ELSEVIER

Expert Systems with Applications 30 (2006) 24–33

Expert Systems with Applications

www.elsevier.com/locate/eswa

Automated trend analysis of proteomics data using an intelligent data mining architecture

James Malone *, Ken McGarry, Chris Bowerman

Centre for Hybrid Intelligent Systems, School of Computing and Technology, University of Sunderland, StPeter's Way, Sunderland, SR6 0DD, UK

Abstract

Proteomics is a field dedicated to the analysis and identification of proteins within an organism. Within proteomics, two-dimensional electrophoresis (2-DE) is currently unrivalled as a technique to separate and analyse proteins from tissue samples. The analysis of post-experimental data produced from this technique has been identified as an important step within this overall process. Some of the long-term aims of this analysis are to identify targets for drug discovery and proteins associated with specific organism states. The large quantities of high-dimensional data produced from such experimentation requires expertise to analyse, which results in a processing bottleneck, limiting the potential of this approach. We present an intelligent data mining architecture that incorporates both data-driven and goal-driven strategies and is able to accommodate the spatial and temporal elements of the dataset under analysis. The architecture is able to automatically classify interesting proteins with a low number of false positives and false negatives. Using a data mining technique to detect variance within the data before classification offers performance advantages over other statistical variance techniques in the order of between 16 and 46%.

Keywords: Data mining; Differential ratios; Protein trend analysis; Neural network; Proteomics; Bioinformatics; Two-dimensional electrophoresis

1. Introduction

Following the explosive growth in research into the genome, the study of the proteome has become fundamental to biochemical research (Righetti, Stoyanov and Zhukov, 2001). Proteomics is defined as the large-scale identification and characterisation of the proteins encoded in an organism's genome (Alberts, Bray, Lewis, Raff, Roberts and Watson, 2002) and is often described in literature as the next step to dramatically advance drug discovery (Whittaker, 2003). More specifically, proteomics is concerned with the analysis of the structure and function of proteins as well as of protein-protein interactions.

Within proteomics, a particular area of interest is the mapping of protein posttranslation modifications (Liebler, 2002). RNA, which is initially transcribed from the genetic details stored in DNA, is translated to protein. Following this translation, the state of a protein can alter during its lifetime, such as from the introduction of a disease (Crenshaw and Cory, 2002). The protein's state within a particular tissue can alter as conditions change and, hence, is indicative of the current physiological state. These posttranslational modifications have a direct effect on the structure, function and turnover of proteins, hence, analysis of these trends of variation may lead to novel avenues to determine how chemical modifications to the proteome affect living systems (Liebler, 2002). Consequently, the analysis of the posttranslational modifications of proteins is particularly important for the study of conditions such as cancer, neurodegenerative diseases, heart disease and diabetes.

In order to perform this analysis, a method of measuring the expression of proteins is required. The most popular, and currently unrivalled, technique to perform protein expression analysis is that of two-dimensional electrophoresis (2-DE) (Jenkins and Pennington, 2001; Pennington, Wilkins, Hochstrasser and Dunn, 1997). This technique uses two successive electrophoresis runs to separate the proteins from a tissue sample with regards to their isoelectric point and molecular weight. The first run separates the proteins in one dimension and the gel is then rotated 90° and the second run is performed to separate into the second dimension. Each protein expressed using this method appears as a dark spot on these gels (see Fig. 1), following the use of staining techniques, and are then individually analysed for features such as relative abundance, shape and appearance and disappearance across an experimental series (such as over time or between different control groups). Such analysis is often assisted with the use of image

* Corresponding author. Tel.: +44 191 515 3268; fax: +44 191 515 3461
E-mail address: james.malone@sunderland.ac.uk (J. Malone).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.09.047

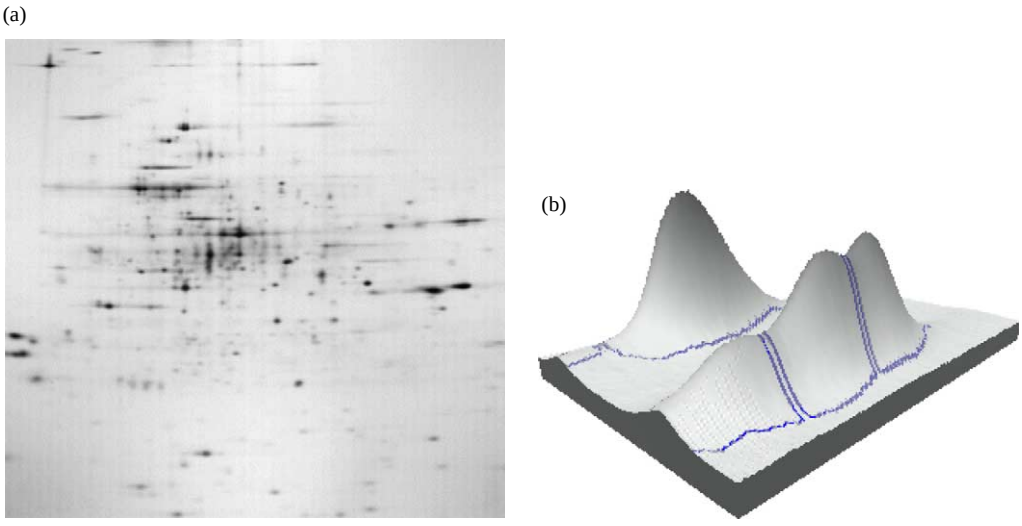

Fig. 1. (a) A single 2-DE Gel image. Each protein is described by a black spot following staining. (b) Individual spots visualised in 3-D using image analysis software. This image produced by Nonlinear's Progenesis software (Marengo et al., 2005)

analysis software which can automatically detect spot correspondence from one gel to the next (Pederson and Ersboll, 2001; Pleissner, Oswald and Wegner, 2001). Following this process, these images can be converted into data which describes each protein, such as volume, area, height and x and y coordinates on a gel. These attributes can be representative of changes to the function of the protein; changes to these attributes can be indicative of an intrinsic link to a particular condition. For example, a protein which has physically altered under a diseased state compared with that of a healthy state may well be intrinsically linked to the physiological state of the organism and, hence, worthy of further investigation.

