Agric Res
https://doi.org/10.1007/s40003-018-0351-5

FULL-LENGTH RESEARCH ARTICLE

Identification of Protein Motifs in Phytoplasma Associated with Root (Wilt) Disease of Coconut (*Cocos nucifera*) Using an Improved Statistical Measure

Sandip Shil[1] · **Kishore K. Das**[2] · **Vijay Kumar Saxena**[3]

Received: 10 July 2015 / Accepted: 14 August 2018

Abstract Protein motif is a conserved short region within larger sequence, and it provides an important way to get functional or structural information about a noble protein sequence. This study primarily provides a new approach to identify motifs which are considered to play an important role to determine functional or structural class of any protein. Coconut palm (*Cocos nucifera* L.) root (wilt) disease (CRWD) became a serious concern for coconut cultivation in the coastal region of South India. It has already been reported that species of the phytoplasma (-belonging to ribosomal group 16SrXI) is primarily associated with CRWD. In this study, our objective is to modify an existing motif identification algorithm, and also identify motifs that describe their conserved region within a CRWD phytoplasma protein sequence. Accordingly, the existing information content measure formulas for a no-gapped and gapped aligned sequence set were re-formulated. Further, an improvement on both these existing measures has been vied by incorporating prior information of BLOSUM90 substitution matrix. Results showed that proposed modifications could successfully identify N-terminal myristoylation motif within an alignment matrix of *Arabidopsis thaliana* sequences; concurrently our modification has also been validated. Both the approaches, namely frequency- and BLOSUM90-based information content, have been compared using the observed results, in terms of identified motifs. Finally, we could also test this noble BLOSUM90-based approach to correctly predict conserved regions at locations, (163–171) and (117–138), in an aligned *secA* gene sequence set of phytoplasma, an etiological agent of CRWD.

Keywords Alignment matrix · Motif discovery · Position weight matrix · Substitution matrix

Introduction

Typically, a sequence motif is a conserved short region within a larger biological sequence, derived from aligned sequences, and represented by a set of shorter ones. The larger sequence may be composed of nucleotides (-DNA/RNA transcripts) or amino acids (-polypeptide chains). Consequently, sequences related DNA/RNA and polypeptide are used to study DNA/RNA and protein motif, respectively. A sequence motif usually provides an important way to get functional or structural information about such a biological sequence [2]. It may also be thought as an alternative way to classify proteins or their functions that can be achieved firstly, via identification of the sequence motifs within known proteins, and then use those motif information to classify an unknown sequence into the specific protein family or functional class. The present study primarily provides a new approach to identify motifs, which are considered to play an important role to determine functional or structural class of any protein.

✉ Sandip Shil
sandip.iasri@gmail.com

[1] Research Centre, ICAR-Central Plantation Crops Research Institute, Guwahati, Assam 781017, India

[2] Department of Statistics, Gauhati University, Guwahati, Assam 781014, India

[3] Division of Animal Physiology and Biochemistry, ICAR-Central Sheep and Wool Research Institute, Avikanagar, Rajasthan 304501, India

Published online: 30 August 2018

The coconut root (wilt) disease (CRWD) became a serious concern for coconut palm (*Cocos nucifera* L.) cultivation in coastal region of India, especially southern state like Kerala. For which, India looses a considerable economic loss of about 968 million nuts, annually. This disease is non-lethal. Various molecular studies have revealed that a specific species of phytoplasma (-belonging to ribosomal group 16SrXI) is primarily associated with this. However, many species of different ribosomal groups of phytoplasma have been reported to be associated with similar kinds of coconut and other related palm diseases across the globe such as lethal yellowing disease of palms in American countries, Caribbean region, New Guinea and Republic of Cuba [21], Cape St Paul wilt of coconut palm in Ghana [22], coconut yellow decline in Malaysia, Weligama coconut leaf wilt disease (WCLWD) in Sri Lanka [24], date palm disease in North Africa and so on. Although coconut yellowing disease symptoms exist in different names across the globe, the studies indicated that causal species of phytoplasma are somehow related and phylogenetic-based studies [7] also confirmed that different 16Sr group phytoplasma are associated with CRWD. Further, a contemporary study on molecular characterization and phylogenetic analysis of less well-conserved *secA* gene of this sequence (GenBank: JX394030) has further been validated that there is an association of 16SrXI group phytoplasma, and this phytoplasma has been identified as *Candidatus* Phytoplasma oryzae closely related strain, belonging to 16SrXI-B group [20]. In this study, our objective is to modify an existing motif identification algorithm and also identify motifs that describe their conserved region within a CRWD phytoplasma protein sequence.

