LECTURE-37

Proteomics and Bioinformatics

Handout

PREAMBLE

Genomics has evolved to study the whole genome of an organism, which provides the information for biological research. Over a period, many advanced technologies such PCR, automated sequencing, DNA microarray has advanced and generated massive amount of data. After completion of human genomic project, there has been an increasing interest to understand the massive amount of genetic information generated and to assign function to the large number of proteins encoded. Proteomics is one of the major disciplines for the development of novel technologies to address protein function/ complexity in global manner. A number of proteome analysis methods have emerged in proteomics research, including two dimensional gel electrophoresis, mass spectrometry, quantitative proteomics, protein microarray, label-free technology and bioinformatics tools. Other omics techniques such as transcriptomics, metabolomics, lipidomics, interactomics etc have also emerged during the last decade. The basic problem with these high throughput technologies is the massive data storage and analysis to interpret the biological problem. In recent years bioinformatics has advanced significantly to provide meaningful information from the high-throughput data generated from omics techniques.

OUTLINE OF LECTURE

1. High-throughput proteomic techniques

   a. 2-DE data analysis

   b. Mass spectrometry data analysis

   c. LC-MS data analysis

2. Bioinformatics analysis of the proteomic data

   a. Basic bioinformatics tools

   b. Interaction network analysis

   c. Metabolic pathway analysis

   d. 3D structure modeling

   e. Molecular docking

BOX FOR TERMINOLOGY

- **Genomics**: The study of whole genome of an organism is called genomics.
- **Proteomics**: The study of an entire proteome complement of an organism, cell or tissue.
- **Bioinformatics**: Deals with storage, retrieval and analysis of the biological data.
- **Panther:** Deals with comprehensive functional gens information to highlight the biological roles of the query proteins.
- **String:** Interaction network database showing the experimental interactions and functional interactions such predictions, co-occurrence, data mining etc.
- **Reactome:** One of the manually curated and peer reviewed database for pathway analysis to understand the biological role of the query molecule.

1. High-throughput proteomic techniques

(a) 2-DE analysis

Two-dimensional electrophoresis (2-DE) resolves hundreds of protein spots across the gel based on molecular weight and pI. There are few bioinformatics tools to simplify the data analysis of the 2-DE gel image to get the no. of spots, its intensity, spot matching etc to obtain meaningful information from gels. Many algorithms have been included in software to provide us the reproducible patterns by generating the raw intensity tables, scattered plots, histograms and 3D views of the spot. There are many commercial and free web based 2-DE analysis tools available **(**Palagi et al., 2006) (Table-1).

Table 1: Commercial and open-source 2-DE and DIGE gel analysis software tools

S N.	Software	Description	URL site
1	Image Master Platinum*	IMP7 and DeCyder software from GE Healthcare for 2-DE and DIGE analysis	www.gehealthcare.com
2	Delta2D	Decodon software for 2-DE and DIGE analysis	http://www.decodon.com/delta2d.html
3	Melanie*	Software from GeneBio for 2-DE and DIGE analysis	http://world-2dpage.expasy.org/melanie/
4.	PDQuest*	Software from BioRad for 2-DE and DIGE analysis	www.bio-rad.com
5.	Progenesis	Software from nonlinear dynamics for 2-DE and DIGE analysis	www.nonlinear.com
6	REDFIN	Software from Ludesi for 2-DE analysis	http://www.ludesi.com/redfin/
7	Flicker$^{\$}$	2-DE analysis software	www.lecb.ncifcrf.gov/flicker/wgFlkPair.html
8	GelScape$^{\$}$	1D and 2-DE gel analysis software	www.gelscape.ualberta.ca

* Free software version available for analysis, $ open source software

b. Mass spectrometry data analysis

Mass spectrometry has succeeded in identification of amino acid sequence of tryptic digested peptide fragments and provides greater than 99% accurate protein prediction. To enhance the accuracy of protein identification; PMF (peptide mass fingerprinting), PFF (peptide fragmenting fingerprinting) and MS/MS ion search have been employed with the help of sophisticated algorithms. The widely used databases for protein identification search are provided in Table-2 (Palagi et al., 2006).

Table 2: Tools and databases for MS data analysis and quantitative information

S. No	Software/database	Description	URL site
1	MASCOT	Search engine for protein identification using mass spectrometry data	http://www.matrixscience.com/
2	MS-Fit	Used for mining the sequence of the protein from MS data	prospector.ucsf.edu
3	SEQUEST	Used for interpretation of tandem mass spectra data for protein identification and amino acid sequence	http://fields.scripps.edu/sequest/
4	X!Tandem	Used for protein identification using tandem mass spectra data	http://www.thegpm.org/tandem/index.html
5	Sequit!	*De novo* sequencing of protein using tandem mass spectrum	http://www.sequit.org/
6	MSQuant	Quantitative proteomic information from MS and LC data	http://msquant.sourceforge.net/

c. LC-MS analysis

There are many online tools available for the analysis of LC-MS/MS data, which are listed in Table-3 (Palagi et al., 2006).

Table 3: Commercial and open-source MS and LC-MS/MS data analysis software

S. No	Software	Description	URL site
1	MapQuant	Used for MS quantification after making two dimensional map	http://arep.med.harvard.edu/MapQuant/
2	XCMS	Used for LC-MS data handling for relative quantization, visualization.	http://metlin.scripps.edu/xcms/
3	MsInspect	Used to combine the LC-MS and LC-MS/MS peptide data and also for peptide array generation	http://proteomics.fhcrc.org/CPL/msinspect/index.html
4	Mzmine	Mainly used for MS and LC-MS data processing purpose	http://mzmine.sourceforge.net/
5	Pep3D	Convert LC-MS or LC-MS/MS data into 2D map as m/z vs time	---------------
6	SpecArray	Used for generation of expression peptide arrays from LC-MS data	tools.proteomecenter.org/SpecArray.php
7	Msight	It represent the data from both MS and separation steps such as chromatography, 2De etc	http://web.expasy.org/MSight/

2. Bioinformatics analysis of proteomics data

a. Basic bioinformatics tools

The enormous amount of data that is generated by the sequencing studies needs to be documented in an organized manner, which is compatible for further processing and curation. These databases are classified depending on information that they store; like genomic or proteomic databases or depending on the organism, cell state, etc. under consideration. The Human Genome Project has been the landmark in field of sequencing and has enabled not only the mapping of all the genes present in the genome, but also annotation. Gene Annotation and Database Management happen to cover the core of the field of Bioinformatics. With the aid of various tools such as BLAST, Pipmaker, Splign; the available sequences can be mined, aligned and further processed in different ways to gain deeper insights into the same. The database can also be searched for matches of the unknown sequence, and thereby deduce information regarding the structural and functional resemblance between the known and unknowns. Some of the commonly available databases and tools available for bioinformatics analysis are described in following simulations.

