In silico Structural and Functional Annotation of *Mycoplasma genitalium* Hypothetical Protein MG_377

Sudip Paul[1*], Moumoni Saha[1], Nikhil Chandra Bhoumik[2], Sattya Narayan Talukdar[3]

[1]*Department of Biochemistry and Molecular Biology*
Jahangirnagar University
Savar, Dhaka-1342, Bangladesh
E-mails: sudippaul.bcmb@gmail.com, moumonisaha@gmail.com

[2]*Wazed Miah Science Research Center*
Jahangirnagar University
Savar, Dhaka-1342, Bangladesh
E-mail: nikhil@juniv.edu

[3]*Department of Biochemistry*
Primeasia University
Banani, Dhaka-1213, Bangladesh
E-mail: sattya.narayan@primeasia.edu.bd

[*]*Corresponding author*

Received: October 31, 2014 **Accepted: March 26, 2015**

Published: April 01, 2015

Abstract: *Mycoplasma genitalium, a Gram-positive sexually transmitted pathogen, has been associated with urethritis in men and several inflammatory reproductive tract syndromes in women including cervicitis, pelvic inflammatory disease, and infertility. The complete sequence of the M. genitalium G37 genome revealed that it consists of 94 hypothetical proteins with unknown function in addition to functional proteins. In the present study, the MG_377 hypothetical protein of M. genitalium was selected for analyzing and modeling by different bioinformatics tools and databases. According to primary and secondary structure analyses, MG_377 is a stable hydrophilic protein containing a significant proportion of α-helices; besides, it is a cytoplasmic protein based on subcellular localization predictions. Homology modeling method was applied to generate its 3D structure using SWISS-MODEL server where the template PDB 1ZXJ with 84.4% sequence identity with the hypothetical protein was exploited. Several evaluations of quality assessment and validation parameters specified the generated protein model as reliable with fairly good quality. Functional genomics analysis carried out by InterProScan, Pfam and NCBI-CDD suggested that the hypothetical protein may contain Trigger factor/SurA domain. Moreover, comparative genomics analysis recommended MG_377 as a non-homologous protein essential for the organism. Further experimental validation would help to identify the actual function of MG_377 as well as to confirm the utility of the protein as drug targets.*

Keywords: *Mycoplasma genitalium, Hypothetical proteins, Homology modeling, Functional genomics, Comparative genomics.*

Introduction

Hypothetical proteins (HPs) are proteins whose existence has been predicted but *in vivo* function has not been established [9]. HPs generally cover around half the protein coding regions in most genomes. Although their functions have yet not been well characterized, they might have their own importance to complete genomic and proteomic information [16, 22]. Proper structural and functional annotations of HPs of particular genome may lead to the

sighting of new structures as well as new functions and help to introduce a list of additional protein pathways and cascades, thus completing our scrappy knowledge on the mosaic of proteins [16]. Elucidating the structural and functional secrets of these HPs may also lead to a better understanding of the protein-protein interactions or networks in different forms of life such as plants, microorganisms, etc. [12]. Furthermore, novel HPs may also serve as markers and pharmacological targets for drug design, discovery and screening [18, 20].

To date, 94 such proteins have been identified within the *Mycoplasma genitalium* G 37 genome with no known function [6]. *M. genitalium* is a Gram-positive bacterium, commonly responsible for acute and chronic nongonococcal urethritis (NGU) in man and considered as an etiologic agent of cervical inflammation and upper tract disease syndromes, including pelvic inflammatory disease (PID) and infertility [13]. Triumphant treatment of *M. genitalium* infection in female patients is of particular importance because protracted inflammation at upper genital tract sites might show the way to considerable reproductive tract morbidity and infertility [24]. The HPs that lack structure-function annotations in the genome of *M. genitalium* would be a good target for drug discovery and design.

In recent years, a number of hypothetical proteins have been discovered in the genome of many organisms. However, due to several limitations such as the cost and time required for experimental approaches, complete genome annotations have not achieved yet. Moreover, the large quantity of hypothetical proteins in a genome makes their study a difficult task. Bioinformatics approaches utilizing different algorithms and databases to estimate protein function would be a good alternative to laboratory-based methods. As these algorithms and databases are based on experimental results, they can be an effective means to perform functional and structural annotation of hypothetical proteins.

