Available online at www.sciencedirect.com
ScienceDirect
Phytochemistry 68 (2007) 521–535
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

Broad screening of the legume family for variability in seed insecticidal activities and for the occurrence of the A1b-like knottin peptide entomotoxins ☆

Sandrine Louis, Bernard Delobel, Frédéric Gressent, Gabrielle Duport, Ousmane Diol, Isabelle Rahioui, Hubert Charles, Yvan Rahbé *

UMR203 BF21 (Biologie Fonctionnelle Insectes et Interactions), IFR41, INRA, INSA-Lyon, F-69621 Villeurbanne, France

Received 5 April 2006; received in revised form 20 November 2006
Available online 12 January 2007

Abstract

Pea albumin 1b (PA1b) is a small sulphur-rich peptide from pea seeds, also named leginsulin because of the binding characteristics of its soybean orthologue. Its insecticidal properties were discovered more recently. By using a combination of molecular, biochemical and specific insect bioassays on seed extracts, we characterised genes from numerous Papilionoideae, but not from Caesalpinioideae or Mimosoideae, although the last group harboured species with partially positive cues (homologous biological activities). The A1b defence peptide family, therefore, appears to have evolved relatively late in the legume lineage, maybe from the sophoroid group (*e.g. Styphnolobium japonicum*). However, unambiguous sequence information is restricted to a group of tribes within the subfamily Papilionoideae (Psoraleae, Millettieae, Desmodieae, Hedysareae, Phaseoleae, Vicieae, and the now clearly polyphyletic "Trifolieae" and "Galegeae"). Recent diversification by gene duplications has occurred in many species, or longer ago in some lineages (*Medicago truncatula*), as well as probable gene or expression losses at different taxonomic levels (Loteae, *Vigna subterranea*).

Keywords: Leguminosae; Fabaceae; Albumins 1b; PA1b; Insect; Weevil; Plant defence; Toxin; Cystine-knot peptide

1. Introduction

Albumins 1b were first discovered in the pea as a set of cysteine-rich peptides without any obvious function other than sulphur-storage (Higgins et al., 1986). Consisting of 37 amino-acids, PA1b is stabilised by three disulfide bridges, which confer on the peptide a high stability to thermal and protease inactivation (Delobel et al., 1998; Hancock et al., 1994). PA1b is encoded by the pea albumin 1 gene, translated as a pre-proprotein with a signal peptide and a small C-terminal linker peptide, followed by another sulphur-rich 53-residue polypeptide termed PA1a, as yet of unknown function (Higgins et al., 1986). A homologous A1b peptide has also been identified from soybean seeds as a natural endogenous ligand to an insulin-binding globulin, and it has been named leginsulin (Watanabe et al., 1994). Recent results detailing the structure of PA1b and leginsulin have led to them being assigned to the knottin fold family (Gelly et al., 2004; Jouvensal et al., 2003; Yamazaki et al., 2003), which includes protease inhibitors

Abbreviations: HCA, hierarchical cluster analysis; IRLC, inverted repeat loss clade; Maldi-Tof MS, matrix assisted laser desorption ionisation time-of-flight detection mass spectrometry; MCA, multiple correspondence analysis; MeOH/MeOH60/H_2O5/H_2O8 methanolic fraction (resp. 60% aqueous methanol/acidic water/basic water extracts see text for details); PA1b/a, pea albumin 1 subunit b (resp. a); PAGE, polyacrylamide gel electrophoresis; PCA, principal component analysis; SDS, sodium dodecyl sulphate.

☆ A1b-like entomotoxic peptides in legume seeds.

* Corresponding author. Tel.: +33 472438476; fax: +33 472438534.
E-mail address: yvan.rahbe@jouy.inra.fr (Y. Rahbé).

0031-9422/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2006.11.032

and channel blockers as well as small lectins. However, PA1b does not seem to exhibit any common enzyme inhibitor activity (Mbaiguinam, 1996), and its mechanism of action involves recognition by a single high-affinity receptor-like membrane protein (Gressent et al., 2003). Work concerning the soybean 7S globulin-binding properties of leginsulin has also been published recently (Hanada et al., 2003; Yamazaki et al., 2003). A few homologous sequences from Papilionoideae species have been characterised in the past decade (Ilgoutz et al., 1997; Louis et al., 2004), attesting to the variability of these A1b proteins among closely related tribes of Papilionoideae.

Being the third largest plant family as regards number of species, legumes (Fabaceae) are divided into three subfamilies, two of which form a monophyletic clade each (Papilionoideae and Mimosoideae) and the third one forming a basal paraphyletic group (Caesalpinioideae). Papilionoideae represents the biggest subfamily and can be divided into one basal grade, the Sophoroids, and four clades, the Genistoids (lupine), Aeschynomenoids (=Dalbergioids: peanut), Galegoids (=Hologalegina including Robinioids (=Loteae *sl*) and the IRLC: pea, alfalfa) and Phaseoloids (bean, soybean) (Wojciechowski, 2003; Wojciechowski et al., 2004; see supplementary material 1). The phylogenetic relationships between and within certain tribes of this large family are still under discussion. This large family is known for its variety of defence compounds, either secondary metabolites, such as alkaloids or terpenoids, or proteins, such as lectins, chitinases, numerous proteases or α-amylase inhibitors (Carlini and Grossi de Sa, 2002; Wink and Mohamed, 2003).

The practice of mixing stored cereals with different edible legume seeds in the granaries has been used for a long time. It was shown to reduce the survival and reproduction of the rice weevil (*Sitophilus oryzae*), a major pest causing heavy losses to stored grains (Coombs et al., 1977; Holloway and Smith, 1985). Whereas the antifeeding activity of pea seeds as regards *S. oryzae* was recently shown to be, in part, due to secondary metabolites such as saponins (Taylor et al., 2004a), the acute mortality of cereal weevils (*Sitophilus* sp.) fed on pea seeds is due to PA1b (Delobel et al., 1998; Louis et al., 2004). Regularly eaten by human beings and cattle without displaying any allergenicity nor toxicity, PA1b is of great interest in the control of cereal weevils and other insect pests, such as the pea aphid *Acyrthosiphon pisum* (Delobel et al., 1998). However, in *S. oryzae*, a few strains have been found to harbour a single recessive gene responsible for full immunity to this peptide (Grenier et al., 1997). The high affinity binding-site identified for PA1b was present in all the susceptible insects screened so far, but absent in the few available resistant weevil strains. This unidentified protein seems, therefore, to be directly involved in PA1b toxicity (Gressent et al., 2003).

Based on this specific biological feature, a comprehensive survey of the Albumin 1b family in the Fabaceae was planned in order to characterise the presence and variability of this toxic trait, and of the underlying peptides, among legumes. We selected a set of 88 species evenly scattered among the three cited subfamilies (Caesalpinioideae, Mimosoideae and Papilionoideae). Using an approach validated on a small subset of species (Louis et al., 2004), combining bioassays on susceptible and PA1b-resistant weevils with biochemical and molecular techniques (serial hydrophobic extractions, HPLC, mass spectrometry and genomic PCR), we quantified various biological activities on seed extracts from 88 plant species. We cloned 19 new homologous genomic sequences from two large clades of the subfamily Papilionoideae (14 species), and confirmed the presence of A1b peptides in extracts with PA1b-like activity from selected species within many other tribes of the Papilionoideae. Similarly, activities were also observed in species within the Mimosoideae (albeit with no sequence available). No evidence for the presence of A1b genes or peptides was obtained from the Caesalpinioideae.

2. Results and discussion

2.1. Variability of biological activity of seed peptide fractions on S. oryzae

To identify plant species displaying A1b-like activities in their seeds, serial solvent extracts of the flours, targeting the hydrophobic peptides, were assayed on PA1b-susceptible and PA1b-resistant weevil strains (Grenier et al., 1997), as previously validated on a small set of plants (Louis et al., 2004). Median survival time (*i.e.* lethal time 50%, based on survival analysis), was chosen as a measure of the insecticidal activity of each fraction on each strain. The full results on 88 legume species are shown in Table 1.

2.1.1. Variability of seed extract toxicity

The overall variability in entomotoxic activity of the different legume seeds is very high, ranging from none (as for many Caesalpinioideae, *Vigna subterranea* or *Cicer arietinum*) to very acute, with species from which all of the extracts tested are highly toxic to both weevil strains (such as *Medicago truncatula*, *Robinia pseudoacacia*, *Bolusanthus speciosus*, or even the less toxic *Canavalia brasiliensis*, *Caragana arborescens* or *Lotus corniculatus*). The toxicity may be regarded as a qualitative variable (presence or absence of acute toxicity) or as a quantitative one, which may be analysed by multivariate methods to derive a plant toxic typology. Taken as a nominal/*qualitative* trait controlled by three factors (plant *species*, *chemical* extract and weevil *strain*), toxicity was shown to be absent in more than 50% of the cases (447 extracts out of 880). A deviance analysis on this trait (generalised linear fit, binomial model) gave similar results to those of a variance analysis on a $\log(1+x)$ fit of the quantitative toxicity trait: any variation was mainly due to plant *species* ($p = 1 \times 10^{-53}$), then weevil

Table 1
Mean survival time (lethal time 50%) of *S. oryzae* susceptible (*S*) and resistant (*R*) strains on diets containing the five seed fractions from selected species within the legume family; dashes denote no or minor toxicity by the end of experiment (15 or 20 days)

