Classification of Potato Varieties Using Isoelectrophoretic Focusing Patterns, Neural Nets, and Statistical Methods

Kirsten Jensen,[†] Thomas K. Tygesen,[†] Can Keşmir,[†] Ib M. Skovgaard,[‡] and Ib Søndergaard*,[†]

Department of Biochemistry and Nutrition, Technical University of Denmark, Building 224, DK-2800 Lyngby, Denmark, and Department of Mathematics and Physics, The Royal Veterinary and Agricultural University, DK-1871 Frederiksberg C, Denmark

Automatic potato variety classification by a multivariate statistical classification method and a combined neural network model is described. Both classification methods were based on digitized isoelectrophoretic patterns of the soluble tuber proteins. The statistical classification algorithm was based on a model that allows for individual stretching as well as displacement along the pH axis. The neural network architecture consisted of two layers: a self-organizing feature map and a feed-forward classifier. Twelve potato varieties were classified. The mean value of the recognition rates were 84.5 and 87.5% obtained by the statistical classification method and the neural network model, respectively. The results confirm the theory stated in earlier classification studies, that the automatic classification systems are well-established, independent of the origin of the samples, and unaffected by pattern deformations and variations in the background level of the electrophoretic gels.

Keywords: *Potato; isoelectric focusing; image processing; classification; neural networks*

INTRODUCTION

For several years, identification of varieties has been an essential factor in the monitoring of plant breeding programs and for settlement of crops. Potatoes are not an exception: Grouping and classification of potato varieties has been performed by a number of different methods based on foliage, flower, and tuber characters; i.e., morphological characteristics of the tubers (Brown, 1973; Brown and Moss, 1976), genetic fingerprints based on restriction fragment length polymorphism analysis (Gebhardt et al., 1989; Görg et al., 1992), and random amplified polymorphic DNA analysis (Demeke et al., 1993). Electrophoretic patterns have been used in routine analysis by German testing stations from the beginning of the seventies (Stegemann, 1979). The identification, which is based upon gel electrophoresis of the soluble tuber-proteins and the esterase patterns, was systematized by Stegemann and Loeschcke (1976). The methods have later been applied for identification of sweet potatoes (*Ipomoea batatas* L.) (Stegemann et al., 1992).

The objective of our study was to examine whether a number of potato cultivars could be classified by an automatic procedure using isoelectric focusing (IEF) combined with image processing and neural networks and a multivariate statistical classification method. In earlier studies, we have shown that distortions of the electrophoretic patterns were the main source of error when they were used for automatic classification purposes (Søndergaard et al., 1994; Jensen et al., 1995). The distortions were observed as linear displacements and nonlinear stretching along the pH axis. To handle this, a classification algorithm based on a model that allows for individual stretching as well as displacement along the pH axis was developed (Skovgaard et al., 1995). In the present study, this algorithm is used along with a two-layered neural network architecture consisting of a self-organizing feature map and a feed-forward classifier, and the results are compared. Both of these classifier systems were developed by using the IEF patterns of several wheat varieties as a model system (Keşmir et al., 1995).

MATERIALS AND METHODS

Experimental Design. The experiment was designed as a complete block design in which a block was defined as 2 gels per run. The number of varieties and the number of lanes per block were maintained at 12. The number of replications per variety was 8. The varieties were randomized within each block by use of the commercially available statistical software package SAS (SAS Institute Inc., Cary, NC), utilizing the procedure PLAN.

Potato Varieties. The experiment comprised 12 potato varieties of *Solanum tuberosum*: Russet Burbank (1), Folva (2), N.88-BGJ-41 (3), Oleva (4), Up to Date (5), Ukama (6), Primula (7), Obelix (8), Berber (9), N.87-BEJ-29 (10), N.87-BDS-7 (11), and N-86-AZX-73 (12). The varieties were obtained from The Danish Institute of Plant and Soil Science, Tylstrup Experimental Station, Denmark, and the Danish Potato Breeding Foundation, Vandel, Denmark.

