

# E-science and Systems Biology - A Revolution in the Life Sciences?

Chris Rawlings  
Head of Department of Biomathematics and  
Bioinformatics

<http://www.rothamsted.ac.uk/bab>

Rothamsted Research  
chris.rawlings@bbsrc.ac.uk

# Outline

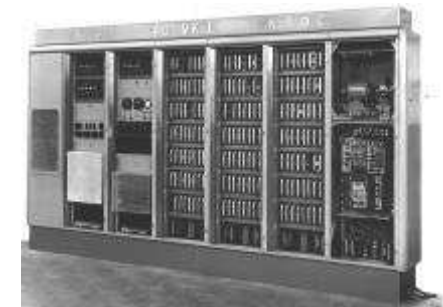
- Rothamsted Research
- Systems Biology, Bioinformatics
- Enriching Biology Data
  - Integrating Data
  - Text Mining to Support Database Curation
- What is different about Systems Biology

# Rothamsted Research



- **Largest agricultural and crop science research institute in UK**
- **Research started in 1853**
- **400 Staff**
- **Funding BBSRC (55%)**
- **Others Defra, EU, Industry**

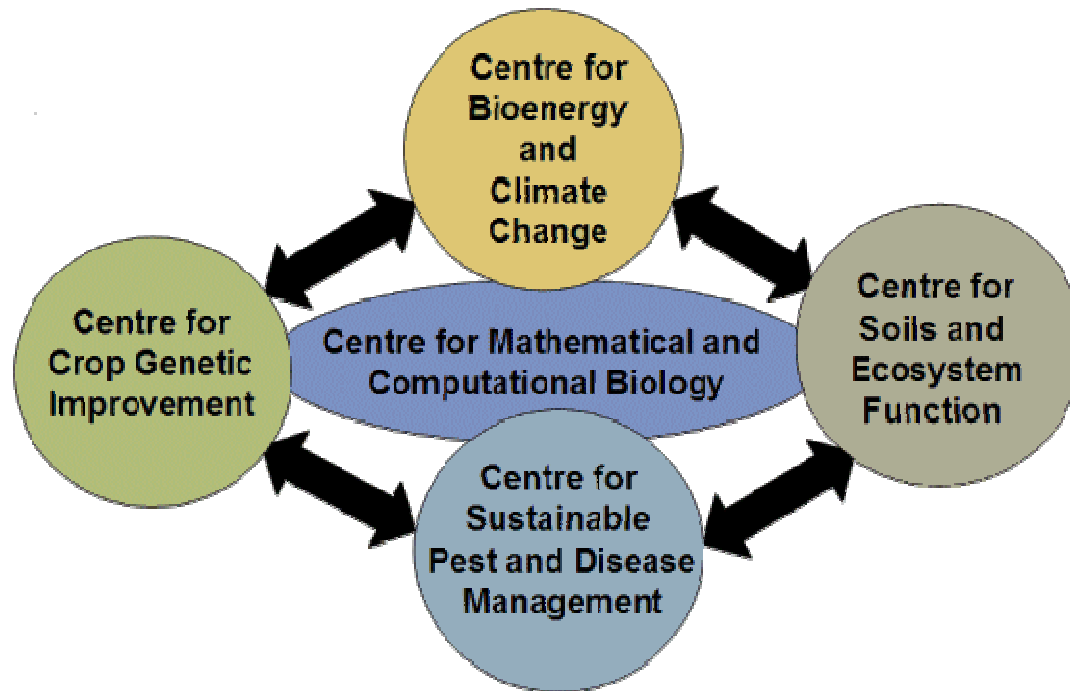
# The classical experiments



# New Approaches – High throughput science in agriculture research



# Rothamsted's Five Research Centres



**The impacts of climate change on agriculture and its mitigation**

**Genetic improvement of arable crops with improved resource use, performance, yield and end-use quality**

**The vital functions performed by soils and agricultural ecosystems – (e.g. role in GHG)**

**Effective and lasting approaches to reducing the impacts of pest and disease**

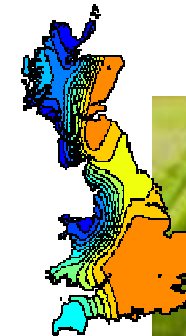
**Use of informatics, mathematics and statistics to derive added value from large volumes of complex noisy data. (E-science)**

# Research style

- Mixture of basic and applied research
  - Translational research important
    - BBSRC-> RRes->Defra->farmers->processors
- Strongly interdisciplinary
  - plant, insect and microbial molecular and cell biology, plant and insect ecology, soil science, chemistry, physics, mathematics, statistics, bioinformatics
- Increasing use of molecular biological approaches to understanding:
  - Interactions between plants and their pests and pathogens including disease resistance
  - Biological diversity in above and below ground ecosystems
  - The mechanisms controlling the productivity of crop plants and their responses to biotic and abiotic stress

# Example Systems

- Molecular basis for crop performance
  - Pathways controlling yield, nutrient usage efficiency
  - Pathways controlling crop architecture
    - Role of plant hormones and signalling
  - From field trials through genetics, genomics, metabolomics
- Plant – Pathogen and Pest Interactions
  - Genomics of fungal pathogens
  - Host resistance mechanisms
    - How the pathogen overcomes pests
  - Pathogen genome sequence annotation
  - From molecular biology to epidemiology
- Interplay between crop plant, nutritional or disease status and weather
  - Impact of climate change
  - Reducing inputs
  - Genetics, genomics, multi-environment trials

