Regulatory Motifs

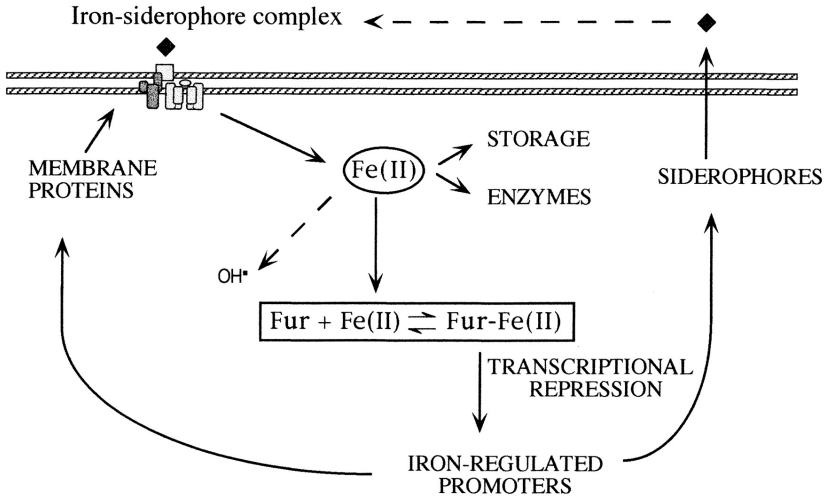
Suppose We Sequence a Genome…

Open Questions

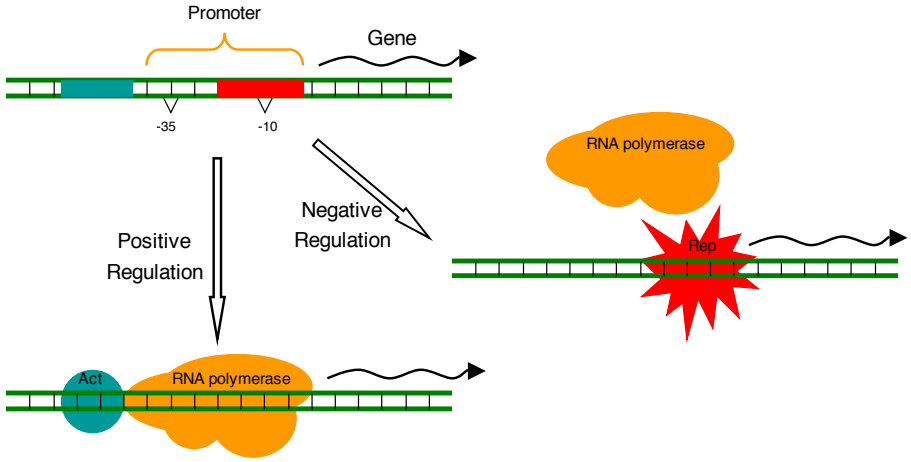
- Which regions of the genome have biological function? (What are the genes?)
- What are the functions of these regions?
- How and when are genes turned on and off?
- How do genes and their products interact with each other?

