# DRASTIC – INSIGHTS:

## Manipulating gene expression data for formation of hypotheses for plant signal transduction

A thesis to be submitted to the
UNIVERSITY OF ABERTAY DUNDEE
for the degree of
DOCTOR OF PHILOSOPHY

by

Davina K. Button

School of Contemporary Science
University of Abertay Dundee

November 2009

I certify that this thesis is the true and accurate version of the thesis approved by the examiners.

Signed [redacted] Date 11th May 2010

(Director of Studies)

## Acknowledgements

First, I would like to thank my supervisors Dr. Louis Natanson and Dr. Les Ball for all their guidance, encouragement and support throughout my PhD.

I would also like to thank Prof. Kevan Gartland, Dr. Jill Gartland and Dr. Jim Bown for their support and guidance.

I would like to give special thanks to Dr Gary Lyon from Scottish Crop Research Institute for the knowledge and inspiration he has given me during my PHD.

Most of all I would like to thank my husband Grant for his time, patience and endless love, without whom I simply would not have finished this PhD.

I would like to dedicate this thesis in memory of my parents Densie and David Button.

## List of Abbreviations

| | |
|---|---|
| ABA | Abscisic acid |
| ACT | Arabidopsis Co-expression Tool |
| ADO | ActiveX Data Objects |
| AGI | Arabidopsis Genome Initiative |
| ASP | Active Server Pages |
| At | Arabidopsis thaliana |
| BTH | benzothiadiazole |
| CEL | Affymetrix CELl Intensity File. |
| DRASTIC | Database Resource for the Analysis of Signal Transduction In Cells |
| EBI | European Bioinformatic Institute |
| ER | Entity Relationship |
| EST | Expression Sequence Tag |
| GEO | Gene Expression Omnibus |
| GO | Gene Ontology |
| GUI | Graphical User Interface |
| HCI | Human-Computer Interaction |
| INSIGHTS | Inference of cell SIGnalling HypoTHeseS |
| IUBMB | Union of Biochemistry and Molecular Biology |
| MAGE | MicroArray Gene Expression |
| MAS | MicroArray Suite |
| MGED | Microarray and Gene Expression Data |
| MGI | Mouse Genome Informatics |
| MIAME | Minimum Information About a Microarray Experiment |
| NASC | Nottingham Arabidopsis Stock Centre |
| NCBI | National Centre for Biotechnology Information |
| Os | Oryza sativa |
| POC | Plant Ontology Consortium |
| RNA | Ribonucleic acid |
| RT-PCR | Reverse Transcription Polymerase Chain Reaction |
| SCRI | Scottish Crop Research Institute |
| SGD | Saccharomyces Genome Database |
| STIF | Stress up-regulated TranscriptIon Factor |
| TAIR | The Arabidopsis Information Resource |
| TIGR | The Institute for Genomics Research |
| XML | Extensible Markup Language |

## Abstract

DRASTIC (Database Resource for the Analysis of Signal Transduction in Cells) and the INSIGHTS (INference of cell Signalling HypoTheseS) web-based suite of tools bring together data on plant responses to pathogens, environmental stresses and chemicals from refereed journal publications. Presenting these data in a unified, searchable format allows the user to extract information beyond that obtained by the authors' single genes, or clusters of similar expression patterns by browsing multiple treatments at once, identifying potential regulatory relationships between multiple treatments and genes. DRASTIC–INSIGHTS overcomes the limitations of other plant expression databases by allowing for updating of information from previous publications, by directly linking to publications and through the tracking of genes with unknown function that have the same accession or AGI (Arabidopsis genome initiative) number, which would otherwise be difficult to link between publications. Additionally, genomic, EST, Northern data and information derived from microarrays from multiple plant species are included, after human curation, to ensure accuracy and to standardize the nomenclature of data. The INSIGHTS tools encourage comparison of gene expression patterns, intelligent mining of information, testing and formulation of novel hypotheses on the complex signal transduction and response pathways used by plants. Identifying common elements in pathways affected by different treatments permits the formation of hypotheses previously opaque to the user.

Genes for proteins involved in the same signal transduction pathway are likely to be co-regulated and show the same response to a range of treatments. Thus, to find for example kinases, transcription factors and calcium-binding proteins that are in the same signal transduction pathway, expression patterns should be compared. Verification that identified genes are truly associated within signal transduction or metabolic pathways requires experimental confirmation, but the database and associated diagrams promote more targeted hypothesis formation. This type of analysis is useful in providing a framework for understanding signal transduction responses and to assist with identifying regulatory gene networks. It is also useful for finding genes associated with plant pathogen infection that are also affected by environmental stresses such as drought and cold in differing ways.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1 Introduction

## 1.1. Overview

Plants are vitally important to our global economy and make a large contribution to the food chain. Climate change and disease are causing increasing problems to the yields of food crops which is adding further pressure on the demands placed for food by an ever increasing global population.

It has been estimated by Agrios (2005) that approx 35% of the world's crop production is lost to disease, weeds and insects causing estimated losses of $550,000,000,000 per annum. Even with advanced crop protection, pesticides and highly advanced breeding programmes scientists have been unable to eradicate the hunger caused by crop diseases. Climate change is also having a negative impact on crop yield particularly in the poorest regions of the world. It has already caused a decrease in yields of most major food crops due to droughts, floods, increasingly salty soils and higher temperatures (Burness Communications, 2008).

One approach to finding solutions to these problems, and the focus of this thesis, is to develop a system to help scientists obtain a better understanding to how plants respond to biotic and abiotic threats by learning more about the signal transduction pathways and how biologists investigate and express them.

The biochemistry of what happens between a stimulus, for example a pathogen or cold (referred to here as a treatment), arriving at the outside of the plant cell, and the change in gene expression it brings about, is only just beginning to be explored. Difficulties in working with signalling mechanisms in plants include the complexity of the responses evoked by a single treatment and the number of different treatments that plants respond to (Dey and Harborne, 1998). In order to respond to a treatment, the plant cell must first be able to perceive it by use of receptors and then transduction of the signal is necessary to have effect on the gene expression. The signal transduction pathway compounds already exist in small amounts in the cell. When the pathway is triggered by a treatment(s), this causes a change to the gene expression in the cell. The genes that produce the gene products involved in the signalling pathway are up regulated which means that the signal transduction pathway is "reinforced". Figure 1 shows a very simplified version of cell signalling.

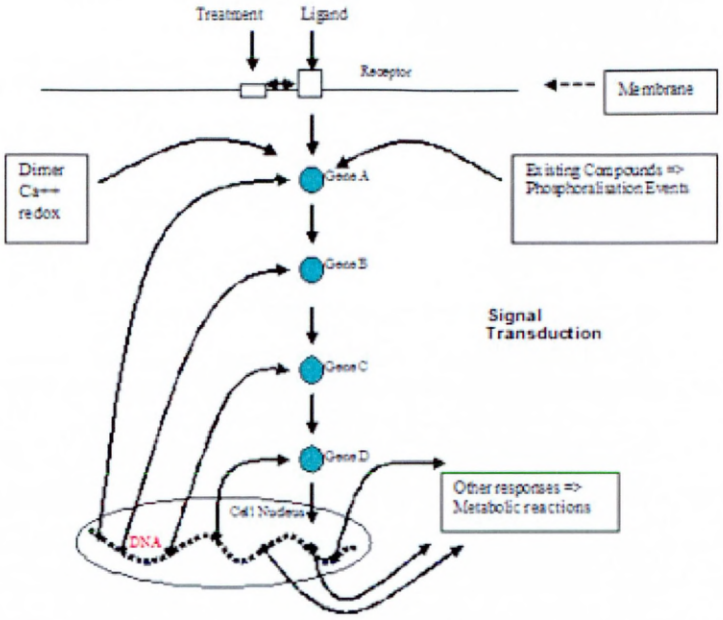



Figure 1: Simplified version of cell signalling

The majority of genes are expressed as the proteins they encode. The gene expression process occurs in two steps:

*Transcription* of the information encoded in DNA into a molecule (*DNA* → *mRNA*) and *Translation* of the information encoded in the nucleotides of mRNA into a defined sequence of amino acids in a protein ( *mRNA* → *protein*).

The EST is a piece of sequence from the transcribed and expressed mRNA which can be used to identify the gene from which the mRNA was most likely expressed. This information can also be used to deduce the protein that the mRNA would likely code for. So, because the ESTs can be traced back to the genes and the transcripts that encode the related proteins, signalling pathways may be constructed from the gene expression data obtained.

### 1.2. The Problem

Plant defence responses are extremely complex reactions and scientists are particularly interested in defence responses for example to drought or pathogen attacks as these can be devastating for crops. Plant signalling response to pathogens produces multiple signals triggering many plant response signal cascades that interact with each other in a complex pattern. The final response of the plant to infection is a consequence of the combined signals from the pathogen and the combined response by the plant. Other biotic and abiotic stresses such as environmental variation, nutritional status, fitness, etc., affecting both host and pathogen, will further modify the response, therefore it is important to consider the impact of

many types of stresses and not just one in isolation. With such a volume and complexity of data on signalling and response in multiple host-pathogen systems to assimilate, a structured framework to store information (data) and systematically build hypotheses, which can then be tested experimentally, is essential (Lyon *et al.,* 2002).

Charts of metabolic pathways in healthy plant cells are freely available (http:/www.genome.ad.jp/kegg/metabolism.html) an example of which is shown in Figure 2.

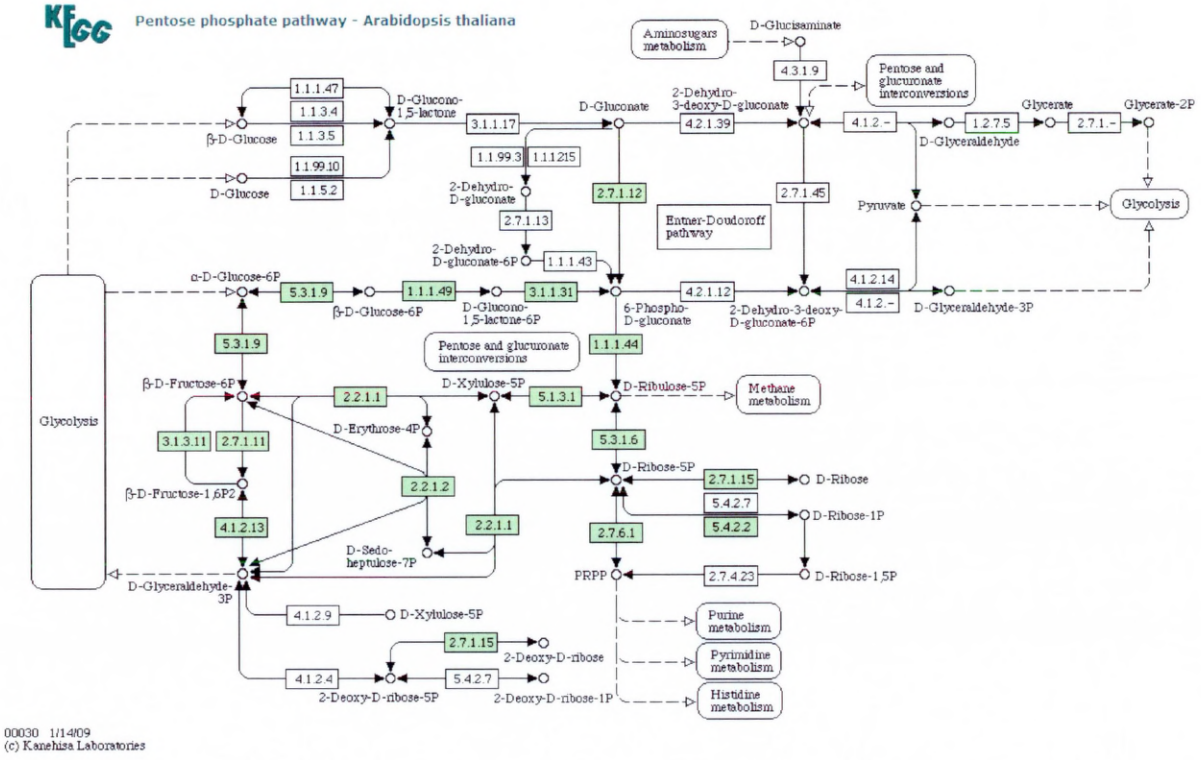



Figure 2: Pentose phosphate pathway for *Arabidopsis thaliana* from Kegg (www.genome.ad.jp/kegg) showing the types of metabolic mapping available for healthy plants.

Biologists can view many of these pathways, and investigate the enzymes highlighted in green by clicking on them to get further details. The problem for plant pathologists is that the published charts on metabolic pathways in plants only refer to healthy cells and do not include any data on signal transduction pathways. To add to the complexity, not all processes are common between plants belonging to different families and therefore it is not possible to create a single metabolic chart although it is possible to present information on secondary metabolism for a single plant species. Depicting the molecular networks involved in signalling pathways that regulate cell function has proven challenging, due to the enormous amount of information that needs to be conveyed for each participant in the network and the cross-connections between pathways (Kohn and Aladjem, 2006).

The concept for the thesis originated from work at the Scottish Crop Research Institute (SCRI) on resistance mechanisms in plants, particularly potatoes. A research project at SCRI had focused on discovery of plant genes up-regulated during resistant and susceptible interactions between potato and the foliar pathogen *Phytophthora infestans*, the stem and tuber pathogen *Erwinia carotovora* and the root pathogens *Globodera pallida* and *G. rostochiensis*. These pathogens are amongst the most economically devastating for potato, the world's fourth major crop (Birch, 2003). Little is known about the biochemical or signalling pathways in potato that are involved in resistance to pathogens. Moreover, there is even less information comparing resistance in leaf tissue with that in root. Knowledge from such comparisons could be crucial in developing broad-range plant disease resistance strategies.

It is the notion of a broad-range approach that is of relevance here. Because of the specialist nature of the research into various aspects of plant signalling, research groups will often focus on one treatment or a small aspect of signalling. Take for example research into the ethylene pathway in Figure 3 which is very specific. There is little in the way of tools that can curate and evaluate this data as a whole which is a problem as many research papers will often contain similar, very specialised results.

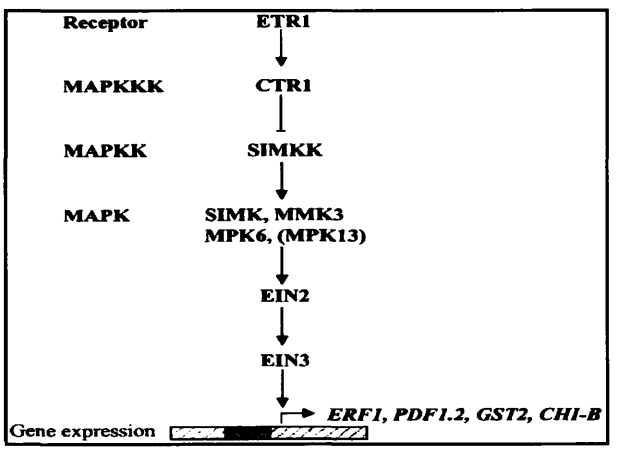



Figure 3: A proposed model of the MAPK pathway mediating ethylene signalling in plants. The histidine kinase ETR1 functions as an ethylene receptor and activates CTR1 in the absence of ethylene. CTR1 is a negative regulator of the MAPKK SIMKK and the MAPKs SIMK/MMK3 in Medicago, and MPK6/13 in Arabidopsis. In the presence of ethylene, ETR1 and CTR1 become inactivated, relieving SIMKK from inhibition. Subsequent activation of the MAPKs activates gene expression of ethylene-responsive genes via direct activation of EIN2 and EIN3 or through other factors.

During the research into disease resistance at SCRI gene expression data was collated from journals that were identified as relevant to defence signalling in plants of all species. While

the data collected was useful for the specific research projects at hand it became apparent that there was no database source available that contained this type of data (Newton *et al.*, 2002) and that the data could potentially be used as a start to create new hypotheses for plant signalling. This thesis investigates whether this data could be used to take a system biology overview. It provides researchers with a generic platform which incorporates gene expression data from multiple experiment types to enable the investigation of the overall signalling pathways rather that investigating niche, specific areas. This enables researchers to identify new areas of interest that would have previously been opaque.

### 1.3. Challenges

There are two separate challenges that must be addressed. The first is determining how to appropriately store and organise biochemical information to allow for future classification and reasoning requirements. This will be particularly important in the areas of biochemical pathways and signal transduction in cells, where the pace of data accumulation has been greatest. The beginning of this study coincided with the rise in use of microarray data and the move away from specialist one gene experiments as discussed in Chapter 4 along with the completion of sequencing projects for several plant species such as Arabidopsis and Tomato. For plant pathologists it is an exciting time but with the new knowledge comes a huge rise in the amount of available data. Plant diseases and defences have become extensively studied and a wealth of data already exists but there has been a slow pace of progress to store, categorise and mine value from the increasing data mountains on the responses of plants to pathogens and other environmental stresses. New data is also continually emerging at a rate much faster that existing data can comprehensively be studied.

The second challenge is to examine the structure of the stored data that will be specific to gene expression results for early defence signal transduction mechanisms in plants and identify ways to synthesise knowledge from this data.

### 1.4. Potential Benefits

The wealth of genomic and microarray data accumulated in the last decade holds enormous promise for understanding the molecular basis for the control of gene expression in all organisms and the functioning of biochemical pathways in all cell types. Understanding and controlling cellular responses to pathogens, disease and environmental stresses is probably

the most significant challenge and opportunity presented by the availability of this data. Important progress has been made through the completion of successive genome sequencing projects, most notably for the human, *Caenorhabditis elegans*, *Escherichia coli* and *Arabidopsis thaliana* genomes. Genomics together with advances in microarray technology provide powerful methods for the global analysis of gene expression and protein content (Forster *et al.*, 2003). There has, however, been much less progress made towards a more unified understanding of biochemical processes across species groups and between kingdoms. This has been due to the volume and complexity of the accumulated data, especially where elements of temporal or ontogenetic control are concerned, which require more sophisticated bioinformatics tools to enable molecular bioscientists and clinicians to integrate biological intuition with rigorous computational data analysis (Quackenbush, 2001). These data suggest significant commonality in key signalling pathways and regulatory networks between plant species and between plants, animals and microbes (Hammond-Kosack and Parker, 2003). Potentially generic solutions to some of the most important food–crop diseases could be found from these networks and pathways, if suitable tools to mine value from them, able to cope with large and diverse datasets, were widely available.

### 1.5. Proposed Solution

An approach to making sense of the volume of genomic data coupled with the complex datasets is to design a knowledge based framework to store information and systematically build hypotheses for signalling in a multiple plant host-pathogen system. This thesis looks at the design of a unique system which will provide a platform for plant pathologists to interrogate early defense gene expression results in a novel way (Button et al., 2005). This is the first database to enable the searching of such data and the focus has been from a system engineering view with the main aims for the study being:

1. Automate and enhance the process of information discovery for the biologist (Chapter 2)
2. Identify suitable data from different sources and different experiments (Chapter 3)
3. Examine the structure of data particularly from journals and make this more accessible (Chapter 4)
4. Provide a method to enable results from different types of experiments to be compared against each other (Chapter 4)

5. Design a data model that takes account of the various data and database standards that exist in the plant biology community (Chapter 5)
6. Improve the speed, efficiency and ability of the biologist to search for information from the gene expression results collected (Chapter 6)

Chapter 2 describes how scientists currently approach the problem of investigating gene expression results and hypotheses for signalling pathways. Chapter 3 examines the possibilities of automation and evaluates what data is currently available to mine for gene expression defence responses. Chapter 4 assesses the quality of the data and looks at how data from a variety of experiments and sources such as online journals, paper-based journals and microarray databases can be integrated to enable searches to be carried out across varying types of data. Chapter 5 builds on the results from the data evaluation and describes the development of a generic model that can hold data from multiple species and experiments. Chapter 6 describes the development of a toolset by building an intelligent and generic system for increased understanding of metabolic and signal transduction pathways based on requirements identified in Chapter 2. Chapter 7 tests and evaluates the toolset and looks at future work.

The DRASTIC-INSIGHTS enabling technology described in this thesis stimulates new interpretations of existing and emerging data, mining value through the identification of elements of commonality between different pathways and diverse organisms, whilst allowing new hypotheses for these pathways to be postulated and tested.

# Chapter 2 Background

## 2.1 Introduction

This chapter provides an overview of the existing work that is being carried out in the plant defence signal transduction pathway domain. Effects of certain treatments and environmental conditions start early signalling events which then trigger gene expression changes within the plant. Understanding signal transduction may help with problems such as disease and drought. There are several current methods used to investigate signalling in plants which are evaluated, but the main focus is on the current methods employed by biologists at SCRI and the strengths and weaknesses of these techniques. These are examined in order to meet the first aim of identifying the best way of automating and enhancing the process of information discovery for the biologist.

## 2.2. Existing Investigative Methods

### 2.2.1. Diagrammatic Modelling

Modelling is a key tool with which scientists develop an understanding of the processes involved in cell signalling and their interactions (Lyon *et al.*, 2002). Biologists use diagrammatic representations to illustrate and summarize molecular interactions between plants and treatments. Gene expression results from experiments where plants are exposed to various environmental stresses and pathogens (treatments) are used along with the biologists own knowledge and educated guesswork to "map" signal transduction pathways into diagram form.

An example of this type of approach is used by scientists at SCRI where information from their gene discovery programmes has been manually collated into a spatial diagrammatic representation of pathogen recognition in potato displayed below in Figure 4. The diagram is created by analysing curated journal data and gradually adding/modifying the diagram manually as new data is reported.

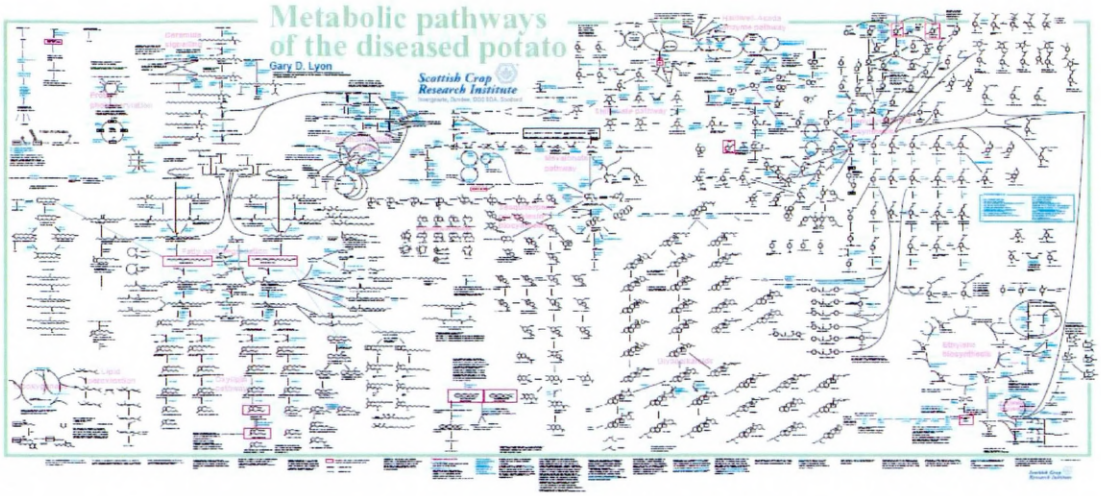



Figure 4: Metabolic pathways of the diseased potato G. Lyon, SCRI.

The diseased potato diagram in Figure 4 shows the complexity of the plant's response to infection but it also shows some of the responses which may not be so obvious when reading primary publications. Papers will report results to one or two specific treatments. When a plant becomes infected by a pathogen for example, it is likely that there will be many defence pathways triggered in respect to combination of stimuli such as the pathogen, the environment and the health of the plant. All these factors will have an impact on the end result of the defence response so it is useful to be able to look at the responses from an overall perspective rather than an isolated view. Figure 4 shows mainly the secondary metabolic reactions that the biologist has hypothesised to occur in a diseased potato. As the level of available data into early signally transduction has increased, this has enabled further research into which early signalling pathways are involved in mediating the metabolic reactions described in the above model. Figure 5 is the next step that the biologist has taken by modelling plant cell signalling in resistance and is the most relevant to this thesis.

The cell signalling diagram in Figure 5 demonstrates the complex defence mechanisms that may occur in primary signal transduction pathways when a plant is infected by a pathogen. Figure 5 highlights some of the different treatments that may come into play in addition to the pathogen as shown in box A and B such as jasmonic acid, arachidonic acid and salycilic acid among others. It has been created using the knowledge gleaned from a variety of papers about gene expression and then illustrated in an A->B->C pathway form built from hypotheses based from the literature along with the specialised knowledge of the biologist.

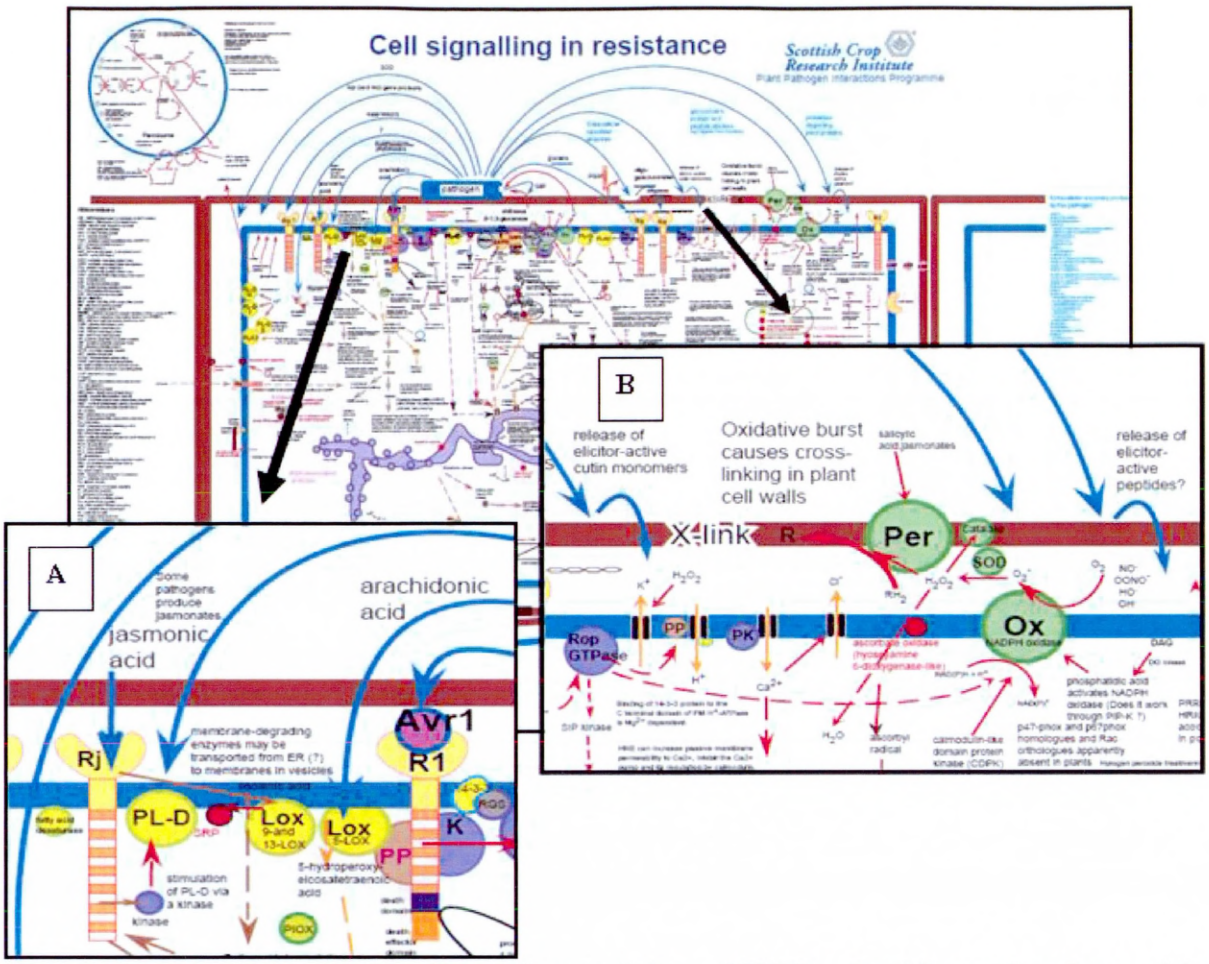



Figure 5: Complex cell signalling in resistance diagram. G. Lyon, SCRI. Box A and B are enlarged areas of the diagram that show the way in which other "treatments" may work together to create a signal cascade.

These data have been collated from results published in peer reviewed journals over a period of several years. The results of the experiments from various stress and pathogenic treatments have been manually examined and results added to the chart to build a potential overview of signal transduction cascades taking a systems biology overview. In contrast, many researchers are interested in a particular pathway or sub pathway due to the level of specialism required to study these areas and will use microarray experiments to identify gene(s) that respond to the treatment(s) that they are interested in. The cell signalling diagram is focused on signalling in resistance due to the nature of research at SCRI, but the thesis aims to broaden the scope of this and enable scientist to hypothesis on the defence mechanisms in plants to a wide range of treatments.

Both of these models are a good example of how new knowledge can be obtained by assembling data from disparate sources visually. They were assembled to assist in understanding the complex biochemical changes that can take place in diseased plants, but as

a wall chart they are fixed both spatially and temporally. Whilst the charts themselves are useful they still only touch the surface of making the information within them readily accessible and more importantly, more easily understood by non-specialists. To exploit the power of IT, ideally the information it contains should be stored in a database and drawn on demand from a set of queries.

To create these charts, in addition to the curated data, the scientist uses additional personal knowledge when sifting through results which is very specific to the specialist area that he/she has worked in. It is not possible to replicate this level of data in a database for a generic structure and so for a system to be useful it must take advantage of the gene expression data and also the processes the scientist uses to piece this data together. For each paper, experimental methods and results are well documented but there is no description in the literature of the cognitive process that the biologist uses when deciding which genes to investigate or where these may be placed in a pathway. In order to better understand this, a case study was made by observing and questioning the biologist on the steps they took when considering how to search for new data and where the results fit into the hypotheses the biologists has or will make.

### 2.2.2. Gene Expression Profiling Analysis

Gene expression profiling studies followed by functional analysis of genes altered in expression has revealed unexpectedly complex interactions between plant signalling pathways (Walters *et al.,* 2007). Microarray analysis has enabled biologists to move past the limitations of understanding single processes such as examining the behaviour of an individual transcription factor in plant defence. Instead researchers are looking at how a single component of a pathway interacts with other components of the same or different signalling pathways and how this interaction contributes to the plant as a whole taking more of a systems biology approach. Two examples of the type of results that microarray experiments can provide are described in papers by Seki *et al.* (2002) and Shenk *et al.* (2000). Seki *et al.* (2002) investigated the expression pattern of 7000*A.thaliana* genes under ABA treatments using microarray technology with the aim of identifying genes induced by environmental stimuli or stress and to analyse their expression profiles in response to these environmental signals. They used microarray experiments to test these genes with no

treatment, ABA, cold, drought and sodium chloride. They identified that there were genes that were induced by several of these treatments and deduced there is crosstalk between these treatments within the signal transduction pathways. This was discovered by comparing genes that were up or down regulated by more that five fold compared to the control set and then identifying genes that were regulated by one or more of these treatments.

Shenk *et al.* (2000) investigated changes in the expression patterns of 2,375 selected genes from *A.thaliana* after inoculation with an incompatible fungal pathogen *Alternaria brassicicola* or treatment with the defence-related signalling molecules salicylic acid, methyl jasmonate, or ethylene. They found that disease resistance, associated with a plant defence response, involves an integrated set of signal transduction pathways. Data analysis revealed a surprising level of coordinated defence responses, including 169 mRNAs regulated by multiple treatments/defence pathways. The largest number of genes co-induced (one of four induced genes) and co-repressed was found after treatments with salicylic acid and methyl jasmonate. In addition, 50% of the genes induced by ethylene treatment were also induced by methyl jasmonate treatment. These results indicated the existence of a substantial network of regulatory interactions and coordination occurring during plant defence among the different defence signalling pathways, notably between the salicylate and jasmonate pathways that were previously thought to act in an antagonistic fashion.

Hein *et al.* (2004) investigated regulation of *bmi* sequences using salicylic acid, methyl jasmonate, ethylene, $H_2O_2$, abscisic acid, wounding and a glucan elicitor. They found no single stimulus up-regulated all genes, suggesting either combinations of these stimuli, or additional stimuli, are involved in characterisation of early transcriptional changes involving multiple signalling pathways. Whereas $H_2O_2$ up- or down-regulated 17 of the transcripts detected in Northern analyses, salicylic acid stimulated only down-regulation of 5 transcripts. These are examples of the type of deductions that can be made when presented with data about multiple treatment stimuli.

There are many papers that present data comparing gene expression data from several treatments and hypothesising on signal transduction pathways based on the results. If all the data from these experiments could be compiled and tools were available to compare treatment sets to identify genes that respond similarly to treatments could this enable biologists to generate new hypothesis for plant defence signalling pathways?

### 2.2.3. Protein-Protein Reactions in Pathways

An alternative option to compiling gene expression results for investigating plant transduction signalling is to track the protein cascades by matching gene protein reactions. In the same way as enzymes can be tracked by matching the product -> substrate reactions and linking pathways using a backtracking search, if known, the protein – protein reactions could be constructed using this manner.

## 2.3. Problems with conventional methods

The diagrammatic modelling approach was a reasonably successful method until the advent of microarray data. Up until this point, biologists were dealing with papers where only one or two gene expression results were reported per paper.

When microarray technologies became widely available, this could yield 22000 results per paper. Even if only 15% of the results reported are relevant to signalling, this is an impossible number of results to manipulate without the aid of some computing assistance. Microarrays profiling analysis techniques are extremely powerful but it is frequently difficult to infer which biological pathways are activated given a list of differentially expressed genes. The biological outcome of a differentially expressed gene is dependant on the simultaneous activation of many more gene products and current knowledge of these dependences and interconnections between biological pathways is incomplete.

Other problems biologists identified with the current modelling methods are:

1. Up-dating (and adding new information) is slow and laborious.
2. There is no temporal dimension within the diagram - No account is taken of time or 'dose/amount' of response and the dynamics of the interaction are very poorly represented.
3. It is difficult to incorporate information on differential induction of certain genes in different plant tissues (e.g. roots vs. leaves).
4. The importance (and interdependence) of proteins in different intracellular locations is sometimes poorly conceptualized.
5. It is not possible to draw separate diagrams for each agonist/response as it would be too time consuming.

6. It is difficult to indicate the source of information i.e. whether it is derived from Arabidopsis or another plant, or whether it is from another eukaryote, nor the source of the publication.
7. It may include varying degrees of uncertainty ('informed guesses') that other scientists may find inappropriate or are wrong (by virtue of having not taken into account some other published information).
8. It is difficult to add information on 'unknown' ESTs.
9. It is not possible to interrogate a diagram.
10. It is not possible to add to the diagram ones own personal or unpublished data.

The process of how biologists deduce pathways is not clear and understanding this will provide insights for the development of a new system. A case study was undertaken to observe the biologist during the manual curation and assimilation of gene expression results.

### 2.3.1. Case Study: A workflow of a typical system is shown in Figure 6:

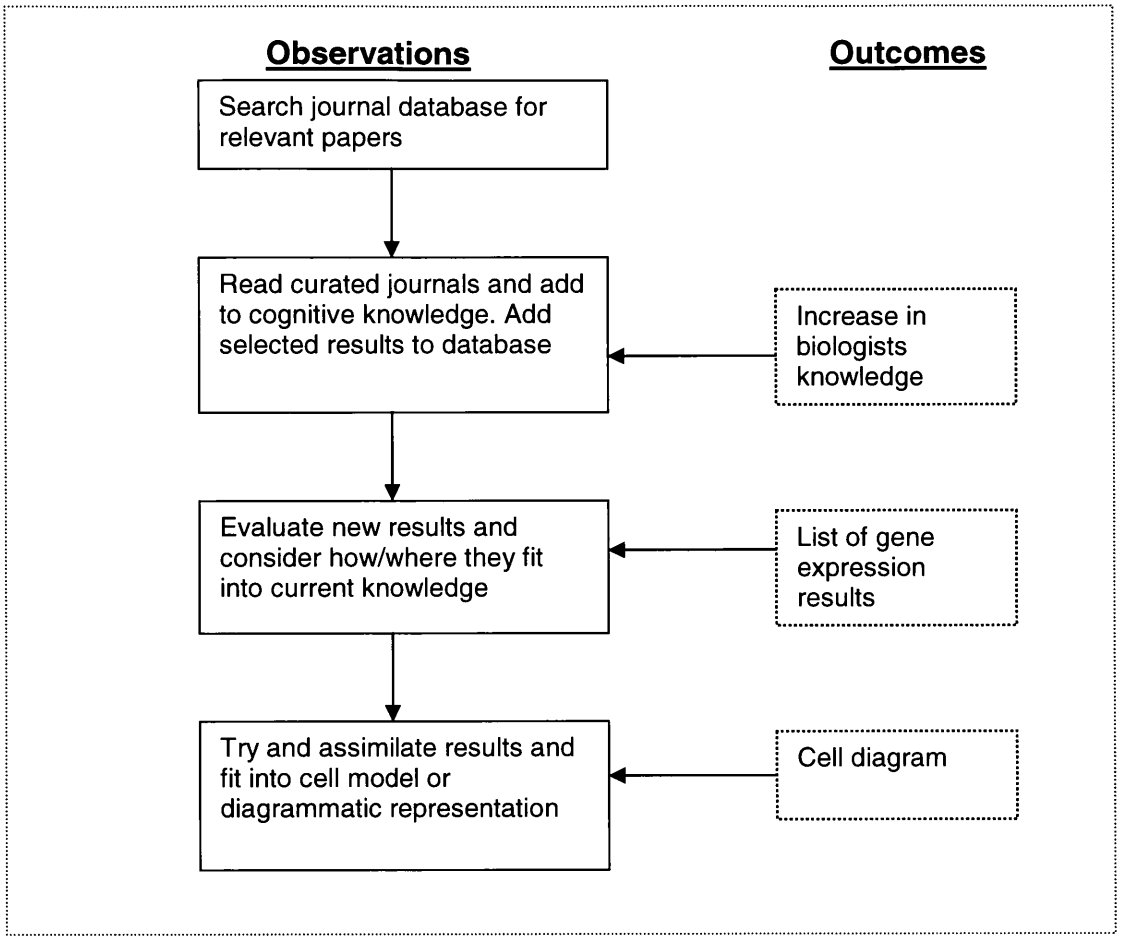



Figure 6: Workflow for signal transduction analysis

This problem can be split into two models – Curation of Data and Data Analysis as described below:

### 2.3.2. The Search and Selection of Journal Papers for the inclusion of data to the DRASTIC database workflow

Several journal databases are searched in order to find relevant journal papers for research. The primary database used is the PubMed database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed) which is a service provided by the U.S. National Library of Medicine that includes over 18 million citations from MEDLINE and other life science journals for biomedical articles. PubMed includes links to full text articles and other related resources and is used to search for papers that may contain gene

expression data that can be used. This is searched on a daily basis using the text strings "plant microarray", "Arabidopsis", "Potato", and sometimes "Rice" and "Barley".

Each of these searches produces a list of papers which are further sifted based on the contents of the title and abstract of the paper. The papers are only accepted to the next stage if they contain data pertaining to gene stress responses and the paper is from a journal with a high impact factor to ensure high quality data.

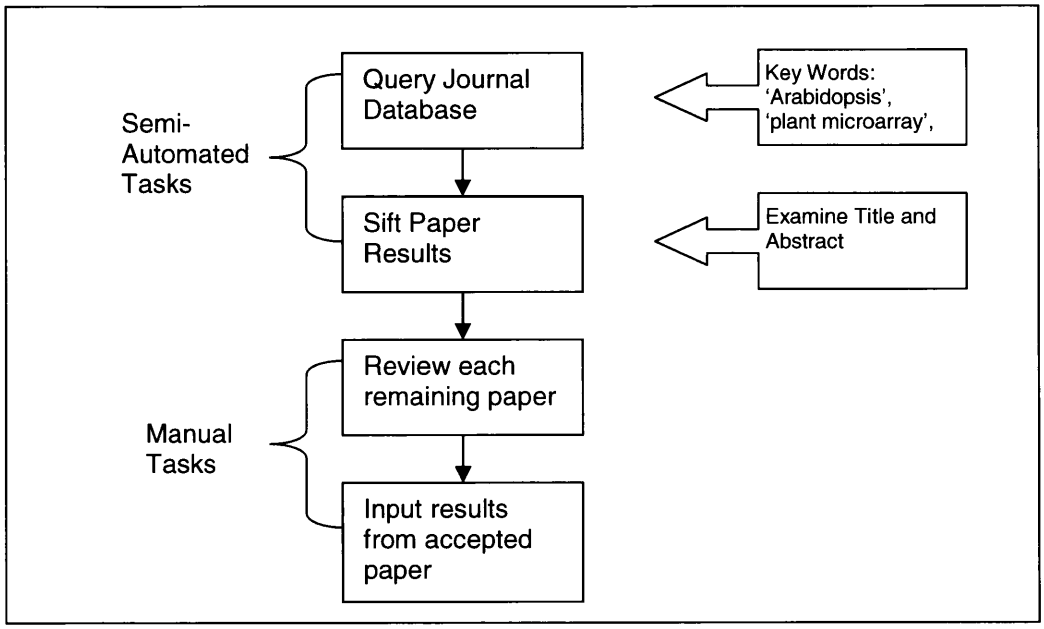



Figure 7: Workflow of paper curation

The papers that are not rejected from this preliminary sift are printed out in full. They are individually reviewed for suitability of data inclusion. The curator is looking with the "eyes of a referee" to ensure that the data is of a high quality. The criteria for this quality check are that the paper must contain enough experimental data and must include accession numbers for each result. Because the biologist is focussed on stress responses, to be included, the results discussed in the paper must relate to early responses (ideally < 1 hour or in the case of pathogens 12-24hrs after inoculation) as these would be more likely to yield information associated with signalling.

If the paper meets all of the criteria it is accepted for inclusion. All the gene expression results that are found to be 'up' or 'down' regulated in each paper are included; there is no selection of genes made by the investigator regardless of their field of interest as this could

bias the data. Because of the time constraints caused by the manual input of the data, 'no change' data is not included when looking at microarray results.

### Data Analysis of the curated data

The analysis model is much harder to define as it is not a transparent process. The observed process workflow from Figure 7 gives no clues as to the type of questions or rules that the biologist uses when analysing the data and it is difficult to relate how the actual data from the paper is converted to the cell signalling diagram.

Strategic knowledge is related to the process by which scientists use the data to construct models and generate hypotheses. In other words, gain new understanding as opposed to simply gathering the facts. These processes or strategies are not always immediately obvious, even to the scientists themselves, and careful, structured research is required to elicit them.

## 2.4. Elicitation of Requirements

The requirements that the biologists have for the system were elicited using semi-structured interview techniques and observation. Previous work carried out under a Carnegie grant demonstrated that by linking these results to a web search page on a simple database, knowledge could be elicited that would previously have been opaque to the biologist. A database approach would seem to resolve a number of the problems identified with the conventional approaches used to tackle the problems of analysing the gene expression data. (Lyon *et al.*, 2002).

In addition to the database, a system is needed to allow the biologist to interrogate the data and build on the model building strategies used to date. Such systems must maintain the stimulation of exploration and hypothesis formation of the current manual methods, while adding the precision and ease necessary for integration across diverse knowledge sources. Importantly, they must retain the links to the underlying data so that the source, quality and reliability of the data can be checked easily. There are several aspects to consider when designing a system that includes scale, scope and complexity of the data, the variability in the level of detail available, in the spatial and temporal expression of genes and in the reliability of the data, as well as differences in the language used to describe such observations.

A diagram for the requirements of the system is shown in Figure 8 and should incorporate:

- Structured database with interface to enable easy data input for all types of experimental results
- Quality/validation data checks and facility to update the data
- Set of web based tools to enable the biologist to intelligently query the data to formulate hypothesis.

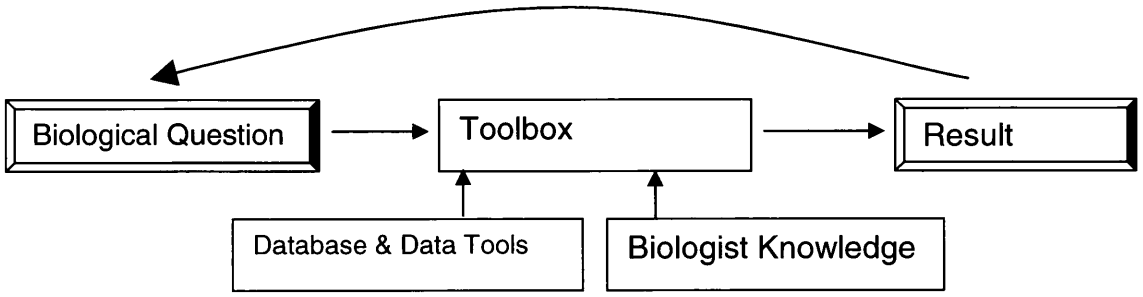



Figure 8: System Diagram

In order to identify the actual tools that the toolbox should contain, prototypes of potential systems were designed to try and capture the type of questions that the biologists were asking of the data. The toolbox should act as a knowledge support tool that will enable them to ask questions in an appropriate manner that will facilitate them to access the relevant data in the system that could be useful to them.

During the case study, the main questions that the biologist evaluated when interrogating their results are listed below and were identified as key elements to be built into the toolbox:

1. Co-regulation of genes - If two or more treatments regulate a gene then the genes may share a signalling pathway.
2. Gene regulation patterns - If a gene shares a similar regulation pattern to another gene then they may be adjacent to each other in the pathway.
3. Regulation types - Within a pathway, a treatment should either up-regulate or down-regulate all genes. If this does not happen, this could indicate a quality control issue or crosstalk in a pathway. This type of result is of great interest as if crosstalk is involved this may indicate a method of switching on/off a particular pathway. For example in Figure 9, if a pathway exists between A->B->C->D where C is up regulated, is C up regulated as a result of A->B->C->D, or is there an alternative pathway that involves C?

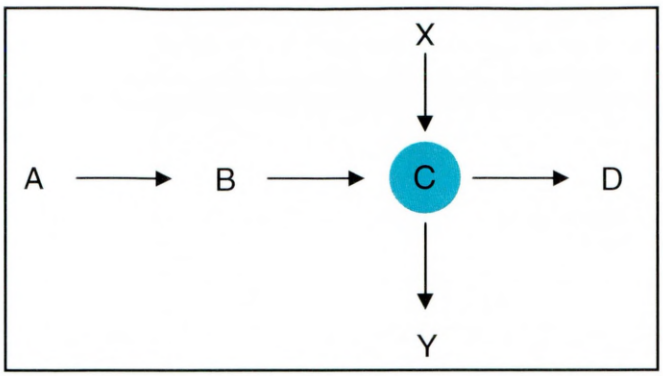



Figure 9: Crosstalk in pathway example where the letters represent genes in a signal cascade pathway.

4. Determine the number of treatments that regulate each gene - Genes that are only regulated by one treatment may be early in the pathway. Genes regulated by many treatments may be later in the pathway.
5. Grouping of treatments according to similarity of the expressed genes – This may indicate the total number of pathways.

## 2.5. Summary

The modelling methods and gene expression studies described in this chapter yield results if taken as isolated pieces of research. But, if these gene expression results can be collated into one system it will enable researchers to take a systems biology approach by curating data across the whole of the plant and may provide new insights into the signal transduction mechanisms of plants. At the moment, efforts to collate enough data to be of use are severely hampered by the sheer scale and time of the manual task and the lack of format from the analysis point of view. There are areas identified in the case studies that may be able to be automated and this is an opportunity to develop a data model and analysis tools that would assist with the hypotheses formation for signal transduction pathways which is investigated in Chapter 3.

# Chapter 3 Opportunities for Automation

## 3.1. Introduction

Chapter 2 describes how and why the biologist has collected the data. Computing has historically been used to automate tasks and there would appear to be an opportunity to do this here. Manual curation of the data from journals as described in case study section **2.3.2** is extremely time consuming and inefficient. The data is not stored in a uniform manner and therefore computational data analysis is near impossible. Database techniques and data mining are standard solutions when considering automation of data collection and analysis of data. There is currently no database that is specifically designed to collect or enable the analysis of plant defence response gene expression, however, databases are commonly used as a means to hold gene expression and associated data. This chapter will focus on the second aim of reviewing existing databases and software in the public domain that hold data relevant to this project or data mining tools. This will establish if there are any other potential sources of data that can be used in conjunction with the collated data and will provide an overview of how others have tackled the analysis of gene expression data.

## 3.2. Automation and Databases

A major problem that biologists are encountering is the overload of data. (Tian *et al.*, 2002). With new data being published in journals and databases on a daily basis, and each of these data consisting of hundreds of thousands of data points, it is not possible that the biologist can manually search and capture this data let alone identify relevancy, quality check the data and then manually input it to a data sheet before analysing and maintaining it. The required data is not readily available and can be found from different sources, in different formats in a non standardised manner and this is a key challenge for automation opportunities. Chapter 3 will investigate what tasks can be automated and how the power of computing can be applied.

## 3.3. Data Sources

Technology advances in the last few years have had a significant impact on the techniques used in signal pathway analysis for defence mechanisms in plants. Scientists have had to adapt from dealing with one or two data points from experiments to thousands of data points.

The increasing precision with which the spatial and temporal expression of genes can be observed concurrently provides vital, additional data. For example, genes that respond in the early stages of an experiment may indicate that a gene is primarily associated with signal transduction from those that are involved in secondary metabolic responses. The project aims to sift through the large amount of data and discover more about signalling pathways with the dataset that is generated. Rensink and Buell (2005) reviews plant genomics trends for microarray expression profiling resources and identifies that currently, plant gene expression data are scattered and stored in multiple databases, often separated between species, which inhibits cross-species comparisons and functional discovery.

This project endeavours to develop a data model suitable for collecting gene expression data from multiple plant species to investigate signal transduction. Plant diseases and defences have been extensively studied and a wealth of data already exists. These data suggest significant commonality in key signalling pathways and regulatory networks between plant species and between plants, animals and microbes (Hammond-Kosack and Parker, 2003). Potentially generic solutions to some of the most important food–crop diseases could be found from these networks and pathways, if suitable tools to mine value from them, able to cope with large and diverse datasets were widely available.

Before any analysis can be done, the data must be collated into a suitably sized dataset. Creating a generic database for the data is not straightforward and this chapter provides a review of existing relevant gene expression for both data content and structure.

The pace of technological advancement in this field is rapid and with this in mind, the reviews of the databases and tools considered here are reflecting the position that they were at the commencement of the project with Chapter six examining the enhancements and discoveries during the project and the impact theses have made with regards to knowledge discovery for plant signalling pathways.

### 3.3.1. Journal Data

Chapter two finds that the biologists require the information from journals as well as the data set to enable them to gain value from the data. Masys (2001) finds that there is a rich source of computer-interpretable information in published literature that describes genes and their

functions which can assist in the interpretation of gene expression patterns confirms this method.

The case study in 2.3.2 shows that the biologist searches for papers that hold new gene expression data relevant to this project. If the biologist is interested in a specific stress response such as "cold" or a particular plant species, journal databases such as Pubmed, science direct and ingenta provide a good selection of journals that are returned from a text search. The searches are designed on a text based structure and do not return useful results if an accession number or an AGI or other unique identifier is entered. This makes it difficult to search for additional gene expression data points using just journals alone.

Text-mining is a huge field in itself and much progress is being made in attempting to automate text processing which may facilitate such a search. One very interesting example of this is pubgene.org (Aubry *et al.*, 2006). Pubgene is designed to accept a gene name or identifier and it searches through Pubmed and finds literature articles where the gene is mentioned. It identifies genes that are related in the same article and builds a literature network, an example of which is shown in Figure 10. This enables the user to discover links in genes that would not be obvious by just looking at one article. The software was developed in response to the large amount of literature that is published and aims to enable researchers to identify links between genes that they may miss by simply reading through manuscripts. Gene and protein names are cross-referenced to each other and to terms that are relevant to understanding their biological function, importance in disease and relationship to chemical substances. Pubgene holds a number of known versions that the gene name is referred to as shown in Figure 10 and it is this non-uniformed manner in which genes and proteins that are referred to that is one of the key issues involved in gathering results from the literature.


Figure 10: Example of PubGene Literature Network for *A. thaliana* gene At5G52310

The reason for highlighting this particular tool is while it is much more successful that the standard text search provided by pubmed, it is still very difficult to quickly identify secondary data such as gene expression results which may be hidden in the journal body using text-mining processes and therefore some relevant articles will not be retrieved. In the example of gene At5G52310, this projects' database holds over 70 journal papers that include gene expression results for this gene whereas Pubgene has identified only 9. Additionally, while these tools are capable of returning a good selection of relevant journal articles, there is no tool that can strip the gene expression results from the article that this project is interested in and the journal articles must still be manually retrieved and examined. The next step to investigate the possibility of increasing the amount of data for this project is by reviewing the main web resources and investigating if there is any suitable data stored there.

### 3.3.2. Data Standards

Prior to examining alternative database data sources, it is important to evaluate what standards exist for the data content and format as these will be relevant to the development of the data model.

#### 3.3.2.1. Microarray Data Content Standards

Minimum Information About a Microarray Experiment (MIAME) is a set of guidelines that describes the data and metadata that authors should provide in order to reproduce individual microarray experiments. The MIAME document was first described in 2001 (Brazma *et al.*, 2001) and is a standard that helps define the level of detail that should exist. MIAME compliance is fast becoming necessary in order to publish or submit microarray data to the majority of journals and public repositories.

