International Journal of Advanced Biotechnology and Research (IJBR)
ISSN 0976-2612, Online ISSN 2278–599X,
Vol 6, Issue4, 2015, pp547-551
http://www.bipublication.com

Research Article

Development of software tool for prediction of signal peptidesand promoter region motifs in defense responsiveSubtilisin-like Protein of *Glycine max* (Soyabean)

Anil Kumar S. Katti* and Rucha M. Wadapurkar

Dept. of PG studies in Bioinformatics,
Walchand Centre for Biotechnology,
Solapur-413006, Maharashtra, India
Corresponding author: Email: anilsk09@gmail.com*
wadapurkarrucha@gmail.com

[Received-19/12/2015, Accepted-24/12/2015, Published-30/12/2015]

ABSTRACT

The defense responses are being activated by small peptides found in crop-derived compounds, whenever there is attack by herbivores and pathogens. However; only some of plant signal peptides have been reported. The amino acid sequence considered in this paper belongs to subtilisin-like protease (subtilase) family. The signal peptide sequence is unique to subtilisin-like protein of Soyabean as it was not found within subtilasesfamily of any other legume crops. Along with the signal peptide, identification of promoter region motifs in genome sequences is one of the major challenges in bioinformatics. The presence of total five promoter region motifs namely, TATA box, CCAAT-box, E-box, GC box and BRE in subtilisin-like protein of Soyabean have been examined. The methodologies used are subsumed of Position Weight Matrix (PWM) and Java Netbeans. The purpose of this analysis is to predict the signal peptide and to check thepresence of promoter region motifs of subtilisin-like proteinin the complete protein source cropSoyabean(Glycine max).The work was carried to develop a Java based software tool for the detection of signal peptide in proteinalong with their cleavage site position and for checking of motifs in nucleotide sequence.

Keywords: Position Weight Matrix (PWM), Java Netbeans,Subtilisin-like protein,*Glycine max*,Signal peptide, Sequence motifs.

[1] INTRODUCTION

The amino acid sequence of a protein contains information about its organelle destination. The information can be considered as zip code that directs the transport of a protein, ensuring its delivery to the correct secretary pathway.

Typically, the information can be found within a short segment of amino acids. These short segments are generally known as sorting-signal sequences, targeting sequences, or signal peptides. After the protein is translocated across the cell membrane, the signal peptide will be *cleaved* off by an extracellular signal peptidase. The location at which thecleave off occurs is called the cleavage site. The mechanism by which a cell transports a protein to its target location within or outside the cell is called the protein sorting process. Defects in the sorting process can causesserious diseases. Therefore, identifying signal peptides and their cleavage sites have both scientific and commercial values.

For instance, to produce recombinant secreted proteins or receptors, it is important to know the exact cleavage sites of signal peptides. The information of signal peptides also allows pharmaceutical companies to manipulate the secretory pathway of a protein by attaching a specially designed tag to it[1]. This ability has opened up opportunity for the design of better drugs.

Signal peptides for the sec pathway generally consist of the following three domains: (i) a positively charged n-region, (ii) a hydrophobic h-region and (iii) an uncharged but polar c-region. The cleavage site for the signal peptidase is located in the c-region. The segment between the h region and the cleavage site is defined as the c region[2]. Based on a statistical analysis, von Heijne (1983) proposed a "(-3,-1) rule": at -1 from the cleavage site (cleavage is defined as occurring between -1 and +1), there is a small, neutral amino acid (Ala, Ser, Gly, Cys, Thr, or Gln); position -3 may not be aromatic, charged, or large polar[3].

The signal peptide which is predicted with the developed tool is of Subtilisin-like protein from soybean contains an embedded, cryptic signal that activates defense related genes[4]. The modification of survival mechanisms during co-evolution of plant hosts with their biotic attackers resulted in the present day complexity of plant–pathogen and plant–insect interactions. Whereas the invading species has developed methods of adhesion, penetration, and feeding, the plant has evolved mechanisms for perception of attack and activation of defense responses, based onsurveillance of its own tissue. Damaged-self recognition occurs when signaling molecules are released from damaged cells andperceived by plant receptors to elicit a defense response.Sequence motifs are short, recurring patterns in DNA that are presumed to have a biological function. Often they indicate sequence-specific binding sites for proteins such as nucleases and transcription factors (TF). Others are involved in important processes at the RNA level, including ribosome binding, mRNA processing (splicing, editing, polyadenylation) and transcription termination.

