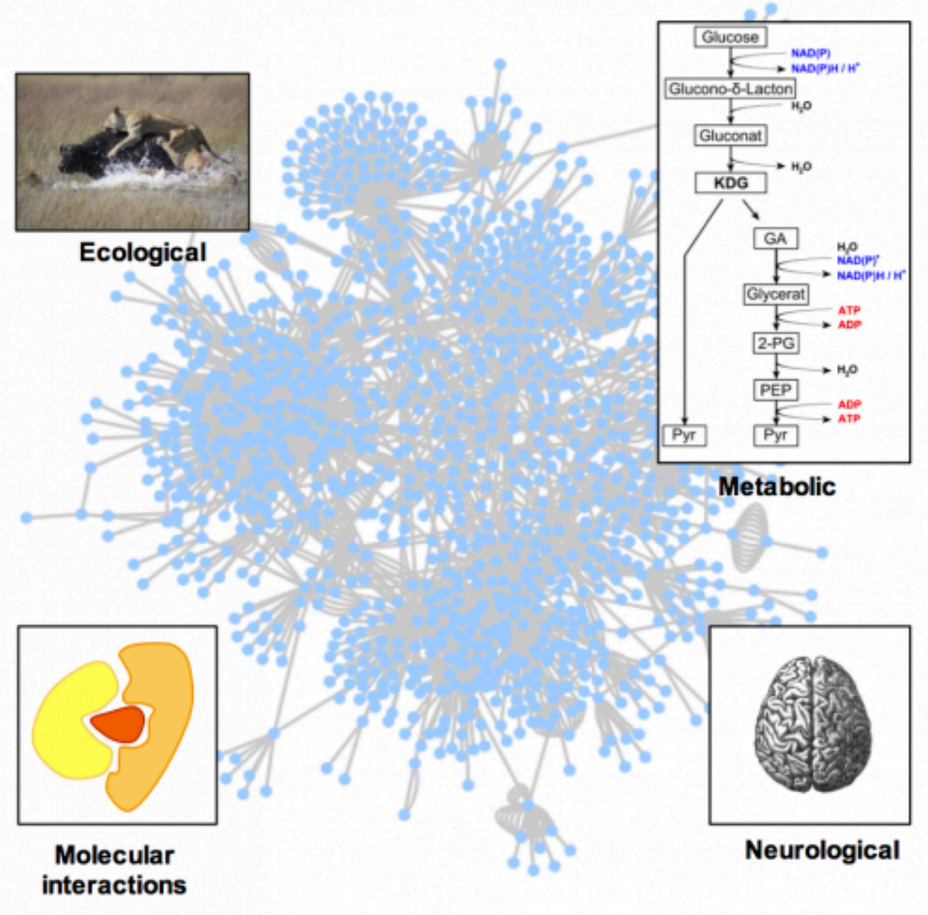
Biological Networks

Networks in Biology

So far we have only talked about sequences

- Many *interactions* in biology are not captured in sequences
- We use graph theory to make biological conclusions

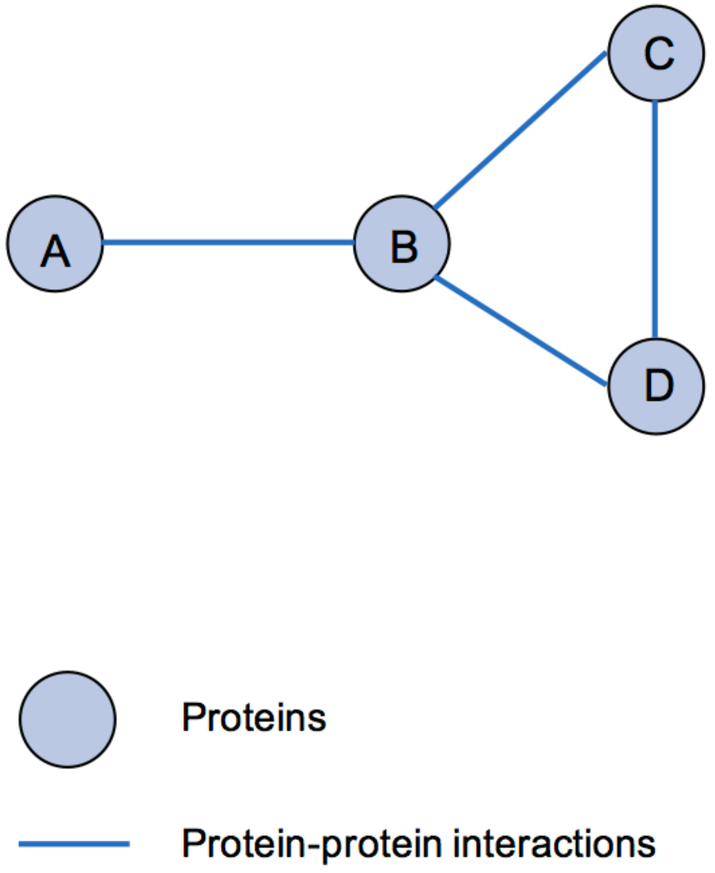


Protein-protein interaction networks

Represent some physical relationships between proteins

- central to practically every process in the cell

Most common type of graph we will look at

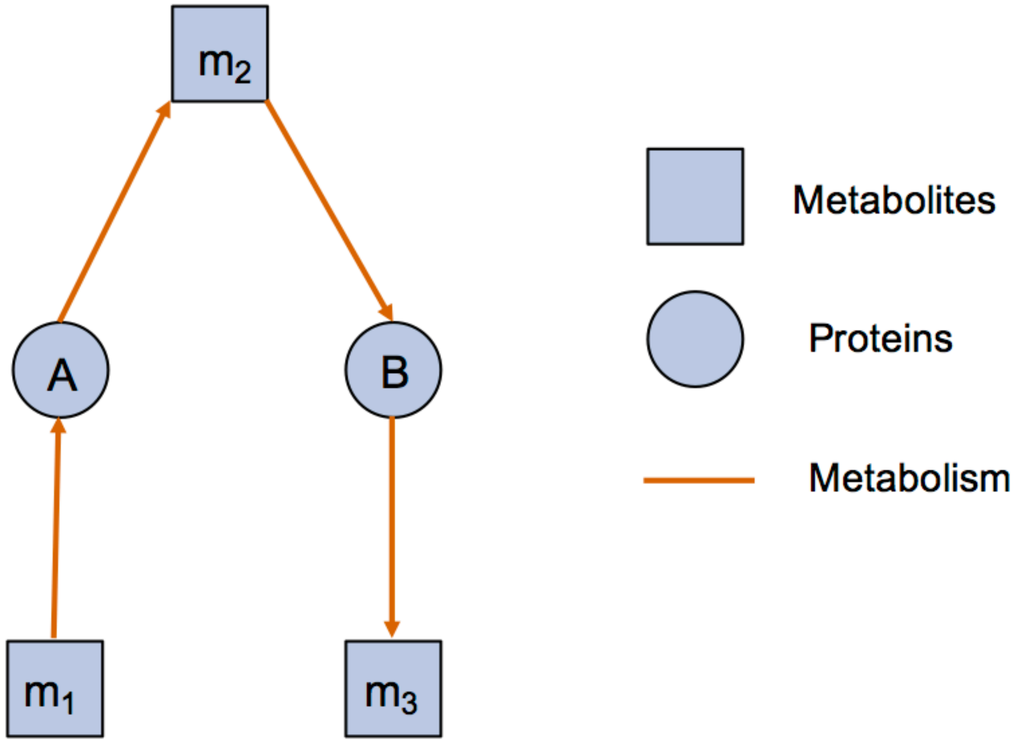


Metabolic Networks

Represent biochemical reactions, allow an organism to

- grow
- reproduce
- respond to the environment
- maintain structure

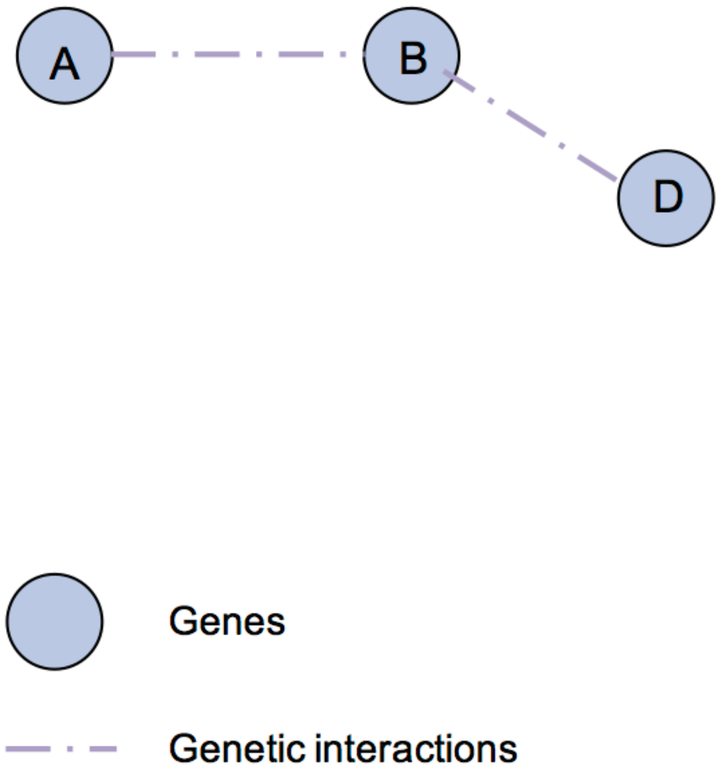
Edge direction shows the metabolic flow or a regulatory effect



Genetic Interaction Networks

Represents the functional relationship between genes rather than physical

"Genetic interaction is the synergistic phenomenon where the phenotype resulting from simultaneous mutations in two or more genes is significantly different from the phenotype that would result from adding the effects of the individual mutations"



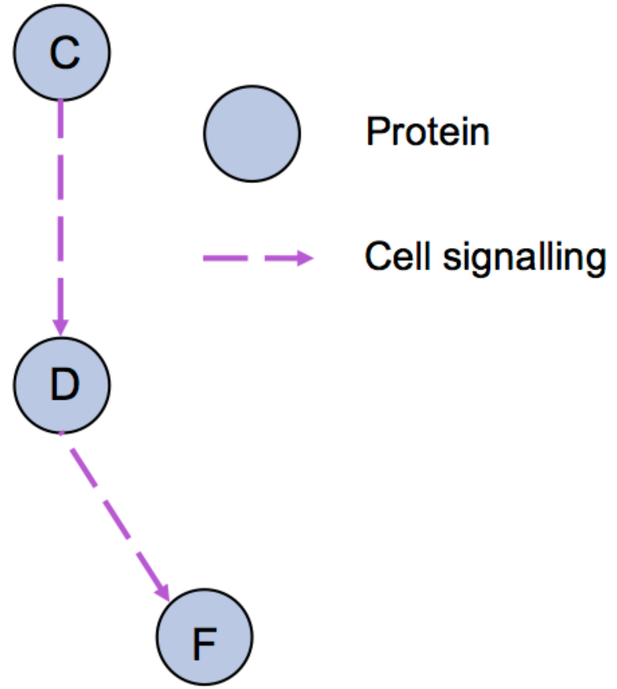
Cell Signaling Networks

Cell signalling is the communication system that controls cellular activities.

