Acta Pharm. 55 (2005) 215–222

Short communication

Comparison of statistical cluster methods in electrophoretic protein pattern analysis

FILIP SPIROVSKI[1]*
KIRO STOJANOSKI[1]
ANGEL MITREVSKI[2]

[1] Institute of Chemistry, Faculty of Natural Sciences, Sts. »Cyril and Methodius« University, POB 162 1000 Skopje, R. Macedonia

[2] Clinic of Neurology, Faculty of Medicine, Sts. »Cyril and Methodius« University, 1000 Skopje, R. Macedonia

Standard electrophoresis methods were used in the qualitative and quantitative protein analyses of cerebrospinal fluid (CSF). Disc electrophoresis was carried out for detection of oligoclonal IgG bands in cerebrospinal fluid on polyacrylamide gel. Pairs of CSF and serum were taken from 30 patients, mainly with multiple sclerosis and other central nervous system dysfunctions, polyradiculoneuritis, known as Guillain-Barre syndrome, encephalitis, paraproteinemia, and analyzed.

ImageMaster 1D Elite and GelPro specialized software packages were used for fast accurate image and gel analysis. The results obtained from different hierarchic cluster analysis methods were compared. In some cases, despite substantial similarities between electropherograms, different cluster methods produced different dendrograms. Therefore, the cluster analysis should be used cautiously. It offers only additional diagnostic information on the inflammatory conditions of the central nervous system.

Received June 30, 2004
Accepted March 8, 2005

Keywords: disc electrophoresis, cerebrospinal fluid, protein analysis, cluster analysis

Identification and determination of different types of proteins play an increasing role in medical diagnosis. Conventional electrophoresis methods are well known for protein detection and analysis in cerebrospinal fluid (CSF). Cerebrospinal fluid analysis, coupled with other methods, remains the cornerstone of the diagnosis of various neurological disorders, including multiple sclerosis (MS) and infectious diseases of the central nervous system (CNS) (1, 2). Electropherograms are classified into different groups according to the qualitative and quantitative composition of cerebrospinal fluid regard to major protein fractions and CSF/serum albumin quotient.

For detection of oligoclonal IgG bands in serum and in unconcentrated spinal fluid, some techniques have been used, such as the isoelectric focusing (IEF) combined with polyethylene-enhanced gel (PEG) immunofixation and silver staining, CSF/serum quotient diagram, different body index, *etc.* (3–5).

* Correspondence, e-mail: filips@iunona.pmf.ukim.edu.mk

However, automation and development of microcomputers and software enabled rapid collection of large amounts of data. Image analysis software is used to extract much more information from the electropherogram for the purpose of comparative analisis between gels generated in-house or available in Web-based databases. Data acquisition, manipulation and computation for electrophoretic protein pattern recognition are performed using standard statistical signal analysis. Cluster analysis, along with artificial neural networks (ANN), is currently the next promising area of interest. Both have been successfully applied to various areas of medicine, such as diagnostic systems (6, 7), biochemical analysis (8) and image analysis (9).

Cluster analysis is a statistical method embedded in the commercial software of most program packages. Gay *et al.* (10) carried out application of cluster analysis in the staging of plaques in early multile sclerosis. They have shown that cluster analysis could be used in identification of distinct lesion groups and prediction of the stage of disease. The novelty of our research is the use of different cluster analysis methods for analysing CSF electrophoresis data.

EXPERIMENTAL

Patients

Pairs of CSF and serum were taken simultaneously during the course of lumbal puncture and venipuncture from 30 patients investigated and treated at the Clinic of Neurology, Faculty of Medicine in Skopje, Macedonia. CSF and serum were sampled under sterile conditions. If they were not analyzed within two days, they were stored at –20 °C. The patients were divided into two groups. Group A (*n* = 19, 10 women and 9 men) consisted of patients with MS and E, and group B (*n* = 11, 7 women and 4 men) consisted of patients with PRN and PP. Patients were aged from 19 to 51 years. Results were compared with the control group, consisting of 18 patients, the majority of which had a history of psychiatric diseases, and none had histories, symptoms or signs of neurological disease: magnetic resonance imaging and electrophysiological investigations showed no abnormalities, and the routine biochemical examinations of blood an dCSF gave normal results. Clinical experiments were performed according to the Regulations of the Macedonian Ethical Committee and Ministry of Health of the Republic of Macedonia.

