Bio Graph Analysis

Lecture 9

CSCI 4974/6971

29 Sep 2016

Today's Biz

1. Reminders
2. Review
3. Biological Network Analysis Topics
4. Hybrid processing - direction optimizing push/pull
5. Assignment 2 solutions

Today's Biz

1. **Reminders**
2. Review
3. Biological Network Analysis Topics
4. Hybrid processing - direction optimizing push/pull
5. Assignment 2 solutions

Reminders

- Project Presentation 1: in class 6 October
  - Email me your slides (pdf only please) before class
  - 5-10 minute presentation
  - Introduce topic, give background, current progress, expected results
- **No class 10/11 October**
- Assignment 3: Thursday 13 Oct 16:00 (social analysis, posted soon)
- Office hours: Tuesday & Wednesday 14:00-16:00 Lally 317
  - Or email me for other availability

Today's Biz

1. Reminders
2. **Review**
3. Biological Network Analysis Topics
4. Hybrid processing - direction optimizing push/pull
5. Assignment 2 solutions

Quick Review

- Balanced graph partitioning:
  - Create k independent subsets of graph
  - Satisfy some balance criteria
- Traditional (mesh-like graph) methods:
  - Coordinate-based methods - inertial bisection, coordinate-based
  - Spectral bisection - compute eigenvector using graph Laplacian
  - KL-refinement - find best cost/gain for vertex swaps
  - Multilevel - iterative coarsening/expanding+refinement

Quick Review

- Drawbacks of tradition methods for small-world/massive scale graphs
  - KL and spectral methods require $O(n^2)$
  - Coarsening occurs a high overhead costs
  - Traditional matching methods perform poorly on skewed graphs
- *Small-world* and large-graph methods:
  - Streaming methods: perform immediate assignment based on some weighted cost/gain function for each vertex/edge encountered in the stream
  - Single-level label propagation: hold full graph in memory, exploit community-like structure of small-world graphs to get quality partitions without a multilevel framework
  - Other: Use distributed label propagation for coarsening in multilevel, tradeoff in quality vs. overhead

Today's Biz

1. Reminders
2. Review
3. **Biological Network Analysis Topics**
4. Hybrid processing - direction optimizing push/pull
5. Assignment 2 solutions

Network Motifs: simple Building Blocks of Complex Networks

Slides from Yoav Lahini, Harvard University

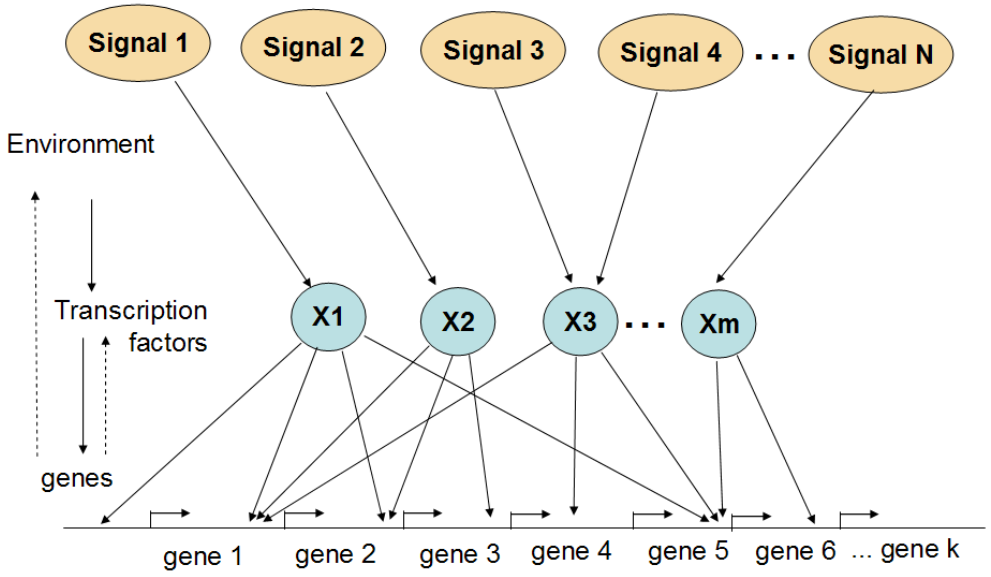
Network Motifs: simple Building Blocks of Complex Networks

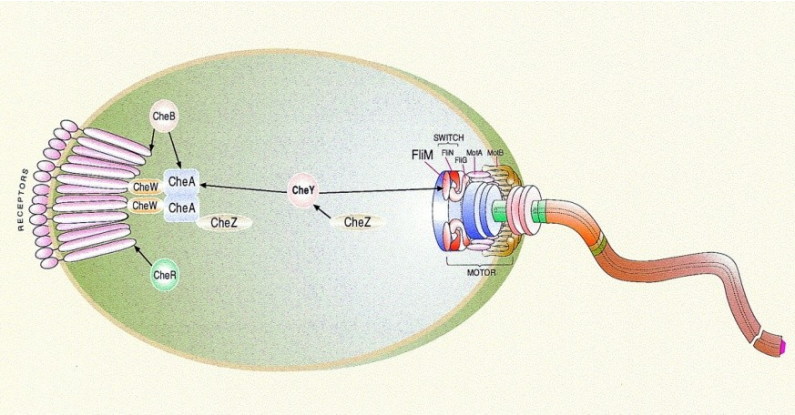
R. Milo *et. al. Science* ***298****, 824 (2002)*

Y. Lahini

The cell and the environment

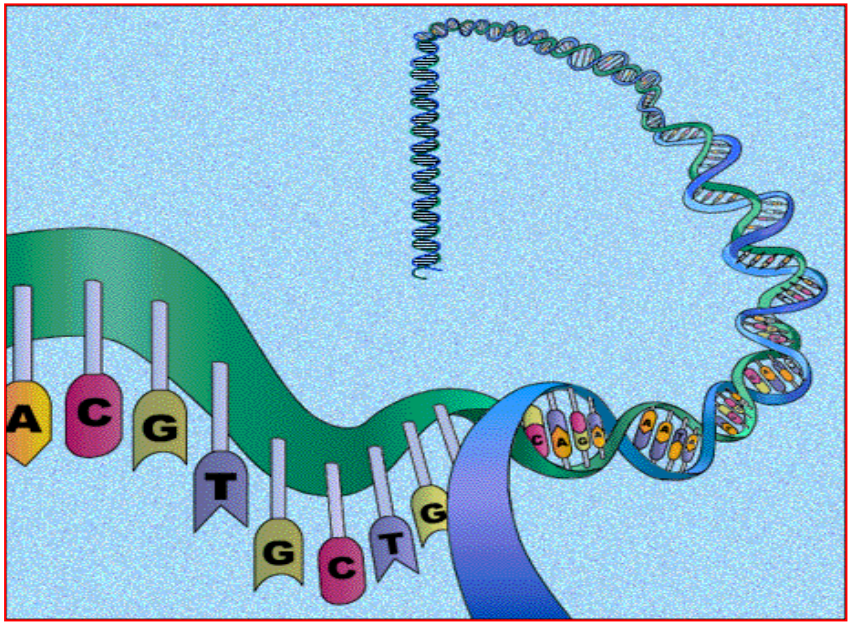
- Cells need to react to their environment
- Reaction is by synthesizing task-specific proteins, on demand.
- The solution – regulated transcription network

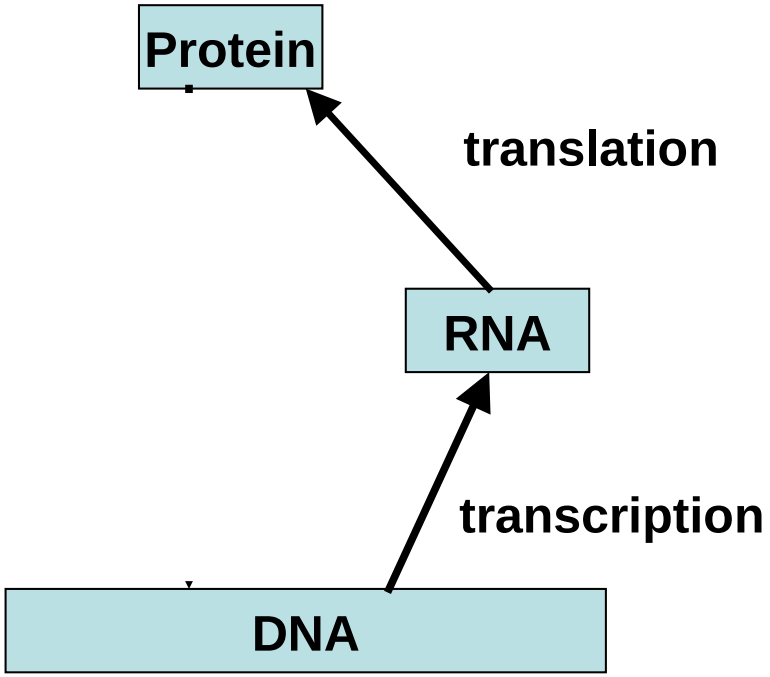




- E. Coli – 1000 protein types at any given moment >4000 genes (or possible protein types) – need regulatory mechanism to select the active set
- We are interested in the design principles of this network

Proteins are encoded by DNA

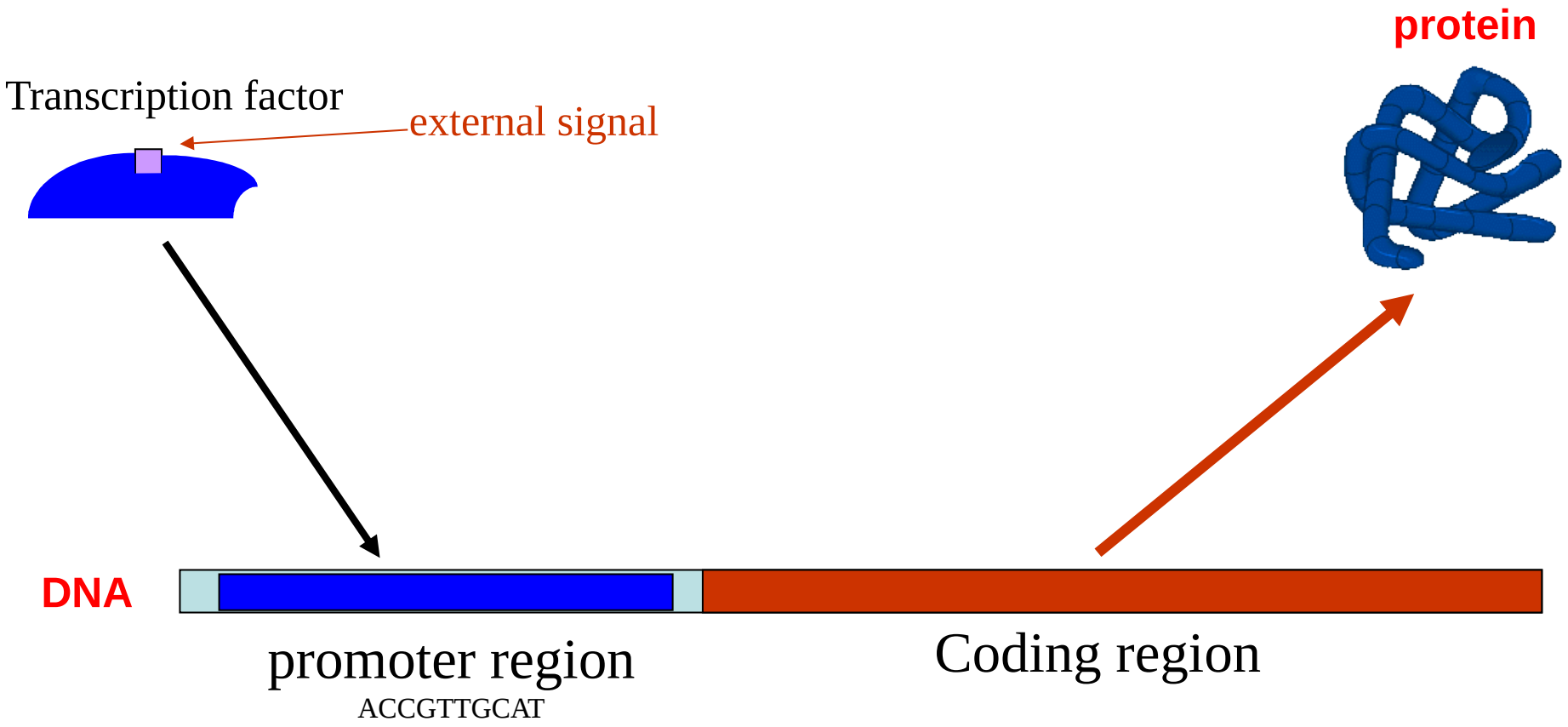




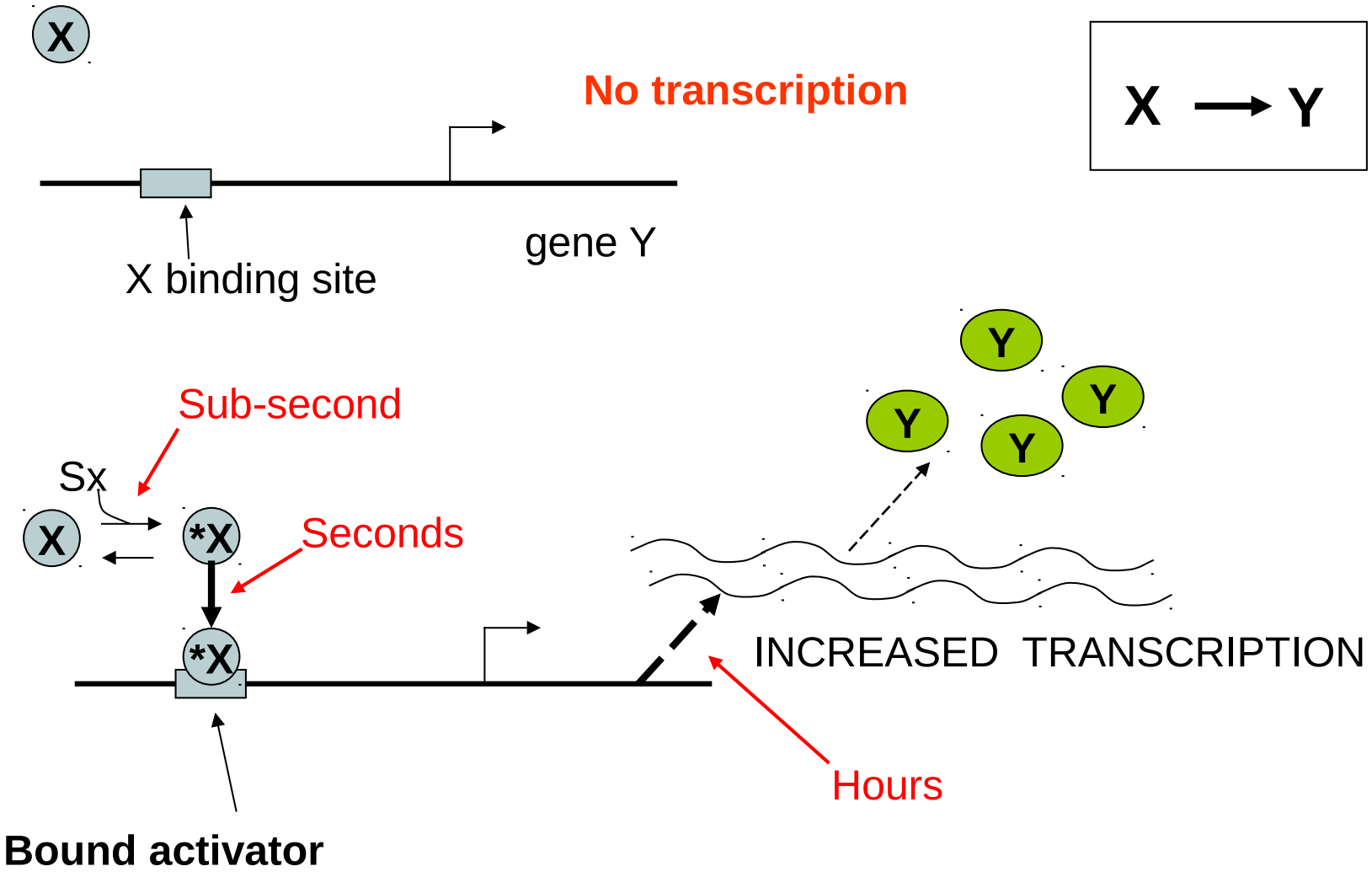
DNA – the instruction manual, 4-letter chemical alphabet – A,G,T,C

Gene Regulation

- Proteins are encoded by the DNA of the organism.
- Proteins regulate expression of other proteins by interacting with the DNA

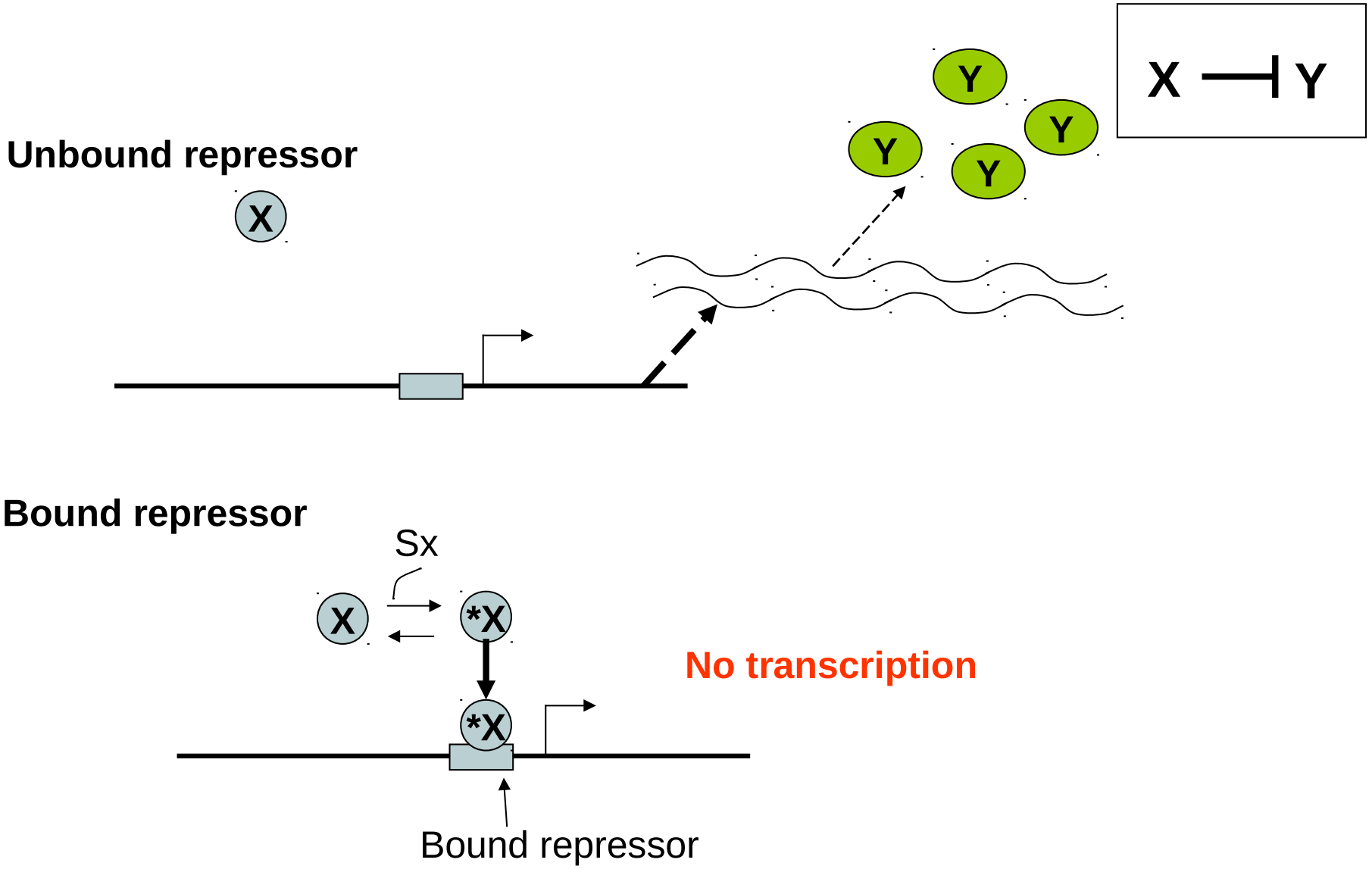


Two types of Transcription Factors: 1.Activators



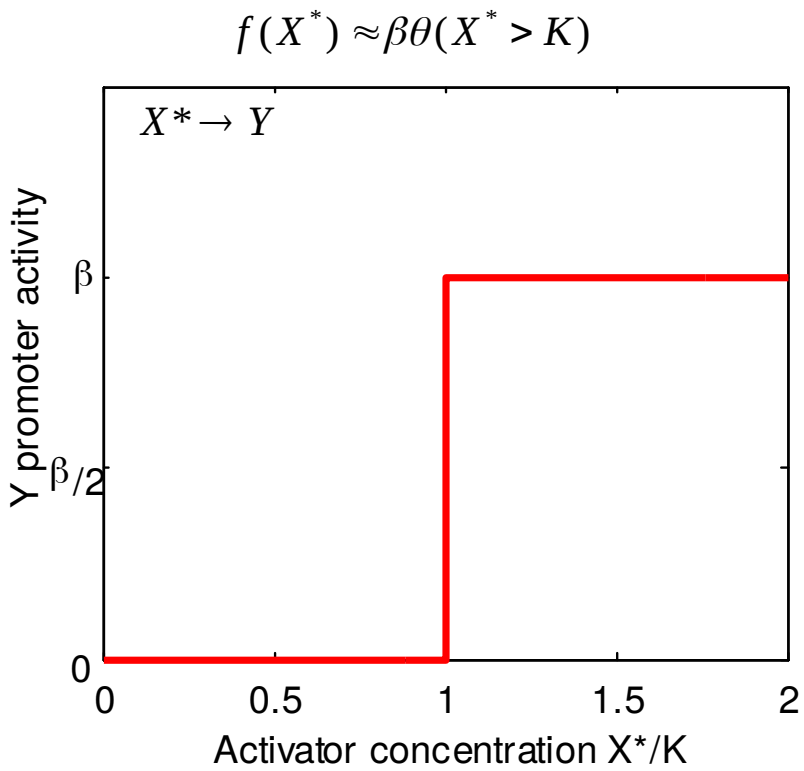
Separation of time scales: TF activation level is in steady state

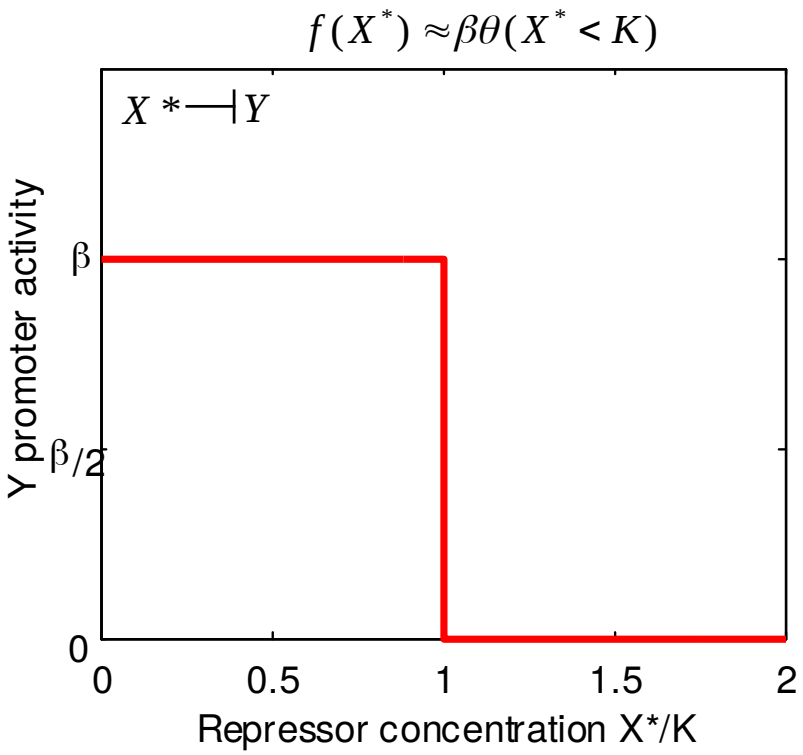
Two types of Transcription Factors: Repressors



Equations of gene regulation

- If X* regulates Y, the net production rate of gene Y is $\frac{dY}{dt} = f(X^*) - \alpha Y$
- α- Dilution/degradation rate

