Protein Engineering vol.14 no.9 pp.615–631, 2001

Pectin degrading glycoside hydrolases of family 28: sequence-structural features, specificities and evolution

Oskar Markovič[1] and Štefan Janeček[2,3]

[1]Institute of Chemistry, Slovak Academy of Sciences, Dúbravská cesta 9, SK-84238 Bratislava and [2]Institute of Molecular Biology, Slovak Academy of Sciences, Dúbravská cesta 21, SK-84251 Bratislava, Slovakia

[3]To whom correspondence should be addressed.
E-mail: umikstef@savba.sk

Family 28 belongs to the largest families of glycoside hydrolases. It covers several enzyme specificities of bacterial, fungal, plant and insect origins. This study deals with all available amino acid sequences of family 28 members. First, it focuses on the detailed analysis of 115 sequences of polygalacturonases yielding their evolutionary tree. The large data set allowed modification of some of the existing family 28 sequence characteristics and to draw the sequence features specific for bacterial and fungal exopolygalacturonases discriminating them from the endo-polygalacturonases. The evolutionary tree reflects both the taxonomy and specificity so that bacterial, fungal and plant enzymes form their own clusters, the endo- and exo-mode of action being respected, too. The only insect (animal) representative is most related to fungal endopolygalacturonases. The present study brings further: (i) the analysis of available rhamnogalacturonase sequences; (ii) the elucidation of relatedness between the recently added member, the endo-xylogalacturonan hydrolase and the rest of the family; and (iii) revealing the sequence features characteristic of the individual enzyme specificities and the evolutionary relationships within the entire family 28. The disulfides common for the individual enzyme groups were also proposed. With regard to functionally important residues of polygalacturonases, xylogalacturonan hydrolase possesses all of them, while the rhamnogalacturonases, known to lack the histidine residue (His223; *Aspergillus niger* polygalacturonase II numbering), have a further tyrosine (Tyr291) replaced by a conserved tryptophan. Evolutionarily, the xylogalacturonan hydrolase is most related to fungal exopolygalacturonases and the rhamnogalacturonases form their own cluster on the adjacent branch.

Keywords: evolution/polygalacturonase/rhamnogalacturonase/xylogalacturonan hydrolase

Introduction

Based on sequence similarities the glycoside hydrolases degrading pectin have been classified into the family 28 (Henrissat, 1991). At present, this family consists of a few enzymes, such as: (i) polygalacturonase (PG; EC 3.2.1.15) catalysing random hydrolytic cleavage of α-1,4 glycosidic bonds in pectate and other galacturonans; (ii) exopolygalacturonase (EPG; EC 3.2.1.67) catalysing the hydrolytic cleavage of one galacturonic acid residue from the non-reducing end of galacturonan; (iii) exo-poly-α-galacturonosidase (EPGD; EC 3.2.1.82) catalysing the hydrolytic cleavage of two galacturonic acid residues from the non-reducing end of galacturonan; (iv) rhamnogalacturonase (RG; EC 3.2.1.-) catalysing the hydrolytic cleavage of α-1,2 glycosidic bonds between D-galacturonic acid and L-rhamnose; and (v) endo-xylogalacturonan hydrolase (XGH; EC 3.2.1.-) catalysing random hydrolytic cleavage of the glycosidic bond between D-galacturonic acid and L-xylose. All the family 28 members act with an inverting mechanism (Henrissat and Davies, 1997).

Pectin, as a heteropolysaccharide, is a major constituent of the middle lamella of primary cell walls of dicotyledonous plants, composed of alternating homogalacturonan-smooth and rhamnogalacturonan-hairy regions (Williamson *et al.*, 1998; Van der Vlugt-Bergmans *et al.*, 2000). The smooth regions are polymers of α-1,4-linked D-galacturonic acid units, partially esterified, which are split by PGs, EPGs, EPGDs, pectin lyases, pectate lyases and de-esterified by pectin methylesterases. The hairy regions consist of three different subunits, as identified in apples (Schols and Voragen, 1996): (i) subunit I is xylogalacturonan, i.e. a galacturonan backbone heavily substituted with xylose (degraded by XGH); (ii) subunit II is a short section of rhamnogalacturonan backbone with many arabinan, galactan or arabinogalactan side-chains (degraded by arabinases and galactanases); and (iii) subunit III is rhamnogalacturonan consisting of alternating rhamnose and galacturonic acid residues (degraded by RGs).

More than 100 amino acid sequences of the family 28 glycoside hydrolases are available from GenBank (Benson *et al.*, 2000) and SWISS-PROT (Bairoch and Apweiler, 2000) sequence databases. Three three-dimensional structures have already been determined, those for the RG A from *Aspergillus aculeatus* (Petersen *et al.*, 1997), PG A from *Erwinia carotovora* ssp. *carotovora* (Pickersgill *et al.*, 1998) and PG II from *Aspergillus niger* (Van Santen *et al.*, 1999). All adopt the so-called parallel β-helix structural domain first observed in pectate lyase C (Yoder *et al.*, 1993). This domain is a characteristic fold for a larger protein family of right-handed parallel β-helix proteins and may consist of 7–12 coils forming either three or four parallel β-sheets (Yoder and Jurnak, 1995; Jenkins *et al.*, 1998).

Comparison of the three-dimensional structure of PG with that of RG enabled the similarities and differences in their active sites to be found (Pickersgill *et al.*, 1998), which should be applicable also for the other members of family 28. The similarities in the presumed active sites as well as the overall structural similarity confirm the original classification of PGs and RGs into one sequence-based family (Henrissat, 1991) despite their very low sequence identity (about 15%).

Several comparisons of amino acid sequences of bacterial, fungal and plant polygalacturonases were performed, but in most cases either a limited number of various enzymes were used for the comparison or attention was focused only on their isolated, best conserved sequence segments (Scott-Craig *et al.*, 1990; Bussink *et al.*, 1991; Ruttkowski *et al.*, 1991; Tebbutt *et al.*, 1994; Kester *et al.*, 1996; Petersen *et al.*, 1996; Tenberge

Table I. The enzymes used in the present study

Abbreviation[a]	Source	EC	SwissProt	GenBank
Bacterial endopolygalacturonases				
Agrtu.pg	*Agrobacterium tumefaciens*	3.2.1.15	P27644	M62814
Agrvi.pg	*Agrobacterium vitis*	3.2.1.15	P77818	U73161
Burce.pg	*Burkholderia cepacia*	3.2.1.15	P94293	U85788
Erwca1.pg	*Erwinia carotovora*	3.2.1.15	P18192	X52944
Erwca2.pg	*Erwinia carotovora*	3.2.1.15	P26509	X51701
Ralso.pg	*Ralstonia solanacearum*	3.2.1.15	P20041	M33692
Thtma.pg	*Thermotoga maritima*	3.2.1.15[b]	Q9WYR8	AE001722
Bacterial exopolygalacturonosidases				
Erwch.epgd	*Erwinia chrysanthemi*	3.2.1.82	P15922	M31308
Ralso.epgd	*Ralstonia solanacearum*	3.2.1.82	Q53241	U60106
Theth.epgd	*Thermoanaerobacterium thermosulfurigenes*	3.2.1.82[b]	Q60045	U50951
Yeren.epgd	*Yersinia enterocolitica*	3.2.1.82[b]	O68975	AF059505
Fungal endopolygalacturonases				
Aspac.pg	*Aspergillus aculeatus*	3.2.1.15	O74213	AF054893
AspflA.pg	*Aspergillus flavus*	3.2.1.15	P41749	U05015
AspflB.pg	*Aspergillus flavus*	3.2.1.15	P41750	U05020
AspniA.pg	*Aspergillus niger*	3.2.1.15	Q9P4W4	Y18804
AspniB.pg	*Aspergillus niger*	3.2.1.15	Q9P4W3	Y18805
AspniC.pg	*Aspergillus niger*	3.2.1.15	Q12554	X64356
AspniD.pg	*Aspergillus niger*	3.2.1.15	Q9P4W2	Y18806
AspniE.pg	*Aspergillus niger*	3.2.1.15	O42809	Y14386
Aspni1.pg	*Aspergillus niger*	3.2.1.15	P26213	X58892
Aspni2.pg	*Aspergillus niger*	3.2.1.15	P26214	X58893
Aspor.pg	*Aspergillus oryzae*	3.2.1.15	P35335	D14282
Asppa.pg	*Aspergillus parasiticus*	3.2.1.15	P49575	L23523
Asptu.pg	*Aspergillus tubigensis*	3.2.1.15	P19805	X58894
Botfu1.pg	*Botryotinia fuckeliana*	3.2.1.15	O94100	U68715
Botfu2.pg	*Botryotinia fuckeliana*	3.2.1.15	Q9Y7V7	U68716
Botfu3.pg	*Botryotinia fuckeliana*	3.2.1.15	Q9Y7V9	U68717
Botfu4.pg	*Botryotinia fuckeliana*	3.2.1.15	Q9Y7W0	U68719
Botfu5.pg	*Botryotinia fuckeliana*	3.2.1.15	Q9Y7W1	U68721
Botfu6.pg	*Botryotinia fuckeliana*	3.2.1.15	Q9Y7W2	U68722
Chopu.pg	*Chondrostereum purpureum*	3.2.1.15	P79074	D45072
Clapu1.pg	*Claviceps purpurea*	3.2.1.15	P78607	Y10165
Clapu2.pg	*Claviceps purpurea*	3.2.1.15	P78608	Y10165
Cocca.pg	*Cochliobolus carbonum*	3.2.1.15	P26215	M55979
Colli1.pg	*Colletotrichum lindemuthianum*	3.2.1.15	Q00446	X89370
Colli2.pg	*Colletotrichum lindemuthianum*	3.2.1.15	O00104	X95457
Crnpa.pg	*Cryptonectria parasitica*	3.2.1.15	Q12593	U49710
Fusmo.pg	*Fusarium moniliforme*	3.2.1.15	Q07181	L02239
Fusox1.pg	*Fusarium oxysporum*	3.2.1.15	O13466	AB000124
Fusox2.pg	*Fusarium oxysporum*	3.2.1.15	O74244	AF078156
Geokl.pg	*Geotrichum klebahnii*	3.2.1.15	P87217	D89650
Kluma.pg	*Kluyveromyces marxianus*	3.2.1.15	O13478	AJ000076
Ophno.pg	*Ophiostoma novoulmi*	3.2.1.15	O59934	AF052061
Penex.pg	*Penicillium expansum*	3.2.1.15	O59925	AF047713
Pengr1.pg	*Penicillium griseoroseum*	3.2.1.15	O93883	AF085238
Pengr2.pg	*Penicilium griseoroseum*	3.2.1.15	Q9UR16	AF195113
Penja.pg	*Penicillium janthinellum*	3.2.1.15	O42824	D79980
Penol1.pg	*Penicillium olsonii*	3.2.1.15	Q9Y834	AJ243521
Penol2.pg	*Penicillium olsonii*	3.2.1.15	Q9Y833	AJ243522
Sacce.pg	*Saccharomyces cerevisiae*	3.2.1.15	P47180	AB011818
Sclsc1.pg	*Sclerotinia sclerotiorum*	3.2.1.15	Q12708	L12023
Sclsc2.pg	*Sclerotinia sclerotiorum*	3.2.1.15	Q11134	L29040
Sclsc3.pg	*Sclerotinia sclerotiorum*	3.2.1.15	Q11135	L29041
Sclsc5.pg	*Sclerotinia sclerotiorum*	3.2.1.15	P87213	Y13669
Fungal exopolygalacturonases				
Asptu.epg	*Aspergillus tubigensis*	3.2.1.67	Q00293	X99795
Botfu.epg	*Botryotinia fuckeliana*	3.2.1.67	Q9UVU0	AF145229
Cocca.epg	*Cochliobolus carbonum*	3.2.1.67	Q00359	L48982
Fusox.epg	*Fusarium oxysporum*	3.2.1.67	O74255	AF083075
Plant endopolygalacturonases:				
Actde.pg	*Actinidia deliciosa*	3.2.1.15	P35336	L12019
Arath1.pg	*Arabidopsis thaliana*	3.2.1.15[b]	O22817	AC002333
Arath2.pg	*Arabidopsis thaliana*	3.2.1.15[b]	O22818	AC002333
Arath3.pg	*Arabidopsis thaliana*	3.2.1.15[b]	O04474	AC001229
Arath4.pg	*Arabidopsis thaliana*	3.2.1.15[b]	O22816	AC004005
Arath5.pg	*Arabidopsis thaliana*	3.2.1.15[b]	O80559	AC004005

