INTELLIGENT HYBRID SPATIO-TEMPORAL DATA MINING FOR KNOWLEDGE DISCOVERY ON PROTEOMICS DATA

James Malone, Ken McGarry and Chris Bowerman

School of Computing and Technology
University of Sunderland
St Peter's Way, Sunderland, SR6 0DD
UNITED KINGDOM
[james.malone ken.mcgarry chris.bowerman] @sunderland.ac.uk

ABSTRACT

The need to extract and subsequently represent meaningful knowledge from biomedical data sets is a rapidly growing area of research. Often, such data sets offer further challenges than more traditional analysis, since many domains contain data that is inherently multi-dimensional and contains spatial and/or temporal elements. This may be further complicated still by the need to incorporate expert's heuristics in order to perform any meaningful analysis which may not be detected by data driven techniques alone. We present a novel hybrid architecture to perform knowledge discovery on spatio-temporal proteomics data. This architecture employs a combination of data and goal driven elements in order to allow the integration of expert's opinions within the process. This approach is able to automatically identify proteins exhibiting interesting behaviour from large proteomics data sets.

1. INTRODUCTION

The development of data mining techniques to effectively and efficiently analyse biomedical data is an increasingly important area of research [1]. Biomedical experiments, such as those in proteomics, can result in large, multi-dimensional data sets, often from heterogeneous sources. Analysis is further complicated when such post-experimental data contains spatial and/or temporal elements.

Within the field of proteomics, two-dimensional electrophoresis (2-DE) [2] is unrivalled as a technique to perform protein expression analysis [3, 4]. However, this technique is not without its limitations. Without the availability of reliable tools for post-experimental data analysis, the technique is essentially a descriptive one [5]. Furthermore, the method is inherently labour-intensive and requires a skill-level such that trained experts perform the analysis manually. This limits the potential for full automation [5]. Since this post-experimental 2-DE analysis is often conducted by experts using heuristic knowledge obtained through experience, an important consideration would be the incorporation of expert domain knowledge within any knowledge discovery method [6].

We demonstrate the use of a novel knowledge discovery architecture which addresses these needs. Using this approach, we achieve the integration of heuristics in the form of fuzzy expert opinions with a spatio-temporal data mining element. We go on to demonstrate how this architecture can automate the process of analysing post-experimental proteomics data and conduct a comparison with other variance identifying techniques. Finally, a qualitative analysis is conducted using a comparison of our novel architecture to the manually derived findings of an expert. The long term aim of such an investigation is to incorporate such an architecture into a single automated process by which 2-DE gels are converted into digital images then subsequently into meaningful and interesting knowledge.

2. 2-DE ELECTROPHORESIS ANALYSIS

The original inspiration for this research comes from the task of interpreting post-experimental, two-dimensional electrophoresis (2-DE) gel data [7], a key component of current proteomics research [8]. 2-DE gels are used to separate thousands of proteins according to their electric charge and molecular weight. With the use of staining, proteins appear as dark spots on such gels (see Figure 1).

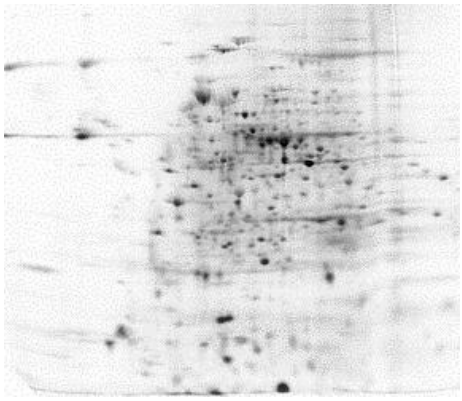


Figure 1. A single 2-DE gel image

Image analysis software is then used to detect spot correspondence from one gel to the next [9, 10]. Such software can perform warping and quality control with the use of intelligent algorithms which can automatically detect whether an entity is a protein spot or superfluous background noise. It can also be used to visualise individual spots in 3-D by extracting detailed information from each gel (Figure 2). Such software converts these images into data which describes each protein, such as volume, area, height, equivalent circularity and x and y coordinates on gel. Each spot is also uniquely numbered to allow ease of identification within the database and corresponding gel image.

