# BSC 4934: Q'BIC Capstone Workshop


Giri Narasimhan

ECS 254A; Phone: x3748

giri@cs.fiu.edu


http://www.cs.fiu.edu/~giri/teach/BSC4934_Su11.html

July 2011

# BLAST Parameters and Output

- Type of sequence, nucleotide/protein
- Word size, $w$
- Gap penalties, $p_1$ and $p_2$
- Neighborhood Threshold Score, $T$
- Score Threshold, $S$
- E-value Cutoff, $E$
- Number of hits to display, $H$
- Database to search, $D$
- Scoring Matrix, $M$
- Score $s$ and E-value $e$
  - E-value $e$ is the expected number of sequences that would have an alignment score greater than the current score $s$.

# BLAST algorithm: Phase 1

Phase 1: get list of word pairs (w=3) above threshold T

Example: for a human RBP query

…FS**GTW**YA…

**GTW** is a word in this query sequence

A list of words (w=3) is:

**FSG SGT GTW TWY WYA**

**YSG TGT ATW SWY WFA**

**FTG SVT GSW TWF WYS**



Fig. 4.11
page 116

# Phase 1: Find list of similar words

- Find list of words of length w (here w = 3) and distance at least T (here T = 11)
  - GTW 22
  - GSW 18
  - ATW 16
  - NTW 16
  - GTY 13
  - GNW 10
  - GAW 9

# Use BLOSUM to score word hits

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | | | | | | | | | | | | | | | | | | |
| R | -1 | 5 | | | | | | | | | | | | | | | | | |
| N | -2 | 0 | 6 | | | | | | | | | | | | | | | | |
| D | -2 | -2 | 1 | 6 | | | | | | | | | | | | | | | |
| C | 0 | -3 | -3 | -3 | 9 | | | | | | | | | | | | | | |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | | | | | | | | | | | | | |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | | | | | | | | | | | | |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | | | | | | | | | | | |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | | | | | | | | | | |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | | | | | | | | | |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | | | | | | | | |
| K | -1 | 2 | 0 | -1 | -1 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | | | | | | | |
| M | -1 | -2 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | | | | | | |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | | | | | |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | | | | |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | | | |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | | |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |
| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |

# BLAST: Phases 2 & 3

- Phase 2: Scan database for exact hits of similar words list and find HotSpots
- Phase 3:
  - Extend good hit in either direction.
  - Keep track of the score (use a scoring matrix)
  - Stop when the score drops below some cutoff.

```
KENFDKARFSGTWYAMAKKDPEG 50   RBP (query)
MKGLDIQKVAGTWYSLAMAASD. 44   lactoglobulin (hit)
```

