MCMC for Sequence Motif Search

Jianlin Cheng, PhD

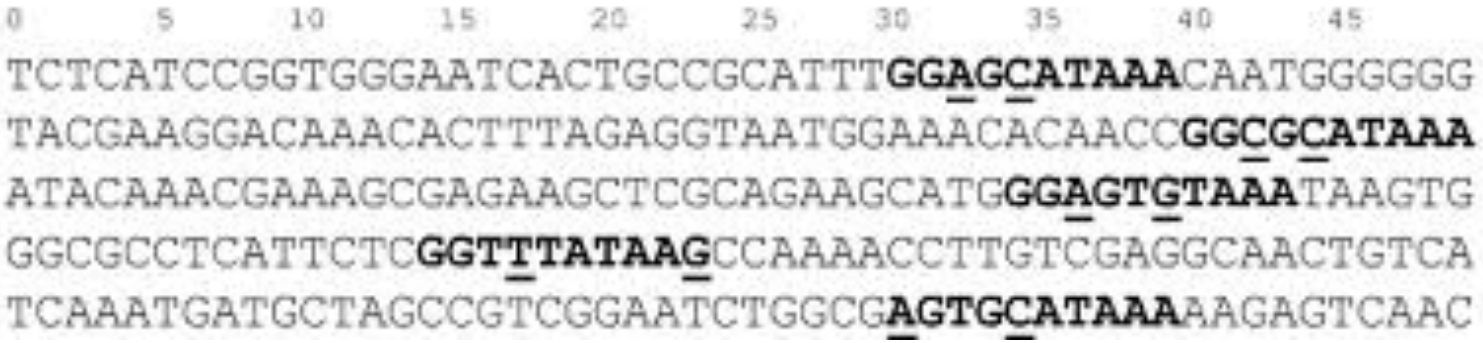
Department of Computer Science
University of Missouri, Columbia

Fall, 2014

Sequence Motif Search Problem

- Find a set of sub-sequences of multiple sequences whose alignment has maximum alignment score (highest similarity).
- It is NP-hard.
- Biologically find highly conserved regions (motifs) of related genes or a protein family

