SciMedCentral
JSM Bioinformatics, Genomics and Proteomics

Research Article

In silico Analysis of Protein

Nishtha Singh, Sonal Upadhyay, Ankur Jaiswar and Nidhi Mishra*

Applied Science Division, Indian Institute of Information Technology, India

***Corresponding author**

Nidhi Mishra, CC-III, Indian Institute of Information Technology Allahabad, Devghat, Jhalwa, Allahabad-211012, Uttar Pradesh, India, Tel: 9450900033; Email: nidhimishra@iiita.ac.in

Submitted: 28 June 2016

Accepted: 31 August 2016

Published: 07 September 2016

Copyright

© 2016 Mishra et al.

OPEN ACCESS

Keywords

- Protein analysis
- In Silico
- Function
- Sequence
- Tools

Abstract

Proteins are inimitable as principal functional agent of living system. Therefore, comprehension of protein sequence and structure and its correlation with its function is equivalent to deciphering almost all of fundamental features of any biological/living system. A treasure of In silico tools is accessible for analysis of protein. Understanding and regeneration of protein function requires comprehension of reliance between protein sequence and its structure, its localization in cell and its interaction with other functional partners. This review provides an insight for various tools for In silico analysis of protein.

ABBREVIATIONS

UNIPROT: Universal Protein Resource; FASTA: FAST Alignment; BLAST: Basic Local Alignment Search Tool; Blastn: Nucleotide Basic Local Alignment Search Tool; Blastp: Protein Basic Local Alignment Search Tool; psi-BLAST: Position-Specific Iterated Basic Local Alignment Search Tool; SOPMA: Self Optimized Prediction Method with Alignment; GOR4:Garnier-Osguthorpe-Robson 4; I-TASSER: Iterative Threading Assembly Refinement; RAMPAGE: Rama chandran Plot Assessment; CASTp: Computer Atlas of Surface Topology of Proteins; T-COFFEE: Tree-based Consistency Objective Function For alignment Evaluation

INTRODUCTION

Tools for *In silico* analysis of protein are of major significance for making use of data for development of drug and health care. Proteins are complex macromolecules which have essence role to cellular activity. They increase biological processes or reactions by expression of catalytic activity. Thus, analysis and knowledge of protein is substantial for determining the protein function and to understand their role and mechanism in human body [1].

Protein sequence retrieval

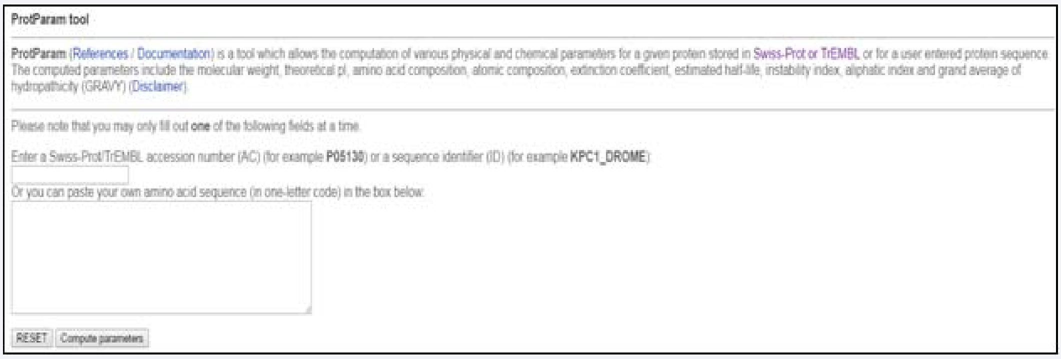
Sequence of protein is retrieved from UniProt (Universal Protein Resource) which is a freely accessible database which contains data of proteins. The required protein sequence is retrieved through their accession number and it is in FASTA format. This sequence is further utilized for primary and secondary structure analysis. To obtain sequence similarity of query protein sequence with the known protein structure BLAST (Basic Local Alignment Search Tool). This tool compares protein or nucleotide query sequence to database of sequences and evaluates analytical importance of matches tool is used. Blast is group of programs including blastn (searches DNA sequences against DNA database) ,blastp (for a given protein query sequence returns similar sequence from database of protein sequences), psi-blast (it establishes distant relationship between proteins), blastx (comparison of six-frame translation product of query sequence of nucleotide against database of protein sequence) , tblastx (comparison of six-frame translation product of query sequence of nucleotide against database of nucleotide sequence) andtblastn (comparison of protein query sequence against six reading frames of database of nucleotide sequence) [2].

