7.91 / 20.490 / 6.874 / HST.506

7.36 / 20.390 / 6.802

C. Burge Lecture #9

Mar. 6, 2014

Modeling & Discovery of Sequence Motifs

Modeling & Discovery of Sequence Motifs

- Motif Discovery with Gibbs Sampling Algorithm
- Information Content of a Motif
- Parameter Estimation for Motif Models (+ others)

Background for today:

NBT Primers on Motifs, Motif Discovery. Z&B Ch. 6.

Optional: Lawrence Gibbs paper, Bailey & Elkan MEME paper

For Tuesday: NBT primer on HMMs, Z&B on HMMs (various pp.)

Rabiner tutorial on HMMs

What is a (biomolecular) sequence motif?

A pattern common to a set of DNA, RNA or protein sequences that share a common biological property, such as functioning as binding sites for a particular protein

Ways of representing motifs

- Consensus sequence
- Regular expression
- Weight matrix/PSPM/PSSM
- More complicated models

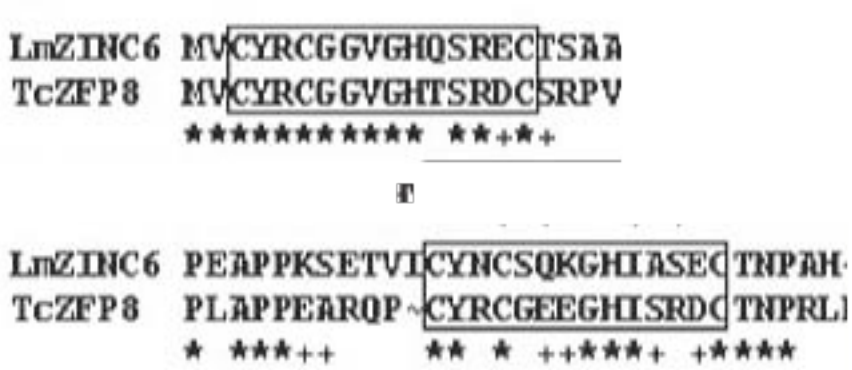
Where do motifs come from?

- Sequences of known common function
- Cross-linking/pulldown experiments
- in vitro binding / SELEX experiments
- Multiple sequence alignments / comparative genomics

Why are they important?

- Identify proteins, DNAs or RNAs that have a specific property
- Can be used to infer which factors regulate which genes
- Important for efforts to model gene expression

Examples of Protein Sequence Motifs

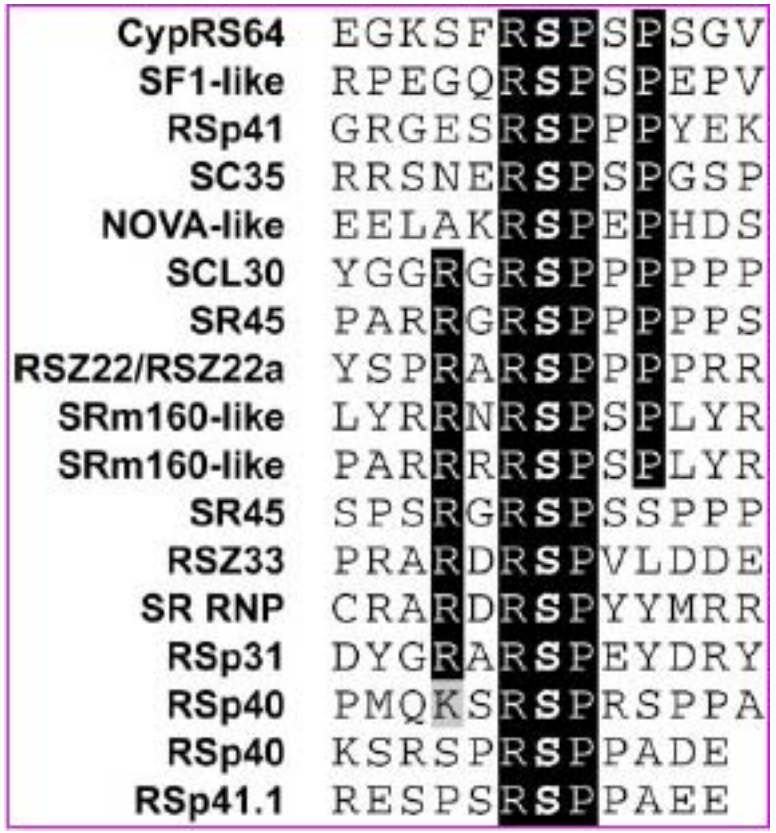


Source: Ericsson, A. O., L. O. Faria, et al. "TcZFP8, A Novel Member of the Trypanosoma Cruzi CCHC Zinc Finger Protein Family with Nuclear Localization." *Genetics and Molecular Research* 5, no. 3 (2006): 553-63.

$$CX_2CX_4HX_4C$$

Zinc finger (DNA binding)

Ericsson et al. Genet. Mol. Res. 2006



Courtesy of the authors. License: CC-BY-NC.
Source: Bentem, Van, Sergio de la Fuente, et al. "Phosphoproteomics Reveals Extensive inVivo Phosphorylation of Arabidopsis Proteins Involved in RNA Metabolism." *Nucleic Acids Research* 34, no. 11 (2006): 3267-78.

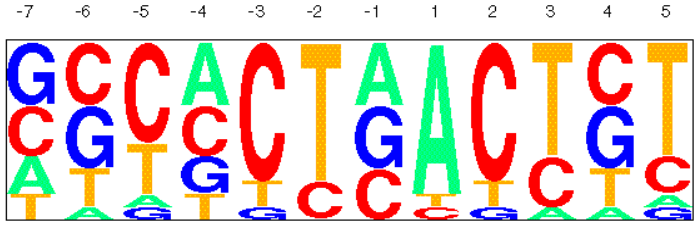
Phosphorylation sites (*Arabidopsis* SRPK4)

de la Fuente van Bentem et al. NAR 2006

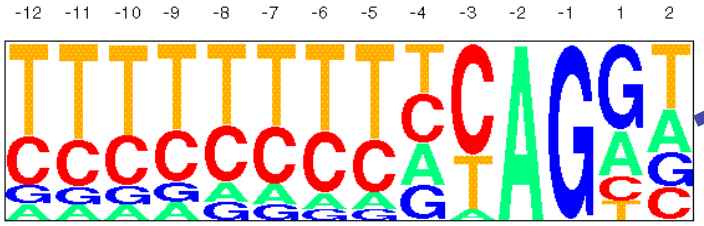
Core Splicing Motifs (Human)

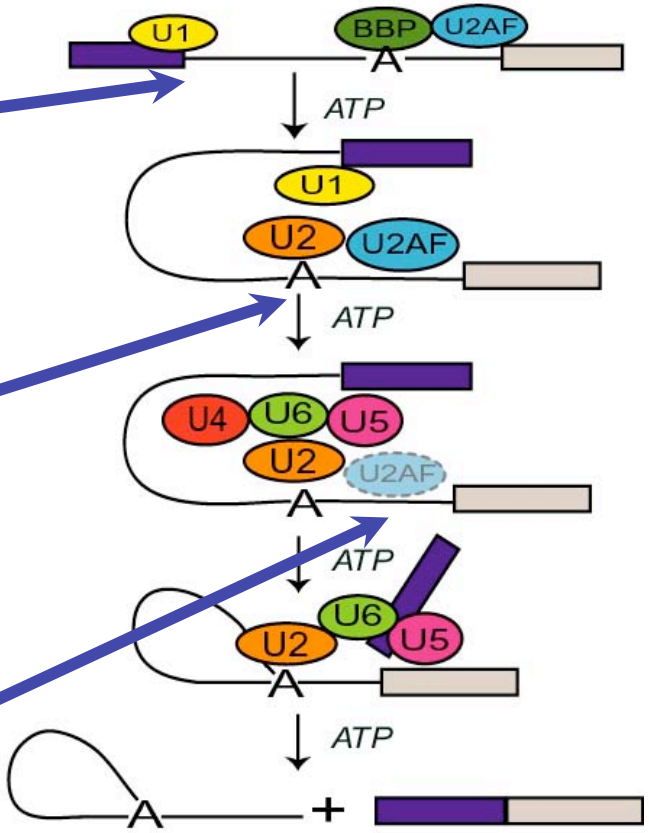
5′ splice site

branch site



3′ splice site





Weight Matrix with Background Model

5′ splice site motif (+)

