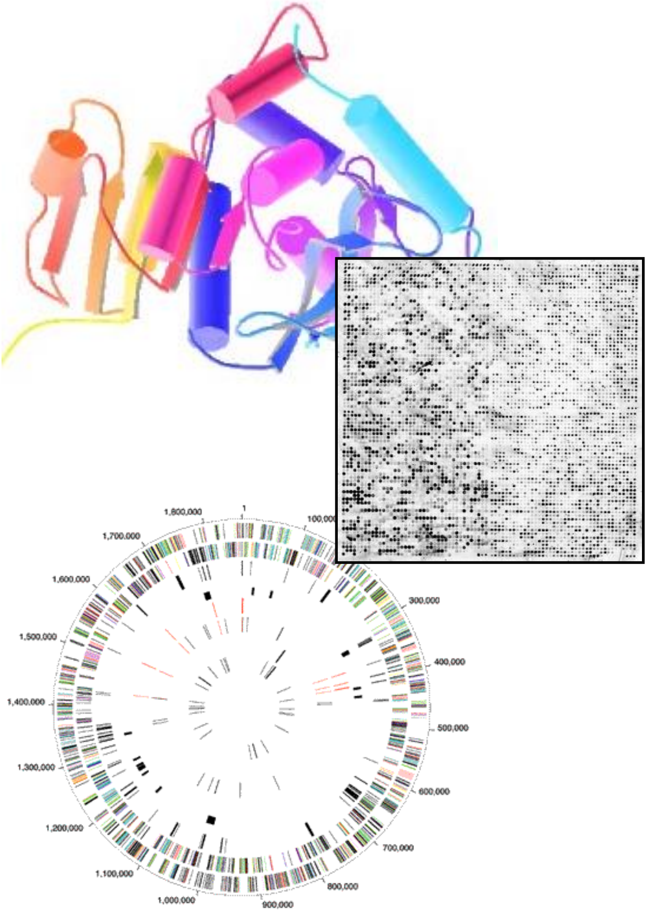
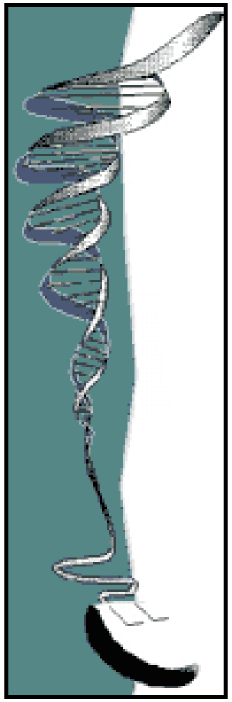
Biomed. Data Science:

Unsupervised Datamining -- Community Detection





Mark Gerstein, Yale University

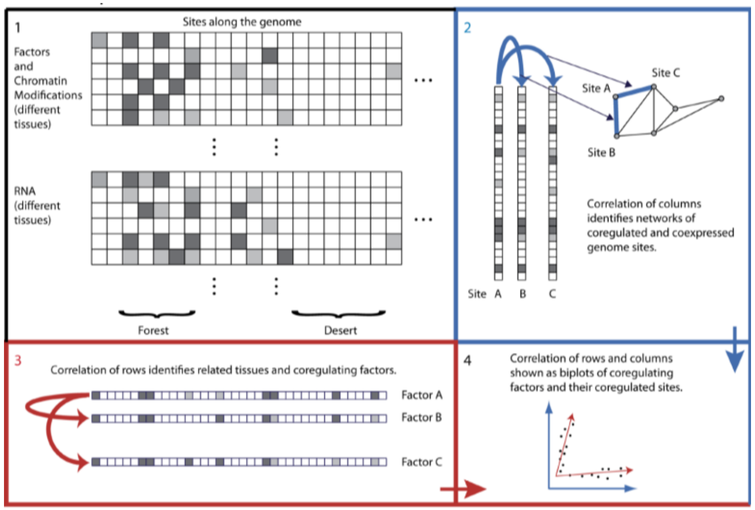
gersteinlab.org/courses/452

(last edit in spring ’22, pack 22m9b; similar to pack M9b from ‘21 with additional slide 2 & edits to the TADs section.)

