The 20S proteasome gene family in *Arabidopsis thaliana*

Yves Parmentier[a], David Bouchez[b], Jacqueline Fleck[a], Pascal Genschik[a,*]

[a]*Institut de Biologie Moléculaire des Plantes du C.N.R.S., Université Louis Pasteur, 12 rue du Général Zimmer, 67084 Strasbourg Cedex, France*
[b]*Laboratoire de Biologie Cellulaire, INRA, 78026 Versailles Cedex, France*

Received 29 July 1997; revised version received 22 September 1997

Abstract The complexity of the proteasome gene family in higher plants was investigated by identification and sequencing cDNA clones from the *Arabidopsis thaliana* database showing homologies to 20S proteasome subunits. We identified plant counterparts for each of the 14 proteasomal subunit subfamilies. Moreover, several of them were highly related isoforms. Mapping data indicate a random distribution of the proteasome genes over the *Arabidopsis* genome.

Key words: Proteasome; Ubiquitin; Phylogenetic tree; Genomic mapping; cDNA; Plant

1. Introduction

Ubiquitin-26S- (Ubi/26S)-dependent protein degradation is an important mechanism involved in many different aspects of cellular regulation, including cell cycle control [1,2] and certain forms of cell death and apoptosis [3–5]. This mechanism which is selective and highly regulated enables specific proteolysis in the cytoplasm and nucleus with no damaging effects in these compartments. The Ubi/26S protein degradation pathway requires first the activation of ubiquitin by an E1 'activating' enzyme and then the transfer of the ubiquitin moiety to an E2 'conjugating' enzyme. Ubiquitin is subsequently conjugated to a target protein, directly or indirectly, via an E3 'ligating' enzyme. Polyubiquitinated proteins are further degraded by the 26S proteolytic complex (26S proteasome) in an ATP-dependent manner. This complex is composed of a cylindrical 20S multicatalytic protease particle (20S proteasome) and other associated proteins that form cap structures at both ends of the cylinder. The cap proteins are involved in the ubiquitinated protein binding, in the unfolding of the substrate, in the transfer to the catalytic 20S core and finally in the re-cycling of ubiquitin monomers [6]. Apart from proteolysis of regulatory proteins, the 20S proteasome has been implicated in misfolded or damaged protein degradation. It also participates in protein processing pathways [7,8] and in MHC class I antigen presentation [1]. The 26S proteasome is also involved in ubiquitin-independent proteolytic pathways [9].

Immuno-electron microscopy as well as X-ray crystallographic data showed that, in the archaebacterium *Thermoplasma acidophilum*, the 20S complex is constituted of 28 subunits arranged in four rings of seven subunits each surrounding a central cavity [10]. The two inner rings are composed of seven identical β-subunits whereas the outer two rings are composed of seven identical α-subunits, giving an α7β7β7α7 structure. Only the β-subunits are proteolytically active, with the active sites facing the central cavity.

Based on the homology with the two archaebacterial subunits, all eukaryotic proteasomal subunits can be classified as either α- or β-type [11]. In yeast seven different α-type and seven different β-type subunits have been characterized and the primary sequence of the proteins established [12,13]. Counterparts of each yeast 20S proteasome subunit were also reported for other organisms including *Caenorhabditis elegans*, *Drosophila melanogaster* and several vertebrates, which suggests a strong conservation in the structure of the particle. Immuno-electron microscopic data have also shown that the eukaryotic 20S proteasome structure is conserved [14], although it is of higher complexity in that each ring of the cylinder may be constituted of seven different subunits [15]. Crystallographic structural analysis of the yeast *Saccharomyces cerevisiae* 20S proteasome were recently described [16].

20S and 26S proteasomes have also been found in several plants [17,18] and purified from pea [19] and spinach [20,21]. It was shown that immunological, structural and biochemical characteristics of the plant particles were very similar to those of other eukaryotes. Indeed, it was no surprise to find, after cloning the first plant 20S subunits [22–24], that a strong protein primary sequence homology exists with the already reported proteasome proteins. However, to our knowledge, only a few plant 20S proteasome cDNAs have been described.

