



## Regulatory Motifs

J-1



## Suppose We Sequence a Genome...

### Open Questions

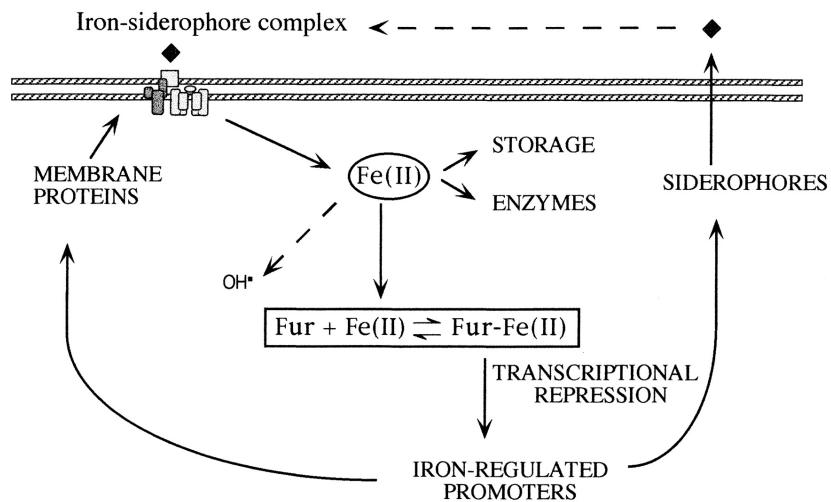
- Which regions of the genome have biological function? (What are the genes?)
- What are the functions of these regions?
- How and when are genes turned on and off?
- How do genes and their products interact with each other?

Bioinformatics methods are often hypothesis generating

J-2



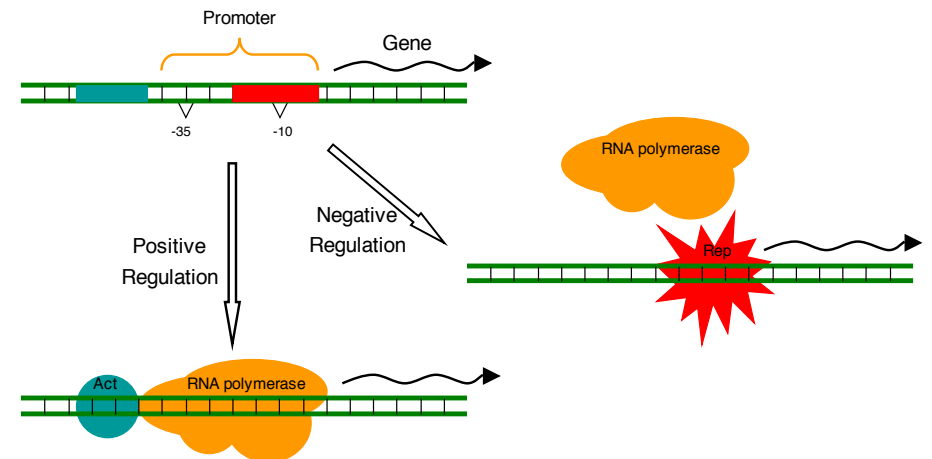
## Gene Regulation: An Example With Iron