In the present study, the *M. genitalium* hypothetical protein MG_377 was selected as the primary amino acid sequence of the protein is available but structural details are not available. The study aimed to analyze the physiochemical and secondary structure features, to generate the first three dimensional (3D) model through homology modeling, and finally to conduct functional and comparative genomics analyses of the *M. genitalium* hypothetical protein MG_377. The outcome of this work will be helpful for better understanding of the mechanism of pathogenesis and finding novel therapeutic targets for *M. genitalium*.

Materials and methods

The sequential work plan was conducted according to a published article [6] with some modifications.

Sequence retrieval

The amino acid sequence of the *M. genitalium* hypothetical protein MG_377 was retrieved from the Uniprot database (http://www.uniprot.org/) using the primary accession number P47617 and the entry name, Y377_MYCGE.

Physiochemical analysis of the protein

Analysis of the physiochemical characteristics of the studied protein such as molecular weight, theoretical pI, amino acid composition, atomic composition, instability index, and grand average of hydropathicity (GRAVY) was performed using ProtParam tool (http://web.expasy.org/protparam/) [10].

Secondary structure analysis

The server SOPMA was employed for secondary structure predictions (helix, sheets, and coils) of the hypothetical protein [11]. In addition to that, the PSIPRED [5, 13] and PredictProtein [25] servers were also exploited to validate the results attained from SOPMA.

Subcellular localization prediction

Subcellular localization of MG_377 was predicted by CELLO v.2.5 [18]. Results were also cross-checked with subcellular localization predictions obtained from PSORTb version 3.0.2 and PredictProtein servers [29].

Homology modeling of the hypothetical protein

The possible 3D structure of the protein MG_377 was built through alignment mode in protein structure homology modeling server SWISS-MODEL [1, 15] using the full amino acid sequence of the protein in FASTA format.

Quality assessment of the 3D model and visualization

The initial structural model was checked for recognition of errors in 3D structure [26] by ERRAT and Verify3D programs included in structural analysis and verification server SAVES (http://nihserver.mbi.ucla.edu/SAVES/) [4, 7]. The final model structure quality of MG_377 was assessed by QMEAN [3] and checked by protein stereology with ProSA program [27]. The Ramachandran plots for all the models were generated using the RAMPAGE server [17], showing the percentage of protein residues in the favored, allowed and outlier regions. Furthermore, the generated and template structures were superimposed and Root Mean Square Deviation (RMSD) was obtained. The superimposition and visualization of generated models were performed by UCSF Chimera 1.8.1 [23].

Functional annotation of the protein

M. genitalium hypothetical protein MG_377 was analyzed for the presence of conserved domains based on sequence similarity search with close orthologous family members. Three different bioinformatics tools and databases including InterProScan [30], Proteins Families Database (Pfam) [8], and NCBI Conserved Domains Database (NCBI-CDD) [19] were used for this purpose.

Comparative genomics analysis of the protein

Comparative genomics analysis of MG_377 was performed by conducting BLASTP search between *Homo sapiens* proteome and MG_377 to test its resemblance to humans. Hits were filtered on the basis of expectation value (e-value) inclusion threshold being set to 0.005, and a minimum bit score of 100. The Database of Essential Genes (DEG) was utilized for obtaining information about essential proteins of *M. genitalium* [31]. E-value cut-off of 10^{-10} and a minimum bit score of 100 were used to scan MG_377 against essential proteins listed in DEG from 17 different Gram-positive and Gram-negative bacteria by DEG microbial BLASTP. Furthermore, KEGG automatic annotation server (KAAS) was used to analyze the contribution of MG_377 in the metabolic pathways of *M. genitalium* [21].

Submission of the model in protein model database (PMDB)

The model generated for *M. genitalium* hypothetical protein MG_377 was successfully submitted in Protein model database (PMDB) (http://bioinformatics.cineca.it/PMDB/).