Illustration: Protein sequence alignment

With sequencing of large number of proteins and subsequent storage of data, it has become easier for researchers to study the proteins. These studies help to provide preliminary insights into the structural and functional aspects of proteins without conducting experiments. Alignment algorithms take two protein sequences and align them residue-by-residue. In this illustration alignment between two sequences of CBR-COL-186 protein of Caenorhabditis briggsae and collagen of Caenorhabditis elegans are shown. For a more detailed study on the types of BLAST tools available, visit http://blast.ncbi.nlm.nih.gov/Blast.cgi Pair-wise alignment gives various kinds of results after alignment. These are alignment views, alignment score, dot-plot, e-value, and *percentage identity amongst many others. Multiple Sequence Alignment tools are used to compare the amino acid sequences of more than two proteins.*

Illustration: Alignment analysis and interpretations

Multiple sequence alignment produces alignment files (.aln), which can be used to determine the evolutionary distances of a set of given protein sequences. Many server-based and stand-alone programs can achieve this. One needs to select the method for calculating the distance. In this illustration the usage of alignment files for phylogenetic analysis is depicted. Cladograms are the graphical representation of the branching during evolution of the proteins that were aligned. Phylograms represent the evolutionary distance tree in a graphical format. In this, the branch lengths correspond to the evolutionary distance between the two proteins. All branches will converge to a common ancestral root. Alignment files can be used for a variety of structural and functional analysis.

Illustration: Structural databases

The protein structural databases contain a basic search box, which requires the input for an identifier of the protein. This identifier can be the protein name, keyword, ID, author, etc. In this illustration the case of Viral Capsid Proteins is presented. These databases have advanced search features, which are optional but help in making the query very specific.

Illustration: Use of structural databases

Two given proteins can be structurally aligned to evaluate the similarity between them. The server requires an input of two protein sequences or their IDs, which are then simulated and aligned based on their 3D coordinates, bond angles and dihedral angles. Few of the various servers available for this are DALI, MAMMOTH, CE/CE-MC, SSAP and ProFit. Since, all known proteins have not been structurally characterized, this provides a useful bioinformatics analysis tool for researchers. The various servers for structure prediction are GOR, HNN, PredictProtein, NNPredict and Sspro.

Given a particular amino acid sequence, the cellular, molecular and biological processes associated with the sequence can be predicted using functional annotation servers. These processes are represented by a unique set of identifiers called “Gene Ontology Terms” or the “GO Terms”. The GO term can be a word or an alphanumeric identifier that includes a definition with cited sources and a namespace indicating the domain to which it belongs. The various servers for this include DbAli Annolite, PFP, ProteomeAnalyst, GOPET, SpearMint and ProKnow.

Illustration: Genome databases

With the advent of the genome sequencing technology, biological research has now become easy and fast access to the complete DNA sequences of many organisms. This DNA sequence information, when stored with the help of databases, can be used for comparative genomics research. To submit a sequence in a nucleotide database, it *must be entered in any one of the sites of the members of the International Nucleotide Sequence Database Collaboration consisting of NCBI, EBI-EMBL and DDBJ. The verified sequence is given an accession number or a gene ID, which acts as the primary key for identifying this entry in the database in future. The user needs to select the database from which sequence has to be retrieved. These databases include: Gene, Genome, EST, SNP, NUCLEOTIDE, GEO DATASETS. Searching it against a suitable nucleotide database can identify an unknown nucleotide sequence. Input the sequence, and then select the database against which the match search is to be performed. Fill the parameter values and then click on the blast tool. Sequence identification through BLAST provides various results after alignment such as identification, alignment views, alignment score, e-value, percentage identity and gaps. Alignment can also be performed between two given nucleotide sequences. To align two sequences, enter them in the input boxes. Enter the necessary parameters, whose values will vary according to query. Then click on the alignment tool. Pair-wise alignment gives various kinds of results after alignment. These are alignment views, alignment score, dot-plot, e-value, and percentage identity.*

Illustration: From wet lab to bioinformatics

The cells present in the tissue culture are lysed open thereby releasing crude extract. Protein of interest must then be isolated from this mixture. The protein of interest is separated from the protein mixture present in the supernatant. This is carried out by suitable techniques such as chromatography or electrophoresis, which makes use of various properties of the proteins such as their charge, mass etc for separation.

Edman degradation employs phenyl isothiocyanate reagent, which reacts with amino terminal residue of the peptides giving rise to phenyl thiocarbamoyl derivative of the amino-acid residue. In mild acidic conditions, this cyclic derivative of the amino acid is released in the form of a PTH-amino acid, which can then be identified by chromatographic techniques. The procedure is then repeated to identify each N-terminal amino acid sequentially. A tandem mass spectrometer can be used for protein sequencing studies.

Illustration: Protein databases

All data related to a protein can be divided into four broad categories namely sequence details, Source, Gene details and References. "Sequence" details contain the features of a protein's amino acid sequence such as the length, location, patterns and identifiers of the protein sequence. The "source" contains information based on the biological source used for retrieving the protein. "Gene" contains details of the gene from which the protein is being expressed. "Reference" contains the details of the research publication in which the study was reported.

Database designing is done at various levels such as Physical, Logical and View. At the physical level, the purpose of the database, which is in accordance with the prospected usage, is defined. At the logical level the tables; attributes of the tables and *relationship between tables are defined. Logical level is the most complex and important schema for databases and requires a thorough understanding of the data and its contexts and relationships. At the View level the views and appearance of the database is defined.*

A typical biological database can be characterized by its "Type" and its "Tools". The "Type" defines the category of data that it includes, such as sequence, domains or structure. This implies that the particular database's most prominent feature includes either sequences, domains or structure and it will primarily be used for their analysis. The analysis tools define the platforms that the site will provide for gaining an insight into the protein data.

