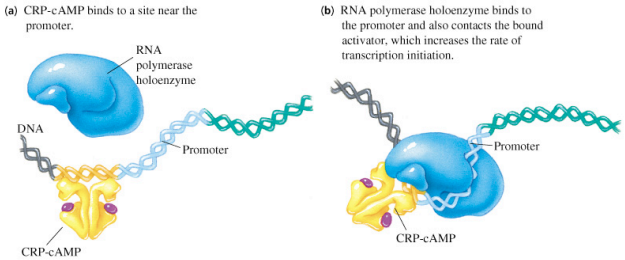
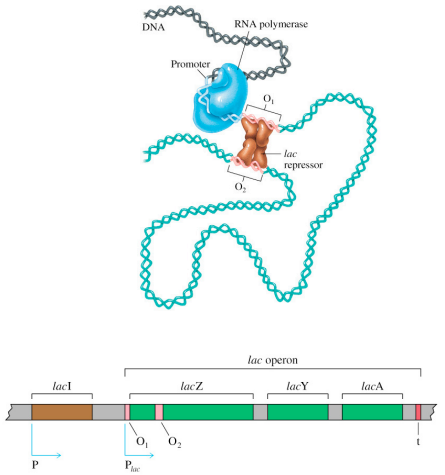
Genetic Networks

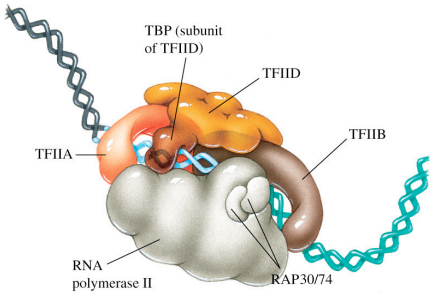
Transcription Regulation

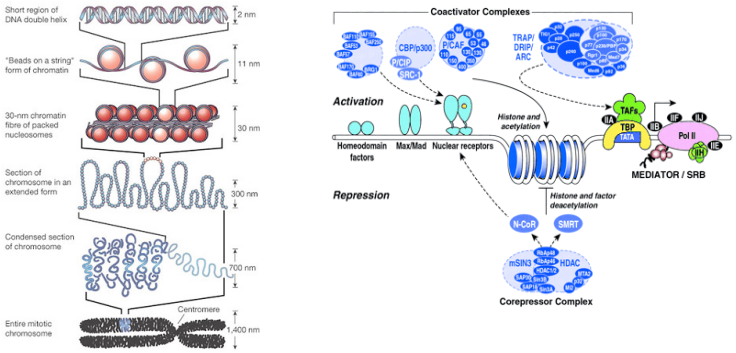
- Interaction between a protein and a specific sequence in the DNA
- The set of proteins that binds the promoter region of a gene will determine its expression
  - In which tissue
  - En which developmental phase
  - Under which environmental conditions
  - etc.



