Identification and Analysis of LEA-3 Protein Motifs

Ralph Arvin De Castro

University of Guelph

Commented [RADC1]: Would like to know if structure of paper is fine

Abstract

LEA-3 proteins belong to the late embryogenesis abundant (LEA) protein family. Although it has been strongly associated with desiccation tolerance, its precise role remains unknown. To search for its function, we determined potential sequence motifs and performed an analysis of these conserved segments. Using sequence similarity search tools, a dataset of LEA-3 proteins from Pfam03242 and Phytozome 12 sequences were created. Analysis of these sequences using a motif discovery tool reveals four potential sequence motifs for LEA-3 proteins. One of the motifs, W- motif, is highly conserved among these proteins and is suggested to be the defining segment of the LEA-3 protein family. Assuming conservation indicates function, the W-motif is also suggested to have a biochemical significance in desiccation tolerance.

Commented [RADC2]: I will be proof reading my writing until I submit it on friday

Commented [RADC3]: Would like to know if this suffice

Keywords: LEA-3 proteins, desiccation tolerance, sequence motifs

Identification and Analysis of LEA-3 Protein Motifs

Late Embryogenesis Abundant (LEA) proteins are proteins that accumulate in the late development stage of plant seeds. They are widely assumed to play a crucial role in cellular dehydration tolerance when plants enter a dry state or desiccation stage (Hinchka and Thalhammer, 2012). They were first characterized in cotton seeds but have also been discovered to protect other non-plant species such as cyanobacteria, brine shrimp and nematodes from abiotic stress (Hand et al. ,2012). To prevent desiccation damage, LEA proteins protect other proteins from aggregation, which happens during dehydration (drought, low temperatures or high levels of salinity) (Goyal et al., 2005). Despite many studies associating this correlation, the mechanism behind LEA protein's biochemical function is still unclear.

In terms of structure, LEA proteins are highly hydrophilic and belong to the structural group called intrinsically disordered proteins (IDP) (Hinchka and Thalhammer, 2012). IDPs are proteins that have no stable or fixed tertiary structure. Despite not having this defined structure, IDPs are often functional and tend to be dynamic when bounded to different ligands (Xue and Uversky, 2016). Regions of these proteins can be conserved or not conserved. A highly conserved protein sequence is a great indication of a functional role (Cooper and Brown, 2008). These regions with biological significance can be defined as protein sequence motifs (Ota and Fukuchi, 2017). Protein sequence motifs are amino-acid sequences that occur repeatedly in a protein family and can be used predict the function of the protein and define protein groups (Bork et al., 1996). For example, dehydrins, a subgroup of the LEA protein family, are generally defined by the presence of the K-segment motif (EKKGIMDKIKEKLPG) (Malik et al., 2017).

LEA proteins have been grouped into various subgroups or families based on sequence similarity. In *Arabidopsis thaliana,* Hundertmark and Hincha classified LEA proteins into nine families including dehydrins. Another subgroup, LEA-3 proteins, are more hydrophobic than the average LEA protein and are predicted to be targeted or localized in mitochondria or chloroplast (reference). Salleh et. al hypothesized that LEA-3 proteins interact with other proteins that are involved in mitochondrial ROS (Reactive Oxygen Species) signalling, which induces biotic responses in root development and other pathogenic reactions. However, in terms of abiotic stress, there is still no novel function for this protein and remains under-researched. We therefore set out to perform an analysis to determine the novel motifs for LEA-3 proteins which can be used in determining its function in dehydration tolerance.

Materials and Methods

Selection of LEA-3 Protein Sequences

The Pfam and Phytozome database, shown in Figure 1, served as the main sources of LEA-3 protein sequences. Pfam (version 31.0) provides a large collection of protein families and allows users to download full-length annotated sequences within each family in FASTA format (Finn et al., 2007). In this database, the dataset of LEA_3(PF03242) protein family functioned as the initial source of LEA-3 protein sequences. Phytozome, on the other hand, is a plant genomic resource that delivers a wide array of tools for plant genomic analysis supported by an extensive library of plant DNA and protein sequences (Goodstein et al., 2012). The primary transcripts of each genome in the Phytozome (version 12.1) catalogue were downloaded and compiled into one large file.

