Multiple Sequence Alignment (II)

Jianlin Cheng, PhD

School of Electrical Engineering and Computer Science
University of Central Florida

2006

Local MSA

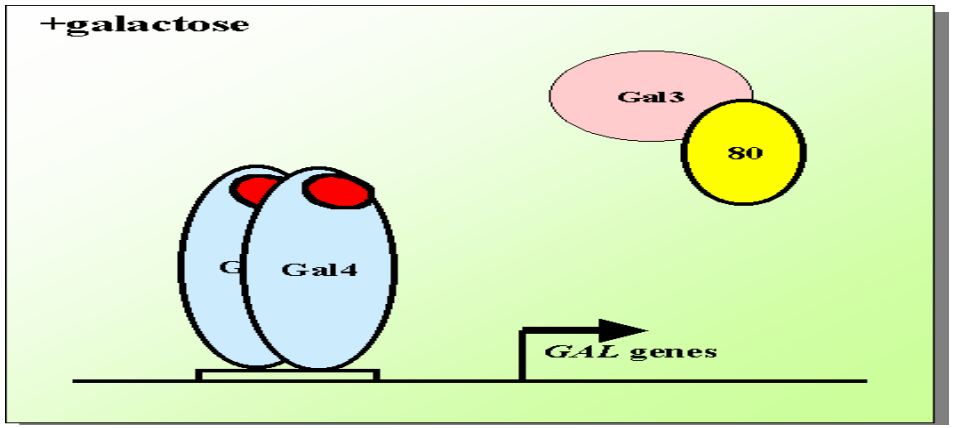
- Most widely used MSA tools are global alignment which generally perform better.
- Local MSA can be useful in finding local conserved regions: functional sites, DNA motifs.
- Local MSA can be useful for alignments of multiple-domain protein sequences.

What's Local MSA?

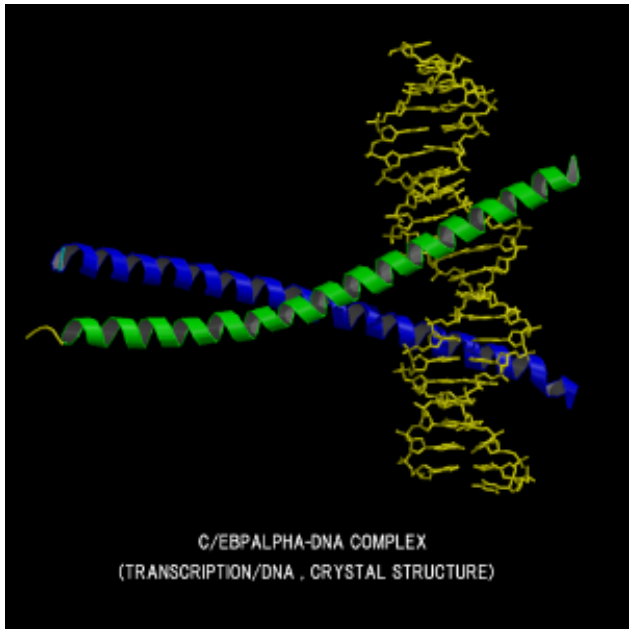
- Generalization of local pairwise alignment
- Find a set of sub-sequences of multiple sequences whose alignment has maximum alignment score.
- It is NP-hard as global MSA.
- Local MSA is to find highly conserved regions (motifs) of related genes or a protein family

Examples: Transcription Factors

- yeast: Gal4
- drosophila
- mammal




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



Kathrina Kechris, 2005

One Example Problem

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

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

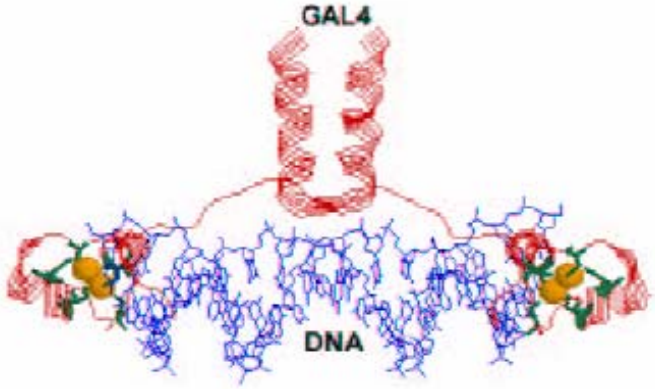
Goals: 1) Estimate motif
2) Align motif / Predict motif locations

	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

Alignment of Transcription Factor Binding Sites




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

Some Motif Databases

- PROSITE: http://www.expasy.org/prosite/
- InterPro: http://www.ebi.ac.uk/interpro/
- BLOCKS: http://blocks.fhcrc.org/
- PRINTS: http://bioinf.man.ac.uk/dbbrowser/PRINTS/
- TRANSFAC: http://www.gene-regulation.com/pub/databases.html#transfac

(DNA binding sites)

Local MSA Methods

- Progressive Local DP approach
- EM algorithm: MEME
- Gibbs sampling algorithms

Progressive Local DP Approach

- Same idea as global progressive MSA
- Three Steps

  1. align sequences pairwisely using local DP and generate similarity and distance matrices

  2. construct a guide tree

  3. align sequences progressively according to guide tree.

