Regulatory Motif Finding

Lectures 12 – Nov 7, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

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

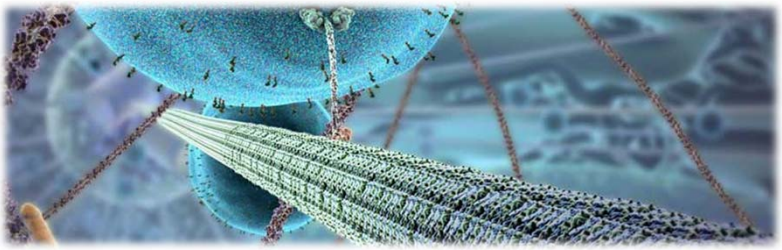
1

Outline

- Biology background
- Computational problem
  - Input data
  - Motif representation
- Common methods
  - Enumeration
  - Expectation-Maximization (EM) algorithm
  - Gibbs sampling methods

2

Cell = Factory, Proteins = Machines

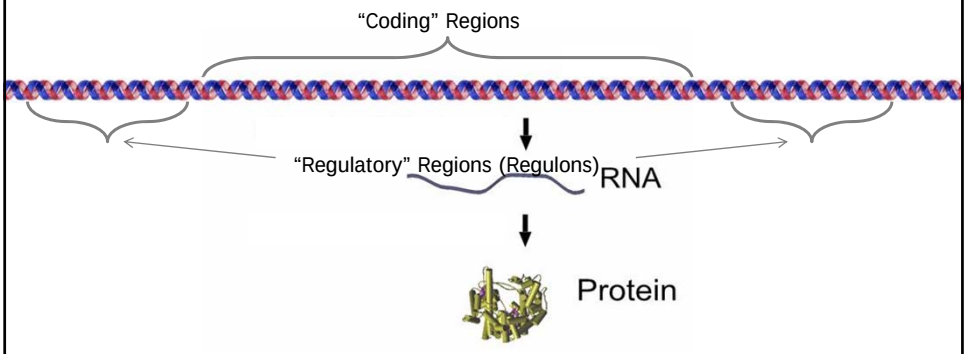


Biovisions, Harvard

3

DNA

- Instructions for making the machines



- Instructions for when and where to make them

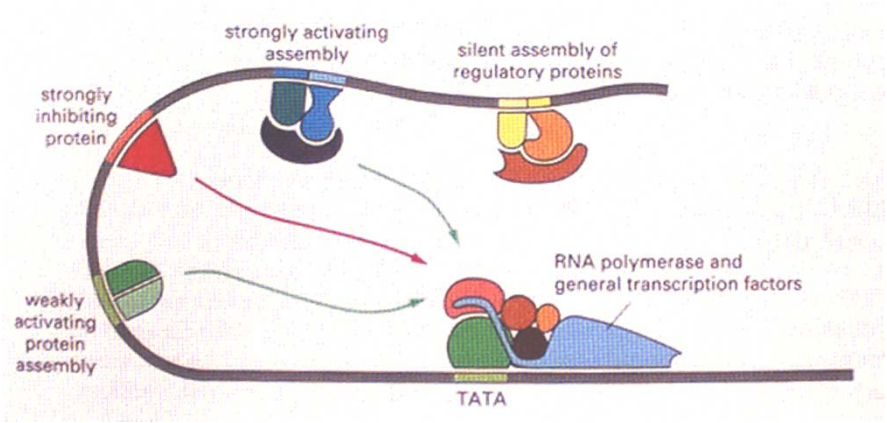
4

Transcriptional Regulation

- Regulatory regions are comprised of “binding sites”
- “Binding sites” attract a special class of proteins, known as “transcription factors”
- A TFBS can be located anywhere within the regulatory region (promoter region)
- Bound transcription factors can also inhibit DNA transcription
  - More realistic picture?

