Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 5, Number 2 (2012) pp. 133-140

http://www.ripublication.com/jcib.htm

SLocP Tool Box-An Integrated Platform for Predicting Sub Cellular Localization of Proteins

Saravanan Vijayakumar and P.T.V. Lakshmi*

Centre for Bioinformatics,
School of Life Sciences,
Pondicherry University, Pondicherry-605014, India
E-mail: brsaran@gmail.com, lakanna@bicpu.edu.in

Abstract

Prediction of subcellular location of a protein remains a challenging task in the field of computational biology. Several approaches, employing amino acid composition, Gene Ontology based, Evolutionary profile based, statistical techniques, and machine learning techniques, are adopted for predicting the exact location of the protein. Although, several approaches exist, there remains the problem in accurate prediction of the sub-localized protein. Moreover, to achieve this task, it has become mandatory to compare the results obtained by different approaches and analyze them manually, which has become a tedious job to perform. Hence, to overcome these difficulties, it is proposed to attempt in collating the available online subcellular localization prediction tools under a single platform to achieve the results with certainty. Thus, the SLocP tool box was developed, using PERL, that provides the user to submit the sequence either singly or in multiples, enabling for easy and quick comparison of results from various prediction server. Perhaps, this tool can be executable under both Linux and Windows platform.

Keywords: SLocP, Protein subcellular prediction, Integrated tool, PERL, comparison.

Introduction

Knowledge about the subcellular location of a protein aids in deducing its function

and its role in interaction with other biomolecules. Traditional way of determining the subcellular location of a protein is expensive time and consuming, in particular the new entries in the Swiss Prot database is increasing rapidly day-by-day [1]. To address this problem various tools with varied approaches are proposed in recent years of which, *PlantmPloc* [2], *YLoc*[3], *WoLF PSORT*[4], *EpiLoc*[5], *TargetP*[6], *Predator* (http://urgi.versailles.inra.fr/predotar/predotar.html), *Euk-mPLoc*[7], *Hum-mPLoc*[8], *ProLocGO*[9], *ESLpred*[10], *HSLpred*[11], SubLoc[12], *Mitoprot*[13], *CELLO*[14], *SubNuclear prediction* [15] , *PSORTb* [16], *CellmPLoc* [8], *SecretomeP* [17], *PSLpred* [18] and Adaboost Subcellular prediction [19] are a few to be named at.

Even though all the tool work towards predicting the subcellular location for a given input, the predicted location from single tool alone could not be trusted with high confidence, as most of the tool have shown prediction accuracy level of only about 60–80%; in fact, each tool can perform well than the other in certain input [21]. So, comparing and analyzing the results manually from various tools could improve the confidence level in considering the predicted subcellular localization of the given input protein, provided that the user is familiar with all the prediction tools. Hence, to implement this, a platform called SLocP Tool Box, which integrates all the available working online servers to predict the sub-cellular localization of both Prokaryotic and Eukaryotic proteins have been developed. The output generated enables for easy comparison of results from different prediction server, without any constraint.

Implementation

SLocP Tool Box is implemented in PERL V5.12. It can be executed in both Windows and Linux. Since the master code is converted into executable, there is no pre-requisite of PERL installation to execute the tool. However, for executing under Linux platform, user has to install BioPerl, Tk and Mechanize module, which can be found at CPAN (Comprehensive Perl Achieve Network) library.

SLocP Tool Box Architecture

The overall architecture of SLocP Tool Box as shown in Fig.1 consists of three main components (a) visible interface for the user to submit the input sequence and to select the desired subcellular location prediction tool (b) an invisible interface to parse the results for the submitted sequences and (c) a module to present the parsed result in comparable way.

