

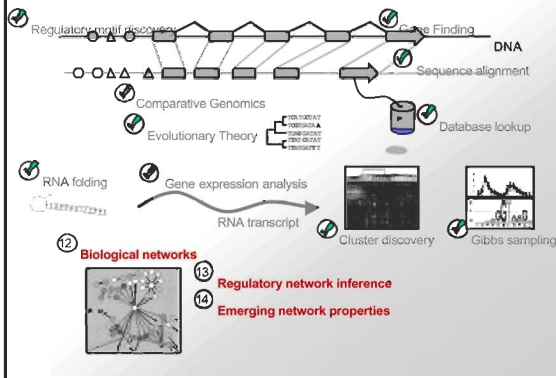
Biological Networks

Microarrays – Expression Clustering – Bayesian nets – Small-world nets

Biological networks

- Lecture 1 - Introduction
- Lecture 2 - Hashing / BLAST
- Lecture 3 - Combinatorial Motif Finding
- Lecture 4 - Statistical Motif Finding
- Lecture 5 - Sequence alignment and Dynamic Programming
- Lecture 6 - RNA structure and Context Free Grammars
- Lecture 7 - Gene finding and Hidden Markov Models
- Lecture 8 - HMMs algorithms and Dynamic Programming
- Lecture 9 - Evolutionary change, phylogenetic trees
- Lecture 11 - Genome rearrangements, genome duplication
- Lecture 12 - Biological networks, expression clustering, small worlds

Challenges in Computational Biology

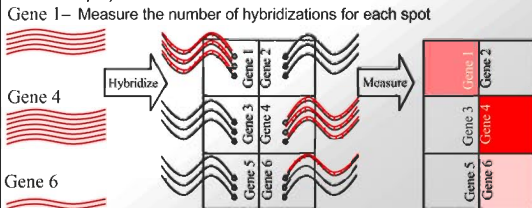


Outline

- Microarray technology
- Clustering gene expression
- TF binding: the controllers
- Bayesian networks
- Network properties
- Scale-free networks

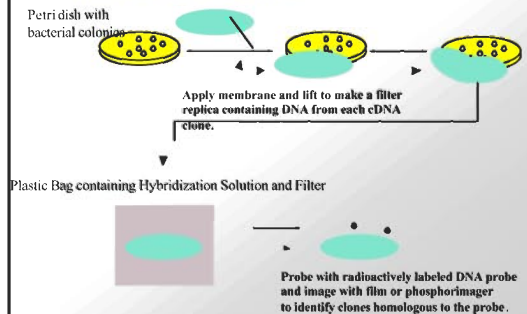
The idea behind DNA MicroArrays

- To measure levels of messages in a cell
 - Construct an array with DNA sequences for 6000 genes
 - Hybridize each RNA in your sample to a sequence in your array (All sequences from the same gene hybridize to the same spot)



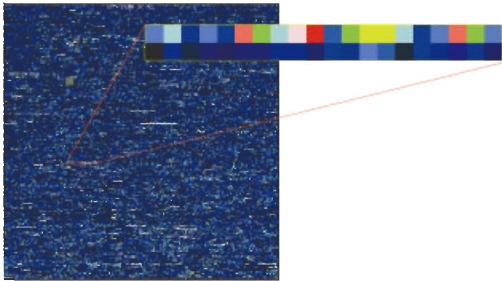
The first RNA expression observation

- Which colonies have expressed a particular gene?



Today

- 6000 genes instead of one
- Entire transcriptome observable in one experiment



Two ways of making DNA Arrays

Printed slides (Stanford)

- PCR amplification of a 1kb portion of the gene of interest (3'most)
- Robotically apply each amplified sample on glass slide
- Chemically attach DNA to glass and thermally denature
- 6000 spots on 2x2cm glass



DNA chips (Affymetrix)

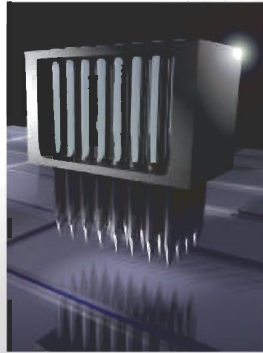
- Each gene provides several neighboring spots on array
- Sampled from various regions within gene (most unique)
- Synthesize oligonucleotides (20b) growing from glass



Printed arrays: fixing DNA onto glass

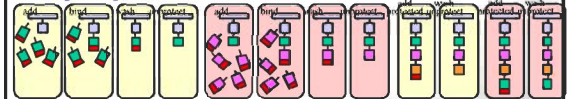
Procedure

1. Array samples onto appropriate spot in glass array
2. Air-dry to fix the samples
3. UV-irradiation forms covalent bonds between T in DNA and positively charged amine groups on the silane slides

