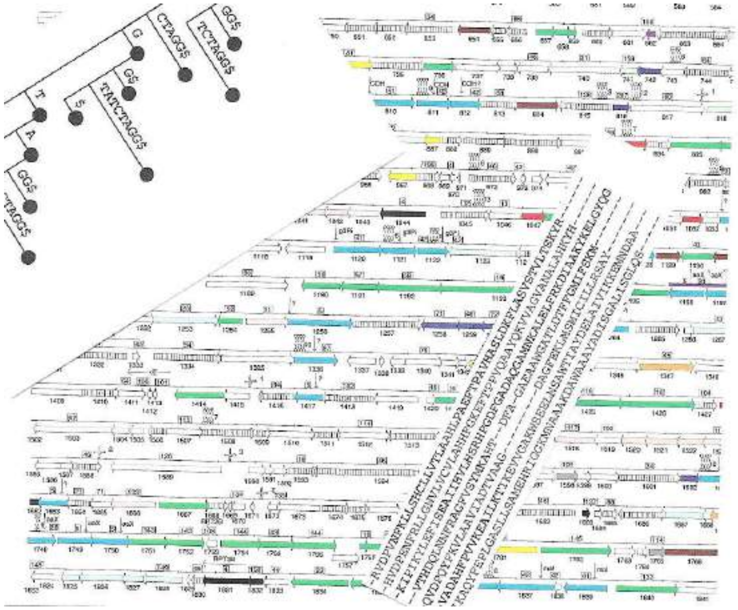
Lecture 10 : Biological Networks – Part I

Principles of Computational Biology

Instructor: Teresa Przytycka, PhD



System Biology

- Integrative approaches in which scientists study and model pathways and networks, with an interplay between experiment and theory
- Integrate biological data as an attempt to understand how biological systems function.
- Study the relationships and interactions between various parts of a biological system
- Study how they connect, infer knowledge using mixed type of data
- Develop a model of the whole system
- Model and predict the behavior of a system upon perturbation.

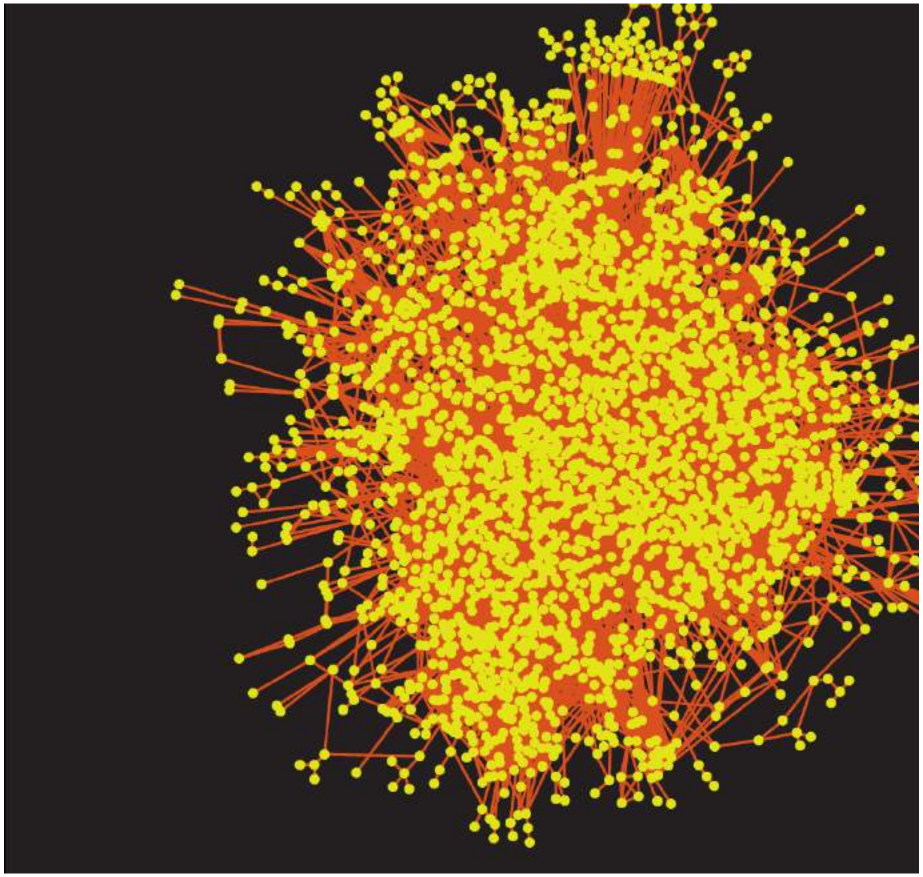
Why networks?

- Networks provide natural description of relation between various components
- Examples:
  - Protein–protein interaction network
  - Protein domain co-occurrence network
  - Metabolic networks
  - Transcription Networks

Networks

- Vertices: elementary units; edges: binary relations between such units
- Example: Protein – protein interaction network:
  - Nodes – proteins
  - Edges – interactions

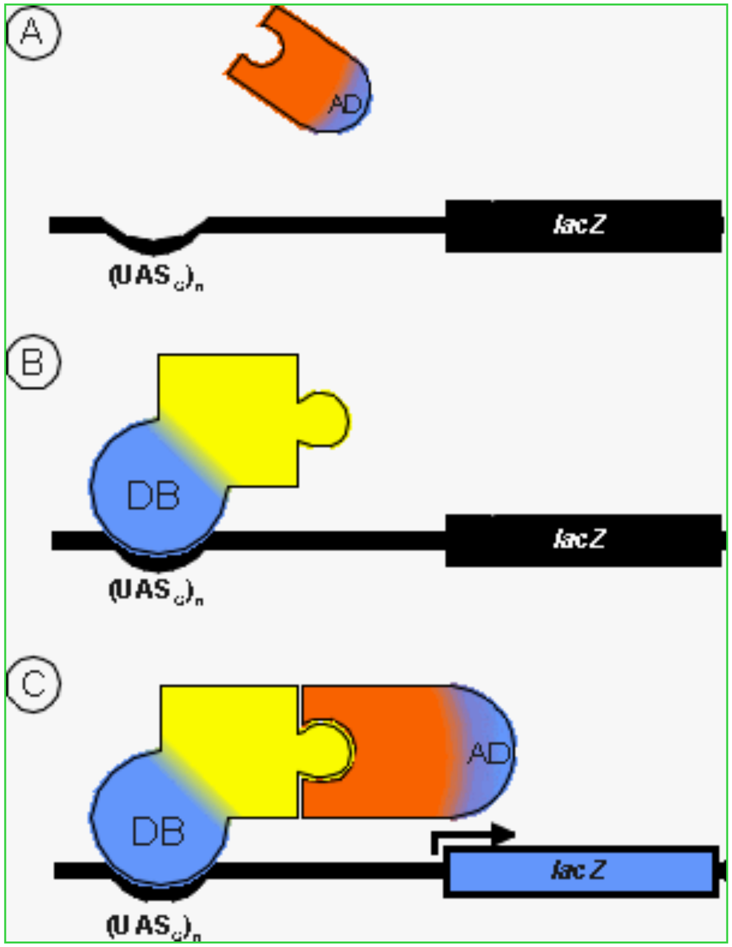
(left yaest PPI)



How do we know that a pair of proteins interact?

- A complex containing these two proteins have been crystallized.
- High throughput interaction screening methods:
  - Yeast two hybrid experiments (Y2H)
  - Protein complex purification (PCP)
- Problem with high throughput method:
  - significant amount of false positives and false negatives