Subfamilly (grade)	Tribe (subtribe or group (Chappill))	Species	#[a]	Clusters[c]		LT50 (SE)[b] for each fraction-strain pair									
						MeOH		MeOH60		H_2O5		H_2O8		Residue	
				PCA	MCA	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*
Polygalaceae		*Polygala myrtifolia*				5.5 (0.6)	10.1 (0.4)	6.2 (0.3)	12.2 (0.4)	–	–	–	–	**10.5** (1.9)	–
Caesalpinioideae	Cercideae	*Cercis siliquastrum*	E03	3	L	–	–	5.7 (0.6)	–	–	–	–	–	4.2 (0.3)	4.8 (0.4)
	Cercideae	*Bauhinia natalensis*	E60	3	L	5.1 (0.5)	8.7 (1.2)	–	-	–	–	–	–	6.4 (0.2)	–
	Detarieae (Hymenostegia gp)	*Intsia bijuga*	E04	1	L	–	–	–	–	–	–	–	–	–	–
	Detarieae (Cynometra gp)	*Schotia afra*	E65	1	L	–	–	–	–	–	–	–	–	–	–
	Detarieae (Amherstia gp)	*Tamarindus indica*	E67	1	L	6.1 (0.4)	11.7 (1.2)	–	–	–	–	–	–	–	–
	Detarieae (Brachystegia gp)	*Brachystegia spiciformis*	E61	1	L	–	–	–	–	–	–	–	–	**12.7** (0.9)	–
	Cassieae (Cassiinae)	*Cassia occidentalis*	E02	1	L	–	–	–	–	–	–	–	–	–	–
	Cassieae (Ceratoniinae)	*Ceratonia siliqua*	E63	3	L	–	–	–	–	–	–	–	–	5.3 (0.6)	5.4 (0.3)
	Cassieae (Dialiinae)	*Dialium guineense*	E64	3	L	–	–	–	–	–	**10.9** (0.5)	–	–	4.8 (0.4)	6.0 (0.3)
	Cassieae (Dialiinae)	*Storckiella pancheri*	E66	3	L	–	–	–	–	13.0 (1.9)	–	–	–	5.4 (0.3)	5.9 (0.2)
	Caesalpinieae (Gleditsia gp)	*Gleditsia triacanthos*	E01	1	L	–	–	–	–	–	–	–	–	–	–
	Caesalpinieae (Peltophorum gp)	*Delonix regia*	E05	1	L	–	–	–	–	–	–	–	–	–	–
	Caesalpinieae (Caesalpinia gp)	*Caesalpinia bonduc*	E62	1	L	**9.0** (1.3)	–	–	–	–	–	–	–	–	–
	Caesalpinieae (Acrocarpus gp)	*Acrocarpus fraxinifolius*	E59	1	L	–	–	–	–	–	–	–	–	11.4 (0.4)	12.7 (1.2)
Mimosoideae	Mimoseae (Prosopis gp)	*Prosopis africana*	E71	3	D	–	–	**8.9** (0.7)	–	**6.5** (0.3)	-	**8.7** (1.9)	–	6.1 (0.4)	7.1 (0.3)
	Mimoseae (Dichrostachys gp)	*Leucaena leucocephala*	E69	4	H	–	–	8.1 (0.5)	–	5.0 (0.6)	7.4 (0.3)	7.9 (0.8)	13.4 (0.4)	5.9 (0.4)	8.3 (0.3)
	Mimoseae (Dichrostachys gp)	*Desmanthus illinoensis*	E68	3	L	–	–	–	–	–	–	–	–	5.9 (0.3)	10.0 (1.3)
	Ingeae	*Acacia dealbata*	E07	3	D	**10.6** (0.7)	–	13.0 (3.9)	–	**12.0** (0.8)	–	13.0 (1.0)	–	6.9 (0.2)	8.4 (0.3)
	Ingeae	*Albizia lebbeck*	E08	2	H	4.1 (0.5)	4.3 (0.4)	4.7 (0.3)	4.7 (0.4)	4.9 (0.3)	6.8 (0.2)	7.1 (1.0)	11.8 (0.7)	6.1 (0.3)	8.4 (0.3)
	Ingeae	*Albizia julibrissin*	E06	4	L	9.3 (0.6)	12.9 (0.4)	6.1 (0.5)	8.1 (0.2)	7.0 (1.0)	9.0 (0.5)	–	–	**10.0** (1.8)	–
	Parkieae	*Parkia biglobosa*	E70	1	L	**7.0** (0.8)	–	–	–	–	–	–	–	–	–
Papilionoideae (Sophoroids)	Swartzieae	*Bobgunnia madagascariensis*	E73	2	H	–	–	5.2 (0.4)	4.8 (0.3)	8.7 (0.6)	8.3 (0.5)	6.0 (0.6)	5.7 (0.4)	6.2 (0.4)	6.1 (0.4)
	Sophoreae (Cladrastis gp)	*Styphnolobium japonicum*	E47	3	D	**3.9** (0.2)	–	**5.2** (0.5)	–	**5.7** (0.5)	–	**5.4** (0.7)	–	4.9 (0.5)	6.9 (0.4)
	Brongnartieae	*Templetonia retusa*	E32	2	H	**5.5** (0.4)	–	5.4 (0.2)	4.6 (0.2)	11.2 (0.9)	–	8.2 (0.5)	9.8 (0.7)	6.9 (0.5)	7.7 (0.4)
Papilionoideae (Aeschynomenoids)	Amorpheae	*Amorpha fruticosa*	E14	1	L	**5.9** (0.3)	–	**10.5** (1.9)	–	–	–	–	–	–	–
	Adesmieae	*Adesmia bicolor*	E48	3	L	–	–	**10.0** (1.3)	–	–	–	–	–	8.9 (0.7)	13.6 (0.9)
	Dalbergieae (Dalbergia gp)	*Dalbergia purpurescens*	E17	3	L	4.7 (0.4)	5.3 (0.4)	7.7 (0.7)	–	–	–	–	–	7.5 (0.7)	7.9 (0.9)
	Aeschynomeneae (Stylosanthinae)	*Arachis hypogaea*	E22	1	L	–	–	–	–	–	–	–	–	9.0	–
	Dalbergieae (Tipuana gp)	*Pterocarpus rotundifolius*	E42	2	H	13.8 (0.7)	–	4.2 (0.3)	4.5 (0.2)	3.8 (0.2)	4.4 (0.4)	3.6 (0.2)	5.1 (0.3)	3.9 (0.2)	5.2 (0.2)

(*continued on next page*)

Table 1 (*continued*)

Subfamilly (grade)	Tribe (subtribe or group (Chappill))	Species	#[a]	Clusters[c]		LT50 (SE)[b] for each fraction-strain pair									
						MeOH		MeOH60		H_2O5		H_2O8		Residue	
				PCA	MCA	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*
Papilionoideae (Genistoids)	Sophoreae (Sophora gp)	*Sophora arizonica*	E34	2	H	5.6 (0.5)	4.6 (0.2)	5.4 (0.3)	5.2 (0.5)	5.3 (0.3)	8.8 (0.7)	7.4 (0.4)	11.7 (0.5)	6.8 (0.3)	9.0 (0.6)
	Sophoreae (Dussia gp)	*Diplotropis purpurea*	E76	1	L	3.6 (0.2)	6.8 (0.4)	5.8 (0.6)	–	–	–	–	–	–	–
	Sophoreae (Sophora gp)	*Bolusanthus speciosus*	E50	2	H	4.6 (0.3)	4.1 (0.3)	4.9 (0.5)	4.0 (0.2)	6.5 (0.4)	5.7 (0.6)	9.1 (0.5)	10.2 (0.5)	5.7 (0.3)	5.4 (0.4)
	Podalyrieae	*Calpurnia aurea*	E51	2	H	3.9 (0.3)	4.5 (0.2)	3.8 (0.2)	4.9 (0.2)	4.6 (0.4)	8.4 (0.7)	6.0 (0.4)	–	4.6 (0.4)	5.0 (0.5)
	Podalyrieae	*Podalyria biflora*	E85	1	L	**6.3** (0.5)	–	**4.9**(0.4)	–	–	–	–	–	–	–
	Thermopsideae (Thermopsis gp)	*Baptisia australis*	E31	1	L	**7.2** (0.7)	–	**9.0** (1.9)	–	–	–	–	–	–	–
	Crotalarieae	*Crotalaria eremaea*	E25	2	H	6.3 (0.4)	5.9 (0.2)	*5.6(0.3)*	*4.3(0.4)*	6.4 (0.3)	6.0 (1.9)	9.7 (0.7)	11.5 (0.3)	5.7 (0.4)	5.0 (0.2)
	Genisteae	*Lupinus albus*	E16	1	H	–	–	–	–	16.7 (0.7)	19.2 (1.2)	15.4 (0.7)	17.5 (1.0)	9.9 (0.6)	15.4 (0.5)
	Genisteae	*Lupinus angustifolius*	E54	2	H	11.3 (0.6)	–	**11.4** (0.4)	–	9.1 (0.5)	12.2 (0.9)	6.8 (0.6)	4.9 (0.3)	6.9 (0.5)	8.2 (0.5)
	Genisteae	*Cytisophyllum sessilifolium*	E74	3	L	–	–	**7.3** (1.2)	–	–	–	–	–	**10.3** (1.2)	–
	Genisteae	*Ulex europaeus*	E41	1	L	–	–	–	–	–	–	–	–	–	–
	Genisteae	*Laburnum anagyroides*	E40	3	L	**9.0** (1.9)	–	**5.9** (0.5)	–	–	–	–	–	8.0 (0.4)	13.0 (1.9)
	Genisteae	*Spartium junceum*	E86	1	L	6.5 (0.9)	–	5.9 (0.2)	10.9 (0.5)	–	–	–	–	–	–
Papilionoideae (Mirbelieae sl)	Mirbelieae	*Goodia lotifolia*	E80	4	H	–	–	5.6 (0.5)	8.6 (0.3)	6.8 (0.2)	9.6 (0.4)	11.5 (3.9)	13.7 (0.5)	9.9 (0.8)	13.4 (0.2)
	Mirbelieae	*Daviesia corymbosa*	E75	3	L	–	–	**6.1** (0.7)	–	–	–	9.5 (1.1)	13.3 (1.2)	5.8 (0.3)	11.0 (0.7)
Papilionoideae **(Phaseoloids)**	Psoraleae	*Bituminaria bituminosa*	E29	3	H	5.6 (0.4)	6.0 (0.4)	4.7 (0.3)	7.7 (0.2)	**11.8** (0.3)	–	**8.3** (0.4)	–	5.2 (0.4)	8.2 (0.3)
	Indigofereae	*Indigofera astragalina*	E35	3	L	–	–	–	–	–	–	–	–	6.3 (0.3)	7.0 (0.6)
	Millettieae	*Lonchocarpus capassa*	E18	2	H	–	–	3.2 (0.3)	4.3 (0.2)	3.4 (0.2)	5.0 (0.2)	3.9 (0.2)	7.6 (0.3)	4.4 (0.2)	6.4 (0.2)
	Millettieae	*Mundulea sericea*	E44	3	H	6.8 (0.3)	6.0 (0.5)	6.2 (0.6)	6.0 (1.3)	–	–	**11.0** (0.8)	–	5.8 (0.4)	14.0 (1.8)
	Desmodieae (Desmodiinae)	*Alysicarpus ovalifolius*	E33	3	L	–	–	**6.7** (0.7)	–	–	–	–	–	7.3 (0.4)	9.1 (1.1)
	Desmodieae (Desmodiinae)	*Desmodium canadense*	E36	4	H	**5.5** (0.4)	–	4.9 (0.2)	7.2 (0.5)	5.4 (0.5)	8.7 (0.4)	–	–	**6.7** (0.6)	–
	Phaseoleae (Glycininae)	*Glycine max*	E10	3	D	**7.2** (0.9)	–	**6.5** (0.9)	–	–	–	**10.8** (0.6)	–	**8.0** (1.3)	–
	Phaseoleae (Diocleinae)	*Dioclea grandiflora*	E57	4	L	–	–	**5.1** (0.3)	–	4.7 (0.3)	4.8 (0.2)	**10.7** (0.9)	–	–	–
	Phaseoleae (Diocleinae)	*Canavalia brasiliensis*	E43	4	H	–	–	3.8 (0.2)	3.8 (0.2)	4.8 (0.4)	5.3 (0.2)	4.4 (0.3)	5.7 (0.2)	–	–
	Phaseoleae (Kennediinae)	*Hardenbergia comptoniana*	E45	2	H	–	–	4.6 (0.2)	5.7 (0.3)	4.5 (0.2)	6.7 (0.1)	4.7 (0.2)	8.4 (0.5)	4.3 (0.2)	7.3 (0.3)
	Phaseoleae (Kennediinae)	*Kennedia nigricans*	E81	2	H	3.9 (0.3)	6.6 (0.1)	–	–	3.8 (0.2)	6.3 (0.2)	3.5 (0.2)	7.0 (0.2)	3.7 (0.3)	6.0 (0.4)
	Phaseoleae (Erythrininae)	*Erythrina crista-galli*	E77	3	L	5.3 (0.5)	11.2 (1.9)	**13.4** (0.9)	–	–	–	–	–	**9.6** (0.5)	–
	Phaseoleae (Cajaninea)	*Rhynchosia pentheri*	E49	4	L	–	–	**5.1** (0.3)	–	**8.5** (0.7)	–	5.2 (0.3)	6.5 (0.1)	–	–

