Your full article (between 500 to 5000 words) - -   Do check for grammatical errors or spelling mistakes

Substractive Genomics approach to identify potential drug targets in *Streptococcus pyogens*

By :Shilpa Shiragannavar

Summary: The present study is carried out to identify potential drug targets in Streptococcus species that might facilitate the discovery of novel drugs in near future. Various steps were adopted to find out novel drug targets. susbtractive genomic approach has been used to identify therapeutic target in *Streptococcus pyogenes*.

INTRODUCTION

The complete genome sequences of many microbes were completed in the past decade. Valuable information on finding the treatment of various infections caused by pathogens can be retrieved using the comparative genomics and subtractive genomics approaches. The critical genes which are crucial for the survival of the pathogens and which are absent in the host can be screened out by using the subtractive genomics approach. The chances of cross reactivity and side effects can be decreased by selecting such non-homologous proteins which are not found in humans. The genes and their products which can be used as a potential drug targets can be identified by analyzing these genes with the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database.

A number of bioinformatics tools are also developed to analyze the genomes.

Completion of Human Genome Project is one of the major revolutions in the field of drug discovery against human pathogen. At present time genomic approach is in tradition. Identification of novel therapeutic targets is one of the major tasks in order to design a novel drug. There are many approaches to identify potential drug target such as virulence genes, uncharacterized essential genes, species–specific gene, unique enzyme and membrane transporter etc. Comparative genomic provide a new approach to identified novel drug target among previously known targets based on their related biological function in pathogen and host. In the proposed work subtractive genomic approach is used, where subtraction dataset comparing two genomes i.e. pathogen and human. This approach is successfully used in many other bacteria such as *Pseudomonas aeruginosa, Helicobacter pylori*, *Burkholderia pseudomalleii* etc.

There are many approaches to identify potential drug target such as virulence genes, uncharacterized essential genes, species–specific gene, unique enzyme and membrane transporter etc. Comparative genomic provide a new approach to identified novel drug target among previously known targets based on their related biological function in pathogen and host. In the proposed work subtractive genomic approach is used, where subtraction dataset comparing two genomes i.e. pathogen and human. This approach is successfully used in many otherbacteria such as *Pseudomonas aeruginosa*, *Helicobacter pylori*, *Burkholderia pseudomalleii.* In the subsequent work subtractive genomics and Database of Essential Gene (DEG) is used to analyze the genes of *Streptococcus pyogenes* for finding potential target at the outer surface of pathogen, might be used as drug target. The search for novel drug targets is relying on the genomics data. Comparative genomics approach can be used for selecting non-homologous genes coding for proteins which are present in pathogen but not in the host

For identifying such genes, Basic Local Alignment Search Tool (BLAST) against the human using BLASTP program can be performed. This will eliminate homologous genes present in the human. Thereafter, the critical genes required for the survival of the pathogen are identified using Database of Essential Genes (DEG). This approach will ensure that drug target is present only in the pathogen and not in the humans. Using this approach, novel targets have been identified successfully for various pathogens.

The involvements of combinatorial Chemistry, high throughp throughput screening, and virtual screening, in silico absorption, distribution, metabolism, and excretion–toxicity screening, and de novo and structure-based drug design serve to expedite as well as economize the modern day drug discovery process. Structure-based computational drug design methods mainly focus on molecules for target site with known 3D structure followed by determination of their affinity for target, based on which hits are obtained. Knowledge of these targets and corresponding drugs, particularly those in clinical uses and trials, is highly useful for facilitating drug discovery.

Table:1 Scientific Classification of S*treptococcus pyogenes*

Scientific classification

Kingdom	Eubacteria
Phylum	Firmicutes
Class	Bacilli
Order	Lactobacillales
Family	Streptococcaceae
Genus	*Streptococcus*
Species	*S. pyogenes*

Table above gives the Scientific classification of the organism *Streptococcus pyogenes,* like to which kingdom, phylum, class, order, family, genus, species the organism belongs to.

- *Streptococcus pyogenes*

Streptococcus pyogenes

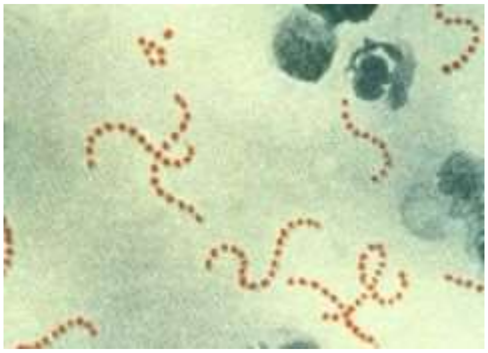


Figure:1 *S. pyogenes* bacteria at 900x magnification

Streptococcus is a genus of spherical Gram-positive bacteria belonging to the phylum Firmicutes, which is nonmotile, nonsporeforming coccus that occurs in chains or in pairs of cells. Individual cells are round-to-ovoid cocci, 0.6-1.0 micrometer in diameter (Figure 1.1). Streptococci divide in one plane and thus occur in pairs or (especially in liquid media or clinical material) in chains of varying lengths. The metabolism of *S. pyogenes* is fermentative, the organism is a catalase-negative aerotolerant anaerobe (facultative anaerobe), and requires enriched medium containing blood in order to grow.

Streptococcus pyogenes is one of the most frequent pathogens of humans. It is estimated that between 5-15% of normal individuals harbor the bacterium, usually in the respiratory tract, without signs of disease. It is estimated that there are more than 700 million infections world wide each year and over 650,000 cases of severe, invasive infections that have a mortality rate of 25%. Early recognition and treatment are critical; diagnostic failure can result in sepsis and death.

Acute Streptococcus pyogenes infections may take the form of pharyngitis, scarlet fever (rash), impetigo, cellulitis, or erysipelas. Invasive infections can result in necrotizing fasciitis, myositis, and streptococcal toxic shock syndrome. Patients may also develop immune-mediated sequelae such as acute rheumatic fever and acute glomerulonephritis.