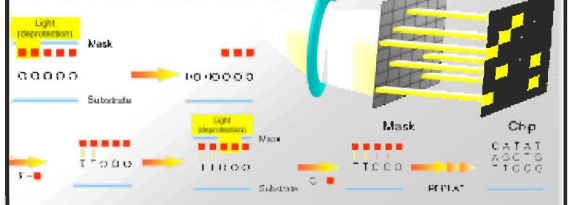


DNA chips: Growing oligos on glass

Making an oligonucleotide

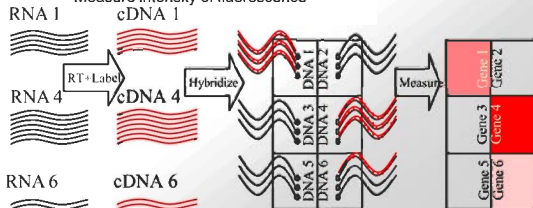


Making a DNA chip (affymetrix.com)

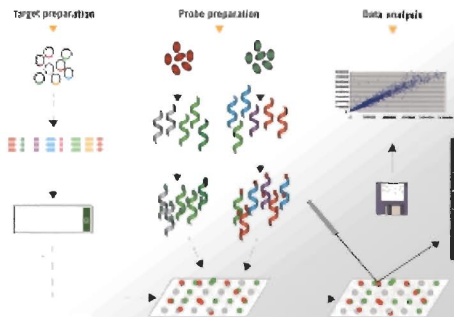


Fluorescently labeling the samples

- From RNA product, reverse transcribe to cDNA
 - Use oligo-dT primer, dNTPs, and a low concentration of nucleotide analog labeled with fluorescent dye
 - After hybridization step, wash away unhybridized cDNA
 - Measure intensity of fluorescence

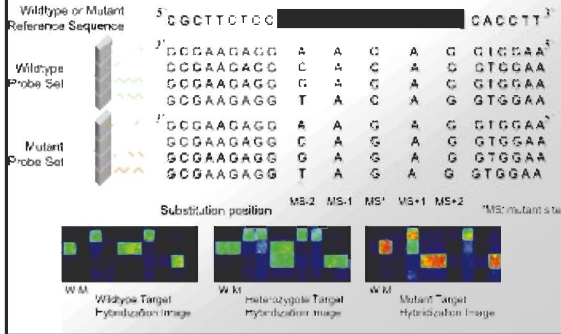


Printed Array: Comparing two conditions



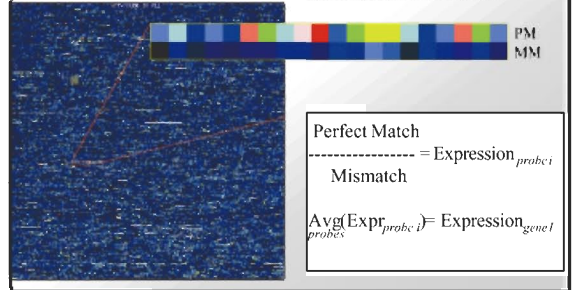
DNA chips: measuring absolute expression

- Compare perfect match affinity to mismatch affinity



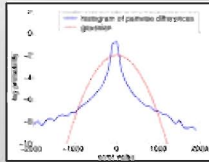
DNA chips: measuring absolute expression

- Statistical models yield gene expression value
 - Separate signal from noise. Estimate noise from multiple observations. Measure uncertainty of each



Modeling noise

- Sources of Noise
 - Cross-hybridization
 - Improper probe sequence
 - Non-uniform hybridization kinetics
 - Non-linearity of array response to concentration
 - Non-linear amplification
- Estimating gene expression value
 - Confidence intervals estimation
 - Model is non-gaussian
 - ML, MAP estimation
 - See Gifford, Young, Jaakkola

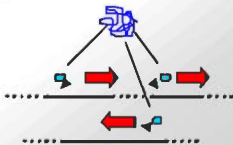


Outline

- Microarray technology
- Clustering gene expression
- TF binding: the controllers
- Bayesian networks
- Network properties
- Scale-free networks

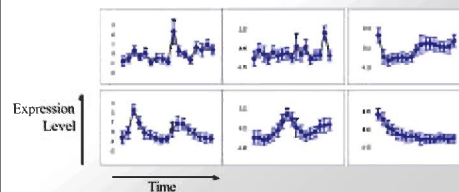
The problem

- Group genes into co-regulated sets
 - Observe cells under different environmental changes
 - Find genes whose expression profiles are affected in a similar way
 - These genes are potentially co-regulated, i.e. regulated by the same transcription factor