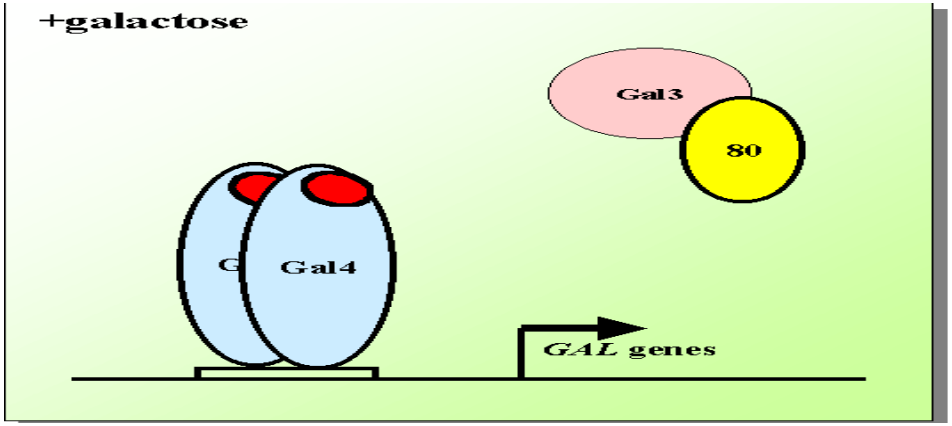
A Motif Example



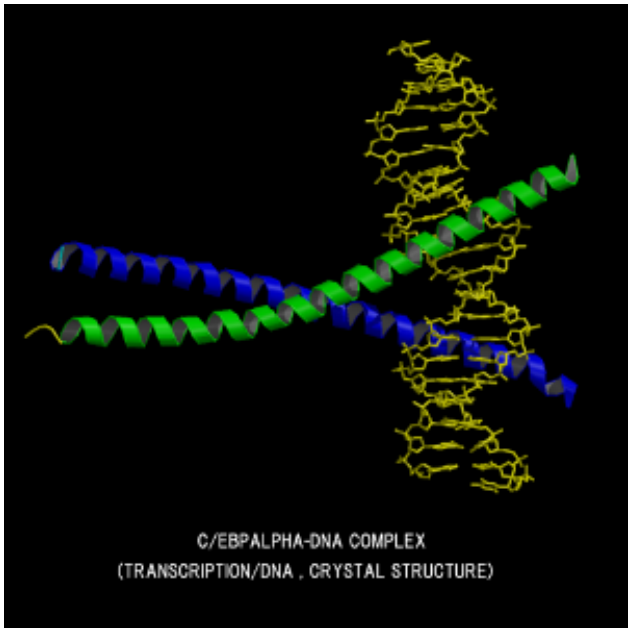
GGAGCATAAA
GGCGCATAAA
GGAGTGTAAA
GGTTTATAAG
AGTGCATAAA

Examples: Transcription Factors

- yeast: Gal4
- drosophila
- mammal




```
1: actcgtcggggcgtacgtacgtaacgtacgtaCGGACAACTGTTGACCG
2: cggagcactgttgagcgacaagtaCGGAGCACTGTTGAGCGgtacgtac
3: ccccgtaggCGGCGCACTCTCGCCCGggcgtacgtacgtaacgtacgta
4: agggcgcgtacgctaccgtcgacgtcgCGCGCCGCACTGCTCCGacgct
```



Kathrina Kechris, 2005

Motif Model

Data: Upstream sequences from co-regulated/co-expressed genes.
Assumption: Binding site occurs in most sequences

```
1: actcgtcggggcgtacgtacgtaacgtacgtacggacaactgttgaccg
2: cggagcactgttgagcgacaagtacggagcactgttgagcggtacgtac
3: ccccgtaggcggcgcactctcgcccgggcgtacgtacgtaacgtacgta
4: agggcgcgtacgctaccgtcgacgtcgcgcgccgcactgctccgacgct
```

Goals: 1) Estimate motif
2) Predict motif locations

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	0	0	0	$\frac{3}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$	0	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0	0	0
C	$\frac{4}{4}$	0	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0	$\frac{1}{4}$	0	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{4}{4}$	0
G	0	$\frac{4}{4}$	$\frac{4}{4}$	0	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0	0	$\frac{3}{4}$	0	0	$\frac{3}{4}$	0	$\frac{1}{4}$	0	$\frac{4}{4}$
T	0	0	0	0	0	0	0	$\frac{1}{2}$	$\frac{3}{4}$	0	$\frac{3}{4}$	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0	0

```
1: actcgtcggggcgtacgtacgtaacgtacgtaCGGACAACTGTTGACCG
2: cggagcactgttgagcgacaagtaCGGAGCACTGTTGAGCGgtacgtac
3: ccccgtaggCGGCGCACTCTCGCCCGggcgtacgtacgtaacgtacgta
4: agggcgcgtacgctaccgtcgacgtcgCGCGCCGCACTGCTCCGacgct
```

Kathrina Kechris, 2005

Initialization of Locations (State)

actcgtcggggcgtacgtacgtaacgtacgta**CGGACAACTGTTGACCG**
cggagcactgttgagcgacaagta**CGGAGCACTGTTGAGCG**gtacgtac
ccccgtagg**CGGCGCACTCTCGCCCG**ggcgtacgtacgtaacgtacgta
agggcgcgtacgctaccgtcgacgtcg**CGCGCCGCACTGCTCCG**acgct

$S = \langle a_1, a_2, a_3, a_4 \rangle$

Motif Size: 17

Construct a probability matrix (profile)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	1/4	1/4	.	.	.	.											
C	2/4	1/4															
G	0	2/4															
T	1/4	0															

Probability of a Position and a State given a Motif Model

i

actcgtcggggcgtacgtacgtaacgtacgta**CGGACAACTGTTGACCG**
cggagcactgttgagcgacaagta**CGGAGCACTGTTGAGCG**gtacgtac
ccccgtagg**CGGCGCACTCTCGCCCG**ggcgtacgtacgtaacgtacgta
agggcgcgtacgctaccgtcgacgtcg**CGCGCCGCACTGCTCCG**acgct

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	1/4	1/4	.	.	.	.											
C	2/4	1/4															
G	0	2/4															
T	1/4	0															

Objective: Find the best position in each sequence that maximize product of probability

Prob (posi_i) = 2/4 * 2/4 * ….