# BLAST: Threshold vs # Hits & Extensions

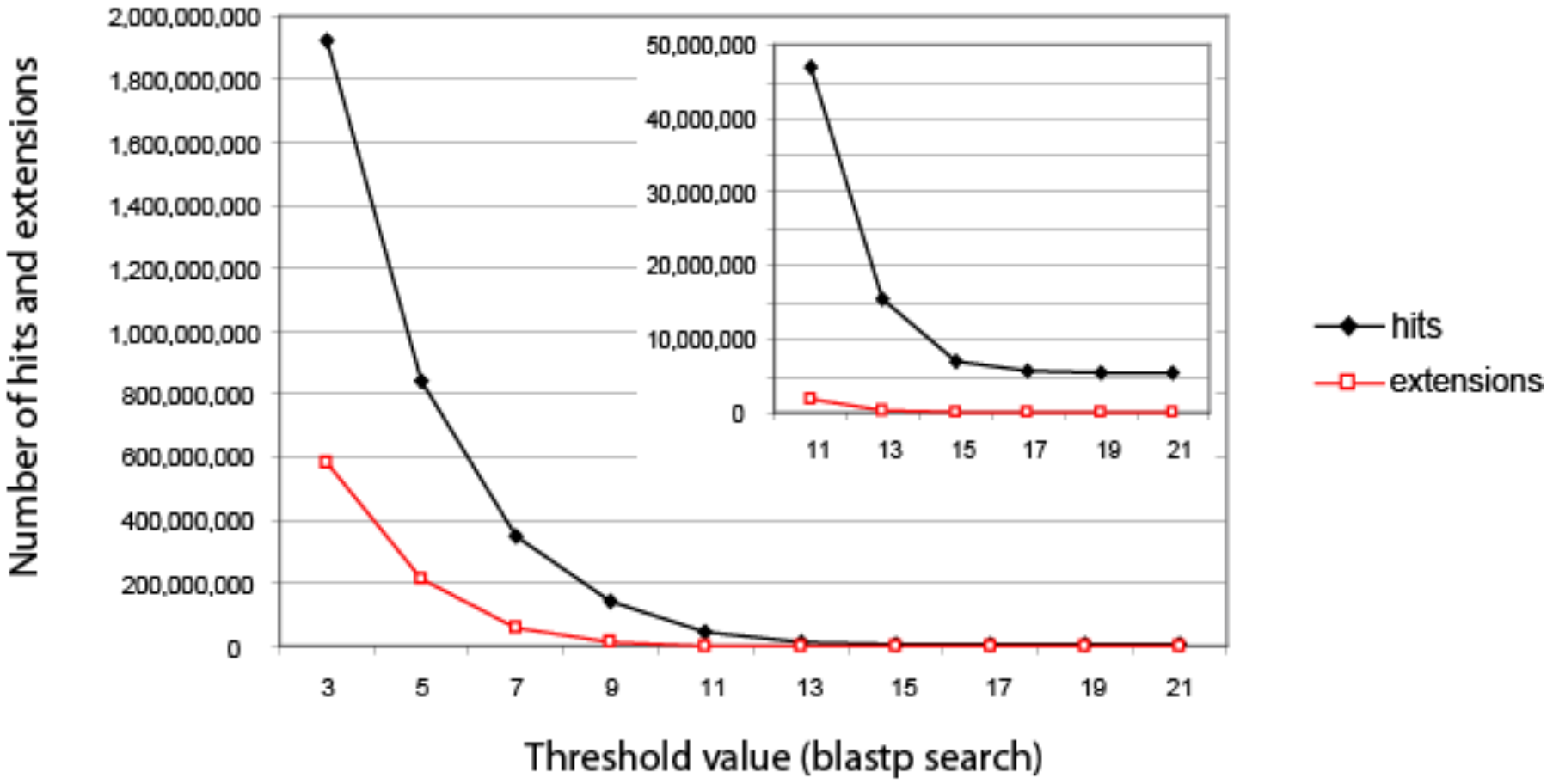


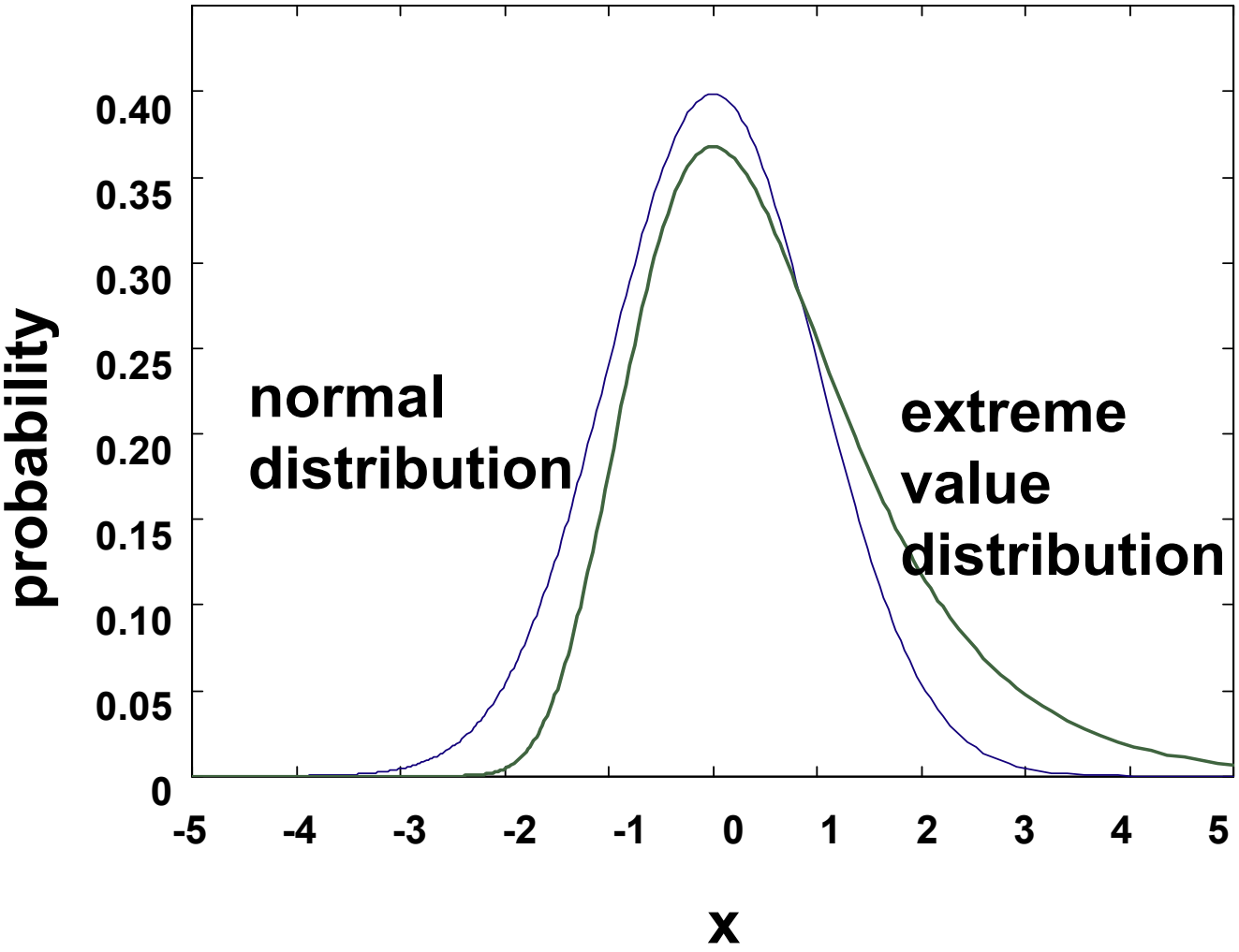



Fig. 4.12
page 118

# Word Size

- Blastn: w = 7, 11, or 15.
  - w=15 gives fewer matches and is faster than w=11 or w=7.
- Megablast: w = 28 to 64.
  - Megablast is VERY fast for finding closely related DNA sequences!

# Scores: Follow Extreme Value Distribution




$E = Kmn\, e^{-\lambda S}$

m,n = seq length
S = Raw Score
K ≈ Search space

$S' = (\lambda S - \ln K) / \ln 2$

S' = Bit Score

$p = 1 - e^{-E}$

p = p-value

# E-value versus P-value

| E-value | P-value |
|---|---|
| 10 | 0.9999546 |
| 5 | 0.99326205 |
| 2 | 0.86466472 |
| 1 | 0.63212056 |
| 0.1 | 0.09516258 |
| 0.05 | 0.04877058 |
| 0.001 | 0.00099950 |
| 0.0001 | 0.0001 |

**E-values are easier to interpret;**

**If query is short aa sequence, then use very large E-value;**

**Sometimes even meaningful hits have large E-values.**

# Assessing whether proteins are homologous

```
>gi|4505583|ref|NP_002562.1| progestagen-associated endometrial protein (placental protein 14,
           pregnancy-associated endometrial alpha-2-globulin, alpha
           uterine protein); Progestagen-associated endometrial
           protein (placental protein 14) [Homo sapiens]
 gi|190215|gb|AAA60147.1| (J04129) placental protein 14 [Homo sapiens]
          Length = 162

 Score = 32.0 bits (71), Expect = 0.49
 Identities = 26/107 (24%), Positives = 48/107 (44%), Gaps = 11/107 (10%)

Query: 26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD- 84
           + K++ +  + +GTW++MA      + L   + A   V  T  +          +L+ W+
Sbjct: 5   QTKQDLELPKLAGTWHSMAMAT-NNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 63

Query: 85  -VCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTY 130
             C +       T +P KFK+ Y  VA       ++  ++DTDYD +
Sbjct: 64  NSCVEKKVLGEKTGNPKKFKINY-TVA-------NEATLLDTDYDNF 102
```

RBP4 and PAEP:
Low bit score, E value 0.49, 24% identity (“twilight zone”).
But they are indeed homologous. Try a BLAST search with PAEP as a query, and find many other lipocalins.

