BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 8(20)

SIGLOCPRED: an algorithm to predict bacterial signal peptides and OMPS

Premnath Dhanaraj*, Jannet vennila James, Patrick Gomez Michael & Indiraleka Muthiah

School of Biotechnology and Health Science, Karunya University, INDIA; Premnath Dhanaraj – Email: prems.bioinfo@gmail.com; *Corresponding author

Received August 21, 2012; Accepted September 09, 2012; Published October 13, 2012

Abstract:
There is a growing interest in Biological investigation to determine the location of proteins, to identify new potentially accessible drug targets. Signal peptide directs the transport of the protein to its location. Bacterial OMPs are essential for their survival in the host organism. SIGLOCPRED a signal peptide predictor for the bacterial proteins as well as OMP prediction has been developed. The signal peptide prediction is done based on the influence of the flanking residues on the signal peptide cleavage. A dataset of proteins with confirmed outer membrane location has being created, and the probable OMP polypeptide sequence is predicted. Since the algorithm uses confirmed datasets the prediction is more reliable and efficient. SIGLOCPRED is as efficient as many of the existing signal peptide predictors and can also predict OMPs in addition.

Keywords: OMP, Signal Peptide, Flanking residues

Background:

The biological process by which a cellular entity such as a protein is transported and maintained in a particular location of the cell such as mitochondria, cytoplasm etc is called cellular localization **[1-2].** Many tools are available till date. These mediate the targeting of secretory precursor proteins to the correct sub cellular compartments in prokaryotes and eukaryotes. Identifying these transient peptides is crucial to the medical, food and beverage and biotechnology industries **[3].** Endoplasmic Reticulum most all proteins that are transported to the ER have a sequence consisting of 5-10 hydrophobic amino acids on the N-terminus. The protein is guided to the ER by a SRP, which moves between the ER and the cytoplasm. It binds to the signal peptide. There are two types of signal peptides directing to peroxisome, which are called PTS. One is PTS1, which is made of three amino acids on the C-terminus. The other is PTS2, which is made of a 9-amino-acid sequence often present on the N-terminus of the protein. (i) N-terminus signal peptides often target the mitochondrial matrix, endoplasmic reticulum; (ii) C-terminus signal peptides often target the peroxisome **[4-6]**.

Structure of a Signal Peptide

A signal peptide consists of three regions: A n-region, A h-region, A c-region. **Figure 1** it shows the n-region is the positively charged hydrophilic region, where the SRP attaches. The h-region is the hydrophobic central region, this part is essential for the interaction with the signal recognition particle and, subsequently, the translocase complex embedded in the membrane of the endoplasmic reticulum. Upon transit through this membrane, the signal peptide is cleaved from the nascent polypeptide chain by signal peptidases and is then in most cases rapidly degraded by other proteases. The c-region consists of the signal peptidase cleavage site **[7-8].**

Outer membrane proteins

OMPs are an important class of proteins usually found in gram-negative bacteria, mitochondria and chloroplasts. The discrimination of these proteins from the other types of proteins is necessary to accelerate drug discovery and genome annotation. Many OMP identification methods already exist and some web servers have also been freely available to the research community **[9-11].**

Functional Importance

Bacterial OMPs are very essential for their survival. They play a major role in the pathogenesis of the organism. The bacterial adaptation to host niches is possible through the outer membrane proteins. They provide resistance against antimicrobial peptides that are produced during the host innate immune response. Some of them like OMPP1, FadL etc are involved in toluene catabolism and degradation of aromatic hydrocarbons. These proteins are also involved in the translocation of long-chain fatty acids across the outer membrane. OmpW belongs to a family of evolutionary related proteins which may form the receptor for S4 colicins in *Escherichia coli.* Until date many localization predictors are available, however such protein localization programs are currently least accurate when predicting OMPs, and so there is a current need for the development of a better OMP classifier **[12]**. Thus the objective of this project is to develop an algorithm that can: (i) Predict the bacterial signal peptides; (ii) Characterize bacterial OMPs.