Table I. continued

Abbreviation[a]	Source	EC	SwissProt	GenBank
Arath6.pg	*Arabidopsis thaliana*	3.2.1.15	O23147	AJ002532
Arath7.pg	*Arabidopsis thaliana*	3.2.1.15[b]	O22935	AC002339
Arath8.pg	*Arabidopsis thaliana*	3.2.1.15[b]	O48576	AC002342
Arath9.pg	*Arabidopsis thaliana*	3.2.1.15[b]	O48577	AC002342
Arath10.pg	*Arabidopsis thaliana*	3.2.1.15[b]	O81798	AL031135
Arath11.pg	*Arabidopsis thaliana*	3.2.1.15[b]	Q38958	X98130
Arath12.pg	*Arabidopsis thaliana*	3.2.1.15[b]	O22699	AC002292
Brana1.pg	*Brassica napus* (PG35-8)	3.2.1.15	Q42399	X95800
Brana2.pg	*Brassica napus* (SAC66)	3.2.1.15	Q42636	Z49971
Cucme1.pg	*Cucumis melo*	3.2.1.15	O81244	AF062465
Cucme2.pg	*Cucumis melo*	3.2.1.15	O81245	AF062466
Cucme3.pg	*Cucumis melo*	3.2.1.15	O81246	AF062467
Glyma1.pg	*Glycine max*	3.2.1.15	Q9SWS3	AF128266
Glyma2.pg	*Glycine max*	3.2.1.15	Q9SWS2	AF128267
LycesA.pg	*Lycopersicon esculentum (fruit; PG2A)*	*3.2.1.15*	*P05117*	*X04583*
Lyces1.pg	*Lycopersicon esculentum* (TAPG1)	3.2.1.15	O22311	AF001000
Lyces2.pg	*Lycopersicon esculentum* (TAPG2)	3.2.1.15	Q96487	AF001001
Lyces3.pg	*Lycopersicon esculentum* (TAPG3)	3.2.1.15	O22310	AF000999
Lyces4.pg	*Lycopersicon esculentum* (TAPG4)	3.2.1.15	Q96488	U70481
Lyces5.pg	*Lycopersicon esculentum* (TAPG5)	3.2.1.15	O22313	AF001003
Lyces6.pg	*Lycopersicon esculentum* (TAPG6)	3.2.1.15	O22610	AF029230
Maldo.pg	*Malus domestica*	3.2.1.15	P48978	L27743
Medsa.pg	*Medicago sativa* (MSPG3)	3.2.1.15	O82019	Y11118
Peram.pg	*Persea americana*	3.2.1.15	Q02096	X66426
Prupe1.pg	*Prunus persica* (PRF5)	3.2.1.15	P48979	X76735
Prupe2.pg	*Prunus persica* (Feicheng)	3.2.1.15	O82586	AF095577
Prupe3.pg	*Prunus persica* (genomic)	3.2.1.15	Q43063	X77231
Rubid.pg	*Rubus idaeus* (fragment)	3.2.1.15	O65886	AJ224147
Plant exopolygalacturonases				
Arath1.epg	*Arabidopsis thaliana* (flower)	3.2.1.67	P49063	X72292
Arath2.epg	*Arabidopsis thaliana* (flower)	3.2.1.67	O65401	X73222
Arath3.epg	*Arabidopsis thaliana* (flower)	3.2.1.67[b]	O65905	AJ003135
Arath4.epg	*Arabidopsis thaliana* (flower)	3.2.1.67	P49062	X72291
Arath5.epg	*Arabidopsis thaliana* (flower)	3.2.1.67[b]	O48729	AC002505
Arath6.epg	*Arabidopsis thaliana*	3.2.1.67[b]	O49721	AL021713
Arath7.epg	*Arabidopsis thaliana*	3.2.1.67[b]	O49319	AC002334
Plant pollen polygalacturonases				
Cryja.pp	*Cryptomeria japonica*	3.2.1.15	P43212	D29772
Gosba.pp	*Gossypium barbadense*	3.2.1.15	Q39766	U09805
Goshi.pp	*Gossypium hirsutum*	3.2.1.15	Q39786	U09717
Medsa.pp	*Medicago sativa* (P73)	3.2.1.15	Q40312	U20431
Nicta.pp	*Nicotiana tabacum*	3.2.1.15	Q05967	X71020
Salgi1.pp	*Salix gilgiana* (flower)	3.2.1.15	Q9MBC0	AB029457
Salgi2.pp	*Salix gilgiana* (flower)	3.2.1.15	Q9MBB9	AB029458
Salgi3.pp	*Salix gilgiana* (flower)	3.2.1.15	Q9MBB8	AB029459
Salgi4.pp	*Salix gilgiana* (flower)	3.2.1.15	Q9MBB7	AB029460
Brana.pep	*Brassica napus* (Sta 44-4)	3.2.1.67[b]	P35337	L19879
Oenor.pep	*Oenothera organensis*	3.2.1.67	P24548	Not available
Phlpr.pep	*Phleum pratense*	3.2.1.67[b]	Q9XG86	AJ238848
Zeama1.pep	*Zea mays*	3.2.1.67	P26216	X57627
Zeama2.pep	*Zea mays*	3.2.1.67	P35338	X64408
Zeama3.pep	*Zea mays*	3.2.1.67	P35339	X66422
Insect polygalacturonase				
Phaco.pg	*Phaedon cochleariae* (mustard beetle)	3.2.1.15	O97400	Y17906
Rhamnogalacturonases				
AspacA.rg	*Aspergillus aculeatus*	3.2.1.-	Q00001	L35499
AspniA.rg	*Aspergillus niger*	3.2.1.-	P87160	X94220
AspniB.rg	*Aspergillus niger*	3.2.1.-	P87161	X94221
Botfu.rg	*Botryotinia fuckeliana*	3.2.1.-	P87247	U62397
Xylogalacturonan hydrolase				
AsptuA.xgh	*Aspergillus tubingensis*	3.2.1.-	Q9UUZ2	AJ249460

[a]The terminations of abbreviations for the individual enzyme specificities are as follows: pg, endopolygalacturonase; epg, exopolygalacturonase; epgd, exopolygalacturonosidase; pp, pollen polygalacturonase; pep, pollen exopolygalacturonase; rg, rhamnogalacturonase; xgh, xylogalacturonan hydrolase.
[b]For these members the EC number was assigned only based on sequence similarities.

et al., 1996; Huang and Allen, 1997; Hadfield *et al.*, 1998; Stratilová *et al.*, 1998; Gognies *et al.*, 1999; Wubben *et al.*, 1999; Torki *et al.*, 2000). Thus, the regions comprising the residues 178_NTD, 201_DD, 222_GHG and 256_RIK (unless otherwise specified, all amino acid numbering throughout the text corresponds to the open reading frame of *A.niger* PG II) (Bussink *et al.*, 1990) have been found to be strictly conserved in all PGs, EPGs and EPGDs, with the Asp180, Asp201, Asp202, His223, Arg256 and Lys258 being probably involved in their active site (Rexová-Benková and Mračková, 1978; Pickersgill *et al.*, 1998; Van Santen *et al.*, 1999; Armand *et al.*, 2000). However, in RGs these regions are 193_glD, 215_De, 237_sgG and 269_miK (*A.aculeatus* RG mature enzyme numbering) (Kofod *et al.*, 1994); the His223 of polygalacturonase is replaced by Gly238 of RG.

Moreover, Stratilová *et al.* (Stratilová *et al.*, 1996) have described the potential role for a tyrosine residue in the function of a PG by chemical modification and spectrophotometric titration. The eventual position of this tyrosine was proposed by comparison of 36 different PG sequences which revealed a strictly conserved tyrosine residue equivalent to Tyr291 of *A.niger* PG II (Stratilová *et al.*, 1998).

Despite all these partial findings a serious deep analysis of all available amino acid sequences of the glycoside hydrolase family 28 is still lacking. Moreover, the family 28 is a quickly growing family of enzymes (more than 100 members). The need for this study may be supported also by running genome projects that yield numerous sequences of putative proteins having similarity to those of the family 28. Therefore, the aim of the present study was to: (i) compare as many as possible amino acid sequences of the members of this family of glycoside hydrolases; (ii) describe their evolutionary relationships in detail; and (iii) reveal their similarities and differences that would allow one to discriminate between them.

Materials and methods

The enzymes belonging to the glycoside hydrolase family 28 involved in the present study are listed in Table I. The listing for this family provided by the CAZy web-server (April 2000) (Coutinho and Henrissat, 2000) served as a base. The following enzyme specificities are represented: polygalacturonase; exopolygalacturonase; exopolygalacturonosidase; rhamnogalacturonase; and endoxylogalacturonan hydrolase. The sequences were retrieved from the GenBank (Benson *et al*., 2000) and SWISS-PROT (Bairoch and Apweiler, 2000) sequence databases.

All sequence alignments were performed using the program CLUSTAL W (Thompson *et al.*, 1994) and then manually tuned where applicable. In some cases the hydrophobic cluster analysis method (Gaboriaud *et al.*, 1987; Callebaut *et al.*, 1997) was applied in order to detect or support weaker sequence similarities. The method used for building the evolutionary trees was the neighbour-joining method (Saitou and Nei, 1987). The Phylip format tree output was applied using the bootstrapping procedure (Felsenstein, 1985); the number of bootstrap trials used was 1000. The trees were drawn with the program TreeView (Page, 1996). The BLAST tool (Altschul *et al.*, 1990) was also used for sequence similarity searches. Three-dimensional structure modelling was performed using the SWISS-MODEL automated protein modelling server (Guex and Peitsch, 1997; Guex *et al.*, 1999) according to the instructions given there (http://www.expasy.ch/swissmod/). The experimentally determined three-dimensional structures were retrieved from the Protein Data Bank (Berman *et al.*, 2000). The protein structures were displayed by the program WebLabViewerLite (Molecular Simulations, Inc.)

Results and discussion

Conserved sequence regions and invariant residues of polygalacturonases

In this study 115 amino acid sequences of PGs, EPGs and EPGDs were compared. One sequence represents the insects (animals), the one from the phytophagous mustard beetle *Phaedon cochleariae* (Girard and Jouanin, 1999), whereas all the others belong to bacteria, fungi and plants (Table I) that form the three main groups. No polygalacturonase from archaeal origin is known. The group of plant enzymes covers 19 members of a gene family encoding PGs and EPGs in *Arabidopsis thaliana* (Torki *et al.*, 2000).

The amino acid sequence alignment of all 115 PG, EPG and EPGD sequences (Figure 1; the colour version of Figure 1 can be found at the URL: http://nic.savba.sk/~umikstef/PGs) confirmed that there are four strictly conserved sequence segments with one invariantly conserved residue, Tyr291, as recognized previously (e.g. Stratilová *et al.*, 1998). Remarkably there was only one further amino acid residue, Gly228, strictly conserved in all these enzymes (Pavenicová, 2000). This glycine is positioned close to the C-terminus of the third conserved region (222_GHG). Structurally, it is located in the seventh turn between the β-sheets PB1 and PB2a of *A.niger* PG II (Van Santen *et al.*, 1999); however, as yet no function has been assigned to it.

The first segment, 178_NTD, consists of two strictly conserved residues, Asn which is substituted in one PG from *Penicillium griseoroseum* (Pengr2.pg) by His and the totally conserved Asp, and the almost invariantly conserved (93.0%) Thr179 in the middle. The threonine was replaced in all bacterial EPGDs by either Gly or Ala, and in four fungal enzymes (Botfu3.pg, Botfu6.pg, Fusox1.pg and Fusox.epg) by serine. Moreover, the enzyme from *Thermotoga maritima*, declared in the GenBank as a putative EPGD, contains the substitution Gly→Asn which is not in accordance with the rest of the bacterial EPGDs. However, as will be shown later, this protein also lacks the other sequence features characteristic of bacterial EPGDs and goes well with the bacterial PGs (Figure 1).

