2013

Uses of phage display in agriculture: Sequence analysis and comparative modeling of late embryogenesis abundant client proteins suggest protein-nucleic acid binding functionality

Rekha Kushwaha
University of Kentucky, rekha.kushwaha@uky.edu

A. Bruce Downie
University of Kentucky, adownie@uky.edu

Christina M. Payne
University of Kentucky, christy.payne@uky.edu

Click here to let us know how access to this document benefits you.

Follow this and additional works at: https://uknowledge.uky.edu/horticulture_facpub

Part of the Horticulture Commons

Repository Citation
Kushwaha, Rekha; Downie, A. Bruce; and Payne, Christina M., "Uses of phage display in agriculture: Sequence analysis and comparative modeling of late embryogenesis abundant client proteins suggest protein-nucleic acid binding functionality" (2013). Horticulture Faculty Publications. 3.
https://uknowledge.uky.edu/horticulture_facpub/3

This Article is brought to you for free and open access by the Horticulture at UKnowledge. It has been accepted for inclusion in Horticulture Faculty Publications by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.
Uses of phage display in agriculture: Sequence analysis and comparative modeling of late embryogenesis abundant client proteins suggest protein-nucleic acid binding functionality

Notes/Citation Information
Published in Computational and Mathematical Methods in Medicine, v. 2013, Article ID 470390, p. 1-11.

Copyright © 2013 Rekha Kushwaha et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Digital Object Identifier (DOI)
http://dx.doi.org/10.1155/2013/470390
A group of intrinsically disordered, hydrophilic proteins—Late Embryogenesis Abundant (LEA) proteins—has been linked to survival in plants and animals in periods of stress, putatively through safeguarding enzymatic function and prevention of aggregation in times of dehydration/heat. Yet despite decades of effort, the molecular-level mechanisms defining this protective function remain unknown. A recent effort to understand LEA functionality began with the unique application of phage display, wherein phage display and biopanning over recombinant Seed Maturation Proteins homologs from Arabidopsis thaliana and Glycine max were used to retrieve client proteins at two different temperatures, with one intended to represent heat stress. From this previous study, we identified 21 client proteins for which clones were recovered, sometimes repeatedly. Here, we use sequence analysis and homology modeling of the client proteins to ascertain common sequence and structural properties that may contribute to binding affinity with the protective LEA protein. Our methods uncover what appears to be a predilection for protein-nucleic acid interactions among LEA client proteins, which is suggestive of subcellular residence. The results from this initial computational study will guide future efforts to uncover the protein protective mechanisms during heat stress, potentially leading to phage-display-directed evolution of synthetic LEA molecules.
The characteristic intrinsically disordered structure and high hydrophilicity of the LEA proteins have been used to argue that they may act in a variety of ways to replace water (or compensate for its loss) in dehydrating tissues [12, 13]. Although there are two known LEA structures [14, 15], many of the proteins belonging to this family are dynamically disordered by design [16–18]. This has reasonably led to difficulties in obtaining structural information despite the use of a variety of techniques [19, 20], temperatures, and additives [17, 21]. Although obtaining crystal structures for most LEAs is not likely in the near future, structures of the preferential LEA client proteins may be estimated through homology modeling [22–24] as the same data allowing client protein identification also permits the identification of the region of the client protein to which the LEAs bind. Understanding which proteins are a particular LEA’s preeminent substrates of the LEA client proteins may provide insights into those functional processes most at risk for dehydration/thermal damage, suggesting novel ways forward in producing more drought-/heat-resistant species. Identification of hallmarks within the bound regions of LEA client proteins will provide the first clues as to which protein topologies are particularly prone to dehydration/heat damage. We hypothesize the regions require protection, which may be achieved through LEA protein binding.

Here, we report functional insights relative to LEA client proteins from application of sequence analysis and comparative modeling. Our examination focuses on identifying commonalities within the set of 21 putative LEA protein segments other than the bound regions of interest; however, for the purposes of this project, we limit discussion to the full-length client protein, whose common thread may be related to dynamic and chemical behavior more than structural or sequence similarity.