Signalling pathways represent the ordered sequences of events and model the information flow within the cell.

Systematically represented by two types of resources:

- **Pathway databases** (also known as 'process description' resources) such as Reactome, KEGG or Wikipathways. These aim to provide a formal representation of the current scientific consensus on cell signalling pathways. They are generated by manual curation and organise the information in the form of reactions, with substrates and products being affected by the action of catalysers. This information must be converted according to specific rules in order to be represented as a network. Some information loss can occur during this process.
- **Reaction network databases** (also known as 'activity flow' resources) such as Signor, SignaLink or SPIKE. These aim to capture known binary relationships in cell signalling, such as activation, phosphorylation, etc. They are generally manually curated, but not always. In contrast with the pathway databases, they are already graphs in the mathematical sense and require no transformation in order to be represented as a network.



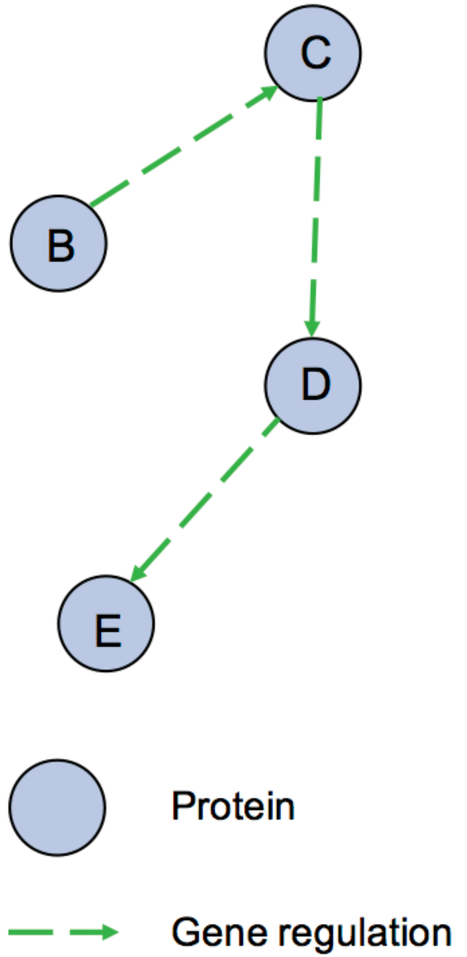
Gene/transcriptional regulatory networks

Represents how gene *expression* is controlled

Gene regulation networks can be considered as a sub-type of cell signalling networks, focusing on a specific signalling event which is often the final stage of a signalling cascade.

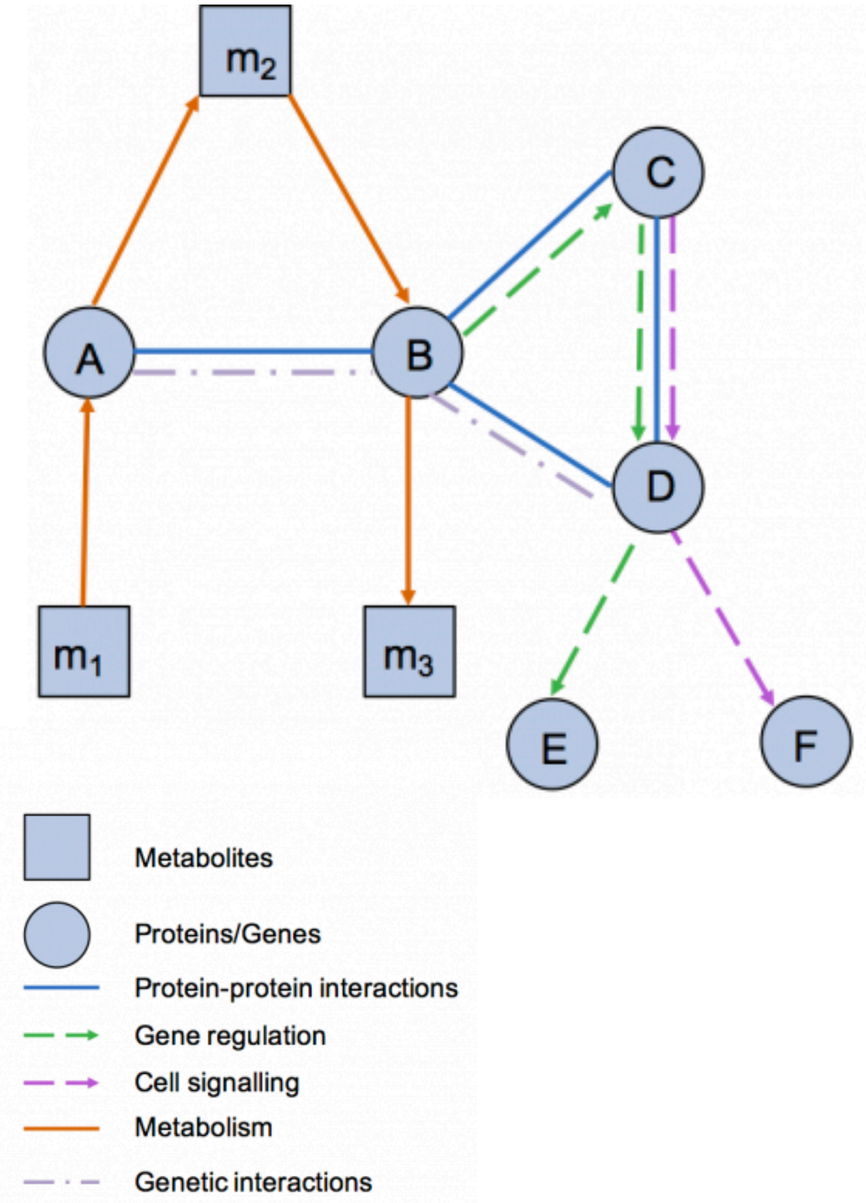
Regulatory RNAs and other mechanisms can also form part of this type of network.

Usually generated via databases representing consensus knowledge of gene regulation (e.g. Reactome or KEGG), although large-scale experimental datasets are increasingly available.



Combined Networks

The meaning of the nodes and edges used in a network representation depends on the type of data used to build the network and this should be taken into account when analysing it.



Data Sources

Network data is inherently noisy and incomplete

- sometimes the evidence from multiple sources will not overlap
- sometimes it will be contradictory

Typical Sources:

- **Manual curation** -- researchers will crawl the scientific literature to look for evidence and manually put it into a large database. Size is limited, cost is high.
- **High-throughput datasets** -- some lab methods can generate large networks, mainly PPI networks. Suffer from bias of the technique. Size is typically large.
- **Computational predictions** -- Using sets of evidence to predict relationships not necessarily in the existing data, but not contradicted.
- **Text mining** -- computationally extract information from literature using NLP.

Protein-protein interaction networks

Protein-protein interactions (PPIs) are essential to almost every process

- understanding PPIs is crucial for understanding cell physiology
- both in normal cells and disease states

PPI networks (PPIN) represent the physical contacts between proteins in the cell. These contacts:

- are specific;
- occur between defined binding regions in the proteins; and
- have a particular biological meaning (i.e., they serve a specific function).

Protein-protein interaction networks

PPI information can represent both transient and stable interactions:

- Stable interactions are formed in protein complexes (e.g. ribosome, haemoglobin).
- Transient interactions are brief interactions that modify or carry a protein, leading to further change (e.g. protein kinases, nuclear pore importins). They constitute the most dynamic part of the *interactome*.

Knowledge of PPIs can be used to:

- assign putative roles to uncharacterised proteins;
- add fine-grained detail about the steps within a signalling pathway; or
- characterise the relationships between proteins that form multi-molecular complexes such as the proteasome.

Interactome

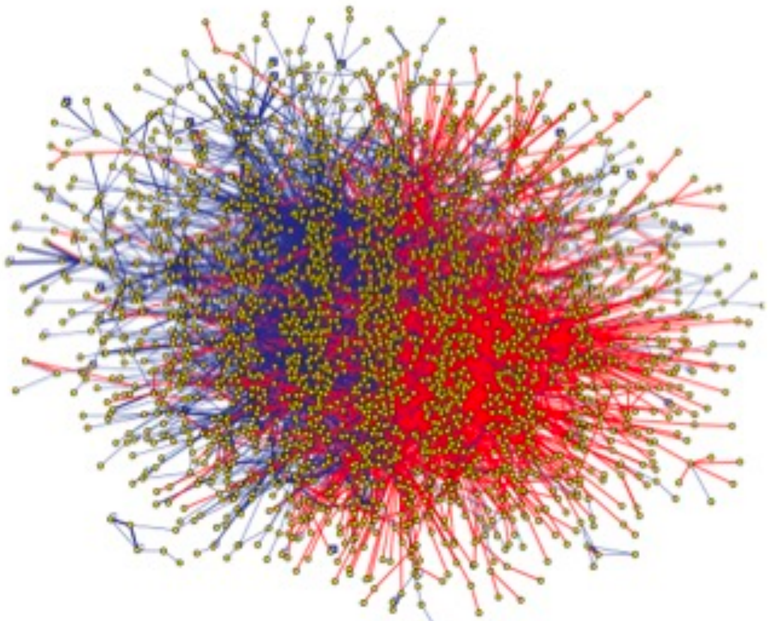
The interactome is the totality of PPIs that happen in a cell, an organism or a specific biological context.

The development of large-scale PPI screening techniques has caused an explosion in the amount of PPI data and the construction of ever more complex and complete interactomes.

This experimental evidence is complemented by the availability of PPI prediction algorithms.

Our current knowledge of the interactome is both *incomplete* and *noisy*.