Unsupervised Mining

Graph Analysis &
Community Detection Approaches

Graph Methods & Community Detection

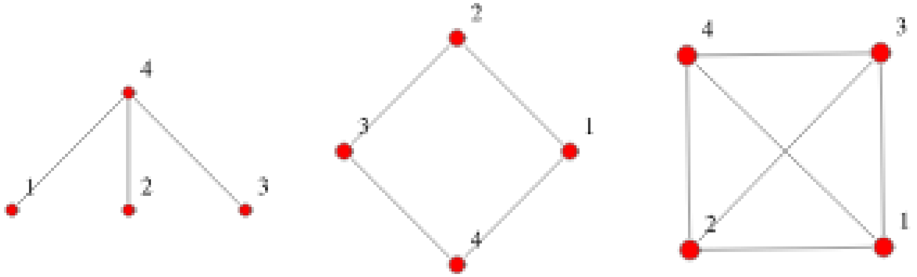


Correlation of columns identifies networks of coregulated and coexpressed genome sites.

Correlation of rows identifies related tissues and coregulating factors.

Correlation of rows and columns shown as biplots of coregulating factors and their coregulated sites.

- Turn data into a graph
- Alternate local and global clustering while optimizing for **modularity**
- Can discover the number of clusters given a resolution
  - Cell type detection
- Fast: O(nlogn)
- E.g. Louvain, Leiden Community detection



$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

Adjacency Matrix & Graphs

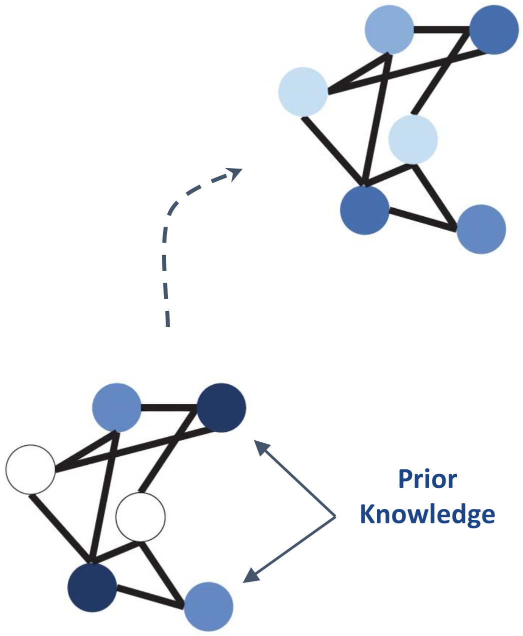
Correlating rows or columns

$R = AA^T$

$C = A^TA$

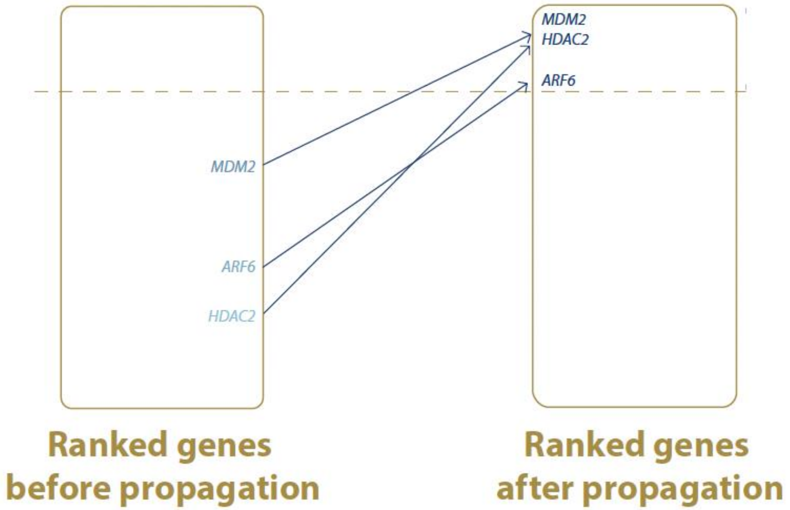
mathworld.wolfram.com/AdjacencyMatrix.html

