Theor Appl Genet (1996) 93: 997–1005

P. Touzet · F. Riccardi · C. Morin · C. Damerval
J.-C. Huet · J.-C. Pernollet · M. Zivy · D. de Vienne

The maize two-dimensional gel protein database: towards an integrated genome analysis program

Received: 8 August 1995 / Accepted: 10 May 1996

Abstract This paper describes the first maize database of proteins separated by two-dimensional electrophoresis. Fifty-six coleoptile proteins and 18 leaf proteins from two maize lines were partially microsequenced. Thirty-six proteins (49%) displayed high similarity with database proteins. Nine of these proteins, representing five different functions, had never been described in maize. No conclusive function could be found for 45 polypeptides (61% of the microsequenced proteins). In addition, an alternative identification method, based on amino acid analysis, allowed candidates to be proposed for 17 proteins out of 44 additional proteins analyzed in the coleoptiles. These results are stored in a database which also includes, when available, genetic information about the chromosomal location of structural genes and regulatory factors of proteins. This database is being used in the context of a project on the genetic mapping of the expressed genome in maize.

Key words Maize · Two-dimensional electrophoresis · Coleoptile proteins · Leaf proteins · Protein database

Introduction

High-resolution two-dimensional electrophoresis (2-DE), first developed by O'Farrell (1975), is the most powerful technique available at present for analyzing complex protein mixtures. Up to 1000 silver-stained proteins can be resolved in a single two-dimensional gel. The applications of this technique are numerous. For instance, in plants 2-DE is the technique of choice for detecting and quantifying modifications in genome expression during development, under different stresses or in response to hormones, pathogen infections and symbiosis (reviewed in Damerval et al. 1988). In genetics, 2-DE is a source of monogenic and codominant markers that are useful for diversity studies (Klose 1982; Zivy et al. 1984; Goldman et al. 1987, etc.) and genome mapping (Bahrman and Damerval 1989; de Vienne et al. 1996). Quantitative variations in proteins may provide predictors for heterosis and traits of agronomical interest (Damerval et al. 1987; Leonardi et al. 1991). Recently, Damerval et al. (1994) showed that in maize it is possible to map the loci involved in quantitative variations of individual proteins using a computer-assisted system for spot quantification and applying the methodology for mapping QTL (quantitative trait loci) (Helentjaris 1987; Paterson et al. 1988).

The routine identification of many proteins excised from 2-DE gels would reinforce interest in the physiological and genetic applications of 2-DE. This can now be achieved using methods for obtaining N-terminal or internal amino acid sequence information from microgram amounts of protein in a single spot (Bauw et al. 1987). We report here the results of a sequencing project for maize proteins extracted from etiolated coleoptiles and leaves. Data obtained from an alternative identification method, based on amino acid composition, are also presented.

Materials and methods

Plant material

Two genetically distant maize lines were used: F2, a flint line from the Institut National de la Recherche Agronomique (France), referred to as Lc; and Io, an American dent line from the Iodent group. The kernels were allowed to germinate in the dark at 24°C for 8 days. The etiolated coleoptiles were then removed and immediately frozen in liquid nitrogen.

The sixth leaf of six-to-seven-leaf-stage plants (25 days after planting) was removed and immediately frozen in liquid nitrogen.

Communicated by G. E. Hart

P. Touzet · F. Riccardi · C. Morin · C. Damerval · M. Zivy · D. de Vienne (✉)
Station de Génétique Végétale,
INRA/UPS/INA P-G/CNRS-URA 1492,
la Ferme du Moulon, 91190 Gif-sur-Yvette, France

J.-C. Huet J.-C. Pernollet
Unité de Recherches de Biochimie et Structure des Protéines,
INRA, Domaine de Vilvert, 78352 Jouy-en-Josas, France

Protein extraction

Denaturing protein extraction (TCA-acetone extraction) was as described in Damerval et al. (1986) for most of the proteins, except that 60 µl of resolubilization solution was used to resuspend 1 mg of pellet. Proteins of low abundancy were concentrated using a TRIS-HCl extraction procedure (Zivy et al. 1983), followed by differential precipitation with acetone.

Gel electrophoresis

Analytical 2-D electrophoresis

The isoelectrofocusing and SDS-PAGE dimensions were as described by Damerval et al. (1987). Approximately 45 µg of protein was loaded onto each gel. The silver staining was as described in Damerval et al. (1987) with the modifications described by Burstin et al. (1993). The various steps of the procedure were carried out at 22°±2°C.

Preparative 2-D electrophoresis

The preparative 2-DE conditions were similar to those of the analytical 2-DE, with the following modifications. About 500 µg of proteins was loaded onto isoelectrofocusing gels (3 mm in diameter); the 2-D gels were 1.5 mm thick, with stacking gels. The gels were either stained with Coomassie Blue, dried and stored at room temperature prior to concentration in a second SDS PAGE (Rasmussen et al. 1991), or stained with Amido Black, following the Chen et al. procedure (1993), prior to in situ gel digestion.

Protein electrotransfer

Proteins were electroblotted onto PVDF ProBlott membranes (Applied Biosystems) using 50 m*M* TRIS-base, 50 m*M* boric acid as the transfer buffer. Prior to transfer, the gels were equilibrated for 10 min in transfer buffer containing 0.1% SDS, whereas the membrane was soaked in methanol for 5 min and then rinsed with distilled water. Protein transfer was carried out using a Bio-Rad Trans-blot Cell overnight at 4°C under 20 V, with the amperage limited to 0.15 A.

Amino acid sequence analysis

N-terminal amino-acid sequence analysis

Prior to sequencing, the pieces of membranes were cut into several fragments and deposited on a glass fiber filter in the cartridge of the Applied Biosystems 470A sequencer.

Internal amino acid sequence analysis

Internal amino acid sequences of coleoptile and leaf proteins, separated by 2-DE, were obtained using several methods: in situ protease cleavage on the membrane, as described by Fernandez et al. (1992); gel in situ protease cleavage as described by Kawasaki et al. (1990) or Rosenfeld et al. (1992). Trypsin (Sigma), modified trypsin (Promega) or lysylendopeptidase (Boehringer-Mannheim) were used.