There are six elements that contribute towards MIAME and these are listed in Table 1

| | |
|---|---|
| 1. | The raw data for each hybridisation e.g. CEL files |
| 2. | The final processed (normalised) data for the set of hybridisations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study) |
| 3. | The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment) |
| 4. | The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates) |
| 5. | Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalogue number) |
| 6. | The essential laboratory and data processing protocols (e.g., what normalisation method has been used to obtain the final processed data) |

Table 1: Six critical elements that contribute towards MIAME

Of the six required MIAME elements listed in Table 1, points 2, 3 and 5 are very important for this project. The gene expression matrix (2) is required as the data may need to be pre-processed in order to be analysed. The experimental factors and values (3) are extremely important as the focus is on experiments which are treatment vs. control therefore the experimental treatment information is vital and sufficient annotation of the array (5) ensures that the author provides enough detail to identify the gene(s) the experiment refers to. When considering data sources, it would seem prudent to ensure that only MIAME compliant or pre-2001 data is included. Importantly, MIAME does not specify a particular format for the data; it only provides guidelines for the *content* of the data. This is a weakness as obviously the data are more usable if it is encoded in a generic manner to enable the data to be easily accessed.

#### 3.3.2.2. Microarray Data Exchange Standards

In response to the lack of format for the data requirements that the MIAME guidelines suggest, The Microarray and Gene Expression Data (MGED, 2007) society began to work on the standardization of the representation of gene expression data and relevant annotations (http://www.mged.org). MAGE, which consists of three parts, The Microarray Gene Expression Object Model (MAGE-OM), an XML-based document exchange format (MAGE-ML), which is derived directly from the object model, and the supporting tool kit MAGEstk; and MO, or MGED Ontology, which defines sets of common terms and annotation rules for microarray experiments, enabling unambiguous annotation and efficient queries, data analysis and data exchange without loss of meaning.

### 3.3.3. Microarray Data Repositories

There are many public sources of gene expression data from full text journal article resources and databases discussed in 3.3.2 such as Pubmed, Ingenta and Science Direct which index journal articles to a more direct source of gene expression data found in Microarray repositories. These are databases which can be accessed via a website and encourage researchers to submit their gene expression experiment results thus enabling all scientists to access this pooled data. There are several of these repositories, but the key repositories considered here are the Gene Expression Omnibus (GEO) database, European Bioinformatics Institute (EBI) ArrayExpress database and species-specific resources, such as NASCArrays database. These repositories typically store data for download and later analysis and are intended to act as central data distribution hubs, not to replace gene expression databases that are constructed to facilitate particular analytic methods or comparisons.

#### 3.3.3.1. National Centre for Biotechnology Information Database

The National Centre for Biotechnology Information (NCBI) hosts the GEO database which is a public repository that archives microarray and other forms of high-throughput data submitted by the scientific community. In addition to data storage, a collection of web-based interfaces and applications are available to help users query and download the experiments and gene expression patterns stored in GEO (Barrett *et al.* 2006). GEO's aim is to provide a central data distribution hub which acts as a general provider of gene expression for all scientists to use. The only restriction for microarray submission to GEO is that the data must be MIAME compliant.

GEO holds data on multiple organisms, but at present there are very few plant data sets. Of those that exist, the plant datasets tend to be mutant gene knock out experiments or time sequence experiments and these are not suitable for the purpose of this project which requires treatment vs. control arrays on wild type plant species. The database contains mathematically comparable data but for the treatment vs. control experiments, the data would still have to be pre-processed to identify genes that have responded to early signalling events.

#### 3.3.3.2. European Bioinformatics Institution (EBI) – ArrayExpress

The European Bioinformatics Institution (EBI) host a database called ArrayExpress which is a public repository for transcriptomics data aimed at storing MIAME compliant data in

accordance with MGED recommendations. The ArrayExpress Warehouse stores gene-indexed expression profiles from a curated subset of experiments in the repository, (EBI Website).

EBI's focus is on creating a database that has strict standards with regards to data content, data format and data transfer (submission and download from database).

(1) To serve as an archive for microarray data associated with scientific publications and other research,

(2) To provide easy access to microarray data in a standard format for the research community, and

(3) To facilitate the sharing of microarray designs and experimental protocols.

The ArrayExpress database is at the centre of a wider microarray informatics system at the EBI (Brazma *et al.*, 2003), which also includes the experiment annotation/submission tool MIAMExpress, data transfer pipelines from other (external) databases and tools, and the online data analysis tool Expression Profiler (Kapushesky *et al.*, 2004). Experiments are biologically related logical groupings of raw and processed data together with annotation of biological samples, the material treatment and data processing steps. Often an experiment corresponds to a particular publication. This initially sounds very promising, but the only restriction on the type of microarray data that they will accept is that it must conform to MIAME guidelines. A lot of the data are unsuitable for this project as again there is very little plant data and many different types of experiments for the more sparse plant data that it holds.

#### 3.3.3.3. Nottingham Arabidopsis Stock Centre (NASC) - NASCArrays

NASCArrays is the Nottingham Arabidopsis Stock Centre's (NASC) microarray database. The majority of the data stored is for *A.thaliana* experiments run by the NASC Affymetrix Facility. All data from the NASC Affymetrix service is made available to the public via the NASCArrays database. The structure of the database is not described in the literature and the database schema is not available but it is possible to download all of the experimental data. The data is provided by means of a subscription service called Affywatch which can be freely downloaded via a web interface or via a paid service whereby the experimental data is sent on CD. Each of the experiments provide data about the purpose of the experiment, CEL files of

the absolute experimental values and excel files of the normalised experimental values in accordance with MIAME guidelines.

For the purposes of this project, NASC houses the largest collection of gene expression data that is suitable for this project as all the data is all for *Arabidopsis thaliana* has been normalised in a unified manner and contains data that relates defence response in plants that is available from one source. Again, there is a large variation in the aims of each experiment but the NASC database contains the most data on plant gene expression from treatment vs. control experiments.

From the subscription to Affywatch 1, approximately 25 experiments were identified as suitable, each which contain either 7000 or 26000 gene expression results. With the expectation of continual releases of data along with results from the AtGenExpress Project which is a multinational effort designed to uncover the transcriptome of the multicellular model organism *A.thaliana* funded by the German Arabidopsis Functional Genomics Network (AFGN), this initially seemed to be a very interesting option for increasing the amount of data. Work was carried out to determine the feasibility of automating the processing of data to enable the gene expression results to be collected and processed for this project.

### 3.3.4. Data Source Discussion

The databases described above are the main large scale microarray databases. There are many smaller databases but these tend to provide very specific data relevant to individual research groups and there were none that were suitable for this project. In order to include gene expression data to the project there are several requirements:

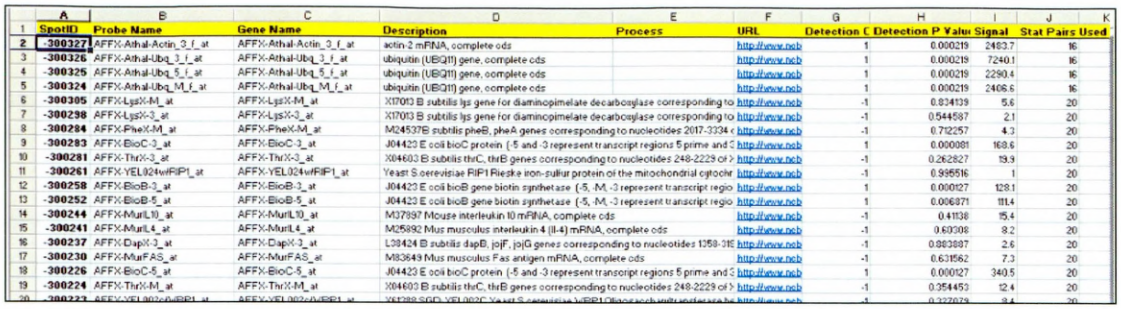
- The data must be from wild type species.
- It must be a treatment vs. control experiment that would elicit a defence response from a plant.
- It must be early from a temporal sense, ideally 0-4 hours after the treatment is applied in order to capture early signalling events
- There must be evidence to support the quality of the experiment for example peer-reviewed results.

In all of the reviewed databases, the treatment vs. control was the experiment type that appeared the least. For the project, there is a requirement to obtain a lot of different treatment expressions for each gene in order to build up enough data to make comparisons between treatments possible. However, in each repository there is a large quantity of experimental data that scientists have carried out to investigate gene function. For gene function experiments, scientists tend to examine the results for one experiment and find genes that respond in a similar manner. They can, for gene function, consider different experiments in isolation whereas, for the results the project is looking for, many experiments need to be analysed together and ideally must have all the results for each gene vs. treatment to be able to get an overview of signal transduction. Rather that looking for genes that respond in a similar manner, we are looking for groups of treatments that make a group of genes respond in a similar manner. Due to this, it would be necessary for gene data to be curated from these repositories on a regular basis rather that a one-off. To establish the feasibility of this curation, the NASCArray Affywatch was used as a sample to investigate the automation of curation from NASCArray datasets into meaningful results that would be suitable for inclusion to the project dataset.

#### 3.3.4.1. Case Study – Curating NASCArray data

This case study describes the task of selecting appropriate experiments from the NASCArray database and curating them into datasets suitable for this project. The NASCArray database enables the user to browse each experiment and all the data content adheres to the MIAME guidelines. The actual microarray value results from each experiment are stored in excel spreadsheets that can be downloaded to a local computer and then processed.

The format and content of the Excel Sheets has changed over time. The sheets from AffyWatch 1 are more difficult to process as they hold less information about each experiment in particular the probe data information is less. Figure 11 shows the Excel Sheet from one of the experiments from the AffyWatch 1 series. Table 2 describes each column heading.



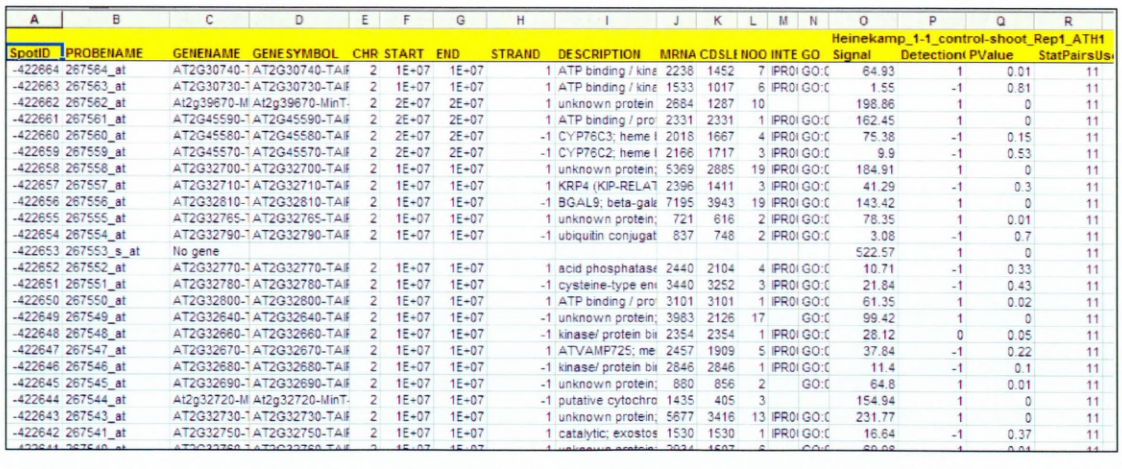
| SpotID | Probe Name | Gene Name | Description | Process | URL | Detection C | Detection P Value | Signal | Stat Pairs Used |
|---|---|---|---|---|---|---|---|---|---|
| -300327 | AFFX-Athal-Actin_3_f_at | AFFX-Athal-Actin_3_f_at | actin-2 mRNA, complete cds | | http://www.ncb | 1 | 0.000219 | 2483.7 | 16 |
| -300326 | AFFX-Athal-Ubq_3_f_at | AFFX-Athal-Ubq_3_f_at | ubiquitin (UBQ11) gene, complete cds | | http://www.ncb | 1 | 0.000219 | 7240.1 | 16 |
| -300325 | AFFX-Athal-Ubq_5_f_at | AFFX-Athal-Ubq_5_f_at | ubiquitin (UBQ11) gene, complete cds | | http://www.ncb | 1 | 0.000219 | 2290.4 | 16 |
| -300324 | AFFX-Athal-Ubq_M_f_at | AFFX-Athal-Ubq_M_f_at | ubiquitin (UBQ11) gene, complete cds | | http://www.ncb | 1 | 0.000219 | 2406.6 | 16 |
| -300305 | AFFX-LysX-M_at | AFFX-LysX-M_at | X17013 B subtilis lys gene for diaminopimelate decarboxylase corresponding to | | http://www.ncb | -1 | 0.834139 | 5.6 | 20 |
| -300298 | AFFX-LysX-3_at | AFFX-LysX-3_at | X17013 B subtilis lys gene for diaminopimelate decarboxylase corresponding to | | http://www.ncb | -1 | 0.544587 | 2.1 | 20 |
| -300284 | AFFX-PheX-M_at | AFFX-PheX-M_at | M24537B subtilis pheB, pheA genes corresponding to nucleotides 2017-3334 | | http://www.ncb | -1 | 0.712257 | 4.3 | 20 |
| -300283 | AFFX-BioC-3_at | AFFX-BioC-3_at | J04423 E coli bioC protein (-5 and -3 represent transcript regions 5 prime and | | http://www.ncb | 1 | 0.000081 | 169.6 | 20 |
| -300281 | AFFX-ThrX-3_at | AFFX-ThrX-3_at | X04603 B subtilis thrC, thrB genes corresponding to nucleotides 248-2229 of | | http://www.ncb | -1 | 0.262827 | 19.9 | 20 |
| -300261 | AFFX-YEL024w/RIP1_at | AFFX-YEL024w/RIP1_at | Yeast S cerevisiae RIP1 Rieske iron-sulfur protein of the mitochondrial cytochr | | http://www.ncb | -1 | 0.995516 | 1 | 20 |
| -300258 | AFFX-BioB-3_at | AFFX-BioB-3_at | J04423 E coli bioB gene biotin synthetase (-5, -M, -3 represent transcript regio | | http://www.ncb | 1 | 0.000127 | 128.1 | 20 |
| -300252 | AFFX-BioB-5_at | AFFX-BioB-5_at | J04423 E coli bioB gene biotin synthetase (-5, -M, -3 represent transcript regio | | http://www.ncb | 1 | 0.006871 | 111.4 | 20 |
| -300244 | AFFX-MurIL10_at | AFFX-MurIL10_at | M37897 Mouse interleukin 10 mRNA, complete cds | | http://www.ncb | -1 | 0.41138 | 15.4 | 20 |
| -300241 | AFFX-MurIL4_at | AFFX-MurIL4_at | M25892 Mus musculus interleukin 4 (Il-4) mRNA, complete cds | | http://www.ncb | -1 | 0.60308 | 8.2 | 20 |
| -300237 | AFFX-DapX-3_at | AFFX-DapX-3_at | L38424 B subtilis dapB, jojF, jojG genes corresponding to nucleotides 1358-315 | | http://www.ncb | -1 | 0.883887 | 2.6 | 20 |
| -300230 | AFFX-MurFAS_at | AFFX-MurFAS_at | M83649 Mus musculus Fas antigen mRNA, complete cds | | http://www.ncb | -1 | 0.631562 | 7.3 | 20 |
| -300226 | AFFX-BioC-5_at | AFFX-BioC-5_at | J04423 E coli bioC protein (-5 and -3 represent transcript regions 5 prime and | | http://www.ncb | 1 | 0.000127 | 340.5 | 20 |
| -300224 | AFFX-ThrX-M_at | AFFX-ThrX-M_at | X04603 B subtilis thrC, thrB genes corresponding to nucleotides 248-2229 of | | http://www.ncb | -1 | 0.354453 | 12.4 | 20 |

Figure 11: Screen shot of excel spreadsheet results from one experiment from AffyWatch 1

| Excel Column Heading | Description |
|---|---|
| SpotID | The Affymetrix ID for the probe set |
| Probe Name | The Affymetrix name for the probe set |
| Gene Name | The name of the gene |
| Description | Description of the probe from NCBI's Nucleotide database based on the Accession Number |
| Process | Description of the function of the gene |
| URL | Provides the link to look up the Accession Number for each probe from the NCBI's Nucleiotide database. |
| Detection Call | The detection call (1=present, -1=absent, 0 = marginal) |
| P Value | A p-value for the detection call |
| Signal | Normalised signal value for the probe set |
| Stats Pairs Used | The number of probe pairs used in the normalisation |

Table 2: Column headings and description for the excel spreadsheets results from the AffyWatch 1 CD's

The change in the data quantity and complexity is demonstrated in Figure 12 which shows the excel spreadsheet of a set of results from Affywatch 3 and Table 3 which describes the columns from this updated format.



| SpotID | PROBENAME | GENENAME | GENESYMBOL | CHR | START | END | STRAND | DESCRIPTION | MRNA | CDSLE | NOO | INTE | GO | Heinekamp_1-1_control-shoot_Rep1_ATH1 Signal | Detection( | PValue | StatPairsUs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -422664 | 267564_at | AT2G30740-1 | AT2G30740-TAIF | 2 | 1E+07 | 1E+07 | 1 | ATP binding / kina | 2238 | 1452 | 7 | IPR0I | GO:( | 64.93 | 1 | 0.01 | 11 |
| -422663 | 267563_at | AT2G30730-1 | AT2G30730-TAIF | 2 | 1E+07 | 1E+07 | 1 | ATP binding / kina | 1533 | 1017 | 6 | IPR0I | GO:( | 1.55 | -1 | 0.81 | 11 |
| -422662 | 267562_at | At2g39670-M | At2g39670-MinT- | 2 | 2E+07 | 2E+07 | 1 | unknown protein | 2684 | 1287 | 10 | | | 198.86 | 1 | 0 | 11 |
| -422661 | 267561_at | AT2G45590-1 | AT2G45590-TAIF | 2 | 2E+07 | 2E+07 | 1 | ATP binding / pro | 2331 | 2331 | 1 | IPR0I | GO:( | 162.45 | 1 | 0 | 11 |
| -422660 | 267560_at | AT2G45580-1 | AT2G45580-TAIF | 2 | 2E+07 | 2E+07 | -1 | CYP76C3; heme I | 2018 | 1667 | 4 | IPR0I | GO:( | 75.38 | -1 | 0.15 | 11 |
| -422659 | 267559_at | AT2G45570-1 | AT2G45570-TAIF | 2 | 2E+07 | 2E+07 | -1 | CYP76C2; heme I | 2166 | 1717 | 3 | IPR0I | GO:( | 9.9 | -1 | 0.53 | 11 |
| -422658 | 267558_at | AT2G32700-1 | AT2G32700-TAIF | 2 | 1E+07 | 1E+07 | 1 | unknown protein; | 5369 | 2885 | 19 | IPR0I | GO:( | 184.91 | 1 | 0 | 11 |
| -422657 | 267557_at | AT2G32710-1 | AT2G32710-TAIF | 2 | 1E+07 | 1E+07 | 1 | KRP4 (KIP-RELAT | 2396 | 1411 | 3 | IPR0I | GO:( | 41.29 | -1 | 0.3 | 11 |
| -422656 | 267556_at | AT2G32810-1 | AT2G32810-TAIF | 2 | 1E+07 | 1E+07 | -1 | BGAL9; beta-gala | 7195 | 3943 | 19 | IPR0I | GO:( | 143.42 | 1 | 0 | 11 |
| -422655 | 267555_at | AT2G32765-1 | AT2G32765-TAIF | 2 | 1E+07 | 1E+07 | 1 | unknown protein; | 721 | 616 | 2 | IPR0I | GO:( | 78.35 | 1 | 0.01 | 11 |
| -422654 | 267554_at | AT2G32790-1 | AT2G32790-TAIF | 2 | 1E+07 | 1E+07 | -1 | ubiquitin conjugat | 837 | 748 | 2 | IPR0I | GO:( | 3.08 | -1 | 0.7 | 11 |
| -422653 | 267553_s_at | No gene | | | | | | | | | | | | 522.57 | 1 | 0 | 11 |
| -422652 | 267552_at | AT2G32770-1 | AT2G32770-TAIF | 2 | 1E+07 | 1E+07 | 1 | acid phosphatase | 2440 | 2104 | 4 | IPR0I | GO:( | 10.71 | -1 | 0.33 | 11 |
| -422651 | 267551_at | AT2G32780-1 | AT2G32780-TAIF | 2 | 1E+07 | 1E+07 | -1 | cysteine-type en | 3440 | 3252 | 3 | IPR0I | GO:( | 21.84 | -1 | 0.43 | 11 |
| -422650 | 267550_at | AT2G32800-1 | AT2G32800-TAIF | 2 | 1E+07 | 1E+07 | 1 | ATP binding / pro | 3101 | 3101 | 1 | IPR0I | GO:( | 61.35 | 1 | 0.02 | 11 |
| -422649 | 267549_at | AT2G32640-1 | AT2G32640-TAIF | 2 | 1E+07 | 1E+07 | -1 | unknown protein; | 3983 | 2126 | 17 | | GO:( | 99.42 | 1 | 0 | 11 |
| -422648 | 267548_at | AT2G32660-1 | AT2G32660-TAIF | 2 | 1E+07 | 1E+07 | -1 | kinase/ protein bi | 2354 | 2354 | 1 | IPR0I | GO:( | 28.12 | 0 | 0.05 | 11 |
| -422647 | 267547_at | AT2G32670-1 | AT2G32670-TAIF | 2 | 1E+07 | 1E+07 | 1 | ATVAMP725; me | 2457 | 1909 | 5 | IPR0I | GO:( | 37.84 | -1 | 0.22 | 11 |
| -422646 | 267546_at | AT2G32680-1 | AT2G32680-TAIF | 2 | 1E+07 | 1E+07 | -1 | kinase/ protein bi | 2846 | 2846 | 1 | IPR0I | GO:( | 11.4 | -1 | 0.1 | 11 |
| -422645 | 267545_at | AT2G32690-1 | AT2G32690-TAIF | 2 | 1E+07 | 1E+07 | -1 | unknown protein; | 880 | 856 | 2 | | GO:( | 64.8 | 1 | 0.01 | 11 |
| -422644 | 267544_at | At2g32720-M | At2g32720-MinT- | 2 | 1E+07 | 1E+07 | -1 | putative cytochro | 1435 | 405 | 3 | | | 154.94 | 1 | 0 | 11 |
| -422643 | 267543_at | AT2G32730-1 | AT2G32730-TAIF | 2 | 1E+07 | 1E+07 | 1 | unknown protein; | 5677 | 3416 | 13 | IPR0I | GO:( | 231.77 | 1 | 0 | 11 |
| -422642 | 267541_at | AT2G32750-1 | AT2G32750-TAIF | 2 | 1E+07 | 1E+07 | 1 | catalytic; exostos | 1530 | 1530 | 1 | IPR0I | GO:( | 16.64 | -1 | 0.37 | 11 |

Figure 12: Screen shot of excel spreadsheet results from one experiment from AffyWatch 3

| Excel Column | Description |
|---|---|
| SpotID | The Affymetrix ID for the probe set |
| ProbeName | The Affymetrix name for the probe set |
| GeneName | The name of the gene, usually the AGI code |
| GeneSymbol | The AGI code for the gene |
| Chromosome | Which chromosome the gene is on |
| Start | The start position of the gene on the chromosome |
| End | The end position of the gene on the chromosome |
| Strand | Which strand of DNA the gene is on |
| Description | A description of the gene |
| MRNALength | The mRNA length |
| CDSLength | The length of the coding sequence |
| NOOFEXONS | Number of exons in the gene |
| InterPro | The INTERPRO accession number for the protein. This is a database of protein families, domains and functional sites |
| GO | The gene ontology numbers for the gene |
| Signal | Normalised signal value for the probe set. |
| Detection Call | The detection call (1=present, -1=absent, 0 = marginal) |
| p Value | A p-Value for the detection call |
| Stats Pair Used | The number of probe pairs used in the normalisation |

Table 3: Column headings and description for the excel spreadsheets results from the AffyWatch 3 CD's

The new format includes several new data types and is much easier to analyse because much more of the required data is there so there is no need, for example, to identify the gene AGI number as this is now provided.

With the majority of the releases of the Affywatch series, the excel sheets tend to change as more data is included or the format modified. This is a problem for developing software to automatically gather the excel data, process it and add it to the project database because if the underlying structure significantly changes on a regular basis, writing generic software is not possible. This highlights the changeability of both the data content and structure.

For the experiments that are suitable for inclusion, the data has to be further processed as each experiment is treatment vs. control so it is not the individual value results that are of interest but the difference in gene expression between the treatment and the control. In the NASCArray database, the results for the treatment and the control are presented in separate excel files. Both these files must be processed and once this difference is obtained, the fold value increase/decrease for each probe in the experiment needs to be calculated. It is now very common for experiments to have three or more replicates for quality purposes and this adds further complications for the calculating the fold value. There is no software available that automates this so a bespoke piece software has been developed for this case study.

All NASCArray data is derived using Affymetrix Microarrays and the results from the experiment are calculated using MAS5.0 Software which is commercial software from Affymetrix. This provides call information, signal values and normalises the data so that it is possible to compare values across different microarray experiments.

The call information provides a Present (P), Absent (A) and Marginal (M) call for each of the probes on the microarray. A present call indicates a probe that is considered as expressed, an absent call is considered as not expressed and probes with a marginal call are possibly expressed. Because there are two experimental results (the treated sample and the control sample) it is possible that a probe had a present call in one experiment and an absent call in another. To process the NASC data, consideration needs to be given as to how to use the call information provided. Table 4 lists the nine possible combinations:

| **Control Call** | **Treated Call** | **"Combined" Call** |
|---|---|---|
| Absent | Absent | AA |
| Present | Absent | PA |
| Marginal | Absent | MA |
| Absent | Present | AP |
| Present | Present | PP |
| Marginal | Present | MP |
| Absent | Marginal | AM |
| Present | Marginal | PM |
| Marginal | Marginal | MM |

Table 4: Possible Call Combinations from NASC dataset

Given the above combinations, the interesting expression data (for the purposes of signalling data), would be genes that have a combined call of AP, PA and PP. Genes that go from Absent in the control to Present in the treated set and vice versa indicate a definite change in gene expression. An increase/decrease in the fold value for a gene expression that has a present call in both experiments is also valid. A gene that is AA (Absent in both the control and the treated set) is not interesting as it simply means that the gene is not expressed in either sample so these can either be removed from the data set or treated as a no change result. In order to maintain quality, combined calls that have a marginal call in them should be regarded as "unsafe results" and removed from the data set.

Once the data rejected from the call result has been removed, the final task is to determine whether each probe's expression increased, decreased or did not change. This is decided by the fold value – if it has increased or decreased by more than two fold then it is classed as up

or down respectively with the remaining results are classed as no change. These were the final results that were included into a database for analysis.

### 3.3.4.2. Dataset Comparison

The number of results from this case study that were suitable for inclusion for the project were disappointingly low. Many of the experiments were rejected as they did not meet the criteria for the temporal or treatment conditions. Of those that were included, many results had to be disregarded as the quality of the call was ambiguous and so after much effort, there were few experimental conditions to compare and even fewer gene results that had a value for more that two of the conditions.

This case study demonstrates the problems that would be encountered when using microarray results from any of the databases described here. The software used for collation and processing would have to be modified continually and this is simply with results from one microarray method. The quality and fold factor decisions are ambiguous and user driven rather than peer reviewed which could introduce errors into the database and the final results are sparse with very few treatment conditions being available. Using the original method of collating journal data which has already had the data outcomes and fold cut-off values decided and reviewed would seem a sensible option.

### 3.1.4.3. Data Structure Comparison

The structure and accessibility of the NASCArray data did hamper the data collation as the data was only available via excel spreadsheets and the format was changeable. The structure that GEO uses to store the data is unconventional in normal database usage as it uses a primary and secondary database to handle the data. This is due to the different formats that microarray experiment results are presented in. This flexibility is largely attributed to the fact that tabular data are not fully granulated in the core database but instead are treated as plain text, tab-delimited tables that may contain any number of rows or columns. The primary database has no knowledge of these tab-delimited tables some columns reserve special meanings and data from selected fields are extracted to secondary databases and used in query and analysis applications, (Barrett *et al.*, 2006). EBI differ again in that they are very focused on software interoperability which was very difficult in the case study of NASC and they focus on the MGED standards. The EBI ArrayExpress uses the MAGE (Microarray gene

expression) object model (MAGE-OM) and MAGE markup language (MAGE-ML) to encode all MIAME required information. The MGED ontology (Stoeckert and Parkinson, 2003) defines sets of common terms and annotation rules for microarray experiments aiming to reduce ambiguous annotation and promote efficient queries, data analysis and data exchange.

#### 3.1.4.4. Data Source Summary

As the NASC data has demonstrated, gathering data from public repository databases is difficult and determining how to calculate the quality of the data and decide if it is possible to compare is ambiguous. These problems stem from commercial vs. research data, lack of standardisation with format and structure and changeability of the data as technology and knowledge advances. This makes it difficult to adopt a current data model or structure. A proprietary data model that is built with consideration of the likelihood of future changeability must be developed and data for this project should continued to be curated from peer-reviewed sources until a uniform method of data collation can be identified.

## 3.4. Data Analysis Tools

There is a wealth of microarray data in existence and there are many different analysis tools available to assist biologists with various aspects of bioinformatics investigation. The number of review articles on gene expression technology exceeds the number of primary research publications in the field but there are a limited number of efficient publicly available tools for data processing and analysing in context of existing knowledge. This is due to the lack of consensus on how to compare results using different technologies and the number of different questions that biologists wish to ask of the available data (Bassett *et al.,* 1999). This section investigates the characteristics and attributes of a selection of these tools that are used to examine gene expression data in order to assist in the development of analytical tools for this project. There are many software tools available ranging from commercial packages to freely available web tools. This section provides an overview of a selection of tools and the types of analysis they provide.

### 3.4.1. TAIR

Two very key plant resources that can be considered as knowledge bases are The Arabidopsis Information Resource (TAIR) and The Institute for Genomics Research (TIGR). TIGR is now part of the J. Craig Venter Institute and houses plant genomics databases for several plant species.

TAIR provides a comprehensive resource for the scientific community working with *A.thaliana*. TAIR is responsible for the annotation of Arabidopsis having taken this over from TIGR. TAIR does provide some microarray data but this is available in tab delimited format rather than a searchable database. However, the site does provide some very interesting generic tools that can be used with expression data. The tools selected for review are

- Java TreeView: an open source, cross-platform gene expression visualization tool for interactive display of clustered microarray data, similar to Eisen's TreeView program. This uses hierarchical clustering algorithms to identify similar gene expression patterns across a microarray experiment and group these together in a dendogram.
- AraCyc: Arabidopsis biochemical pathways visualization and querying tool. Visualisation tool that demonstrates how a list of genes can be transformed into a pathway. Due to the level of data available, this tool focuses on metabolic pathways but will be interesting to see how this type of technology advances with a view to adding temporal gene expression data to the pathways.
- Chromosome Map Tool: provides chromosome maps of the Arabidopsis genome based on a list of Arabidopsis genes entered by the user. This is a generic visualisation tool that does not deal with gene expression, but enables a list of genes to be viewed on a chromosome map.

### 3.4.2. GEO

The GEO database enables users to mine the gene expression profiles using tools described by Barrett *et al.* (2006). The more relevant tools are:

- Cluster heat maps, which enable the user to select from hierarchical and K-means clustering algorithms to investigate microarray experiments

- Query Subset A vs. B, which assists with the identification of genes that display differences in expression level between two specified sets of gene expression profiles
- Subset effects is a feature that retrieves all experiments that are flagged as having a specific experimental type for example 'age' or 'strain'
- Profile neighbours: returns a list of genes that show a similar expression pattern within a given DataSet.

These tools enable the user to select datasets based on search queries, and allow the user to mine through the thousands of pieces of data to home in on experiments that may be of interest. These searches are very general in nature and in order to drill deeper into the experimental results, the user is required to download the experimental datasets of interest and manipulate the data themselves.

### 3.4.3. NASCArrays

The NASCArrays database that was discussed in section 3.3.3.3 has developed tools that can be used to analyse data in the database. The data mining tools consist of

- Gene Swinger – finds experiments that show high variability of a chosen gene,
- Spot History – Histograms of chosen gene,
- Two gene scatter plot – plots the signals of two chosen genes,
- Bulk gene download – enter up to 300 AGI codes and download the signal values over all experiments,
- Super bulk gene download – which is a file containing all genes over all experiments suitable for clustering.

The tools are aimed to provide scientists with a general overview of the data and enable the user to find information about a set of genes across all experiments, but there is no ability to be selective of the experiments included.

### 3.4.4. Genevestigator

The Genevestigator (Zimmerman *et al.* 2004) tool is a data analysis tool rather than a data repository although it uses an in-house database of approximately 2260 microarrays to run

the analysis tools from. It is much more specialised in that it has developed web tools specifically aimed to enhance gene function discovery for *A.thaliana* biologists and so has a more targeted audience than GEO or NASCArrays. Genevestigator uses the raw experimental Affymetrix microarray data from NASCArrays database and processes it to enable data analysis to be carried out in more depth that the NASCArray tools allow. The key tools that are provided via a web-browser are:

- Digital Northern which retrieves the signal intensity values for up to 10 genes input by the user across a set of experiments again chosen by the user,
- Gene Correlator allows the comparison of signal intensity values of two genes across chosen experiments and
- Gene Atlas tool provides the average signal intensity values of a gene of interest across chosen experiments.

The output from these tools tends to be visual in the form of graphs or tables. Only data from one type of microarray can be compared and there is no linking data from the literature but the user can find the original experimental values and MIAME data from the NASCArray database.

### 3.4.5. ACT

The Arabidopsis Co-expression Tool, ACT, ranks the genes across a microarray dataset according to how closely their expression follows the expression of a query gene (Jen *et al.* 2006). The main tools used are:

- Co-expression analysis over available array experiments which shows Pearson Correlation Coefficients for a probe selected by the user,
- Co-correlation scatter plot (2-D Pearson Correlation Coefficients) which shows Pearson Correlation Coefficients for two probes and
- Clique Finder is a tool to find clusters of closely-associated probes within the Pearson correlation coefficient ranked list for a given probe

The dataset is created from the same data as Genevestigator but has been processed differently and is again a data analysis tool rather that a repository. A database stores pre-calculated co-expression results for approximately 21,800 genes based on data from over 300

arrays. ACT's focus is to provide tools to enable users to analyse how gene expression changes with respect to all the other genes on the array and is again based solely on Arabidopsis data. The tool aims to demonstrate novel biological relationships underlying the observed gene co-expression patterns and enable the testing of hypotheses on gene function.

## 3.5. Data Analysis Discussion

From the above review it is shown that there are many different types of tools with different characteristics – tools attached to repository databases, generic tools that can be applied to any dataset with a specific format, data mining tools, website based tools attached to in-house databases and visualisation tools.

### 3.5.1. Data Analysis History

Modern biology has shifted from "one gene" approaches to methods for genomic-scale analysis like microarray technology and in response to this, different analysis techniques have been developed in order to handle the new levels of data. Eisen *et al.* (1998) were instrumental in developing clustering methods and software that enable biologists to analyse thousands of data points and data mine for similarities. Statistical and data mining techniques are now extremely prevalent for the majority of tools including even web-based ones due to the increasing server processing power. As the available data increased, repositories and knowledge bases began to be created, curating data such as TAIR which offers a wide range of data about Arabidopsis with a variety of tools that the user can enter their own data into and GEO and NASCArrays which provide tools aimed at a general audience to simply provide a way of an overall sift through and download the data. Biologists began to embrace the potential of the available data and tailor tools to their own specialism, inducing an explosion of small tools that used pre-processed data from these repositories were developed. These tend to be very specific tools specialised for a particular biological area or species such as Genevestigator and ACT.

### 3.5.2. Data Analysis Characteristics

From the tools described, there are several common characteristics that are shared by a majority of the tools. The tools are web-based and interactive. They attempt to provide the biologist with a method of hypothesis testing and data discovery using data mining

techniques and a large proportion of the tools use visualisation of data to enable the biologist to gain further insight to the results of their search.

### 3.5.3. Data Mining and Hypothesis Testing

The biologists use hypothesis testing to investigate signal pathway transduction where the investigator tests an idea against a body of data to confirm or reject its validity. This will commonly raise new questions that can be tested against the data. This type of software tool can be developed by identifying the common questions that biologists ask of the data and modelling these as demonstrated by the ACT tool.

In addition to this, a further goal must be to include exploratory analysis to find patterns in the data that are not predicted by the biologist's current knowledge or pre-conceptions. Data mining tools such as the clustering tools provided by GEO provide a method of achieving this by providing tools to facilitate large scale interpretation of biological data in "batch" mode. However, such tools often leave the investigator with large volumes of apparently unorganized information and unable to further drill down into the data results. Genevestigator and ACT both use clustering methods to find genes that have a similar expression pattern to a selected start pool of genes. One key difference from the tools described here for this project is the data that the project is using will be binary values of either up or down. This will have a large impact on the design of mining tools as a large number of data mining tools work based on a numeric distance value which will not exist within this projects dataset.

### 3.5.4. Visualisation

Traditional formats of information presentation such as text and tables of data force human analysts into a harder mode of information processing by forcing humans to rely extensively on memory. Visualisation capitalises on cognitive strength as humans excel at processing visual data. Figure 13 shows an example of a list of gene expression results in tabular form. Figure 14 is a Venn diagram displaying the same data but showing the number of genes regulated by each treatment combination. The Venn diagram is much more informative and information can be collected at a glance rather than sifting through lists of gene expression results. This is one small example of how visulisation can produce an instant impact and it is a theme that will be given a high priority when designing tools for this project.

| Name | Gene Name | Accession Number | Regulation | AGINumber | Treatment Name | Compatibility | Genus and Species (Cultivar) | Reference |
|---|---|---|---|---|---|---|---|---|
| 12-oxophytodienoate reductase | AtOPR3 | AV824251; AV785462; AF410322 | Up | At2g06050 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 107 |
| 12-oxophytodienoate reductase | AtOPR3 | AV824251; AV785462 | Up | At2g06050 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 334 |
| 1-aminocyclopropane-1-carboxylate (ACC) oxidase putative | not known | At2g19590 | Down | At2g19590 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 472 |
| 1-aminocyclopropane-1-carboxylate (ACC) oxidase putative | not known | At1g12010 | Down | At1g12010 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 472 |
| 1-aminocyclopropane-1-carboxylate (ACC) synthase | AtACS2 | not available | Up | not available | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 404 |
| 1-aminocyclopropane-1-carboxylate (ACC) synthase like | not known | At4g26200 | Down | At4g26200 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 472 |
| 2-isopropylmalate synthase-like; homocitrate synthase like | not available | AV821148 | Down | At5g23010 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 334 |
| 2-isopropylmalate synthase-like; homocitrate synthase like | not available | AV821148 | Down | At5g23010 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 107 |
| 2-oxoglutarate dehydrogenase, E1 | not available | BE844998 | Up | At3g55410 | sodium chloride | non pathogen | Arabidopsis thaliana (cv | 15 |

Figure 13: Sample of the 500+ data points shown in non-visual list format

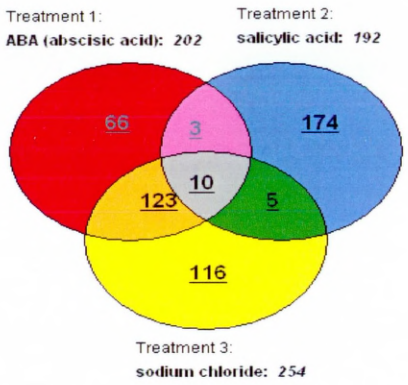



Figure 14: Data shown in visual Venn diagram format

There are several creative uses of visualisation techniques from the tools from the powerful use of dendograms for hierarchical clustering of microarrays in the Java Treeview from TAIR to the simple chromosome mapping tool which converts a simple list of *A.thaliana* genes into a diagram which places them on a chromosome map. However, some of the web-based search tools described do simply produce results as lists of genes. In some cases this is due to the limitations that are caused by web technology but as shown, applying visualisation tools directly to the data enables the user to gain a deeper, faster understanding of the data and is an important consideration when designing tools. There are some new graphing tools

being developed which aim to provide visulisation of pathway data. One such example of this is ArrayXPath (Chung *et al.*, 2004) which takes existing pathways and maps gene expression data onto these pathways. While this is limited at present due to the lack of pathway templates for defense response signalling, as pathway knowledge develops, this technology could be an interesting companion to the DRASTIC database.

### 3.5.5. Web-based Delivery and Interactivity

Web-based software tools attached to an "in-house" database is a popular model for delivery of software tools as it enables scientists to access the tools with no effort and no download or installation requirements. Ideally the data results from one query should be in such a format that these can be easily transferred into another query to promote interactivity. In practice this is very difficult due to formatting issues. One of the problems with microarray data is that in order for it to be analysed for a specific purpose, it tends to require pre-processing into a suitable format for purpose. ACT and Genevestigator in fact use data from the same source but the toolset is not compatible. Another issue with this type of set up is that the dataset tends to remain static and therefore does not reflect updated annotations and from a data mining point of view, once a query is run, the result will remain unchanged.

### 3.5.6. Data Tools Summary

Some of the current web-based services hold similar gene expression data from *A.thaliana* microarray experiments (but not other plant species) and enable the recovery of information for individual genes or gene sets such as NASCArrays tools, Genevestigator and ACT but they are aimed at different users or have different tools. Genevestigator and NASC focus on how expression of selected genes varies with respect to different tissues and experiments. ACT provides tools to enable users to analyse how gene expression changes with respect to all the other genes on the array.

These tools are interesting in that they provide comparative gene analysis services to detect clusters of genes with similar expression patterns across selected or the complete set of treatments. The downside to these tools is that they force the user to start with a given gene of interest to determine similarities in expression patterns to other genes. They do not enable the user to compare treatment selections and there is no capability to select combinations of

different treatments to determine all overlapping genes being up- or down- regulated by these treatments.

From reviewing the data tools, there are many potentially useful tools available but no specific toolset to provide investigators with a toolkit for investigating hypotheses for signalling in plants. There are some excellent tools that enable gene co-expression to be analysed but none that are multi-species with a specific dataset tailored for early signally defence responses, or that enable the user to select treatments or that provide links to literature resources and nomenclature.

### 3.6. Summary

Chapter 3 provides an overview of the type of existing tools and datasets that were current when the project began. There is no single direct data source that is suitable although some data may be extracted from multiple sources that have been examined in this chapter to complement the journal data. Collating and analysing data from multiple sources will entail studying the data and seeking to determine a generic model. SCRI scientists have expressed a wish to find a way to automate the connectivity between resources to enable data to be generically considered by a number of tools while maintaining a close link to the literature and the findings from this chapter would support this model. Chapter 4 investigates the data types to find a uniform system to enable the data sources in different formats to be analysed.

# Chapter 4 Exploring the Gene Expression Data

## 4.1. Introduction

Chapter 3 has found that there are no single direct data sources suitable for specific investigation of plant defence. There are, however, many results from experiments readily available in journal papers and some non-peer reviewed database sources as described but these are not provided in a standardised manner and this is a key challenge to the synthesis of the database. The first step in examining the data is to determine the type and format of the data available from peer reviewed sources, identify the different formats, nomenclature and annotations that may be encountered and construct a model that will be suitable to store data from a variety of sources and formats. Prior to the start of this project, SCRI had been collating journals that contained relevant gene expression data. These journal articles contain data from a variety of experiments in a variety of different formats and it is this diversity which can cause problems when trying to standardise data into a unified structure. The third aim is to examine the structure of data particularly from journals and make this more accessible. The fourth aim is to provide a method to enable results from different types of experiments to be compared against each other. This chapter investigates the different data structures and identifies the key elements that will be required to construct a model that is suitable to store the data relating to stress-response in plant genes from these different formats and in a configuration that will enable the data to be queried thereby assisting knowledge gain in the signal transduction area.

Unfortunately, published gene expression data is not uniformly presented, making truly systematic searching impossible. This is especially so for the vast amounts of microarray data emerging each month. Typically, unknown or poorly characterised genes cannot be compared with prior expression data, rendering up to 15% of database entries valueless, clouding future interpretation and may impede biochemical and technological developments severely. Whilst part of this value paucity undoubtedly comes from the complexity of cell biology, changes in gene names associated with database accession numbers adds to the fog of confusion.

Gene and protein names are often flawed and misleading when naming conventions are not universally adopted and adhered to (Lyon *et al.*, 2002). Value is effectively lost from datasets when the same gene is given different accession numbers in databases. A single *A.thaliana*

gene which has been shown to be down-regulated by chitin, drought, ethylene, low oxygen tension or sodium chloride environmental treatments and up-regulated by salicylic acid treatment has been given five different accession numbers by authors, all ultimately corresponding to the gene of unknown function, Arabidopsis Genome Index (AGI) code At2g10940. This was determined after laborious tracking of identities and codes through the scientific literature and the virtual world. Since it is most unlikely that scientists will systematically search for such similarities for each and every gene studied, the resulting low quality classification of expensively acquired data means that elements of commonality or uniqueness will frequently fail to be identified. Hence reduced scientific value is being realised from complex experiments and potentially important conclusions fail to be drawn with respect to the control of growth, development and host/pathogen responses.

## 4.2. Nomenclature

### 4.2.1. Concept of a Reaction

The data that will be collected for the project is based on the model shown in Figure 15 which captures the basic concept of a reaction.

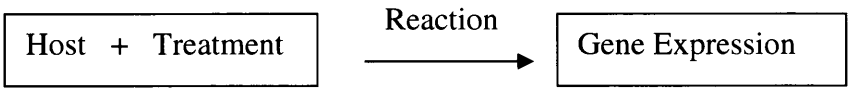



Figure 15: Diagram of basic concept of a reaction

The study is aiming to collect stress response results from multiple plant species and therefore need to identify the data that will be required and how this will be collected to allow for a uniform data model to be built to hold and allow computational investigation of the data. In order to record the gene expression result, the gene that the result corresponds to needs to be recorded.

### 4.2.2. Historical Problems with Gene Name Nomenclature

Researchers are hindered by a lack of standard naming conventions for genes and proteins. The gene name is the form by which a protein object is referred to and communicated in the scientific literature and biological databases. There is a long standing problem of

nomenclature for proteins where 'profligate and undisciplined labelling is hampering communication as discussed in Nature Opinion (1997). Scientists may name a newly discovered or characterised protein based on its function, sequence features, gene name, cellular location, molecular weight or other properties as well as their combinations or abbreviations. The same protein is often named differently in different databases, and occasionally different proteins may share the same name. Only a small fraction of all proteins has standard nomenclature, most notably the enzyme nomenclature of the International Union of Biochemistry and Molecular Biology (IUBMB) www.chem.qmul.ac.uk/iubmb/enzyme.

A later study (Aubourg and Rouzé, 2001) found that there is a clear lack of controlled vocabulary both in the literature and the databases. This problem is linked to sequence redundancy in the databases, which can contain several times the same genes under different names. The resulting loss in time for the search and the annotation is very serious. Furthermore, the multi-origin of the annotations amplifies the diversity of the nomenclature. For example, the American annotators name as 'putative' or '-like' a function deduced from similarities, whereas the Japanese centre and MIPS use 'potential' and 'similar', respectively.

With the advent of large genome sequencing projects, nomenclature has been a main focus in recent years. Concerning the problems of nomenclature, important efforts are in progress. The Gene Ontology consortium (http://www.geneontology.org) and the Mendel database (http://genome-www.stanford.edu/Mendel/) produce a reference vocabulary for the gene names and functions and a common basis for genome annotation. The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org/) is a central site for this model plant. The web site centralizes all the links towards Arabidopsis databases, research laboratories and annotation centres, and also displays regularly updated genetic and physical maps.

### 4.2.3. Standards in Gene Naming Nomenclature

As a result of the problems that non standardised methods of gene naming cause, standards for naming genes were introduced (Price *et al.*, 1996) and curated by the Commission on Plant Gene Nomenclature. Naming conventions for gene families, gene symbols and gene products were standardised in a bid to reduce duplication of names and make them more meaningful. Efforts to improve the gene naming nomenclature have continued to improve

and the biological community is moving towards a universal system for the naming of genes as described for a number of plants such as Arabidopsis (Schlueter *et al.*, 2005), Tomato (Mueller, 2005), Medicagio (VandenBosch and Frugoli, 2001) and maize (MaizeGDB, 2002). In most plant species databases, a gene is identified by a name, a gene family where assigned and a gene symbol or synonym along with the EST and the corresponding Accession Number for EST which acts as a unique identifier for the EST. Having this standard information about all sequences is a huge step in the right direction, but this does not uniquely identify a gene but represents a sequence from a gene of which it is likely that there could be more than one for each gene.

In the plant research domain, *A.thaliana* is the model species of current plant genomic research with a genome size of 125Mb and approximately 28,000 genes. The function of half of these genes is currently unknown (Lan *et al.*, 2007). Arabidopsis was fully sequenced in 2000 in a collaboration project that was headed by The Institute for Genomic Research (TIGR). In order to be able to uniquely identify each of the Arabidopsis genes a naming convention was required and thus AGI numbers were introduced (Haas *et al.*, 2005).

### 4.2.4. Arabidopsis Genome Initiative Nomenclature

The Arabidopsis Genome Initiative (AGI) nomenclature allows for the designation of unique locus (gene) identifiers for Arabidopsis. The syntax of the AGI nomenclature is described below:

The format of the AGI numbers is in three parts:
Part 1 = The organism, in this case At –*A.thaliana*
Part 2 = The chromosome that the gene is found on (1-5 for Arabidopsis)
Part 3 = The gene id – g followed by a five digit number
An example AGI is At4g10020 which means that it represents a gene from *A.thaliana* found on the $4^{th}$ chromosome and the gene id 10020. The fact that there is now unique identifiers for genes is a boon for bioinformatics as it allows direct comparison without the complication and potential errors that the previous systems have introduced. The unique identifier does not remove the need for the gene name (as this provides information about the function of the gene) or the gene symbol and it is wise to record the accession number of the sequence that was used in the experiment as this allows the investigator the opportunity to find the exact coding sequence used for each experiment.

There are rules for adding new genes, deleting genes and editing genes so this means that the AGI numbers will not remain static and may change as annotations are updated. Locus history for each gene is available from the TAIR website so it will be possible to track historical AGI numbers and update these as required.

### 4.2.5. Gene Ontology (GO) Nomenclature

The Gene Ontology (GO) (http://www.geneontology.org/) project was established to provide a common language to describe aspects of a gene product's biology. The use of a consistent vocabulary allows genes from different species to be compared based on their GO annotations. The GO project started as a collaboration between three model organism databases, the *Saccharomyces* Genome Database (SGD), FlyBase (for *Drosophila*), and Mouse Genome Informatics (MGI). The GO Consortium has expanded considerably to include many additional model organism databases and annotation groups including Arabidopsis, each of which contributes to the development of the ontologies, generation of GO annotation files, or development of software tools to utilize GO depending on the nature of its affiliation. The GO annotation can be mapped to the AGI numbers and so can be included for any gene where the AGI identifier is known.

### 4.2.6. Gene Names Nomenclature Discussion

The Arabidopsis genome was the first plant genome to be systematically sequenced and annotated and a large proportion of the gene expression data contained in the curated journals is from Arabidopsis, but the database must be able to deal with data from multiple species to meet the requirements. The sequencing of the rice project was completed after Arabidopsis, and it adopted the Arabidopsis project as a model for the annotation of the rice genome (Aubourg and Rouzé, 2001). The AGI number format transfers over to another species very simply. An example is Os03g44290 which is a gene from rice (*Oryza sativa*). It uses the same convention whereby Os represents the organism, the next numbers represent the chromosome and it is found on the 3$^{rd}$ chromosome and the gene id is 44290. This naming convention enables tools to be designed that are portable from one species to another.

Dealing with plant data from species that have not been fully sequenced or have not adopted the AGI format is not as easy. The only way to uniquely capture the gene is to record the Accession Number, the gene name and the gene symbol for each result reported. This will

have an impact on the comparison prospects for these genes as a gene may be represented by multiple EST's and it may not be possible to identify which group of EST's represent a gene, therefore some of the gene expression data may not be fully transparent. For plants that do not have a unique gene identifier, maintaining an up-to date standardised gene name is imperative as this along with the gene code will be the unique identifier and this will be done through the use of Unigene as discussed in Chapter 3.

In summary, the main obstacles that will be encountered with gene name nomenclature are

1. Not all species have got a defined unique gene nomenclature. Arabidopsis has AGI numbers and this system does seem to be taken up by rice and possibly by barley. For those without gene identifiers, tracking results can be trickier and obtaining an accession number and gene name is vitally important in these cases to enable results to be tracked from journals.
2. Some genes are identified by an EST but the AGI reference may have been omitted or there is no unique identifier for the gene. There can be many EST's per gene and it can be difficult to track the AGI associated with the EST. The database must implement a method to track these genes and maintain up-to date data to increase the knowledge base.
3. There are problems when considering comparing data across species. At present the only method would be to blast accession numbers to find similar genes in different species. It is hoped that with the introduction of unique identifiers that genes which are similar from one species to the next can simply be mapped by the ID and this would enable more powerful queries to be run. At present, the bulk of data is from *A.thaliana* but it is likely that this will change over time, so considerations must be made as to the changeability of the data and the data structures when designing the database and tools.

