IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 7 Issue 7, July 2020
ISSN (Online) 2348 – 7968 | Impact Factor (2020) – 6.72
www.ijiset.com

In-Silico Analysis of Silk Serpin-2 Protein from *Arachnocampa richardsae.*

Amit Chougale[1], Shivani Mantri[1], Sneha Kagale[1], Shruti Vedante[1]

[1] Department of Biotechnology, Kolhapur Institute of Technology's College of Engineering, Kolhapur, Maharashtra, 416234, India.

Abstract

Silk Serpin-2 protein originates in Arachnocampa richardsae species which exhibits the presence of unexplored nucleotide sequence. Whole-genome annotations for Arachnocampa richardsae have not been accomplished and there are no known functions of Silk Serpin-2 until now. Therefore in this paper, we studied specifications like phylogenetic analysis, conservative residue, and conserved domain, secondary and 3D structure, physicochemical properties, and functional importance of Silk Serpin-2. The methodology used for the study involve software tools like MEGA X, Bio Edit, PSIPRED and Swissprot for the overall insilico study of the Silk Serpin-2 protein. Reliable and effective methods for 3D structure modeling were used and validated using the Ramachandran Plot. Results showed that the presence of silk Serpin-2 in the cell is responsible for many processes like regulation of metabolic process (88.08%), catabolic processes (69%), and catalytic activity (85.7%). The tools were used to discover the physicochemical properties from number of amino acids to the instability index. The secondary structure and 3D model of the protein were predicted which represented valid amino acid percentage in the allowed region of Ramachandran plot. The study may prove to be a reference for revealing further applications of the protein Silk Serpin-2 and the implementation of this technique will aid to expose other unknown proteins.

Keywords: *Arachnocampa richardsae; Silk Serpin-2; Phylogenetics; Functional annotations; Homology modeling; Bioinformatics tools; 3D model.*

1. Introduction

Glowworms found in Australia and New Zealand belong to the family Keroplatidae, subfamily Arachnocampinae, genus Arachnocampa are the bioluminescent larvae of flies (Diptera). Their habitat includes a dark and humid environment associated with rainforests and caves. Diptera larvae construct a snare composed of a horizontal mucous tube hung from the substrate by bracing threads. The genus Arachnocampa encompasses nine species categorized into three subgenera. A. luminosa is endemic to New Zealand. Eight species belong to Australia in which five species are newly described based on their morphological features and the remaining three Australian species are A. flava, A. tasmaniensis, and A. richardsae [1]. Arachnocampa richardsae produces a protein named Silk Serpin-2 which is a member of the serpin family.

Serpin-like protease inhibitors have been identified in several species like plants, animals, viruses, archaea, and bacteria and over 1,500 members of this family have been identified till date [2]. Serpins (serine protease inhibitors) are the largest and most broadly distributed superfamily of protease inhibitors. Protease Inhibitory Serpins are well known to function in the processes as diverse as DNA binding, dorsal-ventral axis formation, immune-regulation in Drosophila and other insect's embryo and control of apoptosis. Since Serpins were known to be inhibitors of serine proteases, so their structure would be existing a loop to interact with the protease active site cleft [3]. Further, there are several proteins in the Serpin family whose function(s) is not recognized.

There is a growing need for the automatic annotation of proteins of unknown function, termed 'hypothetical proteins'. In the genome of numerous life systems number of hypothetical proteins have been created. Structural- genomic centers usually lack the resources to involve in detailed functional characterization of each of the resolved structure of many hypothetical proteins [4]. Whole-genome annotations for Arachnocampa richardsae have not been accomplished and there are no known functions of Silk Serpin-2 until now.

Approaches through bioinformatics include algorithms and databases to evaluate the hypothetical proteins, using Bioinformatics is a decent alternative for laboratory explorations. As these algorithms and databases support experimental results analysis, they can be a convincing aim to complete functionality and structural annotation of hypothetical proteins [5]. The accessible Bioinformatics tools and servers have provided the annotation of the genome for revealing the function of a particular gene (protein), to determine the occurrence of the enzymatic conserved domains in the sequences which may contribute in the classifying protein into specific family and three-dimensional structures for protein sequences virtually [6].