The generalized information that users can obtain from protein databases is

1. General description of the protein molecule

2. Annotations of the protein

3. Name and description of the gene that transcribes them

4. ID of the same protein in other relevant databases

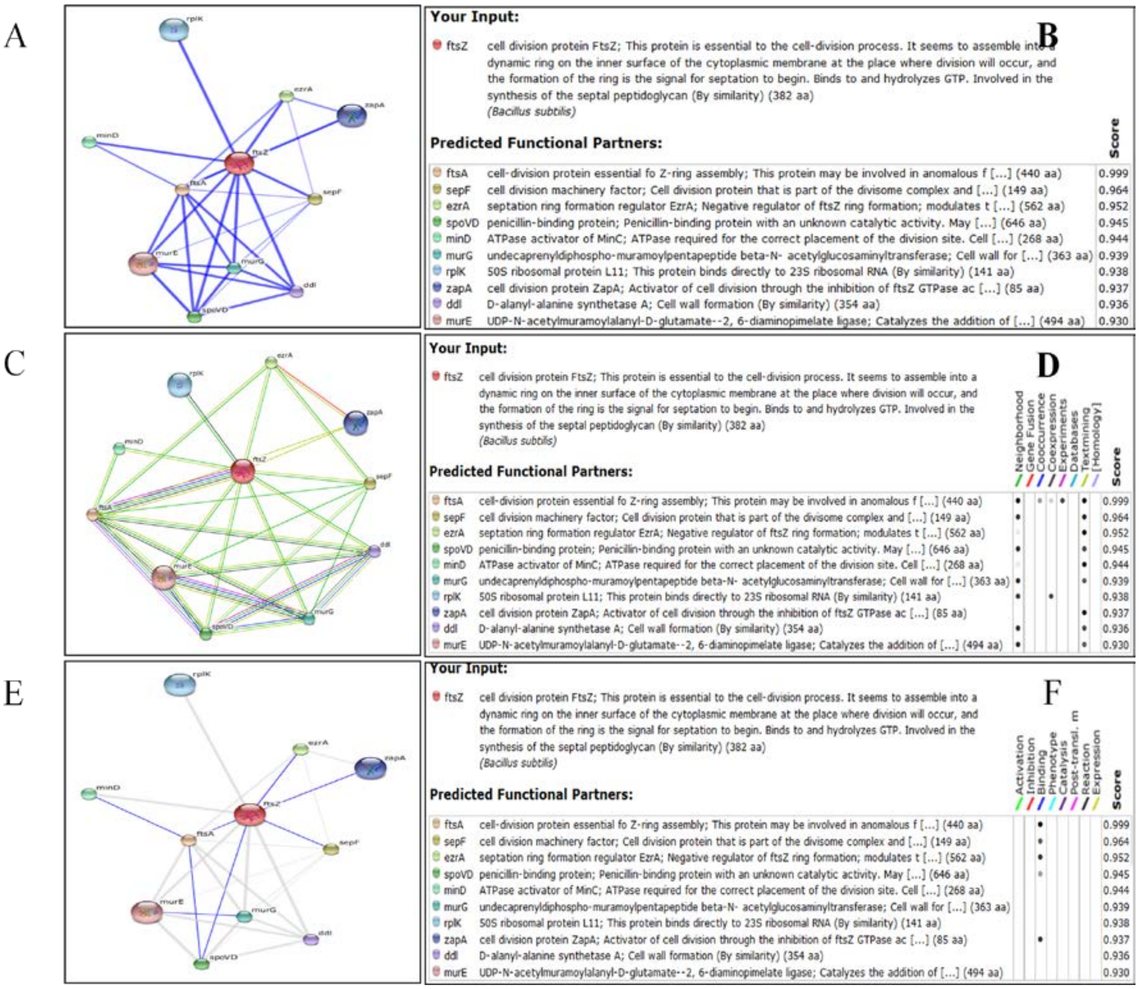
5. Details of the experiment conducted for characterizing proteins

6. Details of protein's secondary structure

7.Patterns occurring within a sequence and their analysis

b. Interaction network analysis

The study of interactions among the bio-molecules in cell is called interactome. Many prediction tools are designed to analyze the interaction network analysis among the proteins or other molecules and built the databases to analyze the data acquired from the proteomic studies. The interaction network can display the information of experimentally derived data, functional interactions, text mining, co-occurrence etc. The enormous data can be sorted into few groups accordingly to provide meaningful information. “String” is one of the online databases to provide the interaction networks among the proteins of interest and display the network with type of interaction and score (**Fig 1**).



B

Your Input:

ftsZ cell division protein FtsZ; This protein is essential to the cell-division process. It seems to assemble into a dynamic ring on the inner surface of the cytoplasmic membrane at the place where division will occur, and the formation of the ring is the signal for septation to begin. Binds to and hydrolyzes GTP. Involved in the synthesis of the septal peptidoglycan (By similarity) (382 aa)
(Bacillus subtilis)

Predicted Functional Partners:

		Score
ftsA	cell-division protein essential fo Z-ring assembly; This protein may be involved in anomalous f [...] (440 aa)	0.999
sepF	cell division machinery factor; Cell division protein that is part of the divisome complex and [...] (149 aa)	0.964
ezrA	septation ring formation regulator EzrA; Negative regulator of ftsZ ring formation; modulates t [...] (562 aa)	0.952
spoVD	penicillin-binding protein; Penicillin-binding protein with an unknown catalytic activity. May [...] (646 aa)	0.945
minD	ATPase activator of MinC; ATPase required for the correct placement of the division site. Cell [...] (268 aa)	0.944
murG	undecaprenyldiphospho-muramoylpentapeptide beta-N- acetylglucosaminyltransferase; Cell wall for [...] (363 aa)	0.939
rplK	50S ribosomal protein L11; This protein binds directly to 23S ribosomal RNA (By similarity) (141 aa)	0.938
zapA	cell division protein ZapA; Activator of cell division through the inhibition of ftsZ GTPase ac [...] (85 aa)	0.937
ddl	D-alanyl-alanine synthetase A; Cell wall formation (By similarity) (354 aa)	0.936
murE	UDP-N-acetylmuramoylalanyl-D-glutamate--2, 6-diaminopimelate ligase; Catalyzes the addition of [...] (494 aa)	0.930

D

Your Input:

ftsZ cell division protein FtsZ; This protein is essential to the cell-division process. It seems to assemble into a dynamic ring on the inner surface of the cytoplasmic membrane at the place where division will occur, and the formation of the ring is the signal for septation to begin. Binds to and hydrolyzes GTP. Involved in the synthesis of the septal peptidoglycan (By similarity) (382 aa)
(Bacillus subtilis)