Primary structure prediction

The primary structure analysis of protein and physicochemical depiction is done using ProtParam tool from ExPasy (Expert Protein analysis system) (Figure 1). And for this number of amino acids, molecular weight, theoretical PI, amino acid composition, total number of atoms, extinction coefficient, instability index, aliphatic index, grand average of hydropathicity, total number of negatively and positively charged residue and estimated half -life is computed. If instability index is below 40 then the protein is predicted as stable and above 40 it may be unstable, the aliphatic index ascertains the thermal stability based on amino acids alanine, valine and leucine of globular proteins. If Gravy value is low then this deciphers that there is better interaction between protein and water. Half- life predicts the amount of time taken by half of protein to disappear after its production in cell, the extinction coefficient predicts amount of light absorbed by a protein at certain wavelength. If amino acids are basic in nature then PI will be high and if amino acids are acidic then the PI will be low [3].

Secondary structure prediction

The secondary structure of protein is predicted using SOPMA (Self- Optimized Prediction Method with Alignment) tool (Figure 2). This tool evaluates the percentage of alpha helices, extended strand, beta turn and random coils. It uses homology methodology. According to percentage secondary structure is predicted. By default it shows output width as 70 which means there will be 70 amino acids in each line. Number of conformational states



ProtParam tool

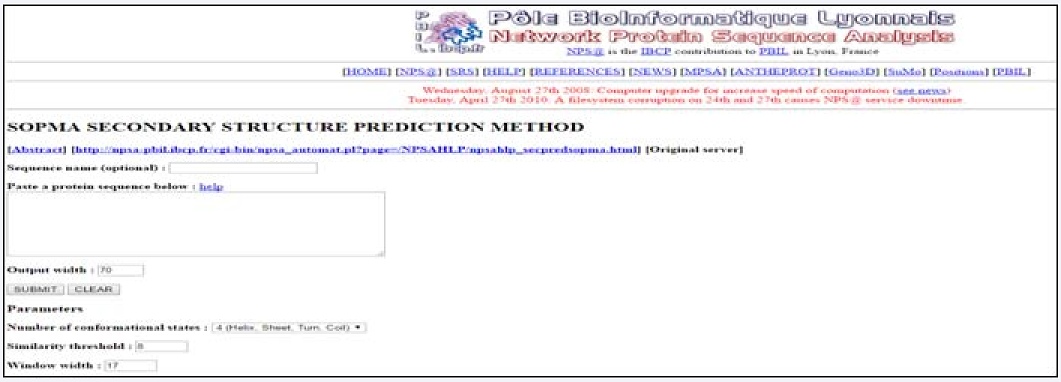
ProtParam (References / Documentation) is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered protein sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) (Disclaimer).

Please note that you may only fill out **one** of the following fields at a time.

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **KPC1_DROME**):

Or you can paste your own amino acid sequence (in one-letter code) in the box below:

Figure 1 Homepage of ProtParam: Paste the sequence in the given box or enter accession number.



Pôle BioInformatique Lyonnais
Network Protein Sequence Analysis
NPS@ is the IBCP contribution to PBIL in Lyon, France

[HOME] [NPS@] [SRS] [HELP] [REFERENCES] [NEWS] [MPSA] [ANTHEPROT] [Geno3D] [SuMo] [Positions] [PBIL]

Wednesday, August 27th 2008: Computer upgrade for increase speed of computation (see news)
Tuesday, April 27th 2010: A filesystem corruption on 24th and 27th causes NPS@ service downtime

SOPMA SECONDARY STRUCTURE PREDICTION METHOD

[Abstract] [http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/HLP/npsahlp_secpredsopma.html] [Original server]

Sequence name (optional) :

Paste a protein sequence below : help

Output width : 70

Parameters

Number of conformational states : 4 (Helix, Sheet, Turn, Coil)

Similarity threshold : 8

Window width : 17

Figure 2 Homepage of SOPMA: Paste the sequence in the given box and choose parameters accordingly.