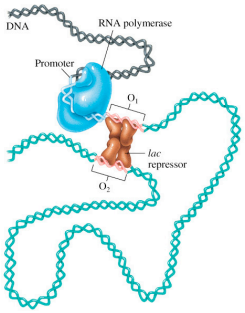


RNA polymerase II Eukaryots

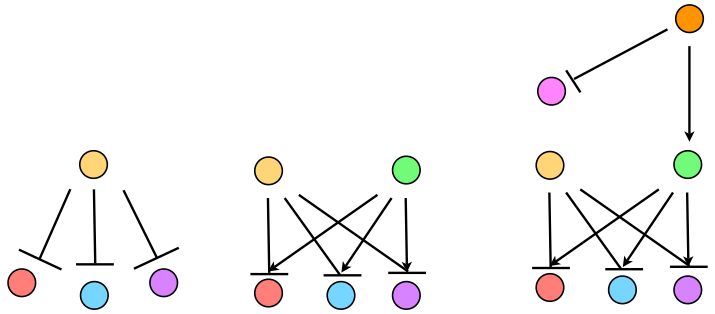


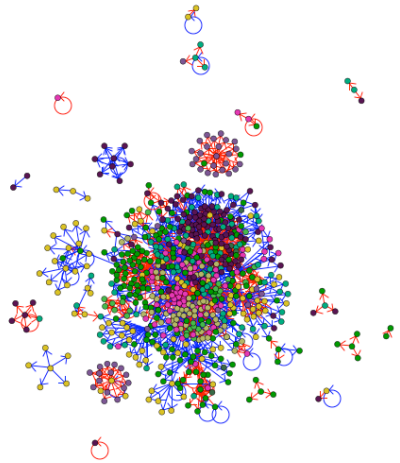


Regulatory Networks



Regulatory Networks





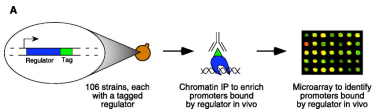
Other sources of regulation

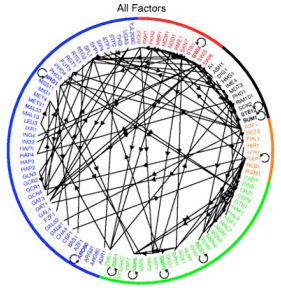
- Elongation
- mRNA stability
- Micro-RNAs
- etc.

Sources of high throughput experimental data

- Chip-on-Chip
- STAGE/SABE
- DNA-arrays
- Prediction
- Text-mining

Chip-On-Chip

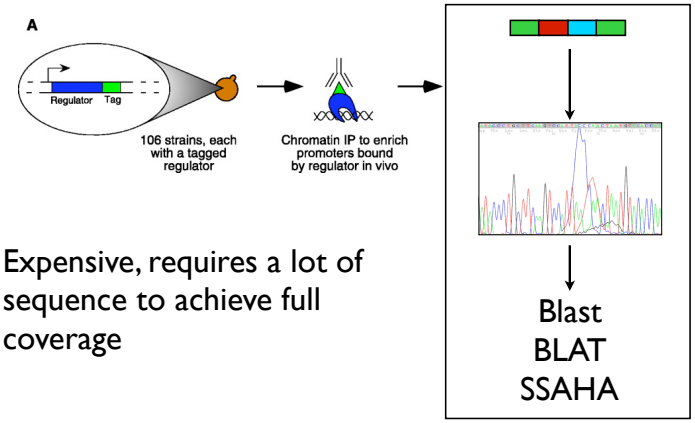




Chip-On-Chip II

- PCR Arrays
  - low resolution
- Oligo Array
  - very expensive
- Normally only regions around genes are included

STAGE/SABE



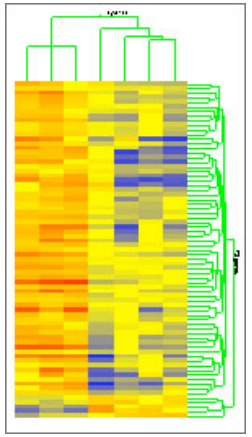
Expensive, requires a lot of sequence to achieve full coverage

Issues

- Depends on the experimental conditions
  - good: context
  - bad: we can not cover all conditions
- We know TF are bound, but don't know if they are active, or how they affect gene transcription

DNA arrays

- Relation between TF expression and other genes.
  - bayesian network, Petri networks, Mutual Information
- So far, only applicable to small subsystems
- Ambiguous, various solutions
- They can be improved with additional info, TF, TFBS, etc.



Prediction

- "pattern matching","pattern discovery"
- Noisy, lots of false positives
- Only Binding sites are predicted, no the time or the conditions, or the action.
- They can be combined with high-throughput experiments.

- Know Sites: "Pattern Matching"
  - We known the pattern a TF is binding:
  - want to know where in the genome
- Unknown Sites: "Pattern Discovery"
  - We know a set of genes that are co-regulated?
  - can we predict the DNA sequences involved?

Pattern Matching

- How to describe a set of binding sites
  - Consensus sequence
  - patterns
  - weight matrices (PSSM)

Consensus Sequence

ATCGTGCTATAGGTAAGT
ATCGTGGTATACGTAAGT
ATCGTGCTTTAGGTAAGA
ATCCTGCTATTGCTAAGT

ATCGTGCTATAGGTAAGT

Consensus Sequence

ACGTA

CGACGTAGATGACCTACGGATGCACGAACG

CGACGTAGATGACCTACGGATGCACGAACG

CGACGTAGATGACCTACGGATGCACGAACG

CGACGTAGATGACCTACGGATGCACGAACG

Patterns

ATCGTGCTATAGGTAAGT
ATCGAGGTATAGGTAAGT
ATGGGGCTACAGGTAAGA
ATGGCGCTATTGGTAAGT

AT[CG]G[ACGT]G[CG]TA[CT][AT]GGTAAG[AT]