2. Comparative Modeling Methods

Previously, phage display with Arabidopsis seed cDNA libraries in T7 phage was used in biopans of recombinant Arabidopsis thaliana seed maturation protein 1 (SMP1) and its Glycine max homologue, GmPM28 [25] (LEA proteins). Biopanning was performed at 25°C and 41°C to identify proteins potentially involved in induction of secondary dormancy of Arabidopsis thaliana as a result of heat stress (see our companion manuscript for a brief synopsis of seed maturation). Figure 1 illustrates the 21 putative LEA client proteins identified through phage display. The proteins are labeled by the Arabidopsis Information Resource (TAIR) locus identifier. Within each plot, the LEA to which the protein binds and the temperature of the biopan are given. These proteins serve as the basis for our sequence analysis and comparative modeling investigation.

The full-length protein sequences to which LEA proteins of the Seed Maturation Protein family bound in phage display [25] were acquired from TAIR. Each protein was used in screens to identify homologs for which suitable three-dimensional (3D) structures had been solved. For the comparative modeling effort, we focused specifically on the regions identified as binding to the LEA homologues. Based on availability of 3D structures similar to these regions, the number of hits was narrowed down to 7 from the original 21 proteins being assessed (AT1G54870.1, AT1G75830.1, AT3G55170.1, AT3G58680.1, AT5G18380.1, AT5G41420.1, and AT5G46430.2).

Homology modeling of the seven LEA client proteins was performed using the Bioinformatics Toolkit from the Max-Planck Institute for Developmental Biology [26]. This suite integrates a number of utilities necessary to complete the modeling process. For each of the seven proteins, HHpred was used to predict secondary structure and sequence homology [27, 28]. HHblits was used to build multiple sequence alignments as input to the homology modeling software [29]. Homology modeling was performed using MODELLER [30].

Each protein used different templates for which the atomic coordinates were obtained from the RCSB Protein Data Bank [31]. Table 1 summarizes the templates used along with a brief description of each. For each of the models, the standard automated MODELLER procedure for structure modeling and optimization was used. This includes the initial rule-based determination of spatial restraints from the alignment and optimization through minimization of restraint violations. Several of the homology models generated include segments other than the bound regions of interest; however, for the purposes of this project, we limit discussion to the phage-display-recovered regions of the LEA client proteins. Visualization of the proteins and templates was accomplished with PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.).

Model quality was determined using the Protein Structure and Model Assessment Tool available through the SWISS-MODEL server [32, 33]. The estimated absolute model quality is reported here using the QMEAN Z-score in Table 1, which is an estimate related to reference X-ray crystallographic structures [34]. The reported Z-scores are standard deviations of the homology model relative to expected values from experimental structures.

3. Results and Discussion

Determination of similarity within the LEA client protein subset begins with analysis of similarities both within the bound region and the full-length client protein. For each of the client proteins identified through phage display as described previously [25], Figure 1 illustrates the hydrophilicity profile (Hopp/Woods analysis from ProtScale [35]) of the entire protein along with any identifiable protein domains. The figure has been divided to categorize the client proteins into those binding the most hydrophilic regions, those in which an identifiable protein motif has been bound, and those with no recognizable attributes having been bound. This preliminary analysis and classification of LEA client proteins have overlapping category members (signified by an asterisk, Figure 1). From this analysis, it is not immediately
Figure 1: Continued.
clear what, if anything, this set of proteins has in common. The full-length proteins are wildly variable in length (79–1608 amino acids) as are the regions containing the portion of the proteins to which the LEAs bind (24–183 amino acids). This latter attribute is consistent with the use of random hexamers to synthesize the phage display libraries [36]. Furthermore, the identifiable protein motifs do not appear to have commonality, though several ribosomal proteins appeared to preferentially bind to the LEA proteins. Interpretation of the hydrophilicity profiles is also mystifying, because while often the most or among the most hydrophilic of the proteins regions, it is placed in (a) (no recognizable attribute bound). If the LEA-bound fragment is neither the most hydrophilic nor encoding a recognizable motif, it is placed in (c) (no recognizable attribute bound). In each graph, the size, in amino acids, of the protein moiety bound by the LEA is provided as well as the number of independently acquired clones. If the clones were of different lengths, then the number of clones of a specific length is provided. Whether the clone was bound by SMP1 or GmPM28 and the temperature at which the binding occurred are also provided.

