Research Article

In silico structural and functional assessment of hypothetical protein L345_13461 from *Ophiophagus hannah*

Zainab Jan¹, Syed Umair Ahmad^{1*}, Amara Qadus², Yasir Ali³, Waseem Sajjad⁴, Fatiha Rais¹, Sitara Tanveer⁵, Muhammad Shehzad Khan⁶ and Ihteshamul Haq⁴

1. Department of Bioinformatics Hazara University Mansehra-Pakistan

2. Department of Bioinformatics and Biotechnology Islamic university Islamabad-Pakistan

3. Department of Computer Science and Bioinformatics, Khushal khan Khattak University Karak-Pakistan

4. Department of Biotechnology and Genetic Engineering, Hazara University Mansehra-Pakistan

5. Department of Biotechnology University of Management and Technology Lahore- Pakistan

6. Department of Physics City University-Hong Kong

*Corresponding author's email: ahmadumair927@gmail.com

Citation

Zainab Jan, Syed Umair Ahmad, Amara Qadus, Yasir Ali, Waseem Sajjad, Fatiha Rais, Sitara Tanveer, Muhammad Shehzad Khan and Ihteshamul Haq. *In silico* structural and functional assessment of hypothetical protein L345_13461 from *Ophiophagus hannah*. Pure and Applied Biology. Vol. 10, Issue 4, pp1109-1118. <http://dx.doi.org/10.19045/bspab.2021.100116>

Received: 16/09/2020

Revised: 05/01/2021

Accepted: 20/01/2021

Online First: 25/01/2021

Abstract

Ophiophagus Hannah is extensively dispersed throughout various portions of the Asia. Many toxin proteins have been identified from their venom which are pharmacology active. Current research goals is to investigate the structural and function assessment of a hypothetical protein L345_13461 of *Ophiophagus hannah* for a better understanding of king cobra venom. Using *in silico* approach, the 3D structure was generated by Homology modelling, while the functions was profiled by ProFunc tool. The primary and secondary structure analysis revealed that the hypothetical protein L345_13461 is a stable and located in cytoplasm comprising of a remarkable number of random coils. Homology modelling was accomplished employing SWISS-MODEL server where the templates identity of 34.76% was observed with PDB ID 5ZZ3. Quite a few evaluations of quality assessment and validation parameters created a stable protein model with prodigious quality. Functional analysis was accomplished through InterProScan, ProFunc, DeepGoPlus and KEGG KAAG, suggesting that the hypothetical protein is a cytoplasmic protein, which plays important roles in protein binding, metabolism, signalling and cellular processes, genetic information processing. Finally, we proposed that experimental support would assist to investigate the structures and functions of other hypothetical proteins of various living organisms.

Keywords: Function predication; hypothetical protein; Structure assessment; *Ophiophagus hannah*

Introduction

Hypothetical proteins (HPs) are proteins whose occurrence has been evaluated although their function but has not been predicted yet via *in vivo* approach [1]. HPs usually cover approximately huge slice of

the protein coding regions in several genomes. Even though their functions are not still categorized well, they could have their own consequence to complete genomic and proteomic material [2, 3] Appropriate structural and functional

observations of HPs of precise genome may shift to the detection of novel structures and its new functions and help to described additional protein pathways and cascades, consequently finishing our patchy information on the mosaic of proteins [4]. Illuminating the structural and functional enigmas of these HPs could promote to a well suitable comprehension of the protein-protein connections and protein-protein networks in various sort of life including plants, microorganisms [5]. Additionally, new HPs could correspondingly serve as markers and pharmacological targets for drug development, finding, and screening [6, 7].

In current era, several hypothetical proteins have been identified in the genome of many life forms. But, because of a few limitations, such as the cost and time desired for experimental procedures, whole genome annotations have not reached yet. Likewise, the wide-ranging number of theoretical proteins in a genome makes their study a problematic mission. Bioinformatics apply Insilco approaches to find out gene loci, predict the transcripts of a particular gene or structure and location of a particular protein inside cell and the disease(s) associated with the abnormal structure or function of that particular protein [8].