### 4.2.7. Plant Name Nomenclature

The host is the plant on which the experiment is centred and there is strict nomenclature for cultivated plants (Brickell *et al.,* 2004), which makes modelling the data straightforward. The requirements elicited from the biologist were that the family, genus, species and cultivar would be required from each experiment as well as the classification of plant-monocot or – dicot . The cultivar has been highlighted by the scientists as important as although most of the experimental results that we anticipate collecting will be from wild type plants, some data

may be obtained from genetically modified plants and it will be important to differentiate these. The Plant Ontology Consortium (POC) are involved in developing, curating and sharing controlled vocabularies that describe plant structures and growth and developmental stages, providing a semantic framework for meaningful cross-species queries across databases which may prove useful as results from experiments can often be from different parts of a plant and the biologists have expressed a requirement to log this (Ilic *et al.*, 2007).

### 4.2.8. Treatment Nomenclature

Lastly, the 'treatment' referred to in this thesis is the method in which the stress response is triggered in the plant. These are classified into type categories of abiotic stress arising from an excess or deficit in the physical or chemical environment, such as heat stress or biotic stress imposed by other organisms, for instance the pathogen *Pseudomonas syringae* pv *maculicola*. For pathogens, a further requirement is the compatibility to the host plant. For each treatment the name and description are required along with the type classification. The gene expression result varies depending on the type of experiment platform used as described in the next section.

## 4.3. Types of Experiments

At present the collection of journal references, which will form the basis of the database data relate to approximately 57% Northern blot experiment results, 13% Microarray results, 8% Reverse Transcription Polymerase Chain Reaction (RT-PCR) and 22% others. This demonstrates that gene expression results related to plant defence signalling are obtained using several types of experimental technique and based on the above findings, this project considers the primary methods to be Northern blotting, RT-PCR and Microarrays. The quality, type and quantity of results from each experiment category varies dramatically as outlined below.

### 4.3.1. Northern Blot

The Northern blot is the oldest technique and was developed in 1977 by James Alwine *et al* (Jackson *et al.*, 2002). mRNA fragments are probed with a labeled DNA probe after separation by electrophoresis and transfered to nylon membranes. Northern blotting is used to detect and quantify mRNA or the levels of gene expression from tissue extracts.

While Northern blots have been superseded in most areas by RT-PCR and microarray approaches it is a widely accepted and straight-forward method and Northern blotting is often used as a confirmation or check of results produced by a microarray experiment (Holmes and Peck, 1998). This explains why there are such a high percentage of journal references that report Northern blot results. The drawback to Northern blotting experiments is that they are time consuming to set up, the number of steps involved creates more opportunity for error and they are low throughput where only one gene can be examined at a time (Roth 2002).

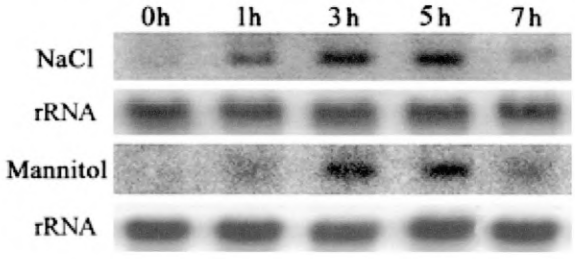



Figure 16: Example result from a Northern blot Experiment (Zhang *et al.* 2007)

Figure 16 is a result from a Northern blot experiment that was published in the Plant Molecular Biology journal by Zhang *et al.* (2007). In this experiment, one gene, namely TSRF1 from a tomato, was treated with sodium chloride (NaCl) as shown in the first horizontal line, rRNA was used as a control in the second line, TSRF1 was treated with Mannitol in the third line and the fourth line is another control experiment. Zhang *et al.* (2007) describe the experiment as "To further investigate the role of TSRF1 in plant abiotic stress response, we analyzed the expression of *TSRF1* under osmotic stress conditions. Results showed that *TSRF1* also responds to ionic osmotic stress caused by NaCl or nonionic osmotic stress caused by mannitol. As shown in Figure 16 above, the expression of *TSRF1* is observed at the first hour, and got the peak at 5 hours after treatment with NaCl or mannitol, indicating the possible regulation of TSRF1 in plant osmotic stress response." The results from Northern blots are generally reported as an up- or down- regulation of gene expression in journals although the quantitative increase or decrease in gene expression can be determined using a densitometer to assess the intensity of each result (Roth, 2002). When collecting data from a Northern blot experiment for inclusion to the database, the host (in this case *Lycopersicon esculentum* cv Lichun), gene (TSRF1 name, AF494201 Accession number), treatment in this case (Mannitol) and gene expression result (obtained from Figure 16) must be collected. The temporal time is included in this experiment and should be

recorded in hours as indicated in section 2.3 of the biologists requirements as this reflects the subject domain where the users are only interested in early stress responses. The gene must be able to be uniquely identified to be included and so it is important to look for EST accession references or a unique identifier such as an AGI where this exists. In this case, the gene name TSRF1 was the only identifier mentioned in the 2007 paper and the reader was directed to a prior 2004 paper. The later paper provided the Unigene accession number for the gene in question and this demonstrates the problems that exist with collation of results from journals.

### 4.3.2. RT-PCR

RT-PCR is an extension of the polymerase chain reaction (PCR) method. mRNA is extracted from the cells or tissue, converted to cDNA using the enzyme reverse transcriptase and PCR is then carried out.


Figure 17: Example result from a RT-PCR Experiment (Lee *et al.* 2001)

This method is used to detect the expression of specific mRNA sequences in cells or tissues. RT-PCR is widely used, quantitatively, in the determination of the abundance of specific different RNA molecules within a cell or tissue as a measure of gene expression. This technique is more sensitive than the Northern blotting technique and can detect levels of mRNA that would be missed using Northern blotting. (Holme and Peck, 1998).

Figure 17 is the result from an RT-PCR experiment (Lee *et al.*, 2001). The result is described as "Treatment with harpin of cultured tobacco cells resulted in transcript accumulation of the *PR* genes *PR1*, *PR2*, acidic chitinase (*PR3*), and chitinase/lysozyme. Transcripts derived from

genes considered HR marker genes (*HSR203*, *HSR201*, *HSR515*, and *HIN1*) also accumulated in harpin treated tobacco cells." In this example, the gene expression result without treatment (-) is compared to the gene expression result with treatment (+). If the untreated sample (in this case the (-) column) has no visible mark and the treated sample (in this case the (+) column) does then the gene expression has been up-regulated. If the (-) column has a mark and the (+) column has no mark or a less bold mark then the gene expression is considered to be down regulated and if the (-) columns and the (+) column are the same then the gene expression is considered unchanged. For the purposes of this database, the collation process for these results is identical to the Northern blot process with the same details being required from the experiment as the results are also reported in journals as up- or down regulated.

### 4.3.3. Microarrays

Microarray technology is a relatively recent innovation and is quickly becoming a standard tool in molecular biology (Lorkowski and Cullen, 2006). Microarrays differ vastly from the previous two techniques discussed as they can provide the gene expression levels of thousands of genes simultaneously by doing thousands of experiments in parallel. The drawback to microarrays is that the creation of each experiment is more complex with more steps than Northern blotting or RT-PCR and this leaves more room for error and an additional problem of variance in protocol from array to array and from experiment to experiment.

There are the two types of microarrays used to measure gene expression which are two-channel cDNA, or Spotted, arrays and single-channel High-density Oligonucleotide arrays and results from both of these types are reported in the papers that are relevant to this project. While both types of the arrays measure gene expression, the two arrays require different experimental and analytical approaches (Baldi and Hatfield, 2002). In the spotted array method, mRNA from two biological samples is copied to cDNA, each cDNA is labeled with a different fluorescent label, and a mixture of the two cDNA's is hybridized to an array that has a single DNA spot for each gene on the array. Each spot is scanned and the ratio of the two labels is determined. Commercial manufacturers of the spotted array use genomic DNA and oligos and based on the papers used for this study, this type of array is used less frequently that the oligonucleotide array. In the oligonucleotide array method, mRNA from a single biological sample is copied to labeled cDNAs, and then hybridized to a set of short 25-mer matching oligonucleotides for each gene and also to another set of the same

oligonucleotides. To compare biological samples using this method, two hybridizations to two separate arrays are needed (Mount, 2004). Commercial manufacturers of the "oligonucleotide" arrays include Affymetrix, Nimblegen and Agilent and they can provide almost complete genome microarray chips for fully sequenced organisms including Arabidopsis. Affymetrix results are the most prevalent and results from the microarray chips for the Arabidopsis AG chip (which can measure expression values of approximately 8000) and its successor the ATH1 chip (which can measure expression values of approximately 22000 genes) appear with greatest frequency based on the sample of journal articles relevant to this project.

There are different ways that microarrays can be used, for example, some microarray experiments are time sequenced where the designer may require the expression results at particular time intervals and then examine how the expression results vary over the time interval while others with be comparing a normal or untreated DNA against chemically treated DNA and the control vs. the treated expression profiles are examined. It is the later type of experiment that this project is interested with.

#### 4.3.3.1. Common Microarray Terminology

The main terminology used for describing the calculation of results from microarray expressions are the same across both the single and two channel microarray techniques although the way in which the samples are processed is different. This section provides a reference for the key terminology that is used for measuring gene expression when comparing a normal sample against a treated sample. Once the experiment has been carried out, the absolute value, which means the numeric value of the level of gene expression, is obtained from the control sample and the treated sample. This value is then normalised which removes any systematic variation, for example differences in power of two lasers or between dyes and therefore brings the data from the different experiments onto a level playing field (Bari *et al.*, 2006). This is imperative as it allows comparison of results from different microarrays. Once the data is normalised, the two samples are then compared and a relative value is calculated. This is simply the ratio of the normalised absolute treated value and the normalised absolute control value. This relative value is also referred to as the log ratio or fold value. The fold value provides an indication of the change in level of expression and allows a cut-off value to be determined after which a change call can be given. This cut-off

value will commonly be set by a software package, and tends to be at least 1.5. For example, if a fold value shows a 1.5 increase, the change call will be up-regulation of the gene expressed and if the fold value shows a 1.5 decrease the change call will be down-regulation of the gene expressed. The majority of all the calculations are done by sophisticated statistical packages which provide additional confidence values about the quality of each individual result. However, as discussed in section 4.3.4, it is necessary to be aware of how to calculate the change call as there are still some papers that only feature absolute values. Relative values or fold values are very suitable for this study's database approach however, as the cut-off value can vary from paper to paper, this study is guided by what the author of the paper considers to be an appropriate cut-off value for up- and down- regulation calls.

### 4.3.3.2. Single Channel Microarray Example

Figure 18 shows a partial table of genes that expressed differently in Arabidopsis plants colonized by *Pseudomonas fluorescens* FPT9601-T5a using a single-channel Affymetrix ATH1 microarray (Wang *et al.*, 2005).

| Annotation | Fold-change ratio | Probe set number | AGI number |
|---|---|---|---|
| Up-regulated genes | | | |
| Metabolism (22.11%) | | | |
| Seed imbitition protein | 3.14 | 246114_at | At5g20250 |
| Putative fatty acid elongase | 2.87 | 263443_at | At2g28630 |
| Putative glutathione *S*-transferase | 2.66 | 266746_s_at | At2g02930 |
| β-Xylosidase | 2.52 | 260914_at | At1g02640 |
| Xylosidase | 2.45 | 248622_at | At5g49360 |
| Invertase inhibitor homolog | 2.45 | 247246_at | At5g64620 |
| Xyloglucan endotransglycosylase | 2.38 | 257203_at | At3g23730 |
| Putative xyloglucan endotransglycosylase | 2.35 | 255433_at | At4g03210 |
| Proline oxidase | 2.33 | 257315_at | At3g30775 |
| β-Galactosidase | 2.33 | 247954_at | At5g56870 |

Figure 18: Example results showing calculated fold changes from a single-channel array (Wang *et al.*, 2005).

The important features to note are the gene name (under the heading annotation) which in this case has been allocated based on the Gene Ontology (GO) nomenclature, the probe set number which is the reference to the individual spot on the microarray, the AGI number which provides the unique ID for the database and the fold-change ratio which allows us to calculate the change call (the cut-off value in this paper is 2). Affymetrix have their own well regarded software called Microarray Suite software (MAS 5.0). This software tool manages both the acquisition and processing of the absolute data and provides fold-change ratio results along with p-value and absent/present calls that give a good indication of the accuracy of each result and therefore if the fold value is not calculated in the paper, it is reasonably

straight forward to calculate the necessary data if this system has been used. Affymetrix have their own published annotation for mapping the probe ID to an AGI number for both the AG and the ATH1 microarray chips. The gene function and description can also be taken from this annotation, however, the unigene database is updated more regularly with gene function information and so the project will use the function and description from the unigene source for all experiments to provide a standard platform unless the primary paper is more up-to-date.

### 4.3.3.3. Two Channel Microarray Example

Figure 19 is a sample of gene expression results from a two channel microarray (Schenk *et al.*, 2000). Here, the change call has been explicitly stated. It shows genes that are significantly induced or up- regulated (positive ratios, shaded in light orange) or repressed or down- regulated (negative ratios, shaded in light blue).

| | Genbank | AtDB | A | SA | MJ | Eth |
|---|---|---|---|---|---|---|
| **PUTATIVE FUNCTION** | Accession | Clone | Ratio | Ratio | Ratio | Ratio |
| **OXIDATIVE BURST/STRESS, APOPTOSIS** | | | | | | |
| copper/zinc superoxide dismutase | H36758 | 178G17T7 | 3.78 | 1.97 | 1.19 | 1.32 |
| L-ascorbate peroxidase | N64977 | 223L16T7 | 2.62 | 1.88 | 1.75 | 1.3 |
| cysteine protease | T04773 | 32C6T7 | 2.53 | 2.29 | 1.95 | 1.55 |
| catalase 3 | H76812 | 203O12T7 | 5.4 | 1.48 | 8 | 1.36 |
| glutathione S-transferase PM24 | N65700 | 229O3T7 | 4.46 | 1.04 | 2.82 | 1.85 |
| DNA-damage-repair/toleration protein | T44979 | 127N10T7 | 5.9 | 2.13 | 1.06 | 7.7 |
| blue copper protein | T44253 | 123N22T7 | 1.27 | 1.39 | 5.59 | 1.47 |
| SAG12, cysteineprotease | #N/A | CI0010 | 1.33 | 0 | 5.57 | 2.74 |
| **ANTI-MICROBIAL GENES** | | | | | | |
| germin-like protein | T22353 | 104D20T7 | 2.1 | 3.17 | -3.93 | 1.13 |
| major latex protein type 3 | T20653 | 89M9T7 | 2.82 | -1.3 | -1.15 | -1.22 |
| PDF1.1 antifungal defensin protein | Z27258 | PAP065 | 8.3 | -2.1 | 4.35 | 2.16 |
| PDF1.2 antifungal defensin protein | T04323 | 37F10T7 | 2.89 | -2.2 | 2.48 | 2.41 |
| PR1 | #N/A | CI0014 | 2.76 | 7.76 | -1.04 | 0 |
| thaumatin-like protein | T46212 | 138H14T7 | 1.8 | 2.39 | 2.55 | 1.21 |
| PAL1, phenylalanineammonia-lyase | #N/A | CI0004 | 2.12 | 2.09 | 2.3 | -1.92 |

Figure 19: Example results showing calculated fold changes from a spotted array by Alternaria (A), salicylic acid (SA), methyl jasmonate (MJ), or ethylene (Eth) treatments. (Schenk et al., 2000).

Here the microarray was of bespoke construction consisting of 2375 ESTs chosen by the paper authors and they imposed an induction or repression ratio cut-off of at least 2.50 for data quality purposes. Of the 9500 results obtained, only 705 of these show up- or down-regulation to the treatments, therefore only 7.5% of the results from the microarray

experiment are suitable for inclusion in the database. This is a very typical figure and as discussed in Chapter 5 has implications on the types of analysis that can be performed on the collected data. The key results for data collection are the host plant, the type of treatment, the gene treated (which is identified in this example by the putative gene name and the Unigene accession number reference) and the fold change result for each gene. The accession number (EST reference) will provide a unique identifier for the experiment, but it makes comparison between different genes impossible as genes can have multiple EST sequences. For inclusion to the database, it is preferable, where the nomenclature exists, to have a unique gene identifier for each gene expression result. As this is an Arabidopsis experiment, further work on the part of the collator would be desirable to obtain the AGI number for each gene as this is not available in the original publication (discussed further in section 4.4).

### 4.3.4. Published Result Formats

There are numerous problems encountered in the curation of the gene expression results from papers and some of the primary causes of this are missing or inaccessible data and varying formats of data. One example of missing data has already been highlighted in section 4.3.1 where the accession number for the results from the experiment published could only be found by searching for a previously published paper. Missing data in the form of missing accession numbers or probe numbers has been a common occurrence in the curation of the data for this study. Another commonly occurring problem is inaccessible data, where the expression data has been created as a word document table, but then saved as a picture file. This makes the data unsearchable and forces manual curation. There seems to be no set standard for the publication of gene expression results and selective results are published in journals in a variety of different formats (such as excel, cvs, .cel and pdf files) and even the supplementary results can be provided in a different levels of depth as well as in different software packages. This is being tackled by some journals by mandating that authors must submit their results to a public microarray repository where they can be accessed in full and enforcing adherence to MIAME standards. The reference papers that have been collected for this project are from over 60 journals. Of this number approximately 50% of the papers come from five journals, namely, Plant Physiology, Plant Molecular Biology, The Plant Journal, Molecular Plant-Microbe Interactions and Molecular Plant Pathology. A case study was

carried out in this study to establish what the journals require from authors that submit papers for publication.

#### 4.3.4.1. Case Study of Author Requirements

The Plant Physiology Journal is the most specific of the journals surveyed for manuscript submission guidelines. It requires plant names in the form of genus, species, and, when appropriate, cultivar. Accession numbers should be provided at the end of the Material and Methods for any data or materials available in a public repository. Novel DNA sequences must be deposited in Unigene and accession numbers provided. For large scale expression data, accession numbers, relevant annotation data, and in the case of Arabidopsis, TAIR locus identifiers. At the time of publication, supplemental data must be placed in a permanent public repository if one is available, or if none is available, in Plant Physiology Online. As a condition of publication in Plant Physiology, submitters of manuscripts that contain gene expression profiling data are required to describe the experiments according to MIAME guidelines and must include replicate experiments (PP, 2008). The Plant Molecular Biology has no requirement to submit data to public repository or no standard format for adding expression data to supplementary files. It does state that standard nomenclature procedures should be followed, but makes no requirement for accession numbers to be provided or AGI or other unique locus ID's. There are no guidelines on how experimental results should be submitted and no mention of MIAME standards (PMB, 2008). The Plant Journal states that authors including microarray analysis should refer to the MIAME recommendations for guidance in preparing their manuscripts and makes no other reference to how strictly this is enforced. There is no reference to any other experiment type and no indication of how results from the experiments should be reported in the paper (PJ, 2008). The journal of Molecular Plant-Microbe Interactions has a very specific policy on large-scale data sets and enforces strict adherence to the MIAME guidelines. There is no guidance for the information required for smaller scale experiments or for the format that results are to be published in. (MPMI, 2008). The journal of Molecular Plant Pathology has no MIAME requirement or guidelines for the information that must be included when reporting experimental results (MPA 2008).

This case study shows that there is a large disparity between journals for author guidelines on reporting results. The introduction of MIAME standards has definitely had a positive impact with three of the five journals surveyed making it mandatory to meet this standard for

acceptance. This means that the data must have unique identifiers and even if the data published is incomplete or in a difficult format to curate, the full microarray results will be publicly available. The downside of this is it that it is likely that the fold values will not have been calculated and it will be absolute data values that are available. With the exception of the Plant Physiology journal, none of the other journals surveyed made any specific requirements for the reporting of any other type of experiment or how the results were to be presented. Until a more unified approach to the reporting of experiments in journals is taken, manual or automated curation of results from certain experiments will remain difficult.

### 4.3.5. Can data from different platforms be compared?

There is much debate in the literature as to whether it is appropriate to compare gene expression results across platforms or from different experiments and this section assesses how other researchers view this. The first question to consider is "can microarrays made by the same manufacturer be compared?"; for example, can we compare Affymetrix ATH1 chip against the AG chip? The AG array contains 8297 probe sets and the new ATH1 array contains 22814 probe sets. Based on annotations compiled by TAIR, 7388 transcripts are targeted by probe sets on both arrays so can these results be analysed against each other?

Genevestigator is a site that allows for comparison of the fold values (relative values) of the results from microarray data (including some Affymetrix AG and ATH1 Arabidopsis microarrays). They state that "Genevestigator uses the Affymetrix MAS5.0 algorithm for data normalization. Provided that scaling factors (SF) are in a similar range and there is no significant skew in the data, our hypothesis is that signal intensity values can be compared. The Genevesigator tools are based on this assumption, and the results seem to indicate that it is a fair assumption. For some tissues (e.g. pollen or embryo) where a large fraction of genes are not or weakly expressed and a few are strongly expressed, signal intensity values tend to be overestimated (and the corresponding scaling factors may differ significantly). Therefore, the results provided by genevestigator reveal trends rather than exact quantitative information" (Zimmerman *et al.*, 2004).

However, the genevestigator tool does not allow the user to compare the ATH1 results against the AG results. Zimmerman *et al.* (2004) state that the reason data from the ATH1 and AG arrays are processed separately is because different sets of oligonucletoide sequences are used to probe identical target genes on the two array types, and thus the efficiencies of the

target to probe hybridization and non target to probe cross-hybidisation makes a direct comparison of signal intensities impossible.

Hennig *et al.* (2003) analysed the reproducibility of the results of transcript profiling between microarrays carrying different probes to a common set of genes and focused on the overlap of more than 7300 targets from the AG array and the ATH1 affymetrix array. They found that the results obtained with ATH1 and AG arrays are very comparable and hence that the analysis is largely independent of probe sets. They summarised by suggesting that analysis should be focused only on genes called Present by MAS5.0 regardless of their actual signal intensities. A further, although smaller improvement of data quality, can be achieved by only including genes called decreased or increased by MAS5.0. Targets producing a fold change of at least 1.5 gave results with the best correlation between array types. Given the fold value of genes identified as up- or down- regulated when reported in journals is > than a 1.5 fold increase or decrease, this would appear to imply that comparison of these results is acceptable but is it suitable to compare results across different microarray platforms by different manufacturers?

There have been approximately 40 studies since 2000 which have evaluated the extent to which data produced by different microarray technologies correlate. Irizarry *et al.* (2005) conducted a multiple laboratory comparison of microarray platforms and found that precision is comparable across platforms and that it is the laboratory that affects the experiments more strongly than the microarray platform used. They also found that it is only relative expression (fold value) that can be compared. Yauk and Berndt (2007) have reviewed the results of the 40 studies and concluded that the vast majority of papers published on this subject support a high degree of correlation among microarray technologies. Both Irizarry *et al.* (2005) and Yauk and Berndt (2007) state that it is the standards and protocols that make the most impact on the correlation of the microarray results. This evidence suggests that it is acceptable to compare microarray data from different platforms provided it is relative values (fold values or call values) that are being compared. It would also appear that only using data from published journals is a sensible approach as it is likely that the required standards will have been applied to these experiments as they have been peer reviewed. The final consideration is whether it is appropriate to compare data from not only different microarray platforms but RT-PCR and Northern blot results as well. Yauk and Berndt, (2007) describe the evaluation

of the microarray gene expression results discussed as being confirmed using RT-PCR and Northern blot techniques. This would imply that it is also suitable to compare these experimental results alongside the microarray results.

Based on the above findings, it is appropriate to compare results from the different experiment techniques described in this chapter. In order to enhance the data quality, the data that we are comparing will only be obtained from a refereed source. It is expected that there will be duplicate results from different experiments/papers which will provide an additional level of confidence in the data. Further more, it is normal practice for any conclusions that researchers may make, while using bioinformatics tools would be presumed 'unproven' and tested using wet science to confirm the hypothesis. Data will be stored in the database in binary form either as up- or down- regulated and for microarray experiments, the actual relative values, where reported, will also be included.

## 4.4. Annotations

From the information obtained about the experiments, it is now important to establish how we can update experimental results to ensure that we can compare genes and that the historical gene expression results gathered relate to the current genome annotation.

### 4.4.1. Overview

Annotations are in every part of bioinformatics from database formats to locus id for genes to probe mapping for microarrays. It is vital to consider what impact annotations may have on historical static results such as those found in journals. To examine this, a case study of the implications of one type of annotation has been undertaken during this study.

### 4.4.2. Microarray Annotation Case Study

This case study will examine the gene to probe mapping annotation of the Affymetrix ATH1 microarray. The ATH1 microarray experiment results are all taken from a probe which has its own ID and then this ID is mapped using annotation to correspond to an AGI number. The problem is that these probe ID to AGI number look-ups change as more up to date data becomes available. There have been five releases of the annotation for the probe to gene mapping for ATH1 microarray chips and Table 5 shows the number of changes that occur between four of these mappings. There are no changes made to the probe ID numbers or to

their extensions, however there is a large variation between the probe ID's and the locus that they are mapped to as shown in Table 5: (The control probes have been disregarded for this purpose so the changes are out of 22746 probe ID's).

| Annotation Date: | 23/12/02 | 30/5/03 | 1/6/04 TIGR V5 | 11/11/05TAIR V6 | 2/5/07 TAIR V7 |
|---|---|---|---|---|---|
| 23/12/02 | | 264 | 1406 | 1713 | 1862 |
| 30/05/03 | 264 | | 1248 | 1611 | 1745 |
| 01/06/04 TIGR V5 | 1406 | 1248 | | 1748 | 1891 |
| 11/11/05 TAIR V6 | 1713 | 1611 | 1784 | | 263 |
| 2/5/07 TAIR V7 | 1862 | 1745 | 1891 | 263 | |

Table 5: Number of changes between the mapping of the probe ID to the locus in the Arabidopsis files on the dates shown in the column and row headings.

Table 5 shows that the probe set to locus mapping is relatively changeable with a 5-8% change between most annotations. Table 6 describes the probe name nomenclature used and contrary to the definition from Redman *et al.* (2004) that probes ending in _at are genes that are represented by unique probe sets there are a number of examples in the TAIR V7 mapping that show this is not the case, for example, probe 259435_at is mapped to both AT1G01448 and AT1G01450. Usadel *et al.* (2005) calculated that there were only 89.5% of the _at genes left that were unique in the TIGR V5. By counting the number of probes ending in _at that had duplicate AGI numbers in the locus fields, the data showed that in the TAIR V7 there were 604 of 21685 probes that were labelled _at but did not uniquely identify a gene. In TAIR V7 there are also 462 "no_matches" for the _at probes which means that previous results linked to these probe sets should now be disregarded. The reason for this change in annotation is a match may have existed in TIGR5 and disappeared in future annotation versions because the gene structure of relevant locus was updated so that the region to which the probe maps is no longer included in the new structure (Pers Com. Berardini, 2006a).

The TAIR annotation is calculated by "The oligonucleotide sequences of the probes were mapped to the Arabidopsis Transcripts dataset from the Arabidopsis genome TAIR7 version. The dataset included mitochondria and chloroplast genes, as well as pseudogenes and non-coding RNAs. The mapping to the TAIR7 Transcripts was performed using the BLASTN program with e-value cut-off < 9.9e-6. For the 25-mer oligo probes used on the Affymetrix chips, the required match length to achieve this e-value is 23 or more identical nucleotides.

To assign a probe set to a given locus, at least 9 of the probes included in the probe set were required to match a transcript at that locus." (Readme file from TAIR)

| Extension | Description |
|---|---|
| **_at** | designates probe sets that uniquely identify a single gene |
| **_s_at** | designates probe sets that share common probes among multiple transcripts from different genes. |
| **_a_at** | designates probe sets that recognize multiple alternative transcripts from the same gene. |
| **_x_at** | designates probe sets where it was not possible to select either a unique probe set or a probe set with identical probes among multiple transcripts. Rules for cross-hybridization were dropped. Therefore, these probe sets may cross-hybridize in an unpredictable manner with other sequences. |
| **_g_at** | similar genes, also unique probe sets elsewhere on the array. |
| **_f_at** | similarity rules dropped, probe set will recognize more than one gene. |
| **_i_at** | designates sequences for which there are fewer than the required numbers of unique probes specified in the design. |
| **_b_at** | all probe selection rules were ignored. Withdrawn from GenBank. |
| **_l_at** | sequence represented by more than 20 probe pairs. |
| **_r_** | designates sequences for which it was not possible to pick a full set of unique probes using Affymetrix probe selection rules. Probes were picked after dropping some of the selection rules. |

Table 6: Probe Name Extension Nomenclature for probes that represent more than one gene or EST. Only the _at, _s_at and _x_at extensions feature on the ATH1 chip.

TAIR are not alone in providing annotation data for the Affymetrix ATH1 microarray. Affymetrix themselves provide data that maps the probe sets to Arabidopsis AGI's and the Affymetrix annotation is different to both the TAIR V7 and the TIGR V5. There is no definitive standard in the literature as to which annotation should be used and is a continuing problem for bioinformatics. TAIRs response to why the Affymetrix data is different to TAIR7: "Unfortunately, I cannot tell from the Affy website what version of the Arabidopsis genome annotation they used to generate their mapping file. Also, there is no mention of the parameters they used to call a match 'legitimate'. This makes it difficult to compare our results and figure out what the differences arise from." (Pers Com. Berardini, 2006b).

To give a direct example, Bari *et al.* (2006) reported the results shown in Table 7 which are genes that displayed 2-fold higher or lower expression in Pi-replete pho2 mutant roots in two independent Affymetrix ATH1 genechip experiments. When using the most up to date release 7 ATH1 annotations to check the probe ID to gene mapping, two are found to be ambiguous and one (267456_at) has no match. The implications of this mean that probe set Ids must be recorded as part of the database. The facility must also be available to check new annotation releases and where necessary remove or update results where the gene mapping has been amended.

| Probe ID | Gene Identifier | Annotation | *pho2* versus Wild Type (+Pi) |
|---|---|---|---|
| 258158_at | At3g17790 | Acid phosphatase 5 (ACP5) | 5.90 |
| 246001_at | At5g20790 | Expressed protein | 5.55 |
| 252414_at | At3g47420 | Glc-3-P permease | 5.45 |
| 246071_at | At5g20150 | SPX domain protein | 4.56 |
| 256597_at | At3g28500 | 60S ribosomal protein P2 | 2.45 |
| 266184_s_at | At2g38940 | Pi transporter Pht1;4 | 2.43 |
| 258856_at | At3g02040 | Glycerophosphodiester phosphodiesterase (SRG3) | 2.41 |
| 260097_at | At1g73220 | Sugar Pi transporter | 2.21 |
| 248770_at | At5g47740 | Expressed protein | 2.20 |
| 245928_s_at | At5g24780 | Acid phosphatase (VSP1) | 2.08 |
| 248970_at | At5g45380 | Sodium:solute symporter | 0.46 |
| 267456_at | At2g33770 | E2 conjugase (PHO2) | 0.17 |

Table 7: Published microarray ATH1 results (Bari *et al.*, 2006)

This is relevant for the collated data for this project as not all papers that describe microarray results will provide the probeset ID for the AGI number or the annotation version that was used and therefore, the data cannot always be updated to reflect the actual AGI that relates to the result. While at present this only affects 7% of results, this is still a high number and seems to be increasing with each new annotation. Additionally, there is a lack of uniformity between public databases as to which version of the annotation is used. NASC has only recently updated their website, but the datafiles are still holding TIGR5 annotation and genevestigator are still using TIGR5. There are also discrepancies between the commercial manufactures as Nimblegen (Nimblegen, 2008), for example, are using probe mapping annotation from the TAIR 6 release whereas Affymetrix have their own version of annotation which is based TAIR 7 that they implemented in Nov 2007. When mapping the probe ID to annotation, Affymetrix use their own method which tracks five levels of relationships

between IVT Probe sets and the current transcript record (Affymetrix, 2008). This annotation example is very typical of the problems that face bioinformatics projects and demonstrates that the data and surrounding models are not static and this is an important factor when developing the data model for this study.

### 4.5. Data Quality

The previous sections describes the difficulties that can be encountered when collating gene expression results. The lack of standard annotation both in print and in gene expression databases is a problem. This will not simply be restricted to microarray data. It is possible that a gene from, for example, a Northern blot experiment identified as for example At1g12345 has been re-annotated as At1g23456 which would have an impact on any Northern blot experiment result. The only way this can be addressed is by looking up the locus history of an AGI number (which is available from TAIR) or by checking the current annotation of the EST (if available) of the gene. The most important point that has emerged from this is that it is not enough to simply rely on the AGI number for the result, but that there must be other ways to track the current gene annotation by either recording the version of annotation used (which is not readily available from the journals) or recording additional data such as probe ID for microarrays and EST's for Northern and RT-PCR experiments. In addition to the above, we need to remember that there is likely to be inaccuracies in the experimental results themselves which is why it is recommended that replicates of experiments are run. When considering results for the purpose of this project, we are assuming that the results from the experiment are correct, but what we are interested in is ensuring that we are getting full value from the results. If for example new updates of a genome are released, can we ensure that we are still holding results for the correct gene and if a gene was of previously unknown function, can we provide more information for this gene? The results accuracy will increase if duplicate references are found. This could be limited as journals require novel data and experiments are expensive therefore experiments are not likely to be routinely repeated. This is however a persuasive argument for ensuring that we keep all the data up to date as some experimental results will not be repeated.

The key reasons for maintaining the data is to ensure that the correct results are being attributed to the gene and to mine more valuable information from experiments that would previously not have been identified. For example, previously unknown function proteins can

be identified and historical gene expression results allocated to them; ESTs when allocated an AGI number can add new data about a gene that was previously unlinked (for example a gene that was given 5 different accession numbers by different authors turns out to be the same one that is regulated by five treatments) (Button *et al.*, 2004).

### 4.6. Data Model Requirements Summary

The previous sections of this chapter have examined the types of experiment used for stress response gene expression and the results each experiment yields and whether it is suitable to compare data across platforms.

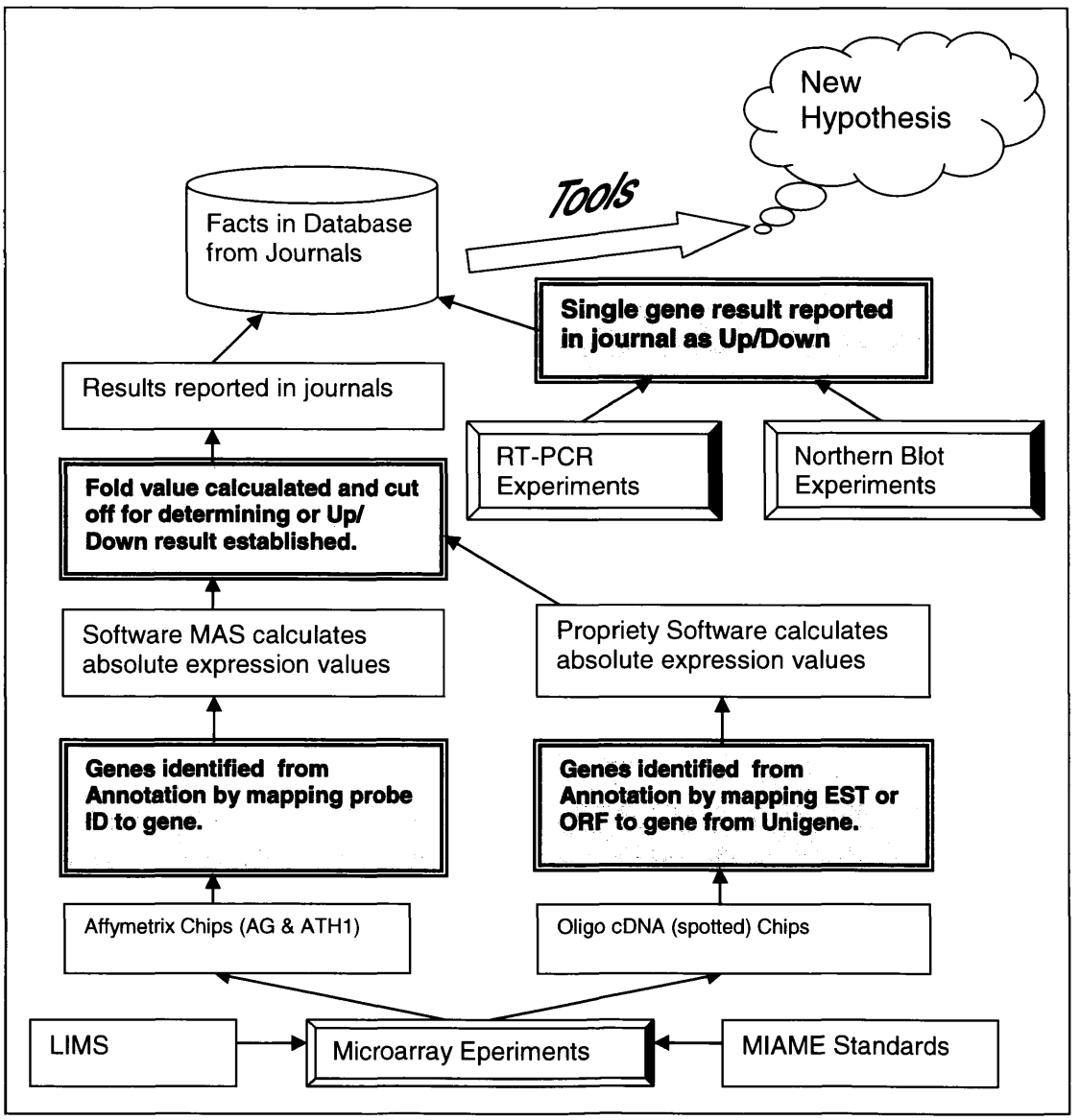


Figure 20: Layers of Gene Expression Data. From the results of the case studies made during this project, the process of obtaining and reporting results has been built into this flow chart. The grey boxes highlight the key points of the process that must be examined to establish whether the most upto date protocol for nomenclature or calculation of fold value have been applied.

It is now necessary to identify the generic elements of the data that will allow for comparison across each platform. Figure 20 provides an overview of the main stages for each type of experiment from construction to publication and shows the layers of data that will need to be considered when defining the data model. The four boxes highlighted in grey are the areas that are identified as a priority to be considered when developing the data model for this projects' database. These are the areas that may change or be partially missing from the paper or require to be updated as new annotations or nomenclatures evolve.

The first consideration is how to record the gene that the expression result is for. While there are many differences between the two-channel and single-channel microarray, for the purposes of this project, the main difference is the way in which each platform annotates a probe to a gene. As Figure 20 shows, the single-channel microarrays use a probe ID which is then mapped to a gene AGI or locus ID whereas the two-channel arrays tend to use longer sequences that relate to known EST's and are identified by an ORF or an accession number. For Northern blots and RT-PCR experiments, these tend to be identified by either the gene or accession number. This means that it is desirable, where available, to record the probe ID, the accession number, the gene function, the gene name and the locus ID.

The second element is the actual gene expression result. The four experiment platforms yield different types of result – two are binary values of up- or down- regulation while the other two can provide quantitive values that are either relative or absolute. The only way to compare the expression results across experiments is to record all the results as binary up- or down-regulation. This potentially could reduce the value of some of the microarray results, so an additional requirement for microarray experiments only would be to record the fold change value and the authors cut-off call. For all experiments, the temporal data for the gene expression result should be included where available. The reference and the experiment type should also be recorded. This chapter has described the main areas to consider when exploring the gene expression data from the published papers. Section 2.3 described the

problems that biologists encounter with their current system and Table 8 assesses if these problems can be addressed by the data requirements established by this chapter.

**Q.** Up-dating (and adding new information) is slow and laborious.

**A. Data will be linked to up to date annotation such as TAIR V7 and Unigene so updating should be an automated process**

**Q.** There is no temporal dimension within the diagram - No account is taken of time or 'dose/amount' of response and the dynamics of the interaction are very poorly represented.

**A. Temporal time for experiments from all platforms will be recorded and fold value will be recorded for all microarray experiments**

**Q.** It is difficult to incorporate information on differential induction of certain genes in different plant tissues (e.g. roots vs. leaves).

**A. Data will be linked to GO ontology and Plant Ontology Consortium information which can provide this information as long as there is a unique identifier for the gene**

**Q.** The importance (and interdependence) of proteins in different intracellular locations is sometimes poorly conceptualized.

**A. Data will be linked to GO ontology information which can provide this information as long as there is a unique identifier for the gene**

**Q.** It is not possible to draw separate diagrams for each agonist/response as it would be too time consuming.

**A. To be addressed at the query building stage**

**Q.** It is difficult to indicate the source of information i.e. whether it is derived from Arabidopsis or another plant, or whether it is from another eukaryote, or the source of the publication.

**A. The database model will be constructed and indexed to allow easy reference to all the specified information sources.**

**Q.** It may include varying degrees of uncertainty ('informed guesses') that other scientists may find inappropriate or are wrong (by virtue of having not taken into account some other published information).

**A. To be addressed at the query building stage**

**Q.** It is difficult to add information on 'unknown' ESTs.

**A. Data will be linked to up to date annotation such as TAIR V7 and Unigene so updating should be an automated process**

**Q.** It is not possible to interrogate a diagram.

**A. To be addressed at the query building stage**

**Q.** It is not possible to add to the diagram ones own personal or unpublished data.

**A. To be addressed at the query building stage**

Table 8: Biologist requirements vs. Data Requirements

The next chapter will describe the process of creating and implementing the data model based on the findings of this chapter.

# Chapter 5 Building the DRASTIC Database

## 5.1. Overview

This Chapter describes how the data requirements identified in Chapter 4 are converted into database design and implementation. The process used to develop the database has been based on an established database application lifecycle model (Connolly and Begg, 2002) and Figure 21 shows the main activities associated with the database design. The previous chapters have described the database planning definition and requirements collection and this chapter focuses on the database design, implementation, user interface design and testing.

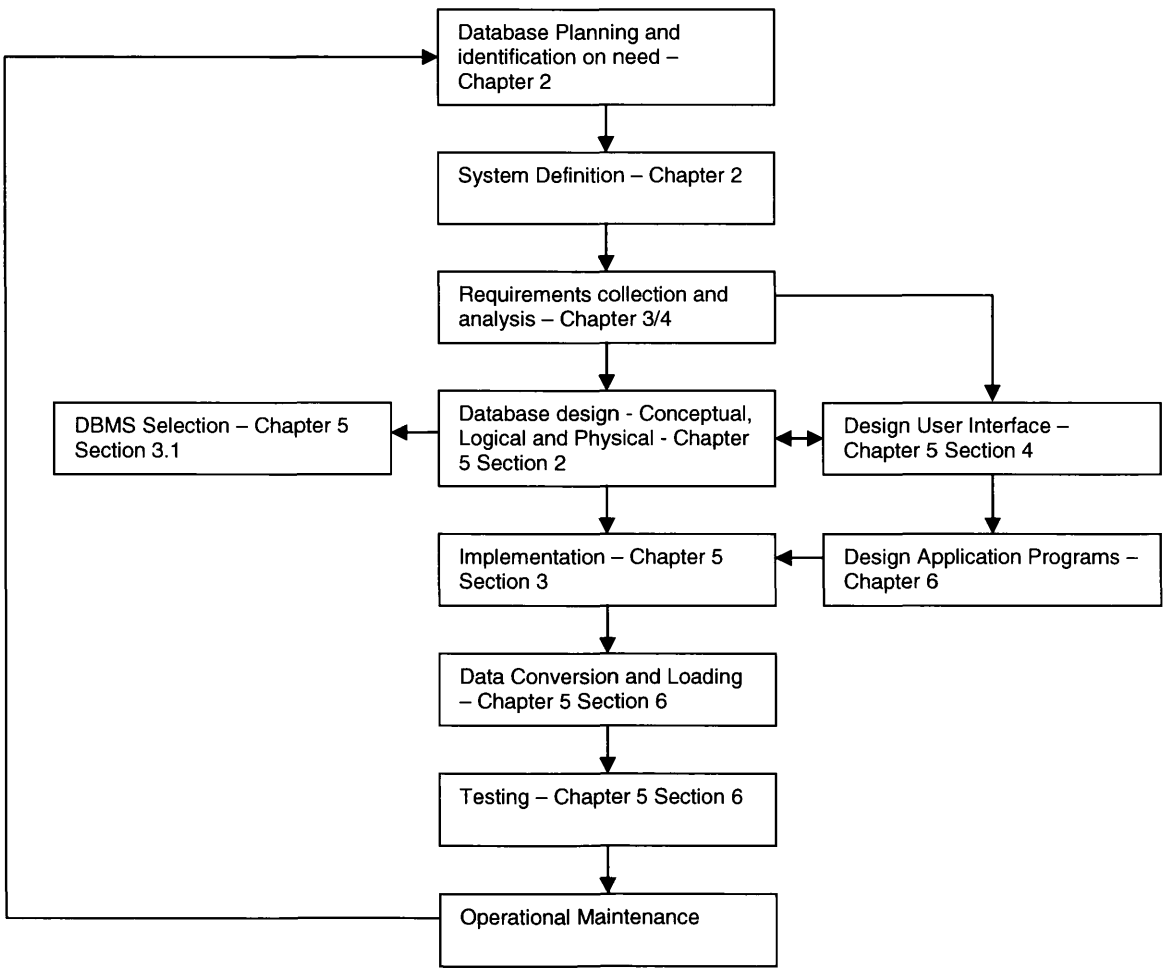



Figure 21: Software development cycle used for developing database application

## 5.2. Database Design

The key purpose of good database design is to ensure that the data is stored in such a way that it will enable the users to retrieve correct information from the data. This section describes the data modelling process of converting the identified user and data requirements into an efficient database structure.

### 5.2.1. Data Modelling Technique

The database design was developed using the principles of entity relationship (ER) modelling. The ER model was devised by Chen in 1976 and is a diagrammatic technique that provides a generalised approach to the representation of data and which is particularly suitable for the design of relational database systems. The key concepts of data modelling which are referred to in this section are the application domain which is the real-world environment in which this database is to be applied, entities which are a group of objects with the same properties which are identified by the application domain as having an independent existence, attributes which are a property of an entity and relationships which are an association or interaction between two or more entities.

This database is more problematic to design than a standard database due to the changing experimental technologies and progression in the subject area. Normally once the data model is established it is anticipated that no major change would be made to its structure once the data is added (Whitehorn and Marklyn, 2002). When designing this database, the data model has to be as flexible and as future proofed as possible due to the different formats and the evolution of data. For example, if we were dealing with an address system, it is unlikely that the attributes for an address would change but this is not the case with the attributes of this system. Due to the changing standards (for example the MIAME standards for microarray experiments) and the advances that are being made in the sequencing of plant data there are now more attributes available compared to the past up/down results of the Northern blots and there is no reason to expect this to stabilise in the future. The database has been robustly designed with this in mind and includes entities such as "protein", "requires" and "produces" which will enable the entry for future protein–protein signalling reactions.

### 5.2.2. Modelling the data

The Entity Relationship Model was developed in conjunction with biologists to model signal transduction. The concept of a Reaction models the data requirements and underpins the model. The main entities that are identified for this database are the References, Treatments, Hosts, Chemicals and the Reaction itself. Each record represents one report of a reaction which consists of the chemical which is uniquely identified by the chemical name, the gene name (if applicable) and the accession number (if applicable). The project is focused on data pertaining to the effects on gene expression of treatments so all reactions contain details of the treatments used and the host for the reaction which is uniquely identified by the genus, species and cultivar. The compatibility of the host is included (where appropriate) for pathogen experiments. The result of the experiment is recorded by the fold change to the gene expression and whether the gene expression was up/down or no change. The reference id is also included and the Reference table provides details of the journal and paper that the results were obtained from. The ER model is shown below along with a brief description of the primary entities and their attributes.

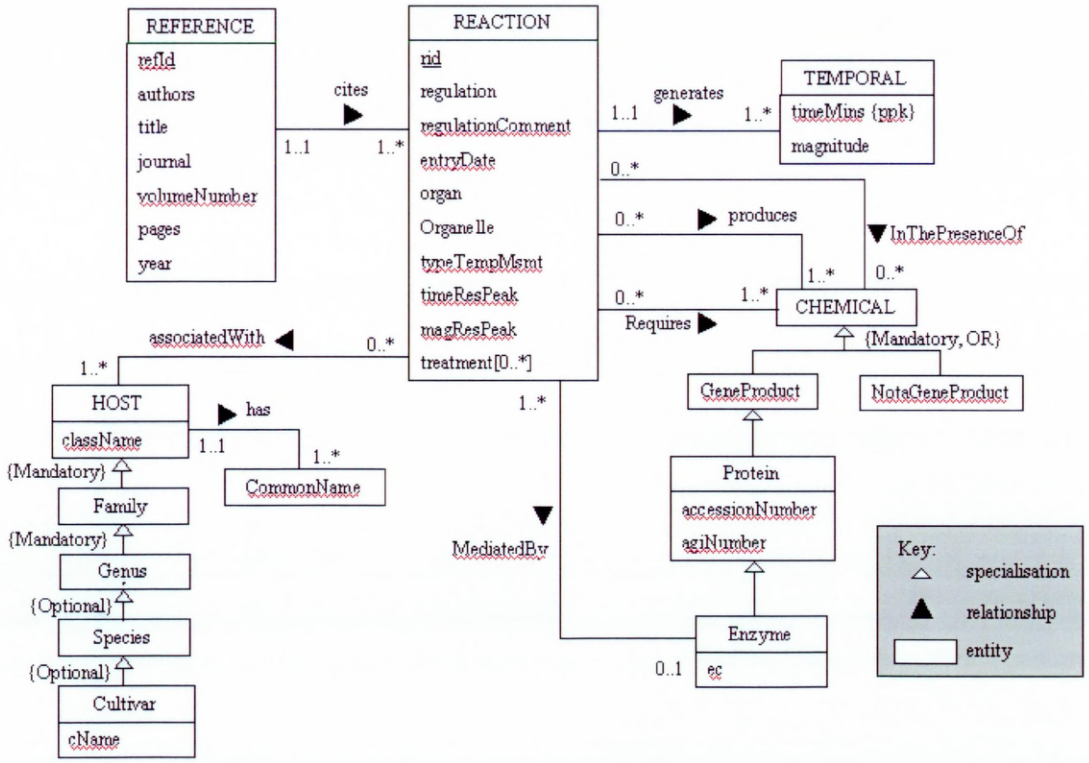



Figure 22: Entity Relationship diagram.

The entities in Figure 22 are represented by the square boxes, and the main attributes are listed in the box. The relationships are represented by the lines that join the boxes and the type of relationship is denoted at the end of each line - for example one to many relationship is represented by 1..1 --------- 1...*. The tables below describe each of the entities and attributes represented in the entity relationship model.

| Table | Description |
|---|---|
| ATH_GO | Contains the Gene Ontology (GO) terms for all the Arabidopsis Gene Index (AGI) numbers |
| Chemical | Details of all genes in the database |
| GeneDictionary | All updates made to the name, geneName or AccessionNumber attributes are stored |
| Host | Contains all the plant details including genus, species and cultivar |
| InThePresenceOf | InThePresenceOf, Produces and Requires contain chemicals which model a protein-protein |
| Reaction | |
| ListCommonName | Taxonomic records of the common name for each host |
| ListCompatibilityClassification | List table that stores all the compatibility classifications |
| ListCultivars | Taxonomic records of all the host cultivars |
| ListGenus | Taxonomic records of all the host genera |
| ListHostClassification | Taxonomic classification of the host e.g. plant-dicot |
| ListKnownTreatments | List table of all treatments |
| ListOrganelles | List table of all organelles |
| ListOrgans | List table of all organs |
| ListRegulationClassification | List table of all regulation classifications |
| ListSpecies | Taxonomic records of all the host species |
| ListTreatmentClassification | List table of treatment classification |
| Produces | InThePresenceOf, Produces and Requires contain chemicals which model a protein-protein reaction |
| Reaction | This contains all the gene expression records. Each record is comprised of expression data for a single gene. |
| Reference | This table contains the full details of each refereed paper that has been used to populate the database. |
| Requires | InThePresenceOf, Produces and Requires contain chemicals which model a protein-protein reaction |

Table 9: Explanatory Table Names

| Attribute | Description | Example |
|---|---|---|
| AccessionNumber | Contains the gene sequence entry number in EMBL format | AY271618 |
| AGINumber | Arabidopsis Gene Index (AGI) number | At5g03650 |
| compatibility | Certain treatment x host combinations can be classified as compatible or incompatible where the treatment is a pathogen. | compatible |
| classification | Defines the taxonomic classification of the host. | Plant-dicot |
| cName | Cultivar Name | cv Columbia |
| EC | Enzyme nomenclature | 2.4.1.18 |
| family | Defines the taxonomic family of the host. | Brassicaceae |
| geneProduct | States whether a chemical is a gene product | Yes/No |
| gName | Genus Name | Arabidopsis |
| geneName | Published or preferred symbolic name of a gene | ALDH7B4 |
| name | A description of the gene function | aldehyde dehydrogenase |
| refid | The identity number of the reference where the gene expression record is cited | Auto Number |
| regulation | In the case of a gene being expressed this attribute indicates whether it is up or down regulated. | Up |
| rid | The unique identity number of the record | Auto Number |
| sName | Species Name | thaliana |
| tDescription | Treatment Description | isolate O-264 |
| tName | Treatment Name | *Alternaria brassicicola* |
| timeMagPeak | Provides a summary of the scale of response | 4 Fold |
| timeResPeak | Provides a summary of the timing of response | 3 Hours |
| type | Treatment Type | Abiotic |
| typeTempMsmt | If temporal data is available then denotes type of data | |

Table 10: Explanatory Attribute Names - The above table does not include attributes such as Year or Author as these are self-explanatory

### 5.2.3. Defining keys for the ER model

Once the entities and their attributes are identified, the next step is to define the relationships between the entities. The important part of defining a relationship is to ensure that no two rows of an entity are identical. This is met by allocating each entity an identifying attribute(s) which is called the primary key. A primary key must contain unique data for each record and not a null value (Whitehorn and Marklyn, 2002). For the majority of the entities in the ER model shown in Figure 22, primary keys are easy to choose. For example the Reference Entity contains an attribute RefID which is a unique auto number field. As this field is unique and cannot be a null value, this satisfies the requirements for a primary key. However, for the chemical entity, the primary key was harder to select. The Chemical entity contains no single attribute that is unique, so a composite key of name, geneName and AccessionNumber was selected as the primary key. This uniquely identifies each record within the table, but because of the nature of the data, it is possible to have missing data or regularly changeable data within these attributes and this impact is discussed in Section 5.5.1.2.