Network propagation-based prioritization



- Changes in gene rank before & after propagation

Gene rank mobility



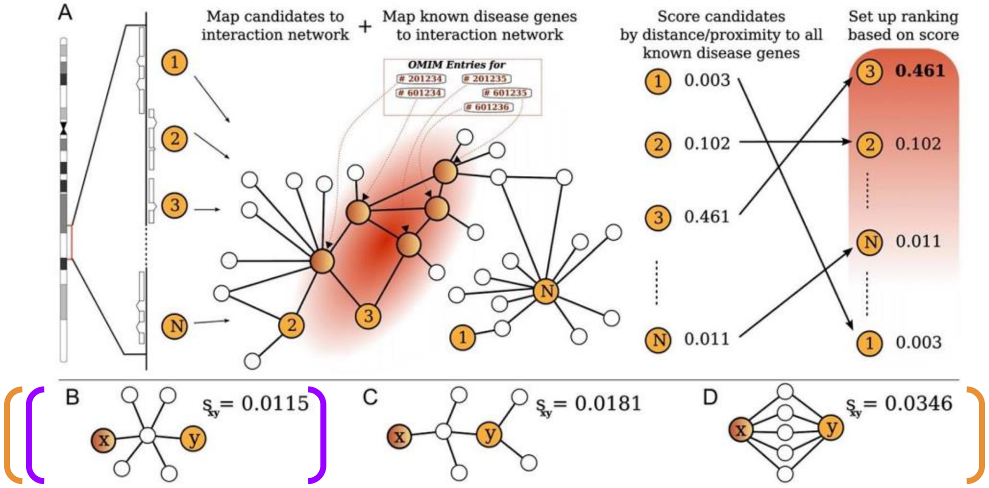
$$S_{(t+1)} = \alpha W' S_{(t)} + (1-\alpha)\, S_{(0)}$$

[Mohsen et al. Genome Biol. (‘21)]

Network Propagation in Biomedicine (Label propagation & Diffusion distance)

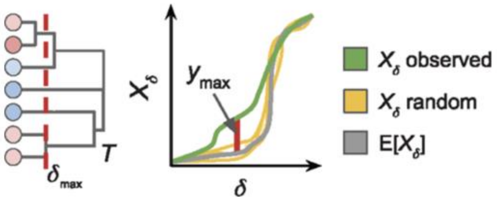
- Starting ~2008
- Limitations in nearest neighbor (B) and shortest distance measures (B-D)
- Leverages local and global network topology
- Mathematically rigorous

- Early methods: function prediction and gene-disease association
- Current methods: gene ranking, subnetwork detection, gene-drug and TF-target associations, patient sample stratification, etc.

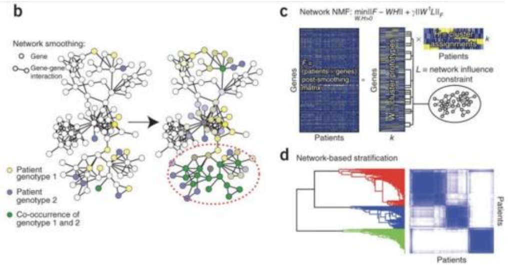


A
Map candidates to interaction network
+
Map known disease genes to interaction network
Score candidates by distance/proximity to all known disease genes
Set up ranking based on score
OMIM Entries for
# 201234
# 201235
# 601234
# 601235
# 601236
1
2
3
N
1 0.003
2 0.102
3 0.461
N 0.011
3 0.461
2 0.102
N 0.011
1 0.003
B
x
y
S_{xy} = 0.0115
C
S_{xy} = 0.0181
D
S_{xy} = 0.0346

Köhler, Bauer, Horn and Robinson (2008)



Reyna, Leiserson, and Raphael (2018)

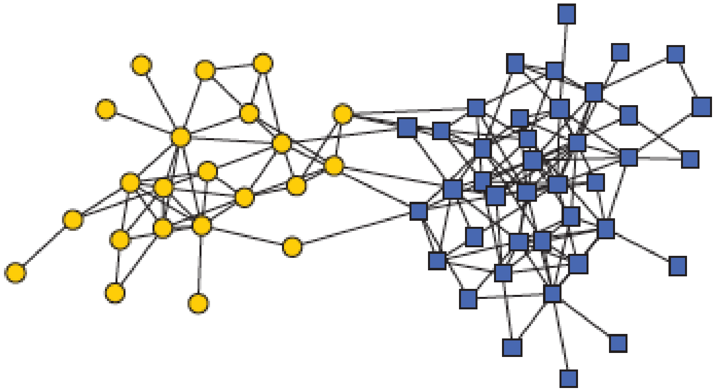


Hofree, Shen, Carter, Gross, and Ideker (2013)

