

## Cluster Analysis Method in Electrophoretic Patterns Classification

F. Spirovski<sup>a,\*</sup>, K. Stojanoski<sup>a</sup> and A. Mitrevski<sup>b</sup>

<sup>a</sup>*Institute of Chemistry, Faculty of Natural Science, Sts. Cyril and Methodius University, 1000 Skopje, Macedonia*

<sup>b</sup>*Clinic of Neurology, Faculty of Medicine, Sts. Cyril and Methodius University, 1000 Skopje, Macedonia*

(Received 8 September 2004, Accepted 16 October 2004)

The conventional electrophoresis methods are well known techniques for protein detection and analysis of cerebrospinal fluid (CSF). Disc electrophoresis (DEP) was carried out for detection of oligoclonal IgG bands in cerebrospinal fluid (CSF) on polyacrilamide gel. However, the advance of automation has made rapid collection of large amounts of data feasible and the development of microcomputers has made sophisticated processing even of old electrophoregrams possible. Automated analysis, data storage and sophisticate data acquisition were carried out with Gel Pro Analyzer 3.1, which is specifically structured to analyze gels and elctrophoregrams: complex band pattern matching (gel variation, dendogram analysis etc.); lane relation studies (sophisticated lane database); general gel analysis (accurate molecular size, quantitative determination of protein mixture etc.). Clustering techniques have been applied for detection of intrathecal immune response. Different hierarchic cluster analysis methods such as single linkage, complete linkage, unweighted pair-group average (UPMGA) were used. In addition, other cluster characteristics such, distance matrix and Euclidean distance matrix were calculated. Pairing of electrophoresis methods and cluster image analysis, could lead to additional diagnostic information of inflammatory conditions of the central nervous system (CNS) or dysfunction of blood-CSF barrier.

**Keywords:** Disc electrophoresis, Cerebrospinal fluid, Immunoglobulin G, Cluster analysis

---

### INTRODUCTION

Polyacrylamide gel electrophoresis (PAGE) is the most widely used method for analyzing protein mixture qualitatively. Protein detection and analysis of cerebrospinal fluid (CSF) was carried out. It is a main target for CSF analysis to detect an intrathecal immune response and to find immunological features which could contribute to differential diagnosis of neurological disease [1,2]. Important difference between plasma and CSF is the protein content, which is on the average about 250 times lower in CSF. More specifically,

the CSF levels of IgG, IgA and IgM are about 500, 100 and 5000 lower than the serum levels, respectively. These differences are due to the presence of a blood-brain and a blood-CSF barrier to proteins. Albumin is the major CSF protein and represents about 55% of the total protein content, a proportion similar to that observed in serum. In normal conditions, CSF immunoglobulin G (IgG) originates from the blood and there is no antibody production within CSF. Immunoglobulin intrathecal production is an important diagnostic parameter in inflammatory conditions of the central nervous system (CNS). On the other hand, any increase of the CSF albumin level, results from an increased transudation from the blood lead to diagnosis of dysfunction of blood-CSF barrier. Quantitative analyses include determination of IgG

---

\* Corresponding author. E-mail: filips@iunona.pmf.ukim.edu.mk

and albumin in CSF and serum, and afterwards out of them IgG index and IgG local synthesis rate are mathematically calculated [3]. It is now widely accepted that the detection of oligoclonal IgG bands specific to the CSF is more sensitive test to demonstrate an intrathecal IgG response than any quantification of the humoral response by any mathematical formulation [4]. The aim of this study was to achieve fast, easy and correct detection of an intrathecal immune response and dysfunction of blood-CSF barrier by using cluster method in electrophoresis gel image analysis. Diagnosis based on medical image data is common in medical decision making and clinical routine [5-8]. We discuss a strategy to derive a classifier with good performance on clinical image data.

