BMC Bioinformatics

Research article

LEAping to conclusions: A computational reanalysis of late embryogenesis abundant proteins and their possible roles

Michael J Wise*

Address: Department of Genetics Cambridge University Cambridge U.K

Email: Michael J Wise* - mw263@cam.ac.uk

* Corresponding author

Published: 29 October 2003

BMC Bioinformatics 2003, **4**:52

Received: 29 May 2003

Accepted: 29 October 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/52

Abstract

Background: The late embryogenesis abundant (LEA) proteins cover a number of loosely related groups of proteins, originally found in plants but now being found in non-plant species. Their precise function is unknown, though considerable evidence suggests that LEA proteins are involved in desiccation resistance. Using a number of statistically-based bioinformatics tools the classification of a large set of LEA proteins, covering all Groups, is reexamined together with some previous findings. Searches based on peptide composition return proteins with similar composition to different LEA Groups; keyword clustering is then applied to reveal keywords and phrases suggestive of the Groups' properties.

Results: Previous research has suggested that glycine is characteristic of LEA proteins, but it is only highly over-represented in Groups 1 and 2, while alanine, thought characteristic of Group 2, is over-represented in Group 3, 4 and 6 but under-represented in Groups 1 and 2. However, for LEA Groups 1 2 and 3 it is shown that glutamine is very significantly over-represented, while cysteine, phenylalanine, isoleucine, leucine and tryptophan are significantly under-represented. There is also evidence that the Group 4 LEA proteins are more appropriately redistributed to Group 2 and Group 3. Similarly, Group 5 is better found among the Group 3 LEA proteins.

Conclusions: There is evidence that Group 2 and Group 3 LEA proteins, though distinct, might be related. This relationship is also evident in the overlapping sets of keywords for the two Groups, emphasising alpha-helical structure and, at a larger scale, filaments, all of which fits well with experimental evidence that proteins from both Groups are natively unstructured, but become structured under stress conditions. The keywords support localisation of LEA proteins both in the nucleus and associated with the cytoskeleton, and a mode of action similar to chaperones, perhaps the cold shock chaperones, via a role in DNA-binding. In general, non-globular and low-complexity proteins, such as the LEA proteins, pose particular challenges in determining their functions and modes of action. Rather than masking off and ignoring low-complexity domains, novel tools and tool combinations are needed which are capable of analysing such proteins in their entirety.

Background

The late embryogenesis abundant (LEA) proteins cover a number of loosely related groups of proteins whose precise function is unknown. While considerable evidence suggests that LEA proteins are involved in desiccation resistance, a variety of mechanisms for achieving this end

have been proposed including protecting cellular structures from the effects of water loss by retention of water, sequestration of ions, direct protection of other proteins or membranes, or renaturation of unfolded proteins [1–4]. LEA proteins are primarily found in plants, where they were originally found in seeds [5–7], and then other plant tissues. In addition, a number of putative LEA genes have been found in a non-plant species, including eubacteria *Haemophilus influenzae* and *Bacillus subtilis* [8], extremophile *Deinococcus radiodurans* [9] and the nematodes *Caenorhabditis elegans* and *Aphelenchus avenae* [10]. Most of the literature to date on LEA proteins has been in the form of reports on individual LEA proteins with general surveys appearing some time ago [1,11,12]. The somewhat more recent survey by Close [13] of Group 2 LEA proteins also includes a discussion of predicted secondary structure for this Group.

LEA proteins are generally grouped on the basis of their similarity to prototypical LEA proteins from the cotton plant *Gossypium hirsutum*. In the Dure naming scheme, LEA protein groups are named after particular *G. hirsutum* cDNA clones, resulting in Group names such as D7, D11, D19, D95 and D113. Many authors since Dure, however, use an assignment to Groups originating with [12], though revised (and to some extent contradictory) assignments also appear in [3] and [4]. There is, however, a consensus only for three LEA protein groups: Group 1 (D19), Group 2 (also known as dehydrins, D11) and Group 3 (D7). Other LEA protein groups from [12] are Group 4 (D113), Group 5 (D29) and Group 6 (D34). Four of the LEA protein groups are also represented by Pfam [14] domain families:

• Small Hydrophilic Plant Seed Protein (PF00477) – Group 1

• Dehydrin (PF00257) – Group 2

• LEA (PF02987) – Group 3

• LEA-1 (PF03760) – Group 4

In addition, there are groups which do not appear in the Bray [1] scheme: Lea5 (D73) and Lea14 (D95) [15], although both are represented by Pfam families: Lea5(D73) by LEA-3, PF03242, and Lea14(D95) by LEA-2, PF03168.

Previous work, using just amino acid percentage composition and the Kyte Doolittle hydrophobicity metric, found that LEA proteins are characterised by a preponderance of hydrophilic amino acids together with high glycine content, resulting in their characterisation as "hydrophilins" [16]. Certain LEA protein Groups are also said to be rich in alanine, but deficient in cysteine and tryptophan [3,4].

However, a significant, though often overlooked, feature of LEA proteins is that the majority are low complexity proteins. This is amply demonstrated through the use of the low complexity sequence demarcation tool, 0j.py [17], which was applied, first to all the sequences above 40aa in SwissProt and SpTrEMBL (also called Swall) and then to a database of 112 LEA proteins, which will be described shortly. The sequences in the large database returned a median score of 3, with 13% having a score of 0 and 32% a score greater than then 3; a low score implies that the protein has high sequence complexity. By contrast, the LEA sequences had a median score of 11.5, and 80% return a score greater than 3 (equivalent to a p-value of 1.1×10^{-25}).

Low complexity sequences pose a particular problem for the local alignment tools such as BLAST which owe much of their discriminative power to scoring schemes based on the extreme value distribution [18]. For example, [19] compares the efficacy of both BLAST and FASTA with an implementation of the Smith-Waterman algorithm, each both with and without the use of scoring schemes based on the extreme-value distribution. The benefit of having statistically based scoring schemes is conclusively demonstrated [19]. However, it is well known that low complexity sequences prejudice extreme value distribution based statistical scoring [20]. The standard way of dealing with low complexity regions in the context of database searches is to mask these off in the query sequence using applications such as SEG [21]. When SEG was run across the set of 112 LEA proteins, 11 high complexity sequences are returned unaltered; the remainder were masked to a greater or lesser extent, with 57 having between 30% and 71% of their amino acids masked. The first effect of masking is to reduce the number of amino acids available for alignment. The second effect is to produce an asymmetry, because only the query sequence is masked, not the target (i.e. database) sequences, so the answer you obtain for an alignment between a masked query and the target sequence depends on which sequence is the query and which is the target.

The aims of this resurvey were therefore twofold. The first aim was to create a sizable set of the LEA proteins spanning all the Groups and then, using a number of software tools to lessen the impact of low sequence complexity, to reexamine the classification of this diverse set of proteins. In the light of this process, the previous findings are reviewed and expanded. Secondly, searches based on peptide composition were used to reveal proteins with similar composition to different LEA Groups; keyword clustering was then applied to the lists of search hits to suggest

Table 1: LEA Protein Group 1 (D19) Exemplar(s): LE19_GOSHI

ID	Species	Tissue	Expression	Pep	SF	Evidence
EM1_ARATH	ARATH	Seed	ABA, **Canon**		4	PF00477_hmm; L194_HORVU (1e-67)
EM1_WHEAT	WHEAT	Seed	ABA, **Canon**	1(1)	6	PF00477_ma
EM2_WHEAT	WHEAT	Seed	ABA, **Canon**	1(1)	6	PF00477_hmm, EMP1_ORYSA (2e-41)
EM6_ARATH	ARATH	Seed	ABA, **Canon**	1(1)	6	PF00477_ma; EMB1_DAUCA (9e-38)
EMB1_DAUCA	DAUCA	Seed	**Canon**	1(1)	4	PF00477_hmm
EMB5_MAIZE	MAIZE	Seed	ABA, **Canon**	1(1)	6	PF00477_hmm
EMP1_ORYSA	ORYSA	Seed	ABA, **Canon**	1(1)	6	PF00477_hmm
L193_HORVU	HORVU	Seed	ABA, notCold, **Canon**, Mannitol	1(3)	4	PF00477_hmm
L194_HORVU	HORVU	Seed	ABA, notCold **Canon**, Mannitol	1(4)	4	PF00477_hmm
L19A_HORVU	HORVU	Seed	ABA, notCold, **Canon** Mannitol, Salt	1(1)	6	PF00477_hmm
L19B_HORVU	HORVU	Seed	ABA, notCold, **Canon**, Mannitol, Salt	1(1)	6	PF00477_hmm
LE10_HELAN	HELAN	Seed	ABA, **Canon**, Mannitol		4	PF00477_hmm
LE19_GOSHI	GOSHI	Seed	ABA, **Canon**		4	PF00477_ma
SEEP_RAPSA	RAPSA	Seed	**Canon**	1(1)	3,6 (280)	PF00477_hmm PF00477_hmm

LEA Group 1 proteins, with the exemplar being LE19_GOSHI. The columns are: 1) the protein identifier, 2) a code for the species (see Table 10), 3) the tissue(s) in which it has been found, 4) the conditions that give rise (or fail to give rise) to the expression of the gene, 5) whether the LEA Group 1 motif is detected using agrep and the number of times it is found, 6) the superfamilies/stand-alone clusters in which the protein is found and 7) other evidence for accepting the protein as LEA Group 1.

keywords and phrases indicative of the Groups' functions. These are the starting point for current and future experimental work.

Results

The Rules Induced by Supervised Learning Application

The input to supervised machine learning application, Ripper, for each LEA protein was therefore 13 values (3 hydrophobicity; 3 predicted secondary structure and 7 amino acid class) plus the Group to which the protein had been assigned. The output was a set of rules for classifying putative LEA proteins into Groups based on the 13 values. When working on real-world (i.e. noisy) data all rule induction algorithms attempt to balance accuracy/correct-predictions with conciseness; at the extreme one could have 100% accuracy by creating a rule for each input protein, while at the other extreme one can achieve maximum conciseness by having a single rule predicting the largest output category, which would in this case mean categorising every input LEA protein as Group 2. Ripper was run several times until the error on the input set was minimised. Extra conditions were then added by hand to the rules to deal with the misclassified proteins until no further rules could be added without generating other misclassifications. The final rule set, which appears in Table 11, should be understood as operating in a top-down, if .. else if, manner.

The reader will have noticed that the table of Group 2 LEA proteins (Table 2) has been partitioned into three subsets; these correspond to the three rules under which Group 2 proteins are classified using the above rule-set. The rules have been labelled 2a, 2b and 2c. Notice that the Group 2 LEA proteins induced by cold stress are predominantly characterised by Rules 2b and 2c (particularly 2c), while the Group 2 proteins which have been shown not to be up-regulated by cold stress and all the canonical LEA proteins are encompassed by Rule 2a.

Four of the proteins would appear to have been misclassified: LE11_HELAN and LE25_LYCES are generally considered to be Group 4 (D113) based on the assignment in Dure (1993), but have here been assigned to Group 3 on the basis of their high predicted percentage alpha-helical content (0.6 and 0.56 versus a threshold of 0.34). While some care needs to be taken because Group 4 is the default category when all others rules have failed, the three Group 4 proteins covered by the default rule all have predicted percentage loop content greater than or equal to 0.25, while the two classified as Group 3 have loop content less than or equal to 0.12. In other words, there would appear to be other grounds for suspecting that LE11_HELAN and LE25_LYCES are not in the same Group as the three proteins assigned to the default, Group 4.

