BLASTing Tools

- BLAST & FASTA: Search database for sequences that can be aligned with query sequence
- ProfileSearch: prepare profile from a multiple sequence alignment (Profilemake) and align profile with database sequence
- MAST: Search in database with profile representing ungapped sequence alignment
- Prosite, Interpro, Pfam: Search query sequence for patterns representative of protein families

More Tools

- PHI-BLAST: Searching for Regular Expressions & Motifs
- PSI-BLAST: Iterative alignment search for similar sequences that starts with a query sequence, builds a gapped multiple alignment, and then uses the alignment to augment the search

PSI-BLAST

- Version of BLAST to find protein families for a given query sequence. Helps to find distant neighbors and sequences with subtle relationships.
- Scheme
  - Perform regular BLAST and inspect results
  - If any interesting results, then iterate:
    - Align high-scoring matches and build profile
    - Perform BLAST using profile to find new hits
    - Show results with new hits highlighted

PHI-BLAST

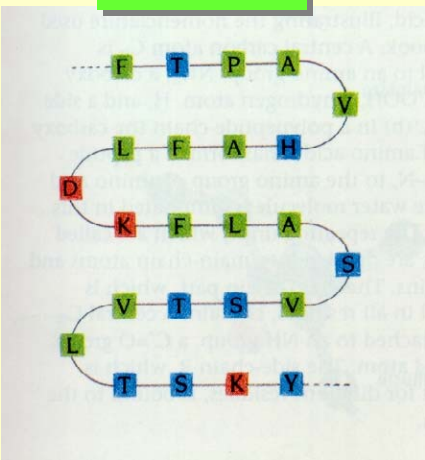
- Pattern-Hit Initiated BLAST: Search for patterns, for example,

  [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV]

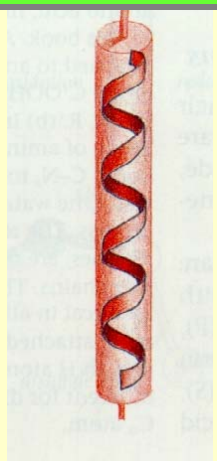
Protein Structures

- Sequences of amino acid residues
- 20 different amino acids

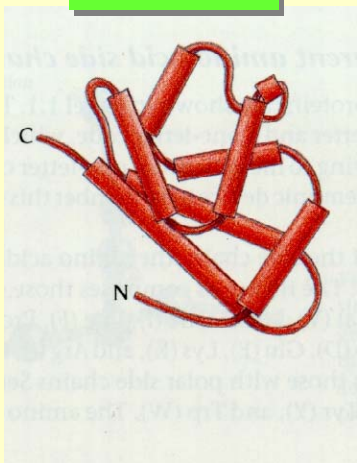
Primary



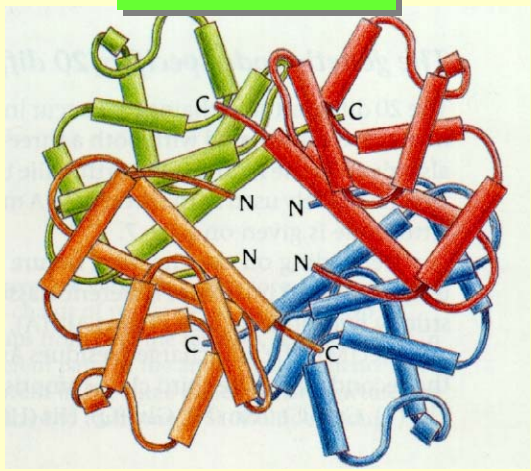
Secondary



Tertiary



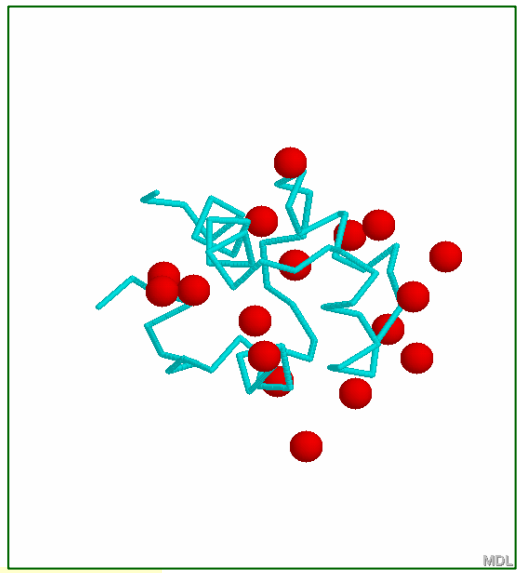
Quaternary



Proteins

- **Tertiary structures** are formed by packing secondary structural elements into a globular structure.

Myoglobin

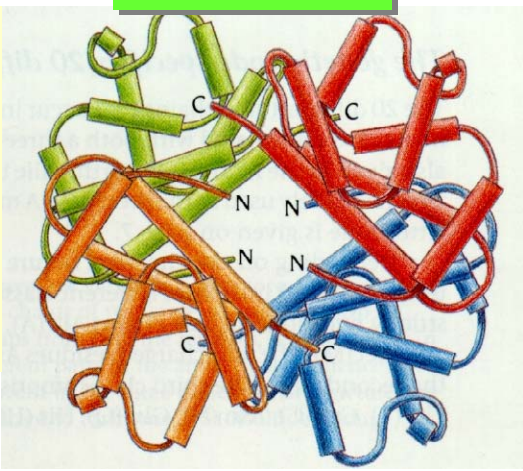


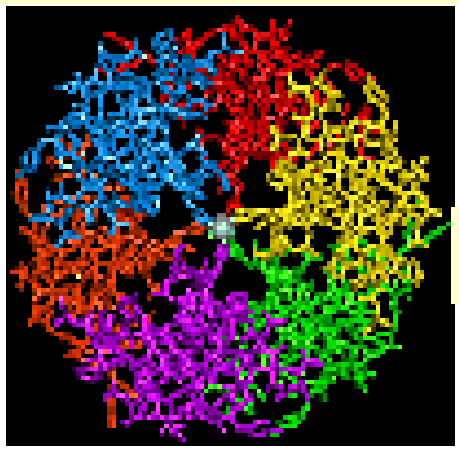
Lambda Cro

Quaternary Structures in Proteins

- The final structure may contain more than one “chain” arranged in a **quaternary structure**.

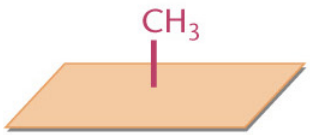
Quaternary



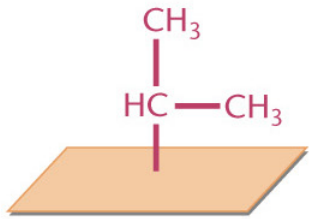


Insulin Hexamer

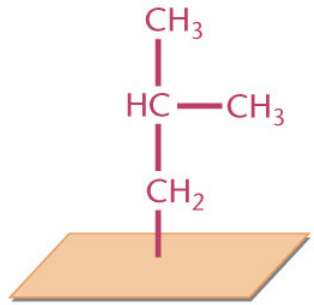
1. Nonpolar: Hydrophobic



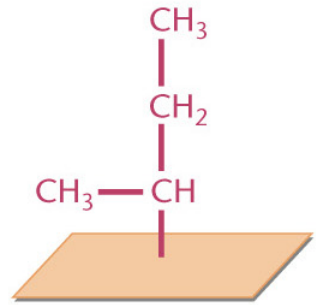
Alanine (ala–A)



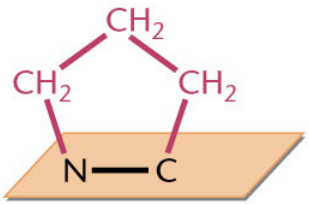
Valine (val–V)



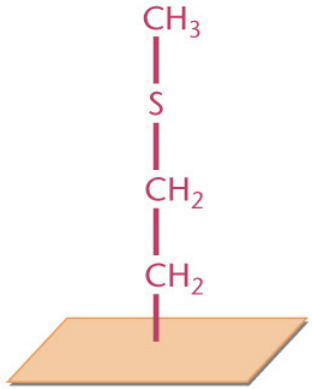
Leucine (leu–L)



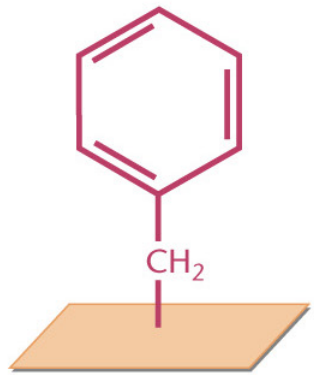
Isoleucine (ile–I)



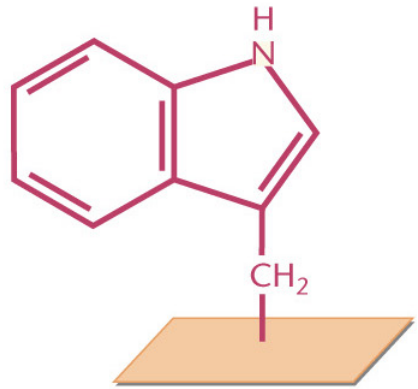
Proline (pro–P)



Methionine (met–M)



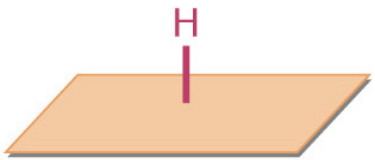
Phenylalanine (phe–F)



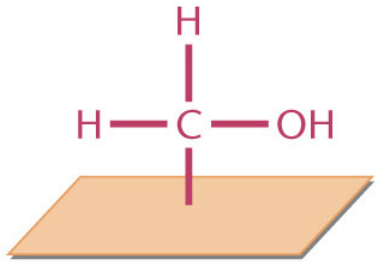
Tryptophan (trp–W)

Amino Acid Structures from Klug & Cummings

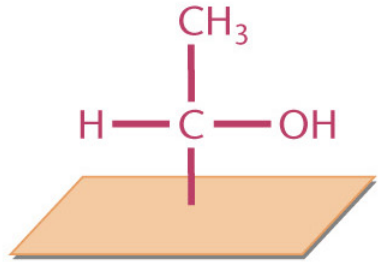
2. Polar: Hydrophilic



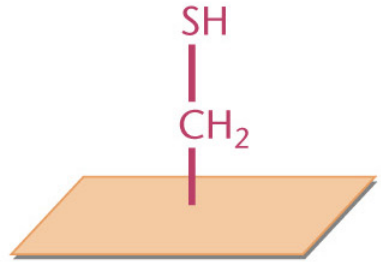
Glycine (gly–G)



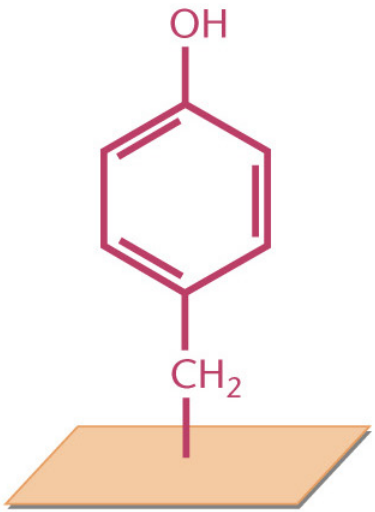
Serine (ser–S)



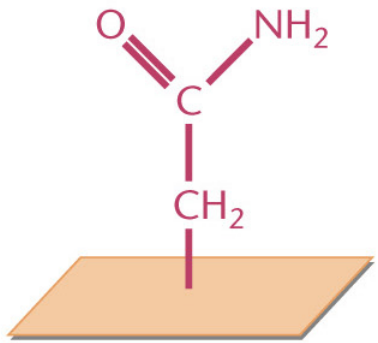
Threonine (thr–T)



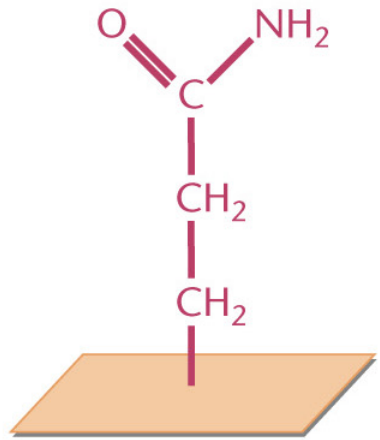
Cysteine (cys–C)



Tyrosine (tyr–Y)



Asparagine (asn–N)

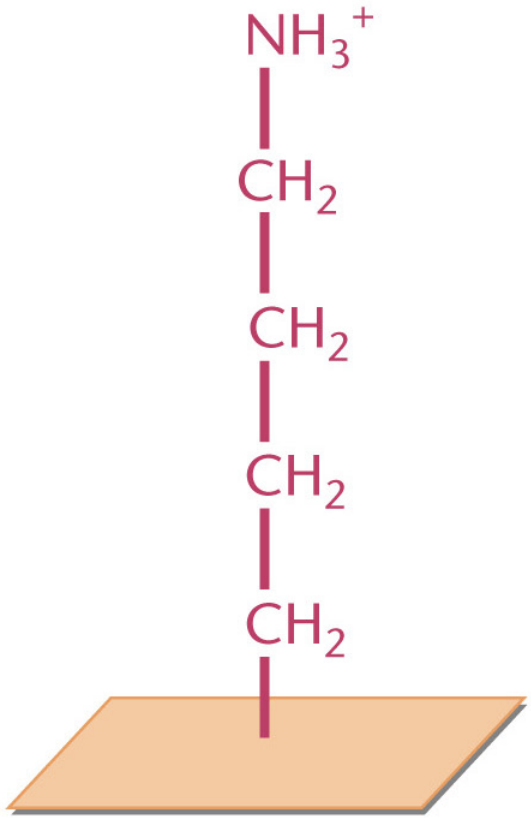


Glutamine (gln–Q)

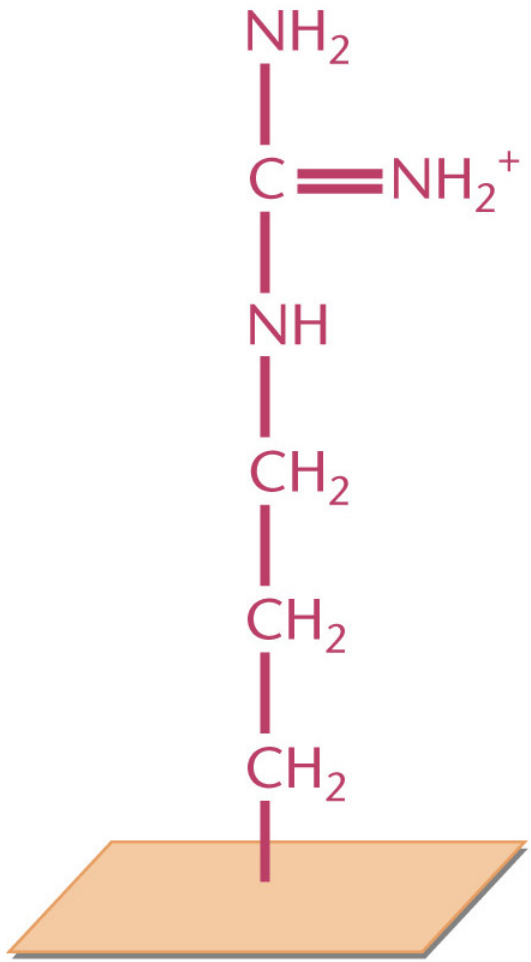
Amino Acid Structures from Klug & Cummings

3. Polar: positively charged (basic)

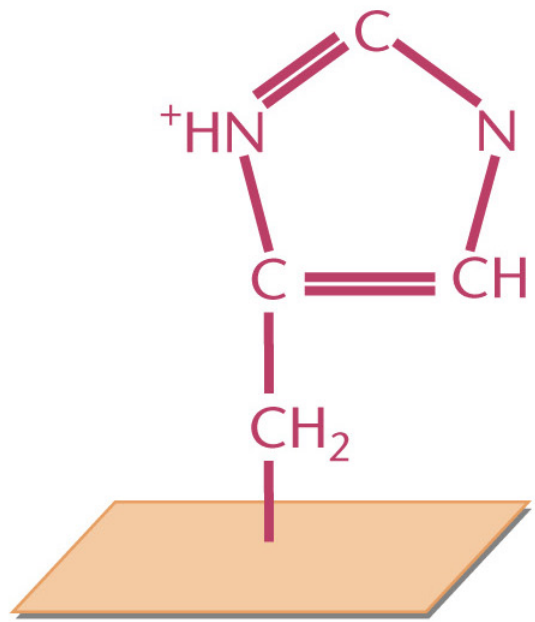
Amino Acid Structures from Klug & Cummings



Lysine (lys–K)

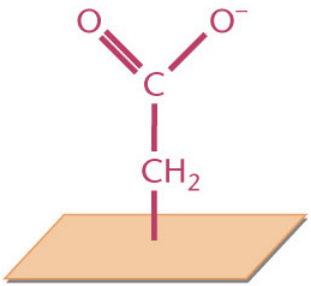


Arginine (arg–R)

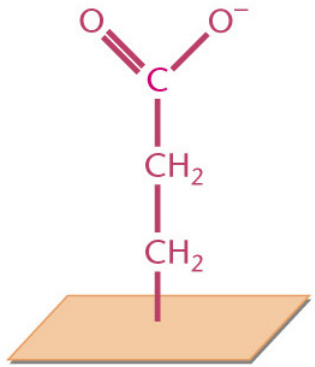


Histidine (his–H)

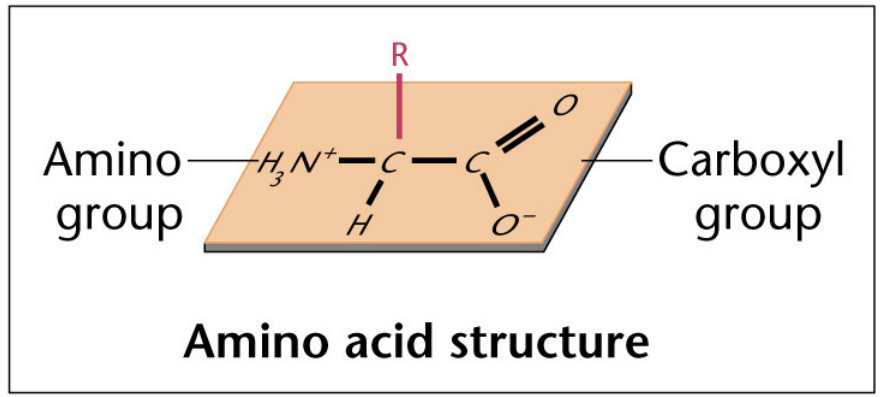
4. Polar: negatively charged (acidic)



Aspartic acid (asp–D)



Glutamic acid (glu–E)



Amino Acid Structures from Klug & Cummings

More on Secondary Structures

- **α-helix**
  - Main chain with peptide bonds
  - Side chains project outward from helix
  - Stability provided by H-bonds between CO and NH groups of residues 4 locations away.
- **β-strand**
  - Stability provided by H-bonds with one or more β-strands, forming β-sheets. Needs a β-turn.