As we have confined our study to polypeptide chains, here onwards we will only concentrate on protein motif. This kind of motif often serves as a signature for a protein family and may further be used as tools for prediction of protein structures or functions. Many works on motif-detection approaches have already been undertaken [10, 12, 15, 33]. An earlier study, [11] has proposed an approach for identifying consensus patterns in a set of unaligned DNA sequences, which have been known to bind a common protein. In its succeeding study, that approach has further been extended for the situation, where sequence alignment contains one or many gaps [9]. Analogous to this, an alternative approach has also been suggested in [32]. A quite recent study, [33] showed that motif characterization and prediction should be done using position weight matrix methods, and they have also explained this with a simulated example, which have been drawn using a Gibbs sampler. However, before searching a motif, an important question also arises, what should be appropriate length of this motif? So far, we could not find any specific answer to this. But, the length may vary within the range (3–20) amino acid residues and may also be determined using biochemical knowledge of the problem [3]. Enormous studies [3, 17, 19, 28, 29] have alreadydisclosed that some protein motifs, with a very small number of specific amino acid residues, are consequentially associated with important biological mechanisms such as: myristoylation, glycosylation and Src homology [SH]2-binding sites, xylose, mitochondrial translational sites, phosphorylation, post-translational modification sites, bio-synthesis/signal transduction for plant hormones and many more, and may also critically take part in building the structures of various proteins/cellular molecules/organelles. It is not necessary to know the accurate length of motifs; but we may try a range of values as recommended by [26].

Methods and Materials

Identification of Motif Using Information Content Measure

A common mathematical presentation in computational biology is to define a matrix of size $K \times L$, where K rows stands for the total number of possible characters in a biological sequence and L columns correspond to the positions of the characters within an alignment matrix of homologous sequences related to that sequence. This is popularly known as a position-specific scoring matrix (PSSM) or position weight matrix (PWM). PWMs are often derived from a set of aligned sequences that are thought to be functionally related, and is a commonly used representation of motifs (patterns) in biological sequences [5, 26, 33]. The sequence may possess four DNA bases {*A, T, G, C*} or, four RNA bases {*A, U, G, C*} or, twenty amino acid residues {*A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y*}. Elements of the matrix are filled with the frequencies (or probabilities) of observing respective character at a given position of that alignment matrix. Due to variations in amino acid residues at given positions within a PWM, the pattern of protein sequence motif may be identified using a multinomial probabilistic model. The basic assumption underlying this model is that the characters within a sequence are independent and identically distributed.

Computation of Information Content when an Aligned Sequence Set Does Not Contain Any Gap

Assuming that an aligned set of n homologous amino acid sequences, $S_1, S_2, \ldots, S_n$, have been given; each contains L-mers (i.e., sequences of length L). For a simple algorithm formulation, we further assume that the aligned sequence

set does not contain any gap. This set has been summarized into a PWM, denoted by $N_{(K\times L)}$, where K corresponds to the size of alphabets and L is the length of that aligned set. Moreover, each entry of the L columns corresponds to one of the positions within that PWM. As our interested is to study protein motifs, K refers to the twenty amino acid residues $\{A, C, \ldots, Y\}$. Elements of this matrix are n_{ij}, the number of occurrence of ith letter at jth position such that $\sum_{i=1}^{K} n_{ij} = n$. Then, for a given jth position (column) within N, it can be thought as a sequence of n independent alphabets; each of which refers to one of K mutually exclusive letters, $\{A, C, \ldots, Y\}$, that must be observed. Further, the letter i is assumed to occurr 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, let us consider n_{ij} as a random variable in these n alphabets. Then, the joint distribution of $n_{1j}, n_{2j}, \ldots, n_{Kj}$ is given by

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

Equation (1) is considered to follow a multinomial distribution with parameters $(n; p_1, p_2, \ldots, p_K)$ as in [14]. Since each position within the alignment is assumed to be independent and identically distributed, an overall probability of the alignment matrix, $Pr[\mathbf{N}]$, can be expressed as a product of multinomial probabilities of all these positions. Therefore, the probability of $\mathbf{N}$ is:

$$\begin{cases} Pr[\mathbf{N}] = Pr\left[\bigcap_{i=1}^{K} n_{i1}\right].Pr\left[\bigcap_{i=1}^{K} n_{i2}\right]\ldots Pr\left[\bigcap_{i=1}^{K} n_{iL}\right] \\ \quad\quad = \prod_{j=1}^{L}\left[n! \prod_{i=1}^{A}\left(\frac{p_i^{n_{ij}}}{n_{ij}!}\right)\right]; n_{ij} \geq 0, \sum_{i=1}^{K} n_{ij} = n \end{cases} \tag{2}$$

Another way, to determine this overall matrix probability, is use of a formula for computing information contents or, relative entropy, originally proposed by [26] and is given as:

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

Here, Info[$\mathbf{N}$] refers the information content value of N. It has already been shown in [9–11] that the probability that an alignment matrix occurred by chance decreases, when an information content of such a matrix increases. One problem arises in Eq. (3), while some f_{ij} entries in N are zero. This is because of $\log_2 0 = \infty$. However, such problem may be easily avoided using pseudo-counts. In these regards, three approaches are suggested in [33]; however, it would be useful to try with second approach. This approach is to use explicit pseudo-counts by defining:

$$\begin{cases} n_{i.\text{pseudo}} = \alpha n_{i.} \\ n_{\text{pseudo}} = \sum_{i=1}^{K} n_{i.\text{pseudo}} \\ f_{ij} = \dfrac{(n_{ij} + n_{i.\text{pseudo}})}{(n + n_{\text{pseudo}})} \end{cases} \tag{4}$$

where $f_{i.}$ is the total frequencies of amino acid residue i, and usually advised to keep α small (example, 0.0001). Thus, f_{ij} might be re-computed to deal with zero entries.

