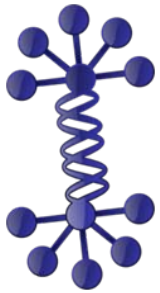


Investigating plant systems using data integration and network analysis

Plant Bioinformatics, Systems and Synthetic Biology Summer School

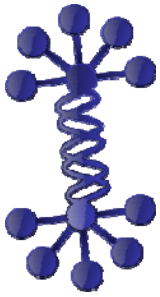
Nottingham, July 30th 2009

Chris Rawlings

Catherine Canevet, Michael Defoin-Platel

ROTHAMSTED RESEARCH

BBSRC bioscience for the future

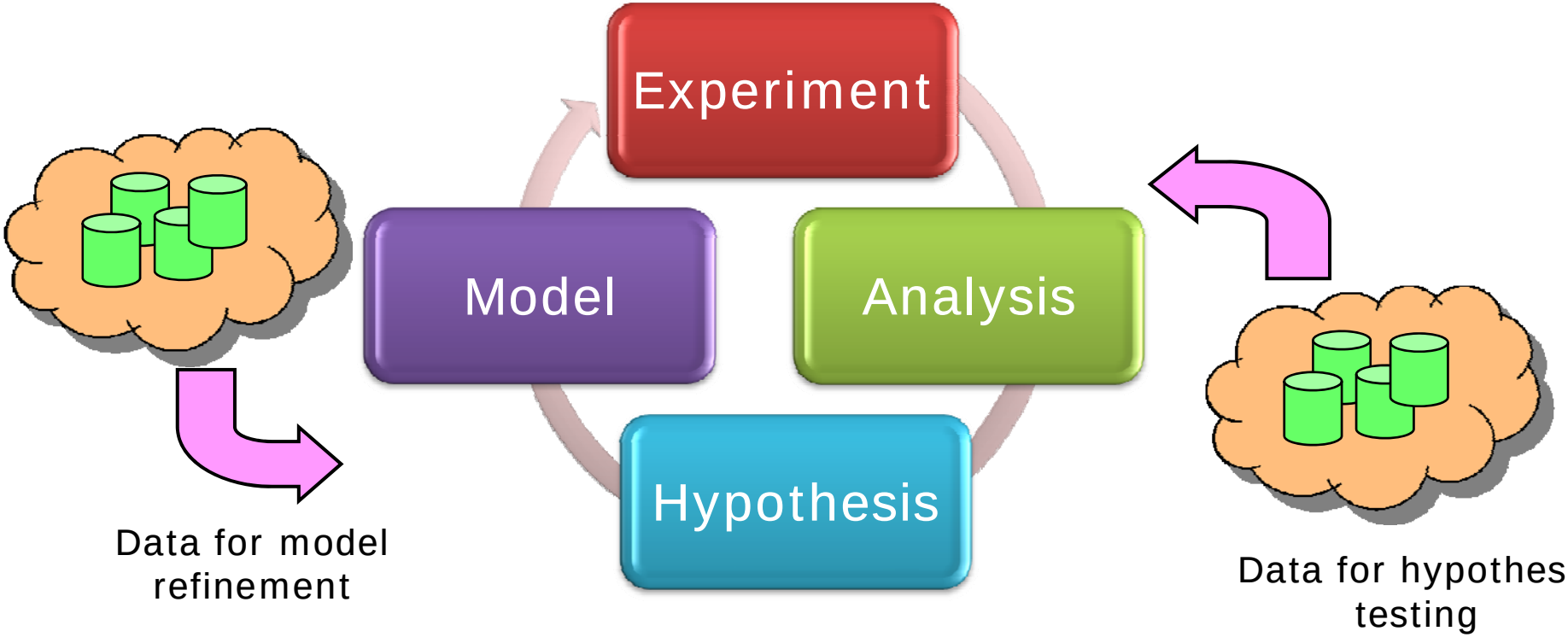


Examples of Plant BioInformatics simple questions requiring data integration

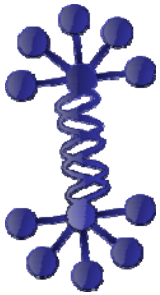
- Which metabolic pathways is a gene or set of genes involved in?
  - Information distributed across several pathway databases
- What transcription factors are involved in that pathway?
  - Information in pathway databases as well as transcription factor databases

ROTHAMSTED RESEARCH

System Biology Needs Data



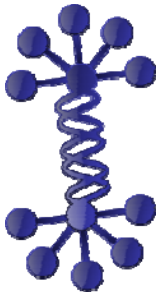
ROTHAMSTED RESEARCH



Outline

- Data integration
  - Needs for data integration
  - Benefits of data integration
  - Challenges
  - Solutions
  - Case study
- Network analysis
  - Definition
  - Visualising biological networks
  - Annotation

ROTHAMSTED RESEARCH



Outline

- Data integration
  - Needs for data integration
  - Benefits of data integration
  - Challenges
  - Solutions
  - Case study
- Network analysis
  - Definition
  - Visualising biological networks
  - Annotation

ROTHAMSTED RESEARCH

Plant databases
many locations, many formats

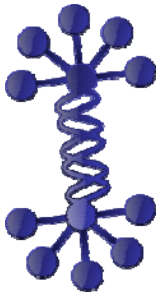
- TAIR
  (The *Arabidopsis* Information Resource)
- AraCyc
  (*Arabidopsis Thaliana* metabolic pathways)
- AtRegNet
  (*Arabidopsis Thaliana* Regulatory Networks)
- Gramene
  (Comparative Grass Genomics)
- Grassius
  (Grass Regulatory Information Services)
- TIGR
  (*Rice* Genome Annotation Resource)

ROTHAMSTED RESEARCH

Plant & crop databases (2/2)

- The Brachypodium distachyon Information Resource
- GrainGenes
- Maize GDB
- Maize Mapping Project
- SGN (Sol Genomics Network)
- SoyBase
- UrMeLDB (Medicago trunculata)
- BeanGenes
- Legume Base
- JGI Poplar database

ROTHAMSTED RESEARCH

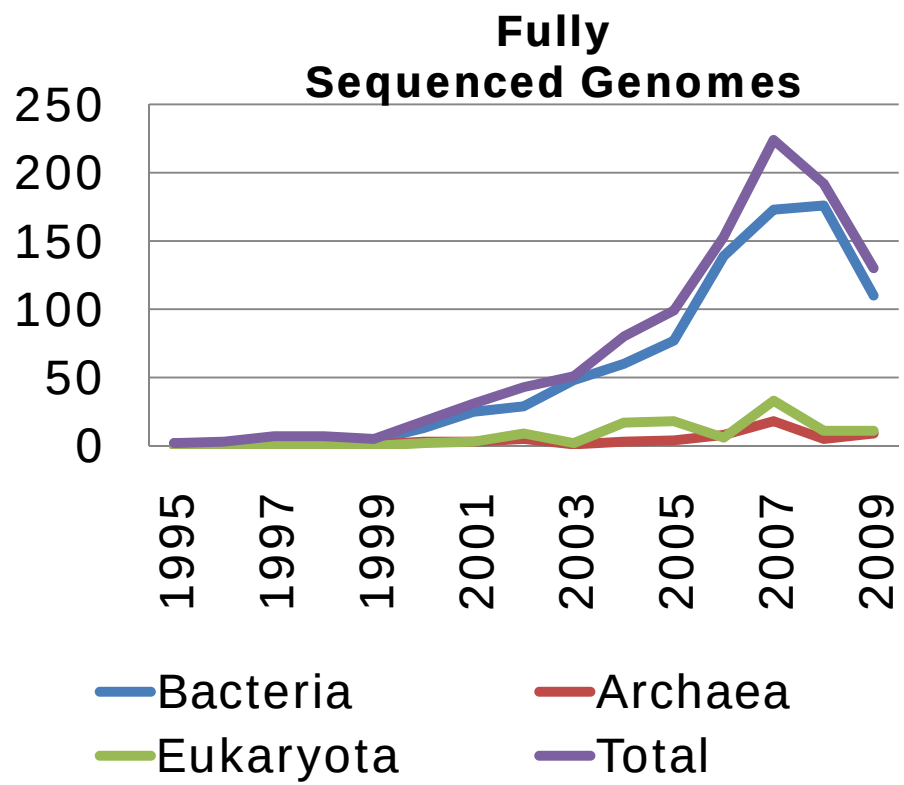


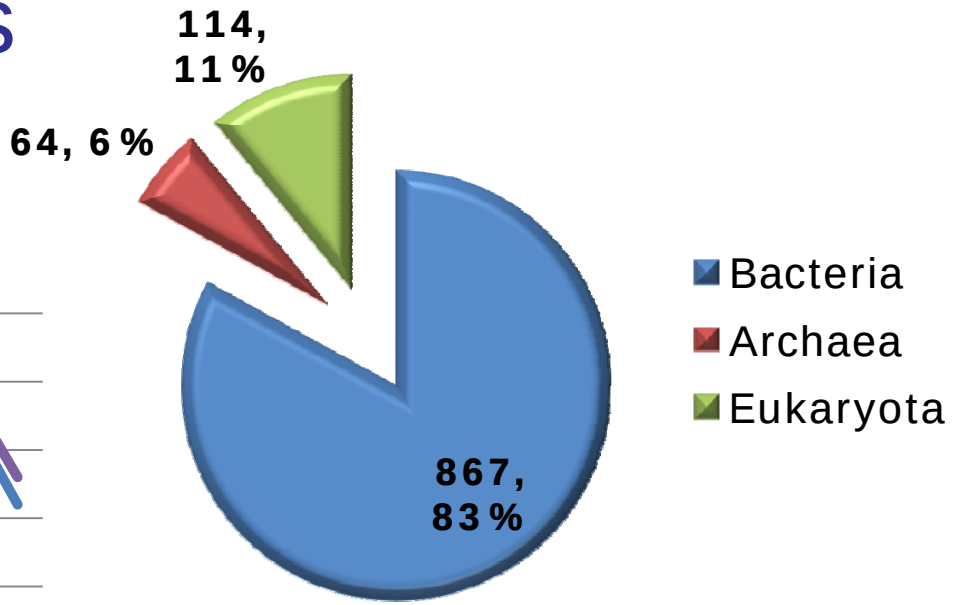
Other bioinformatics databases

- Bcsdb (Bacterial Carbohydrate Structures)
- BioCyc (Pathway/Genome Databases)
- BioGRID (General Repository for Interaction Datasets)
- BRENDA (Enzymes)
- GOA (Gene Ontology Annotation to UniProtKB proteins)
- KEGG (Kyoto Encyclopedia of Genes and Genomes)
- PDB (proteins, nucleic acids and complexes)
- Pfam (protein families)
- SGD (*Saccharomyces* Genome)
- TRANSFAC (transcription factors)
- TRANSPATH (pathways)
- UniProt (proteins)

ROTHAMSTED RESEARCH

Sparseness of plant data (1) – few full genomes





Year of achievement	Plant model organisms
2000	Arabidopsis
2002	Oryza
2006	Populus
2007	Vitis
2008	Physcomitrella

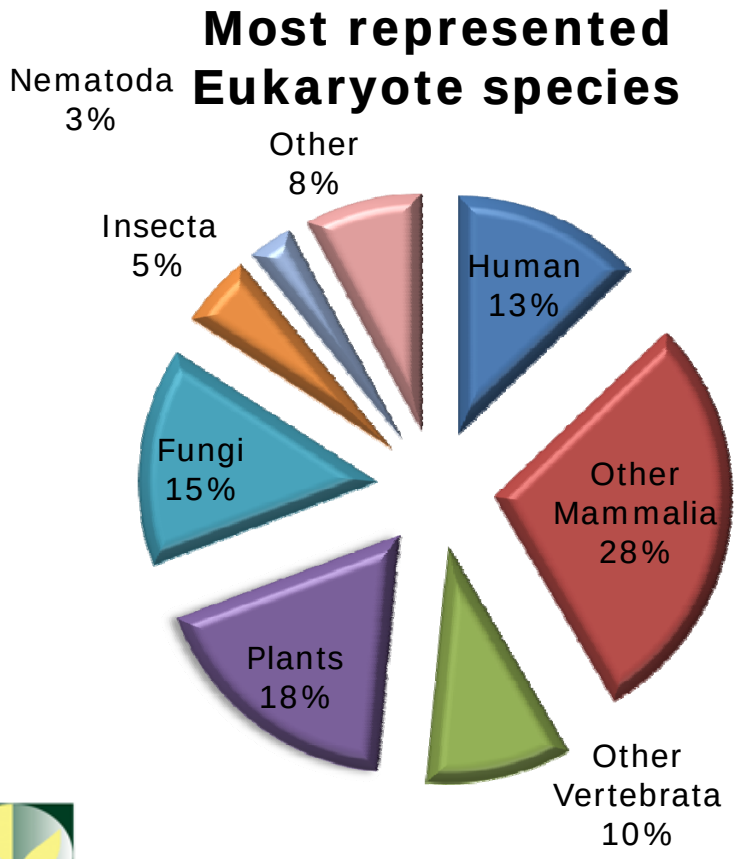
ROTHAMSTED RESEARCH

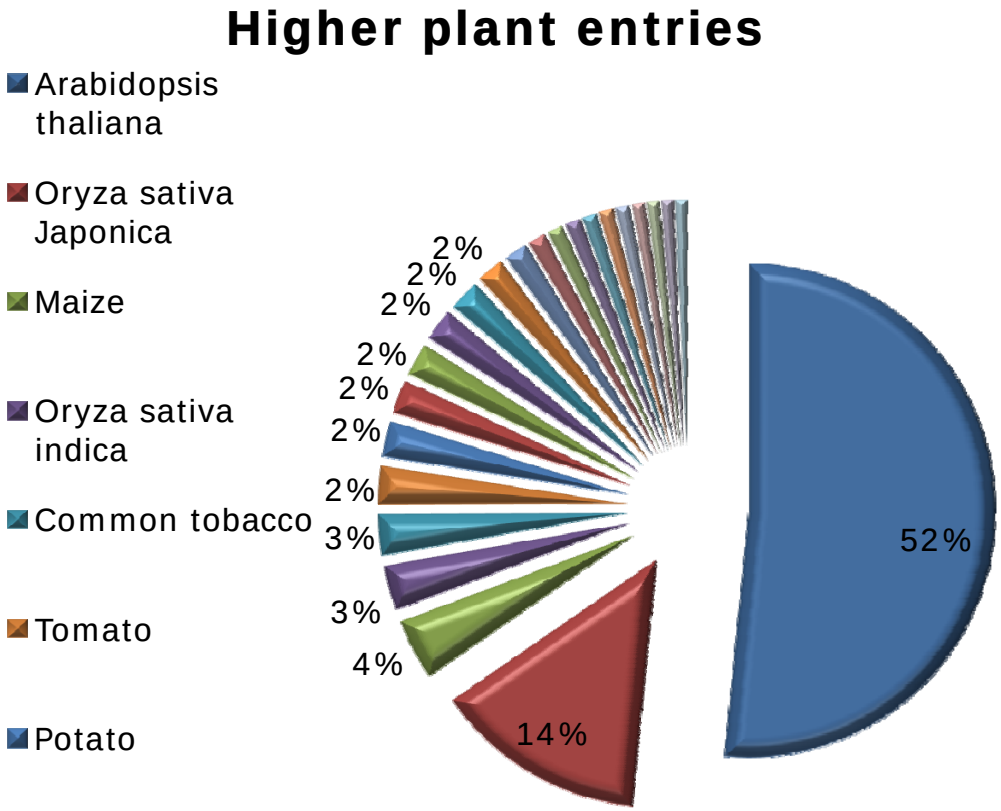
Sparseness of plant data (3)

less sequence data – most from Arabidopsis

In UniProtKB/Swiss-Prot

Eukaryotes represent only 33% of the database



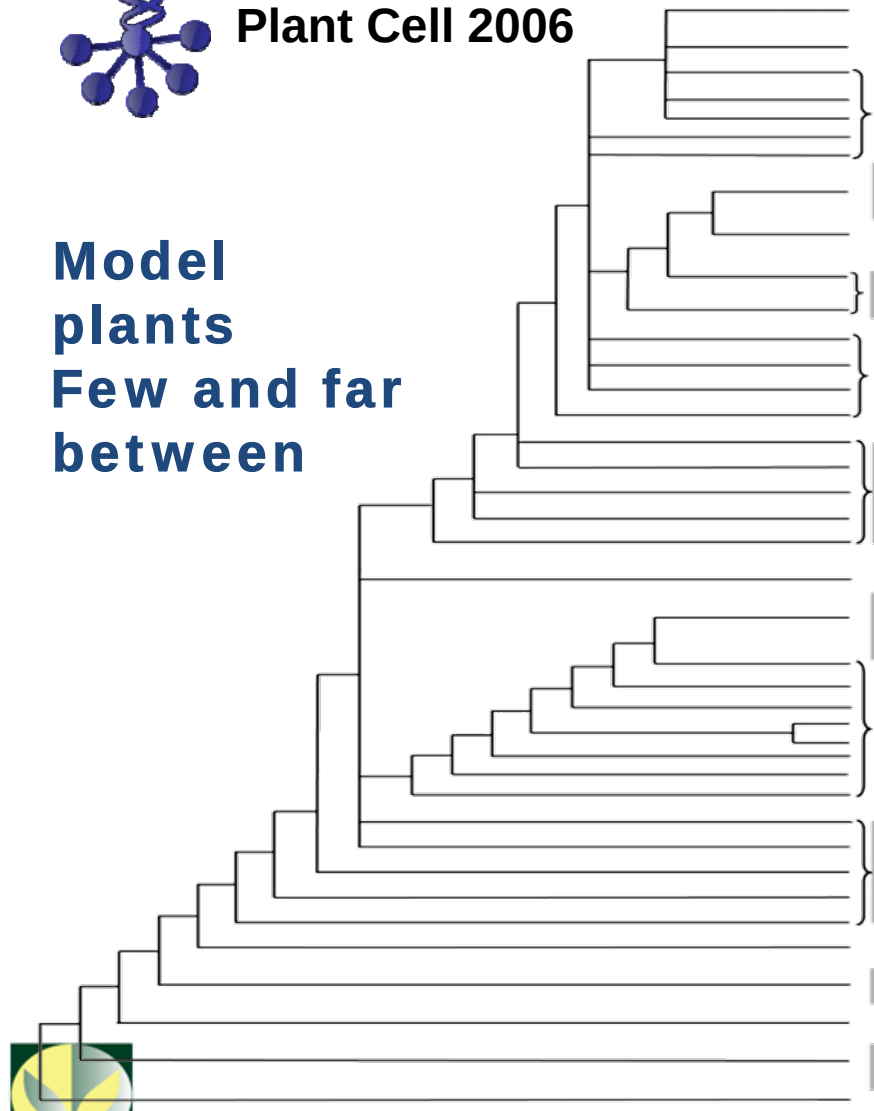


ROTHAMSTED RESEARCH

Sparseness of plant data (2)

Jackson et al.
Plant Cell 2006

Model plants
Few and far between



	Order	Family	Species	Represented Families
Eurosid I (Fabids)	8	64	*Glycine max, Lotus japonicus, Medicago, Populus trichocarpa*	2
Eurosid II (Malvids)	3	34	*Arabidopsis lyrata, A. thaliana, Brassica oleracea, B. rapa, Capsella rubella, Carica papya*	2
Other Rosids	5	35	*Eucalyptus globulus, Vitis vinifera*	2
Euasterid I (Lamiids)	4	40	*Lycopersicon esculentum, Mimulus guttatus, Solanum tuberosum, Triphysaria versicolor*	3
Euasterid II (Campanulids)	4	35		0
Other Asterids	2	30		0
Other core eudicots	4	24		0
Early diverging dicots	2[a]	12		0
Magnoliids	4	18		0
Commelinids	3[b]	29	*Musa acuminata, Oryza sativa, Sorghum bicolor, Triticum aestivum, Zea mays*	2
Other monocots	8	53		0
Basal Angiosperms	1[c]	5		0
Gymnosperms	4	11		0
Ferns	16	55		0
Lycophytes	3	3	*Selaginella moellendorfii*	1
Bryophytes	13	158	*Physcomitrella patens*	1
Green Algae			*Chlamydomonas reinhardtii, Micromonas pusilla, Ostreococcus tauri*	
Total	**84**	**606**		**13**

ROTHAMSTED RESEARCH

Sparseness of plant knowledge (4)

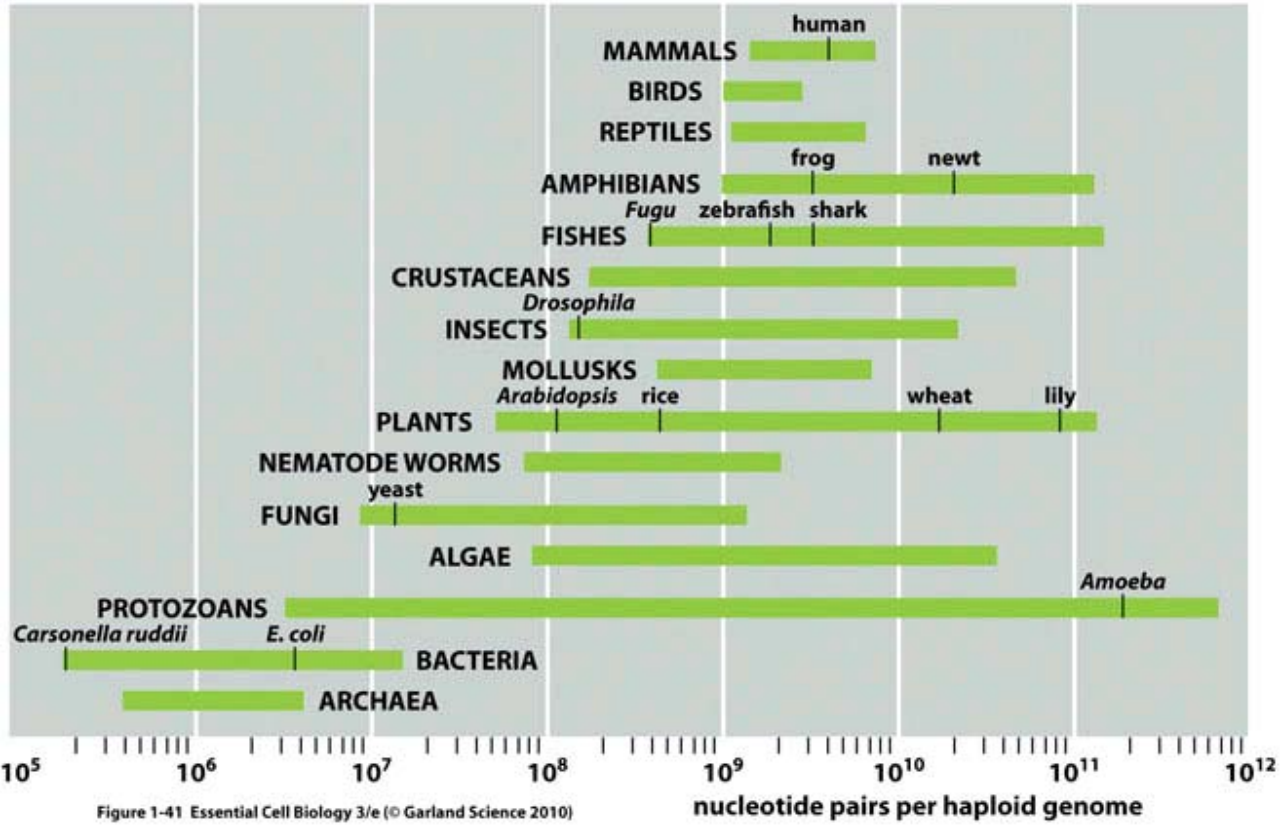


Figure 1-41 Essential Cell Biology 3/e (© Garland Science 2010)

Model plant genomes small and "simple"