Expectation and Maximization Algorithm

- Representative method: MEME (multiple EM for motif elicitation)
- EM algorithm is **a general, powerful method** to maximize likelihood P(D|model)
- Motif is represented as a probability matrix
- Estimate motif locations and the matrix iteratively until converge

EM Algorithm

Assumption: size of motif is fixed

Initialization:

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

Repeat:

E-step: use the matrix to evaluate the probabilities of all positions in each of all sequences (product of probability).

M-step: Select the position with maximum probability in each sequence and recalculate the motif probability matrix

Until matrix is not changed.

Initialization

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

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															

E-step

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															

Find the best position in each sequence
That maximize product of probability

Prob (posi_i) = 2/4 * 2/4 * ….

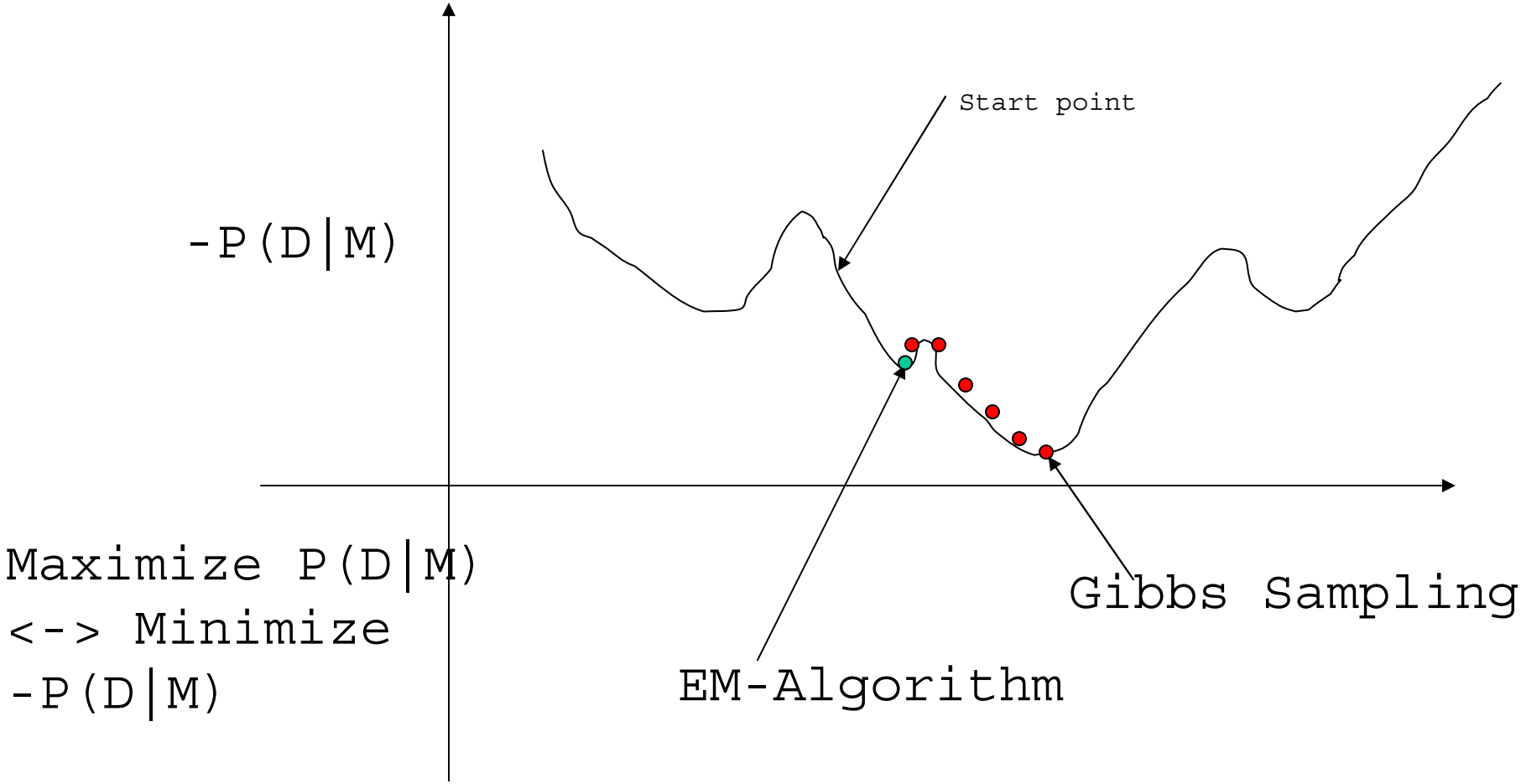
Maximization

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															

Local Minima versus Global Minima



Gibbs Sampling Algorithm

Gibbs sampling algorithm is MCMC (Markov Chain Monte Carlo) method.

It introduces randomness into EM algorithm.