Sample Preparation. After freezing and thawing the potato tubers, sap samples were prepared by pressing small cubes of tuber using a hand-managed press (garlic press). An antioxidant was added (100 μL of 10% sodium sulfite + 7.5% sodium pyrosulfite/10 mL of extract) (Stegemann and Loeschcke, 1976). After centrifugation (2500*g* at 4 °C in 15 min), the supernatant was concentrated by vacuum to one-third of the volume (SpeedVac Concentrator, SVC 100H SAVANT). Finally, the extracts were stored at −18 °C.

Electrophoretic Method. Ultrathin agarose gels (0.5 mm) were prepared by modifying the procedure for casting PAGE gels (Søndergaard et al., 1994; Görg et al., 1978). The gels contain 1% agarose (IsoGel agarose, FMC BioProducts, Rockland, ME), 5% carrier ampholyte (Pharmalyte pH 5−8, Pharmacia Biotech, Uppsala, Sweden), and 30% ethylene glycol (Merck, Darmstadt, Germany) and were cast on Gel Bond film (FMC). IEF was performed on the PhastSystem (Pharmacia Biotech). The gels were prefocused for 75 Vh at 2000 V, 2.0 mA, 3.5 W at 17 °C. Six samples of 1.0 μL extract were applied at the middle of the gel, using a sample applicator (Pharmacia Biotech). In the outer lanes on the gel, p*I* markers (p*I* kit pH

[†] Technical University of Denmark.
[‡] The Royal Veterinary and Agricultural University.

S0021-8561(96)00273-7 CCC: $14.00

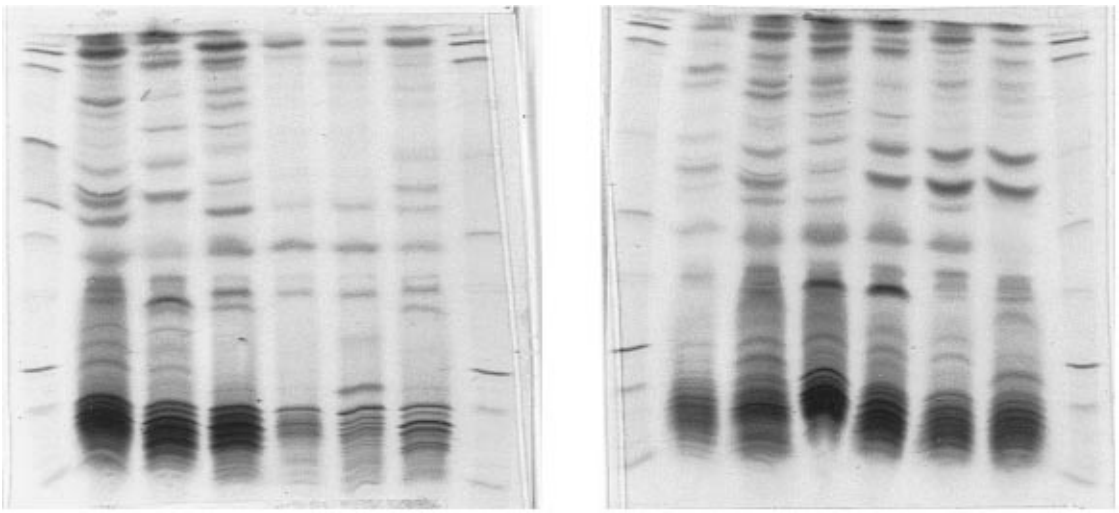

Figure 1. IEF (pH 5−8) of the potato sap proteins from the varieties: Russet Burbank, Folva, N.88-BGJ-41, Oleva, Up to Date, Ukama, Primula, Obelix, Berber, N.87-BEJ-29, N.87-BDS-7, N-86-AZX-73. Lanes 1 and 8 on each gel contain p*I* markers. Cathode at top.