Promoters have guided evolution for millions of years. It appears that they were the main engine responsible forthe integration of different mutations favorable for the environmental conditions. Promoters are critical regionsfor gene regulation in complex genomes and are located upstream of TSS (Transcription Start Site). A typical promoter region is composed of a core promoter and regulatory domains.

The structure of a promoter isrecognized by the presence of known promoter elements, such as TATA box, GC-box, CCAAT-box, BRE and INRbox. Therefore, accurate recognition of a promoter structure relies on a comprehensive list of promoter elements. Nevertheless, using these promoter elements for classification has proven to be difficult and perhaps evendisadvantageous for different functional correlations between promoter sequences. From an evolutionary standpoint,within non-coding regulatory regions, nucleotides can change their order more frequently and these bindingsites often become very small and instable[5]. Following five promoter region motifs were checked with SignalMotifP tool:

1.1 TATA box

TATA box region motifs were identified using Position Weight Matrix algorithm[6]. The **TATA box** (also called **Goldberg-Hogness box**) is a DNA sequence (cis-regulatory element) found in the promoter region of genes in archaea and eukaryotes; approximately 24% of human genes contain a TATA box within the core promoter.Considered to be the core promoter sequence, it is the binding site of either general transcription factors or histones (the binding of a transcription factor blocks the binding of a histone and vice versa) and is involved in the process of transcription by RNA polymerase.

The TATA box has the core DNA sequence 5'-TATAAA-3' or a variant, which is usually followed by three or more adenine bases. It is usually located 25 base pairs upstream of the transcription start site. The sequence is believed to have remained consistent throughout much of the evolutionary process, possibly originating in an ancient eukaryotic organism[5].

1.2 CCAAT-box

In molecular biology, a **CCAAT box** (also sometimes abbreviated a **CAAT box** or **CAT box**) is a distinct pattern of nucleotides with GGCCAATCT consensus sequence that occur upstream by 60-100 bases to the initial transcription site. The CAAT box signals the binding site for the RNA transcription factor, and is typically accompanied by a conserved consensus sequence. It is an invariant DNA sequence at about minus 70 base pairs from the origin of transcription in many eukaryotic promoters.

Genes that have this element seem to require it for the gene to be transcribed in sufficient quantities. It is frequently absent from genes that encode proteins used in virtually all cells. This box along with the GC box is known for binding general transcription factors. CAAT and GC are primarily located in the region from 100-150bp upstream from the TATA box. Both of these consensus sequences belong to the regulatory promoter. Full gene expression occurs when transcription activator proteins bind to each module within the regulatory promoter. Protein specific binding is required for the CCAAT box activation. These proteins are known as CCAAT box binding proteins/CCAAT box binding factors. A CCAAT box is a feature frequently found before eukaryote coding regions, but is not found in prokaryotes[5].

1.3 E-box

An **E-box** (Enhancer Box) is a DNA sequence found in some promoter regions in eukaryotesthat acts as a protein-binding site, and has been found to regulate gene expression in neurons, muscles, and other tissues. Its specific DNA sequence, CANNTG (where N can be any nucleotide), with a palindromic canonical sequence of CACGTG is recognized and bound by transcription factors to initiate gene transcription. Once the transcription factors bind to the promoters through the E-box, other enzymes can bind to the promoter and facilitate transcription from DNA to mRNA[5].

1.4 GC box

In molecular biology, a **GC box** is a distinct pattern of nucleotides found in the promoter region of some eukaryotic genes upstream of the TATA box and approximately 110 bases upstream from the transcription initiation site. It has a consensus sequence GGGCGG which is position dependent and orientation independent. The GC elements are bound by transcription factors andhave similar functions to enhancers[5].

1.5 BRE

The **B recognition element** (BRE) is a DNA sequence found in the promoter region of most genes in eukaryotes and Archaea.The BRE is a cis-regulatory element that is found immediately upstream of the TATA box, and consists of 7 nucleotides.The BRE was discovered in 1998 by Richard Ebright and co-workers. The first two nucleotides of the BRE sequence can be either guanine or cytosine. The third nucleotide is either guanine or adenine. The next four nucleotides are always the same: cytosine, guanine, cytosine, cytosineG/C G/C G/A C G C C.The Transcription Factor IIB (TFIIB) recognizes this sequence in the DNA, and binds to it. The fourth and fifth alpha helices of TFIIB intercalate with the major groove of the DNA at the BRE. TFIIB is one part of the preinitiation complex that helps RNA Polymerase II bind to the DNA[5].