TABLE I

Nuclear genome size in different species

Common name	Scientific name	Nuclear genome size (1)
Wheat	*Triticum aestivum*	15,966
Onion	*Allium cepa*	15,290
Garden pea	*Pisum sativum*	3,947
Corn	*Zea mays*	2,292
Asparagus	*Asparagus officinalis*	1,308
Tomato	*Lycopersicum esculentum*	907
Sugarbeet	*Beta vulgaris*	758
Apple	*Malus X domestica*	743
Common bean	*Phaseolus vulgaris*	637
Cantaloupe	*Cucumis melo*	454
Grape	*Vitis vinifera*	483
Man	*Homo sapiens*	2,910

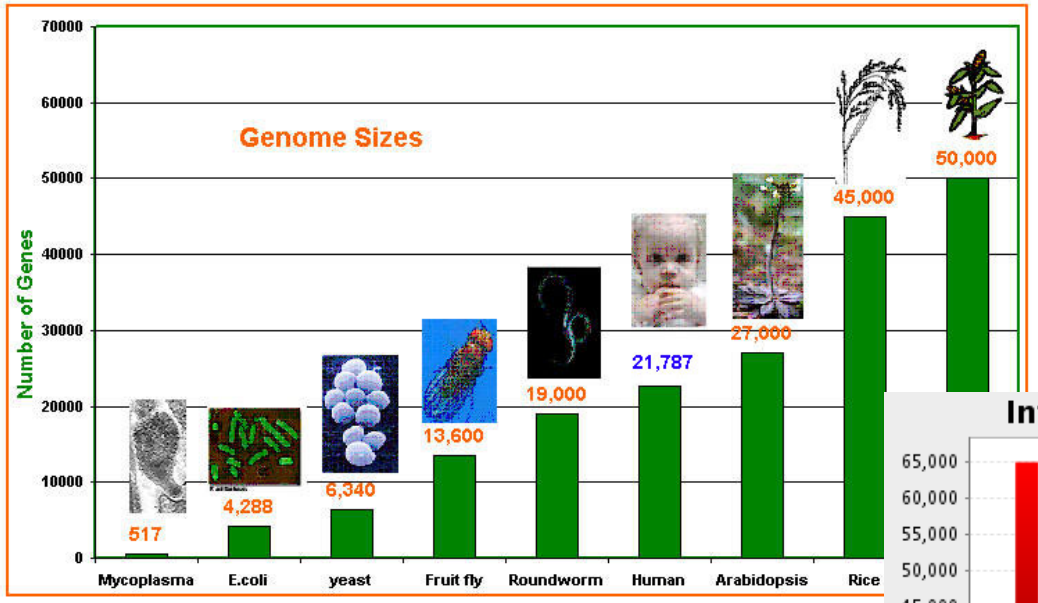
1: Expressed in Megabases (1Mb:1,000,000 bases)

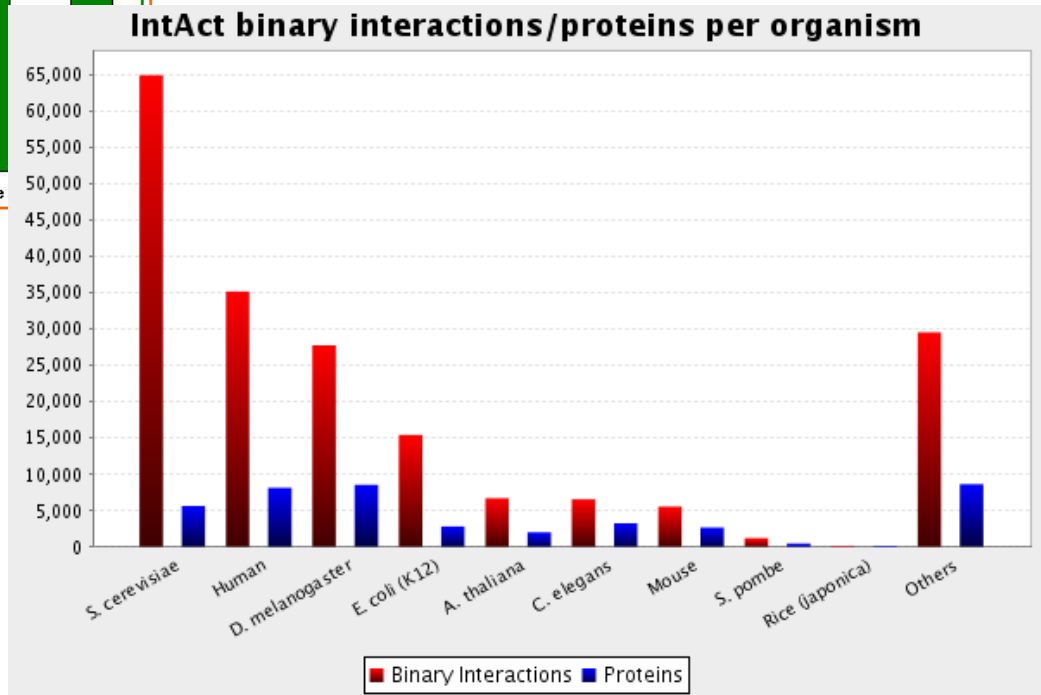
ROTHAMSTED RESEARCH

Sparseness of plant data (5)

Binary interactions from IntAct

S. cerevisiae: 48355
Human: 27969
Fruit fly: 26285
Arabidopsis: 4553
C. elegans: 5738
Mouse: 4454
S. pombe: 685
Rice(japonica): 152





Same story in other 'omics datasets

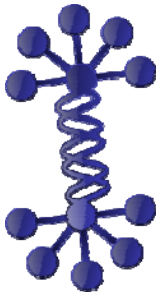
ROTHAMSTED RESEARCH

Conclusion

- Plant and crop scientists, more than others need to make most of all their data
- Exploit data from a greater variety of sources
- Not to mention other challenges
  - Plants in the environment
  - Crop systems – interactions with other organisms

Data Integration

ROTHAMSTED RESEARCH



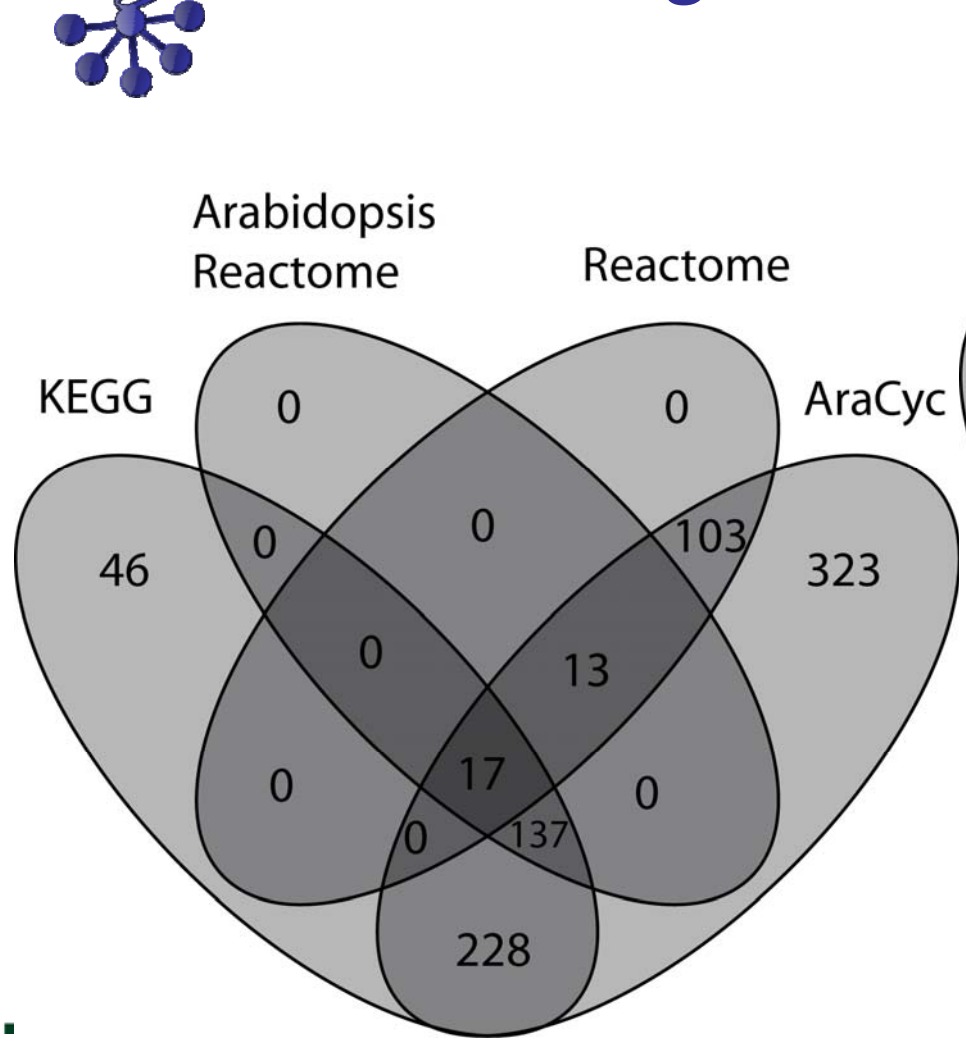
Outline

- Data integration
  - Needs for data integration
  - Benefits of data integration
  - Challenges
  - Solutions
  - Case study

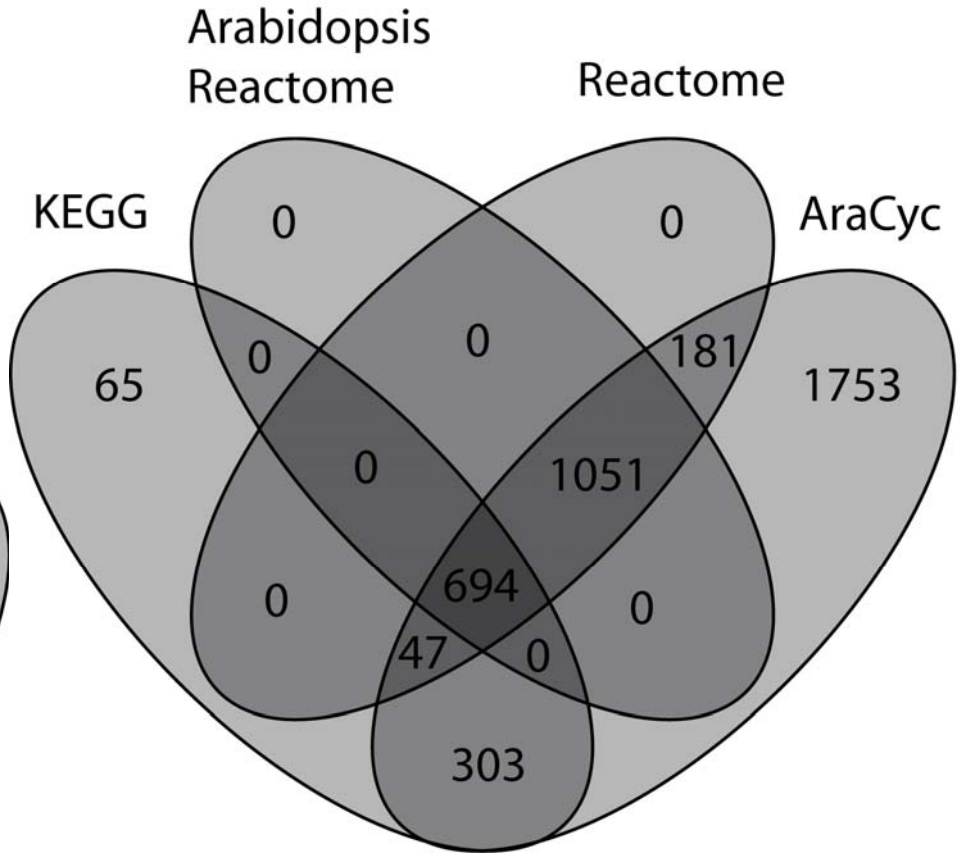
- Network analysis
  - Definition
  - Visualising biological networks
  - Annotation

ROTHAMSTED RESEARCH

Unifying different sources *maximizing coverage*

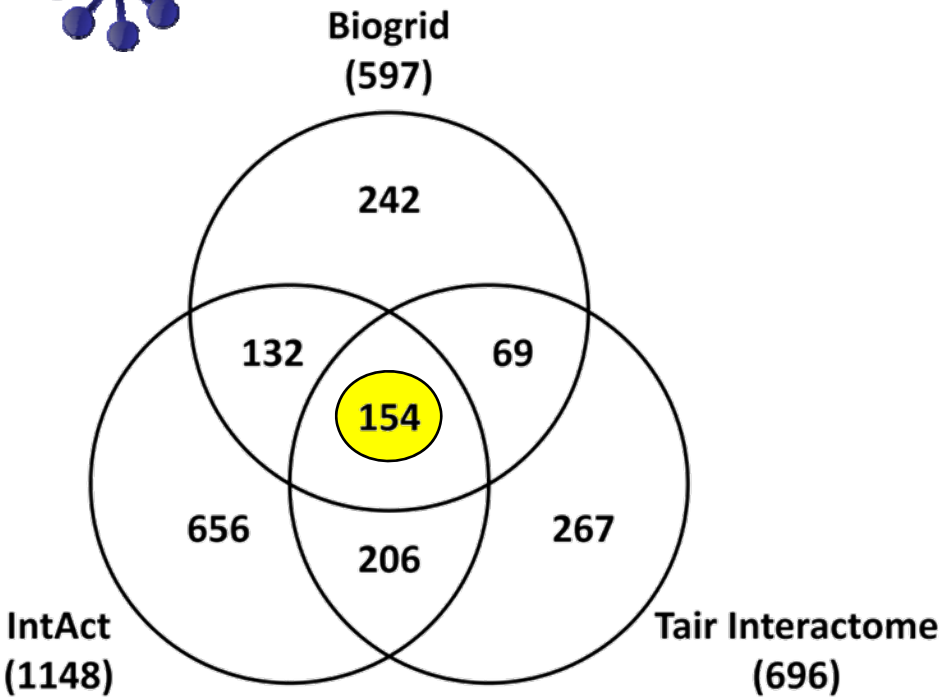


Strict use of EC terms,
e.g. EC1.1.1.6 glycerol dehydrogenase

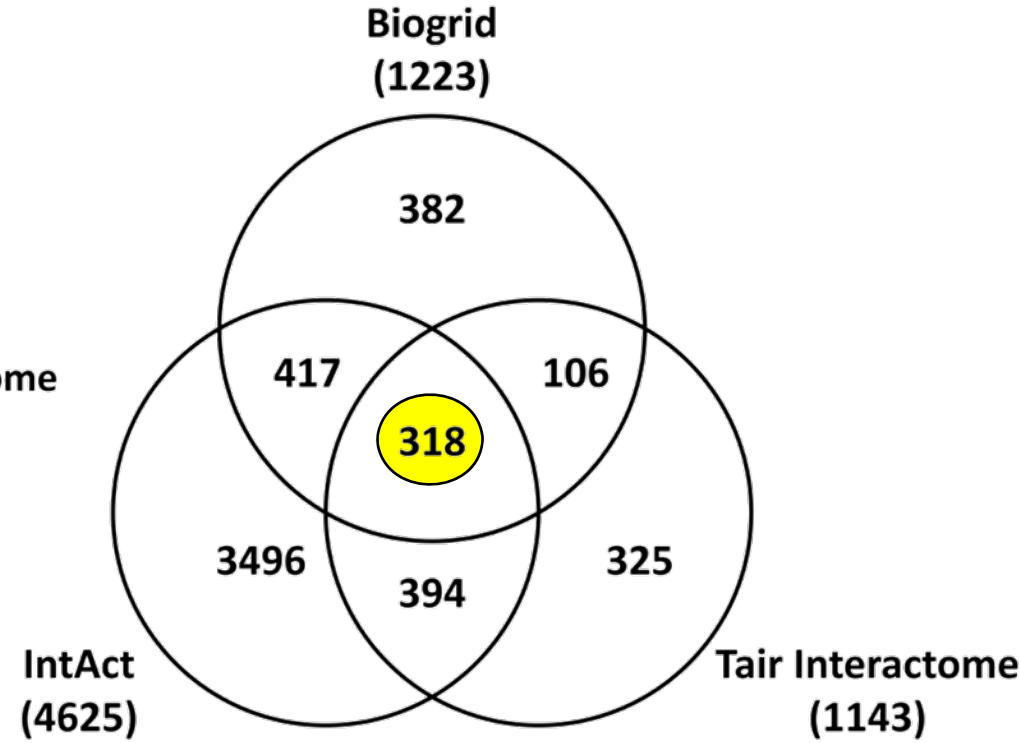


Wider match to enzyme
function EC 1.1.1.-
allow missing substrate term

Consolidating interaction data *maximizing confidence*



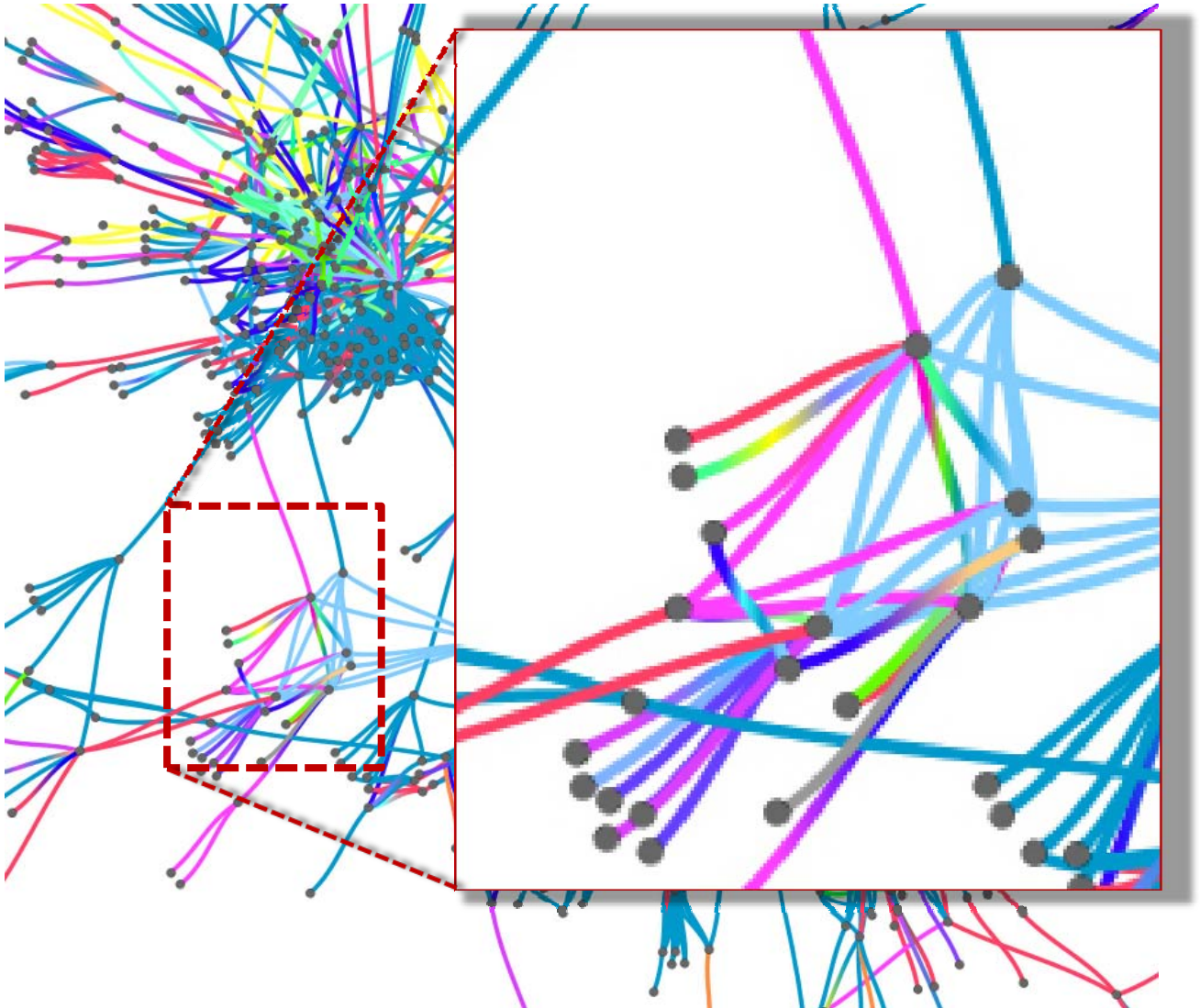
Protein Protein Interactions (PPI)



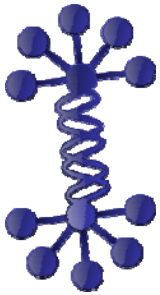
Proteins

ROTHAMSTED RESEARCH

Consolidating knowledge



ROTHAMSTED RESEARCH

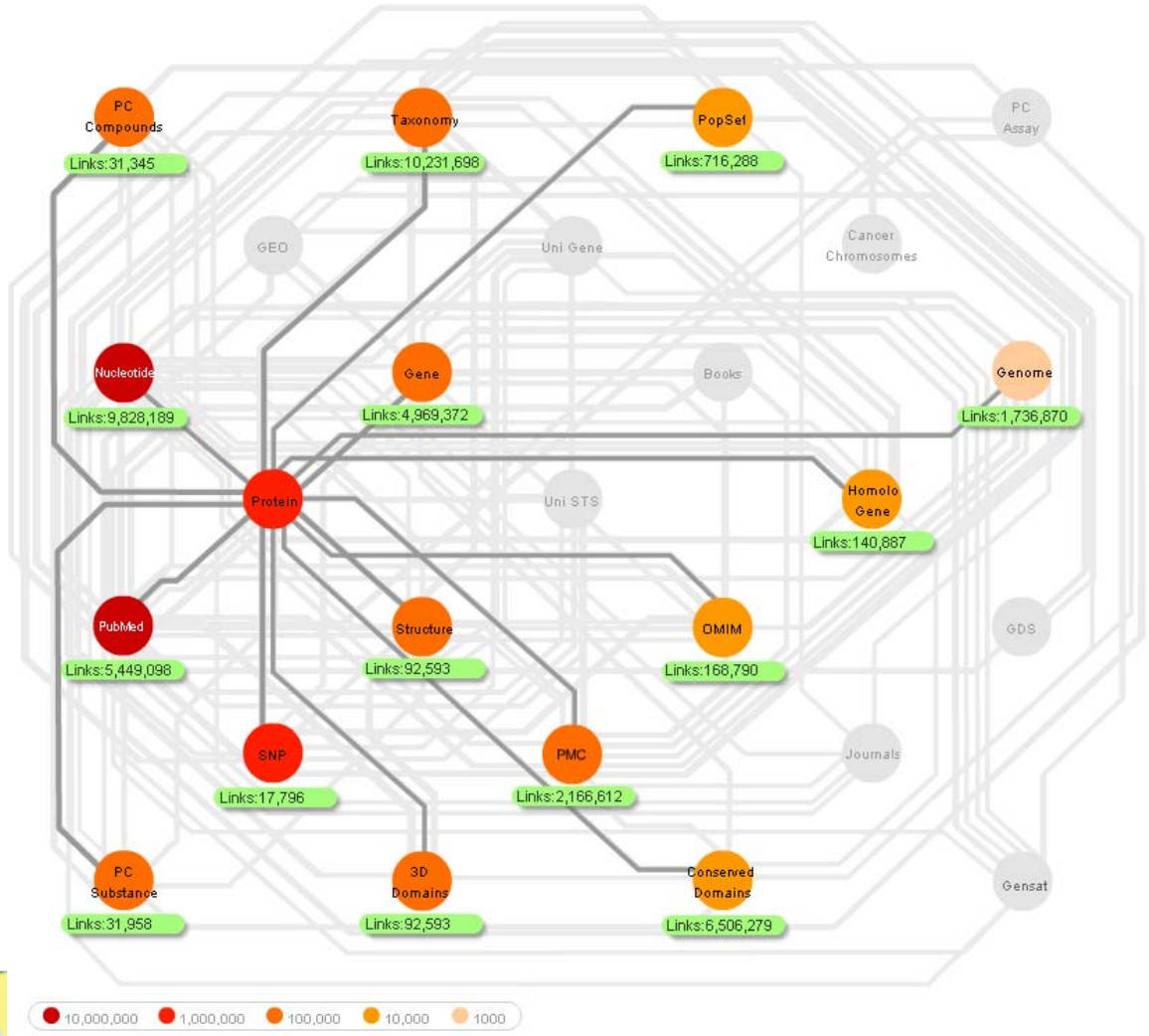


Outline

- Data integration
  - Needs for data integration
  - Benefits of data integration
  - Challenges
  - Solutions
  - Case study
- Network analysis
  - Definition
  - Visualising biological networks
  - Annotation

