Applying Artificial Neural Networks to the Classification of Wheat Varieties Processed Via MALDI-TOF Mass Spectrometry

DAVID SCHIBECI[1], ROB POTTER[2], KEVIN WATHEN-DUNN[2], MIKE JONES[2] and MATTHEW BELLGARD[1]

[1] Centre for Bioinformatics and Biological Computing
School of Information Technology
[2] Western Australian State Agricultural Biotechnology Centre

Murdoch University
Murdoch WA 6150
AUSTRALIA

Abstract: - With the advent of new bio-technologies there is a significant increase in both the variety and amount of data that needs to be analysed in order to extract biological meaning. This is evident from the masses of molecular data being produced as a result of the numerous genome sequencing projects as well as from data now produced from DNA microarray/chip technologies. In this paper, we describe analysis of data obtain from recent advances in time-of-flight mass spectrometry which is a new method for rapid, high-resolution separation of protein mixtures. We employ a feedforward artificial neural network to train the resultant protein profiles processed from a number of varieties of wheat in order to classify them. The results of this study are extremely positive with results ranging from 87.5% to 100% accuracy, and we discuss them in context with a number of challenging problems for further studies.

Key-Words: - MALDI-TOF Mass Spectrometry, wheat variety classification

1 Introduction

Handling and marketing of wheat is moving from a bulk quality classification system to one that is based on specific varieties with the introduction of end-point levies (for collection of breeding royalties) and requirements for variety segregation based on specific grain properties. Grain protein electrophoresis, the method that is now used for variety identification, is relatively slow and labour intensive. In addition, it is now becoming increasingly difficult using this method to distinguish new varieties because of the similarity of the parental lines from which they are being developed. DNA-based identification is the most rigorous approach to identify varieties, because DNA-based molecular markers (eg microsatellites) are now being used to breed current elite cultivars, but a DNA-based approach is still too expensive to be undertaken routinely. Recent advances in time-of-flight mass spectrometry offer an alternative method for rapid, high-resolution separation of protein mixtures, which can be used for variety identification. Coupled with matrix assisted laser desorption/ionisation, this methodology (MALDI-TOF ms) can now be used to analyse complex mixtures of macromolecules such as proteins, which could not previously be analysed.

MALDI-TOFms uses pulses of a high powered laser to vapourise a crystallised matrix in which fragile macro-molecules (proteins and DNA) have been adsorbed. Desorption and ionisation of the macromolecules thus takes place relatively gently and the intact molecules are then accelerated using a high voltage and allowed to drift down a flight tube. The time taken to reach the detector at the end of the flight tube is directly related to the mass/charge ratio of the molecule. The latest equipment is capable of single Dalton resolution in a range up to 50,000 Da and sample data can be collected in as little as 5 seconds. There is potential using this approach to increase the resolution of grain protein analysis as well as the speed and throughput of analysis. However, the complexity of the spectra (protein profiles) obtained make it difficult for the operator to confidently assign variety based on visual inspection of spectra alone. Thus, the motivation of this work is to develop methods to analyse these complex protein profiles. The major issues to be addressed include the identification of

specific wheat varieties and quantifying the proportions of mixtures of varieties of wheat. The application of artificial neural networks (ANNs) to analyse these protein profiles is needed in order to develop an automated, classification system for rapid, high-throughput analysis. If successful, this approach would have the added benefit of increasing throughput and reducing costs by reducing the labour input in a practical agricultural production setting.

ANNs have already been applied successfully to various pattern classification problems in biotechnology and bioinformatics [1, 2, 3, 4]. To date, there has only been one other recent application of ANNs to analyse MALDI-TOFms potein profile data for variety identification [2]. The present study takes this type of analysis further as the aim is to use this approach in a high-throughput analytical system. As a result, it is important to evaluate this type of approach with large data samples, to test a number of ANN parameters (input representation, numbers of hidden units and robust evaluation strategies), and to develop a database to track improvements to the ANN and classification success rates over time for accountability. In this paper, we describe the application of ANN to MALDI-TOFms data for wheat variety identification, addressing these important issues. As is described, the results obtained are very encouraging, with successful classification of wheat varieties ranging from 87.5% to 100%. In addition, the application has been made available online (http://cbbc.murdoch.edu.au/software/nn/) where it is possible to upload a protein profile of a wheat variety and have it classified automatically. The future directions of this work are also discussed.