ATSGNGSTAYWGGTAAGW

W= A or T
R= A or G
K= G or T
S= C or G
Y= C or T
M= A or C

Matrices

A T C G T G C T A T A G G T A A G T
A T C G T G G T A T A C G T A A G T
A T C G T G C T T T A G G T A A G A
A T C C T G C T A T T G C T A A G T

A 4 0 0 0 0 0 0 0 3 0 3 0 0 0 4 4 0 1
C 0 0 4 1 0 0 3 0 0 0 0 0 1 1 0 0 0 0 0
G 0 0 0 3 0 4 1 0 0 0 0 3 3 0 0 0 4 0
T 0 4 0 0 0 4 0 0 4 1 4 1 0 0 4 0 0 0 3

A T G G C T C G A T T G G T A T G T
4+4+0+3+0+4+3+0+3+4+1+3+3+4+4+0+4+3=47

T A G C C A G T T T A T T A G C G T
0+0+0+1+0+0+3+4+1+4+3+0+0+0+0+0+4+3=23

LLR Matrices

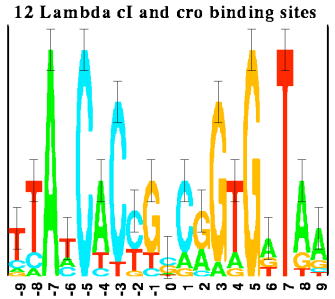
```
A  3  0  0  0  0  0  0  0
C  0  0  4  1  0  0  3  0
G  0  0  0  3  4  0  1  0
T  1  4  0  0  0  4  0  4

A .25
C        1 .25      .25
G          .75 1    .75
T .75 1            1     1
```

$LLR=\log(f_{i\alpha}/f_{0\alpha})$; $f_{0a}=f_{0t}=0.4; f_{0c}=f_{0g}=0.6$

```
A log(0.25/0.4)
C log(0/0.6)   <- pseudocounts
G log(0/0.6)      big negative number
T log(0.75/0.6)
```


Logos

- Graphic representation of an alignment
- Base height is proportional to frequency
- Total height of the column show how conserved it is



http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html
http://weblogo.berkeley.edu/logo.cgi

So far...

- All start with an alignment
- indels are not allowed
- Position-independence assumed

HMM and Neural Networks

Special Binding Sites

- Promoters
  - http://www.fruitfly.org/seq_tools/promoter.html
  - http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb
  - http://www.cbs.dtu.dk/services/Promoter/
- Terminators
  - http://www.softberry.com/berry.phtml?topic=findterm&group=programs&subgroup=gfindb

Matrices Databases

TRANSFAC

- Transcription Factors, binding sites and their matrices (eukaryots)
  - Other realted resuources
    - PathoDB: a database on pathologically relevant mutated forms of transcription factors and transcription factor binding sites
    - S/Mart:collects information about scaffold/matrix attached regions and the nuclear matrix proteins
    - Transcompel: is a database on composite regulatory elements affecting gene transcription in eukaryotes
    - More...

http://www.gene-regulation.com/

RegulonDB

- Transcription Factors, binding sites and operons in *E. coli*
- Visualization and analysis tools
- Integrated in Ecocyc (www.ecocyc.org)

http://www.cifn.unam.mx/Computational_Genomics/regulondb/

Computational Biology and Bioinformatics-CSHL

- TRED: Human and mouse
- CEPDB: C. elegans
- SCPD: Yeast
- Promoters, TF binding sites & matrices

http://rulai.cshl.edu/software/index1.htm

Pattern Discovery

Finding Unknown binding sites

- A set of (supposedly)co-regulated genes
- Take their promoter region
  - Bacteria: 50-300 bp of intergenic region
  - Eukaryot: 1000 - 4000 bp
- Search what they have in common

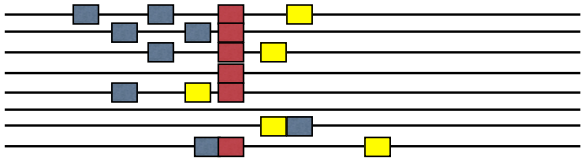
Co-regulated Genes

- Microarrays
- Any other association:
  - Same metabolic Pathway
  - Same functional Class
  - Similar names

phylogenetic footprints

- Use a set of orthologous genes
- Regulation and binding sites are conserved
  - Organisms to far apart: no conservation
  - Organisms to close: sequences haven't diverged enough

Once you have your promoter regions, how to find what is common?