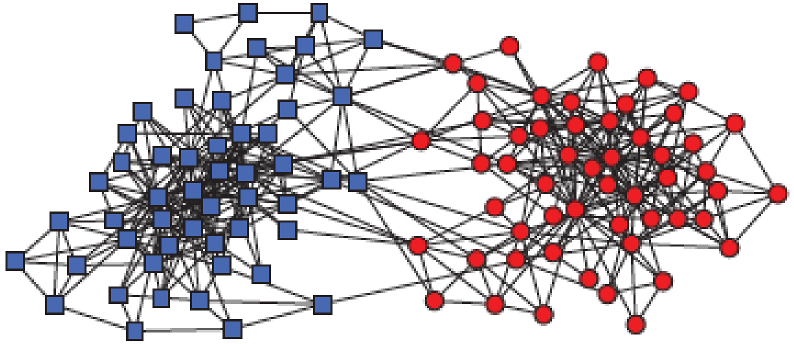
Unsupervised Mining

Community Detection
Application to Hi-C

Network modularity



Dolphin social network



Political books

Newman Phy. Rev. E 2013

degree of node i

adjacency matrix

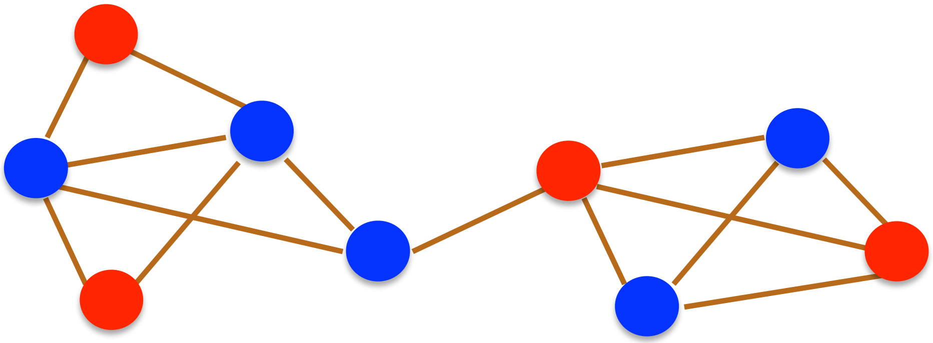
$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \boxed{\frac{k_i k_j}{2m}} \right) \delta_{\sigma_i \sigma_j}$$

whether or not i, j are in the same module

number of edges

expected number of edges between i and j

Network modularity



$$Q \approx 0$$

degree of node i

adjacency matrix

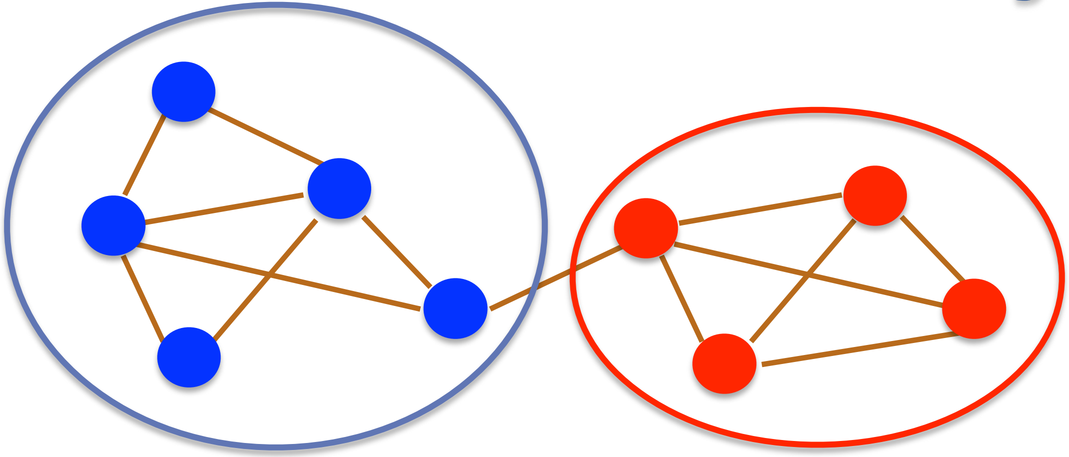
$$Q = \frac{1}{2m}\sum_{i,j}\left(W_{ij} - \boxed{\frac{k_i k_j}{2m}}\right)\delta_{\sigma_i \sigma_j}$$

