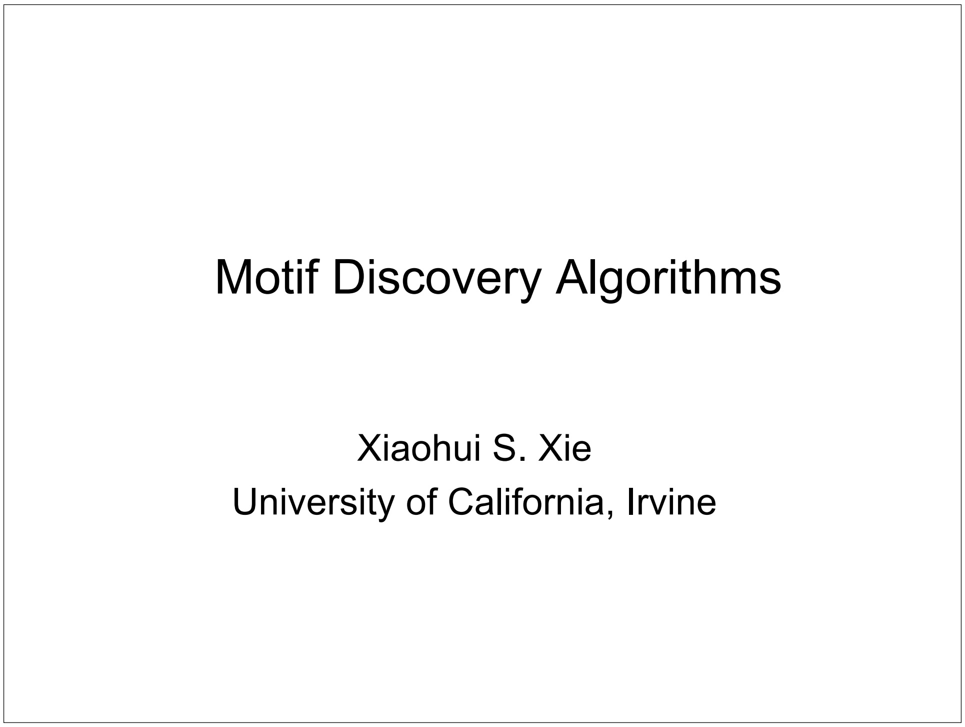
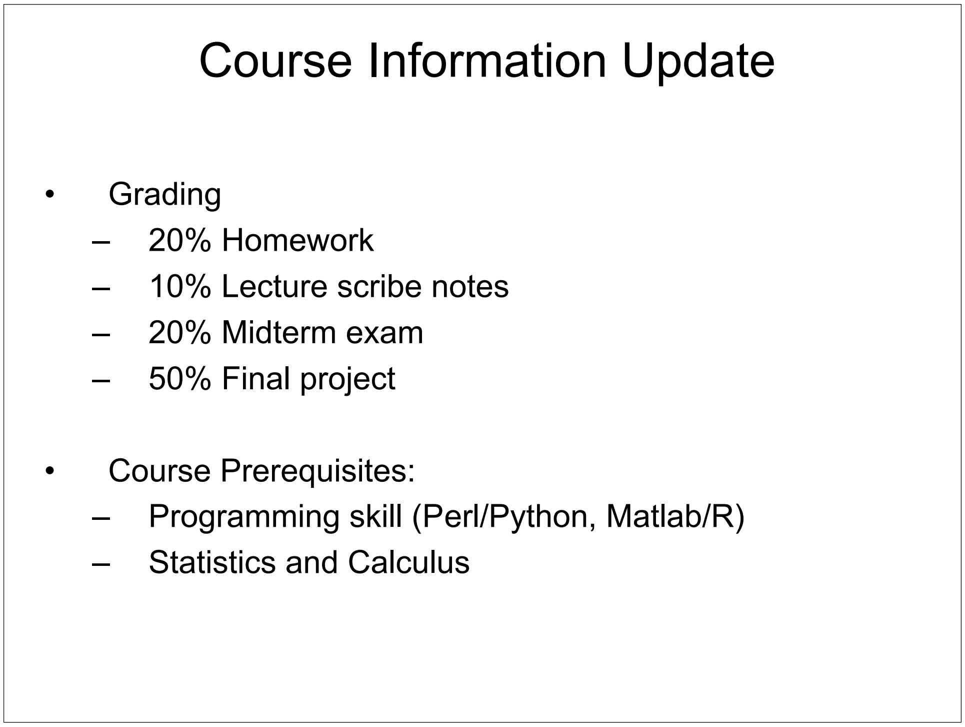