Methods

- Over-represented Motives
  - expensive
  - exhaustive
  - noisy
- Matrix-based methods
  - *Gibbs Sampling*
    - fast
    - non-exhaustive: it can give different results each run
    - http://bayesweb.wadsworth.org/gibbs/gibbs.html
- Symmetrical Motives
  - inverted/direct repeats

Over-represented Motives

- Count the frequency of each n-length word
- Find the word significantly more abundant in our seqeunce set
- you need a good *background* (HMM)

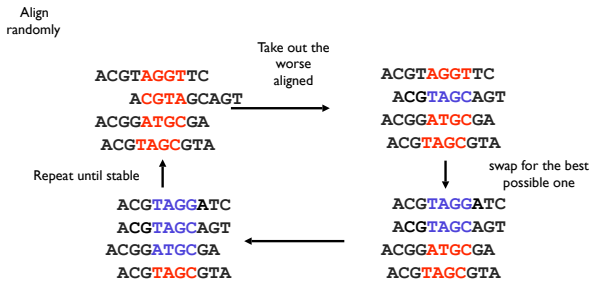
Over-represented Motives

Word (n=5)	expected	Observed
AAAAA	2	3
AAAAC	3	2
AAAAG	5	3
...		
ATGCA	13	17
ATGCC	15	75
ATGCG	17	14
...		
TTTTG	5	3
TTTTT	2	0

4^5=1024, but...

4^{12}=16.777.216

Gibbs Sampling



Clustering Sites

- Overlapping sites

```
AAGTCGGC
 AGTCGGCT
  GTCGGCTT
------------
AAGTCGGCTT
```

- Similar sites

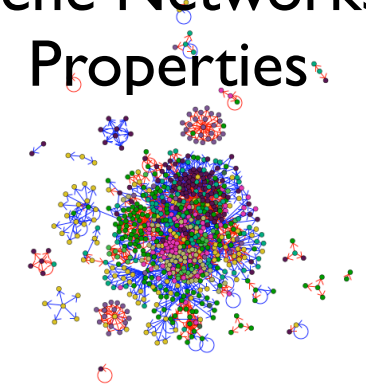
```
AGTCGGCA
AGTCGGCT
TGTCGGCT
------------
AGTCGGCT
```

- Distance
  - Euclidean
  - Pearson correlation
- Clustering
  - k-means
  - Hierarchical clusterin

Software

- Consensus Package
  - Consensus: Pattern discovery (fixed length)
  - WConsensus: Pattern discovery (unknown length)
  - Patser: Pattern Matching
  - http://ural.wustl.edu/software.html
- MEME/MAST
  - MEME: Patter Discovery
  - MAST: Patter Matching
  - http://meme.sdsc.edu/meme/intro.html
- Gibbs Samplers:
  - AlignACE: http://atlas.med.harvard.edu/
  - MotifSampler: http://homes.esat.kuleuven.be/~thijs/Work/MotifSampler.html

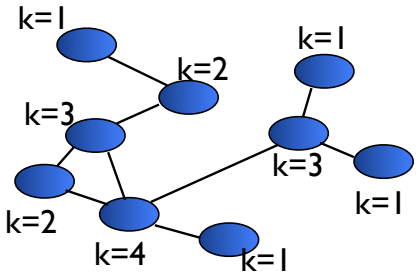
Gene Networks Properties



Characterizing a Network

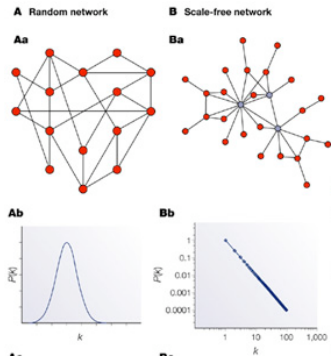
- Degree (k)
- Degree distribution (k/p(k))
- *Degree exponent* (γ)
- Shorter Paths (l)
- Average Lenght of the paths (<l>)
- Clustering coefficient (c)
- Average Clustering Coefficient (<c>)