I-TASSER
Protein Structure & Function Predictions

I-TASSER Online Server

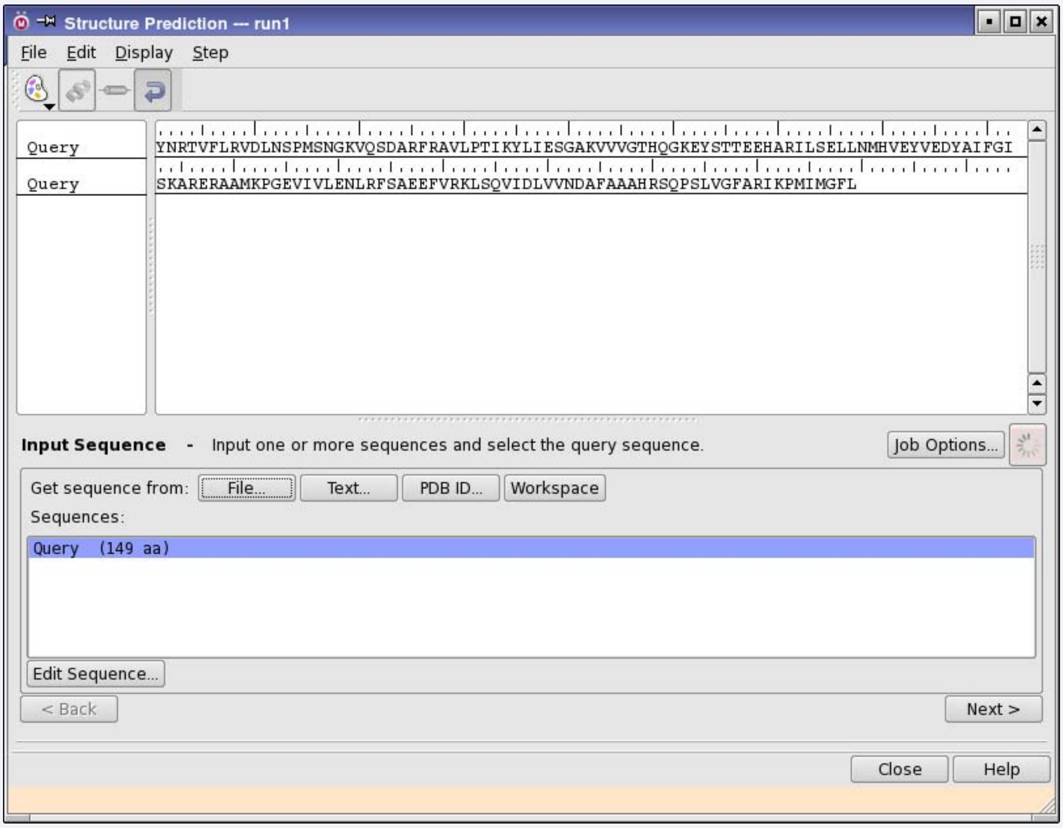
Figure 3 Homepage of I-TASSER: Paste the sequence in the given box.

can be given as either 4(Helix, Sheet, Turn, Coil) or as 3 (Helix, Sheet, Coil). The first graph of sopma result anticipates the prediction and the second graph consist of outcome curves for all of the predicted states. Other tool that can be used for secondary structure prediction is GOR IV [4].

Tertiary structure prediction

Based on availability of template sequence modeling can be Comparative/Homology or Threading or Ab Initio modeling. I – TASSER (Iterative Threading Assembly Refinement) is one of the best tool for automated protein structure prediction [5]. Solvent Accessibility, Normalized B-factor is predicted [6]. Tools provides top 10 threading templates used by I-TASSER and predicts top (Figure 3) ranked 5 models based on C-score, Estimated TM – score and Estimated RMSD. Enzyme commission number and active sites are also predicted [7].

Prime module of Schrödinger LLC, New York, 2014: Taking



```
Query   YNRTVFLRVDLNSPMSNGKVQSDARFRAVLPTIKYLIESGAKVVVGTHQGKEYSTTEEHARILSELLNMHVEYVEDYAIFGI
Query   SKARERAAMKPGEVIVLENLRFSAEEFVRKLSQVIDLVVNDAFAAAHRSQPSLVGFARIKPMIMGFL
```

Input Sequence - Input one or more sequences and select the query sequence.

Sequences:

Query (149 aa)

Figure 4 Input sequence in the structure prediction panel.

