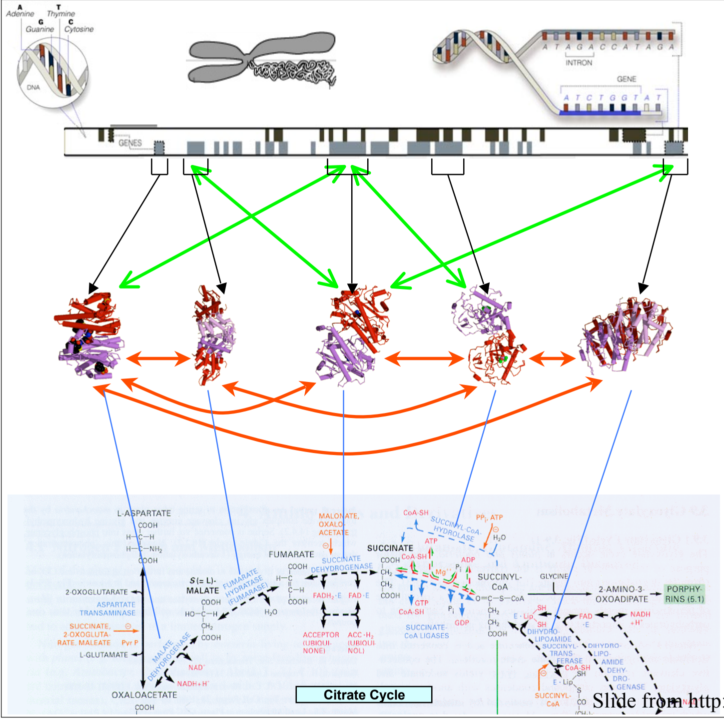
Bioinformatics 2

From genomics & proteomics to biological networks

Aims

- Briefly review functional genomics
- Biological Networks in general
- Genetic Networks
- Briefly review proteomics
- Protein Networks



GENOME

protein-gene interactions

PROTEOME

protein-protein interactions

METABOLISM

Bio-chemical reactions

Slide from http://www.nd.edu/~networks/

Biological Profiling

- Microarrays
  - cDNA arrays
  - oligonucleotide arrays
  - whole genome arrays
- Proteomics
  - yeast two hybrid
  - PAGE techniques

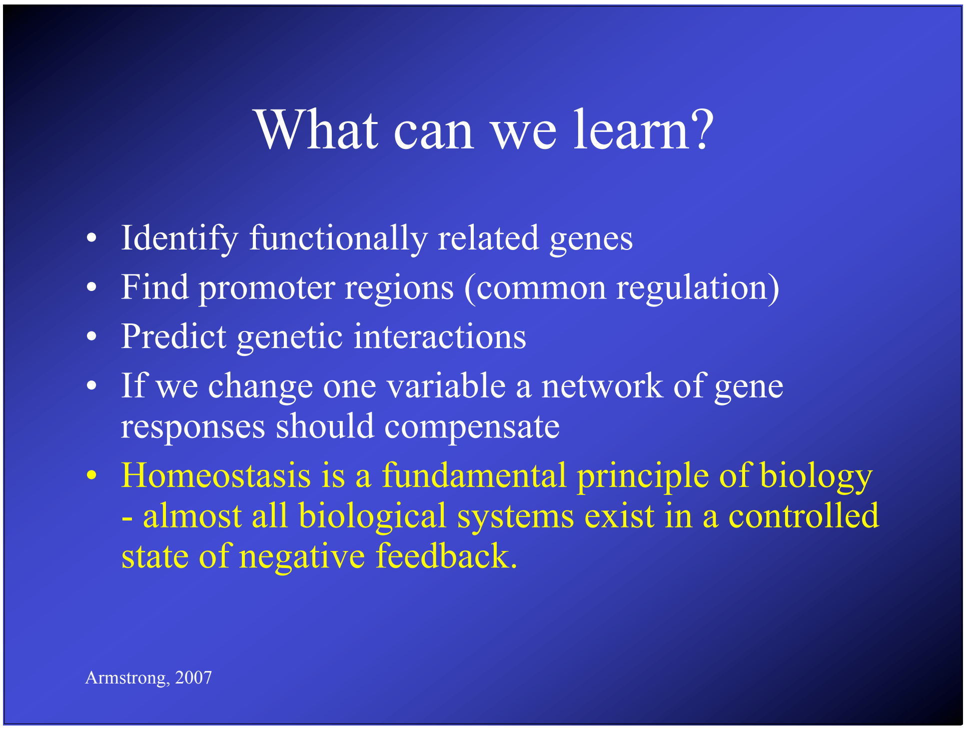
Armstrong, 2007

Why microarrays?

