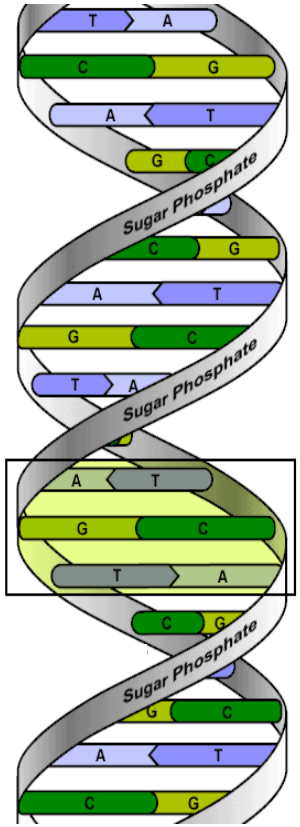
Algorithms for Regulatory Motif Discovery

Xiaohui Xie
University of California, Irvine

Today's Goals

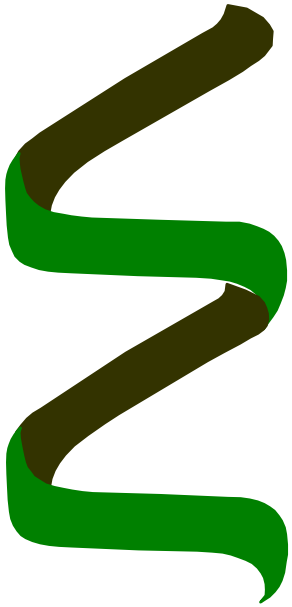
- Enumeration-based method
    - P-value
        - Binomial
        - Hypergeometric
- Probabilistic modeling
    - Position Weight Matrix
    - EM-algorithm