The other apparently misclassified proteins are LE29_GOSHI and Q93Y63, which are classified by Bray (1993) as Group 5, but which have been classified as Group 3 here. This is in line with recent reclassifications of Group 5 (D29) LEA proteins as Group 3 [4,3], although the Group is retained as a separate entity in [22]. Members of the former Group 5 have the same domain composi-

Table 2: LEA Protein Group 2 (D11) Exemplar(s): DH11_GOSHI

ID	Species	Tissue	Expression	Pep	SF	Evidence
DH11_GOSHI	GOSHI	Seed	ABA, **Canon**	2Y(2), 2S(9), k(4)	1	PF00257_ma
DH14_LYCES	LYCES	Root, Stem Leaf	ABA, Salt notCold	2Y(1), 2K(1), 2S(5)	1	PF00257_ma; DH1B_ORYSA (5e-17)
DH15_WHEAT	WHEAT	Root	ABA, Desc	2Y(1), 2K(2) 2S(8)	1	PF00257_ma; DH1B_ORYSA (1e-37)
DH18_ARATH	ARATH	Leaf Stem	ABA, Desc notCold	2Y(2), 2K(2) 2S(5)	1,3	PF00257_ma EC40_DAUCA (3e-25)
DH1B_ORYSA	ORYSA	Seed, Shoot	ABA, **Canon**, Salt	2Y(2), 2K(2), 2S(8)	1,10	PF00257_ma
DH1C_ORYSA	ORYSA	Seed, Shoot	ABA, **Canon**, Salt	2Y(2), 2K(2), 2S(8)	1,10	PF00257_hmm
DH1_MAIZE	MAIZE	Shoot	ABA, Desc	2Y(1), 2K(2) 2S(7)	1	PF00257_ma DH1B_ORYSA (4e-47)
DH25_ORYSA	ORYSA	Callus	ABA, Desc notCold	2Y(2), 2K(2), 2S(9), k(3)	1,8	PF00257_ma DHLE_RAPSA (2e-26)
DHA_CRAPL	CRAPL	Leaf	ABA, Desc	2K(1), 2S(7)	10	PF00257_ma; DHLE_RAPSA (8e-20)
DHB_CRAPL	CRAPL	Leaf	ABA, Desc	2Y(1), 2K(2) 2S(8)	1	PF00257_ma; DH1B_ORYSA
DHLE_RAPSA	RAPSA	Seed	**Canon**	2Y(3), 2K(1) 2S(7)	1	PF00257_ma
DHN1_PEA	PEA	Shoot Cotyledon	ABA, Desc notCanon	2Y(3), 2K(2)	1	PF00257_ma DH1B_ORYSA (1e-17)
EC40_DAUCA	DAUCA	Seed, Embryo cells	ABA, **Canon**	2Y(3), 2K(1) 2S(6)	1	PF00257_hmm;
O22623	VACCO	Floral buds, Leaf	Cold	None	1,3	DH1B_ORYSA (1e-5)
O65216	WHEAT	Leaf, Root Crown, Seed	ABA, Cold, Desc	2K(6)	1,3	PF00257_hmm; DH1B_ORYSA (8e-18)
P93701	VIGUN	Leaf	ABA, notCold Desc, Salt	2Y(2), 2K(1)	1	PF00257_hmm; EC40_DAUCA (6e-23)
Q39937	HELAN	Leaf	ABA, Desc	2Y(3), 2K(1), 2S(7)	1	PF00257_hmm; EC40_DAUCA (1e-38),
Q39938	HELAN	Leaf	ABA, Desc	2K(2), 2S(5), k(3)	1	PF00257_hmm; DH1B_ORYSA (6e-21)
Q40331	MEDFA	Callus	notABA, Cold, notDesc	2K(1)	1	PF00257_hmm; DH1C_ORYSA (1e-7)
Q40968	PRUPE	Bark	Cold, Desc	2Y(2), 2K(4)	1	PF00257_hmm; EC40_DAUCA(3e-13)
Q41306	SOLCO	Leaf, Stem	ABA, Cold	2Y(2), 2K(2), 2S(7), k(3)	1	PF00257_hmm; DHLE_RAPSA (9e-29)
Q41451	SOLTU	Leaf, Stem	ABA, Cold	2Y(2), 2K(2), 2S(7), k(3)	1	PF00257_hmm; DHLE_RAPSA (2e-29)
Q9SBI7	HORVU	Seedling	Desc, ABA, Cold	2K(9)	1,3 (295)	PF00257_hmm; DH1B_ORYSA (5e-16)
Q9SPL8	VIGUN	Seed	Cold	2Y(2), 2K(1)	1	PF00257_hmm; EC40_DAUCA (1.8e-25)
Q9ZTR2	HORVU	Seedling	Desc, notABA notCold	2Y(2), 2K(3), 2S(9)	1	PF00257_hmm; DH1C_ORYSA (4e-33)
Q9ZTR3	HORVU	Seedling	Desc, ABA notCold	2Y(1), 2K(2), 2S(8)	1	PF00257_hmm; DH1B_ORYSA (2e-44)
Q9ZTR4	HORVU	Seedling	Desc, ABA, notCold	2Y(1), 2K(2), 2S(7)	1	PF00257_hmm; DH1C_ORYSA (1e-47)
Q9ZTR5	HORVU	Seedling	Desc, ABA, notCold	2Y(2), 2K(3), 2S(9)	1,9	PF00257_hmm; EC40_DAUCA (1e-43)
COR4_WHEAT	WHEAT	Root, Leaf Crown	ABA, Cold Desc	2K(1), 2S(9), k(7)	3	PF00257_ma, EC40_DAUCA (1e-7)
CS12_WHEAT	WHEAT	Shoot	Cold, notABA notDesc	2K(6)	1,3	PF00257_ma, EC40_DAUCA (8e-17)
CS66_WHEAT	WHEAT	Shoot	Cold, notABA notDesc	2K(5)	1,3	PF00257_hmm, DH1B_ORYSA (6e-17)
DH14_ARATH	ARATH	Leaf, Stem, Root, Seed, Flower	ABA, Desc notCanon Cold	2K(2), 2S(7), k(10)	3	PF00257_ma; EC40_DAUCA (2e-10)
DH1D_ORYSA	ORYSA	Seed Shoot	ABA Salt	2Y(1), 2K(2), 2S(4)	1	PF00257_ma; DH1B_ORYSA (9e-43)
DH1_HORVU	HORVU	Shoot	ABA, Desc Desc	2Y(1), 2K(2), 2S(7)	1	PF00257_ma; DH1B_ORYSA (2e-35)
DH21_ORYSA	ORYSA	Seed	ABA, Desc	2Y(1), 2K(2), 2S(7)	1	PF00257_hmm; DH1B_ORYSA (3e-50)
DH2_HORVU	HORVU	Shoot	ABA, Desc	2Y(1), 2K(2), 2S(7)	1	PF00257_hmm; DH1C_ORYSA (4e-35)
DH3_HORVU	HORVU	Shoot	ABA, Desc	2Y(1), 2K(2), 2S(7)	1	PF00257_ma; DH1C_ORYSA (7e-45)
DH4_HORVU	HORVU	Shoot	ABA, Desc	2Y(1), 2K(2), 2S(7)	1	PF00257_hmm; DH1C_ORYSA (2e-38)

Table 2: LEA Protein Group 2 (D11) Exemplar(s): DH11_GOSHI *(Continued)*

DH47_ARATH	ARATH	Leaf, Stem Seed	ABA, Cold, Desc, notCanon	2K(3), 2S(7), k(4)	3	PF00257_hmm DHLE_RAPSA (2e-12)
DHX2_ARATH	ARATH	Leaf, Stem	Cold, weak ABA weak Desc	2K(3)	1	PF00257_hmm
O64939	LOPEL	Root	Salt	2K(6)	1,3,9	PF00257_hmm; DH1B_ORYSA (7e-16)
Q41347	STELP	Leaf	ABA, Desc, PEG	2K(1), 2S(5), 2S(6), k(3)	8	PF00257_hmm; DHLE_RAPSA (4e-8)
Q42409	TRITU	Root, Shoot	ABA, Desc	2K(2)	1	PF00257_hmm; DH1C_ORYSA (6e-16)
Q43488	HORVU	Leaf	notABA, Cold Desc	2K(1), 2S(9), k(10)	3	PF00257_hmm; DH1B_ORYSA (4e-9)
DH10_ARATH	ARATH	Leaf, Stem, Root, Seeds Flower	weak ABA, weak Desc, notCanon Cold	2K(2), 2S(7), k(11)	3	PF00257_ma EC40_DAUCA (4e-9)
O04232	SOLTU	Tuber	Cold	2K(2), 2S(9), k(9)	3	PF00257_hmm; DH1C_ORYSA (3e-11)
O48622	SPIOL	Shoot	Cold, Desc	2K(4), k(3)	3, (280)	EC40_DAUCA (1e-6)
Q41091	PONTR	Leaf Leaf	Cold, notSalt notDesc	2S(5), k(8)	3	DH1C_ORYSA (1e-8)
Q9XEL3	PICGL	Bud, Stem	ABA, Cold, Desc	2K(3), 2S(8), k(12)	3	PF00257_hmm; DH1C_ORYSA (2e-11)
Q9ZR21	CITUN	Leaf	Cold	2S(5), k(9)	3	DH1C_ORYSA (1e-9)

LEA Group 2 proteins, with the exemplar being DH11_GOSH. The columns are: 1) the protein identifier, 2) a code for the species (see Table 10), 3) the tissue(s) in which it has been found, 4) the conditions that give rise (or fail to give rise) to the expression of the gene, 5) whether any of the Close LEA Group 2 motifs 2Y, 2K or 2S, or poly-lysine stutters are detected using agrep and the number of times each is found, 6) the superfamilies/stand-alone clusters in which the protein is found and 7) other evidence for accepting the protein as LEA Group 2.

tion as Group 3 LEA proteins, but with additional copies of those domains.

The classification rules described above were applied to the set of uncharacterised LEA proteins. As a result, O24439 is predicted to be a member of the first set of Group 2 LEA proteins by Rule 1, while Q9S7S3 is predicted to be in the Lea5/D73 Group by Rule 6 and O81483 is predicted to be Group 6 by Rule 8.

Results from the POPP Analysis of LEA Proteins by Group

Table 12 lists a selection of the most significant peptides which result from placing the sequences corresponding to the different Groups into separate databases and having popp_create.py applied to each such database. A negative p-value indicates a significant under-representation. Some care must be taken interpreting the probabilities generated by the binomial distribution statistic because larger datasets will give rise to much more significant p-values. For that reason, only those p-values that are less than a threshold are now considered, where the threshold is determined from the mean below-threshold log-probability value (i.e. average log probability for p-values less than 0.05) across the respective datasets. For these purposes, p-values above 0.05 are said to be significant, but those above the dataset mean value for each Group will be described as highly significant. If just the first three, more hydrophilic Groups are considered, the list of highly significant peptides found in all the groups is: -C, -F, +GE, -I, -L, -N, +Q and -W, where '+' before a peptide indicates over-representation, while a '-' indicates under-representation. In all three Groups, charged/polar residues feature highly; K is very highly represented in Groups 2 and 3, and moderately so ($9.\ 7 \times 10^{-6}$) in Group 1. Group 1 also evidences highly significant over-representation of R. Similarly E is highly found in Groups 1 and 3, but is not highly over-represented in Group 2 ($4.\ 9 \times 10^{-13}$). Of the other characteristics, glycine is highly represented in Group 1 and Group 2. However, in Group 3 glycine is found only marginally more than expected by chance (p-value 0.012). Overall, the description of these Groups as hydrophilins is not completely borne out; they are indeed characterised by hydrophilic residues, but glycine is only highly expressed in two of the three Groups.

The list of highly significant peptides confirms the previous finding that cysteine is lacking in Group 1, 2, 3 and 4 LEA proteins [3]. In the current dataset, 86 of the 112 sequences had no cysteine residues at all, 17 had just one, six have two and only one each have, respectively, three, four and five, cysteine residues. Similarly for tryptophan, 91 sequences had no tryptophan residues, 12 have one tryptophan residue, seven have two, one sequence has three and one has four.