Predicted Functional Partners:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
ftsA	cell-division protein essential fo Z-ring assembly; This protein may be involved in anomalous f [...] (440 aa)	●		●	●	●		●		0.999
sepF	cell division machinery factor; Cell division protein that is part of the divisome complex and [...] (149 aa)	●						●		0.964
ezrA	septation ring formation regulator EzrA; Negative regulator of ftsZ ring formation; modulates t [...] (562 aa)							●		0.952
spoVD	penicillin-binding protein; Penicillin-binding protein with an unknown catalytic activity. May [...] (646 aa)	●						●		0.945
minD	ATPase activator of MinC; ATPase required for the correct placement of the division site. Cell [...] (268 aa)							●		0.944
murG	undecaprenyldiphospho-muramoylpentapeptide beta-N- acetylglucosaminyltransferase; Cell wall for [...] (363 aa)	●						●		0.939
rplK	50S ribosomal protein L11; This protein binds directly to 23S ribosomal RNA (By similarity) (141 aa)	●			●					0.938
zapA	cell division protein ZapA; Activator of cell division through the inhibition of ftsZ GTPase ac [...] (85 aa)							●		0.937
ddl	D-alanyl-alanine synthetase A; Cell wall formation (By similarity) (354 aa)	●						●		0.936
murE	UDP-N-acetylmuramoylalanyl-D-glutamate--2, 6-diaminopimelate ligase; Catalyzes the addition of [...] (494 aa)	●						●		0.930

F

Your Input:

ftsZ cell division protein FtsZ; This protein is essential to the cell-division process. It seems to assemble into a dynamic ring on the inner surface of the cytoplasmic membrane at the place where division will occur, and the formation of the ring is the signal for septation to begin. Binds to and hydrolyzes GTP. Involved in the synthesis of the septal peptidoglycan (By similarity) (382 aa)
(Bacillus subtilis)

Predicted Functional Partners:

		Activation	Inhibition	Binding	Phenotype	Catalysis	Post-transl. m	Reaction	Expression	Score
ftsA	cell-division protein essential fo Z-ring assembly; This protein may be involved in anomalous f [...] (440 aa)			●						0.999
sepF	cell division machinery factor; Cell division protein that is part of the divisome complex and [...] (149 aa)			●						0.964
ezrA	septation ring formation regulator EzrA; Negative regulator of ftsZ ring formation; modulates t [...] (562 aa)			●						0.952
spoVD	penicillin-binding protein; Penicillin-binding protein with an unknown catalytic activity. May [...] (646 aa)			●						0.945
minD	ATPase activator of MinC; ATPase required for the correct placement of the division site. Cell [...] (268 aa)									0.944
murG	undecaprenyldiphospho-muramoylpentapeptide beta-N- acetylglucosaminyltransferase; Cell wall for [...] (363 aa)									0.939
rplK	50S ribosomal protein L11; This protein binds directly to 23S ribosomal RNA (By similarity) (141 aa)									0.938
zapA	cell division protein ZapA; Activator of cell division through the inhibition of ftsZ GTPase ac [...] (85 aa)			●						0.937
ddl	D-alanyl-alanine synthetase A; Cell wall formation (By similarity) (354 aa)									0.936
murE	UDP-N-acetylmuramoylalanyl-D-glutamate--2, 6-diaminopimelate ligase; Catalyzes the addition of [...] (494 aa)									0.930

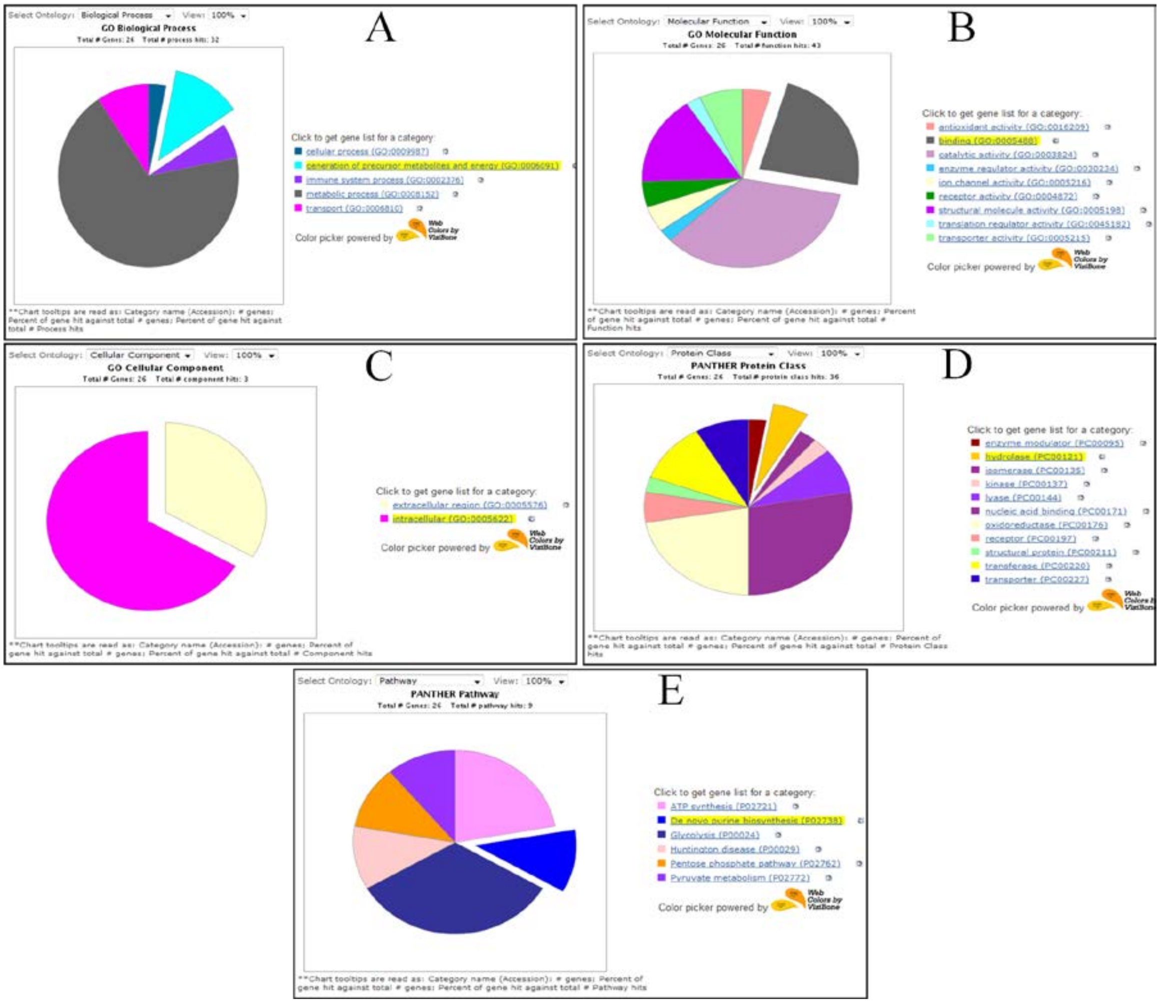
Fig 1. String interaction network in protein mode. (A & B) Confidence view where the thick line indicates the strong interaction and score is given in following image B.

