Data and text mining
digeR: a graphical user interface R package for analyzing 2D-DIGE data
Yue Fan1,2, T. Brendan Murphy3 and R. William G. Watson1,2,*
1UCD School of Medicine and Medical Science, 2UCD Conway Institute of Biomolecular and Biomolecular Research and 3UCD School of Mathematical Sciences, University College Dublin, Dublin 4, Dublin, Ireland

Received on June 24, 2009; revised on August 20, 2009; accepted on August 21, 2009
Advance Access publication August 25, 2009
Associate Editor: John Quackenbush

ABSTRACT
Summary: 2D Difference In-Gel Electrophoresis (2D-DIGE) or 2D gel technology is being used as a routine proteomics technique for biomarker discovery. Analyzing such high-dimensional data requires multivariate analysis techniques to be applied. In addition, protein post-translational modification (PTM) information from the 2D gel data is usually overlooked. We report on an R package, digeR, with an easy to use graphical user interface for analyzing 2D-DIGE (2D gel) data. It provides a tool for visually looking for potential PTM changes from different biological states and support biomarker discovery through multivariate analysis techniques.
Availability: digeR package is freely available from the CRAN: http://cran.r-project.org/web/packages/digeR/index.html.
Contact: yue.fan@ucd.ie
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
2D Difference In-Gel Electrophoresis (2D-DIGE) has been widely used as a general proteomics approach towards biomarker discovery and better understanding of protein composition in various types of biological samples (Byrne et al., 2009; Patz et al., 2007). In addition to its relatively high sensitivity and compatibility with tandem mass spectrometry (MS), the unmatched ability to separate one particular protein into its different isoforms has made 2D-DIGE an ideal tool for studying protein post-translational modification (PTM). Here, we report on an R package, digeR, for analyzing potential protein PTM changes in 2D-DIGE (2D gel) study using spots correlation. digeR provides a graphical user interface (GUI) for the visualization of protein PTM changes between different biological states (e.g. cancer versus health).

Traditional analysis methods which focus on single protein changes usually fail to deliver a protein biomarker with high sensitivity and specificity. Panels of biomarkers are starting to gain recognition as being better in diagnosis and prognosis of disease (Rifai et al., 2006). Due to the high dimensionality of the data, in-depth analysis methods employing multivariate analysis is required for 2D-DIGE studies. Although various packages within R have provided functions that can be utilized for such analysis, biologists usually find it difficult to use these command-line-based packages. digeR provides an easy to use GUI to bridge this user gap.

2 SOFTWARE FEATURES
2.1 Spots correlation analysis
Spots correlation can be visualized on a representation of a 2D gel in the GUI. By changing the correlation coefficient threshold of a selected spot to the rest of the spots on the gel, the spots with the selected level of correlation will appear or disappear. Changes in such correlations can also be seen by switching the dataset from one biological state to another (control to disease), thus facilitating researchers to generate hypotheses for potential PTM changes resulting from the disease.

2.2 Multivariate analysis
2.2.1 Score plot 2D-DIGE data can be visualized using principal component analysis (PCA) and partial least square regression (PLSR) score plot. It supports the options to look at either top N components or pairwise comparison. Different scaling options are also provided. The score plot can be applied to 2D gel data with two or three groups.

2.2.2 Classification The classification can be performed using all or a subset of the spots through five different methods: linear discriminant analysis (LDA), principal component regression (PCR), PLSR, logistic regression and support vector machine (SVM). The classification can be applied through leave-one-out or N-fold cross-validation, or through bootstrapping. The prediction results from different methods can be assessed by ROC curves. Currently, the receiver operating characteristic (ROC) curve can only be generated to support two class classifications.

2.2.3 Feature selection Feature selection can be used to look for subsets of spots with good discrimination power for separating different groups. Four methods are available: LDA, PLSR, RandomForest and Adaboost. The selected features can also be saved and uploaded for the classification and spots correlation analysis.

2.2.4 Power analysis Power analysis can be performed for a single spot given a certain level of significance and sample size. The total gel numbers needed for an entire experiment to achieve certain...
Fig. 1. Correlation changes from control BPH (a) to Gleason 5 (b).

power and significance is calculated using the method proposed by Hwang et al. (2002).

3 EXAMPLE

3.1 Spots correlation analysis

Protein expression and function can change as a result of PTM through attaching various functional groups (e.g. phosphorylation, hydroxylation, glycosylation, etc.) to the peptide. Since the small change of mass is usually trivial to the total mass of the protein, the change of protein isoelectronic point will result in a positional shift horizontally and it can be easily detected on the 2D gel. The expression level of such protein isoforms should be highly correlated (Fig. 1) as the ratio between the different isoforms should be consistent. The changing of such correlation may result from disruptions or alterations in certain PTM pathways and their identification is of significant importance in understanding these disease processes.

A prostate cancer serum 2D-DIGE dataset (J.C. Byrne, unpublished data) is provided with the package as an example. A detailed description of this example is included in the manual that can be found using the vignette ('digeR') command. The correlation between the spots can be visualized on a simulated 2D gel. Optionally, the gel image can also be attached to the GUI for comparison. The spot indicated by the cross-hair (+) in Figure 1a (spot number 1) has been experimentally identified as haptoglobin through MS. Using the correlation analysis, we identified that the protein spots which are highly correlated with this haptoglobin protein spot were also haptoglobin. The differences in their isoelectronic points indicate that they are different isoforms of haptoglobin. By switching the dataset from the BPH group to Gleason 5 group (Fig. 1b), additional stringing of spots with a larger mass are shown to be highly correlated with haptoglobin and subsequently tandem MS experiment has shown that those spots are haptoglobin β chain. It has been reported that haptoglobin β chain is elevated in the prostate cancer serum (Seiichi Saito et al., 2008), which is consistent with the results from this correlation analysis.

4 DISCUSSION

2D-DIGE (gel) techniques have been widely applied in proteomics research, to our knowledge, there is currently no such tool available for analyzing potential PTM changes in 2D gel data using protein spots correlation. We propose a novel GUI R package, digeR, to facilitate hypothesis generation for protein PTM studies using 2D gels by visually exploring the association between spots correlation and disease states. The digeR package also provides multivariate analysis tools through an easy to use GUI for the in-depth analysis of 2D-DIGE (gel) data. digeR package can also be used to support other ‘omic’ data analysis in a similar manner and help biologist to look for panels of biomarkers which would improve the diagnosis and prognosis of the disease.

ACKNOWLEDGEMENTS

The author would like to thank Dr Jennifer C. Byrne for providing the prostate cancer dataset as an example and Dr Kenneth Bryan, Prof. Des Higgins and Dr Lorraine Brennan’s help and advice in the developing of the package.

Funding: Irish Cancer Society via the Prostate Cancer Research Consortium; PhD programme in Bioinformatics and Computation Biomedicine funded by Irish Research Council for Science, Engineering and Technology.

Conflict of Interest: none declared.

REFERENCES

Byrne, J.C. et al. (2009) 2D-DIGE as a strategy to identify serum markers for the progression of prostate cancer. J. Proteome Res., 8, 942–957.