- What genes are expressed in a tissue and how does that tissue respond to one of a number of factors:
  - change in physical environment
  - experience
  - pharmacological manipulation
  - influence of specific mutations

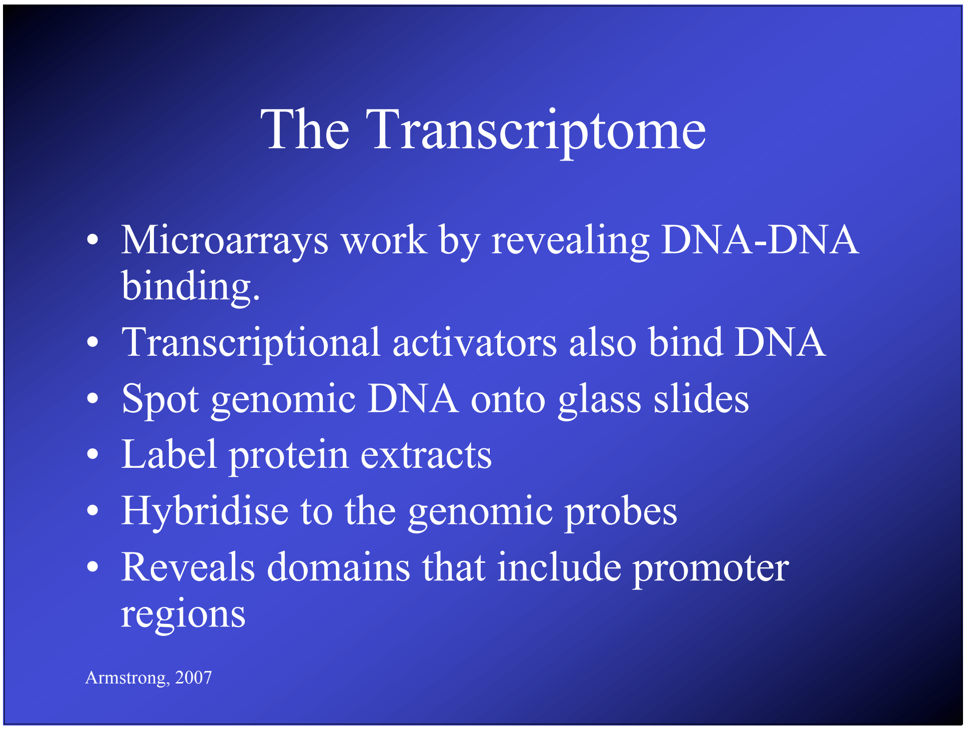
What do we actually get?

- A snap-shot of the mRNA profile in a biological sample
- With the correct experimental conditions we can compare two situations
- Not all biological processes are regulated through mRNA expression levels



What can we learn?

- Identify functionally related genes
- Find promoter regions (common regulation)
- Predict genetic interactions
- If we change one variable a network of gene responses should compensate
- Homeostasis is a fundamental principle of biology - almost all biological systems exist in a controlled state of negative feedback.



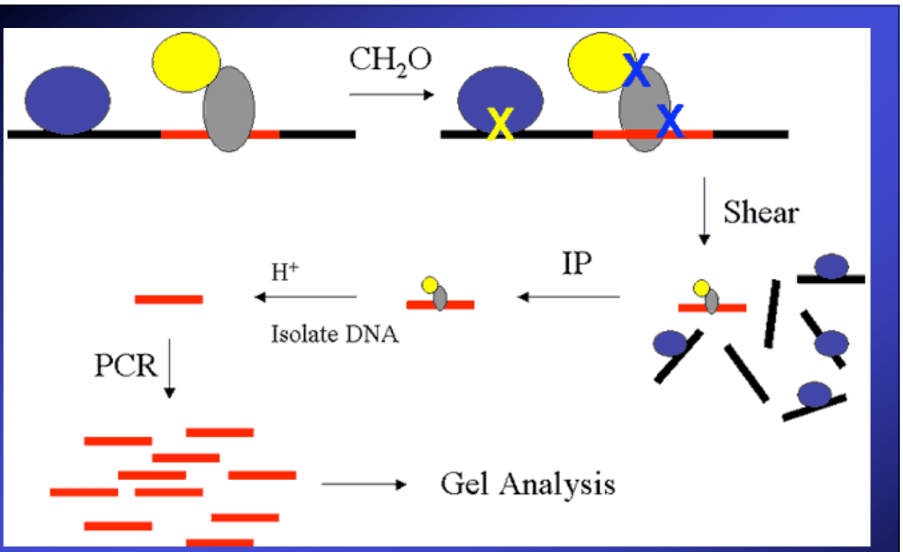
The Transcriptome

- Microarrays work by revealing DNA-DNA binding.
- Transcriptional activators also bind DNA
- Spot genomic DNA onto glass slides
- Label protein extracts
- Hybridise to the genomic probes
- Reveals domains that include promoter regions