Position Resampling

actcgtcggggcgtacgtacgtaacgtacgta**CGGACAACTGTTGACCG**
cggagcactgttgagcgacaagta**CGGAGCACTGTTGAGCG**gtacgtac
ccccgtagg**CGGCGCACTCTCGCCCG**ggcgtacgtacgtaacgtacgta
agggcgcgtacgctaccgtcgacgtcg**CGCGCCGCACTGCTCCG**acgct

Construct a probability matrix (profile) from new positions

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	0	0	.	.	.	.											
C	4/4	0															
G	0	4/4															
T	0	0															

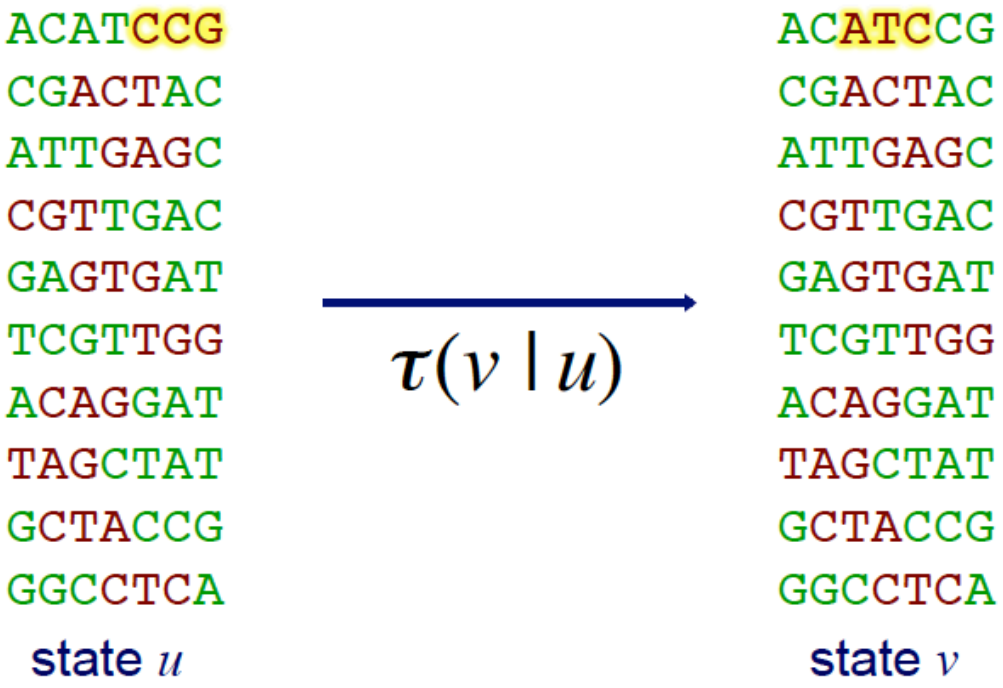
Challenges

- Don’t know the motif model (probability matrix)
- Don’t know the locations of each sub sequence

Markov Chain Monte Carlo (MCMC)

- we can view the motif finding approach in terms of a Markov chain
- each state represents a configuration of the starting positions (a_i values for a set of random variables $A_1 \ldots A_n$)

- transitions correspond to changing selected starting positions (and hence moving to a new state)



Dewey, 2013

Markov Chain Monte Carlo

- for the motif-finding task, the number of states is enormous
- key idea: construct Markov chain with stationary distribution equal to distribution of interest; use sampling to find most probable states
- detailed balance:

$$P(u)\tau(v \mid u) = P(v)\tau(u \mid v)$$

probability of state u

probability of transition $u \rightarrow v$

- when detailed balance holds:

$$\frac{1}{N}\lim_{N\to\infty} count(u) = P(u)$$

MCMC with Gibbs Sampling

Gibbs sampling is a special case of MCMC in which

- Markov chain transitions involve changing one variable at a time

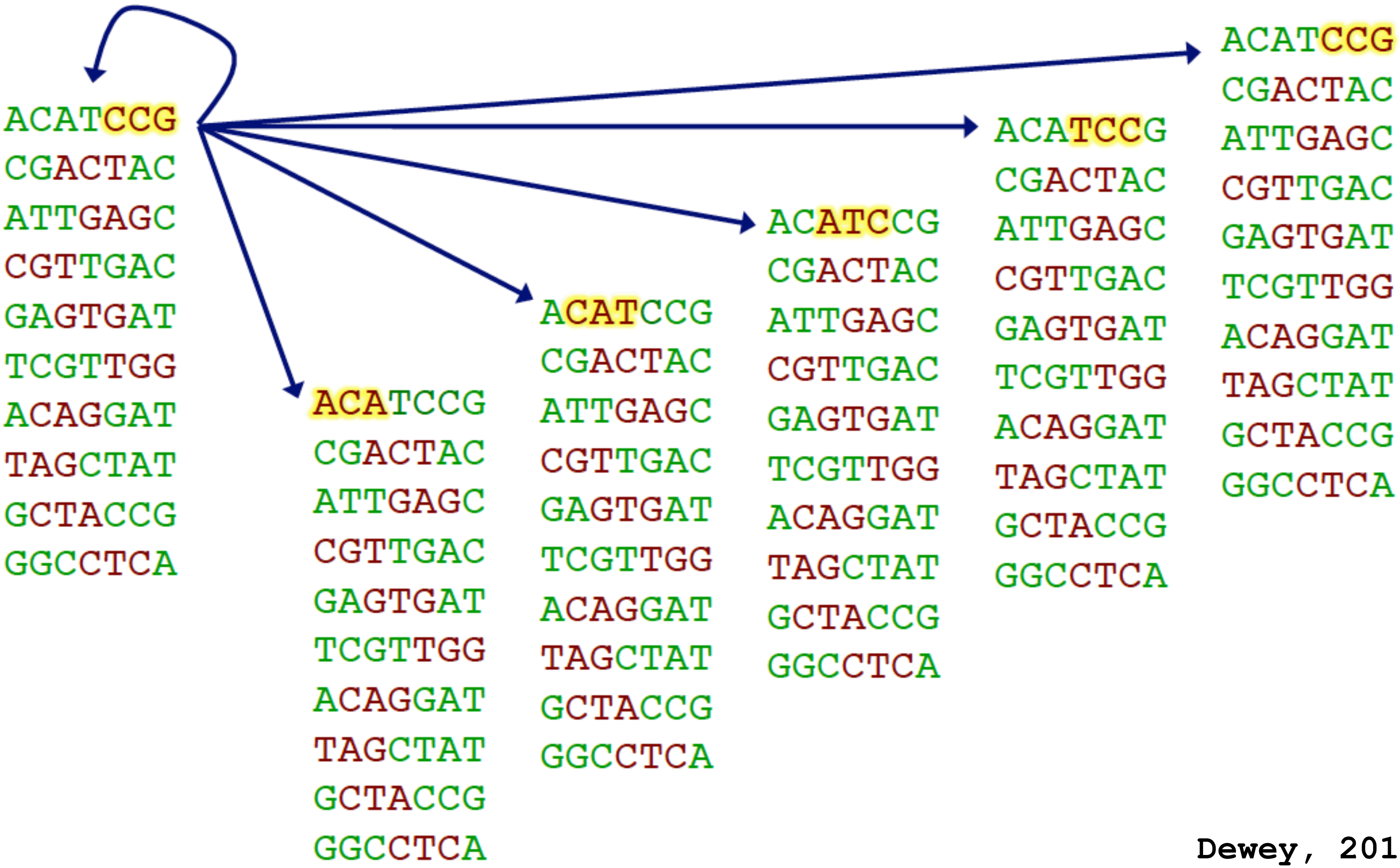
- transition probability is conditional probability of the changed variable given all others

- i.e. we sample the joint distribution of a set of random variables $P(A_1 \ldots A_n)$ by iteratively sampling from $P(A_i \mid A_1 \ldots A_{i-1}, A_{i+1} \ldots A_n)$

Dewey, 2013

Gibbs Sampling Approach

- possible state transitions when first sequence is selected



Dewey, 2013

Project I

- **Objective**: Design and develop a MCMC method to find a motif from a group of DNA sequence
- **Other tools and data**: http://biowhat.ucsd.edu/homer/motif/ ; Download the package to find the data in one of sub directories?
- MEME tool: http://meme.nbcr.net/meme/
- Motif Visualization tool (weblogo):

```
http://weblogo.berkeley.edu/logo.cgi
```

Gibbs Sampling Algorithm for Motif Finding

given: length parameter W, training set of sequences

choose random positions for a

do

pick a sequence X_i

estimate p given current motif positions a

(using all sequences but X_i) (predictive update step)

sample a new motif position a_i for X_i (sampling step)

until convergence

return: p, a

Gibbs Sampling Algorithm II

Assumption: size of motif is fixed

Initialization:

Make an initial guess of the motif locations and compute a probability matrix

Repeat:

Select one sequence randomly

Calculate the motif probability matrix with the new position use all other sequences

Use the matrix to evaluate the probabilities of all positions in the sequence (product of probability)

Select (or sample) a position in the sequence according to their probability

Until matrix converges or other criterion.