Another previous finding is that Group 2 LEA proteins are rich in glycine or alanine and proline [4]. As noted above, G is highly significant for this group (in fact extremely so – p-value 0). On the other hand, A and P are under-represented, respectively $-\ 1.\ 1 \times 10^{-14}$ and $-\ 1.\ 3 \times 10^{-7}$.

Table 3: LEA Protein Group 3 (D7) Exemplar(s): LE7_GOSHI, LE76_BRANA

ID	Species	Tissue	Expression	Pep	SF	Evidence
DRPF_CRAPL	CRAPL	Leaf	ABA, Desc	3(1), k(3)	2	PF02987_ma; LE76_BRANA (3e-12)
EDC8_DAUCA	DAUCA	Seed	ABA, **Canon**	3(5)	2	PF02987_ma
LE76_BRANA	BRANA	Seed	ABA, **Canon**	3(5)	5	PF02987_ma
LE7_GOSHI	GOSHI	Seed	ABA, **Canon**	3(2)	2	PF02987_ma
LEA1_HORVU	HORVU	Aleurone	ABA, Cold, Desc **Canon**, Salt	3(7)	2	PF02987_ma
LEA3_MAIZE	MAIZE	Seed, Leaf Shoot	ABA, Desc, **Canon**	3(4)	2	PF02987_ma
LEA3_WHEAT	WHEAT	Shoot	ABA, Desc	3(7)	2	PF02987_hmm; LEA1_HORVU (1e-100)
LED3_DAUCA	DAUCA	Seed	**Canon**	3(4)	2	LE7_GOSHI (3e-27)
O49816	CICAR	Mesocotyl	notABA, notCold, Desc, Salt	3(5)	5	PF02987_ma; LE76_BRANA (1e-46)
O49817	CICAR	Mesocotyl	notABA, notCold Desc, Salt	3(4)	5	PF02987_ma; LE76_BRANA (5e-36)
Q03967	WHEAT	Shoot	Desc, **Canon**		2	PF02987_ma
Q06540	WHEAT	Shoot	Cold, notABA notDesc, notSalt	2S(3)	2	Q39660 (6e-8) DRPF_CRAPL (5e-5)
Q39058	ARATH	Shoot	ABA, Cold, notDesc		2	Q39873 (7e-8)
Q39660	CHLVU	Whole cells	Cold		2	PF02987_ma; LE76_BRANA (2e-5)
Q39873	SOYBN	Seed, Leaf, Root	ABA, **Canon**, Salt		2	PF02987_ma; EDC8_DAUCA (8e-37)
Q40696	ORYSA	Root	ABA, Salt	3(5)	2	PF02987_ma; LEA1_HORVU (2e-51)
Q40709	ORYSA	Shoot	notABA, Cold Mannitol	3(3)	2	PF02987_ma; LEA3_MAIZE (2e-44)
Q40869	PICGL	Embryo	ABA		2	PF02987_ma; LEA1_HORVU (9e-14)
Q40929	PSEMZ	Seed	Cold, **Canon**	3(1)	5	PF02987_ma; LE76_BRANA (2e-23)
Q41060	PEA	Seed	notABA, Sucrose, notCanon	3(1)	2	PF02987_ma; LE76_BRANA (8e-14)
Q41154	RICFL	Thalli	ABA, Desc		2	PF02987_ma; EDC8_DAUCA (1e-31)
Q41213	BRANA	Shoot, Seed	notABA, Cold notDesc, notCanon		2	PF02987_hmm; EDC8_DAUCA (2e-5)
Q42386	BRANA	Leaf	notABA, Cold		2	PF02987_hmm; EDC8_DAUCA (5e-6)
Q42512	ARATH	Shoot	ABA, Cold, Desc		2	LEA3_MAIZE (6e-5)
Q95V77	APHAV	Whole animal	Desc	3(1)	2	LEA1_HORVU (1e-13)
Q96246	ARATH	Seed, immature silique	ABA, **Canon**	3(1)	2	PF02987_ma; EDC8_DAUCA (1e-73)
Q9M4T9	WHEAT	Shoot	ABA, Cold	3(1)	2	PF02987_ma; LEA1_HORVU (1e-26)
Q9SDV6	WHEAT	Shoot	Cold, notABA notDesc, notSalt	2S(3)	2	PF02987_ma Q39873 (2e-4)
Q9XET0	SOYBN	Seed	**Canon**	3(4)	5	PF02987_ma; LE7_GOSHI (2e-28)
Q9XFD0	WHEAT	Shoot	ABA, Cold	3(4)	2	PF02987_ma; LEA1_HORVU (4e-56)

LEA Group 3 proteins, with the exemplars being LE7_GOSHI and LE76_BRANA. The columns are: 1) the protein identifier, 2) a code for the species (see Table 10), 3) the tissue(s) in which it has been found, 4) the conditions that give rise (or fail to give rise) to the expression of the gene, 5) whether the Group 3 motif or poly-lysine stutters are detected using agrep and the number of times each is found, 6) the superfamilies/stand-alone clusters in which the protein is found and 7) other evidence for accepting the protein as LEA Group 3. Note the presence in two cases of the 2S (i.e. poly-serine) motif.

However, A is highly significant in Groups 3, 4, 5 and 6, which accords with the prediction that Groups 3, 5 and 6 have higher helical secondary structure content – something also seen, for example, in the alanine-rich, alpha-helical antifreeze protein (ANPA_PSEAM) from winter flounder, PDB code 1 wfb. From Table 12 it is evident that the highly significant peptides from Group 4 have disjoint overlaps with Group 2 (+GH, -V) and Group 3 (+A, +AA). Finally, if the four major Groups 1, 2, 3 and 6 are considered, the peptides that are highly significant in all four Groups are: -C, -F, -I, -L, +Q, and -W.

Results from Clustering LEA Protein Probability Profiles

Recalling that the aim of unsupervised machine learning is to cluster the input data so that related objects are associated, while dissimilar objects are in different clusters, a POPP vector was created for each LEA protein sequence, including the three members of the Uncharacterised set. The clustering application, popp_cmp.py, was then used

to cluster the vectors. The significance threshold was set at 0.05. Bearing in mind that POPPs are not constrained to be in any particular cluster, and that the clusters can appear in any number of families and superfamilies, there is a remarkable level of agreement between the membership of the superfamilies versus the Groups derived from the literature and those observed in the supervised learning experiments discussed above.

In Tables 1 to 9, the column labelled SF lists the superfamilies in which each POPP has been placed. Because cluster, family and superfamily identifiers are created and numbered automatically, the specific numbers will bear no relation to LEA Group numbers; instead, what is significant are the sets of POPPs that appear in the same superfamily (i.e. share a superfamily identifier). Where an identifier appears in brackets, the corresponding POPP appears in a free-standing cluster, i.e. a cluster which is not sufficiently similar to any other cluster for it to have been included in a family. Table 13 lists, for each superfamily, the LEA Group it represents and the peptides making up the consensus POPP for the corresponding anchor family.

Scanning the superfamily column (labelled SF) in Tables 1,2,3,4,5,6,7,8,9 a number of observations can be made:

Table 4: LEA Protein Group 4 (D113) Exemplar(s): LE13_GOSHI

ID	Species	Tissue	Expression	Pep	SF	Evidence
LE11_HELAN	HELAN	Seed, Shoot	ABA, Desc, **Canon**		2	PF03760_hmm; PM1_SOYBN (7e-27)
LE13_GOSHI	GOSHI	Seed	ABA, **Canon**		9	PF03760_hmm; PM1_SOYBN (1e-19)
LE25_LYCES	LYCES	Leaf	ABA, Desc		2	PF03760_hmm; PM1_SOYBN (9e-18)
O24442	PHAVU	Root, Embryo	ABA, Desc, **Canon**		1	PF03760_hmm PM1_SOYBN (2e-32)
PM1_SOYBN	SOYBN	Seed	ABA, **Canon**	2Y(1)	1	PF03760_ma

LEA Group 4 proteins, with the exemplars being LE13_GOSHI. The columns are: 1) the protein identifier, 2) a code for the species (see Table 10), 3) the tissue(s) in which it has been found, 4) the conditions that give rise (or fail to give rise) to the expression of the gene, 5) whether any of the LEA Group 1, 2 or 3 motifs or poly-lysine stutters are detected using agrep and the number of times each is found, 6) the superfamilies/stand-alone clusters in which the protein is found and 7) other evidence for accepting the protein as LEA Group 4. Note the presence in one case of the 2Y motif.

Table 5: LEA Protein Group 5 (D29) Exemplar(s): LE29_GOSHI

ID	Species	Tissue	Expression	Pep	SF	Evidence
LE29_GOSHI	GOSHI	Seed	ABA, **Canon**	k(3)	2	PF02987_ma
Q93Y63	MORBO	Cortical parenchymal cells	ABA, Cold, Desc	3(2)	2	LE29_GOSHI (2e-26)

LEA Group 5 proteins, with the exemplar being LE29_GOSHI. The columns are: 1) the protein identifier, 2) a code for the species (see Table 10), 3) the tissue(s) in which it has been found, 4) the conditions that give rise (or fail to give rise) to the expression of the gene, 5) whether any of the LEA Group 1, 2 or 3 motifs are detected using agrep and the number of times it is found, 6) the superfamilies/stand-alone clusters in which the protein is found and 7) other evidence for accepting the protein as LEA Group 5. Note the presence of the Group 3 motif and poly-lysine.

Table 6: LEA Protein Group 6 (D34) Exemplar(s): LE34_GOSHI

ID	Species	Tissue	Expression	SF	Evidence
LE34_GOSHI	GOSHI	Seed	ABA, **Canon**	7	
Q41850	MAIZE	Embryo, Leaf	ABA, Desc, **Canon**	7	LE34_GOSHI (6e-61)
Q43424	DAUCA	Embryo	ABA, **Canon**	7	LE34_GOSHI (4e-75)
Q96245	ARATH	Seed	**Canon**	7	LE34_GOSHI (2e-77)

LEA Group 6 proteins, with the exemplar being LE34_GOSHI. The columns are: 1) the protein identifier, 2) a code for the species (see Table 10), 3) the tissue(s) in which it has been found, 4) the conditions that give rise (or fail to give rise) to the expression of the gene, 5) the superfamilies/stand-alone clusters in which the protein is found and 6) other evidence for accepting the protein as LEA Group 6. None of the LEA Group 1, 2 or 3 motifs match these protein sequences.

Table 7: LEA Protein Group Lea5 (D73) Exemplar(s): LE5A_GOSHI

ID	Species	Tissue	Expression	Pep	SF	Evidence
LE5A_GOSHI	GOSHI	Leaf	Desc	2S(4)	(299)	PF03242_hmm
LE5D_GOSHI	GOSHI	Leaf	Desc	2S(4)	(299)	PF03242_ma
Q39644	CITSI	Leaf, Ovule	Salt, notCold			PF03242_hmm, LE5D_GHOSHI (2.4e-46)

Lea5/D73 proteins – currently not part of any numbering scheme for LEA proteins – with the exemplar being LE5A_GOSHI. The columns are: 1) the protein identifier, 2) a code for the species (see Table 10), 3) the tissue(s) in which it has been found, 4) the conditions that give rise (or fail to give rise) to the expression of the gene, 5) whether any of the LEA Group 1, 2 or 3 motifs or poly-lysine stutters are detected using agrep and the number of times each is found, 6) the superfamilies/stand-alone clusters in which the protein is found and 7) other evidence for accepting the protein as LEA Group Lea5/D73. Note the presence in two cases of the 2S motif (poly-serine stutter). Note also that two of the three proteins are found in a single, stand-alone cluster containing just the pair of proteins, while the other sequence is not found in any cluster.

