International Journal of Bio-resource and Stress Management 2015, 6(2):208-213 DOI: 10.5958/0976-4038.2015.00036.6

Full Research Article

Searching of Conserved Motifs within a Partial secA Gene Sequence of Phytoplasma Associated with Root (Wilt) Disease of Coconut (*Cocos nucifera*) in India: Using a Frequency Based Approach

Sandip Shil[*1], Kishore K. Das[2] and Anamika Dutta[2]

[1]Central Plantation Crops Research Institute (ICAR), Research Centre, Guwahati, Assam (781017), India.
[2]Dept. of Statistics, Gauhati University, Guwahati, Assam (781014), India

Article History

Manuscript No. AR997
Received in 29th October, 2014
Received in revised form 9th March, 2015
Accepted in final form 28th March, 2015

Correspondence to

[*]*E-mail*: sandip.iasri@gmail.com

Keywords

Coconut disease, conserved region, information content, PWM, motif

Abstract

Coconut palm (*Cocos nucifera L.*) root (wilt) disease (CRWD) became a serious concern for coconut cultivation in coastal region of India. Due to which, India looses a considerable economic loss of about 968 million nuts, annually. The disease is non lethal. It has already been reported that species of the phytoplasmas, belonging to ribosomal group 16SrXI, are primarily associated with CRWD. In the current study, our objective is to identify motif that describes the conserved region within protein sequence of CRWD. Due to variations of amino acid residues at given positions within a positional weight matrix (PWM), the pattern of protein sequence motif can be identified using a multinomial probabilistic model. The basic assumption underlying this model is that the amino acid residues within a sequence are independent and identically distributed. To determine the overall probability of that PWM, the formula for computing information content or relative entropy may also be used. Here, such measure has been proposed to identify motifs within PWM, in this paper. Finally, various conserved motifs within the partial *secA* gene sequence of phytoplasma associated with CRWD, using that measure, have been successfully identified. On the basis of obtained results, we reached to conclusions that conserved regions or motifs (of different length) are expected to be found at (163-171) and (120-138) positions of within that partial *secA* gene sequence.

1. Introduction

Thecoconut root (wilt) disease (CRWD) of coconut palm (*Cocos nucifera* L.) became a serious concern for coconut cultivation in coastal region of India, especially southern state like Kerala. Due to which, India loses a considerable economic loss of about 968 million nuts, annually. The disease is non-lethal. Various molecular studiesrevealed that a specific species of phytoplasma, belonging to ribosomal group 16SrXI, are primarily associated with CRWD. However, other species of different ribosomal groups of phytoplasma, which has been reported to be associated with similar kinds of coconut and other related palm diseases across the globe, may also be associated with CRWD. Some of them are lethal yellowing disease of palms in American countries, Caribbean region, New Guinea and Republic of Cuba (Myrie et al., 2014), Cape St Paul wilt of coconut palm in Ghana (Nipah et al., 2007), coconut yellow decline in Malaysia , Weligama coconut leaf wilt disease (WCLWD) in Sri Lanka (Perera et al., 2012), date palm disease in North Africa and so on. The phytoplasma causing CRWD are usually observed in sieve tubes and often found in parietal position, and more frequently closed to the sieve area or, brain and salivary glands of insects like lice (psyllids).

Although coconut yellowing disease symptoms in different names across the globe indicate that causal phytoplasma species are somehow related, but phylogenetic based studies (Harrison et al., 2008) delineated that different 16Sr group phytoplasma may be associated. In an early, molecular characterization and phylogenetic study of CRWD based on nested PCR amplification of 16S rRNA gene operon (Sharmila et al., 2004) reported that the association of 16SrIV group phytoplasma is mainly associated with CRWD. But, no homologous sequences references analogous to that sequence (Gen Bank: AY158660) was found in Gen Bank. In a subsequent study in Sri Lanka disclosed that the association of phytoplasma, belonging

to the 16SrXI '*Candidatus* Phytoplasma *oryzae'* group, is mainly associated with WCLWD(Perera et al., 2012), and their sequence (Gen Bank: EU635503) is almost identical with recent obtained sequence of CRWD (Gen Bank: GQ850122). Recent molecular characterization and phylogenetic analysis of CRWD sequence (Gen Bank: JX394030) based on less well-conserved *secA* gene, further validated that there is an association of 16SrXI group phytoplasma with CRWD and this phytoplasma has been identified as '*Candidatus* Phytoplasma *oryzae'* closely related strain, belonging to 16SrXI-B group (Manimekalai et al., 2014) .