Several antibiotics have many side effects and developed resistant against Streptococcus species. Most of the selected pathogens have resistant mechanism by efflux pump. There are genes which are responsible for the transport of drug molecules in S.agalactiae. Penicillin-binding protein 1a, lb, 2a, and 2b, ABC transporter ATP-binding protein/permease, chloramphenicol-O-acetyltransferase, and MATE efflux family protein are other few proteins responsible for drug resistant in selected pathogens. Hence it is the need of the hour to explore into the possibility of novel drug target identification and drug designing for these pathogens. This can be achieved now due to the availability of the proteomes of these organisms. This study has analyzed the proteome of the S.agalactiae, S. pneumoniae, and S. pyogenes and identified the most suitable and efficient drug-like compounds.

Several antibiotics have many side effects and developed resistant against Streptococcus species. Most of the selected pathogens have resistant mechanism by efflux pump. Hence it is the need to explore into the possibility of novel drug target identification and drug designing for the pathogen. This can be achieved now due to the availability of the proteomes of the organism. This study has analyzed the proteome of the *S.pyogenes* and identified the most suitable and efficient drug-like compounds.

Targets suggested for the pathogen are first screened for non human homologous, then the DEG database checks if the non-human homologous protein is essential for the survival of the organism, if the protein is essential for survival, it is further analyzed for its role in metabolic pathway of the organism. The proteins which are membrane bound and extracellular are considered to be the potential targets for pathogen i.e Streptococcus pyogenes.

The targets found for the pathogen are Glucose dehydrogenase and DNA Polymerase.

Virtual screening (VS) is a computational technique used in drug discovery research. By using computers, it deals with the quick search of large libraries of chemical structures in order to identify those structures which are most likely to bind to a drug target, typically a protein receptor or enzyme. The aim of virtual screening is to identify molecules of novel chemical structure that bind to the macromolecular target of interest.

There are two broad categories of screening techniques: ligand-based and structure-based. Given a set of structurally diverse ligands that binds to a receptor, a model of the receptor can be built by exploiting the collective information contained in such set of ligands. The ligands found after Auto docking and performing ADME and toxicity prediction are Tripdiolide Pomolic acid.

MATERIALS AND METHODOLOGY

Tools used for study: -

1. NCBI (National Center for Biotechnology Information)Genome

The NCBI is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health. The NCBI houses genome sequencing data in GenBank and an index of biomedical research articles in PubMed Central and PubMed, as well as other information relevant to biotechnology. All these databases are available online through the Entrez search engine.

This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

2. NCBI BLASTP:-

The *Basic Local Alignment Search Tool* (*BLAST*) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to the query sequence.

There are five types of BLAST programs, they are- nucleotide blast(Search a **nucleotide** database using a **nucleotide** query) protein blast(Search **protein** database using a **protein** query) blastx(Search **protein** database using a **translated nucleotide** query) tblastn(Search **translated nucleotide** database using a **protein** query) tblastx(Search **translated nucleotide** database using a **translated nucleotide** query).

BLASTP programs search protein databases using a protein query.

3. DEG(Database of essential genes):-

Essential genes are those indispensable for the survival of an organism, and therefore are considered a foundation of life. DEG hosts records of currently available essential genes among a wide range of organisms. For prokaryotes, DEG contains essential genes in more than 10 bacteria, such as *E. coli, B. subtilis, H. pylori, S. pneumoniae, M. genitalium* and *H. influenzae*, whereas for eukaryotes, DEG contains those in yeast, humans, mice, worms, fruit flies, zebra fish and the plant *A. thaliana*.

Users can Blast query sequences against DEG, and can also search for essential genes by their functions and names. Essential gene products comprise excellent targets for antibacterial drugs. Essential genes in a bacterium constitute a minimal genome, forming a set of functional modules, which play key roles in the emerging field, synthetic biology.

4. KEGG *(Kyoto Encyclopedia of Genes and Genomes*):-

KEGG is a bioinformatics resource for linking genomes to life and the environment. KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks.

KEGG PATHWAY mapping is the process to map molecular datasets, especially large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for biological interpretaion of higher-level systemic functions.

5. Cello Server(subCELlular LOcalization predictor):-

CELLO a subcellular Localization Predictor, is a multi-class SVM classification system. CELLO uses 4 types of sequence coding schemes: the amino acid composition, the di-peptide composition, the partitioned amino acid composition and the sequence composition based on the physico-chemical properties of amino acids.

6. Marvin Beans:-

It comes as a suite with 3 programs: Marvin Sketch, Marvin View, Marvin Space. These programs allow user to:

- Chemical structure drawing: Marvin Sketch allows users to quickly draw molecules through basic functions on the GUI and advanced functionalities such as sprout drawing, customizable shortcuts, abbreviated groups, default and user defined templates and context sensitive popup menus
- Query drawing
- Reaction drawing
- Atom and bond properties

7. Autodock Vina:-

AutoDock Vina is a new open-source program for drug discovery, molecular docking and virtual screening, offering multi-core capability, high performance and enhanced accuracy and ease of use. AutoDock Vina has been designed and implemented by Dr. Oleg Trott in the Molecular Graphics Lab at The Scripps Research Institute. AutoDock Vina significantly improves the average accuracy of the binding mode predictions compared to AutoDock 4, judging by our tests on the training set used in AutoDock 4 development.

Additionally and independently, AutoDock Vina has been tested against a virtual screening benchmark called the Directory of Useful Decoys by the Watowich group, and was found to be "a strong competitor against the other programs, and at the top of the pack in many cases". It should be noted that all six of the other docking programs, to which it was compared, are distributed commercially.

For its input and output, Vina uses the same PDBQT molecular structure file format used by AutoDock. PDBQT files can be generated (interactively or in batch mode) and viewed using MGLTools. Other files, such as the AutoDock and AutoGrid parameter files (GPF, DPF) and grid map files are not needed.

8. Hex:-

Hex is an interactive **protein docking** and **molecular superposition** program, written by Dave Ritchie. *Hex* understands protein and DNA structures in PDB format, and it can also read small-molecule SDF files. It is used for determining the best binding pose of the ligand and the receptor. The receptor-ligand complex can be saved and used for further analysis. There are different versions of Hex, the hex used in this project is Hex 6.3.

