygen deprivation, several signaling pathways as well as angiogenesis factor for leading cell proliferation or tumor [13][14]. So the function of the hypothetical protein that exists in the *Litorilituus sediminis* is considerable.

Therefore, this study manifests a reliable interpretation of this hypothetical protein EMK97_00595 (QBG34344.1) by adopting an integrated workflow that can be a potential research interest in the field of tumor suppression study.

Methods

1. **Sequence retrieval and similarity identification**

The hypothetical protein EMK97_00595 [*Litorilituus sediminis*] was chosen by exploring the NCBI database which can act as a significant research interest in numerous cancer research fields in the near future. The sequence of the hypothetical protein (GenBank Accession: QBG34344.1 and NCBI Reference

Sequence: WP_130598461.1) that may contain a tumor suppressor domain was retrieved and collected as a FASTA format and submitted to several prediction servers for the in-silico characterization. Initially, a similarity search was performed using the NCBI BLASTp program [15] against the non-redundant and Swissprot database [16], for predicting the function of the hypothetical protein.

2. **Multiple sequence alignment and phylogeny analysis**

A multiple sequence alignment is a tool used to explore closely related genes or proteins to find the evolutionary relationships between genes and to identify shared patterns among functionally or structurally related genes. Sequence alignment was performed by the MUSCLE server of EBI [17], and an evolutionary relationship was accomplished by Jalview 2.11 software [18], between the hypothetical protein EMK97_00595 and the proteins that had structural similarity with the protein of interest.

3. **Analysis of physicochemical properties**

ProtParam [5] is a tool that computes various physical and chemical parameters of protein sequences. The physicochemical properties of the hypothetical protein were predicted using the ProtParam tool in the ExPASy server [19], which predicts all the relative properties including molecular weight, theoretical pI, amino acid composition, the total number of positive and negative residues, instability index, aliphatic index and grand average of hydropathicity (GRAVY) [20][21][22].

4. **Analysis of the secondary structure**

The servers that are utilized to predict protein secondary structure were SOPMA [23] and PSIPRED [24]. SOPMA is a general secondary structure prediction tool, on the other hand, PSIPRED is a server for comprehensive analysis of protein. The server SOPMA was initially employed to predict the secondary structure and then the result derived from the SOPMA server was validated by exploiting PSIPRED.

5. **3D Structure Modeling and Quality Assessment**

HHpred server [25] that works based on the pairwise comparison profile of hidden Markov models, was used to build the 3-dimensional structure using the best scoring template. The confidence of the predicted structure was also visualized by SWISS-MODEL [26]. Several quality assessment tools of the SAVES and ProFunc [27] server were applied to estimate the reliability of the predicted 3D structure model of the hypothetical protein. The Ramachandran plot for the model was built using the PROCHECK program [28] to visualize the backbone dihedral angles of amino acid residues. The quality of the protein 3D structure was assessed with the help of the ERRAT server [29] and Varify 3D server was used to determine the compatibility of an atomic model (3D) with its amino acid sequence as well as comparing the results to standard structures [30][31].

6. **Active site determination**

Computed Atlas of Surface Topography (CASTp) is an online active site determination server [32] that calculates the location, delineation, and concave surface regions on 3D structures of proteins. CASTp predicted the active site of the selected hypothetical protein that showed the binding sites, amino acid binding regions with area and volume.

7. **Identification of protein subcellular localization and topology**

The subcellular location of the following protein was predicted by using the BUSCA web server [33]. BUSCA amalgamates different tools - DeepSig, TPpred3, PredGPI, BetAware, ENSEMBLE3.0, BaCelLo, MemLoci, and SChloro to predict protein features related to localization. The result was further checked by Cello [34], PsortB [35], Gneg-mPLoc [36], SOSUIGramN [37], and PSLpred [38]. Prediction of signal peptide was done by using PrediSi [39] and SignalP-5.0 Server [40]. The solubility of the hypothetical protein was evaluated by Protein-sol [41] and SOSUI [42] webserver. Protein transmembrane helices were assessed by HMMTOP [43], TMHMM [44] and, Sable [45] webserver. The topology of hypothetical protein was predicted by the ProFunc server [13].

8. **Prediction of protein domain, superfamily, family, coil, and folding pattern**

Domain/Superfamily/Family of the following hypothetical protein was analyzed by using the servers – CDD from NCBI [46], Pfam [47], SMART [48], Interpro [49], SCOP [50][51], Supfam [52], Motif , ProFunc [27], Phyre [53], and CATH-Gene3D [54]. Among them, CDD, Pfam, SMART, Interpro, SCOP, Supfam, MotifFinder were employed to predict function from the sequence of the hypothetical protein, and ProFunc, Phyre 2, and CATH-Gene3D servers were used to predict the function from the 3-dimensional structure of the hypothetical protein. Only the lowest e-value was considered to determine protein classification, which indicates good similarity. The protein folding pattern was determined by using Phyre 2 and PFP-FunDSeqE [55] servers where protein coil nature was determined by using PCoils [56] from the Bioinformatics toolkit server.

9. **Generation of Protein-protein interaction network**

As the proposed investigation seeking a tumor suppressor protein from microorganisms, STRING [57] has been used to summarize the network information of VHL tumor suppressor protein. Because of being a novel microorganism, there is no specific network is available. Here the VHL protein from humans has been used as a supposition model that might give an intellectual knowledge about VHL protein if it may apply to the human.

Results

1. Identification of sequence homology

The BLASTp result of the FASTA sequence shows the sequence homology with other identical proteins (Tables 1 and 2). Construction of phylogenetic tree using multiple sequence alignment generated from

BLASTp result shows the evolutionary relationship of the selected hypothetical protein (WP_130598461.1) in Figure 2.

Table 1: Similar proteins obtained from the non-redundant database.