These experiments can help to identify proteins which may be inherently linked to particular conditions which cause them to be expressed in more or less abundance or with differing characteristics from one control group to the next.

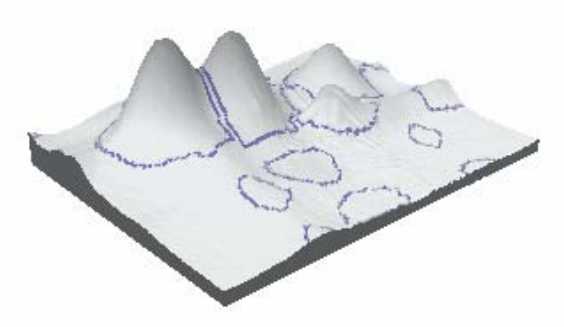


Figure 2. Individual spots can be visualised in 3-D using image analysis software. Such software converts the gels into data which describes the various three-dimensional attributes of the proteins. This image produced by Nonlinear's Progenesis software.

Since such experiments can create large amounts of high-dimensional, spatio-temporal experimental data, manual interpretation of results is impractical [11] and inherently laborious [6]. Since current knowledge discovery methods are unable to handle and analyse such results meaningfully [12], research into novel approaches to automate the task of knowledge discovery is crucial to the progress of the technique.

3. KNOWLEDGE DISCOVERY ARCHITECTURE

The proposed knowledge discovery architecture incorporates both goal driven (expert heuristics) and data driven (data mining) elements of the proteomics data analysis and is illustrated in Figure 3. In the first stage, spatio-temporal data mining is performed on the normalised proteomics data. This entirely data driven element of the knowledge discovery process automatically analyses the data for trends of covariance. The result of this process provides the training data for the second stage which employs a back-propagation, multi-layer perceptron (MLP) neural network as a classifier.

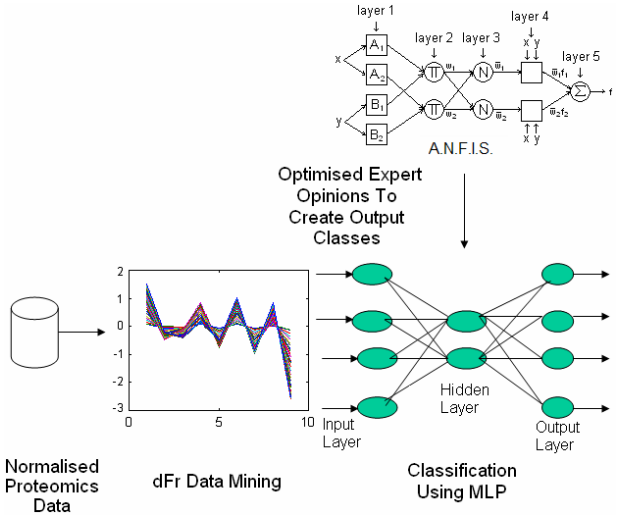


Figure 3. Knowledge Discovery Architecture

The creation of appropriate output classes of 'interesting' proteins is an important task which has been previously tackled by Malone *et al* [6]. 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 of differing interesting protein behaviour.

3.1. Differential Ratio Data Mining

The first stage of the knowledge discovery process is that of spatio-temporal data mining. The technique we employ to perform the proposed spatio-temporal analysis is that of differential ratio (dFr) data mining [1]. This technique draws on certain elements of covariance measures and ratio rules [13]. Covariance measures capture the linear dependencies between variables within a series [14] and have previously been used within biomedical data analysis [15]. 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;

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

Where:

$\bar{x}$ is sample mean of X values

$\bar{y}$ is sample mean of Y values

Ratio rules data mining is a technique that employs *eigensystem analysis* to calculate correlations between values of attributes. Ratio rules can be used to perform 'what-if' type scenarios given an antecedent(s) or consequent(s) and can tackle the issue of reconstructing missing/hidden values. 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 such data.