Ophiophagus hannah (King Cobra) is broadly spread transversely in the Indian subcontinent, south eastern Asia, and the southern portion of China [9]. By way of the common name Ophiophagus (Greek: Ophio mean snakes, -phagous mean eating) recommends, its diet composed of mostly snakes, while small mammals are likewise eaten. The Ophiophagus being a monotypic genus having only one species discovered recently. But its enormous environmental spreading escorted through morphological transformations proposes probable taxonomic deviation and alterations in venom conformation. Subsequently in the beginning of 1970's, the king cobra venom has been studied in detail. Several fragments have been secluded and

categorized, comprising L-amino acid oxidases [10], metalloproteinases [11], three finger toxins (3FTxs) [12, 13], phospholipases A2 (PLA2s) [14, 15], Ohanin [16], kunitz-type protease inhibitors [17] and factor X activator [18]. On the other hand, the implication of the beginning and intraspecific divergence of the king cobra venom has only been lately recognized. In a study beside venom samples obtained from some countries, variances in venom characteristics were determined amongst Chinese and Southeast Asian king cobras [19].

The consequences of current work will be suitable for better understanding of the mechanism of selected protein and determining further novel proteins and their functions through similar technique that we have employing for Hypothetical protein L345_13461 from *Ophiophagus Hannah*.

Material and methods

1. Retrieval of protein sequences

The amino acid sequence of the Hypothetical protein L345_13461 from *Ophiophagus hannah* (Accession number: V8NEU2) was retrieved from the Uniprot database (<http://www.uniprot.org/>).

2. Analysis of physicochemical properties

ProtParam tool (<http://web.expasy.org/protparam/>) was used to explore the physicochemical characteristics of particular hypothetical protein such as molecular weight, theoretical pI, amino acid composition, atomic composition, instability index, negatively Charged Residue, positively Charged Residue and grand average of hydropathicity (GRAVY) [20].

3. Subcellular localization prediction

Subcellular localization of Hypothetical Protein L345_13461 from *Ophiophagus Hannah* was obtained through SOSUI server [21]. Subcellular localization means to discover the exact position of L345_13461 in a cell.

4. Secondary structure predication

An online secondary structure predication tool SOPMA (Self-Optimized Prediction Method with Alignment) was used to

predict secondary structure of Hypothetical Protein L345_13461 which stipulate information about alpha helix, beta sheets and coils [22]. Cofactory-1.0 server was employed to find the potential cofactor binding Rossmann folds and prediction of FAD, NAD and NADP cofactors specificity. DistanceP 1.0 was utilized to predict the protein distance constraints. Finally, we used Easypred 1.0 for the predication of binding motif [23].

5. Tertiary structure predication

Amino acid sequences of Hypothetical Protein L345_13461 was provided as an input to SWISS-MODEL in FASTA Format. In Swiss-model 3D model of protein is built via homology modelling approach [24, 25].

6. Visualization and Quality assessment of 3D structure

The primary structural model was checked for finding of faults in 3D structure [26] through ERRAT and Verify3D programs incorporated in structural assessment and confirmation server SAVES (<http://nihserver.mbi.ucla.edu/SAVES/>).

3D model was hired to illustrate Ramachandran plot using RAMPAGE server [27] and validate by using Procheck. In this mode, we were ingenious to recognize just how residues are locating in favoured sections plus in allowed and outlier sections. All the visual presentations were completed by the help of USCF Chimera 4.1 [26].

7. Functional annotation of the protein

Finally, the function of Hypothetical protein L345_13461 was analyzed by four different bioinformatics tools and databases including KAAS [28], ProFunc [29], InterProScan [30], and DeepGoPlus [31], were utilized for this reason.

8. Submission of 3D model

The 3D model created for Hypothetical protein L345_13461 was successfully submitted in Protein model database (PMDB) (<http://bioinformatics.cineca.it/PMDB/>).