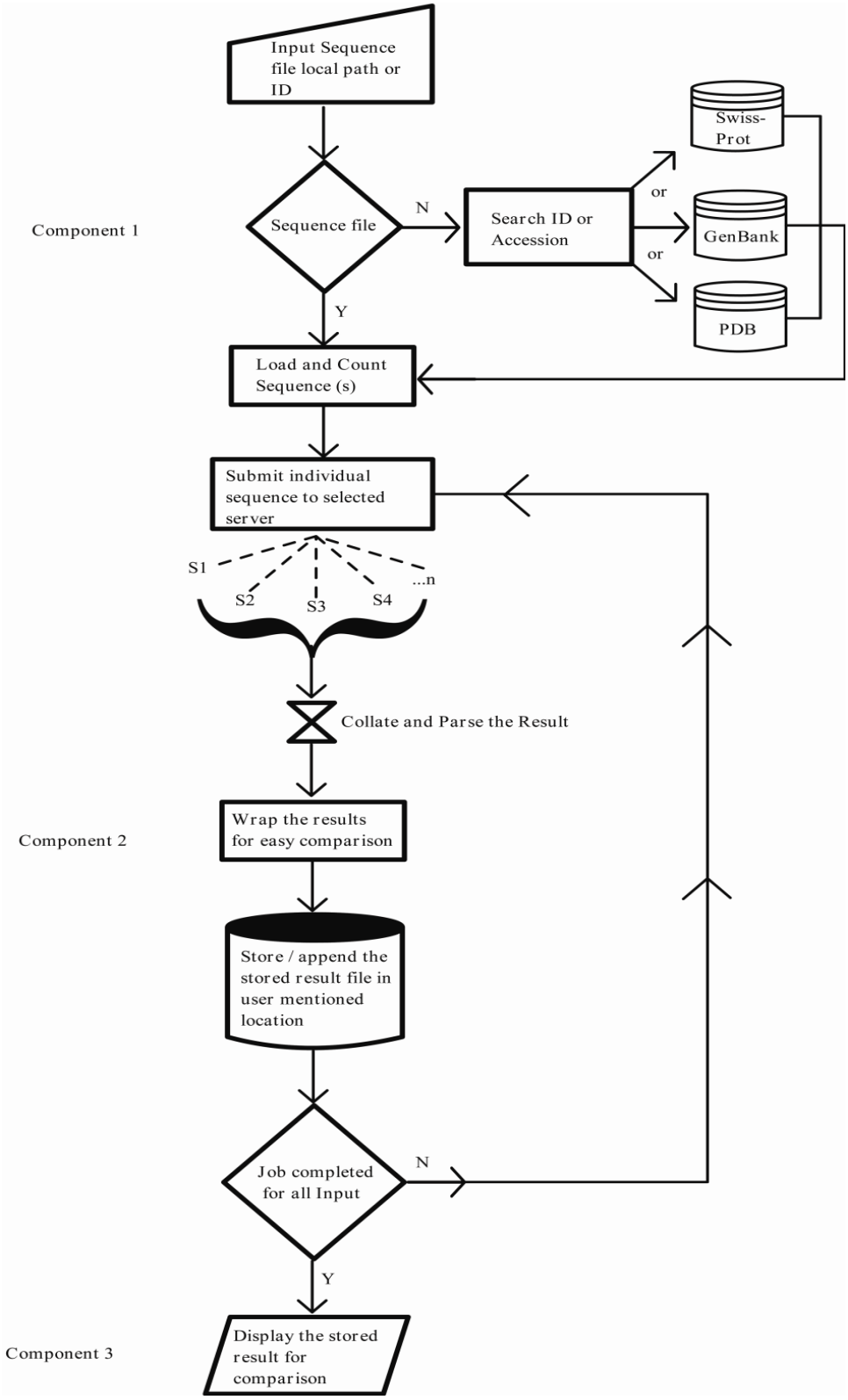


Figure 1: SLocP Tool Box Architecture

Sequence Submission

The input to the system will be protein sequence in FASTA format. User has the option of submitting the input sequence in three different ways; can just paste the sequence in the sequence input box of the tool; can retrieve the sequence(s) from SwissProt, GenBank, or PDB through their corresponding accession number/ID that can be loaded into the retrieve sequence area; and/or by uploading the file containing the sequence through browsing the system. To overcome the restriction of submitting multiple sequences at a time to various prediction servers, the tool is designed in such a way that it could handle a multiple sequences by submitting the single sequence multiple times and hence prevent the server side restrictions. Thus, the designed tool has no restriction for number of sequence input, however the individual's system memory may restrict.