2 Problem Formulation

2.1 MALDI-TOF and Data Extraction

Twenty single grains of five varieties of wheat from two different seasons were ground in a mortar and pestle and transferred to an Eppendorf tube. Protein was extracted by mixing the flour with 100µl of 70% ethanol for 20 minutes and 1µl samples were removed and mixed with 10µl of sinapic acid matrix solution (10mg/ml sinapinic acid in acetonitrile/TFA (300/700 µL)). One microlitre of each sample/matrix mixture was spotted onto a 400 well MALDI plate, air dried and then placed in an Applied Biosystems Voyager DE Pro MADLI-TOF mass spectrometer. Spectra were collected automatically with the following conditions: 25KV acceleration voltage, 0.2% guide wire voltage, 92% plate voltage and 750ns delay time. Collection was from 14,000Da to 45,000Da with a 2,000Da low mass gate. Laser power was set to 2300 max and 1900 minimum (step 50) with a 5-pulse pre-scan to determine optimum level. Twenty spectra of 25 laser pulses were collected from each sample in a random pattern over the sample spot and these were accumulated to provide the final spectrum. Spectra were exported as ascii files with MW and intensity data as x and y values.

The five varieties of wheat used in this study were Camm (CAM), Carnamah (CARN), Spear (SPEAR), Stiletto (STIL) and Westonia (WEST).

2.2 ANNs, Data Normalising, Jack-knifing, Web Site for Testing

The ANN used in this investigation was a standard feed-forward, back propagation trained network [5, 6]. Initially, the data points of the two hundred samples each containing 14,222 sample points (x_i) were standardised using the standard statistical formula:

$$x_i^* = (x_i - x_{mean})s_x. \quad (1)$$

Once standardised, it was possible to visually compare these graphs. Using simple linear statistical measures we determined that it was not possible to identify all strains easily and that the problem required a non-linear solution. This was supported by visual inspection of a number of profiles.

In order to transform the problem suitable for an ANN, it was necessary to reduce the input dimension from 14,222 down to a reasonable size of 250 sample points. To do this, 56 data points were averaged at a time. Visual comparison between the original graphs vs graphs of the sampled data was done to ensure that essential features in the profiles were not lost. We discuss this issue later. Simple statistical methods were again unable to resolve the varieties. Typical protein profiles of two different wheat varieties are shown in Figure 1 (included at the end of this paper). The sample data contained maxima greater than one, so each point was divided by the maxima for that sample to ensure all values were between zero and one. Jack-knifing is given *n* sample, using *(n-1)* samples to create the predictive model and retain one to gauge the accuracy of the prediction. Though it is cpu/resource intensive it is easy to automate.

3 Problem Solution

3.1 Classification of Five Varieties - Performance

As a first pass, 80% of the data was used (160 samples – 32 for each variety) to train the ANN and the remaining data (40 samples – 8 for each variety) to test it. (The parameters used were: input units values ranged from 0 to 1, binary outputs units (0 or 1), one thousand training iterations and five hidden units). These results were extremely encouraging, as shown in Table 1. Even from this initial experiment, the ANN was able perform the classification task with a high degree of accuracy. The ANN seems to have no difficulty in classifying the wheat varieties Camm, Carnamah or Westonia correctly, but does have difficulty in classifying Spear and Stiletto

Table 1 First pass analysis of using an ANN to classify wheat varieties, with very positive results.

	% of samples correctly classified
CAM	100 %
CARN	100 %
SPEAR	93.75 %
STIL	81.25 %
WEST	100 %

Following this, a number of ANN experiments were conducted varying a number of the parameters and employing jack-knifing for testing purposes. For each of these experiments the standardised, sampled version of the wheat samples was used as a starting point into the ANN. The number of hidden layers was varied from five to ten, and the number of iterations from two thousand to ten thousand. Changing the output values from binary to bi-polar (-1, 1) was also investigated. As can be seen from Table 2 (at the end of this paper), none of the variations had a significant effect on the performance of the classification. The largest difference in classification performance between any of the variations was 5%.