9. PreADMET:-

PreADMET program reside entirely on a Web server, and can be accessed by browsers such as Netscape or Internet Explorer. PreADMET consists of four main parts as following:

- Molecular Descriptor Calculation
- Drug-likeness Prediction
- ADME Prediction
- Toxicity prediction

Target information

Target identified for Streptococcus pyogenes: Glucose dehydrogenase & DNA Polymerase

- PDB ID for Glucose dehydrogenase: 1DLJ.pdb

It is involved in two component system as suggested in the KEGG Pathway database. Two-component signal transduction systems enable bacteria to sense, respond, and adapt to changes in their environment or in their intracellular state. Each two-component system consists of a sensor protein-histidine kinase (HK) and a response regulator (RR). In the prototypical two-component pathway, the sensor HK phosphorylates its own conserved His residue in response to a signal(s) in the environment. Subsequently, the phosphoryl group of HK is transferred onto a specific Asp residue on the RR. The activated RR can then effect changes in cellular physiology, often by regulating gene expression. Two-component pathways thus often enable cells to sense and respond to stimuli by inducing changes in transcription.

- PDB ID for DNA Polymerase: 2AVT.pdb

A complex network of interacting proteins and enzymes is required for DNA replication. Generally, DNA replication follows a multistep enzymatic pathway. At the DNA replication fork, a DNA helicase precedes the DNA synthetic machinery and unwinds the duplex parental DNA. On the leading strand, replication occurs continuously in a 5 to 3 direction, whereas on the lagging strand, DNA replication occurs discontinuously by synthesis and joining of short Okazaki fragments. In prokaryotes, the leading strand replication apparatus consists of a DNA polymerase, a sliding clamp, and a clamp loader. The DNA primase is needed to form RNA primers. Normally, during replication of the lagging-strand DNA template, an RNA primer is removed either by an RNase H or by the 5 to 3 exonuclease activity of DNA polymerase, and the DNA ligase joins the Okazaki fragments.

Methodology:-

The main objective of this project is to identify new leads and target for Streptococcus pyogenes through Substractive Genomics approach and Structure based virtual screening which is a step in drug discovery.

Project involves two parts:

1-Substractive Genomics Analysis

2- Virtual Screening

In general, the methodology followed for accomplishing this project are as follows:

1- Substractive Genomics Analysis

1.1 Retrival of proteomes of pathogen from NCBI Genome

1.2 Identification of paralogous proteins in the pathogen by performing BLASTP using UNIPORT database.

1.3 Identification of essential proteins from Non-paralogous proteins through comparison with DEG

1.4 Perform a BLAST search against PDB as the database with these essential proteins

2- Lead and target identification

2.1- Pathway analysis of these proteins using KEGG

2.2- Subcellular location prediction of target using CELLO SERVER

2.3- Finding a good lead by screening library of compounds (Lipinki's screening) using various online, and offline softwares.

a. Substractive Genomics Analysis:

a.1 Retrival of proteomes of pathogen from NCBI Genome:-

The complete genome, genes and protein sequences of *Streptococcus pyogenees* strain were retrieved from the NCBI Genome (National Center for Biotechnology Information) www.ncbi.nlm.nih.gov/gemone/. From the complete genome sequence data, the genes of the organism that coded for proteins, yet be unique to the organism were collected.

Identification of duplicate protein in *pathogen* proteins were eliminated at 60% using CD-HIT suite (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/) to identify the paralogs or duplicates proteins within the proteome of *Streptococcus pyogenees*. The prologs were excluded and the remaining sets of protein were used for further analysis.

a.2 Identification of paralogous proteins in the pathogen by performing BLASTP using UNIPORT database.

The nonparalogs proteins were subjected to NCBI BlastP (http://www.ncbi.nim.nih.gov/blast) against *Homo sapiens* protein sequences using threshold expectation value <=10^{-5} as parameter to find out the non-human homologues proteins of *Streptococcus pyogenes*. The human homologous were excluded and the list of non-homologs was compiled.

The search for novel drug targets is relying on the genomics data. Comparative genomics approach can be used for selecting non-homologous genes coding for proteins which are present in pathogen but not in the host. For identifying such genes, Basic Local Alignment Search Tool (BLAST) against the human using BLASTP program can be performed. This will eliminate homologous genes present in the human.

a.3 Identification of essential proteins from Non-paralogous proteins through comparision with DEG

The selected nonhuman homologues proteins were then subjected to similarity search using standard NCBI against the Database of Essential Genes (DEG) (http://tubic.tju.edu.cu/deg1). A random expectation value (E-value) cut-off of 10-100 and a minimum bit-score cut-off of 100 were used to screen out proteins that appeared to represent essential proteins.

The critical genes required for the survival of the pathogen are identified using Database of Essential Genes (DEG). This approach will ensure that drug target is present only in the pathogen and not in the humans. Using this approach, novel targets have been identified successfully for our pathogen.

a.4 Perform a BLAST search against PDB as the database with these essential Protein

The essential proteins obtained after performing DEG were subjected to BlastP against Streptococcus pyogenes using PDB as database for identifying the proteins PDB ID.

The critical genes which are crucial for the survival of the pathogens and which are absent in the host can be screened out by using the subtractive genomics approach. The chances of cross-reactivity and side effects can be decreased by selecting such non-homologous proteins which are not found in humans.

b. Lead and target identification

b.1 Pathway analysis of these proteins using KEGG

Metabolic pathway analysis of the essential proteins of *Streptococcus pyogenes* was done, KEGG Pathway Database(http://www.genome.jp/kegg/pathway.html) for the identification of potential targets. KEGG provides functional annotation of genes by BLAST comparisons against the manually curated KEGG GENES database. The result contains KO (KEGG Orthology) assignments and automatically generated KEGG pathways.