Disc electrophoresis (DEP)

Disc electrophoresis was carried out on 7% polyacrylamide gel, using the electrophoresis system Canalco (USA). Cerebrospinal fluid (CSF) was used without preconcentration. Proteins were separated on polyacrylamide gels polymerized in glass tubes, approximately 5 mm in diameter and 15 cm in length. The main separating gel, about 8 cm long, was added into the glass tubes. The samples to be separated were loaded directly to the main separating gel. A shorter, approximately 1 cm long, stacking gel was poured on top of the separating gel and into this gel the CSF sample was added. The volume of CSF applied was dependent on the protein concentration of the CSF sample and each sample gel contained about 200 μg of proteins. The purpose of this stacking gel was to concentrate the protein sample into a sharp band before it entered the main separating gel. Tris (5 mmol L^{-1})-glycine (0.038 mol L^{-1}) buffer (pH 8.3) was used. The buffer contained an ionizable tracking dye, bromophenol blue, that allowed the electrophoretic run to be monitored. In addition sucrose solution (40%) was added. Electrophoresis was run at 5 mA per sample. After its completion, the gels were stained with Coomassie-blue and the stains were measured by microdensitometer Canalco Model 8I. Also stained gels were interpreted using the scanner Sharp JX-330 (Japan).

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

Data analysis

Data processing (normalization setting, background subtractions, resolution and smoothing) was performed by standard procedure. Automated, accurate analysis, intelligent data storage and sophisticated data acquisition were carried out with ImageMaster 1D Elite and Gel Pro software. Statistica 6.0 software was used for cluster analysis. Complete linkage (CL), unweighted pair-group average (UPGMA), weighted pair-group average and Ward's statistical cluster methods were used in the statistical analysis (11, 12).

Thirty different curve pattern forms of the electroperograms were digitalized.

RESULTS AND DISCUSSION

Using standard procedures, IgG factors were calculated and used as criteria for electropherogram classification. As positive results for MS cases were taken, those in which there were oligoclonal IgG patterns, the relative values were ≥ 15% with domination of a slow area in G-zone and with IgG quotient < 0.7.

Determination of IgG production was calculated according to Tibbling and Link index (13):

$$\text{IgG quotient} = (\text{CSF IgG}/\text{serum IgG})/(\text{CSF albumin}/\text{serum albumin})$$

$$(\text{reference limit} < 0.7)$$

The state of blood-CSF barrier was evaluated by CSF/serum albumin quotient:

$$\text{albumin quotient} = \text{CSF albumin}/\text{serum albumin}$$

$$(\text{reference limit} < 9)$$

There are therefore four possibilities: a) normal albumin and immunoglobulin quotients, b) normal albumin quotient associated with increased immunoglobulin quotient, that is, an intrathecal synthesis of immunoglobulins without blood-CSF barrier damage, c) an increase of both quotients, in the case of a pure transudation process through an impaired blood-CSF barrier, d) a disproportionate increase of the immunoglobulin quotient compared to the increase of the albumin quotient, in the case of impairment of the blood-CSF barrier associated with an intrathecal synthesis of immunoglobulins (5, 14, 15).

Images of the gels were captured with the scanner and the band pattern was obtained from the electropherograms. Five typical examples of CSF polyacrylamide patterns from patients with different neurological diseases are presented in Fig. 1. Software packages mention above were used to generate the curve pattern form of the electropherograms, and the curves were used for extraction of numerical data. Curve pattern forms of the electropherograms are presented in Fig. 2.

