Original papers

Jean-Claude Autran
Philippe Abbal

Laboratoire de Technologie des Céréales
Institut National de la Recherche Agronomique, Montpellier

Wheat cultivar identification by a totally automatic soft-laser scanning densitometry and computer-aided analysis of protein electropherograms

A computer-aided analysis based on soft-laser densitometry of electrophoretic patterns is described. The system has the potential to automatically identify wheat cultivars from normalized sodium dodecyl sulfate – polyacrylamide gel electropherograms. The principle emcompasses autocalibration of the relative mobilities of components of the unknown pattern using standards of M_r that bracket each pattern on the pictures, automatic identification of peaks and baseline correction, and simplified estimation of the peak intensities. The algorithm provides a normalized array that can be automatically compared with a library of standard cultivar patterns in order to determine those having the highest similarity percentage. The outputs of the program are 1) relative similarity plots which show the percentage of similarity between the unknown cultivar to that of each of the common standard cultivars put in a pattern library, 2) the names of cultivars in the library having the most similar patterns and 3) computer-generated electrophoretic graphics of these cultivars. This system is not intended to identify closely related cultivars but it is particularly recommended for cultivar identification without well-trained personnel, familiar with band location and nomenclature.

1 Introduction

Gel electrophoresis is increasingly being used as an aid in differentiating and identifying cultivars of wheat and other crops for genetic studies or in commercial deliveries [1–3]. Electropherograms of wheat storage proteins produced under well-defined conditions are considered as unique fingerprints of the cultivars. The most polymorphic fractions (gliadins, glutenins) are reliable genetic markers on which a number of discriminating systems has been based [4–8]. The electropherogram of an unidentified cultivar is usually normalized in the form of a "variety formula" or "cultivar array", consisting of a complete set of mobility-density pairs, by using reference bands or a reference pattern. It can then be compared with each of the standard patterns listed in a catalog by using chemotaxonomic keys [9, 10].

In spite of considerable innovations in the computerization of data acquisition and processing, which have found successful applications in the clinical context, most of the laboratories in charge of wheat cultivar analysis are still using a visual examination of the electropherograms, and most of the operations of comparison, classification and identification are still being carried out manually. In the few cases where the use of a computer has been introduced in the process, the most critical steps, especially assignment of the bands to standard mobilities and determination of intensities, are still performed manually. They require skilled co-workers, well aquainted with the main traits of the patterns, who are able to identify those bands that will be used as mobility standards and to decide whether a minor component must be considered as a band or as belonging to the background. For instance, Lookhart *et al.* [11] have reported a simple computer-assisted method for identifying wheat cultivars by calculating the similarity of their gliadin polyacrylamide gel electrophoresis (PAGE) patterns with those of standard varieties stored in the computer memory. However, the distances of the bands had to be measured manually and each band density was subjectively assigned a numerical value from 1 to 5 by visual examination and all values had to be entered on the keyboard. An improvement in the accuracy of this system has been proposed in one of our previous reports [12], adapting it to European wheat cultivars, based as well on PAGE as sodium dodecyl sulfate (SDS)-PAGE protein patterns and compatible with MS-DOS standard systems.

Bushuk *et al.* [13] have first developed a computer analysis of gliadin PAGE patterns that were digitized and put into a computer, but a visual inspection of the patterns was still necessary to calculate the relative mobility data. The same authors developed a more sophisticated and accurate computer-aided analysis of gliadin electropherograms [14–16] using a three-reference band standardization, but the data were acquired from photographic prints in a semiautomatic fashion using a digitizing tablet that could put values directly into a computer file, thus obviating the need to manually transcribe data, but, however, after a manual pointing out of each band of the pattern, as well as of the three components used as standards. In the last issue of the program package, Sapirstein and Bushuk [17] report the possibility of an automatic quantification of electropherograms that looked satisfactory as far as peak detection and reproducibility was concerned, but that did not deal with baseline correction and that seemed to come up against the automatic setting of the reference scale of mobilities: the standardization of relative mobilities was based on a single reference band (the most intense detected by the computer) and not on three as in the manual version.

Correspondence: Dr. Jean-Claude Autran, Laboratoire de Technologie des Céréales, Institut National de la Recherche Agronomique, 9 Place Viala, F-34060 Montpellier Cedex, France

Abbreviations: **CCD**, charge coupled device; **cv**, cultivar; **HMW**, high molecular weight; **LMW**, low molecular weight; M_r, relative molecular mass; **MS-DOS**, Micro Soft Disk Operating System; **PAGE**, polyacrylamide gel electrophoresis; **SDS**, sodium dodecyl sulfate; **SDS-PAGE**, sodium dodecyl sulfate - polyacrylamide gel electrophoresis; **SI**, similarity index

0173-0835/88/0505-0205 $02.50/0

Designing a totally automated system presents multiple difficulties. Obviously, reliable and accurate systems from electrophoretic data acquisition, *e. g.* soft-laser densitometers, charge coupled device (CCD) cameras, video cameras, are now available and the data can be processed by microcomputers. Also, detection of peaks and valleys or quantification of peaks from a given pattern are no longer a major problem since several algorithms for baseline correction, peak fitting and integration have been reported [18–21], some being commercially available. However, experimental variability in factors associated with gel electrophoresis does not usually allow perfect reproducibility in the resolving power of a gel. Identical automatic separations and evaluations of multiple peaks, skewed non-Gaussian peaks or shoulders are not always ensured since there is always an ultimate limit in resolution beyond which automatic evaluation will not be reproducible. All commercially available software allows for manual steps in peak detection, *e.g.* to add missed peaks, to delete extra peaks, to achieve a better separation of multiple peaks (LKB 2190 GELSCAN) or to move the scan in either direction to align it with the standard pattern [22].