ROTHAMSTED RESEARCH

Syntactic integration challenge



http://www.ncbi.nlm.nih.gov/Database

Over **1000 databases** freely available to public

Over **60 million sequences** in GenBank

Over **870 complete genomes** and many ongoing projects

Over **17 million citations** in PubMed

PubMed growth by 600,000 publications each year

Integration of Life Science data sources is essential for Systems Biology research

ROTHAMSTED RESEARCH

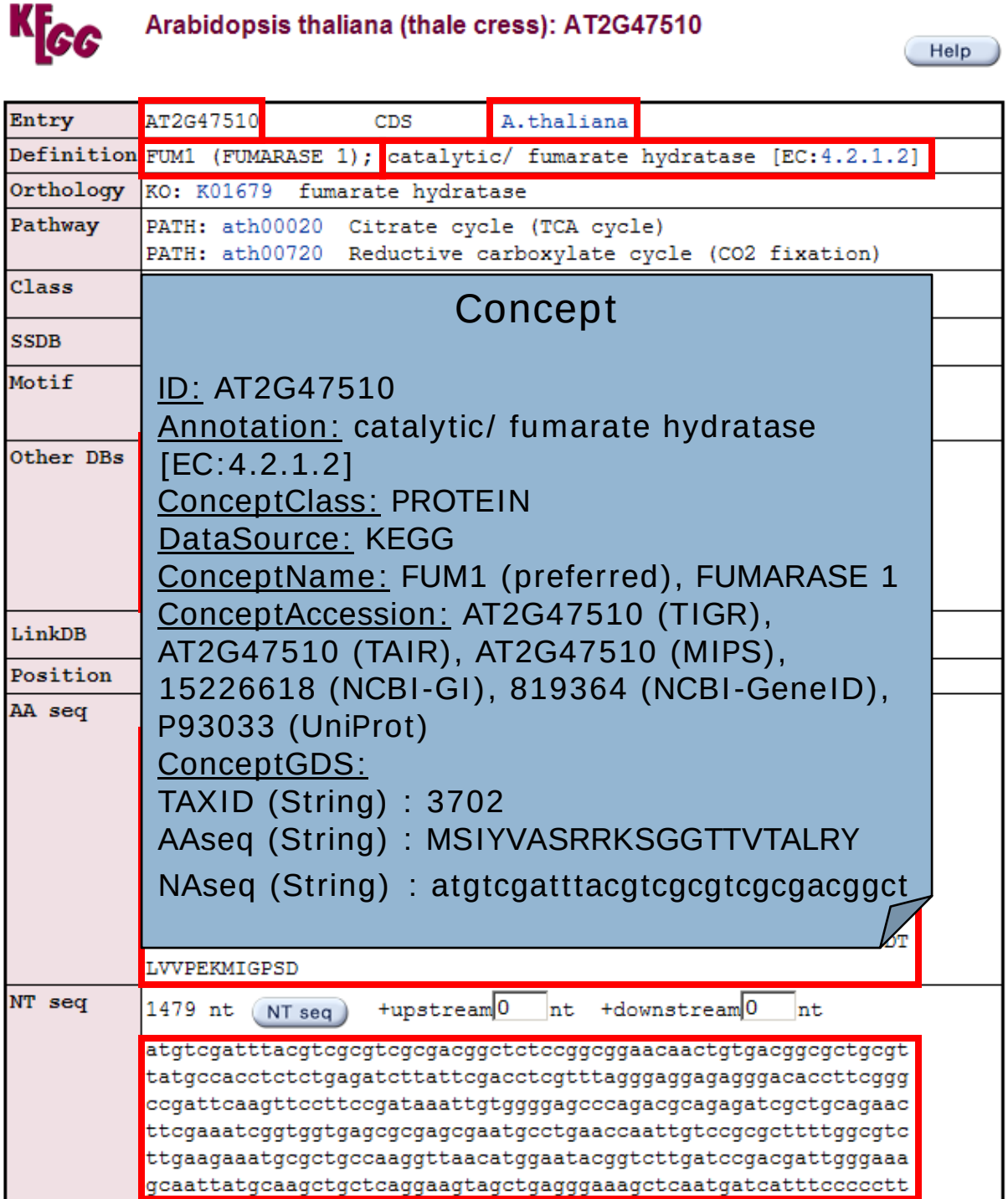
Syntactic integration example

- Databases – widely different formats for same information
  - KEGG (Kyoto Encyclopedia of Genes and Genomes, www.genome.jp/kegg/)
  - TAIR (The *Arabidopsis* Information Resource, www.arabidopsis.org)

- Screenshots for same entry
  - Protein FUMARASE 1
  - *Arabidopsis* AT2G47510

ROTHAMSTED RESEARCH

KEGG - Kyoto Encyclopedia of Genes and Genomes



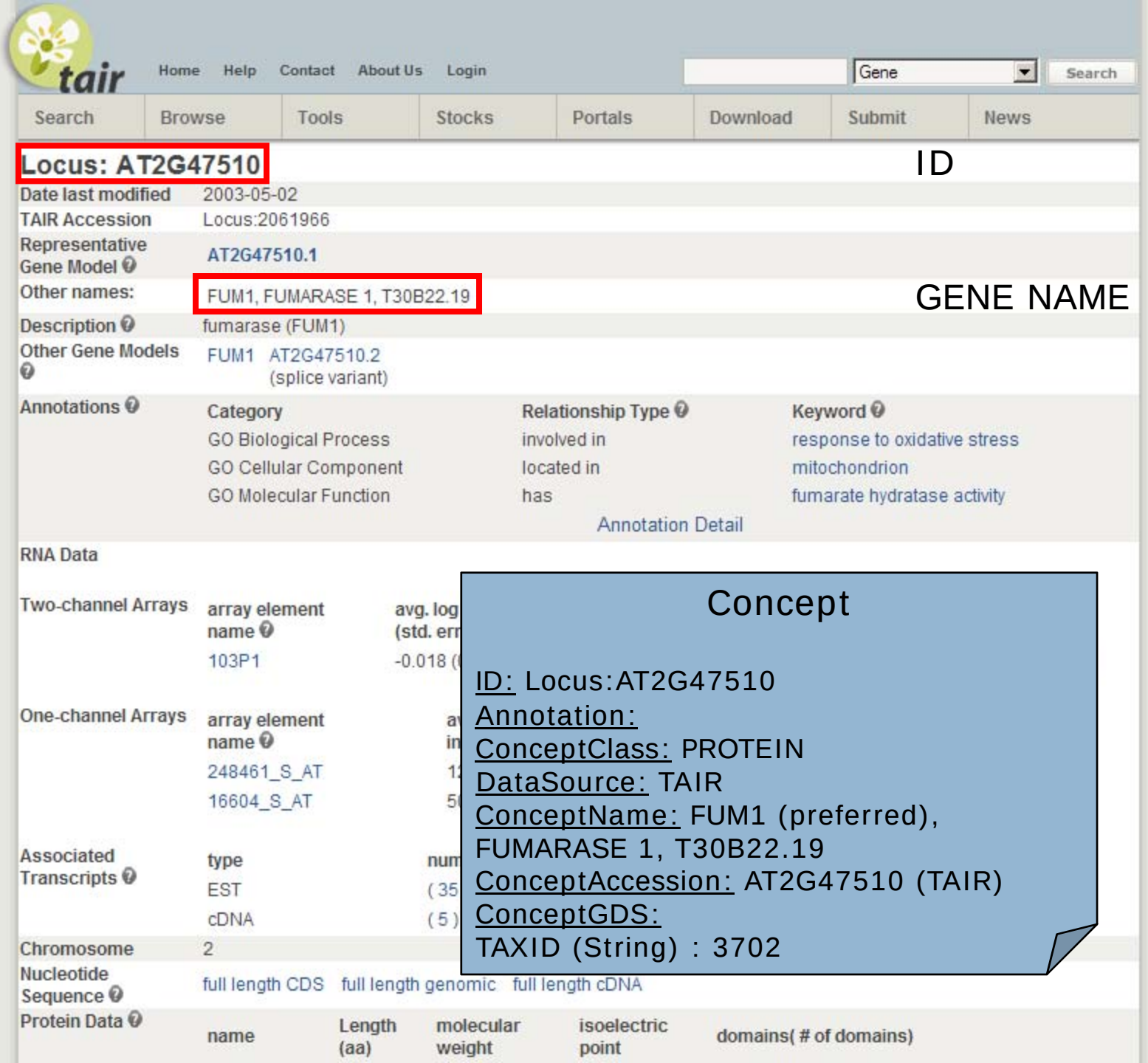
ID/ SPECIES GENE NAME / FUNCTION

REFERENCES TO OTHER DATABASES

PROTEIN SEQUENCE

GENE SEQUENCE

TAIR - The Arabidopsis Information Resource



Semantic Integration challenge

- Same concept different names
  - synonyms
  - ontologies
- Same name different concepts
  - homographs
  - Will expand on these issues later

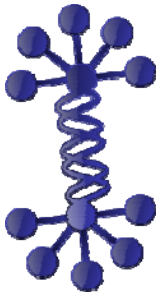
Plant AmiGO Ontology DATABASE: PO_0205

inflorescence

Accession: PO:0009049
Aspect: plant structure
Synonyms:
cob (sensu sorghum)
corymb
cyme
dichasium
drepanium
helicoid cyme
monochasium
panicle
raceme
rhipidium
scorpioid cyme
spike (sensu Triticeae)
umbel
verticillaster
Definition:
That part of the axial system of plants above the uppermost foliage leaf/pair of foliage leaves that bears flowers.
Comment:
Some plants have only solitary flowers, e.g. Magnolia.

Ear

ROTHAMSTED RESEARCH



Outline

- Data integration
  - Needs for data integration
  - Benefits of data integration
  - Challenges
  - Solutions
  - Case study

- Network analysis
  - Definition
  - Visualising biological networks
  - Annotation

ROTHAMSTED RESEARCH

Underpinning Data Integration

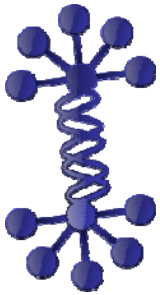
Controlled vocabularies & Ontologies

- List of standardized terms or descriptors whose meanings are specifically defined or authorized by a standards organization

- EC nomenclature, Gene nomenclature
  - http://www.arabidopsis.org/nomencl.html
- Compendium of nomenclature resources
  - http://www.expasy.ch/cgi-bin/lists?nomlist.txt

- CVs help tackle consistency within species/group/taxa

ROTHAMSTED RESEARCH



Underpinning Data Integration

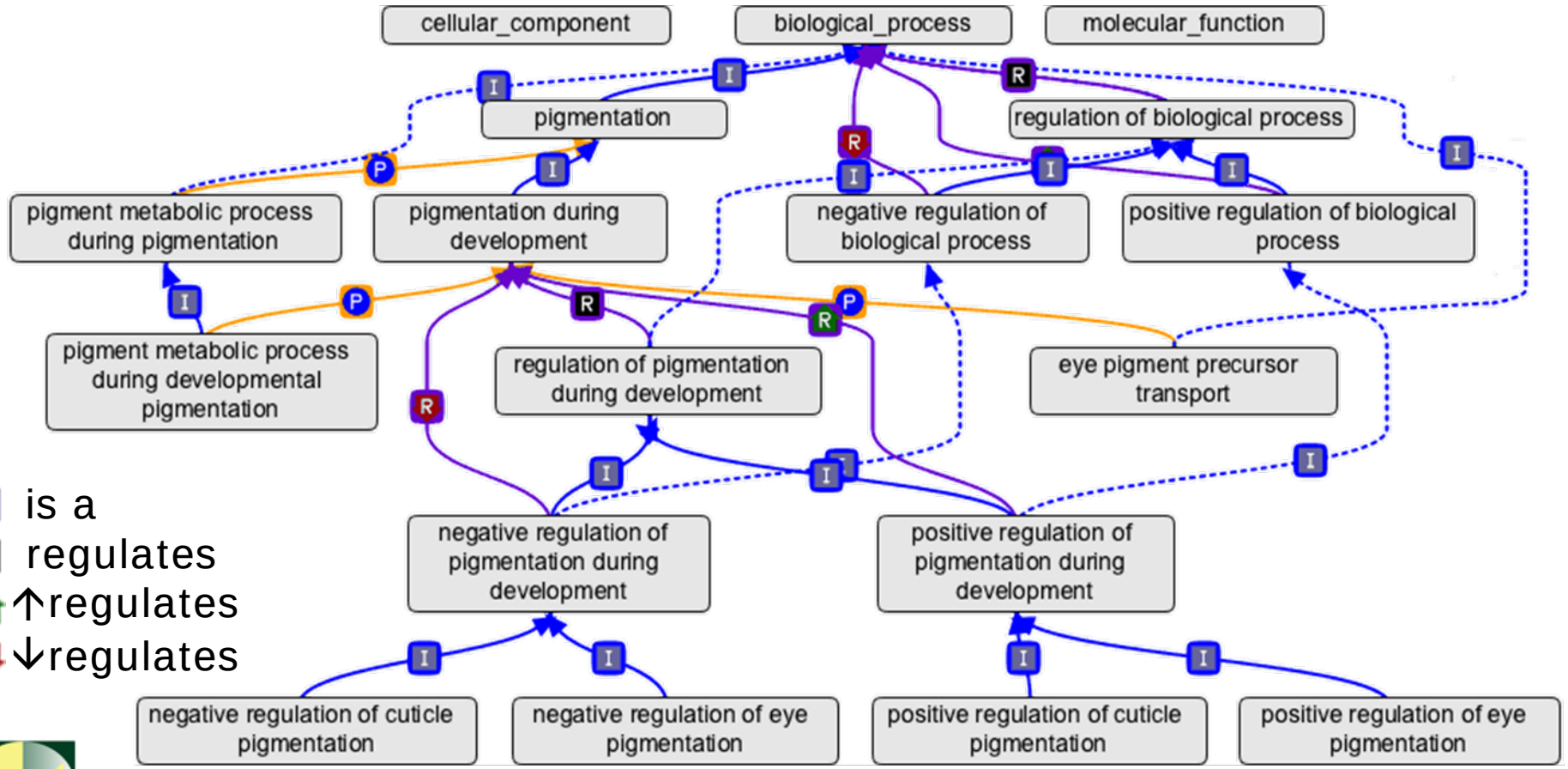
Ontologies

- an **ontology** is a formal representation of a set of concepts within a domain and the relationships between those concepts.
- A semantic structure
- Generally hierarchical
- Provides mechanisms for standardizing across species/groups/taxa

ROTHAMSTED RESEARCH

Underpinning Data Integration

Example: The Gene Ontology

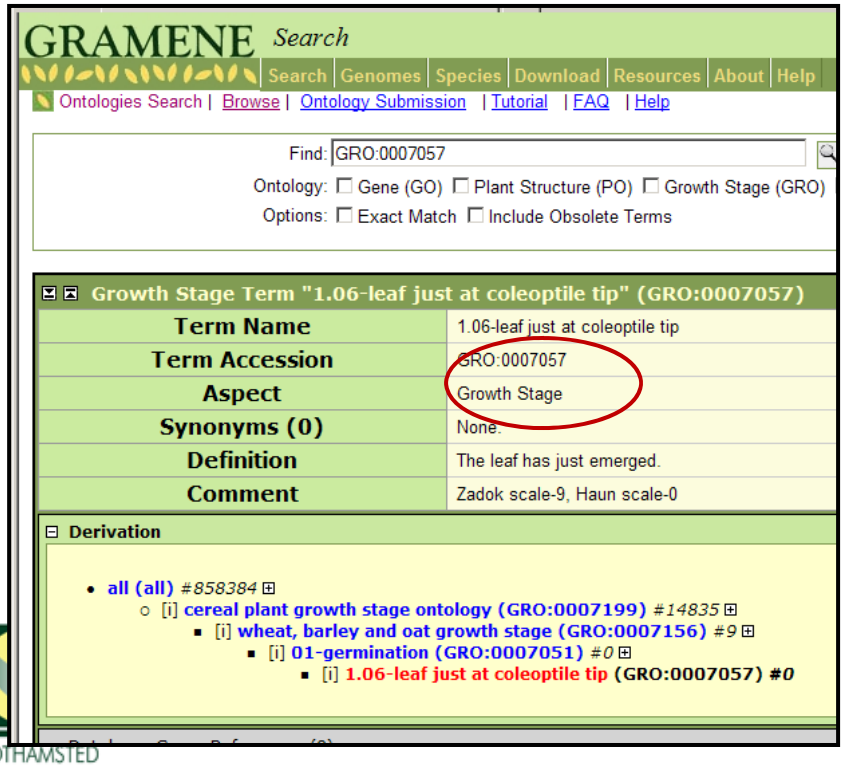


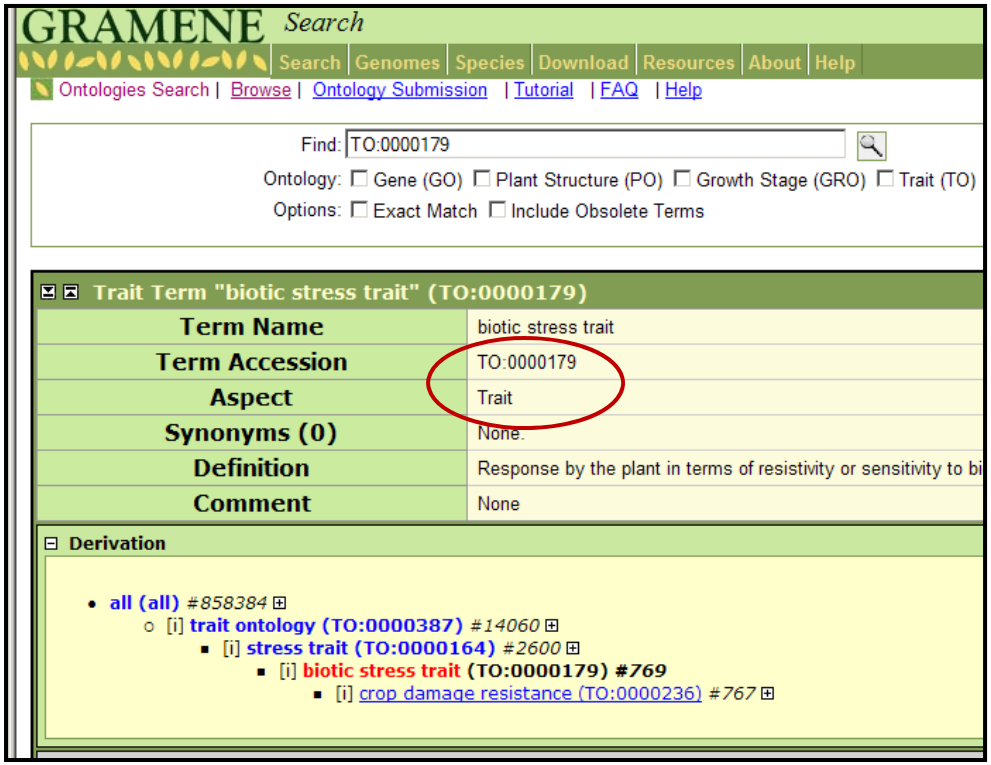
ROTHAMSTED RESEARCH

Plant Ontologies

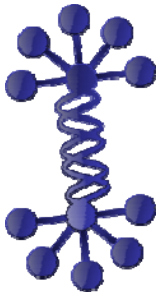
- http://www.gramene.org/plant_ontology/

- Plant structure (PO), Growth stage (GRO)
- Trait (TO), Environment (EO)





ROTHAMSTED RESEARCH



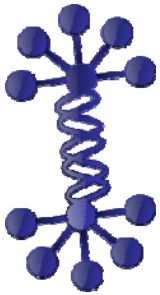
Underpinning Data Integration

Data Standards - Markup Languages

- Extensible Markup Language (XML) facilitates the sharing of data across heterogeneous computer systems and improves the consistency

Markup Language	Purpose
AGAVE	Genomic annotation and visualization
BioML	Experimental information for biopolymers
BSML	Genomic sequences and biological function
CML	Management of molecular information
MAGE-ML	Microarray gene expression data exchange
SBML	Systems biology and biochemical networks

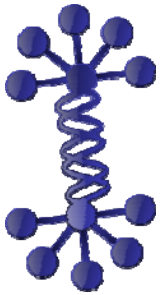
ROTHAMSTED RESEARCH



Minimum Information for Biological and Biomedical Investigations

- Minimum Information Standards for Metadata
  - Data about data
- MIBBI Initiative
- http://www.mibbi.org/
  - MIAME, MIAPE, MIQAS etc
  - Microarray, Proteomics, QTL and Association Experiment

ROTHAMSTED RESEARCH

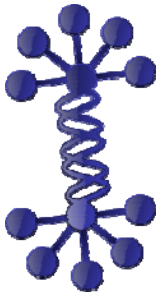


Standards and Underpinnings

- Nomenclatures, Controlled Vocabularies
- Ontologies
- Metadata

- All help tame the complexity of data integration – but they are not enough
  - Legacy information
  - Not widely followed
  - Competing approaches
  - Data integration solutions needed
  - Has been major Bioinformatics challenge

ROTHAMSTED RESEARCH



Solutions for Data Integration (1/4)

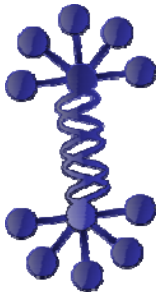
Workflows

- techniques for enacting series of linked processes, useful to systematically automate protocols in bioinformatics
- e.g. InfoSense, Taverna
- generally hard to write – almost programming

Mashups

- Data from different Web services or RSS-feeds are selected by the user and then ‘‘mashed” to form a new Web application
- e.g. UniProt DASTY
- more aggregation than integration

ROTHAMSTED RESEARCH



Solutions for Data Integration (2/4)

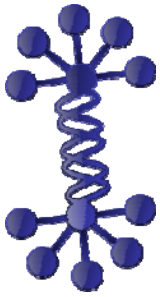
Service oriented architectures

- interconnection of data sources
- e.g. CORBA, Web services
- only “plumbing” techniques
- often poorly documented and constructed and only for programmers

Link integration

- cross-reference data entry from different data sources
- e.g. SRS (40% EMBL-EBI traffic), Entrez, Integr8
- problems with name clashes, ambiguities and updates

ROTHAMSTED RESEARCH



Solutions for Data Integration (3/4)

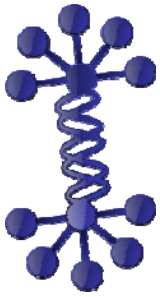
Data Warehousing

- data sources are processed and combined into a new data model to form a new data source
- e.g. eFungi, ATLAS, GIMS
- toolkit: GMOD, BioMART, Intermine, BioWarehouse
- high building and maintenance costs

View integration (“virtual warehouse”)

- data are kept in the original sources but appear to be in a single database
- e.g. BioZon
- complex and slow for users and developers

ROTHAMSTED RESEARCH



Solutions for Data Integration (4/4)

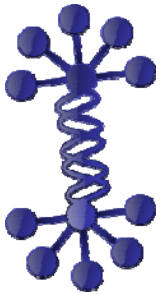
Model-driven service oriented architecture

- data resources and tools are obliged to adhere to a designed model
- e.g. caBIG
- only possible in tightly coupled systems – often difficult to achieve consensus