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 studies is cluster analysis based on similarity measurements. Cluster analysis is a technique used for combining observations into groups or clusters. Each cluster is homogenous with respect to certain characteristics. The first step in cluster analysis is to select a measure of similarity. The joining or tree clustering method uses the dissimilarities or distances between objects when forming clusters. These distances can be based on a single dimension or multiple dimensions. The most straightforward way of computing distances between objects in multidimensional space is to compute Euclidian distances (1). They are computed as:

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

where x_{ik} is the value of the k'th variable for the i'th object (curve pattern form of electropherogram *i*) and x_{jk} is the value of the k'th variable for the j'th object (curve pattern form of electropherogram *j*). x_{ik} and x_{jk} are elements of i and j column vectors.

Four different methods of cluster analysis were used in this work. Complete linkage method determines the distances between clusters by the greatest distance between any two objects in the different clusters. Unweighted pair-group average (UPGMA) is a method in which the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters. Weighted pair-group average is identical to the UPGMA, except in computations. Ward's method is different from all other methods because it uses an analysis of variance approach to evaluate the distance between clusters.

The results of the statistical cluster analysis using different methods are presented in Figs. 3a–d. Complete linkage method (Fig. 3a) produced two large clusters, I and II, each consisting of several clusters. Clusters 1–3 included samples referred to as MS and cluster 4 samples referred to as E cases. Cluster 5 contains PRN cases and cluster 6 is formed of PP cases. This method can be used in diagnosis because cases with different diseases are placed in different clusters. In order to obtain conclusions about diagnosis of a new sample numerical data should be prepared as explained and the next calculations should be performed.

Results of the UPGMA cluster analysis are presented in Fig 3b. Cluster I is formed of almost all samples with multiple sclerosis divided into 3 clusters. Cluster II consists of 19 samples and clusters referred to as cluster 4–7 is expected, the results produced with

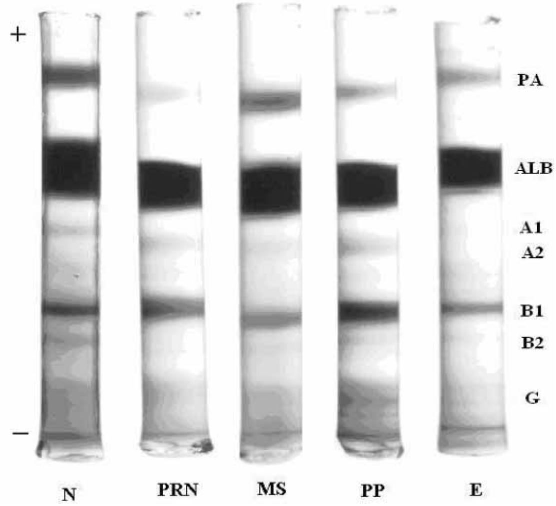

Fig. 1. Disc electrophoresis of CSF from: N – normal subject, PRN – polyradiculoneuritis, MS – multiple sclerosis, PP – paraproteinemia, and E – encephalitis patient. PA – prealbumin, ALB – albumin, A1 – A1-zone, A2 – A2-zone, B1 – B1-zone, B2 – B2-zone and G – G-zone (IgG immunoglobulines).

