# Exploring interactions in *Arabidopsis*

Graph Based Data Integration: exploring interactions in *Arabidopsis* with contributions from multiple data sources

Artem Lysenko[1], Keywan Hassani-Pak[1], Jan Taubert[1], Charlie Hodgman[2], Christopher J. Rawlings[1], and Mansoor Saqi[1]

[1] Centre for Mathematical and Computational Biology, Rothamsted Research, Harpenden, Herts, AL5 2JQ, UK
{artem.lysenko, keywan.hassani-pak, jan.taubert, chris.rawlings, mansoor.saqi}@bbsrc.ac.uk

[2] Centre for Integrative Biology, University of Nottingham, UK
charlie.hodgman@nottingham.ac.uk
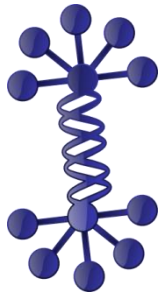
Integration of 3 PPI networks

- IntAct
  - 4625 protein interactions (data derived from literature curation or direct user submissions)
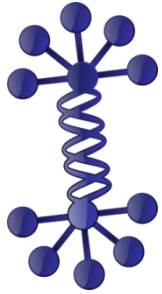
- TAIR (The *Arabidopsis* Information Resource) – 1143 interactions
  - genome sequence, gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications
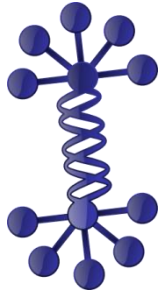
Integration of 3 PPI networks

○ BioGrid (General Repository for Interaction Datasets)

- collections of protein and genetic interactions from major model organism species

- 1223 interactions for *Arabidopsis* derived from high-throughput studies and conventional focused studies

Ondex Accession-based mapping

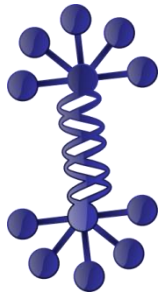○ IntAct

○ TAIR

○ BioGRID

→ Mapping the 3 databases based on TAIR accessions
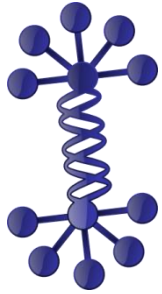
Adding 3 sources of evidence

○ co-expression

○ sequence similarity

○ co-occurrence in scientific literature

→ facilitate the identification of functionally related groups of proteins

Co-expression evidence

○ ATTED II (_Arabidopsis_ _thaliana_ trans-factor and cis-element prediction database)

- provides co-regulated gene relationships in _Arabidopsis_ to estimate gene functions
- gives the Pearson correlation coefficients of co-expressed genes in _Arabidopsis_ calculated from available microarray data
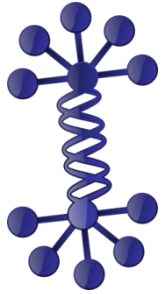
→ Mapping using TAIR accessions
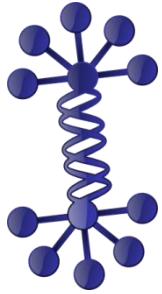
Sequence similarity evidence

○ NCBI PSI-BLAST

- identify similarities between our reference set of proteins

- Matching against *Arabidopsis* subset of UNIPROT
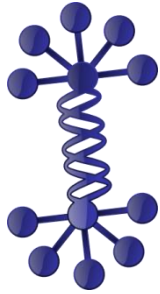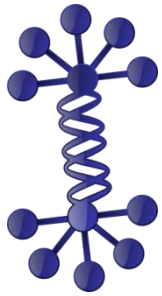
Evidence from the literature