National Center Biotechnology Information's (NCBI) Conserved Domain Database (CDD) imparts domain family models from a variety of external sources. CDD is a resource for the annotation of protein sequences with the location of conserved domain footprints [7]. In life science research, three-dimensional structures of proteins provides valuable perceptions of their function at molecular level and inform a broad spectrum of applications [8]. Instability indices, aliphatic indices and grand averages of hydropathicity (GRAVY) of these proteins were computed using the ProtParam Proteomics tool available at EXPASY server [9]. The PSIPRED Protein Analysis is a web service that facilitates a diverse set of protein prediction and annotation tools that focuses mainly on structural annotations of proteins [10].
In the current computational study, we used fifty sequences of serpin family for analysis of silk serpin 2 (S4TJ40_9DIPT: UniProt Id) by various bioinformatics tools. Phylogenetic and evolutionary studies were performed, along with the structures, functions, subcellular localization, and 3D structures of the hypothetical protein Silk Serpin-2 from Arachnocampa richardsiae were obtained and represented.

2. Materials and Methods

2.1 Sequence Retrieval:

The sequences of proteins were collected from the NCBI Database using Silk Serpin-2 (JQ915215: NCBI Database) as a query sequence in FASTA format.

2.2 Phylogenetic and Evolutionary Analysis with Sequence Alignment:

The Serpin protein sequences from different organisms were aligned using the MUSCLE alignment tool and used for a phylogenetic tree in the form of a cladogram constructed using MEGA X [11]. The percentage of replicate trees in which the associated sequences cluster together in the bootstrap test (500 replicates) were calculated, and branches with, 25% bootstrap cutoff were collapsed for visualization in MEGA X [12].

2.3 Revealing Mutative/Conservative Residues Of Protein:

The retrieved Silk Serpin-2 sequences were aligned and then the partial and imprecise regions at both ends of the sequences were trimmed using BioEdit package v7.1. The variability metrics for these sequences were then characterized based on data content measured as Shannon's entropy (Hx) plot by means of BioEdit software [13].

2.4 Annotation of The Conserved Domain Of Protein:

Conserved domain(s) and superfamily of the Silk Serpin-2 was assigned by searching against the NCBI's Conserved Domains Database v3.15 (CDD) [14]. In CDD, the E-value threshold was set on 10-5 (1e-05), where the values more than 10-5 are considered as putative false positives. CDD is a protein annotation online website that comprises of collection of NCBI-curated domains and also domain models that imported from several external databases such as Pfam and SMART [15].

2.5 Prediction of Protein Secondary Structure:

The secondary structure of the Silk Serpin-2 was predicted using the neural network-based prediction online server, PSSpred (https://zhanglab.ccmb.med.umich.edu/PSSpred). In addition to that, the PSIPRED [10] servers were also exploited to confirm the results achieved from PSSpred. For the prediction of primary structure the ProtParam network analyzer was used. Three states of the secondary structure {α-helix (H), β-strands (S), and coil (C)} were also predicted.

2.6 Physicochemical Analysis of Protein:

Expasy's ProtParam was utilized for the assessment of different physicochemical and structural characteristics of Silk Serpin-2. The parameters like peptide length, molecular weight, theoretical isoelectric point (pI), half-life period, instability index, aliphatic index, extinction coefficient, and grand average of hydropathicity (GRAVY) [16].

2.7 Subcellular Localization Prediction:

The subcellular localization of Silk Serpin-2 was predicted by WOLF PSORT (https://wolfpsort.hgc.jp/) [17]. Results also cross-checked through subcellular localization predictions acquired from PSIPred server and Predict-Protein servers.

2.8 Predict Function of Protein Function:

For predicting the function of protein online web servers like ProteinPredict (https://www.predictprotein.org/), ProFun 2.2, and ProtFunc [18] were used. It provided information about other GO terms. It predicted the biological processes and molecular processes.

2.9 Prediction of 3D Structure Modelling Of Protein:

The Silk Serpin-2 protein sequence was compared with protein sequences in the PDB website using the NCBI blastp (protein-protein BLAST) algorithm. The maximum score in this alignment tool has a similarity of about 28%, therefore for high accurate protein modeling SWISS-MODEL online web-server [19, 20], and for visualizing of the modeled structure Rasmol web-server was used.

2.10 Validation of the Modelled Structure:

For validation of the 3D model of Silk Serpin-2, uploaded the 3D model was on Rampage (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) and was assessed the Ramachandran plot [21]. The 3D structure of Silk Serpin-2 was modeled.