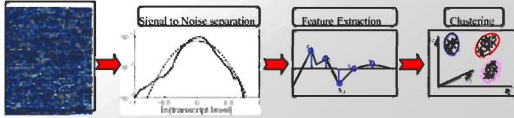
Expression Profiles

- Each coregulated set has a distinct expression profile
 - Expression levels help group genes in a meaningful way
 - In each group, genes have similar expression profiles
 - Shown here are six expression profiles for an experiment. For each of the 18 time steps, the mean and spread of the expression levels in each group are shown



Clustering expression levels

- **Computational problem:**
 - How do we go from expression levels of 6000 genes to meaningful categories of possibly co-regulated genes
- **Method:**
 1. Signal to noise separation. Data normalization.
 2. Feature extraction. Choose features to compare.
 3. Clustering. Group genes into meaningful subsets.
 4. Evaluation. Statistical significance of a grouping.



1. Expression Value Normalization

- **Goal: Combining expression data across experiments**
 - Last lecture: differential expression guaranteed same conditions
 - But values can vary with time, lab, concentrations, chemicals used
 - How do we compare expression values across experiments
- **Noise model** $y_{ij} = n_i [\alpha_{ij}(c_j f_{ij}) + \epsilon_{ij}]$
 - y_{ij} : observed level for gene j on chip i . t_j : true level
 - c_j : gene constant. n_i : multiplicative chip normalization
 - a_j, e_j : multiplicative and additive noise terms
- **Estimating the parameters**
 - n_i : spiked in control probes, not present in genome studied
 - c_j : control experiments of known concentrations for gene j
 - e_j : un-spiked control probes should be zero
 - a_j : spiked controls that are constant across chips

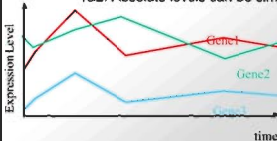
2. Feature extraction

- **Select values which yield maximal group separation**
 - When clustering cell type, select most relevant genes
 - In clustering genes, select most meaningful conditions
- **Pre-process input**
 - Instead of clustering on direct observation of expression values ...
 - ... can cluster based on differential expression from the mean
 - ... or differential expression normalized by standard deviation
- **Sample correlation**
 - 1&3: Absolute levels can be different, yet genes coregulated
 - 1&2: Absolute levels can be similar, yet genes unrelated

$$\sum_{i=1}^{N_{chips}} |x_i - y_i|$$

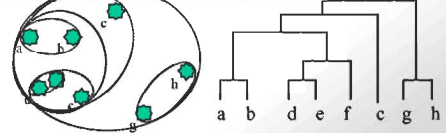
$$\sum_{i=1}^{N_{chips}} \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}$$

$$s(x, y) = \frac{\sum_{i=1}^{N_{chips}} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^p (x_i - \bar{x})^2 \times \sum_{i=1}^p (y_i - \bar{y})^2}}$$

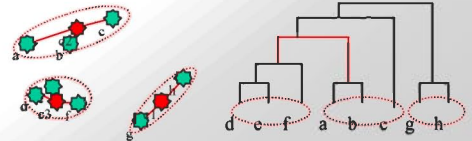


3. Clustering Algorithms

- **Hierarchical: Split data successively to construct tree**

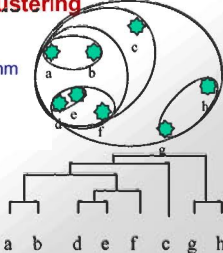


- **Non-Hierarchical: place k-means to best explain data**

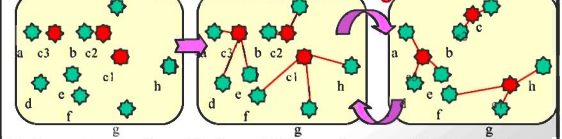


3a. Hierarchical clustering

- **Bottom-up clustering. Greedy algorithm**
 - Construct minimum spanning tree
- **Pre-processing**
 - Create a cluster for every data point
- **Iterative greedy algorithm**
 - Find two nearest clusters
 - Merge them into a single cluster
- **Termination**
 - Return tree of joins
- **$O(n^2)$ comparisons + merge at every step. $O(n)$ steps**
- **Computational cost tradeoffs depends on:**
 - Cluster representation, distance metric, merging algorithm



3b. K-means clustering



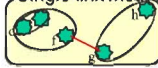
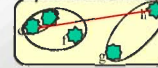
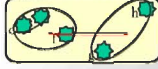
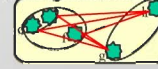
Create random centers Assign points to centers Update centers

- **Iterative algorithm: optimizing random solution**
 - If assignments of points is known, centers easily computed
 - If centers are known, assignment of points easily computed
- **General case: Expectation Maximization (EM algorithm)**
 - In k-means clustering, every point belongs to only one center
 - In general case, probabilistic model. Every point belongs to all centers, to each with probability proportional to distance.