The analysis of this protein data, however, is not a trivial task (Marengo, Leardi, Robotti, Righetti, Antonucci and Cecconi, 2003). Disadvantages of 2-DE include that it is inherently labour-intensive and requires a skill-level such that only trained experts can perform the analysis, often manually. The potentially useful trends are encapsulated within large volumes of multi-dimensional, spatio-temporal post-experimental data, making this manual interpretation of results impractical (Fenyo and Beavis, 2002). Without the availability of reliable tools for post-experimental data analysis, the technique is essentially a descriptive one, limiting the potential for fully automated analysis (Griffin and Aebersold, 2001). The full value of this technique can not then, be realised until this processing bottleneck is resolved; fully automatic approaches for identifying intrinsic trends in gels will go some way towards this goal (Dowsey, Dunn and Yang, 2003).

In this paper, we present an intelligent data mining architecture that is able to analyse post-experimental, 2-DE gel data and identify interesting proteins automatically. This approach uses a combination of a data-driven, data mining technique and a goal-driven, machine learning technique which incorporates expert heuristics, such as those used in manual analysis. Data mining is the process of finding trends and patterns in large data sets (Toroslu and Yetisgen-Yildiz, 2005). The data-mining element employed here is that of differential ratio (dFr) data mining, a technique which measures variance of a given object in terms of the log of pair-wise ratios of the elements describing the data over time (or within any given linear series). The machine-learning element concerns the use of a BackPropagation, Multi-Layer Perceptron (MLP) neural network in order to classify the results of the data mining into discrete classes of interesting behaviour. Such classes are defined using expert heuristics, optimised through the use of an Adaptive Nero-Fuzzy Inference System (ANFIS) as described by Malone et al. (2004b). A comparison is drawn to MLPs trained using Principal Component Analysis (PCA) and Covariance as variance measures. Finally, a comparison to a MLP trained on normalised data alone is conducted to quantify any relative benefits of using a variance analysis measure step before classification of the dataset.

The remainder of this paper is organised as follows. Section 2 discusses current strategies used in the analysis of 2-DE gel data. Section 3 describes the proposed intelligent data mining architecture. Section 4 presents the results of experimentation and discusses these findings. Section 5 outlines the conclusions.

2. Analysis of 2-DE gel data

Studies performing trend analysis have employed techniques including Principal Component Analysis (PCA) (Marengo, Robotti, Righetti, Campostrini, Pascali and Ponzoni, 2004; Picard, Bourgoin-Greneche and Zivy, 1997; Sekiguchi et al., 2002) and Correspondence Analysis (CA) (Krah, Wessel and Pleißner, 2004; Pleissner, Regitz-Zagrosek, Krudewagen, Trenkner, Hocher and Fleck, 1998; Rooney-Varga, Giewat, Savin, Sood, Legresley and Martin, 2005;). PCA is a technique used to reduce the dimensionality of data to summarise the most important (i.e. defining) parts whilst simultaneously reducing noise. Although widely used within 2-DE gel analysis, the technique has the disadvantage of

assuming that data objects are linearly (or at least monotonically) related to each other, and to gradients, which may or may not be true in experiments. CA is related to PCA and is a descriptive technique which analyses two-way and multi-way tables for some measure of correspondence between the rows and columns and can analyse nonlinear data.

The use of three-way PCA has also been applied to proteomic pattern identification. Three-way PCA is a multivariate technique that takes into account the three-way structure of the dataset and is based on the premise that the observed modes can be reduced to more fundamental modes. Marengo et al. (2003) successfully used this method to identify regions within a 2-DE gel that were responsible for differences occurring between sample groups. In this instance, two datasets were tested; control rat serum group and nicotine treated rat serum group; healthy human lymph-nodes group and human lymph-nodes affected by a non-Hodgskin's lymphoma group. Since this approach identified clusters of difference between sample groups using this dimensionality reduction technique, a disadvantage of this approach might be that possibly important individual protein spots which are geographically isolated from the clusters may not be identified as salient to the disease state. Arguably, a better technique would analyse each protein individually without any reduction of specific variable values and, whilst considering geographical elements, would not discriminate against proteins, which are isolated from clusters.

Vohradsky's study (1997) looked into the use of artificial neural networks (ANN) to classify spot profiles. In this work, the author discovers that ANNs outperform both cluster analysis and PCA following comparative analysis. Whilst PCA and clustering could not correctly identify both the up regulation (T+) and down regulation (T−) of proteins, the neural network solution was able to identify almost all correctly. Furthermore, the architecture consisted of a relatively small number of units; 16 input units, 30 units in the first hidden layer, 3 in the second hidden layer and 1 output unit. Experiments contaminated with Gaussian noise still performed well, illustrating a level of tolerance to noise by the neural network. The possibility of implementing a combinatorial approach is also suggested in order to increase accuracy and reliability of results. Such hybrid approaches are usually implemented when the advantages of several techniques are required to fully solve a problem. Since the analysis of 2-DE gel data is such a complex and non-trivial task, the use of a hybrid data mining architecture would appear to be an appropriate choice to tackle this important procedure. Furthermore, the use of neural networks within biomedical data mining has seen a great deal of interest (Chou et al., 2004; Delen, Walker, and Kadam, 2005; Mendyk and Jachowicz, 2005; Wiemer and Prokudin, 2004).