Y2H

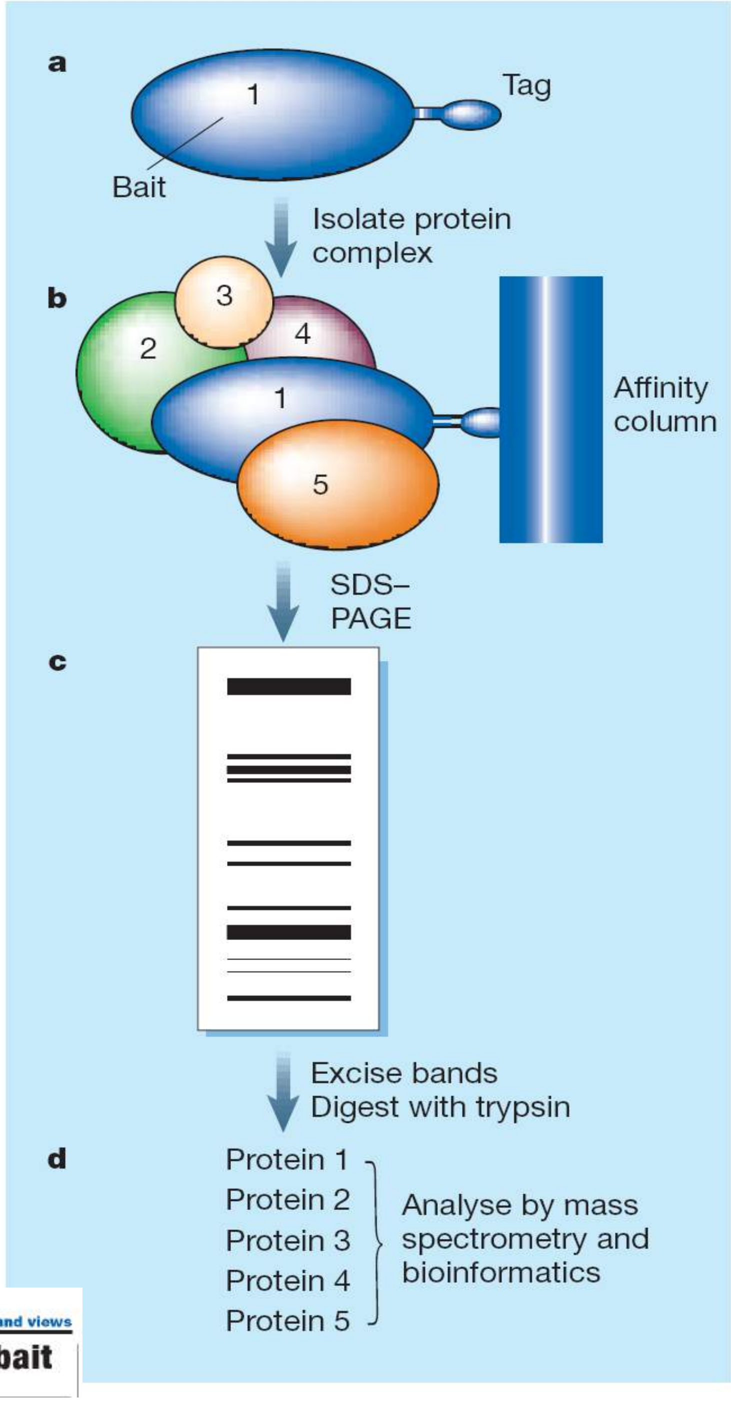


Principle of the Two-hybrid system. (A), (B) Two chimeras, one containing the DNA-binding domain (DB: blue circle) and one that contains an activation domain (AD: half blue circle), are co-transfected into an appropriate host strain. (C) If the fusion partners (yellow and red) interact, the DB and AD are brought into proximity and can activate transcription of reporter genes (here *LacZ*).

From *Yeast Two-Hybrid: State of the Art* *Wim Van Criekinge*[1*] *and Rudi Beyaert*[2]*; http://www.biologicalprocedures.com/bpo/arts/1/16/m16f1lg.htm*

CPC

- Take a set of proteins “baits”
- Expose each “bait” protein so to a set of “pray” proteins that potentially can form complexes with it.
- Allow the complexes to form
- Identify proteins in each complex



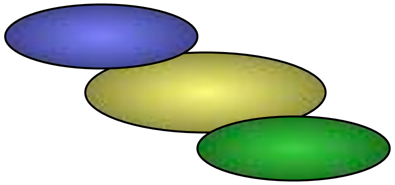
news and views

Protein complexes take the bait

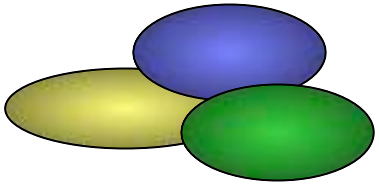
Anuj Kumar and Michael Snyder

Caveats

- For CPC – we don’ t detailed information about contacts e.g.



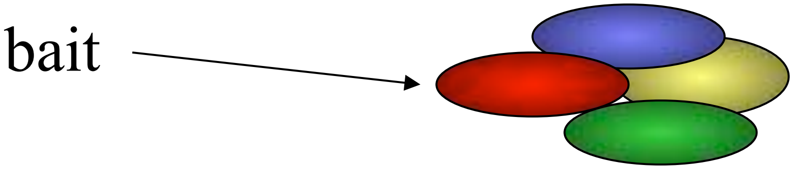
or



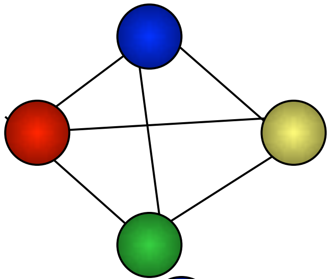
?

- If the second configuration is correct Y2H might never get it (it might require all three to proteins to make a complex)
- Y2H test id the interaction CAN occur nit if it DOES occur (we need to have both proteins at the same time at the same place – Y2H enforces it)

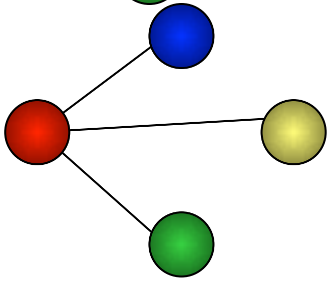
Representing CPC data as graph



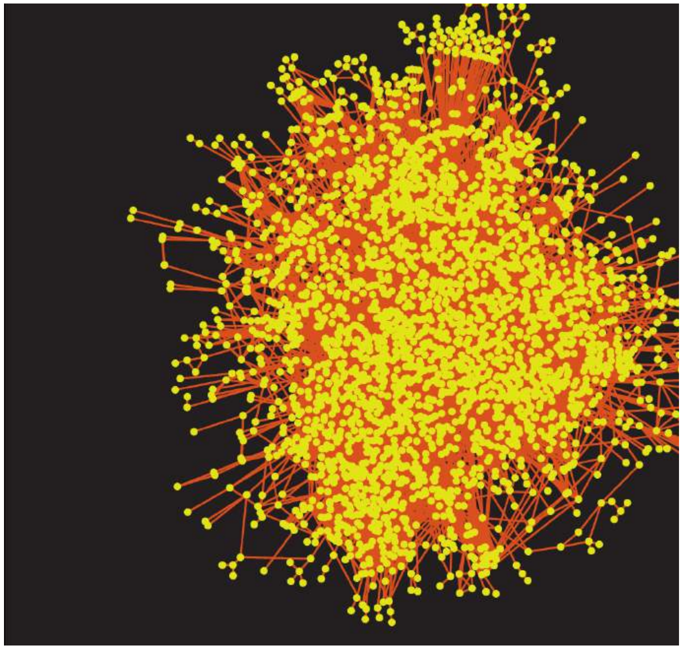
Clique model:



Spike model:



So what we make out of this hairball?



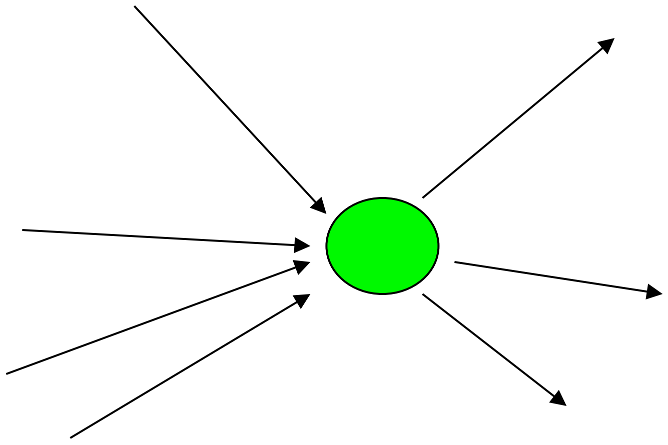
- Can we discover basic principles of its organization or is it random?
- We have concept of sequence evolution, how about network evolution?

Properties of a net

- – Vertex degree distribution
- – Distribution of sizes of connected components
- – Clustering coefficient
- – “Betweeness”
- – Centrality
- – Diameter

Vertex degree distribution

What is probability that a randomly selected vertex has degree k?