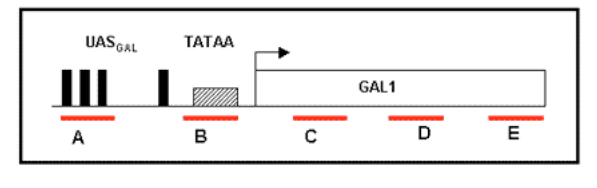
ChIP to Chip

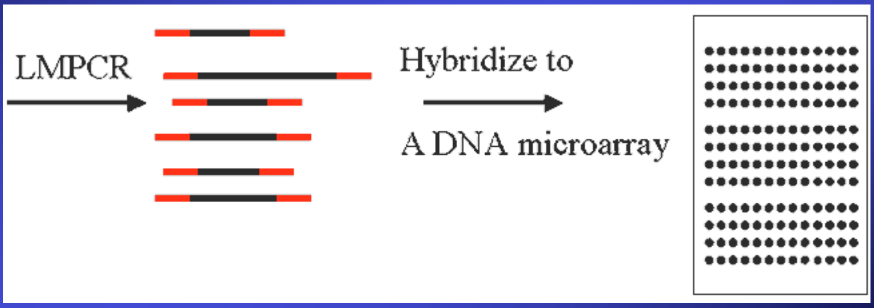
Chromatin Immunoprecipitation to Microarray (Chip)

Protein-DNA interactions
de-novo prediction has many false positives
Which DNA sites do actually bind a specific TF?
Requires an antibody to the protein



http://proteomics.swmed.edu/chiptochip.htm





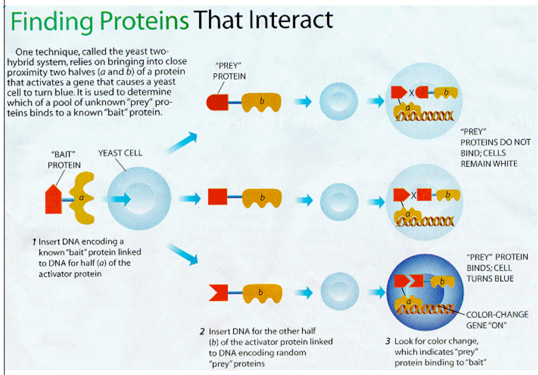
http://proteomics.swmed.edu/chiptochip.htm

Biological networks

Yeast protein network

Nodes: proteins

Links: physical interactions (binding)



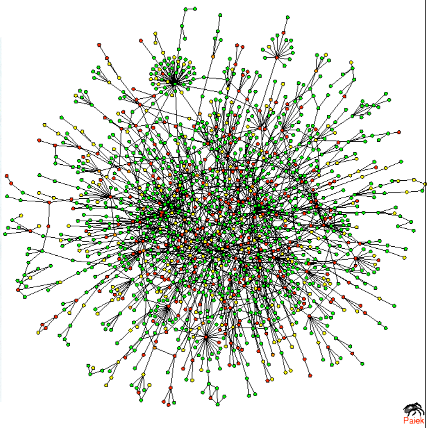
Finding Proteins That Interact

One technique, called the yeast two-hybrid system, relies on bringing into close proximity two halves (*a* and *b*) of a protein that activates a gene that causes a yeast cell to turn blue. It is used to determine which of a pool of unknown "prey" proteins binds to a known "bait" protein.

1 Insert DNA encoding a known "bait" protein linked to DNA for half (*a*) of the activator protein

2 Insert DNA for the other half (*b*) of the activator protein linked to DNA encoding random "prey" proteins

3 Look for color change, which indicates "prey" protein binding to "bait"



P. Uetz, *et al.* *Nature* **403**, 623-7 (2000).

Slide from http://www.nd.edu/~networks/

Building networks…

- Biological Networks
  - Random networks
  - Scale free networks
  - Small worlds
- Metabolic Networks
- Proteomic Networks
- The Mammalian Synapse
- Other synapse models?

Armstrong, 2007

Biological Networks

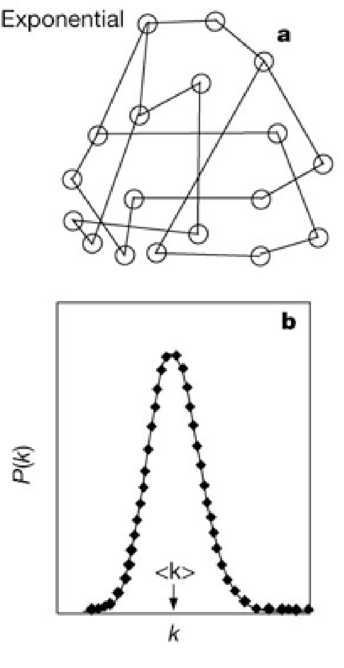
- Genes - act in cascades
- Proteins - form functional complexes
- Metabolism - formed from enzymes and substrates
- The CNS - neurons act in functional networks
- Epidemiology - mechanics of disease spread
- Social networks - interactions between individuals in a population
- Food Chains