3. An intelligent proteomics data mining architecture

Since analysis is usually performed manually by trained specialists, our proposed system incorporates expert judgement, whilst offering a level of automation, also important to decrease the amount of time spent during this analysis phase. The architecture suggested in this paper, therefore, uses a combination of goal-driven (expert heuristics) and data-driven (data mining) elements to perform the proteomics data analysis and is illustrated in Fig. 2.

In the first stage (the first box), the post-experimental, 2-DE gel data is collected from image analysis software and is normalised. In the second stage, spatio-temporal data mining is performed on the normalised proteomics data. This entirely data driven element of the knowledge discovery process

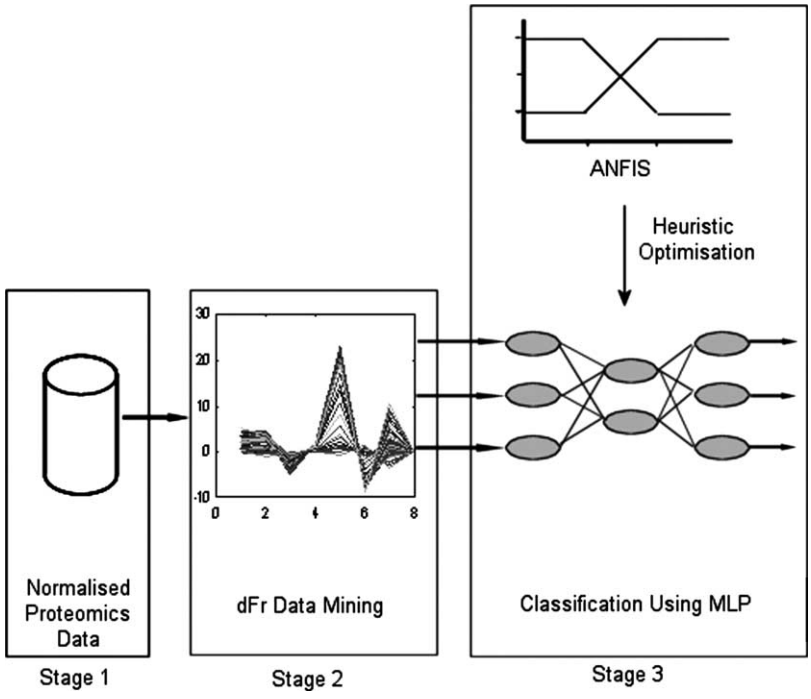

Fig. 2. The proposed 3-stage intelligent data mining architecture.

automatically analyses the data for trends of variance. The result of this process provides the training data for the third stage which employs a BackPropagation, Multi-Layer Perceptron (MLP) neural network as a classifier. The outputs from this classifying stage are classes of 'interesting' protein behaviour, derived from expert's opinions.

3.1. Differential ratio data mining

The technique we propose to perform the spatio-temporal analysis is that of differential ratio (dFr) data mining (Malone, McGarry and Bowerman, 2004a). This technique draws on elements of covariance measures and ratio rules (Korn et al., 2000). Covariance measures the linear dependencies between variables within a given series (Hand, Mannila and Smyth, 2001) and has previously been used within biomedical data analysis (Damcott et al., 2004). Given two variables, X and Y and n observations; X taking values $x(1)$, …, $x(n)$ and Y taking values $y(1)$,…, $y(n)$ the sample covariance between X and Y is defined as;

$$\sigma_{XY} = \mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y}) \tag{1}$$

Where:

$\bar{x}$ sample mean of X values
$\bar{y}$ sample mean of Y values

Data mining using ratio rules is a technique that employs eigensystem analysis to calculate correlations between values of attributes. Ratio rules can tackle the issue of reconstructing missing/hidden values and can be used to perform 'what-if' type scenarios given an antecedent(s) or consequent(s). Although this technique is useful for predicting attribute trends within empirical data, the process does not incorporate either spatial or temporal elements and would therefore have limited applicability to the analysis of datasets such as those created from 2-DE gel experiments.

Differential ratio data mining is used to measure the variance of a given object in terms of the log of pair-wise ratios of the elements describing the data over time (or within any given linear series consisting of two or more elements). Consider two variables x and y as elements of a given object. The calculation of a single differential ratio (herein, differential ratio, or dFr, will be referred to as the measure of difference calculated by this process) between two time points, t and $t+1$ is given by;

$$\mathrm{d}Fr_t = \log\left\{\frac{\left(\frac{x_t}{y_t}\right)}{\left(\frac{x_{t+1}}{y_{t+1}}\right)}\right\} \tag{2}$$

Where: $x \leq y$

When this is not the case, that is $y<x$, the variables are inverted to ensure the measures remain consistent. Since our interest is in the magnitude of difference in ratios, that is how they increase or decrease together, we are not concerned with maintaining the two variable's juxtaposition as numerator and denominator. When considering the instance of $y<x$, then the following is used;

$$\mathrm{d}Fr_t = \log\left\{\frac{\left(\frac{y_t}{x_t}\right)}{\left(\frac{y_{t+1}}{x_{t+1}}\right)}\right\} \tag{3}$$

Such a calculation would be performed for a time series (or any given linear series) $(t=1)$, …, $(t=n)$ and for all pairs of variables that of the dataset. For a single pair of variables, this describes the variance that occurs over time for a given object. For a series of differential ratios (dFr), for variables x and y in a given series, the knowledge extracted can be represented in the form;

$$\text{Object} : x, y[\mathrm{dFr}_t, \mathrm{dFr}_{t+1}, \ldots, \mathrm{dFr}_{t+n}] \tag{4}$$

An actual example of this is given in Eq. (5). This describes the variance for the protein spot numbered 364 (following gel image software labelling) from our proteomics data. The *vol* (spot volume) and *circ* (the equivalent spot circularity) are two variables of the dataset, which form part of the description of each spot. The variance is shown over time within the square brackets. It can be noted that there is a peak of variance at time point 6, shown as a relative increase in dFr value. Such a peak may be representative of an 'interesting' spot, i.e. displaying a particular type of behaviour within the series. In this instance, the growth curve alters within the experiment on which this data mining was performed, between time points 6 and 7, hence proteins (such as this one) altering state at this stage may well be of interest to an expert.

