Subtractive genomics approach towards the identification of novel therapeutic targets against *Methicillin-resistant Staphylococcus aureus* (MRSA)

ABSTRACT

Methicillin-resistant Staphylococcus aureus (MRSA) is caused by the bacteria *Staphylococcus aureus*, which has developed resistance to numerous antimicrobial compounds. An extensive in-silico subtractive genomic technique was used on the genome of *Methicillin-resistant Staphylococcus aureus* to make suggestions for new therapeutic targets (DSM 20231). The pathogen's human homologous proteins and proteins linked to shared metabolic pathways between the pathogen and host were found using a variety of bioinformatics tools and web sites. In pathogen-specific pathways that were further examined to identify membrane proteins that were the targets of drugs, only one protein was found to be related. Protein sequences were discovered to be novel drug and therapeutic targets because they displayed to be a cytoplasmic protein. Additionally, the protein sequence showed broad spectrum conservation with other *Staphylococcus aureus* strains based on the research that was done. As a result, it might be applied to the creation of brand-new drugs and therapeutic elements for the effective management of infections brought by staph infections.

Keywords: *Methicillin-resistant Staphylococcus aureus*, Drug target, proteins, *Staphylococcus aureus*, antibiotic-resistance.

1. Introduction

The Staphylococcaceae family of bacteria includes the bacterium *Staphylococcus aureus*. They are gram-positive bacteria that grows in clusters that resemble grapes-like structure and are known as "staphylo." This particular bacterial species can spread from one organism to another and is known to impact mammalians [1]. Antibiotic treatment regimens are frequently used to treat the majority of staph infections.

Methicillin-resistant Staphylococcus aureus (MRSA) strains, also known as multidrug-resistant S. aureus, is a group of bacteria which is genetically distinct from other strains of *Staphylococcus aureus*. The MRSA is an antibiotic-resistant bacterium that can infect different regions of the body and was first identified as a cause of nosocomial infections in the 1960s. Since then, they have been linked to a number of rapidly progressing, potentially fatal diseases, including osteomyelitis, severe sepsis, toxic shock syndrome, necrotizing fasciitis, and endocarditis [2].

The MRSA bacteria is a gram-positive organism and is characterized by the size of their individual cocci measuring from 0.5 to 0.7 μm in diameter [3]. Its cell wall is mainly made of peptidoglycan, teichoic acid and other fibronectin binding proteins. The MRSA bacteria produces toxins and enzymes such as catalase, coagulase, clumping factor, hyaluronidase, β-lactamase that causes pathogenic factors [3]. The rapid evolution of this bacterium of new genetic linages and resistance to classes of antibodies makes this a public-health crisis. As of 2019, it has been recorded than as many as 80% of the *Staphylococcus aureus* infections are methicillin resistant [2].

The MRSA bacterium is classified into three categories, Healthcare-associated MRSA, Community associated MRSA and Livestock associated MRSA [4]. This is because MRSA is more commonly found in hospitals, prisons, and nursing homes, where people with open wounds, weak immune system and livestock, the HA-MRSA is where MRSA is found in hospital settings where patients are at high risk of the infection. CA-MRSA is the presence of MRSA infections in locations that are not related to healthcare settings such as prison, military and homeless people. MRSA is not limited to humans and is also found in livestock, which falls under LA-MRSA.

The resistance of the MRSA rose in the 1940s and is primarily mediated by the mecA gene and also the β-lactamase gene blaZ. This gene is located in the staphylococcal chromosomes and it encodes for the penicillin binding protein 2a (PBP-2a), the enzyme that is responsible for the crosslinking of the peptidoglycan in the bacterial cell wall [3]. The mecA gene enhances the virulence of the bacterium by causing it resistant to the methicillin antibiotic. Furthermore, the PBP-2a takes over the function of the penicillin binding proteins and has a low affinity for methicillin, allowing the bacterial cell to develop and thrive in the presence of antibiotics. Methicillin resistance is mediated by the mobile genetic element mecA, which is transferred horizontally and known as the staphylococcal cassette chromosome mec (SCCmec) [5].

In the current moment, MRSA can be treated using a glycopeptide antibiotic, vancomycin. However, the probability that the MRSA bacterium develops a resistance against this antibiotic is high. Therefore, in order to predict the phenotypic antibiotic resistance of the bacterium, Next Gen Diagnostics (NDS) bioinformatics tools can be utilised. The tools can be used to predict the antibiotic susceptibility of MRSA [6]. The purpose of this study is to utilise the bioinformatics tools to study the *Methicillin-resistant Staphylococcus aureus*. The main objective is to analyse hypothetical protein of MRSA using bioinformatics tools to identify protein targets.

