ScienSage

Journal of Advanced Scientific Research

Available online through http://www.sciensage.info

ISSN
0976-9595
Research Article

GENOME ANNOTATION OF UNCHARACTERISED PROTEINS IN NEISSERIA MENINGITIDIS K1207

Nikita Sharma* and Kamaldeep Kaur

Department of Biotechnology and Bioinformatics, (Guru Nanak Girls college, Model Town, Ludhiana)

**Corresponding author: niksshrma3@gmail.com*

ABSTRACT

The origin of high-throughput biology has catalyzed a phenomenal improvement in our ability to identify new proteins. Structural and functional characterization of Hypothetical proteins reveal vital roles in bacteria, especially in pathogens related to human diseases. Neisseria meningitidis strain K1207 is a diplococcus Gram-negative bacteria that possesses enzymes that play important role in many cellular mechanisms. However, from its genome many proteins are Hypothetical or un-annotated. These functionally unknown proteins may specify important functions concerning biological role of this bacterium. By using bioinformatics tools the biological understanding of the organisms easily done through functional annotation analysis. There were different bioinformatics tools used for annotation- NCBI, MOTIF SCAN, INTERPRO, BLAST, TMMOD, CRNPRED, SOSUI, PROTPARAM, PSLPRED, PHYRE2. The sequence of 130 Hypothetical proteins were submitted to this workflow resulting in successful identification of proteins based on several parameters. Our study helps to found out two proteins at last, these proteins possess confidence score above 90%. These proteins play important roles in aerobic and anaerobic growth of gram negative bacteria and also in various signaling pathways. Data obtain by bioinformatics tools might facilitate swift identification of prospective therapeutic targets and thereby enabling the search for new inhibitors or vaccines. Our research unlock the chance to analyze the various applications of these functionally predicted proteins in area of biotechnology.

Keywords: Annotation, Hypothetical, Neisseria meningitidis, Analysis, Prediction.

1. INTRODUCTION

Neisseria meningitidis, a gram negative and diplococcus bacteria which is found in upper respiratory tract of human body and belongs to family *Neisseriaceae*. *N. meningitidis* is differentiated into 13 sero-groups and sub-serogroups. *N. meningitides* cause epidemic meningitidis. This disease is spread by droplet infection from person to person, Meningococcal disease can also cause bloodstream infections [1]. The disease caused by N. *meningitidis* is serious communicable disease, associated with substantial mortality rate. The small sub-unit 16s RNA distinguish *N.meningitidis* from other subfamily species, particularly *Moraxella* [2]. The disease caused by this bacteria is multi-step process it start with metabolic adaptations to colonization of host nasopharynx and then continuing with the contribution of core metabolism to replicate in the bloodstream and immune evasion and the last step is with the invasion of subarachnoidal space with subsequent replication in human CSF [3]. On the basis of antigenicity of polysaccharide capsule out of 13, A, B, C, W, Y and X are life threatening disease and these serogroups are most frequently involve in epidemics or outbreaks of meningococcal meningitidis [4]. The 8-15% of the patients die within 24 and 28 hours of symptom onset even when patients diagnosed early and treated properly. The 10-20% of the survivors are having disabilities like brain damage, hearing loss and learning problem [5]. The immune system of human body is play most important role in defence against bacterial colonization. The innate immune response of the body protect irrespectively by antigen through epithelial and phagocytic cells and with antimicrobial action [6]. The symptoms of meningeal infection in body are fever, headache and stiff neck, often accompanied by nausea, vomiting, photo-phobia and altered mental status. There are approximately 10% adults are asymptotic transient carriers of *N. meningitidis* [7]. Meningococcal disease was described dates back to the 16th century, by Vieusseux in 1805 during an outbreak with 33 deaths in the vicinity of Geneva, Switzerland [8]. The Italian pathologists