Con: C A G … G T

Pos	-3	-2	-1	…	+5	+6
A	**0.3**	**0.6**	0.1	…	0.1	0.1
C	**0.4**	0.1	0.0	…	0.1	0.2
G	0.2	0.2	**0.8**	…	**0.8**	0.2
T	0.1	0.1	0.1	…	0.0	**0.5**

Background (-)

Pos	Generic
A	0.25
C	0.25
G	0.25
T	0.25

$$S = S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7\, S_8\, S_9$$

Odds Ratio: $$R = \frac{P(S|+) = P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3) \bullet\bullet\bullet P_5(S_8)P_6(S_9)}{P(S|-) = P_{bg}(S_1)P_{bg}(S_2)P_{bg}(S_3) \bullet\bullet\bullet P_{bg}(S_8)P_{bg}(S_9)}$$

Background model homogenous, assumes independence

Ways to describe a motif

Common motif adjectives:

exact/precise *versus* degenerate

strong *versus* weak (good *versus* lousy)

high information content *versus* low information content

low entropy *versus* high entropy

Statistical (Shannon) Entropy

Motif probabilities: p_k (k = A, C, G, T)

Background probabilities: $q_k = \frac{1}{4}$ (k = A, C, G, T)

$$H(q) = -\sum_{k=1}^{4} q_k \log_2 q_k = ?$$ 2 bits

$$H(p) = -\sum_{k=1}^{4} p_k \log_2 p_k$$ (> or < H(q)?)

Log base 2 gives entropy/information in 'bits'

Relation to Boltzmann entropy: $S = k_B \ln(\Omega)$

Information, uncertainty, entropy

Claude Shannon on what name to give to the “measure of uncertainty” or attenuation in phone-line signals (1949):

“My greatest concern was what to call it. I thought of calling it ‘information’, but the word was overly used, so I decided to call it ‘uncertainty’. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, ‘You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage.”

source: Wikipedia

Information Content of a DNA Motif

Information at position j: $I_j = H_{before} - H_{after}$

Motif probabilities: p_k (k = A, C, G, T)

Background probabilities: $q_k = \frac{1}{4}$ (k = A, C, G, T)

$$I_j = -\sum_{k=1}^{4} q_k \log_2 q_k - -\sum_{k=1}^{4} p_k \log_2 p_k = 2 - H_j$$

If positions in the motif are **independent**, then

$$I_{motif} = \sum_{j=1}^{w} I_j = 2w - H_{motif}$$ (for motif of width w bases)

Otherwise, this relation does not hold in general.

Log base 2 gives entropy/information in 'bits'

The Motif Finding Problem

Unaligned

agggcactagcccatgtgagagggcaaggaccagcggaag
taattcagggccaggatgtatctttctcttaaaaataaca
tatcctacagatgatgaatgcaaatcagcgtcacgagctt
tggcgggcaaggtgcttaaaagataatatcgaccctagcg
attcgggtaccgttcataaaagtacgggaatttcgggtag
gttatgttaggcgagggcaaaagtcatatacttttaggtc
aagagggcaatgcctcctctgccgattcggcgagtgatcg
gatggggaaaatatgagaccaggggagggccacactgcag
ctgccgggctaacagacacacgtctagggctgtgaaatct
gtaggcgccgaggccaacgctgagtgtcgatgttgagaac
attagtccggttccaagagggcaactttgtatgcaccgcc
gcggcccagtgcgcaacgcacagggcaaggtttactgcgg
ccacatgcgagggcaacctccctgtgttgggcggttctga
gcaattgtaaaacgacggcaatgttcggtcgcctaccctg
gataaagaggggggtaggaggtcaactcttccgtattaat
aggagtagagtagtgggtaaactacgaatgcttataacat
gcgagggcaatcgggatctgaaccttctttatgcgaagac
tccaggaggaggtcaacgactctgcatgtctgacaacttg
gtcatagaattccatccgccacgcggggtaatttggacgt
gtgccaacttgtgccgggggggctagcagcttcccgtcaaa
cgcgtttggagtgcaaacatacacagcccgggaatataga
aagatacgagttcgatttcaagagttcaaaacgtgacggg
gacgaaacgagggcgatcaatgcccgataggactaataag
tagtacaaacccgctcacccgaaaggagggcaaatacctt
atatacagccaggggagacctataactcagcaaggttcag
cgtatgtactaattgtggagagcaaatcattgtccacgtg
...

Aligned

gcggaagagggcactagcccatgtgagagggcaaggacca
atctttctcttaaaaataacataattcagggccaggatgt
gtcacgagctttatcctacagatgatgaatgcaaatcagc
taaaagataatatcgaccctagcgtggcgggcaaggtgct
gtagattcgggtaccgttcataaaagtacgggaatttcgg
tatactttaggtcgttatgttaggcgagggcaaaagtca
ctctgccgattcggcgagtgatcgaagagggcaatgcctc
aggatggggaaaatatgagaccaggggagggccacactgc
acacgtctagggctgtgaaatctctgccgggctaacagac
gtgtcgatgttgagaacgtaggcgccgaggccaacgctga
atgcaccgccattagtccggttccaagagggcaactttgt
ctgcgggcggcccagtgcgcaacgcacagggcaaggttta
tgtgttgggcggttctgaccacatgcgagggcaacctccc
gtcgcctaccctggcaattgtaaaacgacggcaatgttcg
cgtattaatgataaagagggggtaggaggtcaactcttc
aatgcttataacataggagtagagtagtgggtaaactacg
tctgaaccttctttatgcgaagacgcgagggcaatcggga
tgcatgtctgacaacttgtccaggaggaggtcaacgactc
cgtgtcatagaattccatccgccacgcggggtaatttgga
tcccgtcaaagtgccaacttgtgccgggggggctagcagct
acagcccgggaatatagacgcgtttggagtgcaaacatac
acgggaagatacgagttcgatttcaagagttcaaaacgtg
cccgataggactaataaggacgaaacgagggcgatcaatg
ttagtacaaacccgctcacccgaaaggagggcaaatacct
agcaaggttcagatatacagccaggggagacctataactc
gtccacgtgcgtatgtactaattgtggagagcaaatcatt
...

...can be posed as an alignment problem

Approaches to Motif Finding

- Enumerative ('dictionary')

- search for a *k*mer/set of *k*mers/regular expression that is statistically over-represented

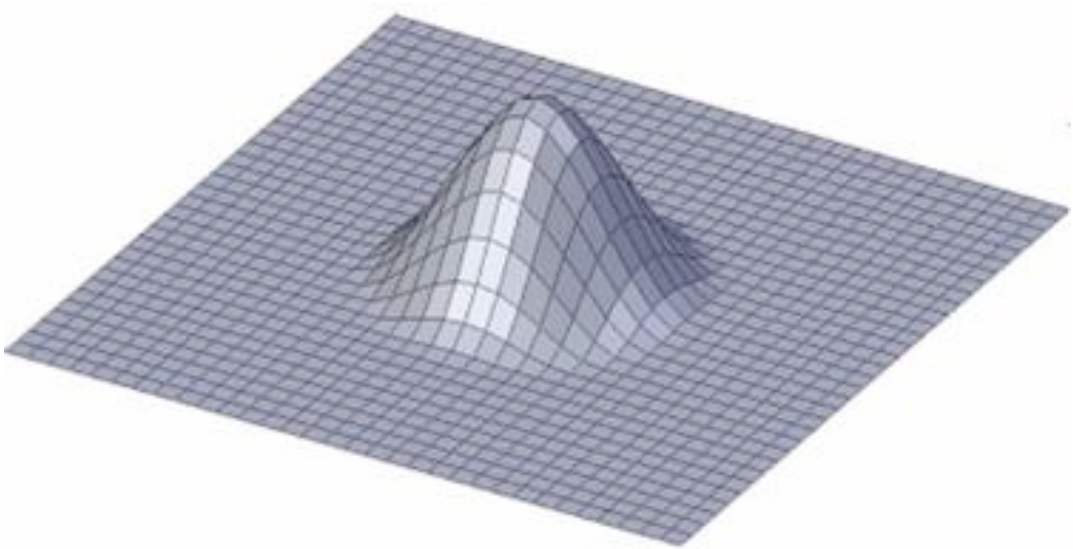
- Probabilistic Optimization (e.g., Gibbs sampler)

