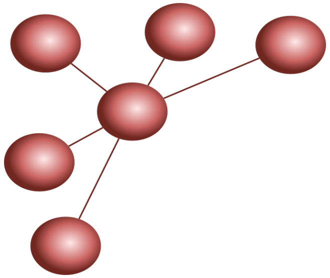
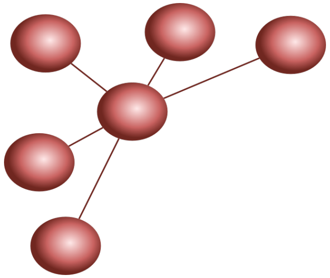


DUBii 2021

Introduction to Network Science
Introduction to Network Biology

Anaïs Baudot anais.baudot@univ-amu.fr

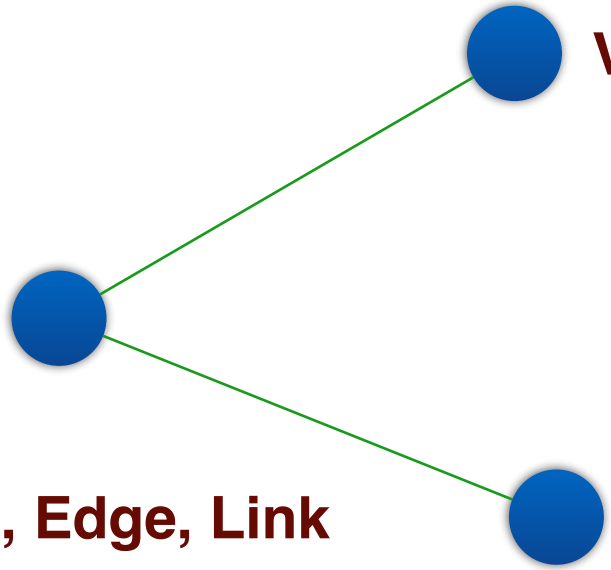


Network Science

- Start 21th century
- Roots on graph theory
- In the context of data production and computer sciences
- closely linked to the study of complex systems

Definition

Graph, Network, Web



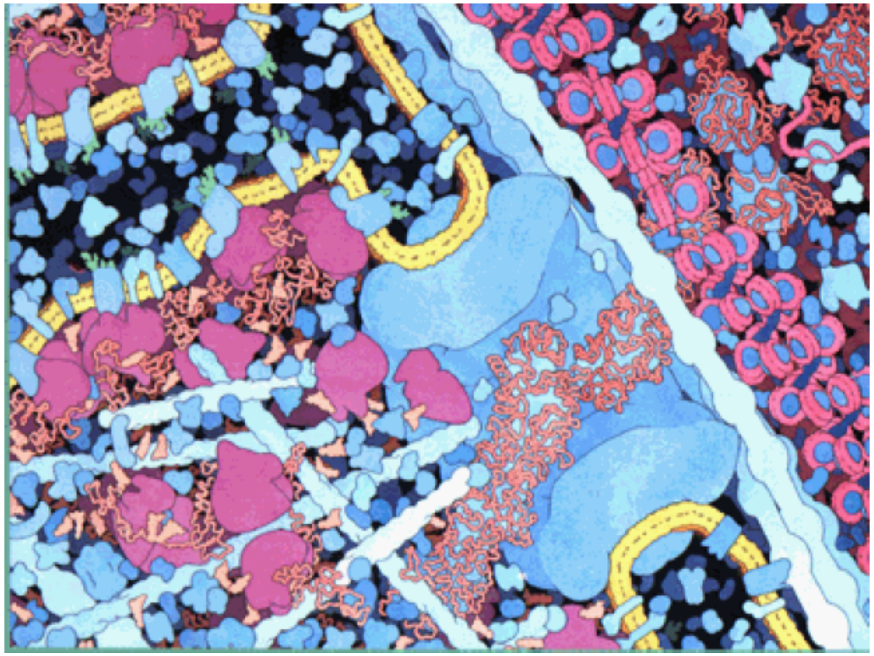
Vertice, Node

Interaction, Edge, Link

G={V, E}

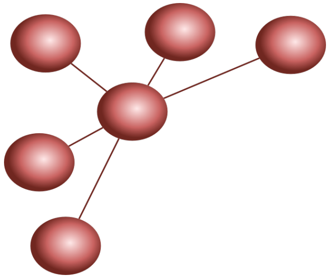
Topology, Motifs …

Macromolecules do not act isolated, but interact with each other to perform their functions

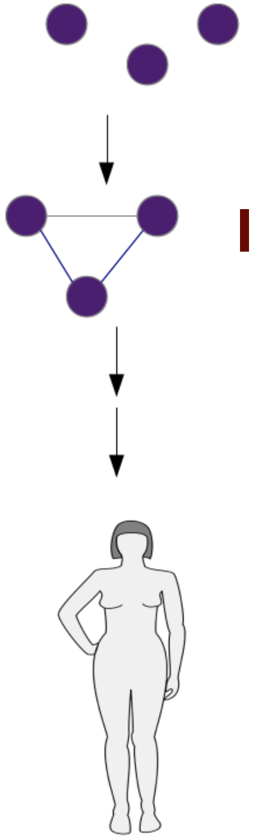


Molecular interactions:
Protein-Protein
Protein-DNA
Protein-RNA
Protein-Lipid
...

Transient, stable, obligatory ...



Systems Biology because living organisms are complex systems



Systems component : genes/proteins

Interaction Networks

Emerging properties : phenotypes

- **Global/collective behaviour cannot be deduced from the knowledge on the components**
- **Phenotype does not emerge from isolated biological molecules but from their interactions**

Part 1: Building Biological Networks

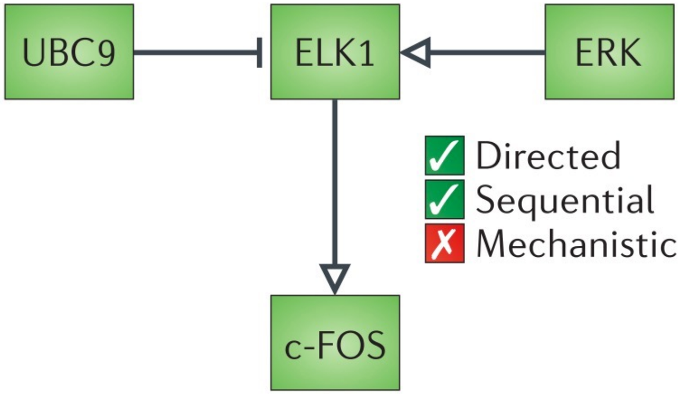
- From literature, knowledge, curation
- From large-scale interaction experiments
- From inference from large-scale omics data

Part 1: Building Biological Networks

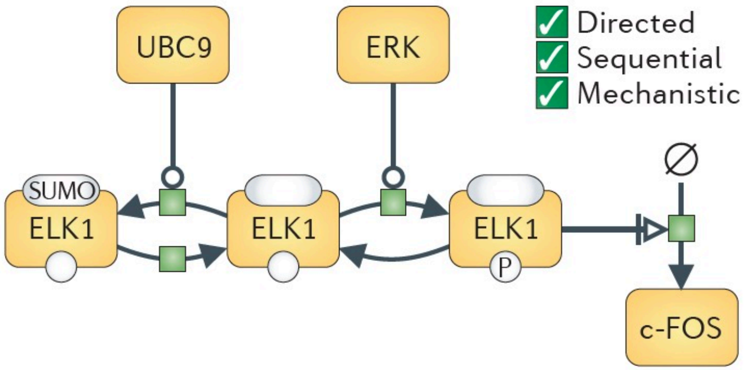
- **From literature, knowledge, curation**
- From large-scale interaction experiments
- From inference from large-scale omics data

Building biological networks from literature, knowledge, curation

b Activity flows

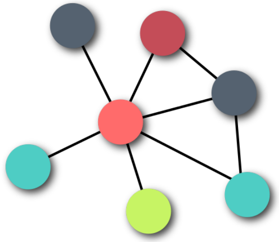


c Process descriptions

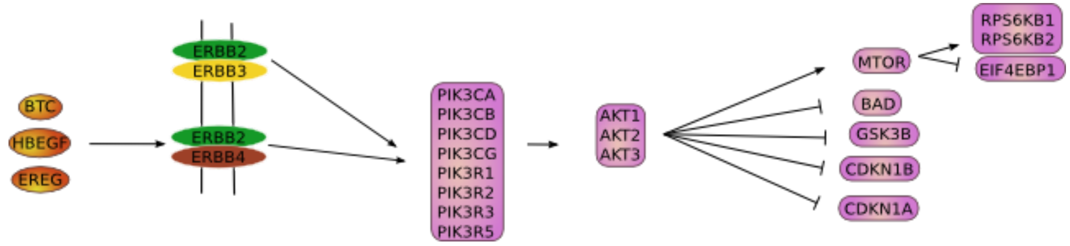


Le Novère et al. Nature Review Genetics. 2015

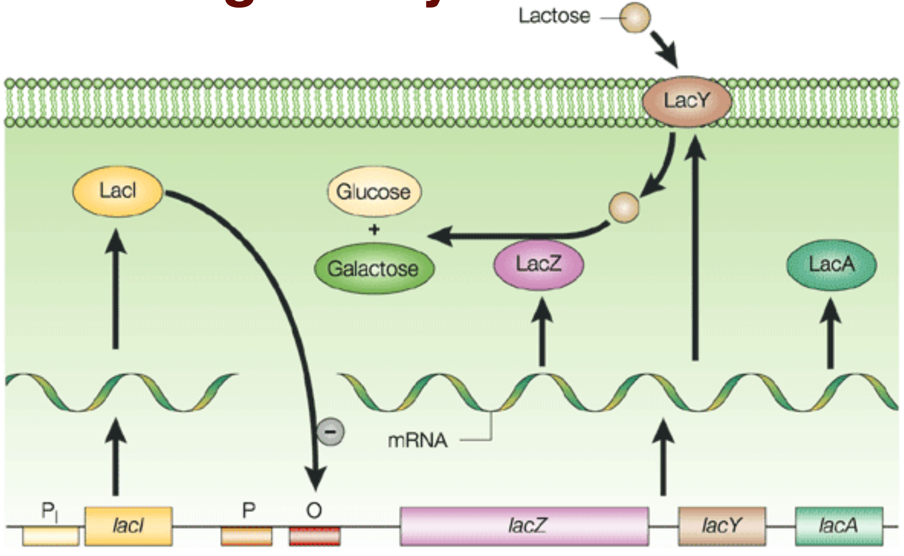
Activity flow / Gene Regulatory Networks / Influence Graphs



Signalling networks

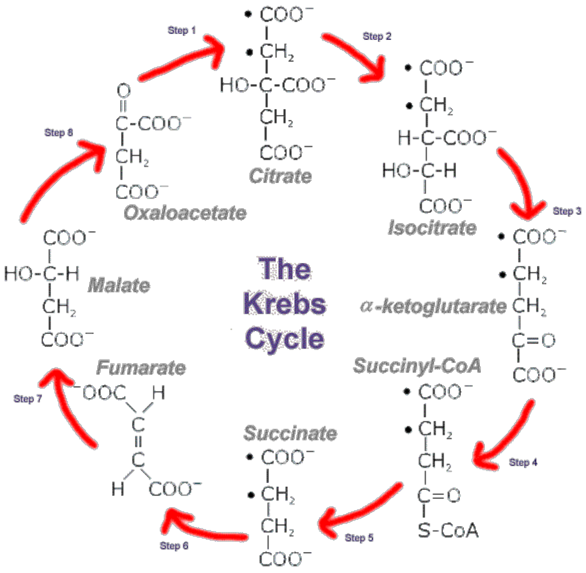


Gene Regulatory networks



Nature Reviews | Genetics

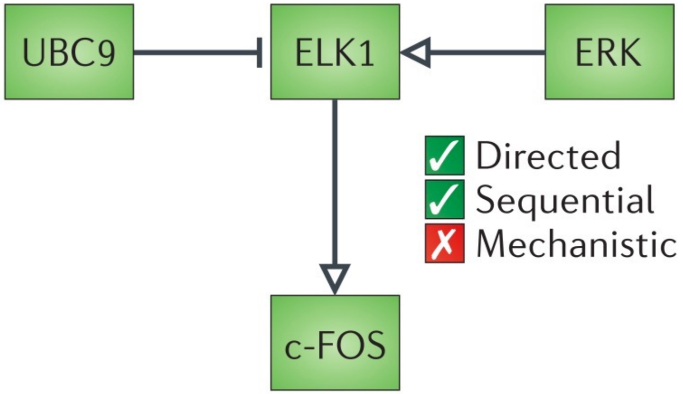
Metabolic Cycle



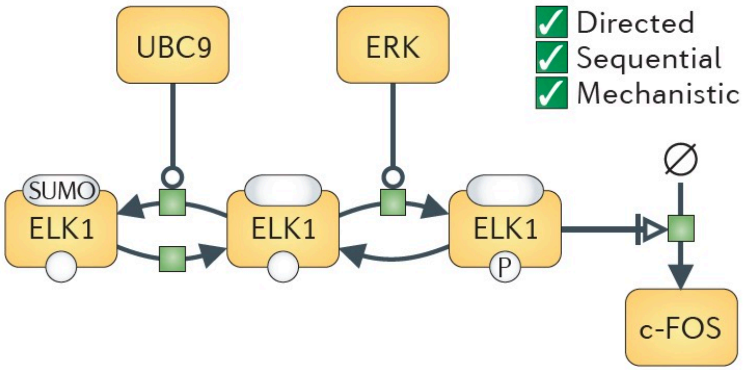
2 types of nodes : enzymes & substrates, reaction directional or bidirectional

Building biological networks from literature, knowledge, curation

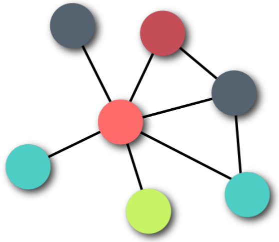
b Activity flows



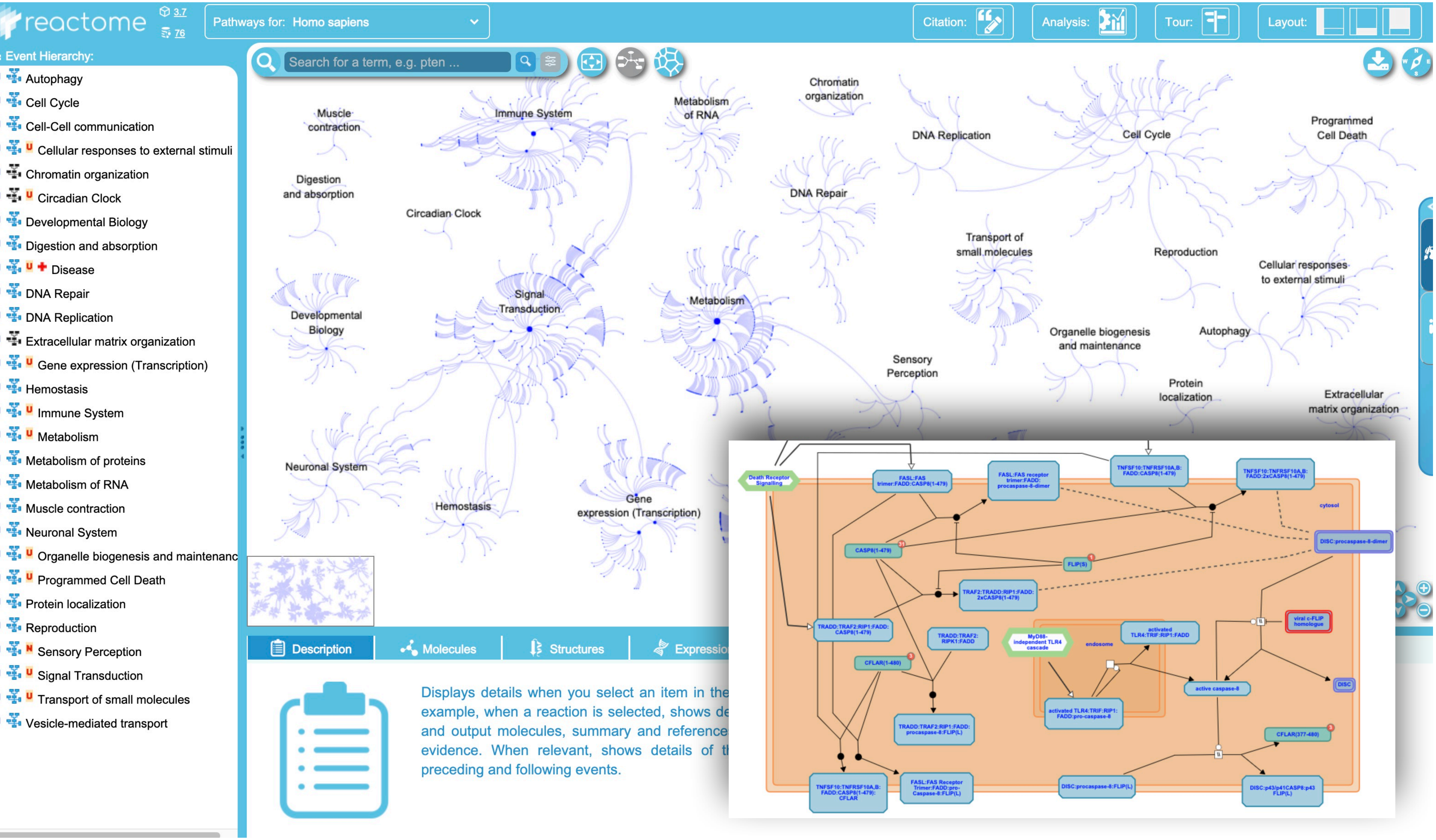
c Process descriptions



=> pathway databases
=> mathematical modelling



pathway databases: Reactome

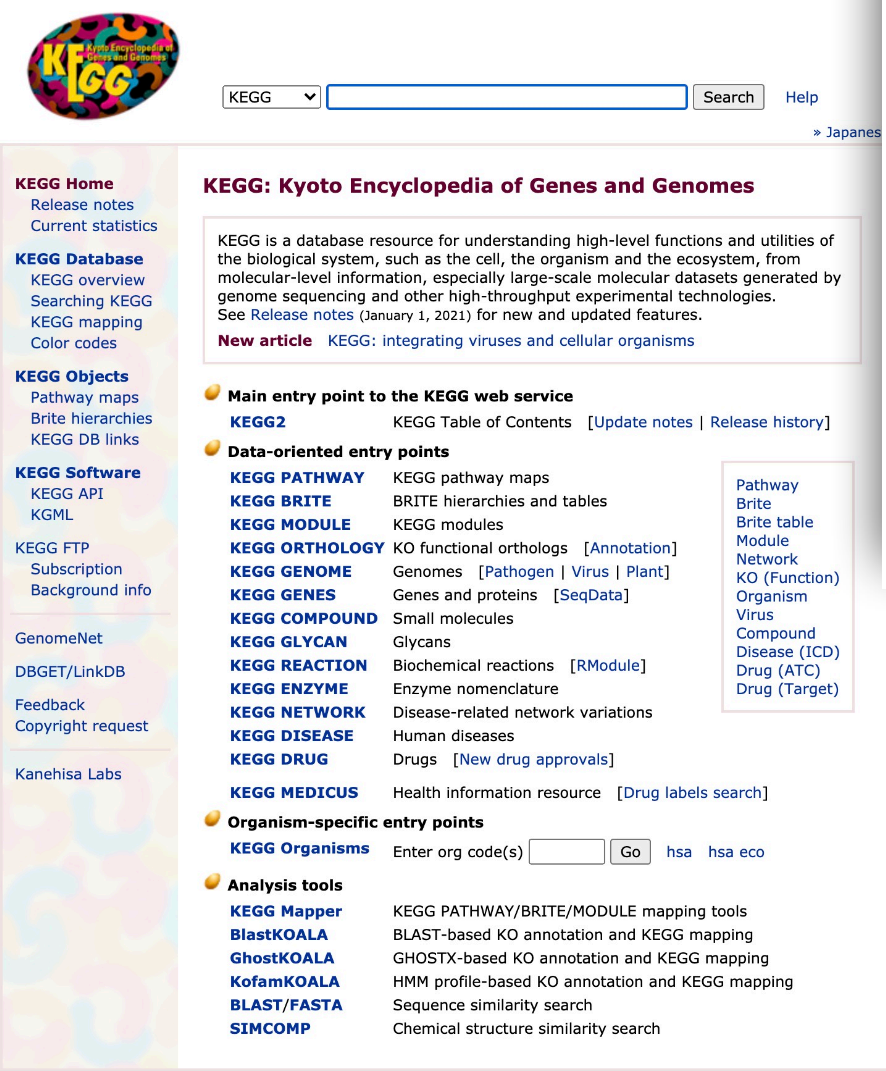


Event Hierarchy:

- Autophagy
- Cell Cycle
- Cell-Cell communication
- U Cellular responses to external stimuli
- Chromatin organization
- U Circadian Clock
- Developmental Biology
- Digestion and absorption
- U ✚ Disease
- DNA Repair
- DNA Replication
- Extracellular matrix organization
- U Gene expression (Transcription)
- Hemostasis
- U Immune System
- U Metabolism
- Metabolism of proteins
- Metabolism of RNA
- Muscle contraction
- Neuronal System
- U Organelle biogenesis and maintenanc
- U Programmed Cell Death
- Protein localization
- Reproduction
- N Sensory Perception
- U Signal Transduction
- U Transport of small molecules
- Vesicle-mediated transport

Displays details when you select an item in the
example, when a reaction is selected, shows de
and output molecules, summary and reference
evidence. When relevant, shows details of th
preceding and following events.

pathway databases: Kegg



KEGG Home
Release notes
Current statistics

KEGG Database
KEGG overview
Searching KEGG
KEGG mapping
Color codes

KEGG Objects
Pathway maps
Brite hierarchies
KEGG DB links

KEGG Software
KEGG API
KGML

KEGG FTP
Subscription
Background info

GenomeNet

DBGET/LinkDB

Feedback
Copyright request

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
See Release notes (January 1, 2021) for new and updated features.

