# Regulatory Motif Finding II

Lectures 13 – Nov 9, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

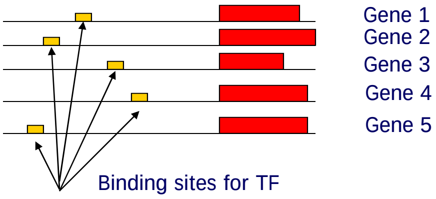
Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

## Outline

- Regulatory motif finding
  - PWM, scoring function
  - Expectation-Maximization (EM) methods (MEME)
  - Gibbs sampling methods (AlignAce, BioProspector)
- More computational methods
  - Greedy search method (CONSENSUS)
  - Phylogenetic foot-printing method
  - Graph-based methods (MotifCut)

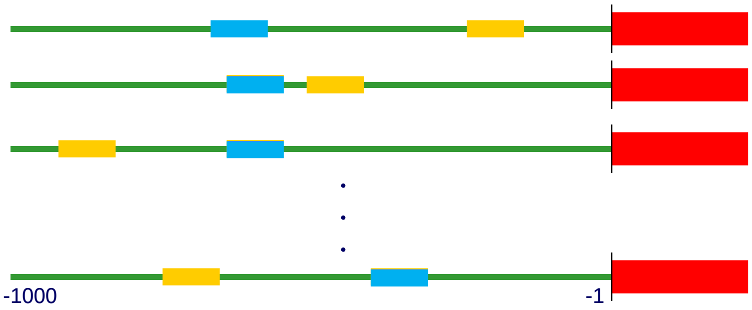
# Finding Regulatory Motifs

- Say a transcription factor (TF) controls five different genes
- Each of the five genes will have binding sites for the TF in their promoter region



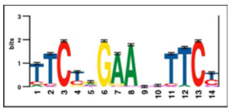
# Finding Regulatory Motifs

- Given the upstream sequences of the genes that seem to be regulated by the same TFs,
- Find the TF-binding sites (motifs) in common

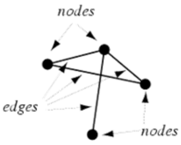


# Motif representation

- Consensus sequence
  - May allow “degenerate” symbols in sequence
  - E.g. N=A/C/G/T; W=A/T; S=C/G; R=A/G; Y=T/C etc
    NTCATWCAS
- Position specific scoring matrix
  - Position weight matrix (PWM)



- A graph
  - Node: k-mer
  - Edge: distance between k-mers



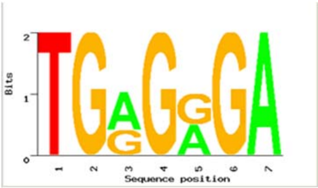
# Position Weight Matrix (PWM)

- The most widely used representation
- Assign probability to (A,G,C,T) in each position
- Example
  - Say that a TF binds to the following 5 sequences:

TGGGGGA
TGAGAGA
TGGGGGA
TGAGAGA
TGAGGGA

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **A** | 0 | 0 | 0.6 | 0 | 0.4 | 0 | 1 |
| **C** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **G** | 0 | 1 | 0.4 | 1 | 0.6 | 1 | 0 |
| **T** | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

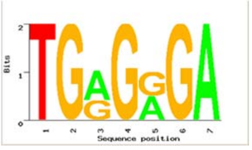
- Representations called motif logos illustrate the conserved and variable regions of a motif



# Position Weight Matrix (PWM)

- Let W be a PWM for a motif of length k, and S be an input sequence.
- How is a subsequence s (of length k) in S evaluated?
  - Probabilistic score P(s|W)
  - e.g. W (k=7):

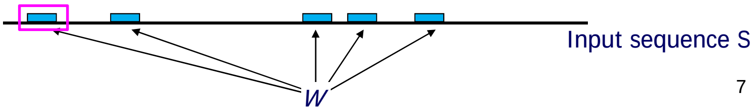
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **A** | 0.1 | 0 | 0.6 | 0 | 0.4 | 0 | 1 |
| **C** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **G** | 0 | 1 | 0.4 | 1 | 0.6 | 1 | 0 |
| **T** | 0.9 | 0 | 0 | 0 | 0 | 0 | 0 |



s : AGAGAGA

P(s|W) = (0.1) x (1) x (0.6) x (1) x (0.4) x (1) x (1)