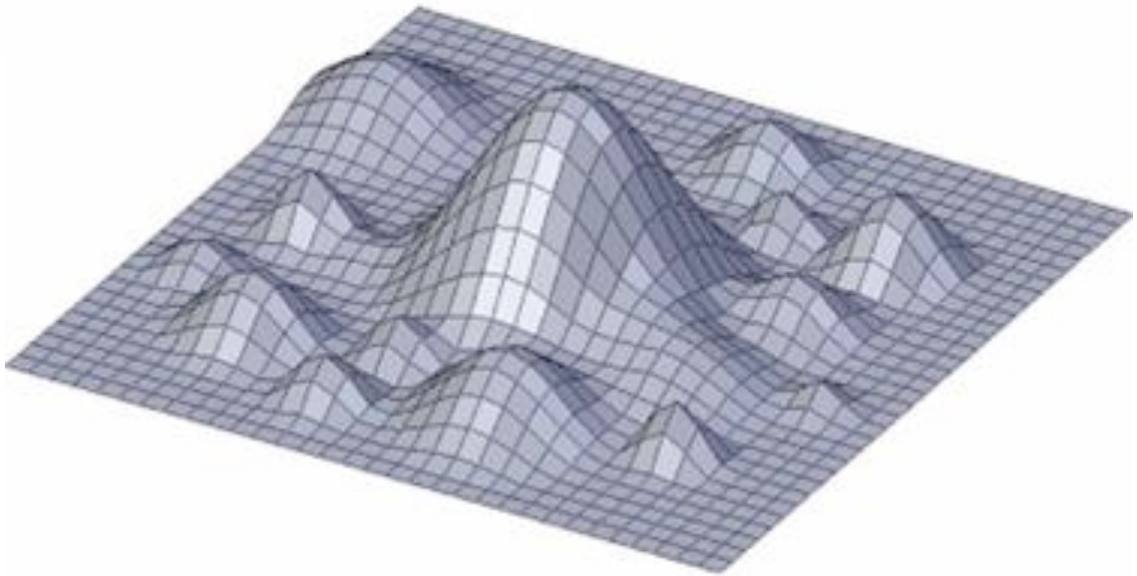
- stochastic search of the space of possible PSPMs

- Deterministic Optimization (e.g., MEME)

  - deterministic search of space of possible PSPMs

What the motif landscape might look like





Monte Carlo Algorithms

The Gibbs motif sampler is a **Monte-Carlo algorithm**

Photograph of people playing craps removed due to copyright restrictions.

General definition: class of computational algorithms that rely on repeated random sampling to compute their results

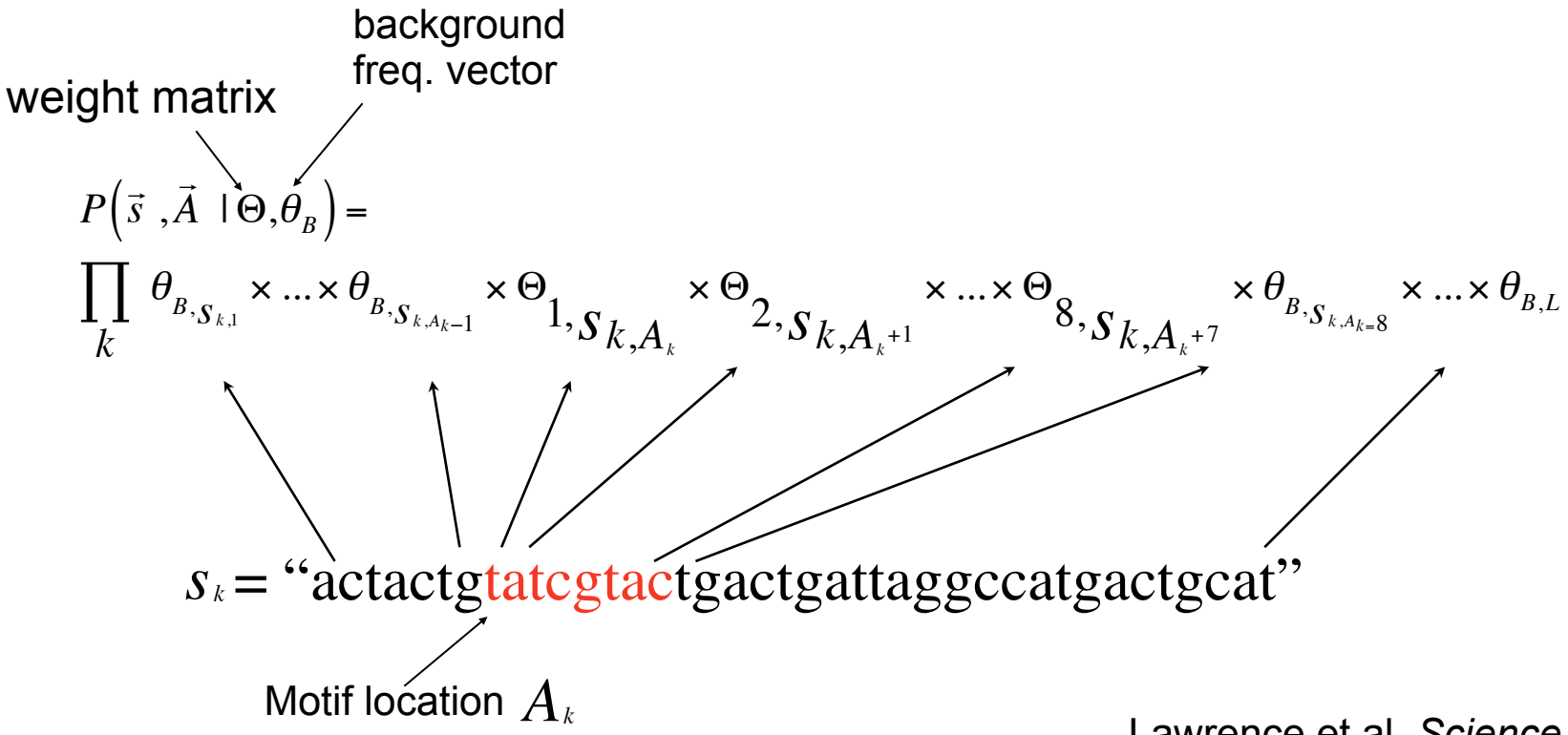
Specific definition: randomized algorithm where the computational resources used are bounded but the answer is not guaranteed to be correct 100% of the time

Related to

Las Vegas algorithm - a randomized algorithm that always gives correct results (or informs about failure)

Example: The Gibbs Motif Sampler

The likelihood function for a set of sequences $\vec{s}$ with motif locations $\vec{A}$



Lawrence et al. *Science* 1993

The Gibbs Sampling Algorithm In Words I

Given **N** sequences of length **L** and desired motif width **W:**

1) Choose a starting position in each sequence at random:

   $\mathbf{a_1}$ in seq 1, $\mathbf{a_2}$ in seq 2, …, $\mathbf{a_N}$ in sequence **N**

2) Choose a sequence at random from the set (say, seq 1).

3) Make a weight matrix model of width **W** from the sites in all sequences *except* the one chosen in step 2.