Accession	Description	Scientific Name	Total Score	Query Cover	E-value	% Identity
WP_118961164.1	hypothetical protein [*Colwellia sp. RSH04*]	*Colwellia sp. RSH04*	349	100%	5.00E-120	74.18%
WP_033081725.1	hypothetical protein [*Colwellia psychrerythraea*]	*Colwellia psychrerythraea*	235	100%	4.00E-75	51.17%
WP_142932219.1	hypothetical protein [*Aliikangiella sp. M105*]	*Aliikangiella sp. M105*	108	94%	2.00E-25	34.78%
WP_155746905.1	hypothetical protein [*Scytonema sp. UIC 10036*]	*Scytonema sp. UIC 10036*	61.2	45%	3.00E-08	34.02%
BAZ36602.1	hypothetical protein NIES4101_25210 [*Calothrix sp. NIES-4101*]	*Calothrix sp. NIES-4101*	57.8	27%	5.00E-07	44.83%

Table 2: Similar proteins obtained from Swissprot database

Entry	Protein names	Identity	Score	E-value
A0A396TZK2	Uncharacterized protein (*Colwellia sp. RSH04*)	74.2%	894	1.3e-120
A0A545UCJ6	VHL domain-containing protein (*Aliikangiella sp. M105*)	34.3%	81	8.3e-28
A0A1Z4R2C0	VHL domain-containing protein (*Calothrix sp. NIES-4101*)	36.6%	150	1.5e-9
A0A1I6H391	Por secretion system C-terminal sorting domain-containing protein (*Robiginitalea myxolifaciens*)	37.1%	133	7e-6
A0A2S7JPT4	VHL domain-containing protein (*Limnohabitans sp. TS-CS-82*)	35.1%	124	2e-5

2. Analysis of physicochemical properties

The physicochemical properties of a protein can be characterized by an analysis of the analogous

properties of the amino acids. The hypothetical protein is negatively charged as the theoretical pI: 4.22 and the total number of positively (Arg + Lys) and negatively charged residues (Asp + Glu) were found to be 10 and 27, respectively. The computed instability index (II) was 32.71 classifying the protein as a stable one. The aliphatic index was 77.37 which gives an indication of proteins' stability over a wide temperature range and all the other properties have been summarized in table 3.

Table 3: Physicochemical properties of the hypothetical protein (WP_130598461.1)

Properties	Value
Molecular weight	23229.44
Theoretical pI	4.22
The total number of negatively charged residues (Asp + Glu)	27
The total number of positively charged residues (Arg + Lys)	10
The instability index (II) is computed to be	32.71
Formula	$C_{1024}H_{1552}N_{262}O_{346}S_5$
The total number of atoms	3189
Aliphatic index	77.37
Grand average of hydropathicity (GRAVY)	-0.261

3. Secondary structure analysis

The secondary structure of a protein can be able to provide some worthy information about the function. The query hypothetical protein shows the percentages of alpha-helix, beta-turn, extended strand, and the random coil of protein 21.13%, 9.91%, 33.33%, and 36.15%, respectively from SOPMA. The results of the secondary structure were also cross-checked by the PRISPRED server which shows a summary of similar results. The representative secondary structure of the hypothetical protein (WP_130598461.1) has been shown in Figure 3.

4. Assessment and validation of protein 3-dimensional structure

PROCHECK program was used for the validation of predicted tertiary structure, where the distribution of φ and ψ angle in the model within the limits are shown (Table 4, Figure. 4). The model was presumed to be a good one according to the Ramachandran Plot Statistics, with 91.1% residues in the most favored regions. Finally, the structure validation server Verifiy3D and ERRAT was implicated to verify the established model of 3D structure for the target sequence. In the Verify3D graph, 93.75% of the residues have averaged a 3D-1D score ≥ of 0.2 which indicates that the environmental profile of the model is good

and the overall quality factor predicted by the ERRAT server was 60.7143 indicates a quality model. From ProFunc, the average G-factors of the hypothetical protein are calculated to be -0.20, which indicates a usual protein model.

Table 4: Ramachandran plot statistics of the predicted 3D model for the target protein EMK97_00595 (WP_130598461.1)

Plot Statistics	Number of amino acid residues	Percentage (%)
Residues in the most favored regions [A, B, L]	51	91.1%
Residues in additional allowed regions [a, b, l, p]	4	7.1%
Residues in generously allowed regions [~a, ~b, ~l, ~p]	0	0.0%
Residues in disallowed regions	1	1.8%
Number of non-glycine and non-proline residues	56	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	4	
Number of proline residues	2	
Total number of residues	64	

5. Active site calculation

The active site of the selected hypothetical protein constituted by 11 amino acids of an area with 52.957 and a volume of 22.609. Chain X of the hypothetical protein shows the amino acids involved in the active site (F, V, Y, Y, T, L, E, V, T, Q, W), supplementary Figure 6 (A & B).

6. Assessment of protein subcellular localization and topology

The subcellular localization of the hypothetical protein seems to be an extracellular secretory signal peptide. Protein-sol and SOSUI both predict the hypothetical protein as a soluble protein. HMMTOP, TMHMM predicted the protein as a non-transmembrane protein (Table 5). The predicted topology of the protein has shown here from N terminal to the C terminal.

Table 5: Assessment of subcellular localization

Prediction	Servers	Results
Prediction of subcellular localization	Busca	Extracellular space, Signal peptide
	Cello	Extracellular
	PsortB	Unknown, Signal Peptide
	Cell-PLoc	Extracellular
	PSLpred	Extracellular protein
	SOSUIgramN	Outer membrane
Signal Peptide prediction	Predisi	Secreted protein, Signal peptide
	SignalP-5.0 Server	Signal Peptide
Prediction of protein solubility	SOSUI	Soluble protein
	Protein-sol	Soluble protein
Prediction of Transmembrane helices	HMMTOP	None
	TMHMM	None
	Sable	No transmembrane domain

7. Functional annotation of the hypotheticalprotein

The initial protein domain was achieved from the Conserved domain database (CDD) of NCBI. The region of the domain, superfamily, and family classifications have been determined by the servers – CDD, Pfam, SMART, Interpro, SCOP, Supfam, MotifFinder, ProFunc, Phyre 2, and CATH-Gene3D. The domain, Superfamily, and Family were selected based on the lowest e-value of the following domain. The higher e-value has been filtered out from the selection procedure. The e-value 9.11e-05 of VHL beta domain from ProFunc, 2.71e-09 of VHL superfamily from SCOP, 8.1e-03 of VHL family from Supfam indicate extremely good protein alignment respectively. The overall alignment range of the VHL beta domain was 133-212, VHL superfamily and Family were 144-200 respectively. Protein coil nature was determined by using PCoils from the Bioinformatics toolkit server. According to Phyre 2, the folding pattern of the following hypothetical protein is pre-albumin-like. On the other hand, PEF-FunSeqE is called the protein immunoglobulin-like. Both are secreted protein as well as soluble protein and hence provide a properly defined similarity indication of VHL protein (Table 6).