3c. Cluster Representation

- Operations supported by cluster data structure
 - Hierarchical clustering
 - point2cluster(point) → cluster
 - clusters_distance(cluster1, cluster2) → distance
 - clusters_merge(cluster1, cluster2) → new_cluster
 - K-means clustering
 - initialize → k clusters
 - points2cluster(points) → cluster
 - cluster_distance(cluster, point) → distance
- Different representations optimize different metrics
 - Mean of cluster points & number of points in cluster
 - Recompute mean when a new point is added to cluster by weighing appropriately

3d. Cluster distance metrics

- Expression level distances for clusters not genes
 - We already talked about different metrics for comparing individual genes. Differential expression. Correlation.
 - Now we defined $d_{cluster}$ in terms of d_{gene}
- Defining the distance between two clusters
 - Single-link method: 
 - Complete-link method: 
 - Centroid method: 
 - Average-link method: 

Depending on application, different methods best fit

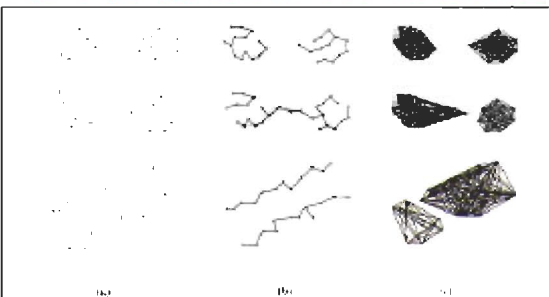


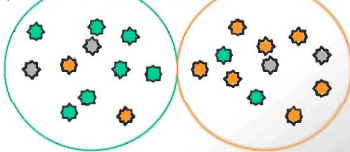
FIG. 4.10. Graphical examples of hierarchical merging. (a) Three data sets. (b) Results of single-link method. (c) Results of complete-link method. (Reproduced with permission from [Duda73]; copyright 1973 by John Wiley & Sons.)

Other clustering methods

- Hierarchical
 - Complete tree structure, but not number of clusters
 - Greedy, hence depends heavily on initial partitioning (two genes that are very similar can be separated)
- K-means clustering
 - Need good estimate of the number of clusters
 - Random assignment of initial centers can bias result
- Parametric methods
 - Model entire density distribution of space. Fit models
- Self-organizing maps
 - Make additional assumptions about geometry of clusters

Evaluating clustering output

- Computing statistical significance of clusters



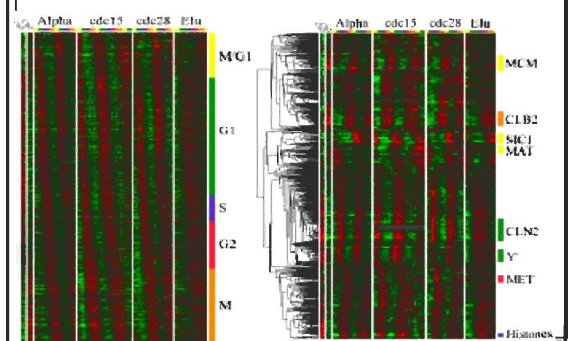
$$P(pos \geq r) = \sum_{m=r}^k \frac{\binom{p}{m} \binom{n}{k-m}}{\binom{N}{k}}$$

p-value of uniformity in computed cluster

- N experiments, p labeled +, (N-p) -
- Cluster: k elements, m positive
- P-value of single cluster containing k elements out of which r are same

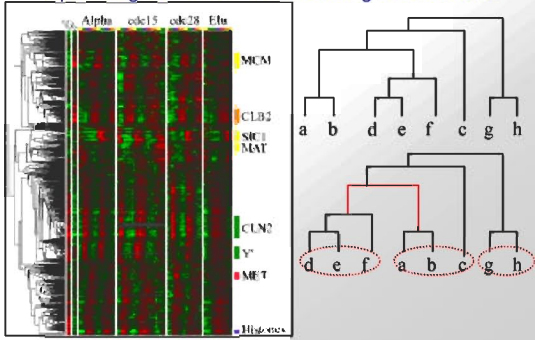
Prob that a randomly chosen set of k experiments would result in m positive and k-m negative

Visualizing clustering output

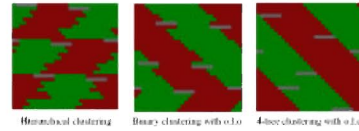


Rearranging tree branches

- Optimizing one-dimensional ordering of tree leaves



$O(n^3)$ algorithm for leaf ordering



- All possible orderings: $O(2^n)$ orderings on n leaves
- Algorithmic improvements:
 - Divide-and-conquer algorithm partitions tree
 - Branch-and-bound allows early termination
 - Works on k -ary trees

Bar-Joseph *et al.* Bioinformatics 2003

What have we learned?

