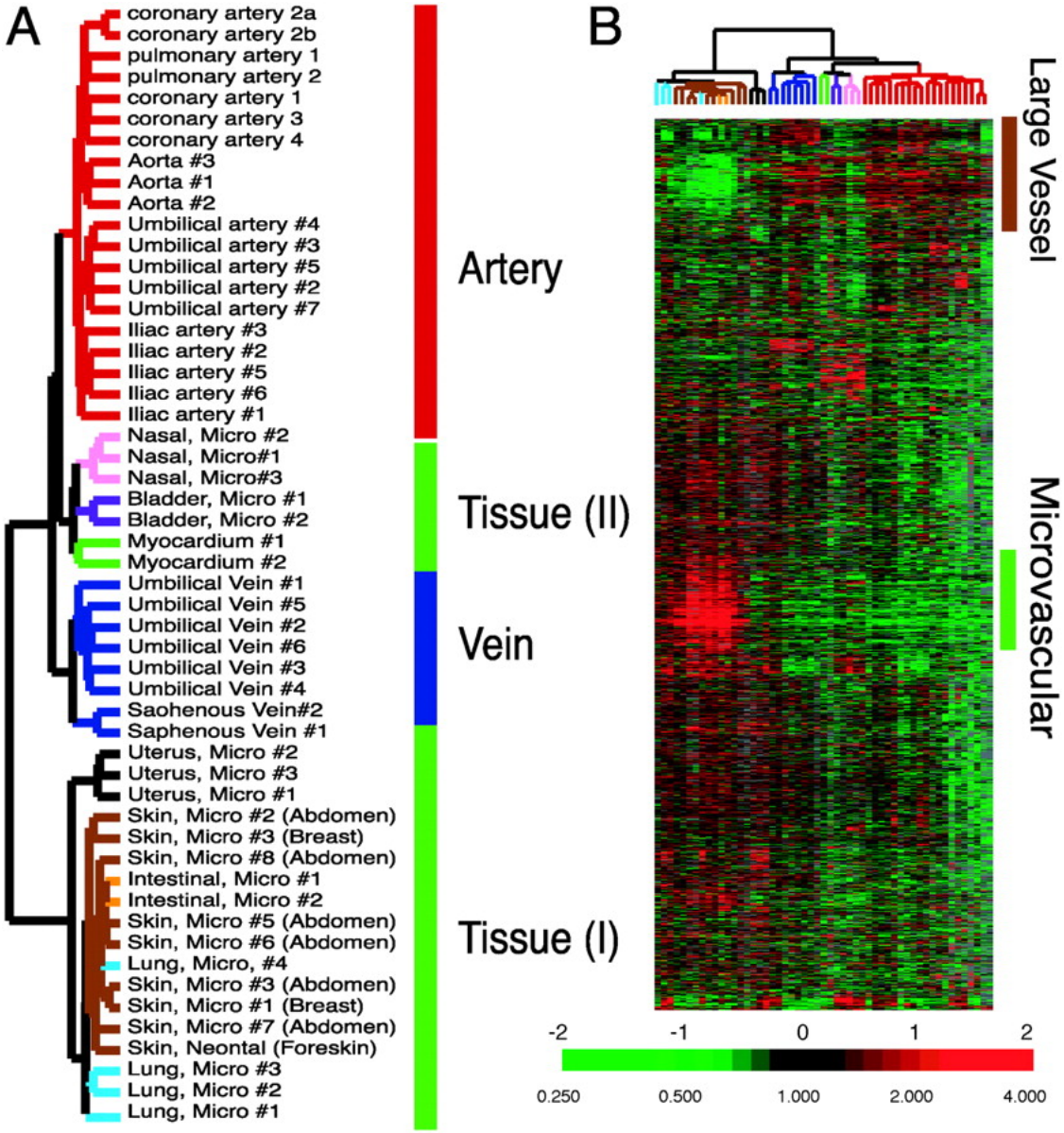
Gene expression 3

ROC curves

Multiple testing

Gene expression networks

Hierarchical clustering results



Chi et al., PNAS | **September 16, 2003** | vol. 100 | no. 19 | **10623-10628**

"Endothelial cell diversity revealed by global expression profiling"

Receiver operating characteristic

- A framework to compare the performance of binary classifiers
- Plot of *false positive rate (sensitivity) vs true positive rate* (1-specificity)