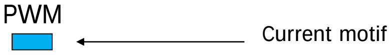
- Given W, we can scan the input sequence S for good matches to the motif


Input sequence S

W

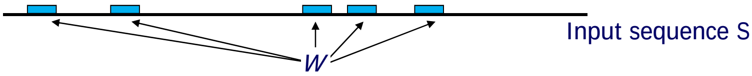
# Motif Finding Using EM Algorithm

- MEME works by iteratively refining PWMs and identifying sites for each PWM
  - 1. Estimate motif model (PWM)
    - Start with a k-mer seed (random or specified)
    - Build a PWM by incorporating some of background frequencies


PWM

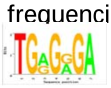
Current motif

  - 2. Identify examples of the model
    - For every k-mer in the input sequences, identify its probability given the PWM model.


Input sequence S

W

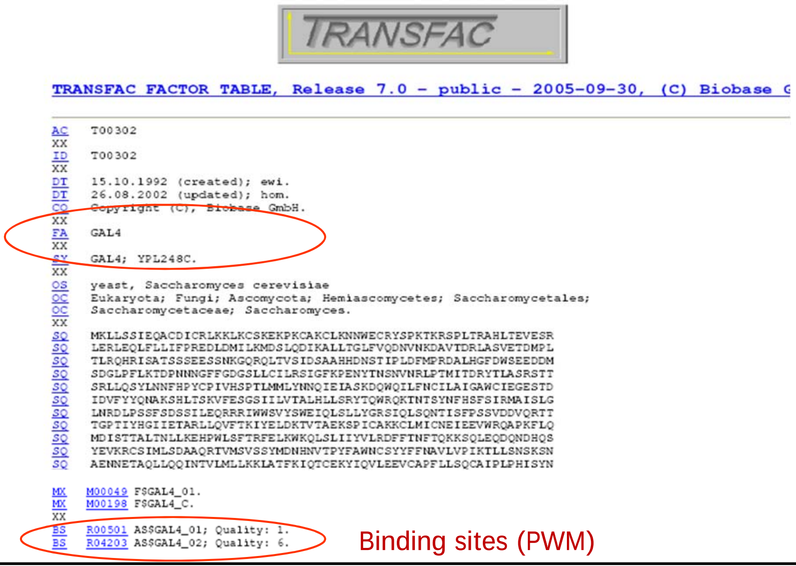
  - 3. Re-estimate the motif model
    - Calculate a new PWM, based on the weighted frequencies of all k-mers in the input sequences



  - 4. Iterate 2 & 3 until convergence.