[II] MATERIALS AND METHODS

2.1 Software Methodologies

The software tool for predictingpeptide and motifs is developed using Java Netbeans software. Netbeans is an integrated development environment (IDE) for developing primarily with Java. It is also an application platform framework for Java desktop applications and others. The Netbeans IDE was written in Java and was run on Windows, it can also be run on OS X, Linux, Solaris and other platforms supporting a compatible JVM. The Netbeans Platform allows applications to be developed from a set of modular software components called modules.

2.2 Algorithm

2.2.1 Weight Matrices

A position weight matrix (PWM), also called position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM)was used for predicting signal peptides in five Glycine max Subtilisin-like proteins. Five sequences were retrieved and stored in fasta format.After that, three different matrices namely Position Frequency Matrix, Position Probability Matrix and Position Weight Matrix were calculated. Finally, threshold values were calculated from Position Weight Matrix.

2.2.2Position Weight Matrix Algorithm

A PWM is a matrix of score values that gives a weighted match to any given substring of fixed length. It has one row for each symbol of the alphabet, and one column for each position in the pattern. PWM score is defined as $\sum_{j=1}^{N} m_{i(j),j}$, where *j* represents position in the substring, *i*(*j*) is the symbol at position *j* in the substring, and $m_{i,j}$ is the score in row *i*, column *j* of the matrix. In other words, a PWM score is the sum of position-specific scores for each symbol in the substring. The score was calculated according to equation below.The elements in PWMs were calculated as log likelihoods. That is, the elements of the PWM were transformed using a background model b so that: $M_{k,j} = \log\left(M_{k,j}/b_k\right).$

The simplest background model shows that each letter appears equally frequently in the dataset. That is, the value of $b_k = 1/|k|$ for all symbols in the alphabet (0.25 for nucleotides and 0.05 for amino acids). A cutoff value wasspecifiedfor sequence to match the motif and signal peptide[6].

2. 3 Identification of signal peptide

Position Weight Matrix algorithm and java software was used to predict signal peptide of five Subtilisin-like proteins[6].In this, signal peptideswere predicted by considering threshold score greater than 0.5.

2. 4 Identification of different promoter region motifs

Position Weight Matrix algorithm and java software was used to predict **TATA box, CCAAT-box, E-box, GC-Box and BRE** promoter region motifsand checking motifs in subtilisin-like proteinaccording to the method developed by Gary Stormo[6]. For prediction of motifs, minimawas calculated from threshold values. Minima values for **TATA box,CCAAT-box, E-box, GC-Box**

and BREwerecalculated -5.7203,0.453,-1.0386,-0.5278 and -0.7101 respectively for five Subtilisin-like proteins.

[III] RESULTS

Attack of herbivores and pathogens to plants initiates defense responsethrough signal peptides.Glycine max Subtilase Peptide (GmSubPep) is a uniqueplant defense peptide signal, cryptically embedded within a plantprotein with an independent metabolic role, providing insights intoplant defense mechanisms.SignalMotifP software was developed to predict signal peptides and motifs of five Glycine max Subtilisin like proteins.

3.1 PWM for prediction of signal peptides

Position Weight Matrix was used to predict signal peptides in five Glycine max Subtilisin like proteins.The score was calculated for 20 standard amino acids for predicting signal peptide sequence and its cleavage position as shown in Table-1.The amino acids having calculated score greater than threshold value were considered as signal peptide sequence.SignalMotifP tool predicts signal peptide sequence from Subtilisin-like proteins. FASTA sequences of five Subtilisin-like proteinswere retrieved from SignalMotifP by clicking on respective Subtilisin-like proteinname buttons. Signal peptide sequences were predicted along with their cleavage position, with this tool, cleavage positions for Subtilisin-like protein were found in the range of 20-30as shown in Fig.1.Similar results were obtained by usingPrediSi tool for predicting signal peptides[2].