2. Methodology

In order to identify immunogenic proteins as possible novel therapeutic targets, the entire proteome of *Methicillin-resistant Staphylococcus aureus* was analysed using the subtractive genomic technique. The overall workflow of this study is illustrated in Fig. 1.

2.1. Retrieval of MRSA Sequence

The bacterium species was searched in the NCBI database. The list from the database was accessed and the protein coding genes (CDS) of the DSM 20231 bacteria strain was selected. From the list of sequences, the hypothetical sequences were selected and filtered. The data was downloaded and was entered into Entrez for further filtering of the sequences. Once the list of sequences has been filtered, the FASTA format for all the sequence was downloaded.

2.2. Identification of paralogous sequences

The protein sequences were run using the CD-Hit suite in order to eliminate the paralogous sequences from *Methicillin-resistant Staphylococcus aureus* (MRSA) [7]. To reduce repetitive sequences using this tool, a cutoff score of 0.6 (60 percent sequence identity) was established. Proteins that are homologous and share more than 60% of their identity were not included in the list of protein sequences. Proteins with fewer than 100 amino acids were also left out of list/set. This set consists of the remaining proteins had non-homologous protein sequences.

2.3. Identification of protein sequences non-homologous to the proteome of human

The purpose of this stage was to avoid functional similarity with the human proteome in order to avoid drug binds to the host homologue proteins' active sites. Using BLASTp against the human (Homo sapiens) Refseq proteome in the Ensemble genome database, non-paralogous proteins of the pathogen were examined [8]. If any pathogen-specific proteins had a significant hit above the threshold of 10^{-4}, they were regarded as host homologous proteins.

2.4. Identification of essential non-homologous proteins

The Database of Essential Genes (DEG) was used to analyse the host non-homologous proteins of the pathogen indicated in set2 [9]. The DEG 15.2 server's entire collection of organism strains was chosen, and BLASTp was run using the parameters of a threshold value of 10^{-10} and a minimum bit score of 100. In this set of proteins, proteins that were identified and considered as essential proteins of the pathogen are proteins that have an expectation value of $\leq 10^{-4}$ and an identity of $\geq 25\%$ were listed. Products of the essential genes makes up a bacterium's minimum genome, which plays an important role in the field of synthetic biology [10].

2.5. Analysis of metabolic pathways

The KEGG server was used to compare the metabolic pathways of *Methicillin-resistant Staphylococcus aureus* to the metabolic processes of humans [11]. Kyoto Encyclopedia of Genes and Genomes, or KEGG, is a database that includes all of the metabolic pathways found in living things. The three-letter KEGG organism codes for *Methicillin-resistant Staphylococcus aureus* and H. sapiens (host) were used to extract all of the metabolic pathways from the KEGG PATHWAY Database [12]; where " sau", "sav", "saw", "sah", "saj", "sam", "sas", "sar", "sac", "sax", "saa", "sao", "sae", "sad", "suu", "sab " for all the *Staphylococcus aureus* and "hsa" for Homo sapiens. By contrasting the metabolic pathways of the host and the

pathogen, distinct metabolic pathways that are exclusively present in the pathogen were identified. The remaining pathogen pathways were categorised as common since they were already present in the metabolism of the host.

Essential host non-homologous proteins of *Methicillin-resistant Staphylococcus aureus* were identified and then subjected to BLASTp searches using the KEEG database's KAAS service [13]. Proteins implicated in widespread metabolic pathways were not included and were only identified as a set of proteins for further research. Pathogen-specific proteins were excluded. Metabolic proteins are indicated by KEGG Orthology (KO) assignments, whereas KAAS sever creates KEGG pathways for metabolic proteins. KO-unassigned proteins were identified as a separate set of proteins.

2.6. Prediction of subcellular localization

The *Methicillin-resistant Staphylococcus aureus* is a gram-positive bacterium. As a result, there are five possible subcellular locations from which to categorise the proteins of this organism. One is the cytoplasm, two are the inner membrane, three are the periplasm, four are the outer membrane, and five are the extracellular space. While membrane proteins can be thought of as prospective drug targets, cytoplasmic proteins can be considered as putative drug targets. The CELLO v.2.5 server [14] was used to forecast the subcellular localization of the shortlisted proteins.

