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

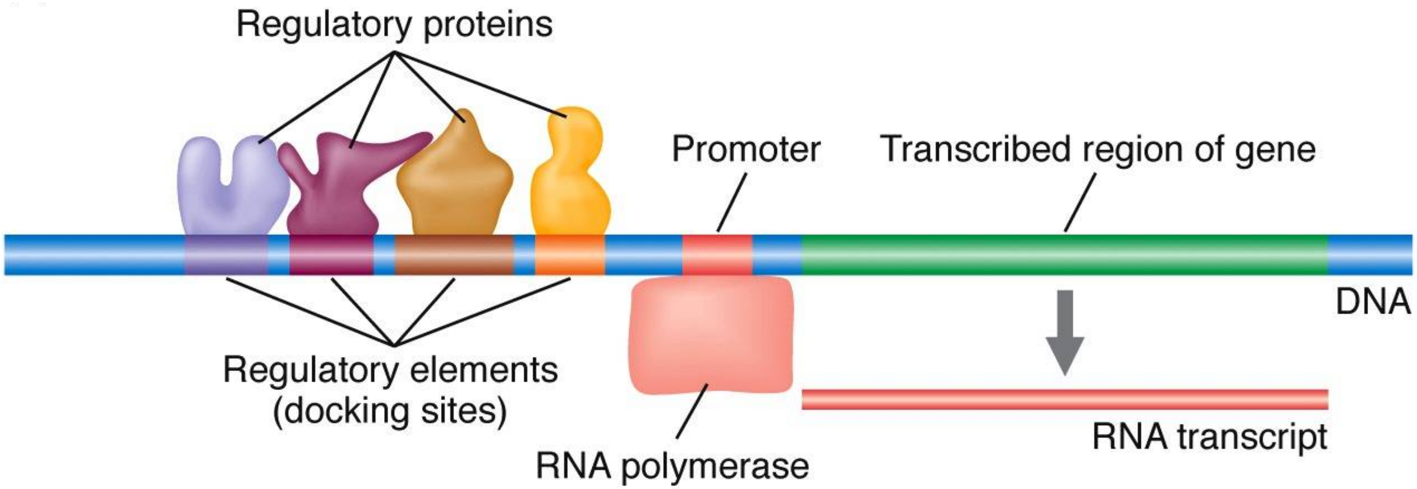
New large-scale experimental data in the form of networks:

- 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:

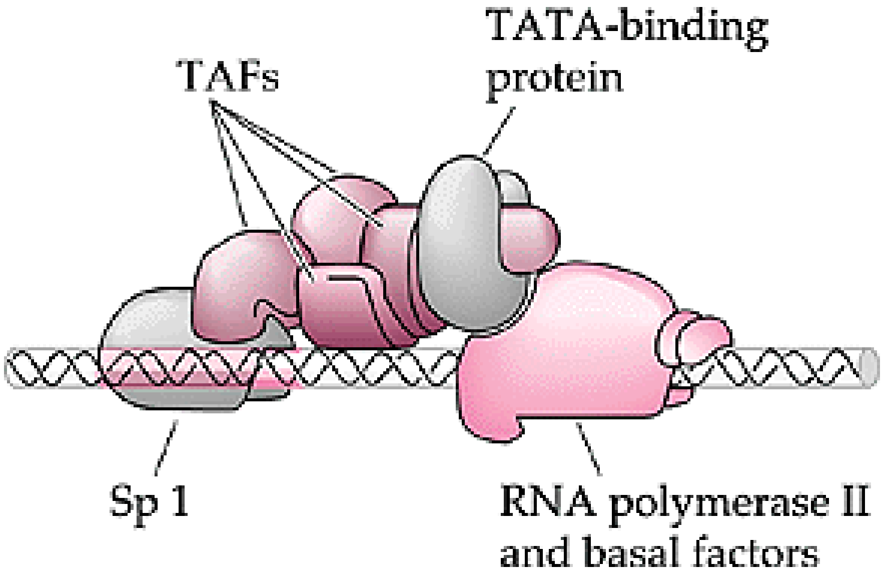
- transcription networks
  - transcription factors bind to regulatory DNA
  - polymerase molecule begins transcription of the gene



Networks in molecular biology

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

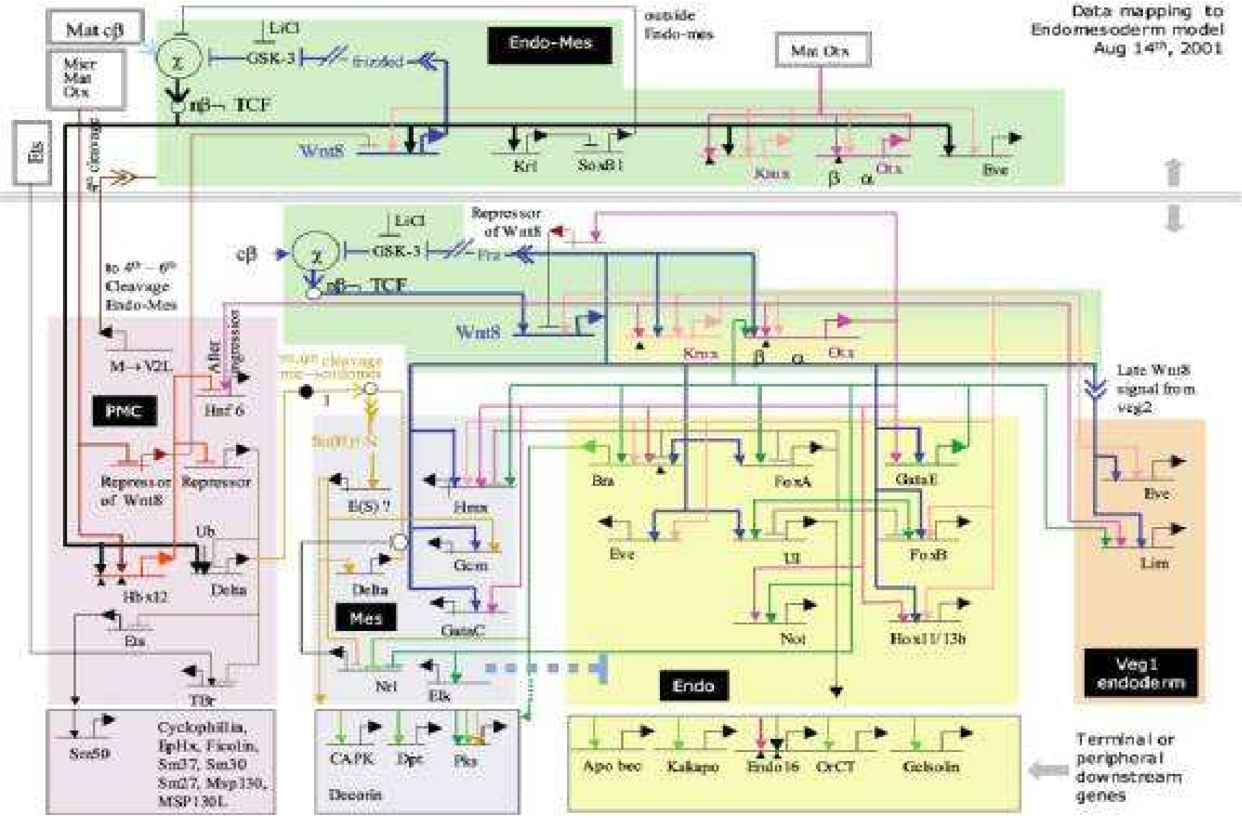
- transcription networks
  - transcription factors bind to regulatory DNA
  - polymerase molecule begins transcription of the gene