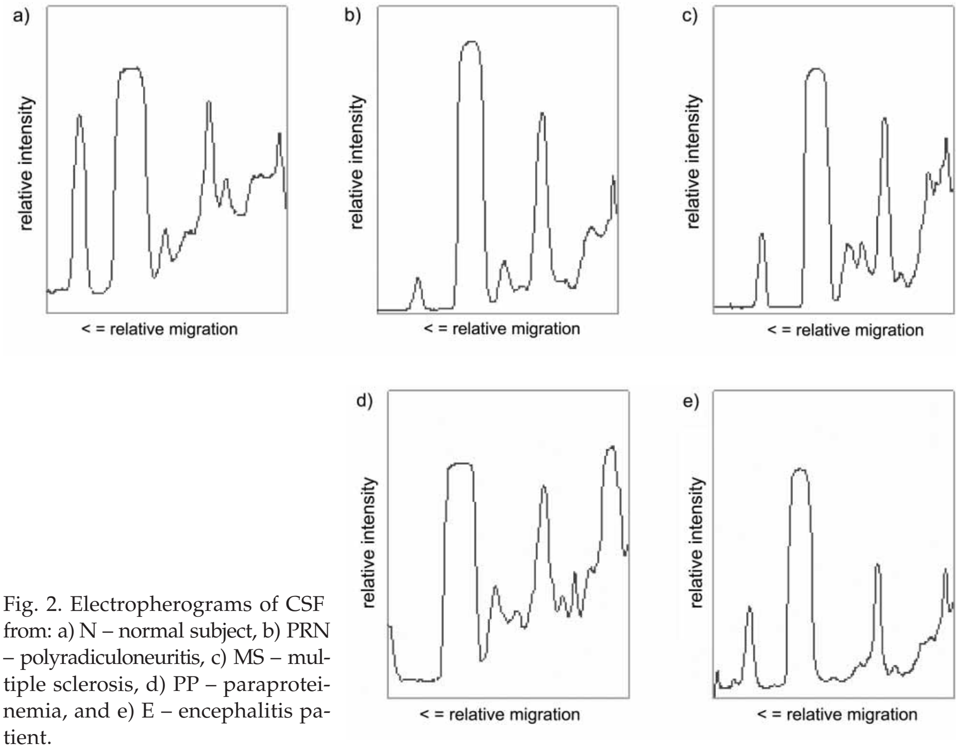

Fig. 2. Electropherograms of CSF from: a) N – normal subject, b) PRN – polyradiculoneuritis, c) MS – multiple sclerosis, d) PP – paraproteinemia, and e) E – encephalitis patient.