Degree



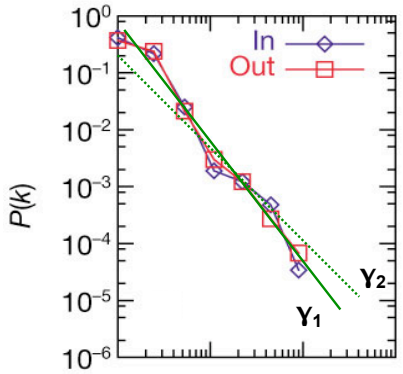
k	*p(k)*
1	4/9
2	2/9
3	2/9
4	1/9

Connectivity



Barabasi et al, Nat Rev Genet 5, 101-13 (2004)

degree exponent (γ)



- **Normally, 2< γ <3.**

Scale Free Network

• hubs, highly connected nodes, bring together different part of the network

• Rubustness: Removing random nodes have little effect

• Low attack resistance: Removing a hub is lethal.

Random Netwok

• No hubs

• Low robustness

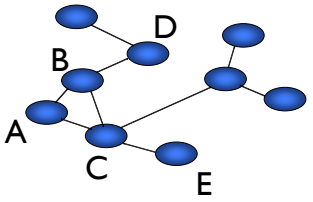
• Low attack resistance

Clustering Coefficient

C/N(N-1)/2

C: links between neighbors

N: Number of neighbors

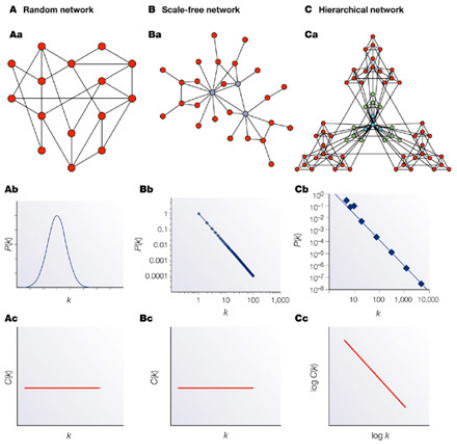


C(A)=1/2(2-1)/2=1

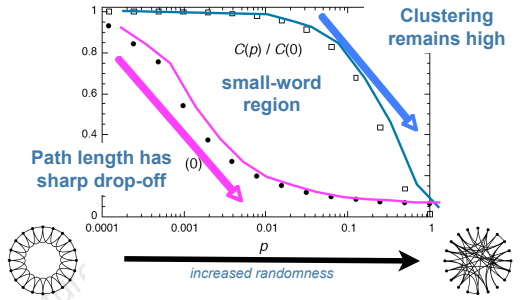
C(B)=1/3(3-1)/2= 1/3

C(C)=1/4(3-1)/2= 1/4

C(D)=0/2(2-1)/2= 0



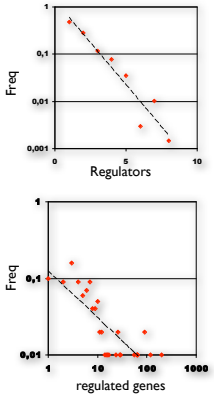
Small-World



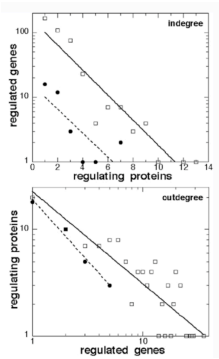
[Watts & Strogatz (1998) *Nature* 393: 440]

Gene Networks are directed

E. coli



S. cerevisiae



Exponential Distribution

Power-Low distribution

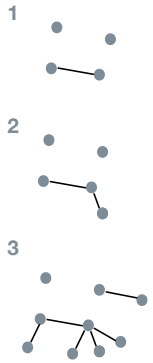
"scale-free"

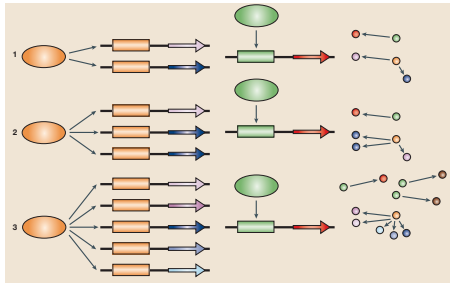
Robust

small world
(signal integration)