Computation of Information Content when an Aligned Sequence Set Contains Gaps

Now, we look for the aligned sequence set containing one or more gaps (or indels). In this case, we have also given a set of n homologous amino acid sequences, $S_1, S_2, \ldots, S_n$, which contains unequal length. However, by performing multiple sequence alignment (MSA), we can obtain such an alignment set, in which all sequences have L-mers. Only difference is that some of sequences within this alignment contain gaps over some positions. Therefore, a little complexity arises in our above algorithm due to the introduction of new gap character (say "-"). Let us define PWM of this aligned set as a $(K+1)\times L$ matrix, which is denoted by $\mathbf{N}'$. The K and L correspond as usual notations (defined in case of $\mathbf{N}$). It can be noted that $(K+1)$ rows contain all residue letters, including a new gap character. Elements of this matrix are defined as n_{ij}, the number of occurrence of ith letter at jth position such that $n_{-j} + \sum_{i=1}^{K} n_{ij} = n$.

One logical problem in computing information content of $\mathbf{N}'$ is estimating a priori probability for the gap element. Typically, this probability for an amino acid is estimated from respective residue frequencies in the aligned set in earlier case. Generally, a gap does not appear before performing MSA. So, this priori probability for gap should not be estimated in similar manner. However, to handle such situation, a solution has already been derived in [9, 27], and instead of Eq. (3), new formula can be obtained as follows:

$$\text{Info}[\mathbf{N}'] = \sum_{j=1}^{L}\left[f_{-j}\log_2(f_{-j}) + \sum_{i=1}^{K} f_{ij}\log_2\left(\frac{f_{ij}}{p_i}\right)\right] \tag{5}$$

where, $f_{ij} = \frac{n_{ij}}{n}$, $f_{-j} = \frac{n_{-j}}{n}$ and n_{-j} is defined as the frequency of occurrence of gap character "-" at jth position such that $n_{-j} + \sum_{i=1}^{K} n_{ij} = n$. It can easily be noted that if there is no gap (i.e.,$f_{-j} \to 0$), the gap term in (5) does not contribute anything to the sum. If we consider that the gap probability is 1, then sum of all these terms (-residues and gap probabilities) is 2 rather than 1, which has been formally claimed in [9] to define the formula of information content. Afterward, these can further be generalized into equation (5) using a large-deviation rate function, which normalize

all priori probabilities with a normalizing factor of 0.5, as follows:

$$\text{Info}[\mathbf{N}'] = \sum_{j=1}^{L}\sum_{i=1}^{K\cup\{-\}} f_{ij}\log_2\left(\frac{f_{ij}}{p'_i}\right) \text{where} \sum_{\mathrm{i}=1}^{\mathrm{K}\cup\{-\}} \mathrm{p}'_{\mathrm{i}} = \left[\mathrm{p}_{\{-\}} + \sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{p}_{\mathrm{i}}\right] = 1 \tag{6}$$

where, $p_{\{-\}}$ is priori probability for gap character, and $p_{\{-\}} = 0.5$ has been considered so that $\sum_{i=1}^{K} p_i = 0.5$. These above two formulas 2 and 6 have been named as frequency-based information content approach.

Proposed Modification Over Existing One

Modifying the Measure when an Aligned Sequence Set Does Not Contain Any Gap

One of the major drawbacks using formula (3) is that calculation of information content is solely based on counting respective residue frequencies at specified positions. A conceptual question arises, if we have some prior knowledge about amino acid residues, whether it can be included in Eq. (3). It is quite logical that capability of the motif discovery algorithm might be enhanced by incorporating such kind of prior knowledge, and probability of getting biologically relevant motifs is increased. Therefore, we have modified the original formula (3) by incorporating prior knowledge of blocks substitution matrix (BLOSUM). BLOSUM matrices are most widely used to score alignments between evolutionarily divergent protein sequences during sequence alignment of proteins. These substitution matrices consist of scores for all possible exchanges of one amino acid residue with another and have been computed using the Dayhoff model of evolutionary rates [8]. A high-numbered BLOSUM matrix finds only those amino acid sequences most relevant (or similar) to a queried amino acid sequence, while lower-numbered matrix find relevant, but distantly related ones. Hence, BLOSUM90 substitution matrix has been chosen for these reasons as suggested in [16].