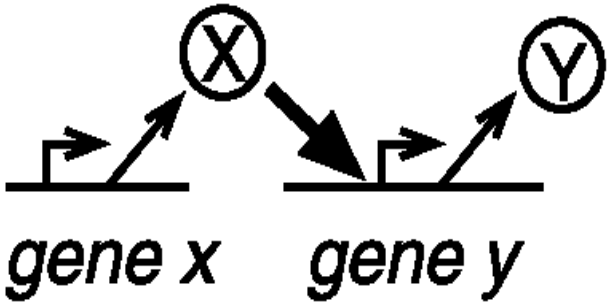


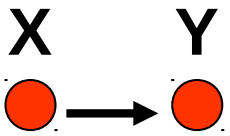


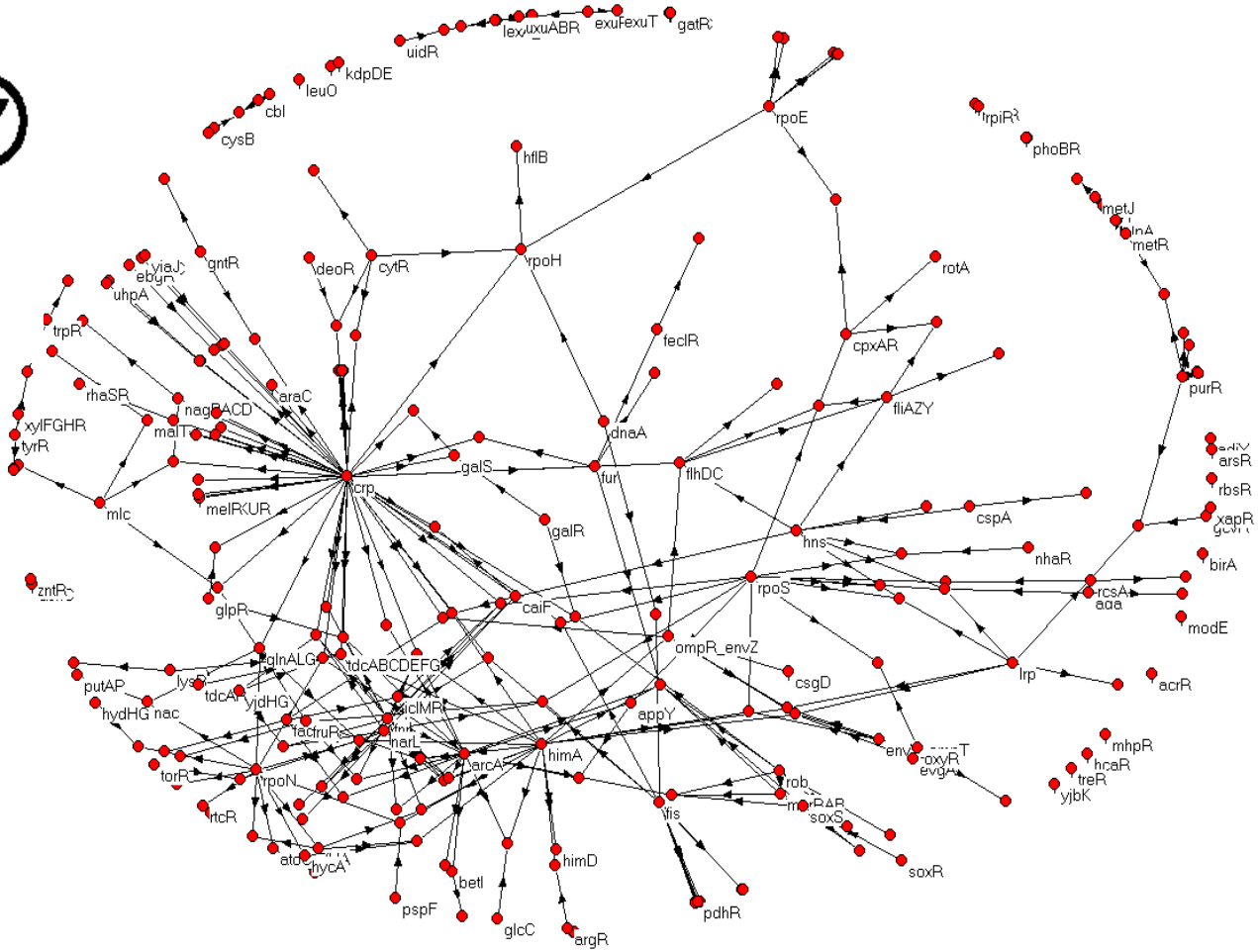
- K – activation coefficient [concentration]; related to the affinity
- *β – maximal expression level*
- Step approximation – gene is on (rate *β*) or off (rate 0) with threshold K

The gene regulatory network of E. coli

- Nodes are proteins (or the genes that encode them)
- Edges = regulatory relation between two proteins

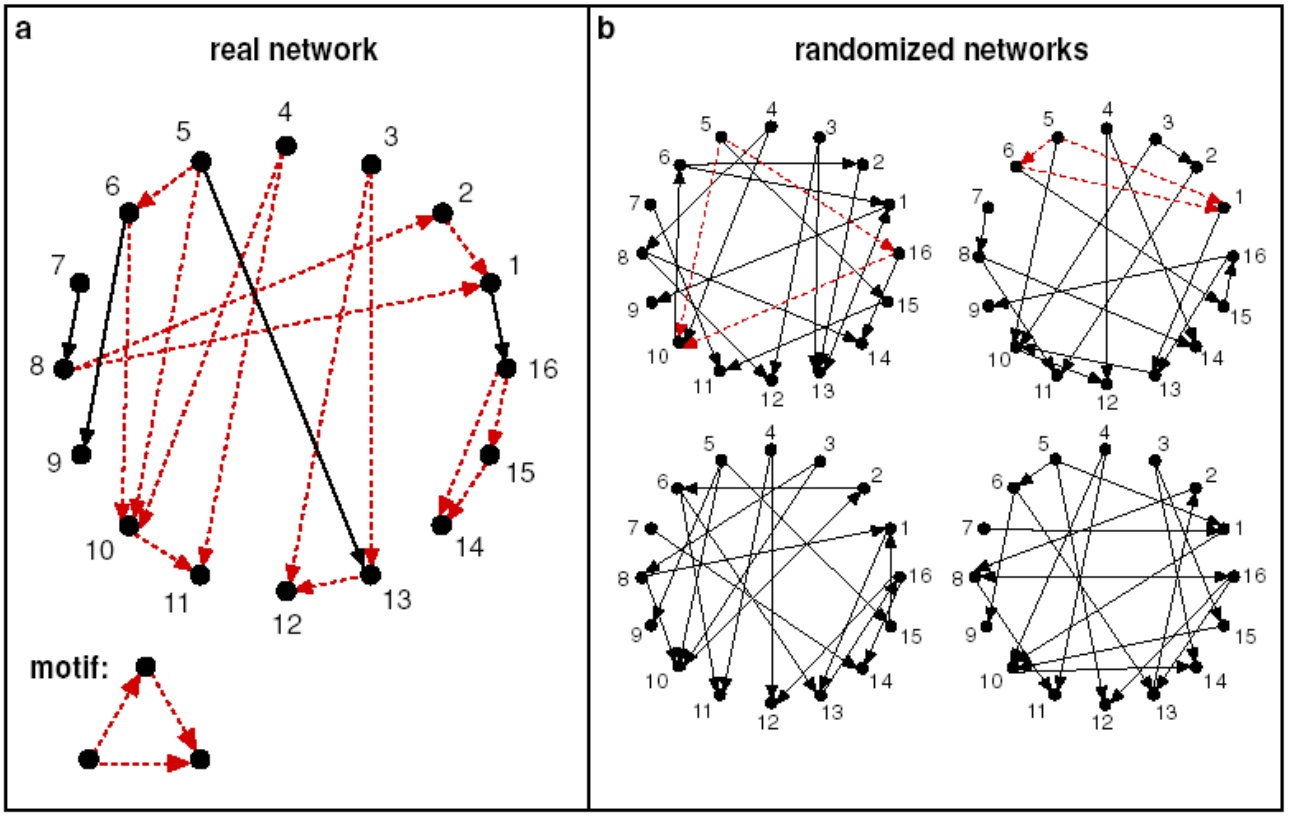




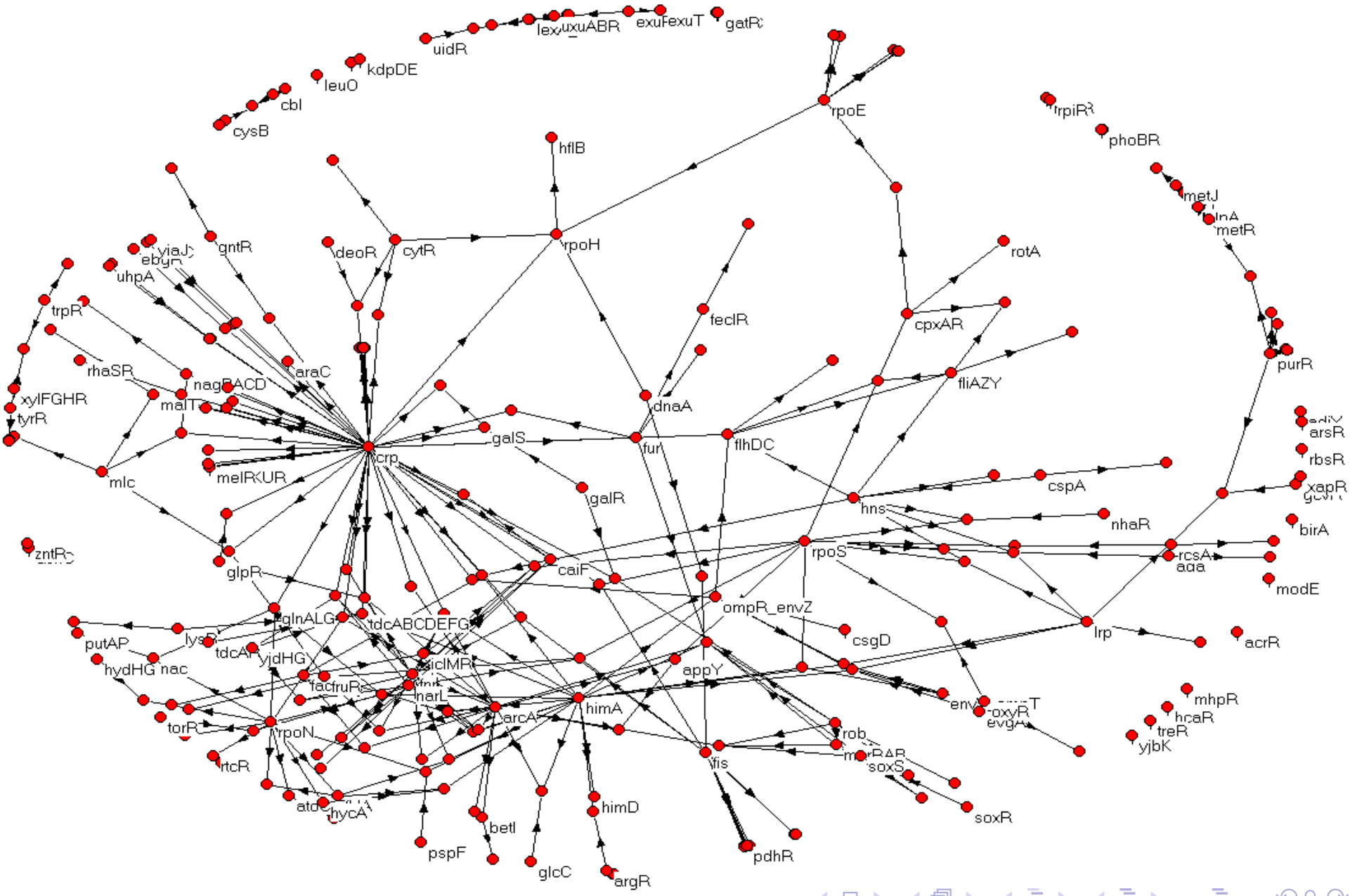


Analyzing networks

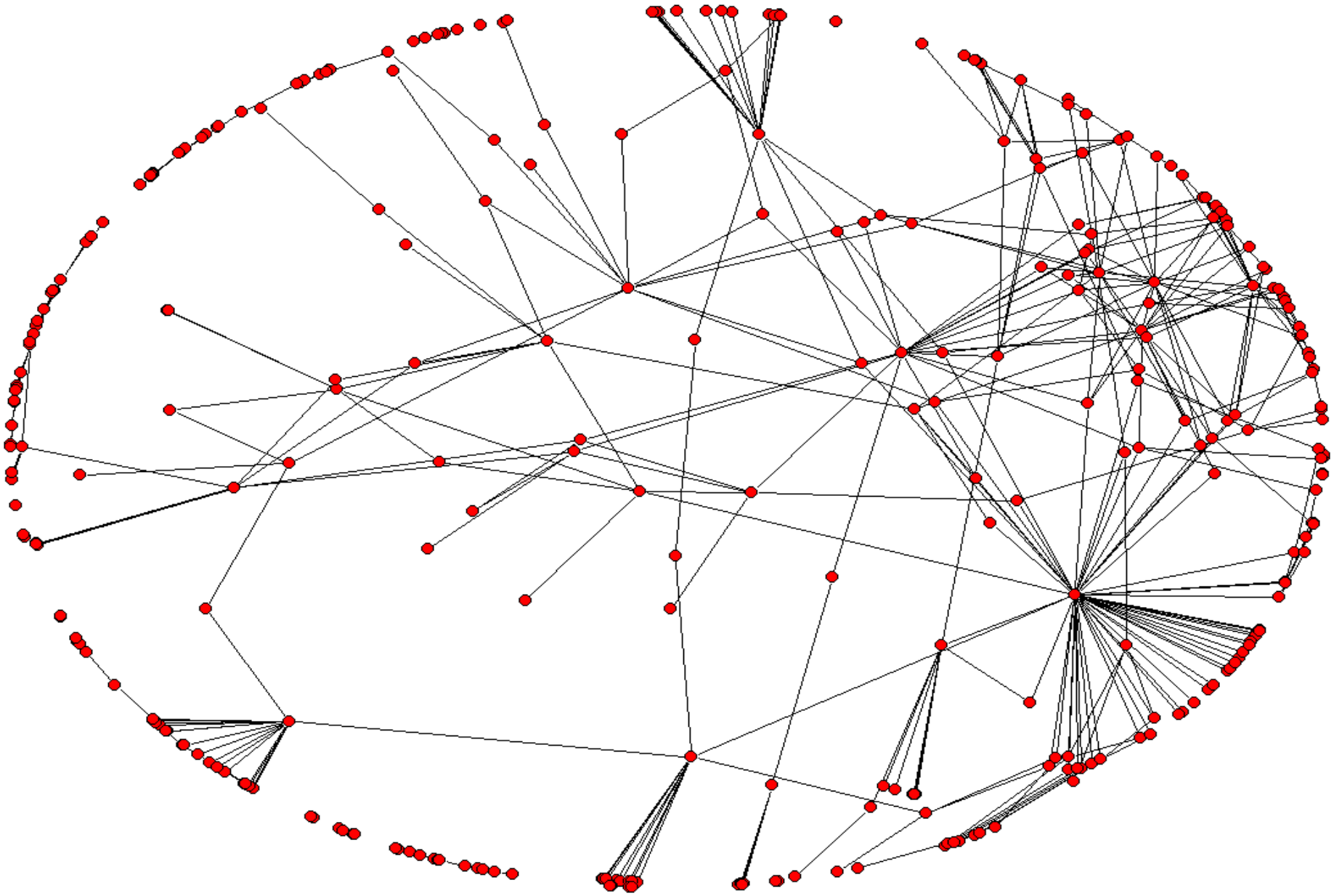
- The idea- patterns that occur in the real network much more then in a randomized network, must have functional significance.
- The randomized networks share the same number of edges and number of nodes, but edges are assigned at random



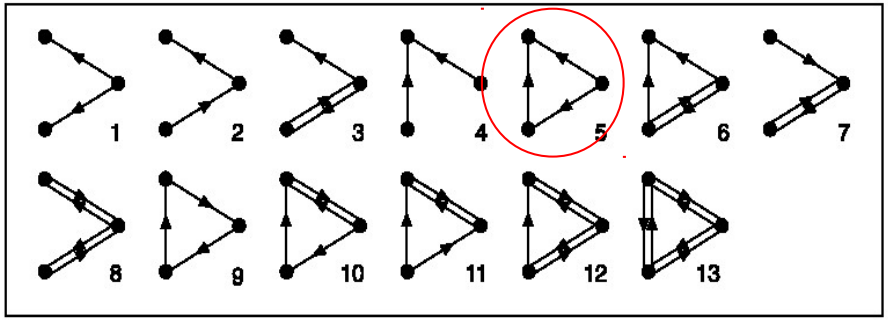
The known E. Coli transcription network

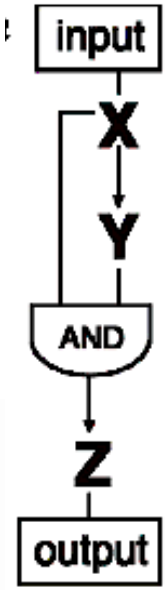


A random graph based on the same node statistics



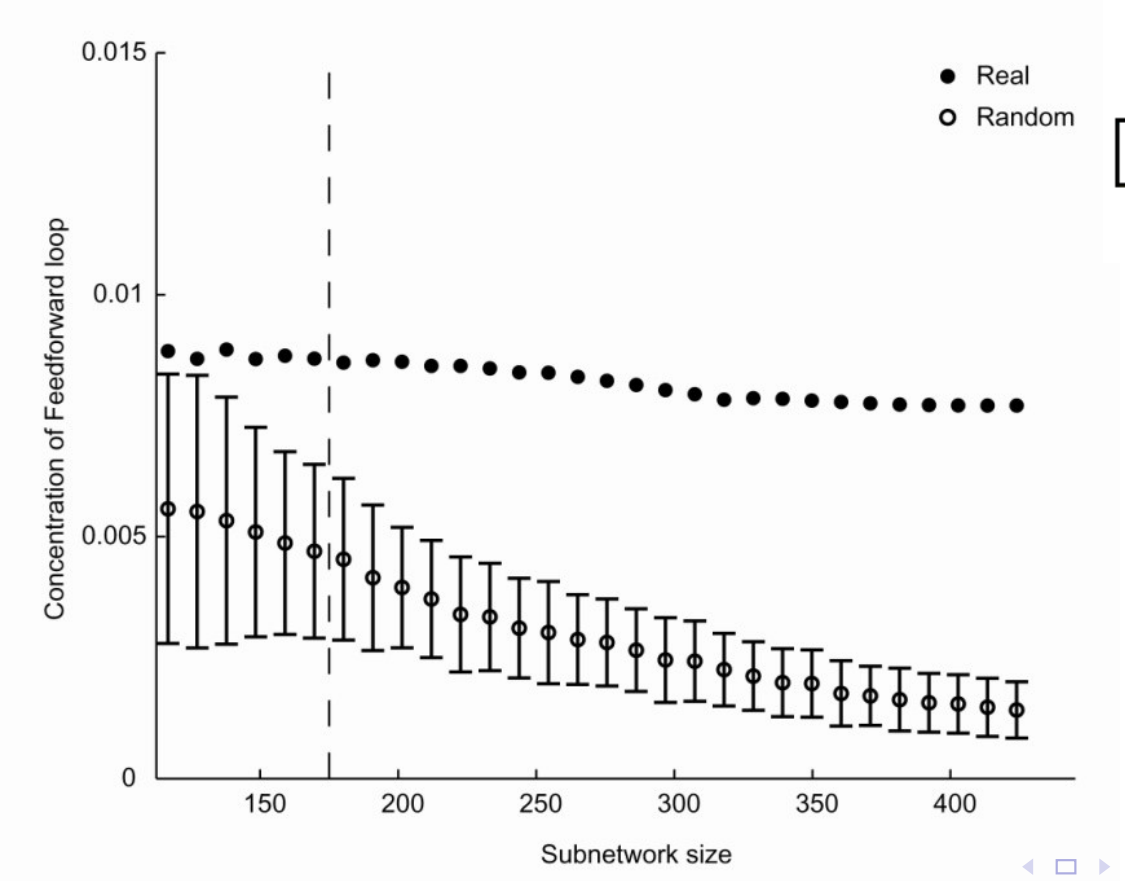
3-node network motif – the feedforward loop



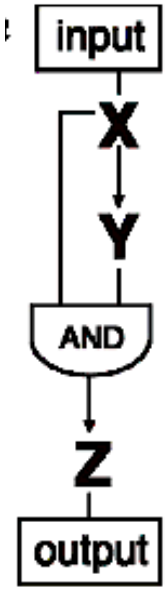


Nreal=40

Nrand=7±3



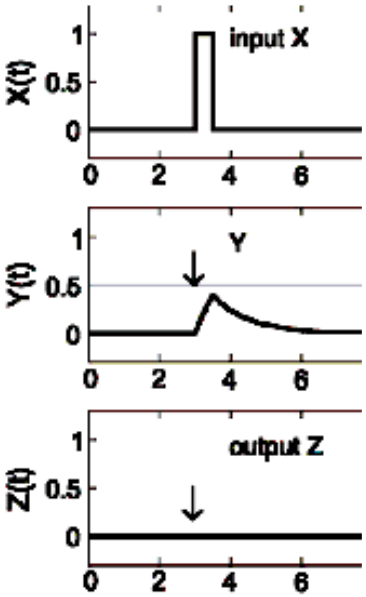
The feedforward loop : a sign sensitive filter



$$X = X(t)$$

$$\frac{dY}{dt} = \theta(X - k_{XY}) - Y$$

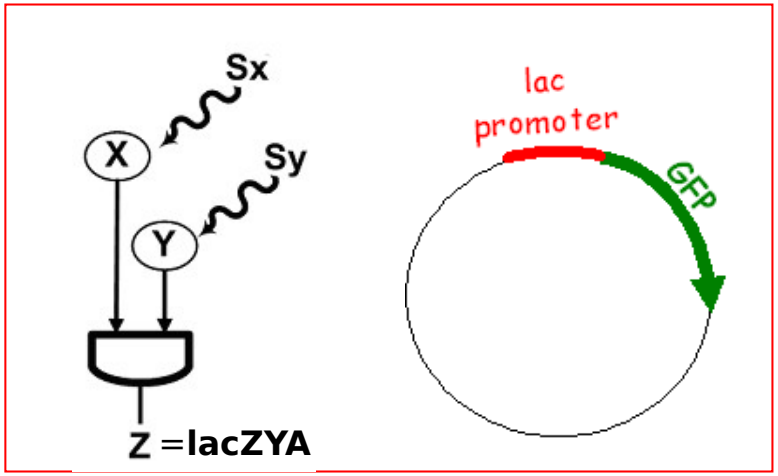
$$\frac{dZ}{dt} = \theta(X - K_{XZ})\theta(Y - K_{YZ}) - Z$$



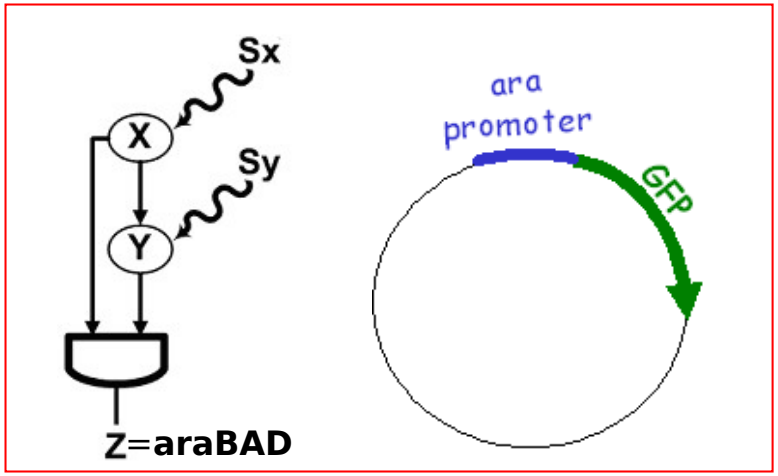
The feedforward loop is a filter for transient signals while allowing fast shutdown

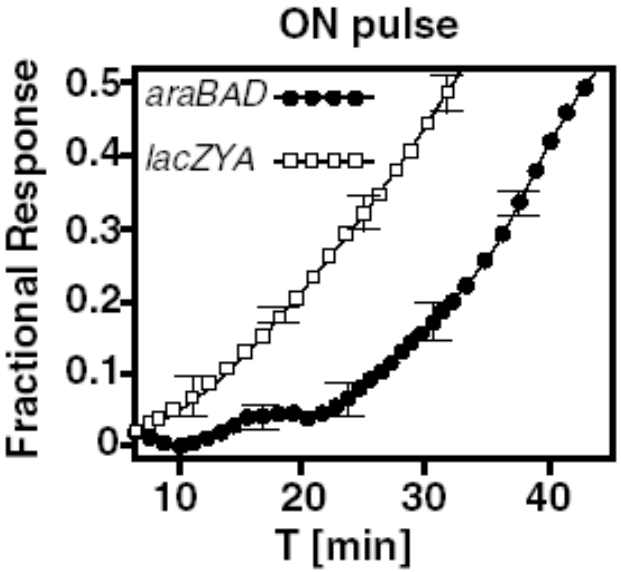
Mangan, Alon, PNAS, JMB, 2003

The Feedforward loop : a sign sensitive filter



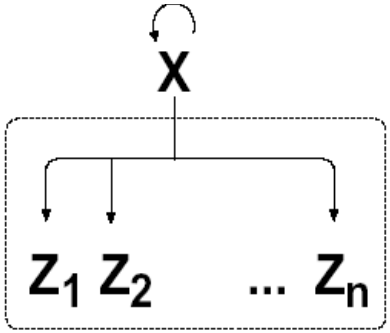
Vs.