(C & D) Evidence view where different colors indicate the different types of interaction and color coding is given in following image D.

(E, F) Action view where it mode of binding and color coding is provided in corresponding table F.

c. Metabolic pathway analysis

There are many pathway databases to construct the pathways from our input data. These databases were built over a period of time after mining the data from various literatures, databases etc. The widely used databases for constructing pathways or network includes Panther (Fig 2), DAVID (Fig 3), Reactome (Fig 4), KEGG (Fig 5) etc. Most of these databases are interconnected with other knowledge databases. In PANTHER, we need to input the gene/protein names or UniProt IDs as a query and the result display as molecular function, biological process, cellular component, Panther protein classes and Panther pathways. Data can be represented in different formats with its *p*-value. Similarly, DAVID, Reactome and KEGG databases need the same input files like panther. DAVID displays the results as gene functional classification chart, functional annotation clustering and gene list report; whereas Reactome and KEGG display metabolic or other physiological pathway network.



Select Ontology: Biological Process View: 100%

GO Biological Process

Total # Genes: 26 Total # process hits: 32

**Chart tooltips are read as: Category name (Accession): # genes: Percent of gene hit against total # genes: Percent of gene hit against total # Process hits

Select Ontology: Molecular Function View: 100%

GO Molecular Function

Total # Genes: 26 Total # function hits: 43

**Chart tooltips are read as: Category name (Accession): # genes: Percent of gene hit against total # genes: Percent of gene hit against total # Function hits

Select Ontology: Cellular Component View: 100%

GO Cellular Component

Total # Genes: 26 Total # component hits: 3

**Chart tooltips are read as: Category name (Accession): # genes: Percent of gene hit against total # genes: Percent of gene hit against total # Component hits

Select Ontology: Protein Class View: 100%

PANTHER Protein Class

Total # Genes: 26 Total # protein class hits: 36

**Chart tooltips are read as: Category name (Accession): # genes: Percent of gene hit against total # genes: Percent of gene hit against total # Protein Class hits

Select Ontology: Pathway View: 100%

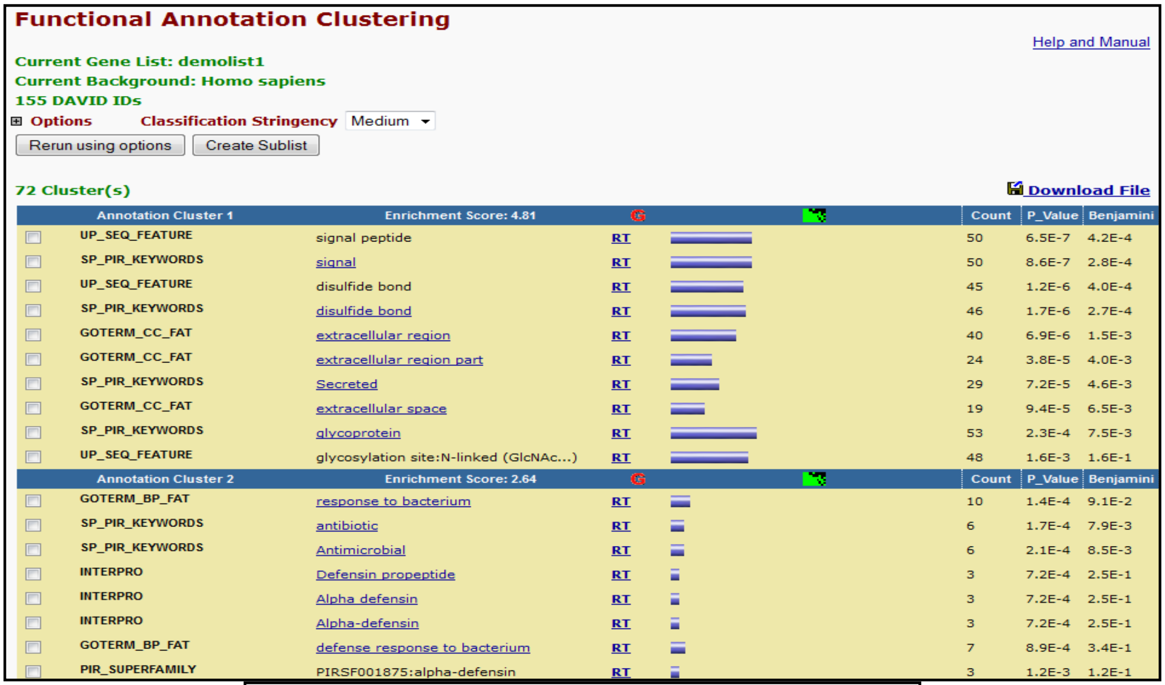
PANTHER Pathway

Total # Genes: 26 Total # pathway hits: 9

**Chart tooltips are read as: Category name (Accession): # genes: Percent of gene hit against total # genes: Percent of gene hit against total # Pathway hits

A
Click to get gene list for a category:
cellular process (GO:0009987)
generation of precursor metabolites and energy (GO:0006091)
immune system process (GO:0002376)
metabolic process (GO:0008152)
transport (GO:0006810)
Color picker powered by Web Colors by VisiBone
B
Click to get gene list for a category:
antioxidant activity (GO:0016209)
binding (GO:0005488)
catalytic activity (GO:0003824)
enzyme regulator activity (GO:0030234)
ion channel activity (GO:0005216)
receptor activity (GO:0004872)
structural molecule activity (GO:0005198)
translation regulator activity (GO:0045182)
transporter activity (GO:0005215)
Color picker powered by Web Colors by VisiBone
C
Click to get gene list for a category:
extracellular region (GO:0005576)
intracellular (GO:0005622)
Color picker powered by Web Colors by VisiBone
D
Click to get gene list for a category:
enzyme modulator (PC00095)
hydrolase (PC00121)
isomerase (PC00135)
kinase (PC00137)
lyase (PC00144)
nucleic acid binding (PC00171)
oxidoreductase (PC00176)
receptor (PC00197)
structural protein (PC00211)
transferase (PC00220)
transporter (PC00227)
Color picker powered by Web Colors by VisiBone
E
Click to get gene list for a category:
ATP synthesis (P02721)
De novo purine biosynthesis (P02738)
Glycolysis (P00024)
Huntington disease (P00029)
Pentose phosphate pathway (P02762)
Pyruvate metabolism (P02772)
Color picker powered by Web Colors by VisiBone

Fig 2. Panther pathway analysis. (A) GO Biological process pie chart indicate the no. of query proteins involved in each biological process. (B) GO molecular process pie chart shows the number of proteins involved in molecular activity of the cell. (C) GO cellular component pie chart shows the localization of query proteins in the cell. (D) Panther protein class shows the distribution of proteins into the panther classes. (E) Panther pathway class shows the involvement of the query proteins into biological pathways.



