Review Article

Open Access

In silico approach for mining of potential drug targets from hypothetical proteins of bacterial proteome

Abstract

An increase in expansion of antibiotic-resistant bacterial pathogens alarms the world's population and creating a wave of the antibiotic apocalypse. The inclination of the death rate due to these antibiotic-resistant superbugs signifies urgency towards a new drug discovery to combat against these bacterial pathogens. The last class of antibiotics developed leaves a huge gap in the antibiotic timeline as the antibiotic development progress failed to kill the bacteria. Current antibiotic targets the central dogma of the bacteria hence finding a new potential drug target could eliminate the superbugs. It is, therefore, crucial to understand the underlying mechanism to identify the root cause of the resistant characteristic by understanding the biological cellular processes. Hypothetical proteins are an uncharacterized protein that is not known for its function which could provide a deeper understanding of the metabolic pathway of the bacterial proteome. This paper will generally provide a guideline for non-bioinformatician to mine potential drug targets from hypothetical proteins of bacterial proteome using a fast and less-cost bioinformatics approach.

Keywords: hypothetical proteins, essential genes, virulent proteins, drug target, subtractive genomic analysis

Volume 4 Issue 4 - 2019

Umairah Natasya Mohd Omeershffudin, Suresh Kumar

Department of Diagnostic and Allied Health Science, Faculty of Health & Life Sciences, Management & Science University, Malaysia

Correspondence: Suresh Kumar, PhD, Department of Diagnostic and Allied Health Science, Faculty of Health & Life Sciences, Management & Science University, 40100 Shah Alam, Selangor, Malaysia, Tel +60-14-2734893, Email sureshkumar@msu.edu.my

Received: August 08, 2019 | **Published:** August 26, 2019

Introduction

Protein consists of four levels structure which is primary, secondary, tertiary and quaternary structure. The structures are built based on the linking amino acids. For sequences of amino acids of lesser than fifty amino acids linked together are known as peptides and are usually the primary structure of the protein.[1] Hence, the primary structure is described as linear chain blocks of amino acids. If the amino acids are more than fifty, it is described as polypeptides. Secondary structures of the protein are known as the "folding protein" where the polypeptides are either folded to α-helices, β-strands and random coil.[2] The structures are folded by linking the C=O and N-H by hydrogen bonds that make the structures more stable.[3] Tertiary structures are known as the whole structure of the protein in a 3-Dimensional shape (3D) of which the protein structures are folded. Quaternary structures are referred to as the spatial relationship or the interaction between the subunits of the protein or also known as the individuals' polypeptide chain.

Protein structure is often studied to understand the contributing factors of diseases, genetics ailment or nutritional composition. In most studies, hypothetical proteins (HPs) often gain notoriety as it is often used in structural genomics with unknown functions as there is a lack of evidence in the *in vivo* experimental setting. The sequence of proteins is classified as hypothetical if the sequence search is not recognizable to the protein that has been functionally characterized.[4] According to Normi,[5] the hypothetical protein is known as the orphan protein however it has been seen that this HPs have a high potential of carrying metal-to protein trafficking and confers antibiotic-resistant.

In 2006, one of the protein databases, NCBI, contained about 19, 85,480 of protein sequence out of which 1/3 of the sequences are hypothetical proteins and 1/10 of the protein sequences are classified as "conserved hypothetical" of which the function is annotated.[6] This approach is called a protein function annotation of which the protein function is predicted by using a bioinformatics approach. Previously, HPs are seen to not be able to be studied thoroughly due to the lack of evidence in the sequence similarity in all the protein databases making it difficult to understand the protein function.[6] However, through the in-silico approach of computational aided drug design, functional annotation of HPs is fully utilized to identify the protein function. HPs are important as these proteins carry excessive involvement in cellular activities and also signalling pathways.[7] The proteins are revealed to contribute its crucial role in various microorganisms. Hence making it as a potential novel drug target for antibiotic-resistant, oncogenic studies, and other automated immune disorders disease by understanding the metabolic pathway presented in HPs. The function predicted proteins are important as it provides the functional characterization of protein sequence for uncharacterized Open Reading Frames (ORFs).[7]

Currently, the world population is teeming with bacteria that becomes a threat to the global. These bacterial pathogens acquire genetic traits that develop an intrinsic mechanism of resistance towards antibiotics. Ongoing research towards developing new antibiotics consumes a lot of cost and time consuming due to the poor experimental outcome. Despite the ongoing research, these antibiotic-resistant superbugs are currently on the rise that is detrimental to the human population which leads to alarming waves of the emergence of antibiotic-resistant pathogens. The root cause of genetic acquirement of resistance towards antibiotics remains enigmatic. However, it can be studied by identifying the HPs that is uncharacterized. The current treatment regimen of antibiotics is unable to combat the bacterial pathogens as most of which target the central dogma of the bacterial pathogens. However, targeting the specific mechanism that

Submit Manuscript | http://medcraveonline.com

©2019 Omeershffudin et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and build upon your work non-commercially.