Marchiafava and Celli in 1884 first isolate intracellular oval micrococci in a sample of CSF [9]. The draft genome sequence for *N. meningitidis* Serogroup C, strains K1207 isolated from the same epidemic cluster which occurred in the Veneto region in northern Italy during the 2007-2008 winter. The method used for sequencing 454 pyrosequencing (Roche), combining shotgun and 30-kb paired-end strategies, according to the manufacturer's recommendations [10]. Currently available meningo - coccal vaccines based on the meningococcal capsule include both polysaccharide vaccines and polysaccharide-protein conjugate vaccines. The development in vaccines has included protein vaccines based on meningococcal outer membrane vesicles for serogroup B; more recently a range of conserved proteins including fHBP and nadA have been used as vaccine components [11]. The main aim of the current study is to analyze the function of hypothetical proteins and three-dimensional structure prediction of proteins using computational methods.

2. MATERIAL AND METHODS

2.1. Obtaining Sequence for Analysis (NCBI)

NCBI (https://www.ncbi.nlm.nih.gov) is a type of database that provides data analysis and retrieval and resources that operate on Genbank data and a variety of other biological data made available through NCBI. [12,13]. The complete genome of Neisseria meningitidis K1207 having Accession no. NZ-KE007332.1 and sequence of total 130 protein was retrieved in Fasta format from the corresponding NCBI protein sequence Database.

2.2. Discovering Domains for Analysis (INTERPRO)

We used INTERPRO for the prediction of domains present in hypothetical proteins. InterPro (https.//www.ebi.ac.uk/interpro/) is a combination of different databases, the PROSITE, PRINTS, Pfam and ProDom databases formed a consortium to amalgamate the predictive signatures they individually produced into a single resource [14].

2.3. Study of Motifs of Protein Sequence (MOTIF SCAN)

Motif scan (https://myhits.isb-sib.ch/cgi-bin/motif_scan) is a tool for finding all known motifs that occur in a sequence. This tool use motif sources or parameters to predict results. The sources are local and global models like Pfam HMMs (Hidden Markov Model) and PROSITE patterns, PROSITE profiles, HAMAP profiles.

2.4. Secondary Structure examination of hypothetical proteins (CRNPRED)

The CRNPRED program that predicts secondary structures (SS) of proteins and residue wise contact orders of a protein given its amino acid sequence. The prediction method is based on large-scale critical random networks. Globular protein domains are usually composed of the two basic secondary structure types, the a-helix and the b-strand, which are easily detectable by using CRNPRED [15].

2.5. Transmembrane Helices identification (TMMOD)

TMMOD(liao.cis.udel.edu/website/servers/TMMOD) is a database predicting the transmembrane protein topology that can be also used for identifying/discriminating helical membrane proteins from other proteins and this can be done by using Forward algorithm to calculate the model likelihood for a given sequence [16].

2.6. Checking Sequence Similarity (BLAST)

The Basic Local Alignment Search Tool (BLAST) is one of the most commonly used tools for comparing sequence information and retrieving sequences from databases. BLAST performs comparisons between pairs of sequences, searching for regions of local similarity. The resulting high-scoring pairs (HSPs) form the basis of the ungapped alignments that characterize BLAST output [17].

2.7. Inspection of protein solubility (SOSUI)

SOSUI (harrier.nagahama-i-bio-ac.jp/sosui/mobile/) is a database for predicting inner membrane proteins from amino acid sequences [18]. SOSUI is a tool which distinguishes between membrane and soluble proteins from amino acid sequences, and predicts the transmembrane helices for the former. This tool has very high accuracy for prediction and can be calculated very quickly [19].

2.8. Survey of Physical and Chemical Parameters (PROTPARAM)

ProtParam (http://web.expasy.org/protparam/) is a type of database can be used to calculate various physicochemical properties that are deduced from a protein sequence. The measurable factors computed by ProtParam include the molecular weight, theoretical

pI,amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index,aliphatic index, and grand average of hydropathicity. Molecular weight and theoretical pI are calculated as in Compute pI/Mw [20].