The second segment, 201_DD, is exclusively conserved in all polygalacturonases with specific amino acid residues neighbouring at both sides of this dipeptide. All bacterial and plant enzymes together with most of fungal EPGs have the glycine at the N-terminal side of 201_DD, while all fungal PGs (including the PG from insect) contain a glutamine in that position. Concerning the C-terminal side of the second segment, there is a cysteine residue conserved in all plant and fungal polygalacturonases, the insect PG as well as the two bacterial enzymes (Agrtu.pg and Yeren.epgd).

In the third segment, 222_GHG, there is an almost invariantly conserved (94.8%) Gly222 followed by two totally conserved His and Gly. The former glycine was replaced in one of the two fungal PGs from *Colletotrichum lindemuthianum* and all four known fungal EPGs by serine as well as in the enzyme from *T.maritima*. The fourth conserved segment, 256_RIK, contains a highly conserved (87.0%) Ile257 in addition to the strictly conserved Lys and Arg which is replaced by His in the insect PG. The isoleucine was not conserved in two

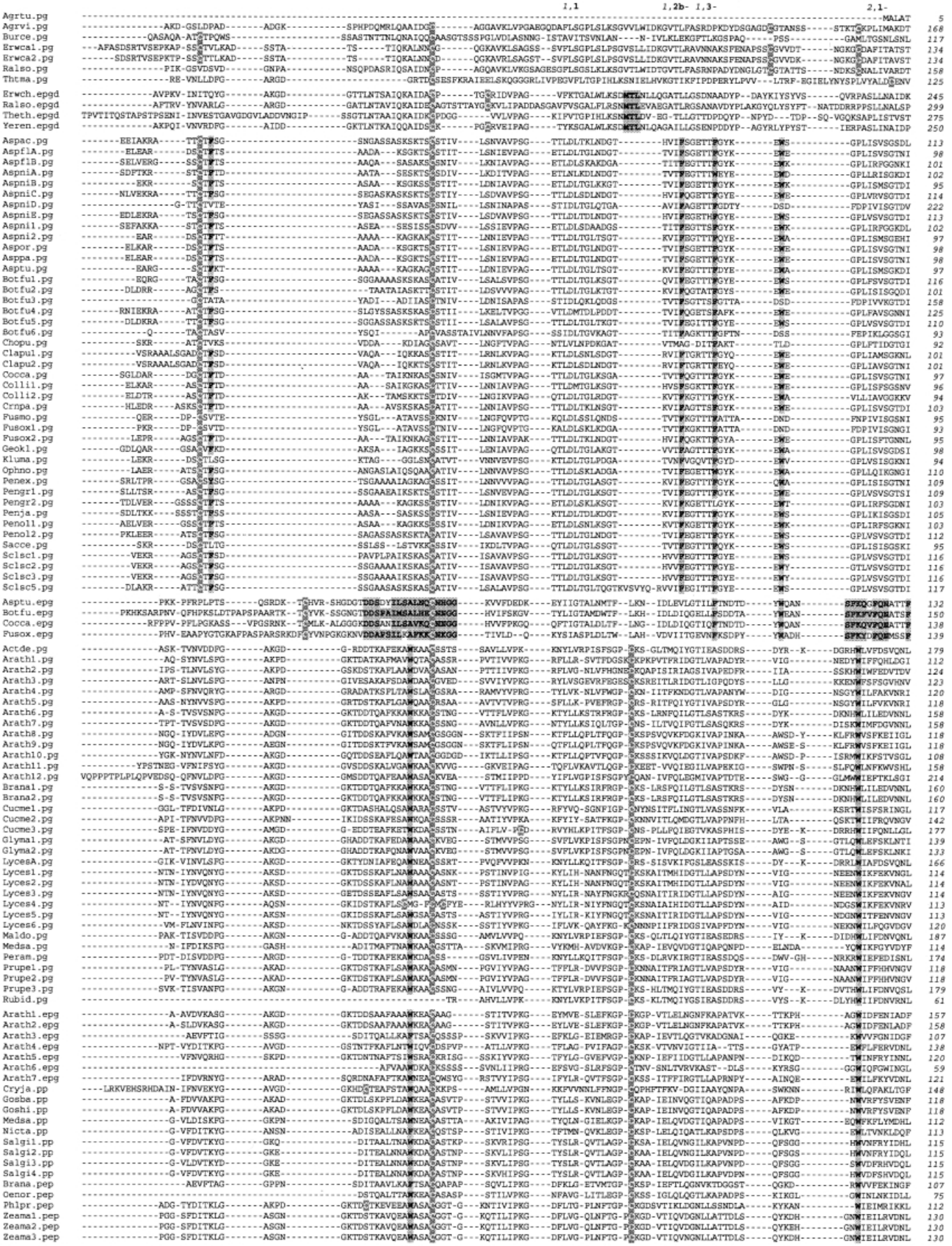

Fig. 1. Legend on page 624.

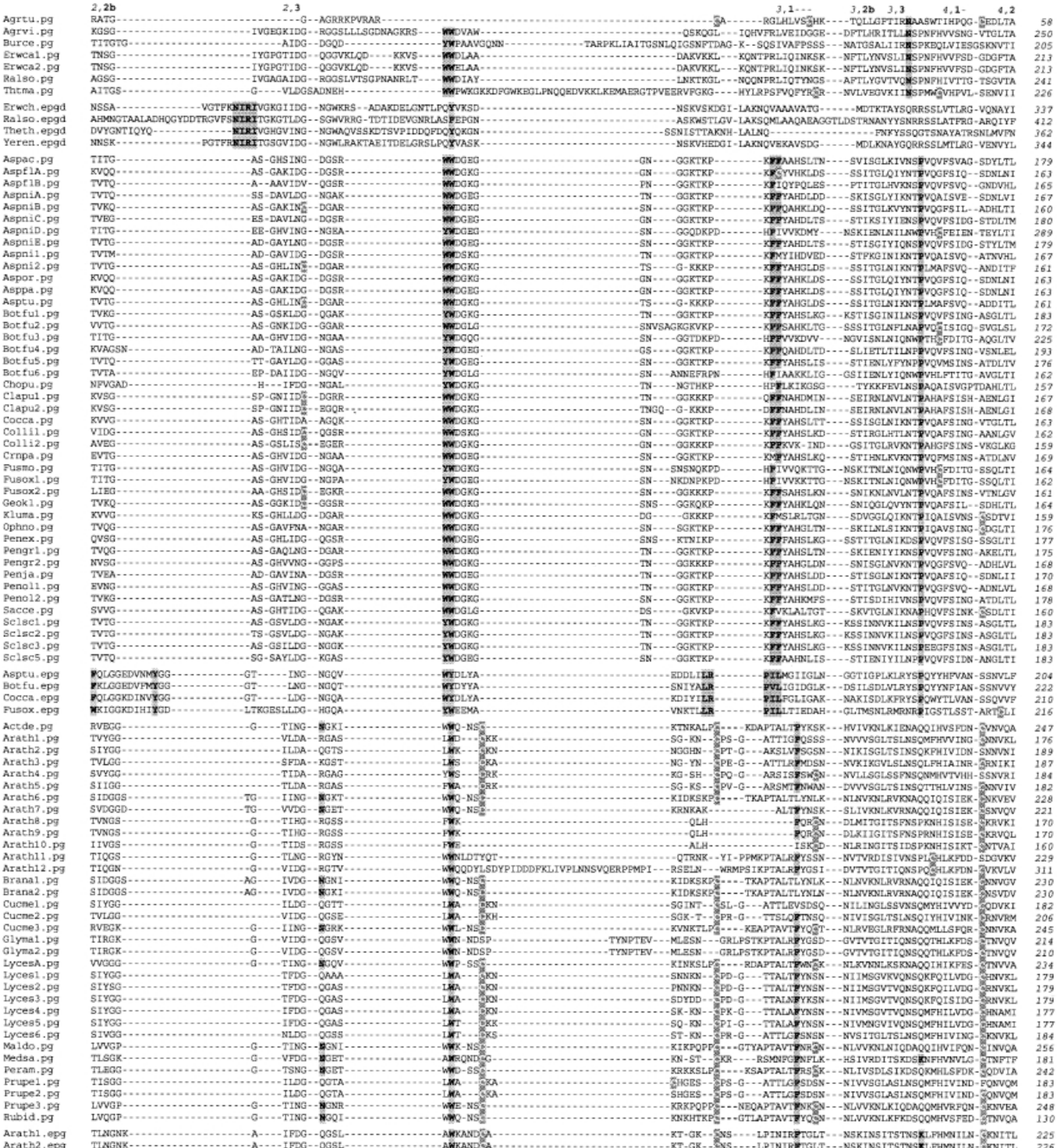

Fig. 1. Continued.

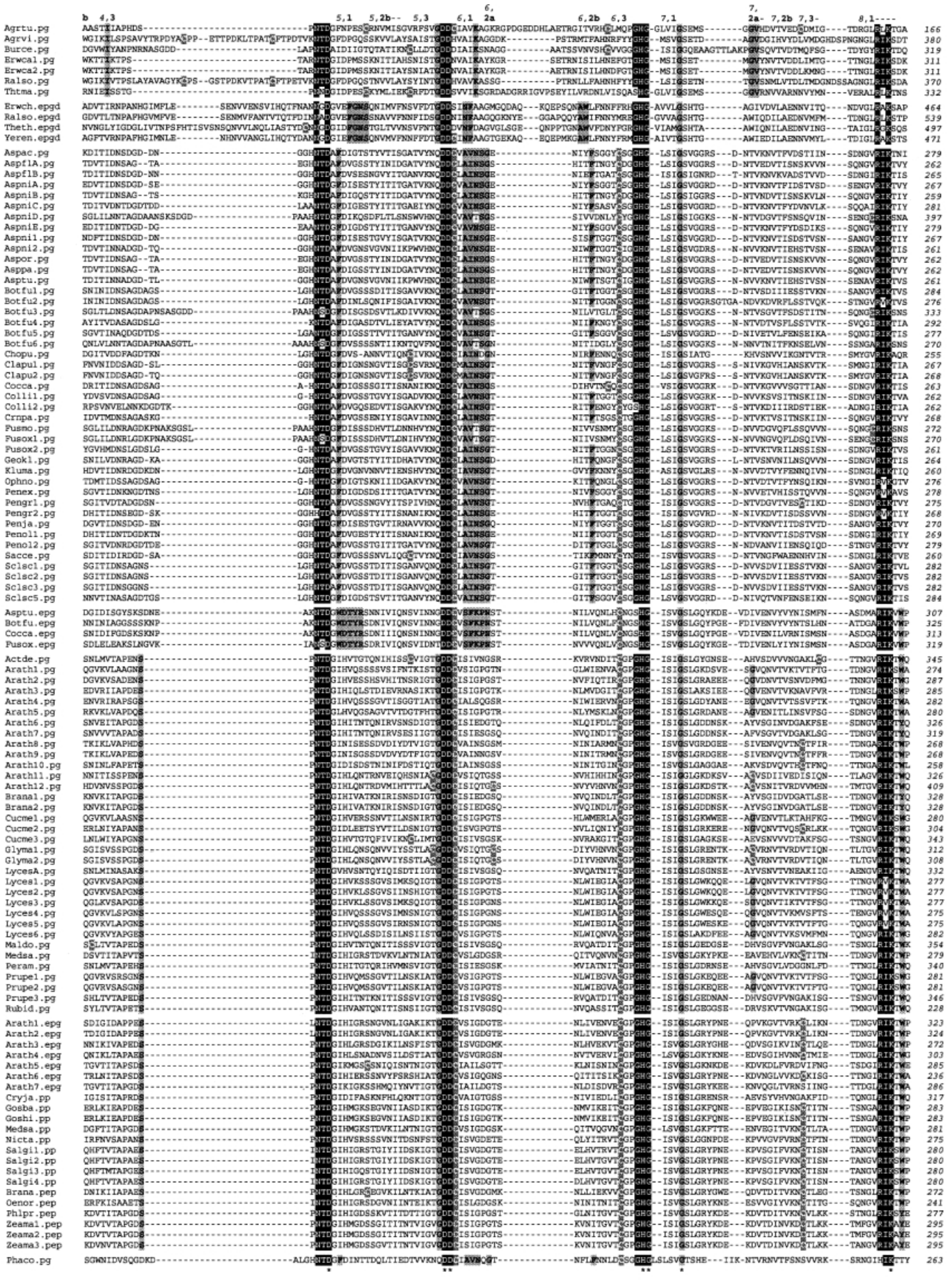

Fig. 1. Continued.

