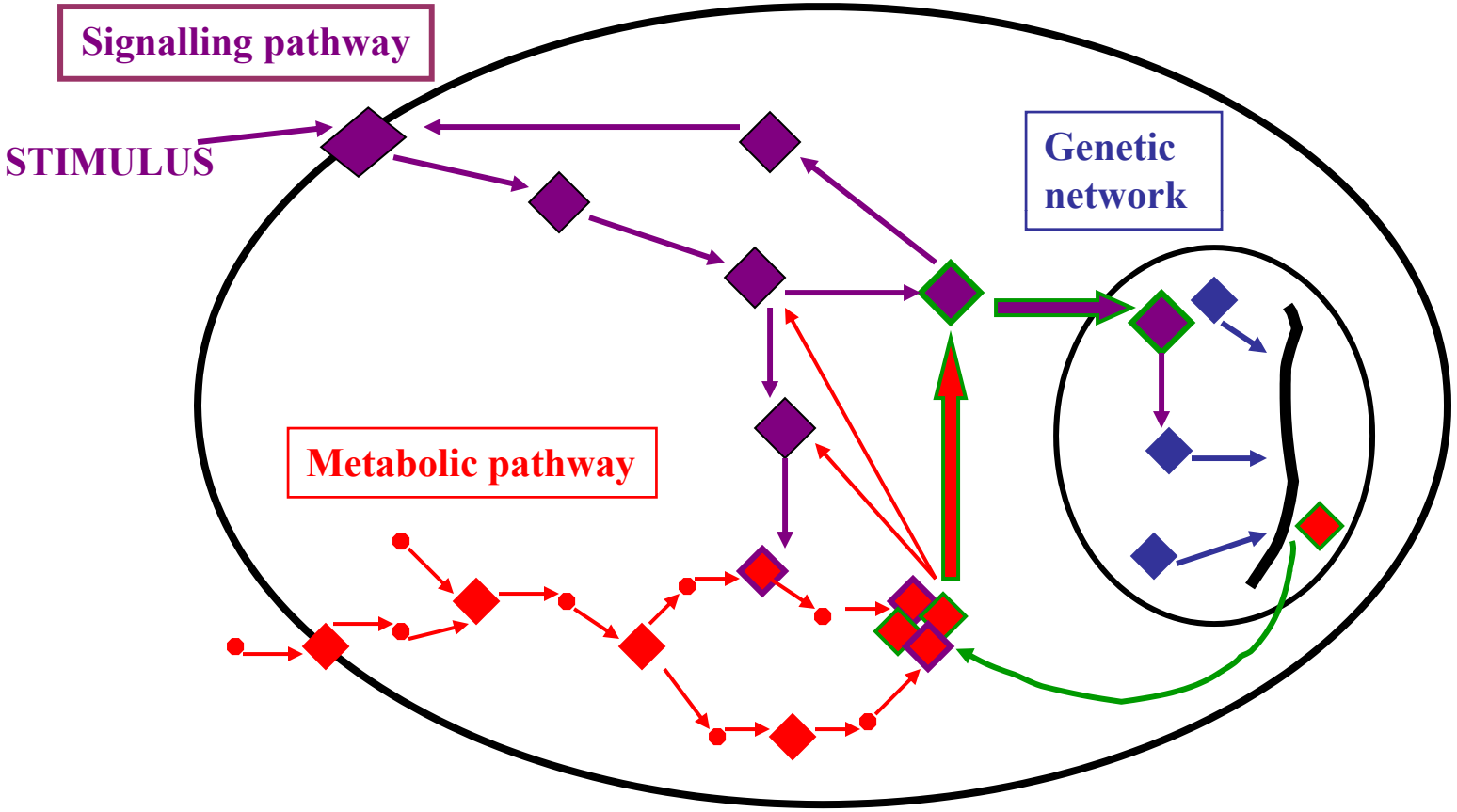
Biological networks

Construction and Analysis

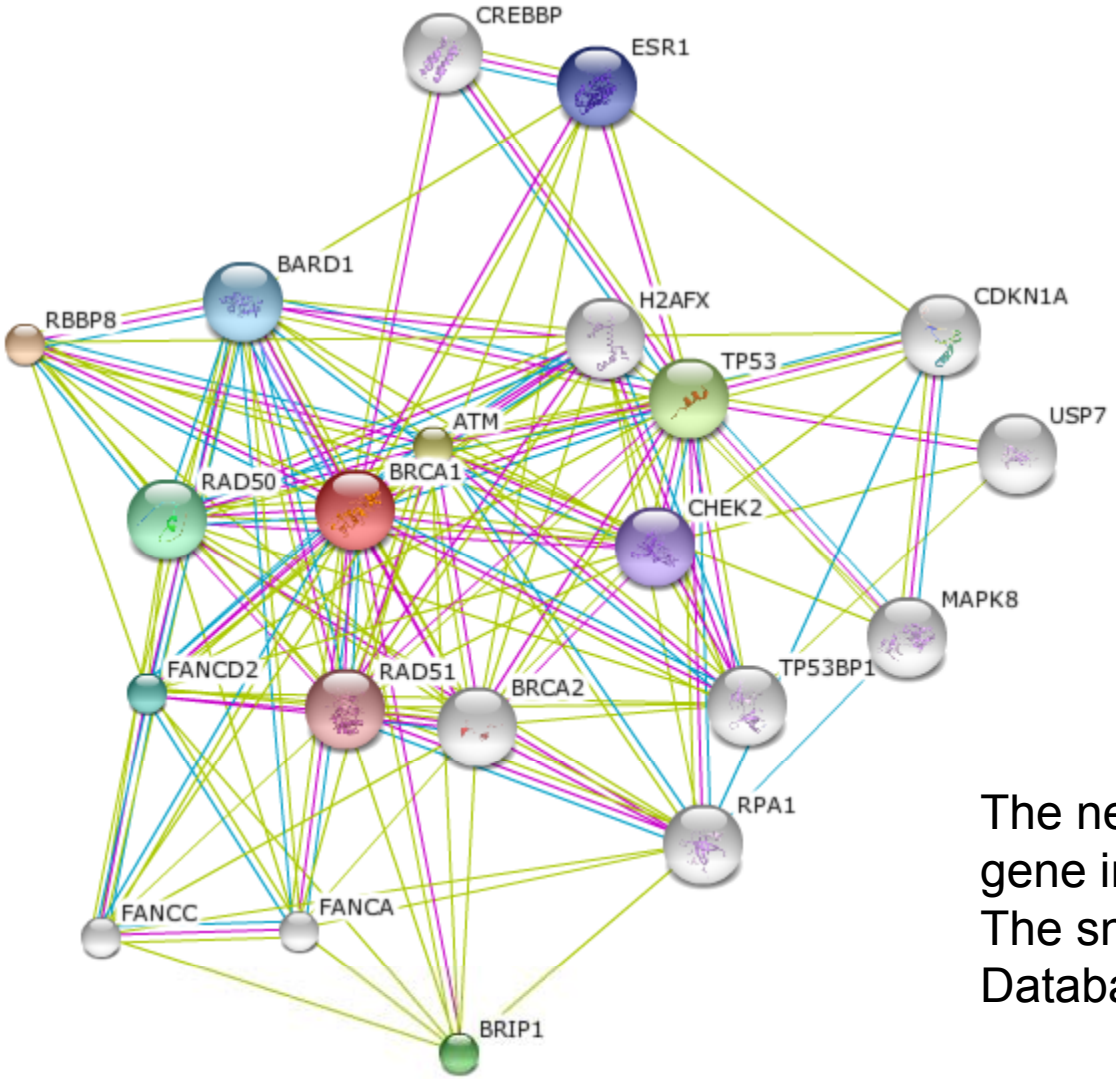
Interactions in a cell



Interactions → Pathways → Network

- A collection of interactions defines a network
- Pathways are subsets of networks
  - All pathways are networks of interactions, however not all networks are pathways!