whether or not
i, j are in the
same module

number of edges

expected number of
edges between i and j

Network modularity



$$Q = Q_{max}$$

Optimization problem for sim. annealing

adjacency matrix

degree of node i

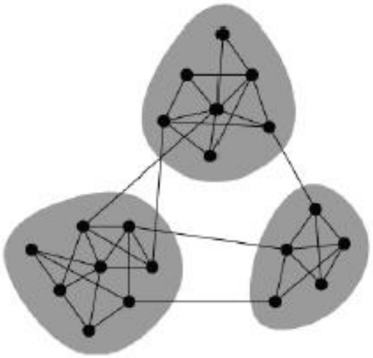
$$Q = \frac{1}{2m}\sum_{i,j}\left(W_{ij} - \boxed{\frac{k_i k_j}{2m}}\right)\delta_{\sigma_i\sigma_j}$$

whether or not i, j are in the same module

number of edges

expected number of edges between i and j

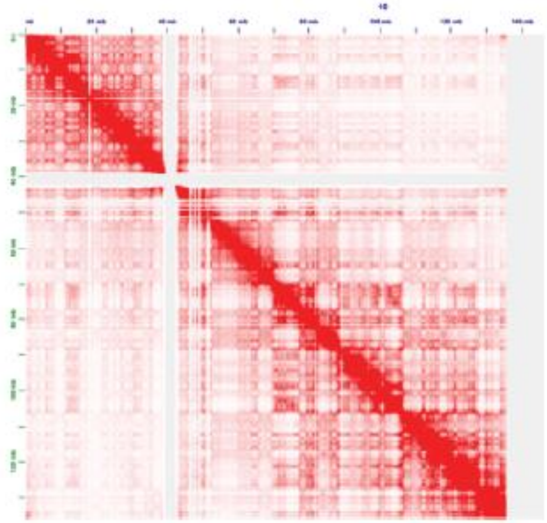
Identifying TADs in multiple resolutions



Modularity maximization

$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

network	contact map
node	chromosome bin
edge	Hi-C contact
# of connections	coverage
module	domain

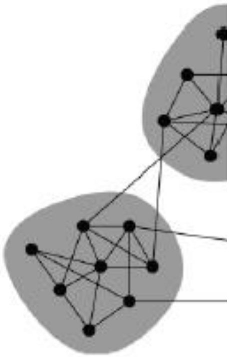


schematic adapted from ref. [2]

…..Resuming from “Multi-omics pack”