We report here the molecular characterization and genomic mapping of the 20S proteasome family constituted in *Arabidopsis thaliana* of at least 20 different expressed subunits with represented members in all the 14 phylogenetic branches.

2. Materials and methods

2.1. Identification and sequence analysis

The 31 234 expressed sequence tags (ESTs) deposited in the *A. thaliana* database (AtDB) at Stanford Genomic Resources in March 1996 [25] were analysed for proteasomal-like sequences using the *Arabidopsis* Blast search service and the two previously identified *Arabidopsis* proteasome sequences [22,23] as queries. About 70 ESTs were identified and the longest cDNAs were sequenced on both strands. Several of them were redundant and excluded from further analysis. Most of the selected cDNAs encoded the complete proteasomal protein sequence, which was furthermore confirmed by the presence of stop codons in the 5′-untranslated regions (not shown). Clones Prc6b_At and Prc6c_At were incomplete at the 3′ end. The sequence alignments and the phylogenetic dendrogram were performed with the UW-GCG PILEUP program [26]. The DDBJ/EMBL/GenBank database accession numbers are indicated between brackets: Prc1_At (Y13691); Prc2a_At (P34066), [24]; Prc2b_At (Y13182); Prc3_At (Y13176); Prc5_At (P42742), [23]; Prc6a_At (P30186), [22]; Prc6b_At (Y13179); Prc6c_At (Y13180); Prc8_At (Y13693); Prc9_At (Y13181); Prcd_At (Y13694); Prce_At (Y13695); Prcfa_At (Y13178); Prcfb_At (Y13174); Prcfc_At (Y13177); Prcga_At (Y13175); Prcgb_At (Y13692); Prch_At (Y13696); Prct_At (Y13173); Prcz_At (Z26556). The different motifs observed in the proteasomal protein primary

*Corresponding author. Fax: +33 (3) 88 61 44 42.
E-mail: pascalg@sylto.U-strasbg.fr

PII S0014-5793(97)01228-3

sequences were determined by using the 'PROSITE' database ([27]; 5th release).

2.2. Genomic mapping

Sequencing, probe labelling, *Arabidopsis* genomic DNA extractions and hybridizations were performed as described by Ausubel et al. [28]. Physical mapping was performed on the CIC library [29] and the RFLP-based genetic mapping by using 100 recombinant inbred (RI) lines [30]. The RI mapping was conducted using Mapmaker [31]. Positioning of the clones was based on published marker data for chromosome II [32], chromosome IV [33,34], chromosome V [35] and on unpublished marker data for chromosome I and III (D. Bouchez, unpublished).

3. Results and discussion

3.1. Identification of 20 different Arabidopsis cDNAs belonging to the 14 eukaryotic proteasome subunit subfamilies

As the proteasome subunits are evolutionarily highly conserved proteins, we used the previously identified *Arabidopsis* α-type: Prc6a_At (TASg64; [22]) and the β-type: Prc5_At (FAFp98; [23]) sequences for database searches. The large-scale analysis of the ESTs from *A. thaliana* revealed at least 20 non-redundant proteasome sequences (Fig. 1).