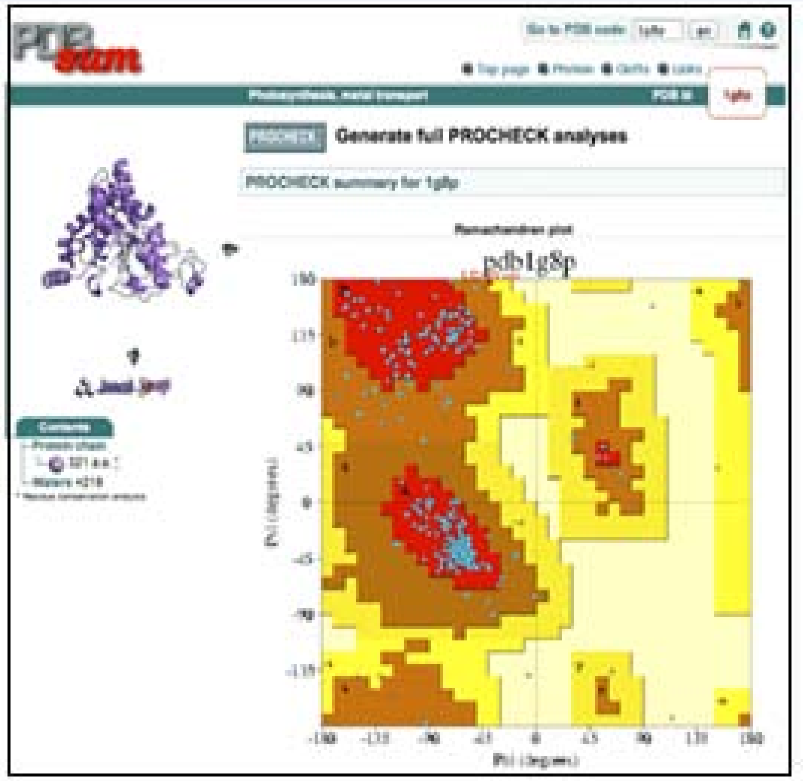


Figure 5 Ramachandran plot of procheck.

into account the degree of similarity in sequence protein can be modeled using this tool. Alignment of sequence can be ameliorated manually and homology methodology is utilized to build the structure of protein [8] (Figure 4). Other tools used are SWISS-MODEL, Modeller for homology modeling.

Validation of predicted models

PROCHECK is a tool to substantiate the spatial arrangement (stereochemistry) of protein structures. The output of this program is number of plots which are in PostScript format (Figure 5). The plot generated by this program is Ramachandran plot which is a plot of phi-psi torsional angles where darkest region is the most favoured region consisting of more than 90% of residues. It also generates Ramachandran plot based on types of residues [9].

RAMPAGE is used to check the stereo chemical properties of protein structure whether it is available through experiments or has been modeled (Figure 6). The plot generated by RAMPAGE provides the percentage of residues in various regions like favoured region, allowed region and outlier region. The more number of residues in favoured region the more stable is the protein [10].

Phylogenetic Tree

Multiple sequence alignment is an important requirement for additional evaluation of families of protein such as comparative modeling and phylogenetic reestablishment. T-COFFEE (Tree based Consistency Objective Function for alignment Evaluation) is for multiple sequence alignment (Figure 7). T-COFFEE is a program to calculate, manipulate and analyse multiple alignments of RNA (ribonucleic acid), DNA (deoxyribonucleic acid) and protein structures. Minimum of two sequences in supported format is entered as an input or file of sequence can be uploaded [11].

Prediction of Active Site of protein

Pockets present on surface of protein and amino acid residues present in those pockets are substantial for generating physiochemical properties which are required for protein to perform its function. CASTp is an online tool which analyze the active site of the protein and the amino acid that are present in those sites (Figure 8). It provides information of those amino acid residues that would be binding with ligand [12].

'SiteMap', Schrödinger LLC, New York, 2014 is also used to predict the active binding site (Figure 9). Site Score is generated and site maps are ranked. Further these site maps are utilized for generating grid of receptor [13].

Predicting phosphorylation sites

Phosphorylation is a vital procedure through which signaling pathways function. The removal or addition of phosphate group may result in alteration in function of protein and its localization. Three major amino acid residues namely Serine, Threonine and Tyrosine are mostly phosphorylated, as they contain hydroxyl group in their side chain and thus are capable of binding phosphate group. NetPhos server is a tool to predict phosphorylation site at threonine, serine and tyrosine (Figure 10). The result of NetPhos consists of three parts. Firstly result consists of length and amino acid that are in the sequence. If the amino acid residue is predicted as not phosphorylated (score is below threshold level) then the position is represented as (.) and if it is phosphorylated (score is above threshold level) then residues are marked as 'T', 'S' and 'Y'. Prediction for residue is delineated in the second part. And a graph describes the prediction in the third part.

Predicting protein ubiquitination sites

Protein ubiquitination is one of the most vital post-translational modifications by covalent attachment of ubiquitin to lysine residues (Figure 11). mUbiSiDa is a universal database for ubiquitination of proteins. The tool provides following functions: advanced retrieval, browse resource and blast search