5

DNA Regulation

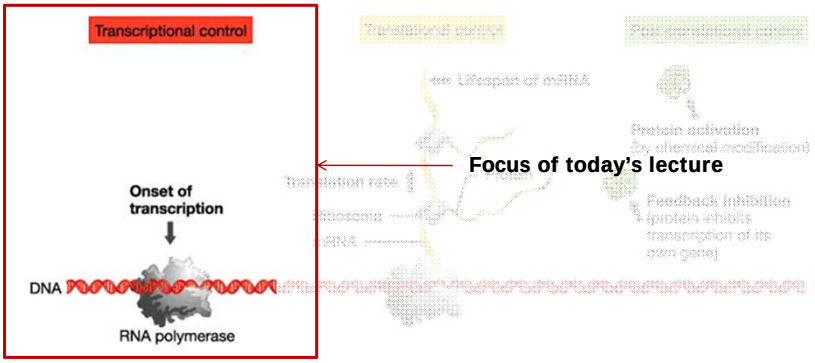


Source: Richardson, University College London

6

Gene Regulation

- Transcriptional regulation is one of many regulatory mechanisms in the cell



Source: Mallery, University of Miami

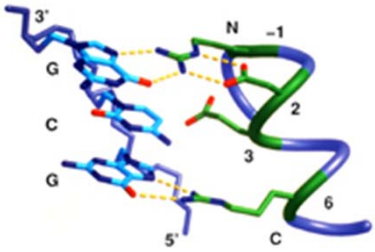
7

Transcriptional Regulation of Genes

- What turns genes on (producing a protein) and off?
- When is a gene turned on or off?
- Where (in which cells) is a gene turned on?
- How many copies of the gene product are produced?

8

Structural Basis of Interaction



9

Structural Basis of Interaction

- **Key Feature:**
  - Transcription factors are not 100% specific when binding DNA because non-essential bases could mutate
- Not one sequence, but family of sequences, with varying affinities

G A C C G	0.54
G A G C G	0.48
G A C T G	0.32
G A C C A	0.25
G G C C G	0.11
G G C T G	0.08

10

What is a motif?

- A subsequence (substring) that occurs in multiple sequences with a biological importance.
- Motifs can be totally constant or have variable elements.
- DNA motifs (regulatory elements)
  - Binding sites for proteins
  - Short sequences (5-25)
  - Up to 1000 bp (or farther) from gene
  - Inexactly repeating patterns

11

Motif Finding

- **Basic Objective:**
  - Find regions in the genome that transcription factors bind to
- Motivations
  - Understanding which TFs regulate which genes
  - Major part of the gene regulation
- Many classes of algorithms, differ in:
  - Types of input data 
  - Motif representation

12

Input Data

- Single sequence

AGCATCAG**CAGCA**CATCATCAGCATACGACTCAGCATAGCCATGGGCTA**CAGCA**GATCGATCGAA**CAGCA**CGGCAGT
AGTCGGGATGCGGATC**CAGCA**GGGAGGGAGCGCGACGCTCTATAGAGGAGGACTTAGCAGAGCGATCGACGATTACG
CAGCAGTACGCAGCAAAAAAAAAACGACGTACGTA**CAGCA**CTGACATCGGACATCTGATCTGTAGCTAGCTACTACACT
CATGACTCAGTCAGTACCGAT**CAGCA**CAGCAGCTACATGCATGCATGCAGTCACGTAGAG...

  - Based on over-representation of short sequences

13

Random Sample

atgaccgggatactgataccgtatttggcctaggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaactttccgaatactgggcataaggtaca

tgagtatccctgggatgacttttgggaacactatagtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgaccttgtaagtgttttccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatggcccacttagtccacttatag

gtcaatcatgttcttgtgaatggatttttaactgagggcatagaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtactgatggaaactttcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttggtttcgaaaatgctctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatttcaacgtatgccgaaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttctgggtactgatagca

14

Implanting Motif AAAAAAAGGGGGGG

atgaccgggatactgatAAAAAAAGGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaactttccgaataAAAAAAAGGGGGGGa

tgagtatccctgggatgacttAAAAAAAGGGGGGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAGGGGGGGcttatag

gtcaatcatgttcttgtgaatggatttAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAGGGGGGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAGGGGGGGa

Where is the Implanted Motif?

atgaccgggatactgataaaaaaaaggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaactttccgaataaaaaaaaaggggggga

tgagtatccctgggatgacttaaaaaaaaggggggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgaaaaaaaaggggggtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaataaaaaaaaggggggcttatag

gtcaatcatgttcttgtgaatggatttaaaaaaaaggggggggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtaaaaaaaaggggggcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttaaaaaaaaggggggctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaaggggggggaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttaaaaaaaaggggggga

Implanting Motif **AAAAAAGGGGGGG** with Four Mutations – (15,4)-motif

atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaactttccgaatacAATtAAAAcGGcGGGa

tgagtatccctgggatgacttAAAAtAATGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag

gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa

Where is the Motif???

atgaccgggatactgatagaagaaaggttgggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaactttccgaatacaataaaacggcggga

tgagtatccctgggatgacttaaaataatggagtggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcaaaaaaagggattgtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatataataaaggaagggcttatag

gtcaatcatgttcttgtgaatggatttaacaataagggctgggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

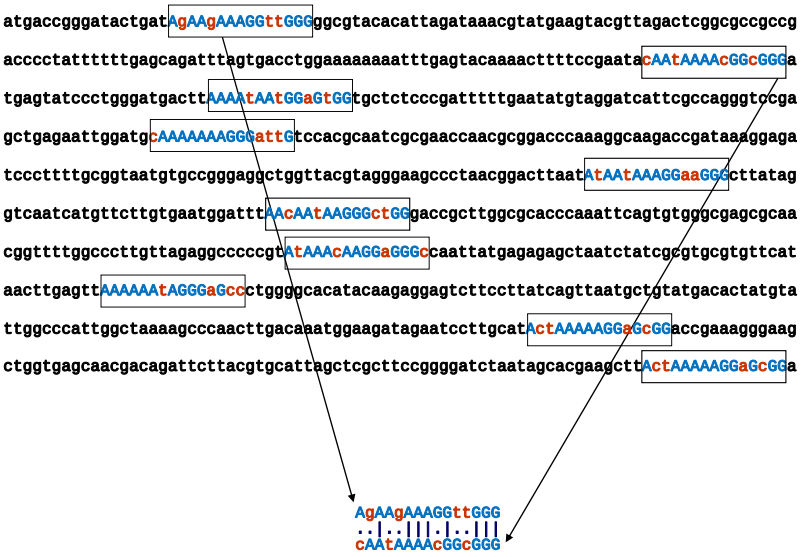
cggttttggcccttgttagaggcccccgtataaacaaggagggccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttaaaaaatagggagccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatactaaaaaggagcggaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttactaaaaaggagcgga

Why Finding (15,4) Motif is Difficult?



Challenge Problem

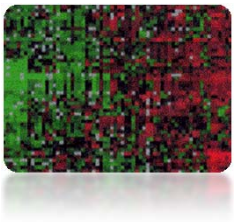
- Find a motif in a sample of
  - 20 “random” sequences (e.g. 600 nt long)
  - each sequence containing an implanted pattern of length 15,
  - each pattern appearing with 4 mismatches as (15,4)-motif.

Input Data

- Single sequence

... AGCATCAG**CAGCA**CATCATCAGCATACGACTCAGCATAGCCATGGGCTA**CAGCA**GATCGATCGAA**CAGCA**CG...

- Sequence + other data
  - Gene expression data
  - ChIP-chip
  - Others...



21

Identifying Motifs

- Genes are turned on or off by regulatory proteins (TFs).
- TFs bind to upstream regulatory regions of genes to either attract or block an RNA polymerase
- So, multiple genes that are regulated by the same TF will have the same motifs in their regulatory regions.
- How do we identify the genes that are regulated by the same TF?