ProtSpot 364

$$: \mathrm{vol}, \mathrm{circ}[0.5,\ 1.9,\ 0.2,\ 1.9,\ 1.3,\ 4.6,\ 2.2,\ 1.0] \tag{5}$$

Crucially, the variables analysed by this technique can include spatial elements. This is achieved by normalising the datasets and then placing values for absolute vectors. In this way, the technique can incorporate spatio-temporal data mining; however, it can also be used with temporal data (or data in any given linear series), without a spatial element.

A further advantageous feature of this technique is that it is also possible to know the total number of differential ratios that can be calculated before data mining is undertaken. For v number of variables, over time series $(t=1)$, …, $(t=n)$ this is given by;

$$\sum_{t=1}^{t=n}\left\{\frac{v_t(v_t-1)}{2}\right\} \tag{6}$$

With such knowledge to hand, a prediction of the length of the data mining process can be estimated, although impacts of CPU speed, etc. would of course also need to be considered, as in all data mining algorithms. Differential ratio data mining also has the feature of requiring only a single sweep of the dataset which can greatly increase the speed of the process (i.e.

```
CreateDRR{
    r[1] = GenerateRatios(D[1])
    for i = 2 to number of datasets in D{
        r[i] = GenerateRatios(D[i])
        for each row in D[i-1]{
            where row index of D[i-1] = row index of D[i]{
                for j = 1 to column size D[i-1]{
                    dFr[i].j = log(D[i-1].j / D[i].j)
                }
            }
        }
    }
}

with function;
ratios GenerateRatios(DataSet d){
    for each row in d{
        for all pairs (d.a, d.b)in D such that a≤b{
            r = d.a / d.b
        }
    }
    return r
}

where;
D[n] is the dataset at time point n
dFr[n] is differential ratio at time point n
r is ratios
```

Fig. 3. Algorithm describing the differential ratio data mining process.

decrease the total time taken to perform the data mining). This is unlike other techniques such as association rule data mining (Agrawal and Srikant, 1994), which requires multiple sweeps over the dataset. A single sweep over each dataset is all that is required since the variance at each time point is calculated only once. Furthermore, this has an additional beneficial impact on memory usage. The technique only requires that, at any one time, the ratios for two datasets are held in memory. Once the differential ratios have been calculated for those two particular time points, the earliest of the two ratios in the series can be removed from memory. Such features increase the efficiency of the data mining process, an important consideration when performing data mining on large datasets.

The algorithm describing the full data mining process is given in Fig. 3.

3.1.1. Interpreting the data mining results

To interpret the algorithm's results, we will define what the measure represents. For each dFr_t extracted the following can be said about the ratio between variables x and y over time point t and $t+1$;

$dFr_t < 0$ Ratio of difference has decreased over time

$dFr_t \sim 0$ Ratio has remained constant

$dFr_t > 0$ Ratio of difference has increased over time

A positive dFr value indicates that the two variable's values are growing *further apart* in terms of the two ratios over time. A negative value is the opposite of this, that is, the two variable's values are becoming *closer together* in terms of the two ratios over time. A value of around 0 indicates that the ratios between the variables has barely altered over time; exactly 0 meaning no difference at all. The magnitude of the measure also has a proportional meaning since the greater the value the more change has occurred. For instance, a larger positive dFr_t value means a larger difference in ratios over time compared with a smaller value.

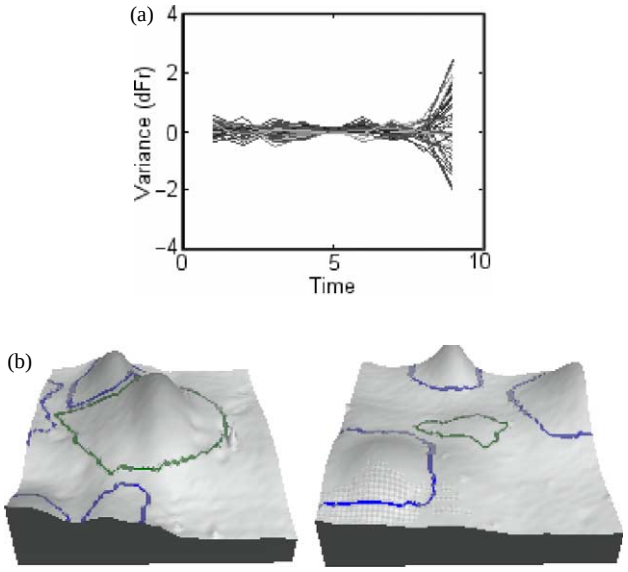

Fig. 4. (a) Visualisation of variance over time following differential ratio data mining. (b) how the protein spot in the centre has physically changed between time points 9 and 10, corresponding to the detected variance shown in the graph.

