Interpretation of two-dimensional electrophoresis gels

Pierre Nugues, Josiane Steinmetz, Jean-Paul Haton, Marie-Madeleine Galteau, Gérard Siest

INRIA

UNITÉ DE RECHERCHE
INRIA-LORRAINE

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 921

Programme 1

INTERPRETATION OF TWO-DIMENSIONAL ELECTROPHORESIS GELS

Pierre NUGUES
Josiane STEINMETZ
Jean-Paul HATON
Marie-Madeleine GALTEAU
Gérard SIEST

Octobre 1988

RR-0921

INTERPRETATION OF TWO-DIMENSIONAL ELECTROPHORESIS GELS

PIERRE NUGUES[1], JOSIANE STEINMETZ[2], JEAN-PAUL HATON[3], MARIE-MADELEINE GALTEAU[2], and GÉRARD SIEST[2].

1. Cognitech, Paris, and INRIA-CRIN, B.P. 239, 54506 VANDŒUVRE-LÈS-NANCY, France.
2. Centre de médecine préventive, VANDŒUVRE-LÈS-NANCY, France.
3. INRIA-CRIN, VANDŒUVRE-LÈS-NANCY, France.

ABSTRACT: The study of human plasma protein polymorphism presents a major advantage in clinical chemistry to discover forms associated with risks or involved in well specified pathological states. If this polymorphism corresponds to structural differences reflected by physical property modifications it can be visualized on a map obtained by 2-dimensional electrophoretic technique.

This study deals with apolipoproteins A-I, A-II, A-IV, C2, C3, and E that can be found on 2-dimensional gels. Each of these proteins appears on images under the form of a *constellation* of several spots. An expertise was extracted on these apolipoproteins in order to itemize all their known variants and their geometrical representations. This knowledge was implemented in a system to recognize what are the constellations we are actually dealing with, on a gel. A management system handles specific techniques taken from *image processing* to extract parameters from 2-dimensional gel images and *artificial intelligence*. Knowledge-based expert systems which perform the matching, allowing a partial interpretation of the gel. The accuracy of quantity measurement and quality control are greatly improved by adding calibration proteins.

This same system will allow us to further implement *conceptual clustering* techniques to identify relevant relations in apolipoprotein metabolism.

Apolipoprotein polymorphism is used in this work as a model which could be extended to other protein constellations in the future.

L'INTERPRÉTATION D'ÉLECTROPHORÈSES BIDIMENSIONNELLES

RÉSUMÉ : La considération du polymorphisme des protéines du plasma permet à la biochimie clinique d'associer certaines formes à des risques ou des états pathologiques spécifiques. Lorsque ce polymorphisme correspond à des modifications de propriétés physiques, il devient visible sur les gels d'électrophorèse.

Notre étude traite des apolipoprotéines A-I, A-II, A-IV, C2, C3 et E présentes sur les gels bidimensionnels. Chacune de ces protéines apparaît sous la forme d'une « *constellation* » de plusieurs taches. Nous avons répertorié toutes les variantes connues de ces apolipoprotéines ainsi que leur représentation géométrique grâce à l'expertise de nos laboratoires. Nous avons implanté cette connaissance afin de reconnaître les constellations avec lesquelles nous sommes en présence sur un gel particulier.

Notre analyseur de gels s'organise à partir d'un système de gestion et de coopération des sources de connaissances : *la société de spécialistes*. Parmi ces spécialistes, on trouve des procédures spécifiques d'*analyse d'image* qui permettent le repérage des taches et l'extraction de leurs paramètres ainsi que d'*intelligence artificielle*, dont des systèmes-experts qui identifient les constellations et les apparient à un modèle. Ils fournissent une interprétation partielle du gel. Nous améliorons la qualité et la précision des mesures grâce à l'inclusion de protéines de calibration.

Nous introduisons dans notre système des techniques de *classification conceptuelle* qui permettront l'identification des relations existant entre les formes géométriques des polymorphismes et les différents états pathologiques ou particuliers que peut présenter un patient.

Notre travail a pour objectif de servir de modèle à l'étude des constellations d'autres protéines.