Secondary Structure Prediction Software

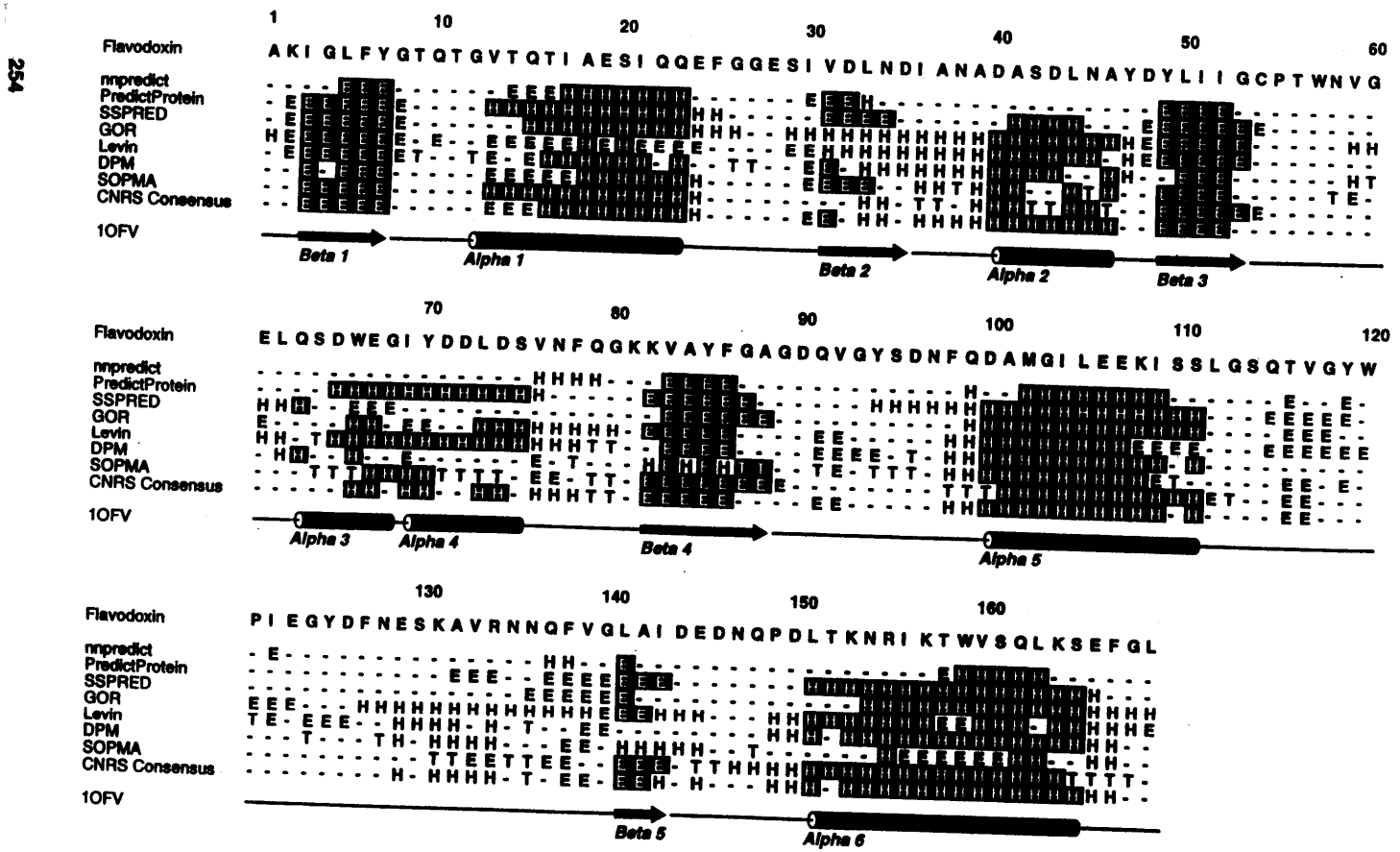


Figure 11.3 Comparison of secondary structure predictions by various methods. The sequence of flavodoxin, an α/β protein, was used as the query and is shown on the first line of the alignment. For each prediction, H denotes an α helix, E a β strand, T a β turn; all other positions are assumed to be random coil. Correctly assigned residues are shown in inverse type. The methods used are listed along the left side of the alignment and are described in the text. At the bottom of the figure is the secondary structure assignment given in the PDB file for flavodoxin (1OFV, Smith et al., 1983).

PDB: Protein Data Bank

- Database of protein tertiary and quaternary structures and protein complexes. http://www.rcsb.org/pdb/
- Over 29,000 structures as of Feb 1, 2005.
- Structures determined by
  - NMR Spectroscopy
  - X-ray crystallography
  - Computational prediction methods
- Sample PDB file: Click here [ ]

Active Sites

Active sites in proteins are usually hydrophobic pockets/crevices/troughs that involve sidechain atoms.

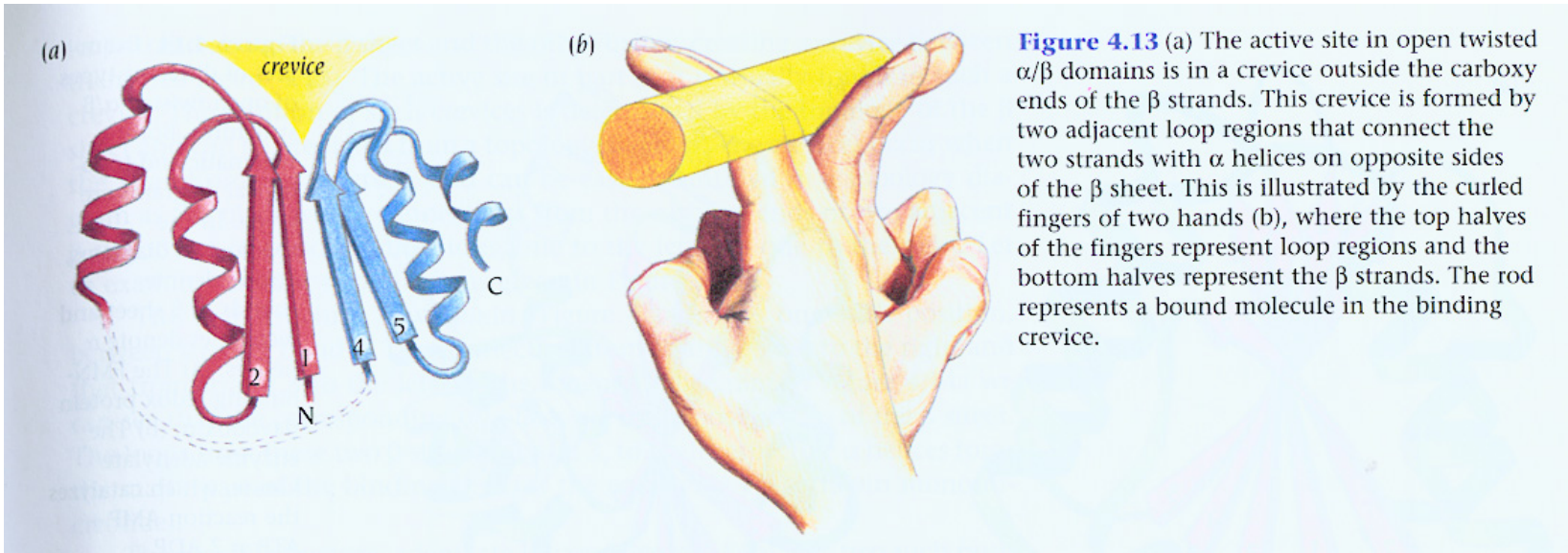
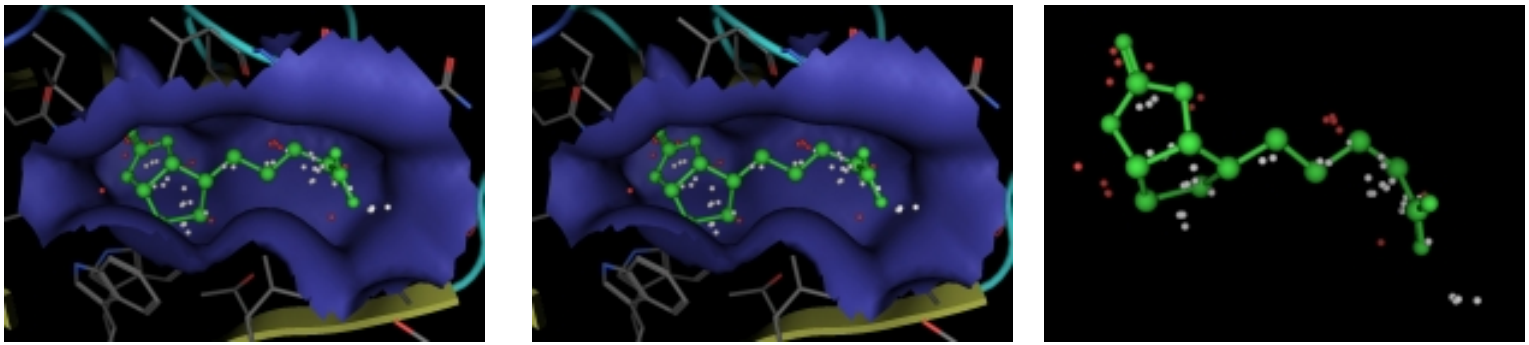


Figure 4.13 (a) The active site in open twisted α/β domains is in a crevice outside the carboxy ends of the β strands. This crevice is formed by two adjacent loop regions that connect the two strands with α helices on opposite sides of the β sheet. This is illustrated by the curled fingers of two hands (b), where the top halves of the fingers represent loop regions and the bottom halves represent the β strands. The rod represents a bound molecule in the binding crevice.

Active Sites



Left PDB 3RTD (streptavidin) and the first site located by the MOE Site Finder. **Middle** 3RTD with complexed ligand (biotin). **Right** Biotin ligand overlaid with calculated alpha spheres of the first site.

Motif Detection Tools

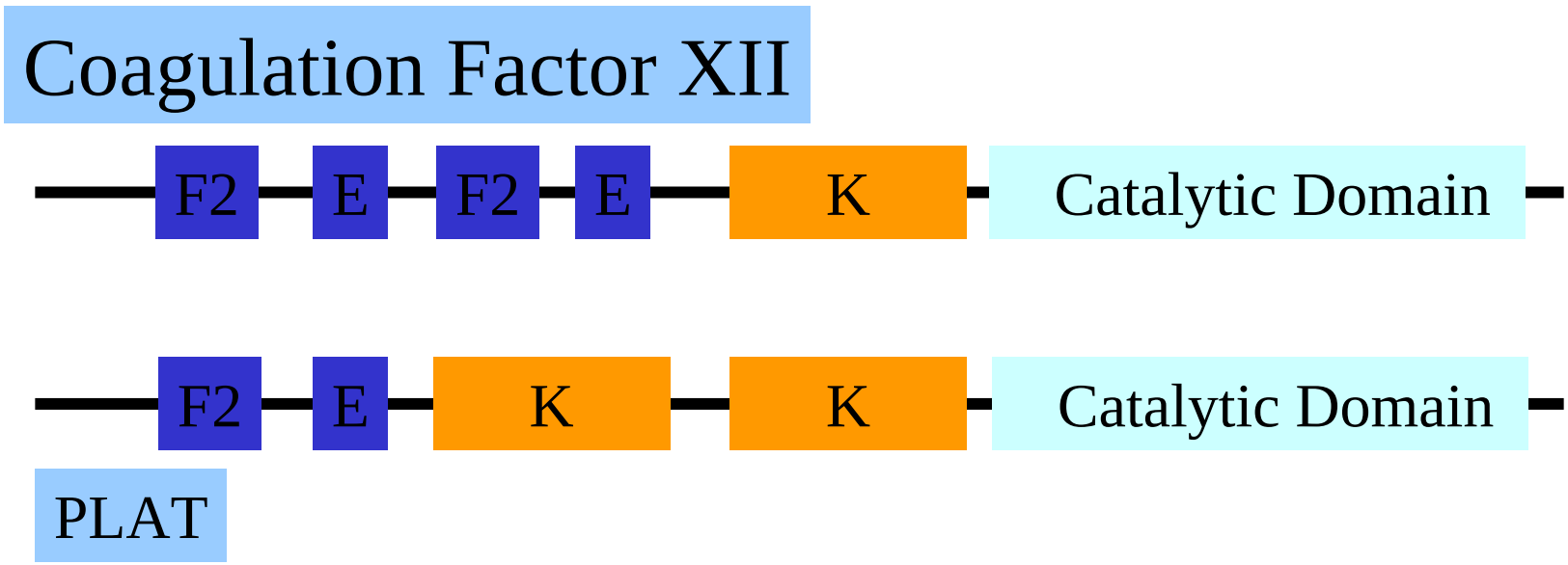
- PROSITE (Database of protein families & domains)
  - Try PDOC00040. Also Try PS00041
- PRINTS Sample Output
- BLOCKS (multiply aligned ungapped segments for highly conserved regions of proteins; automatically created) Sample Output
- Pfam (Protein families database of alignments & HMMs)
  - Multiple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

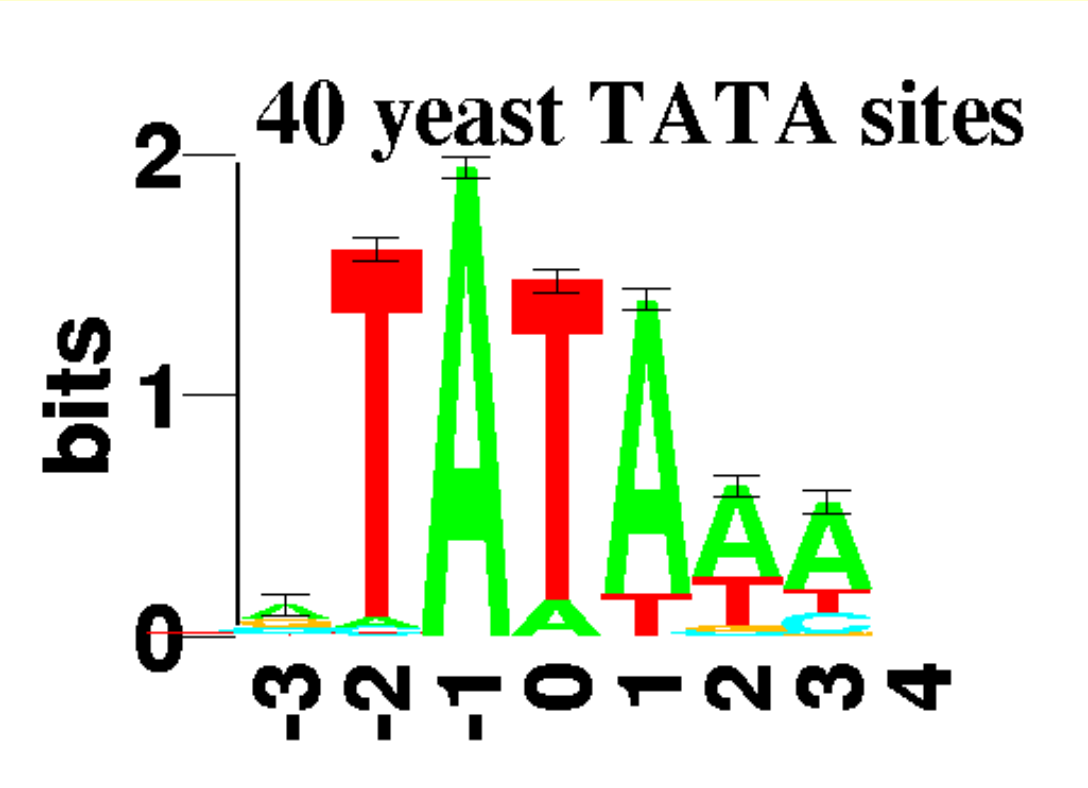
...**Y**K**C**GL**C**ERS**F**VEKSA**L**SR**H**ORV**H**KN...
3 6 19 23