Copyright: ©2019 Omeershffudin et al.

develops the resistant traits could potentially be detrimental to the bacterial pathogens. Uncharacterized proteins could provide answers to elucidate the biological of the bacterial pathogens by integrating bioinformatics approach of the in silico experimental study design through functional annotation. Hence, this paper aims to provide a guideline for mining potential targets of bacterial proteome for the researcher who is not familiar with bioinformatics tools.

Mining of hypothetical protein from database resources of protein sequence

Protein databases are a crucial part of modern biological studies. The open server provides a sufficient amount of protein database that is often used in most biological studies. The major primary protein databases that are often used are UniProt, and Entrez (NCBI). The whole-genome sequence that is completed through sequencing projects for bacteria can be retrieved via the NCBI database where it provides microbial strains.

Entrez

The Entrez gene is a database that provides a gene-specific database that is stored in the National Centre for Biotechnology Information (NCBI). Entrez search engine is built in the NCBI that can be accessed via https://www.ncbi.nlm.nih.gov/. The search engine can be accessible to use by providing the keywords of the microorganism. The database provides gene-specific information related to the bacteria which user can easily retrieve related publications, genome id, genome sequence and etc. The database is encoded with a specific identifier notified as to the GeneID.[8] Entrez is often used as the primary resource to identify bacterial genome species.

NCBI microbial genomes

NCBI Microbial Genomes are embedded to the NCBI database that serves as an open-source that provides resources of the database for a microbial genome that is obtained from sequencing projects. The database includes prokaryotes, eukaryotes, viruses, plasmid, and organelles completed assembly genome. Detailed information related to the protein of the bacterial pathogen is available. During the sequencing projects, any uncharacterized proteins are annotated as "hypothetical". The NCBI Microbial genome can easily be accessed via https://www.ncbi.nlm.nih.gov/genome/microbes/.

Table 1 shows the number of hypothetical proteins of 10 strain-specific most threatening bacterial pathogens. This hypothetical protein of the bacterial pathogens could be further explored to understand the attenuation of virulence factors of the bacterial pathogens. Although hypothetical proteins are progressively being studied, there are still many uncharacterized hypothetical proteins that remain not known.

Table 1 Table shows the number of hypothetical protein in the total proteome of the strain-specific microbial genome that is retrieved from the NCBI microbial genome

Bacterial pathogens	Hypothetical proteins/total proteome
Acinetobacter baumanmii (AB030)	1018 / 3953
Pseudomonas aeruginosa (PA01)	2255/5572
Neisseria gonorrhoeae (FA1090)	413/1886
Staphylococcus aureus (MRSA) (NCT 8325)	1510/2767
Burkholderia cepacia	1146/7614
Streptococcus pyogenes (M1 GA5)	652/1693
Klebsiella pneumoniae (HS11286)	1698/5779
Escherichia coli (IAI 39)	1069/4725
Mycobacterium tuberculosis (H37v)	1055/3906
Clostridium difficile (630)	712/3766

Uniprot

Uniprot also is known as the UniProt knowledgebase are the centerpiece of protein sequences that provides the annotation and functional information. The UniProt comprised of other protein databases which include the SwissProt, TrEMBL PIR-PSD. UniProt is divided into two main sections which contain the fully manually annotated that is obtained from the literature review and also the curator-evaluated which is computationally analyzed.[9] The database is embedded with features to perform sequence alignment, retrieve/ID mapping and peptide search. The protein is tagged with the uniprot ID as reference.

Subtractive genomic analysis

According to Barh,[10] the term subtractive genomic is whereby two sets of genomes is removed or subtract from the genomic data in order to obtain the unique phenotype of the specific genes. Subtractive genomic analysis is a technique of which the hypothetical protein is analyzed by using in-silico approaches. The technique analyses the potential target of hypothetical protein by using various databases and available tools of bioinformatics. It is whereby a technique to identify the essential genes for the survival of the pathogens by identifying the non-homologous sequence that is absent from the human host.[11] The detailed workflow of subtractive genomic analysis is described in Figure 1.

Citation: Omeershffudin UNM, Suresh Kumar. In silico approach for mining of potential drug targets from hypothetical proteins of bacterial proteome. *Int J Mol Biol Open Access.* 2019;4(4):145-152. DOI: 10.15406/ijmboa.2019.04.00110

Copyright: ©2019 Omeershffudin et al.