Bioinformatics methods are often hypothesis generating

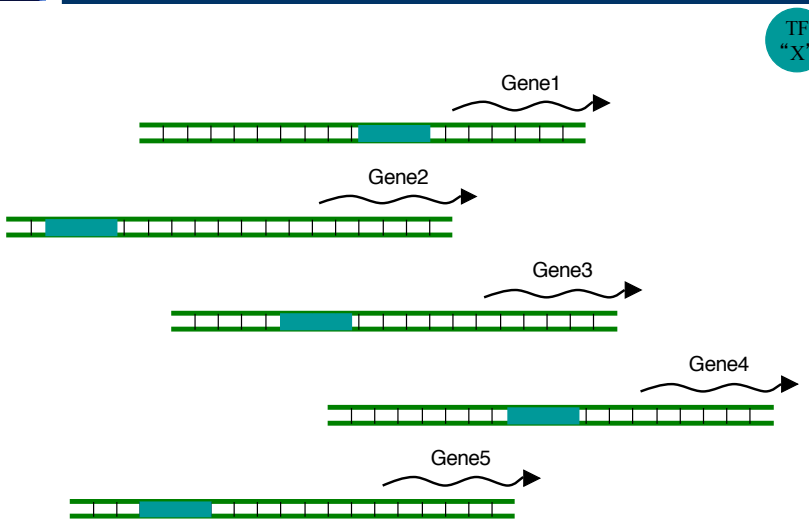
Gene Regulation: An Example With Iron



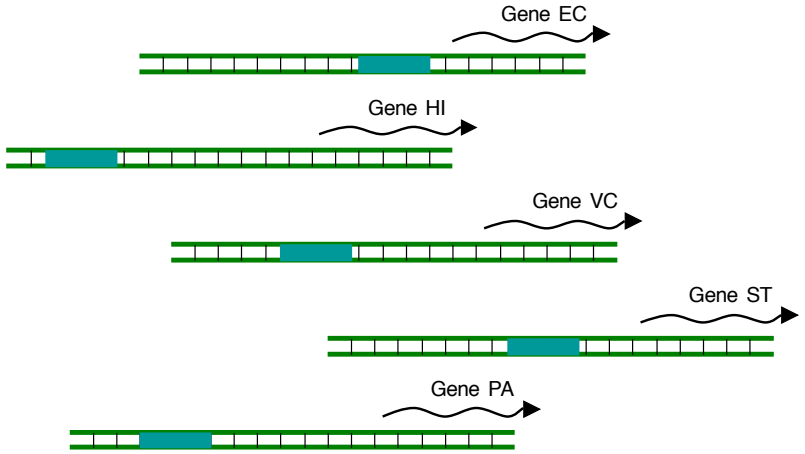
Gene Regulation



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



What If We Believed That a Number of Genes Were Orthologous



If We Knew *Where* the Motif Instances Were Located in Each Sequence…

> Escherichia coli
TTGATTCCCTGAATGCCCGCTTAGT**GTAACACTACTGTAAC**CGGCATTTTCTGCTTTTCC
TGCCGATATTTTTTCTTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCCG
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
ATCTAACGGTACGGATTCTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCGGCATAAGTAGTAAGTGCCGTGCCTCCGCCATT
A**GTTACACTAGTGGGAC**ACC

> Vibrio cholerae
ATTTGTGGCGGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAAC**GTTACACGAGTGTAAC**CGCCGAA
CCATGTTGACACGAATTCTG

> Salmonella typhi
GGTCGG**CTTAGACTAGTGTGAC**CAAAAAGCTTTTGCTGAAGTTTCAGGGTAAGAAGAACC
AGCTCCTAGTAAAAAGACTATTGTGACTGAAAAGCGCGTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATGCCGGCGCCGC
TGCTGGAACGCCTCGACCCCAGG**GCTACACTAGTTTAAC**CGGAACGCCGCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC

Then We Could Determine a Motif Model!

GTAACACTACTGTAAC

GTTACACTAGTGGGAC

GTTACACGAGTGTAAC

CTTAGACTAGTGTGAC

GCTACACTAGTTTAAC

A	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
C	.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
G	.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
T	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

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

A	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
C	.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
G	.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
T	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 = 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

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 * 0.01 * 1.0

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

TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

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

A	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
C	.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
G	.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
T	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 = 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

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

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

TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

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

A	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
C	.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
G	.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
T	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 = $6.77 * 10^{-25}$

TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

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

A	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
C	.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
G	.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
T	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 = $7.16 * 10^{-28}$

TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

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

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

EM - Randomly Guess the Locations of Each Motif Instance

> Escherichia coli
TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCTGCTTTTCC
TGCCGATATTTTTTCTTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAA**AATTCCG**
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
A**TCTAACGGTACGGATT**CTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCGGCATAAGTAGTAAGTGCCGTGCCTCCGCCATT
AGTTACACTAGTGGGACACC