On the other hand, experimental variability generally precludes the comparison of absolute migration distances among gels [23]. Electrophoretic migration is not always a linear function of time, some regions being shrinked or stretched, sometimes only because of handling of the gel slab, making it necessary to use several standards of mobility steadily distributed in the different regions of the pattern and unambiguously picked out by the software. Unfortunately, most of the systems that have currently been developed allow autocalibration in optical densities, but not in relative mobilities. Even in relatively sophisticated systems (Nelson Analytical), calibration in mobilities with redrawing of the curve on a relative mobility scale is seldom proposed and, when available, is based on a single reference band. Unlike serum proteins for which laboratory diagnostic software has been made available [24–26] due to their standard overall electrophoretic profile, wheat storage proteins also present specific difficulties. The various cultivars yield PAGE or SDS-PAGE patterns that do not at all contain the same components. For instance, a gliadin pattern of each wheat cultivar consists of 20–25 gliadin bands that are often closely stacked or fused, taken in a set of about 60 possible bands, not one being common to all cultivars. It is therefore difficult to achieve an automatic and unambiguous identification of those bands that have to be used for the normalization of the unknown pattern. If several internal standards were introduced, the problem would still remain because of the large polymorphism in mobilities and intensities of the wheat proteins bands.

In this paper, we report the first attempt of a totally automated system at wheat cultivar identification that offers the following features: (i) data acquisition of SDS-PAGE protein patterns by a laser densitometer; (ii) automatic normalization of the relative mobilities based on the identification of M_r markers loaded on each side of the unknown pattern; (iii) baseline correction; (iv) detection of the peaks and setting of the cultivar array; (v) comparison with a library of standard cultivar patterns and identification of the unknown sample by similarity calculation. Unlike more sophisticated procedures for peak fitting and accurate quantification, our software has given preference to simplicity, full automation and optimal discrimination of the most extensively grown cultivars in order to make it available for routine use in wheat control laboratories.

2. Materials and methods

2.1 Wheat samples

The wheat samples used in this study were provided by G.E.V.E.S. (Groupe d'Etude des Variétés et des Semences, La Minière, 78280 Guyancourt, France).

2.2 Chemicals and reagents

All chemicals used were of reagent grade. The mixture of protein molecular weight standards (rabbit muscle myosin, M_r *205 000; E coli* β-galactosidase, M_r 116 000; rabbit muscle phosphorylase b, M_r 97 400; bovine serum albumin, M_r 66 000; ovalbumin, M_r 45 000; carbonic anhydrase, M_r 29 000) was from Sigma (St. Louis, MO, USA)

2.3 Electrophoresis

The ground grain (50 mg) was suspended in 800 µL of a medium containing 2 % w/v SDS, 5 % v/v 2-mercaptoethanol, 0.001 % w/v Pyronin Y, 10 % v/v glycerol and 0.063 M Tris-HCl, pH 6.8, according to Payne *et al.* [27]. 10 µL of extracted proteins were loaded onto a 13 % SDS-PAGE gel and electrophoresed as previously described [28]. Wheat protein extracts and molecular weight markers were alternatively loaded onto the gel as illustrated in Fig. 1. Proteins were stained with 0.5 % w/v Coomassie Brilliant Blue R-250 in water-methanol-acetic acid (53-40-7) for 30 min at 80 °C and destained in water-methanol-acetic acid and stored in water. The gels were photographed on high contrast film and 90 × 110 mm prints were used for densitometric scanning.

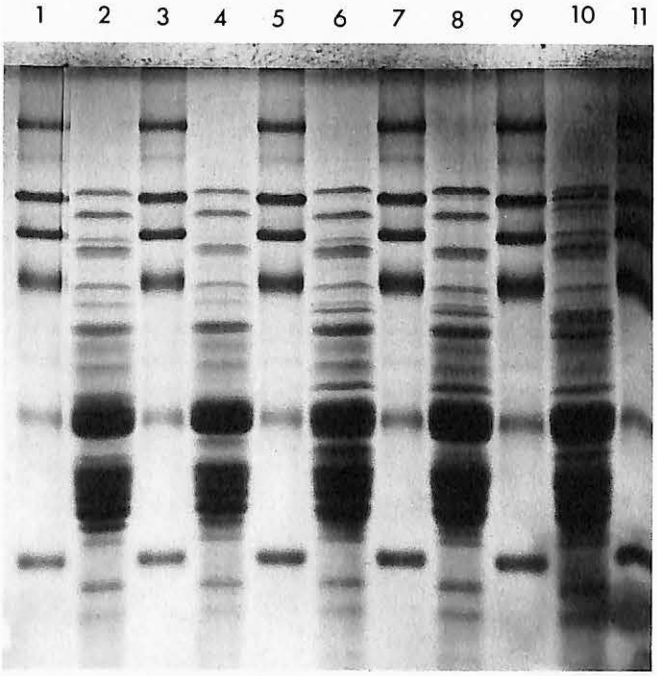

Figure 1. Example of layout of SDS-PAGE patterns for automatic scanning densitometry and computer-aided analysis. Lanes (1), (3), (5), (7), (9), (11) Sigma high M_r standards. Wheat protein extracts from the following cultivars: (2) Talent, (4) Promentin, (6) Capitole, (8) Hardi, (10) Récital.

2.4 Band nomenclature

Band nomenclature was derived from Berger and Le Brun [8] in which reference band No. 100 corresponded to the major alpha-gliadin group [28]. Cultivar formulas including band mobility and intensity were automatically determined as described below.

3 Computer-aided analysis of protein electropherograms

3.1 Hardware description

Electrophoretic patterns were scanned by a soft laser densitometer (Ultroscan 2202, LKB Instruments, Bromma, Sweden) designed for evaluation of high resolution gel electrophoresis and consisting of a helium-neon laser (632.8 nm wavelength) and a unique optical system. Its microprocessor allowed automated scanning processes. For scanning wheat protein patterns, standard settings were 200 mm/min scan speed, 100 mm scanning length, and 1.0 optical density unit. The densitometer provided an analog electric signal of 1000 mV at full scale and a start pulse. It was interfaced with IBM-PC/XT computer 512 KB RAM with a 10 MB hard disk. The analog signal of the densitometer was channeled to the computer by an IBM data acquisition and control card that provided eighteen binary inputs, two digital to analog outputs and four analog to digital input channels. It allowed synchronization and acquisition of different ranges of analog signals (-5/+5 V, 0/10 V and -10/+10 V) at the frequency of 1 MHz by point. The twelve-bit analog-to-digital converter circuits of the IBM card gave 4096 different levels. Final accuracy was about 2.5 mV for an analog input signal varying between 0 and 1000 mV. Hard copies of the densitometric tracings were obtained from a dot matrix printer.