Table 8: LEA Protein Group Lea14 (D95) Exemplar(s): LE14_GOSHI

ID	Species	Tissue	Expression	SF	Evidence
DRPD_CRAPL	CRAPL	Leaf	ABA, Desc		PF03168_ma; LE14_SOYBN (2e-52)
LE14_GOSHI	GOSHI	Leaf	Desc	(297)	PF03168_ma; LE14_SOYBN (2e-64)
LE14_SOYBN	SOYBN	Leaf	ABA, **Canon**	(297)	PF03168_ma
Q40159	LYCES	Root	notABA, notDesc, possible osmotic stress	3	PF03168_ma, LE14_SOYBN (8e-53)

Lea14/D95 proteins with the exemplar being LE5A_GOSHI. The columns are: 1) the protein identifier, 2) a code for the species (see Table 10), 3) the tissue(s) in which it has been found, 4) the conditions that give rise (or fail to give rise) to the expression of the gene, 5) the superfamilies/stand-alone clusters in which the protein is found and 6) other evidence for accepting the protein as LEA Group 6. None of the LEA Group 1, 2 or 3 motifs match these protein sequences. Note that two of the three proteins are found in a single, stand-alone cluster containing just the pair of proteins, one protein is found clustered with Group 2 LEA proteins in SF 3, while the other sequence is not found in any cluster.

Table 9: Uncharacterised LEA Proteins

ID	Species	Tissue	Expression	SF
O24439	PHAVU	Root, Stem, Embryo	ABA, Desc, **Canon**	(295)
O81483	ARATH	Seed	notABA, notDesc, notSalt, **Canon**	(279)
Q9S7S3	ARATH	Seed	notABA, Cold, notDesc, notSalt, **Canon**	(279)

Currently uncharacterised proteins which have expression patterns that are literally late embryogenesis, but which have no similarity to any of the previously described proteins. The columns are: 1) the protein identifier, 2) a code for the species (see Table 10), 3) the tissue(s) in which it has been found, 4) the conditions that give rise (or fail to give rise) to the expression of the gene, and 5) the stand-alone clusters in which the protein is found. Note that one pair only cluster with each other, while the third is found in a stand-alone cluster together with a Group 2 LEA protein.

• The Group 4 LEA proteins are split between superfamilies covering Group 2 LEA proteins (PM1_SOYBN, LE13_GOSHI, O24442) and superfamilies comprising Group 3 LEA proteins (LE11_HELAN, LE25_LYCES).

• The two Group 5 proteins, LE29_GOSHI and Q93Y63 are clustered among the Group 3 LEA proteins (Superfamily 2).

• The Group 1 LEA proteins are split across two superfamilies, with clusters involving EM1_ARATH, EMB1_DAUCA, L193_HORVU, LE19_GOSHI, L194_HORVU and LE10_HELAN appear in Superfamily 4 while clusters involving EM1_WHEAT, EM2_WHEAT, EMB5_MAIZE, EMP1_ORYSA, L19A_HORVU, L19B_HORVU, EM6_ARATH. SEEP_RAPSA are found in a different superfamily, Superfamily 6.

• The Group 2 LEA proteins are split across five superfamilies. Looking at the consensus POPPs of the corresponding anchor families one notices that all the superfamilies have peptides from the 2K motif, while Superfamily 8 and Superfamily 10 have peptides from 2S. None of the anchor families have peptides from the 2Y motif, but they are present in other Families in Superfamily 1 (data not shown).

• Two of the Uncharacterised canonical LEA proteins, O81483 and Q9S7S3 only cluster with each other, while the third in this set, O24439, clusters with a Group 2 LEA protein, Q9SBI7. This situation persists even when the clustering thresholds are lowered to the point where significant numbers of Group 3 LEA proteins were found clustered with Group 2 LEA proteins. Furthermore, it is worth noting that the clustering of O24439 with Q9SBI7 is free-standing, i.e. not in a superfamily, which suggests that the relationship (supported by the supervised machine-learning rules) is a distant one.

Table 10: Mapping from SwissProt Species Codes toSpecies Names Used in LEA Protein Group Tables

Code	Species
APHAV	*Aphelenchus avenae*
ARATH	*Arabidopsis thaliana*
BRANA	*Brassica napus*
CHLVU	*Chlorella vulgaris*
CICAR	*Cicer arietinum*
CITSI	*Citrus sinensis*
CITUN	*Citrus unshiu*
CRAPL	*Craterostigma plantagineum*
DAUCA	*Daucus carota*
DAUCA	*Daucus carota*
GOSHI	*Gossypium hirsutum*
HELAN	*Helianthus annuus*
HORVU	*Hordeum vulgare*
LOPEL	*Lophopyrum elongatum*
LYCES	*Lycopersicon esculentum*
MAIZE	*Zea mays*
MEDFA	*Medicago falcata*
MORBO	*Morus bombycis*
ORYSA	*Oryza sativa*
PEA	*Pisum sativum*
PHAVU	*Phaseolus vulgaris*
PICGL	*Picea glauca*
PONTR	*Poncirus trifoliata*
PRUPE	*Prunus persica*
PSEMZ	*Pseudotsuga menziesii*
RAPSA	*Raphanus sativus*
RICFL	*Riccia fluitans*
SOLCO	*Solanum commersonii*
SOLTU	*Solanum tuberosum*
SOYBN	*Glycine max*
SPIOL	*Spinacia oleracea*
STELP	*Stellaria longipes*
TRITU	*Triticum turgidum subsp. durum*
VACCO	*Vaccinium corymbosum*
VIGUN	*Vigna unguiculata*
WHEAT	*Triticum aestivum*

The codes are those used in forming SwissProt protein identifiers. With a small number of exceptions for the most common species such as PEA, WHEAT and MAIZE, identifiers are generally made up of the first three letters of the genus name followed by the first two letters of the species name.

Table 11: LEA classification rule set induced by supervised learning

Group	Rule
2a	H < = 0.15 and aromatic > = 0.077 and min_hyph < = -1.97 and charged < = 0.42
2b	L > = 0.23 and H < = 0.3 and ave_hyph > = -1.233 and ave_hyph < = -0.978
2c	aromatic > = 0.077 and min_hyph < = -2.743 and charged > = 0.4
3	H > = 0.34
1	E > = 0.02 and ave_hyph < = -1.241
LE5	max_hyph > = 1.0 and ave_hyph < = -0.3
LE14	aliphatic > = 0.25
6	H > = 0.25 and max_hyph > = 0.5
4	Otherwise

The rules are to be applied in a top-down, if .. else if, manner, so, for example, if the percentage of predicted helical conformation (expressed as a number in the range 0 .. 1.0) is greater than or equal to 0.34 then the protein is classified as a Group 3 LEA protein, but only if each of the rules above has failed, e.g. because the percentage of aromatic residues is less than 0.077 in the Group 2 rules. min_hyph and max_hyph are, respectively, the values of minimum and maximum hydrophobicity windows, while ave_hyph is the average across all the hydrophobicity windows. H, E and L refer, respectively, to the percentage composition of amino acids that are found by ProteinPredict (in four-state mode) to be alpha-helical, beta-sheet or loop.

Results from Keyword Clustering of POPP Search Hits

Table 14 summarises some of the keywords and phrases associated with each superfamily (thence Group) through the application of the Protein Annotators' Assistant to the sets of hits returned by popp_search.py when given as queries the consensus POPPs for each anchor family. Lea5 and Lea14 are presented by the consensus POPP for the single cluster respectively representing the two Groups. For compactness, only the most significant, distinct keywords are listed.

When scanning Table 14 it is worth bearing in mind that rather than being understood as the actual functions which the search hits share with the LEA proteins, matches based on shared biases in peptide composition can indicate shared mechanisms or structural elements. In this, POPP searching is similar in spirit to testing a sequence against the motifs in the PROSITE database [23] or against the fingerprints in the PRINTS database [24]. The difference, in principle, is that motifs and fingerprints can be seen as a conjunction of gapped or ungapped patterns and are relatively long, while POPPs are a disjunction of short patterns which are are distinguished by being significantly over- or under-represented.

Table 12: Highly significantly over- and under-represented peptides across LEA Protein Groups

Grp	Threshold Pr	Representation	Sample of Significant Peptides (negative p-values indicate under-representation)
1	3.9e-07	over	G (2.6e-49), E (7.9e-36), Q (2.6e-10), R (4.2e-08), GG (5.8e-47), KGG (1.6e-41), EMG (9.5e-33), QMG (2.1e-19)
		under	I (-3.7e-16), V (-1.1e-14), P (-1.7e-14), F (-7.4e-14), N (-1.3e-12), L (-4.2e-11), C (-6.6e-11), W (-1.7e-09)
2	1.1e-10	over	G (0), TG (0), H (7.4e-291), GG (6.4e-178), T (6.8e-122), K (1.4e-59), Q (8.7e-28), HG (5.6e-187), KLP (2.8e-170), EK (1.1e-120), YG (4.7e-101), SSS (2.0e-43)
		under	L (-5.6e-123), F (-2.2e-81), I (-1.1e-52), V (-9.8e-47), N (-3.7e-43), R (-3.8e-39), C (-1.0e-36), W (-1.7e-35), S (-5.2e-25),
3	4.3e-08	over	A (3.6e-246), K (7.3e-140), T (3.2e-48), E (2.9e-37), Q (7.8e-32), KD (1.6e-93), AKD (1.6e-83), AKE (2.4e-45), KDY (2.1e-46), EK (1.8e-45)
		under	L (-1.2e-89), I (-3.9e-61), P (-2.8e-51), F (-6.1e-35), W (-8.3e-27), C (-4.2e-19), N (-1.5e-14), R (-2.3e-12)
4	4.1e-04	over	TG (1.9e-17), G (4.7e-14), T (9.5e-12), GH (8.7e-09), A (7.8e-08), AKA (1.9e-07), EK (8.6e-07), AA (3.8e-05)
		under	L (-5.1e-19), I (-4.8e-09), F (-1.4e-08), V (-5.6e-06), C (-4.6e-05), W (-1.8e-04), S (-8.5e-04)
5	6.2e-04	over	AKE (1.2e-17), K (9.2e-13), A (4.1e-10), E (1.0e-08), EK (2.7e-05)
		under	L (-5.2e-11), P (-4.3e-08), I (-1.5e-06)
6	8.4e-04	over	A (6.5e-30), AA (3.1e-18), AT (1.7e-08), AE (1.8e-07), QS (2.2e-06), GV (3.7e-06), GG (8.6e-06), Q (4.1e-05), V (1.6e-04), QSA (2e-13)
		under	L (-2.2e-10), F (-8.9e-09), C (-7.9e-07), Y (-6.0e-06), I (-3.5e-05), K (-3.3e-04), W (-3.0e-04)
Lea5	1.2e-03	over	A (4.4e-05), GA (4.1e-05), GY (8.8e-05), SS (1.3e-04), R (2.6e-04), S (6.9e-04)
		under	Q (-4.9e-04)
Lea14	1.3e-03	over	IP (1.1e-07), D (7.7e-05), K (3.3e-04), I (1.2e-03)
		under	R (-4.1e-06), Q (-8.0e-06), F (-3.3e-03)

Applying popp_create.py to each group of LEA protein sequence taken as a whole, the table lists a sample of the peptides that are highly over-represented or highly under-represented, i.e. their probabilities are more stringent than the thresholds listed in the second column. The different thresholds arise due to differences in the numbers of sequences, hence differing amino acid counts, corresponding to each Group.