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



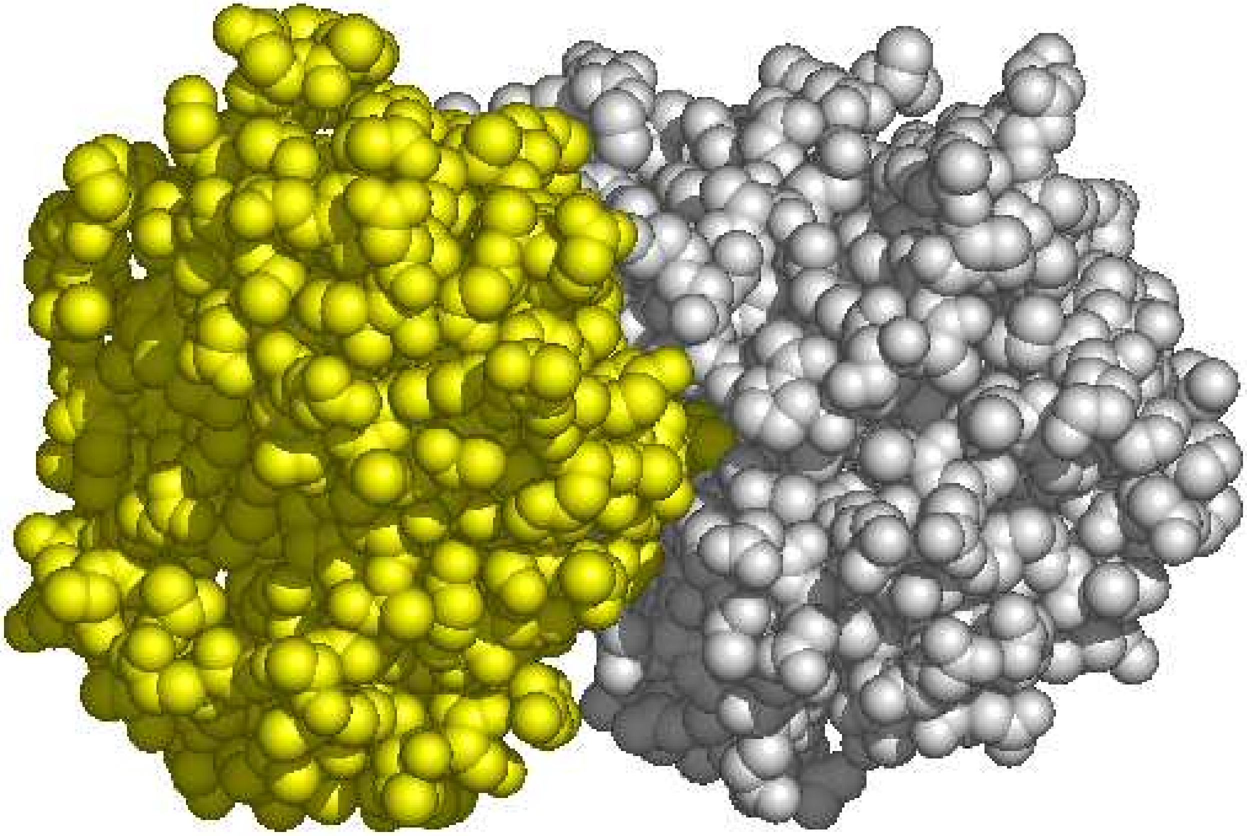
sea urchin

Bolouri &Davidson (2001)

Networks in molecular biology

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

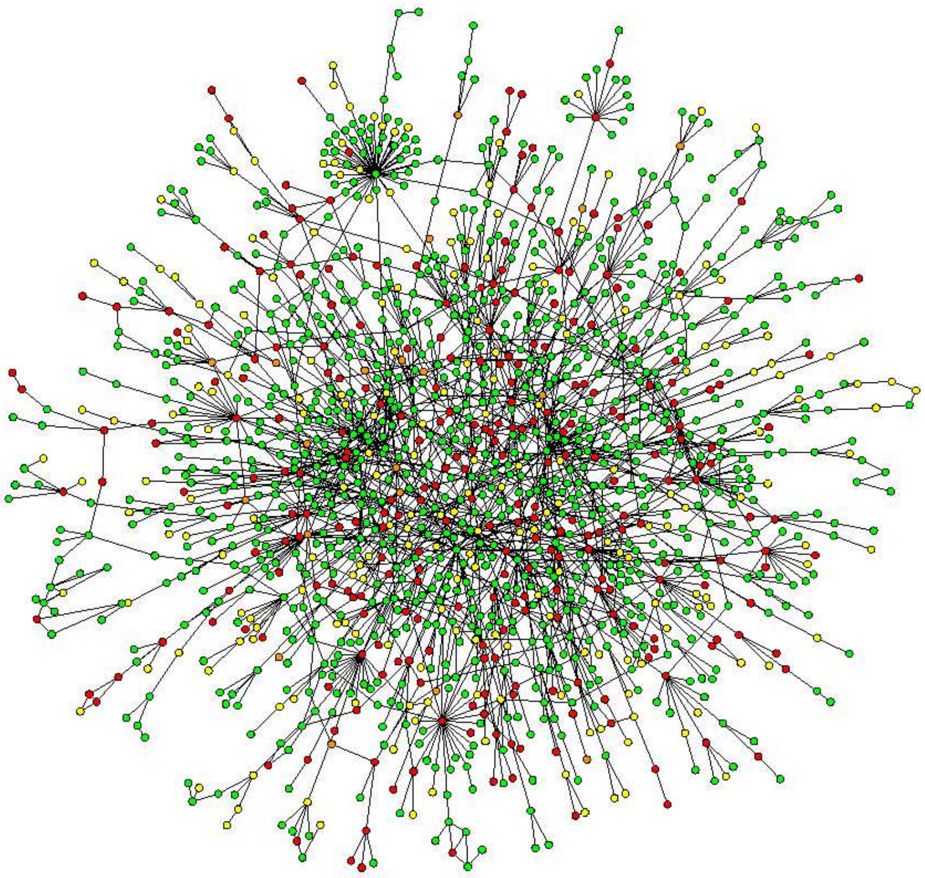
- protein interaction networks
    - proteins interact to form larger units
    - protein aggregates may catalyze reactions *etc.*



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

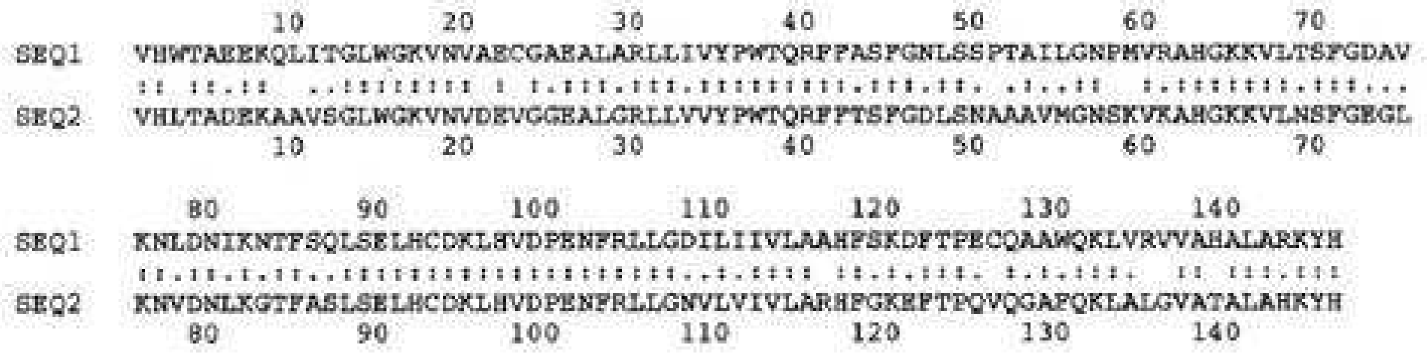
- more than 100 organisms are fully sequenced
- genome sizes range from 3×10^7 to 7×10^{11} basepairs