The combined peptide solution was loaded onto a narrow-bore reversed phase column [C8 0.21×10 cm, (Applied Biosystems) or C18, 0.21×15 cm (Vydac)]. A ion-exchange precolumn [DEAE AX-300 Aquapore 0.46×3 cm, (Applied Biosystems)] was connected in series in the case of in-gel digestion in SDS buffer (Kawasaki and Suzuki 1990). The peptides detected by UV absorbance at 215 nm were collected manually. An internal sequence was obtained by pooling 8–105 spots to reach a total of at least 100 pmol

Automated Edman degradation of the peptides was performed using either an Applied Biosystems 475A sequencer and its on-line model 120A phenylthiohydantoin amino acid analyser or an Applied Biosystems 477A sequencer, carrying out the chemical reactions according to the manufacturer's suggestions.

Internal sequences from coleoptile proteins described as position variants and from leaf proteins were obtained by the Laboratoire de Microséquençage des Protéines, Institut Pasteur (Paris-France).

Similarity search of amino acid sequences

The amino acid sequences were compared to the 175 933 sequences of the Non-redundant Peptide Sequence Database. Feb. 27 1996, of the National Center for Biotechnology Information using the BLAST program (Altschul et al. 1990). This database includes: Brookhaven Protein Data Bank April 1995 Release; SWISS-PROT 32.0; PIR 47 0, CDS translations from Genbank 93; Kabat Sequences of Proteins of Immunological Interest, June 1995; TFD transcription factor database 7; Translations of selected Alu repeats from REPBASE. In addition, weekly or daily updates from SWISS-PROT and Genbank were also interrogated. The BLAST program was run using the default parameters proposed by the NCBI server with a BLOSUM62 matrix.

Detection of glycosylation

Coleoptile proteins separated by 2-DE were transferred onto nitrocellulose membranes (Biorad) and reacted with a peroxidase-coupled concanavalin A (L Faye, CNRS, Rouen) according to the procedure of Faye and Chrispeels (1985).

Amino acid analysis

One to twelve 2-DE spots per protein were pooled for the amino acid analysis. The blotted proteins were hydrolyzed automatically or manually. The autohydrolysis was conducted at 155°C for 75 min, using an Applied Biosystems 420H according to the manufacturer recommendations (User bulletin 9, Applied Biosystems). For manual hydrolysis, the pieces of PVDF membranes were sealed *vacuo* in glass tubes containing 6 *N* HCl and 0.1% phenol. Hydrolysis was conducted overnight at 112°C. Amino acids were derivatized with phenylisothiocyanate (PITC) (quantified as phenylisocarbonyl-amino acids) on an Applied Biosystems 420A derivatizer.

Amino acid contents were measured for 15 amino acids: Phe, Leu, Ile, Val, Ser, Pro, Thr, Ala, Tyr, His, Lys, Arg, Gly, Asp+Asn (Asx), Glu+Gln (Glx). For a given protein, the composition in 1 amino acid corresponded to the ratio of the quantity of this amino acid in picomoles to the sum of all the 15 amino acids in picomoles.

Database of amino acid compositions

Plant protein sequences were extracted from the PIR release 43 using the ACNUC program from the BISANCE server (Dessen et al. 1990).

Sequences referred to as fragments were excluded, and amino acid composition, pI and Mr were computed for each sequence, in its native as well as processed form when necessary, just as in the BISANCE server. For each protein the composition database, which comprised 5657 amino acid compositions, contained the mnemo, computed Mr and pI, rank of first and last amino acid (allowing the user to know whether it is a native or processed protein), the composition for each of the 20 amino acids (in number of amino acids) and the function of the protein, also extracted from PIR.

Identification by comparing amino acid compositions

In order to compare amino acid compositions, the program extracted from the database all of the proteins that belong to a pI-Mr window around the protein that is being analyzed. The pI and Mr intervals were ±1 pH unit and ±10% respectively The pI interval was

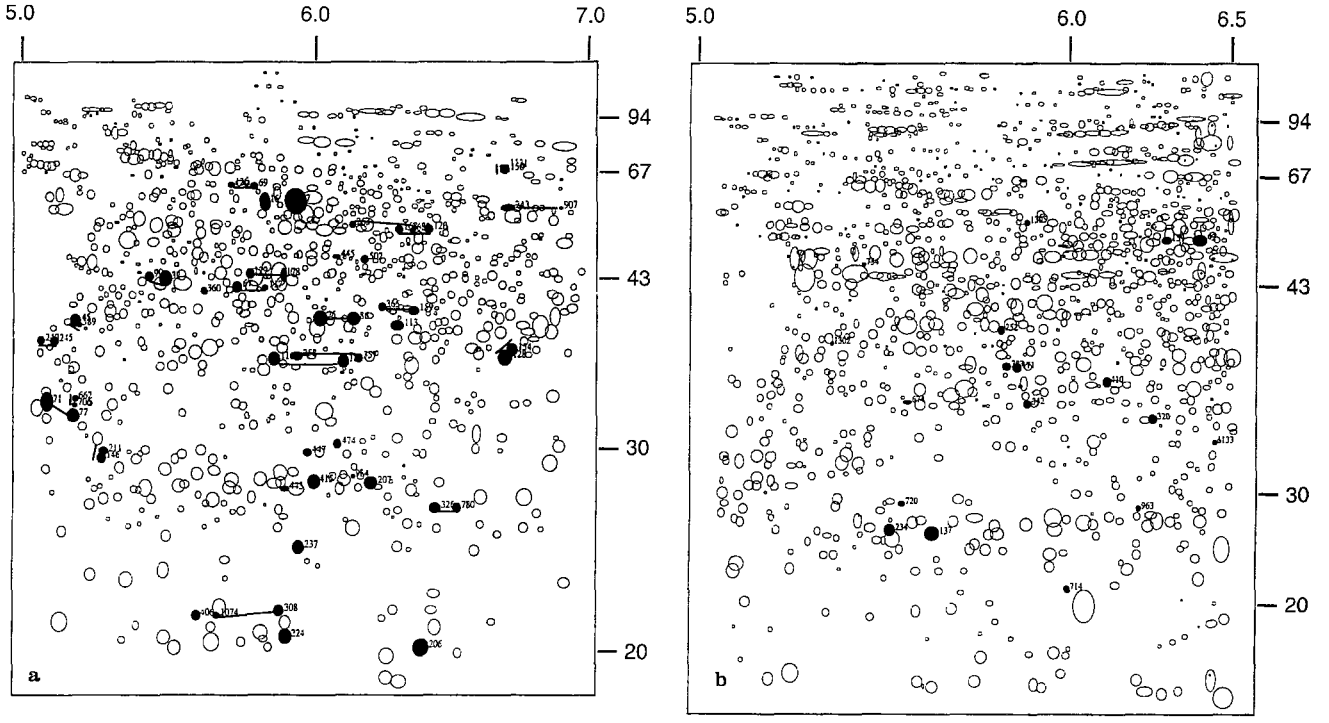

Fig. 1a, b Schematic representation of a silver-stained 2-D gel of a coelectrophoresis 1:1 of proteins from the two maize lines Io and Lc. *Right to left* first dimension. *top to bottom* second dimension. pI and molecular weight (kDa) (*right margin*) scales are indicated. The *black spots* correspond to the microsequenced proteins. The allelic proteins are connected by a *line*. **a** coleoptile proteins. **b** leaf proteins

chosen quite large because of possible effects of sequence variation and/or post-translational modifications on pI.