	Phaseoleae (Phaseolinae)	*Phaseolus vulgaris*	E12	3	D	–	–	**6.3** (0.3)	–	–	–	**8.7** (0.9)	–	**10.6** (0.2)	–
	Phaseoleae (Phaseolinae)	*Vigna radiata*	E88	3	D	–	–	**4.5** (0.3)	–	–	–	**4.6** (0.2)	–	**12.2** (0.6)	–
	Phaseoleae (Phaseolinae)	*Vigna subterranea*	E24	1	L	–	–	–	–	–	–	–	–	–	–
	Phaseoleae (Phaseolinae)	*Vigna unguiculata*	E11	3	D	–	–	**5.5** (0.5)	–	**4.9** (0.2)	–	**4.6** (0.3)	–	**13.4** (0.7)	–
Papilionoideae	Robinieae	*Robinia pseudoacacia*	E15	2	H	3.7 (0.3)	4.9 (0.2)	3.7 (0.2)	6.5 (0.3)	5.8 (0.2)	9.0 (1.9)	3.7 (0.4)	4.9 (0.2)	4.0 (0.8)	5.5 (0.3)
(Hologalegina)	Robinieae	*Sesbania sesban*	E21	2	H	**14.2** (0.9)	–	3.9 (0.2)	6.2 (0.4)	5.4 (0.7)	6.6 (0.2)	5.9 (0.4)	7.2 (0.3)	4.8 (0.3)	7.2 (0.5)
	Abreae	*Abrus precatorius*	E28	3	L	–	–	**6.6** (0.9)	–	–	–	**5.5** (0.1)	–	5.5 (0.2)	7.8 (0.3)
	Loteae	*Anthyllis barba–jovis*	E56	2	H	–	–	3.7 (0.4)	4.9 (0.3)	3.7 (0.4)	4.4 (0.2)	3.9 (0.2)	5.6 (0.7)	4.1 (0.3)	4.5 (0.3)
	Loteae	*Hippocrepis emerus*	E46	4	H	–	–	5.0 (0.4)	8.2 (0.2)	6.6 (0.3)	8.2 (0.2)	–	–	6.9 (0.3)	9.6 (0.3)
	Loteae	*Lotus corniculatus*	E09	4	H	–	–	7.3 (0.6)	9.1 (0.2)	5.0 (0.7)	7.4 (0.2)	–	–	7.7 (0.4)	7.2 (0.5)
	Loteae	*Lotus tetragonolobus*	E53	1	L	–	–	–	–	–	–	–	–	**10.6** (0.3)	–
	Wisterieae	*Wisteria sinensis*	E27	4	H	–	–	4.5 (0.7)	6.4 (0.3)	4.6 (0.4)	6.6 (0.3)	5.8 (0.4)	7.7 (0.2)	–	–
	Galegeae (Glycyrrhiza gp)	*Glycyrrhiza glabra*	E79	2	H	–	–	3.7 (0.2)	5.5 (0.2)	3.5 (0.2)	5.8 (0.1)	5.7 (0.6)	9.4 (0.2)	6.1 (0.5)	10.2 (0.4)
	Galegeae (Astragalinae)	*Caragana arborescens*	E30	4	H	–	–	5.0 (0.4)	6.3 (0.4)	12.7 (0.2)	12.4 (0.2)	–	–	–	–
	Hedysareae	*Onobrychis viciifolia*	E38	3	D	–	–	**5.8** (0.2)	–	**5.6** (0.3)	–	**6.5** (0.5)	–	5.3 (0.3)	8.2 (0.6)
	Galegeae (Astragalinae)	*Alhagi graecorum*	E72	2	H	4.4 (0.5)	–	3.7 (0.4)	5.6 (0.2)	3.2 (0.3)	5.8 (0.2)	3.5 (0.2)	13.5 (0.6)	3.87 (0.2)	–
	Galegeae (Coluteinae)	*Colutea arborescens*	E37	2	H	**6.6** (0.5)	–	5.0 (0.6)	6.2 (0.4)	5.0 (0.5)	6.2 (0.2)	5.0 (0.5)	7.6 (0.2)	6.3 (0.4)	8.3 (0.4)
	Galegeae (Astragalinae)	*Astragalus monspessulanus*	E13	4	L	–	–	6.6 (0.4)	13.0 (0.4)	9.5 (1.8)	11.5 (3.6)	–	–	9.6 (0.6)	–
	Carmichaelieae	*Carmichaelia stevensoni*	E58	2	H	**7.7** (0.6)	–	5.3 (0.4)	6.4 (0.3)	4.7 (0.1)	6.0 (0.2)	4.8 (0.2)	7.5 (0.3)	6.1 (0.4)	8.6 (0.3)
	Cicereae	*Cicer arietinum*	E52	1	L	–	–	13.0 (3.9)	–	–	–	–	–	–	–
	Galegeae (Galeginae)	*Galega officinalis*	E78	3	H	5.2 (0.4)	11.1 (0.3)	3.8 (0.2)	7.5 (0.4)	**7.3** (1.3)	–	–	–	3.7 (0.1)	6.8 (0.2)
	Trifolieae	*Melilotus albus*	E19	3	H	**5.9** (0.5)	–	4.7 (0.4)	9.0 (1.9)	–	–	–	–	**5.2** (0.2)	–
	Trifolieae	*Trigonella foenum-graecum*	E20	2	H	4.8 (0.3)	6.2 (0.5)	4.5 (0.3)	5.2 (0.5)	9.9 (0.9)	12.4 (0.7)	13.7 (0.6)	15.1 (0.3)	6.3 (0.4)	6.7 (0.3)
	Trifolieae	*Medicago lupulina*	E82	2	H	6.8 (0.7)	11.5 (0.9)	4.3 (0.3)	6.3 (0.3)	3.6 (0.2)	6.6 (0.3)	5.0 (0.4)	7.3 (0.5)	5.0 (0.4)	7.1 (0.4)
	Trifolieae	*Medicago truncatula*	E83	2	H	3.9 (0.4)	8.8 (0.6)	3.5 (0.3)	3.6 (0.2)	4.5 (0.2)	6.4 (0.3)	4.6 (0.3)	5.2 (0.2)	4.3 (0.4)	5.2 (0.2)
	Trifolieae	*Trifolium fragiferum*	E87	4	H	–	–	4.2 (0.3)	5.4 (0.3)	6.2 (0.7)	9.6 (0.4)	**3.7** (0.2)	–	4.0 (0.4)	5.7 (0.2)
	Vicieae	*Lens culinaris*	E39	3	D	–	–	**4.6** (0.2)	–	**5.7** (0.4)	–	**4.6** (0.2)	–	**5.4** (0.7)	–
	Vicieae	*Vicia hirsuta*	E55	3	D	–	–	**5.0** (1.3)	–	**6.1** (0.4)	–	**5.0** (0.4)	–	**6.2** (0.7)	–
	Vicieae	*Lathyrus latifolius*	E26	3	D	–	–	**14.7** (1.9)	–	–	–	**3.9** (0.2)	–	**6.7** (0.4)	–
	Vicieae	*Pisum sativum*	E23	3	D	–	–	**4.5** (0.2)	–	**5.7** (0.4)	–	**5.0** (0.4)	–	**5.3** (0.3)	–
	Vicieae	*Pisum sativum arvense*	E84	3	D	–	–	**3.5** (0.2)	–	**4.1** (0.9)	–	**3.7** (0.3)	–	**4.4** (0.4)	–

[a] #: species label used elsewhere (Supplementary materials 2 and 4).

[b] Lethal times 50% of a survival analysis are reported (SE, standard error); assay duration 20 days, $n = 30$; significant R/S differential toxicities are underlined (Breslow–Gehan–Wilcoxon test, $p_{BGW} < 0.001$), and complete R/S differential responses are boldfaced, with low toxicity scored as no or low mortality (–).

[c] Clusters from multidimensional analysis of toxicity profiles, on either quantitative data from this table (HCA hierarchical cluster analysis, and PCA see Fig. 1) or on qualitative data (presence/absence of toxicity for each strain × fraction, MCA, multiple correspondence analysis).

strain ($p = 3 \times 10^{-21}$) and finally *chemical* extract ($p = 9 \times 10^{-17}$), with no significant interactions (*species* × *strain*, $p = 0.21$ and *chemical* × *strain*, $p = 0.18$). As emphasised by the order of the factors, and by the absence of a global interaction with the weevil strain, this high variability in the biological activity of legume seeds was therefore not primarily related to A1b-like activities, assumed to affect differentially the two weevil strains. Legumes, including their seeds, are known to accumulate many compounds of a diverse chemical nature (Carlini and Grossi de Sa, 2002; Duffus and Smith, 1981; Wink and Mohamed, 2003). Our biological screening, therefore, clearly highlights other molecules of potential interest for the control of insect pests. However, another outcome of the global toxicity analysis was that all our extracts were more toxic to our reference strain (Bénin, PA1b-sensitive) than to the PA1b-resistant China strain. The only exception was the *Dialium guineense* H_2O5 fraction (Table 1), and this general behaviour may be taken as an indication of the greater robustness of the China strain (or aggressivity, *sensu* interaction genetics), which may have been selected unintentionally for resistance to other factors present in legume seeds during, or before, its rearing on pea seeds. This feature leads to the natural subdivision of our range of plants, for example after multiple correspondence analysis of the *qualitative* toxicity (MCA, see Table 1), into three well-separated groups of species: (1) those with either no, or only weak, effects (Table 1, Cluster L, Low toxicity), (2) those with a major effect in all of the fractions tested (Cluster H, High toxicity to both strains), and (3) those with a major differential effect on the two weevil strains (Cluster D, Differential toxicity affecting specifically the S strain).

