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
Universidad Católica Boliviana San Pablo
Cochabamba, Bolivia
e-mail: arielcary@hotmail.com

Abstract

This article describes a method, using neural networks, for classifying two-dimensional polyacrylamide gel electrophoretograms, complex biomedical images that contain proteins separated from a biological sample. 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 more than 90% effective.

Key words: Neural networks, proteomics, 2D-PAGE.

1 Introduction

Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) is a powerful biochemical technique for separating proteins contained in a biological sample [2, 15] and is widely used in proteomic research [10].

The proteins migrate on a polyacrylamide gel according to 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 (Figure 1). Typically, a 2D PAGE map is characterized by a thousand or even more protein spots. These protein maps can then be digitalized for further processing.

Due to the high complexity of these biomedical images, visual comparison and classification of 2D PAGE images is almost impossible. Consequently, advanced computer

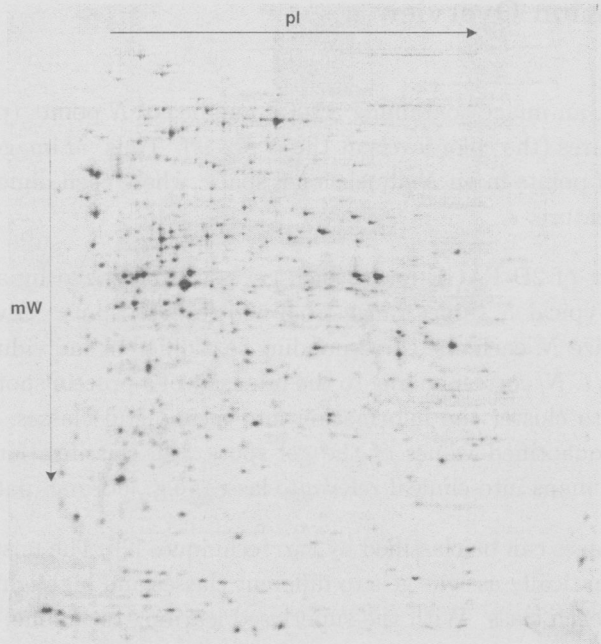

Figure 1: 2D-PAGE image of *escherichia coli* [1]. The spot position indicates the isoelectric point (pI) and the molecular weight (mW), while the spot darkness is proportional to the protein concentration.

systems must be developed in order to detect and quantify the spots contained in a 2D PAGE map [14], and new algorithms must be used for their classification.

In the medical diagnosis field, the classification of 2D PAGE maps is important to 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 possible. Therefore, the classification seeks to correctly assign a gel to a specific class.

In addition to the automatic classification, it is important to identify the characteristic spots within a class. That is, identify those spots that differentiate one class from the others. The characteristic spots may be indicators of specific diseases.

This article describes a method for classifying 2D PAGE images, using feedforward neural networks. Additionally, the sensitivity analysis technique was applied to the trained networks to identify characteristic spots. Finally, an experiment, using real gels confirmed the successful implementation of this method in the context of the MELANIE software for the analysis of 2D-PAGE gels, developed at the Swiss Institute of Bioinformatics (SIB) [3].

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 N values range from 500 to 2000 spots. 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 is how to cluster the information into meaningful classes. Consequently, by considering the quantified values of the gel spots, the classification process seeks to group 2D PAGE maps into clinical relevant classes (e.g., normal, pathological, etc.).

2D PAGE images can be classified by two techniques [2]. The unsupervised learning technique, automatically assigns gels to different classes and highlights the protein patterns specific to each class. With the supervised learning technique, classes are known in advance, for example, class A=disease and class B=normal. Both methods determine disease markers for subsequent classifications of additional gels.

After the classification, it is important to identify significant features, namely spots that differentiate classes. This may indicate which spots are related to 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. Also, feature values of interest such as the spot area, volume, and optical density may be calculated.

To correct possible distortions of the images and make them superimposable, a pixelwise correction (aligning) has to be applied to the gels according to a reference gel, after spot detection and quantization. This reference gel is chosen among all the gels according to image quality.

After the alignment, all images can be matched to identify spot pairs. Thus, spots corresponding to the same protein in various gels are identified and a matched spot list results. Each list is referred to as a spot group and a total of N spot groups result from the matching process (Figure 2).