Guelzim et al. 2002 Nature Genet. 31:60-63

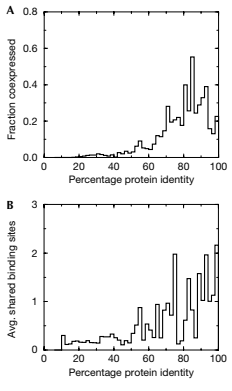
Network Evolution



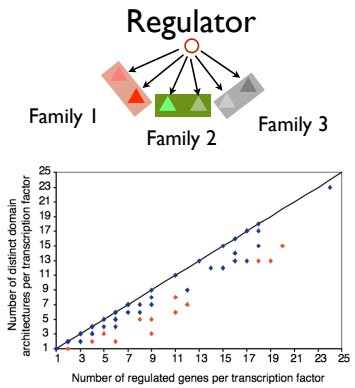


Cases & de Lorenzo 2005 Nature Rev Microb 3:105-118

Duplication and Evolution

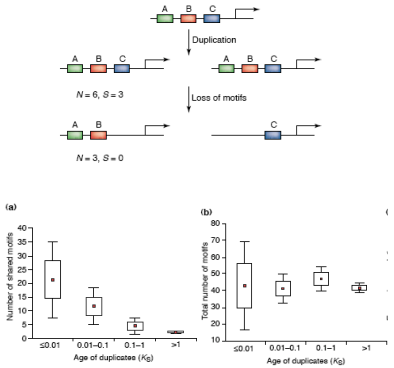


van Noort *et al.*, 2004 EMBO Rep 5(3):280-4



Tecihmann & Babu, 2004 Nature Genetics 36(5):492-6

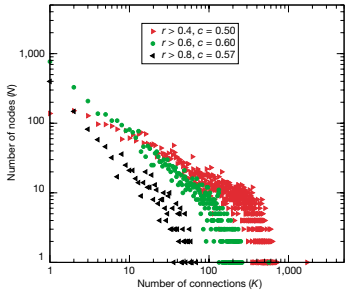
Duplication of TFBS



Papp et al,2003.Trends Genet 19:417

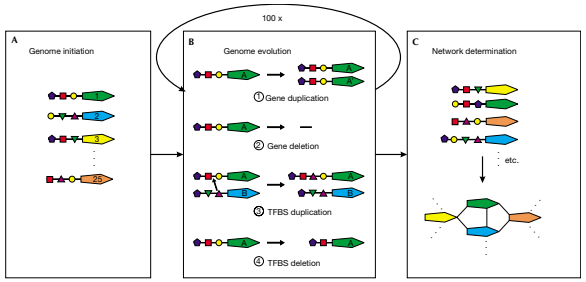
The Co-regulation Network

- gamma≈-1
- c=0.6
- scale-free
- "small world"



van Noort *et al.*, 2004 EMBO Rep 5(3):280-4

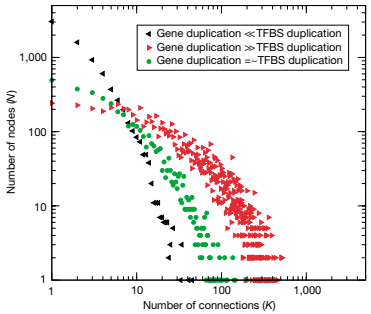
Simulating evolution



van Noort *et al.*, 2004 EMBO Rep 5(3):280-4

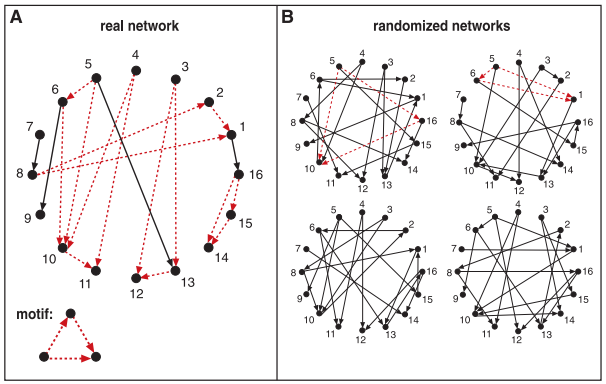
Simulating evolution II

Scale free, small world networks, can appear in the absence of selection



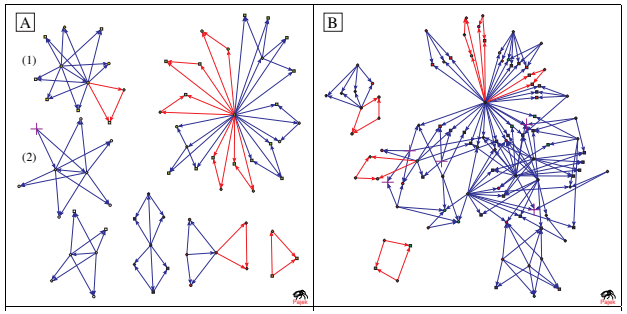
van Noort *et al.*, 2004 EMBO Rep 5(3):280-4