Results from the data mining process technique can also be visualised using a simple line graph plot, which also aids our interpretation. Fig. 4 (a) shows a plot of the various differential ratios produced for a single experiment. Each line within the graph represents a single differential ratio for a pair of variables over time. There are several smaller peaks at which variance is occurring, however at time point 9 there is a clearly a large peak of variance. This may indicate some interesting trend which would be flagged for further analysis. In this instance, the peak represented a large movement and increase in volume by the protein spot under analysis in a time series of gels, which can be seen in the image of the actual protein spot in Fig. 4 (b).

3.2. Classifying with a neural network

Following the data mining stage, the results of this process are then used to train and test a neural network classifier. This offers the advantage of being able to classify new, previously unseen data following successful training and hence introduce a level of automation to the process. The technique selected is that of the BackPropagation, Multi-Layer Perceptron (MLP) neural network. This neural network is a supervised learning approach which involves training the network using both the inputs and the required outputs.

A MLP (McClelland and Rumelhart, 1986) organises computational neurons into at least three layers, the input layer, the middle hidden layer and output layer. The learning rule typically used for the multi-layer neural network is the back-propagation rule that allows the network to learn to classify. This rule creates the output of the network, compares this with the required output and, by propagating the error back

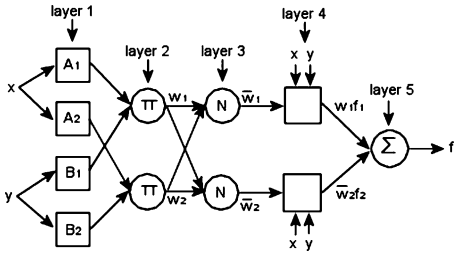

Fig. 5. A 5-layer ANFIS structure (Jang, 1993).

through the network, alters the weights to reduce the error. This supervised neural network was selected, as it is able to learn the salient features produced from the data mining in the first stage and produce output as pre-defined classes corresponding to particular protein behaviour.

These classes are normally an intrinsic part of the data under scrutiny, however in the case of the 2-DE gel data, they are not easily identifiable and can not be intuitively extracted by a non-expert. For this reason, knowledge acquisition sessions were conducted with 20 international experts and an optimisation of these fuzzy expert opinions was conducted in order to provide these output classes.

3.3. An adaptive neuro-fuzzy inference system

The creation of appropriate classes of protein behaviour is an important task, which has been previously tackled by Malone et al. (2004b). This approach uses an Adaptive Neuro-Fuzzy Inference System (ANFIS) to optimise knowledge discovery of expert opinions and extract usable and transparent rules. Such rules form the basis for producing the neural network's output classes.

ANFIS is a fuzzy inference system implemented within the architecture and learning procedure of adaptive networks (Jang, Sun and Mizutani, 1996; Jang, 1993). ANFIS can be used to optimise membership functions to generate stipulated input-output pairs and has the advantage of being able to subsequently construct fuzzy 'if-then' type rules representing these optimised membership functions (Malone, McGarry and Bowerman, 2004b).

The ANFIS structure used in this paper consisted of a 5-layer Sugeno type architecture, a typical example of which is seen in Fig. 5. In this example, two inputs are used (x,y) and one output (f) (which is a limitation of Sugeno-type systems, i.e. that there is only a single output, obtained using weighted average defuzzification (linear or constant output membership functions)).

Following knowledge acquisition sessions with experts from within the field of proteomics, heuristic knowledge is acquired and represented in the form of fuzzy rules. Such rules have the advantage of being able to represent intuitive terms in a form close to natural language and do not have exact thresholds, reducing the likelihood of brittleness. These expert opinions represent the inputs used to train the ANFIS model which proceeds to create optimised membership functions, which will be used to create the output classes of the MLP used in the overall data mining architecture.

The structure used in this paper for experimentation consists of 6 inputs, described in Table 1. These inputs form the basis of the expert opinions and are used in various rules to describe whether or not a particular protein spot is 'interesting' or not and hence worthy of further laboratory analysis. Single or combinatorial rules consisting of these input parameters will form the basis of the output classes for the neural network.

The training dataset consisted of 75% of the total exacted fuzzy expert opinions with the test dataset consisting of the remaining 25%. Table 2 shows the classification accuracy of the ANFIS, with the accuracy for 'interesting' proteins of 96%. That is to say, the network is able to correctly identify 96% of the proteins corresponding to interesting behaviour, characterised by the optimised expert rules.

Following the training and testing, the optimised fuzzy expert rules were extracted, a sample of which is shown in Fig. 6. The membership functions of each input are described in fuzzy terms, such as high and low for all but %_Change which also has a medium membership.

Following this optimisation stage, the proteins corresponding to each of the 14 'interesting' rules were identified from the expert's analysis, previously conducted, and training and testing datasets were created. The results of experiments using the full architecture are discussed in Section 4.

4. Results and discussion

Four strategies were employed to test the effectiveness of the proposed architecture and provide comparative analysis with other variance analysing techniques; (i) using the intelligent data mining architecture, using differential ratios, described in this paper; (ii) using PCA as a variance analysis method to provide data for the neural network; (iii) using covariance as a variance analysis method to provide data for the neural network; (iv) finally, using normalised data only, without any form of variance analysis, to train and test the neural network.

