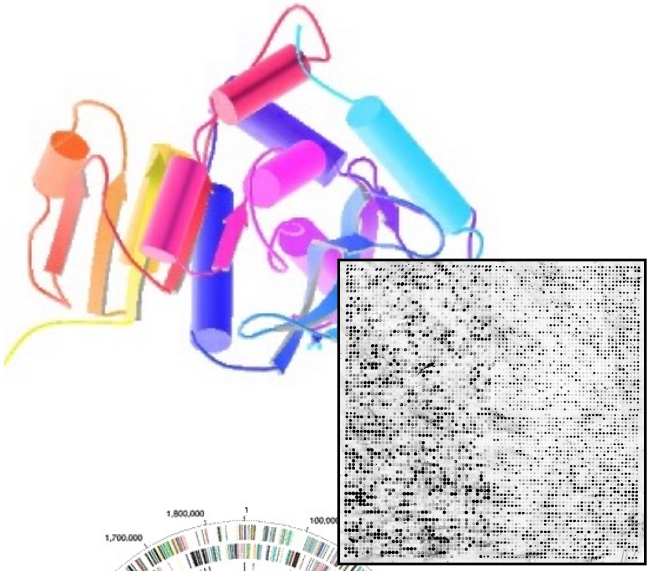
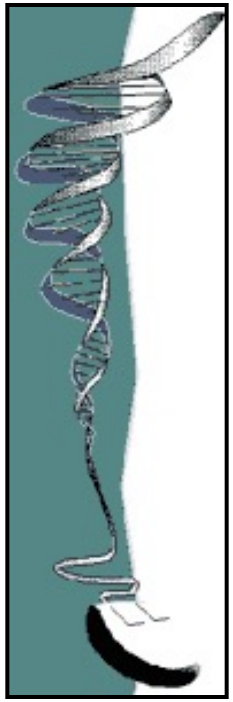
Biomed. Data Science:

Unsupervised Datamining B: Community Detection





Mark Gerstein, Yale University

gersteinlab.org/courses/452

(last edit in spring ’21, pack #9b, final)

Unsupervised Mining

Graph Analysis &
Community Detection Approaches

Graph Methods & Community Detection

- 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

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \frac{s_v s_w + 1}{2}$$

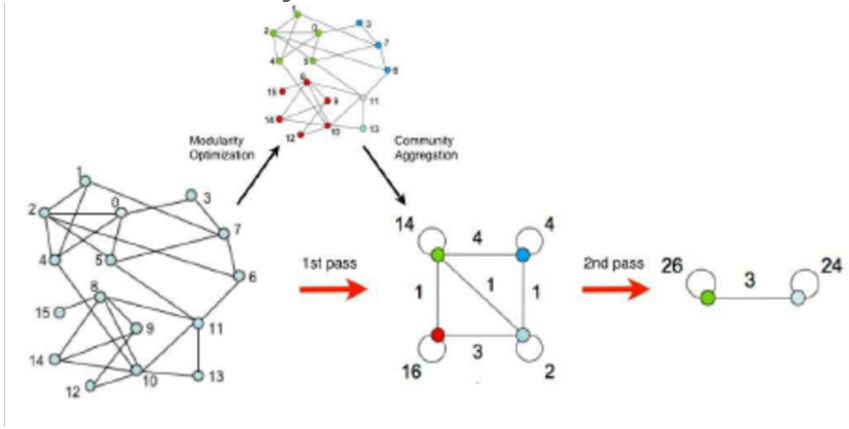
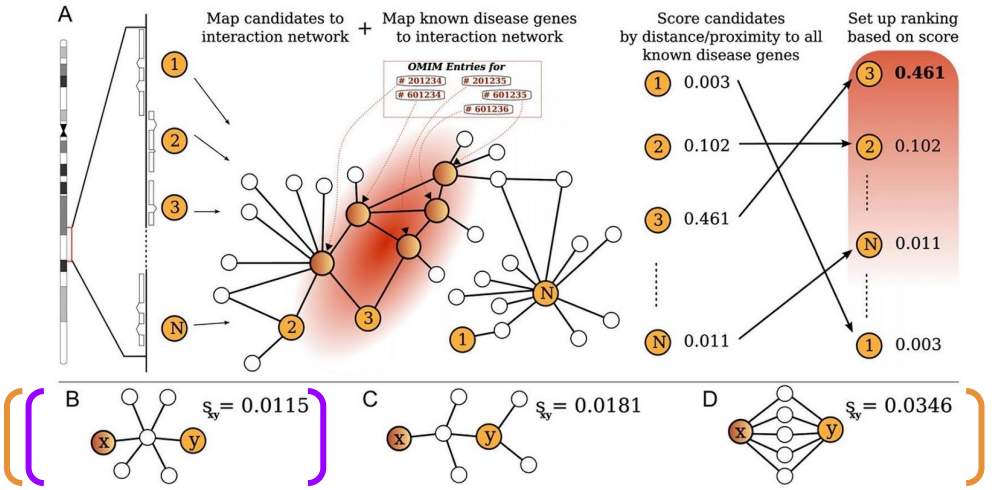