3. RESULTS AND DISCUSSION

3.1 Phylogenetic and evolutionary Analysis:

The maximum likelihood tree was observed on the adaptive evolutionary process of different Serpin proteins. The results were obtained from Mega X software totally diverse from the query sequence. Fig.1 represents the phylogenetic tree for 50 different protein sequences from different species were collected from Blast server. The number on the node of the tree represents the reliability of the tree (greater than 70). It shows that query protein is highly diverse from other Serpin proteins. It assures that Silk Serpin-2 is a member of the Serpin superfamily.
In different species, the structural domains are conserved in the process of species evolution. For Silk Serpin-2 there is no conservative region so, it is different from other types of Serpins. During the evolution, the ancestor of Serpin family is common but after speciation events, huge diversity can be observed.

3.2 Revealing mutative/conservative residues:

The entropy plot H(x) gives the variations and conservative residues from the sequences. The range of entropy values of residues of the proteins is between 0 to 3. The graph showed that there are frequent peaks that indicate the variations in sequences but the length of the query sequence is smaller than comparative sequences. Therefor rest sequence pattern of the graph can be neglected. As no longer a straight regime, means there is no conservative residue. The entropy plot is given in Fig.2.

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 7 Issue 7, July 2020

ISSN (Online) 2348 – 7968 | Impact Factor (2020) – 6.72

www.ijiset.com

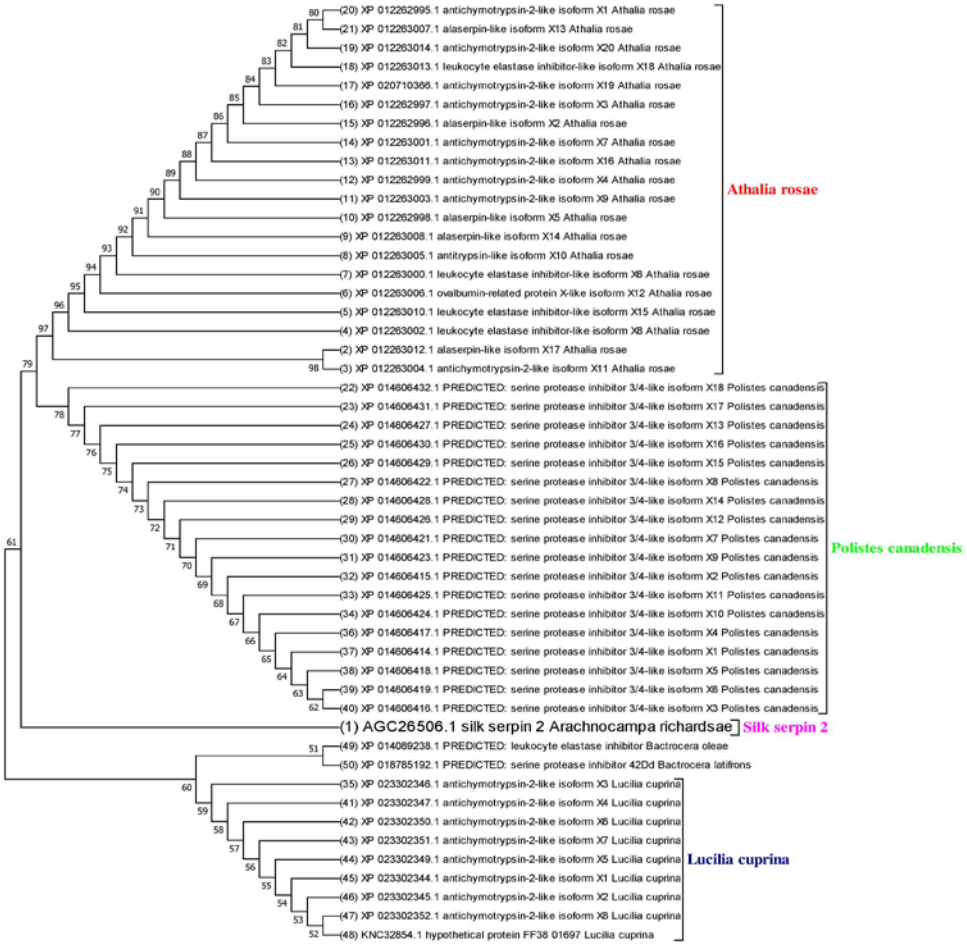

Fig. 1: Phylogenetic tree by bootstrapping method (500 times) is obtained from MEGA X software. It consists of clades of closely related members of the same protein family. The cluster of sequences locating from singulas species.