Tool Selection

The available prediction server in SLocP tool box is listed in Table 1. For the convenience of the user, the prediction tools are classified into three categories viz., Plants, Eukaryotes, and Prokaryotes. Even though plants come under eukaryotes, it was classified into separate category, as some of the prediction servers were specially designed to predict subcellular location based on plant data. Once the input is given to the tool, user could select through any number of prediction servers available in the list for prediction. During selection, a brief detail about the selected tool is displayed in the message window. Fig. 2 shows the interface of SLocP Tool Box.

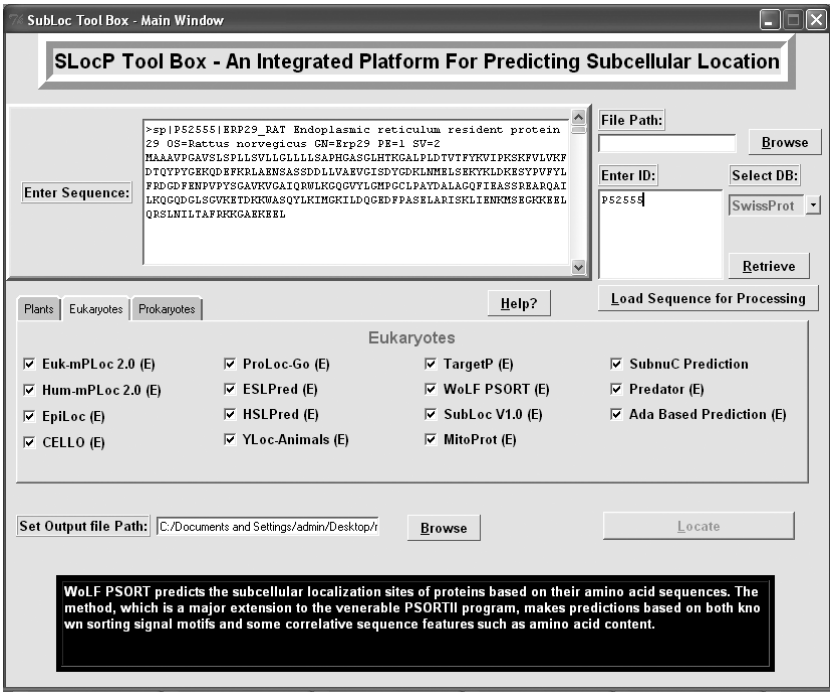


Figure 2: SLocP Tool Box interface

Table 1: List of Prediction Servers Incorporated in SLocP Tool Box

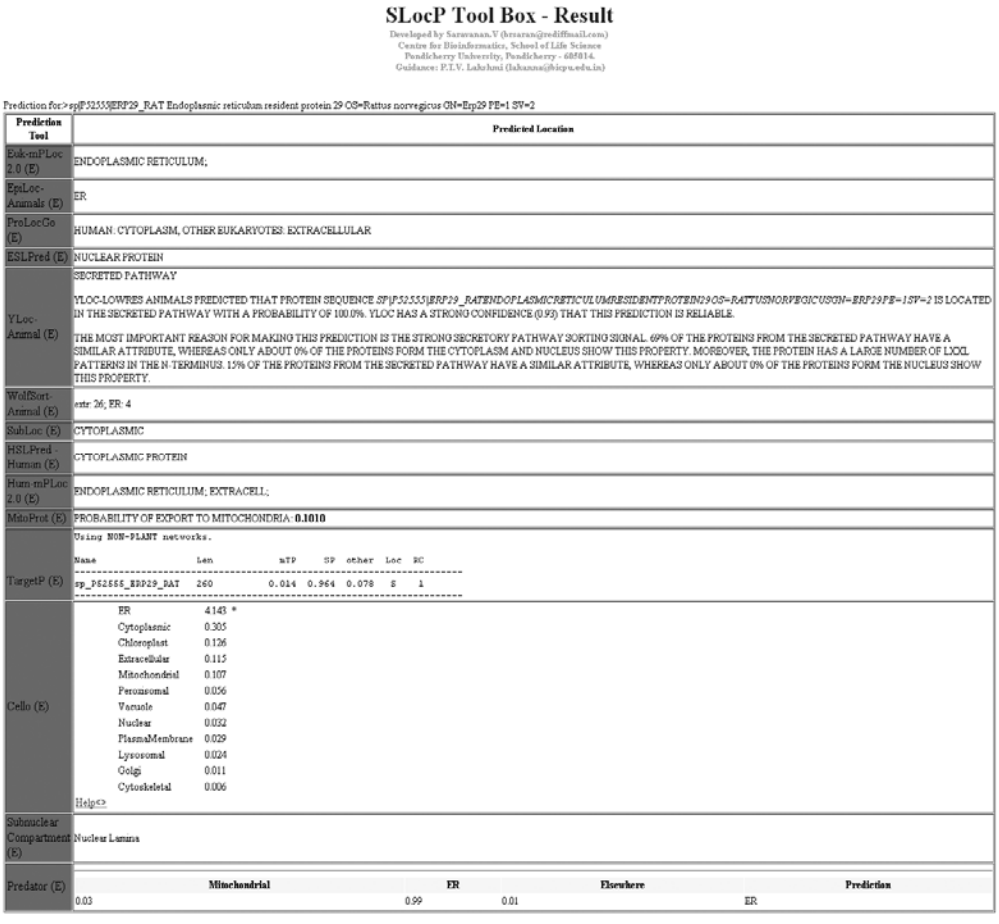
S. No.	Predition Tool [reference]	Computational Method(s)
	PSORTb [16]	Support Vector Machine
	SubLoc V1.0 [12]	Support Vector Machine
	Gneg/pos-mPLoc [8]	*ab-initio*
	SecretomeP [17]	*ab-initio*
	Euk-mPLoc [7]	*ab-initio*
	Hum-mPLoc [8]	*ab-initio*
	EpiLoc [5]	Support Vector Machine
	WoLF PSORT [4]	Weighted K Nearest Neighbors
	ProLoc GO [9]	Genetic algorithm based method combined with SVM
	ESLPred [10]	Support Vector Machine
	HSLPred [11]	Support Vector Machine
	YLoc [3]	Naïve Bayes
	MitoProt [13]	Multivariate Analysis
	TargetP [6]	Neural network and Probability score of different features
	Cello [14]	Support Vector machine
	Plant-mPLoc [2]	*ab-initio*
	PSLpred [18]	Support Vector Machine
	Subnuclear compartment prediction [15]	Support Vector Machine
	Subcellular Prediction [19]	AdaBoost Machine learning
	Predator [20]	Neural network

The results from various servers are parsed and collated in invisible mode, however, the status of the job will be visible to the user. Once the entire job is completed for individual sequence, the result of the performance is saved in user mentioned file path. In the case of multiple sequence submission, once the individual sequence job is completed