# Difficulties with BLAST

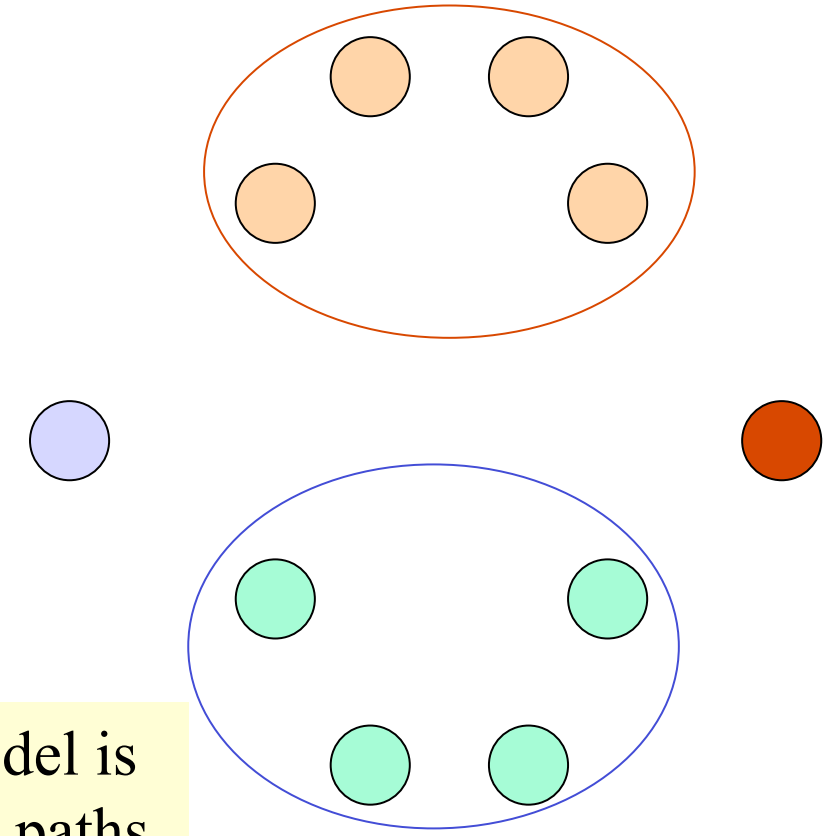
- Use human beta globin as a query against human RefSeq proteins, and blastp does not "find" human myoglobin. This is because the two proteins are too distantly related. PSI-BLAST at NCBI as well as hidden Markov models easily solve this problem.
- How can we search using 10,000 base pairs as a query, or even millions of base pairs? Many BLAST-like tools for genomic DNA are available such as PatternHunter, Megablast, BLAT, and BLASTZ.

# Related Tools

- Megablast
  - For long, closely-related sequences
  - Uses large w and is very fast
- BLAT
  - UCSC tool
  - DB broken into words; query is searched
- PatternHunter
  - Generalized seeds used instead of words
- BLASTZ, Lagan, SSAHA

# Hidden Markov Model (HMM)

- States
- Transitions
- Transition Probabilities
- Emissions
- Emission Probabilities



- What is hidden about HMMs?

Answer: The path through the model is hidden since there are many valid paths.

# Profile Method

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

| Location in Seq. | Sequence 1 | 2 | 3 | 4 | 5 | 6 | 7 | Protein Name |
|---|---|---|---|---|---|---|---|---|
| 14 | G | V | S | A | S | A | V | Ka RbtR |
| 32 | G | V | S | E | M | T | I | Ec DeoR |
| 33 | G | V | S | P | G | T | I | Ec RpoD |
| 76 | G | A | G | I | A | T | I | Ec TrpR |
| 178 | G | C | S | R | E | T | V | Ec CAP |
| 205 | C | L | S | P | S | R | L | Ec AraC |
| 210 | C | L | S | P | S | R | L | St AraC |
| 13 | G | V | N | K | E | T | I | Br MerR |

FREQUENCY TABLE

| | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 2 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 5 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |

7

# Profile Method

FREQUENCY TABLE

| | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 2 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 5 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |

WEIGHT MATRIX

| | A | C | E | G | I | K | L | M | N | P | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 108 | 0 | 101 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 21 | 78 | 0 | 0 | 0 | 0 | 44 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 23 | 0 | 0 | 0 | 0 | 46 | 0 | 0 | 102 |
| 4 | 21 | 0 | 32 | 0 | 38 | 32 | 0 | 0 | 0 | 86 | 39 | 0 |
| 5 | 21 | 0 | 62 | 23 | 0 | 0 | 0 | 74 | 0 | 0 | 0 | 72 |
| 6 | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 69 | 0 |
| 7 | 0 | 0 | 0 | 0 | 98 | 0 | 44 | 0 | 0 | 0 | 0 | 0 |

$$Weight[i, AA] = \log\left(\frac{Freq[i,AA]}{p[AA] \cdot N}\right) \cdot 100$$

8

# Profile Method

WEIGHT MATRIX

| | A | C | E | G | I | K | L | M | N | P | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 108 | 0 | 101 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 21 | 78 | 0 | 0 | 0 | 0 | 44 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 23 | 0 | 0 | 0 | 0 | 46 | 0 | 0 | 102 |
| 4 | 21 | 0 | 32 | 0 | 38 | 32 | 0 | 0 | 0 | 86 | 39 | 0 |
| 5 | 21 | 0 | 62 | 23 | 0 | 0 | 0 | 74 | 0 | 0 | 0 | 72 |
| 6 | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 69 | 0 |
| 7 | 0 | 0 | 0 | 0 | 98 | 0 | 44 | 0 | 0 | 0 | 0 | 0 |

Given the following protein sequence:

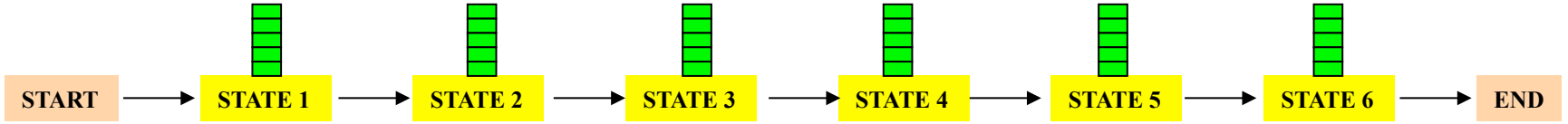
M T E D L F G D L Q D D T I L A H L D N
P A E D T S R F P A L L A E L N D L L R
G E L S R L G V D P A H S L E I V V A I
C K H L G G G Q V Y I P R G Q A L D S L
I R D L R I W N D F N G R N V S E L T T
R Y G V T F N T V Y K A I R R M R R L K

9

# Profile HMMs

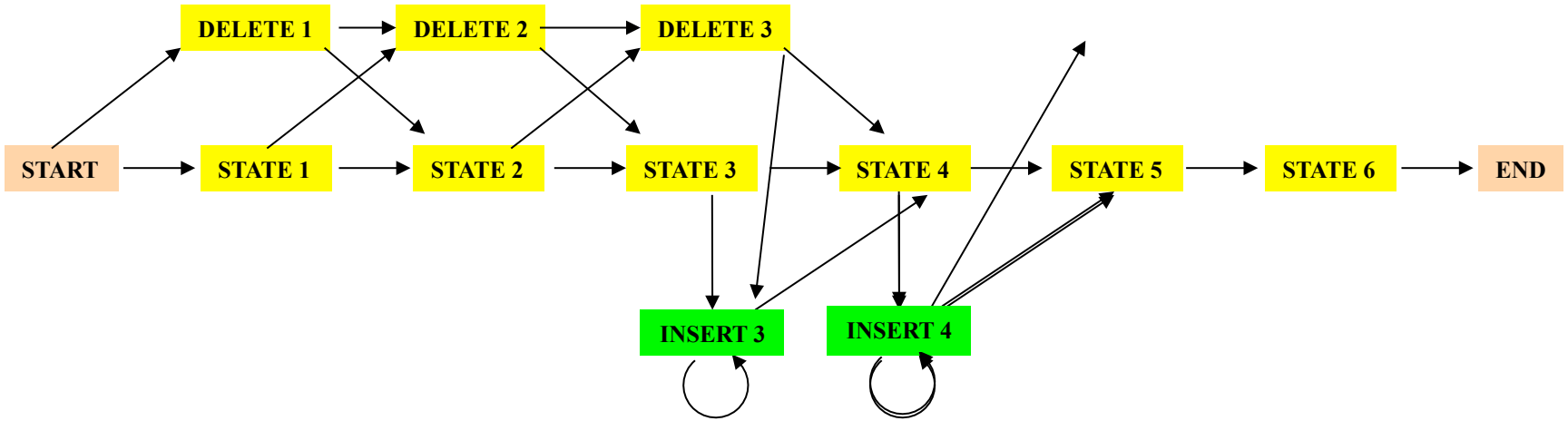
PROFILE METHOD, [M. Gribskov et al., '90]

| Location in Seq. | Sequence 1 2 3 4 5 6 | Protein Name |
|---|---|---|
| 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 |