Results and Discussion

Physiochemical properties of hypothetical protein L345_13461

We employed the ExPASy's ProtParam server to inspect the imaginary physiochemical properties of hypothetical protein L345_13461 using its amino acid sequence. Most of the calculations in this server display protein steadiness and stability because the stability is recognized with its suitable function aptitude [32]. Further, we investigated that the protein overall comprise of 220 amino acids, having a molecular weight of 24506.08 Daltons while its isoelectric point (PI) is 8.90 representing a positively charged protein. The instability index of the protein was calculated 39.27, categorized this protein as stable. The negative GRAVY index of -0.209 which depicts that protein is hydrophobic and soluble in nature. The utmost copious amino acid residue was detected to be Valine (21), Leucine (21) and Glycine (21) tracked by Serine and Lysine (18 each). The lowest was detected as Histidine (1). The sequence comprise of 24 negatively charged residues (Aspartic acid + Glutamic acid) and 28 positively charged residues (Arginine + Lysine). The molecular formula of the protein was investigated as $C_{1114}H_{1726}N_{292}O_{319}S_6$ while the total number of atoms in the protein is 3457.

Subcellular localization of hypothetical protein L345_13461

Protein subcellular localization predictions comprise the computational expectancy of where a protein exists inside a cell. Envisaging subcellular localization of unidentified proteins can provide evidence about their cellular functions. This evidence could be employed in better comprehension of disease mechanism and drugs designing [33]. The subcellular localization of the hypothetical protein L345_13461 protein was evaluated by SOSUI server and find that the protein is cytoplasmic and validated through PSORTb v3.2.0 and Predict Protein servers.

Secondary Structure of hypothetical protein L345_13461

Initially the secondary structure of the protein was identified through SOPMA server. The random coils was detected as the most predominant (36.82%), followed through extended strand (33.64%) and alpha helix (18.64%). Moreover, beta turn was detected as 10.91%. Furthermore, identical outcomes were gained through Predict Protein and PSIPRED servers. The illustrative secondary structure of L345_13461 created from the PSIPRED

server [34] is presented in (Fig. 1). Using Cofactory 1.0 server, we found that 1 Rossmann folds, 0.001 FAD, 0.149 NAD, 0.022 NADP cofactor were observed from residue number 150 to 192. Protein Distance Constraints Matrix [35] was build using DistanceP1.0 and shown in (Fig. 2). EasyPred 1.0 was employed for binding motif predication. Initially clustering was done using Henikoff & Henikoff 1/nr method while visualization is perform using the Kullback-Leibler method with the Seq2Logo-2.0 program (Fig. 3).

Figure 1. Secondary structure of hypothetical protein L345_13461

Figure 2. Contact Distance Matrix; Red represent predicted contact while blue represented predicted non-contact. The upper triangle is a Binary contact matrix while the lower triangle is a Probability matrix

Figure 3. SeqLogo of Binding Motif; Red colour show acid, Green colour show neutral, black colour show hydrophobic while blue colour show basic

Homology modelling of hypothetical protein L345_13461

Homology modelling envisages the 3D structure of a specified protein sequence and build model relating to its alignments to one or more proteins of well-known structure [36].

To accomplish the homology modelling, the sequence of hypothetical protein L345_13461 was provided as input in SWISS-MODEL server [37]. The server immediately did BLASTP hunt for each protein sequence to recognize templates for homology modelling. For each identified template, the quality of the templates has been evaluated from topographies of the target-template alignment. The templates with the maximum quality have then been nominated for model construction. In this specific hunt, PDB ID 5ZZ3 was nominated as the template for homology modelling which is an X-ray diffraction model of a Butyrophilin protein with an 34.76% sequence identity, which was a virtuous score to initiate modelling. The 3D model was observed through UCFS Chimera 1.8.1 and shown in (Fig. 4).

Quality assessment and validation of hypothetical protein L345 13461

Reliability of the created model was originally validated with ERRAT that checked the statistics of non-bonded

associations amongst diverse atom categories based on characteristic atomic associations. The overall quality factor was depicting as 83.5714 which is virtuous to usage this model. While designated through the Verify3D program, the outcomes presented that 87.88% of hypothetical protein residues had an average 3D (atomic model) – 1D (amino acid) score ≥ 0.2 corresponding suggesting that these structures were suitable, well-suited and virtuous (Fig. 5).