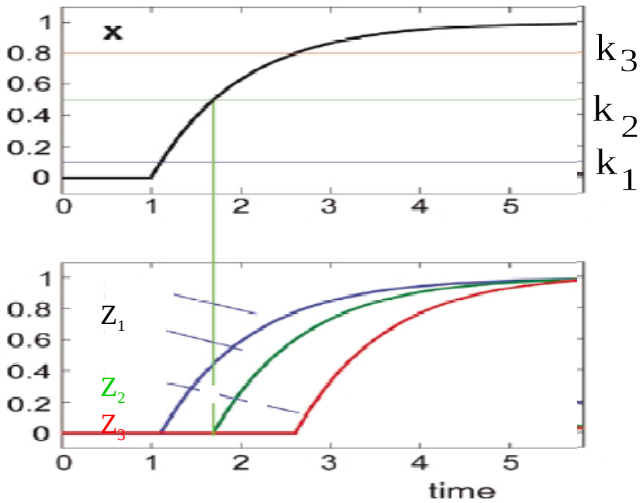




Mangan, Alon, PNAS, JMB, 2003

Single Input Module





$$X = X(t)$$

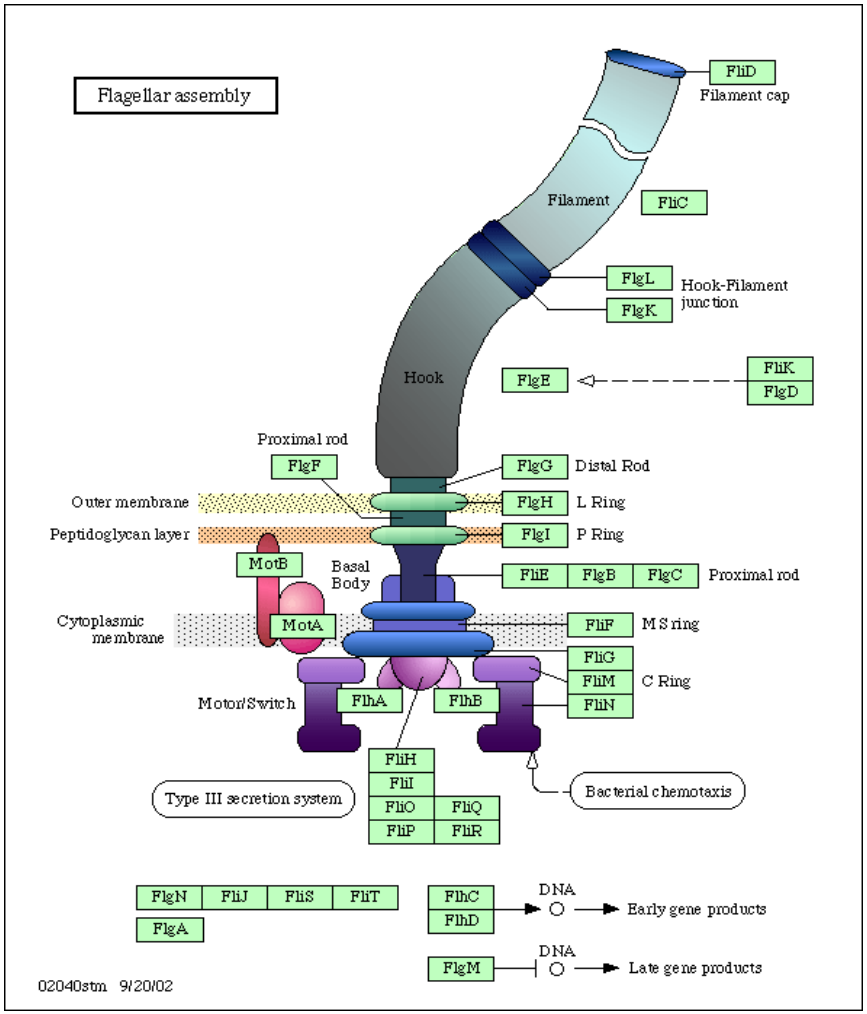
$$\frac{dZ_i}{dt} = \beta_i \theta(X - k_i) - Z_i$$

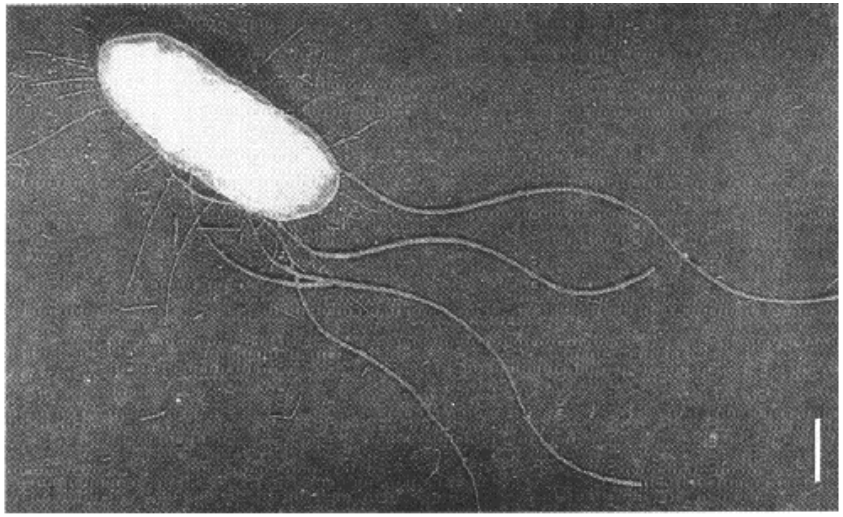
$$\begin{cases} k_1 < k_2 < \ldots < k_n \\ \beta_1 > \beta_2 > \ldots > \beta_n \end{cases}$$

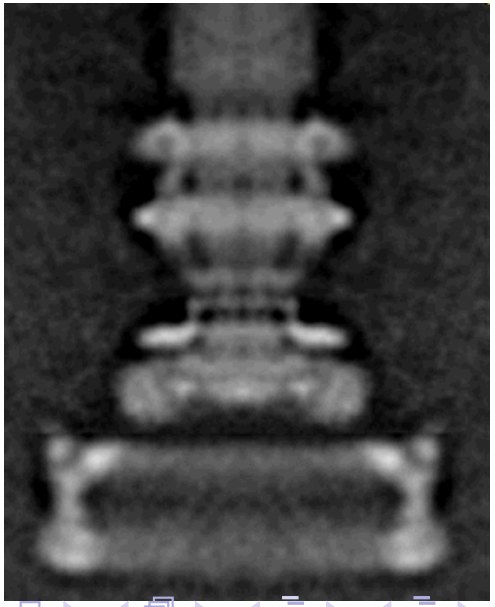
Temporal and expression level program generator

- The temporal order is encoded in a hierarchy of thresholds
- Expression levels hierarchy is encoded in hierarchy of promoter activities

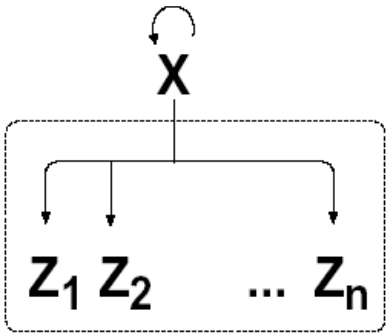
Single Input Module motif is responsible for exact timing in the flagella assembly

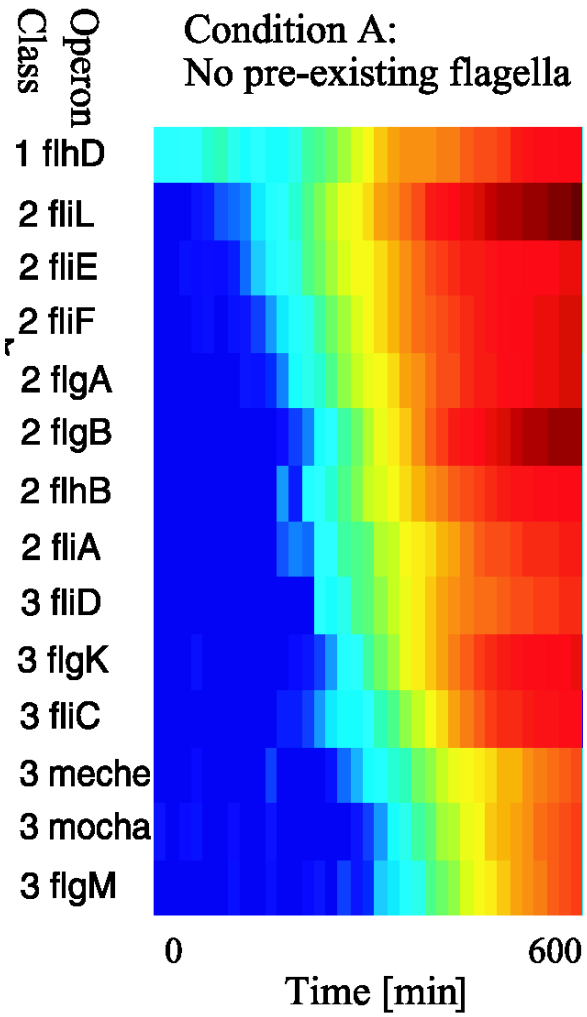


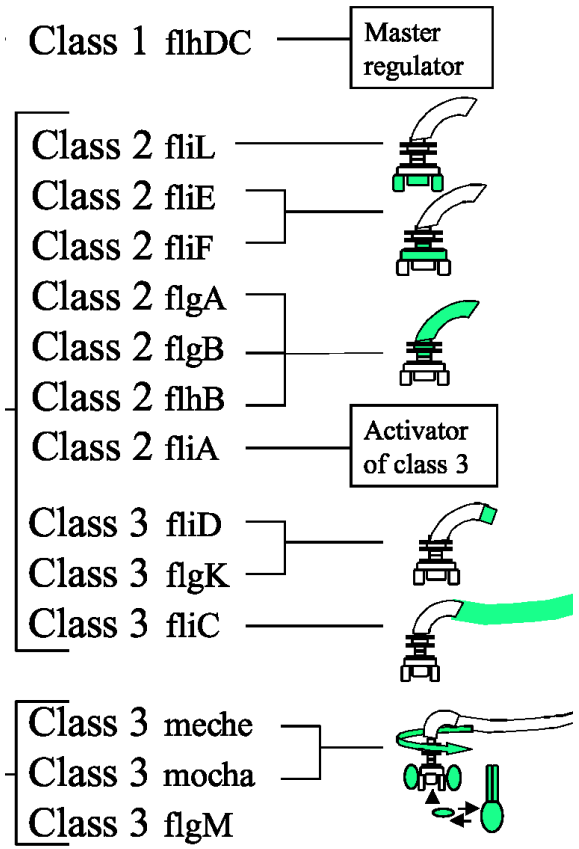




Single Input Module motif is responsible for exact timing in the flagella assembly

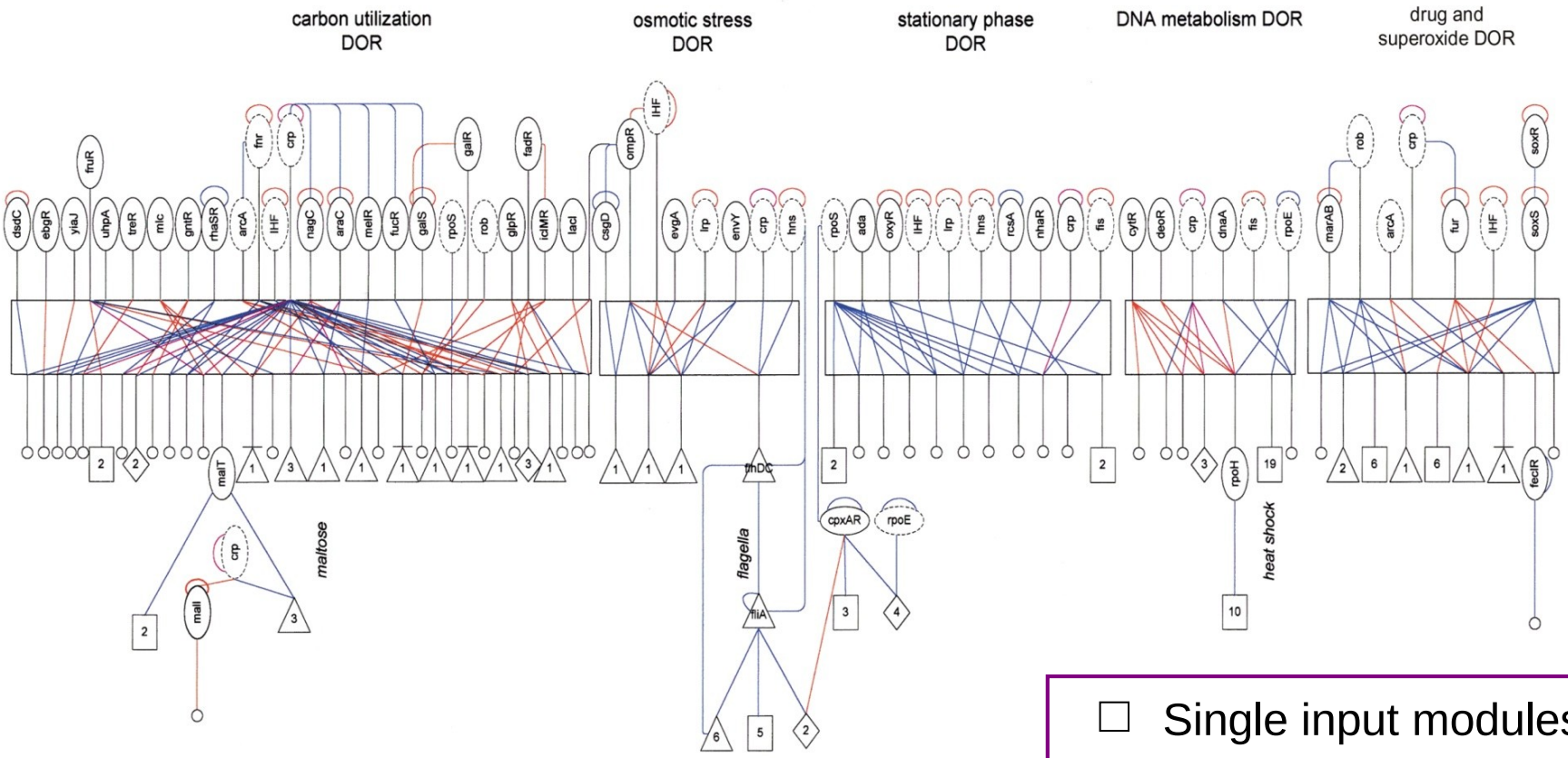






Kalir et. al., science,2001

The gene regulatory network of E. coli



- Shallow network, few long cascades.
- Modular

□ Single input modules

△ Feed-forward loops

Shen-Orr et. al. Nature Genetics 2002

Evolution of transcription networks

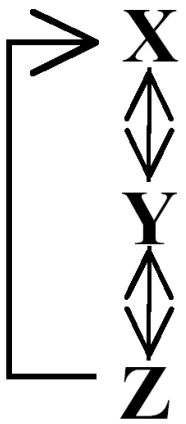
- In 1 day, 10^{10} copies of e-coli, 10^{10} replication of DNA.
- Mutation rate is 10^{-9}
  - 10 mutations per letter in the population per day
- Even single DNA base change in the promoter can change the activation/repression rate

- Edges can be lost or gained (i.e. selected) easily.

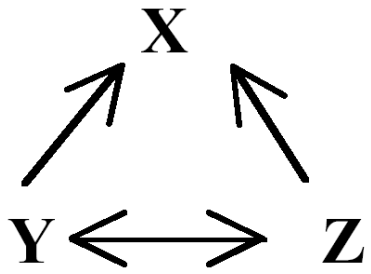
Links between WebPages – a completely different set of motifs is found

- WebPages are nodes and Links are directed edges
- 3 node results:

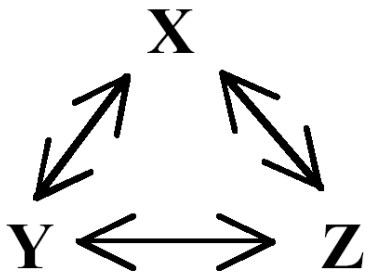
Feedback with two mutual dyads



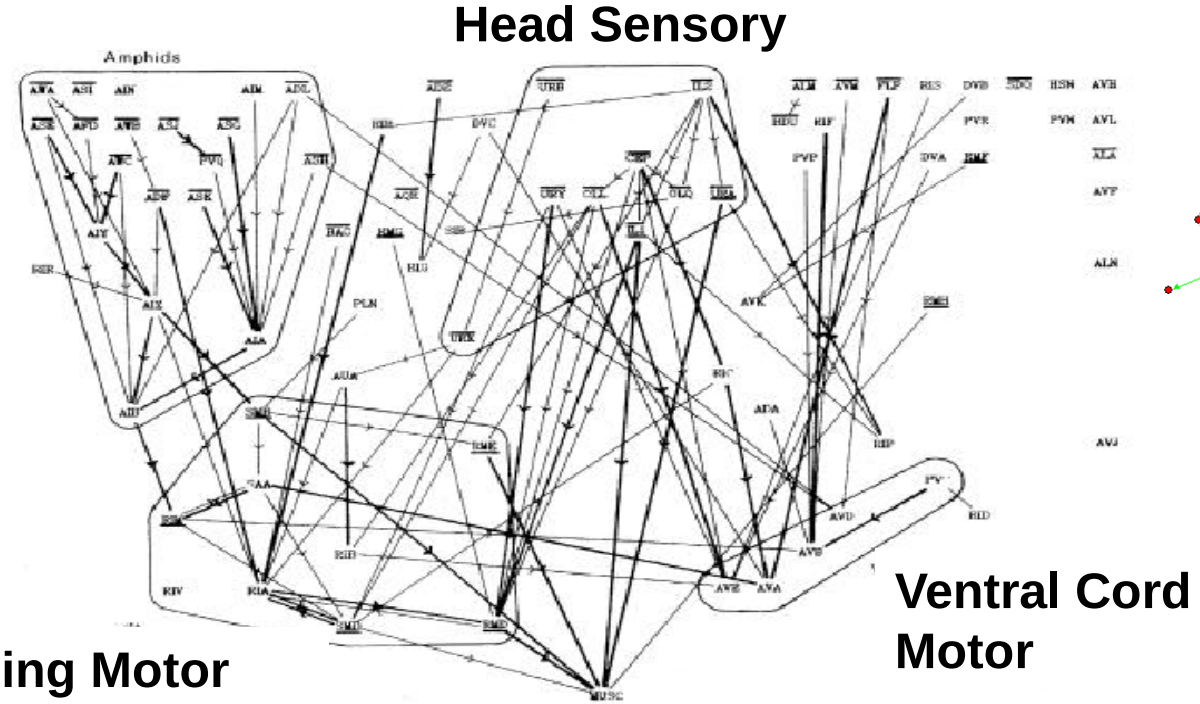
Uplinked mutual dyad

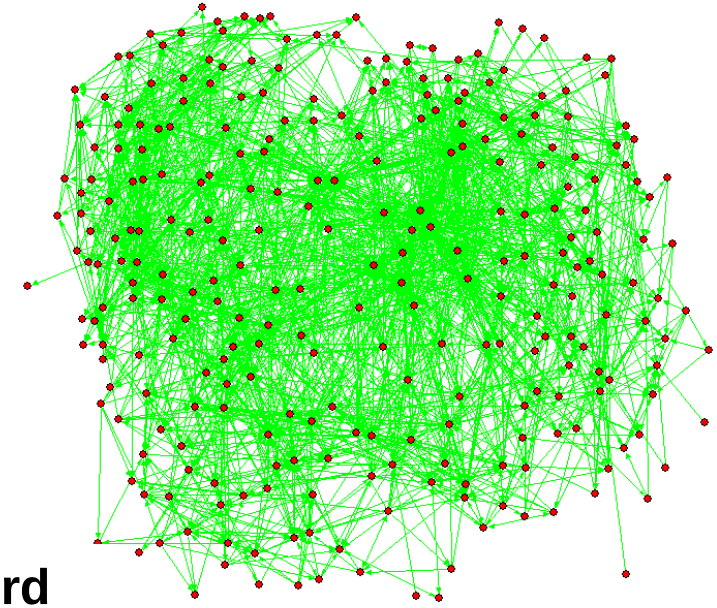


Fully connected triad



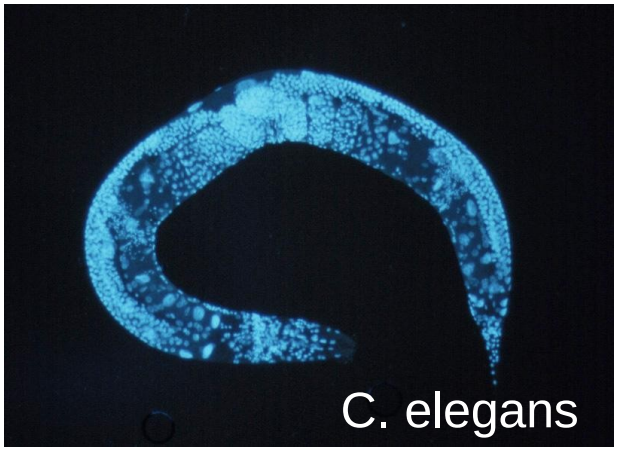
Structure of a nematode neuronal circuitry





[White, Brenner 1986; Durbin, Thesis, 1987]

Neurons and transcription share similar motifs

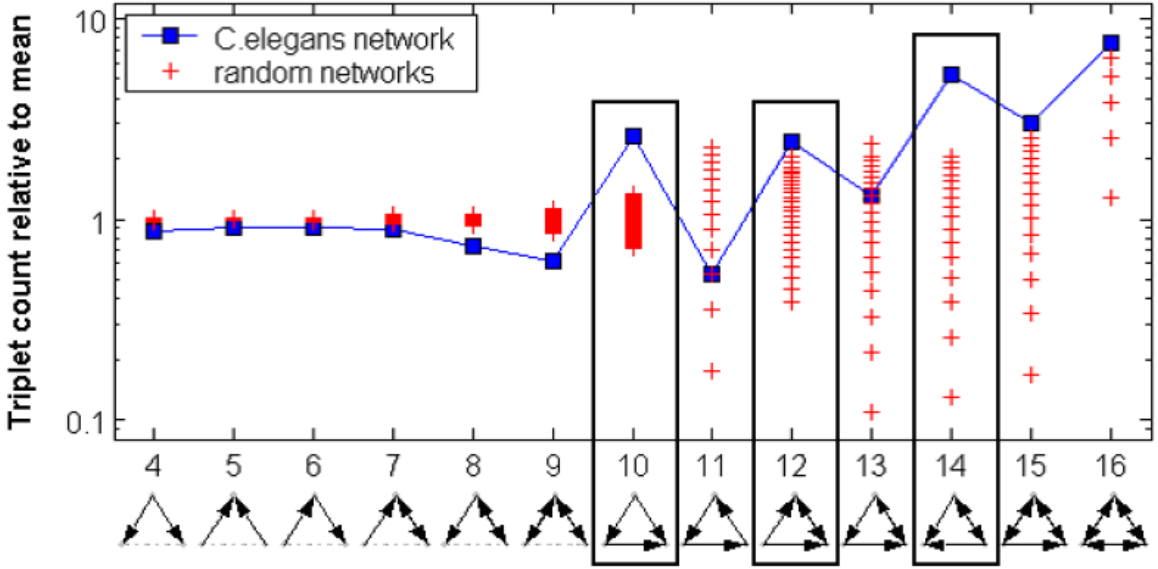


BMC Biology Published: 02 December 2004

Research article

Search for computational modules in the *C. elegans* brain

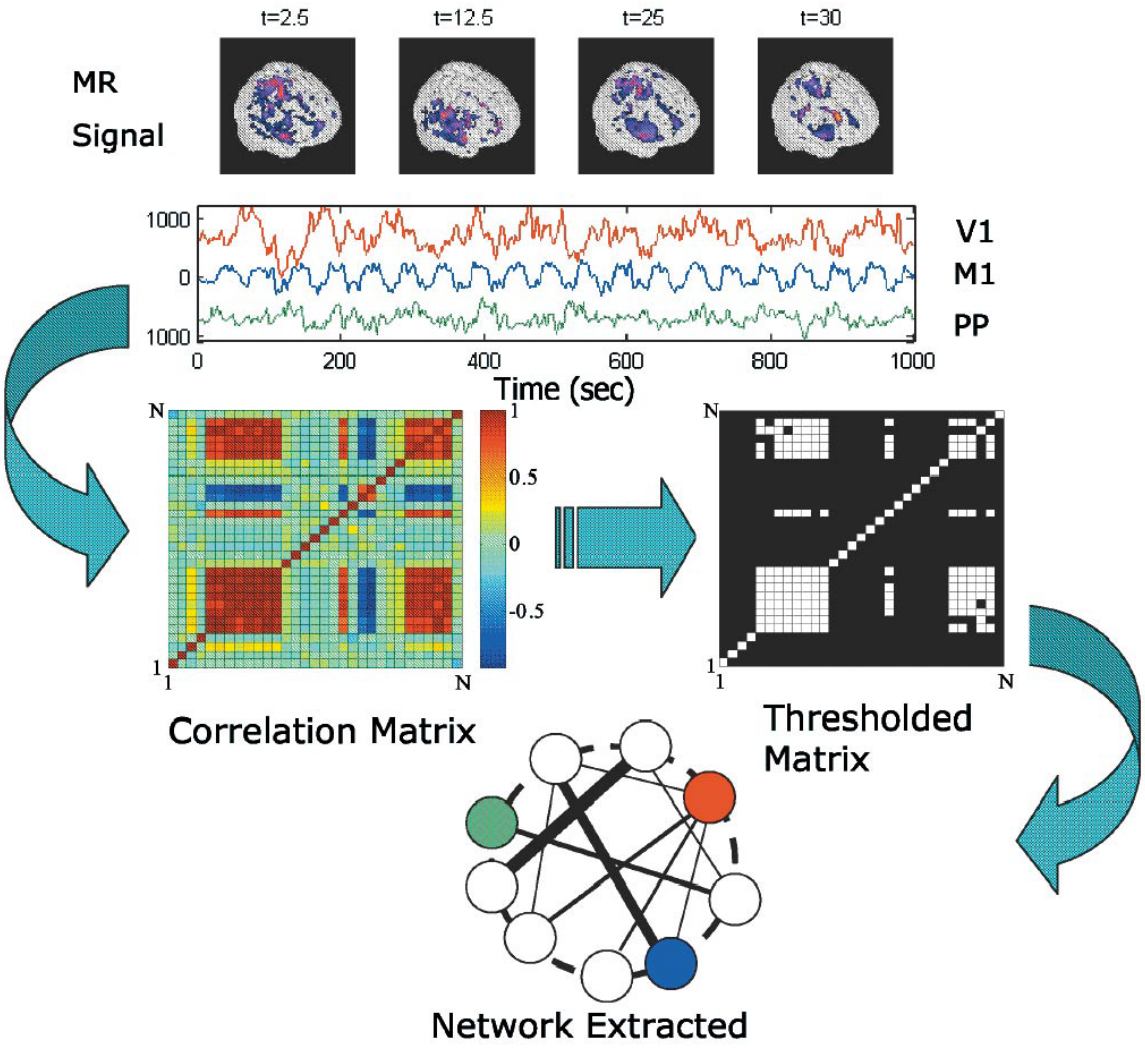
Markus Reigl[1], Uri Alon[2] and Dmitri B Chklovskii*[1]



PRL 94, 018102 (2005) PHYSICAL REVIEW LETTERS week ending 14 JANUARY 2005

Scale-Free Brain Functional Networks

Victor M. Eguíluz,[1] Dante R. Chialvo,[2] Guillermo A. Cecchi,[3] Marwan Baliki,[2] and A. Vania Apkarian[2]



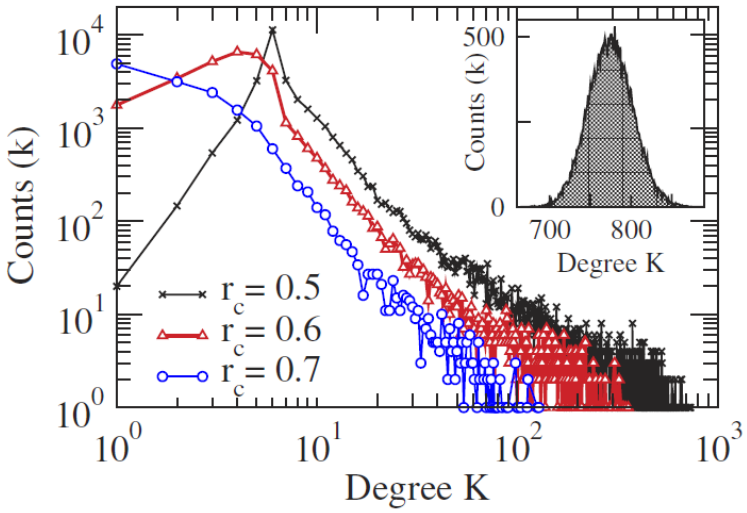
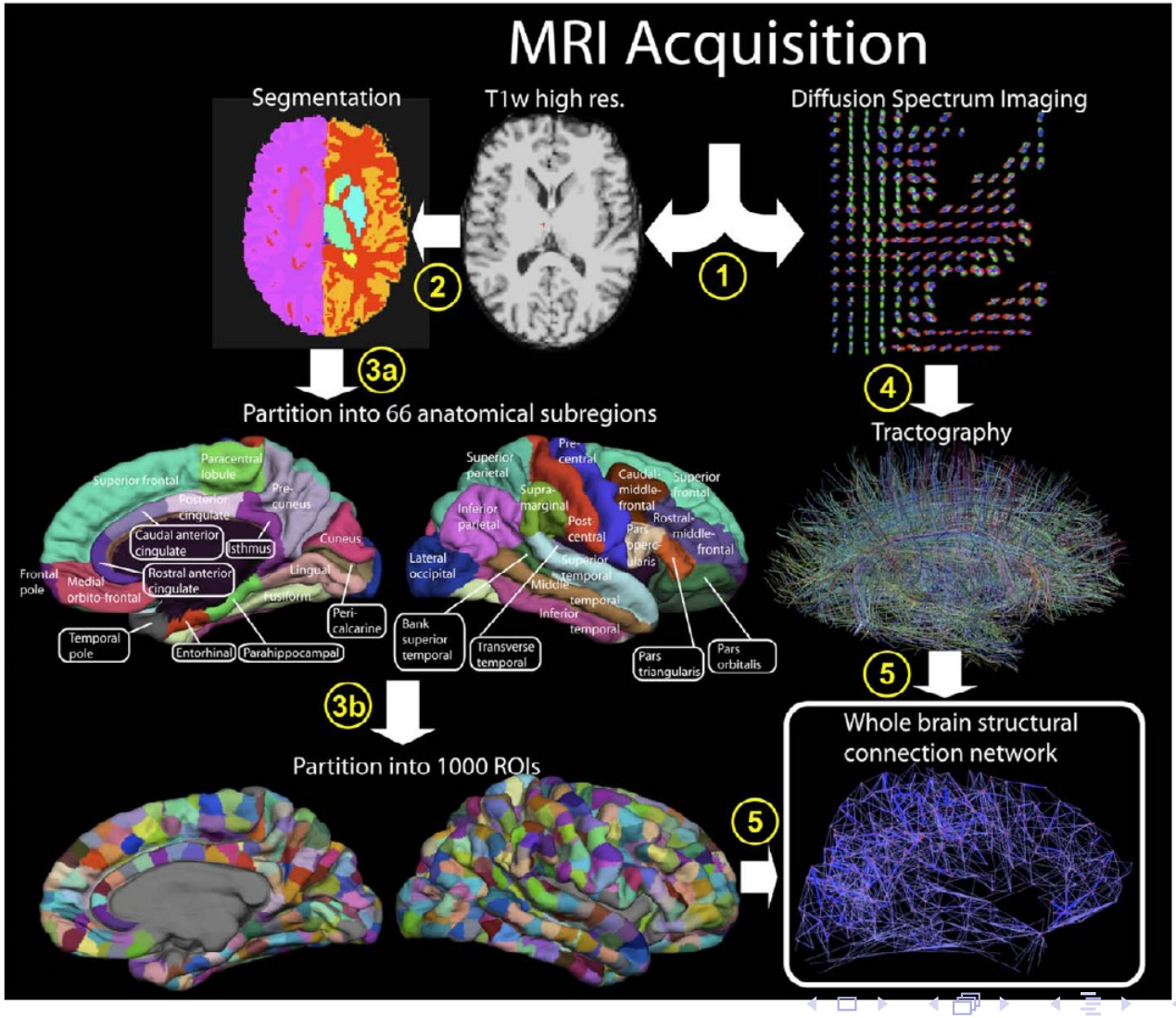


FIG. 2 (color online). Degree distribution for three values the correlation threshold. The inset depicts the degree distrib tion for an equivalent randomly connected network.