# Databases

TRANSFAC: http://www.gene-regulation.com/pub/databases.html#transfac



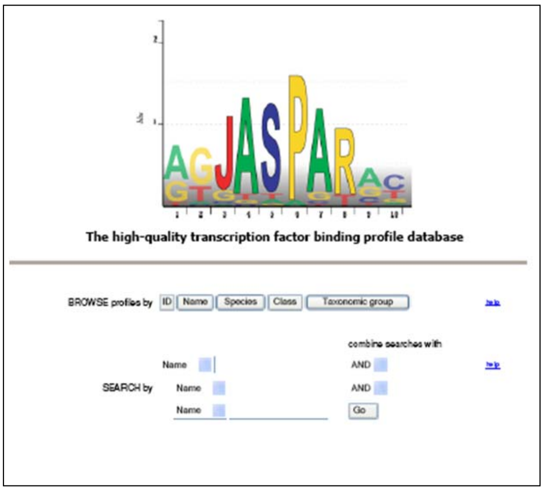
TRANSFAC FACTOR TABLE, Release 7.0 - public - 2005-09-30, (C) Biobase G

```
AC   T00302
XX
ID   T00302
XX
DT   15.10.1992 (created); ewi.
DT   26.08.2002 (updated); hom.
CO   Copyright (C), Biobase GmbH.
XX
FA   GAL4
XX
SY   GAL4; YPL248C.
XX
OS   yeast, Saccharomyces cerevisiae
OC   Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
OC   Saccharomycetaceae; Saccharomyces.
XX
SQ   MKLLSSIEQACDICRLKKLKCSKEKPKCAKCLKNNWECRYSPKTKRSPLTRAHLTEVESR
SQ   LERLEQLFLLIFPREDLDMILKMDSLQDIKALLTGLFVQDNVNKDAVTDRLASVETDMPL
SQ   TLRQHRISATSSSEESSNKGQRQLTVSIDSAAHHDNSTIPLDFMPRDALHGFDWSEEDDM
SQ   SDGLPFLKTDPNNNGFFGDGSLLCILRSIGFKPENYTNSNVNRLPTMITDRYTLASRSTT
SQ   SRLLQSYLNNFHPYCPIVHSPTLMMLYNNQIEIASKDQWQILFNCILAIGAWCIEGESTD
SQ   IDVFYYQNAKSHLTSKVFESGSIILVTALHLLSRYTQWRQKTNTSYNFHSFSIRMAISLG
SQ   LNRDLPSSFSDSSILEQRRRIWWSVYSWEIQLSLLYGRSIQLSQNTISFPSSVDDVQRTT
SQ   TGPTIYHGIIETARLLQVFTKIYELDKTVTAEKSPICAKKCLMICNEIEEVWRQAPKFLQ
SQ   MDISTTALTNLLKEHPWLSFTRFELKWKQLSLIIYVLRDFFTNFTQKKSQLEQDQNDHQS
SQ   YEVKRCSIMLSDAAQRTVMSVSSYMDNHNVTPYFAWNCSYYFFNAVLVPIKTLLSNSKSN
SQ   AENNETAQLLQQINTVLMLLKKLATFKIQTCEKYIQVLEEVCAPFLLSQCAIPLPHISYN

MX   M00049 F$GAL4_01.
MX   M00198 F$GAL4_C.
XX
BS   R00501 AS$GAL4_01; Quality: 1.
BS   R04203 AS$GAL4_02; Quality: 6.
```

Binding sites (PWM)

# More Databases



The high-quality transcription factor binding profile database

Species-specific:

SCPD (yeast) http://rulai.cshl.edu/SCPD/

DPInteract (e. coli) http://arep.med.harvard.edu/dpinteract/

Drosophila DNase I Footprint Database (v2.0) http://www.flyreg.org/

# Outline

- Regulatory motif finding
  - PWM, scoring function
  - Expectation-Maximization (EM) methods (MEME)
  - Gibbs sampling methods (AlignAce, BioProspector)

- More computational methods
  - Greedy search method (CONSENSUS)
  - Phylogenetic foot-printing method
  - Graph-based methods (MotifCut)

# CONSENSUS

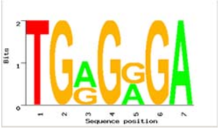
- Popular algorithm for motif discovery, that uses a greedy approach
- Motif model: Position Weight Matrix (PWM)
- Motif score: information content

# Information Content

- PWM W:
  - $W_{\beta k}$ = frequency of base β at position k
  - $q_\beta$ = frequency of base β by chance

$W_{A1}$, $W_{C1}$, $W_{G1}$, $W_{T1}$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **A** | 0.1 | 0 | 0.6 | 0 | 0.4 | 0 | 1 |
| **C** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **G** | 0 | 1 | 0.4 | 1 | 0.6 | 1 | 0 |
| **T** | 0.9 | 0 | 0 | 0 | 0 | 0 | 0 |



- Information content of W:

$$\sum_{k} \sum_{\beta \in \{A,C,G,T\}} W_{\beta k} \log \frac{W_{\beta k}}{q_\beta}$$

# Information Content

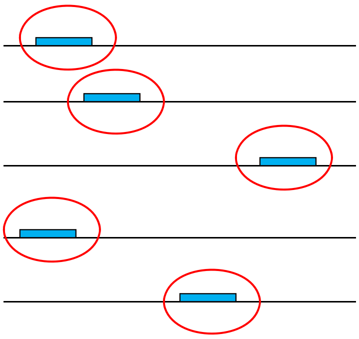
- If $W_{\beta k}$ is always equal to $q_\beta$, i.e., if W is similar to random sequence, information content of W is 0.
- If W is different from q, information content is high.