Sample a position according to probability

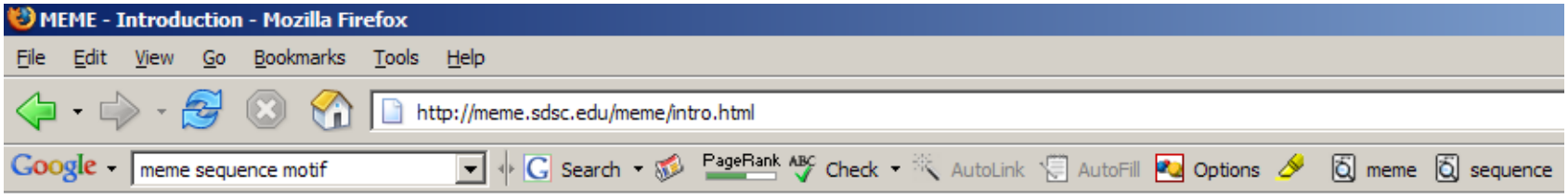
i

actcgtcggggcgtacgtacgtaacgtacgta**CGGACAACTGTTGACCG**
cggagcactgttgagcgacaagta**CGGAGCACTGTTGAGCG**gtacgtac
ccccgtagg**CGGCGCACTCTCGCCCG**ggcgtacgtacgtaacgtacgta
agggcgcgtacgctaccgtcgacgtcg**CGCGCCGCACTGCTCCG**acgct

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	1/4	1/4	.	.	.	.											
C	2/4	1/4															
G	0	2/4															
T	1/4	0															

Compute P_i = 2/4 * 2/4 * …. 1<=i<=n)
Select a position according to its Normalized probability.

Sample probability of i = $\frac{p_i}{\sum_{i=1}^{n} p_i}$



Menu
- Submit A Job
- Resources
- Alternate Servers
- Other Tools

MEME
Multiple Em for Motif Elicitation

MAST
Motif Alignment & Search Tool

THE MEME/MAST SYSTEM

Motif Discovery and Search

Version 3.5.3

The MEME/MAST system allows you to

1. discover motifs (highly conserved regions) in groups of related DNA or protein sequences using MEME and,
2. search sequence databases using motifs using MAST.

- The MEME/MAST system was developed by Timothy Bailey, Charles Elkan, and Bill Noble at the UCSD Computer Science and Engineering department with input from Michael Gribskov at Purdue University.
- MEME and MAST are described in detail in the papers available here.
- Answers to Frequently Asked Questions about MEME and MAST are given in the GENERAL FAQ.
- Visit the MEME user forum for online discussions with the MEME support team memebers and other MEME users.
- You can see sample MEME output or sample MAST output.
- Differences between the current release of the MEME/MAST system and earlier releases are described in the release notes .
- You can download the MEME/MAST software and install it on your own computer. This will allow you to use many features that are not available with the interactive versions of MEME and MAST.
- Meta-MEME combines motif models from MEME into a hidden Markov model framework for use in searching sequence databases.
- MEME and MAST are copyrighted software and can be licensed for commercial use.

Menu
- Submit A Job
- Resources
- Alternate Servers
- Other Tools

MEME
Multiple Em for Motif Elicitation

Version 3.5.3

Use this form to submit DNA or protein sequences to MEME. MEME will analyze your sequences for similarities among them and produce a description (motif) for each pat tern it discovers. Your results will be sent to you by e-mail.