Protein Interactions

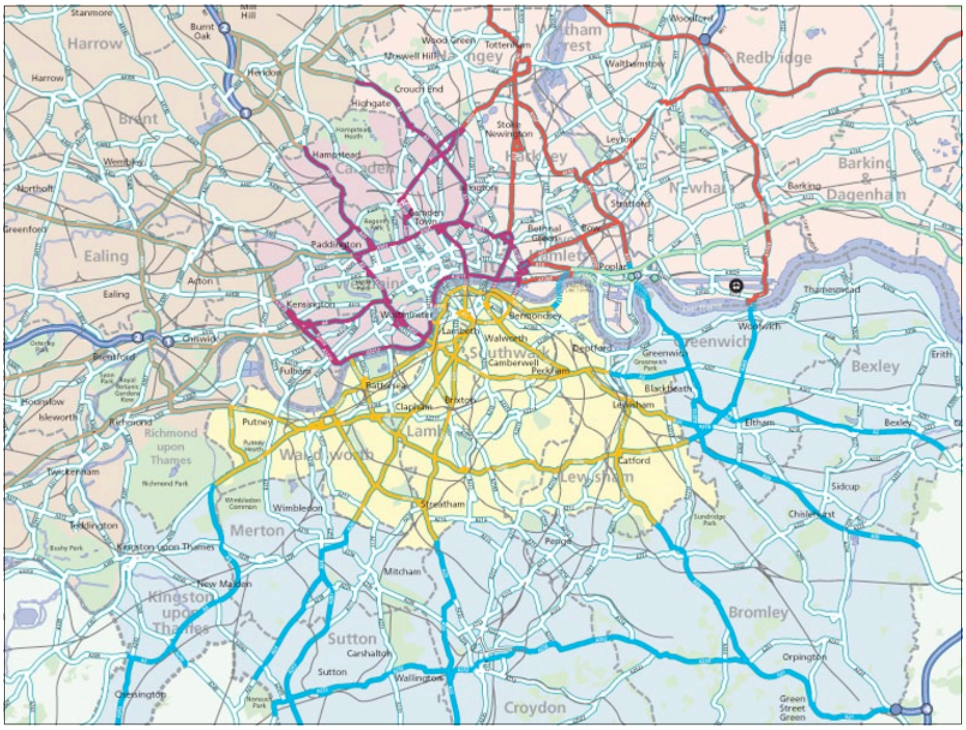
- Individual Proteins form functional complexes
- These complexes are semi-redundant
- The individual proteins are sparsely connected
- The networks can be represented and analysed as an undirected graph

Large scale organisation

- Networks in biology generally modeled using classic random network theory.
- Each pair of nodes is connected with probability *p*
- Results in model where most nodes have the same number of links <*k*>
- The probability of any number of links per node is $P(k) \approx e^{-k}$