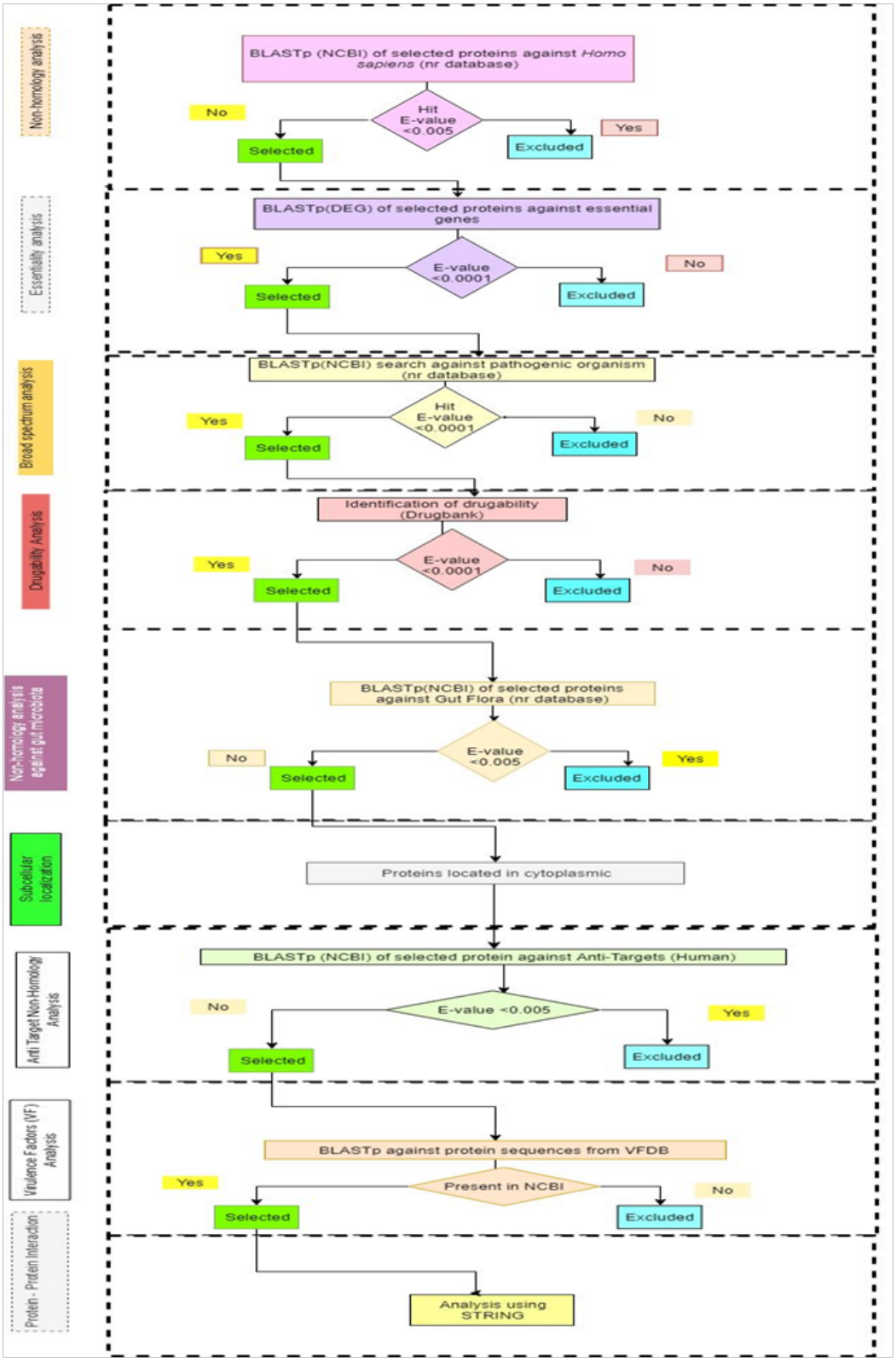

Figure 1 General workflow of subtractive genomic approach to mine drug targets from hypothetical proteins.

Citation: Omeershffudin UNM, Suresh Kumar. In silico approach for mining of potential drug targets from hypothetical proteins of bacterial proteome. *Int J Mol Biol Open Access.* 2019;4(4):145-152. DOI: 10.15406/ijmboa.2019.04.00110

Copyright: ©2019 Omeershffudin et al.

Blast

BLAST is abbreviated from the Basic Local Alignment Search Tool[12] that is crucial in subtractive genomic analysis. The subtractive genomic analysis is mostly screened by using BLAST. The main function of BLAST is to identify regions that show a similarity between two or more biological sequence. The input can either be nucleotide or protein. BLAST is built into the NCBI database and Uniprot database. Further information about BLAST can be accessed via https://blast.ncbi.nlm.nih.gov/Blast.cgi where it provides manual and tutorial to perform a local alignment. The web BLAST performs a search against nucleotide that is tagged as BLASTn, BLASTp for blast against protein sequence, BLASTx for translated nucleotide against protein and tBLASTn for protein against translated nucleotide.