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

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

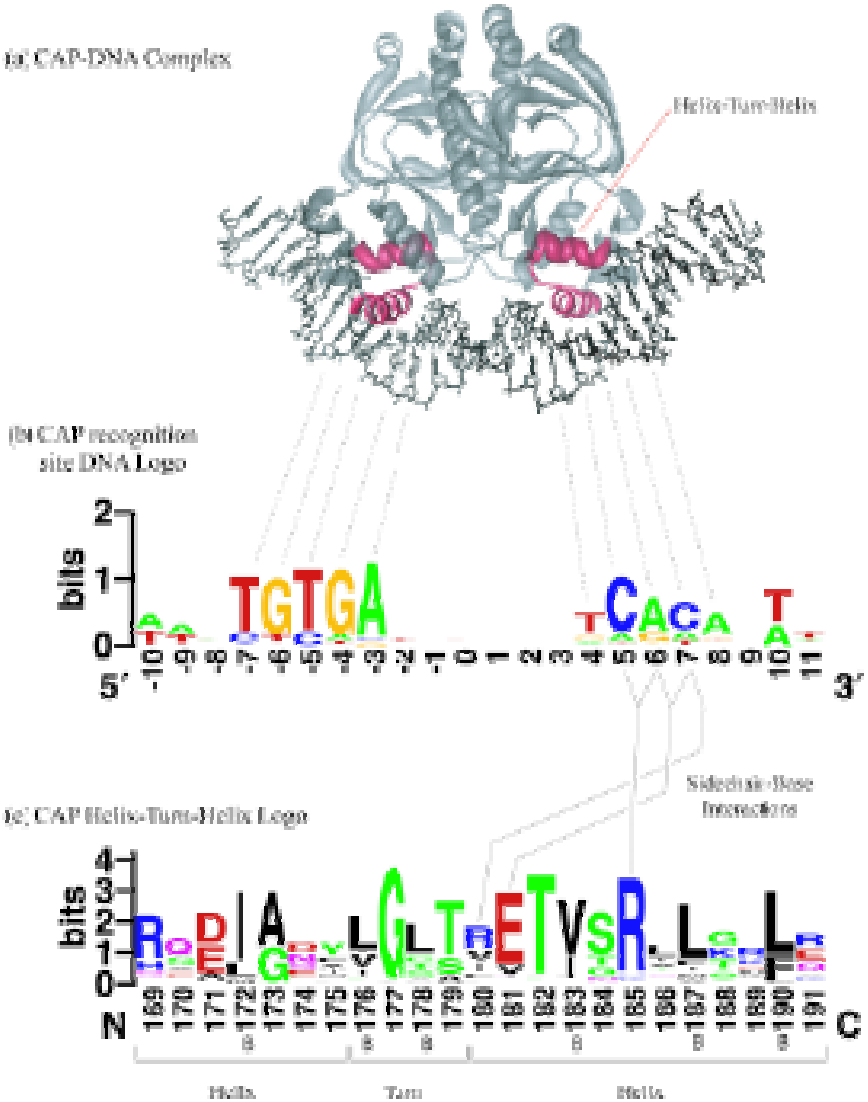
Motifs in DNA Sequences

- 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

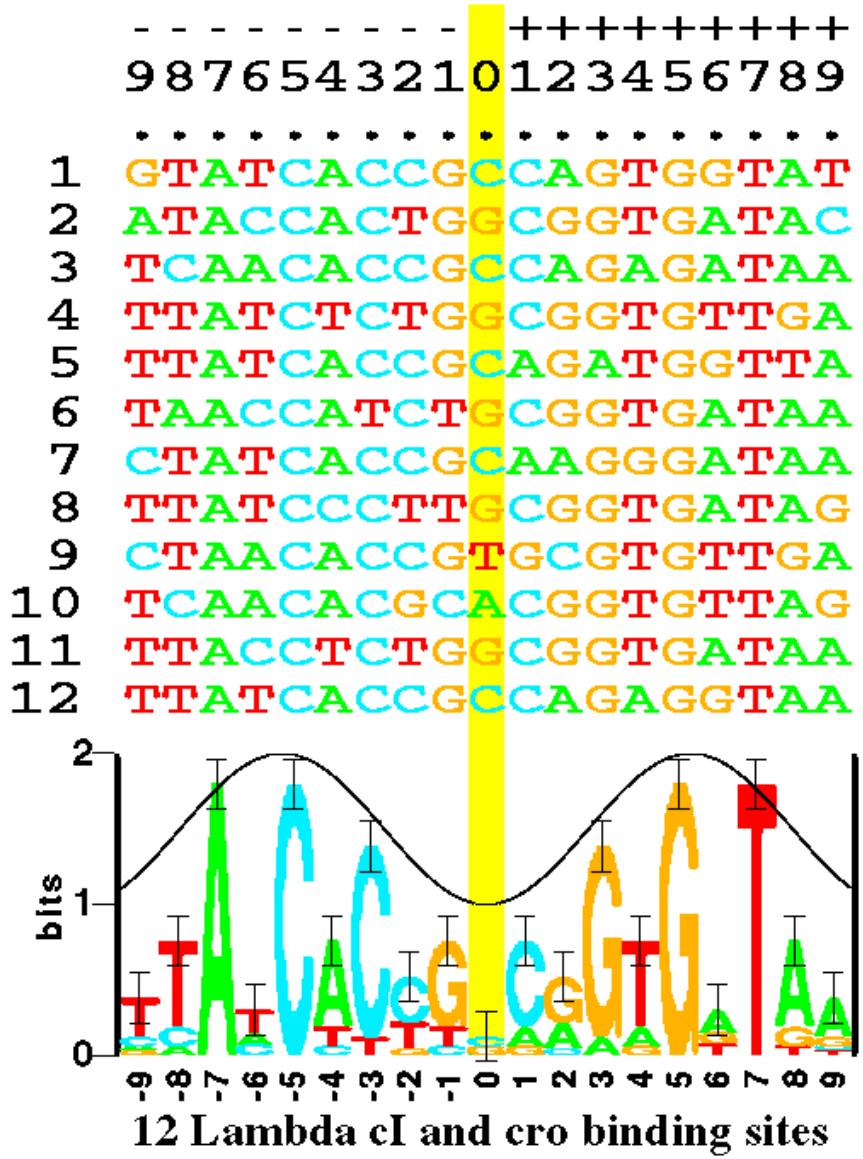
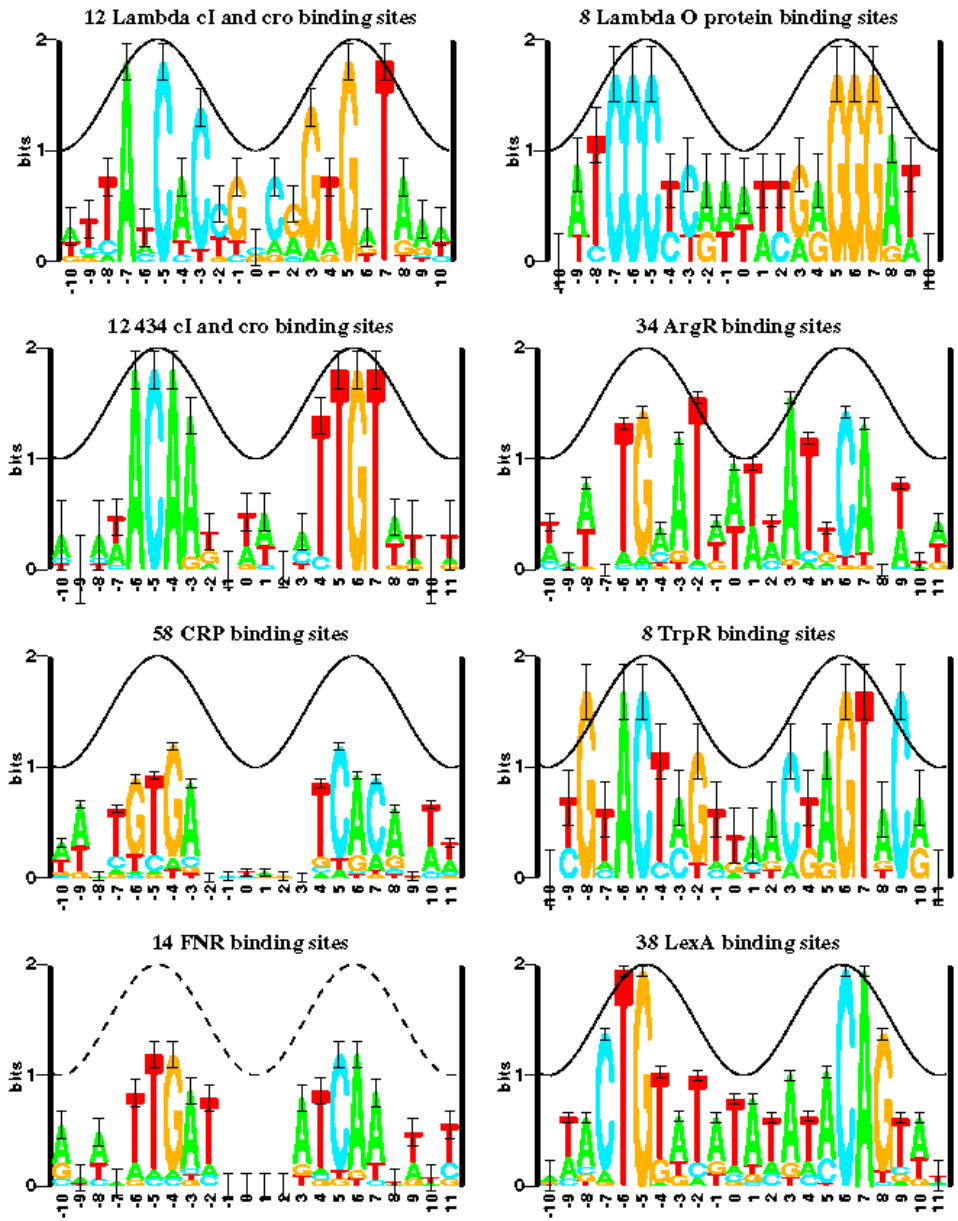


Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered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

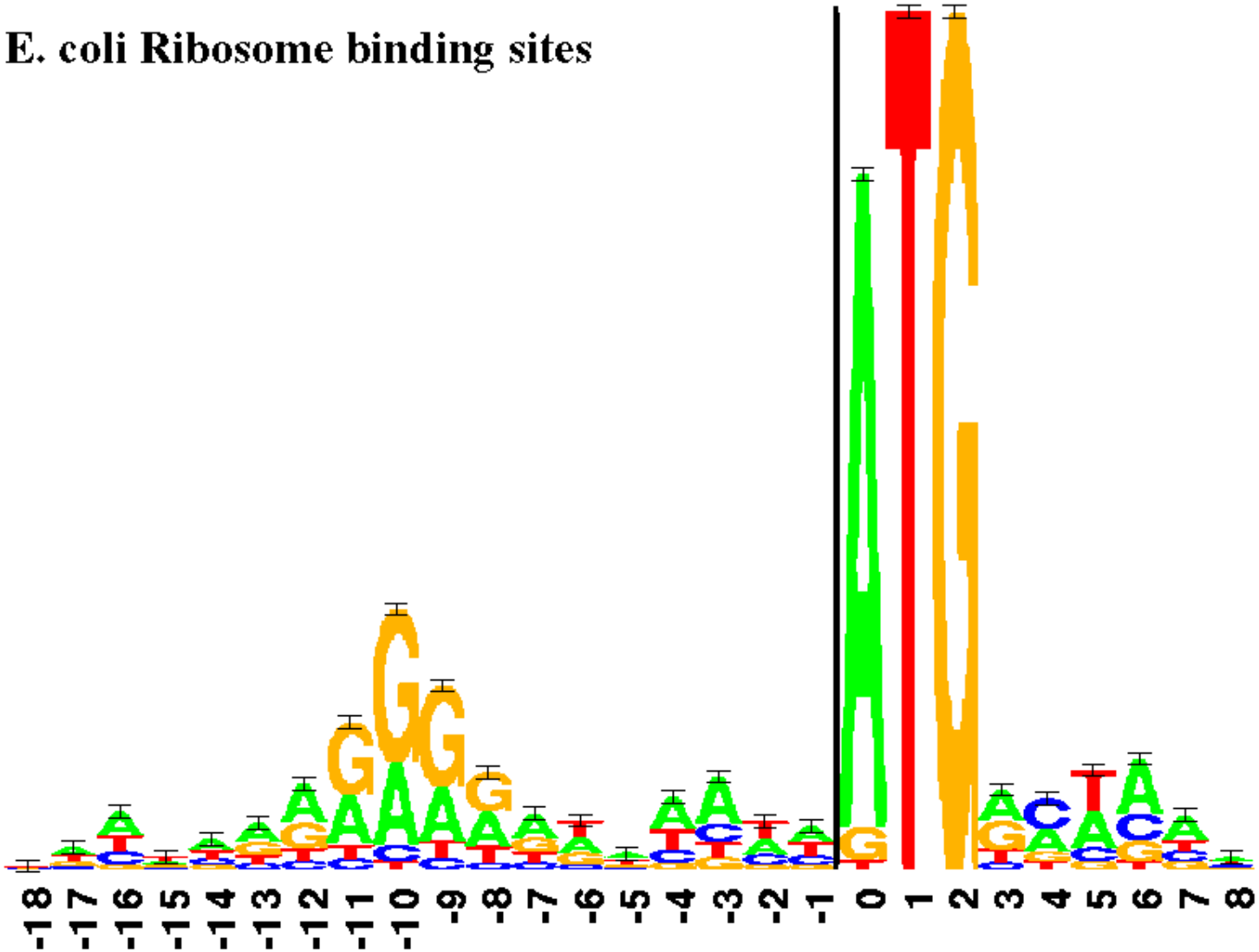
http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

More Motifs in *E. Coli* DNA Sequences



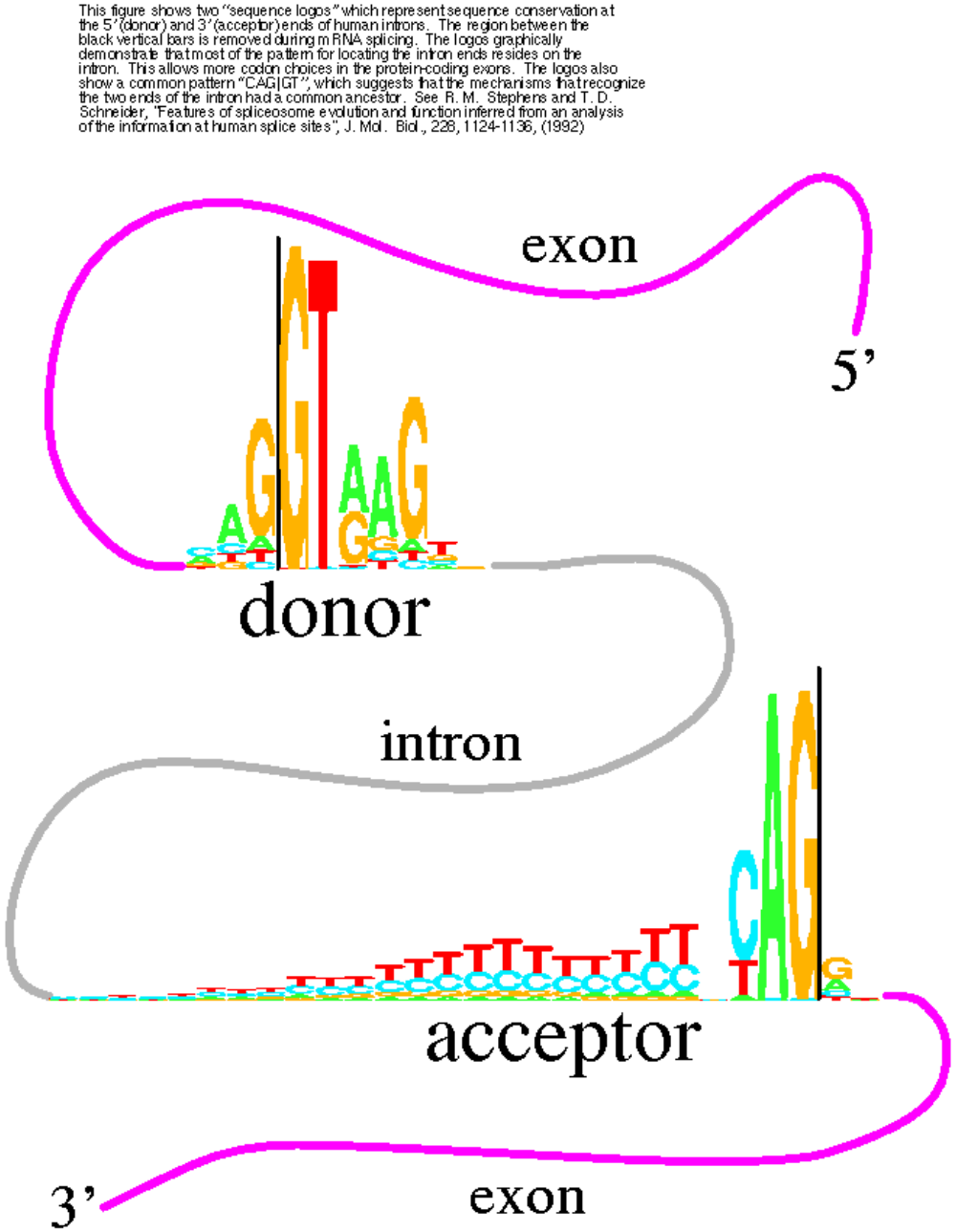
http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