As the MRSA bacterium can be considered as a potential drug target, the scores of the cytoplasm and the inner membrane subcellular locations were taken into consideration as they aid in the development of a new drug or in drug designing. Therefore, only the cytoplasm and inner membrane proteins were selected for further analysis.

2.7. 'Anti-target' analysis of essential, non-homologous and novel drug targets

Drugs or therapeutic substances that are intended to bind and limit the activity of a pathogen protein may dock with some key host proteins, having pharmacological effects on the host. These host proteins are known as "anti-targets." Ether-a-go-go related gene (hGER), P-glycoproteins (P-gp), the pregene X receptor (PXR), and constitutive androstane receptor (CAR) are examples of "anti-targets" in humans. Additionally, a few membrane receptors are listed as "anti-targets," including the muscarinic M1, serotonergic 5-HT2C, dopaminergic D2, and adrenergic 1a. A total of 181 anti-targets were reported in the literature with their accession number and can be found from the European Bioinformatics Institute (EBI). Novel drug targets were subjected to BLASTp analysis in the NCBI blast program against these human 'anti-targets', setting an E-value <0.005 and identity <25% as parameters. Proteins showing <25% identity were listed and proceeded to the other analysis.

2.8. Conservancy analysis of predicted sequences with other strains

For identifying the range of the therapeutic spectrum within the complete homologous bacterial community, conservation pattern of the projected sequences with other traditionally utilised strains is crucial [15]. The NCBI server's BLASTp programme was used to conduct a conservancy analysis of the projected drug target sequences. With the exception of *Methicillin-resistant Staphylococcus aureus* in the organism option, all of the settings in this case of running protein-protein BLAST were left at their default values.