3.5−9.3, Pharmacia Biotech) were applied. The samples were focused for 200 V (15 Vh) during sample application and next at 2000 V (510 Vh). The gels were fixed and pressed twice in 10% w/v trichloracetic acid and after drying were stained in a 1.0% w/v solution of Coomassie Brilliant Blue R 250 for 1 h; for details, see Søndergaard et al. (1994). An example of electrophoretic patterns of potato sap proteins is shown in Figure 1.

Software and Hardware for Digital Image Processing. Digital image processing was performed on a microcomputer (PC) equipped with a VGA display adapter. The images were digitized by a CCD video camera with a resolution of 512 × 512 pixels (Ikegami ICD-290, Ikegami Tsushinki Co., Tokyo, Japan) and a frame grabber board (Matrox PIP 512, Matrox, Quebec, Canada) being able to digitize in 512 × 512 × 8 bit resolution. The images were also presented in real time by using a black and white monitor (Ikegami PM 127, Ikegami Tsushinki Co., Tokyo, Japan) simultaneously connected to the frame grabber, thus enabling alignment and focusing without digitizing.

The image processing software was a commercially available version (CREAM, Kem-En-Tec Software Systems, Denmark) of a previously presented CREAM software package for evaluation of crossed immunoelectrophoretic patterns (Søndergaard et al., 1987, 1992). The software package has been further developed, and among other features, it has the possibility of scanning the lanes in gels from sodium dodecyl sulfate−polyacrylamide gel electrophoresis and IEF.

Scanning Conditions and Preprocessing of Data. For density determination, each lane in the electrophoretic patterns was scanned with a narrow scan in the middle of the lane. The mean values of the gray levels were estimated over each pixel in the full width of the scan, and in this way a spectrum consisting of 401 discrete values between 0 and 255 was obtained (Jensen et al., 1995). The pH interval was established between 8.5 and 5.1, by using p*I* markers. However, due to overloading in the acidic part of the gradient, the individual lanes were scanned in the pH interval 6.0−8.5, in which the most valid information was found. The raw spectra obtained were adjusted by subtracting the mean value of the spectra intensity. In this way, the effects of different background levels in the individual spectra were equalized.

Classification Method with a Random Shift. The shift/stretch algorithm (DA_{ss}) basically uses the method of least squares to shift and stretch the observed spectrum optimally each time it is attempted to match it to a particular variety. This requires that a typical spectrum is first estimated for each variety. Once this is done and a new spectrum has to be classified, it is compared to each variety in turn. The shifting and stretching are then optimized for each such comparison using an iterative numerical algorithm. The "goodness-of-fit" to each variety is then measured as the least squares distance on the scale of intensities plus a penalty term, which becomes large if the shift and stretch transformation is more drastic than usual.

The initial estimation of the typical variety spectrum uses the same basic shift and stretch optimization algorithm to match the spectra of the variety to each other. The typical spectrum is then obtained as the mean of the spectra aligned in this way. Details may be found in Skovgaard et al. (1995).

Neural Network Architecture. The neural network architecture used in this study consisted of two layers: a self-organizing feature map and a feed-forward classifier (NN_{sf}), respectively (Keşmir et al., 1995). The unsupervised learning algorithm used in the first layer enabled us to make a partitioning from the characteristics of the raw data coming directly from the digital image processing (see below for the potato data).

Since this algorithm could partition data of any character as long as there was enough information, the universatility of the NN_{sf} architecture was enlarged, that is, the same system could be used without many modifications within a wide range of data. It should be noted that the first layer was able to make only a rough classification of the data due to the experimental and biological variations in IEF patterns of the same variety.

The partitioned data from the first layer were further classified in the second layer. Here the final tuning for the correct classification was made with a supervised learning algorithm, standard backpropagation (Hertz et al., 1991). The data coming to the second layer are more "clean" than the original one because many of the variations mentioned before disappeared after the partioning process in the first layer. In other words, the input vectors coming from the same partition or cluster in the feature map would resemble each other more than they originally did.