2.9. Recognition of Melting Point (TMPREDICTOR)

TmPredictor (http://tm.life.nthu.edu.tw/) is bioinformatics tool used to calculate the theoretical protein melting temperature. Tm is the temperature at which the protein undergoes the reversible (un)folding transition. To estimate Tm, it based on the sequence similarity between the proteins [21].

2.10. Analysis of Protein Location (PSLPRED)

PSLpred (crodd.osdd.net/raghava/pslpred/) server is better than existing servers for predicting prokaryotic subcellular localization. The prediction accuracy achieved for cytoplasmic proteins by PSLpred is similar to CELLO but 22% higher than that of PSORT-B [22]. For the prediction of subcellular localization of proteins SVM is used, it is machine-learning technique [23].

2.11. Scrutinizing Protein folding (PHYRE2)

Phyre2 (www.sbg.bio.ac.uk/phyre2/) is a web based database which is used to predict and analyze protein structure, function and mutations. This database uses advanced remote homology detection methods to build 3D models, predict ligand binding sites as well as several other features for the user's protein sequence. Phyre2 is easy to use and this is key factor of this database that differentiate it from others [24].

2.12. Verification of protein structure (SAVES V5.0)

Saves V5.0 (services.mbi.ucla.edu/SAVES) is a tool that helps to verify the 3D structures of proteins. The first is a Ramachandran plot, which is simply a scatter plot of the phi, psi values for the amino acids in a single protein structure, which we use here to mean a statistical representation of Ramachandran data, usually in the form of a probability density function (Ψ, Φ values). In SAVES V5.0, the PROCHECK suite of programs used for assessing the "stereo-chemical quality" of a given protein structure [25].

3. RESULTS AND DISCUSSIONS

3.1. Retrieval of Sequence of hypothetical proteins for Annotation (NCBI)

First, the Retrieval of 130 Hypothetical Protein sequences in FASTA format for analysis from NCBI with Accession no. NZ-KE007332.1. The hypothetical proteins are proteins which are having unknown function. The current study was done to annotate the uncharacterised proteins of N. meningitidis K1207.

3.2. Motifs and Domains determination using MOTIFSCAN and INTERPRO database

The Motifs were predicted in 127 Hypothetical Proteins and Domains in 15 Hypothetical Proteins. Motifs and Domains are important in functional annotation of hypothetical proteins because motifs and domains are structural and functional units of proteins, therefore prediction of protein domains are very useful in understanding the role of proteins in cellular pathways. The prediction of Motif and Domain is a base for functional annotation of proteins.

3.3. Secondary Structure of hypothetical proteins of N. meningitidis K1207 (CRNPRED)

We predict type of Secondary Structure present in proteins by using CRNPRED databases. The Secondary Structures are predicted in **127 Hypothetical Proteins**. This is prediction of folding and secondary structure from its primary structure. There were three types of secondary structures predicted coils, α, ß. The prediction of secondary structure is important for drug designing process. The most infinite type of secondary structure is alpha helix and it has 3.6 amino acids per turn. The beta sheets are type of secondary structures formed by H bonds between 5-10 amino acids and a recognizable turn, not a helix or not a beta sheet named as coil.

3.4. Sequence Similarity (BLAST) and Transmembrane helices (TMMOD)

The prediction of number of Transmembrane Helices present was done by using TMMOD and Sequence Similarity using BLAST. The **28 proteins** were analyzed to predict transmembrane helices and sequence similarity. The predicted similarity scores from of blast for most of the proteins was above 90%.

3.5. Protein Solubility investigation (SOSUI)

The Solubility of Hypothetical Proteins was predicted by using SOSUI database. In our research, the solubility of **16 proteins** was analyzed. SOSUI classify the proteins in two types membrane and soluble proteins. We select only membrane proteins for further steps of research because membrane protein are target of many drugs so it is helpful in developing a new drug.