New article KEGG: integrating viruses and cellular organisms

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents [Update notes | Release history]

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps
KEGG BRITE	BRITE hierarchies and tables
KEGG MODULE	KEGG modules
KEGG ORTHOLOGY	KO functional orthologs [Annotation]
KEGG GENOME	Genomes [Pathogen | Virus | Plant]
KEGG GENES	Genes and proteins [SeqData]
KEGG COMPOUND	Small molecules
KEGG GLYCAN	Glycans
KEGG REACTION	Biochemical reactions [RModule]
KEGG ENZYME	Enzyme nomenclature
KEGG NETWORK	Disease-related network variations
KEGG DISEASE	Human diseases
KEGG DRUG	Drugs [New drug approvals]
KEGG MEDICUS	Health information resource [Drug labels search]

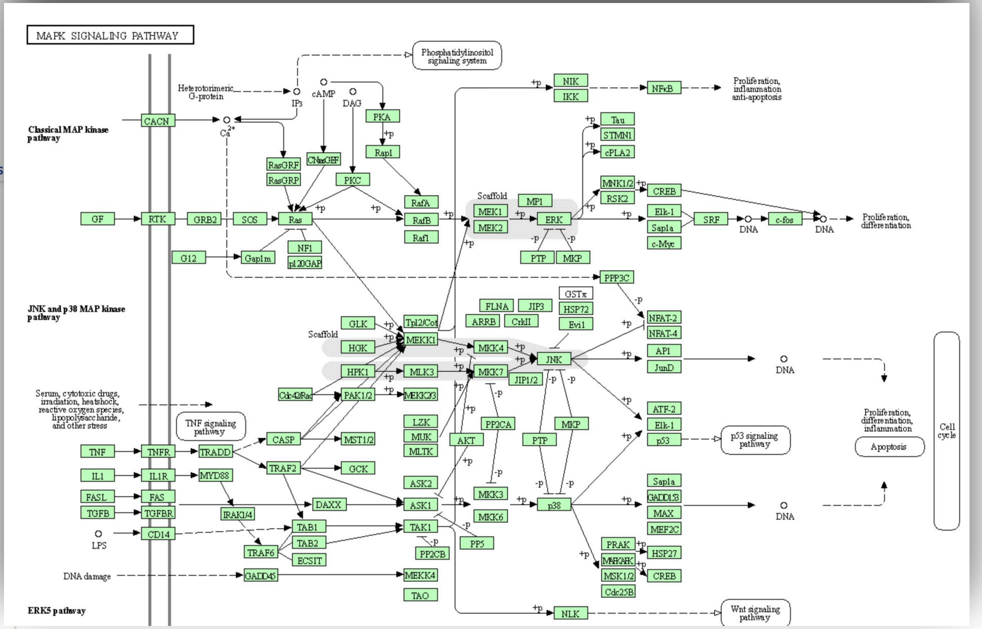
Pathway
Brite
Brite table
Module
Network
KO (Function)
Organism
Virus
Compound
Disease (ICD)
Drug (ATC)
Drug (Target)

Organism-specific entry points

KEGG Organisms Enter org code(s) ☐ Go hsa hsa eco

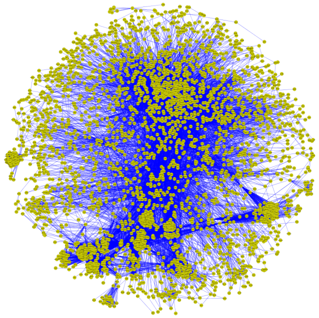
Analysis tools

KEGG Mapper	KEGG PATHWAY/BRITE/MODULE mapping tools
BlastKOALA	BLAST-based KO annotation and KEGG mapping
GhostKOALA	GHOSTX-based KO annotation and KEGG mapping
KofamKOALA	HMM profile-based KO annotation and KEGG mapping
BLAST/FASTA	Sequence similarity search
SIMCOMP	Chemical structure similarity search



pathway databases

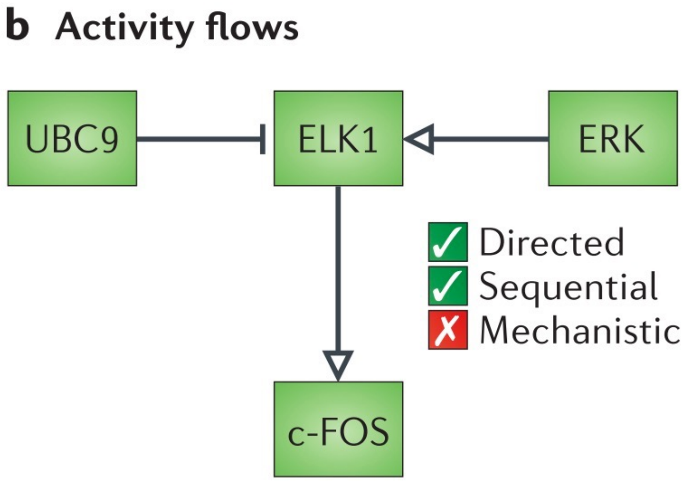
Pathways

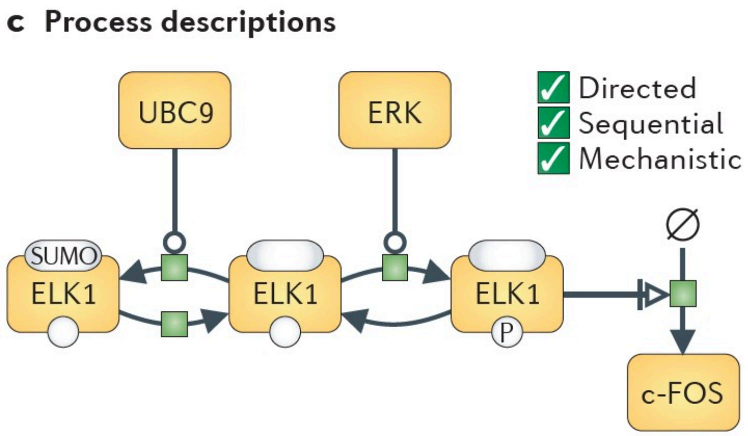


~250 000 edges

Curated networks

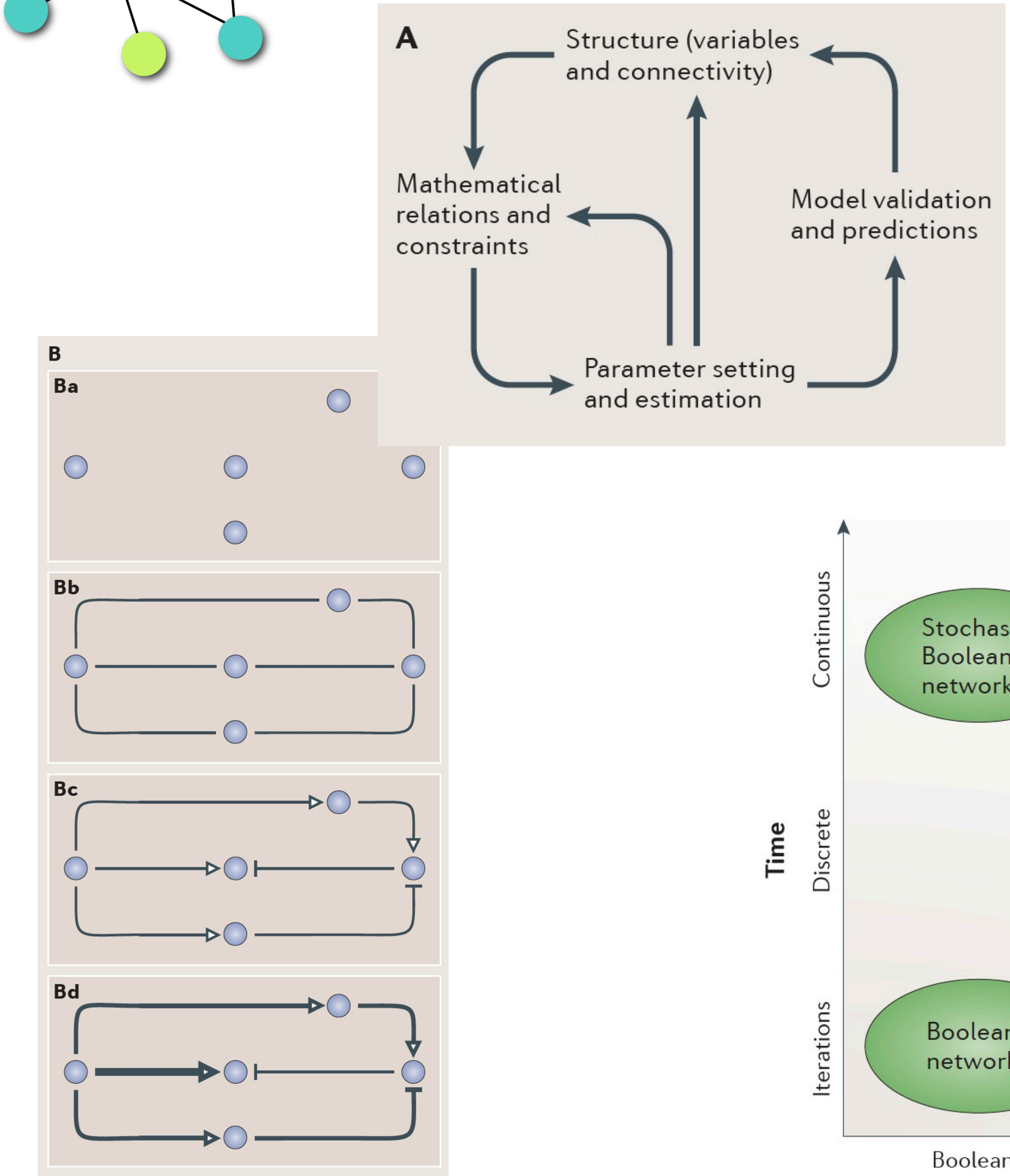
Building biological networks from literature, knowledge, curation

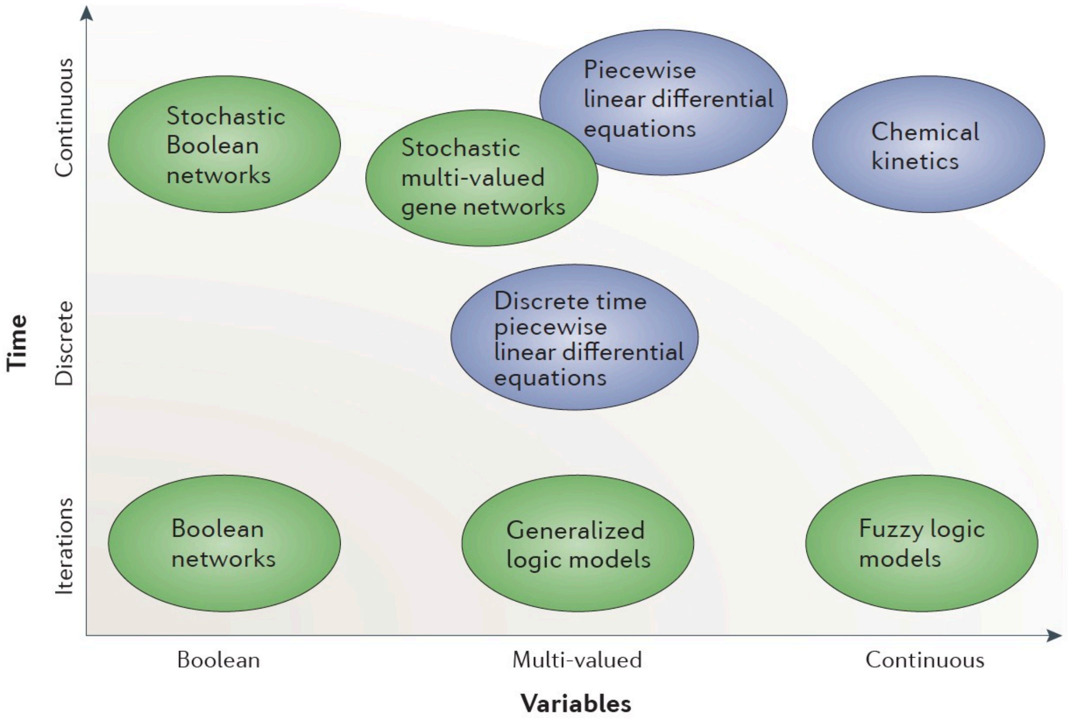


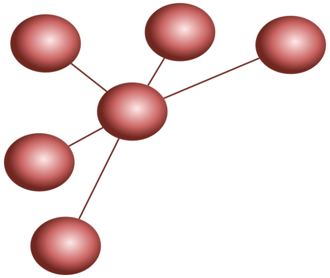


=> pathway databases
=> mathematical modelling

Building mathematical models of biological processes







Part 1: Building Biological Networks

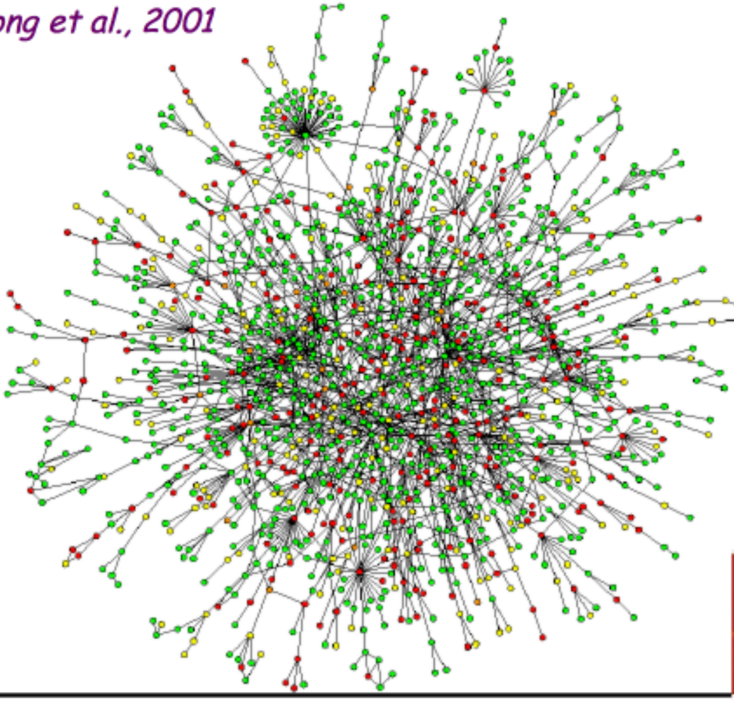
- From literature, knowledge, curation
- **From large-scale interaction experiments**
- From inference from large-scale omics data

Experimental interaction screenings

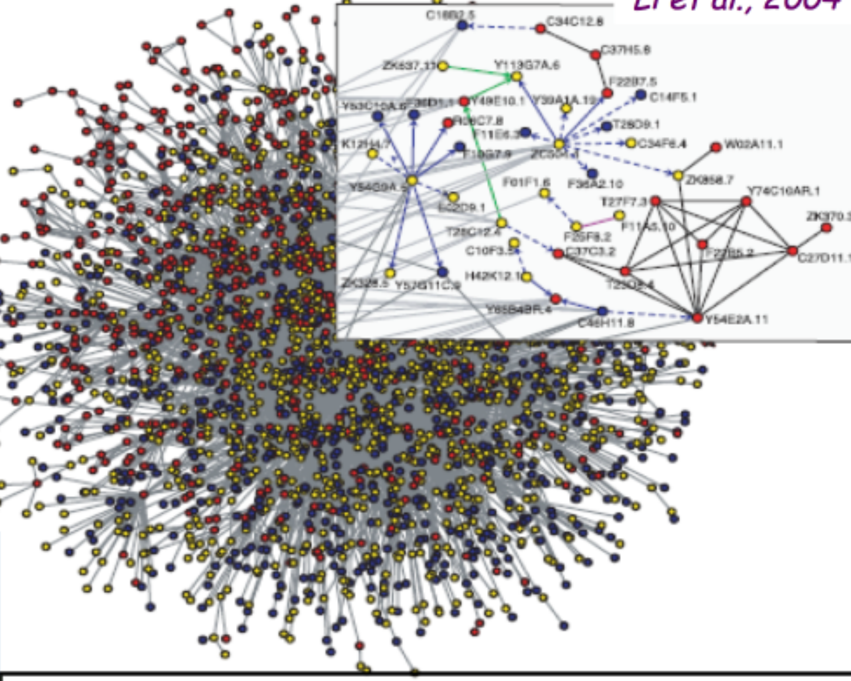
- Protein-protein interactions
- Protein-DNA interactions
- Protein-RNA interactions
- …