Table 1
Inputs used in ANFIS structure

Input	Description	No. Membership Functions
Absence/Presence	The absence and presence of protein spots from gel to gel	2
Budding	The joining or separation of protein spots from gel to gel	2
Percentage Variation	Percentage change in terms of spot abundance from gel to gel	3
Shape Change	Morphological changes in shape, such as height from gel to gel	2
Volume Change	Increase or decrease in the volume of a protein spot from gel to gel	2
X/Y Movement	Geographical movement of a protein spot in x and y dimensions from gel to gel	2

Table 2
Test results on ANFIS

Class	Classification Accuracy	Total No. Optimised Fuzzy Rules
Interesting proteins	96%	14

Experiments were conducted using two post-experimental 2-DE gel datasets with the aim of correctly identifying the 'interesting' proteins, i.e. those corresponding to the 14 classes of interesting behaviour as specified from the ANFIS optimisation process. The first dataset, the '*Biswarup*' concerns the analysis of the proteome of *Methanococcus jannaschii* (Mukhopadhyay, Johnson and Wolfe, 2000), the first of its kind of microorganism to have its genome sequenced. The experiment was performed to identify any changes occurring as it moved through different phases of growth. It was designed to produce a large dataset representing sampling points spanning the entire growth curve; samples were removed at 10 intervals throughout growth. The second, named *'Argonne'*, was designed to produce sample variation for replicate groups of bacteria growing under controlled growth conditions, such as variation in the Nitrogen and Hydrogen content and temperature change. Table 3 provides a summary of the two datasets used.

Some basic assumptions have been made concerning the collection of the data described above. It is assumed that the findings are statistically significant, i.e. that $P<0.05$ within experiments resulting in data collection and that warping and noise reduction are performed before analysis commences.

The results obtained from experiments using the *Biswarup* dataset are described in Table 4. The hidden units used in each experiment were selected because they were found to be the optimum level in terms of performance for that particular strategy.

It is worth noting that false positives, that is, proteins classified in an 'interesting' class, are more acceptable than false negatives, that is, proteins not classified as interesting although they may well be (Vohradsky, 1997). Taking this into consideration, a boundary threshold to indicate the margin by which an output was considered belonging to a class was introduced. For example, for a given class, x, the expected output of that unit should be ~ 1; an output of ≥ 0.7 would be classified as class x if the boundary threshold was 0.3. In this way, we can classify those examples which are not very strong members of the class (within 0.1 of output for that unit) although they may end up being identified as a false positive—however, as noted, this is more acceptable than the converse being true.

Experiments using the *Biswarup* dataset showed reasonable performance for most strategies when the boundary threshold was wider, decreasing, fairly predictably, for all strategies when the margin of acceptance is decreased. The PCA-neural network and the covariance-neural network architectures faired reasonably well, with very similar performances. The classification rates for the normalised data-neural network architecture, where no variance analysis was conducted, performed worst for all thresholds. The results for the differential ratio data mining trained-neural network architecture clearly showed a performance advantage over all other classifiers, regardless of boundary threshold. The comparative classification gains ranged from a 46.4% increase on normalised data-neural network architecture to a 16.7% increase on PCA-neural network architecture.

The results of experiments with the *Argonne* dataset are shown in Table 5. Performance of the four strategies used show a similar pattern of results for this dataset. Again, the differential ratio data mining trained-neural network architecture clearly showed a performance advantage over the other techniques tried, with the neural network trained on normalised data only performing worst.

On comparison of results obtained, it is clearly shown that the use of a variance analysis before training of the neural network classifier provides performance benefits. Furthermore, in every instance of comparable boundary thresholds, the three architectures using variance analysis before training, performed better than the neural network trained on normalised data only, without exception. This indicates that the neural network is not fully able to encapsulate and identify within its structure the salient features within the proteomics data. Therefore, any strategy employing the use of a neural network for datasets of this type, would benefit from the use of a variance analysis to increase accuracy and hence reliability. This is especially true of the differential ratio data mining. One possible explanation for the significant performance benefits of using differential ratio data mining over other variance analysis techniques is the method's ability to fully incorporate spatial and temporal elements of the data. Important trends may well be contained within these elements, which are not fully incorporated in the other variance analysis representations.

One important consideration, as previously discussed, is producing a low number of false negatives, although false positives are more acceptable. Table 6 presents a summary of

If (Absence/Presence is high) and (X/Y_Movement is low) and (Volume_Change is high) and (Budding is low) and (Shape_Change is high) and (%_Change is low) then (output is class1)
If (Absence/Presenceis high) and (X/Y_Movement is high) and (Volume_Change is high) and (Budding is high) and (Shape_Change is low) and (%_Change is low) then (output is class2)
If (Absence/Presence is high) and (X/Y_Movement is high) and (Volume_Change is high) and (Budding is high) and (Shape_Change is high) and (%_Change is low) then (output is class3)
If (Absence/Presence is high) and (X/Y_Movement is high) and (Volume_Change is high) and (Budding is low) and (Shape_Change is low) and (%_Change is low) then (output is class4)
If (Absence/Presence is high) and (X/Y_Movement is low) and (Volume_Change is high) and (Budding is low) and (Shape_Change is low) and (%_Change is high) then (output is class5)

Fig. 6. Optimised fuzzy expert rules.

Table 3
Proteomics datasets used. No. of objects represents the number of training and test examples combined

Dataset	No. of objects	No. of variables	Temporal points
Biswarup	897	16	10
Argonne	607	14	9

Table 4
Results from experimentation using the *Biswarup* dataset

Experimental strategy used to train/test neural network	Hidden units	Boundary threshold	% Correct classification	Mean squared error
Covariance	40	0.1	63.5	0.17
	40	0.2	65.7	0.17
	40	0.3	71.7	0.17
Differential ratio data mining	45	0.1	82.3	0.10
	45	0.2	89.6	0.10
	45	0.3	92.2	0.10
Normalised data	35	0.1	35.9	0.30
	35	0.2	49.3	0.30
	35	0.3	59.8	0.30
PCA	12	0.1	61.1	0.18
	12	0.2	67.1	0.18
	12	0.3	75.5	0.18

the number of false positives and false negatives classified for the two datasets compared to the results of an expert's findings on the dataset. The assumption made within this baseline comparison is that the expert's analysis is fully correct and therefore has 100% accuracy with those identified as interesting. This comparison shows us that there were a number of the more acceptable false positives in all experiments, although, again, a lower number in the differential ratio data mining strategy, as would be expected by high levels of classification accuracy previously discussed. The less acceptable false negatives were also present in most techniques, especially the neural network trained only on normalised data. However, the data mining trained neural network performed best of all with only two false negatives across experiments on both datasets. For reasons previously discussed, this is an important feature since false negatives are in effect determining possibly interesting proteins to be of no importance and, therefore, would be overlooked for further laboratory analysis.