k = 7

$k_i = 4$

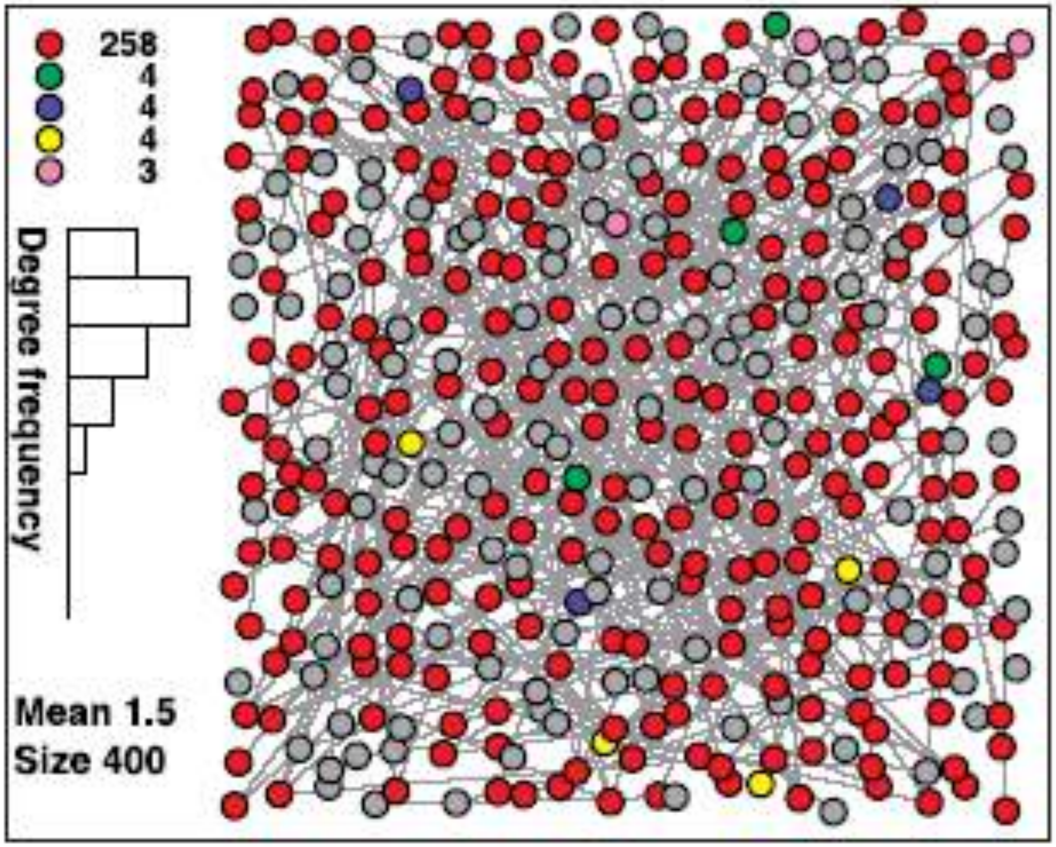
$k_o = 3$

- If edges of the net are not-directed, the degree k of a vertex v is the number of adjacent edges.
- Otherwise we additionally consider k_i and k_o (in-degree and out-degree)

Model 1: Erdos-Renyi model

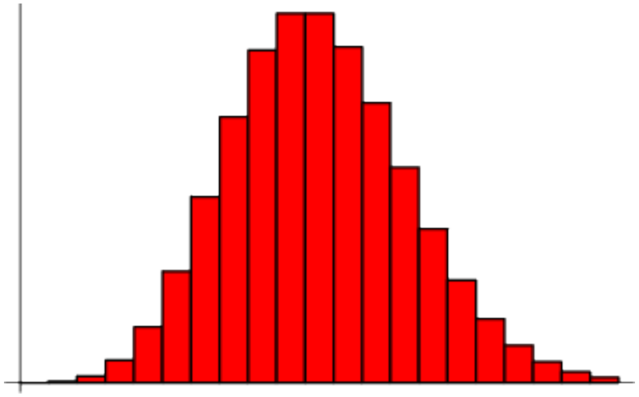
- Erdos-Renyi model:
  - With probability *p* put an edge between any pair of vertices.

http://www.ams.org/notices/200404/comm-networks.pdf



The degree distribution in Erdos-Renyi model: Poisson

$$p(k) = \frac{e^{-\bar{k}}\,\bar{k}^{k}}{k!}$$



P(k) – probability of node having degree k

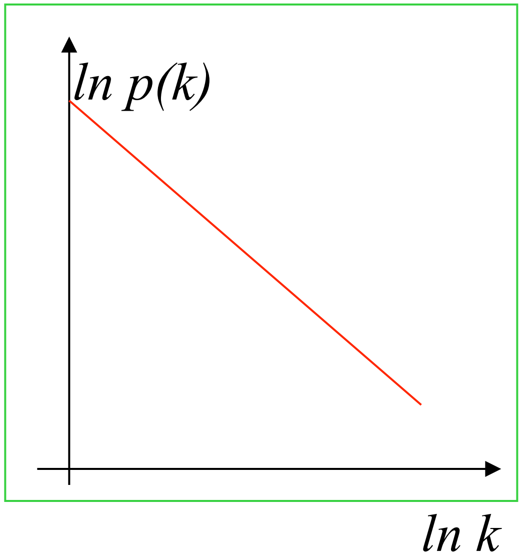
$\bar{k} = \Sigma_k k\, p(k)$ (*average degree*)

This distribution is approached in ER model as #nodes goes to infinity under assumption that $\bar{k}$ is fixed

Vertex degree distribution for biological (and many other “real world” networks) better approximated be is Scale free distribution

$$P(k) \sim k^{-\gamma}$$

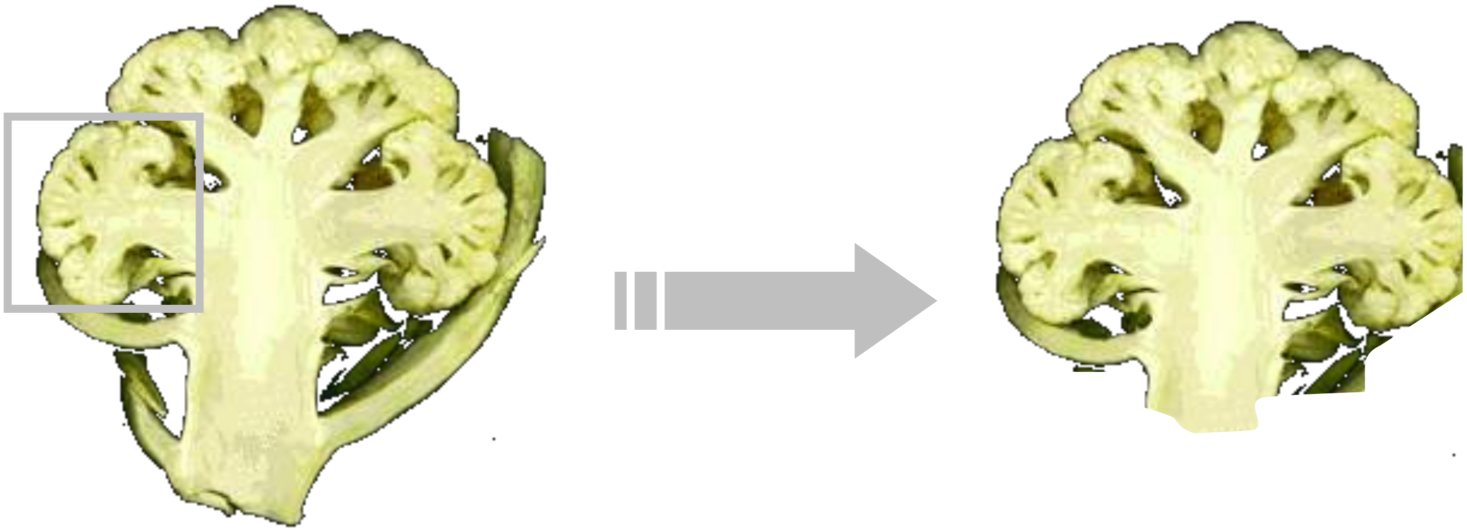
P(k) probability of node of degree k



Power Law (scale free) distribution

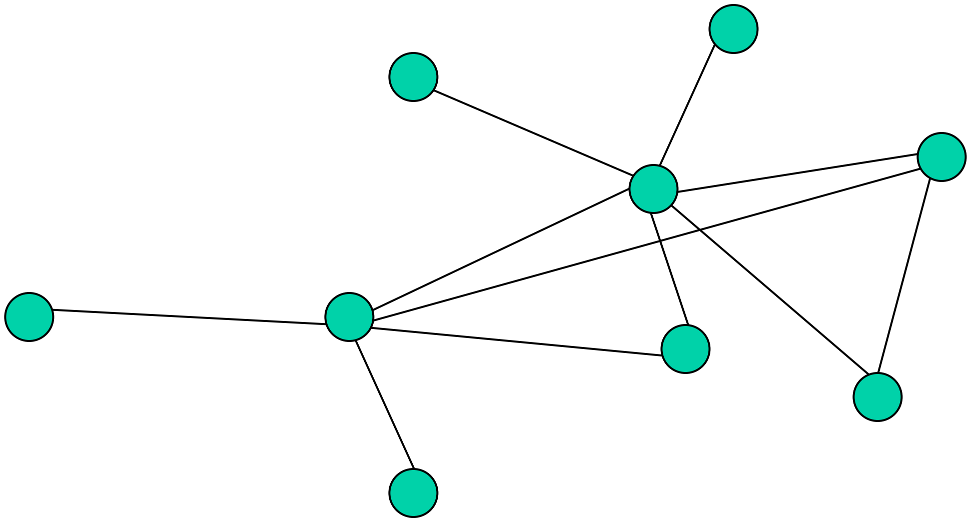
$$P(k) \sim k^{-\gamma} \Longleftrightarrow \boxed{\frac{p(k)}{p(k')} = \frac{p(\alpha\, k)}{p(\alpha\, k')}}$$

Thus no natural scale



Example of a scale free model: Preferential attachment Barabasi-Albert

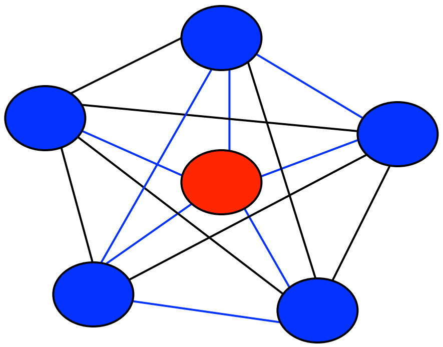
- At each step, a new vertex is added to the graph
- The new vertex is attached to one of old vertices with probability proportional to the degree of that old vertex.