Table 1: Results for Hypothetical Proteins Solubility from SOSUI

Accession No.	Solubility (SOSUI)	Hydrophobicity
WP-002221438.1	Membrane Protein	-0.27
WP-010980990.1	Membrane Protein	-0.12
WP-002236090.1	Membrane Protein	-0.36
WP-002214450.1	Membrane Protein	0.43
WP-002236742.1	Membrane Protein	0.59
WP-002222917.1	Membrane Protein	-0.14
WP-002221314.1	Membrane Protein	0.45
WP-002219540.1	Membrane Protein	0.73
WP-002213977.1	Membrane Protein	271428
WP-041423744.1	Membrane Protein	-0.14
WP-002220648.1	Membrane Protein	0.61
WP-002220650.1	Membrane Protein	0.5
WP-002220674.1	Membrane Protein	-0.32
WP-002220721.1	Membrane Protein	0.49
WP-002217443.1	Membrane Protein	0.3
WP-002220746.1	Membrane Protein	0.56

Table 2: Physical and Chemical Parameters analysis results of Hypothetical Proteins

Accession No.	No. of Amino acids	Moleclular weight	pI	Stability Index	Hydrophobicity	TmPredictor
WP-002221438.1	93	1098.72	9.93	Stable	-0.02	Lower than 55°C
WP-010980990.1	79	9798.38	8.76	Stable	0.02	Lower than 55°C
WP-002236090.1	208	23679.36	10.37	Stable	-0.36	Higher than 65°C
WP-002214450.1	56	6768.36	9.3	Stable	1.02	Higher than 65°C
WP-002236742.1	92	10997.4	9.48	Stable	1.02	Lower than 55°C
WP-002222917.1	152	17839.51	4.97	Stable	-0.09	Lower than 55°C
WP-002221314.1	77	8474.2	10.73	Stable	0.85	Higher than 65°C
WP-002219540.1	201	22405.91	9.37	Stable	0.95	Lower than 55°C
WP-002213977.1	120	13173.55	9.62	Stable	0.49	55°C-65 °C
WP-041423744.1	238	26252.2	6.43	Stable	-0.14	55°C-65 °C
WP-002220648.1	90	10169.54	10.13	Stable	1.01	Higher than 65°C
WP-002220650.1	96	11062.28	8.46	Stable	0.84	Higher than 65°C
WP-002220674.1	312	34508.22	8.73	Stable	-0.31	Higher than 65°C
WP-002220721.1	141	15154.99	6.9	Stable	0.73	Higher than 65°C
WP-002217443.1	91	10250.34	10.59	Stable	0.6	Lower than 55°C
WP-002220746.1	68	7893.31	6.5	Stable	1.12	Lower than 55°C

3.6. Physical and Chemical Parameters (PROTPARAM) and exploration of melting point (TMPREDICTOR)

In our current study, the Physical and Chemical Parameters prediction was important to check stability of proteins. Various physio-chemical properties were predicted like stability index, isoelectric point, hydrophobicity of proteins and also melting point of proteins.

3.7. Prognosis of location of Protein (PSLPRED)

The prediction of location of Hypothetical Proteins in cell done by using PSLPRED. The amino acid sequence of **16 proteins** were entered one by one in database to analyze the location of proteins. If the location of protein is known then it become easier to predict the function of protein because the function of protein is related with its location. In current research, there was two proteins are selected one was inner membrane protein and second was cytoplasmic protein.