The critical genes which are crucial for the survival of the pathogens and which are absent in the host can be screened out by using the subtractive genomics approach. The chances of cross-reactivity and side effects can be decreased by selecting such non-homologous proteins which are not found in humans. The genes and their products which can be used as a potential drug targets can be identified by analyzing these genes with the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database.

b.2 Subcellular location prediction of target using CELLO SERVER

Prediction of protein localization is an important to predict the protein function and genome annotation, and it can assist the identification of targets. Sub-cellular localization analysis of the essential protein sequences has been done by CELLO SERVER, a Subcellular Localization Server (http://cello.life.nctu.edu.tw) to identify the surface membrane proteins which could be feasible vaccine target.

b.3 Finding a good lead by screening library of compounds

Virtual screening to determine best leads to target:

Virtual screening (VS) is done to determine the best lead which binds at appropriate sites with target to enhance the rate of reaction. Virtual screening is a computational technique used in drug discovery research. By using computers, it deals with the quick search of large libraries of chemical structures in order to identify those structures which are most likely to bind to a drug target, typically a protein receptor or enzyme. The aim of virtual screening is to identify molecules of novel chemical structure that bind to the macromolecular target of interest.

Building compound library:

A compound library was made which had around 150 anticancerous compounds and 3D structure which were downloaded from Pubchem, and Chemspider as SDF or Mol files. The target selected for *Streptococcus pyogenes* is based on annotations. Minimum energy conformation of compounds was derived using Marvin Sketch.

Lipinski's screening based on obtained scores like Molecular Weight. H-Bond Donor, H-Bond Acceptor ,XLogP. The mol weight should be less than 500g/mol, H-bond donor should be less than H-acceptor should be less than 10, & Xlog P should not be more than 5. Based on these scores the leads are screened. These screened leads will be used for Auto Docking. Using AutoDock Vina all compounds were docked one by one with the target.

Open AutoDock Vina and select target molecule and add a grid box with particular dimensions by preparing conf.txt file. AutoDock Vina is a new open-source program for drug discovery, molecular docking and virtual screening, offering core capability, high performance and enhanced accuracy and ease of use. Select the ligand and torsions and save the file as .pdbqt files. Docking scores are noted. According to the existing drug docks cores for the disease ,the best dockscores were selected.

Docking was done for both the targets with the lead molecules.

The top 10 best dock scoring ligands were redocked with Hex6.1 software and there energy values were noted and then they were ranked.

To study ADMET properties amd toxicity of ranked compounds, the online tool preADME was used. Results of all tools are attached in RESULTS section.

As human beings genome is more closer to mouse & rat, we check the carcinogenic effect of mouse & rat, if it is positive there is no carcinogenic effect & if negative there is carcinogenic effect exhibited by the lead.

Hence the positive results are selected as best leads.

The best leads as obtained after Toxicity Prediction are Lutein and 1,4-napthoquinone for Glucose dehydrogenase

The best leads as obtained after Toxicity Prediction are Emodin and Plumbagin leads for DNA Polymerase

EXPERIMENTAL RESULTS

The results that were obtained by the subtractive based approach are summarised in Table 4.1. The objective of the work was to identify and locate those essential proteins of *S.pyogenes* that are unique i.e. absent in host and performing normal function within the host and to shortlist them in vaccine development point of view. Identification of non-human homologs essential proteins of *S.pyogenes* with subsequent screening of the proteome to find the corresponding proteins that are likely to lead to development of drugs that specifically interact with the parasite to inhibit its activity. The nonhuman homologs of the membrane proteins would represent ideal vaccine targets. Novel drug targets have been identified successfully for various pathogens with the help of subtractive genomics approach.

The following project analysis has identified 766 non-human homologous proteins and by subjecting these proteins to BLASTP against human genome provided by the NCBI server resulted in 340 essential, non human homolog genes. By further analyzing these essential and non-human homolog genes, it was found 2 proteins that are possibly located on the membrane of the pathogen could be considered as potential drug targets for the pathogen.

A number of approaches for new vaccine development exist, such as sub-unit protein and DNA vaccines, recombinant vaccines, auxotrophic organisms to deliver genes and so on. Testing such candidates is tedious and expensive. *In-silico* approaches enable us to reduce substantially the number of such candidates to test and speed up drug discovery with least toxicity. The use of DEG database is more efficient than conventional methods for identification of essential genes and facilitates the exploratory identification of the most relevant drug targets in the pathogen.

The subtractive genomic approach has been applied in the present study for the identification of several proteins that can be targeted for effective drug design and vaccine development against *S.pyogenes.* The drugs developed against these will be specific to the pathogen, and therefore less or non toxic to the host. Structural modeling of these targets will help identify the best possible sites that can be targeted for drug design by simulation modeling. Virtual screening against these novel targets might be useful in the discovery of novel therapeutic compounds against *S.pyogenes.*

Table: 2 Subtractive proteomic and metabolic pathway analysis result for *S.pyogenes*

Total Number of proteins	**1978**
Duplicates (>60% identical) in CD-HIT	1822
human homologous	256
Non-human homologous proteins (E-value 10^{-5})	766
Essential protein in DEG (E-value 10^{-5})	340
Essential proteins involved in metabolic pathways	18
Proteins involved in unique pathways	2

Membrane associated drug targets finally obtained	2

The critical genes which are crucial for the survival of the pathogens and which are absent in the host can be screened out by using the subtractive genomics approach. The chances of cross-reactivity and side effects can be decreased by selecting such non-homologous proteins which are not found in humans. The genes and their products which can be used as a potential drug targets can be identified by analyzing these genes with the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. The metabolic pathways for DNA Polymerase and Glucose dehydrogenase are shown in fig 1 and 2 respectively.