## EXPERIMENTAL

### Patients

Pairs of CSF and serum were taken at the same time during the course of lumbar puncture and venepuncture from 52 patients investigated and treated at Clinic of Neurology, Faculty of Medicine in Skopje, Macedonia. CSF and serum were sampled under sterile conditions. If they had not been analyzed within two days, they were stored at -20 °C. Clinical experiments were performed according to the Regulation of Macedonian Ethic Committee and Ministry of health of the Republic of Macedonia. Each patient was aware of objective of the paper: each patient was given a code name (no personal name address aiming to ensure privacy).

### Disc Electrophoresis (DEP)

Disc electrophoresis was carried out on 7% polyacrylamide gel, using the electrophoresis system Canalco (Rockville, USA). Cerebrospinal fluid (CSF) was used without preconcentration. Each sample gel contained about 200 µg of proteins and so the volume of CSF applied was dependent on the protein concentration of the CSF sample. 5 mmol l<sup>-1</sup> Tris-0.038 mol l<sup>-1</sup> glycine buffer (pH 8.3) was used. Electrophoresis was running at 5 mA per sample. After its completion, the gels were stained with Coomassie-blue and the stains measured by microdensitometer Canalco Model 8I. Also stained gels were interpreted by means of scanner Sharp JX-330. Gel Pro Analyzer 3.1 specialized software package was used for fast accurate image and gel analysis. All

chemicals (gels, buffer and stain) were purchased from Merck (Germany).

### Cluster Analysis

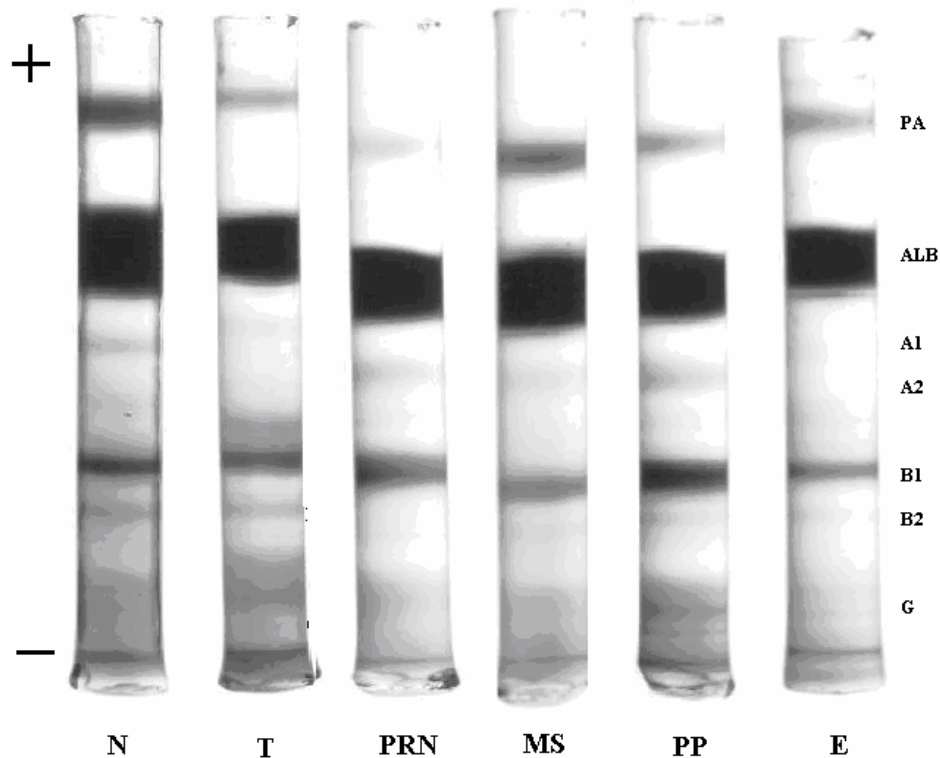
The promise of computer application to electrophoresis gel image analysis is to provide not only accurate and reliable quantification, but also the ability to analyze statistically large quantities of samples. One of the important applications in population study is cluster analysis based on similarity measurements. The clustering problem can be described as finding “natural groupings” in a set of data. This question actually involves two separate issues: how to measure the similarity between samples, and how to partition sets of samples into clusters among a large number of data entry points, such as bands and spots on electrophoresis gels. The term cluster analysis (first used by Tryon, 1939) actually encompasses a number of different classification algorithms. In the context of electrophoresis gel image analysis, the samples can be clustered by either statistical clustering algorithms or neural network approaches [9-11]. A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures.