Hereafter, we have mainly concentrated on formulation of Eq. (3) using matrix notation. Firstly, we have referred our alignment matrix or PWM, $\mathbf{N}_{(K\times L)}$, whose elements are $f_{ij} = \frac{n_{ij}}{n}$; f_{ij} represents relative frequency of ith residue at jth position in $\mathbf{N}_{(K\times L)}$ such that $\sum_{i=1}^{K} n_{ij} = n$. It is to be noted that suffix letters refer to the number of rows and columns in a matrix, respectively. Let us also define a probability vector, $\mathbf{p}_{(K\times L)} = [p_1, p_2, \ldots, p_K]$, where p_i represents a priori (or background) probability of amino acid residue i such that $\sum_{i=1}^{K} p_i = 1$ and is assumed to be already determined from the observed frequencies of the dataset being analyzed. Then, an equivalent form of (3) can be re-formulated as follows:

$$\begin{cases} \text{Info}[\mathbf{N}_{(K\times L)}] = 1_{(1\times K)}\mathbf{I}_{(K\times L)}1^t_{(L\times 1)} \\ \mathbf{I}_{(K\times L)} = \mathbf{N}_{(K\times L)} \odot \log_2\{\mathbf{Q}^{-1}_{(K\times K)}\mathbf{N}_{(K\times L)}\} \\ \mathbf{Q}_{(K\times K)} = \text{Diag}(\mathbf{p}_{(1\times K)}) \end{cases} \tag{7}$$

where $\odot$ is a matrix operator, which has been used to perform the Hadamard product of two matrices [18]. Now, consider, $\mathbf{B}_{(K\times K)}$ is a BLOSUM90 substitution matrix. This matrix has been normalized using coefficient of dispersion measure upon range formula, as suggested [6], followed by an addition of the greatest positive value. Therefore, Eq. (7) can be modified by incorporating information of substitution matrix, $\mathbf{B}_{(K\times K)}$ as follows:

$$\begin{cases} \text{Info}[\mathbf{N}_{(K\times L)}] = 1_{(1\times K)}\mathbf{I}'_{(K\times L)}1^t_{(L\times 1)} \\ \mathbf{I}'_{(K\times L)} = \mathbf{B}_{(K\times K)}\mathbf{I}_{(K\times L)} \\ \mathbf{I}_{(K\times L)} = \mathbf{N}_{(K\times L)} \odot \log_2\left\{\mathbf{Q}^{-1}_{(K\times K)}\mathbf{N}_{(K\times L)}\right\} \\ \mathbf{Q}_{(K\times K)} = \text{Diag}(\mathbf{p}_{(1\times K)}) \end{cases} \tag{8}$$

Modifying the Measure when an Aligned Sequence Set Contains Gaps

As indicated earlier, consider the alignment matrix in Eq. (6), $\mathbf{N}'_{((K+1)\times L)}$, whose element, $f_{ij}(= n_{ij}/n)$, represents relative frequency of ith residue at jth position in $\mathbf{N}'_{((K+1)\times L)}$ such that $n_{(-j)} + \sum_{i=1}^{K} n_{ij} = n$. Let us again define a probability vector, $\mathbf{p}'_{(1\times(K+1))} = [p'_1 p'_2 \ldots p'_K p'_-]$, where p'_i represents the priori (or background) probability of amino acid residue i including the gap character, such that $\sum_{i=1}^{K} p'_i = [p_{\{-\}} + \sum_{i=1}^{K} p_i] = 1$. All these probabilities are assumed to be already determined from the observed frequencies of the dataset being analyzed. Therefore, using matrix notations, an equivalent form of Eq. (8) can be obtained as follows:

$$\begin{cases} \text{Info}[\mathbf{N}'_{((K+1)\times L)}] = 1_{(1\times(K+1))}\mathbf{I}'_{((K+1)\times L)}1^t_{(L\times 1)} \\ \mathbf{I}'_{((K+1)\times L)} = \mathbf{N}'_{((K+1)\times L)} \odot \log_2\{\mathbf{Q}'^{-1}_{((K+1)\times(K+1))}\mathbf{N}'_{((K+1)\times L)}\} \\ \mathbf{Q}'_{((K+1)\times(K+1))} = \text{Diag}(\mathbf{p}'_{(1\times(K+1))}) \end{cases} \tag{9}$$

Further, incorporating the BLOSUM90 substitution matrix, denoted by $\mathbf{B}'^{-1}_{((K+1)\times(K+1))}$, a new formula can be derived as follows:

$$\begin{cases} \text{Info}[\mathbf{N}'_{((K+1)\times L)}] = 1_{(1\times(K+1))}\mathbf{I}''_{((K+1)\times L)}1^t_{(L\times 1)} \\ \mathbf{I}''_{((K+1)\times L)} = \mathbf{B}'_{((K+1)\times(K+1))}\mathbf{I}'_{((K+1)\times L)} \\ \mathbf{I}'_{((K+1)\times L)} = \mathbf{N}'_{((K+1)\times L)} \odot \log_2\{\mathbf{Q}'^{-1}_{((K+1)\times(K+1))}\mathbf{N}'_{((K+1)\times L)}\} \\ \mathbf{Q}'_{((K+1)\times(K+1))} = \text{Diag}(\mathbf{p}_{(1\times(K+1))}) \end{cases} \tag{10}$$

These two formulas 8 and 10 have been named as BLOSUM90-based information content approach.