Results and discussion

Physiochemical characteristics of MG_377

The ExPASy's ProtParam server was used to analyze the theoretical physiochemical characteristics from the amino acid sequence of the hypothetical protein MG_377. The protein was predicted to be consisting of 193 amino acids, with a molecular weight of 22649.1 Daltons and an isoelectric point (P^I) of 6.24 indicating a negatively charged protein. The instability index of the protein was computed to be 27.88, classified this protein as stable. The negative GRAVY index of -0.356 is indicative of a hydrophilic and soluble protein. The most abundant amino acid residue was found to be Lysine (21), followed by Glutamic acid and Isoleucine (19 each). The lowest was found as Tryptophan (1). The sequence had 31 negatively charged residues (Aspartic acid + Glutamic acid) and 29 positively charged residues (Arginine + Lysine). The molecular formula of the protein was found as $C_{1030}H_{1633}N_{269}O_{298}S_3$.

Subcellular localization of MG_377

Predicting subcellular localization of unknown proteins can give information about their cellular functions. This information could be utilized in understanding disease mechanism and developing drugs [31]. The subcellular localization of the query protein was predicted to be a cytoplasmic protein. It was analyzed by CELLO and authenticated by PSORTb v3.2.0 and PredictProtein severs.

Secondary structure of MG_377

The secondary structure of the protein was initially predicted by SOPMA server. The alpha helix was found to be the most predominant (73.58%), followed by random coil (20.73%) and extended strand (4.66%). Also, beta turn was found as 1.04%. Similar results were found from PredictProtein and PSIPRED servers. The representative secondary structure of MG_377 obtained from the PSIPRED server is shown in Fig. 1.

Homology modeling of MG_377

To perform the homology modeling, the query sequence was given as input in SWISS-MODEL server. The server automatically performed BLASTP search for each protein sequence to identify templates for homology modeling. For each identified template, the template's quality has been predicted from features of the target-template alignment. The templates with the highest quality have then been selected for model building. In this particular search, PDB ID 1ZXJ was selected as the template for homology modeling which is an X-ray diffraction model of a *M. pneumoniae* hypothetical protein with an 84.46% sequence identity, which was a good score to begin modeling. The 3D model was viewed by UCFS Chimera 1.8.1 and shown in Fig. 2.

Quality assessment and visualization

Reliability of the generated model was initially checked by ERRAT that analyzed the statistics of non-bonded interactions between different atom types based on characteristic atomic interactions. The overall quality factor was found as 90.00 which good enough to use this model. The overall quality factor of the template was found as 89.08 indicating a structure with good high resolution. As indicated by the Verify3D program, the results showed that 85.79% and 66.67% of MG_377 and 1ZXJ template residues had an average 3D (atomic model) – 1D (amino acid) score ≥ 0.2 also meaning that these structures were compatible and fairly good.

