Clin Chem Lab Med 2008;46(8):1183–1188 © 2008 by Walter de Gruyter • Berlin • New York. DOI 10.1515/CCLM.2008.221

Computer-assisted detection of monoclonal components: results from the multicenter study for the evaluation of CASPER (Computer Assisted Serum Protein Electrophoresis Recognizer) algorithm

Agostino Ognibene[1,*], Maria S. Graziani[2], Anna Caldini[1], Alessandro Terreni[1], Gabriella Righetti[2], Maria C. Varagnolo[3], Ada Campanella[4], Marinella Martelli[5], Rita Mancini[6], Paolo Rizzotti[2], Mario Plebani[3], Marco Mori[4], Giovanni Gaspari[5], Roberto Motta[6], Gianni Galli[7], Massimiliano Fabris[8] and Gianni Messeri[1]

[1] Laboratorio Generale, Dipartimento di Diagnostica di Laboratorio, Azienda Ospedaliero Universitaria Careggi, Firenze, Italy

[2] Laboratorio di Analisi Chimico Cliniche ed Ematologiche, Ospedale Civile Maggiore, Azienda Ospedaliera di Verona, Verona, Italy

[3] Servizio di Medicina di Laboratorio, Azienda Ospedaliero Universitaria di Padova, Padova, Italy

[4] Laboratorio di Analisi, Ente Ospedaliero Ospedali Galliera, Genova, Italy

[5] Laboratorio Analisi, Ospedale Maggiore, Bologna, Italy

[6] Laboratorio Analisi Centralizzato, Azienda Ospedaliera S. Orsola-Malpighi, Bologna, Italy

[7] Beckman-Coulter, Milano, Italy

[8] Sisge, Torino, Italy

Abstract

Background: To investigate the potential use of Artificial Neural Network (ANN) in the evaluation of serum protein electrophoresis, we set up a multicenter study involving six Italian laboratories. For this purpose, we developed an algorithm named CASPER (Computer Assisted Serum Protein Electrophoresis Recognizer).

Methods: A total of 59,516 samples from the six centers were divided into three groups. Training and validation sets were used to develop the neural network, whereas evaluation set was used to test the performance of CASPER in recognizing abnormal electrophoretic profiles.

Results: CASPER showed 93.0% sensitivity and 47.4% specificity. CASPER sensitivity and specificity ranged in the six sites from 88% (site 3) to 97% (site 5) and from 36% (site 6) to 53% (site 3), respectively. Sensitivity for γ zone was 94.6%, for β zone 89.7% and for oligoclonal patterns 92.0%.

Conclusions: The sensitivity of the CASPER algorithm does not allow us to recommend its use as a replacement for the visual inspection, but it could be helpful in avoiding accidental misclassifications by the operator. Moreover, the CASPER algorithm may be a useful tool for training operators and students. This study evidenced a high inter-observer variability, which should be addressed in a dedicated study. Data set to train and validate ANNs should contain a huge range and an adequate number of different abnormalities.
Clin Chem Lab Med 2008;46:1183–8.

Keywords: artificial neural network; capillary electrophoresis; inter-observer variability; multicenter study; serum protein electrophoresis; standardization.

Introduction

Serum protein electrophoresis (SPE) is widely used in clinical laboratories for the detection of immunoglobulin monoclonal components (MCs), due to an abnormal clonal expansion of a single B cell. MCs may reflect the presence of severe, even if relatively rare, lymphoproliferative disorders (multiple myeloma, Waldenström's macroglobulinemia, AL amyloidosis) or can be associated with monoclonal gammopathy of undetermined significance (MGUS) (1). In a recent population-based study, the prevalence of MGUS was reported to be as high as 5%–8% in individuals of 70 years of age and older (2–4). When compared to the general population, subjects with MGUS show a risk of progression towards malignant disease which does not decrease with time, thus requiring an indefinite follow-up (3). As a consequence, a sensitive, rapid and reliable method to screen for the presence of an MCs is essential. This is usually performed by a skilled operator who visually inspects a large number of patterns to select the samples to be typed for MC characterization. At the same time, the inspection should be specific enough to avoid unnecessary second level tests, such as serum and urine immunofixation.