3.2. Differential Ratio Algorithm

Differential ratio data mining is used to measure the covariance of a given object in terms of the log of pairwise ratios of the elements describing the data over time (or within any given linear series). 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;

$$\log\left\{\frac{\left(\frac{x_t}{y_t}\right)}{\left(\frac{x_{t+1}}{y_{t+1}}\right)}\right\} = \mathrm{dFr}_t \qquad (2)$$

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;

$$\log\left\{\frac{\left(\frac{y_t}{x_t}\right)}{\left(\frac{y_{t+1}}{x_{t+1}}\right)}\right\} = \mathrm{dFr}_t \qquad (3)$$

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 covariance 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;

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

An actual example of this is given in Equation (5). This describes the covariance for the protein spot numbered 142 (following gel image software labelling) from our proteomics data. The *vol* (spot volume) and *circ* (the equivalent spot circularity) are two variables of the data set which form part of the description of each spot. The covariance is shown over time within the square brackets. It can be noted that there is a peak of covariance at time point 3, 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 3 and 4, hence proteins (such as this one) altering state at this stage may well be of interest to an expert.

$$\text{ProtSpot 142: } vol,circ[2.1, 1.6, 3.2, 1.9, 1.1, 1.0] \qquad (5)$$

Crucially, the variables used in this calculation can include spatial elements. This would be achieved by first normalising the datasets and then placing values for absolute vectors. Such variables can then be included when performing the trend analysis to ensure comprehensive data mining. 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.

Advantageously, 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;

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

With such knowledge to hand, a prediction of the length of the data mining process can be estimated although impacts of CPU speed, RAM, etc. would also need consideration, as with 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. decrease the total time taken to perform the data mining). This is unlike other techniques such as association rules [16] which require multiple sweeps over the dataset. A single sweep over each dataset is all that is required since the covariance at each time point is calculated only once. Furthermore, 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 time series ratios can be removed from memory.

Within data mining algorithms, there is a large overhead associated with I/O operations to read in rows of data, N. Therefore, the reduction of computation involving N is desirable for any data mining technique. The feature of differential ratio data mining requiring only a single sweep across the data set increases the efficiency of the knowledge discovery process, an important consideration when performing data mining on large data

sets and, hence, large values of *N*. The algorithm describing the full data mining process is given in Figure 4.

```
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
```

Figure 4. The differential ratio data mining algorithm

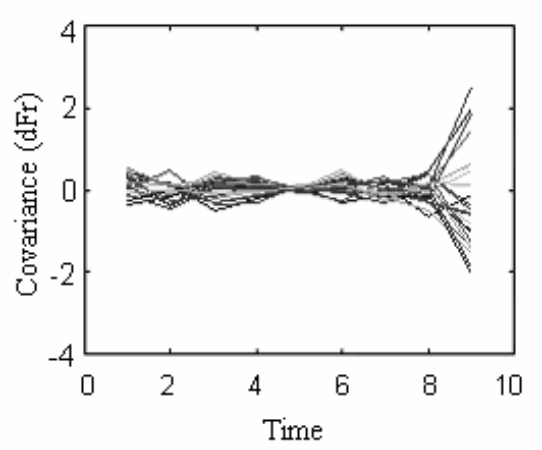
3.3. Interpretation of Data Mining Results

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 variable *x* and *y* over time point *t* and *t+1*;

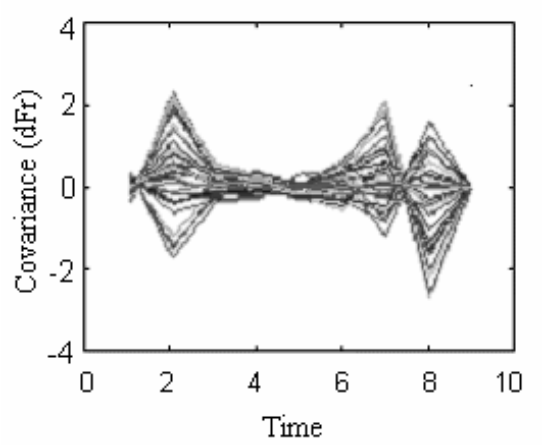
$dFr_t > 0$ Ratio of difference has increased over time
$dFr_t < 0$ Ratio of difference has decreased over time
$dFr_t \sim 0$ Ratio has remained constant

That is, 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.