Functional Annotation Clustering

Help and Manual

Current Gene List: demolist1
Current Background: Homo sapiens
155 DAVID IDs
⊞ Options Classification Stringency Medium
Rerun using options | Create Sublist

72 Cluster(s) Download File

	Annotation Cluster 1	Enrichment Score: 4.81			Count	P_Value	Benjamini
□	UP_SEQ_FEATURE	signal peptide	RT		50	6.5E-7	4.2E-4
□	SP_PIR_KEYWORDS	signal	RT		50	8.6E-7	2.8E-4
□	UP_SEQ_FEATURE	disulfide bond	RT		45	1.2E-6	4.0E-4
□	SP_PIR_KEYWORDS	disulfide bond	RT		46	1.7E-6	2.7E-4
□	GOTERM_CC_FAT	extracellular region	RT		40	6.9E-6	1.5E-3
□	GOTERM_CC_FAT	extracellular region part	RT		24	3.8E-5	4.0E-3
□	SP_PIR_KEYWORDS	Secreted	RT		29	7.2E-5	4.6E-3
□	GOTERM_CC_FAT	extracellular space	RT		19	9.4E-5	6.5E-3
□	SP_PIR_KEYWORDS	glycoprotein	RT		53	2.3E-4	7.5E-3
□	UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT		48	1.6E-3	1.6E-1
	Annotation Cluster 2	Enrichment Score: 2.64			Count	P_Value	Benjamini
□	GOTERM_BP_FAT	response to bacterium	RT		10	1.4E-4	9.1E-2
□	SP_PIR_KEYWORDS	antibiotic	RT		6	1.7E-4	7.9E-3
□	SP_PIR_KEYWORDS	Antimicrobial	RT		6	2.1E-4	8.5E-3
□	INTERPRO	Defensin propeptide	RT		3	7.2E-4	2.5E-1
□	INTERPRO	Alpha defensin	RT		3	7.2E-4	2.5E-1
□	INTERPRO	Alpha-defensin	RT		3	7.2E-4	2.5E-1
□	GOTERM_BP_FAT	defense response to bacterium	RT		7	8.9E-4	3.4E-1
□	PIR_SUPERFAMILY	PIRSF001875:alpha-defensin	RT		3	1.2E-3	1.2E-1

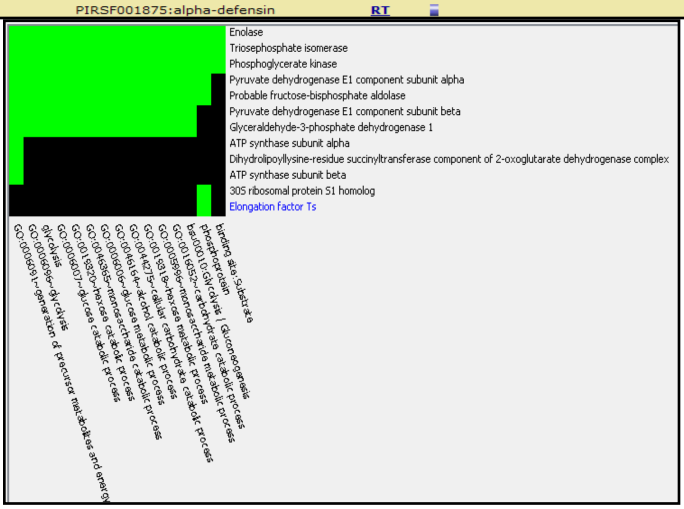


Fig 3. DAVID functional annotation clustering tool categorizes the query proteins into clusters and give corresponding score and cluster image. It also provides the statistical analysis information such as P-value.

