PERFORMING TREND ANALYSIS ON SPATIO-TEMPORAL PROTEOMICS DATA USING DIFFERENTIAL RATIO DATA MINING

James Malone, Ken McGarry, Chris Bowerman
School of Computing and Technology, University of Sunderland, St Peter's Campus, Sunderland, SR6 0DD, UK

Key words to describe the work: Data Mining, Differential Ratio Data Mining, Bioinformatics, Spatio-Temporal Data

Key Results: Differential Ratio data mining was used to perform knowledge discovery within the 2-DE proteomics data, incorporating the spatial and temporal components.

How does the work advance the state-of-the-art?: Development of data mining technique that performs automatic discovery of interesting trends within large spatio-temporal data incorporating both spatial and temporal elements, and non-spatial/temporal elements that describe the data. A measure is also introduced to evaluate and rank discoveries.

Motivation (problems addressed): Analysis of 2-DE proteomics data is presently undertaken manually since current automated techniques are unable to identify interesting trends successfully, due to the inability to incorporate the spatial and temporal elements. An automated knowledge discovery technique could prove very useful within important areas of proteomics research and other spatio-temporal datasets.

Introduction

The development of data mining techniques to efficiently and effectively analyse biomedical data is an increasingly important area of research. One such application area is that of proteomics, in which two-dimensional electrophoresis (2-DE) is unrivalled as a technique to analyse protein expression [3, 4] and is a key component of current proteomics research [1]. At present, analysis of 2-DE gel biomedical data is normally conducted manually which is both time consuming and requires considerable expertise. To date, the use of data mining to extract meaningful knowledge from such data has seen very limited success. In this paper we demonstrate the use of differential ratio data mining to perform trend analysis in such proteomics data, whilst incorporating the important spatial and temporal elements of the data.

2-DE Gel Data

2-DE is an important and popular technique used within the field of Proteomics. It is capable of separating thousands of proteins according to their electrical charge and molecular weight. Such experiments can create large amounts of high-dimensional, spatio-temporal experimental data making manual interpretation of results impractical [2]. Current data analysis methods are unable to handle and analyse such results meaningfully [5].

The data set used to perform these experiments concerns the analysis of the proteome of *Methanococcus jannaschii*, 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 spatio-temporal data set representing sampled points spanning the entire growth curve; samples were removed as the microorganism progressed through these 10 intervals – giving data for 10 time points.

Differential Ratio Data Mining

The data described was analysed using differential ratio data mining. The technique incorporates all aspects of the spatio-temporal data and was used to detect changes in proteins within the time series. Such proteins were flagged as interesting and ranked according to the amount they had altered. Such alterations include morphological variations, absence/presence of proteins over time and spatial variations.

Results and Discussion

We performed two sets of data mining experiments to automatically extract interesting trends from the biomedical data. The first was to investigate correlation of events over time. This involved calculating the total differential ratios for the data at each time point and simply determining the points at which the greatest variation occurred. This could be used to identify areas of variation, and hence interest, or could simply be used to confirm expected theories. The second was to highlight specific proteins which had altered over time and give a measure of the magnitude of variation over time to allow ranking.

Results from the first differential ratio data mining are shown in fig. 1 (variation is measured in differential ratio). The results show that at time points 2 (between time points 2 and 3) and 7 (between time points 7 and 8) we have the largest variation in terms of total differential ratio. This 2-DE experiment was designed to map three areas of change, corresponding with periods of growth of the microorganism. These are between 1-3, 4-6, 7-9 and 10. However, the data mining performed gives some further, novel insight into the data by showing that large areas of variance occur at other unexpected time points. Experts point out that, such unexpected points of variation within the proteome can be characteristic of an organism 'preparing' for the various stages of growth.