Table 3: The Protein Localization result of Hypothetical Proteins from PSLPRED

Accession No.	Subcellular Localization (PSLPRED)
WP-002221438.1	Cytoplasmic Protein
WP-010980990.1	Inner Membrane Protein
WP-002236090.1	Inner Membrane Protein
WP-002214450.1	Inner Membrane Protein
WP-002236742.1	Inner Membrane Protein
WP-002222917.1	Cytoplasmic Protein
WP-002221314.1	Inner Membrane Protein
WP-002219540.1	Inner Membrane Protein
WP-002213977.1	Inner Membrane Protein
WP-041423744.1	Cytoplasmic Protein
WP-002220648.1	Inner Membrane Protein
WP-002220650.1	Inner Membrane Protein
WP-002220674.1	Cytoplasmic Protein
WP-002220721.1	Inner Membrane Protein
WP-002217443.1	Inner Membrane Protein
WP-002220746.1	Inner Membrane Protein

3.8. Three dimensional structure remembrance of protein (PHYRE2)

The prediction of Fold Recognition of Hypothetical proteins of N. meningitidis done by using Phyre2. We build 3D Structure of **6 hypothetical proteins** out of them, **2 hypothetical proteins** have shown confidence score above **98%**.

The 3D structure of proteins help to understand the conformation of protein and interactions of protein with neighboring proteins and help to determine the active site of protein, by knowing the active site of protein it become easy to design a drug against it.

Table 4: The Confidence score results of Hypothetical Proteins from PHYRE2

Accession No.	Confidence score (PHYRE2)
WP-010980990.1	20.6%
WP-002236090.1	99.4%
WP-002236742.1	16.6%
WP-002213977.1	6.2%
WP-002220674.1	98.6%
WP-002217443.1	40.9%

3.9. Ramachandran Plot evaluation by SAVESV5.0

At last, it is important to confirm the quality of model, SAVESV5.0 tool allow us to check the quality of protein structures of N. meningitidis by Ramachandran Plot.

The final models produced were perfectly reliable and 100% of the residues come in allowed region. The table below shows the final results of the research.

Table 5: The Ramachandran Plot analysis in 2 Hypothetical proteins by SAVESV5.0

Accession no.	Residues in most favored regions
WP-002236090.1	91.5%
WP-002220674.1	89.5%

Annotated Protein with function in N. meningitidis K1207

First protein which is functionally annotated is **BIOTIN** and lipoic acid share many common properties. Both vitamins are essential for aerobic growth of gram negative bacteria biotin is also required for growth of these bacteria under anaerobic conditions.

Both biotin and lipoic acid must be covalently attached to their coupled proteins to perform their roles in cellular enzymology, although free biotin plays an indirect regulatory role.

It is involved in a wide range of metabolic processes, both in humans and in other organisms, primarily related to the utilization of fats, carbohydrates, and amino acids. The disruption in gluconeogenesis, amino acid catabolism and fatty acid metabolism leads affected individuals to develop life threating ketoacidosis and organic acidemia that requires life-long pharmacological doses of biotin to be resolved.

Second predicted protein is MYRISTYL, a wide variety of stimuli can induce TNF-α production by immune cells. MYRISTYL also play role in induction of TNF-α production in at least some cell types. The TNF-α gene is one of the first genes expressed in T or B lymphocytes after these cells have been stimulated through their antigen receptors. In humans, the MYRISTYL group functions in conserved basic residues to facilitate membrane anchoring and assembly of Gag. Myristyl plays an essential role in membrane targeting, protein-protein interactions and functions widely in a variety of signal transduction pathways.

4. CONCLUSION

In current research, 130 hypothetical proteins have been taken from NCBI of Neisseria meningitidis K1207 and then functional and structural analysis of 130 hypothetical proteins of Neisseria meningitidis K1207 performed using different bioinformatics tools. In past years, the outbreaks of meningitidis disease has increased and this bacteria is reported as cause of death among many persons in U.S and Africa. Based on the Domain analysis, 127 out of 130 different hypothetical proteins under functional classification by using different freely available domain search tools such as InterPro and MOTIF SCAN. Out of 16 hypothetical processes, characterization of 2 proteins finally done, by analyzing the transmembrane helices, solubility, sub-cellular localization and Ramachandran plot. These properties can importantly helping in improving target recognition during drug-discovery process. One of these final predicted proteins having Accession no. **WP-002236090.1** with confidence score 99% is responsible for the aerobic and anaerobic growth of Neisseria meningitidis. Second predicted protein having Accession no. **WP-002220674.1** with confidence score 98% is responsible in membrane targeting and in many signal transduction pathways. Finally conclude that, the recent study carried out by computational approach, that may help us to understand the biology of Neisseria meningitidis K1207 and recognition of probable therapeutic target at molecular level.