Connectivity/clustering coef.

- Characterizes "density" of connections
- For each node, consider its neighbors and see what percentage of possible connections between them are realized

z = #neighbors ;
y = #connections between the neighbors
C=2y/z(z-1)



black – potential connections
blue – existing connection
connectivity in red node is

C= 3/10

$\bar{C}$ = average connectivity = prob. there exists a connection between two neighbors

For ER model $\bar{C} = \bar{k}/N$

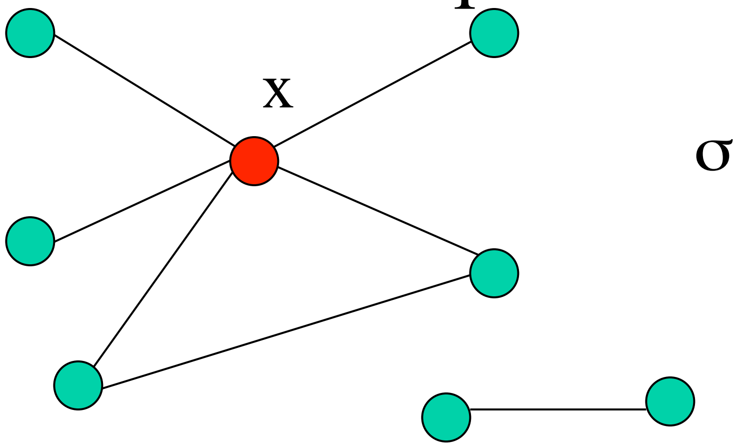
Application

Hierarchical organization of Modularity in Metabolic Networks

Ravasz, Somera, Mongru, Oltavi, Barabasi (Science 2002)

Betweenness

- The total number of shortest paths that pass trough a vertex x (normalized by #pairs in the connected component containing x).
- Indicates if vertex is important for the traffic



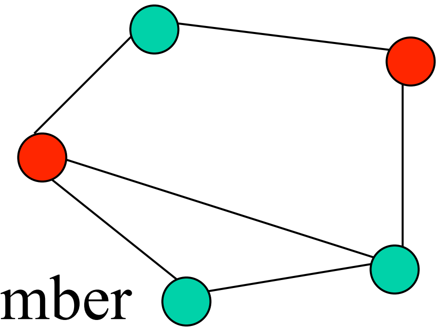
$\sigma(x) = 14/15$

$\sigma(x) = 14/15$

- Also called betweenness centrality

Small worlds

- Distance between two nodes in a net = the smallest number of steps one can take to reach on node from the other



The distance between red nodes = 2

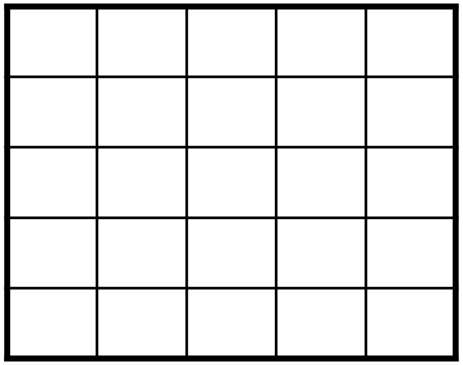
Example: Erdos number

l_{max} = longest shortest path = diameter of the network

Small worlds- cont.

p(l) = the probability that the shortest path between two random nodes is l

$\bar{l} = \Sigma_l \, lp(l)$



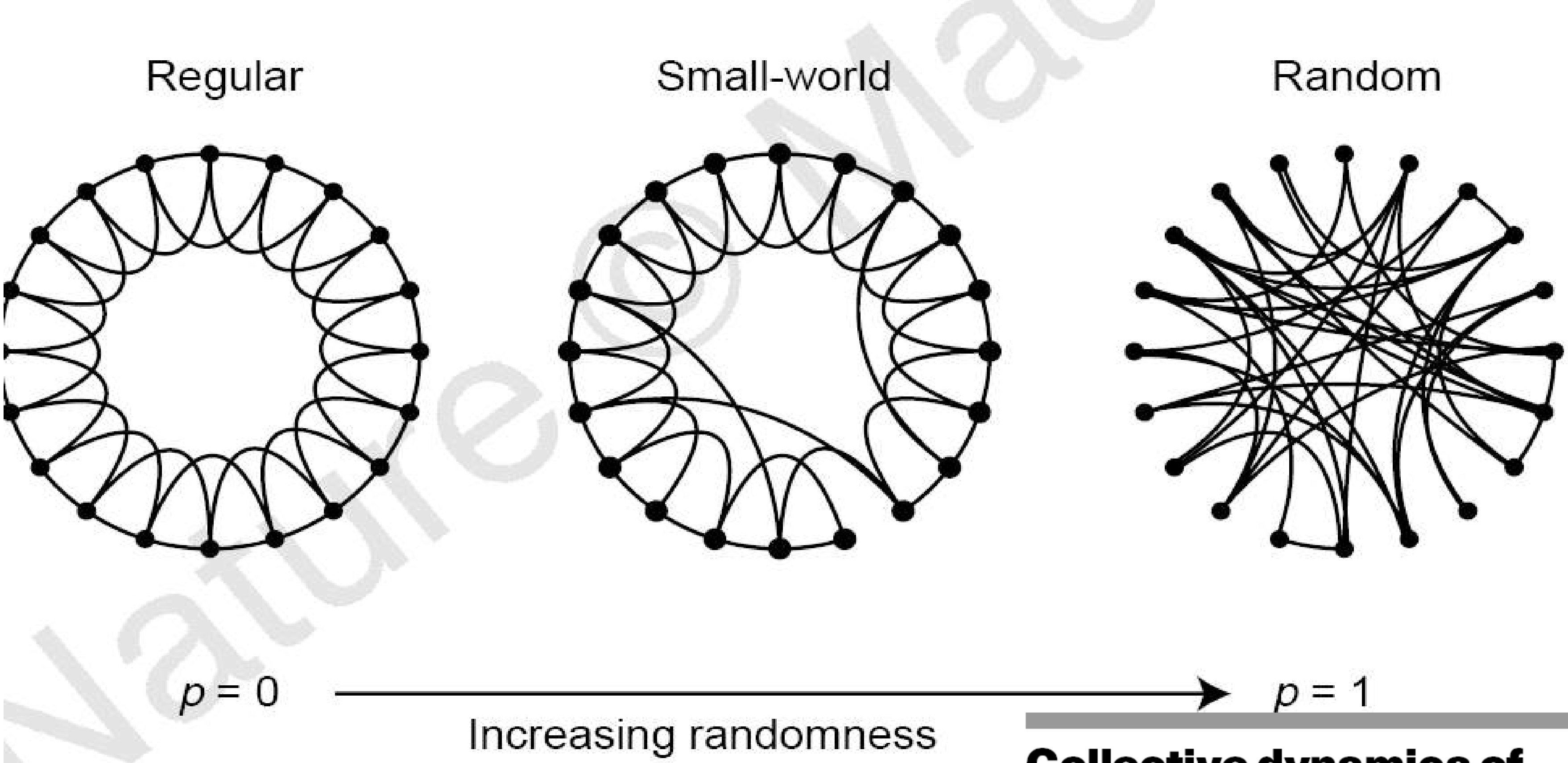
In a grid $\bar{l} \sim sqrt(N)$

In a tree $\bar{l} \sim \ln N / \ln \bar{z}$ *where* $\bar{z}$ *– average degree*

The fact that $\bar{l}$ *that tends to be small (relative to a regular greed) is called small world effect.*

Small world:
C(small world) >> C(random)
L (small world) ~ L(random)