Techniques for spot detection, quantification, image aligning, and matching have been proposed and successfully tested [14].

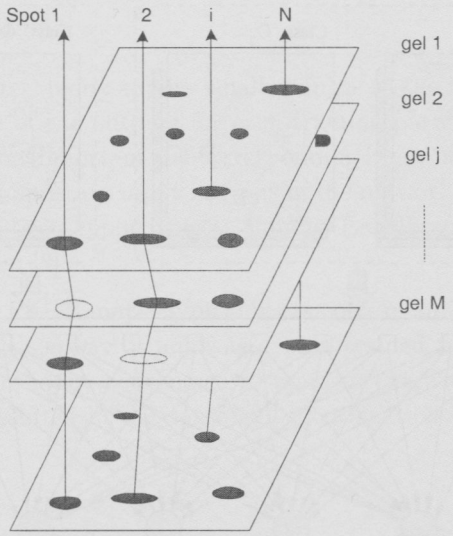

Figure 2: Building the spot groups [12]. Spots representing the same protein in different images are identified by matching the gels.

4 Two-D PAGE Image Classification Using Feedforward Neural Networks

Feedforward neural networks [9], using the backpropagation algorithm [13] as the learning rule (supervised learning), can be trained to classify 2D PAGE images because 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 learns the subjacent concept of each class through the samples contained in each class. Then, the knowledge acquired by the network can be used to classify new gels not previously presented.

This task is achieved in two phases. In the training phase, the network learns to classify the gels, using a portion of the M gels as the training set. Then, in the testing phase, the performance of the network is evaluated to determine its ability to generalize the knowledge acquired during training. This is achieved by classifying a new set of gels chosen from the remaining M gels (Figure 3).

Once the network reaches an acceptable performance, it can be used to classify additional 2D PAGE images without further training or testing. Additionally, the sensitivity analysis technique can be used to identify the most important image features.

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 number

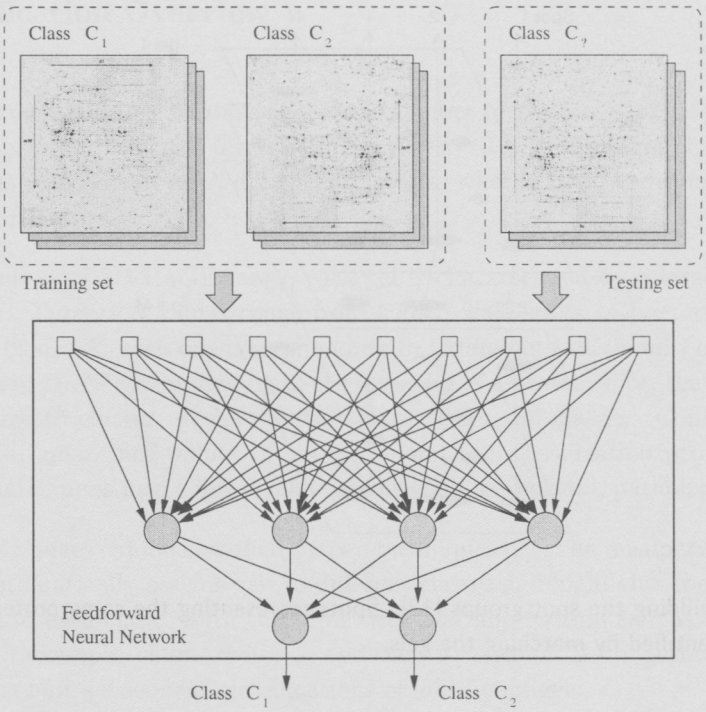

Figure 3: Classification of 2D PAGE images using neural networks. The network learns the concept of each class through the samples contained in the training set. Then, images from the testing set are evaluated using the knowledge acquired during training.

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 or two hidden layers can be considered with an arbitrary number of neurons in each layer.

4.2 Training, Testing, and Classification

The training set consists of a percentage p of the M gels. The training is then carried out, using the backpropagation learning algorithm [13] so that the corresponding output neuron of the gel's class being fed at the input layer is highly active (on), while the rest of the output neurons are off (near zero). The rest of the gels $(1-p) \times M$ are used for testing the overall performance of the network after the training has been completed. Afterwards, following the same procedure, completely new gels, with unknown association classes, can be classified.