Integration applications

- built specifically to integrate data
- e.g., ToolBus, Youtopia, **Ondex**
- more often for a single application domain

ROTHAMSTED RESEARCH

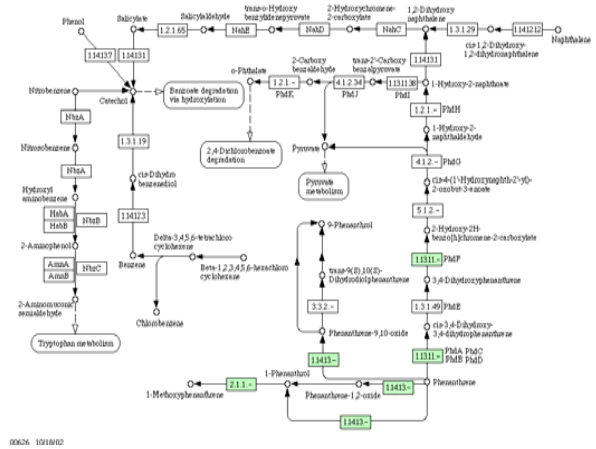


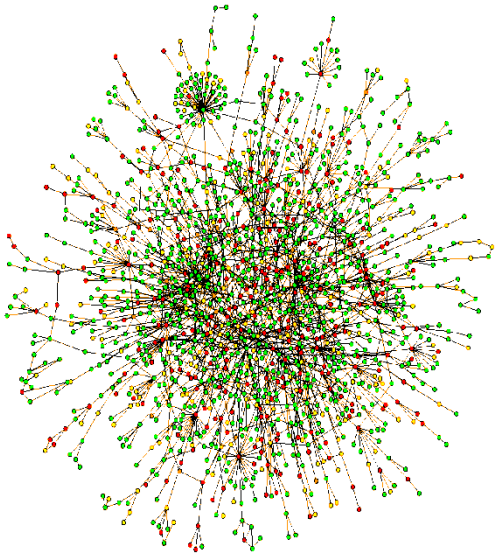
Ondex as an example data integration system

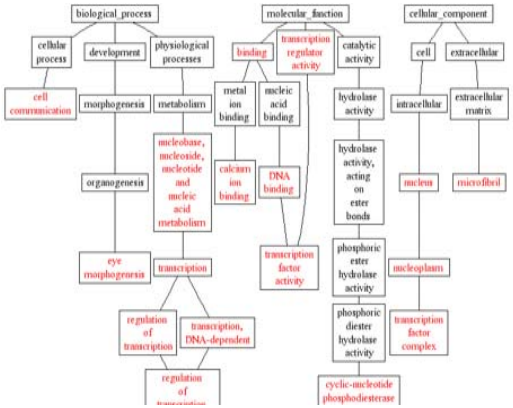
- Three closely coupled aspects:
  - Data input – technical integration and transformation into data domain graph
  - Mapping across graphs – semantic integration
  - Data visualisation and quantitative analysis
- In use in all BBSRC-funded systems biology Centres
- Open source

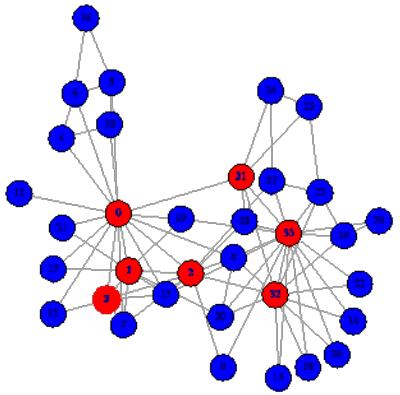
ROTHAMSTED RESEARCH

Everything is a network



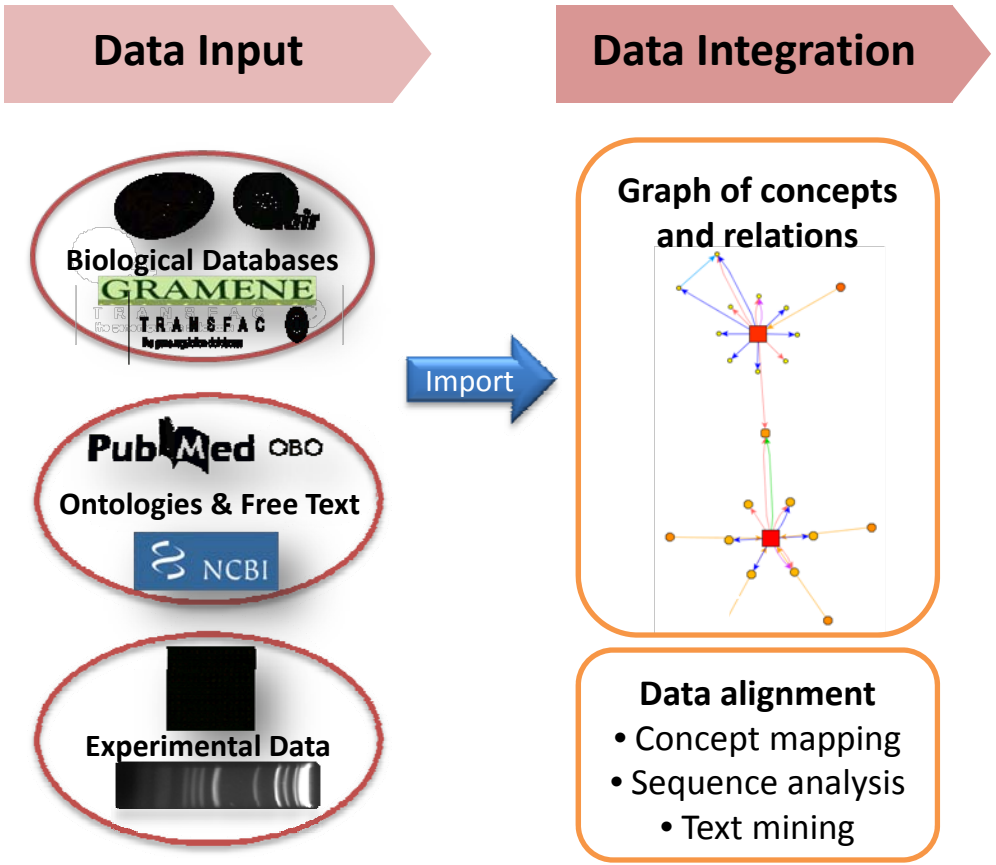






ROTHAMSTED RESEARCH

Data integration in Ondex

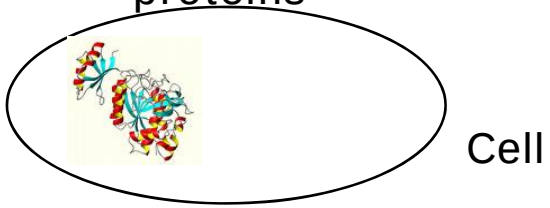


ROTHAMSTED RESEARCH

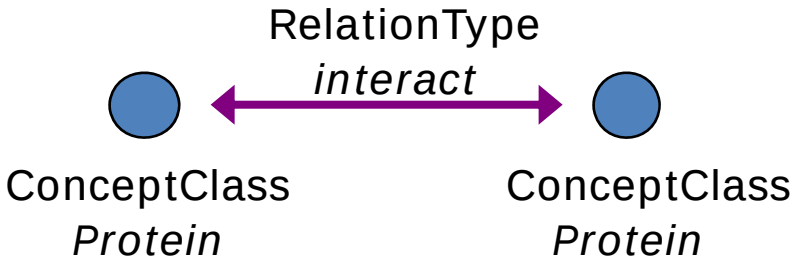
Concepts and relations (1/2)

Protein – Protein interaction network (PPI)

Cellular location of proteins



e.g. Network of Concepts and Relations

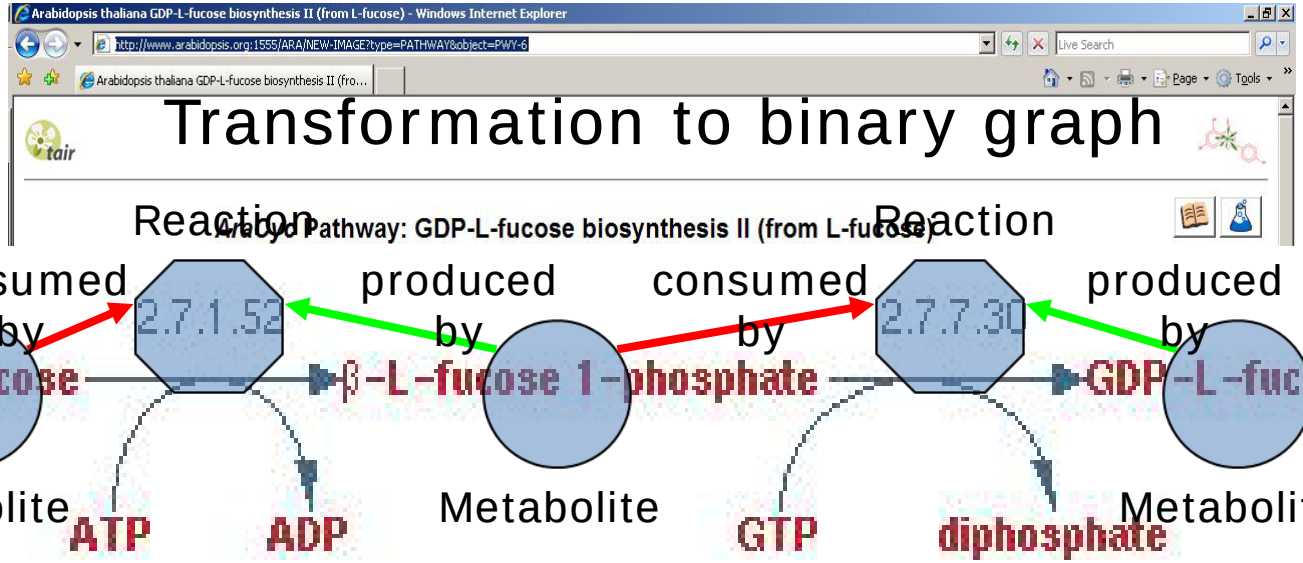


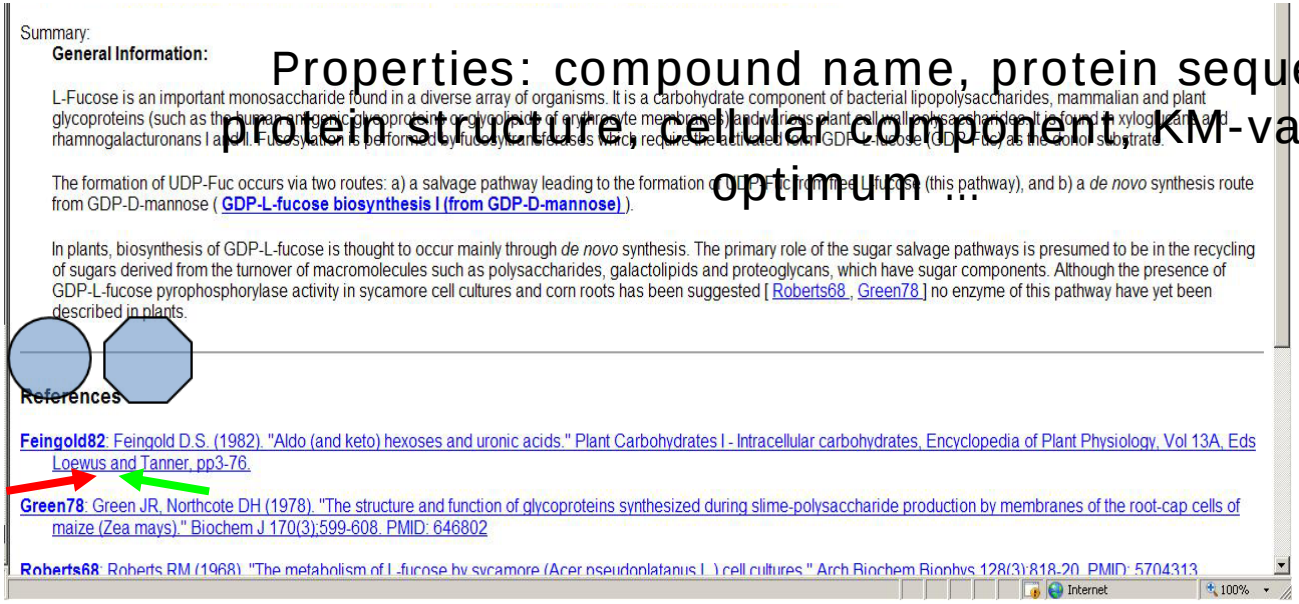
Properties: compound name, protein sequence, protein structure, cellular component, KM-value, PH optimum …

Ontology of Concept Classes, Relation Types and additional Properties

ROTHAMSTED RESEARCH

Concepts and relations (2/2)

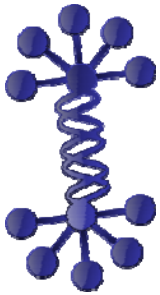




Properties: compound name, protein sequence, protein structure, cellular component, KM-value, PH optimum ...

Concepts:

Relations:

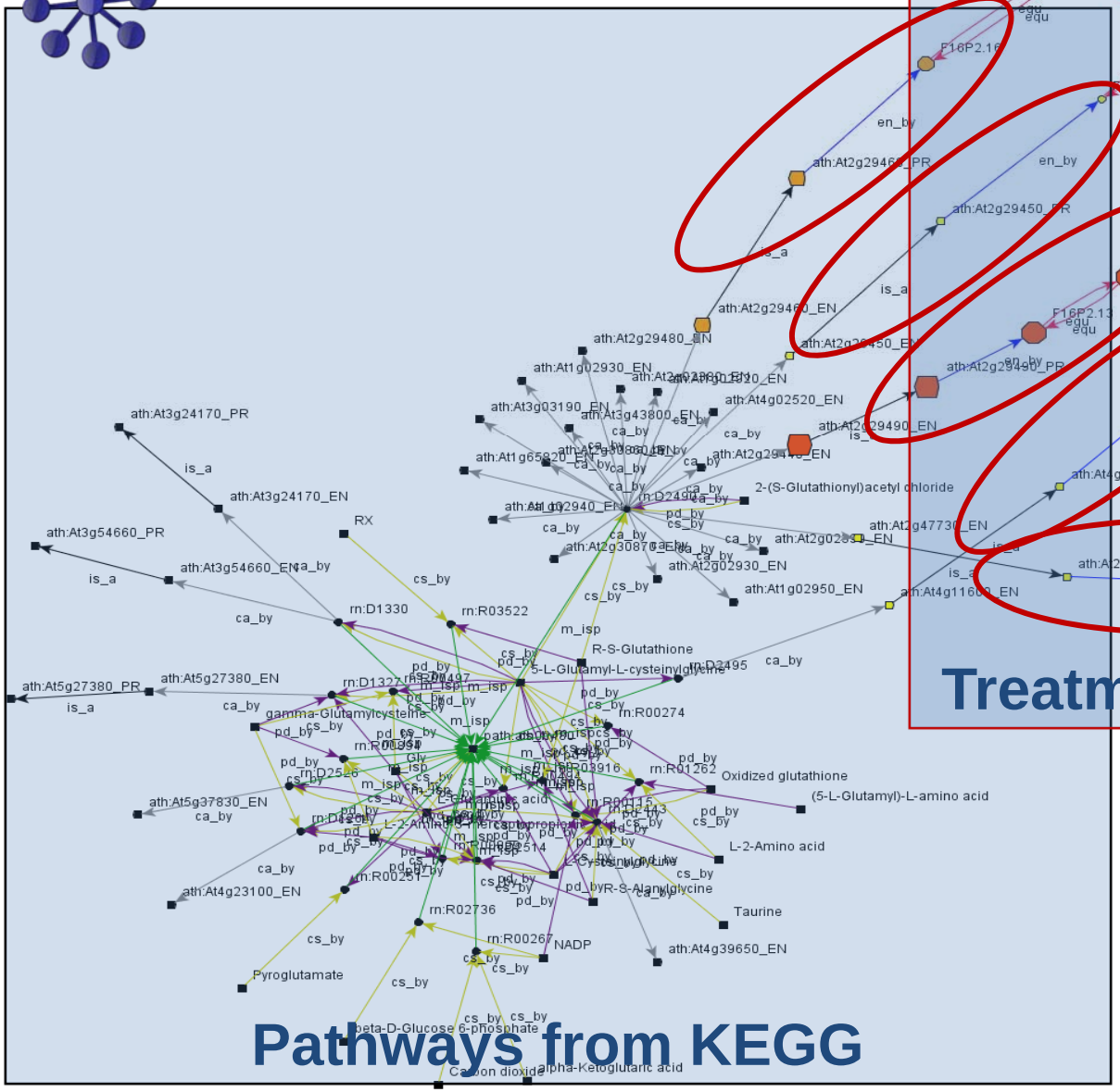


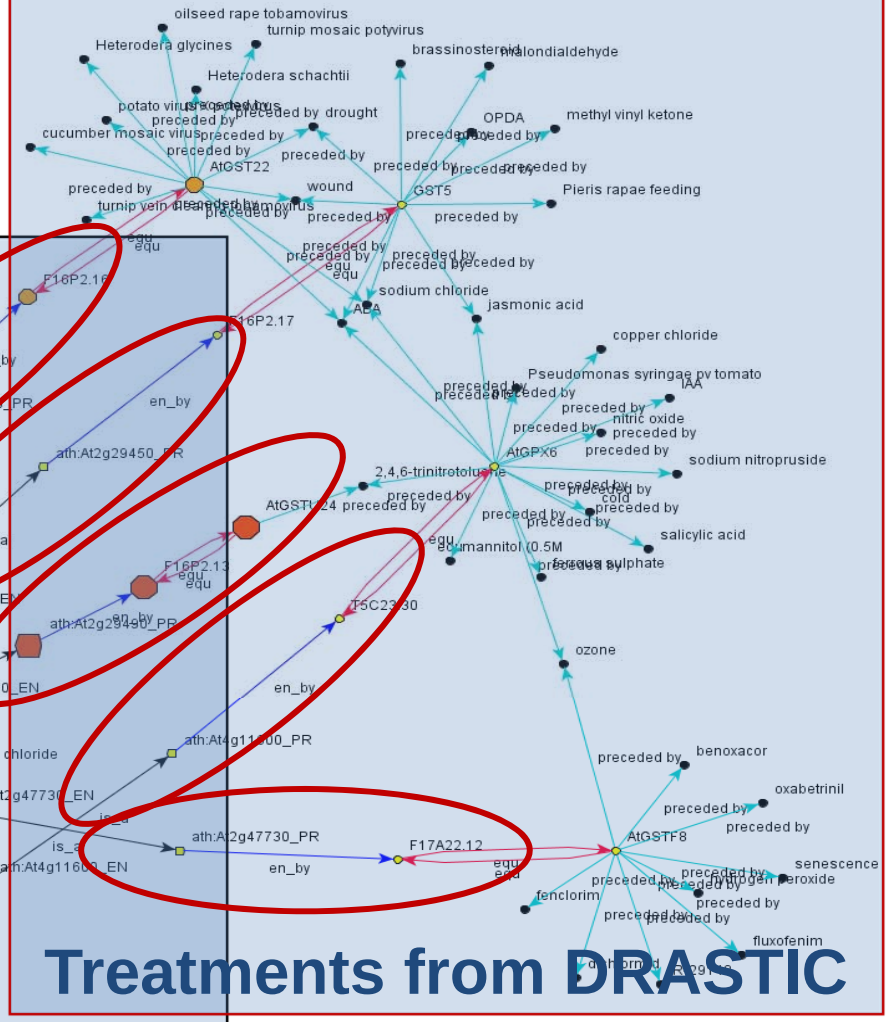
Semantic Integration by Graph Alignment

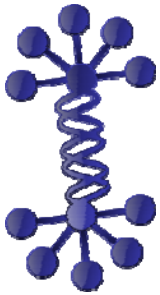
- Going beyond technical data integration
  - Normalising data formats
- Create relations between equivalent entries from different data sources
- Identified by *Mapping methods*
  - Concept name (gene name), synonyms
  - Concept accessions (UniProt ID)
  - Graph neighbourhood
  - Sequence methods

ROTHAMSTED RESEARCH

Graph alignment







Data integration – accession matching

- Matching of accession
- Within the context of data source

Concept

ID: AT2G47510
Annotation: catalytic/ fumarate hydratase [EC:4.2.1.2]
ConceptClass: PROTEIN
DataSource: KEGG
ConceptName: FUM1 (preferred), FUMARASE 1
ConceptAccession: AT2G47510 (TIGR), AT2G47510 (TAIR), AT2G47510 (MIPS), 15226618 (NCBI-GI), 819364 (NCBI-GeneID), P93033 (UniProt)
ConceptGDS:
TAXID (String) : 3702
AAseq (String) : MSIYVASRRKSGGTTVTALRY
NAseq (String) : atgtcgatttacgtcgcgtcgcgacggct

Concept

ID: Locus:AT2G47510
Annotation:
ConceptClass: PROTEIN
DataSource: TAIR
ConceptName: FUM1 (preferred), FUMARASE 1, T30B22.19
ConceptAccession: AT2G47510 (TAIR)
ConceptGDS:
TAXID (String) : 3702

ROTHAMSTED RESEARCH

Data integration – name matching

- At least two concept names have to match

Concept

ID: AT2G47510
Annotation: catalytic/ fumarate hydratase [EC:4.2.1.2]
ConceptClass: PROTEIN
DataSource: KEGG
ConceptName: FUM1 (preferred), FUMARASE 1
ConceptAccession: AT2G47510 (TIGR), AT2G47510 (TAIR), AT2G47510 (MIPS), 15226618 (NCBI-GI), 819364 (NCBI-GeneID), P93033 (UniProt)
ConceptGDS:
TAXID (String) : 3702
AAseq (String) : MSIYVASRRKSGGTTVTALRY
NAseq (String) : atgtcgatttacgtcgcgtcgcgacggct

Concept

ID: Locus:AT2G47510
Annotation:
ConceptClass: PROTEIN
DataSource: TAIR
ConceptName: FUM1 (preferred), FUMARASE 1, T30B22.19
ConceptAccession: AT2G47510 (TAIR)
ConceptGDS:
TAXID (String) : 3702

ROTHAMSTED RESEARCH

Principle of alignment of concepts as nodes in a data graph

- How can we map 2 concepts (proteins)
  - from 2 different data sources (KEGG & TAIR)
- Examine attributes (name, accession, graph structure, sequence)

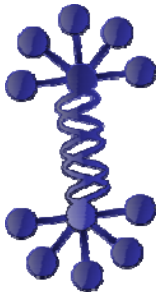
Concept

ID: AT2G47510
Annotation: catalytic/ fumarate hydratase [EC:4.2.1.2]
ConceptClass: PROTEIN
DataSource: KEGG
ConceptName: FUM1 (preferred), FUMARASE 1
ConceptAccession: AT2G47510 (TIGR), AT2G47510 (TAIR), AT2G47510 (MIPS), 15226618 (NCBI-GI), 819364 (NCBI-GeneID), P93033 (UniProt)
ConceptGDS:
TAXID (String) : 3702
AAseq (String) : MSIYVASRRKSGGTTVTALRY
NAseq (String) : atgtcgatttacgtcgcgtcgcgacggct

Concept

ID: Locus:AT2G47510
Annotation:
ConceptClass: PROTEIN
DataSource: TAIR
ConceptName: FUM1 (preferred), FUMARASE 1, T30B22.19
ConceptAccession: AT2G47510 (TAIR)
ConceptGDS:
TAXID (String) : 3702