Table 5
Results from experimentation using the *Argonne* dataset

Experimental strategy used to train/test neural network	Hidden units	Boundary threshold	% Correct classification	Mean squared error
Covariance	35	0.1	61.5	0.19
	35	0.2	68.5	0.19
	35	0.3	70.2	0.19
Differential ratio data mining	40	0.1	79.6	0.07
	40	0.2	80.3	0.07
	40	0.3	88.7	0.07
Normalised data	30	0.1	42.3	0.28
	30	0.2	47.3	0.28
	30	0.3	56.1	0.28
PCA	12	0.1	59.5	0.21
	12	0.2	61.9	0.21
	12	0.3	65.2	0.21

Table 6
False positive and false negative classifications

Dataset	Experimental strategy	No. false positives	No. false negatives
Biswarup	Covariance	8	10
	Differential ratio data mining	5	0
	Normalised data	7	15
	PCA	11	7
Argonne	Covariance	10	5
	Differential ratio data mining	4	2
	Normalised data	7	14
	PCA	9	7

The major contribution of this work is to enable the post-experimental data analysis to be performed automatically, thereby greatly decreasing the amount of time spent during this important step. Furthermore, the technique will help to identify potentially interesting proteins, which may well have been otherwise missed through manual analysis. Although time is taken to perform the data mining and train and test the neural network, this took under 15 min for each network, and less again for the neural network trained on normalised data only. The ANFIS stage is seen as a once only procedure for the experiments since the output classes remain the same once the network is trained. The obvious advantage from this type of architecture is that once the network is trained and optimised, it can then be used at a later time to classify previously unseen data.

5. Conclusion

In this paper we presented an intelligent data mining architecture and performed experiments using two post-experimental, 2-DE gel datasets. Three variance analysis methods were applied to the datasets to use as training and testing data for a BackPropagation, Multi-Layer Perceptron (MLP) neural network in order to classify the results of the data mining into discrete classes of interesting behaviour. The neural network was also trained and tested using normalised data only to assess the benefits of using a variance analysis step before machine learning. Of the three variance analysis methods employed and tested, the differential ratio data mining proved to be the most successful in identifying and representing the salient trends within the data. The intelligent data mining architecture also provided the lowest number of false negatives and false positives of all strategies, an important consideration when attempting a comprehensive and accurate analysis of the data.

The architecture also allows the encapsulation of expert opinions through the use of an adaptive fuzzy logic system (ANFIS). This offers the advantage of optimising initially approximate data in an effective manner whilst, following training, allowing fuzzy rules to be extracted which represent the optimised fuzzy membership functions. Such membership functions form the basis of our output classes, which correspond to interesting features of protein behaviour.

This research goes some way to addressing the processing bottleneck that exists within post-experimental 2-DE gel data analysis by providing a technique that automatically extracts potentially interesting proteins from within the datasets. Since the technique involves the use of a supervised neural network, normal considerations of suitability apply, i.e. that empirical data must be available in order to train and test the network's ability to learn and classify correctly.

Future work will concentrate on expanding the technique to further proteomics data sets. We also aim to show that this approach is suitable more generally as a spatio-temporal data mining technique by expanding to other spatio-temporal datasets such as robotics and meteorological data.

Acknowledgements

The authors would like to acknowledge the support of EPSRC (grant GR/P01205), NonLinear Dynamics Ltd and Biswarup Mukhopadhyay of the Virginia Bioinformatics Institute. Any errors or omissions remain those of the named authors.

References

Agrawal, R., & Srikant, R., (1994). Fast algorithms for mining association rules. *Proceedings of the 20th international conference on very large databases* (pp. 487–499).

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., & Watson, J. (2002). *Molecular biology of the cell*. New York: Garland Science.

Chou, S.-M., Lee, T.-S., Shao, Y. E., & Chen, I. F. (2004). Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, *27*(1), 133–142.

Crenshaw, T., & Cory, J. (2002). Overexpression of protein disulfide isomerase-like protein in a mouse leukemia L1210 cell line selected for resistance to 4-methyl-5-amino-1-formylisoquinoline thiosemicarbazone, a ribonucleotide reductase inhibitor. *Advances in Enzyme Regulation*, *42*, 143–157.

Damcott, C. M., Moffett, S. P., Feingold, E., Barmada, M. M., Marshall, J. A., Hamman, R. F., et al. (2004). Genetic variation in fatty acid-binding protein-4 and peroxisome proliferator-activated receptor gamma interactively influence insulin sensitivity and body composition in males. *Metabolism*, *53*(3), 303–309.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127.

Dowsey, A. W., Dunn, M. J., & Yang, G. Z. (2003). The role of bioinformatics in two-dimensional gel electrophoresis. *Proteomics*, *3*(8), 1567–1596.

Fenyo, D., & Beavis, R. C. (2002). Informatics and data management in proteomics. *Trends in Biotechnology*, *20*(12), S35–S38.

Griffin, T. J., & Aebersold, R. (2001). Advances in proteome analysis by mass spectrometry. *Journal of Biological Chemistry*, *276*(45), 497–500.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge: MIT Press.