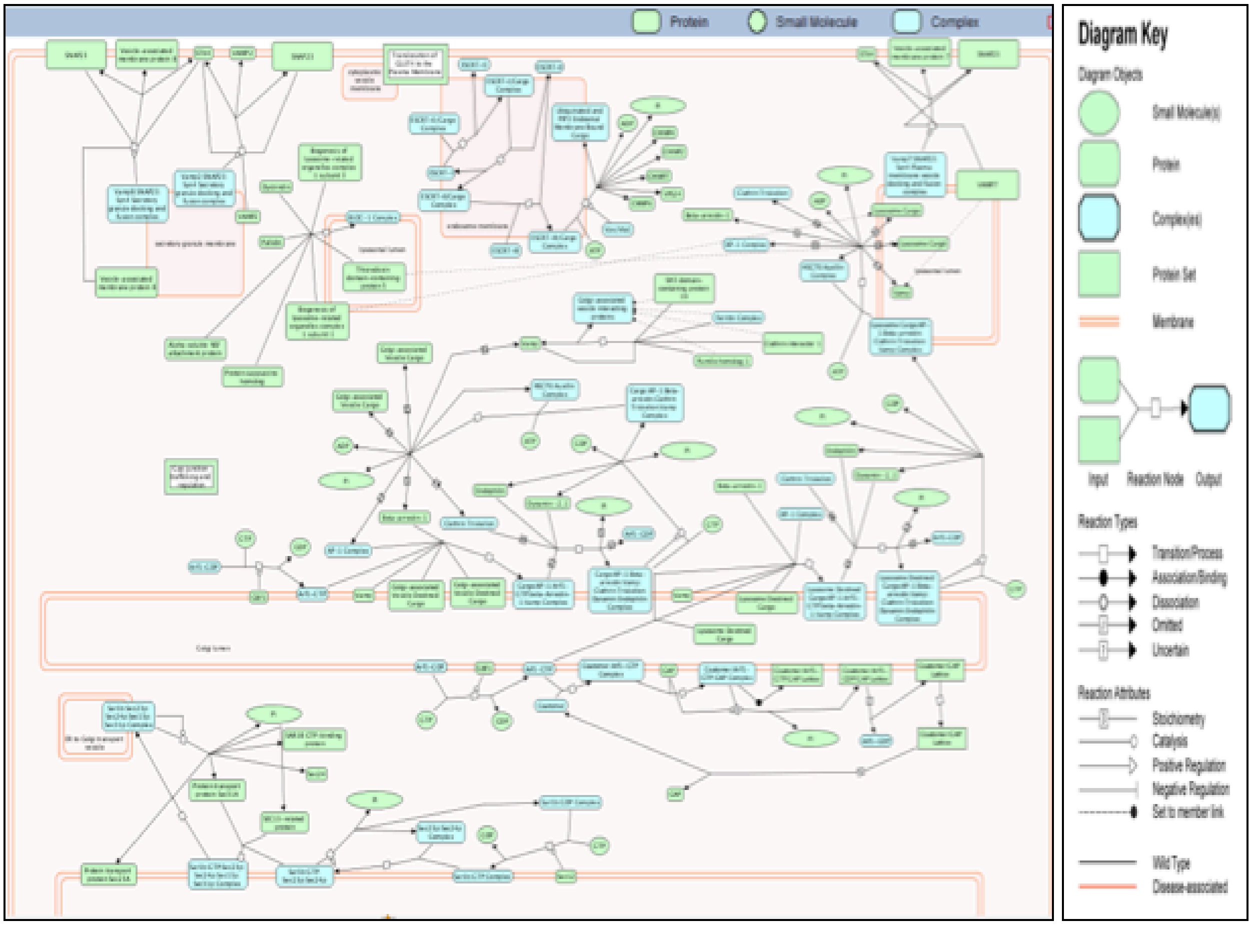


Fig 4. Reactome pathway analysis tool used to display the query molecules in biological pathway. This pathway shows both the metabolic network and protein network with proteins, small molecules, complexes and also provides information such as transition, catalytic reaction, binding etc.

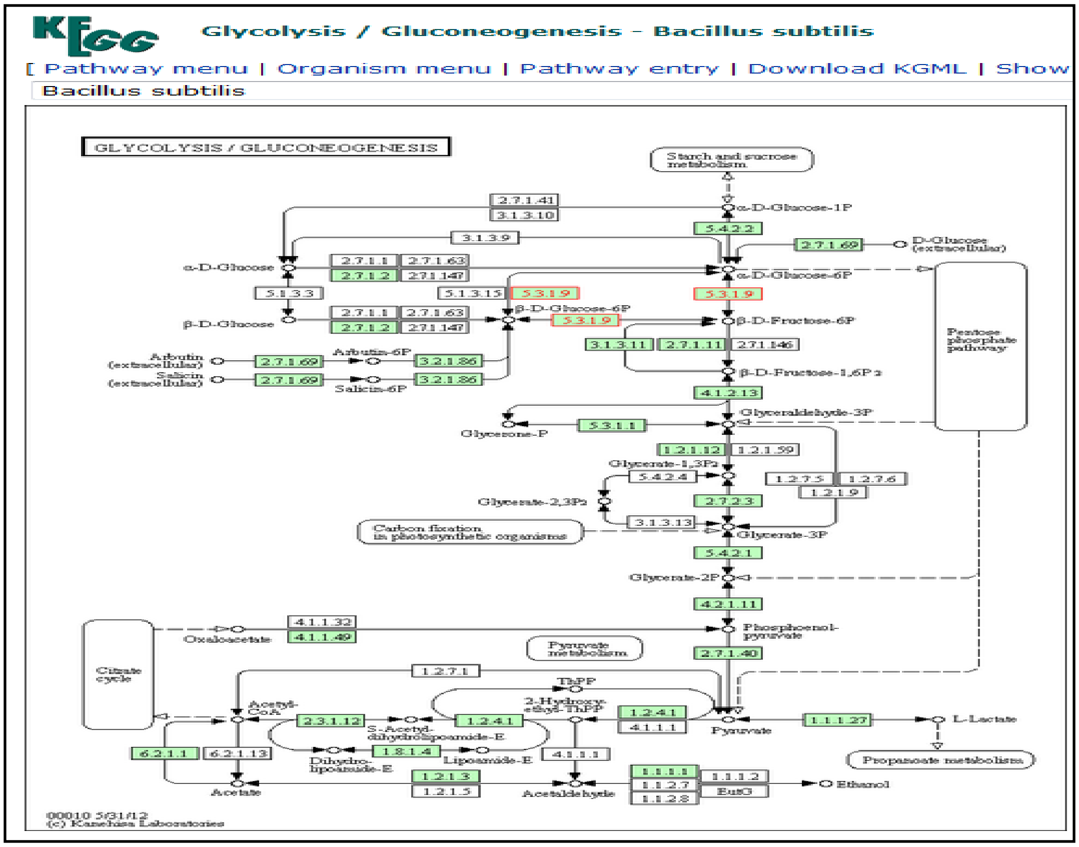


Fig 5. KEGG metabolic pathway. It shows the glycolytic/ gluconeogenesis converting glucose to pyruvate or lactate or ethanol by series of enzymatic reaction. The enzyme that comes from the query is highlighted in red color as shown in figure.

d. **3D structure modeling**

The 3D structure of proteins discovered so far is very limited and homology modeling is one of the ways to model the 3D structure of protein. I-TASSER server provides one of the best online platforms for structural prediction simulations. TM score and RMSD are considered for the good prediction along with C-score. I-TASSER models the structure using the combination of hidden Markov model, PSI-BLAST profiles, the Needleman-Wunsch and Smith-Waterman alignment algorithms for the backbone structure model and *ab initio* modeling for the loops. Further validation of each protein can verify using Verify_3D, WHAT_IF server. Verify_3D give us the distribution of amino acids in Ramachandran plot statistics whether amino acids are in accepted regions or not for further studies.