A large parameter set was used for this combined system (Keşmir et al., 1995). To get a high performance from the NN_{sf} classifier, parameters should be adjusted according to the nature and quality of IEF patterns. In the present NN_{sf} system, only variations in the positions of protein bands are handled in the first layer. Thus, the variations coming from the background of the IEF plates effected results drastically: Clear backgrounds consistent in the whole of the data set always result in a better performance than dominating backgrounds. To prevent dominating background, the simple preprocessing of data explained earlier was applied to potato data. The interested reader will find the details of this classifier system and the analysis of parameters in Keşmir et al. (1995).

Validation of the Classification Results. Error count estimates were calculated counting the number of misclassified

Table 1. Classification Results for 12 Potato Varieties[a]

variety no.	$Q_{c_{DA}}$	$C(X)_{DA}$	$Q_{c_{NN}}$	$C(X)_{NN}$
1	100	0.93	75	0.78
2	50	0.69	75	0.86
3	75	0.78	88	0.81
4	100	1.00	100	0.84
5	75	0.73	75	0.86
6	88	0.76	100	0.94
7	100	1.00	88	0.86
8	88	0.81	100	1.00
9	88	0.81	100	0.94
10	88	0.81	88	0.81
11	88	0.93	88	0.86
12	75	0.86	75	0.86
mean	84.5	0.84	87.5	0.87

[a] Q_c was calculated on basis of error count estimates from the cross-validation of the discriminant analysis ($Q_{c_{DA}}$) and the neural network model ($Q_{c_{NN}}$). The correlation coefficients $C(X)_{DA}$ and $C(X)_{NN}$ estimate the correlation between prediction and observation for the variety in question, for the discriminant analysis and the neural network classifier, respectively.

observations in a test set, by using the classification criterion derived from a training set. To make cross-validation, the data set was divided into two parts. For training the system, eight training sets consisting of 12 varieties in seven replications were composed. Similiarly, eight test sets were created using the remaining replications of the same 12 varieties. In other words, each training set contained 84 data vectors, and each test set contained 12 data vectors. The recognition rates for each variety, Q_c, is given by the mean value of correctly classified observations. An alternative measure of classification results, which takes into account the relation between correctly predicted positives and negatives as well as false positives and false negatives, was used too (Matthews, 1975): The correlation coefficient, which is given by

$$C(X) = \frac{(P_X N_X) - (P^{f}_X N^{f}_X)}{\sqrt{[(N_X + N^{f}_X)(N_X + P^{f}_X)(P_X + N^{f}_X)(P_X + P^{f}_X)]}} \quad (1)$$

estimates the correlation between prediction and observation for the population of the replicates (variety in question), X. P_X and N_X are the correctly predicted positives and negatives, and P^{f}_X and N^{f}_X are similarly the incorrectly predicted positives and negatives, respectively. For a perfect prediction, $C(X) = 1$; for a random prediction, $C(X) = 0$ and for a totally negatively correlated prediction, $C(X) = -1$.