Table 6: Function annotation of hypothetical protein through the analysis of protein domain/superfamily/Family

Servers	Domain/Superfamily/Family	e-value/ Confidence	Region/ Alignment
Functional annotation from sequence			
Conserved Domain Database (CDD)	Superfamily: pVHL	6.22e-05	146-197
Pfam	Family: VHL (VHL beta domain)	1.3e-02	144-200
SMART	VHL	1.2e-02	133-205
Interpro	VHL superfamily	-	144-199
	VHL beta domain	-	131-212
Superfamily 1.75 (SCOP)	Superfamily: VHL	2.71e-09	144-199
	Family: VHL	8.1e-03	
Supfam	Superfamily: VHL	1.54e-09	144-199
	Family: VHL	8.1e-03	
Motif (From Pfam)	VHL beta domain	8.1e-03	146-200
Functional annotation from the 3D structure			
ProFunc	VHL beta domain	9.11e-05	131-191
Phyre 2	Superfamily: VHL Family: VHL	99.8% (Confidence)	135-212
CATH-Gene3D (From Interpro)	VHL beta domain	-	131-212

8. Analysis of protein network

The STRING interaction of VHL protein from *Homo sapiens* has been shown in Figure 8 as a model. VHL interacts with various proteins based on their combined score (table 7). The network has 11 nodes, 40 edges, average node degree 7.27, local clustering coefficient 0.819, expected number of edges 18, and the p-value of protein-protein interaction enrichment 7.07e-06 indicates the network has significantly more interactions than expected.

Table 7: Interacting proteins and their combined score from STRING 11.0 server

Interacted protein	Combined score
AKT1 (RAC-alpha serine/threonine-protein kinase)	0.997
AKT2 (RAC-beta serine/threonine-protein kinase)	0.994
CUL2 (Cullin-2; Core component of multiple cullin-RING-based ECS E3 ubiquitin-protein ligase complexes)	0.999
EGLN1 (Egl nine homolog 1)	0.989
EPAS1 (Endothelial PAS domain-containing protein 1)	0.994
HIF1A (Hypoxia-inducible factor 1-alpha)	0.999
PPP2CA (Serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform)	0.993
RBX1 (E3 ubiquitin-protein ligase RBX1)	0.982
TCEB1 (Elongin-C)	0.999
TCEB2 (Elongin-B)	0.998

Discussion

The sequence information as well as the structural information contributes to understanding the function of a hypothetical protein. This study aims to characterize a hypothetical protein, which showed strong homology with VHL superfamily, involved in tumor suppressor. Therefore, the amino acid sequence of the hypothetical protein EMK97_00595 [*Litorilituus sediminis*] was retrieved, and initially, the physicochemical properties were obtained by ExPASy's ProtParam tool and the prediction results are the deciding factors for the hydrophilicity, stability, and function of the protein [58]. The protein was considered as a stable one even in a wide temperature range as the instability index (II) and the aliphatic index were 32.71 and 77.37, respectively. And the query protein seems to be hydrophilic as the GRAVY was -0.261, supplementary table 3.

Protein structure is closely associated with its function. The secondary structure, viz. helix, sheet, turn and therefore the coil of any protein has an excellent association with the structure, function, and interaction of the protein. The query hypothetical protein contains the percentages of alpha-helix, beta-turn, extended strand, and the random coil 21.13%, 9.91%, 33.33%, and 36.15%, respectively. Findings from SOPMA revealed that the protein has an abundance of coiled regions that contributes to higher stability and conservation of the protein structure [58]. Moreover, the protein features a reliable helices percentage in its structure, which may facilitate folding by providing more flexibility to the structure, thus protein interactions could be increased [59].

For the prediction of the protein 3D model, HHpred was employed, where the highest identical template was selected for getting an acceptable model. The query protein WP_012259469.1 showed the highest

template identity of 25% with von Hippel-Lindau disease tumor suppressor; E3 ubiquitin ligase, transcription factor, hypoxic signaling, transcription; [*Homo sapiens*] with lowest E-value: 1.1e-11. Ramachandran plot analysis revealed that 91.1% of residues were located in the most favored regions. Moreover, residues in additional allowed regions and generously allowed regions were 7.1%and 0.0%, respectively, which evaluated the quality of the model to be good and reliable as it is generally accepted that if 90% of residues are in the most favored regions, it is likely to be a reliable model [60], shown in Fig. 4(B). The model is compatible with its sequence as Verify 3D analysis implies that 93.75% of the residues had an average 3D–1D score of ≥0.2. "Overall quality factor" was estimated by ERRAT, which is used to evaluate the amino acid environment for non-bonded atomic interactions. Higher scores indicate higher quality, and the query protein's quality factor was 60.7143, which is greater than the generally accepted range (>50) for a high-quality model [61]. The average G-factor of the query protein is -0.20 obtained from ProFunc analysis, which indicates a usual protein model.

Protein's active site was determined by CASTp, containing 11 amino acids (F, V, Y, Y, T, L, E, V, T, Q, W) of an area with 52.957 and a volume of 22.609, shown in figure 5(A & B). The subcellular localization obtained from CELLO, BUSCA, and other similar servers, seems to be an extracellular secretory signal peptide and non-transmembrane (Table 5). As the functions of secreted proteins are diverse, the query hypothetical protein may work like paracrine, autocrine, endocrine, or neuroendocrine depending on the target [62]. Solubility is the most important factor and an excellent index for protein functionality. Protein-sol and SOSUI both predict the hypothetical protein as a soluble one, so it may possess good dispersibility and lead to the formation of finely dispersed colloidal systems.

The superfamily, family, and domain information have been determined by a combinational sequence and structural informative approach based on the e-value of different sequence and structure analysis servers. These servers suggested the following hypothetical protein EMK97_00595 from the organism *Litorilituus sediminis* to be a VHL beta domain from the VHL superfamily. VHL tumor suppressor protein can play role in tumor suppression in multiple ways and the most common of them is targeting the hypoxia-inducible transcription factor (HIF) that mediated tumor suppression activity through polyubiquitylation and proteasomal degradation [63]. The major contribution of Von Hippel-Lindau tumor suppressor protein (pVHL) is to suppress clear-cell renal cell carcinoma in kidney cancer [63][64].

Litorilituus sediminis is a novel species and the investigated protein EMK97_00595 is also novel so there is no specific STRING derived protein-protein network is available for this organism. The protein-protein interaction network analysis shown here from *Homo sapiens* is just for a supposition model to evaluate how the protein interacted in humans. The protein-protein interaction of VHL-HIF1A (Hypoxia-inducible factor 1-alpha) with a combined score of 0.999 indicated a strong relationship between these two proteins. The interaction between VHL and HIF1A indicating the involvement of the same pathway to suppress tumor activity [12].

Overall, the combinational strategy of computing physicochemical properties, evaluating the secondary structure and tertiary structure information, and domain information analysis denoted the protein as Von

Hippel–Lindau tumor suppressor protein that is associated with Von Hippel–Lindau disease.