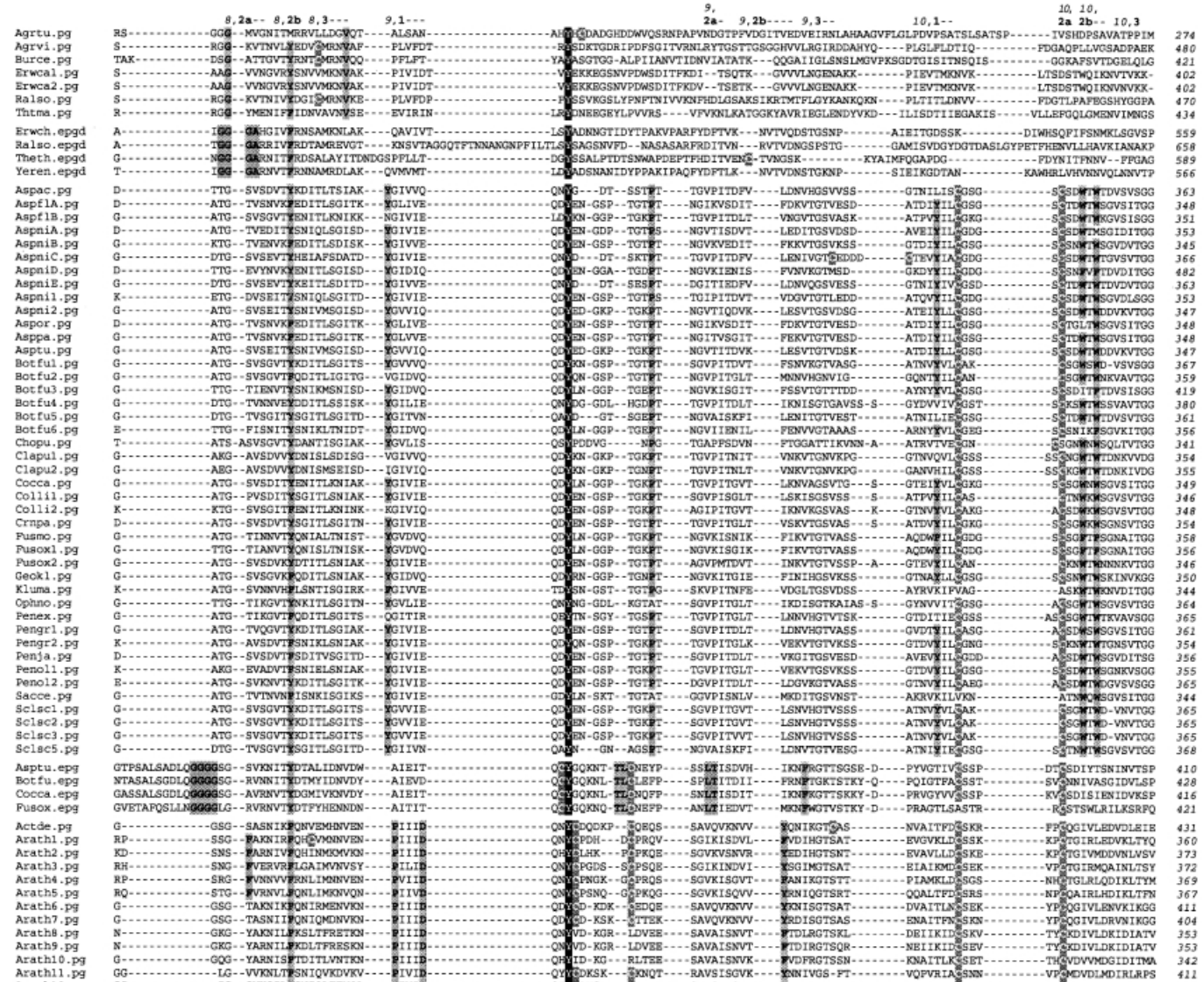

Fig. 1. Continued.

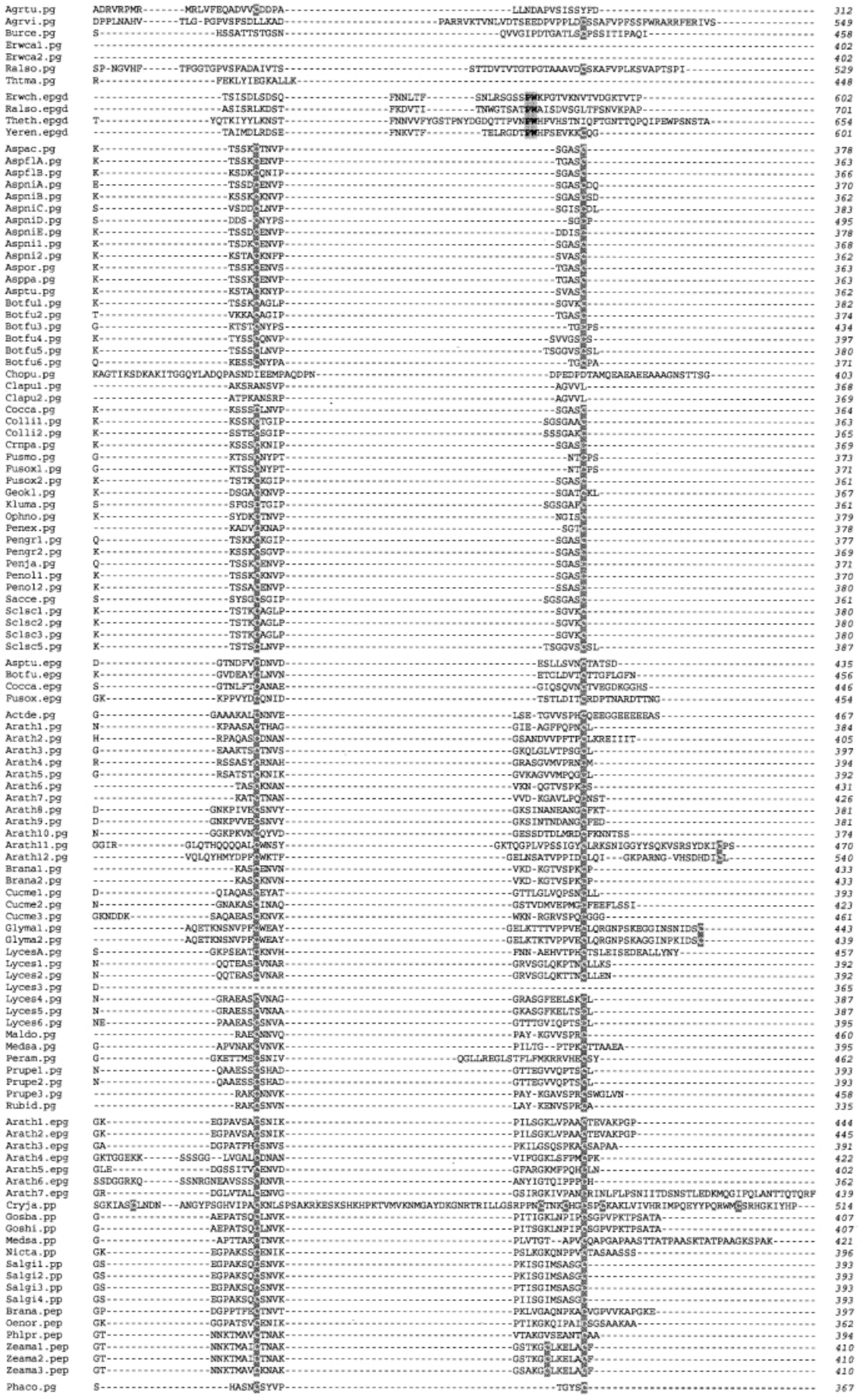

Fig. 1. Continued.

bacterial PGs (Agrtu.pg and Thtma.pg; Ile→Leu), all bacterial EPGDs (by Ala, Gly and Leu), four fungal PGs (Botfu2.pg, Ophno.pg, Penex.pg and Pengr2.pg; Ile→Val) and five plant PGs (tomato abscission zone; Ile→Val).

With regard to the Tyr291 found previously to be conserved in polygalacturonases (Stratilová *et al.*, 1998), the alignment of all 115 PGs, EPGs and EPGDs known at present confirmed that this tyrosine belongs to the invariantly conserved residues of these enzymes (Figure 1).

Bacterial PGs and EPGDs

The group of bacterial PGs contains the sequences with a very low degree of mutual similarity. The seven PGs involved in the present study (Table I) exhibit only 4.7 and 11.2% identity and similarity, respectively, except for the two PGs from *E.carotovora* that share 96.0% sequence identity (Hinton *et al.*, 1990; Saarilahti *et al.*, 1990). Pair-wise similarity varies in the range between 11 and 50%, but in most cases does not reach 20%. The bacterial PG sequences have only 10 invariant residues (Asn197, Ile218, Gly248, Lys255, Gly282, Gly315, Val330 and the dipeptide 287_GV; Erwca2.pg numbering) in addition to the four well recognized conserved regions and presumably functional invariant tyrosine (Figure 1). It is worth mentioning that in *Ralstonia solanacearum* EPGD, this tyrosine was identified only with help of the hydrophobic cluster analysis method (data not shown) due to an inserted oligopeptide segment (Huang and Allen, 1997) absent in all other polygalacturonases (Figure 1).

On the other hand, the overall sequence similarity among bacterial EPGDs is higher (23.7% identity and 37.3% similarity), the pair-wise similarity ranging from 25 to 60%. These sequences contain not only isolated invariant residues but also several identical segments that have been found to be characteristic of bacterial EPGDs only (Figure 1). The most important segments are 202_MTL, 255_NIRI, 378_FGNS, 399_NF, 417_AW, 467_GGGA and 584_PW (Erwch.epgd numbering).

Fungal PGs and EPGs and the insect PG

Forty-three fungal PGs form a substantial part of the entire set of polygalaturonases studied in this work (Table I). Their sequences are 8.9% identical and 17.4% similar with the average pair-wise similarity of about 60% ranging from lower than 20% to higher than 90%. In the amino acid sequence alignment (Figure 1) several aromatic residues can be found as characteristic of these fungal PGs: Phe32, Phe74, Phe80, Trp85, Trp114, Trp115, Phe128, Phe129, Phe182, Phe214, Tyr272, Tyr283, Tyr326, Trp337 and Trp339 (Aspni2.pg numbering). Not all of them are conserved strictly (except of Trp115 and Phe182), but in most cases there are conservative (aromatic→aromatic) substitutions. Of these Phe80, Trp85 and Tyr272 (or their correspondences) are present also in the fungal EPGs, and the equivalents of Trp115 (though replaced by tyrosines) may be found in the whole set of sequences shown in Figure 1. There is also one longer segment specific for fungal PGs (205_AinSG) positioned from the C-terminal side close to the active-site dipeptide 201_DD. Among the other conserved residues the two prolines, Pro148 (conserved also in the fungal EPGs) and Pro300 (not strictly conserved), could be of interest.

Fungal EPGs exhibit quite high degrees of identity (36.7%) and similarity (51.0%) with several longer conserved stretches, characteristic dipeptides and isolated invariant residues. In the N-terminal part, there is a segment starting with 60_DD and ending with 75_GG (*Aspergillus tubigensis* EPG numbering). The following regions are also typical for the EPGs from fungi: 121_SFKxxFQN, 166_LRPiL, 225_WDTYR, 248_SFKPN, 319_GGGG. In addition, Phe132, Phe133, Tyr143, Trp306, Tyr331 and Phe374 (although not invariantly conserved in all cases) as well as the two dipeptides, 355_TL and 364_LT (located in the C-terminal part), should be of importance for fungal EPGs.

