Metabolic profiling of plant disease:  
From data alignment to pathway predictions

Submitted by Munasinhage Venura Lakshitha Perera to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Biological Sciences  
In September 2011

This thesis is available for the Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgment.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature:......................................................................................................................
Abstract

Understanding the complex metabolic networks present in organisms, through the use of high throughput liquid chromatography coupled mass spectrometry, will give insight into the physiological changes responding to stress. However the lack of a proper work flow and robust methodology hinders verifiable biological interpretation of mass profiling data.

In this study a novel workflow has been developed. A novel Kernel based feature alignment algorithm, which outperformed Agilent’s Mass profiler and showed roughly a 20% increase in alignment accuracy, is presented for the alignment of mass profiling data. Prior to statistical analysis post processing of data is carried out in two stages, noise filtering is applied to consensus features which were aligned at a 50% or higher rate. Followed by missing value imputation a method was developed that outperforms both at model recovery and false positive detection. The use of parametric methods for statistical analysis is inefficient and produces a large number of false positives. In order to tackle this three non-parametric methods were considered. The histogram method for statistical analysis was found to yield the lowest false positive rate.

Data is presented which was analysed using these methods to reveal metabolomic changes during plant pathogenesis. A high resolution time series dataset was produced to explore the infection of Arabidopsis thaliana by the (hemi) biotroph Pseudomonas syringe pv tomato DC3000 and its disarmed mutant DC3000 hrpA, which is incapable of causing infection. Approximately 2000 features were found to be significant through the time series. It was also found that by 4h the plants basal defence mechanism caused the significant ‘up-regulation’ of roughly 400
features, of which 240 were found to be at a 4-fold change. The identification of these features role in pathogenesis is supported by the fact that of those features found to discriminate between treatments a number of pathways were identified which have previously been documented to be active due to pathogenesis.
# Table of Contents

Chapter 1  **General Introduction** ................................................................. 19  
  1.1  The aim of this thesis ................................................................. 20  
  1.2  Motivation for this research ................................................................. 21  
  1.3  The Pseudomonas-Arabidopsis Pathosystem .................................. 22  
      1.3.1  Current knowledge of plant metabolomics during pathogen attack 24  
  1.4  **Mass spectrometry** ...................................................................... 27  
      1.4.1  Different types of Mass Spectrometry ..................................... 28  
      1.4.2  Sample Injection .................................................................... 28  
      1.4.3  Ionisation Techniques .............................................................. 30  
      1.4.4  Output and Isotope distribution .............................................. 31  
      1.4.5  The nature of compounds ....................................................... 33  
  1.5  **Extraction media** ......................................................................... 34  
  1.6  The MS output .................................................................................. 36  
  1.7  **Data analysis** .............................................................................. 36  
  1.8  Challenges of MS analysis .............................................................. 38  
  2.1  **Introduction** .................................................................................. 40  
  2.2  **Wet Lab Experimentation** ............................................................ 41  
  2.3  Sample Extraction ........................................................................... 42  
      2.3.1  Initial Study ............................................................................ 42  
      2.3.2  High resolution time course .................................................... 42  
  2.4  **Machine Setup** ............................................................................ 43  
      2.4.1  Nominal Mass QQQ-MS/MS (nm-QQQ-MS/MS) ..................... 43  
      2.4.2  Accurate Mass QtoF-MS/MS .................................................... 43  
  2.5  **Results** ........................................................................................ 44  
      2.5.1  Discriminate Feature extraction .............................................. 44  
  2.6  **Conclusion** .................................................................................. 46
Chapter 3  Mass Profile Alignment (MPA) .............................................. 48

3.1  Introduction .................................................................................. 49

3.2  Alignment ..................................................................................... 50

3.2.1 GeneSpring MS(GS) & Mass Profiler(MPr) .................................. 50

3.2.2 Kernel based Feature Alignment (KFA) ....................................... 52

3.3  Experimental Design ..................................................................... 62

3.3.1 Experiment 1: Toy Model ............................................................ 62

3.3.2 Experiment 2: Biological Samples .............................................. 64

3.4  Deconvolution and Data extraction using MassHunter© .................. 64

3.5  Results: Alignment Accuracy Testing ............................................. 66

3.5.1 Toy Experiments ........................................................................ 66

3.5.2 Biological Model Alignment ....................................................... 71

3.6  Conclusion ..................................................................................... 73

Chapter 4  Data Filtering and Missing Value Imputation (MVI) ............ 76

4.1  Introduction .................................................................................. 77

4.2  Experimental Design ..................................................................... 77

4.3  Method .......................................................................................... 78

4.3.1 Noise Filtering ........................................................................... 78

4.3.2 Missing Value Imputation (MVI) ............................................... 80

4.3.3 k-th Nearest Neighbour (kNN) Imputation .................................. 82

4.3.3.1 Algorithm ............................................................................. 84

4.3.4 Gaussian EM (GEM) Imputation .............................................. 85

4.3.4.1 Maximum likelihood and EM procedure ................................ 85

4.3.5 Weighted EM-kNN (WEk) ......................................................... 87

4.3.5.1 Algorithm ............................................................................. 88

4.4  Results and Discussion ................................................................. 89

6
Chapter 5  Discriminate Ion identification through density analysis and Feature Clustering........................................................................105

5.1  Introduction.......................................................................................106

5.2  Experimental Design ........................................................................106

5.3  Notation Used ..................................................................................106

5.4  Dataset Used ..................................................................................107

5.5  Transformation, Distance Metric and Resampling..........................107

5.5.1  Transformation ...........................................................................107

5.5.2  Distance Metric .........................................................................109

5.5.3  Re-sampling ...............................................................................110

5.6  Probability Density Estimation (PDE) ..............................................111

5.6.1  Histogram Approach to PDE ......................................................112

5.6.2  k\textsuperscript{th} Nearest Neighbour (kNN) ..................................114

5.6.3  Kernel Density Estimation .........................................................116

5.6.4  Bandwidth optimization .............................................................117

5.6.5  p-Value Approximation .............................................................118

5.6.6  The Student’s t-test ..................................................................120

5.7  Venn Diagrams and Cluster Analysis ..............................................121