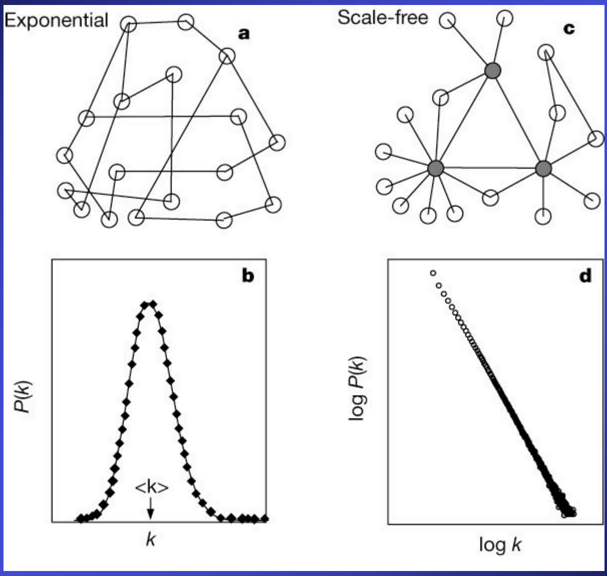
Armstrong, 2007



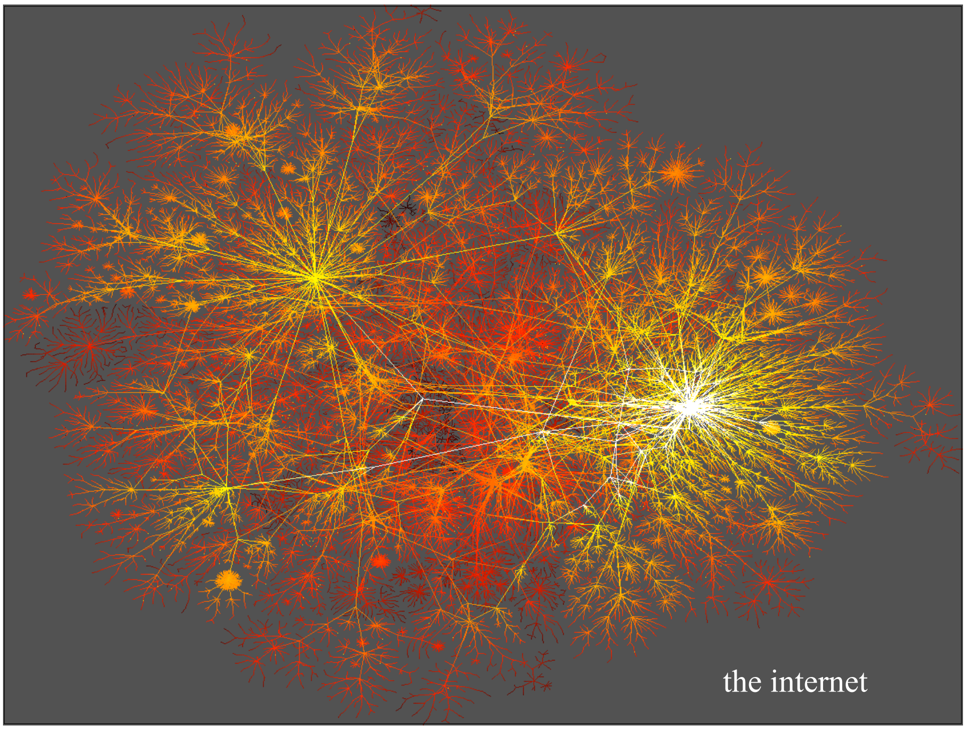
Non-biological networks

- Research into WWW, internet and human social networks observed different network properties
  - ‘Scale-free’ networks
  - P(k) follows a power law: P(k)≈$k^{-\gamma}$
  - Network is dominated by a small number of highly connected nodes - hubs
  - These connect the other more sparsely connected nodes

Armstrong, 2007



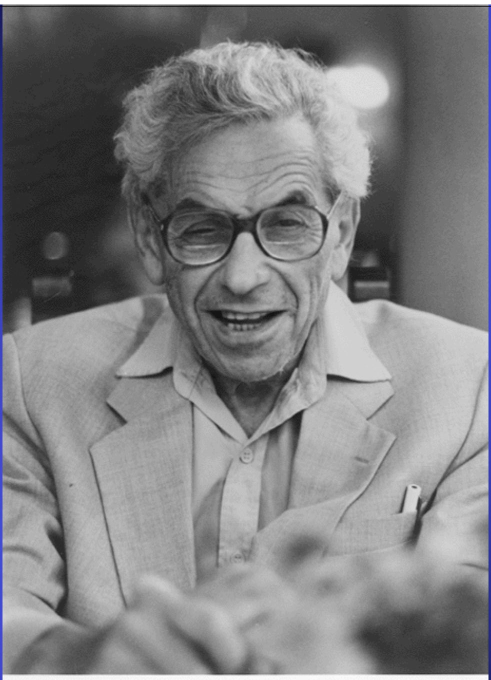
Armstrong, 2007



the internet

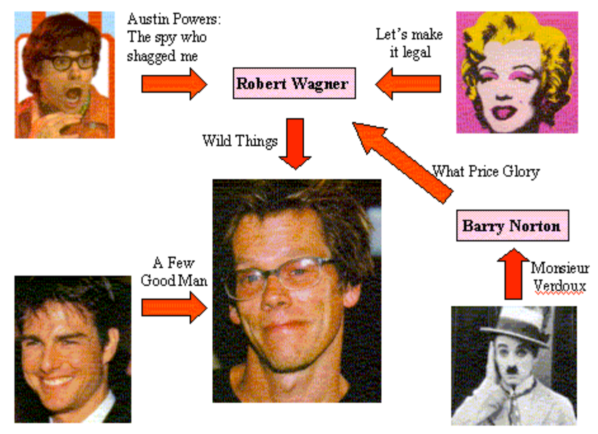
Small worlds

- General feature of scale-free networks
  - – any two nodes can be connected by a relatively short path
  - – average between any two people is around 6
    - What about SARS???
  - – 19 clicks takes you from any page to any other on the internet.