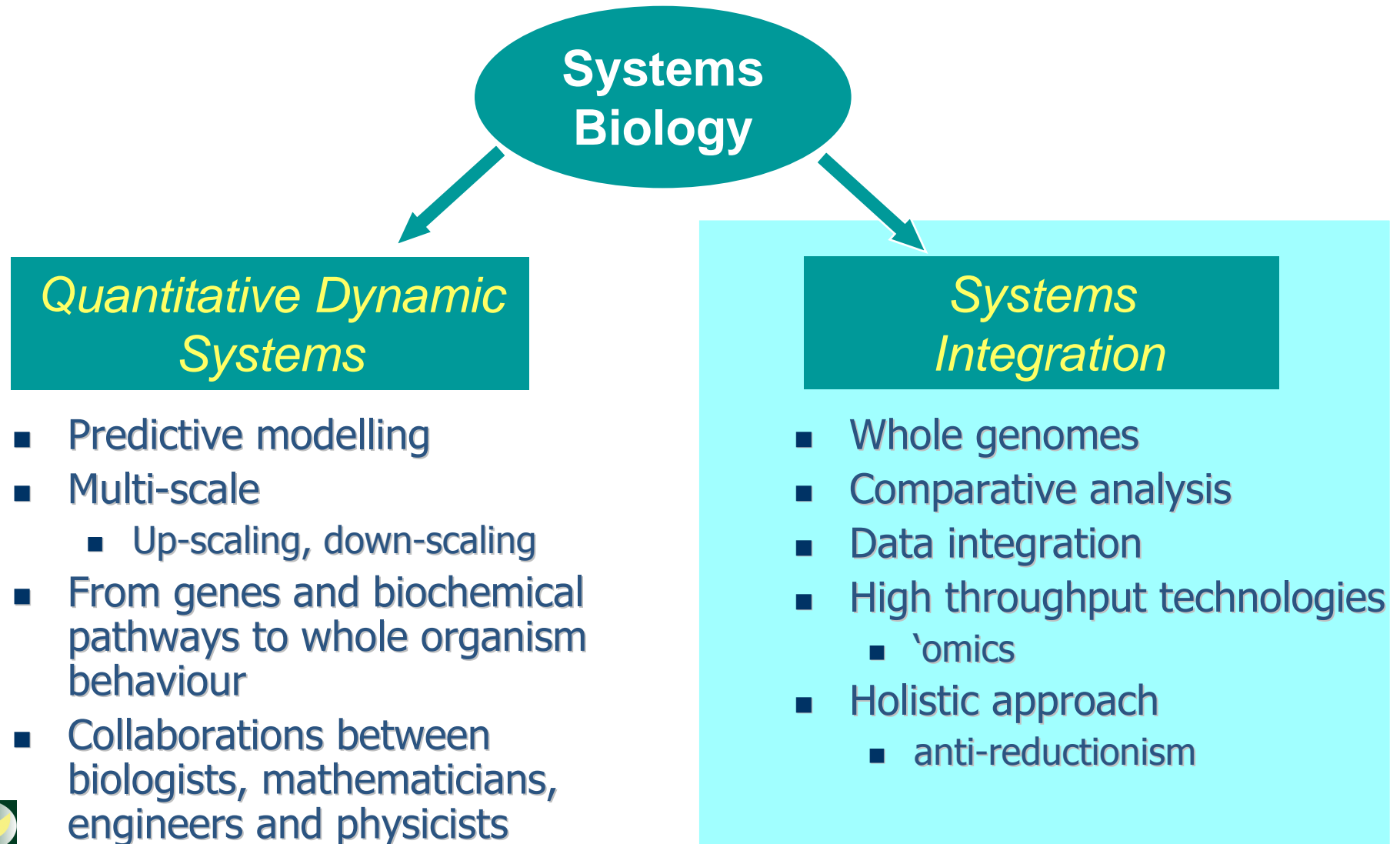




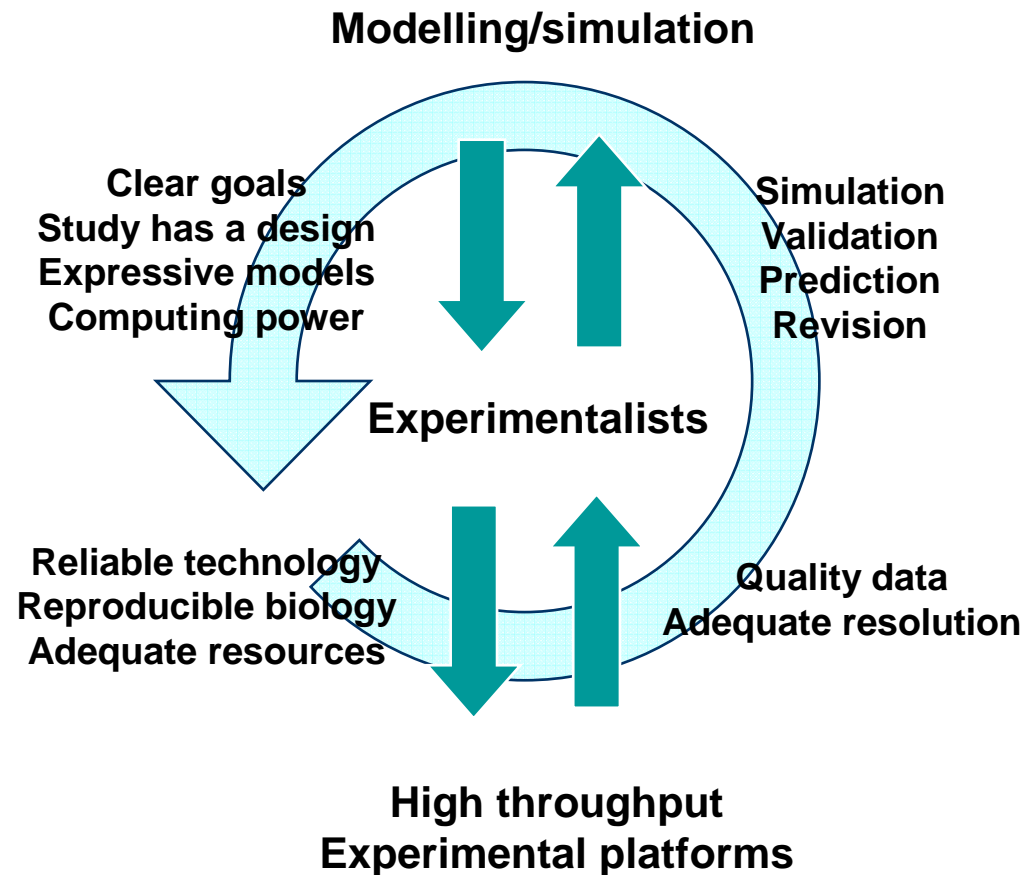
ROTHAMSTED  
RESEARCH

# Systems Biology

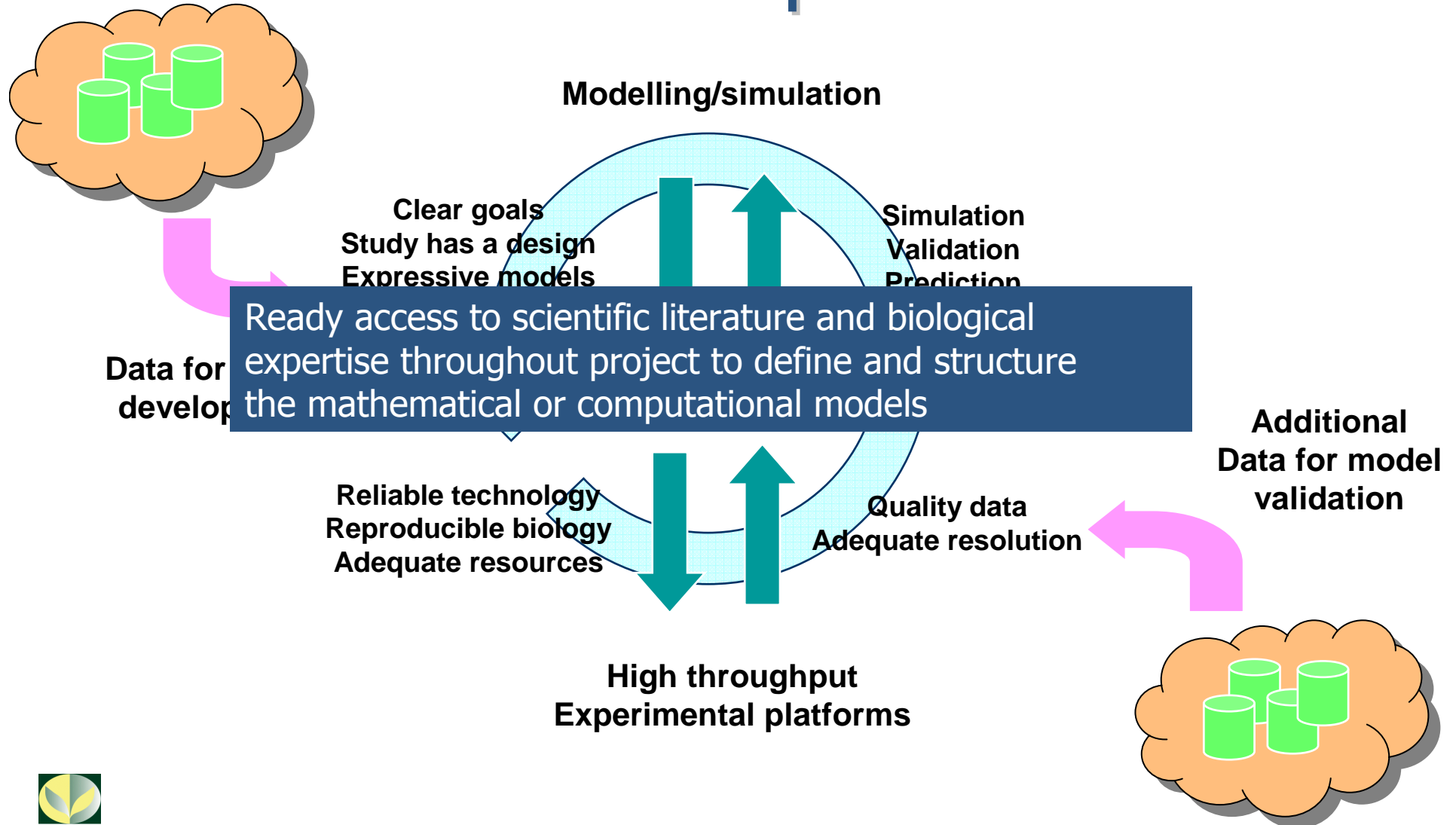
# Systems Biology - Two Definitions



# An Ideal Systems Biology Project



# Common Requirements

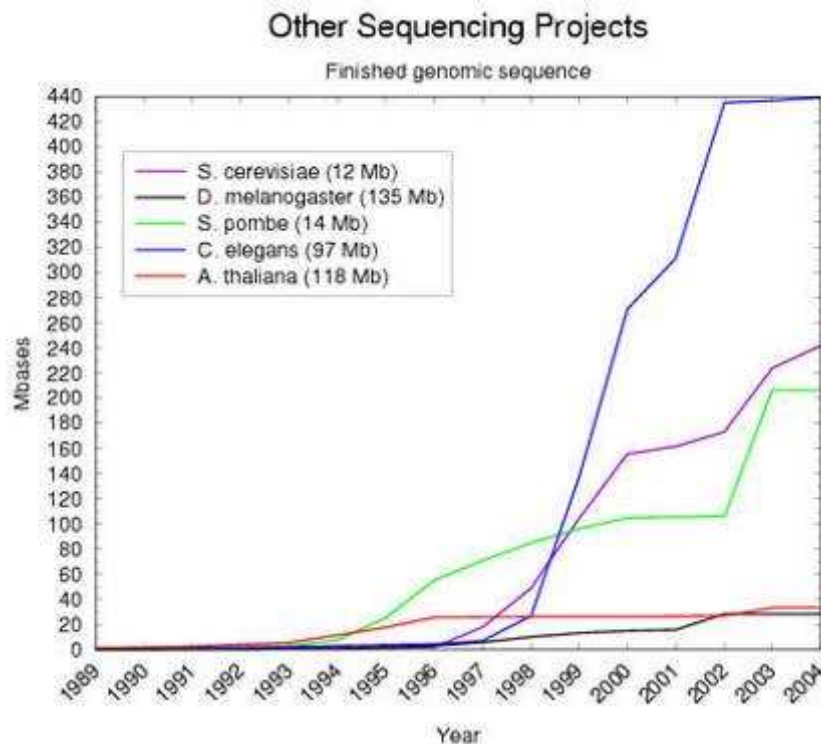


# Enriching Data for Systems Biology

## Bioinformatics and Systems Biology Data Sources

1. Data integration
2. Expert curation

