chemistry proceedings

MDPI

Proceeding Paper

Bioinformatics Approaches for Molecular Characterization of CT670 Hypothetical Protein of *Chlamydia pneumoniae* †

Abu Saim Mohammad Saikat [1,*], Tazin Afrose [2], Umme Saoda [3], Kazi Nur Uddin [4], Mir Monir Hossain [5] and Md. Lutful Kabir [6,*]

[1] Department of Biochemistry and Molecular Biology, Life Science Faculty, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj 8100, Bangladesh
[2] Pioneer Dental College and Hospital, Dhaka 1229, Bangladesh; dr.afrosetazin26@gmail.com
[3] Department of Virology, Dhaka Medical College, Dhaka 1000, Bangladesh; saodaumme@gmail.com
[4] Department of Mathematical Science, Kent State University, Kent, OH 44240, USA; kuddin@kent.edu
[5] Department of Chemistry, Cleveland State University, Cleveland, OH 44115, USA; m.m.hossain39@vikes.csuohio.edu
[6] Department of Chemistry and Biochemistry, Kent State University, Kent, OH 44240, USA
* Correspondence: asmsaikat.bmb@gmail.com (A.S.M.S.); mkabir3@kent.edu (M.L.K.)
† Presented at the 28th International Electronic Conference on Synthetic Organic Chemistry (ECSOC 2024), 15-30 November 2024; Available online: https://sciforum.net/event/ecsoc-28.

Abstract: Researchers have linked *Chlamydia pneumoniae* (*C. pneumoniae*), a type of bacteria that cannot survive outside of cells and is resistant to gram staining, to many autoimmune diseases. People hypothesized that *C. pneumoniae* had a harmful function due to its tendency to inhabit human endothelium and epithelial tissue. This study implemented multiple bioinformatics tools and databases to understand the possible function of the CT670 hypothetical protein of *C. pneumoniae*. The physicochemical parameters showed the protein's half-life in different media. These parameters also displayed the protein's theoretical isoelectric point, aliphatic index, GRAVY value, extinction coefficient, instability index, as well as the amino acids and atoms that comprise it. Amino acid composition measured the percentage of amino acids present in the selected protein, with glutamate demonstrated as the greatest proportion. Moreover, hydrogen was the most abundant ratio in terms of the atomic composition of the protein, followed by carbon, oxygen, nitrogen, and sulfur. The PPI networks reveal its potential primary and secondary interactions with other proteins. We modeled and assessed the secondary and tertiary structures to understand the nature of the selected protein. Computational functional analysis predicted that the protein would be a chaperone effector. By designing and developing drugs and vaccines, we can use this protein as a target for further analysis to combat diseases caused by *C. pneumoniae*.

Keywords: *Chlamydia pneumoniae*; CT670; protein-protein interactions; Ramachandran plot; computational chemistry

Citation: Saikat, A.S.M.; Afrose, T.; Saoda, U.; Uddin, K.N.; Hossain, M.M.; Kabir, M.L. Bioinformatics Approaches for Molecular Characterization of CT670 Hypothetical Protein of *Chlamydia pneumoniae*. *Chem. Proc.* **2024**, *6*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 15 November 2024

1. Introduction

C. pneumoniae, a kind of bacteria that cannot survive outside of cells and has a negative reaction to the Gram stain, is a common respiratory infection. It is often responsible for respiratory disorders in humans, such as pneumonia [1]. Despite the complete sequencing of the *C. pneumoniae* genome, there is still a lack of knowledge about the processes of acute infection, target cell activation, and the discovery of possible chlamydial virulence factors. Interestingly, several groups of patients with atherosclerosis have shown that the existing antibiotic treatments for acute chlamydial infection are ineffective in achieving positive clinical outcomes [2].

However, in addition to respiratory infections, *C. pneumoniae* also plays a role in the development of various inflammatory conditions, including asthma, COPD, lung cancer,

neurological disorders like multiple sclerosis, Alzheimer's disease, and schizophrenia, as well as arthritis and atherosclerosis. Therefore, the healthcare professional must possess the ability to immediately identify, assess, and manage this disease in order to prevent any associated problems [3–5].

2. Methods

2.1. Sequence Retrieval

The amino acid sequence of the CT670 protein of *C. pneumoniae* was collected from the NCBI protein database with the accession number BAA88656 (version number: BAA88656.1) [6].