=> Small-scale / Large-scale

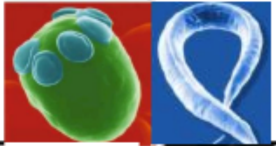
Y2H screens

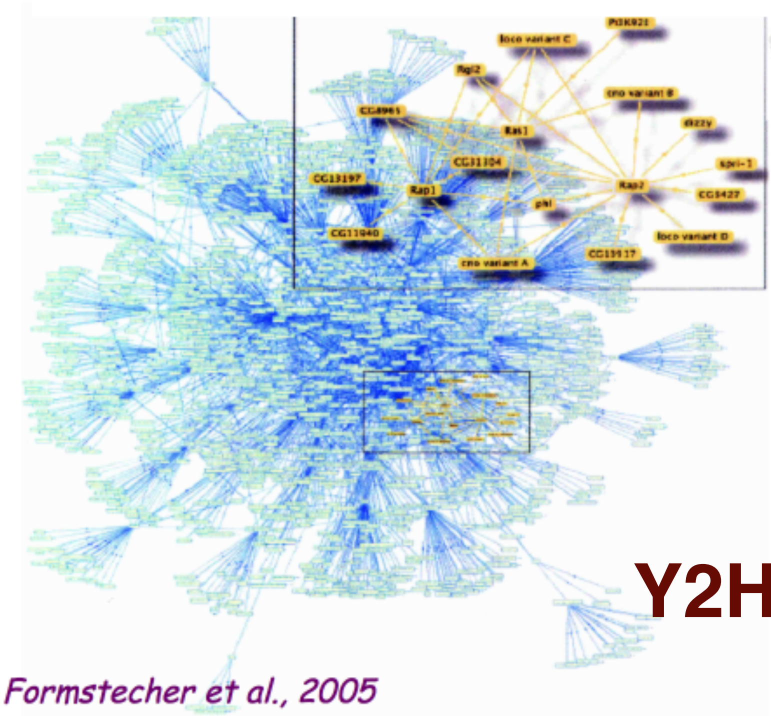


Jeong et al., 2001

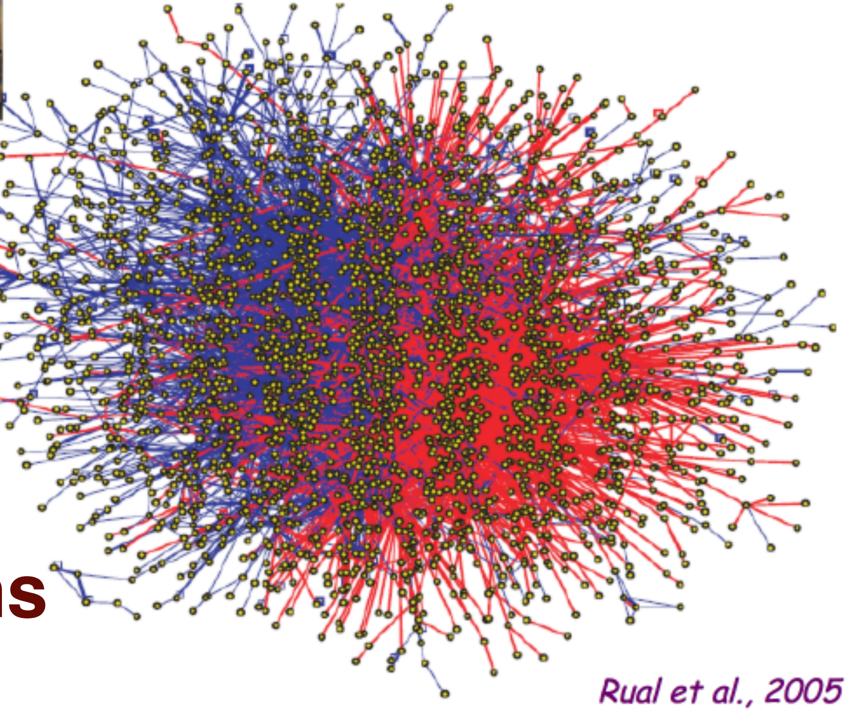


Li et al., 2004

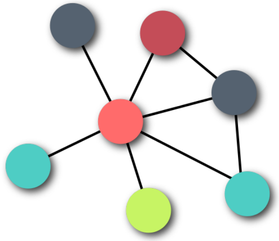


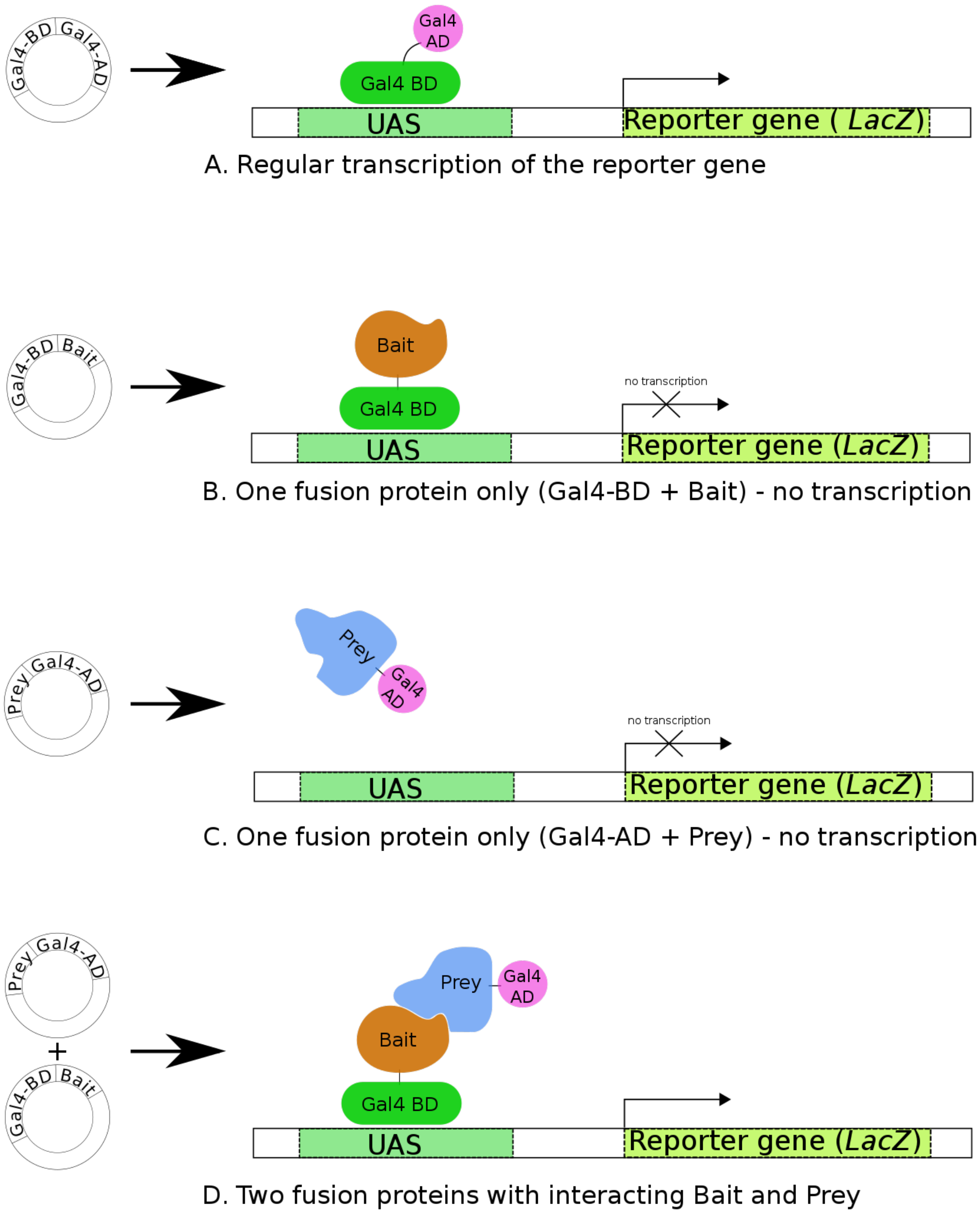


Formstecher et al., 2005



Rual et al., 2005





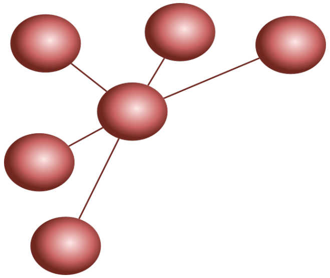
A. Regular transcription of the reporter gene

B. One fusion protein only (Gal4-BD + Bait) - no transcription

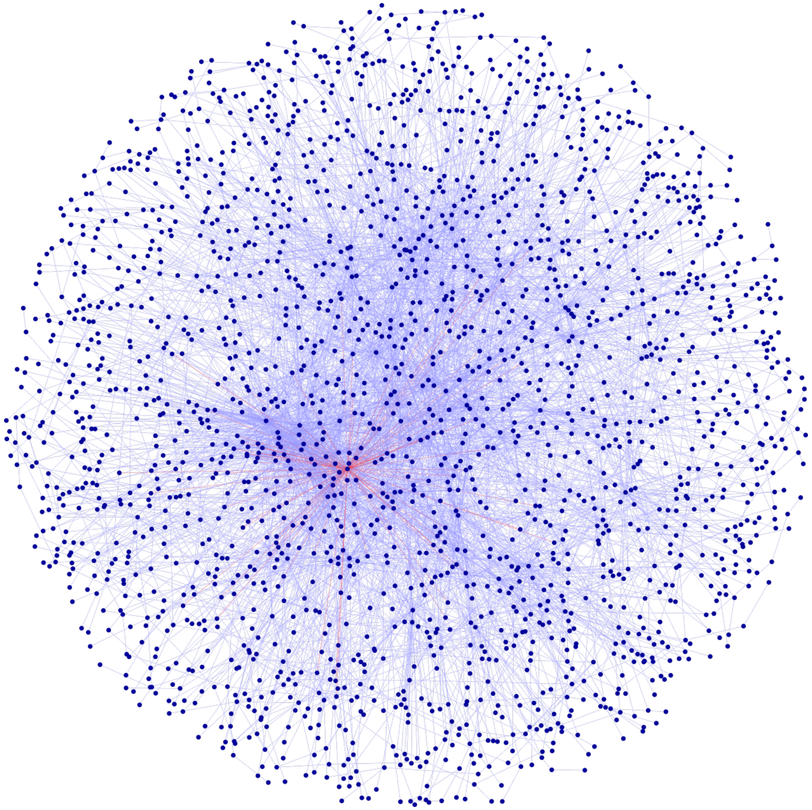
C. One fusion protein only (Gal4-AD + Prey) - no transcription

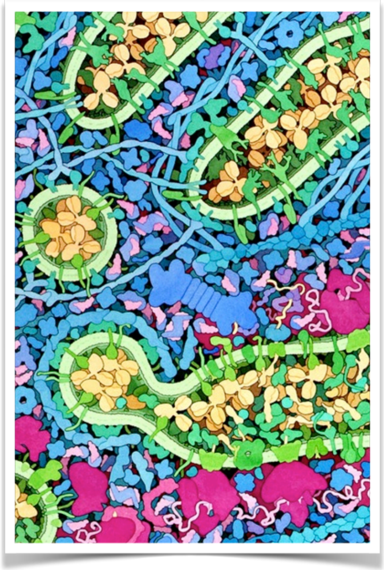
D. Two fusion proteins with interacting Bait and Prey

By Anna - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=2890233



Interactomes: protein-protein interactions



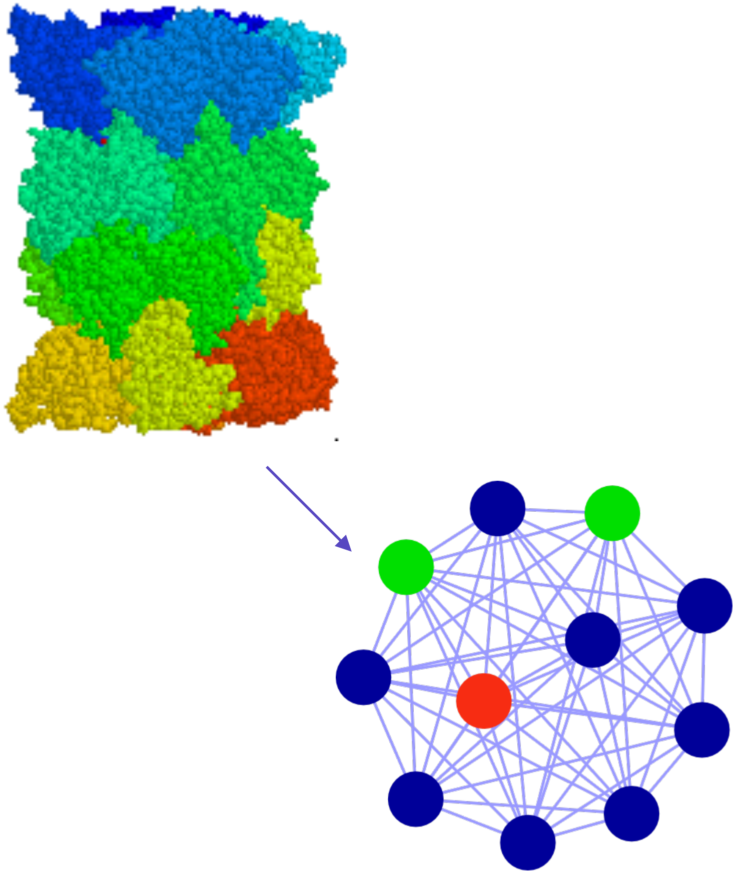


Set of detected protein-protein interactions

Physical interactions, but physiological interactions ?

Interactomes are devoid of spatio-temporal information

Interactomes: Pull-Down



Protein complexes

False-negatives and “sparse networks”

Interaction space (to be discovered)

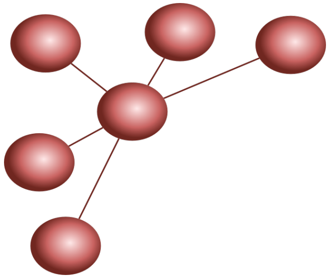
Interactions discovered by method 1

Interactions discovered by method 2

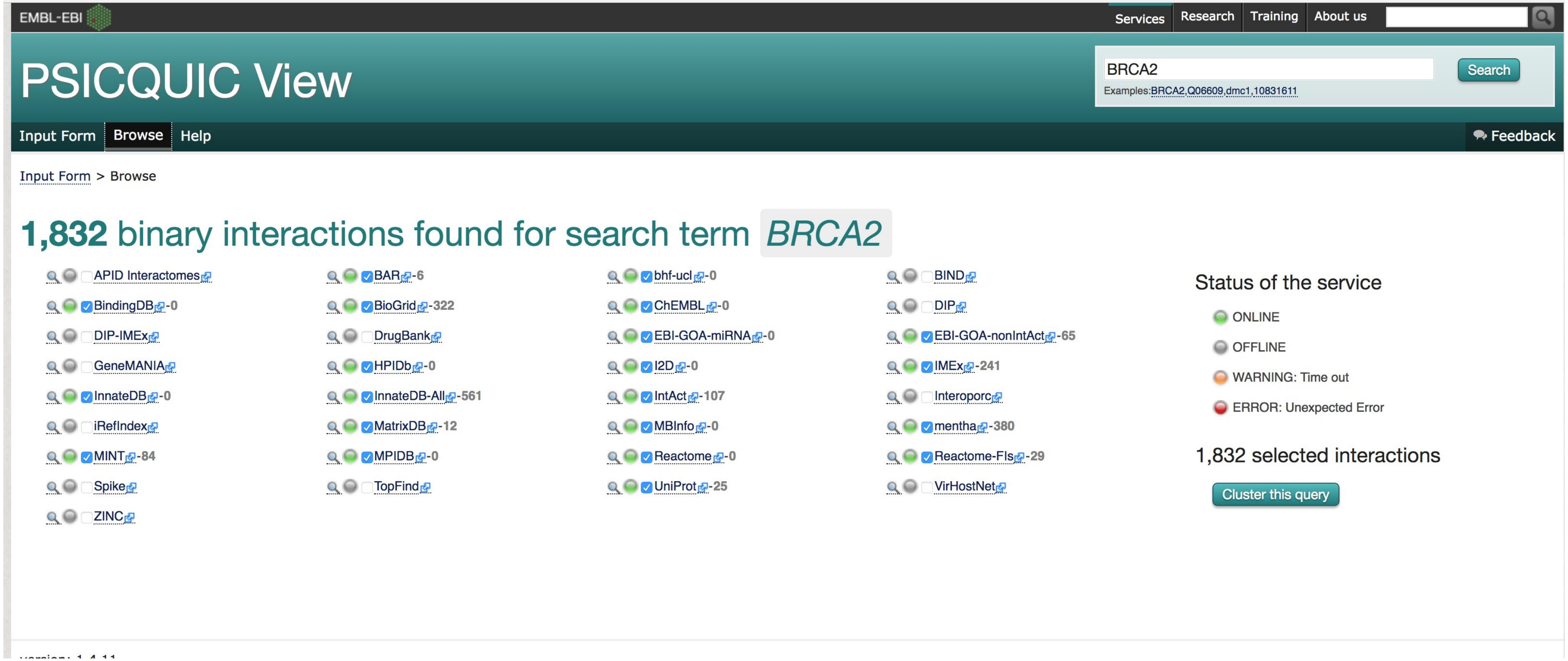
Interaction databases

Multi-organisms:
- DIP (dip.doe-mbi.ucla.edu)
- **IntAct** (www.ebi.ac.uk/intact)
- MINT (mint.bio.uniroma2.it/mint)
- BioGRID (www.thebiogrid.org)
- BIND (www.blueprint.org)

International
Molecular
Exchange
Consortium



PSICQUIC portal



PSICQUIC View

BRCA2

Examples: BRCA2, Q06609, dmc1, 10831611

Input Form > Browse

1,832 binary interactions found for search term *BRCA2*

- ○ ☐ APID Interactomes
- ● ☑ BAR-6
- ● ☑ bhf-ucl-0
- ○ ☐ BIND
- ● ☑ BindingDB-0
- ● ☑ BioGrid-322
- ● ☑ ChEMBL-0
- ○ ☐ DIP
- ○ ☐ DIP-IMEx
- ○ ☐ DrugBank
- ● ☑ EBI-GOA-miRNA-0
- ● ☑ EBI-GOA-nonIntAct-65
- ○ ☐ GeneMANIA
- ● ☑ HPIDb-0
- ● ☑ I2D-0
- ● ☑ IMEx-241
- ● ☑ InnateDB-0
- ● ☑ InnateDB-All-561
- ● ☑ IntAct-107
- ○ ☐ Interoporc
- ○ ☐ iRefIndex
- ● ☑ MatrixDB-12
- ● ☑ MBInfo-0
- ● ☑ mentha-380
- ● ☑ MINT-84
- ● ☑ MPIDB-0
- ● ☑ Reactome-0
- ● ☑ Reactome-FIs-29
- ○ ☐ Spike
- ○ ☐ TopFind
- ● ☑ UniProt-25
- ○ ☐ VirHostNet
- ○ ☐ ZINC

Status of the service

- 🟢 ONLINE
- ⚪ OFFLINE
- 🟠 WARNING: Time out
- 🔴 ERROR: Unexpected Error

1,832 selected interactions

version: 1.4.11

Part 1: Building Biological Networks

- From literature, knowledge, curation
- From large-scale interaction experiments
- **From inference from large-scale omics data**

Network inference from -omics data

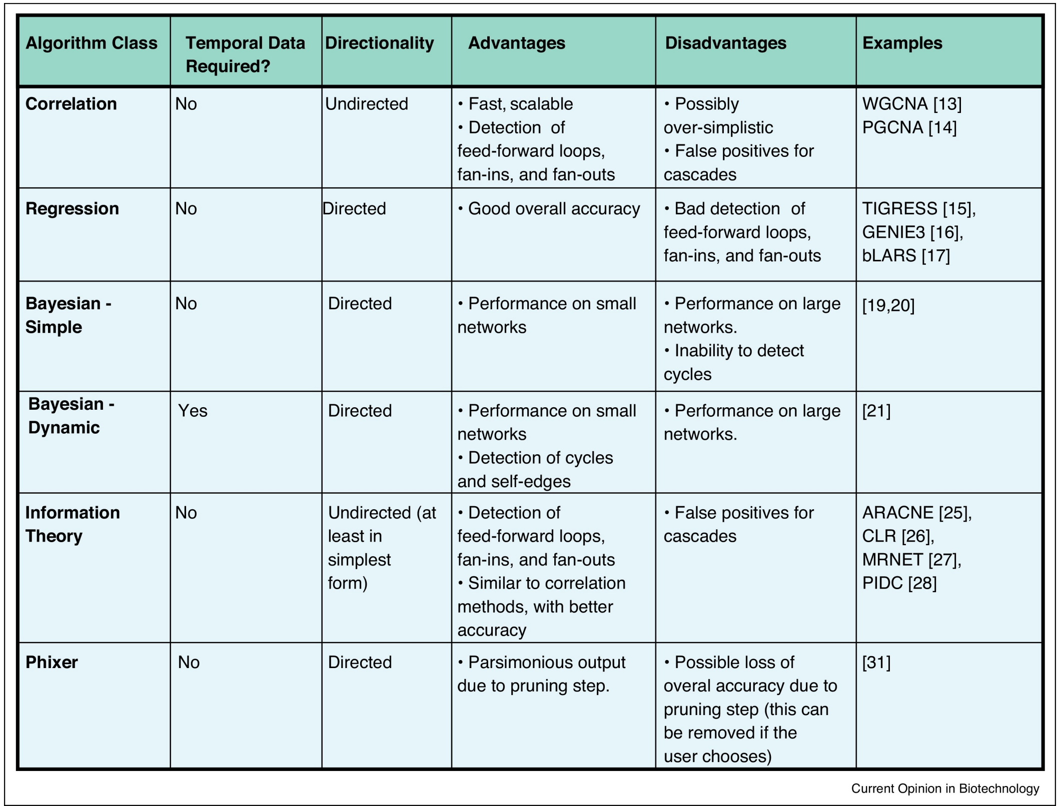
- Inferring/learning regulatory interactions from gene expression data (time-series, perturbation experiments)

- Famous methods: WGCNA, GENIE3

- Now on single-cells

Greenfield A, Madar A, Ostrer H, Bonneau R (2010) DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. PLoS ONE 5(10): e13397.