These results were analysed further to show misclassification as well as successful classifications (Table 3) and it can be seen that the neural network still has difficulty in distinguishing between Spear and the Stiletto. Visual analysis of the spectra confirms that these two varieties are the hardest to distinguish.

Table 3 Two specific varieties of wheat are the main cause for the ANN to make misclassifications.

	% classified as SPEAR by ANN	*% classified as STIL by ANN*
SPEAR	93.75 %	6.25 %
STIL	18.75 %	81.25 %

3.3 MSE Curve

For each iteration of the algorithm, MBP [6] provides five pieces of information for each iteration: iteration number; gradient norm; error; maximum absolute error; and digital error. Using the maximum absolute error the performance of the learning process is plotted (Figure 2). After eight hundred iterations the algorithm improves steadily, however there is always danger of overtraining.

4 Conclusion

Although the results obtained so far are very promising, there is still additional work to be done. A neural network is only as good as the training data. The training data used in this work was of a very high, reproducible qulaity. These were protein profiles from single seeds from pure stocks grown at the same site. Of interest would be whether variation between sites would make the analysis more difficult and is the subject of current studies. It would also be of interest to examine the effect of using different kinds of samples: different methods of grinding bulk samples (for a low labour, high throughput extraction system); mixtures of varieties; and to test additional wheat varieties - both as certified samples from breeders/grain handlers and also samples grown in different areas. The aim is to establish a large database of protein profiles from authenticated wheat samples with which to train the NN, to produce some confidence levels for field samples, so that this method can be used as a routine test to identify the identity of wheat varieties and mixtures of varieties.

As more data is being processed, other investigations are being conducted such as evaluating the level of compression of data before classification degrades substantially. Currently, every 56^{th} samples points are averaged, as there is little biological knowledge of what each peak specifically represents.

From the experiments above, having 5 outputs meant that it was possible to determine if the network was confident in its classification. In the case of the miclassifications of spear and stil varieties, the output of the ANN provided evidence

of this with both corresponding output units have a higher value than the other 3 output units (representing the other 3 varieties). It will be important to resolve this classification accuracy for varieties like these, in order to conduct more complex problems. For example, determining variety mixtures, differences between yields of the same variety.

References:

[1] Y Xu et. al. Recognising exons in genomic sequence using grail II. *Genetic Engineering* 16, 1994, 241-253.

[2] HA Bloch, C Kesmir, M Peterson, S Jacobsen and I Søndergaard, Identification of Wheat Varieties Using Matrix-assisted Laser Desorption/Ionisation Time-of-flight Mass Spectrometry and an Artificial Neural Network. *Rapid Communications In Mass Spectrometry* **13**: 1999, 1535-1539.