Phylogenetic analysis, established together with the known yeast proteasomal subunits present in *Saccharomyces* genome database (SGD) at Stanford Genomic Resources, revealed a 14 branch tree with 7 α-type and 7 β-type subunit subgroups (Fig. 2). The nomenclature we adopted for *Arabidopsis* was thus based on the yeast locus names in the databases. This extensive cDNA characterization clearly demonstrates that plant homologs exist for all 14 different proteasome subunit subfamilies. A search of the fully sequenced yeast genome with the different proteasomal protein sequences as a query did not identify other related subunit sequences. Nevertheless, while yeast counts a single member in each of the 14 subgroups, which fits exactly with the 14 different protein spots seen on two-dimensional protein gels of purified yeast proteasomes [36], *Arabidopsis* has different isoforms in several subfamilies (Prc2_At, Prc6_At, Prcf_At and Prcg_At), indicating that the gene family is more complex. These results even suggest that the 10 different α-type and 10 different β-type expressed proteasomal subunits reported here may not represent the complete gene family in *A. thaliana* and we cannot exclude the existence of more isoforms. The heterogeneity between the different isoforms is weak and consists only of highly related subunits which have probably arisen from recent gene duplication and mutation events. The most related subunit isoforms are Prc6b_At and Prc6c_At (with only one nucleotide difference) and the most distant subunits are Prcga_At and Prcgb_At (93% identity and 95% similarity between the primary sequences of the proteins). Secondary structure predictions according to Chou and Fasman [37], indicate that the few amino acid substitutions between the different isoforms do not change the general structure of the proteins except at the very C-terminal end of the Prc2a_At and Prc2b_At subunits (data not shown).

In yeast, all subunits except one α-type subunit (Prc9_Sc) are essential for viability as demonstrated by gene disruption experiments, suggesting that the different α- or β-type subunits cannot complement each other. It is noteworthy that while deletion of the yeast Prc2_Sc subunit is lethal, deletion of the *Arabidopsis* counterpart does not affect the viability of the radiation-induced mutant *tt3* (M218) and thus does not seem to be essential for plant survival [24]. The existence of close relatives (Prc2b_At and Prc2c_At) in this subfamily may explain the non-lethality of the *Arabidopsis* mutant. However, if there is a complementation by the highly related isoform, it implicates its constitutive housekeeping expression. It will be of interest to determine if the different redundant proteasomal genes are co-regulated or are controlled by distinct regulatory mechanisms.

N-terminal sequencing of the mature β-type subunits revealed a mechanism of N-terminal processing [38,39]. Structural as well as biochemical studies [10,40] have shown that the catalytic active site is the hydroxyl group of the N-terminal threonine residue. This residue is present, after the autocatalytic removal of the propeptide sequence, in the three

Table 1

Locus	Subunit type	Mol. mass (kDa)	(p*I*)	Glycosylation	Phosphorylation	Nuclear localization signals (NLS)
Prc2_At	α-type	30.4	(4.99)	1×ASN_glycos	5×PKC_PS; 4×CK2_PS	–
Prc6a_At	α-type	27.3	(6.86)	1×ASN_glycos	1×cAMP_PS; 6×PKC_PS; 3×CK2_PS; 1×Tyr_PS	–
Prc3_At	α-type	25.7	(5.53)	1×ASN_glycos	1×cAMP_PS; 6×PKC_PS; 3×CK2_PS; 1×Tyr_PS	NLS (RKSRK)
Prc9_At	α-type	27.5	(6.60)	–	1×cAMP_PS; 5×PKC_PS; 3×CK2_PS	–
Prc1_At	α-type	nd		–	2×PKC_PS; 4×CK2_PS	NLS (KKMK)
Prc8_At	α-type	27.2	(5.77)	1×ASN_glycos	5×PKC_PS; 5×CK2_PS; 2×Tyr_PS	–
Prcfa_At	β-type	29.5	(6.66)	–	4×PKC_PS; 6×CK2_PS; 1×Tyr_PS	–
Prc5_At	β-type	24.6	(6.95)	–	1×PKC_PS; 4×CK2_PS	–
Prct_At	β-type	22.6	(5.31)	1×ASN_glycos	1×PKC_PS; 4×CK2_PS	–
Prcga_At	β-type	21.9	(6.21)	2×ASN_glycos	1×PKC_PS; 4×CK2_PS	–
Prcgb_At	β-type	22.5	(5.95)	2×ASN_glycos	2×PKC_PS; 4×CK2_PS	–
Prcd_At	β-type	25.1	(5.31)	1×ASN_glycos	2×PKC_PS; 4×CK2_PS	–
Prce_At	β-type	31.4	(6.66)	–	1×cAMP_PS; 4×PKC_PS; 5×CK2_PS; 2×Tyr_PS	2×NLS (KKAK and KKKK)

