Proceedings of the V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
pp. 37–42, April 2–5, 2001 - Rio de Janeiro - RJ - Brazil

Feedforward Neural Networks for the Classification of Two-dimensional Polyacrylamide Gel Electrophoresis Images

Ariel Cary, Davor Pavisic, Reynaldo Vargas
Instituto de Investigación en Informática Aplicada (IIIA),
Universidad Católica Boliviana, Cochabamba, BOLIVIA.
E-mail: dp@ucbcba.edu.bo

Abstract

This article describes a method for classifying two-dimensional polyacrylamide gel electrophoretograms, complex biomedical images that contain proteins separated from a biological sample, using neural networks. The classification aims at grouping images and identifying their most significant features. The gel image processing part is first summarized. The details on how the classification is accomplished using neural networks are then presented. After that, an experiment using real gels of rat cells is carried out, showing the successful implementation and application of this method. Finally, experimental results show that this neural network based method is fairly effective.

1. Introduction

Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) is a powerful biochemical technique for separating proteins contained in a biological sample [1, 2] and is widely used in proteomic research [3].

The proteins migrate on a polyacrylamide gelaccording to two important characteristics: their isoelectric point and their molecular weight. After staining the gel, one can observe spots that are spread over the gel according to these two characteristics (Fig. 1). Typically, a 2D PAGE map is characterized by a thousand or even more protein spots. These protein maps can then be digitized for further processing in a computer.

Due to the high complexity of these biomedical images, visual comparison and classification of 2D PAGE images is almost impossible. Consequently, advanced powerful computer systems must be developed in order to detect and quantify the spots contained in a 2D PAGE map [4], and new algorithms must be used for classifying them.

In the field of medical diagnosis, the classification of 2D PAGE maps turns out important for clinicians since it provides means for grouping these images in accordance with their similarities and differences. These groups may be associated with certain diseases, and an early diagnosis may be stated. The classification, therefore, aims at correctly separating gels among a specific set of classes.

Besides an automatic classification, what is also important is the identification of the characteristic spots of the classes. That is, those spots that differentiate the most one class from other. This would indicate what the spots linked to a particular disease are.

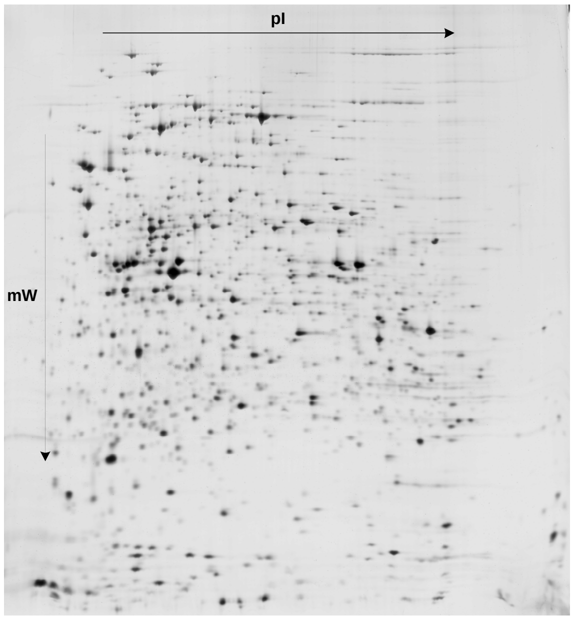

Figure 1: Two-D PAGE image of escherichia coli [5]. The spot position indicates the isoelectric point (*pI*) and the molecular weight (*mW*), and the spot darkness is related with the concentration of the protein.

In this article, a method for classifying 2D PAGE images using feedforward neural networks is described. Also, the sensitivity analysis technique is applied over the trained networks for identifying the characteristic spots. Finally, an experiment using real gels is presented showing the successful implementation of this method in the context of the MELANIE software for the analysis of two-dimensional polyacrylamide gel electrophoresis developed at the Swiss Institute of Bioinformatics (SIB) [6].

2 Classification Overview

In general terms, an image containing a total number of N points (pixels) can be seen as having N features (the color levels at the N pixels). Thus, M images of size N can be represented as M points in an N-dimensional space, where each dimension corresponds to a particular feature.

In the context of 2D PAGE maps, each gel is characterized by a large quantity of protein spots; typical values of *N* range from 500 to 2000 spots. Therefore, given a set of *M* 2D PAGE maps of size *N*, each one corresponding to a different individual *j* in *[1..M]*, and each feature *i* in *[1..N]* corresponding to the intensity of a protein spot, the classification problem can be stated as how can these points, each corresponding to one 2D PAGE image, be clustered into meaningful classes [7]. Consequently, the classification process aims at grouping 2D PAGE maps into clinical relevant classes (e.g., normal, pathological, etc.) by considering the quantified values of their spots.