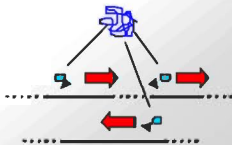
- What have we done?
 - Took expression values
 - Normalized them
 - Clustered them
- What have we obtained?
 - List of possibly co-regulated genes
- What is missing?
 - Causality: Identify transcription factor(s) responsible for the observed co-regulation
 - Molecular basis: what promoter sequences are recognized by these transcription factors

Outline

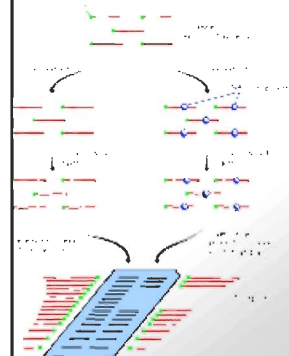
Microarray technology
 Clustering gene expression
TF binding: the controllers
 Bayesian networks
 Network properties
 Scale-free networks

The question

- Which factor binds to which upstream region
 - Gives causality of regulation.
 - Initial networks can be built
- Correlating the binding location data with regulation
 - Able to draw single-link connections

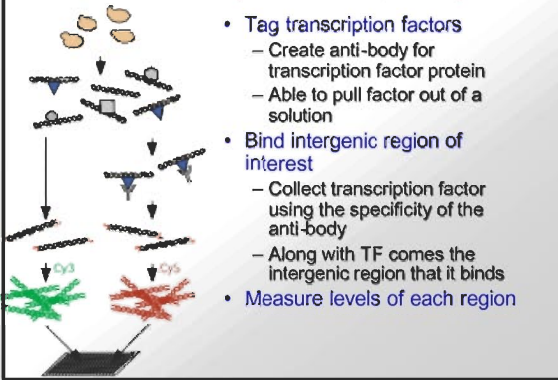


Footprint experiments



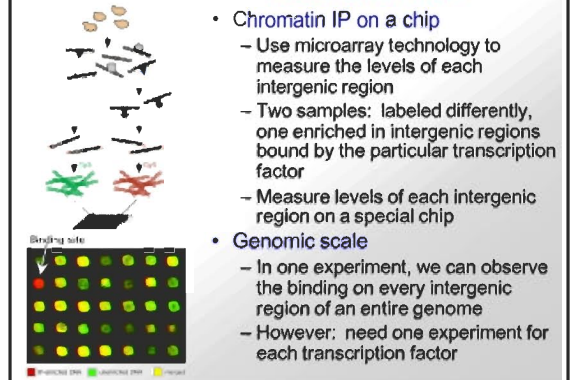
- Most direct observation of binding
 - Protection assay: digest nucleotides that are not protected by the presence of the transcription factor
 - This is how molecular interactions of DNA and regulatory proteins was first described
 - Gives the exact sequence at binding site

Chromatin IP (ImmunoPrecipitation)



- Tag transcription factors
 - Create anti-body for transcription factor protein
 - Able to pull factor out of a solution
- Bind intergenic region of interest
 - Collect transcription factor using the specificity of the anti-body
 - Along with TF comes the intergenic region that it binds
- Measure levels of each region

Location Analysis



- Chromatin IP on a chip
 - Use microarray technology to measure the levels of each intergenic region
 - Two samples: labeled differently, one enriched in intergenic regions bound by the particular transcription factor
 - Measure levels of each intergenic region on a special chip
- Genomic scale
 - In one experiment, we can observe the binding on every intergenic region of an entire genome
 - However: need one experiment for each transcription factor

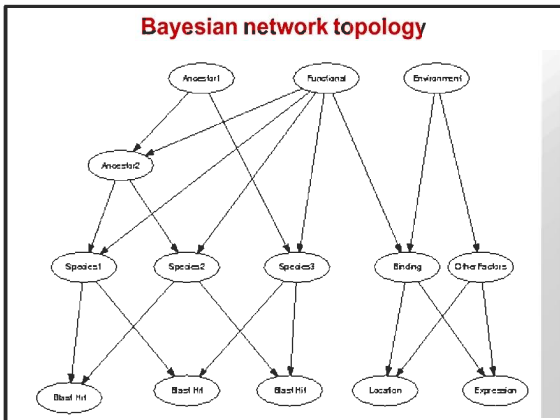
Outline

Microarray technology
 Clustering gene expression
 TF binding: the controllers
Bayesian networks
 Network properties
 Scale-free networks