In this study, our objective is to identify the motifs (or, conserved regions) that describe their conserved region within protein sequence of CRWD. Amotif is a conserved short region within larger sequence, derived from a sequence alignment, and represented by a set of short sequences. Itusually provides an important way to get functional or structural information about anamino acid (or, nucleotide). In case of amino acid sequences, motifs serve as signatures of protein families, and can further be used as tools for prediction of protein structure or function. Before searching a protein motif, an important question arises, what should be appropriate length of the motif? So far, we could not find any specific answer to this. The length usually varies from 3 to 20 amino acid residues, and may also be determined using biochemicalknowledge of the problem. There are short functional motifs that only consist of a very small number of specific residues, and such motifs are found to be associated with important roles in structural context like myristilation sites, glycosylation sites, Src homology [SH]2-binding sites (Bork and Koonin, 1996). It is not necessary to know accuratelength of motifs; but we can try a range of values as recommended by (Schneider et al., 1986).

2. Materials and Methods

2.1. Data preparation

To identify the motifs within protein sequence of CRWD, we chose a partial amino acid based on less well-conserved *secA* genesequence(Gen Bank: AFS50101)of phytoplasma, associated with CRWD,as our query sequence. This partial gene sequence of phytoplasma has beenidentified as '*Candidatus* Phytoplasm*a oryzae'* closely related strain(Manimekalai et al., 2014). Now, our interest is to find out all the highest-scoring alignments analogous to that query sequence, those are mostly likely to represent homologous sequencesof this kind of phytoplasma.Therefore, weexplored a homologous protein sequencesetwith this query sequenceinthe well knownlarge repository, Gen Bank of National Centre for Biotechnology Information (NCBI) (Pruitt et al., 2007), and retrieved only highly homologous sequencesusing the standard nucleotide basic local alignment search tool (BLAST)(Altschul et al., 1990) against Non-Redundant (NR) protein sequences of Gen Bank database, along with '*Candidatus* Phytoplasma' (taxonomy id:33926) organismsearch specification. The 'blastp' algorithm was used in program selection. As per recommendations (Ladunga, 2003), some algorithm parameters were changed according to our need. Further, BLOSUM90 substitution matrix was chosen as protein scoring matrix and segments of that query sequencehavinglow compositional complexity was masked off.

2.2. Multiple sequence alignment (MSA)

Once a highly homologous sequence set is available, our next task is to find out the highest-scoring alignments of those selected sequences. This process of aligning such a sequence set is commonly known as MSA. In short, MSA orders a set of sequences in such a manner that homologous residues between sequences are placed in the same columns of the alignment via introducing indels or gaps. The highest-scoring alignment was generated using a widely-used progressive MSA program, Clustal X 2.1, which is a windows interface of Clustal W (Thompson et al., 1997).

2.3. Identification of motif using information content measure

A common mathematical presentation is matrix, which is also popularly known as position weight matrix (PWM) (Schneider et al., 1986, Hertz et al., 1990). The 20 rows correspond to the 20 amino acid residues {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, and the L columns correspond to the positions of the amino acid residues within the alignment matrix. Elements of the matrix are filled with the frequencies (or probabilities) of observing respective residue at respective position from the aligned sequences.Due to variations of amino acid residues at given positions within a PWM, the pattern of protein sequence motif can be identified using a multinomial probabilistic model. The basic assumption underlying this model is that the amino acid residues within a sequence are independent and identically distributed.