J-3



## Gene Regulation

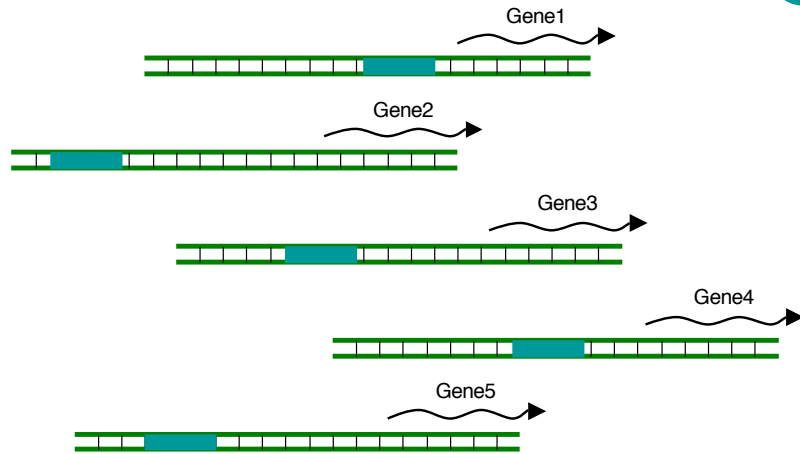


J-4



## What If We Believed That a Number of Genes Were Regulated By the Same Transcription Factor?

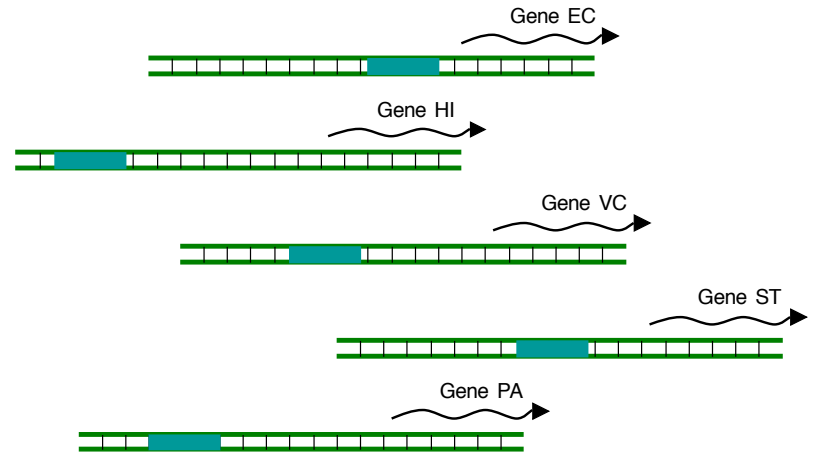
TF  
"X"



J - 5



## What If We Believed That a Number of Genes Were Orthologous



J - 6



## If We Knew Where the Motif Instances Were Located in Each Sequence...

> Escherichia coli  
TTGATTCCTGAATGCCGCTTAGT**GTAACTACTGTAA**CGGCATTTCTGCTTTTCC  
TGCCGATATTTTCTTATCTACCTCACAAAGTTAGCAATAACTGCTGGGAAAATCCG  
AGTTAGTCGTTATATCTAT

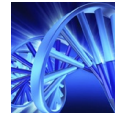
> Haemophilus influenzae  
ATCTAACGGTACGGATCTCCAAAGGCCTATGGAACTCTGTAGAATATGAACGTTCTAA  
TAAATCATAAAGTTGGAGCAACGCTCGGCATAAGTAGTAGTGCCGTGCCCTCCGCCATT  
**AGTTAACTAGTGGGAC**ACC

> Vibrio cholerae  
ATTTGTGGCGTTTCAAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
GAGCTGGCAGCAGCGAAGTCTGGGGCTGTTTGAAC**GTTACACGAGTGTAA**CCCGCGAA  
CCATGTTGACACGAATTCG

> Salmonella typhi  
GGTCGG**CTTAGACTAGTGTGAC**CAAAAAGCTTTTCTGCTGAAGTTTCAGGTAAGAAGAAC  
AGCTCCTAGTAAAAGACTATTTGTGACTGAAAAGCGCTCAGCGCAAAAGCCGACCGCAC  
AAACGCACAAGGAGTTACG

> Pseudomonas aeruginosa  
ACCGGCCAGGGTCTTCTCCTGCGAGATCATCGCGGGCGCGCGCATGCCGGCGCCG  
TGCTGGAACCCCTCGACCCAG**GCTACACTAGTTTAA**CCGGAACGCCGCACTGGATCG  
GCCTGCCCCAGCTATTGCTC

J - 7



## Then We Could Determine a Motif Model!

**GTAACTACTGTAA**  
**GTTAACTAGTGGGAC**  
**GTTACACGAGTGTAA**  
**CTTAGACTAGTGTGAC**  
**GCTAACTAGTTTAA**

<b>A</b>	0.0	0.0	.20	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	.60	1.0	0.0
<b>C</b>	.20	.20	0.0	0.0	.80	0.0	1.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	0.0	1.0
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	.20	0.0	.80	0.0	.80	.20	.40	0.0	0.0
<b>T</b>	0.0	.80	.80	0.0	0.0	0.0	0.0	.80	0.0	0.0	1.0	.20	.80	0.0	0.0	0.0

**G T T A C A C T A G T G T A A C**

Consensus Sequence

J - 8



## We Could Determine the Location of the Motif Instance That Best Matches the Model...

<b>A</b>	0.0	0.0	.20	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	.60	1.0	0.0
<b>C</b>	.20	.20	0.0	0.0	.80	0.0	1.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	0.0	1.0
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	.20	0.0	.80	0.0	.80	.20	.40	0.0	0.0
<b>T</b>	0.0	.80	.80	0.0	0.0	0.0	0.0	.80	0.0	0.0	1.0	.20	.80	0.0	0.0	0.0

