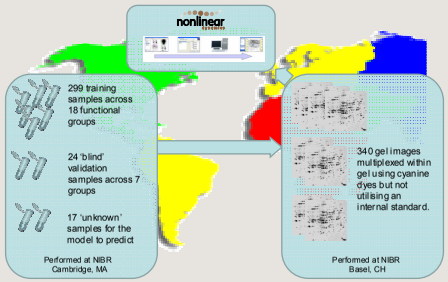
The Application of Multivariate Model Building to Derive Predictive 'signatures' from Proteomics Data

D. Bramwell[1], I. Morns[1], M. O'Gorman[1], S. Hoving[2], B. Wiedmann[2], H. Voshol[2];
[1]Nonlinear Dynamics, Newcastle upon Tyne, United Kingdom, [2]Novartis Institutes for BioMedical Research, Basel (CH) and Cambridge (MA, USA)

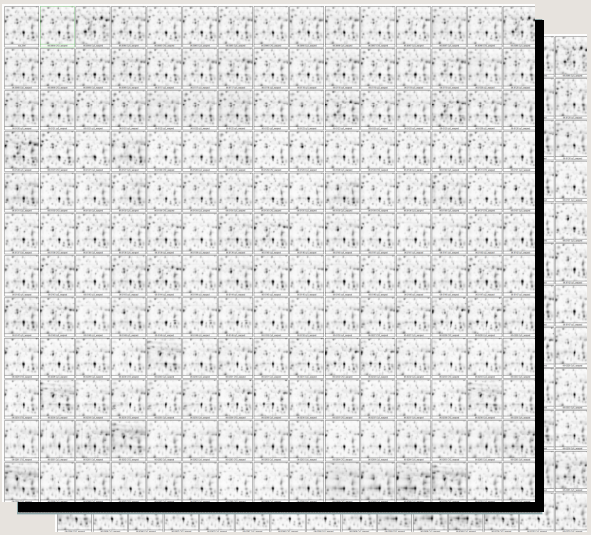
Data preparation

Research goal: To explore the process and application of predictive multivariate statistical models in complex proteomics data and explore the potential of using predictive signatures over single markers.

1) Samples prepared in the US, Gel images run in Basel. Data analysed in Newcastle



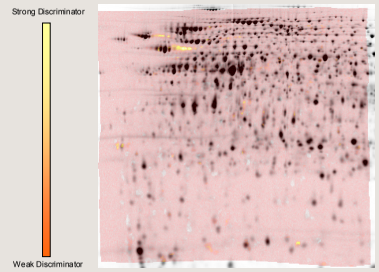
2) Gel images geometrically corrected to single reference gel at pixel levels



3) Images normalised via a third order polynomial correction based on pixel level data

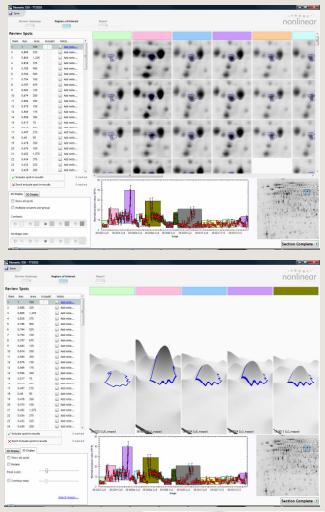
Finding the 'best' areas to use

Simply applying advanced statistical learning procedures to data sets with huge numbers of variables and few replicates (i.e most proteomics experiments) is fraught with difficulties and dangers. One primary 'quality control' step is to ensure the analysis is based on biologically relevant variables and not simply technical artefacts.



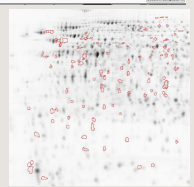
A novel pixel level, multivariate model building procedure was used. This procedure results in a pixel level heat map for each defined group. 'Hot' areas highlight the best pixels to use if you wanted to tell a given group from all of the others.

No attempt is made to segment 'spot' material from background during this analysis process. This procedure not only starts from a lower assumption basis but also provides powerful quality control aspects. Any areas predictive of a given grouping that do not overlay valid spot material signal some technical issue – without ambiguity. This is not the case if you consider spot areas alone.



Each of the 'best discriminatory' areas was manually examined both as a quality control process and also to validate it as valid 'biological' material (i.e. from a protein spot).

Up to 10 discriminatory spots were chosen based on the results for each of the groups (some groups did not produce 10 strong discriminatory spots so for these cases all were used).



At the end of this process we have 117 validated spot areas that that we can then explore further by building predictive models on this subset of data.

Predictive 'signatures'

We can now explore many questions including:

What's the difference between a 'signature' and a single marker?

A single marker takes one measure and derives a prediction from this. The prediction is usually based on an absolute threshold. As such its sensitivity is usually reduced by biological variation amongst individuals and can be strongly affected by other confounding factors. A signature is not simply using more than one marker (i.e. applying several thresholds independently to several markers and combining the results in some way). A signature combines multiple measures to produce a more robust measure that can be used in the prediction. The optical illusion to the left shows this clearly. Looking at any single spot does not allow you to tell if there is a dog there!

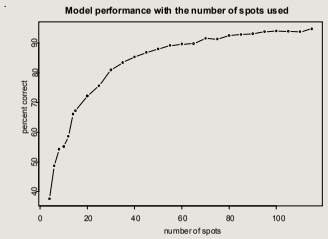
Can I build a model that perfectly groups my training data?

A nonlinear discriminant analysis model was built using the 117 selected spots and all of the training group data (not including the blinds and unknowns). This model could perfectly predict the training set data (i.e. which of the 18 groups each of the 299 samples belonged to).

Can the model based on the training data predict the 'blind' samples?

By predicting the training samples we have shown that we can predict the group of a training sample within the data set. To gain more confidence that this model has captured aspects not simply dependent on features of the training set we wanted an independent test. The blind samples were predicted and sent back to the US for verification. The model correctly predicted the groupings of the 'blind' sample images. Up until this point the blind samples had not be used in any of the analysis and their groupings were unknown to the researchers performing the analysis.

How many spots do I need?



In the current data set we can get perfect prediction performance with 117 spots. We can repeat the analysis process and by restricting the number of spots we allow the model to use we can systematically explore how prediction performance varies with the number of spots used. The graph to the left shows the trade off between complexity (i.e. the number of spots) and predictive performance.

Which are the most important ones?

As an extension to the procedure above we can be selective about the spots used to build the models and derive which ones in general allow us to build the best models. This allows us to not only select the optimal number of spots but also the best spots to use if we select this number.

Conclusions: Proteomics data, when correctly prepared, is a rich source of data that can support the application of advanced statistical learning techniques. These advanced techniques open up far more powerful approaches over normal analysis procedures. Intuitive procedures, such as blind prediction, can add new levels of understanding and confidence to the data analysis procedures.

Presented at:
ABRF 2007 Charting the Biological Roadmap
March 31–April 3, 2007
Tampa Convention Center
Tampa, FL, USA