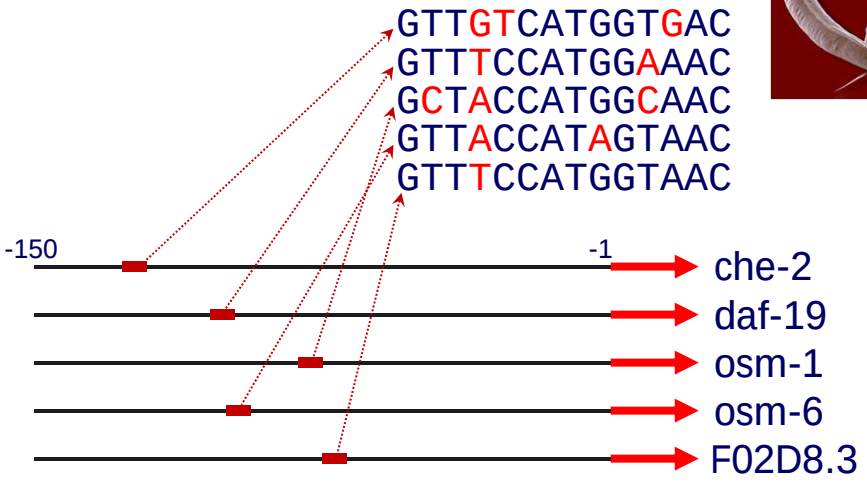
Sequence + Gene Expression Data

- Say that a microarray experiment showed that when gene X is knocked out, 20 other genes are not expressed.
  - How can one gene have such drastic effects?
- Say that 5 different genes are co-expressed across many experiments in a gene expression data.
  - These genes are likely to share the same binding sites.

23

daf-19 Binding Sites in *C. elegans*

- Motifs and transcriptional start sites



source: Peter Swoboda

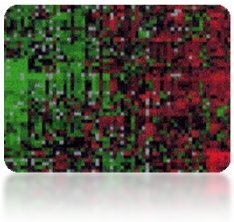
24

Input Data

- Single sequence

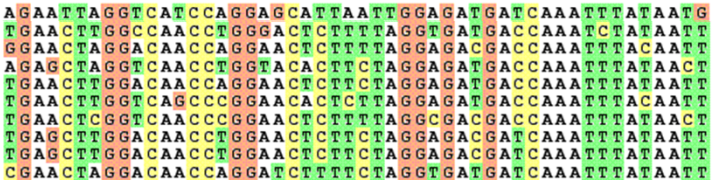
... AGCATCAG**CAGCA**CATCATCAGCATACGACTCAGCATAGCCATGGGCTA**CAGCA**GATCGATCGAA**CAGCA**CG...

- Sequence + other data
  - Gene expression data
  - ChIP-chip
  - Others…



- Evolutionarily related set of sequences

species ↓



25

Motif Finding

- **Basic Objective:**
  - Find regions in the genome that bind transcription factors
- Motivations
  - Understanding which TFs regulate which genes
  - Major part of the gene regulation
- Many classes of algorithms, differ in:
  - Types of input data
  - Motif representation ⇦

26

Structural Basis of Interaction

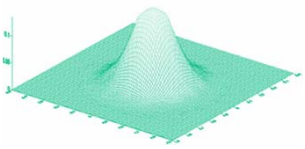
- **Key Feature:**
  - Transcription factors are not 100% specific when binding DNA
- Not one sequence, but family of sequences, with varying affinities

G A C C G	0.54
G A G C G	0.48
G A C T G	0.32
G A C C A	0.25
G G C C G	0.11
G G C T G	0.08

27

Motif Representation

- Structural discussion immediately raises difficulties
- Least expressive: G A C C G
  - Single sequence



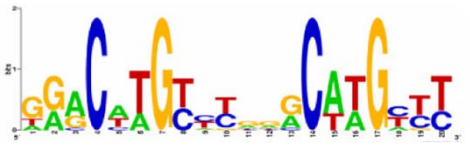
- Most expressive:
  - 4^k-dimensional probability distribution
  - Independently assign probability for each of the possible k-mers*

*k-mer refers to a specific n-tuple of nucleic acid or amino acid sequences that can be used to identify certain regions within DNA or protein.