## 5.3. Implementation

Once the ER model had been designed and agreed with the users, the database needed to be constructed. Consideration was given as to which Database Management System (DBMS) to use for the implementation of the database.

### 5.3.1. DBMS Selection

There were several aspects to take into account when deciding which database package to use.

1. The database needs to be free or have an existing license for use on both Abertay and SCRI systems
2. It needs to be easily transferable from one machine to another. SCRI will be hosting the planned website that will provide search tools for the database, but there will be no administrative access to the server from Abertay therefore the database will be constructed at Abertay and uploaded to the SCRI server. In addition, data input will be processed at SCRI so there must be access to the database front-end there.

3. Needs to have the facility to create a front-end that can be used to input data and provide robust validation checks for the data.

Microsoft Access was chosen as the database application as it is easy to use as a stand alone product, would be transferable via the servers used and would serve as a good prototyping medium. Microsoft Access also has its own programming language called Visual Basic for Applications (VBA) which can be used to construct data entry forms and thus enabled the database and front-end to be encapsulated as one package making transferring the program from machine to machine very simple.

### 5.3.2. Database Construction

The database structure was created in accordance to the ER model with empty tables ready for the data to be input. Each of the tables was set up with the primary/foreign keys, the relationships were created and all the data types set.

## 5.4. User Interface Design

Good user design is critical to the success of a system (Sommerville, 2001). An interface that is difficult to use will, at best, result in a high level of user errors. At worse users will simply refuse to use the software system. Because this interface is being built specifically for the task of collecting data, it is imperative that the interface design meets the user needs to ensure data integrity. The key principles of user interface design are user familiarity, consistency, minimal surprise, recoverability and user guidance and these have all been considered in the design process of the interfaces along with Human-Computer Interaction (HCI) as discussed below (Dix *et al.*, 1998).

### 5.4.1. Design Rules

Designing a user interface is never a trivial task, but for this database, particular attention had to be paid to how the input interface was designed due to the variety of ways that the data to be capture was presented as illustrated in Chapter 4. There are therefore two main parts to the design considerations: the design appearance and how the information can be most easily input irrespective of which method the data is presented.

During the design process of the database, the users were interviewed to assess what their user requirements were of the data entry screens. The decision to use a Graphical User

Interface (GUI) was based on the fact that the users were all familiar with the HCI of windows based systems and it would be inappropriate to expect users to be able to enter data directly into the database. A sample screen of the input GUI is shown in Figure 23.

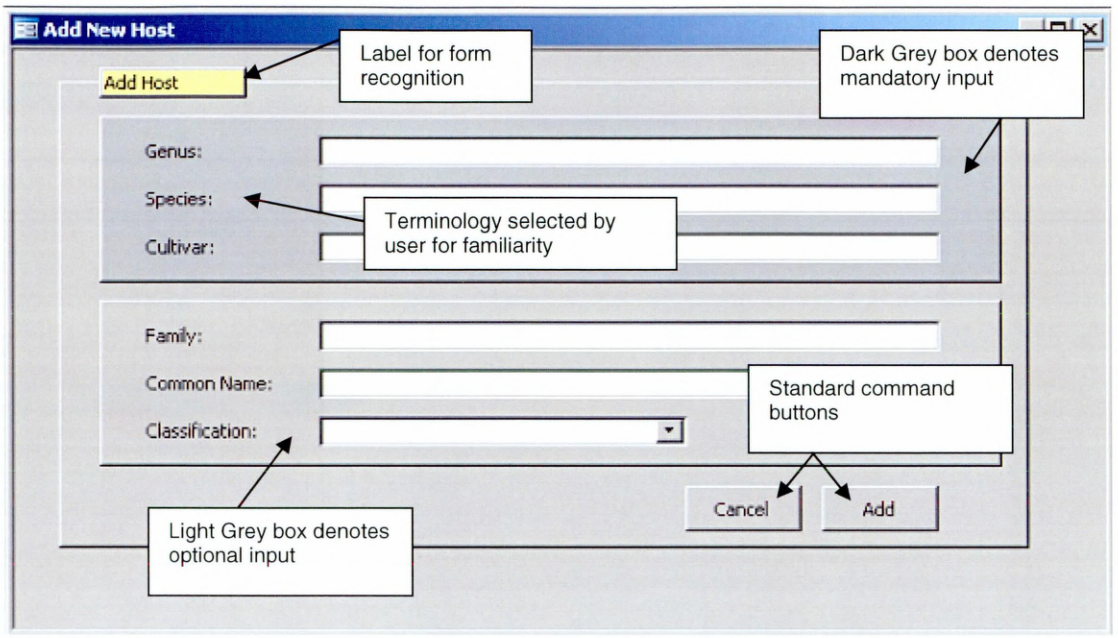



Figure 23: Example of the User Interface Design – the Add New Host form

In order to provide consistency and user familiarity, each of the data entry forms were split into blocks that represented the entities that the users of the system were familiar with. A yellow label is used to highlight the area that each section represents. The names used are the ones allocated by the biologist to assist with quick recognition and help users who are unfamiliar with the system. Each of the entry sections is in a 3-D box – the dark grey box indicates mandatory fields and the light grey box indicates optional fields. The fields are all labelled using the terminology selected by the user. In order to meet the requirement of minimal surprise, all buttons included on the forms use either a standard windows image or the same wording throughout the application. There is always a cancel button available and all complete records are automatically saved to provide recoverability. Error messages are not generalist, but will actually inform the user which of the fields they have not completed or the cause of the error giving the user good guidance. The adding of records is hidden to the user and even if a form has collected data which needs to be add records to several different tables, this is not visible to the user and they only have to press one update button to add their

entry. This GUI scheme is used on every form in the application for consistency and to make using the application easier.

### 5.4.2. Input Screen Design

The design rules that are described in Section 5.4.1 provide the basis for the appearance of all the forms. The next step is to decide how to present the screens to the user in a logical sequence that will minimise the chance of user error and to provide the fastest means of user entry. With such a richly structured database, data entry is difficult due to the order data is required to be input (dependant on the primary/foreign keys). Data input screens were designed around these issues. In addition, consideration was given to the different types of format the data came in. Chapter 4 describes what data is required for each record and the different formats that the user may find the data. Chapter 3 describes some of the ways in which data is uploaded into databases, for example, tab delimited files with identifying columns. Automatic loading of data is not appropriate here as the data is found in many different formats (including diagrams which would be unconvertible) and the time taken to convert the files would negate the efforts made to convert them.

Therefore, until there is more uniformity in the way results are published in journals either in the article itself or the supplementary material, the data must be manually input. Several different types of input process were identified:

1. Journal reports one result
2. Journal reports multiple results from same gene, same species but different treatment
3. Journal reports multiple results from different gene, same species, same treatment
4. Journal reports multiple results from same gene, different species, same treatment

In case 1, there would be no data that would be the same as another records, however in cases 2-4 there is duplicate data input for multiple records. As cases 2-4 are most likely, the input screens needed to be designed to enable the user to reuse data that they had entered from the previous record to prevent the need for re-entry.

Because of this requirement it is likely that there will be several elements of each reaction that are already in existence in the database and need to be found rather that added to the current reaction. As the user placed an extremely high priority on data quality, reducing

errors introduced to the system during the input had to be a main concern in the design of the forms. Common types of user error are transposition of numerical digits, misspelling of names, repetition of characters etc, and if these errors are introduced to the database they are very difficult to pick up and will result in searches missing out potentially vital data (Ritchie, 2002). For example, as previously mentioned, the only way to uniquely identify genes was to create a composite primary key comprising of gene description, gene name and gene accession number. Given the complex names that are allocated to gene descriptions and the likelihood of change to gene name (as shown later through the gene dictionary), it would be unwise to allow the user to enter each gene in manually each time.

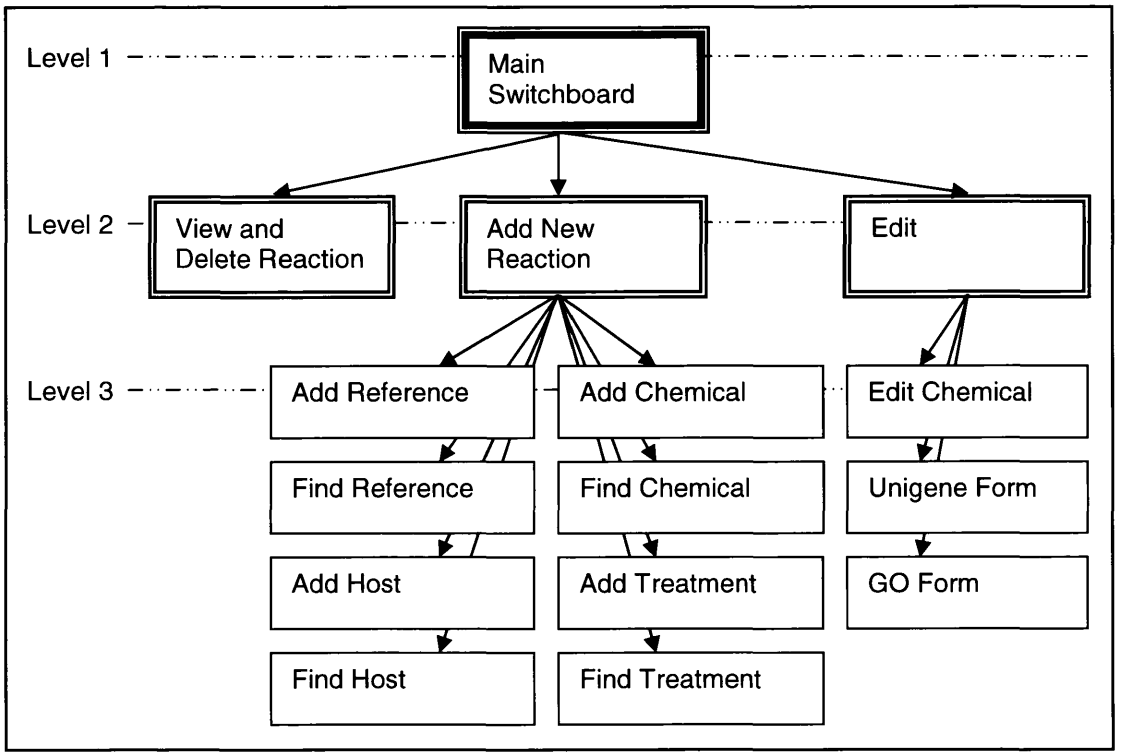



Figure 24: Map of forms

In order to satisfy these requirements, the input screen has been designed to have one main screen that allows access to all the other input screens and also displays all the current information that the user has entered. The main reasons for this choice are:

i. Obvious to the user what information is not input

ii. More robust as the order in which information is input is controlled

iii. Enables better typographical error control by enforcing the user to search for existing records first

iv. Allows the data entered from the previous record to be kept so data can be re-used where applicable for multiple reactions.

v. The database structure requires that a related record exists in the host, chemical, reference, organ, organelle and treatment entities before a new reaction can be entered and this provides an easy way to enforce this.

There are three types of form that the user will encounter when entering a record: The main Add New Reaction form shown in Figure 25 which is the master form (described below), there are 'sub forms' which the user can add new records for a specific entity (an example Add Chemical is shown in Figure 25 and 'search forms' where the user can search to see if the entry they are looking for in a sub form already exists (an example Find Host is shown in Figure 27). The map of the data entry forms is shown in Figure 24.

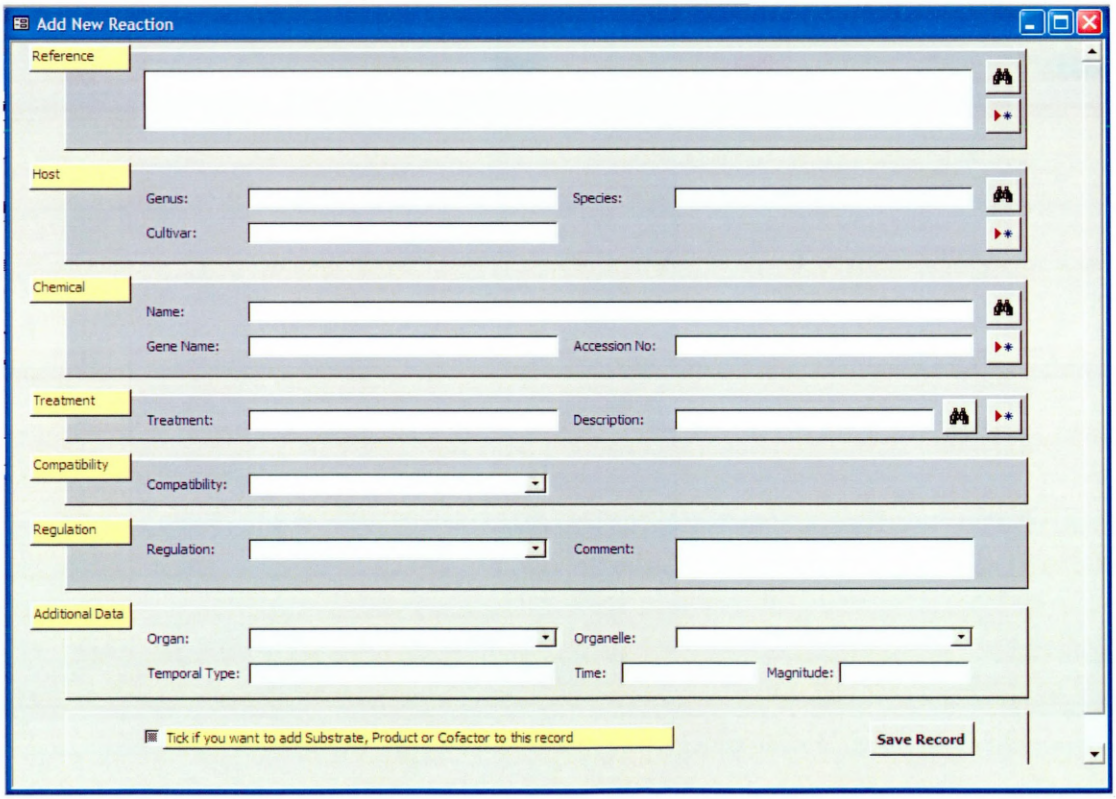



Figure 25: Picture of frmAddNewReaction. The form has been designed so that all sections are clearly separated and mandatory fields are highlighted in dark grey as described in the design rules section.

The Add New Reaction form contains some fields that are mandatory but the user cannot directly input data into the main form. They must choose to search for an existing record or add new both which launch a new sub form. When the data is input to the subform, a record is created in a corresponding table and the data is then copied into to the fields on the main form. For example, if a user wanted to add a reference, they would have to use the form Find Reference (described below) to find the appropriate reference or form New Reference to add the new reference. This ensures that there is no duplication and that a correct record exists. Once the user has chosen/entered the reference, the details are passed back to the form Add New Reaction to be used in the creation of the new reaction record. This is the same procedure used for Genes, Treatments, Chemicals and Host.

The Search command buttons and the Add command buttons open new forms. The Save Record command button performs several validation checks before saving a new reaction. It:

- checks that there is an entry in all mandatory fields or produces an error message
- checks for any blank optional entries and if found inserts the correct response according to the rules for null entries
- updates the lookup tables that are related to the non mandatory entries for example ListOrgans
- inserts a record to the Reaction table. There will already be a record for Chemical, Reference, Treatment and Host as these need have already been entered in the process of completing the form.

Validation rules are also built into the form where applicable for example the Magnitude field uses the validation rule: *txtMagnitude].[Text]="" Or IsNumeric([txtMagnitude].[Text])=True* to enforce numeric input to this field and all fields that require input to be first entered to the sub forms and then copied to the main form are disabled.

The Add New Chemical form is shown in Figure 26 and is typical of all the sub forms that enable the user to enter details about specific entities. This particular form allows the user to add new chemicals to the database. The user simply enters the details and then presses the Add command to return to the main form.

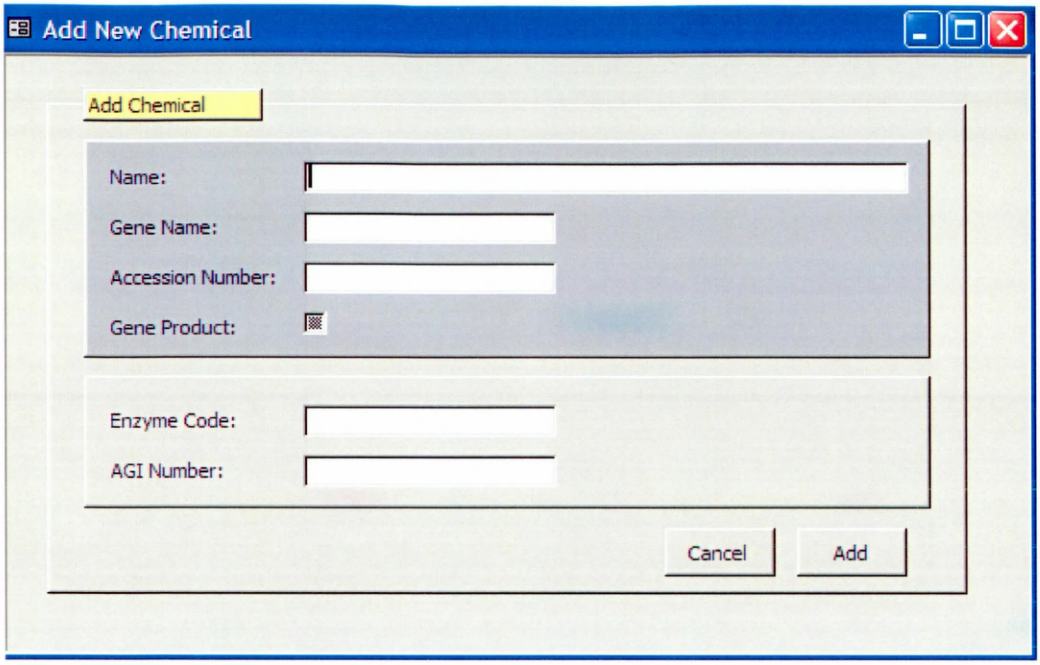



Figure 26: Picture of form Add New Chemical.

The Add command button performs the following functions:

- It checks that there is an entry in all mandatory fields or produces an error message
- It checks for any blank optional entries and if found inserts "not available"
- It inserts a new record to the Chemical Table. If at this point a matching record is found, the insert new record command is cancelled and the data is copied across to the Add New Reaction form.

The values from the Name, GeneName and Accession Number are passed to textboxes in form Add New Reaction. The Cancel command button closes the form.

The Find Host form is shown in Figure 27 and is typical of all the sub forms that enable the user to search for details about specific entities. This particular form allows the user to search for a Host that is already entered in the database.

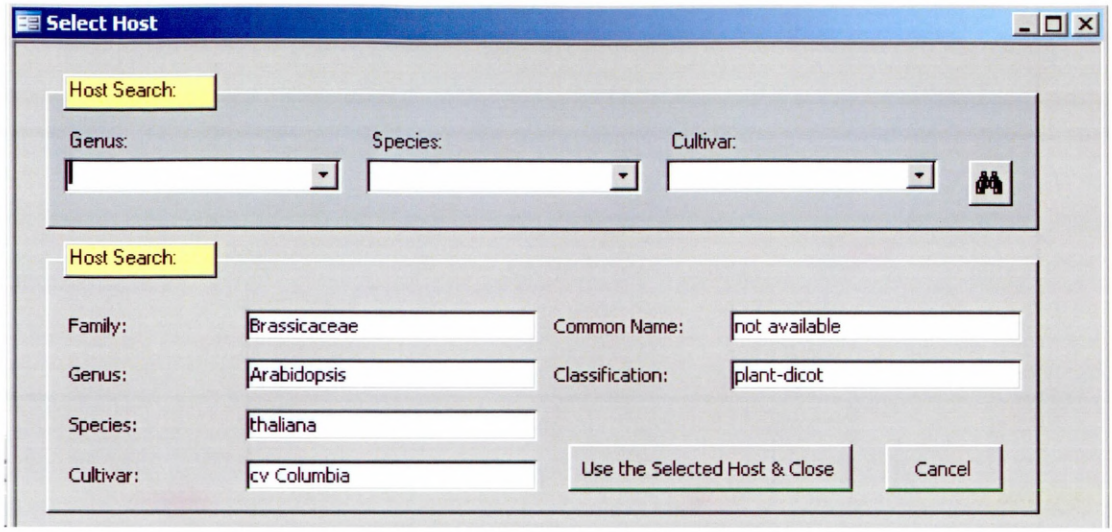



Figure 27: Form Find Host

The search method is very versatile to enable the user to quickly find the record that they are looking for in several different ways. The three combo boxes contain all the genus, species and cultivars in the database, but if the user selects a genus from the combo box, the remaining combo boxes are updated to contain only records relating to the selected genus. The user can type in the Host that they are looking for in the combo boxes and then click the search button to find the record(s) they are looking for. The system also supports the use of wild cards to enhance the search facility. Once the user has found the record they want, they simply click the Use Selected Host button and this transfers the data to the main Add Reaction Form. If the host cannot be found, the user can click the close button and return to the main form to add the host. This is the basis of all the search sub forms to enhance the usability and learnability of the system.

### 5.4.3. Editing Screen Design

As described in Chapter 4, it is very common for gene names to evolve over time, and unknown or null values be filled (for example gene function) due to new experimental discoveries. Part of the requirements of this project is to enable this data which affects primarily data in the chemical entity of the ER model to be updated. However, due to the data modelling constraints, editing and deleting data are not trivial tasks and screens have been developed to automate this. This editing facility provides this database with much richer information than data simply published in journals as it allows simple tracking and updating of names and experimental results that would be unable to manually be traced.

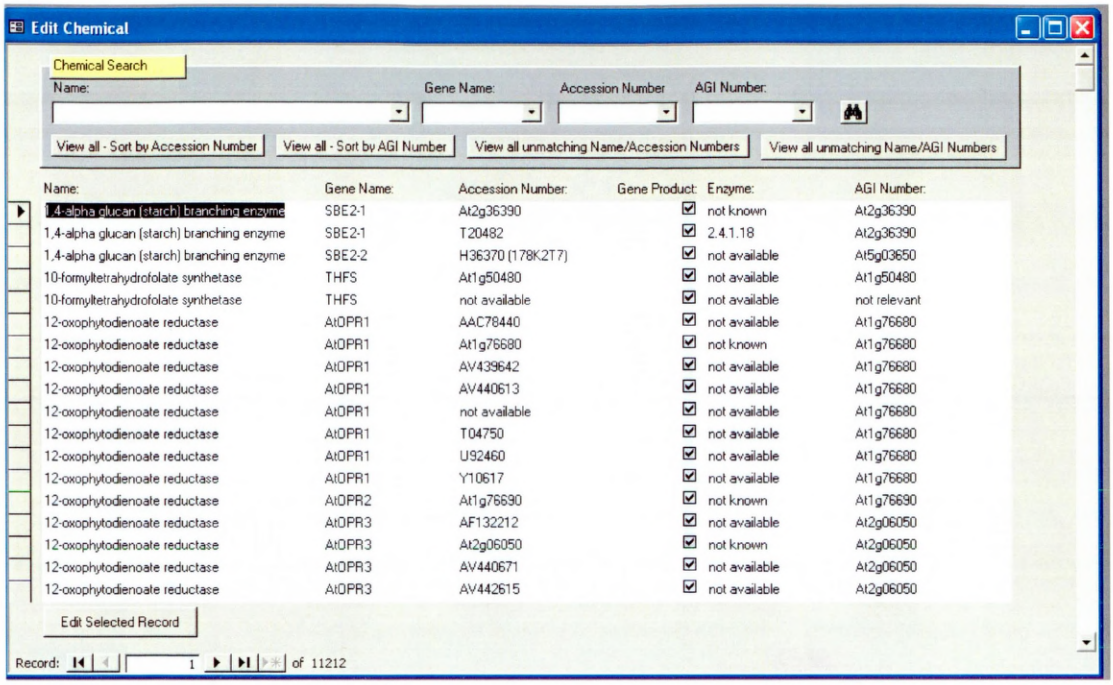



Figure 28:The Edit Chemical Main Form

The Edit Chemical Screen shown in Figure 28 allows the user to view the elements from the Chemical table. The screen has been designed using the same design principles as described in 5.4.1. The top box allows the user to search using typed or selected items from combo boxes to find the chemical they wish to edit. Alternatively they can scroll through the records displayed in the bottom part of the screen.

In order to edit a record, the user must click to highlight a field and then click on the Edit Selected Record command button. This displays the Edit Chemical sub form shown in Figure 29 and enables the user to edit all aspects of the selected chemical.

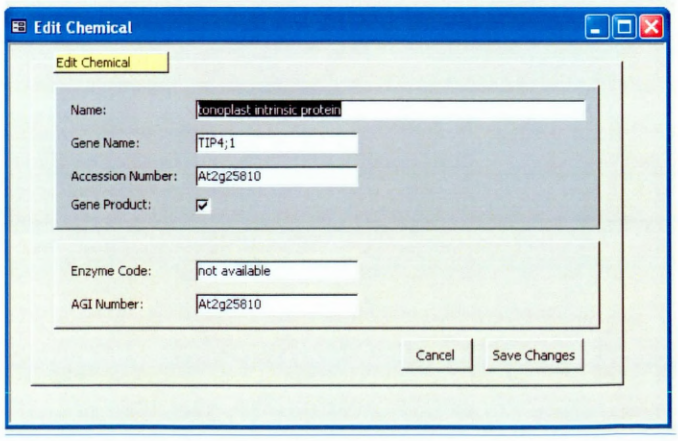



Figure 29: Edit Chemical Sub Form

When the gene name is altered, a check is made to ensure that this will not cause any errors with the keys in other tables. Providing this check is satisfied, an automated process rolls out the changes to all applicable tables within the database.

The delete function couples as a view record facility as shown in Figure 30. The user can scroll through the records or enter the record number that they are looking for. When the delete command is clicked, the current record in view is deleted and where appropriate, any associated records from other tables.

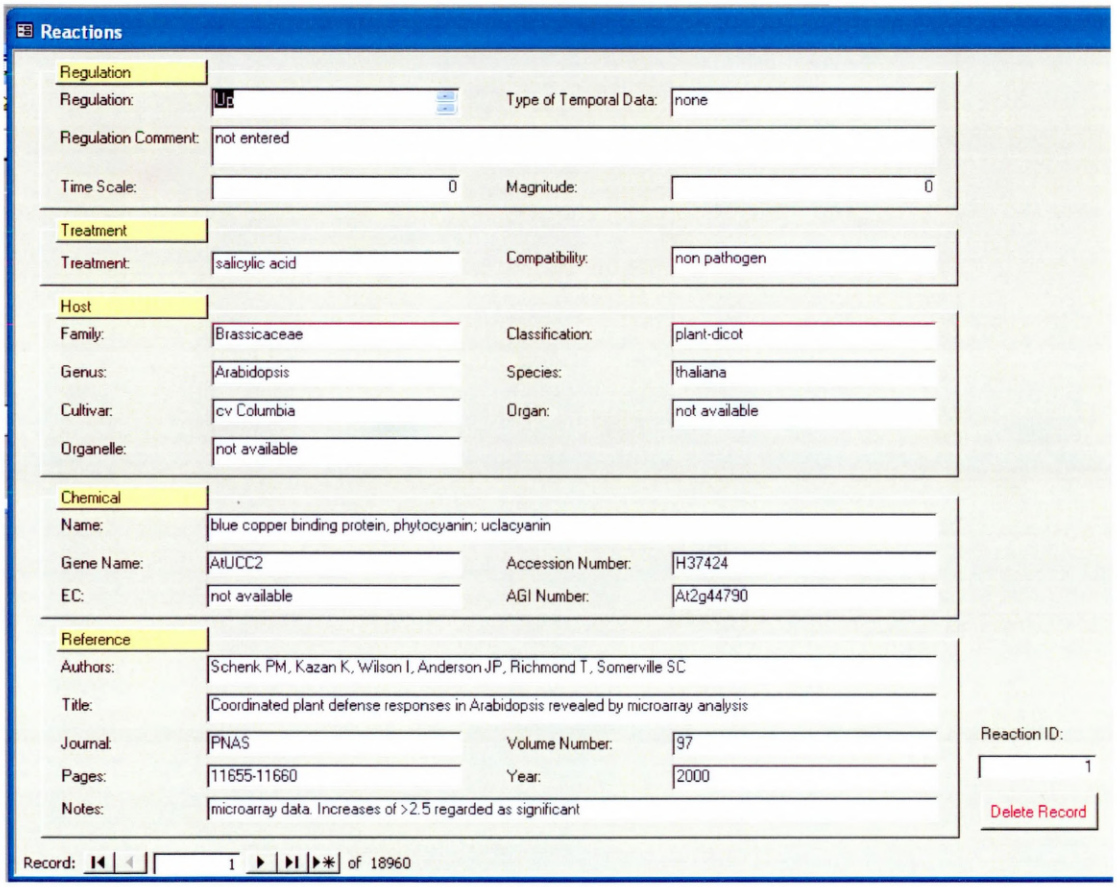



Figure 30: View/ Delete Reaction form

## 5.5. Database Integrity

One of the key reasons for developing this database has been to create a store of gene results that are consistent and up-to date in order to search and gain information from this data. As the level of accuracy of the information that is retrieved is directly related to the level of accuracy of the data that is stored in the database it is imperative that the data is of high quality. Database integrity refers to the correctness and consistency of stored data (Connolly

and Begg, 2004). There are several levels of consistency to consider: domain-level integrity, entity integrity, and referential integrity which are discussed in the context of this database in Section 5.5.1. Database correctness is addressed in Section 5.5.2 where some of the validation tools that enhance the consistency of the database are explained.

### 5.5.1. Database Consistency

Database consistency implies that the data held in the various tables of the database is consistent with the concept of the relational model. The following sections examine how this consistency is applied to this database.

#### 5.5.1.1. Domain–level integrity

Domain-level integrity ensures that the structure of each field is sound and the values in each field are valid, accurate and consistently defined throughout the database. For example, the GeneName attribute is repeated in the database and it is always a text data type. Each attribute has a domain or set of values that are legal – for example the Microarray attribute can only contain the values 'Yes', 'No' or 'Not Known'. This means that the attribute domain for Microarray is a Text data type consisting of only the three aforementioned strings. These constraints are identified during the design phase where each attribute is examined and assigned a domain. To ensure that this integrity is not violated during the input of new data, integrity constraints are developed and implemented in both the database and the input screens. This reduces the chance of user introduced errors in the system.

In the database itself, fields that are required are tagged, and default values are set for non-required fields where it would not be appropriate to have null values. Additional tables have been included in the database to act as lookup lists. For example, table ListOrganelles contains all known organelles within the database. In the input screens, the user is presented with a combo box that contains this list. The user may only select an organelle from this list. If a new organelle is found, it must first be added to the table. This technique is used for several entities within the database to prevent typographical errors being introduced and also ensures that the same object is not defined twice (for example American spelling/ English spelling) which assists with future data mining. Validation rules are built into some of the input fields to enhance data integrity (for example to prevent users entering letters where numbers are required).

#### 5.5.1.2. Entity Integrity

Entity integrity is dealt with by the primary keys and ensures that there are no duplicate records held in tables. Section 5.2.3 refers to how the primary keys were assigned and outlined the problem that is caused by the changeable nature of the data. There is a large variance with respect to the amount of data available for each published experiment. This means that it is very likely that many records that the biologist will want to include will not contain entries for all the attributes. Another problem is that some of the data is missing because it has not been discovered experimentally yet. This causes problems with both domain-level and entity integrity and rules need to be established to decide how the missing data will be dealt with. Missing data causes problems for several reasons:

1. If there is missing data in a primary key it will cause an error in the database as null entries are not allowed
2. If there is missing data in a required field that is not necessarily a key field this will again cause an error.
3. If there is missing data in a non-required field it needs to be decided if the field will be null or substitute entry made as null fields in databases can cause conflict errors.

The problem with null fields is that null represents missing or unknown values – null does not represent zero or a string of text or one or more blank spaces. This means that nulls do not have a data type (for example they are not classed as a text or a number type) and if you were to use a mathematical sum with a null e.g. 1 * null, this would create an error.

The data is such that some of the fields that make up the primary key may be unknown at present but will be allocated a name or description at a future date, for example in the chemical table, the attribute name is occasionally missing or not yet allocated. This is an unusual situation in database design as for entity integrity where no component of a primary key is allowed to have a missing value of any type (Ritchie, 2002).

Consideration then had to be given as to what should be input into the fields where the experiment had been accepted for data inclusion, but there was missing data entries in the primary key fields. A set of rules was created which included rules for handling missing data in the primary key fields and also the non-mandatory fields. It was decided that it would be more consistent to avoid null attributes. This is due in part to the problems that null fields can

cause in a database, and also because it was observed that the biologists already had some existing unwritten rules of using "not applicable" or "unknown" rather than leaving empty fields. In order to ensure consistency of data, a list of rules was drawn up to provide guidance about what to enter in fields where there is the potential for null input e.g. description = "unknown" and these rules were also built into the database design. It is vital for the success of the database that these rules are adhered to as for the planned searches to be successful, if the user is looking for "unknown" but "not applicable" has been used in its place this will cause the search to fail. This is also important as it allows us to identify records with missing data and query external data sources on a regular basis to find out if new information is available.

#### 5.5.1.3. Referential Integrity

Referential integrity ensures that a pair of tables are synchronised whenever data is entered, updated or deleted from either table and should ensure that the data of one table does not contradict the data of another table. Specifically, every foreign key value in a table must have a matching primary key value in the related table. The input screen and edit screen design ensures that referential integrity is maintained as they control the order in which data is added to the database tables. The input screens are set up to prevent attempts to add duplicate records to tables and ensure that a record containing the primary key exists before a record is added to a child table. The edit screen was more problematic as the user could potentially be altering primary key values which would cause a referential integrity error in the corresponding foreign key values. The inbuilt procedure for the editing chemical records is:

- Check if the change will cause a duplicate record entry in the chemical table
- If not, then add the new record to the chemical table and then amend all related records in the reaction table by running an update query to find all the records that match the old primary key values (the foreign keys) and updating them. Finally delete the old record from the Chemical table.
- If adding the edited record to the chemical table will cause a duplicate record entry in Chemical simply check to ensure that the most up to date data is held in the none key fields of the chemical record and update this record, update the reaction table records as previously described and in both cases, update the Gene Dictionary table.

### 5.5.2. Data Correctness

Data correctness implies that the data capture for entry into the database does in fact correctly represent the 'real world' data that it is supposed to. This database has to tackle many quality issues including naming conventions, annotation updates, errors in the experiments themselves and gene name updates and this section describes some of the tools that have been developed to provide further validation and correctness to the captured data.

#### 5.5.2.1. Gene Name Updating

The need to standardise gene nomenclature is important (Lyon *et al.*, 2002), thus the names used in the database correspond with current National Centre for Biotechnology Information (NCBI) Unigene classification rather than those cited in the original publication, unless a more recent primary publication indicates otherwise. Sometimes changes in gene identification are small but in other cases they can be dramatic and critical if signal transduction pathways are to be correctly understood. Figure 31 shows the unigene form which is a subform of the Edit Chemical form described in section 5.4.3. The unigene form provides the user with a helpful check facility when they are updating gene details. If the user clicks on one of the Accession Numbers of one of the chemicals displayed in the Edit Chemical form, the Unigene form is displayed. The Unigene form selects the Accession Number of the particular chemical and using Extensible Markup Language (XML) technology connects to NCBI's Unigene database and retrieves the gene name, Unigene ID, gene function and AGI number that Unigene holds. This allows the user to easily find out comparative data from a separate source during the editing process of a gene and is particularly useful for genes that may have an unknown name or AGI number or genes that have conflicting data. As previously mentioned, NCBI's Unigene database has been chosen as the gene details standard because it is the only large database that contains gene standardised data for multiple plant species (but not gene expression data).

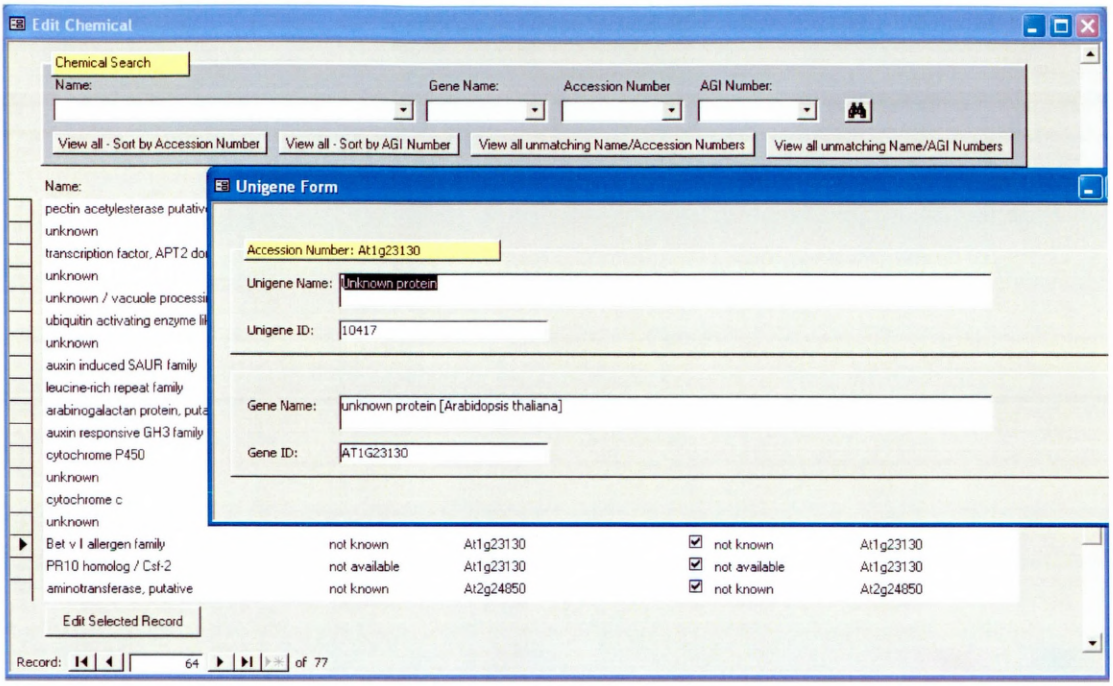



Figure 31: Unigene sub form from the Edit Chemical form of the database.

Feedback from the user was very positive for the Unigene search and further investigation showed that there are around 12% of genes with unknown function in the database. Knowing the function of the genes is very important and can give insight into what role it may play in the signalling process. Therefore regularly checking and updating information for these genes is vital and is where the database differentiates itself from simply collecting static results reported in journals. To enhance this updating process, a routine to select all genes with unknown function and batch process these records through NCBI's Unigene database to establish if there are any updates has been created. UnigeneSearch is a companion program for the database that allows the user to search for all the Accession Numbers that have 'unknown' in the corresponding gene name attribute or 'not available' in the AGI attribute. The selected Accession Numbers are loaded into a text file and the program accesses NCBI's Unigene and Gene databases to find out if the gene function has been established or an AGI number found for each of the Accession Numbers. In order to use the package, the user is prompted to select a text file to load. The file name the user selects is used as the name of the excel file the results are saved in – which is saved to the same directory as the UnigeneSearch.exe file resides in.

The results, of which a sample is shown in Table 11, display the Accession number, the gene id, the unigene name, the gene name and links to the geo, homolog, nucleotide and gene databases for the particular Accession Number. The program is separate to the database as the results need to be manually checked and not automatically updated.

| Data from database | | Results from NCBI | | |
|---|---|---|---|---|
| **Accession Number** | **AGI** | **Gene ID** | **Unigene Name** | **Gene Name** |
| 103C7T7 | At5g42530 | At5g42530 | Expressed protein (At5g42530) mRNA, complete | expressed protein [Arabidopsis thaliana] |
| 104A7T7 | not available | Not Found | Not Found | Not Found |
| 105P15T7 | At3g15450 | At3g15460 | Expressed protein (At3g15450) mRNA, complete | brix domain-containing protein [Arabidopsis thaliana] |
| 110F10T7 | At5g45500 | At5g45500 | Expressed protein (At5g45500) mRNA, complete | expressed protein [Arabidopsis thaliana] |
| AI618746 | At5g51550 | At5g51550 | Phosphate-responsive 1 family protein (At5g51550) mRNA, complete cds | phosphate-responsive 1 family protein [Arabidopsis thaliana] |
| AI618753 | At3g46540 | At3g46540 | Epsin N-terminal homology (ENTH) domain-containing protein / clathrin assembly protein-related (At3g46540) mRNA, complete cds | epsin N-terminal homology (ENTH) domain-containing protein / clathrin assembly protein-related [Arabidopsis thaliana] |
| AI618755 | not available | At1g66860 | Expressed protein (At1g66860) mRNA, complete | expressed protein [Arabidopsis thaliana] |

Table 11: Sample results from UnigeneSearch.exe program. The first two columns on the left show the data from the project database. The three columns on the right shown the results obtained from the NCBI databases.

The reason that the results are quality checked first and not simply uploaded is that the unigene results sometimes returns "Expressed protein" or similar phrases for the gene function which is the equivalent of 'not known'. Genes with this result must be maintained as 'not known' in accordance with the data rules and checked again at a later date.

The results in the last three rows of Table 11 which are highlighted in grey demonstrate the usefulness of this search as new information is uncovered. In two examples we now have a much more detailed description for the gene function and in one case, we now have an AGI number for an accession number. For the Arabidopsis data, the database is very useful for linking together reactions based on Accession Numbers described as genes of unknown function. By using the unigene tool to convert accession numbers into AGI numbers (as

demonstrated in the Unigene example in Table 11) we have established that that all the following accession numbers reported in separate journal articles are either down-regulated by chitin (H37231, R90140, T41806), drought (AV823744), ethylene (R90140), low oxygen (At2g10940), or sodium chloride (AV823744), or up-regulated by salicylic acid (R90140, H37231) are actually the same gene i.e. At2g10940. This provides much more insight to how this gene responds in the signal transduction process and provides a set of related data (Button *et al.*, 2004).

#### 5.5.2.2. Gene Dictionary

There is a gene dictionary which tracks all changes made to the name or AGI number of a gene during the editing process. This was designed for two reasons, one to capture the type of changes that are made to the gene data, and two so that we can backtrack our gene expression results to show how we reached the currently assigned name by providing historical terminology, and if necessary correct an error made in the editing process.

When a chemical record is edited only one record is altered in the Chemical table. However, because the gene name is a foreign key in the reaction table, if this is modified then all child records in the reaction table must also be amended which can mean modifying hundreds of records. The Microsoft Access database is designed so that once a record is changed it is saved and it can be difficult or impossible to return the database to the pre-edited state. Because of this it seemed a wise precaution to construct a data trail of the alterations to the chemical table as it has been identified at the planning stage that it was likely there would be a large number of changes. At the time of writing there have been over 7000 modifications to the chemical name or AGI attributes. The contents of the Gene Dictionary are interesting as it shows that the gene name changes vary from subtle semantics such as removing a comma or modifying abbreviations to a huge leap for example from an unknown gene name to an identified protein. Table 12 shows some sample records from the gene dictionary. The changes made are underlined in bold and show the complexity of the gene names. There are several examples of small changes to the gene name format and some typographical errors which demonstrate the problems that the lack of naming conventions bring and also raises issues that need to be considered when creating SQL, searches as straight comparison searches may not suffice. It is also worth noting that many of the gene names have changed several times in the time the database has been operational and each change can be a major

descriptive change. This data can also be used as a knowledge base of old to new historical gene names that has the potential for use in text mining.

| ID | Date Changed | Old Name | Old Gene Name | Old Accession Number | New Name | New Gene Name | New Accession Number |
|---|---|---|---|---|---|---|---|
| 6902 | 16-Feb-07 | adenine **phosphoribosyltransferase, putative** | not known | At4g22570 | adenine **phosphoribosyltransferase** | not known | At4g22570 |
| 6900 | 16-Feb-07 | ACT **domain-containing** protein | ACR1 | At5g65890 | ACT **domain containing** protein | ACR1 | At5g65890 |
| 6898 | 16-Feb-07 | **aalic** enzyme/ oxidoreductase | not known | At5g11670 | **malic** enzyme/ oxidoreductase | not known | At5g11670 |
| 6897 | 16-Feb-07 | 4-coumarate**-CoA** ligase | not known | AU093458, AU093459 | 4-coumarate**:CoA** ligase | not known | AU093458, AU093459 |
| 6895 | 16-Feb-07 | 3-deoxy-arabino-heptulosonate-7-phosphate (DAHP) s | **DHS1** | M74819 | 3-deoxy-arabino-heptulosonate-7-phosphate (DAHP) s | **AtDHS1** | M74819 |
| 6887 | 15-Feb-07 | lipid-binding serum glycoprotein | not known | At1g04970 | lipid binding serum glycoprotein | not known | At1g04970 |
| 6886 | 15-Feb-07 | **lipid-transfer** protein | **OsLTP5** | AU063656, AU172383 | **lipid transfer** protein | **OsLTP5** | AU063656, AU172383 |
| 6868 | 06-Feb-07 | fatty acid hydroxylase **(FAH1)** | **not known** | At2g34770 | fatty acid hydroxylase | **FAH1** | At2g34770 |
| 6862 | 06-Feb-07 | chalcone synthase **(naringenin-chalcone synthase)** | **not known** | At5g13930 | chalcone synthase | **AtCHS** | At5g13930 |
| 6860 | 01-Feb-07 | **unknown** | not available | AV795353 | **glutamate-ammonia ligase** | not available | AV795353 |
| 6857 | 31-Jan-07 | xylosidase | **not known** | BAB09906 | xylosidase | **AtBXL1** | BAB09906 |

Table 12: Sample of contents from the Gene Dictionary

### 5.5.2.3. Validation Screen

There are four other useful functions that enable the data to be checked and edited as appropriate. These functions are additional facilities that are available in the Edit Chemical form described in Section 5.4.3 and are:

1. View all – Sort by Accession number
2. View all – Sort by AGI number
3. View all unmatching Name/Accession numbers

4. View all unmatching Name/AGI numbers

The first two functions were requested by the biologist to simply sort the chemicals by accession number or AGI number. This allows the experienced user to scan through the results for interesting or unusual data.

The third and fourth functions are the most useful in terms of maintaining the data. They provide the facility to identify records that have the same Accession Number but a different or unknown Name or to identify records that have the same Accession Number but a different or unidentified AGI number. Table 13 shows a sample of the results from the unmatching Name/Accession number search. The three pairs of matching Accession numbers are highlighted in black outline boxes. One would expect that as the Accession numbers are the same, the name would be the same so this prompts further investigation to establish what the most up-to date nomenclature is for this gene using the unigene form and then update the database. The same procedure is used for the unmatching AGI/Accession numbers. The reason that this is useful is that it links reactions in the database and enables us to find more results that relate to a specific gene thereby giving more information about how each individual gene reacts under multiple conditions.

| Name: | Gene Name: | Accession Number: | Gene Product: | Enzyme: | AGI Number: |
|---|---|---|---|---|---|
| cytochrome P450 | DWF4/CYP90B1 | AF044216 | ☑ | not available | At3g50660 |
| steroid 22-hydroxylase | CYP90B1 / DWF4 | AF044216 | ☑ | not available | At3g50660 |
| late embryogenesis abundant LEA SAG21 homo | not available | AF069298 | ☑ | not available | At4g02380 |
| pectinesterase putative | BRU18 | AF069298 | ☑ | not available | not available |
| chitinase (glycosyl hydrolase family 19) | not available | AF104919 | ☑ | not available | At4g01700 |
| DNA binding protein | not available | AF104919 | ☑ | not available | not available |

Table 13: Sample of unmatching Name/ Accession numbers.

For completeness, another checking facility available to the user is a subform of the Edit Chemical that displays the Gene Ontology (GO) terms for a selected AGI number.

## 5.6. Testing of database

Once the data structure and interfaces have been implemented, the old data needs to be converted and loaded into the new database and the database needs to be tested to ensure that it meets the requirements and design specifications.

### 5.6.1. Data Conversion & Loading

The procedure of converting the data was particularly difficult as the gene expression data had previously been stored in a single table and there was not data for all the required attributes nor was there any structure. The data had to be backtracked to the original journals that it was captured from, appropriate attributes collected and then the data restructured and validated before being loaded into the new database. To test the conversion process, a sub-set of the added records were randomly selected and checked against the database to ensure that the records were stored correctly.

### 5.6.2. Testing Procedure

Connelly and Begg (2004) describe the purpose of testing in the context of a database application as the process of running the database system with the intent of finding errors. Testing should analyse the functional and structural aspects of the database as well as the usability of the interface and the tests for the database were designed to consider:

- Interface Testing - ensure the user interface behaves as expected, is useable and validation rules work correctly
- Verification Testing -ensure all functional specifications are accurately met and check there are no bugs in the code from either the Structured Query Language (SQL) which is used to search the data or the VBA which is used in designing the interface.
- Validation Testing - show that the whole application meets the original formal requirement specifications

### 5.6.3. Testing Results

A sample of some of the tests that were carried out on the database are shown below in Table 14.

| Test No. | Test Description | Test Data | Expected Output | Actual Output |
|---|---|---|---|---|
| 1 | Search for all treatments | SQL search "Select * from Treatments" | All treatments returned | Fail - All bar one treatment returned. |
| 2 | Add new records to | 1st Record | Both records | Fail - Unable to |

| | | | | |
|---|---|---|---|---|
| | chemical table | Name: Test<br>GeneName: not available<br>AccessionNumber: not known<br>ECNum: unknown<br>AGINum: not relevant<br><br><u>2nd Record</u><br>Name: Test<br>GeneName: not available<br>AccessionNumber: not known<br>ECNum: not known<br>AGINum: At1g11111 | added | add 2nd record. |
| 3 | Enter new reaction – check interface operates correctly | Sample data for a full gene expression result was selected from a journal | Reaction Added and all records in other entities added | Success - All records successfully added. |
| 4 | Leave blank data in required fields for adding reaction – check domain integrity | Sample data for a full gene expression result was selected from a journal | Error message to prompt for more data entry | Success – Error message prompt displayed. |
| 5 | Leave blank data in non-required fields for adding reaction | Sample data for a full gene expression result was selected from a journal | Expect that the rules for handling missing data should be applied and 'unknown' or 'not available' should be inserted to missing attributes where appropriate and records added | Success – missing data correctly filled and record added. |
| 6 | Edit a Chemical – Check referential | Change the Name from Test to Test1 | Expect that the Chemical record should be updated and all | Success – all data updated. |

| | | | | |
|---|---|---|---|---|
| | integrity | | records in reaction containing the foreign key should be updated | |

Table 14: A small selection of some of the tests that were carried out on the database

There were a few errors picked up in the testing process of the database ranging from data formatting to domain specific – two are described below:

1. Square brackets [ ] are accepted as an attribute value by Microsoft Access but are not picked up during the SQL search. One treatment sodium nitropruside [an NO donor] had these brackets. It is considered good practice to avoid characters such as ampersands, percentages, asterisks, brackets and quotation marks fields or field names but this is difficult to avoid with the make up of the gene names. The record and all related reaction records have been amended to sodium nitropruside (an NO donor) and a new validation rule to reject any user entries with square brackets has been added to the input screens.
2. The user is unable to create a new record in the chemical table: It was found during the testing process in the majority of cases it was possible to add a record to the Chemical table, but if a gene from Arabidopsis has the same name as a gene from another species and an unknown accession number then a record will already exist and the user will not be able to create a new record to include the AGI number (which is a none key field). This would be a rare occurrence as most records have an accession number which would be different for different species. However, a new business rule had to be created to avoid this situation. This rule stipulates that if the Accession Number is unknown, but the AGI number is known, the AGI number is input to the Accession Number field.

On completion of the testing, the test results showed that the database and the input screens were working correctly.

### 5.7. Summary

An independent published review of the newly created database (Samson, 2005) included in Appendix II found "that there are no plant genomes represented in Ensembl, and even Medline, surely an essential tool for all biologists, deals, deliberately, with 'very broadly medically related' journals only". The review finds that "the plant communities are, at last, setting up some unique resources. DRASTIC, a Database Resource for Signal Transduction

in Cells is a useful, and relatively new, plant-specific bioinformatics resource". With the database now successfully constructed and tested to the users and to requirements specifications standards, the next step in the process is to develop tools to enable the user to analyse the data.