2.2. Physicochemical Properties

The ProtParam web-based tool was used for the determination of the physicochemical properties of the CT670 protein with default parameters [7].

2.3. Protein-Protein Interaction (PPI) and Functional Analysis

The PPI was determined by utilizing the STRING database (v.12) [8] and visualized by Cytoscape software (v.3.10.2) [9]. Functional analysis was performed using the CD-Search tool of the NCBI [10].

2.4. Secondary and Tertiary Structural Assessment

SOPMA web-based software was used with default options (width for output: 70, similarity threshold: 8, and window width: 17) to determine the secondary structural parameters [11]. The 3D structure was predicted using the Swiss-Model server [12]. Besides, the Swiss-Model's assessment tool and the ProSA-Web were implemented to verify the modelled structure of the protein [12–14].

3. Results and Discussion

3.1. Protein Sequence Retrieval and Physicochemical Properties Determination

Protein sequence analysis is a fundamental and essential approach for evaluating biological information. Given the significance of proteins, protein sequence benchmarks play a particularly vital role [15,16]. The obtained sequence is consisted of 168 amino acids while retrieved in FASTA format from the NCBI database (GenBank accession: BAA88656.1). It located to the locus BAA88656 of *C. pneumoniae* [6].

The chemical composition of the constituent amino acids determines the physicochemical characteristics of proteins [15]. The ProtParam anticipated the protein's molecular weight 19,882.89 Dalton and the isoelectric point (pI) of 8.84. There was an increased amount of protein solubility and electrophoretic separation, as measured by the isoelectric point (pI), which is greater than 7.0 [17,18]. Moreover, the amino acid composition (Figure 1a) and the atomic composition (Figure 1b) also measured. The negatively charged restudies including Asp and Glu were 38, while the positively charged (Arg and Lys) were 41. One way to quantify the degree to which a protein absorbs light at a given wavelength is by looking at its extinction coefficient, which is another name for its molar absorption or absorption coefficient. For every given protein molecule, there is an individual extinction coefficient [19]. The extinction coefficient of the selected protein was anticipated as 9970 M^{-1} cm^{-1}. Additionally, the estimated half-life were closely 30 h (mammalian reticulocytes, in-vitro), >20 h (yeast, in-vivo), and >10 h (*Escherichia coli*, in-vivo). In a laboratory setting, one can evaluate proteins for their potential stability using the instability index [20]. The predicted instability index of the targeted protein was 62.24. The aliphatic index refers to the percentage of a protein's total volume that its aliphatic side chains account for [21,22]. Computational methods measured the aliphatic index at 83.57. The GRAVY

value is one way to quantify the overall hydrophobicity of a protein or peptide [23,24]. The GRAVY value was measured at −1.151 of the selected hypothetical protein.

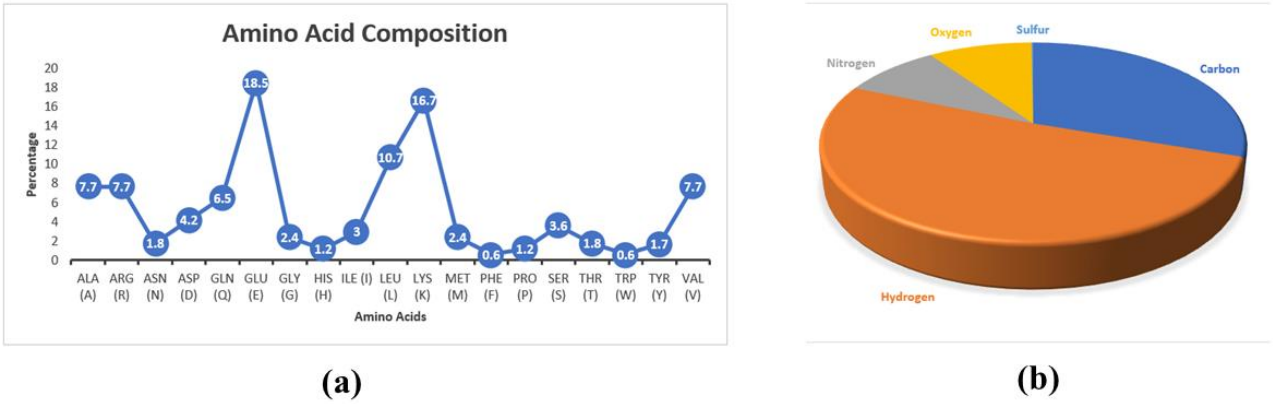

Figure 1. (**a**) Amino acid composition of the selected protein. (**b**) Atomic composition of the protein. Hydrogen (n = 1460) is the most abundant chemical element followed by carbon (n = 865), oxygen (n = 271), nitrogen (n = 254), and sulfur (n =4).