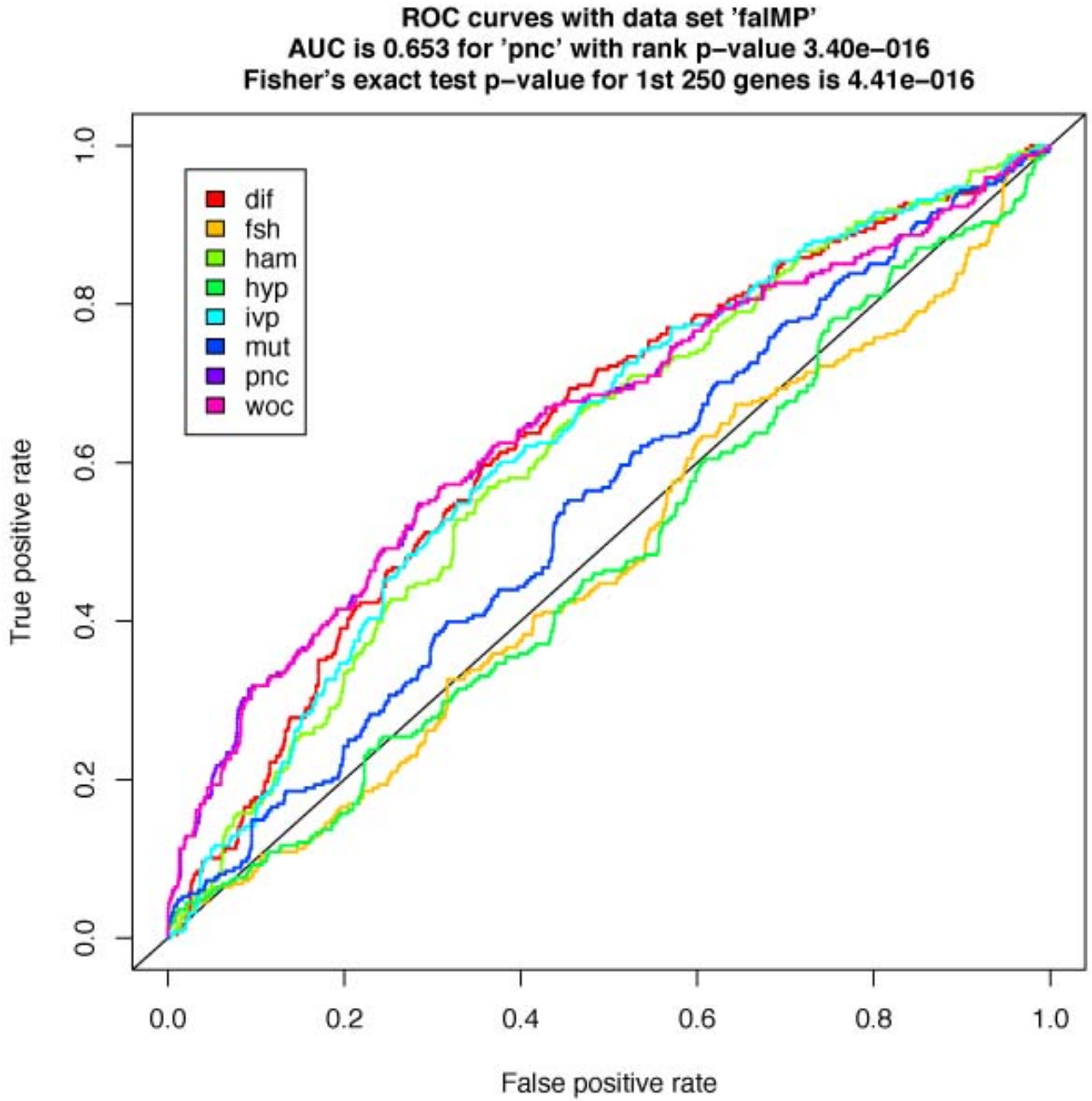
Gütemaße

- Sensitivität/*Recall*
  TPR = TP/P = TP/(TP+FN)

- Spezifizität
  FPR = FP/N = FP/(FP+TN)

- *Precision (positive predictive value)*
  PPV= TP/(TP+FP)

- *False discovery rate*
  FDR=FP/(TP+FP)