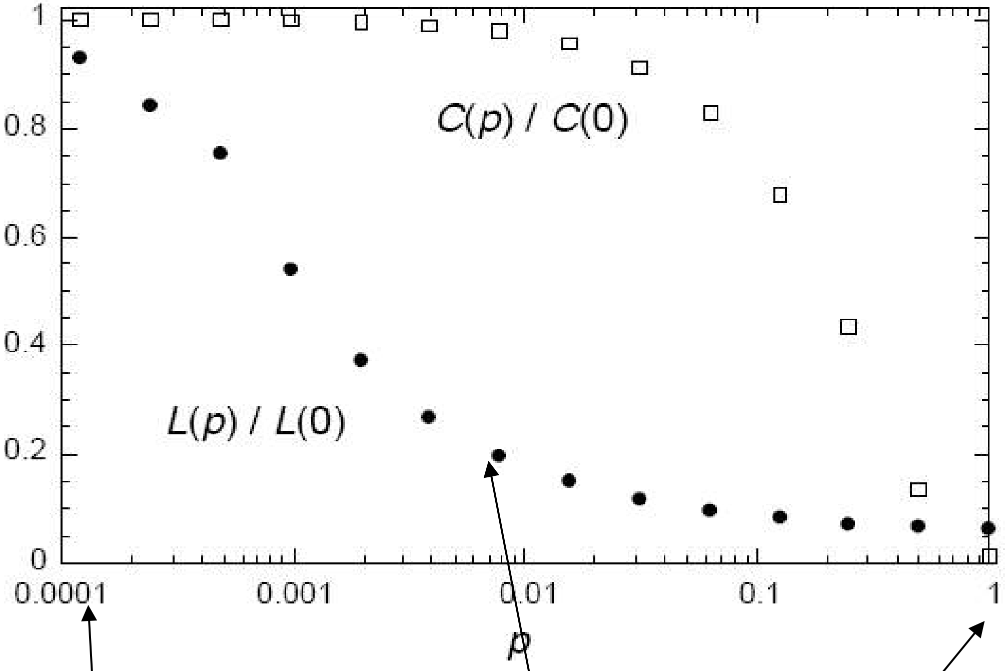
Connecting randomly chosen vertices of regular lattice converts it to a small world net.



Collective dynamics of 'small-world' networks

Duncan J. Watts[*] & Steven H. Strogatz

Department of Theoretical and Applied Mechanics, Kimball Hall, Cornell University, Ithaca, New York 14853, USA



regular

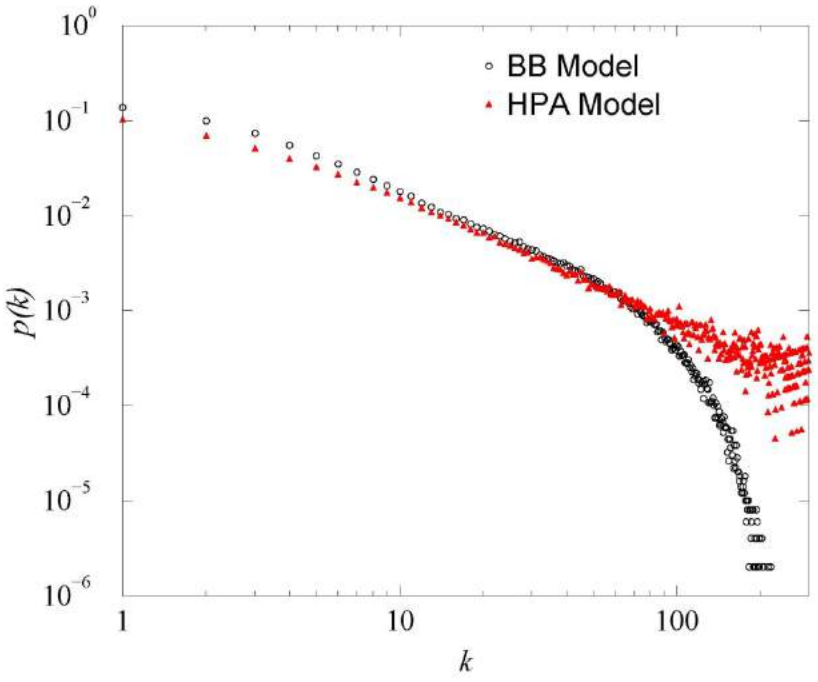
small world

random

How did the biological network evolve?

- We can measure a number of properties of a network. Can we infer anything about network evolution?
- Possible models, (assume protein protein interaction network):
  - E-R (this is rejected by degree distribution)
  - Preferential attachment (makes no biological sense)
  - Gene duplication and assume that duplicate interacts with approximately the same proteins as the original gene, and numerous variants of this model

Degree distribution is not specific

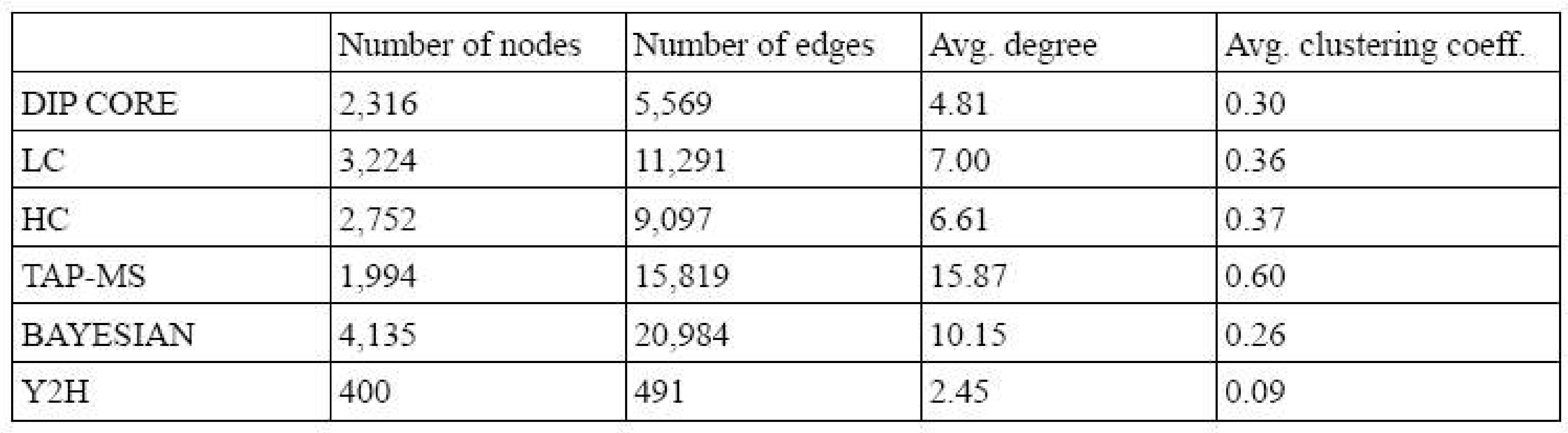


- We have two different models (doesn't matter what their exact definition is).
- Both models agree not only with other on a large interval but also with the data (data points not shown).
- The real data is in the interval (1,80)

T. Przytycka, Yi-Kou Yu, *short paper* ISMB 2004
J. Comp. Biol. and Chem. 2004

Protein Interaction Networks

- DIP CORE Deane et.al. 2002
  - high-confidence interactions from the DIP database
- LC (Literature Curated) Reguly et.al. 2006
  - interactions reported in small-scale experiments
- HC (High Confidence) Batada et.al. 2006
  - interactions reported by several independent studies
- TAP-MS Collins et.al. 2007
  - interactions derived from two high-throughput complex purification experiments
- BAYESIAN Jansen et.al. 2003
  - interactions derived in-silico (from experimental data) using Bayesian Networks formalism

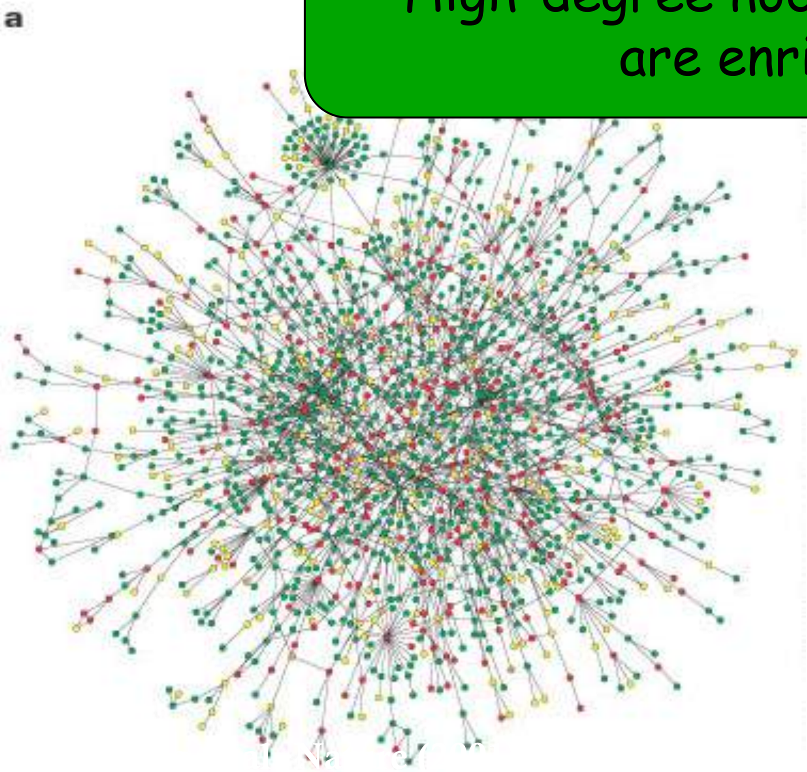


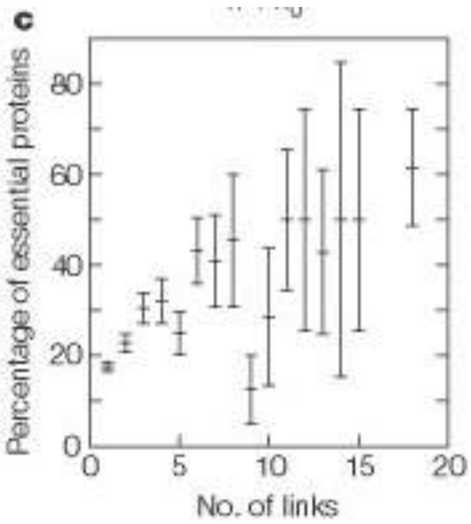
	Number of nodes	Number of edges	Avg. degree	Avg. clustering coeff.
DIP CORE	2,316	5,569	4.81	0.30
LC	3,224	11,291	7.00	0.36
HC	2,752	9,097	6.61	0.37
TAP-MS	1,994	15,819	15.87	0.60
BAYESIAN	4,135	20,984	10.15	0.26
Y2H	400	491	2.45	0.09

Is there a relation between graph-theoretical properties of a network and function?