The Central Dogma



DNA

Transcription

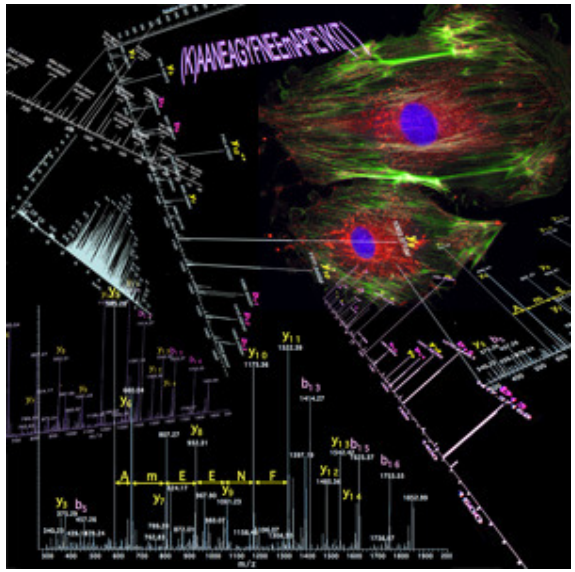


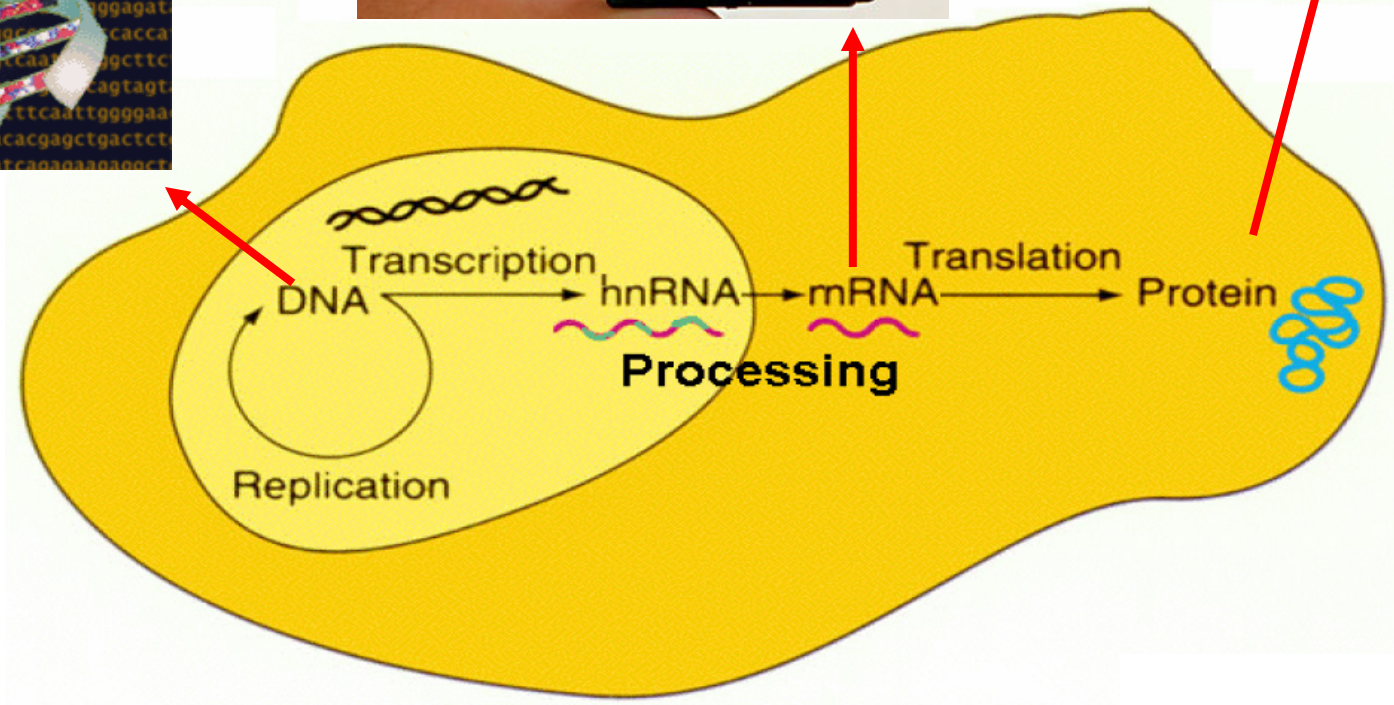
RNA

Translation

Protein

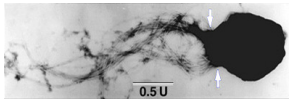
Readout from the genome

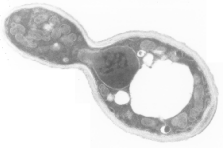
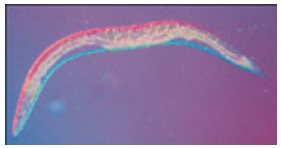