3.2 Software description

The package consists of seven sections: (1) raw data acquisition; (2) noise reduction; (3) scaling of the raw densitometric curve on a relative mobility basis; (4) baseline correction; (5) peak detection and scaling of height coordinates; (6) setting the cultivar formula; (7) quantification of the extent of homology between the unknown pattern and all reference patterns encoded in the data base. The development tools are Microsoft Basic compiler, IBM data acquisition and programming support and Microsoft macroassembler. In the main (fully automatic) option, the pattern is processed from M_r standards run on each gel. In secondary options, the patterns can be processed either upon real time scanning or from stored data using manual reference bands.

3.2.1 Data acquisition

To scale the raw peak positions of each unknown sample on the basis of standard proteins, the densitometer has been programmed to scan three patterns one after another: (i) first standard, (ii) unknown wheat pattern, (iii) second standard. Unlike previous reports using one (or three) standard band(s) belonging to the wheat protein pattern [11, 14, 17], this system was based on molecular weight calibration kits loaded on both sides of the unknown pattern in order to control eventual error across the gel, curvatures or unlinearities. The system made it possible to acquire and digitize electric signal data as a function of the running distance. The full scale corresponded to a 500 mV signal. For each densitometric curve, 640 points were acquired with a frequency of 10 s^{-1} and, therefore, each densitometric curve was stored as a file of 640 raw coordinate data.

3.2.2 Noise reduction

A simple smoothing filter has been used in averaging a number of neighboring points [18]: the ordinate of each of the 640 points was replaced by the mean of the ordinates of the points located within a moving neighborhood of length $2j+1 = 5$ (called "window" of size $2j+1$) (see Section 3.2.3).

3.2.3 Scaling in relative mobilities

This operation was based on the position of 6 standard proteins that were selected from preliminary runs in such a way that their positions frame the major wheat gliadin and glutenin bands (Table 1). The abscissa corresponding to the maximum height of each standard peak was assigned a relative mobility in agreement with the mean position of the different wheat protein bands according to the nomenclature proposed by Berger and Le Brun [8]. (A scale based on apparent molecular weights could be used alternatively).

From the two densitometric curves of standard proteins, a resulting curve was inferred by averaging the ordinates of each of their 640 points. This curve essentially consisted of 6 well defined peaks with little background since it corresponded to a mixture of 6 pure molecular weight markers. Therefore, the abscissa corresponding to the maximum height of each of the 6 peaks could be simply used for scaling in relative mobilities the wheat protein pattern located between them (Fig. 1), without going into quantification or baseline correction.

These abscissa were automatically determined as follows: (i) Elimination of all minor peaks or residual electric noise by removing the points with ordinates lower than a threshold of 25 mV (5 % of the full scale). (ii) Selection of a first subassembly of maximum points by using a moving window of length $2j+1 = 9$ and storing the abscissa A_j of the points having an ordinate higher than the arithmetic mean between the one A_{j-4} and A_{j+4} points and higher than the one of both the A_{j-4} and A_{j+4} points (see Section 3.2.4). These points make up the maxima regions of the peaks and are similar to the "convex kernels" reported by Lasters *et al.* [21]. (iii) In a second step, the six maxima of the standard proteins were determined by selecting the absolute maximum within each of the 6 "convex kernels". The abscissa of these 6 points were stored and were used for transforming the wheat protein curve into a relative

Table 1. Assigned relative mobility of molecular weight standard proteins

Standard protein	Apparent molecular weight	Assigned relative mobility
Myosine	205 000	15
β-Galactosidase	116 000	29
Phosphorylase b	97 400	40
Bovine serum albumin	66 000	56
Ovalbumin	45 000	87
Carbonic anhydrase	29 000	115

mobility scale (between 29 to 115) by simple linear interpolation that allows aligning and stretching the pattern within each of the five intervals determined by the six abscissa values.

3.2.4 Baseline correction

A baseline correction is required since the visually observable bands are often superimposed on a non-constant background and the density envelope between peaks does not usually return to zero. The reasons are the presence of protein aggregates that give rise to streaks in the patterns, or the inhomogeneous absorbance of the gel. This baseline was determined as follows: (i) Identification of the minima of the curve (by which the baseline curve will be approximated) by a similar algorithm as for maximum peak detection, but using a different size of moving window (length $2j+1 = 11$). Instead of keeping all "concave kernels" thus determined, a simple system consisted in storing a limited number of minimum points evenly spread out in the whole pattern. Preliminary trials have shown that the determination of seven intervals in the whole curve and storing one minimum point per interval allowed a satisfactory baseline calculation. (ii) Generation of the baseline curve as a third degree polynomial, from the 7 minimum point coordinates, according to a Gauss-Newton method for resolving linear equations [29]. (iii) Subtraction of the baseline curve from the normalized wheat densitometric curve (see Section 3.2.3), plotting and hard disk storage of the background corrected curve.

3.2.5 Identification of peaks from the unknown sample curve

The principle of automated identification of the peak locations was the same as mentioned in Section 3.2.3 for peak detection from the standard protein curve, using a moving window of $2j+1 = 9$ and storing one mobility at most per kernel of length 11 abscissa units in order to remove the eventually closely located peaks that would not correspond to reproducibly detectable bands. This step made it possible to obtain a well-defined list (with values approximated to the closest integer) of peak mobilities.