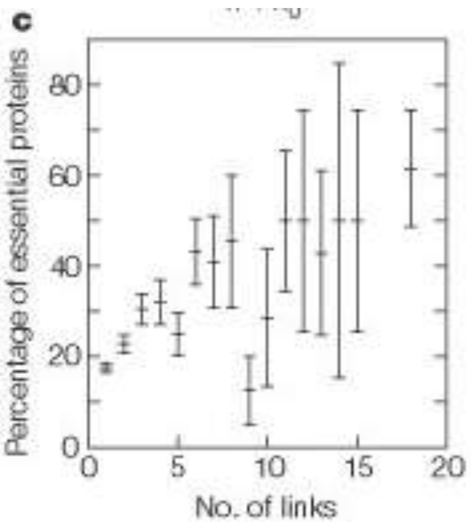
The Centrality-Lethality Rule

High-degree nodes in a protein interaction network are enriched in essential proteins.

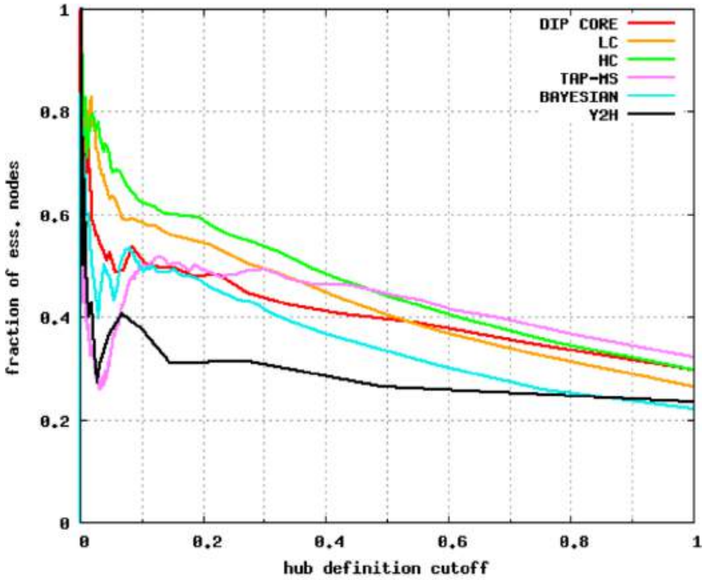




Essentiality and vertex degree, centrality



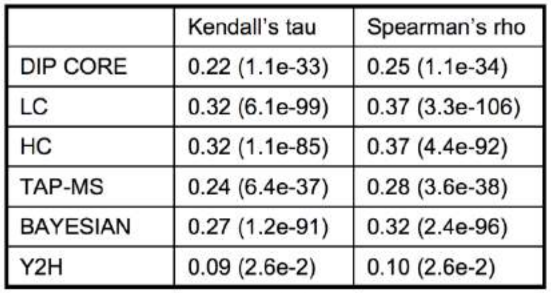
Gene essentiality is correlated with node degree (Jeong et al, 2001) proposed Centrality –Lethality rule



(a)

Observation confirmed in many but NOT all networks

No relation between vertex and lethality in the new, larger set of Y2H Yu et al. Science 2008



	Kendall's tau	Spearman's rho
DIP CORE	0.22 (1.1e-33)	0.25 (1.1e-34)
LC	0.32 (6.1e-99)	0.37 (3.3e-106)
HC	0.32 (1.1e-85)	0.37 (4.4e-92)
TAP-MS	0.24 (6.4e-37)	0.28 (3.6e-38)
BAYESIAN	0.27 (1.2e-91)	0.32 (2.4e-96)
Y2H	0.09 (2.6e-2)	0.10 (2.6e-2)

(b)

Zotenko et al 2008

Modularity and essentiality

- Essentiality – lethality rule (for networks other than Y2H) can be explained by Essential of Complex Biological Modules (ECOBIM)- densely connected subnetworks (presumably protein complexes) enriched in essential proteins (Zotenko et al. 2008)

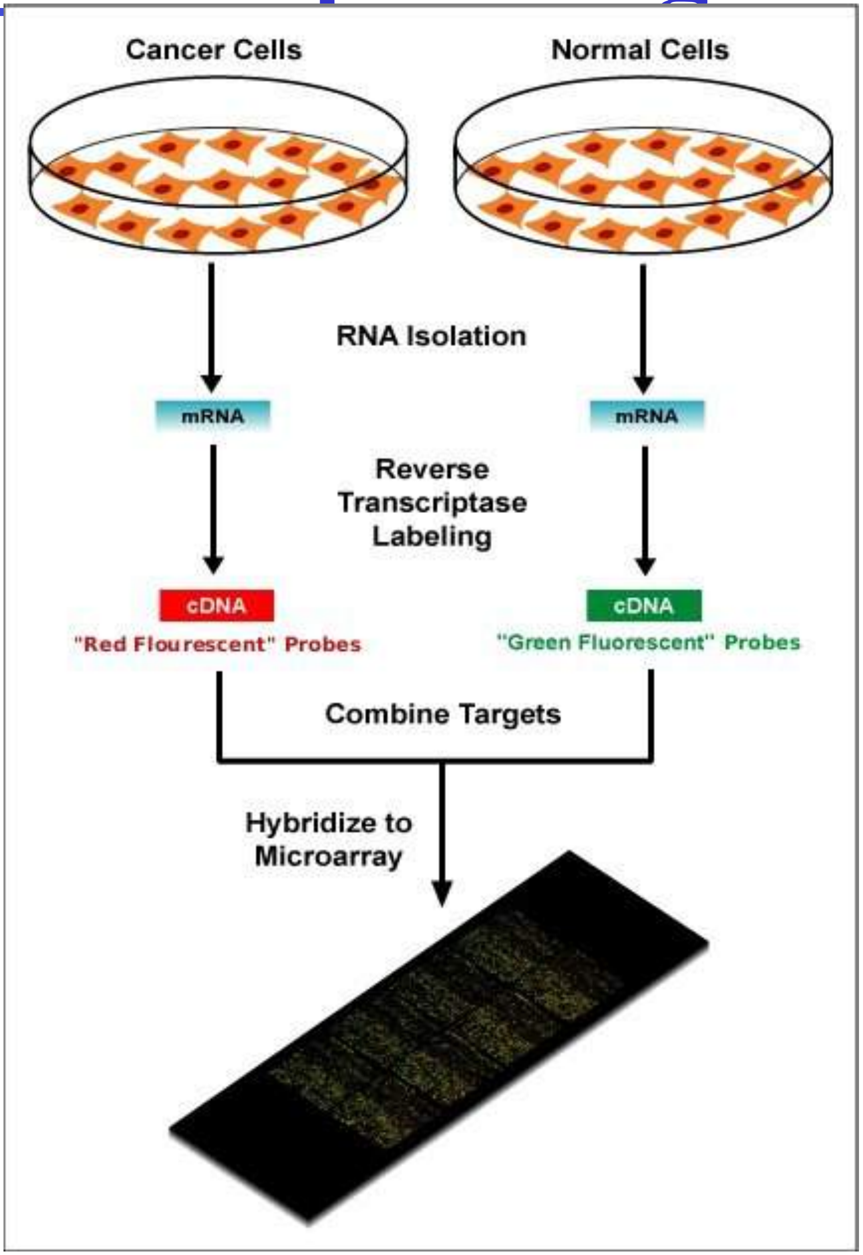
- large complexes are enriched in essential proteins and the enrichment is increasing with complex size (Wang et al. 2009)