28

Motif Representation

- Standard Solution:
  - Position-Specific Scoring Matrix (PSSM)
  - Assuming independence of positions, assign a probability distribution for each position

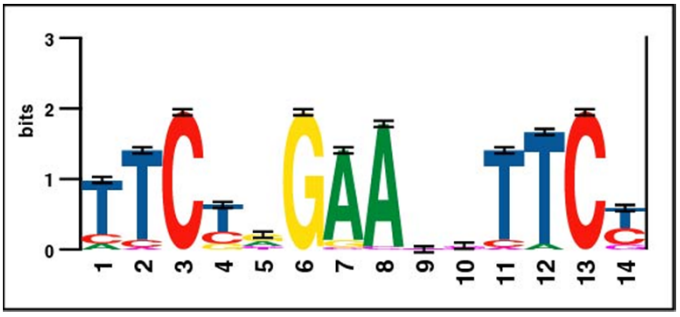


  - This is a too simple representation

29

Position Weight Matrix (PWM)

- Assign probability to (A,G,C,T) in each position

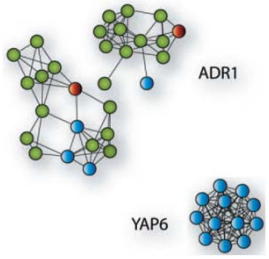


G	0.1	0.01	0	0.1	0.25	1	0.1	0	...
A	0.2	0.99	0	0	0.25	0	0.9	1	...
T	0.6	0.7	0	0.5	0.25	0	0	0	...
C	0.3	0.2	1	0.3	0.25	0	0	0	...

30

Oversimplicity of PSSMs

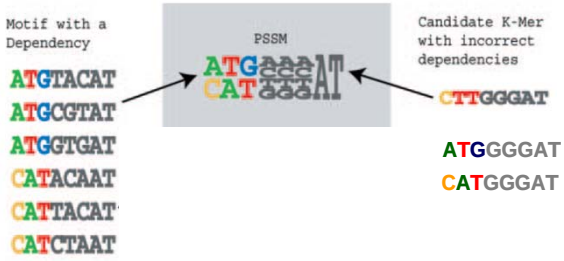
- Assumes independence between positions
- ~25% of TRANSFAC motifs have been shown to violate this assumption
  - Two Examples: ADR1 and YAP6



31

Oversimplicity of PSSMs

- Assumes independence between positions
- Generates potentially unseen motifs



- We need to model dependency between positions.
  - Revisit later

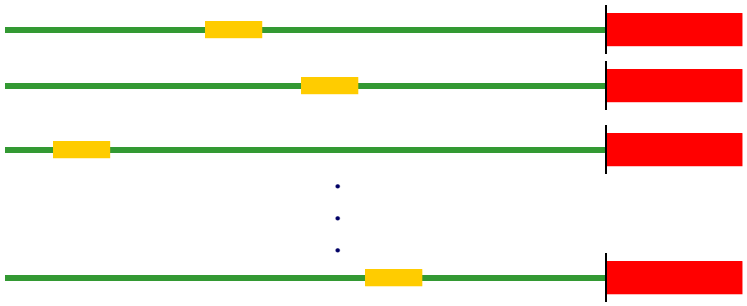
Outline

- Biology background
- Computational problem
  - Input data
  - Motif representation
- Common methods
  - Enumeration
  - Expectation-Maximization (EM) algorithm
  - Gibbs sampling methods

33

Finding Regulatory Motifs

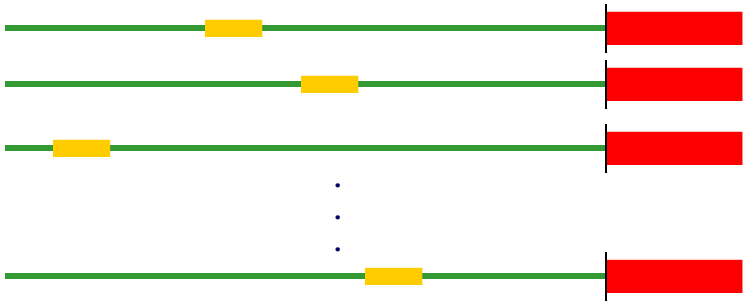
- Given a collection of genes that are likely to be regulated by the same TFs,
- Find the TF-binding motifs in common



34

Identifying Motifs: Complications

- We do not know the motif sequence
- We do not know where it is located relative to the genes start
- Motifs can differ slightly from one gene to another
- How to discern it from “random” motifs?