# Chapter 6 DRASTIC-INSIGHTS: Data Toolset

## 6.1. Introduction

This chapter describes and evaluates the Drastic-Insight toolset as described in the paper published in the Nucleic Acid Research Journal (Button *et al.*, 2006). A downloadable guide for using DRASTIC–INSIGHTS has been developed to assist users and is available in Appendix VI. The tools were developed based on the requirements analysis and key objectives from Chapter 2

## 6.2. Gene RoadMap

The requirements process identified several methods of investigating data that may provide insights into signal transduction pathways. These have been implemented in the *RoadMap* tool which has four types of searches (Common Genes, Unique Genes, Gene RoadMap & Pathway RoadMap) providing summary representation on all or selected genes from the DRASTIC database and the ability to drill deeper:

Common Genes - This search enables the user to determine genes that are co-regulated by the treatments that they have selected.

Unique Genes - This utility identifies all genes that are only regulated by one treatment.

Gene RoadMap - To operate the Gene RoadMap (Figure 32), the user must select the genes, species and regulation (up, down or both) to include in the search. An initial check is run to establish the treatments to include in the search and populate the column and row headings, after which the tool processes and displays the data. The map itself (shown in Figure 32 Part A) is dynamic and will allow the user to view results ranging from a small number through to results covering all species within the database. Both row and column headings of the Gene RoadMap are the same and display each treatment in the current search. The cells that are highlighted in red indicate the results that correspond to only one treatment. The cells highlighted in yellow indicate potential new discoveries of interest which would otherwise remain opaque.

The Gene RoadMap can be used as a lookup table. For example if we want to know if there are any genes that are regulated by cold and abscisic acid (ABA), we can scan along the cold row until we come to the ABA column. The cell shows that there are fifty five genes that are regulated by both treatments. This tool provides the facility for the user to mine through different layers of data. For example, functionality to display all gene names for any result is found if the user double clicks on a cell as shown in Figure 32 Part B where the cold/ABA cell has been selected.

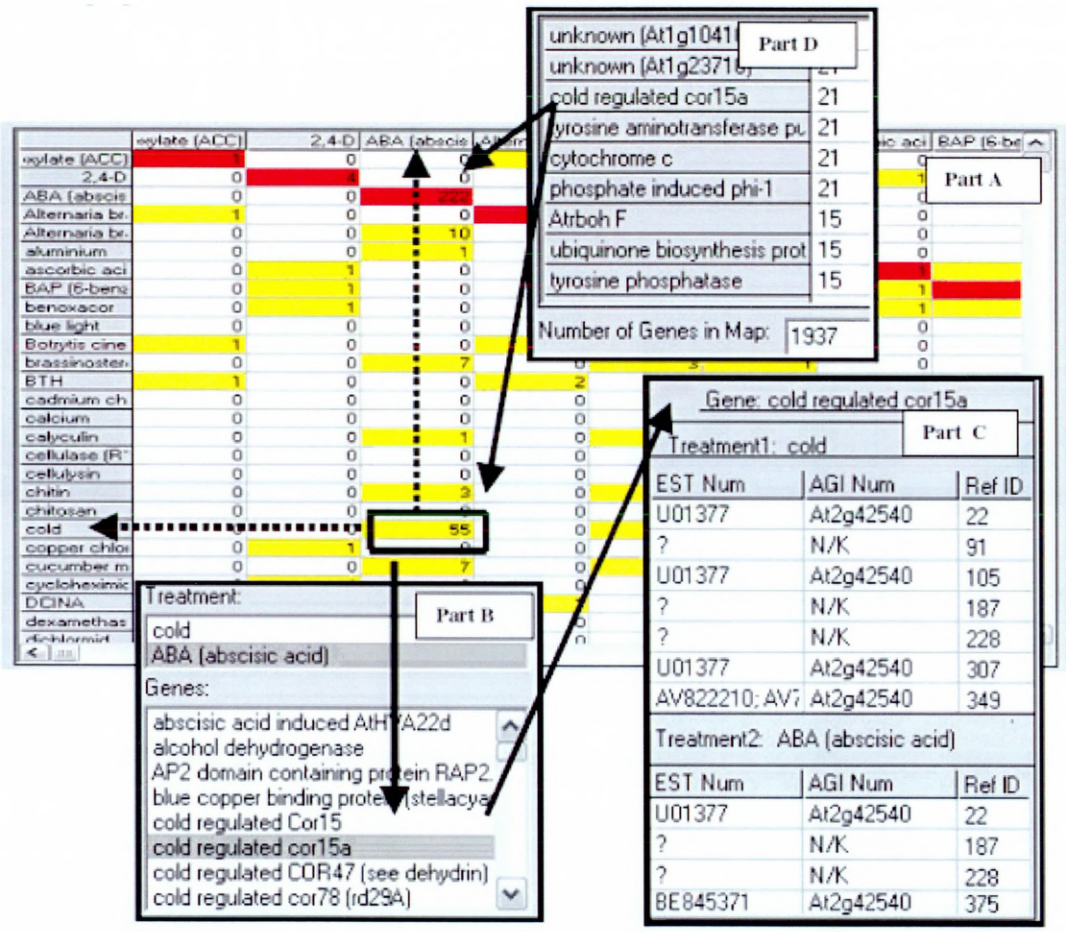



Figure 32: Gene RoadMap screen showing results from a search of all genes from *Arabidopsis thaliana* which are up-regulated. The different Parts (A-D) demonstrate the diverse ways in which the RoadMap tool can be used to examine the dataset.

The user can then further investigate individual genes by clicking on the gene name. This allows all the references that support the individual gene expression data for each treatment to be obtained, thus providing the user with a level of confidence for each result (Figure 32 Part

C). In addition, the total number of genes in the map along with the number of times each gene occurs is displayed. The user can locate every entry for a particular gene by clicking on the gene name. The tool will then highlights all cells in which the gene is featured allowing for further investigation (Figure 32 Part D). The Gene RoadMap tool demonstrates that it is possible to identify groups of treatments that appear to produce similar regulatory results. This has yielded both expected and unexpected grouping results.

### 6.3. Clustering Analysis of Treatments

From the results provided by the Gene RoadMap, the next stage of research focused on identifying groups of 'similar' treatments. We sought to find a method that could identify any groupings that existed between treatments as these could indicate shared signal transduction pathways. Narayanan *et al.* (2002) described the use of Hierarchical Clustering to establish the similarity of two biosequences across all attributes. This data mining technique has been applied to the database. The data required preparation in advance of applying the clustering technique as follows:

1. Selection of a species and compilation of an array of all treatments that have results for the species.

2. Compilation of an array of all genes that have results for the treatments and species.

3. Creation of an array for each treatment which holds the response result for each selected gene. The resulting gene expression data is categorical and has four possibilities classifications: up, down, same, not known (N/K). Each treatment will have an array of results of equal number as shown in Table 15.

| **Treatment** | **Gene 1** | **Gene 2** | **Gene 3** | **Gene *n*** |
|---|---|---|---|---|
| A | Up | Up | N/K | Up |
| B | Up | Up | Up | Up |
| C | N/K | N/K | N/K | Down |
| D | Up | N/K | Up | Up |

Table 15: Example of the results produced from the cluster search. Each treatment can now be compared to identify similar treatments.

The 'distance' between treatments is calculated using a matching coefficient for each pair of treatments. It is computed by dividing the total number of genes by the number of matching genes in each pair of treatments. This produces a 2-D matrix of treatments by treatments with the corresponding matching coefficient for each. In order to produce a visual result, the treatment matrix is processed using the statistics package plclust method (Ihaka and Gentleman, 1996). Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each stage distances between clusters are recomputed by the Lance–Williams dissimilarity update formula according to the particular clustering method being used. The complete linkage hierarchical clustering method produced the similarity tree for *A.thaliana* shown in Figure 33. Each branch represents a treatment. The magnified sample of the diagram shows the treatment results for one sub-cluster of the tree. The diagram illustrates treatment groupings that have been indicated in the literature. For example, jasmonates have been implicated in the wounding response and abscisic acid has been associated with cold tolerance.

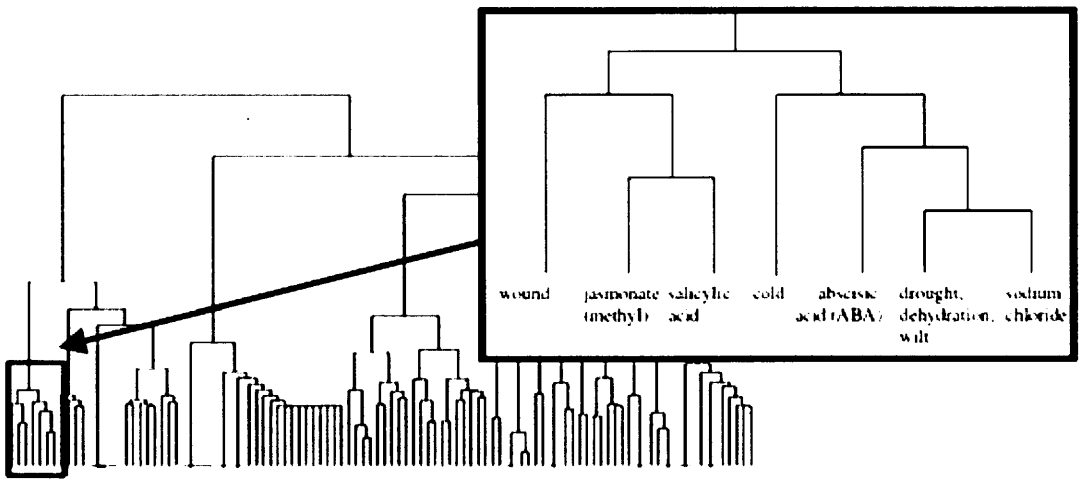



Figure 33: Dendrogram created using *A.thaliana* data and R Stats Package. As shown in the enhanced portion of the diagram, each branch in the cluster represents a treatment.

Thus the results demonstrate that this tool is capable of quickly producing interesting clustering output, but to be truly effective a more even distribution of data is required. The results from the clustering yielded the requirements for the next tool which provides a method of comparing the regulatory patterns of groups of genes.

## 6.4. Expression Patterns

Expression Patterns show all relationships between selected genes and the treatments (based on the data from the database). The treatments are placed around the circle edges of the diagram predominantly alphabetically. This order may in the future be informed by the results from the clustering tool. Genes that are found to be regulated by only one treatment are noted outside the circle. Genes that are co-regulated are placed inside the circle with lines drawn to each treatment to which they respond. The data used in the creation of the diagram is obtained through an interactive web tool. This tool enables the user to select a group of genes such as 'all kinases' or 'all transcription factors', a species and a regulation from the database. A table is produced for each group of genes selected displaying all the genes and treatments that regulate the specified group. The cells on the table that have a result have been hyperlinked to allow rapid retrieval of the records from the database. However, in this format, the data is very sparsely spread and difficult for the user to interpret. To provide a more insightful diagrammatic view, the resulting data has been used to manually create the diagrams shown in Figure 34.

Figure 34 is an example of an expression pattern diagram for stress-responsive kinases up-regulated in *A.thaliana*. The magnified Part A shows MAP kinase 3 gene which in this diagram is shown to be up-regulated by calyculin, chitin, cold, sodium chloride, drought/dehydration/wilt, hydrogen peroxide, jasmonate (methyl), mechanical stimulation, salicylic acid, UV and wounding. Part B shows gene MAP kinase kinase kinase up-regulated by cold, sodium chloride, jasmonate (methyl), mechanical stimulation, salicylic acid and wounding. It can be seen in Figure 34 that some kinases (e.g. MAP kinase 3 (AtMPK3; At3g09010) and MAP kinase kinase kinase (AtMEKK1; At4g08500)) are up-regulated by a number of treatments whilst others have only been reported to be up-regulated by a single treatment. As more information is put into the database, some of the kinases currently shown as up-regulated by a single treatment will probably be up-regulated by other treatments. Genes for proteins involved in the same signal transduction pathway are likely to be co-regulated and show the same response to a range of treatments.

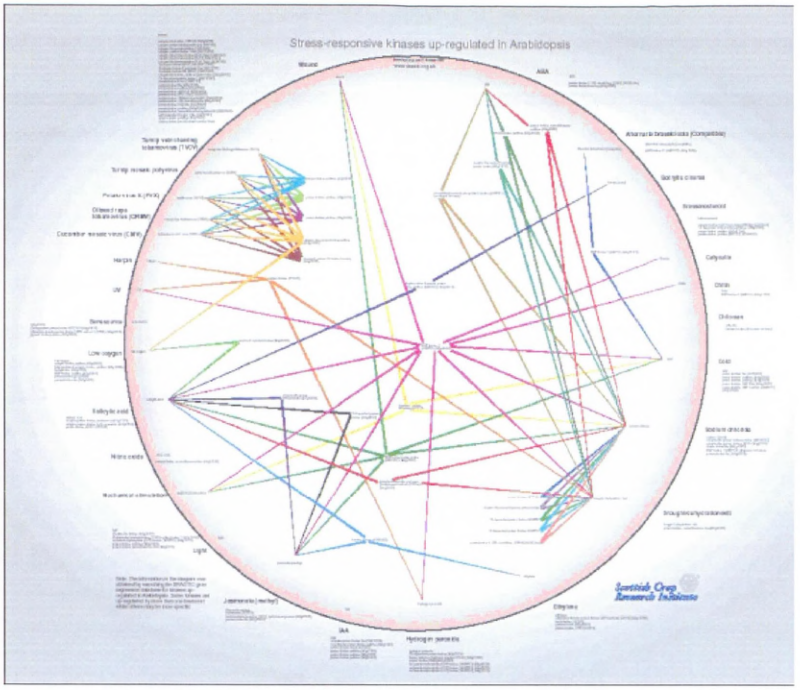



Figure 34: Expression Pattern Diagram for kinases

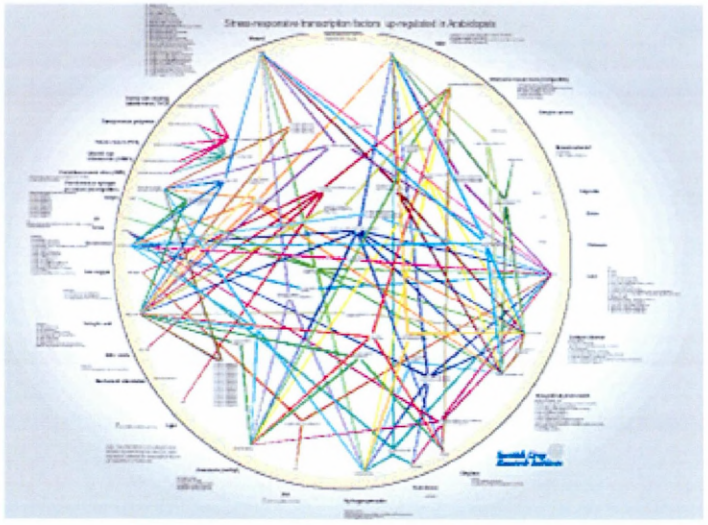


Figure 35: Expression Pattern Diagram for transcription factors

Thus, to find kinases, transcription factors, and calcium-binding proteins that are in the same signal transduction pathway one should compare expression patterns between these diagrams. Figure 35 shows the Expression Pattern Diagram for stress-responsive transcription factor genes up-regulated in *A.thaliana*. There are several instances where genes are showing similar patterns of regulation one of which is highlighted in Figure 35. Verification that these genes are really associated requires experimental confirmation, but the database, and these

diagrams, promote more targeted hypothesis formation. This type of analysis is useful for providing a framework for understanding signal transduction responses and to assist with identifying regulatory gene networks. This approach is also useful for finding genes which are associated with infection by plant pathogens and that are also affected by environmental stresses such as drought and cold.

Figure 35 shows an expression diagram for stress-responsive transcription factors up-regulated in *A.thaliana*. The magnified part of the diagram shows the gene for the Transcription Factor (TF), DREB2A (At5g05410) which is up-regulated by cold, sodium chloride, drought/dehydration/wilt, hydrogen peroxide, UV, harpin and wounding.

## 6.5. INSIGHTS data tools

Following on from the development of PC based tools, a web interface was developed to enable many scientists to use the features of the DRASTIC-INSIGHT tools. At a simple level the web interface (software available in attached CD) permits users to find published information on expression data for plant genes of interest. More importantly, INSIGHTS offers a number of tools to mine further information and create new knowledge and formulation of hypotheses. Some mining tools use AGI numbers where expression data correctly identify a specific member of a gene family. Through the INSIGHTS integrated toolkit users may investigate data in the following ways:

### 6.5.1. General Database Search

General database search provides a basic query function for the database. The user can select the following parameters: treatments, species, gene, regulation and date. The search returns the results in tabular format which can be sorted on all parameters and provides links to the primary references.

### 6.5.2. DRASTIC Statistics

DRASTIC statistics provides an up-to-date list of statistics for the database including the total number of records, species and treatments. It also provides a breakdown of both records per species and records per treatment, which can be ordered alphabetically or numerically. To gain a more in-depth view, a table of data providing statistics on the number of records by

species or treatment can be obtained. These can be further mined to view individual records with bibliographic references.

### 6.5.3. Accession Number Search

Accession number search provides a query function specifically for the accession numbers in the DRASTIC database. Selectable parameters include accession number, treatment, regulation type and date. The results are displayed in tabular format which can be sorted, providing links to references.

### 6.5.4. AGI Search

Arabidopsis genome initiative search provides a query function specifically for AGI numbers in the DRASTIC database. The user can select from AGI number, treatment, regulation and date.

### 6.5.5. Venn Diagram

Venn diagrams enables the creation of Venn diagrams using the *A.thaliana* data from the DRASTIC database. The user can select two or three treatments and the tool will process the selections and output the results as a Venn diagram. The Venn diagram tool displays the number of genes regulated by each individual treatment or by multiple treatments based on the DRASTIC data. Records where genes have been up-regulated, down-regulated or both (up or down) can be included. The diagrams can be mined further by clicking on a segment of the diagram to view the individual records and relevant bibliographies.


Figure 36: Web interface for the Pathway tool.

Figure 36 is an example of the web interface for the pathway tool. (A) shows the search page for a set of AGI numbers. The pathway result is shown in (B). Up-regulated genes are shown in red. Down-regulated genes are shown in blue. Green cells indicate that both up- and down-regulation record(s) are held in DRASTIC. The pathway can be further mined by choosing any coloured cell which will display all the records for the AGI/treatment combination as shown in (C). The references for each record can be selected as shown in (D).

### 6.5.6. TAIR AGI Search

TAIR AGI search enables the user to search records that include the AGI number and directly use them with the TAIR chromosome mapping and functional categorization tools, which are specifically designed to analyse AGI data. The user can select a subset of records from

DRASTIC using a search on a treatment, multiple treatment or gene group (such as kinases) and regulation type. The selected data are then formatted for use with the TAIR tools.

### 6.5.7. Pathway Tool

Pathway tool enables the user to extract and visualize knowledge from the database to hypothesize potential relationships between signalling elements. It includes a search facility to allow selection of a number of *A.thaliana* genes by AGI numbers. A 'pathway' is produced to display the regulation of selected genes in response to different treatments (Figure 36). Any groups of genes that are always co-regulated are identified, suggesting that they are likely to occur in the same signal transduction pathway. The pathway tool can be used to indicate the relatedness of induction patterns for selected genes. For instance, it can be shown that up-regulation of calreticulin 3 (At1g08450) in Arabidopsis has been shown to be associated with the up-regulation of a number of potential signalling genes (including kinases), which does not occur if calreticulin 1 (At1g56340) and calreticulin 2 (At1g09210) are down-regulated. The pathway tool can also be used in a hypothesis testing manner or as a quality control check tool for data in known signal transduction pathways.

### 6.5.8. RoadMap Tool

Roadmap tool creates lookup tables to find genes that are co-regulated by different treatments. The user can 'drill down' through the map to investigate individual genes and view all references that support each data point providing a level of confidence for each result. To operate the roadmap, the user selects an AGI number and a regulation (up-, down- or both) to include in the search. The tool establishes which treatments regulate expression of the selected gene and then displays in a map all the genes in DRASTIC that are regulated by these treatments (Figure 37). This tool demonstrates that it is possible to identify groups of treatments that appear to produce similar regulatory results in *A.thaliana*. Roadmap results can be used in conjunction with the Pathway tool.

### 6.5.9. Unique Genes Tool

Unique genes tool identifies all the *A.thaliana* genes that are regulated by a single treatment. Full details including references for each gene are linked to each record.

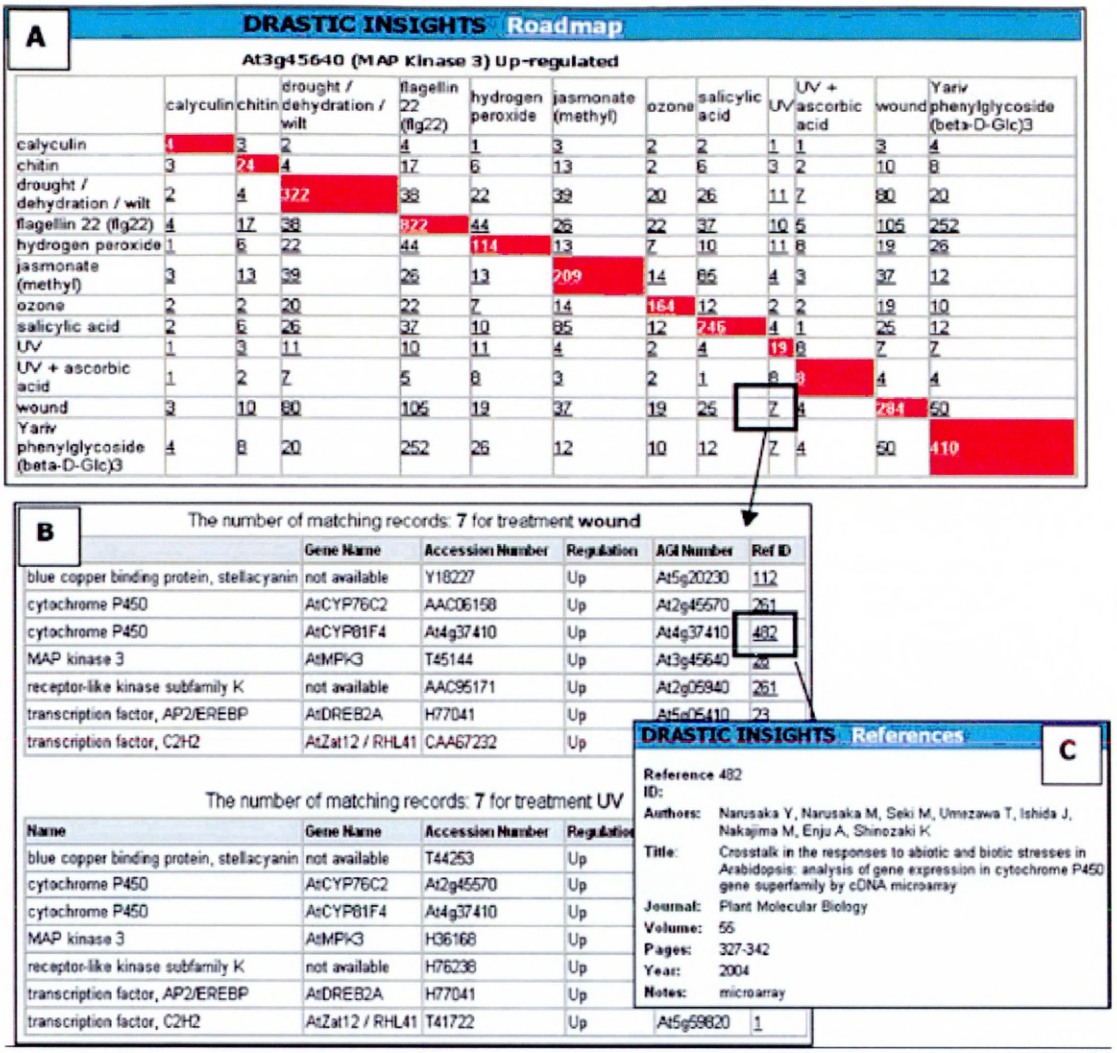



Figure 37: Web interface for the Roadmap tool.

In the example shown in Figure 37, treatments up-regulating At3g45640 (MAP Kinase 3) were selected for investigation. (A) The resulting roadmap. From the DRASTIC data, 12 treatments up-regulate Atg45640. Using these treatments as the 'lookup co-ordinates', the map displays the total number of unique AGIs up-regulated by these treatments. The red squares hold the total number of genes up-regulated by a single treatment, and the numbers in the unshaded squares show the number of genes co-regulated by treatments. This map can be further mined by clicking on any of the squares to display the supporting records (see (B) where the co-ordinates wound and UV have been selected). Each record has a link to the reference it was curated from as shown in (C).

### 6.6. Technical ToolSet Information

The Drastic-Insight toolset that is available through the website has been developed using ActiveX Data Objects (ADO) and Active Server Pages (ASP) technology to enable the user to dynamically interface between the toolset and the database as demonstrated in Figure 38.

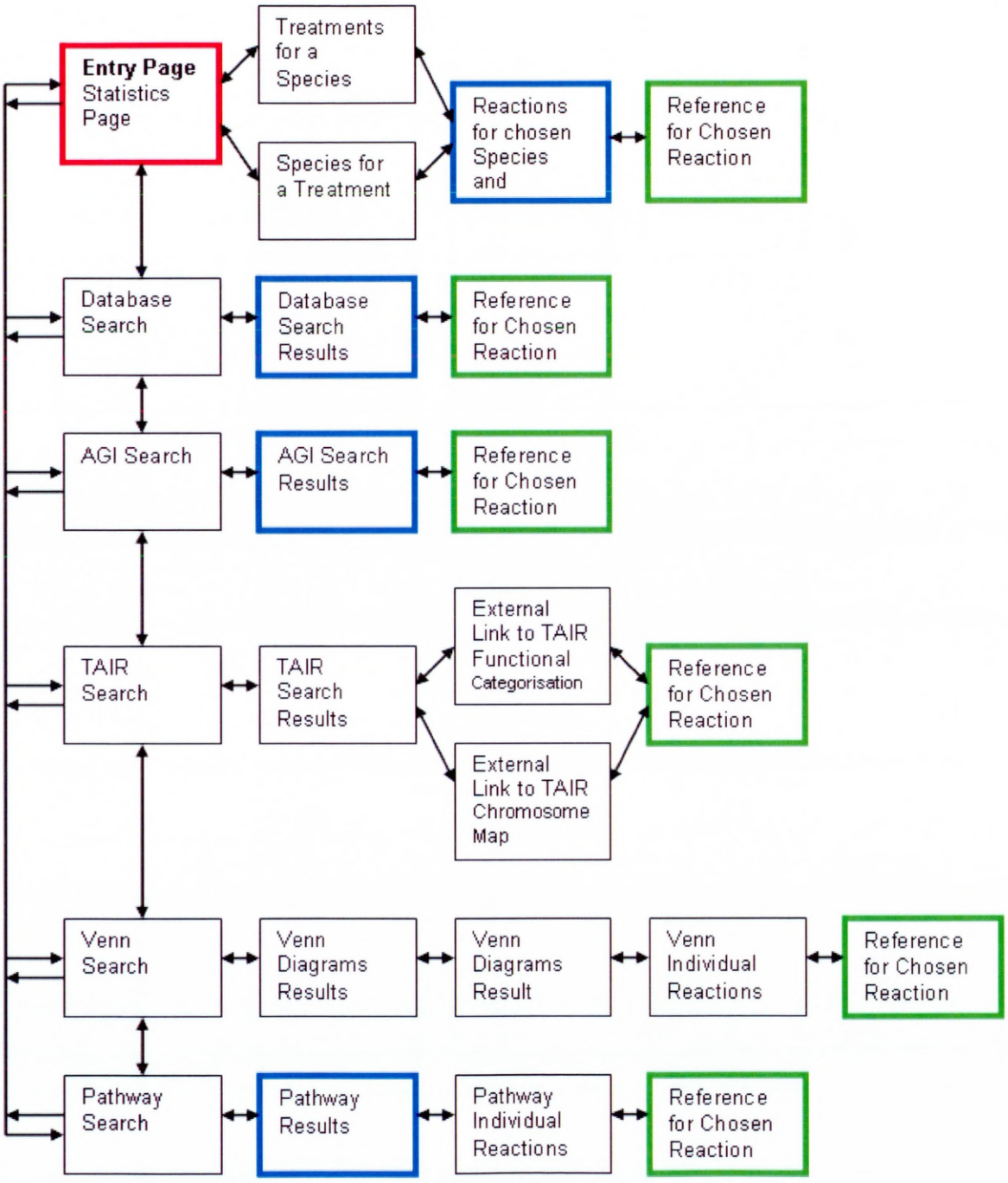



Figure 38: Drastic-Insight SiteMap. This figure shows the pages and links that are needed for the website. The boxes with the green border are one page template that is used many times and the blue bordered boxes are

another example of the use of dynamically generated content based on user request. The red box is the entry page for the Drastic-Insight toolset and provides links to the other tools. This illustration does not include the internal links on a page such as sorting functions or clearing results that require the page to be reloaded.

This technology enables the creation of webpages *"on the fly"* as the ASP technology allows programmers to use functions, variables and control structures to build a web page dynamically at the moment it is requested by the browser. This enables users to "ask" different questions of the database and be presented with individual results which will include the most up-to-date information from the database.

## 6.7. Summary

The INSIGHTS tools encourage comparison of gene expression patterns, intelligent mining of information, testing and formulation of novel hypotheses on the complex signal transduction and response pathways used by plants. Identifying common elements in pathways affected by different treatments permits the formation of hypotheses previously opaque to the user as demonstrated in Chapter 7

This type of analysis is useful in providing a framework for understanding signal transduction responses and to assist with identifying regulatory gene networks. It is also useful for finding genes associated with plant pathogen infection that are also affected by environmental stresses such as drought and cold in differing ways. Chapter 7 provides details of testing, evaluation and analysis of the toolset and results that it produces.

# Chapter 7 Testing and Analysis

## 7.1. Introduction

This chapter analyses the DRASTIC-INSIGHT toolset and summarises the study findings. The toolset is very difficult to benchmark due to the fact that it is an investigative tool and is likely to be subjectively assessed based on the experience of the individual user. The toolset has been comprehensively tested as detailed in section 7.2. In order to evaluate the tool effectiveness, plant pathologists were asked to utilise the database and report back their findings. The results from this test are evaluated against the overall objectives of this project.

## 7.2. Testing Procedure

The testing approach taken for the DRASTIC-INSIGHT toolset incorporated the standard software testing protocols of Unit Testing, Integration Testing, System Testing and Acceptance Testing (Kappel, 2006). As the majority of the toolset is a web based application, the testing was adapted to include content, hypertext structure, design aesthetics, usability and page loading.

### 7.2.1. Testing Results

A sample of some of the tests that were carried out on the database are shown below in Table 16.

| Test No. | Test Type | Test Description | Expected Output | Actual Output |
|---|---|---|---|---|
| 1 | Usability | Test the internal and external navigational links between the webpages. | All links work correctly. | Pass |
| 2 | Reliability | Test the time that pages take to load. | Pages load in under 15 seconds. | Fail – Venn Diagram search is extremely slow. Progress bar added to page. |
| 3 | Functionality | Test the pathway software with pre-chosen data. | Expect that the results will match the predicted result | Pass – However, user noted that it would be useful for the treatment number to be a pre-allocated number rather than a |

| | | | | |
|---|---|---|---|---|
| | | | | different number for each search as it is easier to compare diagrams. |
| 4 | Learnability | Test if the site navigation and design is consistent and easy to use | Pass | Pass |
| 5 | Database Connectivity | Test if the database is connecting and accessing data correctly by testing the return of the statistics page and comparing the results against the actual database results | Expect that statistics from the website and the database will match. | Fail – One SQL statement in the Individual Reaction Page was incorrect and has been amended. |

Table 16: A small selection of some of the tests that were carried out on the Drastic-Insight tools

### 7.2.2. System Objective Testing

With the system functioning correctly, the next test was to evaluate if the system met its original objectives in section 2.4. The requirements of the main components of the system were:

- Structured database with interface to enable easy data input for all types of experimental results described in Chapter 5.
- Quality/validation data checks and facility to update the data described in Chapter 4 and
- Set of web based tools to enable the biologist to intelligently query the data to formulate hypothesis described in Chapter 6.

The original objectives were then examined as detailed below in Table 17.

**Objective 1.** Co-regulation of genes - If two or more treatments regulate a gene then the genes may share a signalling pathway

**The user can use the Venn Diagram search to identify co-regulation of genes by up to three treatments or they can use the roadmap to search for co-regulation across the whole database.**

**Objective 2.** Gene regulation patterns - If a gene shares a similar regulation pattern to another gene then they may be adjacent to each other in the pathway.

**The clustering program enables the biologists to identify genes with a similar regulation pattern and**

**groups the results by treatment. The user can alternatively use the pathway tool to view regulation patterns of selected genes and visually compare the patterns.**

**Objective 3.** Regulation types - Within a pathway, a treatment should either up-regulate or down-regulate all genes. If this does not happen, this could indicate a quality control issue or crosstalk in a pathway.

**The user can select all the genes within the pathway and use the pathway tool to display the gene expression values for all the treatment/gene combinations in the database. The pathway tool uses colour to highlight the regulation making it easy to identify any anomalies in the pathway result.**

**Objective 4.** Determine the number of treatments that regulate each gene - Genes that are only regulated by one treatment may be early in the pathway. Genes regulated by many treatments may be later in the pathway.

**The user can go to the statistics page and drill down into the database results to find the number of genes per treatment among many other useful figures. There are sort functions available to enhance the usability of this tool.**

**Objective 5.** Grouping of treatments according to similarity of the expressed genes – This may indicate the total number of pathways.

**The clustering tool enables the biologist to group treatments according to similarity and produce a dendogram of results.**

Table 17: Each of the original objectives are listed along with the tools that meet them.

In addition to the objective testing, the biologists at SCRI used the web tool for a two week period as part of the acceptance testing and with the exception of some small modifications to the navigation menu to enhance usability, the system functioned correctly and met the requirements identified in Chapter 2.

## 7.3. Discussion

From the outset of the research the goal for this project was to develop a novel system that would enable the plant pathologists to search and collate gene expression data and section 4 demonstrates the benefits of this system. Chapter 1 set out the key aims of this research and how these are met is described in the preceding Chapters, however, some of the most relevant and interesting results or findings from these aims are discussed below.

1. Automate and enhance the process of information discovery for the biologist

From the outset it as been clear that one of the main issues that plant biologists are facing is data overload. A key aspect of this research has been enabling the biologists to quickly

identify useful data, provide a system to collate the data and lastly make information discoveries from the data. One of the ways that this study has approached the information discovery task is to interview the biologists and model the way in which they currently work with a view to automating tasks and modeling cognitive responses.

The system provides the biologist with an easy data entry system along with tools that will automate the process for certain datasets. When searching for key papers, this system quickly identifies relevant research papers and the results from the analysis in 7.4 demonstrate that the information discovery is successful. There are also tools that enable connectivity with public databases such as INCUBI to automate data discovery and identify newly annotated gene function or AGI number updates thus maintaining the quality and currency of the data.

2. Identify suitable data from different sources and different experiments

At the outset of the project it was anticipated that there would be much more suitable data available than there infact is. Scientists working in human and animal equivalents of this area are much more advanced and have not only got gene expression results but also protein – protein interactions. While the database is set up to collate these type of experiments when they become available it was surprising how few suitable data sources there were for data on plant defence signalling. This research also found that there are many different experiment formats and types and has explored how to identify suitable experiments and modeled ways to store these diverse types. There were no databases or tools that were available for plant signalling in defence and very few suitable sources of data in database repositories. In order to maintain quality and to provide additional background on gene interaction, peer reviewed results were sought for this project and the only suitable plan for this research was to create and curate a novel database source for this niche discipline.

3. Examine the structure of data particularly from journals and make this more accessible

For this project, a requirement was that the data should be from peer reviewed sources, thus journals are a natural focus for data acquisition. Once the data model had been developed the next challenge was how to collect the data that is published in journals. It is relatively difficult to search for publications containing results relating to specific genes. In a comparison study described in Chapter 4, DRASTIC returned far more relevant papers than Pubmed or text mining tools. This seems to stem from the way in which the journal data is stored. In addition to key words, it would be useful to have the facility to "tag" relevant points in an article in a way similar to the function that on-line forums use to link similar

threads. For example, authors could tag the unique gene identifier and gene function to enable these to be searched for. It would also be interesting to see a focus placed on updating journal articles - for example where a gene is stated as "unknown", and this status is subsequently updated, what impact does this have on the article? At present, journal articles remain untouched, however, now that the majority of these are available on-line, it is increasingly possible to enhance the search facilities and quality of the results contained within them. There are constantly new discoveries being made in the biology and from this research, it would seem that the data and the previous conclusions detailed in the papers are vital for biologists to hypothesis further. With the exception of the Plant Physiology journal, none of the other journals surveyed made any specific requirements for the reporting of any other type of experiment or how the results were to be presented. This means that key results can be contained within an unsearchable image or a poorly formatted file. Until a more unified approach to the reporting of experiments in journals is taken, manual or automated curation of results from certain experiments will remain difficult.

4. Provide a method to enable results from different types of experiments to be compared against each other

As discussed in section 4.3.5 there is debate in the literature as to whether it is appropriate to compare results from different types of experiments. Some database platforms such as Genevestigator have taken the decision to not enable users to compare results from different types of experiment (for example AG chip versus ATH1 chip). For this research, I opted to allow the users to carry out comparison searches between results from many experiments and developed a schema to enable this search. This was because the tool was designed to enable scientists to investigate and create hypothesis and any findings would be tested using wet science. It also uniquely enabled the biologist to have a general overview of all of the results from different treatments and species rather focusing in on a smaller more niche area as tends to be the case in this discipline. This provides scientists with a new way to view and manipulate the data. In order to combat the potential issues of inaccurate data, multiple reports of gene expression results for a gene and treatment are stored where possible and any differences are highlighted to the user at the results stage.

At the start of this project, there were some differing views about DRASTIC allowing different types of experimental results to be compared as it inevitably meant reducing the data to a binary set. This is because the only way to enable comparison from experiments across

the board was to simply consider the results as up and down rather than use any numeric or factor values. This concern was raised as compressing the data like this potentially loses some data value. However, due to the simpler view of up and down, it is much easier to identify genes that respond in a similar manner than if a large set of numerical values are presented to the user. As already mentioned, the journal results can be in many different formats and the only common denominator for these results at present is to convert them to up or down thus it has also enabled the key principle of using peer reviewed quality data to be maintained. This approach is increasingly being validated with some of the main data repositories including EMBL recently enabling users to search for results that are converted to these binary values.

5. Design a data model that takes account of the various data and database standards that exist in the plant biology community

The data model for this research needed to have the ability to enable storage of diverse data formats while meeting the protocols and data standards and nomenclature of the plant community. There are two well documented types of standard which are applied to plant gene expression results (MIAME) and protocols for software developers (MAGE). MAGE is only useful if dealing with the full MIAME data which most tool providers will not be. There are no other standards produced for small niche database sites which make data and tool sharing difficult. In addition, these standards do not address how to integrate the data they describe with different data and nomenclature. Microarray technology is relatively young, but it would be wrong to assume that there will not be a new generation of technology. Standards such as these implemented in this data model will be required to future proof the data and enable new formats and standards to be integrated with old.

If we consider only Arabidopsis, there are publicly funded sites such as TAIR and TIGR that release updates on the annotation and nomenclature of this species. However, private companies such as Affymetrix have their own annotation which does differ slightly from the TAIR version and there is no collaboration on this. Further more, of the public repository databases surveyed here, there was no updating of the results when new annotations were released so certain gene expression results are actually now referring to different genes. While there is so much effort in the statistical correctness of microarray probe calls and absolute values, once these are released into the database repositories there is very little evidence of these results being maintained. Due to the nature of publication only allowing

novel findings to be published, there may be no up-to-date results for some gene expression values uploaded and these therefore may always remain out of date and incorrect. DRASTIC overcomes these issues by data quality checks to key database repositories and updating annotation as they are released. GO annotation references are maintained where appropriate for relevant genes in the database. Only data that has met the MIAME standard is included. In addition, all modifications are stored in a "gene dictionary" so any discrepancies can be checked and the data and data model is available to any developers or biologists.

6. Improve the speed, efficiency and ability of the biologist to search for information from the gene expression results collected

Through the development of the data model that holds the diverse data formats, automation of the simpler search queries, for example search for one gene, have vastly improved the speed and efficiency of these searches compared to the previous manual curation and investigation techniques.

Developing more sophisticated search functions, such as the pathway finder, for the project was a demanding process due to the structure of the data stored. While there is data overload in terms of the hundreds of thousands of individual gene expression results that are available, when each experiment is stripped down, it was discovered that approximately only 5% of the results from a microarray experiment that was identified as suitable for inclusion are of value to the project.

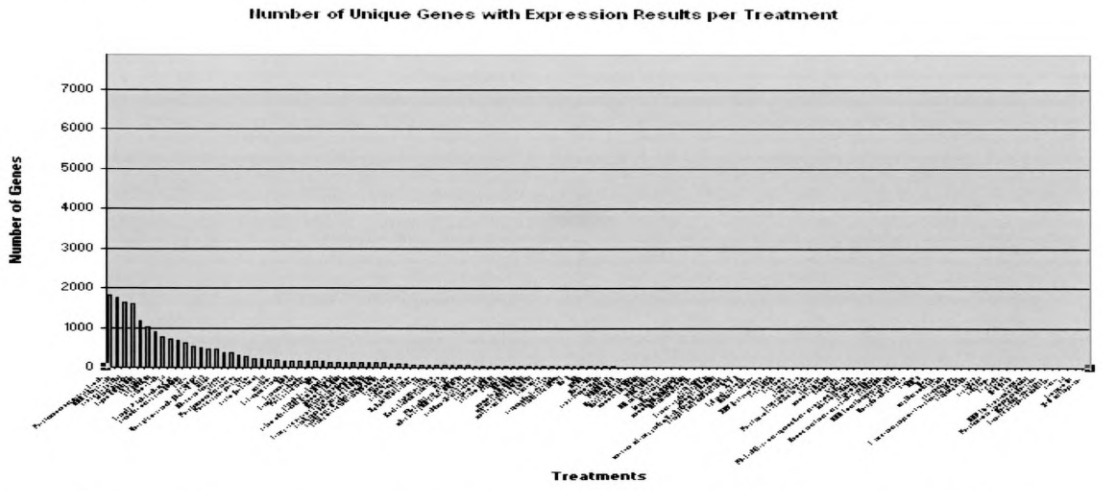



Figure 39: Shows the distribution of the number of genes with gene expression results per treatment.

This along with the inclusion of single result experiments causes the data matrix to be sparse and unevenly distributed as shown in Figure 39 which therefore presents a great challenge when considering how to search this data.

At the outset of the project, the expectation was that it would be possible to use data mining tools such as Bayesian, Decision Trees and Clustering. The data is binary so only hierarchical clustering was suitable. Experiments were undertaken in collaboration with researchers at Aberdeen University (Pers Comm. Wiranga, 2005) to use their algorithms which were developed specifically for use with a sparse dataset. The results from these experiments showed a limited improvement but disappointingly did not enable anymore significant results to be obtained from that of the clustering tool discussed in 6.3.

Bayesian methods were also explored as it was hoped that these may enable linkage of related genes to be discovered, but again the data sparsity meant that the results were of little use. Other conventional data mining methods were considered, but the data sparsity and lack of attributes did not make the dataset suitable for these techniques. Consideration was given to utilising the absolute value of the microarray data of each result, but the biologists focus for this research was primarily in whether the gene was up or down regulated and not by how much.

The DRASTIC INSIGHTS system overcomes these challenges by providing advanced search functionality for the biologists even with the sparse data set. During the research process, the focus has been to creating a knowledge discovery tool which would enable hypothesis generation based on the cognitive way in which the biologist processes the information. The process was modelled in Chapter 2 and the search functionality has been built based on this. Appendix V details some of the successful results that have been discovered using the search protocols developed for this system and shows that this provides the biologist with the ability to quickly search through many data points and discover new knowledge that was previously opaque.

### 7.4. Analysis of Results

The tools have been shown to meet the original system objectives, but to evaluate this study, researchers at SCRI were asked to utilise the database and toolset and report back any interesting results. The results were then analysed in order to establish whether the database was useful and if it assisted with the formation of hypothesis for signal transduction.

Appendix V shows some of the sample results that the biologists discovered during testing of the system. While these data discoveries would need to be experimentally tested, they clearly demonstrate the synthesis of knowledge that was previously opaque or hidden in unsearchable data journals or disparate microarray databases. The results from the use of the toolset show that it is possible for a biologist to hypothesis, test the hypothesis using the toolset and based on the results re-hypothesis or devise a wet science experiment to formally test the results.

The number of matching records: 37

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| At2g40000 (nematode resistance protein putative / Hs1pro-1 like) | | | Up | | | Up | Up | Up | | Down | | | Up | Up |
| At3g50930 (ATPase family AAA-type) | | | | | | Up | Up | | Up | | | Up | | Up |
| At4g12720 (growth factor-like) | | | | | | Up | Up | | | | | | | Up |
| At4g20780 (calcium-binding protein putative) | | Up | | | | Up | Up | | | | | | | Up |
| At4g39640 (glutamyltransferase, gamma, putative) | | | | | | Up | Up | Up | | | | | | Up |
| At5g25930 (leucine-rich repeat family protein / protein kinase family RLK5 like) | | | | | | Up | Up | | | | | | | Up |
| At5g57560 (xyloglucan endotransglycosylase related) | Up | Up | Up | Down | Up | Up | Up | | Up | | Up | | Up | Up |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | ABA (abscisic acid) |
| 2 | brassinosteroid |
| 3 | cold |
| 4 | cucumber mosaic virus (CMV) |
| 5 | drought / dehydration / wilt |
| 6 | flagellin 22 (flg22) |
| 7 | gamma irradiation |
| 8 | hydrogen peroxide |
| 9 | IAA |
| 10 | low oxygen |
| 11 | malondialdehyde (MDA) |
| 12 | methanol |
| 13 | wound |
| 14 | Yariv phenylglycoside (beta-D-Glc)3 |

Figure 40: Evaluation Experiment 1.

The result in Figure 40 demonstrates how several of the DRASTIC-INSIGHT tools can be used to find new information from the data set. The researchers comment from this test was *"Another set of genes possibly in the same pathway. I have just put some information into the database on genes regulated by gamma irradiation and noticed that some were also regulated by Yariv phenylglycosides. I therefore had a closer look at everything that was regulated by both of those treatments and noticed that the following subset of genes are also all up-regulated by flagellin. These must therefore be likely candidates for being involved in the same pathway, (possibly the last gene i.e. At5g57560 may be regulated by too many different treatments and may therefore not be that close in terms of signalling distance to the other genes)".*

Here the researcher has used the Venn Diagram tool to find genes that are regulated by both Yariv phenylglycosides and gamma irradiation based on curiosity from a research article he was interested in. This query has yielded a subset of genes which have been put into the

pathway tool. The results find that the subset of genes are also all up-regulated by flagellin. This may indicate that they share the same pathway and enable the biologist to further investigate this set of genes. Gene At5g57560 is shown to be regulated by 11 treatments and this leads the biologist to hypothesis that it may be further down the signalling pathway and therefore not as specific to the grouping of genes. This is new information that has been discovered because of the design of the toolset which enables the biologists to "ask" questions of the data in various ways.

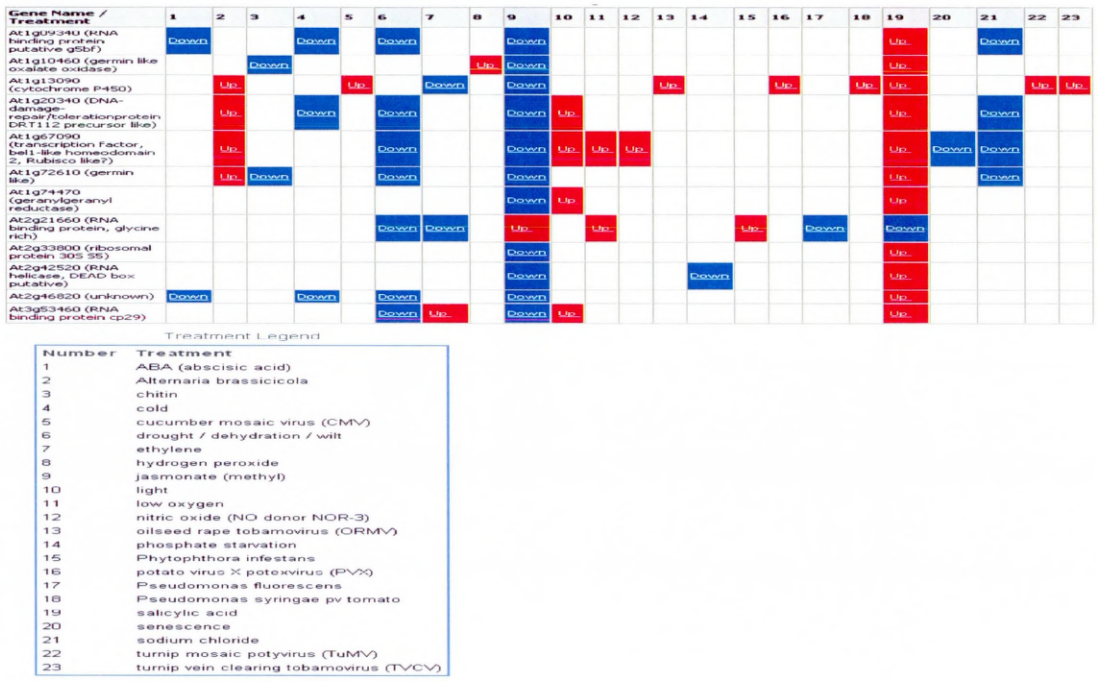



Figure 41: Evaluation Experiment 2.

The result in Figure 41 is an exciting result for the researcher whose comment on this result was *"This is one to rave about"* as it may indicate that there is cross talk within the pathways. In column 9, all genes are down regulated except gene At2g21660 and in column 19, all genes are up regulated except gene At2g21660 again. The reason that this is of interest is it may mean that this gene can be used as a switch to turn a pathway on or off thereby inhibiting or enabling a plant response. This would require further investigation using wet science. Neither of these results could have been discovered using querying alone or without the additional knowledge / curiosity of the researcher who was operating the system.

The number of matching records: 15

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| At1g20850 (protease, cysteine, papain-like (C1A-3 family)) | | | | Up | Down | | |
| At3g12610 (DNA damage-repair/toleration protein) | Up | Down | | Up | Down | | |
| At3g53460 (RNA binding protein cp29) | | | Down | Up | Down | Up | Up |
| At4g04460 (protease, aspartic, pepsin-like (A1-4 family)) | | | | Down | Up | | |
| At4g13660 (pinoresinol-lariciresinol reductase, putative) | | | | Down | Up | | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | Alternaria brassicicola |
| 2 | chitin |
| 3 | drought / dehydration / wilt |
| 4 | ethylene |
| 5 | jasmonate (methyl) |
| 6 | light |
| 7 | salicylic acid |

Figure 42: Evaluation Experiment 3. Researcher Comment: "Another small set to consider. Not sure if this is a good one or not. Maybe this is a set where one could make some predictions to test in the lab.

This experiment shown in Figure 42 shows two groupings of genes that respond in the opposite manner to each other when induced by ethylene or jasmonate(methyl). This is indicative of cross-talk in the pathway. It is believed that crosstalk between jasmonate and ethylene pathways enables plants to optimize their defence strategies more efficiently and economically (Zhao, 2004). The biologist may be interested to find how these genes respond to other treatments that the two plant hormones are known to react to in order to further investigate this gene grouping.