  - Difference in the level of annotation or understanding
- We can define a pathway as a biological network that relates to a **known** physiological process or complete function

A biological network

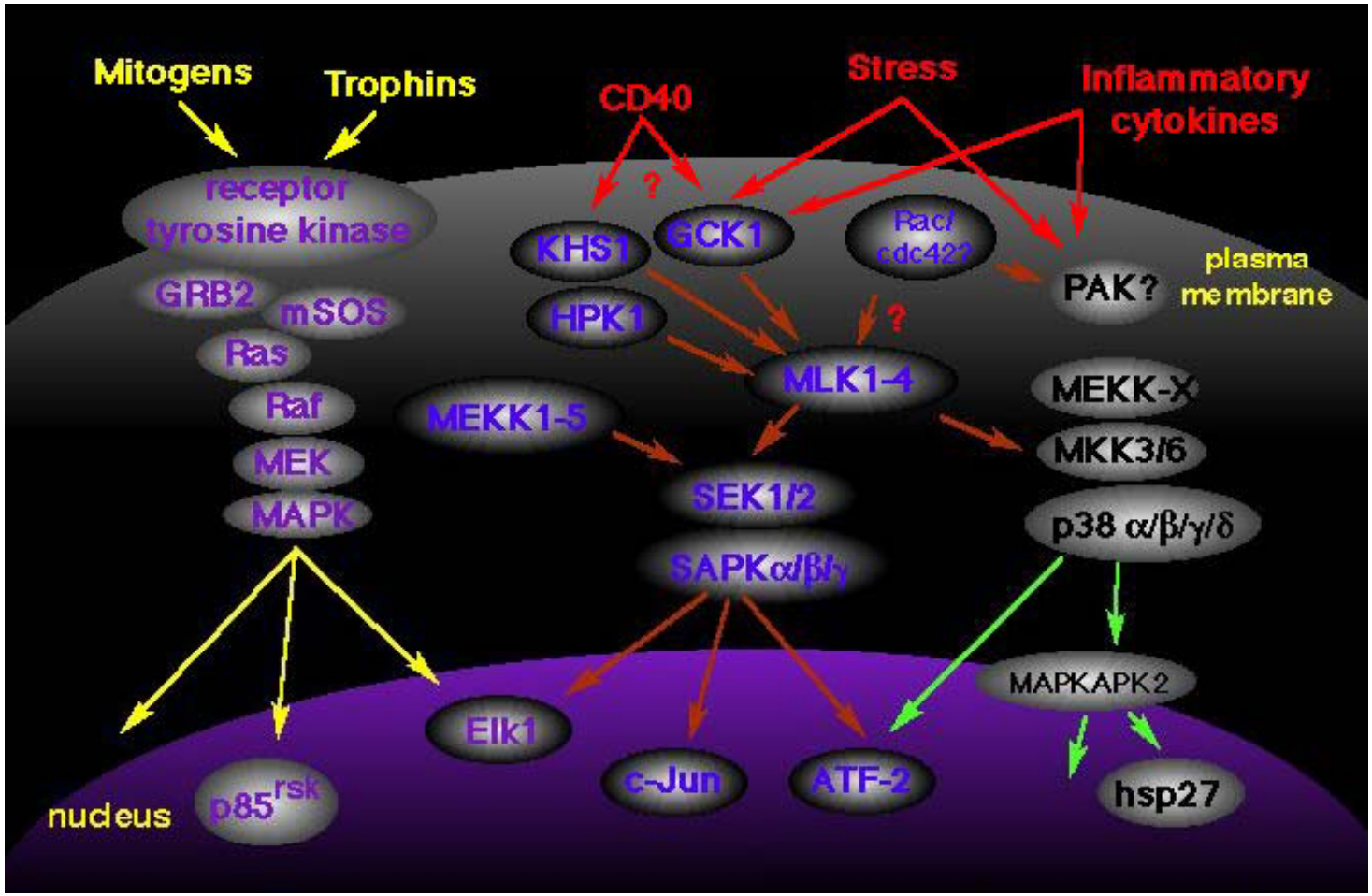


The network around the BRCA1 gene in human.
The snapshot is from the STRING Database at string.embl.de