Network Motives



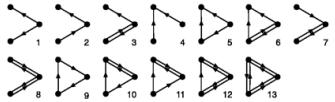
Milo *et al*,2002. Science 298:824

Overlapping Motives



Dobrin *et al*,2004. BMC Bioiformatics 5:10

Regulatory Networks Motives



Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	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

Milo *et al*,2002. Science 298:824

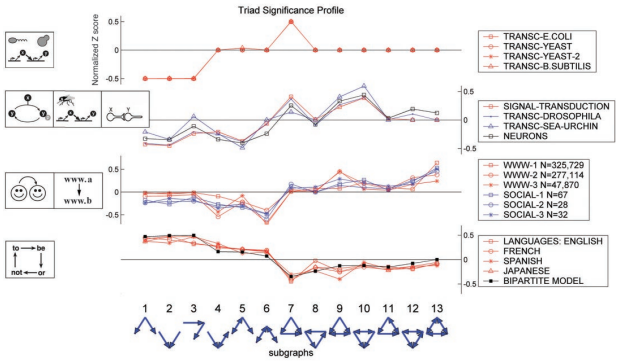
Regulatory Networks Motives are specific

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

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

Milo *et al*,2002. Science 298:824

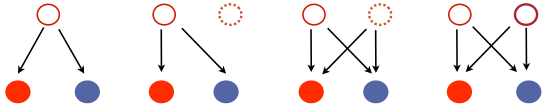
Motive-based Profiles



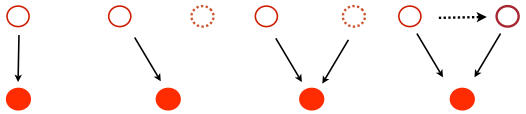
Milo et al. 2004 Science 303:1538-1542

Motives evolution

Bi-fans



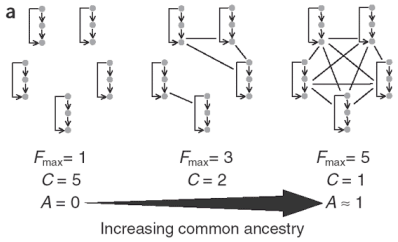
FFL



NO: There is no homologous regulators in he same regulatory motif in *E. coli*

Tecihmann & Babu, 2004 Nature Genetics 36(5):492-6

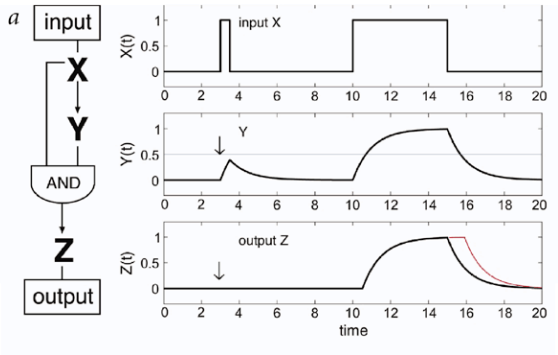
Motives Evolution II



E. coli	Feed-forward	11	11	0	1
	Bi-fan	27	27	0	1

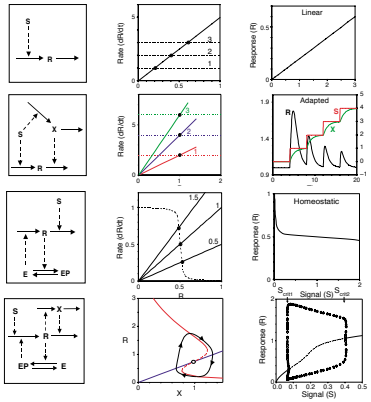
Conant & Wagner, 2003. Nat Genet. 34:264

Motives Properties



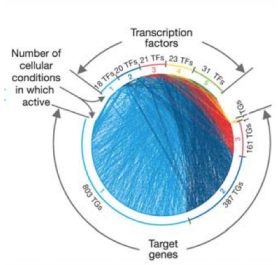
Shen-Orr *et al.*, 2002. Nat Genet. 31:64