With regard to the one representative of animal polygalacturonases, the insect PG from *P.cochleariae*, its sequence goes well with the fungal PGs and contains almost all sequence features characteristic of this group of PGs (Figure 1). There are 20 single residues or short segments conserved in fungal PGs that are identical with those from the insect PG.

Plant PGs and EPGs

In the group of plant polygalacturonases analysed in this study, the exact enzyme specificity has not been determined strictly due to the fact that many of them were not biochemically characterized in detail or were taken as putative proteins from sequencing the whole genome. It is not possible to say clearly in all cases whether the enzyme is a PG or EPG. There are (Figure 1) again several well conserved aromatic amino acid residues, such as Trp99 (LycesA.pg numbering), Trp157, Phe203, Trp331, Phe343 and Tyr382, as well as the other residues Ser245, Gly250, Gly269 (strictly conserved also in all bacterial PGs and EPGDs and almost in all fungal EPGs), Pro354 and Asp358 (substituted in three cases by Asn).

In agreement with the proposed classification system (Hadfield and Bennett, 1998; Torki *et al.*, 2000) the present set of 56 plant polygalacturonases can be divided into five clades: A, B, C, D, E plus the gymnosperm PG from *Cryptomeria japonica* (cedar) with a sequence without resemblance to the rest (except for the conserved sequence regions covering the active-site residues discussed above). This division is based on the evolutionary tree (Figure 2) of all polygalacturonases (Table I) reflecting the alignment shown in Figure 1. The tree will be discussed later.

The characterization of the clades and the numbering of the residues according to the consensus alignment of Torki *et al.* is used here (Torki *et al.*, 2000). For clade A, there are two exclusively specific, invariantly conserved residues Gly264 and Phe294. For clade B, there is also a characteristic Asn104 present in the Medsa.pg (from clade C), which thus exhibits an intermediary nature of clades B and C. Clade C, covering all pollen and flower PGs and plant EPGs, contains the invariant Lys176. Clades D and E are without exclusively conserved residues.

Fig. 1. Amino acid sequence alignment of all polygalacturonases. The colour version of this figure can be found at the URL: http://nic.savba.sk/~umikstef/ PGs. The abbreviations of enzyme sources are given in Table I. The 115 sequences of polygalacturonases are ordered according to their groups (from the top): bacterial PGs, bacterial EPGDs, fungal PGs, fungal EPGs, plant PGs, plant EPGs with plant pollen polygalacturonases and the insect PG. All selected residues are signified by bold. Four conserved active-site segments (178_NTD, 201_DD, 222_GHG, 256_RIK; Aspni2.pg numbering) and the invariant tyrosine (Tyr291) are highlighted by black-and-white inversion. Cysteines are coloured white and highlighted in dark grey. The residues characteristic for the individual group of polygalacturonases are highlighted in light grey. The invariant residues are signified by asterisks. The β-strands forming the 10 coils of parallel β-helix (four parallel β-sheets PB1, PB2a, PB2b and PB3) of the PG from *A.niger* (van Santen *et al.*, 1999) are indicated above the alignment blocks (the number of the coil is written in italics, while the number of the β-sheet is written in bold).


Colli2.pg
Cocca.pg
Crnpa.pg
Colli1.pg
Fusox2.pg
Geokl.pg
AspniB.pg
Asppa.pg
AspflA.pg
Aspor.pg
Aspni2.pg
Asptu.pg
Pengr2.pg
Penol1.pg
AspflB.pg
Penja.pg
AspniA.pg
Aspni1.pg
Clapu1.pg
Clapu2.pg
Botfu2.pg
Penex.pg
Pengr1.pg
Penol2.pg
Botfu1.pg
Sclsc1.pg
Sclsc2.pg
Sclsc3.pg
Botfu4.pg
Ophno.pg
Botfu5.pg
Sclsc5.pg
Aspac.pg
AspniC.pg
AspniE.pg
Insect PG
Kluma.pg
Sacce.pg
Phaco.pg
Botfu6.pg
AspniD.pg
Botfu3.pg
Fusmo.pg
Fusox1.pg
Chopu.pg
Fungal and insect PGs
Burce.pg
Agrtu.pg
Thtma.pg
Agrvi.pg
Ralso.pg
Erwca1.pg
Erwca2.pg
Bacterial PGs
Theth.epgd
Ralso.epgd
Erwch.epgd
Yeren.epgd
Bacterial EPGDs
Fusox.epg
Botfu.epg
Asptu.epg
Cocca.epg
Fungal EPGs
Cryja.pp
Plant gymnosperm PG
Arath11.pg
Arath12.pg
Glyma1.pg
Glyma2.pg
Plant clade E
Arath10.pg
Arath8.pg
Arath9.pg
Plant clade D
Arath4.pg
Arath5.pg
Arath3.pg
Arath1.pg
Arath2.pg
Plant clade A
Arath7.pg
Arath6.pg
Brana1.pg
Brana2.pg
Peram.pg
Cucme3.pg
LycesA.pg
Actde.pg
Rubid.pg
Maldo.pg
Prupe3.pg
Plant clade B
Cucme2.pg
Cucme1.pg
Lyces6.pg
Prupe1.pg
Prupe2.pg
Lyces4.pg
Lyces5.pg
Lyces3.pg
Lyces1.pg
Lyces2.pg
Plant clade A
Arath5.epg
Arath7.epg
Phipr.pep
Zeama3.pep
Zeama1.pep
Zeama2.pep
Arath4.epg
Arath6.epg
Arath1.epg
Arath2.epg
Medsa.pg
Medsa.pp
Arath3.epg
Brana.pep
Gosba.pp
Goshi.pp
Oenor.pep
Nicta.pp
Salgi1.pp
Salgi2.pp
Salgi3.pp
Salgi4.pp
Plant clade C
a

Fig. 2a.

Table II. Sequence identity and similarity for the clades of plant polygalacturonases

Clade	Members[a]	Identity (%)	Similarity (%)
A	15	18.3	35.6
B	11	25.4	35.1
C	22	12.2	23.1
D	3	53.4	66.0
E	4	35.1	51.8

[a]The only gymnosperm polygalacturonase from cedar (Cryja.pp) has not been classified in any of the above clades.

With regard to sequence identity and similarity in the frame of the individual clades, the values are shown in Table II. However, in general, the conserved sequence regions containing the active-site residues are conserved in plant polygalacturonases as longer segments (Figure 1).

Conserved cysteines and aromatic residues of polygalacturonases

As pointed out by the published crystal structures of polygalacturonases (Pickersgill *et al.*, 1998; Van Santen *et al.*, 1999) these enzymes contain some disulfide bridges stabilizing their molecules. However, the conservation of cysteines reflects taxonomy, i.e. the corresponding disulfides could be conserved only in the frames of the respective bacterial, fungal and plant groups as described above. There is only one cysteine residue conserved throughout all the polygalacturonases (Figure 1) in the position of Cys45 (Aspni2.pg).

Bacterial PG from *E.carotovora* has two S–S bridges, Cys41–Cys62 and Cys115–Cys125 (Pickersgill *et al.*, 1998), but there is no conservation of cysteines in the respective positions for all bacterial PGs. Bacterial EPGDs contain an even smaller number of cysteine residues which are also without specific arrangement (Figure 1).

On the other hand, the cysteines among fungal PGs are very well conserved. *Aspergillus niger* PG II (Van Santen *et al.*, 1999) has four disulfides: Cys30–Cys45, Cys203–Cys219, Cys329–Cys334 and Cys353–Cys362. While the first two bridges should be present in all fungal PGs, the one corresponding with Cys329–Cys334 is missing in the PGs from yeasts (Cys→Val and Cys→Ala substitutions). With regard to the fourth S–S bridge, the corresponding cysteines are absent in both PGs from *Claviceps purpurea* and the one from *Chondrostereum purpureum*. The insect PG from *P.cochleariae* has all the cysteines in accordance with those present in the group of fungal PGs. Based on the alignment shown in Figure 1 it is possible to suppose that the fungal EPGs could contain all the four disulfides present in fungal PGs. However, the position of the first cysteine from the first disulfide (Cys30–Cys45 in the fungal PGs) is shifted and corresponds to Cys50 of *A.tubigensis* EPG, and there is Cys→Ala substitution in EPG from *Fusarium oxysporum* in the position corresponding to the first cysteine of the third disulfide (Cys329; Aspni2.pg numbering). Fungal EPGs possess two additional conserved cysteines, Cys348–Cys357 (*A.tubigensis* EPG numbering), forming probably an extra disulfide bridge.

Since the three-dimensional structure of a plant polygalacturonase has still not been determined and the presence of disulfides in these enzymes has not been experimentally proved, the eventual S–S bridges can be proposed by analogy with fungal PGs only. Thus, plant enzymes could contain the three disulfides corresponding with the second, third and fourth disulfides of fungal PGs, positioned at Cys272–Cys289, Cys399–Cys405 and Cys427–Cys442, respectively (*Lycopersicon esculentum* PG A numbering). There are two exceptions, the pollen PG from *Nicotiana tabacum* with the Cys→Arg substitution in the position corresponding with the Cys399 and the one from *L.esculentum* (TAPG3) with a shorter polypeptide chain. With regard to the first disulfide present in fungal PGs, all plant enzymes contain only the second cysteine, Cys103, which corresponds to Cys45 of *A.niger* PG II. However, there is a strictly conserved cysteine residue in all plant enzymes, Cys130 (LycesA.pg numbering), which could eventually form the S–S bridge equivalent to the first disulfide of fungal PGs. Most of the plant polygalacturonases have further cysteines, Cys186, Cys194, Cys229, Cys362 and Cys368 (LycesA.pg numbering), four of which (except for the Cys229) are absent in the PGs from clade D. Clade D, on the other hand, contains an extra cysteine corresponding with Ala317 in the LycesA.pg, which was proposed as a pollen-specific cysteine residue (Tebbutt *et al.*, 1994; Petersen *et al.*, 1996). This cysteine is further present in most members of clade C and in the PGs from *Cucumis melo* (Cucme3.pg) and *Medicago sativa* (Medsa.pg). Cys186 is not present in clades D, E and in the pollen PG from *M.sativa* (Medsa.pp). This is also the case for Cys194, which is absent in two more pollen PGs, those from *Gossypium barbadense* and *Gossypium hirsutum*. Cys229 was not observed in the sequences of PGs from *A.thaliana* (Arath2.pg, Arath4.pg and Arath11.pg) and *Prunus persica* (Prupe1.pg) as well as of EPGs from *A.thaliana* (Arath3.epg, Arath4.epg, Arath5.epg and Arath7.epg) and *Brassica napus* (Brana.pep). Both Cys362 and Cys368 are present in all plant polygalacturonases except for those from clade D and Arath4.epg, thus indicating the possibility of forming a disulfide bridge.

