CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS11.html

MSA: Progressive Method

- Perform global pairwise alignments
- Build guide tree
- Progressively align the sequences

How to Score Multiple Alignments?

- Sum of Pairs Score (SP)
  - Optimal alignment: $O(d^N)$ [Dynamic Prog]
  - Approximate Algorithm: Approx Ratio 2
    - Locate Center: $O(d^2N^2)$
    - Locate Consensus: $O(d^2N^2)$

Consensus char: char with min distance sum

Consensus string: string of consensus char

Center: input string with min distance sum

Multiple Alignment Methods

- Phylogenetic Tree Alignment (NP-Complete)
  - Given tree, task is to label leaves with strings
- Iterative Method(s)
  - Build a MST using the distance function
- Clustering Methods
  - Hierarchical Clustering
  - K-Means Clustering

Multiple Alignment Methods (Cont'd)

- Gibbs Sampling Method
  - Lawrence, Altschul, Boguski, Liu, Neuwald, Winton, *Science*, 1993
- Hidden Markov Model
  - Krogh, Brown, Mian, Sjolander, Haussler, *JMB*, 1994

Multiple Sequence Alignments (MSA)

- Choice of Scoring Function
  - Global vs local
  - Gap penalties
  - Substitution matrices
  - Incorporating other information
  - Statistical Significance
- Computational Issues
  - Exact/heuristic/approximate algorithms for optimal MSA
  - Progressive/Iterative/DP
  - Iterative: Stochastic/Non-stochastic/Consistency-based
- Evaluating MSAs
  - Choice of good test sets or benchmarks (BAliBASE)
  - How to decide thresholds for good/bad alignments

MSA Properties/Features/Uses

- No "correct" alignment; Hard to identify good alignments
- Good MSAs suggest shared secondary structures, motifs, good structure alignment, similar functions and strong homology
- For two proteins with 30% amino acid identity, 50% of structure may align
- Some residues (e.g., C) are highly conserved
- Many regions share consistent patterns of indels
- SNPs can be detected from MSA of population data
- Query can be searched against database of MSAs (Pfam)
- Regulatory regions of genes may have consensus sequences identifiable by MSA (TFBS)

MSA: Conclusions

- Significant uses
  - Phylogenetic analyses
  - Identify members of a family, SNPs, motifs, TFBS
  - Protein structure prediction
- No perfect methods
- Popular
  - Progressive methods: CLUSTALW
  - Recent interesting ones: Prrp, SAGA, DiAlign, T-Coffee
- Review of Methods [C. Notredame, *Pharmacogenomics,* **3**(1), 2002]
  - CLUSTALW works reasonably well, in general
  - DiAlign is better for sequences with long insertions & deletions (indels)
  - T-Coffee is best available method

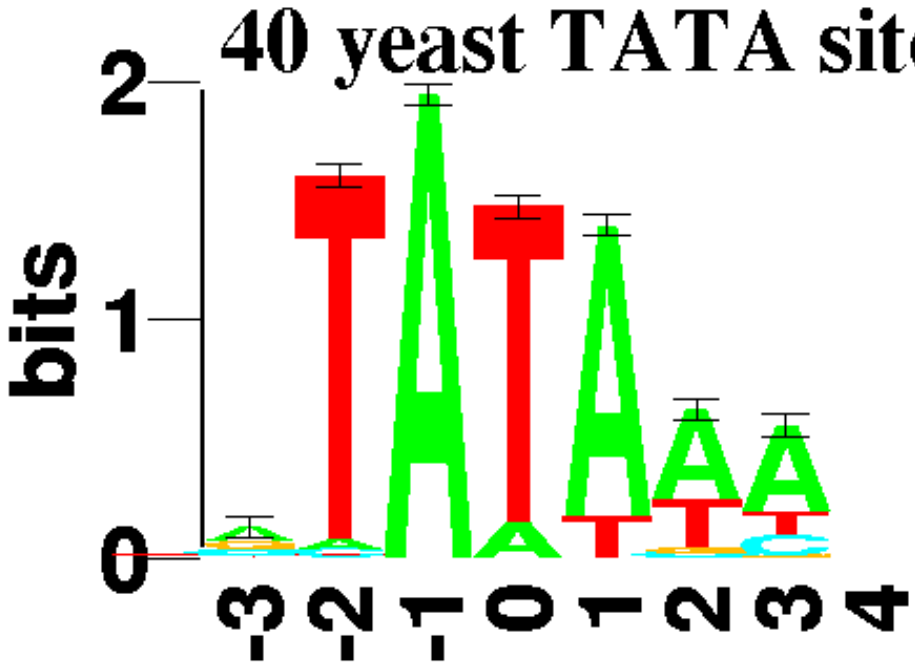
Describing & Modeling Patterns

Patterns in DNA Sequences

- Signals in DNA sequence control events
  - Start and end of genes
  - Start and end of introns
  - Transcription factor binding sites (regulatory elements)
  - Ribosome binding sites
- Detection of these patterns are useful for
  - Understanding gene structure
  - Understanding gene regulation

Motifs in DNA Sequences

- Given a collection of DNA sequences of promoter regions, describe the transcription factor binding sites (also called regulatory elements)
  - Example:



http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

Motifs in DNA Sequences

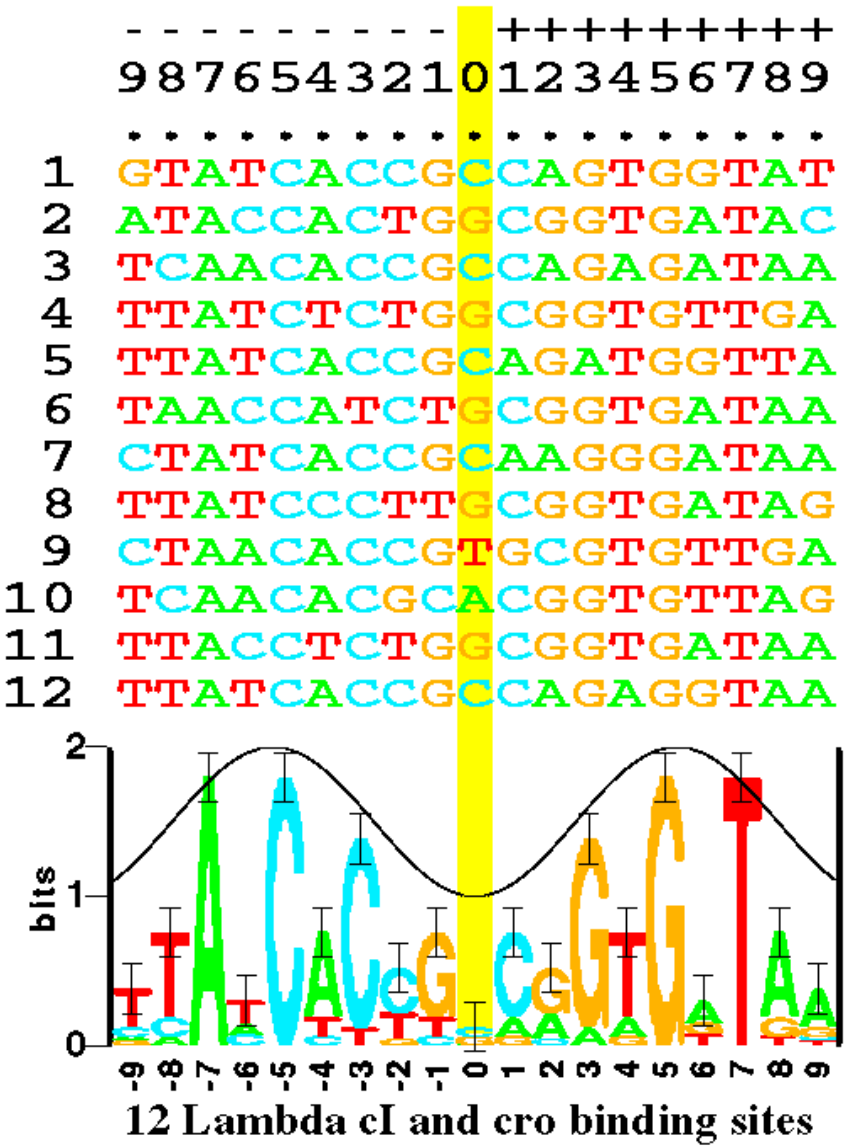
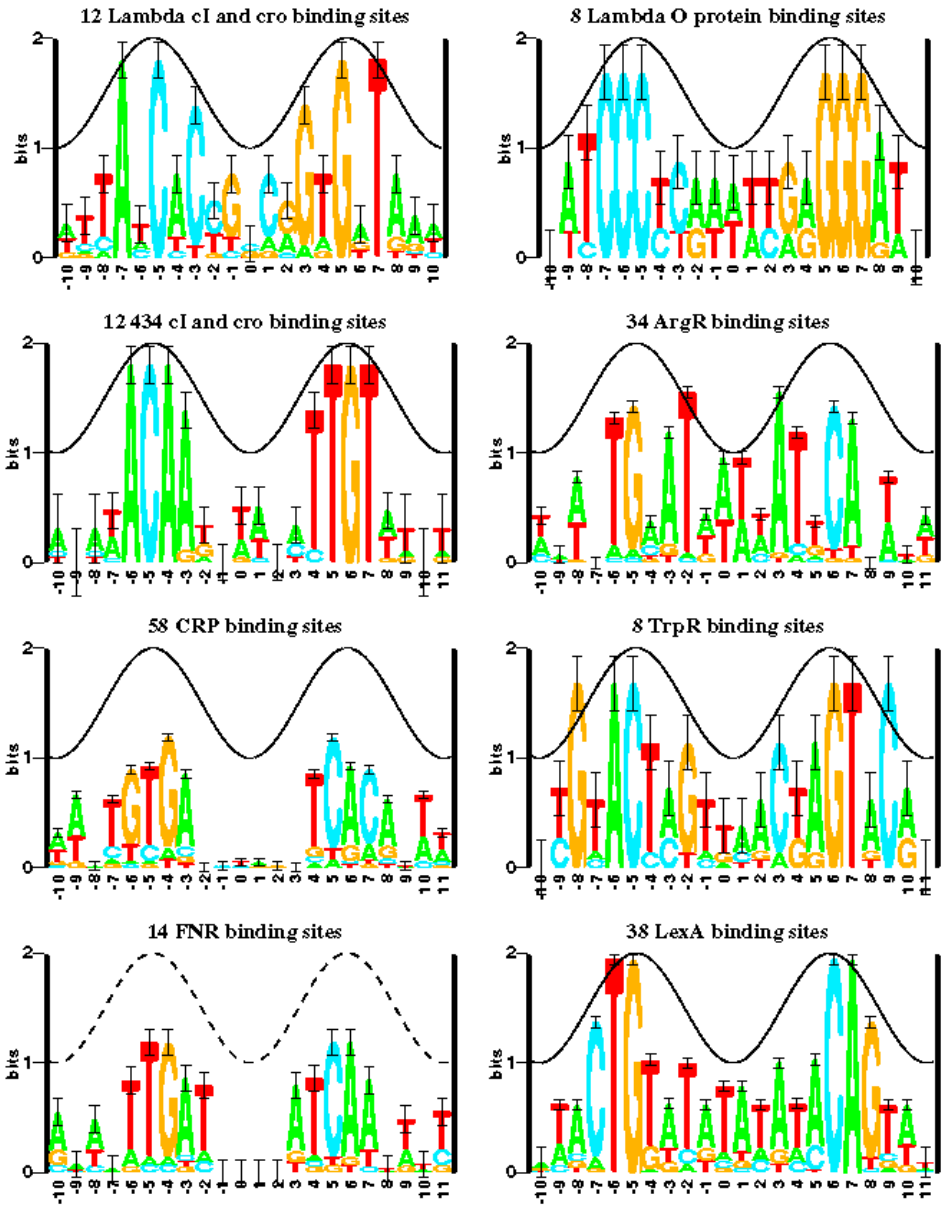


Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered sequence. The sequence logo, described in detail in the text, is at the bottom of the figure. The cosine wave is positioned to indicate that a minor groove faces the center of each symmetrical protein. Data which support this assignment are given in reference [17].

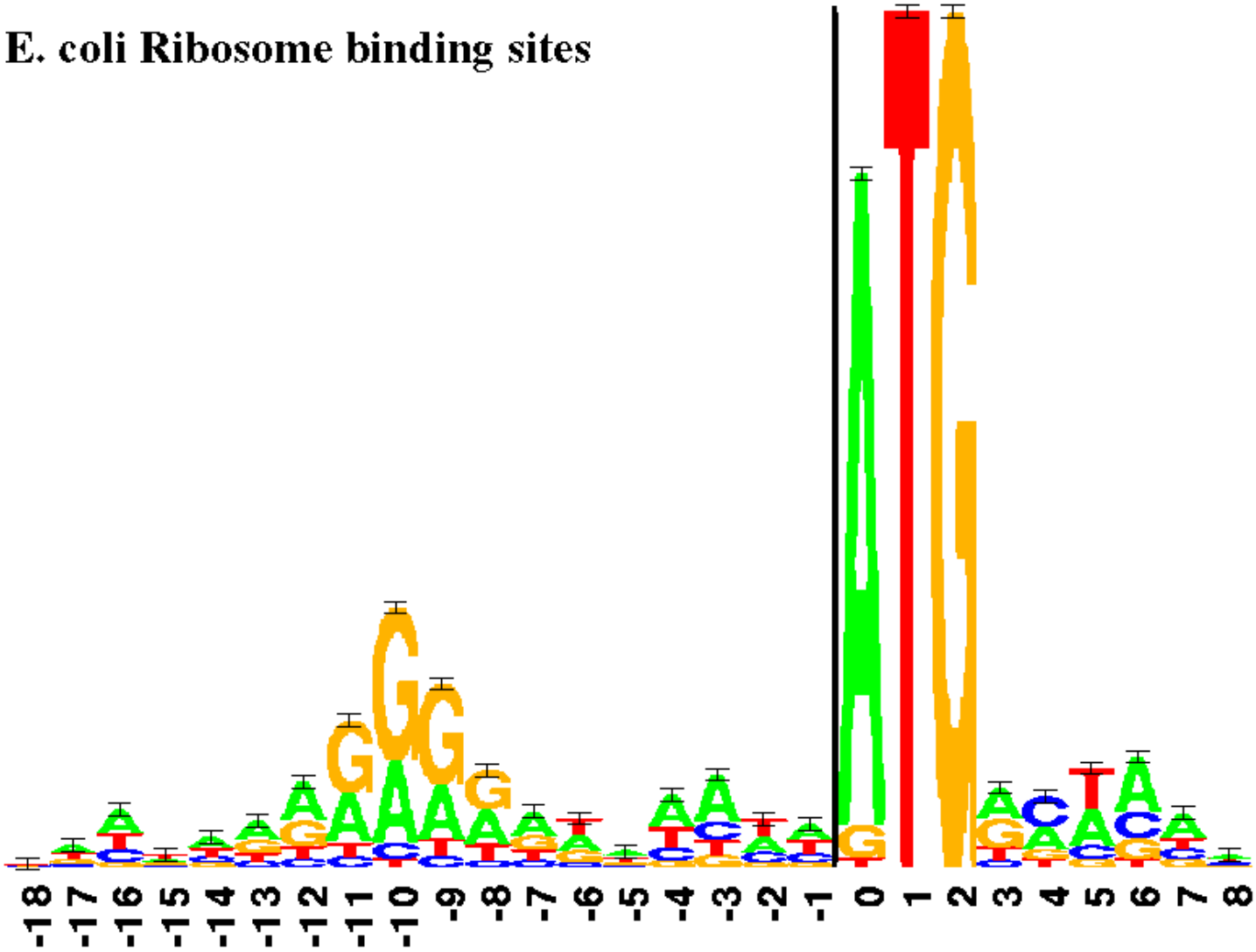
http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

More Motifs in *E. Coli* DNA Sequences



http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

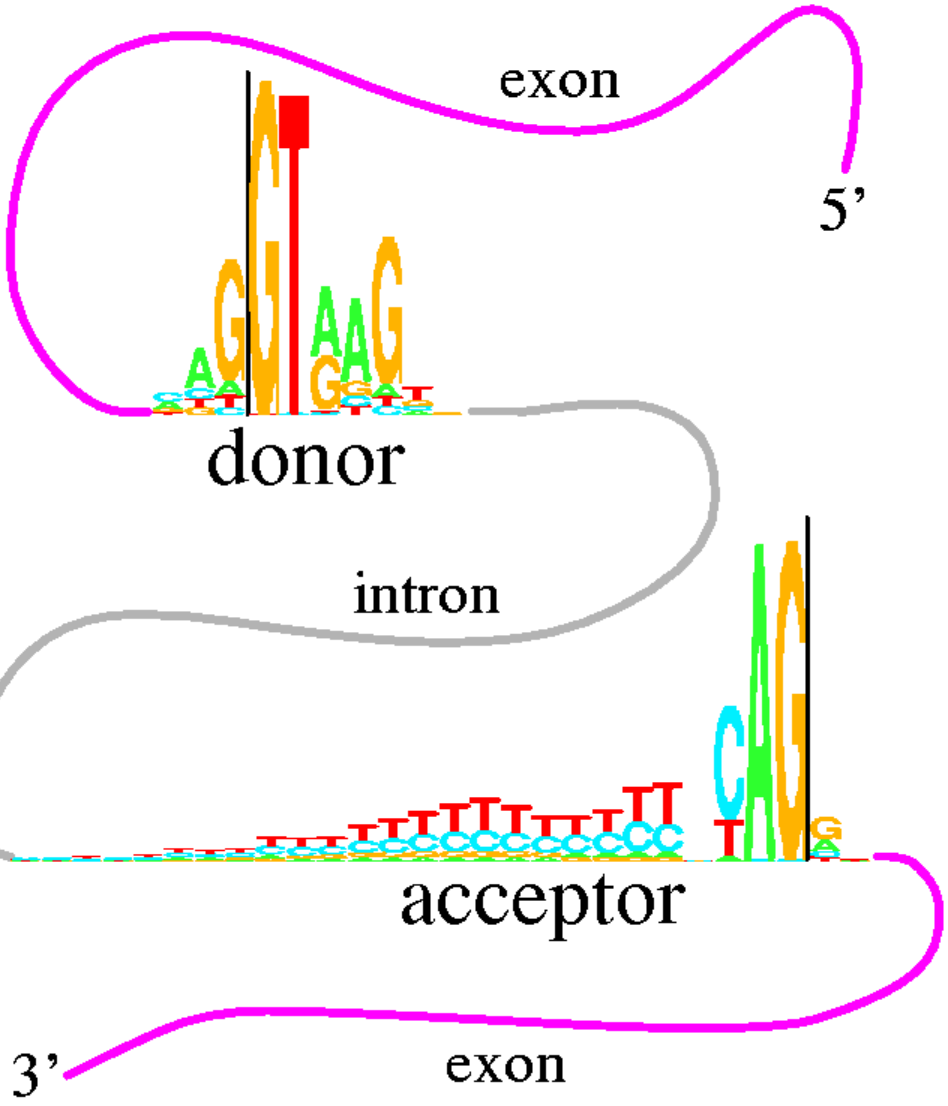
E. coli Ribosome binding sites



http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

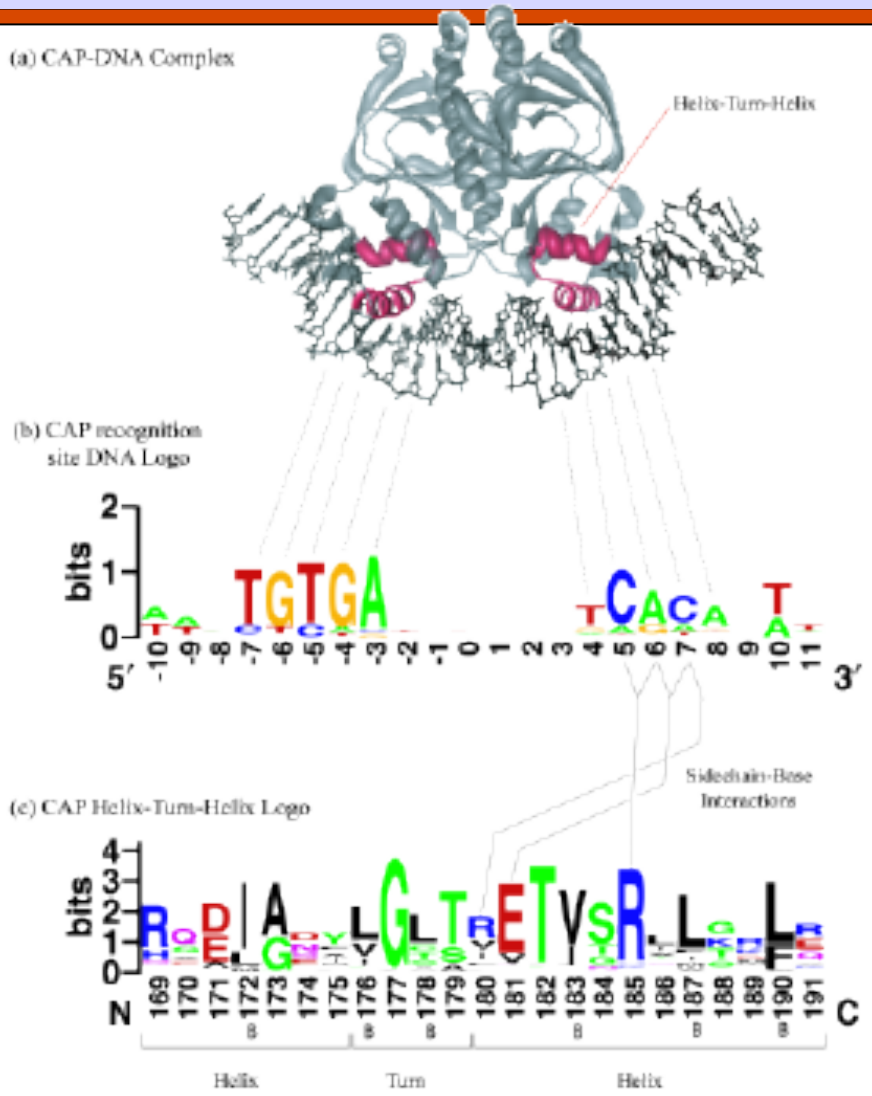
Other Motifs in DNA Sequences: Human Splice Junctions

This figure shows two "sequence logos" which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern "CAG|GT", which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992)



http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

Motifs



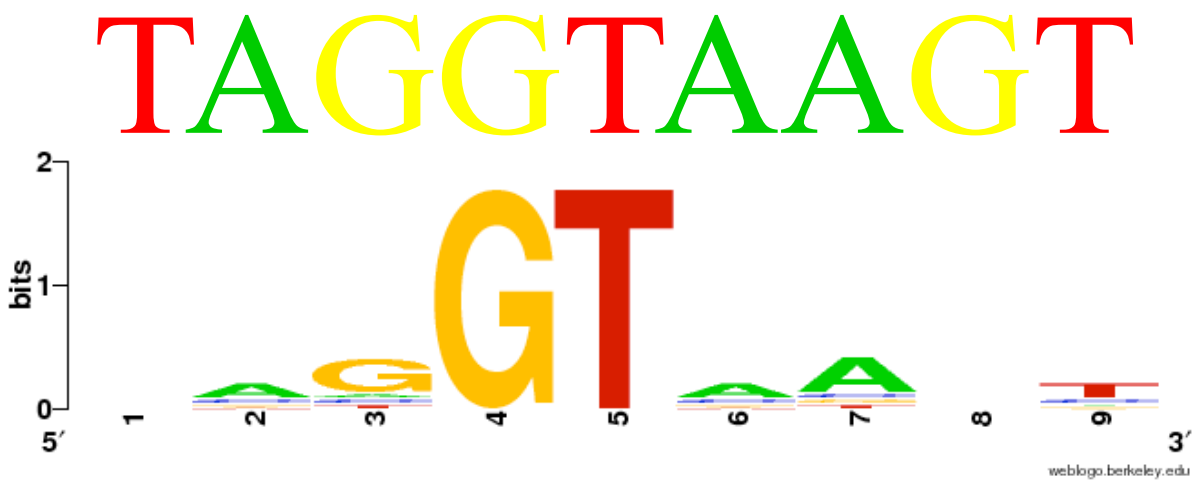
http://weblogo.berkeley.edu/examples.html

Pattern: Representations

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

TAGGTAAGT

- Alignments
- Consensus Sequences
- Logo Formats
- ...



Profiles

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

	1	2	3	4	5	6	7	8	9
A	3	6	1	0	0	6	7	2	1
C	2	2	1	0	0	2	1	1	2
G	1	1	7	10	0	1	1	5	1
T	4	1	1	0	10	1	1	2	6

Frequency Matrix

	1	2	3	4	5	6	7	8	9
A	.3	.6	.1	0	0	.6	.7	.2	.1
C	.2	.2	.1	0	0	.2	.1	.1	.2
G	.1	.1	.7	1	0	.1	.1	.5	.1
T	.4	.1	.1	0	1	.1	.1	.2	.6

Relative Frequencies

Profiles

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

	1	2	3	4	5	6	7	8	9
A	.3	.6	.1	0	0	.6	.7	.2	.1
C	.2	.2	.1	0	0	.2	.1	.1	.2
G	.1	.1	.7	1	0	.1	.1	.5	.1
T	.4	.1	.1	0	1	.1	.1	.2	.6

Relative Frequencies

	1	2	3	4	5	6	7	8	9
A	0.14	0.72	-0.61	-1.43	-1.43	0.72	0.86	-0.16	-0.61
C	-0.16	-0.16	-0.61	-1.43	-1.43	-0.16	-0.61	-0.61	-0.16
G	-0.61	-0.61	0.86	-0.61	-1.43	-0.61	-0.61	0.57	-0.61
T	0.38	-0.61	-0.61	-1.43	1.19	-0.61	-0.61	-0.16	0.72

Profiles

Profile entries:

$$P_{ij} = \ln (f_{ij}/b_i)$$

Zero counts:

$$f_{ij} = (c_{ij}+\alpha b_i)/(n+\alpha)$$

	1	2	3	4	5	6	7	8	9
A	.3	.6	.1	0	0	.6	.7	.2	.1
C	.2	.2	.1	0	0	.2	.1	.1	.2
G	.1	.1	.7	1	0	.1	.1	.5	.1
T	.4	.1	.1	0	1	.1	.1	.2	.6

Relative Frequencies

	1	2	3	4	5	6	7	8	9
A	0.14	0.72	-0.61	-1.43	-1.43	0.72	0.86	-0.16	-0.61
C	-0.16	-0.16	-0.61	-1.43	-1.43	-0.16	-0.61	-0.61	-0.16
G	-0.61	-0.61	0.86	1.19	-1.43	-0.61	-0.61	0.57	-0.61
T	0.38	-0.61	-0.61	-1.43	1.19	-0.61	-0.61	-0.16	0.72

http://coding.plantpath.ksu.edu/profile/