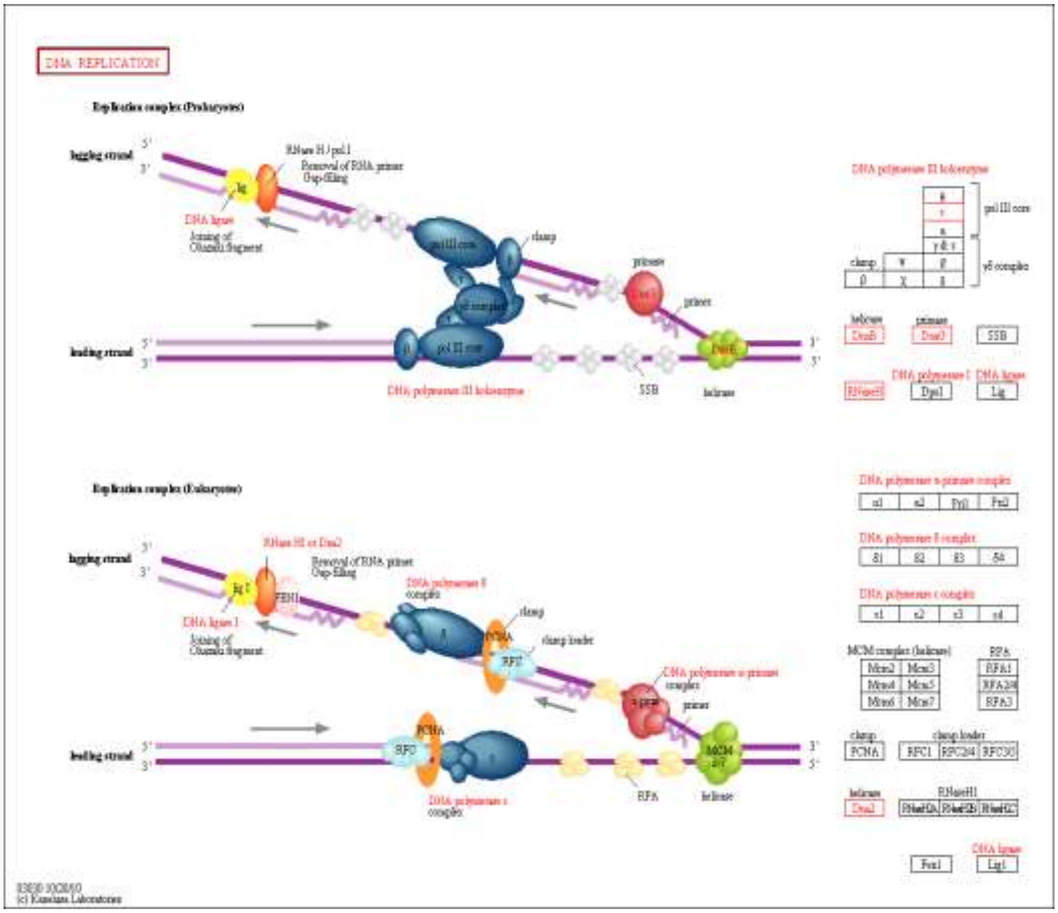


Figure:2 KEGG Pathway results for DNA Polymerase, its role in DNA Replication pathway

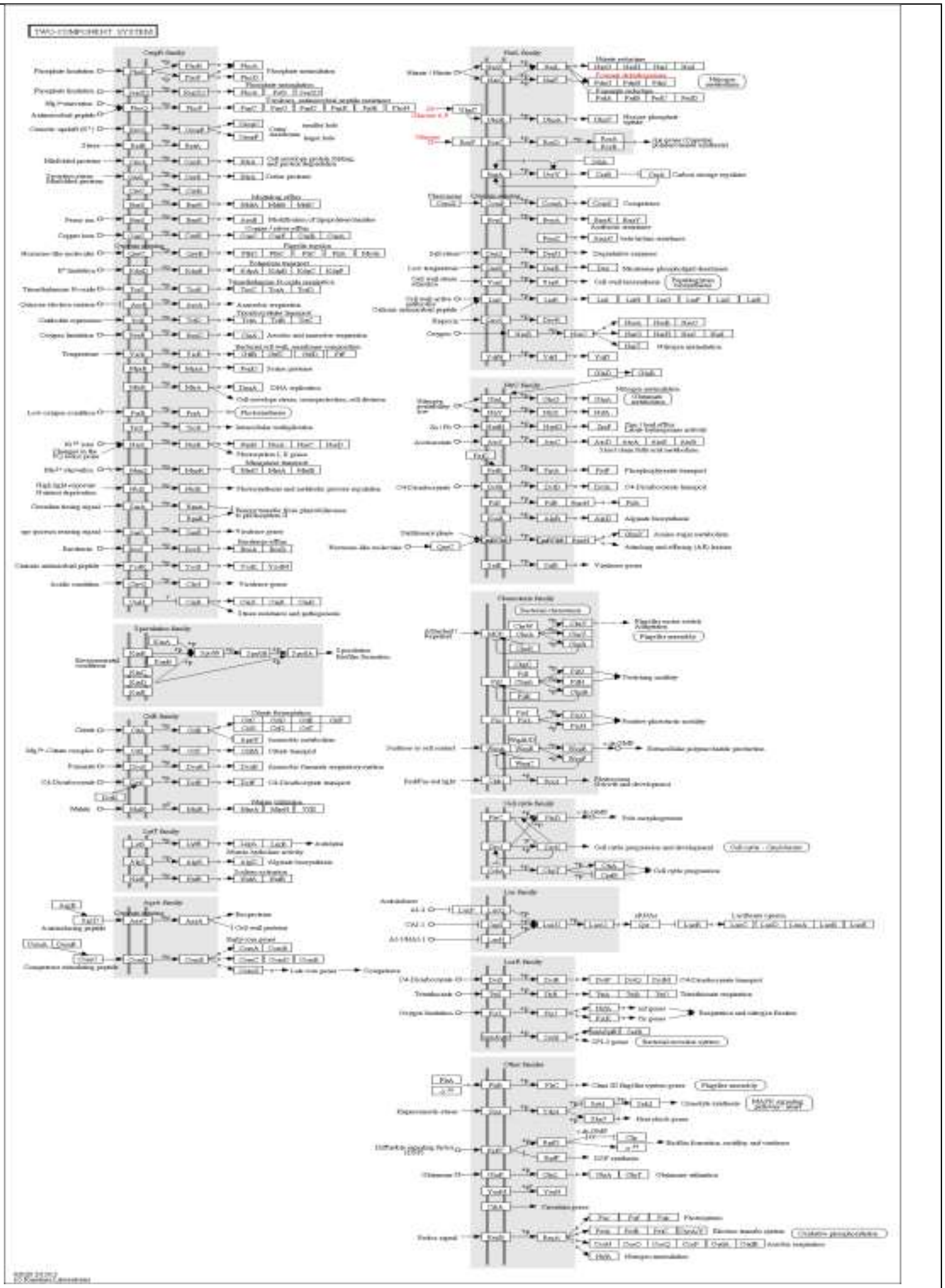


Figure:3 KEGG Pathway results for Glucose Dehydrogenase, its role in pathway Two Component System