the result of the completed job will be appended to the user mentioned file. This will prevent the loss of completed sequence result due to network failure during the course of the job. After the entire job is completed, the results for the input would be automatically made available for the user to interpret.

Example

The efficiency of the tool was tested with a number of protein sequences. For an instance, protein sequence with the Uniprot ID P52555 that encoded to localize in endoplasmic reticulum when was given as an input to the SLocP tool box revealed different pattern of localization. Out of the fifteen eukaryotic prediction tools selected from the SLocP tool box, EukmPloc, Epiloc, Cello, Hum-mPloc, and Predator

predicted the location as “Endoplasmic Reticulum” while SubLoc, Adaboost prediction, and HSLPred predicted the protein to be localized in “Cytoplasm”, while still others such as ProLocGo and WolfSORT predicted the location as “Extracellular”. However, TargetP and YLoc strongly suggested it as a “Secretory protein”, while MitoProt probability index recommended with very less probability of being in mitochondria. Moreover, ESLPred and SubNuc prediction revealed to be localized in Nucleus and Nuclear Lamina respectively. This variation in the results may be due to the underlying principle or algorithm behind the individual tool. But still, by comparing the result from the entire listed tools, one could unravel the location with higher confidence. In fact, majority of the tools (five) predicted the location as “Endoplasmic reticulum” and also Wolf PSORT suggested four nearest neighbors as “Endoplasmic reticulum” apart from predicting the location as “Extracellular”. Some of the modules like YLoc and TargetP also suggested the given protein to be a “Secretory protein” adds support to evidence for the input protein to be synthesized in “Endoplasmic reticulum” [22] as the MitoProt probability index strongly suggested its absence in mitochondria. The result generated by SLocP tool box for the test input was shown in the Fig 3.



