SameSpots

Advanced Statistical Tools for 2DGE Data Analysis

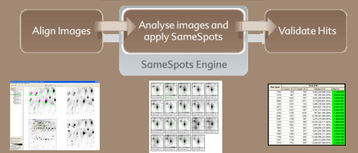
Martin O'Gorman, Andy Borthwick, Peter Kraus and David Bramwell, Nonlinear Dynamics.

Introduction

***SameSpots**: A unique technology for the analysis of single stain or DIGE 2D images allowing you to match and compare all the proteins in your experiment.*

The data produced by the SameSpots workflow is fully matched and has improved noise properties. This increases the power of univariate statistical methods and facilitates the application of multivariate statistical tools.

Method



DIGE or single stained gels are aligned before detection to correct for positional variation.

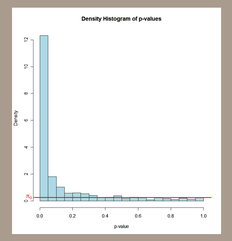
Images are analysed and identical spot outlines are transferred to all gels.

Statistics are performed automatically.

Simply work down the list and decide whether to pick or not.

False Discovery Rate and Q-values

Corrections for multiple comparisons are required to control for the number of false positives reported in a statistical analysis of your gel data. Traditionally, the Bonferroni correction has been used, but this is recognized as being too conservative. The False Discovery Rate (FDR), on the other hand, controls the proportion of false discoveries at a specified α level, e.g. α = 0.05. This gives us greater power to find truly significant features. Q-values are an extension of FDR, taking into account the proportion of null p-values amongst all tests carried out.



The density histogram of p-values allows us to estimate the proportion π_0 of null p-values amongst all tests. In our study we get $\pi_0 = 0.1596$

$$p_1 < p_2 < \ldots < p_m, m = 1405$$
$$\tilde{q}_m = p_m$$
$$\tilde{q}_i = \min(\tilde{q}_{i+1}, \pi_0 * (m/i) * p_i),\ i = m-1, \ldots, 1$$

Q-values are calculated using the m ordered p-values and take into account the factor π_0

Spots can be ordered by q-value. In this case, the q-value for a spot can be interpreted as the proportion of false positives amongst all spots with lower q-value than the chosen spot, when this spot is chosen as the threshold feature.

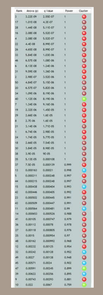


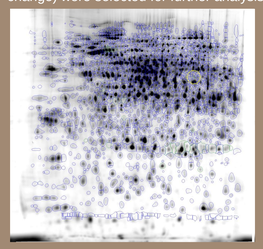
Table showing p-values and q-values per spot, coloured coded according to hierarchical clustering results

"Can I improve the power to find true discoveries while controlling against false positives?"

Progenesis SameSpots Experiment:

4 groups,
5 replicates per group

Automated analysis, no spot or match editing is performed. The top 50 (of 1405) spots (ranked by ANOVA p < 0.05 and fold change) were selected for further analysis.

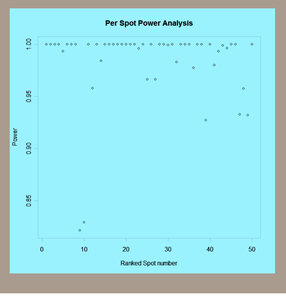


"How many replicates should I run?"
"What is the per-spot power?"

Power Analysis

The power of a statistical test can be defined as the probability that it will find a significant expression change when it exists. With more powerful tests we increase our chances of finding those proteins which are changing due to treatment or disease states. A generally accepted power threshold is 80%. The SameSpots approach has been shown to reduce protein expression variance and thus, increase power. Power analysis can also be used to determine how many replicates are required in order to be able to detect a specified change in mean expression level, were it to occur.

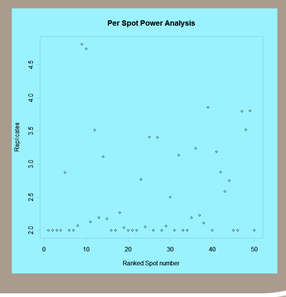
We can use power analysis with pilot experiments to help us power future studies. We can calculate the observed power for each spot and also, based on the observed data how many replicates we would need to run to achieve a target power of 80 %.