As far as the aromatic amino acid residues are concerned, those characteristic for the individual groups of polygalacturon-

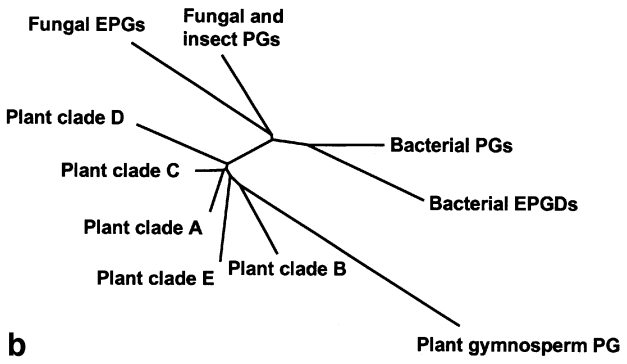

Fig. 2. Evolutionary trees of all polygalacturonases. Both trees are based on the alignment shown in Figure 1. The abbreviations of enzyme sources are given in Table I. The branch lengths are proportional to the sequence divergence. (**a**) The complete tree, calculated with involving the positions with gaps in the sequence alignment, showing the relationships among the individual taxonomic (bacteria, fungi and plants) and specificity (endo- and exo-mode of action) groups as well as in the frames of all these groups. (**b**) The simplified tree, calculated with excluding the positions with gaps in the sequence alignment, showing the basic relationships among the individual group of polygalacturonases.

ases (PGs, EPGs and EPGDs as well as bacteria, fungi and plants) were briefly described above. The interest in these residues is due to the fact that they may be involved in binding of substrate not only in polygalacturonases (Rao *et al.*, 1996) but also in other glycoside hydrolases, e.g. in amylases (Clarke and Svensson, 1984; Gibson and Svensson, 1986; Williamson *et al.*, 1997). The invariantly conserved tyrosine, Tyr291 (Stratilová *et al.*, 1993, 1998), to which also the function has been proposed (Stratilová *et al.*, 1996), was recently confirmed by site-directed mutagenesis to be indispensable for effective catalysis constituting the subsite +1 (Pagès *et al.*, 2000). There are two further aromatic positions conserved among the different groups of polygalacturonases. These are not conserved strictly but only aromatic residues (Trp, Phe and Tyr) occur there. The first one is at Trp115, which is in all bacterial EPGDs, most fungal EPGs and three plant pollen PGs replaced by tyrosine and phenylalanine (Figure 1). The only exception is the PG from *Agrobacterium tumefaciens* (with a shorter polypeptide chain) that evidently does not possess an aromatic residue equivalent to Trp115. The second aromatic position corresponds with Phe271 which alternates with tyrosine (60:40%, respectively) only. The *A.tumefaciens* PG with a methionine residue equivalent to Phe271 exhibits an exceptional behaviour again.

Evolutionary tree of all endo- and exo-polygalacturonases

The evolutionary tree showing for the first time the relationships of a complete as possible set of sequenced polygalacturonases belonging to bacteria, fungi, plants and animals (represented by an insect) is presented in Figure 2A. The tree is based on the sequence alignment shown in Figure 1 and thus reflects the sequence similarities and differences discussed above.

Basically, the tree manifests that there are three main groups: bacteria, fungi and plants (Figure 2A), bacteria being positioned, however, between the fungal PGs and fungal EPGs. The only one representative of the animal kingdom, the insect PG from *P.cochlearie*, is included in the cluster of fungal PGs. Following this basic division of the tree, one can further see the clustering according to the endo- and exo-mode of action of these enzymes especially among the bacterial and fungal polygalacturonases and into the plant clades.

While bacterial EPGDs form their own cluster, the group of bacterial PGs is not so homogeneous reflecting the lower degree of mutual sequence similarity. The two PGs from *E.carotovora* are very closely related to each other and are located next to the pair of PGs from *Agrobacterium vitis* and *Ralstonia solanacearum*. There are three further PGs located on long branches. Two of them, those from *A.tumefaciens* and *T.maritima*, have been determined as putative PGs only (Rong *et al.*, 1991; Nelson *et al.*, 1999) so that it is not possible to correlate their position in the tree with their exact enzyme specificity. The last bacterial PG from *Burkholderia cepacia* is even more isolated which may reflect the fact that this PG is a plasmid-encoded protein (Gonzalez *et al.*, 1997).

The consequence of clear sequence differences between fungal PGs and EPGs (Figure 1) is that these two groups of polygalacturonases are well separated in the evolutionary tree (Figure 2A), both forming quite homogeneous and isolated groups. The most remarkable feature of the fungal PG part of the tree is the location of the insect PG from *P.cochlearie* (representing the animal kingdom) directly among the fungal PGs. However, this is based on the resemblance of the insect sequence to the fungal ones described above. The PG from *C.purpureum* seems to be the most distantly related member of the fungal PG group (Figure 2A) in agreement with several non-conservative substitutions in characteristic positions (Figure 1). The two PGs from yeasts are both positioned adjacent to each other thus indicating that all eventual yeast PGs would form their own separate yeast cluster in the frame of all fungal PGs. Several sub-clusters or sub-groups can be found in the fungal PG part of the tree. Based on the analysis of 35 sequences of fungal PGs Wubben *et al.* (Wubben *et al.*, 1999) have proposed five monophyletic groups of closely related PGs. The present study covering more fungal PGs (43 sequences) indicates that the number of the so-called monophyletic groups is probably higher and will even rise as more sequences become available.

As mentioned in the section dealing with the sequence comparison of plant PGs and EPGs, these enzymes have been classified into the five clades (Hadfield and Bennett, 1998; Torki *et al.*, 2000) plus the PG from cedar (Table II). This division can also be seen from the tree (Figure 2A) where the plant polygalacturonases form clusters according to their clades and one long isolated branch leading to the cedar PG (Cryja.pp), which is the only plant gymnosperm PG. The detailed analysis of the branch arrangement in the plant angiosperm part of the tree suggested that the 'plant' branch leads to the node separating clades E, D and B on the one side from those of C and A on the other side. The only 'exception' is the clustering of the five PGs from *A.thaliana* (Arath1.pg-Arath5.pg) from clade A together with the three further PGs from this plant that should form clade D (Torki *et al.*, 2000). The largest clade C (Table II) contains polygalacturonases expressed mainly in flower buds, flowers and pollen that are thought to encode the EPGs (Torki *et al.*, 1999, 2000) with one exception, the PG from *M.sativa* induced by a *Rhizobium* strain, which was, however, originally revealed to exhibit extremely high sequence similarity to its pollen counterpart (Muñoz *et al.*, 1998).

In order to re-analyse the plant part of the evolutionary tree with respect to the plant clades, a further tree was constructed (based on the alignment shown in Figure 1), however, excluding the positions with gaps. The simplified version of this tree is shown in Figure 2B. The detailed analysis of the branch arrangement in the plant part of the tree in this case suggested that the 'plant' branch leads to the node separating clade D (three PGs from *A.thaliana*) from the rest of the plant enzymes. The rest was then divided into well separated clades (without dividing any of them) so that clades C and A were on the one side and clades E and B were on the other side. The only gymnosperm PG from cedar was on its own long branch (adjacent to clade B) which reflected its higher dissimilarity (discussed above) with the other plant PGs which are angiosperm. With regard to bacterial and fungal parts of the tree, in the case when the positions with gaps were excluded, taxonomy was fully respected so that there were two separate branches in the tree (Figure 2B): one leading to fungal PGs and EPGs, and the other one leading to bacterial PGs and EPGDs.

Rhamnogalacturonases

The alignment of all sequenced rhamnogalacturonases is presented in Figure 3. As is well known (Coutinho and Henrissat, 2000) the RGs belong to glycoside hydrolase family 28 together with polygalacturonases although the conserved active-site regions of PGs are slightly modified in the sequences of RGs. Thus, 193_GLD (*A.aculeatus* RG numbering) corresponds

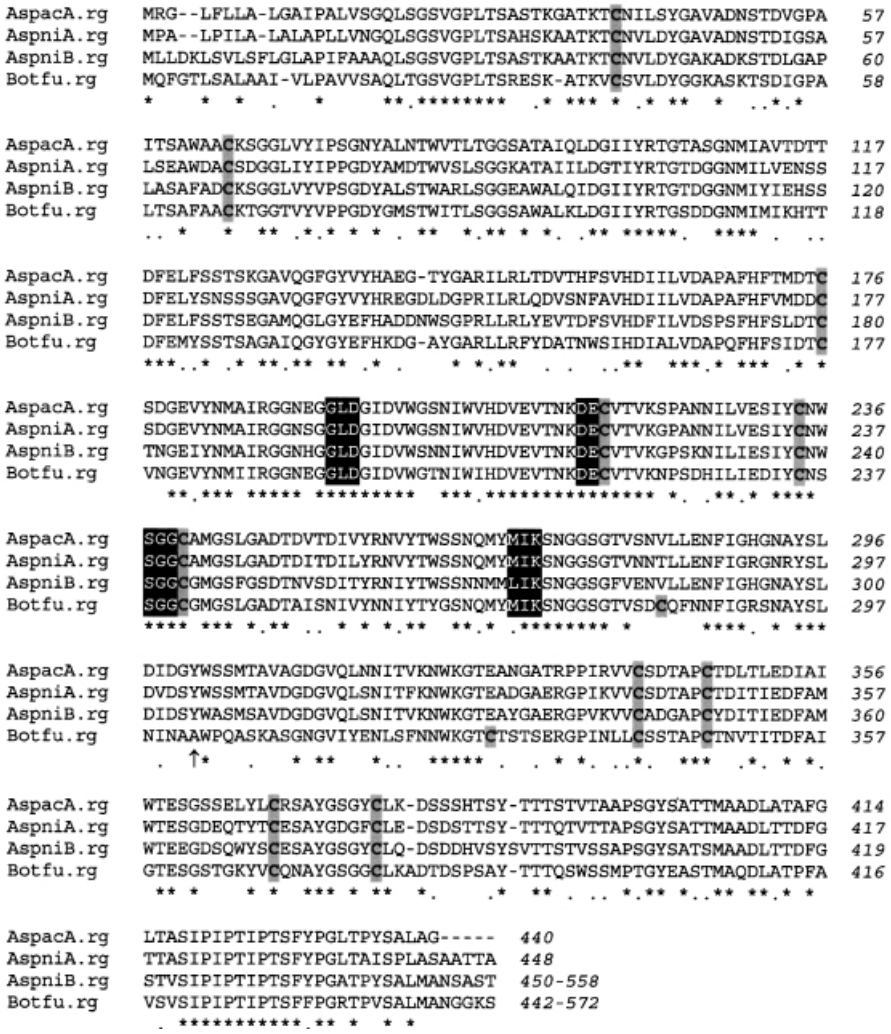

Fig. 3. Amino acid sequence alignment of rhamnogalacturonases. The abbreviations of enzyme sources are given in Table I. The asterisks and dots signify the identical amino acid residues and conservative substitutions, respectively. Gaps are indicated by dashes. Cysteines are highlighted in grey and signified by bold. The four conserved active-site segments are highlighted in black-and-white inversion. The vertical arrow marks the tyrosine position (not invariantly conserved) which could correspond with the invariant tyrosine of PGs (Tyr291 of *A.niger* PG II; cf. Figure 2).

with 178_NTD (*A.niger* PG II numbering), 215_DE with 201_DD, 237_SGG with 222_GHG, and 269_MIK with 256_RIK. It means that the most significant amendments in the sequences of RGs in comparison with polygalacturonases are the lack of His223 in the third region (SGG/GHG) and hydrophilic→hydrophobic substitution in the fourth region (MIK/RIK). With regard to the invariant tyrosine residue, Tyr291, present in polygalacturonases (Figure 1), this residue may have its equivalent in RGs (Tyr301; *A.aculeatus* RG numbering); however, it seems that there is no corresponding tyrosine in the RG from *Botryotinia fuckeliana* (Figure 3).

In general, the amino acid sequences of RGs are highly similar: they exhibit mutual 48.9% sequence identity and 57.1% sequence similarity. Ten cysteines were found to be conserved in the alignment of RGs (Figure 3). Based on the determined three-dimensional structure of the RG from *A.aculaetus* (Petersen *et al.*, 1997), all the four RGs analysed in the present study should be stabilized by four disulfide bridges (Cys39–Cys65, Cys217–Cys234, Cys340–Cys346 and Cys368–Cys377) and contain two free cysteine residues (Cys176 and Cys240). Furthermore, there are 13 tyrosines, 6 phenylalanines and 4 tryptophans invariantly conserved in all four RGs, Phe169, Trp200 and Trp302 being found present in the active site of the *A.aculeatus* RG (Petersen *et al.*, 1997). Taking into account the lack of the equivalent residue of the PG-active-site histidine, there are four invariant histidines (His138, His159, His170 and His207) in the sequences of RGs (Figure 3), however, it is possible that there is no histidine in the active site of RG (Pickersgill *et al.*, 1998). This is consistent with the site-directed mutagenesis study (Armand *et al.*, 2000) indicating that His223 is not a catalytic residue in the entire glycoside hydrolase family 28, but may play an indirect role in catalysis of polygalacturonases.