Circuits



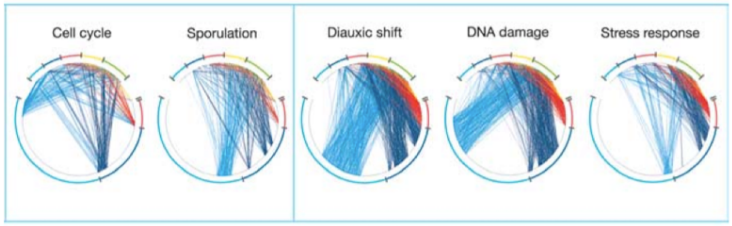
Tyler *et al.* 2003 Current Opinion in Cell Biology 15:221–

Regulatory Networks are dynamic

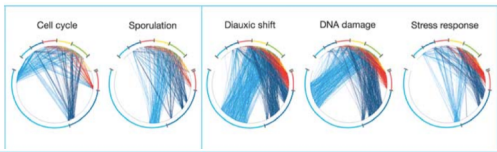


Luscombe *et al.*, 2004 Nature 431:308

Different Networks are active under different conditions

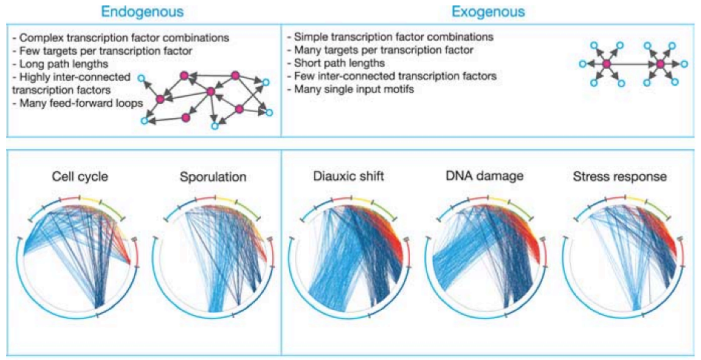


Endogenous Exogenous



		Cell cycle	Sporulation	Diauxic shift	DNA damage	Stress response
Size	No. of transcription factors	142	70	74	71	72	63
	No. of target genes	3,420	280	257	748	678	362
	No. of regulatory interactions	7,074	550	481	1,217	1,082	566
Topological measures	In-degree ($<k_{in}>$)	2.1	2.0	1.9	1.6	1.6	1.6
	Out-degree ($<k_{out}>$)	49.8	7.9	6.5	17.1	15.0	9.0
	Path length (<*l*>)	4.7	4.5	3.4	2.1	2.0	2.2
	Clustering coefficient (<*c*>)	0.11	0.15	0.14	0.09	0.09	0.08
Motifs (%)	Single input (SIM)	1,748 (37.6%)	130 (32.0%)	117 (38.9%)	438 (57.4%)	462 (55.7%)	228 (59.1%)
	Multiple input (MIM)	325 (7.0%)	96 (23.7%)	50 (16.6%)	180 (23.6%)	226 (27.3%)	78 (20.2%)
	Feed-forward loop (FFL)	2,581 (55.5%)	180 (44.3%)	134 (44.5%)	145 (19.0%)	141 (17.0%)	80 (20.7%)
	Total	4,654	406	301	763	829	386

Luscombe *et al.*, 2004 Nature 431:308



Luscombe *et al.*, 2004 Nature 431:308

Summary

- Building regulatory networks from experiments is tedious and expensive... but it can be done.
- Computational methods are noisy and generate many false positives.
- Two main questions:
  - Pattern matching
  - Pattern Discovery (phylogenetic footprint)

Summary II

- Regulatory Networks are directed. Outgoing connectivity follows a power law, but not the incoming one
- The network is scale-free and small world
  - Robust and good for signal integration
- The network could have grown by duplication, but there are some contradictory evidences
- Regulatory motives are specific and provides some convenient properties. Motives are under strong selective pressure.
- The network is dynamic. Different stimuli require different networks with different properties.

Identification of Transcription Factor Binding Sites

- Go to: http://rsat.scmbb.ulb.ac.be/rsat/
- Misc>Tutorials
  - 1. Sequence retrieval
  - 3. Pattern Matching
    - 2 patser
  - 4. Patten Discovery
    - 1.1 oligo-analysis
    - 2.1 Gibbs Motif Sampler
    - 4.3 Microarrays