Characterisation of an *Arabidopsis* cDNA encoding a thylakoid lumen protein related to a novel 'pentapeptide repeat' family of proteins

Thomas Kieselbach[a], Alexandra Mant[b], Colin Robinson[b], Wolfgang P. Schroder[a,*]

[a] *Arrhenius Laboratory, Department of Biochemistry, Stockholm University, S-106 91 Stockholm, Sweden*
[b] *Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK*

Received 17 February 1998; revised version received 22 April 1998

Abstract **We have cloned an *Arabidopsis* cDNA encoding a novel thylakoid lumen protein, P17.4, that has been previously isolated from lumen extracts of spinach chloroplasts. The protein is synthesised with a bipartite presequence containing a Sec-type lumen-targeting signal peptide and the precursor protein is imported into the lumen of pea chloroplasts. The encoded protein is homologous to an *Anabaena* protein that is essential for correct glycolipid localisation, and is also related to at least 16 unassigned open reading frames in *Synechocystis*. This family of proteins is characterised by the presence of numerous pentapeptide repeats with the consensus structure AXLXX, and its members are predicted to be located in the cytosol, plasma membrane and periplasm/lumen. P17.4 is therefore the first higher plant member of an extended family of putative cyanobacterial proteins that may serve important roles in lipid transport or assembly.**

Key words: Chloroplast; Membrane biogenesis; Photosynthesis; Protein translocation; Thylakoid lumen

1. Introduction

The thylakoid membranes of chloroplasts and cyanobacteria carry out the light reactions of oxygenic photosynthesis, in which sunlight is used to drive the transfer of electrons from water to $NADP^+$ and the resultant proton gradient is harnessed to drive ATP synthesis. These processes are carried out by four major complexes in the thylakoid membrane (photosystems I and II, the cytochrome *bf* complex and the ATP synthase), together with a small number of ancillary carriers. The complexes themselves have been analysed in great detail and the component proteins have, for the most part, been characterised and cloned. Photosynthetic proteins account for the vast majority of chloroplast membrane protein and the thylakoid membranes of chloroplasts account for approximately 99% of the total chloroplast lipid content.

Although the photosynthetic process of thylakoid membranes have been studied in intricate detail, other functions of this highly abundant biomembrane have been slow to emerge and remarkably few non-photosynthetic thylakoid proteins have been identified. This applies particularly to the soluble lumen phase enclosed by the thylakoid membrane. Several lumenal photosynthetic proteins have been well-characterised, of which plastocyanin is probably the best known, and others identified recently include the D1 processing peptidase [1] and polyphenol oxidase [2,3]. However, most lumenal proteins have yet to be characterised in any real detail. In order to gain insight into the resident proteins of this phase, and hence understand more about the functions of the thylakoid network, we recently devised a fractionation procedure which allows the preparation of lumen contents in a highly purified form [4]. At least 25 polypeptides could be identified by SDS-polyacrylamide electrophoresis, of which only six correspond to known lumen photosynthetic proteins. The majority of the remainder are of unknown identity but one has recently been further purified and shown to be a prolyl cis/trans isomerase [5]. N-terminal sequence data were obtained for several of the other proteins and in this report we have used these data to identify and characterise an *Arabidopsis* cDNA encoding one of the proteins. We show that the clone encodes a novel 17 kDa lumenal protein with significant homology to a new protein family from *Synechocystis* and an *Anabaena* protein that is essential for correct localisation of glycolipids.

2. Materials and methods

An *Arabidopsis* cDNA encoding a 17.4 kDa lumenal protein from spinach (P17.4) was identified by searching the expressed sequence tag (EST) database using the N-terminal 16-residue sequence of P17.4 [4]. The cDNA (clone 104N18T7) was obtained from the *Arabidopsis* Biological Resource Center at Ohio State University and both strands were fully sequenced. The EMBL accession number for the cDNA encoding P17.4 is T21992. The encoded precursor protein was synthesised in vitro by transcription translation and imported into intact pea chloroplasts as described [6]. Prediction of signal peptide cleavage sites was carried out using the signalP programme [7].