The classification of 2D PAGE images can be carried out using two types of classification techniques [1]. The first type, known as unsupervised learning, automatically forms classes from a set of images and highlights the protein patterns specific to each class. The second category is known as supervised learning. In this case, classes are known in advance (e.g., class of disease A, and class B of normal gels). The programs here are used to determine disease markers for subsequent classifications of additional gels.

After the classification is done, what is also important is the identification of significant features differentiating the individuals, namely the identification of the characteristic spots. This would show what spots are responsible for certain classes (diseases).

3 Global 2D PAGE Image Analysis

The spots on a 2D PAGE image can be detected, and the intensity of each spot can be quantified. Several feature value types of interest may be calculated; for example, the spot area, volume, and optical density.

After spot detection and quantization, in order to correct possible distortions on the images and so that all gels become superimposable, a pixelwise correction (aligning) has to be applied over the gels according to a reference gel, which may be the best gel chosen visually.

Once the gels get superimposable, a matching between the images can be executed in order to find spot pairs from the different gels. As a result, spots corresponding to the same protein in various gels are found, and a list of spots that have been matched among the gels can be formed. This list is referred to as a spot group, and a total of N spot groups are formed after the matching process (Fig. 2).

For carrying out the 2D PAGE map protein spot detection and quantization, image aligning and matching processes, some techniques have been proposed and successfully tested in [4].

4 Two-D PAGE Image Classification Using Feedforward Neural Networks

Feedforward neural networks [8], using the backpropagation algorithm [9] as the learning rule, can be trained to classify 2D PAGE images. This neural model belongs

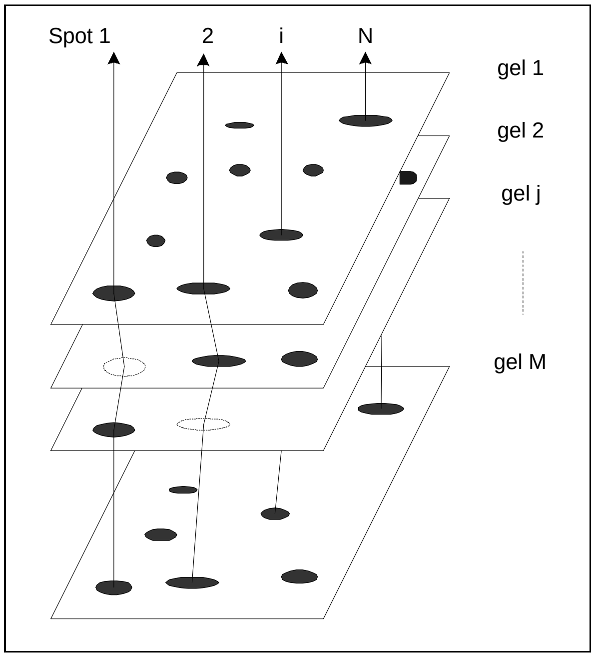

Figure 2: Building the spot groups [7]. Spots describing the same protein in different images are found by matching the gels.

to the category of supervised learning methods. That is, the corresponding class *c* in *[1..C]* of each image of a set of M 2D PAGE maps is known.

Each class is made up of a set of gels and represents a specific concept (e.g., normal, pathological). The goal of the training is that the network learn the subjacent concept of each class by means of the samples contained in each class. Then, the knowledge acquired by the network can be used to classify new gels not seen during training.

This task is achieved in two phases. In the first one, the training phase, the network learns to classify the gels using a portion of the M gels as its training set. In the second phase, the testing phase, the performance of the network is measured to check whether it is able to generalize the knowledge acquired during the training phase to classify a new set of gels, the remaining of the M gels (testing set), not seen during training (Fig. 3).

Furthermore, once the network reaches an acceptable performance, it can be used to classify additional 2D PAGE images (neither training nor testing pattern) into one of the learned classes. Additionally, the sensitivity analysis technique can be used over the trained network in order to identify what the most important image features are. The rest of this section describes some important details to be considered when implementing such a network.