Database Glx and Asx and composition relative to the 15 measured amino acids were computed so that database and experimental compositions were directly comparable.

Relatedness parameters

Two parameters were used for the comparisons between compositions. The first one was a relative Euclidian distance.

$$d = \sqrt{\sum_{i}^{15}\left(\frac{A_i - B_i}{B_i}\right)^2}$$

where A_i is the content in amino acid of the analyzed protein A. B_i is the content in the same amino acid of the database protein B. The second parameter r is the correlation between protein A and protein B values for the 15 amino acid contents. Two other parameters were computed: r' and d'. They were computed in the same way as r and d, except that the 2 amino acids with the highest relative difference $\left(\frac{A_i - B_i}{B_i}\right)$ were not taken into account. This was empirically done to limit possible biases due to poorly estimated amino acid contents.

The program produces a result file in which all the database proteins within the pI-Mr window *and* satisfying the following conditions (*and*) or [(*or*) and (d'<0.7 and r'>0.92)] are selected as "candidate" functions. These thresholds are discussed in Touzet et al. (1996). Briefly, the combination of the four thresholds allowed us to select less than 10% false positives while retaining a good sensitivity (approximately 85% good candidates selected).

Results

Choice of the proteins

Some of the proteins analyzed were selected because they were present in abundance, but for most of them the choice was based on genetic criteria. When 2-D gels from homozygous lines were compared, two kinds of polymorphism were observed:

– allelic position shifts (PS), which are most likely due to coding sequence polymorphism in the structural genes of the proteins (Touzet et al. 1995; de Vienne et al. 1996);
– quantitative variations of proteins, i.e. the variable proteins are not identically expressed in the two genotypes. Loci responsible for these variations (PQL for protein quantity loci) can be mapped using a marker-based quantitative genetics approach (Damerval et al. 1994).

Six genetically unvariable proteins, 43 proteins corresponding to 23 PS variants (in 3 cases, only one member of the pair was microsequenced), 7 quantitatively variable proteins from coleoptile and 18 leaf proteins possibly associated with maize responses to drought stress were microsequenced. Forty-four additional coleoptile proteins were tentatively identified by comparing their amino acid composition with the amino acid composition of plant proteins stored in databases.

Sequence similarity

The 56 coleoptile proteins and 18 leaf proteins that were analysed are shown in Fig. 1. Table 1 contains the sequences and the results of similarity searches.

Table 1 List of the microsequenced proteins and identity scores. The type of variation, position shift or quantitative, is indicated. For PS proteins, the allele protein is indicated between parenthesis when it has been microsequenced. Only the non-redundant sequence information is listed [*Glyc* glycosylated (y) or non-glycosylated (n) protein (see Materials and methods), *Enz* protease used for the cleavage; lysylendopeptidase (KC) or trypsin (T)]