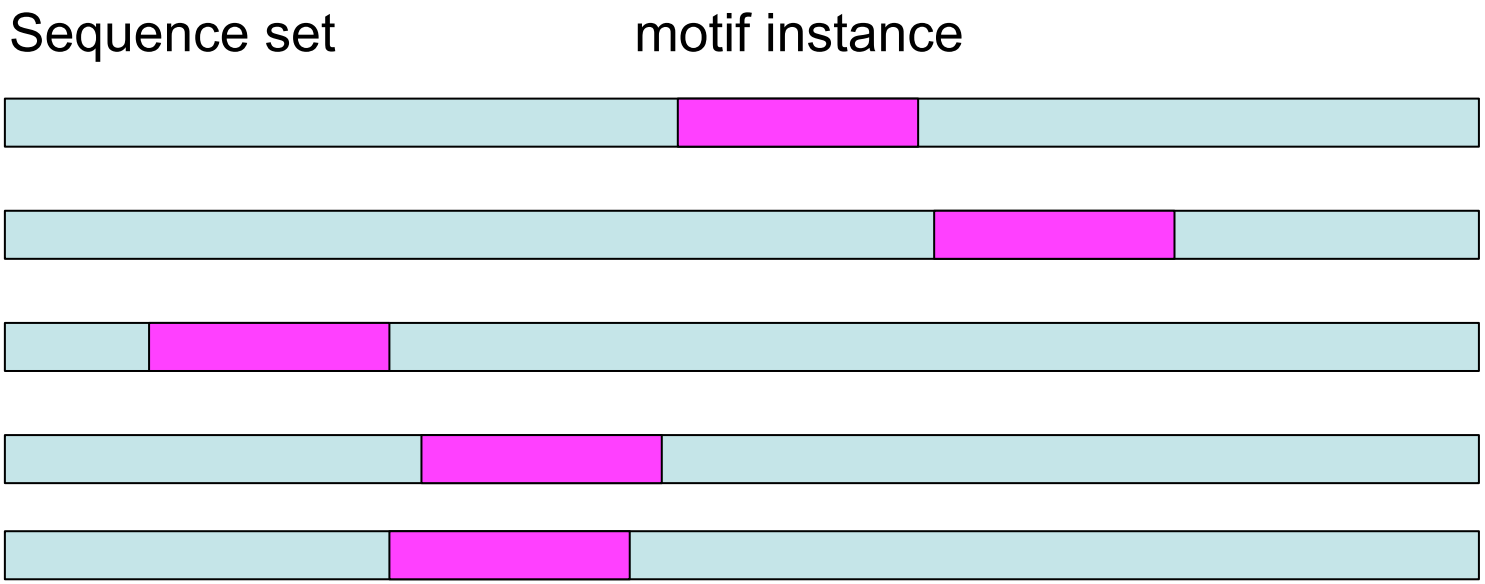
4) Assign a probability to each position in seq 1 using the weight matrix model constructed in step 3:

   $\mathbf{p} = \{ \mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}, \ldots, \mathbf{p_{L-W+1}} \}$