Non-homology analysis

The homologous sequence is a sequence that is similar in terms of ancestry. The homologous sequence protein is inferred based on their sequence similarity of the DNA protein. This homologous protein might carry any undesirable cross-reactivity and will interfere with the binding between the active sites.[13] This non-homology analysis is the primary steps for every subtractive genomic analysis. The analysis is subjected to BLASTp against *Homo sapiens* proteome with an expected threshold value of 0.005.[14]

Essentiality analysis

Essential genes are a principal drug target criterion to identify the potential bacterial proteins for an identifying therapeutics drug target. The essential genes are genes that are necessary to support the cellular life of living microorganisms.[15] The identification of essential proteins is identified by performing a BLASTp search against the Database of Essential Genes (DEG) with an e-value of 0.0001.

Druggability analysis

The druggable analysis is one of the important steps of which the protein was analyzed for their ability as a potential drug target. Druggability is one of the most important features for potential drug target protein as it shows whether the protein is susceptible towards binding with small inhibitor molecules. The HPs will be screened against Drugbank database to identify the protein that can be potentially developed as a drug target. The database is known as both bioinformatics and cheminformatics primary resource that merges both comprehensive information on drugs.[16] Drugbank is extensively used in silico studies as a tool for drug discovery. In subtractive genomic approach, hypothetical proteins that can be found in drug bank are selected for further analysis.

Anti-target non-homology analysis

Anti-target is protein receptor or host protein that when it binds with the drug molecules it may react causing and adverse effects which can be toxic to the human cells. The anti-target protein will also lead to severe pharmacokinetics effects. The anti-targets are considered as the following constituents the human ether-a-go-related gene (hERG), the pregnane X, constitutive androstane receptors (PXR and CAR, respectively) and also the P-glycoprotein (P-up).[17] Hence, it is crucial to remove anti-targets protein. This analysis is performed by BLASTp with an e-value of 0.005.[18]

Gut-flora non-homology analysis

The microbiota is described as the entire population of microorganisms which includes bacteria, fungi, archaea, viruses, and protozoans.[19] The gut microbiota plays a significant role in the association of a broad array of diseases related to the bowel systems.[20] Evidently, antibiotics could result in short or long term implications towards the ecological system of the normal gut microbiota.[20] In addition to that, the gut microbiota plays an important role in preventing the colonization of bacteria pathogens.[21] Therefore, it is important to eliminate proteins that show similarities to the gut flora as any unintentional disruption of gut microbiota can lead to severe effects. In order to avoid such circumstances, the proteins of the bacterial pathogens are subjected to BLASTp with an e-value of 0.0001.[22]

Subcellular localization

Subcellular localization is one of the main criteria of potential bacterial drug target. The bacterial proteins can be presented in 5 feasible regions of subcellular which is the cytoplasm, plasma membrane, periplasm, outer membrane and in the extracellular membrane.[14] Each bacterium located at a different subcellular location can provide different functionality. The subcellular localization is performed by using PsortB.[23] The PsortB predicts the bacterial proteome of gram-negative based on 3 localizations type; cytoplasmic, membrane and extracellular proteins.[24] CELLO2GO[25] is a also a protein subcellular localization tools provided that it is more accurate Protein located at the cytoplasmic and the membrane of the protein channels are targeted for drug target whereas protein located at the membrane, exo-membrane and also proteins that are secreted is developed for potential peptide vaccines.[10]

Virulence factor analysis

Virulence factors (VFs) are one important determinant to identify drug target. VFs are secreted by the bacterial pathogens which makes them causing disease. The database provides information that determines the related factors of virulent of the bacterial pathogens. These factors can include adhesion, colonization, the capsule of the bacterial pathogens, exoenzyme and exotoxins that involves the bacterial attachment and the host cells.[26] The analysis can be performed by finding similarities against the Virulence Factor Database (VFDB).[27]

Broad-spectrum analysis

The broad-spectrum analysis is one of the important criteria for multiple infections caused by a broad spectrum of bacterial pathogens. The step is to analyze the bacterial pathogens in broad-spectrum proteins target which can be beneficial for developing drug target for antimicrobial-resistant pathogens.[10] The bacterial proteins are subjected to BLASTp with e-value of 0.005 against a wide-range of bacterial proteomes.[14]

Functional annotation

Functional annotation is a process of which the information related to a certain gene biological information is gathered based on its aliases, molecular function, its biological roles, protein subcellular localization and the domain expression.[28] Functional annotation

Copyright: ©2019 Omeershffudin et al.

is performed to define the protein functionality of "conserved hypothetical" proteins or protein with undefined function. These approaches include the identification of the similarities between sequences, profiling of phylogenetic trees, the analysis of protein-protein interactions, defining the protein complexes and also profiling of gene expression.[6]