Spot	Line	Tissue	Type	Glyc	pI	Mr (kDa)	Enz	Similar protein	Identity (%)	Sequence		Chromosome[b]	Marker[c]	PQL[d]
										N-terminal	internal[a]			
COL0007	Io/Lc	Coleoptile	Unvariable		5.93	58.4	KC	Beta-glucosidase	100	SARVGSQNGVQMLSP	YGGFLDK	10	UMC64	
							KC	Beta-glucosidase	100		SIVEDYTYFAK			
							KC	Beta-glucosidase	100		XGGINPDGIK			
							KC	Beta-glucosidase	100		VCFDNFGD			
COL0016	Io/Lc	Coleoptile	Unvariable		5 83	58.3	KC	Beta-glucosidase	100	SARVGSQNGVQMLSP		10	UMC64	
COL0032	Io	Coleoptile	PS2 (90)		5 54	42.7	KC				ALSVPVFAVAPLNKK	1	UMC53	
COL0045	Io	Coleoptile	PS33 (589)		5 26	37.6	KC				NPNPVPIPLVDIDYL	1	BNL632	
COL0067	Io	Coleoptile	PS35 (117)	n	5 76	41 7	KC				WILHDWDEDK	4	UMC31	
							KC				XXXPYYNTI			
COL0069	Io	Coleoptile	PS16 (126)	n	5 8	61.1	KC	Phosphoglycerate mutase	100		RGWDAQVLGEAPYK	3	UMC60	1 (3)
COL0075	Io	Coleoptile	PS11 (86)		6 01	38 0	KC				XADGAMINYVE	2	UMC16	
							KC				AFPHIK			
							KC				CTVLAP			
							KC				EFGAYVAFEVYP			
							KC				GFSEYK			
							KC				EWSELFTK			
COL0077	Io	Coleoptile	PS4 (71)		5 25	31.6	KC				AIGGLSRSFPVEAFE	3	UMC102	
COL0113	Io	Coleoptile	Unvariable		6.33	37.4	KC				VPEGFDYELYNRNDI			
COL0116	Lc	Coleoptile	PS18 (181)	n	5 87	34 3	KC	Cysteine synthase	100		DVTELIGNTPLVYLN	1	UMC53	
COL0128	Io	Coleoptile	PS25 (154)	n	6 77	34.6	KC				ADEGFSATVRNGAV	1	UMC67	1(1)
COL0146	Io	Coleoptile	PS32 (211)		5.35	29.3	KC				AAVSLLQNK	2	UMC5	3 (2.3)
							KC				LAYDGFLSK			
COL0154	Lc	Coleoptile	PS25 (128)	n	6.80	35.3	KC				ILP(W)GDEAYAAGG	1	UMC67	1 (1)
							KC				(D)HPGVRDGTNIVLXK			
COL0159	Io	Coleoptile	PS23 (365)	n	6.40	38.8	KC	Cinnamyl alcohol dehydrogenase	69		NDWRNAMYPVVPGHE	7	UMC116	
COL0168	Lc	Coleoptile	PS22 (120)	n	6.34	48 9	KC				IILELAAVDSASGK	9	BNL510	5 (5,6,8, 9)
							KC				LLINGNFK			
COL0178	Io	Coleoptile	PS10 (129)	n	5.90	43.3	KC	Methionine adenosyl transferase	100		RPEEIGAGDQGHMFG	3	UMC102	
COL0206		Coleoptile	Quantitative		6.45	20 4	KC	ABA- and ripening-induced protein	81		ITEEVAAAAAVGAGGYV			
							KC				XXLGEAG			
							KC	ABA- and ripening-induced protein	75		HHHLFNHE			
COL0207	Lc	Coleoptile	Quantitative		6 23	27.1	KC	Auxin-induced protein	54		GLAYEYLEQDLGK			3 (1,9)
							KC	Auxin-induced protein	84		KSELLLAANPVHK			
							KC				VYDFVXGMK			
							KC				PEVAK			
COL0224	Lc	Coleoptile	Quantitative		5 91	20.9	T	Initiation factor IEF-5A	100		XLEDIVPS			
							T	Initiation factor IEF-5A	66		LPTDETLVAQIK			
COL0237	Io	Coleoptile	Quantitative		5 95	21 7	KC	ABA- and ripening-induced protein	87		IEEEVAAAAAVGSGG	10	UMC64	
							KC	ABA- and ripening-induced protein	100		XHHHHLFHHK			
							KC	ABA- and ripening-induced protein	84		XPEHAHRHKIEEEVA			
COL0243	Io	Coleoptile	PS39 (907)		6 78	56 9	KC	Glucose-6-phosphate 1-dehydrogenase	92		XGRNEFVIRLQXSEA	5	UMC54	
COL0245	Lc	Coleoptile	PS8 (249)		5 19	35 7	KC				VSTLLPVVAAEEPA	5	BNL771	
COL0258	Lc	Coleoptile	PS19 (357)	n	5.94	34.6	KC				AYGGDGGAYYEWSPA	8	UMC89	
COL0263	Io	Coleoptile	PS47 (688)		6.14	50 3	KC				(D)VHSNTGIFGIHTS	3	UMC92	
COL0308	Io	Coleoptile	PS29 (1074)		5 89	18.6	KC				QVWYDREVTAFVEPG	6	UMC65	
COL0326	Io	Coleoptile	PS28 (780)		6.50	23.7	T	Glutathione S-transferase III	100		VLDVYEAHLAR	3	UMC102	
COL0360	Io	Coleoptile	Quantitative		5 65	41 1	T				SIXEPLAL			2 (1.5)
							T				SVFDEP			
COL0365	Lc	Coleoptile	PS23 (159)		6 27	39 2	KC	Cinnamyl alcohol dehydrogenase	93		HLGVVGLGGLGHVAV	7	UMC116	
							KC	Cinnamyl alcohol dehydrogenase	71		XQEAIENLXADEFLI			
COL0406	Lc	Coleoptile	PS43		5.64	18.3	KC				NGRRYTTYG(C)SPPVT	5	UMC43	
COL0415	Io/Lc	Coleoptile	Unvariable		6 00	27 2	KC	L-ascorbate peroxidase	100	AKNYPTVS	(GS)DHLRQVF			
							KC	L-ascorbate peroxidase	100		IMGLSDQ			
							KC	L-ascorbate peroxidase	100		ALLSDPVFRPLVEK			
							KC	L-ascorbate peroxidase	83		XFFDDYA			
							KC		100		XRSGFEG			
COL0443	Io/Lc	Coleoptile	Unvariable		5 93	26 5	KC	L-ascorbate peroxidase	100	AKNYPTVSAEYSXAVEKA				

COL0445	Io/Lc	Coleoptile	Unvariable		6.08	45.3	KC	Methionine adenosyl transferase	86	TQVTVEYVNEGGAMV				
COL0447	Lc	Coleoptile	Quantitative		5.98	30 0	KC	Trithorax protein	87 5	STAKSTA				2 (7, 10)
COL0474	Lc	Coleoptile	Quantitative		6 09	30 4	KC	Glycosyl hydrolase signature	100	IFFEV				1 (10)
COL0502	Io	Coleoptile	PS15		6.20	45.0	KC			AGDKPGDALLDEWLG	6	UMC65		
COL0662	Lc	Coleoptile	PS5 (706)		5.25	32.3	KC			SGTSPLLPAITFILD	3	UMC32		
COL0688	Lc	Coleoptile	PS47 (263)		6 40	48 4	KC	Histone-binding protein	93	EREQLRDQVYDAMAE	3	UMC92		
COL0907	Lc	Coleoptile	PS39 (243)	n	6.99	57 1	KC	Glucose-6-phosphate 1-dehydro-genase	75	AEPRDQFK	5	UMC54		
COL0984	Io	Coleoptile	PS42		6 16	27 7	KC			XLEGAFVLNQSQ(D)AE	8	BNL911		7 (1,3,5,
COL1131	Lc	Coleoptile	PS21 (150)	y	6 76	71 0	KC			AAAAPPRRGPSGPDA	5	UMC54		6,8)
LEF0092	Io	Leaf			6 36	52 0	KC	Beta-glucosidase	100	GENGXIGLAFDVMGR				
							KC	Beta-glucosidase	100	NWLTFNEPQTFTSFS				
LEF0137	Lc	Leaf			5 62	26.8	KC	Ferritin	100	EVLSGVVFQPFEEIK				
LEF0159	Io	Leaf			6.24	52 0	KC	Beta-glucosidase	100	VCFDNFGDK				
LEF0171	Lc	Leaf			5 82	37.0	KC	Malate dehydrogenase	100	MELVDAAFPLLK				
LEF0234	Lc	Leaf			5 52	27.0	KC	Triosephosphate isomerase	93	DWSNVVLAYEPVWAI				
LEF0252	Io	Leaf			5.78	40 4	KC	Glutamate 1-semialdehyde 2,1-aminotransferase	92	SVGGQPVVFDSVK				
LEF0320	Lc	Leaf			6.22	33 1	KC			HSLGQSHPVLLTRHN				
LEF0342	Lc	Leaf			5.84	33 7	KC			VLEGAEERLQLLK				
LEF0383	Io	Leaf			5.79	36.8	KC	Catechol O-methyltransferase	100	WILHDWSDAHXATLL				
LEF0410	Io	Leaf			6 06	34.6	KC	Fructose 1,6-bisphosphate aldolase	100	TIASPGRGILAMDES				
LEF0674	Io	Leaf			5 55	33.7	KC	Cysteine synthase	60	KVXIFGGIGTGGT				
LEF0714	Io	Leaf			5.98	21.7	KC	ABA stress and ripening protein	93	QHLGEAGAIAAGAFA				
LEF0720	Lc	Leaf			5 54	29.7	KC	Soluble inorganic pyrophosphatase	81	RAPKLNERILSSLSR				
LEF0754	Io	Leaf			5.43	48 1	KC	Enolase	100	YNQLLRIEEELGDAA				
LEF0963	Lc	Leaf			6.17	28 6	KC			SLEGAFVLNQHQPAE				
LEF1133	Io	Leaf			6 47	32.1	KC			IATVEPVTMK				
							KC			AVHNLVEAVSQHGVA				
							KC			YVHVVTVDSHDFWFM				
LEF1302	Io	Leaf			5 35	39.2	KC	Phosphoribulokinase	100	ASIEARKPDFDAFID				
LEF1305	Io	Leaf			5 83	58.5	KC	Beta-glucosidase	100	AGSYNMLGLNYY				