Evaluation exclusive to the regions containing those bound by the LEA proteins provides more insight into what the LEA client proteins may have in common. Amino acid composition, normalized by length of the region (Figure 2(a)), reveals that the bound regions have relatively low occurrences of aromatic residues, Phe, Trp, Tyr, and His, and sulfur-containing residues, Cys and Met, lending a general hydrophilicity to the region. Such an attribute would be consistent with the solvent-exposed exterior of a globular protein to which the LEA protein is presumed to bind. Lack of surface-exposed, thiol-containing, amino acids is not surprising given their tendency towards oxidation. Interestingly, the amino acid composition profile is consistent with that of a protein-nucleic acid complex data set examined by Baker and Grant [37]. Baker and Grant postulate that despite the low prevalence of aromatic residues within the binding sites of protein-nucleic acid complexes, aromatic residues still play a critical role in nucleic acid recognition. Relative to our observations, however, the binding site amino acid frequency of protein-nucleic acid complexes appears to be dominated by Arg, Lys, Asn, Glu, Gly, Ser, Thr, and Asp residues, providing us with a common thread potentially linking this set of LEA client proteins.

Further analysis of the hydrophobicity of the bound regions provides additional insight into potential functional relationships of the LEA client proteins. The grand average hydrophobicity (GRAVY) was determined for each of the LEA client proteins as shown in Figure 2(b). The GRAVY hydrophobicity is calculated based on the Kyte and Doolittle [38] hydrophathy values for each amino acid, the total of which is subsequently divided by the number of amino acids in the sequence to arrive at an average [35]. A negative value indicates hydrophilicity, and likewise, a positive value indicates hydrophobicity. We see that for a vast majority of
our LEA client proteins, the bound region is overwhelmingly hydrophilic, lending credence to the putative role of LEA proteins in the protection of client proteins from dehydration. However, two regions, part of ATIG75830.1 and AT4G04470.1, are identified as hydrophobic, and two others, ATIG65090.2 and AT5G44120.1, are only mildly hydrophilic. For all four of these bound regions, we confirmed that a single residue or subset of residues was not dominating the average, and rather, the hydrophobicity or mild hydrophilicity is indicative of the nature of the entire bound region (see Figure 1).

The entire set of full-length LEA client proteins was also analyzed using WOLF-pSORT (invoking the plant option), a program designed to predict protein localization sites [39]. Of the set, all but three were predicted to reside within a subcellular compartment containing nucleic acid polymers, with most predicted as either nuclear or cytoplasmic. The three outliers in the WOLF-pSORT analysis included ATIG75830.1, AT2G36640.1, and AT5G44120.1. ATIG75830.1 was predicted as extracellular. AT2G36640.1 was predicted to be peroxisomal, and AT5G44120.1 was predicted to be vacuolar. It is noteworthy that two of the three WOLF-pSORT outliers correspond to the hydrophobic or only mildly hydrophilic binding regions. This suggests, as does the amino acid composition, that perhaps an overall commonality of the remaining LEA client proteins is an ability to bind nucleic acids or at the very least promote interaction.

The amino acid sequences of the full-length LEA client proteins were analyzed using two separate protein motif/pattern and signature identification utilities with the aim of uncovering unifying motifs or functionality within the set. Table 2 summarizes the motifs and patterns uncovered, delineating between those that belong to the sequence region containing the bound moiety and those belonging to the full-length protein excluding this region. Putative amidation motifs (x-G-[RK]-[RK]) were identified using the patmat-motif utility within the EMBOSS software suite, which searches amino acid sequences against the PROSITE motif database [40]. PROSITE defines an amidation site (PS00009) as situated at the carboxy terminus of an active peptide in a larger precursor protein at the site of cleavage. Typically, peptidylglycine α-amidating enzyme (α-AE) can utilize the amino group from the C-terminal glycine in this motif to effect the conversion of the amino acid “x” to an amidated-(CO–NH$_2$) rather than a carboxylated-(COOH) terminus [41]. Nearly 60% of the full-length sequences contain at least one amidation domain (25% within the binding regions); however, relevance is difficult to determine at this point given the high natural probability of occurrence of this tetrapeptide sequence.