Endoxylogalacturonan hydrolase

The sequence of this new member of the glycoside hydrolase family 28 was determined only recently (Van der Vlugt-Bergmans *et al.*, 2000). In contrast to RGs, the XGH sequence exhibits better similarity in the four active-site segments to polygalacturonases (205_NTD, 228_DD, 250_SHG and 284_GIK; *A.tubigensis* XGH numbering). It has the active-site DD dipeptide as well as the His251 equivalent to His223

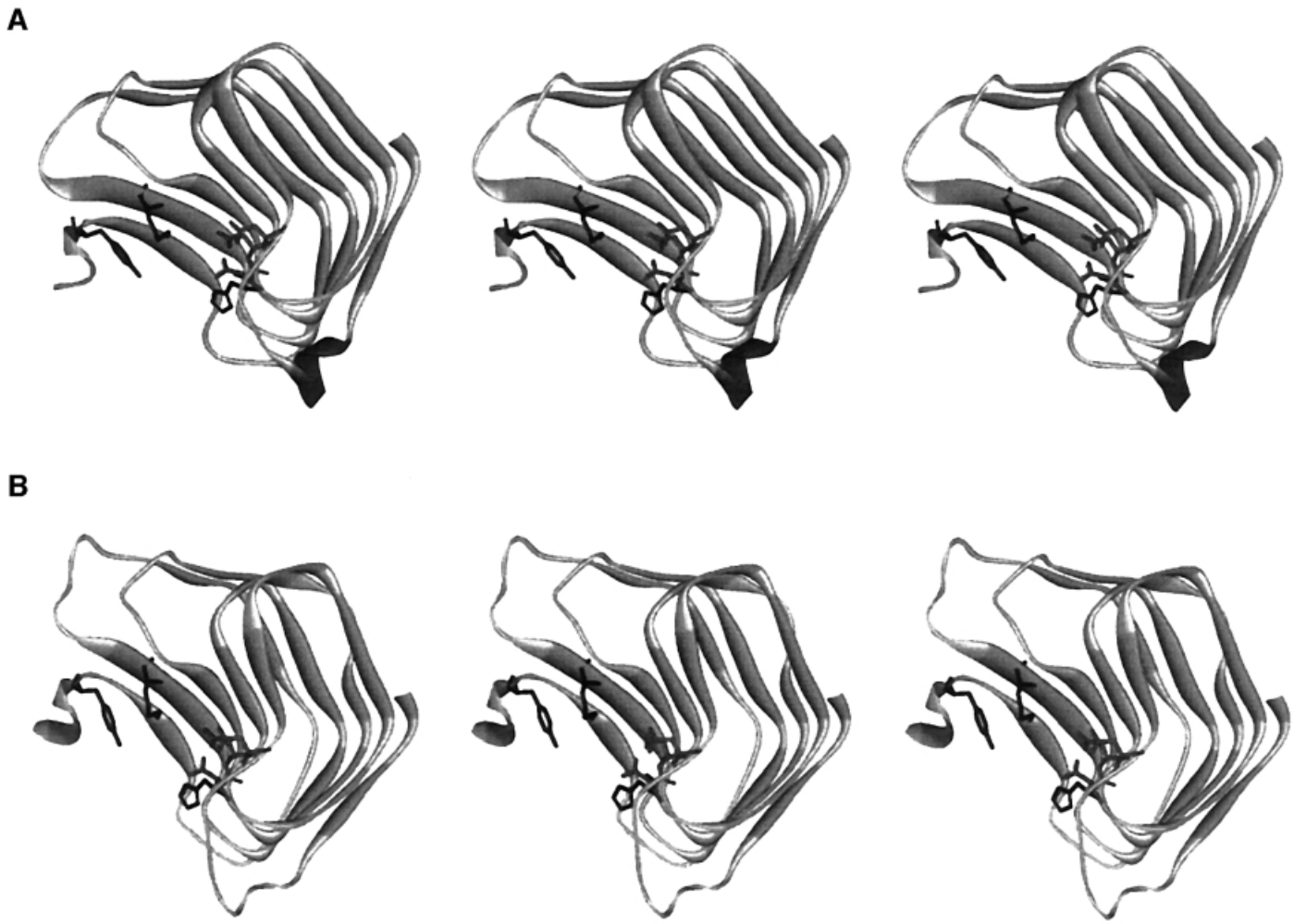

Fig. 4. Location of the functionally important residues of polygalacturonase II from *A.niger* (**A**) and xylogalacturonan hydrolase from *A.tubigensis* (**B**). The selected residues in PG (XGH): Asp180 (Asp207), Asp201 (Asp228) and Asp202 (Asp229) coloured dark grey, and His223 (His251), Lys258 (Lys286) and Tyr291 (Tyr322) coloured black. The PG structure was retrieved from the Protein Data Bank (code: 1CZF), while the structure of XGH was modelled on the SWISS-MODEL server using the PG X-ray coordinates (1CZF) as template.

of *A.niger* PG II. It also possesses the tyrosine residue corresponding with the Tyr291 present in polygalacturonases. All this can be supported by the model of the three-dimensional structure of XGH (Figure 4) constructed using the X-ray coordinates of *A.niger* PG II (van Santen *et al.*, 1999; Protein Data Bank code: 1CZF) as template.

There is a further change in the hydropathic character of the residue in position $i - 2$ with respect to the invariant lysine in the four segment (256_RIK in the *A.niger* PG II versus 284_GIK in the *A.tubigensis* XGH). The transition in this position from the hydrophilic residue in polygalacturonases (Arg or His) to the hydrophobic residue in RGs (Met or Leu) via the neutral side-chain of glycine, found in XGH, should be of interest.

In order to find the most closely related family 28 sequence to that of XGH the BLAST search was used (Altschul *et al.*, 1990). It was found that the sequence of XGH exhibits the highest similarity to that of EPG from *Cochliobolus carbonum*. These two sequences have 39.9% identity and 55.4% similarity (the alignment not shown). In general, the sequence of *A.tubigensis* XGH exhibits higher similarity to the sequences of fungal EPGs than to those of fungal PGs (Van der Vlugt-Bergmans *et al.*, 2000). Remarkably, the similarity is lower to the EPG from the same organism *A.tubigensis* (53.5%) than to the taxonomically more distantly related EPG from *C.carbonum* (55.4%). Despite this pronounced sequence similarity to fungal EPGs, the XGH sequence does not contain most of the conserved regions characteristic of fungal EPGs, thus indicating its enzymatic uniqueness.

All glycoside hydrolase family 28 enzymes

Based on the analysis of available amino acid sequences of PGs, EPGDs, EPGs, RGs and XGH discussed above, a set of sequences of the family 28 members representing all the individual groups was aligned (Figure 5). It is evident that despite the overall rather low sequence similarity, each representative contains its functionally important residues in the segments equivalent to the four conserved active-site segments of PGs (178_NTD, 201_DD, 222_GHG and 256_RIK) as well as at least the conservative substitution of the Tyr291. This makes from them a common family in the frame of all glycoside hydrolases (Coutinho and Henrissat, 2000). On the other hand, there are some important changes of the residues adjacent to the residues constituting the active site, especially in RGs. In fact there are only three strictly conserved residues in common that could be functionally important in the family

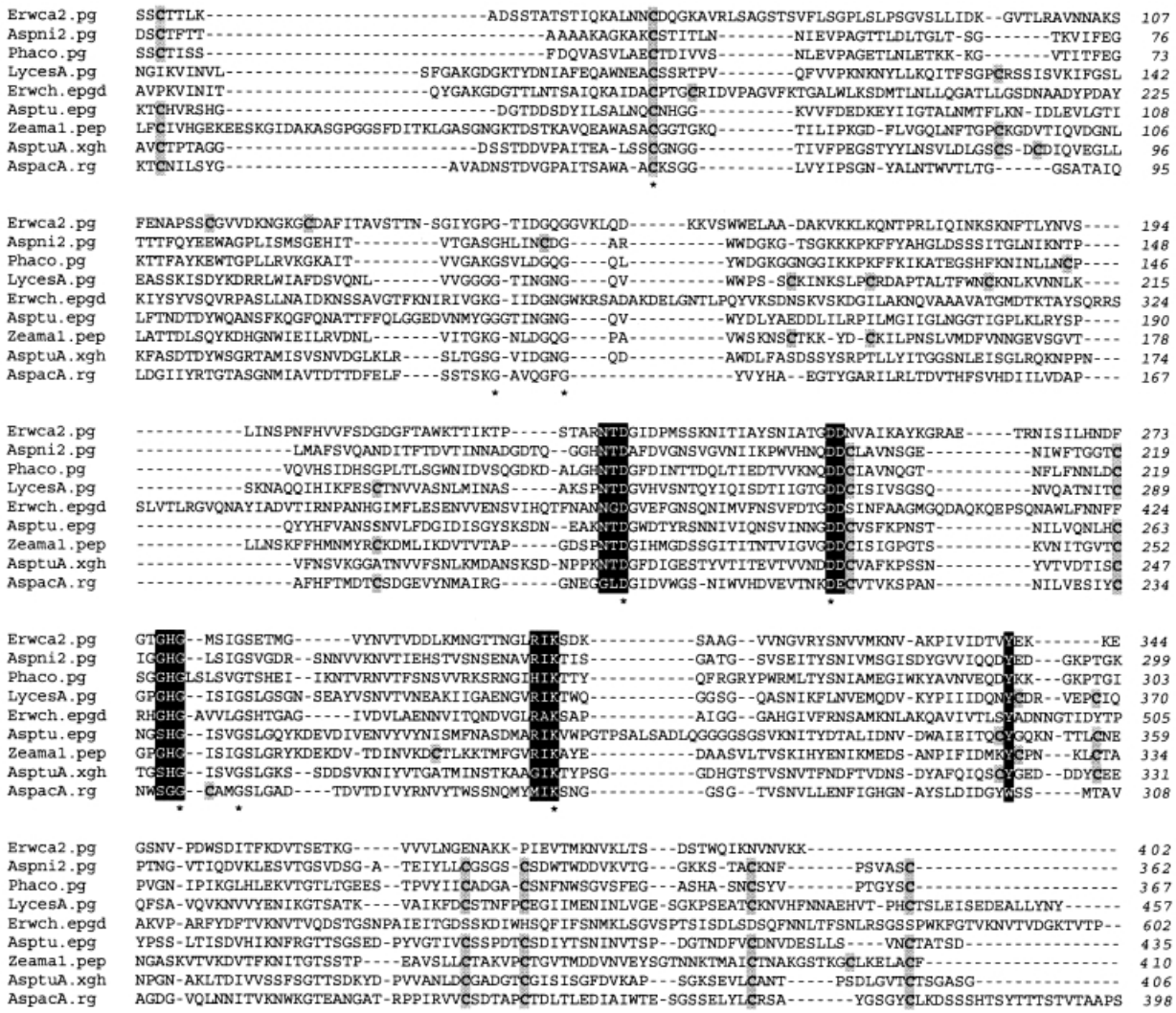

Fig. 5. Amino acid sequence alignment of representative members of glycoside hydrolase family 28. The abbreviations of enzyme sources are given in Table I. Erwca2.pg represents all bacterial PGs, Aspni2.pg all fungal PGs, LycesA.pg all plant PGs, Erwch.epg all bacterial EPGDs, Asptu.epg all fungal EPGs, Zeama1.pep all plant EPGs (including pollen PGs), AspacA.rg all RGs. Phaco.pg and AsptuA.xgh are the only representatives of PGs from insects and XGHs, respectively, so that these two enzymes are also used for comparison. The asterisks signify the identical amino acid residues and gaps are indicated by dashes. Cysteines are highlighted in grey and signified by bold. The four conserved active-site segments as well as the tyrosine invariant in all PGs (cf. Figure 2) are highlighted in black-and-white inversion. However, in RGs the strictly conserved tryptophane (Trp302 of *A.aculeatus* RG) following the Tyr301 might rather be the equivalent of the tyrosine conserved in all PGs and XGH (cf. Figure 3; for details, see text).