In the last 10 years, capillary electrophoresis (CE) has been introduced in clinical laboratories and has proved to be a reliable tool to detect MCs (5–7). The CE reading device produces digital data accessible to mathematical analysis; it is therefore possible to create a computerized algorithm that may be able to

*Corresponding author: Agostino Ognibene, Laboratorio Generale, Dipartimento di Laboratorio, Piastra dei Servizi, Azienda Ospedaliero Universitaria Careggi, Viale Morgagni 85, 50139 Florence, Italy
Phone: +39-055-4279424, Fax: +39-055-4279416,
E-mail: ognibenea@aou-careggi.toscana.it
Received November 30, 2007; accepted April 1, 2008

identify abnormal electrophoretic patterns with the aim of lowering the cost and time of the examination and to decrease inter- and intra-observer variability.

Despite CE systems being common in clinical laboratories, few efforts have been made in the development and validation of such computerized programs and the reported results have not been fully satisfactory (8–12).

The main drawbacks encountered in the evaluation of these algorithms are: the huge range of electrophoretic characteristics of MCs and the lack of unequivocal criteria for the classification of electrophoretic patterns. To overcome the first problem, the set of samples used for Artificial Neural Network (ANN) training should include a sufficient number of different MCs in terms of electrophoretic mobility, shape and concentration of the peak. Visual inspection by a skilled operator is assumed to be the ''gold standard'' when evaluating the performance of computer-aided procedures, but, to the best of our knowledge, no data exist about the reliability of the human operator and about intra- and inter-observer variability. It is generally believed that classification of electrophoretic patterns by visual examination is highly dependent on the operator's skill and is therefore highly subjective.

In a previous paper (13), coupling CE with an ANN-based algorithm, we obtained a sensitivity of 98.4% and a specificity of 80.6% in a set of 4971 samples consecutively collected in a single center. To further investigate the possible potential use of ANN, we set up a multicenter study involving six Italian laboratories. For this purpose, the algorithm, now named CASPER (Computer Assisted Serum Protein Electrophoresis Recognizer), has been reviewed and trained again. The aim of this paper is to present and discuss the results of the study.

Materials and methods

Participants

A total of six Italian laboratories participated in the study: Laboratorio Generale, Azienda Ospedaliero Univesitaria Careggi, Firenze (site 1), Laboratorio di Analisi Chimico-Cliniche ed Ematologiche, Ospedale Civile Maggiore, Verona (site 2), Servizio di Medicina di Laboratorio, Azienda Ospedaliero Universitaria, Padova (site 3), Laboratorio di Analisi, Ente Ospedaliero Ospedali Galliera, Genova (site 4), Laboratorio Analisi, Ospedale Maggiore, Bologna (site 5), and Laboratorio Analisi Centralizzato, Azienda Ospedaliera S. Orsola-Malpighi, Bologna (site 6).

Capillary electrophoresis system

CE was performed on the Paragon CZE 2000® (Beckman Coulter, Fullerton, CA, USA) according to the manufacturer's instructions. Detection is accomplished via an optic fiber connected to a UV detector, for direct reading of protein fractions at 214 nm. Results are automatically transferred from the instrument software to LabItUp Millennium (Sisge s.r.l., Turin, Italy) middleware.

CASPER algorithm

The neural network was constructed by means of general-purpose software (NeuralSolutions, ver. 4.0, Neural Dimension, Gainesville, FL, USA) and executed under Windows XP (Microsoft, Redmond, WA, USA). The number of input values determined the number of units in the input layers. The network was constructed with two middle layers with 22 total units. The hyperbolic tangent activation function was used with continuous output on the interval (−1, +1).

The variables to train neural networks to classify sample abnormalities in β and/or γ regions were selected as follows:

- Absorbance variables: the Paragon CZE 2000 software includes an option to export aligned curve information through the RS-232 interface. Raw absorbance data were distributed in 250 time points along the time axis and fed into the ANN. To analyze the data, mathematical algorithms were developed and tested using LabItUp Millennium software.
- Statistical variables: for γ region analysis, kurtosis and skewness of the curve were added as additional variables and elaborated by the ANN together with absorbance data. These statistical parameters were calculated assuming that the absorbance measurements represent a distribution curve of absorbance (protein concentration) versus time (electrophoretic mobility).

The learning rule used for training was back-propagation of error; this procedure adjusts the weight of the connections in the network to minimize square error between actual output vector and desired output vector. To avoid overtraining, neural network training was stopped when the sum of squared error compared to the validation data set was at the minimum. The network was trained and validated many times with different and randomly selected training and validation data sets.