3. Results and discussion

3.1. *Arabidopsis P17.4 is synthesised as a preprotein with a lumen-targeting, bipartite presequence*

The DNA and predicted protein sequence of an *Arabidopsis* cDNA encoding P17.4 have been deposited in the EMBL database. The 16-residue N-terminal sequence determined for spinach P17.4 [4] is highly homologous to residues 78–93 in the *Arabidopsis* sequence, with 12 of the 16 residues identical in this region. The *Arabidopsis* protein is thus synthesised with a 77-residue presequence. As with all known cytosolically synthesised thylakoid lumen proteins, P17.4 is synthesised with a bipartite presequence in which two distinct domains can be recognised. The first domain corresponds to an 'envelope transit' peptide which directs translocation into the stroma by an envelope-localised protein translocase. These peptides are usually hydrophilic, overall positively charged and enriched in hydroxylated residues [8], and the first ca. 40 residues of the *Arabidopsis* precursor protein fit these criteria. The second targeting signal specifies translocation across the thylakoid membrane, and these peptides invariably contain a

*Corresponding author. Fax: +46 (8) 1153679.
E-mail: Wolfgang@biokemi.su.se

PII S0014-5793(98)00517-1

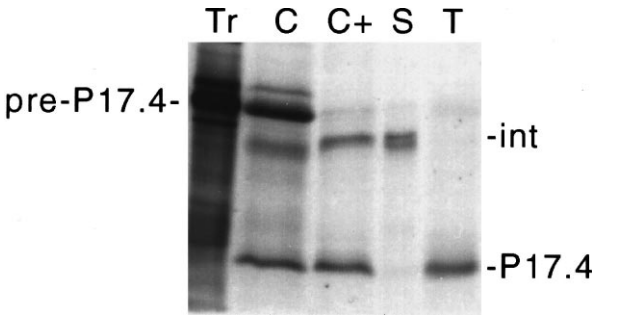

Fig. 1. Pre-P17.4 is imported into the thylakoid lumen of pea chloroplasts. Pre-P17.4 was synthesised in vitro by transcription translation of a cDNA clone and the precursor protein (lane Tr) was incubated with isolated intact pea chloroplasts. After incubation, samples were analysed of the chloroplasts (C), protease-treated chloroplasts (C+) and the stromal (S) and thylakoid (T) fractions after lysis of the organelles and pelleting of the membranes. Int denotes stromal intermediate form.

hydrophobic core region followed by a motif specifying cleavage by the thylakoidal processing peptidase, in which a short-chain residue is present at the −3 position and alanine is essential at the −1 position [8,9]. The *Arabidopsis* presequence is typical in these respects (see Fig. 2B), containing a hydrophobic core region in the second domain with the start of the mature protein preceded by the sequence Val-X-Ala. Confirmation of these targeting properties, and the lumenal location of the mature *Arabidopsis* protein, was obtained using in vitro assays for the import of proteins by intact chloroplasts. Fig. 1 shows that the protein is synthesised as a 28 kDa precursor protein which is imported into the organelles, targeted into the thylakoids (lane T) and processed to a 17 kDa polypeptide. The mature protein is resistant to protease digestion as expected of a lumenal protein (data not shown). Some intermediate-size polypeptide is apparent in the stromal fraction (lane S), suggesting that the envelope transit peptide is removed by the stromal processing peptidase, as is the case with most bipartite presequences.

Although all known lumenal proteins are synthesised with superficially similar bipartite presequences, the thylakoid-targeting peptides have been found to direct translocation by two distinct pathways once inside the chloroplast (reviewed in [10]). Some lumenal proteins are targeted by an ATP-dependent, Sec-related system whereas others use a very different system that is driven by the thylakoidal ΔpH. The choice of pathway is dictated by the type of targeting peptide [11,12] and, while the precise sorting characteristics of the signals remain to be fully defined, targeting by the ΔpH-dependent pathway is known to depend totally on the presence of a twin-arginine motif immediately prior to the hydrophobic region of the targeting signal [13]. No such motif is present before the hydrophobic region of the presequence, strongly suggesting that translocation takes place by the Sec-dependent pathway. Other tests (data not shown) have confirmed that dissipation of the thylakoidal ΔpH does not inhibit translocation across the thylakoid lumen whereas the SecA inhibitor, azide, does inhibit translocation to an extent. We conclude that this pro-