2.3.1. Computation of information content measurewhen an aligned sequence set does not contain any gap

Assume that an aligned set of n homologous amino acid sequences, S_1, S_2,....., S_n, each of L-mers (i.e., sequences of length) were given. For the simplicity of initial problem formulation, we further assume that the aligned sequence set does not contain any gap. This aligned set can be summarized into a PWM, of size ($K \times L$), denoted by $\boldsymbol{N}$. The K corresponds to the size of the alphabet of interest-in this case, 20 amino acid residues $\{A, C, ..., Y\}$: each of the rows corresponds to one of

the letters, representing an amino acid residue, of the alphabet. The L corresponds to the length of the aligned sequences: each of the columns corresponds to one of the positions within the PWM. The elements of the matrix are n_{ij}, the number of occurrence of i^{th} letter at j^{th} position such that $\sum_{i=1}^{K} n_{ij} = n$. Then, for a given j^{th} position (column) within $\boldsymbol{N}$, it can be thought as a sequence of n independent alphabets, in each of which just one of K mutually exclusive letters, $\{A, C, ..., Y\}$, must be observed, and also in which letter i is assumed to be occurred with some a priori (or, background) probability p_i such that $\sum_{i=1}^{K} p_i = 1$. These p_i are usually determined from the observed frequencies of the dataset being analyzed or, some other sources. Now, considering n_{ij} as a random variable in these n alphabets, the joint distribution of n_{1j}, n_{1j}, ..., n_{kj} can be given by

$$Pr\left[\bigcap_{i=1}^{K} n_{ij}\right] = n! \prod_{i=1}^{K} \left(\frac{p_i^{n_{ij}}}{n_{ij}!}\right); \quad n_{ij} \geq 0, \quad \sum n_{ij} = n \tag{1}$$

Equation (1) is considered to follow a multinomial distribution with parameters (n; p_1, p_2, ..., p_K as in (Johnson et al., 1997). Since each position within the aligned sequences is assumed to be independent and identically distributed, the overall probability of the PWM, denoted by Pr [$\boldsymbol{N}$], can be expressed as the product of the multinomial probabilities for each position. Therefore, the probability of $\boldsymbol{N}$ is:

$$Pr\left[\boldsymbol{N}\right] = Pr\left[\bigcap_{i=1}^{K} n_{i1}\right] . Pr\left[\bigcap_{i=1}^{K} n_{i2}\right] .... Pr\left[\bigcap_{i=1}^{K} n_{iL}\right]$$

$$= \prod_{j=1}^{L} \left[n! \prod_{i=1}^{A} \left(\frac{p_i^{n_{ij}}}{n_{ij}!}\right)\right] \quad n_{ij} \geq 0, \quad \sum_{i=1}^{K} n_{ij} = n \tag{2}$$

Another way, to determine the overall probability of that PWM, is use of the formula for computing information content or relative entropy, originally proposed by (Schneider et al., 1986), and is given as:

$$Info\ [\boldsymbol{N}] = \sum_{i=1}^{L} \sum_{i=1}^{K} f_{ij} \log_2 \left(\frac{f_{ij}}{p_i}\right) \quad where\ f_{ij} = n_{ij}/n \tag{3}$$

Here, $Info$ [$\boldsymbol{N}$] refers the information content value of $\boldsymbol{N}$. It has already been shown in (Hertz et al., 1990) that when information content of thePWM increases, a probability that the PWM occurred by chance decreases. One problem arises in equation (3), when some f_{ij} entries in $\boldsymbol{N}$ are zero. This is because of $\log_2 = \infty$. However, such problem can be easily avoided using pseudocounts. With regards to pseudocounts, three approaches were suggested in (Xia, 2011); however, it can be useful to try with second approach. The second approach is to use explicit pseudo counts by defining:

$$\left.\begin{aligned} n_{i.pseudo} &= \alpha\, n_{i.}, \\ &= \sum_{i=1}^{K} n_{i.pseudo} \end{aligned}\right\} \tag{4}$$

where f_i is the overall frequency of amino acid residue, and it is usually advised to keep α small (example, 0.0001). Then, can be re-computed as follows:

$$f_{ij} = \frac{(n_{ij} + n_{i.pseudo})}{(n + n_{i.pseudo})} \tag{5}$$

2.3.2. Computation of information content measurewhen an aligned sequence set contains gaps

Consider the case of the aligned sequence set containing gaps (or, indels). Here again, assume that we are given a set of n homologous amino acid sequences, S_1, S_2,....., S_n, some of having different lengths. However, after alignment, each sequence in aligned set becomes L-mers, but only difference is that some of sequences within this alignment contain gaps over some positions. Therefore, a little complexity arises in our above calculation, which is due to introduction of gaps. Due to incorporation of new gap character (say "-"), this aligned set can be further summarized into a $(K + 1) \times L$ matrix, which is denoted by $\boldsymbol{N'}$. The K and L correspond as usual (defined in case of $\boldsymbol{N}$), but only difference is that the K rows now correspond to all letters, that representing amino acid residues, of the alphabet plus a new gap character. The elements of this matrix are n_{ij}, the number of occurrence of i_{th} letter at j_{th} position such that $n_{-j} + \sum_{i=1}^{K} n_{ij} = n$