Modularity and essentiality, continued

- Gene essentiality is modular: that essentiality is a product of the protein complex rather than the individual protein - Hart, Lee and Edward M Marcotte, 2007.
- The best predictor of a protein's knockout phenotype is the knockout phenotype of other proteins that are present in a protein complex with it. Farser and Plotkin 2007
- Complex Biological Modules (densely connected sub networks) are clearly partitioned in ones that are depleted of essential proteins and ones that are enriched (ECOBINS)) Zotenko et. al. 2008

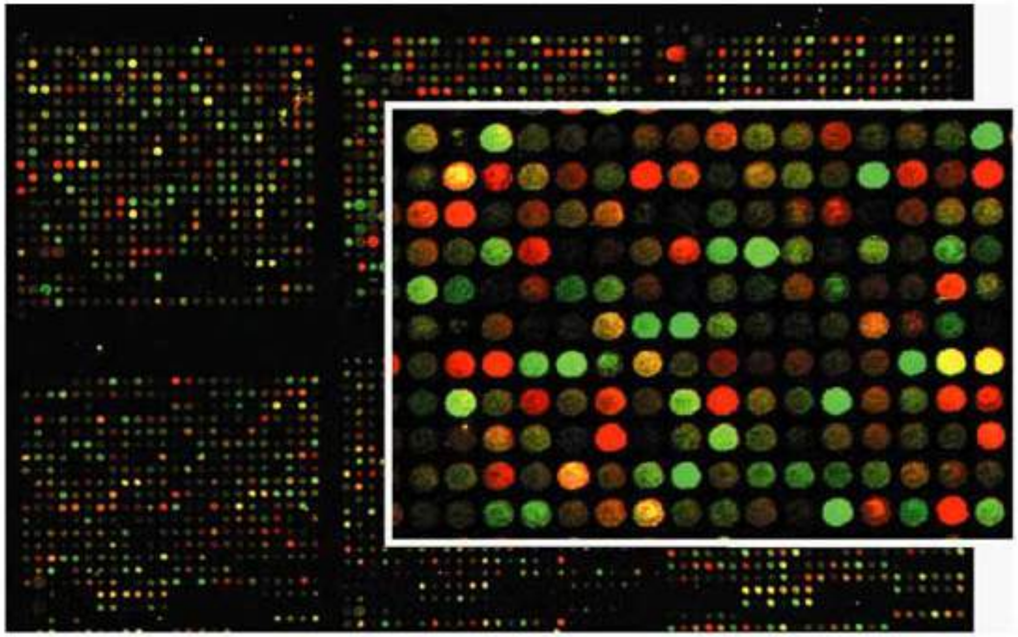
mRNA gene expression profiling

- Monitoring expression levels for thousands of genes simultaneously to study the effects of certain treatments
- Illustration on the left is from wikipedia



- We can learn which genes change expression as a result of treatment
- We can monitor gene expression in different conditions or in different time steps after treatment. This will give us a set of arrays.
- Then we can ask which genes have similar expression patterns
- Gene products of genes that are co-expressed often interact (physically or functionally)

If labeled DNA binds to sequence on microarray, that sequence is being transcribed in the cell



Distance Metric (finding co-expressed genes)

X,Y genes; X_i, Y_i expression of X and Y respectively in condition *i*

N- number of conditions

Φ-standard deviation

G_{offset} is the mean of observations on gene *G*,

$$S(X, Y) = \frac{1}{N} \sum_{i=1,N} \left(\frac{X_i - X_{\mathrm{offset}}}{\Phi_X} \right) \left(\frac{Y_i - Y_{\mathrm{offset}}}{\Phi_Y} \right)$$

$$\Phi_G = \sqrt{\sum_{i=1,N} \frac{(G_i - G_{\mathrm{offset}})^2}{N}}.$$

The basic algorithm from class 9 provides means of hierarchical clustering

Input: distance array d; cluster to cluster distance function

Initialize:

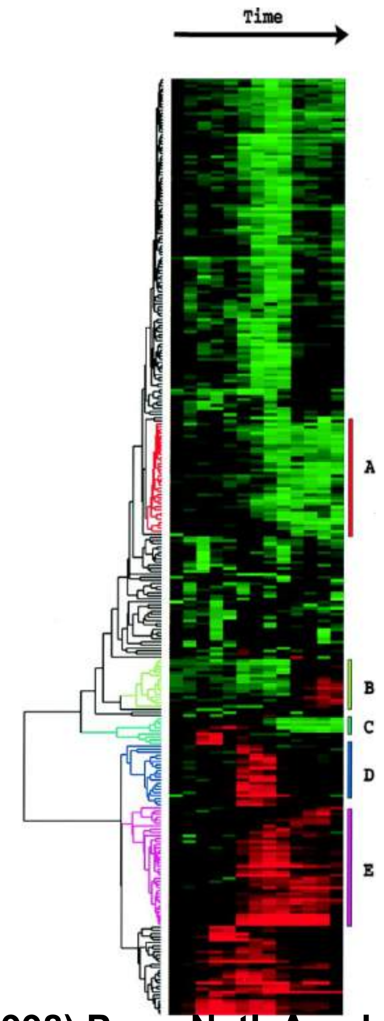
1. **Put every element in one-element cluster**
2. **Initialize a forest T of one-node trees (each tree corresponds to one cluster)**

while there is more than on cluster

1. **Find two closest clusters C_1 and C_2 and merge them into C**
2. **Compute distance from C to all other clusters**
3. **Add new vertex corresponding to C to the forest T and make nodes corresponding to C_1, C_2 children of this node.**
4. **Remove from d columns corresponding to C_1,C_2**
5. **Add to d column corresponding to C**

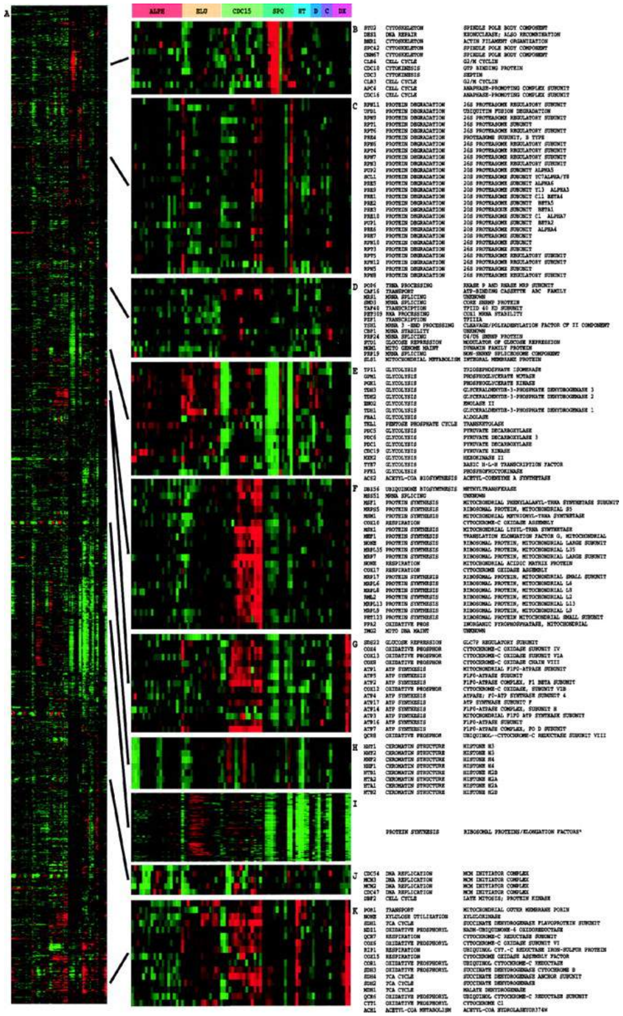
Clustering expression profiles genes with similar pattern of expression