Z-score of the hypothetical protein L345_13461 model was identified through PROSAweb. The Z-score is used to guess the model quality by means of structured resolved proteins as orientations [33]. The z-score of the hypothetical protein was found to be -6.02, signifying the model is virtuous, presented in (Fig. 6). The stereo chemical quality of the model hypothetical protein L345_13461protein was observed by means of Ramachandran plots through the Procheck server and validated by using RAMPAGE. Ramachandran plot analysis detected 93.3% of residues of the protein's model structure in the favoured region, with 5.5% and 1.2% residues in allowed and outlier regions, exclusively, demonstrating that the model was reliable and of better quality shown in (Fig. 7). The final protein

structure was placed in PMDB and is available under ID PM0083165.

Figure 4. 3D structure of hypothetical protein L345_13461; Yellow colour represent strand, Cyan colour represent coils, Green colour represent helix

Figure 5. Verification of a 3D model via Verify3D

Figure 6. Relationship of HP with other protein based on z-score where HP shows in black dots while grey colour shows other protein

Figure 7. Ramachandran Plot of HPs- L345_13461

Function Annotation of hypothetical protein L345_13461

Prediction of protein functions is a key mission in bioinformatics that is significant in considerate the functions of proteins in disease pathobiology, the role of metagenomes, or discovery drug targets. An extensive variety of approaches have been established for envisaging protein functions computationally [38]. We employed four online web tools to explore the possible functions of selected hypothetical protein. Considering

predictions completed through InterProScan, ProFunc, DeepGoPlus and KEGG KAAG, we suggest that hypothetical protein L345_13461 comprise of SPRY domain and concanavalin A-like lectin-glucanase domain and play various function. The protein was found to be crucial for the cellular signalling process, metabolism, molecular and biological functions, and protein binding. The function of hypothetical protein L345_13461 is presented in (Table 1).

Table 1. Functional annotation of hypothetical protein L345_13461

Cellular Components and cellular processes	Molecular function	Biological Function
outer membrane\bounded periplasmic space periplasmic space ATP\binding cassette (ABC) transporter complex ATP\binding cassette (ABC) transporter complex\ substrate\binding subunit\ containing Transporters Secretion System Bacterial Toxins Two-Component System Bacterial Motility Proteins Cytoskeleton Proteins <u>Exosome</u> Prokaryotic Defense System	Protein Binding Carbohydrate Binding Galactoside Binding Enzymes Protein Kinases Protein Phosphatases and Associated Proteins Peptidases and Inhibitors Glycosyltransferases Lipopolysaccharide Biosynthesis Proteins Peptidoglycan Biosynthesis and Degradation Proteins Lipid Biosynthesis Proteins Polyketide Biosynthesis	Response to stimulus Biological Regulation Metabolic Process Defense Response positive regulation of cellular process immune system process nitrogen compound metabolic process organic substance metabolic process response to stress cellular process response to DNA damage stimulus Maltose Transport Carbohydrate Transport Transcription Factors Transcription Machinery Messenger RNA Biogenesis

G Protein-Coupled Receptors Cytokine Receptors Pattern Recognition Receptors Nuclear Receptors Ion Channels GTP-Binding Proteins Cytokines and Growth Factors Cell Adhesion Molecules Proteoglycans Glycosaminoglycan Binding Proteins CD Molecules Glycosylphosphatidylinositol (GPI)-anchored Proteins Lectins Antimicrobial Resistance Genes Viral Proteins Non-coding RNAs	Proteins Prenyltransferases Amino Acid Related Enzymes Cytochrome P450 Photosynthesis Proteins	Spliceosome Ribosome Ribosome Biogenesis Transfer RNA Biogenesis Translation Factors Chaperones and Folding Catalysts Membrane Trafficking Ubiquitin System Proteasome DNA Replication Proteins Chromosome and Associated Proteins DNA Repair and Recombination Proteins Mitochondrial Biogenesis
--	---	---

Conclusion

The current research was focused to produce the 3D structure and suggest possible functions of the hypothetical protein L345_13461 belong to *Ophiophagus Hannah*. The 3D model of the protein was created by means of Homology Modelling and polished through few structural assessment approaches and the concluding result was sincerely prodigious. We detected that this novel protein is a stable cytoplasmic protein having SPRY domain and concanavalin A-like lectin-glucanase domain. The protein involved in various biological, molecular, metabolic, genetic, and cellular signalling processing. Additionally, this kind of procedure could be supportive in the structure and functions annotation of further hypothetical proteins.