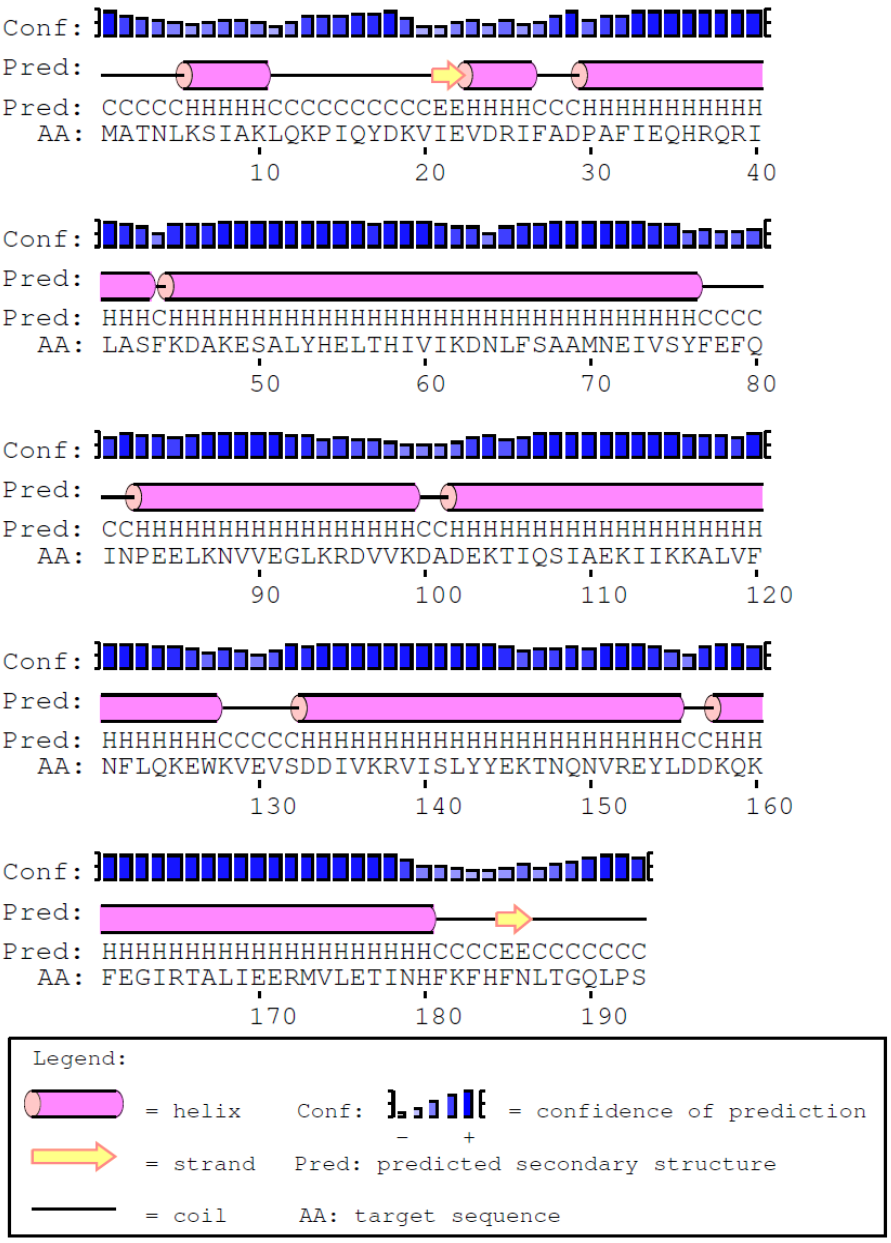


Fig. 1 Predicted secondary structure of *Mycoplasma genitalium* hypothetical protein MG_377 by PSIPRED server

The quality and reliability of the predicted 3D model was further checked by several structure assessment methods including Q values, Z-scores, Ramachandran plots and RMSD value. The Q values attained from the QMEAN analysis for MG_377 and 1ZXJ structures were found as 0.678 and 0.642, respectively. The values indicate the suitability of them for subsequent analyses. The Z-score is used to estimate the quality of model using structured solved proteins as references [2]. Z-scores of the template and query model were obtained from PROSAweb. The template Z-score was -4.05 (Fig. 2e) and for the MG_377 homology model it was -5.79 (Fig. 2d), suggesting significant similarity between template and query structures. The stereochemical quality of the model protein was analyzed using Ramachandran plots through the RAMPAGE server (Fig. 3). Ramachandran plot analysis found 92% of residues of the protein's model structure in the favored region, with 5.9%, and 2.1% residues in allowed and outlier regions, respectively, indicating that the model was reliable and of good quality. The Ramachandran plot for the template structure showed the

amino acid residues to be 89.7%, 9.2% and 1.1%, in favored, allowed and outlier regions respectively (Fig. 3).