RESULTS AND DISCUSSION

The classification results for the 12 potato varieties are shown in Table 1. The mean value of the recognition rates were 84.5 and 87.5%, obtained by the statistical classification method and the neural network model, respectively. There is a good agreement between the recognition rates obtained by the two classification methods, only the results calculated by DA_{ss} for variety no. 2 can be considered as less succesful. The correlation coefficients estimated between the individual predictions and observations support that both classification methods are well suited for the current problem. When the original raw spectra were classified without adjusting for different background levels and displacements along the pH axis, the mean value of the cross-validation results obtained by a general linear discriminant analysis was less than 50%. This indicates that the preprocessing of data by the pattern matching algorithm and the combined neural network model causes a marked improvement of the recognition rates. This study includes a relatively small number of varieties as compared to the earlier classification studies of potatoes. However, our intention has been to evaluate the usefulness of the developed classification methods for different purposes. Thus, there are several possibilities for improving the amount of information in the IEF patterns of the potato sap proteins and in this way take a greater number of varieties into account without decreasing the recognition rate. The acidic proteins were not included in the classification due to overloading. However, as is obvious in Figure 1, a great proportion of the proteins are focused in the acidic area of the pH gradient, and essential information concerning the individual variety is probably inaccessible. These proteins constitute more than 20% of the total soluble potato protein, determined as glycoproteins with the trivial name patatins (Racusen and Foote, 1980; Racusen, 1983). The isoelectric points of the patatins were found between pH 4.2 and pH 5.3 (Seibles, 1979; Racusen and Foote, 1980). Therefore, it will be desirable to use a commercial manufactured carrier ampholyte, nonlinear in the pH interval 4−6, to stretch the acidic part of the gradient and to get a higher resolution of these bands. Additionally, stretching the pH gradient probably will make the concentration of the potato sap superfluous too, and to prevent this will simplify the procedure and decrease the time of analysis. Today such a possibility is only accessible by a manual mixing of carrier ampholytes stretching between different pH intervals (Pharmacia LKB Biotechnology, 1992). However, this will not be an optimal procedure in a routine analysis.

There is no substantial difference between the results obtained by the two classification methods, which is in agreement with the earlier classification problems based on IEF patterns of wheat. However, these type of problems do not have high complexity, i.e., the classification is based on a unique pattern with deformations for each entity. If the problem has a more complex character, the neural network classifier, which shows high performance in nonlinear problems, is expected to be more succesful. This is also what we have observed in a study to classify wheat varieties according to their baking quality using IEF patterns of gliadins and glutenins (Jensen et al., 1996). Here the aim of the classifier is to extract the minimum information needed from complex patterns of different character. Currently, we are working on a problem of similar complexity, namely, to decide the degree of purity in a lot of grain using IEF patterns of random samples taken.

CONCLUSION

The classification results obtained in this study were satisfying, in spite of the fact that the resolution of the individual protein bands are markedly inferior as compared to the model developed on basis of the gliadin fraction of the wheat endosperm. Therefore, the conclusion of the earlier studies is confirmed, namely, the developed automatic classification systems are well-established, independent of the origin of the samples, and unaffected by the pattern deformations and variations in the background level on the electrophoretic gels (Keşmir et al., 1995; Skovgaard et al., 1995).

ABBREVIATIONS USED

IEF, isoelectric focusing; DA_{ss}, shift/stretch algorithm; NN_{sf}, self-organizing feature map/feed-forward classifier; Q_c, recognition rate; $C(X)$, correlation coefficient.

ACKNOWLEDGMENT

We wish to thank Jørgen Christiansen, Department of Forage Crops and Potatoes, The Danish Institute of Plant and Soil Science, Foulum, Denmark, and Karl Tholstrup, Danish Potato Breeding Foundation, Vandel, Denmark, for kindly discussing the technical aspects of this study and for providing the potato cultivars.

LITERATURE CITED

Brown, E. Preliminary survey on the identification of potato varieties by tuber characters. *J. Natl. Inst. Agric. Bot.* **1973**, *13*, 67−86.

Brown, E.; Moss, J. The identification of potato varieties from tuber characters. *J. Natl. Inst. Agric. Bot.* **1976**, *14*, 49−69.

Demeke, T.; Kawchuk, L. M.; Lynch, D. R. Identification of potato cultivars and clonal variants by random amplified polymorphic DNA analysis. *Am. Potato J.* **1993**, *70*, 561−570.

Gebhardt, C.; Blomendahl, C.; Schachtschabel, U.; Debener, T.; Salamini, F.; Ritter, E. Identification of 2n breeding lines and 4n varieties of potato (*Solanum tuberosum*, spp. tuberosum) with RFLP-fingerprints. *Theor. Appl. Genet.* **1989**, *78*, 16−22.

Görg, A.; Postel, W.; Westermeier, R. Ultrathin-layer isoelectric focusing in polyacrylamide gels on cellophane. *Anal. Biochem.* **1978**, *89*, 60−70.