The joining or tree clustering method uses the dissimilarities or distances between objects when forming the clusters. These distances can be based on single dimension or multiple dimensions. The most straightforward way of computing distances between objects in multidimensional space is to compute Euclidian distances. They are computed as:

$$d_{ij} = \sqrt{\left\{ \sum_{k=1}^p (x_{ik} - x_{jk})^2 \right\}}$$

where  $x_{ik}$  is the value of variable  $X_k$  for individual  $i$  and  $x_{jk}$  is the value of the same variable for individual  $j$ .

Also there are possibilities to compute various types of distance measures as: squared euclidian distances, city-block distance, chebychev distance and power distances. In our work we compute Euclidian distances. When each object represents its own cluster, the distances between those objects are defined by the chosen distance measure. However, once several objects have been linked together, we have to determine the



**Fig. 1.** Disc electrophoresis of CSF from: N-normal subject, Thrombosis cerebri patient, polyradiculoneuritis patient with PRN, multiple sclerosis patient with MS, paraproteinemia-patient with PP and Encephalitis patient with E.

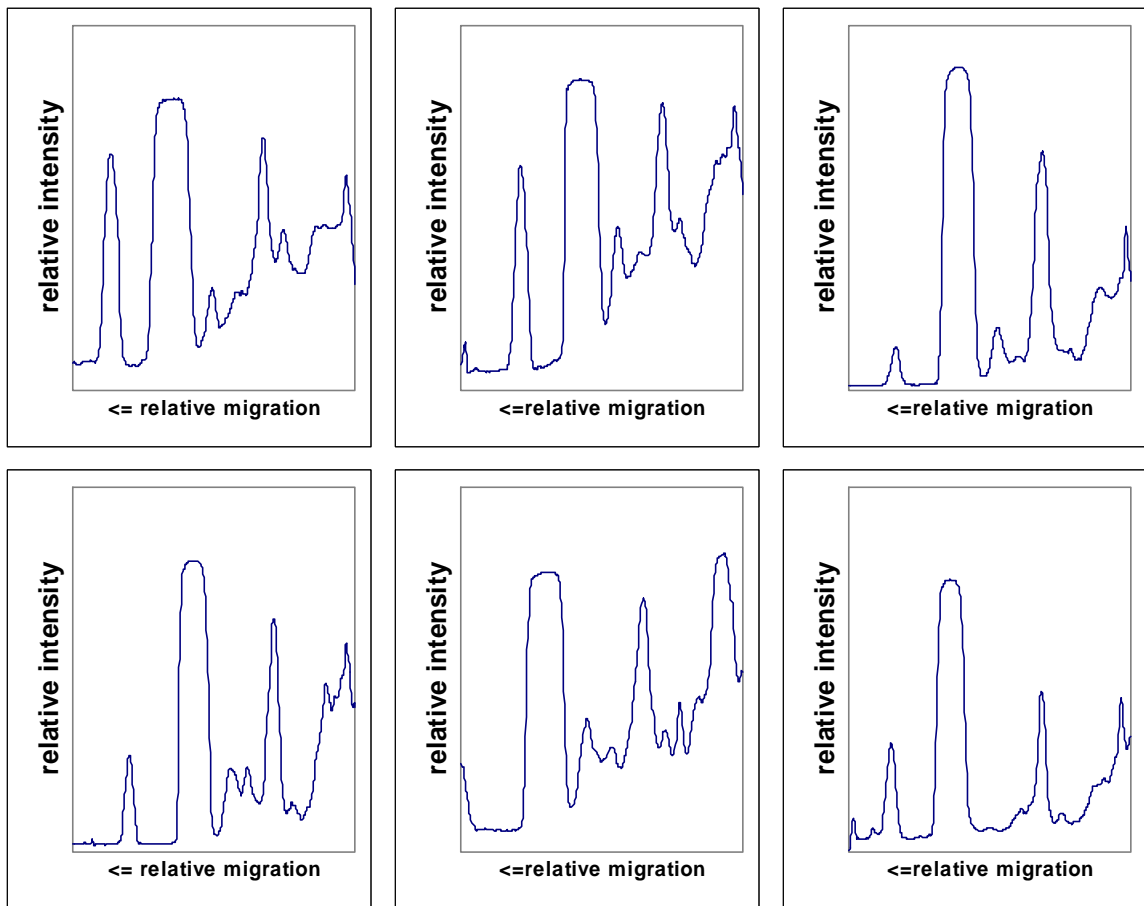
distances between those new clusters. In other words we need a linkage or amalgamation rule to determine when two clusters are sufficiently similar to be linked together. There are numerous linkage rules that have been proposed: single linkage, complete linkage, unweighted pair-group average (UPGMA), weighted pair-group average and etc. [12,13].