Motif Discovery Algorithms

Xiaohui S. Xie
University of California, Irvine



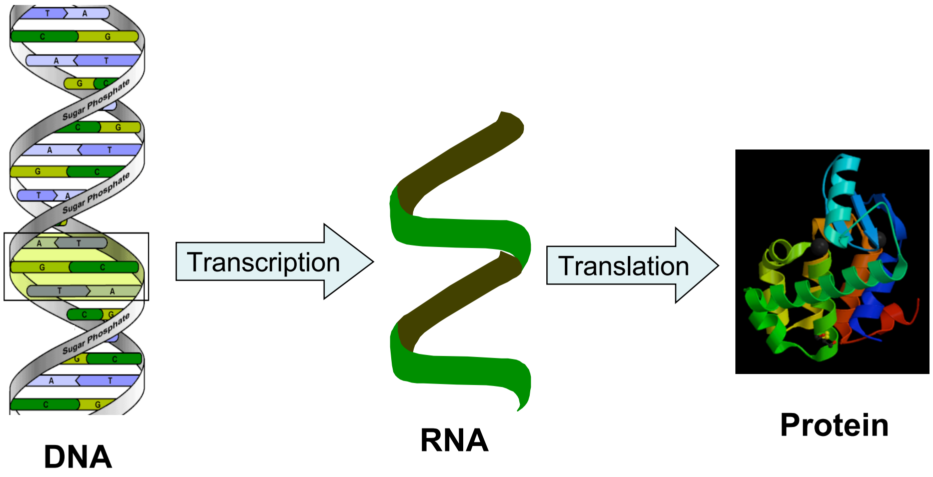
Course Information Update

- Grading
  - 20% Homework
  - 10% Lecture scribe notes
  - 20% Midterm exam
  - 50% Final project
- Course Prerequisites:
  - Programming skill (Perl/Python, Matlab/R)
  - Statistics and Calculus

Today's Goals

- Gene regulation
- Motif discovery algorithms
  - Enumeration
  - Statistical significance
  - Expectation-Maximization
  - Gibbs Sampling