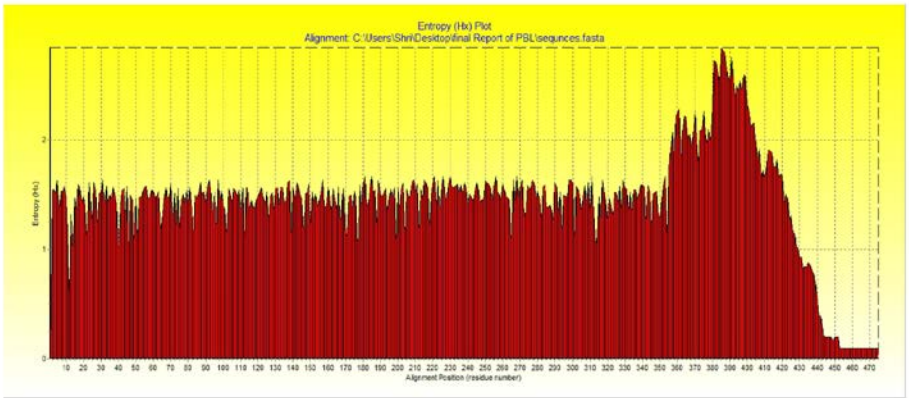

Fig. 2: Entropy plot H(x) (range: 0 to 3) which is obtained from BioEdit package v7.1. The length of Silk Serpin-2 is 275 amino acids. In that region, the peaks represent the variations after aligning with other protein sequences. The flat region shows conserved residues.

3.3 The conserved domain annotation of Silk Serpin-2:

NCBI-CDD's outputs are based on the RPS-BLAST algorithm to compare protein sequence against position-specific score matrices (PSSMs), resulting in faster identification of conserved domains in protein sequences, which include three different intervals and E-values in the sequence. Based on these results, the Silk Serpin-2 is counted as a member of the Serpin superfamily.

3.4 Secondary structure of Silk Serpin-2:

The secondary structure of the protein predicted by PSIPRED has alpha helix, the most predominant (44.73%), followed by a random coil (40%) and extended strand (15.27%). Similar results stayed found from Predict-Protein and Polyview servers. Different servers showed the presence of protein structure in the cell. A certain part of the structure in which 57.82% residues are buried, 34.18% residues are exposed to the solvent while 8% residues are intermediate contained by the cell. The representative secondary structure of Silk Serpin-2 obtained from the PSIPRED server is shown in Fig.3.

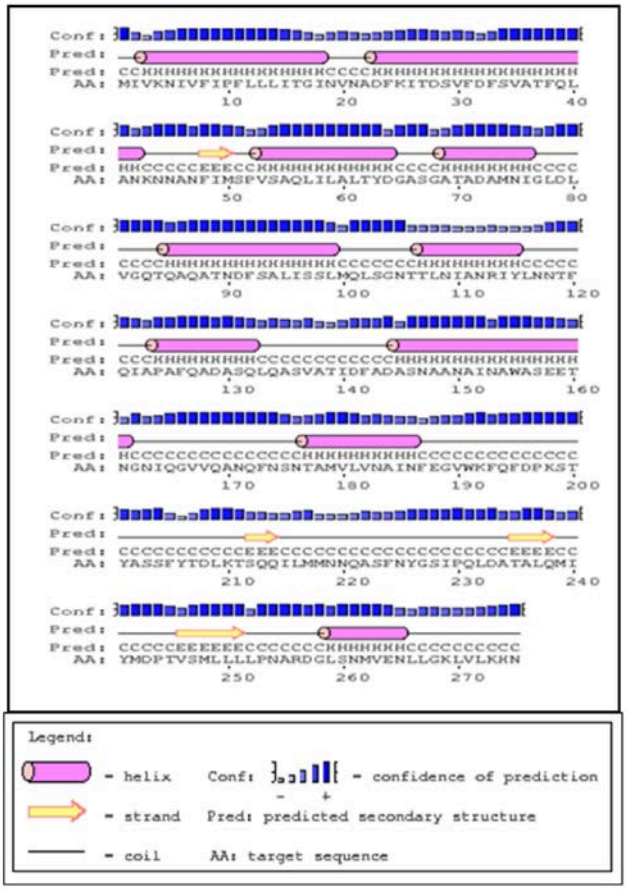

Fig. 3: Secondary structure of Silk Serpin-2; Pink Cylinder: Helix; Yellow Arrow: Strands; Line: Coils; Blue columns: confidence level proportional to the height of column predicted by software.