### Results and Discussion

Six examples of CSF polyacrilamide patterns from patients with different disease are presented on Fig. 1. Using standard procedures IgG factors were calculated and used as criteria for the electrophoregram classification. DEP was used, and as positive results for multiple sclerosis (MS) cases were taken those in which oligoclonal IgG patterns, the relative values were  $\geq 15\%$  ( $\gamma$ -globuline type of electrophoregram), with domination of slow area in G-zone. For other cases standard procedures were used which have been described elsewhere

[14,15]. Collection of large amounts of data by automated analysis was carried out with Gel Pro Analyzer 3.1, which is specifically structured to analyze gels and electrophoregrams. Fig. 2 presents electrophoregrams collected by analyzing gels from same six patients respectively.

Cluster analysis on the different types of electrophoregram curves data set, extracted after the digitalization process, was carried out. Results from cluster analysis using unweighted pair-group average method (UPGMA) are presented in Fig. 3. As we can see, cluster analysis revealed two main clusters, first consist of all patients with multiple sclerosis and one with encephalitis (case E3) and second cluster consist of patients with polyradiculoneuritis (PRN), also known as Guillain-Barre syndrom, and paraproteinemia (PP). This method of cluster analysis shows that we can create classifier with good performances [16,17]. Knowing that multiple sclerosis is inflammatory disease with strong intrathecal immune response



**Fig. 2.** Electrophoregrams of CSF from: N-normal subject, Thrombosis cerebri patient, polyradiculoneuritis patient with PRN, multiple sclerosis patient with MS, paraproteinemia-patient with PP and Encephalitis patient with E.