3.2 PWM for prediction of motifs:

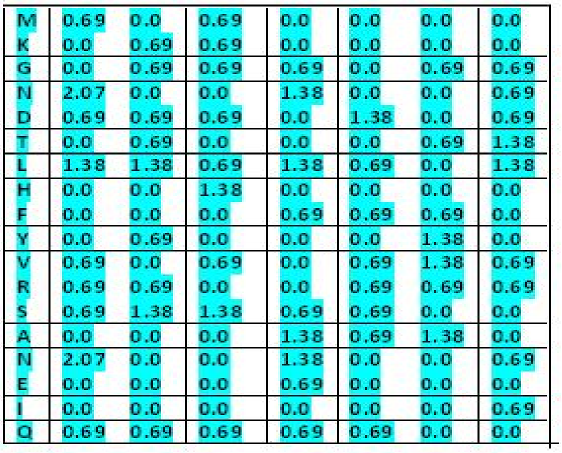
SignalMotifP predicts five different promoter region motifs namely TATA- box, CCAAT-box, E-box, GC-box and BRE –boxby clicking on motifs buttonas shown in Fig.2.The score was calculated for five motif nucleotidessequences as shown in Table-2.The presence and absence of motif regions were determined by calculating the minima from threshold values.If calculated threshold value for each motif found greater than minima then, the particular motif waspresent in the given sequence.

[IV] CONCLUSION

The SignalMotifP tool is offline tool, useful in various bioinformatics applications to detect signal peptides in five Subtilisin-like proteins which are involved in defense response mechanism. It also identifies presence and absence of promoter region motifs in sequence. This tool was developed using Java platform which is user-friendly, easy to build & deploy the software.

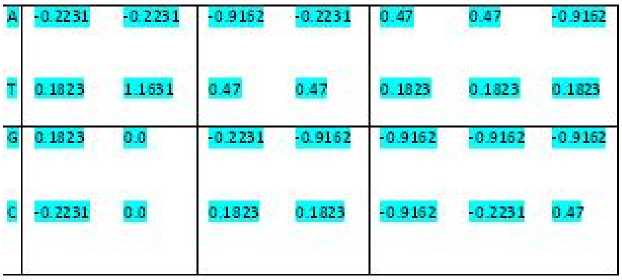
REFERENCES:

1. Man-Wai Mak and Sun-Yuan Kung,(2009), Conditional random fields for the prediction of signal peptide cleavage sites, *Proc. ICASSP*, 1605–1608
2. Karsten Hiller, Andreas Grote, Maurice Scheer, Richard Munch and Dieter Jahn, (2004), PrediSi: prediction of signal peptides and their cleavage positions, *Nucleic Acids Research*, 32, W375–W379
3. Gunnar VON HEIJNE,(2005), Patterns of Amino Acids near Signal-Sequence Cleavage Sites, *Eur. J. Blochem*, 133, 17-21
4. Gregory Pearce, Yube Yamaguchi, Guido Barona, and Clarence A. Ryan,(2010), A subtilisin-like protein from soybean contains an embedded, cryptic signal that activates defense-related genes, *Proc Natl Acad Sci USA*, 107, 14921–14925
5. Paul Gagniuc and Constantin Ionescu-Tirgoviste,(2012), Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters, *BMC Genomics*, 13, 1-16
6. Xuhua Xia, (2012), Position Weight Matrix, Gibbs Sampler, and the Associated Significance Tests in Motif Characterization and Prediction, *Scientifica*, 2012, 1-15



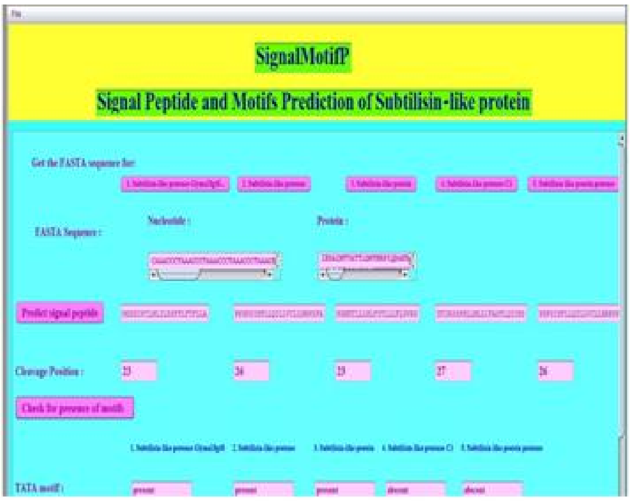
M	0.69	0.0	0.69	0.0	0.0	0.0	0.0
K	0.0	0.69	0.69	0.0	0.0	0.0	0.0
G	0.0	0.69	0.69	0.69	0.0	0.69	0.69
N	2.07	0.0	0.0	1.38	0.0	0.0	0.69
D	0.69	0.69	0.69	0.0	1.38	0.0	0.69
T	0.0	0.69	0.0	0.0	0.0	0.69	1.38
L	1.38	1.38	0.69	1.38	0.69	0.0	1.38
H	0.0	0.0	1.38	0.0	0.0	0.0	0.0
F	0.0	0.0	0.0	0.69	0.69	0.69	0.0
Y	0.0	0.69	0.0	0.0	0.0	1.38	0.0
V	0.69	0.0	0.69	0.0	0.69	1.38	0.69
R	0.69	0.69	0.0	0.0	0.69	0.69	0.69
S	0.69	1.38	1.38	0.69	0.69	0.0	0.0
A	0.0	0.0	0.0	1.38	0.69	1.38	0.0
N	2.07	0.0	0.0	1.38	0.0	0.0	0.69
E	0.0	0.0	0.0	0.69	0.0	0.0	0.0
I	0.0	0.0	0.0	0.0	0.0	0.0	0.69
Q	0.69	0.69	0.69	0.69	0.69	0.0	0.0