3.2. Protein-Protein Interaction (PPI) and Functional Analysis

Electrostatic forces, hydrogen bonds, and hydrophobic influence are some of the interactions that make PPIs possible. PPIs are specific physical contacts between two or more protein molecules, while interactions influence metabolic processes [25,26]. While performing PPI network (Figure 2), the STRING program demonstrated 21 nodes, 128 edges, average node degree of 12.2, median clustering coefficient value as of 0.777, and the enrichment *p*-value was less than 1.0×10^{-16}. Moreover, the CD-Search tool predicted the protein was a YscO-like protein. Some proteins, including Chlamydia trachomatis's CT670, have unknown cellular roles despite having genes in a type III secretion gene cluster. The structures and/or functionalities of these proteins are yet unknown; however, CT670 has many similarities with the Yersinia pestis YscO protein, such as the adjacent genes, size, charge, and secondary structure. CT670 interacts with CT671, which is a potential YscP homolog. It is possible that CT670 and CT671 function together as a chaperone-effector pair [27].

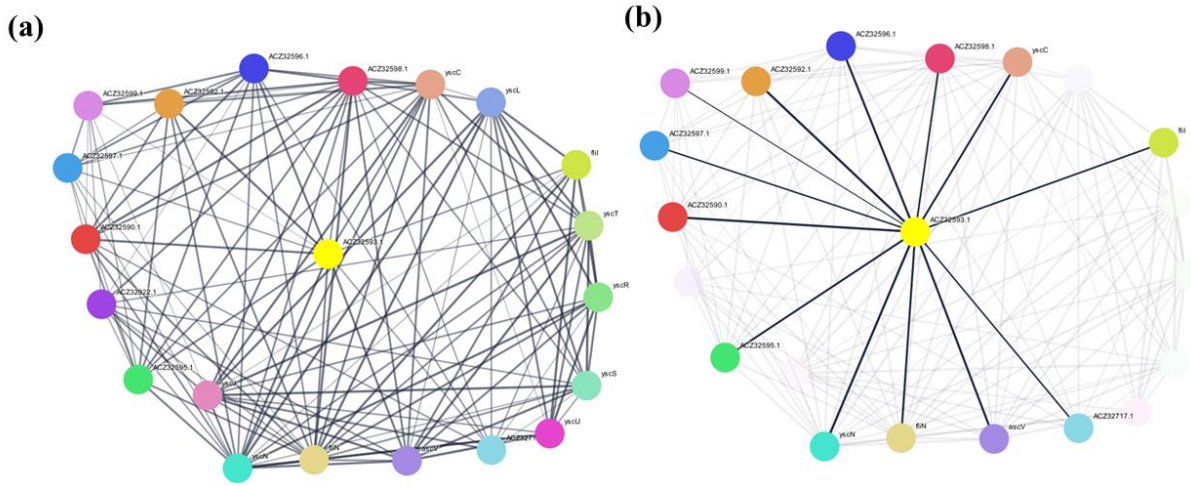

Figure 2. (**a**) The PPIs network of the selected hypothetical protein. The program selected ACZ32593.1 as an input to generate the network. It interacted with multiple proteins in both primary and secondary interactions, for example with ACZ32592.1, yscN, ACZ32590.1, ACZ32596.1,

ascV, ACZ32595.1, fliN, yscC, fliI, and ACZ32598.1. (**b**) The first-neighbor-interactions of the selected protein.

3.3. Secondary and Tertiary Structural Assessment

The sequence of amino acids is the main building block. The term 'secondary structure' refers to the local interactions between parts of a polypeptide chain. It includes both α-helix and β-pleated sheet structures (Figure 3). Interactions between R groups are the primary drivers of tertiary structure, which is the overall three-dimensional folding [28,29]. The 3D structure was predicted based on the most suitable template following GMQE (0.95) and identity (84.43%) compared to the selected hypothetical protein. Moreover, the predicted structure was assessed by Z-score and Ramachandran plots (Figure 3).