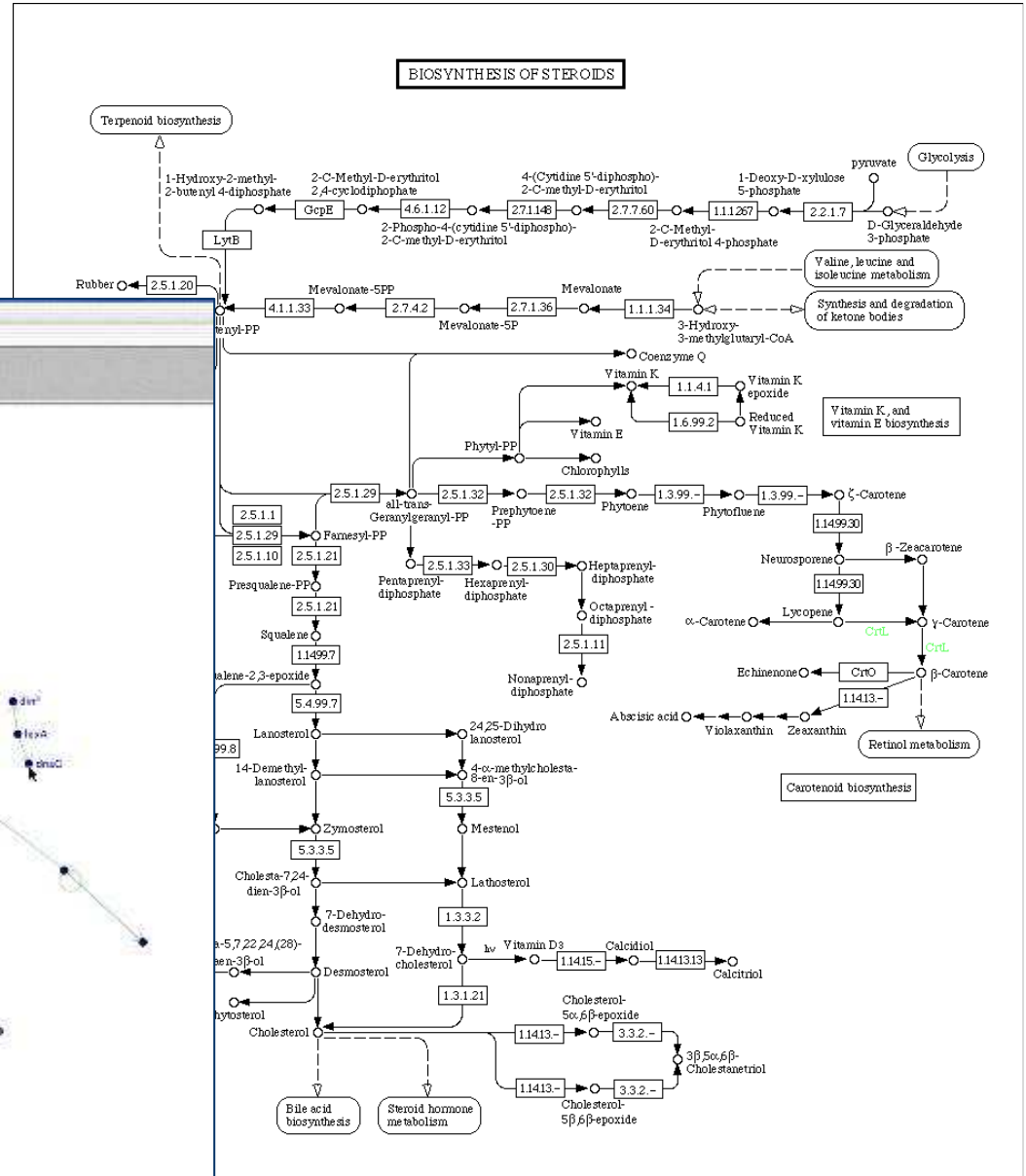
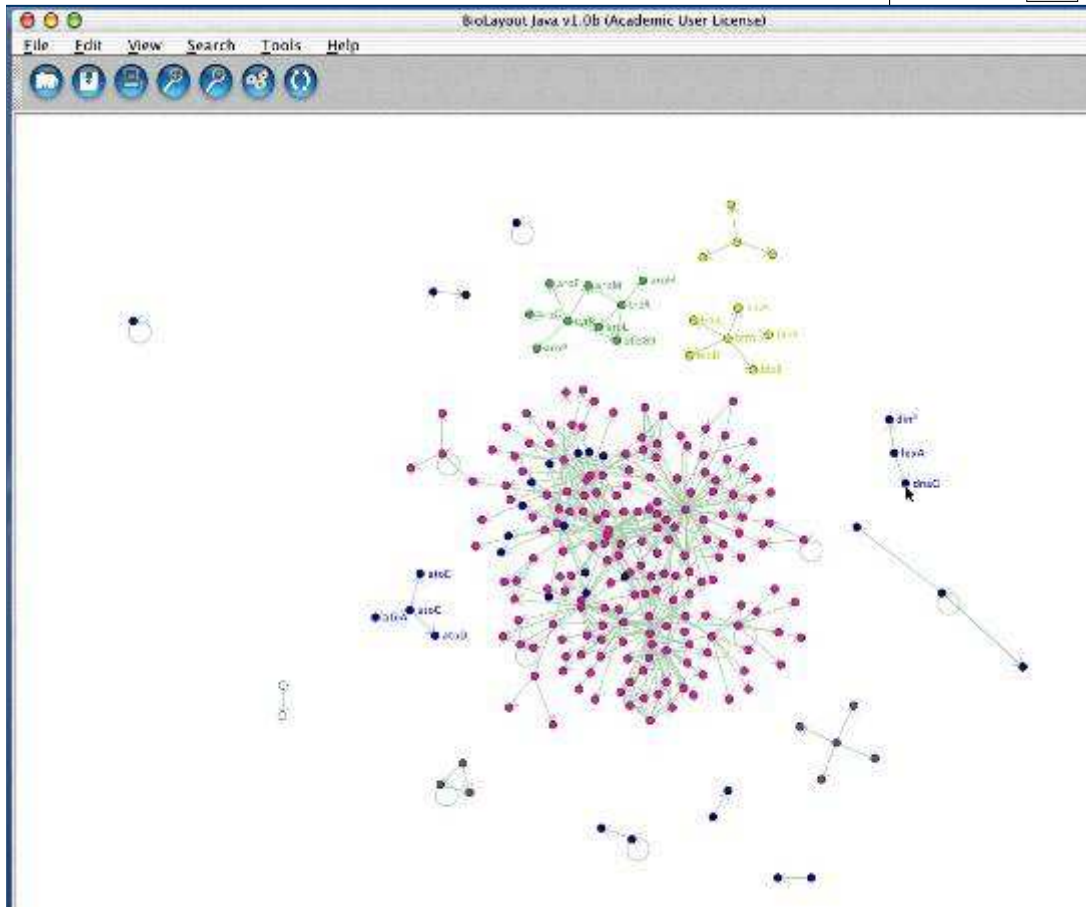
# Data volume is a challenge - but not the most important



- By comparison with other domains, the volume of data is not that great
- The real challenges are:
  - The interrelatedness of all these data
  - The complexity of the dependencies
  - The incompleteness of the data



# Complexity of interactions

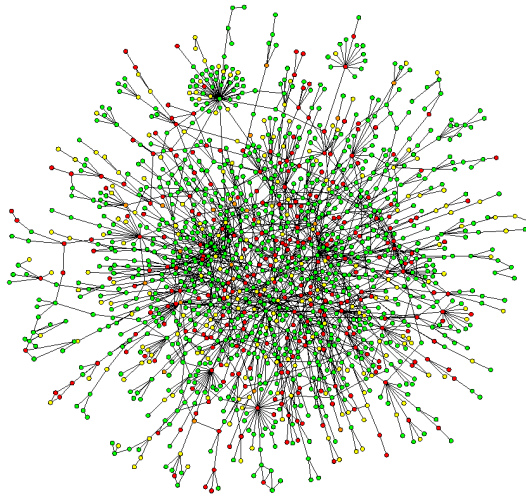


# Data Integration

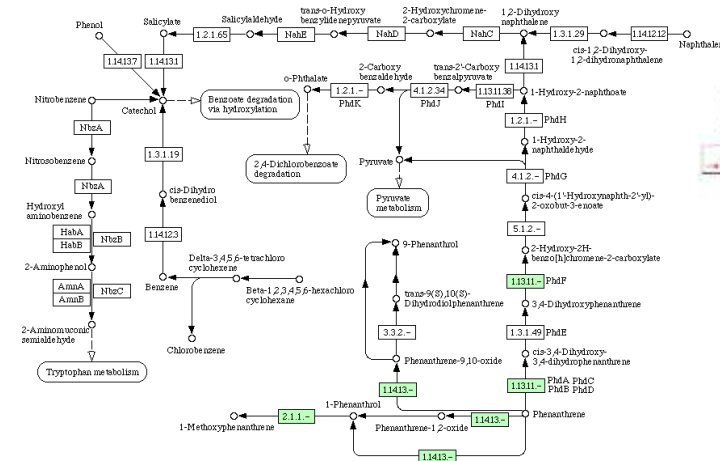
- ONDEX system
  - <http://ondex.sourceforge.net>
  - Key features:
    - Treats all data as components in a graph of concepts linked by edges with defined semantics
    - All information is a network
      - Ontologies provide key to linking across information types
      - Specialist treatment of text and sequence information
    - Recent version exploits emerging GRID technologies

# ONDEX principles

everything is a network...