ROC curves with data set 'falMP'
AUC is 0.653 for 'pnc' with rank p−value 3.40e−016
Fisher's exact test p−value for 1st 250 genes is 4.41e−016
dif
fsh
ham
hyp
ivp
mut
pnc
woc
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Hypothesis Testing

- H0 : Null hypotheis vs. H1 : Alternative Hypothesis
- T : test statistics C : critical value
- If |T|>C, H0 is rejected. Otherwise H0 is retained
- Example
  H0 : $\mu 1 = \mu 2$ vs. H1 : $\mu 1 \neq \mu 2$
  T = (x1- x2) / pooled standard error (se)
- If |T| > z(1- $\alpha/2$), H0 is rejected at the significance level α
- Cα

Hypothesis Testing

		Hypothesis Result	
		Retained	Rejected
Truth	H0		Type I error
	H1	Type II error	

- Type I error rate = false positives (α : significance level)
- Type II error rate = false negatives
- Power : 1–Type II error rate

- P-values : $p=\inf\{\alpha \mid H0 \text{ is rejected at the significance level } \alpha\}$

Issues in Multiple Comparison

- Given n treatments, which two treatments are significantly different ? (simultaneous testing)
- Is treatment A different from treatment B ?

- m treatment means : $\mu_1,\ldots,\mu_n$

  $H_j : \mu_i = \mu_j$ where $i \neq j$ $T_j = (x_i - x_j)$ / pooled SE

- Type I error when testing each at 0.05 significance level one by one : $1 - (0.95)^n$
- Inflated Type I error, ex) $\alpha = 1 - (0.95)^{10} = 0.401263$

- Remedies : Bonferroni Method

  Type I error rate = α / # of comparison

Type I Error Rates

		Hypothesis Result		
		#retained	#rejected	Total
Truth	H0	U	V	m0
	H1	T	S	m1
	Total	m-R	R	m

- Per-comparison error rate (PCER) = E(V) / m
- Per-family error rate (PFER) = E(V)
- Family-wise error rate = pr (V ≥ 1)
- False discovery rate (FDR) = E(Q), $Q \begin{cases} V/R \text{ , if } R > 0 \\ 0, \quad \text{if } R = 0 \end{cases}$

Type I Error Rates

Under the complete null hypothesis, each H_j has Type I error rate α_j.