- Information content of W:

$$\sum_{k} \sum_{\beta \in \{A,C,G,T\}} W_{\beta k} \log \frac{W_{\beta k}}{q_\beta}$$

# CONSENSUS: Basic Idea

- Find a set of subsequences, one in each input sequence

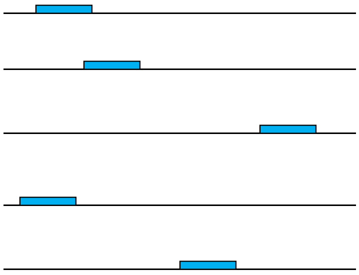


Set of subsequences define a PWM.

Goal: This PWM should have high information content.

High information content means that the motif “stands out”.

# CONSENSUS: Basic Idea



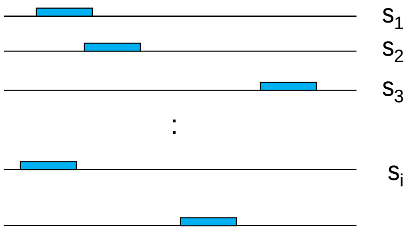
Start with a subsequence in one input sequence

Build the set of subsequences incrementally, adding one subsequence at a time

Until the entire set is built

# CONSENSUS: the greedy heuristic

- Suppose we have built a partial set of subsequences $\{s_1,s_2,\ldots,s_i\}$ so far.
- Have to choose a subsequence $s_{i+1}$ from the input sequence $S_{i+1}$
- Consider each subsequence s of $S_{i+1}$
- Compute the score (information content) of the PWM made from $\{s_1,s_2,\ldots,s_i,s\}$
- Choose the s that gives the PWM with highest score, and assign $s_{i+1} \leftarrow s$

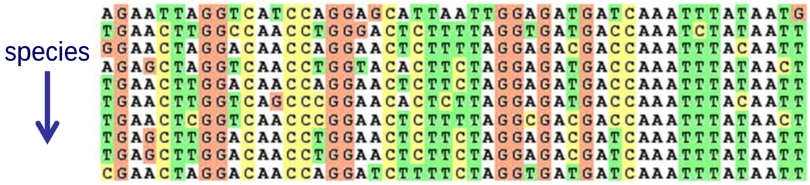


# Outline

- Regulatory motif finding
  - PWM, scoring function
  - Expectation-Maximization (EM) methods (MEME)
  - Gibbs sampling methods (AlignAce, BioProspector)
- More computational methods
  - Greedy search method (CONSENSUS)
  - Phylogenetic foot-printing method
  - Graph-based methods (MotifCut)