protein interactions



00626 1018102

metabolic pathways

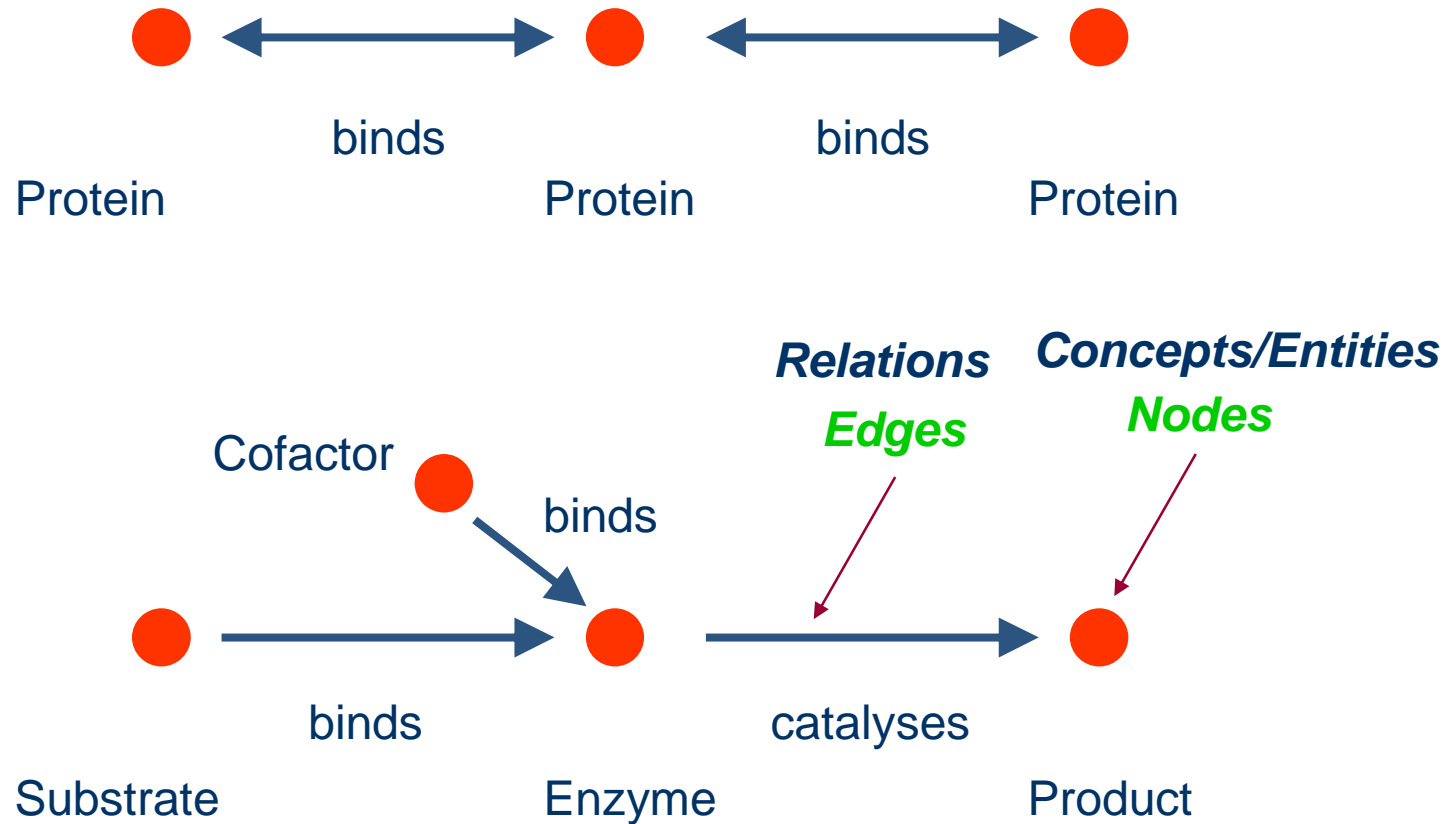


ontologies

... in which the nodes and edges have different properties

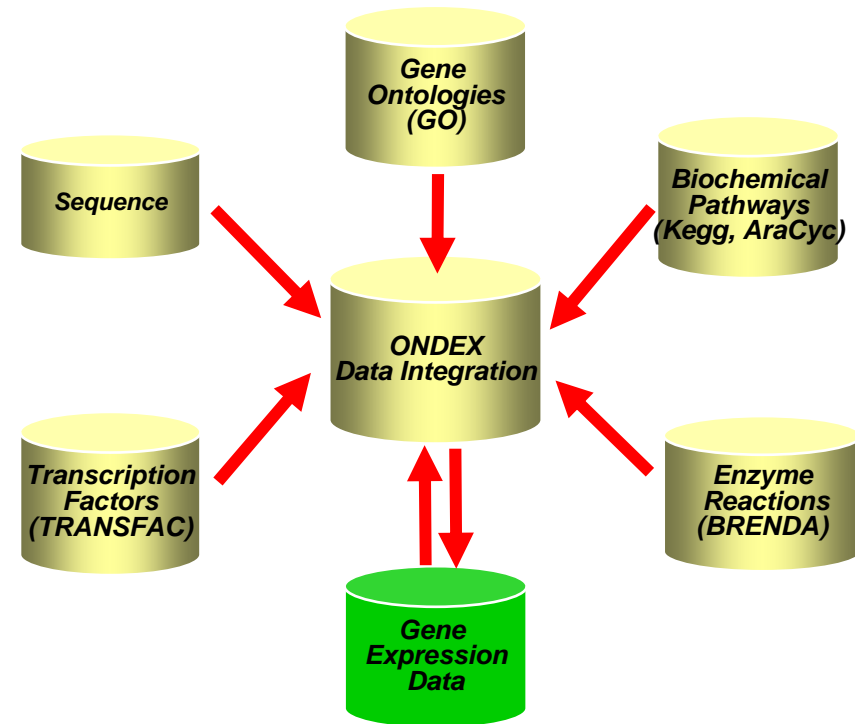
# Main idea

## Simple graphs

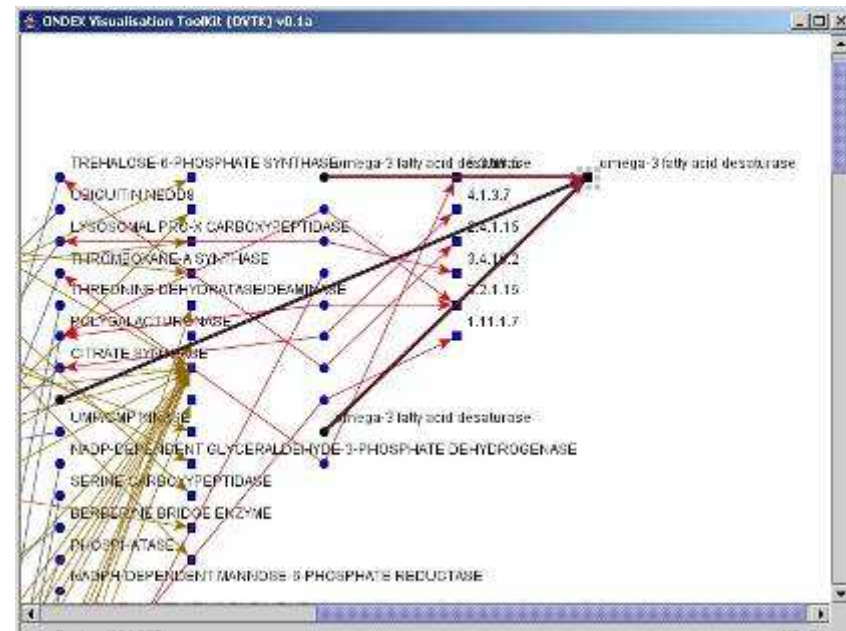
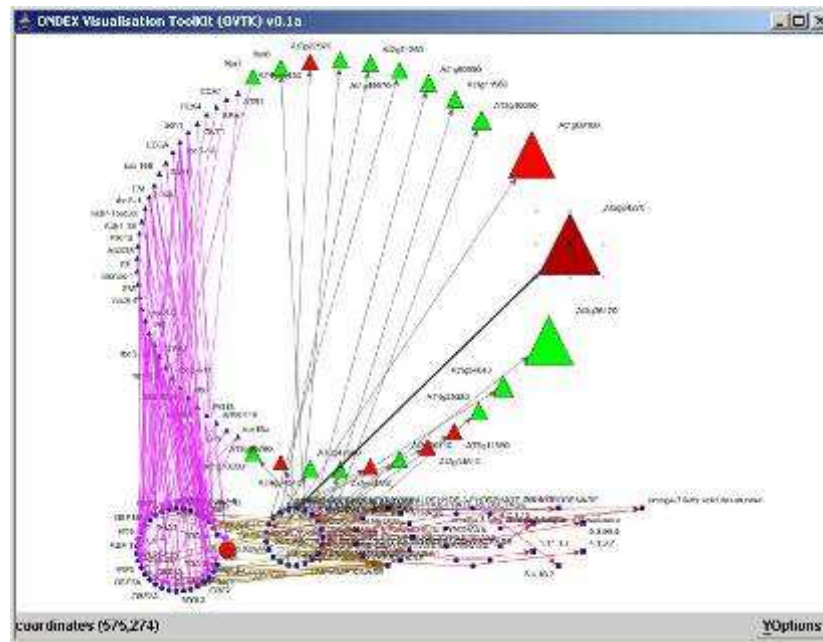