The values and motifs indicated in the table were given for the proteasomal subunits for which the protein sequences are believed to be full length, and for Prc1_At, where a few amino acids are missing. The motifs were determined by using the 'PROSITE' database [27] except for the nuclear localization signals. Abbreviations: ASN-glycos, *N*-glycosylation site; cAMP_PS, cAMP and cGMP-dependent protein kinase phosphorylation site; PKC_PS, protein kinase C phosphorylation site; CK2_PS, casein kinase II phosphorylation site; Tyr_PS, tyrosine kinase phosphorylation site; NLS, were only considered short stretch of basic amino acids (SV 40-like nuclear localization signal); nd, not determined (in case of incomplete protein sequences).

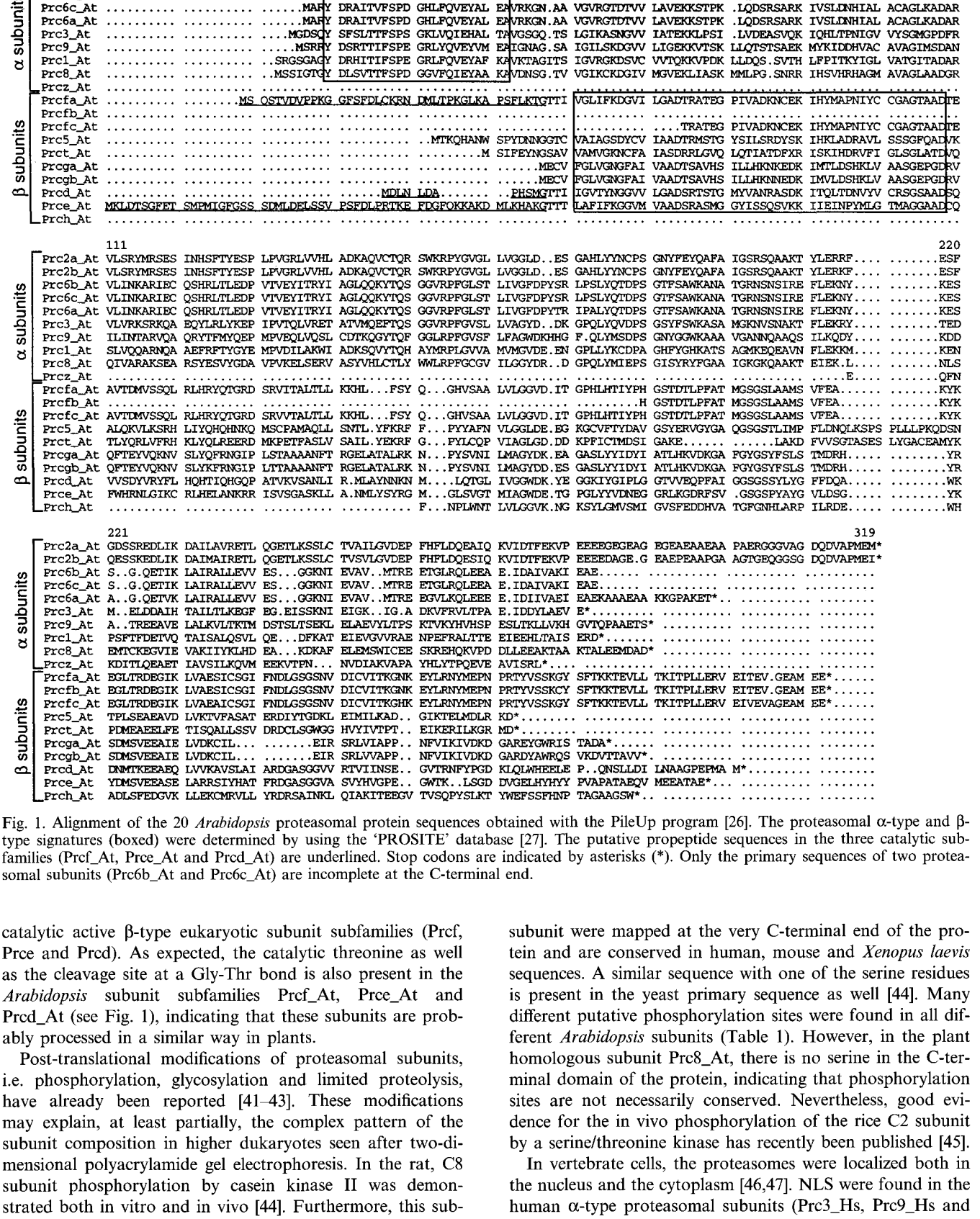

Fig. 1. Alignment of the 20 *Arabidopsis* proteasomal protein sequences obtained with the PileUp program [26]. The proteasomal α-type and β-type signatures (boxed) were determined by using the 'PROSITE' database [27]. The putative propeptide sequences in the three catalytic subfamilies (Prcf_At, Prce_At and Prcd_At) are underlined. Stop codons are indicated by asterisks (*). Only the primary sequences of two proteasomal subunits (Prc6b_At and Prc6c_At) are incomplete at the C-terminal end.

catalytic active β-type eukaryotic subunit subfamilies (Prcf, Prce and Prcd). As expected, the catalytic threonine as well as the cleavage site at a Gly-Thr bond is also present in the *Arabidopsis* subunit subfamilies Prcf_At, Prce_At and Prcd_At (see Fig. 1), indicating that these subunits are probably processed in a similar way in plants.