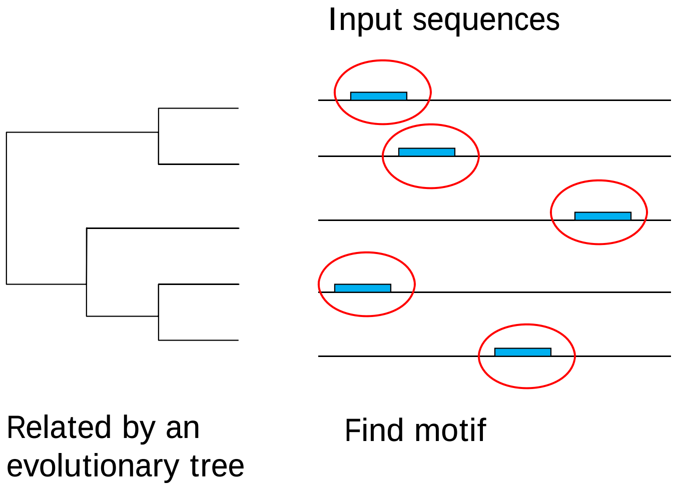
# Phylogenetic footprinting

- So far, the input sequences were the “upstream” (promoter) regions of genes believed to be “co-regulated”
- A special case: the input sequences are promoter regions of the same gene, but from multiple species.
  - Such sequences are said to be “orthologous” to each other.



```
AGAATTAGGTCATCCAGGAGCATTAATTGGAGATGATCAAATTTATAATG
TGAACTTGGCCAACCTGGGACTCTTTTAGGTGATGACCAAATCTATAATT
GGAACTAGGACAACCAGGAACTCTTTTAGGAGACGACCAAATTTACAATT
AGAGCTAGGTCAACCTGGTACACTTCTAGGAGATGACCAAATTTATAACT
TGAACTTGGACAACCAGGAACTCTTCTAGGAGATGACCAAATTTATAATT
TGAACTTGGTCAGCCCGGAACACTCTTAGGAGATGACCAAATTTACAATT
TGAACTCGGTCAACCCGGAACTCTTTTAGGCGACGACCAAATTTATAACT
TGAGCTTGGACAACCTGGAACTCTTCTAGGAGACGATCAAATTTATAATT
TGAGCTTGGACAACCTGGAACTCTTCTAGGAGACGATCAAATTTATAATT
CGAACTAGGACAACCAGGATCTTTTCTAGGTGATGATCAAATTTATAATT
```

# Phylogenetic Footprinting



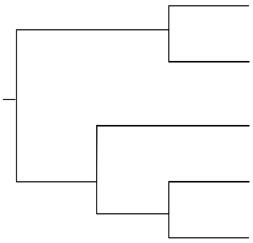
# Phylogenetic Footprinting

- Formally speaking,

- Given:
  - Phylogenetic tree *T*,
  - set of orthologous sequences at leaves of *T*,
  - length *k* of motif
  - threshold *d*

- Problem:
  - Find each set *S* of *k*-mers*,* one *k*-mer from each leaf, such that the “parsimony” score of *S* in *T* is at most *d*.

# Small Example



**AGTCGTACGTGAC**... (Human)

**AGTAGACGTGCCG**... (Chimp)

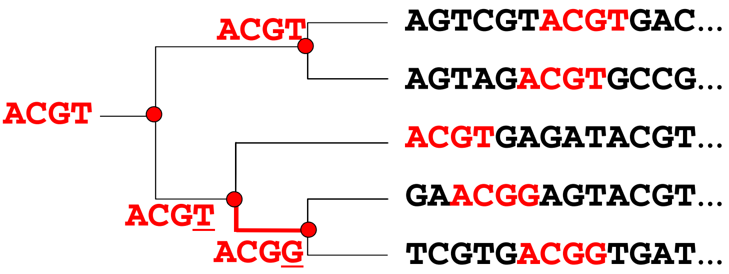
**ACGTGAGATACGT**... (Rabbit)