A

```
Anab.     SVGSASPEREARQKRLQINQQTINKILNVYSGKTLGLDLSRTQLGQSGTV
P17.4     ------------------------------GKTLSAAL---------MV
                                         ****.  *         *

Anab.     GGSFFNLILDNIDLSGIKFKSANLNQASFKGSRFRSVGDDGRWDTYDDAI
P17.4     GAKF-----DGADMTEVVMSKAYAVEASFKG--------------------
          *. *     *  *.. .    *   .*****

Anab.     ADLSQAQMKQANFTDANLSRVLMTRSDLSRATLNRANLSNARLIGANLSS
P17.4     ----------VNFTNAVIDRVNFGKSNL----------------------
                    ***.* . **   .*.*

Anab.     AQLVGADLRGTVLENASLTGADLGDAKLQEANLYGARLSRVIAIGAQLSF
P17.4     --------KGAVFRNTVLSGSTFEEANLED----------VVFEDTIIGY
                 .*.*. *. *.*. . .*.*..          *.  .. ...

Anab.     ANLTKTDWQSSDLSGADLERANLSNADLSATRMTGAILRSAQLENANLRN
P17.4     IDLQKI-------------------------------------------
           .* *

Anab.     ADLSLVDLRGANVAGADFKDTILTPSRQDPADQFVQTPELGSVSAVVKGV
P17.4     ------------------------CRNESINE-----------------
                                  *... ..

Anab.     DFSQAKNLDGKQLAYICTQGGVHPRCP
P17.4     -----------------EGRLVLGCR
                          .*  .   *
```

B

```
(P17.4 pre-   MASLPVQFTRNQISSPFFSVKLRREPRSLVTVHCSAGENRENGEGVKKSLF
sequence)     PLKELGSIACAALCACTLTIASPVIA*

spinach P17.4 ANQ                RLPPLSNDPKR
P17.4         ANQ-----------------RLPPLSTEPDR-----------------C
slr1819       MEAKELVQRYRNGETLFTGLKLPGINLEAADLIGIVLNEADLRGANLLFC
                .                  .** . *.                   *

P17.4         --EKAFVGNTIGQANGVYDKPL--------DLRFCDYTNDQ---------
slr1819       YLNRANLGQANLVAANLSGASLNQADLAGADLRSANFHGAMLQGAILRDS
                ..* .*..   *  . . .*       ***  ..

P17.4         --------------TNLKGKTLSAALMVG----------------------
slr1819       DMTLATLQDTNLIGADLRGADLSGATLTGACLRGANMRQEKKGYYTNLQA
                        ..*.*  **.*  .  *

P17.4         -----AKFDGADMTEVVMSKAYAVEASFKGVNFTNAVIDRVNFGKSNLKG
slr1819       AILGRADLQGANMKGVDLSRADLSYANLKEANLRDVDLRKADLSYANLKG
                   * ..**.*  * .*.*   *..*  *. . . . . . .****

P17.4         AVFRNTVLSGSTFEEANLEDVVFED-------------EDTIIGYIDLQK
slr1819       ALLTDANLSGAKLNGADLQNANLMRAKISEAEMTAVNCQGAIMTHVNLNR
              *.. .. ***. . .*.*..  .                  ...*.  ..*..

P17.4         -----------------ICRNESIN---EEGRLV----------------
slr1819       TNLTGSNLSFTRMNSADLSRANLTKANLQEAELIEAFFARANLTEANFIN
                          . * .  .  .*. *.

P17.4         --------------------------LGCR
slr1819       ANLVRADLMSANMVGADFQGATMPDGQVRHH
```