# Integrated Analysis of 'Omics Data

- ONDEX for Gene Expression
- Use integrated information to help provide biological context/explanation for the pattern of up/down regulated genes



# Graph Visualisation & Analysis



Gene expression signal strength expressed as colour and size of glyph  
Relationship between genes/proteins shown as lines  
Circular layout designed to display maximum number of concepts/relations

# Pilot Study

Arabidopsis data with 120 “novel” genes

New observations not in original paper made because of access to integrated data:

- provided annotation to 50 “novels”
- an important “unspotted” gene (a TF)
- drought stress
- jasmonic acid biosynthesis

**Köhler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Rueegg, A., Rawlings, C., Verrier, P. and Philippi, S. (2006)**  
***Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics 22(11):1383-90.***

# Enriching Biological Data by Expert Curation

Experiments in text mining

# The Pathogen-Host Interaction Database

<http://www.phi-base.org/>

## Database of genes from plant fungal pathogens

- List of “hot” target genes from literature
- Genes validated by gene disruption experiments
- Spreadsheet developed by hand
- Research question - use of text mining to improve search for additional genes
- Extended to other pathogens
- Need to supplement manual methods



# Why a database?

- support analysis of experimental results
- identify key pathogen genes and families across species
- Pathogen genome annotation
- how are the genes related?
- pathway analysis
- starting point for fungicide/drug target identification

# Original Curation Process

Papers



Curator

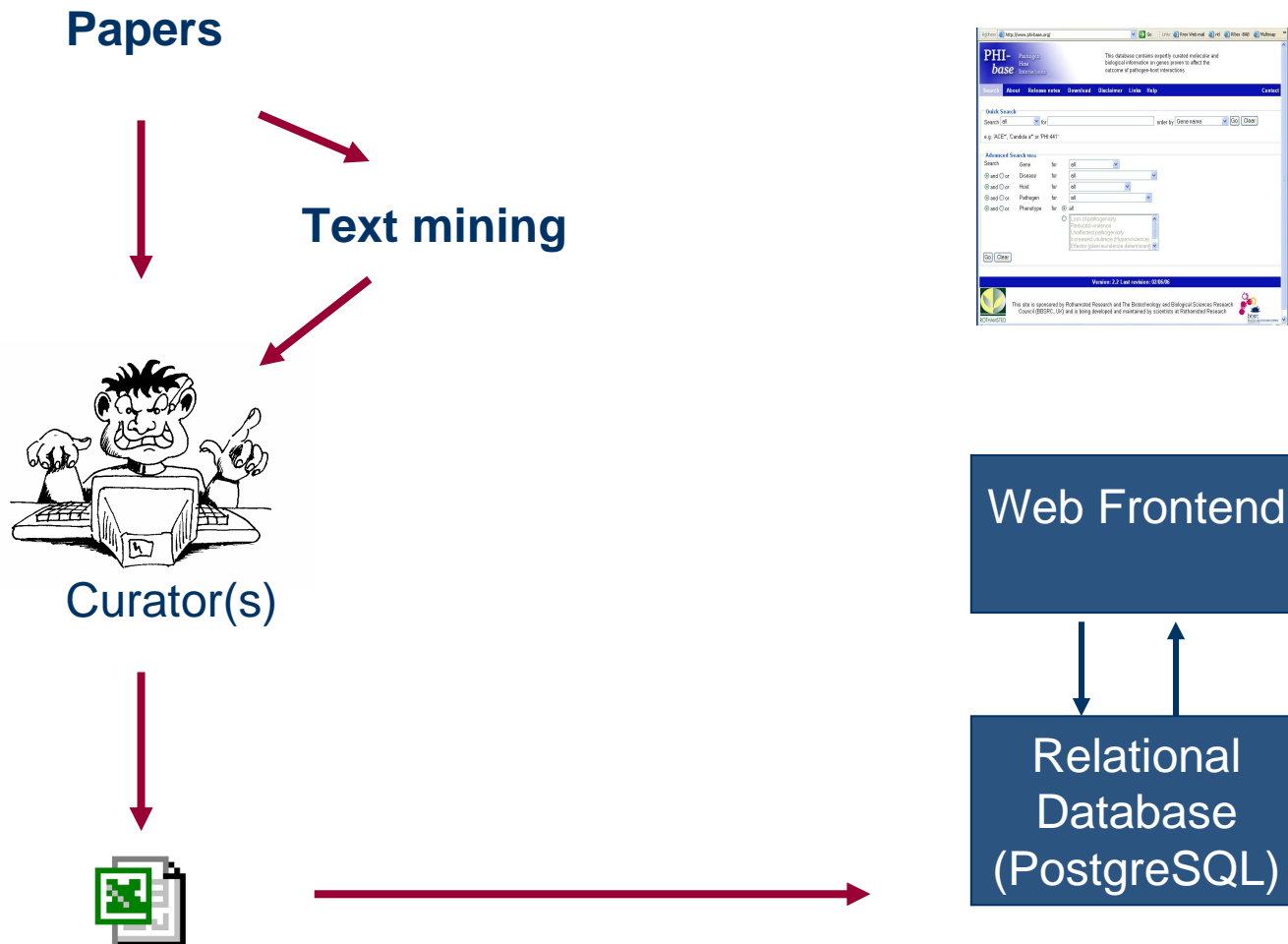


## Original situation

Post-doc and PhD Student curators  
Simple literature search terms  
Read abstracts to select relevant articles  
Read paper to abstract detailed information

Time consuming  
Potential for missing genes  
Free text, no controlled vocab  
No links to other database  
Not scalable  
Capture in spreadsheet not suitable for DB

# Text Mining to Support Curation



# PHI-base

Pathogen  
Host  
Interactions

This database contains expertly curated molecular and biological information on genes proven to affect the outcome of pathogen-host interactions

[Search](#) [About](#) [Release notes](#) [Download](#) [Disclaimer](#) [Links](#) [Help](#) [Contact](#)

### Quick Search

Search  for  order by

e.g. 'ACE\*', 'Candida a\*' or 'PHI:441'