One logical problem, in computing information content from this PWM, is that what priori probability we should consider for gap elements. The priori probability for an amino acid residue is typically estimated via counting that respective residue frequency from original sequence set or, aligned set that does not contain any gap. But, the gaps do not appear before the alignment is made, so the priori probability for gap should not be estimated in similar manner. However, to handle such situation, a solution has already been derived by (Hertz and Stormo, 1995), and instead of equation (3), the formula can be written as follows:

$$Info\ [\boldsymbol{N'}] = \sum_{j=1}^{L} \left[f_{-j} \log_2 (f\text{-}j) + \sum_{i=1}^{K} f_{ij} \log_2 \left(\frac{f_{ij}}{p_i}\right)\right] \tag{8}$$

where, $f_{ij} = n_{ij}/n$, $f_{-j} = n_{-j}/n$ and n_{-j} is defined as the frequency of occurrence of gap character "---" at j_{th} position such that $n_{-j} + \sum_{i=1}^{K} n_{ij} = n$. It can easily be observed that if there is no gap (*i.e.*, $f_{-j} \rightarrow 0$), the gap term in (8) does not contribute anything to the sum, and the gap priori probability is also one so that all the priori probabilities sum to 2 rather than one,

which was formally claimed by (Hertz and Stormo, 1995) to define information content measure. But, this problem was also resolved by (Hertz and Stormo, 1995) using a large-deviation rate function, which normalize all priori probabilities with a normalizing factor of 0.5. Therefore, the formula in (8) can be generalized for any priori probability including gap as follows:

$$Info\,[\boldsymbol{N'}] = \sum_{j=1}^{L}\left[\sum_{i=1}^{KU\{-\}} f_{ij}\log_2\left(\frac{f_{ij}}{p_i}\right)\right] \quad (9)$$

$$where \quad \sum_{i=1}^{KU\{-\}} p'_i = \left[p_{\{-\}} + \sum_{i=1}^{K} p_i \log_2\right] = 1$$

Here, $p_{\{-\}}$is the priori probability of gaps. In this study, we considered p=0.5 so that $\sum_{i=1}^{K} p_i = 0.05$.

2.4. Motif-detection program implementation

The above proposed computational measure, to identify sequence motif within a large sequence, was written in R environment (Team, 2014). Finally, a MOTIF-DETECTION programwas implemented using "seqinr" package (Charif and Lobry, 2007). A graphical representation of a motif, popularly known as sequence logo, was generated using "motifStack" package (Ou and Zhu, 2013). A sequence logo typically consists of a stack of letters at each position; the relative sizes of the letters indicates their frequency count in the sequences and the total height of the letters depicts information content of the position, which is measured in bits.

3. Results and Discussion

In this section, an important question usually arises, how to ensure that our MOTIF-DETECTION program correctly identify the protein motifs that are biologically meaningful or validated. Whether our MOTIF-DETECTION program could be applied to all kinds of protein sequences for identification of motifs? Answer is affirmative.

N-myristoylation motif of *Arabidopsis thaliana* proteins usually plays important roles in various cellular activities in eukaryotic organisms; such activities include altering the lipophilicity of a target protein so that the target protein can interact with membranes, interacting with nucleotide-binding proteins, Ca^{2+}-binding proteins, participating in signal transduction pathways, and so on. Further, the biochemical studies (Johnson et al., 1994, Yamauchi et al., 2010) already confirmed that most myristoylated proteins contain a myristoylation motif at the N-terminal end of sequences. Here, we tested our MOTIF-DETECTION program to identify the motif, which is responsible for N-myristoylationin *Arabidopsis thaliana* proteins. We found that our developed program performed the identification job, successfully. The positions of the consensus region, within the PWM of *Arabidopsis thaliana* sequences, was identified on the basis of maximum information content (measured in bits), and also located at the N-terminal end of given protein sequences, as per our expectation (Table 1).