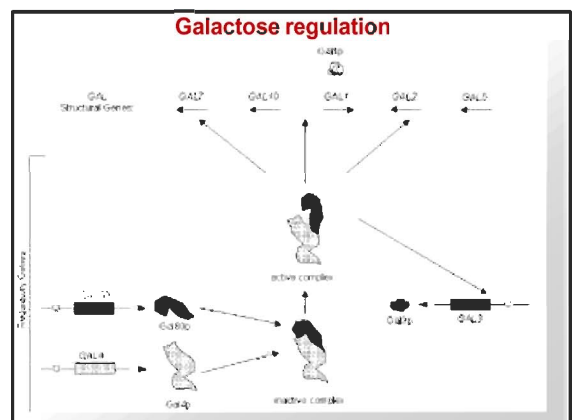
Modeling the dependencies

- Binding and Regulation
 - Regulation data depends on presence binding
 - Location data depends on **binding but also other factors**
- Conservation data
 - Multiple species provide extra predictive power
 - However, species observations are **not independent**
 - Dependencies modeled with a **phylogenetic tree**
- Binding and motif conservation
 - The conservation of a regulatory motif, and the binding of the factor specific to that motif are dependent on functionality of motif
- Environmental factors
 - Binding may occur only in some conditions, not in others

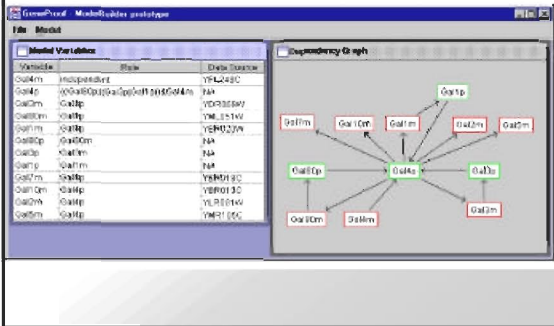
Bayesian network topology



Galactose regulation



Evaluating Alternative Hypotheses



Scoring Bayesian models

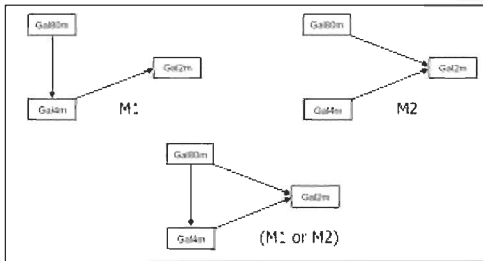
Bayesian scoring metric allows us to compare models with statistical rigor

- Bayesian approach: we score model structure as an ensemble, with a distribution over all possible parameter settings
- Score of model structure S is proportional to average likelihood of observing data D over all possible parameter settings θ :

$$P(D | S) = \int P(D | \theta, S) P(\theta | S) d\theta$$

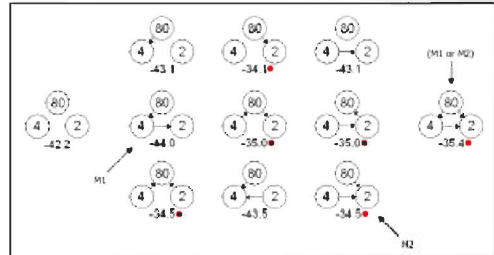
Hartemink et al.

Model Comparison



Scoring all possible models

- Combinatorially many models
- Score variations point to models that best explain data



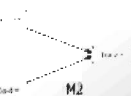
Summary: Inferring regulatory networks



- Expression Clustering**
- Microarray technology allows genome-wide measurements
 - Cluster co-regulated genes according to expression patterns

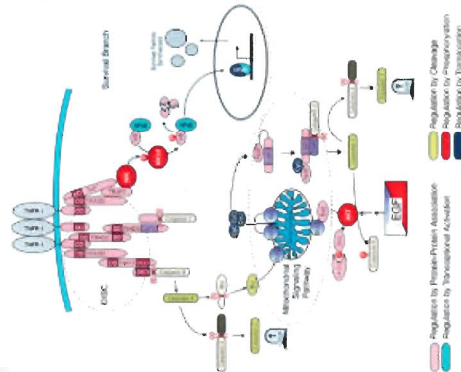


- Location analysis**
- Determine intergenic regions of TF binding
 - Scan identified regions for common motifs