SLocP Tool Box - Result

Developed by Saravanan.V (hrsaran@rediffmail.com)
Centre for Bioinformatics, School of Life Science
Pondicherry University, Pondicherry - 605014.
Guidance: P.T.V. Lakshmi (lakshmi@bicpu.edu.in)

Prediction for:>sp|P52555|ERP29_RAT Endoplasmic reticulum resident protein 29 OS=Rattus norvegicus GN=Erp29 PE=1 SV=2

Prediction Tool	Predicted Location
Euk-mPLoc 2.0 (E)	ENDOPLASMIC RETICULUM;
EpaLoc-Animals (E)	ER
ProLocGo (E)	HUMAN: CYTOPLASM, OTHER EUKARYOTES: EXTRACELLULAR
ESLPred (E)	NUCLEAR PROTEIN
YLoc-Animal (E)	SECRETED PATHWAY YLOC-LOWRES ANIMALS PREDICTED THAT PROTEIN SEQUENCE *SP|P52555|ERP29_RATENDOPLASMICRETICULUMRESIDENTPROTEIN29OS=RATTUSNORVEGICUSGN=ERP29PE=1SV=2* IS LOCATED IN THE SECRETED PATHWAY WITH A PROBABILITY OF 100.0%. YLOC HAS A STRONG CONFIDENCE (0.93) THAT THIS PREDICTION IS RELIABLE. THE MOST IMPORTANT REASON FOR MAKING THIS PREDICTION IS THE STRONG SECRETORY PATHWAY SORTING SIGNAL. 69% OF THE PROTEINS FROM THE SECRETED PATHWAY HAVE A SIMILAR ATTRIBUTE, WHEREAS ONLY ABOUT 0% OF THE PROTEINS FORM THE CYTOPLASM AND NUCLEUS SHOW THIS PROPERTY. MOREOVER, THE PROTEIN HAS A LARGE NUMBER OF LXXL PATTERNS IN THE N-TERMINUS. 15% OF THE PROTEINS FROM THE SECRETED PATHWAY HAVE A SIMILAR ATTRIBUTE, WHEREAS ONLY ABOUT 0% OF THE PROTEINS FORM THE NUCLEUS SHOW THIS PROPERTY.
WolfSort-Animal (E)	extr: 26; ER: 4
SubLoc (E)	CYTOPLASMIC
HSLPred - Human (E)	CYTOPLASMIC PROTEIN
Hum-mPLoc 2.0 (E)	ENDOPLASMIC RETICULUM; EXTRACELL;
MitoProt (E)	PROBABILITY OF EXPORT TO MITOCHONDRIA: **0.1010**
TargetP (E)	Using NON-PLANT networks. Name Len mTP SP other Loc RC sp_P52555_ERP29_RAT 260 0.014 0.964 0.078 S 1
Cello (E)	ER 4.143 * Cytoplasmic 0.305 Chloroplast 0.126 Extracellular 0.115 Mitochondrial 0.107 Peroxisomal 0.056 Vacuole 0.047 Nuclear 0.032 PlasmaMembrane 0.029 Lysosomal 0.024 Golgi 0.011 Cytoskeletal 0.006 Help<>
Subnuclear Compartment (E)	Nuclear Lamina
Predator (E)	Mitochondrial: 0.03; ER: 0.99; Elsewhere: 0.01; Prediction: ER

Figure 3: Output form SLocP Tool Box for the UniProt entry P52555

Conclusion

SLocP Tool Box was framed in a way to integrate various subcellular localization prediction tools under one roof. The key features of the tool box include no restriction to the number of sequence input, provides flexibility in tool selection modules, and simple output display for easy comparison. SLocP tool box perform simple but repetitive task in an ambiguity free manner, thereby providing scope for analyzing and considering the result from various prediction server at a given time for both single and multiple inputs. The tool is also designed in a way, such that a new prediction server in future can be incorporated into this tool box.