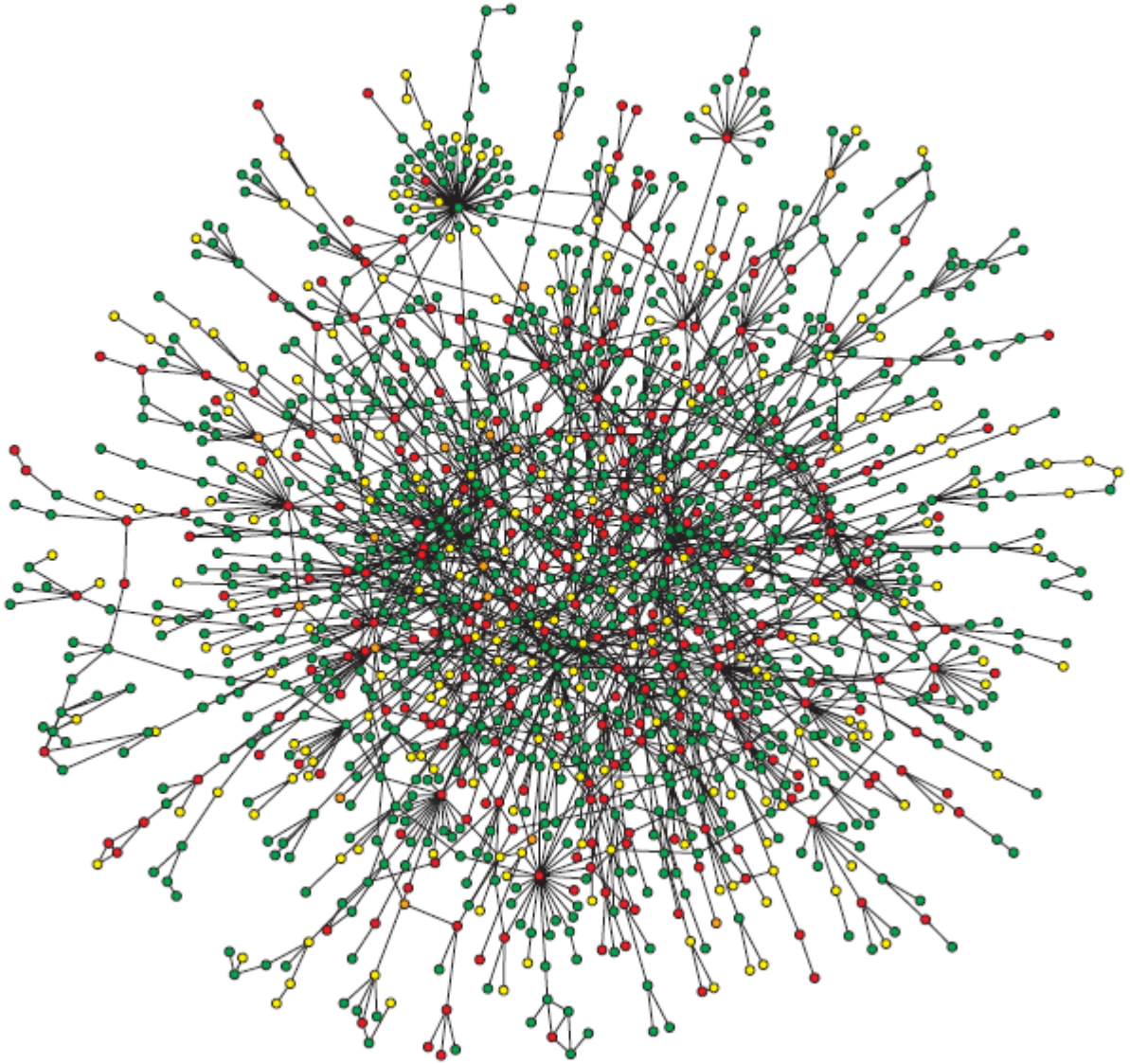
- PCER = E(V) / m = $(\alpha_1+...+\alpha_m)/m$
- PFER = E(V) = $\alpha_1+...+\alpha_m$
- FWER= pr (V ≥ 1) = 1 - Pr (H_j , j=1, …, m, not rejected)
- FDR = E(V / R) = FWER

PCER = $(\alpha_1+...+\alpha_m)/m$ ≤ max $(\alpha_1+...+\alpha_m)$ ≤ PWER = FDR ≤ PFER= $\alpha_1+...+\alpha_m$

Types of comparisons

- Assume Hj , j=1, …, m, with their test statistics Tj , j=1,…, m, which has a MN with mean $\mu=(\mu 1,\ldots,\mu m)$ and identity covariance vector
- Let Rj = I (Hj is rejected) and rj is observed value of Rj
- Let γj = Pr (Hj rejected under Hj).

- PFER = $\sum j=1m\ \gamma j$ (Per family error rate)
- PCER = $\sum j=1m\ \gamma j$ / m (Per comparison error rate)
- FWER = 1- $\prod j=1m$ (1- γj) (Family wise error rate)
- FDR = $\sum r1=01\ldots\sum r1=01(\sum j=1m0rj$ / $\sum j=1mrj)\prod \gamma jrj$ (1- γj) 1-rj (False discovery rate)



Networks

Considerations for the analysis

- Directed vs undirected graphs
- Analysis of confounding factors
- How to assign weights?
  - Repetitions in screen
  - Outgoing and incoming edges
  - External data
- Hubs

→b

Topological analysis

- Small worlds
  - Shortest path lengths are small
  - Degrees of separation
- Modular
  - Clustering co-efficient