Visualising the results of this data mining technique can also help in interpretation. Figure 5 shows two simple line graph plots of the various differential ratios produced for a single experiment. Each line within the graphs represents a single differential ratio for a pair of variables over time. In Figure 5 (a) there are several smaller peaks at which covariance is occurring, however at time point 9 there is clearly a large peak of covariance. 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.



(a)



(b)

Figure 5. Visualisations of covariance over time. (a) A peak of covariance is apparent at the final time point, in this case caused by variations in volume and protein shape. (b) There are three peaks of covariance in this plot at time points 2, 8 and 9 corresponding to an increase in volume and shape change (points 2 and 8) and the protein being absent at point 9.

3.4. Creating Output Classes

Following 20 knowledge acquisition sessions with international experts and the use of an ANFIS model to optimise the acquired expert heuristics, rules were constructed as discussed in Malone *et al* [7]. An example of these rules is given in Table 1. These rules formed the basis of the output classes defined in our neural network architecture. Following this knowledge optimisation

process, the spots were categorised according to which set of rules they corresponded to. For those spots not falling into any of the extracted classes, an 'other' class was created.

Table 1. An extract of optimised fuzzy expert opinions

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) then (output is high)
If (Absence/Presence is high) and (X/Y_Movement is high) and (Volume_Change is high) and (Budding is high) and (Shape_Change is low) then (output is high)
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) then (output is high)
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) then (output is high)
(Absence/Presence is high) and (X/Y_Movement is high) and (Volume_Change is high) and (Budding is low) and (Shape_Change is high) then (output is high)
(Absence/Presence is low) and (X/Y_Movement is high) and (Volume_Change is high) and (Budding is low) and (Shape_Change is high) then (output is high)

4. RESULTS AND DISCUSSION

Four sets of experiments were performed using two post-experimental 2-DE data sets with the aim of correctly identifying the 'interesting' proteins. The first data set, the '*Biswarup*' concerns the analysis of the proteome of *Methanococcus jannaschii* [17], 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 the different phases of growth. It was designed to produce a large data set 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 Bacterial growing under controlled growth conditions, such as variation in the Nitrogen and Hydrogen content and temperature change. Table 2 provides a summary of the two data sets used.

Table 2. Summary of data sets. No. of objects represents the number of training and test data combined

Data Set	No. of Objects	No. of Variables	Temporal Points	Spatial Element
Biswarup	897	16	10	X, Y coordinate within image Vector of movement from previous image
Argonne	607	14	9	X, Y coordinate within image Vector of movement from previous image

The variables describing the data sets, and hence used in the experimentation, are described in Table 3. The variables were common to both data sets and characterise each spot present on a gel within a series.

Table 3. Summary of attributes used to describe each protein spot within the Proteomics data sets

Variable	Description	Value
1st dimension	Molecular Weight (MW) of spot	Real number
2nd dimension	Isoelectric Point (pI) of spot	Real number
Area	The total surface area of the spot	Real number
Background	(area covered by boundary of each spot)	Real number
Circularity	The roundness of a spot (0-1). Higher value, the rounder the spot	Real number
Equivalent radius (width of image on gel)	Equivalent radius of a circle with the same area as the protein	Real number
Peak height	Maximum height of protein spot	Real number
Vector length	Distance from a matched spot to the spot to which it is matched.	Real number
Volume	Volume of protein spot	Real number
x coordinate	Coordinate of protein on gel on x-axis	Integer
y coordinate	Coordinate of protein on gel on y-axis	Integer

In order to further evaluate the results of the knowledge discovery process, a comparative analysis was also conducted with Principal Components Analysis (PCA) to compare the network's performance trained using a different variance measure. The network was also tested using only the normalised data to determine whether or not there were any benefits of using a variance measure within the knowledge discovery architecture. Finally, as an alternative classifier, the C5 decision tree generator was tested. For training and testing purposes, the data sets were partitioned; 75% was used for training and 25% for testing to evaluate the technique's ability to generalise on previously unseen data.