Using the InterProScan protein signature recognition software, potentially meaningful motifs, though not discernibly mutual, were established. The identification of the microbiomes C-terminal targeting signal domain, a tripeptide C-terminal consensus sequence occasionally found in peroxisomal proteins [42], in AT2G36640.1 is consistent with the prediction from WOLF-pSORT of this protein as peroxisomal. This is not unexpected, as pSORT algorithms use the SKL motif as recognition mechanism for peroxisomal proteins. The RGD tripeptide sequence motif was also returned in three separate instances, which is thought to promote binding to integrins and similar proteins [43] and appears to be critical in mediation of cell attachment [44]. The leucine zipper and coiled coil motifs were also repeatedly returned by InterProScan searches. The leucine zipper is a protein-protein motif of α-helices that dimerizes to form a coiled coil. The leucine zipper is known to participate in DNA-binding and regulation of gene expression [45], and the coiled coil is suspected to more generally participate in protein-protein interactions [46]. Less often, though interesting, nonetheless, the ATP/GTP A motif was returned by InterProScan. This motif, ATP/GTP A, is a glycine-rich loop sequence connecting a β-strand and an α helix, which has been identified as a conserved region of ATP- and GTP-binding proteins through observation of crystallographic data [47–52]. The loop region is known to interact with the phosphate groups of nucleotides. Finally, several proteins were identified as ribosomal which, along with RNA in protein-RNA interactions, assemble to form ribosomal subunits [53]. While there does not seem to be a single unifying motif or pattern among the set of phage-display-identified LEA client proteins, there does appear to
be a common thread of nucleotide interaction based upon known functionality of the identified patterns and motifs.

Within the amino acid sequences of the bound regions alone, PRATT was used to identify recurring patterns within the unaligned sequences (multiple sequence alignment of the diverse proteins being infeasible) [54]. Figure 3 illustrates the sequences of the bound LEA client protein regions, identified by the TAIR locus identifier. The sequences are coded with red and blue characters according to the two most commonly occurring patterns in the set of proteins. The K-x(2,4)-V-x(4)-[ACDGNSTV] pattern, represented in red, is found in 75% of the bound protein regions identified using phage display. In blue, the R-x(1,2)-R-x(0,1)-S pattern is common to 50% of the bound protein regions. PDBeMotif, a search algorithm providing statistics from 3D structural data, was used to interpret the significance of these two patterns [55]. For both patterns, PDBeMotif suggests—based on existing 3D structural data of proteins containing these sequence patterns—that glycerophosphate, ribose, and deoxyribose are among the structure-bound ligands. These sugars comprise the backbone of nucleic acid, and the presence of the phosphate in glycerophosphate is chemically consistent with an ester-linked phosphate of a nucleotide dimer (Figure 4). While this is by no means confirmation of the common functionality of the LEA client proteins, which can only be guaranteed through experimental means, sequence analysis and pattern/motif algorithms based on existing structural and functional data continually return to the theme of protein-nucleotide interactions.

The computational analysis of LEA client proteins concludes with homology modeling, where feasible, of the phage-display-bound regions of the LEA client proteins. As with the bioinformatics-based investigation, the intent here was to identify defining characteristics, either structural or chemical, which may yield insight as to why these proteins in particular are consistently returned as LEA-binding partners. Homology modeling methodology and identified structural templates are described in the methods section. As alluded to above, we were only able to successfully identify suitable 3D structural templates for seven of the LEA client proteins. Many of the LEA client proteins, including some of those modeled here, exist as membrane-bound proteins and are thus difficult to resolve structurally. The seven LEA client proteins include AT1G54870.1, AT1G75830.1, AT3G55170.1, AT3G58680.1, AT5G18380.1, AT5G44120.1, and AT5G46430.2. We anticipate that as crystallographic methods continue to develop, additional structures will become available to serve as templates to the remaining client proteins. Several other templates were available for portions of the full-length proteins; however, we are restricting this homology modeling study to that within the bound regions, as this should intuitively provide the most information.
> AT1G5280.1
> AT1G5280.1
> AT1G5310.1