Comparison of genome size

Organisms

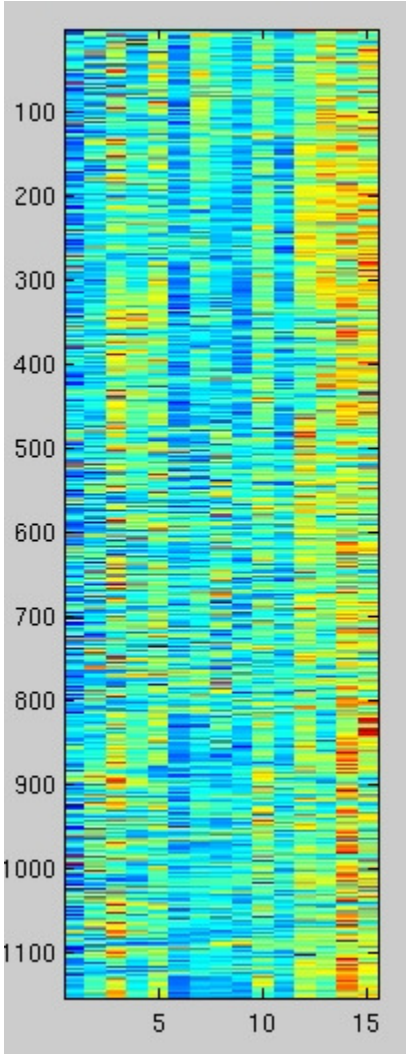



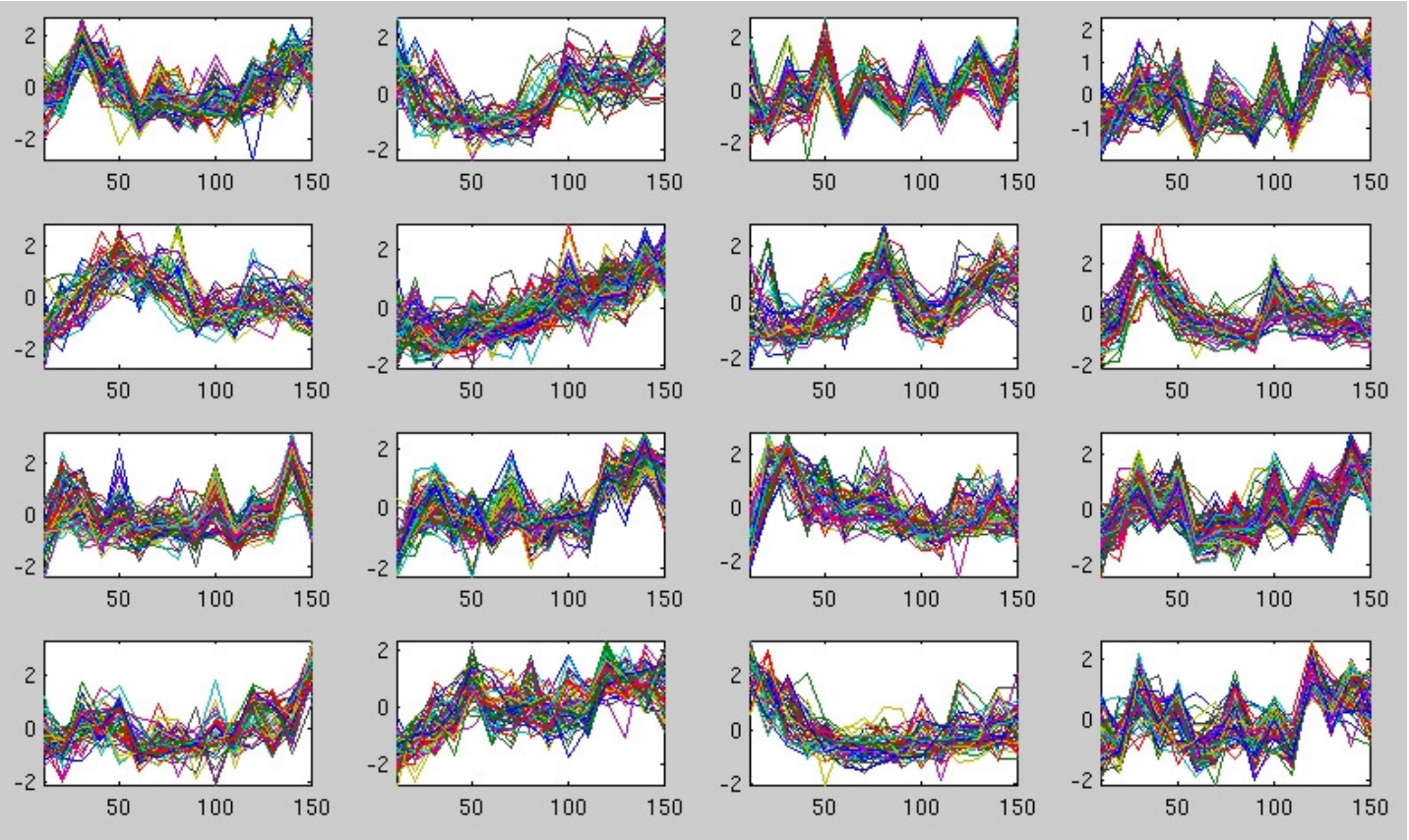




Genomes

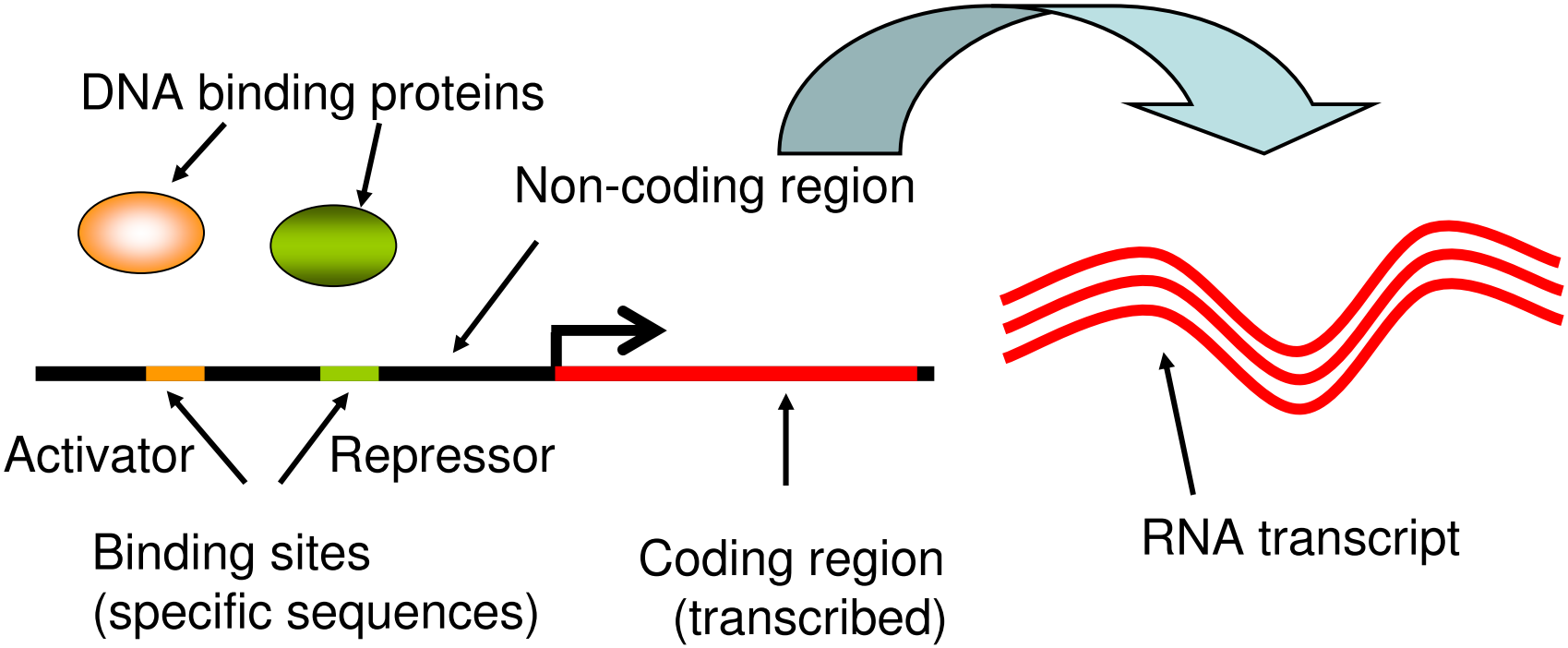
	Haemophilus influenzae	Methannococcus jannaschii	Saccharomyces cerevisiae (baker's yeast)	Caenorhabditis elegans (nematode worm)	Drosophila Melanogaster (fruit fly)	Mus musculus (laboratory mouse)	Homo sapiens (man)
Genome (MB)	1.83	1.66	13	97	180	3200	3500
Number of genes	1709	1682	6241	18,424	13,500	~30,000	~30,000