Table 13: Consensus POPPs for the anchor families of each superfamily

Group	SF	Anchor Family Consensus POPP
1	4	+E, +G, +EG, +GE, +GG, +KG, +QE, +RK, +GGE, +KGG
1	6	+E, +G, +DE, +EG, +ES, +GG, +GQ, +RE, +RK, +ARE, +DES, +REG
2	1	+G, -L, +EK, +GG, +GT, +EKL, +IKE, +KEK, +KIK, +KKG, +KLP, +LPG
2	3	-F, -I, -L, -R, -W, +DK, +EK, +KK, +KL, +LP, +TH, +EKK, +KEK, +KLP, +LPG
2	8	+EK, +SS, +EKI, +KEK, +KIK, +SSS
2	9	-F, +G, -I, -L, -V, +AG, +EK, +GG, +GH, +GT, +TA, +TG, +GGT, +GTG, +TAG, +TGG
2	10	-F, +G, -I, +AG, +EK, +GG, +GQ, +KE, +SS, +EKL, +GAG, +IKE, +KEK, +KLP, +LPG, +SSS
3	2	+A, -C, +E, -F, -I, +K, -L, -P, +AE, +AK, +EK, +ET, +GE, +GK, +KE, +AAE, +AKD, +EKA
3	5	+A, -I, +K, -L, -P, +Q, +T, -V, +AA, +AQ, +EK, +KE, +KT, +QA, +QQ, +QS, +QT, +TQ, +AAK, +AQA, +EKT, +QAA, +TQQ
6	7	+A, -F, -L, +AA, +AE, +MQ, +QS, +VA, +AAA, +GVA, +QSA, +SAA
Lea5	299	+A, +R, +S, +AM, +GA, +GY, +RP, +SF, +SS, +YS
Lea14	297	+D, -R, +AS, +IP, +KV, +VS, +TIP

Clusters, families and superfamilies closely mirror the structure of the LEA Groups, with the exception of Group 4 and Group 5. Against each LEA Group are listed the superfamilies that contain proteins from that Group (column 2) and the peptides forming the consensus POPP of the anchor (i.e. most typical) family in the superfamily. '+' before a peptide indicates significant over-representation; '-' indicates significant under-representation.

Discussion

As mentioned in the Introduction, one source of confusion in the coverage to date of LEA proteins has been the overlapping and sometimes contradictory assignments to Groups. For example, if [12] is taken as a starting point, [3] differs from the former by coalescing the proteins corresponding to LEA protein Group 6 and Lea14 into a single Group (which in that paper is called Group 5); Lea5 is not found in any Group in this scheme. On the other hand, in [4], the Group 4 of [12] has been renamed Group 5, while the Group labelled Lea14 in this study is called Group 4. There is agreement, however, on the first three

Table 14: Keywords/Phrases for each Group and Superfamily

Group	SF	Principal Keywords/Phrases
1	4	histone H4, chromosomal protein, nuclear protein, DNA binding
1	6	dsRNA binding, DNA gyrase, breakage, CLP, ATP binding
2	1	break, ATP binding, DNA topoisomerase, protein biosynthesis, topoisomerase, repair
2	3	coiled, coil, nuclear protein, caldesmon, histone H1, chaperone, tropomyosin filament, break, DNA topoisomerase
2	8	DNA topoisomerase, nuclear protein, HMG box, coiled coil
2	9	transcriptional inhibition, glycosyl hydrolase, nuclear protein,
2	10	nuclear protein, DNA binding, transcription regulation, intermediate filament, keratin, chaperone, homeobox, coiled coil, HMG box domain, cytoskeletal
3	2	chaperone, coiled coil, tropomyosin, stress, filament, phosphorylation, caldesmon elongation factor, neurofilament, actin binding, cytoskeleton, rotamase
3	5	coiled coil, histone H1, filament, nuclear protein, neurofilament, flagella, HAMP domain, synuclein, DNA binding, hsp70
6	7	groel protein, nuclear protein, histone H1, chaperonin, DNA binding, HAMP domain, synuclein, transcription regulation
Group	Cluster	Principal Keywords/Phrases
Lea5	299	DNA binding, transcription regulation, nuclear protein, gata, zinc finger, homeobox
Lea14	297	esterase, gapdh, chaperone protein DNA, glycoprotein

For each superfamily, the consensus POPP (Table 13) has been set as a query against a database of POPP vectors representing SwissProt. The protein hits, excluding LEA proteins for each query were submitted to the Protein Annotators' Assistant, which returns a list of keywords and phrases shared by sets of the submitted proteins. A sample of the most prominent are listed against the Group/Class and the corresponding superfamilies/clusters. Rather than being understood as the actual functions which the search hits share with the LEA proteins, matches based on shared biases in peptide composition can indicate shared mechanisms or structural elements.

Groups. Given the new findings on this sizable sample taken from the spectrum of LEA proteins, it is now possible to revisit the different LEA Groups.

Group 1 LEA proteins are strongly hydrophilic and each cluster has the peptides E and RK over-represented (not found in any other Group). The phrase DNA binding appears in various guises connected with this group (Table 14). As can be seen from the respective entries in Table 13, consensus POPPs for the two superfamilies representing Group 1 LEA proteins are in fact very similar. In addition, from the input data used for the supervised machine learning experiments (not shown) it is noted that the members of Superfamily 4 generally have a higher percentage of charged amino acids than Superfamily 6 (and some of the highest percentages overall). The LEA proteins covered by Superfamily 4 also include those with repeats of the Group 1 motif.

Analysing the Group 2 LEA proteins exposes a difficulty with the methodology of retrospective reanalysis; the data that would be required to settle questions of group membership are often not available from the original publications. However, Group 2 appears to split into three subgroups, labelled 2a, 2b and 2c, with the line of demarcation being between those Group 2 LEA proteins which are cold-tolerant versus those which are sensitive to cold stress. The split is evident in the three rules proposed by the classification engine Ripper. Subgroup 2a has low predicted helix content and medium to high percentage of aromatic residues while Subgroup 2b has high predicted loop content. All three subgroups are hydrophilic, but the third and smallest subgroup, Subgroup 2c, is very hydrophilic. The eight proteins which were found not to be up-regulated by cold stress are in 2a, while all members of 2c are up-regulated by cold stress. The proteins in Subgroup 2a, and in particular the proteins not up-regulated by cold stress, are covered by the Superfamily 1 and Superfamily 9. All the members of Subgroup 2c have poly-lysine stutters (versus 5 in the muchlarger Subgroup 2b and 4 in Subgroup 2a), and most of those with those with poly-lysine stutters are found to be cold tolerant; for the remainder, data on cold tolerance has not been presented. In general, tolerance of cold is found associated with Superfamily 3.

The entire Group is characterised by an over-representation of either H or SSS (often both); O24442 and PM1_SOYBN (from Group 4, though arguably Group 2) and EM1_ARATH, also have an over-representation of H, while Q06540 and Q9SDV6 from Group 3 and LE5A_GOSHI and LE5D_GOSHI from the Lea5 have poly-serine stutters of at least 3aa. The poly-serine stutters are all the more remarkable when one notes that serine by itself is highly under-represented. PM1_SOYBN also matches the 2Y motif corresponding to the Close (1997) Y-segment, which accounts in part for its presence in Superfamily 3. Fourteen of the 22 Subgroup 2a proteins have an over-representation of GNP or YGN, corresponding to the Y-segment of [13], which suggests that Subgroup 2a is distinct from the other Subgroups. On the other hand, K is over-represented in all six the Subgroup2c

LEA proteins and 9 out of 16 Subgroup 2b LEA proteins. (It is also over-expressed in 20 of 23 Group 3 LEA proteins and the subset of Group 1 LEA proteins discussed above.) The suggestion, therefore, is that while most Subgroup 2a LEA proteins have a Close (1997) K-segment, it is is less significant than those of Subgroups 2b and 2c LEA proteins (cf. DH11_GOSHI and DH14_LYCES versus DH47_ARATH, which suggests a role in cold stress resistance. It is therefore likely that many of the Subgroup 2b proteins which are found in Superfamily 1 but which are not specifically cold induced, such as DH1_HORVU and DH21_HORVU, might in fact also be induced by cold stress. The association of the K-segment with cold tolerance has been noted by other researchers [13]. Finally, as mentioned above the non-ABA dependent protein Q40159, characterised by sequence similarity and the classification rules as Lea14, ends up clustered with Group 2 LEA proteins associated with resistance to cold stress, in particular DH14_ARATH, but also DH10_ARATH, DH47_ARATH and O04232, so the role of this protein, which is neither induced by ABA nor desiccation stress, might in fact be related to cold-stress resistance.

The picture with the Group 3 LEA proteins is rather more straight forward, with a crisp rule encompassing all members of this group, namely that they have high helix content. The similarity of members of this Group is also borne out by the fact that they are all clustered in a single superfamily. It should be noted, however, that Group 4 LEA proteins LE11_HELAN and LE25_LYCES are clustered in families within Superfamily 3, which mirrors what was observed with rules induced by Ripper.

Looking at the major LEA Groups it is interesting to note that when the threshold scores for adding to a cluster and for merging two clusters are reduced (see [25] for more details), the Group 1 and Group 6 LEA proteins remain distinct with unique superfamilies being created for each, but clusters representing Group 2 and Group 3 LEA proteins merge into a single superfamily. This is a little less surprising when one notes the number of Group 3 proteins that are also up-regulated by cold stress. In addition, if the number of mismatches allowed for the Group 3 motif TAQAAKEKAXE is increased by 1 to 5, the set of matching LEA proteins includes the Group 2 LEA proteins DH1_HORVU, DH1D_ORYSA and DH2_HORVU, Group 4 LEA proteins LE13_GOSHI and LE25_LYCES and Group 5 LEA protein LE29_GOSHI. In addition, the Group 3 LEA protein DRPF_CRAPL has a poly-lysine stutter, while the Group 3 LEA proteins Q06540 and Q9SDV6 have poly-serine stutters (i.e. the 2S motif). Taken together, it would appear that Group 2 and Group 3 LEA proteins might be related. K is over-represented across both Groups, while L and, generally I, are both under-represented, suggesting a connection with charged amino acids. Theconnection with Group 2 LEA proteins is, perhaps, less surprising if one considers the association of the (Group 2) K-segment with cold tolerance (noted above), the fact that many Group 3 LEA proteins are associated with cold tolerance (see Table 3), and that the K-segment consensus for gymnosperms differs in up to six places from the canonical K-segment (noted in [13]), versus the five that were allowed in the motif-search described above.

Returning to Table 14, there is considerable overlap across the sets of keywords, particularly across Group 2 and Group 3 LEA proteins. A remarkable, and seemingly paradoxical, recent result has been the demonstration that a nematode Group 3 LEA protein, AavLEA1 (Q95V77), is unstructured in the native state, but then becomes structured on desiccation, showing significant alpha-helical content and possible coiled-coil structures [26]. In other words, the consistent prediction of high alpha-helical content for Group 3 LEA proteins appears to be borne out, but only in response to desiccation stress. Coiled coil is one of the phrases evident from the keyword analysis of Group 3 LEA proteins; it is also characteristic of Group 2. The keyword filament and related keywords such as keratin and neurofilament are also prominent in the list, mirroring a suggestion in [26] that the coiled coils might form larger structures related to intermediate filaments, which would provide mechanical support to plant cells undergoing desiccation stress. The conundrum of some keywords being associated with the cytoskeleton while others are nuclear has already been noted via localisation experiments reported in [13], at least for the Group 2 LEA proteins. Table 14 would suggest that the observation is generally true. A number of other themes are also apparent in the list of keywords and phrases: DNA binding, stress and chaperone activity. While dealing with stress, particularly cold stress, has long been associated with LEA proteins, mechanisms suggested by the keywords "DNA binding" and "chaperone" require experimental verification.