[a] X within a sequence means an undetermined residue; a residue between parentheses indicates the most probable one

[b] Chromosomal location of the structural gene (putatively determined by genetic mapping of loci responsible for PS)

[c] The closest RFLP marker to the mapped PSL (Causse et al. 1996; de Vienne et al. 1996)

[d] PQL=Protein quantity loci. The number present is indicated, with the chromosomal location indicated between parentheses (Damerval et al. 1994)

Coleoptile proteins

Genetically unvariable polypeptides. Some of those peptides present in abundance in the coleoptile that were common to the two genotypes and not quantitatively variable were chosen at the beginning of our program in order to test the methodology.

Polypeptides COL0007 and COL0016 were found to share the same N-terminal sequence. The N-terminal sequence and several internal sequences allowed us to identify them as being a β-glucosidase (EC 3.2.1.21) (Brzobohaty et al. 1993). This protein is abundant in the coleoptile since it represents 5–8% of the total soluble proteins (Esen et al. 1992). It plays a role in the maturation of phytohormones (Brzobohaty et al. 1993), and its activity may be highly correlated with growth (Esen et al. 1992). In another genetic background, polypeptides COL0007 and COL0016 exhibit a position shift in the same direction that is controlled by a single locus (PSL84 and 86 on chromosome 10; de Vienne et al. 1996). The occurrence of these two different forms could then be explained by post-translational modifications. In addition, COL0007 is more expressed than polypeptide COL0016 in the coleoptile, whereas in the leaf polypeptide COL0016 (or LEF1305) is more expressed than COL0007. This might suggest that post-translational modifications of the product of the β-glucosidase gene are also organ-specific. Two additional forms have been reported in the leaf (LEF0159 and LEF0092) that have a smaller weight and a more basic pI (see below).

Polypeptides COL0415 and COL0443 share the same N-terminal sequence, which is similar to that of a maize cytosolic L-ascorbate peroxidase (EC 1.11.1.11) (Koshiba 1993). The presence of the four internal sequences of COL0415 confirm this identity with other L-ascorbate peroxidases from different species. This cytosolic ascorbate peroxidase could be involved in the indole-3-acetic acid biosynthesis pathway from tryptophan (Koshiba 1993). Another ascorbate peroxidase has been reported in maize, but it does not share the same N-terminal sequence and is not strictly identical to the internal sequences of COL0415 (Van Breusegem et al. 1995). The polypeptides COL0415 and COL0443 could be the products, post-translationally modified, of the same gene.

Lastly, polypeptide COL0445 shows a similarity to S-adenosyl methionine synthetase (EC 2.5.1.6) in several plant species. *Hordeum vulgare*, *Actinidia chinensis*, *Populus deltoides*, *Dianthus caryophyllus* and *Arabidopsis thaliana*.

Position shift variants (PS). Among the 23 PS variants characterized, the amino-acid sequences of 3 variants were found to be strictly identical to those of maize proteins already characterized: PS16 is identical to bisphosphoglycerate-independent phosphoglycerate mutase (EC 5.4.2.1) (Grana et al. 1992), an enzyme of the glycolytic pathway; PS28 to glutathione S-transferase III (EC 2.5.1.18) (Grove et al. 1988); and PS18 to a chloroplastic isoform of cysteine synthase (EC 4.2.99.8) (EMBL ZMCSOATL).

Four PS variants were found to be similar to proteins from plants other than maize or from other species out of the plant kingdom: methionine adenosyltransferase (EC 2.5.1.6) of *Lycopersicon esculentum*, *Populus deltoides* and *Oryza sativa* (PS10); cinnamyl alcohol dehydrogenase of *Arabidopsis thaliana* and *Nicotiana tabacum* (PS23); glucose-6-phosphate 1-dehydrogenase (EC 1.1.1.49) of *Solanum tuberosum* (PS39). A maize expressed sequence tag (csuh00350) sharing the same homology with this last protein has been described (Keith et al. 1993). Surprisingly, PS47 shares similarity to the histone-binding protein N1/N2 of *Xenopus laevis*. This last protein is described as being nuclear and involved in the nucleosome assembly. Nevertheless, an important discrepancy exists between the pI of the PS variant and the pI of this protein. In addition only one out of two internal sequences shows such a strong similarity (14/15 AA) to the histone binding protein (Touzet et al. 1995).

Quantitative variants. Proteins COL0207, COL0206 and COL0237 show a similarity with proteins that are induced by several types of stress, but their functions are not well-defined.

Two out of the five sequences of COL0207 show similarity with the auxin-regulated genes of tobacco, soybean and potato that have a conserved N-terminal region, which might confer a glutathione S-transferase function (Droog et al. 1993). The three remaining dissimilar sequences would then correspond to the less conserved regions.