Post-translational modifications of proteasomal subunits, i.e. phosphorylation, glycosylation and limited proteolysis, have already been reported [41–43]. These modifications may explain, at least partially, the complex pattern of the subunit composition in higher dukaryotes seen after two-dimensional polyacrylamide gel electrophoresis. In the rat, C8 subunit phosphorylation by casein kinase II was demonstrated both in vitro and in vivo [44]. Furthermore, this subunit was shown to be the main phosphate acceptor of the proteasome. The two phosphorylated serine residues of this subunit were mapped at the very C-terminal end of the protein and are conserved in human, mouse and *Xenopus laevis* sequences. A similar sequence with one of the serine residues is present in the yeast primary sequence as well [44]. Many different putative phosphorylation sites were found in all different *Arabidopsis* subunits (Table 1). However, in the plant homologous subunit Prc8_At, there is no serine in the C-terminal domain of the protein, indicating that phosphorylation sites are not necessarily conserved. Nevertheless, good evidence for the in vivo phosphorylation of the rice C2 subunit by a serine/threonine kinase has recently been published [45].

In vertebrate cells, the proteasomes were localized both in the nucleus and the cytoplasm [46,47]. NLS were found in the human α-type proteasomal subunits (Prc3_Hs, Prc9_Hs and Prc1_Hs) and all three sequences are able to direct reporter molecules to the nucleus of permeabilized HeLa and 3T3 cells