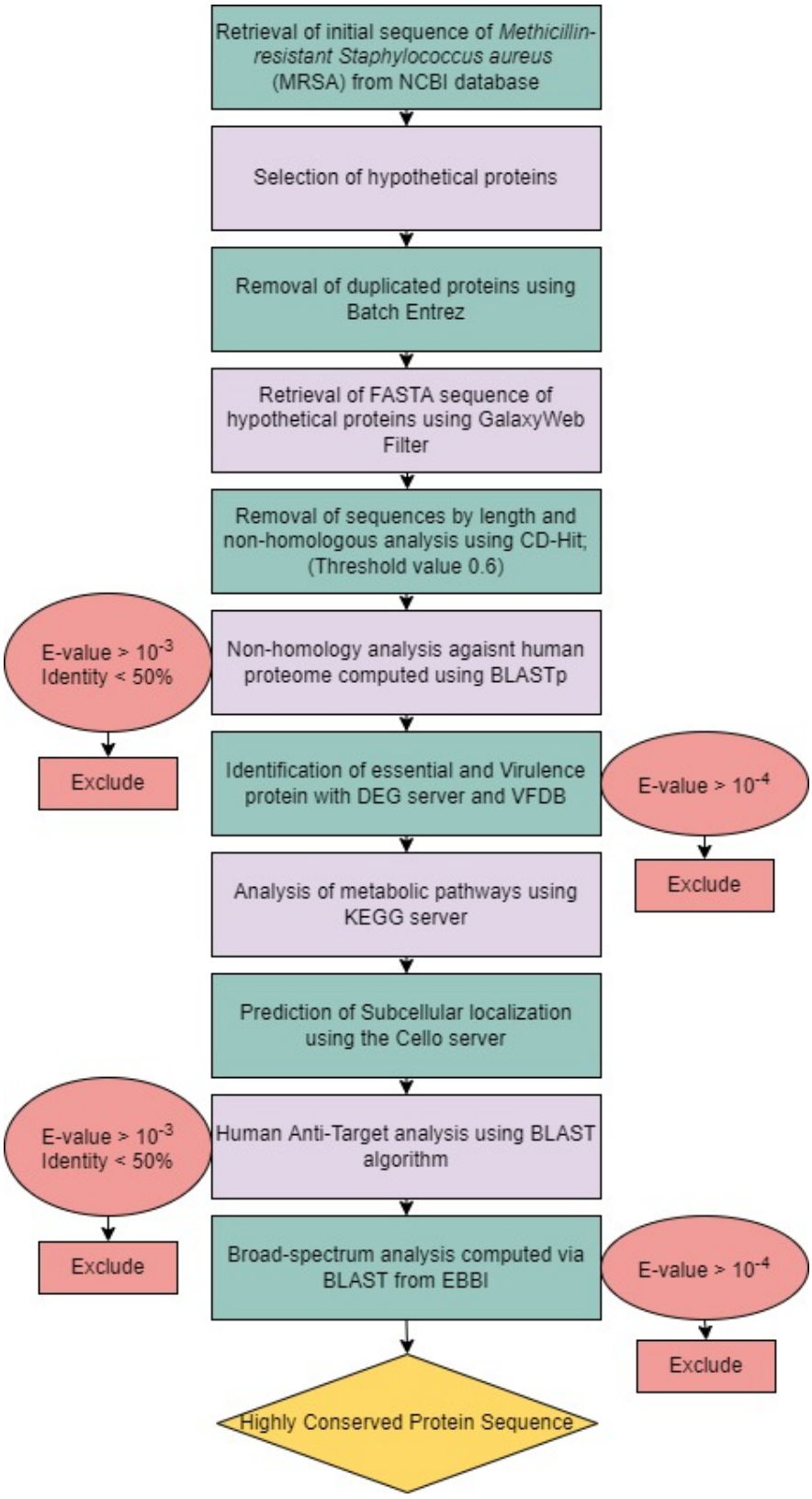

Figure 1 – Schematic representation of overall workflow employed in this study.

3. Results and discussion

The main goal of this study was to investigate potential new treatment targets for *Methicillin-resistant Staphylococcus aureus*. In order to accomplish this, a subtractive genomic method was used for the entire proteome by utilising several online databases and computational tools. The summary of the findings of the conducted study is presented in Table 1 below.

Table 1
Subtractive genomic analysis scheme towards the identification of novel therapeutic targets against *Methicillin-resistant Staphylococcus aureus*.

SI. No.	Subtractive Approaches	User Bioinformatics Server and Tools	Number of Proteins
1	The whole proteome of *Methicillin-resistant Staphylococcus aureus*	NCBI database	14594
2	Hypothetical sequences	Batch Entrez	317
3	Removed non-homologous (>60% identical) and smaller proteins (>100 amino acids)	CD-Hit suite	99
4	Proteins non-homologous to H. sapiens	BLASTp (E-value 10^{-3})	99
5	Essential and virulence of protein sequence	DEG Server & VFDB	8
6	Essential Proteins involved only in unique metabolic pathways	KAAS at KEGG	1
7	Proteins assigned KO (KEGG Orthology) but not in any pathway	KEGG	1
8	Essential membrane proteins	CELLO Server	1
9	Novel drug target proteins non-homologous to 'anti-targets'.	using BLASTp (E-value < 0.0001, Identity > 25%)	1
10	Broad-spectrum analysis	EBI (E-value 10^{-4})	1
11	Highly conserved proteins		1

3.1. Exclusion of homologous sequences

The proteome of contained *Methicillin-resistant Staphylococcus aureus* 14594 proteins. Among all the protein sequences, only the hypothetical sequences were selected for this antibody-resistant bacteria. The NCBI database was used to retrieve the hypothetical sequences from the MRSA bacteria strain, DSM 20231. A total of 317 hypothetical sequences were available. The genome records of the 317 protein sequences were retrieved using Batch Entrez. CD-Hit server was utilized to screen out homologous sequences of the pathogen. The server reported a total of 99 sequences out of the 317 hypothetical sequences. The homologous sequence >60% similarity were removed leaving the remaining 99 protein sequences as non-homologous. Furthermore, total proteins containing <100 amino acids were excluded from 99 non-homologous protein sequences assuming that the smaller proteins are unlikely to represent the essentiality.

3.2. Selection of human non-homologous proteins

Proteins involved in common cellular systems had emerged as homologous in course of time between bacteria and human [15–17]. Therefore, therapeutics developed and administered to bind target proteins of pathogens must avoid cross-reactivity with host homologous proteins. BLASTp was performed for the 99 proteins of the pathogen against Homo sapiens reference proteome. All the 99 proteins have showed significant hits with human proteins above the threshold. Considering as human homologous, no proteins were excluded and all the 99 sequences remained as non-homologous proteins.