Proteins COL0206 and COL0237 are similar to abscisic stress and ripening proteins of tomato and might belong to the same gene family. These proteins, induced by ABA or drought stress, might reveal a genetic ability of the plant, already expressed in its early stage, to face these kinds of stress.

Proteins COL0224, COL0447 and COL0474, the quantities of which are regulated by loci on chromosome 10 (Damerval et al. 1994 and Damerval, unpublished data), are not the products of the same gene. The two sequences of COL0224 show similarity with the initiation factor 5A of *Medicago sativa*, a protein that promotes the first peptide bond during protein biosynthesis (Pay et al. 1991). Protein COL0447 shows a surprising similarity with the trithorax protein of Drosophila (7 successive amino acids): however, the important size discrepancy between the two proteins (30 kDa for COL0447 against 145 kDa for the trithorax protein) does not allow us to identify them together. In addition, the fact that the similar region does not seem to play a known functional role makes the identification of the function of COL0447 difficult (Mazo et al. 1990). Lastly, the short sequence of COL0474 corresponds to the signature of a glycosyl hydrolase family (Prosite, Bairoch 1991).

Leaf proteins

The 18 leaf proteins were sequenced in the context of a drought-stress project. Among them, 9 were similar or identical to proteins previously characterized in maize:

β-glucosidase (EC 3.2.1.21) (3 forms), ferritin, triose phosphate isomerase (EC 5.3.1.1), catechol *O*-methyl transferase (EC 2.1.1.6), malate dehydrogenase (EC 1.1.1.37), enolase (EC 4.2.1.11) and fructose 1.6-biphosphate aldolase (EC. 4.1.2.13). The three β-glucosidases seem to be the products of two different genes: peptide LEF1305 might correspond to the product of a first gene also expressed in the coleoptile (COL0007 and COL0016) (Brzobohaty et al. 1993), whereas peptides LEF0159 and LEF0092, with a more basic pI and a smaller weight, could be the products of another gene (Bandaranayake and Esen 1996). Three other proteins were found to be similar to glutamate 1-semialdehyde 2,1-aminotransferase (EC 5.4.3.8) of *Hordeum vulgare*, phosphoribulokinase (EC 2.7.1.19) of sorghum and soluble inorganic pyrophosphatase (EC 3.6.1.1) of *Solanum tuberosum*.

Conclusion on sequence similarity

With respect to the two organs, the coleoptile and the leaf, we identified 11 functions previously reported in maize: β-glucosidase (4 forms, possibly 2 genes), 2,3-biphosphoglycerate-independent phosphoglycerate mutase (2 allelic forms), glutathione *S*-transferase III (2 allelic forms), glucose 6-phosphate 1-dehydrogenase (2 allelic forms), L-ascorbate peroxidase (2 forms), ferritin, triose phosphate isomerase, catechol *O*-methyl transferase, malate dehydrogenase, enolase and fructose 1,6-biphosphate aldolase. We tentatively identified 5 proteins which had never been described in maize: methionine adenosyl transferase (3 forms, possibly 2 genes), initiation factor IEF-5A, cinnamyl alcohol dehydrogenase (2 allelic forms), glutamate 1-semialdehyde 2,1-aminotransferase, phosphoribulokinase and soluble inorganic pyrophosphatase. Finally we were unable to propose any conclusive function for 45 polypeptides (61% of the microsequenced proteins), due either to the lack of strong similarity or because the proteins matched are physiologically poorly characterized (e.g. proteins induced by ABA or auxins). Surprisingly, functions found in the leaves under water-stress conditions were also found in the coleoptile: β-glucosidase, ABA-induced proteins. Actually more than a third of the identified coleoptile proteins (9/24) are related to hormonal pathways. This is the case for β-glucosidase, auxin-regulated proteins, ABA-regulated proteins, ascorbate peroxidase, glutathione *S*-transferase. It must be noted that the biases that we introduced in the choice of the proteins are the relative abundancy in the coleoptile and their genetic variability. These proteins are also revealed in coleoptiles from non-etiolated plants and thus cannot result from possible stress caused by growth under etiolated conditions (not shown).

Amino acid composition analysis

Amino acid analysis as a tool for rapid identification of proteins was recently investigated in our lab (Touzet et al. 1996).

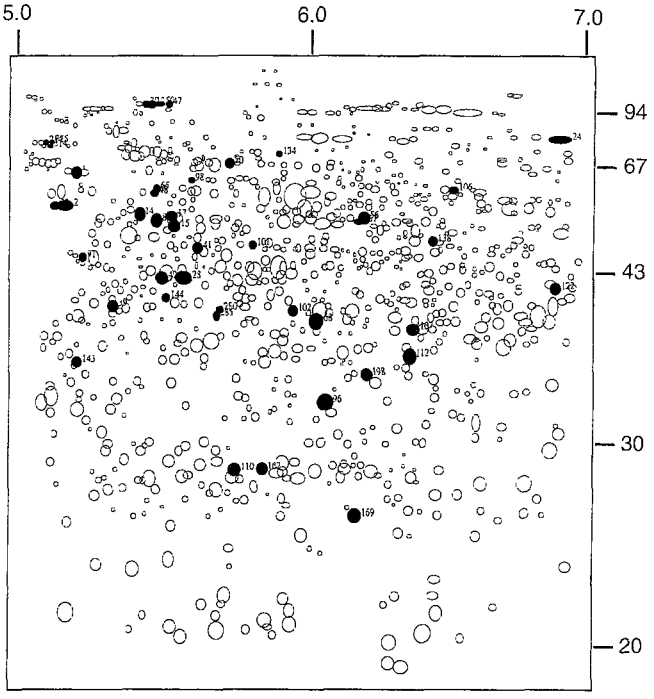

Fig. 2 Schematic representation of a silver-stained 2-D gel of a co-electrophoresis 1:1 of proteins from the two maize lines Io and Lc. The *black spots* correspond to the 44 additional polypeptides that were analyzed for their amino acid composition

Amino acid compositions were compared through the use of two independent parameters, a relative Euclidian distance and the correlation coefficient between the amino acid compositions. These were used to assess the relatedness between proteins separated by 2-DE and plant proteins extracted from a database (see Materials and methods).