$$C_i = \frac{2E_i}{k_i(k_i - 1)}$$

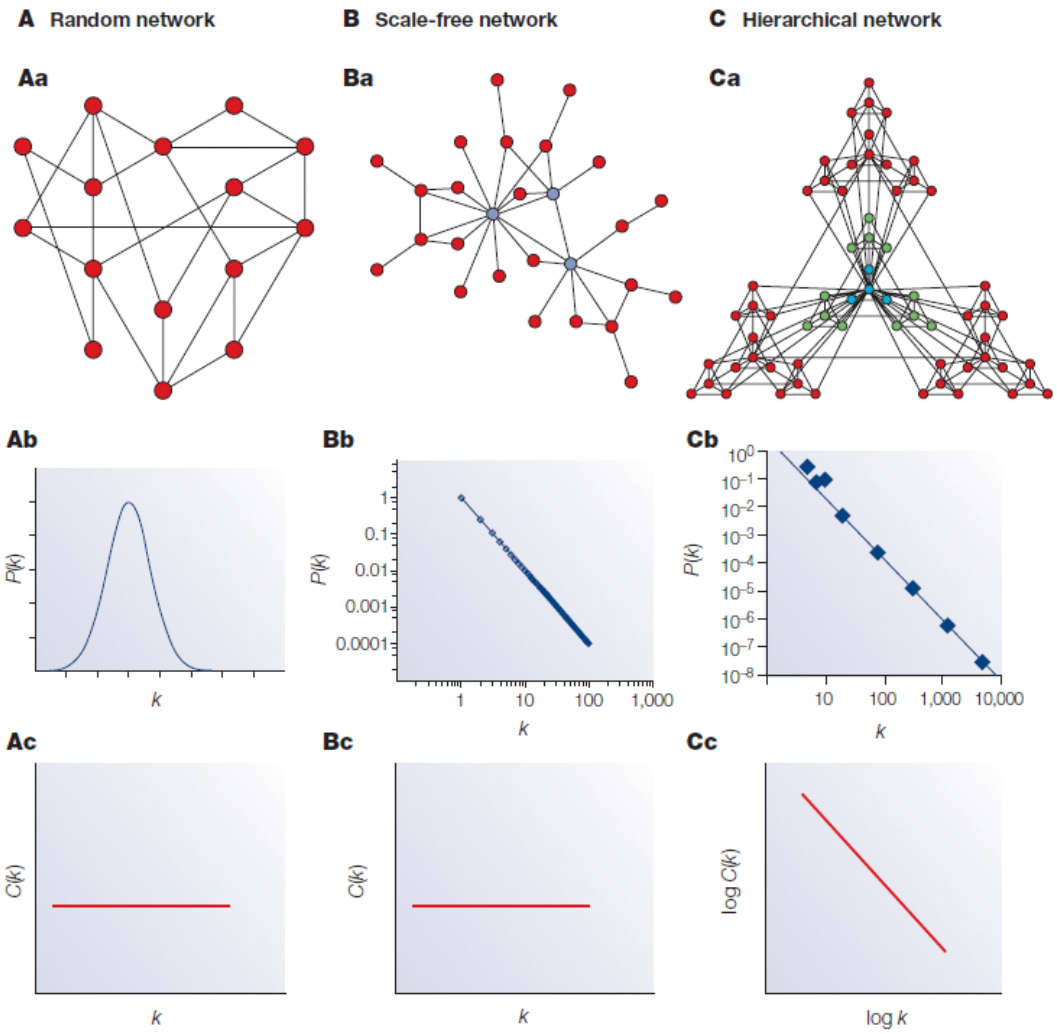
- Degree distribution
  - Random model

$$P(\deg(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k},$$

    - Poisson with max = P (<k>)
  - Scale free
    - $P(k) \sim k^{-r}$
    - $-1.5 > r > -3$