Image reference: https://sites.google.com/site/findcommunities/

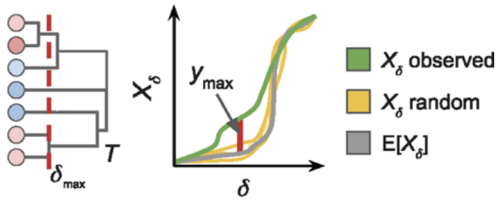
Network Propagation in Biomedicine (Label propagation & Diffusion distance)

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

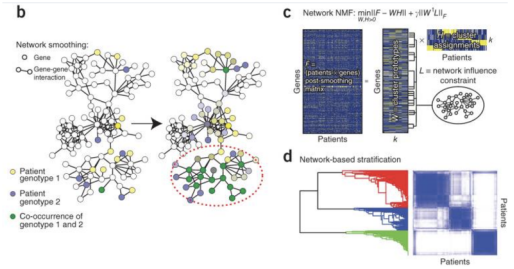
- Early methods: function prediction and gene-disease association
- Current methods: gene ranking, subnetwork detection, gene-drug and TF-target associations, patient sample stratification, etc.



Köhler, Bauer, Horn and Robinson (2008)



Reyna, Leiserson, and Raphael (2018)



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

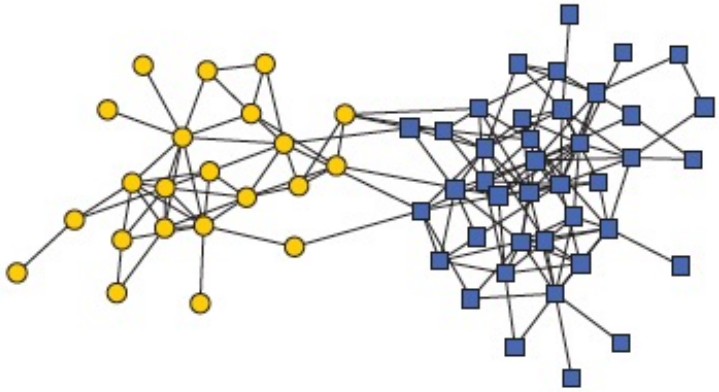
Dimensionality Reduction & Spectral Methods Outline & Papers

- Based on affinity matrix
- PCA/SVD
- Extensions: biplot, RCA, CCA….
- Related papers
  - O Alter et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." PNAS 97: 10101
  - Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 2007, 1:54
  - Z Zhang et al. (2007) "Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions." Genome Res 17: 787
  - TA Gianoulis et al. (2009) "Quantifying environmental adaptation of metabolic pathways in metagenomics." PNAS 106: 1374.