Fig. 2. P17.4 is homologous to an *Anabaena* protein, HglK, essential for heterocyst-specific glycolipid localisation, and to a large family of hypothetical *Synechocystis* proteins. A: The sequence of the mature-size P17.4 is shown aligned against the C-terminal region of *Anabaena* HglK (Anab.) using the Clustal V programme. Identical residues are denoted by asterisks and conserved residues by dots. B: The entire sequence of *Arabidopsis* pre-P17.4 is shown aligned against the sequence of the *Synechocystis* slr1819 gene product; the bipartite presequence of the plant protein is shown separately above the alignment, with the hydrophobic region underlined and the terminal processing site denoted by an asterisk. The N-terminal sequence of spinach P17.4 (*italicised*) is shown above the homologous region of the *Arabidopsis* protein. AXLXX-type pentapeptide repeat regions are underlined with the first and third residues of each motif shown in bold.

Table 1
Synechocystis homologues of P17.4

Gene	AAs	Predicted location
slr1819	331	cytoplasmic
sll0301	169	lumenal/periplasmic
slr0516	166	membrane
slr1152	331	cytoplasmic
sll0577	169	lumenal/periplasmic
slr1851	162	membrane
sll1446	320	cytoplasmic
slr0719	388	membrane
sll0183	259	membrane
slr0967	150	lumenal/periplasmic
slr1697	574	membrane
sll1011	270	cytoplasmic
sll0274	196	membrane
sll0414	286	cytoplasmic
sll1350	398	membrane
slr1519	245	putative membrane

The table lists the open reading frames within the *Synechocystis* genome that are highly homologous to the mature-size *Arabidopsis* P17.4 sequence. The hydrophobicity of each sequence was assessed using the TopPred 2 programme [15] and the encoded proteins are given putative locations in the cytoplasm or membrane according to the prediction obtained. Those proteins containing a single N-terminal hydrophobic region were further assessed for the possible possession of a cleavable signal peptide using the signalP programme [7] and three members have been given a predicted lumen/periplasm location accordingly.

tein is transported by the Sec pathway as predicted from the presequence structure.

3.2. *P17.4 is homologous to an Anabaena protein required for glycolipid localisation, and to a large family of hypothetical Synechocystis proteins containing AXLXX pentapeptide repeats*

A search of the databases reveals that P17.4 bears high homology to only one known protein: the product of the *hglK* gene in *Anabaena* sp strain PCC 7120 [14]. Interestingly, it was found that disruption of this gene led to a defect in the localisation of heterocyst-specific glycolipids, and the authors proposed that the *hglK* gene product may be involved in the transport or assembly of these lipids. The HglK protein is much larger than P17.4 (727 amino acids compared with 159 residues for the P17.4 *mature* protein) and it is the C-terminal half of the protein that is most homologous to P17.4 (Fig. 2A). A very surprising, and thus far unexplained, feature of the HglK protein is the presence of 36 degenerate AXLXX pentapeptide repeats [14].

Although HglK is the only *known* protein bearing a resemblance to P17.4, several open reading frames within prokaryotic genomes are highly homologous. Among these are a protein encoded by a plasmid in *Erwinia stewartii* (accession number L42525) and, most notably, a large family of 16 hypothetical proteins in *Synechocystis*. Several of these proteins are highly homologous to P17.4 and one such alignment is shown in Fig. 2B. As with the *Anabaena* HglK protein, this *Synechocystis* protein contains multiple copies of the pentapeptide repeat, as do all members of this particular family of hypothetical proteins. Close examination of the mature P17.4 sequence shows that this protein is also a member of this 'pentapeptide repeat' family, containing at least 16 repeat motifs that are shown underlined in Fig. 2B. However, the first and third amino acids of the *Arabidopsis* P17.4 repeats are more variable. Within the 16 repeats, the first position is likewise occupied in most cases by Ala (9 of 16) but Val is found in four cases, Ser in two and Thr in one. The third position is again invariably hydrophobic but Leu is found in only four cases and Phe is in fact most commonly found (seven cases) with the other repeats containing Met (three cases), Ile (one example) and in one case, Ala. Since this entire region is also conserved in the *Synechocystis* slr1819 sequence, it can be seen that the cyanobacterial motifs also diverge in a few cases from the AXLXX motif, with Val present in some cases at the first position. However, it is striking that the Leu is so highly conserved. As yet, we have no information concerning the function of these repeat units although there is evidence that they are essential for the functioning of the *hglK* gene product [14].

