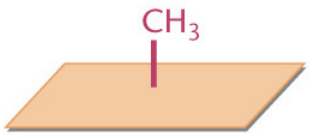
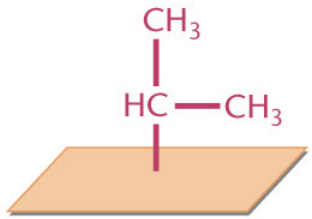
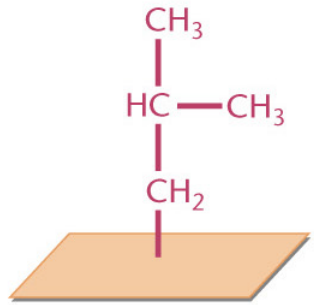
1. Nonpolar: Hydrophobic



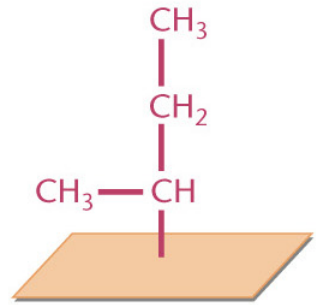
Alanine (ala–A)



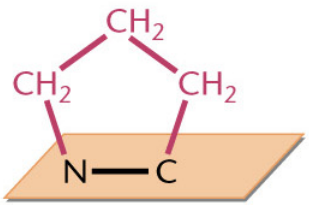
Valine (val–V)



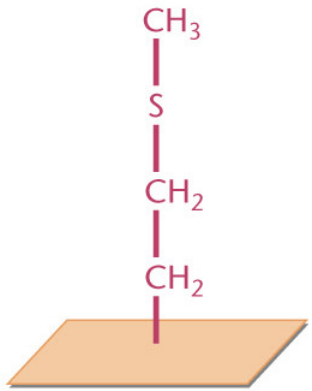
Leucine (leu–L)



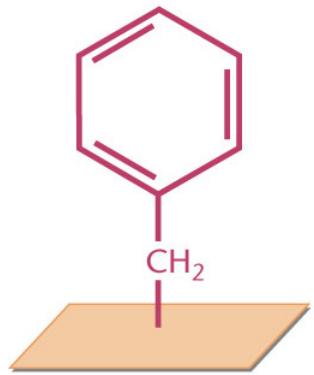
Isoleucine (ile–I)



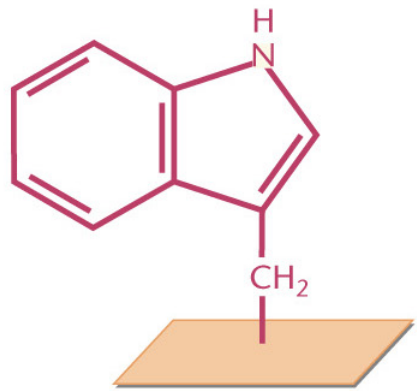
Proline (pro–P)



Methionine (met–M)



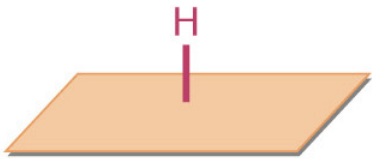
Phenylalanine (phe–F)



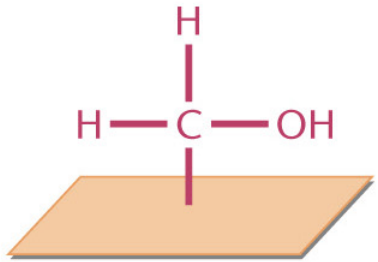
Tryptophan (trp–W)

Amino Acid Structures from Klug & Cummings

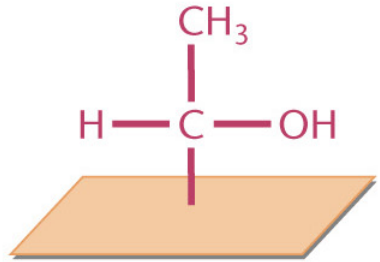
2. Polar: Hydrophilic



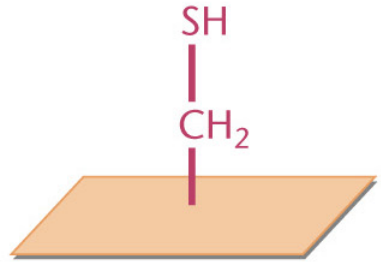
Glycine (gly–G)



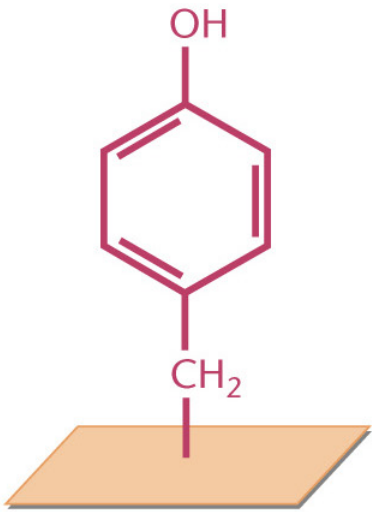
Serine (ser–S)



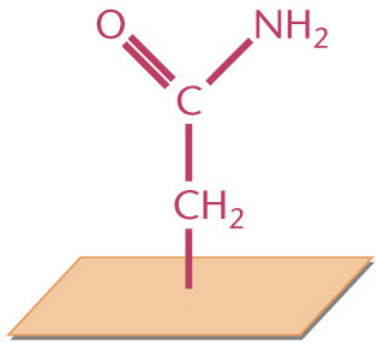
Threonine (thr–T)



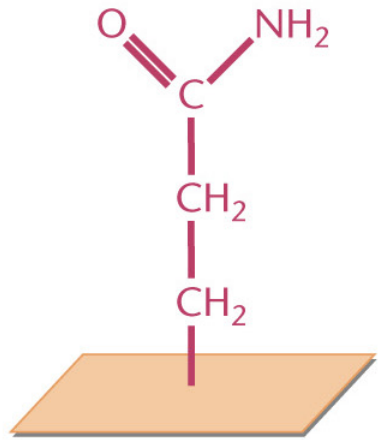
Cysteine (cys–C)



Tyrosine (tyr–Y)



Asparagine (asn–N)