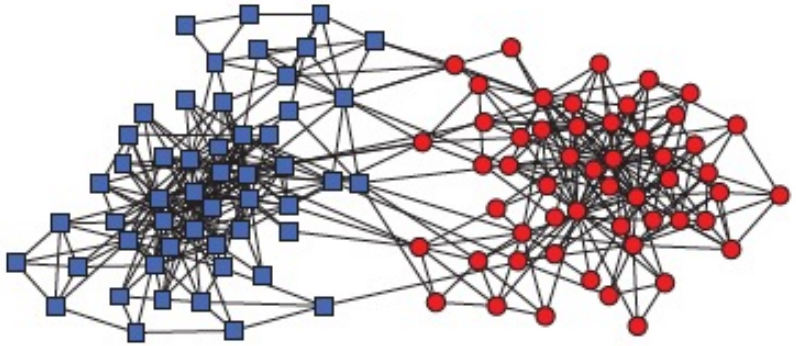
Unsupervised Mining

Community Detection
Application to Hi-C

Network modularity



Dolphin social network



Political books

Newman Phy. Rev. E 2013

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

adjacency matrix

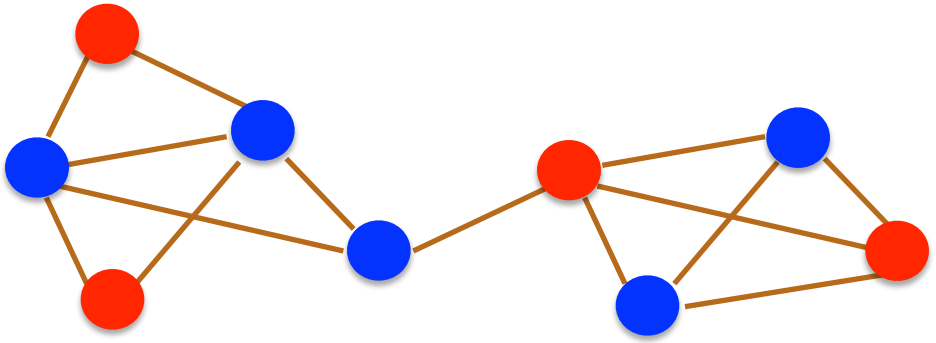
degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module

Network modularity



$$Q \approx 0$$

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

adjacency matrix

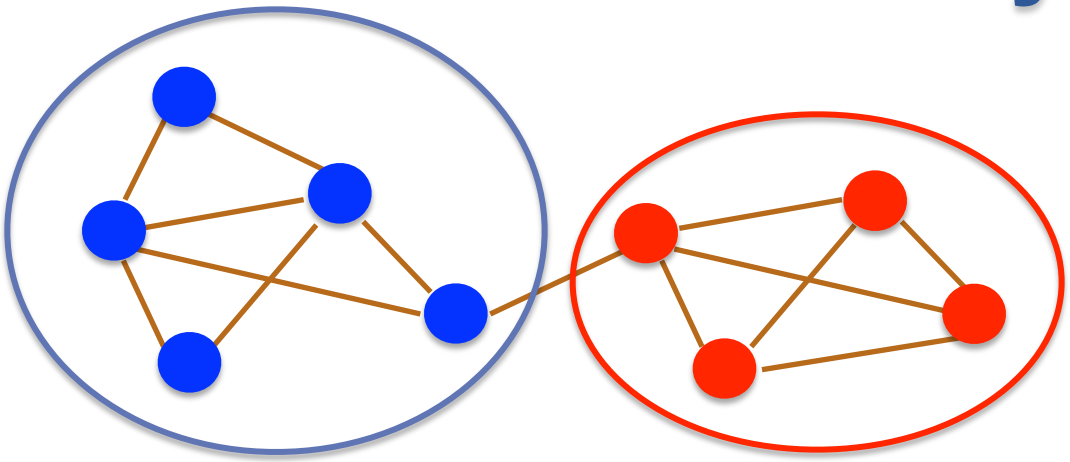
degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module