July 2008 | Volume 6 | Issue 7 | e159 PLOS BIOLOGY

Mapping the Structural Core of Human Cerebral Cortex

Patric Hagmann[1,2], Leila Cammoun[2], Xavier Gigandet[2], Reto Meuli[1], Christopher J. Honey[3], Van J. Wedeen[4], Olaf Sporns[3*]



Summary

- The production of proteins in cells is regulated using a complex regulation network
- Network motifs: simple building blocks of complex networks
- An algorithm to identify network motifs
- Example: the transcription network of E. coli.
- The feed forward loop as a sign sensitive filter
- The single input module: exact temporal ordering of protein expression

Biological Network Alignment

Slides from Johannes Berg, University of Cologne

Graph Alignment and Biological Networks

Johannes Berg

http://www.uni-koeln.de/~berg

Institute for Theoretical Physics

University of Cologne

Germany

Networks in molecular biology

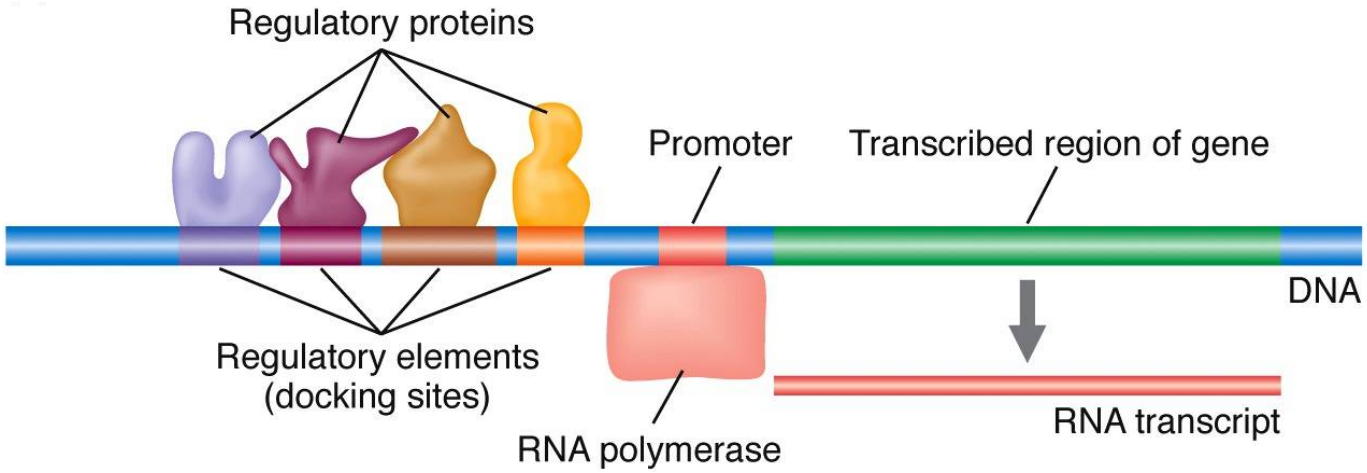
New large-scale experimental data in the form of networks:

- transcription networks
- protein interaction networks
- co-regulation networks
- signal transduction networks, metabolic networks, *etc.*

Networks in molecular biology

New large-scale experimental data in the form of networks:

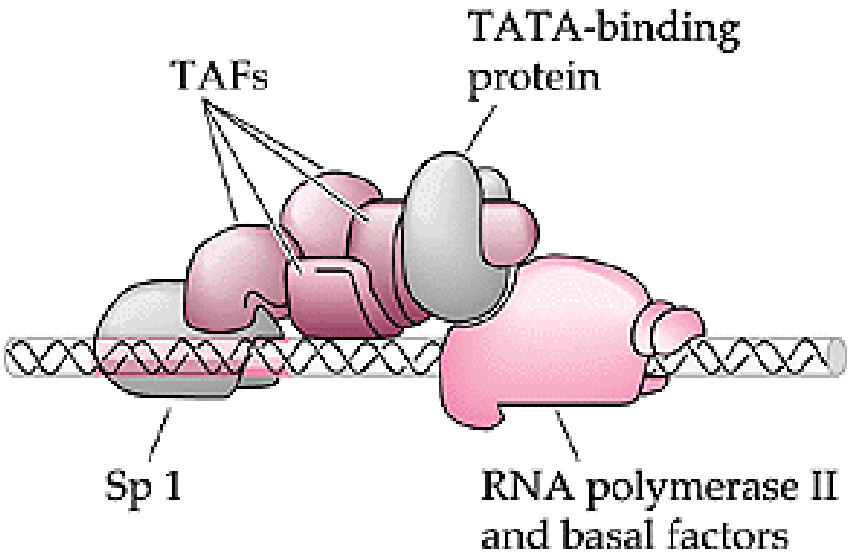
- transcription networks
  - transcription factors bind to regulatory DNA
  - polymerase molecule begins transcription of the gene



Networks in molecular biology

New large-scale experimental data in the form of networks:

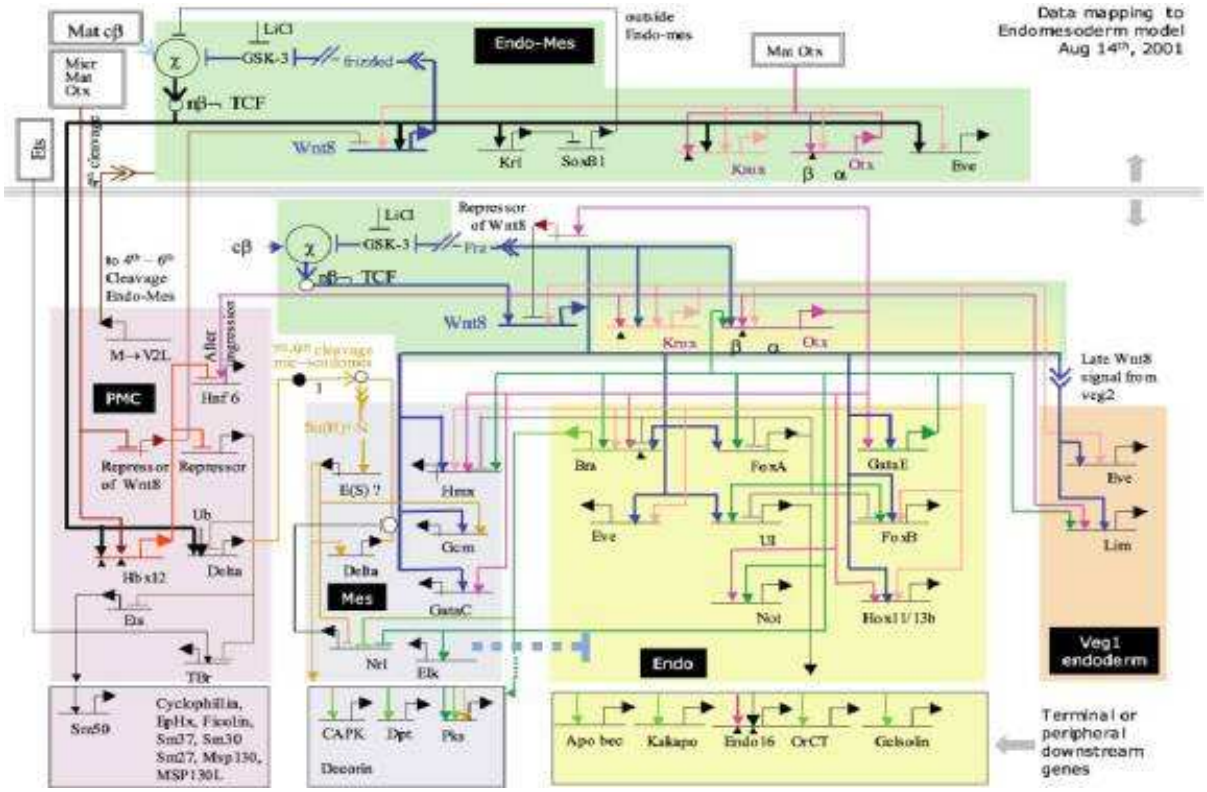
- transcription networks
  - transcription factors bind to regulatory DNA
  - polymerase molecule begins transcription of the gene



Networks in molecular biology

New large-scale experimental data in the form of networks:

- transcription networks
  - transcription factors bind to regulatory DNA
  - polymerase molecule begins transcription of the gene

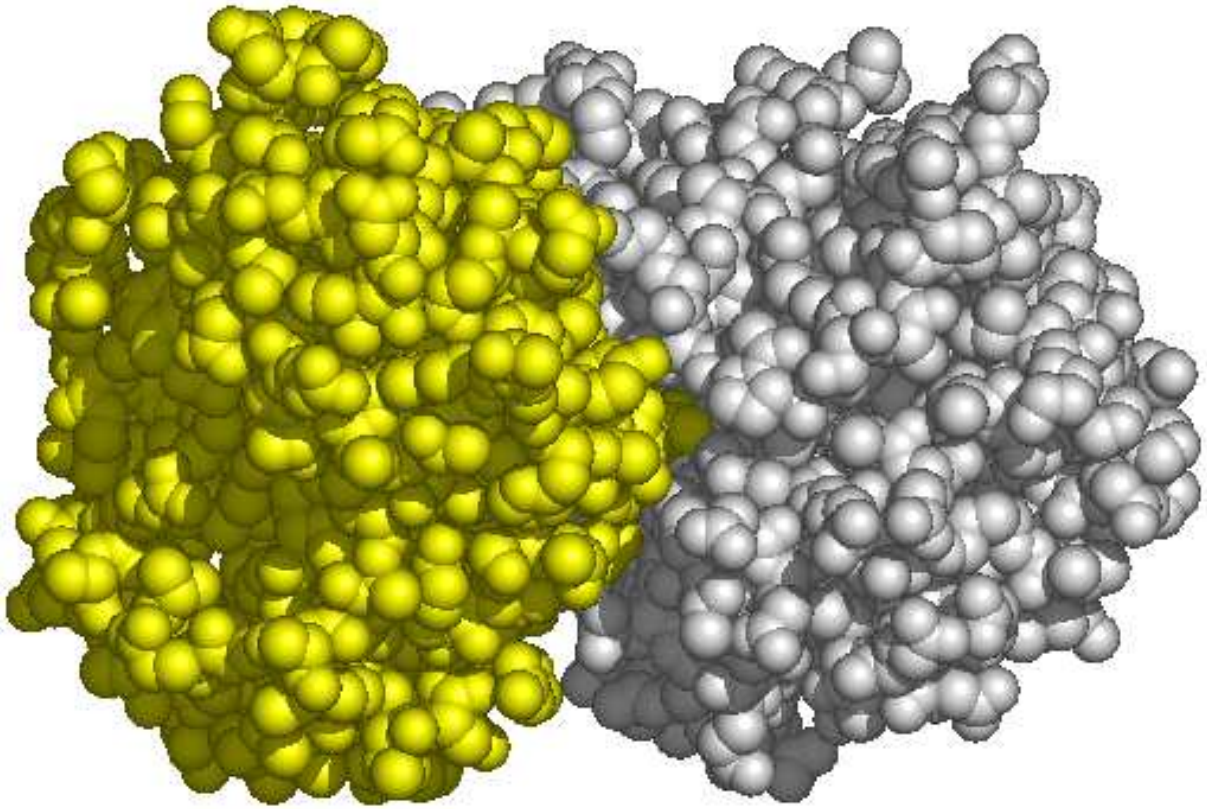


sea urchin
Bolouri &Davidson (2001)

Networks in molecular biology

New large-scale experimental data in the form of networks:

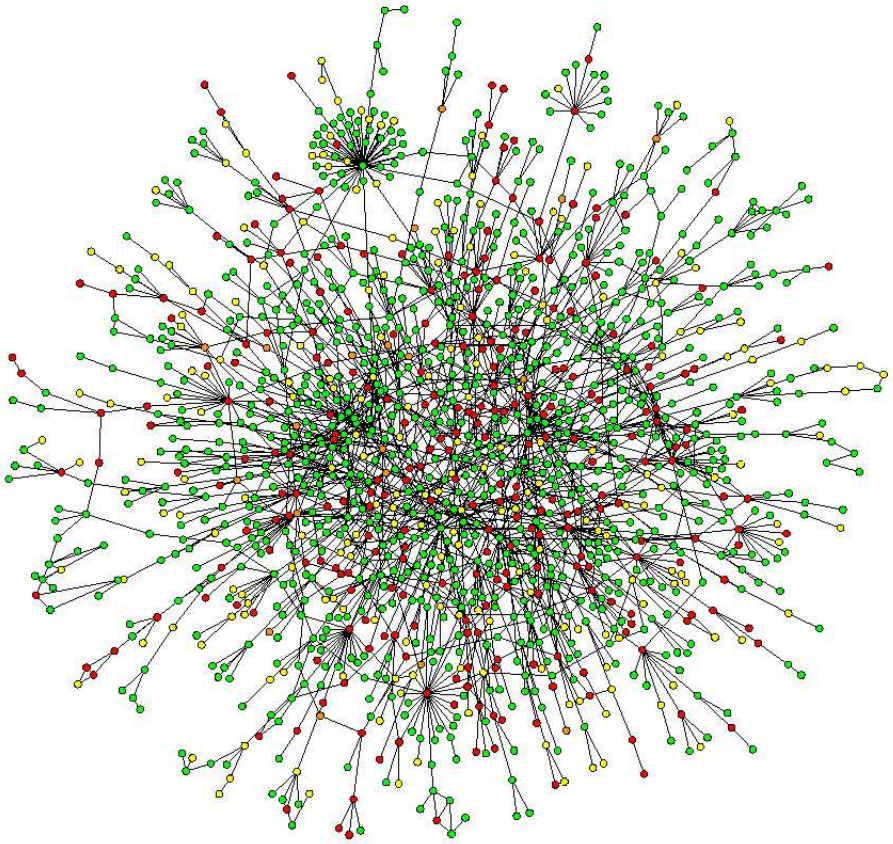
- protein interaction networks
  - proteins interact to form larger units
  - protein aggregates may catalyze reactions *etc.*



Networks in molecular biology

New large-scale experimental data in the form of networks:

- protein interaction networks
  - proteins interact to form larger units
  - protein aggregates may catalyze reactions *etc.*



protein interactions in yeast
Uetz *et al.* (2000)

Sequence alignment in molecular biology

- more than 100 organisms are fully sequenced
- genome sizes range from 3×10^7 to 7×10^{11} basepairs

Sequence alignment in molecular biology

- more than 100 organisms are fully sequenced
- genome sizes range from 3×10^7 to 7×10^{11} basepairs

Global alignment: search for related sequences across species

- evolutionary relationships
- hints at common functionality

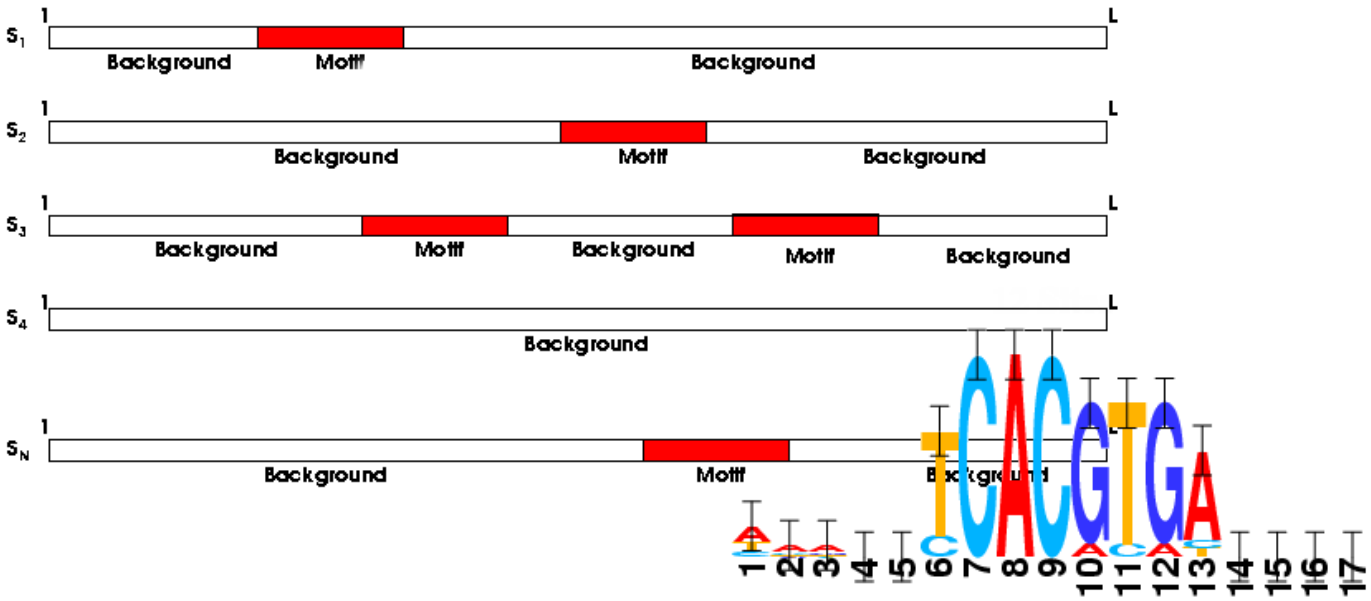
```
                10        20        30        40        50        60        70
SEQ1   VHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAV
       :: ::.::  ..::::::::: : :.:::.:::.::::::::::.::::.::. .:..:: :.::::::::.:::...
SEQ2   VHLTADEKAAVSGLWGKVNVDEVGGEALGRLLVVYPWTQRFFTSFGDLSNAAAVMGNSKVKAHGKKVLNSFGEGL
                10        20        30        40        50        60        70

          80        90       100       110       120       130       140
SEQ1   KNLDNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH
       ::.::.:.::..::::::::::::::::::::::..:.:::: ::.:.:::. :.:.:::.  :: :::.:::
SEQ2   KNVDNLKGTFASLSELHCDKLHVDPENFRLLGNVLVIVLARHFGKEFTPQVQGAFQKLALGVATALAHKYH
          80        90       100       110       120       130       140
```

Sequence alignment in molecular biology

- more than 100 organisms are fully sequenced
- genome sizes range from 3×10^7 to 7×10^{11} basepairs

Motif search: search for short repeated subsequences

- binding sites in transcription control



Sequence alignment in molecular biology

- more than 100 organisms are fully sequenced
- genome sizes range from 3×10^7 to 7×10^{11} basepairs

Tools

- statistical models are used infer non-random correlations against a background
- build score function from statistical models
- design efficient algorithms to maximize score
- evaluate statistical significance of a given score

Sequence alignment in molecular biology

- more than 100 organisms are fully sequenced
- genome sizes range from 3×10^7 to 7×10^{11} basepairs

Tools

- statistical models are used infer non-random correlations against a background
- build score function from statistical models
- design efficient algorithms to maximize score
- evaluate statistical significance of a given score

organism	number of genes
worm *C. elegans*	19 000
fruit fly *drosophila*	17 000
human *homo sapiens*	$\lesssim$ 25 000

Graph alignment

What can be learned from network data?
Can we distinguish functional patterns from a random background?

1. Search for network motifs [Alon lab]
   - patterns occurring repeatedly within a given network
2. Alignment of networks across species
   - identify conserved regions
   - pinpoint functional innovations

Graph alignment

What can be learned from network data?
Can we distinguish functional patterns from a random background?