Conclusion

Protein is the building block of life that serves both biological processes and molecular functions in living organisms. Hence, this study investigated the functional role of a hypothetical protein from a novel bacterium, (*Litorilituus sediminis*) that possesses a significant tumor suppression activity. The employment of highly recommended bioinformatics tools to analyze the combinational sequence and structural information revealed the underlying molecular function of the examined hypothetical protein. The current investigation suggested that the hypothetical protein may exhibit a VHL beta-domain that is similar to the human VHL beta-domain and is also a part of Von Hippel–Lindau tumor suppressor protein (pVHL). Therefore, this finding with the aid of bioinformatics tools can soften our viewpoint for further investigation and experimental validation of this hypothetical protein containing VHL beta domain, and the use of this hypothetical protein with the aid of modern biotechnology might be utilized to suppress tumor progression in higher organisms such as human as an alternative to human defective or mutated VHL protein in the near future.

Abbreviations

VHL: von Hippel– Lindau

PE: Phosphatidylethanolamine

PG: phosphatidylglycerol

PN: aminophospholipid

AL: amino lipids

ORF: open reading frame

DUF: domain of unknown functions

CDD: Conserved Domain Database

GRAVY: grand average of hydropathicity

CASTp: Computed Atlas of Surface Topography

AKT: RAC-alpha serine/threonine-protein kinase

CUL2: Cullin-2

EPAS: Endothelial PAS domain-containing protein

HIF1A: Hypoxia-inducible factor 1-alpha

PPP2CA: Serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform

References

1. Wang Y, Zhao R, Ji S, Li Z, Yu T, *et al.* Litorilituus sediminis gen. nov. sp. nov., isolated from coastal sediment of an amphioxus breeding zone in Qingdao, China. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol* 2013; 104:423–430.
2. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, *et al.* Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* 2010; 28:1248–1250.
3. Naveed M, Tehreem S, Usman M, Chaudhry Z, Abbas G. Structural and functional annotation of hypothetical proteins of human adenovirus: prioritizing the novel drug targets. *BMC Res Notes* 2017; 10:706.
4. Bashir Z, Rizwan M, Mushtaq K, Munir A, Ali I. In Silico Structural and Functional Prediction of Phaseolus vulgaris Hypothetical Protein PHA VU_004G136400g. *J Proteomics Bioinform* 2017; 10:207–211.
5. Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, *et al.* Exploration of uncharted regions of the protein universe. *PLoS Biol* 2009; 7. doi:10.1371/journal.pbio.1000205
6. Bharat Siva Varma P, Adimulam YB, Kodukula S. In silico functional annotation of a hypothetical protein from Staphylococcus aureus. *J Infect Public Health* 2015; 8:526–532.
7. Mudgal R, Sandhya S, Chandra N, Srinivasan N. De-DUFing the DUFs: Deciphering distant evolutionary relationships of Domains of Unknown Function using sensitive homology detection methods. *Biol Direct* 2015; 10:1–23.
8. Nimrod G, Schushan M, Steinberg DM, Ben-Tal N. Detection of Functionally Important Regions in “Hypothetical Proteins” of Known Structure. *Structure* 2008; 16:1755–1763.
9. Shahbaaz M, Hassan MI, Ahmad F. Functional annotation of conserved hypothetical proteins from Haemophilus influenzae Rd KW20. *PLoS One* 2013; 8. doi:10.1371/journal.pone.0084263
10. Mohan R. Computational structural and functional analysis of hypothetical proteins of Staphylococcus aureus. *Bioinformation* 2012; 8:722–728.
11. Mills CL, Beuning PJ, Ondrechen MJ. Biochemical functional predictions for protein structures of unknown or uncertain function. *Comput Struct Biotechnol J* 2015; 13:182–191.
12. Haase VH. The VHL/HIF oxygen-sensing pathway and its relevance to kidney disease. *Kidney Int* 2006; 69:1302–1307.
13. Zhang Q, Yang H. The roles of VHL-dependent ubiquitination in signaling and cancer. *Front Oncol* 2012; 2:1–7.
14. Blankenship C, Naglich JG, Whaley JM, Seizinger B, Kley N. Alternate choice of initiation codon produces a biologically active product of the von Hippel Lindau gene with tumor suppressor activity.

Oncogene 1999; 18:1529–1535.

15. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res* 2008; 36:5–9.
16. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003; 31:365–370.
17. Madeira F, Park YM, Lee J, Buso N, Gur T, *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 2019; 47:W636–W641.
18. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009; 25:1189–1191.
19. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, *et al.* The Proteomics Protocols Handbook. *Proteomics Protoc Handb* 2005; :571–608.
20. Atsushi I. Thermostability and aliphatic index of globular proteins. *J Biochem* 1980; 88:1895–1898.
21. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982; 157:105–132.
22. Guruprasad K, Reddy BVB, Pandit MW. Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng Des Sel* 1990; 4:155–161.
23. Geourjon C, Deléage G. Sopma: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Bioinformatics* 1995; 11:681–684.
24. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* 2000; 16:404–405.
25. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* 2018; 430:2237–2243.
26. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 2003; 31:3381–3385.
27. Laskowski RA, Watson JD, Thornton JM. ProFunc: A server for predicting protein function from 3D structure. *Nucleic Acids Res* 2005; 33:89–93.
28. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 1993; 26:283–291.
29. Colovos C, Yeates TO. Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Sci* 1993; 2:1511–1519.
30. Kihara D, Chen H, Yang Y. Quality Assessment of Protein Structure Models. *Curr Protein Pept Sci* 2009; 10:216–228.
31. Luthy R, Bowei J, Einsenberg D. Verify3D: Assessment of protein models with three-dimensional profiles. *Methods Enzymol* 1997; 277:396–404.
32. Tian W, Chen C, Lei X, Zhao J, Liang J. CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res* 2018; 46:W363–W367.

33. Savojardo C, Martelli PL, Fariselli P, Profiti G, Casadio R. BUSCA: An integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res* 2018; 46:W459–W466.
34. Yu C-S, Lin C-J, Hwang J-K. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n -peptide compositions . *Protein Sci* 2004; 13:1402–1406.
35. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, *et al.* PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 2010; 26:1608–1615.
36. Shen H Bin, Chou KC. Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *J Theor Biol* 2010; 264:326–333.
37. Imai K, Asakawa N, Tsuji T, Akazawa F, Ino A, *et al.* SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in Gram-negative bacteria. *Bioinformation* 2008; 2:417–421.
38. Bhasin M, Garg A, Raghava GPS. PSLpred: Prediction of subcellular localization of bacterial proteins. *Bioinformatics* 2005; 21:2522–2524.
39. Hiller K, Grote A, Scheer M, Münch R, Jahn D. PrediSi: Prediction of signal peptides and their cleavage positions. *Nucleic Acids Res* 2004; 32:375–379.
40. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 2019; 37:420–423.
41. Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. Protein-Sol: A web tool for predicting protein solubility from sequence. *Bioinformatics* 2017; 33:3098–3100.
42. Mitaku S, Hirokawa T. Physicochemical factors for discriminating between soluble and membrane proteins: Hydrophobicity of helical segments and protein length. *Protein Eng* 1999; 12:953–957.
43. Tusnády GE, Simon I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* 2001; 17:849–850.
44. Möller S, Croning MDR, Apweiler R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 2001; 17:646–653.
45. Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. *J Comput Biol* 2005; 12:355–369.
46. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, *et al.* CDD/SPARCLE: The conserved domain database in 2020. *Nucleic Acids Res* 2020; 48:D265–D268.
47. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* 2019; 47:D427–D432.
48. Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res* 2021; 49:D458–D460.
49. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 2021; 49:D344–D354.
50. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* 2001;

313:903–919.