Data Submission Form

Required

Your e-mail address:
jianlin.cheng@gmail.com
Re-enter e-mail address:
jianlin.cheng@gmail.com

Please enter the sequences which you believe share one or more motifs. The sequences may contain no more than **60,000 characters** total in any of a large number of formats.
Enter the name of a file containing the sequences here:
Browse...
or
The actual sequences here (Sample Input Sequences):

```
>seq1
actcgtcggggcgtacgtacgtaacgtacgtaCGGACAACTGTTGA
CCG
>seq2
cggagcactgttgagcgacaagtaCGGAGCACTGTTGAGCGgt
```

How do you think the occurrences of a single motif are distributed among the sequences?
- (•) **One** per sequence
- ( ) **Zero or one** per sequence
- ( ) **Any number** of repetitions

MEME will find the optimum width of each motif within the limits you specify here:
17 **Minimum** width (>= 2)
17 **Maximum** width (<= 300)

3 Maximum number of motifs to find

Optional

Description of your sequences:

MEME will find the optimum number of sites for each motif within the limits you specify here:
Minimum sites(>= 2)
Maximum sites (<= 300)

- [ ] Text output format
- [ ] Shuffle sequence letters

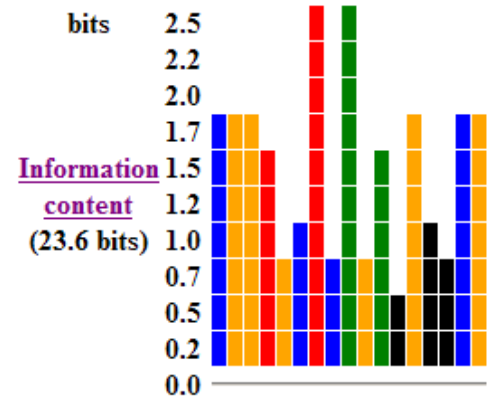
For DNA sequences only:
- [ ] Search given strand only
- [ ] Look for palindromes only

Start search Clear Input

P N **MOTIF 2** **width = 17** **sites = 4** **llr = 65** **E-value = 1.4e+000**

Simplified pos.-specific probability matrix

```
A : : : 8: 3a: : : : : 5: : :
C a: : 338: 8: 333: 55a:
G : aa: 8: : 3: 8: 3a: 5: a
T : : : : : : : : a: 85: : : : :
```



Multilevel consensus sequence

```
CGGAGCACTGTTGACCG
   CCA G CCC CG
        G
```

NAME	STRAND	START	P-VALUE	SITES		
seq2	+	1	2.44e-10		CGGAGCACTGTTGAGCG	ACAAGTACGG
seq1	+	33	5.18e-09	AACGTACGTA	CGGACAACTGTTGACCG	
seq4	-	28	1.08e-07	AGCGT	CGGAGCAGTGCGGCGCG	CGACGTCGAC
seq3	+	10	1.08e-07	CCCCGTAGG	CGGCGCACTCTCGCCCG	GGCGTACGTA

Motif 2 block diagrams

Gibbs Motif Sampler

http://bayesweb.wadsworth.org/gibbs/gibbs.html

The Gibbs Motif Sampler

(for DNA)

Show advanced options | How to enter data?

Email Address:

Please enter the data sequence: (FASTA format) *

Browse...

Prokaryotic Defaults | Prokaryotic Defaults

Sampler Mode: ○ Site Sampler

No. of different motifs (patterns):

Motif Width(s):*

Eukaryotic Defaults | Eukaryotic Defaults

○ Motif Sampler ◉ Recursive Sampler

Max sites per seq: (recursive sampler)

Est. total sites for each motif type:

Submit | Clear

Gibbs Motif Sampler
http://bayesweb.wadsworth.org/gibbs/gibbs.html

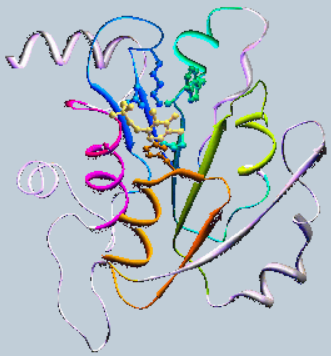
Email Address: jianlin.cheng@gmail.com

The Gibbs Motif Sampler

(for DNA)

Please enter the data sequence: (FASTA format) *

```
>seq1
actcgtcggggcgtacgtacgtaacgtacgtaCGGACAACTGTTGACCG
>seq2
cggagcactgttgagcgacaagtaCGGAGCACTGTTGAGCGgtacgtac
>seq3
ccccgtaggCGGCGCACTCTCGCCCGggcgtacgtacgtaacgtacgta
>seq4
agggcgcgtacgctaccgtcgacgtcgCGCGCCGCACTGCTCCGacgct
```



Browse...