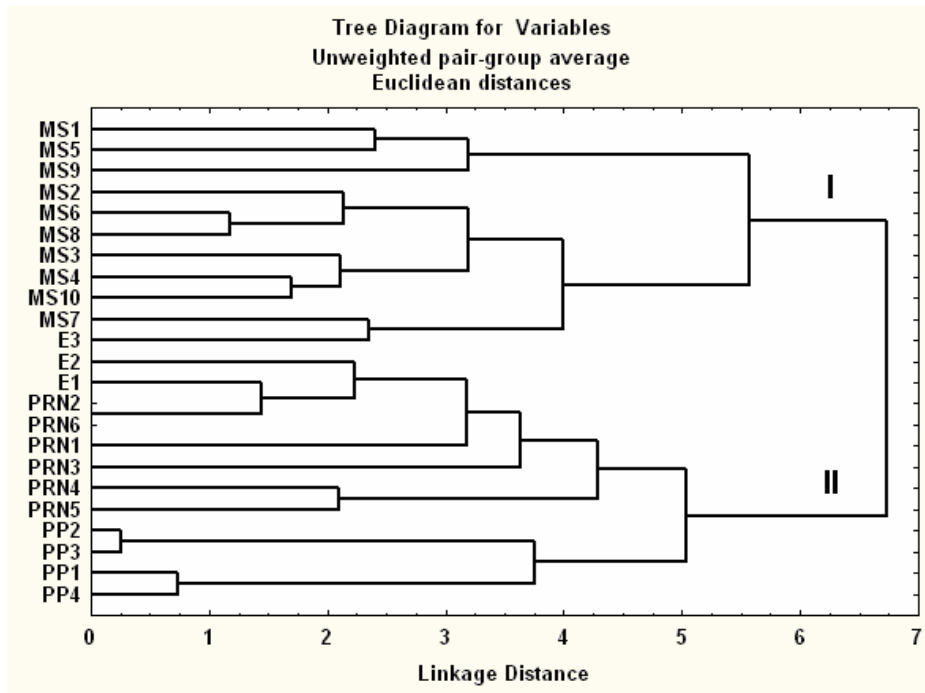
we expect new data from patients with intrathecal immune response to be include in same cluster. Also there are three distinct groups in which entries revealed a greater similarity.

Another cluster analysis, using method of complete linkage is presented in Fig. 4. This is a different type of showing results, by using vertical icicle plot, instead of horizontal hierarchical tree plot. Cluster analysis revealed two main clusters with two subgroups, looking from left to right, first consist of all patients with PP, then other subgroup consist of patients with PRN and E. Second cluster, which is in right part of Fig. 4, assembly electrophoregrams from patients with T and MS and one with E. Knowing that patients with paraproteinemia have dysfunction of blood-CSF barrier, position where should be classified new, unknown patterns,

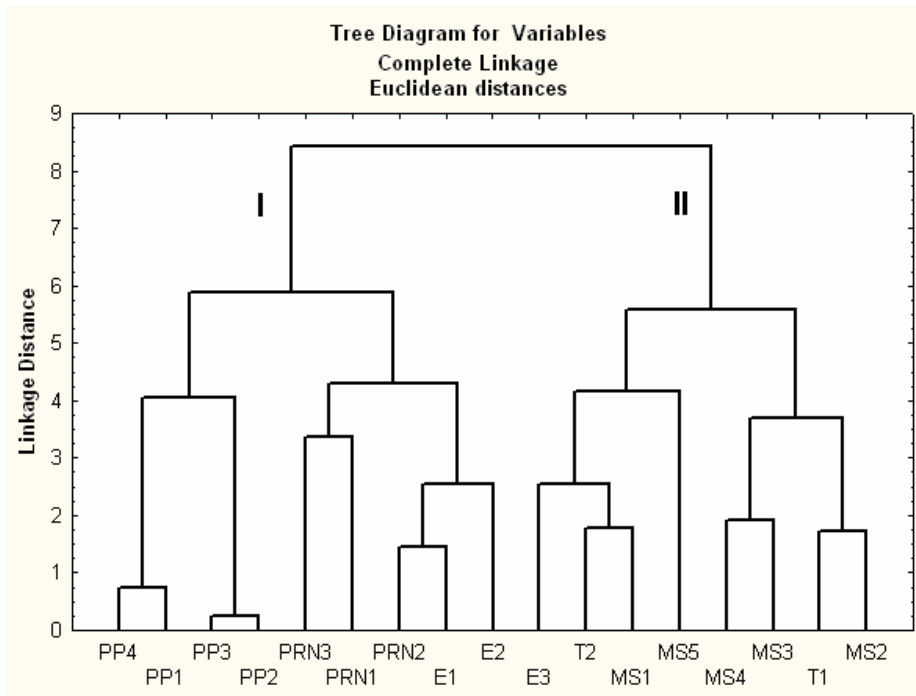
could lead to conclusion of presence of intrathecal immune response or dysfunction of blood-CSF barrier.