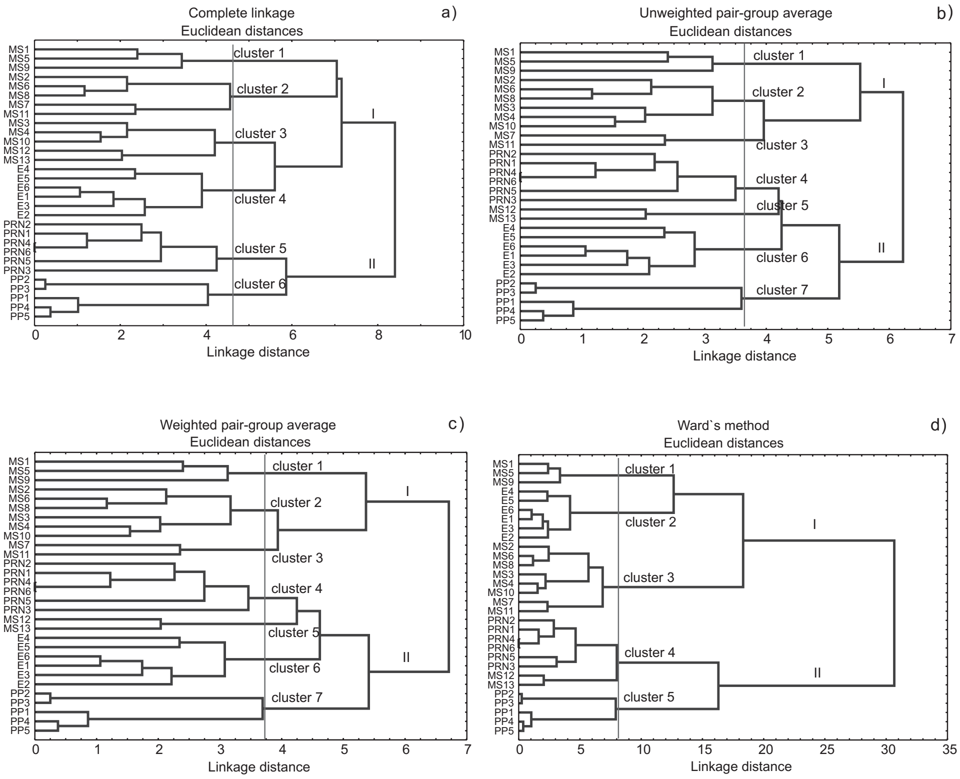

Fig. 3. Cluster analysis (joining tree model) using different methods: a) complete linkage, b) UPGMA, c) weighted pair-group average and d) Ward's method. PRN – polyradiculoneuritis, MS – multiple sclerosis, PP – paraproteinemia, E – encephalitis.

the weighted pair-group average method (Fig. 3c) are almost identical to UPGMA method. The only difference is in linkage distances due to differences in computation. Ward's method is presented in Fig. 3d. Cluster I is formed of samples with multiple sclerosis (clusters 1 and 3) and encephalitis (cluster 2). Cluster II is formed of samples with polyradiculoneuritis and paraproteinemia which are grouped in two distinct clusters, 4 and 5. Unfortunately, cluster 4 also contains two cases with multiple sclerosis. This fact, in addition to improper order of clusters 1–3, could lead to uncertain diagnosis and make Ward's method unsuitable for diagnostic purposes.

The results obtained by cluster methods show that application of the hierarchic cluster analysis (HCA) produced different dendrograms. However, the obtained dendrograms show that the analyzed electropherograms belong mainly to two distinct clusters with many subgroups. Euclidean distances between objects in multidimensional space, which is a measure of the dissimilarity, and the results of the Euclidean distance matrices for different cluster methods are available as supplemental material from the authors.

It is a well know fact 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, which is in concordance with our results (Figs. 3a–d). As one can see, different cluster methods lead to different shapes of dendrograms and will therefore cause difficulties in medical practice. In some cases, difficulties will arise because of the shape of the clusters. In the case of great similarities between samples, some algorithms might even fail to detect two clusters because of the intermediate points (similar values in the column vectors in the data matrix) (16).

In some cases of neurological diseases clustering is farily good for their differentiations and valuable information is obtained from dendrograms (Fig. 3a). However, when great similarities in the electroperograms exist and when undesirable mixing between clusters is present (Fig. 3d), additional analysis is needed. Clearly, the unsupervised classification method is not good enough to be used for clinical screening of neurological cases.

CONCLUSIONS

The cluster analysis has proved to be an attractive approach for the classification in electrophoretic protein or DNA pattern analysis. However, comparison of the results obtained by different cluster analysis methods has shown that different dendrograms were obtained and that the cluster analysis should be used cautiously. Therefore, there is a need of testing more than one cluster method and choosing one with the best prediction characteristics. From our critical point of view cluster analysis offers only additional diagnostic information on the inflammatory conditions of the central nervous system, and only coupled with conventional electrophoresis can lead to better medical relevance of the method.

REFERENCES

1. C. Sindic, M. Antwerpen and S. Goffette, The intrathecal humoral immune response: laboratory analysis and clinical relevance, *Clin. Chem. Lab. Med.* **39** (2001) 333–340.
2. H. Reiber, M. Otto, C. Trendelenburg and A. Wormek, Reporting cerebrospinal fluid data: knowledge base and interpretation software, *Clin. Chem. Lab. Med.* **39** (2001) 324–332.
3. A. Mitrevski, K. Stojanoski and P. Korneti, Detection of oligoclonal IgG bands in cerebrospinal fluid on polyacrylamide support media: Comparison of isoelectric focusing and disc electrophoresis, *Acta Pharm.* **3** (2001) 163–171.
4. W. W. Tourtellotte, A. R. Potvin, J. O. Fleming, K. N. Murthy, J. Levy, K. Syndulko and J. H. Potvin, Multiple sclerosis: measurement and validation of central nervous system IgG synthesis rate, *Neurology* **30** (1980) 240–244.
5. C. J. M. Sandic, P. Monteyne, G. Bigaignon and E. C. Laterre, Polyclonal and oligoclonal IgA synthesis in the cerebrospinal fluid of neurological patients. An immunoaffinity-mediated capillary blot study, *J. Neuroimmunol.* **94** (1994) 103–111.
6. W. Vogt and D. Nagel, Cluster analysis in diagnosis, *Clin. Chem.* **38** (1992) 182–198.
7. J. M. Jerez-Aragones, A combined neural network and decision trees model for prognosis of breast cancer relapse, *Artf. Intell. Med.* **27** (2003) 45–63.