3.2.6 Scaling peak intensity and setting up of the cultivar array

The system of peak intensity scaling was aimed at automatically yielding a pattern consistent with the manually e-stablished SDS-PAGE patterns [7, 30]. These usually consist of about 15 bands with numerical values for intensities ranging from 1 to 5. It has been observed that, for any cultivar, the cumulated value of the intensity levels of the bands was equal to or near the value 45. Accordingly, the scaling of peak intensities was obtained as follows: (i) the raw heights of the peaks identified in Section 3.2.5 from the background corrected curve were stored, (ii) these height intensities were cumulated and the sum was assigned a value of 45, and (iii) the raw intensity of each peak was then converted and approximated by rule of three into an intensity numerical value (1, 2, 3, 4, or 5). The set of mobility/intensity pairs calculated from the unknown pattern without any manual intervention corresponded to the cultivar array. It had the same file structure and can be processed in a similar way as manually established arrays.

3.2.7 Calculation of electrophoretic pattern homology and identification of the unknown cultivar

The software previously described for comparison of manual data [12] was used. It allowed the comparison of the unknown pattern with counterpart protein patterns in the cultivar data base, taken one at a time, and was based on the calculation of three similarity indexes: (i) The first index (SI_1) was based on the number of matching bands (MB) or nearly matching bands (NMB) (mobility values differing by + 1 unit). For each pair of matching or nearly matching bands, a match value (MV) was assigned according to Lookhart *et al.* [11]. The more similar the matching bands were in density, the higher the match value. For instance, when the intensity levels of the pair of matching bands were 5 and 5, 5 and 3, or 5 and 1, the calculated MV were, respectively, 2.00, 0.80 and 0.30. The SI_1 index was then obtained by summing the MV for all matching protein bands of the unknown and standard cultivars. (ii) Since matching bands are more likely to correspond to the same protein species than nearly matching bands, a second index (SI_2) was introduced in order to give less weight to the latter. SI_2 was based on the ratio: $MB/(MB+NMB)$. (iii) Because different cultivars may contain different numbers of other bands (beside MB and NMB) that have not been taken into account in SI_1 and SI_2, the third index (SI_3) allowed making a correction when the compared patterns contain different total numbers of bands. Whereas UNB and SNB, respectively, refer to the total number of bands of unknown and standard cultivars, SI_3 calculation was based on the ratio (comprised between 0 and 1) of the least to the highest of the two numbers of UNB and SNB. The overall similarity index SI was the product: $SI_1 \times SI_2 \times SI_3$.

Finally, the computer outputs are: (i) A graph of the relative percent similarities of the unknown cultivar *versus* all the cultivars of the data base in the order of decreasing similarity. (ii) A listing of the name of the identified cultivar (when a couterpart with similarity index higher than 90 % was found) or, if such is not the case, the names and the SI of the three next most similar cultivars. (iii) A scheme of the electrophoretic patterns of these cultivars for visual control of the results.

3.2.8 Other options

In a semiautomatic subroutine, the operator could store raw data or normalized curves on the hard disk and delay their processing and their comparison with the data base. A manual processing of the curve was also provided: the position of mobility standards could be determined on the screen and the curve could be redrawn according to the normalized relative mobility.

4 Results

4.1 Setting of the cultivar array from the densitometric curve

Fig. 2 illustrates a typical densitometric curve obtained after baseline subtraction from an SDS-PAGE pattern of wheat proteins. From the numerical data, a normalized array, consisting of mobility/intensity pairs, was calculated by using the M_r standard bands. The calculated arrays of the most extensively grown French cultivars are shown in Table 2, using a solid data display system as in Zillman and Bushuk [30].

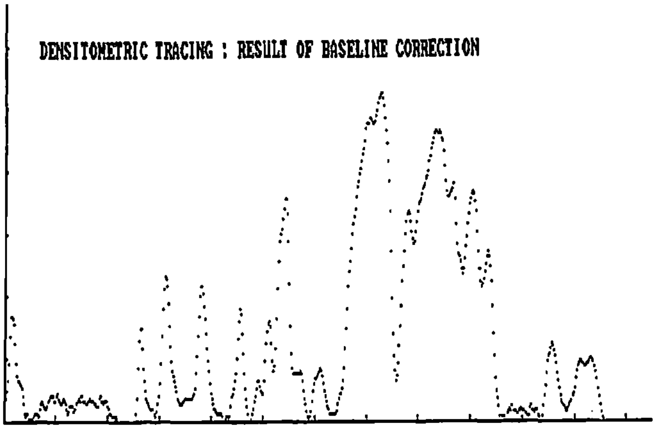

Figure 2. Densitometric tracing of the SDS-PAGE pattern of proteins extracted from wheat cv. Fidel after baseline correction.

4.2 Identification of an unknown cultivar

The interpretative outputs from this program were compared with previous routine manual interpretations of wheat protein electropherograms. Fig. 3 a, b, c illustrate the different cases that occur when an "unknown" cultivar pattern was selected and compared to the data base. In Fig. 3a, the sample has a similarity index of 95 % with one of the cultivars (Prinqual), and then only 65 % and below with all the others. It is concluded that, in such a case (see Section 5.4), the unknown can be unambiguously identified as cv. Prinqual. In Fig. 3b, the output indicates that cv. Hardi is identified. However, it can be seen that the similarity of cv. Hardi over several others pertains to only a few points. In Fig. 3c, no cultivar has a pattern similar to the one of the unknown, the most related having about 70 % similarity only, and the output indicates that, in that case, no conclusion is drawn.

4.3 Schemes of the most similar electrophoretic patterns

For a more accurate control of the diagnostic, the schemes of the three next most similar patterns are given (Fig. 4).