It is harder to detect convergence of alignment.

For each run, it is stochastic instead of deterministic. Less susceptible to local minima.

Gibbs Sampling

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

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

Recalculate the motif probability matrix with the new position

Until matrix converges.

Sample a position according to probability instead of choosing best

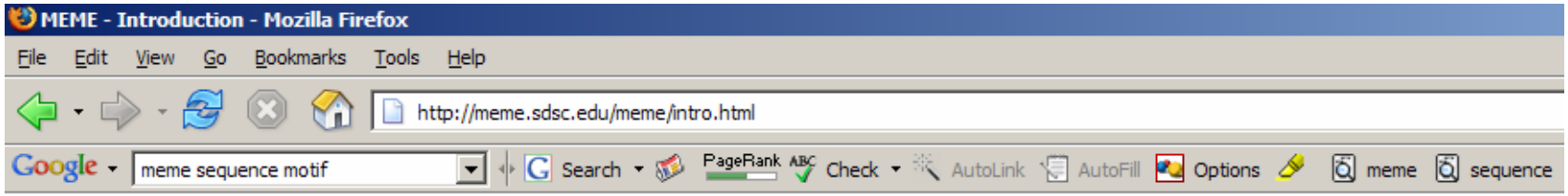
i

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.

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



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

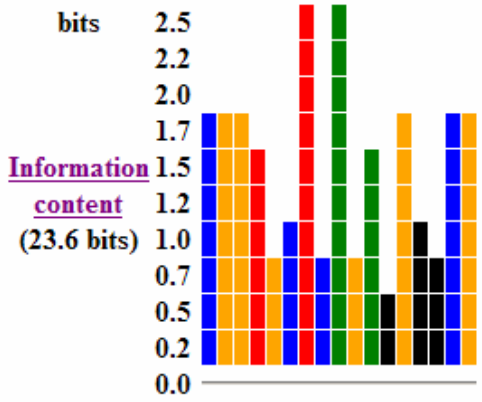
For DNA sequences only:
- [ ] Search given strand only
- [ ] Look for palindromes only

Start search | Clear Input

PN **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

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

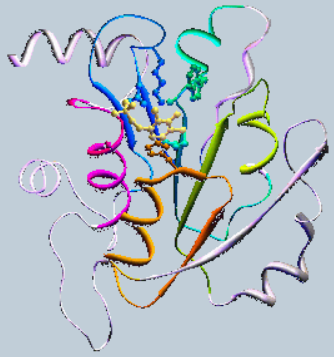
Email Address: jianlin.cheng@gmail.com

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

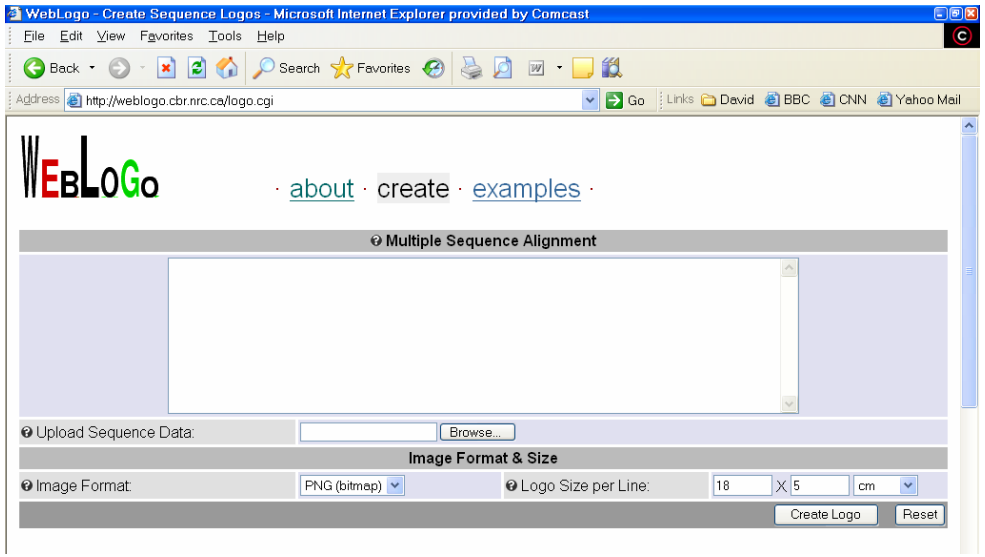
17 columns
Num Motifs: 5

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:	☐

Ten Topics

- 1. Introduction to Molecular Biology and Bioinformatics
- 2. Pairwise Sequence Alignment Using Dynamic Programming
- 3. Practical Sequence/Profile Alignment Using Fast Heuristic Methods (BLAST and PSI-BLAST)
- 4. Multiple Sequence Alignment
- 5. Gene and Motif Identification
- 6. Phylogenetic Analysis
- 7. Protein Structure Analysis and Prediction
- 8. RNA Secondary Structure Prediction
- 9. Clustering and Classification of Gene Expression Data
- 10. Search and Mining of Biological Databases, Databanks, and Literature