> Vibrio cholerae
ATTTGTGGCGGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAACGTTACACGAGTGTAACCGCCGAA
CC**ATGTTGACACGAATTC**TG

> Salmonella typhi
GGTCGGCTTAGACTAGTGTGACCAAAAAGCTTTTGCTGAAGTTTCAGGGTAAGAAGAACC
AGCTCCTAGTAAAAAGACTAT**TGTGACTGAAAAGCGC**GTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATG**CCGGCGCCGC**
TGCTGGAACGCCTCGACCCCAGGGCTACACTAGTTTAACCGGAACGCCGCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC

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

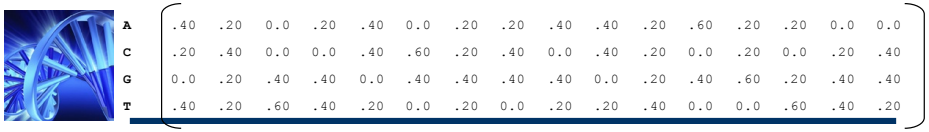
> Escherichia coli
TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCTGCTTTTCC
TGCCGATATTTTTTCTTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAA**AATTCCG**
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
A**TCTAACGGTACGGATT**CTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCGGCATAAGTAGTAAGTGCCGTGCCTCCGCCATT
AGTTACACTAGTGGGACACC

> Vibrio cholerae
ATTTGTGGCGGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAACGTTACACGAGTGTAACCGCCGAA
CC**ATGTTGACACGAATTC**TG

> Salmonella typhi
GGTCGGCTTAGACTAGTGTGACCAAAAAGCTTTTGCTGAAGTTTCAGGGTAAGAAGAACC
AGCTCCTAGTAAAAAGACTAT**TGTGACTGAAAAGCGC**GTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATG**CCGGCGCCGC**
TGCTGGAACGCCTCGACCCCAGGGCTACACTAGTTTAACCGGAACGCCGCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC



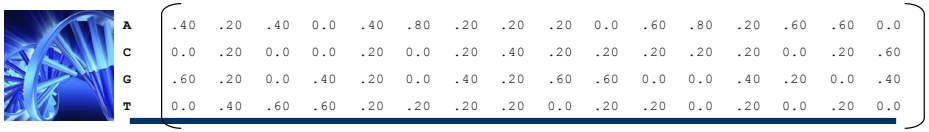
> Escherichia coli
TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCTGCTTTTCC
TGCCGATATTTTTTCTTATCTACCTCAC**AAAGGTTAGCAATAAC**TGCTGGGAAAATTCCG
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
ATCTAACGGTACGGATTCTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCG**GCATAAGTAGTAAGTG**CCGTGCCTCCGCCATT
AGTTACACTAGTGGGACACC

> Vibrio cholerae
ATTTGTGGCGGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAACGTTACACG**AGTGTAACCGCCGAA**
CCATGTTGACACGAATTCTG

> Salmonella typhi
GGTCGGCTTAGACTAGTGTGACCAAAAAGCTTTTGCTGAA**GTTTCAGGGTAAGAAG**AACC
AGCTCCTAGTAAAAAGACTATTGTGACTGAAAAGCGCGTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATGCCGGCGCCGC
TGCTGGAACGCCTCGACCCCAGGGCTACACTA**GTTTAACCGGAACGCC**GCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC



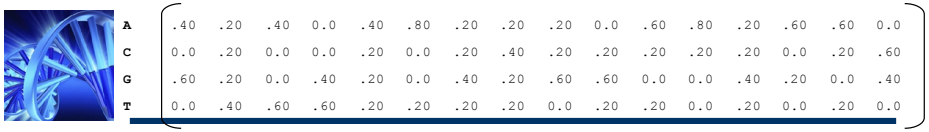
> Escherichia coli
TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCTGCTTTTCC
TGCCGATATTTTTTCTTATCTACCTCAC**AAAGGTTAGCAATAAC**TGCTGGGAAAATTCCG
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
ATCTAACGGTACGGATTCTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCG**GCATAAGTAGTAAGTG**CCGTGCCTCCGCCATT
AGTTACACTAGTGGGACACC