5 Discussion

5.1 Data acquisition

An illustration of the resolving capacity and sensitivity of soft-laser densitometry after baseline subtraction is shown in Fig. 2. With the exception of the fast-migrating region, most of the bands were well resolved, confirming that a soft-laser densitometer brought significant improvements in both resolution and sensitivity over previous systems [31], even when scanning photographic prints, and was adapted to the analysis of wheat protein electropherograms consisting of many closely stacked bands. The different levels of processing of the curve (noise reduction, baseline correction, detection of maxima) are based on the use of a moving window that includes j points on both sides of the moving abscissa. The window length ($2j+1$ points) has been optimized according to the type of problem to be solved, using previous results [18], sampling theorem [32] and preliminary continual approaches. A compromise had to be determined between a minimum of noise in the processed signal and a loss of significant peaks. For instance, the most reliable representation for the initial noise reduction step (Section 3.2.3) was found for a j value equal to 2 (window length $2j+1 = 5$). For the detection of maxima kernels (Sections 3.2.3 and 3.2.5), the optimum was found when $2j+1 = 9$, while for the detection of minima kernels (Section 3.2.4), which correspond to regions of the curve flatter than the maxima of the peaks, the optimum was found when $2j+1 = 11$.

5.2 Data base

The data base previously described [12] which consists of manually determined patterns can be used. However, because of slight differences between manual and computer-aided patterns, more accurate identification can be achieved when using calculated patterns for setting the data base. Accordingly, several samples of each cultivar were submitted to different electrophoretic runs and the mobility/intensity values were averaged and entered once and for all on the keyboard to make a reference pattern of the cultivar in the data base (see below: cv. Primadur-S in Table 3).

Table 2. SDS-PAGE protein arrays of the 12 most extensively grown French wheat cultivars[a)]

	20	30	40	50	60	70	80	90	100	110
Mondur	1	5 4	1 1	1	3	5	3 5 3	4 4	5	
Cando		5 4		2	3 4	2 3	3 4 1	1 2 4	4	
Capdur		5 3 2		1	4	5	3 5 3	2 3 5	4	
Hardi	5	5	3	2	2 4	2 2 3	5	1 5	4	
Capitole	5	5	1 2	2	2 3	1 2 2	4	3 4	4 4	
Prinqual	5	5	3	1	2 2 2	1 2	5	3 5	5	
Talent	5	5	1 3	2	1 3	2 2	5	55	5	
Beauchamp	5	5 2	3	1 2	1 2 3	1 2	5	5	5	
Arminda	4	5	2	2	1 2 4	1 1	5 5	3 5	4	
Camp remy	4	5 1	2	1	1 2	1 1 2	4 5	4	4 4	
Fidel	3	5 2	2	1	2 3	1 1 1	5 5	1 3	5 4	
Festival	4	5	2 3	2	3	1 1 2	5	2 4	4 4	
	20	30	40	50	60	70	80	90	100	110

a) The 3 upper patterns are from durum wheat type, the other 9 patterns are from soft wheat type. Relative band intensity is expressed in the 1 to 5 scale (1 is lightest, 5 is strongest).

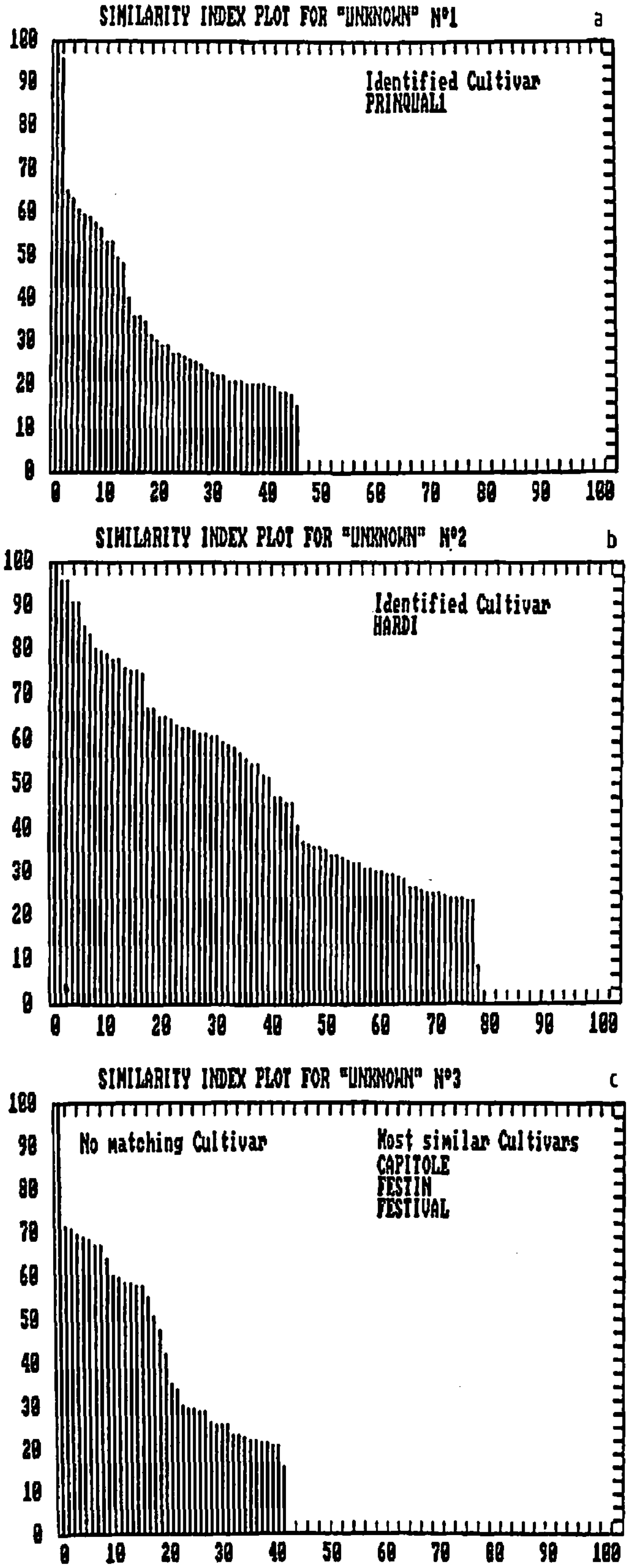

Figure 3. Relative percent similarity plot of an unknown pattern *versus* the standard cultivar database. (a) The computer program identified the unknown as cv. Prinqual. (b) The computer program identified the unknown as cv. Hardi, but no unique identification can be achieved because several other cultivars fall into a 5 % similarity span. (c) No counterpart protein pattern could be found in the database.