1. INTRODUCTION

The evidence of the role of lipid and apolipoproteins in the atherogenic process has been widely documented in human. It has been established that a large proportion of the variation on lipid and lipoprotein levels is associated with genetic differences among individuals. We have recently investigated the contribution of apoliprotein E polymorphism to the variability of plasma cholesterol and triglycerides [1]. The lipid constituents are also dependent on apolipoprotein A-IV genetic polymorphism [2].

One of the most elegant technique to study the genetic polymorphism of plasma apolipoproteins is two-dimensional electrophoresis. This method has been widely used in our laboratory for several years. Our goal, in this study, is to develop an expert system able to help clinical chemists in the interpretation of apolipoprotein polymorphism and relationship between lipid metabolism and cardio-vascular risks.

2. SAMPLES

This survey is conducted on a sample of unrelated, supposedly healthy subjects coming for routine family health screening at the Center for Preventive Medicine, Vandœuvre-lès-Nancy, France. About 100 families (which corresponds approximatively to 400 subjects) will be analyzed. During the health screening, much information is collected, concerning:

- blood analysis (measurement of cholesterolemia, triglyceridemia and other chemical enzymological constituents as well as hematological tests.)

- health status:
  - functional tests useful for the detection of cadiovascular disorders (blood pressure, electrocardiogram etc.)
  - questionnaire orientated towards life habits, drugs intake, familial and personal antecedents, etc.)

3. ELECTROPHORESIS PROCESSING TECHNIQUE

Two-dimensional electrophoresis method in presence of denaturing agents according to O'Farrell [3] and modified in our laboratory [4] was used.

Isoelectric focusing was run in pH 3 - 9.5 gradient and electrophoresis in 10 - 20% acrylamide linear gradient. Calibration proteins of known molecular weight (kits from Pharmacia, Uppsala, Sweden) were loaded in SDS slab gels in order to improve quality control and to semi-quantify results.

4. A MANAGEMENT SYSTEM: THE SPECIALIST SOCIETY

Due to the specific nature of the problem (important number of gels, their relative heterogeneity, and the existence of a variable background) a new system architecture was to be designed.

A Managing System that could meet the specifications for analyzing 2-D gels has to:

- integrate all the necessary techniques related to image processing;

• incorporate the background knowledge thanks to expert-system techniques.

Our system is based on an architecture which elicits Knowledge Sources in a non deterministic and versatile way, from the beginning where the knowledge is reduced to background expertise. It builds partial solutions by handling and reducing the uncertainty, generates forward and backward chaining inferences or backtracks in case of failure in the interpretation, until a final answer is delivered. Besides it can integrate heterogeneous types of gels, reducing the need for a severe standardization.

The knowledge sources can be conventional computer procedures as well as production rule systems able to manipulate data created dynamically during a session or belonging to the background (static) knowledge, in our case, the apolipoprotein domain.

The management system is based on the concept of *specialist society* [5]. It is implemented with the expert-system generator IROISE@CEWB [6] which provides a graphic environment and allows for a modular design. The basic principle is as follows:

A managing module, the *administration*, activates knowledge sources grouped in *associations*. The administration can receive and send messages. The functioning of each of the associations is determined by an *agenda*. This agenda is defined dynamically by the association itself or with the help of the messages emitted by the administration. For each of the constellations of apolipoprotein studied on the gel, four different associations were defined:

• standardizing, image zoning and peak detection,

• structural matching which recognizes overall patterns and gives the possible candidates,

• syntactic qualifier and interpretation,

• numerical fitting.

Other associations and specialists will be included later on, particularly concerning data clustering and parallel or beam searching.

Here is an example of the functioning of the system:

In plasma gels, the apoliporotein A-I is composed of several spots (figure 1). The expertise reports that two small spots are located between the major spot and the pro-apoliporotein A-I spot. (figure 2). The expertise adds that these two small spots can merge with the major one leaving only one maximum. .

Suppose a gel was found with the following configuration: the two small spots merged partly with the major one. Only one maxima is detected in area which could correspond to the major A-I. (figure 3). The associations are activated in the following way for interpretating this gel:

1. the zoning specialist extracts the apolipoprotein A-I region. All maxima above a definite threshold are registered.