Turning to the Group 4 LEA proteins, as noted in discussion of the supervised classification experiments, two of the five Group 4 LEA proteins, LE11_HELAN and LE25_LYCES, are subsumed into Group 3. In the unsupervised clustering, those same proteins are also subsumed in Group 3, while the remainder, PM1_SOYBN, LE13_GOSHI and O24442, appear in Group 2. Even when the probability threshold is made more stringent – 0.005 – the five putative Group 4 LEA proteins do not cluster separately. In addition, as was noted above, PM1_SOYBN has a hit against the 2Y motif, while LE11_HELAN and LE25_LYCES each have hits against the Group 3 motif once the number of allowed mismatches is increased by 1 (a level which still leaves out some acknowledged Group 3 LEA proteins). In other words,

Table 15: Comparison of New LEA Protein Classes with Previous Group Classifications

Class	Baker/Dure	Bray 1994	Bray 2000	Cuming	Comments
I	D19	1	1	1	
IIa	D11	2	2	2	Includes some Group 4 (D113); Subgroup 2a from Rules
IIb	D11	2	2	2	Subgroups 2b and 2c from the Rules
III	D7	3	3	3	Includes Group 5 (D29) and remainder of Group 4 (D113)
IV	D34	6	-	5	
V	-	-	-	-	Lea5/D73 – Named in [15]
VI	-	-	4	5	Lea14/D95 – Named in [15]

A proposed LEA Class numbering scheme (column 1), encompassing all the Groups listed above with the exception of Group 4 and Group 5, is compared with past numbering schemes from: Baker/Dure (column 2), Bray 1994 (column 3), Bray 2000 (column 4) and Cuming (column 5).

there is mounting evidence that Group 4 should not be considered as a separate Group, but that its members be absorbed into Group 2 and Group 3. This stands in apparent contrast to the evidence from sequence alignments which suggests that the five members of this group should remain together. However, the weight to be given to this evidence must be tempered by the knowledge that each of these is a low complexity protein and numbers of the amino acids will need to be masked: PM1_SOYBN (15.6% masked), LE25_LYCES (18.2%), O24442 (28.6%), LE13_GOSHI (47.3%) and LE11_HELAN (53.8%). The effect of this is that when LE13_GOSHI is run as a BLAST query with SEG masking in place, the only hits returned (at p-value of 0.79) are Group 2 LEA sequences, Q39876 and Q39805. While on balance the Group 4 proteins are best reassigned to Group 2 and Group 3 it is also arguable on the basis of motif hits and the weak alignment evidence that the Group 4 LEA proteins form a link between the Group 2 and Group 3 LEA proteins, particularly PM1_SOYBN, which matches the Group 3 motif twice at the N terminal and the 2Y motif from Group 2 at the C terminal; LE13_GOSHI and LE25_LYCES have their Group 3 motif matches also at the N terminal.

A similar line of reasoning – in this case supported by other investigators authors [3,4] applies tothe former Group 5 (D29) LEA proteins, which were folded into the Group 3 LEA proteins by both the supervised and unsupervised algorithms.

By contrast, it is proposed in [3] that proteins corresponding to LEA protein Group 6 and Lea14 form a single Group (which in that paper is called Group 5), while Lea5 LEA proteins are not mentioned. In this study, all three groups appear at the top of the list of average hydrophobicity scores (either just over 0 or just below it, with Lea14 > Group 6 > Lea5). They also gather at the bottom of the list for percentage polar residues. On the other hand, Group 6 proteins are just behind Group 3 in predicted helix content, with Lea5 and Lea14 some way below, while in the Lea5 Group, long loop segments are evident. Group 6 have an over-representation of both MQ and AAA, while the three Lea5 LEA proteins have an over-representation of A and R. By contrast, the Lea14 LEA proteins have an over-representation of IP and an under-representation of R. The three groups are sufficiently different for crisp classification rules to have been created, although the rules must be treated with caution due to the small numbers of examples on which they are based. In addition, the clusters involving Group 6 LEA proteins persist even when cluster-merging thresholds are lowered or significance thresholds made less stringent. At the same time the Lea5 and Lea14 proteins form independent clusters neither of which merge with Group 6.

Conclusions

The study of a carefully selected set of 112 LEA protein sequences has revealed a number of aspects of these proteins, which can be summarised in the following conclusions:

• There is a high level of agreement between the different machine learning methods on the one hand, and the previous assignments on the other. However, given the previous contradictory revisions and current findings a new scheme for naming groups of LEA proteins is proposed, based on Classes. In particular, while it is generally accepted that the former LEA Group 5 is not distinct from Class III, the balance of evidence is that the members of former Group 4 are more appropriately housed in Class II and Class III.

• There is evidence from overlapping motifs, overlapping POPP clusters, from the split of former LEA Group 4 and from similarities in the modes of induction related to cold stress that Class II and Class III LEA proteins, though distinct might be related, perhaps through the LEA Class II K-segment motif, which mirrors the Class III motif. The major difference between Class II and Class III is that the

former contains different combinations of three motifs/domains, while Class III has often multiple instances of the one motif/domain.

• In the same way that not all sequence alignment hits are necessarily relevant, it is possible that not all the keywords will turn out to be relevant. However, there is confirmation in the keywords concerning subcellular localisation which sees LEA proteins being associated with the cytoskeleton, the cytoplasm and with the nucleus (though these are unlikely to apply to the same protein). However, each possibility has been noted for dehydrins [13].

• Keywords related to chaperones and to DNA-binding are also present, suggesting a role similar to the DNA-binding cold-shock proteins found in bacteria, but also in eukaryotes, e.g. DBPA_HUMAN (P16989). DBPA_HUMAN is found both in the nucleus and in cytosol. However, such suggestions await experimental verification.

• Keywords emphasising alpha-helical structure (coiled coil) and, at a larger scale, filaments also support the recent finding that Class III LEA proteins show high alpha helical content, and possibly coiled-coil structures, except that this occurs under conditions of desiccation stress; the protein has no defined structure in its native state [26]. High alpha helical content is also consistent with the over-representation of alanine, particularly in Class III and Class IV (former Group 6) LEA proteins.

• Apart from the near total lack of cysteine and tryptophan, the study has found that isoleucine, leucine and phenylalanine are highly under-represented across the four major Classes, while glutamine is highly over-represented. Glutamate and lysine are highly over-represented in two of the first three LEA Classes, and moderately in the third, so the description of these as hydrophilins [16] is borne out.

• Glycine is highly over-represented in Class I and overwhelmingly so in Class II, but only in line with chance in Class III, which is consistent with the first two Classes having the highest predicted loop content, particularly Class II LEA proteins. The high proportion of predicted loop content is supported by the observation that at least one dehydrin has no defined structure in its nature state [27]. However, as with the Class III LEA proteins, Class II LEA proteins acquire alpha-helical content under stress conditions, e.g. application of sodium dodecyl sulfate (SDS) [28].

In general, non-globular and, particularly, low-complexity proteins such as the LEA proteins pose special challenges in determining their functions and modes of action. Therefore, rather than relying solely on evidence from sequence alignments, a combination of data sources can be used, particularly software tools less affected by such unusual proteins. Further work involves expanding the analysis to examine the large number of putative LEA proteins found in genomic sequences, particularly from non-plant species.

Methods

Defining a LEA Protein for this Study

There are two parts to a working definition of what constitutes a LEA protein. The first is that a LEA protein is a plant protein which has no – or at most limited – expression in the stages up to and including maturation of the ovule, and sharply rising expression post-abscission, peaking at desiccation, with expression disappearing at germination [7]. In other words, LEA proteins are characterised in the first instance by raised levels of expression in mature seeds, with expression disappearing at germination. However, proteins homologous to LEA proteins have also been found in other plant tissues, so although they are not involved in embryogenesis, let alone late in embryogenesis, they too are now considered to be LEA proteins. The latter set are characterised by sharply raised expression due to desiccation, raised salinity, cold or induction by abscisic acid (ABA), followed by a sharp decline in expression once the stress condition has been removed [29]. As a result, where the distinction is useful, the former set of LEA proteins will be termed "canonical LEA" proteins in this study.

Unfortunately, sharply raised expression under the conditions such as desiccation or cold stress is not sufficient to unambiguously characterise a protein as an LEA protein because plants use a number of metabolic pathways to respond to such abiotic stresses and there are a number of other protein families which are induced under similar conditions. For example, the *Arabidopsis thaliana* gene *RD22* (RD22_ARATH) is expressed in the early and middle stages of seed maturation, but is also induced by desiccation, salinity or application of ABA [30]. Similarly, the gene *PCC13-62* (DRPE_CRAPL) is up-regulated in the leaves of the resurrection plant, *Craterostigma plantagineum*, by desiccation or the application of ABA [31]. Neither of these have any sequence similarity to LEA proteins.

On the other hand, sequence similarity to canonical LEA proteins, by itself, is also not sufficient to accurately classify all putative non-canonical LEA proteins because there are several proteins with significant similarity to canonical LEA proteins which are not expressed under conditions typical of LEA proteins. Examples are: Q06431 (BP8 protein) – which is among the "seed" proteins underpinning Pfam family PF02987 – and Q43430 (Dehydrin cognate), which is found among the proteins recovered by the Hidden Markov Model for Pfam family PF00257. In the case

of Q39846 (labelled as: LEA Protein) there is some evidence of similarity to Group 3 LEA proteins via BLAST hits to Q41060 and EDC8_DAUCA, but the level and timing of expression is such that [32] concludes: "Since the GmPM4 proteins do not appear to fulfil the biochemical properties of LEA proteins, their messages are not very abundant in mature seeds and will not express in water-stressed seedlings, we suggest that the physiological roles of GmPM4 protein might differ from those of the LEA proteins, i.e. desiccation protection." (pg 489). However, the most striking case of this problem is the putative LEA protein DHX1_ARATH, which has been classified as a D11 (Group 2) LEA protein in the Dure survey [11], but which is only expressed constitutively, i.e. not as a stress response nor late in embryogenesis. In a related manner, the protein O48672, superficially a Group 2 LEA protein, is largely constitutively expressed although there is some increased expression due to cold stress.

The problem of interpreting purely sequence-based data becomes more acute for the putative LEA proteins found in non-plant species, e.g. the LEA Group 3 motif found on avian developmental gene *px19* [33]. As a second example, while no claim is made that gene *gvpQ* of *Bacillus megaterium* is a LEA protein – it is thought to be a negative regulator of gas vesicle synthesis – the corresponding sequence, O68678, is annotated as a Group 3 LEA protein by Pfam and is one of the sequences used in the multiple alignment that defines the Pfam family PF02987. In other words, significant sequence similarity to known (and in particular canonical) LEA proteins might indicate homology, but once the functions have diverged doubts can arise – proteins with different functions, arising perhaps due to paralogy, face different conservation pressures. Automated classification studies, also known as machine learning, require a strict notion of which objects are members of the categories under study (the "universe of discourse") and which are not. Therefore, a conservative strategy in building a database of sequences for categorisation experiments is to only accept proteins that have related functions or, as a surrogate, related mRNA expression patterns when the functions are not known. In the latter case there is the assumption that proteins which have expression patterns unrelated to LEA proteins will turn out to have different functions.

In summary, to ensure that only true members of the set of LEA proteins are used in this study, a LEA protein is either a canonical LEA protein or one whose expression is sharply up-regulated by desiccation, salinity, cold or exogenous ABA and which has sequence similarity to canonical LEA proteins.

Obtaining the Sequences

The sequences were drawn in the first instance from the SwissProt and SpTrEMBL databases (containing between them around 700,000 proteins) using the SRS sequence retrieval system [34]. Because different authors have, over time, used different words to describe LEA proteins a number of keywords were used to extract the sequences from the databases, including: "LEA", "small hydrophilic plant seed", "late embryogenesis abundant", "dehydrin" and "seed maturation". A second source of LEA protein sequences were those revealed by BLAST similarity searches using other LEA proteins as search queries. However, irrespective of the path by which a putative sequence was uncovered, as discussed above there also needed to be evidence of expression of the protein under conditions associated with LEA proteins, as revealed in the cited literature. In other words, the literature corresponding to the sequence had to be examined for evidence, typically via Northern blots, of expression patterns conforming to the definition outlined above; in order to have confidence in the provenance of the hits, putative LEA proteins unsupported by expression evidence were passed over.