genes



Eisen, Michael B. et al. (1998) Proc. Natl. Acad. Sci. USA 95, 14863-14868

Clustering of expression profiles of yeast genes

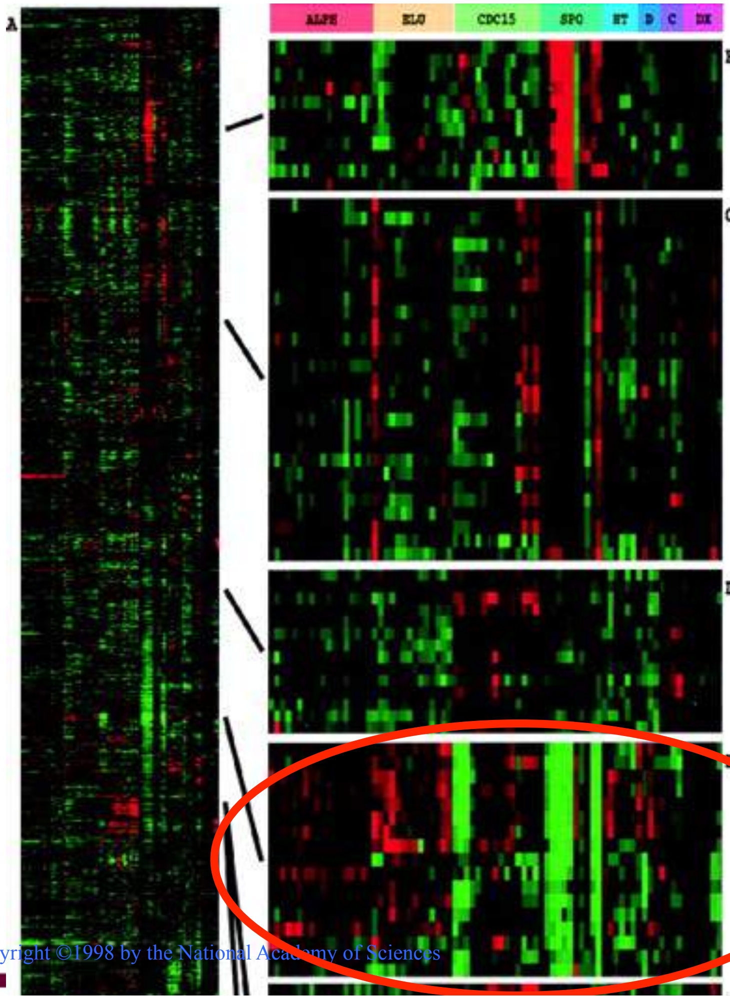


Data were drawn from time courses during the following processes: the cell division cycle after synchronization by alpha factor arrest (ALPH; 18 time points); centrifugal elutriation (ELU; 14 time points), and with a temperature-sensitive cdc15 mutant (CDC15; 15 time points); sporulation (SPO, 7 time points plus four additional samples); shock by high temperature (HT, 6 time points); reducing agents (D, 4 time points) and low temperature (C; 4 time points) (P. T. S., J. Cuoczo, C. Kaiser, P.O. B., and D. B., unpublished work); and the diauxic shift (DX, 7 time points).

Eisen, Michael B. et al. (1998) Proc. Natl. Acad. Sci. USA 95, 14863-14868

Eisen, Michael B. et al. (1998) Proc. Natl. Acad. Sci. USA 95, 14863-14868

Full gene names are shown for representative clusters containing functionally related genes involved in (*B*) spindle pole body assembly and function, (*C*) the proteasome, (*D*) mRNA splicing, (*E*) glycolysis, (*F*) the mitochondrial ribosome, (*G*) ATP synthesis, (*H*) chromatin structure, (*I*) the ribosome and translation, (*J*) DNA replication, and (*K*) the tricarboxylic acid cycle and respiration.



B	STU2	CYTOSKELETON	SPINDLE POLE BODY COMPONENT
	DNS1	DNA REPAIR	EXONUCLEASE; ALSO RECOMBINATION
	BNR1	CYTOSKELETON	ACTIN FILAMENT ORGANIZATION
	SPC42	CYTOSKELETON	SPINDLE POLE BODY COMPONENT
	CNM67	CYTOSKELETON	SPINDLE POLE BODY COMPONENT
	CLB4	CELL CYCLE	G2/M CYCLIN
	CDC10	CYTOKINESIS	GTP BINDING PROTEIN
	CDC3	CYTOKINESIS	SEPTIN
	CLB3	CELL CYCLE	G2/M CYCLIN
	APC4	CELL CYCLE	ANAPHASE-PROMOTING COMPLEX SUBUNIT
	CDC16	CELL CYCLE	ANAPHASE-PROMOTING COMPLEX SUBUNIT
C	RPN11	PROTEIN DEGRADATION	26S PROTEASOME REGULATORY SUBUNIT
	UFD1	PROTEIN DEGRADATION	UBIQUITIN FUSION DEGRADATION
	RPN9	PROTEIN DEGRADATION	26S PROTEASOME REGULATORY SUBUNIT
	RPT1	PROTEIN DEGRADATION	26S PROTEASOME SUBUNIT
	RPT6	PROTEIN DEGRADATION	26S PROTEASOME REGULATORY SUBUNIT
	PRE4	PROTEIN DEGRADATION	PROTEASOME SUBUNIT, B TYPE
	RPN6	PROTEIN DEGRADATION	26S PROTEASOME REGULATORY SUBUNIT
	RPT4	PROTEIN DEGRADATION	26S PROTEASOME REGULATORY SUBUNIT
	RPN7	PROTEIN DEGRADATION	26S PROTEASOME REGULATORY SUBUNIT
	RPN3	PROTEIN DEGRADATION	26S PROTEASOME REGULATORY SUBUNIT
	PUP2	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT ALPHA5
	SCL1	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT YC7ALPHA/Y8
	PRE5	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT ALPHA6
	PRE9	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT Y13 ALPHA3
	PRE1	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT C11 BETA4
	PRE2	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT BETA5
	PRE3	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT BETA1
	PRE10	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT C1 ALPHA7
	PUP1	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT BETA2
	PRE6	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT ALPHA4
	PRE7	PROTEIN DEGRADATION	20S PROTEASOME SUBUNIT
	RPN10	PROTEIN DEGRADATION	26S PROTEASOME SUBUNIT
	RPT3	PROTEIN DEGRADATION	26S PROTEASOME SUBUNIT
	RPT5	PROTEIN DEGRADATION	26S PROTEASOME REGULATORY SUBUNIT
	RPN12	PROTEIN DEGRADATION	26S PROTEASOME REGULATORY SUBUNIT
	RPN5	PROTEIN DEGRADATION	26S PROTEASOME SUBUNIT
	RPN8	PROTEIN DEGRADATION	26S PROTEASOME REGULATORY SUBUNIT
D	POP6	TRNA PROCESSING	RNASE P AND RNASE MRP SUBUNIT
	CAF16	TRANSPORT	ATP-BINDING CASSETTE ABC FAMILY
	MRS1	MRNA SPLICING	UNKNOWN
	SMD3	MRNA SPLICING	CORE SNRNP PROTEIN
	TAF40	TRANSCRIPTION	TFIID 40 KD SUBUNIT
	PET309	RNA PROCESSING	COX1 MRNA STABILITY
	PZF1	TRANSCRIPTION	TFIIIA
	YSH1	MRNA 3 -END PROCESSING	CLEAVAGE/POLYADENYLATION FACTOR CF II COMPONENT
	CBP1	MRNA STABILITY	UNKNOWN
	PRP24	MRNA SPLICING	U4/U6 SNRNP PROTEIN
	STD1	GLUCOSE REPRESSION	MODULATOR OF GLUCOSE REPRESSION
	MGM1	MITO GENOME MAINT	DYNAMIN FAMILY PROTEIN
	PRP19	MRNA SPLICING	NON-SNRNP SPLICEOSOME COMPONENT
	SLS1	MITOCHONDRIAL METABOLISM	INTEGRAL MEMBRANE PROTEIN
E	TPI1	GLYCOLYSIS	TRIOSEPHOSPHATE ISOMERASE
	GPM1	GLYCOLYSIS	PHOSPHOGLYCERATE MUTASE
	PGK1	GLYCOLYSIS	PHOSPHOGLYCERATE KINASE
	TDH3	GLYCOLYSIS	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 3
	TDH2	GLYCOLYSIS	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 2
	ENO2	GLYCOLYSIS	ENOLASE II
	TDH1	GLYCOLYSIS	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 1
	FBA1	GLYCOLYSIS	ALDOLASE
	TKL1	PENTOSE PHOSPHATE CYCLE	TRANSKETOLASE
	PDC5	GLYCOLYSIS	PYRUVATE DECARBOXYLASE
	PDC6	GLYCOLYSIS	PYRUVATE DECARBOXYLASE 3
	PDC1	GLYCOLYSIS	PYRUVATE DECARBOXYLASE
	CDC19	GLYCOLYSIS	PYRUVATE KINASE
	HXK2	GLYCOLYSIS	HEXOKINASE II
	TYE7	GLYCOLYSIS	BASIC H-L-H TRANSCRIPTION FACTOR
	PFK1	GLYCOLYSIS	PHOSPHOFRUCTOKINASE
	ACS2	ACETYL-COA BIOSYNTHESIS	ACETYL-COENZYME A SYNTHETASE

So you got some clusters now what?

Check whether a cluster is enriched in some biological function/process etc.

Example – GO terms annotation

http://www.geneontology.org/

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data

Gene Ontology(GO)

The Ontologies:

Cellular component

Biological process

Molecular function

Superimposing protein interaction data end expression data

Date hubs, party hubs, family hubs

•'party' hubs - Hubs highly coexpressed with their neighbors and that are therefore in modules as 'party' hubs (Han *et al*, 2004).

•'date' hubs those that are not coexpressed with their neighbors (Han *et al*, 2004).

•'family' hubs - hubs located in static neighborhoods that is constitutively expressed in the network in condition-independent manner and interact with their partners constitutively (they interact with their neighbors constitutively Komurov and White 2007

•Proteins in dynamic/party modules are almost twice as likely to be essential as proteins in static modules (Komurov and White 2007)