35

Common Methods

- Problem statement:
  - Given a set of n promoters of n co-regulated genes, find a motif common to the promoters
  - Both the PWM (defined in page 30) and the motif sequences are unknown.
- Enumeration (simplest method)
  - Look at the frequency of all k-mers*
- EM algorithm (MEME)
  - Iteratively hone in on the most likely motif model
- Gibbs sampling methods (AlignAce, BioProspector)

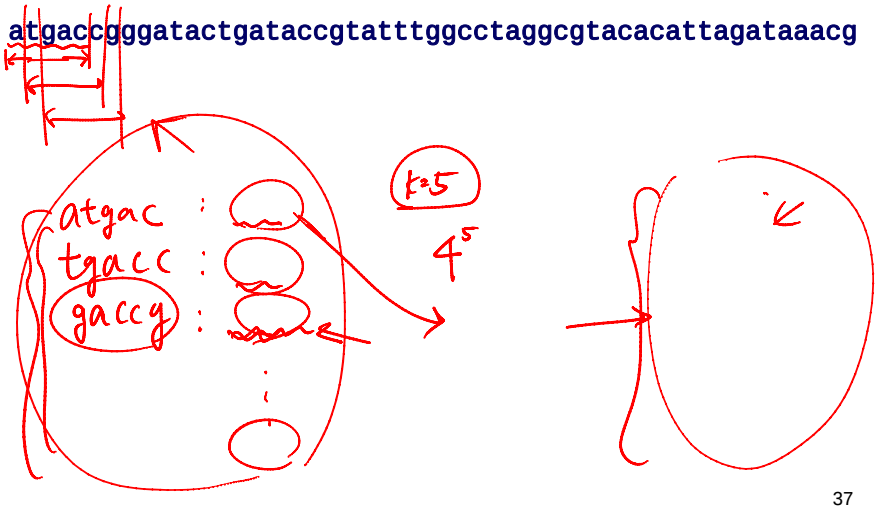
*k-mer refers to a specific n-tuple of nucleic acid that can be used to identify certain regions within DNA or proteins.

36

Generating k-mers

- Example:

```
atgaccgggatactgataccgtatttggcctaggcgtacacattagataaacg
```



Motif Finding Using EM Algorithm

- MEME (Multiple EM for Motif Elucidation)
  - Bailey and Elkan (1995), Bailey et al. (2006)
  - http://meme.sdsc.edu/meme/intro.html
- Expectation-Maximization
  - In each iteration, it learns the PWM model and identifies examples of the matrix (sites in the input sequences)

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
  - 2. Identify examples of the model
    - For every k-mer in the input sequences, identify its probability given the PWM model.
  - 3. Re-estimate the motif model
    - Calculate a new PWM, based on the weighted frequencies of all k-mers in the input sequences
  - 4. Iteratively refine the PWMs and identify sites until convergence.

Example: MEME

- Find a 6-mer motif in 4 sequences

  S_1: GGCTATTGCAGATGACGAGATGAGGCCCAGACC

  S_2: GGATGACAATTATATAAAGGACGATAAGAGATGAC

  S_3: CTAGCTCGTAGCTCGTTGAGATGCGCTCCCCGCTC

  S_4: GATGACGGAGTATTAAAGACTCGATGAGTTATACGA

  - 1. MEME uses an initial EM heuristic to estimate the best starting-point PWM matrix:

G	0.26	0.24	0.18	0.26	0.25	0.26
A	0.24	0.26	0.28	0.24	0.25	0.22
T	0.25	0.23	0.30	0.25	0.25	0.25
C	0.25	0.27	0.24	0.25	0.25	0.27

40

- 2. MEME scores the match of all 6-mers to current matrix

gGCTATTGCATATGACGAGATGAGGCCCAGACC

GGATGACAATTATATAAAGGACCGTGATAAGAGATTAC

CTAGCTCGTAGCTCGTTGAGATGCGCTCCCCGCTC

GATGACGGAGTATTAAAGACTCGATGAGTTATACGA

Here, just consider the underlined 6-mers, Although in reality all 6-mers are scored

- 3. Re-estimate the PWM based on the weighted contribution of all 6-mers.