4.1 Architecture

The feedforward network architecture depends on how many classes *C* are defined in the set of *M* images, and how many features *N* characterize each gel. Thus, the

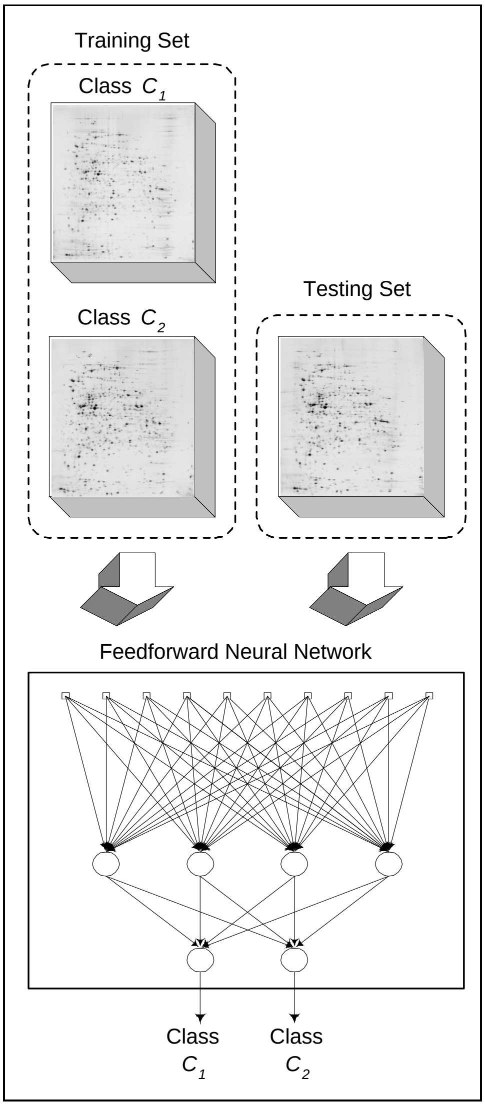

Figure 3: Classification of 2D PAGE images using neural networks. The network learns the concept of each class by means of the samples contained in the training set. Then, images from the testing set are evaluated by the network using the knowledge acquired during training.

number of input neurons is defined by *N*, one input neuron for each image feature. The size of the output layer is determined by *C*, one output neuron for each class. Finally, one can consider either one or two hidden layers.

4.2 Training, Testing, and Classification

A percentage *p* of the *M* gels is taken for making up the training set. The training is then carried out using the backpropagation learning algorithm [9] and so that the corresponding output neuron of the gel's class being presented at the input is highly activated (near to one), whereas the rest of the output neurons are off (near to zero). The rest of the gels *(1-p)xM* are used for testing the overall performance of the network after the training has been completed. Afterwards and in the same way, completely new gels, for which their associated class is unknown, can be classified to one (only one output neuron is highly activated), many (two or more output neurons are highly activated meaning that the gel shares characteristics of various classes), or none (all output neurons are nearly off meaning that the gel does not resemble to any class) of the learned categories, and an early diagnostic may be stated according to which class the gel belongs.

4.3 Sensitivity Analysis

Basically, the sensitivity analysis technique applies small changes at the inputs, one training pattern at a time, to study variations on the outputs [10]. One can say that the bigger the output variation, the more important the corresponding input neuron is. In this way, one can identify what the most important, influencing input neurons are for the functioning of the network. Although it is a simple technique, it has been shown that it is fairly effective, and it has been applied successfully several times [11, 12, 13].

In the context of 2D PAGE maps, typical spots corresponding to particular classes can be determined using the sensitivity analysis technique over the trained networks. One expects that the most influencing input neurons correspond to the characteristic spots. As the results of the experiment presented in section 5 will show, this is true in that this technique has successfully found several features that are visually discriminating between classes.

5 Experiment

At the Researching Institute on Applied Computing (IIIA) of the Catholic University of Bolivia, the method presented in the previous section has been implemented and tested [14]. In this section, an experiment using real gels of rat vessel smooth mussel cells is presented for showing the behavior of the implemented neural network in a supervised classification task.

The 2D PAGE images used in this experiment were taken from a larger experiment carried out at the Department of Pathology of the Faculty of Medicine at the University of Geneva, Switzerland [15]. The samples come from two different populations: newborn rats (*NN*) and two-year old ones (*VE*). The goal of this experiment is to train a feedforward neural network to discriminate these two categories of gels and to identify protein spots that are differentially expressed in the two populations.

5.1 Data

Twenty 2D PAGE images (10 of each population) were used in two phases. In the first phase, the best five gels of each population, chosen visually, were analyzed by computer and used to make up the training set. The network was then trained accordingly. In the second phase, the remaining 10 gels were also analyzed and used to verify the results of the training phase, namely used as the testing set.