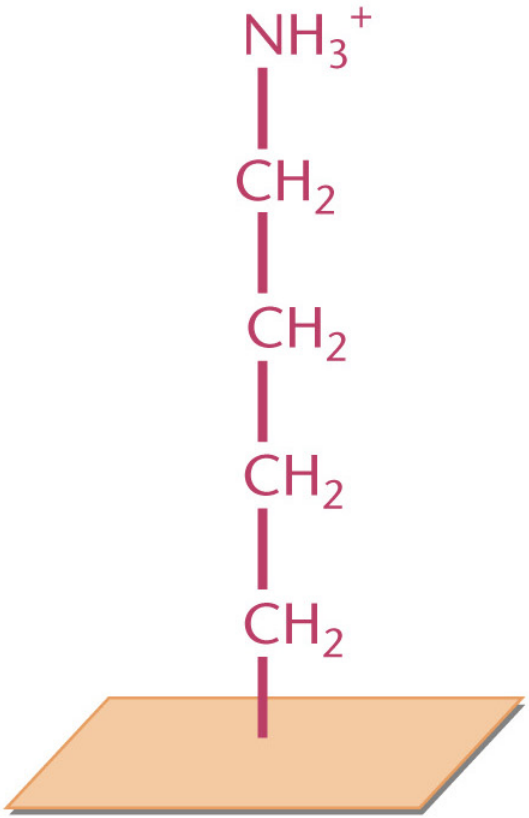


Glutamine (gln–Q)

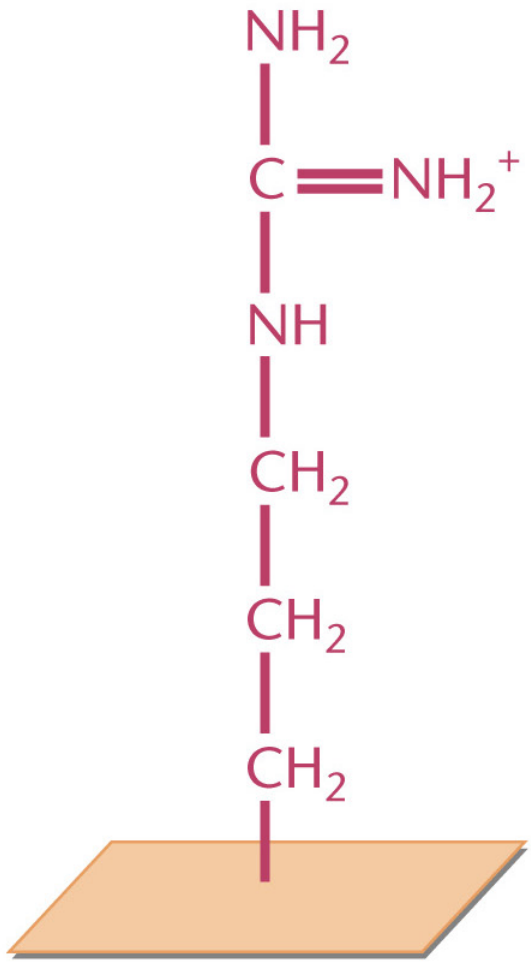
Amino Acid Structures from Klug & Cummings

3. Polar: positively charged (basic)

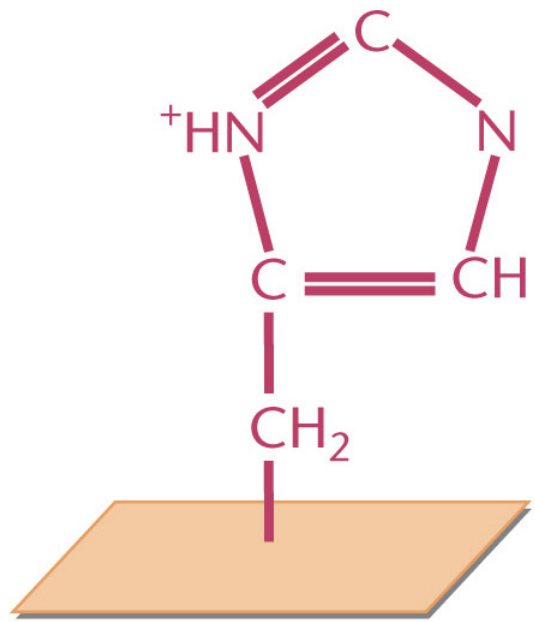
Amino Acid Structures from Klug & Cummings



Lysine (lys–K)

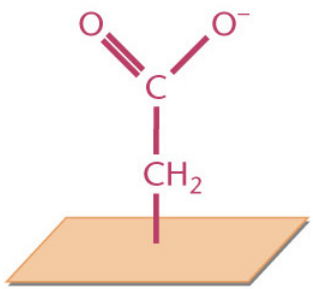


Arginine (arg–R)

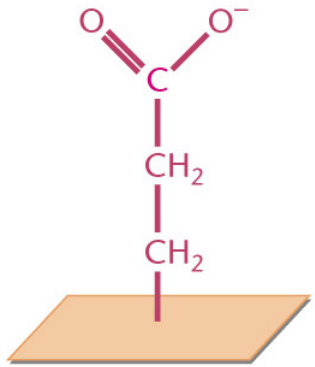


Histidine (his–H)

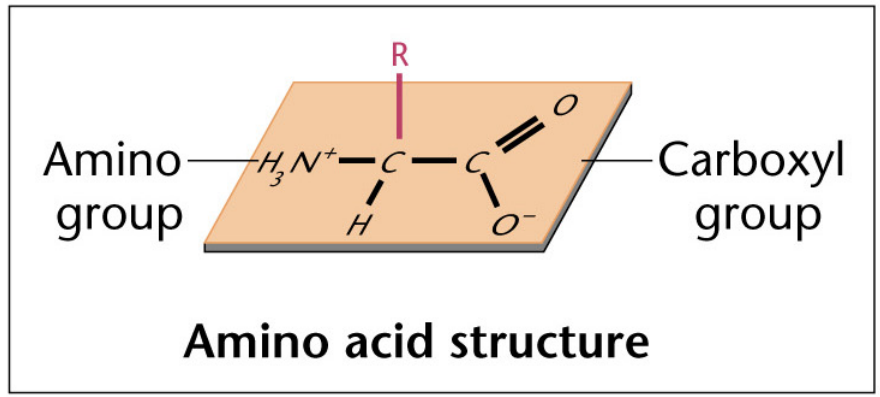
4. Polar: negatively charged (acidic)



Aspartic acid (asp–D)



Glutamic acid (glu–E)