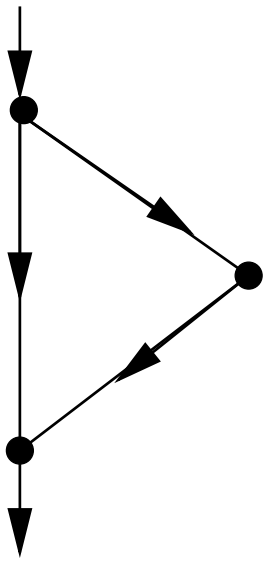
1. Search for network motifs [Alon lab]
   - patterns occurring repeatedly within a given network
2. Alignment of networks across species
   - identify conserved regions
   - pinpoint functional innovations

Tools

- scoring function based on statistical models
- heuristic algorithms: algorithmic complexity

Graph alignment I: The search for network motifs

- patterns occurring repeatedly in the network
- building blocks of information processing [Alon lab]



Graph alignment I: The search for network motifs

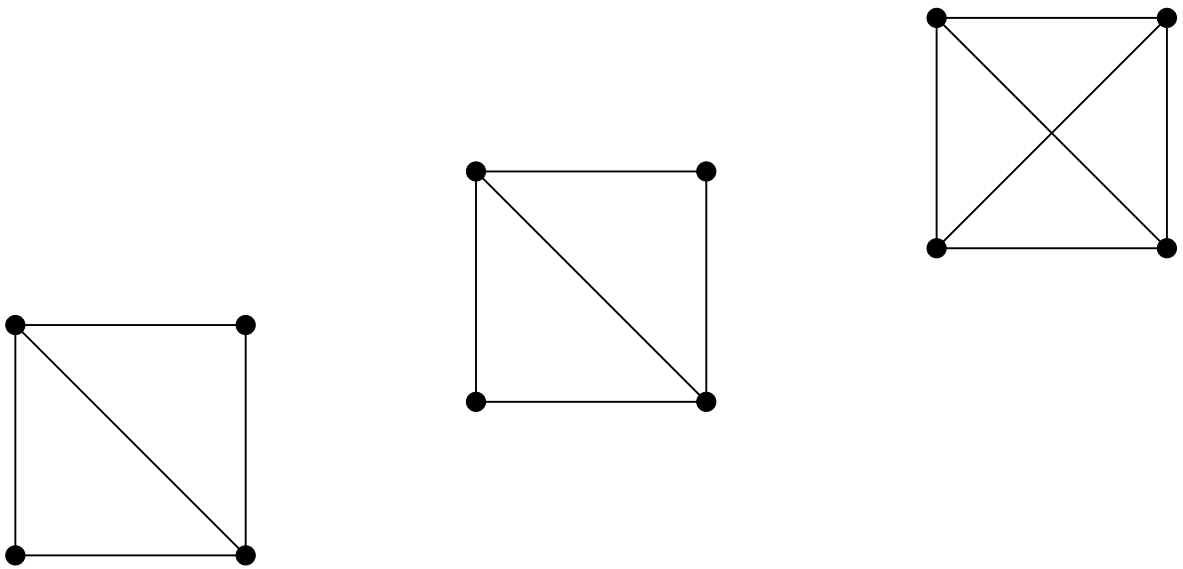
- patterns occurring repeatedly in the network
- building blocks of information processing [Alon lab]

- counting of identical patterns: Subgraph census
- alignment of topologically similar regions of a network
- allow for mismatches
- construct a scoring function comparing the aligned subgraphs to a background model

Graph alignment I: The search for network motifs

- patterns occurring repeatedly in the network
- building blocks of information processing [Alon lab]

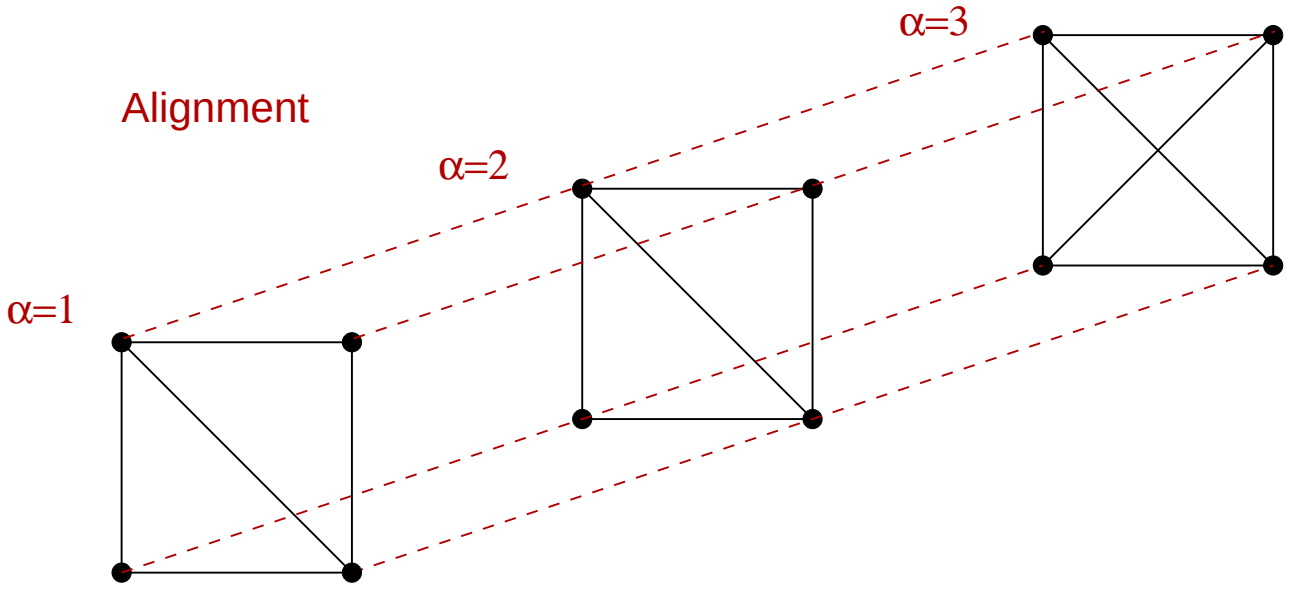
- counting of identical patterns: Subgraph census
- alignment of topologically similar regions of a network
- allow for mismatches
- construct a scoring function comparing the aligned subgraphs to a background model



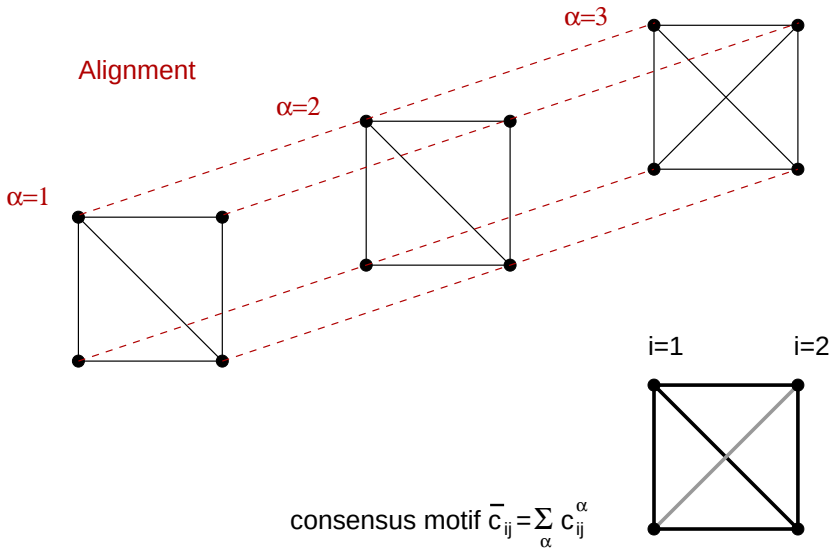
Graph alignment I: The search for network motifs

- patterns occurring repeatedly in the network
- building blocks of information processing [Alon lab]

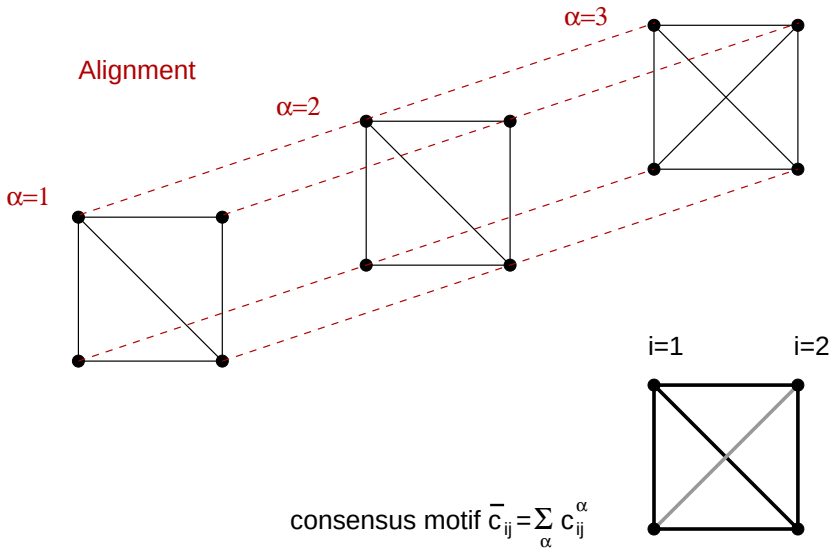
- counting of identical patterns: Subgraph census
- alignment of topologically similar regions of a network
- allow for mismatches
- construct a scoring function comparing the aligned subgraphs to a background model



Statistical properties of alignments



Statistical properties of alignments



- consensus motif $\overline{\mathbf{c}} = \frac{1}{p} \sum_{\alpha=1}^{p} \mathbf{c}^{\alpha}$
- number of internal links
- average correlation between two subgraphs fuzziness of motif

Statistics of network motifs

null model:

- ensemble of uncorrelated networks with the same connectivities as the data

Statistics of network motifs

null model:

- ensemble of uncorrelated networks with the same connectivities as the data

model describing network motifs

- ensemble with enhanced number of links
- enhanced correlation of subgraphs divergent vs convergent evolution?

Statistics of network motifs

null model:

- ensemble of uncorrelated networks with the same connectivities as the data

model describing network motifs

- ensemble with enhanced number of links
- enhanced correlation of subgraphs divergent vs convergent evolution?

Log likelihood score

$$
\begin{aligned}
S(\mathbf{c}^1, \ldots, \mathbf{c}^p) &= \log\left(\frac{Q(\mathbf{c}^1, \ldots, \mathbf{c}^p)}{\prod_{\alpha=1}^{p} P_\sigma(\mathbf{c}^\alpha)}\right) \\
&= (\sigma - \sigma_0)\sum_{\alpha=1}^{p} L(\mathbf{c}^\alpha) - \frac{\mu}{2p}\sum_{\alpha,\beta=1}^{p} M(\mathbf{c}^\alpha, \mathbf{c}^\beta) - \log Z
\end{aligned}
$$

Statistics of network motifs

null model:

- ensemble of uncorrelated networks with the same connectivities as the data

model describing network motifs

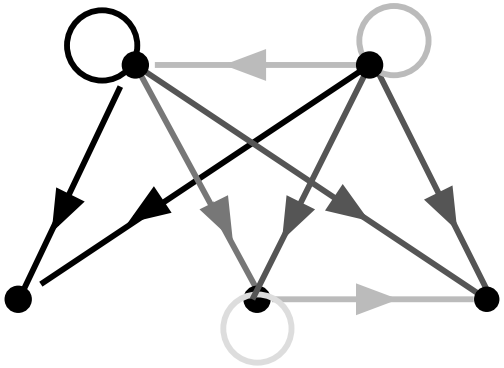
- ensemble with enhanced number of links
- enhanced correlation of subgraphs divergent vs convergent evolution?

Log likelihood score

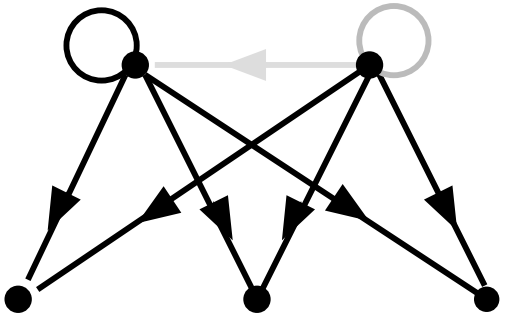
$$
\begin{aligned}
S(\mathbf{c}^1, \ldots, \mathbf{c}^p) &= \log\left(\frac{Q(\mathbf{c}^1, \ldots, \mathbf{c}^p)}{\prod_{\alpha=1}^{p} P_\sigma(\mathbf{c}^\alpha)}\right) \\
&= (\sigma - \sigma_0)\sum_{\alpha=1}^{p} L(\mathbf{c}^\alpha) - \frac{\mu}{2p}\sum_{\alpha,\beta=1}^{p} M(\mathbf{c}^\alpha, \mathbf{c}^\beta) - \log Z
\end{aligned}
$$

Algorithm: Mapping onto a model from statistical mechanics (Potts model)

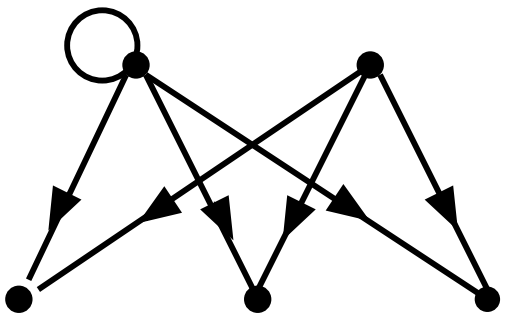
Consensus motif of the *E. coli* transcription network



$\mu = \mu^* = 2.25$

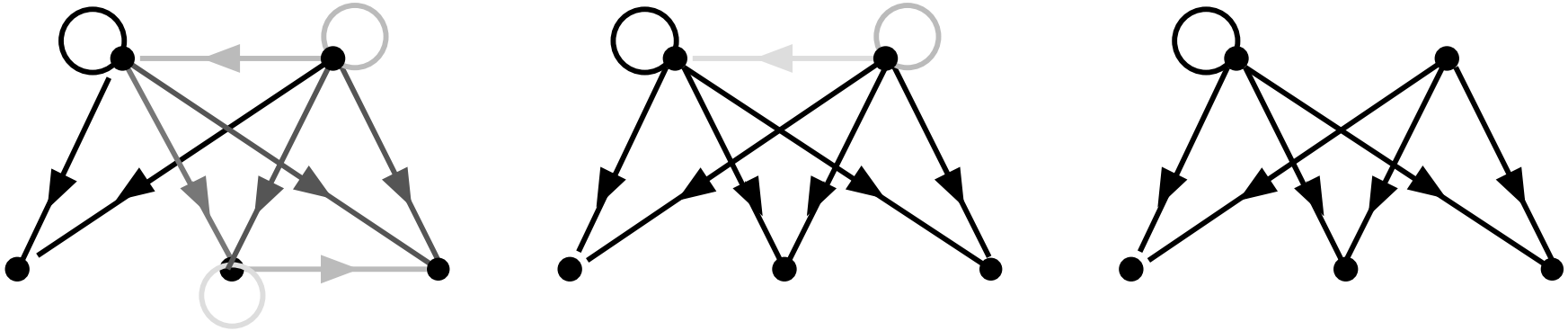


$\mu = 5$

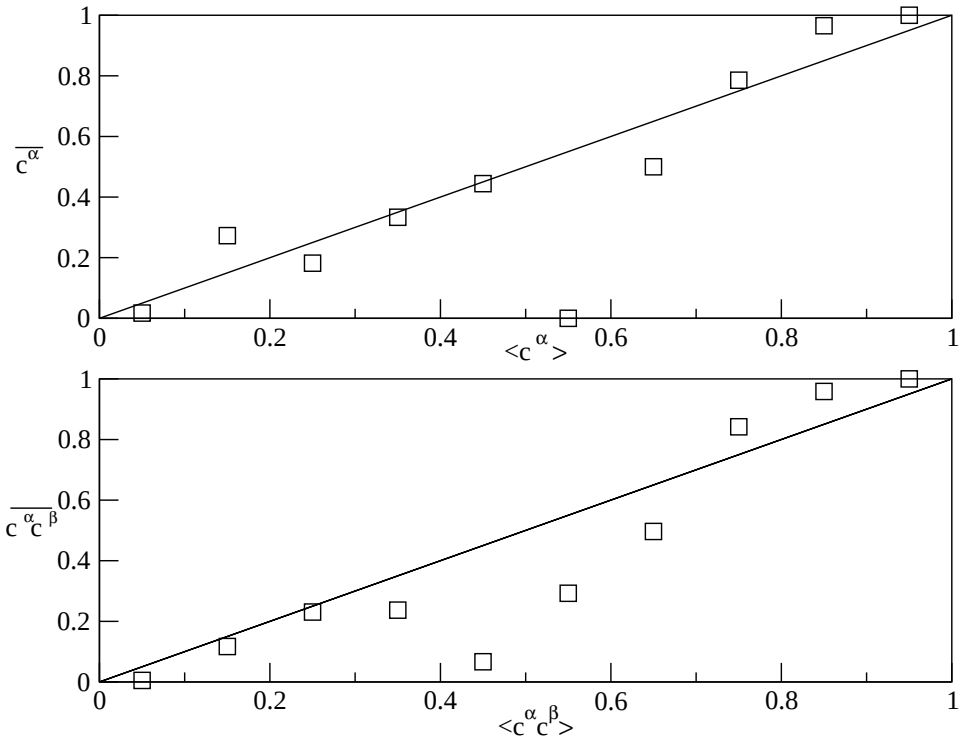


$\mu = 12$

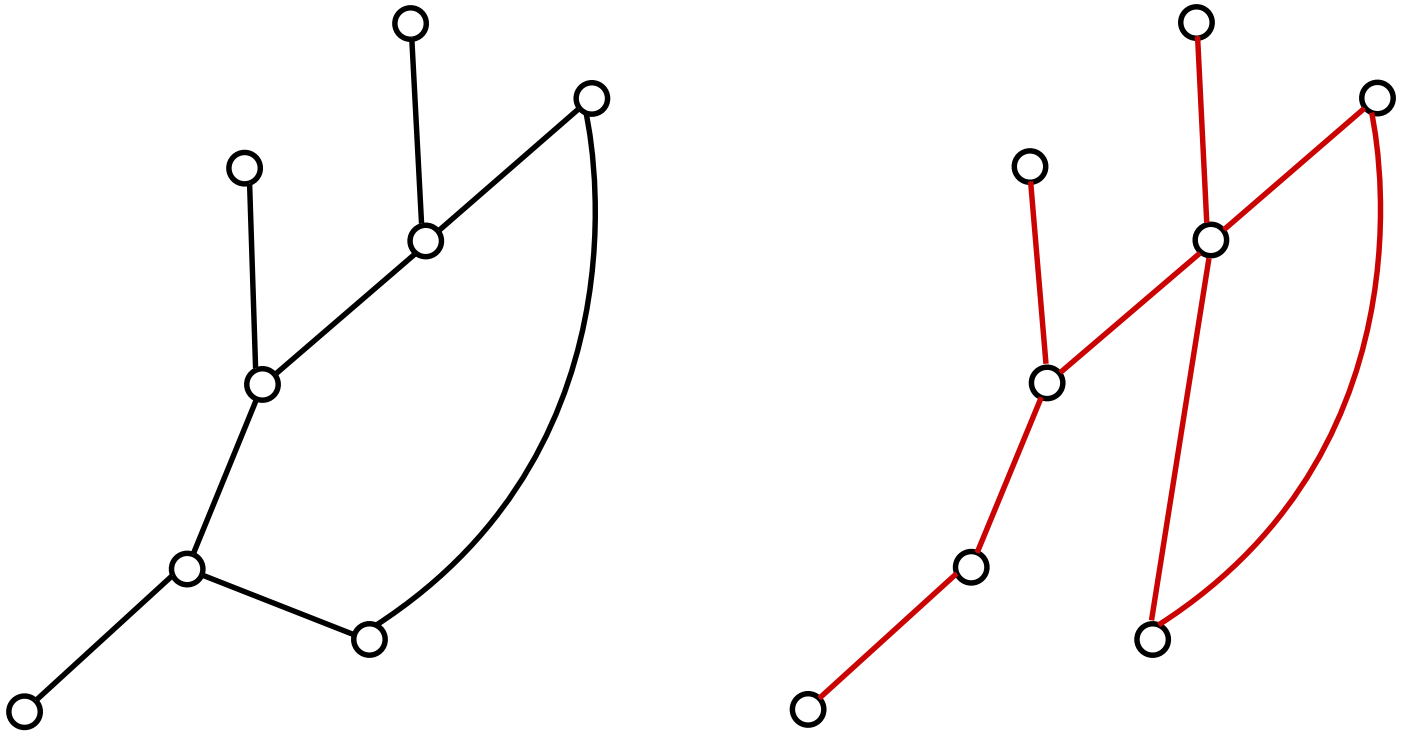
Consensus motif of the *E. coli* transcription network



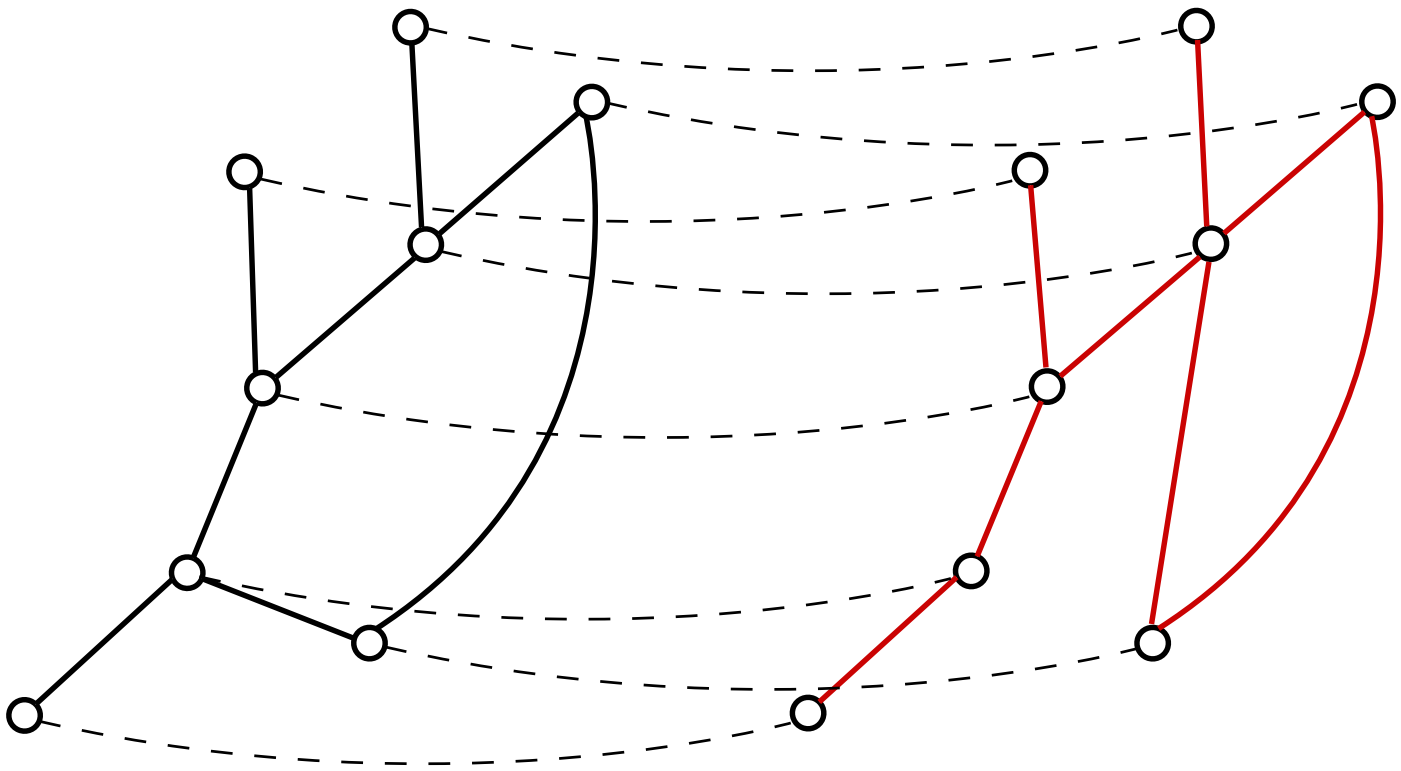
$\mu = \mu^* = 2.25$ $\mu = 5$ $\mu = 12$



Graph alignment II: Comparing networks across species

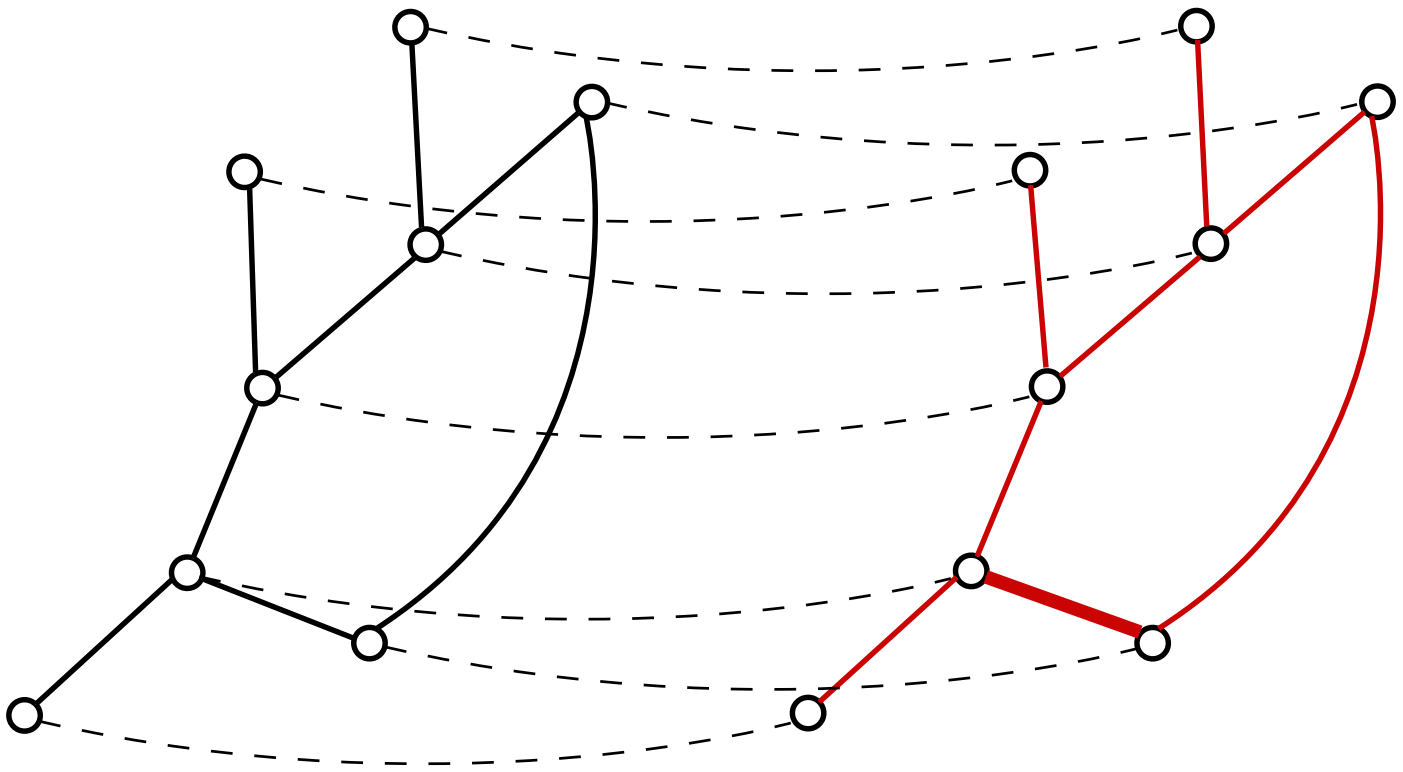


Graph alignment II: Comparing networks across species



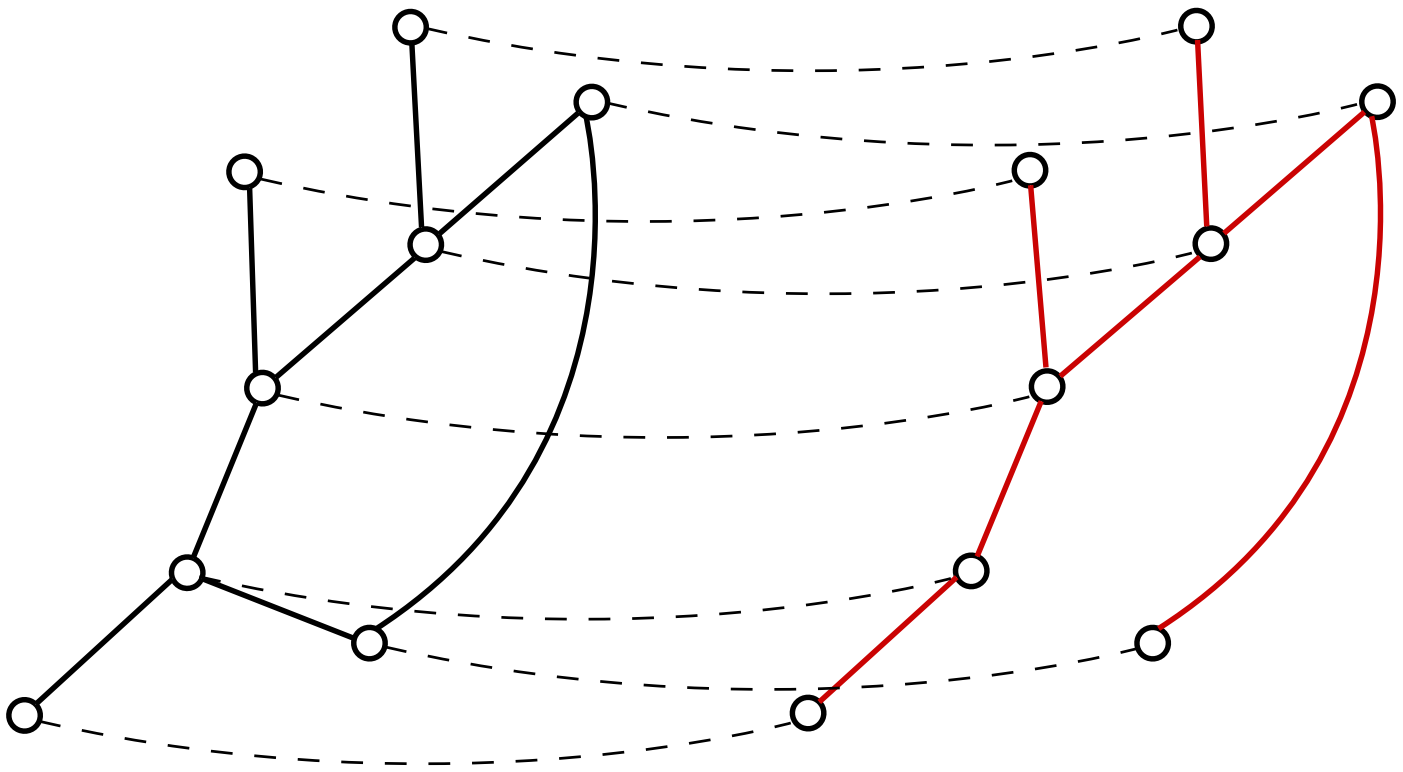
Alignment: Pairwise association of nodes across species