Authors' contributions

Conceived and designed the experiments: Z Jan & SU Ahmad, Performed the experiments: A Qadus & S Tanveer, Analyzed the data: MS Khan, W Sajjad, Y Khan & I Haq, Contributed materials/analysis/ tools: Z Jan & AQF Rais, Wrote the paper: Z Jan & SU Ahmad.

References

- Galperin, MY & Koonin, EV (2004). Conserved hypothetical proteins: prioritization of targets for

experimental study. *Nucl Acids Res* 32(18): 5452-5463.

- Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial, M & Tramontano A. (2009). Protein function annotation by homology-based inference. *Genome Biol* 10(2): 1-8.
- Nimrod G, Schushan M, Steinberg DM & Ben-Tal N (2008). Detection of functionally important regions in "hypothetical proteins" of known structure. *Str* 16(12): 1755-1763.
- Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, & Tramontano A (2009). Protein function annotation by homology-based inference. *Genome Biol* 10(2): 1-8.
- Idrees S, Nadeem S, Kanwal S, Ehsan B, Yousaf A, Nadeem S, & Rajoka M I (2012). In silico sequence analysis, homology modeling and function annotation of Ocimum basilicum hypothetical protein G1CT28_OCIBA. *Int J Bioautomat* 16(2): 111-118.
- Lubec G, Afjehi-Sadat L, Yang JW & John JPP (2005). Searching for hypothetical proteins: theory and practice based upon original data and literature. *Progress in Neurobiol* 77(1-2): 90-127.
- Minion FC, Lefkowitz EJ, Madsen ML, Cleary BJ, Swartzell SM & Mahairas

- GG (2004). The genome sequence of *Mycoplasma hyopneumoniae* strain 232, the agent of swine mycoplasmosis. *J of Bacteriol* 186(21): 7123-7133.
8. Ahmad SU, Qadus A, Ahmad B, Khan JA, Shah ZW, Saeed A & Mahmood T (2020). In-silico investigation of EGFR network in kidney cancer: a drug discovery approach. *Pure and Appl Biol* 9(2): 1583-1595.
9. Wallach V, Williams KL & Boundy J (2014). Snakes of the world: a catalogue of living and extinct Species. CRC press.
10. Tan NH & Saifuddin MN (1989). Isolation and characterization of an unusual form of L-amino acid oxidase from King cobra (*Ophiophagus hannah*) venom. *Biochem Inter* 19(4): 937.
11. Guo XX, Zeng L, Lee WH, Zhang & Jin Y (2007). Isolation and cloning of a metalloproteinase from king cobra snake venom. *Toxicon* 49(7): 954-965.
12. He YY, Lee WH & Zhang Y (2004). Cloning and purification of α -neurotoxins from king cobra (*Ophiophagus hannah*). *Toxicon* 44(3): 295-303.
13. Li J, Zhang H, Liu J & Xu K. (2006). Novel genes encoding six kinds of three-finger toxins in *Ophiophagus hannah* (king cobra) and function characterization of two recombinant long-chain neurotoxins. *Biochem J* 398(2): 233-242.
14. Tan NH & Saifuddin MN (1990). Purification and characterization of two acidic phospholipase A2 enzymes from king cobra (*Ophiophagus hannah*) snake venom. *The Inter of Biochem* 22(5): 481-487.
15. Wang QY, Shu YY, Zhuang MX & Lin ZJ (2001). Cloning and Sequence Analysis of cDNAs Encoding Two Acidic PLA (2) from venom of *Ophiophagus hannah* (King Cobra), Guangxi Species. *Sheng wu hua xue yu Sheng wu wu li xue bao Acta Biochimica et Biophysica Sinica* 33(3): 340-344.
16. Pung YF, Wong PT, Kumar PP, Hodgson WC & Kini RM (2005). Ohanin, a novel protein from king cobra venom, induces hypolocomotion and hyperalgesia in mice. *J of Biol Chem* 280(13): 13137-13147.
17. He YY, Liu SB, Lee WH, Qian JQ & Zhang Y (2008). Isolation, expression and characterization of a novel dual serine protease inhibitor, OH-TCI, from king cobra venom. *Peptides* 29(10): 1692-1699.
18. Lee WH, Zhang Y, Wang WY, Xiong YL & Gao R (1995). Isolation and properties of a blood coagulation factor X activator from the venom of king cobra (*Ophiophagus hannah*). *Toxicon* 33(10): 1263-1276.
19. Chang HC, Tsai TS & Tsai IH (2013). Functional proteomic approach to discover geographic variations of king cobra venoms from Southeast Asia and China. *J of Proteom* 89: 141-153.
20. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD & Bairoch A. (2005). Protein identification and analysis tools on the ExPASy server. In *The Proteomics Protocols Handbook*, Humana press. pp. 571-607.
21. Lubec G, Afjehi-Sadat L, Yang JW & John JPP (2005). Searching for hypothetical proteins: theory and practice based upon original data and literature. *Prog in Neurobiol* 77(1-2): 90-127.
22. Geourjon C & Deleage G (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Bioinform* 11(6): 681-684.
23. Colovos C & Yeates TO (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* 2(9): 1511-1519.
24. Arnold K, Bordoli L, Kopp J & Schwede T (2006). The SWISS-MODEL workspace: a web-based