A further look at the toxicity data from a multivariate *quantitative* point of view (*e.g.* through PCA and hierarchical clustering, Fig. 1) allows for an additional typology and the separation of our range of species into four groups (Clusters 1–4; Table 1: most "1" belong to the L cluster, all "2" to the H cluster, but the intermediate species are grouped in two additional clusters according to their quantitative toxic profiles). When looking at the variables used in such an analysis (see columns of Table 1, Fig. 1c), and their resulting transformation in factorial axes (Fig. 1b), the striking features are:

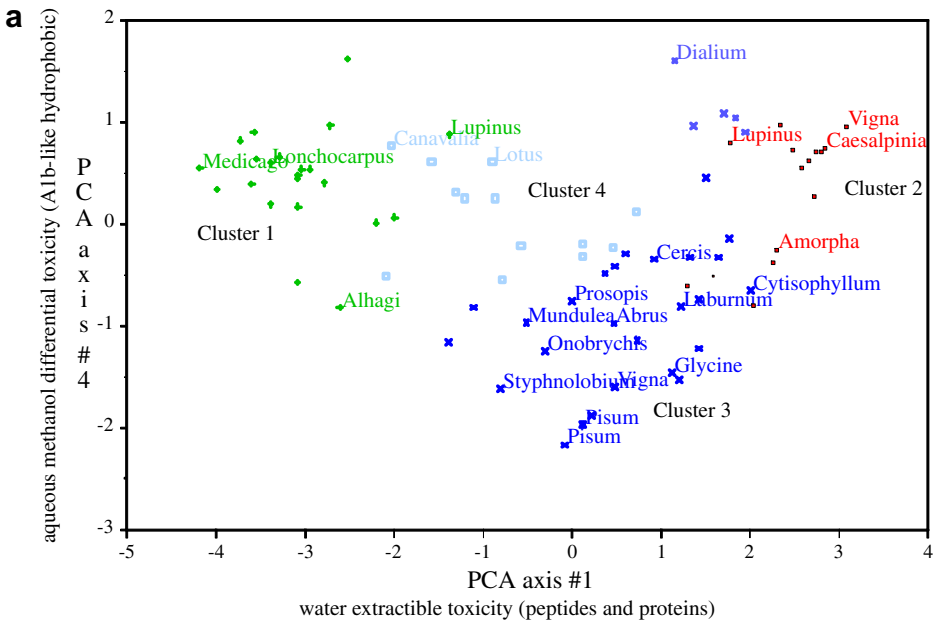
i. The strong chemical structuration of the toxic profiles, with axis 1 representing the global toxicity of the plant, resuming the effects of all the aqueous extractible components (Fig. 1b), axis 2 showing mainly the methanol-soluble information (Fig. 1b), and axis 3 representing the residual non-extractible compounds (Fig. 1b).
ii. The overall grouping of *R*/*S* variables within their chemical fractions (*i.e.* the global absence of interaction between chemical extract and weevil strain, as mentioned above), except for the aqueous fractions containing the archetypal A1b peptides (Fig. 1c).
iii. The detection, as a fourth level of variability (axis 4, Fig. 1b), of the activity related to the aqueous methanolic fraction (strict A1b-like activities), which clearly differentiated the weevil strain-specific effects (plants from clusters 3 and, more strictly, D, separated from cluster 4 plants, Fig. 1a and b, Table 1).

It is noteworthy that the aqueous fractions (MeOH60, H_2O5, H_2O8) are the most discriminating extracts between the plant clusters (Fig. 1a and b). These fractions contain mostly small proteins and peptides (as shown by amino-acid analyses in pea and some other species: H_2O5 and H_2O8 extracts contain more than 90% peptidic material). Therefore, such polypeptide components should account for a significant part of the insecticidal activities described here.

2.1.2. Phylogenetic analysis of toxicity and PA1b-like biological activity

The significance of the phylogenetic component in our toxicity indices was evaluated using a modified Abouheif test (1999). The phylogenetic structuration was not very strong on global indices (*e.g.* on axis 1–3 from PCA: $0.001 < p < 0.01$), or even on the characterised clusters ($p = 0.01$ for clusters 1–4) but was much more significant on some chemical extract variables, such as S1 or S3 (aqueous/methanolic extracts on susceptible weevil strain, $p = 1.0 \times 10^{-5}$, $p = 1.6 \times 10^{-4}$, respectively, see Supplementary material 3), or on the PCA fourth axis ($p = 3.0 \times 10^{-5}$), that is consistent with its strong correlation with the variable S1. This is an indication that global toxicity is a complex trait that has evolved many times in the Leguminosae, although some of its components (remarkably the canonical A1b peptides underlying S1 toxicity) are still phylogenetically recognisable.

Many seed fractions display PA1b-like activity (toxicity to S but not to R weevils), as observed in extracts from *Vigna unguiculata*, *V. radiata*, *Dioclea grandiflora*, *Desmodium canadense*, *Colutea arborescens*, *Onobrychis viciifolia*, *Melilotus albus*, *Trifolium fragriferum*, *Lens culinaris* (Phaseoloid and IRLC species) and also from *Styphnolobium japonicum*, *Amorpha fruticosa* (Sophoroid), *Abrus precatorius* (Loteae *sl*), *Laburnum anagyroides*, *Cytisophyllum sessilifolium*, *Baptisia australis* (Genistoid), or *Prosopis africana* and *Acacia dealbata* (Mimosoideae). Few Caesalpinioideae species display toxicity, and only two of them harbour a fraction with more toxicity to *S* than to *R* weevils (*Cercis siliquastrum*, *Tamarindus indica*). As the acute toxicity of pea seeds is due mostly to PA1b (Delobel et al., 1998; Louis et al., 2004), and as the resistance of *S. oryzae* to this toxicity has been shown to be a monogenic character (Grenier et al., 1997), we first hypothesised that a differential *R*/*S* mortality was due to A1b-like peptides. Thus the results of our bioassays alone indicate the potential presence of A1bs in species from the three subfamilies of Fabaceae.



b **Principal Components / Factor Analysis**

	#1	#2	#3	#4
Eigenvalue	4.80	1.60	1.14	0.78
Percent	48.02	15.98	11.36	7.85
Cum Percent	48.02	64.01	75.37	83.22
Eigenvectors				
S0	0.1518	**0.6485**	-0.2013	0.1343
R0	0.1857	**0.6360**	-0.0231	-0.0184
S1	**0.3188**	-0.0730	-0.1962	**0.6432**
R1	**0.3747**	0.0855	-0.2581	-0.1405
S2	**0.3881**	-0.2084	-0.0908	0.0995
R2	**0.3808**	-0.1213	-0.2035	-0.3848
S3	**0.3400**	-0.2645	0.0142	0.3077
R3	**0.3535**	-0.1453	-0.1547	-0.4094
S4	0.2782	0.0425	**0.6575**	0.2211
R4	0.2947	0.1062	**0.5897**	-0.2812

c **Variable clustering**

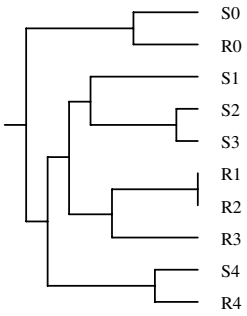


Fig. 1. Multivariate analysis of toxicity profiles of the 88 legume species. Table 1 data were submitted to a principal component analysis (PCA), from which clustering of species ($n = 88$) and variables ($n = 10$, strain × extract factors) were performed. (a): Factorial plane 1 × 4 showing species clustering (clusters 1–4, also reported in Table 1); (b): variable contributions to factorial axes 1–4, showing the global discrimination between water extractible factors (axis 1), methanol soluble factors (axis 2), residual non-extractible toxicity (axis 3) and aqueous methanolic differential *S*/*R* toxicity (axis 4, mainly contributed by hydrophobic peptides putatively linked to the A1b family); (c): variable clustering showing the specificities of methanolic and residual fractions (0 and 4, respectively, toxic to both strains), the contribution of water extracts to *R* and *S* strain discrimination (showing the chemical dispersion of differentially toxic factors between solvents 1–3), and the specificity of S1 variable (deep branching, corresponding to the canonical A1b peptide toxic phenotype).

To confirm that the differential toxicity observed in certain extracts is linked to the presence of A1b peptides, we looked both for PA1b homologous genes in the genome of various legume species and for homologous peptides in the corresponding extracts.

2.2. PA1b homologous genes from species of two Papilionoideae clades

2.2.1. Taxonomic distribution of the homologous genes

PCR with degenerate primers gave products of expected size (220 bp) for species from two clades of Papilionoideae, Hologalegina and Phaseoloid. Inside these two brother clades, a few species did not yield any PCR product. These were *Cicer* (which does not display any acute toxicity), and the five species tested within the Loteae–Robinieae clade (*Lotus corniculatus*, *L. tetragonolobus*, *Abrus precatorius*, *Sesbania sesban* and *Robinia pseudoacacia*). Despite several PCR attempts, no A1b sequences were characterised in any species outside these two clades. Nonspecific products (as checked by sequencing) were sometimes obtained with species from other clades or subfamilies.

These results indicate that A1b genes are widely present in at least two clades of the Papilionoideae. The presence of A1b genes correlated well with the observed toxicity, except for *V. subterranea* which does not exhibit any insecticidal

activity towards weevils. Interestingly, however, the gene characterised for this species might be a recent pseudogene as a stop codon was detected in the N-terminal sequence encoding the putative signal peptide. The absence of homologous amplification on DNA from species in the other clades may be explained by two hypotheses. Either A1b genes are restricted to these two related Papilionoideae clades, which contrasts with our bioassay results, or A1b genes from more distant species are too divergent to be recognised by our degenerate primers. The latter methodological limitation is the most likely, as our primers were designed using pea and soybean sequences (two species belonging to the IRLC and the Phaseoloid clade, respectively). Concerning the Loteae, which forms a distinct brother clade to the IRLC, a specific evolution (mutation in the primer anchor region, reorganisation of the gene structure, loss of the gene) may have occurred in the common ancestor, and could explain our negative results. It should be noted, however, that ESTs from *Lotus japonicus* do not identify any single A1b homologue from this species, despite comprehensive coverage of target tissues by cDNA libraries (growing seeds/pods). This differs from the extensive hits for A1b-like peptides in the two other legume species for which extensive genomic data are available: two different clusters exist in soybean, of which only one (leginsulin, TIGR TC225846) is the clear orthologue of PA1b, showing a dominant expression in developing seeds. In contrast, *M. truncatula* harbours more than 20 homologous clusters, none of which have been designated as clearly orthologous to PA1b by a high and specific expression in growing seeds (not shown, all analyses are from the TIGR plant gene indices database). From the available data, including the absence of full differential toxicity in all seeds from the Robinieae/Loteae, one likely hypothesis is that albumin A1b was present in the ancestor from Hologalegina and Phaseoloids and has lost its expression (and presence ?) in the Robinieae/Loteae clade but not in the IRLC (see supplementary material 1 for legume phylogeny).

2.2.2. Sequence variability

Nineteen novel sequences were characterised, corresponding to fourteen species from various tribes of the two clades (seven from IRLC and seven from Phaseoloids). Eleven were completed for their coding 5′ end, using the gene walking technique. For four of the species, two (*Bituminaria bituminosa*, *Canavalia brasiliensis*, *Onobrychis viciifolia*) to three (*Mundulea sericea*) different genes were sequenced. The deduced protein sequences are aligned in Fig. 2.