Graph alignment II: Comparing networks across species



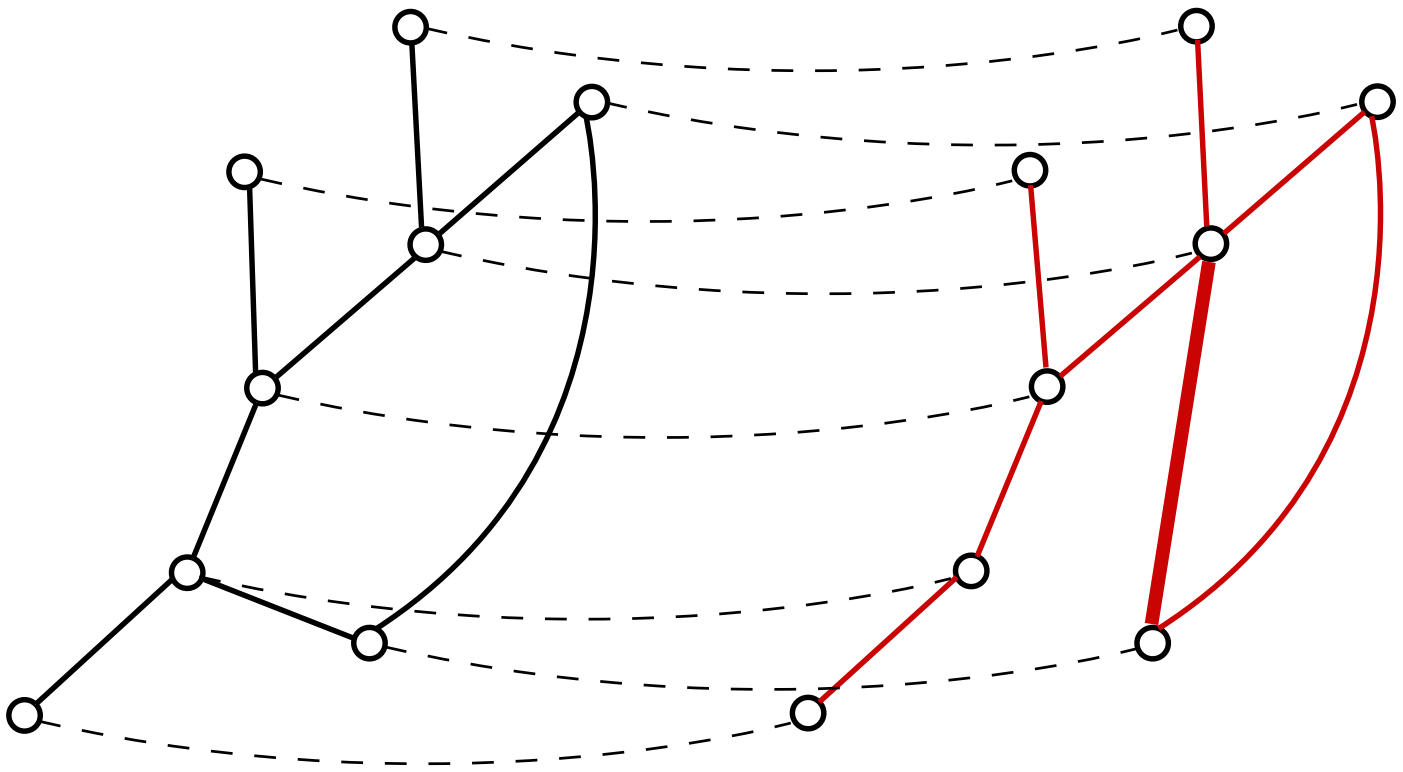
Last common ancestor

Graph alignment II: Comparing networks across species



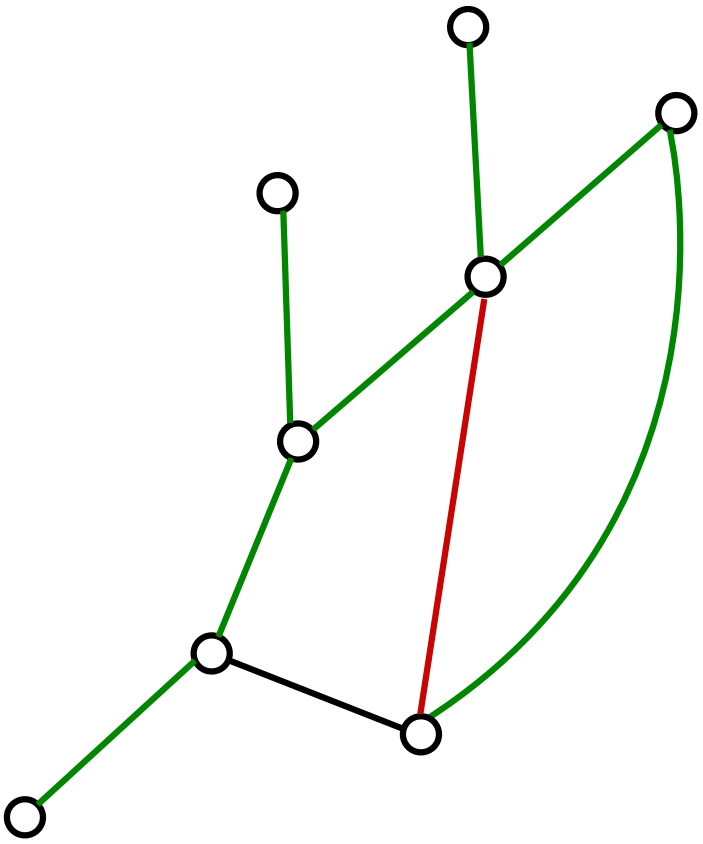
Evolutionary dynamics: Link attachment and deletion

Graph alignment II: Comparing networks across species



Evolutionary dynamics: Link attachment and deletion

Graph alignment II: Comparing networks across species



Representation of the alignment in a single network. Conserved links are shown in green.

Scoring graph alignments across species

null model P:

- ensemble of uncorrelated networks with the same connectivities as the data

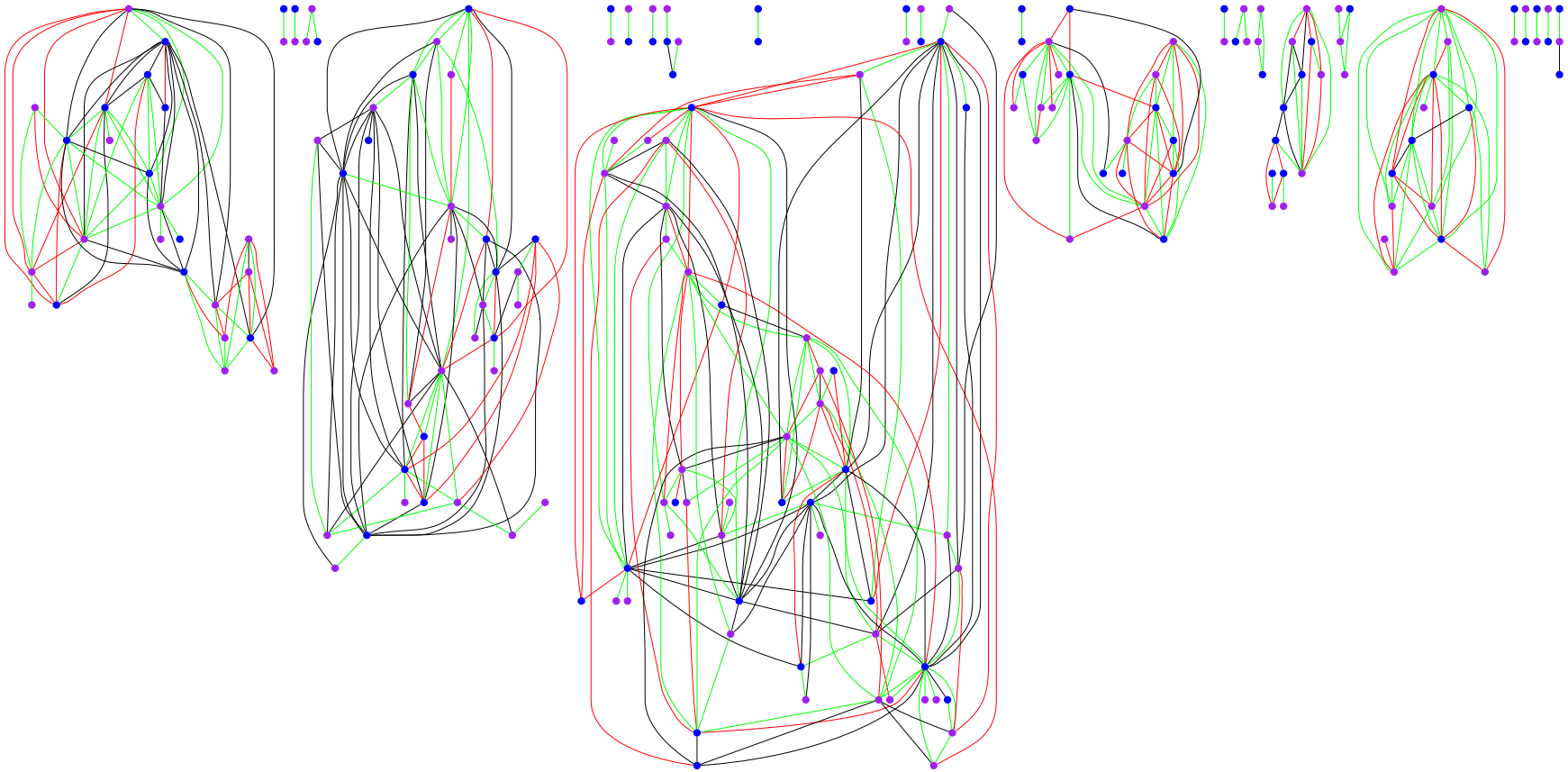
Q-model

- correlated networks (due to functional constraints or common ancestry)
- statistical assessment of orthologs: interplay between sequence similarity and network topology

Scoring alignments

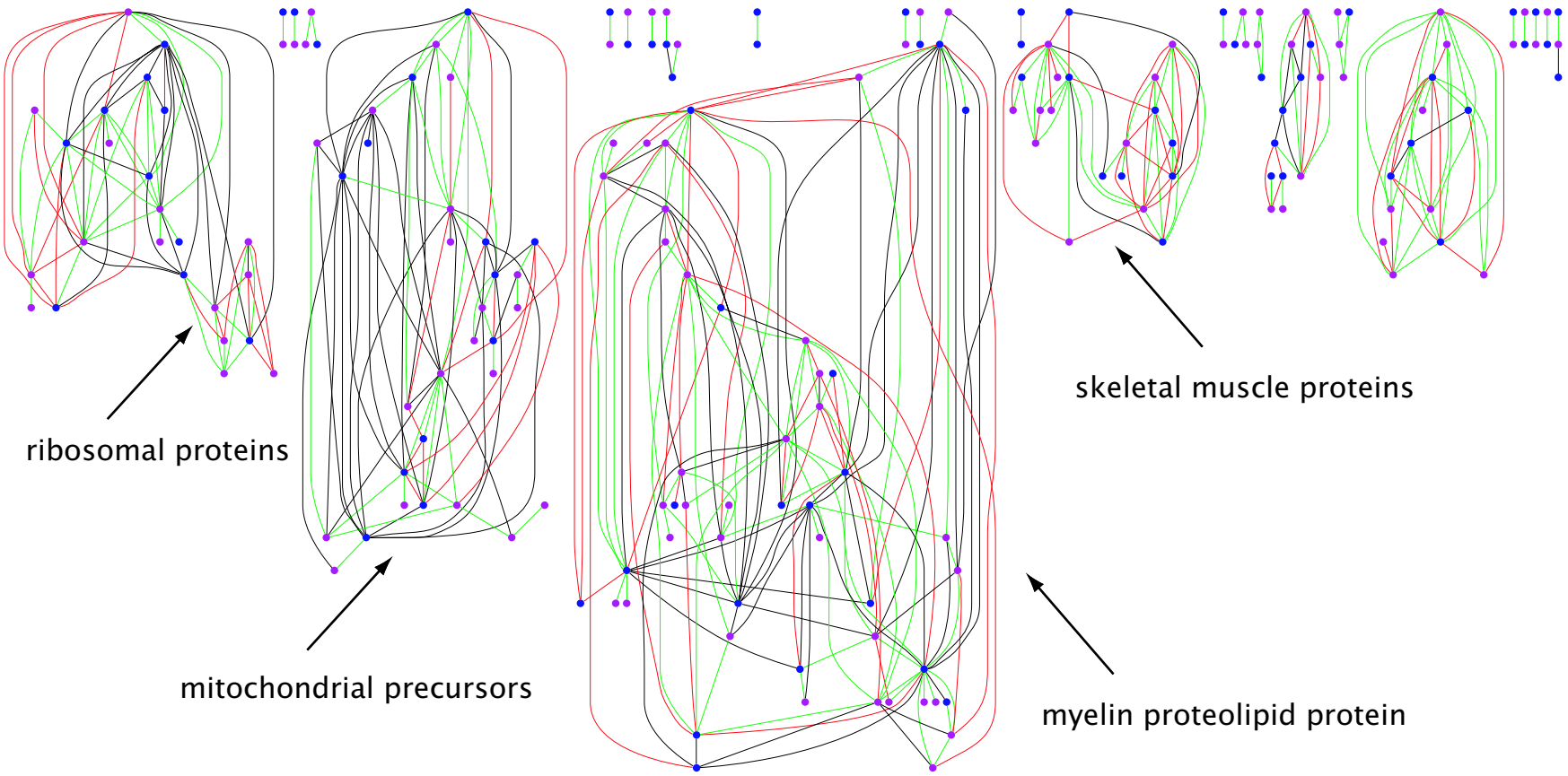
- log-likelihood score $S = \log(Q/P)$ is used to search for conserved parts of the networks

Application to Co-Expression networks



alignment of H. sapiens and M. musculus

Application to Co-Expression networks



alignment of H. sapiens and M. musculus

Genomic systems biology and network analysis

New concept and tools are needed to fully utilize high-throughput data

- functional design versus noise: statistical analysis
- evolutionary conservation indicates function

Topological conservation versus sequence conservation

- genes may change functional role in network with small corresponding change in sequence
- the role of a gene in one species may be taken on by an entirely unrelated gene in another species

References:

- J. Berg and M. Lässig, "Local graph alignment and motif search in biological networks", *Proc. Natl. Acad. Sci. USA*, **101** (41) 14689-14694 (2004)
- J. Berg, M. Lässig, and A. Wagner, “Structure and Evolution of Protein Interaction Networks: A Statistical Model for Link Dynamics and Gene Duplications”, *BMC Evolutionary Biology* **4**:51 (2004)
- J. Berg, S. Willmann und M. Lässig, “Adaptive evolution of transcription factor binding sites”, *BMC Evolutionary Biology* **4**(1):42 (2004)
- J. Berg and M. Lässig, "Correlated random networks", *Phys. Rev. Lett.* **89**(22), 228701 (2002)

Detecting Signaling Pathways using Color-coding

Slides from Hüffner et al., Friedrich-Schiller-Universität Jena

Algorithm Engineering for Color-Coding to Facilitate Signaling Pathway Detection

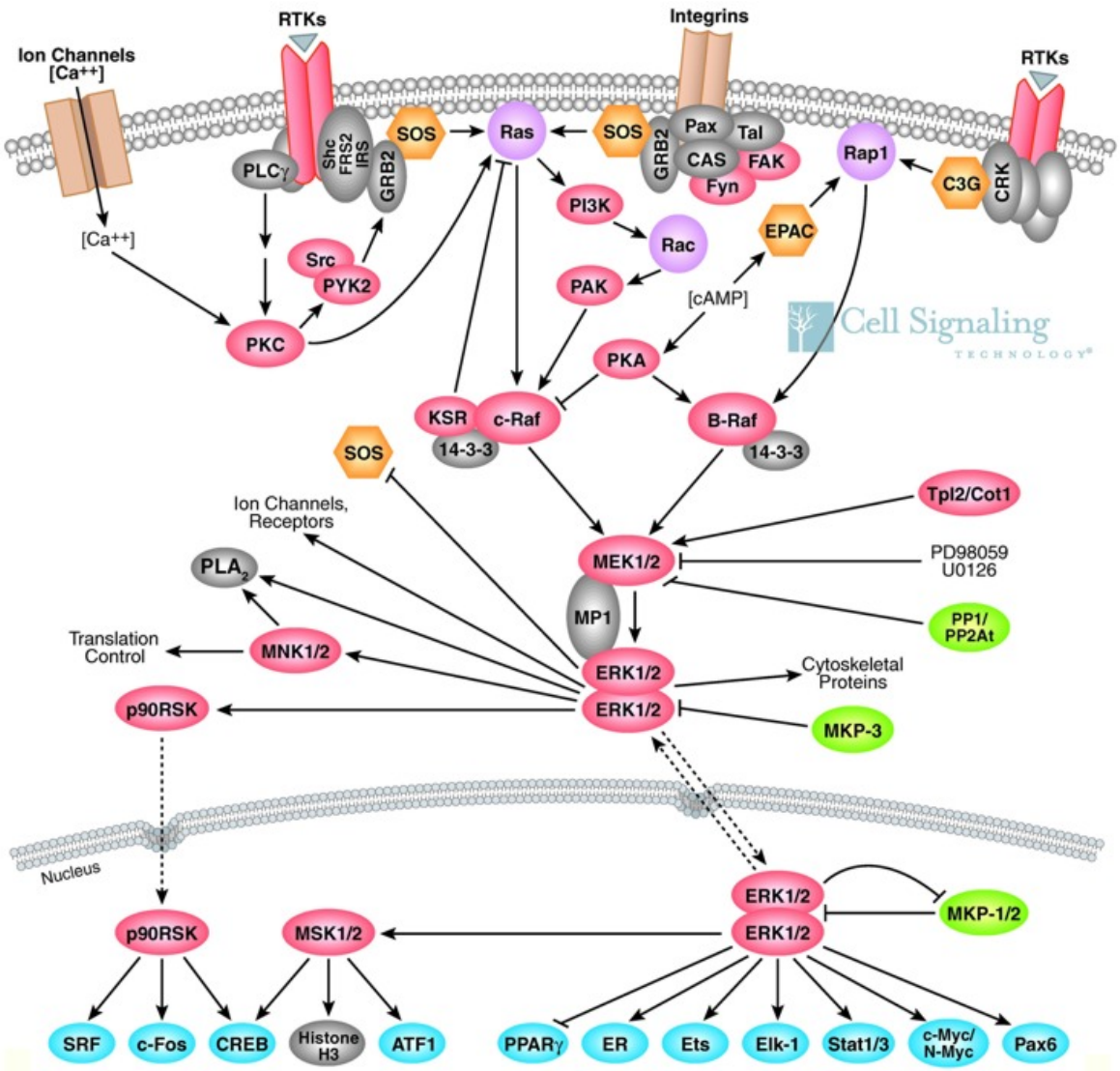
Falk Hüffner Sebastian Wernicke Thomas Zichner

Friedrich-Schiller-Universität Jena

Fifth Asia Pacific Bioinformatics Conference
January 17, 2007

Outline

Protein Interaction Networks



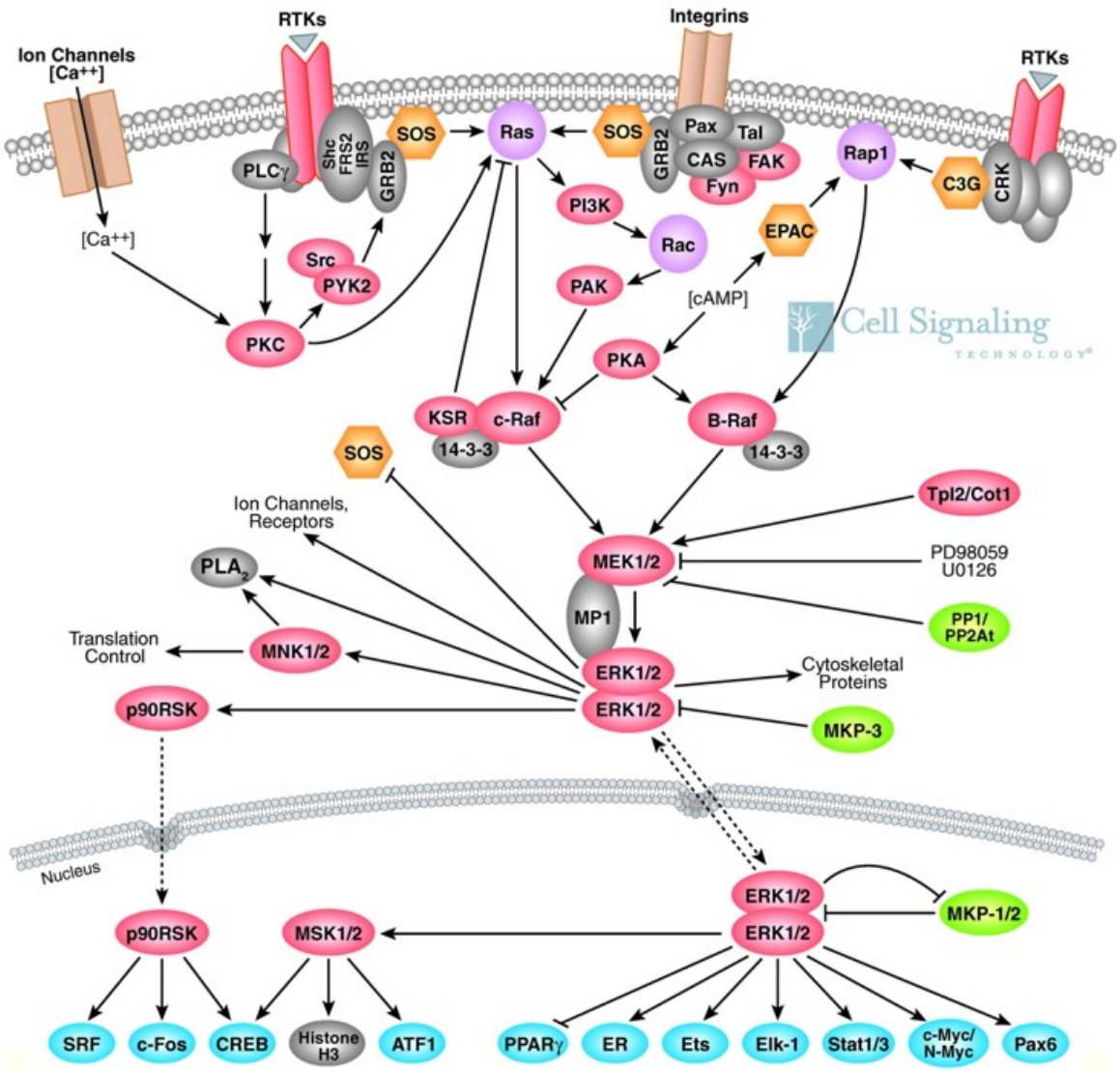
[www.cellsignal.com]

Protein Interaction Networks

Representation of protein interactions as a graph:

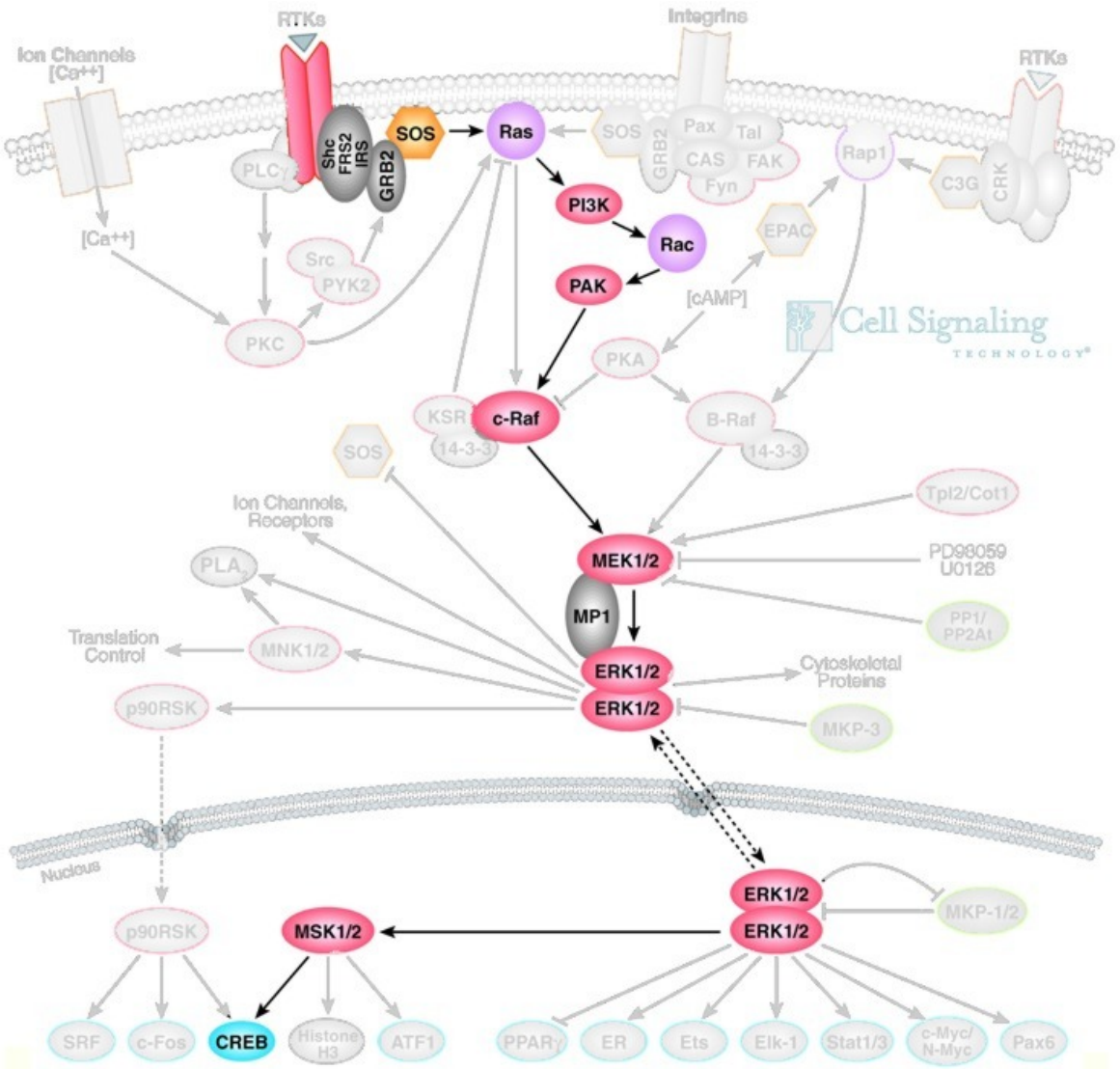
- Proteins are nodes
- Interactions are edges
- Edges are annotated with interaction probability (obtained by two-hybrid screening)

Signaling Pathways



[www.cellsignal.com]

Signaling Pathways



[www.cellsignal.com]

Signaling Pathways

Sequence of distinct proteins, where each interacts strongly with the previous one.