Saint-Antoine, M. M. & Singh, A. Network inference in systems biology: recent developments, challenges, and applications. *Current Opinion in Biotechnology* **63**, 89–98 (2020).

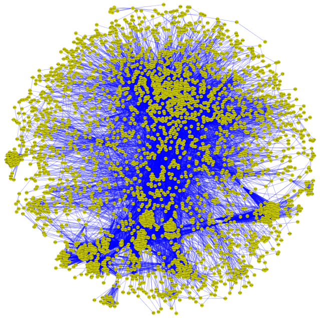


Algorithm Class	Temporal Data Required?	Directionality	Advantages	Disadvantages	Examples
Correlation	No	Undirected	• Fast, scalable • Detection of feed-forward loops, fan-ins, and fan-outs	• Possibly over-simplistic • False positives for cascades	WGCNA [13] PGCNA [14]
Regression	No	Directed	• Good overall accuracy	• Bad detection of feed-forward loops, fan-ins, and fan-outs	TIGRESS [15], GENIE3 [16], bLARS [17]
Bayesian - Simple	No	Directed	• Performance on small networks	• Performance on large networks. • Inability to detect cycles	[19,20]
Bayesian - Dynamic	Yes	Directed	• Performance on small networks • Detection of cycles and self-edges	• Performance on large networks.	[21]
Information Theory	No	Undirected (at least in simplest form)	• Detection of feed-forward loops, fan-ins, and fan-outs • Similar to correlation methods, with better accuracy	• False positives for cascades	ARACNE [25], CLR [26], MRNET [27], PIDC [28]
Phixer	No	Directed	• Parsimonious output due to pruning step.	• Possible loss of overal accuracy due to pruning step (this can be removed if the user chooses)	[31]

Current Opinion in Biotechnology

Interactome(s)

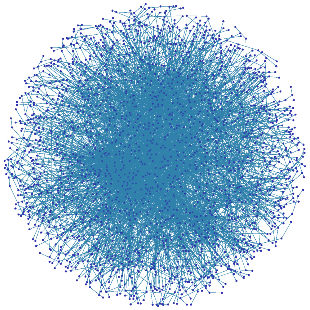
Pathways



~250 000 edges

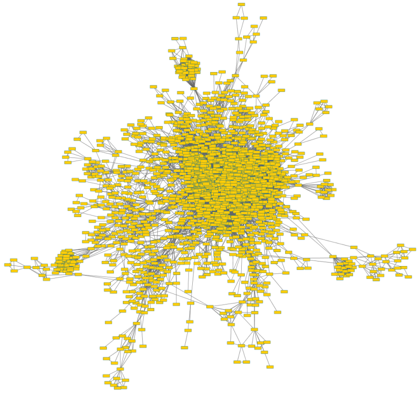
Curated
networks

PPI



~60 000 edges

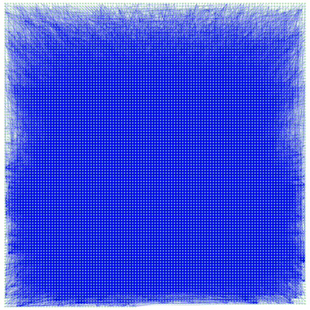
Complexes



~40 000 edges

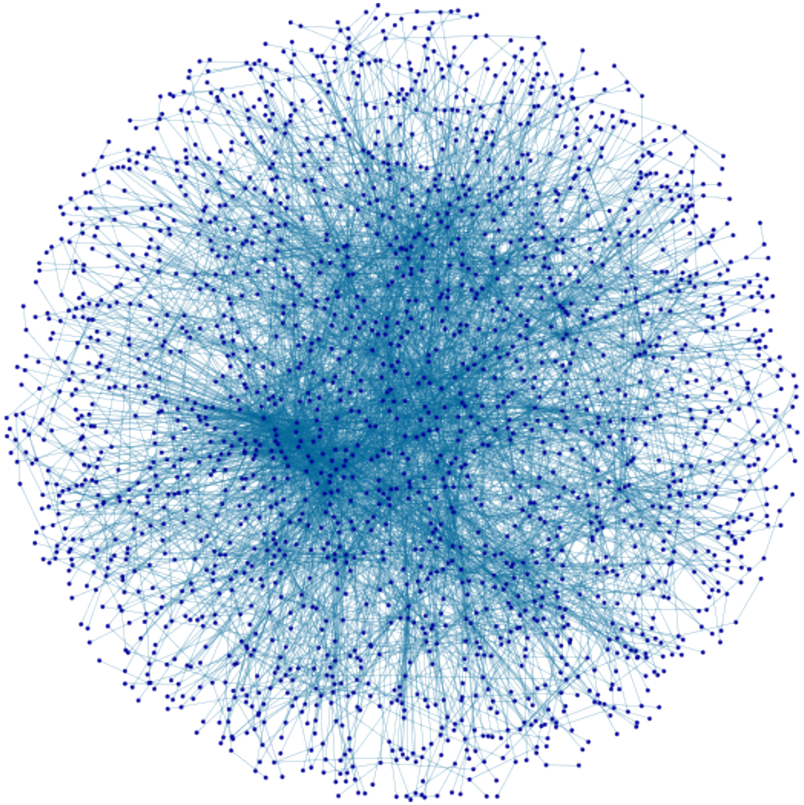
Measured
networks

Correlation of expression



~1 400 000 edges

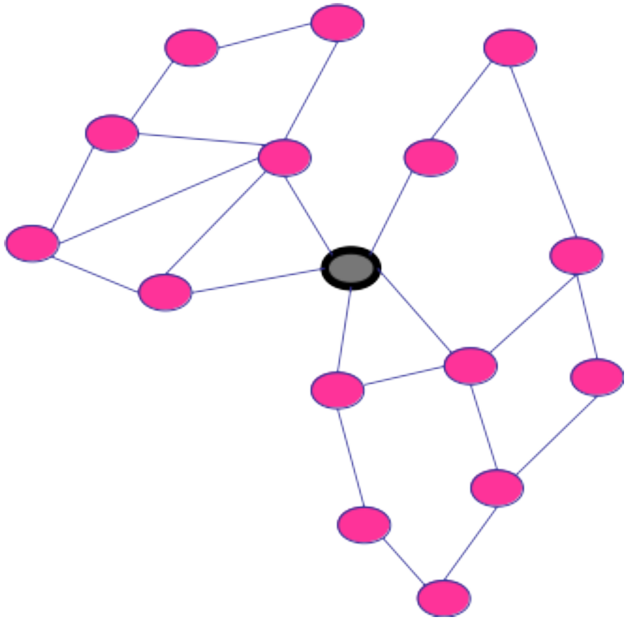
Inferred
networks



Part2: What’s next ?
=> Network Analysis
Network metrics / Network algorithms

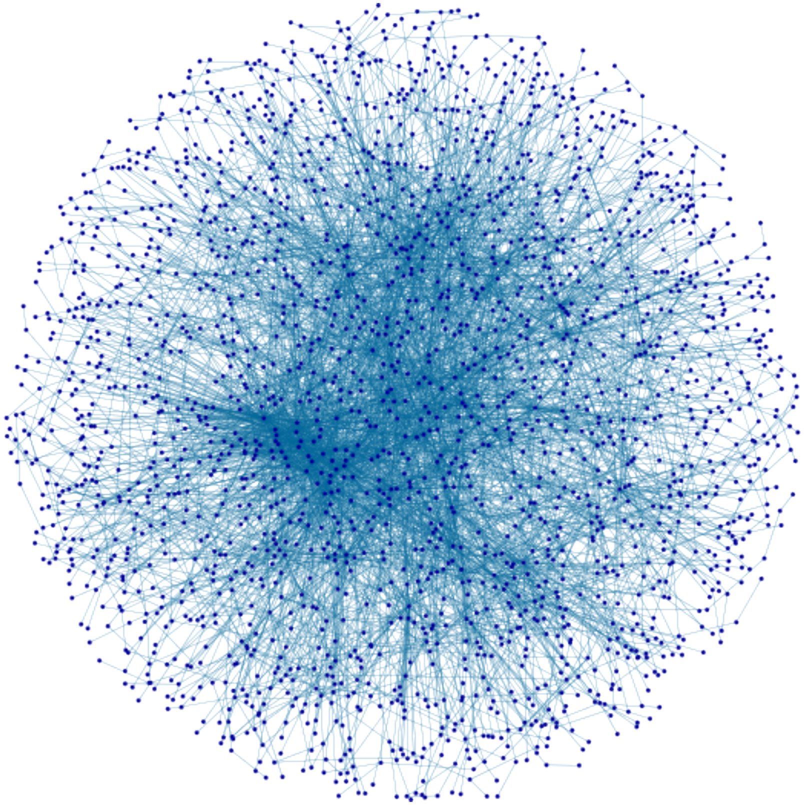
How to use large-scale biological networks ?

Local approaches



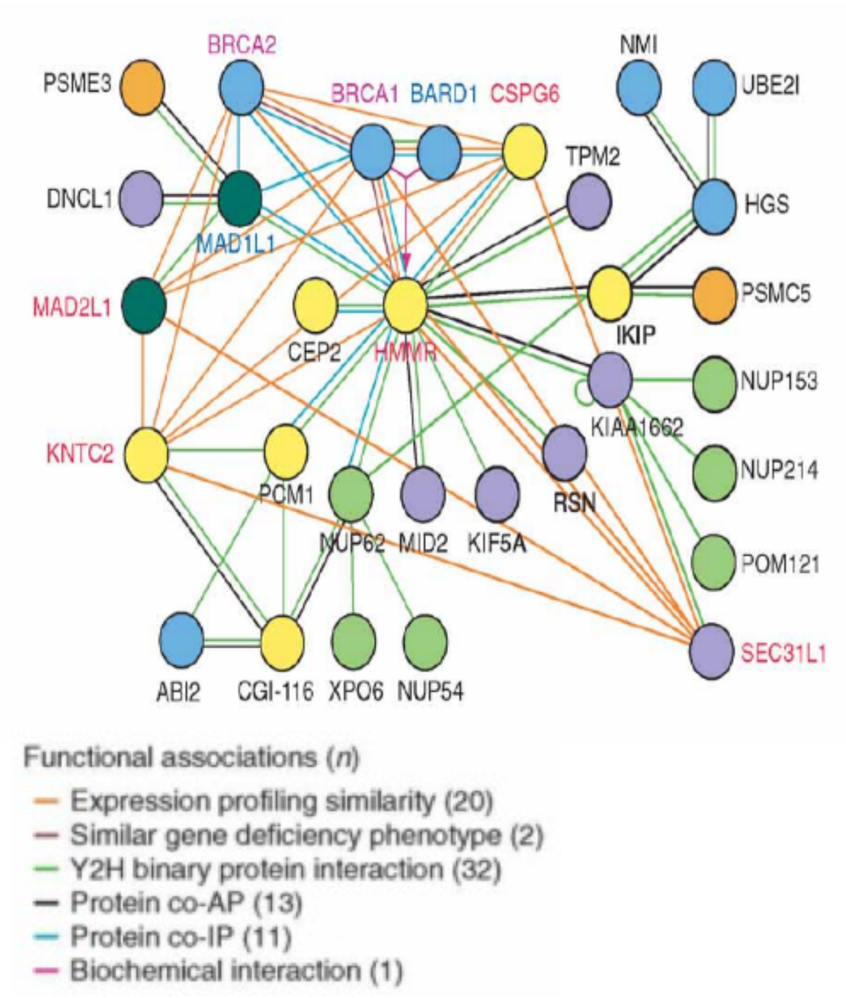
- mathematical modelling
- “guilt by association”

Global approaches



- Topological features
- Network analysis algorithms

Local approach: identification of a new gene involved in breast cancer

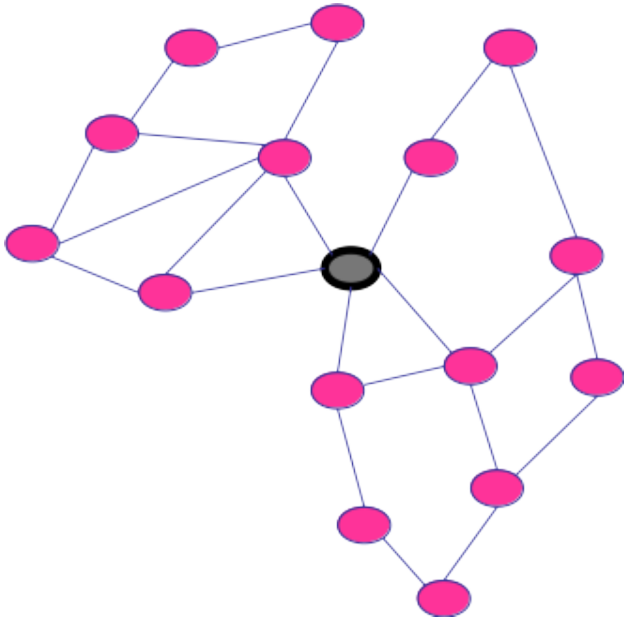


Nodes correspond to proteins, edges to interactions identified by different experimental techniques

Pujana et al. 2007

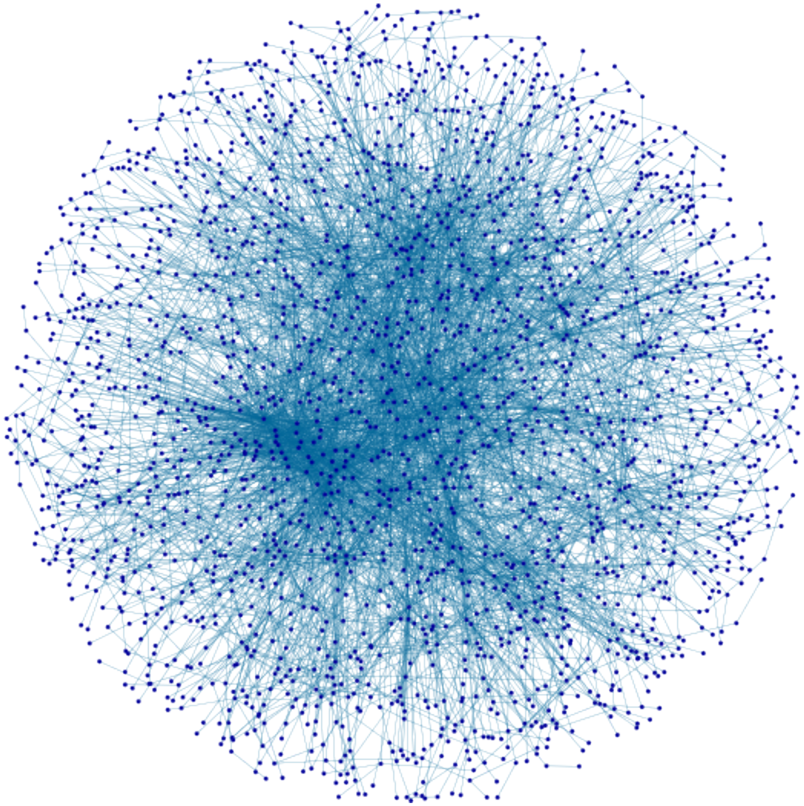
How to use large-scale biological networks ?

Local approaches



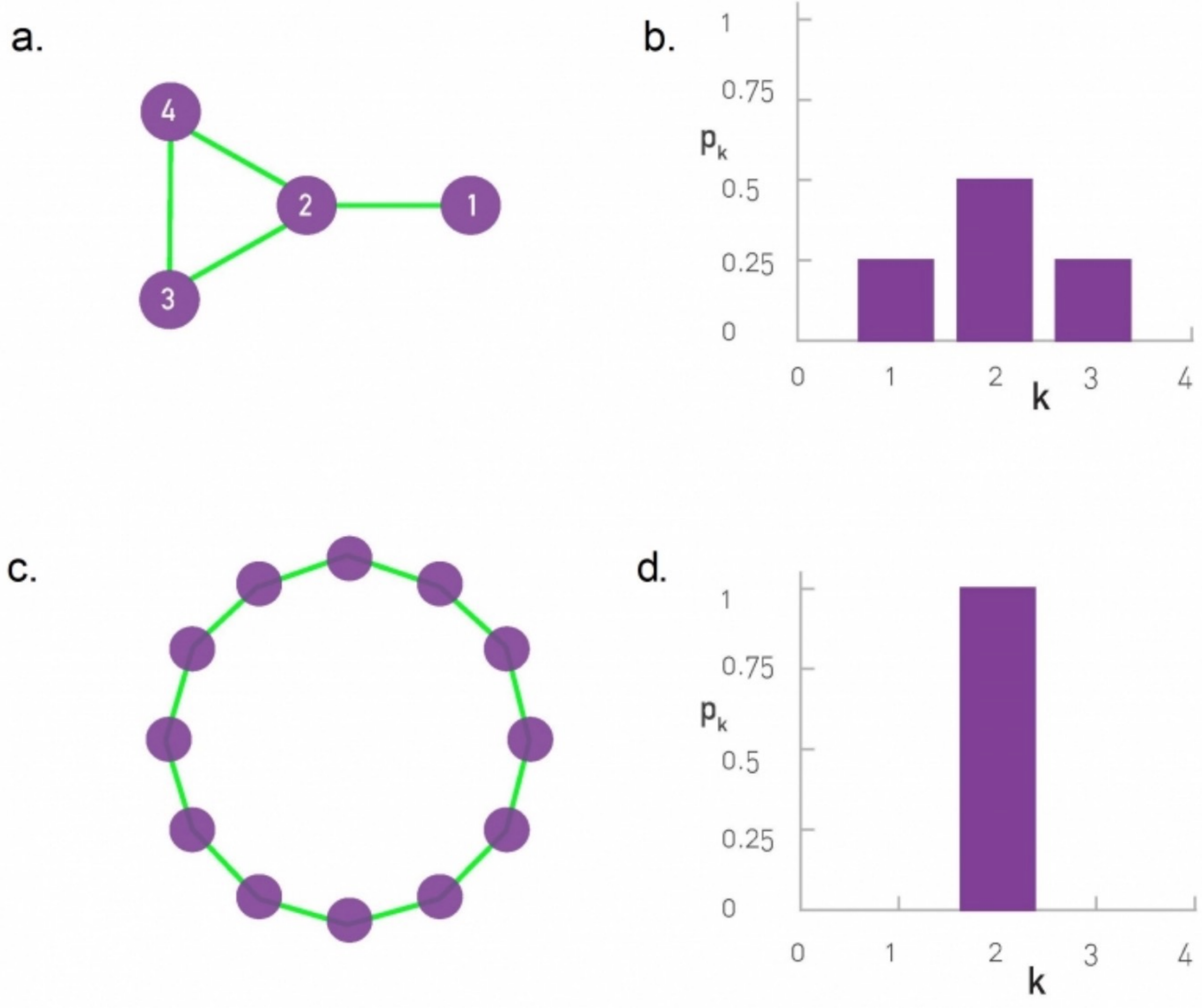
- mathematical modelling
- “guilt by association”