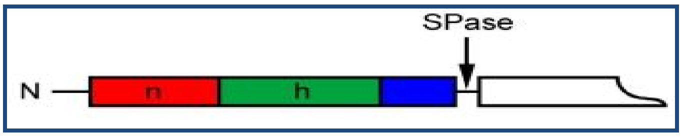


Figure 1: Signal Peptide Structure.

Methodology:

Many databases are available for signal peptide as well as localization. The signal peptide was predicted with the help of datasets from the signal sequence database. The datasets used by SigLocPred for the prediction of the bacterial OMPs was taken from the PSORTdb. PSORTdb is a database for the protein subcellular localizations for both bacteria and archaea. The database is subdivided into two other databases called the cPSORTdb and ePSORTdb. The cPSORTdb contains precomputed genomes or proteomes while the ePSORTdb contains experimentally derived localizations. The dataset contains about more than 1500 sequences whose localization is experimentally determined as outer membrane **[13-15].**

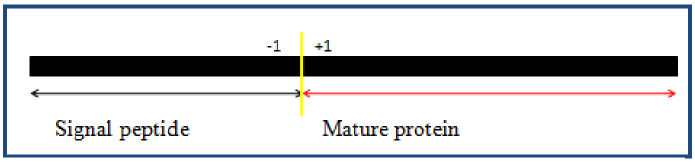


Figure 2: Protein showing The Signal Peptide and Mature Protein.

Frequencies of Amino Acids in Signal Peptide

Proteins have specific amino acids residue positions. The **(Figure 2)** explains the residue position just before the cleavage position is said to be +1 position or P1, similarly the position just after the cleavage site is said to be in -1 position or P1'position. The **(Figure 3)** shows some specific amino acids that tend to occur in these positions at a certain frequency. In a study done with 2352 eukaryotic and bacterial signal peptides it was found that, in bacteria, P1 and P3 favors small, aliphatic residues like Ala and Val. P2 of gram-positive bacteria exhibits Ser (12.5%), Gln (11.9%),Phe (11.9%), Ala (11.3%) and for gram-negative bacteria Leu (17.6%), Gln (14.3%), Phe (11.4%), His (11.4%) are preferred. In the case of signal peptide of gram-positive bacteria P1' is mostly occupied by Ala (36.3%), Asp (11.3%), Ser (10.7%) and Glu (9.5%). While, P2' is populated by Thr (14.3%), Glu\ (13.7%), (Pro) (13.1%), Ser (10.7%) and Asp (10.7%). Similarly in the case of gram-negative bacteria P1' is populated by Ala (41.7%), Gln (12.1%), Asp (7.2%) and Glu (6.2%) whereas P2' is largely distributed between Asp (17.3%), Glu (16.9%), Pro (10.8%) and Thr (10.8%) **[16-18].**