Most Probable Path

Input: Graph $G = (V, E)$, interaction probabilities $p : E \to [0, 1]$, integer $k > 0$.

Task: Find a non-overlapping path $v_1, \ldots, v_k$ of length k in G that maximizes $p(v_1, v_2) \cdot \ldots \cdot p(v_{k-1}, v_k)$.

Signaling Pathways

Sequence of distinct proteins, where each interacts strongly with the previous one.

Most Probable Path

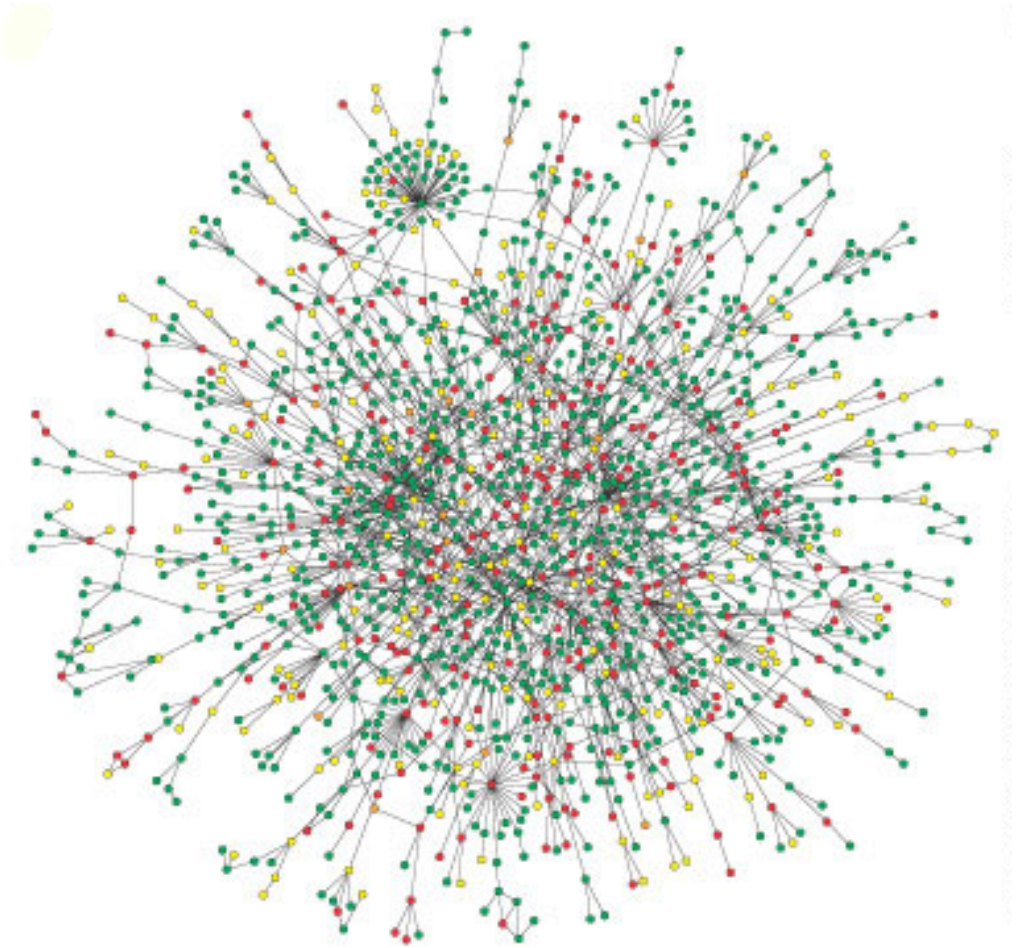
Input: Graph $G = (V, E)$, interaction probabilities $p : E \to [0, 1]$, integer $k > 0$.
Task: Find a non-overlapping path $v_1, \ldots, v_k$ of length k in G that maximizes $p(v_1, v_2) \cdot \ldots \cdot p(v_{k-1}, v_k)$.

Setting $w(e) := -\log(p(e))$:

Minimum-Weight Path

Input: Graph $G = (V, E)$, weights $w : E \to [0, 1]$, integer $k > 0$.
Task: Find a non-overlapping path $v_1, \ldots, v_k$ of length k in G that minimizes $w(v_1, v_2) + \cdots + w(v_{k-1}, v_k)$.

Yeast Network



4 400 proteins, 14 300 interactions, looking for paths of length 5–15

Minimum-Weight Path

Theorem

MINIMUM-WEIGHT PATH *is NP-hard* [GAREY&JOHNSON *1979*].

For an exact algorithm, we have to accept exponential runtime.

Idea

Exploit the fact that the paths sought for are rather short ($\approx$ 5–15): restrict the exponential part of the runtime to k (parameterized complexity).

Color-Coding

Color-coding [Alon, Yuster&Zwick J. ACM 1995]:

- randomly color each vertex of the graph with one of k colors
- hope that all vertices in the subgraph searched for obtain different colors (colorful)
- solve the Minimum-Weight Path under this assumption (which is much quicker)
- repeat until it is reasonably certain that the path was colorful at least once

Result: exponential part of the runtime depends only on k

Dynamic Programming for Minimum-Weight Colorful Path

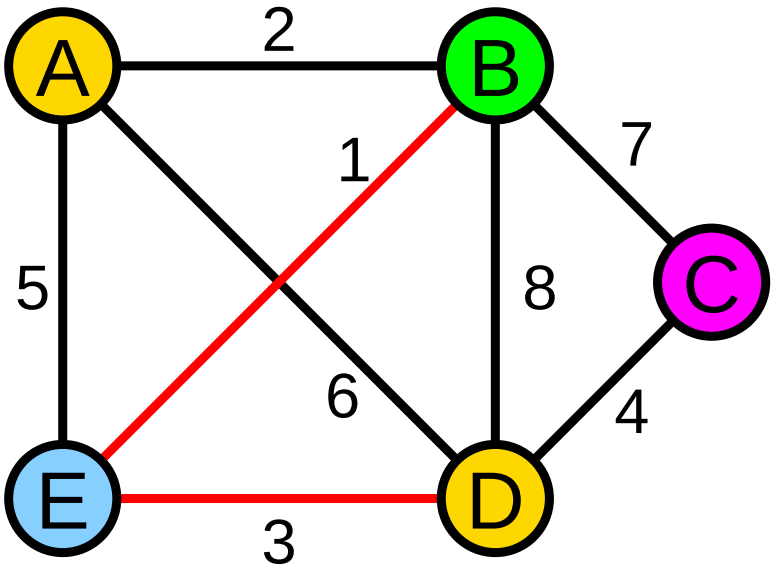
Idea

Table entry $W[v, C]$ stores the minimum-weight path that ends in v and uses exactly the colors in S.

Dynamic Programming for Minimum-Weight Colorful Path

Idea

Table entry $W[v, C]$ stores the minimum-weight path that ends in v and uses exactly the colors in S.



$W[B, \{$ ●, ●, ● $\}] = 4$

Dynamic Programming for Minimum-Weight Colorful Path

Coloring $c : V \to \{1, \ldots, k\}$

Recurrence

$$W[v, C] = \min_{u \in N(v) | c(u) \in C \setminus \{c(v)\}} (W[u, C \setminus \{c(v)\}] + w(u, v))$$

Dynamic Programming for Minimum-Weight Colorful Path

Coloring $c : V \to \{1, \ldots, k\}$

Recurrence

$$W[v, C] = \min_{u \in N(v) | c(u) \in C \setminus \{c(v)\}} (W[u, C \setminus \{c(v)\}] + w(u, v))$$

- Each table entry can be calculated in $O(n)$ time
- $n2^k$ table entries

$\rightsquigarrow$ Runtime: $O(n \cdot n2^k) = n^2 \cdot 2^k$

Color-coding Runtime

- $O(n^2 \cdot 2^k)$ time per trial
- To obtain error probability ε, one needs $O(|\ln \varepsilon| \cdot e^k)$ trials

Theorem ([Alon et al. JACM 1995])

Minimum-Weight Path *can be solved in* $O(|\ln \varepsilon| \cdot 5.44^k |G|)$ *time).*

Color-coding Runtime

- $O(n^2 \cdot 2^k)$ time per trial
- To obtain error probability ε, one needs $O(|\ln \varepsilon| \cdot e^k)$ trials

Theorem ([Alon et al. JACM 1995])

Minimum-Weight Path *can be solved in* $O(|\ln \varepsilon| \cdot 5.44^k |G|)$ *time).*

Color-coding can find minimum-weight paths of length 10 in the yeast protein interaction networks within 3 hours ($n = 4\,400, k = 10$) [Scott et al., RECOMB'05]

Increasing the Number of Colors

Idea

Use $k + x$ colors instead of k colors.

Trial runtime:

$$O(2^k|G|) \rightarrow O(2^{k+x}|G|)$$

Increasing the Number of Colors

Idea

Use $k + x$ colors instead of k colors.

Trial runtime:

$$O(2^k|G|) \to O(2^{k+x}|G|)$$

Probability P_c for colorful path ($k = 8$, $\varepsilon = 0.001$):

x	0	1	2	3	4	5
P_c	0.0024	0.0084	0.0181	0.0310	0.0464	0.0636
trials	2871	816	378	220	146	106

Increasing the Number of Colors

Idea

Use $k + x$ colors instead of k colors.

Trial runtime:

$$O(2^k|G|) \to O(2^{k+x}|G|)$$

Probability P_c for colorful path ($k = 8$, $\varepsilon = 0.001$):

x	0	1	2	3	4	5
P_c	0.0024	0.0084	0.0181	0.0310	0.0464	0.0636
trials	2871	816	378	220	146	106

Theorem

Minimum-Weight Path *can be solved in* $O(|\ln \varepsilon| \cdot 4.32^k|G|)$ *time by choosing* $x = 0.3k$.

Increasing the Number of Colors

Idea

Use $k + x$ colors instead of k colors.

Trial runtime:

$$O(2^k|G|) \to O(2^{k+x}|G|)$$

Probability P_c for colorful path ($k = 8$, $\varepsilon = 0.001$):

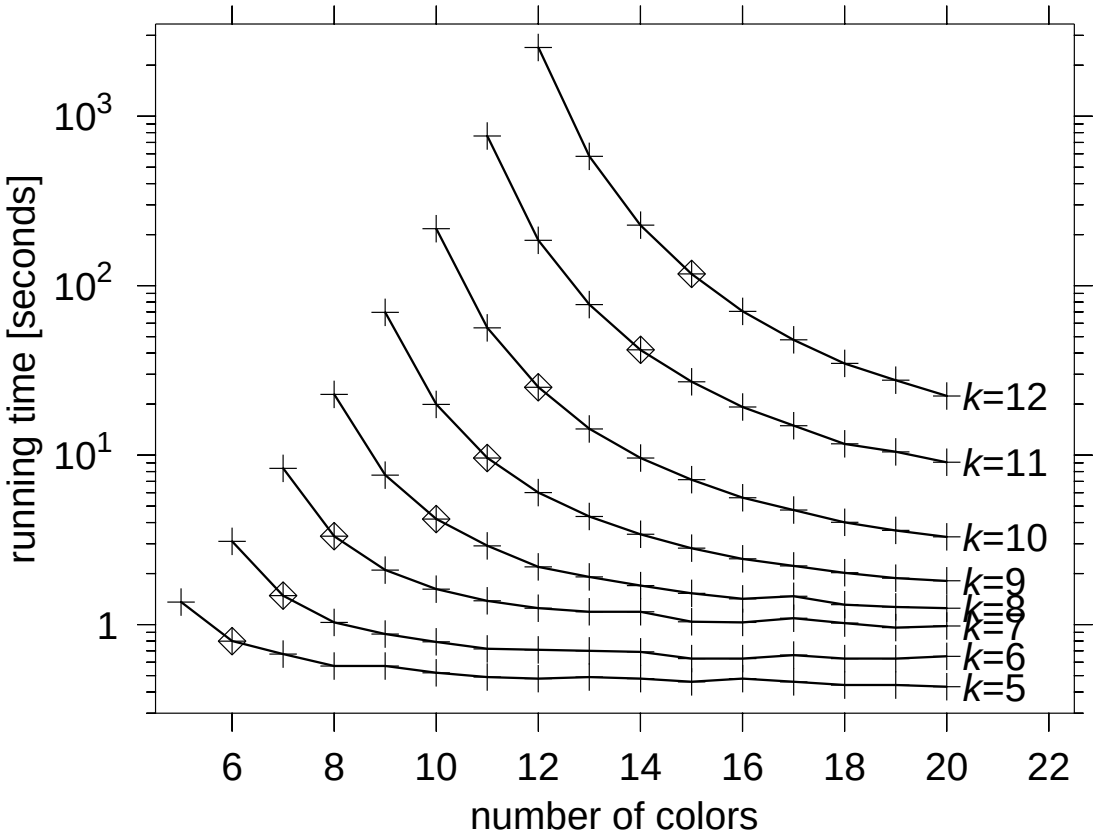
x	0	1	2	3	4	5
P_c	0.0024	0.0084	0.0181	0.0310	0.0464	0.0636
trials	2871	816	378	220	146	106

Theorem

Minimum-Weight Path *can be solved in* $O(|\ln \varepsilon| \cdot 4.32^k|G|)$ *time by choosing* $x = 0.3k$.

But: Higher memory usage

Increasing the Number of Colors



Runtimes for the yeast protein interaction network (highlighted point of each curve marks worst-case optimum)

Exploiting Lower Bounds

Idea

Use a known solution to prune "hopeless" table entries.

- Discard entries that already have a weight higher than the known solution.

Exploiting Lower Bounds

Idea

Use a known solution to prune "hopeless" table entries.

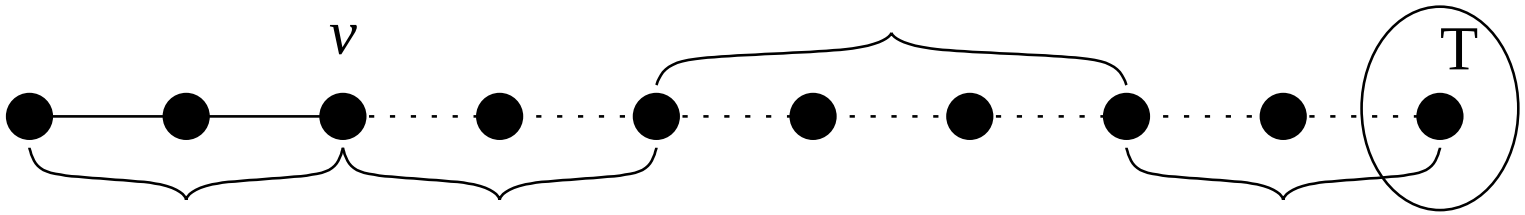
- Discard entries that already have a weight higher than the known solution.
- Discard entries when

$$\text{weight} + (\text{minimum edge weight} \cdot \text{edges left})$$

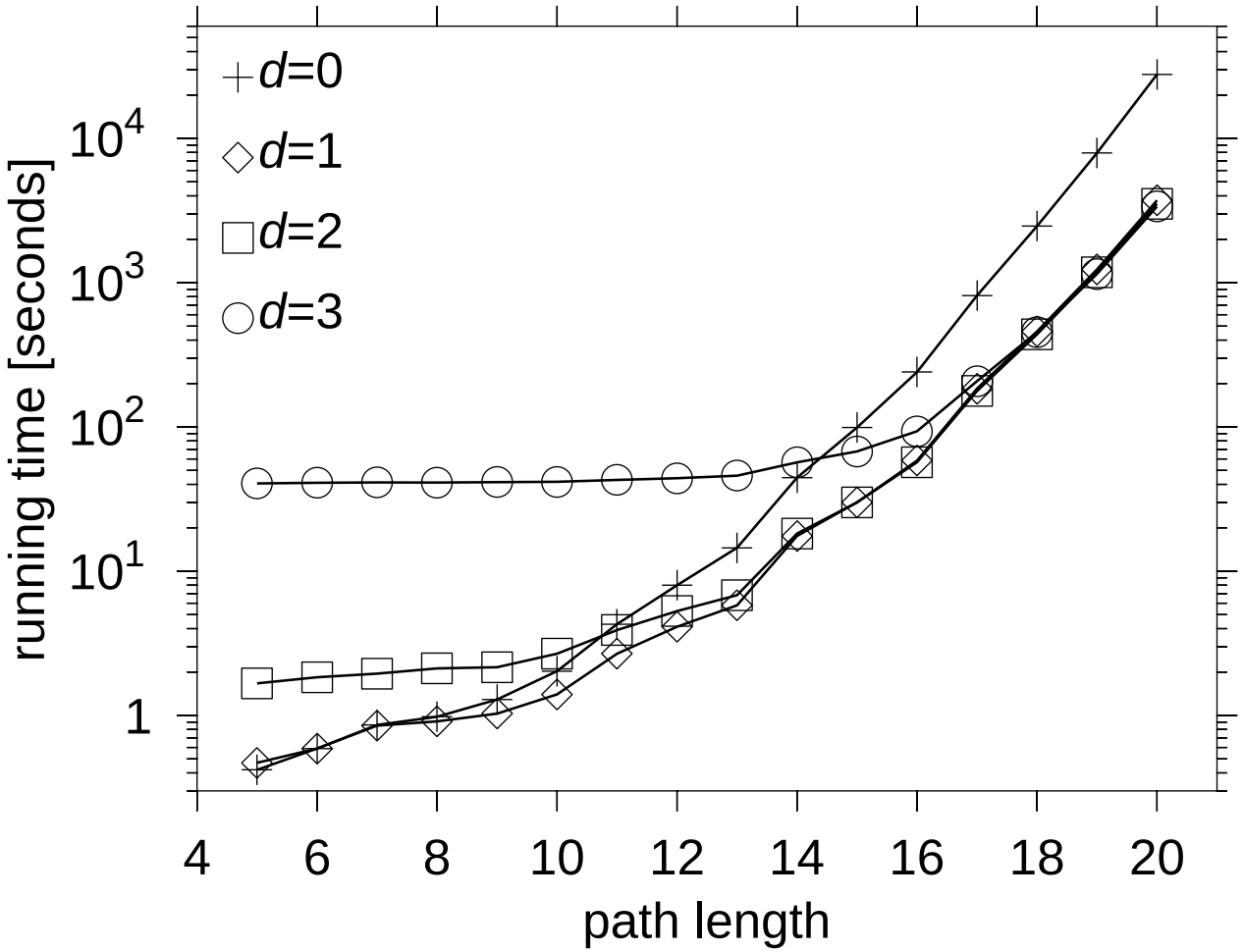
is higher than the weight of the known solution.

Precalculated Lower Bounds

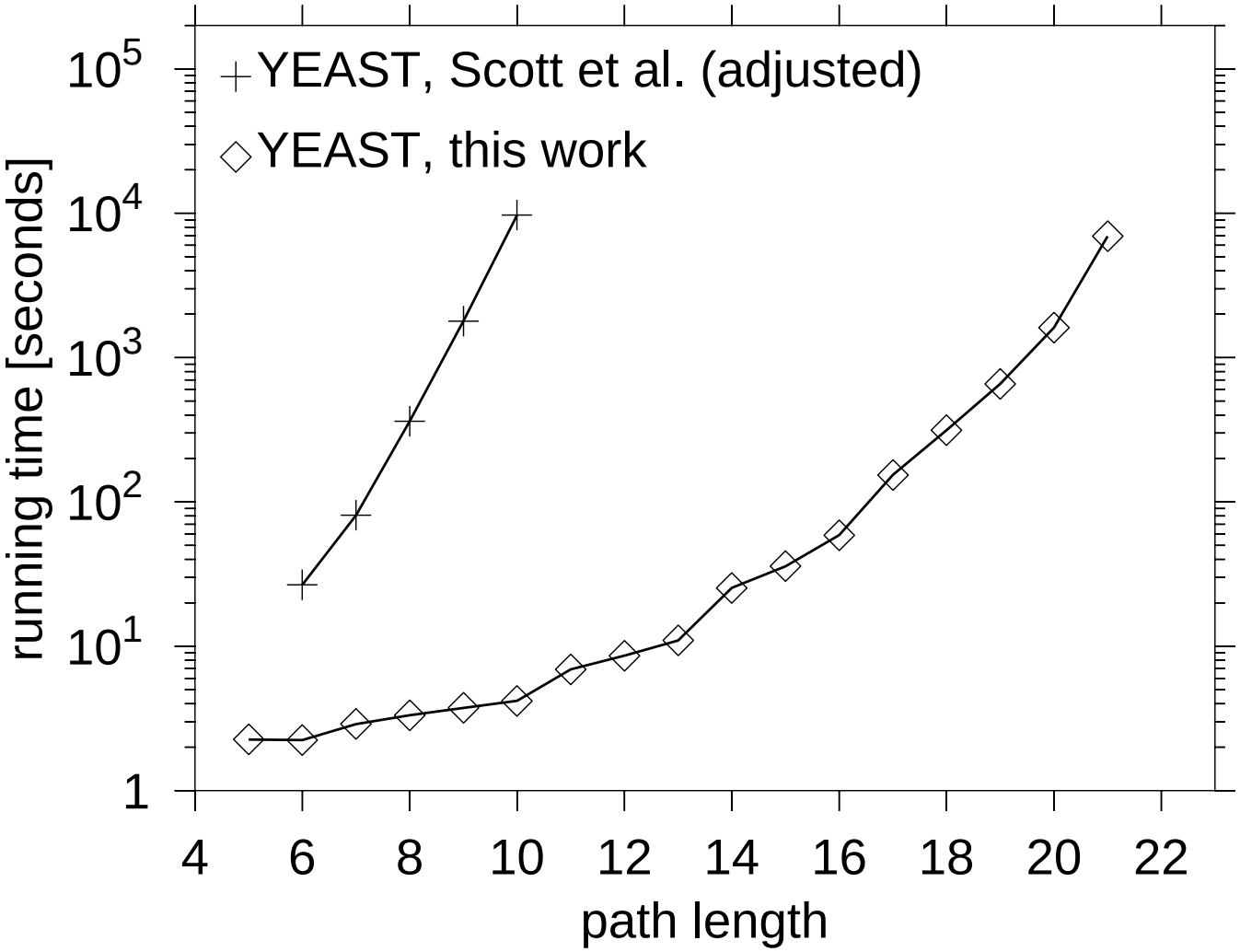
For each vertex u and a range of lengths $1 \leq i \leq d$, determine the minimum weight of a path of i edges that starts at u.