The query with that partial gene sequence of phytoplasma of CRWD (Gen Bank: AFS50101) returned a result, consists of 100 homologous protein sequences from Gen Bank NR database. Of which, we chose only 36 sequences of interests, which mostly include the partial secA gene sequences of phytoplasma related to economically various important crops such as coconut (Gen Bank: ABY48828.1, ACD10534.1, ABY48831.1), napier grass (Gen Bank: ABY48841.1), arecanut (Gen Bank: AFS50100.1), sugarcane (Gen Bank: AFG28541.1), brinjal (Gen Bank: ABY48834.1), *Brassica rapa* (Gen Bank: ADJ67448.1), potato (Gen Bank: ABY48833.1), faba bean (Gen Bank: ABY48816.1), soybean (Gen Bank: ABY48818.1), pepper (Gen Bank: ABY48843.1) etc. All the selected sequences were highly significant, determined based on E-values (-range between 3E-91 and 4E-63). E-value is an important statistical homology measure, which is defined as the number matches with equal or greater scores that are expected by chance. E-value is essentially the same as p-values, the probability of an equal or greater score by chance, if E-value is less than one. The only difference is that an E-value can exceed one, whereas a p-value cannot (Ladunga, 2003). The amino acid residue lengths and percentages of identities with the query sequences also varied from (132 to 161) and (68% to 99%), respectively. Combinations of these important statistical measures essentially help us to identify the statistically and highly significant homologous sequences.

Based on the obtained sequences, an aligned set of sequences

Table 1: information content information table for identification of N-myristoylation motif in *Arabidopsis thaliana* proteins

Serial Number	Using information contents	
	Identified positions in aligned sequences	Information contents (in bits)
1	1-8	20.71300
2	2-9	17.04812
3	3-10	14.57468
4	4-11	13.74983
5	102-109	13.74792
6	5-12	13.41664
7	72-79	13.31144
8	103-110	13.30100
9	106-113	13.27276
10	6-13	13.12654

was generated for further analysis using Clustal X 2.1. Alignment showed that our sequence set containsmany gaps. Therefore, we decide to use the formula of equation (9) to identify motifs. Further, we applied our MOTIF-DETECTION program to identify the motif within CRWD. In this case, there was no prior information about the protein motifs of CRWD that has been biologically meaningful or validated. Therefore, we generated motif sets, which consist of different length size (such as- 3, 5, 7, 9, 11, 13, 15, 17 and 19, respectively), and further computed their information contents using our MOTIF-DETECTION program. Table 2 contains the results of identified conserved regions or motifs (of different length) within the PWM of CRWD protein sequences.

For each motif length, we identified best three motifs within that PWM. The best motif was chosen on the basis of maximum information content, measured in bits and the sequence logo of the best motif was also generated for respective motif length. On the basis of obtained results, we reached to conclusions that conservedregions or motifs (of different length) are expected to be found at (163-171) and (120-138) positions of within that partial *secA* gene sequence.

Table 2: information content information table for identification of conserved regions or motifs (of different length) within the PWM of CRWD

Motif length	Identified positions in aligned sequences	Information contents (in bits)	Sequence logo
3	12-14	16.57830	
	163-165	16.50922	
	164-166	16.02576	
5	163-167	25.92641	
	164-168	25.33625	
	165-169	25.09350	
7	163-169	34.98379	
	9-15	33.89068	
	12-18	33.83341	
9	163-171	42.48157	
	127-135	42.47045	
	9-17	42.04386	
11	125-135	51.84508	
	128-138	51.58375	
	126-136	51.56091	
13	123-135	61.92203	
	126-138	61.52360	
	124-136	61.11364	
15	121-135	72.64754	
	122-136	71.78733	
	124-138	71.06628	
17	122-138	81.73027	
	121-137	80.87405	
	119-135	79.30508	
19	120-138	89.21063	
	121-139	88.91153	
	117-135	87.72159	

4. Conclusion

A program, MOTIF-DETECTION,which thought to be most important development,was successfully implemented. This study provides important formulasfor computing information content and also a generic way to identify motifs within any unaligned or, aligned sequences, applicable to protein sequences related to any organism. As we do not have prior information of motifs about CRWD protein sequences, we could not further able to validate the identified motifs, biologically. But, our program correctly identified the N-myristoylation motif in *Arabidopsis thaliana* proteins.