CASPER tool

For the multicenter study, a dedicated program, CASPER tool, was developed by Sisge (Turin, Italy) and installed in the manager software of the six laboratories. Data from the LabItUp Millennium archives were exported and formatted for the CASPER tool blinded by any information of the patient and previous SPE classification. The operator was asked to classify the samples in two ways: normal (no evidence or suspect of MC) or abnormal (MC in the β region, MC in the γ region, oligoclonal pattern, further investigation required). After the classification was given, the program showed the result of the CASPER algorithm classification and the operator's choice could not be modified.

Sample collection

Data obtained from fresh samples submitted to the laboratories for serum protein electrophoresis from January 2005 to May 2006 were retrieved from the archives of the six laboratories and imported into the CASPER tool. For each sample, the absorbance values and the operator's classification were registered.

Sample allocation

A total of 59,516 samples from the six centers were divided into three groups and used for training, validation, and evaluation of the neural network. Training and validation sets were used to develop the ANN; evaluation set was used to test the performance of the CASPER algorithm in recognizing abnormal serum protein profiles. Table 1 shows the sample distribution in the three groups:

- Training and validation: 22,650 samples from sites 1 and 2; both absorbance variables and operator's classification were provided to the system. These samples were selected by the same operator (A.O.), to include an adequate number of abnormal samples with different migration patterns.
- Evaluation: 36,866 samples were classified by different operators in the six centers using the CASPER tool software. Only absorbance variables were provided to the neural network to compare the CASPER algorithm and operator's classifications.

Results

During ANN development, the network was trained and validated many times with randomly selected training and validation data sets and very similar performances were obtained in different training and validation exercises (data not shown).

The number of the samples obtained, the percentage of abnormal samples and the type of abnormalities show huge differences among the sites; the lowest number of abnormalities was observed in site 5 (8%) and the highest in site 1 (28%) (Table 2).

When the CASPER algorithm was used to classify the samples of the evaluation group, it showed 93.0% sensitivity and 47.4% specificity (Table 3). Out of 562 false negative samples, 182 (32%) were classified as ''oligoclonal patterns'' by the operator. CASPER sensitivity and specificity varied significantly among the six sites ranging from 88% (site 3) to 97% (site 5) and from 36% (site 6) to 53% (site 3), respectively (Table 4). Sensitivity is different when calculated considering the type of abnormality: for γ zone, MC sensitivity was 94.6%, for β zone MC it was 89.7%, and for oligoclonal patterns 92.0%, respectively.

Figures 1 and 2 show examples of samples worthy of further investigation in the operator's judgment and considered normal by the CASPER algorithm (false negative samples). Samples reported in Figure 1 show evident abnormalities in the β or γ region (real false negative samples), whereas the electrophoretic patterns of samples reported in Figure 2 are apparently normal (possibly indicating an excess of precaution by the single operator).

Table 1 Number and classification of samples included in the three groups.

	Training	Validation	Evaluation	Total
Normal	12,250	2600	28,964	43,814
Abnormal	6300	1500	7902	15,702
Total	18,550	4100	36,866	59,516

Table 3 Comparison of results obtained by visual inspection and CASPER classification on the evaluation group of samples (n=36,866).

ANN classification	Visual inspection classification	
	Normal	Abnormal
Normal	13,734	562
Abnormal	15,230	8393
	Sensitivity 93.0%	Specificity 47.4%

Table 4 Sensitivity and specificity of CASPER algorithm in the six centers.

	Sensitivity, %	Specificity, %
Site 1	96	44
Site 2	89	50
Site 3	88	53
Site 4	95	52
Site 5	97	39
Site 6	94	36

Discussion

The major clinical indication for serum protein electrophoresis is the investigation of plasma cell dyscrasia producing monoclonal immunoglobulins. MCs are detected by visual inspection of electrophoretic patterns; when an abnormality is evident or suspected, the sample is immunotyped to verify the class and the type of the monoclonal immunoglobulin. This screening activity is usually performed by a skilled operator, but to date no data are present in the literature concerning variability and reliability affecting the procedure, although both are usually perceived as significant. ANN can be instructed to detect electrophoretic abnormalities, utilizing CE analog absorbance signals, with the aim of assuring more uniform behavior than the human operator. Before this software can be used in laboratory practice, the algorithm performances have to be fully investigated. In a pre-

Table 2 Number and operator's classification of samples collected in the six sites and used for the evaluation of CASPER algorithm.