5. REFERENCES

1. Gabutti G, Stefanati A, Kuhdari P. *J Prev Med Hyg,* 2015; **(56):**E116-E120.
2. Khater WS and Elabd SH. *International Journal of Microbiology*, 2016; 4197187.
3. Schoen C, Kischkies L, Elias J, et al. *Frontiers in Cellular and Infection Microbiology*, 2014; **(4)114:**16.
4. Munguambe AM, Cardoso de Almeida AEC, Nhantumbo AA, et al. *PLoS ONE*, 2014; **13(8):** e0197390.
5. Anouk M Oordt-Speets, Bolijn R, Rosa C van Hoorn, Bhavsar A, Moe H Kyaw. *Clinics and Practice,* 2017; **(7):**943.
6. Gasparini R, Amicizia D, Lai PL, Panatto D. *Journal of Preventive Medicine and Hygiene*, 2012; **53(2):** 50-5.
7. Chapter–Meningococcal Disease, Atkinson W, Centers for Disease Control and Prevention, Epidemiology and Prevention of Vaccine-Preventable Diseases, The Pink Book: Updated 11th Edition, 2009; 177-188.
8. Vieusseux M. *J Med Chir Pharmacol*, 1805; **(11):**163.
9. Weichselbaum A. *Fortschr Med*, 1887; **(5):**573-83.
10. Lavezzo E, Toppo S, Barzon L, Cobelli C, et al. *Journal of Bacteriology*, 2010; **192(19):**5270-5271.
11. Jokhdar H, Borrow R, Sultan A, Adi M, Riley C, Fuller E, et al. *Clin Diagn Lab Immunol,* 2014; **11(1):**83-8.
12. David L Wheeler, Chappey C, Alex E Lash, Detlef D. Leipe, Thomas L. Madden, et al. *Nucleic Acids Research,* 2000; **28(1):**10-14.
13. Dennis A Benson, Boguski M, David J Lipman and Ostell J. *Nucleic Acids Research*, 2013; **(41):**D36-D42.
14. Hunter S, Apweiler R, Teresa K Attwood, Bairoch A, Bateman A, et al. *Nucleic Acids Research,* 2009; **(37):**D211-D215.
15. Cymerman IA, Feder M, Paw Ł owski M, Kurowski MA, Bujnicki JM. *Nucleic Acids and Molecular Biology*, 2004; **15:**21.
16. Robel Y Kahsay, Gao G and Liao L. *Bioinformatics*, 2005; **21(9):**1853-1858.
17. Syngai GG, Barman P, Bharali R & Dey S. *Keanean Journal of Science*, 2013; **2:**67-76.
18. Imai K, Asakawa N, Tsuji T, Akazawa F, et al. *Bioinformation*, 2008; **2(9):**417-421.
19. Hirokawa T, Boon-Chieng S and Mitaku S. *Bioinformatics,* 1998; **14(4):**378-379.

20. Appaiah P and Vasu P. *J Proteomics Bioinform*, 2016; **9:**287-297.
21. Pucci F, Bourgeas R & Rooman M. *Scientific Reports*, 2016; **(6):**23257.
22. Bhasin M, Garg A and Raghava GPS. *Bioinformatics,* 2005; **21(10):**2522-2524.
23. Somvanshi P, Singh V, Seth PK. *Internet Journal of Genomics and Proteomics,* 2008; **3(2):**9.
24. Darby N and Creighton TE. *Humana Press*, 1995; **(40):**219-252.
25. Ting D, Wang G, Shapovalov M, Mitra R, Michael I Jordan, Roland L Dunbrack. *PLOS Computational Biology,* 2010; **6(40):**e1000763.