Interproscan

The tools provide a vast integration of protein analysis that combines different protein signature.[29] Interproscan can be used to analyze the protein functionality that can be beneficial to understand the underlying mechanism that contributes to the survival of bacterial pathogens. The InterPro is the integration of major protein resource databases which includes PROSITE, PRINTS, SMART, Pfam PIRSF and SUPERFAMILY.[29]

Pfam

Pfam is one of the databases that consist of large protein families databases and domains. The database contains approximately 12,000 families that are utilized in an experimental biological setting, computational sequence organization and evolution of proteins origin.[30] Pfam is one for the Interpro consortium that provides protection functionality annotation, the construction of structural data, non-domain annotation of protein, active site protein analysis, the architectural evolution of domain and taxonomy.[30]

Gene ontology & pathway analysis

KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) is known as an open-source database that contains 16 major databases that cover a broad categorization of systematic information of the genome and chemical information. KEGG database can be used to identify the pathway analysis of the hypothetical protein. Pathway analysis can be divided into 3 which is metabolic pathways, gene regulatory and signaling pathway.[31] In silico approach observes the metabolic pathway that is presented between the hypothetical protein of the bacteria and the host.

Gene Ontology (GO)

According to Neuhaus[32] the term ontology is defined as an obvious parameter of a shared conceptualization. GO database is curated organism database that provides the gene specification of the bacterial proteomes. The database provides annotation that is supported by experimental evidence and is termed by GO codes. It provides detailed reports of correlation of the products of the genes with the biological types. Gene ontology helps to characterize the specific genomic function of the hypothetical protein of the bacterial pathogens which is important in drug-target based method.

Protein_protein Interaction Analysis-STRING

Bacterial proteomes may carry some specific functions influenced by the neighboring protein which can lead to a certain function. STRING is an open server site that provides a unique scoring-framework of interactome analysis by analyzing the protein-protein interaction based on the physical of the functional interactions of the bacterial proteome.[33] In order to understand the responsible interactors towards the survival of the bacterial proteomes, only high confidence score greater than 0.700 is included in the protein network frame. The protein frameworks are derived from various experimental data, analysis of gene; the gene fusion neighborhood, co-occurrence, coexpression that is curated from various pathway databases.[33] Protein interactions scores might result in either false positives or negatives, hence low confidence interactors are pruned out from the network framework.[14]

Homology modeling

Hypothetical protein does not have a 3D structure model that is available on Protein Data Bank(PDB) due to the unknown characterization of the conserved regions. Hence, homology is performed to structuring 3D model. Hypothetical proteins that show 40% identical similarities scoring to PDB are most likely to use SWISS-MODEL, RaptorX, and Modeller to perform homology modeling. Lesser than 40% similarities with PDB are more compatible to perform homology modeling by using Phyre2. The concept of homology modeling is to build the 3D model of the proteins by referring to the templates of related family members. The homology model web server is observed under the Continuous Automated Model EvaluatiOn (CAMEO) project to analyze its accuracy.

Swiss-model

SWISS-MODEL relies on building 3D models of protein structures by evolving protein structures based on templates of the particular protein families of which the structures are readily solved in the experimental setting by X-ray or Nuclear Magnetic Resonance(NMR).[34] SWISS-Model is readily built with high sensitive sequence-based template identification by using the Hidden Markovnikov Model(HMM) averse to the SWISS-MODEL template library (SMTL).[34] The model quality built by using SWISS-MODEL is accessed by the composite score of QMEAN.[34] In order to build a homology model at SWISS-MODEL interface, the input only requires a FASTA sequence of the bacterial proteome.

RaptorX

RaptorX is another web server that can be used to build the protein homology model of the bacterial proteome. Unlike SWISS-MODEL, RaptorX runs the prediction of the structure-property without any templates reference.[35] The web server predicts protein secondary structure, tertiary structure base template, and sampling probability outcome for alignment. It consists of 3 main constituent that includes, threading of single-template, prediction for alignment quality and threading of multiple-template.[36]

Phyre2

Phyre2 web portal provides a prediction for protein mode and analysis where it includes building 32 models, ligand binding site prediction and also analyzing the amino acid variants for protein query by users.[37] The server predicts the secondary and tertiary structure of the protein and accesses the model quality of the protein. Phyre2 uses 5 main facilities for the 3D structure prediction which includes; searching structure averse to various genomes by using backphyre, submitting a larger number of the protein sequence by batch, threading technique of one-to-one for user sequence to user structure, scanning proteins that are difficult through weekly basis by using Phyrealarm and a thorough analysis for the quality of the model, protein functionally and the mutational effects by Phyre Investigator.[37]

Citation: Omeershffudin UNM, Suresh Kumar. In silico approach for mining of potential drug targets from hypothetical proteins of bacterial proteome. *Int J Mol Biol Open Access.* 2019;4(4):145-152. DOI: 10.15406/ijmboa.2019.04.00110

Copyright: ©2019 Omeershffudin et al.