The sequences exhibit a high intra-family variability, with percentages of nucleic acid identity ranging from 61% (between soybean and *Alysicarpus ovalifolius*) to 82% (between pea and *Vicia hirsuta*). Some domains of the deduced proteins are more variable in nature and in number of amino acids, such as the segment between the

```
Spe ID (AC-EMBL)[a]                                nC1   CC2    CC3   CC4 CC5    CC6       Cc7 (PA1b segments)
                                  <-------   signal   -------> <-----------------   PA1b   ---------------->   propep   <-----------   PA1a   ----/partial
Psa P08687 (M13709)               MASVKLA SLIVLF ATLGMFLTKNVGA ASCNGV   CSPFEMPPCGTSA CRCIPVGLVV  GY  CRNPSG     VFLRTN DEHPNLCESDADCRKKGSGNFCGHYPNP
Psa Q40999 (M81864)               ------- ------ ------------- ------   ---------S-- --------LI  --  ------     ---KG- ------------K-------------
PsaA1b005[b] P62931                                             ------   ---------S-- --------I  --  ------
PsaA1b015[c] (AJ574796)                                         I-----   ----DI----SPL -----A---I  -N  ----Y-
PsaA1b014 (AJ574795)              ------- ------ ------------- ------   ---------S-- --------LI  --  ------     ---KG- ------------K-------------
PsaA1b012 (AJ574794)              ------- ------ ------------- ------   ---------S-- --------FI  --  ------     ---KA- -------------------------
LlaA1b001 (AJ784955)                                                     -----V---RSKD ----------  --  --H---     I----T -K-------------E----------
VhiA1b001 (AJ784951)              --Y-Q-- Y-V--- --F-----Q--E- -E----   ---------S-D ----------  --  --Y---      LRN-I E------K-NL--IE-E--K--------
TfoA1b001 (AJ784950)              --Y-R-THLVVF-L S-FSL-PM-K--- TD-S-I   ---------RS-D -----IV--G  --  -I--IS P   AATKMV K------H-HT--T-----S--AR----
MalA1b001 (AJ784946)              --Y----PLVVF-L -AFSI-PMMKGE  EE-S-I   --S------RS-S ------V-LG  -N  -VD--S P   TIMKMV E--S---Q-HFE-M-----S--AR----
MtrA1b006 (AJ574789)              --YIRF-HLVVF-L -AFSLVP--K--- TD-S-A   ---------RS-D -----I---A  --  -TY--S P   TVMKMV E------Q-H---T--E--S--AR----
MtrA1b007 (AJ574790)              --YLR--HLVVF-H --FSLIFPMMKA- ED-S-I   ---------PS-S ------ILIG  -N  YVD--S P   TITKMV EK-A---Q-H---T-----S--AR----
AmoA1b001 (AJ784941)              -G-----PLALF-- --FL--PA---E- TD-S--   ---------STD -----W--F-  -E  -KY-T-    SERVTKMV EK-----Q-HD--VH-E--S--AR---A
CarA1b001 (AJ784953)                                                     -----V----S-S -----W--FW  -T  --Y---    SGRVIKKV G--------H---M--E--S---R----
OviA1b002 (AJ784949)              -PNYFKLTP-VLFLI--FL--PM-K-E-   -D--   ---------STD ---V-W--F-  -Q  -I--TS     ATVEKMV K------Q-HV--ME----S---R---A
OviA1b001 (AJ784957)                                                     -----V---RS-D -----IA--M  --  --Y--S VSNDAVT-MA K------Q-H---KN-R--S--AR----
PvuA1b002                                                                ---------S-D ---V-Y--F-  -S  -IH-T-    LSAAAKMI -------QPHEE-M--------AR---H
PvuA1b001 (AJ574792)              --N-RV-PLALF-L --SI--PM-KTE- VV-S--   -----R----S-RD-----Y--FI  -A  -TY-S-    LSSVAKTI -----I-QTHEE-T--------AR---D
VsuA1b001 (AJ784958)                    (stop?)               -E-S-A   -----V----SRD -C---I--F-  -F  -IY-T-    LSSVSKMI -------Q-HDE-M--------AR---N
VunA1b001 (AJ784959)                                                     -----V----SRD ----------  -F  -IY-T-    LSSVAKMI E------Q--DE-V--------AR---N
GmaA1b005 (AJ574791)                    MAVF-L --STIMFPTKIE- -D---A   ---------RSRD ---V-I---A  -F  -IH-T-    LSSVAKMI -------Q--DE-M--------AR---N
CbrA1b001 (AJ784944)              --YA-F-PLAVF-- --FL-LSM-KIE-   -S-G   ---------S-D -----W---A  --  -I--TS    LAAAAKMI EA--H--Q-NDE-L--------ARF---
CbrA1b002 (AJ784954)                                                     ---------SGE ---V-A----  -F  -T--TS    LASAAKMI EA-LH--Q-NDE-L--------ARF---
AovA1b001 (AJ784942)                    MEAVFWYSL   L-T---IE- -D-S-A   -----R-L--STD -----IV-LA  -F  -IY-T-    LASVANLI -N-----KTHDE-I--------ARF---
BbiA1b001 (AJ784952)                                                     ---------STD ---V-W--FA  -Q  -ID-I-  GVAS-AKMI -D-----Q-NED-I--R--D--AR---N
BbiA1b002 (AJ784943)              -VC-RVPPLALF-L --SL--S--KIE-   -S-A   -F--Q-----STD ---V-W--F-  -Q  -ID-I-  GVAS-AKMI -D-----Q-NED-I--R--D--ARF--N
LcaA1b001 (AJ784945)              --Y-R-VPLAVF-L --SV-LPM-IK   -----RDV ---------DDATN-----W----  -Q  -VH---    LPSVAKMI -K-L---Q-NDE-M--------AR----
MseA1b003 (AJ784948)              ---LRV-PLAVFFL --SV--PMIR    GD---RVV ---------DAQK-----YL-AG  -L  YTP-TP     SVT-MI -Q-----Q-NDE-M--------AR----
MseA1b002 (AJ784956)                                                     ---------DAQN-----IL-FG  -L  -IH-T-     LDTKMI -Q-----Q-NEE-T--------AR----
MseA1b001 (AJ784947)              ---LRV-PLAVFFL --SV-LPMTR    G----RDW ---------DAQN------V--G  --  --H---     LDTKKI -Q-----Q-NDE-M--------AR----

Van Q9FRT9 (AB052880)                                                    A-D---A   ----Q-----STD -L---A--LFV --  -TY---    LSSVAKMI -------Q--DE-M--------AR---N
Vra Q9FRT8 (AB052881)                                                    A-D---A   ---------RSTD -----IA-FG  -F  -I--T-    LSSVAKMI -------Q--DE-L--------AR---H
Gso (AJ011935)                          MAVF-L --STIMFPTKIE- -D---A   -----V---RSSD ---V-I--F-  -F  -IH-T-    LSSVAKMI -------Q--DE-M--------AR---N
Lan U74383 (Q96474)                                                                 I--SRS-D ---V-IT-I-  -F  -IH-T-    LSSVAKMI -------Q--DE-M--------AR---N
Mtr MTR389043c (EST, AJ389043)    -TY----ILAVLHL TIFLI-Q----E- ---PN-GAV-----TK--NVKD---L-W--FF  -T  -I--T-    SKYNMKMI E------QTHGE-I--------AR-A-A
Mtr TC18698 (EST, AL388929)       --N----PFAVFCL -AFL--PM-KIEG E--ESRG  -IFYINDS-PSG  -V-D-IDP-TWA-V  -VSYS      SIKKKV E----Y--THTE-T--------ARFA-S
```

a : Species/isoform identification : the three first letters identify plant species (see Table 1) ; published accession identifiers are Swissprot IDs or EMBL accessions (between parentheses).
b : isoform labeling, as introduced in the present work and published in the corresponding EMBL accessions (*e.g.* : MtrA1006 retrieves the corresponding entry from EMBL ; note : PvuA1b002 unpublished).
c : variant pea protein sequence published as a separate isoform (peptide sequenced) in the original paper by Higgins *et al.* (1986), recently cloned and deposited as PsaA1015 isoform (EMBL:AJ574796).

Fig. 2. Sequence alignment of reference PA1b peptides, and cloned genes from the studied legume species. First column indicates peptide identification (isoform labels/EST TC and Swissprot/EMBL accessions), and plant source (3-letter identification, see Table 1; additional plant species: Gso *Glycine soya*, Van *Vigna angularis* and Vra *Vigna radiata*). First line indicates canonical (pea-type) processing of the pre-propeptide (Higgins et al., 1986). Conserved bridged cysteine residues are highlighted (1–4, 2–5, 3–6 knottin-type bridges; relative cysteine numbering within PA1b) and define different fragments, indicated above the alignment, from the N-terminal amino acid (nC1) to the C-terminal one (Cc7).

first two cysteines of PA1b (indicated as CC2 on the alignment), which displays two extra residues in the sequences from the Millettieae species. The segment between the third and fourth cysteine (CC4) is particularly divergent, being sometimes highly charged (RSKD, GDAQK) in contrast to PA1b (GTSA). The linking peptide between A1b and A1a is the most variable part of the deduced proteins, with one to four extra amino acids of different kinds, as compared to sequences from pea. However, some amino acids are particularly well conserved, such as the structural cysteines, prolines and some glycines, as well as some bulky charged (R21, as numbered from the PA1b sequence) or hydrophobic (L27) residues.

The structure of the A1 genes appears to be conserved, with only one putative intron in the coding sequence of the signal peptide. Intron length varies from 75 nucleotides for *Vicia hirsuta* (close to the 83 nucl. observed for pea) to 298 nucl. for *Melilotus albus* (within the range of 271–304 nucl. characterised for *Medicago truncatula*). The position of this intron is highly conserved between the first and the second nucleotide of the codon corresponding to the 18th amino acid of the signal peptide. Exceptions are *V. hirsuta* (17th codon, as for pea), *O. viciifolia* (19th) and *A. ovalifolius* (9th). However, for all these species, the definition of the signal peptide needs to be confirmed experimentally by cDNA analysis.

2.2.3. Phylogenetic analysis of A1b sequences

A distance-based nucleic acid phylogeny was performed on the common partial sequences obtained (most of PA1b plus part of the PA1a coding sequence), including the few sequences already published. The unrooted tree obtained is shown in Fig. 3. Even if the sequences are short, and therefore include a limited set of informative sites (104), the resulting tree clearly shows the grouping of genes from species of the same tribe, as for the Vicieae (Psa, Vhi, Lla), the Trifolieae (Mtr, Mal, Tfo), the Millettieae (Mse, Lca) and the Phaseoleae (Pvu, Vra, Van, Vsu, Gso, Gma, Cbr). It is also striking that sequences from the same species cluster together, except for *Medicago truncatula* whose sequence MtrA1-7 appears to be closer to the *Melilotus albus* gene than to MtrA1-6. Globally, the congruence with species phylogeny is high and this indicates that the sequenced genes might be recently diversifying orthologues, as exemplified in pea for which at least six very similar gene variants are present (Louis et al., 2004). One has to keep in mind that this view is probably methodologically biased by our PCR-based approach, for which the primers