Different networks

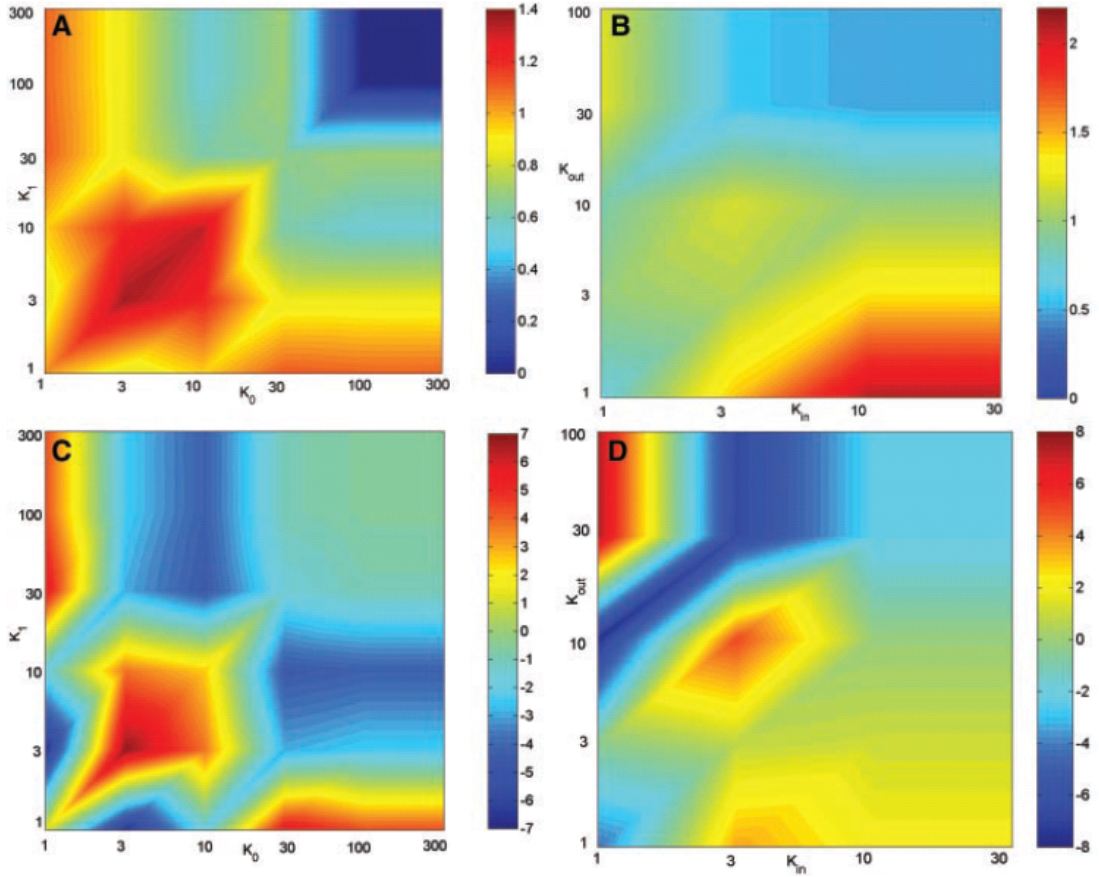
From Barabási (2004), Nature Reviews Genetics