Prediction of methylation and acetylation

Prediction of potential methylation and acetylation of protein

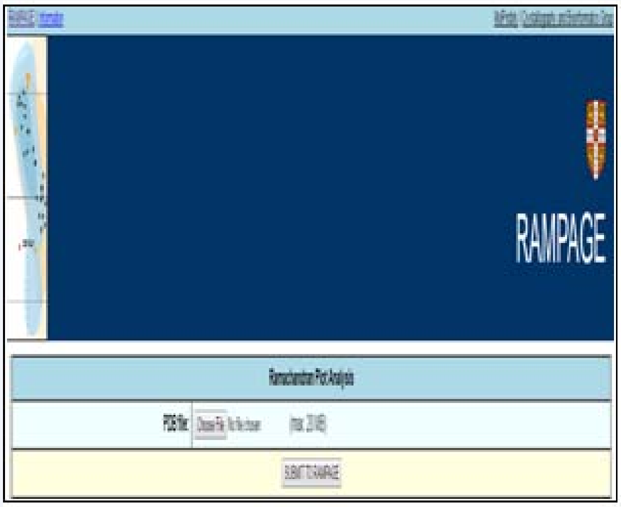
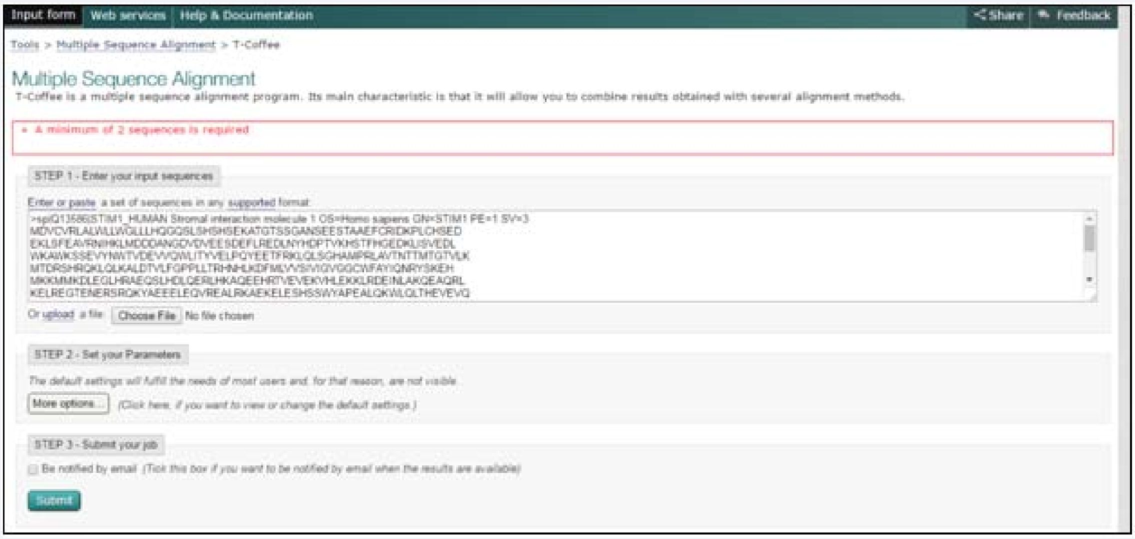


Figure 6 Homepage of RAMPAGE: Upload PDB file.



Tools > Multiple Sequence Alignment > T-Coffee

Multiple Sequence Alignment

T-Coffee is a multiple sequence alignment program. Its main characteristic is that it will allow you to combine results obtained with several alignment methods.

- A minimum of 2 sequences is required

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

```
>sp|Q13586|STIM1_HUMAN Stromal interaction molecule 1 OS=Homo sapiens GN=STIM1 PE=1 SV=3
MDVCVRLALWLLWGLLLHQGQSLSHSHSEKATGTSSGANSEESTAAEFCRIDKPLCHSED
EKLSFEAVRNIHKLMDDDANGDVDVEESDEFLREDLNYHDPTVKHSTFHGEDKLISVEDL
WKAWKSSEVYNWTVDEVVQWLITYVELPQYEETFRKLQLSGHAMPRLAVTNTTMTGTVLK
MTDRSHRQKLQLKALDTVLFGPPLLTRHNHLKDFMLVVSIVIGVGGCWFAYIQNRYSKEH
MKKMMKDLEGLHRAEQSLHDLQERLHKAQEEHRTVEVEKVHLEKKLRDEINLAKQEAQRL
KELREGTENERSRQKYAEEELEQVREALRKAEKELESHSSWYAPEALQKWLQLTHEVEVQ
```

Or upload a file: Choose File No file chosen

STEP 2 - Set your Parameters

The default settings will fulfill the needs of most users and, for that reason, are not visible.

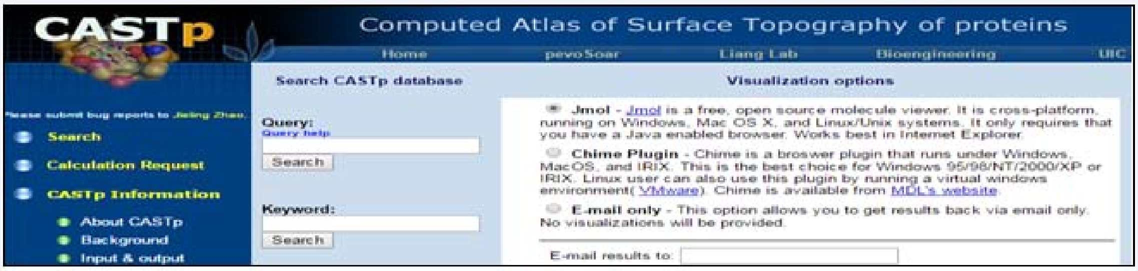
More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

Figure 7 Homepage of T-COFFEE, Paste the sequence in the given box, set parameters if required and submit your job.