The prime objective of this step was to search for potential LEA-3 sequences in the Phytozome database as indicated in Figure 1. To achieve this goal, two sequence similarity search tools were utilized. Firstly, Basic Local Alignment Search Tool (BLAST) is a widely-used bioinformatics search tool that allows researchers to compare and analyze biological sequence information such as DNA, RNA and protein sequences. It utilizes a series of algorithms that locally aligns sets of sequences, finds regions of similarities between these sets and calculates the statistical significance (E-value) for each matching sequence (Altschul et al, 1990). This provides the user the functionality of searching for sequences from any given database, using any query sequence as an input. Its local standalone software (BLAST+) also allows a user to build their own BLAST database using the makeblastdb command. For this study, the Phytozome dataset (version 12.1) was used as the source of our own BLAST database and the Pfam LEA_3 protein dataset was utilized as the query sequence to search for LEA-3 protein sequences. Finally, a BLAST search entry was executed by the blastp command with an E-value of 1e-06.

Meanwhile, Find Individual Motif Occurrences (FIMO) is a motif-based sequence analysis tool from the software, MEME Suite (Version 5.0.2) (Grant et al, 2011). It has a similar search functionality as BLAST but requires a MEME motif instead of a query sequence as its input. The motifs are created from the Multiple Em for Motif Elicitation (MEME) tool, which discovers novel, ungapped motifs from a sequence dataset. Therefore, an intermediate step of generating MEME motifs from Pfam sequences was conducted to fulfill the input requirements. FIMO was then executed for each MEME motif generated against the Phytozome dataset. Subsequently, these FIMO outputs were filtered to find sequences containing all motifs.

The BLAST and FIMO output were then both separately concatenated with Pfam sequences. To check for the accuracy of these datasets, the length and fold index (measure of disorder probability) of the sequences were investigated. A python script, FoldIndex, created by Matthew Stoodley was used to determine the disorder of the proteins and verify the validity of sequences. Datasets that represent a great sample of LEA-3 proteins were then used for subsequent analyses.

Motif Discovery and Analysis

The MEME tool was utilized again for the generation of LEA-3 protein sequence motifs. Running this command generates a MEME output which includes a LOGO representation, position weight matrix and other properties of a motif, which will all be discussed more in the results section. This command was also used in discovering the motifs of Pfam sequences from the last section. Location of these motifs in the sequences were also analyzed using the MEME output.

Results

Selection of Protein Sequences

Primary transcripts of sixty-five annotated genomes from Phytozome 12.1 were downloaded, containing approximately two million protein sequences. The LEA_3 (PF03242) family dataset from Pfam included 490 true LEA-3 protein sequences. Seven duplicates were found and removed, filtering it to 483 sequences.

Running the Pfam sequences as the input of MEME program generated three sequence motifs. Executing FIMO on the first, second and last motif against the Phytozome database

resulted in 67180, 62344 and 76964 sequences respectively. To reduce the number of sequences for analysis, the intersection of these three results were determined, filtering it to 913 sequences. All these sequences are assumed to be LEA-3 proteins since they contain motifs generated by MEME. On the other hand, running the blastp command against the Phytozome database with an E-value of 1e-6 resulted in 889 potential LEA-3 protein sequences, combined with Pfam sequences.

After investigating the length and the disorder probability of these set of sequences, we decided to discard the sequence output from FIMO for the following reasons. Firstly, the lengths of sequences have a wide range of variability and does not follow a normal distribution. Secondly, there were many sequences that are excessively long in length and assumed to only contain all those motifs due to chance. This could lead to false positives and a wrong representation of LEA-3 protein sequences. Thirdly, running the sequences through the FoldIndex script resulted in many sequences that have a positive fold index. A positive fold index is more structured and, a negative fold index means that the sequence is disordered. As mentioned, LEA-3 proteins are disordered and therefore, should have a negative fold index. Considering these factors, we therefore concluded that these sequences do not represent a good sample of LEA-3 proteins. The BLAST output, however, gave us a better representation of LEA-3 proteins. Therefore, we used the 889 sequences results from BLAST as our proposed LEA-3 protein sequences and will be used for motif discovery analysis.

Commented [RADC4]: Not really sure about this.