Lower Bounds Experiments



Yeast Network

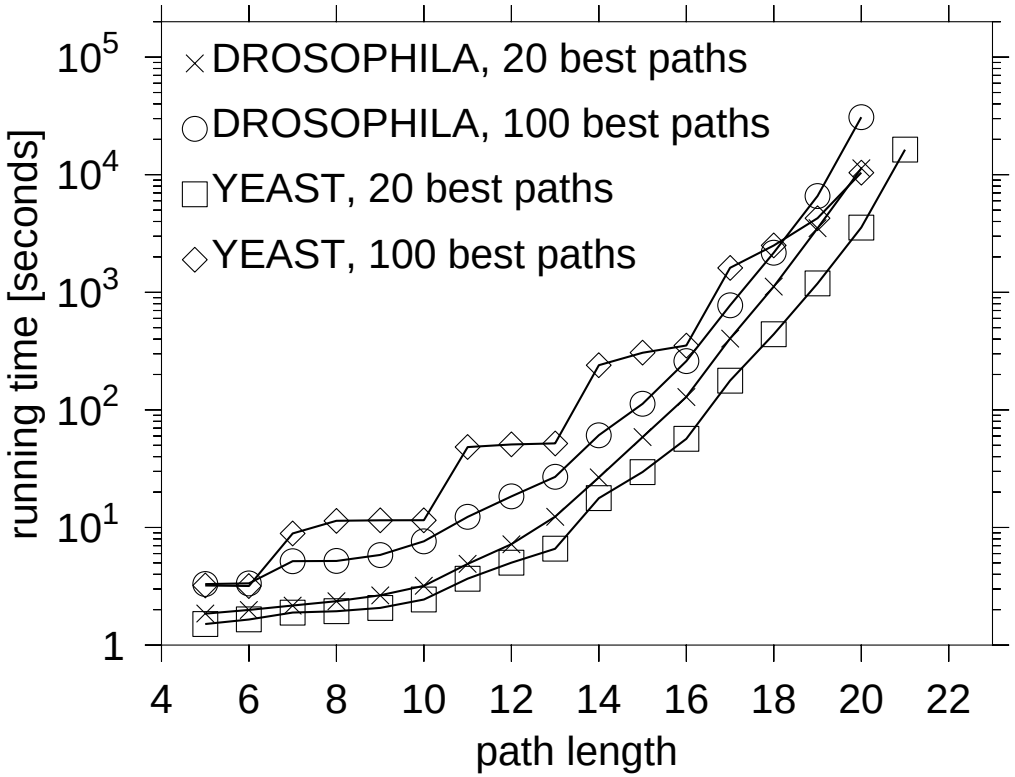


Network Comparison

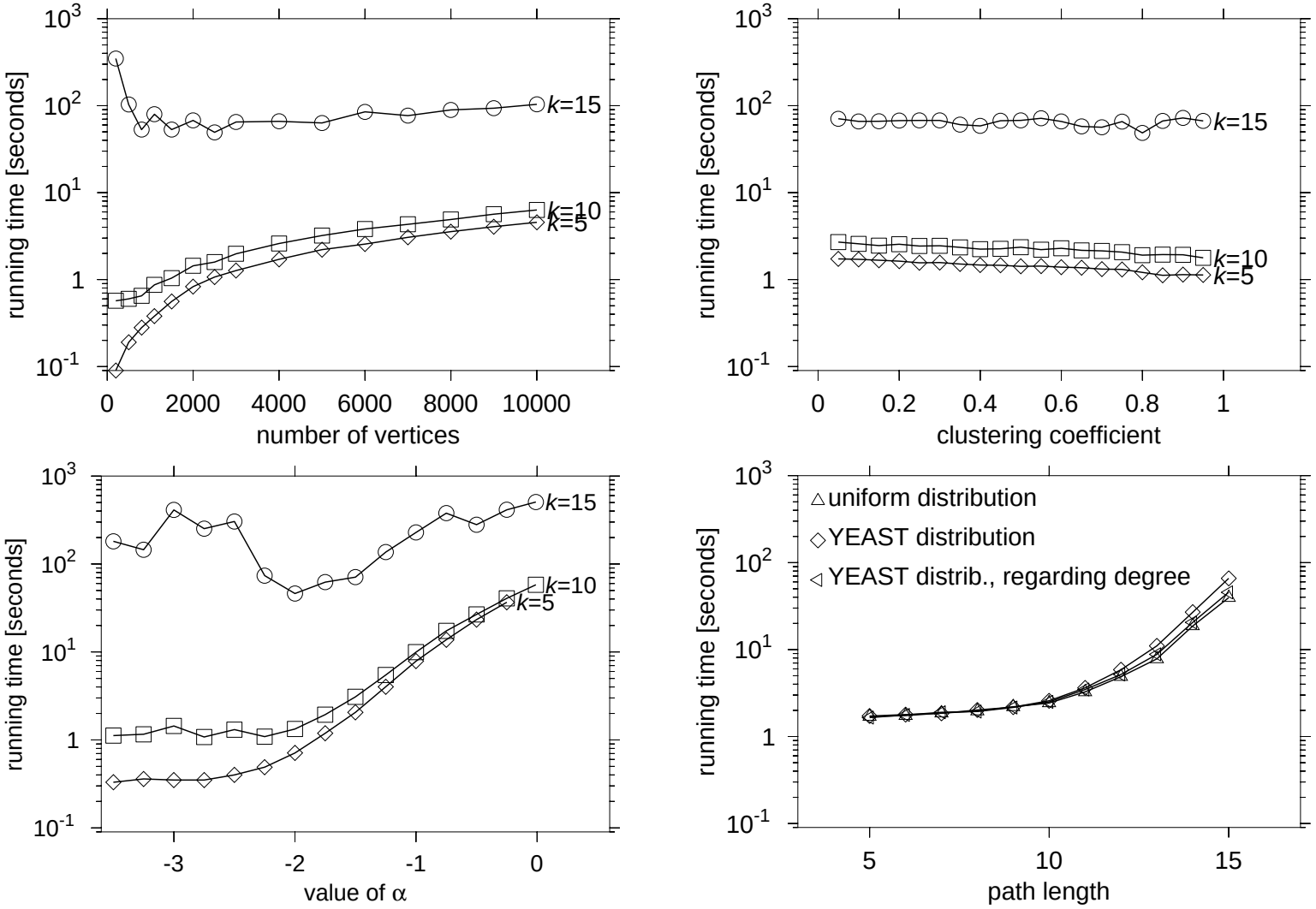
	$\|V\|$	$\|E\|$	clust. coeff.	avg. degree	max. degree
	4 389	14 319	0.067	6.5	237
	7 009	20 440	0.030	5.8	175

Network Comparison

	$\|V\|$	$\|E\|$	clust. coeff.	avg. degree	max. degree
	4 389	14 319	0.067	6.5	237
	7 009	20 440	0.030	5.8	175



Simulations: Robustness of Algorithm



Conclusion & Outlook

Color-coding, with some algorithm engineering, is a practical and reliable method for finding signaling pathways in protein interaction networks.

Conclusion & Outlook

Color-coding, with some algorithm engineering, is a practical and reliable method for finding signaling pathways in protein interaction networks.

Future work:

- Pathway queries
- Richer motifs (cycles, trees, . . .)
- Derandomization

Today's Biz

1. Reminders
2. Review
3. Biological Network Analysis Topics
4. **Hybrid processing - direction optimizing push/pull**
5. Assignment 2 solutions

Direction-optimizing BFS

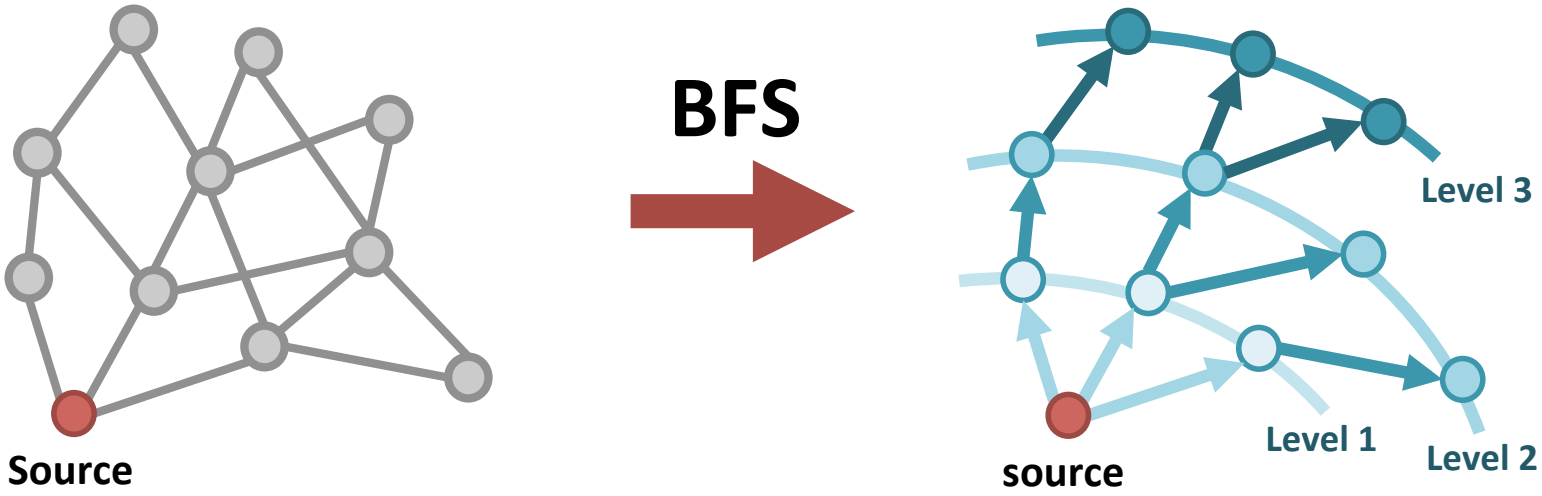
Slides from Yasui et al., Chuo University & JST CREST and Intel

Outline

1. Background
2. **Breadth-first Search (BFS)**
3. NUMA architecture
4. Proposal : NUMA-optimized parallel BFS
5. Numerical Results

Breadth-first Search (BFS)

- Obtains level of each vertices from source vertex
- **Level** = certain # of hops away from the source

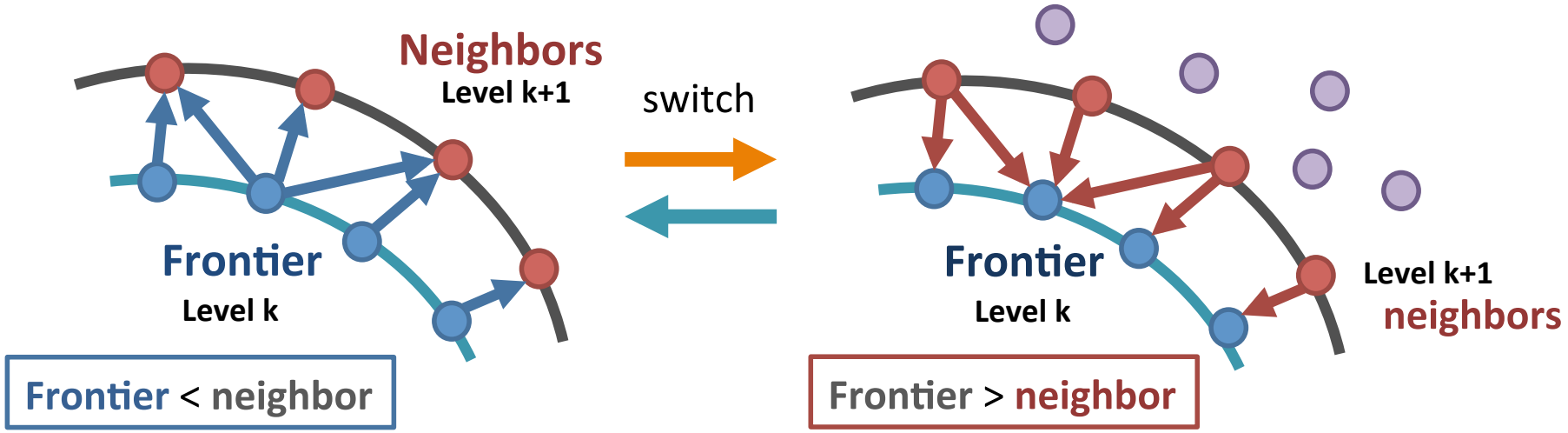


Input:
Graph **G** and **source**

Output:
Tree with root as source

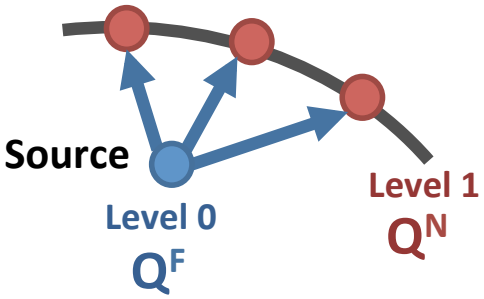
Hybrid BFS for low-diameter graph

- **Efficient** for **Low-diameter** graph **[Beamer2011, 2012]**
  - **scale-free** and/or **small-world** property such as social network.
- At higher ranks in Graph500 benchmark
- Hybrid algorithm
  - combines top-down algorithm and bottom-up algorithm
  - reduces unnecessary edge traversal



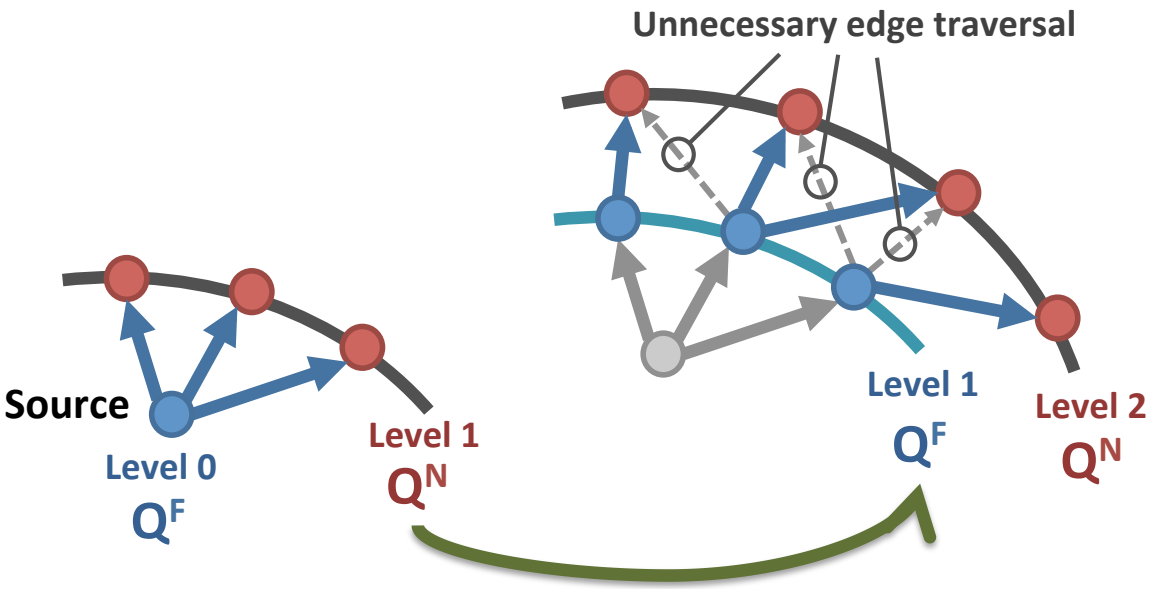
Top-down algorithm

- Explores outgoing edges of **frontier queue Q^F**
- Appends unvisited vertices into **neighbor queue Q^N**



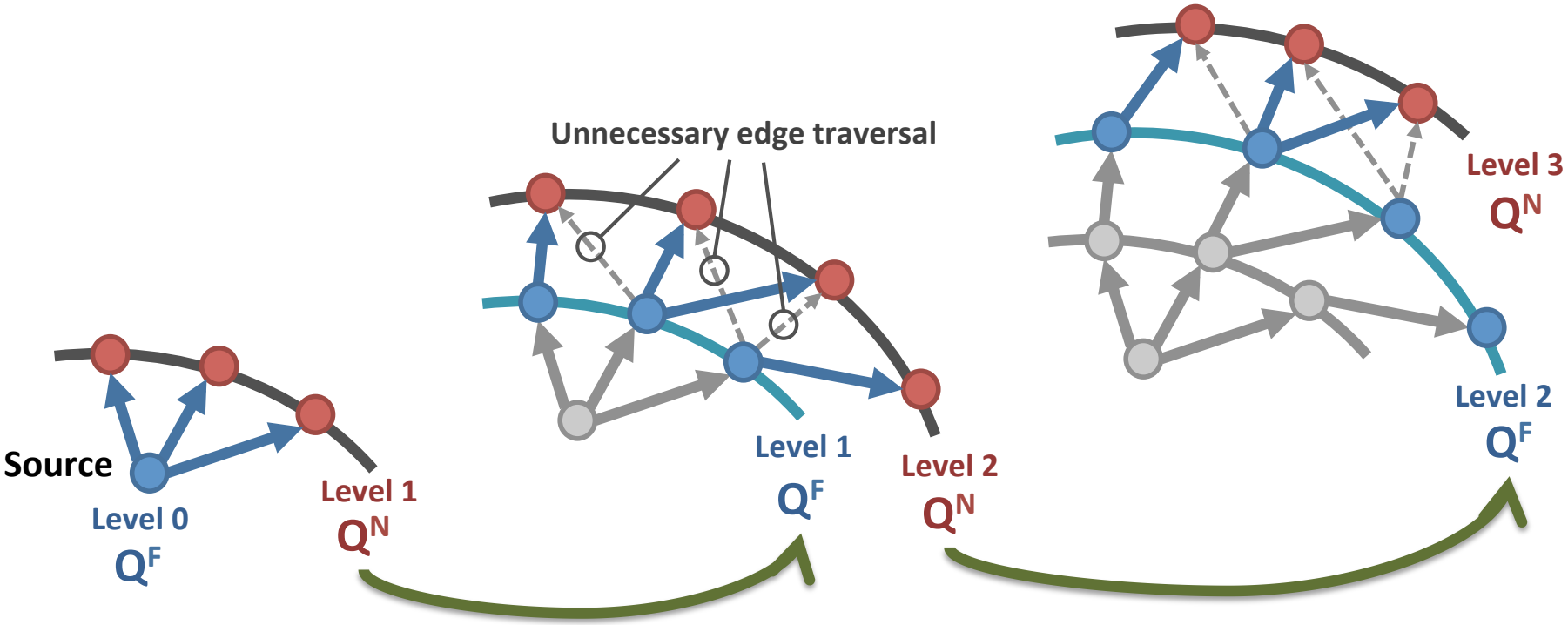
Top-down algorithm

- Explores outgoing edges of **frontier queue Q^F**
- Appends unvisited vertices into **neighbor queue Q^N**



Top-down algorithm

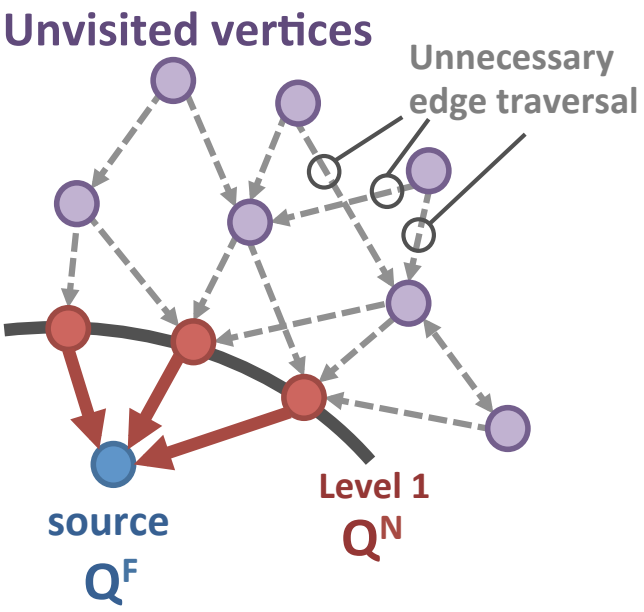
- Explores outgoing edges of **frontier queue Q^F**
- Appends unvisited vertices into **neighbor queue Q^N**



- **Efficient for a small frontier**
- Has an unnecessary edge traversal for a large frontier

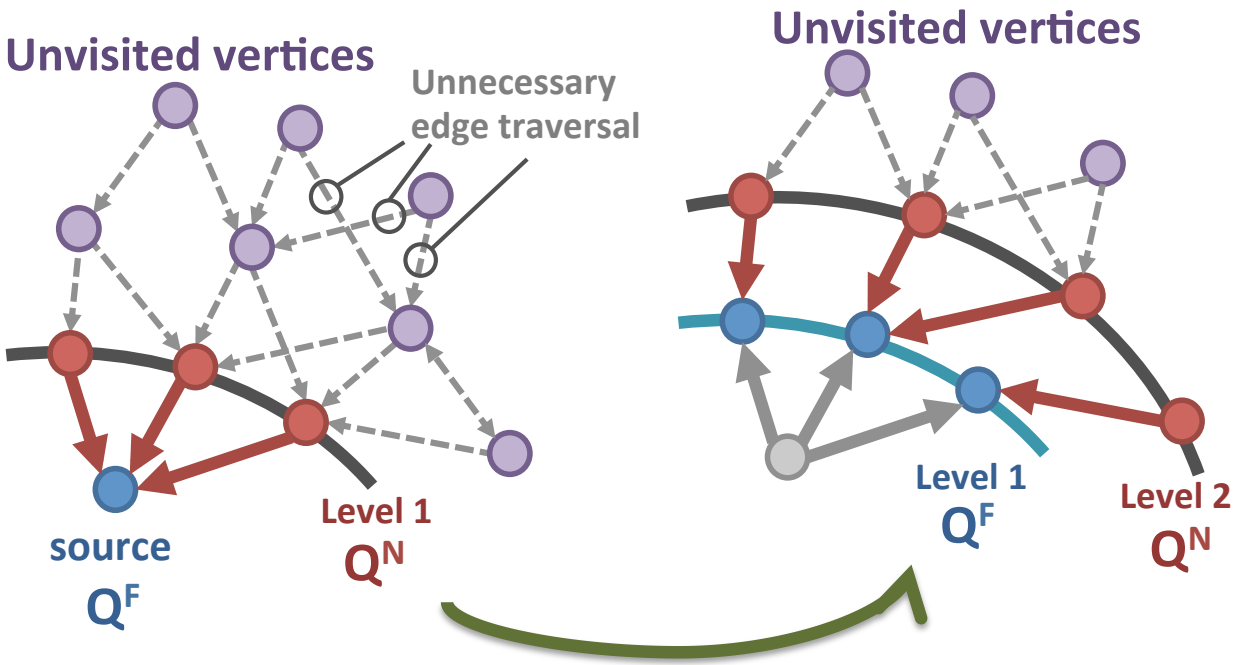
Bottom-up algorithm

- Explores **frontier queue Q^F** from **unvisited vertices**
- Appends adjacent vertices into **neighbors Q^N**



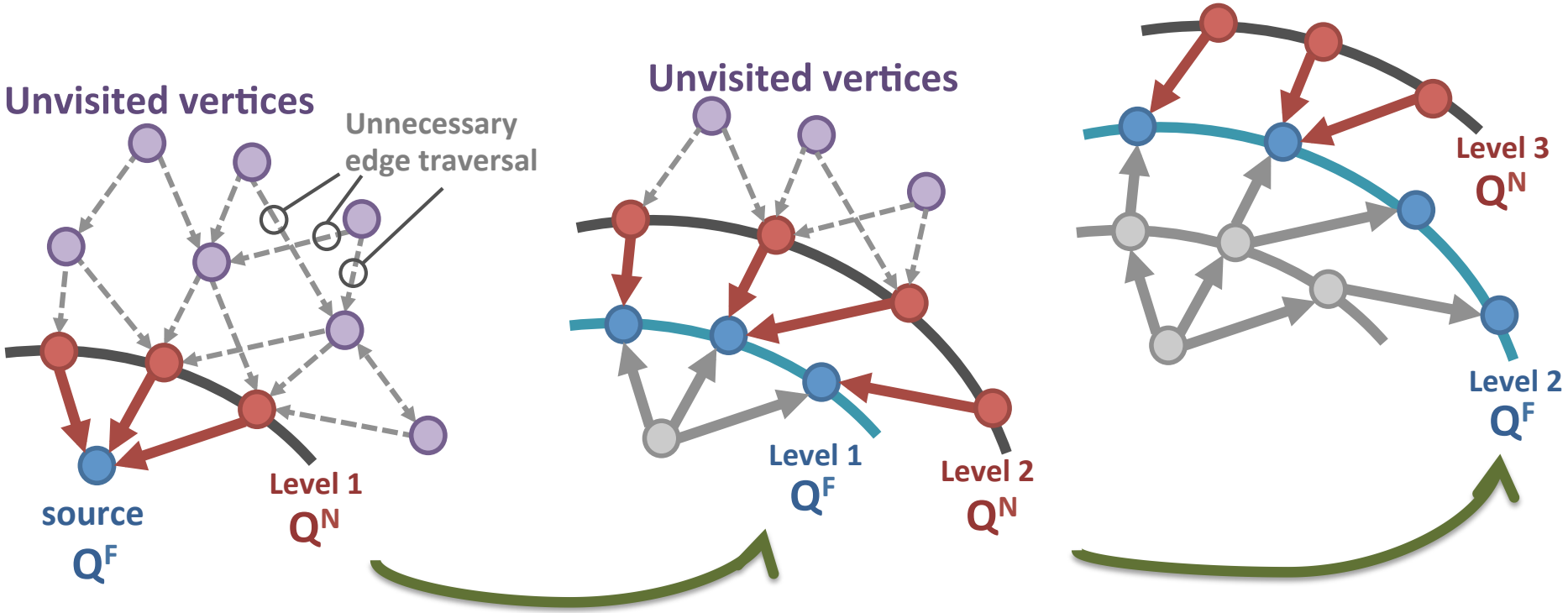
Bottom-up algorithm

- Explores **frontier queue Q^F** from **unvisited vertices**
- Appends adjacent vertices into **neighbors Q^N**



Bottom-up algorithm

- Explores **frontier queue Q^F** from **unvisited vertices**
- Appends adjacent vertices into **neighbors Q^N**



- **Efficient for a large frontier**
- Has unnecessary edge traversal for a small frontier

Hybrid BFS combines Top-down and Bottom-up

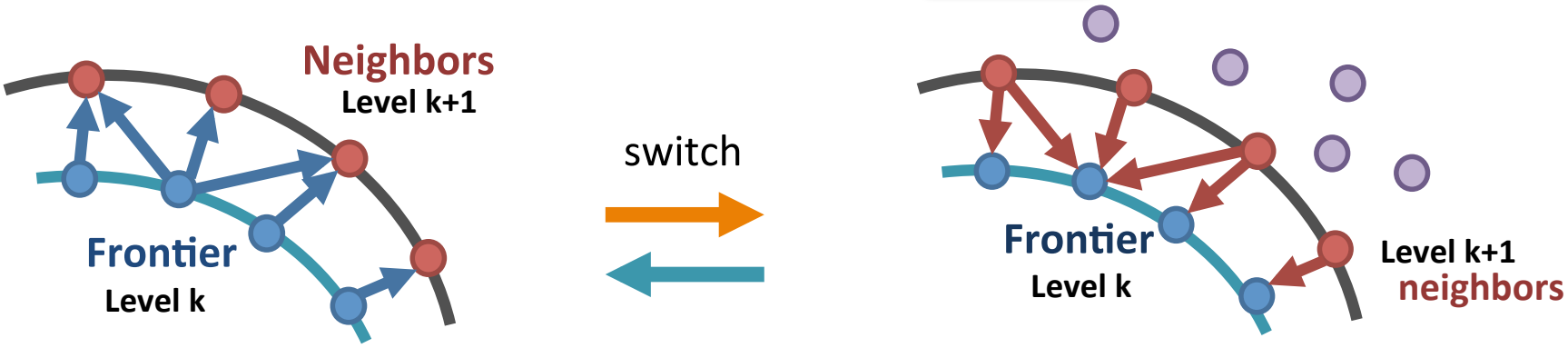
Level	Top-down $m_{\mathcal{F}}$	Bottom-up $m_{\mathcal{B}}$	Hybrid $min(m_{\mathcal{F}}, m_{\mathcal{B}})$
0	2	2,103,840,895	2
1	66,206	1,766,587,029	66,206
2	346,918,235	52,677,691	52,677,691
3	1,727,195,615	12,820,854	12,820,854
4	29,557,400	103,184	103,184
5	82,357	21,467	21,467
6	221	21,240	227
Total	2,103,820,036	3,936,072,360	65,689,631
Ratio	100.00%	187.09%	3.12%

switch

switch

Traversal edges of Kronecker graph (SCALE 26)

only



Top-down algorithm

Bottom-up algorithm

Today's Biz

1. Reminders
2. Review
3. Biological Network Analysis Topics
4. Hybrid processing - direction optimizing push/pull
5. **Assignment 2 solutions**