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Total
Total	12,052	11,237	5330	3162	3132	1575	38,275
Normal	8591	9721	4113	2378	2873	1288	28,964
Abnormal[a]	3461	1516	1217	784	259	287	7902
γ zone MC	1801	1048	787	294	131	181	5597
β zone MC	815	375	107	53	107	22	1479
Oligoclonal pattern	1115	132	364	458	23	89	2181

[a] Since a single sample can show more than one abnormality (i.e., γ MC plus oligoclonality), the data in this row can be higher than the sum of specific abnormalities.

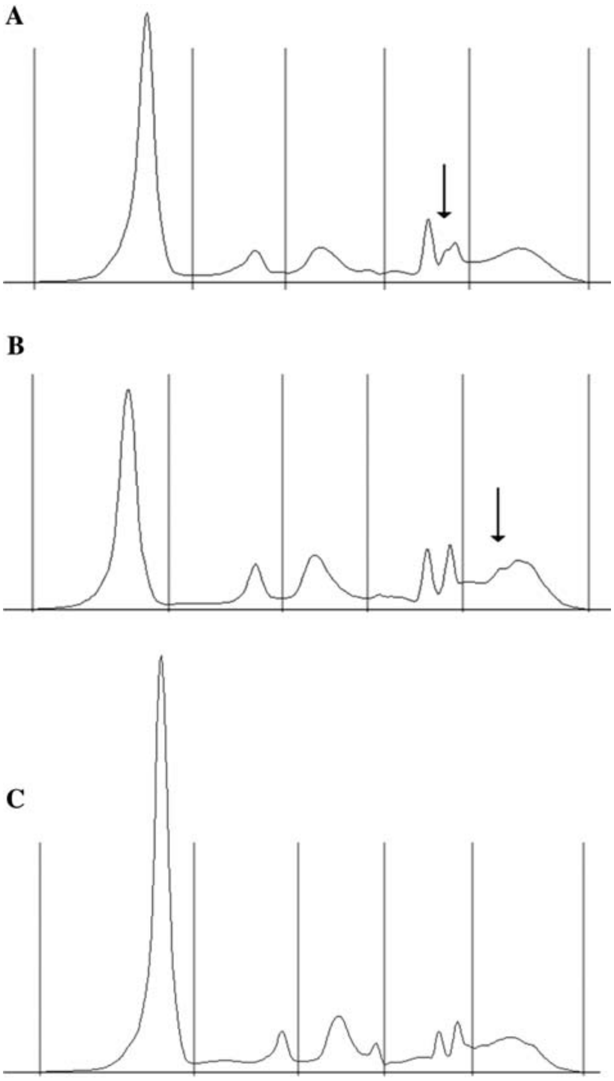

Figure 1 Examples of false negative samples: electropherograms evaluated normal by CASPER and in which an abnormality was checked by the operator in β region (A), γ region (B) or as oligoclonality (C).