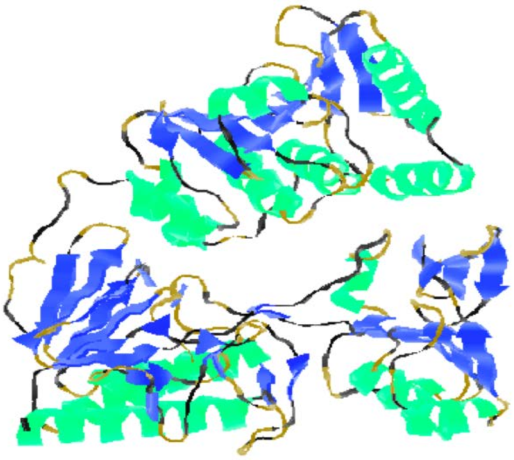


Fig 6. I-TASSER model prediction of protein. This protein is modeled by using I-TASSER server and the structure shows the arrangement of amino acid sequence in 3D pattern with the help of functional prediction.

e. Molecular Docking

3D structures can be obtained from the PDB databases if available; otherwise one needs to model protein using homology modeling. Accurate starting structures are a pre-requisite for successful docking studies. Prior to the docking, preliminary processing of protein and ligand structure is performed to add the missing hydrogen atoms and deleting the non-interacting heteroatom and water molecules from the ligand and protein. Docking can be performed either blindly by selecting the grid to whole protein or by selecting the grid for specific part of the protein. Tools such as autoDock provide docking energy and inhibitor constants (Fig 7).

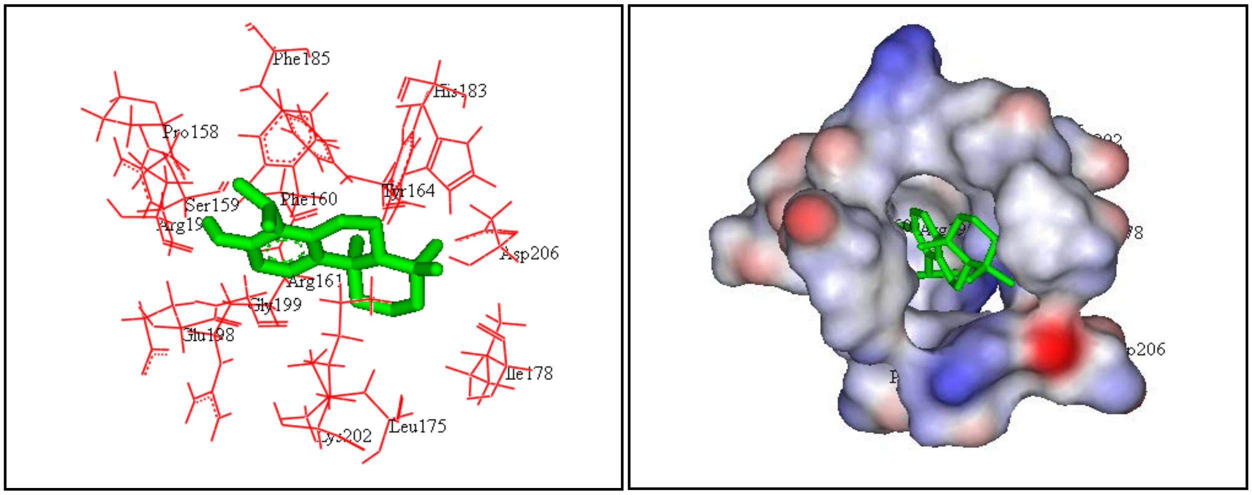


Fig 7. Docking of the ligand and protein (A) The stick model generated from AutoDock showing the docking of ligand in the protein with hydrogen bonds. (B) Crystal structure to show the docking of ligand to the protein. The ligand molecule is shown in green color and amino acids are shown in red color.

Table 4: Bioinformatic tools/databases for network, pathway and modeling of the proteins (Tsui IF et al., 2007)

S. No	**Name of the database/tool**	**Description**	**URL link**
Interaction network			
1	String database	Both physical and functional interactions can display with score	http://string-db.org/
2	InterDom	Interactions among the putative domains	http://interdom.i2r.a-star.edu.sg/
3	IntAct	Information of protein interaction among the literature curated or directly submitted proteins	http://www.ebi.ac.uk/intact/
4	MINT	Display the interaction among the experimentally proved proteins	http://mint.bio.uniroma2.it/mint/Welcome.do
5	HPRD	Interaction information which is provided experimentally and also provides information regarding domain architecture, PTM and its connection with disease	http://www.hprd.org/
Metabolic pathway analysis			
1	Panther	Provides information of pathways besides it also provide the gene ontology terms	http://www.pantherdb.org/
2	DAVID	Provides information regarding pathways and also gives the functional annotation information	http://david.abcc.ncifcrf.gov/
3	KEGG	Provides the pathway analysis of wide range of organisms	http://www.genome.jp/kegg/
4	MetaCyc	Experimentally proved, non redundant pathway database	http://metacyc.org/
5	Reactome	Manually curated and peer reviewed pathway database	http://www.reactome.org/ReactomeGWT/entrypoint.html
Modeling and Docking			
1	ProteinModelPortal	Comparative modeling protein with the help of computational methods	http://www.proteinmodelportal.org/
2	ModBase	Comparative modeling of the protein with the help of theoretically calculated modeling	http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi
3	I-TASSER	Online 3D modeling with the help of multi-threading alignment	http://zhanglab.ccmb.med.umich.edu/I-TASSER/

4	AutoDock	To find docking site on protein for small molecule by making grid	http://autodock.scripps.edu/
5	Glide	Docking of protein-ligand molecules	http://www.schrodinger.com/
6	GOLD	Docking of protein-ligand	http://www.ccdc.cam.ac.uk/products/life_sciences/gold/

Reference:

- Snel B, Lehmann G, Bork P, Huynen MA.STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene.Nucleic Acids Res. 2000 Sep 15;28(18):3442-4.
- Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57
- Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13
- Palagi PM, Hernandez P, Walther D, Appel RD.Proteome informatics I: bioinformatics tools for processing experimental data.Proteomics. 2006 Oct;6(20):5435-44.
- Tsui IF, Chari R, Buys TP, Lam WL.Public databases and software for the pathway analysis of cancer genomes.Cancer Inform. 2007 Dec 12;3:379-97.
- Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M et al. The IntAct molecular interaction database in 2012.Nucleic Acids Res. 2012 Jan;40(Database issue):D841-6.
- Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G.MINT, the molecular interaction database: 2009 update.Nucleic Acids Res. 2010 Jan;38(Database issue):D532-9.

- Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S et al., Human Protein Reference Database--2009 update.Nucleic Acids Res. 2009 Jan;37(Database issue):D767-72.
- Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A et al., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2012 Jan;40(Database issue):D742-53.
- Arnold K, Kiefer F, Kopp J, Battey JN, Podvinec M, Westbrook JD, Berman HM, Bordoli L, Schwede T.The Protein Model Portal.J Struct Funct Genomics. 2009 Mar;10(1):1-8.
- Yang Zhang. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, vol 9, 40 (2008)