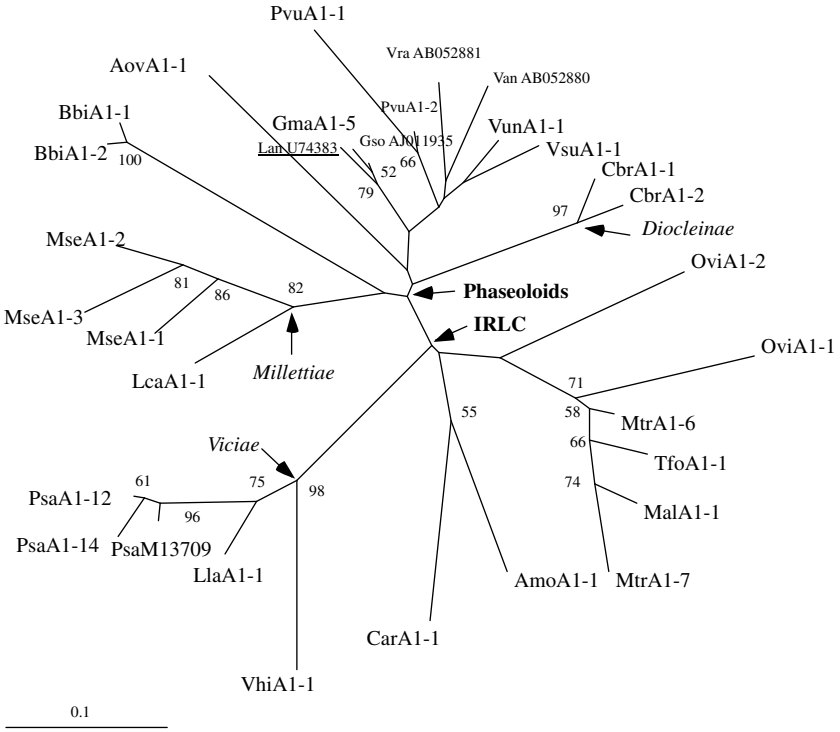


Fig. 3. Phylogenetic analysis of the A1b/leginsulin family. Molecular phylogeny built from a BioNJ/Galtier-Gouy distance method on the nucleic acid sequences of the partial protein alignments shown in Fig. 2 (104 informative sites out of the 160 complete, 500 bootstrap replicates; only bootstraps higher than 50% are reported at corresponding nodes). Isoform numberings are simplified from Fig. 2. Species codes as follows (see Table 1, and Supplementary material 1 for phylogenetic relationships; identified clades are reported at corresponding nodes): Psa = E23, Lla = E26, Vhi = E55, Car = E30, Amo = E13, Ovi = E38, Mtr = E83, Tfo = E20, Mal = E19, Cbr = E43, Vun = E11, Vsu = E24, Pvu = E12, Aov = E33, Bbi = E33, Mse = E44, Lca = E18.

probably target a specific subset of homologues. The occurrence of more than 20 A1b-like genes in *M. truncatula* is an indication of sustained gene duplications in this lineage. Since the number of PA1 transcripts/genes within species is quite variable (from no expressed gene in *L. corniculatus*, 1–2 in soybean, 6-7 in pea and more than 20 in *M. truncatula* – data from EST analysis –), it may be unlikely that the identification of single pseudogenes, such as in *V. subterranea*, would just lead to a complete lack of activity; this could also reflect post-transcriptional gene silencing and warrants further investigation.

The phylogenetic analysis also highlights the odd positioning of the lupin sequence U74383 (Ilgoutz et al., 1997), which is clustered with soybean although the species belongs to a quite separate and more basal clade (Genistoid). Despite many attempts, including the use of the same primers, PCR conditions, and plant genotype as described in Ilgoutz et al. (1997), we were not able to amplify the expected product from the DNA of *Lupinus angustifolius*.

Finally, the limitations of this molecular approach obliged us to use a biochemical approach to check for the presence of A1b peptides in the seeds of species from the two positive Papilionoideae clades, and to test for their potential presence outside these two clades.

2.3. Peptide characterisation in hydrophobic peptide fractions

To confirm the presence of A1bs in some extracts displaying PA1b-like activity, and belonging to species inside and outside the two clades of Papilionoideae from which homologous genes had been characterised, such extracts were analysed by mass spectrometry as described earlier (Louis et al., 2004). Major detected masses (pick intensity >10% of major pick intensity) are reported in Table 2, and a representative selection of mass spectra is shown in Supplementary material 4.

2.3.1. Extracts from IRLC and Phaseoloid species: agreement between gene-peptide presence and biological activity

A1b peptide like masses were detected in the control extract MeOH60 of *B. bituminaria*, from which one complete A1b sequence was characterised. Among these masses, one corresponds to the deduced product of this gene and one to the same peptide lacking a glycine. Such variation in peptide maturation at the C-terminal end (or even at the N-terminal) has already been described in pea (Higgins et al., 1986; Louis et al., 2004; Taylor et al., 2004b).

The mass profile comparison of *V. unguiculata* and *V. subterranea* MeOH60 extracts also confirms the result of the bioassays. Indeed, whereas masses corresponding to PA1b-like peptides were detected in the *V. unguiculata* MeOH60 fraction, none were ever observed from the *V. subterranea* extract (Supplementary material 4, I and J). This fraction does not exhibit any activity on the two weevil strains, in contrast to *V. unguiculata* extract which is highly toxic for susceptible insects. This is also confirmed by gel electrophoresis of the two fractions, where a 4 kDa band is observed from *V. unguiculata* and not from *V. subterranea* (Table 2). The conclusive interpretation of the whole set of data from *V. subterranea* is that this species, unlike its sister taxon *V. unguiculata*, harbours a signal-peptide mutant of the A1b gene (Fig. 2) that precludes its efficient expression in seeds (Supplementary material 4) and results in the absence of toxicity (Table 1). This simplified situation (or interpretation) is, however, a good example of the evolutionary, and ecologically significant, variations that might occur in the plant peptide defence of legume seeds.

2.3.2. Extracts from other legumes: A1b and other compounds responsible for differential toxicity

Concerning species outside the Hologalegina and Phaseoloid clades, the mass spectrometry results are noticeably different. Indeed, whilst no typical PA1b-like masses were observed from species in the two subfamilies Mimosoideae and Caesalpinioideae, as well as from species of the Aeschynomenoid clade (Papilionoideae), significant mass signals ranging from 3600 to 4100 Da were clearly detected from extracts of a Genistoid and a Sophoroid species (*Laburnum anagyroides* and *Styphnolobium japonicum*; Table 2, Supplementary material 4 H and C to F). As Sophoroids are actually basal Papilionoideae, these results tend to confirm the wide representation of A1bs among Papilionoideae, but also indicate that this peptide family might be restricted to this subfamily. The absence of A1b in seeds from the Caesalpinioideae was to be expected, as many of the species tested display no toxicity to the weevils. The differential toxicity observed for *Cercis siliquastrum* may be due to unrelated compounds, indicating a probable multiple resistance of the China strain. This explanation is likely to apply to *A. julibrissin* too. Indeed, fractionation of MeOH and MeOH60 extracts of *A. julibrissin* and MeOH60 of *C. siliquastrum* by HPLC leads to five groups of peaks, according to their retention time (data not shown). The first peak (RT 4–5 min), very polar and unlikely to contain hydrophobic peptides, was the group bearing all the differential toxicity in these and other non-Papilionoid species, whereas PA1b isoforms elute normally in group 5 (RT 18–20 min), as was observed for the *B. bituminosa* MeOH60 fraction. This HPLC fractionation was coupled, for a limited set of species, to a PA1b-receptor binding-inhibition assay (Gressent et al., 2003), as already used in this screening context (Louis et al., 2004). Globally, this experiment only identified IRLC (*Lathyrus*) or Phaseoloid (*Bituminaria*, *Vigna*) fractions that were able to compete unambiguously with the insect PA1b binding-site (Table 2), and demonstrated that many partially purified fractions were clearly inducing differential *S*/*R* toxicity, without being able to compete at the molecular level with the pea albumin PA1b (and thus considered as non-homologous toxicity). Altogether, these data also highlight the

Table 2
Biochemical and associated molecular data obtained on species/fractions from the analysed panel of legume species

Sub-family/ Clade	Species	Fraction	Sequence[a] *c/p/−*	Binding assay[b]	SDS–PAGE [c]	MALDI-TOF[d]
IRLC, Phaseoloid	*P. sativum, G. max, P. vulgaris,*	MeOH, MeOH60	*c*	+	+	yes, see Louis et al., 2004
IRLC	*M. truncatula*	MeOH, MeOH60	*c*	–	–	no hit, see Louis et al. (2004)
IRLC	*T. foenum-graecum*	MeOH60	*c*		+	no conclusive hit in spectrum range 1000–10000
IRLC	*T. foenum-graecum*	TA2B5C4	”			not conclusive
Phaseoloid	*B. bituminosa*	MeOH60	*c/p*	+	+	2576; **3629; 3644; 3685; 3701; 3806; 3863**; 4151
Phaseoloid	*B. bituminosa*	H2O8	”			1647; 2132; 3014; **3684; 3716; 3907**; 4385; 7370; 9472
IRLC	*L. latifolius*	H2O8	*p*	+	+	**3894; 3931**; 5188; 6808; 6831; 6849; 6873; 6887; 7295; 7410; 7868
Abreae/Rob-Lotae	*A. precatorius*	MeOH60	–		–	3182; *4094*; 4208; 5426; 7972
Aeschynomenoid	*D. purpurescens*	MeOH60	–		–	1129; 5396; 5497; 5561; 5585; 5605; 5665; 5684; 5752; 5765; 5852; 5981
Sophoroid	*S. japonicum*	MeOH	–			**3767; 3821; 3866; 3879; 3917; 3980; 3995**
Sophoroid	*S. japonicum*	MeOH60	”		+	*3614; 3629; 3806; 3822; 3851; 3866; 3879; 3918; 3936; 3954; 3965; 3980*
Sophoroid	*S. japonicum*	H2O5	”			2374; 3050; **3613; 3628; 3639; 3686; 3696; 3726; 3742; 3804; 3864; 3963; 3978**
Sophoroid	*S. japonicum*	H2O8	”			**3641; 3744; 3806; 3821; 3864; 3879; 3918; 3964; 3980**; 7138; 7282; 7397
Aeschynomenoid	*A. fruticosa*	MeOH60	–		–	3984; 4062; 4071; 7972; 8129; 8144; 8279
Caesalpinoidae	*C. siliquastrum*	MeOH60	–	–	–	no hit, no peptide in spectrum range 1000-10000
Caesalpinoidae	*C. bonduc*	MeOH	–			no hit, no peptide in spectrum range 1000-10000
Mimosoidae	*P. africana*	H2O5	–			2045; 2157; *3655; 3737; 3748; 3842*; 7310; 7496
Genistoid	*C. sessilifolium*	MeOH60	-		+	**3711; 3727; 3752;** *3765; 3872*
Genistoid	*L. anagyroides*	MeOH60	–			**3709; 3724;** *3762; 3886*
Mimosoidae	*A. julibrissin*	MeOH60	–	–	–	2774; 5119; 5131; 5133; 5134; 5478; 5549; 5564
Phaseoloid	*V. unguiculata*	MeOH60	*p*	+	+	**3645; 3691; 3707; 3729; 3748; 3764; 3785; 3801; 3838**; 5406
Phaseoloid	*V. unguiculata*	H2O5	”			**3647; 3692; 3707; 3749; 3765; 3787; 3801**; 5366; 5390; 5407
Phaseoloid	*V. unguiculata*	H2O8	”			2096; **3691; 3707; 3748; 3764; 3801**; 4396; 5405
Phaseoloid	*V. subterranea*	MeOH60	*c*	–	–	2737; 4242; 5434; 5475
Phaseoloid	*V. subterranea*	H2O5	”			2738; 4456; 4514; 5436; 5459; 5475; 6135; 8914; 9011; 9028
Phaseoloid	*V. subterranea*	H2O8	”			2737; 5434; 5458; 5474; 8914
IRLC	*M. albus*	MeOH	*c*	–	–	no hit, no peptide in spectrum range 1000-10000
IRLC	*M. truncatula*	H2O5	*c*	–	–	**3607; 3624; 3654; 3906; 3972**; 5118; 5190; 6891
IRLC	*M. truncatula*	H2O8	”	–	–	4476; 4520; 4528; 4538; 4548; 4555; 4559; 4570; 4597; 4605; 4618; 4625; 4633; 4649; 6637; 6652; 6676

[a] *c* = complete A1b, *p* = partial A1b (N-terminal missing), – = no positive hit by genomic PCR.
[b] Positive results in radiolabelled-ligand (^{125}I PA1b) binding competition assays, on whole fraction and further HPLC-purified peak pools.
[c] Peptide signal in the expected mass range in silver or Coomassie-stained electrophoreses (SDS–PAGE or Tris–Tricine gels).
[d] In bold, major masses (>10% max signal) falling in the A1b 3600–4100 Da range; in italics, minor masses in the A1b range.

existence of other toxic compounds, much more polar than A1bs and unrelated to them, which induce similar biological effects and may be of future interest for insect control.