Search CASTp database

Query:
Query help

Search

Keyword:

Search

Visualization options

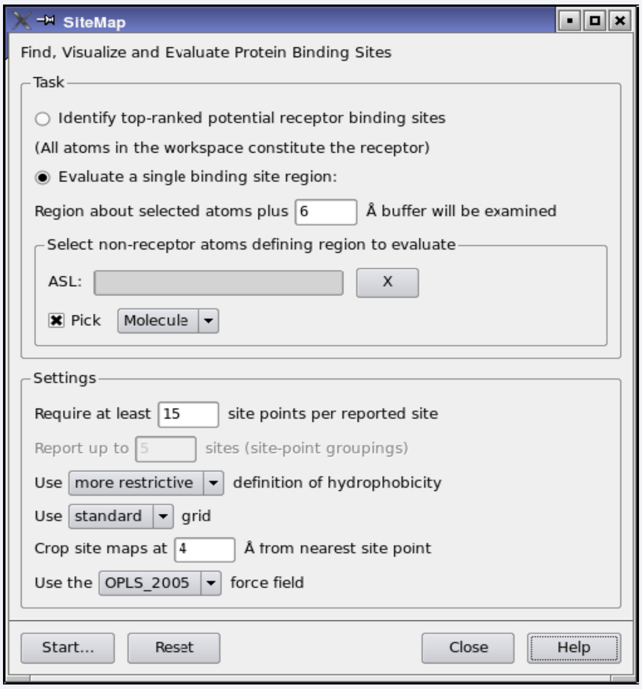
● **Jmol** - Jmol is a free, open source molecule viewer. It is cross-platform, running on Windows, Mac OS X, and Linux/Unix systems. It only requires that you have a Java enabled browser. Works best in Internet Explorer.

○ **Chime Plugin** - Chime is a broswer plugin that runs under Windows, MacOS, and IRIX. This is the best choice for Windows 95/98/NT/2000/XP or IRIX. Linux user can also use this plugin by running a virtual windows environment(VMware). Chime is available from MDL's website.

○ **E-mail only** - This option allows you to get results back via email only. No visualizations will be provided.

E-mail results to:

Figure 8 Homepage of CASTp.



Find, Visualize and Evaluate Protein Binding Sites

Task

○ Identify top-ranked potential receptor binding sites

(All atoms in the workspace constitute the receptor)

● Evaluate a single binding site region:

Region about selected atoms plus 6 Å buffer will be examined

Select non-receptor atoms defining region to evaluate

ASL: X

☒ Pick Molecule

Settings

Require at least 15 site points per reported site

Report up to 5 sites (site-point groupings)

Use more restrictive definition of hydrophobicity

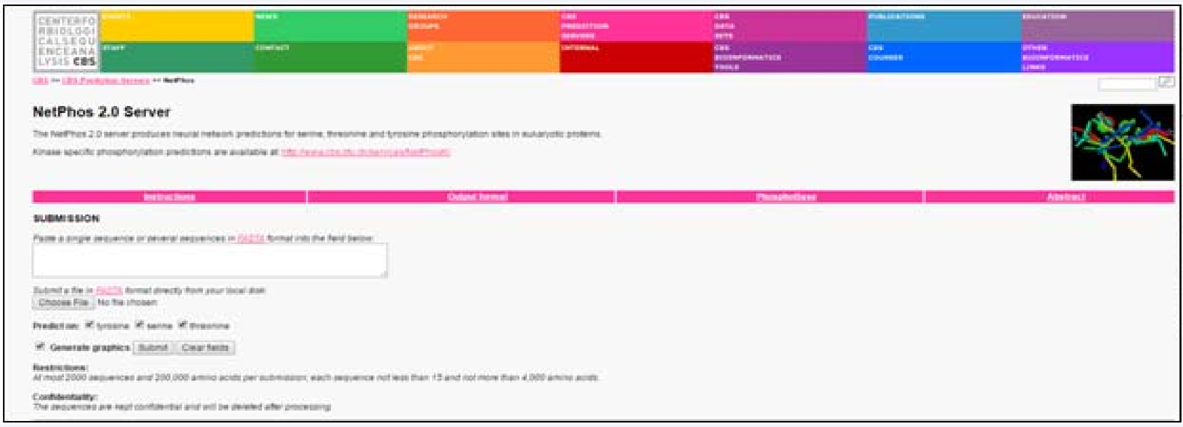
Use standard grid

Crop site maps at 4 Å from nearest site point

Use the OPLS_2005 force field

Start... Reset Close Help

Figure 9 SiteMap panel: Select the parameters accordingly and run the job.