Network modularity



$$Q = Q_{max}$$

Optimization problem for sim. annealing

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

adjacency matrix

degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module

3D organization of genome

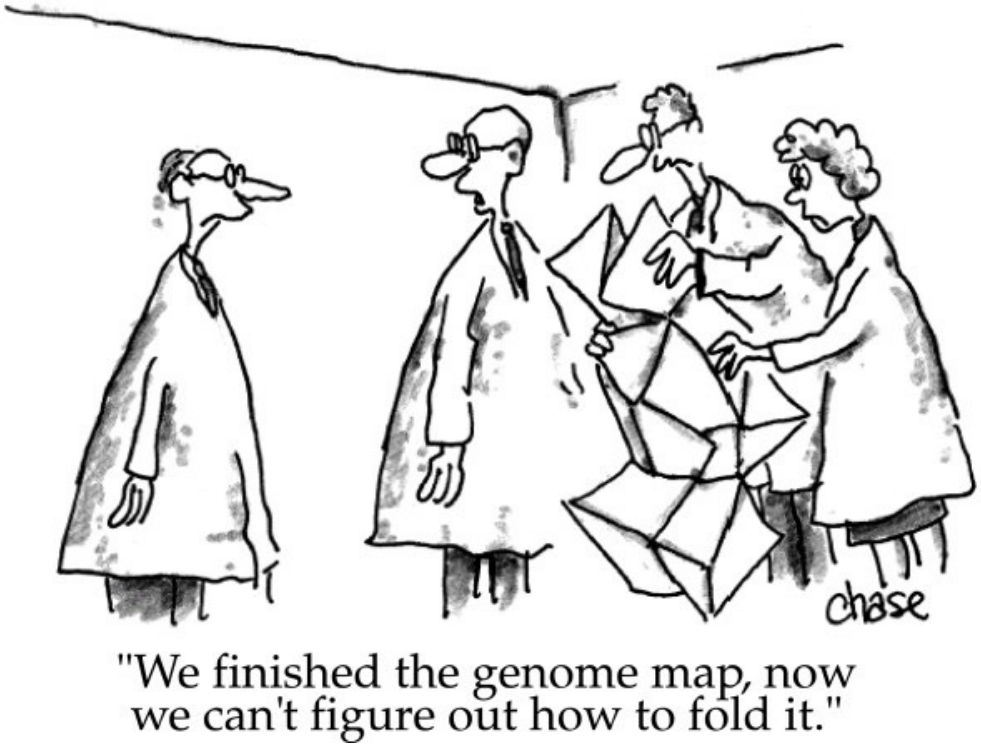


image credit: Iyer et al. BMC Biophysics 2011, cartoonist John Chase

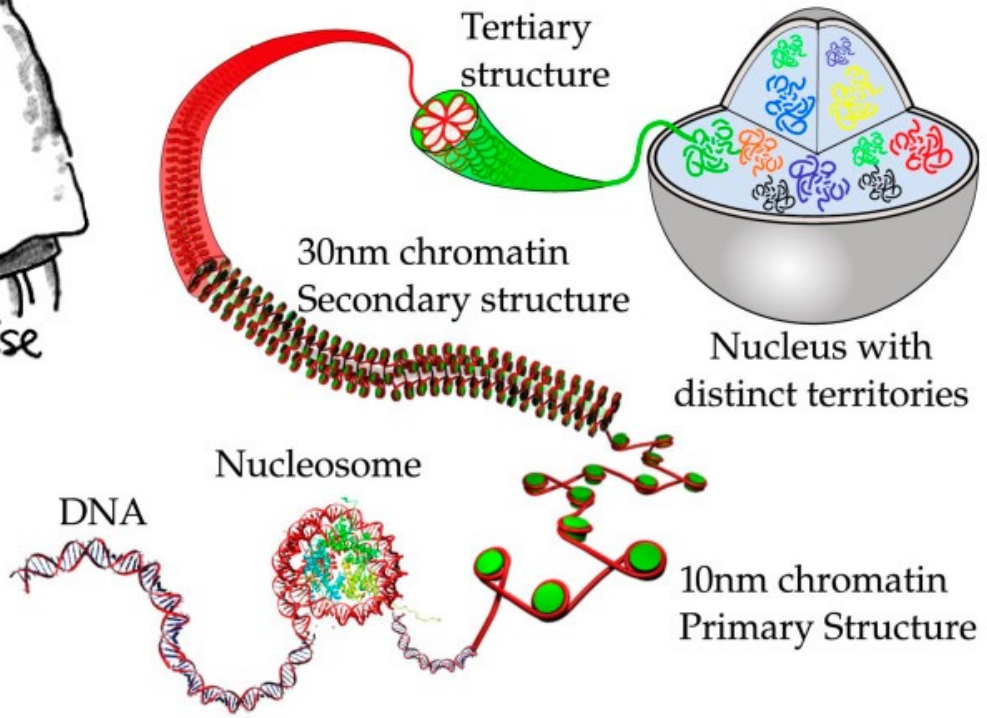
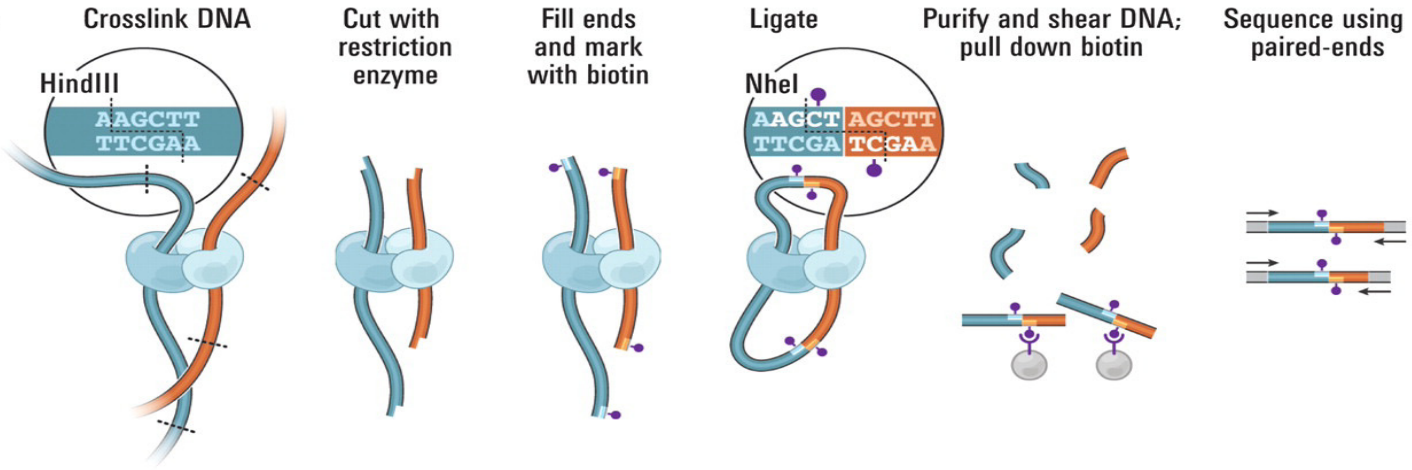
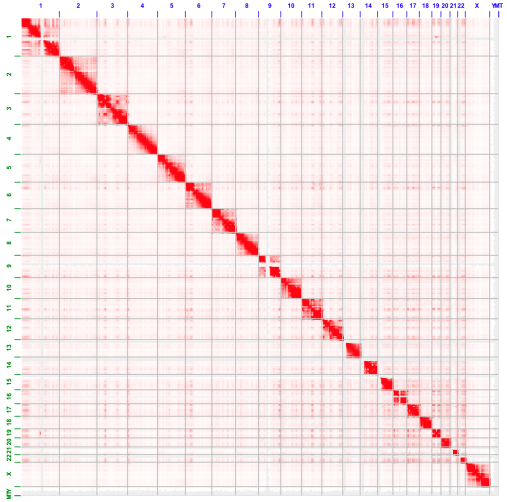


