

# Using multiple p-values to study missing proteomic data

Graham Horgan

Biomathematics & Statistics Scotland, University of Aberdeen Rowett Institute of Nutrition and Health  
g.horgan@bioss.ac.uk



## Background

Gel electrophoresis images are a widely used tool for studying differences in protein expression. Each sample leads to a gel image in which each spot represents a different protein, the darkness of the spot being proportional to the expression of that protein. In a typical dataset, many values will be missing because some spots were not found in some gels. This may be because the expression is low, or the spot is obscured by a larger nearby spot, or a software error means it is not found. They cannot be assumed missing at random. That some spots are missing contains information.

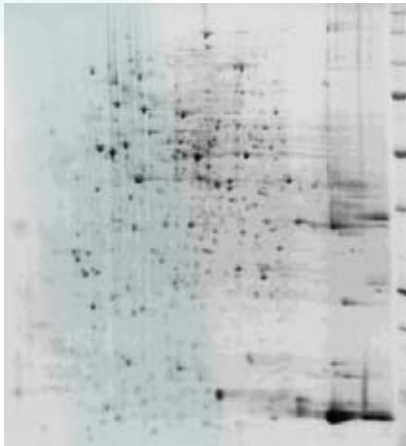


Fig 1. Example of a gel image

## How to treat missing spots.

There are several possibilities. The simplest is to ignore them and treat as if missing at random. Or we may assume that they all correspond to very low expression, and set them to zero. There are also approaches that treat them as a separate type of information, and then combine this with the data in the spots which are present. (Wood *et al*, 2004; Krogh *et al*, 2007)

## Use of multiple p-values

In gel proteomics experiments, it is typical for very many spots to be recorded, usually 300-1000. If each is tested separately for treatment effects, we have many p-values. Their distribution and behaviour is a source of information, and is already commonly used for calculating false discovery rates. We propose here to use them to study how best to treat the missing spots.

## Setting the value to assign to missing spots.

Omitting missing spots and setting them to zero are two extremes, and we might imagine that the best approach lies somewhere between these.

We first note that omitting data which does contain information should lead to an increase in type II (false negative) errors, but not in type I (false positive). So the optimal way to treat missing spots is the one which will lead to the greatest number of significant results, provided that no assumptions or approximations made cause an increase in type I errors.

We illustrate this by doing a series of tests on an example data set (not yet published), with 581 protein spots recorded in two treatment groups of 24 mice. About 2/3 of spots had some missing values, with 20.5% overall being missing. The groups were compared by standard univariate t-tests. Missing values, are replaced by  $(1 - \alpha)$  times the group mean. Setting  $\alpha = 0$  is approximately equivalent to the missing value imputation often used in small samples, whereas setting it to 1 is equivalent to replacing missing spot intensities with zero.

When this is done, the greatest number of significant spots occurs when  $\alpha$  is close to 1. ( $p < 0.1$  is chosen since these experiments are seen more as screening large number of proteins for further study than for hypothesis testing).

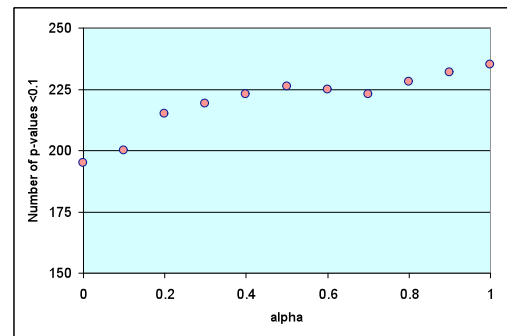


Fig 2. Number of proteins (of 581) with  $p < 0.1$ , as a function of  $\alpha$ . Omitting missing spots from the test gave 202 spots with  $p < 0.1$

## Assumptions.

A crucial part of the above approach is that the results are not affected by introduction of type I errors due to inappropriate assumptions. For example, the tests above were simple t-tests assuming a Gaussian distribution, which is not valid when considerable numbers of values are replaced with  $(1 - \alpha)$  times the group mean. We have used simulations, and randomisations of the data, to verify that altering the distribution in this way does not introduce extra type I errors.

## Further work.

The above is just a first attempt to make use of the large number of variables which are being tested. Further refinement can be done by examining tests other than the t-test, such as nonparametric or randomisation tests, which should be less sensitive to potential distributional issues. More generally, we can study the overall pattern of between-group differences in the number of missing spots, and in the means of the spots that are detected. If evidence for treatment effects is calculated separately from the difference in mean intensities of the non-missing spots, and from the different proportions of spots missing in the two treatment groups (Fig 3), then the association between them is not as clear as might be expected.

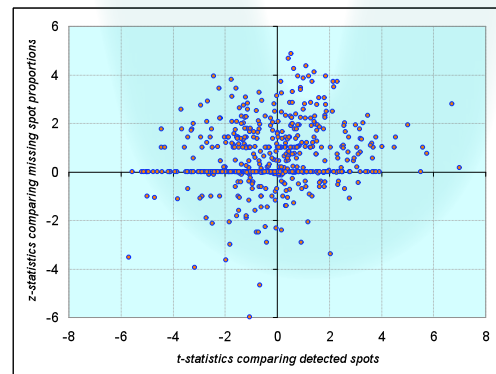


Fig 3. Plot of evidence for treatment effects calculated separately from detected and missing spots.

## Acknowledgement.

This work was funded by the Scottish Government Rural and Environment Research and Analysis Directorate. Data were obtained from studies at the Rowett Institute of Nutrition and Health, University of Aberdeen.

## References.

J Wood, R White, P Cutler (2004). A likelihood-based approach to defining statistical significance in proteomic analysis where missing data cannot be disregarded. *Signal Processing* 84: 1772-1780.  
M Krogh, C Fernandez, M Tatum, S Bengtsson, P James (2007). A Probabilistic Treatment of the Missing Spot Problem in 2D Gel Electrophoresis Experiments. *Journal of Proteome Research*, 6, 3335-3343 3335