We used this method to try to identify 44 additional coleoptile proteins separated by 2-DE (Fig. 2). An identification was considered as reliable when only one function was proposed among the first rank candidates on the four parameters. Table 2 gives the list of the 17 proteins that were putatively identified. It can be noted that new functions were proposed, e.g. H^+-transporting ATPases, tubulins, actin, enolase, glutamate ammonia ligase, pyruvate orthophosphate dikinases, etc. Interesting enough, a similar pyruvate orthophosphate dikinase cluster has also been described in sorghum (Monroy et al. 1988). Other isoforms of proteins previously identified by microsequencing, e.g. a phosphoglycerate mutase, an ascorbate peroxidase, were also detected.

Conclusion

We have described the first maize protein database including sequence and amino acid composition data. This database is currently being used in the context of a genetic mapping of the expressed genome in maize project. The large

Table 2 Candidate proteins for the 44 additional coleoptile proteins analyzed for their amino acid composition

Spot	Putative function
COL0008	H^+-transporting ATP synthase
COL0012	Pyruvate orthophosphate dikinase
COL0015	Enolase
COL0023	Oryzain
COL0024	Beta-fructofuranosidase
COL0039	3-Isopropylmalate dehydrogenase
COL0041	Tubulin beta chain
COL0042	Actin
COL0047	Pyruvate orthophosphate dikinase
COL0056	H^+-transporting ATP synthase
COL0065	Chaperonin hsp60
COL0098	Phosphoglycerate mutase
COL0101	Tubulin alpha chain
COL0102	Nitrilase
COL0144	Glutamate ammonia ligase
COL0162	Ascorbate peroxidase
COL0280	Ribonucleoprotein

scale cDNA sequencing projects developed in maize (Keith et al. 1993; Shen et al. 1994) and rice (Kurata et al. 1994) are valuable sources of probes in mapping the structural genes coding for the identified proteins. In addition, whenever the protein is positionally variable, the location of the locus responsible for the position shift can be compared to the mapping data from restriction fragment length polymorphism analyses. In case of multilocus profiles, the PS locus reveals which locus is expressed and translated in the studied organ. This program consequently provides a complementary insight on genome organization and expression by revealing the portion that is actually translated. In an F_2 progeny from the two maize lines Io and Lc, 42 position shift variants were observed by 2-D PAGE, and this enabled the mapping of the loci responsible for 39 of these position shifts (Damerval et al. 1994). More generally, de Vienne et al. (1996), reviewing studies on maize, barley, pea and maritime pine, showed that about 10% of the polypeptides displayed PS variation between distant genotypes for a given organ. Increasing the number of genotypes and analyzing several contrasted organs would dramatically increase this percentage. Thus, a significant number of structural genes could be mapped using this approach.

The novelty of this database consists in connecting the protein data to genetic and genome mapping data, from structural genes to loci controlling protein quantity. This approach could prove to be a useful tool to reveal candidate genes involved in the quantitative variation of polygenic traits. The peptide sequences were submitted to SWISSPROT, and this maize protein database is accessible via internet (http://moulon.inra.fr/imgd).

Acknowledgements We wish to thank J. d'Alayer and M. Davi from the Institut Pasteur for the peptide sequences, and M. Mansion for technical expertise. This work was in part funded by grants from Rhône-Poulenc Agrochimie, BioAvenir Programme, a grant from the Action Génomes 1991 (Ministère de la Recherche et de l'Espace, France), Groupement de Recherche et d'Etude des Génomes (GIP GREG, Ministère de l'Enseignement Supérieur et de la Recherche, France), and a grant from the European Communities BIOTECH Programme, as part of the Project of Technological Priority 1993–1996. P. Touzet's salary was from the Institut National de la Recherche Agronomique (INRA, France) and from Rhône-Poulenc Agrochimie.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bahrman N, Damerval C (1989) Linkage relationships of loci controlling protein amounts in maritime pine (*Pinus pinaster Ait.*). Heredity 63:267–274

Bairoch A (1991) PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acid Res 19:2241–2245

Bandaranayake H, Esen A (1996) Nucleotide sequence of a beta-glucosidase (glu2) cDNA from maize. Plant Physiol (in press)

Bauw G, De Loose M, Inze D, Van Montagu M, Vandekerckhove J (1987) Alterations in the phenotype of plant cells studied by NH_2-terminal amino acid-sequence analysis of proteins electroblotted from two-dimensional gel-separated total extracts. Proc Natl Acad Sci USA 84:4806–4810

Brzobohaty B, Moore I, Kristoffersen P, Bako L, Campos N, Schell J, Palme K (1993) Release of active cytokinin by a β-glucosidase localized to the maize root meristem. Science 262:1051–1054

Burstin J, Zivy M, de Vienne D, Damerval C (1993) Analysis of scaling methods to minimize experimental variations in two-dimensional electrophoresis quantitative data. Applications to the comparison of maize inbred lines. Electrophoresis 14:1067–1073

Causse M, Santoni S, Damerval C, Maurice A, Charcosset A, Deatrick J, de Vienne D (1996) A composite map of expressed sequences in maize. Genome (in press)

Chen H, Cheng H, Bjerknes M (1993) One-step Coomassie brilliant blue R-250 staining of proteins in polyacrylamide gel. Anal Biochem 212:295–296

Damerval C, de Vienne D, Zivy M, Thiellement H (1986) Technical improvements in two-dimensional electrophoresis increase the level of genetic variation detected in wheat-seedling proteins. Electrophoresis 7:52–54

Damerval C, le Guilloux M, Blaisonneau J, de Vienne D (1987) A simplification of Heukeshoven and Dernick's silver staining of proteins. Electrophoresis 8:158–159

Damerval C, Zivy M, Granier F, de Vienne D (1988) Two-dimensional electrophoresis in plant biology. In: Chrambach A, Dunn MJ, Radola BJ (eds) Advances in electrophoresis. VCH Verlagsgesellschaft, Weinheim, pp 263–340

Damerval C, Maurice A, Josse JM, de Vienne D (1994) Quantitative trait loci underlying gene product variation: a novel perspective for analysing regulation of genome expression. Genetics 137: 289–301

Dessen P, Fondrat C, Valencien C, Mugnier C (1990) BISANCE: a French service for access to biomolecular sequence databases. Compl Appl BioSci 6 355–356

Droog FNJ, Hooykaas PJJ, Libbenga KR, van der Zaal EJ (1993) Proteins encoded by an auxin-regulated gene family of tobacco share limited but significant similarity with glutathione S-transferase and one member indeed shows in vitro GST activity. Plant Mol Biol 21:965–972