- Bayes Networks**
- Evaluate alternative hypotheses
 - Select network topology

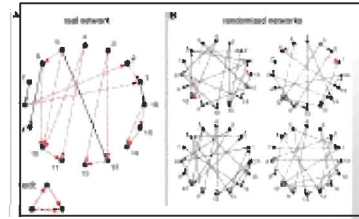
Regulatory Networks: Example



Outline

- Microarray technology
- Clustering gene expression
- TF binding: the controllers
- Bayesian networks
- Network properties**
- Scale-free networks

Recurring network motifs

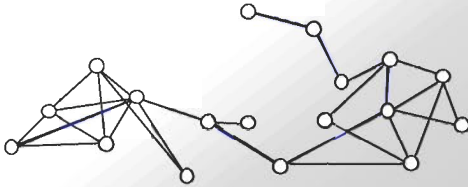


- What are common patterns of interconnectivity?



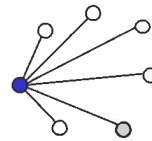
Geodesic distance

- What is the shortest path between any two nodes?
- What is the diameter of the network?
- How many connected components are there?
- What is the size of the largest component?

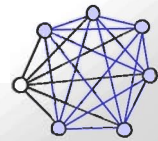


Clustering coefficient

- How likely are my friends to know each other?



C=0

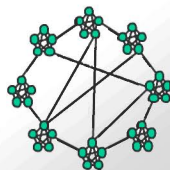


C=1

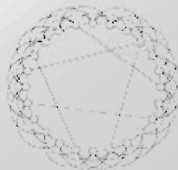
Small world networks

- High clustering coefficient
- Small path lengths

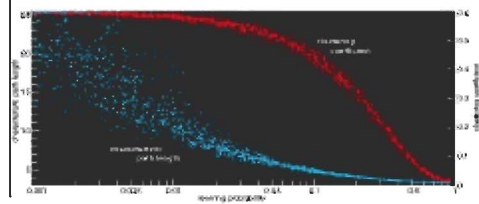
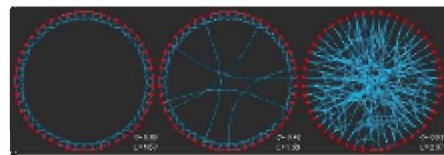
In a highly clustered, ordered network, a single random connection will create a shortcut that lowers L dramatically



Watts demonstrates that small world properties can occur in graphs with a surprisingly small number of shortcuts

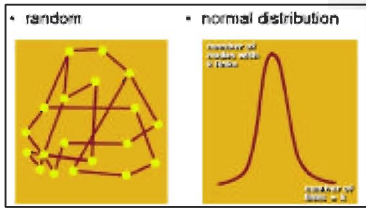


Emergence of small-world phenomenon



Degree distribution

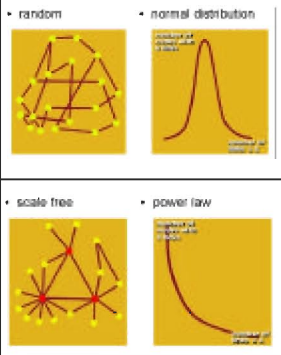
- What's the average number of friends anyone has
- Is this average representative of a typical person



Scale Free Networks

- **DEFINITION:** Scale-free networks, including the Internet, are characterized by an uneven distribution of connectedness. Instead of the nodes of these networks having a random pattern of connections, some nodes act as "very connected" hubs, a fact that dramatically influences the way the network operates.
- Barabasi and his colleagues mapped the connectedness of the Web.
- Their experiment yielded a connectivity map that they called "scale-free".

Scale Free vs. Random



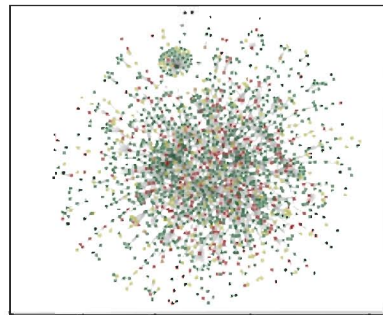
Random networks suffer from random failures because each node important as any other

"scale free" networks are more immune to random failure due to the redundancy of paths linking nodes

connectivity ensured by few highly connected nodes

"scale free" networks are prone to catastrophic failure when key "hubs" are attacked