Bioinformatics enables us to reduce substantially the number of the candidates to test. The computational genomics approach stated here is likely to speed up drug discovery process by removing hindrances like dead ends or toxicity that are encountered in classical approaches. The membrane associated proteins of *Streptococcus pyogenes strain* are invariably linked with essential metabolic and signal transduction pathways. The localization of the protein also plays vital role in determining the target, hence the targets suggested based on the CELLO results, i.e location and reliability the targets selected were Glucose dehydrogenase and DNA Polymerase.

Table: 3 Subcellular Localization results obtained by CELLO localization predictor.

Target Protein	**Localization**	**Reliability**
Glucose Dehydrogenase	Cytoplasmic	4.813
DNA Polymerase	Membrane	2.055
	Extracellular	1.926

Target information:

Target identified for Streptococcus pyogenes:

- Glucose dehydrogenase &
- DNA Polymerase

- PDB ID for Glucose dehydrogenase: 1DLJ.pdb

It is involved in two component system as suggested in the KEGG Pathway database. Two-component signal transduction systems enable bacteria to sense, respond, and adapt to changes in their environment or in their intracellular state. Each two-component system consists of a sensor protein-histidine kinase (HK) and a response regulator (RR). In the prototypical two-component pathway, the sensor histidine kinase phosphorylates its own conserved His residue in response to a signal(s) in the environment.

Subsequently, the phosphoryl group of histidine kinase is transferred onto a specific Asp residue on the response regulator. The activated response regulator can then effect changes in cellular physiology, often by regulating gene expression. Two-component pathways thus often enable cells to sense and respond to stimuli by inducing changes in transcription.

PDB ID for DNA Polymerase: 2AVT.pdb

A complex network of interacting proteins and enzymes is required for DNA replication. Generally, DNA replication follows a multistep enzymatic pathway. At the DNA replication fork, a DNA helicase precedes the DNA synthetic machinery and unwinds the duplex parental DNA. On the leading strand, replication occurs continuously in a 5 to 3 direction, whereas on the lagging strand, DNA replication occurs discontinuously by synthesis and joining of short Okazaki fragments.

In prokaryotes, the leading strand replication apparatus consists of a DNA polymerase, a sliding clamp, and a clamp loader. The DNA primase is needed to form RNA primers. Normally, during replication of the lagging-strand DNA template, an RNA primer is removed either by an RNase H or by the 5 to 3 exonuclease activity of DNA polymerase, and the DNA ligase joins the Okazaki fragments.

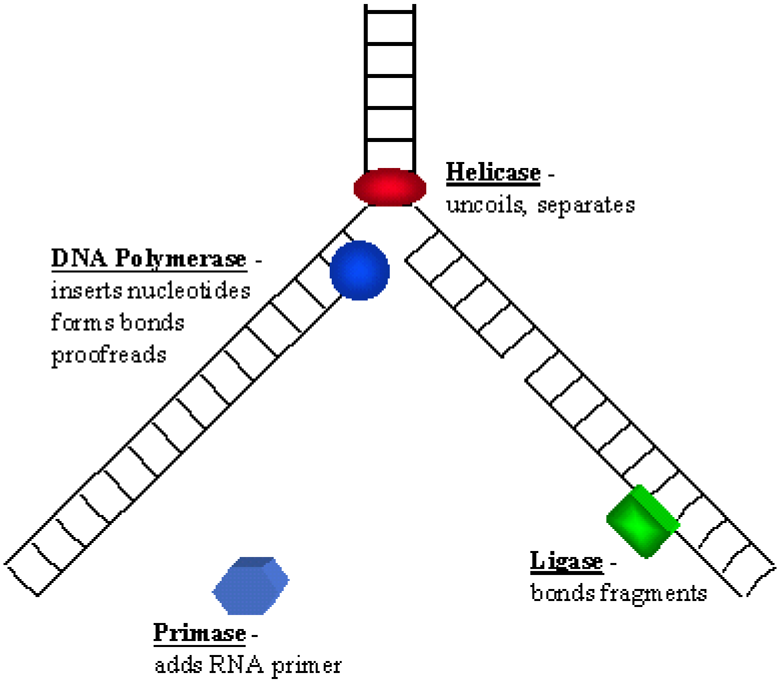


Figure: 4 Role of DNA Polymerase in Replication

Virtual screening to find best lead:

A compound library was made which had 150 antibacterial compounds and 3D structure were downloaded from Pubchem, and Chemspider as SDF or Mol files.

The target selected for *Streptococcus pyogenes* based on annotations from KEGG Pathway are Glucose Dehydrogenase and DNA-Polymerase. Minimum energy conformation of compounds was derived using Marvin Sketch. Using AutoDock Vina all compounds were docked one by one with the target, open AutoDock Vina and select target molecule and add a grid box with particular dimensions by preparing conf.txt file. Select the ligand and torsions and save the file as .pdbqt files. Docking score are noted. According to the existing drug dock scores for the pathogen, the best dock scores were selected.