Lawrence et al. *Science* 1993

Gibbs Sampling Algorithm I

1. Select a random position in each sequence

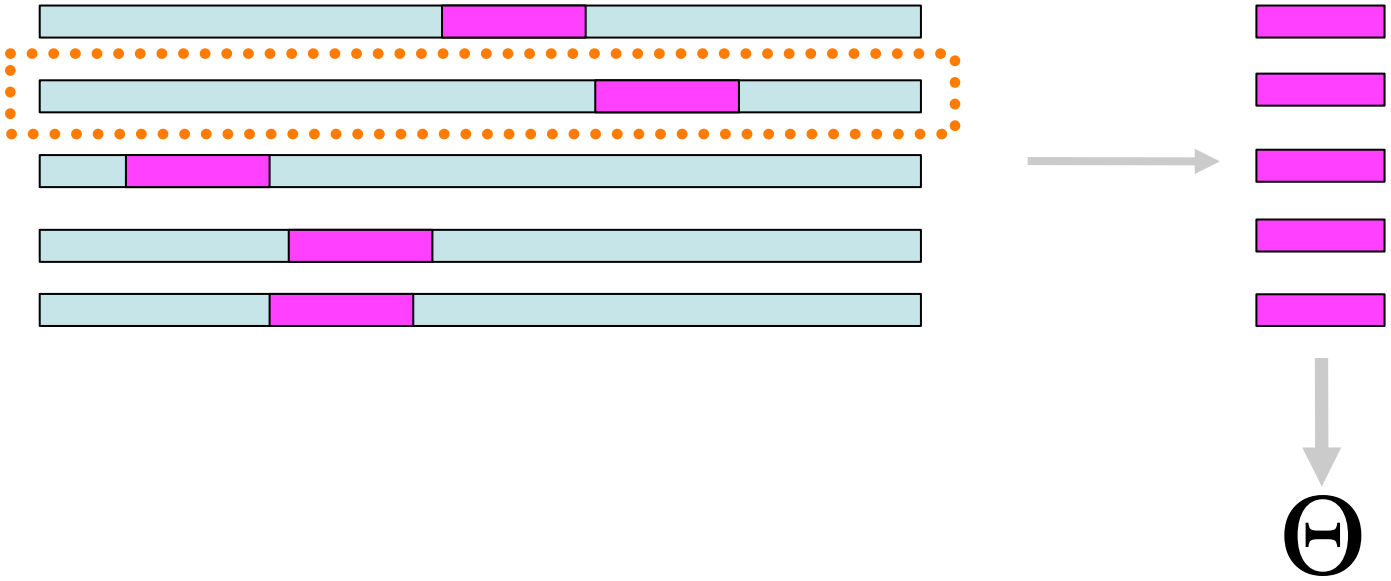


Gibbs Sampling Algorithm II

2. Build a weight matrix

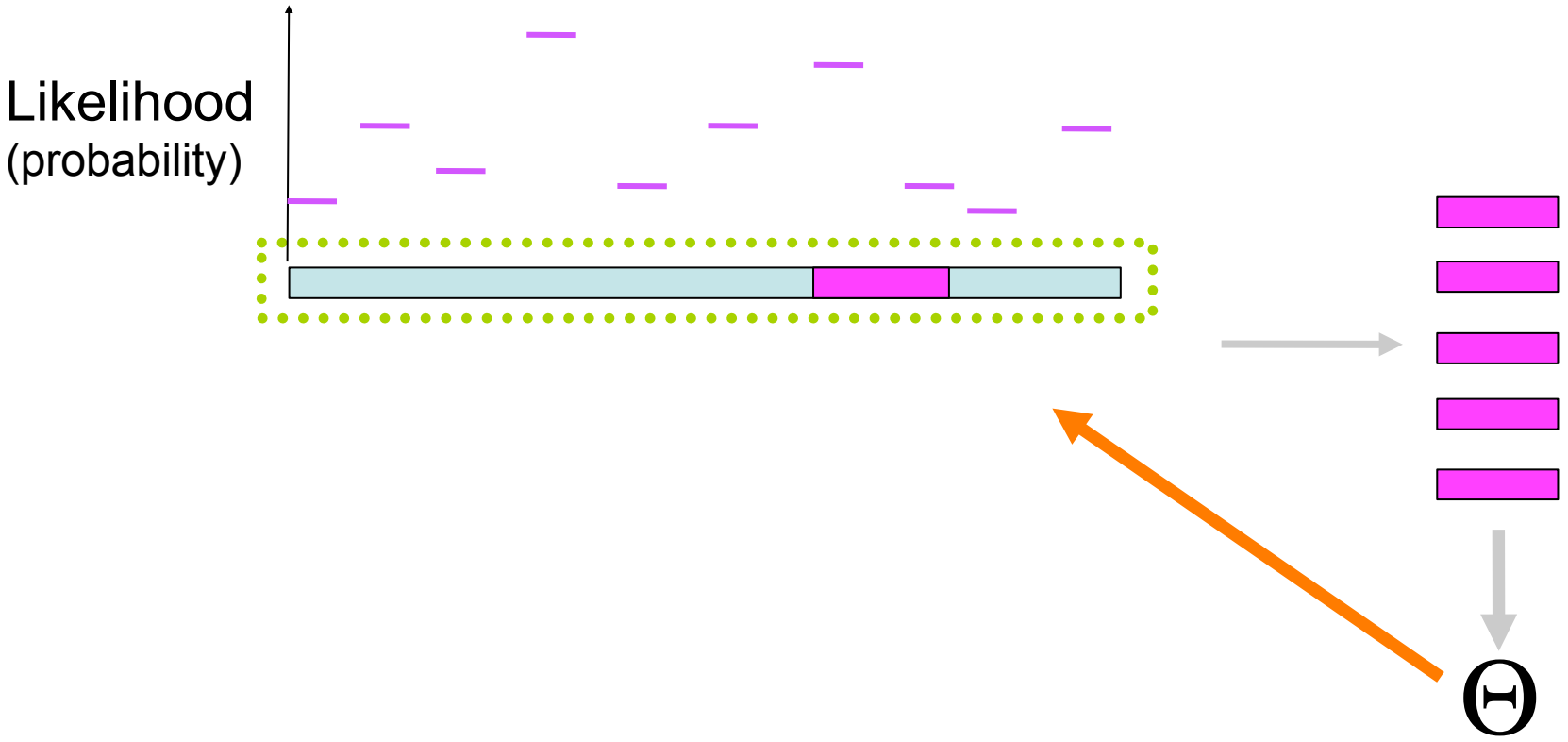
Gibbs Sampling Algorithm III

3. Select a sequence at random



Gibbs Sampling Algorithm IV

4. Score possible sites in the sequence using weight matrix



The Gibbs Sampling Algorithm In Words, II

Given **N** sequences of length **L** and desired motif width **W:**

5) Sample a starting position in seq 1 based on this probability distribution and set $\mathbf{a_1}$ to this new position.

6) Choose a sequence at random from the set (say, seq 2).

7) Make a weight matrix model of width **W** from the sites in all sequences *except* the one chosen in step 6.

8) Assign a probability to each position in seq 2 using the weight matrix model constructed in step 7.

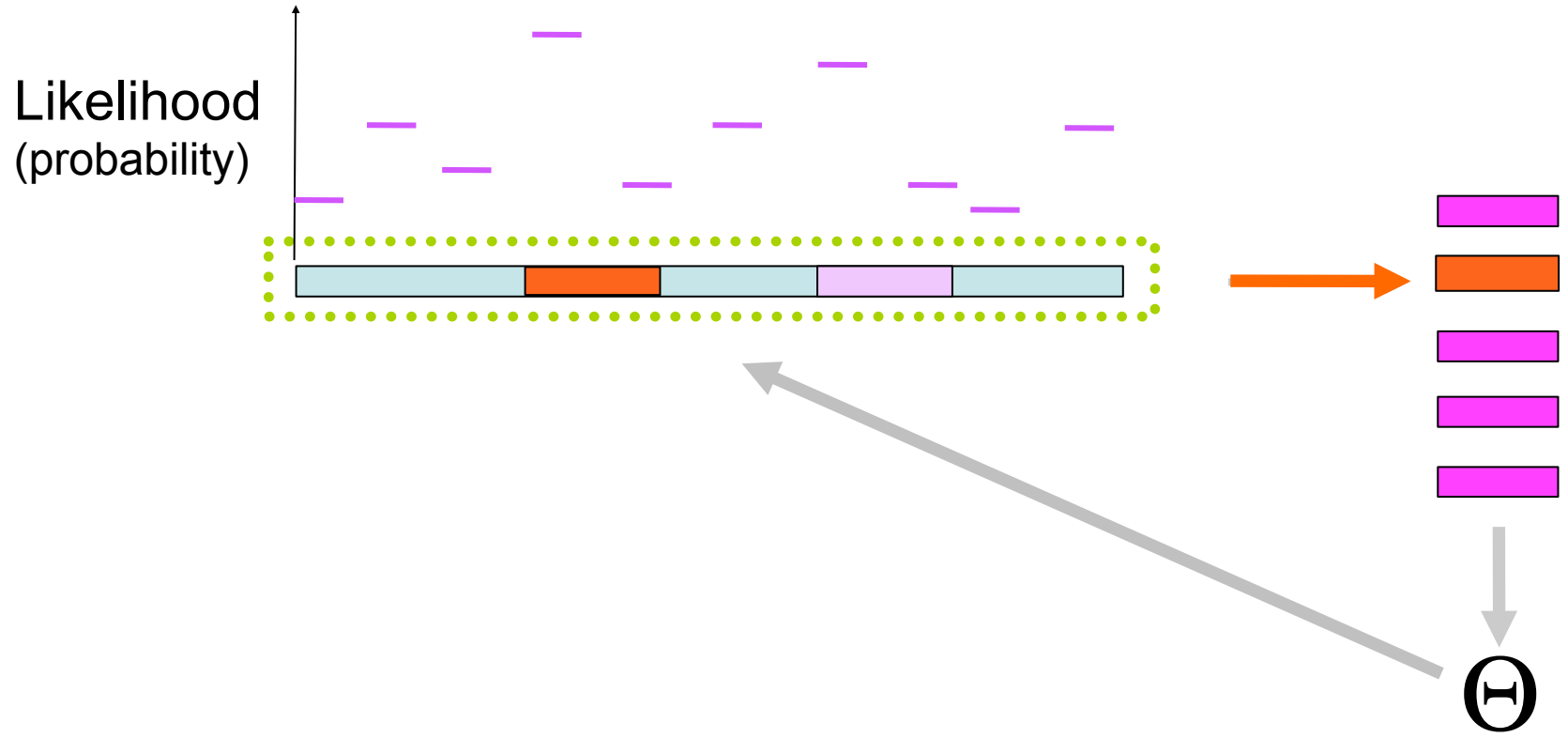
Step 9) Sample a starting position in seq 2 based on this dist.

Step 10) Repeat until convergence (of positions or motif model)