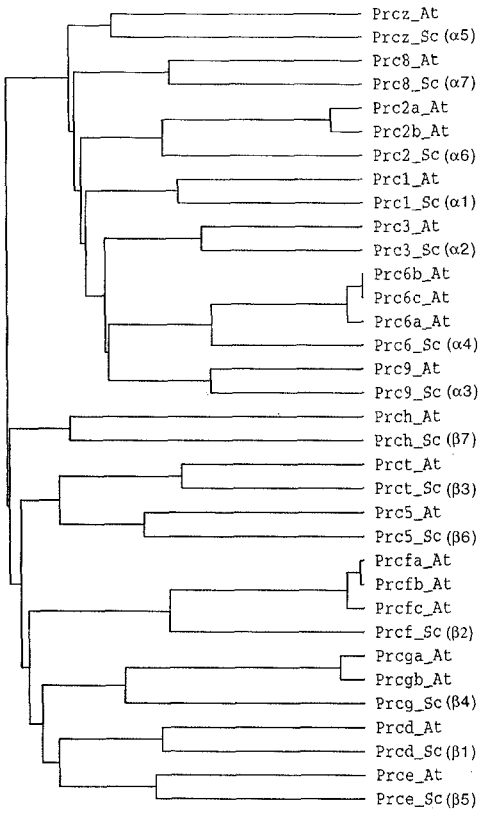

Fig. 2. Dendrogram showing the relationships among all the yeast and *Arabidopsis* 20S proteasome protein sequences. The new nomenclature of yeast proteasomal subunits according to Groll et al. [16] is indicated between brackets.

[48]. Putative NLS were also found in two *Arabidopsis* α-type subunits Prc3_At and Prc1_At and in the β-type subunit Prce_At (Table 1). Indications for both nuclear and cytoplasmic localization of proteasomes in plants have already been reported [49].

3.2. Genomic mapping of the identified proteasomal cDNAs

The mapping data clearly show a random distribution of the α-type and β-type proteasome subunit genes over the *Arabidopsis* genome (Fig. 3). Chromosome I and III show the highest accumulation of proteasome genes. Furthermore, most of the highly related isoforms are not clustered on the chromosomes, but are localized on different chromosomes. This is the case for subunit Prcga, which is located on chromosome IV, and Prcgb, which is located on chromosome III. Similarly, Prc2a is on chromosome V [24] and Prc2b on chromosome I and Prc6a is on chromosome III and Prc6b on chromosome V (Prc6c may be localized in the same region as Prc6b). Only isoforms of the proteasomal gene subfamily Prcf (at least Prcfa and Prcfb) seem to be localized in the same chromosomal region.

Acknowledgements: The authors gratefully acknowledge financial support from the Centre National de la Recherche Scientifique, the Groupement de Recherche (GDR): grant no. 1003, the Groupe de Recherche et d'Etude des Génomes (GREG): grant no. 31-94 and the European Community, the European Scientist Sequencing Arabidopsis project (ESSA): grant no. BIO2-CT93-0075. We would like to thank Thomas Newman and the *Arabidopsis* Biological Resource Center at Ohio State University (Columbus, OH) for providing the *Arabidopsis* EST clones. We also thank Kenneth Richards for critical reading of the manuscript.

References

[1] Hochstrasser, M. (1995) Curr. Opin. Cell Biol. 7, 215–223.
[2] Murray, A. (1995) Cell 81, 149–152.
[3] Dawson, S.P., Arnold, J.E., Mayer, N.J., Reynolds, S.E., Billett, M.A., Gordon, C., Colleaux, L., Kloetzel, P.M., Tanaka, K. and Mayer, R.J. (1995) J. Biol. Chem. 270, 1850–1858.
[4] Sadoul, R., Fernandez, P.-A., Quiquerez, A.-L., Martinou, I., Maki, M., Schröter, M., Becherer, J.D., Irmler, M., Tschopp, J. and Martinou, J.-C. (1996) EMBO J. 15, 3845–3852.
[5] Grimm, L.M., Goldberg, A.L., Poirier, G.G., Schwartz, L.M. and Osborne, B.A. (1996) EMBO J. 15, 3835–3844.