Acknowledgement

Saravanan Vijayakumar is supported by the DBT-BINC junior research fellow and author thank DBT-BINC for the fellowship provide to carry out the research.

References

[1] The UniProt Consortium, (2010). The Universal Protein Resource (UniProt), Nucleic Acids Res. **38**(Database issue), D142–D148.

[2] Kuo-Chen Chou and Hong-Bin Shen.(2010). "Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization", PLoS ONE 5, e11335.

[3] Briesemeister S, Rahnenfuhrer J, Kohlbacher O (2010). YLoc—an interpretable web server for predicting subcellular localization. Nucleic Acids Res 38, W497-W502.

[4] Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. (2007). WoLF PSORT: protein localization predictor. Nucleic Acids Res 35, W585–W587.

[5] Brady S and Shatkay H (2008). EpiLoc: a (working) text-based system for predicting protein subcellular location. Pac Symp Biocomput 604-615.

[6] Olof Emanuelsson, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen (2007). Locating proteins in the cell using TargetP, SignalP, and related tools. Nature Protocols 2, 953-971.

[7] Kuo-Chen Chou and Hong-Bin Shen (2007). Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. Journal of Proteome Research 6(5), 1728-1734.

[8] Kuo-Chen Chou and Hong-Bin Shen (2008). Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3, 153-162.

[9] Huang, W.-L., Tung, C.-W., Ho, S.-W., Hwang, S.-F. and Ho, S.-Y (2008). ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. BMC Bioinformatics 9, 80.

[10] Bhasin,M. and Raghava (2004). G.P.S., ESLpred: SVM Based Method for Subcellular Localization of Eukaryotic Proteins using Dipeptide Composition and PSI-BLAST. Nucleic Acids Reasearch 32, W414-W419.

[11] Aarti Garg, Manoj Bhasin, and Gajendra P. S. Raghava (2005). SVM-based method for subcellular localization of human proteins using amino acid compositions, their order and similarity search. J. Biol. Chem 280, 14427-14432.

[12] Chen, H., Huang, N., and Sun, Z. (2006). SubLoc: A server/client suite for protein subcellular location based on SOAP. Bioinformatics 22, 376–377.

[13] M.G. Claros, and P. Vincens (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur. J. Biochem 241, 779-786.

[14] Yu CS, Chen YC, Lu CH, and Hwang JK (2006). Prediction of protein subcellular localization. Proteins: Structure, Function and Bioinformatics 64, 643-651.

[15] Zhengdeng Lei and Yang Dai (2006). Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. BMC Bioinformatics 7, 491

[16] L. Gardy, C. Spencer, K. Wang, M. Ester, G.E. Tusnady and I. Simon, et al (2003). PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. Nucl. Acids Res, 3613–3617.

[17] J. Dyrløv Bendtsen, L. Juhl Jensen, N. Blom, G. von Heijne and S. Brunak (2004). Feature based prediction of non-classical and leaderless protein secretion. Protein Eng. Des 17(4), 349-356.

[18] Bhasin, M., Garg, A. and Raghava, GPS (2005). PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics 21(10), 2522-2524.

[19] Bing Niu, Yu-Huan Jin, Kai-Yan Feng, Wen-Cong Lu, Yu-Dong Cai and Guo-Zheng Li (2008). Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol. Divers 12, 41-45

[20] Small I, Peeters N, Legeai F, Lurin C (2004). Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4: 1581–1590

[21] Patarroyo M. A (2009). Validating subcellular localization prediction tools with mycobacterial proteins. BMC Bioinformatics 10, 134.

[22] Johnson, A. E. and M. A. van Waes. (1999). The translocon: a dynamic gateway at the ER membrane. Ann. Rev. Cell Devel. Biol. 15, 799–842.