5. References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. 1990. Basic local alignment search tool. Journal of Molecular Biology 215(3), 403-410.

Bork, P., Koonin, E.V. 1996. Protein sequence motifs. Current Opinion in Structural Biology 6(3), 366–376.

Charif, D., Lobry, J., 2007. Seqin(R) 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla, U., Porto, M., Roman, H., Vendruscolo, M., (Eds.), Structural approaches to sequence evolution: Molecules, networks, populations, Biological and Medical Physics, Biomedical Engineering, Springer Verlag, New York, 207-232.

Harrison, N.A., Helmick, E.E., Elliott, M.L., 2008. Lethal yellowing-type diseases of palms associated with phytoplasmas newly identified in Florida, USA. Annals of Applied Biology 153(1), 85-94.

Hertz, G.Z., Hartzell, G.W., Stormo, G.D., 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Computer applications in the Biosciences: CABIOS 6(2), 81-92.

Hertz, G.Z., Stormo, G.D., 1995. Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps. In: Proceedings of the Third International Conference on Bioinformatics and Genome Research 2, 201-216.

Johnson, D.R., Bhatnagar, R.S., Knoll, L.J., Gordon, J.I.,

1994. Genetic and biochemical studies of protein N-myristoylation. Annual review of biochemistry 63(1), 869-914.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1997. Discrete multivariate distributions, vol. 165. Wiley, New York.

Ladunga, I.S., 2003. Finding homologs in amino acid sequences using network BLAST searches. Current Protocols in Bioinformatics, 3-4.

Manimekalai, R., Nair, S., Thomas, G.V., 2014. Molecular characterization identifies 16srxi-b group phytoplasma ('*Candidatus* Phytoplasma *oryzae*'-related strain) associated with root wilt disease of coconut in india. Scientia Horticulturae 165, 288-294.

Myrie, W., Harrison, N., Douglas, L., Helmick, E., Gore-Francis, J., Oropeza, C., McLaughlin, W., 2014. First report of lethal yellowing disease associated with subgroup 16SrIV-A phytoplasmas in Antigua, West Indies. New Disease Reports 29(1), 12.

Nipah, J.O., Jones, P., Dickinson, M.J., 2007. Detection of lethal yellowing phytoplasma in embryos from coconut palms infected with cape stpaul wilt disease in Ghana. Plant Pathology 56, 777-784.

Ou, J., Zhu, L.J., 2013. motifStack: Plot stacked logos for single or multiple DNA, RNA and amino acid sequence. R package version 1.8.1.

Perera, L., Meegahakumbura, M.K., Wijesekara, H.R.T., Fernando, W.B.S., Dickinson, M.J., 2012. A phytoplasma is associated with the weligama coconut leaf wilt disease in Sri Lanka. Journal of Plant Pathology 94(1), 205-209.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 35(suppl 1), D61-D65.

Schneider, T.D., Stormo, G.D., Gold, L., Ehrenfeucht, A., 1986. Information content of binding sites on nucleotide sequences. Journal of Molecular Biology 188(3) 415-431.

Sharmila, L.B., Bhasker, S., Thelly, M.T., Edwin, B.T., Mohankumar, C., 2004. Cloning and sequencing of phytoplasma ribosomal DNA (rDNA) associated with Kerala wilt disease of coconut palms. Journal of Plant Biochemistry and Biotechnology 13, 1-5.

Team, R.C., 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Available from http://www.R-project.org.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 25(24), 4876-4882.

Xia, X., 2012. Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. Scientifica 2012, 15. http://dx.doi.org/10.6064/2012/917540.

Yamauchi, S., Fusada, N., Hayashi, H., Utsumi, T., Uozumi, N., Endo, Y., Tozawa, Y., 2010. The consensus motif for N-myristoylation of plant proteins in a wheat germ cell-free translation system. FEBS journal 277(17), 3596-3607.