### Advanced Search Menu

Search	Gene	for	<input type="text" value="all"/>
<input checked="" type="radio"/> and <input type="radio"/> or	Disease	for	<input type="text" value="all"/>
<input checked="" type="radio"/> and <input type="radio"/> or	Host	for	<input type="text" value="all"/>
<input checked="" type="radio"/> and <input type="radio"/> or	Pathogen	for	<input type="text" value="all"/>
<input checked="" type="radio"/> and <input type="radio"/> or	Phenotype	for	<input checked="" type="radio"/> all <input type="radio"/> Loss of pathogenicity <input type="radio"/> Reduced virulence <input type="radio"/> Unaffected pathogenicity <input type="radio"/> Increased virulence (Hypervirulence) <input type="radio"/> Effector (plant avirulence determinant)

Version: 2.2 Last revision: 02/06/06



This site is sponsored by Rothamsted Research and The Biotechnology and Biological Sciences Research Council (BBSRC, UK) and is being developed and maintained by scientists at Rothamsted Research



# Text Mining Results

- Compared with manual curators – trying to recreate same content
    - 3 Concept groups: gene symbols, pathogens and hosts
      - Precision 41% (41 / 100 extracted abstracts)  
(60 different genes, 7 new genes)
      - Recall 70% (104 / 150 extracted abstracts)
  - Mixed results:
    - Reduced recall and precision – but not that bad for first attempts with simple term co-occurrence
    - Found new genes
  - Combined manual and text mining
- 

New project in collaboration with National Centre for Text Mining will take this study further

# Is Systems Biology a Paradigm Shift?

# What Characterises Systems Biology Research

- Access to wide variety of data from many different sources
- Wide variety of data analysis methods for different types of data
  - combine and interpret data
- Create structured quantitative model of system –
  - Mathematical – differential equations
  - Computational – Petri nets, Pi Calculus
- Validate quantitative dynamic behaviour of model by simulation

# What Systems Biology Requires

- Open access to life science databases
  - Challenge: number and variety
- Access to scientific literature and especially the quantitative information embedded there
  - Reaction rates, time course information etc

# Particular Challenges

- Integrating data to facilitate analysis and interpretation
- Identification and extraction of relevant information from scientific literature
  - Currently manually intensive and requires moderate domain expertise
- Finding all the information necessary to parameterise highly complex models
  - Parameter estimation methods for under-determined models

# Issues

- Public databanks capture high volume data
  - Generally low “value” until high volume
  - Exception - protein structure database
- Increasing number of databases that synthesize richer views
  - Database equivalent of review
  - E.g. KEGG (Kyoto Encyclopedia of Genes and Genomes), EBI Genome Reviews database
- No general problem to the small volume, high value interpreted data
  - E.g. in supplementary data lodged with journals publishers
- Data in Online Publications
  - Poor links between additional data and text – for data mining
  - Information in other presentation forms – graphs, tables
  - Images

# E-science and Systems Biology

## what is different

- Highly dependent on 3<sup>rd</sup> party “public” data
  - Open access is vital
  - Even for primary data producer in lab – interpretation in context of 3<sup>rd</sup> party data is essential
- Rapid change in methods with higher sensitivity and throughput makes (some) information ephemeral
  - E-science = Ephemeral-science?
  - Cheaper to run experiment again ( gene expression )
- Peer-reviewed literature important but needs are different
  - Online publication model (2 column PDF) unsatisfactory
  - More structure / improved information extraction
  - Methods/protocols/metadata
  - Publications more for scientific career development than as a true record of scientific progress?
- Evolution – not Revolution

# Acknowledgements

- Funding – BBSRC
- Rothamsted Colleagues
  - Jacob Koehler
  - Rainer Winnenber
  - Jan Taubert
  - Tully Yates
  - Peter Heddon
  - Andy Phillips
  - Kim Hammond-Kosack
  - Martin Urban
  - Thomas Baldwin