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



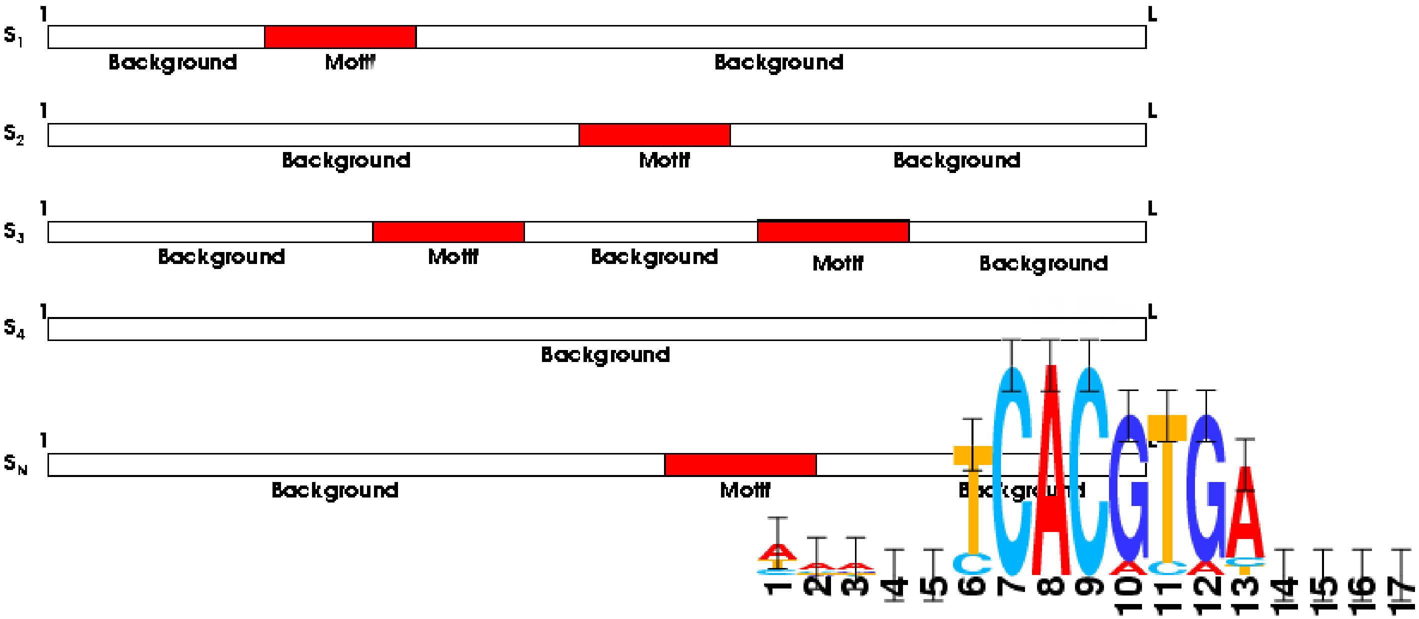
```
               10        20        30        40        50        60        70
SEQ1   VHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAV
       :: ::.::  ..:::::::: : :.:::.:::.:::::::::.:::.::. .:..::  :.:::::::.:::...
SEQ2   VHLTADEKAAVSGLWGKVNVDEVGGEALGRLLVVYPWTQRFFTSFGDLSNAAAVMGNSKVKAHGKKVLNSFGEGL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

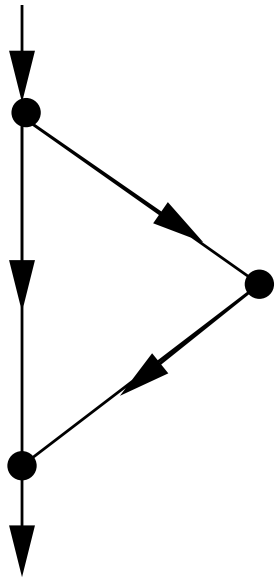
1. Search for network motifs [Alon lab]
   - patterns occurring repeatedly within a given network