> AT1G5487.0.1

> AT1G58210.1

> AT1G6600.1.1

> AT1G5301.3

> AT1G6290.2

> AT1G7300.1

> AT2G02160.1

> AT2G04470.1

> AT2G12400.2

> AT2G18380.1

> AT2G31410.1

> AT2G36640.1

> AT3G0690.1

> AT3G5170.1

> AT3G58680.1

> AT3G63400.3

> AT4G04470.1

> AT4G41200.2

> AT5G18380.1

> AT5G44120.1

> AT5G46430.2

> AT5G52240.1

---

**Figure 3**: Amino acid sequences of the bound regions of the LEA client proteins. Recurring patterns within the set of sequences have been identified by red and blue text. The red characters indicate the K-x(2,4)-V-x(4)-[ACDGNSTV] pattern. Blue characters indicate the R-x(1,2)-R-x(0,1)-Spattern.

**Figure 4**: Chemical structure of a nucleotide dimer, left, and glycerol 3-phosphate (glycerophosphate), right. The red lettering on the nucleotide dimer represents the chemical similarity to the glycerophosphate molecule.
regarding features contributing to the protein-protein interactions. Figure 5 illustrates the sequences of the seven homology models annotated according to the predicted secondary structure. Below the sequences, cartoon representations of each model are provided.

Structural or sequential representation of these seven protein regions does not provide a striking explanation as to which attribute is acting as a functional link. Several of the structures exhibit relatively large expanses of disordered loop regions, which seem rather uncharacteristic of globular proteins. This is almost certainly related to the hydrophilic nature of the bound regions (Figure 2(a)). We do find it somewhat intriguing, however, that of the three proteins returned by WOLF-pSORT as being neither nuclear nor cytoplasmic, two (AT3G55170.1 and AT1G75830.1) homologous protein structures are available through the Protein Data Bank, though we cannot ascribe significance to this based on the data presented here. With our limited subset of binding site homology models, we can only state that the structures appear to vary significantly from one another and that commonality may lie more in the chemical and dynamical rather than the structural nature of the region.

### 4. Conclusion
A great deal of information relating both sequence and structure of the LEA client protein bound regions and how this contributes to binding remains to be determined. From this initial computational study aiming to shed light on the functional role of LEA proteins through similarity in their bound substrates, we have uncovered what seems to be a predilection for protein-nucleic acid interaction in the

---

**Figure 5:** Seven homology models of LEA client proteins, focused on the regions containing the protein moiety to which the LEA proteins bound, were developed. The sequences, numbered by the full-length protein TAIR locus identifier, are shown annotated by secondary structure elements. Secondary structure annotation was accomplished using the ESPript web utility [59]. β-Sheets are labeled with a solid black arrow, α-helices with medium curly script, β-turns with TT, and 3_10-helices (η) with small curly script. Sequence number is also indicated in frequency of ten and corresponds to that of the full-length sequence. Below the sequences, the seven homology models of the bound regions only are shown in cartoon representation. The homology models are labeled, as with the sequences, according to the TAIR locus identifier of the full-length protein to which they belong. The homology model PDB files have been included in Supplementary Materials available online at http://dx.doi.org/10.1155/2013/470390.
Table 2: Patterns and motifs identified using InterProScan and the patmatmotifs utility as part of the EMBoss package. The full-length protein is identified by the TAIR locus identifier. Patterns and motifs have been separated according to their position either within the binding region or exclusive of the binding region.