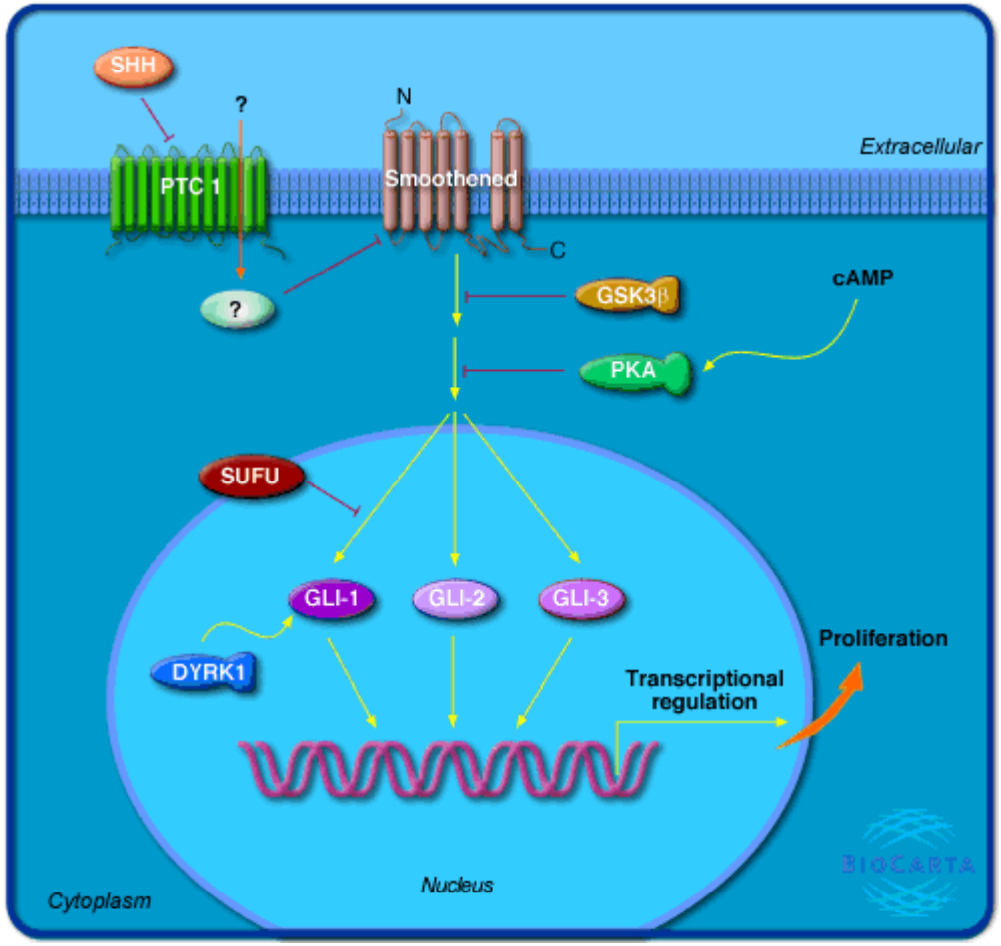
Types of protein interactions

- Metabolic and signaling (genetic) pathways
- Morphogenic pathways in which groups of proteins participate in the same cellular function during a developmental process
- Structural complexes and molecular machines in which numerous proteins are brought together