Table: 4 AutoDock results for 1DLJ (Glucose dehydrogenase)

Compound id	Affinity	Rank
cid_259577	-10.3	1
cid_5351344	-10.2	2
cid_83843	-10.2	3
cid_9064	-10.2	4
cid_382831	-9.0	4
cid_9817550	-8.3	5
cid_105111	-7.9	6
cid_ 54678486	-7.2	7
cid_4042	-6.9	8
cid_8530	-5.4	9
cid_10041259	-5.2	10

Table: 5 AutoDock results for 2AVT (DNA Ploymerase)

Compound id	Affinity	Rank
cid_259577	-10.9	1
cid_5351344	-10.1	2
cid_83843	-10.0	3
cid_9064	-10.0	4
cid_382831	-8.8	4
cid_9817550	-8.2	5
cid_105111	-7.2	6
cid_ 3220	-6.7	7
cid_6249	-6.2	8
cid_10205	-5.8	9
cid_10041259	-5.6	10

The top 10 best dock scoring ligands were redocked with Hex6.1 software and there energy values were noted and then they were ranked. The ligand receptor complexes obtained from Hex were analysed using online tool called Q-site finder, to know residues in binding site. The complexes were viewed in PyMol to see the interacting residues (target and ligand). To study ADMET properties amd toxicity of ranked compounds, the online tool preADME was used. The best leads found after toxicity prediction are Amoxicillin and Cephalexin .The results are as follows.

Table: 6 Admet test results for Absorption & Distribution of leads

Ligand	**Absorption**				**Distribution**	
	Human intestinal absorption (HIA,%)	In vitra caco-2 cell permeability (nm/sec)	In vitro MDCK cell permeability (nm/sec)	In vitro skin permeability(logKp, cm/hour)	In vitro plasma protein binding(%)	In vivo blood-brain barrier penetration(C.brain/C.blood)
cid_259577	45.59979	.708826	0.554786	-5.17424	41.346348	0.0495777
cid_5351344	21.01295	19.8908	0.511919	-5.16274	33.358920	0.0329153
cid_83843	89.16193	25.529	0.594227	-4.77099	37.450953	0.012638
cid_9064	69.55503	20.963	7.22203	-4.64722	69.654442	0.0953538
cid_382831	76.17051	20.7669	0.561361	-5.54442	38.112273	0.0272161
cid_9817550	88.04429	21.2524	0.043762	-3.35153	100.00000	3.37177
cid_105111	24.66360	17.2848	0.520026	-5.15877	28.548405	0.0364744
cid_5467 8486	94.96857	43.8235	0.646969	-2.64562	74.842892	0.0979118
cid_4042	30.67044	21.0277	0.533005	-5.33045	50.194045	0.0476871
cid_8530	85.93853	21.0867	45.6504	-3.04242	92.545090	3.68124
cid_10041259	17.2848	94.96857	43.8235	-3.04242	74.842892	0.0979118
cid_	43.8235	69.55503	0.047583	-4.48626	89.46	0.02859744

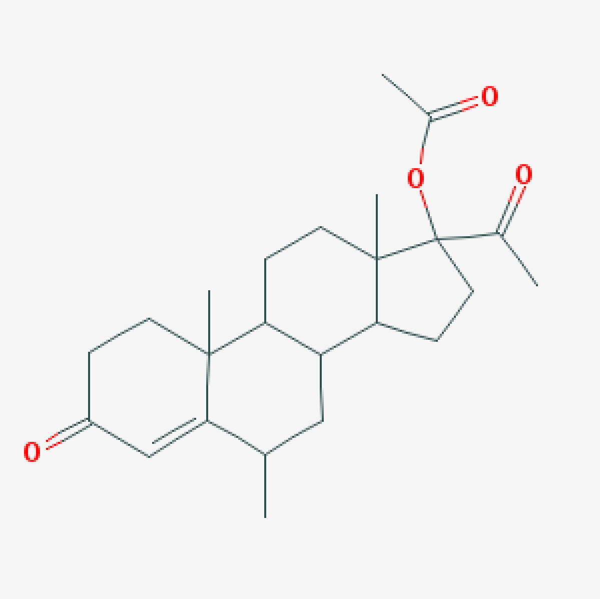
3220					5786	
cid_10205	25.529	88.04429	0.364866	-8.76322	35.689778	0.03658694

Table: 7 Toxicity results indicating the carcinogenic effects of the leads

ligand	**Amest test**							**Carcinogenicity**	
	Ames TA100 (+S9)	**Ames TA100(-S9)**	**Ames TA1535(+S9)**	**Ames TA1535(-S9)**	**Ames TA98(+S9)**	**Ames TA98(-S9)**	**Ames test**	**Carcinogenicity(Mouse)**	**Carcinogenicity(Rat)**
cid_259577	Positive	negative	negetive	negetive	Positive	negative	Mutagen	negetive	negative
cid_5351344	negative	negative	negetive	negetive	negetive	Positive	mutagen	negetive	Negative
cid_83843	negative	negative	negetive	negetive	negetive	Negative	Non mutagen	negetive	Negative
cid_9064	negative	negative	negetive	negetive	negetive	Negative	Non mutagen	negetive	Negative
cid_382831	negative	negative	negetive	negetive	negetive	Negative	Non mutagen	negetive	negative
cid_9817550	negative	negative	negetive	negetive	negetive	Negative	Non mutagen	negetive	negative
cid_105111	negative	negative	negetive	negetive	negetive	Positive	mutagen	negetive	Negative
cid_5467 8486	negative	positive	negetive	negetive	positive	Negative	mutagen	negetive	Negative
cid_404	negati	nega	negative	negative	negative	negative	Non-	**Positive**	**Positive**

2	ve	tive					mutagen		
cid_8530	Positive	Positive	Positive	Positive	Positive	Positive	mutagen	**Positive**	**Positive**
cid_10041259	negative	negative	positive	positive	positive	Positive	mutagen	negetive	negative
cid_3220	negative	positive	negative	negative	positive	positive	mutagen	negative	**positive**
cid_10205	Positive	Positive	Positive	negative	Positive	Positive	mutagen	negative	**positive**

The best leads as obtained after Toxicity Prediction are Lutein and 1,4-napthoquinone for Glucose dehydrogenase