51. Pandurangan AP, Stahlhacke J, Oates ME, Smithers B, Gough J. The SUPERFAMILY 2.0 database: A significant proteome update and a new webserver. *Nucleic Acids Res* 2019; 47:D490–D494.
52. Pandit SB, Bhadra R, Gowri VS, Balaji S, Anand B, *et al.* SUPFAM: A database of sequence superfamilies of protein domains. *BMC Bioinformatics* 2004; 5:1–5.
53. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. Trabajo práctico Nº 13 . Varianzas en función de variable independiente categórica. *Nat Protoc* 2016; 10:845–858.
54. Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, *et al.* CATH: Expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res* 2019; 47:D280–D284.
55. Shen H Bin, Chou KC. Predicting protein fold pattern with functional domain and sequential evolution information. *J Theor Biol* 2009; 256:441–446.
56. Gruber M, Söding J, Lupas AN. Comparative analysis of coiled-coil prediction methods. *J Struct Biol* 2006; 155:140–145.
57. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019; 47:D607–D613.
58. Hasan A, Mazumder HH, Chowdhury AS, Datta A, Khan A. Molecular-docking study of malaria drug target enzyme transketolase in Plasmodium falciparum 3D7 portends the novel approach to its treatment. *Source Code Biol Med* 2015; :1–14.
59. Butt AM, Batool M, Tong Y. functional annotation of Mycoplasma genitalium hypothetical protein MG _ 237. 2011; 7.
60. Hooda V, Gundala P, Chinthala P. Sequence analysis and homology modeling of peroxidase from Medicago sativa Abstract : Background : 2012; 8.
61. Messaoudi A, Belguith H, Hamida J Ben. Evolutionary Bioinformatics Three-Dimensional Structure of Arabidopsis thaliana Lipase Predicted by Homology Modeling Method. ; :99–105.
62. Farhan H, Rabouille C, Farhan H, Rabouille C. Signalling to and from the secretory pathway Signalling to and from the secretory pathway. Published Online First: 2011. doi:10.1242/jcs.086991
63. Article O. Treatment of Kidney Cancer. Published Online First: 2009. doi:10.1002/cncr.24232
64. Clark PE. The role of VHL in clear-cell renal cell carcinoma and its relation to targeted therapy. *Kidney Int* 2009; 76:939–945.

Figures

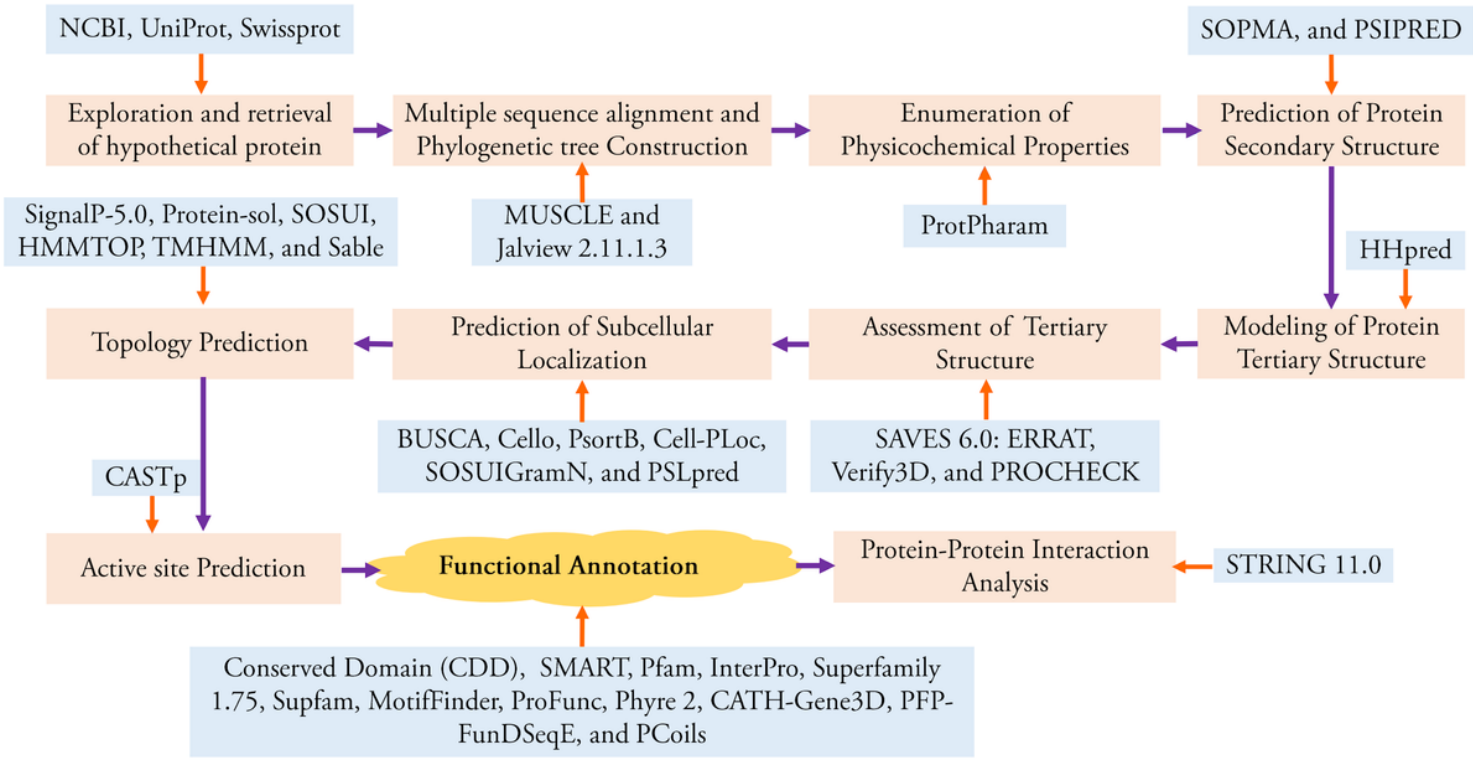

Figure 1

A schematic representation of the overall experimental design.