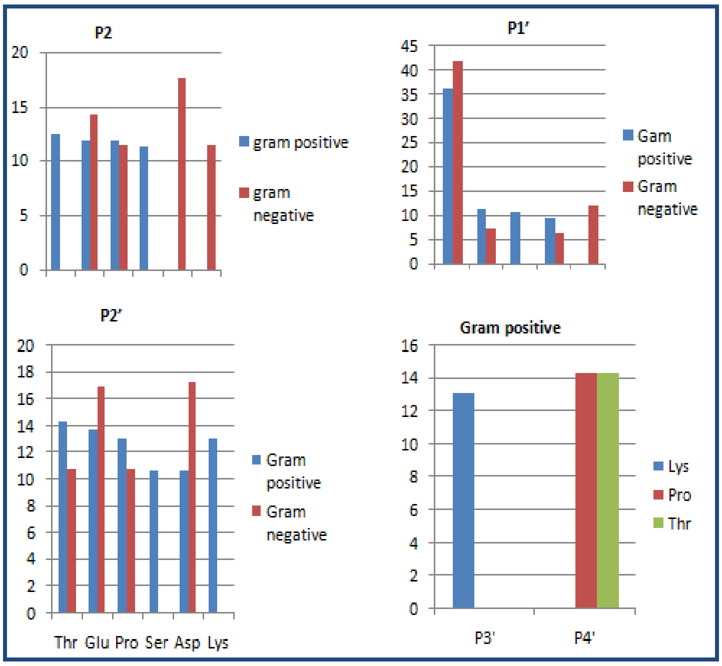


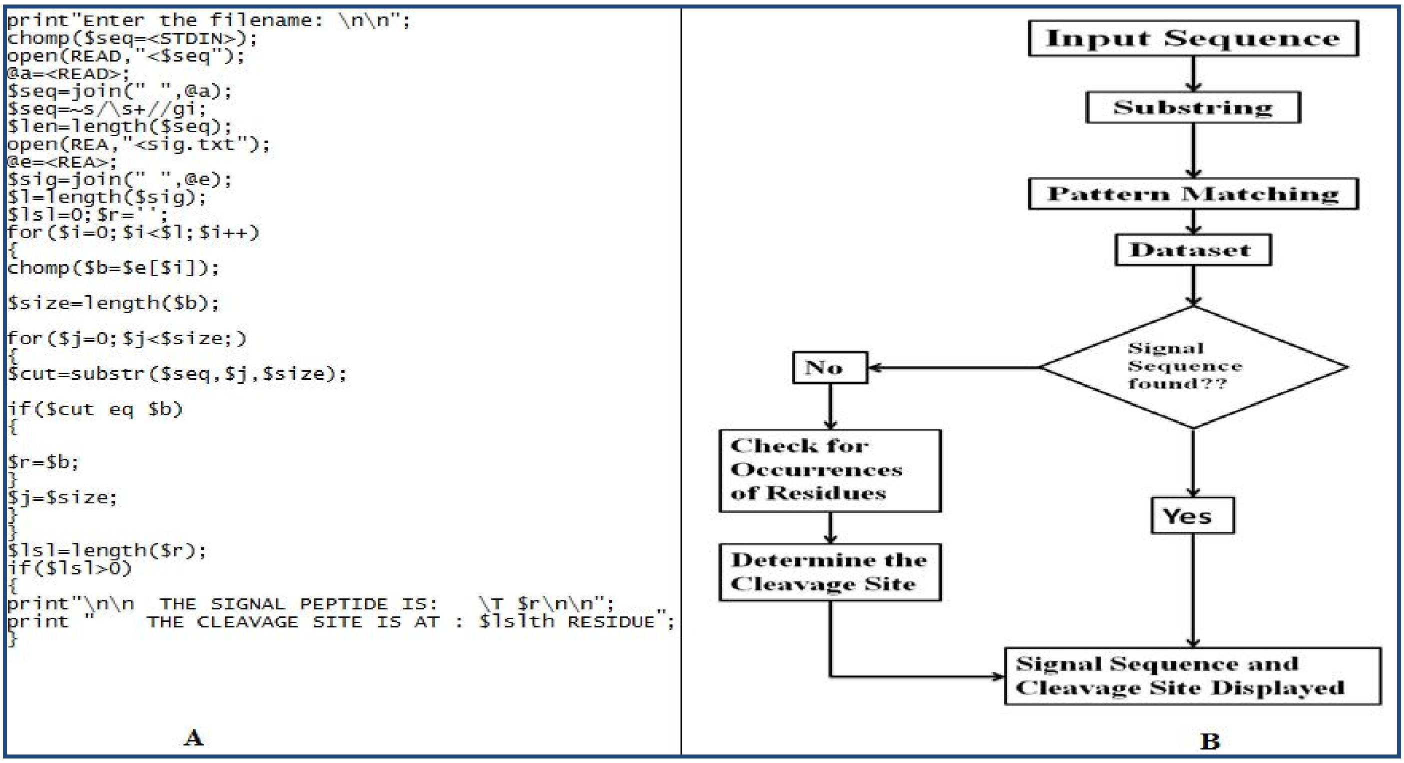
Figure 3: Frequencies of Amino Acids in each Position.

PERL

PERL is a scripting language. **Practical Extraction and Reporting Language** is a high-level, general-purpose, interpreted, dynamic programming language PERL is nicknamed "the Swiss Army chainsaw of programming languages" because of its flexibility and power **[19-20]**.