Paul Erdős, the most prolific mathematician who ever lived, has no home and no job, but he has wandered the world for over fifty years, inspiring other mathematicians. From the documentary **N is a Number: A Portrait of Paul Erdős** © 1993 by George Csicsery

Armstrong, 2007



Austin Powers: The spy who shagged me

Robert Wagner

Let's make it legal

Wild Things

What Price Glory

Barry Norton

A Few Good Man

Monsieur Verdoux

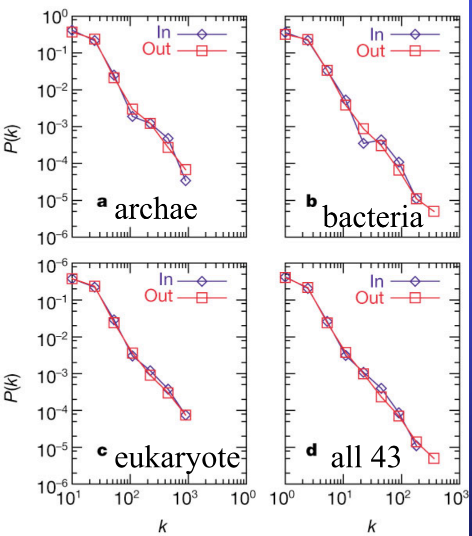
Armstrong, 2007

Biological organisation

Jeong et al., 2000 The large-scale organisation of metabolic networks. Nature 407, 651-654

- Pioneering work by Oltvai and Barabasi
- Systematically examined the metabolic pathways in 43 organisms
- Used the WIT database
  - 'what is there' database
  - http://wit.mcs.anl.gov/WIT2/
  - Genomics of metabolic pathways

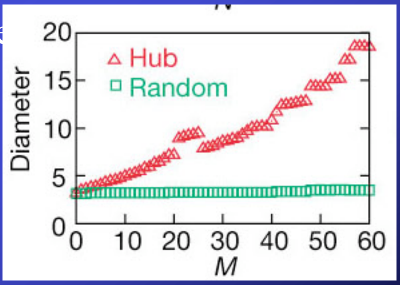
Using metabolic substrates as nodes



=scale free!!!

Random mutations in metabolic networks

- Simulate the effect of random mutations or mutations targeted towards hub nodes.
  - Measure network diame
  - Sensitive to hub attack
  - Robust to random



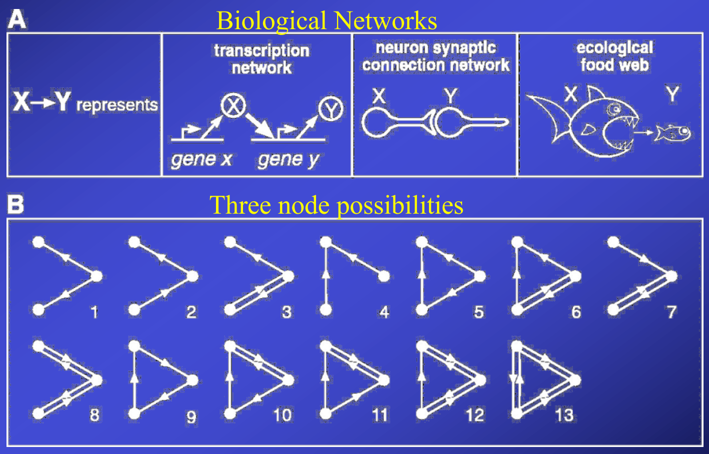
Consequences for scale free networks

- Removal of highly connected hubs leads to rapid increase in network diameter
  - Rapid degeneration into isolated clusters
  - Isolate clusters = loss of functionality
- Random mutations usually hit non hub nodes
  - therefore robust
- Redundant connectivity (many more paths between nodes)

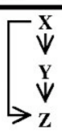
Network Motifs

- Do all types of connections exist in networks?
- Milo et al studied the transcriptional regulatory networks in yeast and E.Coli.
- Calculated all the three and four gene combinations possible and looked at their frequency

Milo et al. 2002 Network Motifs: Simple Building Blocks of Complex Networks. Science 298: 824-827