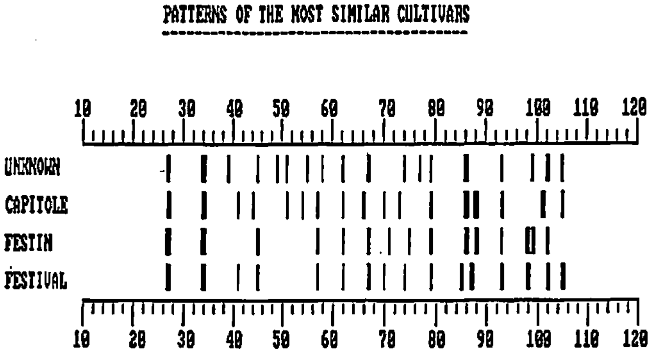

Figure 4. Mobility-density schemes of the cultivars having the 3 next most similar patterns to the unknown of Fig. 3c.

5.3 Reproducibility and accuracy of the cultivar array

The arrays shown in Table 2 are similar to those visually established and previously reported [8]. The only differences involve minor slow-moving components and, except for the more poorly resolved fast moving region, all major bands previously identified appear unambiguously in the automatically calculated array. In order to specify the limits of the method, experimental fluctuations were checked in calculating the arrays of a given cultivar from several scannings of the same pattern, from scanning of the same protein extract loaded at different places on the same SDS-PAGE gel, or from different protein extracts corresponding to different growing locations of the same cultivar and run on different gels. These results are illustrated in Table 3 for the durum wheat cultivar Primadur. The variation was quite low on repeated scanning of the same pattern along slightly different track positions or on scanning the same cultivar sample (or several cultivar samples) loaded at different positions of the same gel. In most cases, the differences in the calculated arrays only involve minor bands detected with an integer value of intensity which can fall either to 1 or to 0. Another difference is caused by poorly resolved peaks, in which one or several maxima may be detected by the software when analyzing different scans. Also, the approximation to the closest integer of the calculated mobilities may bring some fluctuation and affect the similarity index. However, in such experiments, the variation of the relative mobilities of the peaks was never higher than 1 unit and similarity indexes between the different repeats did not drop below 90 or 95 %. When different electrophoretic runs were compared, a higher variation was noticed (especially in the fast moving regions), indicating that an important limiting factor of the method was the constancy in the resolution of the patterns. However, when comparing a set of routinely run patterns, the similarity index remained within a range of 85–95 %.

Such a minor variation in the *IS* output has been made possible by adjusting the software to reduce the effect of the experimental differences between automatically acquired SDS-PAGE patterns. As reported above, some regions of the pattern, (especially those with M_r lower than 35 000) consist of large and heavy peaks with apparently little interest in cultivar discrimination. Therefore, the similarity indexes were calculated by giving special emphasis to the slow-moving (HMW-glutenin) and intermediate (LMW-glutenin, beta-, gamma- and omega-gliadin) components, the most important in cultivar differences. For instance, when comparing the *IS* between different cultivars, taking several repeated patterns

Table 3. Reproducibility of the automatically determined cultivar array from 10 different scannings of the cv. Primadur pattern[a)]

	20				30			40	50		60	70			80			90						100	110
Primadur-1					4	1	4			1	33			5		5	2	2		3		4		4	
Primadur-2				1	4	1	3			1	33			5		5	2	1		3		4		4	
Primadur-3				1	5	1	4			1	33			5		5	2	1		3		5		4	
Primadur-4					5	1	4			1	33			5		5	2	2		3		4		4	
Primadur-5					5	11	4			1	33			5		5	2	1		3	5			4	
Primadur-6					4	1	4			1	33			5		5	2	1		3		5		4	
Primadur-7		1		1	4	1	3			1	33			5		45	22	1		3		5		3	
Primadur-8					4	11	4			1	33		5			5	2	2		3		5	2	4	
Primadur-9			1		5	1	4			1	33			5		25	2	2	3		4		2	4	
Primadur-10					5	1	4			1	33		5			5	12	2		3		5		4	
Primadur-S					5	1	4			1	33			5		5	2	2		3		5		4	
Primadur-M					3		2			1	32			5		5	3	2	2				5	2	
	20				30			40	50		60	70			80			90						100	110

a) References 1 to 3 corresponded to scan repeats of the same lane; 4 to 6, to scans of different samples loaded on the same SDS-PAGE gel; 7 to 10, to scans of samples analyzed in different SDS-PAGE experiments. Primadur-S was the standard averaged pattern, as it was put into the data base. Primadur-M corresponded to the visually established pattern.

from each, the variation within each cultivar remains in the range of 10–15 %. This is illustrated in Fig. 5, with the 3 cvs. Prinqual, Hardi and Fidel, with 10 repeats from each. The *IS* of the patterns compared to one sample of cv. Prinqual taken as "unknown" is clearly distributed in three levels, the mean and standard deviation being, respectively, 92.3 (4.0), 47.1 (4.5) and 19.2 (3.3). This indicates that significant differences occur between these three groups of patterns. Similar experiments with patterns of the major French cultivars have shown that reliable identification was possible only if one cultivar could be favored by at least 15 percent points of similarity over any other cultivar.