Signaling pathways

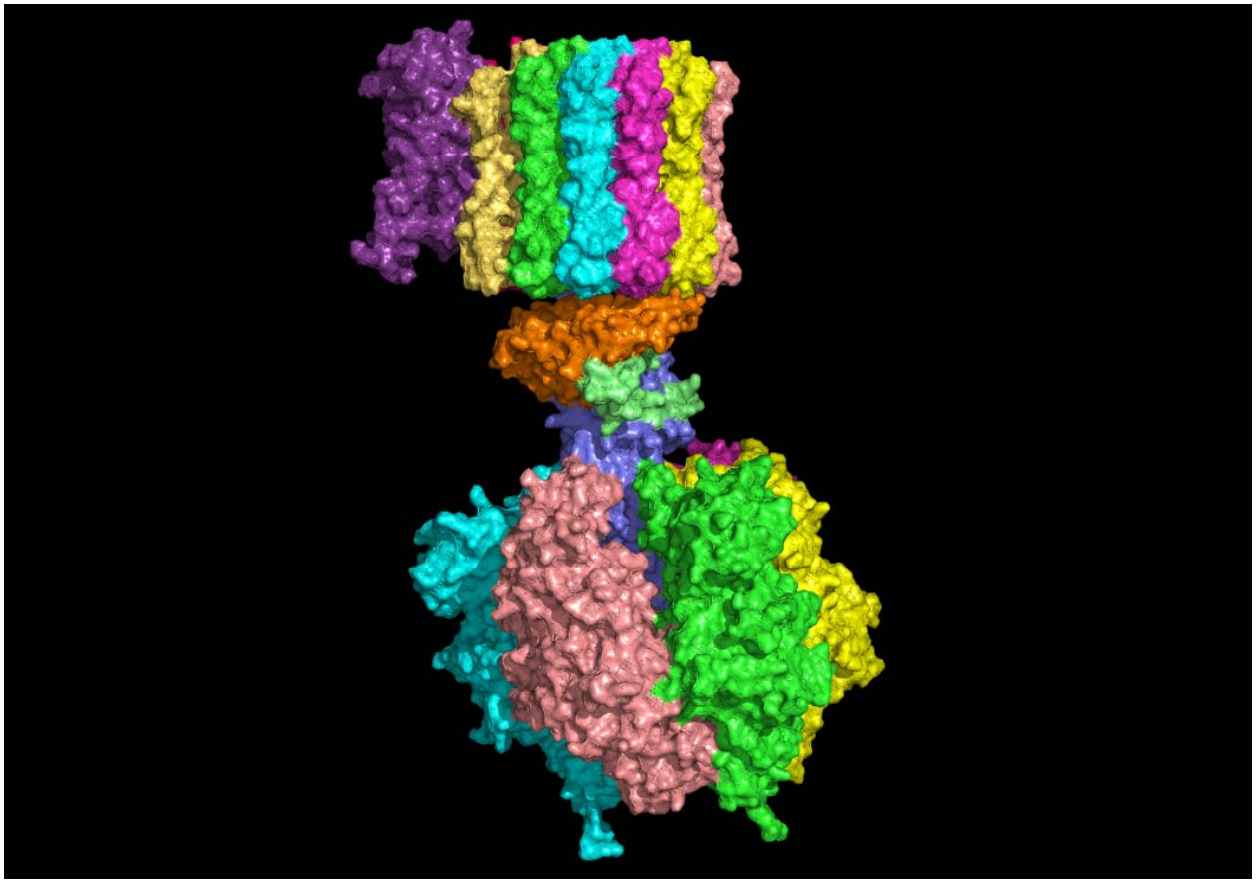


Morphogenic pathways



Structural complexes and molecular machines

ATPase



Experimental methods

- Tagged Fusion Proteins
- Coimmunoprecipitation
- Yeast Two-hybrid
- Biacore
- Atomic Force Microscopy (AFM)
- Fluorescence Resonace Energy Trasfer (FRET)
- X-ray Diffraction

Where is the data?

- Results of high-throughput experiments are usually collected in databases
- What about low-throughput experiments?

The literature

- Thousands of small scale, low throughput experiments performed in labs worldwide for years
  - The results are published as articles
- So we can collect this information to get individual data about pairs of proteins/genes
- What is the difficulty?

Text mining

- Hundreds of thousands of unstructured free text articles should be processed automatically to extract this information
- Challenges
  - Non standard naming of genes, proteins, processes
  - Understanding natural language
- Concerns
  - Accuracy?
  - Coverage?

BioCreative Challenge

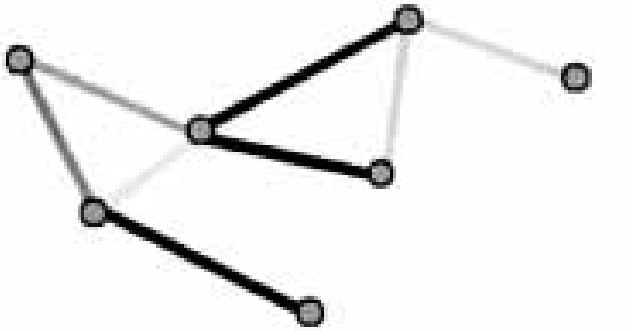
- A competition of algorithms for text mining
- Problems
  - Identify whether an article contains the relevant information or not
  - Extract the information

What else can we do?

- Computational prediction of relationships between pairs of genes/proteins
- Data sources for prediction
  - Sequence data
  - Genome data:
    - Interologs
    - Existence of genes in multiple organisms
    - Locations of the genes
  - Bio-image data
  - Gene Ontology annotations
  - Microarray experiments
  - Sub-cellular localization data

Probabilistic network approach

- Each “interaction” link between two proteins has a posterior probability of existence, based on the quality of supporting evidence.



Computing the posterior

- Using Bayes' rule and with naïve Bayes assumption that different evidence types are independent of one another given the truth about interaction:

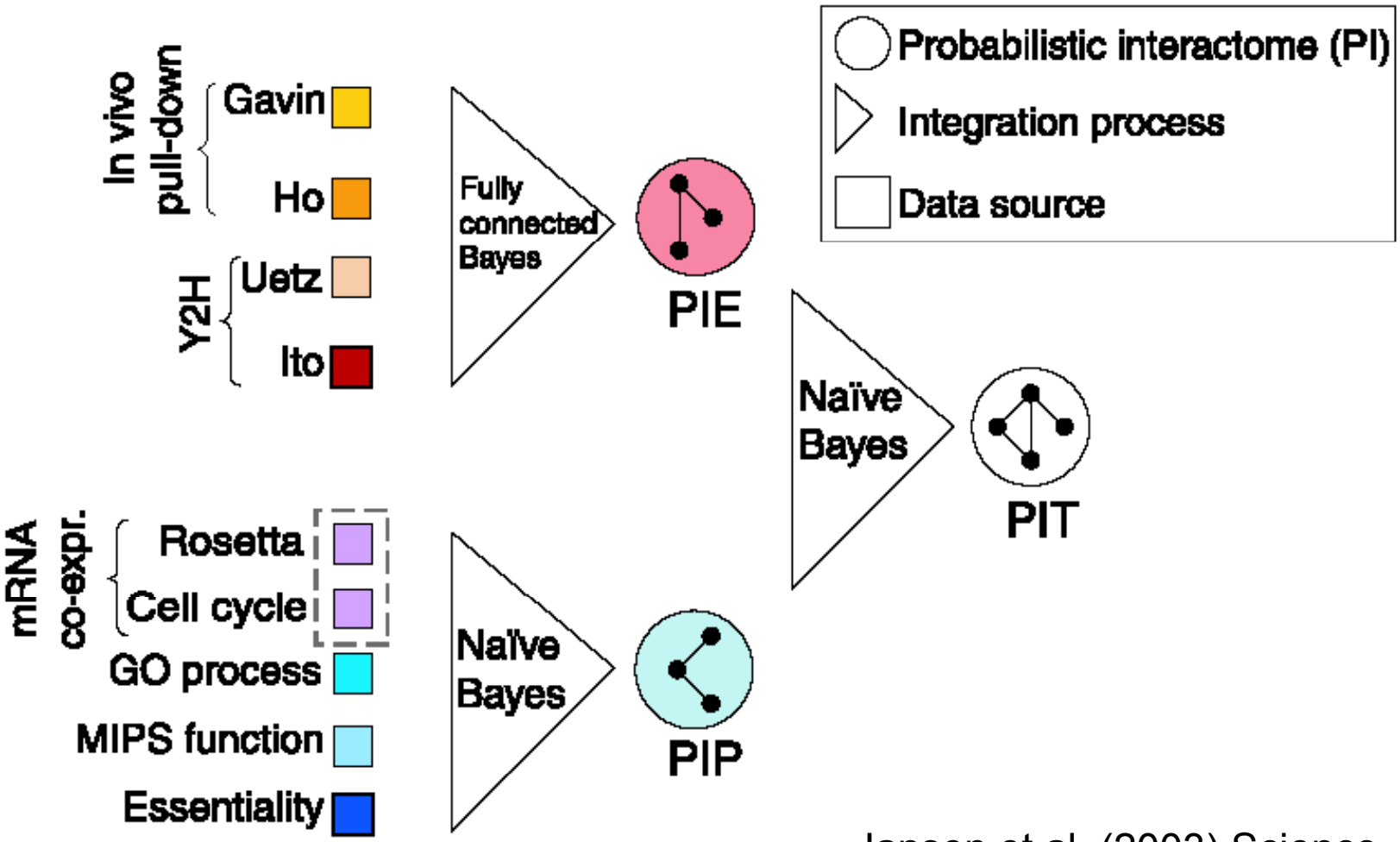
$$p(y=1 \mid \mathbf{z}) = \frac{\left(\prod_{i=1}^{T} p(z_i \mid y=1)\right) \cdot p(y=1)}{\sum_{j \in 0,1}\left(\left(\prod_{i=1}^{T} p(z_i \mid y=j)\right) \cdot p(y=j)\right)}$$

Asthana *et al. Genome Research*, **13**:1170:1174 (2004)

Bayesian Network approach

- Jansen *et al.* (2003) *Science*. Lee *et al.* (2004) *Science*.
- Combine individual probabilities of likelihood computed for each data source into a single likelihood (or probability)
- Naïve Bayes:
  - Assume independence of data sources
  - Combine likelihoods using simple multiplication

Bayesian Network approach



Jansen et al. (2003) Science

Bayesian Approach

- A scalar score for a pair of genes is computed separately for each information source.
- Using gold positives (known interacting pairs) and gold negatives (known non-interacting pairs) interaction likelihoods for each information source is computed.
- The product of likelihoods can be used to combine multiple information sources
  - Assumption: A score from a source is independent from a score from another source.

Naïve Bayes vs. Fully Connected Bayes

- In Naïve Bayes approach we can find the correlation of each data source with the gold standards separately and then compute the combined likelihood of a protein pair by just multiplying the individual likelihoods.

$$L(f_1 \cdots f_N) = \prod_{i=1}^{N} L(f_i) = \prod_{i=1}^{N} \frac{P(f_i \mid pos)}{P(f_i \mid neg)}$$

Computing the likelihoods

- Partition the pair scores of an information source into bins and provide likelihoods for score-ranges
- E.g. Using the microarray information source and using Pearson correlation for scoring protein pairs you may get scores between -1 and 1. You want to know what is the likelihood of interaction for a protein pair that gets a Pearson correlation of 0.6.

Partitioning the scores

pearson corr.	likelihood
(0.8,1.0]	
(0.6,0.8]	
(0.4,0.6]	
(0.2,0.4]	
(0.0,0.2]	
(-0.2,0.0]	
(-0.4,-0.2]	
(-0.6,-0.4]	
(-0.8,-0.6]	
[-1.0,-0.8]	

Computing the likelihood

- $$L = \frac{P(\text{Score} \mid \text{Interaction}) / P(\text{Interaction})}{P(\text{Score} \mid \sim\text{Interaction}) / P(\sim\text{Interaction})}$$

- Example

Example

- Calculating the likelihood ratio for expression dataset.