Lawrence et al. *Science* 1993

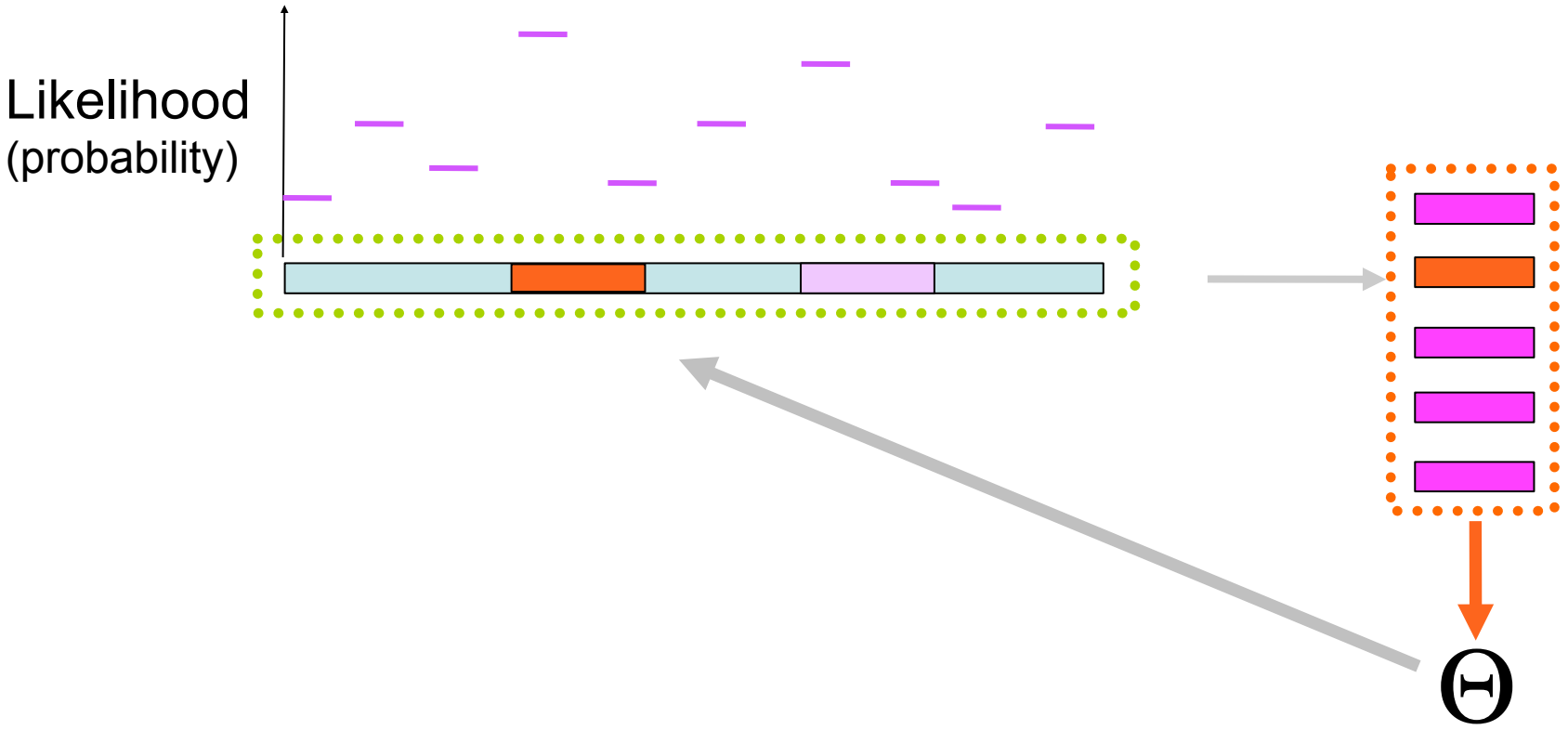
Gibbs Sampling Algorithm V

5. Sample a new site proportional to likelihood and update motif instances



Gibbs Sampling Algorithm VI

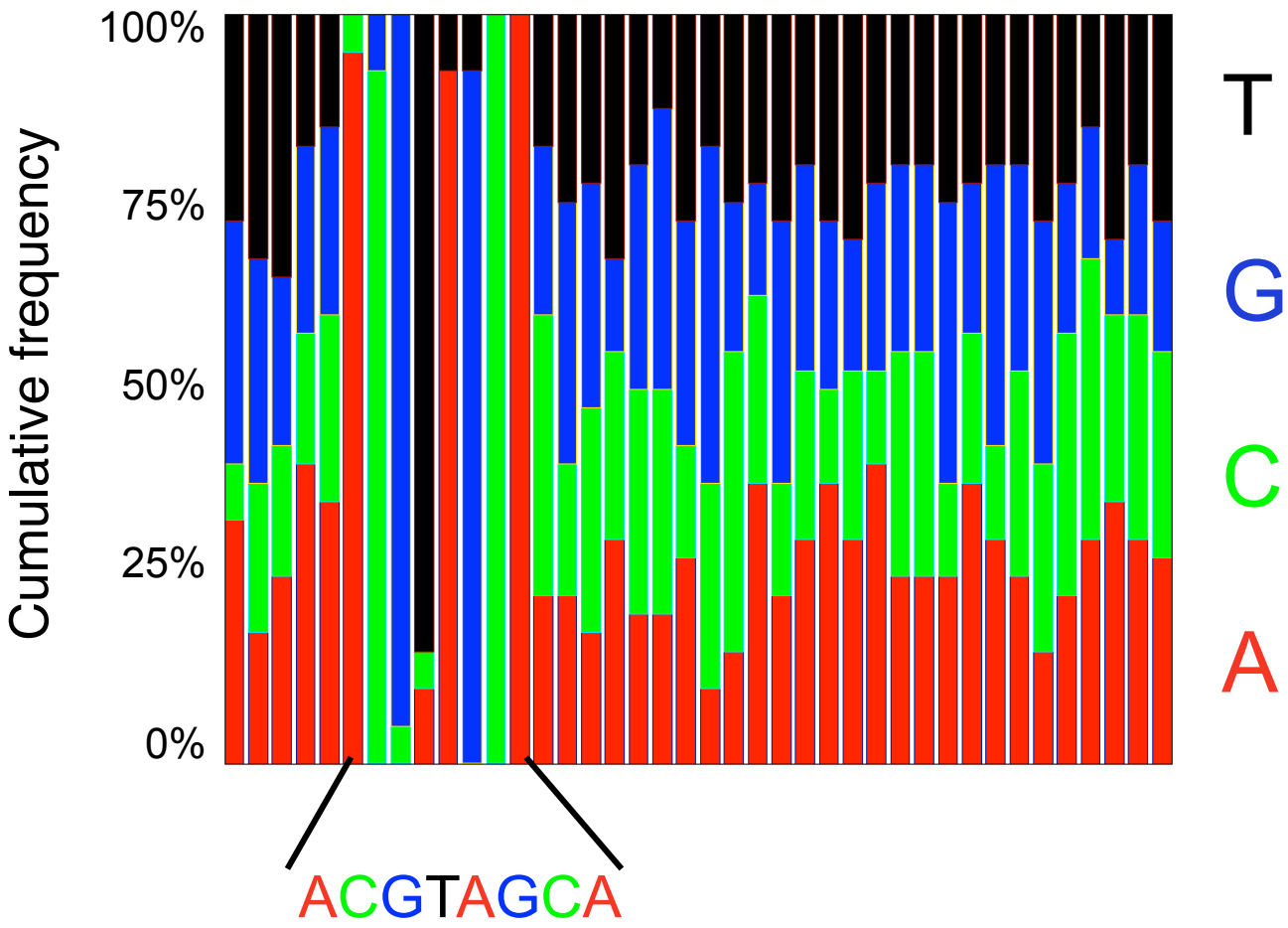
6. Update weight matrix



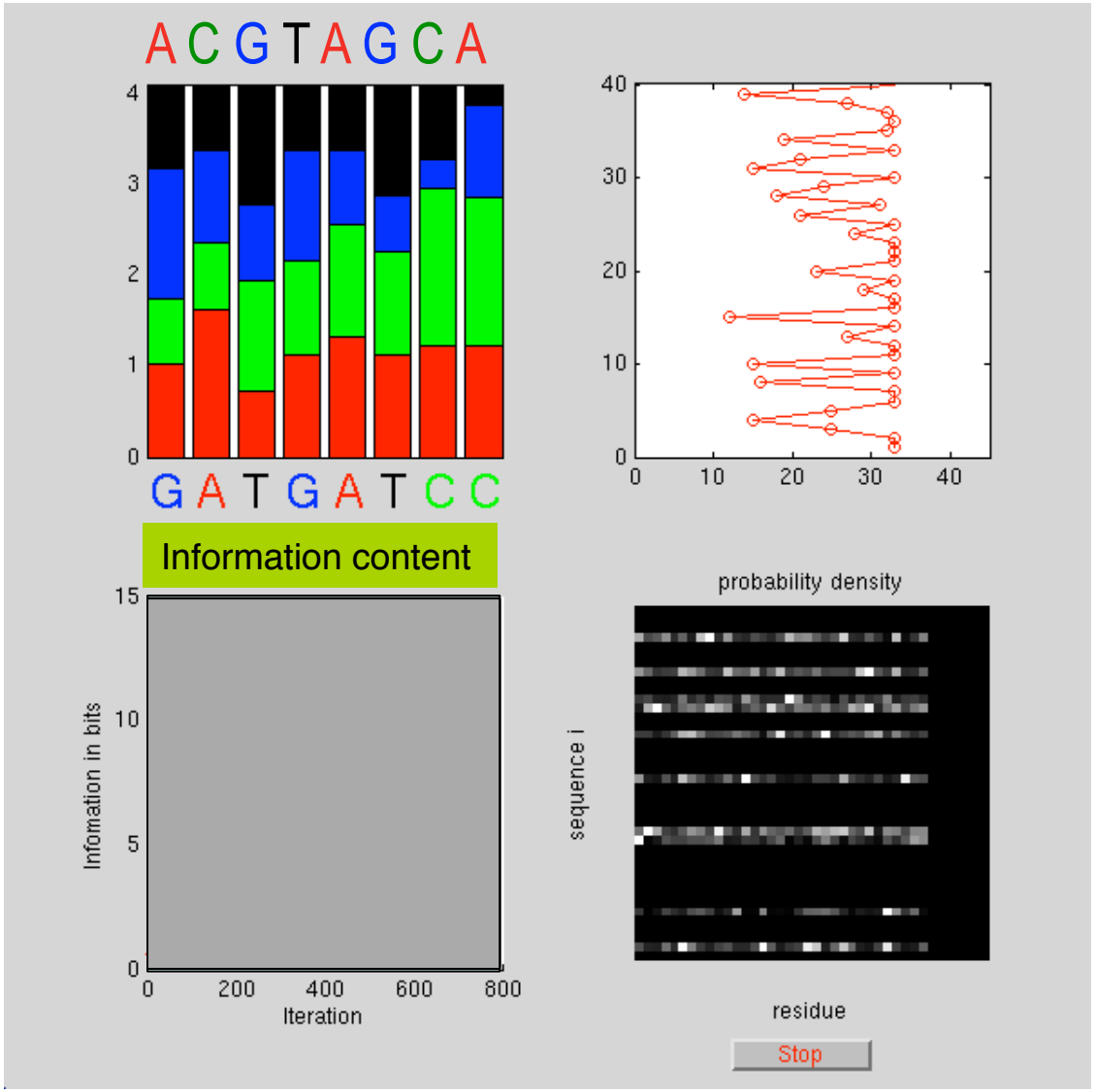
Gibbs Sampling Algorithm VII

7. Iterate until convergence ($\triangle$sites = 0 or $\triangle\Theta \sim 0$)

Input Sequences with Strong Motif



Gibbs Sampler - Strong Motif Example



Current weight matrix

Motif strength

Current Position in seq

Probability density

Gibbs sampler animation by M. Yahyanejad

Iteration

Position in seq

Courtesy of Mehdi Yahyanejad. Used with permission.

Input Sequences (Weak Motif)

```
gcggaagagggcactagcccatgtgagagggcaaggacca
atctttctcttaaaaataacataattcagggccaggatgt
gtcacgagctttatcctacagatgatgaatgcaaatcagc
taaaagataatatcgaccctagcgtggcgggcaaggtgct