If only one output neuron becomes active a match with a single class has been made. However, if two or more output neurons are activate, the gel shares characteristics of various classes). Finally if all neurons are inactive, the gel does not match any known class within learned categories. Thus, an early diagnostic may only be possible if a class or classes have been established.

4.3 Sensitivity Analysis

The sensitivity analysis technique applies small changes at the inputs and observes the corresponding variations of the outputs for each training pattern [11]. The larger the output variation, the more important the corresponding input neuron. Thus, it is possible to identify the most important input neurons of the network. This simple technique is quite effective and has been applied successfully by Engelbrecht and Cloete [6], Frost and Karri [7] and Hashem [8].

In the context of 2D PAGE maps, typical spots corresponding to particular classes can be determined using sensitivity analysis over a trained network. It is expected that the most influencing input neurons correspond to characteristic spots. Application of this technique resulted in the successful identification of several features that discriminate classes.

5 Experiment

An experiment, using real gels of rat vessel smooth muscle cells, was used to evaluate the proposed neural network in a supervised classification task [4].

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 [5]. The samples came from populations of newborn (*NN*) and two-year old rats (*VE*). The goal of this experiment was to train a feedforward neural network to discriminate these two gel categories and to identify protein spots that differentiate 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, were analyzed and used to build the training set. The network was then trained.

In the testing phase, the remaining 10 gels were also analyzed and used to verify the results of the training phase.

5.2 Image Analysis

Image analysis (spot detection and quantification, image aligning and matching) was done using the MELANIE software package for 2D PAGE image analysis [3]. In the first phase, the protein spots of 20 2D PAGE images were automatically detected and quantified. Between 519 and 1509 spots were detected on the gels. The optical density value of each spot was computed. This value is directly related to the concentration of the protein. In the second phase, a reference gel was selected from the set of images and each of the remaining gels was then corrected pixelwise so that all gels became superimposable. In the third phase, the gels were automatically pairwise matched; i.e.,

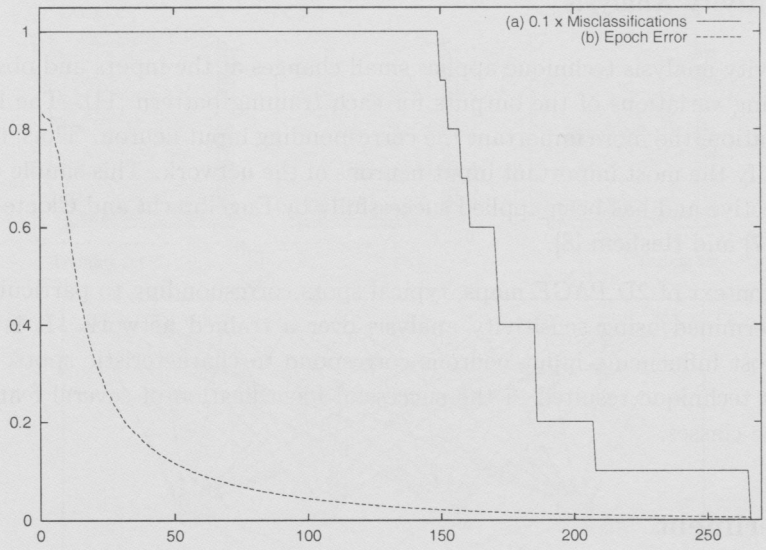

Figure 4: Neural network learning curves. (a) The misclassification curve: at the beginning of the training, all gels are misclassified. The training phase is completed when all gels are correctly classified. (b) The epoch error curve shows how the global error of the network decreases as the training progresses.

correspondences between the spots in different gels were identified. 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 ten lines (one per gel) and 867 columns (one per spot). The two matrices contain the spot values representing the optical density.

5.3 Architecture

In this experiment, the size of the input layer is 867 neurons, one input neuron for each feature spot. The number of output neurons is two since there are two different categories. Two hidden neurons haven been used in one hidden layer. Thus, 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 needed (Figure 4).