Expression correlation		# protein pairs	Gold standard overlap					P(exp|pos)	P(exp|neg)	L
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	0.9	678	16	45	16	45	0.36	2.10E-03	1.68E-05	124.9
	0.8	4,827	137	563	153	608	0.25	1.80E-02	2.10E-04	85.5
	0.7	17,626	530	2,117	683	2,725	0.25	6.96E-02	7.91E-04	88.0
	0.6	42,815	1,073	5,597	1,756	8,322	0.21	1.41E-01	2.09E-03	67.4
	0.5	96,650	1,089	14,459	2,845	22,781	0.12	1.43E-01	5.40E-03	26.5
	0.4	225,712	993	35,350	3,838	58,131	0.07	1.30E-01	1.32E-02	9.9
	0.3	529,268	1,028	83,483	4,866	141,614	0.03	1.35E-01	3.12E-02	4.3
	0.2	1,200,331	870	183,356	5,736	324,970	0.02	1.14E-01	6.85E-02	1.7
	0.1	2,575,103	739	368,469	6,475	693,439	0.01	9.71E-02	1.38E-01	0.7
	0	9,363,627	894	1,244,477	7,369	1,937,916	0.00	1.17E-01	4.65E-01	0.3
	-0.1	2,753,735	164	408,562	7,533	2,346,478	0.00	2.15E-02	1.53E-01	0.1
	-0.2	1,241,907	63	203,663	7,596	2,550,141	0.00	8.27E-03	7.61E-02	0.1
	-0.3	484,524	13	84,957	7,609	2,635,098	0.00	1.71E-03	3.18E-02	0.1
	-0.4	160,234	3	28,870	7,612	2,663,968	0.00	3.94E-04	1.08E-02	0.0
	-0.5	48,852	2	8,091	7,614	2,672,059	0.00	2.63E-04	3.02E-03	0.1
	-0.6	17,423	-	2,134	7,614	2,674,193	0.00	0.00E+00	7.98E-04	0.0
	-0.7	7,602	-	807	7,614	2,675,000	0.00	0.00E+00	3.02E-04	0.0
	-0.8	2,147	-	261	7,614	2,675,261	0.00	0.00E+00	9.76E-05	0.0
	-0.9	67	-	12	7,614	2,675,273	0.00	0.00E+00	4.49E-06	0.0
Sum		18,773,128	7,614	2,675,273	-	-	-	1.00E+00	1.00E+00	1.0

Example

- Calculating the likelihood ratio for the Biological Process (GO) dataset.

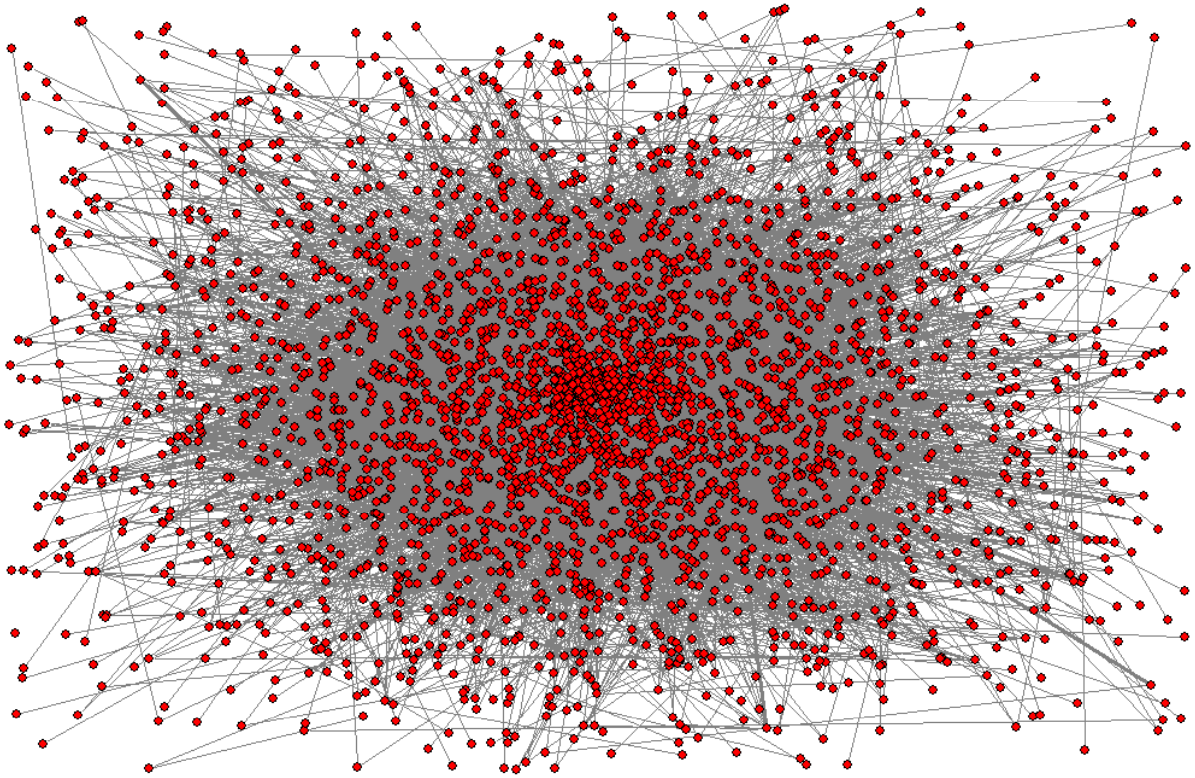
GO biological process similarity		# protein pairs	Gold standard overlap					P(GO|pos)	P(GO|neg)	L
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	1 -- 9	4,789	88	819	88	819	0.11	1.17E-02	1.27E-03	9.2
	10 -- 99	20,467	555	3,315	643	4,134	0.16	7.38E-02	5.14E-03	14.4
	100 -- 1000	58,738	523	10,232	1,166	14,366	0.08	6.95E-02	1.59E-02	4.4
	1000 -- 10000	152,850	1,003	28,225	2,169	42,591	0.05	1.33E-01	4.38E-02	3.0
	10000 -- Inf	2,909,442	5,351	602,434	7,520	645,025	0.01	7.12E-01	9.34E-01	0.8
Sum		3,146,286	7,520	645,025	-	-	-	1.00E+00	1.00E+00	1.0

- Given a pair of proteins with microarray Pearson correlation 0.65 and GO biological process similarity 2500, what is the likelihood of interaction?