Analysis of Sequence Motifs

The list of LEA-3 sequences was run through the MEME Suite Software which lead to the discovery of 4 protein sequence motifs. Each motif has its own corresponding LOGO

representation and position weight matrix (reference). A position weight matrix shows the probability of encountering each amino acid in each position. The first motif, shown in Figure 2, can be designated as the W-motif because of its highly conserved tryptophan in its first position. Other highly conserved amino acids are Pro at positions 5 and 12, Thr at position 5 and Gly at position 8. Other positions vary in terms of amino acid type and properties. The second motif, shown in Figure 3, can be labeled as the LL-motif due to the presence of two conserved leucines at its C-terminus. Another conserved Leu is located at position 7. A conserved Asp and Arg can be seen at positions 3 and 8 respectively. Overall, this motif consists of mostly hydrophobic and negative charged amino acids. In Figure 4, the R-A motif, is named after the presence of two arginine at its N-terminus and an alanine tract at the C-terminus. Between those are highly conserved glycine and tyrosine. Lastly, the last motif (Figure 5) is the MARS-motif, which consists of mostly polar and hydrophobic amino acids. Positive amino acids such as arginine and lysine are also conserved at third and ninth position.

The MEME output also provides an E-value, a measure of statistical significance, for each sequence motif generated. Motifs with an E-value higher than 0.05 are not considered statistically significant. In table 1, the E-values for all motifs are very significant with RA- motif, $1.7x10^{-1324}$ having the highest E-value. The motif with highest statistical significance is W-motif with lowest E-value of$1.3x10^{-8373}$. The LL-motif and MARS-motif has an E-value of $7.5x10^{-3738}$ and $1.7x10^{-1324}$ respectively.

The number of sites in which the motifs occur were also stated in the MEME output. Note that the number of sites is the total number of occurrences in the whole set of sequences and does not necessarily equate to one site per sequence. Protein sequences can have duplicates

or more copies of a single motif. However, in our results, there are no cases of multiple copies of a motif in a single sequence. Among the four motifs, The W-motif spans the highest number of sites. This is not surprising as it is the most significant. It occurs 873 times of our sequences. The other motifs, however, are not as conserved as they do not occur in most of the sequences.

MEME also tells us the position of motifs in each sequence, which is shown in Figure 6. Note that this is only a consensus and each motif could be present or absent in the sequence. The MARS-motif is located at the N-terminus followed by the RA-motif. The LL-motif is located very near at the C-terminus and W-motif occurs before it.

Discussion

Our dataset of 889 sequences and our analysis of MEME-defined motifs provide us insights about the general properties of LEA-3 proteins. The most remarkable motif discovered by the MEME program is the W-motif, which contains the highest statistical significance in terms of E-value and occurs in 871 out of 889 sequences. We speculate that the absence of W-motif in other 16 sequences is a product of misannotated Pfam sequences. Incorrectly annotated sequences in PF03242 family could result to the inclusion of non-LEA-3 proteins in our sequence dataset since BLAST searches through the Phytozome database for each query sequence. Without considering the 16 sequences in our analysis, the results would suggest that a LEA-3 protein can be defined or be identified by the presence of W-motif. Assuming that highly conserved amino-acid sequences have functional value, then it can also be suggested that the W-motif has an important role in the LEA-3 protein function.

A comprehensive functional analysis, studying the biochemical properties of these motifs would be a logical direction for this study. However, some initial predictions could be observed

about the function of these motifs based on their position in the sequence. For instance. the MARS-motif located at the N-terminus could be a targeting signal for LEA-3 protein mitochondrial translocation. As mentioned before, LEA-3 proteins are predicted to be targeted to mitochondria (reference). The LL-motif located near at the C-terminus could be a retention signal (reference). Interactions between motifs could also occur since W-motif is very close to the LL-motif and the RA-motif is close to the MARS-motif.

As mentioned above, a functional analysis of these motifs looking at the biochemical properties such as isoelectric point, hydrophobicity, molecular mass and fold index would be a great direction for this research study.

References

Hincha, D., & Thalhammer, A. (2012). LEA proteins: IDPs with versatile functions in cellular dehydration tolerance: Figure 1. *Biochemical Society Transactions*, 40(5), 1000-1003.

Hand S. C., Menze M. A., Toner M., Boswell L., Moore D. (2011). LEA proteins during water stress: not just for plants anymore. *Annu. Rev. Physiol.* 73 115–134.