Concerning the Aeschynomenoid species (*Amorpha*, *Dalbergia*), the negative MS results are more surprising. The presence of A1bs in the basal group of Papilionoideae indicates that they could be present in the ancestor of all the current Papilionoideae species, including Aeschynomenoids. It is, however, possible that in the course of evolution the A1b gene was lost in the ancestor of the existing Aeschynomenoid species, as no 4 kDa peptide was detected by electrophoreses of the two tested species (*Dalbergia purpurescens* and *Amorpha fruticosa*, Table 2). This situation has already been discussed at the gene level for the Robinieae/Loteae clade to which *Abrus precatorius* is presumed to belong (Wink and Mohamed, 2003), a species which also seems devoid of positive mass hits (in the range 3600–4100) in its active MeOH60 fraction (Table 2).

3. Concluding remarks

The main objective of our work was to clarify the origin of the seed-accumulating albumin 1b toxins (A1b) within the legume family, to which these cystine-knot peptides seem to be confined. In contrast to other sulphur-rich peptides, such as the Bowman-Birk or the γ-thionin families (Broekaert et al., 1997), sequences homologous to the PA1b-leginsulin archetypes had, up to now, never been found outside a limited set of species within the Vicieae or the Phaseoleae tribes. The present work unambiguously ascribes the origin of this peptide family to a common

ancestor of the Hologalegina and the Phaseoloid clades, which date back to 35–40 MY ago in the history of the Papilionoids (Wojciechowski, 2003). These sequence-based data add many new tribes (Psoraleae, Millettieae, Desmodieae, Hedysareae for example), from 16 species, to the previous A1b containing clades. In addition, an enlarged set of results, including those of canonical mass spectral hits and seed extract/bioassay behaviour, tentatively ascribes the origin of the family to the basal Sophoroid group. Putative standard peptide mass spectra from aqueous methanolic extracts of the sophoroid *Styphnolobium japonicum* (Supplementary material 4) would date their origin to the early eocene period, around −50 MY (Wojciechowski, 2003). Due to intrinsic limitations of the PCR-based approach, this claim can only be confirmed by future peptide purification and sequencing.

In addition to these A1b positive hits, our work also provides different sets of evidence that the A1b family was (i) probably not present in non-papilionoid species (even though fractions were identified with very clear differential insecticidal activity on PA1b-sensitive and PA1b-resistant strains), and (ii) was probably lost at different taxonomic levels within the papilionoid clade (Aeschynomenoids, Robinieae/ Loteae, or even lost at the protein-expression level in a single species, such as *Vigna subterranea*).

Finally, this wide variation in A1b-based activity is only a small part of the global variation in insecticidal compounds present in legume seeds (Louis, 2004). Again, at different taxonomic levels (Caesalpinoideae, Mimosoideae, Genistoids, "Trifolieae"), species with extreme intensities, or a large chemical range, of bioactive compounds, unrelated to A1bs, are highlighted by our survey (Cluster 1 on the left of Fig. 1a).

4. Experimental

4.1. Insects

Rice weevils (*Sitophilus oryzae*, Coleoptera, Curculionidae) were reared on wheat seeds at 27.5 °C and 60–70% r.h. Two strains were used, differing in their genetic ability to thrive on pea seeds and resist the toxic activity of pea albumin PA1b: a control susceptible strain "Benin" (S) and a fully resistant strain "China" (R) harbouring the recessive pea-resistance allele (Grenier et al., 1997).

4.2. Plant material and peptide extractions

Plant species were chosen in order to get the largest phylogenetic coverage of the different tribes in the three Fabaceae subfamilies (Polhill, 1994; Wojciechowski et al., 2004). Concerning Papilionoideae, the largest subfamily, we considered five major groups, namely the basal paraphyletic Sophoroid (Swartzieae, basal Sophoreae, Aeschynomeneae, Amorpheae), the Genistoid clade (including genistoid Sophoreae, Podalyrieae, Thermopsideae and Genisteae), the Phaseoloid clade (including Indigofereae, Millettieae, Desmodieae and Phaseoleae), the Hologalegina clade including two subclades, Loteae sl clade (including Loteae, Robinieae and Abreae), and its brother clade called IRLC (including Vicieae, Trifolieae, Cicereae, "Galegeae", Hedysareae) for lacking an inverted repeat of 25 kb in the chloroplast DNA (Wojciechowski et al., 2004). Information on seed providers is available upon request. Two grams of shelled seeds were crushed and extracted, as described in Louis et al. (2004), adjusting the volumes of solvents to 20 mL. After extraction with pentane, five fractions were obtained which were submitted to biological assays on both susceptible (Benin) and PA1b-resistant (China) *S. oryzae* strains: MeOH (fr. 0: soluble in methanol 100%), MeOH60 (fr. 1: soluble in aqueous methanol 60%), H_2O5 (fr. 2: soluble in water pH 5), H_2O8 (fr. 3: soluble in water pH 8), residue (fr. 4: insoluble). Fraction numbering is the label used in our biological data analysis (*e.g.* Fig. 1). A few seeds of each species were also germinated in a greenhouse to obtain young leaf material used for DNA extraction.

4.3. Bioassays

Insects used for bioassays were adults aged 2–3 weeks, collected from experimental 1-week cohorts and deposited in batches of 30 individuals (for each *S* and *R* strain) on food pellets incorporating tested flour fractions in a whole-wheat based diet, as fully described in Louis et al. (2004). Concentrations of the fractionated material were always given as % equivalent of their relative abundance in the original legume flour (% total meal equiv., TME); the full screening presented in this work was performed with fractions incorporated at 100% TME. Bioactivity was evaluated by scoring daily insect survival during the first two weeks of contact with test food (27.5 °C, 70% r.h.), and through standard survival analysis, allowing LT50s to be calculated (lethal time 50%, or median life duration; Statview software, actuarial analysis and associated non-parametric statistics).

4.4. Bioassay data analysis and link to Fabaceae phylogeny

The primary data (Table 1) was used either as presented (88 species/lines, 10 toxicological variables) for quantitative multivariate analysis (PCA, principal component analysis; or MCA, multiple correspondence analysis after transformation of the 10 variables in 5 nominal classes), or as a measure of a response variable (toxicity outcome, either qualitative or quantitative) to a set of 3 factors (plant *species*, *chemical* extract or weevil *strain*). Data were analysed with the JMP statistical software (SAS Institute) or with the ADE4 statistical package (Thioulouse et al., 1997).

According to their global toxic typology, plant species were grouped into a number of classes/clusters depending on the multivariate analysis performed. These classes, as

well as raw or composite indices of toxicity, were mapped on the phylogenetic tree of the Fabaceae (see below) according to specific analyses and statistical tests (Abouheif, 1999; Ollier et al., 2006) using the R-package and its ade4 library (Chessel et al., 2004).

We considered the phylogenetic relationships between the tribes of the three subfamilies according to the work of Wink and Mohamed (parsimony phylogeny of the *rbcL* gene, 2003), completed with the Wojciekowski analysis of the IRLC clade (2003) and, for a few taxa, with other literature sources (Chappill, 1995; Doyle and Luckow, 2003; Hu et al., 2000; Kass and Wink, 1994). Our reference consensus tree is available as supplementary material (item 1 – dendrogram – and item 2 – numerical Newick tree –). Some connections are still controversial but will not be considered further (for example Aeschynomenoids clade sometimes placed between Mirbelieae and Genisteae, or positioning of *Bolusanthus*).

4.5. HPLC and binding assays

Reverse-phase HPLCs were performed on a Nucleosil® 300 C18 column (250 × 4.6 mm; 5 μm particle size, 300 Å porosity). Proteins were eluted at 1 ml/min with a 22 min gradient from 20% to 60% acetonitrile in water (0.1% TFA), and monitored by UV diode-array detection between 210 and 350 nm. Fractions were collected manually for mass spectra analyses, and for binding inhibition assays. For the latter assays, a scale-up procedure was used for semi-preparative HPLC, on a column of 10 mm sectional size and a 5 mL/min flow rate. Binding inhibition assays were performed as described elsewhere (Gressent et al., 2003; Louis et al., 2004).

4.6. Mass spectrometry

Two hundred microgram of target plant fractions containing toxicity and the potential A1b peptide (mainly MeOH60 fractions, see results) were submitted to mass spectrometry on a Voyager DE-PRO spectrometer (PerSeptive Biosystems, Farmingham, MA, USA). Positive ion mass spectra were recorded in the linear mode of this time-of-flight MALDI mass spectrometer. All mass spectra were externally calibrated with a calibration kit (Pep Mix 2, LaserBio Labs, Sophia Antipolis France). Samples were mixed with a sinapinic acid matrix (3.5-dimethoxy 4-hydroxy-cinnamic acid , LaserBio Labs) at a ratio ranging from 1:1 to 1:10, then spotted onto the target, dried, and submitted to MALDI-TOF analysis.

4.7. DNA extraction and degenerate primer genomic PCR

DNA was extracted from fresh leaves using a CTAB based protocol (Ausubel et al., 1997) adding, during grinding, 15% (w/v) polyvinylpolypyrrolidone (MW = 40,000) to eliminate phenolics (Page, 1999). PCR amplifications (35 cycles : 94 °C 30 s, 50 °C 45 s, 72 °C 45 s) were performed on 10 ng of genomic DNA with for/rev pairs (final concentration of 3 μM each) of the following degenerated primers, designed from sequences of soybean and pea, either in the A1b (For 1) or the A1a peptides (Rev 1 and Rev 6): For 1: $^{5'}$TGYTCICCITTYGARRTICCICCITG$^{3'}$, Rev 1: $^{5'}$CRAARCACCAICCRTAITCIATRTM$^{3'}$, Rev 6: $^{5'}$GCTTCAGAWTYAGAGKCAAARCACCAICCRTA$^{3'}$. PCR amplified fragments were purified, ligated to pCR2.1 vectors and used to transform *E. coli* Top10 electrocompetent cells. Recombinant plasmids were purified and inserts sequenced (Genome Express, Grenoble France). Four clones were sequenced from each first round of PCR in each plant species to select potential variants. For a few species, RNA was extracted from mid-growth seeds (Salzman et al., 1999), to perform Northern-blot analyses.