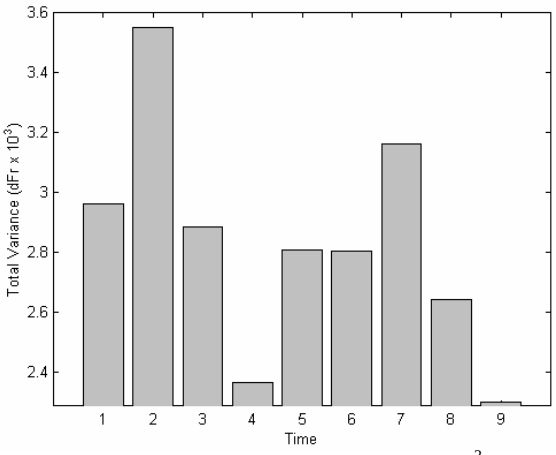

Fig 1. Total Variation over time (dFr x 10^3).

Results from the second experiment are illustrated in fig. 2. This shows the protein spots with the highest variance (in differential ratios) and at the time points at which this occurred. For example, spot no 726 has high variance at time point 6.

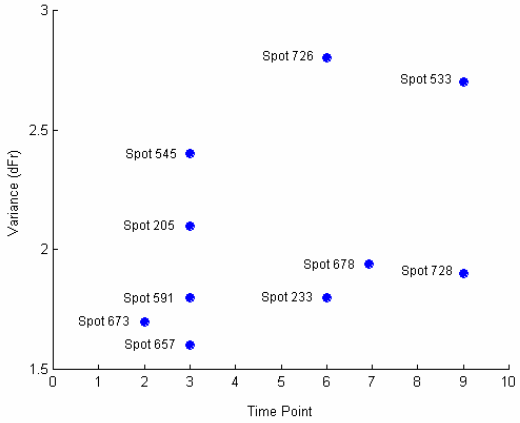

Fig 2. Ten protein spots with highest variation

Using a traditional, manual visual analysis of each spot on the gel, it was confirmed that the data mining results shown in fig. 2 were accurate. The spots identified as having variation at the time points specified were accurate. Such variation was identified visually as changes in physical attributes (such as height), changes in volume and spatial movements. The data mining results could then be easily mapped back to the physical gel (shown in fig.3) using the spot number allocated by the imaging software, allowing the experts to analyse such interesting spots using further experiments.

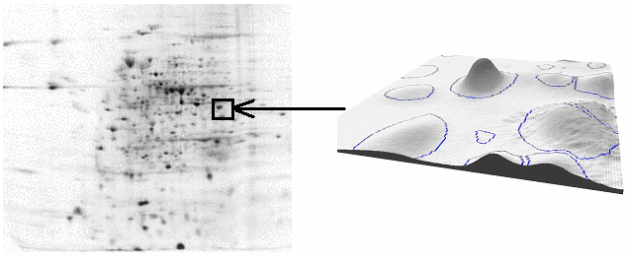

Fig. 3. 2-DE Gel and enlarged section of gel shown in 3-D

Conclusion

Analysis of 2-DE gels is an involving and impractical task for an expert to undertake. We have demonstrated the novel use of differential ratio data mining to perform trend analysis on the complex spatio-temporal data produced by such experiments. The technique is able to identify time points of maximum variance and individual spots with the most variance at particular time points.

Acknowledgements

The authors acknowledge the support of EPSRC, NonLinear Dynamics and Biswarup Mukhopadhyay of Virginia Bioinformatics Institute.

References

1. Beranova-Giorgianni, S. (2003) Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations, *TrAC Trends in Analytical Chemistry*, Vol 22, Issue 5, p. 273-281

2. Fenyo, D. and Beavis, R.C. (2002) Informatics and data management in proteomics, *Trends in Biotechnology*, Vol 20, Iss 12 (suppl), p. S35-S38

3. Jenkins, R.E. and Pennington, S.R. (2001) Novel Approaches To Protein Expression Analysis. *Proteomics: From Protein Sequence To Function*, BIOS Scientific Publishers: Oxford, p. 207-224

4. Pennington, S.R., Wilkins, S.R., Hochstrasser, D.F and Dunn, M.J. (1997) Proteome analysis: from protein characterisation to biological function. *Trends In Cell Biology*, Vol 17, Issue 4, p. 168-173

5. Vihinen, M. (2001) Bioinformatics in Proteomics.*Biomolecular Engineering*, vol 18, p. 241-248