28, i.e. Asp180, Asp201, Lys258. This reflects very probably the fact that even closely related fungal PGs from *A.niger* have different specific kinetic parameters on polygalacturonic acid and a specific mode of action (Pavěnicová, 2000). Therefore, for example, the Met150 of PG II from *A.niger* located at the subsite –2 has no strictly conserved equivalents in the frame of the entire family 28 (Figure 5) although its mutation to glutamine affected catalysis (Pagès *et al.*, 2000).

With regard to the Tyr291, which seems to be invariantly conserved in all polygalacturonases and in the XGH, the alignment of four RGs (Figure 3) indicated that there is no corresponding tyrosine in the RG from *B.fuckeliana*. However, adjacent to the Tyr301 (*A.aculeatus* RG numbering) there is a tryptophan (Trp302) which is strictly conserved in all RGs (Figure 3). Moreover, the comparison of the hydrophobic cluster analysis plots of the RG sequences with those of the PG II from *A.niger* and XGH A from *A.tubigensis* (data not shown) supports that in RGs, a tryptophan (Trp302 in *A.aculeatus* RG) replaces the role of the conserved tyrosine in polygalacturonases and XGH (Figure 5).

As far as the cysteine residues are concerned there is only one cysteine, Cys45, conserved invariantly throughout the family 28 (Figure 5). Except for the bacterial PGs and EPGDs, all the members of this family could have six cysteine residues in common corresponding with three of the four disulfide bridges (Cys203–Cys219, Cys329–Cys334 and Cys353–Cys362) present in the PG II from *A.niger.*

Based on the alignment of representative members an evolutionary tree was constructed (Figure 6) showing the mutual relationships in the frame of the entire family 28. The long branches reflect the overall rather low sequence similarity among the groups of bacterial, fungal, plant and insect PGs, EPGDs and EPGs as well as RGs and XGH. However, the taxonomy is respected so that bacterial PGs and EPGDs as well as plant PGs and EPGs are on the bacterial and plant nodes, respectively, on the neighbouring branches. As was discussed above, the insect PG is most closely related to fungal PGs represented by the PG II from *A.niger* and the XGH to fungal EPGs represented by the EPG from *A.tubigensis*. The RG A from *A.aculaetus* positioned on the longest isolated

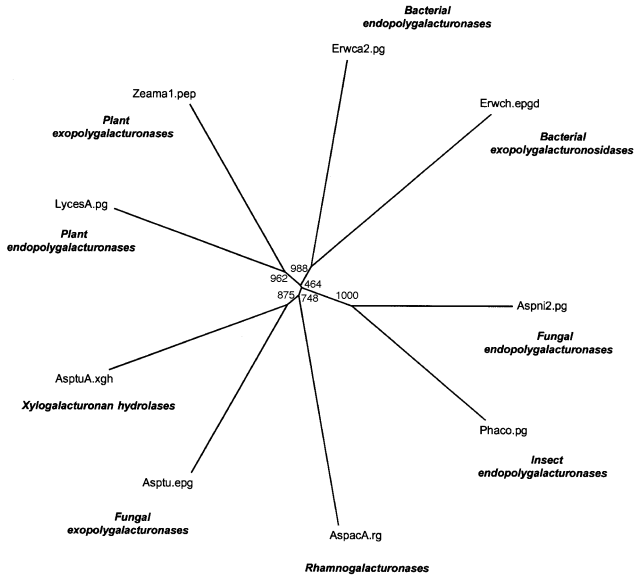

Fig. 6. Evolutionary tree of the family 28 representatives. The tree is based on the alignment shown in Figure 5. The abbreviations of enzyme sources are given in Table I and the choice of the representatives is explained in the legend to Figure 5. The branch lengths are proportional to the sequence divergence. Numbers along branches are bootstrap values (1000 replicates). In the future the RGs and XGHs may be expected to be also divided according to the taxonomy (like the PGs, EPGs and EPGDs).

branch manifests the amendments in the sequences of RGs even in the four active-site segments.

Acknowledgements

Dr D.Falconet is thanked for fruitful discussion. This work was financially supported by the VEGA grants Nos 2/6045/99 and 2/7142/20 from the Slovak Grant Agency for Science and by the Slovak Literary Fund.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.

Armand,S., Wagemaker,M.J.M., Sánchez-Torres,P., Kester,H.C.M., van Santen,Y., Dijkstra,B.W., Visser,J. and Benen,J.A.E. (2000) *J. Biol. Chem.*, **275**, 691–696.

Bairoch,A. and Apweiler,R. (2000) *Nucleic Acids Res.*, **28**, 45–48.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) *Nucleic Acids Res.*, **28**, 15–18.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.

Bussink,H.J.D., Kester,H.C.M. and Visser,J. (1990) *FEBS Lett.*, **273**, 127–130.

Bussink,H.J.D., Buxton,F.P. and Visser,J. (1991) *Curr. Genet.*, **19**, 467–474.

Callebaut,I., Labesse,G., Durand,P., Poupon,A., Canard,L., Chomilier,J., Henrissat,B. and Mornon,J.P. (1997) *Cell Mol. Life Sci.*, **53**, 621–645.

Clarke,A.J. and Svensson,B. (1984) *Carlsberg Res. Commun.*, **49**, 111–122.

Coutinho,P.M. and Henrissat,B. (2000) URL: http://afmb.cnrs-mrs.fr/~pedro/CAZY/ghf_28. html

Felsenstein,J. (1985) *Evolution*, **39**, 783–791.

Gaboriaud,C., Bissery,V., Benchetrit,T. and Mornon,J.P. (1987) *FEBS Lett.*, **224**, 149–155.

Gibson,R.M. and Svensson,B. (1986) *Carlsberg Res. Commun.*, **51**, 295–308.

Girard,C. and Jouanin,L. (1999) *Insect Biochem. Mol. Biol.,* **29**, 1129–1142.

Gognies,S., Gainvors,A., Aigle,M. and Belarbi,A. (1999) *Yeast*, **15**, 11–22.

Gonzalez,C.F., Pettit,E.A., Valadez,V.A. and Provin,E.M. (1997) *Mol. Plant Microbe Interact.*, **10**, 840–851.

Guex,N. and Peitsch,M.C. (1997) *Electrophoresis*, **18**, 2714–2723.

Guex,N., Diemand,A. and Peitsch,M.C. (1999) *Trends Biochem. Sci.*, **24**, 364–367.

Hadfield,K.A. and Bennett,A.B. (1998) *Plant Physiol.*, **117**, 337–343.

Hadfield,K.A., Rose,J.K.C., Yaver,D.S., Berka,R.M. and Bennett,A.B. (1998) *Plant Physiol.*, **117**, 363–373.

Henrissat,B. (1991) *Biochem. J.*, **280**, 309–316.

Henrissat,B. and Davies,G. (1997) *Curr. Opin. Struct. Biol.*, **7**, 637–644.

Hinton,J.C., Gill,D.R., Lalo,D., Plastow,G.S. and Salmond,G.P. (1990) *Mol. Microbiol.*, **4**, 1029–1036.

Huang,Q. and Allen,C. (1997) *J. Bacteriol.*, **179**, 7369–7378.

Jenkins,J., Mayans,O. and Pickersgill,R. (1998) *J. Struct. Biol.*, **122**, 236–246.

Kester,H.C.M., Kusters-van Someren,M.A., Muller,Y. and Visser,J. (1996) *Eur. J. Biochem.*, **240**, 738–746.

Kofod,L.V., Kauppinen,S., Christgau,S., Andersen,L.N., Heldt-Hansen,H.P., Dörreich,K. and Dalbøge,H. (1994) *J. Biol. Chem.*, **269**, 29182–29189.

Muñoz,J.A., Coronado,C., Pérez-Hormaeche,J., Kondorosi,A., Ratet,P. and Palomares,A.J. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 9687–9692.

Nelson,K.E. *et al*. (1999) *Nature,* **399**, 323–329.

Page,R.D.M. (1996) *Comput. Applic. Biosci.*, **12**, 357–358.

Pagès,S., Heijne,W.H.M., Kester,H.C.M., Visser,J. and Benen,J.A.E. (2000) *J. Biol. Chem.*, **275**, 29348–29353.

Pařenicová,L. (2000) PhD Thesis, Wageningen University, The Netherlands.

Petersen,M., Sander,L., Child,R., van Onckelen,H., Ulvskov,P. and Borkhardt,B. (1996) *Plant Mol. Biol.*, **31**, 517–527.

Petersen,T.N., Kauppinen,S. and Larsen,S. (1997) *Structure*, **5**, 533–544.

Pickersgill,R., Smith,D., Worboys,K. and Jenkins,J. (1998) *J. Biol. Chem.*, **273**, 24660–24664.

Rao,M.N., Kembhavi,A.A. and Pant,A. (1996) *Biochim. Biophys. Acta*, **1296**, 167–173.

Rexová-Benková,L. and Mračková,M. (1978) *Biochim. Biophys. Acta*, **523**, 162–169.

Rong,L.J., Karcher,S.J. and Gelvin,S.B. (1991) *J. Bacteriol.*, **173**, 5110–5120.

Ruttkowski,E., Khanh,N.Q., Wientjes,F.J. and Gottschalk,M. (1991) *Mol. Microbiol.*, **5**, 1353–1361.

Saarilahti,H.T., Heino,P., Pakkanen,R., Kalkkinen,N., Palva,I. and Palva,E.T. (1990) *Mol. Microbiol.*, **4**, 1037–1044.

Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.

Schols,H.A. and Voragen,A.G.J. (1996) In Visser,J. and Voragen,A.G.J. (eds), *Progress in Biotechnology*, *Vol. 14*, *Pectins and Pectinases.* Elsevier Science, Amsterdam, pp. 3–19.

Scott-Craig,J.S., Panaccione,D.J., Cervone,F. and Walton,J.D. (1990) *Plant Cell*, **2**, 1191–2000.

Stratilová,E., Markovič,O., Škrovinová,D., Rexová-Benková,L. and Jörnvall,H. (1993) *J. Protein Chem.*, **12**, 15–22.

Stratilová,E., Dzurová,M., Markovič,O. and Jörnvall,H. (1996) *FEBS Lett.*, **382**, 164–166.

Stratilová,E., Mislovičová,D., Kačuráková,M., Machová,E., Kolarová,N., Markovič,O. and Jörnvall,H. (1998) *J. Protein Chem.*, **17**, 173–179.

Tebbutt,S.J., Rogers,H.J. and Lonsdale,D.M. (1994) *Plant Mol. Biol.*, **25**, 283–297.

Tenberge,K.B., Homann,V., Oeser,B. and Tudzynski,P. (1996) *Phytopathology*, **86**, 1084–1097.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.

Torki,M., Mandaron,P., Thomas,F., Quigley,F., Mache,R. and Falconet,D. (1999) *Mol. Gen. Genet.*, **261**, 948–952.

Torki,M., Mandaron,P., Mache,R. and Falconet,D. (2000) *Gene*, **242**, 427–436.

Van der Vlugt-Bergmans,C.J.B., Meeuwsen,P.J.A., Voragen,A.G.J., van Ooyen,A.J.J. (2000) *Appl. Environ. Microbiol.*, **66**, 36–41.

Van Santen,Y., Benen,J.A., Schröter,K.H., Kalk,K.H., Armand,S., Visser,J. and Dijkstra,B.W. (1999) *J. Biol. Chem.*, **274**, 30474–30480.

Williamson,M.P., Le Gal-Coeffet,M.F., Sorimachi,K., Furniss,C.S., Archer,D.B. and Williamson,G. (1997) *Biochemistry,* **36***,* 7535–7539.

Williamson,G., Kroon,P.A. and Faulds,C.B. (1998) *Microbiology*, **144**, 2011–2023.

Wubben,J.P., Mulder,W., ten Have,A., van Kan,J.A. and Visser,J. (1999) *Appl. Environ. Microbiol.*, **65**, 1596–1602.

Yoder,M.D. and Jurnak,F. (1995) *FASEB J.*, **9**, 335–342.

Yoder,M.D., Keen,N.T. and Jurnak,F. (1993) *Science*, **260**, 1503–1507.

Received December 15, 2000; revised March 20, 2001; accepted June 15, 2001