NetPhos 2.0 Server

The NetPhos 2.0 server produces neural network predictions for serine, threonine and tyrosine phosphorylation sites in eukaryotic proteins.

Kinase specific phosphorylation predictions are available at http://www.cbs.dtu.dk/services/NetPhosK/

SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:

Submit a file in FASTA format directly from your local disk:
Choose File No file chosen

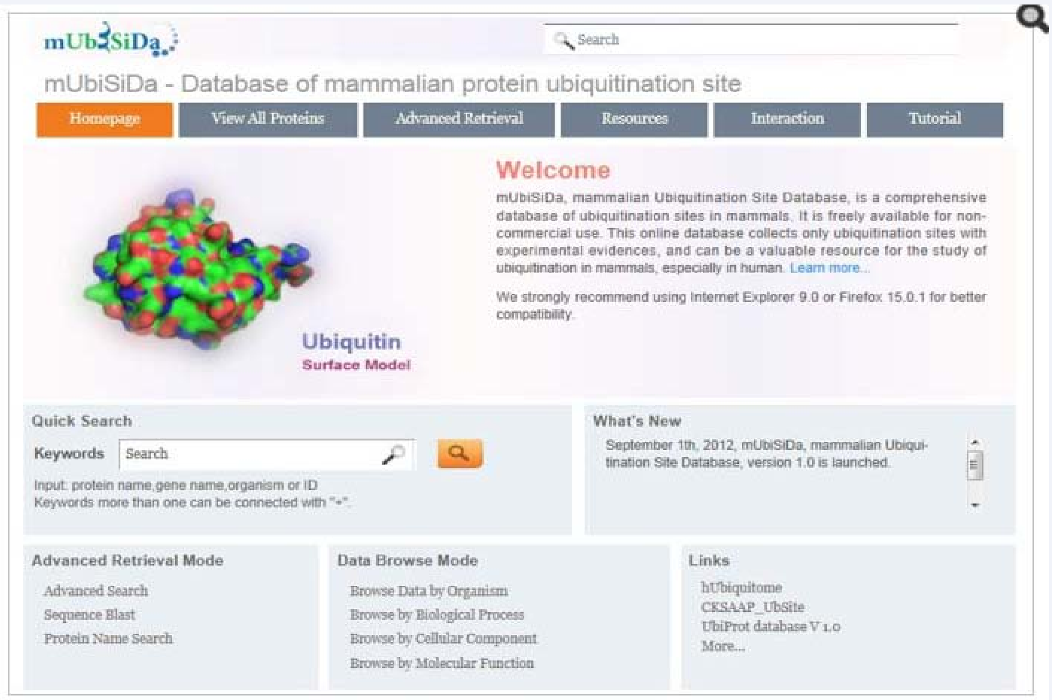
Predict on: ☑ tyrosine ☑ serine ☑ threonine

☑ Generate graphics Submit Clear fields

Restrictions:
At most 2000 sequences and 200,000 amino acids per submission; each sequence not less than 15 and not more than 4,000 amino acids.

Confidentiality:
The sequences are kept confidential and will be deleted after processing.

Figure 10 Homepage of NetPhos server: Paste the sequence in the given box and submit the job.



mUbiSiDa - Database of mammalian protein ubiquitination site

Welcome

mUbiSiDa, mammalian Ubiquitination Site Database, is a comprehensive database of ubiquitination sites in mammals. It is freely available for non-commercial use. This online database collects only ubiquitination sites with experimental evidences, and can be a valuable resource for the study of ubiquitination in mammals, especially in human. Learn more...

We strongly recommend using Internet Explorer 9.0 or Firefox 15.0.1 for better compatibility.

Quick Search

Keywords Search

Input: protein name,gene name,organism or ID
Keywords more than one can be connected with "+".

What's New

September 1th, 2012, mUbiSiDa, mammalian Ubiquitination Site Database, version 1.0 is launched.