Görg, R.; Schachtschabel, U.; Ritter, E.; Salamini, F.; Gebhardt, C. Discrimination among 136 tetraploid potato varieties by fingerprints using highly polymorphic DNA markers. *Crop Sci.* **1992**, *32*, 815−819.

Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the theory of neural computation*; Addison-Wesley Publishing Company: Redwood City, 1991; pp 1−327.

Jensen, K.; Søndergaard, I.; Skovgaard, I. M.; Nielsen, H. B. From image processing to classification: I. Modelling disturbances of isoelectric focusing patterns. *Electrophoresis* **1995**, *16*, 921−926.

Jensen, K.; Keşmir, C.; Søndergaard, I. From image processing to classification: IV. Classification of electrophoretic patterns by neural networks and statistical methods enable quality assessment of cereals for breadmaking. *Electrophoresis* **1996**, *17*, 694−698.

Keşmir, C.; Søndergaard, I.; Jensen, K. From image processing to classification: II. Classification of electrophoretic patterns using self-organizing feature maps and feed-forward neural networks. *Electrophoresis* **1995**, *16*, 927−933.

Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.

Pharmacia LKB Biotechnology. Isoelectric focusing with PhastGel Dry IEF. *Sep. Tech. File* **1992**, *No. 101*.

Racusen, D. Occurrence of patatin during growth and storage of potato tubers. *Can .J. Bot.* **1983**, *61*, 370−373.

Racusen, D.; Foote, M. A major soluble glycoprotein of potato tubers. *J. Food Biochem.* **1980**, *4*, 43−52.

Seibles, T. S. Studies on potato proteins. *Am. Potato J.* **1979**, *56*, 415−425.

Skovgaard, I. M.; Jensen, K.; Søndergaard, I. From image processing to classification: III. Matching electrophoretic spectra by shifting and stretching. *Electrophoresis* **1995**, *16*, 1385−1389.

Stegemann, H. Characterization of proteins from potatoes, and the "Index of European Varieties". In *The Biology and Taxonomy of the Solanaceae*; Hawkes, J. G., Lester, R., Skelding, A. D., Eds.; Linnean Society Symposium Series 7; Academic Press: London, 1979.

Stegemann, H.; Loeschcke, V. Index Europäischer Kartoffelsorten. *Mitt. aus der Biol. Bundesanst. Land Forstwirtsch. Berlin-Dahlem* **1976**, *No. 168*.

Stegemann, H.; Shah, A.; Kroegerrecklenfort, E.; Hamza, M. Sweet potato (*Ipomoea batatas L.*): genotype identification by electrophoretic methods and properties of their proteins. *Plant Var. Seeds* **1992**, *5*, 83−91.

Søndergaard, I.; Poulsen, L. K., Hagerup, M.; Conradsen, K. Image processing and pattern recognition algorithms for evaluation of crossed immunoelectrophoretic patterns (Crossed Radioimmunoelectrophoresis Analysis Manager; CREAM). *Anal. Biochem.* **1987**, *165*, 384−391.

Søndergaard, I.; Hagerup, M.; Krath, B. N. Classification of crossed immunoelectrophoretic patterns using digital imageprocessing and neural network simulations. *Electrophoresis* **1992**, *13*, 411−415.

Søndergaard, I.; Jensen, K.; Krath, B. N. Classification of wheat varieties by isoelectric focusing patterns of gliadins and neural network. *Electrophoresis* **1994**, *15*, 584−588.

Received for review April 18, 1996. Revised manuscript received September 3, 1996. Accepted September 3, 1996.[⊗] Financial support was given by The Danish Agricultural and Veterinary Research Council under the PIFT program, The Danish Natural Science Research Council, and the DINA informatics research center. This research is also part of the FØTEK program sponsored by the Ministry of Research through LMC−Center for Advanced Food Studies, Denmark.

JF9602737

[⊗] Abstract published in *Advance ACS Abstracts*, November 1, 1996.