2. Alignment of networks across species
   - identify conserved regions
   - pinpoint functional innovations

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

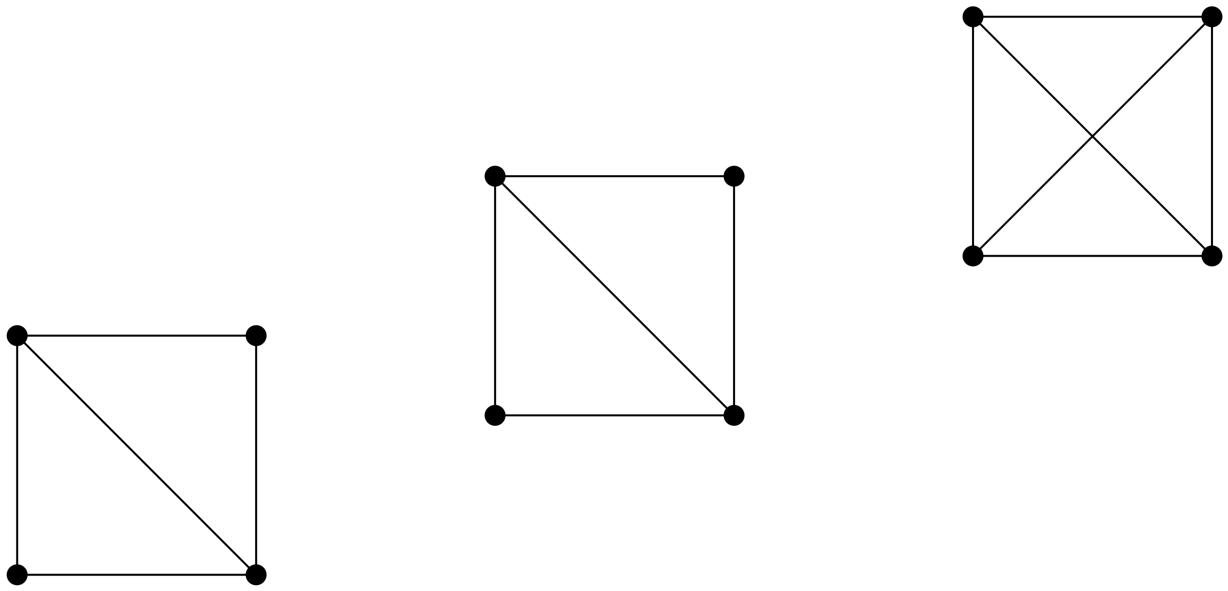
- patterns occurring repeatedly in the network
- building blocks of information processing [Alon lab]

- 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]

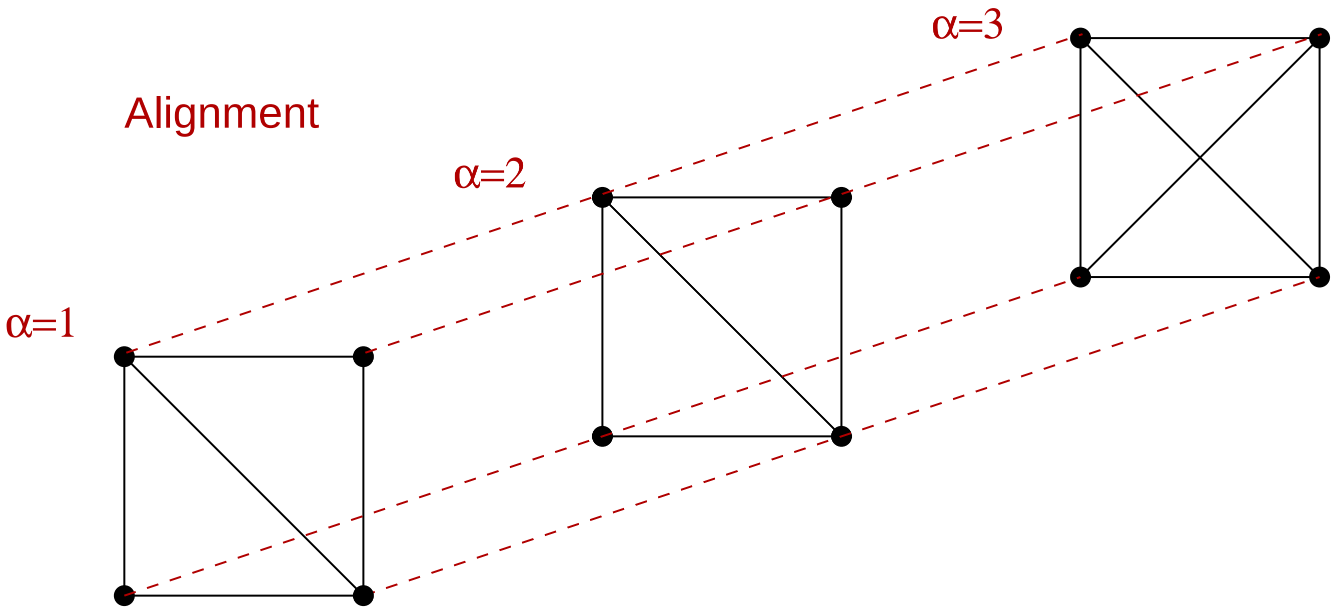
- counting of identical patterns: Subgraph census
- alignment of topologically similar regions of a network
- allow for mismatches
- construct a scoring function comparing the aligned subgraphs to a background model