Motif-Detection Program Implementation

Both the approaches, namely frequency- and BLOSUM90-based information content, have been implemented under using the 'seqinr' package [4] in R environment [30]. The program has been named as a Motif-Detection program. Further, a graphical representation of a motif, popularly known as sequence logo, has also been generated using the 'motifStack' package [23]. Moreover, a sequence logo typically consists of a stack of letters at each position, where relative sizes of letters indicate their frequency count within a motif and total height of each one depicts information content of a respective position, measured in bits.

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 same columns of the alignment via introducing gaps [16]. A highest-scoring alignment is usually constructed using a widely used progressive MSA program, namely Clustal X 2.1, which is a windows interface of Clustal W [31].

Preparation of Aligned CRWD Protein Sequences

We have chosen a partial protein sequence of phytoplasma associated with CRWD (GenBank: AFS50101) that is based on less well-conserved *secA* gene, and this is our query sequence. This sequence has been identified as *Ca.* Phytoplasma oryzae closely related strain in [20]. Now, our interest is to find out all the highest-scoring alignments analogous to that query sequence, those are most likely to represent homologous sequences of same kinds of phytoplasma. Therefore, we have explored a homologous protein sequence set with this query in the well-known large repository, namely National Center for Biotechnology Information (NCBI) [25] and retrieved only highly homologous sequences using the standard nucleotide basic local alignment search tool (BLAST) [1] against non-redundant (NR) protein sequences of GenBank database, along with *Candidatus* Phytoplasma (taxonomy id:33926) organism as search specifications. The *BLASTP* algorithm has been selected. As per recommendations [16], some parameters have also been tuned, according to our need. BLOSUM90 substitution matrix has been chosen as protein scoring matrix, and segments of that query sequence having low compositional complexity has also been masked off.

Results and Discussion

Validating Algorithms with an Already Known Motif

In this section, an important question arises, how to ensure the validity of these algorithms? Can these have the ability to identify biologically meaningful protein motifs correctly? Whether could these algorithms be applied to all kinds of biological sequences for identification? Answer is affirmative. It is well known that protein N-myristoylation usually plays important roles in various cellular activities in eukaryotic organisms. These include: altering lipophilicity of a target protein so that it can interact with membranes, formulating protein kinases/phosphatases/guanine nucleotide-binding proteins/Ca^{2+}-binding proteins, participating in signal transduction pathways and also promoting the membrane association that is essential for appropriate protein localization. We have employed both these algorithms to identify this consensus region, which is responsible for myristoylation of plant proteins. However, the biochemical studies [13, 34] have already confirmed that most myristoylated proteins contain a myristoylation motif at the N-terminal end of sequences, and using regular expression notations, this motif is written as: {Met–Gly–*X*–*X*–*X*–Ser/Thr –*X*–*X*}, where *X* refers any amino acid. We have taken a pre-validated set of sequences (Table 1) of *Arabidopsis thaliana* known to contain N-terminal myristoylation motif [34]. Since, these sequences have a validated N-terminal motif, they can be taken as a gold standard to assess the veracity of our modified algorithm. We have decided to search for this consensus region within the alignment matrix of *Arabidopsis thaliana* protein sequences, with the existing as well as proposed algorithms. It has been found that both of them have correctly detected the motif (shown in Table 2). The sequence logo (Fig. 1) for that motif is also matched with the regular expression, which has been explained in [13].

Identifying Motifs Within CRWD Phytoplasma Sequence

The BLAST query against NR database with the CRWD phytoplasma sequence (GenBank: AFS50101) has retrieved a result, comprising of 100 homologous protein sequences. Of which, we have only segregated 36 sequences of our interest, and this list pertains to phytoplasma *secA* gene sequences of economically important

Table 1 This table presents information about the motifs within protein sequences of the eight *Arabidopsis thaliana*, which are known to be responsible for myristoylation of proteins

GenBank accession number	Nucleotides (in base pairs)	Amino acids (number of residues)	N-myristoylation motif	Functions
NM_105159	678	162	MGQVFNKLRG	Ca2+ -binding EF-hand family protein
NM_105319	965	225	MGNSITVKRK	Plastid movement impaired 2 (PMI2)
NM_112728	813	175	MGNTSSMLTQ	Ca2+ -binding EF-hand family protein
NM_115403	1649	389	MGSCLSSRVL	Protein kinase
NM_116252	381	126	MGLSYSGAGV	Zinc-finger (C3HC4-type RING finger) family protein
NM_120398	1281	337	MGNLISLIFC	Zinc-finger (C3HC4-type RING finger) family protein
NM_120468	1392	344	MGCVSSKLGK	Glutaredoxin family protein
NM_125865	1310	384	MGNCAIKPKV	Neurofilament triplet H protein