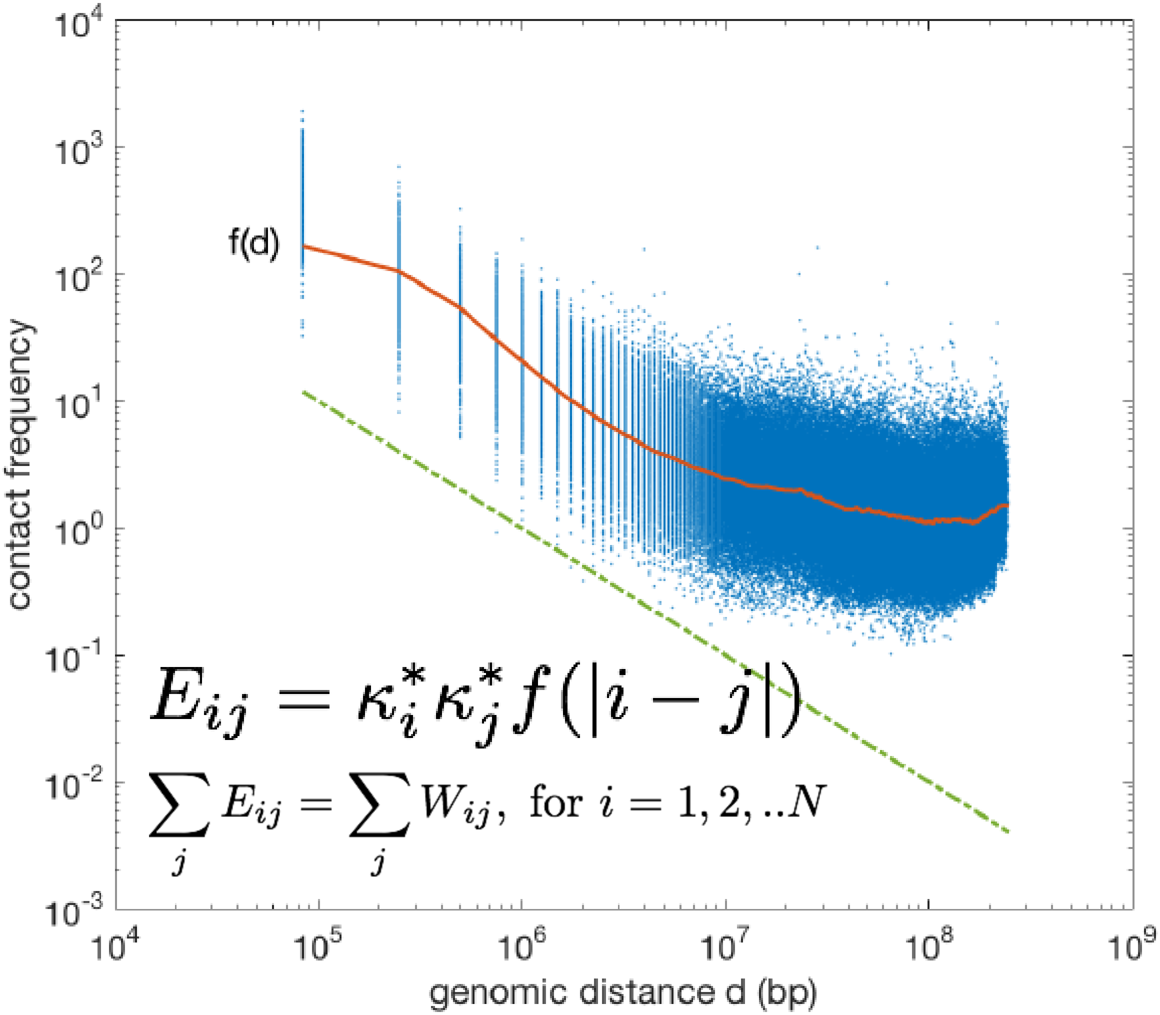
[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

Identifying TADs in multiple resolutions



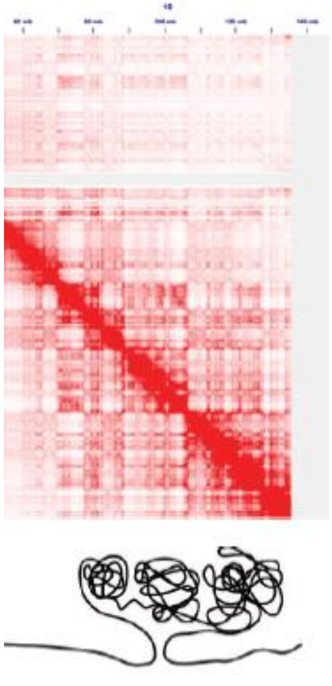
Modularity m

$$Q = \frac{1}{2m}\sum_{i,j}$$



$$E_{ij} = \kappa_i^* \kappa_j^* f\left(|i-j|\right)$$

$$\sum_j E_{ij} = \sum_j W_{ij}, \text{ for } i = 1, 2, ..N$$



: adapted from ref. [2]

[Yan et al., *PLOS Comp. Bio.* (in revision, '17); bioRxiv 097345]