ROTHAMSTED RESEARCH

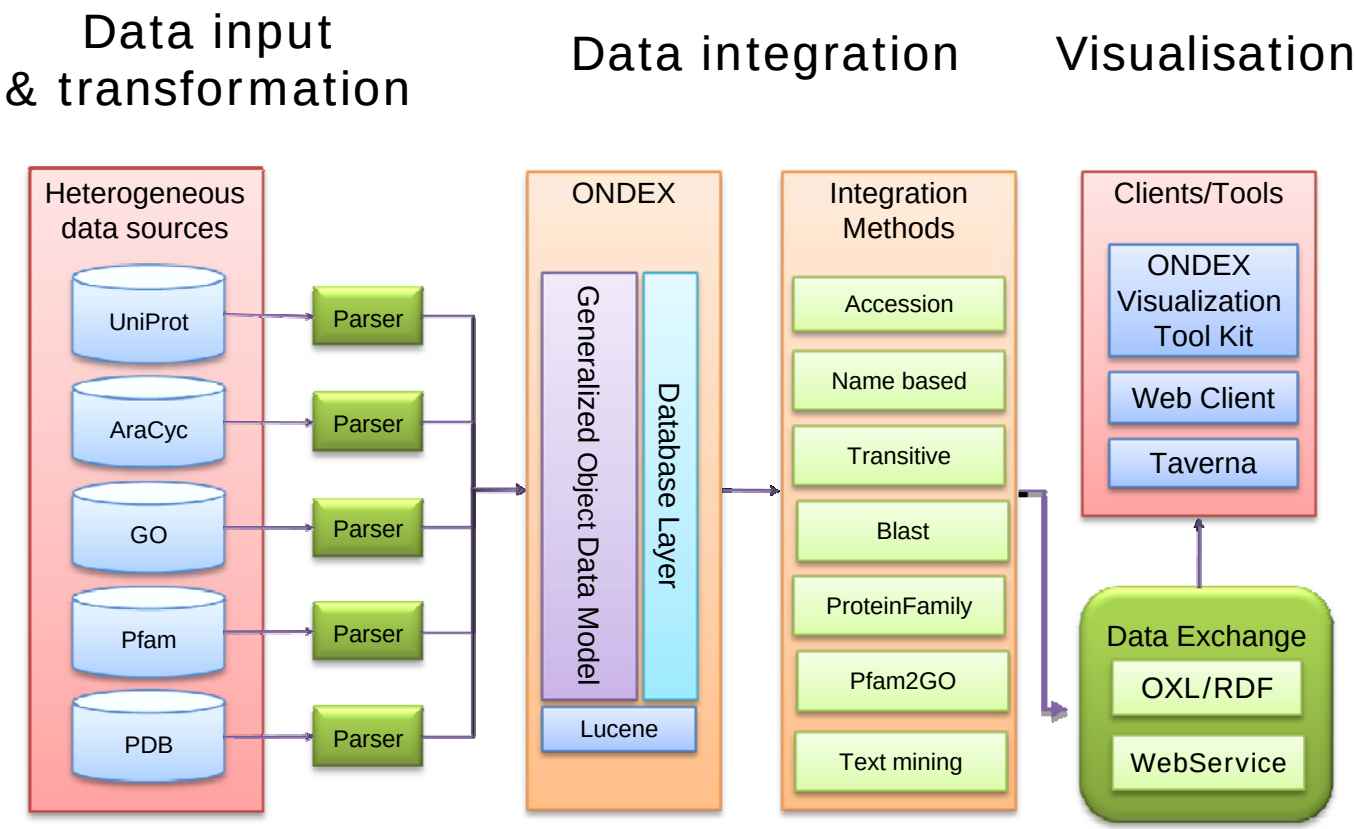


Data integration – sequence matching

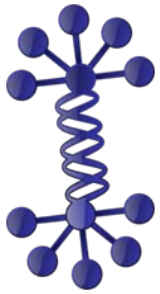
- Sequence similarity methods
  - e.g. BLAST
- Link to a new resource
  - Without any concept attributes in common
- Other similarity matches include
  - EC2GO
  - Pfam2GO

ROTHAMSTED RESEARCH

Ondex data integration scheme



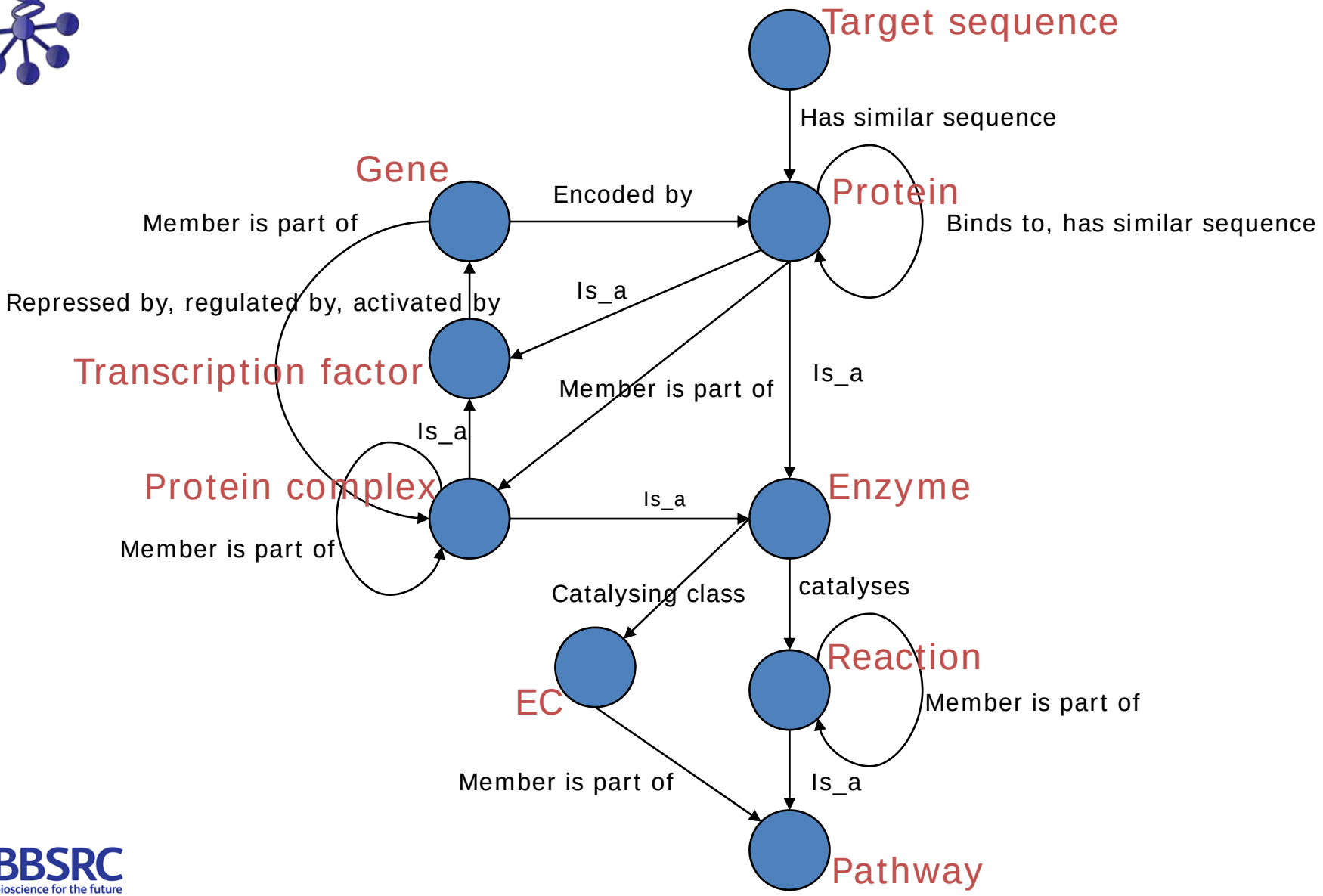
ROTHAMSTED
RESEARCH

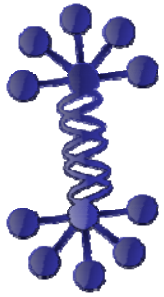


Importing data into Ondex

- What databases to import
- What format these are in
- Ondex parsers already written
  - Generic
    - OBO, PSI-MI, SBML, Tab-delimited, Fasta
  - Database-specific
    - Aracyc, AtRegNet, BioCyc, BioGRID, Brenda, Drastic, EcoCyc, GO, GOA, Gramene, Grassius, KEGG, Medline, MetaCyc, Oglycbase, OMIM, PDB, Pfam, SGD, TAIR, TIGR, Transfac, Transpath, UniProt, WGS, WordNet

Example of resulting graph





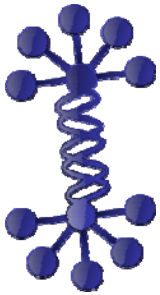
Outline

- Data integration
  - Needs for data integration
  - Benefits of data integration
  - Challenges
  - Solutions
  - Case study

Tutorial

- Network analysis
  - Definition
  - Visualising biological networks
  - Annotation

ROTHAMSTED RESEARCH

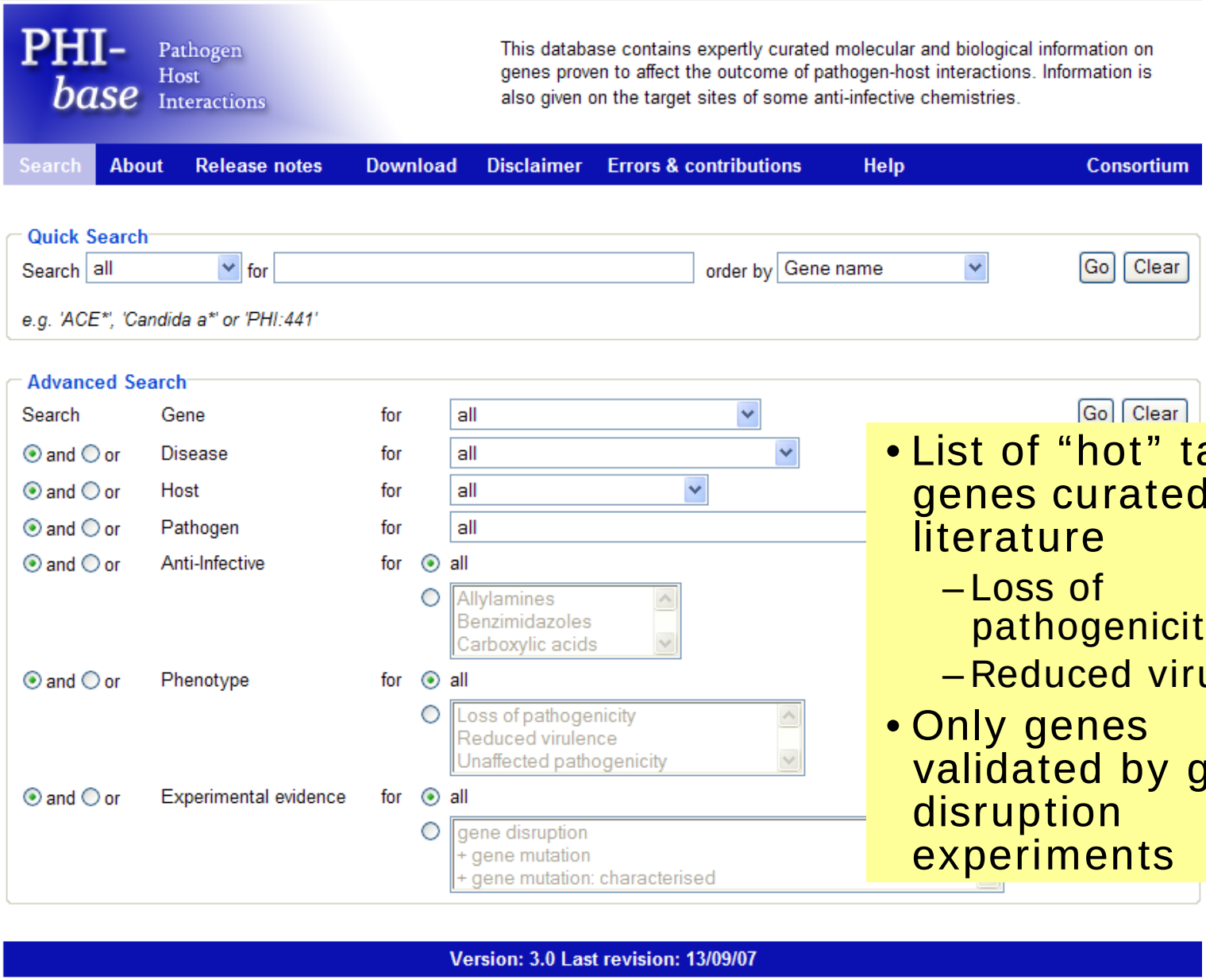


Predicting fungal pathogenicity genes

- Reference database of virulence and pathogenicity genes validated by gene disruption experiments
  - Literature mining
  - http://www.phi-base.org/
- Sequence comparison – orthology and gene cluster analysis

ROTHAMSTED RESEARCH

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

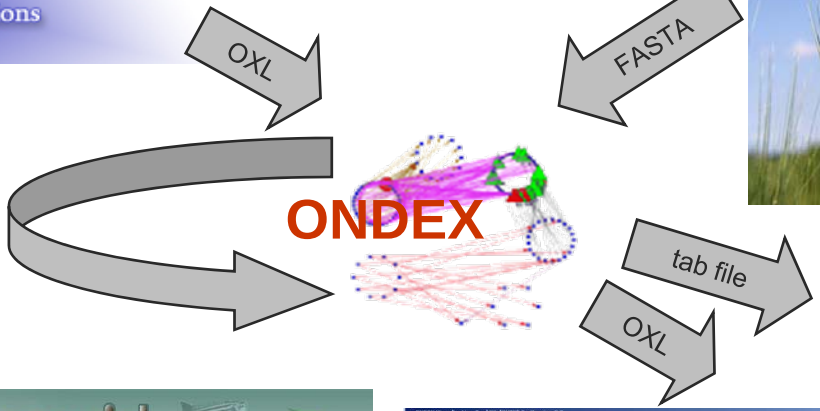


- List of "hot" target genes curated from literature
  - Loss of pathogenicity
  - Reduced virulence
- Only genes validated by gene disruption experiments

Integrated phenotype and comparative genome information

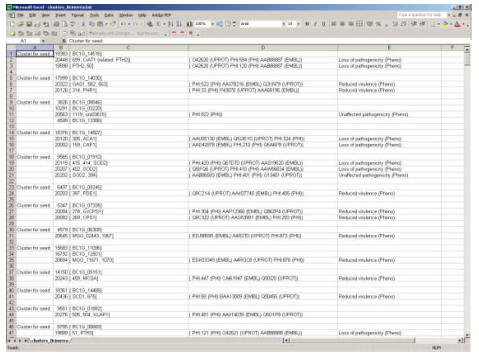
PHI-base Pathogen Host Interactions

Fusarium graminearum (*microscopic fungus*) genome

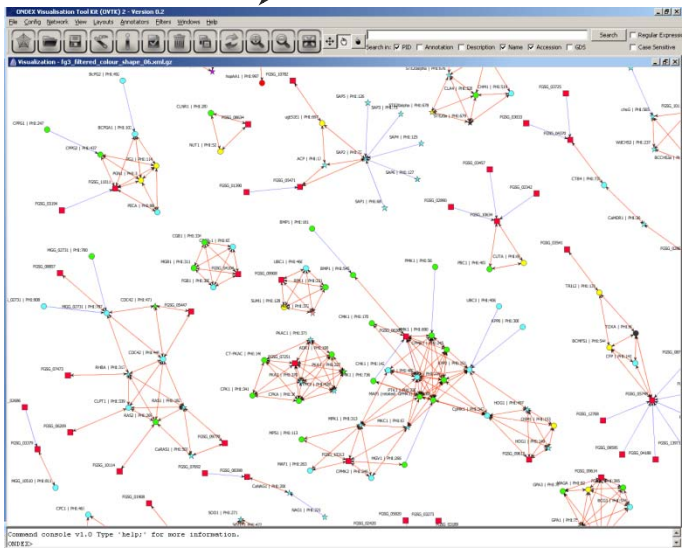


Inparanoid

clusters of orthologs and paralogs between entries of PHI-base and *Fusarium graminearum*

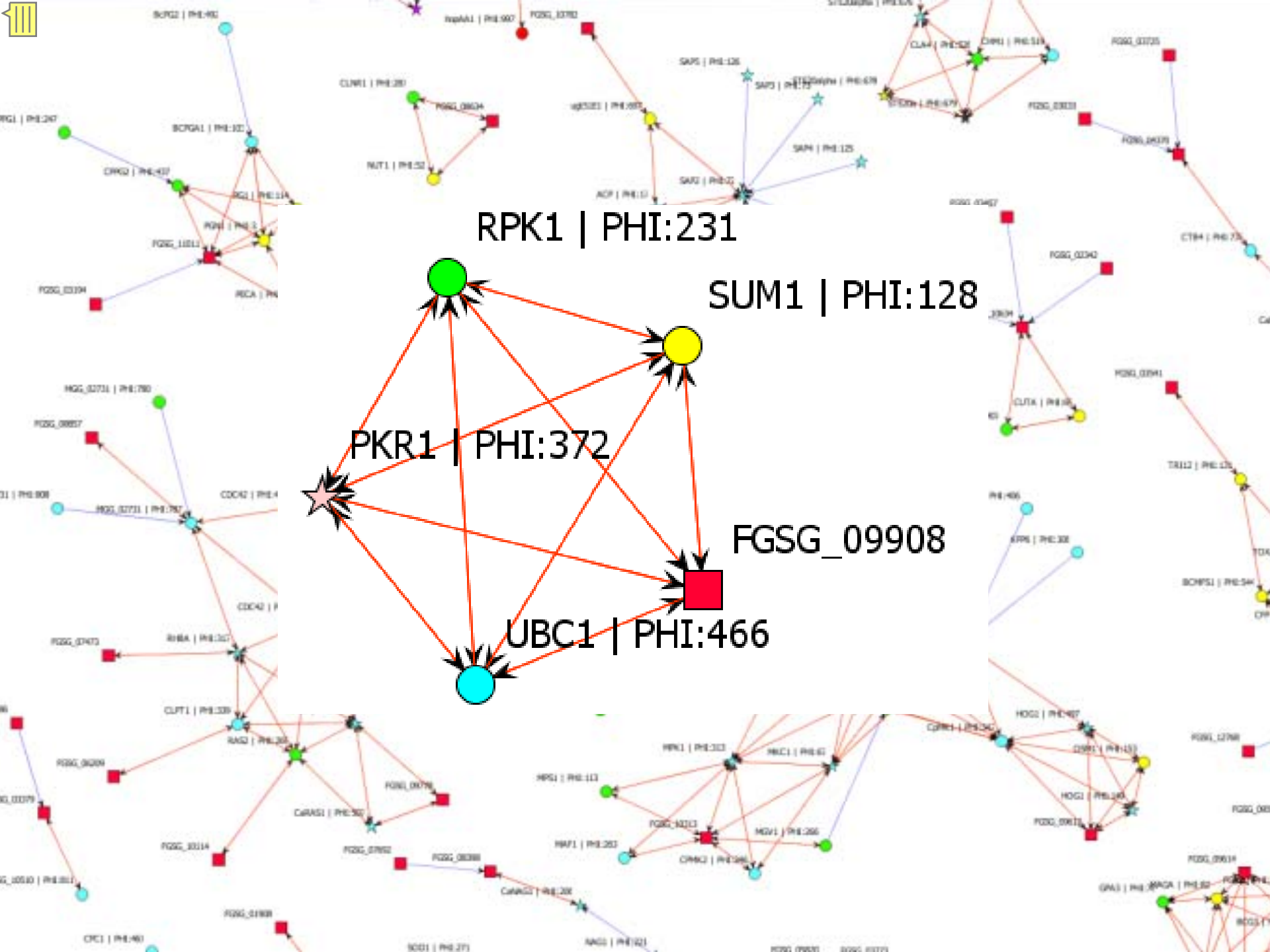


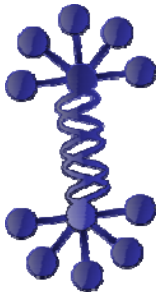
tab separated text file of clusters loaded in Excel



Ondex front-end

ROTHAMSTED RESEARCH

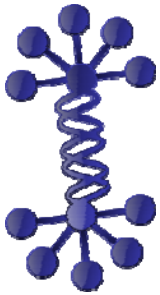

RPK1 | PHI:231
SUM1 | PHI:128
PKR1 | PHI:372
FGSG_09908
UBC1 | PHI:466



Data integration - Summary

- Choose and import data of interest
- Address semantic and syntactic challenges
- Integration methods generate new merged data
- Need to analyse the results of integration
- → Network Analysis

ROTHAMSTED RESEARCH



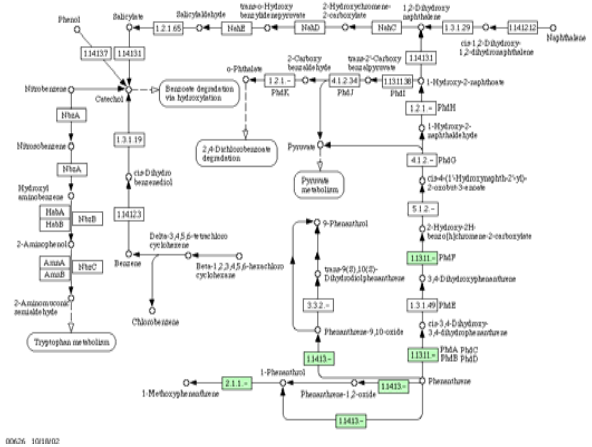
Outline

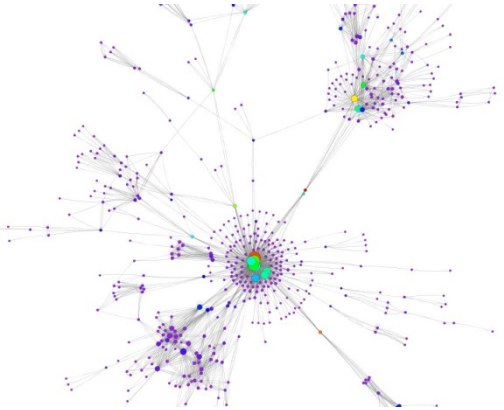
- Data integration
  - Needs for data integration
  - Benefits of data integration
  - Challenges
  - Solutions
  - Case study

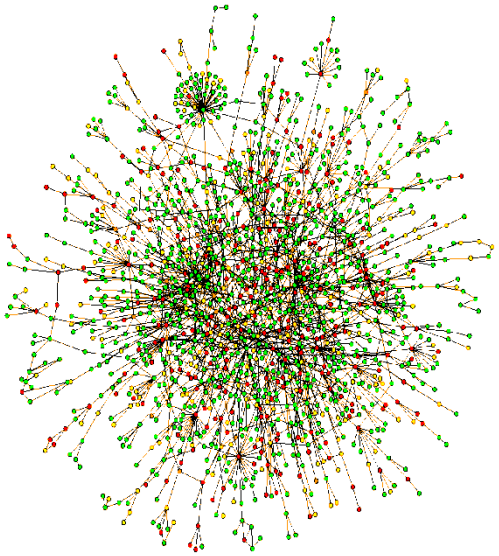
- Network analysis
  - Definition
  - Visualising biological networks
  - Annotation

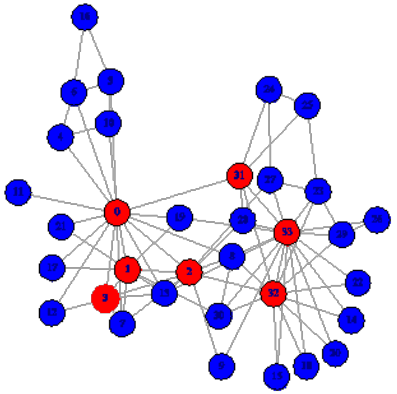
ROTHAMSTED RESEARCH

Everything is a network

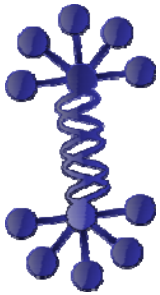








ROTHAMSTED RESEARCH

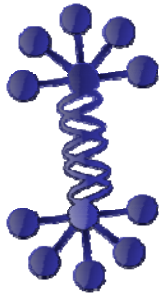


Outline

- Data integration
  - Data types and sources
  - Benefits of data integration
  - Challenges
  - Solutions
  - Case study