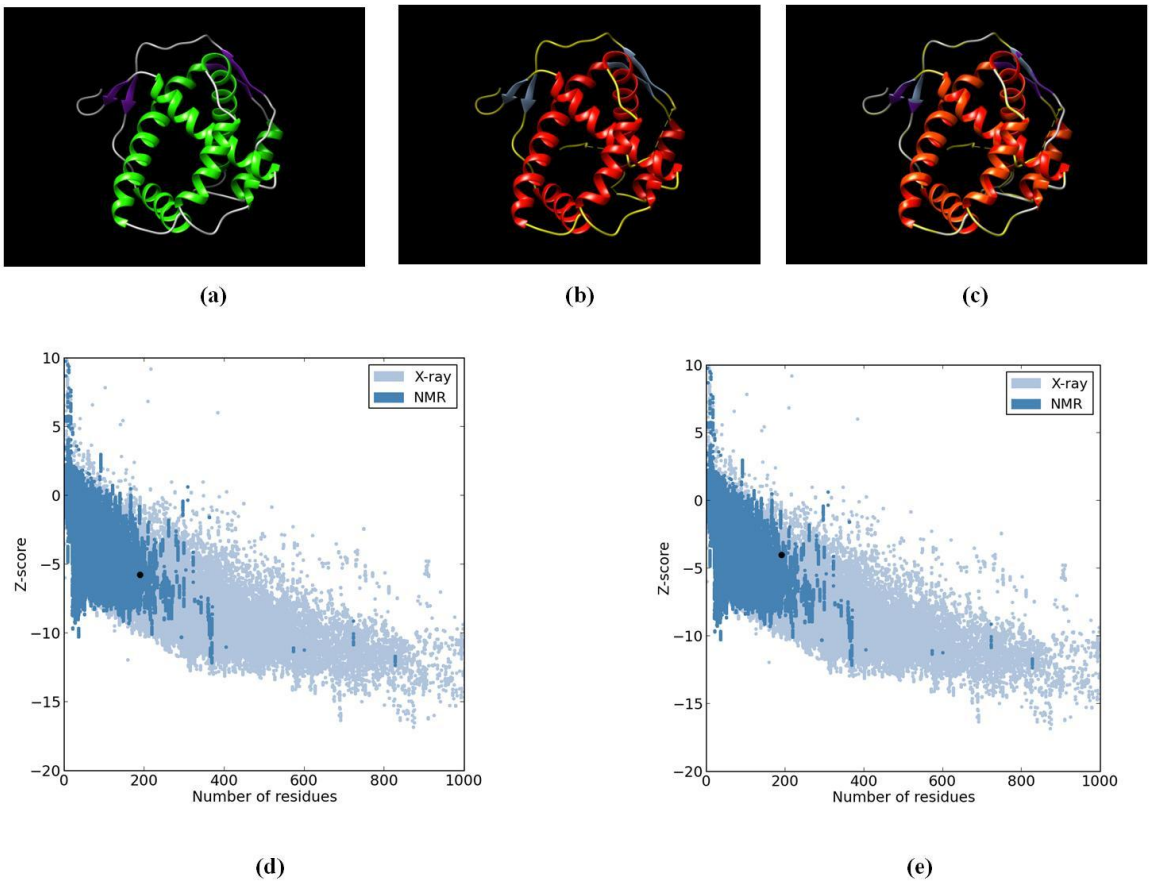


Fig. 2 Structural analysis of *Mycoplasma genitalium* hypothetical protein MG_377 (a) the predicted 3D model of MG_377; (b) the 3D model of the template protein (1ZXJ) used for homology modeling; (c) Superimposed structure of hypothetical and template proteins; (d) and (e) the Z-scores for the hypothetical and template proteins respectively.

The RMSD value indicates the degree to which extent two 3D structures are similar. The lower the value, the more similar are the structures. Both template and query structures were superimposed for the calculation of RMSD (Fig. 2c). The RMSD value obtained from superimposition of MG_377 and 1ZXJ in UCSF Chimera was found to be 0.0.087Å, suggesting a reliable 3D structure.