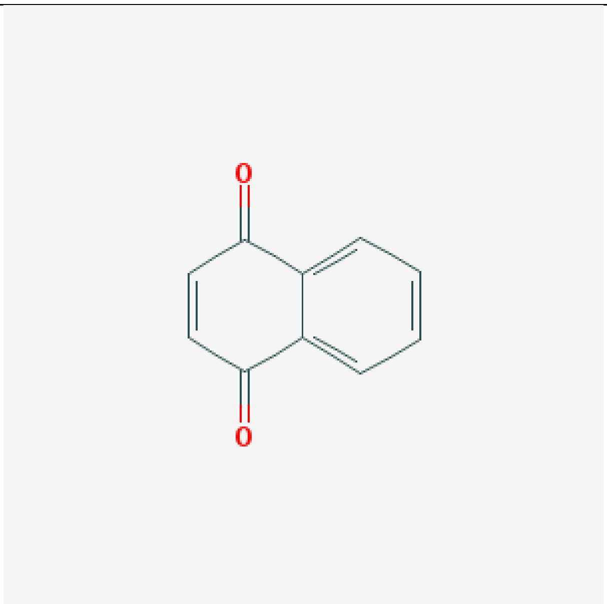
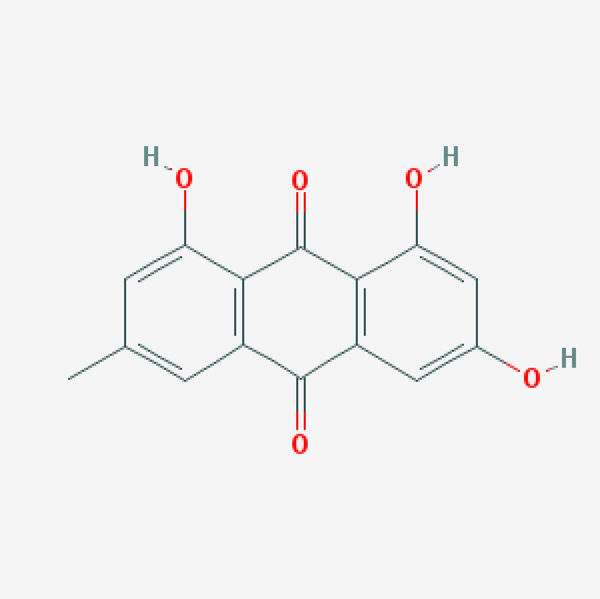


Fig: 5 Structure of Lutein and 1,4-napthoquinone leads for Glucose dehydrogenase

The best leads as obtained after Toxicity Prediction are Emodin and Plumbagin leads for DNA Polymerase



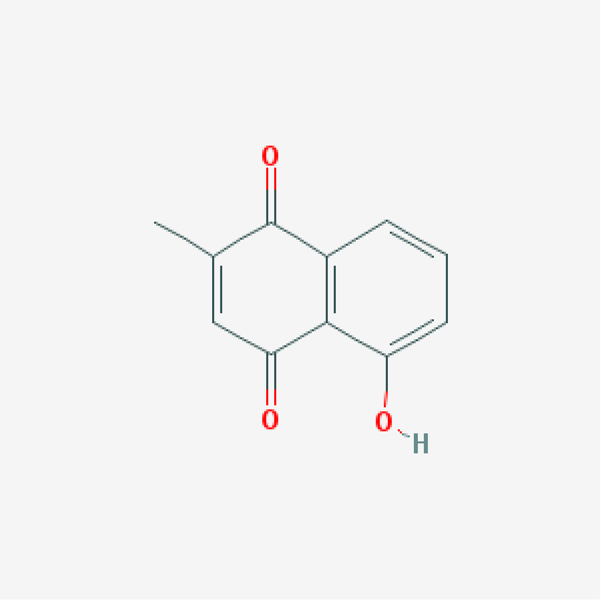


Fig: 6 Structure of Emodin and Plumbagin leads for DNA Polymerase

Table: 8 Lipinski's Screening For Leads

Lead	Compound Id	Mol.Wt	Xlogp	H-Bond Donor	H-Bond Receptor
Lutein	Cid_4042	386.524	4.1	0	4
1,4-Napthoquinone	Cid-8530	158.153	2.7	3	5
Emodin	Cid_3220	270.236	2.3	1	3
Plumbagin	Cid-10205	188.179	1.7	0	2

Lipinski's screening based on obtained scores like Molecular Weight. H-Bond Donor, H-Bond Acceptor ,XLogP. The mol weight should be less than 500g/mol, H-bond donor should be less than H-acceptor should be less than 10, & Xlog P should not be more than 5. Based on these scores the leads are screened. These screened leads will be used for Auto Docking.

Among the 150 anti bacterial samples collected the above four compounds which satisfy Lipinski's screening are proved to be the best leads for the targets.

CONCLUSION

The computational genomic approach has greatly accelerated the identification of potential drug targets against a variety of parasites. Use of the subtractive genomics is more efficient than traditional methods to identify unique proteins and facilitates the exploratory identification of the most relevant drug targets in the parasite. This computational analysis has thus led to the identification of several proteins that can be targeted for effective drug design and vaccine development against *Streptococcus pyogenes* as the threat caused by it is becoming a serious concern for developing countries and effective drugs and vaccines are yet to be developed. It is being identified that two best possible enzyme drug targets are expected to be unique for the parasite from various metabolic pathways.

Glucose dehydrogenase and DNA Polymerase are proposed that they may be the better option for drug design against *Streptococcus pyogenes*. Further, prediction of best leads for these targets will help identify the best possible lead that can be targeted for drug design by AutoDock. Streptococcus are highly heterogeneous species, therefore, based on the homology these proteins share with those of other *Streptococcus* species, best targets can be used for the drug development so that these can be used in other species as well.

References (if any)

1. Dutta A., Singh S.K., Ghosh P.Mukherjee R., Mitter S. and Bandyopadhyay D. (2006) *In Silico Biology* 6, 0005.
2. Sarangi A.N., Aggarwal R., Rahman Q., Trivedi N. (2009) *B. J Comput Sci Syst Biol* 2: 255-258.

About Author:

Your Full Name (published)	Shilpa Shiragannavar
A few lines about you: (published)	M E in Bioinformatics

Terms - Do not remove or change this section (It should be emailed back to us as it is)