$$\text{Score} = 0.0 * .80 * 0.0 * 1.0 * 0.0 * 0.0 * 1.0 * 0.0 * 0.0 * 0.0 * 0.0 * 0.0 * 0.0 * 0.0 * 0.0 * 1.0$$

$$\text{Score} = 0.01 * .80 * 0.01 * 1.0 * 0.01 * 0.01 * 1.0 * 0.01 * 0.01 * 0.01 * 0.01 * 0.01 * 0.01 * 0.01 * 1.0$$

$$\text{Score} = 8.0 * 10^{-27}$$

TTGATTCCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

J-9



## We Could Determine the Location of the Motif Instance That Best Matches the Model...

<b>A</b>	0.0	0.0	.20	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	.60	1.0	0.0	
<b>C</b>	.20	.20	0.0	0.0	.80	0.0	1.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	1.0	
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	.20	0.0	.80	0.0	.80	.20	.40	0.0	0.0
<b>T</b>	0.0	.80	.80	0.0	0.0	0.0	0.0	.80	0.0	0.0	1.0	.20	.80	0.0	0.0	0.0

$$\text{Score} = 0.0 * 0.0 * .20 * 0.0 * 0.0 * 0.0 * 1.0 * 0.0 * 0.0 * .80 * 0.0 * 0.0 * .80 * .40 * 0.0 * 1.0$$

$$\text{Score} = 0.01 * 0.01 * .20 * 0.01 * 0.01 * 0.01 * 1.0 * 0.01 * 0.01 * .80 * 0.01 * 0.01 * .80 * .40 * 0.01 * 1.0$$

$$\text{Score} = 5.12 * 10^{-22}$$

TTGATTCCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

J-10



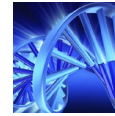
## We Could Determine the Location of the Motif Instance That Best Matches the Model...

<b>A</b>	0.0	0.0	.20	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	.60	1.0	0.0
<b>C</b>	.20	.20	0.0	0.0	.80	0.0	1.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	0.0	1.0
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	.20	0.0	.80	0.0	.80	.20	.40	0.0	0.0
<b>T</b>	0.0	.80	.80	0.0	0.0	0.0	0.0	.80	0.0	0.0	1.0	.20	.80	0.0	0.0	0.0

$$\text{Score} = 6.77 * 10^{-25}$$

TTGATTCCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

J-11



## We Could Determine the Location of the Motif Instance That Best Matches the Model...

<b>A</b>	0.0	0.0	.20	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	.60	1.0	0.0	
<b>C</b>	.20	.20	0.0	0.0	.80	0.0	1.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	1.0	
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	.20	0.0	.80	0.0	.80	.20	.40	0.0	0.0
<b>T</b>	0.0	.80	.80	0.0	0.0	0.0	0.0	.80	0.0	0.0	1.0	.20	.80	0.0	0.0	0.0

$$\text{Score} = 7.16 * 10^{-28}$$

TTGATTCCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

J-12



## k-means Clustering Algorithm

- Randomly assign each point (gene) to one of the  $k$  clusters
- Repeat until convergence
  - Calculate model of each of the  $k$  clusters
  - Assign each point (gene) to the cluster with the closest model

J - 13



## Expectation-Maximization (EM)

- Randomly guess the locations of each motif instance
- Repeat until convergence
  - Calculate a new motif model from the motif instances
  - Calculate new locations for the motif instances from the motif model

J - 14



## EM - Randomly Guess the Locations of Each Motif Instance

> Escherichia coli  
 TTGATTCGGTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCGCTTTTCC  
 TGCCGATATTTTTCTTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAA**AAATCCG**  
**AGTTAGTCG**TTATATTCTAT

> Haemophilus influenzae  
**ATCTAACGGTACGGATT**CTCCAAAGGCCTATGGAATCTGTAGAATATGAAACGTTCTAA  
 TAAATCATAAAGTTGGAGCAACGCTCGGCATAAGTAGTAGTGAAGTGCCTGCCATCCGCCATT  
 AGTTACACTAGTGGGACACC

> Vibrio cholerae  
 ATTTGTGGCGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
 GAGGCTGGCAGCAGCGAAGGTCGGGGCTGTTTGAACGTTACACGAGTGAACCCGCCAA  
**CCATGTTGACACGAATTC**TG

> Salmonella typhi  
 GGTCGGCTTAGACTAGTGTGACCAAAAAGCTTTTCTGCTGAAAGTTTCAGGGTAAAGAACC  
 AGCTCCTAGTAAAAGACTAT**TGTGACTGAAAAGCGC**GTACGGCAGCAAGCCGACCGCAC  
 AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa  
 ACGGGCCAGGGTCTTCTCCTGCGAGATCATGCGGGCGCGCCGCGCATG**CCGGGCGCCG**  
**TGCTGG**AACGCCCTCGACCCAGGGCTACACTAGTTTAAACGGAAACGCCGAGTGGATCG  
 GCCTGCCCCAGCTATTGCTC