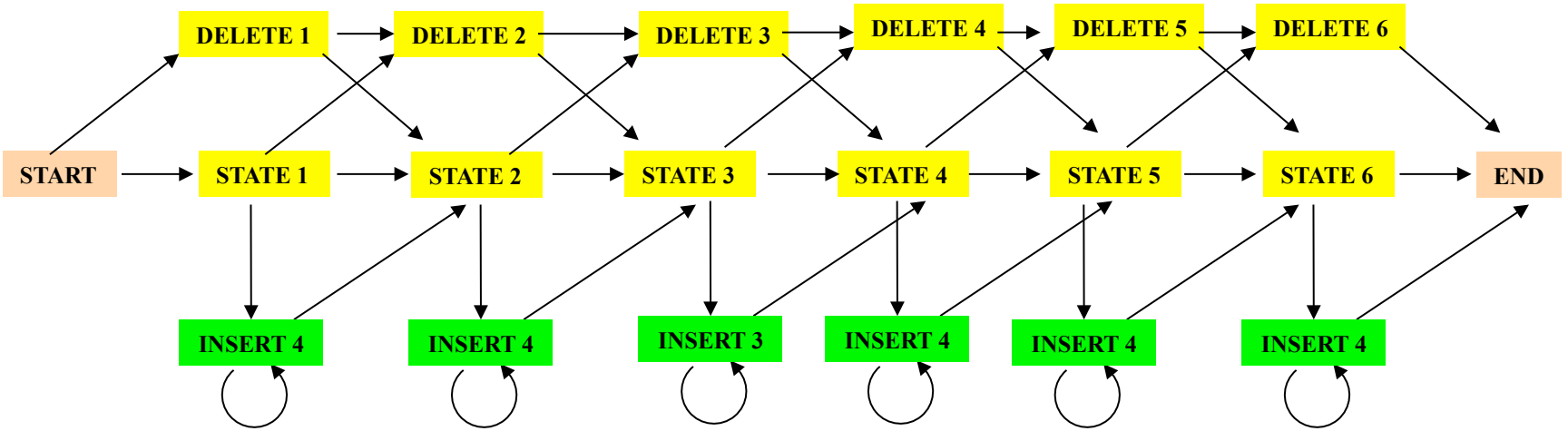


# Profile HMMs with InDels

- Insertions
- Deletions
- Insertions & Deletions



# Profile HMMs with InDels

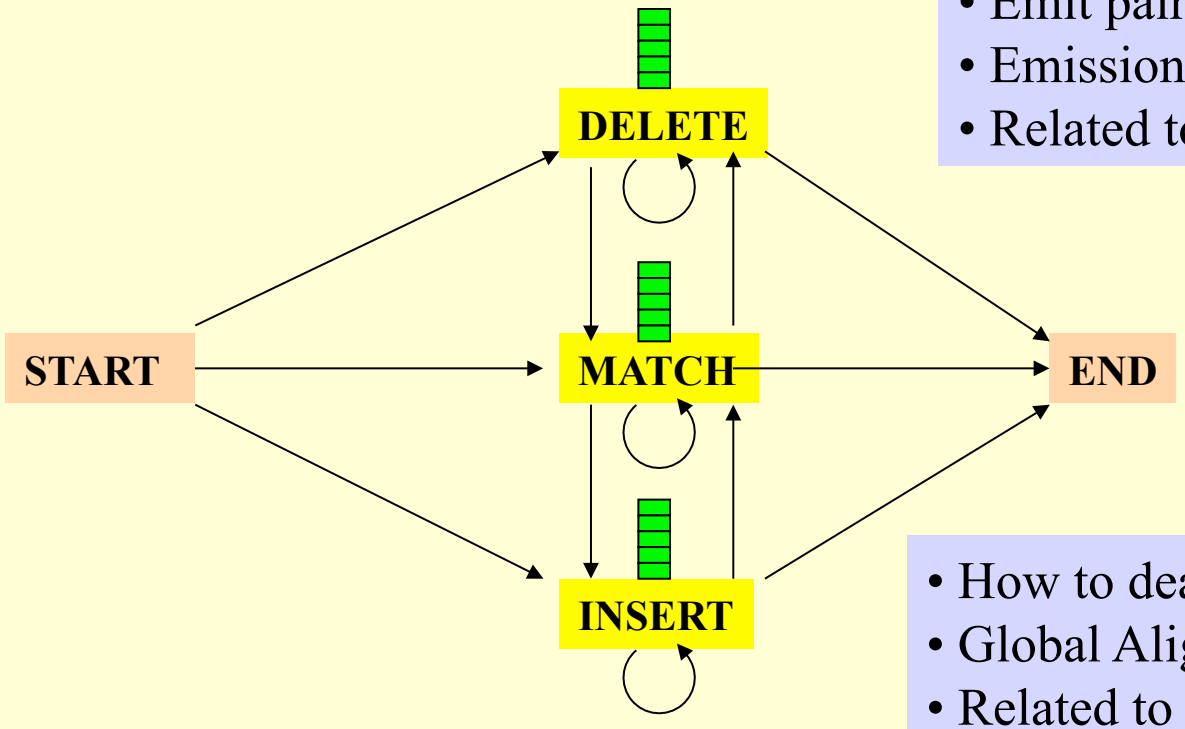



Missing transitions from DELETE j to INSERT j and from INSERT j to DELETE j+1.