Global approaches

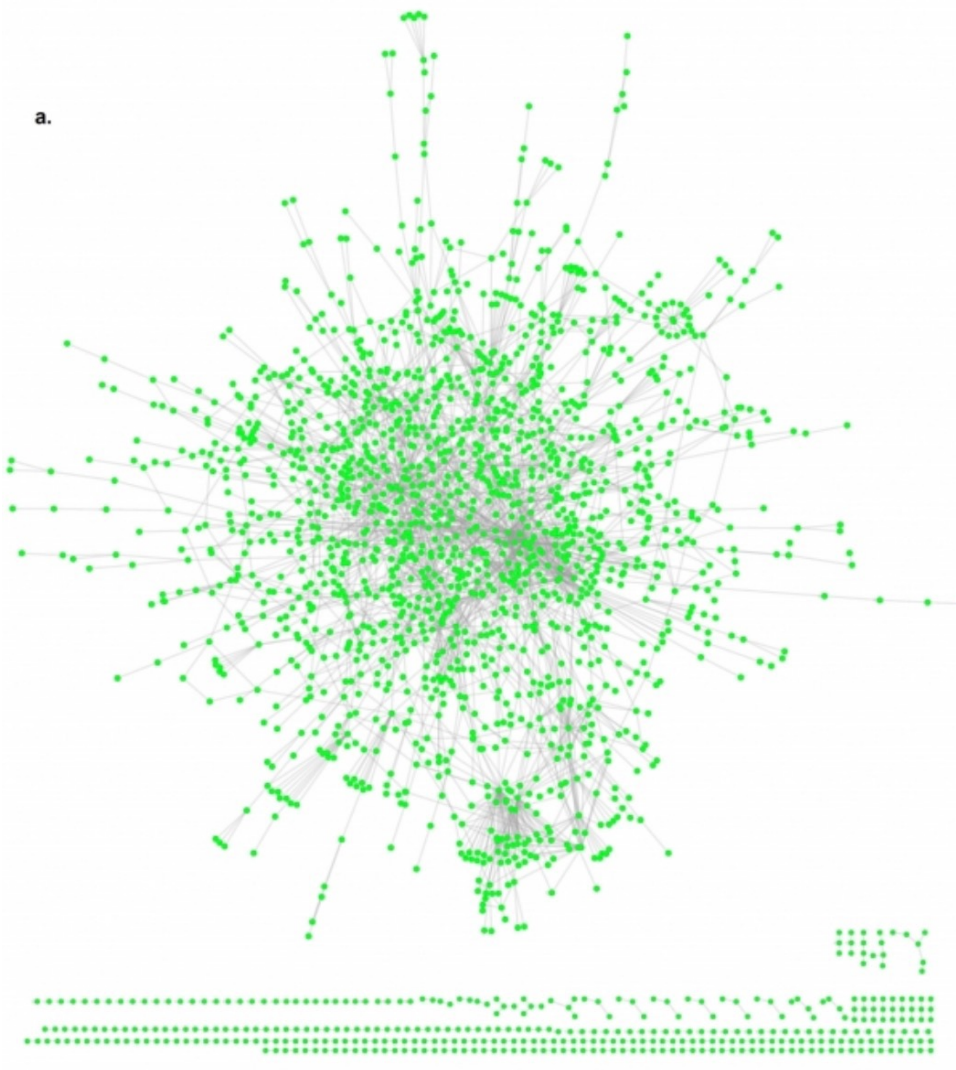


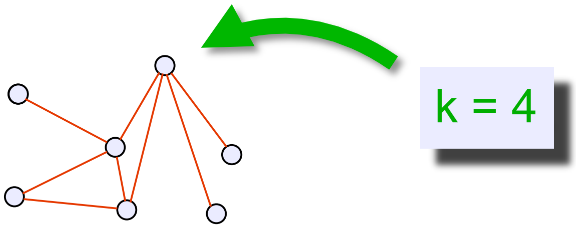
- Topological features
- Network analysis algorithms

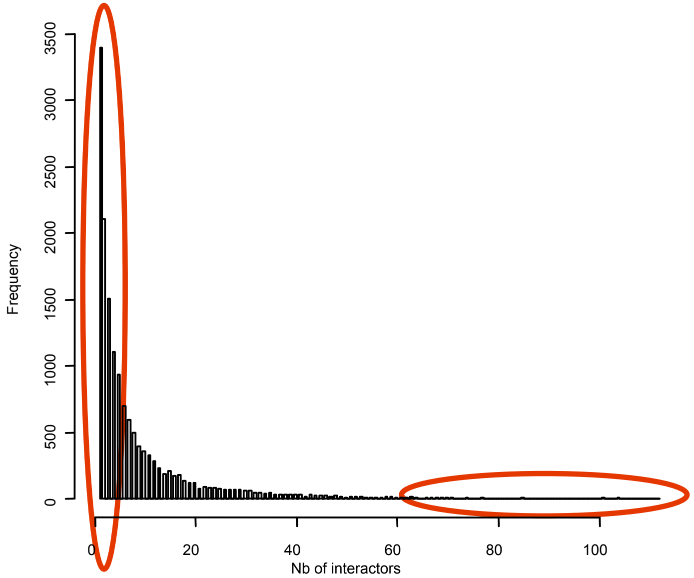
Degree distribution



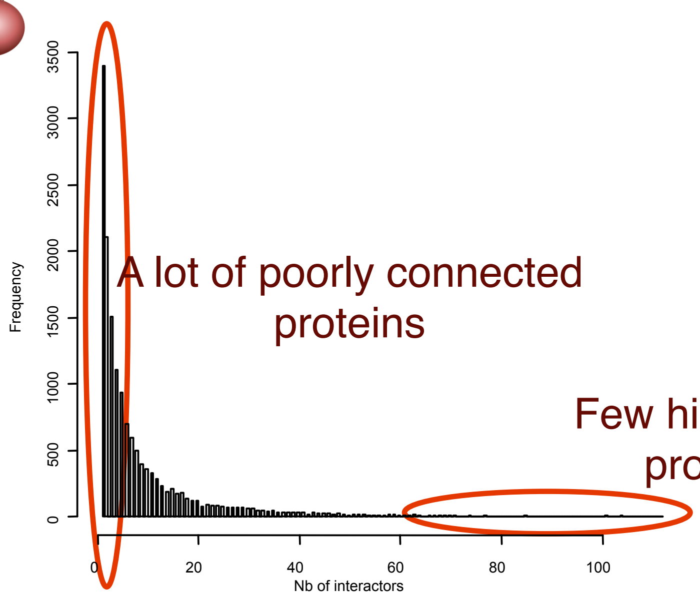
Degree distribution







Protein degree distribution : interactomes are scale-free and small-world

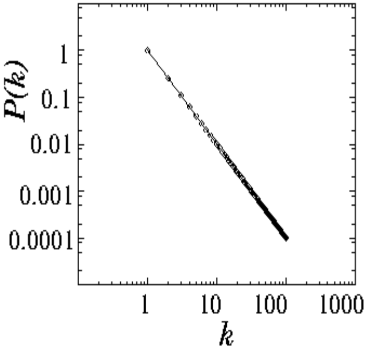


A lot of poorly connected proteins

Few highly connected proteins = hub

Scale-free

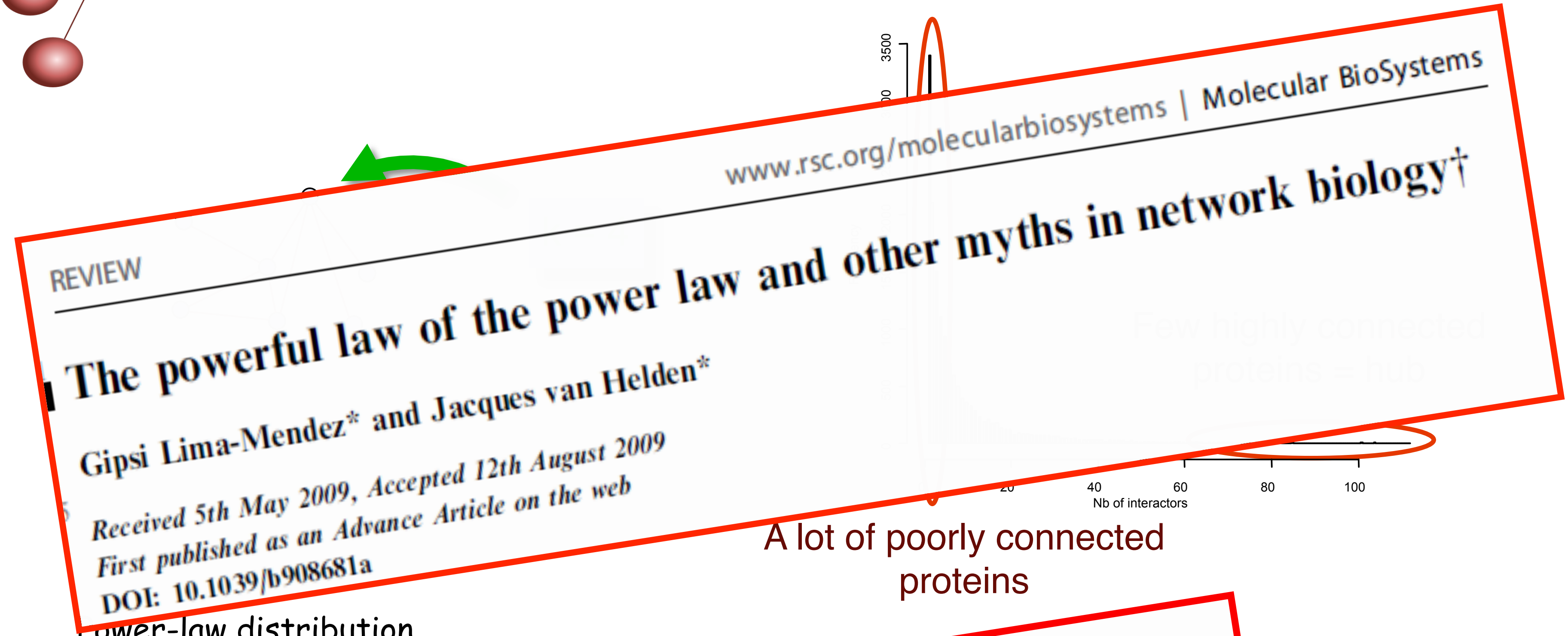
Power-law distribution



Biological interpretation?

- **Growth with preferential attachement (“rich get richer”) => create “hubs”**
- **Robust to random attack, sensitive to targeted attacks**

Protein degree distribution : interactomes are scale-free and small-world



REVIEW

www.rsc.org/molecularbiosystems | Molecular BioSystems

The powerful law of the power law and other myths in network biology†

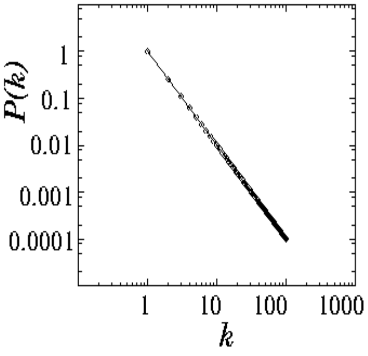
Gipsi Lima-Mendez* and Jacques van Helden*

Received 5th May 2009, Accepted 12th August 2009
First published as an Advance Article on the web
DOI: 10.1039/b908681a

Few highly connected proteins = hub

A lot of poorly connected proteins

ower-law distribution



ARTICLE

https://doi.org/10.1038/s41467-019-08746-5

OPEN

Scale-free networks are rare

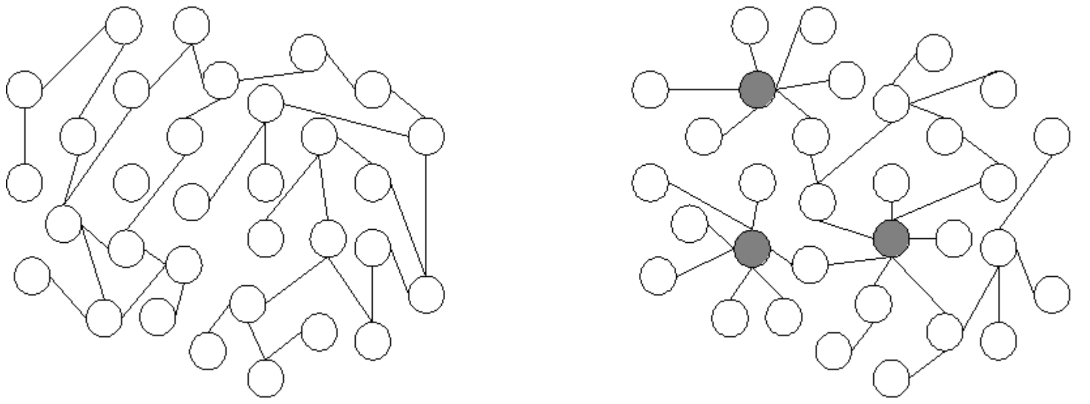
Anna D. Broido[1] & Aaron Clauset[2,3,4]

targetted

attachement (“rich get

) => create “hubs”

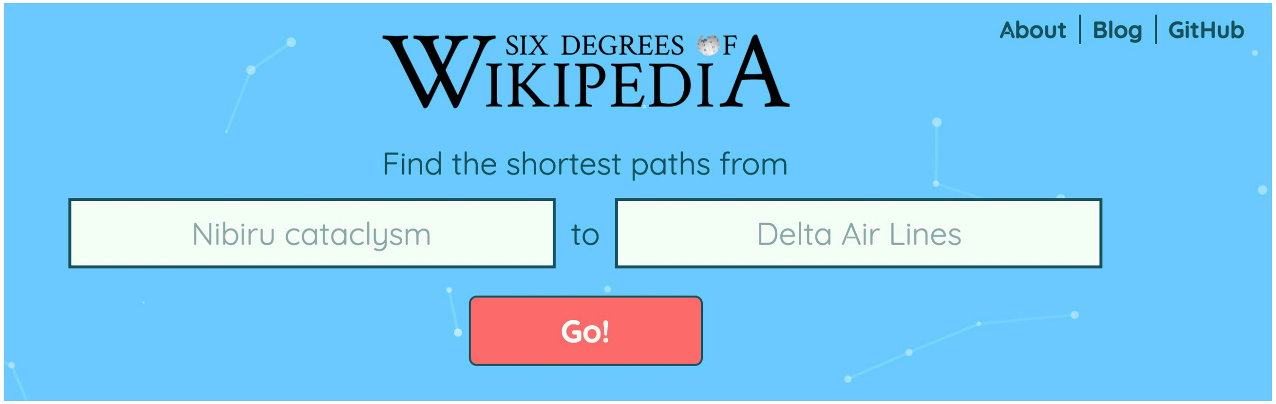
Network topological structure : Small-world property



(a) Random network

(b) Scale-free network

- Milgram, 6 degrees of separation



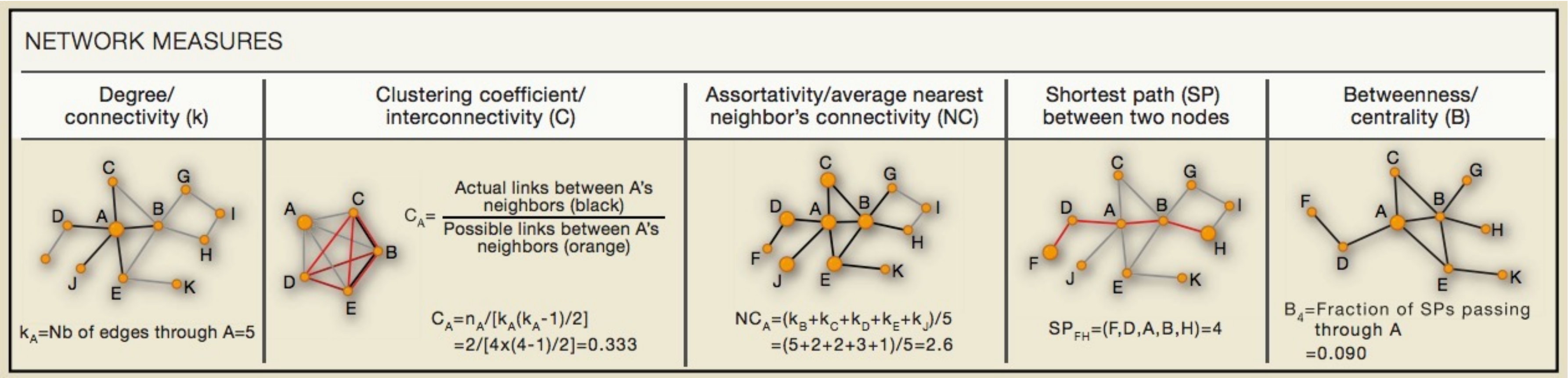
SIX DEGREES OF WIKIPEDIA

Find the shortest paths from

Nibiru cataclysm to Delta Air Lines

Metrics on graphs

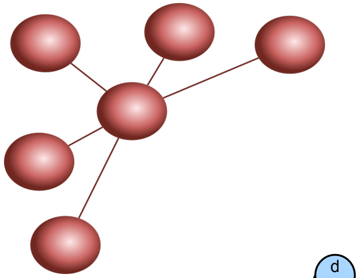
- N nodes, V edges
- Network size
- Adjacency matrix
- Degree, degree distribution
- Path, shortest path, distances
- Connectivity, clustering coefficient
- Betweenness
- Motifs



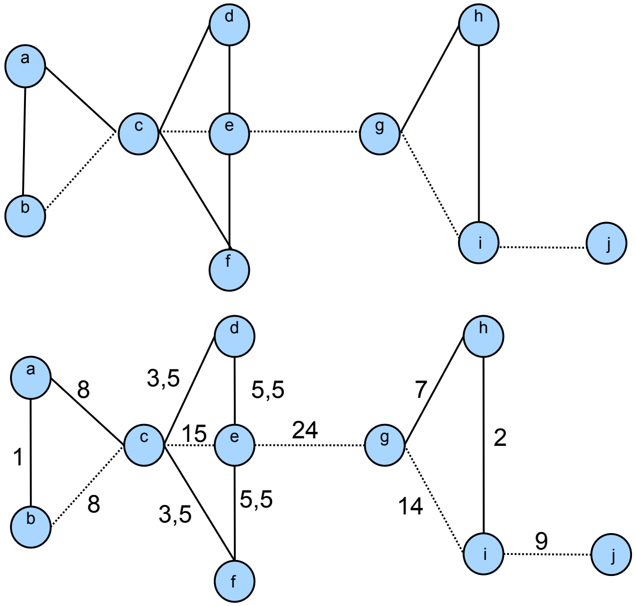
NETWORK MEASURES

Degree/ connectivity (k)	Clustering coefficient/ interconnectivity (C)	Assortativity/average nearest neighbor's connectivity (NC)	Shortest path (SP) between two nodes	Betweenness/ centrality (B)
k_A=Nb of edges through A=5	$C_A=\frac{\text{Actual links between A's neighbors (black)}}{\text{Possible links between A's neighbors (orange)}}$ $C_A=n_A/[k_A(k_A-1)/2]$ $=2/[4x(4-1)/2]=0.333$	$NC_A=(k_B+k_C+k_D+k_E+k_J)/5$ $=(5+2+2+3+1)/5=2.6$	SP_{FH}=(F,D,A,B,H)=4	B_4=Fraction of SPs passing through A =0.090

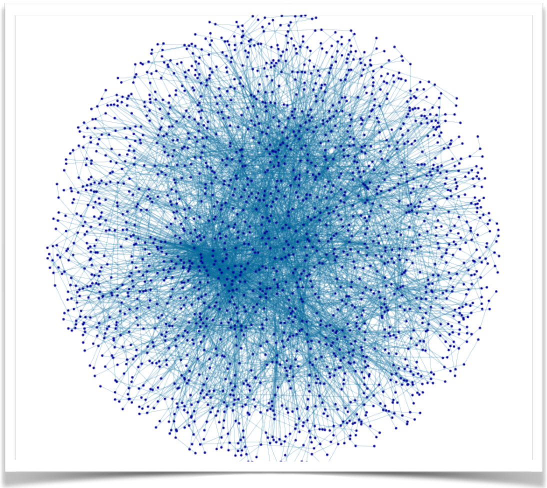
Gavin, Cell



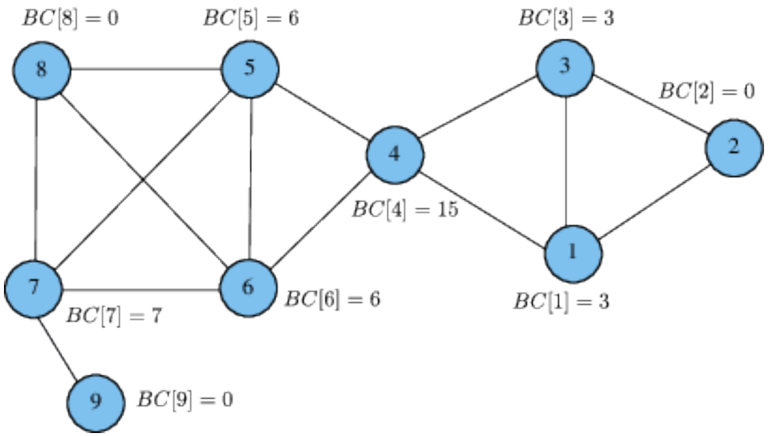
“Edge Betweenness”



Number of shortest paths running through an edge
=
“bottleneck”