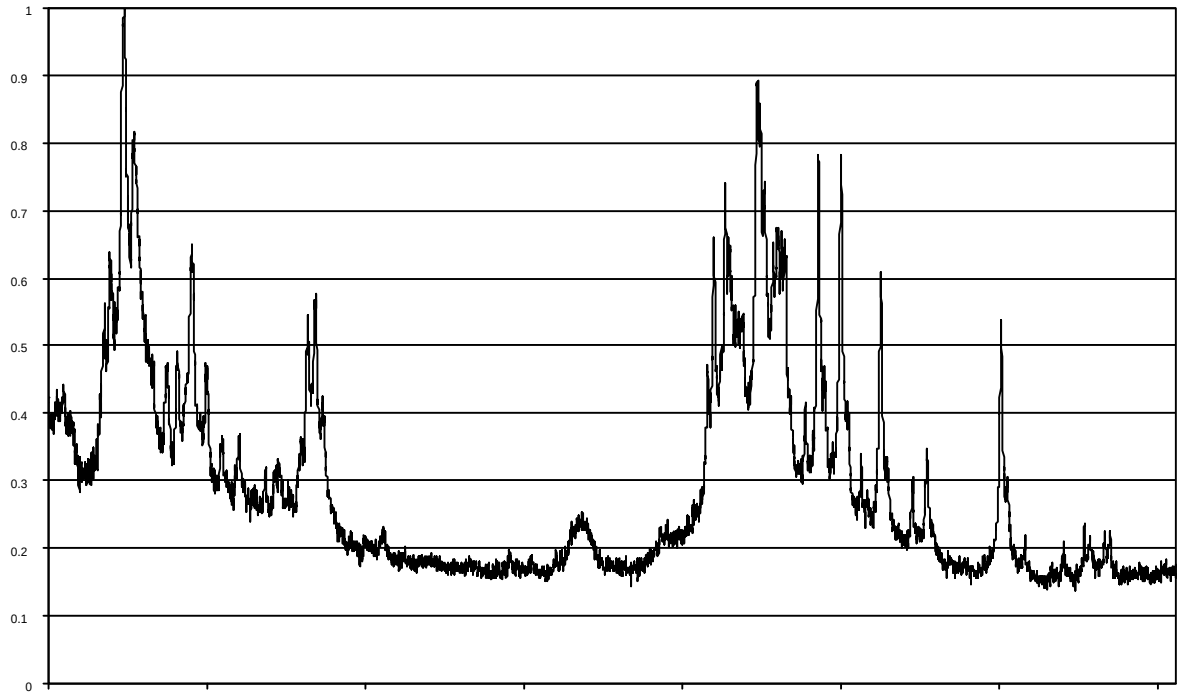
[3] MI Bellgard et. al. MHC Haplotype Analysis by Artifical Neural Networks, *Human Immunology* 59, 1998, 56-62.

[4] K Jensen., C Kesmir and I Søndergaard, From image processing to classification: IV. Classification of electrophoretic patterns by neural networks and statistical methods enable quality assessment of wheat varieties for breadmaking, *Electrophoresis*, Apr;17(4), 1996, 694-698.

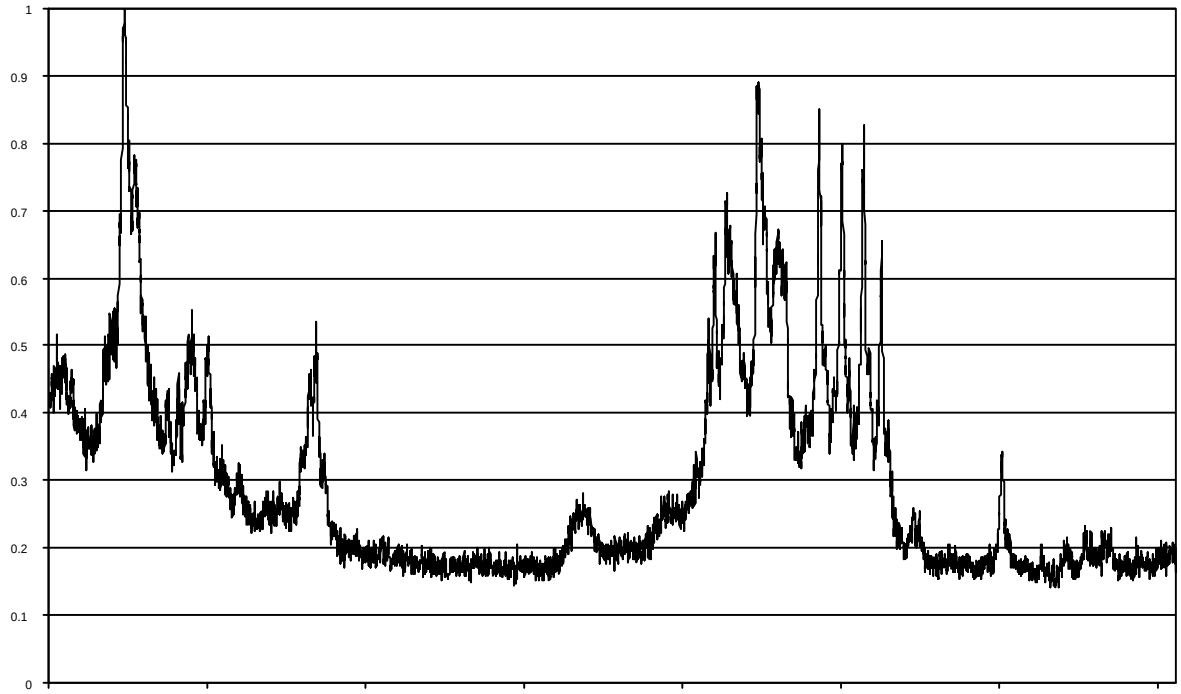
[5] MH Hassoun, Adaptive multilayer Neural Networks I. In MH Hassoun (Ed): *Fundamentals of artificial neural networks*. Cambridge, Bradford, 1995, 197-276.