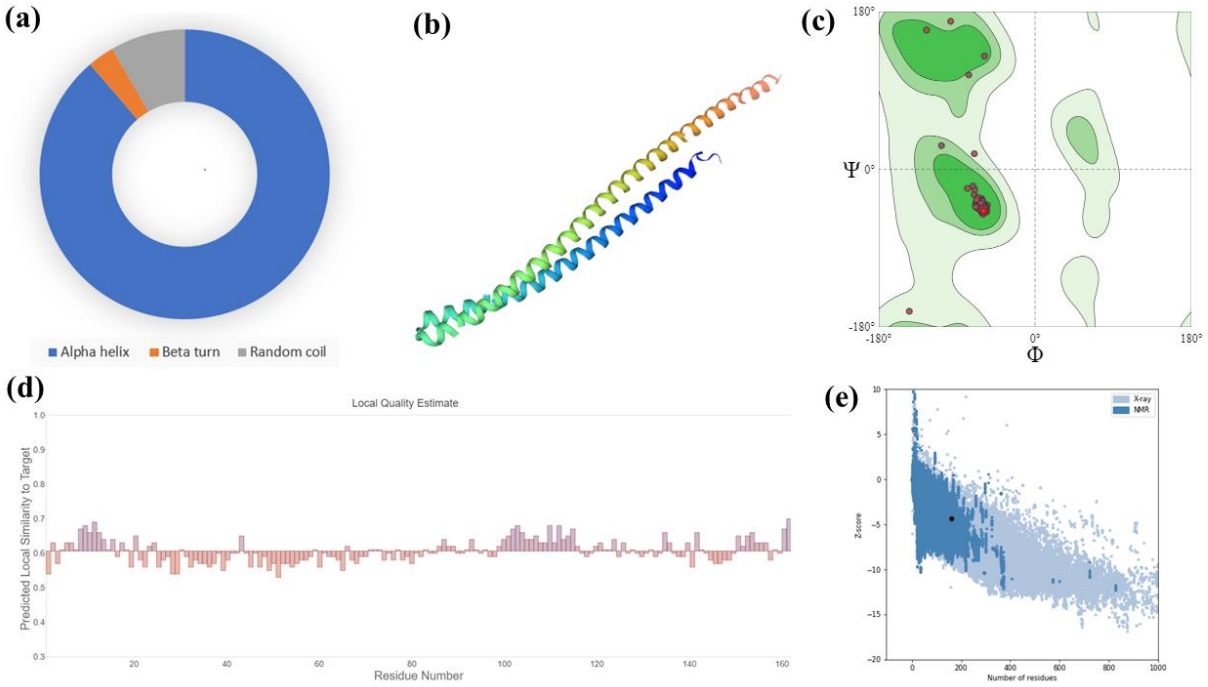

Figure 3. (**a**) Secondary structural elements revealed that alpha helix (n = 149, 88.69%) was the most abundant following by random coil (n = 14, 8.33%), and beta turn (n = 5, 2.98%). (**b**) The 3D structure of the selected hypothetical protein. (**c**) Structural assessment by the Swiss-Model program. Ramachandran plots measured 98.75% amino acids were in the most favored regions and there were no Ramachandran outliers. (**d**) The 'Local Quality Estimate' values of the selected protein with other parameters including QMEAN (3.13), Cβ (5.75), solvation (4.40), and torsion (0.09). (**e**) Z-score obtained from the ProSA-web demonstrated as −4.3.

4. Conclusions

Researchers have linked *C. pneumoniae*, which cannot survive outside of cells and is resistant to gram staining. Researchers postulated that *C. pneumoniae*'s propensity to reside in human endothelium and epithelial tissue played a detrimental role. This study utilized various bioinformatics methods and databases to elucidate the potential functionality of the CT670 putative protein of *C. pneumoniae*. The physicochemical properties indicated the protein's half-life in various mediums. These metrics also indicated the protein's theoretical isoelectric point, aliphatic index, GRAVY value, extinction coefficient, instability index, as well as the amino acids and atoms that make up the protein. The amino acid composition quantified the relative abundance of different amino acids in the chosen protein, with glutamate shown to be the most prevalent. Furthermore, hydrogen had the highest proportion in the atomic makeup of the protein, with carbon, oxygen, nitrogen, and sulfur following in descending order. The PPI networks demonstrate the possibility

for both primary and secondary interactions with other proteins. We employed computational modeling techniques to analyze and evaluate the secondary and tertiary structures of the chosen protein in order to gain insights into its characteristics. Computational functional analysis anticipated that the protein would serve as a chaperone effector. Through the process of designing and developing medications and vaccines, we can utilize this protein as a specific target for in-depth examination in order to treat diseases caused by *C. pneumoniae*.