It has already been mentioned that there are many algorithms for cluster analysis. However, there is no generally accepted “best” method. Unfortunately, different algorithms do not necessarily produce the same results on a given set of data. There is usually rather a large subjective component in the assessment of the results from any particular method. A fair test of any algorithm is to take a set of data with a known group structure and see whether the algorithm is able to reproduce this structure. It seems to be the case that this test only works in cases where the groups are very distinct. When there is a considerable overlap between the initial groups, a cluster analysis may produce a solution that is quite different

Cluster Analysis Method in Electrophoretic Patterns Classification



**Fig. 3.** Cluster analysis-joining tree clustering model-UPGMA  
MS-multiple sclerosis, E-encephalitis, PRN-polyradiculoneuritis, PP-paraproteinemia



**Fig. 4.** Cluster analysis-joining tree clustering model-complete linkage  
vertical icicle plot.

MS-multiple sclerosis, E-encephalitis, PRN-polyradiculoneuritis, PP-paraproteinemia, T-thrombosis cerebri

from the true situation.

## CONCLUSIONS

Cerebrospinal fluid analysis, coupled with other methods, remains cornerstone for the diagnosis of various neurological disorders, including multiple sclerosis and other infectious and inflammatory diseases of central nervous system (CNS). Pairing of electrophoretic methods and cluster image analysis could lead to additional diagnostic information. As we shown in our study, we can determine presence of intrathecal immune response or dysfunction of blood-CSF barrier and also detect different types of patterns. Unfortunately certain diagnosis of one disease could not be achieved by only analyzing patterns from 1D electrophoresis. Our next step is to investigate method of analyzing patterns from 2D electrophoresis, which we hope could give diagnosis of different inflammatory and infectious diseases of CSF.

## REFERENCES

- [1] H. Reiber, M. Otto, C. Trendelenburg, A. Wormek, *Clin. Chem. Lab. Med.* 39 (2001) 324.
- [2] C. Sindic, M. Antwerpen, S. Goffette, *Clin. Chem. Lab. Med.* 39 (2001) 333.
- [3] W.W. Tourtellotte, A.R. Potvin, J.O. Fleming, K.N. Murthy, J. Levy, K. Sydulko, J.H. Potvin, *Neurology*. 30 (1980) 240.
- [4] A. Mitrevski, K. Stojanoski, P. Korneti, *Acta. Pharm.* 3 (2001) 163.
- [5] J.M. Jerez-Aragones, *Artf. Intell. Med.* 27 (2003) 45.
- [6] M.F.G. Boriollo, *Rev. Inst. Med. Trop. S. Paulo.* 45 (2003) 249.
- [7] H. Wang, D. Chen, Y. Chen, *Artf. Intell. Med.* 70 (2004) 23.
- [8] B. A. Mobley, *Artf. Intell. Med.* 18 (2000) 187.
- [9] Neuro solutions-software, Curt Lefebvre and Co., NeuroDimensions Inc., 1994-2003.
- [10] P.J.G. Lisboa, *Neural Netw.* 15 (2002) 11.
- [11] I. Aizenberg, N. Aizenberg, J. Hiltner, C. Moraga, E.M. zu Bexten, *Image and Vision Comp.* 19 (2001) 177.
- [12] N. Saitou, M. Nei, *Mol. Biol. Evol.* 4 (1987) 406.
- [13] R.R. Sokal, P.H.A. Sneath, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, Freeman, London, 1973.
- [14] L. Thomas, *Laboratory diagnosis of neurological diseases. Clinical Laboratory Diagnostics: Use and Assessment of Clinical Laboratory Results*, TH-Books Verlagsgesellschaft, Frankfurt, 1998.
- [15] H. Reiber, *Lab. Med.* 19 (1995) 444.
- [16] F.W. Gay, T.J. Drye, G.W. Dick, M.M. Esiri, *Brain* 120 (1997) 1461.
- [17] B.F.J. Manly, *Multivariate Statistical Methods-A Primer*, Chapman and Hall, New York, 1986.