Expression of genes at different time points of cell cycle in yeast cells

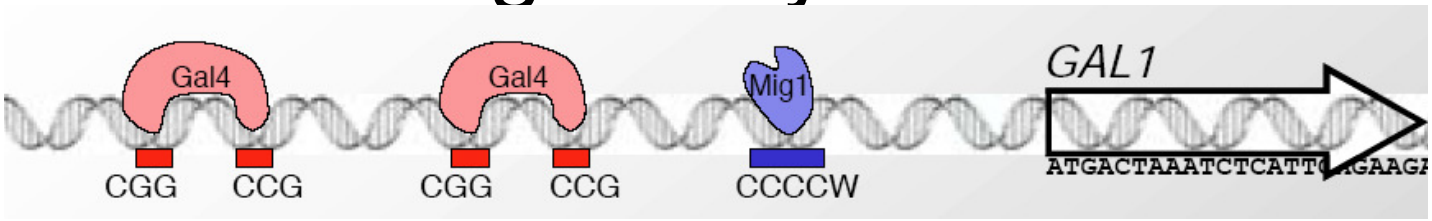




Transcriptional Regulation



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

Regulatory motif discovery

```
1  CAAACTCCTGCACGTGTCTCAAGGAATTTCCCGCCTCTGTCTTCTGAGTT
2  GGCTACAGATGTGTACCACGCACGTGGAACCCAGCTGATTTCCCACCTTT
3  TTATCACGTGGAGCAAACGATTAGGGAGAATTAATTATTCTCTTCCTCTT
4  AGGAAATGATGTTTACCCTAACCCAAAATGTAAGACACGTGATTTATCAG
5  ACTACCTATAAAGAAGACACGTGAATCTGTCCTGCTTGGTGGTGTAAGGA
6  TTGGATTTGGTCCCACGTGATGTCAGAAGATTGCGGACCAAATCCCACTA
7  GCACACGTGGGGTCATTTGGAGAAAGATACTTTGTAACATTGGACCTCTG
8  CATCTGTAAAACACGTGTGGGAATAGTAAGAATAATAATACTTGTCTCAC
9  ATGTGAAGGTAAAATGAGGTCATGCACGTGTGTGCACAGAATCTAGTCCA
10 AGAACATACCTGGCACTCAATTAATATGAGATAATTGTGCCATGCCTTAA
11 GTATAAGATTTGTTATTACCGCACGTGTAAACACTACAGCATGAATTTGC
12 ACTGCCAAACACGTGTGGAGGTTTAAGTTCTGATTCCTGATGATGAAATA
13 CTCTGGCCTGCTACGTTAACACGTGAAACAGCACTGATGGTAAAGGCTAA
14 TTGGATTTGGTCCCACGTGATGTCAGAAGATTGCGGACCAAATCCCACTA