gtagattcgggtaccgttcataaaagtacgggaatttcgg
tatacttttaggtcgttatgttaggcgagggcaaaagtca
ctctgccgattcggcgagtgatcgaagagggcaatgcctc
aggatggggaaaatatgagaccaggggagggccacactgc
acacgtctagggctgtgaaatctctgccgggctaacagac
gtgtcgatgttgagaacgtaggcgccgaggccaacgctga
atgcaccgccattagtccggttccaagagggcaactttgt
ctgcgggcggcccagtgcgcaacgcacagggcaaggttta
tgtgttgggcggttctgaccacatgcgagggcaacctccc
gtcgcctaccctggcaattgtaaaacgacggcaatgttcg
cgtattaatgataaagaggggggtaggaggtcaactcttc
aatgcttataacataggagtagagtagtgggtaaactacg
tctgaaccttctttatgcgaagacgcgagggcaatcggga
tgcatgtctgacaacttgtccaggaggaggtcaacgactc
cgtgtcatagaattccatccgccacgcggggtaatttgga
tcccgtcaaagtgccaacttgtgccgggggggctagcagct
acagcccgggaatatagacgcgtttggagtgcaaacatac
acgggaagatacgagttcgatttcaagagttcaaaacgtg
cccgataggactaataaggacgaaacgagggcgatcaatg
ttagtacaaacccgctcacccgaaaggagggcaaatacct
agcaaggttcagatatacagccaggggagacctataactc
gtccacgtgcgtatgtactaattgtggagagcaaatcatt
…
```