  67.4*3.0 = 202.2

Protein interaction networks

- Large scale (genome wide networks):



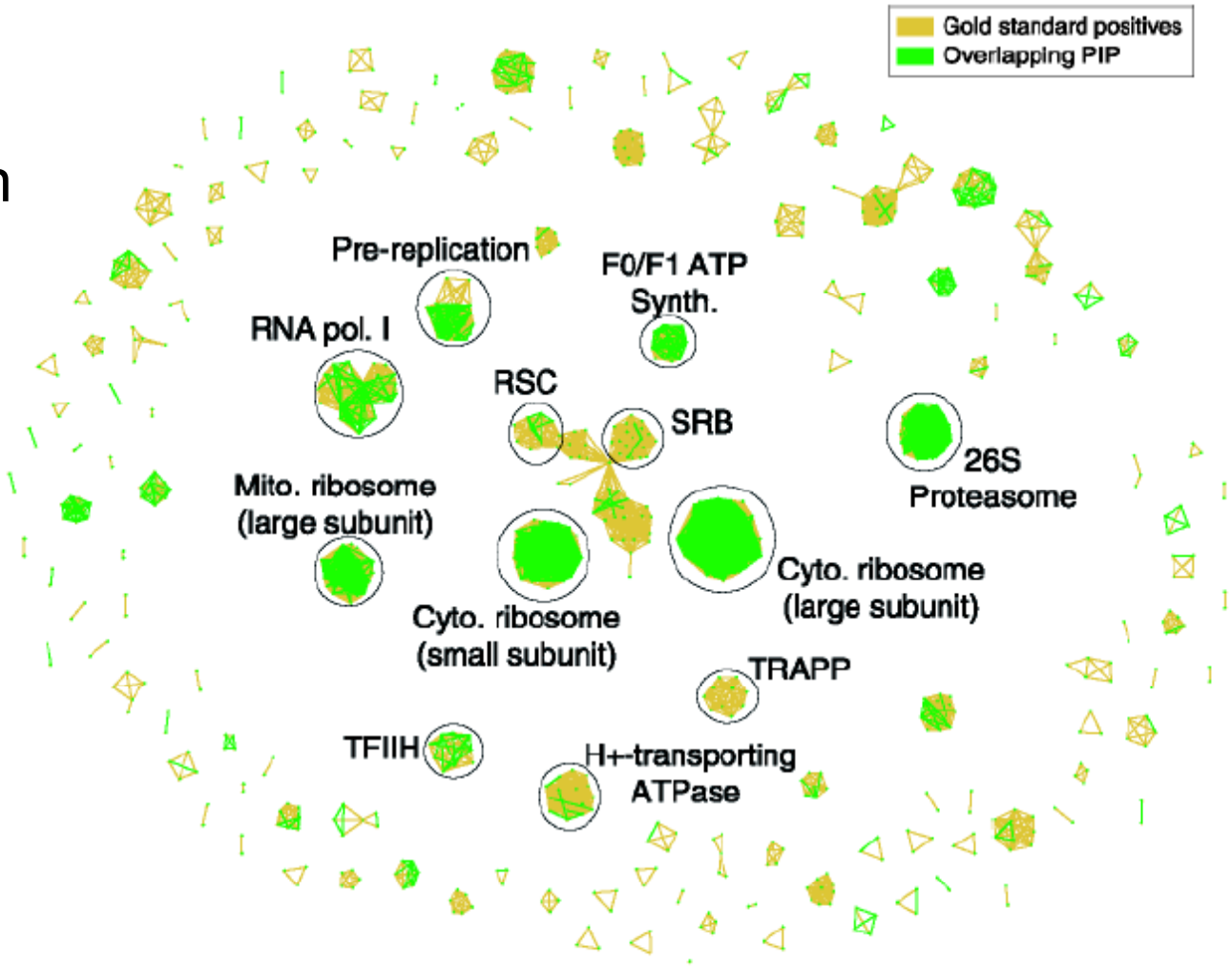
ProNet (Asthana et al.)
Yeast
3,112 nodes
12,594 edges

Analyzing Protein Networks

- Predict members of a partially known protein complex/pathway.
- Infer individual genes' functions on the basis of linked neighbors.
- Find strongly connected components, clusters to reveal unknown complexes.
- Find the best interaction path between a source and a target gene.

Simple analysis

The network can be ***thresholded*** to reveal clusters of interacting proteins



Advanced Analysis

- Clustering algorithms
  - MCL (Markov CLustering)
  - RNSC (Restricted Neighborhood Search Clustering)
  - SPC (Super Paramagnetic Clustering)
  - MCODE (Molecular COmplex DEtection)
  - and many more
  - “Evaluation of clustering algorithms for protein-protein interaction networks,” by Brohee and van Helden in BMC Bioinformatics, November 2006.

Markov Cluster Algorithm

- Simulates a flow on the graph.
- Calculates successive powers of the adjacency matrix
- Parameters
  - One parameter: *inflation parameter*
- The process partitions the graph (i.e., no overlapping clusters)
- The inflation parameter influence the number of clusters generated

Restricted Neighborhood Search Clustering

- Starts with an initial random clustering
- Tries to minimize a cost function by iteratively moving vertices between neighboring clusters.
- Parameters:
  - Number of iterations
  - Diversification frequency
  - …. and 5 other parameters

Super Paramagnetic Clustering

- Hierarchical algorithm inspired from an analogy with the physical properties of a ferromagnetic model subject to fluctuation at nonzero temperature.
- Parameters:
  - Number of nearest neighbors
  - Temperature

MCODE

- Weight each vertex by its local neighborhood density (using a modified version of clustering coefficient using k-cores)
- Starting from the top weighted vertex, include neighborhood vertices with similar weights to the cluster
- Post-process to remove or add new vertices
- Continue with the next highest weight vertex in the network
- May provide overlapping clusters