[6] D Anguita, MBP (Matrix Back Propagation), ftp.esng.dibe.unige.it as /neural/MBP/MBPv1.1.tar.Z, 1993.

Fig. 1 – Typical protein profiles of particular wheat samples (a) Stiletto, (b) Spear produced by the MALDI-TOF Mass Spectrometry



(a) Stiletto



(b) Spear

Fig.2 - Mean Squared Error Plot for a training set obtain from Jack-knifing over 2000 iterations.

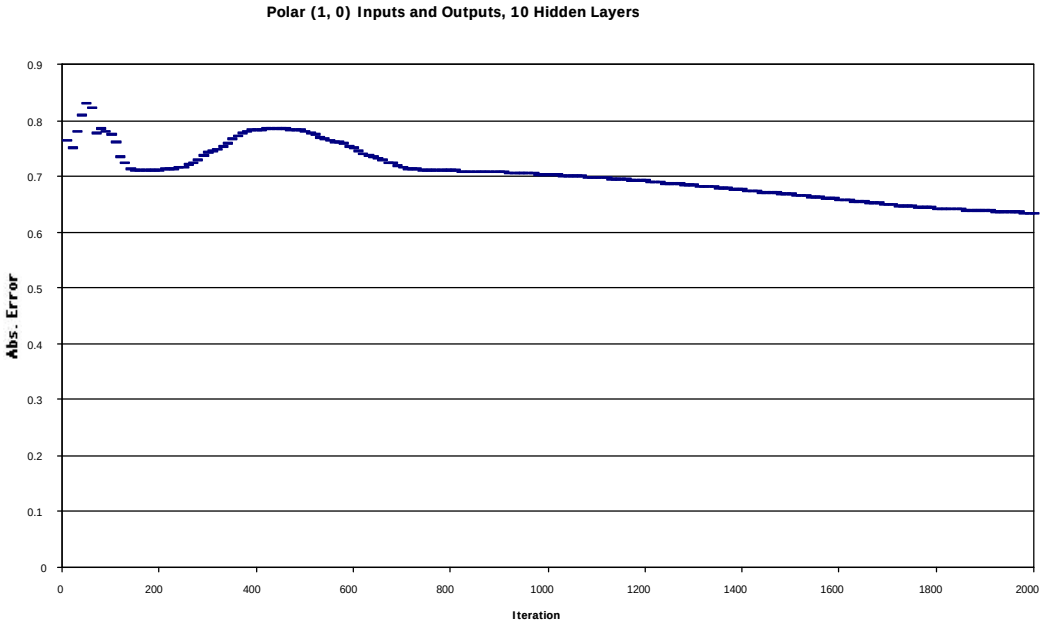


Table 2 – Effects on classification performance using different ANN parameter settings

No. of Hidden Units	Output Representation	Iterations	CAM	CARN	SPEAR	STIL	WEST
5	Binary	2,000	97.5 %	100 %	90 %	92.5 %	100 %
10	Binary	2,000	100 %	97.5 %	85 %	92.5 %	100 %
5	Bi-Polar	2,000	100 %	100 %	82.5 %	82.5 %	100 %
10	Bi-Polar	2,000	100 %	100 %	87.5 %	87.5 %	100 %
10	Binary	10, 000	100 %	100 %	90 %	90 %	100 %