Amino Acid Structures from Klug & Cummings

Motif Detection Tools

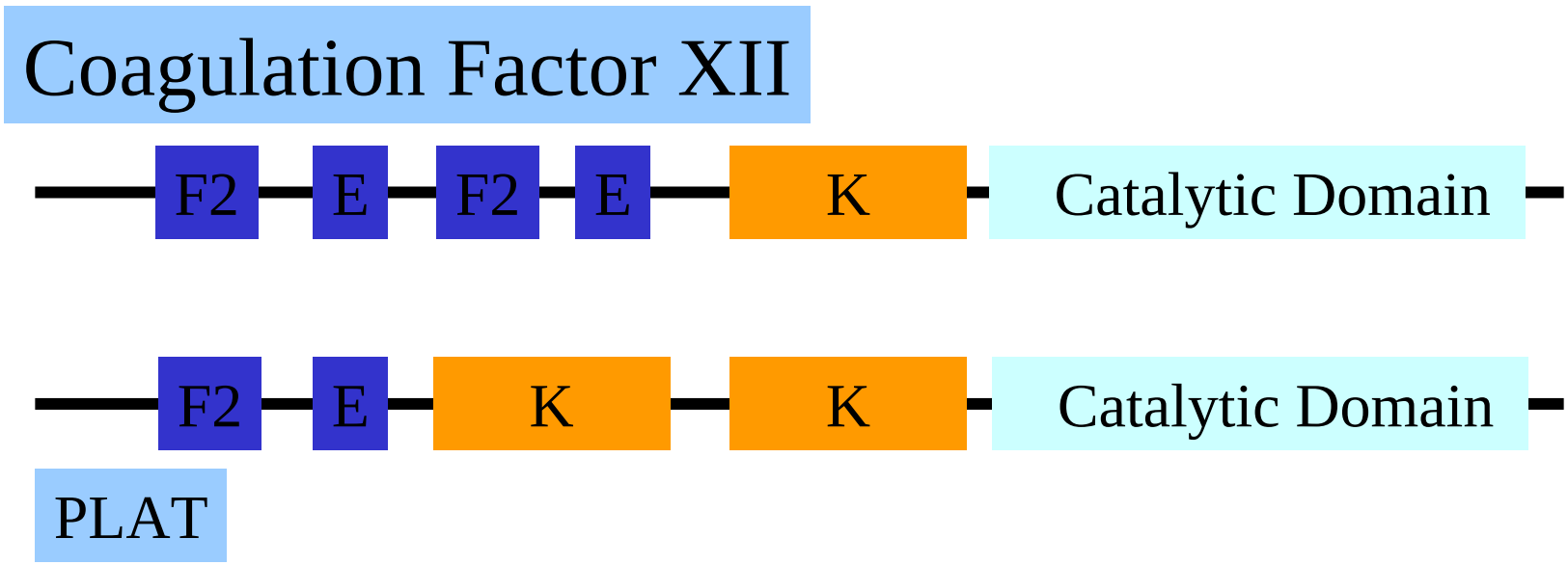
- PROSITE (Database of protein families & domains)
  - Try PDOC00040. Also Try PS00041
- PRINTS Sample Output
- BLOCKS (multiply aligned ungapped segments for highly conserved regions of proteins; automatically created) Sample Output
- Pfam (Protein families database of alignments & HMMs)
  - Multiple 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

Motifs 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 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