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

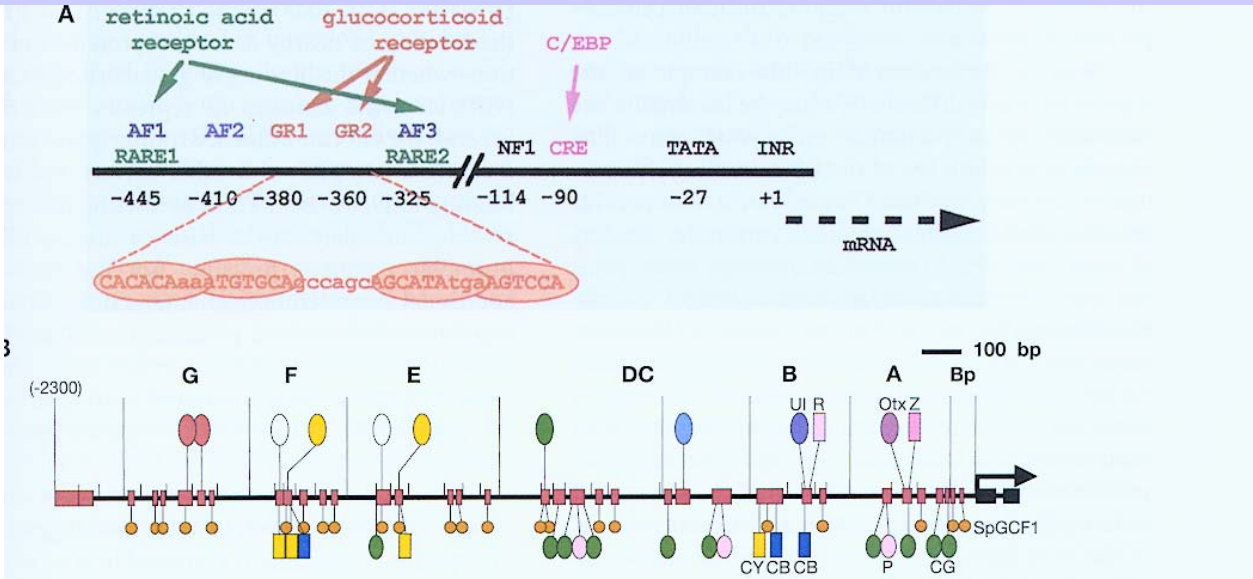


FIGURE 9.13. Regulatory elements of two promoters. (*A*) The rat *pepCK* gene. The relative positions of the TF-binding sites are illustrated (Yamada et al. 1999). The glucocorticoid response unit (GRU) includes three accessory factor–binding sites (AF1, AF2, and AF3), two glucocorticoid response elements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

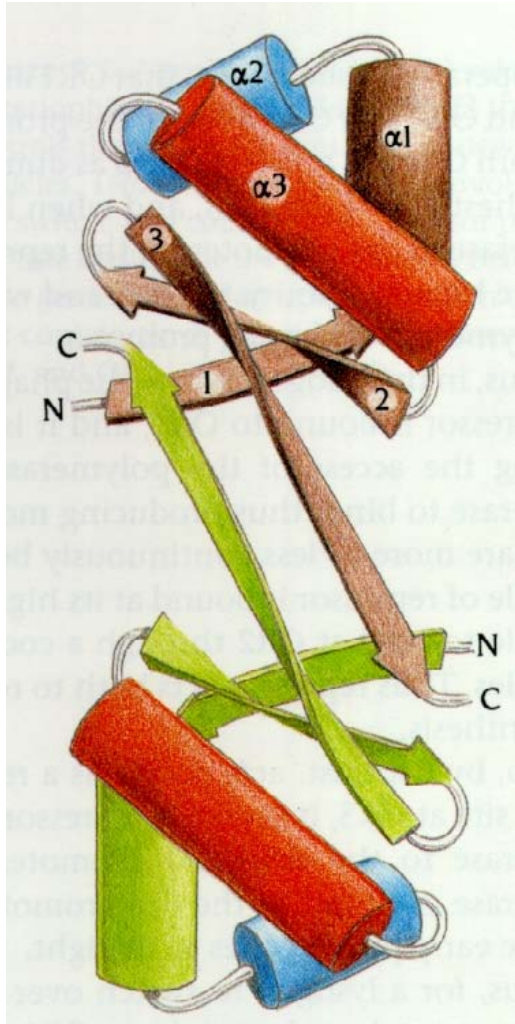
- Profile Method
  - If many examples of the motif are known, then
    - **Training**: build a **Profile** and compute a **threshold**
    - **Testing**: **score** against profile
- **Gibbs Sampling**
- **Expectation Method**
- **HMM**
- **Combinatorial Pattern Discovery Methods**

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

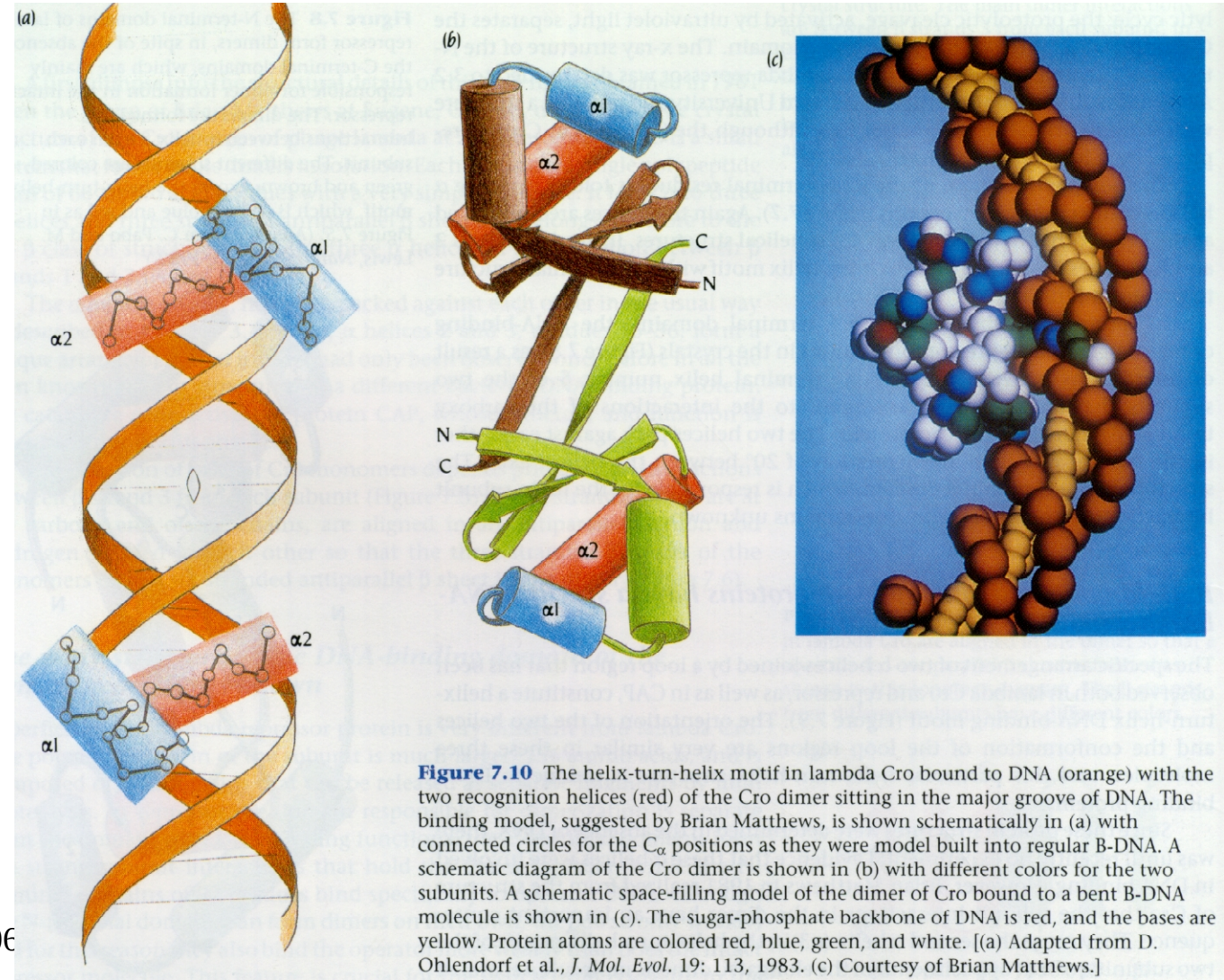


Figure 7.10 The helix-turn-helix motif in lambda Cro bound to DNA (orange) with the two recognition helices (red) of the Cro dimer sitting in the major groove of DNA. The binding model, suggested by Brian Matthews, is shown schematically in (a) with connected circles for the C_α positions as they were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

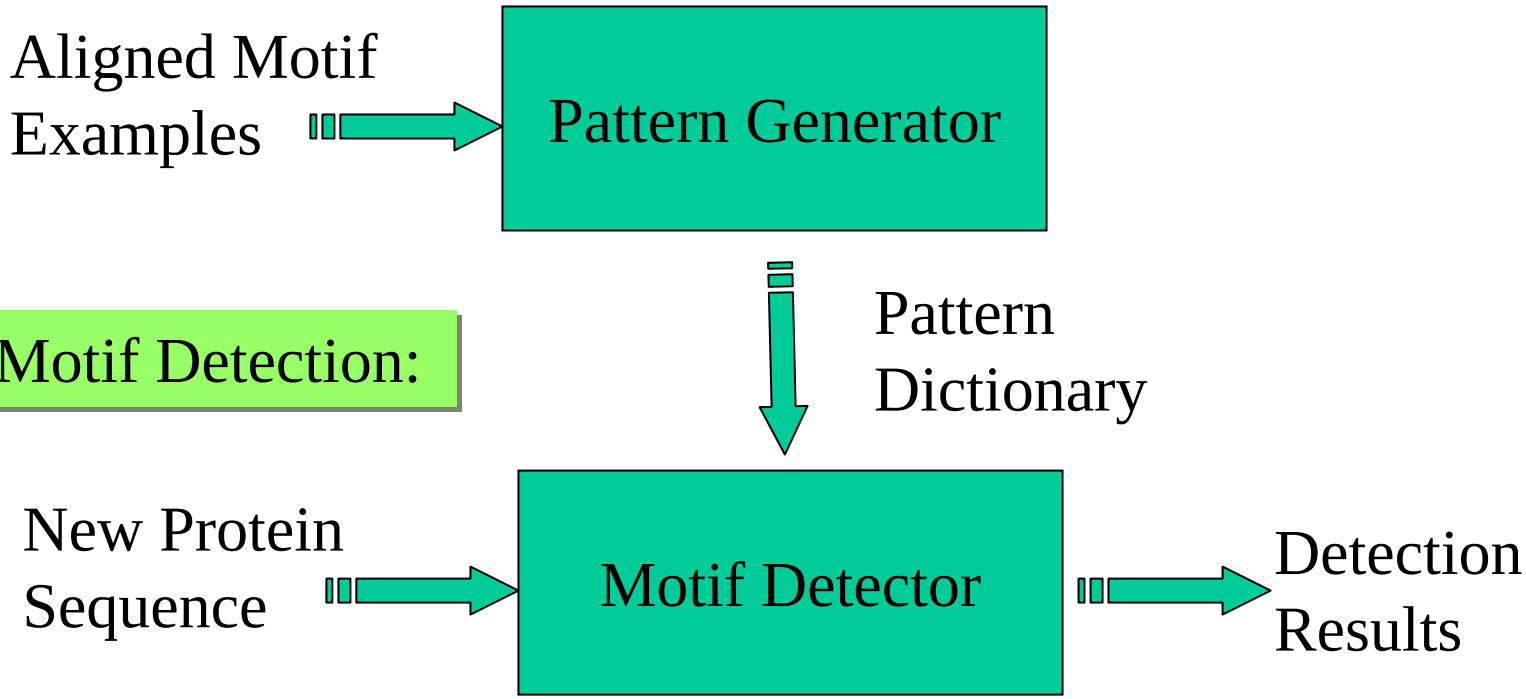
Loc	*Protein*	*Helix 2*									*Turn*				*Helix 3*								
	Name	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	**Cro**	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	**434 Cro**	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	**P22 Cro**	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	**Rep**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

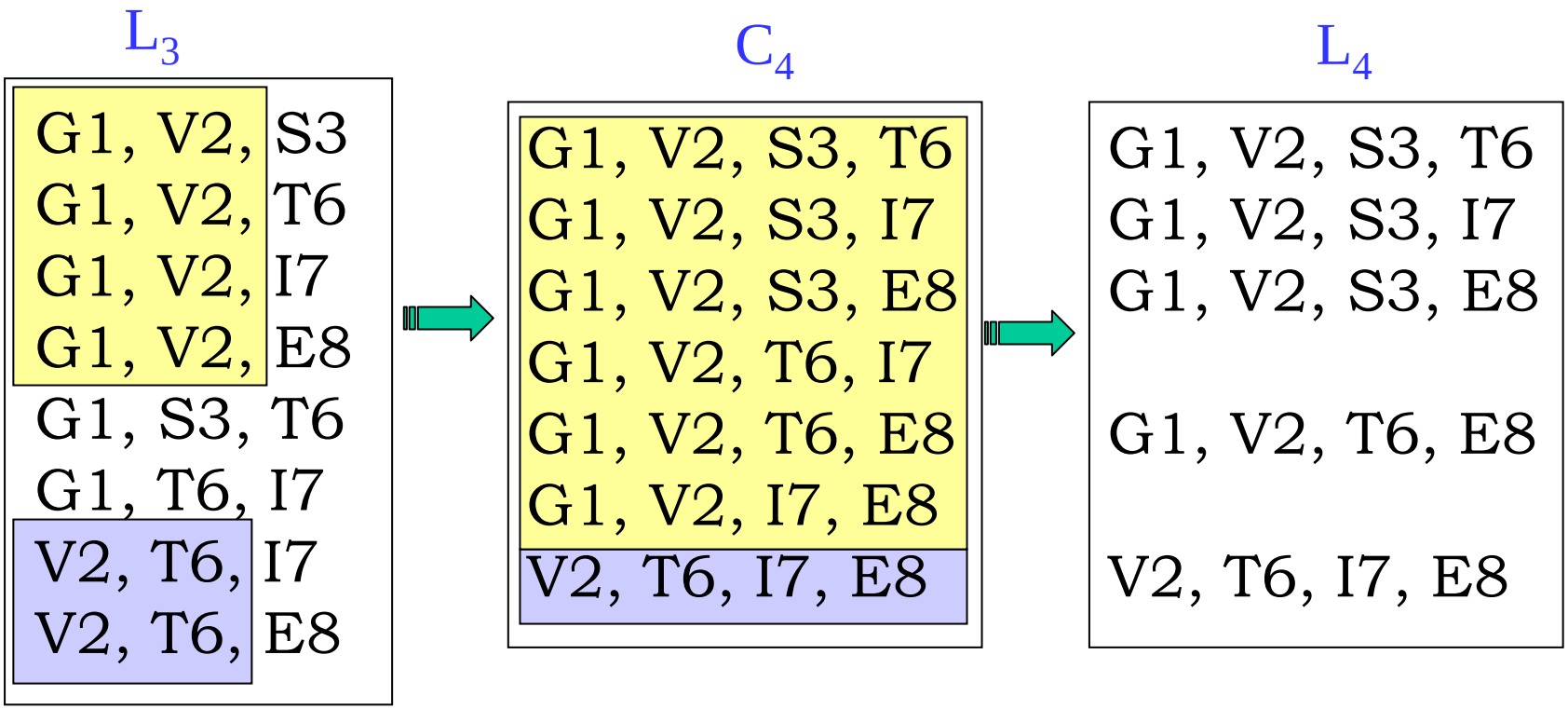
- Q1 G9 N20
- A5 G9 V10 I15

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 %)

