CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS07.html

Pattern Discovery

Patterns

- ❑ Nature stumbles upon recipes to accomplish tasks.
- ❑ With high probability, such recipes are reused.
- ❑ This causes the recipe to be conserved through evolution.
- ❑ Such recipes give rise to **patterns**.

Why Pattern Discovery?

- Modern Biomedical Research
  - Generates a “ton of data”.
  - Use analytical tools to find patterns in data.
- Pattern Discovery facilitates this process!
  - Pattern Discovery in sequences
  - Pattern Discovery in structures
  - Pattern Discovery in quantitative data
- Patterns help to detect members of a class
- Patterns help to characterize classes

Sequence Patterns: Examples

- Protein active sites and functional domains
  - For e.g., Zinc-finger motifs & Helix-turn-helix motifs
- Protein family signatures
- Signals in DNA e.g., protein binding sites
- MicroRNA and Anti-sense RNA

Example 1: Protein Motifs

- ❑ DNA-binding motifs
  - Helix-turn-Helix
- ❑ Motifs in Cys_2His_2-Zinc-binding proteins

Example: Zinc Finger Motif

...**Y**K**C**GL**C**ERS**F**VEKSA**L**SR**H**ORV**H**KN...

3 **6** **19** **23**

- ❑ Motifs in proteins that bind to [4Fe-4S]-complex

Example: Ferredoxin subfamily

...**C**xx**C**xx**C**xxx**CP**...

How to Represent Patterns

- ❑ Consensus sequence
- ❑ Alignments
- ❑ LOGO format
- ❑ Frequency Matrices
- ❑ Weight Matrices (Profiles, PSSMs, PWMs)

Pattern Representations

- Consensus sequences

[Pribnow, 1975]

```
TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT
------
TATAAT Consensus
```

```
TATRNT Consensus
       w/ IUPAC
```

```
TATAAT Multi-level
G CGC  Consensus
    T
```

Needs Alignment

Pattern Representations

- Consensus sequences
- Weight Matrices (Profiles, PSSMs)
  - Frequency Counts
  - Relative Frequency Measures
  - Normalized Measures
  - Log-transformed Measures
  - Information content
  - "Logo" technique
  - HMMs

Pattern Representation: Weight Matrix

Alignment

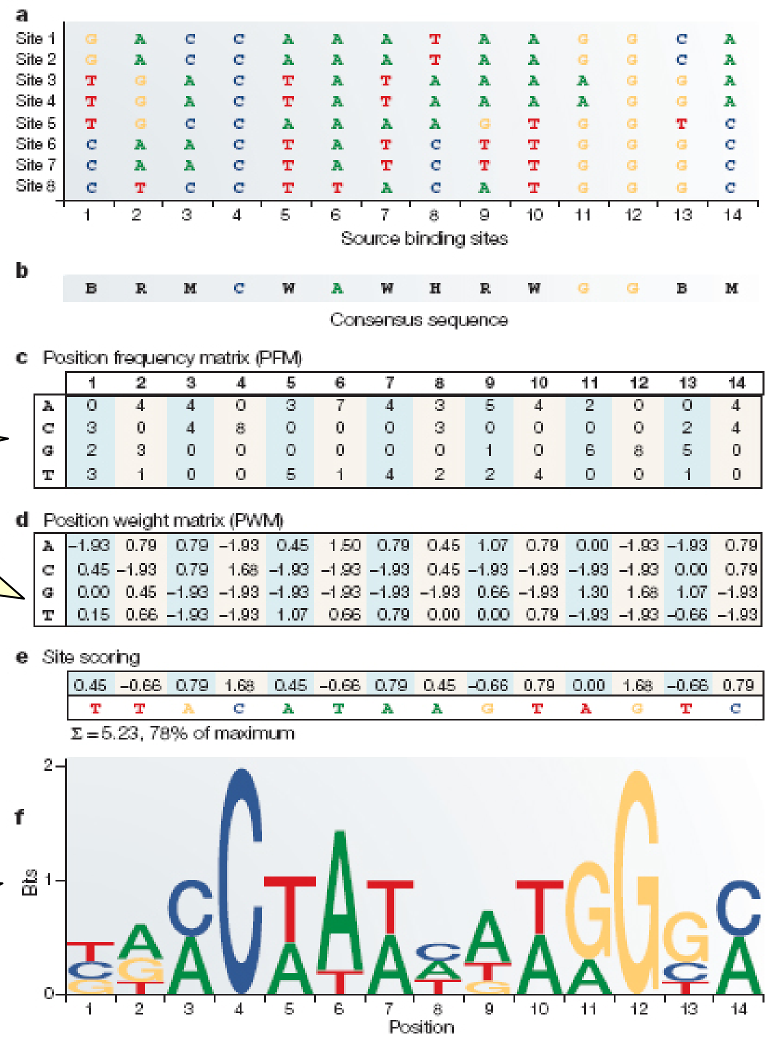
Consensus

Frequencies

Profile/ PSSM/PWM

Scoring a sequence against a profile

Visualizing a profile



a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

d Position weight matrix (PWM)

A	–1.93	0.79	0.79	–1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	–1.93	–1.93	0.79
C	0.45	–1.93	0.79	1.68	–1.93	–1.93	–1.93	0.45	–1.93	–1.93	–1.93	–1.93	0.00	0.79
G	0.00	0.45	–1.93	–1.93	–1.93	–1.93	–1.93	–1.93	0.66	–1.93	1.30	1.68	1.07	–1.93
T	0.15	0.66	–1.93	–1.93	1.07	0.66	0.79	0.00	0.00	0.79	–1.93	–1.93	–0.66	–1.93

e Site scoring

0.45	–0.66	0.79	1.68	0.45	–0.66	0.79	0.45	–0.66	0.79	0.00	1.68	–0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

Σ = 5.23, 78% of maximum

f

[Wasserman, Sandelin, Nat Genet, 2004]

Formulae

- ❑ Prob of char b in position i:

$$p(b,i) = \frac{f_{b,i}}{N}$$

Frequency

# Sequences

- ❑ Corrected prob:

$$P(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{a \in \mathbf{A}} s(a)}$$

PseudoCount

- ❑ Weight matrix entry:

$$W_{b,i} = \log_2 \frac{P(b,i)}{BP(b)}$$

Background Frequency

- ❑ Information content of position of i:

$$D_i = 2 + \sum_b P(b,i) \log_2 P(b,i)$$

[Wasserman, Sandelin, Nat Genet, 2004]

Statistical Evaluation Fundamentals

- Probability of finding a sequence w in some position of a DNA/protein sequence (assuming independence at each position)

$$\Pr(w) = \prod_{i=1}^{m} \Pr(w_i)$$

- $\Pr(w_i)$ = BP(b) [Background Frequency]

Statistical Evaluation

- ❑ Z-score of a motif with a certain frequency:

$$z(w) = \frac{Obs(w) - Exp(w)}{\sqrt{Var(w)}}$$

- ❑ Information Content or Relative Entropy of an alignment or profile:

$$IC(M) = \sum_{i=1}^{4}\sum_{j=1}^{m} m_{i,j} \log \frac{m_{i,j}}{b_i}$$

- ❑ Maximum a Posteriori (MAP) Score:

$$MAP(M) = -\sum_{i=1}^{4}\sum_{j=1}^{m} n_{i,j} \log \frac{m_{i,j}}{b_i}$$

- ❑ Model Vs Background Score:

$$L(w) = \frac{\Pr(w \mid M)}{\Pr(w \mid Bg)} = \prod_{j=1}^{m} \frac{m_{i,j}}{b_i}$$

Pattern Discovery in Protein Sequences

Motifs are combinations of secondary structures in proteins with a specific **structure** and a specific **function**. They are also called **super-secondary structures**.

Examples: Helix-Turn-Helix, Zinc-finger,
Homeobox domain, Hairpin-beta motif,
Calcium-binding motif, Beta-alpha-beta motif,
Coiled-coil motifs.

Several motifs may combine to form **domains**.

- Serine proteinase domain, Kringle domain, calcium-binding domain, homeobox domain.

Motif Detection

- Profile Method
  - If many examples of the motif are known, then
    - Training: build a Profile and compute a threshold
    - Testing: score against profile
- Combinatorial Pattern Discovery Methods
- Gibbs Sampling
- Expectation Method
- HMM

How to evaluate these methods?

- Calculate TP, FP, TN, FN
- Compute sensitivity fraction of known sites predicted, specificity, and more.
  - Sensitivity = TP/(TP+FN)
  - Specificity = TN/(TN+FN)
  - Positive Predictive Value = TP/(TP+FP)
  - Performance Coefficient = TP/(TP+FN+FP)
  - Correlation Coefficient =

Motif Detection Problem

Input: Set, S, of known (aligned) examples of a motif M,
A new protein sequence, P.

Output: Does P have a copy of the motif M?

Example: Zinc Finger Motif

...**Y**K**C**GL**C**ERS**F**VEKSA**L**SR**H**ORV**H**KN...

3 6 19 23

Input: Database, D, of known protein sequences,
A new protein sequence, P.

Output: What interesting patterns from D are present in P?

Supervised Pattern Discovery

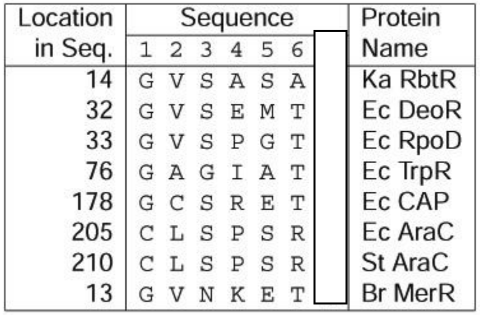
- Input: Alignment of known motifs, and Query sequence

  Output: Is the query sequence a motif?

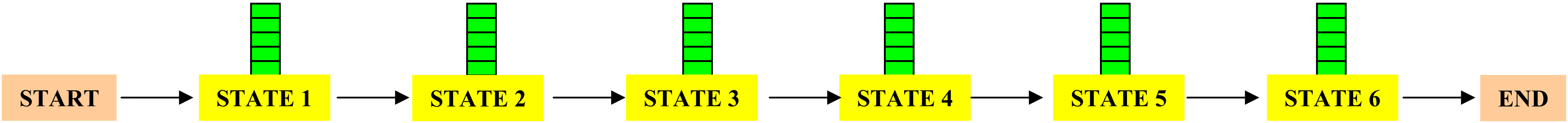
  - Profile Method [Gribskov et al., 1996]
    - Build a profile from the alignment and score query sequence against the profile to decide if it "fits the profile".
    - Need to pick a threshold score.
  - Enumerative/Combinatorial Methods

Profile HMMs

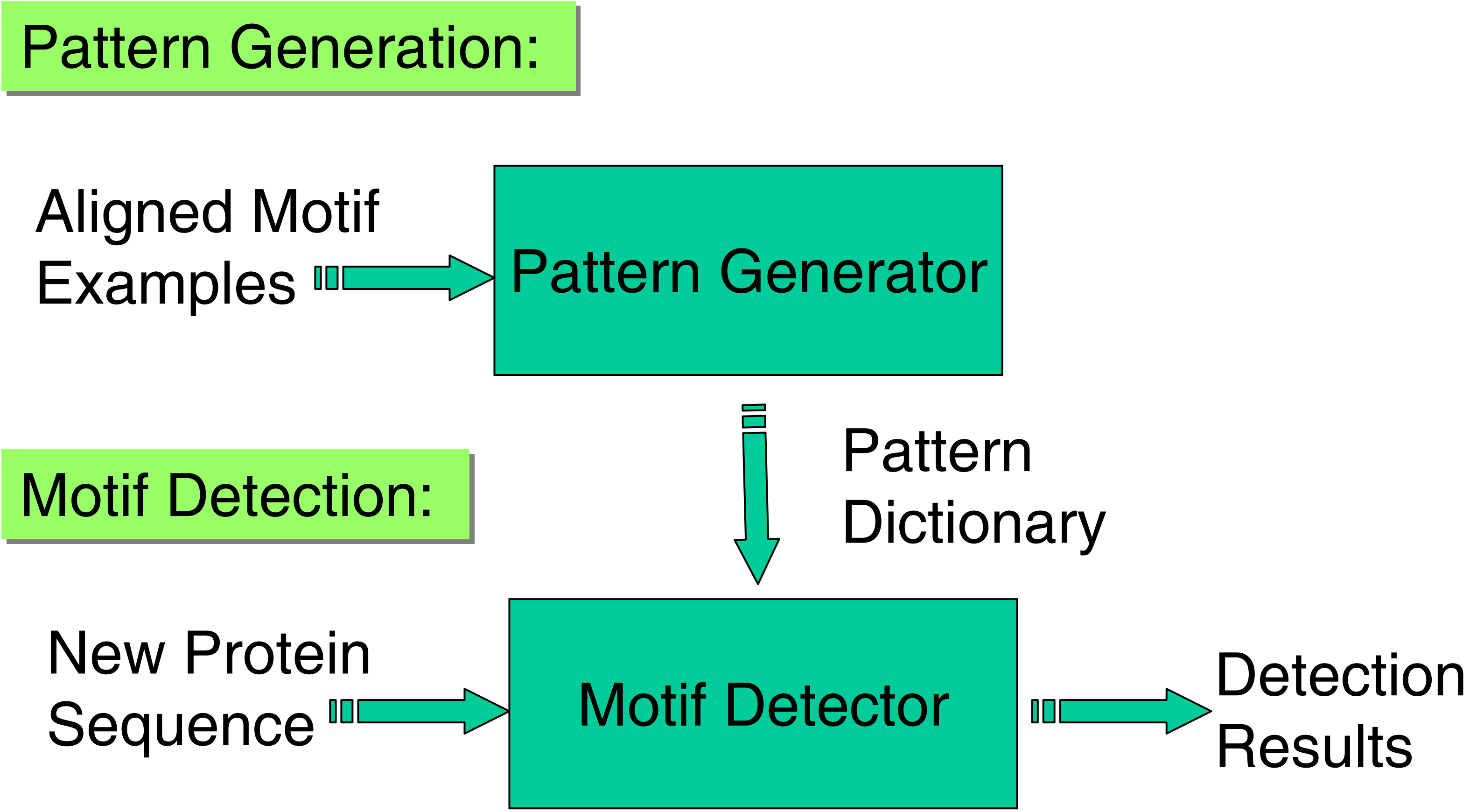
PROFILE METHOD, [M. Gribskov et al., '90]



Location in Seq.	Sequence						Protein Name
	1	2	3	4	5	6	
14	G	V	S	A	S	A	Ka RbtR
32	G	V	S	E	M	T	Ec DeoR
33	G	V	S	P	G	T	Ec RpoD
76	G	A	G	I	A	T	Ec TrpR
178	G	C	S	R	E	T	Ec CAP
205	C	L	S	P	S	R	Ec AraC
210	C	L	S	P	S	R	St AraC
13	G	V	N	K	E	T	Br MerR



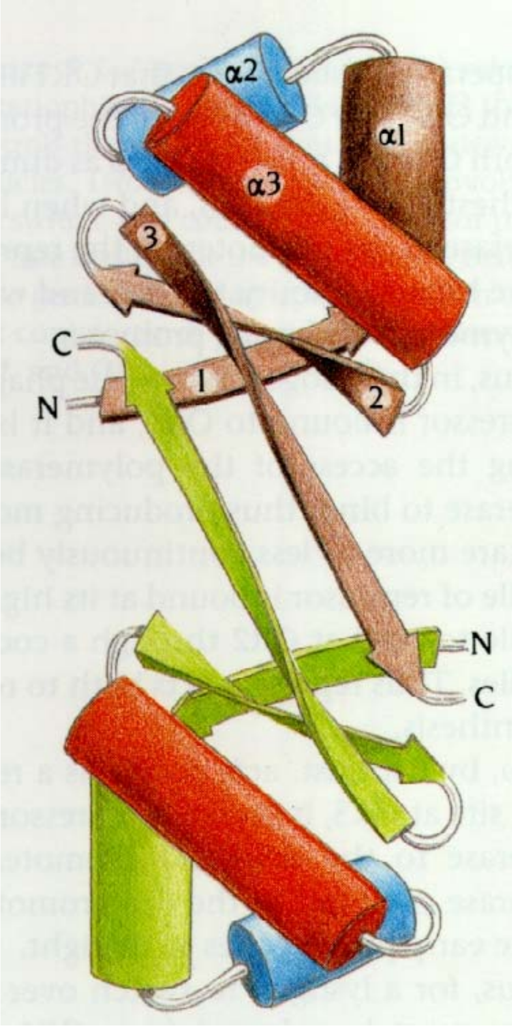
Combinatorial Method: GYM



[Narasimhan, Bu, Wang, Xu, Yang, Mathee, J Comput Biol, 2002]

Helix-Turn-Helix Motifs

- Structure
  - 3-helix complex
  - Length: 22 amino acids
  - Turn angle
- Function
  - Gene regulation by binding to DNA



Branden & Tooze

DNA Binding at HTH Motif

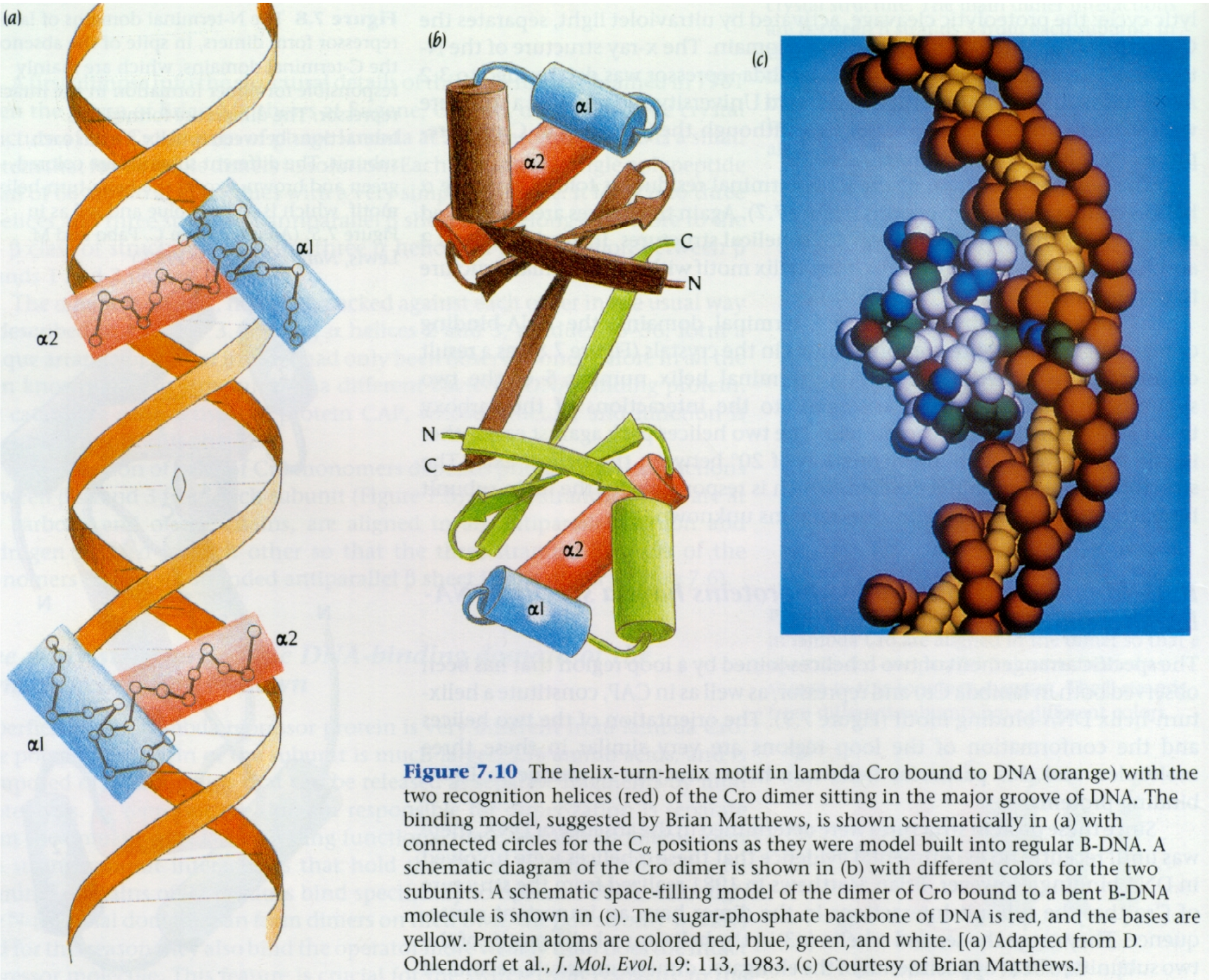


Figure 7.10 The helix-turn-helix motif in lambda Cro bound to DNA (orange) with the two recognition helices (red) of the Cro dimer sitting in the major groove of DNA. The binding model, suggested by Brian Matthews, is shown schematically in (a) with connected circles for the C_α positions as they were model built into regular B-DNA. A schematic diagram of the Cro dimer is shown in (b) with different colors for the two subunits. A schematic space-filling model of the dimer of Cro bound to a bent B-DNA molecule is shown in (c). The sugar-phosphate backbone of DNA is red, and the bases are yellow. Protein atoms are colored red, blue, green, and white. [(a) Adapted from D. Ohlendorf et al., *J. Mol. Evol.* 19: 113, 1983. (c) Courtesy of Brian Matthews.]

Branden & Tooze

HTH Motifs: Examples

Loc	*Protein*	*Helix 2*									*Turn*				*Helix 3*								
	Name	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	**Cro**	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	**434 Cro**	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	**P22 Cro**	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	**Rep**	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
16	**434 Rep**	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
19	**P22 Rep**	I	R	Q	A	A	L	G	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
24	**CII**	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
4	**LacR**	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
167	**CAP**	I	T	R	Q	E	I	G	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
66	**TrpR**	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
22	**BlaA Pv**	L	N	F	T	K	A	A	L	E	L	Y	V	T	Q	G	A	V	S	Q	Q	V	R
23	**TrpI Ps**	N	S	V	S	Q	A	A	E	Q	L	H	V	T	H	G	A	V	S	R	Q	L	K

Combinatorial Method: GYM

- ❑ Combinations of residues in specific locations (may not be contiguous) contribute towards stabilizing a structure.
- ❑ Some reinforcing combinations are relatively rare.
- ❑ GYM algorithm is inspired by the APriori algorithm [Agrawal et al., 1996]

[Narasimhan, Bu, Wang, Xu, Yang, Mathee, J Comput Biol, 2002]

Patterns

Loc	*Protein Name*	*Helix 2*									*Turn*				*Helix 3*								
		-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	**Cro**	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	**434 Cro**	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	**P22 Cro**	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	**Rep**	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
16	**434 Rep**	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
19	**P22 Rep**	I	R	Q	A	A	L	G	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
24	**CII**	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
4	**LacR**	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
167	**CAP**	I	T	R	Q	E	I	G	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
66	**TrpR**	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
22	**BlaA Pv**	L	N	F	T	K	A	A	L	E	L	Y	V	T	Q	G	A	V	S	Q	Q	V	R
23	**TrpI Ps**	N	S	V	S	Q	A	A	E	Q	L	H	V	T	H	G	A	V	S	R	Q	L	K

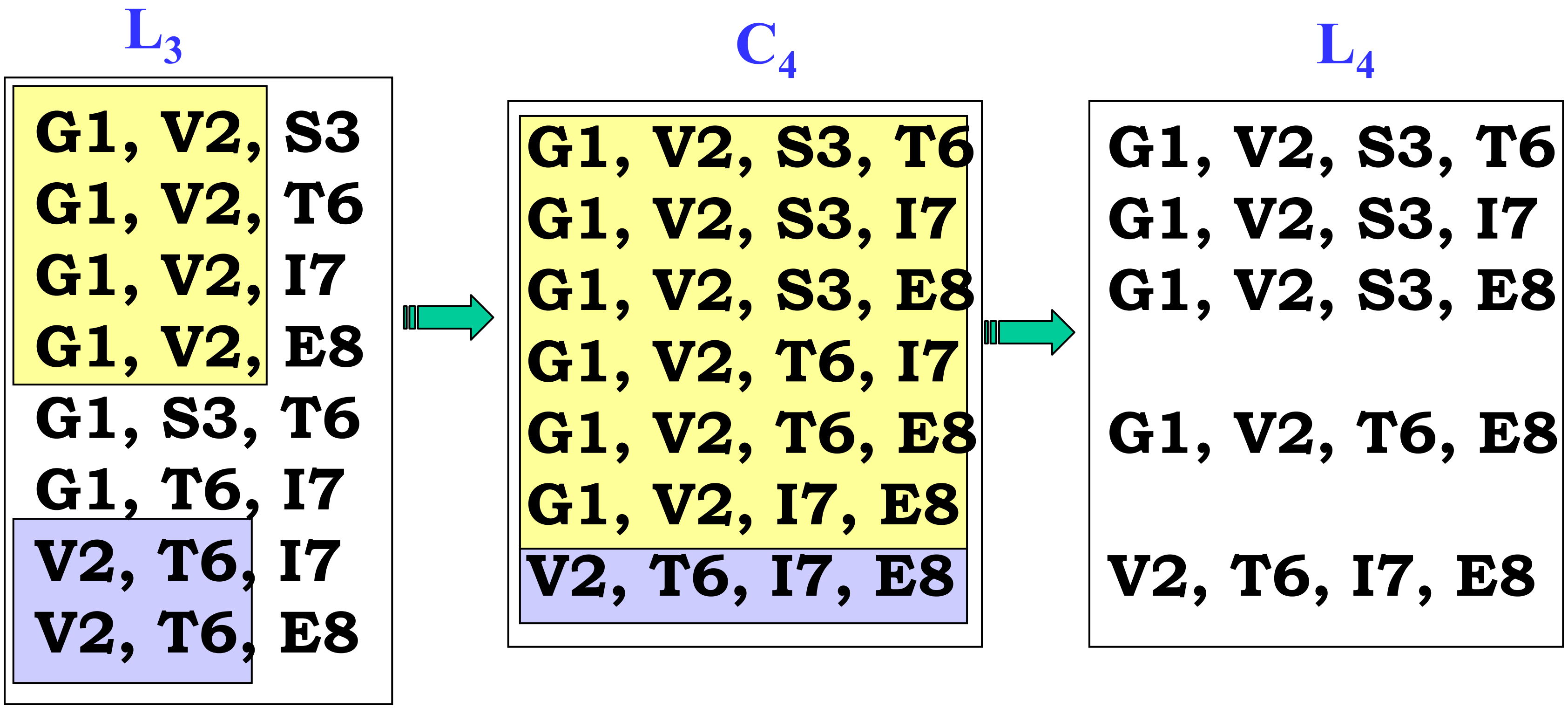
- **Q1 G9 N20**
- **A5 G9 V10 I15**

Pattern Mining Algorithm

Algorithm Pattern-Mining
Input: Motif length m, support threshold T,
list of aligned motifs M.
Output: Dictionary L of frequent patterns.

```
$L_1$ := All frequent patterns of length 1
for $i = 2$ to $m$ do
    $C_i$ := Candidates($L_{i-1}$)
    $L_i$ := Frequent candidates from $C_i$
    if ($|L_i| <= 1$) then
        return $L$ as the union of all $L_j$ , $j <= i$.
```

Candidates Function



L_3

G1, V2, S3
G1, V2, T6
G1, V2, I7
G1, V2, E8
G1, S3, T6
G1, T6, I7
V2, T6, I7
V2, T6, E8

C_4

G1, V2, S3, T6
G1, V2, S3, I7
G1, V2, S3, E8
G1, V2, T6, I7
G1, V2, T6, E8
G1, V2, I7, E8
V2, T6, I7, E8

L_4

G1, V2, S3, T6
G1, V2, S3, I7
G1, V2, S3, E8
G1, V2, T6, E8
V2, T6, I7, E8

Motif Detection Algorithm

Algorithm Motif-Detection

Input : Motif length m,
threshold score T,
pattern dictionary L,
and input protein sequence P[1..n].

Output : Detected motif(s).

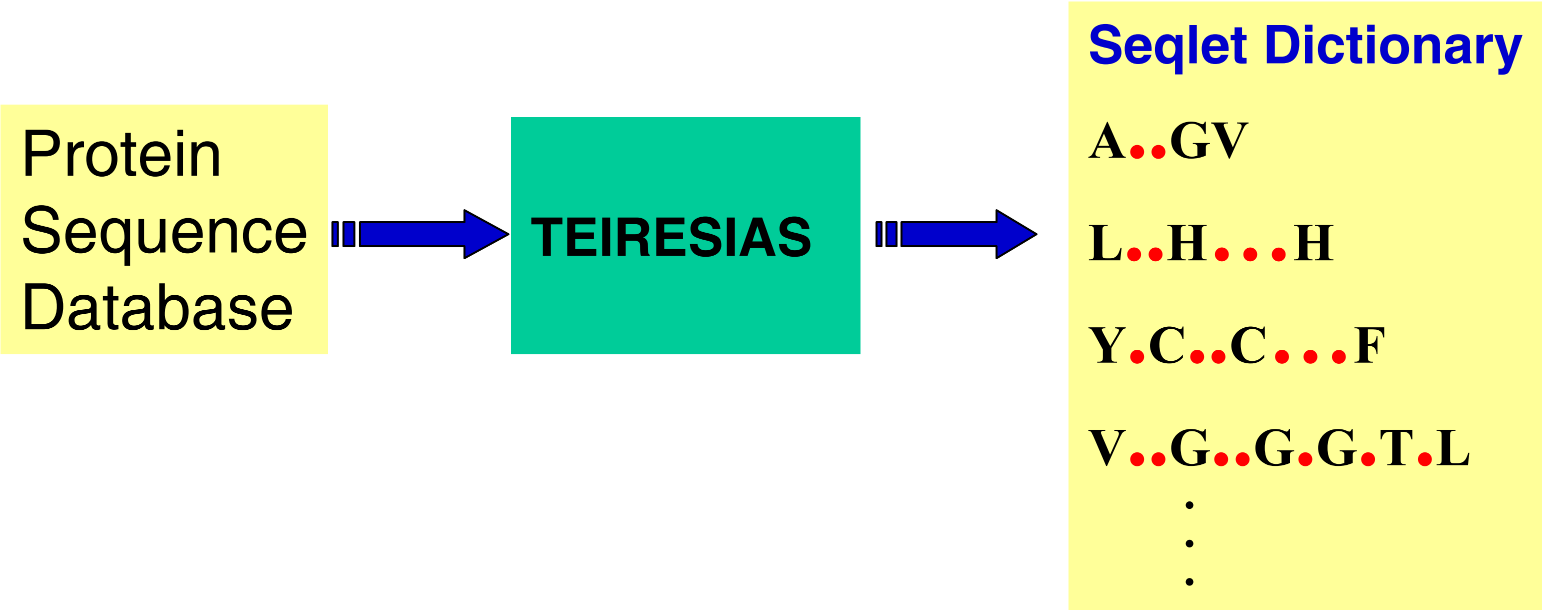
```
for each location i do
    S := MatchScore(P[i..i+m-1], L).
    if  (S > T) then
        Report it as a possible motif
```

Experimental Results: **GYM 2.0**

Motif	*Protein Family*	*Number Tested*	*GYM = DE Agree*	*Number Annotated*	*GYM = Annot.*
HTH Motif (22)	Master	88	88 (100 %)	13	13
	Sigma	314	284 + 23 (98 %)	96	82
	Negates	93	86 (92 %)	0	0
	LysR	130	127 (98 %)	95	93
	AraC	68	57 (84 %)	41	34
	Rreg	116	99 (85 %)	57	46
	Total	675	653 + 23 (94 %)	289	255 (88 %)

Unaligned Pattern Discovery

TEIRESIAS:
The algorithm is similar to that used in GYM for aligned Pattern discovery.



Protein Sequence Database

TEIRESIAS

Seqlet Dictionary

A..GV

L..H...H

Y.C..C...F

V..G..G.G.T.L

.
.
.

Rigoutsos & Floratos, Bioinformatics, '98

TEIRESIAS: Key Features

- Starts with a set of **seed** patterns (Enumeration step)
- **Convolution** operator applied to all pairs of patterns:

  A..GV.S ⊕ V.S.GR = A..GV.S.GR
- **Order of Evaluation** carefully chosen so that long patterns get longer first
- Finds **all maximal patterns.**
- **Combinatorial explosion** avoided by generating only relevant maximal patterns.

Rigoutsos & Floratos, Bioinformatics, '98

SPLASH

- ❑ Structural Pattern Localization Analysis by Sequential Histogram (SPLASH)
- ❑ Not limited to fixed alphabet size
- ❑ Patterns are modeled by a homology metric and thus allow mismatches
- ❑ Early pruning of inconsistent seed patterns, leading to increased efficiency.
- ❑ Easily parallelized with availability of extra resources.

Califano, Bioinformatics, '00; Califano et al., J Comput Biol, '00

Precomputed Sequence Patterns

- PROSITE
- BLOCKS and PRINTS
- eMOTIF
- SPAT
- PRODOM
- Pfam

Motif Detection Tools

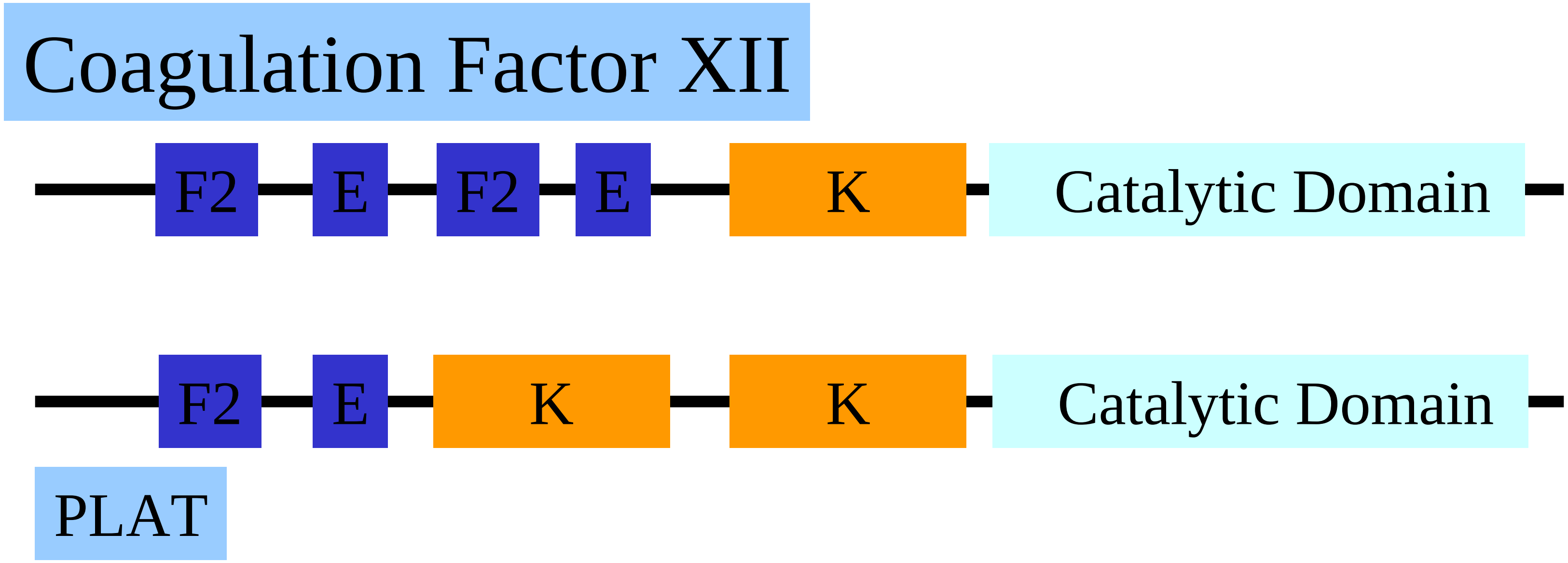
- PROSITE (Database of protein families & domains)
  - Try PDOC00040. Also Try PS00041
- PRINTS Sample Output
- BLOCKS (multiply aligned ungapped segments for highly conserved regions of proteins; automatically created) Sample Output
- Pfam (Protein families database of alignments & HMMs)
  - Multiple Alignment, domain architectures, species distribution, links: Try
- MoST
- PROBE
- ProDom
- DIP

Protein Information Sites

- SwissPROT & GenBank
- InterPRO is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences. See sample.
- PIR Sample Protein page

Modular Nature of Proteins

- Proteins are collections of “modular” domains. For example,



Domain Architecture Tools

- CDART
  - Protein AAH24495 ; Domain Architecture;
  - It's domain relatives;
  - Multiple alignment for 2nd domain
- SMART

Predicting Specialized Structures

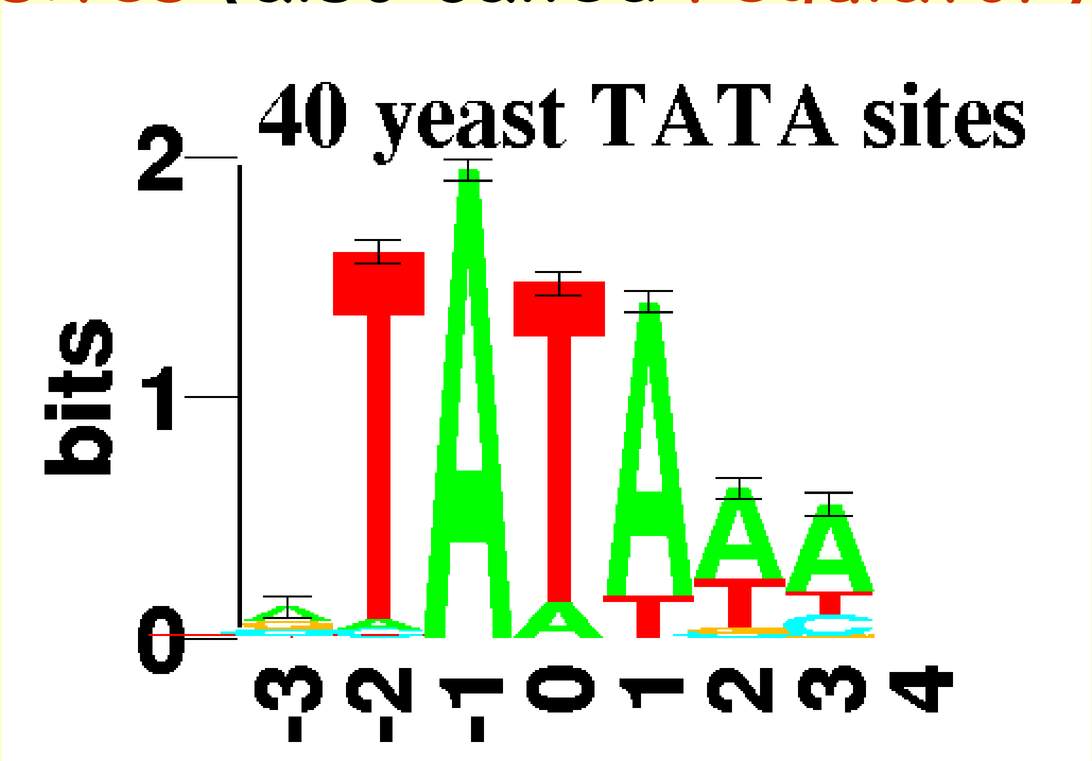
- ❑COILS – Predicts coiled coil motifs
- ❑TMPred – predicts transmembrane regions
- ❑SignalP – predicts signal peptides
- ❑SEG – predicts nonglobular regions

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

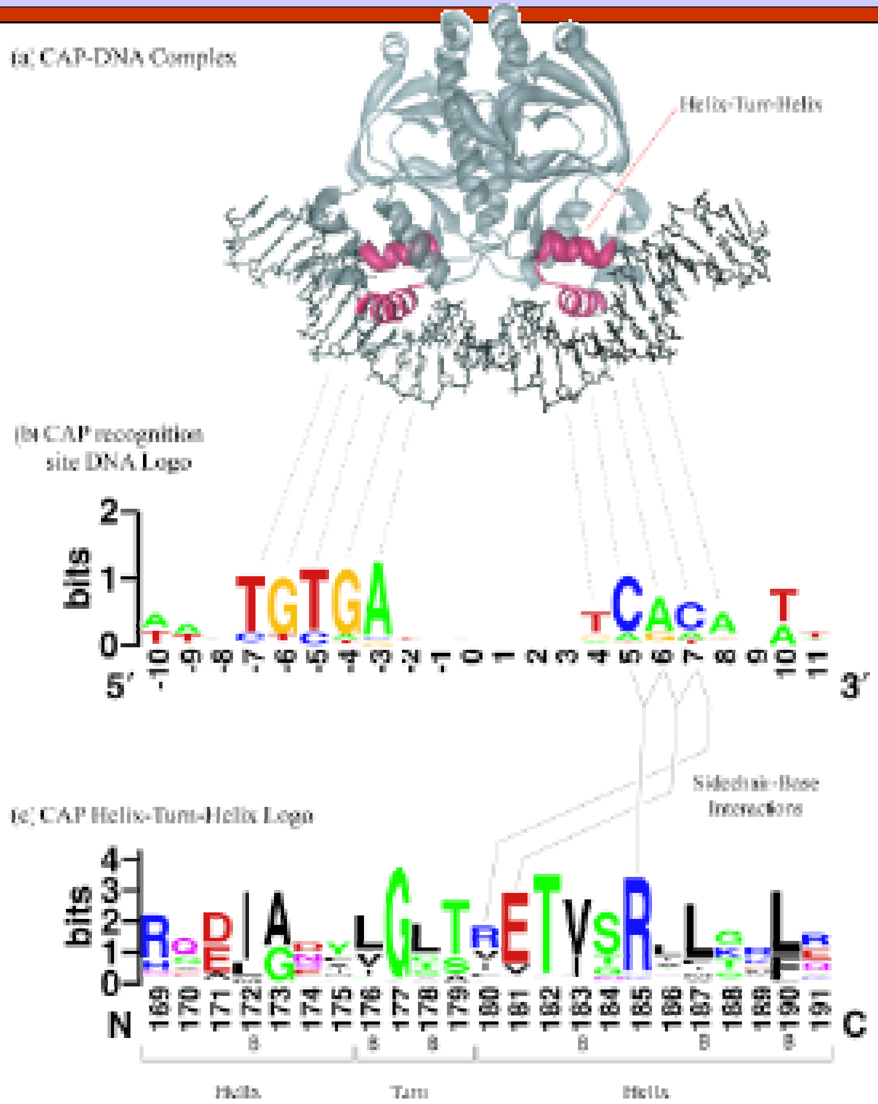
Motifs in DNA Sequences

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



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

Motifs



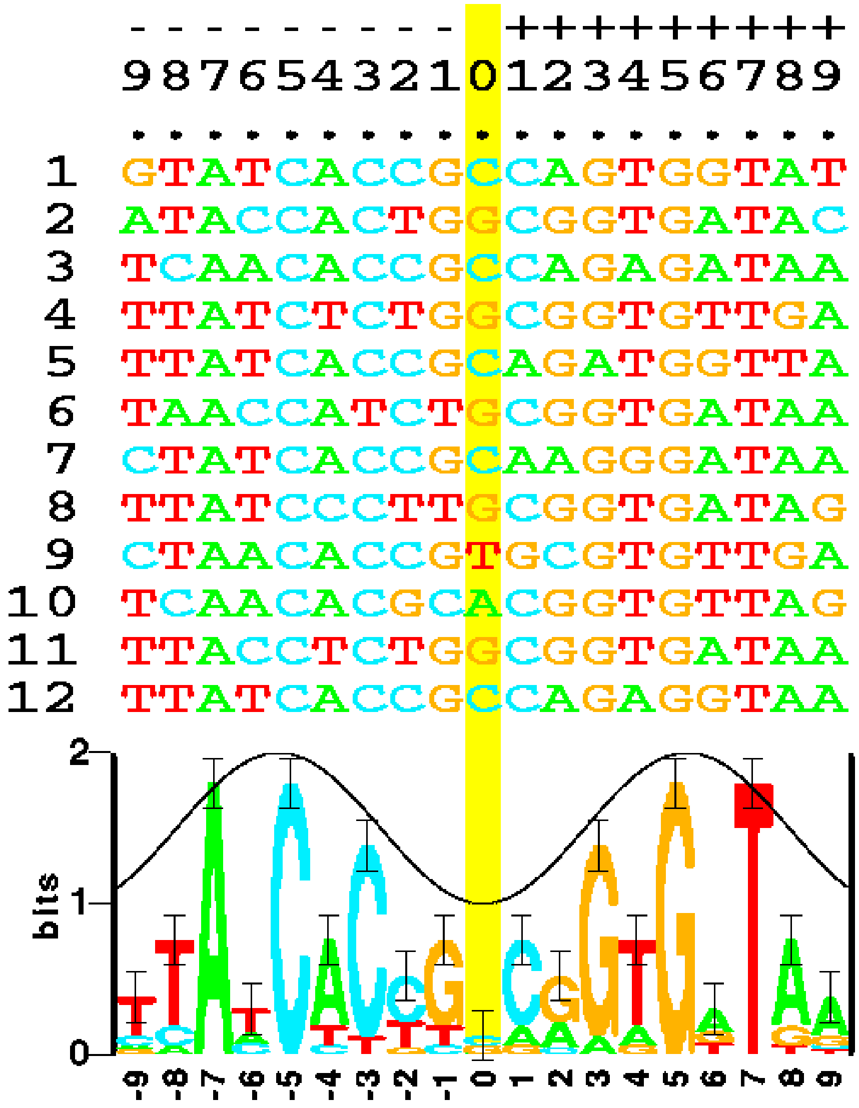
(a) CAP-DNA Complex

(b) CAP recognition site DNA Logo

(c) CAP Helix-Turn-Helix Logo

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

Motifs in DNA Sequences



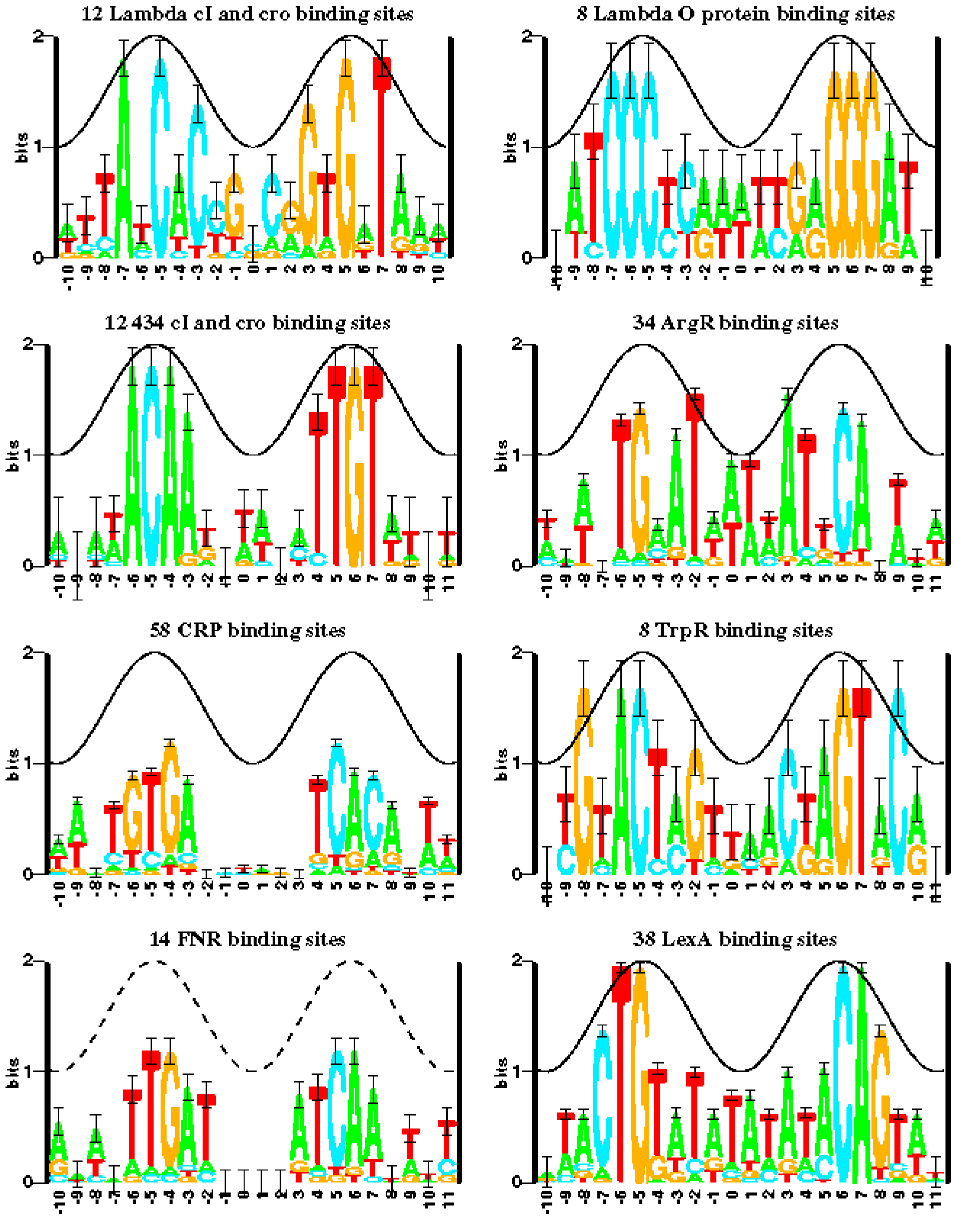
```
    ---------   +++++++++
    9876543210123456789
    ...................
 1  GTATCACCGCCAGTGGTAT
 2  ATACCACTGGCGGTGATAC
 3  TCAACACCGCCAGAGATAA
 4  TTATCTCTGGCGGTGTTGA
 5  TTATCACCGCAGATGGTTA
 6  TAACCATCTGCGGTGATAA
 7  CTATCACCGCAAGGGATAA
 8  TTATCCCTTGCGGTGATAG
 9  CTAACACCGTGCGTGTTGA
10  TCAACACGCACGGTGTTAG
11  TTACCTCTGGCGGTGATAA
12  TTATCACCGCCAGAGGTAA
```

12 Lambda cI and cro binding sites

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



12 Lambda cI and cro binding sites

8 Lambda O protein binding sites

12 434 cI and cro binding sites

34 ArgR binding sites

58 CRP binding sites

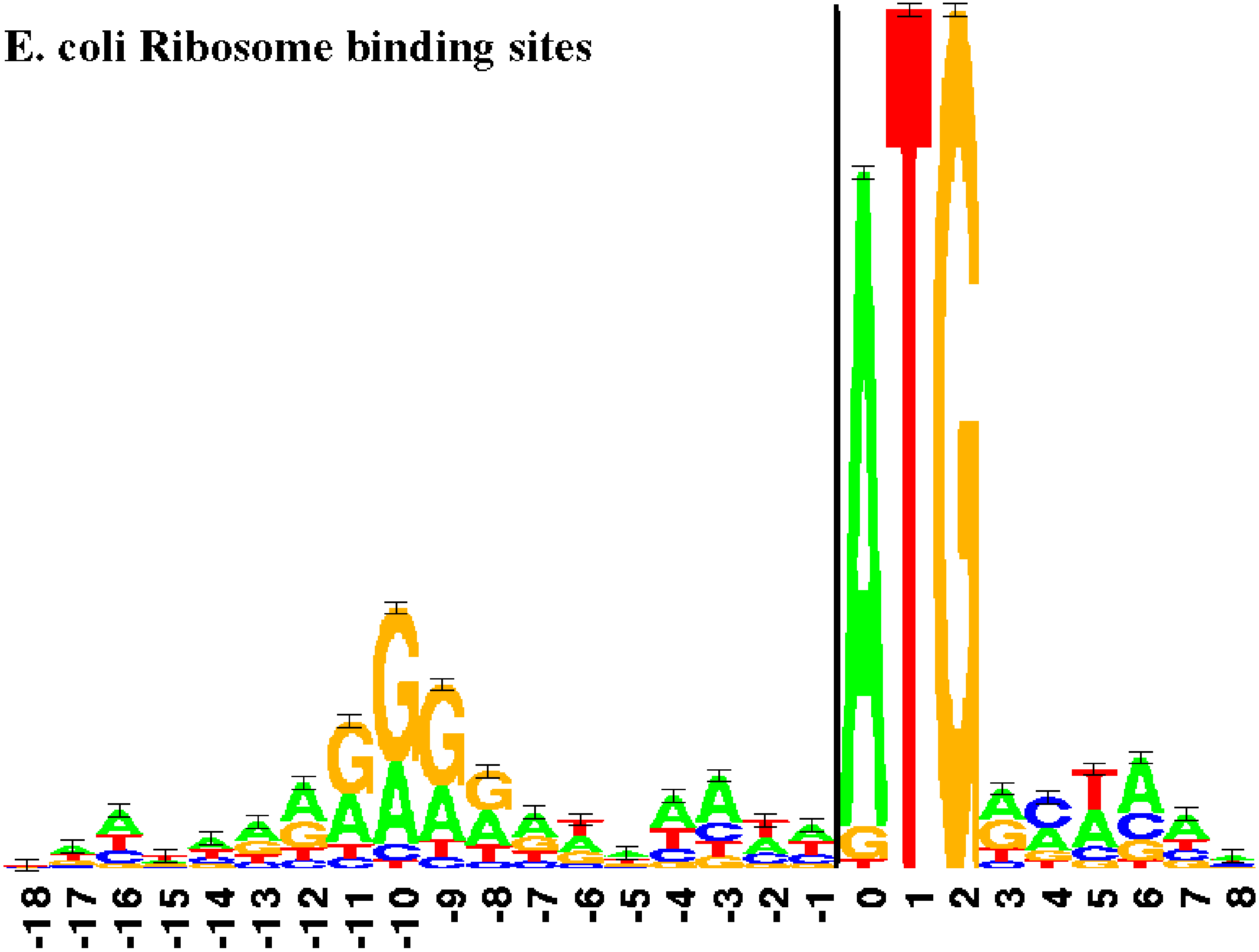
8 TrpR binding sites

14 FNR binding sites

38 LexA binding sites

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

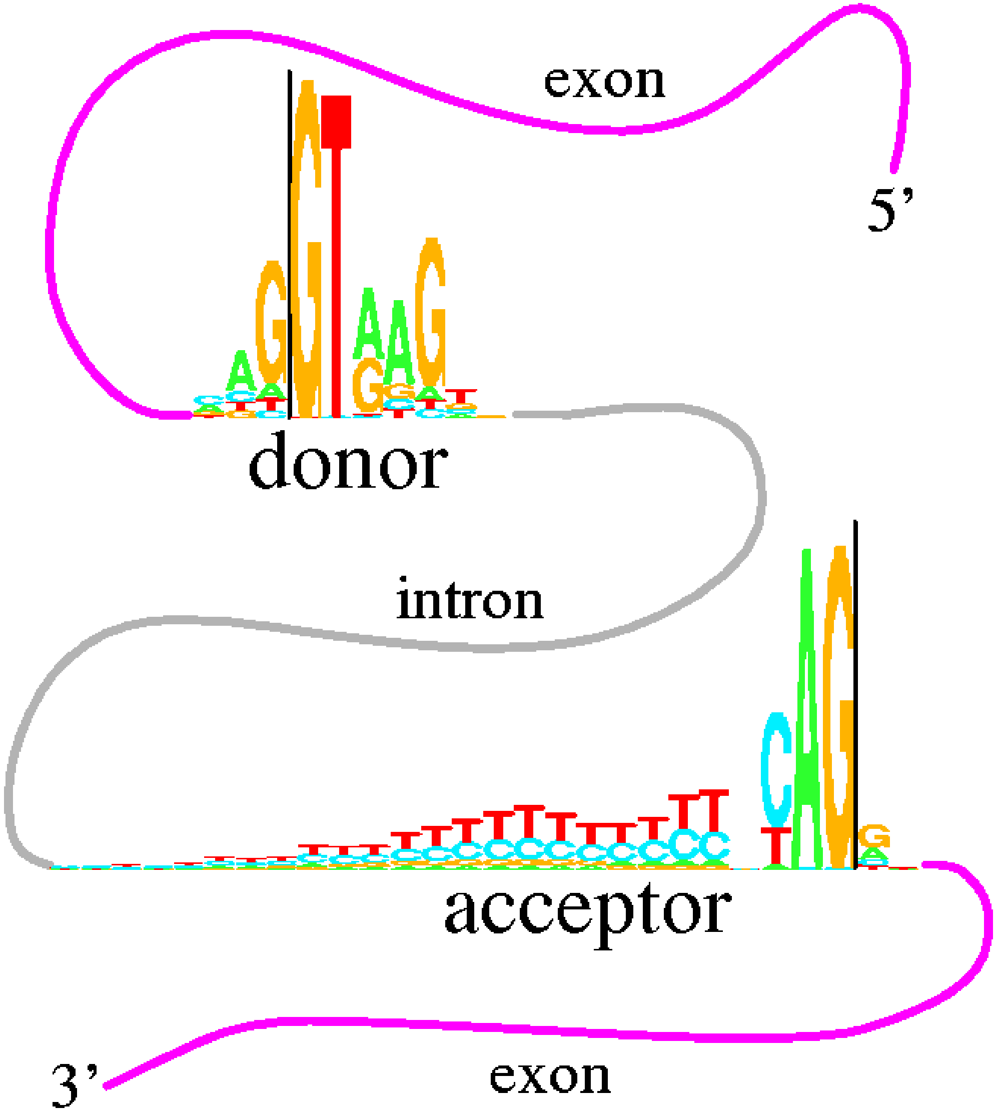
E. coli Ribosome binding sites



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 in DNA Sequences

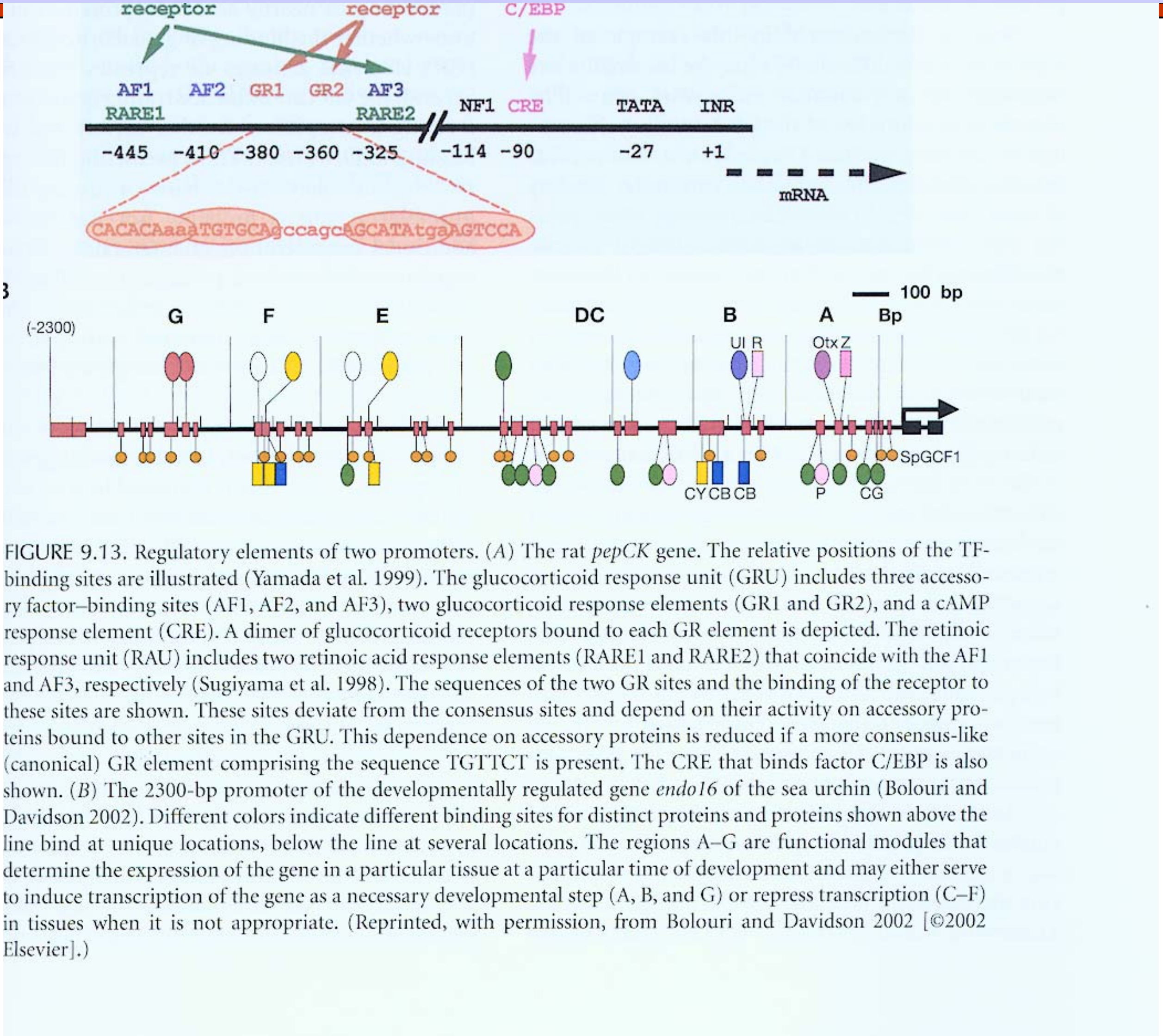


FIGURE 9.13. Regulatory elements of two promoters. (*A*) The rat *pepCK* gene. The relative positions of the TF-binding sites are illustrated (Yamada et al. 1999). The glucocorticoid response unit (GRU) includes three accessory factor–binding sites (AF1, AF2, and AF3), two glucocorticoid response elements (GR1 and GR2), and a cAMP response element (CRE). A dimer of glucocorticoid receptors bound to each GR element is depicted. The retinoic response unit (RAU) includes two retinoic acid response elements (RARE1 and RARE2) that coincide with the AF1 and AF3, respectively (Sugiyama et al. 1998). The sequences of the two GR sites and the binding of the receptor to these sites are shown. These sites deviate from the consensus sites and depend on their activity on accessory proteins bound to other sites in the GRU. This dependence on accessory proteins is reduced if a more consensus-like (canonical) GR element comprising the sequence TGTTCT is present. The CRE that binds factor C/EBP is also shown. (*B*) The 2300-bp promoter of the developmentally regulated gene *endo16* of the sea urchin (Bolouri and Davidson 2002). Different colors indicate different binding sites for distinct proteins and proteins shown above the line bind at unique locations, below the line at several locations. The regions A–G are functional modules that determine the expression of the gene in a particular tissue at a particular time of development and may either serve to induce transcription of the gene as a necessary developmental step (A, B, and G) or repress transcription (C–F) in tissues when it is not appropriate. (Reprinted, with permission, from Bolouri and Davidson 2002 [©2002 Elsevier].)

Motif Detection (TFBMs)

- See evaluation by Tompa et al.
  - [bio.cs.washington.edu/assessment]
- Gibbs Sampling Methods: AlignACE, GLAM, SeSiMCMC, MotifSampler
- Weight Matrix Methods: ANN-Spec, Consensus,
- EM: Improbizer, MEME
- Combinatorial & Misc.: MITRA, oligo/dyad, QuickScore, Weeder, YMF

Gibbs Sampling for Motif Detection