The number of matching records: 160

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| At1g01140 (SNF1-related kinase) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At1g01720 (no apical meristem / transcription activator, NAC domain, ATAF1) | Up | | Down | Up | | | | | | | | | | | | | Up | Up | |
| At1g07040 (unknown) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At1g13990 (unknown) | Up | | | Up | | | | | | | | | Up | | | | Up | | |
| At1g32640 (protein kinase; RD22BP1; transcription factor, bHLH putative) | Up | Up | | Up | Down | | Up | Up | | | | | | | | | Up | | |
| At1g45249 (ABA-responsive cis-acting element) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At1g52980 (GTP-binding protein putative) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At1g61890 (MATE efflux family) | Up | | Down | Up | | | | | | | | | | | | | Up | | Up |
| At1g66760 (MATE efflux family protein, putative) | Up | | Down | Up | | | | | | | | | | | | | Up | | |
| At1g68440 (unknown) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At1g72800 (unknown, nucleolin like? (NuM1-related)) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At1g75500 (nodulin MtN21 family) | Down | | | Down | Down | | | | Down | | | | | | | | Down | | |
| At2g02710 (receptor-like serine/threonine protein kinase) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At2g05540 (glycine-rich protein putative) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At2g06050 (12-oxophytodienoate reductase) | Up | | Down | Up | | Up | | Up | | Up | | | | | | | Up | | |
| At2g30550 (lipase putative) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At2g42890 (Mei2) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At2g46680 (transcription factor, homeobox leucine zipper) | Up | | | Up | | | | | | | | | | | | Up | Up | Up | Down |
| At2g47190 (transcription factor, Myb) | Up | | | Up | | Up | | | | | | | | | | | Up | | |
| At3g02480 (ABA-responsive protein-related) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g17000 (ubiquitin conjugating enzyme like) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g19290 (transcription factor, bZIP (AREB2, ABF4)) | Up | | | Up | | | | | | | | | | | | | Up | | Down |
| At3g20300 (unknown) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g22830 (heat shock transcription factor-like protein) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g29575 (unknown) | Up | | Down | Up | | | | | | | | | | | | | Up | | |
| At3g53710 (zinc finger-containing protein ARF GAP-like ZIGA2) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At3g55530 (zinc finger (C3HC4-type RING finger) family) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g59930 (unknown) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g62030 (peptidyl-prolyl cis-trans isomerase, chloroplast / cyclophilin / rotamase) | Down | | | Down | | | | | | | | | | Down | | | Down | | |
| At4g01020 (pentatricopeptide (PPR) repeat-containing protein) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At4g20830 (reticuline oxidase homolog) | Up | | | Up | | | | | Up | | Up | | Up | | | | Up | | |
| At4g23050 (protein kinase, serine/threonine putative) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At4g37510 (ribonuclease III family) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At5g01520 (zinc finger (C3HC4-type RING-HCa)) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At5g09440 (phosphate-responsive protein, putative) | Up | Up | | Up | | | | | | | | | | | Up | | Up | | |
| At5g11420 (unknown) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At5g23060 (unknown) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At5g25460 (unknown) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At5g39610 (NAC, NAM (no apical meristem) like) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At5g42010 (WD-40 repeat protein family) | Down | | | Down | | | | | | | | | | | | | Both | | |
| At5g48180 (jasmonate inducible (myrosinase binding)) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At5g52300 (responsive to dessication RD29B) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At5g61820 (MtN19 like) | Up | | | Up | | | Up | | | | | Up | | | | | Up | | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | ABA (abscisic acid) |
| 2 | Alternaria brassicicola |
| 3 | cucumber mosaic virus (CMV) |
| 4 | drought / dehydration / wilt |
| 5 | ethylene |
| 6 | flagellin 22 (flg22) |
| 7 | hydrogen peroxide |
| 8 | jasmonate (methyl) |
| 9 | low oxygen |
| 10 | malondialdehyde (MDA) |
| 11 | methanol |
| 12 | nitric oxide (NO donor NOR-3) |
| 13 | ozone |
| 14 | phosphate starvation |
| 15 | salicylic acid |
| 16 | senescence |
| 17 | sodium chloride |
| 18 | wound |
| 19 | Yariv phenylglycoside (beta-D-Glc)3 |

Figure 43: Evaluation Experiment 4. Researcher Comment: "May be the genes that are down-regulated in this list are linked whilst those that are up-regulated are linked together. The genes down-regulated by treatment 3 also look linked."

The experiment example depicted in Figure 43 has been created from genes that the researcher knows respond to ABA which were found using a search function from the database. It clearly shows groups of genes that are up-regulated by three treatments and another group that are down regulated. This indicates cross-talk in the pathway and enables the researcher to hypothesise that these results may mean that there are two or more gene groups that are interlinked in two separate pathways. The treatments involved are also interesting to the research such as drought and sodium chloride seem to share the same gene group which makes sense as drought would increase the salt levels in soil.

Figure 44: Evaluation Experiment 5. Pathway result enabling researcher to hypothesis about the involvement of calreticulin 1 and calreticulin 2 in signalling.

The result shown in Figure 44 was found by a researcher who was interested at looking for groups of genes that are co-regulated in response to different treatments suggesting that they are likely to occur in the same signal transduction pathway. Here the up-regulation of calreticulin 3 (At1g08450) in Arabidopsis is associated with the up-regulation of a number of potential signalling genes including kinases which do not occur if calreticulin 1 (At1g56340) and calreticulin 2 (At1g09210) are down-regulated.

In addition to results found from using the pathway tools, the database toolset has also been used for gaining new knowledge of genes that have been described in publications as unknown. For example 12% of entries in the database are described as 'unknown' function. The toolset is useful for linking together information on ESTs described as genes of unknown function. For example, by converting accession numbers into AGI numbers we have shown that the following ESTs that are down-regulated by chitin (H37231, R90140, T41806), drought (AV823744), ethylene (R90140), low oxygen (At2g10940), or sodium chloride (AV823744), or up-regulated by salicylic acid (R90140, H37231) are all the same gene i.e. At2g10940.

The tools have also been used by researchers to gain an insight into the interaction between biotic and abiotic stress responses by looking at gene expression data within this database using the search or Venn diagram tools. For example, the Arabidopsis genes At2g14560 and At5g14920 of unknown function are down-regulated by cold and drought respectively but are both up-regulated by BTH and by infection with an incompatible isolate of *Peronospora parasitica,* thus suggesting how resistance could be affected detrimentally by environmental conditions.

These results demonstrate that DRASTIC has met its aim of enabling researchers to create new hypotheses for signal transduction. There are several interesting results; one result has prompted the researcher to consider exploring hypothesised results in a laboratory setting and all have given the researcher new insight and enabled new thoughts about possible gene grouping and placement in pathways.

# Chapter 8 Summary and Future Work

## 8.1. Summary

The aim of the project was to create a generic toolset for scientists to assist with investigating gene expression for defence signal transduction mechanisms in plants. Anecdotal evidence from web statistics of visitors to the DRASTIC-INSIGHTS website have shown a steady increase in both number and location base of users (Pers Comm. Lyon, 2007). DRASTIC-INSIGHTS has been used by biologists at SCRI in their research towards induced resistance for plant defence (Walters *et al.*, 2007 and Button *et al.*, 2005). The DRASTIC-INSIGHTS website was described in the book "The Epidemiology of Plant Diseases" (Cooke *et al*, 2006) as a useful tool for epidomiologists:

"*For example, an important aspect of understanding the epidemiology of plant-microbe interactions is to understand how abiotic stress caused by climate, e.g. drought and heat stresses, can affect the susceptibility of a plant to infection. A web resource called DRASTIC (Database Resource for the Analysis of Signal Transduction in Cells) may not at first seem like an epidemiologist's favourite site. Likewise, one of its main resources, a database of plant gene expression data which provides valuable information on the potential interaction between biotic and abiotic stresses, may be difficult at first to relate to epidemiology. However, DRASTIC does not simply provide data from microarrays. The database contains information from a wide range of published papers on whether plant genes are up- or down-regulated in response to various biotic and abiotic stresses. Much of the information is based on experiments with the model plant Arabidopsis thaliana and because the database uses Arabidopsis Genome Initiative (AGI) numbers it is possible to be confident about which gene within a family of genes is actually being regulated.*"

Furthermore, DRASTIC has been used by the ONDEX project (Kohler *et al.*, 2006) which is a tool for biological network analysis. ONDEX provides 2D representations of directed, undirected and weighted networks. It can handle large scale networks of hundred thousands of nodes and edges. Data for integration is modelled as a suitable framework of concepts (such as gene, pathway, and protein) and relations (such as 'belongs_to', 'is_a') describe the mapping between them. In addition, a powerful filter is available to import microarray expression level data to globally analyze the relations between the different genes being

expressed. ONDEX uses data from DRASTIC-INSIGHTS AMONG other sources as described by Pavlopoulos (2008).

Independent use of DRASTIC-INSIGHTS was most recently illustrated in a paper published by Sundar *et al.*, (2008) who have developed a computer algorithm to identify key transcription factor binding sites upstream of a gene of interest. They used the DRASTIC-INSIGHT toolset to identify stress responsive genes based on their consistent up-regulation in response to abiotic stress signals. They found that several genes could be up-regulated during multiple stresses, such as cold, salinity, drought etc. Experimental biochemical validations have proved the involvement of several transcription factors could be involved in the up-regulation of these stress responsive genes. In order to follow the intricate and complicated networks of transcription factors and genes that respond to stress situations in plants, they developed the Stress up-regulated TranscriptIon Factor (STIF) algorithm. This demonstrates the wide and generic application possibilities for the DRASTIC-INSIGHT tool.

### 8.2. Future Work

The database and toolset have met its original objectives and testing has demonstrated that it meets its requirement specification. DRASTIC-INSIGHTS is built on a structured design that stores, and importantly facilitates the maintenance and updating of data from peer reviewed journals and databases that relate to stress response genes that have been shown to be active in early signal transduction events. During this study, there have been many nomenclature issues with the data as described in Chapter 4 and thus the database has been developed to be a generic data container which should be adaptable to what ever the next generation of experimental protocols are.

While there are progressive movements towards unified ontologies and nomenclatures such as GO and the AGI, there is more work required in this field. The data inaccuracy problems encountered during the work on this thesis appears to emanate from the rapid advances that are made in the technologies and knowledge in the field and the staticness of journals. Microarray technologies are constantly evolving and there are still updates being made on gene sequences but these updates do not seem to filter through to the older public data that is being processed and mined by many different enterprises. Findings from observing biologists showed that they were willing to accept the results from computing tools without delving further into the background data which as demonstrated can be flawed or out of date. More

knowledge is lost from journals by the fact that old publications cannot be updated in instances where a gene that at time of publication had an unknown function and as a result knowledge is lost or not recognised. It seems that in the rush to evaluate new data, the older data is ignored and allowed to degenerate. Although unexpected at the start of the project, there has been a focus on evaluating how to maintain and store both new and legacy data in this thesis and DRASTIC is a way forward in an attempt to bridge the gap between journals and microarray databases.

The DRASTIC knowledge base enables these genes to be updated by revised annotation / nomenclature / function as knowledge improves while preserving traceable links to all previous annotation and references. Scientists can access multiple references for the majority of genes and accession numbers that are held in DRASTIC. Searches in pubmed, science direct and ingenta when using accession numbers or AGI numbers do not return these results and as chapter 4 shows, there is no consistent submission format to address this in journals. Drastic uniquely enables scientist to search for genes which are co-expressed by multiple stimuli therefore enabling researchers to establish a start group of genes that they may be interested in (Bülow *et al.* 2007). Drastic has already been used as a basis by ONDEX (Koehler *et al.* 2006) and as many data mining techniques require "bait genes" for example network analysis and learning sets for Soms and neural nets and DRASTIC can easily provide these as it does not have the overhead of hundreds of thousands of data sets that need to be calculated (pair-wise comparison) in order to create a starting group.

Although many of the results presented in section 7.3 to evaluate the tools are based on one species which is *A.thaliana* the system is designed to be able to process multi-species data. The database does hold information on multiple species and the tools are designed to retrieve data from multiple species. For some of the searches (Venn diagram and pathway), a unique identifier is required which in the case of Arabidopsis is the AGI number. The unique locus identifier is needed as it is important to be able to distinguish between genes and not mistakenly report multiple responses for many genes which later turn out to be one gene with many accession numbers. As more progress is made in sequencing different species, more focus is paid to naming conventions as discussed in section 4.2.6. The *Oryza sativa* (rice) nomenclature for locus id are similar in design to the AGI numbers (McCouch, 2008) and analogous naming conventions are being adopted for tomato (Mueller, 2005) and maize (MaizeGDB, 2002). The use of AGI's in DRASTIC-INSIGHTS is proof of concept and the

system is set up to enable the user to select any gene that has a unique identifier of this type of nomenclature to compare genes across multiple species.

An interesting development from NCBI is the HomoloGene database which enables the user to enter a gene of interest and the database retrieves homologs among the annotated genes of several completely sequenced eukaryotic genomes using an automated procedure (Wheeler et al, 2006). This has more recently included *A.thaliana* and *O.sativa* and future work could include making use of this facility to compare potentially related genes in different species and examining their gene expression as the data levels increase.

The biologists ultimate aim would be to create an interactive chart containing all the hypothesised pathways and the biologists issues with current systems that were not fully met from Chapter 2 are:

1. It is not possible to draw separate diagrams for each agonist/response as it would be too time consuming.
2. It may include varying degrees of uncertainty ('informed guesses') that other scientists may find inappropriate or are wrong (by virtue of having not taken into account some other published information).
3. It is not possible to interrogate a diagram.
4. It is not possible to add to the diagram ones own personal or unpublished data.

DRASTIC-INSIGHTS has developed the infrastructure to store and query gene expression data but this could be improved by focussing on the visualisation aspect in the future. A version of the database which included the facility for a single biologist to input hypothesised results and store pathways for further investigation has been developed to prototype level and this partially covers some of these requests but it would be useful to have the ability to view and manipulate these newly found hypothesised pathways.

Pathoplant is a database project which aims to give a comprehensive overview about specific plant-pathogen interactions and to link this information to signal perception and transduction components. This may allow identification of missing links in signal transduction pathways, deduction of the function of novel proteins by comparison with known signal transduction pathways. It stores sequences of known molecules and corresponding reactions and it facilitates easy access to published data. The problem that they have encountered is lack of protein interaction data which it relies on to create the signal transduction pathways. In 2004 the database contains one signalling pathway (ethylene) available with104 interaction records and 26 reaction records over 47 plant species. In 2008 the database contains 1 signalling

pathway, 350 interaction records and 26 reaction records over 96 species although this approach may be very successful when more data becomes available. The database does however have a small scale visualisation of a single pathway (ethylene). Collaboration between this database, DRASTIC-INSIGHT and potentially ONDEX or some similar pathway viewing software could move us nearer to the ultimate goal of being able to visually represent the whole of the data as an interrogative chart.

Appendix I-IV contain conference posters and a published paper for this thesis along with an review from The Biochemist which reviews the work as "an excellent example of a database, probably constructed on a limited budget, that, once it is complete, will prove invaluable to one specialist research community and be of great interest to those outside it. With the 'data swamp' described by the authors showing few signs of becoming more tractable, it is an approach that other specialist research communities might do well to copy."

Drastic-Insights will in the future be extended by linking both public and private domain data to enable scientists to hypothesize using personal and published data. Optional access to Nottingham Arabidopsis Stock Centre (NASC) microarray data will also be made available via individual user domains. Development of text mining and data capture tools in a separate Carnegie project to automatically identify suitable publications and datasets for inclusion in DRASTIC have been undertaken.

# Bibliography


Affymetrix. (2007). Data Sheet GeneChip Arabidopsis Genome Array. Available via: http://www.affymetrix.com/support/technical/datasheets/arab_datasheet.pdf. Cited 27th February 2008.

Affymetrix. (2008). Annotation method description. Available via https://www.affymetrix.com/support/help/IVT_glossary/index.affx#methoddescription. Cited 28th February 2008.

Agrios, G. N. (2005). *Plant Pathology Forth Edition.* Harcourt Academic Press.

Alwine, J.C., Kemp D.J., and Stark, G.R. (1977). *Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.* Proc. Natl. Acad. Sci. U.S.A. 74 (12): pp. 5350- 5354

Aubourg, S. and Rouzé, P. (2001). Genome annotation. Plant Physiology and Biochemistry 39(3-4): pp. 181-193.

Aubry, M., Monnier, A., Chicault, C., de Tayrac, M., Galibert, M. D., Burgun, A. and Mosser, J. (2006). *Combining evidence, biomedical literature and statistical dependence: new insights for functional annotation of gene sets.* BMC Bioinformatics. **7**: pp 241.

Baldi, P. and Hatfield, G. W. (2002). *DNA microarrays and gene expression.* Cambridge Universtiy Press.

Bari, R., Pant, B. D., Stitt, M. and Scheible, W. (2006). *PHO2, MicroRNA399, and PHR1 Define a Phosphate-Signalling Pathway in Plants.* Plant Physiology, 141(3): pp. 988–999.

Barrett, T., Troup, D. B., Wilhite, S., E., Ledoux, P., Rudney, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2006). *NCBI GEO: mining tens of millions of*

*expression profiles—database and tools update.* Nucleic Acids Research. January 35(Database issue): D760–D765.

Bassett Jr, D. E., Eisen, M. B. and Boguski, M. S. (1999). *Gene expression informatics – it's all in your mine.* Nature Genetics. Vol 21(1) pp: 51-55.

Berardini, T. (2006). Personal Communication.

Birch, P. R. J, Blok, V., Philips, M., Jones, J. T., Stewart, H. E., Duncan, J. M., Bryan, G. J., Waugh, R., Lyon, G. D., MacFarlane, S., Avrova, A. O., Whisson, S. C. and Toth, I. K. (2003). Current and future research at SCRI on the molecular genetics of the interactions between potato and its major pathogens. ISHS Acta Horticulturae 619: XXVI International Horticultural Congress: Potatoes, Healthy Food for Humanity: International Developments in Breeding, Production, Protection and Utilization.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J and Vingron M. (2001). *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nature Genetics. Dec; 29(4) pp: 373.

Brickell, C.D., Baum, B.R., Hetterscheid, W.L.A., Leslie, A.C., McNeill, J., Trehane, P., Vrugtman, F. and Wiersema, J.H. (2004). *International Code Of Nomenclature For Cultivated Plants.* Lubrecht & Cramer Ltd.

Bülow, L., Schindler, M., Choi, C., and Hehl, R. (2004). *PathoPlant®: A Database on Plant-Pathogen Interactions. In Silico* Biology.

Burness Communications. (2008). *Scientists Behind 'Doomsday Seed Vault' Ready World's Crops For Climate Change.* ScienceDaily. Available via: http://www.sciencedaily.com/releases/2008/09/080917145518.htm. Cited 14 January 2009.

Button, D. K., Gartland K. M. A., Ball, L. D., Natanson, L., Gartland, J. S., Ghazal, P., Duncan, L., Newton, A. C., Marshall, B. and Lyon, G. D. (2004). Mining Value From Gene Expression Data. *12th International conference on intelligent systems for molecular biology (ISMB 2004)* and *3rd European conference on computational biology (ECCB 2004).* July 31st - August 4th, Glasgow. Abstract A-70.

Button D. K., Heilbronn J., Ball L., Natanson L., Gartland J., Gartland K. M. A., Marshall B., Newton A. and Lyon G. (2005). *Drastic: A Database Resource for the Analysis of Signal Transduction In Cells. (www.drastic.org.uk).* XII International Congress on Molecular Plant-Microbe Interactions 2005 in Cancun, México.

Button, D. K. *et al.* (2006). DRASTIC – INSIGHTS: Querying Information in a Plant Gene Expression Database. *Nucleic Acids Research*, Preprint.

Chen, P. P. (1976). *The entity-relationship model: towards a unified view of data*, ACM Trans on Database Systems 1:1.

Chung, H., Kim, M., Park, C. H., Kim, J., and Kim, J. H. (2004). *ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics.* Nucleic Acids Research. 32: pp.460-464.

Connolly, T. and Begg, C. (2002). Database Systems: A practical approach to design, implementation and management 3rd Edition. Addison-Wesley.

Connolly, T. and Begg, C. (2004). *Database Solutions: A step-by-step guide to building databases* 2nd Edition. Pearson: Addison-Wesley.

Cooke, B. M., Jones, D. G. and Kaye, B. (2006). *The Epidemiology of Plant Diseases* 2nd Edition. Springer.

Dey, P. and Harborne, J. B. (1998). *Plant Biochemistry.* Academic Press.

Dix, A. Findlay, J. Abowd and G. Beale, R. (1998). *Human-Computer Interaction* 2nd Edition. Prentice Hall.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences USA. **95**: pp. 14863-14868.

Forster, T., Roy, D. and Ghazal, P. (2003) Experiments using microarray technology: limitations and standard operating procedures. *Journal of Endocrinology* **178**: 195-204.

Haas, B., J., Wortman, J., R., Ronning, C., M., Hannick, L., I., Smith, R., K. Jr., Maiti, R., Chan, A. P., Yu, C., Farzad, M., Wu, D., White, O. and Town, C., D. (2005). *Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release.* BMC Biology 22(3): pp. 7.

Hammond-Kosack, K.E. and Parker, J.E. (2003) Deciphering plant-pathogen communication: fresh perspectives for molecular resistance breeding. *Current Opinion in Biotech* **14**: 177-193.

Hennig, L., Menges, M., Murray, J. A. H. and Gruissem W. (2003). *Arabidopsis transcript profiling on Affymetrix GeneChip arrays.* Plant Molecular Biology. 53(4): pp. 457-465.

Hein, I., Campbell, E. I., Woodhead, M., Hedley, P. E., Young, V., Morris, W. L., Ramsay, L., Stockhaus, J., Lyon, G. D. and Birch, P. R. J. (2004). *Characterisation of early transcriptional changes involving multiple signalling pathways in the Mla13 barley interaction with powdery mildew ( Blumeria graminis f. sp. hordei).* Planta 218(5) : pp. 803-813.

Holme, D. J. and Peck, H. (1998). *Analytical Biochemistry* 3rd Edition. Longman.

Ihaka, R. and Gentleman, R. (1996). *R: A Language for Data Analysis and Graphics.* Journal of Computational and Graphical Statistics 5: pp. 299-314.

Ilic, K., Kellogg, E. A., Jaiswal, P., Zapata, F., Stevens, P., F., Vincent, L., P., Avraham, S., Reiser, L., Pujar, A., Sachs, M., M., Whitman, N., T., McCouch, S., R., Schaeffer, M., L., Ware, D., H., Stein, L., D. and Rhee, S., Y. (2007). *The Plant Structure Ontology, a Unified Vocabulary of Anatomy and Morphology of a Flowering Plant.* Plant Physiology. 143: pp. 587-599.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jodlicka, A. E., Kawasaki, E., Martinez Murillo, F., Morsberger, L., Lee, H., Peterson, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q. and Yu, W. (2005). Multiple laboratory comparison of microarray platforms. Nature Methods. 2: pp. 345-350.

Jackson, J., Strachan, B., von Schack, D. and Sylvers, L. (2002). Detection of Nucleic Acids Using Chemiluminescence: From Northerns to Southerns and Beyond, *In* Luminescence Biotechnology: Instruments and Applications, Van Dyke, K. and Woodfolk, K. CRC Press, pp. 223-230.

Jen, C. H., Manfield, I. W., Michalopoulos, I., Pinney, J. W., Willats, W. G., Gilmartin, P. M. and Westhead, D. R. (2006). *The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis.* Plant Journal. April; 46(2) pp: 336-348.

Kappel, G., Proll, B., Reich, S. and Retscitzegger, Werner. (2006). Web Engineering. Wiley.

Kapushesky, M., Kemmeren, P., Culhane, A. C., Durinck, S., Ihmels, J., Körner, C., Kull, M., Torrente, A., Sarkans, A., Vilo, J. and Brazma A. (2004) *Expression Profiler: next generation—an online platform for analysis of microarray data* Nucleic Acids Research 32: pp. 465-470.

Kim, K., Cheong, Y. H., Grant, J. J., Pandey, G. K. and Luan, S. (2003). *CIPK3, a Calcium Sensor–Associated Protein Kinase That Regulates Abscisic Acid and Cold Signal Transduction in Arabidopsis.* Plant Cell, 15: pp. 411-423.

Köhler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Rüegg, A., Rawlings, C., Verrier, P. and Philippi, S. (2006). *Graph-based analysis and visualization of experimental results with ONDEX.* Bioinformatics. 22: pp.1383 - 1390.

Kohn, K. W. and Aladjem, M. I. (2006). *Circuit diagrams for biological networks.* Molecular Systems Biology 2: 2006.0002.

Kothapalli, R., Yoder, S. J., Mane, S. and Loughran Jr, T. P. (2002). *Microarray results: how accurate are they?* BMC Bioinformatics 3: pp. 22-32.

Lan, H., Carson, R., Provart, N. J., and Bonner, A. J. (2007). Combining classifiers to predict gene function in Arabidopsis thaliana using large-scale gene expression measurements. BMC Bioinformatics. 8: pp.358.

Lee, J., Klessig, D. F. and Nürnberger, T. (2001). *A Harpin Binding Site in Tobacco Plasma Membranes Mediates Activation of the Pathogenesis-Related Gene HIN1 Independent of Extracellular Calcium but Dependent on Mitogen-Activated Protein Kinase Activity.* Plant Cell 13: pp. 1079-1093.

Lorkowski, S. and Cullen, P. (2006). *From Analysing Gene Expression – A handbook of Methods, Possibilities and Pitfalls* Volume 2. Wiley-VCH.

Lyon G. D., Newton A. C. and Marshall B. (2002). *The need for a standard nomenclature for gene classification and a generic, automated tool to assist in hypothesis formulation in cell signalling.* Molecular Plant Pathology, 3(2): pp. 103-109.

MaizeGDB. (2002). A Standard For Maize Genetics Nomenclature. http://www.maizegdb.org/maize_nomenclature.php. Cited 5th March 2009.

Masys D. (2001) *Linking microarray data to the literature [editorial]*. Nature Genetics 27(6): pp. 9–10.

McCouch, S. R. and CGSNL. (2008). *Gene nomenclature system for rice*. Rice 1(1): pp. 72-84.

MGED (2007). A non-exhaustive list of journals requiring MIAME compliant data as a condition for publishing microarray based papers. Available via: http://www.mged.org/Workgroups/MIAME/journals.html Cited 26th February 2008.

Mount, D. W. (2004). *Bioinformatics – Sequence and Genome Analysis* 2nd Edition. Cold Spring Harbour Laboratory Press.

MPA (2008). Molecular Plant Pathology Author Guidelines. Available via: http://www.blackwellpublishing.com/submit.asp?ref=1464-6722&site=1. Cited 26th February 2008.

MPMI. (2008). Molecular Plant–Microbe Interactions. Instructions for Authors, 2008. Available via: http://apsjournals.apsnet.org/userimages/ContentEditor/1173402237082/mpmi_author_instructions.pdf. Cited 26th February 2008.

Mueller L. (2005) SOL Project Sequencing and Bioinformatics Standards and Guidelines. http://www.sgn.cornell.edu/documents/solanaceae-project/docs/tomato-standards.pdf. Cited 5th March 2009.

Narayanan, A., Keedwell, E. C. and Olsson, B. (2002). *Artificial intelligence techniques for bioinformatics*. Applied Bioinformatics: pp. 191-222.

Nature Opinion. (1997). *Obstacles of Nomenclature*. Nature 1: pp 389.

Newton A. C., Lyon G. D. and Marshall, B. (2002). DRASTIC: a Database Resource for Analysis of Signal Transduction in Cells. BSPP Newsletter 42, 36-37.

Nimblegen. (2008). Gene Expression Microarrays and Services. Available via: http://www.nimblegen.com/products/exp/#eukaryotic. Cited 27th February 2008.

Pavlopoulos, G. A., Wegener, A. and Schneider, R. (2008). *A survey of visualization tools for biological network analysis.* BioData Mining. 1 pp. 12.

PP. (2008). INSTRUCTIONS FOR AUTHORS Plant Physiology 2008. Available via: http://www.plantphysiol.org/misc/ifora.shtml Cited 26th February 2008.

Price CA, Reardon EM, Lonsdale DM. (1996). *A guide to naming sequenced plant genes. Plant Mololecular Biolology.* 30(2): pp. 225-7.

Quackenbush, J. (2001) Computational analysis of microarray data. *Nature Review Genetics* **2**: 418-427.

Redman J. C., Haas B. J., Tanimoto G. and Town C. D. (2004). *Development and evaluation of an Arabidopsis whole genome Affymetrix probe array.* The Plant Journal. 38: pp. 545-561.

Rensink W. A. and Buell, C. R. (2005). *Microarray expression profiling resources for plant genomics trends in plant science.* Trends in Plant Science. 10(12) pp: 603-609.

Ritchie, C. (2002). Relational Database Principles 2nd Edition. Continuum.

Roth, C. M. (2002). *Quantifying Gene Expression.* Current Issues in Molecular Biology. 4(3): pp. 93-100.

Sansom, C. (2005). *DRASTIC: www.drastic.org.uk.* The Biochemist, August: pp. 47-48.

Schenk, P. M., Kazan, K., Wilson, I., Anderson, J. P., Richmond, T., Somerville, S. C. and Manners, J. M. (2000). *Coordinated plant defense responses in Arabidopsis revealed by microarray analysis.* Proceedings of the National Academy of Sciences of the United States of America. 97(2): pp. 11655-11660.

Schlueter, SD, Wilkerson, MD, Huala, E, Rhee, SY, and Brendel, V (2005) Community-based gene structure annotation. TRENDS in Plant Science 10(1):9-14

Seki, M., Ishida, J., Narusaka, M., Fujita, M., Nanjo,T., Umezawa, T., Kamiya, A., Nakajima, M., Enju, A., Sakurai, T., Satou, M., Akiyama, K., Yamamguchi-Shinozaki, K., Carninci, P., Kawai, J., Hayashizaki, Y. and Shinozaki, K. (2002). *Monitoring the expression pattern of around 7,000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray.* Functional Integrated Genomics. 2 pp: 282-291.

Sommerville, I. (2001). *Software Engineering* 6th Edition. Addison-Wesley.

Stoeckert C. J., and Parkinson, H. (2003). *The MGED Ontology: A Framework for Describing Functional Genomics Experiments.* Computational Functional Genomics. 4(1): pp.127–132.

Sundar, A. S., Varghese, S. M., Shameer, K., Karaba, N., Udayakumar, M., and Sowdhamini, R. (2008). *STIF: Identification of stress-upregulated transcription factor binding sites in Arabidopsis thaliana.* Bioinformation. 2(10) pp: 431-437.

Tian, Q., Uhlir, N. J. and Reed, J. W. (2002). Arabidopsis SHY2/IAA3 inhibits auxin-regulated gene expression. *Plant Cell.* 14: pp. 301–319.

Usadel B., Nagel A., Thimm O., Redestig H., Blaesing O. E., Palacios-Rojas N., Selbig J., Hannemann J., Piques M. C., Steinhauser D., Scheible W., Gibon Y., Morcuende R., Weicht D., Meyer S. and Stitt M. (2005). *Extension of the Visualization Tool MapMan to Allow Statistical Anaysis of Arrays, Display of Corresponding Genes, and Comparison with Known Responses.* Plant Physiology. 138: pp. 1195-1204.

Walters, D., Newton, A. and Lyon, G. D. (2007). *Induced Resistance for Plant Defence: A Sustainable Approach to Crop Protection.* Blackwell Publishing.

Wang, Y., Ohara, Y., Nakayashiki, H., Tosa, Y. and Mayama, S. (2005). *Microarray Analysis of the Gene Expression Profile Induced by the Endophytic Plant Growth-Promoting Rhizobacteria, Pseudomonas fluorescens FPT9601-T5 in Arabidopsis.* Molecular Plant-Microbe Interactions. 18(5): pp. 385-396.

Wheeler D. L., Barrett T., Benson D. A., Bryant S. H., Canese K., Chetvernin V., Church D. M., DiCuccio M., Edgar R., Federhen S., Geer L. Y., Helmberg W., Kapustin Y., Kenton D. L., Khovayko O., Lipman D. J., Madden T. L., Maglott D. R., Ostell J., Pruitt K. D., Schuler G. D., Schriml L. M., Sequeira E., Sherry S. T., Sirotkin K., Souvorov A., Starchenko G., Suzek T. O., Tatusov R., Tatusova T. A., Wagner L. and Yaschenko E. (2006). *"Database resources of the National Center for Biotechnology Information."*, Nucleic Acids Research. 34:D173-D180.

VandenBosch, K. A. and J. Frugoli. (2001). *Guidelines for genetic nomenclature and community governance for the model legume Medicago truncatula.* Molecular Plant Microbe Interactions. 14: pp.1364-1367

Whitehorn, M. and Marklyn, B. (2002). *Inside Relational Databases* 2nd Edition. Springer.

Yauk, C. L. and Berndt, M. L. (2007). *Review of the Literature Examining the Correlation Among DNA Microarray Technologies.* Environmental and Molecular Mutagenesis 48: pp.380-394.

Zhao, J. , Zheng, S. , Fujita, K. , and Sakai, K. (2004). Jasmonate and ethylene signalling and their interaction are integral parts of the elicitor signalling pathway leading to ß-thujaplicin biosynthesis in *Cupressus lusitanica* cell cultures. *Journal of Experimental Botany.* 55: pp. 1003-1012.

Zhang, H., Wenzheng, L., Yang, Y. and Zhang, Z. (2007). *Transcriptional activator TSRF1 reversely regulates pathogen resistance and osmotic stress tolerance in tobacco.* Molecular Plant Biology, 63(1): pp. 63-71 .

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. and Gruissem, W. (2004). *GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox.* Plant Physiology. 136(1): pp. 2621-32.

# Appendix I – Nucleic Acids Ressearch Published Paper

*D712–D716 Nucleic Acids Research, 2006, Vol. 34, Database issue*
*doi:10.1093/nar/gkj136*

# DRASTIC—INSIGHTS: querying information in a plant gene expression database

**Davina K. Button[1,*], Kevan M. A. Gartland[1,4], Leslie D. Ball[2], Louis Natanson[2], Jill S. Gartland[1] and Gary D. Lyon[3]**

[1]Abertay Centre for the Environment and [2]School of Computing and Creative Technologies, University of Abertay Dundee, Dundee DD1 1HG, Scotland, UK, [3]Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK and [4]School of Life Sciences, Glasgow Caledonian University, Glasgow G4 0BA, Scotland, UK

Received August 15, 2005; Revised and Accepted October 27, 2005

## ABSTRACT

**DRASTIC—Database Resource for the Analysis of Signal Transduction In Cells (http://www.drastic.org.uk/) has been created as a first step towards a data-based approach for constructing signal transduction pathways. DRASTIC is a relational database of plant expressed sequence tags and genes up- or down-regulated in response to various pathogens, chemical exposure or other treatments such as drought, salt and low temperature. More than 17 700 records have been obtained from 306 treatments affecting 73 plant species from 512 peer-reviewed publications with most emphasis being placed on data from *Arabidopsis thaliana*. DRASTIC has been developed by the Scottish Crop Research Institute and the University of Abertay Dundee and allows rapid identification of plant genes that are up- or down-regulated by multiple treatments and those that are regulated by a very limited (or perhaps a single) treatment. The INSIGHTS (INference of cell SIGnaling HypoTheseS) suite of web-based tools allows intelligent data mining and extraction of information from the DRASTIC database. Potential response pathways can be visualized and comparisons made between gene expression patterns in response to various treatments. The knowledge gained informs plant signalling pathways and systems biology investigations.**

## INTRODUCTION

Recovering value from the burgeoning mass of genomics and gene expression data now being accumulated is a major task for biologists and computer scientists (1). Increasing amounts of gene sequence, expressed sequence tag (EST), northern blot and microarray data provide fertile ground for the mining of expression data, extracting information and adding value by evaluating how gene expression is regulated and biochemical pathways function (2). DRASTIC (Database Resource for the Analysis of Signal Transduction in Cells) and the INSIGHTS (INference of cell SIGnaling HypoTheseS) web-based suite of tools bring together data on plant responses to pathogens, environmental stresses and chemicals (treatments) from refereed journal publications. Presenting these data in a unified, searchable format allows the user to extract information beyond single genes, or clusters of similar expression patterns by browsing multiple treatments at once, identifying potential regulatory relationships between multiple treatments and genes. DRASTIC–INSIGHTS overcomes the limitations of other plant expression databases by allowing for updating of information from previous publications, by directly linking to publications and through the tracking of genes with unknown function that have the same accession or AGI (*Arabidopsis* genome initiative) number, which would otherwise be difficult to link between publications (3,4). Additionally, genomic, EST, northern data and information derived from microarrays from multiple plant species are included, after human curation, to ensure accuracy and to standardize the nomenclature of data (5). The INSIGHTS tools encourage comparison of gene expression patterns, intelligent mining of information, testing and formulation of novel hypotheses on the complex signal transduction and response pathways used by plants (6). Identifying common elements in pathways affected by different treatments permits the formation of hypotheses previously opaque to the user (7).

### Database content

DRASTIC is a gene expression relational database developed by SCRI (Scottish Crop Research Institute) and UAD (University of Abertay Dundee) to record responses to treatments, which are defined as exposure to experimental conditions such

*To whom correspondence should be addressed. Tel: +44 1382 308000; Fax: +44 1382 308626; Email: davina.button@abertay.ac.uk

as pathogens, chemicals and other environmental stresses. More than 17 700 records are included with information on 73 species and 306 treatments obtained from 512 references. Each record contains expression data for a single gene, from a single host, subjected to a treatment obtained from a single refereed journal publication. Manually curated records include data from plant northern blots, ESTs, cDNA-AFLPs, quantitative RT–PCR, massively parallel signature sequencing and information derived from microarrays. These expression data are recorded as up- or down-regulated compared with control values, and, where applicable, the time and magnitude of expression are also recorded. DRASTIC makes it possible, for example, to rapidly identify plant genes that are up-regulated by multiple treatments and those that are up-regulated by a single treatment (see http://www.scri.sari.ac.uk/TiPP/PPS/DRASTIC/mpage/countoftreatment.asp for numbers of records per treatment). Such information represents important knowledge to assist in constructing putative signalling pathways for systems biology research (8). Database requirements were elicited using semi-structured interviews with computer scientists and bioscientists. The complex ERM (entity relationship model) consists of over 20 tables (see explanatory tables and ERM diagram in Supplementary Data). The ERM is implemented in Microsoft RDBMS (relational database management system) and is searched using the public web-based interface hosted by SCRI on Microsoft 2000 Advanced Server. The web toolkit was developed using SQL (structured query language) embedded in ASP (active server pages) to dynamically create HTML result pages based on user queries. All records in the DRASTIC–INSIGHTS database are accessible through the publically available website http://www.drastic.org.uk or can be freely downloaded in a comma delimited text file from the website download page (http://www.scri.sari.ac.uk/TiPP/PPS/DRASTIC/mpage/downloads.asp).

## Data quality

Several methods have been implemented to ensure that the data stored in DRASTIC are of high quality. Inclusion in the database is solely following expert human curation of expression data from refereed publications. No expression data have been included by direct submission from laboratories. Accession numbers for ESTs are preferred, as the nomenclature of such sequences can be updated in the future. Information from some papers has not been included because accession numbers were not provided. The need to standardize nomenclature is important (5), thus the names used in the database correspond with current Unigene classification (9) rather than those cited in the original publication, unless a more recent primary publication indicates otherwise. Sometimes changes in gene identification are small but in other cases they can be dramatic and critical if signal transduction pathways are to be correctly understood. For example, one plant gene originally described as senescence-associated is now described as inositol-1,4,5-triphosphate 5-phosphatase, and another gene originally described as 'no homology' is now known to be a protein kinase. In addition, with genes from *Arabidopsis*, the Unigene system has been used to provide AGI (*Arabidopsis* genome initiative) numbers where known. This has proven particularly useful for genes classified as 'unknown', as 'unknowns' from different publications and with different accession numbers can be shown to be the same gene. For example, At2g36220 is classified as a gene of unknown function but by using information from many references we can see that it is up-regulated by abscisic acid, brassinosteroid, benzothiadiazol (BTH), cold, flagellin-22, hydrogen peroxide, low oxygen, *Peronospora parasitica* and sodium nitroprusside treatments. A backtracking facility has been included for historical gene names as all updates are stored in a data dictionary. In addition, a software routine called AGIDetect has been developed to check for mismatches between AGI, gene names and accession numbers to assist in maintaining data accuracy.

## INSIGHTS data tools

At a simple level the web interface (http://www.drastic.org.uk) permits users to find published information on expression data for plant genes of interest. More importantly, INSIGHTS offers a number of tools to mine further information and create new knowledge. Some mining tools use AGI numbers where expression data correctly identify a specific member of a gene family. Through the INSIGHTS integrated toolkit users may investigate data in the following ways:

(i) *General database search* provides a basic query function for the database. The user can select the following parameters: treatments, species, gene, regulation and date. The search returns the results in tabular format which can be sorted on all parameters and provides links to the primary references.

(ii) *DRASTIC statistics* provides an up-to-date list of statistics for the database including the total number of records, species and treatments. It also provides a breakdown of both records per species and records per treatment, which can be ordered alphabetically or numerically. To gain a more in-depth view, a table of data providing statistics on the number of records by species or treatment can be obtained. These can be further mined to view individual records with bibliographic references.

(iii) *Accession number search* provides a query function specifically for the accession numbers in the DRASTIC database. Selectable parameters include accession number, treatment, regulation type and date. The results are displayed in tabular format which can be sorted, providing links to references.

(iv) *Arabidopsis genome initiative search* provides a query function specifically for AGI numbers in the DRASTIC database. The user can select from AGI number, treatment, regulation and date.

(v) *Venn diagrams* enables the creation of Venn diagrams using the *Arabidopsis thaliana* data from the DRASTIC database. The user can select two or three treatments and the tool will process the selections and output the results as a Venn diagram. The Venn diagram tool displays the number of genes regulated by each individual treatment or by multiple treatments based on the DRASTIC data. Records where genes have been up-regulated, down-regulated or both (up or down) can be included. The diagrams can be mined further by clicking on a segment of the diagram to view the individual records and relevant bibliographies.

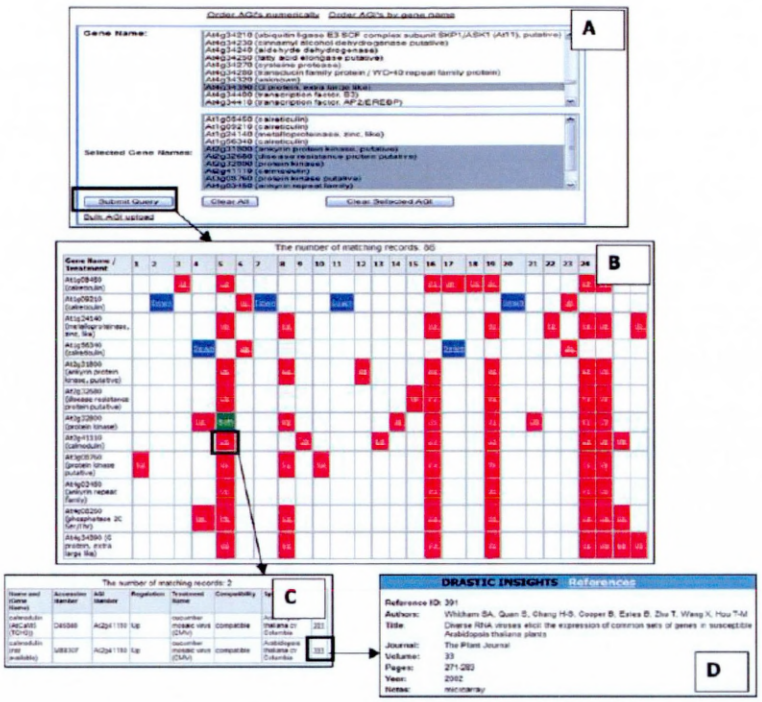



**Figure 1.** Web interface for the Pathway tool. (A) The search page for a set of AGI numbers. The pathway result is shown in (B). Up-regulated genes are shown in red. Down-regulated genes are shown in blue. Green cells indicate that both up- and down-regulation record(s) are held in DRASTIC. The pathway can be further mined by choosing any coloured cell which will display all the records for the AGI/treatment combination as shown in (C). The references for each record can be selected as shown in (D).

(vi) *TAIR AGI search* enables the user to search records that include the AGI number and directly use them with the TAIR (the *Arabidopsis* information resource) chromosome mapping and functional categorization tools, which are specifically designed to analyse AGI data (10). The user can select a subset of records from DRASTIC using a search on a treatment, multiple treatment or gene group (such as kinases) and regulation type. The selected data are then formatted for use with the TAIR tools.

(vii) *Pathway tool* enables the user to extract and visualize knowledge from the database to hypothesize potential relationships between signalling elements. It includes a search facility to allow selection of a number of *A.thaliana* genes by AGI numbers. A 'pathway' is produced to display the regulation of selected genes in response to different treatments (Figure 1). Any groups of genes that are always co-regulated are identified, suggesting that they are likely to occur in the same signal transduction pathway. The pathway tool can be used to indicate the relatedness of induction patterns for selected genes. For instance, it can be shown that up-regulation of calreticulin 3 (At1g08450) in *Arabidopsis* has been shown to be associated with the up-regulation of a number of potential signalling genes (including kinases), which does not occur if calreticulin 1 (At1g56340) and calreticulin 2 (At1g09210) are down-regulated. The pathway tool can also be used in a hypothesis testing manner or as a quality control check tool for data in known signal transduction pathways (11).

(viii) *Roadmap tool* creates lookup tables to find genes that are co-regulated by different treatments. The user can 'drill down' through the map to investigate individual genes and view all references that support each data point providing a level of confidence for each result. To operate the roadmap, the user selects an AGI number and a regulation (up-, down- or both) to include in the search. The tool establishes which treatments regulate expression of the selected gene and then displays in a map all the genes in DRASTIC that are regulated by these treatments (Figure 2). This tool demonstrates that it is possible to identify groups of treatments that appear to produce similar regulatory results in *A.thaliana*. Roadmap results can be used in conjunction with the Pathway tool.

(ix) *Unique genes tool* identifies all the *A.thaliana* genes that are regulated by a single treatment. Full details including references for each gene are linked to each record.

Genes for proteins involved in the same signal transduction pathway are likely to be co-regulated and show the same

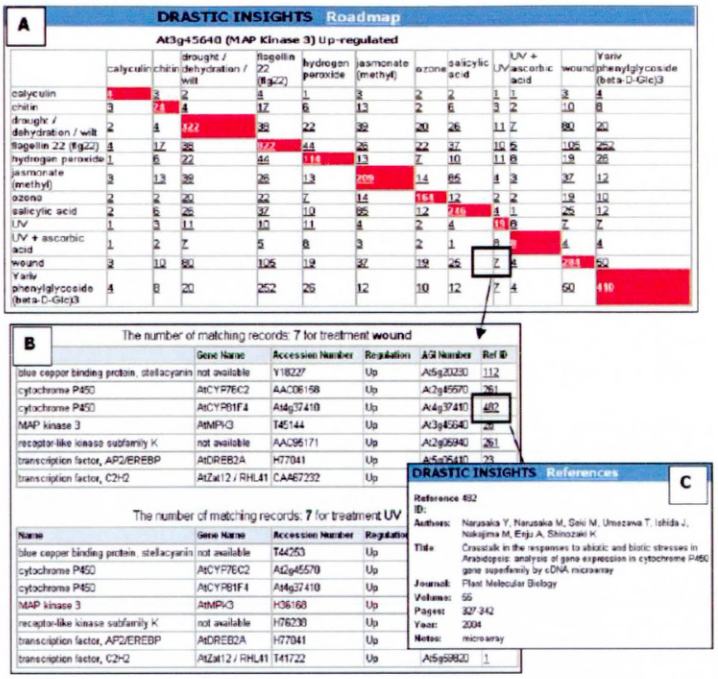



**Figure 2.** Web interface for the Roadmap tool. In this example, treatments up-regulating At3g45640 (MAP Kinase 3) were selected for investigation. (**A**) The resulting roadmap. From the DRASTIC data, 12 treatments up-regulate Atg45640. Using these treatments as the 'lookup co-ordinates', the map displays the total number of unique AGIs up-regulated by these treatments. The shaded squares hold the total number of genes up-regulated by a single treatment, and the numbers in the unshaded squares show the number of genes co-regulated by treatments. This map can be further mined by clicking on any of the squares to display the supporting records (see (**B**) where the co-ordinates wound and UV have been selected). Each record has a link to the reference it was curated from as shown in (**C**).

response to a range of treatments. Thus, to find e.g. kinases, transcription factors and calcium-binding proteins that are in the same signal transduction pathway expression patterns should be compared. Verification that identified genes are truly associated within signal transduction or metabolic pathways requires experimental confirmation, but the database and associated diagrams promote more targeted hypothesis formation. This type of analysis is useful in providing a framework for understanding signal transduction responses and to assist with identifying regulatory gene networks. It is also useful for finding genes associated with plant pathogen infection that are also affected by environmental stresses such as drought and cold in differing ways (12, 13). A downloadable guide to using DRASTIC–INSIGHTS has been developed to assist users and is available at http://www.scri.sari.ac.uk/TiPP/PPS/DRASTIC/helpfiles/index.html.

### Future work

DRASTIC–INSIGHTS will in the future be extended by linking both public and private domain data to enable scientists to hypothesize using personal and published data. Optional access to Nottingham Arabidopsis Stock Centre (NASC) microarray data will also be made available via individual user domains. Development of text mining and data capture tools to automatically identify suitable publications and datasets for inclusion in DRASTIC is currently being undertaken.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The contributions of Bruce Marshall, Adrian Newton, Peter Ghazal, Ishbel Duncan and Michael Idowu to the conceptual development of DRASTIC are acknowledged. DRASTIC–INSIGHTS is supported by funding from the Scottish Executive Environment Rural Affairs Department (SEERAD), Carnegie Trust, the Forestry Commission and the University of Abertay Dundee (UAD). The website was funded by Mylnefield Research Services Ltd. Infrastructure support was provided by Abertay Centre for the Environment

(ACE). Funding to pay the Open Access publication charges for this article was provided by JISC.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Tian,Q., Uhlir,N.J. and Reed,J.W. (2002) *Arabidopsis* SHY2/IAA3 inhibits auxin-regulated gene expression. *Plant Cell*, **14**, 301–319.
2. Kazic,T. (1994) Biochemical Databases: Challenges and Opportunities. New Data Challenges in Our Information Age. In Glaeser,P.S. and Millward,M.T.L. (eds), *Proceedings of the 13th International CODATA Conference*. CODATA Secretariat, Paris, pp. C133–C140.
3. Mueller,L.A., Zhang,P. and Rhee,S.Y. (2003) AraCyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiol.*, **132**, 453–460.
4. Thimm,O., Blaesing,O., Gibon,Y., Nagel,A., Meyer,S., Krueger,P., Selbig,J., Mueller,L.A., Rhee,S.Y. and Stitt,M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914–939.
5. Lyon,G.D., Newton,A.C. and Marshall,B. (2002) The need for a standard nomenclature for gene classification (a Nucleotide Function code) and an automated data based tool to assist in understanding the molecular associations in cell signalling in plant pathogen interactions. *Mol. Plant Pathol.*, **3**, 103–109.
6. Cheong,Y.H., Chang,H. S., Gupta,R., Wang,X., Zhu,T. and Luan,S. (2002) Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in *Arabidopsis*. *Plant Physiol.*, **129**, 661–677.
7. Kunkel,B.N. and Brooks,D.M. (2002) Cross talk between signaling pathways in pathogen defense. *Curr. Opin. Plant Biol.*, **5**, 325–331.
8. Zimmermann,P., Hirsch-Hoffmann,M., Hennig,L. and Gruissem,W. (2004) GENEVESTIGATOR. *Arabidopsis thaliana* microarray database and analysis toolbox. *Plant Physiol.*, **136**, 2621–2632.
9. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
10. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
11. Gonzali,S., Loreti,E., Novi,G., Poggi,A., Alpi,A. and Perata,P. (2005) The use of microarrays to study the anaerobic response in *Arabidopsis*. *Ann. Bot.*, **96**, 661–668.
12. Norman-Setterblad,C., Vidal,S. and Palva,E.T. (2000) Interacting signal pathways control defense gene expression in *Arabidopsis* in response to cell wall-degrading enzymes from *Erwinia carotovora*. *Mol. Plant Microbe Interact.*, **13**, 430–438.
13. McDowell,J.M. and Woffenden,B.J. (2003) Plant disease resistance genes: recent insights and potential applications. *TRENDS Biotechnol.*, **21**, 178–183.

Appendix II (pp. 156-157) - **Independent Review of DRASTIC** by Clare Sansom, Birkbeck College, London, has been removed from this e-thesis due to copyright restrictions.

The review was published in *The Biochemist,* August 2005 , pp 47-48.

# Appendix III – Conference Poster

Button D. K., Heilbronn J., Ball L., Natanson L., Gartland J., Gartland K. M. A., Marshall B., Newton A. and Lyon G. (2005). *Drastic: A Database Resource for the Analysis of Signal Transduction In Cells. (www.drastic.org.uk).* XII International Congress on Molecular Plant-Microbe Interactions 2005 in Cancun, México

**Poster Abstract**

In recent years there has been an explosion in data acquisition regarding the molecular biology of plant-microbe interactions. Unfortunately much of it is ignored or lost as it cannot readily be searched, gene names have changed with time, and many genes of unknown function are being ignored. Consequently our understanding of the implications of that data has not been fully exploited. We have therefore set up a relational database containing stress-responsive gene expression data derived from Northerns and microarrays that have been published in refereed papers. The database currently contains approximately 14,000 entries covering 67 plant species and 289 stress treatments, both abiotic and biotic, and can be queried from our web site at www.drastic.org.uk. This database has enabled us to up-date the nomenclature of genes cited in 'old papers' and to link publications on genes classified as unknown function. Querying the database can rapidly identify treatments that regulate expression of selected genes. For example, the database has enabled us to identify genes that are down-regulated by cold and drought but which are up-regulated in response to infection by a pathogen thus providing a molecular insight into how environmental factors may affect disease resistance. Importantly, we are developing a suite of software tools to extract and visualise knowledge from the database to hypothesise possible relationships between signalling genes that can then be tested experimentally. For instance, up-regulation of a number of signalling genes in response to virus infection correlates with up-regulation of calreticulin 3 (At1g08450) but not with the other calreticulins.

# Drastic: A Database Resource for the Analysis of Signal Transduction In Cells

www.drastic.org.uk

Davina Button[1], Jacqueline Heilbronn[2], Les Ball[1], Louis Natanson[1], Jill Gartland[1], Kevan Gartland[1], Bruce Marshall[2], Adrian Newton[2], Gary Lyon[2]
[1] University of Abertay, Dundee DD1 1HG, Scotland, UK
[2] Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK.

Scottish Crop Research Institute

UNIVERSITY OF ABERTAY DUNDEE

Although there has been an explosion in data acquisition regarding the molecular biology of plant-microbe interactions in recent years our understanding of the implications of that data has not been fully exploited. As a consequence of this 'data swamp' much published information is ignored or 'lost'.

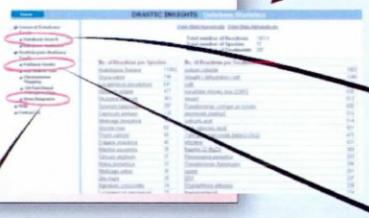
Whilst we have been able to draw 'static' cell signalling diagrams containing a limited amount of (sometimes speculative) information such diagrams are of limited value when trying to interpret large amounts of data from microarray experiments. Thus in future, we expect such diagrams to be drawn dynamically in response to a query and to be driven by a combination of public and private domain databases.