2. The structural association registers a candidate for pro-apolipoprotein A-I and aligned spots which are more acidic and with an approximatively equal molecular weight.

3. The syntactic qualifier checks the consistency of the configuration. It detects a major spot, showing only one maximum. It fails matching the two minor spots. It sends this possible candidate for the apolipoprotein A-I chain to the administrator.

4. The numerical fitting association is activated yielding an important approximation error.

5. A new agenda is prepared for the syntactic association, asking it to predict the location of the missing minor spots.

6. The numerical fitting association is activated again yielding the good parametrization of these spots: one major and two minor peptides.

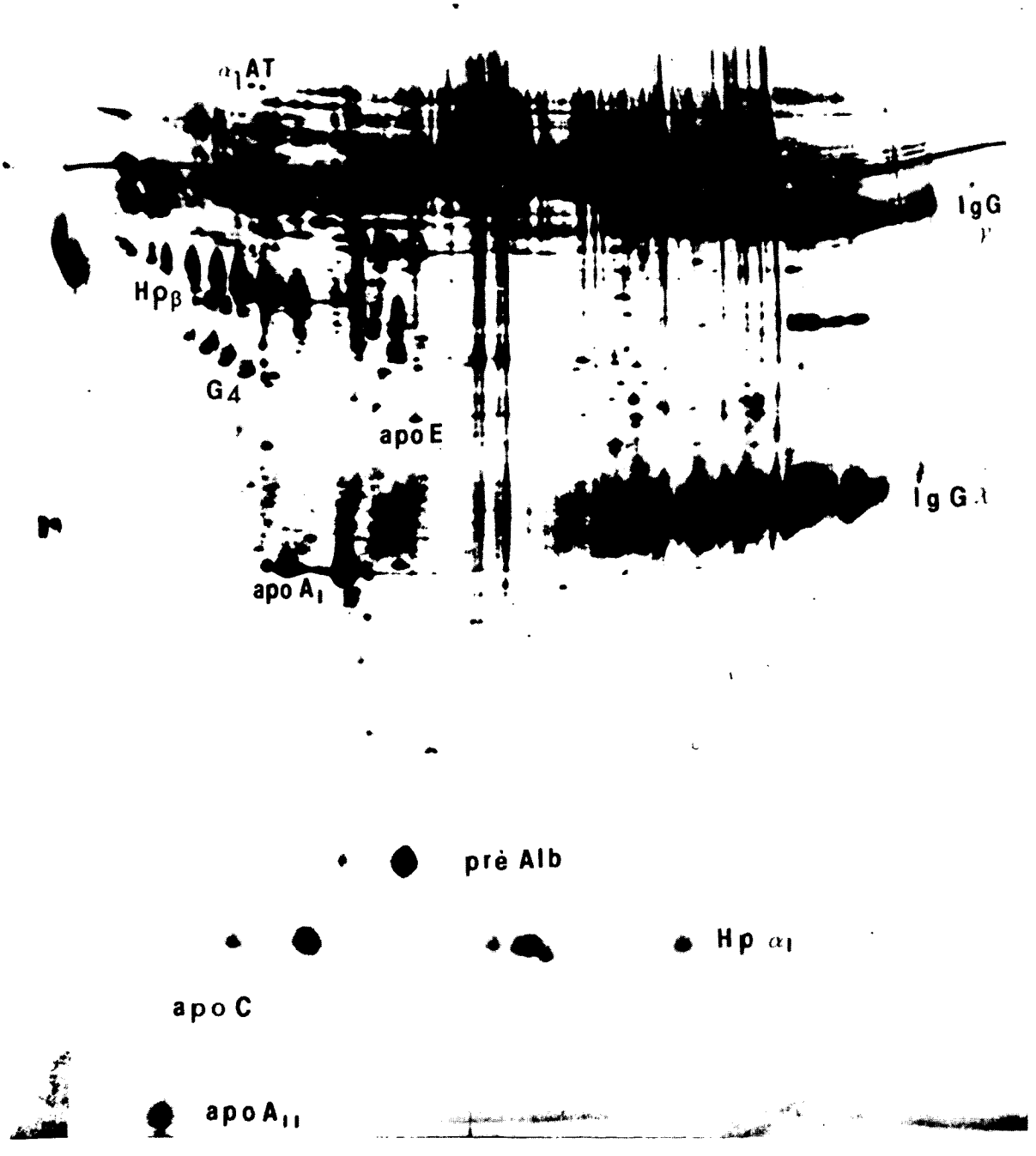

Figure 1: 2-D electrophoresis pattern of human plasma proteins.