15 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?

Definition:

- Given *n* sequences $\{S_1, S_2, \ldots, S_n\}$.
- Use $|S_i|$ to denote the length of sequence S_i
- Consider a particular *k*-mer *m* (e.g. *m*=CACGTG, *k*=6)

Assumption:

Each of the four letters (A,C,G,T) is equally likely to occur at any position of each sequence, that is

$$p(S_{ij}=A)=p(S_{ij}=C)=p(S_{ij}=G)=p(S_{ij}=T)=1/4$$

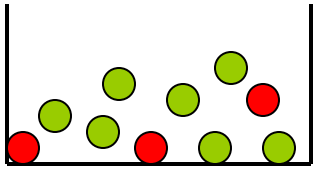
Question:

What is the likelihood of observing at least *x* sequences that contain the k-mer *m*?

Binomial distribution

Experiment 1: Sampling with replacement

A box with 3 red balls and 7 green balls:



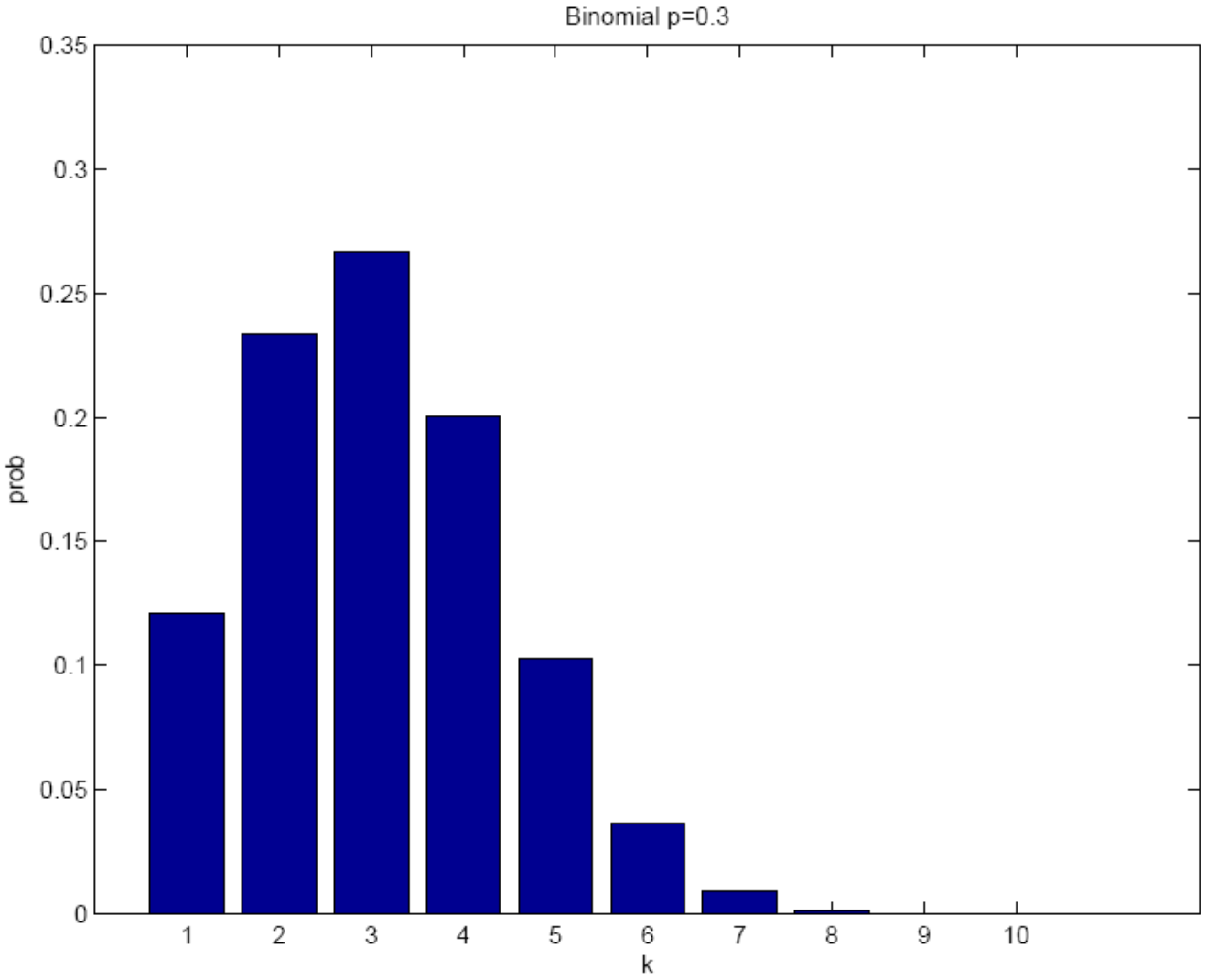
Q1: Randomly pick one ball. What's the chance that the ball is red?

p = 3/10

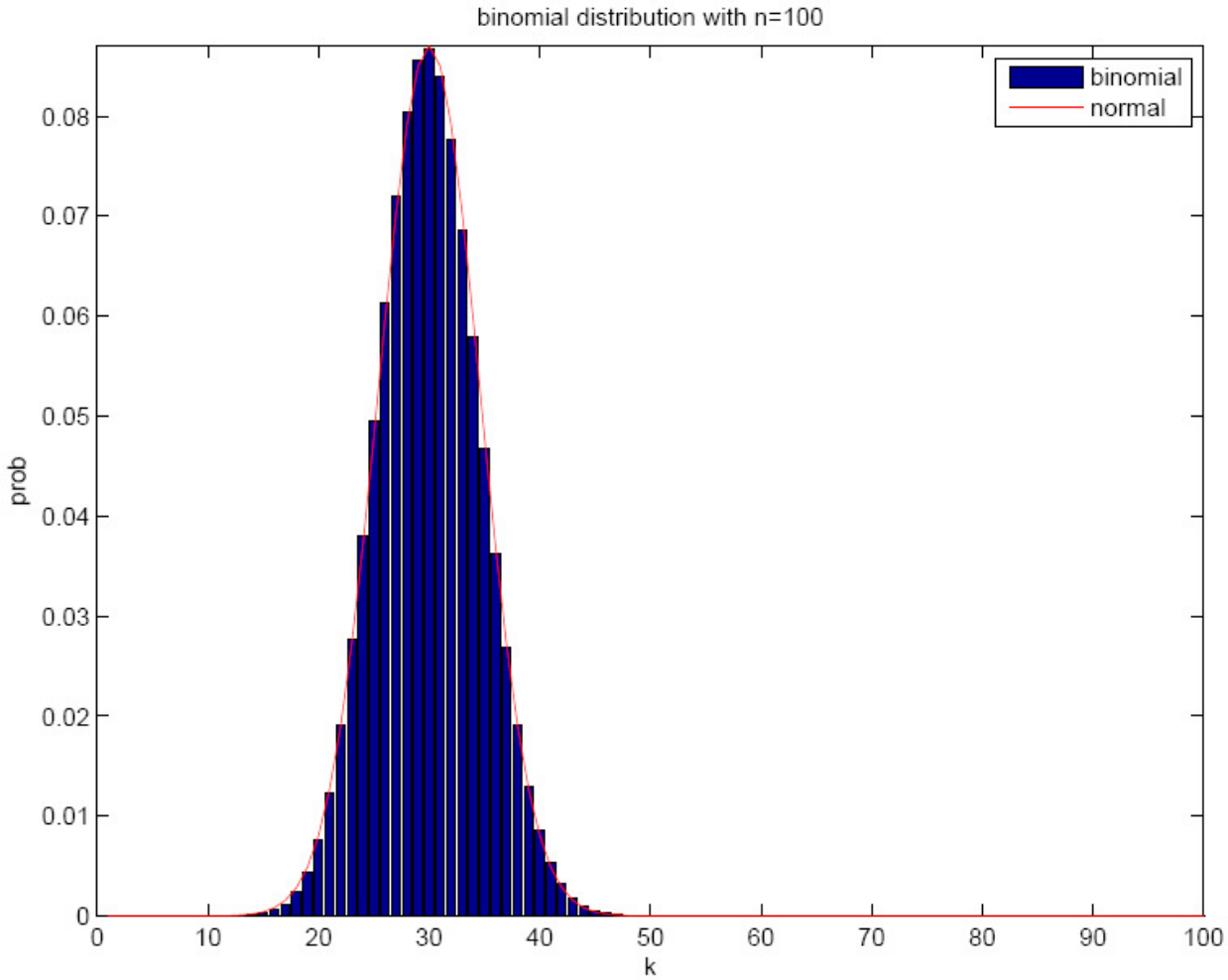
Q2: Randomly pick one ball.
Place back a ball with the same color.
Repeat *n* times.
What's the chance that the total number of red balls you pick is *k*?