**GAACGGAGTACGT**... (Mouse)

**TCGTGACGGTGAT**... (Rat)

Size of motif sought: k = 4

# Solution

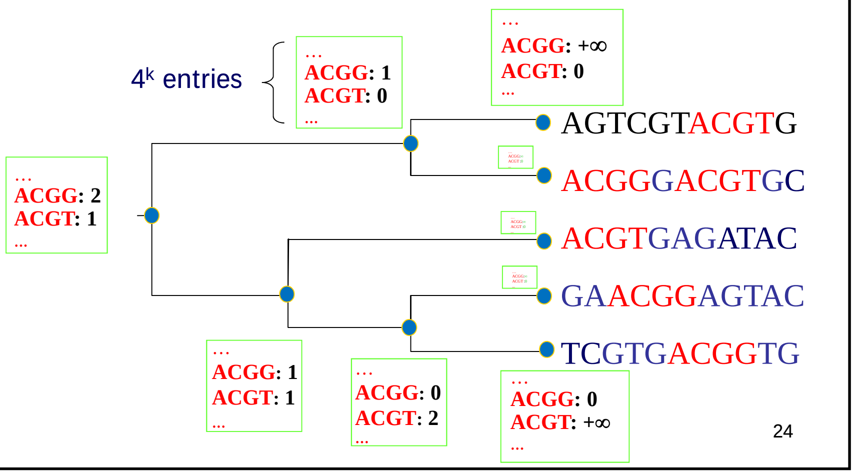


Parsimony score: 1 mutation

# An Exact Algorithm

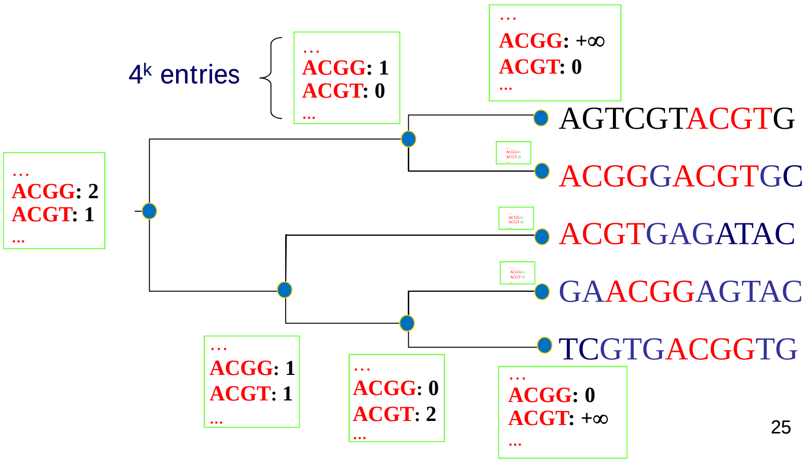
(Blanchette's algorithm)

$W_u$ [s] = best parsimony score for subtree rooted at node u, if u is labeled with string s.



## Recurrence

$$W_u[s] = \sum_{v:\text{ child of } u} \min_{t} \left( W_v[t] + d(s,t) \right)$$



## Running Time

$$W_u[s] = \sum_{v:\text{ child of } u} \min_{t} \left( W_v[t] + d(s,t) \right)$$

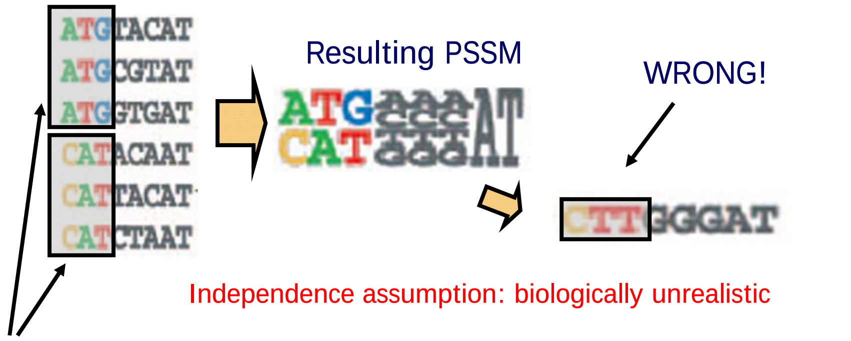
$O(k \cdot 4^{2k})$
time per node

# Outline

- Regulatory motif finding
  - PWM, scoring function
  - Expectation-Maximization (EM) methods (MEME)
  - Gibbs sampling methods (AlignAce, BioProspector)

- More computational methods
  - Greedy search method (CONSENSUS)
  - Phylogenetic foot-printing method
  - Graph-based methods (MotifCut)

# Drawbacks of Existing Methods



Resulting PSSM

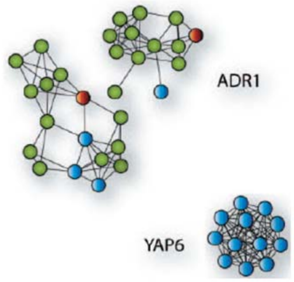
WRONG!