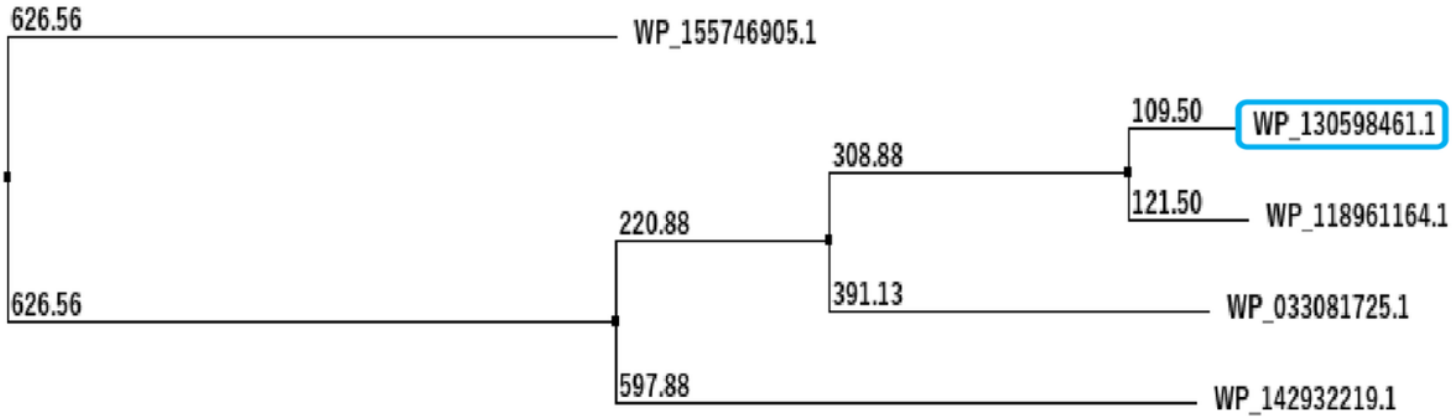

Figure 2

Evolutionary analysis of different VHL proteins with the target protein shown in the blue box (WP_130598461.1).

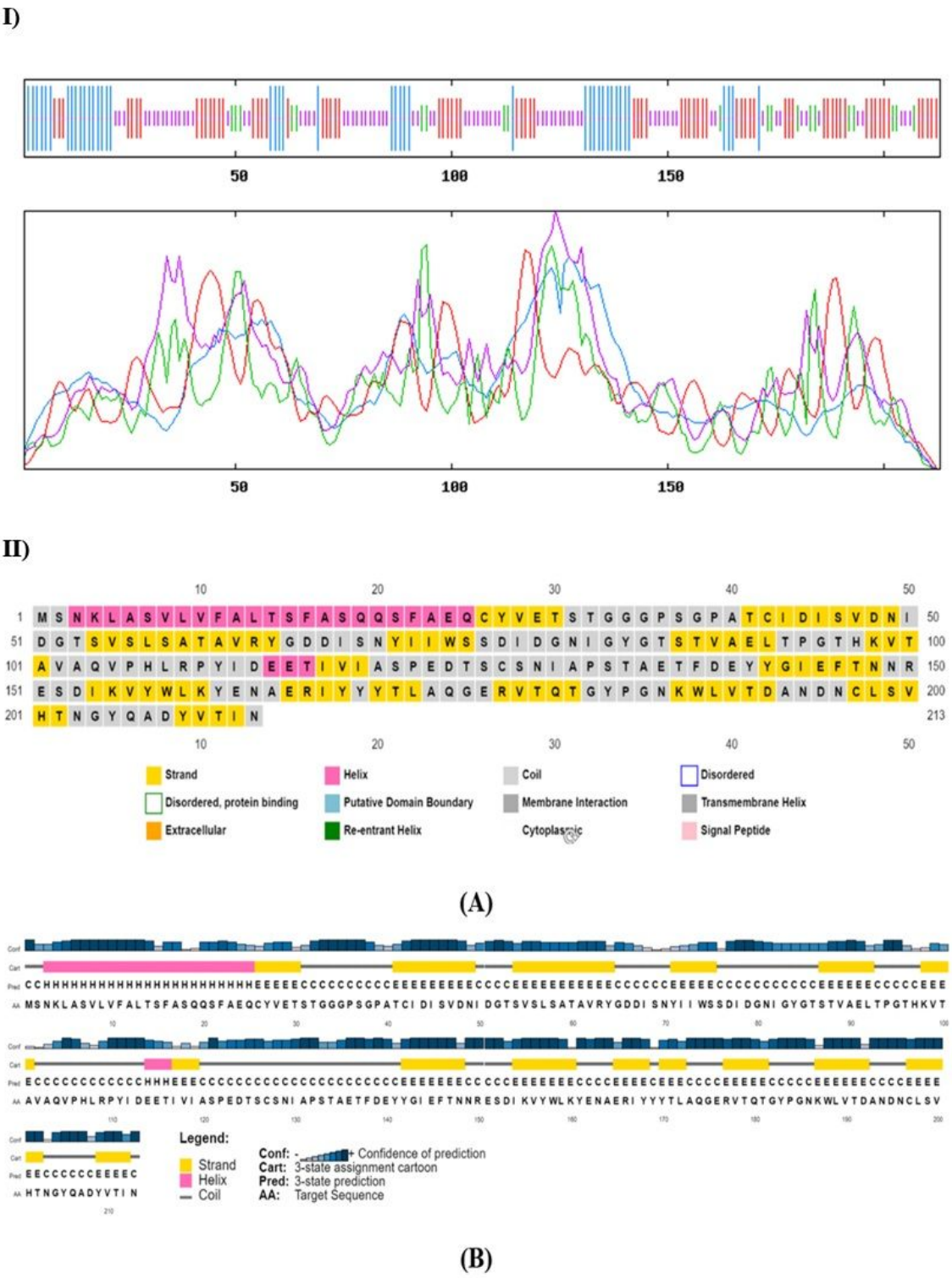

Figure 3

Model of secondary structure. (I) Secondary structure information from SOPMA server, and II (A) and (B) Sequential organization and graphical visualization of secondary structure from PSIPRED.