The height of the bases above corresponds to how much that 6-mer counts in calculating the new matrix

G	0.29	0.24	0.17	0.27	0.24	0.30
A	0.22	0.26	0.27	0.22	0.28	0.18
T	0.24	0.23	0.33	0.23	0.24	0.28
C	0.24	0.27	0.23	0.28	0.24	0.24

41

- 4. MEME scores the match of all 6-mers to current matrix

gGCTATTGCATATGACGAGATGAGGCCCAGACC

GGATGACAATTATATAAAGGACCGTGATAAGAGATTAC

CTAGCTCGTAGCTCGTTGAGATGCGCTCCCCGCTC

GATGACGGAGTATTAAAGACTCGATGAGTTATACGA

- 5. Re-estimate the PWM based on the weighted contribution of all 6-mers.

The height of the bases above corresponds to how much that 6-mer counts in calculating the new matrix

G	0.40	0.20	0.15	0.42	0.24	0.30
A	0.30	0.30	0.20	0.24	0.46	0.18
T	0.15	0.30	0.45	0.16	0.15	0.28
C	0.15	0.20	0.20	0.16	0.15	0.24

42

- 6. MEME scores the match of all 6-mers to current matrix

  gGCTATTGCATATGACGAGATGAGGcccagaCC

  GGATGACTTATATAAAGGACCGTGATAAGAGATTAC

  CTAGCTCGTAGCTCGTTGAGATGCGCTccccgctC

  GATGACGGAGTATTAAAGACTCGATGAGTTATACGA

- Iterations continue until convergence
  - Numbers do not change much between iterations

Final motif

G	0.85	0.05	0.10	0.80	0.20	0.35
A	0.05	0.60	0.10	0.05	0.60	0.10
T	0.05	0.30	0.70	0.05	0.20	0.10
C	0.05	0.05	0.10	0.10	0.10	0.35

43

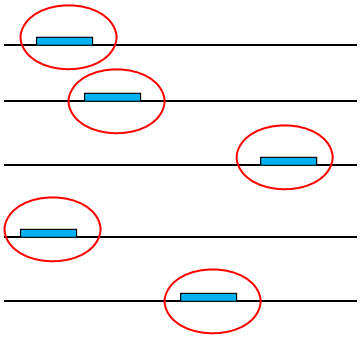
Gibbs Sampling

- References
  - AlignAce by Hughes et al. 2000
    http://atlas.med.harvard.edu/download/index.html,
  - BioProspector by Liu et al. 2001
    http://motif.stanford.edu/distributions/r

- Procedure
  - 1. Start by randomly choosing sites and creates an initial PWM matrix
  - 2. Sample other sites
    - Remove some set of matrix examples (sites)
    - Randomly choose other sites and calculate P given matrix
    - If they have a high score to the matrix, keep the new site
  - 3. Iterate until convergence

44

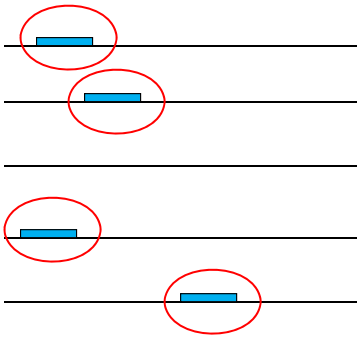
Gibbs Sampling: Basic Idea



Current motif = PWM formed by circled substrings

Slides generously and unknowingly provided by S. Sinha, Urbana-Chamaign CS Dept.

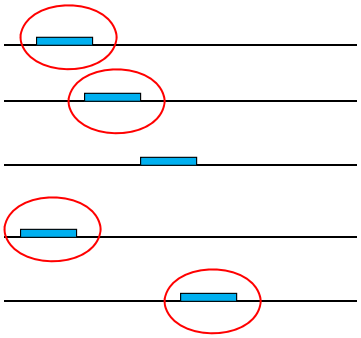
Gibbs Sampling: Basic Idea



Delete one substring

Slides generously and unknowingly provided by S. Sinha, Urbana-Chamaign CS Dept.

Gibbs sampling: Basic Idea



Try a replacement:
Compute its score,
Accept the replacement
depending on the score.

Slides generously and unknowingly provided by S. Sinha, Urbana-Chamaign CS Dept.

Outline

- Biology background
- Computational problem
  - Input data
  - Motif representation
- Common methods
  - Enumeration
  - Expectation-Maximization (EM) algorithm
  - Gibbs sampling methods

48