3.5 Physiochemical Characteristics of Silk Serpin-2:

The ExPASy's ProtParam server provided the result of theoretical physiochemical characteristics of the amino acid sequence of protein Silk Serpin-2 as shown in Table 1 and Table 2. The values got from sever shows vital relevancy in data which aid further studies.

Table 1: Physicochemical Properties of Silk Serpin-2.

Sr. no.	***Items***	***Values***
1.	Number of amino acids	275
2.	Molecular weight	29722.76
3.	Theoretical pI	4.29
4.	Formula	$C_{1325}H_{2075}N_{343}O_{410}S_{11}$
5.	Total number of atoms	4164
6.	Half-life period	30 hours
7.	Instability index	27.55
8.	Aliphatic index	99.09
9.	Grand average of hydropathicity (GRAVY)	0.212
10.	Extinction coefficient	19940
11.	Abs 0.1% (=1 g/l) at 280 nm	0.671
12.	total number of negatively charged residues (Asp + Glu)	19
13.	total number of positively charged residues (Arg + Lys)	10

Table 2: Amino acid composition of Silk Serpin-2.

Amino Acid	***Composition %***	***Amino Acid***	***Composition %***
Ala (A) 35	12.7%	Gln (Q) 19	6.9%
Arg (R) 2	0.7%	Glu (E) 4	1.5%
Asn (N) 31	11.3%	Gly (G) 12	4.4%
Asp (D) 15	5.5%	His (H) 1	0.4%
Cys (C) 0	0.0%	Ile (I) 20	7.3%
Leu (L) 29	10.5%	Lys (K) 8	2.9%
Met (M) 11	4.0%	Trp (W) 2	0.7%
Phe (F) 17	6.2%	Tyr (Y) 6	2.2%
Pro (P) 7	2.5%	Val (V) 16	5.8%
Ser (S) 22	8.0%	Pyl (O) 0	0.0%
Thr (T) 18	6.5%	Sec (U) 0	0.0%

3.6 Subcellular localization of Silk Serpin-2:

Protein subcellular localization predictions contribute to computational probability of survival of protein inside the cell. Prediction of subcellular localization of unidentified proteins can provide information about their cellular functions. The reliability represented in table give authentic presence of protein using support vector machine algorithm. The subcellular localization of the query protein was anticipated to be a nuclear protein, analyzed by WOLF PSORT, and confirmed by CELLO Prediction and Predict Protein severs. Table 3 classified into two parts, one of them represents the analysis report for protein and other predicted locations all over the cell for same.

Table 3: Subcellular localization of Silk serpin 2 (CELLO Prediction).

SUPPORT VECTOR MACHINE	*LOCALIZATION*	*RELIABILITY*
Analysis Report		
Amino Acid Comp.	Extracellular	0.655
N-peptide Comp.	Extracellular	0.694
Partitioned seq. Comp.	Plasma Membrane	0.686
Physico-chemical Comp.	Plasma Membrane	0.847
Neighboring seq. Comp.	Extracellular	0.502
CELLO Prediction		
-	PlasmaMembrane	2.120 *
-	Extracellular	2.114 *
-	Nuclear	0.183
-	Cytoplasmic	0.132
-	Chloroplast	0.131
-	Mitochondrial	0.079
-	Vacuole	0.073
-	Golgi	0.065
-	Lysosomal	0.034
-	Peroxisomal	0.030
-	Cytoskeletal	0.022
-	ER	0.017

3.7 Prediction of a function of Silk Serpin-2:

PSIPRED was used to perform biological processes like regulation of metabolic process (probability is 80.8%) and catabolic process (probability 69%). It gave value for molecular functions like catalytic activity (probability 85.7%) and peptidase activity (probability 73.1%). The reliability score of predicted functions was very low. The results obtained were confirmed using ProteinPredict and ProtFunc web tools.