image credit: Iyer et al. BMC Biophysics 2011

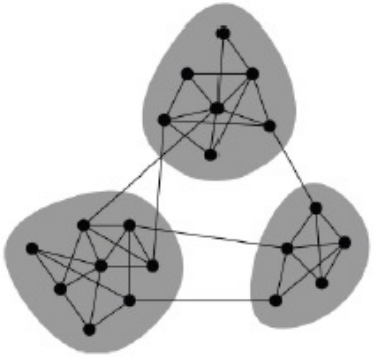
Hi-C contact map





Science 2009, 5950: 289-293

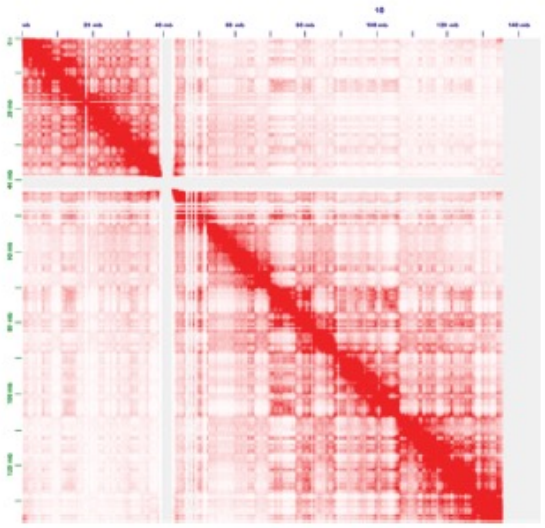
Identifying TADs in multiple resolutions



Modularity maximization

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

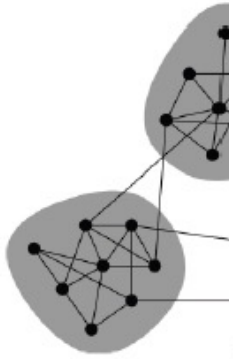
network	contact map
node	chromosome bin
edge	Hi-C contact
# of connections	coverage
module	domain



schematic adapted from ref. [2]