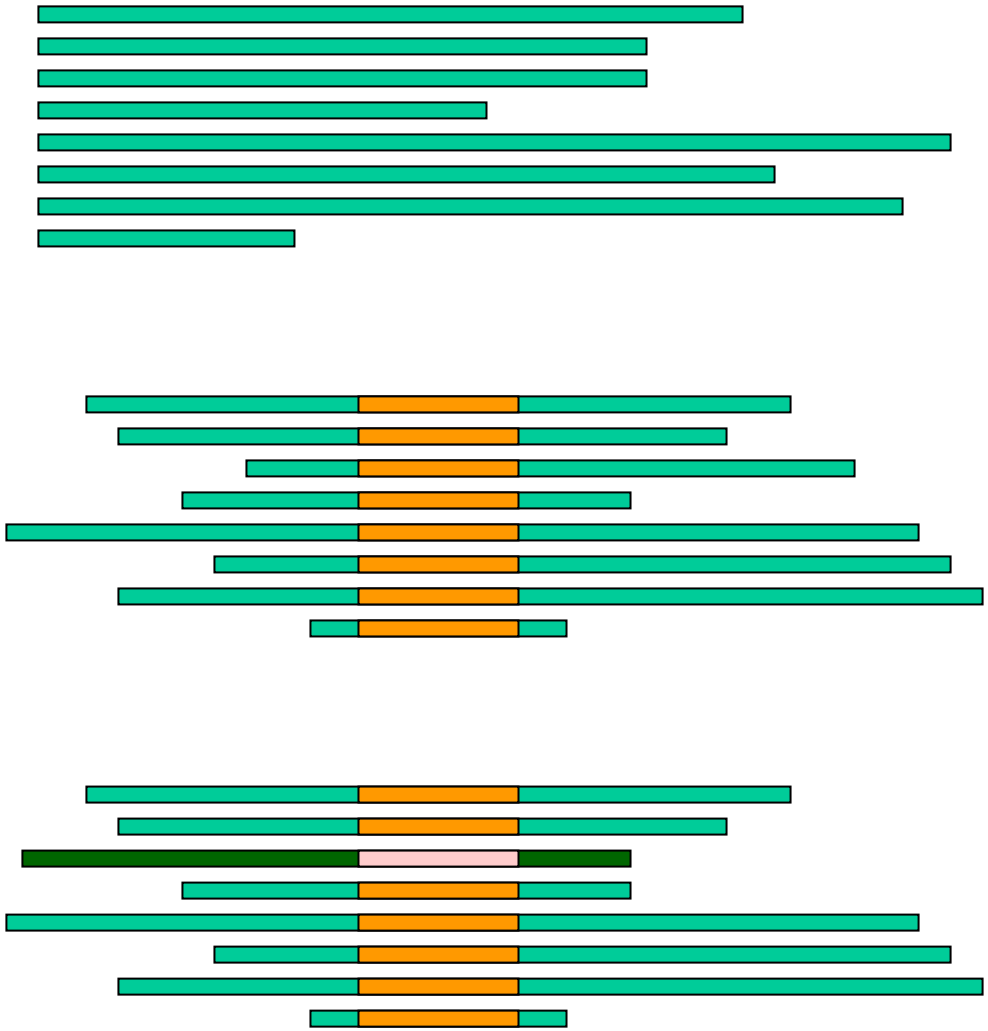
Experiments

- 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).

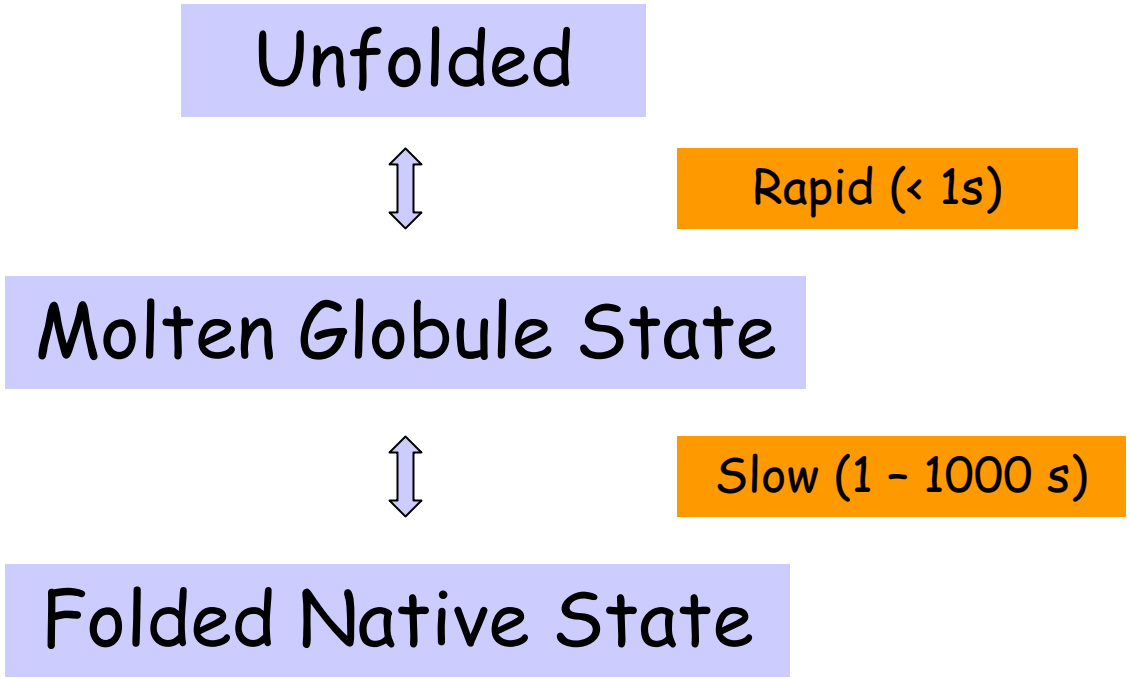
Motif Detection (TFBMs)

- 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



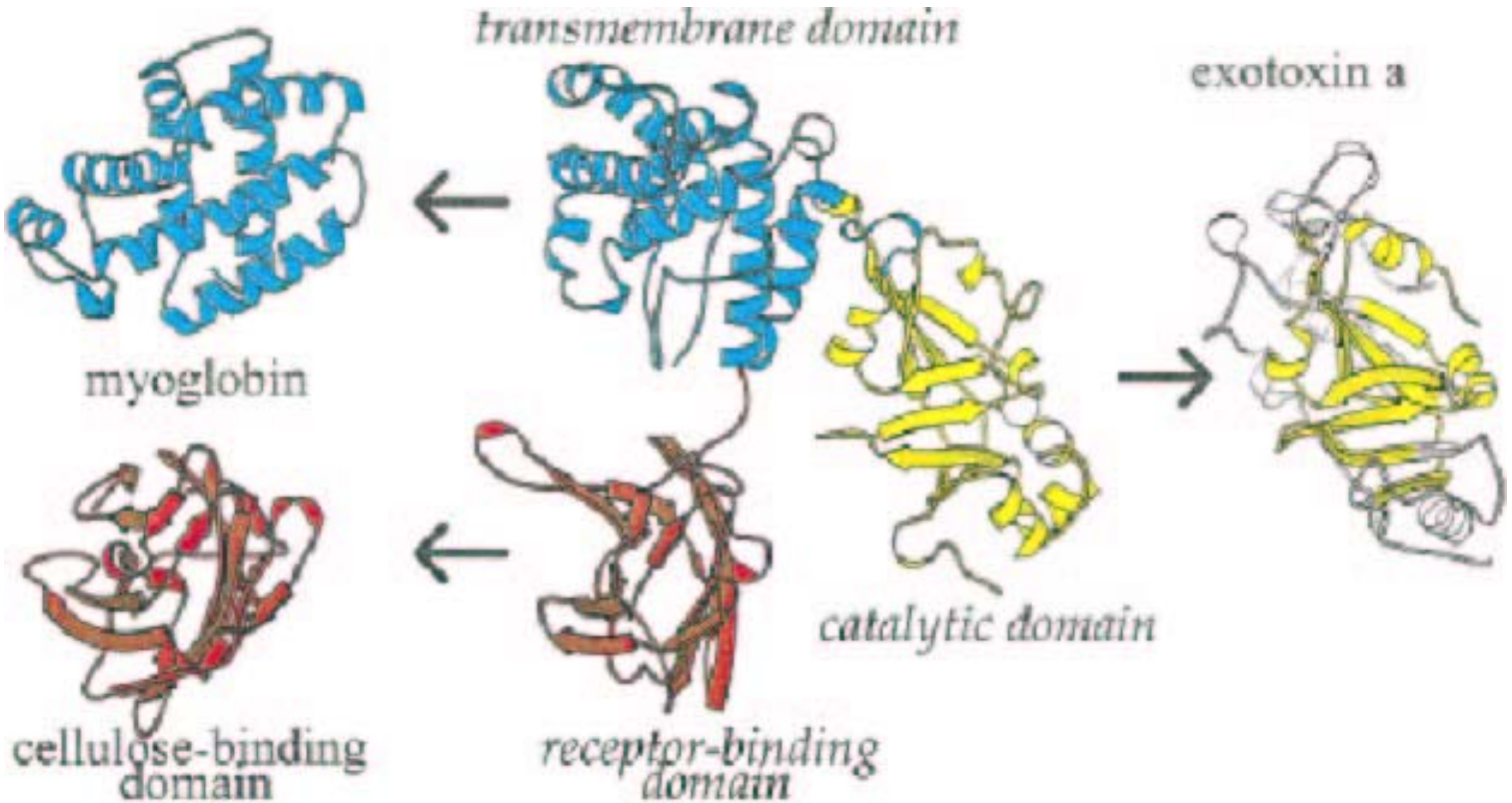
Protein Folding



- How to find minimum energy configuration?

Modular Nature of Protein Structures

Example: Diphtheria Toxin



Protein Structures

- Most proteins have a hydrophobic core.
- Within the core, specific interactions take place between amino acid side chains.
- Can an amino acid be replaced by some other amino acid?
  - Limited by space and available contacts with nearby amino acids
- Outside the core, proteins are composed of loops and structural elements in contact with water, solvent, other proteins and other structures.

Viewing Protein Structures

- SPDBV
- RASMOL
- CHIME

Structural Classification of Proteins

- Over 1000 protein families known
  - Sequence alignment, motif finding, block finding, similarity search
- SCOP (Structural Classification of Proteins)
  - Based on structural & evolutionary relationships.
  - Contains ~ 40,000 domains
  - Classes (groups of folds), Folds (proteins sharing folds), Families (proteins related by function/evolution), Superfamilies (distantly related proteins)

SCOP Family View

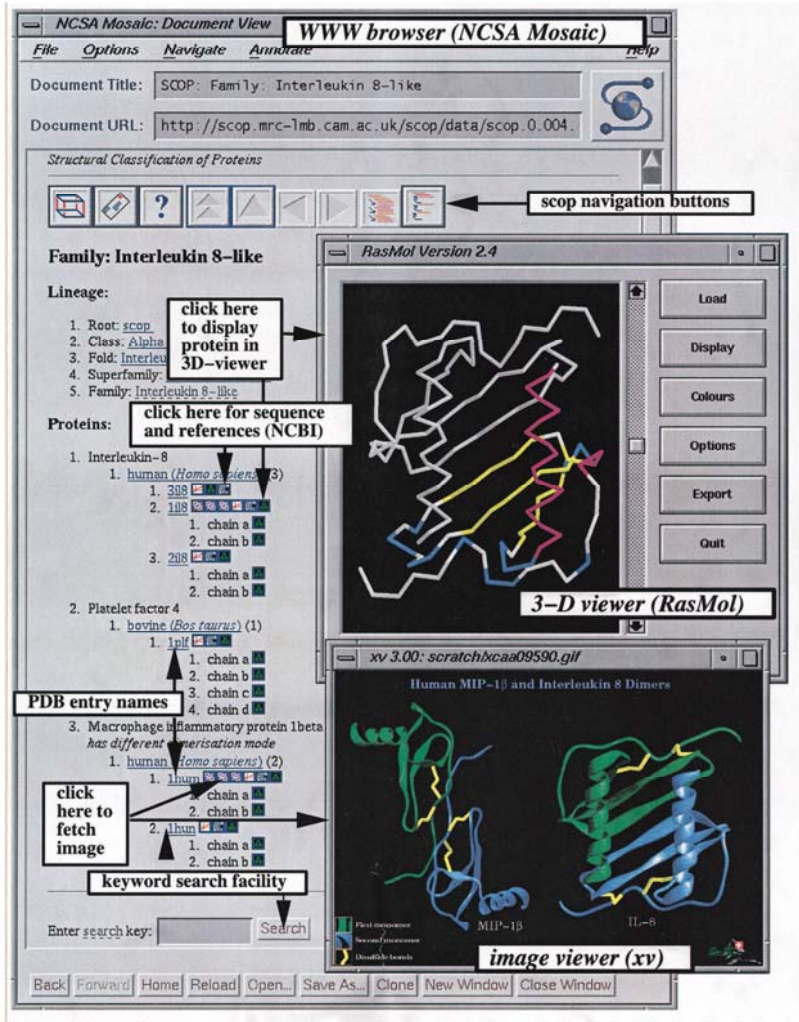


Figure 2. A typical scop session is shown on a unix workstation. A scop page, of the Interleukin 8-like family, is displayed by the *WWW browser program (NCSA Mosaic)* (Schatz & Hardin, 1994). Navigating through the tree structure is accomplished by selecting any underlined entry by clicking on buttons (at the top of each page) and by keyword searching (at the bottom of each page). The static image comparing two proteins in this family was downloaded by clicking on the icon indicated and is displayed by image-viewer program *xv*. By clicking on one of the green icons, commands were sent to a molecular viewer program (*RasMol*) written by Roger Sayle (Sayle, 1994), instructing it to automatically display the relevant PDB file and colour the domain in question by secondary structure. Since sending large PDB files over the network can be slow, this feature of scop can be configured to use local copies of PDB files if they are available. Equivalent WWW browsers, image-display programs and molecular viewers are also available free for Windows-PC and Macintosh platforms.

CATH: Protein Structure Classification

- Semi-automatic classification; ~36K domains
- 4 levels of classification:
  - Class (C), depends on sec. Str. Content
    - α class, β class, α/β class, $\alpha+\beta$ class
  - Architecture (A), orientation of sec. Str.
  - Topolgy (T), topological connections &
  - Homologous Superfamily (H), similar str and functions.

DALI/FSSP Database

- Completely automated; 3724 domains
- Criteria of compactness & recurrence
- Each domain is assigned a Domain Classification number DC_l_m_n_p representing fold space attractor region (l), globular folding topology (m), functional family (n) and sequence family (p).

Structural Alignment

- What is structural alignment of proteins?
  - 3-d superimposition of the atoms as "best as possible", i.e., to minimize RMSD (root mean square deviation).
  - Can be done using VAST and SARF
- Structural similarity is common, even among proteins that do not share sequence similarity or evolutionary relationship.