3.8 Prediction of 3D Structure:

Homology modeling predicts the 3D structure of an assumed protein sequence which builds principally with respect to its alignments to one or more proteins of known structure. SignalP server conformed that signal peptide as amino acids 1-22, these amino acids would be cleaved in the ER and mature secreted protein would start at amino acid 23. SWISS-MODEL gives the 3D structure of the protein and allows us to assist the structure according to the Ramachandran plot. The server consequently performed BLASTP search for each protein sequence to identify templates for homology modeling. BLASTP results show that the putative serpin sequence is 38% identical to alaserpin-like isoform X17[Athalia rosae] but it is 399 amino acid long sequence. Therefore, highest template identity was 28.40%. Fig.4 represents the 3D structure of Silk Serpin-2.

3.9 Validation of the 3d model of Silk Serpin-2:

Ramachandran plot of the model revealed that about 98.4% of the Silk Serpin-2 residues are located in the allowed region of the plot (Fig.4). The number of residues in the favored region is expected to be 98%. So, the structure obtained from 3D

modeling is considered as a valid model. Fig.4 gives the quality estimation result for the predicted protein structure. Few negligible bad bonds and angles are also presente in the structure.

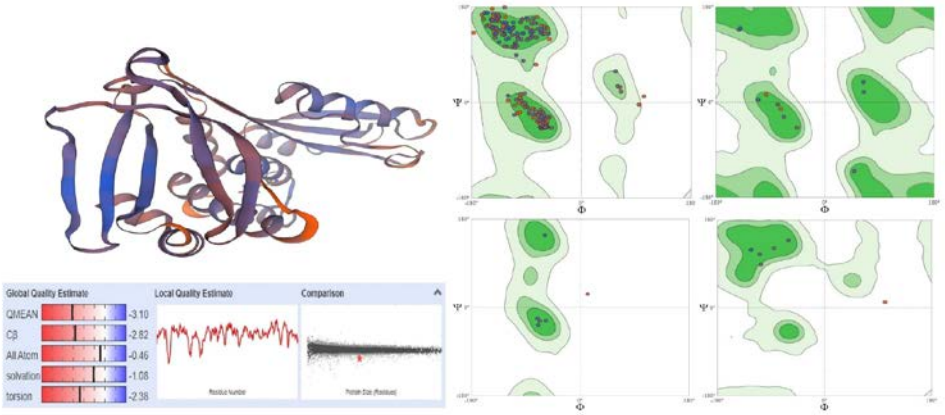

Fig. 4: 3D structure of Silk Serpin-2 obtained from the SWISS-MODEL online web-server. The four graphs of Ramachandran plot analyzes the more reliable region for the 3D structure of Silk Serpin-2. Quality estimation of the 3D structure of the protein. In which the graph describes the local quality estimation. In global quality estimation, the red side gives the errors in the structures while the blue side shows more reliability of protein structure.

4. Conclusions

Silk Serpin-2 protein was thoroughly studied which shows great diversity from other Serpins when it's Phylogenetic study was performed. So, there were low conservative residues from Silk Serpin-2 as compared to all the different Serpins from different species. The Silk Serpin-2 showed that it belongs to Serpin superfamily. Furthermore, we predicted all the parameters of Silk Serpin-2 of which PSIPRED provided more precise 2D structure. The chemical and physical properties were appropriately given through. The location of protein sequence mostly found in plasma membrane also the protein is predicted to be localized in extracellular space in the secretary form.
It involves processes like regulation of metabolic process (Reliability score: 0.8808), catabolic processes (Reliability score: 0.69), catalytic activity (Reliability score: 0.857). Silk Serpin-2 may act as a surface binding protein which helps in the cell signaling process. The 3D model of the protein was built using alaserpin-like isoform X17[Athalia rosae] in SWISS-MODEL server which shows 38% identity and confirmed by the Ramachandran plot method with 98.4% permissible region. The study completed with certain online and offline Bioinformatics tool for discovery of novel proteins similarly this study may prove to be a reference for studying further applications of Silk Serpin-2 Protein.

Acknowledgments

We would like to thank the Biotechnology department from Kolhapur Institute of Technology College of Engineering, Kolhapur for the technical support.