The Central Dogma



DNA

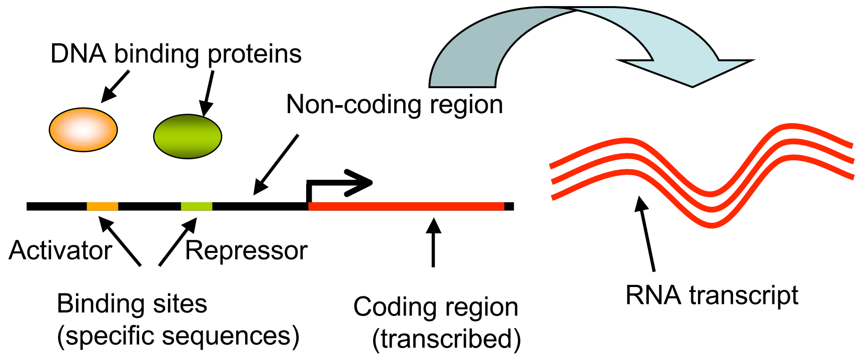
Transcription

RNA

Translation

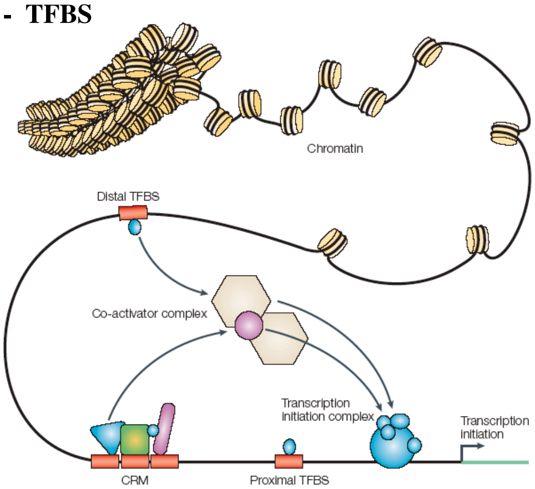
Protein

Transcriptional Regulation



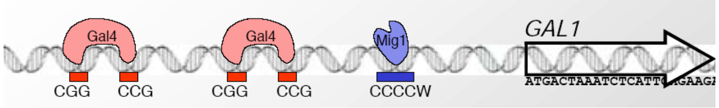
Regulation in Eukaryotes

- •Promotor
- •Transcription Factors - TF
- •Transcription Factor binding Sites - TFBS
- •Cis-regulatory modules - CRM
- •Transcription Start Site - TSS
- •TATA boxes
- •CG richness
- •Phylogenetic Footprinting
- •Combinatorial Interaction
- •Enhancers



Wasserman and Sandelin (2004) 'Applied Bioinformatics for the Identification of Regulatory Elements" Nature Review Genetics 5.4.276

Regulatory motifs



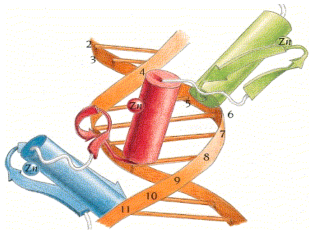
- Motifs are fundamental units of gene regulation:
  - What turns genes on (producing a protein) and off?
  - When is a gene turned on or off?
  - Where (in which cells) is a gene turned on?
  - How many copies of the gene product are produced?
- Specialized proteins (transcription factors) recognize these motifs

What we know about regulatory motifs:

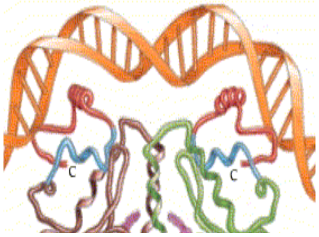
- Motifs are short (6-20 bp), sometimes degenerate
- Can contain any set of nucleotides (no ATG or other rules)
- Act at variable distances upstream (or downstream) from target gene (could be 100 Kb upstream or downstream)
- Human genome contains roughly 2000 motifs

Transcription Factor Binding Sites

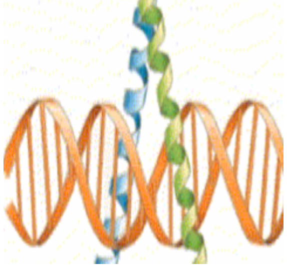
- Gene regulatory proteins contain structural elements that can “read” DNA sequence “motifs”
- The amino acid – DNA recognition is not straightforward
- Experiments can pinpoint binding sites on DNA



Zinc finger



Helix-Turn-Helix



Leucine zipper

Regulatory motif discovery

```
CAAACTCCTGCACGTGTCTCAAGGAATTTCCCGCCTCTGTCTTCTGAGTT
GGCTACAGATGTGTACCACGCACGTGGAACCCAGCTGATTTCCCACCTTT
TTATCACGTGGAGCAAACGATTAGGGAGAATTAATTATTCTCTTCCTCTT
AGGAAATGATGTTTACCCTAACCCAAAATGTAAGACACGTGATTTATCAG
ACTACCTATAAAGAAGACACGTGAATCTGTCCTGCTTGGTGGTGTAAGGA
TTGGATTTGGTCCCACGTGATGTCAGAAGATTGCGGACCAAATCCCACTA
GCACACGTGGGGTCATTTGGAGAAAGATACTTTGTAACATTGGACCTCTG
CATCTGTAAAACACGTGTGGGAATAGTAAGAATAATAATACTTGTCTCAC
ATGTGAAGGTAAAATGAGGTCATGCACGTGTGTGCACAGAATCTAGTCCA
AGAACATACCTGGCACTCAATTAATATGAGATAATTGTGCCATGCCTTAA
GTATAAGATTTGTTATTACCGCACGTGTAAACACTACAGCATGAATTTGC
ACTGCCAAACACGTGTGGAGGTTTAAGTTCTGATTCCTGATGATGAAATA
CTCTGGCCTGCTACGTTAACACGTGAAACAGCACTGATGGTAAAGGCTAA
TTGGATTTGGTCCCACGTGATGTCAGAAGATTGCGGACCAAATCCCACTA
ACTACCTATAAAGAAGACACGTGAATCTGTCCTGCTTGGTGGTGTAAGGA
```