Table 1: Position Weight Matrix (PWM) Score for Signal Peptide Prediction



A	-0.2231	-0.2231	-0.9162	-0.2231	0.47	0.47	-0.9162
T	0.1823	1.1631	0.47	0.47	0.1823	0.1823	0.1823
G	0.1823	0.0	-0.2231	-0.9162	-0.9162	-0.9162	-0.9162
C	-0.2231	0.0	0.1823	0.1823	-0.9162	-0.2231	0.47

Table 2: Position Weight Matrix (PWM) Score for Motif Prediction



SignalMotifP

Signal Peptide and Motifs Prediction of Subtilisin-like protein

Get the FASTA sequence for:

1. Subtilisin-like protease Glyma18g8... 2. Subtilisin-like protease 3. Subtilisin-like protein 4. Subtilisin-like protease C1 5. Subtilisin-like protein protease

FASTA Sequence : Nucleotide : [illegible] Protein : [illegible]

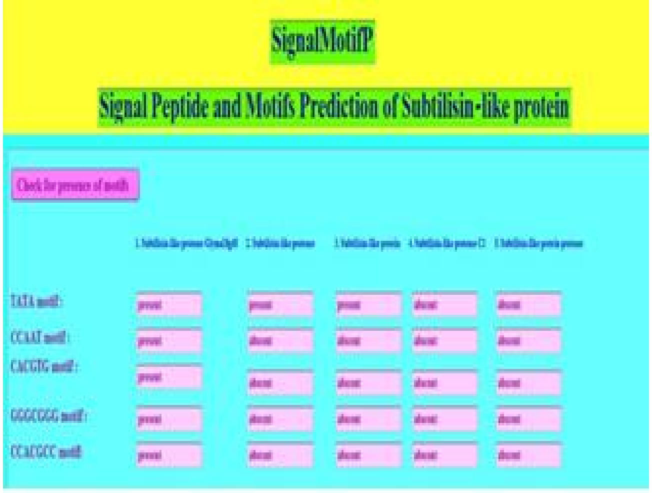
Predict signal peptide [illegible] [illegible] [illegible] [illegible] [illegible]

Cleavage Position : 23 24 23 27 26

Check for presence of motifs

	1. Subtilisin-like protease Glyma18g8	2. Subtilisin-like protease	3. Subtilisin-like protein	4. Subtilisin-like protease C1	5. Subtilisin-like protein protease
TATA motif :	present	present	present	absent	absent

Fig. 1: Screenshot of the Result page of SignalMotifP Software Tool



SignalMotifP

Signal Peptide and Motifs Prediction of Subtilisin-like protein

Check for presence of motifs

	1. Subtilisin-like protease Glyma18g8	2. Subtilisin-like protease	3. Subtilisin-like protein	4. Subtilisin-like protease C1	5. Subtilisin-like protein protease
TATA motif :	present	present	present	absent	absent
CCAAT motif :	present	absent	absent	absent	absent
CACGTG motif :	present	absent	absent	absent	absent
GGGCGGG motif :	present	absent	absent	absent	absent
CCACGCC motif	present	absent	absent	absent	absent

Fig.2: Screenshot of the Result page of SignalMotifP Software Tool