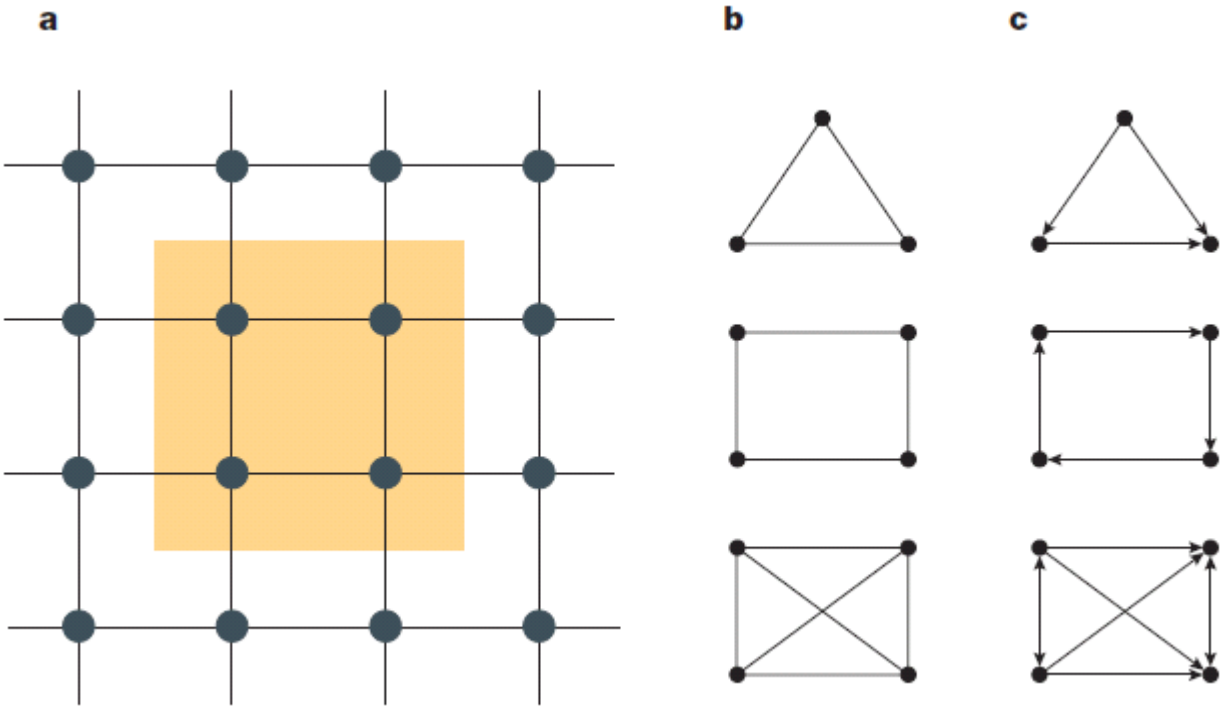
Connections between hubs

Maslov and Sneppen (2002)
Science

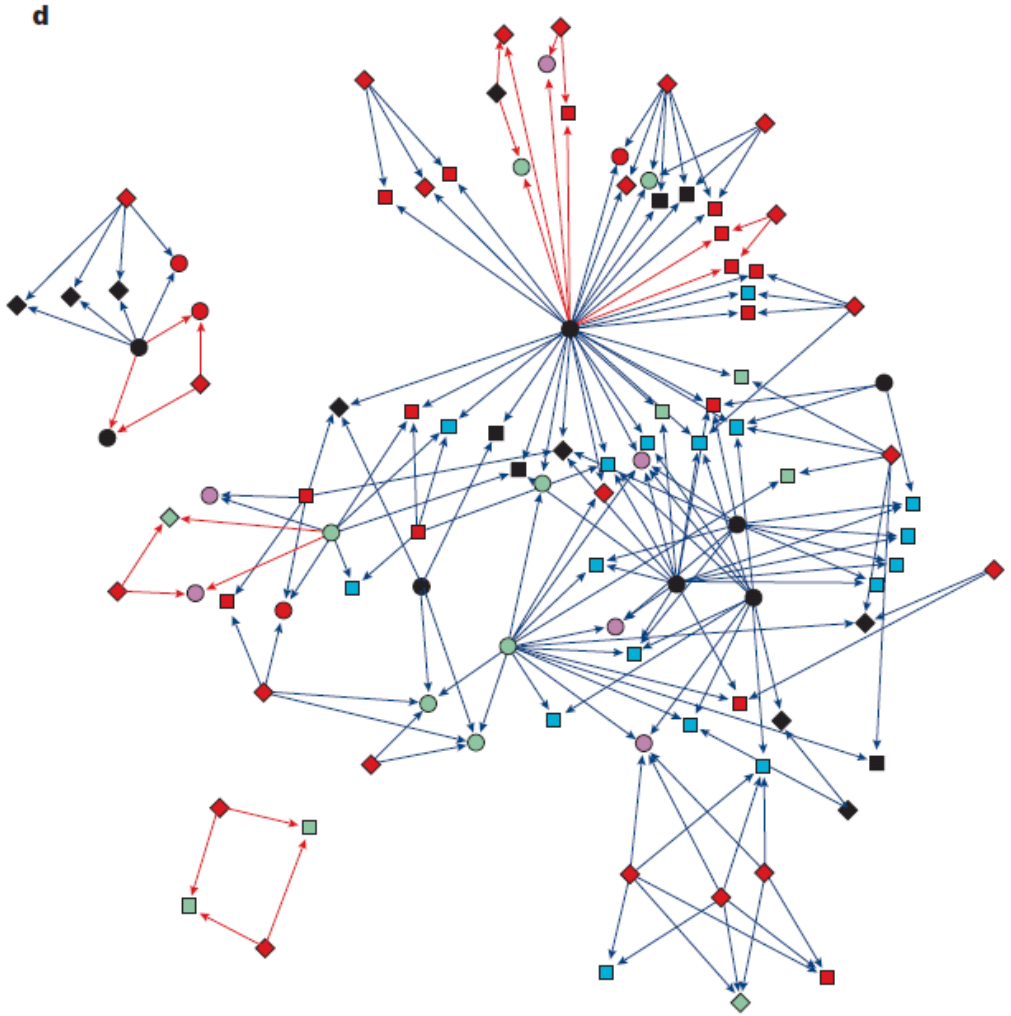
Hubs are connected to proteins of low degree, not between each other