# How to model Pairwise Sequence Alignment

**LEAPVE**

**LAPVIE**

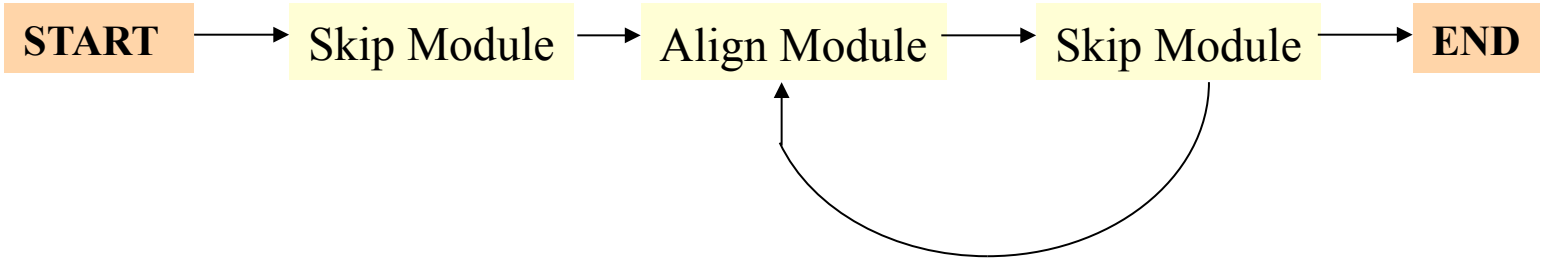



Pair HMMs
- Emit pairs of synbols
- Emission probs?
- Related to Sub. Matrices

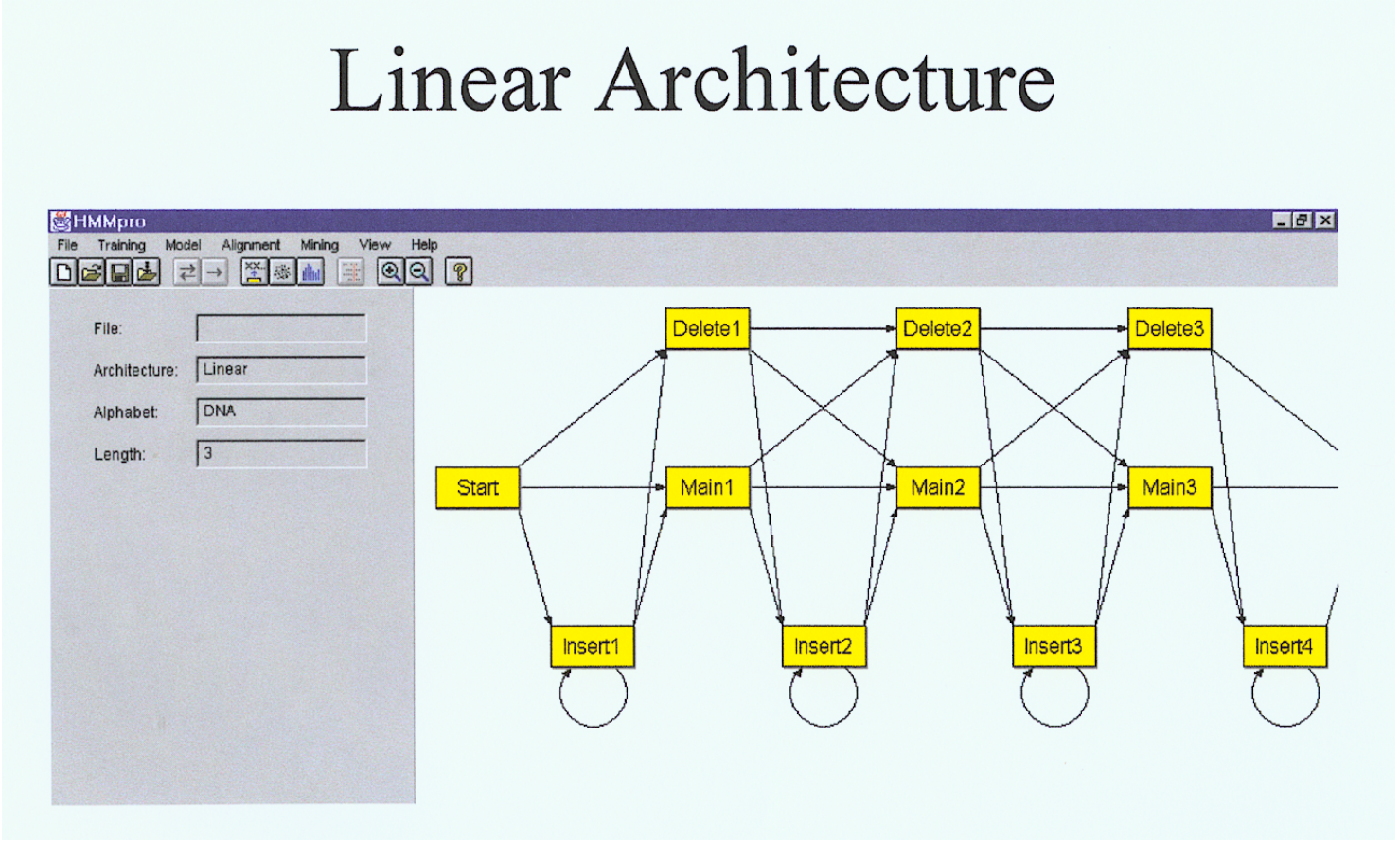
- How to deal with InDels?
- Global Alignment? Local?
- Related to Sub. Matrices