Prediction of Signal Peptides

Figure 4 shows the algorithm and perl code for The SigLocPred algorithm uses the datasets of confirmed signal sequences in combination with the frequency of occurrence of amino acids flanking the cleavage site. The program written in the PERL language allows the user to input the file name of the input sequence. The filename which contains the bacterial polypeptide sequence whose signal peptide have to be determined. The input sequence is taken as a string and a substring of first 30 amino acids is considered. This is because in bacteria the average length of amino acids is about 25 amino acids. At first, this substring is pattern searched against the dataset for confirmed bacterial signal peptides. By comparing the substring with the dataset position by position the program finds out whether the signal peptide for the given input amino acid sequence is confirmed. If the pattern is found the program displays the signal peptide. If the pattern is not found in the dataset it means that the signal peptide for the user given protein is not yet confirmed and then the position-specific amino acid frequency is used to predict the signal peptide for the input sequence. Here the substring is searched for specific amino acids that tend to appear frequently in certain positions. And thus the signal peptide is predicted. In both the cases the cleavage site of the signal peptidase will be determined.



```
print"Enter the filename: \n\n";
chomp($seq=<STDIN>);
open(READ,"<$seq");
@a=<READ>;
$seq=join(" ",@a);
$seq=~s/\s+//gi;
$len=length($seq);
open(REA,"<sig.txt");
@e=<REA>;
$sig=join(" ",@e);
$l=length($sig);
$lsl=0;$r='';
for($i=0;$i<$l;$i++)
{
chomp($b=$e[$i]);

$size=length($b);

for($j=0;$j<$size;)
{
$cut=substr($seq,$j,$size);

if($cut eq $b)
{

$r=$b;
}
$j=$size;
}
}
$lsl=length($r);
if($lsl>0)
{
print"\n\n  THE SIGNAL PEPTIDE IS:   \T $r\n\n";
print "     THE CLEAVAGE SITE IS AT : $lslth RESIDUE";
}
```

A
Input Sequence
Substring
Pattern Matching
Dataset
Signal Sequence found??
No
Yes
Check for Occurrences of Residues
Determine the Cleavage Site
Signal Sequence and Cleavage Site Displayed
B

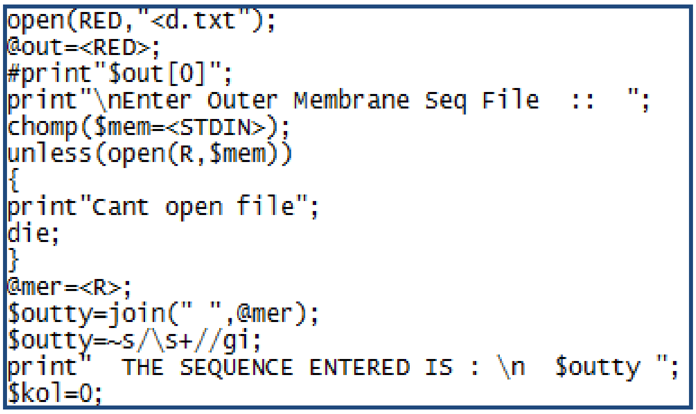
Figure 4: (A) Sample Code for Signal Peptide Prediction; **(B)** Flowchart of Algorithm for Signal Peptide Prediction.

Prediction of Bacterial Omps

The SigLocPred algorithm uses the dataset collected from the PSORTdb. This dataset contains around 1500 sequences. These sequences are those bacterial proteins whose location is confirmed as outer membrane. The user gives the name of the file which consist s of the input sequences. The (**Figure 5)** shows the perl code structure to predict whether the user inputted sequence is an outer membrane protein or not. The sequence that is being entered by the user is treated as an array. Now this array is being pattern matched with the dataset. This is done by using a loop. If the user given sequence is found in the dataset the sequence is an OMP. Otherwise, the input sequence is not an OMP. If the user wants to find a homologous sequence, stand alone blast can be used. Using stand alone blast one can convert the datasets into database against which the query sequence can be blasted. A threshold value can be mentioned by the user according to his interest. And, since the homologous proteins tend to occur in the same location, one can predict whether the query sequence is probably an OMP.