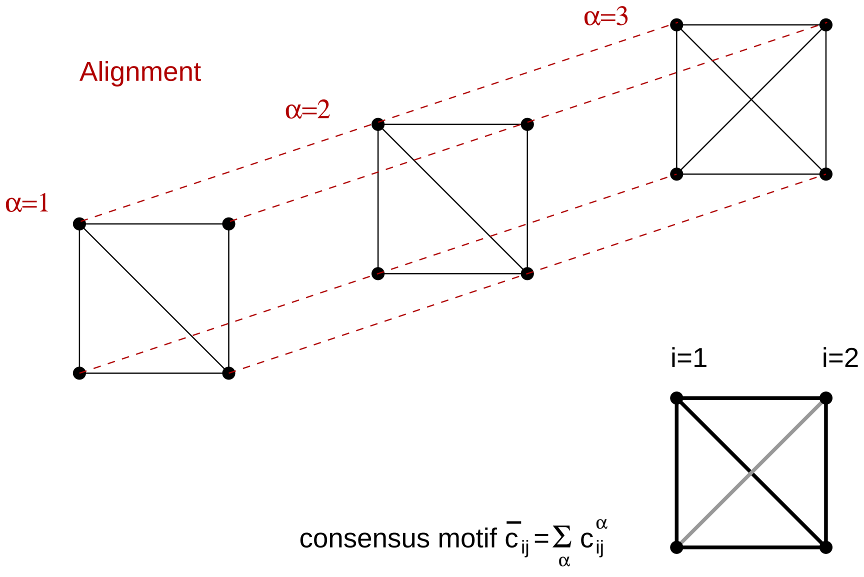
Graph alignment I: The search for network motifs

- 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

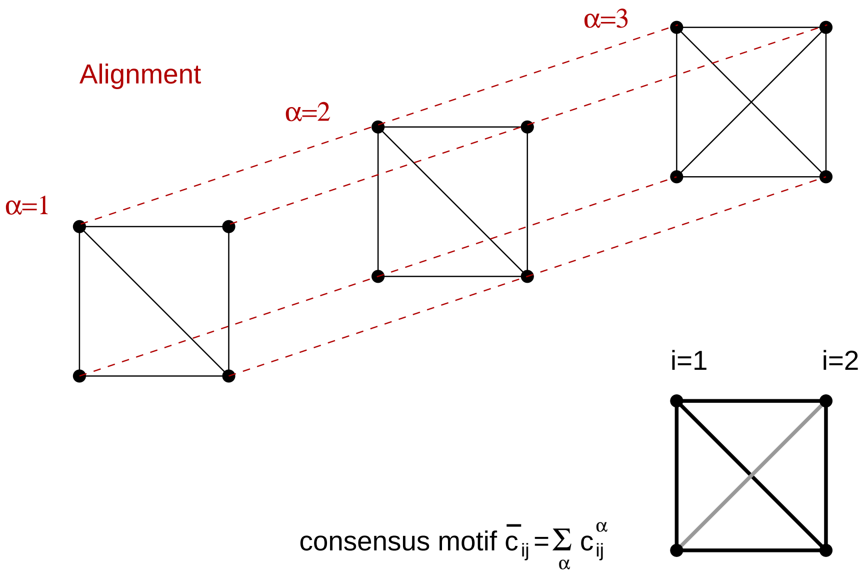


Statistical properties of alignments



consensus motif $\bar{c}_{ij} = \sum_{\alpha} c_{ij}^{\alpha}$

Statistical properties of alignments



Alignment

α=1

α=2

α=3

i=1

i=2

consensus motif $\bar{c}_{ij} = \sum_{\alpha} c_{ij}^{\alpha}$

- 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

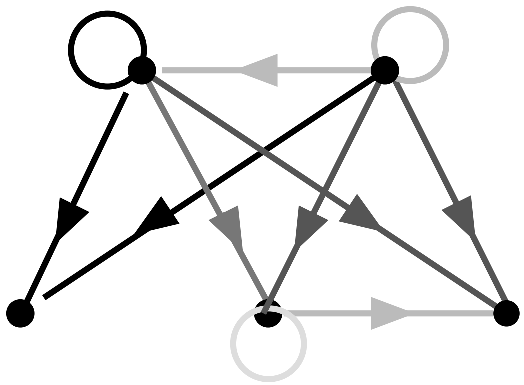
- ensemble with enhanced number of links