A list of the other *Synechocystis* proteins is shown in Table 1, together with a description of their predicted sizes. Although some of the *Synechocystis* proteins are less homologous, all 16 hypothetical proteins are sequence-related because all display high homology to one another and each contains numerous copies of the characteristic pentapeptide repeat (not shown). We have also made an estimate of the locations of the proteins as deduced from their hydrophobicity profiles combined with an analysis of the N-terminal sequences. Several of the family are strongly predicted to be membrane-bound but others appear to be soluble cytosolic proteins and three are synthesised with N-terminal hydrophobic regions that are strongly predicted to be signal peptides. The latter proteins are thus probably transported into the periplasm or thylakoid lumen (or both) and may be functionally analogous to P17.4. The combined data therefore suggest that this family contains members that are cytosolic, membrane-bound and soluble in the periplasm/lumen. The functions of these proteins are as yet unknown but, given the importance of the *hglK* gene for glycolipid localisation, it may well be the case that at least some members are involved in membrane biogenesis. This possibility remains to be tested but we would point out that the sheer number of the proteins in *Synechocystis* argues for an important physiological function in most, if not all, of the cellular compartments.

Acknowledgements: T.K. was a recipient of a guest research grant by the Swedish Institute. This work was supported by Biotechnology and Biological Sciences Research Council grant C07900 to C.R. and by grants from the Swedish Forestry and Agriculture Research Council and the Swedish Natural Sciences Research Council to W.P.S. We are also very grateful to the *Arabidopsis* Biological Resource Center for their support in this work.

References

[1] Oelmüller, R., Herrmann, R.G. and Pakrasi, H.B. (1996) J. Biol. Chem. 271, 21848–21852.
[2] Sommer, A., Neéman, E., Steffens, J.C., Mayer, A.M. and Harel, E. (1994) Plant Physiol. 105, 1301–1311.
[3] Sololenko, A., Fulgosi, H., Gal, A., Altschmied, L., Ohad, I. and Herrmann, R.G. (1995) FEBS Lett. 371, 176–180.
[4] Kieselbach, T., Hagman, Å., Andersson, B. and Schröder, W.P. J. Biol. Chem. 273, 6710–6716.
[5] Fulgosi, H., Vener, A., Altschmied, L., Herrmann, R.G. and Andersson, B. EMBO J., in press.
[6] Kim, S.J., Robinson, D. and Robinson, C. (1996) FEBS Lett. 390, 175–178.
[7] Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Protein Eng. 10, 1–6.
[8] von Heijne, G., Steppuhn, J. and Herrmann, R.G. (1989) Eur. J. Biochem. 180, 535–545.

[9] Shackleton, J.B. and Robinson, C. (1991) J. Biol. Chem. 266, 12152–12156.
[10] Robinson, C. and Mant, A. (1997) Trends Plant Sci. 2, 431–437.
[11] Robinson, C., Cai, D., Hulford, A., Brock, I.W., Michl, D., Hazell, L., Schmidt, I., Herrmann, R.G. and Klösgen, R.B. (1994) EMBO J. 13, 279–285.
[12] Henry, R., Kapazoglou, A., McCaffery, M. and Cline, K. (1994) J. Biol. Chem. 269, 10189–10192.
[13] Chaddock, A.M., Mant, A., Karnauchov, I., Brink, S., Herrmann, R.G., Klösgen, R.B. and Robinson, C. (1995) EMBO J. 14, 2715–2722.
[14] Black, K., Buikema, W.J. and Haselkorn, R. (1995) J. Bacteriol. 177, 6440–6448.
[15] Claros, M.G. and von Heijne, G. (1994) CABIOS 10, 685–686.