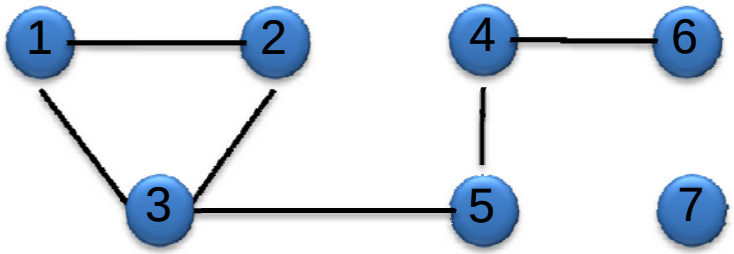
- Network analysis
  - Definition
  - Visualising biological networks
  - Annotation

ROTHAMSTED RESEARCH



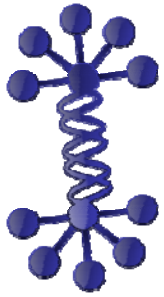
Biological networks can be described using Graph Theory

- A graph $G = (V, E)$ is
  - a set of vertices V (nodes)
  - a set of edges E (links between the nodes)



- $V = \{1,2,3,4,5,6,7\}$
- $E = \{ (1,2), (1,3), (2,3), (3,5), (5,4), (4,6) \}$

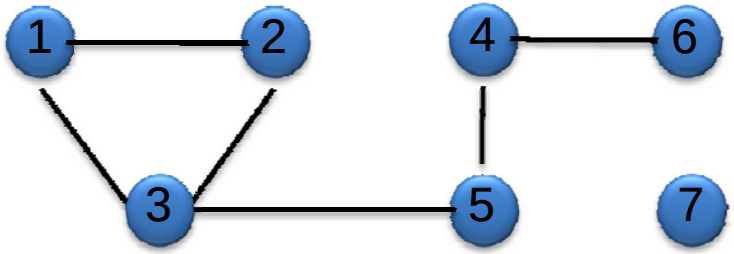
ROTHAMSTED RESEARCH



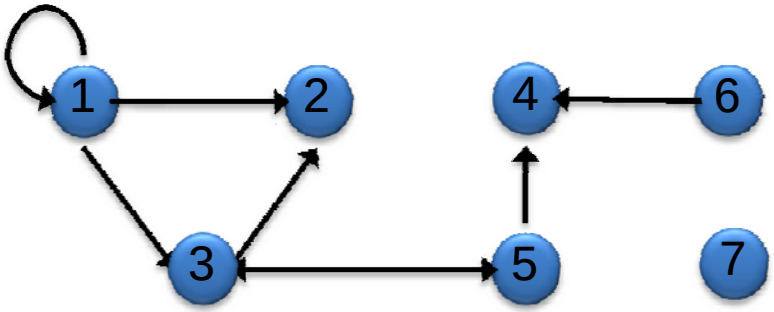
Biological networks can be described using Graph Theory

- Graphs can be :

Undirected



Directed



- Graphs can also be more complex:

multigraphs, bipartite graphs, hypergraphs

ROTHAMSTED RESEARCH

Biological networks can be described using Graph Theory

- Node degree

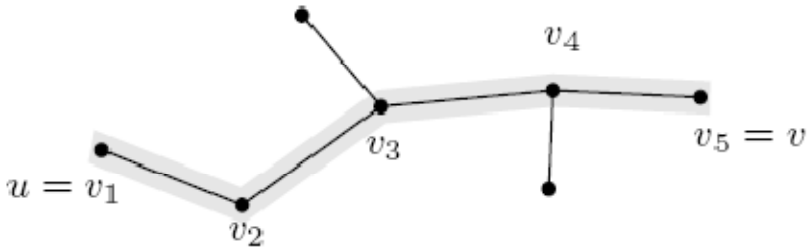
  Is the number of edges of a node. In a directed graph, we can also define the in-degree and out-degree.

- Adjacent nodes

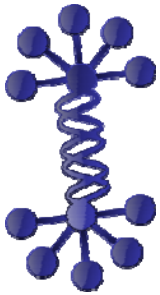
  Are nodes joined by an edge

- Path

  Is a sequence of adjacent nodes between two nodes



ROTHAMSTED RESEARCH



Biological networks can be described using Graph Theory

- Connected graph

  If there is a path between all nodes in a graph

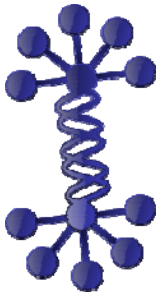
- Distance

  Between two nodes *u* and *v* is the length of the shortest path between *u* and *v* (can be infinite)

- Graph diameter

  is the maximum distance in a graph

ROTHAMSTED RESEARCH



Biological networks can be described using Graph Theory

- Clustering coefficient

  characterizes the overall tendency of nodes to form clusters or groups

- Centrality measures

  Identify “most important” nodes in a graph (hubs). The importance can be assessed for example using :

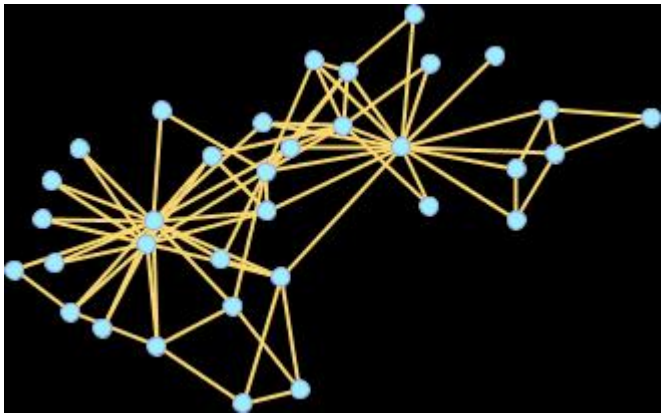
  - the number of connections (degree centrality)
  - the proximity (closeness centrality)
  - the number of paths (betweenness centrality)

ROTHAMSTED RESEARCH

Social Networks

Zachary's karate club; J. Anthropological Res. 33, 452-473 (1977)

- 34 members of a karate club
- Visual inspection shows some substructure
- Some very 'popular' individuals (hubs)

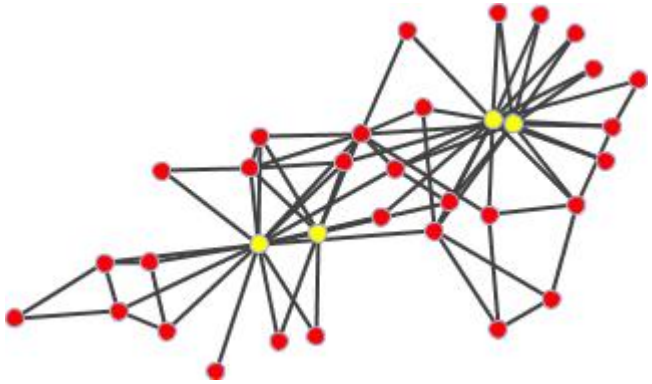


ROTHAMSTED RESEARCH

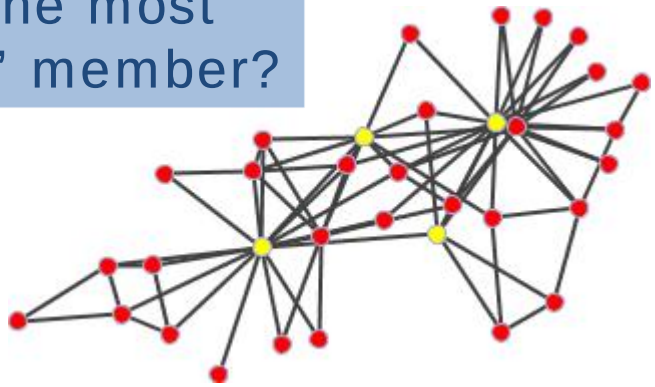
Social Networks

Zachary's karate club; J. Anthropological Res. 33, 452-473 (1977)

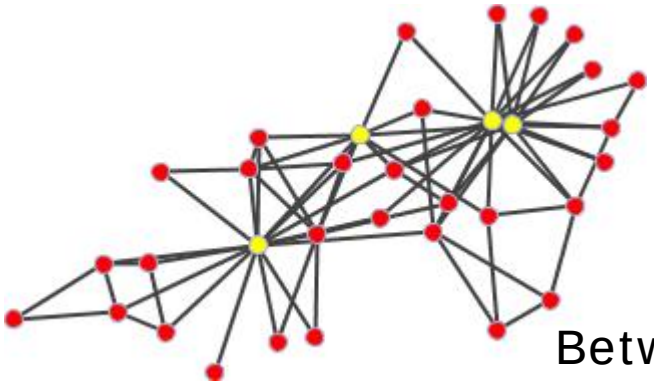
Who is the most 'influential' member?



Degree: 34, 1, 33, 2



Closeness: 1, 3, 34, 32



Betweenness: 1, 34, 33, 3

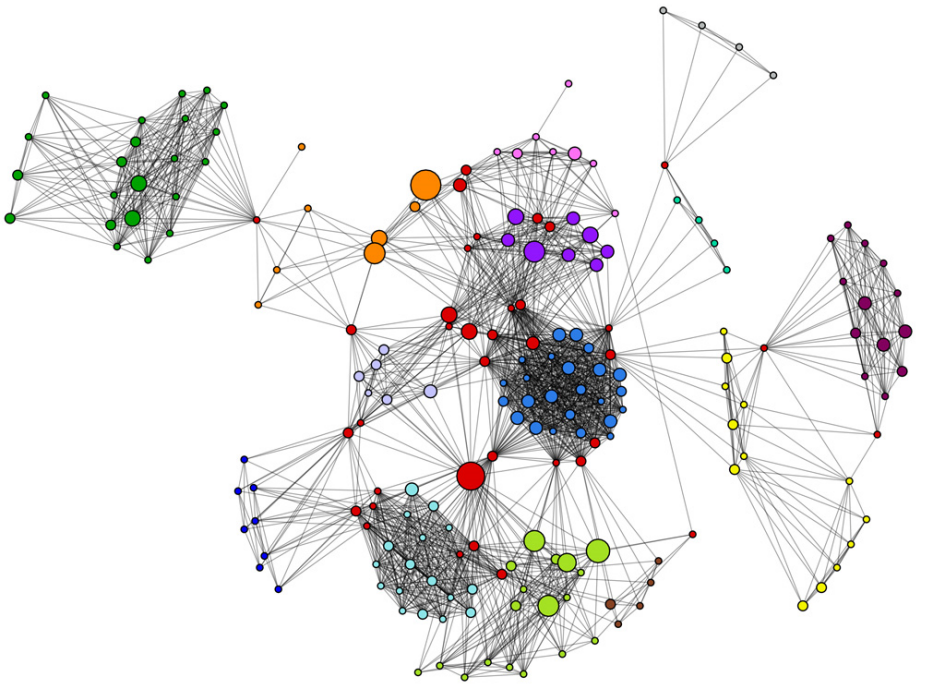
Data from http://www-personal.umich.edu/~mejn/netdata/

ROTHAMSTED RESEARCH

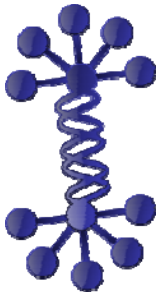
Biological networks can be described using Graph Theory

- Clique

  is a subset of nodes such that every pair of nodes in the subset is joined by an edge



ROTHAMSTED RESEARCH



Outline

- Data integration
  - Needs for data integration
  - Benefits of data integration
  - Challenges
  - Solutions
  - Case study

- Network analysis
  - Definition
  - Visualising biological networks
  - Annotation

ROTHAMSTED RESEARCH

Existing data visualisation tools

Review

Open Access

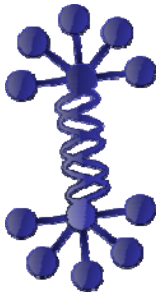
A survey of visualization tools for biological network analysis

Georgios A Pavlopoulos*, Anna-Lynn Wegener and Reinhard Schneider

- http://www.biodatamining.org/content/1/1/12

- Medusa, Cytoscape, BioLayout Express 3D, Osprey, Proviz, Ondex, PATIKA, PIVOT, Pajek

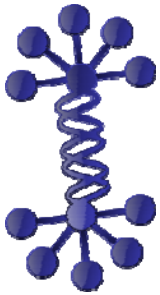
ROTHAMSTED RESEARCH



Outline

- Data integration
- Network analysis
  - Definition
  - Visualising biological networks
    - Layouts
    - Case study
    - Filters
  - Annotation

ROTHAMSTED RESEARCH

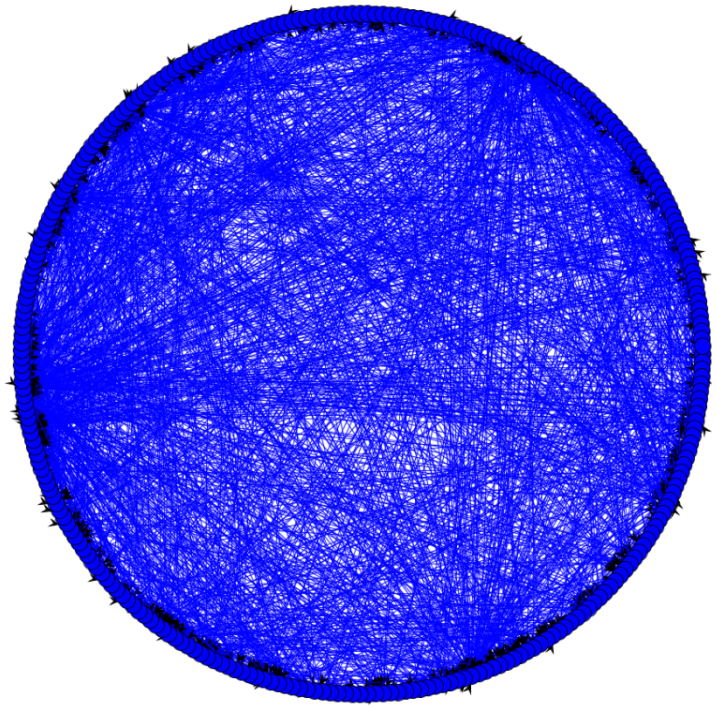


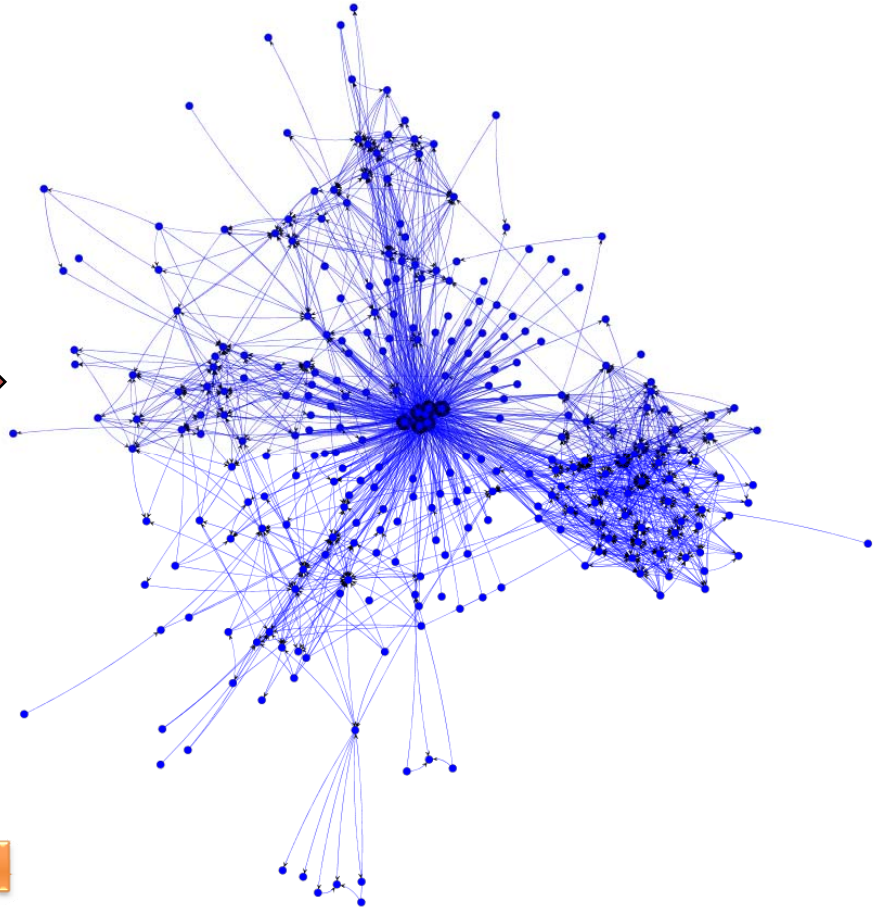
Outline

- Data integration
- Network analysis
  - Definition
  - Visualising biological networks
    - Layouts
    - Case study
    - Filters
  - Annotation

ROTHAMSTED RESEARCH

Visualising biological networks e.g. PPI network



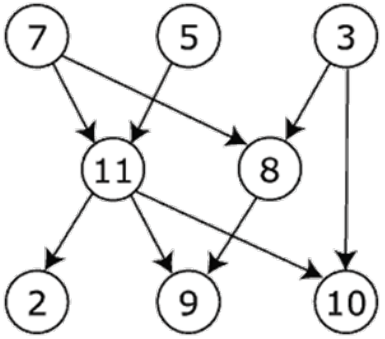


Tutorial

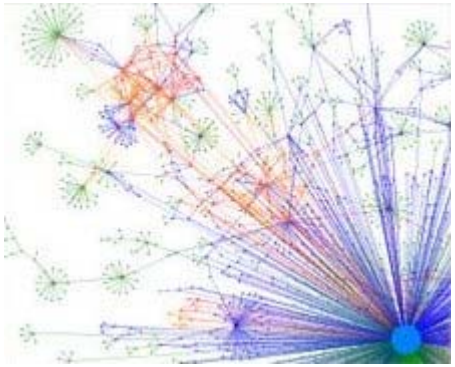
ROTHAMSTED RESEARCH

Layouts

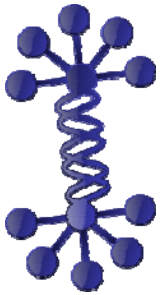
- topology → shape → geometry
- Draw 2-dimensional representations
  - Trees
  - Directed acyclic graph (DAG)
  - General graph (*e.g.* force-directed)



- Case study in Ondex
  - bioenergy crop improvement



ROTHAMSTED RESEARCH

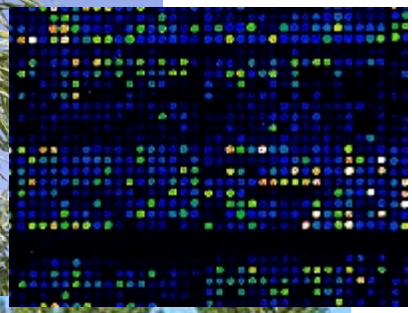


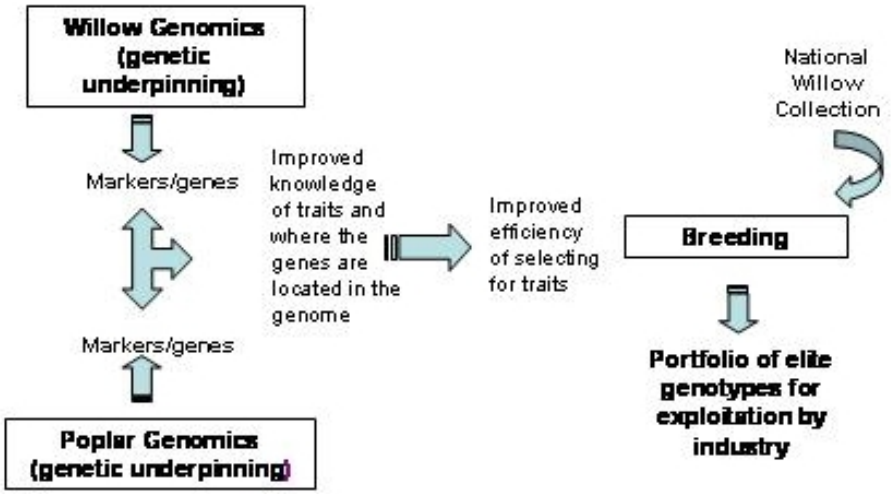
Outline

- Data integration
- Network analysis
  - Definition
  - Visualising biological networks
    - Layouts
    - Case study
    - Filters
  - Annotation