Advanced Retrieval Mode

Advanced Search
Sequence Blast
Protein Name Search

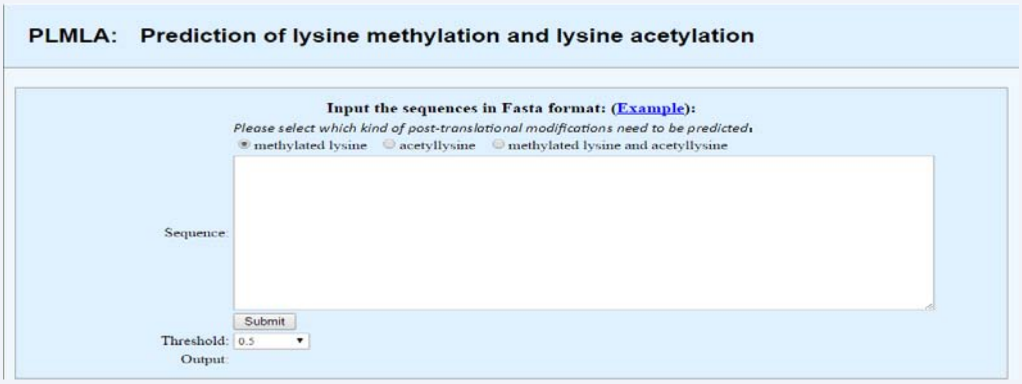
Data Browse Mode

Browse Data by Organism
Browse by Biological Process
Browse by Cellular Component
Browse by Molecular Function

Links

hUbiquitome
CKSAAP_UbSite
UbiProt database V 1.0
More...

Figure 11 Homepage of mUbiSiDa: Options are selected accordingly.



PLMLA: Prediction of lysine methylation and lysine acetylation

Input the sequences in Fasta format: (Example):

Please select which kind of post-translational modifications need to be predicted:
● methylated lysine ○ acetyllysine ○ methylated lysine and acetyllysine

Sequence:

Submit

Threshold: 0.5 ▾

Output:

Figure 12 Homepage of PLMLA tool. Sequence is pasted in FASTA format.

sequence is done using *In silico* tool PLMLA (Prediction of potential lysine methylation and lysine acetylation) (Figure 12). Sequence is submitted in FASTA format and appropriate option is selected based on post-translational modification requirement for prediction. Name of protein, site position predicted result is returned.

CyMate: It is a tool to perform *In silico* evaluation of DNA methylation at cytosine site. It is a simple, quick and automated tool. It is a comprehensive tool.

CONCLUSION

The computational or *In silico* approach that has been highlighted in this review for predicting the structure and function of unknown protein apprehends the efficacy of various tools of bioinformatics. These tools are pre-requisite in predicting structural and functional features thereby facilitating experimental analysis of proteins. The study of proteins made here can be explored and utilized further so that it can be beneficial for therapeutic purposes.

REFERENCES

1. Pruess M, Apweiler R. Bioinformatics Resources for *In Silico* Proteome Analysis. J Biomed Biotechnol. 2003; 2003: 231-236.
2. Ginnis Scott Mc, Madden Thomas L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004; 32: 20-25.
3. Pradeep NV, Anupama A, Vidyashree KG, Lakshmi P. *In silico* Characterization of Industrial Important Cellulases using Computational Tools. Advances in Life Science and Technology. 2012; 4.
4. Anshul T, Monika S, Sandeep S, Pant AB, Prachi S. *In silico* Characterization of Retinal S-antigen and Retinol Binding Protein-3: Target against Eales' Disease. Int J. Bioautomation. 2014; 18: 287-296.
5. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010; 5: 725-738.
6. Geetika J, Mishra A K, Pandey P S, Chandrasekharan H. Structure and function prediction of unknown wheat protein using LOMETS and I-TASSER. Indian Journal of Agricultural Sciences. 2012; 82: 867-874.
7. Priyadarshini P, Kumar NP, Dipankar S, Kumar SS, Chanderdeep T. Mode of interaction of calcium oxalate crystal with human phosphate cytidylyl transferase 1: a novel inhibitor purified from human renal stone matrix. J. Biomedical Science and Engineering, 2011; 4: 591-598.
8. Laskowsk RA, Macarthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemicai quality of protein structures. J. Appl. Cryst. 1993; 26: 283-291.
9. Ertugrul F, Ibrahim K. In silico sequence analysis and homology modeling of predicted beta-amylase 7-like protein in Brachypodiumdistachyon L. J BioSci Biotech. 2014; 3: 61-67.
10. Notredame C, Higgins DG, Heringa J. T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. JMB. 2000; 302: 205-217.
11. Sabitha K, Rajkumar T. Identification of small molecule inhibitors against UBE2C by using docking studies. Bioinformation. 2012; 8: 1047-1058.
12. Raj U, Varadwaj PK. Flavonoids as Multi-target Inhibitors for Proteins Associated with Ebola Virus: in-silico Discovery Using Virtual Screening and Molecular Docking Studies. Interdiscip Sci. 2015; 7: 1-10.
13. Palmeri A, Gherardini PF, Tsigankov P, Ausiello G, Späth GF, Zilberstein D, et al. PhosTryp: a phosphorylation site predictor specific for parasitic protozoa of the family trypanosomatidae. BMC Genomics.2011; 12: 614.

Cite this article

Singh N, Upadhyay S, Jaiswar A, Mishra N (2016) In silico Analysis of Protein. J Bioinform, Genomics, Proteomics 1(2): 1007.