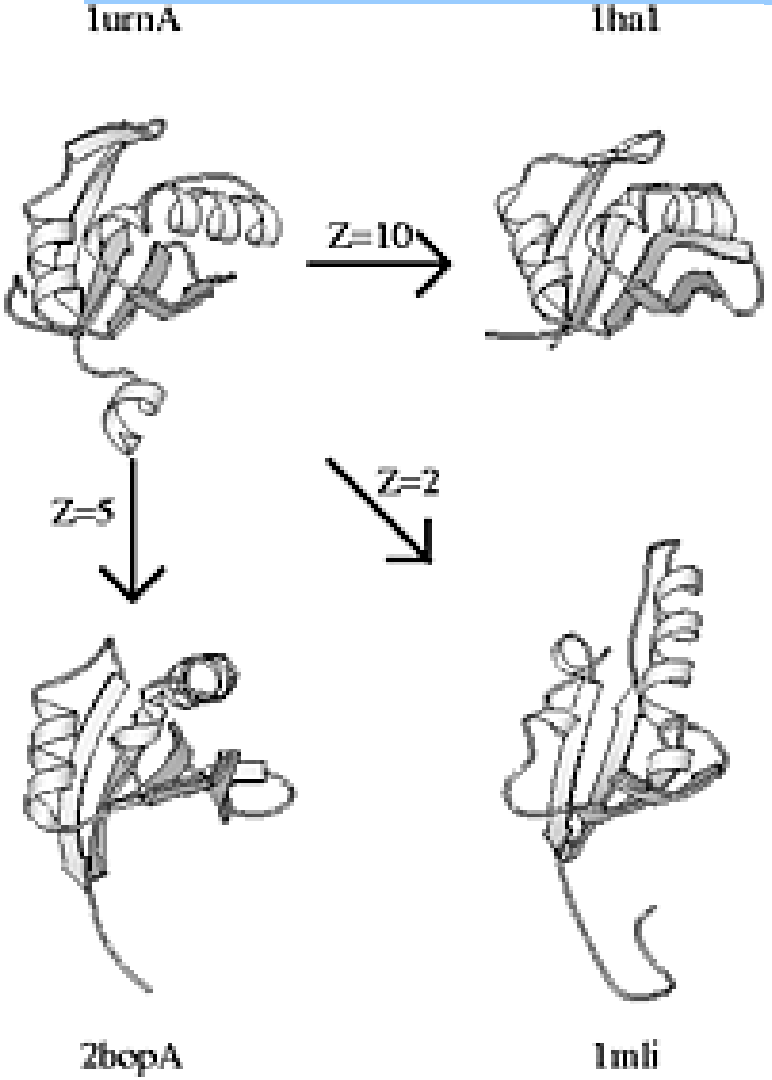
Other databases & tools

- MMDB contains groups of structurally related proteins
- SARF structurally similar proteins using secondary structure elements
- VAST Structure Neighbors
- SSAP uses double dynamic programming to structurally align proteins

5 Fold Space classes

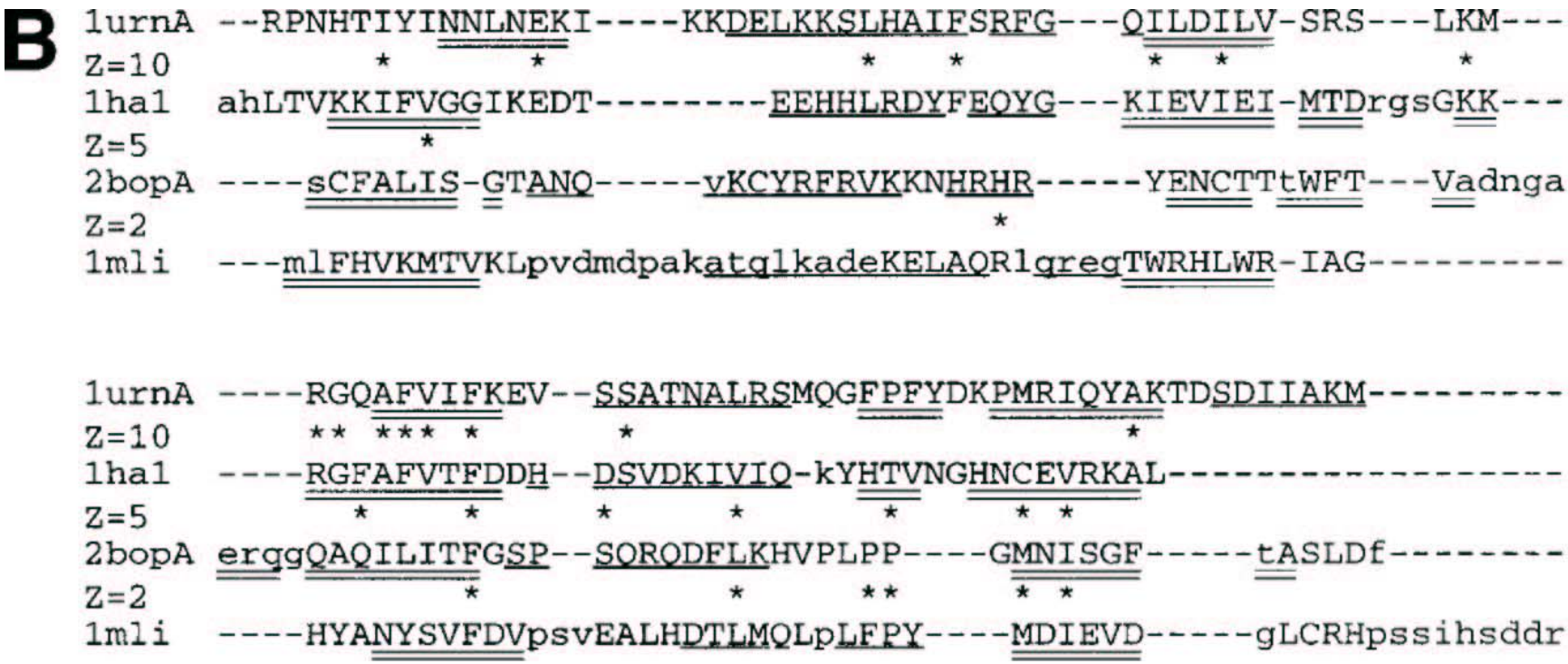
Attractor 1 can be characterized as alpha/beta,
attractor 2 as all-beta, attractor 3 as all-alpha,
attractor 5 as alpha-beta meander (1mli), and
attractor 4 contains antiparallel beta-barrels e.g. OB-fold (1prtF).

Fold Types & Neighbors

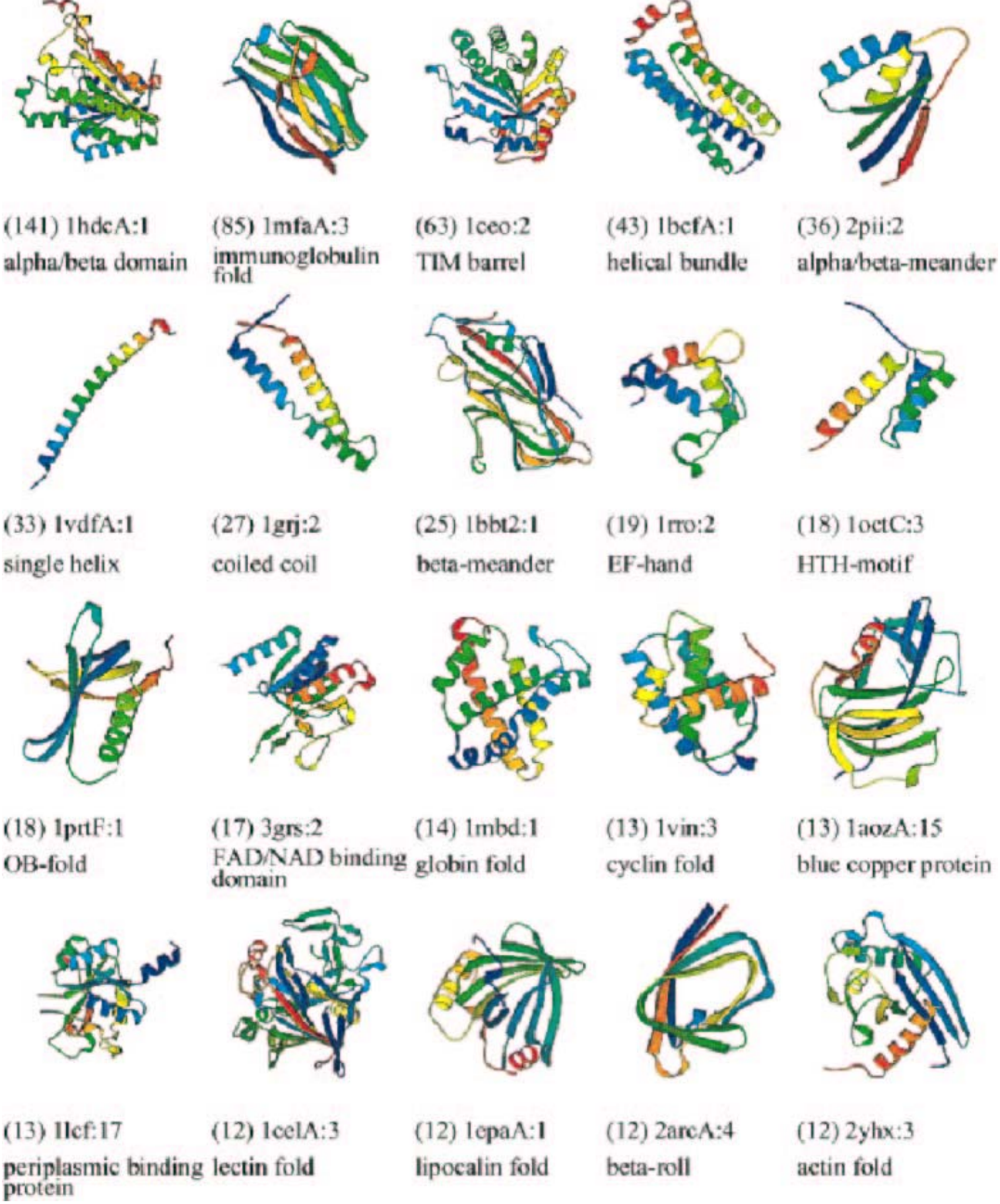


Structural neighbours of 1urnA (top left). 1mli (bottom right) has the same topology even though there are shifts in the relative orientation of secondary structure elements.

Sequence Alignment of Fold Neighbors



Frequent Fold Types



Protein Structure Prediction

- Holy Grail of bioinformatics
- Protein Structure Initiative to determine a set of protein structures that span protein structure space sufficiently well. WHY?
  - Number of folds in natural proteins is limited. Thus a newly discovered proteins should be within modeling distance of some protein in set.
- CASP: Critical Assessment of techniques for structure prediction
  - To stimulate work in this difficult field

PSP Methods

- homology-based modeling
- methods based on fold recognition
  - Threading methods
- *ab initio* methods
  - From first principles
  - With the help of databases

ROSETTA

- Best method for PSP
- As proteins fold, a large number of partially folded, low-energy conformations are formed, and that local structures combine to form more global structures with minimum energy.
- Build a database of known structures (I-sites) of short sequences (3-15 residues).
- Monte Carlo simulation assembling possible substructures and computing energy

Threading Methods

- See p471, Mount
  - **http://www.bioinformaticsonline.org/links/ch_10_t_7.html**

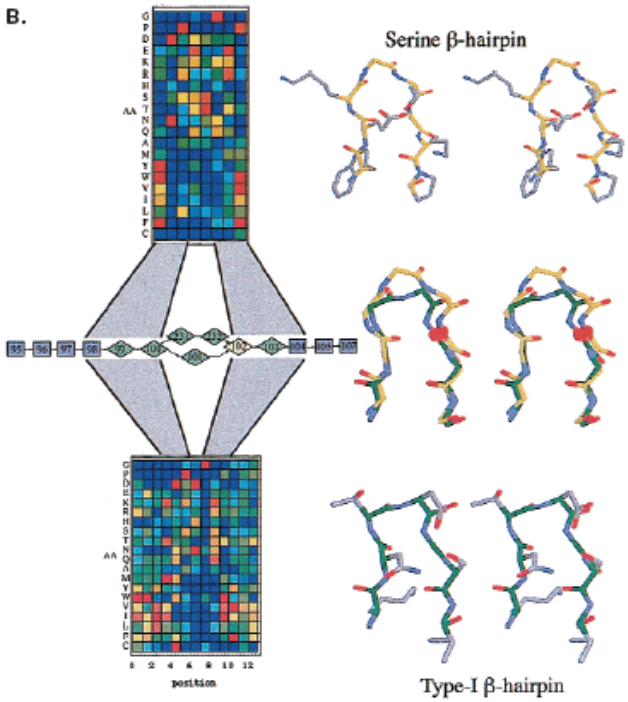


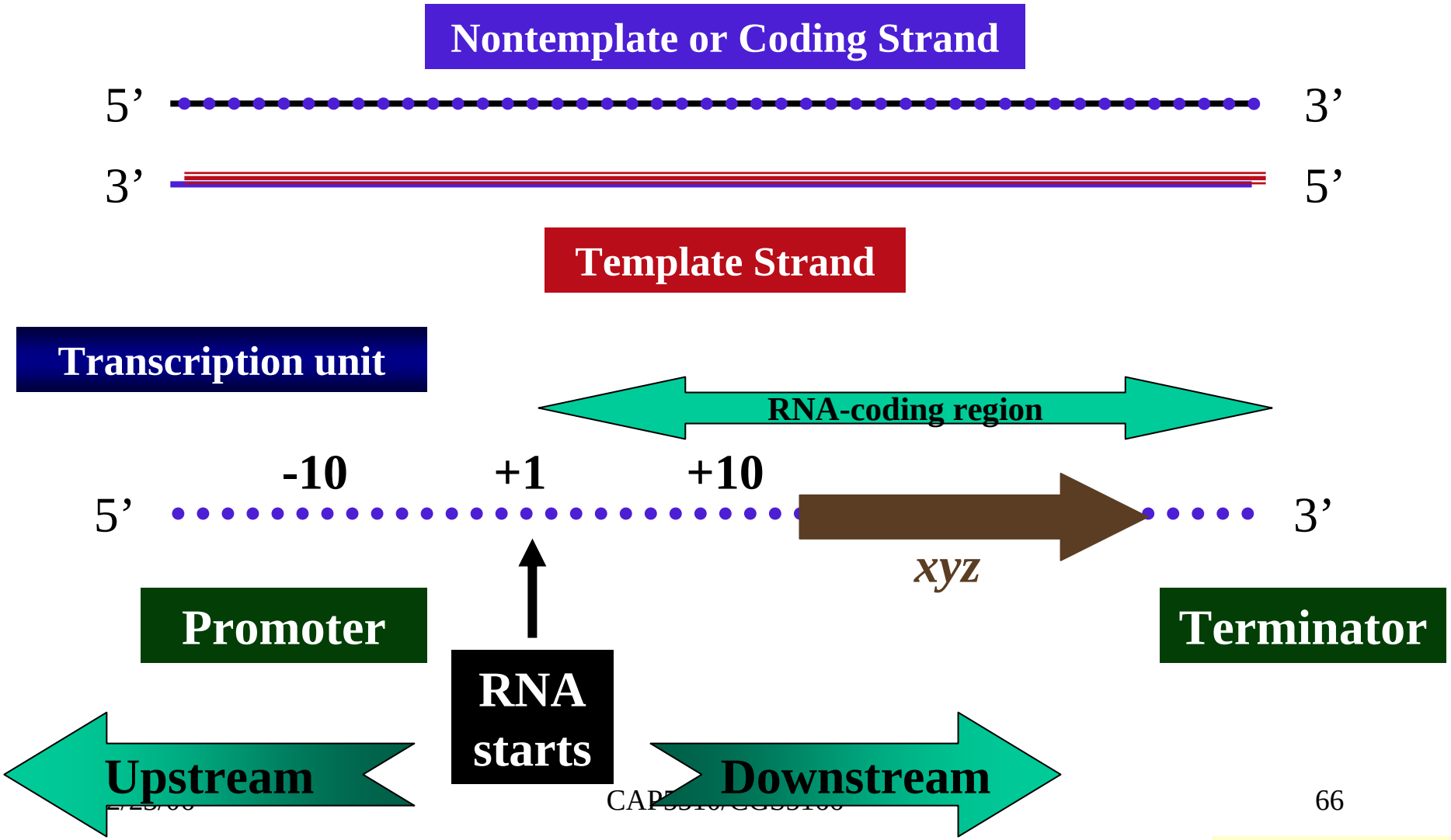
FIGURE 10.30. A hidden Markov model (discrete state-space model) of protein three-dimensional structure. (*B*) HMM called HMMSTR based on I-sites, 3- to 15-amino-acid patterns that are associated with three-dimensional structural features. The two matrices with colored squares represent alignment of sets of patterns that are found to be associated with a structure, in this case the hairpin turns shown on the right. Each column in the table corresponds to the amino acid variation found for one structural position in one of the turns. (*Blue* side chains) Conserved nonpolar residues; (*green*) conserved polar residues; (*red*) conserved proline; and (*orange*) conserved glycine. The two hairpins are aligned structurally in the middle structure on the right and the observed variation in the corresponding amino acid positions is represented by the HMM between the matrices on the left. The HMM represents an alignment of the two hairpin structural motifs in three-dimensional space and an alignment of the sequences. A short mismatch in the turn is represented by splitting the model into two branches. The shaped icons represent states, each of which represents a structure and a sequence position. Each state contains probability distributions about the sequence and structural attributes of a single position in the motif, including the probability of observing a particular amino acid, secondary structure, Φ-Ψ backbone angles, and structural context, e.g., location of β strand in a β sheet. Rectangles are predominantly β-strand states, and diamonds are predominantly turns. The color of the icon indicates a sequence preference as follows: (*blue*) hydrophobic; (*green*) polar; and (*yellow*) glycine. Numbers in icons are arbitrary identification numbers for the HMM states. There is a transition probability of moving from each state in the model to the next, as in HMMs that represent msa's. This model is a small component of the main HMMSTR model that represents a merging of the entire I-sites library. Three different models, designated λ^D, λ^C, and λ^R, are included in HMMSTR, which differ in details as to how the alignment of the I-sites was obtained to design the branching patterns (topology) of the model and which structural data were used to train the model. HMMSTR may be used for a variety of different predictions, including secondary structure prediction, structural context prediction, and Φ-Ψ dihedral angle prediction. Predictions are made by aligning the model with a sequence, finding if there is a high-scoring alignment, and deciphering the highest-scoring path through the model. The HMMSTR program may be downloaded or used on a server that can be readily located by a Web search. (*B*, reprinted, with permission, from Bystroff et al. 2000 [©2000 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