Goyal, K., Walton, L. J., & Tunnacliffe, A. (2005). LEA proteins prevent protein aggregation due to water stress. *The Biochemical journal*, *388*(Pt 1), 151-7.

Battaglia, M., Olvera-Carrillo, Y., Garciarrubio, A., Campos, F., & Covarrubias, A. A. (2008). The enigmatic LEA proteins and other hydrophilins. *Plant physiology*, *148*(1), 6-24.

Xue, B., & Uversky, V. N. (2016). Unfoldomes and Unfoldomics: Introducing Intrinsically Disordered Proteins. *Molecular Science of Fluctuations Toward Biological Functions*, 125-150.

Cooper, G. M., & Brown, C. D. (2008). Qualifying the relationship between sequence conservation and molecular function. *Genome Research*, *18*(2), 201-205.

Ota, H., & Fukuchi, S. (2017). Sequence conservation of protein binding segments in intrinsically disordered regions. *Biochemical and Biophysical Research Communications*, *494*(3-4), 602-607.

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

Hundertmark, M., & Hincha, D. K. (2008). LEA (late embryogenesis abundant) proteins and their encoding genes in Arabidopsis thaliana. BMC genomics, 9, 118.

Malik, A. A., Veltri, M., Boddington, K. F., Singh, K. K., & Graether, S. P. (2017). Genome Analysis of Conserved Dehydrin Motifs in Vascular Plants. Frontiers in plant science, 8, 709.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H., . . . Bateman, A. (2007). The Pfam protein families database. *Nucleic Acids Research*, *36*(Database).

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., … Rokhsar, D. S. (2011). Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, *40*(Database issue), D1178-86.

Altschul, S. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215*(3), 403-410. doi:10.1006/jmbi.1990.9999

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)*, *27*(7), 1017-8.

Furuki T, Sakurai M. Group 3 LEA protein model peptides protect liposomes during desiccation. Biochimica et biophysica acta. 2014;1838(11):2757–66. Epub 2014/07/19. doi: 10.1016/j.bbamem.2014.07.009

Commented [RADC5]:

Appendix

Table 1

E-value, number of sites and width of the motifs discovered

Motif	E-Value	Sites	Width
W-motif	1.3e-8373	871	14
LL-motif	7.5e-3738	574	12
RA-motif	1.7e-1324	544	8
MARS-motif	3.9e-1818	290	15

Table 2

Probability-weighted matrix (PWM) of the W-motif

		AMINO ACID																		
		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
POSITION	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0
	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.2	0.0	0.0	0.0	0.0	0.0
	4	0.0	0.0	0.8	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.4	0.0	0.0
	7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
	8	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	9	0.0	0.1	0.0	0.0	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
	10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.8
	11	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.1	0.0	0.1
	12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	13	0.1	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
	14	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0

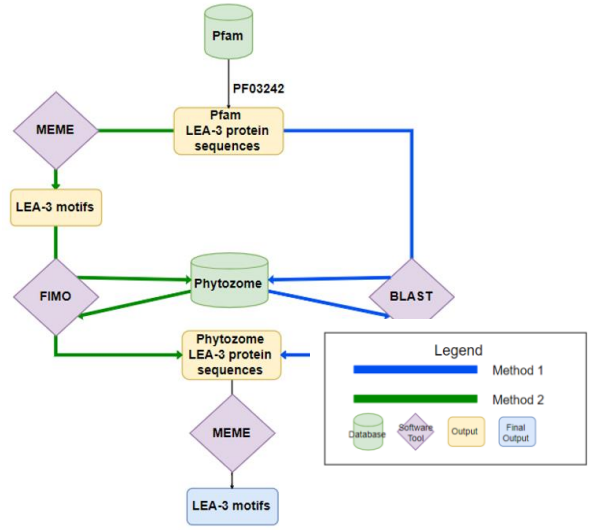
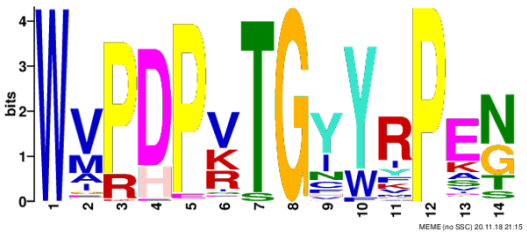


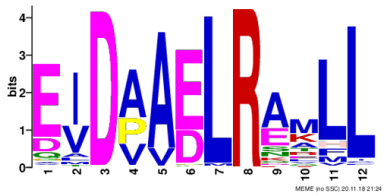
Figure 1 Selection of LEA-3 Proteins and Discovery of Protein Motifs.