5.4 Assessment of the computerized diagnostic program for cultivar identification

The previous results throw light on the interpretation of the different interpretative outputs shown in Figs. 3a, b, c. Thus, in Fig. 3a, considering that a normal fluctuation within a cultivar allowed the index to remain in the 85–100 % range, the unknown No. 1 could be unambigously identified as cv. Prinqual. Fig. 3b reflects the fact that the unknown cultivars and several others are likely to be members of the same group of patterns, which is usually the case when considering closely related genotypes. Although cv. Hardi is identified as the most similar to the unknown, its similarity advantage over the others is of a few points only, so that no accurate identification can be achieved. In Fig. 3c, no matching cultivar can be found, the most related having about 70 %.similarity only. In that case, no conclusion can be drawn, either because the corresponding cultivar was not introduced in the data base, or because an insufficient resolution of the pattern brought about a failure of the system in calculating an accurate or reliable array.

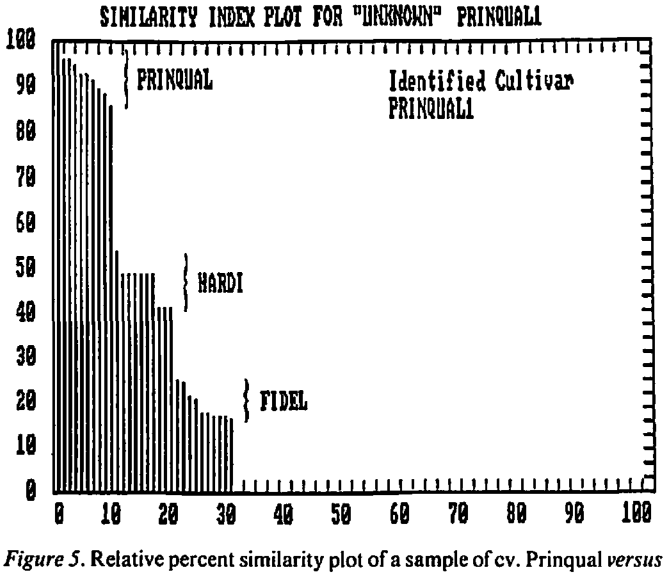

Figure 5. Relative percent similarity plot of a sample of cv. Prinqual *versus* 10 repeated patterns of each of the cvs. Prinqual, Hardi and Fidel.

Generally, unrelated genotypes, such as spring wheats (cvs. Prinqual, Florence-Aurore) or cultivars having uncommon patterns because of unrelated parents (cvs. Fidel, Avalon) were unambigously detected with a 20 or 30 % similarity span over others. On the contrary, several winter wheats (cvs. Capitole, Festival) that have similar and sometimes identical protein electrophorograms could not be discriminated. In practice a data base, consisting of extensively grown cultivars such as Arminda, Beauchamp, Camp-Rémy, Festival, Fidel or Prinqual, allows unambigous identification. It therefore seems adapted to the normal deliveries of a mill plant. However, when closely related cultivars are analysed or introduced into the data base, the discrimination achieved by the similarity index may become insufficient. For instance, mistakes may occur between cultivars with similar patterns, such as Capitole, Festival and Frandoc, or Aminda and Festin.

5.5 Advantages and limitations of the new system

Unlike previous systems for wheat cultivar identification, occasionally involving manual or visual operations, our system is totally automatic. The operator need only place the photograph of the patterns into the densitometer and store the positions of the lanes to be scanned. After pressing the start button, no additional human operation is required until the name of the identified cultivar is displayed on the screen. From the acquired data, the software automatically allows all the following operations: detection of the main peaks of the electrophoretic pattern, calculation of the peak coordinates and scaling into mobility/intensity pairs in order to produce a normalized cultivar array, comparison of the unknown array to the data base with a classification by decreasing similarity, leading to a diagnosis of the cultivar.

The preference given to a totally automatic system, easy to use on a personal computer and with short calculation time, has directed the different solutions adopted in the software. A certain number of limitations, the penalty of total automation, will be discussed below. A first limitation originates in the usual drawbacks of densitometry itself, especially when working from a paper print of the electrophoretic gels, already discussed elsewhere [33]. While there is no difficult in using the algorithm and automatically determining a cultivar array (assuming that cultivar differences are obvious enough to be picked up by the densitometer), another problem may arise with reproducibility of the automatic array since it depends on the quality of band separation. The major limiting factor for reliable identification is therefore the experimental step. An automatically determined array cannot be as accurate or representative of the cultivar as the averaging of several manually picked and controlled patterns. As expected, from poorly resolved electropherograms, the computer may fail in establishing an accurate array and may produce a wrong diagnosis.

With regard to the specific difficulties of wheat protein patterns (presence of both sharp HMW bands and stacked or fused peaks with no well-defined maximum position, and variable streaks due to partly aggregated fractions), it was illusory to look for overly high precision that could not be reproducibly retained between different electrophoretic runs or scanning experiments. This explains the adjustment of the software to deal with possible variations in the array, for instance: (i) relatively low accuracy in mobility calculation: values simply approximated to the closest integer, (ii) fast and simple baseline correction, (iii) simplified peak detection and peak height computation that takes into account the experimental variation in approximating to 5 intensity levels only, instead of using a highly accurate peak integration or densitometric quantification based on Gaussian peak fitting, (iv) flexible computation of the similarity index that bears some differences in the automatically acquired pattern, for instance, in considering the bands whose mobility values differed by +1 unit as not necessarily different and in giving particular emphasis to the most discriminating regions.

Other limitations of the system result from the use of M_r markers, restricted to evaluation of SDS-PAGE gels. Most of the steps of the algorithm were optimized for the particular features of SDS-PAGE wheat protein patterns. For an extension to other proteins, there is no doubt that some parameters (range of M_r markers, speed of data aquisition, threshold of peak detection, length of the windows used in peak detection) would have to be adjusted to the specific characteristics of the patterns. Also compared to the short computation time (a few seconds for a data base consisting of 80 arrays), the scanning time is limiting (about 2 min for scanning the three patterns). The new systems, based on video or CCD camera [34, 35], should be introduced for data acquisition intended for fast cultivar identification. This should allow almost instantaneous and possibly more accurate data acquisition due to the possibility of automatically averaging and processing several passes of each lane.