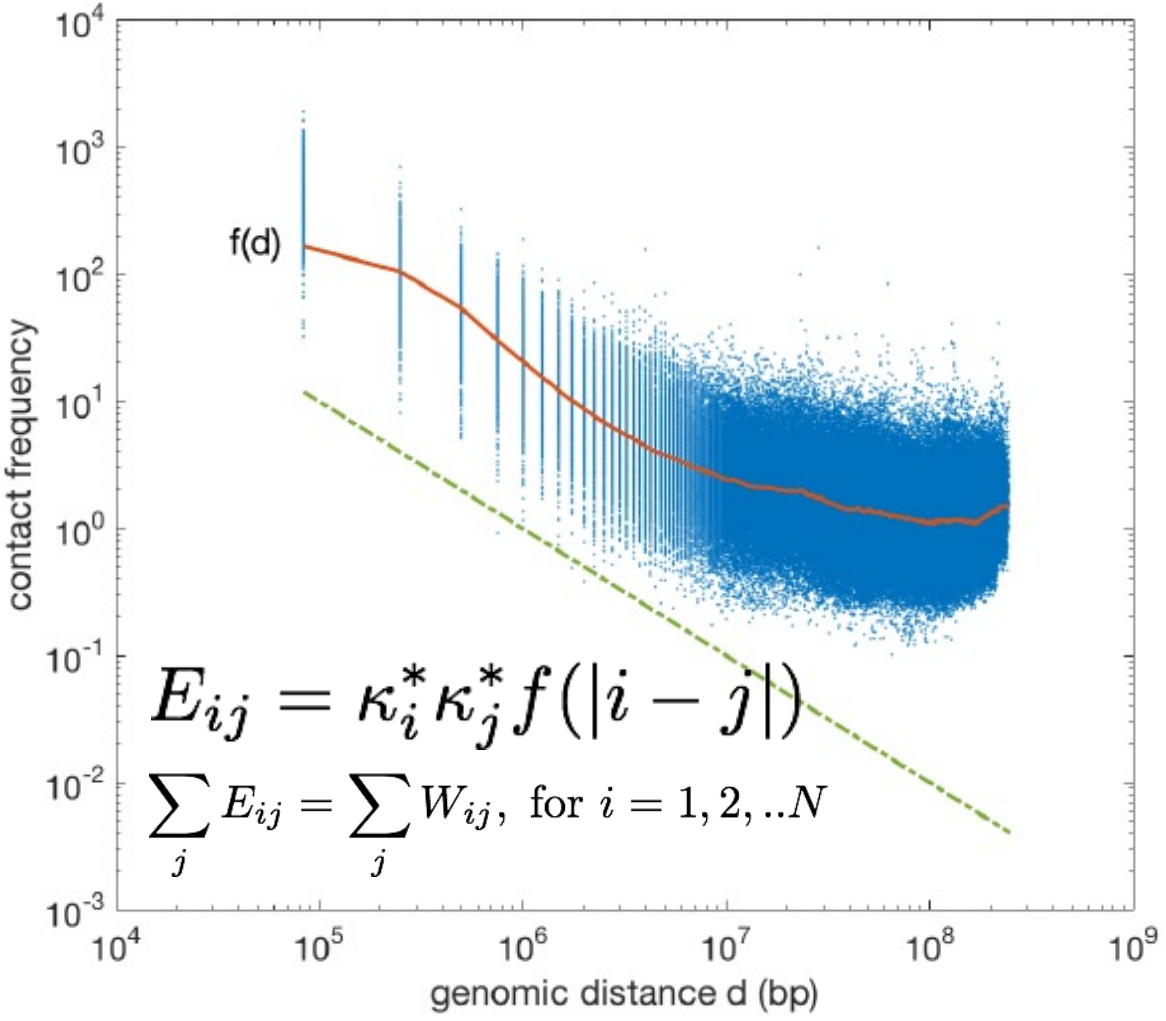
[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

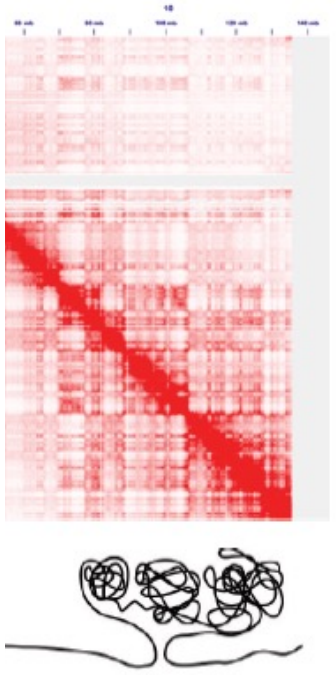
Identifying TADs in multiple resolutions



Modularity m

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





: adapted from ref. [2]

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