ROTHAMSTED RESEARCH

Bioenergy Crop Improvement
Candidate Gene Identification





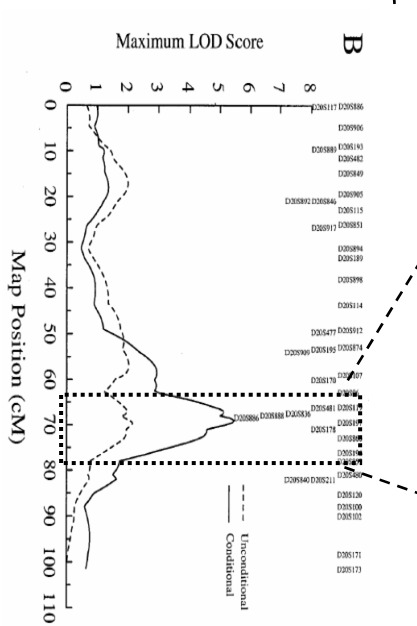
ROTHAMSTED RESEARCH

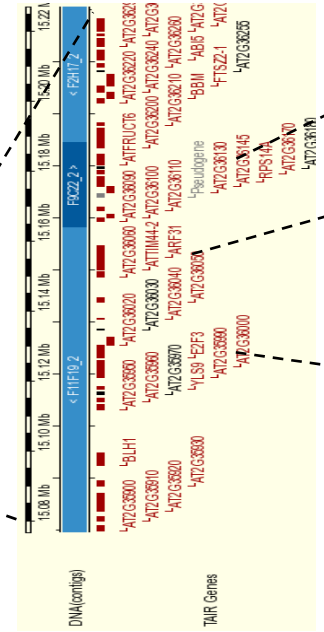
Identifying Candidate Genes

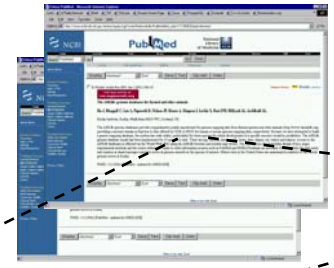
Willow

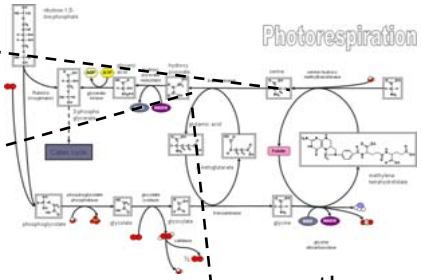
model
e.g. Poplar,
Arabidopsis

Linked
References

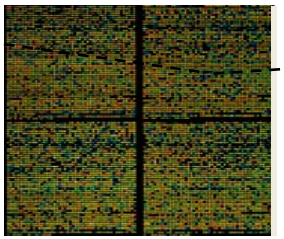




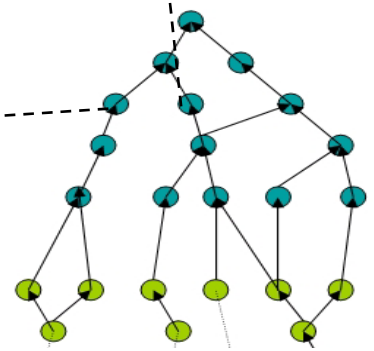




pathways



Expression Patterns



Ontologies

QTL Map

Orthologous
Markers

Physical
map

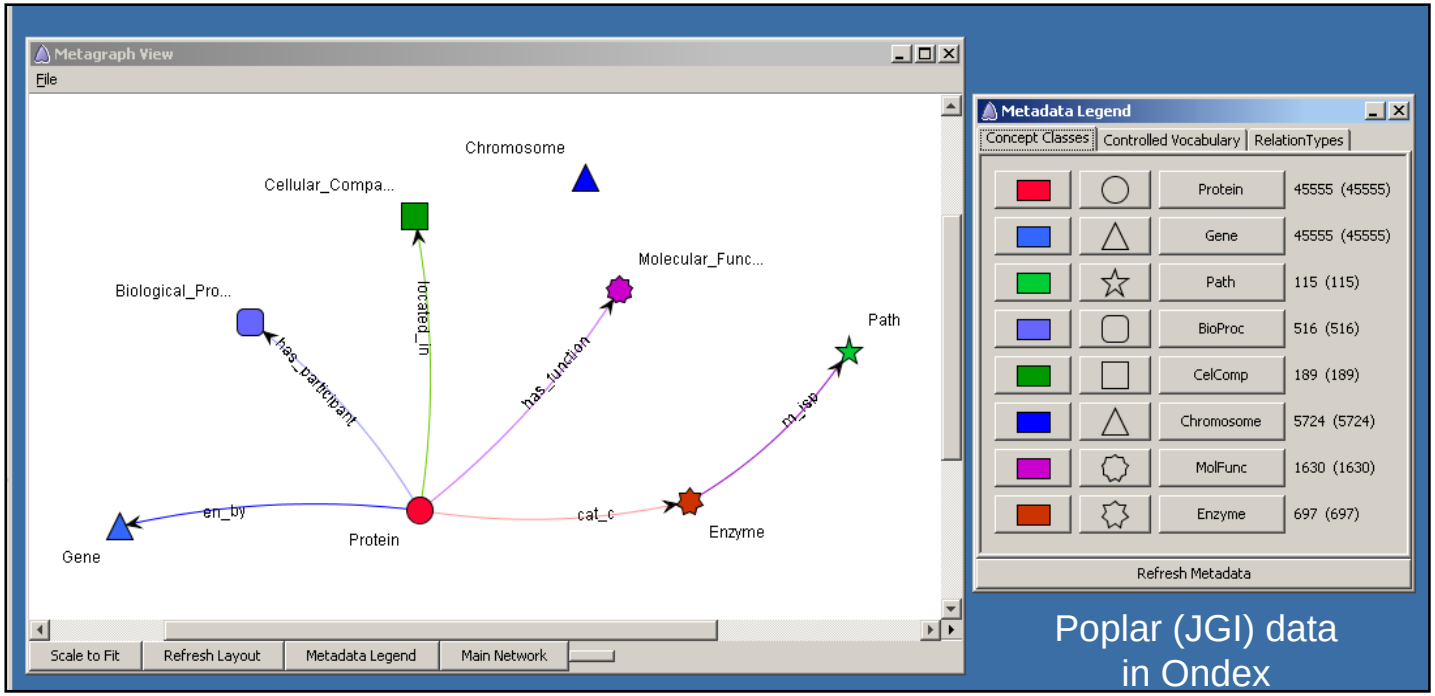
Genes

List of candidate genes linked to biological processes

ROTHAMSTED
RESEARCH

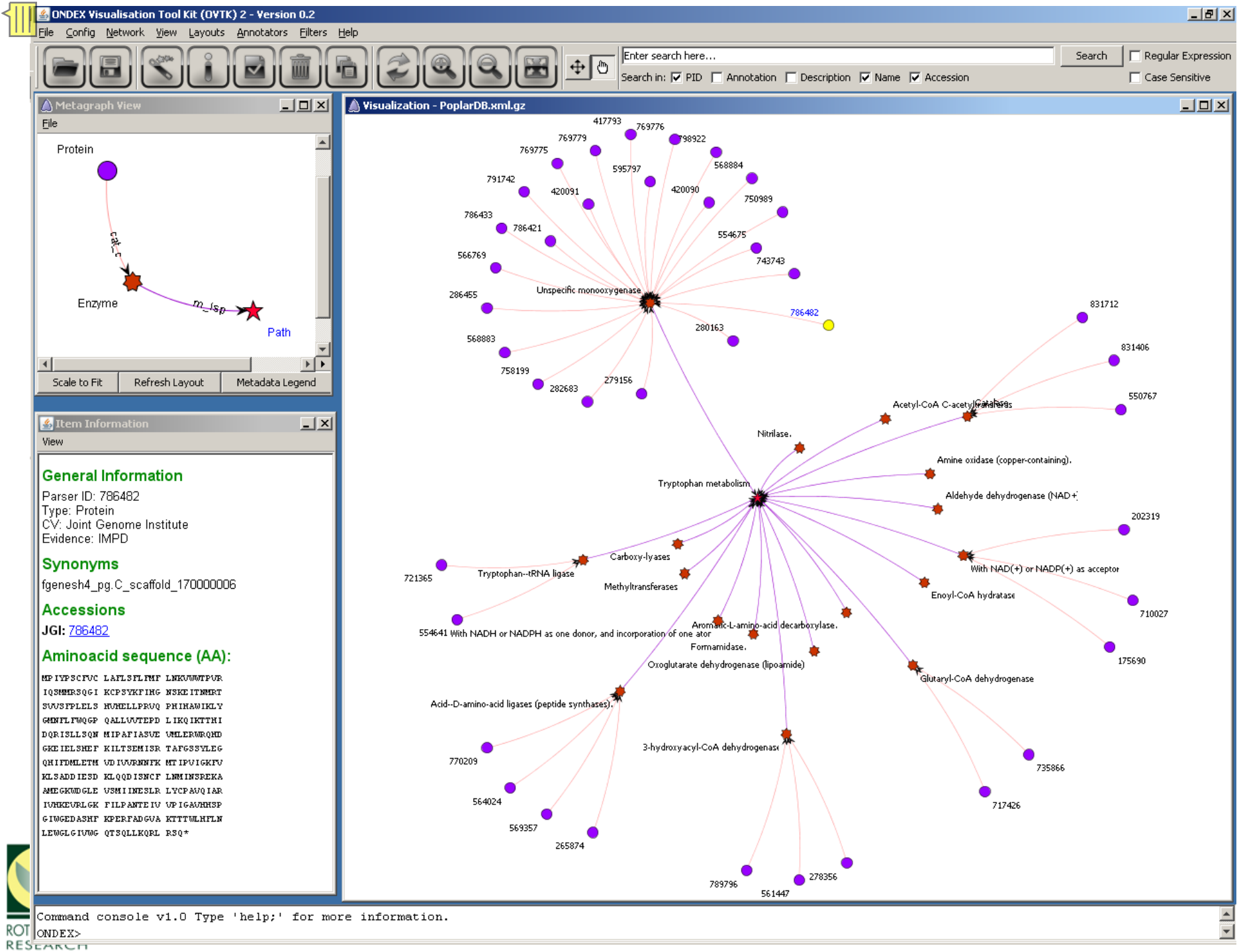
Integrating Relevant Data Sources

- Poplar (JGI) and Arabidopsis (TAIR) genomes
- Linking genes to KEGG, AraCyc, GO, MEDLINE
- Arabidopsis Hormone Database for trait/hormones/genes relations

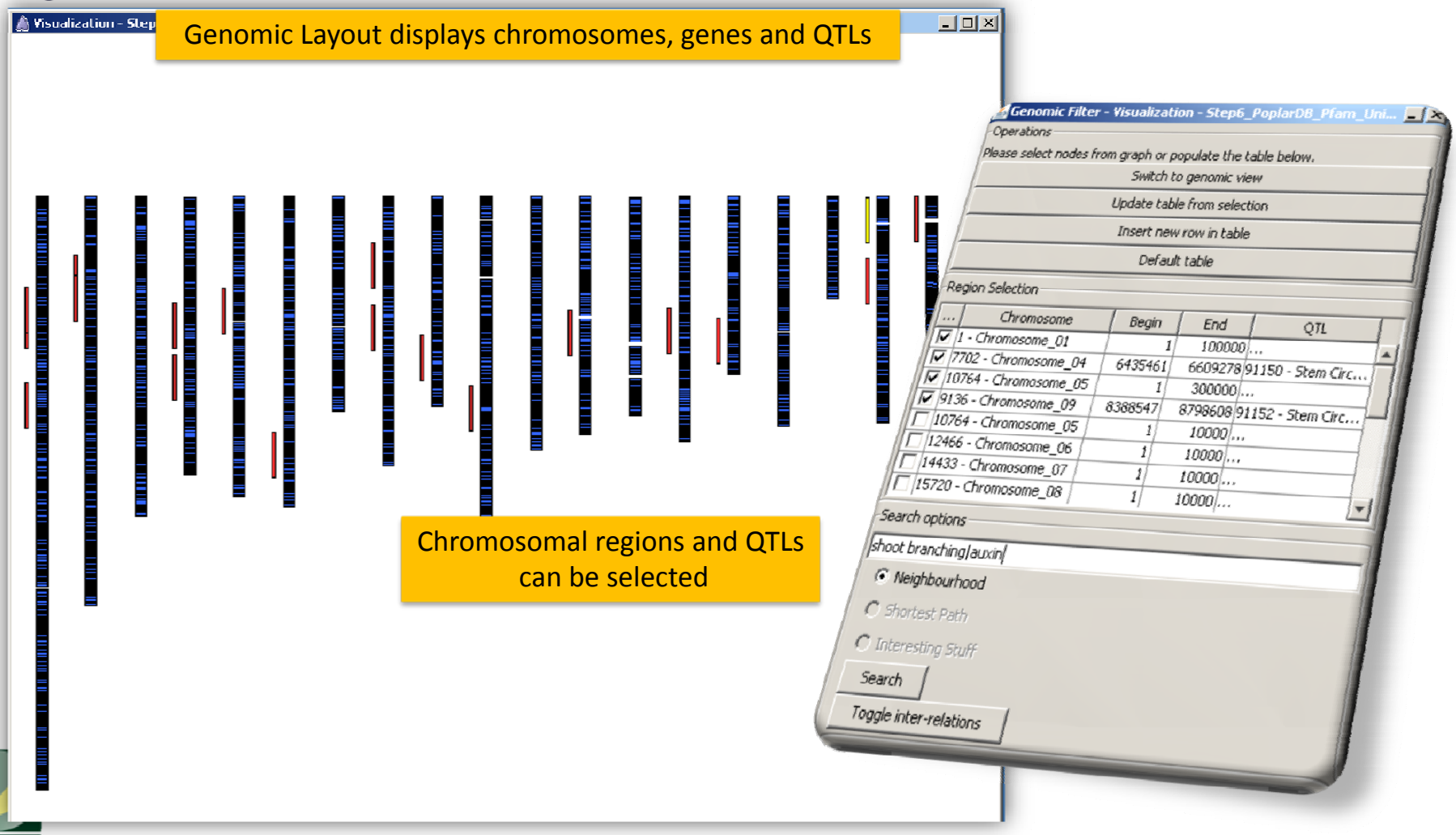


Poplar (JGI) data in Ondex

ROTHAMSTED RESEARCH

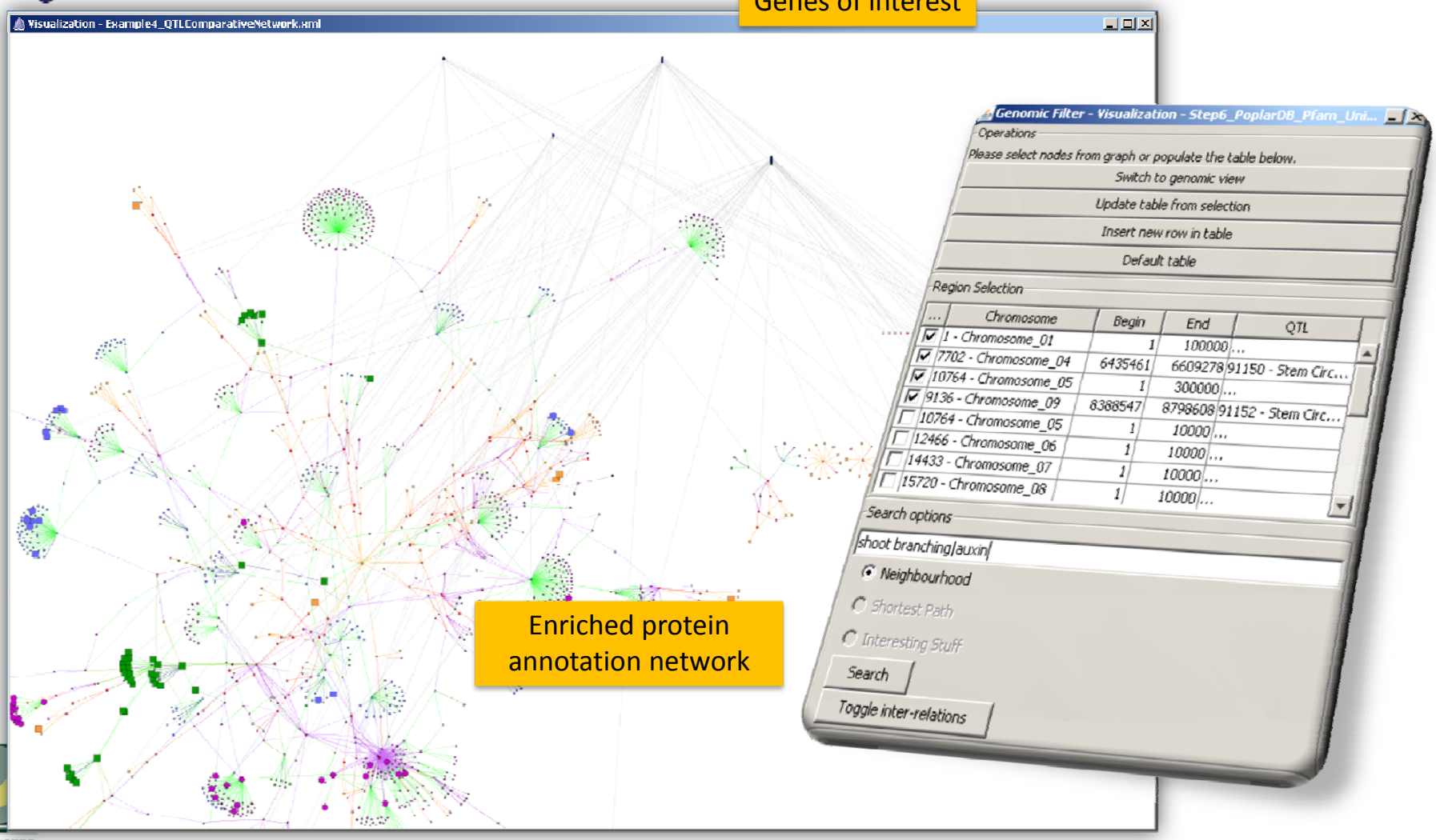

ONDEX Visualisation Tool Kit (OVTK) 2 - Version 0.2
File Config Network View Layouts Annotators Filters Help
Enter search here...
Search
Regular Expression
Search in: PID Annotation Description Name Accession
Case Sensitive
Metagraph View
File
Protein
Enzymе
Path
Scale to Fit
Refresh Layout
Metadata Legend
Item Information
View
General Information
Parser ID: 786482
Type: Protein
CV: Joint Genome Institute
Evidence: IMPD
Synonyms
fgenesh4_pg.C_scaffold_170000006
Accessions
JGI: 786482
Aminoacid sequence (AA):
MPIYPSCFVC LAFLSFLFMF LNKVWWTPVR
IQSMMRSQGI KCPSYKFIHG NSKEITNMRT
SVVSFPLELS HVHELLPRVQ PHIHAWIKLY
GMNFLFWQGP QALVVTEPD LIKQIKTTHI
DQRISLLSQN MIPAFIASVE VMLERWRQHD
GKEIELSHEF KILTSEMISR TAFGSSYLEG
QHIFDMLETM VDIVVRNNFK MTIPVIGKFV
KLSADDIESD KLQQDISNCF LNMINSREKA
AMEGKWDGLE VSMIINESLR LYCPAVQIAR
IVHKEVRLGK FILPANTEIV VPIGAVHHSP
GIWGEDASHF KPERFADGVA KTTTWLHFLN
LEWGLGIVWG QTSQLLKQRL RSQ*
Visualization - PoplarDB.xml.gz
Unspecific monooxygenase
Tryptophan metabolism
Nitrilase.
Acetyl-CoA C-acetyltransferase
Amine oxidase (copper-containing).
Aldehyde dehydrogenase (NAD+)
Carboxy-lyases
Tryptophan--tRNA ligase
Methyltransferases
With NAD(+) or NADP(+) as acceptor
Enoyl-CoA hydratase
Aromatic-L-amino-acid decarboxylase.
Formamidase.
Oxoglutarate dehydrogenase (lipoamide)
Glutaryl-CoA dehydrogenase
Acid--D-amino-acid ligases (peptide synthases).
3-hydroxyacyl-CoA dehydrogenase
554641 With NADH or NADPH as one donor, and incorporation of one ator
721365
770209
564024
569357
265874
789796
561447
278356
735866
717426
175690
710027
202319
550767
831406
831712
786482
280163
743743
554675
750989
568884
798922
595797
420090
420091
791742
769775
769779
769776
417793
786421
786433
566769
286455
568883
758199
282683
279156
Command console v1.0 Type 'help;' for more information.
ONDEX>

Genomic View in Ondex

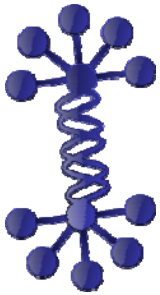


ROTHAMSTED RESEARCH

Network View in Ondex



ROTHAMSTED RESEARCH



Outline

- Data integration
- Network analysis
  - Definition
  - Visualising biological networks
    - Layouts
    - Case study
    - Filters
  - Annotation

ROTHAMSTED RESEARCH

Filters

- Integrating different datasets
  → large resulting graph
- Need to narrow down
- Select meaningful areas of the graph Tutorial

- Example in Ondex
  - protein-protein interaction network

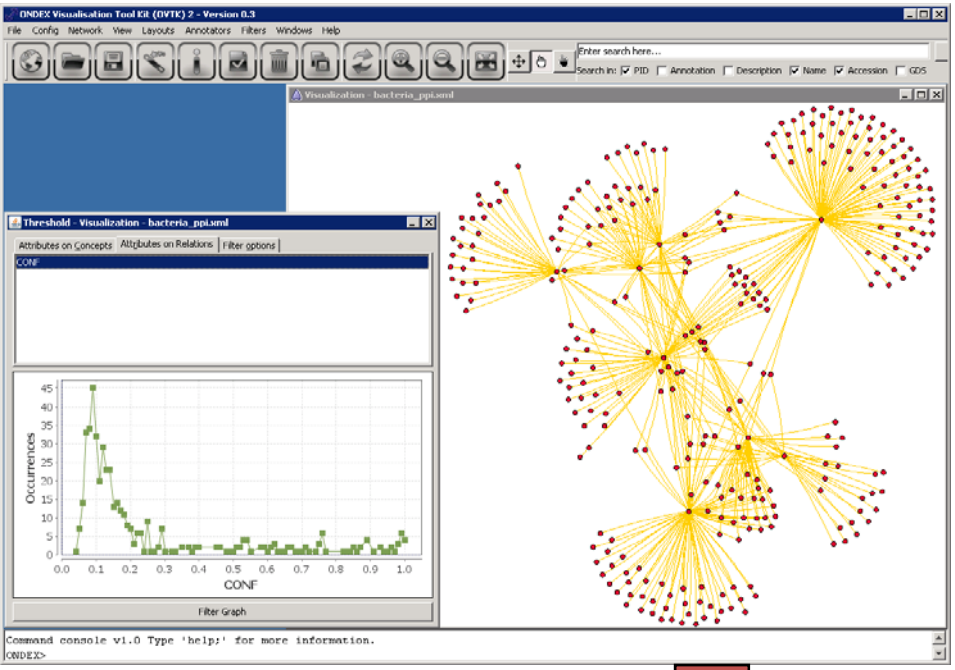
ROTHAMSTED RESEARCH

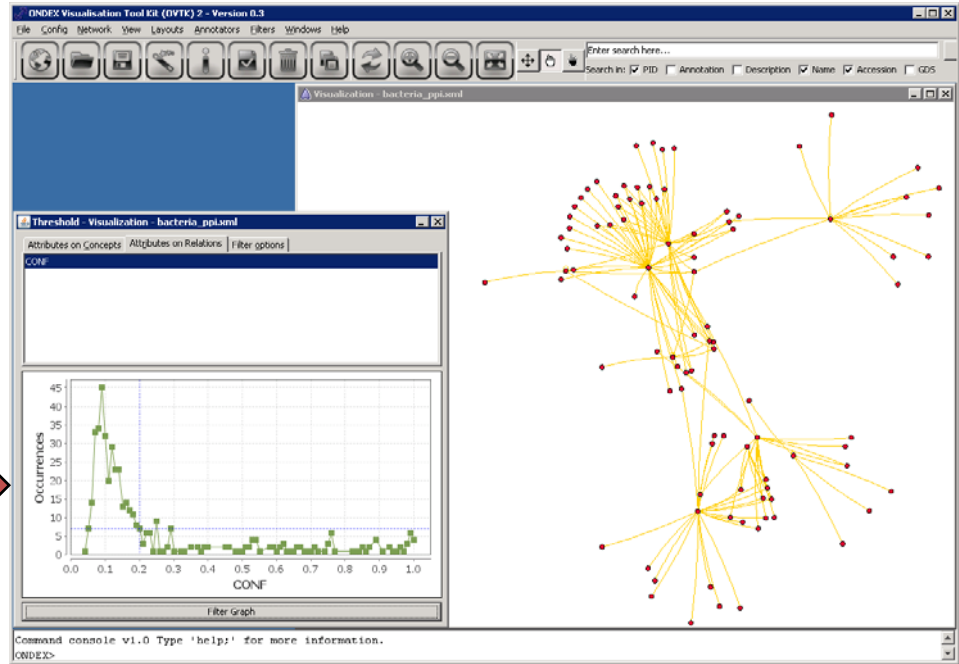
Filters in Ondex

Protein protein interactions measured using quantitative techniques

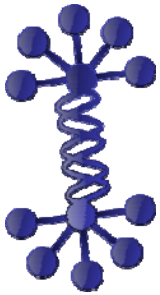
→ Relations on graph have confidence values (confidence)

→ Threshold filter





ROTHAMSTED RESEARCH

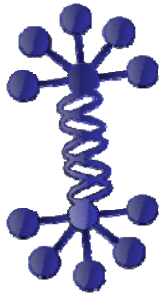


Outline

- Data integration
  - Needs for data integration
  - Benefits of data integration
  - Challenges
  - Solutions
  - Case study

- Network analysis
  - Definition
  - Visualising biological networks
  - Annotation

ROTHAMSTED RESEARCH



Annotators (1/3)

- Visualise concepts and relations using their attributes/properties
  - Colour
  - Shape
  - Size