Nomenclature

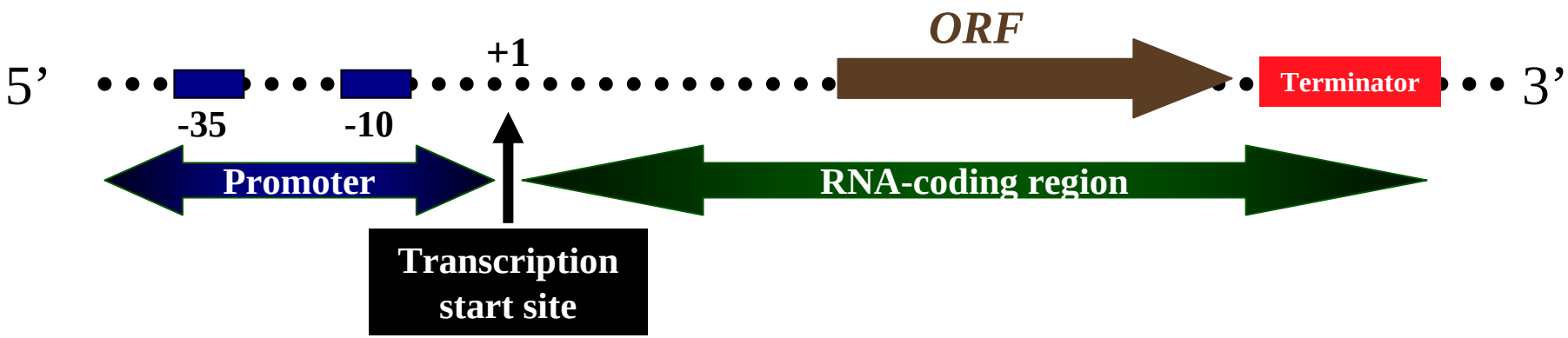
RNA Polymerization occurs 5’ to 3’



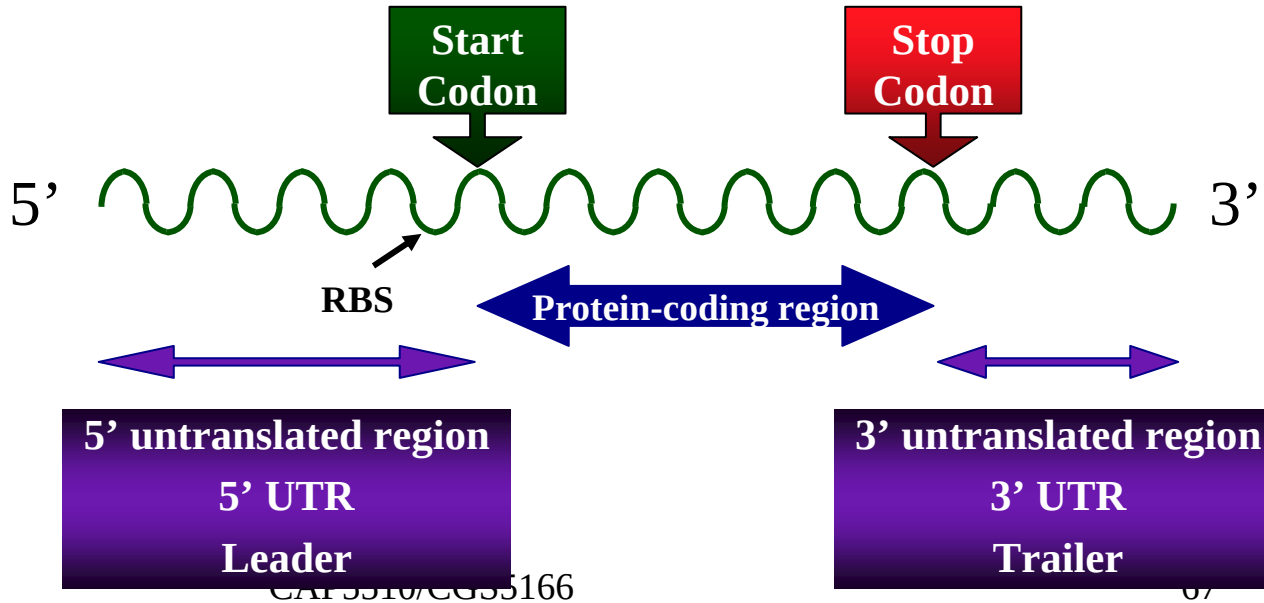
Slide courtesy Prof. Mathee

Transcriptional unit and single gene mature mRNA

Transcriptional unit



mRNA



RBS

Ribosome binding site

Slide courtesy Prof. Mathee

Prokaryotic Gene Characteristics

DNA PATTERNS IN THE *E. coli lexA* GENE

	GENE SEQUENCE	PATTERN
1	GAATTCGATAAATCTCTGGTTTATTGTGCAGTTTATGGTT	CTGNNNNNNNNNNCAG
	TT	TTGACA
41	CCAAAATCGCCTTTTGCTGTATATACTCACAGCATAACTG	CTGNNNNNNNNNNCAG
	CCAA -35 -10 TATACT >	TATAAT, > mRNA start
81	TATATACACCCAGGGGGCGGAATGAAAGCGTTAACGGCCA	CTGNNNNNNNNNNCAG
	+10 GGGGG Ribosomal binding site	GGAGG
121	GGCAACAAGAGGTGTTTGATCTCATCCGTGATCACATCAG	
161	CCAGACAGGTATGCCGCCGACGCGTGCGGAAATCGCGCAG	ATG
201	CGTTTGGGGTTCCGTTCCCCAAACGCGGCTGAAGAACATC	
241	TGAAGGCGCTGGCACGCAAAGGCGTTATTGAAATTGTTTC	
281	CGGCGCATCACGCGGGATTCGTCTGTTGCAGGAAGAGGAA	
321	GAAGGGTTGCCGCTGGTAGGTCGTGTGGCTGCCGGTGAAC	
361	CACTTCTGGCGCAACAGCATATTGAAGGTCATTATCAGGT	OPEN READING FRAME
401	CGATCCTTCCTTATTCAAGCCGAATGCTGATTTCCTGCTG	
441	CGCGTCAGCGGGATGTCGATGAAAGATATCGGCATTATGG	
481	ATGGTGACTTGCTGGCAGTGCATAAAACTCAGGATGTACG	
521	TAACGGTCAGGTCGTTGTCGCACGTATTGATGACGAAGTT	
561	[illegible]	
601	TGTTGCCAGAAAATAGCGAGTTTAAACCAATTGTCGTTGA	
641	CCTTCGTCAGCAGAGCTTCACCATTGAAGGGCTGGCGGTT	
681	GGGGTTATTCGCAACGGCGACTGGCTGTAACATATCTCTG	TAA
721	AGACCGCGATGCCGCCTGGCGTCGCGGTTTGTTTTTCATC	
761	TCTCTTCATCAGGCTTGTCTGCATGGCATTCCTCACTTCA	
801	TCTGATAAAGCACTCTGGCATCTCGCCTTACCCATGATTT	
841	TCTCCAATATCACCGTTCCGTTGCTGGGACTGGTCGATAC	
881	GGCGGTAATTGGTCATCTTGATAGCCCGGTTTATTTGGGC	
921	GGCGTGGCGGTTGGCGCAACGGCGGACCAGCT	

Shown are matches to approximate consensus binding sites for LexA repressor (CTGNNNNNNNNNNCAG), the -10 amd -35 promoter regions relative to the start of the mRNA (TTGACA and TATAAT), the ribosomal binding site on the mRNA (GGAGG), and the open reading frame (ATG...TAA). Only the second two of the predicted LexA binding sites actually bind the repressor.

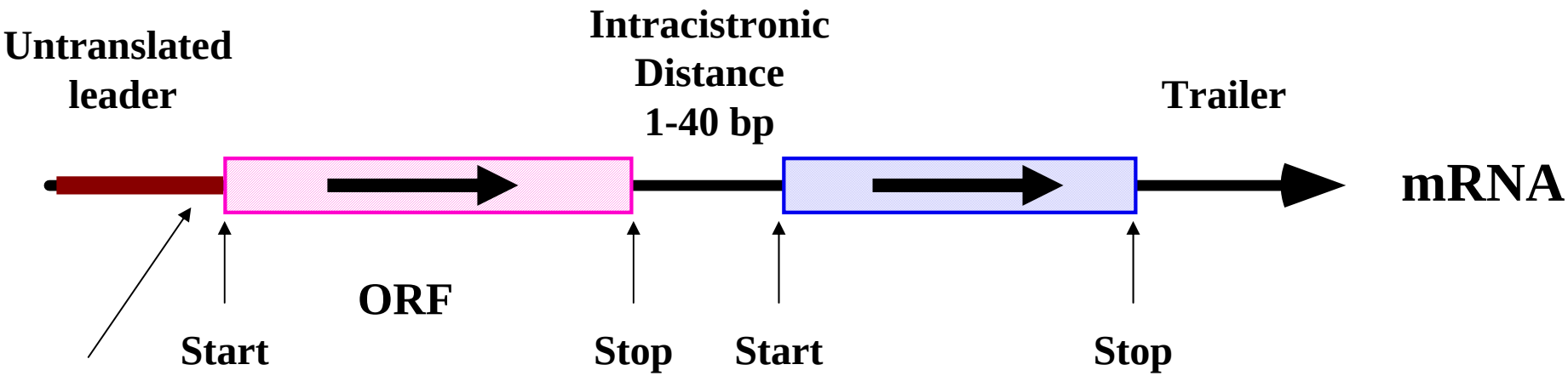
FIGURE 9.6. The promoter and open reading frame of the *E. coli lexA* gene.

Messenger RNA or mRNA

Initation Codon	AUG	Methionine

Termination Codons	
UAA	Ochre
UAG	Amber
UGA	Opal

Others:		
	GUG	Valine
	UUG	Leucine
	AUU	Isoleucine



RBS
Ribosome Binding Site
Shine-Dalgarno Sequence

7 bp upstream of start codon
5'--AGGAGG--3'

Coding region
Open Reading Frame (ORF)

Reading frame is one of three possible ways of reading a nucleotide sequence as a series of triplets.

Slide courtesy Prof. Mathee

Start and Stop Codon Distribution

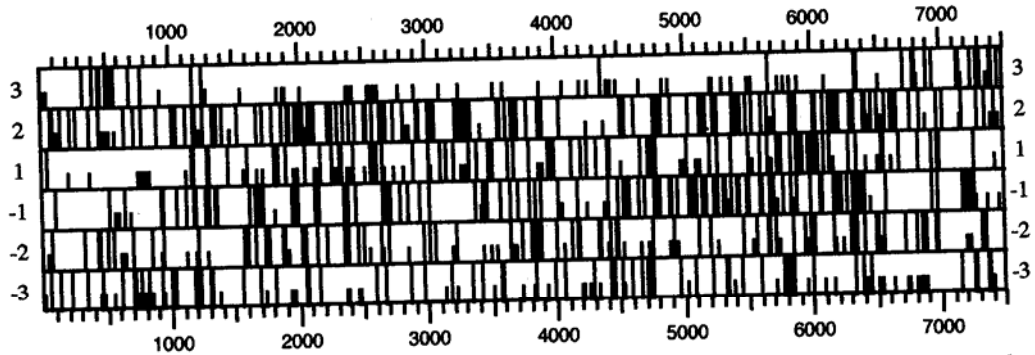
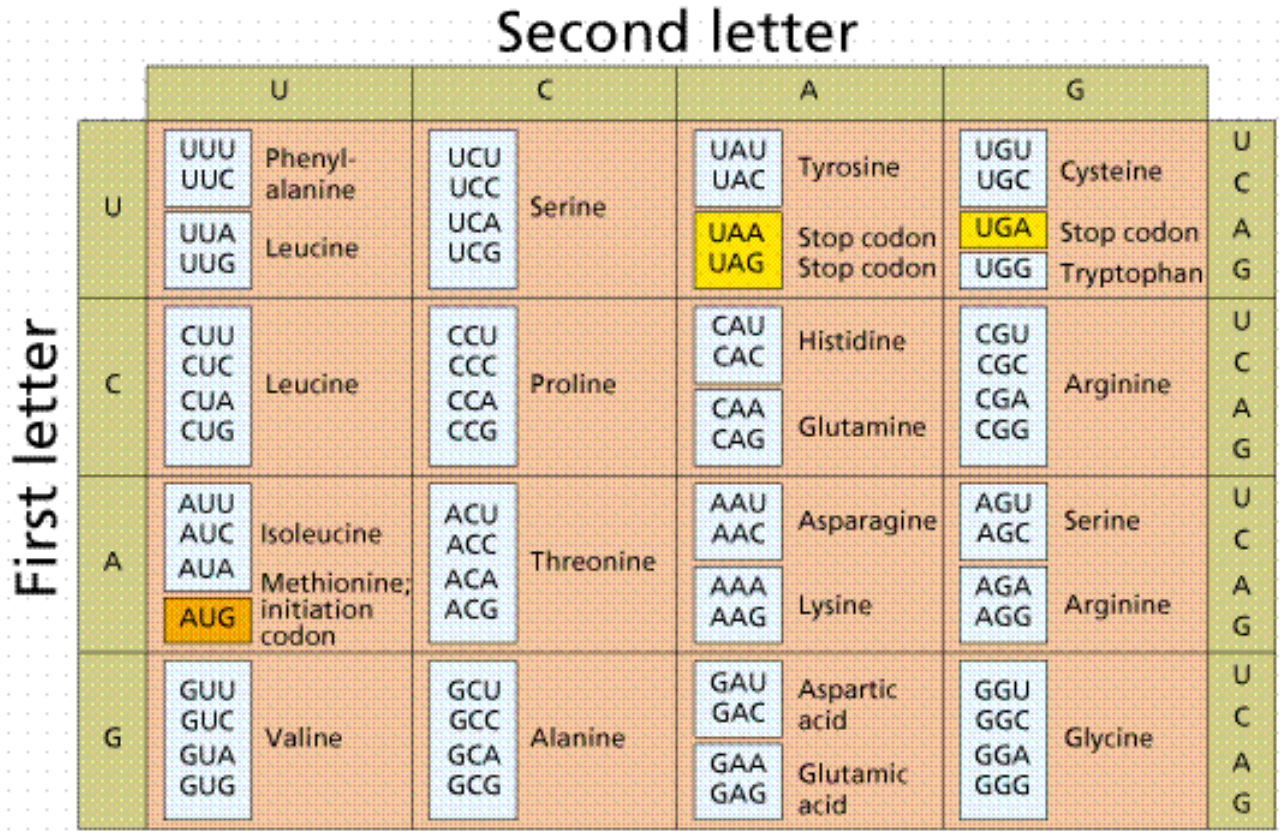


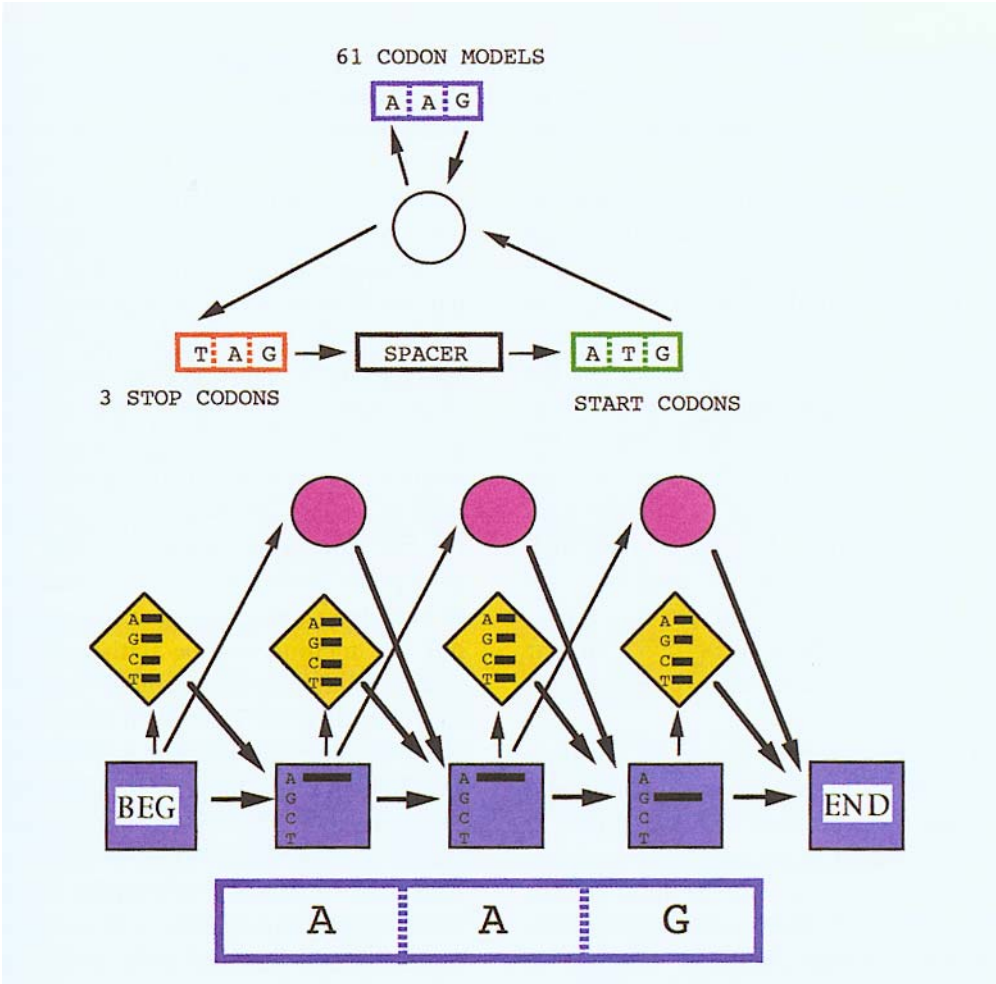
FIGURE 9.1. ORF map of a portion of the *E. coli lac* operon using the DNA STRIDER program (Marck 1988). Shown are AUG and termination codons as one-half and full vertical bars, respectively, in all six possible reading frames. The *lacZ* gene is visible as an ORF that runs from positions 1284 to 4355 in frame 3.