Validation of protein model

Homology model build is subjected to the validation of the 3D protein model. Few benchmarks what is readily available for the structure validation assessment are Root Median Square Deviation(RMSD), Ramachandran plot assessment and ERRAT. The validation of the protein model is done to analyze the protein structure model whether it is programmed correctly and if the implementation of the algorithm is applied.[38] Generally, the structure validations observe non allowed and allowed conformations.[39] Most structures validation aims to oversee the resolution and the R-value as higher resolution shows a higher accuracy of the structures of molecules.[40]

ProSA-web

ProSA or Protein Structure Analysis is one of the tools that consist of a wide base of the user and often applied for analysis and validation of predicted protein structures and model. ProSA specifies on the analysis of the X-ray and NMR spectroscopy.[41] The server recognizes the error of the protein structure. HPs structures may be interfered by machine artifacts throughout the homology process which leads to error in the structures. To identify the regions to facilitate the interpretation process. This is a crucial step for structure depositors before the homology model of the 3D structure is submitted to PDB.

Procheck

Procheck perform protein structure validation through the Ramachandran plot. Ramachandran plot access the structure of the quality of the stereochemical of the protein model. The analysis is performed to identify the residues of the protein structures of which within the favored region and disallowed regions.[40] The highlighted regions could either be an error or be further analyzed. The residue is listed based on the parameters of the stereochemical. The good quality model should demonstrate a total of over 90% score for residues in the core and allowed regions.[42]

Errat

Some errors of protein structures resulted in errors causing higher randomized distributions of the atom; carbon (C), nitrogen (N), and oxygen (0). The ERRAT plot is used to validate the protein structure by outputting the "overall quality factor" of the atomic interactions that are non-bonded. This step is to validify the protein model that is built by using homology modeler and high-quality model are based on the higher score.[43]

Ligand preparation

Ligand compound can be resourced by using available chemical databases which provide the information of the chemical compound of the ligand, 2D structure and SMILES formatted of the chemical compound. Some of the chemical curated databases that are used for chemical ligand preparation is PubCHEM[44] and ZINC.[45] Both are often used for ligand preparation however, ZINC is available for commercial purpose.

Binding site prediction

Protein is not independent whereby their function are expressed upon interacting with other molecules. In functional annotation, it is crucial to understand the protein-ligand interaction as it plays a vital role in drug discovery.[46] The binding site prediction identifies the relationship of the protein-ligand based interaction which is divided into two methods; geometry and energy-based method.[46] Binding site prediction software works to detect a site of which the site has the highest potential to induce the binding interaction with other molecules. Some predictors provide data such as the binding pocket.

Virtual screening

Virtual screening (VS) is one of the most recent progressive studies of drug discovery which applies computational methods. Generally, the concept of virtual screening is to identify potential drug target against a large scale of libraries of chemicals and understand the underlying mechanism of whether the molecules are able to dock or bind with the target molecules. VS are included as the structure-based drug design(SBDD) which are seen as highly efficient towards the identification of drug in the pharmaceutical field.[47]

Docking

Molecular docking is currently one of the most progressive computational methods in aided drug design. Docking can be divided into ligand-protein interaction or protein-protein interaction. In ligand-protein interaction it follows the concept of lock-and-key to visualize whether the target protein is amenable to small molecule inhibition; ligand. The method outputs binding affinity score (Kcal/mol) where the least binding affinity scores show that the inhibition consumes the least energy to bind (Kcal/mol). Docking can be performed when the 3D structure of the protein is known. This process provides information on how the protein structure or macromolecule interacts with small molecules.[48] There are currently few available docking tools that can freely be used such as AutoDock, AutoDock Vina and etc.

ADME toxicity test

Absorption, distribution, metabolism, elimination, and toxicity (ADMET)[49] is one of the important aspects of identifying the potential drug. Initially, drugs were identified through *in vivo* screening where it takes years and time-consuming. As the drug development field is progressing, more promising compounds are screened for further pharmacokinetic properties and whether the compound could induce cell toxicity to the human body. ADMET test would underline the drug properties of the compound as drug interactions are based on the properties of the compound that it carries. In some cases of which a small portion and flexible scaffold against random receptors will have a tendency to change the molecular shapes and also the conformations which may lead to severe effects or cell toxicity. In recent years, more *in silico* studies towards analyzing the ADMET property have been integrated for rapid analysis. Some available open server includes ADME- Tox[50] and preADMET.[51]

Conclusion

The spread of antibiotic resistant bacterial pathogens kept on increasing causing a huge cost to discover new antibiotics. Most research is based on laboratory approaches where it can be time-consuming and unable to obtain the result while conducting the experimental procedures. Hence, a low-cost and highly accurate method could help to overcome this. It is suggested that both in vitro and in silico technique could be performed to acquire more accurate result. Practicing in silico approach can help to minimize the cost, time and increase accuracy.