*GenBank accession number: NCBI GenBank accession number of the complete CDS encoding the protein, Nucleotides (in base pairs): Number of nucleotides present in the CDS, Amino acids (number of residues): Number of amino acid residues present in the protein encoded by the correspondig CDS, N-myristoylation motif: Consensus motif present in the N-terminal of the predicted protein containing the Glycine residue for the covalent attachment of myristoyl group, Functions: Molecular function performed by the protein

Table 2 This table represents obtained 10 best identified locations of the motif within the alignment matrix of *Arabidopsis thaliana* proteins using both the approaches, namely frequency- and BLOSUM90-based information content

Sl. No.	Using existing approach		Using new approach	
	location	IC	location	IC
1	**1–8**	**20.713**	**1–8**	**141.52493**
2	2–9	17.04812	2–9	113.87958
3	3–10	14.57468	3–10	102.69776
4	4–11	13.74983	4–11	96.87761
5	102–109	13.74792	5–12	96.56891
6	5–12	13.41664	6–13	92.69498
7	72–79	13.31144	115–122	87.90456
8	103–110	13.301	14–21	87.63864
9	106–113	13.27276	102–109	87.45031
10	6–13	13.12654	91–98	86.82878

IC information contents measured in bits

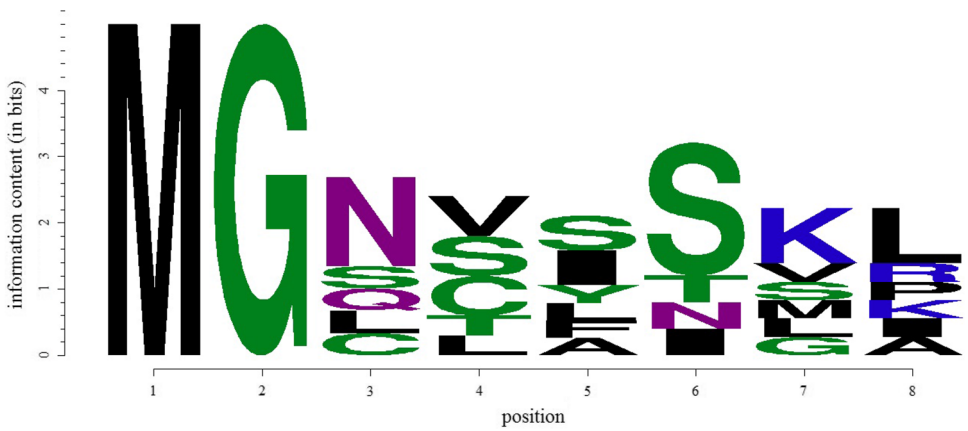

Fig. 1 This figure shows the graphical representation, in the form of sequence logo, of the identified motif, which contains maximum information content (measured in bits) and is located at positions (1–8) within the alignment matrix of *Arabidopsis thaliana* protein sequence

crops, viz. coconut (ABY48828.1, ACD10534.1, ABY48831.1), napier grass (ABY48841.1), arecanut (AFS50100.1), sugarcane (AFG28541.1), brinjal (ABY48834.1), Brassica rapa (ADJ67448.1), faba bean (ABY48816.1), soybean (ABY48818.1), pepper (ABY48843.1). Interestingly, all these sequences are significantly homologous to those that are very well evident from their E–values (–range between $3E-91$ and $4E-63$). E-value is an important statistical homology measure and computes the number matches with equal or greater scores that are expected by chance. This value is essentially the same as p-values that measures the probability of an equal or greater score by chance, provided that $E-\text{value} \leq 1$; only difference is that an E-value can exceed one, whereas a p-value cannot [16]. Chosen sequence residue lengths and percentages of identities with the query have also varied from (132 to 161) and (68 to 99%), respectively. Combinations of these important measures essentially help us to identify the statistically and highly significant homologous amino acid sequences.

Based on these unaligned sequences, MSA has been performed to obtain an aligned sequence set analysis using Clustal X 2.1. But, it has been found that our aligned set contains many gaps, and equation (5) has been chosen to resolve this problem. Applying this, we have generated probable motif sets of different lengths, viz. − 3, 5, 7, 9, 11, 13, 15, 17 and 19, respectively. Also, we have computed their information contents using both those approaches. Tables 3 and 4 contain the location information about three probable motifs within the matrix along with their information content values (in bits) and sequence logo of the best motif for length, viz. − 3, 5, 7, 9, 11, 13, 15, 17 and 19, respectively.