Author Contributions: Conceptualization, A.S.M.S. and M.L.K.; methodology, A.S.M.S. and M.L.K.; software, A.S.M.S., T.A., U.S., K.N.U., M.M.H. and M.L.K.; validation, A.S.M.S., T.A., U.S., K.N.U., M.M.H. and M.L.K.; formal analysis, A.S.M.S., T.A. and U.S.; investigation, A.S.M.S., K.N.U., M.M.H. and M.L.K.; resources, A.S.M.S., K.N.U., M.M.H. and M.L.K.; data curation, A.S.M.S., T.A., U.S., K.N.U., M.M.H. and M.L.K.; writing—original draft preparation, A.S.M.S., T.A., U.S., K.N.U., M.M.H. and M.L.K.; writing—review and editing, A.S.M.S., K.N.U., M.M.H. and M.L.K.; visualization, A.S.M.S., T.A., U.S., K.N.U., M.M.H. and M.L.K.; supervision, M.L.K.; project administration, A.S.M.S. and M.L.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

COPD = Chronic obstructive pulmonary disease, GRAVY = Grand average of hydropathy, NCBI = National Center for Biotechnology Information, CD-Search = Conserved Domain Search, SOPMA = Self-Optimized Prediction Method with Alignment, FASTA = Fast Adaptive Shrinkage Threshold Algorithm, GMQE = Global Model Quality Estimation, QMEAN = Qualitative Model Energy Analysis, ProSA-web = Protein Structure Analysis.

References

1. Premachandra, N.M.; Jayaweera, J. *Chlamydia pneumoniae* infections and development of lung cancer: Systematic review. *Infect. Agent. Cancer* **2022**, *17*, 11.
2. Krüll, M.; Maass, M.; Suttorp, N.; Rupp, J. *Chlamydophila pneumoniae*. Mechanisms of target cell infection and activation. *Thromb. Haemost.* **2005**, *94*, 319–326.
3. Puri, B.K.; Lee, G.S.; Schwarzbach, A. The Role of *Chlamydia pneumoniae* in the Aetiology of Autoimmune Diseases. *Cureus* **2023**, *15*, e49095.
4. Hashemian, S.M.M.; Madani, S.A.; Allymehr, M.; Talebi, A. A molecular survey of *Chlamydia* spp. infection in commercial poultry and detection of *Chlamydia pneumoniae* in a commercial turkey flock in Iran. *Vet. Med. Sci.* **2023**, *9*, 2168–2175.
5. Chen, Q.; Lin, L.; Zhang, N.; Yang, Y. Adenovirus and *Mycoplasma pneumoniae* co-infection as a risk factor for severe community-acquired pneumonia in children. *Front. Pediatr.* **2024**, *12*, 1337786.
6. Sayers, E.W.; Beck, J.; Bolton, E.E.; Bourexis, D.; Brister, J.R.; Canese, K.; Comeau, D.C.; Funk, K.; Kim, S.; Klimke, W.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2021**, *49*, D10–D17.
7. Wilkins, M.R.; Gasteiger, E.; Bairoch, A.; Sanchez, J.C.; Williams, K.L.; Appel, R.D.; Hochstrasser, D.F. Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* **1999**, *112*, 531–552.
8. Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; Mehryary, F.; Hachilif, R.; Gable, A.L.; Fang, T.; Doncheva, N.T.; Pyysalo, S.; et al. The STRING database in 2023: Protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **2023**, *51*, D638–D646.
9. Doncheva, N.T.; Morris, J.H.; Gorodkin, J.; Jensen, L.J. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J. Proteome Res.* **2019**, *18*, 623–632.
10. Marchler-Bauer, A.; Anderson, J.B.; Cherukuri, P.F.; DeWeese-Scott, C.; Geer, L.Y.; Gwadz, M.; He, S.; Hurwitz, D.I.; Jackson, J.D.; Ke, Z.; et al. CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Res.* **2005**, *33*, D192–D196.
11. Combet, C.; Blanchet, C.; Geourjon, C.; Deléage, G. NPS@: Network protein sequence analysis. *Trends Biochem. Sci.* **2000**, *25*, 147–150.
12. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F.T.; De Beer, T.A.P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303.