# How to model Pairwise Local Alignments?

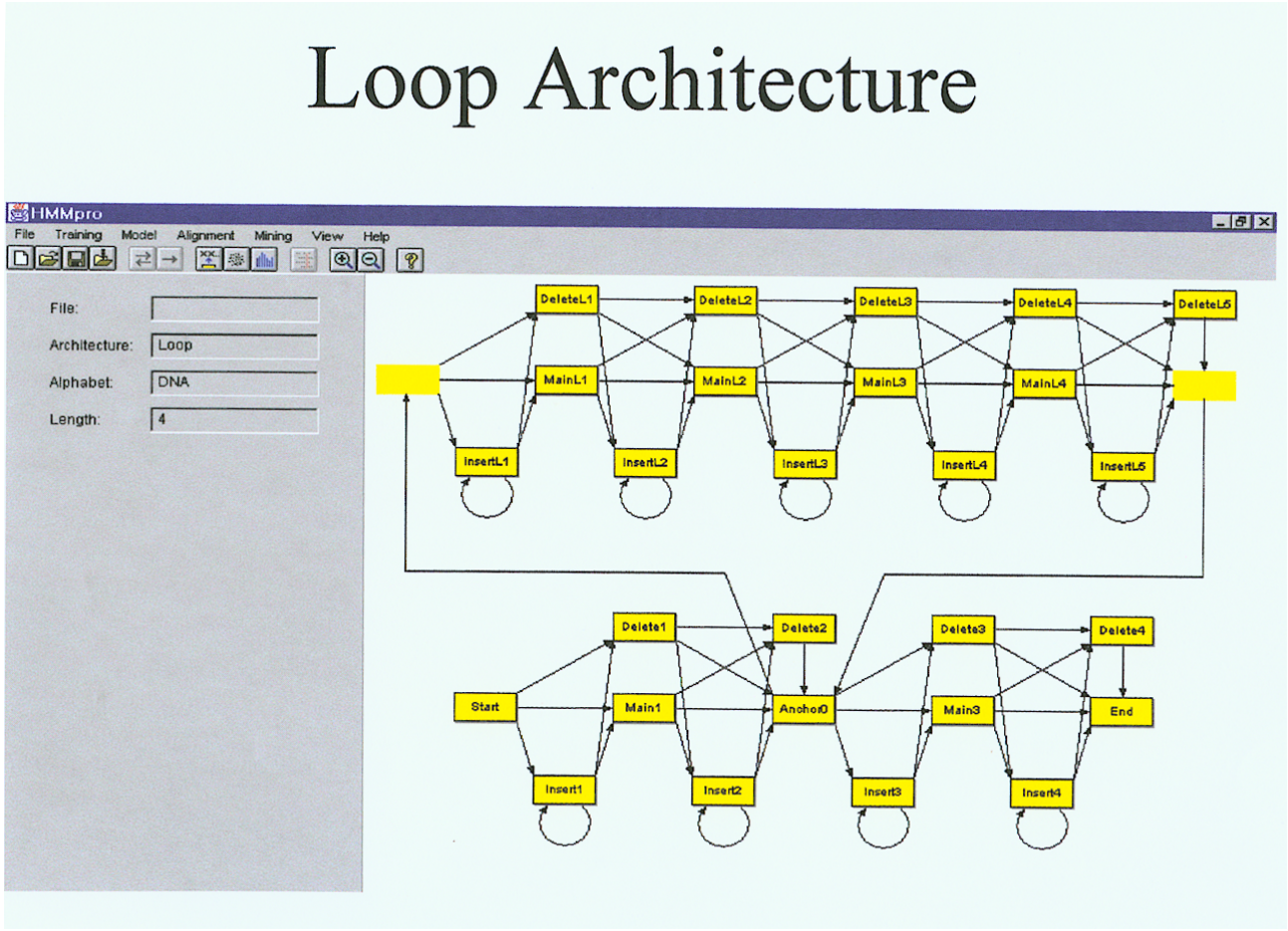

# How to model Pairwise Local Alignments with gaps?



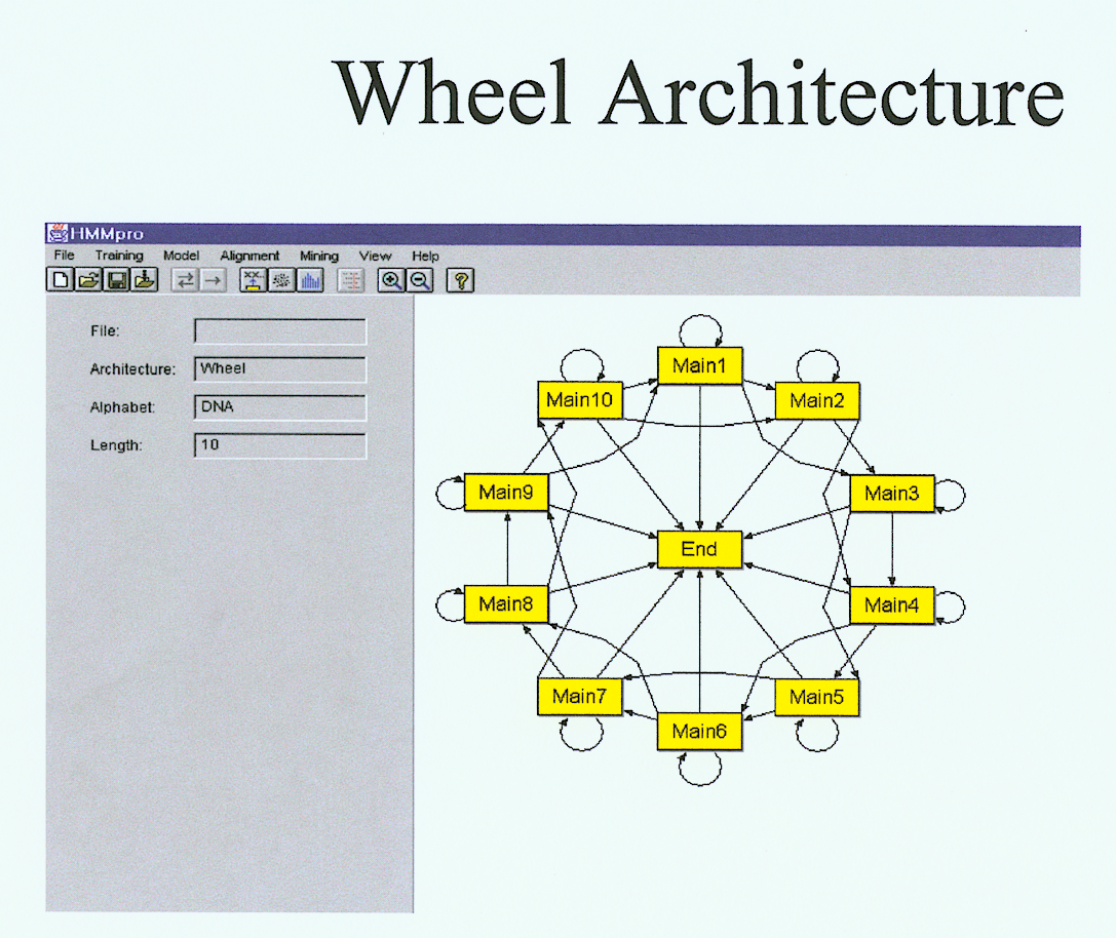
# Standard HMM architectures



# Standard HMM architectures



# Standard HMM architectures



# Profile HMMs from Multiple Alignments

```
HBA_HUMAN    VGA--HAGEY
HBB_HUMAN    V----NVDEV
MYG_PHYCA    VEA--DVAGH
GLB3_CHITP   VKG------D
GLB5_PETMA   VYS--TYETS
LGB2_LUPLU   FNA--NIPKH
GLB1_GLYDI   IAGADNGAGV
```

Construct Profile HMM from above multiple alignment.