Yeast



Human

Images from Jeong et al. Nature 2001. 411 and Rual et al. Nature 2005: 437

The Small World Effect

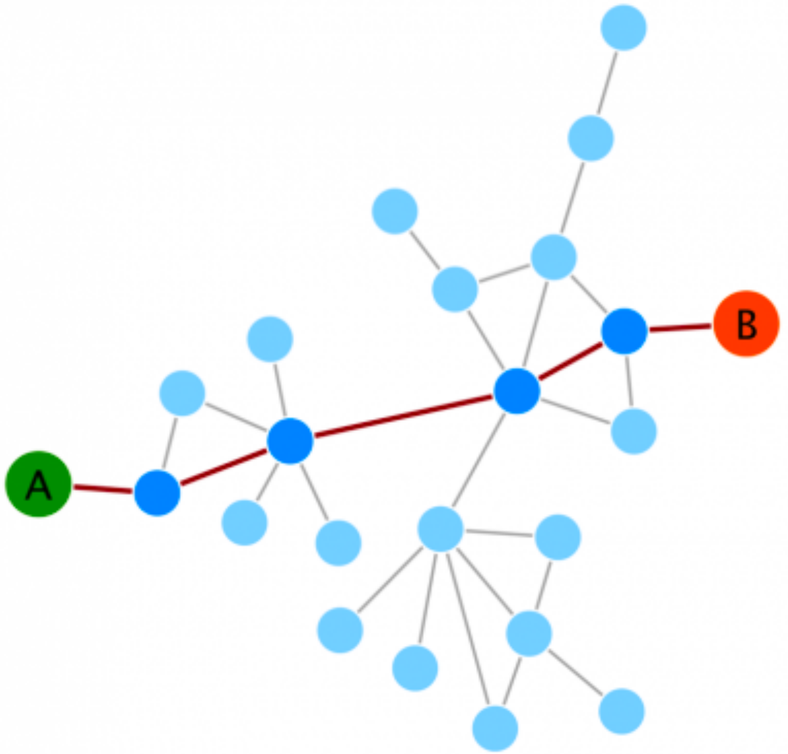
There is great connectivity between proteins

The network's diameter (the maximum number of steps separating any two nodes) is small, no matter how big the network is

Think "the 6 degrees of Kevin Bacon"

This it poses an interesting question:

- if the network is so tightly connected, why don't perturbations in a single gene or protein have dramatic consequences for the network?

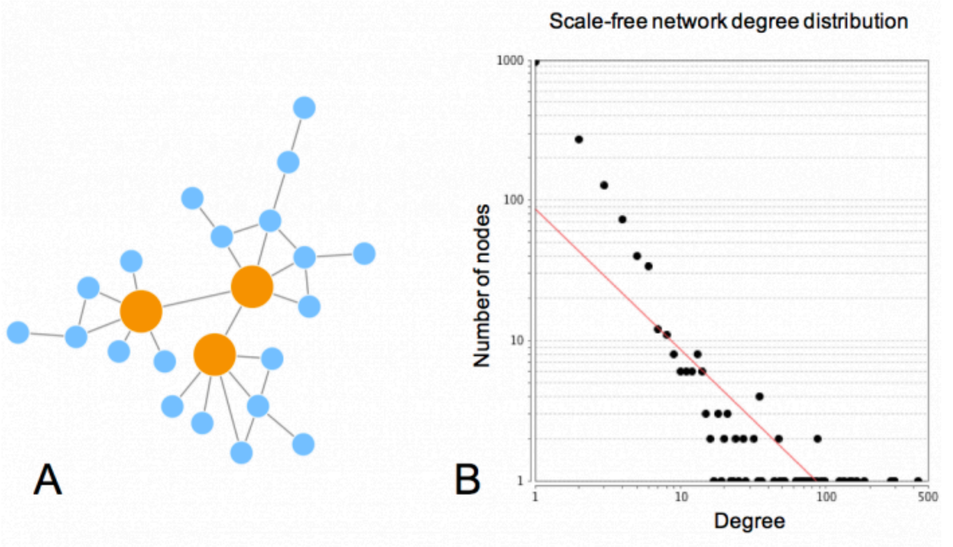


Scale-free networks

The majority of nodes (proteins) in scale-free networks have only a few connections to other nodes, whereas some nodes (hubs) are connected to many other nodes in the network

In PPI networks this "scale-free-ness" allows for:

- stability
- invariance in changes to scale
- vulnerability to targeted attack



Transitivity

The *transitivity* or *clustering coefficient* of a network is a measure of the tendency of the nodes to cluster together

- high transitivity means that the network contains communities or groups of nodes that are densely connected internally
- finding these communities is very important, because they can reflect functional modules and protein complexes

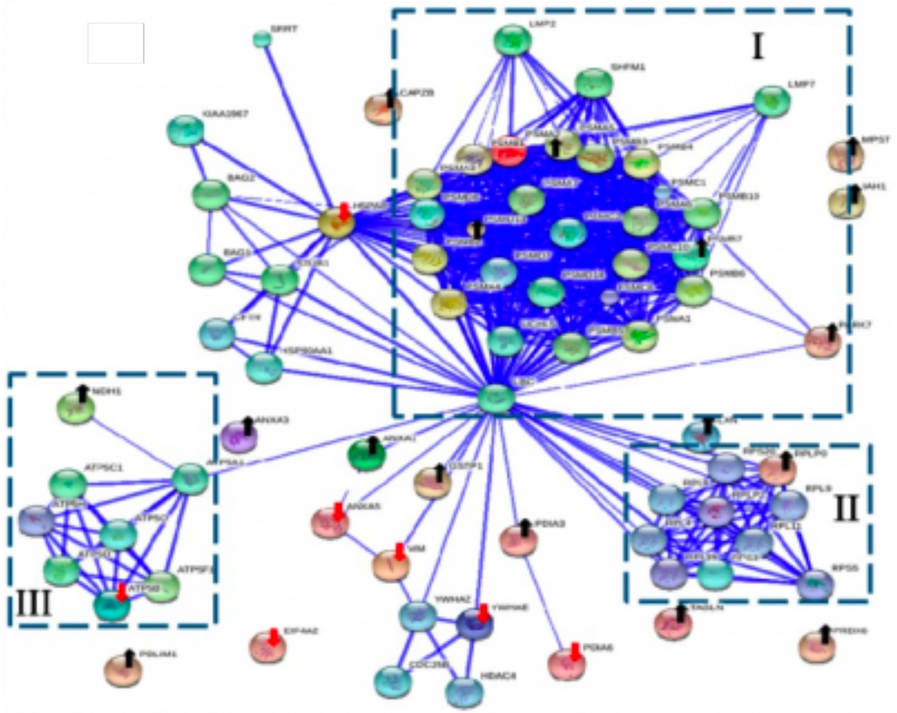
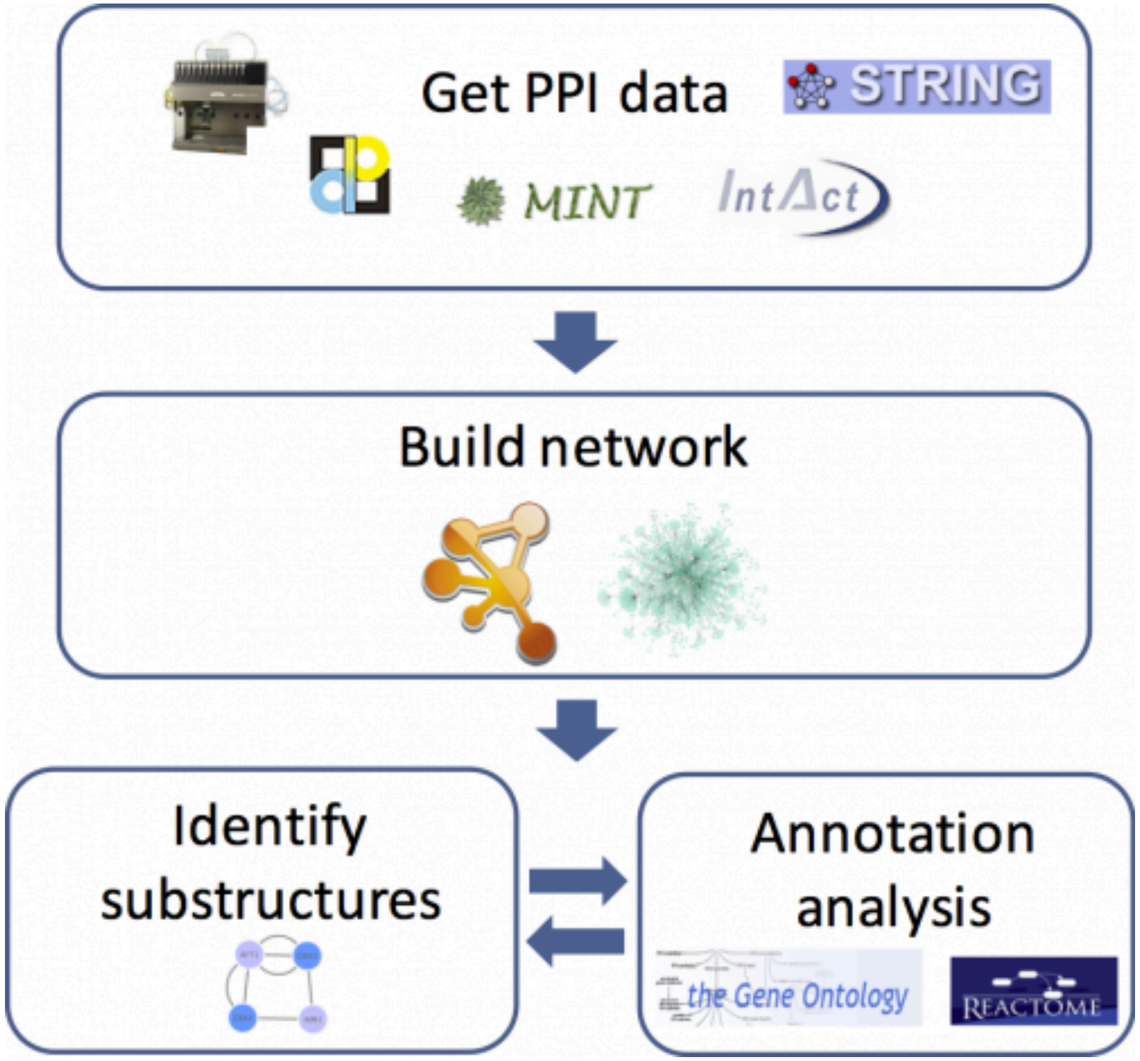


Image from Hsia et al. Int J Mol Sci. 2015: 16

Building a PPI Network



Measuring Network Confidence

It's important to know whether the interaction network can be trusted to represent a “real” biological interaction

- there is lots of noise, it's important to be stringent when evaluating the data
- also, can't just filter stuff since the interactome coverage is incomplete