Promoter sequences for 15 genes

Method 1: Enumeration

List all potential motifs with a given length

For instance, 6-mer motifs

```
AAAAAA   0
AAAAAC   1
AAAAAG   2
AAAAAT   1
…
CACGTG   15
…
TTTTTT   1
TTTTTT   0
```

Total: 4^6=4096 6-mers

Regulatory motif discovery

```
CAAACTCCTGCACGTGTCTCAAGGAATTTCCCGCCTCTGTCTTCTGAGTT
GGCTACAGATGTGTACCACGCACGTGGAACCCAGCTGATTTCCCACCTTT
TTATCACGTGGAGCAAACGATTAGGGAGAATTAATTATTCTCTTCCTCTT
AGGAAATGATGTTTACCCTAACCCAAAATGTAAGACACGTGATTTATCAG
ACTACCTATAAAGAAGACACGTGAATCTGTCCTGCTTGGTGGTGTAAGGA
TTGGATTTGGTCCCACGTGATGTCAGAAGATTGCGGACCAAATCCCACTA
GCACACGTGGGGTCATTTGGAGAAAGATACTTTGTAACATTGGACCTCTG
CATCTGTAAAACACGTGTGGGAATAGTAAGAATAATAATACTTGTCTCAC
ATGTGAAGGTAAAATGAGGTCATGCACGTGTGTGCACAGAATCTAGTCCA
AGAACATACCTGGCACTCAATTAATATGAGATAATTGTGCCATGCCTTAA
GTATAAGATTTGTTATTACCGCACGTGTAAACACTACAGCATGAATTTGC
ACTGCCAAACACGTGTGGAGGTTTAAGTTCTGATTCCTGATGATGAAATA
CTCTGGCCTGCTACGTTAACACGTGAAACAGCACTGATGGTAAAGGCTAA
TTGGATTTGGTCCCACGTGATGTCAGAAGATTGCGGACCAAATCCCACTA
ACTACCTATAAAGAAGACACGTGAATCTGTCCTGCTTGGTGGTGTAAGGA
```

Promoter sequences for 15 genes

How to measure significance?

Suppose we observe that among the *n* promoter sequences, the motif occurs in k of them.

How surprise is the observation?

1. Curate a set of control sequences (total number: *N*) that the motif is not enriched
2. Count the number of sequences that contain the motif (*K*)

Representation of motifs, PWM

Site 1	A	G	A	T	G	G	A	T	G	G
Site 2	T	G	A	T	T	G	A	T	G	T
Site 3	T	G	A	T	G	G	A	T	G	G
Site 4	A	G	A	T	T	G	A	T	C	G
Site 5	T	G	A	T	G	G	A	T	T	G
Site 6	T	G	A	T	G	G	A	T	T	G
Site 7	A	G	A	T	G	G	A	T	T	G

PWM represents frequencies of each base at each position in the motif *

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0

* These days, PWM/PSSM can correspond to the frequency matrix or a likelihood matrix

Positional weight matrix representation

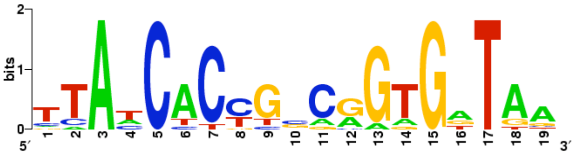
1 GTATCACCGCCAGTGGTAT
2 ATACCACTGGCGGTGATAC
3 TCAACACCGCCAGAGATAA
4 TTATCTCTGGCGGTGTTGA
5 TTATCACCGCAGATGGTTA
6 TAACCATCTGCGGTGATAA
7 CTATCACCGCAAGGGATAA
8 TTATCCCTTGCGGTGATAG
9 CTAACACCGTGCGTGTTGA
10 TCAACACGCACGGTGTTAG
11 TTACCTCTGGCGGTGATAA
12 TTATCACCGCCAGAGGTAA

Lambda cI/cro binding sites

W_{ij}

A:	9	9	94	25	1	71	1	1	1	9	17	32	9	17	1	48	1	71	63
C:	17	17	1	25	94	9	86	55	9	40	71	9	1	1	1	1	1	1	9
G:	9	1	1	1	1	1	1	9	71	40	9	55	86	9	94	25	1	17	17
T:	63	71	1	48	1	17	9	32	17	9	1	1	1	71	1	25	94	9	9

Sequence Logo



Information content

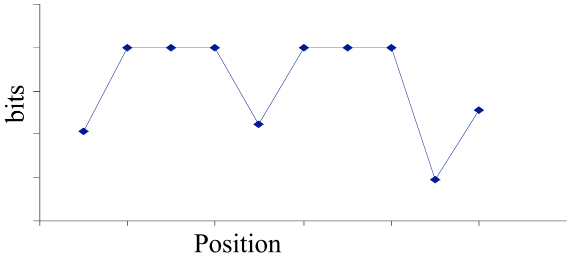
The least variable positions likely are important for specifying the protein-DNA interaction

Therefore high information content = low sequence variation at that position.