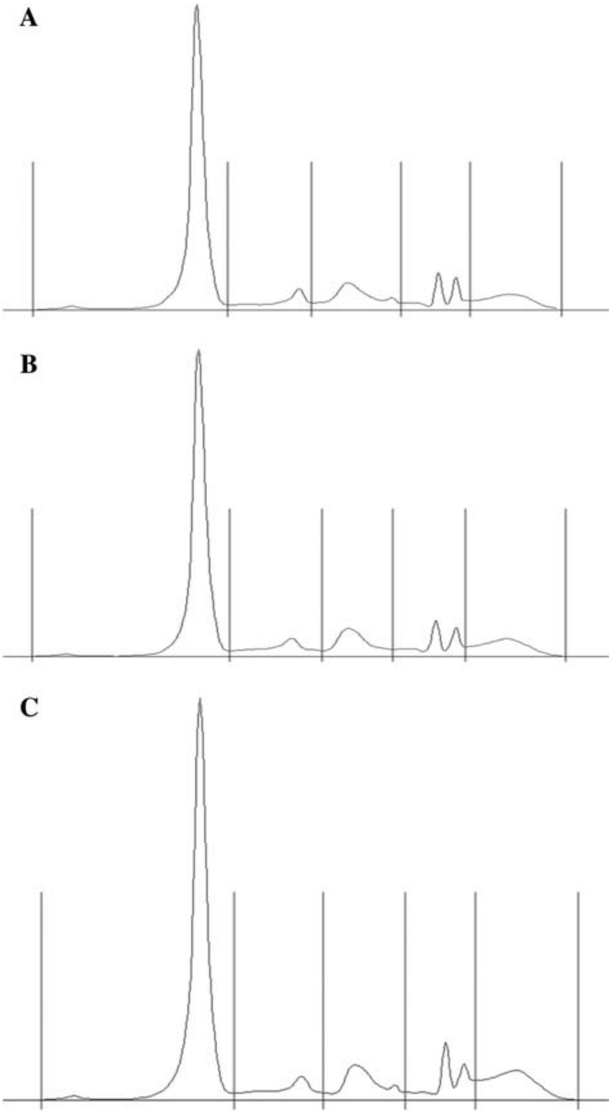

Figure 2 Examples of false negative samples: electrophoretic patterns evaluated normal by CASPER and in which the abnormality checked by the operator in β region (A), γ region (B) or as oligoclonality (C) is hardly visible.

liminary paper (13), we obtained a sensitivity of 98.4% and a specificity of 80.6% analyzing 4971 samples from a single center classified by the same operator. In our opinion, the better performance of the algorithm used in that study was due to the small number of abnormal samples included, and to the lack of certain types of abnormalities which are more difficult to detect, such as alterations of the β zone and oligoclonal banding. In fact, this study analyzed 1479 samples with abnormalities in the β region compared to less than 100 in our previous study (13). On the basis of the preliminary results, the ANN was modified, trained with a new set of samples and used in the multicenter study.

Concerning the sensitivity showed by the CASPER algorithm (94.6% and 89.7% for MCs in γ and β regions, respectively), it should be noted that missed MCs (380 out of 562 discordant samples) were ''difficult'' MCs (of low concentration, co-migrating with normal bands, with a heavy polyclonal background), whereas ''well-represented'' MCs have always been correctly classified by the algorithm.

The remaining third of false negative samples (n=182) was represented by the so-called ''oligoclonal patterns'' where multiple small and very small bands alter the normal electrophoretic profile. It is well recognized that the alteration is difficult to define and even among expert operators, a degree of uncertainty exists in the classification of these samples. These gammopathies have been associated with bone marrow or solid organ transplantations or viral infections, and are usually of low concentration and short-lived and their clinical significance remains obscure (14).

Analyzing samples from different sites and classified by different operators, we were able to examine the ANN performances in a situation very close to the daily routine practice in clinical laboratories and to highlight the important inter-observer variability in the pattern classification. The non-uniform distribution of pattern abnormalities among the centers (i.e., oligoclonal banding ranging from 0.7% of the site 5 to 14% of site 4; Table 2) could only be explained in part by differences in the clinical situation of patients admitted to different hospitals; really, it could be largely attributable to inter-observer variability. The different approach of different operators to the electrophoretic pattern classification is also evident from other observations. First, sensitivity calculated according to the type of abnormality increased to 94.6% when considering γ region, and decreased to 89.7% when considering β region. So, the sensitivity is obviously better when the ANN examines evident and well-represented abnormalities, such as the ones visible in the γ zone. On the other hand, the detection of the less clear ones is highly dependent on the operator's skill and to the subjective approach to the specific problem. Furthermore, in the β region, MCs can be superimposed upon other proteins, making their detection less easy compared to the γ region. Second, the examples of false negative samples presented in Figure 2 clearly show that the operator's classification of these samples as ''abnormal'' is highly questionable. Third, as shown in Table 4, the difference observed in specificity and sensitivity between sites is surprising. As the performance of CASPER is obviously constant, this variation is attributable, at least in part, to inter-observer variability. These remarks emphasize one of the main problems encountered when evaluating the diagnostic performances of such ANN: the use of a subjective evaluation as the ''gold standard''.

A number of studies have been published some years ago (8–10) aimed at evaluating a rule-based system for MC detection, and the reported sensitivities ranged from 76% (8) to 97% (9). These studies are difficult to compare to the present one, because they utilized cellulose acetate to perform electrophoresis, because the number of samples was quite low ($<$1000), and because the type and the number of abnormalities included was not clearly stated. More recently, Jonsson et al. (11), using a rule-based system to evaluate CE protein profiles, obtained a sensitivity of 98% and a specificity of 99% in detecting ''well-represented'' MCs in a set of 711 samples. This study differs from ours, because it was performed in a single center, and because of the much lower number of samples examined.