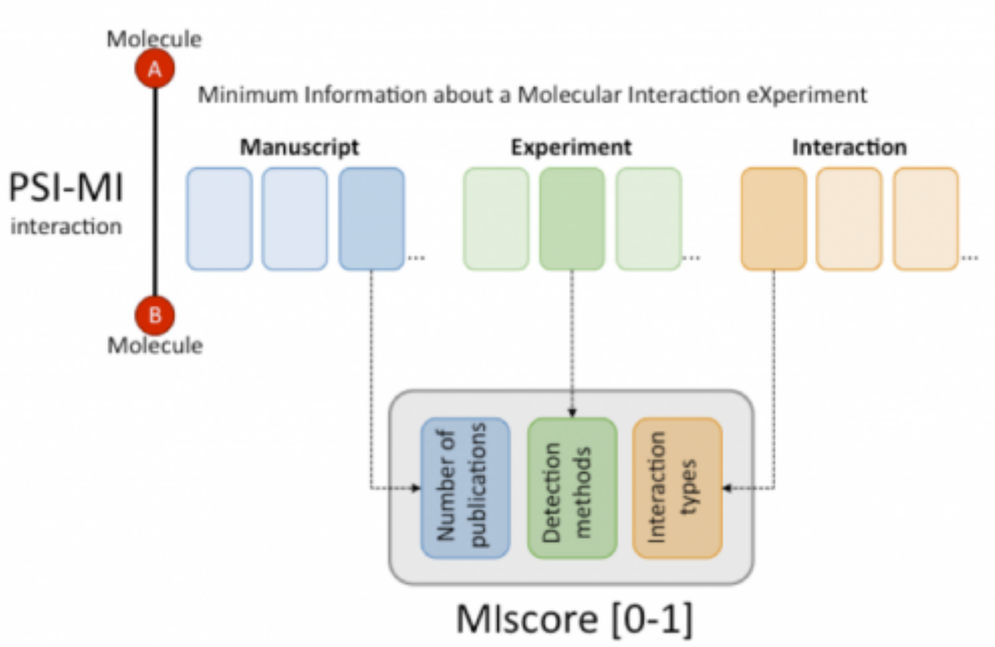
Strategies for measuring reliability:

- Contextual biological information regarding the proteins in the interaction.
  - For example, overlapping co-expression patterns.
- Count the number of times times an interaction is reported in the literature.
  - This is a popular and straightforward approach.
- Ensemble methods that integrate different strategies into a single score.

Measuring Network Confidence

MIscore assesses the reliability of protein-protein interaction data based on the use of standards.

- It gives an estimation of confidence weighting on all available evidence for an interacting pair of proteins.
- The method weights evidence provided from:
  - number of publications;
  - detection method;
  - interaction evidence type.



Different interaction detection methods and interaction types have different weights, assigned by a group of expert curator

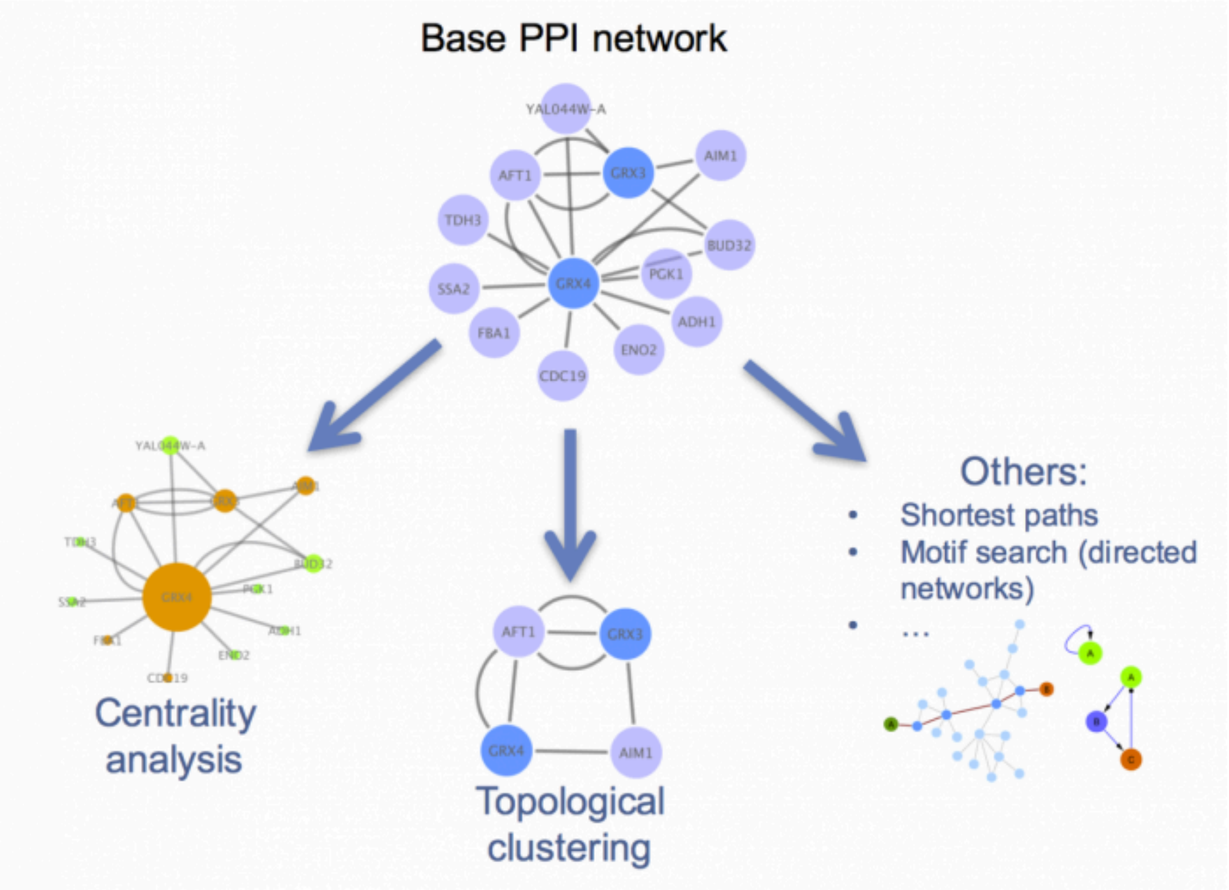
- These parameters are aggregated for each interacting pair and then normalized, giving a quantitative measure of how much experimental evidence for a given interaction.

Topology Analysis

Analyzing the topological features of a network is a useful way of identifying relevant participants and substructures that may be of biological significance.

Some methods

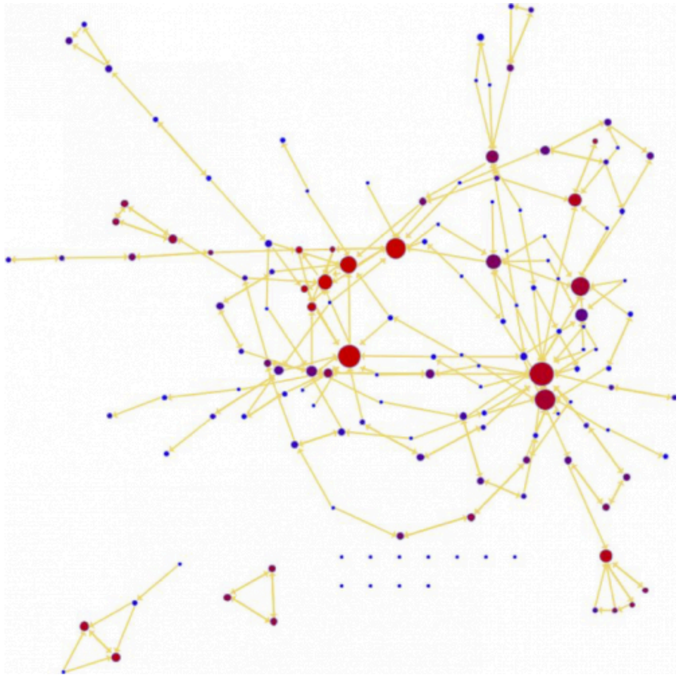
- centrality analysis
- topological clustering
- search for shortest paths
- motifs that are more often applied to networks with directionality



Centrality Analysis

The definition of ‘central’ varies with the context or purpose of our analysis.
Centrality can be measured using different metrics and criteria:

- Degree of the nodes
  - As we saw earlier, nodes with a high degree (hubs) are key in maintaining some characteristics of scale-free networks
  - This is a local measure since it does not take into account the rest of the network and the importance we give to its value depends strongly on the network's size.
- Global centrality measures
  - Global centrality measures take into account the whole of the network.
  - Two of the most widely used global centrality measures are closeness and betweenness centralities.
- Other measures of centrality
  - Often calculated using random walks.
  - Can be combined with the weights assigned to nodes or edges in the graph to influence the centrality calculation derived from other features.
  - The method used by the Google PageRank.



Closeness

Closeness centrality is estimates how fast the flow of information would be through a given node to other nodes.

Measures how short the shortest paths are from a node to all other nodes.

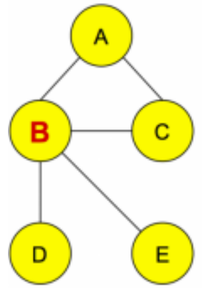
- Typically expressed as the normalized inverse of the sum of the topological distances in the graph.

In the example shown the distance matrix for the graph on the left and the calculations to get the closeness centrality on the right.

- Node B is the most central node according to these parameters.

$$CC(i) = \frac{N-1}{\sum_j d(i,j)}$$

where
$i \neq j$,
d_{ij} is the length of the shortest path between nodes i and j in the network,
N is the number of nodes.



	A	B	C	D	E	farness $\sum_{j=1}^{n} d(i,j)$	$CC(i) = \frac{N-1}{\sum_j d(i,j)}$
A	0	1	1	2	2	6	(5-1)/6= 0.67
B	1	0	1	1	1	4	1.00
C	1	1	0	2	2	6	0.67
D	2	1	2	0	2	7	0.57
E	2	1	2	2	0	7	0.57

N = 5 (# of nodes)

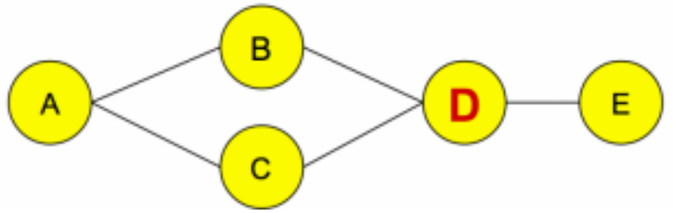
Betweenness

Betweenness centrality is based on communication flow.

- Nodes with a high betweenness centrality are interesting because they lie on communication paths and can control information flow.
- These nodes can represent important proteins in signaling pathways and can form targets for drug discovery.
- It is basically defined as the number of shortest paths in the graph that pass through the node divided by the total number of shortest paths.
- Measures how often a node occurs on all shortest paths between two nodes.
  - The betweenness of a node N is calculated considering couples of nodes (v1, v2) and counting the number of shortest paths linking those two nodes, which pass through node N.