Per spot power analysis for the 50 selected spots.

The left panel shows the calculated post hoc power for each spot.

The right panel shows estimates of the number of replicates that would be required to obtain 80% power based on the measured group variance.



Conclusion

A major obstacle to the analysis of 2D gel data is due to the positional bias introduced by the electrophoresis process. Traditionally this resulted in matching difficulties across gel samples and thus impacted on any statistical analysis of the expression data.

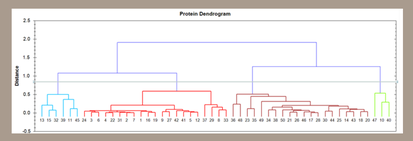
The SameSpots workflow solves these issues and leaves us with a complete data set, i.e. no missing values. We can confidently apply both univariate and multivariate statistical techniques for a complete exploration of the protein expression data. It is recommended that experimenters utilise the multiple analytics shown here as each has value in highlighting various facets of the data.

This study has demonstrated the value of applying advanced statistical tools in the quest for knowledge retrieval from 2D data.

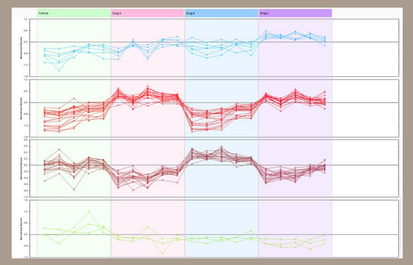
"Group my proteins together according to how similar their expression profiles are."

Correlation Analysis and Hierarchical Clustering

Traditionally, 2D gel data is analysed in a univariate manner where each protein spot is considered independently. Considering the complexity of the underlying biological system, it is unlikely that this level of independence is a true representation. We expect that many proteins spots are involved in the same systems or processes. This can be examined using correlation analysis and hierarchical clustering.



Here we see a hierarchical clustering of the spots based on protein expression correlation across all the gels. The cut-off line can be used to "cut""sub-clusters at a chosen value of distance (where distance = 1- correlation value). Spots are represented by numbers on the horizontal axis and those spots with similar expression profile will cluster together on the dendrogram.



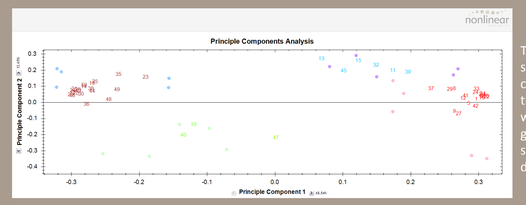
The expression profiles corresponding to the sub-clusters reflect the relatively high level of correlation in the data. The profiles show log normalised volume and are grouped according to treatments. Such graphs allow us to quickly see which spots have similar profiles and therefore may be involved in the same biological process. We can also see which spot clusters are under- or over expressed for which particular treatment groups

"How does my data cluster?"
"Are there outliers in my data?"

Principle Component Analysis

Principle Component Analysis (PCA) exploits the strong correlation in 2D gel data to allow us to capture a large percentage of the data variation in a greatly reduced number of dimensions. In this way, we can visualise both our gel samples and our spot data on, for example, a two-dimensional graph.

PCA is an "unsupervised" technique (i.e. it does not use any knowledge of the grouping of the data) and as such, is useful in determining if your gel samples have the groupings you expect or if there are outliers in these samples. Plotting the spot and sample data on a bi-plot allows us to determine which spots are over- or under expressed for which gel group.



The bi-plot shows the gel data (dots) and spot data (numbers) on a single graph. The axis represent the first and second principle components and these account for 81.19% of the variation seen in the data. The gel samples are coloured according to treatment group while the spot data is coloured according to the sub-clusters generated by the cut-off line in the dendrograph above. The closer a spot cluster is to a treatment group, the greater its influence on discriminating this group from the other treatment groups.

Presented at:

HUPO Human Proteome Organisation
US HUPO 3rd Annual Conference, Seattle 2007
March 4-8th, 2007

www.samespots.com

nonlinear dynamics