Jang, J. S. R. (1993). ANFIS: Adaptive-network-based fuzzy inference systems. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*(3), 665–685.

Jang, J. S. R., Sun, C. T., & Mizutani, E. (1996). *Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence*.

Jenkins, R. E., & Pennington, S. R. (2001). *Novel approaches to protein expression analysis Proteomics: From protein sequence to function* (pp. 207–224).

Korn, F., Labrinidis, A., Kotidis, Y., & Faloutsos, C. (2000). Quantifiable data mining using ratio rules. *The International Journal on Very Large Data Bases* , 254–266.

Krah, A., Wessel, R., & Pleißner, K. P. (2004). Assessment of protein spot components applying correspondence analysis for peptide mass fingerprint data. *Proteomics*, *4*(10), 2982–2986.

Liebler, D. (2002). *Introduction to proteomics*. Totowa, NJ: Humana Press.

Malone, J., McGarry, K., & Bowerman, C. (2004a). Performing trend analysis on spatio-temporal proteomics data using differential ratio data mining. *Proceedings of the sixth EPSRC conference on postgraduate research in electronics, photonics, communications and software (Prep 2004)* (pp. 103–105).

Malone, J., McGarry, K., & Bowerman, C. (2004b). Using an adaptive fuzzy logic system to optimise knowledge discovery in proteomics. *Fifth international conference on recent advances in soft computing (RASC) 2004* (pp. 80–85).

Marengo, E., Leardi, R., Robotti, E., Righetti, P. G., Antonucci, F., & Cecconi, D. (2003). Application of three-way principal component analysis to the evaluation of two-dimensional maps in proteomics. *Journal of Proteome Research*, *2*(4), 351–360.

Marengo, E., Robotti, E., Antonucci, F., Cecconi, D., Campostrini, N., & Righetti, P. G. (2005). Numerical approaches for quantitative analysis of two-dimensional maps: A review of commercial software and home-made systems. *Proteomics*, *5*(3), 654–666.

Marengo, E., Robotti, E., Righetti, P. G., Campostrini, N., Pascali, J., Ponzoni, M., et al. (2004). Study of proteomic changes associated with healthy and tumoral murine samples in neuroblastoma by principal component analysis and classification methods. *Clinical Chimica Acta*, *345*(1–2), 55–67.

McClelland, J., & Rumelhart, D., (1986). Mechanisms of sentence processing: Assigning roles to constituents of sentences. *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 272–325). Cambridge, MA: MIT Press.

Mendyk, A., & Jachowicz, R. (2005). Neural network as a decision support system in the development of pharmaceutical formulation—focus on solid dispersions. *Expert Systems with Applications*, *28*(2), 285–294.

Mukhopadhyay, B., Johnson, E. F., & Wolfe, R. S. (2000). A novel pH2 control on the expression of flagella in the hyperthermophilic strictly hydrogenotrophic methanarchaeaon Methanococcus jannaschii. *Proceedings of the National Academy of Sciences USA*, *97*(21), 11522–11527.

Pederson, L., & Ersboll, B., (2001). Protein spot correspondence in two-dimensional electrophoresis gels. *Proceedings of 12th Scandinavian conference on image analysis* (pp. 118–215).

Pennington, S. R., Wilkins, S. R., Hochstrasser, D. F., & Dunn, M. J. (1997). Proteome analysis: From protein characterisation to biological function. *Trends in Cell Biology*, *17*(4), 168–173.

Picard, P., Bourgoin-Greneche, M., & Zivy, M. (1997). Potential of two-dimensional electrophoresis in routine identification of closely related durum wheat lines. *Electrophoresis*, *18*(1), 174–181.

Pleissner, K., Oswald, H., & Wegner, S. (2001). *Image analysis of two-dimensional gels Proteomics: From protein sequence to function* (pp. 131–149). Oxford: BIOS Scientific Publishers.

Pleissner, K., Regitz-Zagrosek, V., Krudewagen, B., Trenkner, J., Hocher, B., & Fleck, E. (1998). Effects of renovascular hypertension on myocardial protein patterns: Analysis by computer-assisted two-dimensional gel electrophoresis. *Electrophoresis*, *19*(11), 2043–2050.

Righetti, P., Stoyanov, A., & Zhukov, M. (2001). *The proteome revisited: Theory and practice of all relevant electrophoretic steps*. Amsterdam: Elsevier.

Rooney-Varga, J. N., Giewat, M. W., Savin, M. C., Sood, S., Legresley, M., & Martin, J. L. (2005). Links between phytoplankton and bacterial community dynamics in a coastal marine environment. *Microbial Ecology*, *49*(1), 163–175.

Sekiguchi, H., Watanabe, M., Nakahara, T., Xu, B., & Uchiyama, H. (2002). Succession of bacterial community structure along the Changjiang river determined by denaturing gradient gel electrophoresis and clone library analysis. *Applied and Environmental Microbiology*, *68*(10), 5142–5150.

Toroslu, I. H., & Yetisgen-Yildiz, M. (2005). Data mining in deductive databases using query flocks. *Expert Systems with Applications*, *28*(3), 395–407.

Vohradsky, J. (1997). Adaptive classification of two-dimensional gel electrophoretic spot patterns by neural networks and cluster analysis. *Electrophoresis*, *18*(15), 2749–2754.

Whittaker, P. A. (2003). What is the relevance of bioinformatics to pharmacology? *Trends in Pharmacological Sciences*, *24*(8), 434–439.

Wiemer, J. C., & Prokudin, A. (2004). Bioinformatics in proteomics: Application, terminology, and pitfalls. *Pathology-Research and Practice*, *200*(2), 173–178.