Commented [RADC6]: Should I add more detail about the figure description

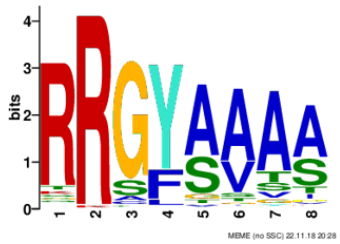
		Amino Acid																			
		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Position	1						0.00				0.00									1.00	
	2	0.12			0.04		0.03		0.08	0.01	0.01	0.21				0.00	0.01	0.00	0.50		
	3									0.01	0.00			0.82		0.17					
	4	0.00		0.79	0.02			0.18					0.00								
	5	0.00			0.02		0.01							0.96			0.01				
	6	0.00		0.02	0.02			0.00	0.03	0.23	0.01	0.00	0.01		0.03	0.20			0.44		
	7																0.05	0.95			
	8						1.00														
	9		0.09	0.02		0.05		0.05	0.18				0.10					0.00	0.02		0.48
	10					0.01														0.18	0.81
	11	0.03	0.00			0.06	0.01	0.00	0.09	0.08	0.00	0.04			0.00	0.54			0.05		0.09
	12													1.00							
	13	0.10		0.02	0.52	0.00	0.01		0.03	0.12	0.01	0.01		0.00	0.02	0.02	0.08		0.05		
	14	0.00		0.01			0.27		0.00	0.01			0.48		0.00		0.08	0.14			

Figure 2 Conservation of W-motif (A) LOGO representation of the W-motif created by MEME program. Amino acids are coloured based on their group type. Blue – hydrophobic (A, V, L, I, F, M,W), Green – polar (S, T, N, Q), Purple – negative (D, E), Red – positive (K, R, H), Cyan – Y, Orange – G, Yellow - P (B) Probability-weighted matrix (PWM) of the W-motif



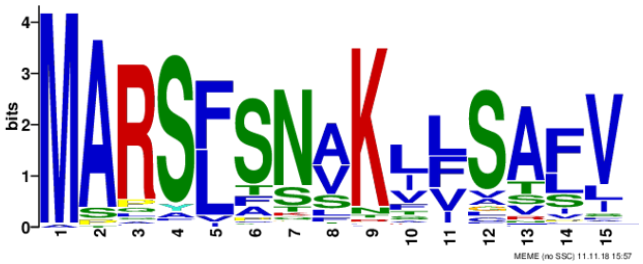
		Amino Acid																			
		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Position	1	0.03		0.13	0.72		0.02			0.01		0.00		0.00	0.06		0.01	0.00	0.02		
	2	0.04		0.00			0.00		0.51		0.04	0.04		0.00	0.00			0.02	0.33		
	3	0.00	0.01	0.99	0.00		0.00												0.00		
	4	0.48		0.01	0.00		0.01		0.00			0.00		0.26				0.00	0.23		
	5	0.85					0.00							0.00			0.00	0.00	0.14		
	6	0.00		0.27	0.70		0.02			0.00			0.00		0.01						0.00
	7										0.97	0.02		0.01	0.01		0.00		0.00		
	8		0.00					0.00			0.00				0.00	0.99					
	9	0.49		0.02	0.23			0.00		0.05			0.08		0.03	0.02	0.06	0.00	0.00		
	10	0.11			0.06				0.02	0.17	0.06	0.29	0.00		0.04	0.10	0.01	0.07	0.05	0.00	
	11					0.07		0.07	0.00		0.75	0.06		0.00	0.00	0.01		0.00	0.03		
	12	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.93	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00

Figure 2 Conservation of LL-motif (A) LOGO representation of the LL-motif created by MEME program. Amino acids are coloured based on their group type. Blue – hydrophobic (A, V, L, I, F, M,W), Green – polar (S, T, N, Q), Purple – negative (D, E), Red – positive (K, R, H), Cyan – Y, Orange – G, Yellow - P (B) Probability-weighted matrix (PWM) of the LL-motif