Yeast protein-protein interaction network

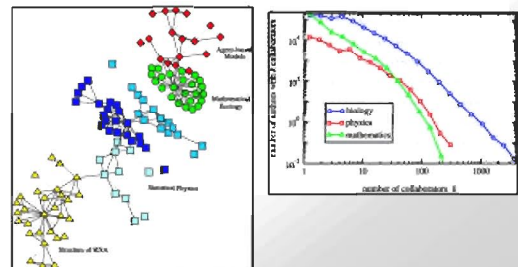


- High clustering coefficient / Short paths

Scale-free networks are ubiquitous

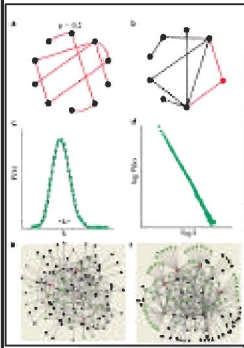
- Web pages
- Internet routers
- Airports
- Power grid
- Social networks
- Boards of directors
- Scientific co-authorship
- Medline citations
- US patents
- Movie database
- Metabolic networks
- Protein interactions
- Regulatory networks
- Predator-prey networks
- Neuron connections
- Blood vessels

Scientific authorship



- Hubs play central role in network connectivity
- Small number of cross-cluster interconnections

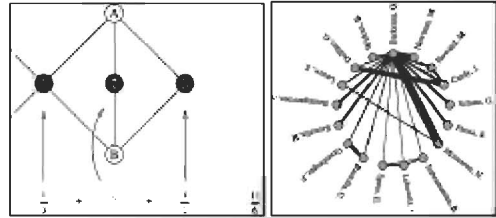
How do scale-free networks emerge?



- (a) constructed by laying down N nodes and connecting each pair with probability p . This network has $N = 10$ and $p = 0.2$.
- (b) A new node (red) connects to two existing nodes in the network (black) at time $t + 1$. This new node is much more likely to connect to highly connected nodes, a phenomenon called **preferential attachment**.
- (c) The network connectivity can be characterized by the probability $P(k)$ that a node has k links. For random graphs $P(k)$ is strongly peaked at $k = \langle k \rangle$ and decays exponentially for large k .
- (d) A scale-free network does not have a peak in $P(k)$, and decays as a power law $P(k) \sim k^{-g}$ at large k .
- (e) A random network - most nodes have approximately the same number of links.
- (f) The majority of nodes in a scale-free network have one or two links, but a few nodes (hubs) have a large number of links; this guarantees that the system is fully connected.

Scale-free networks from bi-partite graphs

- Person belongs to multiple social groups
- Protein acts in multiple functional categories
- Author publishes to multiple fields



- Loose connections from group membership

Implications of scale-free networks

- Hubs become important
 - Random networks are subject to random failures
 - Scale-free networks are unlikely to lose a hub
 - Scale-free networks subject to directed attacks
- Biological implication
 - Essential proteins in yeast often correspond to hubs

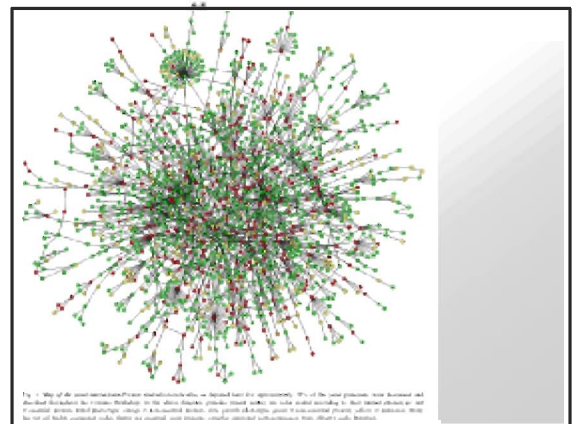
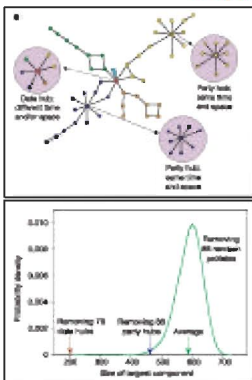


Fig. 1. Map of all yeast interactions. Proteins are color-coded by the approximately 30% of the yeast proteome that has been identified as essential for growth. Proteins that are essential for growth are shown in red, while non-essential proteins are shown in green. The graph is a scale-free network, with a few hubs and many tails.

Two types of hubs



- "Date hubs"
 - Interconnections at different times
- "Party hubs"
 - Interconnections are coordinated
- Different effects on network connectivity
 - Date hubs bring together distinct components of the network

Outline

- Microarray technology
- Clustering gene expression
- TF binding: the controllers
- Bayesian networks
- Network properties
- Scale-free networks