- enhanced correlation of subgraphs divergent vs convergent evolution?

Log likelihood score

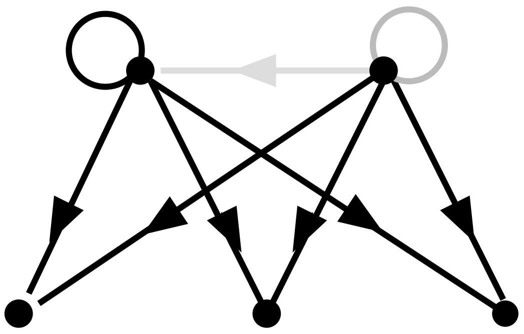
$$
\begin{aligned}
S(\mathbf{c}^1, \ldots, \mathbf{c}^p) &= \log \left(\frac{Q(\mathbf{c}^1, \ldots, \mathbf{c}^p)}{\prod_{\alpha=1}^{p} P_\sigma(\mathbf{c}^\alpha)} \right) \\
&= (\sigma - \sigma_0) \sum_{\alpha=1}^{p} L(\mathbf{c}^\alpha) - \frac{\mu}{2p} \sum_{\alpha,\beta=1}^{p} M(\mathbf{c}^\alpha,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

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

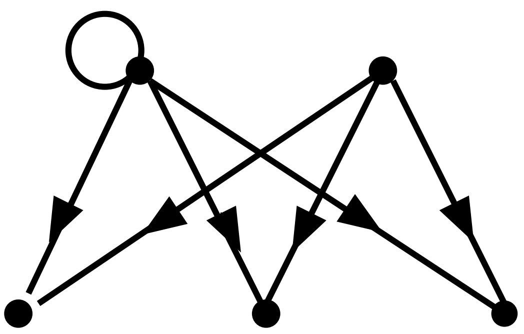
Consensus motif of the *E. coli* transcription network