Results and Discussion:

Until date many localization predictors are available, however such protein localization programs are currently least accurate when predicting OMPs, and so there is a current need for the development of a better OMP classifier. There is a current need for the development of a better OMP classifier since the existing methods were least accurate in it. And, SIGLOCPRED is one such algorithm that meets the need of Bioinformatician. SIGLOCPRED predicts the signal peptide for all bacterial proteins and also predict whether the input sequence is an outer membrane protein or not. The user just has to give the input as the filename containing the query sequence and the algorithm predicts the signal peptide and the OMP. The output of the signal peptide prediction includes: (i) The signal peptide of the query sequence; (ii) The cleavage site. Thus in the OMP prediction, the algorithm output is simply the query sequence along with the statements “THE PROTEIN IS LOCATED IN THE OUTER MEMBRANE” and “THE PROTEIN IS NOT LOCATED IN THE OUTER MEMBRANE” for OMPs and other proteins respectively.



```
open(RED,"<d.txt");
@out=<RED>;
#print"$out[0]";
print"\nEnter Outer Membrane Seq File  ::  ";
chomp($mem=<STDIN>);
unless(open(R,$mem))
{
print"Cant open file";
die;
}
@mer=<R>;
$outty=join(" ",@mer);
$outty=~s/\s+//gi;
print"  THE SEQUENCE ENTERED IS : \n  $outty ";
$kol=0;
```

Figure 5: Sample Code for OMP Prediction.

Conclusion:

The prediction of signal peptide was done on the basis of occurrence of certain amino acids in specific positions in the signal peptide. The OMP prediction is done such that it predicts whether the entered query sequence is a probable OMP or not. SigLocPred provides efficient prediction of the signal peptides of all bacterial, both gram-negative and gram-positive proteins. In addition it also specifies between which two amino acids the cleavage site occurs. The current need of the research community is a method that predicts the OMPs, and SigLocPred, a program written in PERL language is such a method. It confirms whether the query sequence is an OMP or not. Moreover, SigLocPred uses confirmed signal peptides and sequences with outer membrane as location as datasets, thus giving way to an efficient prediction.

References:

[1] Rusch SL & Kendall DA, *Mol Membr Biol.* 1995 **12:** 295 [PMID: 8747274]
[2] Rapoport TA, *Science.* 1992 **258**: 931 [PMID: 1332192]
[3] Nielsen H *et al. Protein Eng*. 1997 **10**: 1 [PMID: 9051728]
[4] Park K *et al. Protein Transport into the Endoplasmic Reticulum.* 2009 **13**: 6219
[5] Bannai H *et al. J Mol Biol*. 2000 **300**: 1005
[6] *Emanuelsson O et al. J Mol Biol.* 2000 **300**: 1005 [PMID: 10891285]
[7] Vogt J & Schulz GE, *Structure.* 1999 **7**: 1301 [PMID: 10545325]
[8] Reinhardt A & Hubbard T, *Nucleic Acids Res.* 1998 **26**: 2230 [PMID: 9547285]
[9] Bendtsen JD *et al. J Mol Biol.* 2004 **340**: 783 [PMID: 15223320]
[10] Gardy JL *et al. Bioinformatics*. 2005 **5:** 617 [PMID: 15501914]
[11] Mecsas J *et al. J Bacteriol.* **177**: 799
[12] Gardy LJ *et al. Nucleic Acids Research.* 2003 **31**: 3613 [PMID: 12824378]
[13] Rusch SL & Kendall DA, *Mol Membr Biol.* 1995 **12:** 295 [PMID: 8747274]
[14] Horton P *et al. Nucleic Acids Research.* 2004 **35**: W585 [PMID: 17517783]
[15] Horton P & Nakai K, *Proc Of the Fifth ISMB, AAAI Press.* 1997 **89**: 298
[16] Kall L *et al. J Mol Biol.* 2004 **338**: 1027 [PMID: 15111065]
[17] Cirillo MD *et al. Infect immun.* 1996 **64**: 2019 [PMID: 8675302]
[18] Lu Z *et al. Bioinformatics*. 2004 **20:** 547 [PMID: 14990451]
[19] Choo HK & Ranganathan S, *BMC Bioinformatics.* 2008 **9**: S15 [PMID: 18315846]
[20] Foy BD, *O'Reiley & Associates Inc.* 2004 ISBN: 99-7566

Edited by P Kangueane
Citation: Dhanaraj *et al.* Bioinformation 8(20): 970-973 (2012)