Assignment to Historical Groups

The LEA proteins were initially assigned to a Group based on a number of criteria. The first is an assessment by the authors and/or inclusion in the 1993 survey by Dure. A second is whether the protein is covered by one of the Pfam families listed above. Finally, BLAST was used to determine if there are any close hits against one or other canonical LEA protein or, in default, to known members of a Group. (Given the problems outlined earlier, low complexity sequence masking was not used for this.)

Members of each LEA protein Group are listed in Tables numbered from 1 to 9. The first two columns in the tables are the protein's SwissProt/SpTrEMBL identifier and the species from which the protein was taken, represented by a SwissProt species code. (A mapping from the SwissProt codes to the species names can be found in Table 10.) This is followed by the tissues used for the expression evaluation and a list of the conditions that give rise (or fail to give rise) to the expression of the gene. The possible conditions are: ABA (application of abscisic acid to aerial parts of the plants), Cold, Desc (desiccation) and Salt. As mentioned above, the descriptor canonical is used to indicate that high levels of the mRNA are to be found in dry seeds, i.e. the protein is literally late embryogenesis (abbreviated Canon). The appearance of 'not' before any of these descriptors indicates that expression has been tested for this condition and no significant expression was seen. For example, notDesc indicates that there was no significant increase in the expression of the corresponding gene under conditions of desiccation stress.

For LEA protein Groups 1, 2 and 3, consensus sequence motifs have been reported [3]: GGQTRREQLGEEGYSQMGRK (Group 1), DEYGNP and EKKGIMDKIKEKLPG (Group 2, patterns 2Y and 2K, in the nomenclature of [13]) and TAQAAKEKAXE (Group 3). Being consensus sequences, matching against any particular protein sequence implies accepting a certain number of insertions, deletions or substitutions. Using an implementation of the string searching application, Agrep [35], each consensus peptide was tested against the LEA protein sequences, allowing up to 5, 2, 4 and 4 mismatches, respectively, for the four consensus patterns. In addition, Group 2 LEA proteins generally have a poly-serine stutter. If a consensus peptide matches without exceeding the stated maximum number of amino acid mismatches, or a poly-serine stutter is found (which is labelled 2S after [13]), it is noted in the fifth column, with the number of repetitions noted in brackets (or the length of the poly-serine stutter, which must be at least 4aa). While the 2S segment is highly characteristic of Group 2 LEA proteins (occurring in 36 of the 50 sequences in the set used in this study, versus an expected count of 1.98 sequences – corresponding to a probability of $1.\ 7 \times 10^{-39}$) it was noticed that poly-lysine stutters with a length of at least 3aa are also relatively common, although the stutters are generally not contiguous. The label k(N), with N in the range 3 to 11 is the sum of the lengths of the poly-lysine stutters, assuming a minimum of 3aa. Of the set of Group 2 LEA proteins, 16 have at poly-lysine stutters totalling at least 3aa (versus an expected count of 4.93, corresponding to a probability of $1.\ 5 \times 10^{-5}$). The application 0j.py [17] was used to find the poly-serine and poly-lysine stutters. The lists of hits against the different sequence motifs is followed by a column labelled SF (short for SuperFamily). This will be discussed in the section below on automated clustering of the LEA proteins. The final column in the tables, labelled Evidence, lists evidence supporting the protein's inclusion in the particular Group, beyond the articles cited in the SwissProt record. If the protein is included in a Pfam family, the family's identifier is listed, followed by either '_ml' or '_hmm'. The suffix '_ml' is used to indicate that the protein has been included in the edited multiple-sequence (or "seed") alignment that forms the basis for the family. The proteins annotated with '_hmm' are those recovered by the hidden Markov model that has been trained from the multiple sequence alignment (called by Pfam the "full" family). This is somewhat weaker evidence than the curated multiple alignment. Finally, if a SwissProt or SpTrEMBL identifier is shown, it is followed by a p-value and represents the closest match found by BLAST (without masking) from among the canonical LEA proteins in that Group or, in default, to a protein that in turn matches a canonical LEA protein.

The tables of sequences by Group are:

1 LEA protein Group 1 (D19) Exemplar: LE19_GOSHI

2 LEA protein Group 2 (D11) Exemplar: DH11_GOSHI

The set of Group 2 LEA proteins is subdivided into three parts. The reasons for this are canvassed below.

3 LEA protein Group 3 (D7) Exemplars: LE7_GOSHI, LE76_BRANA

4 LEA protein Group 4 (D113) Exemplar: LE13_GOSHI

5 LEA protein Group 5 (D29) Exemplar: LE29_GOSHI

6 LEA protein Group 6 (D34) Exemplar: LE34_GOSHI

7 LEA protein Group Lea5 (D73) Exemplar: LE5A_GOSHI

8 LEA protein Group Lea14 (D95) Exemplar: LE14_GOSHI

9 Uncharacterised LEA proteins

Three proteins where uncovered which are canonical LEA proteins but for which little or no similarity exists with known LEA protein sequences. One of this group also has expression levels due to ABA or desiccation/cold stress which closely follow the patterns viewed as characteristic of LEA proteins.

Machine Learning Applied to the LEA Protein Sequence Sets

Machine learning software takes a set of descriptions of objects, in this case proteins, and brings related ones together to form groups. There are two basic sorts of machine-learning algorithms-supervised and unsupervised learning [36,37]. Both sorts have been employed in this study. Supervised algorithms are given values for an array of features, such as maximum hydrophobicity or percentage composition of aliphatic residues, and an output class, e.g. Group 1, Group 2, etc. Rules are then induced which categorise each of the input examples into one of the set of output classes. The aim of the rule induction process is to minimise miscategorisation. In unsupervised machine-learning, (also known as "classification" or "data mining"), similar objects are clustered based on a metric, e.g. sequence similarity score. The aim is to maximise scores between members of clusters, while minimising inter-cluster scores.

Supervised Machine Learning Applied to LEA proteins – Ripper

From the surveys listed above different protein properties have been used to characterise the various LEA protein Groups. The most commonly noted are hydrophilicity and predicted secondary structure. To these have now been added percentage composition by amino-acid class, i.e. acid, basic, aliphatic, etc. Scores summarising these attributes, calculated from the protein sequences, formed the input to the supervised learning application Ripper [38].

Hydrophilicity

The EMBOSS [39] application Pepinfo was used to calculated hydrophobicity values based on the method of Kyte and Doolittle. A larger window, 21aa versus the default 9aa, was used at each amino acid in order to favour larger structures over smaller ones. That is, an average hydrophobicity value was calculated at each amino acid based on the hydrophobicity values of that amino acid, the previous 10 and the following 10. Three values were returned for each sequence: the minimum and maximum windows together with the average across all the windows. The ranges of these values were, respectively: -3.21 .. 0, -0.73 .. 2.25 and -1.70 .. 0.07; negative hydrophobicity values indicate hydrophilicity.

Predicted Secondary Structure Percentage Composition

No structures have been determined for any of the LEA proteins, so all analyses of structure for these proteins have been done on the basis of predictions based on the amino acid sequence. In this study, four-state predictions were obtained for each amino acid in the LEA proteins using PHDsec from the ProteinPredict server [40,41]. PHDsec takes a neural network approach. The ProteinPredict server returns two predictions for each amino acid: a three-state prediction (H/E/L) together with a value indicating the degree of confidence in that value, or a more stringent, four-state prediction, with the additional option of none of H, E or L being recorded if none prove significant. This is indicated by a '.'. The four-state predictions used in this study were converted to percentage composition values (e.g. the count of H predictions divided by the protein length), which minimises effects due to differences in length across the sequences. However, before the percentage composition values were calculated, some preprocessing was done to remove possible prediction artefacts, in particular predicted features encompassing a single amino acid, though beta-sheets of spanning just one amino acid could be beta-turns. Remembering that values must be in the range 0. . 1. 0, the ranges of values for H, E and L were respectively: 0. . 0. 85, 0. . 0. 17 and 0. 04. . 0. 60.

A number of alternative secondary structure prediction servers were tried, including NPS@ secondary structure consensus server [42], Prof, which combines different classifiers with a neural network [43,44] and SAM-T02 which uses Hidden Markov Model methods [45,46]. It is worth noting that all secondary structure predictors have been trained on the relatively small number of distinct globular proteins for which structures have been determined, typically from X-ray or NMR data. Bearing in mind that most of the LEA proteins have low sequence complexity and are probably not globular, any predications need to be viewed a little skeptically. In addition, three-state predictors have the problem that coil or loop is the default category so will tend to be over-predicted. Building a consensus of such values might therefore compound the problem. For example, when Prof was used to examine the Group 1 LEA protein EM1_ARATH, 150 of the 152 amino acids were labelled as coil. For the same protein, the NPS@ gave a percentage of 25.7% for helix and 67.8 for coil, PHDsec in its three-state mode returned 26.3% helix and 53.3% coil, while SAM-T02 returned 34.2% helix and 65.8% coil. By contrast, the PHDsec four-state mode gave 11.2% helix and 23.7% loop. The four-state prediction returned by PHDsec is more conservative and therefore was used for this study. In addition, use of percentage composition values should average out any point inaccuracies.

Amino Acid Class Percentage Composition

While issues of biases in the peptide composition of LEA proteins will be more fully explored using unsupervised machine learning, it was believed that a general classification could provide added detail to that afforded by the hydrophobicity values. The amino acid types and the ranges in their values are: Aliphatic (0. 03. . 0. 29), Aromatic (0. 01. . 0. 15), Non-polar (0. 32. . 0. 59), Polar (0. 41. . 0. 68), Charged (0. 19. . 0. 52), Basic (0. 08. . 0. 28) and Acidic (0. 07. . 0. 28). The only point to note in the membership of the different sets is that the set of Aromatic residues includes histidine, as well as phenylalanine, tryptophan and tyrosine.

Unsupervised Machine Learning Applied to LEA Proteins – The POPPs

The method of choice for most biologists faced with protein sequence data is to compare their sequences against those in a protein database such as SwissProt using the Smith-Waterman algorithm, e.g. Scanps [47] or approximations to the Smith-Waterman algorithm, such as BLAST [48]. The POPPs suite of tools [25], available under license from the author, employs an alternative approach, based on comparisons of sets of peptides that are "unusual" in the proteins under comparison.

Significant LEA Protein Peptides

The first application in the suite is called popp_create.py. Given one or more sequences or files of sequences popp_create.py compares the distributions of peptides of length 1aa – 3aa (typically), found in the individual sequences or across files of sequences, versus their distributions across a suitably large database (currently Swiss-Prot plus SpTrEMBL, also called Swall). A single-sided binomial distribution statistic is used to produce a list of those peptides that are either significantly over-represented in the samples versus the database or significantly under-represented, both with respect to a user-specified threshold p-value. Peptides whose absolute probability is greater than the threshold are not reported. This list, called a Protein or Oligonucleotide Probability Profile, or "POPP", can provide useful information about the sorts of peptides that are characteristic of the sequence or group of sequences. Sequences corresponding to the different Groups were placed into separate databases and popp_create.py was then applied to each database.

Clustering LEA proteins

An alternative output format available to popp_create.py is the creation of a POPP vector for each input sequence. POPP vectors contain the same information as the profiles but in a compressed form; the profiles are formatted for inspection by users while the vectors are used by the second component of The POPPs, popp_cmp.py. popp_cmp.py applies a clustering algorithm to the POPP vectors so that related proteins are formed into groups around a consensus POPP, i.e. a POPP composed of those peptides that are significantly under or over represented in all the component POPPs. Details of the algorithm can be found in [25]. However, from the user's point of view an important feature is that POPP vectors are not forced to belong to a single cluster but can appear in any cluster where this is appropriate.