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) / g_{jk}$$

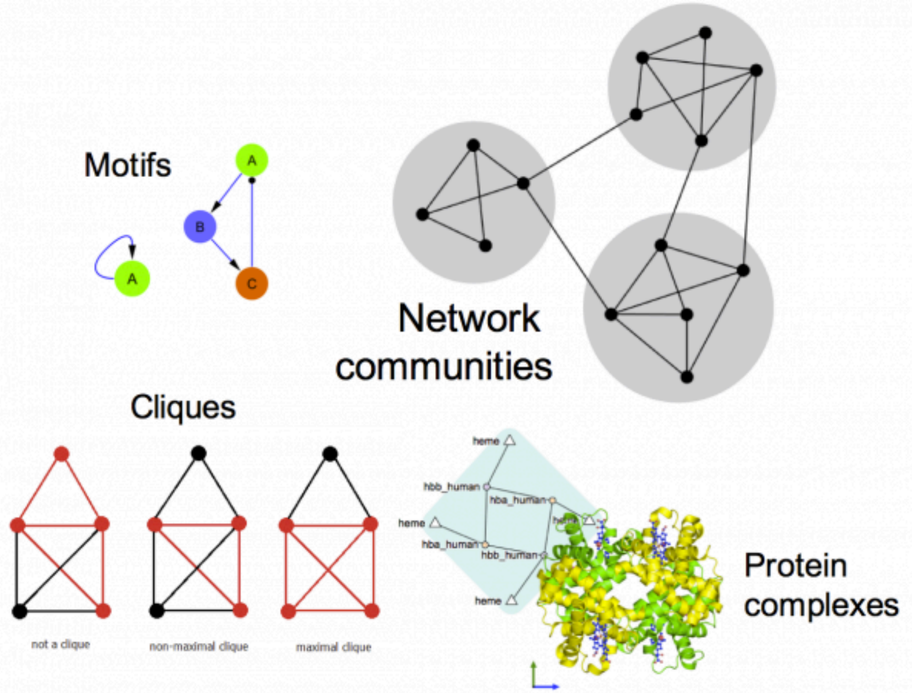
Where g_{jk} = the number of geodesics (shortest paths) connecting *jk*, and $g_{jk}(n_i)$ = the number that node *i* is on.



Clustering analysis

Community analysis help reduce network complexity and extract functional modules

- **Community / Cluster**
  - A group of nodes that are more connected within themselves than with others.
  - Two categories in PPI Networks: functional modules and protein complexes.
- **Module**
  - Exchangeable functional units in which the nodes (proteins) do not have to be interacting in the same time or space.
  - Its functional properties do not change when it is placed in a different context.
- **Complex**
  - A group of proteins that interact with each other at the same time and in the same space, forming relatively stable multi-protein machinery.
- **Clique**
  - A subset of nodes in which every node is connected with every other member.
- **Motif**
  - Statistically over-represented sub-graphs in a network.
  - They correspond with a pattern of connections that generates a characteristic dynamical response (e.g. a negative feedback loop).
  - The goal is often to find functional modules or protein complexes that execute defined biological functions.



Clustering analysis

Methods that exclusively use the topology of the network to find closely-connected components are known as 'community detection methods'.

- No assumptions are made about the internal structure of these communities.

Finding the best community structure is algorithmically extremely complex, only possible for very small networks.

Many approximation methods have been developed.

- Clique-percolation method
- Markov Clustering Algorithm (MCL)
- Fuzzy C-Means
- Affinity Propagation
- Chinese Whispers Clustering
- Label Propagation Clustering
- **Newman-Girvan fast greedy algorithm,** and
- **the MCODE algorithm.**

Other methods combine topology of the network and some external property, such as protein expression values.

- For instance connected regions within a network with differential expression.

Newman-Girvan

Developed for the study of networks in general, with a special focus on social and biological networks.

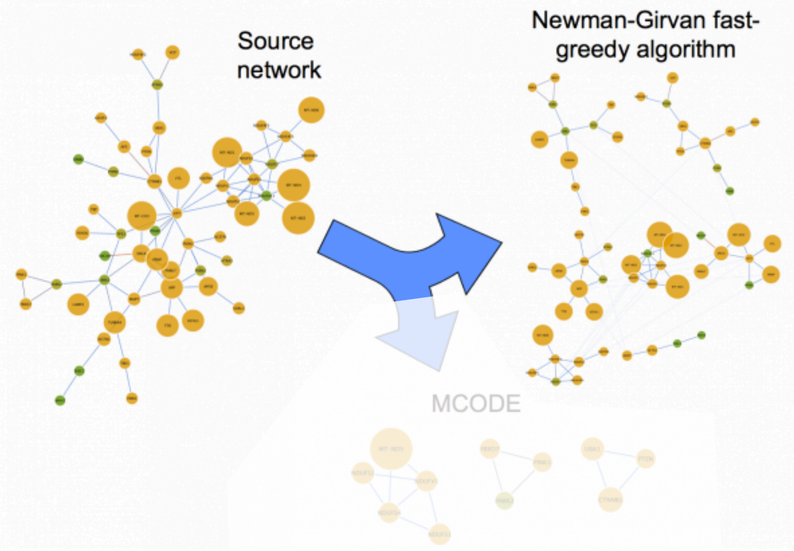
Uses the edge betweenness centrality measure.

- Edges that connect different communities have higher centrality values, since a larger proportion of shortest paths will pass through them.

Method

- use the edge betweenness centrality scores to rank the edges,
- remove the most central edges
- re-calculates the betweenness scores until no edges are left.
- Edges affected by the removal are part of the same community.

Can be considered a ‘naïve’ approach that will define communities even when they are only marginally more connected than the rest of the network.



MCODE

Developed to find protein complexes in PPI networks.

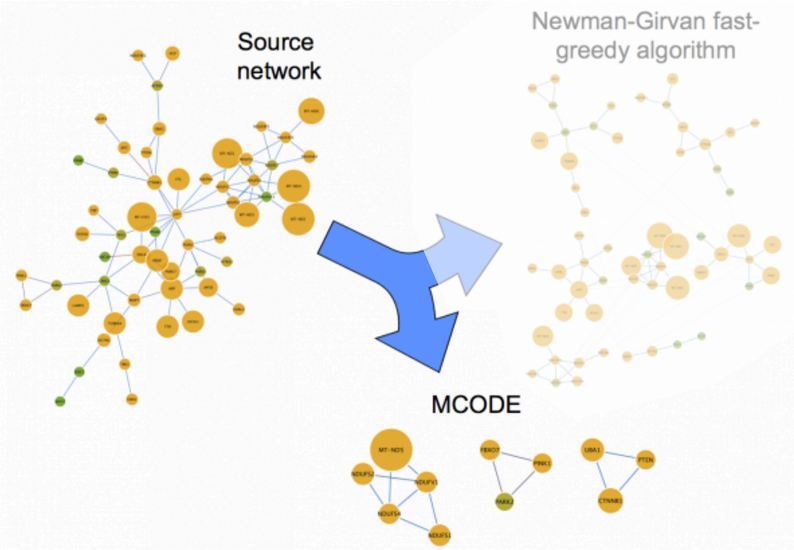
- Considered to be more stringent than the Newman-Girvan algorithm,
- It aims to find only those sub networks that are very highly interconnected, representing relatively stable, multi-protein complexes that function as a single entity in time and space.

The algorithm uses a three-stage process:

1. **Weighting:** a higher score is given to those nodes whose neighbors are more interconnected.
2. **Molecular complex prediction:** starting with the highest-weighted node (seed), recursively move out, adding nodes to the complex that are above a given threshold.
3. **Post-processing:** applies filters to improve the cluster quality.

Stringency -- how interconnected the nodes within a sub-network must be in order to be considered a separate community.

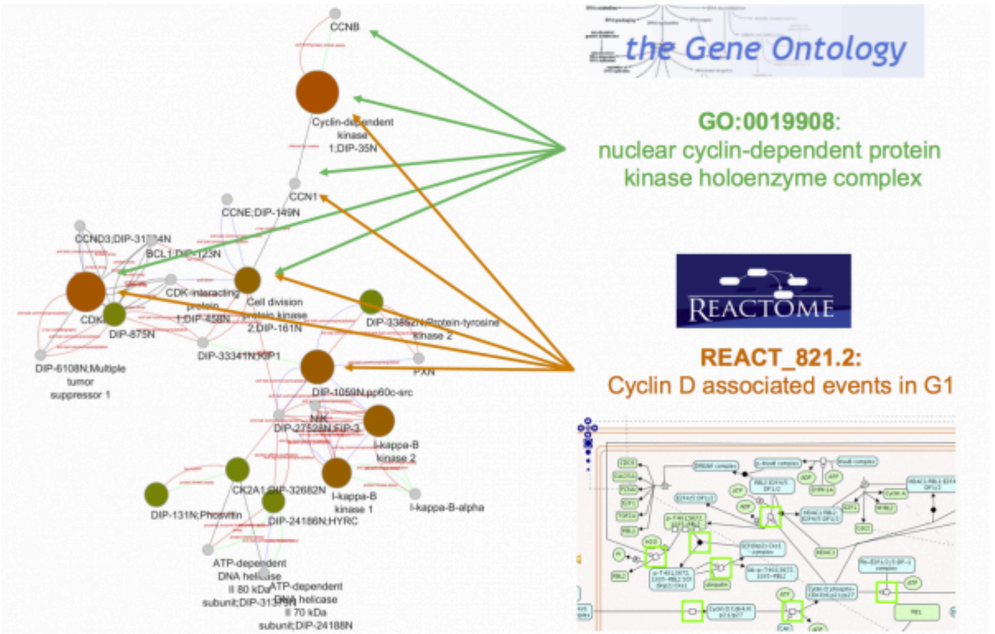
- Changes depending on the biological question underlying the analysis.



Annotation enrichment analysis

Annotation enrichment analysis uses gene/protein annotations to infer which annotations are over-represented in a list of genes/proteins taken from a network.

- Annotation tools perform statistical test tries to that answer:
  - *When sampling X proteins (test set) out of N proteins (reference set; graph or annotation), what is the probability that x, or more, of these proteins belong to a functional category C shared by n of the N proteins in the reference set.*
- The result of this test provides us with a list of terms that describe the list/network, or rather a part of it, as a whole.



This analysis is most frequently performed using GO annotation as a reference.

- This is a widely used technique that helps characterize the network as a whole or sub-sets of it, such as inter-connected communities found through topological clustering analysis.

More complex versions of this technique can factor in continuous variables such as expression fold change.

Limitations of annotation enrichment

Annotation enrichment is limited by the annotations themselves.

- Certain areas of biology are more thoroughly annotated and better described than others, with more detail and more accurate terms for well-known processes.
- At the level of the proteins, more "popular" proteins are better annotated. This introduces a certain bias into the statistical analysis.

GO terms can be assigned either by

- a human curator who performs careful, manual annotation or
- by computational approaches that use the basis of manual annotation to infer which terms would properly describe uncharted gene products.
- They use a number of different criteria that always refer to annotated gene products, such as sequence or structural similarity or phylogenetic closeness.
- The importance of the computationally derived annotations is quite significant, since they account for roughly 99% of the annotations that can be found in GO.