$\mu = \mu^* = 2.25$

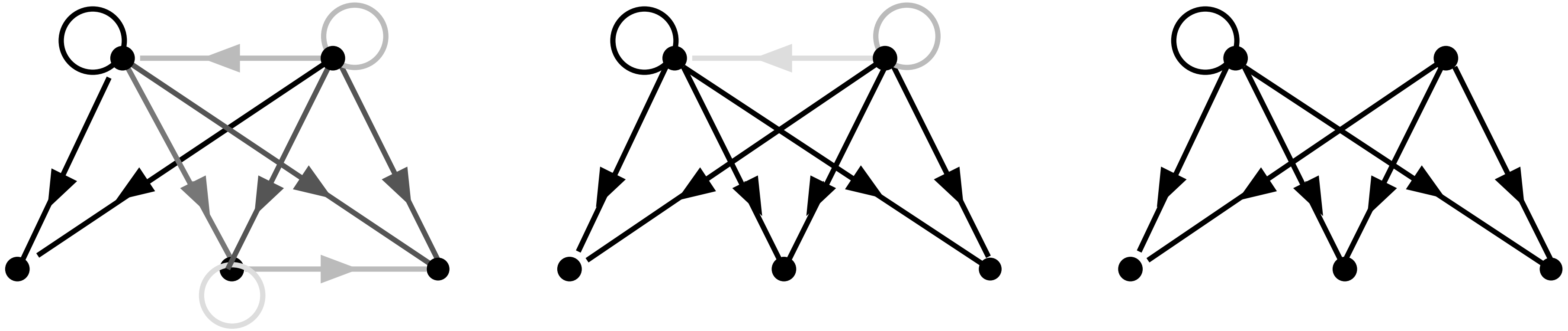


$\mu = 5$



$\mu = 12$

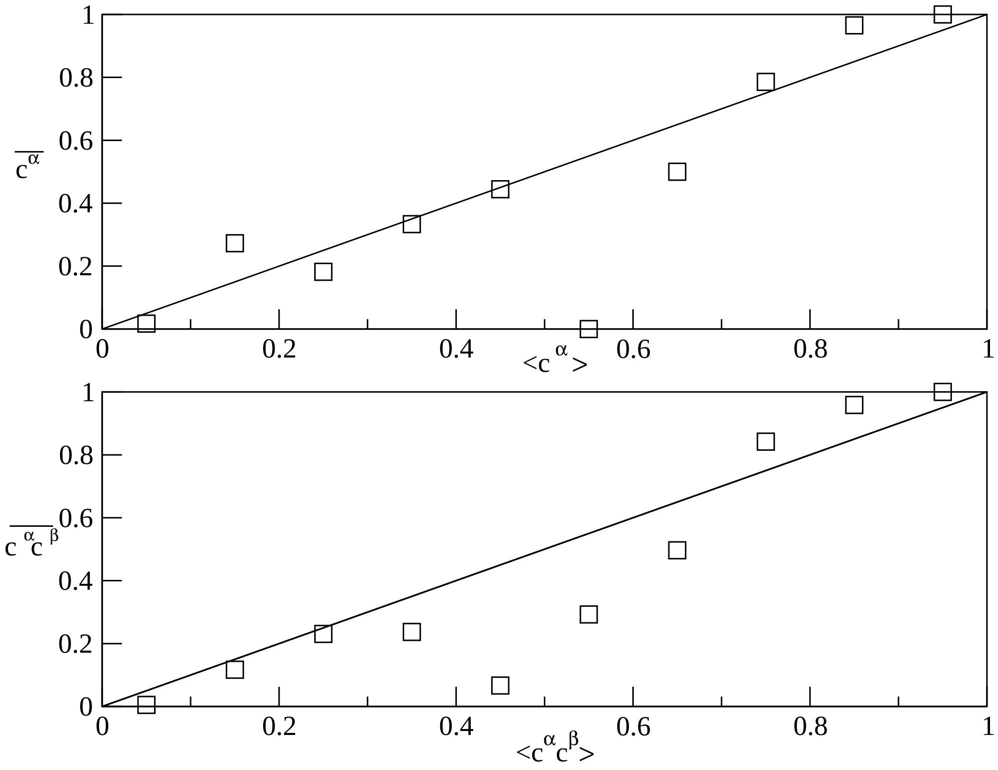
Consensus motif of the *E. coli* transcription network



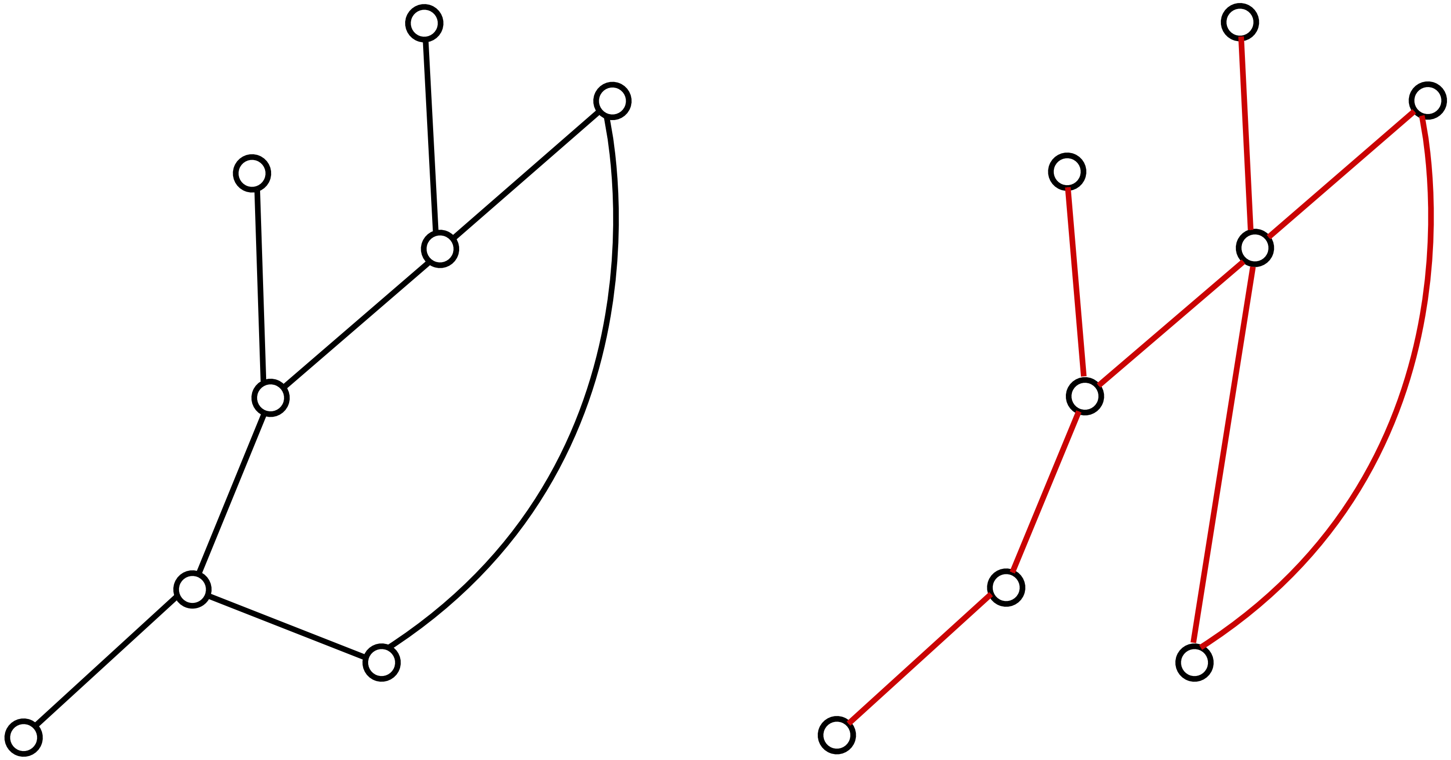
$\mu = \mu^* = 2.25$

$\mu = 5$

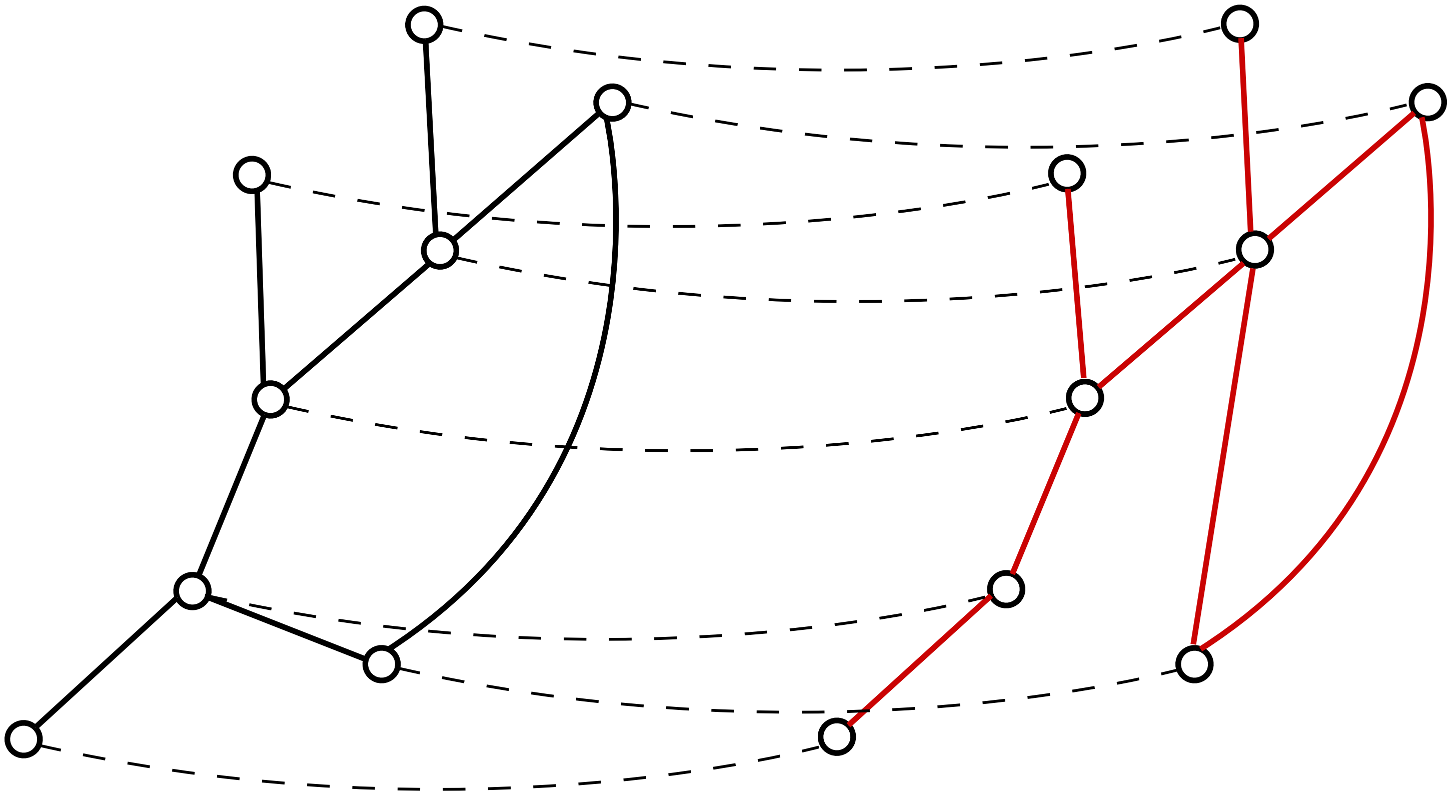
$\mu = 12$



Graph alignment II: Comparing networks across species

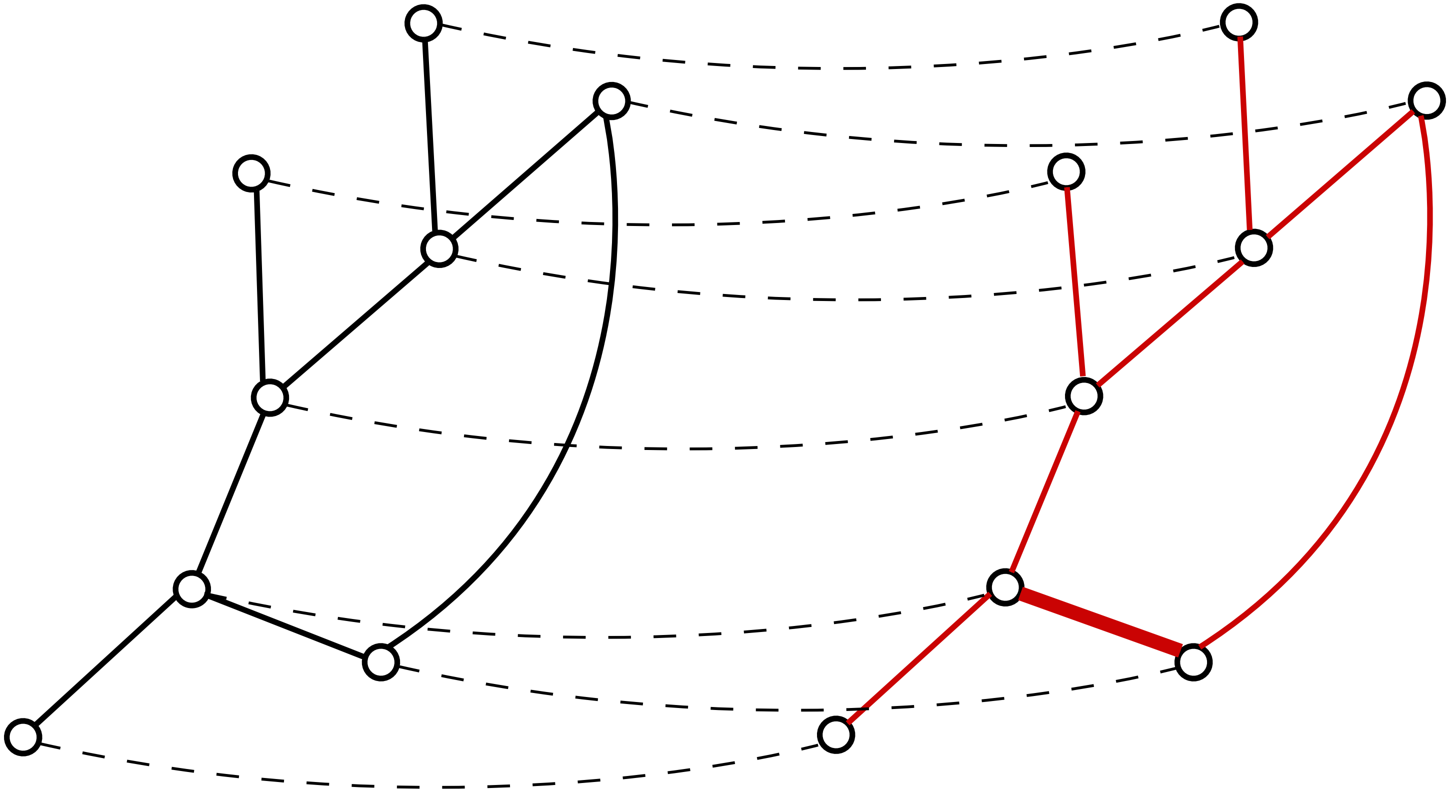


Graph alignment II: Comparing networks across species



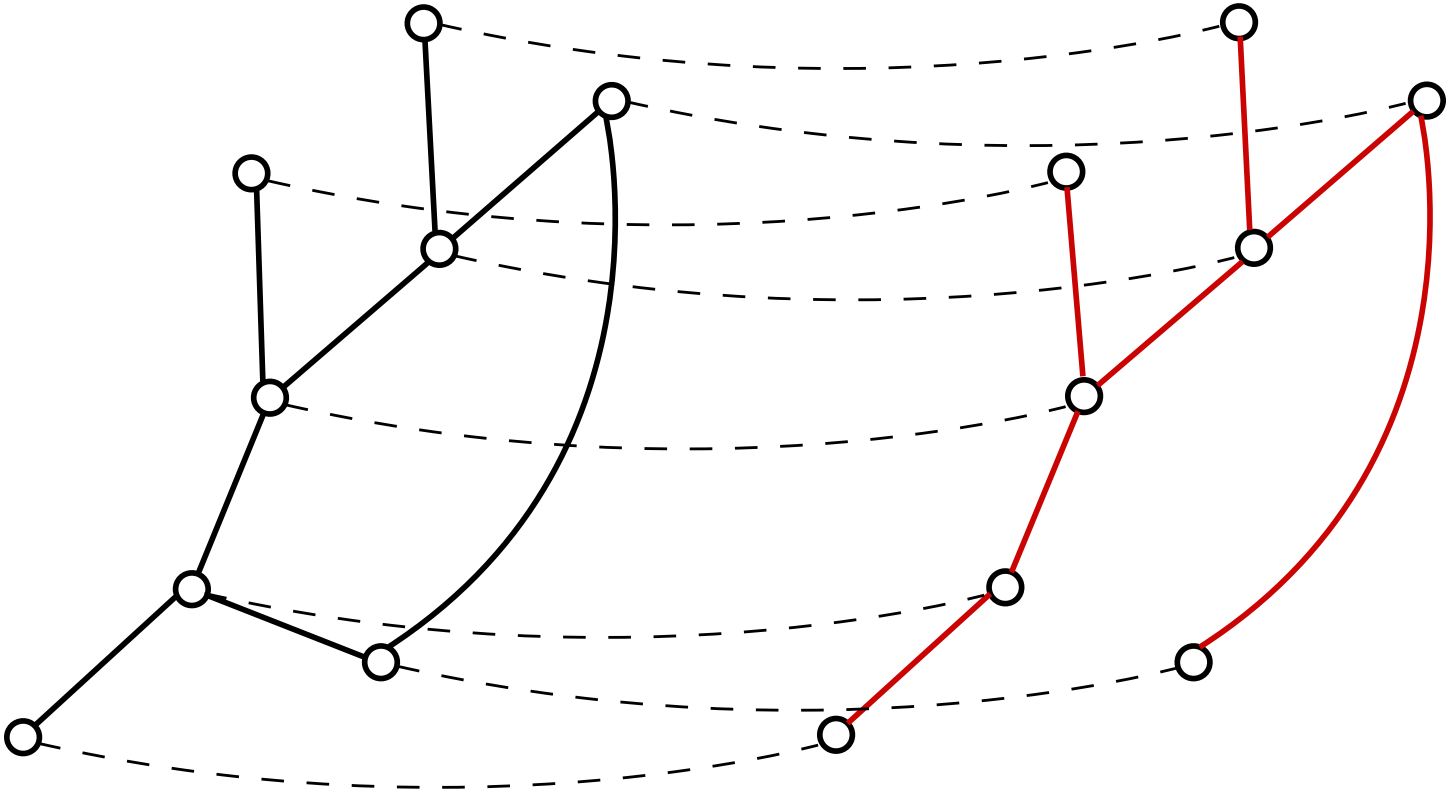
Alignment: Pairwise association of nodes across species

Graph alignment II: Comparing networks across species