13. Bienert, S.; Waterhouse, A.; de Beer, T.A.P.; Tauriello, G.; Studer, G.; Bordoli, L.; Schwede, T. The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res.* **2017**, *45*, D313–D319.
14. Wiederstein, M.; Sippl, M.J. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **2007**, *35*, W407–W410.
15. Akbari Rokn Abadi, S.; Abdosalehi, A.S.; Pouyamehr, F.; Koohi, S. An accurate alignment-free protein sequence comparator based on physicochemical properties of amino acids. *Sci. Rep.* **2022**, *12*, 11158.
16. Saikat, A.S.M.; Ripon, A. Structure Prediction, Characterization, and Functional Annotation of Uncharacterized Protein BCRIVMBC126_02492 of Bacillus cereus: An In Silico Approach. *Am. J. Pure Appl. Biosci.* **2020**, *2*, 104–111.
17. Audain, E.; Ramos, Y.; Hermjakob, H.; Flower, D.R.; Perez-Riverol, Y. Accurate estimation of isoelectric point of protein and peptide based on amino acid sequences. *Bioinformatics* **2016**, *32*, 821–827.
18. Saikat, A.S.M. Computational approaches for molecular characterization and structure-based functional elucidation of a hypothetical protein from *Mycobacterium tuberculosis*. *Genom. Inform.* **2023**, *21*, e25.
19. Gill, S.C.; von Hippel, P.H. Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **1989**, *182*, 319–326.
20. Gamage, D.G.; Gunaratne, A.; Periyannan, G.R.; Russell, T.G. Applicability of Instability Index for In vitro Protein Stability Prediction. *Protein Pept. Lett.* **2019**, *26*, 339–347.
21. Ikai, A. Thermostability and aliphatic index of globular proteins. *J. Biochem.* **1980**, *88*, 1895–1898.
22. Alam, M.M.; Saikat, A.S.M.; Uddin, M.E. Bioinformatics Approaches for Structural and Functional Annotation of an Uncharacterized Protein of *Helicobacter pylori*. *Eng. Proc.* **2023**, *37*, 61.
23. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132.
24. Yousuf, M.; Saikat, A.S.M.; Uddin, M.E. A Bioinformatic Approach for Molecular Characterization and Functional Annotation of an Uncharacterized Protein from *Vibrio cholerae*. *Eng. Proc.* **2023**, *37*, 62.
25. Soleymani, F.; Paquet, E.; Viktor, H.; Michalowski, W.; Spinello, D. Protein-protein interaction prediction with deep learning: A comprehensive review. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 5316–5341.
26. Saikat, A.S.M.; Al-Khafaji, K.; Akter, H.; Choi, J.-G.; Hasan, M.; Lee, S.-S. Nature-Derived Compounds as Potential Bioactive Leads against CDK9-Induced Cancer: Computational and Network Pharmacology Approaches. *Processes* **2022**, *10*, 2512.
27. Lorenzini, E.; Singer, A.; Singh, B.; Lam, R.; Skarina, T.; Chirgadze, N.Y.; Savchenko, A.; Gupta, R.S. Structure and protein-protein interaction studies on Chlamydia trachomatis protein CT670 (YscO Homolog). *J. Bacteriol.* **2010**, *192*, 2746–2756.
28. Wardah, W.; Khan, M.; Sharma, A.; Rashid, M.A. Protein secondary structure prediction using neural networks and deep learning: A review. *Comput. Biol. Chem.* **2019**, *81*, 1–8.
29. Al Asad, M.; Shorna, S.A.; Saikat, A.S.M.; Uddin, E. Computational Approaches for Structure-Based Functional Annotation of an Uncharacterized Conserved Protein of *Acinetobacter baumannii*. *Eng. Proc.* **2023**, *37*, 25.