8. P. J. G. Lisboa, A review of evidence of health benefit from artificial neural networks in medical intervention, *Neural. Net.* **15** (2002) 11–39.
9. I. Aizenberg, N. Aizenberg, J. Hiltner, C. Moraga and E. Meyer zu Bexten, Cellular neural networks and computational intelligence in medical image processing, *Image Vision Comp.* **19** (2001) 177–183.
10. F. W. Gay, T. J. Drye, G. W. Dick and M. M. Esiri, The application of multifactorial cluster analysis in the staging of plaques in early multiple sclerosis. Identification and characterization of primary demyelinating lesion, *Brain* **120** (1997) 1461–1483.
11. N. Saitou and M. Nei, The neighbour-joining method: A new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.* **4** (1987) 406–425.
12. R. R. Sokal and P. H. A. Sneath, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, Freeman, London 1973.
13. G. Tibbling, H. Link and S. Ohman, Principles of albumin and IgG synthesis in neurological disorders. I. Establishment of reference values, *Scand. J. Clin. Lab. Invest.* **37** (1977) 385–390.
14. L. Thomas, *Laboratory Diagnosis of Neurological Diseases. Clinical Laboratory Diagnostics: Use and Assessment of Clinical Laboratory Results*, TH-Books Verlagsgeselschaft, Frankfurt 1998, pp. 1308–1326.
15. H. Reiber, Die diagnostische Bedeutung neuroimmunologischer im Liquor cerebrospinalis, *Lab. Med.* **19** (1995) 444–462.
16. B. F. J. Manly, *Multivariate Statistical Methods – A Primer*, Chapman and Hall, New York 1986.

SAŽETAK

Usporedba statističkih klaster metoda u analizi proteina elektroforezom

FILIP SPIROVSKI, KIRO STOJANOSKI i ANGEL MITREVSKI

Standardne metode elektroforeze upotrebljene su u kvalitativnoj i kvantitativnoj analizi proteina cerebrospinalne tekućine (CSF). Detekcija oligoklonalnih IgG proteina u cerebrospinalnoj tekućini provedena je disk elektroforezom na poliakrilamidnom gelu. Analizirani su uzorci CSF i seruma od 30 pacijenata s multiplom sklerozom i drugim oboljenjima središnjeg živčanog sustava kao što su poliradikuloneuritis, poznat kao Guillain-Barre sindrom, encefalitis i paraproteinemia. ImageMaster 1D Elite i GelPro specijalizirani kompjutorski programi upotrebljeni su za brzu analizu slike i gela. Usporedbom rezultata dobivenih iz različitih hijerarhijskih klaster analiza utvđeno je da različite klaster metode ne daju iste rezultate. Usprkos sličnostima elektroferograma različite klaster metode u nekim slučajevima daju različite dendrograme pa je potreban oprez u interpretaciji rezultata. Klaster analiza daje samo dodatne dijagnostičke informacije o upalnom stanju središnjeg živčanog sustava.

Ključne riječi: disk elektroforeza, cerebrospinalna tekućina, analiza proteina, klaster analiza

Institute of Chemistry, Faculty of Natural Science, Skopje, Macedonia

Clinic of Neurology, Faculty of Medicine, Skopje, Macedonia