3.3. Selection of essential proteins

Most antibacterial compounds are made to dock and inhibit essential gene products. Thus, the most potential therapeutic targets are thought to be essential proteins [18]. A total of only 8 proteins showed significant hits with the dataset of bacterial essential proteins deposited in the DEG 15.2 server. The remaining 91 sequences were not considered as essential proteins for the survival of the *Methicillin-resistant Staphylococcus aureus* bacteria. Essential gene products that are particular to an organism may be used as species-specific therapeutic targets [19].

3.4. Identification of pathogen specific pathways proteins

The 8 remaining protein sequences were retrieved from the KEGG database and they were compared. Among these 8 specific pathways, only one protein sequence was present in the pathogen. Therefore, it was termed as the pathogen-specific unique pathway. The proteins involved only in pathogen specific pathways are considered for screening potential novel therapeutic targets. All 8 essential proteins of the MRSA pathogen were not assigned with KO with specific metabolic pathways. However, only one sequence was assigned a KO assignment and is involved in unique pathways. This protein sequence is listed in Table 2 and was used to identify the novel drug targets in the study.

The 7 proteins among 8 sequences that were not assigned KO, infer that they are not involved in any metabolic pathway of the pathogen and the host.

Table 2
Proteins involved only in pathogen-specific unique pathways.

SI. No.	KO Assignment	Protein ID	Protein Name	Pathways
1	K04766	WP_001015996.1	Acetoin utilization protein AcuA	Metabolism

3.5. Analysis of subcellular localization

Due to the fact that many proteins can exist in numerous localizations, localization is one of the most crucial therapeutic target criteria [20]. Proteins that passed through the KEGG analysis were screened through CELLO v.2.5 server. Localization of the protein was finalized based on the prediction conducted by the CELLO server. Even though only one protein sequence remained from the previous step, it was still scanned for its subcellular localization. It was found from the CELLO server that the protein sequence was predicted as cytoplasmic proteins.

These cytoplasmic proteins are known to be putative drug targets and were proceeded to be tested for it [15]. The following are some benefits of using membrane proteins as therapeutic targets: 1) their functions can be determined without the need for in-vitro and in-vivo laboratory trials; and 2) secondary structure membrane proteins can be predicted and generated with ease because they have a distinctive structure and exhibit a propensity to form secondary structure [21]. If there is no experimentally confirmed knowledge of high resolution three dimensional (3D) structure, structure-based therapies development will be feasible in the near future [22] through the ab initio modelling method [21,23].

3.6. Anti-Targets analysis of novel drug targets

Cross-reactivity and carcinogenic screening are essential for creating an effective therapeutic molecule because several drug candidates were pulled from the market due to displaying carcinogenic effects [24–26]. Although this pathogen's non-homologous host proteins were cut from non-homologous sequences, this action was taken to prevent lethal side effects from accidental therapeutic interactions with host "anti-targets." Because there was no degree of similarity with "anti-target" proteins, the protein sequence was considered to be a host non-anti-targets.

3.7. Conservancy analysis of predicted sequences with other strains

The remaining one protein sequence had undergone the broad-spectrum analysis and resulted with a total of 23 homologous alignments with an E-value of 0.0001 and alignment length cut-off of 1%. The Protein-protein BLASTp had revealed the conservancy pattern of Acetoin utilization protein (A6QHR7_STAAE), Putative uncharacterized protein (Q2G293_STAA8), Acetoin utilization protein AcuA (Q2FG04_STAA3) and Acetoin utilization protein AcuA (Q5HF40_STAAC) from the *Methicillin-resistant Staphylococcus aureus*. When compared to similar proteins from other commonly used *Methicillin-resistant Staphylococcus aureus*, all four of these proteins displayed 100% conservation. For efficient therapeutic and drug targets against a particular bacterial disease, the predicted protein must always be conserved in a large variety of bacterial strains [15,27]. They are thus a potent and versatile therapeutic target for *Methicillin-resistant Staphylococcus aureus*.

4. Conclusion

To summarise the following information above, it can be concluded that one protein from *Methicillin-resistant Staphylococcus aureus* was identified in our work using the investigated subtractive genomics technique as a potential therapeutic target. Since this cytoplasmic protein participates in the metabolic pathway unique to the *Staphylococcus aureus* pathogen, inhibiting it will aid in the fight against infectious illnesses. Furthermore, as they displayed no similarity against the human proteome and "anti-targets," the likelihood of drug and human protein cross-reactivity has been diminished. Therefore, the creation of new drugs or therapeutic substances may mark a promising first step in the prevention of *Methicillin-resistant Staphylococcus aureus*, an antibiotic-resistant bacteria, from spreading disease throughout the world.