Citation: Omeershffudin UNM, Suresh Kumar. In silico approach for mining of potential drug targets from hypothetical proteins of bacterial proteome. *Int J Mol Biol Open Access.* 2019;4(4):145-152. DOI: 10.15406/ijmboa.2019.04.00110

Copyright: ©2019 Omeershffudin et al.

Acknowledgments

The authors acknowledge Department of Diagnostic and Allied Health Sciences, Faculty of Health and Life Sciences, Management & Science University, Shah Alam, Selangor Darul Ehsan, Malaysia for providing necessary infrastructure facility to carry out this research.

Conflicts of interest

The author declares that there are no conflicts of interest.

Funding

None.

References

1. Particle Sciences Drug Development Services. Protein Structure: Technical Brief; 2009.
2. Jez JM. Revisiting protein structure, function, and evolution in the genomic era. *J Invertebr Pathol*. 2016;142:11–15.
3. Jessica May, David S Goodsell, Laskowski RA, et al. Protein Structure. *Endocrine*. 2016;5:546–547.
4. Watson JD, Sanderson S, Ezersky A, et al. Towards Fully Automated Structure-Based Function Prediction In Structural Genomics: A Case Study. *J Mol Biol*. 2007;367(5):1511–1522.
5. Normi MY. Hypothetical proteins: Can they be the next drug targets? *3rd International Conference on Integrative Biology.* 2015;4(2):6577.
6. Sivashankari S, Shanmughavel P. Functional annotation of hypothetical proteins-A review. *Bioinformation*. 2006;1(8):335–8.
7. Naveed Muhammad, Zoma Chaudary, Zeeshan Ali, et al. Annotation and curation of hypothetical proteins: prioritizing targets for experimental study. *Adv Life Sci*. 2018;5(3):73–87.
8. Maglott D, Ostell J, Pruitt KD, et al. Entrez Gene : gene-centered information at NCBI. *Nucleic Acids Res.* 2005;33:54–58.
9. Apweiler R, Bairoch A, Wu CH, et al. UniProt : the Universal Protein knowledgebase. *Nucleic Acids Res.* 2004;32:115–119.
10. Barh D, Tiwari S, Jain N, et al. *In Silico* Subtractive Genomics for Target Identification in Human Bacterial Pathogens. *Drug Dev Res.* 2011;72(2):162–177.
11. Barh D, Jain N, Parida BP, et al. A Novel Comparative Genomics Analysis for Common Drug and Vaccine Targets in Corynebacterium pseudotuberculosis and other CMN Group of Human Pathogens. *Chem Biol Drug Des.* 2011;78(1):73–84.
12. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–410.
13. Sarkar M, Maganti L, Ghoshal N, et al. *In silico* quest for putative drug targets in Helicobacter pylori HPAG1: molecular modeling of candidate enzymes from lipopolysaccharide biosynthesis pathway. *J Mol Model.* 2012;18(5):1855–1866.
14. Shanmugham B, Pan A. Identification and Characterization of Potential Therapeutic Candidates in Emerging Human Pathogen Mycobacterium abscessus: A Novel Hierarchical *In Silico* Approach. *PLoS One.* 2013;8(3):59126.
15. Zhang R, Ou H, Zhang C. DEG: a database of essential genes. *Nucleic Acids Res.* 2004;32:271–272.
16. Wishart DS, Knox C, Guo AC, et al. Drug Bank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 2008;36:901–906.
17. Recanatini M, Bottegoni G, Cavalli A. Lead optimization *In silico* antitarget screening. *Drug Discovery Today: Technologies.* 2004;1(3):209–215.
18. Anishetty S, Pulimi M, Pennathur G. Potential drug targets in *Mycobacterium tuberculosis* through metabolic pathway analysis. *Comput Biol Chem.* 2005;29(5):368–378.
19. Sekirov I, Russell SL, Antunes LCM, et al. Gut Microbiota in Health and Disease. *Physiological reviews.* 2010;90(3):859–904.
20. Jandhyala SM, Talukdar R, Subramanyam C, et al. Role of the normal gut microbiota. *World J Gastroenterol.* 2015;21(29):8787–8803.
21. Joseph M Pickard, Melody Y Zeng, Roberta Caruso, et al. Gut Microbiota: Role in Pathogen Colonization, Immunol Reponses and Inflammatory Disease. *Immunol Rev.* 2017;279(1):70–89.
22. Raman K, Yeturu K, Chandra N. target TB : A target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Syst Biol*. 2008;2:109.
23. Yu NY, Wagner JR, Laird MR, et al. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*. 2010;26(13):1608–1615.
24. Mondal SI, Ferdous S, Jewel NA, et al. Identification of potential drug targets by subtractive genome analysis of Escherichia coli O157 : H7 : an *in silico* approach. *Adv Appl Bioinform Chem*. 2015;8:49–63.
25. Yu CS, Cheng CW, Su WC, et al. CELLO2GO: A web server for protein subcellular localization prediction with functional gene ontology annotation. *PLoS One*. 2014;9(6):99368.
26. Hasan MA, Khan MA, Sharmin T, et al. Identification of putative drug targets in Vancomycin-resistant *Staphylococcus aureus* (VRSA) using computer aided protein data analysis. *Gene*. 2016;575(1):132–143.
27. Chen L, Yang J, Yu J, et al. VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Res*. 2005;33:325–328.
28. Berardini TZ, Mundodi S, Reiser L, et al. Functional Annotation of the Arabidopsis Genome Using Controlled Vocabularies. *Plant Physiol*. 2004;135(2):745–755.
29. Mulder NJ, Apweiler R, Attwood TK, et al. InterPro, progress and status in 2005. *Nucleic Acids Res.* 2005;33:201–205.
30. Bateman A, Coin L, Durbin R, et al. The Pfam protein families database. *Nucleic Acids Res.* 2009;32:138–141.
31. Habermann B, Villaveces J, Koti P. Tools for visualization and analysis of molecular networks, pathways, and -omics data. *Adv Appl Bioinforma Chem*. 2015;8:11–22.
32. Neuhaus F. On the Definition of 'Ontology'. 2018.
33. Jensen LJ, Kuhn M, Stark M, et al. STRING 8 — a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* 2009;37:412–416.
34. Biasini M, Bienert S, Waterhouse A, et al. SWISS-MODEL : modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 2014;42:252–258.
35. Wang S, Li W, Liu S, et al. Raptor X-Property : a web server for protein structure. *Nucleic Acids Res*. 2016;44(W1):430–435.
36. Xu J, Peng J. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins.* 2012;79(Supply 10):161–171.
37. Kelley LA, Mezulis S, Yates CM, et al. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*. 2015;10(6):845–858.