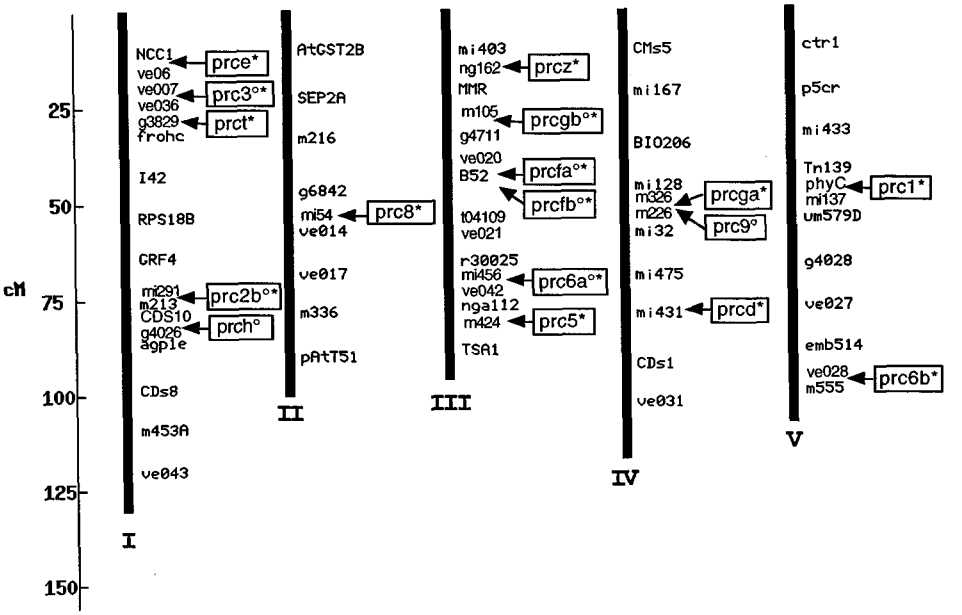

Fig. 3. Genomic locations of the different *Arabidopsis* proteasome subunits: °, determined by restriction fragment length polymorphisms (RFLP); *, determined by physical mapping. Prc2a is localised on chromosome V (not indicated here) as established by Shirley and Goodman [24].