Prokaryotic Defaults Prokaryotic Defaults

Eukaryotic Defaults Eukaryotic Defaults

Sampler Mode: (•) Site Sampler ( ) Motif Sampler ( ) Recursive Sampler

No. of different motifs (patterns): 3

Max sites per seq: (recursive sampler)

Motif Width(s):* 17

Est. total sites for each motif type:

Submit Clear

Browse the Gibbs Motif Sampler Manual

Output of Gibbs Sampler

actcgtcggggcgtacgtacgtaacgtacgtaCGGACAACTGTTGACCG
cggagcactgttgagcgacaagtaCGGAGCACTGTTGAGCGgtacgtac
ccccgtaggCGGCGCACTCTCGCCCGggcgtacgtacgtaacgtacgta
agggcgcgtacgctaccgtcgacgtcgCGCGCCGCACTGCTCCGacgct

Motif probability model

Pos. #	a	t	c	g
1	0.014	0.013	0.949	0.024
2	0.014	0.013	0.023	0.950
3	0.014	0.013	0.023	0.950
4	0.755	0.013	0.209	0.024
5	0.014	0.013	0.209	0.765
6	0.199	0.013	0.764	0.024
7	0.940	0.013	0.023	0.024
8	0.014	0.013	0.764	0.209
9	0.014	0.939	0.023	0.024
10	0.014	0.013	0.209	0.765
11	0.014	0.754	0.209	0.024
12	0.014	0.568	0.209	0.209
13	0.014	0.013	0.023	0.950
14	0.570	0.013	0.394	0.024
15	0.014	0.013	0.394	0.579
16	0.014	0.013	0.949	0.024
17	0.014	0.013	0.023	0.950

Prob Matrix

Confidence

Motif

Start pos

End pos

Background probability model
0.225 0.189 0.279 0.306

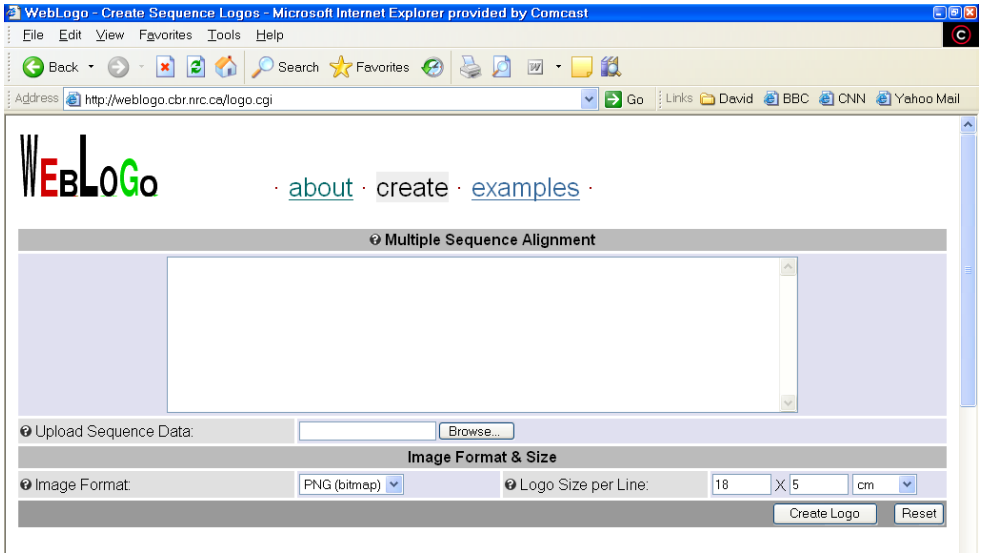
17 columns
Num Motifs: 5

1,	1	33	acgta	CGGACAACTGTTGACCG	49	1.00	F	seq1
2,	1	1		CGGAGCACTGTTGAGCG	acaag	17	0.49	F seq2
2,	2	25	aagta	CGGAGCACTGTTGAGCG	gtacg	41	0.51	F seq2
3,	1	10	gtagg	CGGCGCACTCTCGCCCG	ggcgt	26	1.00	F seq3
4,	1	44	agcgt	CGGAGCAGTGCGGCGCG	cgacg	28	1.00	R seq4

****

- Graphical representation of nucleotide base (or amino acid) conservation in a motif (or alignment)
- Information theory $$2 + \sum_{b=\{A,C,G,T\}} p(b) \log_2 p(b)$$
- Height of letters represents relative frequency of nucleotide bases

http://weblogo.berkeley.edu/



Kathrina Kechris, 2005

Entropy and Information

Visualization goals

(1) The height of the position is proportional to the information contained at the position

(2) The height of a letter is proportional to the probability of the letter appearing at the position

Two new concepts related to probability matrix:

Entropy

Information

- Entropy is a measure of uncertainty of a distribution $\sum_i - p_i \log_2 p_i$

	A	C	G	T
1	1/4	1/4	1/4	1/4
2	0	1	0	0
3	1/2	1/2	0	0
4				
⋮				

What is the entropy
Of positions 1,2,3?