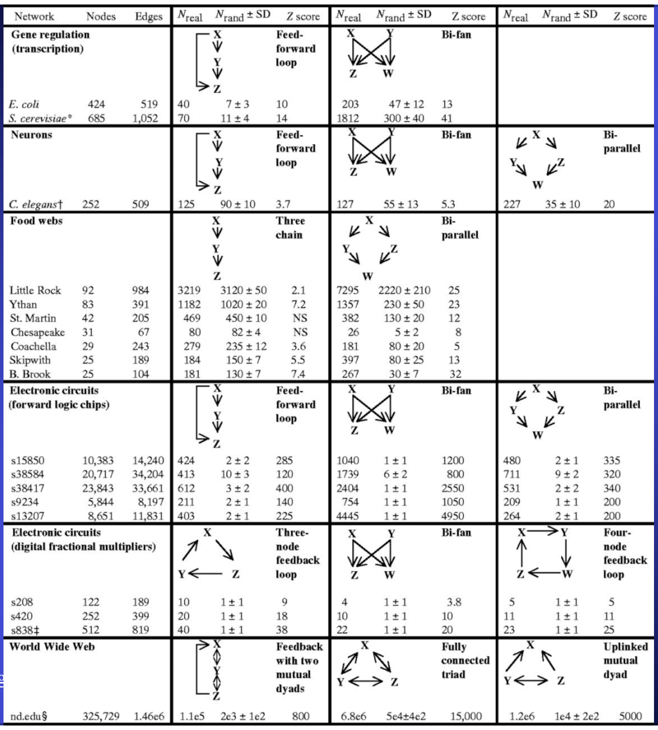


Gene sub networks

Network	Nodes	Edges	N_{real}	N_{rand} ± SD	Z score	N_{real}	N_{rand} ± SD	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41

Heavy bias in both yeast and E.coli towards these two sub network architectures

Armstrong, 2007



Network	Nodes	Edges	N_{real}	N_{rand} ± SD	Z score	N_{real}	N_{rand} ± SD	Z score	N_{real}	N_{rand} ± SD	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			
Electronic circuits (forward logic chips)					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

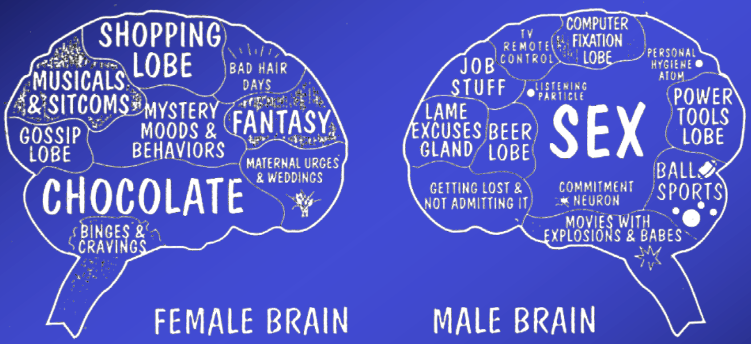
Armstrong

Gene networks and network inference

What is a gene network

- Genes do not act alone.
- Gene products interact with other genes
  - Inhibitors
  - Promoters
- The nature of genetic interactions in complex
  - Takes time
  - Can be binary, linear, stochastic etc
  - Can involve many different genes

What makes boys boys and girls girls?



Sugar, Spice and synthetic Oestrogens?

Armstrong, 2007

Sex determination: a gene cascade (in flies…)

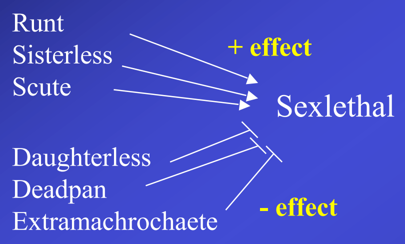
6 Genes detect X:A ratio

	Females	Males
Runt Sisterless Scute	↑	↓
Daughterless Deadpan Extramachrochaete	↓	↑

Armstrong, 2007

Sex determination
(in flies…)

6 Genes regulate ‘Sexlethal’



Armstrong, 2007

Sex determination
(in flies…)

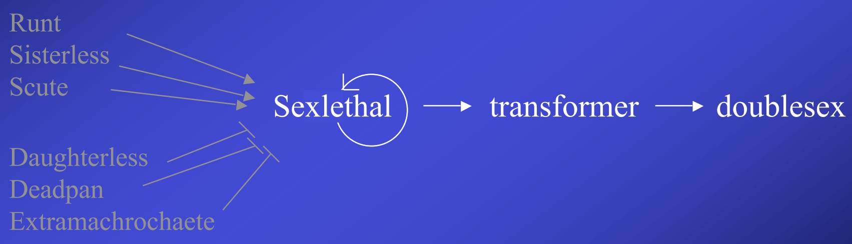
Sexlethal can then regulate itself...

Armstrong, 2007

Sex determination

(in flies…)

Downstream cascade builds...