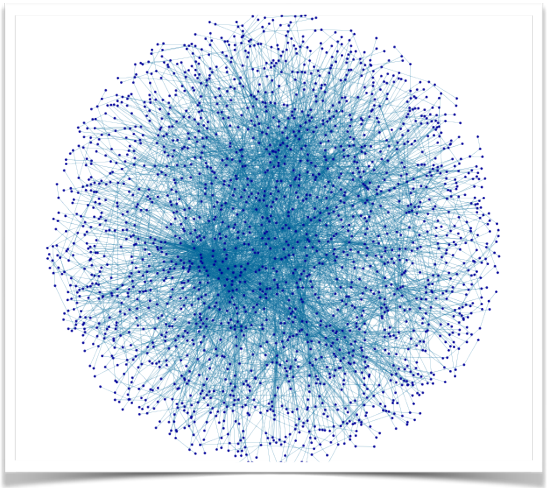
“Node Betweenness”



Number of shortest paths running through a node

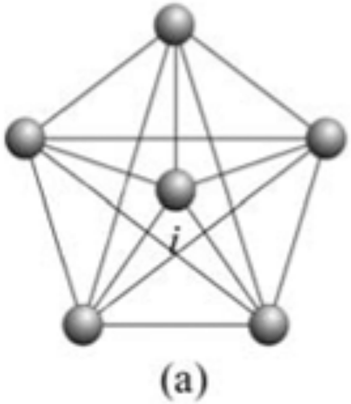
=

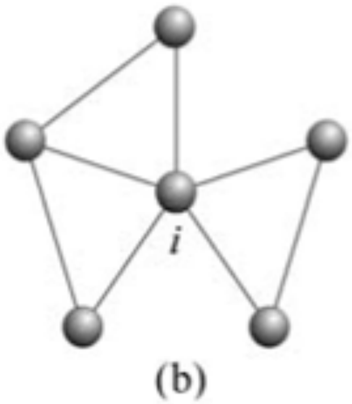
“bottleneck”

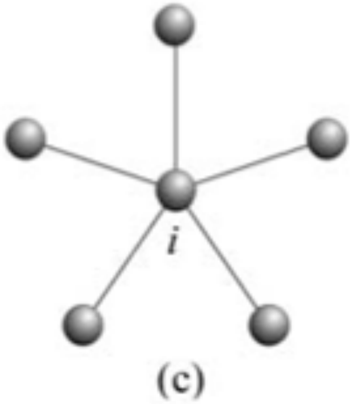


Biological interpretation ?
Correlation with gene essentiality, gene involvement in diseases, importance in flux transmission …

Clustering coefficient / modularity





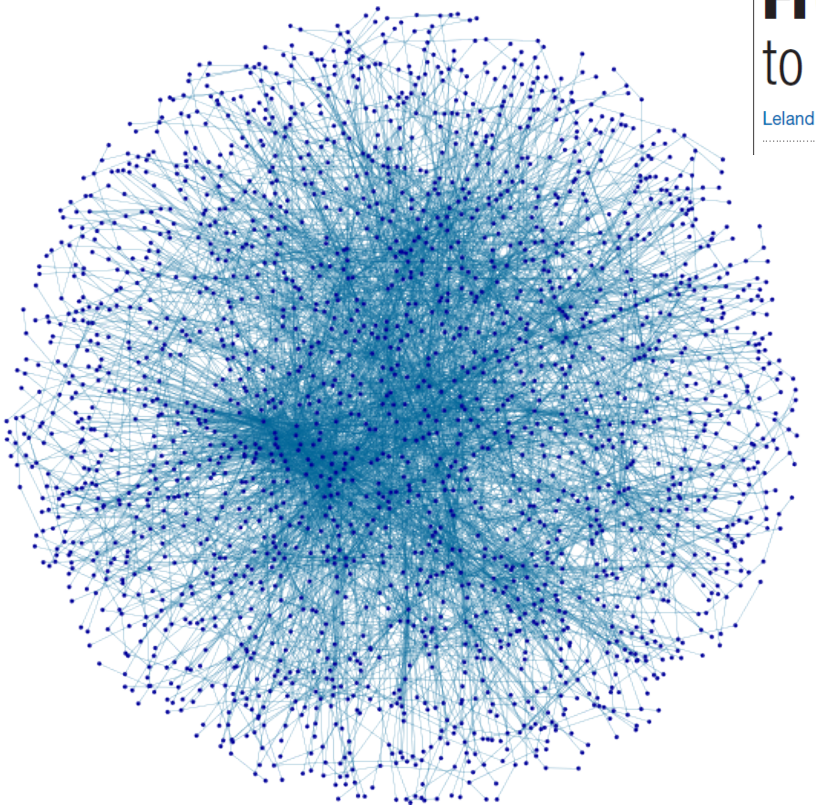


Actual links between neighbours / Possible links between neighbours

Algorithms for Network Analysis

- Clustering
- Exploration with Random Walk with Restart
- Integration of expression data to find active modules