J - 15



A	.40	.20	0.0	.20	.40	0.0	.20	.20	.40	.40	.20	.60	.20	.20	0.0	0.0
C	.20	.40	0.0	0.0	.40	.60	.20	.40	0.0	.40	.20	0.0	.20	0.0	.20	.40
G	0.0	.20	.40	.40	0.0	.40	.40	.40	.40	0.0	.20	.40	.60	.20	.40	.40
T	.40	.20	.60	.40	.20	0.0	.20	0.0	.20	.20	.40	0.0	0.0	.60	.40	.20

> Escherichia coli  
 TTGATTCGGTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCGCTTTTCC  
 TGCCGATATTTTTCTTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAA**AAATCCG**  
**AGTTAGTCG**TTATATTCTAT

> Haemophilus influenzae  
**ATCTAACGGTACGGATT**CTCCAAAGGCCTATGGAATCTGTAGAATATGAAACGTTCTAA  
 TAAATCATAAAGTTGGAGCAACGCTCGGCATAAGTAGTAGTGAAGTGCCTGCCATCCGCCATT  
 AGTTACACTAGTGGGACACC

> Vibrio cholerae  
 ATTTGTGGCGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
 GAGGCTGGCAGCAGCGAAGGTCGGGGCTGTTTGAACGTTACACGAGTGAACCCGCCAA  
**CCATGTTGACACGAATTC**TG

> Salmonella typhi  
 GGTCGGCTTAGACTAGTGTGACCAAAAAGCTTTTCTGCTGAAAGTTTCAGGGTAAAGAACC  
 AGCTCCTAGTAAAAGACTAT**TGTGACTGAAAAGCGC**GTACGGCAGCAAGCCGACCGCAC  
 AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa  
 ACGGGCCAGGGTCTTCTCCTGCGAGATCATGCGGGCGCGCCGCGCATG**CCGGGCGCCG**  
**TGCTGG**AACGCCCTCGACCCAGGGCTACACTAGTTTAAACGGAAACGCCGAGTGGATCG  
 GCCTGCCCCAGCTATTGCTC

J - 16



A	.40	.20	0.0	.20	.40	0.0	.20	.20	.40	.40	.20	.60	.20	.20	0.0	0.0
C	.20	.40	0.0	0.0	.40	.60	.20	.40	0.0	.40	.20	0.0	.20	0.0	.20	.40
G	0.0	.20	.40	.40	0.0	.40	.20	.40	0.0	0.0	.20	.40	.60	.20	.40	.40
T	.40	.20	.60	.40	.20	0.0	.20	0.0	.20	.20	.40	0.0	0.0	.60	.40	.20

> Escherichia coli

TTGATTC CCTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCTGCTTTTCC  
TGCCGATATTTTTCTTATCTACCTCACAAAGGTAGCAATAACTGCTGGGAAAATCCG  
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae

ATCTAACGGTACGGATCTCCAAAGGCCTATGGAATCTGTAGAATATGAAACGTTCTAA  
TAAATCATAAAGTTGGAGCAAAGCTCGCATAAGTAGTAGTCCGTCGCTCCGCCATT  
AGTTACACTAGTGGACACC

> Vibrio cholerae

ATTTGTGGCGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
GAGGCTGGCAGCAGCGAAGGTCGGGGCTGTTGAACGTTACACGAGTGAACCGCGAA  
CCATGTTGACACGAATTCG

> Salmonella typhi

GGTCGGCTTAGACTAGTGTGACCAAAAAGCTTTTCTGCTGAA GTTTCAGGGTAAAGAACC  
AGCTCCTAGTAAAAGACTATTGTGACTGAAAAGCGCTCAGCGCAAAGCGACCGCACA  
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa

ACCGGCCAGGGTCTTCTCCTGCGAGATCATGCGGGCGCGCCGCGCATGCCGGCGCCG  
TGCTGGAACGCCCTCGACCCAGGGCTACACTAGTTTAACCGGAACGCCGCCAGTGGATCG  
GCCTGCCCCAGCTATTGGCT