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 classifier. Initially, for each neural network, 2/3 of the number of input units was used to determine the number of hidden units. This was then increased and decreased until the optimum numeber was reached. 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 ≥0.8 would be classified as class *x* if the boundary threshold was 0.2. Table 4 describes the results obtained from knowledge discovery experiments using the *Biswarup* data set.

Table 4. Knowledge discovery results using *Biswarup* data set

Experiment	Hidden Units	Boundary Threshold	% Correct Classification	Mean Squared Error
Neural network trained on normalised data	40	0.2	58.4%	0.34
	40	0.1	37.2%	0.34
Neural Network trained on PCA results	10	0.2	64.7%	0.21
	10	0.1	59.7%	0.21
Neural network trained on differential ratio data mining results	50	0.2	87.6%	0.11
	50	0.1	80.3%	0.11
C5 Decision Tree on Normalised Data	N/A	N/A	61.6%	-

The first *Biswarup* data set experiments using the normalised data-neural network architecture showed reasonable performance for wider boundaries, rapidly deteriorating as the boundary was constrained further. The PCA-neural network architecture faired better, with improved classification rates over the normalised data-neural network architecture for both boundary 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 43% increase on normalised data-neural network architecture to a 20% increase on PCA-neural network architecture. The C5 Decision Tree performed averagely (the boundary value was not applicable to this technique), but was again outperformed by the differential ratio data mining-neural network architecture.

Table 5. Knowledge discovery results using *Argonne* data set

Experiment	Hidden Units	Boundary Threshold	% Correct Classification	Mean Squared Error
Neural network trained on normalised data	30	0.2	59.9%	0.28
	30	0.1	47.3%	0.28
Neural Network trained on PCA results	15	0.2	68.8%	0.19
	15	0.1	63.7%	0.19
Neural network trained on differential ratio data mining results	40	0.2	89.7%	0.07
	40	0.1	80.3%	0.07
C5 Decision Tree on Normalised Data	N/A	N/A	64.5%	-

If we consider the results from knowledge discovery experiments using the *Argonne* data set, described in Table 5, we see similarities reflected from the first set of experiments performed. Specifically, the classification rate of the normalised data-neural network architecture and PCA-neural network architecture performed reasonably well, but were out-performed by the differential ratio-neural network architecture for both boundary constraints. The comparative classification gains using our knowledge discovery architecture for this data set ranged from a 29% increase on normalised data-neural network architecture to a 20% increase on PCA-neural network architecture.

One important consideration which has been previously discussed is that such proteomics data sets are usually analysed manually by a trained expert to highlight 'interesting' spots. Therefore, a comparison of the results produced automatically from the knowledge discovery process with that of an expert manually analysing the data sets would appear a valid qualitative measure.

Following an expert's analysis of the *Biswarup* data set we compared the number of highlighted spots from this process to those extracted from our knowledge discovery architecture. The results are summarised in Table 6.

Table 6. Comparison of expert manual analysis and automated data mining-neural network knowledge discovery

Experiment	Boundary Constraint	No. Extracted Spots	No. of expert spots highlighted (%)	% Correct Classification
Expert's Manual Analysis	N/A	42	N/A	(assumed to be 100%)
Neural network trained on normalised data	0.2	109	95%	87.6%
Neural network trained on normalised data	0.1	97	95%	87.6%

The assumption made within this comparison is that the expert's analysis is fully correct and therefore has 100% accuracy with those identified as interesting. The results indicate that 95% of the interesting spots highlighted from manual analysis were also captured from our knowledge discovery process. This figure did not decrease when the boundary constraint was decreased which indicates that these expert discovered spots identified by our knowledge discovery technique are not merely examples lying near the fringes of a class.

5. CONCLUSIONS

In this paper we have presented the use of a spatio-temporal data mining architecture to perform knowledge discovery on post-experimental 2-DE proteomics data. The approach tackles the issues of incorporating knowledge from heterogeneous sources using both; (i) a data

driven element, using differential ratio data mining, and (ii) a goal driven element, using an Adaptive Neuro-Fuzzy Inference System to optimise expert's heuristics. These two elements are used to construct a hybrid classifier using data from heterogeneous sources in order to automate the process of knowledge discovery. This knowledge is ultimately represented as protein spots which are members of discrete, labelled classes. Differential ratio data mining is able to identify salient trends of covariance within data whilst incorporating both the spatial and temporal elements during analysis.