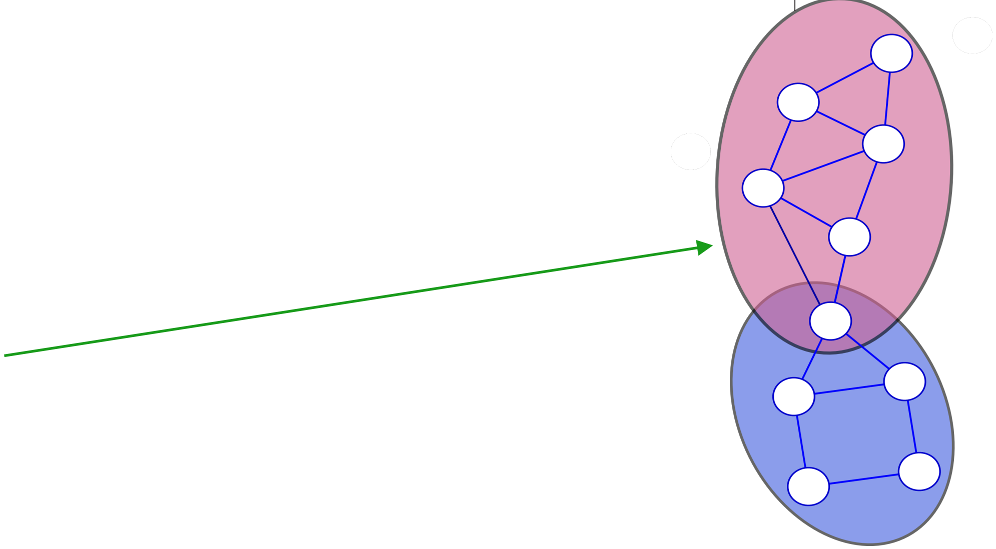
Network Analysis - Clustering



impacts

From molecular
to modular cell biology

Leland H. Hartwell, John J. Hopfield, Stanislas Leibler and Andrew W. Murray

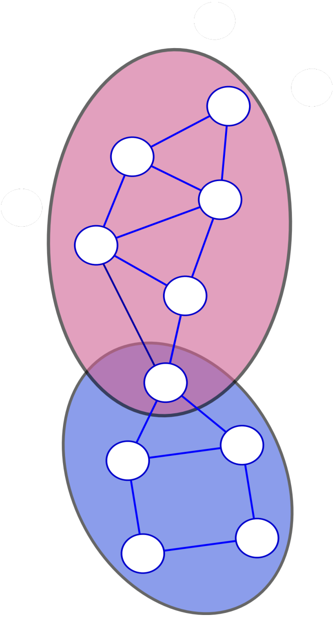


Functional module / community / cluster / class : discrete function

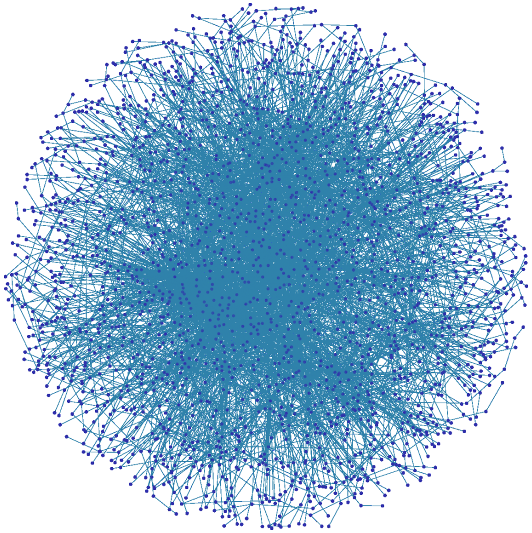
Modules can be isolated or connected

Groups of proteins involved in a common cellular function

Binary interaction



Functional modules

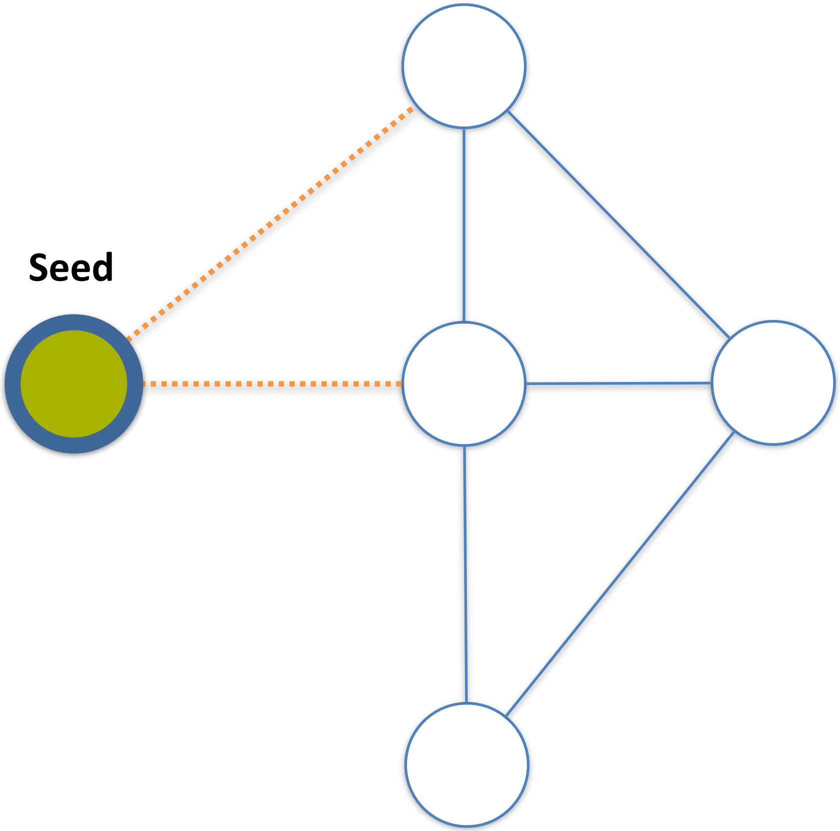


Interaction Networks

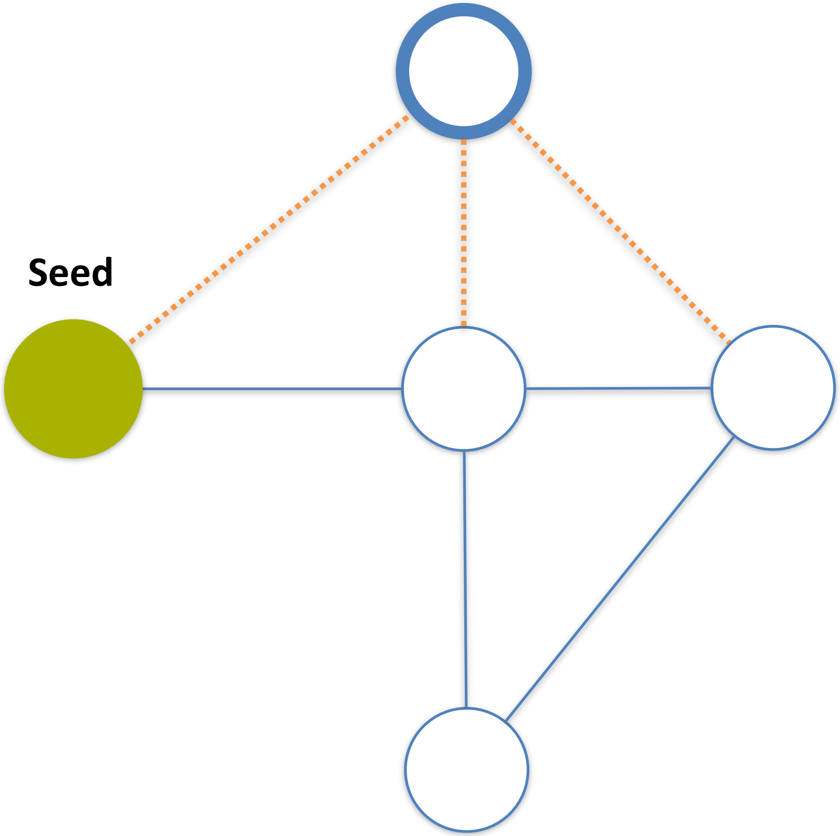
Inspired from P. Aloy, ECCB 2014

Network Exploration: Random Walk with Restart

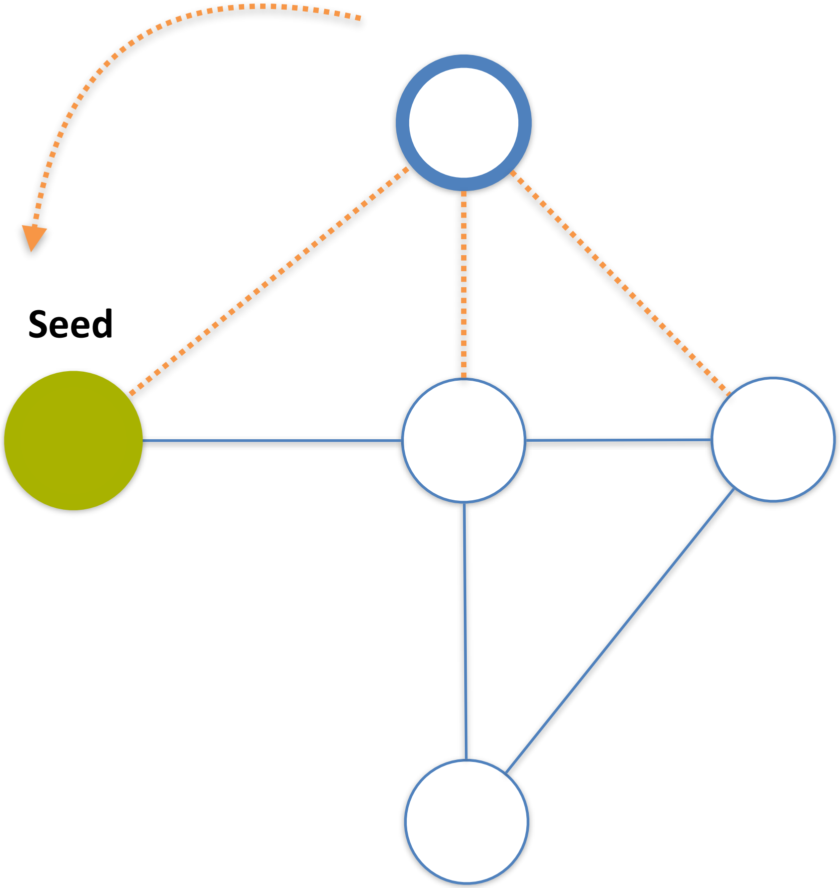
The Random Walk with Restart algorithm



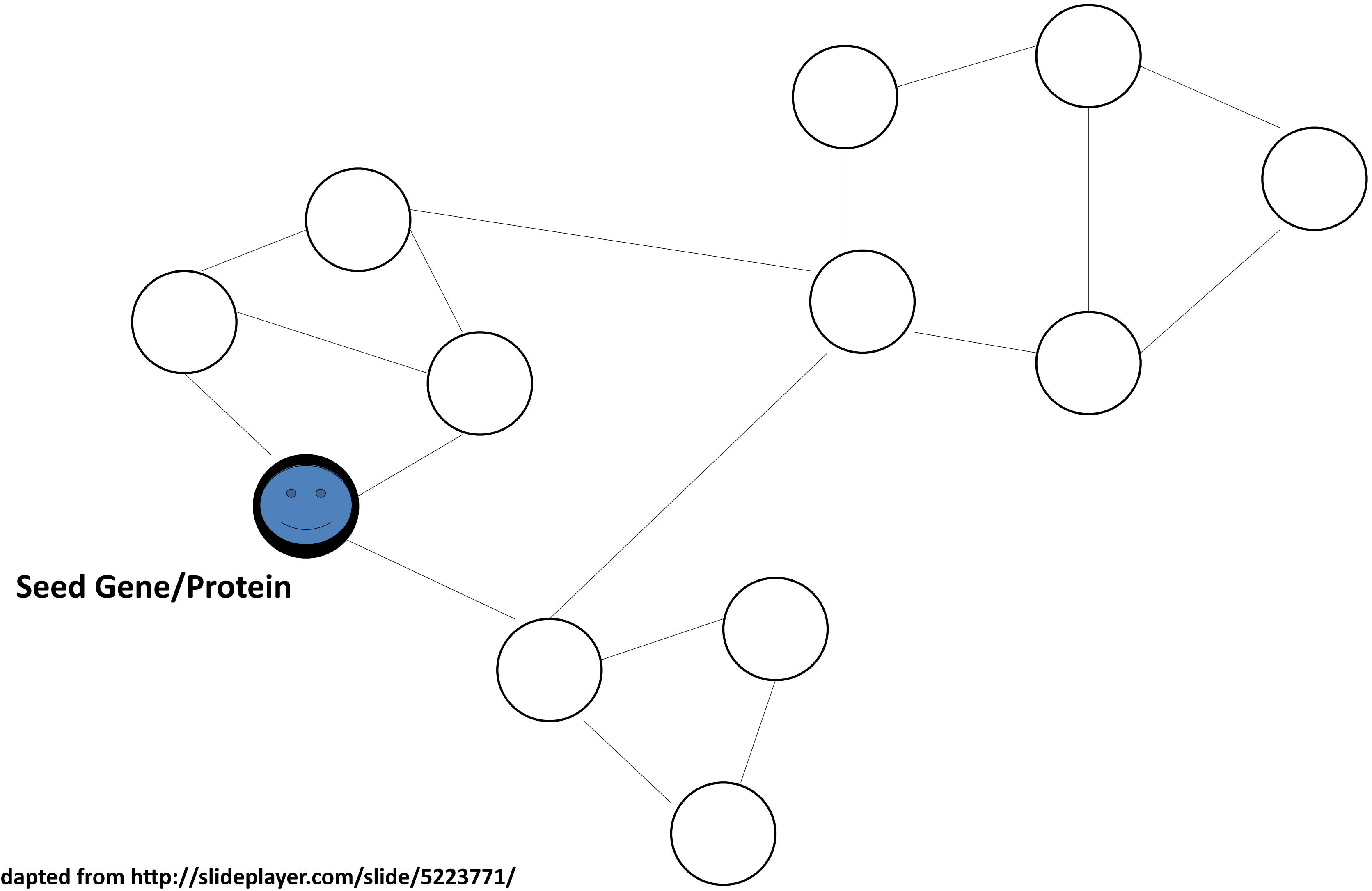
The Random Walk with Restart algorithm



The Random Walk with Restart algorithm



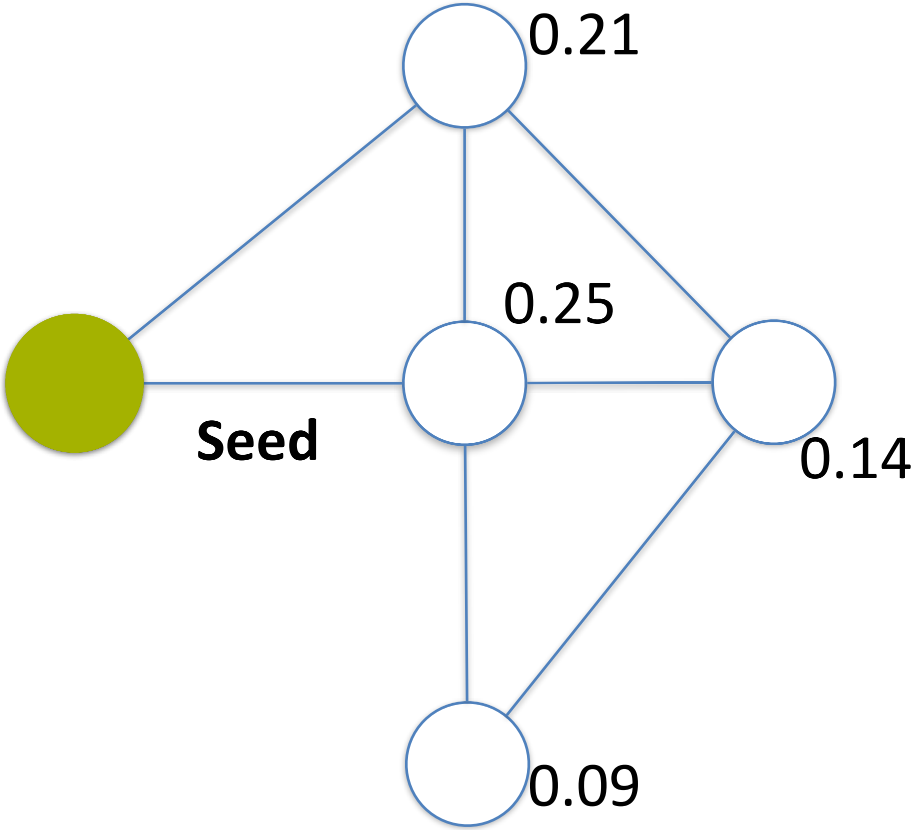
The Random Walk with Restart algorithm



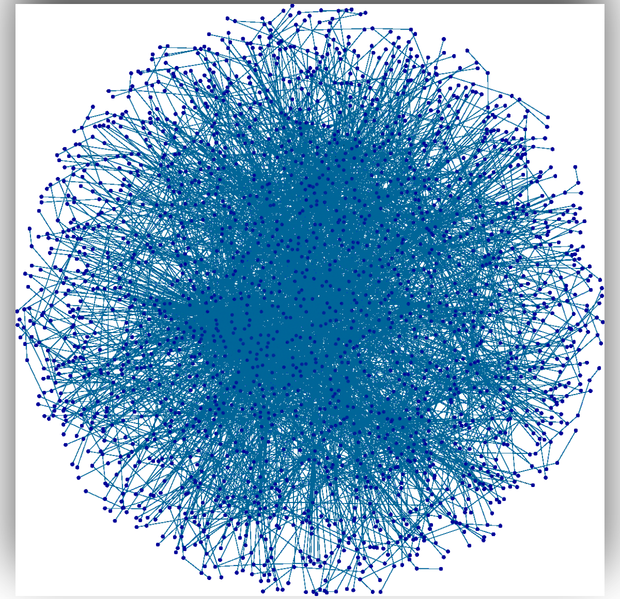
Seed Gene/Protein

Adapted from http://slideplayer.com/slide/5223771/

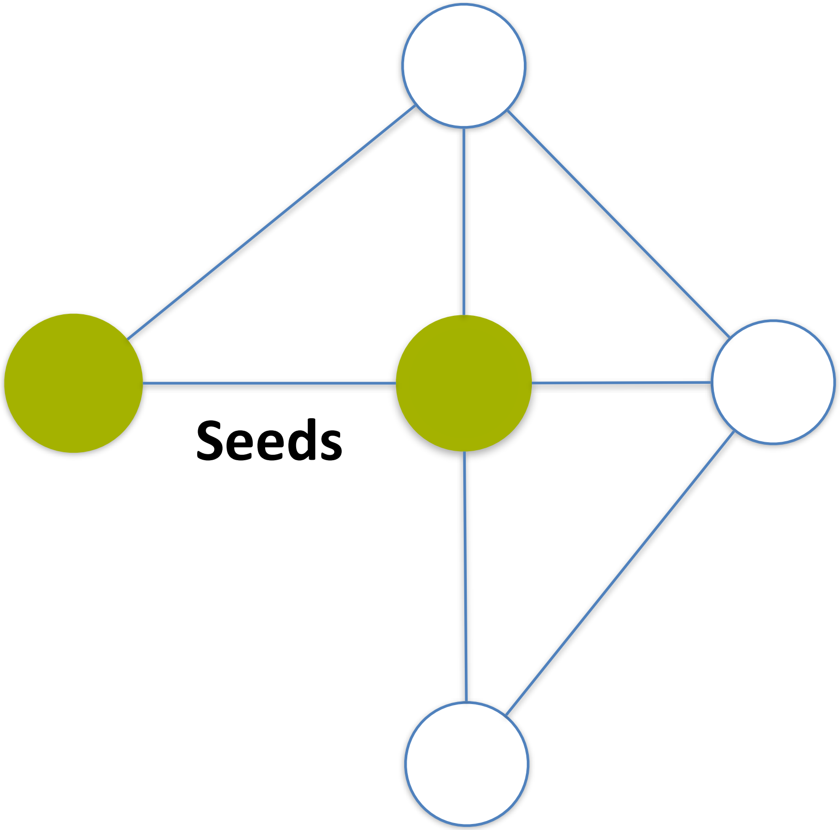
The Random Walk with Restart algorithm



- Local exploration
- Proximity/pertinence score *wrt* the seed

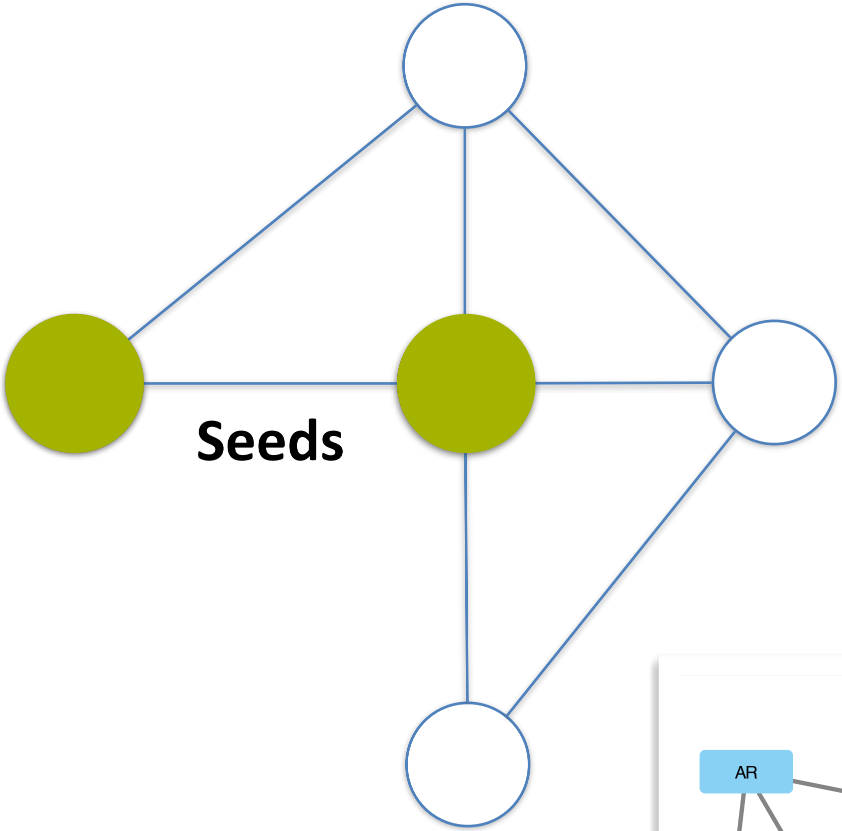


The Random Walk with Restart algorithm

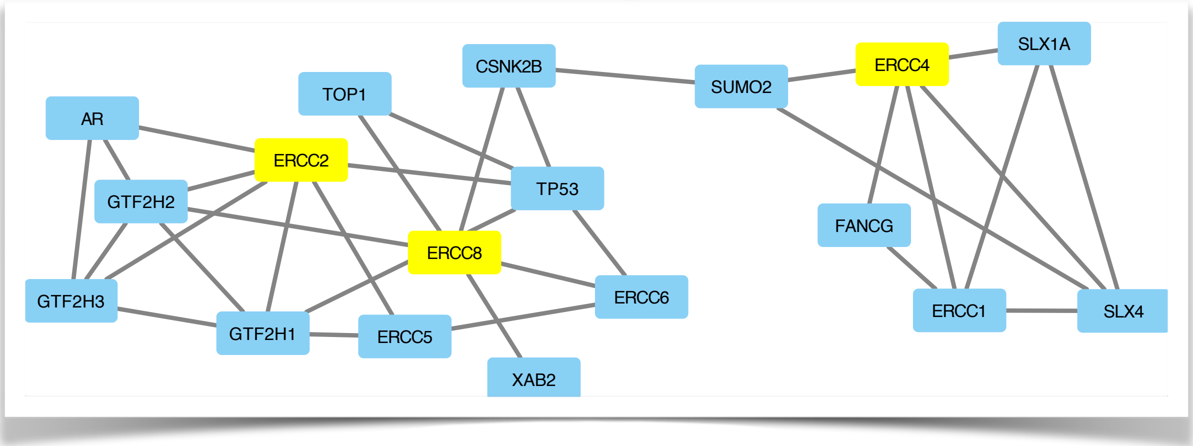


- Local exploration
- Proximity/pertinence score *wrt* the seed(s)

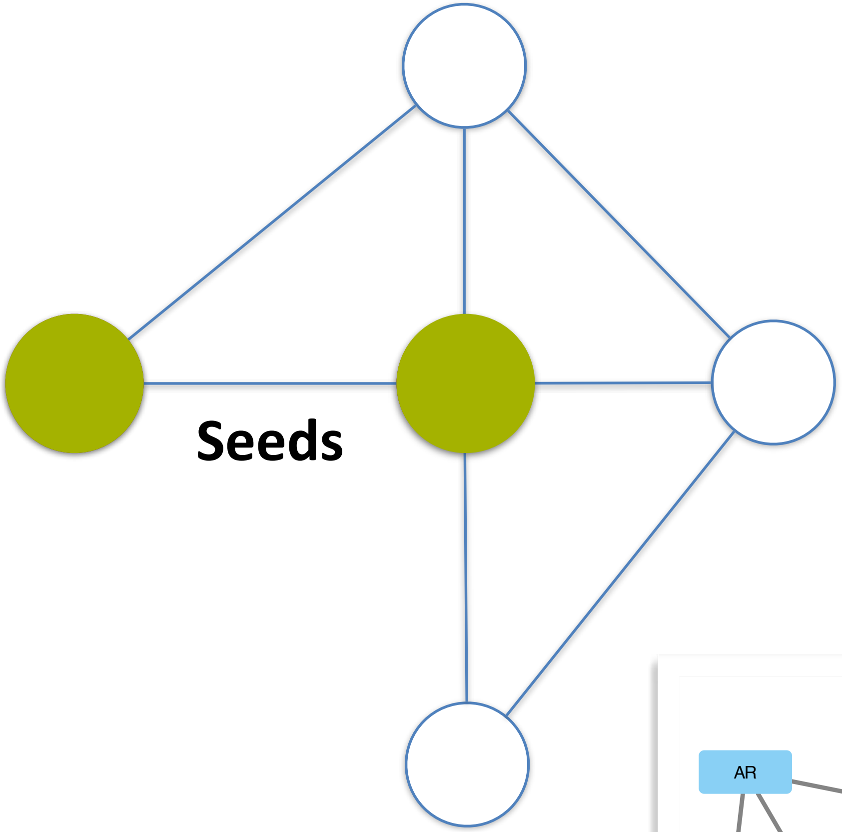
The Random Walk with Restart algorithm



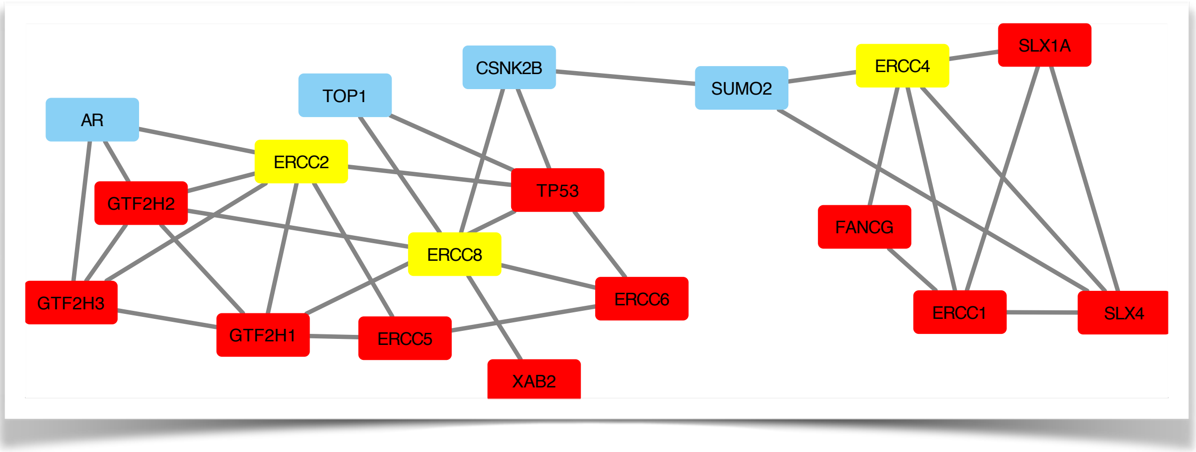
- Local exploration
- Proximity/pertinence score *wrt* the seed(s)
- Guilt-by association



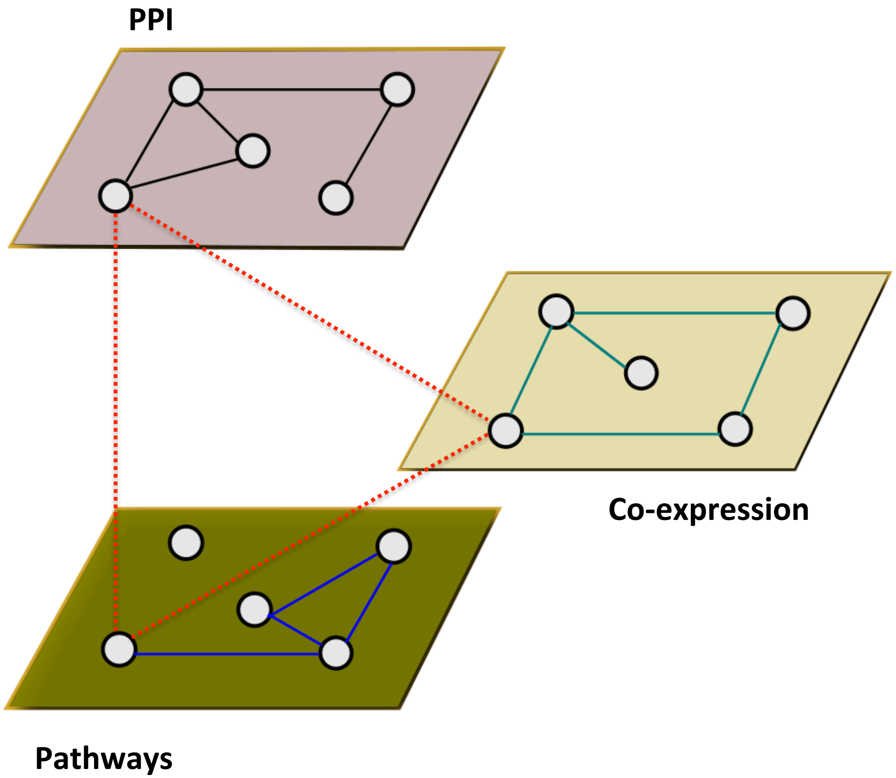
The Random Walk with Restart algorithm



- Local exploration
- Proximity/pertinence score *wrt* the seed(s)
- Guilt-by association

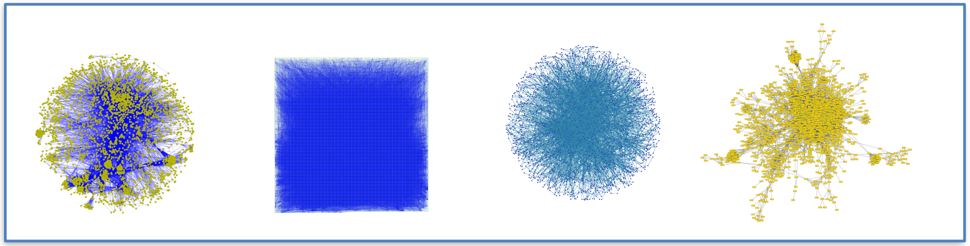


Random Walk with Restart on Multiplex Networks (RWR-M)

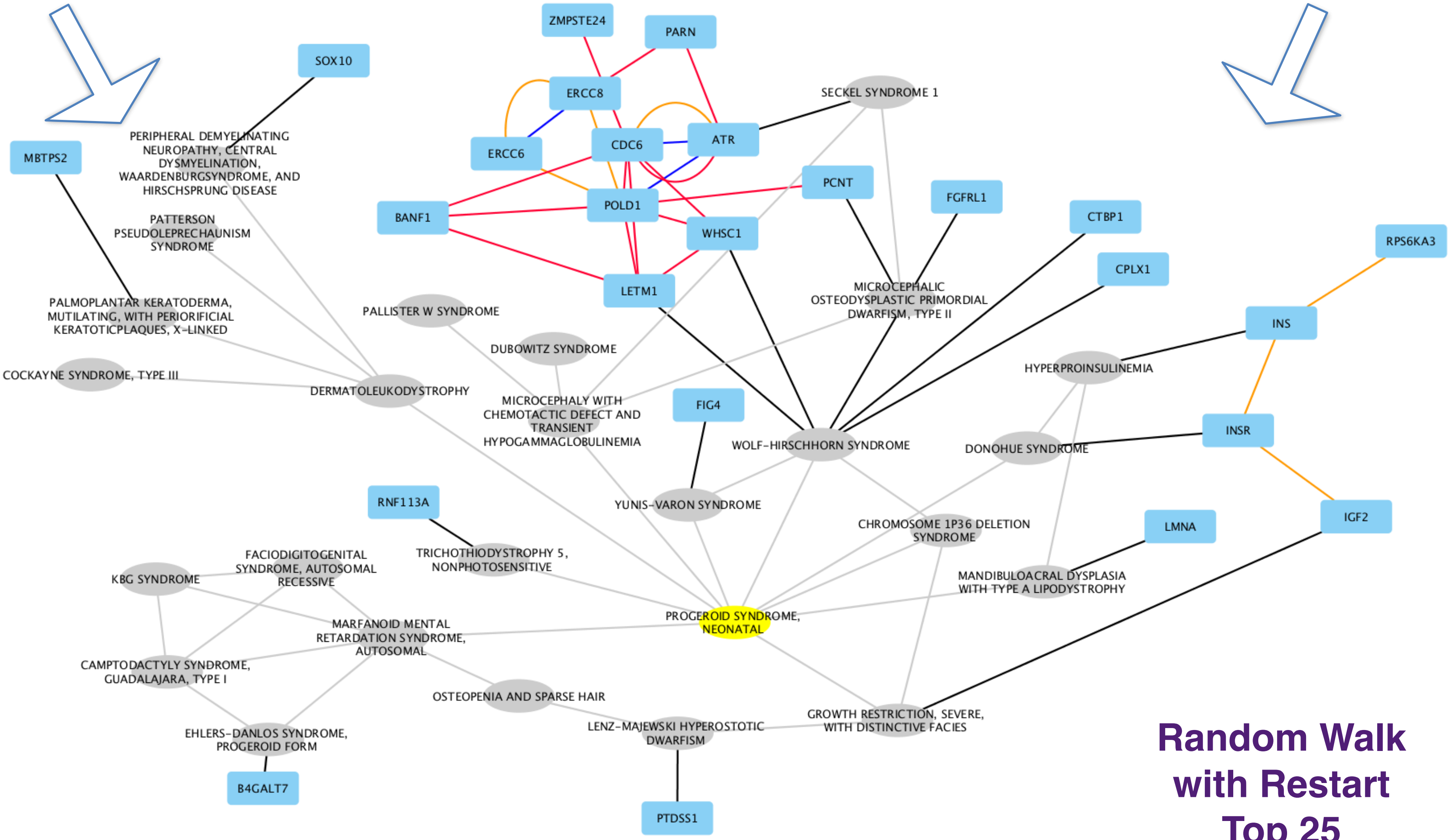


- Walk one layer
- Jump across layers
- Gene/Protein Seed(s)