…YKCGLCERSFVEKSALSRHORVHKN…

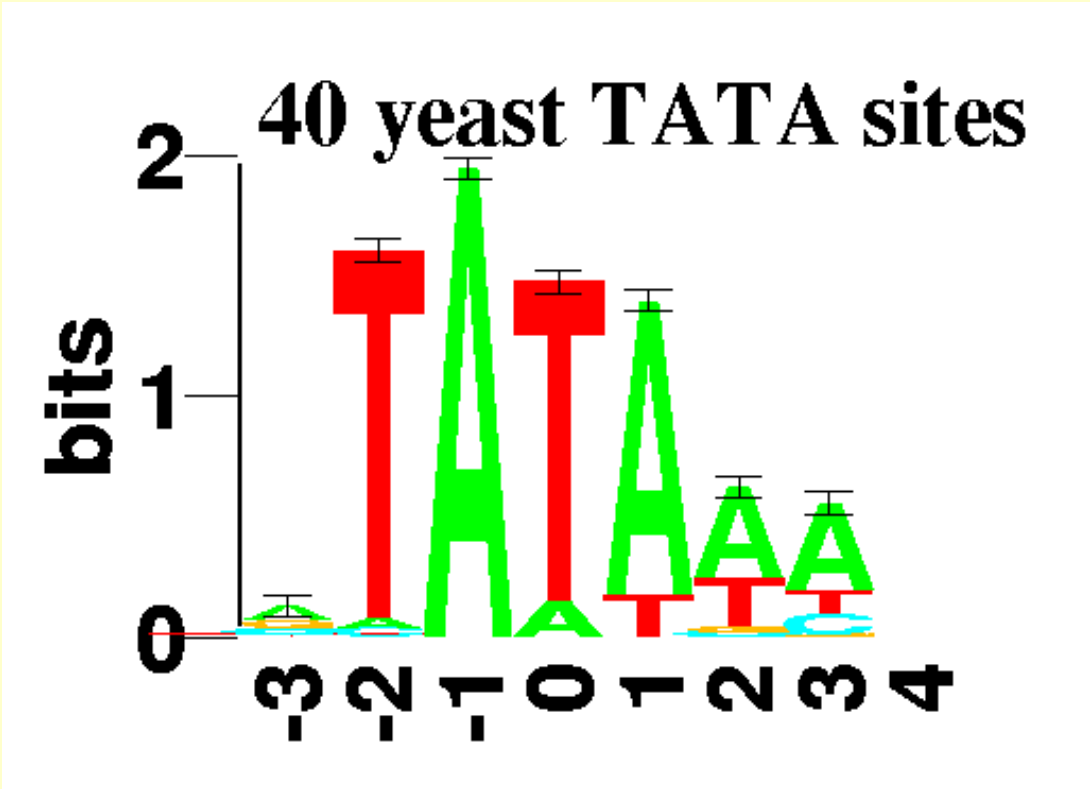
3 6 19 23

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

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

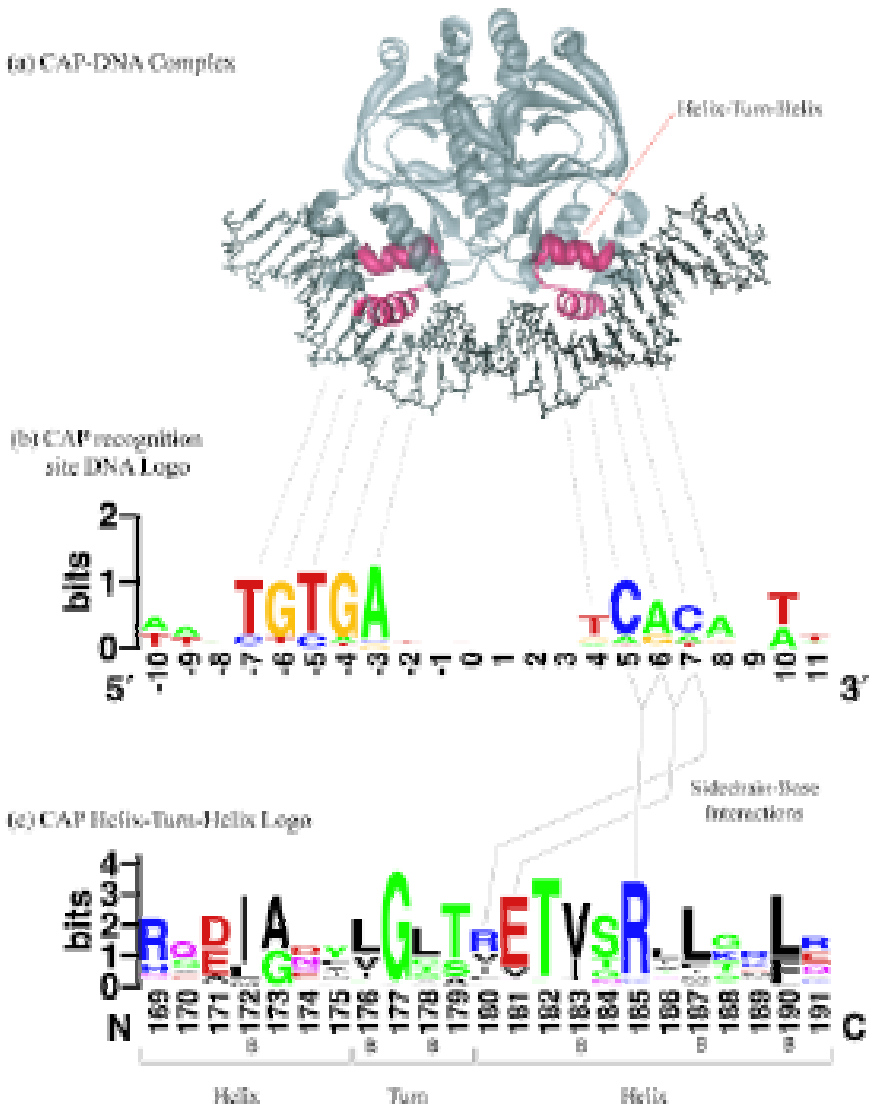
Motifs in DNA Sequences

- 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



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

Motifs in DNA Sequences

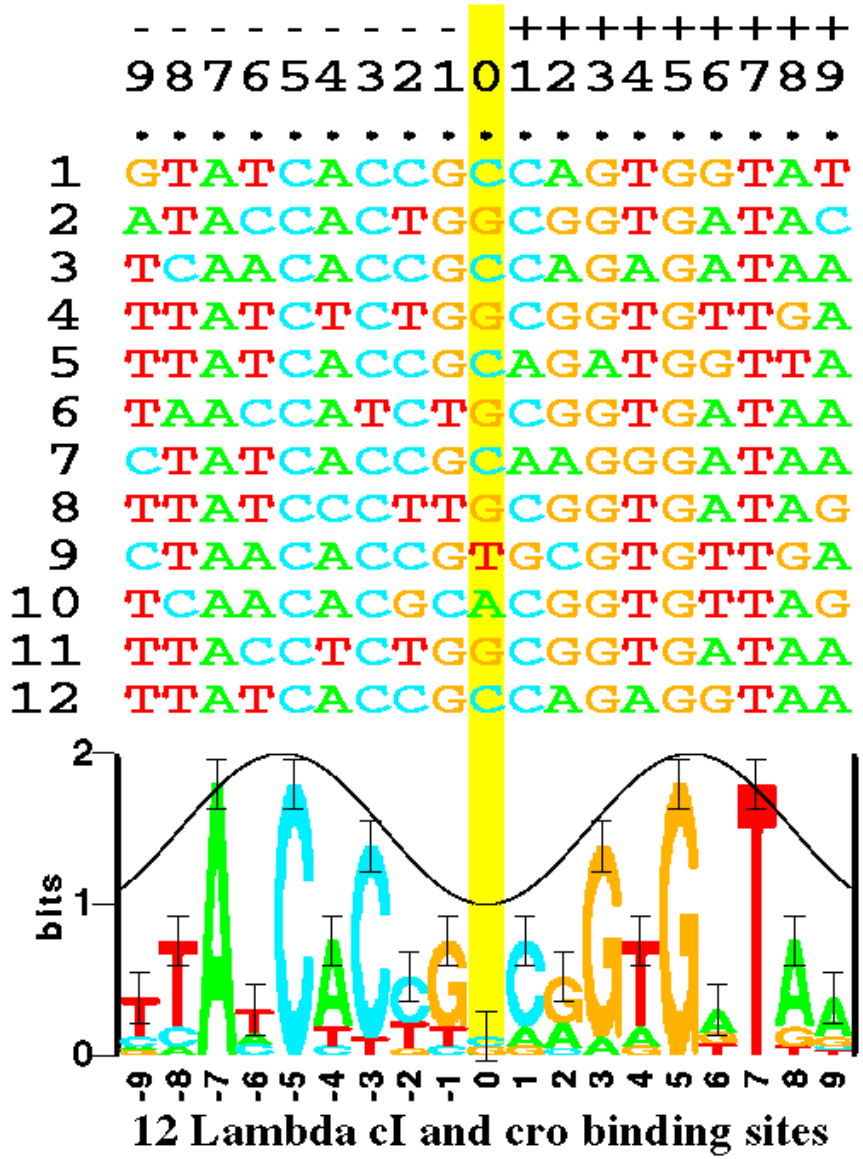
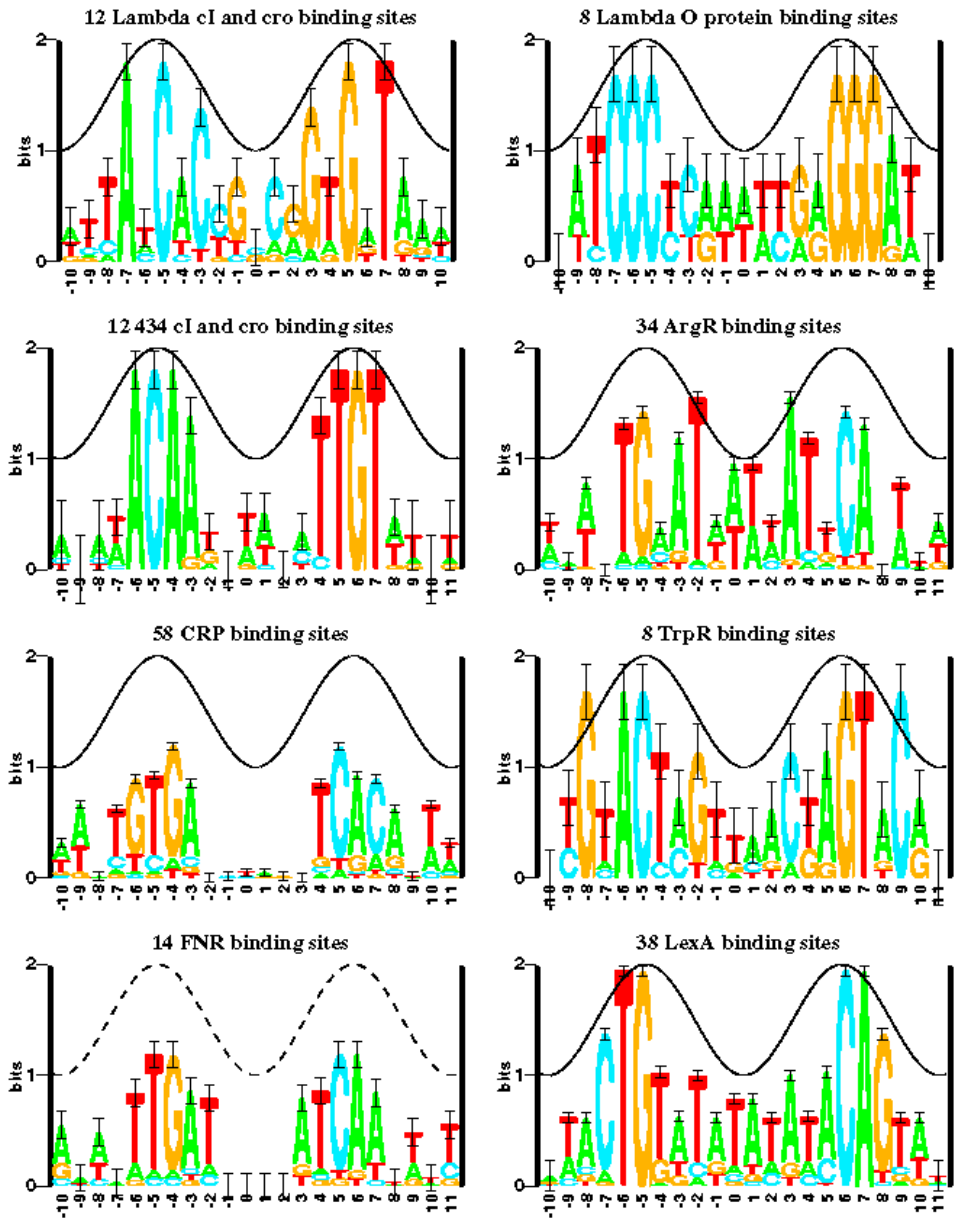


Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered 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].

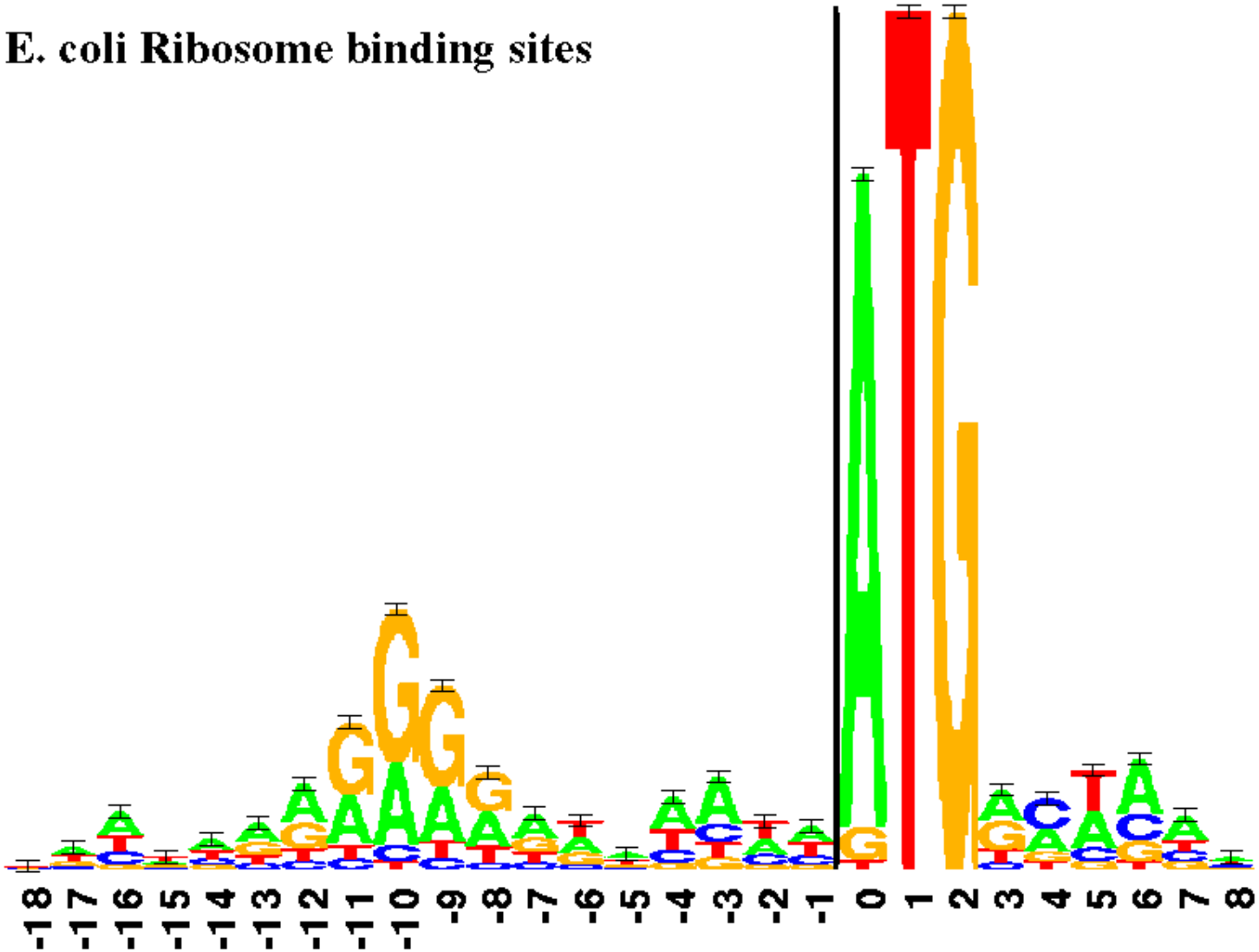
http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