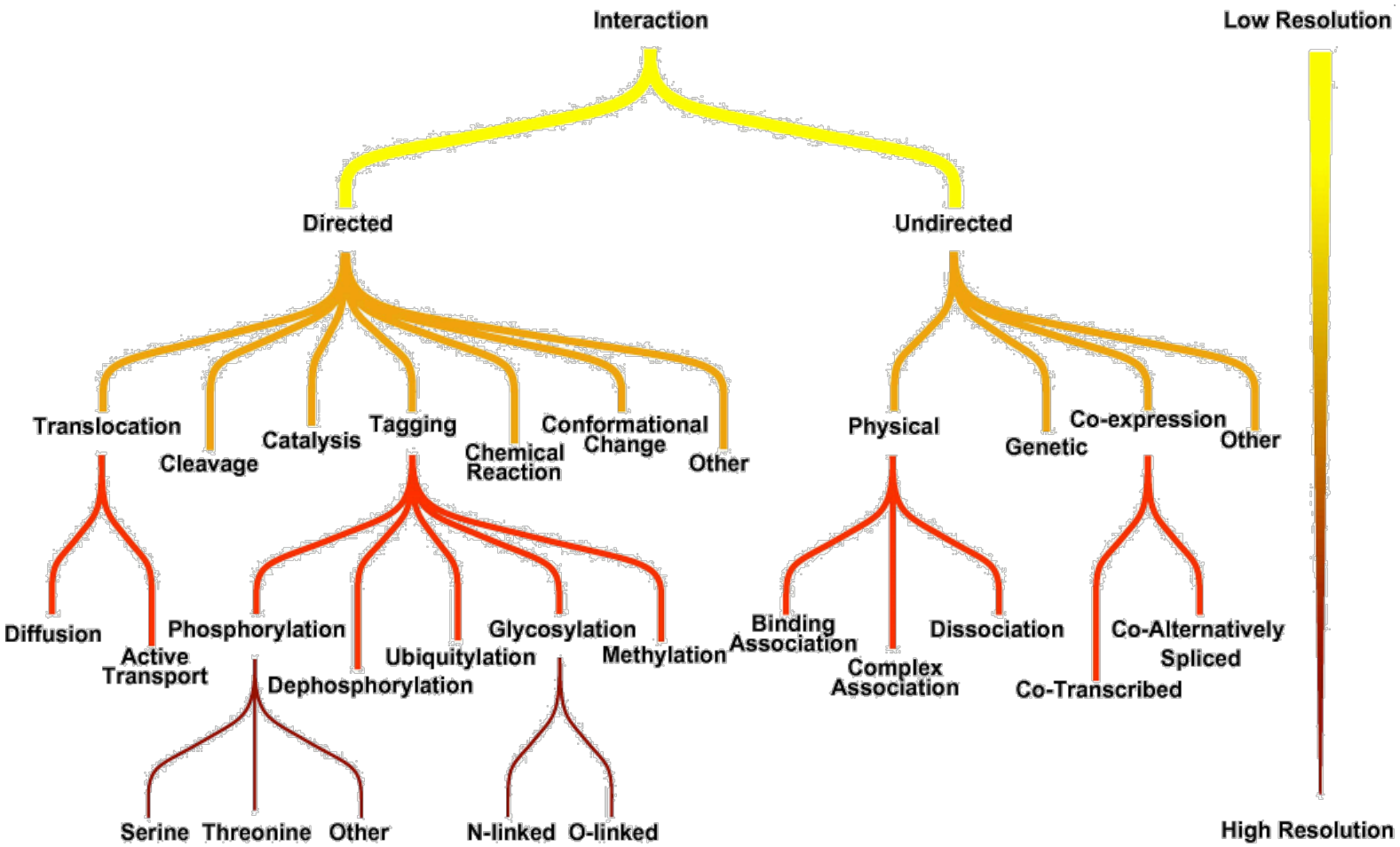
Motifs and subgraphs



Motifs in real networks



Biological types of interactions



A proposed ontology for interactions (Lu et al.)

- Unweighted graphs
  - Hamming distance
- Weighted graphs
  - Euclidean distance
  - Correlation
    - Pearson
    - Spearman

- Boolean networks
- Probabilistic networks
  - Markov Random Fields
  - Bayesian networks