Alberto Valdeolivas



Disease-disease network



Random Walk with Restart Top 25

Integrating expression data and networks: finding active modules

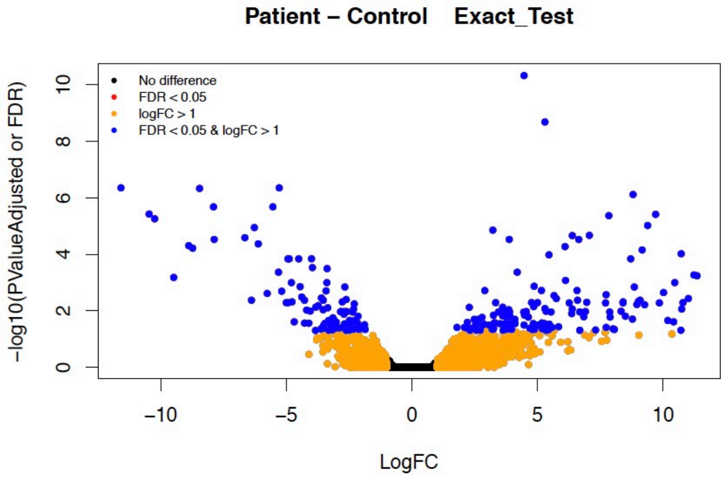
Interactomes are devoid of spatio-temporal information

RNA-seq transcriptomics analyses

RNA-seq transcriptomics data

↓

Identify significant DE mRNAs

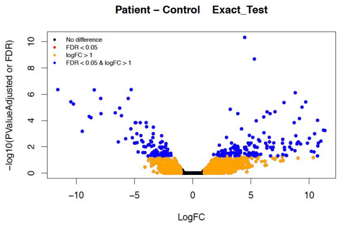


→

Functional enrichments

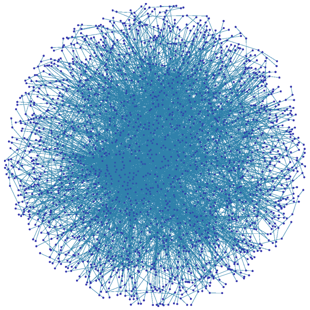
Active module identification

RNA-seq transcriptomics data



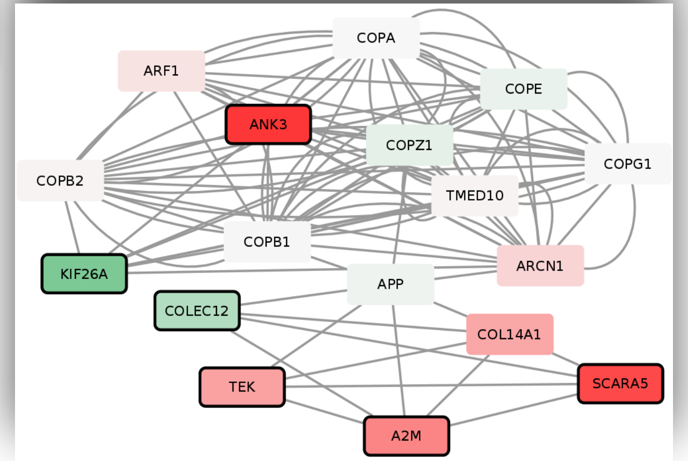
+

Biological Network



↓

Find “active” subnetworks



Algorithms: Greedy searches (PinnacleZ), Simulated Annealing (jActiveModules), Genetic Algorithms (COSINE) (Ideker et al. 2002, Chuang et al. 2007, Ma et al. 2011, Ozisik et al. 2017…)

Active module identification

- Few methods consider the density of interactions
- Methods are using only one (usually protein-protein) interaction networks

=> We propose a **Multi-Objective Genetic Algorithm** to identify active modules from Multiplex Networks

Elva Novoa

2 objectives to maximize

1

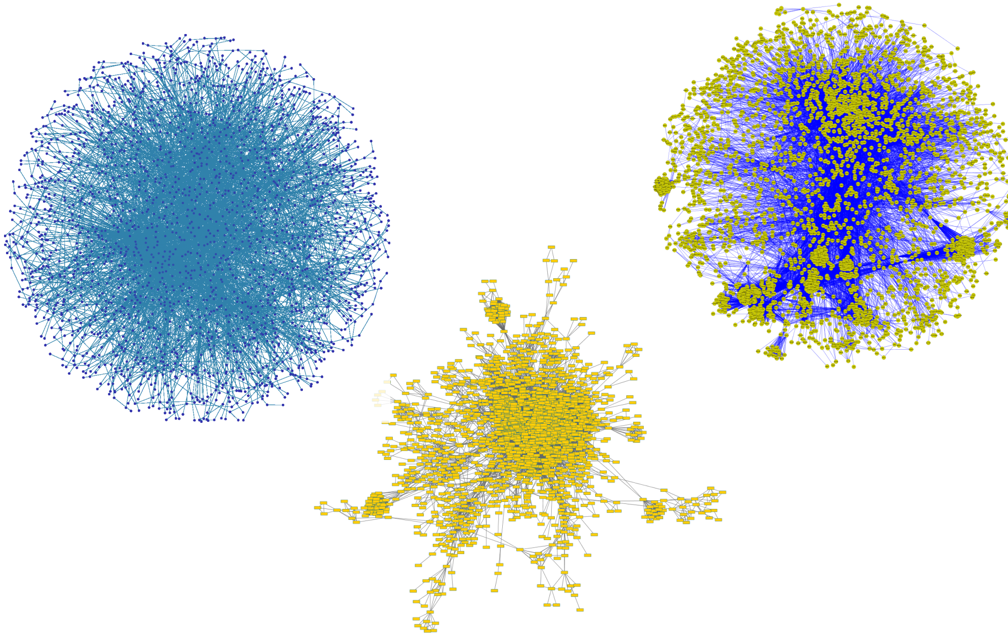
Average nodes score $$\overline{NodesScore} = \frac{1}{n}\sum_{i=1}^{n}(Score_i^{norm})$$

$$Score_i^{norm} = \frac{Score_i - min(Score)}{max(Score) - min(Score)} \qquad Score_i = \Phi^{-1}(1 - p_i)$$

2

Density

$$\boxed{D_{norm} = \sum_{l=1}^{L} \frac{d_s}{d_l}}$$



2 objectives to maximize

1

Average nodes score

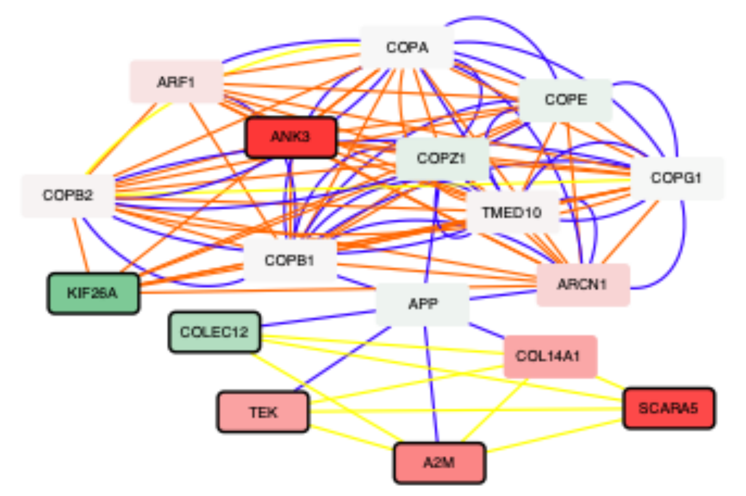
$$\overline{NodesScore} = \frac{1}{n}\sum_{i=1}^{n}(Score_i^{norm})$$

$$Score_i^{norm} = \frac{Score_i - min(Score)}{max(Score) - min(Score)} \qquad Score_i = \Phi^{-1}(1 - p_i)$$

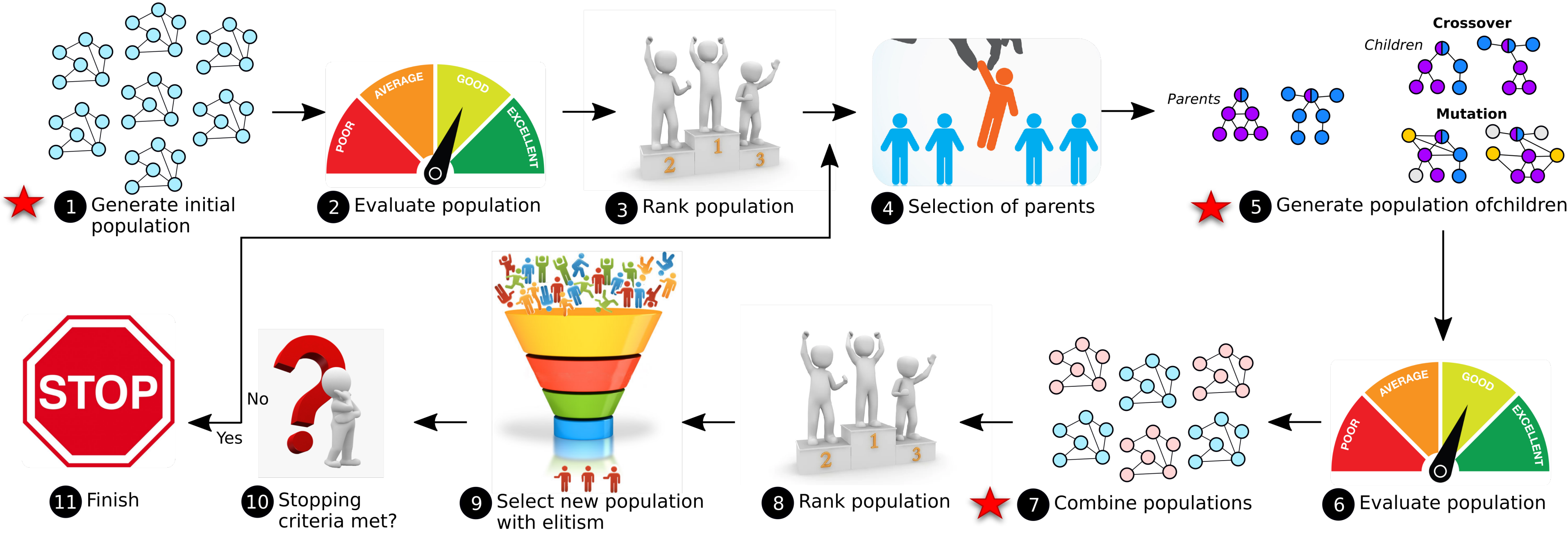
2

Density

$$D_{norm} = \sum_{l=1}^{L} \frac{d_s}{d_l}$$



Multi-Objectives Genetic Algorithm



1 Generate initial population

2 Evaluate population

3 Rank population

4 Selection of parents

5 Generate population ofchildren

6 Evaluate population

7 Combine populations

8 Rank population

9 Select new population with elitism

10 Stopping criteria met?

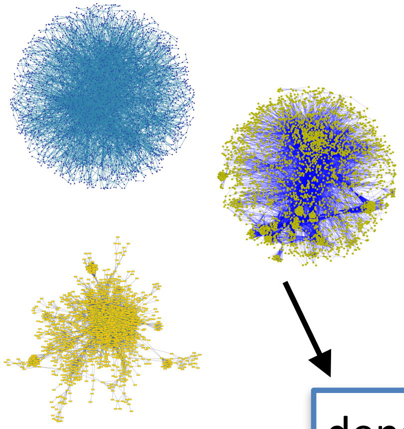
11 Finish

Based on NSGA-II (Deb et al., 2002)

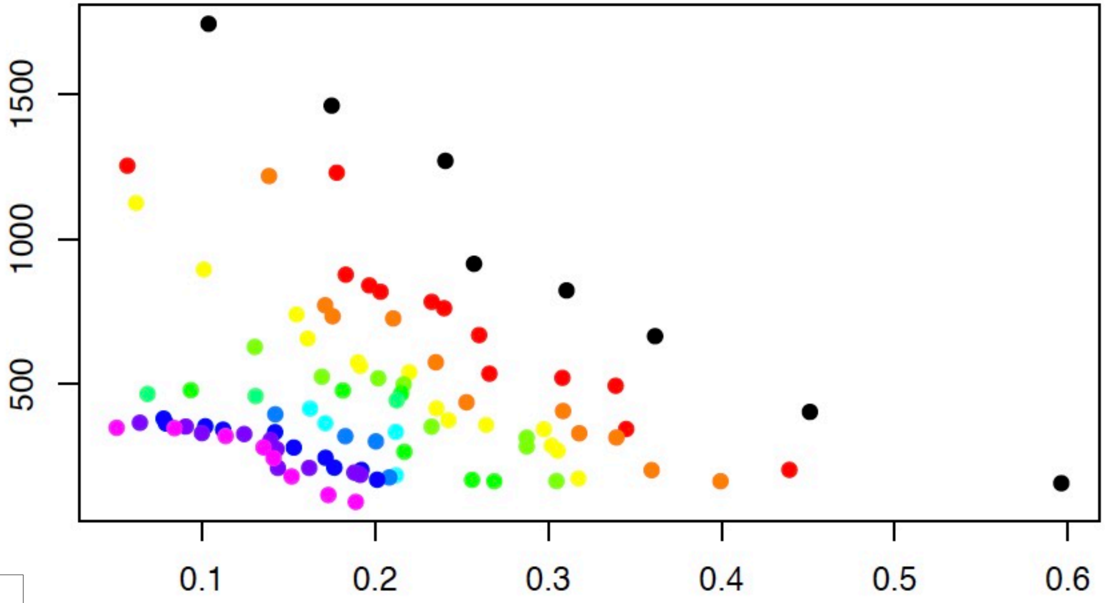
MOGAMUN

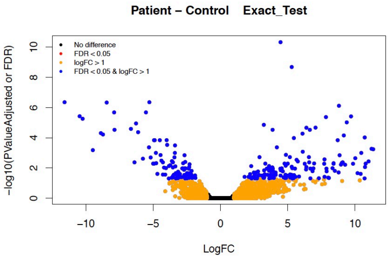
Multi-objective Genetic Algorithm to find Active Modules in Multiplex Biological Networks

Maximize 2 objectives



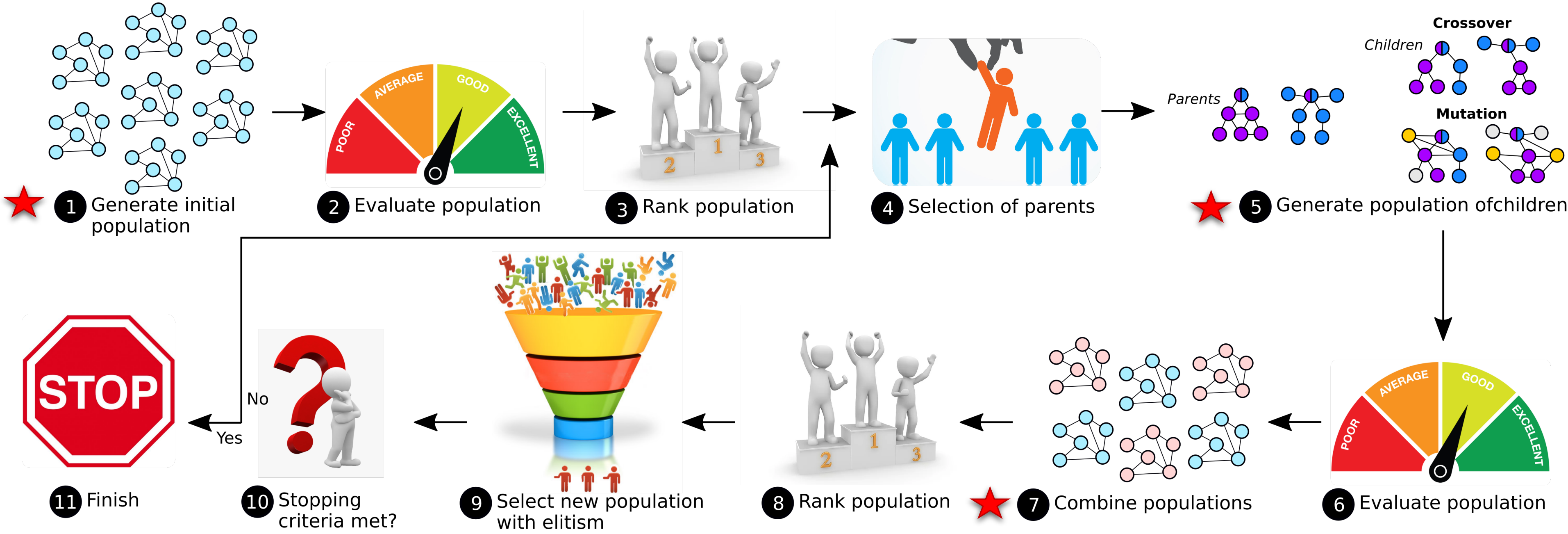
density of interactions from the multiplex network





Differential expression (average node-scores)

Multi-Objectives Genetic Algorithm



1 Generate initial population

2 Evaluate population

3 Rank population

4 Selection of parents

5 Generate population ofchildren

6 Evaluate population

7 Combine populations

8 Rank population

9 Select new population with elitism

10 Stopping criteria met?

11 Finish

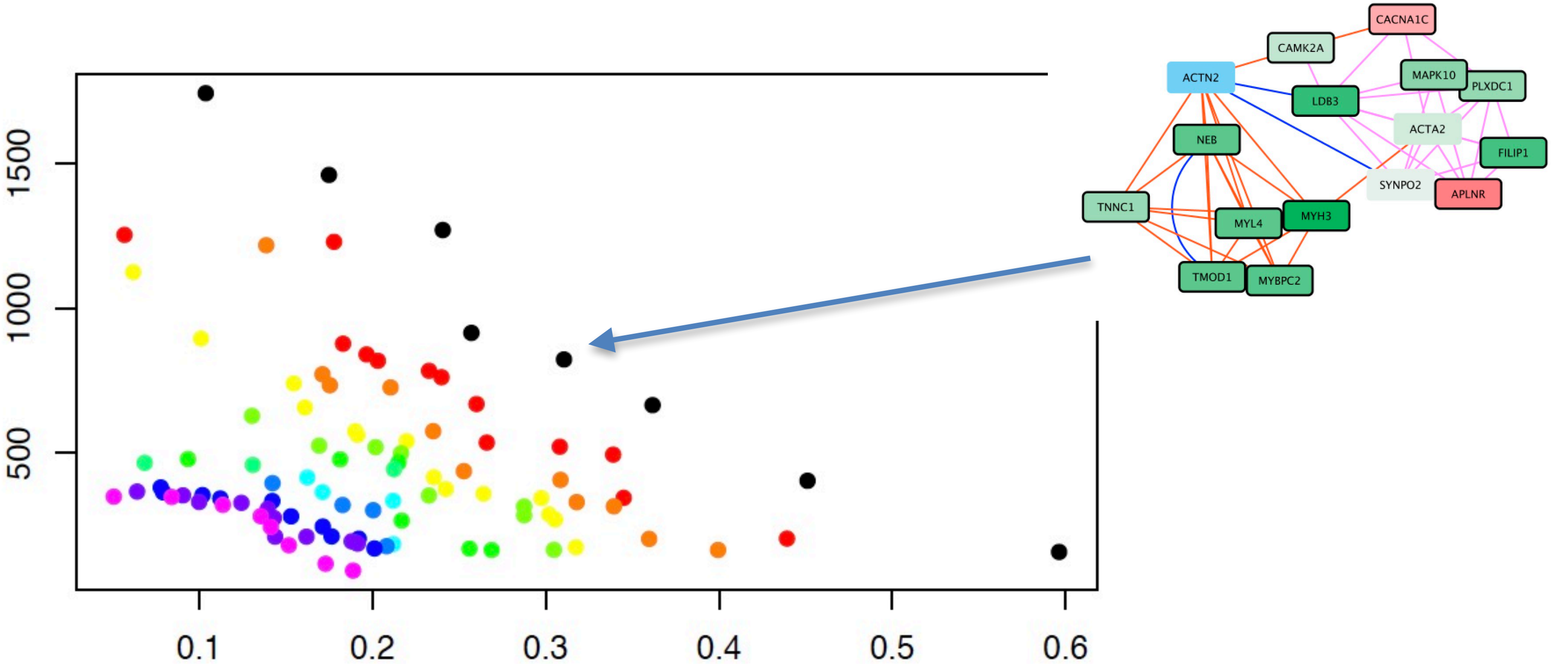
Based on NSGA-II (Deb et al., 2002)

MOGAMUN

Multi-objective Genetic Algorithm to find Active Modules in Multiplex Biological Networks

MOGAMUN

density of interactions from the multiplex network



Differential expression (average node-scores)

Novoa et al. BioRxiv, 2020
Bioconductor
https://github.com/elvanov/MOGAMUN