The same clustering algorithms are also used to perform meta-clustering. That is, the consensusPOPPs found in the first pass are themselves clustered into families. Furthermore, if the various families are sufficiently similar, groups of families are brought together into superfamilies, which are distinguished by the fact that each family in a superfamily shares at least one cluster with at least one of the other families. The most highly connected (i.e. most representative) family is selected as the "anchor" of its superfamily.

In the context of the current investigations, the application popp_create.py was used to create a POPP vector for each of the LEA protein sequences – Group 1 to Group 6, plus Groups Lea5 and Lea14 – together with the Uncharacterised set. The application popp_cmp.py was then used to cluster the POPP vectors; the results are discussed below.

Keyword Clustering Applied to Sets of Related POPPs Vectors

When POPPs are gathered into clusters, families and superfamilies a consensus POPP is also reported. The consensus POPP contains the peptides that significantly under- or over-represented in all the POPPs making up the cluster, family or superfamily. Another POPP analysis tool, popp_search.py, can then be used to search a POPP-vector database (in this case created from SwissProt) for proteins related to a query sequence by similar biases in their peptide compositions. Searches were undertaken based on the consensus POPPs from the anchor family in each superfamily. In the final step of this process, ignoring the hits against the sequences forming the consensus (i.e. search) POPPs, the remaining hits were submitted to the protein keyword clustering application, Protein Annotators' Assistant [49,50]. This web-based application takes a list of SwissProt identifiers or accession numbers and returns a list of keywords or phrases that characterise subsets of the input proteins, automating a process that is typically done by hand, e.g. from BLAST hits.

Additional Material

Additional material can be found by unzipping the Additional file: 1. The resulting web pages list the data used in the experiments and the outputs that resulted, in particular from the unsupervised machine learning experiments using The POPPs suite.

Additional material

Additional file 1

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-52-S1.zip]

Acknowledgements

I would like to thank Dr Alan Tunnacliffe, Institute of Biotechnology, Cambridge University, for making me aware of the LEA proteins, and for making extremely useful comments on the results of the investigations that I have undertaken on them. This paper has also benefited greatly from his comments, and from the comments of the reviewers. I would also like to acknowledge the generous support for my Fellowship provided by Bristol-Myers Squibb.

References

1. Bray EA: **Molecular Responses to Water Deficit.** *Plant Physiol* 1993, **103:**1035-1040.
2. Ingram J and Bartels D: **The Molecular Basis of Dehydration Tolerance in Plants.** *Annu Rev Plant Physiol Plant Mol Biol* 1996, **47:**377-403.

3. Cuming AC: **LEA Proteins,.** In *Seed Proteins* Edited by: *Peter R. Shewry and Rod Casey. Kluwer Academic Publishers*; 1999:753-780.
4. Bray EA, Bailey-Serres J and Weretilnyk E: **Responses to Abiotic Stress,.** In *Biochemistry and Molecular Biology of Plants* Edited by: *Bob B. Buchanan, Wilhelm Gruissem and Russell L. Jones. American Society of Plant Physiologists*; 2000:1158-1203.
5. Baker J, Steele C and Dure L III: **Sequence and Characterization of 6 *Lea* Proteins and their Genes from Cotton.** *Plant Mol Biol* 1988, **11:**277-291.
6. Dure L III, Crouch M, Harada J, Ho T.-HD, Mundy J, Quatrano R, Thomas T and Sung ZR: **Common Amino Acid Sequence Domains among the LEA Proteins of Higher Plants.** *Plant Mol Biol* 1989, **12:**475-486.
7. Hughes DW and Galau GA: **Temporally Modular Gene Expression During Cotyledon Development.** *Genes Dev* 1989, **3:**358-369.
8. Stacy RAP and Aalen RB: **Identification of Sequence Homology Between the Internal Hydrophilic Repeated Motifs in Group 1 Late-Embryogenesis-Abundant Proteins in Plants and Hydrophilic Repeats of the General Stress Protein GsiB of *Bacillus subtilis.*** *Planta* 1998, **206:**476-478.
9. Makarova KS, Aravind L, Wolf YI, Tatusov RL, Minton KW, Koonin EV and Daly MJ: **Genome of the Extremely Radiation-Resistant Bacterium *Deinococcus radiodurans* Viewed from the Perspective of Comparative Genomics.** *Microbiol Mol Biol Rev* 2001, **65:**44-79.
10. Browne J, Tunnacliffe A and Burnell A: **Plant Desiccation Gene Found in a Nematode.** *Nature* 2002, **416:**38.
11. Dure III L: **Structural Motifs in LEA Proteins,.** In *Plant Responses to Cellular Dehydration During Environmental Stress* Edited by: *Timothy J. Close and Elizabeth A. Bray. American Society of Plant Physiologists*; 1993:91-103.
12. Bray EA: **Alterations in Gene Expression in Response to Water Deficit,.** In *Stress-Induced Gene Expression in Plants* Edited by: *Amarjit S. Basra. Harwood Academic*; 1994:1-23.
13. Close TJ: **Dehydrins: A Commonalty in the Response of Plants to Dehydration and Low Temperature.** *Physiol Plant* 1997, **100:**291-296.
14. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy S, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer ELL: **The Pfam Protein Families Database.** *Nucleic Acids Res* 2002, **30:**276-280.
15. Galau GA, Wang HY.-C and Hughes DW: **Cotton *Lea5* and *Lea14* Encode Atypical Late Embryogenesis-Abundant Proteins.** *Plant Physiol* 1993, **101:**695-696.
16. Garay-Arroyo A, Colmenero-Flores JM, Garciarrubio A and Covarrubias AA: **Highly Hydrophilic Proteins in Prokaryotes and Eukaryotes are Common during Conditions of Water Deficit.** *J Biol Chem* 2000, **275:**5668-5674.
17. Wise MJ: **0j.py: A Software Tool for Low Complexity Proteins and Protein Domains.** *Bioinformatics* 2001, **Suppl17:**288-295.
18. Altschul SF and Gish W: **Local Alignment Statistics,.** In *Computer Methods for Macromolecular Sequence Analysis* Edited by: *Russell F. Doolittle. Academic Press*; 1996:460-480.
19. Brenner SE, Chothia C and Hubbard TJP: **Assessing Sequence Comparison Methods with Reliable Structurally Identified Distant Evolutionary Relationships.** *Proc Natl Acad Sci USA* 1998, **95:**6073-6078.
20. Altschul SF, Boguski MS, Gish W and Wootton JC: **Issues in Searching Molecular Sequence Databases.** *Nat Genet* 1994, **6:**119-129.
21. Wootton JC and Federhen S: **Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases.** *Comput Chem* 1993, **17:**149-163.
22. Dure L III: **Occurrence of a Repeating 11-mer Amino Acid Sequence Motif in Diverse Organisms.** *Protein Pept Lett* 2001, **8:**115-122.
23. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJA, Hofmann K and Bairoch A: **The PROSITE Database, its Status in 2002.** *Nucleic Acids Res* 2002, **30:**235-238.
24. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A and Zygouri C: **PRINTS and its Automatic Supplement, prePRINTS.** *Nucleic Acids Res* 2003, **31:**400-402.
25. Wise MJ: **The POPPs: Clustering and Searching Using Peptide Probability Profiles.** *Bioinformatics* 2002, **Suppl18:**38-45.
26. Goyal K, Tisi L, Basran A, Browne J, Burnell A, Zurdo J and Tunnacliffe A: **Transition from Natively Unfolded to Folded State Induced by Desiccation in an Anhydrobiotic Nematode Protein.** *J Biol Chem* 2003, **278:**12977-12984.
27. Lisse T, Bartels D, Kalbitzer HR and Jaenicke R: **The Recombinant Dehydrin-Like Desiccation Stress Protein from the Resurrection Plant *Craterostigma plantagineum* Displays No Defined Three-Dimensional Structure in Its Native State.** *Biol Chem* 1996, **377:**555-561.
28. Ismail AM, Hall AE and Close TJ: **Purification and Partial Characterization of a Dehydrin Involved in Chilling Tolerance during Seedling Emergence of Cowpea.** *Plant Physiol* 1999, **120:**237-244.
29. Berge SK, Bartholomew DM and Quatrano RS: **Control of the Expression of Wheat Embryo Genes by Abscisic Acid,.** In *The Molecular Basis of Plant Development* 1989:193-201.
30. Yamaguchi-Shinozaki K and Shinozaki K: **The Plant Hormone Abscisic Acid Mediates the Drought-Induced Expression but not the Seed-Specific Expression of *rd22*, a Gene Responsive to Dehydration Stress in.** *Arabidopsis thaliana. Mol Gen Genet* 1993, **238:**17-25.
31. Bartels D, Schneider K, Terstappen G, Piatkowski D and Salamani F: **Molecular Cloning of Abscisic Acid-Modulated Genes which are Induced during Desiccation of the Resurrection Plant.** *Craterostigma plantagineum. Planta* 1990, **181:**27-34.
32. Hsing YC, Tsou C, Hsu T, Chen Z, Hsieh K, Hsieh J and Chow T: **Tissue and Stage-Specific Expression of a Soybean (*Glycine max* L.) Seed Maturaion, Biotinylated Protein.** *Plant Mol Biol* 1998, **38:**481-490.
33. Niu S, Antin PB and Morkin E: **Cloning and Sequencing of a Developmentally Regulated Avian mRNA Containing the LEA Motif Found in Plant Seed Proteins.** *Gene* 1996, **175:**187-191.
34. Zdobnov EM, Lopez R, Apweiler R and Etzold T: **The EBI SRS Server – Recent Developments.** *Bioinformatics* 2002, **18:**368-373.
35. Wu S and Manber U: **Fast Text Searching Allowing Errors.** *Commun ACM* 1992, **35:**83-91.
36. Shavlik JW and Dietterich TG: **General Aspects of Machine Learning,.** In *Readings in Machine Learning* Edited by: *Jude W. Shavlik and Thomas G. Dietterich. Morgan Kaufmann*; 1990:1-10.
37. Mitchell TM: *Machine Learning McGraw Hill*; 1997.
38. Cohen WW: **Fast Effective Rule Induction.** In *Twelfth International Conference on Machine Learning: July 9–12, 1995 Lake Tahoe, U.S.A. Morgan Kaufmann*; 1995:115-123.
39. **EMBOSS** [http://www.hgmp.mrc.ac.uk/Software/EMBOSS/]
40. Rost B: **PHD: Predicting 1D Protein Structure by Profile Based Neural Networks.** In *Methods in Enzymology 266* Edited by: *Russell F. Doolittle. Academic Press*; 1996:525-539.
41. **ProteinPredict** [http://cubic.bioc.columbia.edu/predictprotein]
42. **NPS@ (Network Protein Sequence @nalysis) Server** [http://npsa-pbil.ibcp.fr/]
43. Ouali M and King RD: **Cascaded Multiple Classifiers for Secondary Structure Prediction.** *Protein Sci* 2000, **9:**1162-1176.
44. **PROF – Secondary Structure Prediction System** [http://www.aber.ac.uk/~phiwww/prof/]
45. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M and Hughey R: **Combining Local-Structure, Fold-Recognition, and new-Fold Methods for Protein Structure Prediction.** *Proteins* 2003.
46. **HMM-based Protein Structure Prediction, SAM-T02** [http://www.soe.ucsc.edu/research/compbio/SAM_T02/T02-query.html]
47. Barton GJ: **An Efficient Algorithm to Locate all Locally Optimal Alignments between Two Sequences Allowing for Gaps.** *CABIOS* 1993, **9:**729-734.
48. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ: **Basic Local Alignment Search Tool.** *J Mol Biol* 1990, **215:**403-410.
49. Wise MJ: **Protein Annotators' Assistant.** *Trends Biochem Sci* 2000, **25:**252-253.
50. **Protein Annotators' Assistant** [http://www.ebi.ac.uk/paa]