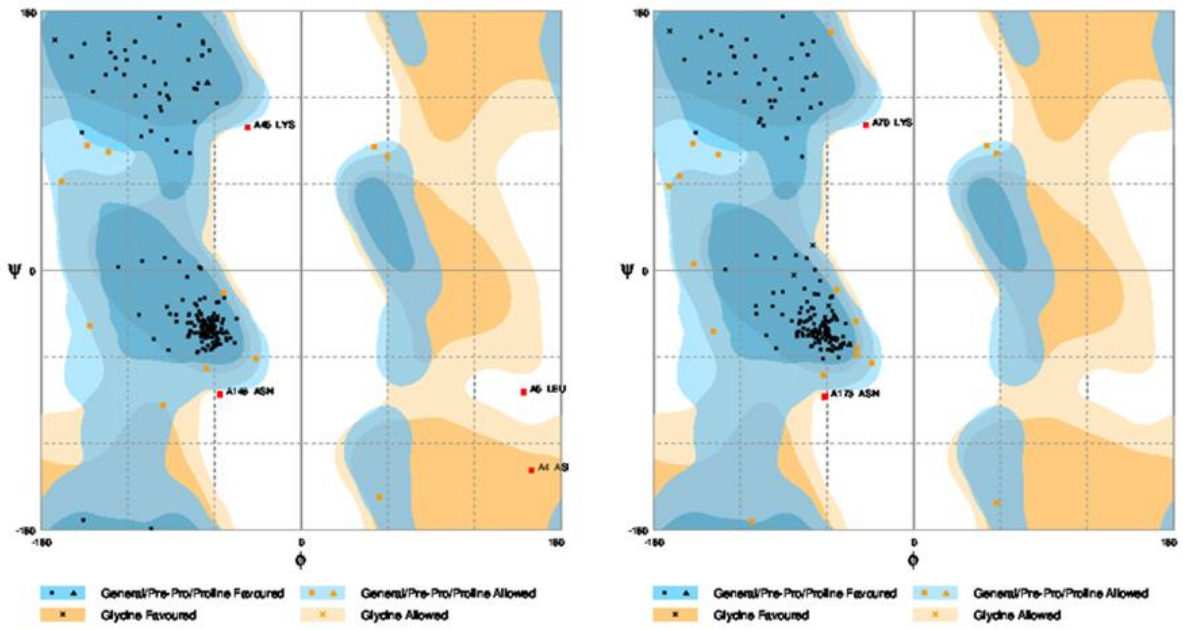
The comparable Q values, Z-scores, Ramachandran plot characteristics and the small RMSD value confirm the quality of the homology model of MG_377.

The final protein structure was deposited in PMDB and is available under ID: PM0079764.

Functional annotation and comparative genomics analysis of MG_377

The function of MG_377 has yet not been confirmed. In the present study, we used three web tools to search the conserved domains and potential functions of MG_377. Based on consensus predictions made by InterProScan, Pfam and NCBI-CDD, MG_377 was suggested to contain Trigger factor/SurA domain. Trigger Factor is an ATP-independent chaperone and displays chaperone and peptidyl-prolyl-cis-trans-isomerase (PPIase) activities *in vitro*.

It is required for folding of newly synthesized proteins to native state. The porin chaperon SurA facilitates correct folding of outer membrane proteins in Gram-negative bacteria.



(a) (b)

Fig. 3 Ramachandran plot analysis of 3D protein model by RAMPAGE server
(a) Ramachandran plot for the 3D model of the studied hypothetical protein MG_377;
(b) Ramachandran plot for the 3D model of template protein.

After performing functional annotation of the hypothetical protein, we applied comparative genomics approach to further characterize MG_377. At first, a BLASTP search against human proteome was performed to identify whether MG_377 has any human homologues. MG_377 was identified as a unique protein of *M. genitalium* and showed no homology to any of the human proteins. Targeting microbial proteins that are non homologous to human proteins for drug action would be a relatively effective and secured option in terms of generating any side effects [6].

Essential proteins of a pathogen regulate key factors, such as metabolism, nutrient uptake, virulence and pathogenicity. Therefore these proteins are of immense importance in disrupting pathogen functions and existence. Again, not all essential proteins are non-homologous in nature. Therefore, pathogenic proteins that are both unique and essential at the same time represent more striking drug targets. We retrieved the information about essential genes of *M. genitalium* from the DEG database. Microbial BLASTP search as per selection criteria mentioned in materials and methods section, suggested that it is an essential protein for the organism.

Finally, the screened hypothetical protein was found as "unassigned" by the KAAS server, therefore involvement of this protein in microbial metabolic pathways could not be predicted.

Conclusion

The present study was conducted to create the first 3D structure and propose possible functions of the *M. genitalium* hypothetical protein MG_377. The 3D model of the protein was generated using homology modeling method as well as refined by several structural

assessment methods and the final outcome was fairly good. We found that the protein is a stable cytoplasmic protein with probable Trigger factor/SurA domain. The protein was found to be essential for the organism and non-homologous to human; therefore would a potential candidate for drug design. Furthermore, this type of approach could be helpful in drug discovery for characterizing putative drug targets for other clinically important pathogens.

Acknowledgements

Authors greatly acknowledge the support provided by the Department of Biochemistry and Molecular Biology, Jahangirnagar University for conducting the research work.