A	.40	.20	.40	0.0	.40	.80	.20	.20	.20	0.0	.60	.80	.20	.60	.60	0.0
C	0.0	.20	0.0	0.0	.20	0.0	.20	.40	.20	.20	.20	0.0	.20	0.0	.20	.60
G	.60	.20	0.0	.40	.20	0.0	.40	.20	.60	.60	0.0	0.0	.40	.20	0.0	.40
T	0.0	.40	.60	.60	.20	.20	.20	.20	0.0	.20	.20	0.0	.20	0.0	.20	0.0

> Escherichia coli

TTGATTC CCTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCTGCTTTTCC  
TGCCGATATTTTTCTTATCTACCTCACAAAGGTAGCAATAACTGCTGGGAAAATCCG  
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae

ATCTAACGGTACGGATCTCCAAAGGCCTATGGAATCTGTAGAATATGAAACGTTCTAA  
TAAATCATAAAGTTGGAGCAAAGCTCGCATAAGTAGTAGTCCGTCGCTCCGCCATT  
AGTTACACTAGTGGACACC

> Vibrio cholerae

ATTTGTGGCGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
GAGGCTGGCAGCAGCGAAGGTCGGGGCTGTTGAACGTTACACGAGTGAACCGCGAA  
CCATGTTGACACGAATTCG

> Salmonella typhi

GGTCGGCTTAGACTAGTGTGACCAAAAAGCTTTTCTGCTGAA GTTTCAGGGTAAAGAACC  
AGCTCCTAGTAAAAGACTATTGTGACTGAAAAGCGCTCAGCGCAAAGCGACCGCACA  
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa

ACCGGCCAGGGTCTTCTCCTGCGAGATCATGCGGGCGCGCCGCGCATGCCGGCGCCG  
TGCTGGAACGCCCTCGACCCAGGGCTACACTAGTTTAACCGGAACGCCGCCAGTGGATCG  
GCCTGCCCCAGCTATTGGCT



A	.40	.20	.40	0.0	.40	.80	.20	.20	.20	0.0	.60	.80	.20	.60	.60	0.0
C	0.0	.20	0.0	0.0	.20	0.0	.20	.40	.20	.20	.20	.20	0.0	.20	0.0	.60
G	.60	.20	0.0	.40	.20	0.0	.40	.20	.60	.60	0.0	0.0	.40	.20	0.0	.40
T	0.0	.40	.60	.60	.20	.20	.20	.20	0.0	.20	.20	0.0	.20	0.0	.20	0.0

> Escherichia coli

TTGATTC CCTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCTGCTTTTCC  
TGCCGATATTTTTCTTATCTACCTCACAAAGGTAGCAATAACTGCTGGGAAAATCCG  
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae

ATCTAACGGTACGGATCTCCAAAGGCCTATGGAATCTGTAGAATATGAAACGTTCTAA  
TAAATCATAAAGTTGGAGCAAAGCTCGGCATAAGTAGTAGTCCGTCGCTCCGCCATT  
AGTTACACTAGTGGACACC

> Vibrio cholerae

ATTTGTGGCGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
GAGGCTGGCAGCAGCGAAGGTCGGGGCTGTTGAACGTTACACGAGTGAACCGCGAA  
CCATGTTGACACGAATTCG

> Salmonella typhi

GGTCGGCTTAGACTAGTGTGACCAAAAAGCTTTTCTGCTGAA GTTTCAGGGTAAAGAACC  
AGCTCCTAGTAAAAGACTATTGTGACTGAAAAGCGCTCAGCGCAAAGCGACCGCACA  
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa

ACCGGCCAGGGTCTTCTCCTGCGAGATCATGCGGGCGCGCCGCGCATGCCGGCGCCG  
TGCTGGAACGCCCTCGACCCAGGGCTACACTAGTTTAACCGGAACGCCGCCAGTGGATCG  
GCCTGCCCCAGCTATTGGCT



A	.20	0.0	0.0	.40	.20	.20	.20	.60	.20	.20	.60	0.0	.40	0.0	.20	0.0	.60	0.0
C	.20	0.0	0.0	0.0	.20	.20	.60	.40	0.0	0.0	0.0	.40	.20	.40	.20	.40	.20	.80
G	.60	0.0	0.0	0.0	.20	0.0	0.0	.20	0.0	0.0	.20	.60	.20	.40	.20	.40	.20	.20
T	0.0	1.0	1.0	.60	.40	.20	.20	.40	.20	.40	.40	.40	.40	.40	.20	.40	0.0	0.0

> Escherichia coli

TTGATTC CCTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCTGCTTTTCC  
TGCCGATATTTTTCTTATCTACCTCACAAAGGTAGCAATAACTGCTGGGAAAATCCG  
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae

ATCTAACGGTACGGATCTCCAAAGGCCTATGGAATCTGTAGAATATGAAACGTTCTAA  
TAAATCATAAAGTTGGAGCAAAGCTCGGCATAAGTAGTAGTCCGTCGCTCCGCCATT  
AGTTACACTAGTGGACACC

> Vibrio cholerae

ATTTGTGGCGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
GAGGCTGGCAGCAGCGAAGGTCGGGGCTGTTGAACGTTACACGAGTGAACCGCGAA  
CCATGTTGACACGAATTCG

> Salmonella typhi

GGTCGGCTTAGACTAGTGTGACCAAAAAGCTTTTCTGCTGAA GTTTCAGGGTAAAGAACC  
AGCTCCTAGTAAAAGACTATTGTGACTGAAAAGCGCTCAGCGCAAAGCGACCGCACA  
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa

ACCGGCCAGGGTCTTCTCCTGCGAGATCATGCGGGCGCGCCGCGCATGCCGGCGCCG  
TGCTGGAACGCCCTCGACCCAGGGCTACACTAGTTTAACCGGAACGCCGCCAGTGGATCG  
GCCTGCCCCAGCTATTGGCT



<b>A</b>	.20	0.0	0.0	.40	.20	.60	.20	.20	.60	0.0	.40	0.0	.20	0.0	.60	0.0
<b>C</b>	.20	0.0	0.0	0.0	.20	.20	.60	.40	0.0	0.0	0.0	.40	.20	.40	.20	.80
<b>G</b>	.60	0.0	0.0	0.0	.20	0.0	0.0	0.0	.20	.60	.20	.20	.40	.20	.20	.20
<b>T</b>	0.0	1.0	1.0	.60	.40	.20	.20	.40	.20	.40	.40	.40	.20	.40	0.0	0.0

> Escherichia coli

TTGATTC CCTGAATGCCCGCTTAGT**GTAACACTACTGTAAC**CGGCATTTTCTGCTTTTCC  
TGCCGATATTTTTCCTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCGG  
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae

ATCTAACGGTACGGATCTCCAAAGGCCTATGGAACTCTGTAGAATATGAACGTTCTAA  
TAAATCATAAAGTTGGACAAACGCTCGGCATAAGTAGTAGTGCCGCTGCCGCCATT  
**AGTTACACTAGTGGGAC**ACC

> Vibrio cholerae

ATTTGTGGCG**GTTTCAAAATGCTTGG**AGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
GAGGCTGGCAGCAGCGAAGGTTGGGGCTGTTGAACGTTACACGAGTGAACCGCCGAA  
CCATGTTGACACGAATTCG

> Salmonella typhi

GGTCGG**CCTAGACTAGTGTGAC**CAAAAAGCTTTTCTGCTGAAAGTTTCAGGGTAAAGAACC  
AGCTCCTAGTAAAAGACTATTGTGACTGAAAAGCGCTCAGCGCAAAAGCCGACCGCACA  
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa

ACCGGCCAGGGTCTTCTCCTGCGAGATCATGCGGGCGCGCCGCGCATGCCGGCGCCG  
TGCTGGAACGCCCTCGACCCAG**GCTACACTAGTTTAAAC**CGGAACGCCGCGAGTGGATCG  
GCCTGCCCCAGCTATTGGCT



<b>A</b>	0.0	0.0	.20	.80	0.0	.80	.20	.20	1.0	0.0	0.0	0.0	0.0	0.0	.40	.80	0.0
<b>C</b>	.20	0.0	0.0	0.0	.60	.20	.80	0.0	0.0	.20	0.0	.20	0.0	0.0	0.0	0.0	.80
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	.60	.20	.60	.20	.40	.20	.20	.20
<b>T</b>	0.0	1.0	.80	.20	.20	0.0	0.0	.80	0.0	.20	.80	.20	.80	.20	0.0	0.0	0.0

> Escherichia coli

TTGATTC CCTGAATGCCCGCTTAGT**GTAACACTACTGTAAC**CGGCATTTTCTGCTTTTCC  
TGCCGATATTTTTCCTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCGG  
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae

ATCTAACGGTACGGATCTCCAAAGGCCTATGGAACTCTGTAGAATATGAACGTTCTAA  
TAAATCATAAAGTTGGACAAACGCTCGGCATAAGTAGTAGTGCCGCTGCCGCCATT  
**AGTTACACTAGTGGGAC**ACC

> Vibrio cholerae

ATTTGTGGCG**GTTTCAAAATGCTTGG**AGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
GAGGCTGGCAGCAGCGAAGGTTGGGGCTGTTGAACGTTACACGAGTGAACCGCCGAA  
CCATGTTGACACGAATTCG