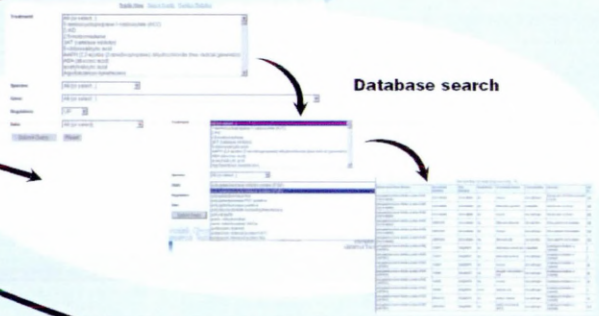
As a first step in this process we have set up a relational database of 'stress-responsive' gene expression data populated with information from published refereed papers containing gene expression data (Northerns and Microarrays). This database can be searched from our web site at **www.drastic.org.uk.**


## The database

The database currently contains approx 15,300 entries covering 70 plant species and 297 treatments.

The database can be searched for published information on these 'stress-responsive' genes.

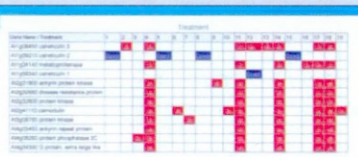




**Database search**

## Pathways

We are developing a suite of software tools to extract and visualise knowledge from the database to hypothesise possible relationships between potential signalling genes. Looking at the regulation of genes in response to different treatments it is possible to see that certain groups of genes are always co-regulated suggesting that they are likely to occur in the same signal transduction pathway. For instance, using a prototype of a 'PATHWAY' software we have developed we can show that up-regulation of calreticulin 3 (At1g08450) in Arabidopsis is associated with the up-regulation of a number of potential signalling genes (inc. kinases) which does not occur if calreticulin 1 (At1g56340) and calreticulin 2 (At1g09210) are down-regulated.



**Venn diagrams**

## Unknown genes

12% of entries in the database are described as 'unknown' function.

The database is useful for linking together information on ESTs described as genes of unknown function. For example, by converting accession numbers into AGI numbers we have shown that the following ESTs that are down-regulated by chitin (H37231, R90140, T41806), drought (AV823744), ethylene (R90140), low oxygen (At2g10940), or sodium chloride (AV823744), or up-regulated by salicylic acid (R90140, H37231) are all the same gene i.e. At2g10940.

## Biotic and abiotic interactions

It is possible to gain an insight into the interaction between biotic and abiotic stress responses by looking at gene expression data within this database. For example, the Arabidopsis genes At2g14560 and At5g14920 of unknown function are down-regulated by cold and drought respectively but are both up-regulated by BTH and by infection with an incompatible isolate of *Peronospora parasitica*, thus suggesting how resistance could be affected detrimentally by environmental conditions.

**References**
Lyon GD, Newton AC and Marshall B. 2002. The need for a standard nomenclature for gene classification (a Nucleotide Function Code) and an automated data-based tool to assist in understanding the molecular associations in cell signalling in plant-pathogen interactions. Molecular Plant Pathology, 3, 103-109.

**Acknowledgements**
JH, GL, AN and BM are grateful to SEERAD for funding

Travel award from BSPP

Sponsored by The Biochemical Society

# Appendix IV – Conference Poster

Button D. K., Gartland K. M. A., Ball L. D., Natanson L., Gartland J. S., Ghazal P., Duncan I., Newton A. C., Marshall B. and Lyon G. D. (2004). Mining Value From Gene Expression Data. 12th International conference on intelligent systems for molecular biology (ISMB 2004) and 3rd European conference on computational biology (ECCB 2004). July 31$^{st}$ - August 4$^{th}$ , Glasgow. Abstract A-70

**Poster Abstract**

DRASTIC is a relational database for gene expression developed by University of Abertay Dundee and Scottish Crop Research Institute to record molecular plant responses to treatments. The INSIGHT Project is focused on developing a suite of tools to intelligently mine the database. Techniques used include Data Mining , clustering and visualisation. INSIGHT tools are available at http://www.drastic.org.uk.

UNIVERSITY of ABERTAY DUNDEE

Scottish Crop Research Institute

www.drastic.org.uk

Database Resource for Analysis of Signal Transduction In Cells  Inference of Cell Signalling Hypotheses

# DRASTIC INSIGHTS

Mining Value from Gene Expression Data

Davina K. Button[1], Kevan M. A. Gartland[2], Leslie D. Ball[1], Louis Natanson[1], Jill S. Gartland[2], Peter Ghazal[4], Isobel Duncan[5], Adrian C. Newton[3], Bruce Marshall[6] and Gary D. Lyon[1]

[1]School of Computing and Advanced Technologies, University of Abertay Dundee, UDT1HG, UK, [2]School of Contemporary Sciences, University of Abertay Dundee, UDT1HG, UK, [3]Scottish Crop Research Institute, Invergowrie, Dundee, DD2 5DA, UK, [4]Scottish Centre for Genomic Technology and Informatics, University of Edinburgh, Edinburgh EH9 3JT, UK, [5]School of Computing, The Robert Gordon University, Aberdeen, AB25 1HG, UK

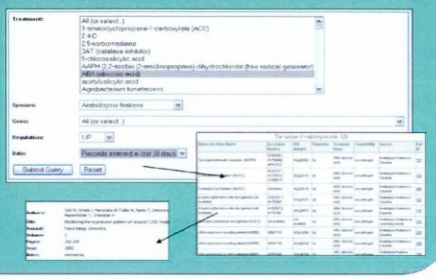
## 1. Introduction

Making sense of the enormous quantities of gene expression data now becoming available is a central problem for all bioscientists. Developing easy to use tools to enable valuable information to be extracted from complex datasets is crucial to achieving this aim.

- The DRASTIC INSIGHTS project builds an intelligent and generic system for new hypothesis formulation from complex biochemical pathway databases.
- DRASTIC contains a relational database for gene expression developed by Scottish Crop Research Institute (SCRI) and University of Abertay Dundee (UAD) to record molecular plant responses to pathogen and other environmental stress treatments e.g. drought, sodium chloride, high and low temperatures.
- INSIGHT develops a suite of tools to intelligently mine the database allowing novel hypotheses for response regulation to be created.

## 2. Gene Expression Database

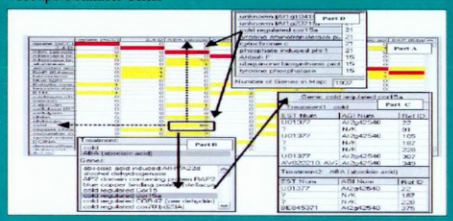
- **www.drastic.org.uk** offers web-based access to the providing tools to search the database
- Incorporates published microarray and Northern data of stress responsive plant genes
- Tracks Arabidopsis Genome Initiative (AGI) numbers through NCBI
- Permits updating of gene names
- Quality control software
- Currently contains information on
  - ❖300 references
  - ❖270 treatments
  - ❖65 plant species
  - ❖1160 responses



## 3. Results

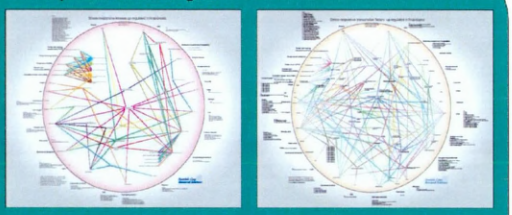
### 3.1 **Roadmap** is a suite of search tools that:

- Lists all genes that are only regulated by one treatment
- Displays a "Roadmap" of all genes that are co-regulated by corresponding treatments
- Groups Common Genes



Gene Roadmap screen showing results from a search of all genes from Arabidopsis thaliana which are up-regulated

### 3.2 Expression Pattern Diagrams

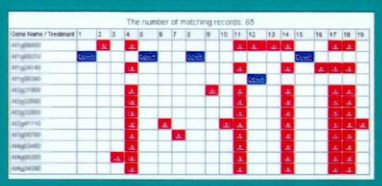


Proteins involved in the same signal transduction pathway are likely to be co-regulated and show the same response to a range of treatments. To find kinases, transcription factors, and calcium-binding proteins that are in the same signal transduction pathway one should compare expression patterns between these diagrams.

### 3.3 Pathway Builder

Genes with similar expression patterns are likely to be in the same signal transduction pathway.

- The Pathway Search allows the user to build up "pathways" of genes.
- The results display the gene regulation for each treatment that the database holds results for.
- The user can identify genes that have similar response patterns which permits hypotheses postulation which can be tested experimentally.

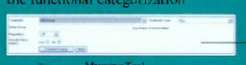


Output from Pathway Tool

Potential shared signalling pathways

### 3.4 Chromosome Mapping

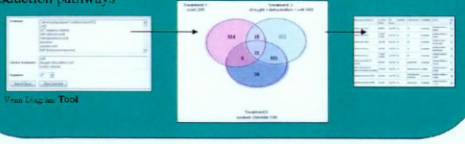
For Arabidopsis genes regulated by a specific treatment can be mapped on the chromosome using software available at TAIR (www.Arabidopsis.org) as shown below or in the TAIR GO program that provides pie charts showing the functional categorization.



Chromosome Mapping Tool

### 3.5 Venn Diagrams

The Venn Tool shows us which genes are involved in overlapping signal transduction pathways.



Venn Diagram Tool

## 4. Summary

- Enables one to retrieve published information on gene expression data for stress responsive plant genes.
- INSIGHT's interface has produced visualisations giving rise to novel hypotheses which where previously opaque to the user of the database and include identification of common elements in pathways affected by different treatments.
- Hypotheses can be used to inform further experimentation by bioscientists whose findings could enrich the database.

Acknowledgements: Carnegie Trust and SEERAD for funding

# Appendix V – Experiment Results (13 Experiments)

## Result 1

The number of matching records: 78

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| At1g08450 (calreticulin) | | Up | | Up | | | | | | | | | | Up | Up | Up | Up | | | | Up | Up | | |
| At1g09210 (calreticulin) | Down | | | | Down | | | | Down | | | | | | | | | Down | | | | | | |
| At1g24140 (metalloproteinase, zinc, like) | | | | Up | | Up | | | | | | | | Up | | | Up | | | Up | Up | Up | | Up |
| At1g56340 (calreticulin) | | | Down | | | | | | | | | | | | Down | | | | | | | | | |
| At2g31800 (ankyrin protein kinase, putative) | | | | Up | | | | | | Up | | | | Up | | | Up | | | | Up | Up | | |
| At2g32680 (disease resistance protein putative) | | | | Up | | | | | | | | | Up | Up | | | Up | | | | Up | Up | | |
| At2g32800 (protein kinase) | | | Up | Both | | | | | | | | Up | | Up | | | Up | | Up | | Up | Up | | |
| At2g41110 (calmodulin) | | | | Up | | Up | | | | | Up | | | Up | | | Up | | | | Up | Up | Up | |
| At3g08760 (protein kinase putative) | | | | Up | | Up | | Up | | | | | | Up | | | Up | | | | Up | Up | | |
| At4g03450 (ankyrin repeat family) | | | | Up | | | | | | | | | | Up | | | Up | | | | Up | Up | | |
| At4g08260 (phosphatase 2C Ser/Thr) | | | Up | Up | | Up | | | | | | | | Up | | | Up | | | | Up | Up | Up | |
| At4g34390 (G protein, extra large like) | | | | Up | | Up | | | | | | | | Up | | | Up | | | | Up | Up | | Up |

| Number | Treatment |
|---|---|
| 1 | Alternaria brassicicola |
| 2 | BTH |
| 3 | cold |
| 4 | cucumber mosaic virus (CMV) |
| 5 | ethylene |
| 6 | flagellin 22 (flg22) |
| 7 | heat stress |
| 8 | hydrogen peroxide |
| 9 | jasmonate (methyl) |
| 10 | light |
| 11 | malondialdehyde (MDA) |
| 12 | mannitol (0.2M) |
| 13 | methanol |
| 14 | oilseed rape tobamovirus (ORMV) |
| 15 | ozone |
| 16 | Peronospora parasitica |
| 17 | potato virus X potexvirus (PVX) |
| 18 | salicylic acid |
| 19 | sodium chloride |
| 20 | sodium nitropruside (an NO donor) |
| 21 | turnip mosaic potyvirus (TuMV) |
| 22 | turnip vein clearing tobamovirus (TVCV) |
| 23 | wound |
| 24 | Yariv phenylglycoside (beta-D-Glc)3 |

Researcher Comment:

## Result 2

The number of matching records: 90

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| At1g09340 (RNA binding protein putative g5bf) | Down | | Down | | Down | | Down | | | | Up | | Down | | |
| At1g16850 (unknown) | Up | | Up | | Up | | | | | | | | Up | | |
| At1g20340 (DNA-damage-repair/tolerationprotein DRT112 precursor like) | | Up | Down | | Down | | Down | Up | | | Up | | Down | | |
| At1g43160 (AP2 domain containing protein RAP2.6, ERF family) | Up | | Up | | Up | | | | | | | | Up | | |
| At1g62570 (monooxygenase, flavin-containing) | Up | | Up | | Up | | | | Up | | | | Up | | |
| At1g78070 (WD-40 repeat family) | Up | | Up | Down | Up | | | | | | | | Up | | |
| At2g23120 (unknown) | Up | | Up | | Up | | | | | Down | | | Up | | |
| At2g26980 (SNF1-related kinase) | Up | | Up | | Up | | | | | | | | Up | Up | |
| At2g30360 (protein kinase 11 CBL-interacting (CIPK11) / SNF1-related kinase) | Up | | Up | | Up | | | | Up | | | | Up | Up | |
| At2g33380 (calcium-binding EF-hand protein) | Up | | Up | | Up | | | | Up | | | | Up | Up | |
| At2g45820 (remorin (DNA binding protein) putative) | | | Up | | Up | | | | Up | | | | Up | Up | |
| At2g46270 (transcription factor, bZIP G box binding factor GBF3) | Up | | Up | | Up | Up | Up | | | | Up | Up | Up | | |
| At2g47770 (benzodiazepine receptor-related) | | | Up | Down | Up | | | | Up | | | | Up | | |
| At3g61890 (transcription factor, homeobox leucine zipper) | Up | | Up | Down | Up | | | | Up | | | | Up | Up | Down |
| At4g26080 (protein phosphatase 2C ABI1 / PP2C ABI1 / abscisic acid-insensitive 1 (ABI1)) | Up | | Up | | Up | Up | | | | | | | Up | | Down |
| At4g27410 (NAC family) | Up | | Up | | Up | | | | Up | | | | Up | | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | ABA (abscisic acid) |
| 2 | Alternaria brassicicola |
| 3 | cold |
| 4 | cucumber mosaic virus (CMV) |
| 5 | drought / dehydration / wilt |
| 6 | ethylene |
| 7 | jasmonate (methyl) |
| 8 | light |
| 9 | mannitol (0.2M) |
| 10 | ozone |
| 11 | salicylic acid |
| 12 | senescence |
| 13 | sodium chloride |
| 14 | wound |
| 15 | Yariv phenylglycoside (beta-D-Glc)3 |

Researcher Comment: “Put these into pathway search. Lots of possibly interesting stuff. More to follow later probably”

## Result 3

The number of matching records: 52

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| At4g34000 (bZIP abscisic acid responsive elements-binding factor) | Up | Up | Down | Up | Up | | | | | Up | | | | Up | | | |
| At5g05410 (DRE-binding protein (DREB2A)) | | Up | Down | Up | | | Up | Up | Up | | | | | Up | Up | Up | Up |
| At5g15850 (zinc finger protein CONSTANS like 1) | | Up | | Up | | Down | | | | | | | | Up | | | |
| At5g15960 (kin1) | Up | Up | | Up | | | | | | Up | | Down | Up | Up | | | Up |
| At5g15970 (kin2 (cold regulated cor6.6)) | Up | Up | | Up | | | | | | Up | | | | Up | | | |
| At5g17300 (transcription factor, Myb) | Up | Up | | Up | | | | | | | Down | | | Up | | | |
| At5g17460 (unknown) | Up | Up | | Up | | | | | | | | | | Up | | | |
| At5g17465 (unknown) | Up | Up | | Up | | | | | | | | | | Up | | | |
| At5g53140 (protein phosphatase 2C putative) | Down | Down | | Down | Up | | | | | | | | | Down | | | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | ABA (abscisic acid) |
| 2 | cold |
| 3 | cucumber mosaic virus (CMV) |
| 4 | drought / dehydration / wilt |
| 5 | ethylene |
| 6 | flagellin 22 (flg22) |
| 7 | harpin |
| 8 | hydrogen peroxide |
| 9 | malondialdehyde (MDA) |
| 10 | mannitol (0.2M) |
| 11 | ozone |
| 12 | Peronospora parasitica |
| 13 | Pseudomonas syringae pv tomato |
| 14 | sodium chloride |
| 15 | UV |
| 16 | UV + ascorbic acid |
| 17 | wound |

Researcher Comment: "They should be related to the last set (Pathway 2) I sent but there are too many to keep togther"

**Result 4**

The number of matching records: 13

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| At1g69490 (transcription factor, NAC family, NAP like putative) | Up | Up | Down | Up | Down | Up |
| At2g01010 (unknown) | Up | | Down | Up | | Up |
| At2g01030 (unknown) | | | Down | Up | | Up |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | ABA (abscisic acid) |
| 2 | brassinosteroid |
| 3 | cold |
| 4 | drought / dehydration / wilt |
| 5 | Pseudomonas fluorescens |
| 6 | sodium chloride |

Researcher Comment: "A small group"

## Result 5

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| At1g09340 (RNA binding protein putative g5bf) | Down | | | Down | | Down | | | Down | | | | | | | | | | Up | | Down | | |
| At1g10460 (germin like oxalate oxidase) | | | Down | | | | | Up | Down | | | | | | | | | | Up | | | | |
| At1g13090 (cytochrome P450) | | Up | | | Up | | Down | | Down | | | | Up | | | Up | | Up | Up | | | Up | Up |
| At1g20340 (DNA-damage-repair/tolerationprotein DRT112 precursor like) | | Up | | Down | | Down | | | Down | Up | | | | | | | | | Up | | Down | | |
| At1g67090 (transcription factor, bel1-like homeodomain 2, Rubisco like?) | | Up | | | | Down | | | Down | Up | Up | Up | | | | | | | Up | Down | Down | | |
| At1g72610 (germin like) | | Up | Down | | | Down | | | Down | | | | | | | | | | Up | | Down | | |
| At1g74470 (geranylgeranyl reductase) | | | | | | | | | Down | Up | | | | | | | | | Up | | | | |
| At2g21660 (RNA binding protein, glycine rich) | | | | | | Down | Down | | Up | | Up | | | | Up | | Down | | Down | | | | |
| At2g33800 (ribosomal protein 30S S5) | | | | | | | | | Down | | | | | | | | | | Up | | | | |
| At2g42520 (RNA helicase, DEAD box putative) | | | | | | | | | Down | | | | | Down | | | | | Up | | | | |
| At2g46820 (unknown) | Down | | | Down | | Down | | | Down | | | | | | | | | | Up | | | | |
| At3g53460 (RNA binding protein cp29) | | | | | | Down | Up | | Down | Up | | | | | | | | | Up | | | | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | ABA (abscisic acid) |
| 2 | Alternaria brassicicola |
| 3 | chitin |
| 4 | cold |
| 5 | cucumber mosaic virus (CMV) |
| 6 | drought / dehydration / wilt |
| 7 | ethylene |
| 8 | hydrogen peroxide |
| 9 | jasmonate (methyl) |
| 10 | light |
| 11 | low oxygen |
| 12 | nitric oxide (NO donor NOR-3) |
| 13 | oilseed rape tobamovirus (ORMV) |
| 14 | phosphate starvation |
| 15 | Phytophthora infestans |
| 16 | potato virus X potexvirus (PVX) |
| 17 | Pseudomonas fluorescens |
| 18 | Pseudomonas syringae pv tomato |
| 19 | salicylic acid |
| 20 | senescence |
| 21 | sodium chloride |
| 22 | turnip mosaic potyvirus (TuMV) |
| 23 | turnip vein clearing tobamovirus (TVCV) |

Researcher Comment: "This is one to rave about Print this one out and show it to Kevan"

**Result 6**

The number of matching records: 15

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| At1g20850 (protease, cysteine, papain-like (C1A-3 family)) | | | | Up | Down | | |
| At3g12610 (DNA damage-repair/toleration protein) | Up | Down | | Up | Down | | |
| At3g53460 (RNA binding protein cp29) | | | Down | Up | Down | Up | Up |
| At4g04460 (protease, aspartic, pepsin-like (A1-4 family)) | | | | Down | Up | | |
| At4g13660 (pinoresinol-lariciresinol reductase, putative) | | | | Down | Up | | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | Alternaria brassicicola |
| 2 | chitin |
| 3 | drought / dehydration / wilt |
| 4 | ethylene |
| 5 | jasmonate (methyl) |
| 6 | light |
| 7 | salicylic acid |

Researcher Comment: "Another small set to consider. Not sure if this is a good one or not. Maybe this is a set where one could make some predictions to test in the lab."

**Result 7**

The number of matching records: 19

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| At1g09210 (calreticulin) | Down | | | Down | | Down | Down | | |
| At1g21550 (calcium-binding protein) | | | | | | | Down | Up | Up |
| At1g62480 (calcium-binding protein, vacuolar, putative) | Up | Up | Down | Up | | Up | Up | | |
| At3g57530 (calcium-dependent protein kinase) | | | | | Up | Down | Down | | Up |
| At4g04720 (calcium-dependent protein kinase) | | | | | | Up | Up | | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | Alternaria brassicicola |
| 2 | BTH |
| 3 | cucumber mosaic virus (CMV) |
| 4 | ethylene |
| 5 | hydrogen peroxide |
| 6 | jasmonate (methyl) |
| 7 | salicylic acid |
| 8 | sodium nitropruside (an NO donor) |
| 9 | Yariv phenylglycoside (beta-D-Glc)3 |

Researcher Comment: "More AGI for pathways"

**Result 8**

The number of matching records: 16

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| At1g09210 (calreticulin) | Down | | | Down | | Down | Down | |
| At1g62480 (calcium-binding protein, vacuolar, putative) | Up | Up | Down | Up | | Up | Up | |
| At3g57530 (calcium-dependent protein kinase) | | | | | Up | Down | Down | Up |
| At4g04720 (calcium-dependent protein kinase) | | | | | | Up | Up | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | Alternaria brassicicola |
| 2 | BTH |
| 3 | cucumber mosaic virus (CMV) |
| 4 | ethylene |
| 5 | hydrogen peroxide |
| 6 | jasmonate (methyl) |
| 7 | salicylic acid |
| 8 | Yariv phenylglycoside (beta-D-Glc)3 |

Researcher Comment:

## Result 9

The number of matching records: 35

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| At1g09210 (calreticulin) | Down | | | | Down | | Down | | | | | Down | |
| At1g20850 (protease, cysteine, papain-like (C1A-3 family)) | | | | | Up | | Down | | | | | | |
| At2g21660 (RNA binding protein, glycine rich) | | | | Down | Down | | Up | | Up | Up | Down | Down | |
| At3g12610 (DNA damage-repair/toleration protein) | Up | Down | | | Up | | Down | | | | | | |
| At3g53460 (RNA binding protein cp29) | | | | Down | Up | | Down | Up | | | | Up | |
| At3g57530 (calcium-dependent protein kinase) | | | | | | Up | Down | | | | | Down | Up |
| At4g04460 (protease, aspartic, pepsin-like (A1-4 family)) | | | | | Down | | Up | | | | | | |
| At4g04720 (calcium-dependent protein kinase) | | | | | | | Up | | | | | Up | |
| At4g13660 (pinoresinol-lariciresinol reductase, putative) | | | | | Down | | Up | | | | | | |
| At4g30170 (peroxidase) | | | Up | | Down | | Up | | | | | | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | Alternaria brassicicola |
| 2 | chitin |
| 3 | cold |
| 4 | drought / dehydration / wilt |
| 5 | ethylene |
| 6 | hydrogen peroxide |
| 7 | jasmonate (methyl) |
| 8 | light |
| 9 | low oxygen |
| 10 | Phytophthora infestans |
| 11 | Pseudomonas fluorescens |
| 12 | salicylic acid |
| 13 | Yariv phenylglycoside (beta-D-Glc)3 |

Researcher Comment: "This list contains some existing numbers but is slightly different and perhaps better than some earlier ones."

## Result 10

The number of matching records: 37

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| At2g40000 (nematode resistance protein putative / Hs1pro-1 like) | | | Up | | | Up | Up | Up | | Down | | | Up | Up |
| At3g50930 (ATPase family AAA-type) | | | | | | Up | Up | | Up | | | Up | | Up |
| At4g12720 (growth factor-like) | | | | | | Up | Up | | | | | | | Up |
| At4g20780 (calcium-binding protein putative) | | Up | | | | Up | Up | | | | | | | Up |
| At4g39640 (glutamyltransferase, gamma, putative) | | | | | | Up | Up | Up | | | | | | Up |
| At5g25930 (leucine-rich repeat family protein / protein kinase family RLK5 like) | | | | | | Up | Up | | | | | | | Up |
| At5g57560 (xyloglucan endotransglycosylase related) | Up | Up | Up | Down | Up | Up | Up | | Up | | Up | | Up | Up |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | ABA (abscisic acid) |
| 2 | brassinosteroid |
| 3 | cold |
| 4 | cucumber mosaic virus (CMV) |
| 5 | drought / dehydration / wilt |
| 6 | flagellin 22 (flg22) |
| 7 | gamma irradiation |
| 8 | hydrogen peroxide |
| 9 | IAA |
| 10 | low oxygen |
| 11 | malondialdehyde (MDA) |
| 12 | methanol |
| 13 | wound |
| 14 | Yariv phenylglycoside (beta-D-Glc)3 |

Researcher Comment: "Another set of genes possibly in the same pathway. I have just put some information into the database on genes regulated by gamma irradiation and noticed that some were also regulated by Yariv phenylglycosides. I therefore had a closer look at everything that was regulated by both of those treatments and noticed that the following subset of genes are also all up-regulated by flagellin. These must therefore be likely candidates for being involved in the same pathway. (possibly the last gene i.e. At5g57560 may be regulated by too many different treatments and may therefore not be that close in terms of signalling distance to the other genes)"

## Result 11

The number of matching records: 15

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| At1g19670 (chlorophyllase 1, coronatine-induced protein 1) | | Down | Down | Down | | Up | |
| At2g21210 (auxin-induced putative) | | | Down | Down | | | Down |
| At2g46690 (auxin-responsive family) | | | Down | Down | | | Down |
| At4g38840 (auxin induced SAUR-13 like) | Down | | Down | Down | Up | | Down |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | cold |
| 2 | cucumber mosaic virus (CMV) |
| 3 | flagellin 22 (flg22) |
| 4 | gamma irradiation |
| 5 | IAA |
| 6 | jasmonate (methyl) |
| 7 | wound |

Researcher Comment: "This small group looks interesting when put into the pathway search."

## Result 12

The number of matching records: 133

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| At1g09340 (RNA binding protein putative g5bf) | Down | Down | | Down | | | Down | | | | | | | Up | | Down | | |
| At1g16850 (unknown) | Up | Up | | Up | | | | | | | | | | | | Up | | |
| At1g43160 (AP2 domain containing protein RAP2.6, ERF family) | Up | Up | | Up | | | | | | | | | | | | Up | | |
| At1g49450 (transposon like protein, En/Spm-like) | Up | Up | | Up | | | | | | | | | | | | Up | | Down |
| At1g78070 (WD-40 repeat family) | Up | Up | Down | Up | | | | | | | | | | | | Up | | |
| At2g01010 (unknown) | Up | Down | | Up | | | | | | | | | | | | Up | | |
| At2g15970 (low temperature-regulated membrane protein putative) | Up | Up | | Up | | | | Up | | | | | | | | Up | | |
| At2g23120 (unknown) | Up | Up | | Up | | | | | | | Down | | | | | Up | | |
| At2g26980 (SNF1-related kinase) | Up | Up | | Up | | | | | | | | | | | | Up | Up | |
| At2g30360 (protein kinase 11 CBL-interacting (CIPK11) / SNF1-related kinase) | Up | Up | | Up | | Up | | Up | | | | | | | | Up | Up | |
| At2g33380 (calcium-binding EF-hand protein) | Up | Up | | Up | | | | Up | | | | | | | | Up | Up | |
| At2g46270 (transcription factor, bZIP G box binding factor GBF3) | Up | Up | | Up | Up | | Up | | | | | | | Up | Up | Up | | |
| At3g61890 (transcription factor, homeobox leucine zipper) | Up | Up | Down | Up | | Down | | Up | | | | | | | | Up | Up | Down |
| At4g26080 (protein phosphatase 2C ABI1 / PP2C ABI1 / abscisic acid-insensitive 1 (ABI1)) | Up | Up | | Up | Up | | | | | | | | | | | Up | | Down |
| At4g27410 (NAC family) | Up | Up | | Up | | | | Up | | | | | | | | Up | | |
| At4g34000 (bZIP abscisic acid responsive elements-binding factor) | Up | Up | Down | Up | Up | | | Up | | | | | | | | Up | | |
| At5g15960 (kin1) | Up | Up | | Up | | | | Up | | | | Down | Up | | | Up | Up | |
| At5g15970 (kin2 (cold regulated cor6.6)) | Up | Up | | Up | | | | Up | | | | | | | | Up | | |
| At5g17300 (transcription factor, Myb) | Up | Up | | Up | | | | | | | Down | | | | | Up | | |
| At5g17460 (unknown) | Up | Up | | Up | | | | | | | | | | | | Up | | |
| At5g17465 (unknown) | Up | Up | | Up | | | | | | | | | | | | Up | | |
| At5g52310 (responsive to dessication RD29A, cor78, lti140, LTI78) | Up | Up | | Up | | | Down | Up | Up | Up | | | | Both | | Up | Up | |
| At5g53140 (protein phosphatase 2C putative) | Down | Down | | Down | Up | | | | | | | | | | | Down | | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | ABA (abscisic acid) |
| 2 | cold |
| 3 | cucumber mosaic virus (CMV) |
| 4 | drought / dehydration / wilt |
| 5 | ethylene |
| 6 | gamma irradiation |
| 7 | jasmonate (methyl) |
| 8 | mannitol (0.2M) |
| 9 | mannitol (0.4M) |
| 10 | mannitol (0.6M) |
| 11 | ozone |
| 12 | Peronospora parasitica |
| 13 | Pseudomonas syringae pv tomato |
| 14 | salicylic acid |
| 15 | senescence |
| 16 | sodium chloride |
| 17 | wound |
| 18 | Yariv phenylglycoside (beta-D-Glc)3 |

Researcher Comment: "I have had a closer look at genes involved in ABA-associated signalling. Lots of them. I think there are sub-groups in this list eg look at responses to treatment 8"

## Result 13

The number of matching records: 160

| Gene Name / Treatment | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| At1g01140 (SNF1-related kinase) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At1g01720 (no apical meristem / transcription activator, NAC domain, ATAF1) | Up | | Down | Up | | | | | | | | | | | | | Up | Up | |
| At1g07040 (unknown) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At1g13990 (unknown) | Up | | | Up | | | | | | | | | Up | | | | Up | | |
| At1g32640 (protein kinase; RD22BP1; transcription factor, bHLH putative) | Up | Up | | Up | Down | | Up | Up | | | | | | | | | Up | | |
| At1g45249 (ABA-responsive cis-acting element) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At1g52980 (GTP-binding protein putative) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At1g61890 (MATE efflux family) | Up | | Down | Up | | | | | | | | | | | | | Up | | Up |
| At1g66760 (MATE efflux family protein, putative) | Up | | Down | Up | | | | | | | | | | | | | Up | | |
| At1g68440 (unknown) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At1g72800 (unknown, nucleolin like? (NuM1-related)) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At1g75500 (nodulin MtN21 family) | Down | | | Down | Down | | | | Down | | | | | | | | Down | | |
| At2g02710 (receptor-like serine/threonine protein kinase) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At2g05540 (glycine-rich protein putative) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At2g06050 (12-oxophytodienoate reductase) | Up | | Down | Up | | Up | | Up | | Up | | | | | | | Up | | |
| At2g30550 (lipase putative) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At2g42890 (Mei2) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At2g46680 (transcription factor, homeobox leucine zipper) | Up | | | Up | | | | | | | | | | | | Up | Up | Up | Down |
| At2g47190 (transcription factor, Myb) | Up | | | Up | | Up | | | | | | | | | | | Up | | |
| At3g02480 (ABA-responsive protein-related) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g17000 (ubiquitin conjugating enzyme like) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g19290 (transcription factor, bZIP (AREB2, ABF4)) | Up | | | Up | | | | | | | | | | | | | Up | | Down |
| At3g20300 (unknown) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g22830 (heat shock transcription factor-like protein) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g29575 (unknown) | Up | | Down | Up | | | | | | | | | | | | | Up | | |
| At3g53710 (zinc finger-containing protein ARF GAP-like ZIGA2) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At3g55530 (zinc finger (C3HC4-type RING finger) family) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g59930 (unknown) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At3g62030 (peptidyl-prolyl cis-trans isomerase, chloroplast / cyclophilin / rotamase) | Down | | | Down | | | | | | | | | | Down | | | Down | | |
| At4g01020 (pentatricopeptide (PPR) repeat-containing protein) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At4g20830 (reticuline oxidase homolog) | Up | | | Up | | | | | Up | | Up | | Up | | | | Up | | |
| At4g23050 (protein kinase, serine/threonine putative) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At4g37510 (ribonuclease III family) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At5g01520 (zinc finger (C3HC4-type RING-HCa)) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At5g09440 (phosphate-responsive protein, putative) | Up | Up | | Up | | | | | | | | | | | Up | | Up | | |
| At5g11420 (unknown) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At5g23060 (unknown) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At5g25460 (unknown) | Down | | | Down | | | | | | | | | | | | | Down | | |
| At5g39610 (NAC, NAM (no apical meristem) like) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At5g42010 (WD-40 repeat protein family) | Down | | | Down | | | | | | | | | | | | | Both | | |
| At5g48180 (jasmonate inducible (myrosinase binding)) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At5g52300 (responsive to dessication RD29B) | Up | | | Up | | | | | | | | | | | | | Up | | |
| At5g61820 (MtN19 like) | Up | | | Up | | | Up | | | | | Up | | | | | Up | | |

Treatment Legend

| Number | Treatment |
|---|---|
| 1 | ABA (abscisic acid) |
| 2 | Alternaria brassicicola |
| 3 | cucumber mosaic virus (CMV) |
| 4 | drought / dehydration / wilt |
| 5 | ethylene |
| 6 | flagellin 22 (flg22) |
| 7 | hydrogen peroxide |
| 8 | jasmonate (methyl) |
| 9 | low oxygen |
| 10 | malondialdehyde (MDA) |
| 11 | methanol |
| 12 | nitric oxide (NO donor NOR-3) |
| 13 | ozone |
| 14 | phosphate starvation |
| 15 | salicylic acid |
| 16 | senescence |
| 17 | sodium chloride |
| 18 | wound |
| 19 | Yariv phenylglycoside (beta-D-Glc)3 |

Researcher Comment: "May be the genes that are down-regulated in this last are linked whilst those that are up-regulated are linked together. The genes down-regulated by treatment 3 also look linked."

# Appendix VI – Drastic Insight User Guide

Welcome to **DRASTIC INSIGHTS** Help Pages

Drastic Insights is a suite of web tools that enable you to mine the gene expression data in the DRASTIC database.

The tools are listed below. Click on the links to view how to use them:
General Database Search
Drastic Statistics

The following tools work with only the *Arabidopsis thaliana* data:
AGI Search
Chromosome Mapping
GO Functional Categorisation
Venn Diagrams
Pathway Finder
Roadmap
Unique Genes

**DRASTIC Database**
The Drastic Database is a relational database developed by SCRI (**S**cottish **C**rop **R**esearch **I**nstitute) and UAD ( **U**niversity of **A**bertay **D**undee). It incorporates data from published refereed papers containing plant molecular responses, microarrays, Northerns and ESTs regulated by a wide range of chemicals, environmental stresses, pathogens and elicitors referred to here as treatments. The effects of each treatment on the up- or down- regulation of gene expression have been used to populate the database. At the time of writing there are more than 16,000 gene accessions relating to over 295 different treatments in over 60 plant species taken from over 400 references.
This data can be searched and manipulated using the web tools provided by the Drastic Insight project available from www.drastic.org.uk. The Drastic database is updated with new data on a daily basis.

**Drastic Statistics**
Drastic Statistics provides an up-to-date list of statistics for the database including the total number of reactions, species and treatments. It also provides a breakdown of both reactions per species and reactions per treatment. You can order the lists alphabetically or numerically. There are several ways to view the data:

Sort Reactions
In depth view of Reactions
View Individual Reaction Results
View Journal References

Order Stats Numerically Order Stats Alphabetically

**Total number of Reactions:** 13824
**Total number of Species:** 67
**Total number of Treatments:** 287

| Number of Reactions per Species: | |
|---|---|
| Arabidopsis thaliana | 10510 |
| Oryza sativa | 683 |
| Lycopersicon esculentum | 641 |
| Hordeum vulgare | 473 |
| Nicotiana tabacum | 363 |
| Solanum tuberosum | 297 |
| Capsicum annuum | 113 |
| Medicago truncatula | 103 |
| Glycine max | 69 |
| Pisum sativum | 49 |
| Fragaria ananassa | 46 |

| Number of Reactions per Treatment: | |
|---|---|
| sodium chloride | 1050 |
| drought / dehydration / wilt | 1037 |
| cold | 913 |
| cucumber mosaic virus (CMV) | 635 |
| wound | 609 |
| Pseudomonas syringae pv tomato | 561 |
| jasmonate (methyl) | 513 |
| salicylic acid | 511 |
| ABA (abscisic acid) | 496 |
| Yariv phenylglycoside (beta-D-Glc)3 | 475 |
| ethylene | 419 |

**Sort Function**

The Drastic Statistic page provides an easy way of browsing the data. The number of reactions represents the number of records held in the Drastic database for each category.

You can click on the 'Order Stats Numerically' button if you want to see the table(s) numerically ordered. This option sorts the data according to the number of reactions per category in ascending order.

You can click on the 'Order Stats Alphabetically' button which will sort the data into alphabetically ascending order.

**View Reactions by Species for a selected Treatment**

To gain a more in depth view on a particular treatment, click on the treatment that you are interested in. This will produce a new table of data providing statistics on the number of reactions by species for the treatment you selected. The reactions can be sorted Numerically or Alphabetically. This can be mined further by clicking on a species from the new list - see View Individual Reactions

**View Reactions by Treatment for a selected Species**

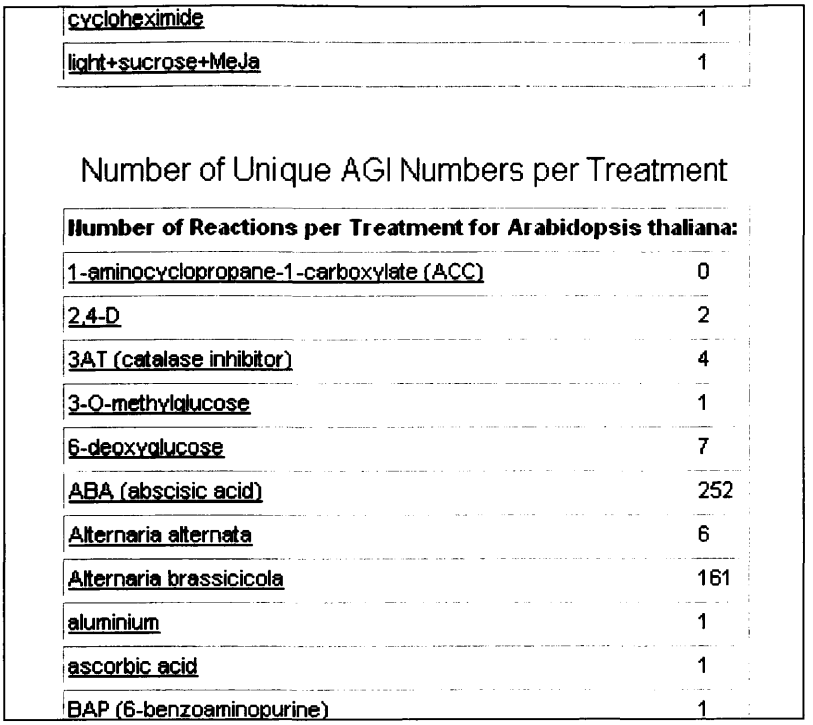
To view further information on a particular species, click on the species that you are interested in. This will navigate you to a new page which will provide statistics on the number of reactions by treatment for the species you selected as shown below:

Order Stats Numerically Order Stats Alphabetically

**Total number of Reactions:** 683
**Total number of Treatments:** 40

| Number of Reactions per Treatment for Oryza sativa: | |
|---|---|
| Magnaporthe grisea | 166 |
| sodium chloride | 120 |
| drought / dehydration / wilt | 58 |
| ABA (abscisic acid) | 48 |
| jasmonic acid | 47 |
| cold | 42 |
| BTH | 24 |
| elicitor, cell wall from M grisea | 20 |
| wound | 19 |

For *Arabidopsis thaliana* only, in addition to the number of reactions per treatment, there is an additional table which displays the number of unique AGI numbers per treatment. This table can be found by scrolling down the page past the first table as shown below:



| | |
|---|---|
| cycloheximide | 1 |
| light+sucrose+MeJa | 1 |

Number of Unique AGI Numbers per Treatment

| Number of Reactions per Treatment for Arabidopsis thaliana: | |
|---|---|
| 1-aminocyclopropane-1-carboxylate (ACC) | 0 |
| 2,4-D | 2 |
| 3AT (catalase inhibitor) | 4 |
| 3-O-methylglucose | 1 |
| 6-deoxyglucose | 7 |
| ABA (abscisic acid) | 252 |
| Alternaria alternata | 6 |
| Alternaria brassicicola | 161 |
| aluminium | 1 |
| ascorbic acid | 1 |
| BAP (6-benzoaminopurine) | 1 |

The reactions can be sorted Numerically or Alphabetically or mined further by clicking on a treatment from the new list - see View Individual Reactions

**View Individual Reaction Results**

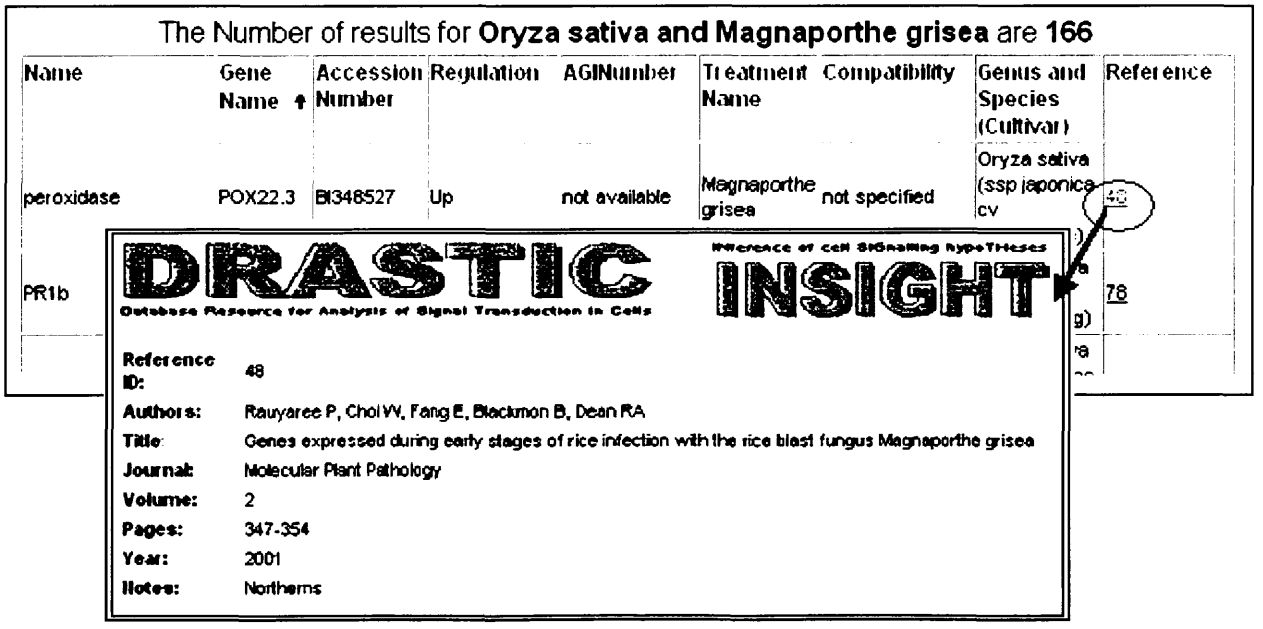
This screen enables you to view all the individual reactions held on drastic for the treatment and species that you selected from the Drastic Statistic pages as shown in the diagram below:
The number of results and the species/treatment chosen are displayed at the top of the page.
Each table of records can be sorted alphabetically (ascending or descending) by clicking on the column heading that you want to sort the data by. The first time you click, it will sort the contents in ascending order. The second time you click it will sort the table contents in descending order. This is indicated by a black arrow that will appear in the sorted column heading.
To view the full journal reference data, click on the Reference ID number of the record you are interested in. This will open a new window which displays the full reference information including the journal it was published in.

The Number of results for **Arabidopsis thaliana and sodium chloride** are **786**

| Name | Gene Name | Accession Number | Regulation | AGINumber | Treatment Name | Compatibility | Genus and Species (Cultivar) | Reference |
|---|---|---|---|---|---|---|---|---|
| 12-oxophytodienoate reductase | AtOPR3 | AV824251; AV785462; AF410322 | Up | At2g06050 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 107 |
| 12-oxophytodienoate reductase | AtOPR3 | AV824251; AV785462 | Up | At2g06050 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 334 |
| 1-aminocyclopropane-1-carboxylate (ACC) oxidase putative | not known | At2g19590 | Down | At2g19590 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 472 |
| 1-aminocyclopropane-1-carboxylate (ACC) oxidase putative | not known | At1g12010 | Down | At1g12010 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 472 |
| 1-aminocyclopropane-1-carboxylate (ACC) synthase | AtACS2 | not available | Up | not available | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 404 |
| 1-aminocyclopropane-1-carboxylate (ACC) synthase like | not known | At4g26200 | Down | At4g26200 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 472 |
| 2-isopropylmalate synthase-like; homocitrate synthase like | not available | AV821148 | Down | At5g23010 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 334 |
| 2-isopropylmalate synthase-like; homocitrate synthase like | not available | AV821148 | Down | At5g23010 | sodium chloride | non pathogen | Arabidopsis thaliana (cv Columbia) | 107 |
| 2-oxoglutarate dehydrogenase, E1 | not available | BE844998 | Up | At3g55410 | sodium chloride | non pathogen | Arabidopsis thaliana (cv | 15 |

## View Journal References

To view the full journal reference data, click on the Reference ID number of the record you are interested in. This will open a new window which displays the full reference information including the journal it was published in as shown below:

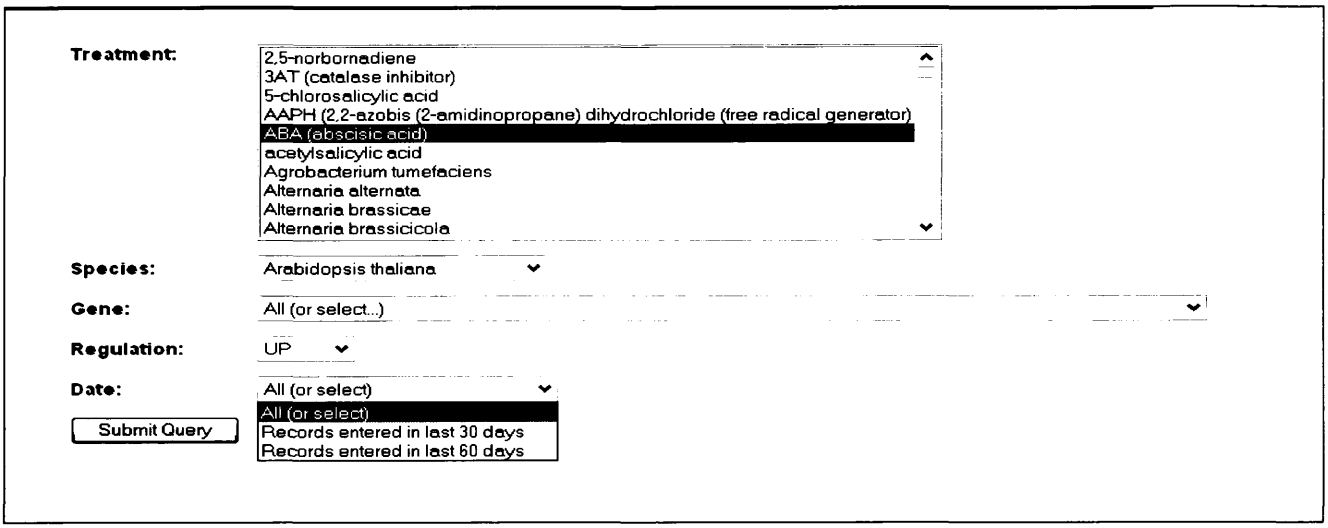


## Database Search

This provides a basic query function for the database. You can select the following parameters: Treatments, Species, Gene, Regulation and Date. The search returns the results in tabular format which can be sorted, providing links to the references.
Database Search Page
Database Search Results

## General Search Page

To select a subset of reactions to view from the Drastic Database, select your search criteria from the Search Page as shown below: Click the 'Submit Search' button to view your results.




If you would like to use the AND operator with treatments, use the Venn Diagram tool.

## General Search Results

This screen enables you to view all the individual reactions held on the drastic database according to your search criteria as shown in the diagram below:
The number of results are displayed at the top of the page.
The whole table of records can be sorted alphabetically (ascending or descending) by clicking on the column heading that you want to sort the data by. The first time you click, it will sort the contents in ascending order. The second time you click it will sort the table contents in descending order. This is indicated by a black arrow that will appear in the sorted column heading.
To view the full journal reference data, click on the Reference ID number of the record you are interested in. This will open a new window which displays the full reference information including the journal it was published in.

The number of matching records: 207

| Name and (Gene Name) ↓ | Accession Number | AGI Number | Regulation | Treatment Name | Compatibility | Species | Ref ID |
|---|---|---|---|---|---|---|---|
| 2-oxoisovalerate dehydrogenase, putative (not known) | At5g09300 | At5g09300 | Up | cucumber mosaic virus (CMV) | incompatible | Arabidopsis thaliana C24 | 485 |
| 4-coumarate:CoA ligase (At4-CL1) | AY133582 | At1g51680 | Up | ozone | non pathogen | Arabidopsis thaliana cv Columbia | 486 |
| 4-coumarate:CoA ligase (At4-CL2) | BX003958 | At3g21230 | Up | ozone | non pathogen | Arabidopsis thaliana cv Columbia | 486 |
| amino acid permease 3 (not available) | At1g77380 | At1g77380 | Up | cucumber mosaic virus (CMV) | incompatible | Arabidopsis thaliana C24 | 485 |
| ammonium transporter (AtAMT1.1) | At4g13510 | At4g13510 | Up | cucumber mosaic virus (CMV) | incompatible | Arabidopsis thaliana C24 | 485 |
| arabinogalactan protein fasciclin-like (AtFLA12) | At5g60490 | At5g60490 | Up | cucumber mosaic virus (CMV) | incompatible | Arabidopsis thaliana C24 | 485 |
| ascorbate oxidase putative (not known) | At5g21100 | At5g21100 | Up | cucumber mosaic virus (CMV) | incompatible | Arabidopsis thaliana C24 | 485 |

**Accession Number Search**

This provides a query function specifically for the Accession numbers in the Drastic database. You can select the following parameters: Accession number, Treatment Regulation and Date. The search returns the results in tabular format which can be sorted, providing links to the references.

Accession Number Search Page

Accession Number Search Results

**Accession Number Search Page**

To select a subset of reactions to view from the Drastic Database, select your search criteria from the Accession Number Search Page

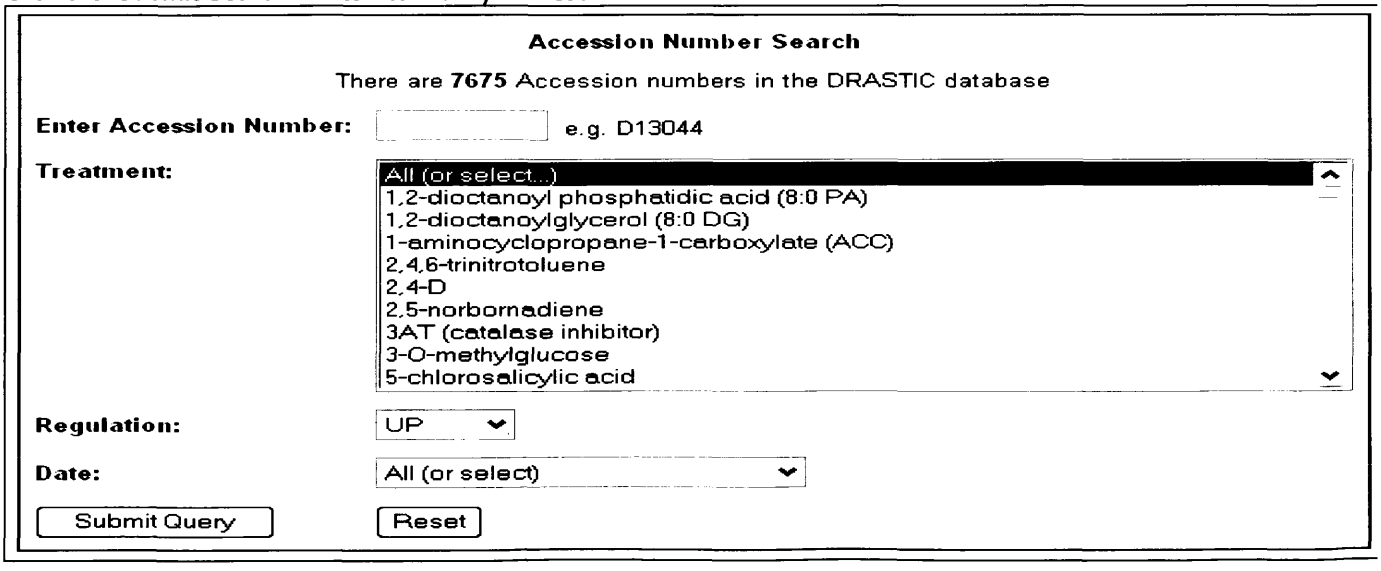
**Accession Number:** This box can be left blank to include all Accession numbers or you can input one Accession number to the search.