[6] Coux, O., Tanaka, K. and Goldberg, A.L. (1996) Annu. Rev. Biochem. 65, 801–847.
[7] Palombella, V.J., Rando, O.J., Goldberg, A.L. and Maniatis, T. (1994) Cell 78, 773–785.
[8] Alkalay, I., Yaron, A., Hatzubai, A., Orian, A., Ciechanover, A. and Ben-Neriah, Y. (1995) Proc. Natl. Acad. Sci. USA 92, 10599–10603.
[9] Murakami, Y., Matsufuji, S., Kameji, T., Hayashi, S., Igarashi, K., Tamura, T., Tanaka, K. and Ichihara, A. (1992) Nature 360, 597–599.
[10] Löwe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W. and Huber, R. (1995) Science 268, 533–539.
[11] Zwickl, P., Grziwa, A., Pühler, G., Dahlmann, B., Lottspeich, F. and Baumeister, W. (1992) Biochemistry 31, 964–972.
[12] Heinemeyer, W., Tröndle, N., Albrecht, G. and Wolf, D.H. (1994) Biochemistry 33, 12229–12237.
[13] Hilt, W. and Wolf, D.H. (1995) Mol. Biol. Rep. 21, 3–10.
[14] Kopp, F., Dahlmann, B. and Hendril, K.B. (1993) J. Mol. Biol. 229, 14–19.
[15] Koster, A.J., Walz, J., Lupas, A. and Baumeister, W. (1995) Mol. Biol. Rep. 21, 11–20.
[16] Groll, M., Ditzel, L., Löwe, J., Stock, D., Bochtler, M., Bartunik, H.D. and Huber, R. (1997) Nature 386, 463–471.
[17] Arrigo, A.P., Tanaka, K., Goldberg, A.L. and Welch, W.J. (1988) Nature 331, 192–194.
[18] Schliephacke, M., Kremp, A., Schmid, H.-P., Kohler, K. and Kull, U. (1991) Eur. J. Cell Biol. 55, 114–121.
[19] Skoda, B. and Malek, L. (1992) Plant Physiol. 99, 1515–1519.
[20] Ozaki, M., Fujinami, K., Tanaka, K., Amemiya, Y., Sato, T., Ogura, N. and Nakagawa, H. (1992) J. Biol. Chem. 267, 21678–21684.
[21] Fujinami, K., Tanahashi, N., Tanaka, K., Ichihara, A., Cejka, Z., Baumeister, W., Miyawaki, M., Sato, T. and Nakagawa, H. (1994) J. Biol. Chem. 269, 25905–25910.
[22] Genschik, P., Philipps, G., Gigot, C. and Fleck, J. (1992) FEBS Lett. 309, 311–315.
[23] Genschik, P., Jamet, E., Philipps, G., Parmentier, Y., Gigot, C. and Fleck, J. (1994) Plant J. 6, 537–546.
[24] Shirley, B.W. and Goodman, H.M. (1993) Mol. Gen. Genet. 241, 586–594.
[25] Cooke, R. et al. (1996) Plant J. 9, 101–124.
[26] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
[27] Bairoch, A. (1990) PROSITE: A Dictionary of Protein Sites and Patterns, 5th release, University of Geneva, Switzerland.
[28] Ausubel, F., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (1990) Current Protocols in Molecular Biology, Greene Publishing, New York.
[29] Creusot, F. et al. (1995) Plant J. 8, 763–770.
[30] Lister, C. and Dean, C. (1993) Plant J. 4, 745–750.
[31] Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Day, M.J., Lincoln, S.E. and Newberg, L. (1987) Genomics 1, 174–181.
[32] Zachgo, E.A., Wang, M.L., Dewdney, J., Bouchez, D., Camilleri, C., Belmonte, S., Huang, L., Dolan, M. and Goodman, H.M. (1996) Genome Res. 6, 19–25.
[33] Schmidt, R., West, J., Love, K., Lenehan, Z., Lister, C., Bouchez, D. and Dean, C. (1995) Science 270, 480–483.
[34] Schmidt, R., West, J., Cnops, G., Love, K., Balestrazzi, A. and Dean, C. (1996) Plant J. 9, 755–765.
[35] Schmidt, R., Love, K., West, J., Lenehan, Z. and Dean, C. (1997) Plant J. 11, 563–573.
[36] Heinemeyer, W., Kleinschmidt, J.A., Saidowsky, J., Escher, C. and Wolf, D.H. (1991) EMBO J. 10, 555–562.
[37] Chou, G. and Fasman, G.D. (1978) in: Advances in Enzymology and Related Areas of Molecular Biology (M. Alton, Ed.), Vol. 47, pp. 45–147, John Wiley, New York.
[38] Lee, L.W., Moomaw, C.R., Orth, K., McGuire, M.J., DeMartino, G.N. and Slaughter, C.A. (1990) Biochim. Biophys. Acta 1037, 178–185.
[39] Chen, P. and Hochstrasser, M. (1995) EMBO J. 14, 2620–2630.
[40] Seemüller, E., Lupas, A., Stock, D., Löwe, J., Huber, R. and Baumeister, W. (1995) Science 268, 579–582.
[41] Haass, C. and Kloetzel, P.M. (1989) Exp. Cell. Res. 180, 243–252.
[42] Pereira, M.E. and Wilk, S. (1990) Arch. Biochem. Biophys. 283, 68–74.
[43] Arribas, J., Arizti, P. and Castaño, J.G. (1994) J. Biol. Chem. 269, 12858–12864.
[44] Castaño, J.G., Mahillo, E., Arizti, P. and Arribas, J. (1996) Biochemistry 35, 3782–3789.
[45] Umeda, M., Manabe, Y. and Uchimiya, H. (1997) FEBS Lett. 403, 313–317.
[46] Rivett, A.J., Palmer, A. and Knecht, E. (1992) J. Histochem. Cytochem. 40, 1165–1172.
[47] Peters, J.M., Franke, W.W., Kleinschmidt, J.A. and Baumeister, W. (1994) J. Biol. Chem. 269, 7709–7718.
[48] Nederlof, P.M., Wang, H.-R. and Baumeister, W. (1995) Proc. Natl. Acad. Sci. USA 92, 12060–12064.
[49] Vallon, U. and Kull, U. (1994) Protoplasma 182, 15–18.