More Motifs in *E. Coli* DNA Sequences



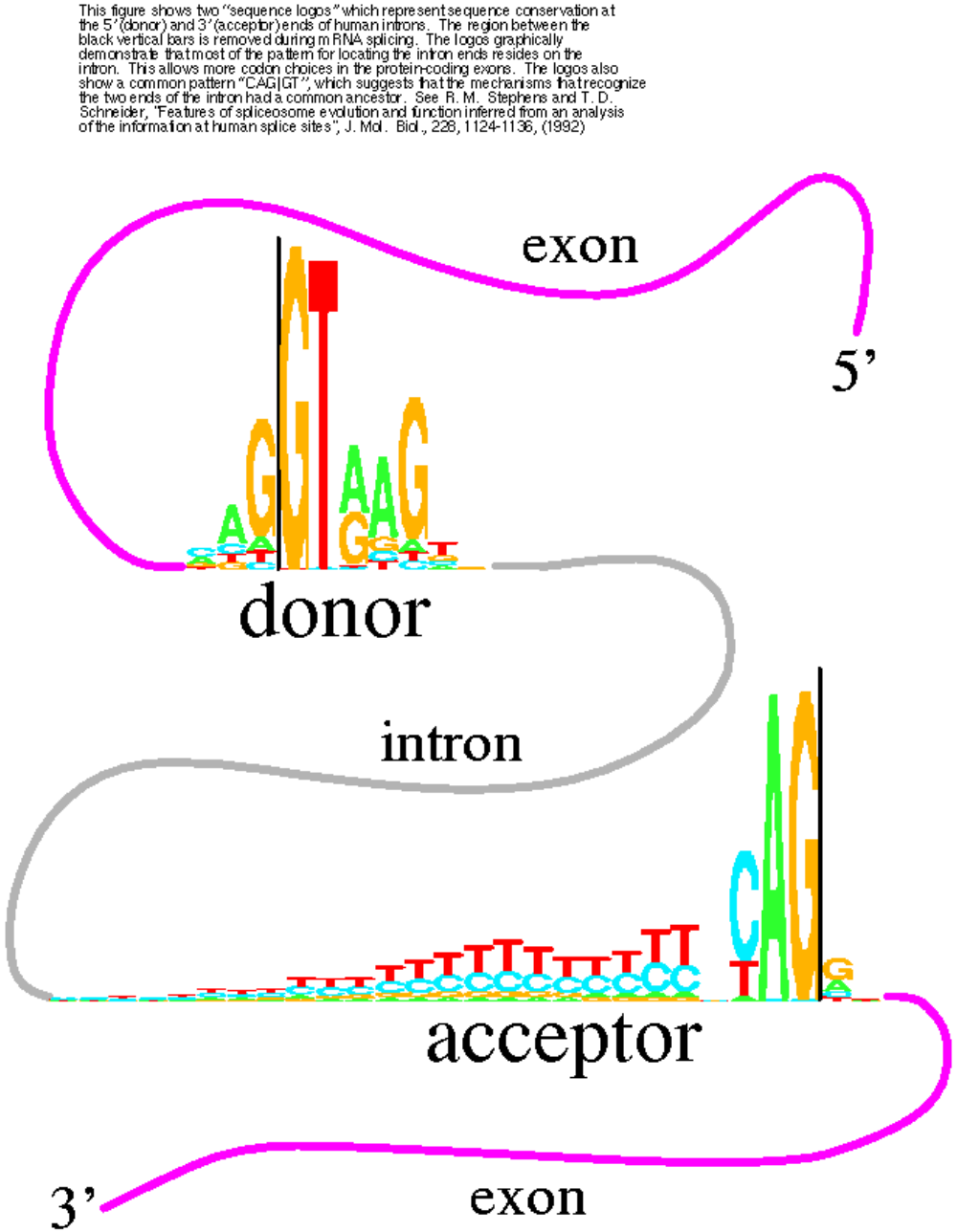
http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

E. coli Ribosome binding sites



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

Other Motifs in DNA Sequences: Human Splice Junctions



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

Motifs in DNA Sequences

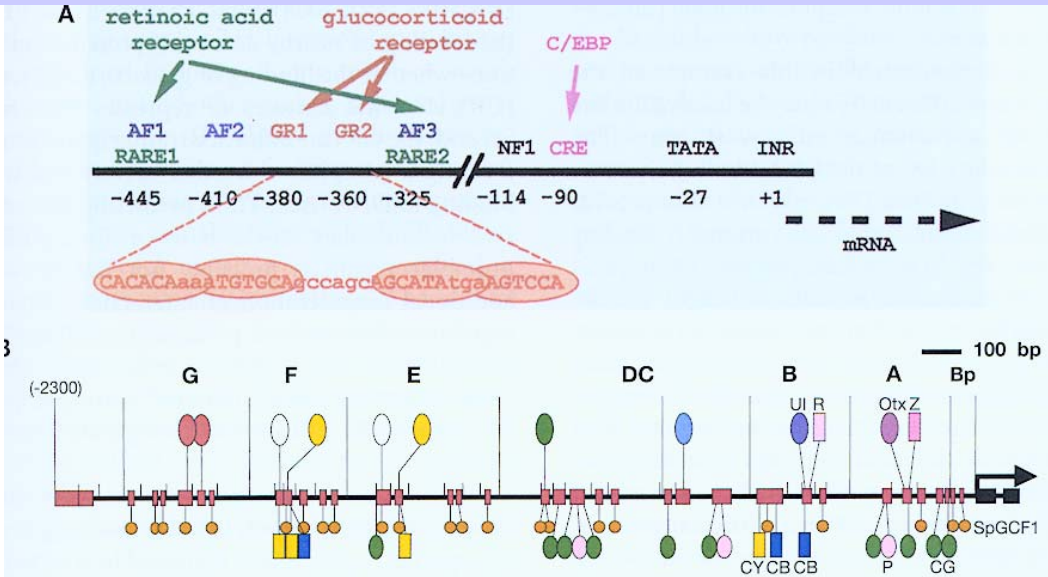


FIGURE 9.13. Regulatory elements of two promoters. (*A*) The rat *pepCK* gene. The relative positions of the TF-binding sites are illustrated (Yamada et al. 1999). The glucocorticoid response unit (GRU) includes three accessory factor–binding sites (AF1, AF2, and AF3), two glucocorticoid response elements (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

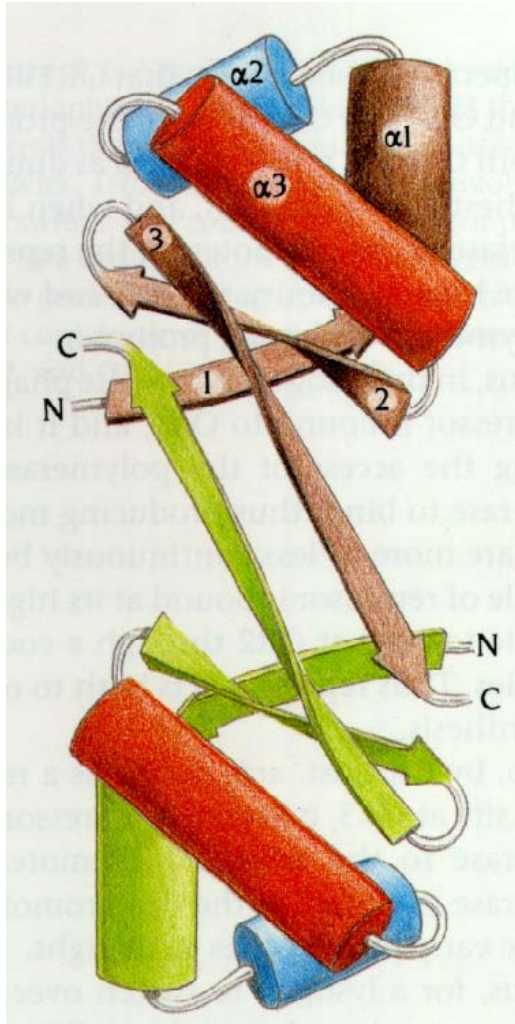
- Profile Method
  - If many examples of the motif are known, then
    - **Training**: build a **Profile** and compute a **threshold**
    - **Testing**: **score** against profile
- **Gibbs Sampling**
- **Expectation Method**
- **HMM**
- **Combinatorial Pattern Discovery Methods**