Simplifying the interpretation of annotation enrichment results

The data sources can be very complex and detailed in their annotation leading to the generation of overwhelmingly complicated networks of inter-related and similar terms.

The simplest approach is to use simplified ontologies.

- Use ontologies where fine detailed terms are removed and assigned to broader, more general parent terms
- Represent the results as a network of terms, where directed edges represent term relationships as defined in the ontology used.
  - Allows tools from graph theory to be used to reorganize the layout of the network to uncover communities inside these terms networks which helps to simplify the output. BiNGO only provides the network view, so other tools are required to further simplify the analysis.
- Use the output from some of the most popular annotation enrichment tools and apply clustering and automatic layout techniques to overlap similar gene sets and provide a simplified representation of annotation enrichment results.
  - Especially useful when comparing results from different sets, i.e. two different conditions.

It is important to know the limitations of the annotation resource, be aware of the inherent complexity of the results. Network analysis techniques can help simplify the interpretation of these results.

Summary

Biological Networks

- Network biology makes use of the tools provided by graph theory to represent and analyse complex biological systems.
- There are several types of biological networks: genetic, metabolic, cell signaling etc.
- Networks are represented by nodes and edges. Nodes represent different entities (e.g. genes or proteins) and edges convey information about how the nodes are linked.

Summary

Protein-Protein Interaction Networks

- **Small-world effect:** Network diameter is usually small (~ 6 steps), no matter how big the network is.
- **Scale-free:** A small number of nodes (hubs) are lot more connected than the average node.
- **Transitivity:** The networks contain communities of nodes that are more connected internally than they are to the rest of the network.

Summary

Analyzing PPI Networks

- Several tools are available for PPIN analysis.
- It is important to be aware of the type and quality of the data used. Confidence scoring tools such as MIscore can help select the best characterized interactions.
- Two of the most used topological methods to analyze PPINs are:
  - **Centrality analysis:** Which identifies the most important nodes in a network, using different ways to calculate centrality.
  - **Community detection:** Which aims to find heavily inter-connected components that may represent protein complexes and machineries
- Annotation enrichment analysis is a tool often used when analyzing PPINs.
  - It uses resources such as the Gene Ontology (GO) or Reactome to infer which annotations are over-represented in a list of genes or proteins.
  - It can produce complex results that can be simplified using network analysis tools.

PathLinker

npj | Systems Biology and Applications

Article | Open Access | Published: 03 March 2016

Pathways on demand: automated reconstruction of human signaling networks

Anna Ritz, Christopher L Poirel, Allison N Tegge, Nicholas Sharp, Kelsey Simmons, Allison Powell, Shiv D Kale & TM Murali

npj Systems Biology and Applications **2**, Article number: 16002 (2016) | Cite this article

Pathway reconstruction problem

Given

- weighted, directed interactome, *G*, with physical & regulatory interactions
- receptors, *S*, in a signaling pathway of interest
- transcriptional regulators (TRs), *T*, in the same pathway
- a parameter *k*

Find

- the *k* highest scoring loopless paths that begin at any receptor in *S* and end at any TR in *T*
- the score of the path is the product of the edge weights (all in *[0,1]*)

Method Setup

Modify the graph

- Add an extra source node *s* and an extra sink node
- add edges *(s,x)* for $x \in S$
- add edges *(y,t)* for $y \in T$
- assign the following *costs* to each edge *(u,v)*

$$c_{uv} = \begin{cases} -log(w_{uv}) & \text{if } u, v \in V \backslash \{s, t\} \\ 0 & \text{if } u = s \text{ or } v = t \end{cases}$$

- Let the cost of a path be the sum of the edges on the path.

The least costly $s \rightarrow t$ path will be the highest weight $s \rightarrow t$ path

Yen's Algorithm

Assume you know some set of *i-1* shortest *s-t* paths in *G* and you want to find the *i*th

- the new path will deviate from one of the previous path at some vertex

For each *i'<i* and for all vertices *j* in the path

- create a new graph *G'* that removes all vertices on the path from *s* to *j*
- remove all outgoing edges from *j* that are in a previously found path
- run Dijkstra's algorithm to find the smallest *j-t* path

Path *i* is then the shortest path found by extending from all possible *i'* and *j*

This is $O(kn(m+n \log n))$-time

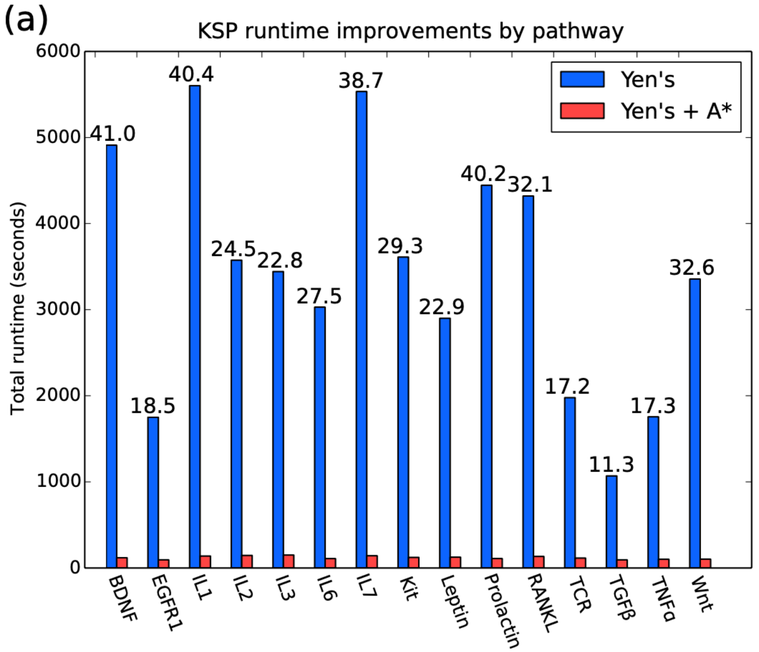
Using the A* Algorithm

Let $h(v) = d_G(v)$ where d_G is the shortest v-t path in G

- computed once in advance for all v

Change the priority queue in Dijkstra's algorithm from c_{uv} to $c_{uv} + h(v)$

Does not change the asymptotic run time of Yen's algorithm, but improves running time in practice



PathLinker

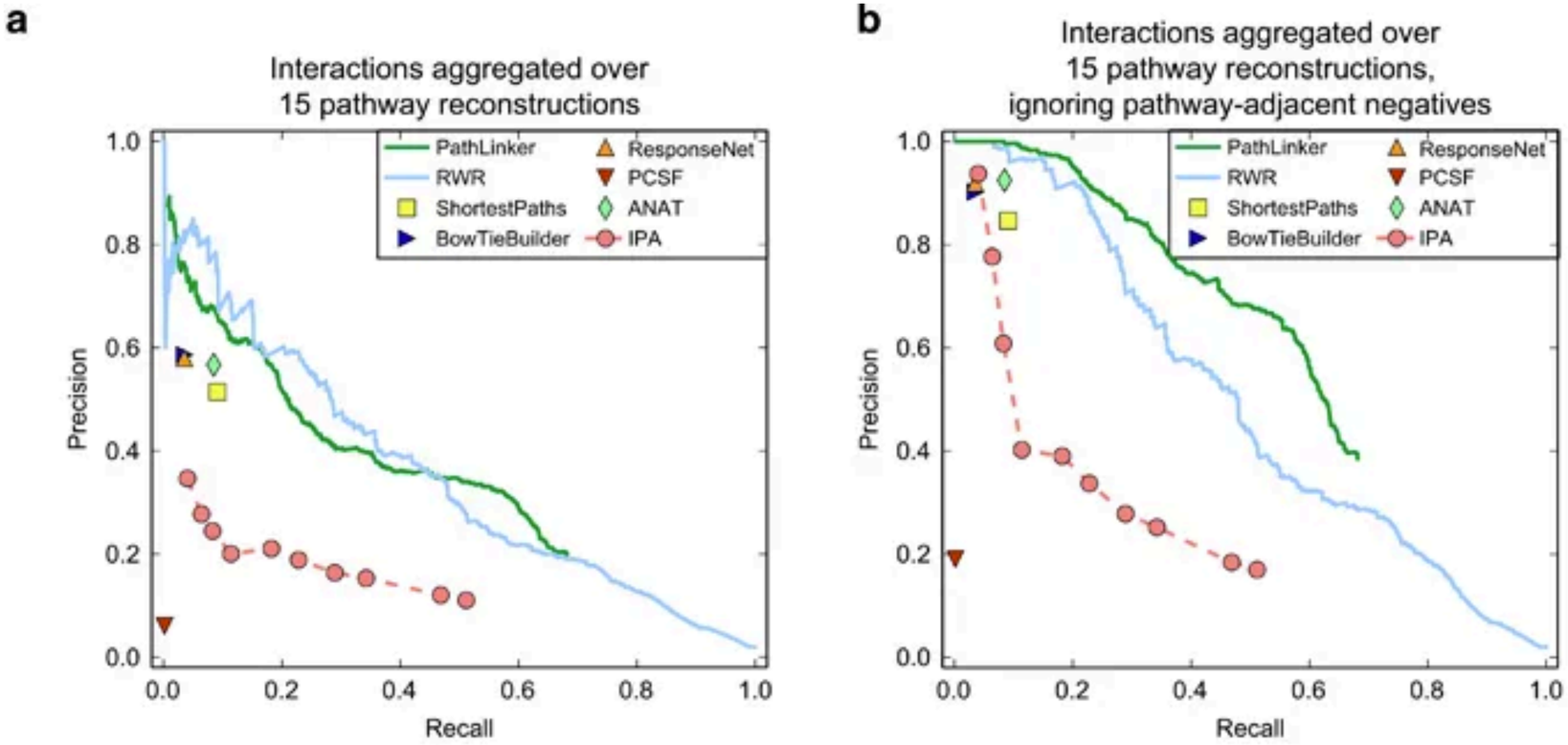
Algorithm

- Find the set of *k* highest scoring paths $P_1, P_2, \ldots, P_k$ where each $P_i = (V_i, E_i)$
- Return $G_k = \left(\cup_{1 \leq i \leq k} V_i, \cup_{1 \leq i \leq k} E_i \right)$