**Treatment:** You can select one treatment from the list or choose them all.

**Regulation:** You can choose to include only up-regulation reactions or down-regulation reactions or both.

**Date:** You can view all reactions inputted to the database, or select to only view those inputted in the last 60 or 30 days.

Click the 'Submit Search' button to view your results



**View Accession Number Results**
This screen enables you to view all the individual reactions held on the drastic database according to your search criteria as shown in the diagram below:
The number of results are displayed at the top of the page.
The whole table of records can be sorted alphabetically (ascending or descending) by clicking on the column heading that you want to sort the data by. The first time you click, it will sort the contents in ascending order. The second time you click it will sort the table contents in descending order. This is indicated by a black arrow that will appear in the sorted column heading.
To view the full journal reference data, click on the Reference ID number of the record you are interested in. This will open a new window which displays the full reference information including the journal it was published in.

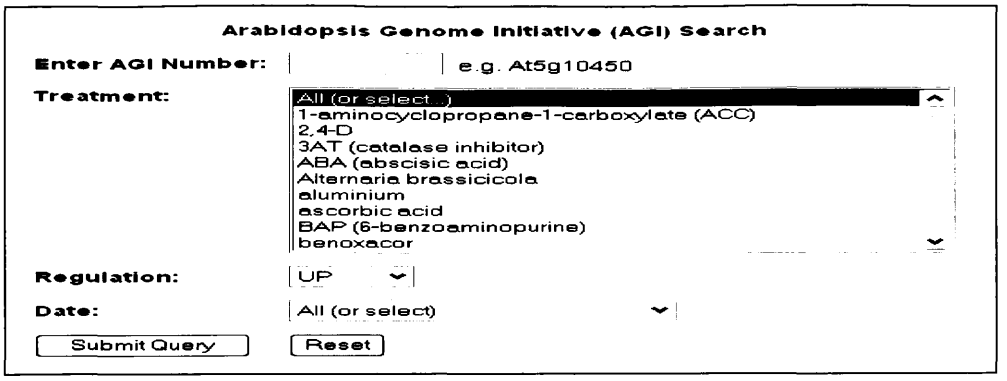
The number of matching records: 23

| Name and (Gene Name) | Accession Number | AGI Number | Regulation | Treatment Name | Compatibility | Species | Ref ID |
|---|---|---|---|---|---|---|---|
| responsive to dessication RD29A, cor78, lti140, LTI78 (Atrd29A) | D13044 | At5g52310 | Up | drought / dehydration / wilt | non pathogen | Arabidopsis thaliana cv Columbia | 279 |
| responsive to dessication RD29A, cor78, lti140, LTI78 (Atrd29A) | D13044 | At5g52310 | Up | sodium chloride | non pathogen | Arabidopsis thaliana C24 expressing RDA29A-LUC transgene | 255 |
| responsive to dessication RD29A, cor78, lti140, LTI78 (Atrd29A) | D13044 | At5g52310 | Up | ABA (abscisic acid) | non pathogen | Arabidopsis thaliana C24 expressing RDA29A-LUC transgene | 255 |
| responsive to dessication RD29A, cor78, lti140, LTI78 (Atrd29A) | D13044 | At5g52310 | Up | sodium chloride | non pathogen | Arabidopsis thaliana cv Columbia | 214 |
| responsive to dessication RD29A, | D13044 | At5g52310 | Up | cold | non pathogen | Arabidopsis thaliana cv | 132 |

## ***Arabidopsis thaliana*** **Only Tools**

**AGI Search**
This provides a query function specifically for AGI numbers in the Drastic database. You can select the following parameters: AGI number, Treatment Regulation and Date. The search returns the results in tabular format which can be sorted, providing links to the references.
AGI Search Page
AGI Search Results

**AGI Search Page**
To select a subset of reactions to view from the Drastic Database, select your search criteria from the AGI Search Page
**AGI Number:** This box can be left blank to include all AGI numbers or you can input one AGI number to the search.
**Treatment:** You can select one treatment from the list or choose them all.
**Regulation:** You can choose to include only up-regulation reactions or down-regulation reactions or both.
**Date:** You can view all reactions inputted to the database, or select to only view those inputted in the last 60 or 30 days.
Click the 'Submit Search' button to view your results




**View AGI Results**

This screen enables you to view all the individual reactions held on the drastic database according to your search criteria as shown in the diagram below:

The number of results are displayed at the top of the page.

The whole table of records can be sorted alphabetically (ascending or descending) by clicking on the column heading that you want to sort the data by. The first time you click, it will sort the contents in ascending order. The second time you click it will sort the table contents in descending order. This is indicated by a black arrow that will appear in the sorted column heading.

To view the full journal reference data, click on the Reference ID number of the record you are interested in. This will open a new window which displays the full reference information including the journal it was published in.

The number of matching records: 10

| Name and (Gene Name) | Accession Number | AGI Number | Regulation | Treatment Name ↓ | Compatibility | Species | Ref ID |
|---|---|---|---|---|---|---|---|
| 14-3-3 (RCI1B homolog ATF1 (GF14 lambda)) | T75872 | At5g10450 | Up | Alternaria brassicicola | compatible | Arabidopsis thaliana cv Columbia | 2 |
| 14-3-3 (RCI1B homolog ATF1 (GF14 lambda)) | H36693 | At5g10450 | Up | Alternaria brassicicola | compatible | Arabidopsis thaliana cv Columbia | 2 |
| 14-3-3 (RCI1B homolog ATF1 (GF14 lambda)) | X74141 | At5g10450 | Up | cold | non pathogen | Arabidopsis thaliana cv Columbia | 370 |
| 14-3-3 (RCI1B | | | | | | Arabidopsis | |

**TAIR AGI Search**

The Drastic database has a large number of records for reactions on *Arabidopsis thaliana* . The majority of these records include the AGI (**A**rabidopsis **G**enome **I**nitiative) number for the gene involved.

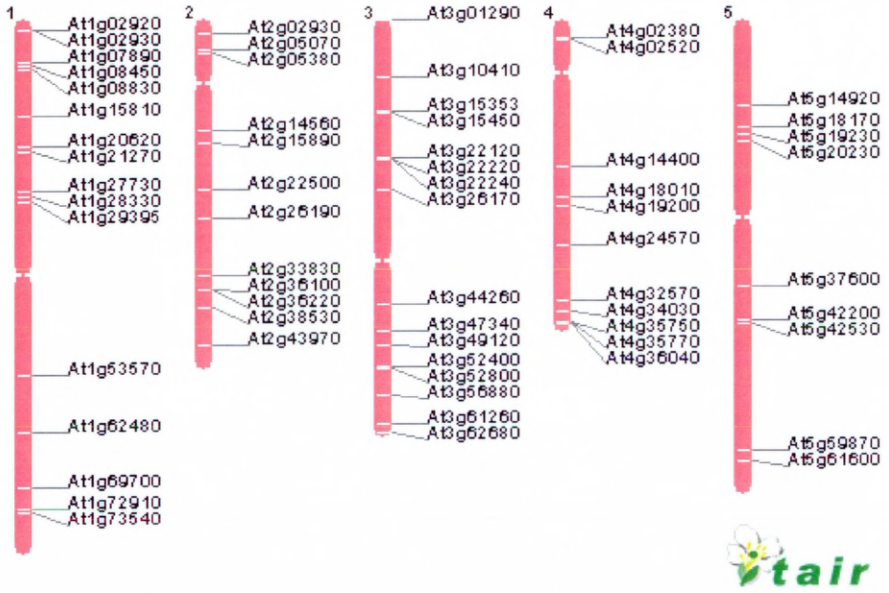
The TAIR (**T**he **A**rabidopsis **I**nformation **R**esource) site has a number of tools specifically designed to analysis AGI data. We enable you to make use of these tools by selecting a subset of records from Drastic, and formatting the data so that it can be used with the TAIR tools. The tools that are available to use are:

**Chromosome Mapping** and
**Functional Categorisation**

How to use the **Chromosome Map Tool**

This search enables you to produce an image of the position of your selected genes on TAIR's Arabidopsis Chromosome Map.

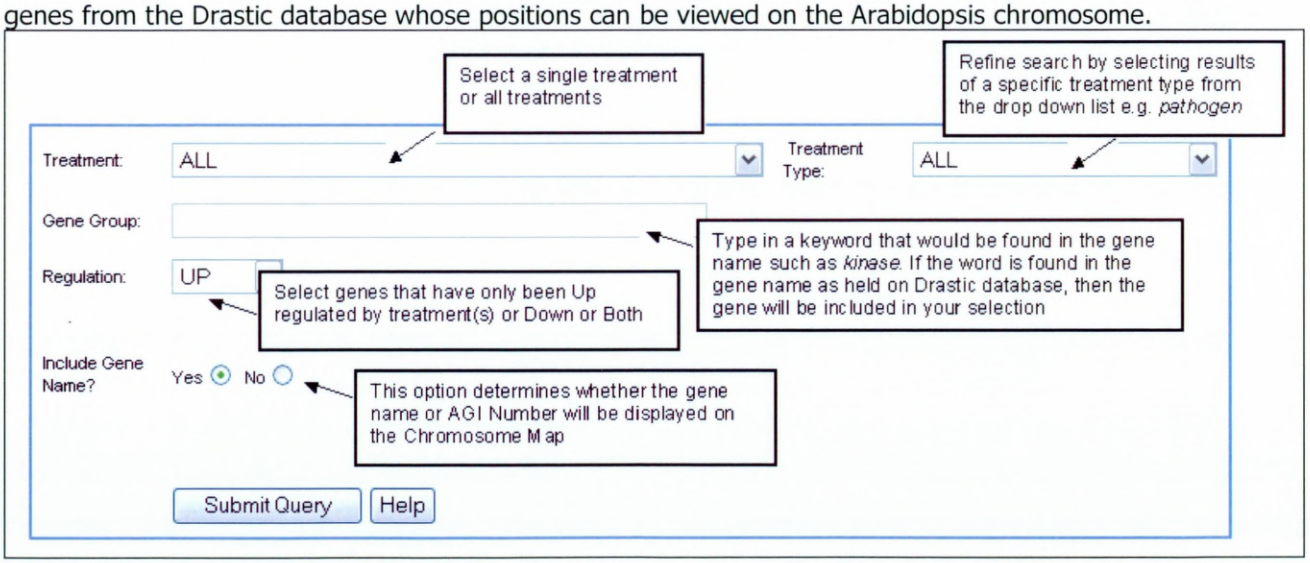
Step 1: Use the search options to select genes. Click 'Submit Query
Step 2: A list of the genes resulting from your search will be displayed. Select 'TAIR Chromosome Map' button.
Step 3: A pop up window should appear for the TAIR (www.arabidopsis.org) site. Right Click in list box and select paste. Click 'Display on Chromosome' button to view Chromosome map of your selected genes. A sample output is shown below.




**Search Page**

Step 1: Use search option to select genes. Click 'Submit Query'

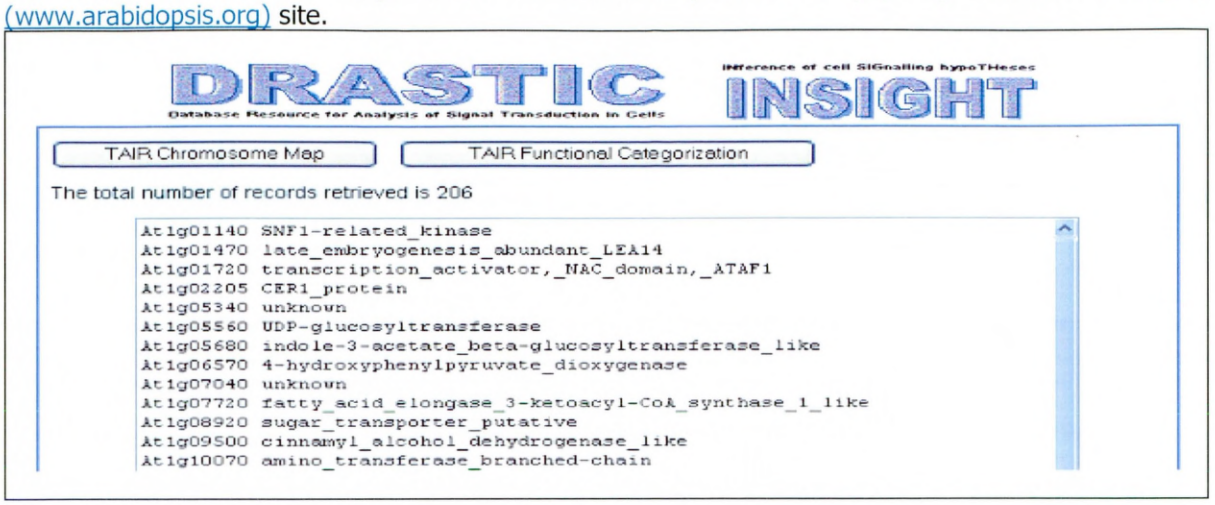
The screen shot below details each search option. This enables you to select a treatment or treatment type or gene group (such as kinases) and regulation. The search produces a group of genes from the Drastic database whose positions can be viewed on the Arabidopsis chromosome.




**Search Results**

Step 2: Click 'TAIR Chromosome Map' button

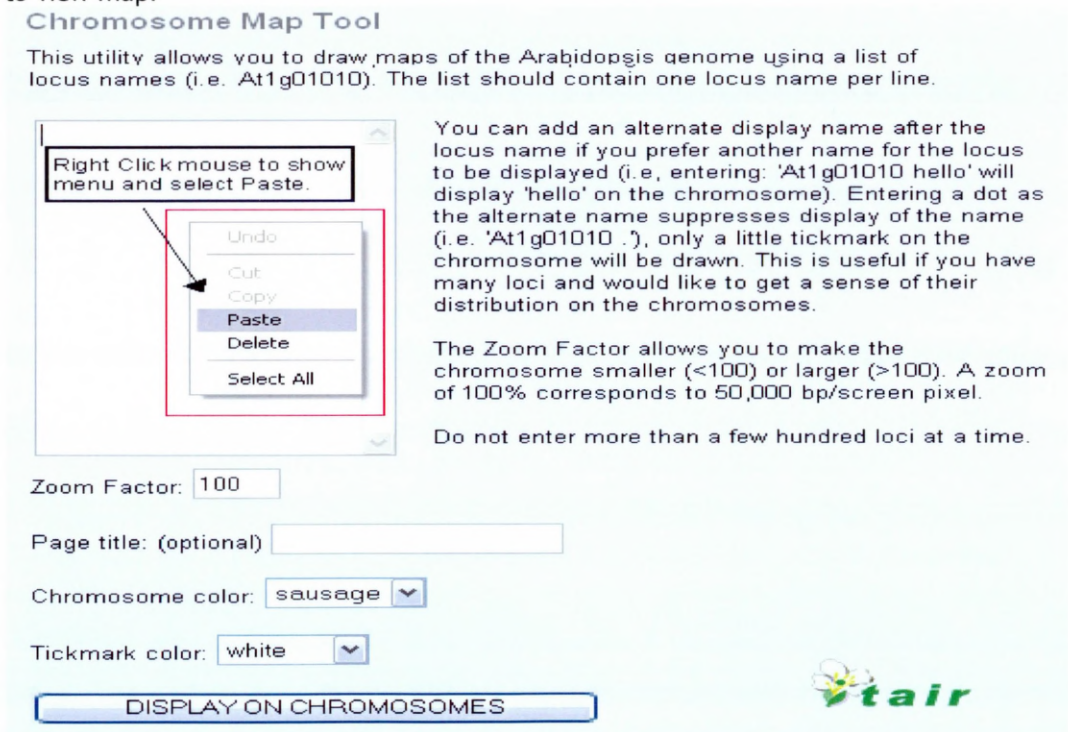
This is your selected list of genes from DRASTIC based from your selected search options. Click on the 'TAIR Chromosome Map' button which will produce a pop up window to TAIR (www.arabidopsis.org) site.




**View Chromosome Map**

Step 3: Input results to TAIR Chromosome Mapping tool

A pop up window will have appeared looking like the one below. Place the mouse pointer in the list box as shown in image. Right click on the mouse and select Paste from the menu. This will copy the results from the search into the TAIR tool to be processed. Click the 'Display on Chromosomes' button to view map.



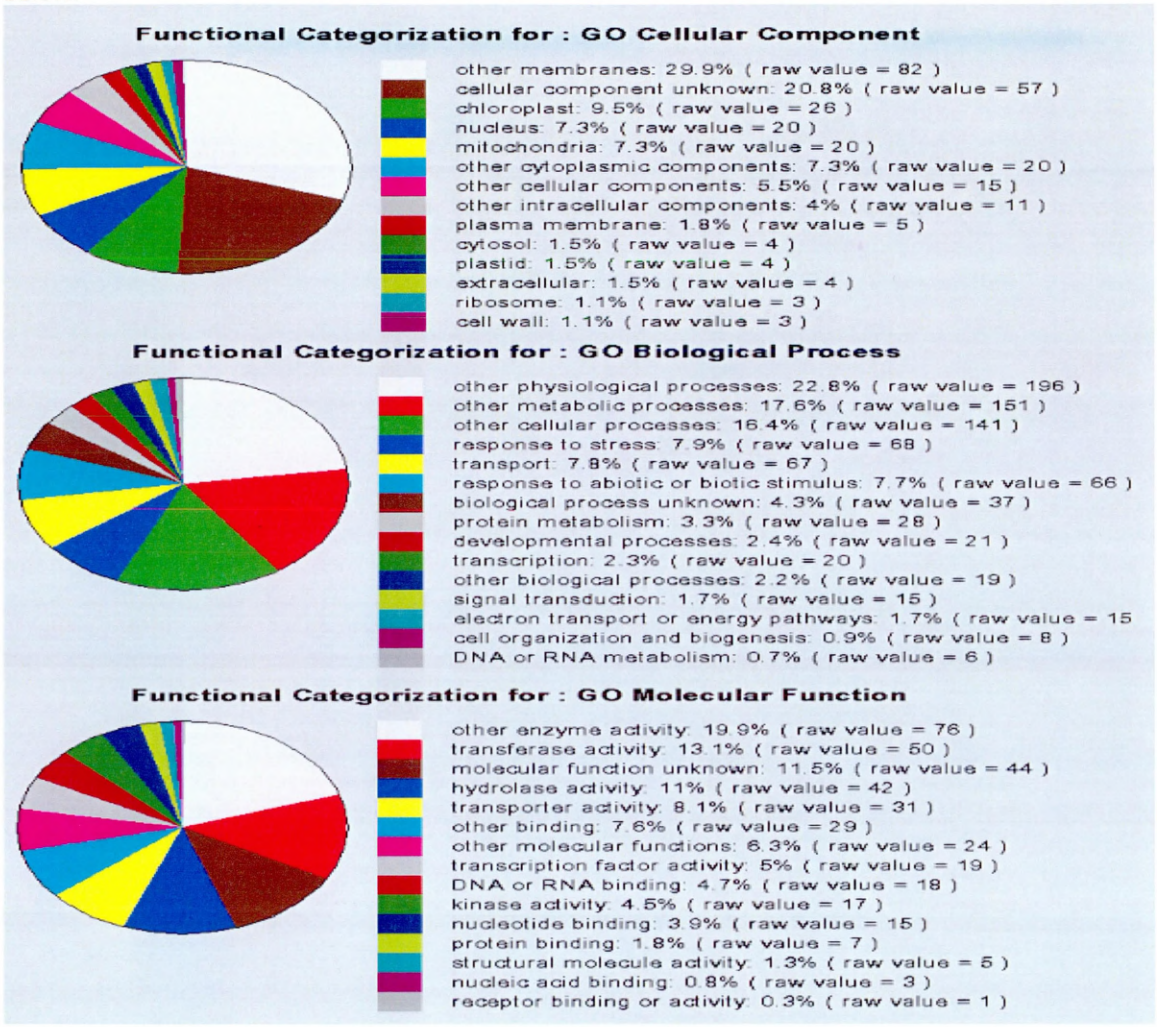

How to use the **Functional Categorisation Tool**

This search enables you to generate pie charts of the GO slim/Functional Classification set using TAIR GO tool.

Step 1: Use the search options to select genes. Click 'Submit Query'

Step 2: A list of the genes resulting from your search will be displayed. Select 'TAIR Functional Categorization' button.

Step 3: A pop up window should appear for the TAIR www.arabidopsis.org) site. Right Click in list box and select paste. Click 'Functional Categorization' button to obtain a list of functional categories for your selected of genes.
Step 4: Click 'Create Pie Charts' button to display the graphs. The results will be displayed as three separate graphs, each representing the functional classification according to keyword category (GO Molecular Function, GO Biological Process and GO Cellular Component). - an example output is shown below:




For Step 1 and Step 2 see instructions for Chromosome Map

**Process Results**

Step 3: Input results to TAIR Functional Categorisation tool.

A pop up window should appear for the TAIR www.arabidopsis.org) site. Right Click in the empty list box and select paste as illustrated in the screen shot below. This will copy your results into the TAIR tool. You should see text appear in the list box. Click the 'Functional Categorisation' button to obtain a list of functional categories for your selected of genes.

**View Functional Categorisation Pie Charts**

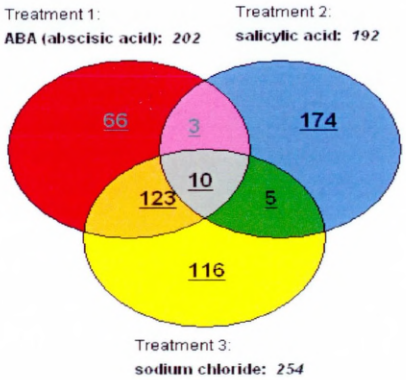
Step 4: Click 'Create Pie Charts' button to display the graphs.


The results will be displayed as three separate graphs, each representing the functional classification according to the keyword category (GO Molecular Function, GO Biological Process and GO Cellular Component). The created graphs are in GIF format and can be saved to your PC by right clicking on the image and selecting "Save Picture As.." option.

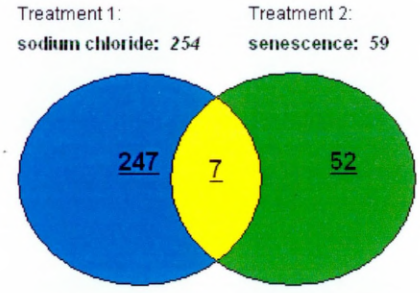
**Venn Diagrams**

This tool enables you to create Venn Diagrams using the *Arabidopsis thaliana* data from the DRASTIC database. You can select two or three treatments and the tool will process the selections and output the results as a Venn Diagram. The Venn Diagram displays the number of genes regulated by each individual treatment or by multiple treatments based on the Drastic data. You can choose to include records where genes have been up regulated, down regulated or both. The diagrams can be mined further by clicking on a segment of the diagram to view the individual reactions.

Select Treatments
View Venn Diagram
View Individual Reaction Results
View Journal References




Venn Diagram with three Treatments

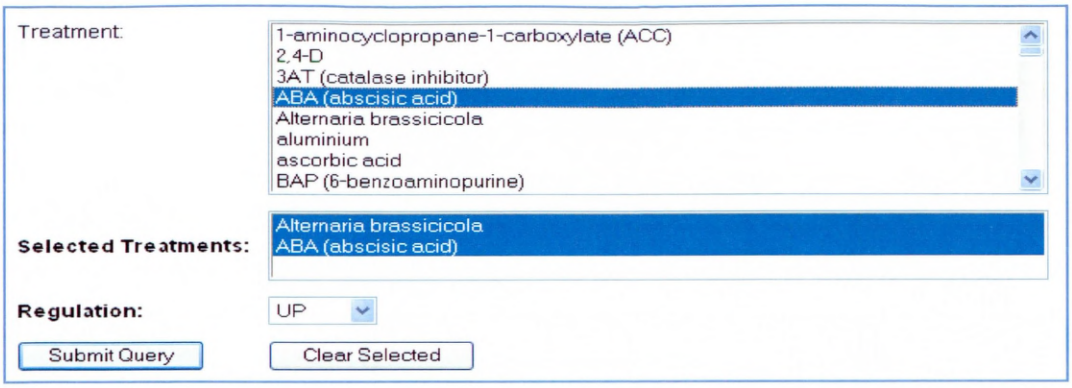



Venn Diagram with twoTreatments

**Search Page**

Use search option to select treatments. Click 'Submit Query'

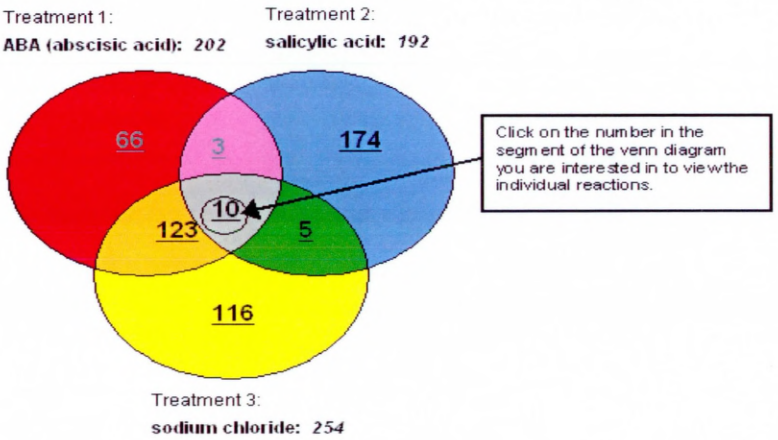
You can create a Venn diagram by selecting either two or three treatments. To select a treatment, double click on the treatment from the list box and it will appear in the 'Selected Treatment' box. To select the regulation of the included genes, select up, down or both from the list. Click 'Submit Query' to create the Venn diagram. To change your selection click 'Clear Selected'. Due to the large amount of data processed, this search may take a few minutes, but a progress bar will be displayed.




The search does not include any results in the database that do not have an AGI number associated with them.

**View Venn Diagram**

A Venn diagram is used to display the results of your search. Each segment of the Venn diagram has a number, which represents the number of genes that are regulated by the treatment combinations. The reactions for each segment of the Venn diagram can be viewed by clicking on underlined number.




**View Individual Reaction Results**

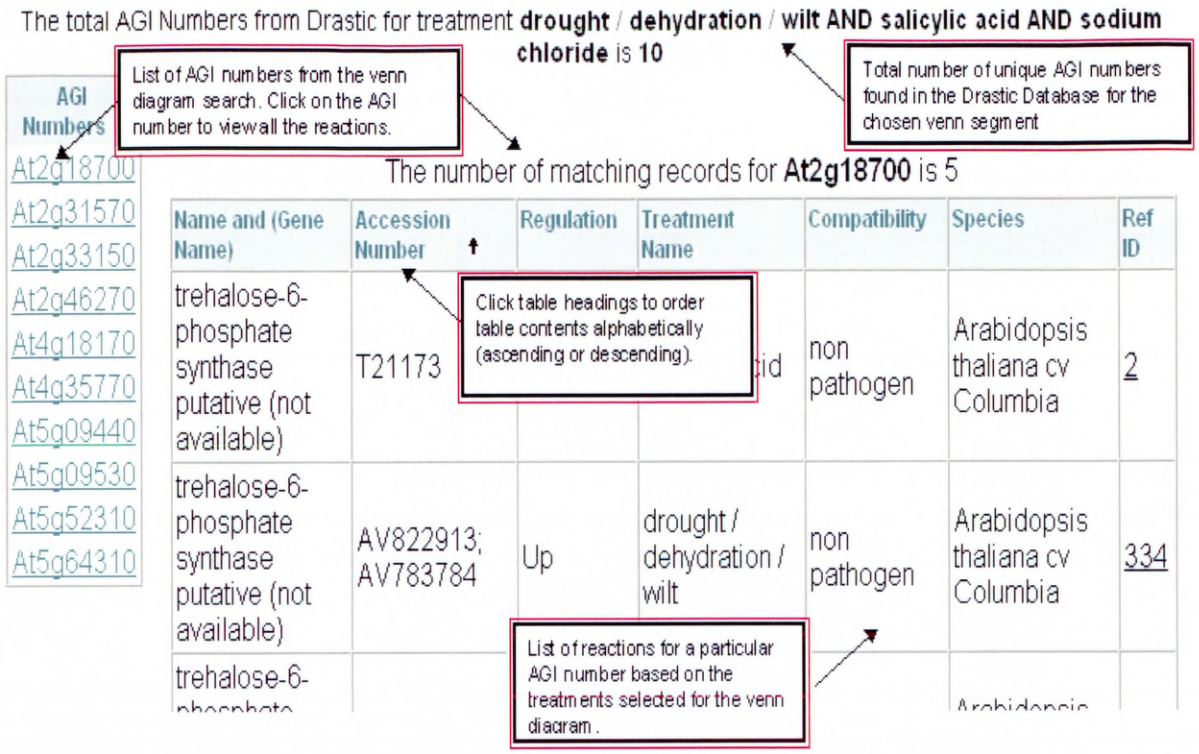
This screen enables you to view all the individual reactions held on drastic for the treatment(s) and regulation that you selected from the Venn diagram search as shown in the diagram below:

The number of unique AGI numbers and the treatments chosen are displayed at the top of the page.

The AGI numbers are displayed in a table on the left hand side of the screen to enable you easily navigate between the reactions.

Each table of records can be sorted alphabetically (ascending or descending) by clicking on the column heading that you want to sort the data by. The first time you click, it will sort the contents in ascending order. The second time you click it will sort the table contents in descending order. This is indicated by a black arrow that will appear in the sorted column heading.

To view the full journal reference data, click on the Reference ID number of the record you are interested in. This will open a new window which displays the full reference information including the journal it was published in.




**Pathway Tool**

The pathway tool enables you to extract and visualise knowledge from the database to hypothesise possible relationships between potential signalling genes.

It has a search facility that allows you to select a number of *Arabidopsis thaliana* genes using the AGI numbers. This produces a 'pathway' which enables you to look at the regulation of genes in response to different treatments.

It is possible to see that certain groups of genes are always co-regulated suggesting that they are likely to occur in the same signal transduction pathway. An example of this is shown below:

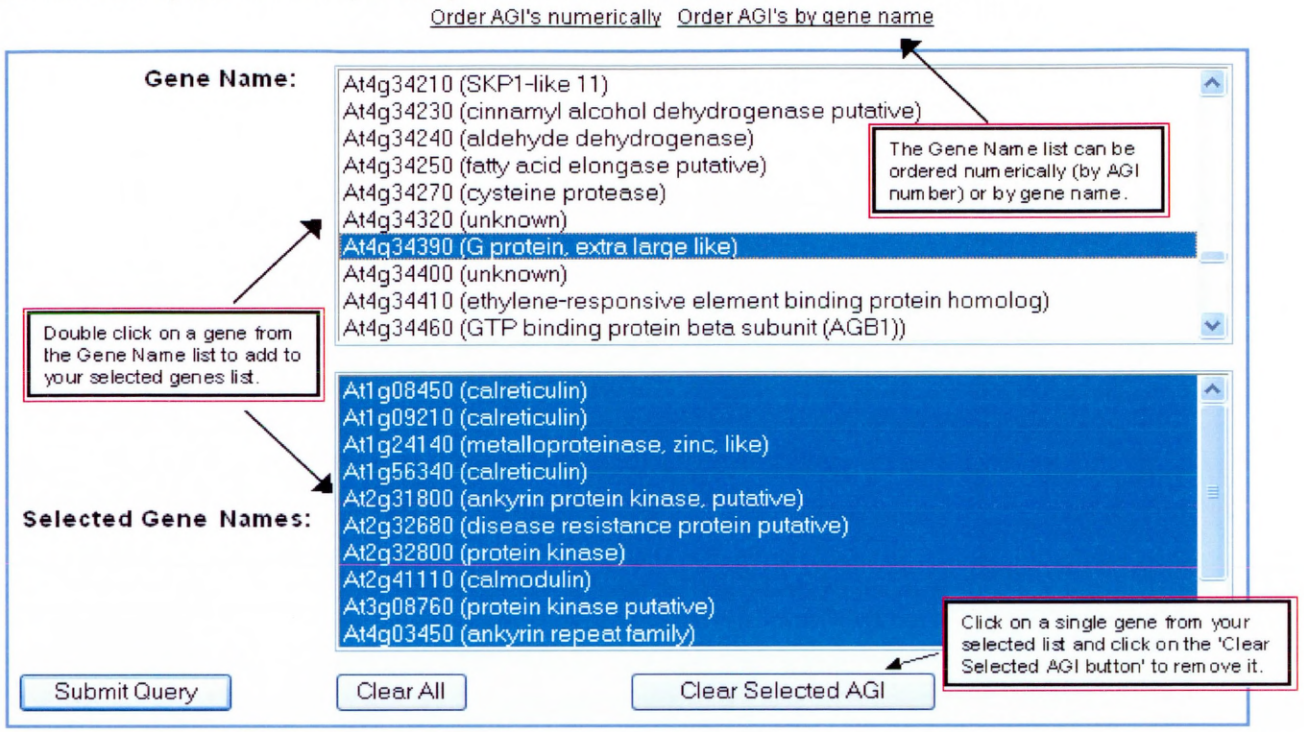

**Pathway Search Page**

The Pathway Search enables you to select genes by AGI number or gene name (only includes genes which have an AGI number in the Drastic database can be selected to ensure accuracy of results).

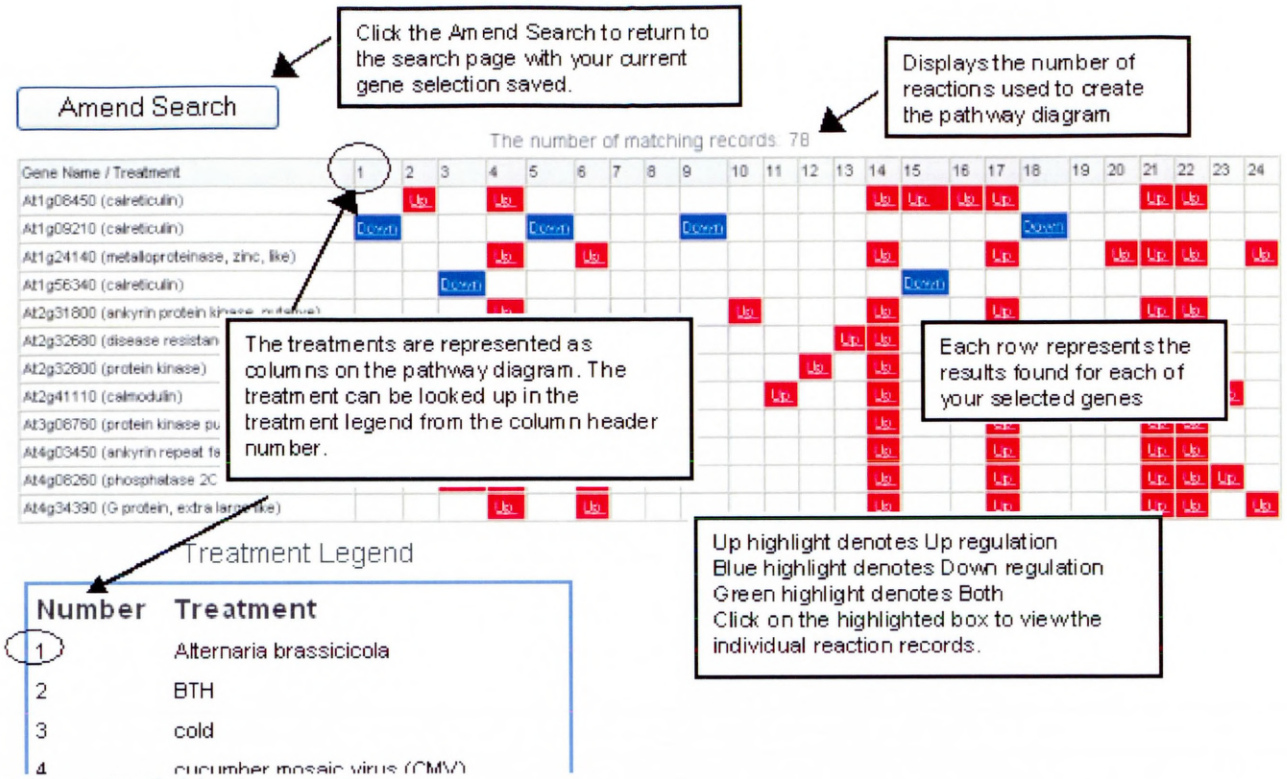
**Selecting genes:** Double click on the gene in the Gene Name and it will be copied to the lower box and included in your search. Click 'Submit Query' when you have completed your selection.

**Remove genes from search:** To remove one gene from your search, click on the gene in the lower box and then click the 'Clear Selected AGI'. To clear all your current selection click 'Clear All'

**Sorting gene list:** To make finding your selection easier, you can choose to sort the gene list by name or AGI number. NOTE: Sort the list prior to selecting the genes to include in the search or you may have to re-select your genes.
**Bulk Upload:** If you know the AGI numbers of the genes you want to include in your search, select the 'AGI Bulk Upload' link and you can paste the AGI numbers into the text box.




**Pathway Results**
The diagram below shows a sample pathway result.
The column headers represent the treatments (see treatment legend) for which the database holds records.
The rows hold your selected genes and the corresponding results.
The blue highlight denotes down-regulation, red highlight denotes up-regulation and the green highlight denotes that reactions for both up- and down- regulation are held.
The number of reactions used to create the pathway diagram is displayed at the top.
**Amend Search:** The 'Amend Search' button enables you to return to the search page with the genes you previously selected saved to enable you to add to your current search.



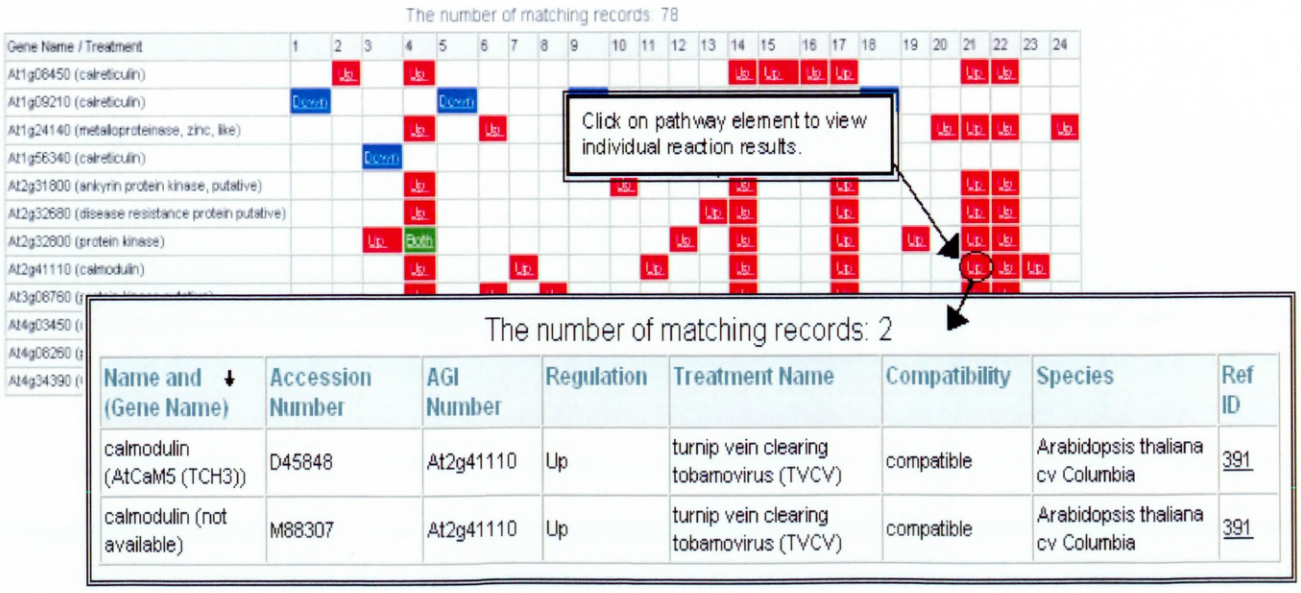

**Pathway Individual Reactions**

This screen enables you to view all the individual reactions held on the drastic database for the specific pathway element selected as shown in the diagram below:

The number of results are displayed at the top of the page.

The whole table of records can be sorted alphabetically (ascending or descending) by clicking on the column heading that you want to sort the data by. The first time you click, it will sort the contents in ascending order. The second time you click it will sort the table contents in descending order. This is indicated by a black arrow that will appear in the sorted column heading.

To view the full journal reference data, click on the Reference ID number of the record you are interested in. This will open a new window which displays the full reference information including the journal it was published in.




| Name and (Gene Name) | Accession Number | AGI Number | Regulation | Treatment Name | Compatibility | Species | Ref ID |
|---|---|---|---|---|---|---|---|
| calmodulin (AtCaM5 (TCH3)) | D45848 | At2g41110 | Up | turnip vein clearing tobamovirus (TVCV) | compatible | Arabidopsis thaliana cv Columbia | 391 |
| calmodulin (not available) | M88307 | At2g41110 | Up | turnip vein clearing tobamovirus (TVCV) | compatible | Arabidopsis thaliana cv Columbia | 391 |

## Roadmap Tool

The Roadmap Tool enables the user to create look-up tables to find genes that are co-regulated by treatments (only includes genes which have an AGI number in the Drastic database can be selected to ensure accuracy of results).

It has a search facility that allows you to select one *Arabidopsis thaliana* gene using the AGI numbers. This produces a 'roadmap' that enables you to look at the regulation of genes in response to different treatments. This can assist in identifying groups of treatments that appear to produce similar regulatory results.

The search operates by identifying all treatments that regulated the selected AGI number. These treatments are then used to create the roadmap, and all AGIs that are regulated by these treatments are included in the map (see example below):

Roadmap Search Page
Roadmap Results
View Individual Results

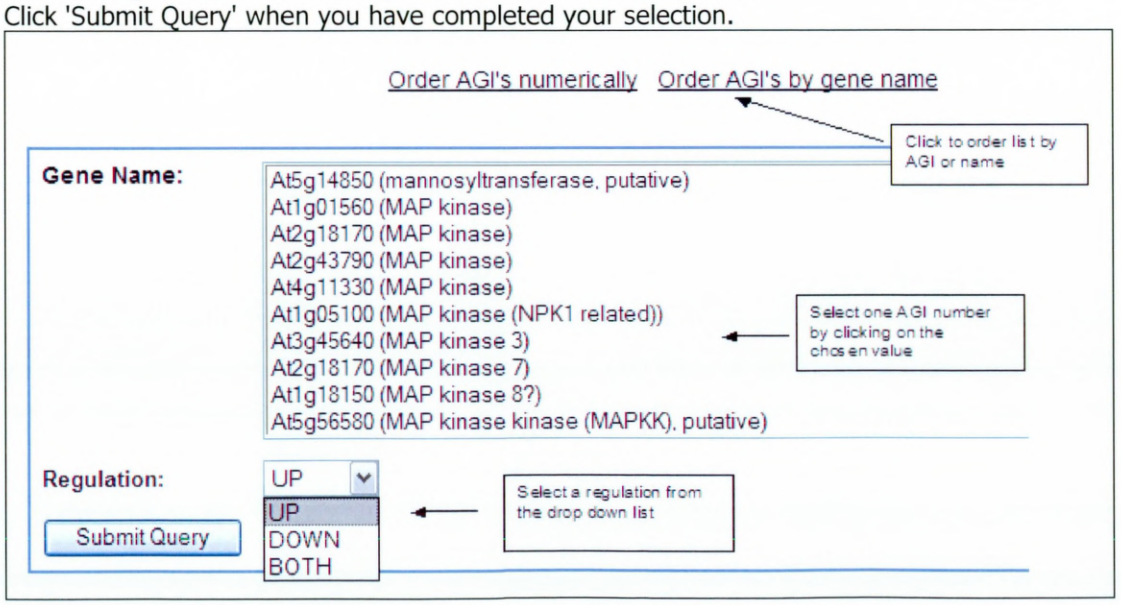
**At4g11330 (Map Kinase) - Up Regulation**

| | chitin | flagellin 22 (flg22) | Yariv phenylglycoside (beta-D-Glc)3 |
|---|---|---|---|
| chitin | 24 | 17 | 8 |
| flagellin 22 (flg22) | 17 | 822 | 252 |
| Yariv phenylglycoside (beta-D-Glc)3 | 8 | 252 | 410 |

## Roadmap Search Page

The Roadmap search enables you to select an AGI number and a regulation.

**Select Gene:** The list of genes can be ordered by AGI or name. Click on the AGI number to select it.
**Select Regulation:** Select a regulation from the drop down list.
Click 'Submit Query' when you have completed your selection.



## Roadmap Results

The diagram below shows a sample pathway result.
The column and row headers represent the treatments for which the database holds records and are used as a look-up table.
The red highlight indicates cells showing the total AGIs that are regulated by a single treatment. E.g. in the roadmap below, there are 58 genes that are regulated by chitin.
The underlined results in the white cells show the number of genes that are regulated by both treatments. E.g. in the roadmap below, it can be seen that there are 14 genes that are co-regulated by chitin and wound.

**At3g45640 (Map Kinase 3) – Up or Down Regulation (Both)**

| | calyculin | chitin | drought / dehydration / wilt | flagellin 22 (flg22) | hydrogen peroxide | jasmonate (methyl) | ozone | salicylic acid | UV | UV + ascorbic acid | wound | Yariv phenylglycoside (beta-D-Glc)3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| calyculin | 4 | 3 | 2 | 4 | 1 | 3 | 2 | 2 | 1 | 1 | 3 | 4 |
| chitin | 3 | 58 | 8 | 19 | 7 | 23 | 3 | 19 | 3 | 2 | 14 | 8 |
| drought / dehydration / wilt | 2 | 8 | 401 | 40 | 27 | 62 | 27 | | | | | 26 |
| flagellin 22 (flg22) | 4 | 19 | 40 | 912 | 48 | 37 | 26 | | | | | 254 |
| hydrogen peroxide | 1 | 7 | 27 | 48 | 160 | 20 | 10 | | | | | 27 |
| jasmonate (methyl) | 3 | 23 | 62 | 37 | 20 | 313 | 19 | 165 | 4 | 3 | 45 | 15 |
| ozone | 2 | 3 | 27 | 26 | 10 | 19 | 213 | 18 | 2 | 2 | 24 | 11 |
| salicylic acid | 2 | 19 | | | | 165 | 18 | 376 | 4 | 1 | 36 | 17 |
| UV | 1 | 3 | | | | 4 | 2 | 4 | 19 | 8 | 7 | 7 |
| UV + ascorbic acid | 1 | 2 | | | | 3 | 2 | 1 | 8 | 8 | 4 | 4 |
| wound | 3 | 14 | 83 | 125 | 21 | 45 | 24 | 36 | 7 | 4 | 342 | 54 |
| Yariv phenylglycoside (beta-D-Glc)3 | 4 | 8 | 26 | 254 | 27 | 15 | 14 | | | | 54 | 473 |

Click on a number and the supporting records for both treatments will be displayed (along with references)

The white boxes show the total number of AGIs that are regulated by BOTH treatments

The boxes in reds how the total number of records (unique AGIs) for a single treatment.

Treatments are displayed in rows and columns and are used as a look up table e.g. there are 14 AGIs that are co-regulated by wound and chitin

## View Individual Roadmap Results

This screen enables you to view all the individual reactions held on the drastic database from BOTH treatments and the supporting references
The number of results is displayed at the top of each table. (You may have to scroll down the page for the second treatment results. The number of results may be different for each treatment as this depends how many supporting records are held by Drastic.
Each table of records can be sorted alphabetically (ascending or descending) by clicking on the column heading that you want to sort the data by. The first time you click, it will sort the contents in ascending order. The second time you click it will sort the table contents in descending order. This is indicated by a black arrow that will appear in the sorted column heading.
To view the full journal reference data, click on the Reference ID number of the record you are interested in. This will open a new window that displays the full reference information including the journal it was published in.

The number of matching records: **39** for treatment **drought / dehydration / wilt**

| Name | Gene Name | Accession Number | Regulation | AGI Number | Ref ID |
|---|---|---|---|---|---|
| calmodulin related putative | not available | AV824175 AV785354 | Up | At5g42380 | 334 |
| calmodulin related putative | not available | R30557 | Up | At5g42380 | 1 |
| cytochrome P450 | AtCYP707A1 | AV825284 AV790857 | Up | At4g19230 | 334 |
| cytochrome P450 | AtCYP707A1 | AV825284 AV790857 AY050980 | Up | At4g19230 | 107 |
| cytochrome P450 | AtCYP707A3 | At5g45340 | Up | At5g45340 | 482 |
| dehydrin like | lti30 | Z18121 | Up | At3g50970 | 26 |
| late embryogenesis abundant LEA D113 type 1 | not available | AV826209 AV794009 | Up | At5g06760 | 334 |

The number of matching records: **26** for treatment **Yariv phenylglycoside (beta-D-Glc)3**

| Name | Gene Name | Accession Number | Regulation | AGI Number | Ref ID |
|---|---|---|---|---|---|
| calmodulin related putative | not known | At5g42380 | Up | At5g42380 | 455 |
| cytochrome P450 | AtCYP707A1 | At4g19230 | Up | At4g19230 | 455 |
| cytochrome P450 | AtCYP707A3 | At5g45340 | Up | At5g45340 | 455 |
| dehydrin like | lti30 | At3g50970 | Up | At3g50970 | 455 |
| late embryogenesis abundant LEA D113 type 1 | not available | At5g06760 | Down | At5g06760 | 455 |
| late embryogenesis abundant LEA SAG21 homolog | not available | At4g02380 | Up | At4g02380 | 455 |

## Unique Gene Tool

The Unique Genes Tool identifies all of the *Arabidopsis thaliana* genes that are regulated by a single treatment based on the Drastic data. Full details (including the reference) of each gene are available by selecting a result. The results will dynamically change with time as more data is added to the database.

Unique Gene Results
View Individual Results

## Unique Gene Results

This screen displays all the *Arabidopsis thaliana* genes that Drastic has identified as being regulated by a single treatment.

The number of unique AGI numbers is displayed at the top of the page.
To make viewing your results easier, you can choose to sort the list by name, AGI number or treatment.
This data can be mined further by viewing the Individual Results by clicking on the AGI number.

Order AGI's numerically Order AGI's by gene name Order Treatments alphabetically

The number of matching records: **3117**

| AGI Number (Name) | Treatment |
|---|---|
| At1g01010 (no apical meristem) | ABA (abscisic acid) |
| At1g01070 (nodulin MtN21 family) | Pseudomonas fluorescens |
| At1g01120 (fatty acid elongase 3-ketoacyl-CoA synthase 1) | cucumber mosaic virus (CMV) |
| At1g01130 (unknown) | ABA (abscisic acid) |
| At1g01300 (aspartyl protease) | zeatin (cytokinin) |
| At1g01370 (histone H3 HTR12) | ABA (abscisic acid) |
| At1g01480 (1-aminocyclopropane-1-carboxylate (ACC) synthase) | cucumber mosaic virus (CMV) |
| At1g01490 (heavy-metal-associated domain-containing protein) | Pseudomonas fluorescens |

**View Individual Results**

This screen enables you to view all the individual reactions held on the drastic database according to your search criteria as shown in the diagram below:

The number of results is displayed at the top of the page.

The whole table of records can be sorted alphabetically (ascending or descending) by clicking on the column heading that you want to sort the data by. The first time you click, it will sort the contents in ascending order. The second time you click it will sort the table contents in descending order. This is indicated by a black arrow that will appear in the sorted column heading.

To view the full journal reference data, click on the Reference ID number of the record you are interested in. This will open a new window that displays the full reference information including the journal it was published in.

# Appendix VII – Contents of CD

1. Website Files (These will need an ASP Server to view the site)
Folder Drastic_Website:

AccNumResponse.asp
AGIResponse.asp
BasicQueryReference.asp
DatabaseStats.asp
Pathway.asp
PathwayCheck.asp
PathwayP.asp
PathwayResult.asp
PathwayResultp.asp
Search.asp
searchAccNum.asp
SearchAGI.asp
StatsReactions.asp
StatsResult.asp
Tairagi.asp
TairAGIResponse.asp
VennDetails.asp
VennQuery.asp
VennQueryResponse.asp

Folder Drastic_Website/Files

DatabaseDetails.inc
Footer.inc
headerA.inc
headerB.inc
insight.css
List.js
Sorttable.js
Treearr.js

2. Drastic Database including User Entry Forms

DrasticV6.mdb

3. Unigene Search Program

UnigeneSearch.vbp
UnigeneSearchIngo.doc