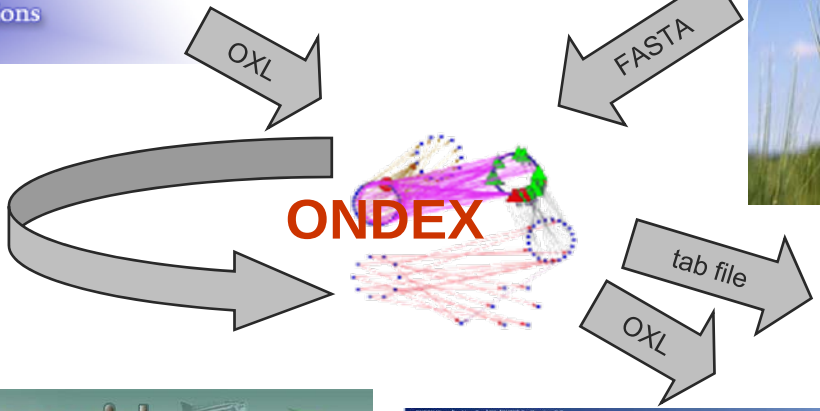
Tutorial

ROTHAMSTED RESEARCH

Integrated phenotype and comparative genome information

PHI-base Pathogen Host Interactions

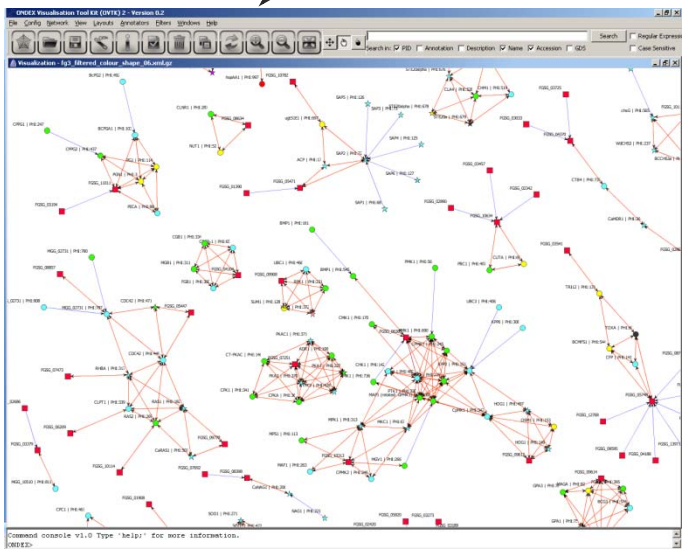
Fusarium graminearum (*microscopic fungus*) genome



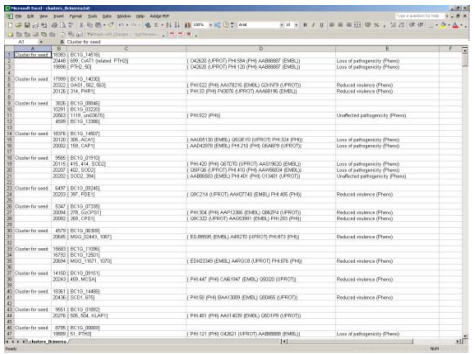
Inparanoid mapping

Inparanoid

clusters of orthologs and paralogs between entries of PHI-base and *Fusarium graminearum*



Ondex front-end



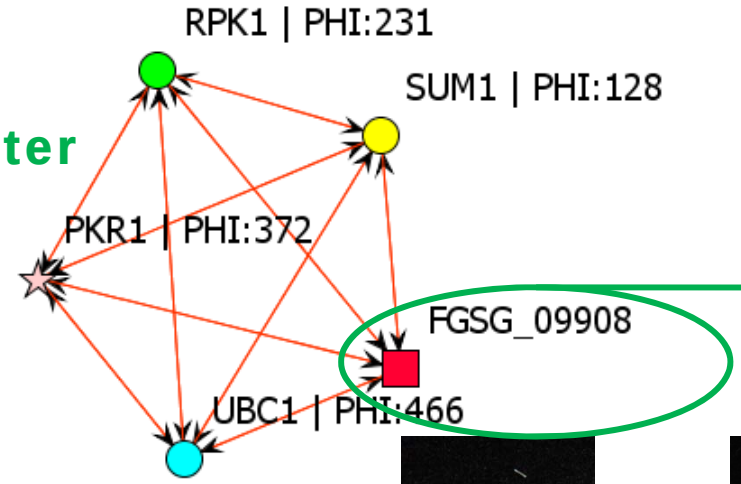
tab separated text file of clusters loaded in Excel

ROTHAMSTED RESEARCH

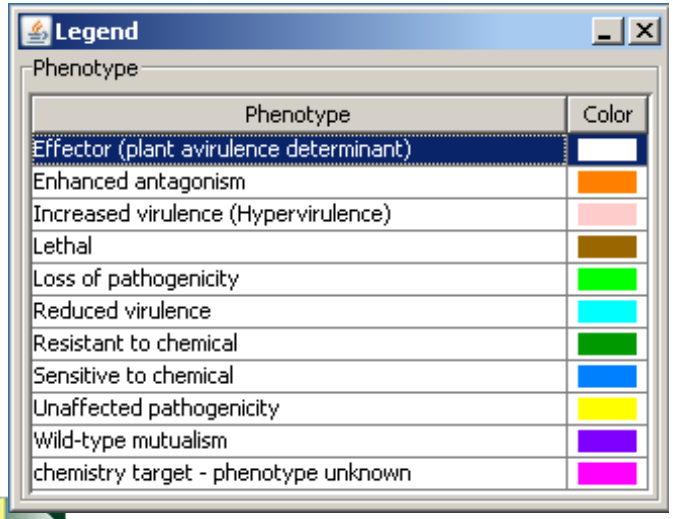
Protein kinase A- regulatory subunit

Mixed phenotype cluster

Shape Legend
Star: animal
Circle: plant
Red square: Fusarium



Reduced virulence



Uninfected

Wild-type PH-1

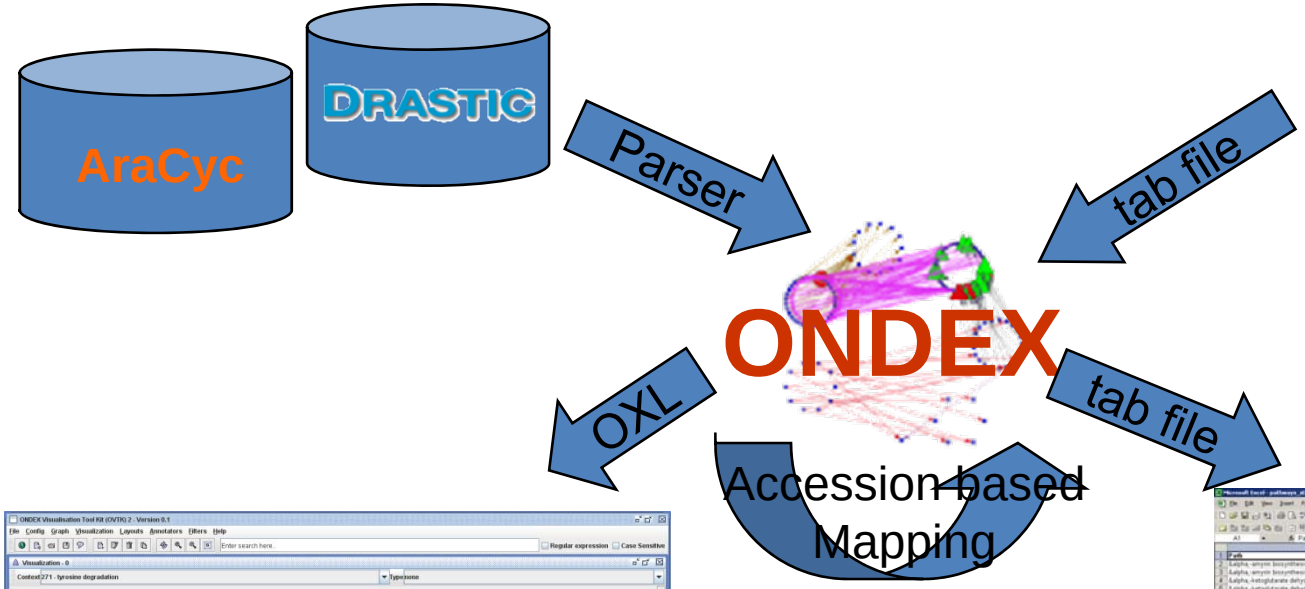
FGSG_9908 pkar PH-1

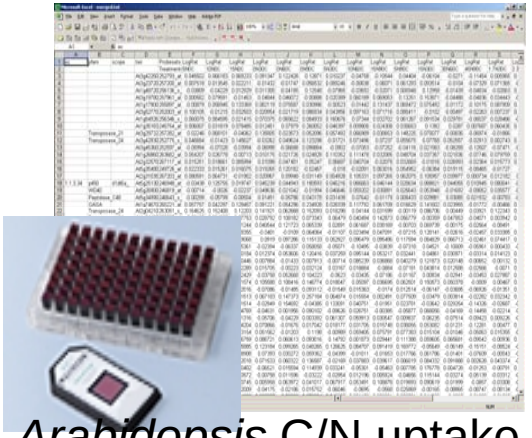
ROTHAMSTED RESEARCH

Annotators (2/3)

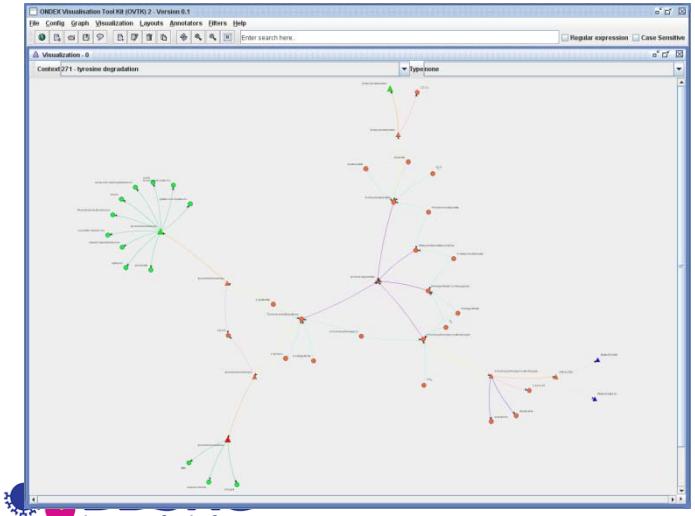
- Virtual Knock-out
  - Annotator to see how important a single concept is to all possible paths contained in a network
  - Ondex resizes the concepts based on this score
- Scale Concept by Value
  - Pie charts
  - Up/down regulation is indicated in red/green

Mapping microarray expression data to integrated pathways

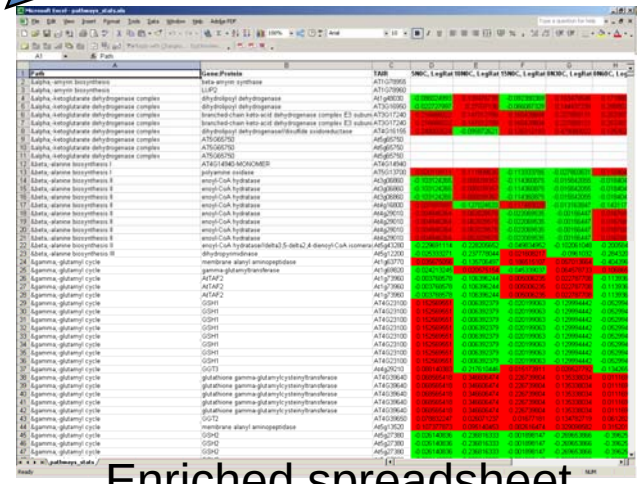




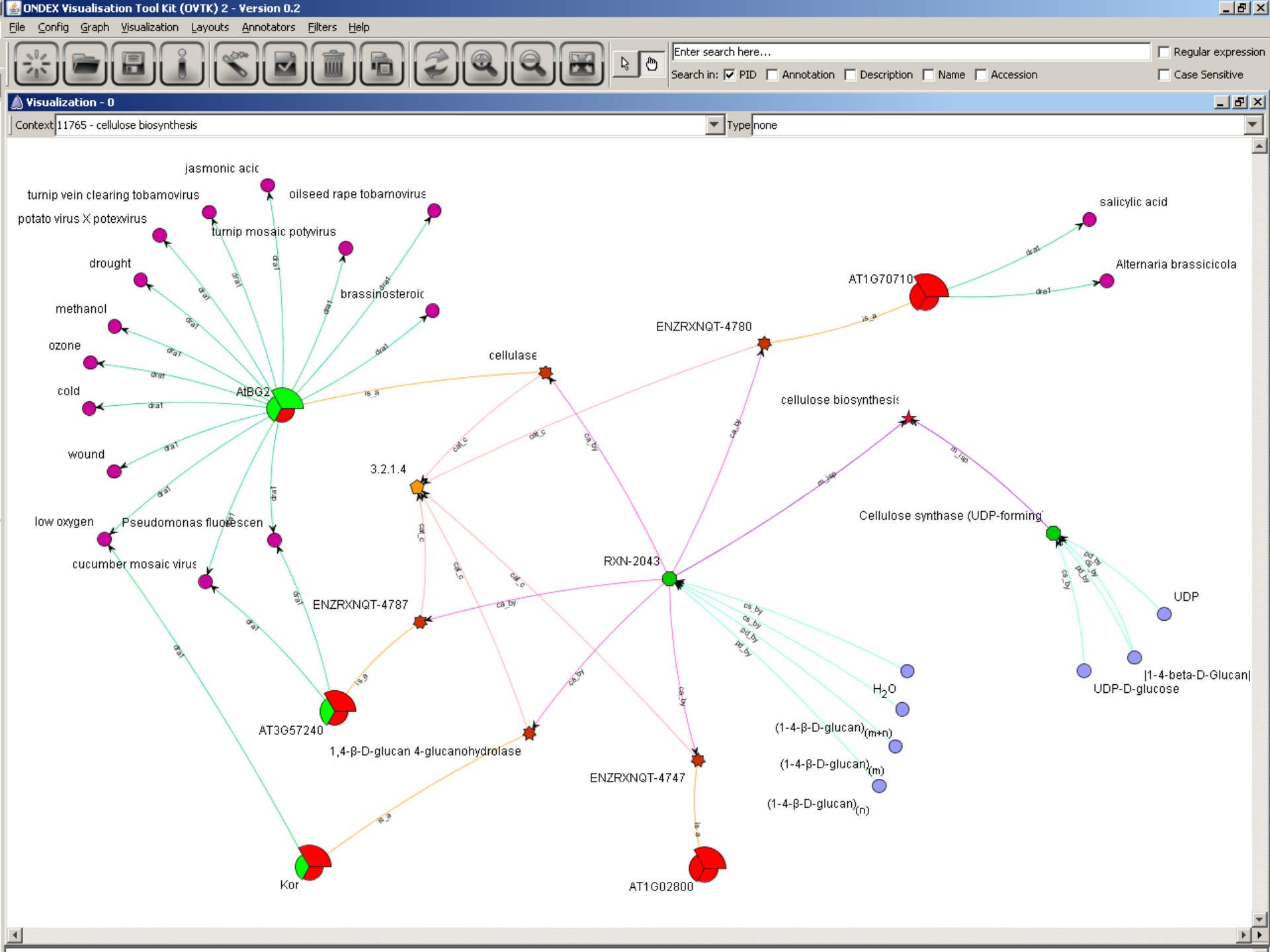
Arabidopsis C/N uptake

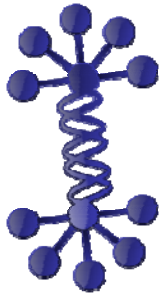


OVTK
Interactive
exploration



Enriched spreadsheet,
e.g. AraCyc pathways


ONDEX Visualisation Tool Kit (OVTK) 2 - Version 0.2
File Config Graph Visualization Layouts Annotators Filters Help
Enter search here...
Regular expression
Search in: PID Annotation Description Name Accession
Case Sensitive
Visualization - 0
Context 11765 - cellulose biosynthesis
Type none
jasmonic acid
turnip vein clearing tobamovirus
oilseed rape tobamovirus
potato virus X potexvirus
turnip mosaic potyvirus
drought
brassinosteroid
methanol
ozone
cold
wound
low oxygen
Pseudomonas fluorescen
cucumber mosaic virus
AtBG2
cellulase
3.2.1.4
ENZRXNQT-4780
salicylic acid
Alternaria brassicicola
AT1G70710
cellulose biosynthesis
Cellulose synthase (UDP-forming
RXN-2043
ENZRXNQT-4787
AT3G57240
1,4-β-D-glucan 4-glucanohydrolase
ENZRXNQT-4747
AT1G02800
Kor
UDP
UDP-D-glucose
|1-4-beta-D-Glucan|
H2O
(1-4-β-D-glucan)(m+n)
(1-4-β-D-glucan)(m)
(1-4-β-D-glucan)(n)



Annotators (3/3)

- Run network statistics such as:
  - Connectivity
  - Centrality
  - Clustering
  - Network diameter

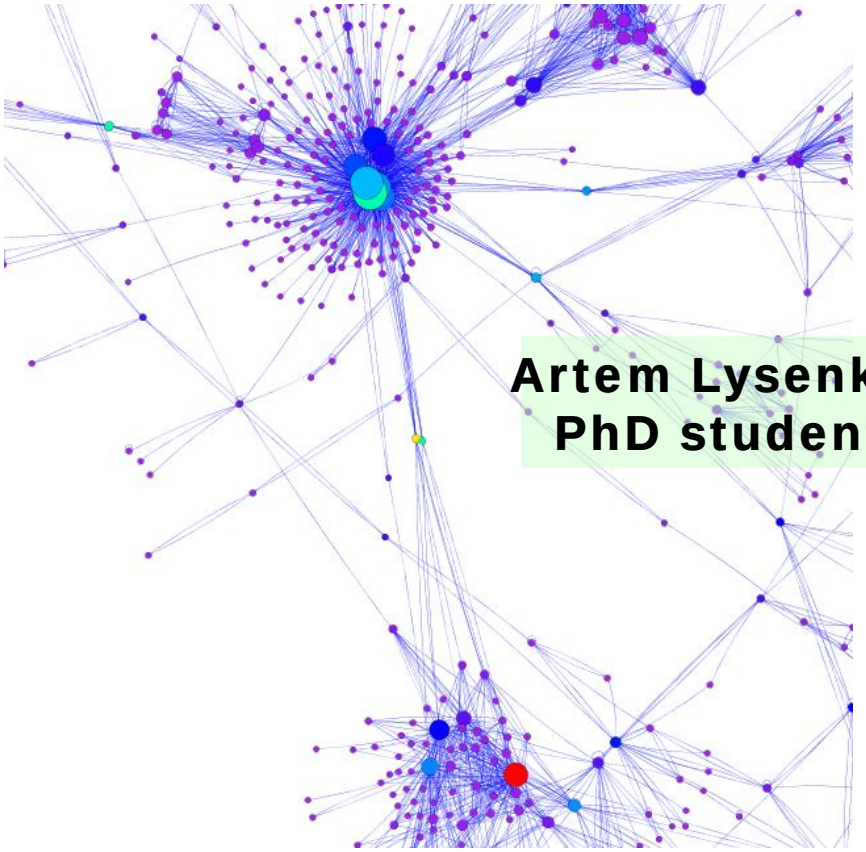
→ Add annotation to the graph

Tutorial

ROTHAMSTED RESEARCH

Arabidopsis PPI network

- Calculate *betweenness* and *degree* centralities
- Identify '*influential*' nodes with low degree and high betweenness
  - yellow, green nodes
  - communication between functional modules?

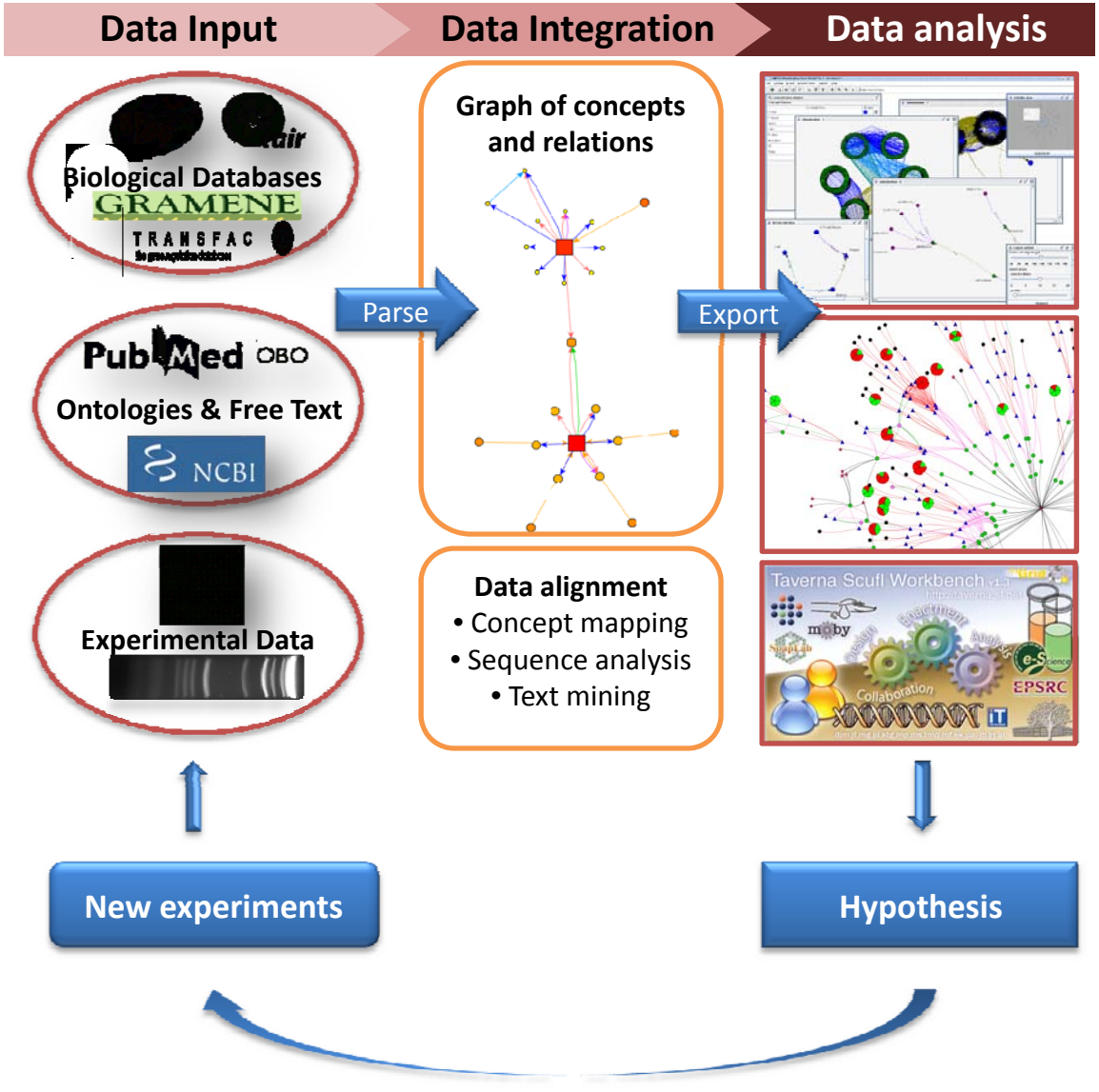


Artem Lysenko, PhD student

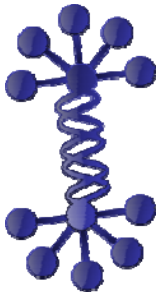
Node size = degree
Node colour = betweenness

ROTHAMSTED RESEARCH

Dynamic process



ROTHAMSTED RESEARCH



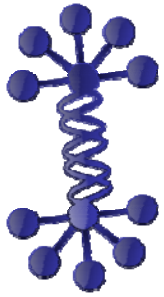
Investigating plant systems Conclusions (1/2)

- Data integration
  - Integrate various heterogeneous data sources
  - Unify and consolidate knowledge
  - Overcome sparseness of data

  - Is difficult

    → semantic and syntactic challenges

ROTHAMSTED RESEARCH



Investigating plant systems
Conclusions (2/2)

- Network analysis
  - Once data is integrated → large volume
    → need to filter down to regions of interest
  - Use various layouts to study merged data
  - Annotate data with network statistics to analyse further

- Ondex = 1 solution
  → Tutorial tomorrow 9-11am

ROTHAMSTED RESEARCH

Acknowledgements

- Catherine Canevet
- Michael Defoin Platel

Rothamsted members:

- Keywan Hassani-Pak
- Matthew Hindle
- Shao Chih Kuo
- Artem Lysenko
- Chris Rawlings
- Mansoor Saqi
- Andrea Splendiani
- Jan Taubert

Former member:

- Jacob Köhler

Biological collaborators:

- Kim Hammond-Kosack
- Martin Urban
- Dimah Habash
- David Wild
- Katherine Denby
- Roxane Legaie