The main advantage of the system is complete automation. It is therefore practical for operators with minimal skill or little knowledge of the wheat cultivar patterns and of the standard band positions. However, it requires working on relatively well resolved electrophoretic patterns, and the number of cultivars that can be unambiguously identified is lower than when working with manually corrected patterns via keyboard input. The system is certainly not recommended for the exhaustive and accurate characterization of the whole set of cultivars of, for instance, the Official Catalogue of a country, or for carrying out genetic studies with closely related cultivars, nor does it seem that the present algorithm is adapted to deal with identification of mixtures of cultivars. It could be convenient, however, for laboratories dealing with a limited number of cultivars. It is especially suitable for the specifications of wheat control laboratories of millers or storage companies who make extensive use of wheat protein electrophoresis, with local wheat supplies and deliveries consisting of less than 20 (and sometimes less than 10) main cultivars.

The excellent technical assistance and personal engagement of Ms. Renée Berrier and Ms. Joëlle Dusfour are greatly acknowledged. This work was supported by the Direction Scientifique I.N.R.A. des Industries Alimentaires et Agricoles and by the Contrat de Plan Etat – Région (Pôle de Recherche – Industrie pour l'Agro-Alimentaire Méditerranéen).

Received November 5, 1987

6 References

[1] Autran, J. C., in: Gregory, C. (Ed.), *Universalia*, Encyclopaedia Universalis France, Paris 1976, pp. 171-174.

[2] Autran, J. C., *Biofutur* 1986, *51*, 121-126.

[3] Wrigley, C. W., Autran, J. C. and Bushuk, W., *Adv. Cereal Sci. Technol.* 1982, *5*, 211-259.

[4] Bushuk, W. and Zillman, R. R., *Can. J. Plant Sci.* 1978, *58*, 505-515.

[5] Shewry, P. R., Faulks, A. J., Pratt, H. M. and Miflin, B. J., *J. Sci. Food Agric.* 1978, *29*, 847-849.

[6] Du Cros, D. L. and Wrigley, C. W., *J. Sci. Food Agric.* 1979, *30*, 785-794.

[7] Jones, B. L., Lookhart, G. L., Hall, S. B. and Finney, K. F., *Cereal Chem.* 1982, *59*, 181-188.

[8] Berger, M. and Le Brun, J., *Industr. Céréales* 1985, *37*, 17-25.

[9] Autran, J. C. and Bourdet, A., *Ann. Amélior. Plantes* 1975, *25*, 277-301.

[10] Dal Belin Peruffo, A., Pallavicini C., Varanini, Z. and Pogna, N. E., *Genetica Agraria* 1981, *35*, 195-208.

[11] Lookhart, G. L., Jones, B. L., Walker, D. E., Hall, S. B. and Cooper, D. B., *Cereal Chem.* 1985, *60*, 111-115.

[12] Autran, J. C. and Abbal, P., *Ind. Alim. Agric.* 1986, *103*, 535-545.

[13] Bushuk, W., Sapirstein, H. D. and Zillman, R. R., *Cereal Foods World* 1978, *23*, 496-501.

[14] Sapirstein, H. D. and Bushuk, W., *Cereal Chem.* 1985, *62*, 372-377.

[15] Sapirstein, H. D. and Bushuk, W., *Cereal Chem.* 1985, *62*, 377-392.

[16] Sapirstein, H. D. and Bushuk, W., *Cereal Chem.* 1985, *62*, 392-398.

[17] Sapirstein, H. D. and Bushuk, W., *Seed Sci. and Technol.* 1986, *14*, 489-517.

[18] Yakin, H. M., Kronberg, H., Zimmer, H. G. and Neuhoff, V., *Electrophoresis* 1982, *3*, 244-254.

[19] Stanley, K. K. and Pitt, T. J., *Anal. Biochem.* 1983, *133*, 476-481.

[20] Matthews, H. R., *Methods Mol. Biol.* 1984, *1*, 127-139.

[21] Lasters, I., Leyns, F. and Jackman, P. J. H., *Electrophoresis* 1985, *6*, 508-511.

[22] Schilling, J. J., Allan, B. B. and White, T. T., *Int. J. Bio-Medical Computing* 1983, *14*, 321-332.

[23] Plikaytis, B. D., Carlone, G. M. and Plikaytis, B. B., *J. Gen. Microbiol.* 1986, *132*, 2653-2660.

[24] Weiss, S. M., Kulikowski, C. A. and Galen, R. S., *J. Clin. Automation* 1983, *3*, 383-387.

[25] McLester, J. and Leung, F. Y., *Clin. Chem.* 1983, *29*, 2000-2001.

[26] Neeley, W. E., *Clin. Chem.* 1984, *30*, 794-797.

[27] Payne, P. I., Law, C. N. and Mudd, E. D., *Theor. Appl. Genet.* 1980, *58*, 113-120.

[28] Autran, J. C., Laignelet, B. and Morel, M. H., *Biochimie* 1987, *20*, 699-711.

[29] Lipschutz S., *Mathématiques pour mathématiciens*, Mc Graw-Hill, Paris 1986, Vol. *10*, pp. 245-253.

[30] Zillman, R. R. and Bushuk, W., *Can. J. Plant. Sci.* 1979, *59*, 287-298.

[31] Zeineh, R. A., Kyriakidis, G. and Bhatti, A. R., *Appl. Biochem. Biotechnol.* 1986, *13*, 111-117.

[32] Brigham, E. O., *The Fast Fourier Transform*, Englewood Cliffs, New Jersey, Prentice Hall 1974.

[33] Zimmer, H.-G. and Neuhoff, V., in: Radola, B. J. (Ed.), *Electrophoresis '79* de Gruyter, Berlin 1980, pp. 513-516.

[34] Tracy, R. P. and Young, D. S., *Clin. Chem.* 1984, *30*, 462-465.

[35] Haselgrove, J. C., Lyons, G., Rubenstein, N. and Kelly, A., *Anal. Biochem.* 1985, *150*, 449-456.