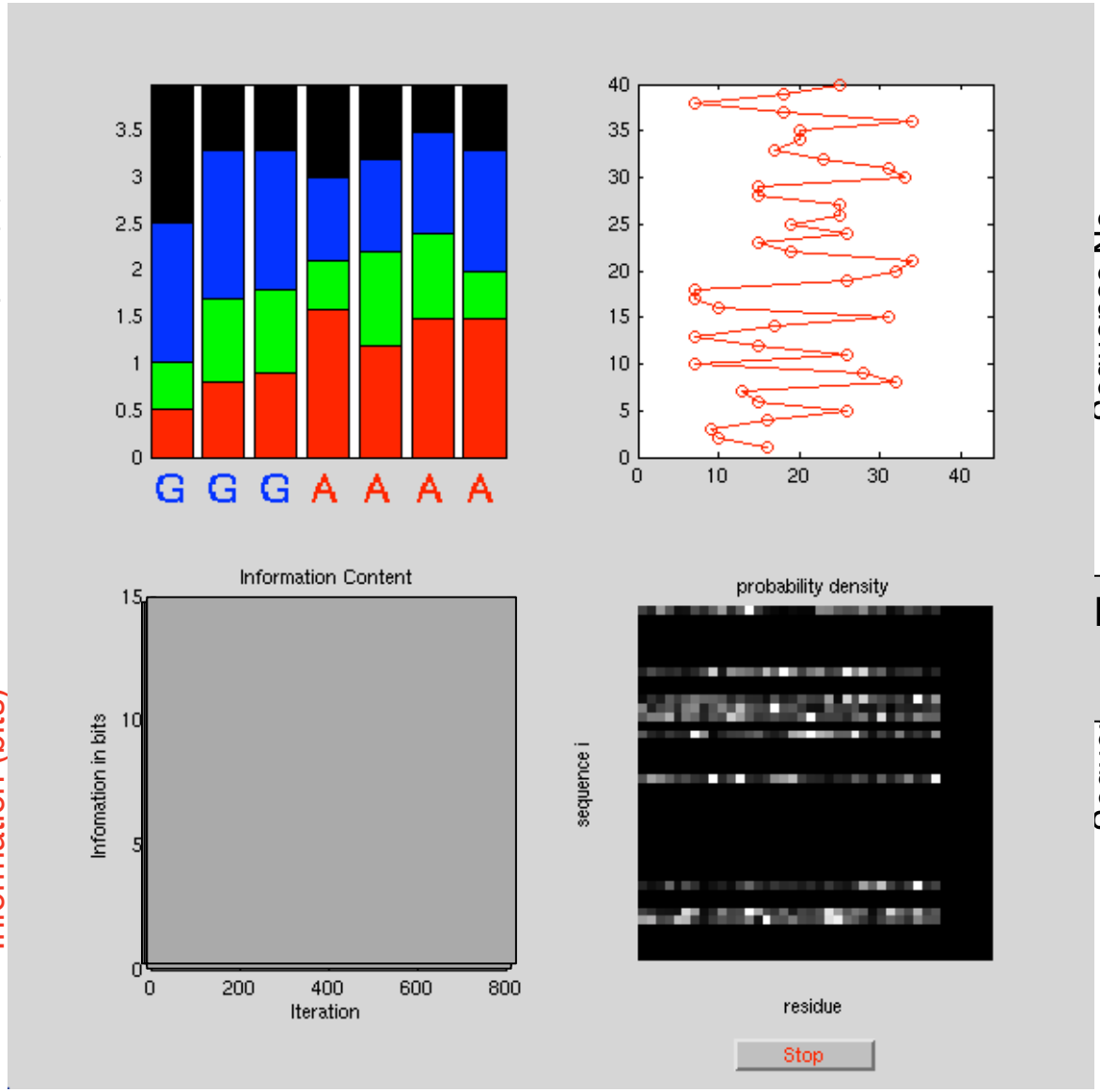
Gibbs Sampler - Weak Motif Example

Current weight matrix

Current Position in seq

Probability density

Motif strength



Gibbs sampler movie by M. Yahyanejad

Iteration Position in seq

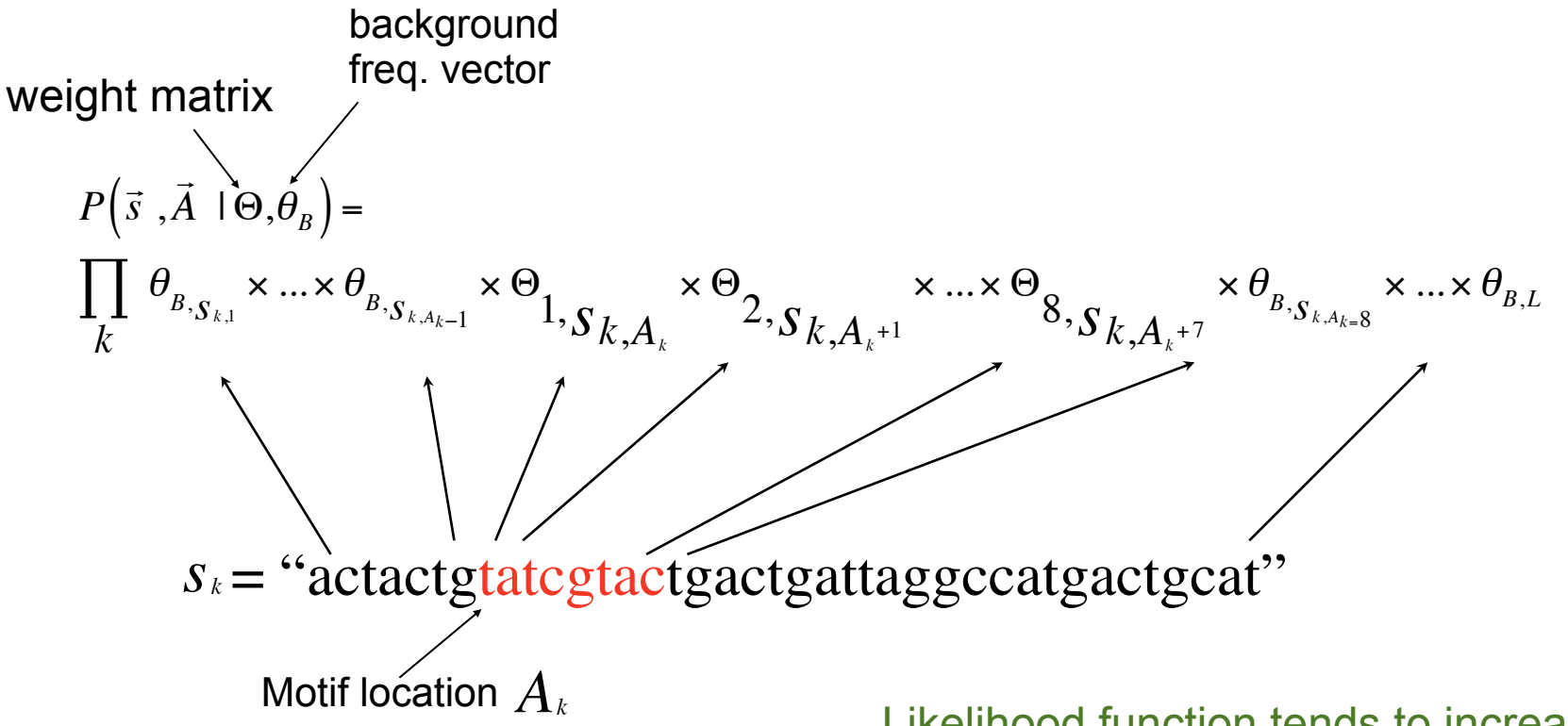
Courtesy of Mehdi Yahyanejad. Used with permission.

Gibbs Sampler Summary

- A stochastic (Monte Carlo) algorithm for motif finding
- Works by 'stumbling' onto a few motif instances, which bias the weight matrix, which causes it to sample more motif instances, which biases the weight matrix more, … until convergence
- Not guaranteed to converge to same motif every time - run several times, compare results
- Works for protein, DNA, RNA motifs

What does this algorithm accomplish?

The likelihood function for a set of sequences $\vec{s}$ with motif locations $\vec{A}$



Likelihood function tends to increase

```
agggcactagcccatgtgagagggcaaggaccagcggaag
taattcagggccaggatgtatctttctcttaaaaataaca
tatcctacagatgatgaatgcaaatcagcgtcacgagctt
tggcgggcaaggtgcttaaaagataatatcgaccctagcg
attcgggtaccgttcataaaagtacgggaatttcgggtag
gttatgttaggcgagggcaaaagtcatatacttttaggtc
aagagggcaatgcctcctctgccgattcggcgagtgatcg
gatggggaaaatatgagaccaggggagggccacactgcag
ctgccgggctaacagacacacgtctagggctgtgaaatct
gtaggcgccgaggccaacgctgagtgtcgatgttgagaac
attagtccggttccaagagggcaactttgtatgcaccgcc
gcggcccagtgcgcaacgcacagggcaaggtttactgcgg
ccacatgcgagggcaacctccctgtgttgggcggttctga
gcaattgtaaaacgacggcaatgttcggtcgcctaccctg
gataaagaggggggtaggaggtcaactcttccgtattaat
aggagtagagtagtgggtaaactacgaatgcttataacat
gcgagggcaatcgggatctgaaccttctttatgcgaagac
tccaggaggaggtcaacgactctgcatgtctgacaacttg
gtcatagaattccatccgccacgcggggtaatttggacgt
gtgccaacttgtgccgggggggctagcagcttcccgtcaaa
cgcgtttggagtgcaaacatacacagcccgggaatataga
aagatacgagttcgatttcaagagttcaaaacgtgacggg
gacgaaacgagggcgatcaatgcccgataggactaataag
tagtacaaacccgctcacccgaaaggagggcaaatacctt
atatacagccaggggagacctataactcagcaaggttcag
cgtatgtactaattgtggagagcaaatcattgtccacgtg
...
```

Features that affect motif finding

No. of sequences

Length of sequences

Information content of motif

Match between expected length and actual length of motif

Motif finding issues

“shifted” motifs

biased background composition

Practical Motif Finding

• MEME is a classic method
Deterministic - like Gibbs, but uses expectation maximization
Bailey & Elkan 1995 paper is posted.
Run MEME at:

http://meme.nbcr.net/meme/

The Fraenkel lab's WebMotifs combines
AlignACE (similar to Gibbs), MDscan, MEME, Weeder, THEME
Described in Romer et al. and references therein

http://fraenkel.mit.edu/webmotifs.html

Mean Log-odds (bit-) Score of a Motif

bit-score: $\log_2(\frac{p_k}{q_k})$ mean bit-score: $\sum_{k=1}^{n} p_k \log_2(\frac{p_k}{q_k})$

motif width w, n = 4^w

If $q_k = \frac{1}{4^w}$ *then* mean bit-score = $2w - H_{motif} = I_{motif}$

What is the use of knowing the information content of a motif?

Rule of thumb*: a motif with m bits of information will occur about once every 2^m bases of random sequence

* Strictly true for regular expressions, approximately true for general motifs

For more on information theory, see: Elements of Information Theory by T. Cover

Relative Entropy*

Relative entropy, D(p||q) = mean bit-score: $\sum_{k=1}^{n} p_k \log_2 (\frac{p_k}{q_k})$

If $q_k = \frac{1}{4^w}$ *then* mean RelEnt = $2w - H_{motif} = I_{motif}$

RelEnt is a measure of **information**, not entropy/uncertainty. In general RelEnt is different from $H_{before} - H_{after}$ and is a better measure when background is non-random

Example: $q_A = q_T = 3/8$, $q_C = q_G = 1/8$

Suppose: $p_C = 1$. $H(q) - H(p) < 2$

But RelEnt $D(p||q) = \log_2(1/(1/8)) = 3$

Which one better describes frequency of C in background seq?

* Alternate names: “Kullback-Leibler distance”, “information for discrimination”

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014