G	0	1.0	0	0	0.7	1.0	0	0	0.4	0.8
A	0.4	0	1.0	0	0	0	1.0	0	0	0
T	0.6	0	0	1.0	0.3	0	0	1.0	0.4	0.2
C	0	0	0	0	0	0	0	0	0.2	0
IC	1.0	2.0	2.0	2.0	1.1	2.0	2.0	2.0	0.5	1.3

= bit score of 15.9

Information Profile:



Weight matrix, sequence logos

Corrected probabilities of observing a given nucleotide can be calculated using equation 1.

Corrected probability calculation:

$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in \{A,C,G,T\}} s(b')} \quad (1)$$

$f_{b,i}$ = counts of base b in position i; N = number of sites; $p(b,i)$ = corrected probability of base b in position i; $s(b)$ = pseudocount function

A position weight matrix (PWM) is constructed by dividing the nucleotide probabilities in (1) by expected background probabilities and converting the values to a log-scale (see equation 2).

PWM conversion:

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)} \quad (2)$$

$p(b)$ = background probability of base b; $p(b,i)$ = corrected probability of base b in position i; $W_{b,i}$ = PWM vaue of base b in position i

The quantitative PWM score for a putative site is the sum of the PWM values for each nucleotide in the site (see equation 3)

Evaluation of sequences:

$$S = \sum_{i=1}^{w} W_{l_i,i} \quad (3)$$

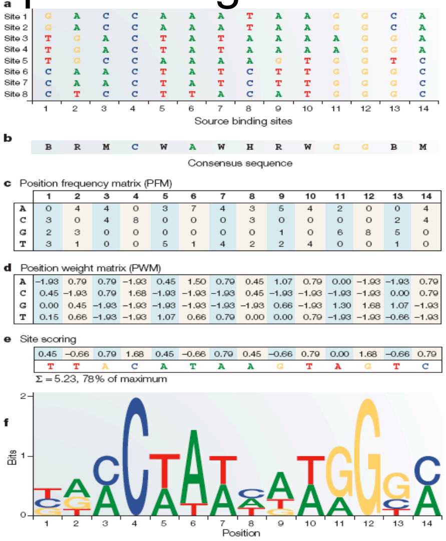
l_i = the nucleotide in position i in an input sequence; S = PWM score of a sequence; w = width of the PWM

Probability values (1) can be used to determine the total information content (in bits) in each position (see equation 4).

Information content calculation:

$$D_i = 2 + \sum_b p_{b,i} \log_2 p_{b,i} \quad (4)$$

D_i = information content in position i; $p(b,i)$ = corrected probability of base b in position i



a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

d Position weight matrix (PWM)

A	–1.93	0.79	0.79	–1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	–1.93	–1.93	0.79
C	0.45	–1.93	0.79	1.68	–1.93	–1.93	–1.93	0.45	–1.93	–1.93	–1.93	–1.93	0.00	0.79
G	0.00	0.45	–1.93	–1.93	–1.93	–1.93	–1.93	–1.93	0.66	–1.93	1.30	1.68	1.07	–1.93
T	0.15	0.66	–1.93	–1.93	1.07	0.66	0.79	0.00	0.00	0.79	–1.93	–1.93	–0.66	–1.93

e Site scoring

0.45	–0.66	0.79	1.68	0.45	–0.66	0.79	0.45	–0.66	0.79	0.00	1.68	–0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

Σ = 5.23, 78% of maximum

Very high frequency of false positives. A model for binding of MyoD will yield 10^6 binding sites, while only 10^3 might be real.

Wasserman and Sandelin (2004) 'Applied Bioinformatics for the Identification of Regulatory Elements" Nature Review Genetics 5.4.276