Esen A (1992) Purification and partial characterization of maize (*Zea mays L.*) β-glucosidase. Plant Physiol 98:174–182

Faye L, Chrispeels MJ (1985) Characterization of N-linked oligosaccharides by affinoblotting with concanavalin A – peroxidase and treatment of the blots with glycosidases. Anal Biochem 149:218–224

Fernandez J, DeMott M, Atherton D, Mische SM (1992) Internal protein sequence analysis: enzymatic digestion for less than 10 mg of protein bound to polyvinylidene difluoride or nitrocellulose membranes. Anal Biochem 201:255–264

Goldman D, Rathna GP, O'Brien SJ (1987) A molecular phylogeny of the hominoid primates as indicated by two-dimensional protein electrophoresis. Proc Natl Acad Sci USA 84:3307–3311

Grana X, de Lecea L, El-Maghrabi MR, Urena JM, Caellas C, Carreras J, Puigdomenech P, Pilkis SJ, Climent F (1992) Cloning and sequencing of a cDNA encoding 2,3-biphosphoglycerate-independent phosphoglycerate mutase from maize. J Biol Chem 267:12797–12803

Grove G, Zarlengo RP, Timmerman KP, Li NQ, Tam MF, Tu CPD (1988) Characterization and heterospecific expression of cDNA clones of genes in maize GSH S-transferase multigene family Nucleic Acids Res 16:425–438

Helentjaris T (1987) A genetic linkage map for maize based upon RFLPs. TIG 3:217–221

Kawasaki H, Suzuki K (1990) Separation of peptides dissolved in a sodium dodecyl sulfate solution by reversed-phase liquid chromatography: removal of sodium dodecyl sulfate from peptides using an ion-exchange precolumn. Anal Biochem 186:264–268

Kawasaki H, Emori Y, Suzuki K (1990) Production and separation of peptides from proteins stained with coomassie brilliant blue R-250 after separation by sodium dodecyl sulfate-polyacrylamide gel electrophoresis. Anal Biochem 191:332–336

Keith CS, Hoang DO, Barrett BM, Feifgelman B, Nelson MC, Thai H, Baysdorfer C (1993) Partial sequence analysis of 130 randomly selected maize cDNA clones. Plant Physiol 101:329–332

Klose J (1982) Genetic variability of soluble proteins studied by two-dimensional electrophoresis on different inbred mouse strains and on different mouse organs. J Mol Evol 18:315–328

Koshiba T (1993) Cytosolic ascorbate peroxidase in seedlings and leaves in maize (*Zea mays*). Plant Cell Physiol 34:713–721

Kurata N, Nagamura Y, Yamamoto K, Harushima Y, Sue N, Wu J, Antonio BA, Shomura A, Shimizu T, Lin S-Y, Fukuda A, Shimano T, Kuboki Y, Toyama T, Miyamoto Y, Kirihara T, Hayasaka K, Miyao A, Monna L, Zhong HS, Tamura Y, Wang Z-X, Momma T, Umehara Y, Yano M, Sasaki T, Minobe Y (1994) A 300 kilobase interval genetic map of rice including 883 expressed sequences. Nat Genet 8:365–372

Leonardi A, Damerval C, Hébert Y, Gallais A, de Vienne D (1991) Association of protein amount polymorphism (PAP) among maize lines with performances of their hybrids. Theor Appl Genet 82:552–560

Mazo AM, Huang DH, Mozer BA, Dawid IB (1990) The trithorax gene, a *trans*-acting regulator of the bithorax complex in Drosophila, encodes a protein with zinc-binding domains. Proc Natl Acad Sci USA 87:2112–2116

Monroy AF, Gardiner WE, Schwartzbach SD (1988) Photocontrol of organelle and cell type specific changes in the polypeptide composition of *Euglena* and sorghum. Electrophoresis 9:764–773

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. J Biol Chem 250:4007–4021

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factor using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721–726

Pay A, Heberle-Bors E (1991) Isolation and sequence determination of the plant homologue of the eukaryotic initiation factor 4D cDNA from alfafa, *Medicago sativa*. Plant Mol Biol 17:927–929

Rasmussen HH, Van Damme J, Puype M, Gesser B, Celis JE, Vandekerckhove J (1991) Microsequencing of proteins recorded in human two-dimensional gel protein databases. Electrophoresis 12:873–882

Rosenfeld J, Capdevielle J, Guillemot J-C, Ferrara P (1992) In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. Anal Biochem 203:173–179

Shen B, Carneiro N, Torres-Jerez I, Stevenson B, McCreery T, Helentjaris T, Baysdorfer C, Almira E, Ferl RJ, Habben JE, Larkins B (1994) Partial sequencing and mapping of clones from two maize cDNA libraries. Plant Mol Biol 26:1085–1101

Touzet P, Morin C, Damerval C, Le Guilloux M, Zivy M, de Vienne D (1995) Characterizing allelic proteins for genome mapping in maize Electrophoresis 16:1289–1294

Touzet P, de Vienne D, Huet J-C, Ouali O, Bouet F, Zivy M (1996) Amino acid analysis of proteins separated by two-dimensional electrophoresis in maize. isoform detection and function identification. Electrophoresis 17:1393–1401

Van Breusegem F, Villarroel R, Van Montagu M, Inzé D (1995) Ascorbate peroxidase cDNA from maize. Plant Physiol 107: 649–650

Vienne D de, Burstin J, Gerber S, Leonardi A, Le Guilloux M, Murigneux A, Beckert M, Bahrman N, Damerval C, Zivy M (1996) Two-dimensional electrophoresis of proteins as a source of monogenic and codominant markers for population genetics and mapping the expressed genome. Heredity 76:166–177

Zivy M, Thiellement H, de Vienne D, Hofmann J-P (1983) Study on nuclear and cytoplasmic genome expression in wheat by two-dimensional gel electrophoresis. 1. First results on 18 alloplasmic lines. Theor Appl Genet 66:1–7

Zivy M, Thiellement H, de Vienne D, Hofmann J-P (1984) Study on nuclear and cytoplasmic genome expression in wheat by two-dimensional gel electrophoresis. 2. Genetic differences between two lines and two groups of cytoplasms at five developmental stages or organs. Theor Appl Genet 68:335–345

Note added in proof

Accession numbers in SWISS-PROT of the peptide sequences: P80607–P80642