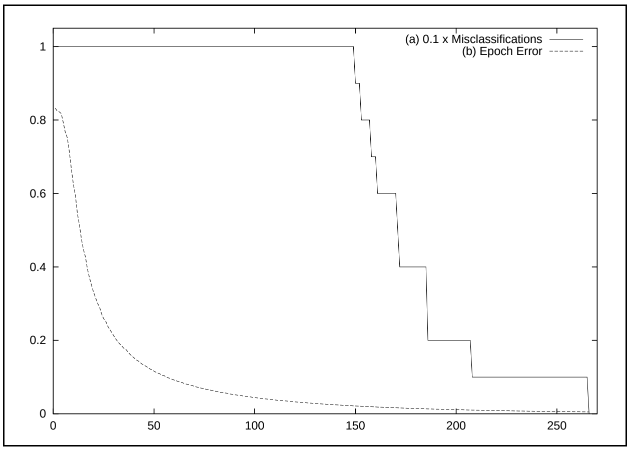

Figure 4: Neural network learning curves: (a) The misclassification curve, at the beginning the ten training gels are misclassified, the training ends when all of these gels are learned; (b) The epoch error curve.

5.2 Image Analysis

The image analysis part (spot detection and quantization, image aligning and matching) has been carried out using the MELANIE software package for 2D PAGE image analysis [6]. First, the protein spots of the twenty 2D PAGE images were automatically detected and quantified. Between 519 and 1509 spots were detected on the gels, and the optical density value of each spot was computed; this value is directly related to the concentration of the protein. Second, a reference gel was selected from the set of images and each of the remaining gels were then corrected pixelwise so that all gels become superimposable. Third, the gels were automatically matched pairwise; i.e., correspondences between the spots in different gels have been searched. A total number of 867 correspondences were found. Thus, each gel is characterized by 867 protein spot values. Finally, two matrices were built for the training and testing sets respectively. Each matrix contains 10 lines (one per gel) and 867 columns (one per spot). The two matrices contain the spot values representing the optical density measure.

5.3 Architecture

In this experiment, the size of the input layer is 867, one input neuron for each feature spot characterizing the images. The number of output neurons is two since there are two different categories. Two hidden neurons haven been used in one hidden layer. In summary, the network layout is 867-2-2.

5.4 Training and Classification

The 867-2-2 network was trained with the backpropagation algorithm using the ten cases (five of each class) of the training set. The training phase was stopped when all the training patterns were learned. A total number of 266 epochs were executed (Fig. 4).

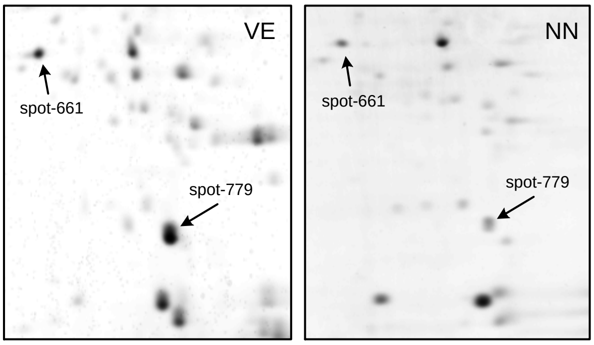

Figure 5: Characteristic spots found by sensitivity analysis. The protein spots 661 and 779 are elevated in 2D PAGE of old rats (*VE*).

Table 1: Actual network outputs. The class to which the gel is supposed to belong is indicated by the highest output value in boldface.

Training Gel	Def. Class	NN	VE
93008ve-w	VE	0.101778	**0.898324**
93013ne-w	NN	**0.878340**	0.121558
93019ce-w	NN	**0.746901**	0.253135
93028ce-w	VE	0.126711	**0.873548**
93029ve-w	VE	0.105309	**0.895055**
93030ve-w	VE	0.359482	**0.640692**
93031ve-w	VE	0.137114	**0.863317**
93036ne-w	NN	**0.899081**	0.100806
93038ne-w	NN	**0.899333**	0.100556
93039ne-w	NN	**0.898943**	0.100925

During the testing phase, the ten gels of the testing set, each one being associated to one class, were presented to the network, and 100% of them were correctly classified. The actual network outputs for the ten testing patterns are shown in Table 1.

5.5 Characteristic Spots

Once the network was trained, the sensitivity analysis technique, as described in section 4, was applied to find the most important inputs. In this way, ten protein spots were identified as the most differentiating features between the classes. Some of those discriminating spots are shown in Figure 5 in two gels of different classes.

5.6 Results and Discussion

The 867-2-2 network learned the characteristics and regularities of the training set, discriminating the gels of the two populations. Furthermore, during the testing phase, all of the testing gels were correctly classified, reaching an overall performance of 100% over gels not seen during training.