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 =

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

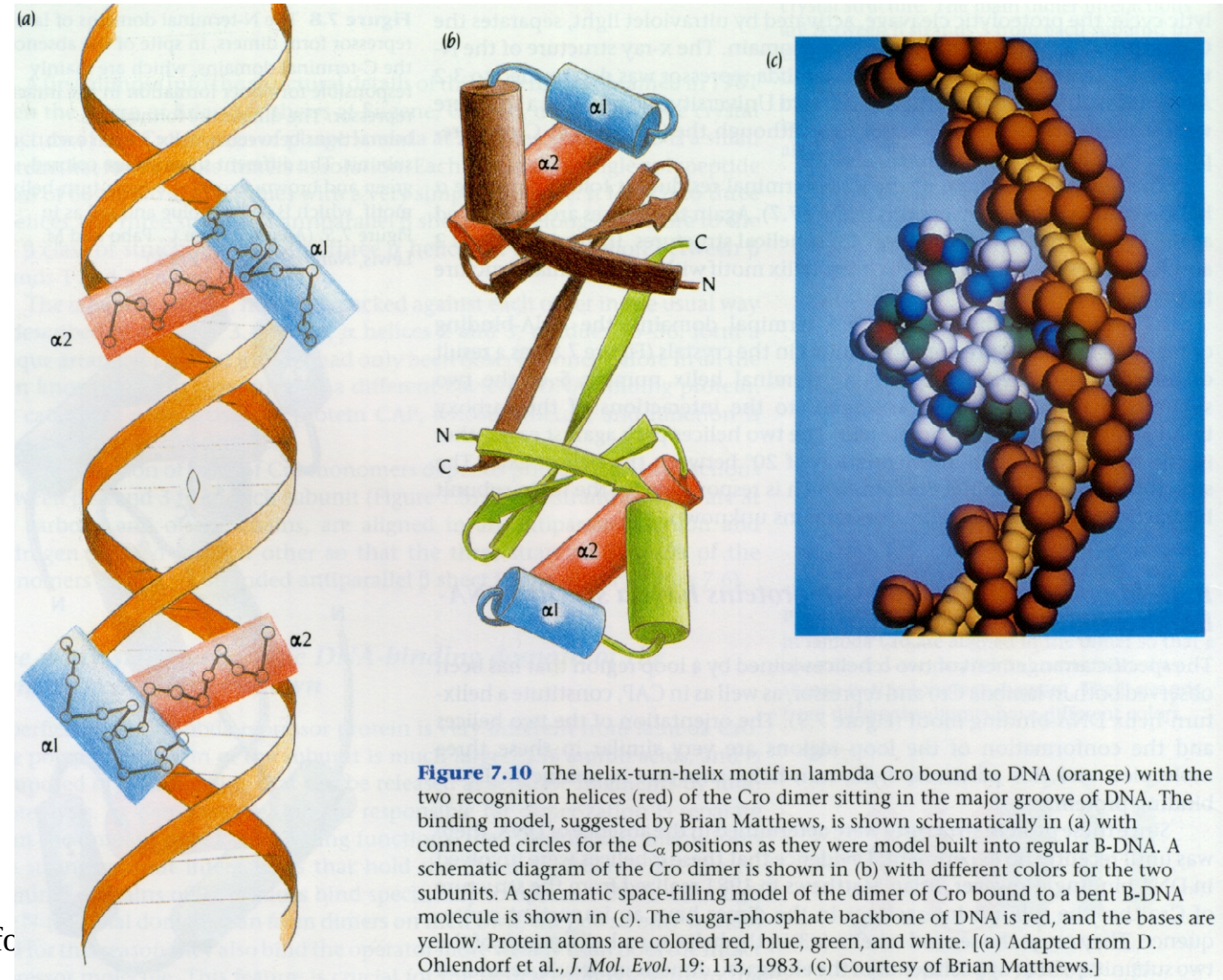


Figure 7.10 The helix-turn-helix motif in lambda Cro bound to DNA (orange) with the two recognition helices (red) of the Cro dimer sitting in the major groove of DNA. The binding model, suggested by Brian Matthews, is shown schematically in (a) with connected circles for the C_α positions as they were 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.]

HTH Motifs: Examples

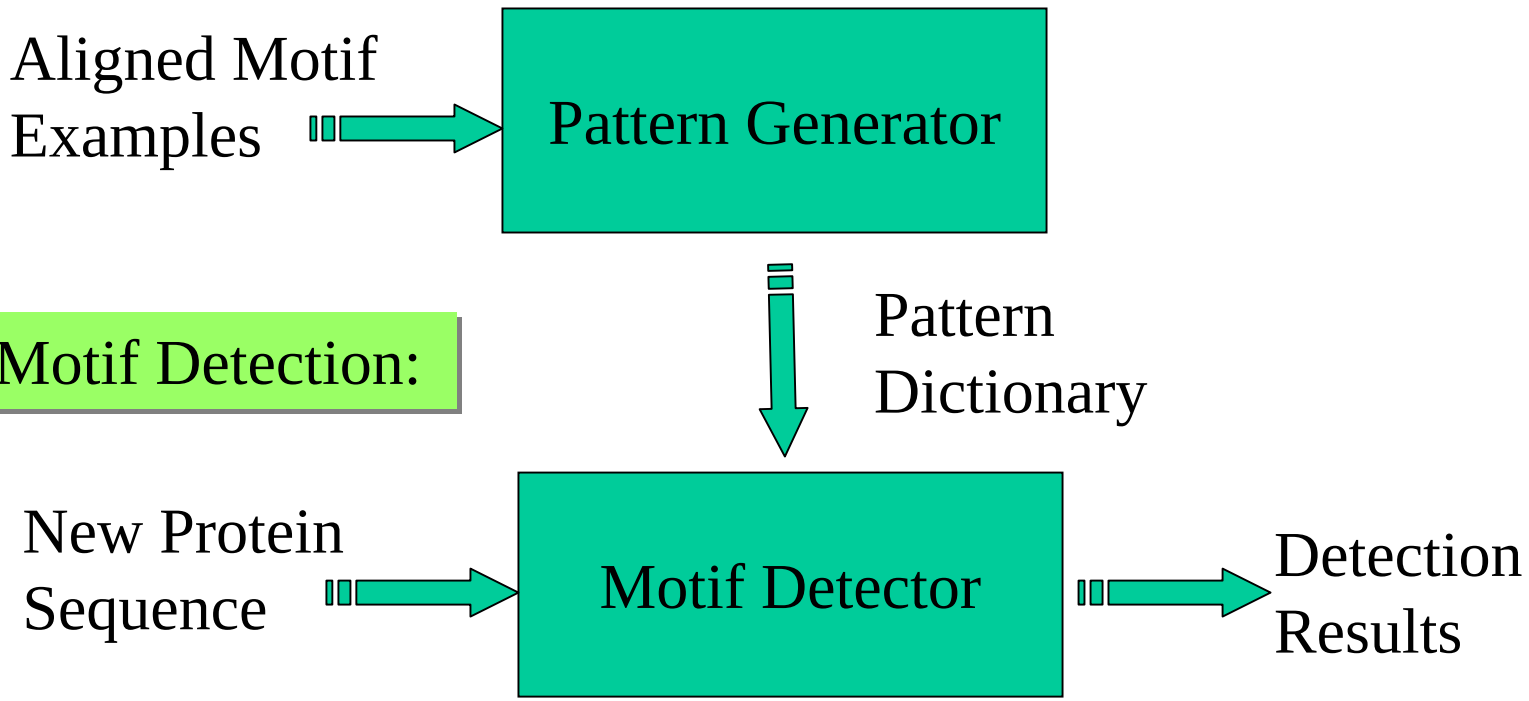
Loc	*Protein*	*Helix 2*									*Turn*				*Helix 3*								
	Name	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	**Cro**	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	**434 Cro**	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	**P22 Cro**	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	**Rep**	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

Basis for New Algorithm

- Combinations of residues in specific locations (may not be contiguous) contribute towards stabilizing a structure.
- Some reinforcing combinations are relatively rare.

New Motif Detection Algorithm

Pattern Generation:



Patterns

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

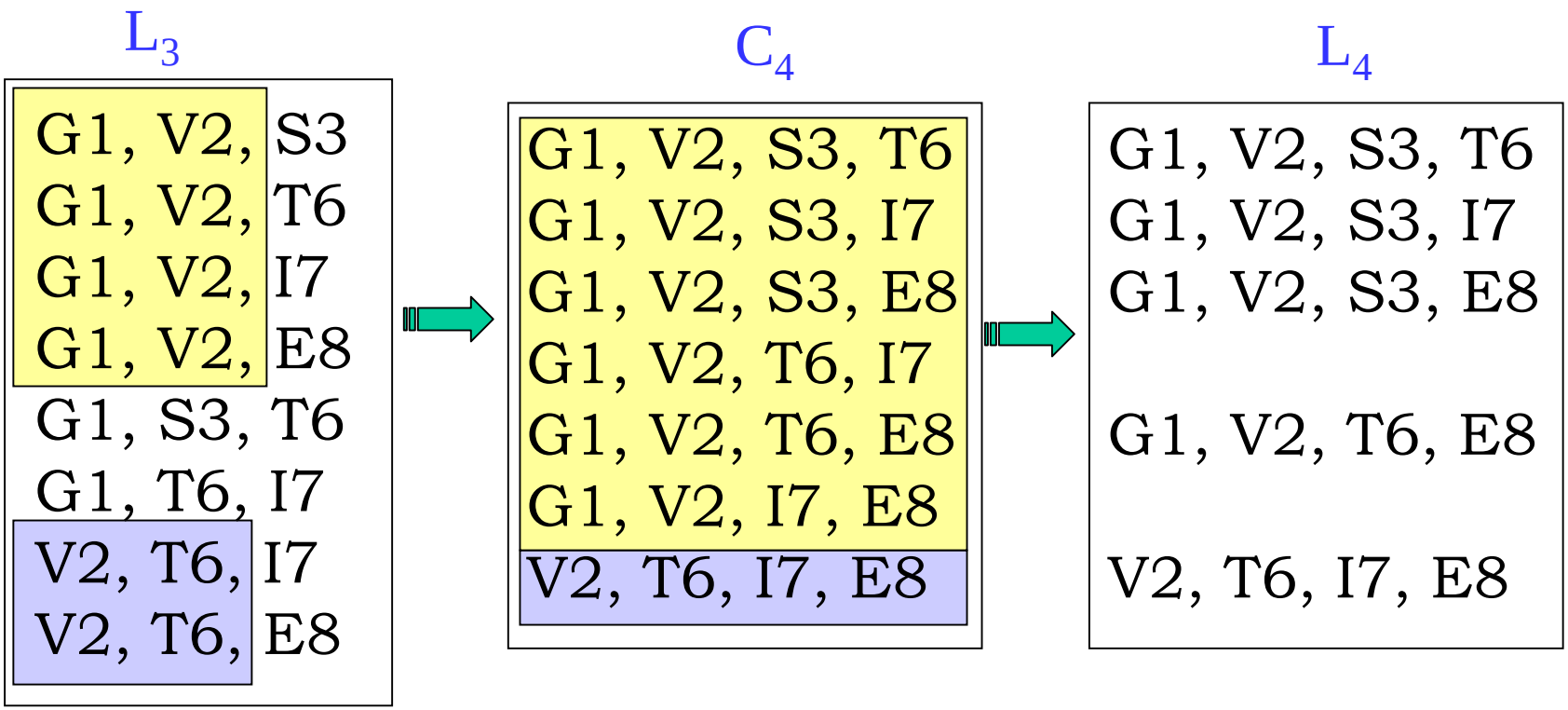
- Q1 G9 N20
- A5 G9 V10 I15

Pattern Mining Algorithm

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

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

Candidates Function



Motif Detection Algorithm

Algorithm Motif-Detection

Input : Motif length m, threshold score T, pattern dictionary L, and input protein sequence P[1..n].
Output : Information about motif(s) detected.

1. **for** each location i **do**
2. S := **MatchScore**(P[i..i+m-1], L).
3. **if** (S > T) **then**
4. 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 %)

Experiments

- Basic Implementation (Y. Gao)
- Improved implementation & comprehensive testing (K. Mathee, GN).
- Implementation for homeobox domain detection (X. Wang).
- Statistical methods to determine thresholds (C. Bu).
- Use of substitution matrix (C. Bu).
- Study of patterns causing errors (N. Xu).
- Negative training set (N. Xu).
- NN implementation & testing (J. Liu & X. He).
- HMM implementation & testing (J. Liu & X. He).

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