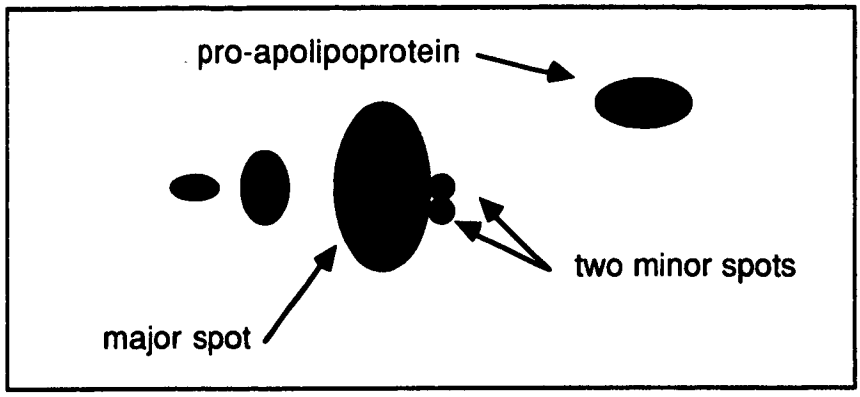

Figure 2: Human serum apolipoprotein A-I chain.

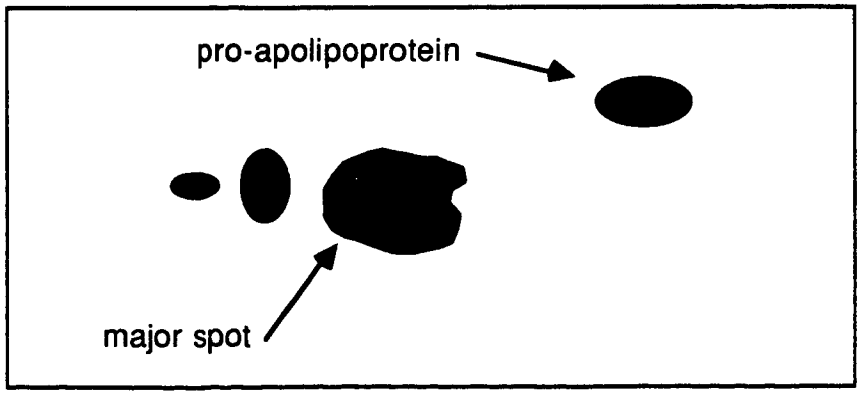

Figure 3: Human serum defective apolipoprotein A-I chain.

5. ANALYZING A GEL

Gel analysis involves the following operations:

5.1 Acquisition. The first stage of the computer processing is the digitization of the gel. To carry out this task, a scanning densitometer is used, with a resolution of 10 points per mm. The digitizing is performed on 256 gray levels. Since our gel dimensions are approximatively of 20 *cm* $\times$ 20 *cm*, it leads to a size of 4 megabytes. This size is very large but it can be reduced by means of image compression [7]. It allows a theoretical detection of spots the full width of which at half maximum is above 0.4 mm [7]. Such a resolution is not necessary for most of the apoliproteins and so we operate a scaling of order 2 or 4 on the original image before the image is loaded in memory.

5.2 Preprocessing. Digital treatments can be performed on the image to remove worthless information, coming from high-frequency noise, image background, and protein streaks. A method based on mathematical morphology to remove high frequency noise was implemented [8]. It ensures a further good detection of the local maxima by eliminating artifactual spikes. It may introduce a minor bias by severing a part of information beyond the cut-off frequency. No method for removing background and streaks was implemented, because on all the experiments which were carried out, we could not restore a 'clean' image without distorting it severely or losing a significant amount of information.