References

1. Arnold K., L. Bordoli, J. Kopp, T. Schwede (2006). The SWISS-MODEL Workspace: A Web-based Environment for Protein Structure Homology Modelling, Bioinformatics, 22(2), 195-201.
2. Benkert P., M. Biasini, T. Schwede (2011). Toward the Estimation of the Absolute Quality of Individual Protein Structure Models, Bioinformatics, 27(3), 343-350.
3. Benkert P., S. C. E. Tosatto, D. Schomburg (2008). QMEAN: A Comprehensive Scoring Function for Model Quality Assessment, Proteins: Structure, Function and Genetics, 71(1), 261-277.
4. Bowie J. U., R. Luthy, D. Eisenberg (1991). A Method to Identify Protein Sequences that Fold into a Known Three-dimensional Structure, Science, 253(5016), 164-170.
5. Buchan D. W. A., F. Minneci, T. C. O. Nugent, K. Bryson, D. T. Jones (2013). Scalable Web Services for the PSIPRED Protein Analysis Workbench, Nucleic Acids Research, 41(W1), W340-W348.
6. Butt A. M., M. Batool, Y. Tong (2011). Homology Modeling, Comparative Genomics and Functional Annotation of *Mycoplasma genitalium* Hypothetical Protein MG_237, Bioinformation, 7(6), 299-303.
7. Colovos C., T. O. Yeates (1993). Verification of Protein Structures: Patterns of Nonbonded Atomic Interactions, Protein Science, 2(9), 1511-1519.
8. Finn R. D., J. Mistry, J. Tate, P. Coggill P, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman (2010). The Pfam Protein Families' Database, Nucleic Acids Research, 38(Database issue), D211-D222.
9. Galperin M. Y., E. V. Koonin (2004). 'Conserved Hypothetical' Proteins: Prioritization of Targets for Experimental Study, Nucleic Acids Research, 32, 5452–5463.
10. Gasteiger E., C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, A. Bairoch (2005). Protein Identification and Analysis Tools on the ExPASy Server, In: The Proteomics Protocols Handbook, Walker J. M. (Ed), Humana Press, 571-607.
11. Geourjon C., G. Deléage (1995). SOPMA: Significant Improvements in Protein Secondary Structure Prediction by Consensus Prediction from Multiple Alignments, Computer Applications in Biosciences, 11(6), 681-684.
12. Idrees S., S. Nadeem, S. Kanwal, B. Ehsan, A. Yousaf, S. Nadeen, M. I. Rajoka (2012). *In silico* Sequence Analysis, Homology Modeling and Function Annotation of *Ocimum basilicum* Hypothetical Protein G1CT28_OCIBA, International Journal of Bioautomation, 16(2), 111-118.
13. Jensen J. S. (2004). *Mycoplasma genitalium*: The Aetiological Agent of Urethritis and Other Sexually Transmitted Diseases, Journal of the European Academy of Dermatology and Venereology, 18(1), 1-11.
14. Jones D. T. (1999). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices, Journal of Molecular Biology, 292, 195-202.