> Vibrio cholerae
ATTTGTGGCGGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAACGTTACACG**AGTGTAACCGCCGAA**
CCATGTTGACACGAATTCTG

> Salmonella typhi
GGTCGGCTTAGACTAGTGTGACCAAAAAGCTTTTGCTGAA**GTTTCAGGGTAAGAAG**AACC
AGCTCCTAGTAAAAAGACTATTGTGACTGAAAAGCGCGTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATGCCGGCGCCGC
TGCTGGAACGCCTCGACCCCAGGGCTACACTA**GTTTAACCGGAACGCC**GCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC



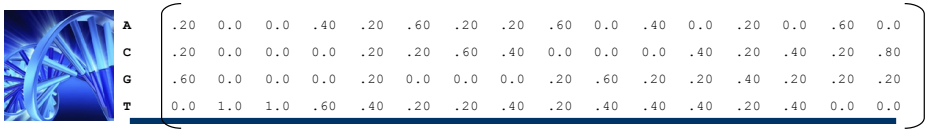
> Escherichia coli
TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCTGCTTTTCC
TGCCGAT**ATTTTTTCTTATCTAC**CTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCCG
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
ATCTAACGGTACGGATTCTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCGGCATAAGTAGTAAGTGCCGTGCCTCCGCCATT
A**GTTACACTAGTGGGAC**ACC

> Vibrio cholerae
ATTTGTGGCG**GTTTTCAAATGCTTGG**AGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAACGTTACACGAGTGTAACCGCCGAA
CCATGTTGACACGAATTCTG

> Salmonella typhi
GGTCGG**CTTAGACTAGTGTGAC**CAAAAAGCTTTTGCTGAAGTTTCAGGGTAAGAAGAACC
AGCTCCTAGTAAAAAGACTATTGTGACTGAAAAGCGCGTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATGCCGGCGCCGC
TGCTGGAACGCCTCGACCCCAGGGCTACACTA**GTTTAACCGGAACGCC**GCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC



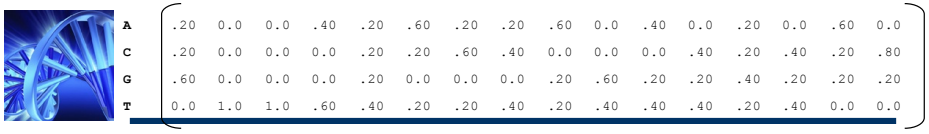
> Escherichia coli
TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAACCGGCATTTTCTGCTTTTCC
TGCCGAT**ATTTTTTCTTATCTAC**CTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCCG
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
ATCTAACGGTACGGATTCTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCGGCATAAGTAGTAAGTGCCGTGCCTCCGCCATT
A**GTTACACTAGTGGGAC**ACC

> Vibrio cholerae
ATTTGTGGCG**GTTTTCAAATGCTTGG**AGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAACGTTACACGAGTGTAACCGCCGAA
CCATGTTGACACGAATTCTG

> Salmonella typhi
GGTCGG**CTTAGACTAGTGTGAC**CAAAAAGCTTTTGCTGAAGTTTCAGGGTAAGAAGAACC
AGCTCCTAGTAAAAAGACTATTGTGACTGAAAAGCGCGTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATGCCGGCGCCGC
TGCTGGAACGCCTCGACCCCAGGGCTACACTA**GTTTAACCGGAACGCC**GCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC



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

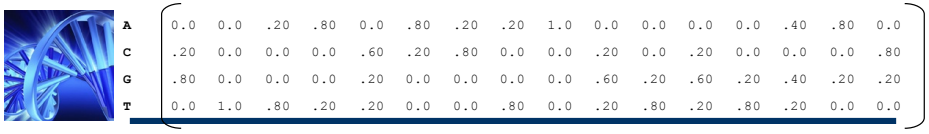
> Escherichia coli
TTGATTCCCTGAATGCCCGCTTAGT**GTAACACTACTGTAAC**CGGCATTTTCTGCTTTTCC
TGCCGATATTTTTTCTTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCCG
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
ATCTAACGGTACGGATTCTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCGGCATAAGTAGTAAGTGCCGTGCCTCCGCCATT
A**GTTACACTAGTGGGAC**ACC

> Vibrio cholerae
ATTTGTGGCG**GTTTTCAAATGCTTGG**AGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAACGTTACACGAGTGTAACCGCCGAA
CCATGTTGACACGAATTCTG

> Salmonella typhi
GGTCGG**CTTAGACTAGTGTGAC**CAAAAAGCTTTTGCTGAAGTTTCAGGGTAAGAAGAACC
AGCTCCTAGTAAAAAGACTATTGTGACTGAAAAGCGCGTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATGCCGGCGCCGC
TGCTGGAACGCCTCGACCCCAGG**GCTACACTAGTTTAAC**CGGAACGCCGCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC



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

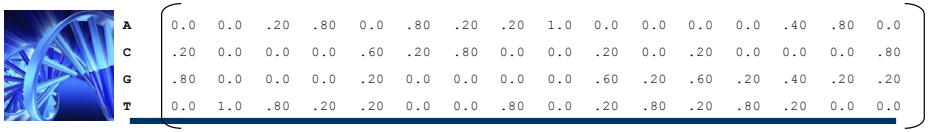
> Escherichia coli
TTGATTCCCTGAATGCCCGCTTAGT**GTAACACTACTGTAAC**CGGCATTTTCTGCTTTTCC
TGCCGATATTTTTTCTTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCCG
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
ATCTAACGGTACGGATTCTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCGGCATAAGTAGTAAGTGCCGTGCCTCCGCCATT
A**GTTACACTAGTGGGAC**ACC

> Vibrio cholerae
ATTTGTGGCG**GTTTTCAAATGCTTGG**AGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAACGTTACACGAGTGTAACCGCCGAA
CCATGTTGACACGAATTCTG

> Salmonella typhi
GGTCGG**CTTAGACTAGTGTGAC**CAAAAAGCTTTTGCTGAAGTTTCAGGGTAAGAAGAACC
AGCTCCTAGTAAAAAGACTATTGTGACTGAAAAGCGCGTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATGCCGGCGCCGC
TGCTGGAACGCCTCGACCCCAGG**GCTACACTAGTTTAAC**CGGAACGCCGCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC



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

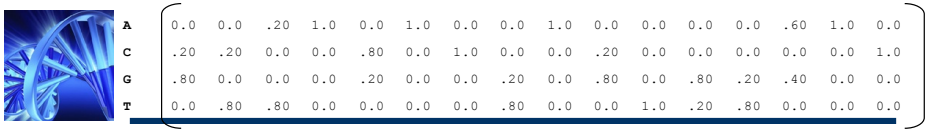
> Escherichia coli
TTGATTCCCTGAATGCCCGCTTAGT**GTAACACTACTGTAAC**CGGCATTTTCTGCTTTTCC
TGCCGATATTTTTTCTTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCCG
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
ATCTAACGGTACGGATTCTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCGGCATAAGTAGTAAGTGCCGTGCCTCCGCCATT
A**GTTACACTAGTGGGAC**ACC

> Vibrio cholerae
ATTTGTGGCGGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAAC**GTTACACGAGTGTAAC**CGCCGAA
CCATGTTGACACGAATTCTG

> Salmonella typhi
GGTCGG**CTTAGACTAGTGTGAC**CAAAAAGCTTTTGCTGAAGTTTCAGGGTAAGAAGAACC
AGCTCCTAGTAAAAAGACTATTGTGACTGAAAAGCGCGTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATGCCGGCGCCGC
TGCTGGAACGCCTCGACCCCAGG**GCTACACTAGTTTAAC**CGGAACGCCGCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC



A	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
C	.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
G	.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
T	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
TTGATTCCCTGAATGCCCGCTTAGT**GTAACACTACTGTAAC**CGGCATTTTCTGCTTTTCC
TGCCGATATTTTTTCTTATCTACCTCACAAAGGTTAGCAATAACTGCTGGGAAAATTCCG
AGTTAGTCGTTATATTCTAT

> Haemophilus influenzae
ATCTAACGGTACGGATTCTCCAAAGGCCTATGGAATCTTGTAGAATATGAAACGTTCTAA
TAAATCATAAAGTTGGAGCAAACGCTCGGCATAAGTAGTAAGTGCCGTGCCTCCGCCATT
A**GTTACACTAGTGGGAC**ACC

> Vibrio cholerae
ATTTGTGGCGGTTTTCAAATGCTTGGAGAATGGGTACATGATCCGCTTGGCATTGAAGGT
GAGGCTGGCAGCAGCGAAGGTCTGGGGCTGTTTGAAC**GTTACACGAGTGTAAC**CGCCGAA
CCATGTTGACACGAATTCTG

> Salmonella typhi
GGTCGG**CTTAGACTAGTGTGAC**CAAAAAGCTTTTGCTGAAGTTTCAGGGTAAGAAGAACC
AGCTCCTAGTAAAAAGACTATTGTGACTGAAAAGCGCGTCAGCGCAAAGCCGACCGCACA
AAACGCACAAGGAGTTACAG

> Pseudomonas aeruginosa
ACGCGGCCAGGGTCTTCTCCTGCGAGATCATGCGCGGCGCGCCGCGCATGCCGGCGCCGC
TGCTGGAACGCCTCGACCCCAGG**GCTACACTAGTTTAAC**CGGAACGCCGCCAGTGGATCG
GCCTGCCCCAGCTATTGCTC

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

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

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

A	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
C	.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
G	.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
T	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 = $8.0 * 10^{-27}$

TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

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

A	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
C	.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
G	.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
T	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 = $5.12 * 10^{-22}$

TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

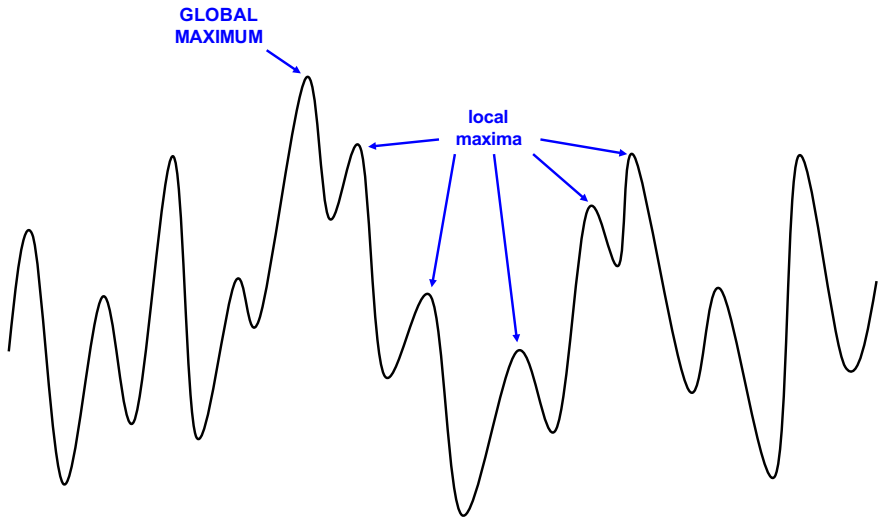
We Could Determine the Location of the Motif Instance That Best Matches the Model…

A	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
C	.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
G	.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
T	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 = $7.16 * 10^{-28}$

TTGATTCCCTGAATGCCCGCTTAGTGTAACACTACTGTAA

Exponential Search Space



How Do You Sample?