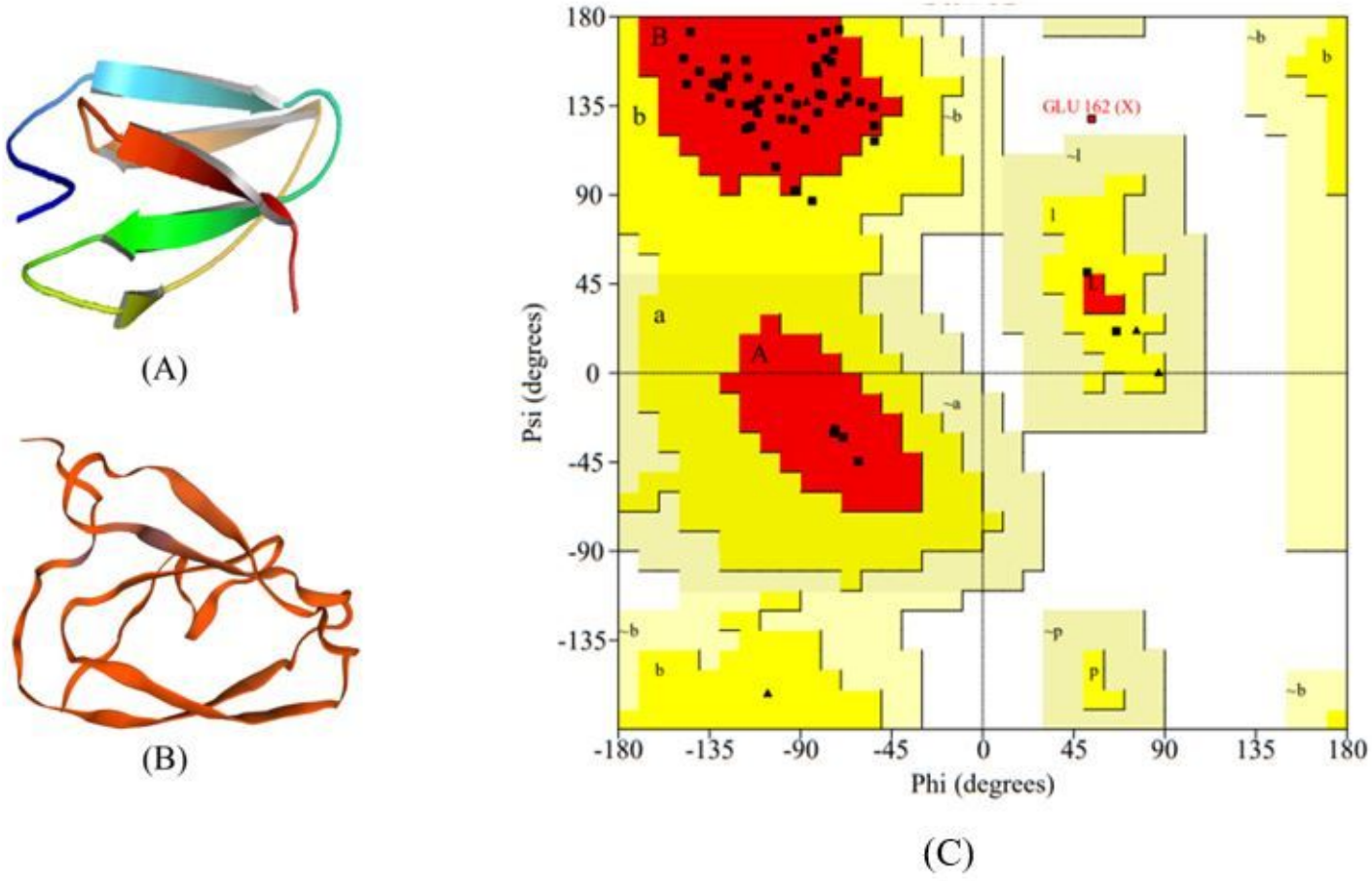

Figure 4

Graphical representation and assessment of protein 3D structure. (A) Predicted 3-dimensional structure from SAVES server (Pymol view), (B) from SWISS-MODEL, and (C) Ramachandran plot analysis of 3D modeled structure validated by PROCHECK program.

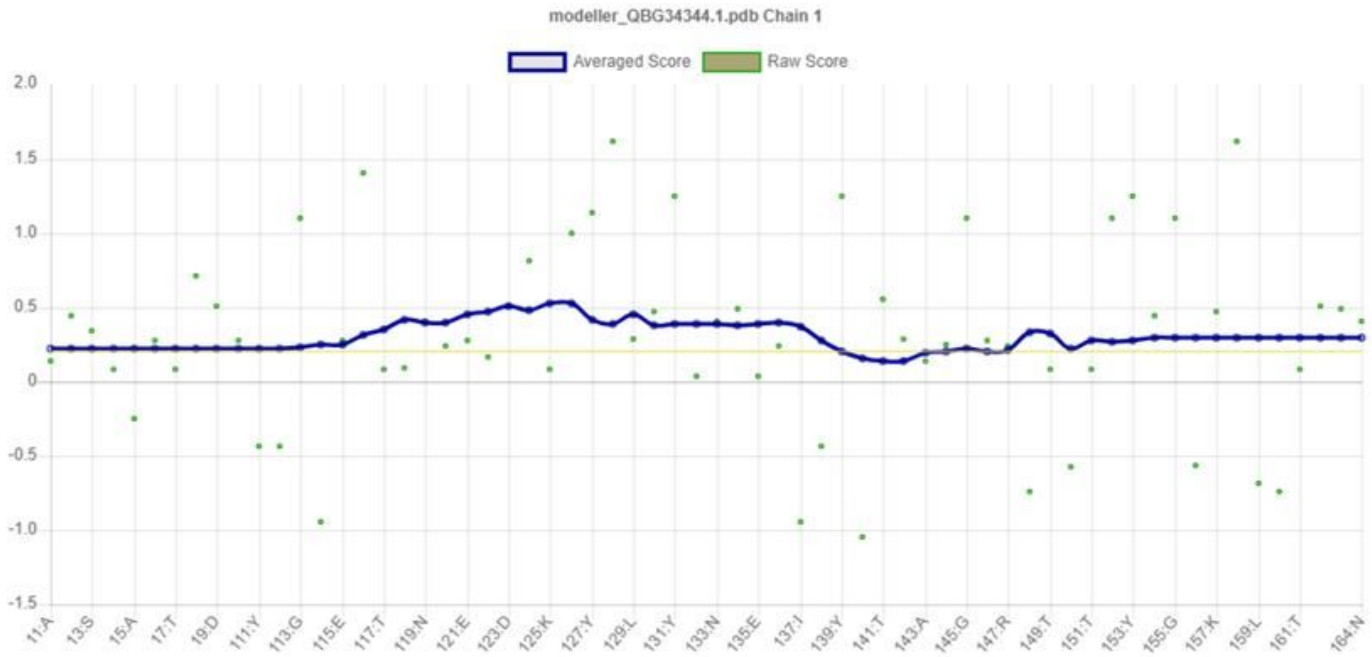

Figure 5

3D-structure validation by Verifiy3D.