5.3 Local Maxima Detection. The matching procedure starts with the detection of spots. We first assume that each local maximum of the smoothed image, beyond a determined level corresponds to a protein. A mathematical morphology method was implemented [8]. This method yields good results but is not error-free. Some configurations can appear that mask the real number of spots. For instance close spots could merge into one peak showing only one local maximum or artifactual spikes could generate a local maxima in spite of the smoothing. Our knowledge-based management system allows to get rid of most of these deficiencies.

5.4 Matching. The procedure used is not properly a matching between gels as it is described in the literature [8-11] but rather a matching of an expertise about proteins, with a gel.

Our study will involve a large number of images (approximatively 400). This raises a formidable difficulty when a blind matching is performed, but it can be tractable if background knowledge is taken into account. Apolipoproteins can be sorted in major groups. Inside these 'constellations', different patterns can be found. All types of encountered clusters for the studied proteins have been itemized by the description of patterns, and their contingent variants. The exact identification of each spot involves its location in comparison with other spots of the same group and with remarkable spots. This kind of description eliminates the influence of most of the variations that may occur during the chemical process.

The structural identification is performed by matching partially well identified spots to patterns of pre-recorded protein groups. They determine 'confidence islands' on which the further processing relies. A syntactic module parses the configuration determined, checking for the relevance of the relationships in the group and the dependencies with the environment Eventually, apolipoprotein

spots on the image are fitted to the proper models yielding the apolipoprotein configurations we are dealing with. Thus a partial interpretation of the state of the patient can be given if one of the configurations can be related to a specific disease.

5.5 Parameter Extraction. The two-dimensional Gaussian modeling of spots was chosen. This model can easily handle overlapping spots. Every protein is determined by five parameters: its amplitude A, its center coordinates (x_0, y_0) and its standard deviations (σ_x, σ_y). Then a least square approximation is performed on the connectivity area determined by a single spot or cluster of overlapping proteins [12].

A normalization procedure was introduced to determine the exact value of A. The quantitation (i.e. determination of A) of proteins on a gel image shows up a certain difficulty, as the raw integral of levels over the surface of each spot is of no significance for most silver-stained gels. The normalization is independent and intrinsic to each gel. We assume that studied proteins are submitted to the same influences as calibration proteins loaded on the gel, then the quantity of each studied protein is approximated thanks to the included known amount of markers.

6. RELATING A GEL WITH OTHER GELS AND INTERPRETATING EXPERIMENTS

This module of the system will make classication or partition a set of patients according to the quantity of apolipoproteins and to the results of other collected information, using numerical and conceptual clustering.

Principal Component Analysis and Factor Analysis will identify major numerical axes. These analyses will consider all numerical data available from patients. Vincens [13] showed up the aid that could be brought from Multivariate Analysis to determine the functions of cells, to classify diseases' specific proteins.

However such conventional data analyses are unable to handle symbolic concepts or to integrate *a priori* knowledge. Besides, there is a lack in descriptive power which means that an extra interpretation will have to be given after the computer output. Conceptual clustering is very expensive in term of computation time but description for a class such as: [apo C value = elevated] and [cardiac heredity = strong] and [smoker] can be expected.

In our analyzer, this type of conceptual processing will be integrated. A method similar to these described by Michalski and his school [14] is now being implemented. Using these techniques, Appel [15] pioneered in the domain of automated analysis applied to 2-dimensional electrophoresis.

The method can be summarized as follows:

1. The number of classes is entered first, and one seed (i.e. one element from the set) is chosen for each class;
2. A description is given for each seed by its differences with others. (i.e only attributes which are different from others are kept)
3. The conjunction of all these disjunctions is sequentially performed, inverting the formula and simplifying it;

4. At each step the total number of conjunctions is reduced by taking the best ones according to evaluation formulas. The remaining sets are then generalized;
5. Once the total description obtained, each class can be determined by several possible conjunctions. The partition is obtained by choosing a formula amongst each class and by disjoining the described set; The best set (according to evaluation criteria) for the partition is kept.
6. This algorithm is iterated, if needed, by selecting, for each class, a new seed amongst its elements.
At the end of the process a description of each class is yielded by the discriminant values of its attributes.