PathLinker Results

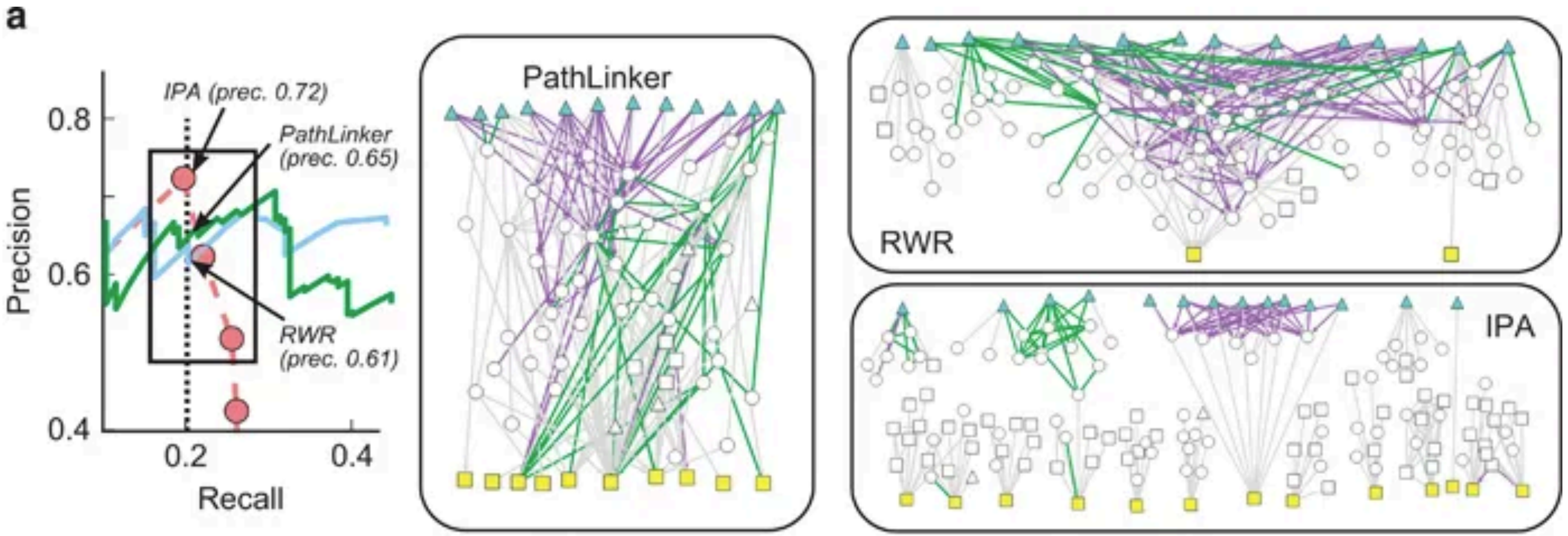
Since $G_{(k-1)} \subseteq G_k$ the graphs grow smoothly with k.

Precision and recall are measured by ranking the nodes and edges in order of when they first appear and compare them to some reference.



(**a**) Precision and recall of the interactions in pathway reconstructions computed by PathLinker and other algorithms. (**b**) Precision and recall of PathLinker and RWR without considering interactions adjacent to the pathway (distance=1).

PathLinker Results



Blue triangles: Wnt receptors;
yellow squares: Wnt TRs,
green edges: NetPath interactions,
purple edges: KEGG interactions that are not present in NetPath

NetBox

PLOS ONE

PUBLISH ABOUT BROWSE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Automated Network Analysis Identifies Core Pathways in Glioblastoma

Ethan Cerami, Emek Demir, Nikolaus Schultz, Barry S. Taylor, Chris Sander

Published: February 12, 2010 • https://doi.org/10.1371/journal.pone.0008918

NetBox

Basic Algorithm

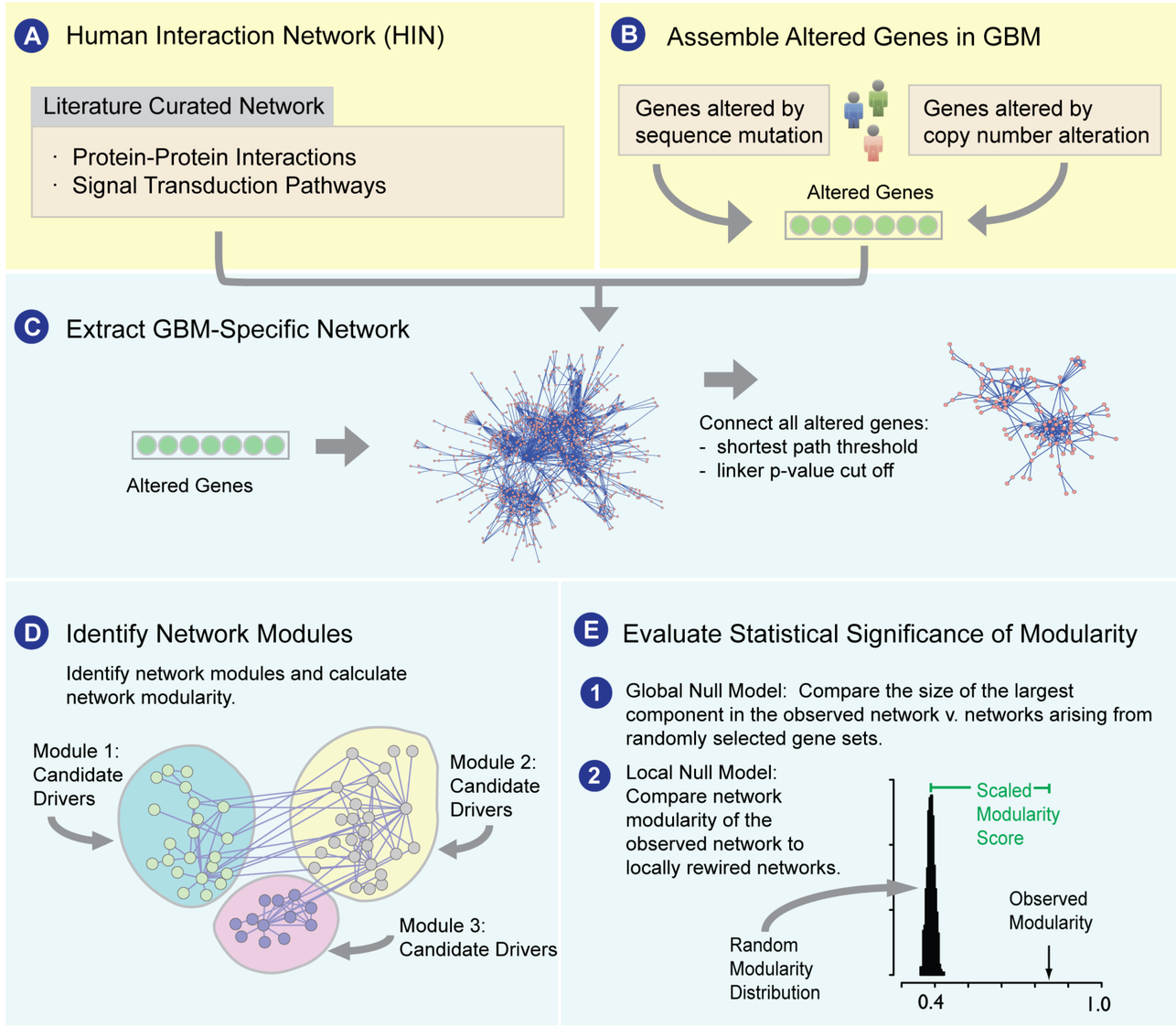
A. create human interactome (both interaction and pathway information)

B. find mutated or copy number variant genes for condition in question

C. extract these genes and their neighbors from the interactome

D. run the Newman-Girvan algorithm to find modules

E. analyze statistical significance

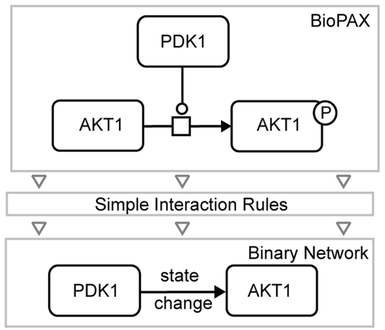


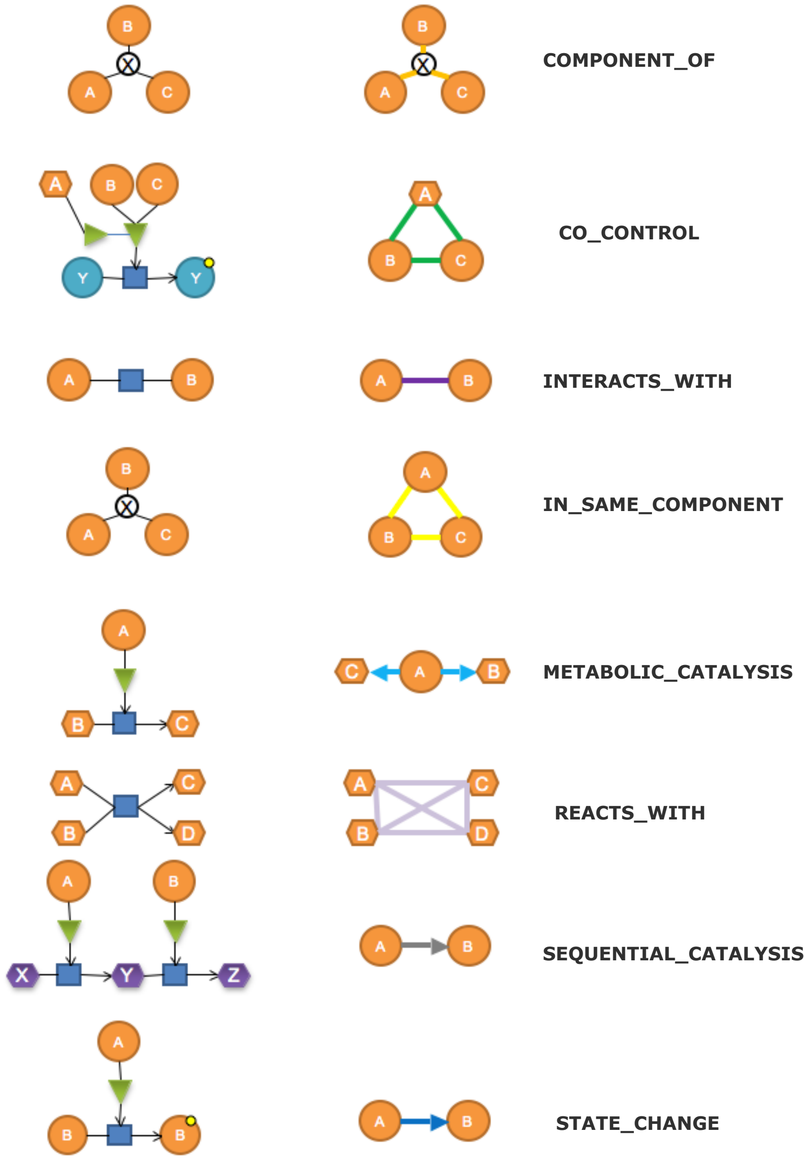
Curating the Interactome

Data was sourced from:

- the Human Protein Reference Database (interaction data)
- Reactome (pathways)
- NCI/Nature Pathway Interaction Map (pathways)
- MSKCC Cancer Cell Map (pathways)