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

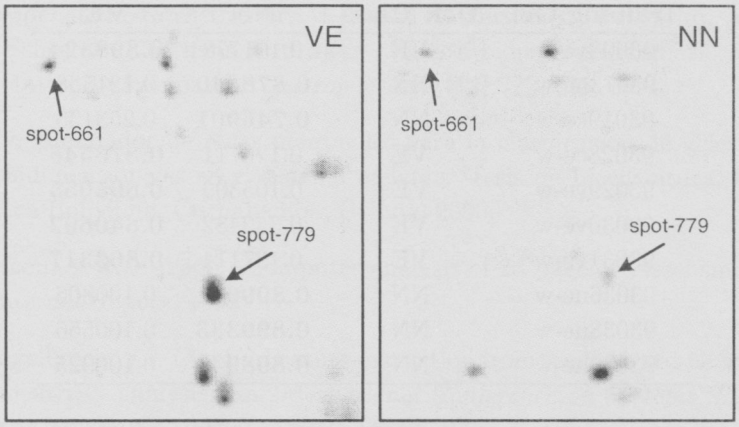

Figure 5: Characteristic spots found by sensitivity analysis. In old rats, the protein spots 661 and 779 are highlighted in 2D PAGE(*VE*).

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. Ten protein spots were identified as the most differentiating features between the classes. Some discriminating spots for two different gels (different classes) are shown in Figure 5.

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, not seen previously, were correctly classified.

Also ten characteristic spots were identified. It was then verified that these protein spots alone determine the class of each gel. The protein concentration may be higher in one class than in the other, and, in some cases, a protein spot is found only in one class. The sensitivity analysis has shown to be effective in these cases. Thus, this technique may help physicians identify disease markers.

It is important to note that the size of the training set was small for training a network with such large number of parameters. Thus the validity of the results may be questioned. However, the recognition of gels not seen during training is a good indicator of the validity of the neural network approach. Nevertheless, the characteristic spots found by the sensitivity analysis varied between runs (different network initialization and training affected the outcome). However, it has been observed that a group of differentiating spots tend to remain constant.

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

Table 1: Actual network outputs. Class association of each gel is presented in boldface (highest output value).

6 Conclusion

A neural network based classification approach of two-dimensional polyacrylamide gel electrophoresis was developed. The objective was to group 2D PAGE images and identifying their most significant features. The method was successfully applied to the classification of real 2D PAGE images of rat cells.

Not only did this technique correctly discriminate gels into different categories, but it also identified typical spots corresponding to specific classes. Our results indicated that this technique may be effective as a diagnostic tool.

Acknowledgements

This study was supported by the Melanie Group of the Swiss Institute of Bioinformatics (SIB), the Central Clinical 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 to Prof. Ron Appel of the SIB, Prof. Denis Hochstrasser of the HUG, and Prof. Christian Pellegrini of CUI.

References

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

[2] R.D. Appel, G. Bologna, y D.F. Hochstrasser. Classification tools for diagnostic rule formation from protein maps. En A. Reichert et al, editor, *MIE 93*, pp 40–44, Jerusalem, April 18-22, 1993. Freund Publishing House, Ltd.

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

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

[5] O. Cremona y R.D. Appel. Computer analysis of 2d page of vessel smooth mussel cells. Internal Report, 1993.

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

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

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

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

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

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

[12] T. Pun, D. Hochstrasser, y C. Pellegrini. Correspondence analysis and hierarchical classification of complex images: Application to two-dimensional gel electrophoretograms. En J.L. Lacoume, A. Chehikian, N. Martin, y J. Malbos (Eds.), editores, *Signal Processing IV, Theories and Applications*, North Holland, 1988.

[13] D.E. Rumelhart, G.E. Hinton, y R.J. Williams. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cap. Learning Internal Representations by Error Propagation, pp 318–362. MIT Press, Cambridge, MA, 1986.

[14] J.R. Vargas. *Two-Dimensional Gel Electrophoresis Computer Analysis System: From Image Acquisition to Protein Identification.* Tesis Doctoral, Geneva University, 1996.

[15] M.R. Wilkins, K.L. Williams, R.D. Appel, y D.F. Hochstrasser, editores. *Proteome Research: New Frontiers in Functional Genomics.* Springer Verlag, 1997.