Independence assumption: biologically unrealistic

Perfectly conserved nucleotide dependency — ATG and CAT

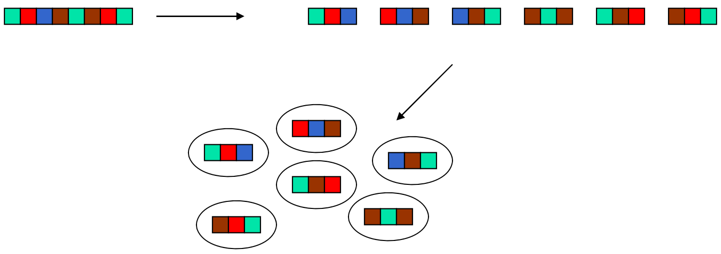
# Overview

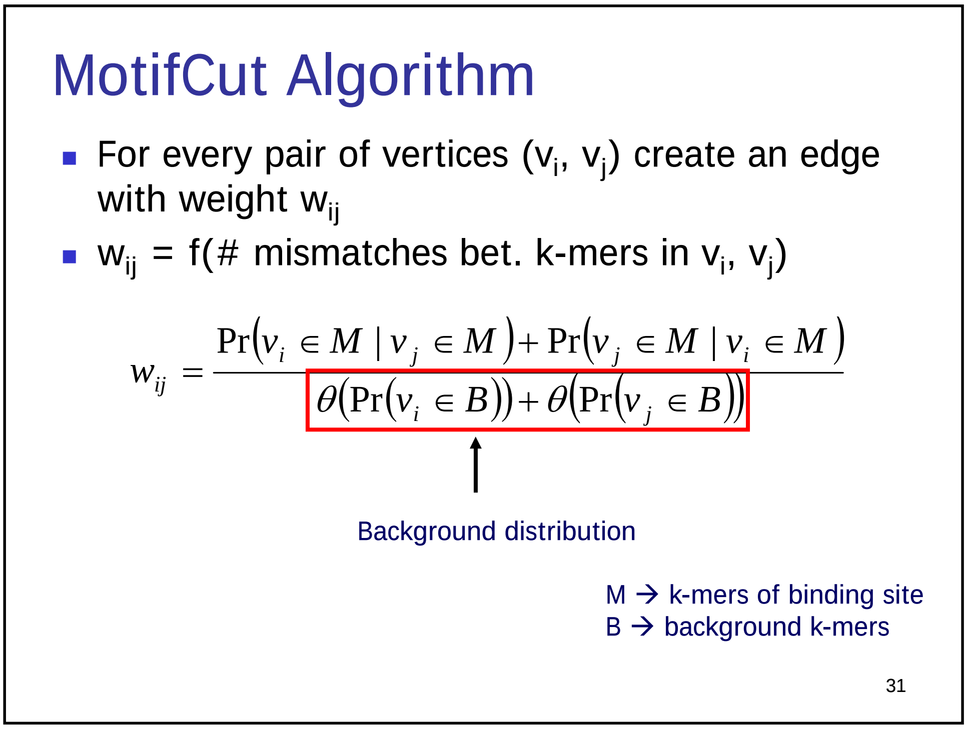
- Nodes: k-mers of input sequence
- Edges: pairwise k-mer similarity
- Motif search → maximum density subgraph



# MotifCut Algorithm

- Convert sequence into a collection of k-mers
  - Each overlap/duplicate considered distinct