7. CONCLUSION

Each electrophoresis image bears a huge quantity of information which is not presently totally exploited. Electrophoresis results can be kept in data bases under the form of parameters or better under the form of images. They represent a major source of information for the researcher, even if no interpretation is given on them or if the proteins are not properly identified.
Facing an unknown protein, the researcher will have the possibility to retrieve it on gel images by describing its relationships with respect to its environment, even if the gel images are only weakly standardized. Then, a further processing of the data base will relate the protein to specific properties by means of conceptual clustering. These properties could bear on other proteins of same gels or other of biological nature that could be recorded in the database.
This may be the future design and use of electrophoresis data bases. This model, first tested on apolipoprotein constellations, could be applied for other protein constellations.

REFERENCES

1. Boerwinkle E., Visvikis S., Welsh D., Steinmetz J., Hanash S.M., and Sing C.F. The use of measured genotype information in the analysis of quantitative phenotype in man. II: The role of the apoliprotein E polymorphism in determining levels, variability and co-variability of cholesterol, β-lipoprotein and triglycerides in a sample of unrelated individuals. Amer. J. Med. Genet., 27, 567-582. (1987).

2. Visvikis S., Steinmetz J., Boerwinkle E., Galteau M.M., and Siest G. Frequency of apolipoprotein E and A-IV polymorphisms and effects on lipid levels. A study by two-dimensional electrophoresis. An. Clin. Biochem., 24, supp. 2, (1987).

3. O'Farrell P.H. High resolution two-dimentional gel electrophoresis of proteins. J. of Biol. Chem., 250, 4007-4021, (1975).

4. Visvikis S., Steinmetz J., Cuvelier I., Galteau M.M., and Siest G. Study of apoliporotein A-I

and E polymorphism using 2-D electrophoresis. In *Recent progress in two-dimensional electrophoresis*, Eds. Galteau M.M. and Siest G. , Presses Universitaires de Nancy, 1986.

5. Gong Y. and Haton J.-P. A specialist society for continuous speech understanding. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, New York, 1988.

6. CEWB. *Reference Manual.* Cognitech, 1988. To be published.

7. Vincens P., Paris N., Pujol J.L., Gaboriaud C., Rabilloud T., Pennetier J.L., Matherat P., and Tarroux P. HERMeS: A second generation approach to the automatic analysis of two-dimentional electrophoresis gels. Part I: Data acquisition. *Electrophoresis*, 7, 347-356, (1986)

8. Skolnick M.M. Application of morphological transformations to the analysis of two-dimensional electrophoretic gels of biological materials. *Computer Vision, Graphics, and Image Processing*, 35:306--332, (1986).

9. Miller M.J., Vo P.K., Nielsen C., Geiduschek E.P., and Xuong N.H. Computer analysis of two-dimensional gels: semi-automatic matching.Clin.Chem., 28 867-875, (1982).

10. Garrels J.I., Farrar J.T., and Burwell IV C.B. The QUEST system for computer-analyzed two-dimensional electrophoresis of proteins. In *Two-Dimensional Gel Electrophoresis of Proteins* Eds. J.E. Celis and R. Bravo, pp. 37--91, Academic Press, 1984.

11. Vincens P. and Tarroux P. HERMeS: A second generation approach to the automatic analysis of two-dimentional electrophoresis gels. Part III: Spot list matching. *Electrophoresis*, 8 100-107, (1987).

12. Taylor J., Anderson N.L., and Coultern B.P. Estimation of 2-dimentional electrophoretic spot intensities and positions by modelling. In *Electrophoresis'79* Eds. Radola B.J., pp 329--339, De Gruyter, 1979.

13. Tarroux P., Vincens P., and Rabilloud T. HERMeS: A second generation approach to the automatic analysis of two-dimentional electrophoresis gels. Part V: Data analysis. *Electrophoresis*, 8 187-199, (1987).

14. Michalski R.S., Stepp R.E. Learning from observation: conceptual clustering. In *Machine Learning: An Artificial Intelligence Approach*.. Eds. Michalski, Carbonnell and Mitchell. pp 331-363 Springer Verlag, 1984.

15. Appel. R.D. Melanie. Un système d'analyse et d'interprétation automatique d'image de gels

d'électrophorèse bidimentionnelle. Systèmes-experts et apprentissage automatique. PhD thesis, Université de Genève.1987. (French)

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique