Identifying and representing DNA binding sites

Biol4559 Tues, April 21, 2015
Bill Pearson wrp@virginia.edu 4-2818 Jordan 6-057

- Looking for functional sites: promoters, regulatory elements, modification sites
- Products of *convergent*, not divergent evolution
- Weak spacing constraints
- Usually represented as a consensus sequence
- If alignment is given, consensus is obvious
- If consensus is given, alignment is obvious
- Search for consensus and alignment together
- `consensus, meme, gibbs`

To learn more:

- Overview of multiple alignment and motif finding – Mount (2001) Chapter 4
- Schneider et al. (1986) J. Mol. Biol. Information content of binding sites on nucleotide sequences 188:415-431
- Stormo and Hartzell (1989) Identifying protein binding sites from unaligned DNA fragments
- Lawrence and Reilly (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences PROTEINS 7:41-51
- Lawrence et al. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment Science 262:208-214

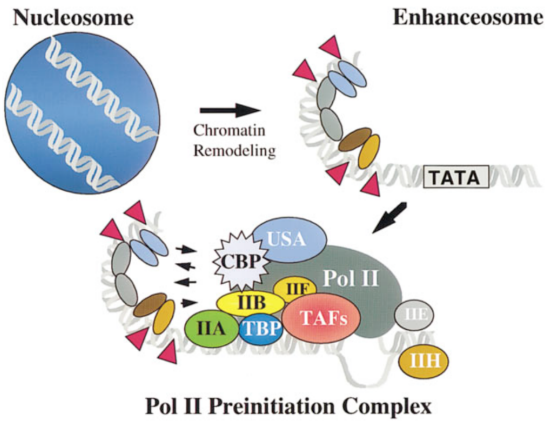
Regulation of transcription: RNA polymerase, promoters, enhancers, transcription factors, DNA binding proteins

- Gene expression is regulated at many levels:
  - production of RNA (transcription)
    - promoter occupancy, transcription rate, termination
  - transcript RNA splicing
  - mRNA stability
  - mRNA translation
  - post-translational processing/stability

Regulation of transcription: RNA polymerase, promoters, enhancers, transcription factors, DNA binding proteins

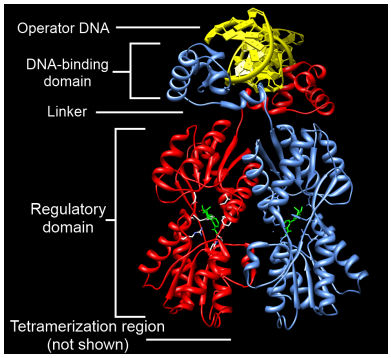
- Transcription factors interact with DNA to increase/decrease transcription levels
  - lacR (bacterial lac Repressor)
  - hundreds of transcription factors modify expression in response to signals
    - FOS/JUN
    - nuclear receptors
    - homeobox proteins
    - Myc
    - etc.,etc.

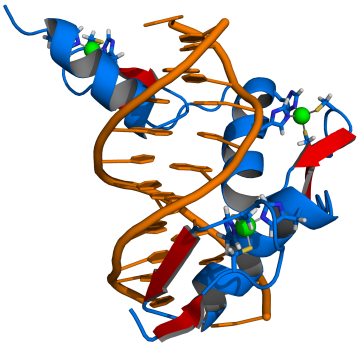
Regulation of transcription: RNA polymerase, promoters, enhancers, transcription factors, DNA binding proteins



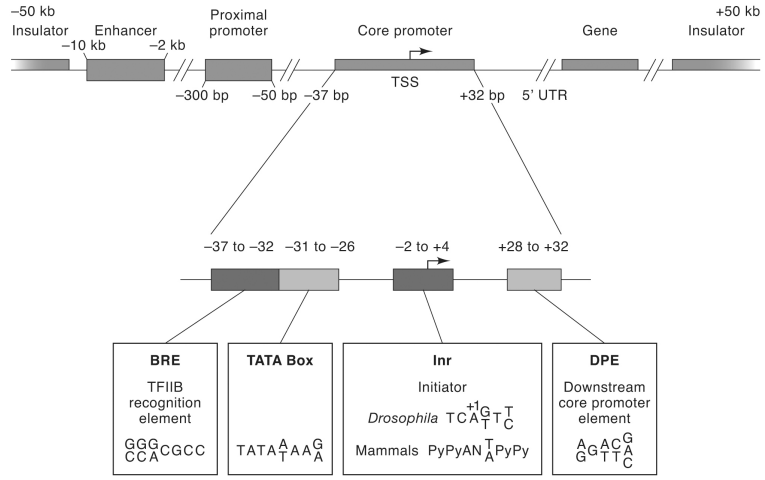
Pol II Preinitiation Complex

Regulation of transcription: RNA polymerase, promoters, enhancers, transcription factors, DNA binding proteins





Regulation of transcription: RNA polymerase, promoters, enhancers, transcription factors, DNA binding proteins



Transcriptional regulation in eukaryotes, concepts, strategies, and techniques. CSHL Press 2009

7

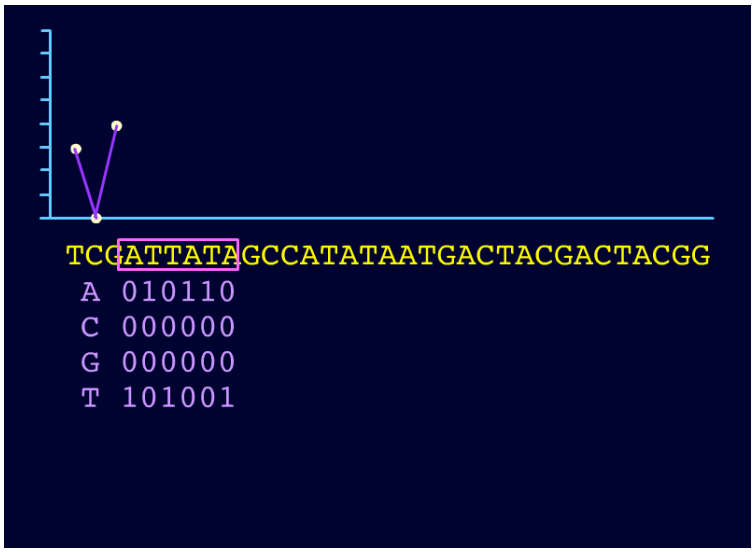
Consensus Patterns and Motifs

- How to represent motifs
  - consensus patterns
  - weight matrices
- How to "weight" positions?
  - frequency
  - information content (S log odds)
- How to search for motifs
  - Heuristic greedy (consensus, wconsensus)
  - Expectation-Maximization (MEME)
  - Gibbs sampling

fasta.bioch.virginia.edu/biol4559

8

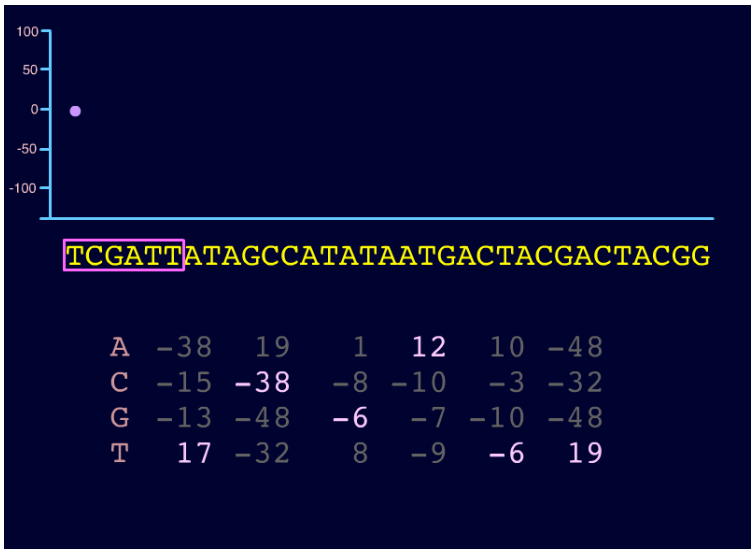
Scanning a sequence with PATSER



fasta.bioch.virginia.edu/biol4559

9

Scanning a sequence with PATSER



fasta.bioch.virginia.edu/biol4559

10

Consensus Patterns and Motifs

- How to represent motifs
  - consensus patterns
  - weight matrices
    - do not require identity
    - include consensus (1's, 0's)
    - allow mismatches

How to represent motifs – information content

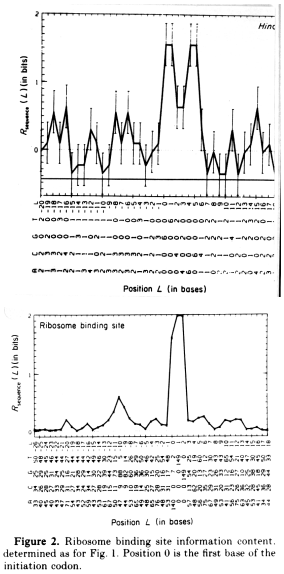


Figure 2. Ribosome binding site information content, determined as for Fig. 1. Position 0 is the first base of the initiation codon.

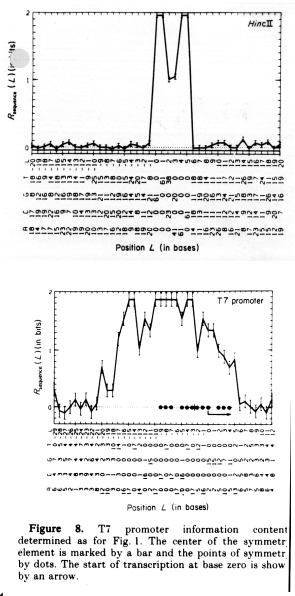


Figure 8. T7 promoter information content determined as for Fig. 1. The center of the symmetr element is marked by a bar and the points of symmetr by dots. The start of transcription at base zero is show by an arrow.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. *J Mol Biol.* (1986) **188**:415-431
Information content of binding sites on nucleotide sequences.

Representing Consensus Sequences

A	0	1	0	1	1	0
C	0	0	0	0	0	0
G	0	0	0	0	0	0
T	1	0	1	0	0	1

A	0	11	4	7	9	0
C	1	0	1	2	1	0
G	1	0	1	2	1	0
T	10	1	6	1	1	12

A	2	95	26	59	51	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96

A	-38	19	1	12	10	-48
C	-15	-38	-8	-10	-3	-32
G	-13	-48	-6	-7	-10	-48
T	17	-32	8	-9	-6	19

Consensus Patterns and Motifs

- How to represent motifs
  - consensus patterns
  - weight matrices
- How to "weight" positions?
  - frequency
  - information content (S log odds)
- How to search for motifs
  - heuristic greedy
  - Gibbs sampling

Representing Consensus Sequences

A	9	214	63	142	118	8
C	22	7	26	31	52	13
G	18	2	29	38	29	5
T	193	19	124	31	43	216

A	0.04	0.88	0.26	0.59	0.49	0.03
C	0.09	0.03	0.11	0.13	0.22	0.05
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89

A	-2.76	1.82	0.06	1.23	0.96	-2.92
C	-1.46	-3.11	-1.22	-1.00	-0.22	-2.21
G	-1.76	-5.00	-1.06	-0.67	-1.06	-3.58
T	1.67	-1.66	1.04	-1.00	-0.49	1.84

$$I_p = \sum_{b}^{A..T} f_b \log\left(\frac{f_b}{p_b}\right)$$

I	0.98	1.33	0.28	0.38	0.22	1.35

15

Ranking criterion: Information Content

$$\text{Information content} = \frac{1}{N} LogLikelihood\ \ Ratio$$

$$I_{total} = \sum_{pos=1}^{n} \sum_{b}^{A..T} f_{pos,b} \log_2 \frac{f_{pos,b}}{p_b}$$

Information content and binding energy

If we denote *Hi* the *binding energy* for a DNA site *Si*, then the probability that the protein would be bound to *Si* (at equilibrium) is given by the Boltzman distribution:

$$P_i = \frac{e^{-Hi}}{Z}$$

where Z is the *partition function* and is defined as the sum of the e^{-Hi} over all possible sites *Sx*:

$$Z = \sum_x e^{-Hx}$$

The average binding energy for this protein over all sites *Si* would be:

$$\langle H \rangle = \sum P_i H_i = -\sum_i (P_i \ln P_i) - \ln Z$$

The sum on the right is called the entropy of the probability distribution.

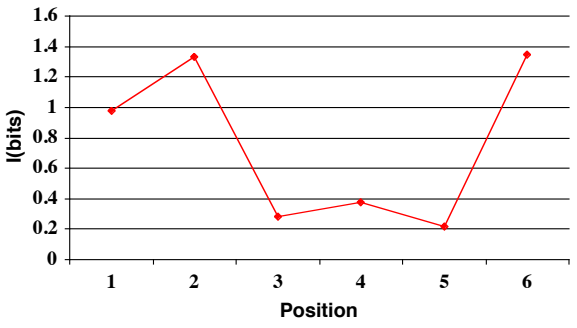
A useful measure of difference between two probability distributions is the relative entropy, which is defined as:

$$H(P,Q) \equiv \sum_i P_i \ln \frac{P_i}{Q_i}$$

Conensus Information Content

$$I_p = \sum_b^{A..T} f_b \log\left(\frac{f_b}{p_b}\right)$$

I	0.98	1.33	0.28	0.38	0.22	1.35



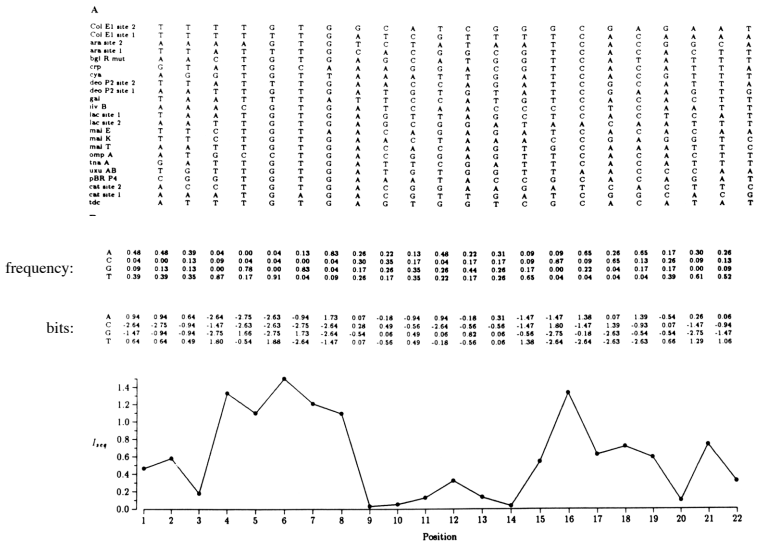


FIG. 1. CRP binding sites. (*A*) Twenty-three sites identified as binding to CRP (8, 9). (*B*) The frequency with which each base occurs at each position in the CRP binding sites. (*C*) The specificity matrix for the protein, based on the binding sites, which is calculated as $\log_2(f_b/p_b)$, where f_b is the observed frequency of each base (from the matrix above) and p_b is the *a priori* probability of obtaining base *b*. In this example p_b = 0.25 for all *b*, approximating the *E. coli* genomic composition. At positions for which f_b = 0, an estimated frequency of 0.5/23 is used in the calculation. (*D*) The "information content" at each position of the CRP binding sites is plotted (10). The sum of all positions is 13.06 bits.

fasta.bioch.virginia.edu/biol4559 Stormo GD, Hartzell GW (1989) Identifying protein-binding sites from unaligned DNA fragments.Proc Natl Acad Sci U S A. 86:1183-1187. 19

Consensus Patterns and Motifs

- How to represent motifs
  - consensus patterns
  - weight matrices
- How to "weight" positions?
  - frequency
  - information content (S log odds)
- How to search for motifs
  - heuristic greedy (consensus, wconsensus)
  - Expectation-Maximization (MEME)
  - Gibbs sampling

fasta.bioch.virginia.edu/biol4559 20

Identifying Functional Domains in Biological Sequences

A problem in:
Feature Detection
or
Multiple Alignment

Two parts to the problem:

1. Can't look at all possible alignments, pick a subset likely to contain the answer
2. Need criterion for ranking alignments that is reasonable and efficient

E. coli CRP binding regions

```
CE1CG      taatgtttgtgctggtttttgtggcatcgggcgagaatagcgcgtggtgtgaaagactgttttttgatcgttttcacaaaaatggaagtccacagtcttgacag
ECOARABOP  gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattatttgcacggcgtcacactttgctatgccatagcatttttatccataag
ECOBGLR1   acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaataactgtgagcatggtcatatttttatcaat
ECOCRP     cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtatgcaaaggacgtcacattaccgtgcagtacagttgatagc
ECOCYA     acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcaaggtgttaaattgatcacgttttagaccatttttcgtcgtgaaactaaaaaaacc
ECODEOP2   agtgaattatttgaaccagatcgcattacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
ECOGALE    gcgcataaaaaacggctaaattcttgtgtaaacgattccactaatttattccatgtcacacttttcgcatctttgttatgctatggttatttcataccataagcc
ECOILVBPR  gctccggcggggttttttgttatctgcaattcagtacaaaacgtgatcaacccctcaattttccctttgctgaaaaattttccattgtctcccctgtaaagctgt
ECOLAC     aacgcaattaatgtgagttagctcactcattaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
ECOMALBA   acattaccgccaattctgtaacagagatcacacaaagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ECOMALBA   ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
ECOMALT    gatcagcgtcgttttaggtgagttgttaataaagatttggaattgtgacacagtgcaaattcagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ECOOMPA    gctgacaaaaaagattaaacatacctatacaagacttttttttcatatgcctgacggagttcacacttgtaagttttcaactacgttgtagactttacatcgcc
ECOTNAA    ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacgattgtgattcgattcacatttaaacaatttcaga
ECOUXU1    cccatgagagtgaaattgttgtgatgtggttaacccaattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
PBR322     ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggtgtgaaataccgcacagatgcgtaaggagaaaataccgcatcaggcgctc
TRN9CAT    ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaatgagacgttgatcggcacg
TDC        gatttttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatttgtgagtggtcgcacatatcctgtt
```

E. coli CRP binding sites

```
CE1CG      taatgtttgtgctggtttttgtggcatcgggcgagaatagcgcgtggtgtgaaagactgtttttTTTGATCGTTTTCACAAAaatggaagtccacagtcttgacag
ECOARABOP  gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattaTTTGCACGGCGTCACACTttgctatgccatagcatttttatccataag
ECOBGLR1   acaaatcccaataacttaattattgggatttgttatatataacttttataaattcctaaaattacacaaagttaataacTGTGAGCATGGTCATATTtttatcaat
ECOCRP     cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtaTGCAAAGGACGTCACATTaccgtgcagtacagttgatagc
ECOCYA     acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcaaggTGTTAAATTGATCACGTTttagaccatttttcgtcgtgaaactaaaaaaacc
ECODEOP2   agtgaattaTTTGAACCAGATCGCATTacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
ECOGALE    gcgcataaaaaacggctaaattcttgtgtaaacgatttccactaatttattccaTGTCACACTTTTCGCATCtttgttatgctatggttatttcataccataagcc
ECOILVBPR  gctccggcgggggttttttgttaTCTGCAATTCAGTACAAAacgtgatcaacccctcaattttcccttgctgaaaaattttccattgtctccccgtaaagctgt
ECOLAC     aacgcaattaaTGTGAGTTAGCTCACTCAttaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
ECOMALBA   acattaccgccaattcTGTAACAGAGATCACACAaagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ECOMALBA   ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttCGTGATGTTGCTTGCAAAaatcgtggcgattttatgtgcgca
ECOMALT    gatcagcgtcgttttaggtgagttgttaataaagatttggaatTGTGACACAGTGCAAATTcagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ECOOMPA    gctgacaaaaaagattaaacatacctatacaagacttttttttcatatgCCTGACGGAGTTCACACTtgtaagttttcaactacgttgtagactttacatcgcc
ECOTNAA    ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacgatTGTGATTCGATTCACATTtaaacaatttcaga
ECOUXU1    cccatgagagtgaaattgtTGTGATGTGGTTAACCCAattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
PBR322     ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggTGTGAAATACCGCACAGAtgcgtaaggagaaaataccgcatcaggcgctc
TRN9CAT    ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaaTGAGACGTTGATCGGCACg
TDC        gattttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaattTGTGAGTGGTCGCACATAtcctgtt
```


E. coli CRP binding sites – location gives alignment

```
CE1CG      taatgtttgtgctggtttttgtggcatcgggcgagaatagcgcgtggtgtgaaagactgtttttTTTGATCGTTTTCACAAAaatggaagtccacagtcttgacag
ECOARABOP  gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattaTTTGCACGGCGTCACACTttgctatgccatagcatttttatccataag
ECOBGLR1   acaaatcccaataacttaattattgggatttgttatatataacttttataaattcctaaaattacacaaagttaataacTGTGAGCATGGTCATATTtttatcaat
ECOCRP     cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtaTGCAAAGGACGTCACATTaccgtgcagtacagttgatagc
ECOCYA     acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcaaggTGTTAAATTGATCACGTTttagaccatttttcgtcgtgaaactaaaaaaacc
ECODEOP2   agtgaattaTTTGAACCAGATCGCATTacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
ECOGALE    gcgcataaaaaacggctaaattcttgtgtaaacgatttccactaatttattccaTGTCACACTTTTCGCATCtttgttatgctatggttatttcataccataagcc
ECOILVBPR  gctccggcgggggttttttgttaTCTGCAATTCAGTACAAAacgtgatcaacccctcaattttcccttgctgaaaaattttccattgtctccccgtaaagctgt
ECOLAC     aacgcaattaaTGTGAGTTAGCTCACTCAttaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
ECOMALBA   acattaccgccaattcTGTAACAGAGATCACACAaagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ECOMALBA   ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttCGTGATGTTGCTTGCAAAaatcgtggcgattttatgtgcgca
ECOMALT    gatcagcgtcgttttaggtgagttgttaataaagatttggaatTGTGACACAGTGCAAATTcagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ECOOMPA    gctgacaaaaaagattaaacatacctatacaagacttttttttcatatgCCTGACGGAGTTCACACTtgtaagttttcaactacgttgtagactttacatcgcc
ECOTNAA    ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacgatTGTGATTCGATTCACATTtaaacaatttcaga
ECOUXU1    cccatgagagtgaaattgtTGTGATGTGGTTAACCCAattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
PBR322     ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggTGTGAAATACCGCACAGAtgcgtaaggagaaaataccgcatcaggcgctc
TRN9CAT    ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaaTGAGACGTTGATCGGCACg
TDC        gattttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaattTGTGAGTGGTCGCACATAtcctgtt
```

Aligned CRP binding sites

```
CE1CG    cgcgtggtgtgaaagactgtttttTTTGATCGTTTTCACAAAaatggaagtccacagtcttgacag
ECOARABOP gcagaaaagtccacattgattaTTTGCACGGCGTCACACTttgctatgccatagcatttttatccataag
ECOBGLR1 taaaattacacaaagttaataacTGTGAGCATGGTCATATTtttatcaat
ECOCRP   aatacattgatgtactgcatgtaTGCAAAGGACGTCACATTaccgtgcagtacagttgatagc
ECOCYA   ttctttacggtcaatcagcaaggTGTTAAATTGATCACGTTTtagaccatttttcgtcgtgaaactaaaa
ECODEOP2              agtgaattaTTTGAACCAGATCGCATTacagtgatgcaaacttgtaagtagatttccttaat
ECOGALE  aacgattccactaatttattccaTGTCACACTTTTCGCATCtttgttatgctatggttatttcataccata
ECOILVBPR gctccggcggggttttttgttaTCTGCAATTCAGTACAAAacgtgatcaacccctcaattttccctttgctg
ECOLAC              aacgcaattaaTGTGAGTTAGCTCACTCAttaggcaccccaggctttacactttatgcttccggct
ECOMALBA        acattaccgccaattcTGTAACAGAGATCACACAaagcgacggtggggcgtaggggcaaggaggatggaaa
ECOMALBA aaccgaggtcatgtaaggaatttCGTGATGTTGCTTGCAAAaatcgtggcgattttatgtgcgca
ECOMALT gagttgttaataaagatttggaatTGTGACACAGTGCAAATTcagacacataaaaaaacgtcatcgcttgcattagaaa
ECOOMPA tatacaagacttttttttcatatgCCTGACGGAGTTCACACTtgtaagttttcaactacgttgtagactttacatcgcc
ECOT    atttaatattgctccccgaacgatTGTGATTCGATTCACATTtaaacaatttcaga
ECOUXU1      cccatgagagtgaaattgtTGTGATGTGGTTAACCCAattagaattcgggattgacatgtcttaccaaaaggta
PBR322  gtactgagagtgcaccatatgcggTGTGAAATACCGCACAGAtgcgtaaggagaaaataccgcatcaggcgctc
TRN9CAT ccctgggccaacttttggcgaaaaTGAGACGTTGATCGGCACg
TDC     tctggaaagtattgaaagttaatttTGTGAGTGGTCGCACATAtcctgtt
```


Ranking criterion: Information Content

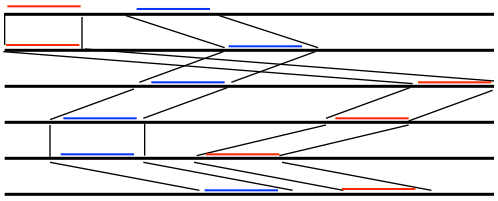
$$\text{Information content} = \frac{1}{N} LogLikelihood\ \ Ratio$$

$$I_{total} = \sum_{pos=1}^{n} \sum_{b}^{A..T} f_{pos,b} \log_2 \frac{f_{pos,b}}{p_b}$$

Consensus Patterns and Motifs

- How to represent motifs
  - consensus patterns
  - weight matrices
- How to "weight" positions?
  - frequency
  - information content (S log odds)
- How to search for motifs
  - heuristic greedy (consensus, wconsensus)
  - Expectation-Maximization (MEME)
  - Gibbs sampling

Finding a consensus sequence: `consensus`



Finding a consensus sequence: `consensus`

```
A C T G A A T
A G C G T C C
C T T G C C G
```

l=12.0

	A	C	T	G	A	A
A	1	0	0	0	1	1
C	0	1	0	0	0	0
G	0	0	0	1	0	0
T	0	0	1	0	0	0

l=12.0

	C	T	G	A	A	T
A	0	0	0	1	1	0
C	1	0	0	0	0	0
G	0	0	1	0	0	0
T	0	1	0	0	0	1

l=8.0

	A A	C G	T C	G G	A T	A C
A	2	0	0	0	1	1
C	0	1	1	0	0	1
G	0	1	0	2	0	0
T	0	0	1	0	1	0

l=7.0

	A G	C C	T G	G T	A C	A C
A	1	0	0	0	1	1
C	0	2	0	0	1	0
G	1	0	1	1	0	1
T	0	0	1	1	0	0

l=6.0

	C A	T G	G C	A G	A T	T C
A	1	0	0	1	1	0
C	1	0	1	0	0	1
G	0	1	1	1	0	0
T	0	1	0	0	1	1

l=7.0

	C G	T C	G G	A T	A C	T C
A	0	0	0	1	1	0
C	1	1	0	0	1	1
G	1	0	2	0	0	0
T	0	1	0	1	0	1

l=6.1

	A A	C G	T C	G T	A C	A C
	C	T	T	G	C	C
A	2	0	0	0	1	1
C	1	1	1	0	1	2
G	0	1	0	3	0	0
T	0	1	2	0	1	0

l=3.8

	A	C	T	G	A	A
	A	G	C	G	T	C
	T	T	G	C	C	G
A	2	0	0	0	1	1
C	0	1	1	1	1	1
G	0	1	1	2	0	1
T	1	1	1	0	1	0

l=5.8

	C G	T C	G G	T T	C G	C C
A	0	0	0	1	1	0
C	2	1	0	0	2	2
G	1	0	2	1	0	0
T	0	2	1	1	0	1

l=5.4

	C G	T C	G G	A T	A C	T C
	T	T	G	C	C	G
A	0	0	0	1	1	0
C	1	1	0	1	2	1
G	1	0	3	0	0	1
T	1	2	0	1	0	1

E. coli CRP binding sites

```
CE1CG      taatgtttgtgctggtttttgtggcatcgggcgagaatagcgcgtggtgtgaaagactgtttttTTTGATCGTTTTCACAAAaatggaagtccacagtcttgacag
ECOARABOP  gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattaTTTGCACGGCGTCACACTttgctatgccatagcatttttatccataag
ECOBGLR1   acaaatcccaatcaacttaattattgggatttgttatatatatcaactttataaattcctaaaattacacaaagttaataacTGTGAGCATGGTCATATTtttatcaat
ECOCRP     cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtaTGCAAAGGACGTCACATTaccgtgcagtacagttgatagc
ECOCYA     acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcaaggTGTTAAATTGATCACGTTttagaccatttttcgtcgtgaaactaaaaaaacc
ECODEOP2   agtgaattaTTTGAACCAGATCGCATTacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
ECOGALE    gcgcataaaaaacggctaaattcttgtgtaaacgattccactaatttattccaTGTCACACTTTTCGCATCtttgttatgctatggttatttcataccataagcc
ECOILVBPR  gctccggcggggttttttgttaTCTGCAATTCAGTACAAAacgtgatcaacccctcaattttcccttgctgaaaaattttccattgtctccccgtaaagctgt
ECOLAC     aacgcaattaaTGTGAGTTAGCTCACTCAttaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
ECOMALBA   acattaccgccaattcTGTAACAGAGATCACACAaagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ECOMALBA   ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttCGTGATGTTGCTTGCAAAaatcgtggcgattttatgtgcgca
ECOMALT    gatcagcgtcgttttaggtgagttgttaataaagatttggaatTGTGACACAGTGCAAATTcagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ECOOMPA    gctgacaaaaaagattaaacatacctatacaagacttttttttcatatgCCTGACGGAGTTCACACTtgtaagttttcaactacgttgtagactttacatcgcc
ECOTNAA    ttttttaaacatttaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacgatTGTGATTCGATTCACATTtaaacaatttcaga
ECOUXU1    cccatgagagtgaaattgtTGTGATGTGGTTAACCCAattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
PBR322     ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggTGTGAAATACCGCACAGAtgcgtaaggagaaaataccgcatcaggcgctc
TRN9CAT    ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaaTGAGACGTTGATCGGCACg
TDC        gatttttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatttTGTGAGTGGTCGCACATAtcctgtt
```

Aligned CRP binding sites

```
CE1CG     cgcgtggtgtgaaagactgttttTTTGATCGTTTTCACAAAaatggaagtccacagtcttgacag
ECOARABOP gcagaaaagtccacattgattaTTTGCACGGCGTCACACTttgctatgccatagcatttttatccataag
ECOBGLR1 taaaattacacaaagttaataacTGTGAGCATGGTCATATTtttatcaat
ECOCRP    aatacattgatgtactgcatgtaTGCAAAGGACGTCACATTaccgtgcagtacagttgatagc
ECOCYA    ttctttacggtcaatcagcaaggTGTTAAATTGATCACGTTttagaccatttttcgtcgtgaaactaaaa
ECODEOP2               agtgaattaTTTGAACCAGATCGCATTacagtgatgcaaacttgtaagtagatttccttaat
ECOGALE   aacgattccactaatttattccaTGTCACACTTTTCGCATCtttgttatgctatggttatttcataccata
ECOILVBPR gctccggcgggggttttttgttaTCTGCAATTCAGTACAAAacgtgatcaacccctcaattttccctttgctg
ECOLAC              aacgcaattaaTGTGAGTTAGCTCACTCAttaggcaccccaggctttacactttatgcttccggct
ECOMALBA        acattaccgccaattcTGTAACAGAGATCACACAaagcgacggtggggcgtaggggcaaggaggatggaaa
ECOMALBA aaccgaggtcatgtaaggaatttCGTGATGTTGCTTGCAAAaatcgtggcgattttatgtgcgca
ECOMALT gagttgttaataaagatttggaatTGTGACACAGTGCAAATTcagacacataaaaaaacgtcatcgcttgcattagaaa
ECOOMPA tatacaagactttttttttcatatgCCTGACGGAGTTCACACTtgtaagttttcaactacgttgtagactttacatcgcc
ECOT    atttaatattgctccccgaacgatTGTGATTCGATTCACATTtaaacaatttcaga
ECOUXU1      cccatgagagtgaaattgtTGTGATGTGGTTAACCCAattagaattcgggattgacatgtcttaccaaaaggta
PBR322  gtactgagagtgcaccatatgcggTGTGAAATACCGCACAGAtgcgtaaggagaaaataccgcatcaggcgctc
TRN9CAT ccctgggccaacttttggcgaaaaTGAGACGTTGATCGGCACg
TDC     tctggaaagtattgaaagttaattTGTGAGTGGTCGCACATAtcctgtt
```

fasta.bioch.virginia.edu/biol4559

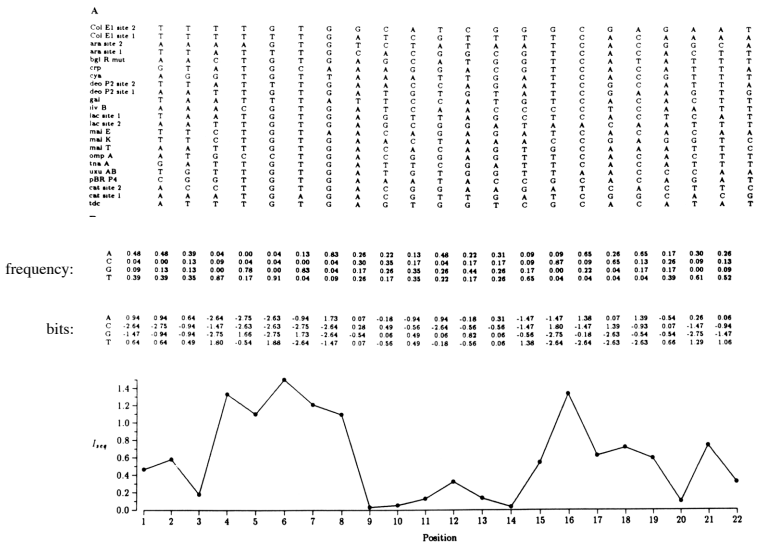


FIG. 1. CRP binding sites. (*A*) Twenty-three sites identified as binding to CRP (8, 9). (*B*) The frequency with which each base occurs at each position in the CRP binding sites. (*C*) The specificity matrix for the protein, based on the binding sites, which is calculated as $\log_2(f_b/p_b)$, where f_b is the observed frequency of each base (from the matrix above) and p_b is the *a priori* probability of obtaining base *b*. In this example p_b = 0.25 for all *b*, approximating the *E. coli* genomic composition. At positions for which $f_b = 0$, an estimated frequency of 0.5/23 is used in the calculation. (*D*) The "information content" at each position of the CRP binding sites is plotted (10). The sum of all positions is 13.06 bits.

Stormo GD, Hartzell GW (1989) Identifying protein-binding sites from unaligned DNA fragments.Proc Natl Acad Sci U S A. 86:1183-1187.

fasta.bioch.virginia.edu/biol4559

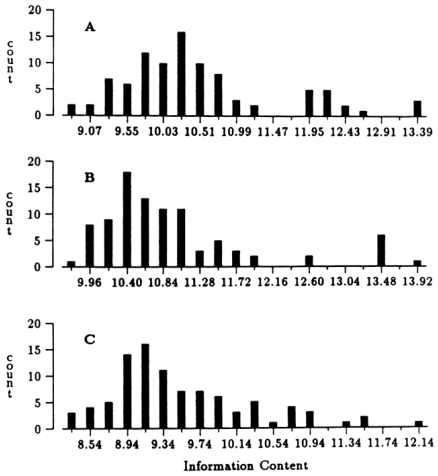


FIG. 3. Distribution of "information content" of the matrices at the end of each analysis, described in the text. Count is the number of matrices with "information content" in the interval shown. (*A*) The 94 20-wide matrices. (*B*) The 93 22-wide matrices. (*C*) The 93 16-wide matrices.

Stormo GD, Hartzell GW (1989) Identifying protein-binding sites from unaligned DNA fragments Proc Natl Acad Sci U S A. 86:1183-1187.

A

A	0.00	0.00	0.00	0.17	0.78	0.28	0.17	0.11	0.50	0.06	0.11	0.00	0.06	0.83	0.11	0.89
C	0.22	0.06	0.06	0.00	0.06	0.22	0.50	0.28	0.00	0.28	0.22	0.06	0.72	0.00	0.72	0.06
G	0.06	0.61	0.00	0.72	0.00	0.11	0.22	0.33	0.17	0.56	0.28	0.11	0.17	0.11	0.00	0.00
T	0.72	0.33	0.94	0.11	0.17	0.39	0.11	0.28	0.33	0.11	0.39	0.83	0.06	0.06	0.17	0.06

B

A	-3.43	-3.43	-3.43	-0.82	1.38	-0.10	-0.82	-1.45	0.74	-2.32	-1.45	-3.43	-2.32	1.47	-1.45	1.57
C	0.29	-1.59	-1.59	-0.97	-1.59	0.29	1.47	0.64	-2.70	0.64	0.29	-1.59	2.00	-2.70	2.00	-1.59
G	-1.61	1.54	-2.92	1.78	-2.92	-0.93	0.07	0.65	-0.31	1.42	0.42	-0.93	-0.31	-0.93	-2.92	-2.92
T	1.22	0.09	1.60	-1.50	-0.87	0.33	-1.50	-0.15	0.09	-1.50	0.33	1.42	-2.37	-2.37	-0.87	-2.37

C

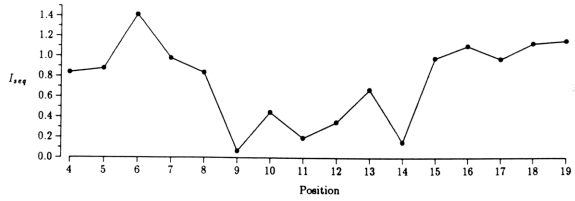


FIG. 4. The best 16-wide matrix. The positions are numbered 4–19, corresponding to the central positions of Fig. 1. (*A*) The frequency of each base for the sites included in the best matrix, as in Fig. 1*B*. (*B*) The specificity matrix determined from the frequency matrix, as in Fig. 1*C*. In this case the *a priori* values were determined from the data set shown in Fig. 2: $p_A = 0.30$; $p_C = 0.18$; $p_G = 0.21$; and $p_T = 0.31$. Analyses were also performed using $p_b = 0.25$ for all b. In that case, essentially the same matrix remained the best, but the distribution had more high-scoring matrices due to the high probability of adenine and thymine matches. The specificity matrix values for positions with $f_b = 0$ were estimated using $f_b = 0.5/18$ from the 18 sequences in the data set. (*C*) The "information content" at each position of the matrix is plotted. The sum from all positions is 12.15 bits.

Stormo GD, Hartzell GW (1989) Identifying protein-binding sites from unaligned DNA fragments Proc Natl Acad Sci U S A. 86:1183-1187.

Consensus (CRP)

```
consensus -L 16 -q 1000 -c0 -pr2 -pt 4 -pf 4

L-mer Width: 16

Top Matrices saved from each cycle: 4
    Matrices Saved from the last cycle: 4

sequence 1: CE1CG     fragments: 1-105
sequence 18: TDC      fragments: 1-105
Total number of sequences: 18.
Total number of sequence fragments: 18.

# Observed frequency and occurrence of each letter.
#anumber of letters in the input sequences = 1890
A  0.302646; observed occurrence =   572 (letter   1)
C  0.182540; observed occurrence =   345 (letter   2)
G  0.208995; observed occurrence =   395 (letter   3)
T  0.305820; observed occurrence =   578 (letter   4)

PRIOR FREQUENCIES DETERMINED BY OBSERVED FREQUENCIES.
* Information for the alphabet from the command line.
letter   1: A (complement: T)  prior freq = 0.302646
letter   2: C (complement: G)  prior freq = 0.182540
letter   3: G (complement: C)  prior freq = 0.208995 CYCLE
letter   4: T (complement: A)  prior freq = 0.305820 ------

INFORMATION CONTENT IS CALCULATED USING NATURAL
LOGARITHMS (i.e. BASE e). DIVIDE BY ln(2) = 0.693 TO
CONVERT TO BASE 2, WHICH WAS USED INPREVIOUS VERSIONS
OF THIS PROGRAM.
```

MATRICES SAVED FOR NEXT CYCLE

CYCLE	total number	top adjusted information	ln top p-value	ln expected frequency
1	1620	2.9492	0.0000	7.3902
2	748	6.8222	-13.5825	0.4475
3	849	8.2051	-20.1461	0.0577
4	832	8.7882	-26.3882	-0.3628
5	848	8.9275	-31.8728	-0.3179
6	857	9.1860	-38.8902	-2.0624
7	877	9.2908	-45.6680	-3.8014
8	864	9.1929	-51.3100	-4.6251
9	876	9.1152	-57.2498	-5.9597
10	879	9.0644	-63.5596	-7.8751
11	864	8.8973	-68.7012	-8.8353
12	854	8.8738	-75.4184	-11.5917
13	850	8.6817	-80.0738	-12.5205
14	873	8.5267	-85.0113	-13.9878
15	852	8.2955	-88.6578	-14.4562
16	865	8.1066	-92.6288	-15.6014
17	857	7.8793	-95.7075	-16.3204
18	836	7.6066	-97.6650	-16.6684


```
THE LIST OF TOP MATRICES FROM EACH CYCLE-- (total of 17):

MATRIX 1
number of sequences = 18
unadjusted information = 9.03227
sample size adjusted information = 7.60657
ln(p-value) = -97.665   p-value = 3.84279E-43
ln(expected frequency) = -16.6684   expected frequency = 5.76782E-08
A |   0   0   1   2  16   6   6   1   7   1   5   0   1  14   1  14
C |   2   2   1   1   2   5   4   4   0   4   4   0  15   0  15   2
G |   0  13   0  14   0   3   5   6   4  10   3   4   0   4   1   1
T |  16   3  16   1   0   4   3   7   7   3   6  14   2   0   1   1


 1|11  :   1/64    TTTGATCGTTTTCACA
 2|9   :   2/58    TTTGCACGGCGTCACA
 3|14  :   3/79    TGTGAGCATGGTCATA
 4|8   :   4/66    TGCAAAGGACGTCACA
 5|18  :   5/53    TGTTAAATTGATCACG
 6|10  :   6/10    TTTGAACCAGATCGCA
 7|13  :   7/54    TGTCACACTTTTCGCA
 8|1   :   8/23    TCTGCAATTCAGTACA
 9|4   :   9/12    TGTGAGTTAGCTCACT
10|5   :  10/17    TGTAACAGAGATCACA
11|15  :  11/64    CGTGATGTTGCTTGCA
12|6   :  12/44    TGTGACACAGTGCAAA
13|7   :  13/51    CCTGACGGAGTTCACA
14|12  :  14/74    TGTGATTCGATTCACA
15|16  :  15/20    TGTGATGTGGTTAACC
16|2   :  16/56    TGTGAAATACCGCACA
17|17  :  17/87    TGAGACGTTGATCGGC
18|3   :  18/81    TGTGAGTGGTCGCACA
```

Statistics for Consensus

```
consensus -L 16 -q 1000 -c0 -pr2 -pt 4 -pf 4

L-mer Width: 16

Top Matrices saved from each cycle: 4
    Matrices Saved from the last cycle: 4

sequence 1: CE1CG     fragments: 1-105
sequence 18: TDC      fragments: 1-105
Total number of sequences: 18.
Total number of sequence fragments: 18.

# Observed frequency and occurrence of each letter.
#anumber of letters in the input sequences = 1890
A  0.302646; observed occurrence =   572 (letter   1)
C  0.182540; observed occurrence =   345 (letter   2)
G  0.208995; observed occurrence =   395 (letter   3)
T  0.305820; observed occurrence =   578 (letter   4)

PRIOR FREQUENCIES DETERMINED BY OBSERVED FREQUENCIES.
* Information for the alphabet from the command line.
letter   1: A (complement: T)  prior freq = 0.302646
letter   2: C (complement: G)  prior freq = 0.182540
letter   3: G (complement: C)  prior freq = 0.208995
letter   4: T (complement: A)  prior freq = 0.305820

INFORMATION CONTENT IS CALCULATED USING NATURAL
LOGARITHMS (i.e. BASE e). DIVIDE BY ln(2) = 0.693 TO
CONVERT TO BASE 2, WHICH WAS USED INPREVIOUS VERSIONS
OF THIS PROGRAM.
```

MATRICES SAVED FOR NEXT CYCLE

CYCLE	total number	top adjusted information	ln top p-value	ln expected frequency
1	1620	2.9492	0.0000	7.3902
2	748	6.8222	-13.5825	0.4475
3	849	8.2051	-20.1461	0.0577
4	832	8.7882	-26.3882	-0.3628
5	848	8.9275	-31.8728	-0.3179
6	857	9.1860	-38.8902	-2.0624
7	877	9.2908	-45.6680	-3.8014
8	864	9.1929	-51.3100	-4.6251
9	876	9.1152	-57.2498	-5.9597
10	879	9.0644	-63.5596	-7.8751
11	864	8.8973	-68.7012	-8.8353
12	854	8.8738	-75.4184	-11.5917
13	850	8.6817	-80.0738	-12.5205
14	873	8.5267	-85.0113	-13.9878
15	852	8.2955	-88.6578	-14.4562
16	865	8.1066	-92.6288	-15.6014
17	857	7.8793	-95.7075	-16.3204
18	836	7.6066	-97.6650	-16.6684

Consensus from random sequences

MATRICES SAVED FOR NEXT CYCLE

CYCLE	total number	top adjusted information	ln top p-value	ln expected frequency
1	2040	1.4133	0.0000	7.6207
2	648	6.5863	-14.0510	0.4547
3	866	8.2103	-21.0667	-0.1259
4	819	8.6626	-26.3960	0.6457
5	849	8.6417	-30.6613	2.2093
6	871	8.7530	-36.3388	2.1271
7	872	8.7389	-41.7409	2.1122
8	878	8.6112	-46.5558	2.4937
9	875	8.4596	-51.2305	2.8370
20	937	6.4446	-88.4984	9.6257
21	943	6.3042	-91.2183	9.6902
22	955	6.1507	-93.4632	9.8955
23	947	5.9713	-94.9290	10.4300
24	969	5.8093	-96.4855	10.1381

```
MATRIX 1    number of sequences = 3
unadjusted information = 18.0219
sample size adjusted information = 8.21031
ln(p-value) = -21.0667   p-value = 7.09314E-10
ln(expected frequency) = -0.125938   expected frequency = 0.88167
A |   0   0   0   3   1   0   0   3   0   0   0   0   3   0   1   0
C |   0   0   3   0   0   2   0   0   0   3   1   0   0   1   2   0
G |   0   3   0   0   0   0   1   0   3   0   1   0   0   2   0   3
T |   3   0   0   0   2   1   2   0   0   0   1   3   0   0   0   0
```

Consensus from random sequences

```
      []            MATRICES SAVED FOR NEXT CYCLE             []
      []------------------------------------------------------[]
      []  total    | top adjusted |   ln top    [] ln expected []
CYCLE []  number   | information  |   p-value   []  frequency  []
------[]----------|--------------|-------------[]-------------[]
    1 []     2040 |       1.4133 |      0.0000 []      7.6207 []
    2 []      648 |       6.5863 |    -14.0510 []      0.4547 []
    3 []      866 |       8.2103 |    -21.0667 []     -0.1259 []
    4 []      819 |       8.6626 |    -26.3960 []      0.6457 []
    5 []      849 |       8.6417 |    -30.6613 []      2.2093 []
    6 []      871 |       8.7530 |    -36.3388 []      2.1271 []
    7 []      872 |       8.7389 |    -41.7409 []      2.1122 []
    8 []      878 |       8.6112 |    -46.5558 []      2.4937 []
    9 []      875 |       8.4596 |    -51.2305 []      2.8370 []
   20 []      937 |       6.4446 |    -88.4984 []      9.6257 []
   21 []      943 |       6.3042 |    -91.2183 []      9.6902 []
   22 []      955 |       6.1507 |    -93.4632 []      9.8955 []
   23 []      947 |       5.9713 |    -94.9290 []     10.4300 []
   24 []      969 |       5.8093 |    -96.4855 []     10.1381 []
```

```
MATRIX 1    number of sequences = 3
unadjusted information = 18.0219
sample size adjusted information = 8.21031
ln(p-value) = -21.0667   p-value = 7.09314E-10
ln(expected frequency) = -0.125938   expected frequency = 0.88167
A |   0   0   0   3   1   0   0   3   0   0   0   0   3   0   1   0
C |   0   0   3   0   0   2   0   0   0   3   1   0   0   1   2   0
G |   0   3   0   0   0   0   1   0   3   0   1   0   0   2   0   3
T |   3   0   0   0   2   1   2   0   0   0   1   3   0   0   0   0
```


Statistical Strategies for Consensus Alignment - EM and Gibbs

- A problem of estimation with hidden data - the positions are easy to find if the consensus is known, and the consensus is easy to find if the positions are known
- Start with random positions, build a consensus estimate
- Apply consensus to sequences, assign probability of being a consensus, repeat
- Gibbs is similar, but a target sequence is left out and scanned at each stage

Expectation Maximization for Consensus Alignment

(1) Begin with a set of sequences:

```
CE1CG      taatgtttgtgctggtttttgtggc
ECOARABOP  gacaaaaacgcgtaacaaaagtgtc
ECOBGLR1   acaaatcccaataacttaattattg
ECOCRP     cacaaagcgaaagctatgctaaaac
ECOCYA     acggtgctacacttgtatgtagcgc
ECODEOP2   agtgaattatttgaaccagatcgca
ECOGALE    gcgcataaaaaacggctaaattctt
ECOILVBPR  gctccggcggggttttttgttatct
```

(2) Select consensus sites at random:

```
CE1CG      taatGTTTGtgctggtttttgtggc
ECOARABOP  gacaaaaacgcgTAACAaaagtgtc
ECOBGLR1   acaaatcccAATAActtaattattg
ECOCRP     cacaaagcgaaagctatgctAAAAC
ECOCYA     ACGGTgctacacttgtatgtagcgc
ECODEOP2   agtgAATTAtttgaaccagatcgca
ECOGALE    gcgcataaAAAACggctaaattctt
ECOILVBPR  gctccggcggggttttttgtTATCT
```

(3) Use the consensus to build a matrix

```
CE1CG      GTTTG
ECOARABOP  TAACA
ECOBGLR1   AATAA
ECOCRP     AAAAC
ECOCYA     ACGGT
ECODEOP2   AATTA
ECOGALE    AAAAC
ECOILVBPR  TATCT
```

A	5	6	3	3	3
C	0	1	0	2	2
G	1	0	1	1	1
T	2	1	4	2	2

f_{ns}		f_1	f_2	f_3	f_4	f_5
0.3	A	0.6	0.8	0.4	0.4	0.4
0.2	C	0.0	0.1	0.0	0.2	0.3
0.2	G	0.2	0.0	0.1	0.1	0.1
0.3	T	0.2	0.1	0.5	0.3	0.2

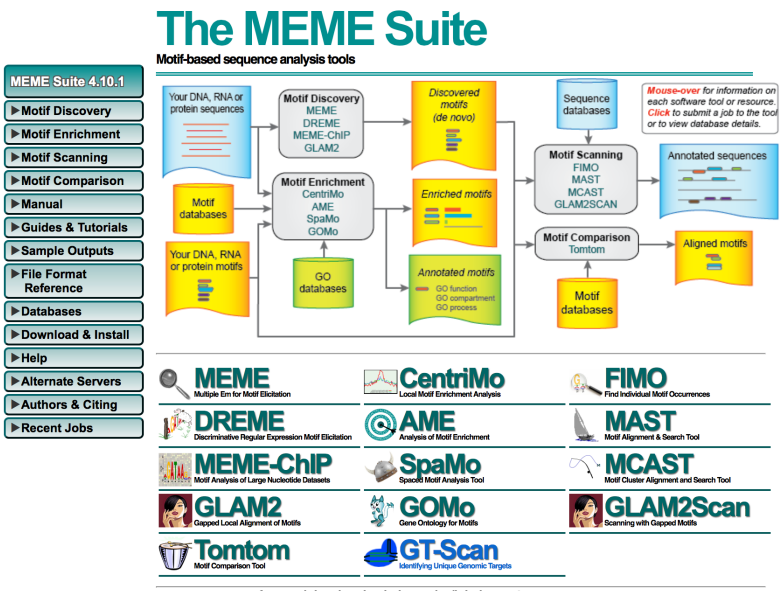
(4) Use the consensus to weight each position

$$p_i = \sum_{j=0..5} f_{sb} + \sum_{non-site} f_{ns}$$

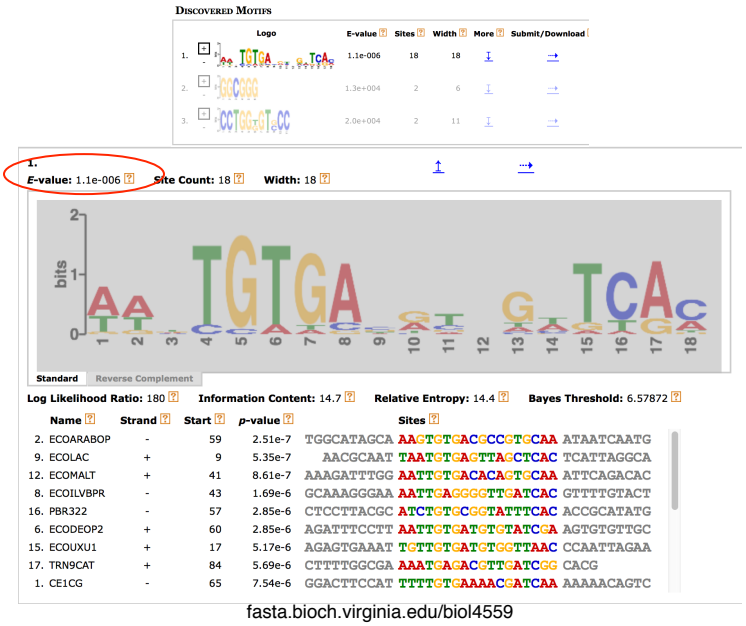
(5) Use the weighted site to build a new consensus

(6) Repeat (3..5)

MEME: http://meme-suite.org



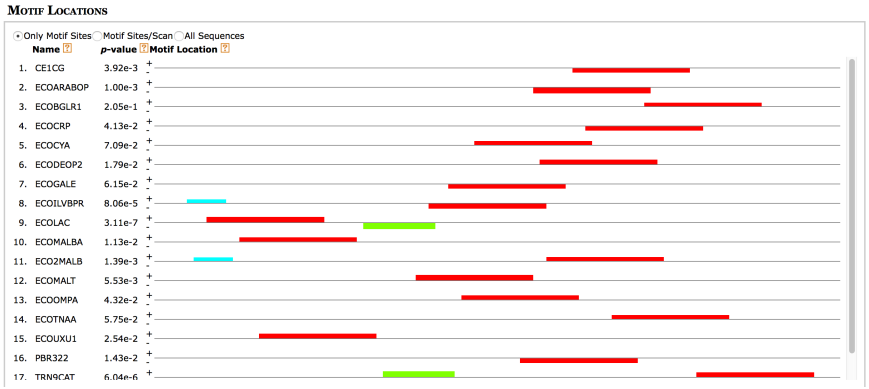
MEME Results: CRP sites



fasta.bioch.virginia.edu/biol4559

43

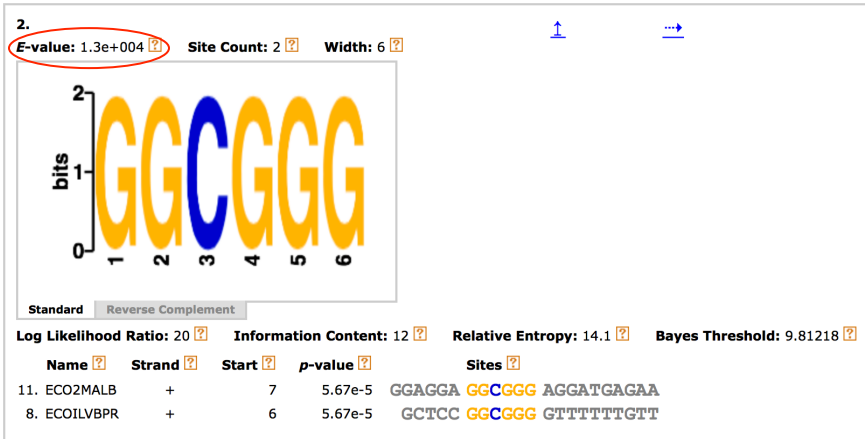
MEME Results: CRP site positions



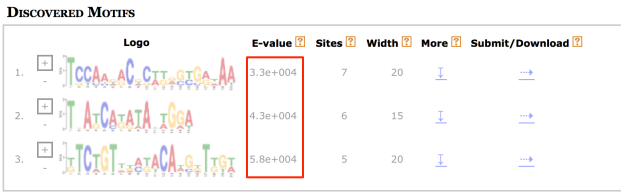
fasta.bioch.virginia.edu/biol4559

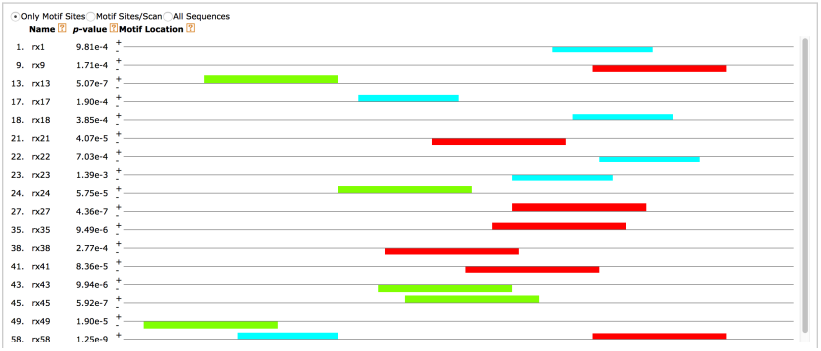
44

MEME Results: CRP - Motif 2



MEME Results: Random





Gibb's sampling for Consensus Alignment

(1) Begin with a set of sequences with randomly located motifs:

```
CE1CG      taatGTTTGtgctggtttttgtggc
ECOARABOP  gacaaaaacgcgTAACAaaagtgtc
ECOBGLR1   acaaatcccAATAActtaattattg
ECOCRP     cacaaagcgaaagctatgctAAAAC
ECOCYA     ACGGTgctacacttgtatgtagcgc
ECODEOP2   agtgAATTAtttgaaccagatcgca
ECOGALE    gcgcataaAAAACggctaaattctt
ECOILVBPR  gctccggcggggttttttgtTATCT
```

(2) Exclude one of the sequences at random, and build a consensus matrix from the other motifs

```
CE1CG      taatGTTTGtgctggtttttgtggc
ECOARABOP  gacaaaaacgcgTAACAaaagtgtc
ECOBGLR1   acaaatcccAATAActtaattattg
ECOCRP     cacaaagcgaaagctatgctaaaac
ECOCYA     ACGGTgctacacttgtatgtagcgc
ECODEOP2   agtgAATTAtttgaaccagatcgca
ECOGALE    gcgcataaAAAACggctaaattctt
ECOILVBPR  gctccggcggggttttttgtTATCT
```

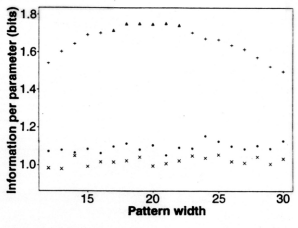
(3) Using the probability matrix from the included sequences, calculate the probability of each site on the excluded sequence

```
ECOCRP     cacaaagcgaaagctatgctaaaac
```

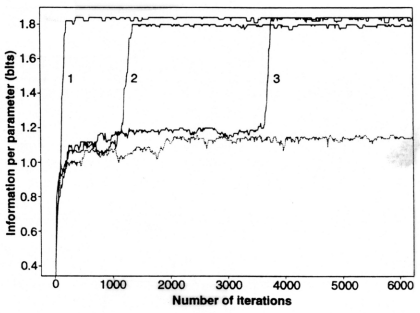
(4) Select a site at random, using weights from the probabilities in (3)

(5) Repeat steps (2) - (4)

g. 2. Information per parameter as the criterion pattern width for helix-turn-helix (HTH) pro- ns. The points indicate the maximum values of ormation per parameter found by the algo- ım. The upper points (▲ and +) used the mplete sequences of the 30 HTH proteins ted in Fig. 1A. (▲) All of the sequences in the ta set were aligned in the correct register (as Fig. 1A). (+) One or more of the sequences in data set were incorrectly aligned. All com- tely correct alignments in the width range m 17 to 22 residues gave greater values of ormation per parameter than any incorrect gnments outside this width range. (●) The nsites" sequence data of the 30 HTH proteins, constructed by deleting the 18 residues of the H pattern itself (Fig. 1A) from each of the sequences. (×) A shuffled data set (*46*) of the 30 HTH quences. The alignments from the nonsites background of the HTH proteins give values slightly ater than random expectation.

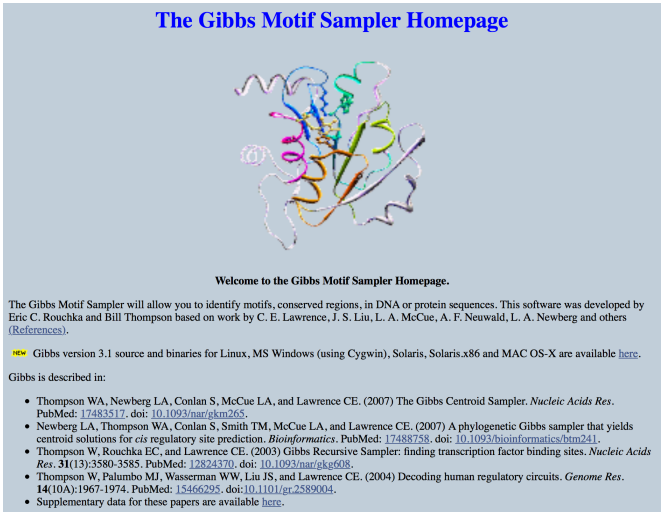


. 3. Convergence behav- of the Gibbs sampling orithm. Because the bs sampler, when run for e time, is a heuristic rath- han a rigorous optimiza- procedure, one cannot rantee the optimality of results it produces. refore, the best solution nd in a series of runs will called "maximal." A sin- pattern of width 18 resi- s was sought in the data of 30 HTH proteins wn in Fig. 1A. Solid lines w the course of three pendent runs with dif- nt random seeds. Evolv-



Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.Science 262:208-214.

Gibbs: ccmbweb.ccv.brown.edu/gibbs/gibbs.html



The Gibbs Motif Sampler Homepage

Welcome to the Gibbs Motif Sampler Homepage.

The Gibbs Motif Sampler will allow you to identify motifs, conserved regions, in DNA or protein sequences. This software was developed by Eric C. Rouchka and Bill Thompson based on work by C. E. Lawrence, J. S. Liu, L. A. McCue, A. F. Neuwald, L. A. Newberg and others (References).

NEW Gibbs version 3.1 source and binaries for Linux, MS Windows (using Cygwin), Solaris, Solaris.x86 and MAC OS-X are available here.

Gibbs is described in:

- Thompson WA, Newberg LA, Conlan S, McCue LA, and Lawrence CE. (2007) The Gibbs Centroid Sampler. *Nucleic Acids Res*. PubMed: 17483517. doi: 10.1093/nar/gkm265.
- Newberg LA, Thompson WA, Conlan S, Smith TM, McCue LA, and Lawrence CE. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for *cis* regulatory site prediction. *Bioinformatics*. PubMed: 17488758. doi: 10.1093/bioinformatics/btm241.
- Thompson W, Rouchka EC, and Lawrence CE. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res*. **31**(13):3580-3585. PubMed: 12824370. doi: 10.1093/nar/gkg608.
- Thompson W, Palumbo MJ, Wasserman WW, Liu JS, and Lawrence CE. (2004) Decoding human regulatory circuits. *Genome Res*. **14**(10A):1967-1974. PubMed: 15466295. doi:10.1101/gr.2589004.
- Supplementary data for these papers are available here.

Finding consensus regions in unaligned sequences

- Some introduction: regulation of transcription
- Looking for functional sites: promoters, regulatory elements, modification sites
- Products of *convergent*, not divergent evolution
- Weak spacing constraints
- Usually represented as a consensus sequence
- If alignment is given, consensus is obvious
- If consensus is given, alignment is obvious
- Search for consensus and alignment together
- `consensus, meme, gibbs`