Motif prediction/identification within a primary protein sequence sheds valuable insights into the structural and functional characteristics of a protein. In this study, we have attempted to improve the frequency-based information content approach algorithm. According to this, existing formulas in our modified algorithms have been improved by incorporating substitution matrix information so that all possible amino acid exchange may be taken care during a motif prediction. The proposed algorithm has been first run on *Arabidopsis thaliana* sequences known to contain N-terminal myristoylation motif, and this motif has already been validated in [34]. Interestingly, it has also been observed that the positions of motif within the alignment matrix are located at the N-terminal end of given protein sequences, and their corresponding information content values have also achieved the maximum (see

Table 3 This table represents resultant motifs within the CRWD alignment matrix using frequency-based information content approach

Motif length	location	Information contents (in bits)	Sequence logo
	12-14	**16.5783**	
	163-165	16.50922	
3	164-166	16.02576	
	163-167	**25.92641**	
	164-168	25.33625	
5	165-169	25.0935	
	163-169	**34.98379**	
	8-15	33.89068	
7	12-18	33.83341	
	163-171	**42.48157**	
	127-135	42.47045	
9	8-17	42.04386	
	125-135	**51.84508**	
	128-138	51.58375	
11	126-136	51.56091	
	123-135	**61.92203**	
	126-138	61.5236	
13	124-136	61.11364	
	121-135	**72.64754**	
	122-136	71.78733	
15	124-138	71.06628	
	122-138	**81.73027**	
	121-137	80.87405	
17	119-135	79.30508	
	120-138	**89.21063**	
	121-139	88.91153	
19	117-135	87.72159	

Table 4 This table represents the results of motifs within the CRWD alignment matrix using BLOSUM90-based information content approach

Motif length	location	Information contents (in bits)	Sequence logo
	163-165	**119.5584**	
	164-166	112.8299	
3	165-167	112.8299	
	165-169	**190.2501**	
	163-167	183.8792	
5	165-169	178.9300	
	163-169	**261.2192**	
	165-171	252.3056	
7	129-135	250.6858	
	163-171	**323.2008**	
	162-170	312.4843	
9	128-136	312.0174	
	125-135	**375.0782**	
	161-171	370.1528	
11	147-157	366.9625	
	123-135	**443.4882**	
	124-136	435.3300	
13	125-137	428.6608	
	122-136	**513.6999**	
	121-135	510.7623	
15	123-137	497.0052	
	122-138	**564.5986**	
	121-137	564.2355	
17	119-135	557.4954	
	117-135	**618.7024**	
	151-169	617.2760	
19	148-166	616.2824	

Table 2). Further, comparing the results of both those approach, namely frequency- and BLOSUM90-based information content, we found some feeble differences. We have computed the best 10 identified motifs' information contents for these sequences (Tables 2). It can be noted that the positions of identified motifs are varied between fifth to tenth in size, and the obtained information content values by applying improved approach are quite high. This is due to incorporation of prior knowledge of BLOSUM90 matrix, which consists of positive scores for all possible exchanges of one amino acid residue with another. Moreover, BLOSUM90 usually contains negative scores for all possible exchanges of one amino acid residue with another and positive scores for matched amino acid residues. When we have used raw BLOSUM90, all obtained information contents for motifs are negative values. This is a serious violation rule that overall information content of a matrix is a negative value. So, we have decided to normalize this matrix using coefficient of dispersion measure upon range formula, and subsequently, adding the greatest positive value. For CRWD phytoplasma sequences, we have identified best three motifs within the alignment matrix with respect to defined motif lengths (Table 3 and 4). Unlike earlier case, we noticed that by applying improved approach, obtained information content values are quite high, and probable motif of specified lengths is found to be more consistent than existing algorithm. Although, both these algorithms have predicted probable conserved regions at the same positions within the alignment matrix with the respective length, the results are deviated for motif length of 3, 5, 15 and 19. For motif length of 3, proposed algorithm is predicted more consistent motif at (163–165) positions. For motif length of 19, position for the best motif by proposed algorithm is occupied in the results by frequency-based information content approach as third rank. The reason has already been delineated. Thereafter, on the basis of overall results (Tables 3 and 4), motifs for CRWD phytoplasma sequences are expected to be found at (163–171) and (117–138) positions. The veracity of predicted positions in terms of probable functionalities may be further studied and verified. From this study, we have found that both those approaches provide equally powerful and competent base to develop motif-detection algorithm.