4.8. Genome walking for cloning of A1b 5′ ends and sequence analysis

Gene walking was performed using the Universal GenomeWalker™ kit (Clontech, USA) with primers designed from specific parts of the sequences obtained in the first degenerated PCR round. Each genomic DNA was digested by four restriction enzymes (*Dra*I, *Eco*RV, *Stu*I, *Pvu*II). Adaptors were ligated to the restriction fragments and two successive nested PCR amplifications were performed on the products of ligation, following the manufacturer's instructions. The major PCR product obtained over 600 bp from the four restriction libraries was cloned, and two clones were double-sequenced for each product. Sequences were assembled and analysed with the MacMolly software, checked for sequencing errors, for agreement with first-round PCR results, annotated for cds, intron position and signal peptide positions (using SignalP, Netgene2, Netstart at http://www.cbs.dtu.dk/, and manual refinement with published pea and soybean genomic sequences as templates), and finally submitted to EMBL with the following accession numbers: EMBL: AJ784941–AJ784959. A loci/allele identification system was defined as follows: GspA1nnn (Genus species, Albumin1-like peptide, isoform numbering running after any already published sequence, uniquely identified by either nucleic or amino acid 100% identity); for example, an all-text EMBL search for OviA1002 will retrieve Albumin 1 isoform 2 from the species *Onobrychis viciifolia*. Search for A1b homologues in model legumes was performed through the TIGR plant gene index portal (http://www.tigr.org/tdb/tgi/plant.shtml).

4.9. Phylogenetic analyses

Protein and associated nucleic sequences were retrieved from the EMBL database. Protein alignments were performed using ClustalW and refined manually, with the Sea_View software (Galtier et al., 1996), essentially for conserved cysteine positions as they appear in Fig. 2. The resulting nucleic alignments were analysed with the

Phylo_win software (Galtier et al., 1996), and trees generated through a BioNJ algorithm (Gascuel, 1997) associated with a Galtier and Gouy distance (Galtier et al., 1995).

Acknowledgements

We are grateful to Dr. Christophe Chambon (INRA Clermont-Theix) for MALDI-TOF/MS analysis. We are also indebted to Prof. Daniel Chessel (UCB Lyon 1 Univ.) for his invaluable help with the statistical analysis of phylogenetical structures in our bioassay data. We thank Dr. Valerie James for the language revision. We fully acknowledge financial support from INRA Plant Health (SPE) department to Y.R., and from INRA (Departments SPE and CEPIA) for a PhD Grant to S.L. We are also indebted to all the providers of seeds and taxonomic information on the plant material used in this work (contact Bernard Delobel, our local botanical expert; please e-mail to the corresponding author, or directly to delobelbe@wanadoo.fr). We thank an anonymous reviewer for his valuable suggestions on the present manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.phytochem.2006.11.032.

References

Abouheif, E., 1999. A method for testing the assumption of phylogenetic independence in comparative data. Evol. Ecol. Res. 1, 895–909.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., Struhl, K., 1997. Current Protocols in Molecular Biology. John & Wiley, New York.

Broekaert, W.F., Cammue, B.P.A., Debolle, M.F.C., Thevissen, K., Desamblanx, G.W., Osborn, R.W., 1997. Antimicrobial peptides from plants. Crit. Rev. Plant Sci. 16, 297–323.

Carlini, C.R., Grossi de Sa, F., 2002. Plant toxic proteins with insecticidal properties. A review on their potentialities as bioinsecticides. Toxicon 40, 1515–1539.

Chappill, J.J., 1995. Cladistic analysis of the Leguminosae: the development of an explicit hypothesis. In: Cripps, M.D., Doyle, J.J. (Eds.), Advances in Legume Systematics, Part 7, Phylogeny. Royal Botanic Gardens, Kew, pp. 1–10.

Chessel, D., Dufour, A.-B., Thioulouse, J., 2004. The ade4 package: I – one-table methods. R. News 4, 5–10.

Coombs, C.W., Billings, C.J., Porter, J.E., 1977. The effect of yellow split-peas (*Pisum sativum* L.) and other pulses on the productivity of certain strains of *Sitophilus oryzae* (L.) (Col. Curculionidae) and the ability of other strains to breed thereon. J. Stored Prod. Res. 13, 53–58.

Delobel, B., Grenier, A.M., Gueguen, J., Ferrasson, E., Mbaiguinam, M., 1998. Utilisation d'un polypeptide dérivé d'une albumine PA1b de légumineuse comme insecticide, Patent-98/05877, Paris, pp. 1–25.

Doyle, J.J., Luckow, M.A., 2003. The rest of the Iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiol. 131, 900–910.

Duffus, C.M., Smith, P.M., 1981. Legumes and their toxins. Span 24, 63–65.

Galtier, N., Gouy, M., Gautier, C., 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA 92, 11317–11321.

Galtier, N., Gouy, M., Gautier, C., 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Applic. Biosci. 12, 543–548.

Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple sequence data. Mol. Biol. Evol. 14, 685–695.

Gelly, J.C., Gracy, J., Kaas, Q., Le-Nguyen, D., Heitz, A., Chiche, L., 2004. The KNOTTIN website and database: a new information system dedicated to the knottin scaffold. Nucleic Acids Res. 32, D156–D159.

Grenier, A.M., Mbaiguinam, M., Delobel, B., 1997. Genetical analysis of the ability of the rice weevil *Sitophilus oryzae* (Coleoptera, Curculionidae) to breed on split peas. Heredity 79, 15–23.

Gressent, F., Rahioui, I., Rahbé, Y., 2003. Characterization of a high affinity binding site for the pea albumin 1b (PA1b) entomotoxin in the weevil *Sitophilus*. Eur. J. Biochem. 270, 2429–2435.

Hanada, K., Nishiuchi, Y., Hirano, H., 2003. Amino acid residues on the surface of soybean 4-kDa peptide involved in the interaction with its binding protein. Eur. J. Biochem. 270, 2583–2592.

Hancock, K.R., Ealing, P.M., White, D.W., 1994. Identification of sulphur-rich proteins which resist rumen degradation and are hydrolysed rapidly by intestinal proteases. British. J. Nutr. 72, 855–863.

Higgins, T.J.V., Chandler, P.M., Randall, P.J., Spencer, D., Beach, L.R., Blagrove, R.J., Kortt, A.A., Inglis, A.S., 1986. Gene structure, protein structure, and regulation of the synthesis of a sulfur-rich protein in pea seeds. J. Biol. Chem. 261, 11124–11130.

Holloway, G.J., Smith, R.H., 1985. Inheritance of the ability of *Sitophilus oryzae* (L.) (Col. Curculionidae) to feed and breed on yellow split-pea (*Pisum sativum*). Bull. Entomol. Res. 75, 367–375.

Hu, J.M., Lavin, M., Wojciechowski, M.F., Sanderson, M.J., 2000. Phylogenetic systematics of the tribe Millettieae (Leguminosae) based on chloroplast *trnK/matK* sequences and its implications for evolutionary patterns in Papilionoideae. Am. J. Bot. 87, 418–430.

Ilgoutz, S.C., Knittel, N., Lin, J.M., Sterle, S., Gayler, K.R., 1997. Transcription of genes for conglutin gamma and a leginsulin-like protein in narrow-leafed lupin. Plant Mol. Biol. 34, 613–627.

Jouvensal, L., Quillien, L., Ferrasson, E., Rahbé, Y., Gueguen, J., Vovelle, F., 2003. PA1b, an insecticidal protein extracted from pea seeds (*Pisum sativum*): (1)H-2D-NMR study and molecular modeling. Biochemistry 42, 11915–11923.

Kass, E., Wink, M., 1994. Molecular phylogeny of the Papilionoideae (Family Leguminosae): *rbcL* gene sequences versus chemical taxonomy. Bot. Acta 108, 149–162.

Louis, S., 2004. Diversité structurale et d'activité biologique des albumines entomotoxiques de type 1b des graines de légumineuses. Variability in structure and biological activity of A1b insecticidal albumins from legume seeds. PhD Thesis, INSA de Lyon, Lyon, pp. 1–264.

Louis, S., Delobel, B., Gressent, F., Rahioui, I., Quillien, L., Rahbé, Y., 2004. Molecular and biological screening for insect-toxic seed albumins from four legume species. Plant Sci. 167, 705–714.

Mbaiguinam, M., 1996. Détermination des causes de la résistance des légumineuses aux charançons des céréales du genre *Sitophilus*. PhD Thesis, University of Claude Bernard Lyon1, Lyon, pp. 1–164.

Ollier, S., Couteron, P., Chessel, D., 2006. Orthonormal transform to decompose the variance of a life-history trait across a phylogenetic tree. Biometrics 62, 471–477.

Page, D., 1999. Genetic improvement of the pea seed storage protein: heredity of its composition and molecular study of the tri locus, QTL candidate of the trypsin inhibitor activity. PhD Thesis, I.N. Polytechnique Lorraine, Nancy, pp. 1–122.

Polhill, R.M., 1994. Classification of the Leguminosae. In: Southon, I.W. (Ed.), Phytochemical Dictionary of the Leguminosae. Chapman & Hall, London, pp. 35–57.

Salzman, R.A., Fujita, T., Zhu-Salzman, K., Hasegawa, P.M., Bressan, R.A., 1999. An improved RNA isolation method for plant tissues

containing high levels of phenolic compounds or carbohydrates. Plant Mol. Biol. Rep. 17, 11–17.

Taylor, W.G., Fields, P.G., Sutherland, D.H., 2004a. Insecticidal components from field pea extracts: soyasaponins and lysolecithins. J. Agric. Food Chem. 52, 7484–7490.

Taylor, W.G., Sutherland, D.H., Olson, D.J.H., Ross, A.R.S., Fields, P.G., 2004b. Insecticidal components from field pea extracts: sequences of some variants of pea albumin 1b. J. Agric. Food Chem. 52, 7499–7506.

Thioulouse, J., Chessel, D., Dolédec, S., Olivier, J.M., 1997. ADE-4: a multivariate analysis and graphical display software. Stat. Comput. 7, 75–83.

Watanabe, Y., Barbashov, S.F., Komatsu, S., Hemmings, A.M., Miyagi, M., Tsunasawa, S., Hirano, H., 1994. A peptide that stimulates phosphorylation of the plant insulin-binding protein. Isolation, primary structure and cDNA cloning. Eur. J. Biochem. 224, 167–172.

Wink, M., Mohamed, G.I.A., 2003. Evolution of chemical defense traits in the Leguminosae: mapping of distribution patterns of secondary metabolites on a molecular phylogeny inferred from nucleotide sequences of the *rbcL* gene. Biochem. Syst. Ecol. 31, 897–917.

Wojciechowski, M.F., 2003. Reconstructing the phylogeny of Legumes (Leguminosae): an early 21st century perspective. In: Klitgaard, B.B., Bruneau, A. (Eds.), . In: Advances in Legume Systematics, vol. 10. Royal Botanic Gardens, Kew, pp. 5–35.

Wojciechowski, M.F., Lavin, M., Sanderson, M.J., 2004. A phylogeny of legumes (Leguminosae) based on analyses of the plastid *matK* gene resolves many well-supported subclades within the family. Am. J. Bot. 91, 1846–1862.

Yamazaki, T., Takaoka, M., Katoh, E., Hanada, K., Sakita, M., Sakata, K., Nishiuchi, Y., Hirano, H., 2003. A possible physiological function and the tertiary structure of a 4-kDa peptide in legumes. Eur. J. Biochem. 270, 1269–1276.