- Information is the opposite of entropy. It measures the certainty of a distribution
- Information = maximum entropy – the entropy of a position (or distribution)

Maximum entropy for n characters is the Entropy when n characters are uniformly Distributed. $\log_2 n$

```
Info. Of pos 1= 2 - 2 = 0
Info. Of pos 2 = 2 - 0 = 2
Info. Of pos 3 = 2 -1 = 1
```

http://weblogo.berkeley.edu/logo.cgi

WEBLOGO

· about · create · examples ·

Multiple Sequence Alignment

```
>seq1
CGGACAACTGTTGACCG
>seq2
CGGAGCACTGTTGAGCG
>seq3
CGGCGCACTCTCGCCCG
>seq4
CGCGCCGCACTGCTCCG
```

Upload Sequence Data: Browse...

Image Format & Size

Image Format: PNG (bitmap) | Logo Size per Line: 18 X 5 cm

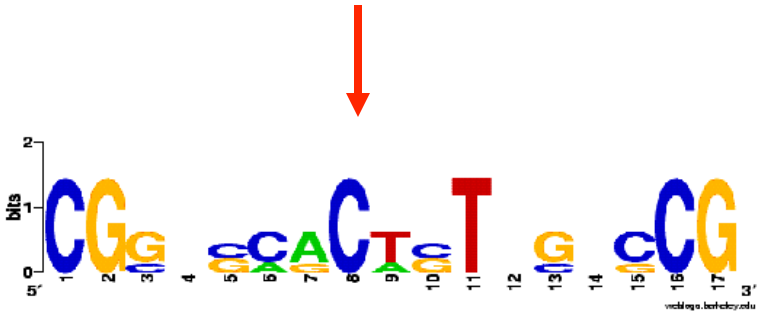
Create Logo

Advanced Logo Options

Sequence Type: amino acid / DNA / RNA / Automatic Detection

First Position Number: 1 | Logo Range: -

Small Sample Correction: ☑ | Frequency Plot: ☐



Project I

- Develop your Gibbs sampling program to find motifs in a group of gene sequences
- Test the program on a number of sequence groups
- Analyze the algorithm, program and results (speed, robustness, accuracy, quality, visualization, comparison, alternative solutions)

Data at Course Website

Projects

1. Search DNA sequence motif using MCMC

(Discussion of Plan (Sept. 10, Wed), Presentation of Plan (Sept. 12, Fri), and report on 9/19 (Friday)); Motif data set (Reference: Brown et al., MEME-LaB: motif analysis in clusters. Bioinformatics, 2013). One sample file (You can test your program on a few sequence files to search moti with different lengths (6 - 15 nucleotides)).

Project Groups

Group 1: Jie Hou, Minguan Song, Tuan Trieu, Meng Zhang, Hao Sun

Group 2: Abhishek Shah, Mike Phinney, Chao Fang, Matt England

Group 3: Xinjian Yao, Yuxiang Zhang, Rui Xie, Muxi Chen, Xinwei Du

Group 4: Kevin Melkowski, Michael Pieper, Mary Sheahen, Kristofferson Culmer

Others: participating in discussion

Items to Discuss (Wednesday, Sept. 10)

Group Discussion

- Problem definition
- Algorithm
- Implementation
- Evaluation
- Visualization
- Task Assignment
- Select Coordinator
- Data Sharing
- Communication

- **Brief Introduction by Deb (5 – 10 minutes)**
- **Discussion of Plan (40)**
- **Informal Presentation on whiteboard (5-6 minutes per group)**

Presentation of Plan (Friday, Sept. 12)

- Make an official plan in PPT
- Present your plan by coordinator: 15 minutes presentation + 3 minutes question-answer
- Create your data sharing account on public cloud
- Send your revised plan to mumachinelearning@gmail.com after presentation
- **Present your results on Sept. 19 (Friday)**