Conclusions

This study mainly provides an improved algorithm for the identification of motifs. We found that existing information content formula can be improved by incorporating

BLOSUM90 substitution matrix as prior knowledge. The study also depicts that both the approaches, namely frequency- and BLOSUM90-based information content, provide equally powerful and competent base to develop a good motif-detection algorithm. The developed methodology may also be used as a generic one in any kind of motif-detection study.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
2. Bork P, Gibson TJ (1996) Applying motif and profile searches. Methods Enzymol 266:162–184
3. Bork P, Koonin EV (1996) Protein sequence motifs. Curr Opin Struct Biol 6(3):366–376
4. Charif D, Lobry J (2007) SeqinR 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, pp 207–232 ISBN: 978-3-540-35305-8
5. Dreos R, Ambrosini G, Périer RC, Bucher P (2015) The eukaryotic promoter database: expansion of EPDnew and new promoter analysis tools. Nucleic Acids Res 43(D1):D92–D96
6. Gupta S, Kapoor V (1982) Fundamentals of mathematical statistics: a modern approach. Sultan Chand and Sons, New Delhi
7. Harrison NA, Helmick EE, Elliott ML (2008) Lethal yellowing-type diseases of palms associated with phytoplasmas newly identified in Florida, USA. Ann Appl Biol 153(1):85–94
8. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89(22):10,915–10,919
9. Hertz GZ, Stormo GD (1995) Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps. Proc Third Int Conf Bioinf Genome Res 2:201–216
10. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15(7):563–577
11. Hertz GZ, Hartzell GW, Stormo GD (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput Appl Biosci: CABIOS 6(2):81–92
12. Huan HX, Tuyet DT, Ha DT, Hung NT (2015) An efficient ant colony algorithm for DNA motif finding. In: Nguyen VH, Le AC, Huynh VN (eds) Knowledge and systems engineering, vol 326. Advances in Intelligent Systems and Computing, Springer, pp 589–601
13. Johnson DR, Bhatnagar RS, Knoll LJ, Gordon JI (1994) Genetic and biochemical studies of protein N-myristoylation. Annu Rev Biochem 63(1):869–914
14. Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions, vol 165. Wiley, New York
15. Kaiser F, Eisold A, Labudde D (2015) A novel algorithm for enhanced structural motif matching in proteins. J Comput Biol 22:698–713
16. Ladunga IS (2003) Finding homologs in amino acid sequences using network BLAST searches. Curr Protoc Bioinf 25:3–4
17. Li F, Liu J, Valls L, Hiser C, Ferguson-Miller S (2015) Identification of a key cholesterol binding enhancement motif in translocator protein 18 kDa. Biochemistry 54(7):1441–1443
18. Liu S, Trenkler G (2008) Hadamard, Khatri-Rao, Kronecker and other matrix products. Int J Inf Syst Sci 4(1):160–177
19. Liu Z, Li F, Zhang B, Li S, Wu J, Shi Y (2015) Structural basis of plant homeodomain finger 6 (PHF6) recognition by retinoblastoma binding protein 4 (RBBP4) component of the nucleosome remodeling and deacetylase (NuRD) complex. J Biol Chem 290:6630–6638
20. Manimekalai R, Soumya V, Nair S, Thomas GV, Baranwal V (2014) Molecular characterization identifies 16srxi-b group phytoplasma (*candidatus phytoplasma oryzae*-related strain) associated with root wilt disease of coconut in india. Sci Hortic 165:288–294
21. Myrie W, Harrison N, Douglas L, Helmick E, Gore-Francis J, Oropeza WC, McLaughlin (2014) First report of lethal yellowing disease associated with subgroup 16SrIV-A phytoplasmas in Antigua, West Indies. New Dis Rep 29(1):12
22. Nipah JO, Jones P, Dickinson MJ (2007) Detection of lethal yellowing phytoplasma in embryos from coconut palms infected with cape St Paul wilt disease in Ghana. Plant Pathol 56:777–784
23. Ou J, Zhu LJ (2013) motifStack: plot stacked logos for single or multiple DNA, RNA and amino acid sequence. R Package Version 1(8):1
24. Perera L, Meegahakumbura MK, Wijesekara HRT, Fernando WBS, Dickinson MJ (2012) A phytoplasma is associated with the Weligama coconut leaf wilt disease in Sri Lanka. J Plant Pathol 94(1):205–209
25. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35(suppl 1):D61–D65
26. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. J Mol Biol 188(3):415–431
27. Shil S, Das KK, Dutta A (2015) 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. Int J Bio-resour Stress Manag 6(2):208–213
28. Sun T, Shi X, Friso G, Van Wijk K, Bentolila S, Hanson MR (2015a) A zinc-finger motif-containing protein is essential for chloroplast rna editing. PLoS Genetics 11(3):e1005,028–e1005,028
29. Sun W, Chen H, Wang J, Sun HW, Yang SK, Sang YL, Lu XB, Xu XH (2015b) Expression analysis of genes encoding mitogen-activated protein kinases in maize provides a key link between abiotic stress signaling and plant reproduction. Funct Integr Genom 15(1):107–120
30. Team RC (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org/
31. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25(24):4876–4882
32. Wu T, Brutlag D (1994) Identification of protein motifs using conserved amino acid properties and partitioning techniques. In: Proceedings. International conference on intelligent systems for molecular biology; ISMB, vol 3, pp 402–410
33. Xia X (2011) Position weight matrix, Gibbs sampler, and the associated significance tests in motif characterization and prediction. Scientifica 2012:917,540–917,540
34. 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 J 277(17):3596–3607