<table>
<thead>
<tr>
<th>TAIR locus</th>
<th>Full-length (excludes binding region)</th>
<th>Binding Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATIG15280.1</td>
<td>Amidation motif</td>
<td>Amidation motif</td>
</tr>
<tr>
<td>ATIG15290.1</td>
<td>Leucine zipper Coiled coil</td>
<td>Amidation motif</td>
</tr>
<tr>
<td>ATIG16610.1</td>
<td>Amidation motif</td>
<td>RGD</td>
</tr>
<tr>
<td>ATIG30610.1</td>
<td>ATP/GTP A Ribosomal L32e</td>
<td>Zinc finger plant homeodomain</td>
</tr>
<tr>
<td>ATIG32810.2</td>
<td>Amidation motif (4)</td>
<td>Coiled coil</td>
</tr>
<tr>
<td>ATIG54870.1</td>
<td>—</td>
<td>Short-chain dehydrogenase</td>
</tr>
<tr>
<td>ATIG5530.3</td>
<td>Amidation motif (2)</td>
<td>—</td>
</tr>
<tr>
<td>ATIG58210.1</td>
<td>Leucine zipper</td>
<td>Amidation motif (2)</td>
</tr>
<tr>
<td>ATIG65090.2</td>
<td>Amidation motif</td>
<td>—</td>
</tr>
<tr>
<td>ATIG75830.1</td>
<td>—</td>
<td>Gamma thionin</td>
</tr>
<tr>
<td>ATIG15260.1</td>
<td>Amidation motif Coiled coil</td>
<td>Amidation motif</td>
</tr>
<tr>
<td>ATIG19540.1</td>
<td>Amidation motif</td>
<td>—</td>
</tr>
<tr>
<td>ATIG3140.1</td>
<td>Coiled coil</td>
<td>—</td>
</tr>
<tr>
<td>ATIG36640.1</td>
<td>Microbodies C-ter Coiled coil (2)</td>
<td>Coiled coil</td>
</tr>
<tr>
<td>ATIG51690.1</td>
<td>—</td>
<td>Amidation motif</td>
</tr>
<tr>
<td>ATIG55170.1</td>
<td>—</td>
<td>Ribosomal L29</td>
</tr>
<tr>
<td>ATIG58680.1</td>
<td>Coiled coil</td>
<td>Amidation motif</td>
</tr>
<tr>
<td>ATIG63400.3</td>
<td>Prolyl-peptidyl isomerase ATP/GTP A (2) Amidation motif (3)</td>
<td>—</td>
</tr>
<tr>
<td>ATIG40470.1</td>
<td>Amidation motif</td>
<td>—</td>
</tr>
<tr>
<td>ATIG12400.2</td>
<td>Coiled coil</td>
<td>RGD</td>
</tr>
<tr>
<td>ATIG51800.1</td>
<td>Amidation motif Ribosomal</td>
<td>—</td>
</tr>
<tr>
<td>ATIG4120.1</td>
<td>I1s seed storage</td>
<td>—</td>
</tr>
<tr>
<td>ATIG4630.2</td>
<td>—</td>
<td>Ribosomal L32e</td>
</tr>
<tr>
<td>ATIG52240.1</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

LEA client proteins. While this does not yet tell us how the LEA proteins function relative to the bound protein regions, it does suggest hypotheses to be tested concerning the subcellular residence of the LEA proteins under study. An evolutionary relationship between the LEA protein and the substrate protein dictates that the SMPI and GmPM28 homologs be located in subcellular compartments containing the nucleic acid polymers to which their client proteins apparently bind (i.e., the nucleus, cytoplasm, plastids, and/or mitochondria). Phenotypic consequences for specific LEA protein (or LEA protein family) reductions [36, 56], as well as a demonstration that LEA protein homologs from the Seed Maturation Protein family have preferred client proteins to which they bind [25], suggest that at least some LEA proteins are not redundantly backed up, indiscriminate spacer molecules, and lead to the conclusion that other LEA proteins will also have preferred binding partners. The elucidation of the subfunctionalization of specific LEA proteins concerning which client proteins they bind to is most efficaciously performed using phage display.

In the near term, molecular dynamics simulations of the LEA client protein homology models, including the full-length domains, may provide additional insight into the flexibility and solvation dynamics of the proteins, in addition to directing ongoing experimental phage display efforts. The long-term focus will be on the development of additional homology models as more crystallographic structures become available as well as de novo protein design using rapidly developing structure prediction methods. Our continuing aim is the effective integration of computational modeling with phage display for the prediction of protein structures at risk for dehydration or heat damage, uncovering the mechanisms by which LEA proteins perform their protective function. Future endeavors could conceivably encompass phage-display-directed evolution of synthetic LEA proteins engineered to protect labile proteins.

Acknowledgments

This project was partially funded by an NSF IOS (0849230), Hatch, McIntire-Stennis (AD421 CRIS), USDA Seed Grant (2011-04375), and Sir Frederick McMaster Research Fellowship to A. Bruce Downie. The authors thank Stephen Chmely for his assistance with figure preparation.

References