38. Jamil K, Mustafa SM. Thioredoxin System : A Model for Determining Novel Lead Molecules for Breast Cancer Chemotherapy. *Avicenna J Med Biotechnol*. 2012;4(3):121–130.

39. Colovos C, Yeates T. Verification of protein structures : Patterns of nonbonded atomic interactions. *Protein Sci.* 1993;2(9):1511–1519.

40. Laskowski RA, Molecular E, Thornton J, et al. PROCHECK : A program to check the stereo chemical quality of protein structures. *Journal of Applied Crystallography.* 1993;26:283–291.

41. Wiederstein M, Sippl MJ. ProSA-web : interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*. 2007;35:407–410.

42. Kumar S. Computational identification and binding analysis of orphan human cytochrome P450 4X1 enzyme with substrates. *BMC Res Notes*. 2015;8:9.

43. Messaoudi A, Belguith H, Ben Hamida J. Homology modeling and virtual screening approaches to identify potent inhibitors of VEB-1 β-lactamase. *Theor Biol Med Model.* 2013;10:22.

44. Bolton EE, Wang Y, Thiessen PA, et al. Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry.* 2008;4:217–241.

45. Irwin JJ, Shoichet BK. ZINC - A free database of commercially available compounds for virtual screening. *J Chem Inf Model*. 2005;45(1):177–182.

46. Wang K, Gao J, Shen S, et al. An Accurate Method for Prediction of Protein-Ligand Binding Site on Protein An Accurate Method for Prediction of Protein-Ligand Binding Site on Protein Surface Using SVM and Statistical Depth Function. *BioMed Research International.* 2013.

47. Lionta E, Spyrou G, Vassilatis DK, et al. Structure-Based Virtual Screening for Drug Discovery : Principles , Applications and Recent Advances. *Curr Top Med Chem*. 2014;14(16):1923–1938.

48. Morris GM, Lim Wilby M. Molecular docking. *Methods Mol Biol*. 2008;443:365–382.

49. Zhong HA. ADMET Properties: Overview and Current Topics. *Drug Design: Principles and Applications.* 2017:113–133.

50. Yu H, Adedoyin A. ADME-Tox in drug discovery: Integration of experimental and computational technologies. *Drug Discov Today*. 2003;8(18):852–861.

51. Nursamsiar, Surantaatmadja S, H Tjahjono D. Absorption, Distribution and Toxicity Prediction of Curculigoside A and its Derivatives. *Advances in Computer Science Research.* 2015:32–35.