15. Kiefer F., K. Arnold, M. Kunzli, L. Bordoli, T. Schwede (2009). The SWISS-MODEL repository and associated resources, Nucleic Acids Research, 37(1), D387-D392.
16. Loewenstein Y., D. Raimondo, O. C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton, A. Tramontano (2009). Protein Function Annotation by Homology-based Inference, Genome Biology, 10, 207.
17. Lovell S. C., I. W. Davis, W. B. Arendall IIIrd, P. I. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, D. C. Richardson (2003). Structure Validation by Calpha Geometry: Phi, Psi and Cbeta Deviation, Proteins, 7, 437-450.
18. Lubec G., L. Afjehi-Sadat, J. W. Yang, J. P. John (2005). Searching for Hypothetical Proteins: Theory and Practice Based upon Original Data and Literature, Progress of Neurobiology, 77(1-2), 90-127.
19. Marchler-Bauer A., S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. N. Zhang, C. Zheng, S. H. Bryant (2011). CDD: A Conserved Domain Database for the Functional Annotation of Proteins, Nucleic Acids Research, 39(Database Issue), D225-D229.
20. Minion F. C., E. J. Lefkowitz, M. L. Madsen, B. J. Cleary, S. M. Swartzell, G. G. Mahairas (2004). The Genome Sequence of *Mycoplasma hyopneumoniae* Strain 232, the Agent of Swine Mycoplasmosis, Journal of Bacteriology, 186(21), 7123-7133.
21. Moriya Y, M. Itoh, S. Okuda, A. C. Yoshizawa, M. Kanehisa (2007). KAAS: An Automatic Genome Annotation and Pathway Reconstruction Server, Nucleic Acids Research, 35(Web Server Issue), W182-W185.
22. Nimrod G., M. Schushan, D. M. Steinberg, N. Ben-Tal (2008). Detection of Functionally Important Regions in "Hypothetical Proteins" of Known Structure, Structure, 16(12), 1755-1763.
23. Pettersen E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin (2004). UCSF Chimera: A Visualization System for Exploratory Research and Analysis, Journal of Computational Chemistry, 25(13), 1605-1612.
24. Quayle A. J. (2002). The Innate and Early Immune Response to Pathogen Challenge in the Female Genital Tract and the Pivotal Role of Epithelial Cells, Journal of Reproductive Immunology, 57(1-2), 61-79.
25. Rost B., G. Yachdav, J. Liu (2004). The PredictProtein Server, Nucleic Acids Research, 32(Web Server Issue), W321-W326.
26. Sippl M. J. (1993). Recognition of Errors in Three-dimensional Structures of Proteins, Proteins, 17, 355-362.
27. Wiederstein M., M. J. Sippl (2007). ProSA-web: Interactive Web Service for the Recognition of Errors in Three-dimensional Structures of Proteins, Nucleic Acids Research, 35, W407-W410.
28. Yu C. S., Y. C. Chen, C. H. Lu, J. K. Hwang (2006), Prediction of Protein Subcellular Localization, Proteins: Structure, Function and Bioinformatics, 64, 643-651.
29. Yu N. Y., J. R. Wagner, M. R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S. C. Sahinalp, M. Ester, L. J. Foster, F. S. Brinkman (2010). PSORTb 3.0: Improved Protein Subcellular Localization Prediction with Refined Localization Subcategories and Predictive Capabilities for All Prokaryotes, Bioinformatics, 26(13), 1608-1615.
30. Zdobnov E. M., R. Apweiler (2001). InterProScan: An Integration Platform for the Signature-recognition Methods in InterPro, Bioinformatics, 17(9), 847-848.
31. Zhang R., H. Y. Ou, C. T. Zhang (2004). DEG: A Database of Essential Genes, Nucleic Acids Research, 32(Database Issue), D271-D272.

Sudip Paul, M.Sc.
E-mail: sudippaul.bcmb@gmail.com

Sudip Paul is working as a Lecturer in the Department of Biochemistry and Molecular Biology, Jahangirnagar University, Savar, Dhaka. He completed B.Sc (Hons.) and M.Sc. in Biochemistry and Molecular Biology from Jahangirnagar University, Bangladesh. His areas of research are Bioinformatics, Medicinal Chemistry, Molecular Pharmacology, Cancer Genetics and Public Health.

Moumoni Saha, M.Sc.
E-mail: moumonisaha@gmail.com

Moumoni Saha completed her M.Sc. in Biochemistry and Molecular Biology from Jahangirnagar University, Bangladesh. Her scientific interests are Molecular Pharmacology, Clinical Biochemistry, Medicinal Chemistry and Bioinformatics.

Nikhil Chandra Bhoumik, M.Sc.
E-mail: nikhil@juniv.edu

Nikhil Chandra Bhoumik completed M.Sc. in Chemistry from Jahangirnagar University (JU), Bangladesh. Currently he is working as a Research Officer in Wazed Miah Science Research Center of JU. His research interests are Bioinformatics, Analytical and Inorganic Chemistry, etc.

Sattya Narayan Talukdar, M.Sc.
E-mail: sattya.narayan@primeasia.edu.bd

Sattya Narayan Talukdar completed his M.Sc. in Biochemistry and Molecular Biology from Jahangirnagar University, Bangladesh. He is presently working as a Lecturer in Department of Biochemistry, Primeasia University, Dhaka. His fields of interests are Phytopharmacology, Molecular Biology, Bioinformatics, and Public Health and Nutrition.