Converted from complex mappings (input, output, catalyst) to a binary network





Create the condition graph

Create an empty graph *G*

For each altered gene *X*

- find all neighbors of *X* in the interactome
- add the neighbors and *X* to *G*

Prune

- any gene with degree 1 (they don't lead to any links between genes)
- for all other (non-altered genes) use Benjiamini-Horchberg correction to find the probability that these genes are included by random chance
- remove any genes with p-value > 0.05

Modularity Score

The fraction of edges that connect noces within modules minus the expected value of that quantity:

$$M = \sum_{1 \le s \le N_M} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right]$$

- N_M -- number of modules
- l_s -- number of edges in module s
- L -- number of edges in the graph
- d_s -- sum of the degrees of the nodes in s

Random network $M \to 0$, strong modularity $M \to 1$

Module detection

Using a modified Newman-Girvan:

1. use the edge betweenness centrality scores to rank the edges,
2. remove the most central edge
3. calculate *M* on the remaining graph
4. go to 1 until no edges are left
5. return the graph that lead to the highest value of *M*

Assessing Significance

Global Random Gene Set

- **assesses the level of global connectivity in the network being tested**
- generate a random set of genes the size of the set of altered genes
- compare the size of the largest connected component in the networks
- compute the probability the random network component is smaller

Local Random Rewiring

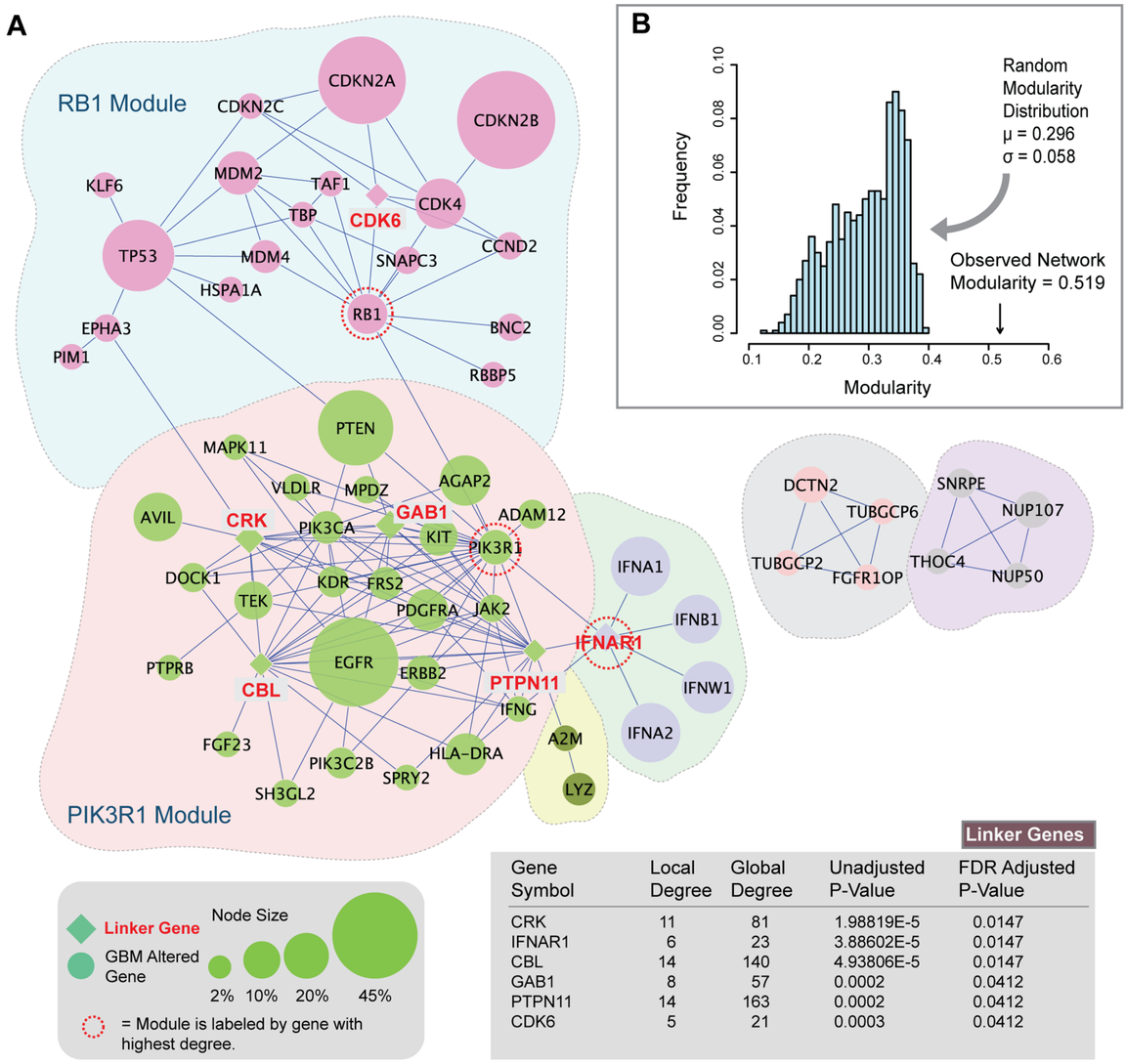
- **assesses the significance of the modularity observed**
- generate random links between genes in the network
- each node will still have the same degree and network size is the same
- measure the modularity score of this new network

NetBox Results

Network modules identified in Glioblastoma.

A. Modules are densely connected sets of altered genes that may reflect oncogenic processes.
   - 10 modules were identified, the largest of which are shown.
   - Linker genes, indicated in red, are not altered in Glioblastoma, but are statistically enriched for connections to Glioblastoma-altered genes.

B. The observed modularity of the GBM network (0.519) is compared with 1000 randomly rewired networks (average 0.296, standard deviation 0.058).
   - z-score, or scaled modularity score, of 3.84.



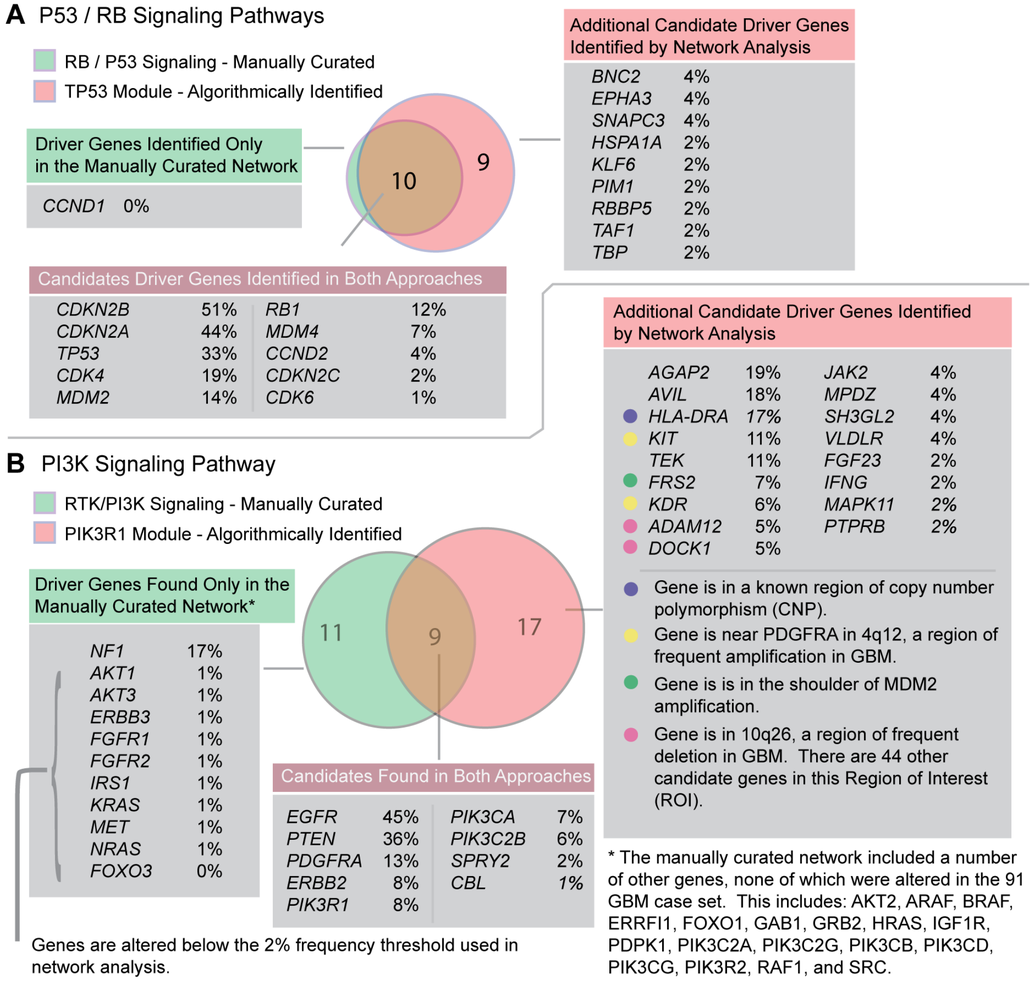
Linker Genes

Gene Symbol	Local Degree	Global Degree	Unadjusted P-Value	FDR Adjusted P-Value
CRK	11	81	1.98819E-5	0.0147
IFNAR1	6	23	3.88602E-5	0.0147
CBL	14	140	4.93806E-5	0.0147
GAB1	8	57	0.0002	0.0412
PTPN11	14	163	0.0002	0.0412
CDK6	5	21	0.0003	0.0412

NetBox Results

Automated network analysis approach is in close agreement with previous manually curated pathway analysis approach.

- The original pathway analysis of TCGA glioblastoma datasets was derived by mapping observed gene alterations onto a manually curated GBM-specific network, based on the glioblastoma literature.
- This non-algorithmic analysis identified driver alterations in the p53, RB and PI3K pathways.
- Our automated network analysis approach is in close agreement with these results (top: P53/Rb; bottom: PI3K). The one main exception is that network analysis does not identify NF1 as a participant in the PI3K module. Additional candidate driver genes identified by network analysis, including AGAP2, are identified and annotated on the right.
- Percentage values after each newly identified candidate driver indicate percent of cases with genetic alterations (sequence mutations, homozygous deletions, or multi-copy amplifications) across the 84 TCGA GBM cases analyzed.



NetBox Results

Network analysis identifies three additional altered modules, including the DCTN2 module, which is involved in microtubule organization.

- Each of the altered modules is implicated by homozygous deletions or multi-copy amplifications across the 84 analyzed GBM cases.
- Each module is annotated with Gene Ontology enrichment, chromosome location, statistical significance of copy number alteration against a background model of random aberrations, as determined by RAE copy number analysis; assessment of correlation between copy number and mRNA expression, and genomic signature across 84 GBM cases.