In order for 2-DE to become more than just a descriptive technique, reliable tools for the identification and analysis of protein spots need to be developed and integrated into the overall process. One possible solution would be a 'workflow' architecture, in which gels are converted to digitised images and subsequently automatically analysed for potentially interesting proteins, identified using a form of knowledge discovery, such as that described in this research.

The use of techniques such as data mining and machine learning may help to overcome some of the limitations of the manual analysis, such as reducing the noise intrinsic to such experiments. Such techniques will also help to decrease the labour-intensive elements of analysis whilst decreasing the total time taken to perform analysis by automating knowledge discovery.

6. ACKNOWLEDGEMENTS

The authors would like to 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.

7. REFERENCES

[1] J. Malone, K. McGarry and C. Bowerman, "Performing trend analysis on spatio-temporal proteomics data using differential ratio data mining," in Proc. 6th EPSRC Conference on Postgraduate Reasearch in Electronics, Photonics, Communications and Software (PREP 2004), 2004, pp. 103-105.

[2] P. H. O'Farrell, *Journal Biological Chemistry* 250, 4007-4021, 1975.

[3] R. E. Jenkins and S. R. Pennington, "Novel approaches to protein expression analysis," Proteomics: From Protein Sequence to Function, pp. 207-224, 2001.

[4] S. R. Pennington, S. R. Wilkins, D. F. Hochstrasser and M. J. Dunn, "Proteome analysis: from protein characterisation to biological function," Trend In Cell Biology, Vol 17(4), pp. 168-173, 1997.

[5] T. J. Griffin and R. Aebersold, "Advances in proteome analysis by mass spectrometry," Journal of Biological Chemistry, Vol 276(45), pp. 497-500, 2001.

[6] J. Malone, K. McGarry and C. Bowerman, "Using an adaptive fuzzy logic system to optimise knowledge discovery in proteomics". 5th International Conf on Recent Advances in Soft Computing (RASC) 2004, pp.80-85.

[7] F. Duffes, P. Jenoe and P. Boyaval, "Use of two-dimensional electrophoresis to study differential protein expression in divercin V41-resistant and wild-type strains of Listeria monocytogenes," Appl Environ Microbiol., Vol 66(10) pp. 4318-24, 2000.

[8] S. Beranova-Giorgianni, "Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations," TrAC Trends in Analytical Chemistry, Vol 22(5), pp. 273-281, 2003.

[9] L. Pederson and B. Ersboll, "Protein spot correspondence in two dimensional electrophoresis gels," Proceedings of 12th Scandinavian Conference on Image Analysis, 2001, pp. 118-215.

[10] K. Pleissner, H. Oswald and S. Wegner, "image analysis of two-dimensional gels," Proteomics: From Protein Sequence to Function, pp. 131-149, 2001.

[11] D. Fenyo and R. C. Beavis, "Informatics and data management in proteomics", Trends in Biotechnology, Vol 20 (12), pp. S35-S38, 2002.

[12] M. Vihinen, "Bioinformatics in Proteomics," Biomolecular Engineering, Vol 18, pp. 241-248, 2001.

[13] F. Korn, A. Labrinidis, Y. Kotidis and C. Faloutsos, "Quantifiable data mining using ratio rules," The VLDB Journal, pp. 254-266, 2000.

[14] D. Hand, H. Mannila and P. Smyth, *Principles of data mining,* MIT Press: Cambridge, 2001.

[15] D. J. Rigden, "Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments," Protein Engineering, Vol 15(2), pp. 65-77, 2002.

[16] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," Proceedings 20th International Conf. on Very Large Databases, 1994, pp. 487-499.

[17] B. Mukhopadhyay, E. F. Johnson and R. S. Wolfe, "A novel pH2 control on the expression of flagella in the hyperthermophilic strictly hydrogenotrophic methanarchaeaon Methanococcus jannaschii," Proceedings Natl Acad Sci U S A, Vol 97(21), 2000, pp.11522-11527.