# MotifCut Algorithm

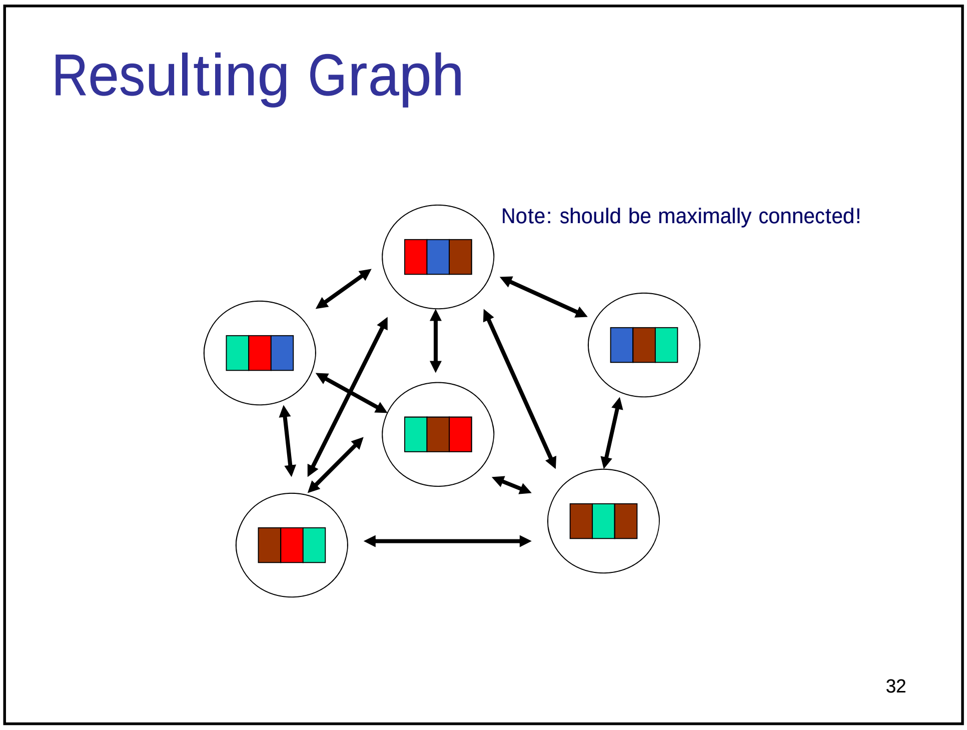
- For every pair of vertices ($v_i$, $v_j$) create an edge with weight $w_{ij}$
- $w_{ij}$ = f(# mismatches bet. k-mers in $v_i$, $v_j$)

$$w_{ij} = \frac{\Pr(v_i \in M \mid v_j \in M) + \Pr(v_j \in M \mid v_i \in M)}{\boxed{\theta(\Pr(v_i \in B)) + \theta(\Pr(v_j \in B))}}$$

Background distribution

M → k-mers of binding site<br>
B → background k-mers

31



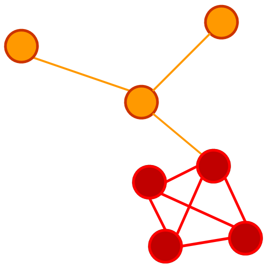
# Resulting Graph

Note: should be maximally connected!

32

# Motif Finding

- Find highest density subgraph



- Density is defined as sum of edge weights per node
- Find the maximum density subgraph (MDS)

# What After Motif Finding ?

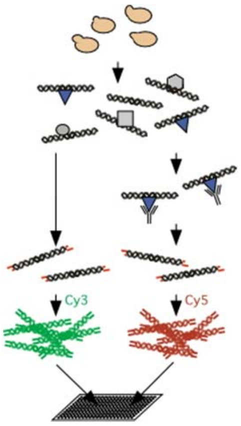
- Experiments to confirm results
- DNaseI footprinting & gel-shift assays
- Tells us which subsequences are the binding sites

## Before Motif Finding

- How do we obtain a set of sequences on which to run motif finding ?
- In other words, how do we get genes that we believe are regulated by the same transcription factor ?
- Two high-throughput experimental methods: ChIP-chip and microarray.

## Before Motif Finding

- ChIP-chip
  - Take a particular transcription factor TF
  - Take hundreds or thousands of promoter sequences
  - Measure how strongly TF binds to each of the promoter sequences
  - Collect the set to which TF binds strongly, do motif finding on these



- Gene expression data
  - Collect set of genes with similar expression (activity) profiles and do motif finding on these.