Our study was the first to evaluate the ANN performance in a multicenter study, and with a clear definition of the number and type of abnormalities of the samples examined. The design of the study allows us to highlight the importance of the inter-observer variability when examining the morphology of electrophoretic patterns.

In conclusion, the main points of the present study can be summarized as follows:

- Given the differences observed between the preliminary study (13) and this study, it is clear that the data set to train and validate ANN should contain a huge range and an adequate number of different abnormalities. Really, the performances of ANNs seem highly dependent on the characteristic (number, type, size and migration position) of the abnormal samples included in the training set.
- The inter-observer variability is higher than expected, considering that only well-trained operators who have used CE for many years were involved in the study. In our opinion, this point deserves to be addressed in a dedicated study, to quantify the size of this variability. The CASPER tool seems to be appropriate to do this.
- The sensitivity of the CASPER algorithm as verified in this study does not allow us to recommend it as a replacement for visual inspection, considering that false negative samples due to less evident abnormalities might represent a major clinical problem, i.e., amyloidosis diagnosis missing.
- ANN could instead be useful in avoiding gross operator's misclassification errors, especially in medium to large laboratories where a large number of patterns have to be visually inspected every day. It is, however, worthy to note that appropriate communication from the clinician to the laboratory is essential in focusing attention on critical samples.
- The CASPER algorithm may be a useful tool for training operators and students. Considering the diagnostic performance of the ANN, any discordance should be discussed with an expert tutor.

References

1. Kyle RA, Rajkumar SV. Monoclonal gammopathy of undetermined significance. Br J Haematol 2006;134: 573–89.
2. Kyle RA, Terry MT, Therneau S, Rajkumar SV, Larson DR, Plevak MF, et al. Prevalence of monoclonal gammopathy of undetermined significance. N Engl J Med 2006;354:1362–9.
3. Kyle RA, Rajkumar SV. Monoclonal gammopathies of undetermined significance. Rev Clin Exp Hematol 2002;63:225–52.
4. Aguzzi F, Bergami MR, Gasparro C, Bellotti V, Merlini GP. Occurrence of monoclonal components in general practice: clinical implications. Eur J Haematol 1992;48:192–5.
5. Bienvenu J, Graziani MS, Arpin F, Bernon H, Blessum C, Marchetti C, et al. Multicenter evaluation of the Paragon CZE™ 2000 capillary zone electrophoresis system for serum protein electrophoresis and monoclonal component typing. Clin Chem 1998;44:599–605.
6. Gay-Bellile C, Bengoufa D, Houze P, Le Carrer D, Benlakehal M, Bousquet B, et al. Automated multicapillary electrophoresis for analysis of human serum proteins. Clin Chem 2003;49:1909–15.
7. Roudiere L, Boularan AM, Bonardet A, Vallat C, Cristol JP, Dupuy AM. Evaluation of a capillary zone electrophoretic system versus a conventional agarose gel system for routine serum protein separation and monoclonal component typing. Clin Lab 2006;52:19–27.
8. Knüppel W, Neumeier D, Fateh-Moghadam A, Knedel M. Computer-assisted findings on protein electrophoresis on cellulose acetate film. J Clin Chem Clin Biochem 1984;22:407–17.
9. Kratzer MA, Invandic B, Fateh-Moghadam A. Neuronal network analysis of serum electrophoresis. J Clin Pathol 1992;45:612–5.
10. Manner GA, Schweiger CR, Söregi G, Pohl AL. Detection of monoclonal gammopathies in serum electrophoresis by neural networks. Clin Chem 1993;39:1984–5.
11. Jonsson M, Carlson J, Jeppsson J-O, Simonsson P. Computer-supported detection of M-components and evaluation of immunoglobulins after capillary electrophoresis. Clin Chem 2001;47:110–7.

12. Jonsson M, Carlson J. Computer-supported interpretation of protein profiles after capillary electrophoresis. Clin Chem 2002;48:1084–93.
13. Ognibene A, Motta R, Caldini A, Terreni A, Dalla Dea E. Fabris M, et al. Artificial neural network-based algorithm for the evaluation of serum protein capillary electrophoresis. Clin Chem Lab Med 2004;42:1451–2.
14. Alexanian R, Weber D, Liu F. Differential diagnosis of monoclonal gammopathies. Arch Pathol Lab Med 1999; 123:108–13.