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial distribution



Binomial distribution



When n is large, the binomial distribution can be approximated by the normal distribution with

Mean: ***np***

Variance: ***np(1-p)***

How to measure significance?

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

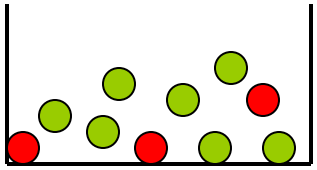
How surprise is the observation?

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*)

Hypergeometric distribution

Experiment 2: Sampling without replacement

A box with 30 red balls and 70 green balls:



Q1: Randomly pick one ball. What's the chance that the first ball is red?

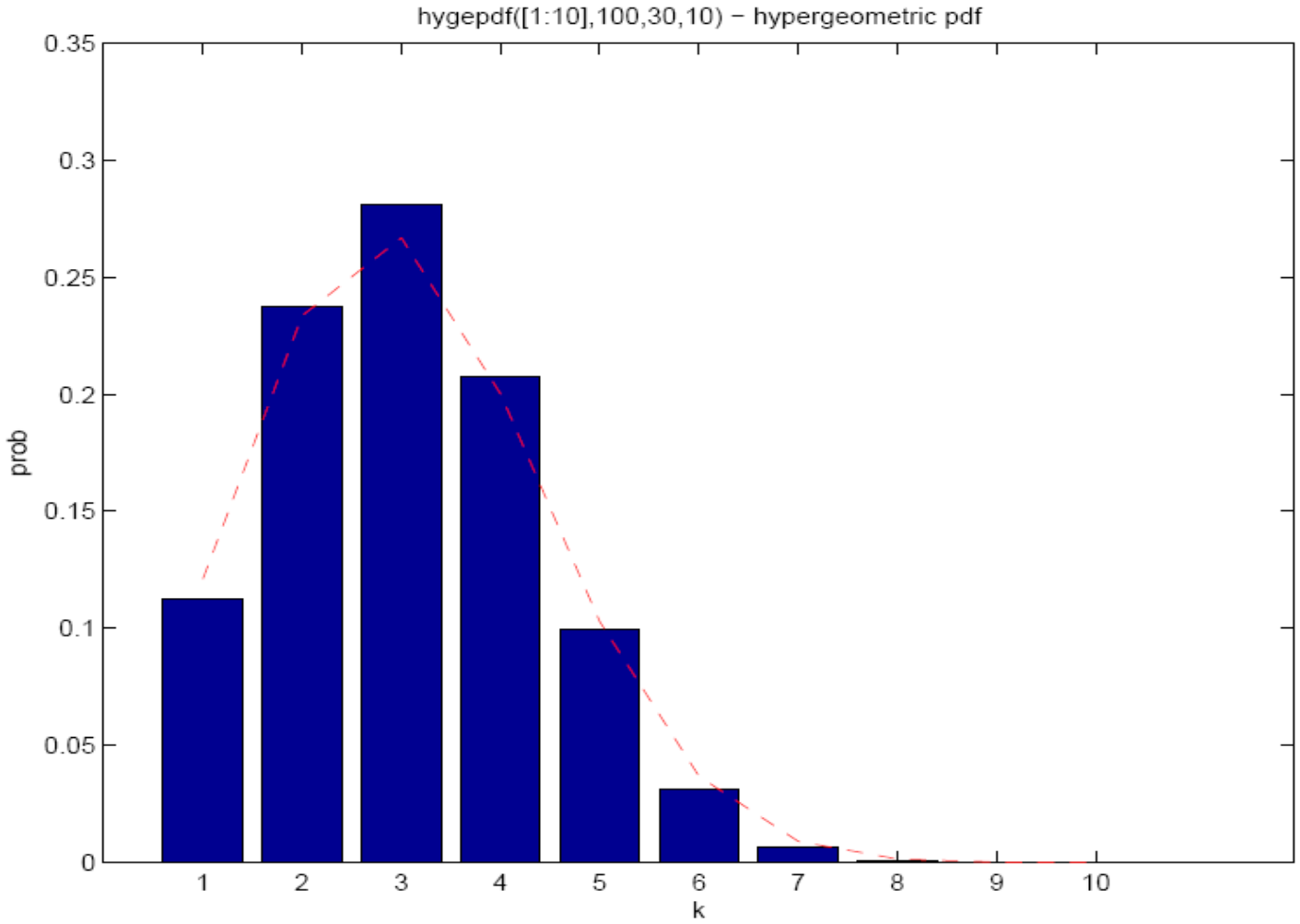
p = 3/10

Q2: Randomly pick one ball. Repeat *n* times.
What's the chance that the total number of red balls you pick is *k*?

Total number of red balls: K
Total number of balls: N

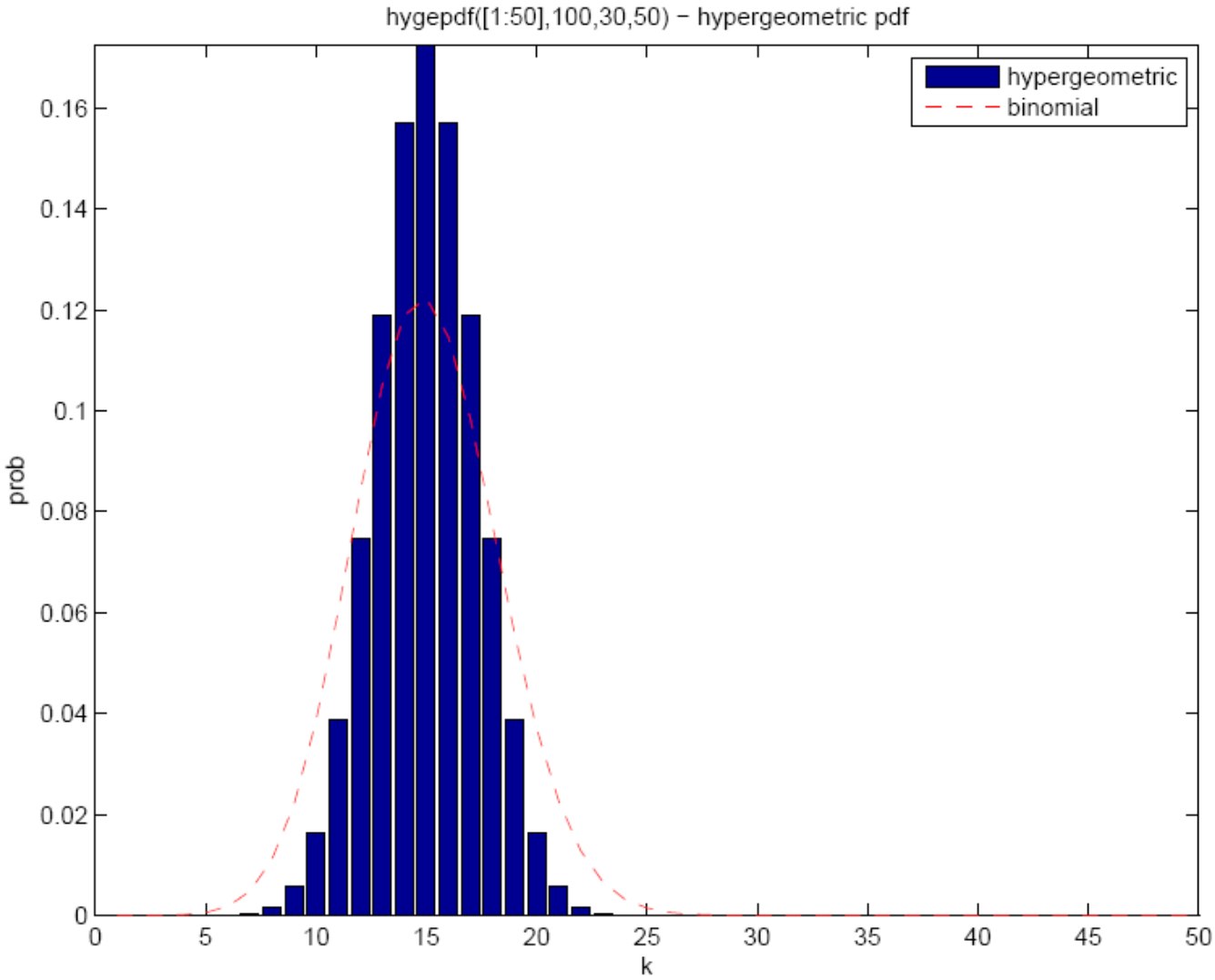
$$P(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Hypergeometric