> Salmonella typhi

GGTCGG**CCTAGACTAGTGTGAC**CAAAAAGCTTTTCTGCTGAAAGTTTCAGGGTAAAGAACC  
AGCTCCTAGTAAAAGACTATTGTGACTGAAAAGCGCTCAGCGCAAAAGCCGACCGCACA  
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa

ACCGGCCAGGGTCTTCTCCTGCGAGATCATGCGGGCGCGCCGCGCATGCCGGCGCCG  
TGCTGGAACGCCCTCGACCCAG**GCTACACTAGTTTAAAC**CGGAACGCCGCGAGTGGATCG  
GCCTGCCCCAGCTATTGGCT



<b>A</b>	0.0	0.0	.20	.80	0.0	.80	.20	.20	1.0	0.0	0.0	0.0	0.0	.40	.80	0.0
<b>C</b>	.20	0.0	0.0	0.0	.60	.20	.80	0.0	0.0	.20	0.0	.20	0.0	0.0	0.0	.80
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	.60	.20	.60	.20	.40	.20	.20
<b>T</b>	0.0	1.0	.80	.20	.20	0.0	0.0	.80	0.0	.20	.80	.20	.80	.20	0.0	0.0

> Escherichia coli

TTGATTC CCTGAATGCCCGCTTAGT**GTAACACTACTGTAAC**CGGCATTTTCTGCTTTTCC  
TGCCGATATTTTTCCTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCGG  
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae

ATCTAACGGTACGGATCTCCAAAGGCCTATGGAACTCTGTAGAATATGAACGTTCTAA  
TAAATCATAAAGTTGGACAAACGCTCGGCATAAGTAGTAGTGCCGCTGCCGCCATT  
**AGTTACACTAGTGGGAC**ACC

> Vibrio cholerae

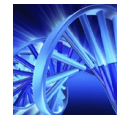
ATTTGTGGCG**GTTTCAAAATGCTTGG**AGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
GAGGCTGGCAGCAGCGAAGGTTGGGGCTGTTGAAC**GTTACACGAGTGTAAAC**CGCCGAA  
CCATGTTGACACGAATTCG

> Salmonella typhi

GGTCGG**CCTAGACTAGTGTGAC**CAAAAAGCTTTTCTGCTGAAAGTTTCAGGGTAAAGAACC  
AGCTCCTAGTAAAAGACTATTGTGACTGAAAAGCGCTCAGCGCAAAAGCCGACCGCACA  
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa

ACCGGCCAGGGTCTTCTCCTGCGAGATCATGCGGGCGCGCCGCGCATGCCGGCGCCG  
TGCTGGAACGCCCTCGACCCAG**GCTACACTAGTTTAAAC**CGGAACGCCGCGAGTGGATCG  
GCCTGCCCCAGCTATTGGCT



<b>A</b>	0.0	0.0	.20	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	.60	1.0	0.0
<b>C</b>	.20	.20	0.0	0.0	.80	0.0	1.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	0.0	1.0
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	0.0	.20	0.0	.80	0.0	.80	.20	.40	0.0
<b>T</b>	0.0	.80	.80	0.0	0.0	0.0	0.0	.80	0.0	0.0	1.0	.20	.80	0.0	0.0	0.0

> Escherichia coli

TTGATTC CCTGAATGCCCGCTTAGT**GTAACACTACTGTAAC**CGGCATTTTCTGCTTTTCC  
TGCCGATATTTTTCCTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCGG  
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae

ATCTAACGGTACGGATCTCCAAAGGCCTATGGAACTCTGTAGAATATGAACGTTCTAA  
TAAATCATAAAGTTGGACAAACGCTCGGCATAAGTAGTAGTGCCGCTGCCGCCATT  
**AGTTACACTAGTGGGAC**ACC

> Vibrio cholerae

ATTTGTGGCG**GTTTCAAAATGCTTGG**AGAATGGGTACATGATCCGCTTGGCATTGAAGGT  
GAGGCTGGCAGCAGCGAAGGTTGGGGCTGTTGAAC**GTTACACGAGTGTAAAC**CGCCGAA  
CCATGTTGACACGAATTCG

> Salmonella typhi

GGTCGG**CCTAGACTAGTGTGAC**CAAAAAGCTTTTCTGCTGAAAGTTTCAGGGTAAAGAACC  
AGCTCCTAGTAAAAGACTATTGTGACTGAAAAGCGCTCAGCGCAAAAGCCGACCGCACA  
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa

ACCGGCCAGGGTCTTCTCCTGCGAGATCATGCGGGCGCGCCGCGCATGCCGGCGCCG  
TGCTGGAACGCCCTCGACCCAG**GCTACACTAGTTTAAAC**CGGAACGCCGCGAGTGGATCG  
GCCTGCCCCAGCTATTGGCT



## Expectation-Maximization (EM)

- Randomly guess the locations of each motif instance
- Repeat until convergence
  - Calculate a new motif model from the motif instances
  - Calculate new locations for the motif instances from the motif model

Each motif instance is *best scoring* match to motif model

J - 25



## Gibbs Sampling

- Randomly guess the locations of each motif instance
- Repeat until convergence
  - Calculate a new motif model from the motif instances
  - Calculate new locations for the motif instances from the motif model

Each motif instance is *sampled* from scores of matches to motif model

J - 26



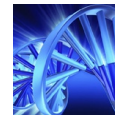
## We Could Determine the Location of the Motif Instance That Best Matches the Model...

<b>A</b>	0.0	0.0	.20	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	.60	1.0	0.0
<b>C</b>	.20	.20	0.0	0.0	.80	0.0	1.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	0.0	1.0
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	.20	0.0	.80	0.0	.80	.20	.40	0.0	0.0
<b>T</b>	0.0	.80	.80	0.0	0.0	0.0	0.0	.80	0.0	0.0	1.0	.20	.80	0.0	0.0	0.0

Score =  $8.0 * 10^{-27}$

TTGATTC CCTGAATGCCCGCTTAGTGTAACACTACTGTAA

J - 27



## We Could Determine the Location of the Motif Instance That Best Matches the Model...

<b>A</b>	0.0	0.0	.20	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	.60	1.0	0.0
<b>C</b>	.20	.20	0.0	0.0	.80	0.0	1.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	0.0	1.0
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	.20	0.0	.80	0.0	.80	.20	.40	0.0	0.0
<b>T</b>	0.0	.80	.80	0.0	0.0	0.0	0.0	.80	0.0	0.0	1.0	.20	.80	0.0	0.0	0.0

Score =  $5.12 * 10^{-22}$

TTGATTC CCTGAATGCCCGCTTAGTGTAACACTACTGTAA

J - 28



## We Could Determine the Location of the Motif Instance That Best Matches the Model...

<b>A</b>	0.0	0.0	.20	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	.60	1.0	0.0
<b>C</b>	.20	.20	0.0	0.0	.80	0.0	1.0	0.0	0.0	.20	0.0	0.0	0.0	0.0	0.0	1.0
<b>G</b>	.80	0.0	0.0	0.0	.20	0.0	0.0	.20	0.0	.80	0.0	.80	.20	.40	0.0	0.0
<b>T</b>	0.0	.80	.80	0.0	0.0	0.0	0.0	.80	0.0	0.0	1.0	.20	.80	0.0	0.0	0.0

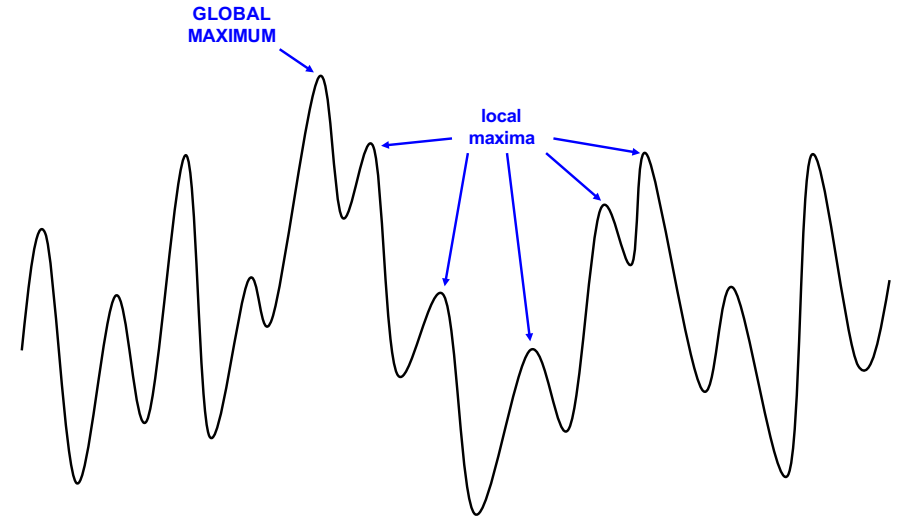
Score =  $7.16 * 10^{-28}$

TTGATTCCCTGAATGCCCGCTTAG**TGTAACACTACTGTAA**

J - 29



## Exponential Search Space



J - 30



## How Do You Sample?

J - 31