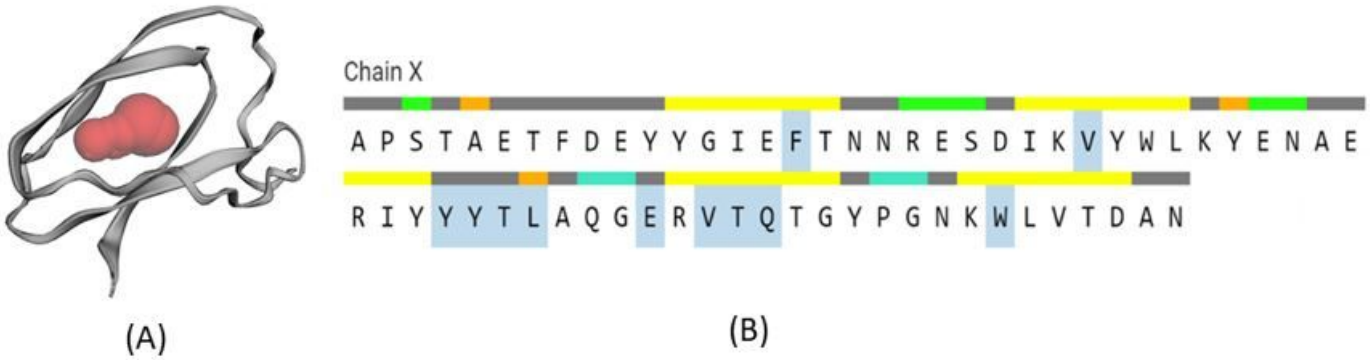

Figure 6

Active site of the hypothetical protein, (A) Binding site of the hypothetical protein indicated by red region, and (B) Amino acids involved in the active site.

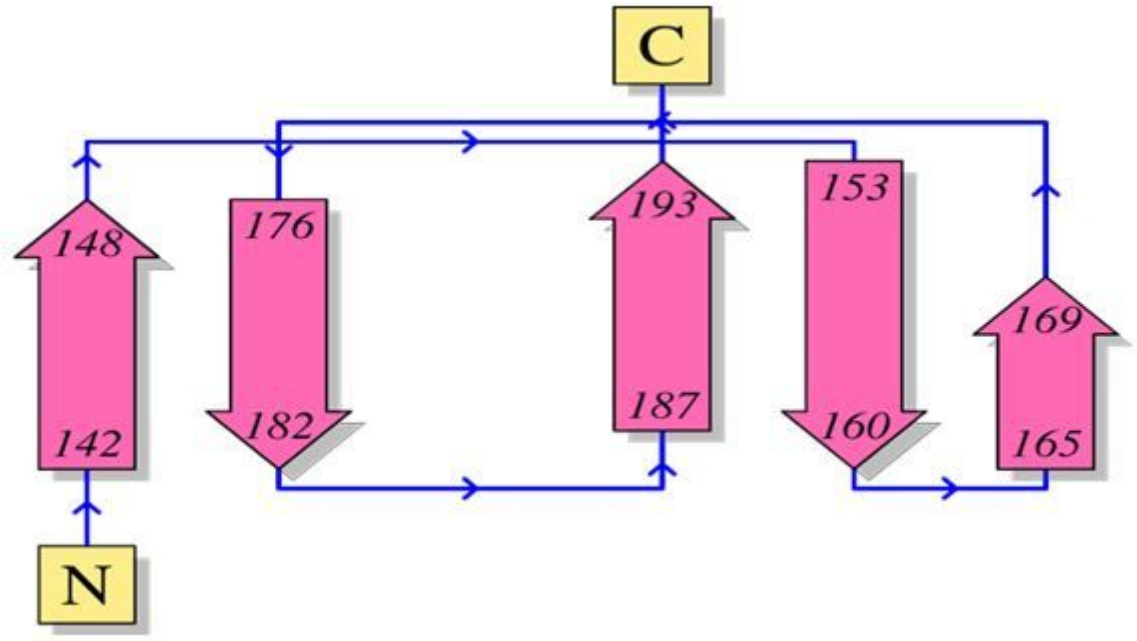

Figure 7

Topology of hypothetical protein

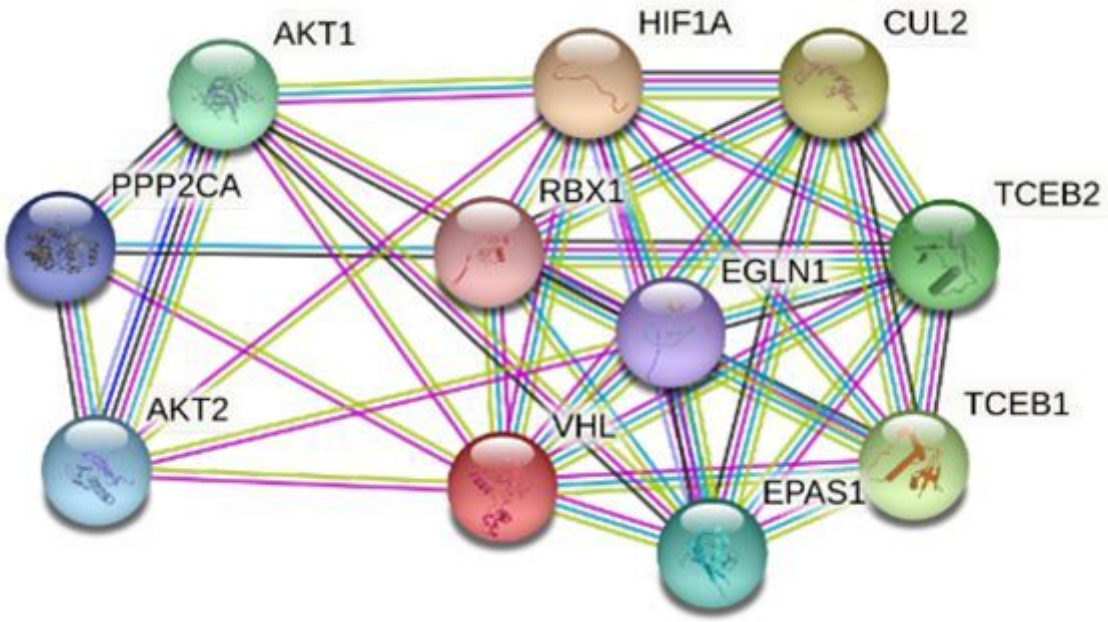

Figure 8

Protein-protein interaction network of the hypothetical VHL protein

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile.docx