Genetic Code



Slide courtesy http://www.emc.maricopa.edu/faculty/farabee/BIOBK/code.gif

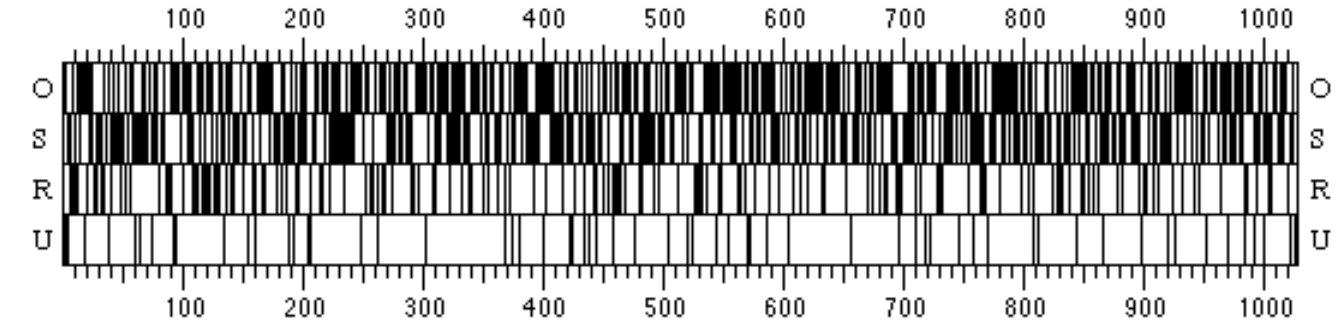
Recognizing Codons



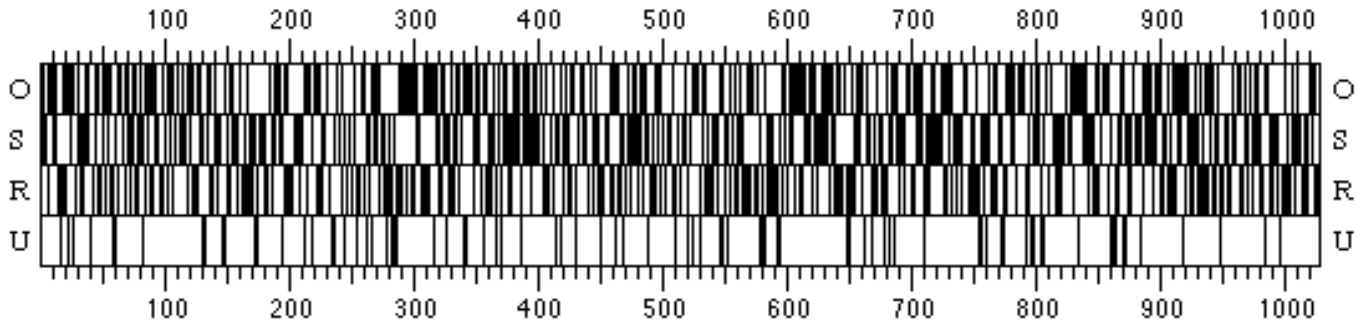
Codon Bias

- Some codons preferred over others.

O = optimal
S = suboptimal
R = rare
U = unfavorable



Frame Shift 1

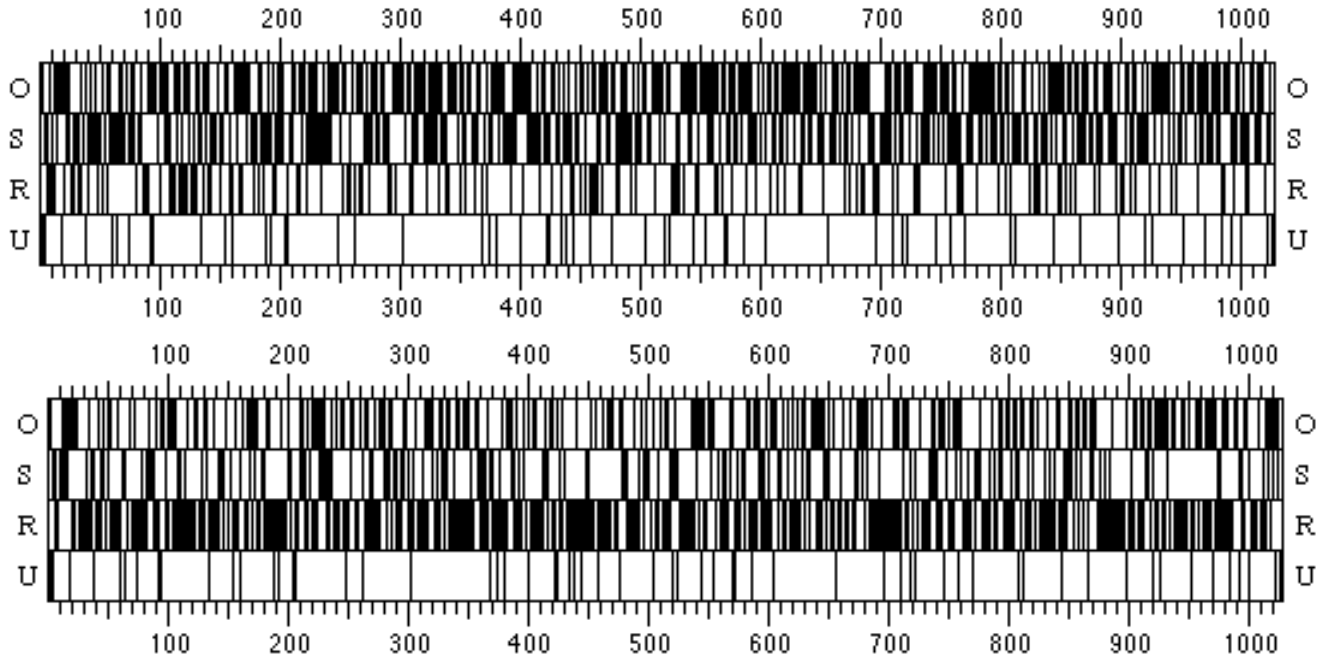


Frame Shift 2

Codon Bias

- Codon biases specific to organisms

O = optimal
S = suboptimal
R = rare
U = unfavorable



Same Frames; Different labeling of codon types (i.e., from yeast)

Eukaryotic Gene Prediction

- Complicated by introns & alternative splicing
- Exons/introns have different GC content.
- Many other measures distinguish exons/introns
- Software:
  - GENEPARSER Snyder & Stormo (NN)
  - GENIE Kulp, Haussler, Reese, Eckman (HMM)

Introns/Exons in *C. elegans*

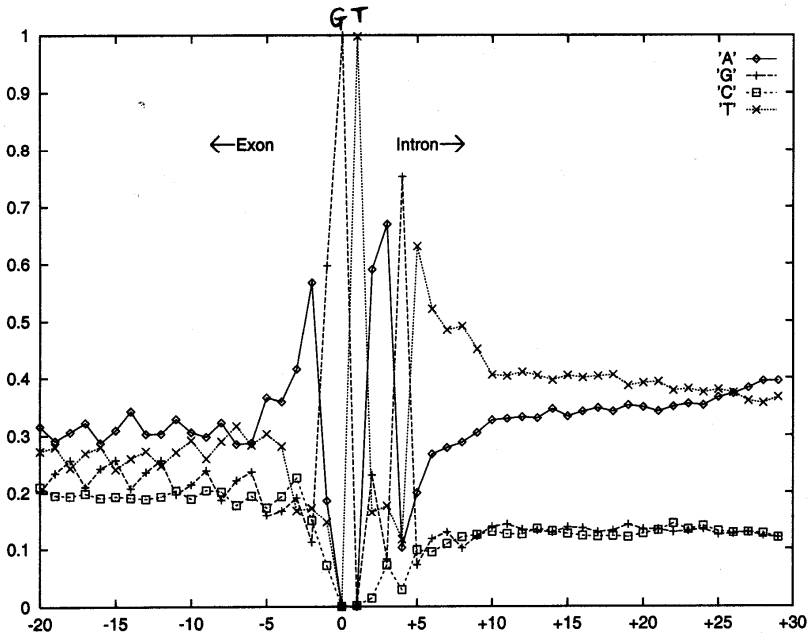


Figure 2: Profile of the same 5' collection but around a larger window.

A/T

G/C

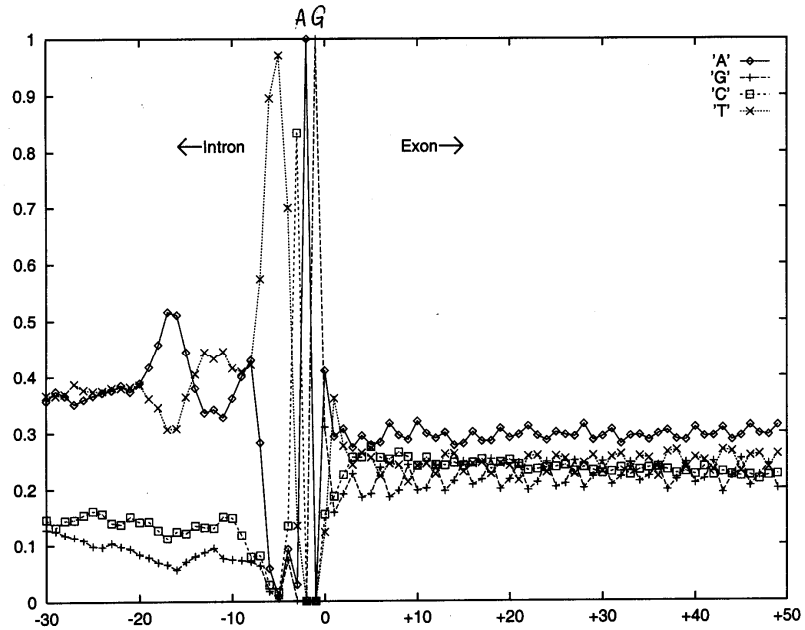
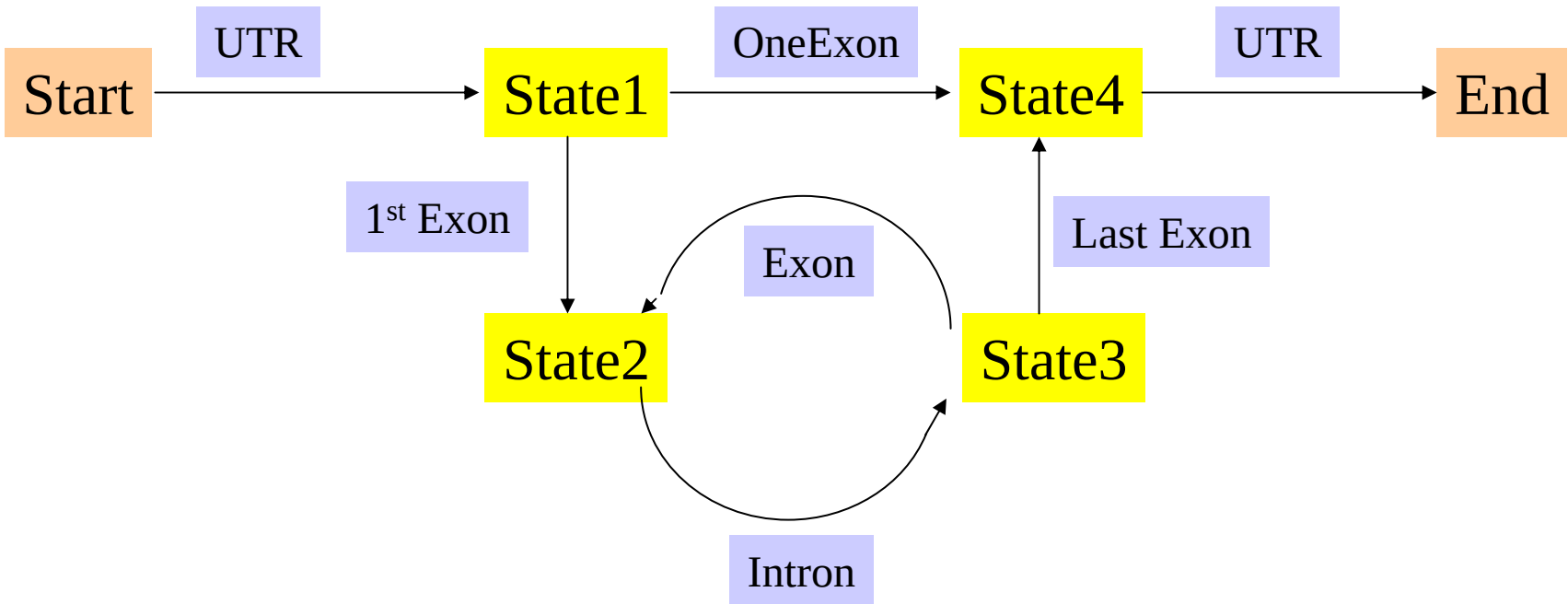


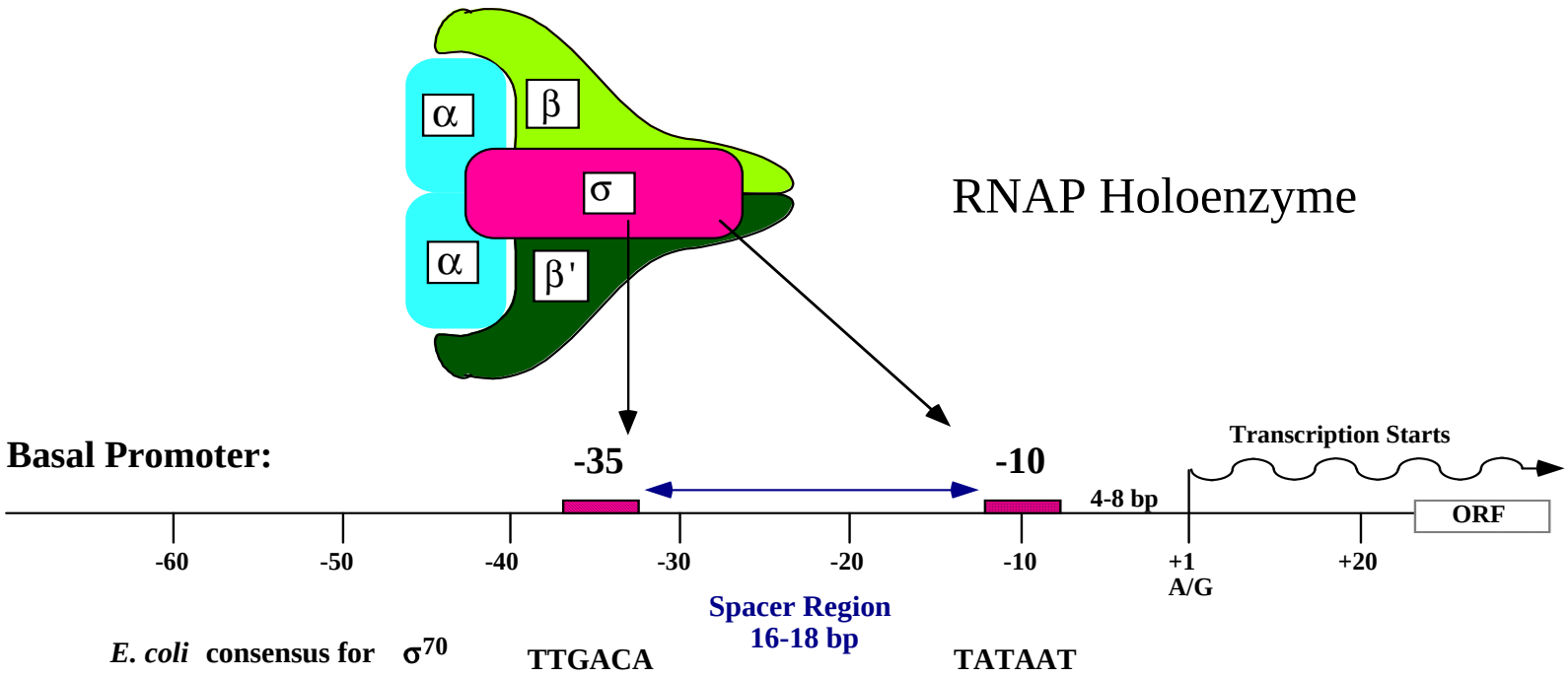
Figure 4: Profile of 8,192 sequences of length 80 around the 3' site. The first position in the exon is labeled 0.

- 8192 Introns in *C. elegans* : **[GT...AG]**
- Vary in lengths from 30 to over 600; Complexity varies

HMM structure for Gene Finding



Transcriptional machinery: RNA Polymerase and DNA



Slide courtesy Prof. Mathee