- environment for protein structure homology modelling. *Bioinform* 22(2): 195-201.
25. Lovell SC, Davis IW, Arendall WB, De Bakker PI, Word JM, Prisant MG & Richardson DC (2003). Structure validation by C α geometry: ϕ , ψ and C β deviation. *Proteins: Structure, Function, and Bioinform* 50(3): 437-450.
 26. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C & Heger A (2012). The Pfam protein families database. *Nuc Acids Res* 40(D1): D290-D301.
 27. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC & Ferrin TE (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *J of Comput Chem* 25(13): 1605-1612.
 28. Moriya Y, Itoh M, Okuda S, Yoshizawa AC & Kanehisa M (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nuc Acids Res* 35(suppl-2): W182-W185.
 29. Zdobnov EM & Apweiler R (2001). InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinform* 17(9): 847-848.
 30. Kulmanov M & Hoehndorf R (2020). DeepGOPlus: improved protein function prediction from sequence. *Bioinform* 36(2): 422-429.
 31. Zhang R, Ou HY & Zhang CT (2004). DEG: a database of essential genes. *Nucl Acids Res* 32(Suppl-1): D271-D272.
 32. Oany AR, Jyoti TP & Ahmad SAI (2014). An in silico approach for characterization of an aminoglycoside antibiotic-resistant methyltransferase protein from *Pyrococcus furiosus* (DSM 3638). *Bioinform and Biol Insights* 8: BBI-S14620.
 33. Benkert P, Biasini M & Schwede T (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinform* 27(3): 343-350.
 34. Oany AR, Ahmad SAI, Siddikey MAA, Hossain MU & Ferdoushi A (2014). Computational structure analysis and function prediction of an uncharacterized protein (I6U7D0) of *Pyrococcus furiosus* Com1. *Austin J Comput Biol Bioinform*, 1(2): 5.
 35. Gorodkin J, Lund O, Andersen CA & Brunak S (1999). Using sequence motifs for enhanced neural network prediction of protein distance constraints. In *ISMB* 99: 95-105.
 36. Fa R, Cozzetto D, Wan C & Jones DT (2018). Predicting human protein function with multi-task deep neural networks. *Plos One* 13(6): e0198216.
 37. Bowie JU, Luthy R & Eisenberg D (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Sci* 253(5016): 164-170.
 38. Geertz-Hansen HM, Blom N, Feist AM, Brunak S & Petersen TN (2014). Cofactory: sequence-based prediction of cofactor specificity of Rossmann folds. *Proteins* 82(9):1819-1828.