The distribution can be approximated by the binomial distribution, if K and N are much larger than n and K/N is not close to 0 or 1.

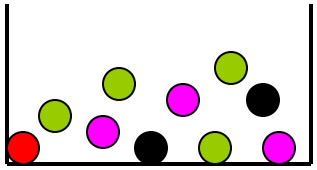
Hypergeometric



Multinomial distribution

Experiment 3: Sampling with replacement

A box with 1 red balls, 2 **black** balls, 3 pink balls, and 4 green balls:



Q1: Randomly pick one ball. What's the chance that the ball is red?

- $p_1=1/10$
- $p_2=2/10$
- $p_3=3/10$
- $p_4=4/10$

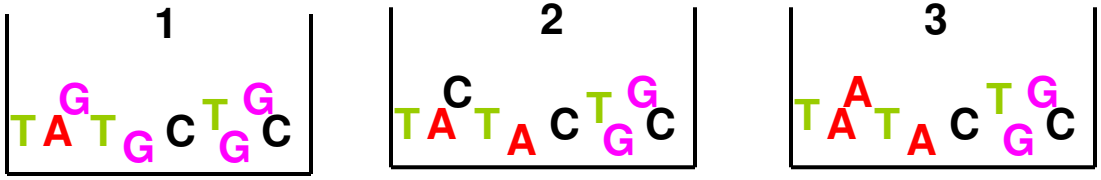
Q2: Sampling with replacement, what's the chance of picking 3 balls with colors in the order of: ● ● ●

$$p_1 * p_4 * p_2$$

Position Weight Matrix

Experiment 4: Sampling with replacement from ***W=3*** boxes

Boxes with nucleotides: A, **C**, G, and T::



A	θ_{11}	θ_{21}	θ_{31}
C	θ_{12}	θ_{22}	θ_{32}
G	θ_{13}	θ_{33}	θ_{33}
T	θ_{14}	θ_{44}	θ_{34}

Q: Pick one letter from each box in the order of 1,2,3. What's the chance of picking: **ACT**?

$$\theta_{11}\,\theta_{22}\,\theta_{34}$$

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