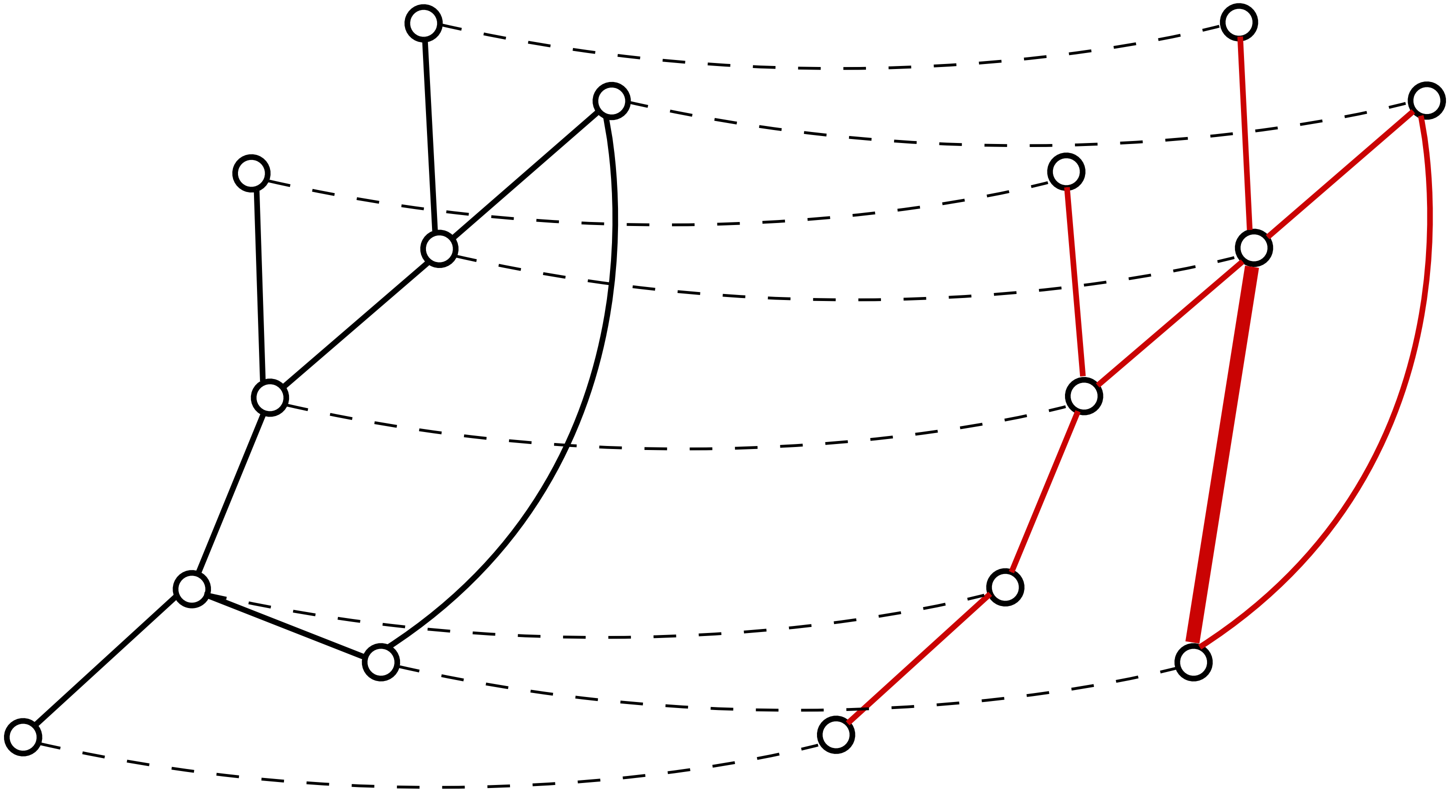
Last common ancestor

Graph alignment II: Comparing networks across species



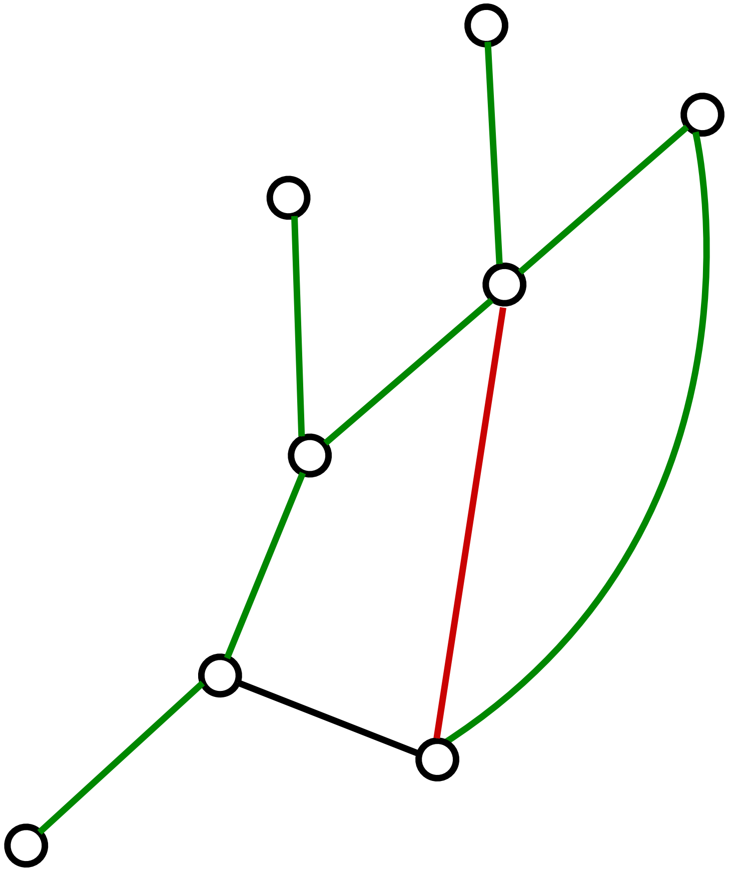
Evolutionary dynamics: Link attachment and deletion

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

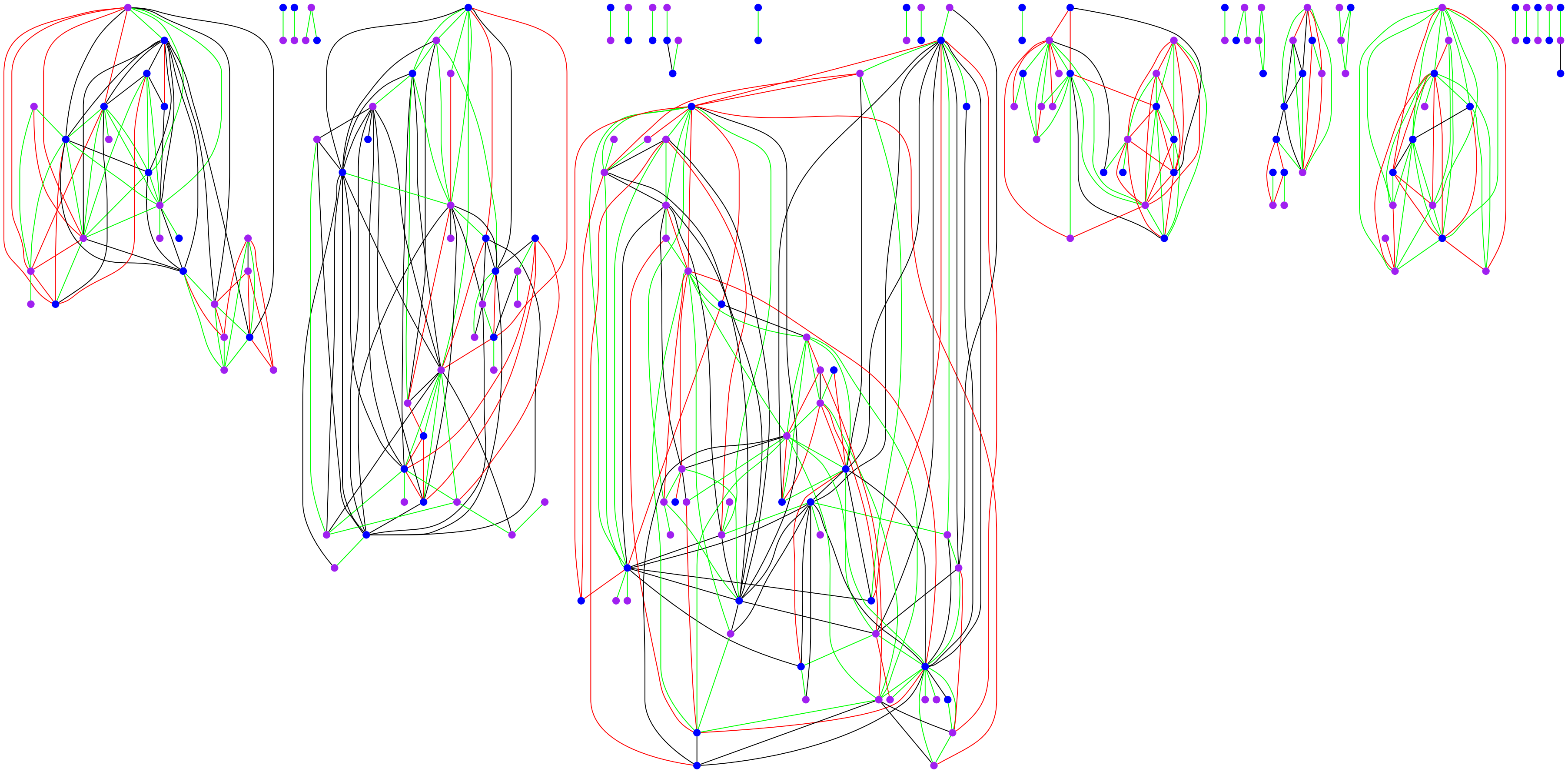
Q-model

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

Scoring alignments

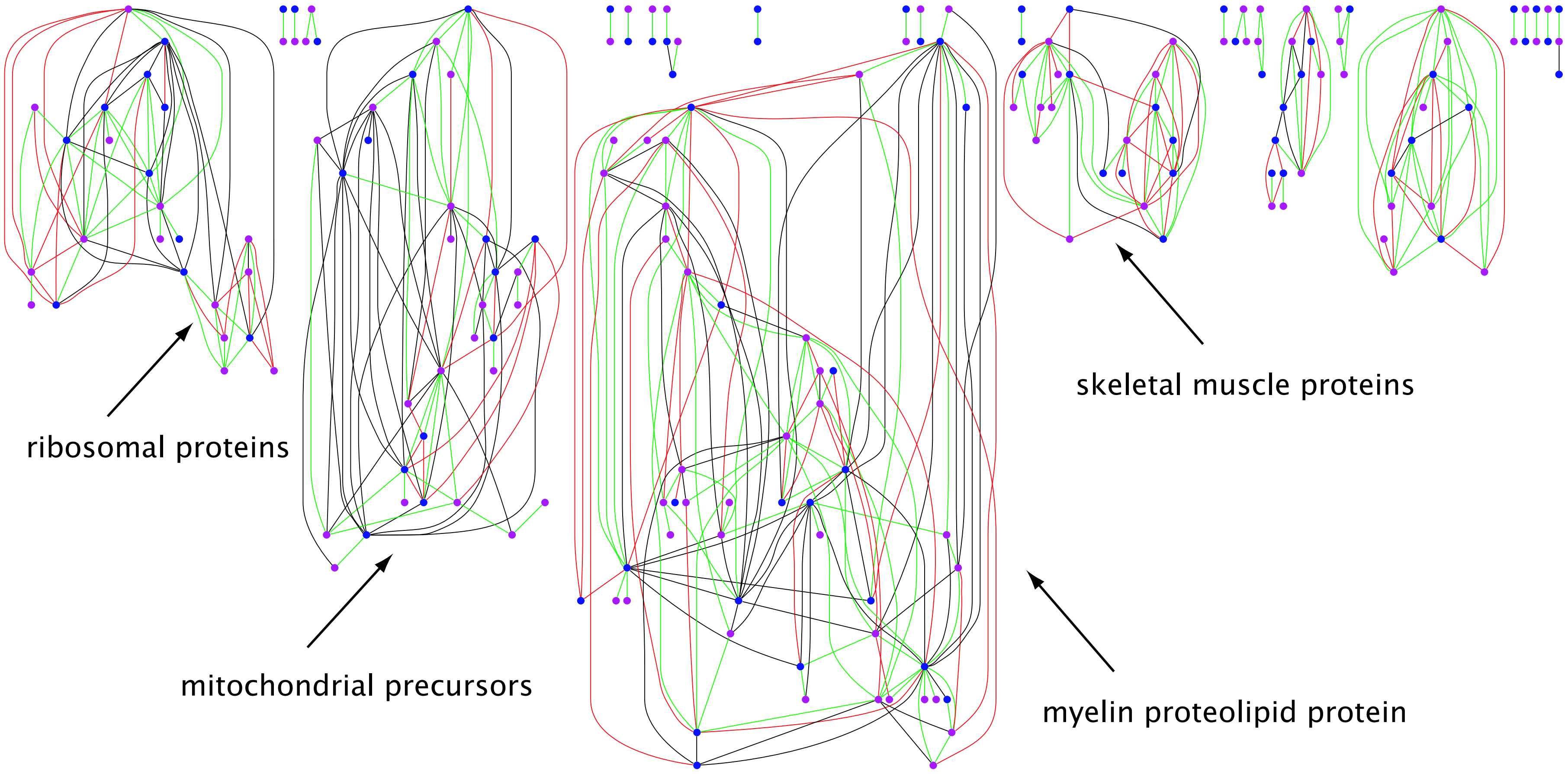
- log-likelihood score $S = \log(Q/P)$ is used to search for conserved parts of the networks

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)