○ Co-occurrence of protein names

- 25,900 Medline abstracts related to *Arabidopsis Thaliana*

- Integrated Lucene-based mapping method

Integrated *Arabidopsis* PPI

- Nodes are connected if there is evidence from
  - Interaction
  - co-expression
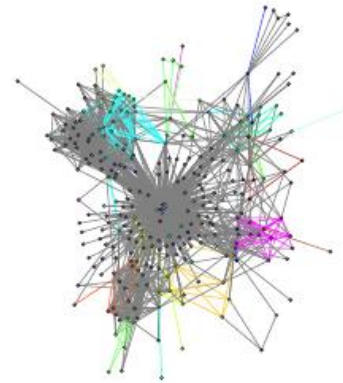 - sequence similarity
  - co-occurrence
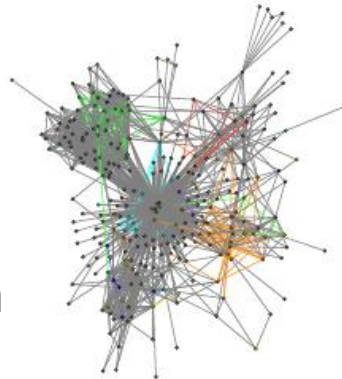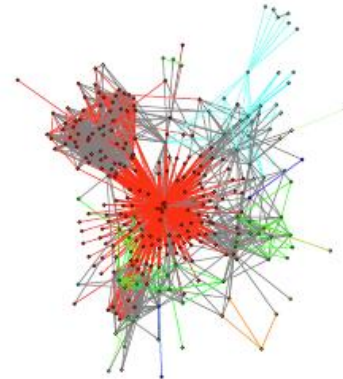
Evidence networks in isolation
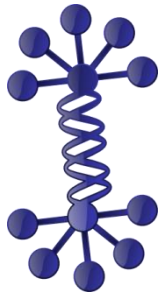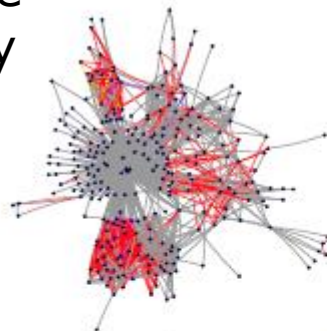
Sequence similarity
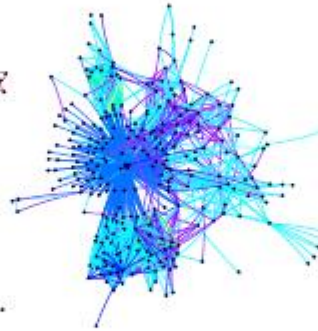
Co-occurrence

Co-expression

Protein interaction
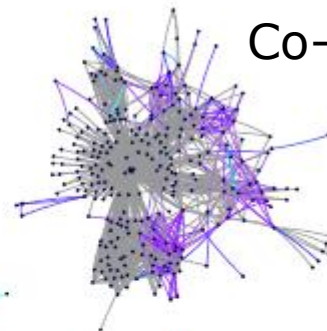
Relations colour = strength of evidence
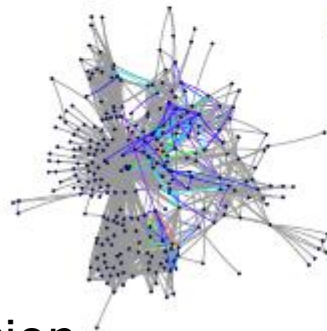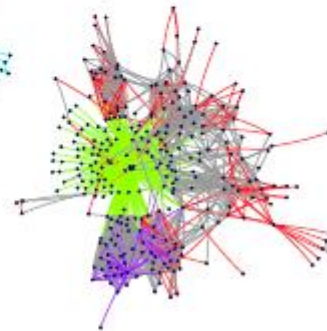

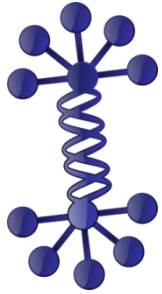Sequence similarity

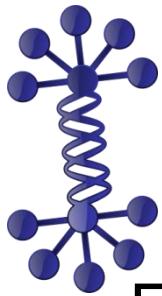Combined

Co-occurrence

Co-expression

Protein interaction

Added attributes to nodes/edges

○ Network stats

- Betweenness centrality (BWC)
  → How influential (bridge)
- Degree centrality (DC)
  → Hub likeness

○ Markov Clustering

- Identifies strongly connected groups of proteins in the network

three_ppi_added_evidence.xml.gz / ABA_cluster.xml.gz (only 1 cluster)

- Example 1:
  - Ann → Scale/Colour Relations by Numerical Value (BLAST weight) (size min 4 max 4) [only BLAST network left]
  - Ann → Colour Concepts by General Attribute (BLAST Cluster)

- Example 2:
  - Ann → Scale/Colour Relations by Numerical Value (interaction weight) (size min 4 max 4)
  - Filter → More → Threshold, relations (interaction EBWC)

- Example 3:
  - Ann → Scale/Colour Relations by Numerical Value (interaction weight) (size min 4 max 4)
  - Ann → Scale/Colour Concepts by Numerical Value (Combined DC) (size min 10 max 100)