References

[1] Baker, C. H., Graham, G. C., Scott, K. D., Cameron, S. L., Yeates, D. K., & Merritt, D. J. Distribution and phylogenetic relationships of Australian glow-worms Arachnocampa (Diptera, Keroplatidae). Molecular Phylogenetics and Evolution, 48(2), 506-514(2008).
[2] Law, R. H., Zhang, Q., McGowan, S., Buckle, A. M., Silverman, G. A., Wong, W. & Whisstock, J. C. An overview of the serpin superfamily. Genome biology, 7(5), 1-11 (2006).
[3] Huntington, J. A. Serpin structure, function and dysfunction. Journal of thrombosis and haemostasis, 9, 26-34(2011).

[4] Nimrod, G., Schushan, M., Steinberg, D. M., & Ben-Tal, N. Detection of functionally important regions in "hypothetical proteins" of known structure. Structure, 16(12), 1755-1763(2008).
[5] Huntington, J. A. Serpin structure, function and dysfunction. Journal of thrombosis and haemostasis, 9, 26-34(2011).
[6] Thakare, H. S., Meshram, D. B., Jangam, C. M., Labhasetwar, P., Roychoudhary, K., & Ingle, A. B. Comparative genomics for understanding the structure, function and sub-cellular localization of hypothetical proteins in Thermanerovibrio acidaminovorans DSM 6589 (tai). Computational biology and chemistry, 61, 226-228(2016).
[7] Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., ... & Gwadz, M. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic acids research, 39(suppl_1), D225-D229(2010).
[8] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., ... & Lepore, R. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic acids research, 46(W1), W296-W303(2018).
[9] Dutta, S. K., Bhattacharya, T., & Tripathi, A. Chikungunya virus: genomic microevolution in Eastern India and its in-silico epitope prediction. 3 Biotech, 8(7), 318(2018).
[10] Buchan, D. W., & Jones, D. T. The PSIPRED protein analysis workbench: 20 years on. Nucleic acids research, 47(W1), W402-W407(2019).
[11] Kumar, A., Bhandari, A., Sarde, S. J., & Goswami, C. Sequence, phylogenetic and variant analyses of antithrombin III. Biochemical and biophysical research communications, 440(4), 714-724(2013).
[12] Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular biology and evolution, 28(10), 2731-2739(2011).
[13] Elkins, K. M. Chapter 15—Analysis of Deoxyribonucleic Acid (DNA) Sequence Data Using BioEdit. Forensic DNA Biology; Academic Press: San Diego, CA, USA, 129-132(2013).
[14] Pourseif, M. M., Moghaddam, G., Naghili, B., Saeedi, N., Parvizpour, S., Nematollahi, A., & Omidi, Y. A novel in silico minigene vaccine based on CD4+ T-helper and B-cell epitopes of EG95 isolates for vaccination against cystic echinococcosis. Computational biology and chemistry, 72, 150-163(2018).
[15] Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., ... & Sonnhammer, E. L. Pfam: the protein families database. Nucleic acids research, 42(D1), D222-D230(2014).
[16] Gao, X. Y., Zhang, Y. F., Zheng, W. L., & Lu, B. L. Evaluating driving fatigue detection algorithms using eye tracking glasses. In 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER) (pp. 767-770). IEEE(2015, April).
[17] Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. WoLF PSORT: protein localization predictor. Nucleic acids research, 35(suppl_2), W585-W587(2007).
[18] Sunil, L., & Vasu, P. In silico designing of therapeutic protein enriched with branched-chain amino acids for the dietary treatment of chronic liver disease. Journal of Molecular Graphics and Modelling, 76, 192-204(2017).
[19] Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., ... & Tramontano, A. Protein function annotation by homology-based inference. Genome biology, 10(2), 1-8(2009).
[20] Jahangiri, A., Rasooli, I., Owlia, P., Fooladi, A. A. I., & Salimian, J. An integrative in silico approach to the structure of Omp33-36 in Acinetobacter baumannii. Computational biology and chemistry, 72, 77-86(2018).
[21] Tamboli, A. S., Waghmare, P. R., Khandare, R. V., & Govindwar, S. P. Comparative analyses of enzymatic activity, structural study and docking of fungal cellulases. Gene Reports, 9, 54-60(2017).