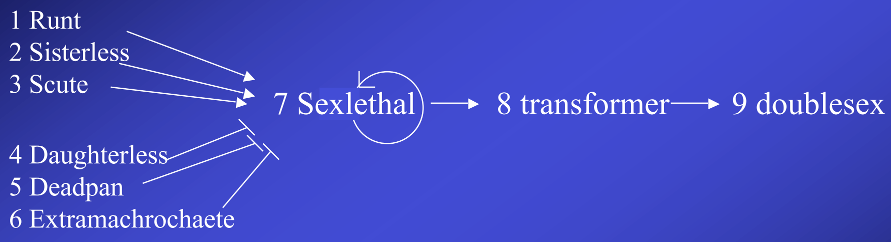
Runt
Sisterless
Scute

Daughterless
Deadpan
Extramachrochaete

Sexlethal → transformer → doublesex

Armstrong, 2007

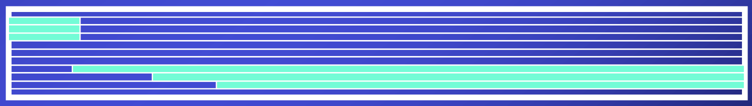
Gene expression and time



1 Runt
2 Sisterless
3 Scute

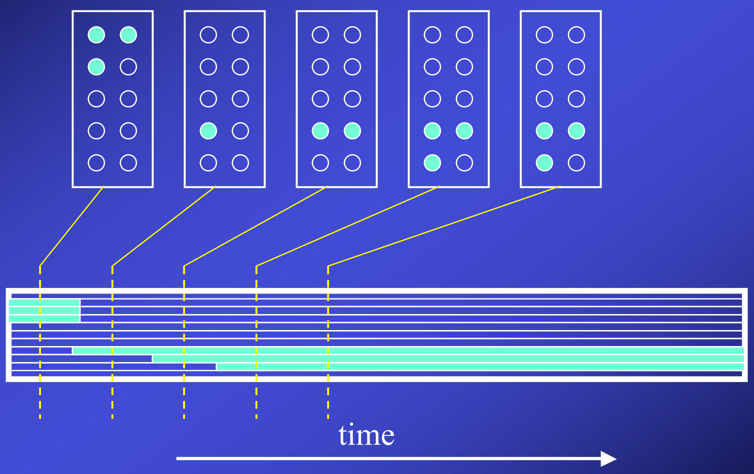
4 Daughterless
5 Deadpan
6 Extramachrochaete

7 Sexlethal → 8 transformer → 9 doublesex



Armstrong, 2007

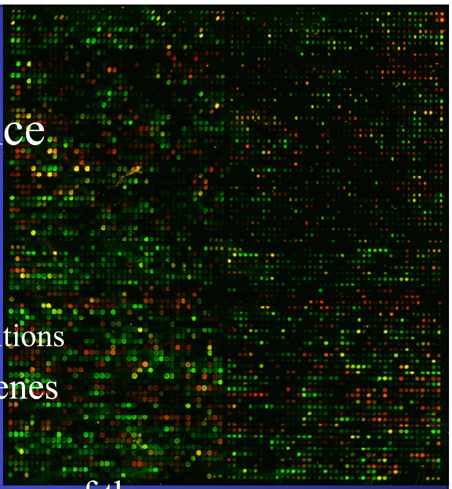
Gene microarrays



Gene Network Inference

- Gene micro-array data
- Learning from micro-array data
- Unsupervised Methods
- Supervised Methods
- Edinburgh Methods

Gene Network Inference



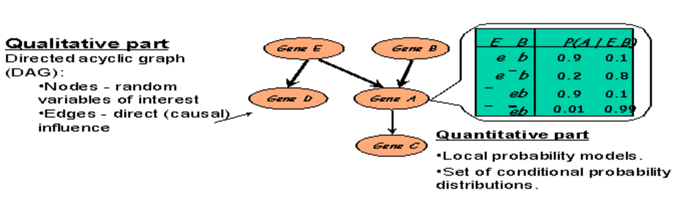
- Gene micro-array data
  - Time Series array data
  - Tests under ranges of conditions
- Unlike example - 1000s genes
- Lots of noise
- Clustering would group many of these genes together
- Aim: To infer as much of the network as possible

Learning from Gene arrays

- Big growth industry but difficult problem
- Initial attempts based on unsupervised methods:
  - Basic clustering analysis - related genes
  - Principal Component Analysis
  - Self Organising Maps
  - Bayesian Networks

Bayesian 'gene' networks

- Developed by Nir Friedman and Dana Pe'er
- Can be easily adapted to a supervised method



Qualitative part

Directed acyclic graph (DAG):

- Nodes - random variables of interest
- Edges - direct (causal) influence

E	B	P(A | E,B)	
e	b	0.9	0.1
e	$\bar{b}$	0.2	0.8
$\bar{e}$	b	0.9	0.1
$\bar{e}$	$\bar{b}$	0.01	0.99

Quantitative part

- Local probability models.
- Set of conditional probability distributions.

Learning Gene Networks

- The field is generally moving towards more supervised methods:
  - Bayesian networks can use priors
  - Support Vector machines
  - Neural Networks
  - Decision Trees

Can we combine network knowledge with gene inference?

- Scale free architecture
  - Chance of new edges is proportional to existing ones
  - Highly connected nodes may well be known to be lethal

- Network motifs
  - Constrain the types of sub networks

- Prior Knowledge
  - Many sub networks already known

Conclusions

- Gene network analysis is a big growth area
- Several promising fields starting to converge
  - Complex systems analysis
  - Using prior knowledge
  - Application of advance machine learning algorithms
  - AI approaches show promise