		Amino Acid																			
		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Position	1	0.01			0.00	0.00	0.02		0.00	0.03	0.01	0.01			0.02	0.84	0.02	0.03	0.01	0.01	0.00
	2							0.00		0.02			0.00		0.00	0.98					
	3	0.04	0.03				0.80					0.00		0.00		0.01	0.11	0.01	0.00		0.00
	4	0.01				0.21		0.00	0.01		0.04		0.01						0.01		0.72
	5	0.57	0.01				0.05	0.01	0.00	0.01							0.30	0.03	0.02		
	6	0.60					0.02	0.00	0.01	0.01	0.00	0.00			0.01		0.04	0.03	0.26		0.01
	7	0.67	0.00	0.00	0.01		0.04		0.00	0.01	0.00				0.01	0.00	0.08	0.10	0.06		0.02
	8	0.50			0.00		0.01		0.05	0.00		0.01		0.03	0.01	0.01	0.27	0.07	0.02		

Figure 3 Conservation of LL-motif (A) LOGO representation of the RA-motif created by MEME program. Amino acids are coloured based on their group type. Blue – hydrophobic (A, V, L, I, F, M,W), Green – polar (S, T, N, Q), Purple – negative (D, E), Red – positive (K, R, H), Cyan – Y, Orange – G, Yellow - P (B) Probability-weighted matrix (PWM) of the RA-motif



Position	Amino Acid																			
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	0.01							0.00			0.98									
2	0.89					0.00							0.02			0.07	0.02			
3	0.01	0.01				0.00	0.00			0.03			0.06	0.06	0.81	0.02				
4	0.03	0.00			0.02			0.01	0.01			0.01	0.01			0.87	0.01			0.04
5					0.47		0.01	0.06		0.39			0.00			0.01		0.06		0.00
6	0.12	0.02			0.07			0.00		0.01		0.01	0.02	0.01		0.65	0.08			
7	0.00		0.01			0.02		0.00	0.03	0.01		0.64	0.01	0.04		0.15	0.07			0.01
8	0.46	0.01			0.04			0.01		0.07			0.01			0.10	0.02	0.27		
9	0.02	0.01		0.01	0.01	0.01			0.89			0.03			0.02		0.02			0.00
10	0.04				0.08			0.17		0.31	0.03	0.02	0.01		0.05	0.06	0.06	0.16		
11	0.04				0.27			0.11		0.37			0.00	0.01		0.01		0.19		0.00
12	0.09	0.03			0.02	0.03				0.05			0.01		0.01	0.68	0.01	0.07		0.01
13	0.58	0.02	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.03	0.10	0.10	0.08	0.00	0.00
14	0.03			0.01	0.40	0.00		0.05		0.29			0.02	0.01	0.01	0.12	0.01	0.04		0.00
15	0.04	0.01			0.03	0.01		0.15		0.14			0.00			0.03		0.57		0.01

Figure 4 Conservation of LL-motif (A) LOGO representation of the MARS-motif created by MEME program. Amino acids are coloured based on their group type. Blue – hydrophobic (A, V, L, I, F, M,W), Green – polar (S, T, N, Q), Purple – negative (D, E), Red – positive (K, R, H), Cyan – Y, Orange – G, Yellow - P (B) Probability-weighted matrix (PWM) of the MARS-motif

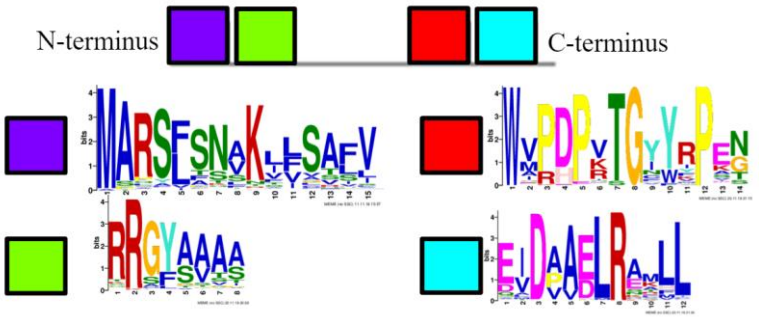


Figure 5 Position of the motifs generated by MEME program