Vertex weighting

- Clustering coefficient

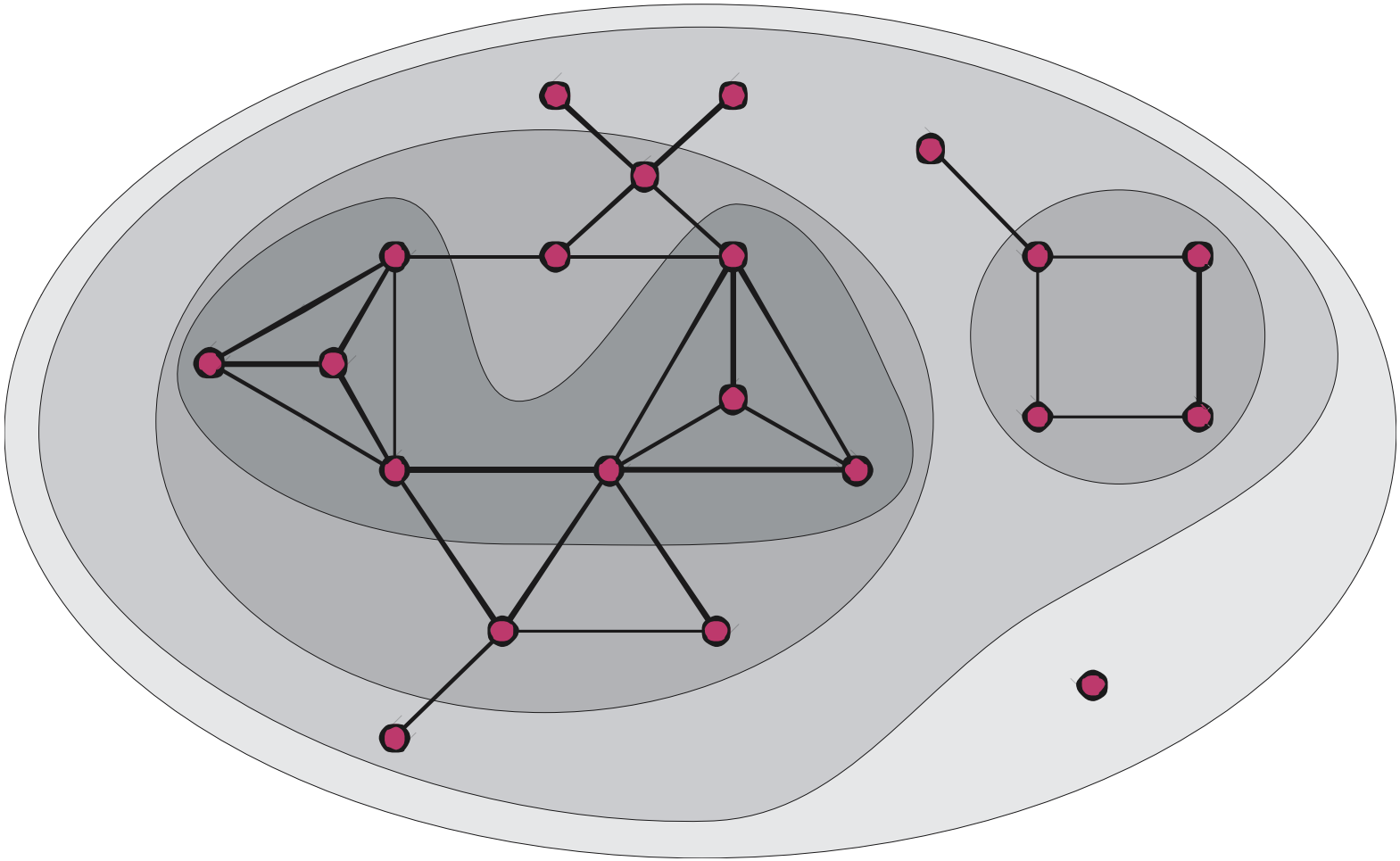
$$CC_i = \frac{2e_i}{d_i(d_i - 1)}$$

where e_i is the number of edges between the neighbors of node i and d_i is the number of neighbors of node i.

k-core

- A part of a graph where every node is connected to other nodes with at least k edges (k=0,1,2,3...)
- Finding a k-core in a graph proceeds by progressively removing vertices of degree < k until all remaining vertices are connected to each other by degree k or more. Complexity: $O(n^2)$. The highest k-core is found by trying to find k-cores from one up until the highest degree in the neighborhood graph. Overall complexity: $O(n^3)$

k-core example



Core-clustering Coefficient

- Product of the clustering coefficient of the highest k-core in the neighborhood of a vertex and *k*.

Features of the algorithms

	Restricted Neighborhood Search Clustering (RNSC)	Markov Clustering (MCL)	Molecular Complex Detection (MCODE)	Super-paramagnetic clustering (SPC)
Type	Local search cost based	Flow simulation	Local neighbourhood density search	Hierarchical
Allow multiple assignations	No	No	Yes	No
Allow unassigned nodes	No	No	Yes	No
Edge-weighted graphs supported	No	Yes	No	Yes
First application	Protein complex prediction	Protein family detection	Protein complex detection	
Other applications	/	Identification of ortholog groups, protein complexes, peer-to-peer node clustering, image retrieval, Word Sense Discrimination, molecular pathway discovery, structural domains, ...	/	Image clustering, microarray data clustering, protein complexes detection, protein structure classification, identification of ortholog groups, ...
Availability	Upon request	http://micans.org/mcl/	ftp://ftp.blueprint.org/pub/BIND/README	Upon request
Developper	King AD	Van Dongen S	Bader GD and Hogue CWV	Blatt M, Wiseman S, Domany E
References	[21]	[35]	[19]	[18]