Ten characteristic spots were also identified as the differentially expressed protein spots of the classes. These spots were then verified as being differentiating ones; the concentration of the protein is elevated in gels of one class, and in some cases a protein is found in only one category. Consequently, the sensitivity analysis technique has shown to be effective in this case. This technique, therefore, may help physicians find disease markers.

At this point, two final remarks are worth to be discussed. First, the size of the training set is certainly small for training a network with such large number of parameters. So it may be thought that the validity of the obtained results is dubious. However, the results obtained on gels not seen during training are encouraging. Second, note also that the characteristic spots found by the sensitivity analysis may vary between runs (different network initialization and training session). Nevertheless, it has been observed that a group of differentiating spots tends to stay the same between different runs.

6 Conclusion

In this article, a neural network based classification approach of two-dimensional polyacrylamide gel electrophoresis has been presented, aiming at grouping 2D PAGE images and identifying their most significant features. The method has been successfully applied to the classification of real 2D PAGE images of rat cells. Not only does this technique correctly discriminate gels into different categories, but it also identifies typical spots corresponding to specific classes. Experimental results show that this technique is fairly effective and it becomes a potential diagnostic tool from 2D PAGE.

Acknowledgements

This work has been developed with the help of the Melanie Group of the Swiss Institute of Bioinformatics (SIB), the Central Clinal Chemistry Laboratory of the University Hospital of Geneva (HUG), the Computer Science Department (CUI) of the University of Geneva, Geneva Bioinformatics (GeneBio) and the University of Geneva (UNIGE). Special thanks go to Prof. Ron Appel of the SIB, Prof. Denis Hochstrasser of the HUG, and Prof. Christian Pellegrini of CUI.

References

[1] R. Appel, G. Bologna, and D. Hochstrasser. Classification tools for diagnostic rule formation from protein maps. In A. R. et al, editor, *MIE 93*, pages 40–44, Jerusalem, April 18-22 1993. Freund Publishing House, Ltd.

[2] M. Wilkins, K. Williams, R. Appel, and D. Hochstrasser, editors. *Proteome Research: New Frontiers in Functional Genomics*. Springer Verlag, 1997.

[3] B. Herber, J. Sanchez, and L. Bini. *Proteome Research: New Frontiers in Functional Genomics*, chapter Two-dimensional electrophoresis: the state of the art and future directions, pages 13–30. Springer Verlag, 1997.

[4] J. Vargas. *Two-Dimensional Gel Electrophoresis Computer Analysis System: From Image Acquisition to Protein Identification*. PhD thesis, Geneva University, 1996.

[5] D. H. R.D. Appel, A. Bairoch. A new generation of information retrieval tools for biologists: The example of the expasy www server. *Trends Biochem. Sci.*, (19):258–260, 1994.

[6] R. Appel, D. Hochstrasser, M. Funk, R. Vargas, C. Pellegrini, A. Muller, and J.-R. Scherrer. The melanie project - from a biopsy to automatic protein map interpretation by computer. *Electrophoresis*, (12):722–735, 1991.

[7] C. P. T. Pun, D. Hochstrasser. Correspondence analysis and hierarchical classification of complex images: Application to two-dimensional gel electrophoretograms. In N. M. . J. M. E. J.L. Lacoume, A. Chehikian, editor, *Signal Processing IV, Theories and Applications*, North Holland, 1988.

[8] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, Inc., second edition, 1999.

[9] R. W. D.E. Rumelhart, G.E. Hinton. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, chapter Learning Internal Representations by Error Propagation, pages 318–362. MIT Press, Cambridge, MA, 1986.

[10] T. Masters. *Practical Neural Network Recipes in C++*. Academic Press, 1993.

[11] I. C. A.P. Engelbrecht. Feature extraction from feedforward neural networks using sensitivity analysis. In *International Conference on Systems, Signals, Control, and Computers*, Durban, South Africa, 1998.

[12] V. K. F. Frost. Determining the influence of input parameters on bp neural network output error using sensitivity analysis. In *IEEE Proceedings of the Third International Conference on Computational Intelligence and Multimedia Application*, 1999.

[13] S. Hashem. Sensitivity analysis for feedforward artificial neural networks with differentiable activation functions. In *IEEE Proceedings of the 1992 International Joint Conference on Neural Networks (IJCNN'92)*, volume 1, pages 419–424, 1992.

[14] A. Cary. Simulador de redes neuronales para la clasificación de geles de electroforesis bidimensionales en el sistema melanie. Master's thesis, Universidad Cat/'olica Boliviana, Cochabamba, Bolivia, 2000.

[15] R. A. O. Cremona. Computer analysis of 2d page of vessel smooth mussel cells. Internal Report, 1993.