References

1. Garoy, E. Y., Gebreab, Y. B., Achila, O. O., Tekeste, D. G., Kesete, R., Ghirmay, R., Kiflay, R., & Tesfu, T. (2019). Methicillin-Resistant *Staphylococcus aureus* (MRSA): Prevalence and Antimicrobial Sensitivity Pattern among Patients—A Multicenter Study in Asmara, Eritrea. *Canadian Journal of Infectious Diseases and Medical Microbiology*, *2019*, 1–9. https://doi.org/10.1155/2019/8321834
2. Grema, H. A. (2015). Methicillin Resistant Staphylococcus aureus (MRSA): A Review. *Advances in Animal and Veterinary Sciences*, *3*(2), 79–98. https://doi.org/10.14737/journal.aavs/2015/3.2.79.98
3. Haddadin, A. S. (2002). Methicillin resistant Staphylococcus aureus (MRSA) in the intensive care unit. *Postgraduate Medical Journal*, *78*(921), 385–392. https://doi.org/10.1136/pmj.78.921.385
4. Kumar, N., Raven, K. E., Blane, B., Leek, D., Brown, N. M., Bragin, E., Rhodes, P. A., Parkhill, J., & Peacock, S. J. (2020). Evaluation of a fully automated bioinformatics tool to predict antibiotic resistance from MRSA genomes. *Journal of Antimicrobial Chemotherapy*, *75*(5), 1117–1122. https://doi.org/10.1093/jac/dkz570
5. Mandal, A., MD. (2019, June 5). *What is Staphylococcus Aureus?* News-Medical.Net. Retrieved August 11, 2022, from https://www.news-medical.net/health/What-is-Staphylococcus-Aureus.aspx
6. Turner, N. A., Sharma-Kuinkel, B. K., Maskarinec, S. A., Eichenberger, E. M., Shah, P. P., Carugati, M., Holland, T. L., & Fowler, V. G. (2019). Methicillin-resistant Staphylococcus aureus: an overview of basic and clinical research. *Nature Reviews Microbiology*, *17*(4), 203–218. https://doi.org/10.1038/s41579-018-0147-4
7. Huang Ying, BeifangNiu Ying Gao, Fu Limin, Li Weizhong. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26 (5):680–2
8. Zerbino Daniel R, Achuthan Premanand, Akanni Wasiu, RidwanAmode M, Barrell Daniel, Bhai Jyothish, Billis Konstantinos, et al. Ensembl 2018. Nucleic Acids Res 2017;46(D1):D754–61.
9. Luo Hao, Lin Yan, Gao Feng, Zhang Chun-Ting, Zhang Ren. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res 2013;42(D1):D574–80.
10. Kanehisa Minoru, Goto Susumu. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27–30.
11. Paduano Francesco, Forbes Angus Graeme. Extended LineSets: a visualization technique for the interactive inspection of biological pathways. BMC Proc 2015;9 (6):S4. BioMed Central.
12. Moriya Yuki, Itoh Masumi, Okuda Shujiro, Yoshizawa Akiyasu C, Kanehisa Minoru. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 2007;35(suppl_2):W182–5.
13. Damte Dereje, Suh Joo-Won, Lee Seung-Jin, Yohannes Sileshi Belew, Hossain MdAkil, Park Seung-Chun. Putative drug and vaccine target protein identification using comparative genomic analysis of KEGG annotated metabolic pathways of Mycoplasma hyopneumoniae. Genomics 2013;102(1):47–56.
14. Yu Chin-Sheng, Chen Yu-Ching, Lu Chih-Hao, Hwang Jenn-Kang. Prediction of protein subcellular localization. Proteins: Struct Funct Bioinf 2006;64(3):643–51.
15. Khan, M. T., Mahmud, A., Iqbal, A., Hoque, S. F., & Hasan, M. (2020). Subtractive genomics approach towards the identification of novel therapeutic targets against human Bartonella bacilliformis. *Informatics in Medicine Unlocked*, *20*, 100385. https://doi.org/10.1016/j.imu.2020.100385
16. Hediger Matthias A, Turk Eric, Wright Ernest M. Homology of the human intestinal Naþ/glucose and Escherichia coli Naþ/proline cotransporters. Proc Natl Acad Sci Unit States Am 1989;86(15):5748–52.