

1. Introduction

The use of statistical techniques to corroborate or contradict any given experimental hypothesis is a widely used - if not widely understood - process. For separation techniques such as 2D-PAGE and LC-MS, several methods, including PCA and cluster analysis, have become popular. However, the application of these methods has always been something of a 'black art' and the interpretation of the results not clearly understood.

Most statistical techniques can only give you reliable results when applied to a given hypothesis, for example: I expect some proteins to be affected in similar ways by different drug treatments when compared to a control sample. It is therefore important to know what questions you are trying to ask and design your experiment accordingly, so that you can pinpoint those proteins of interest in a statistically robust and repeatable way.

2. Validation

Ensuring that it is possible to extract useful information from a given data set is essential in all analysis. To this end, validation can be used to assess the technical processes employed to run and scan the physical gels, as well as examining the experimental groupings.

Validation looks at all of the available data, without any user-bias being introduced.

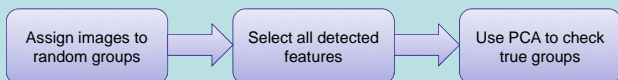


Figure 1: Process for validation

On the PCA graph, replicates should be tightly clustered together. In the example shown (Figure 2), clear grouping of the Control and two sets of treated images is evident. The PCA has also clearly indicated that the Drug B treatment and the Control seem to exhibit similar behaviour which is quite different from that of the Drug A treatment:

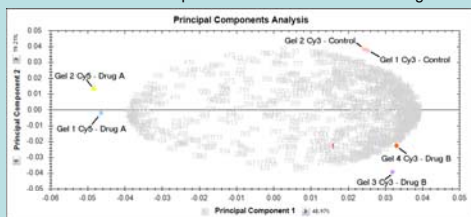


Figure 2: PCA graph showing relative positions of replicate gel images.

These clear groupings indicate that the expected groups are valid. This result has been reached without any assumptions being made about significant differences in protein expression between the groups. As the technical process of running the gels as well as the experimental design appear sound, it is possible to begin the exploration of the results.

3a. Exploration

Once gel images have been assigned to their correct experimental groups, a list of proteins ranked by significant change is obtained. Using both PCA and correlation analysis it is possible to explore this data in greater detail to find candidate proteins of interest.

Selecting only those proteins displaying a significant change (and noting that this is adding assumptions about the data, and therefore changing the answer expected from the statistical analysis), three clusters can be clearly seen in the data (Figure 3) – reflecting the original hypothesis and the current protein selection.

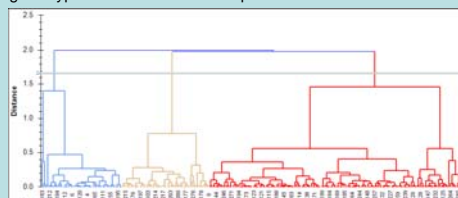


Figure 3: Correlation analysis dendrogram highlighting three distinct clusters of protein behaviour.

It is now possible to examine the behaviour of these clusters in order to isolate those proteins that are suppressed by both Drug A and Drug B. By looking at the expression profiles of the clusters shown in Figure 3, it becomes clear that proteins of the middle cluster display the behaviour of interest (Figure 4).



Figure 4: Protein expressions of central cluster from Figure 3.

Selecting to focus on only these proteins allows closer investigation on the processes happening within this group, and where these proteins appear on the physical gels.

3b. Exploration...

This approach gives 58 proteins of interest. Performing the same procedure as above on this subset of proteins allows investigation of the subtle differences in behaviour within that subset. Notice however that the question being asked has changed to: what are the differences in behaviour of those proteins that are suppressed by both Drugs A and B?

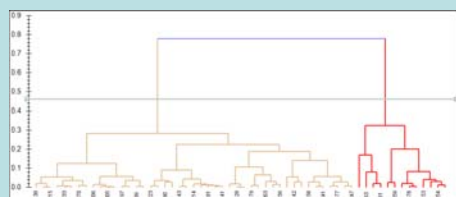


Figure 5: Correlation analysis dendrogram highlighting two main modes of protein behaviour.

Here it can be seen that there are two clear behaviour patterns. The expression profiles of these two patterns, given in Figure 6, shows the differences are between proteins more highly expressed under Drug B conditions and those that have almost equal expression with either drug.



Figure 6: Expression graphs of clusters highlighted in Figure 5.

Again, selecting those proteins which look most interesting - in this case those proteins that behave the same regardless of drug treatment - it is possible to focus on a small set of proteins following our required behavioural pattern. This process eventually yields 7 spots that exhibit this particular behaviour.

3c. Exploration...

It can be seen in the gel image in Figure 7, that the selected proteins of interest are located mainly in the same area of the gel, leading to further research into how these proteins interact with both the treatments and each other.

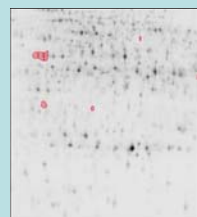


Figure 7: Gel image showing location of the 7 final proteins of interest.

The three features that are co-located at the top-left of the gel could in fact be multiple isoforms of the same protein. Again, this opens up a new avenue to explore further.

4. Confirmation

A robust method of confirmation would be to re-run the experiment, locally or in another lab, and follow the above process again. In particular, given the search is really about finding proteins that are similarly suppressed with either Drug A or Drug B, attention should be paid to the design of the new experiment to ensure its design reflects the hypothesised output. In this way, if you again get the same proteins you have far greater confidence in your experimental results.

5. Summary

This is a simple introduction to statistical workflows which can be expanded to include, for example: full technical trials and pilot experiments. However, it should hopefully cover the basic approach to managing the following questions:

- Can I ask the question? (Technical validation)
- Does my experiment match my hypothesis? (Biological validation)
- If I ask the question, what is the answer? (Exploration)
- Can I be sure of the answer I have got? (Confirmation)