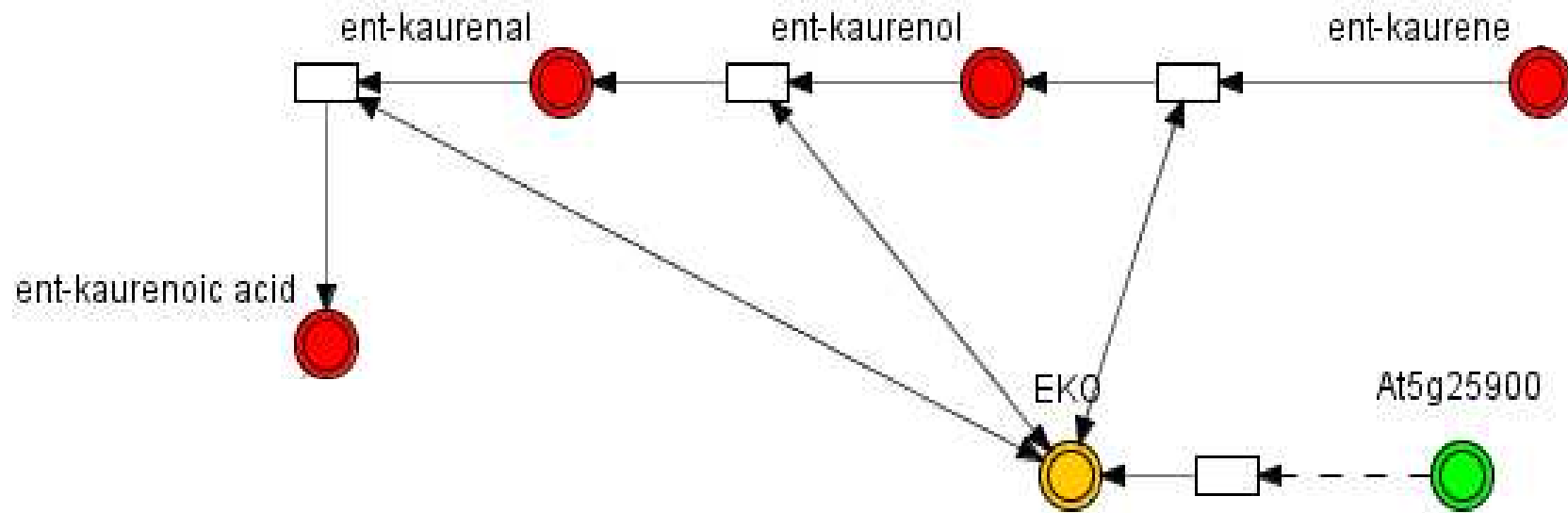




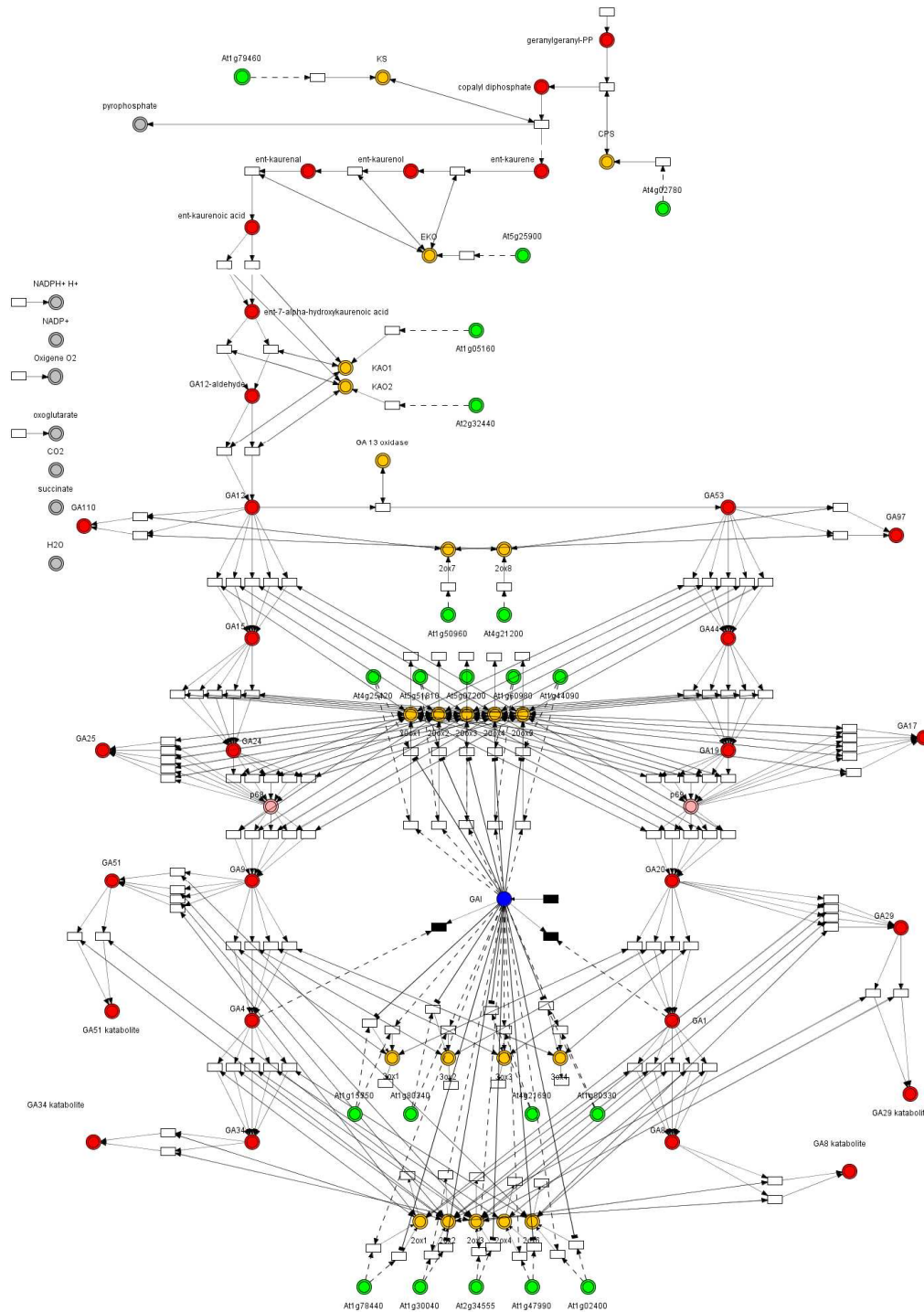
# Modelling Plant Biochemical Systems

- Many groups in RRes study complex signalling and metabolic pathways
- Create mutant plants
  - Single targetted gene knocked-out
  - Phenotype not always easy to predict
- Develop predictive biochemical systems models
- Formalise pathways and biological hypothesis
- Use to predict phenotype from model

# Biological pathways represented as Petri nets

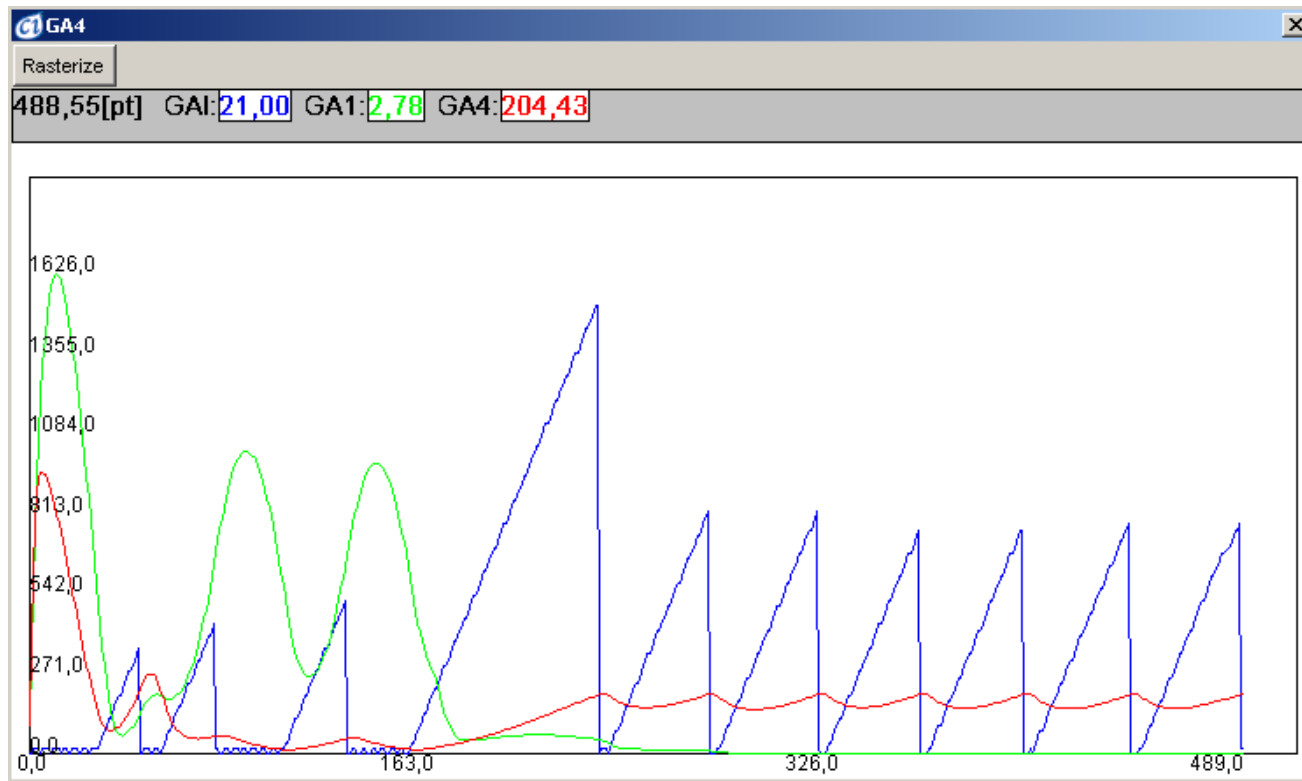


# Gibberellin biosynthesis





# Gibberellin biosynthesis



		GA <sub>4</sub>	GA <sub>34</sub>	GA <sub>1</sub>	GA <sub>8</sub>	GA <sub>20</sub>	GA <sub>19</sub>
Experiment	Control	99	89	38	47	18	111
	GA 2-ox	16	3	3	6	2	2
Petri Net	Control	205	5861	3	4439	3	115
	GA 2-ox	9	453	0.3	0.8	3	102
ODE	Control	103-201	4061	17.8-30.7	7584.6	249.2	0.00052
	GA 2-ox	0.73	2.3	0.6	1.4	98	0.00052