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

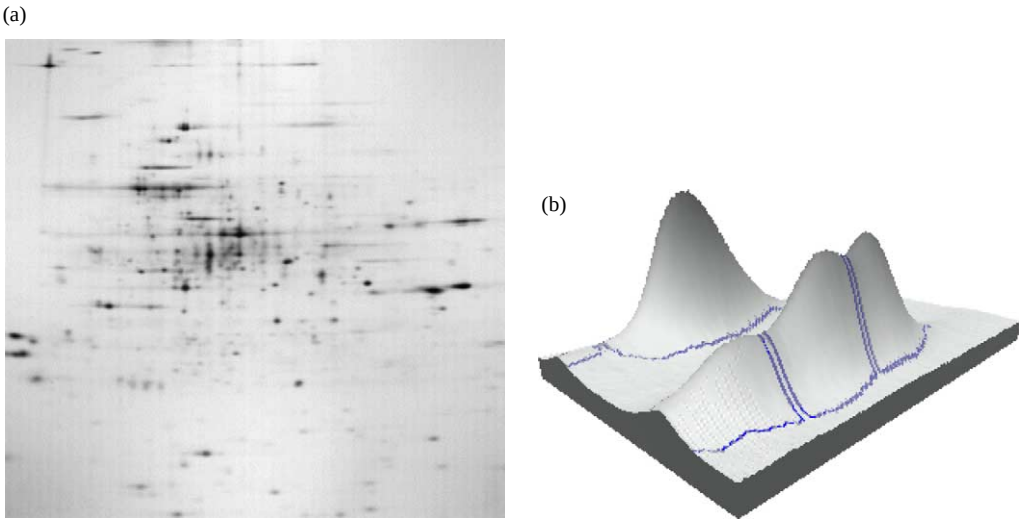

Fig. 1. (a) A single 2-DE Gel image. Each protein is described by a black spot following staining. (b) Individual spots visualised in 3-D using image analysis software. This image produced by Nonlinear's Progenesis software (Marengo et al., 2005)

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