# HMM for Sequence Alignment

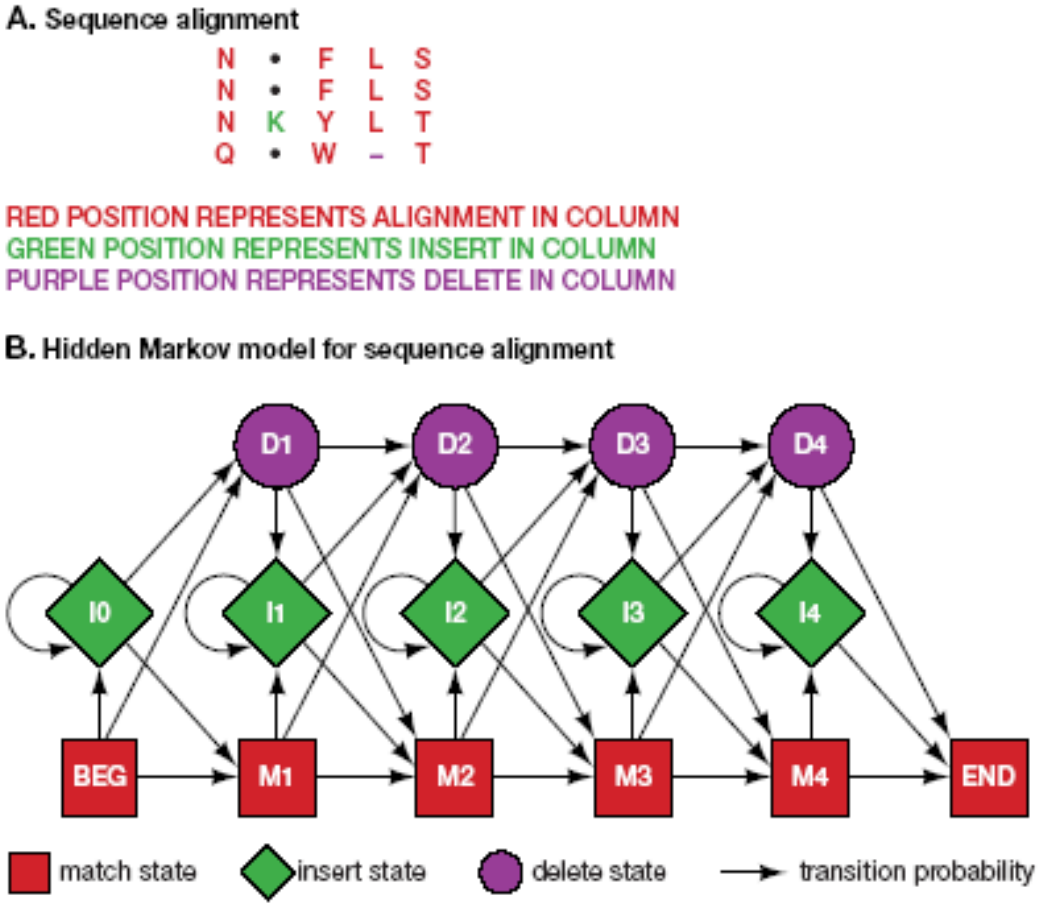



FIGURE 5.16. Relationship between the sequence alignment and the hidden Markov model of the alignment (Krogh et al. 1994). This particular form for the HMM was chosen to represent the sequence, structural, and functional variation expected in proteins. The model accommodates the identities, mismatches, insertions, and deletions expected in a group of related proteins. (*A*) A section of an msa. The illustration shows the columns generated in an msa. Each column may include matches and mismatches (*red* positions), insertions (*green* positions), and deletions (*purple* positions). (*B*) The HMM. Each column in the model represents the possibility of a match, insert, or delete in each column of the alignment in *A*. The HMM is a probabilistic representation of a section of the msa. Sequences can be generated from the HMM by starting at the beginning state labeled BEG and then by following any one of many pathways from one type of sequence variation to another (states) along the state transition arrows and terminating in the ending state labeled END. Any sequence can be generated by the model and each pathway has a probability associated with it. Each square match state stores an amino acid distribution such that the probability of finding an amino acid depends on

## Problem 3: LIKELIHOOD QUESTION

- Input: Sequence S, model M, state i
- Output: Compute the probability of reaching state i with sequence S using model M
  - Backward Algorithm (DP)

## Problem 4: LIKELIHOOD QUESTION

- Input: Sequence S, model M
- Output: Compute the probability that S was emitted by model M
  - Forward Algorithm (DP)

## Problem 5: LEARNING QUESTION

- Input: model structure M, Training Sequence $S$
- Output: Compute the parameters $\Theta$
- Criteria: ML criterion
    - maximize $P(S \mid M, \Theta)$ HOW???

## Problem 6: DESIGN QUESTION

- Input: Training Sequence $S$
- Output: Choose model structure M, and compute the parameters $\Theta$
    - No reasonable solution
    - Standard models to pick from

# Iterative Solution to the LEARNING QUESTION (Problem 5)

- Pick initial values for parameters $\Theta_0$
- Repeat

  Run training set $S$ on model $M$

  Count # of times transition $i \Rightarrow j$ is made

  Count # of times letter x is emitted from state $i$

  Update parameters $\Theta$
- Until (some stopping condition)

# Entropy

- Entropy measures the variability observed in given data.

$$E = -\sum_{c} p_c \log p_c$$

- Entropy is useful in multiple alignments & profiles.

- Entropy is max when uncertainty is max.

# HMM for Sequence Alignment

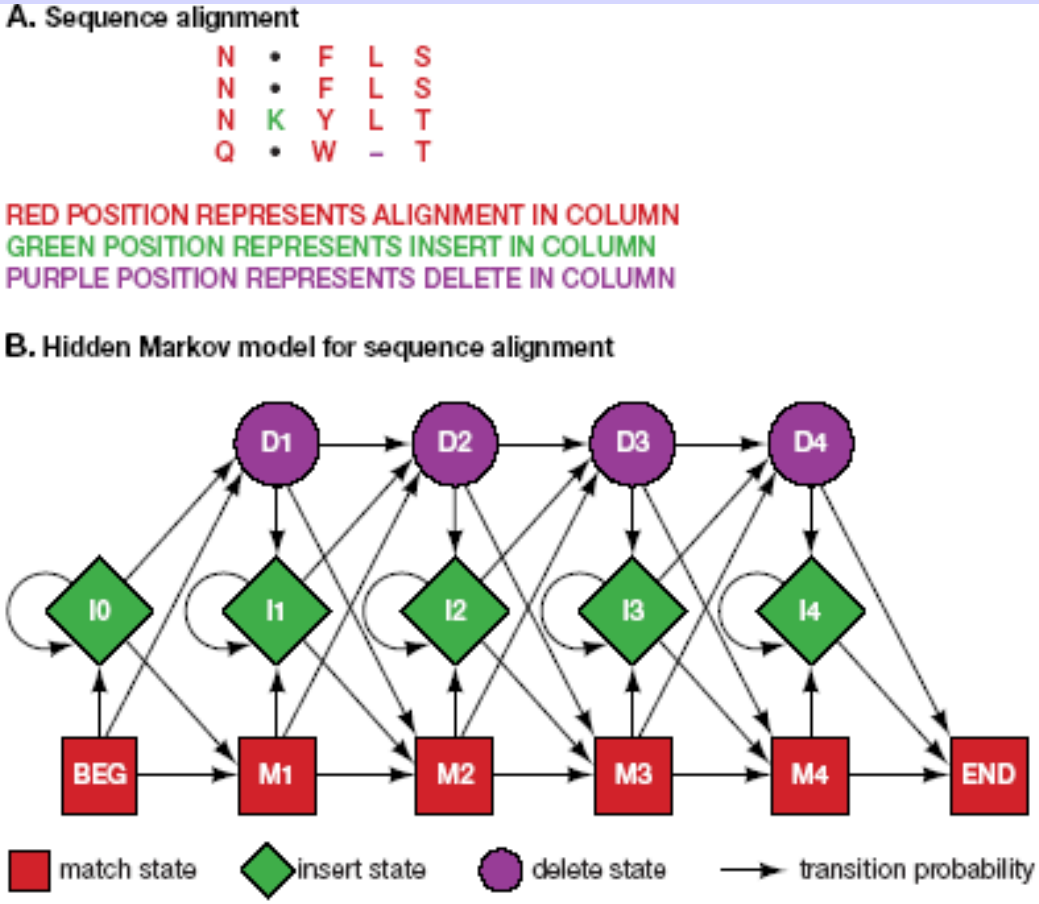



FIGURE 5.16. Relationship between the sequence alignment and the hidden Markov model of the alignment (Krogh et al. 1994). This particular form for the HMM was chosen to represent the sequence, structural, and functional variation expected in proteins. The model accommodates the identities, mismatches, insertions, and deletions expected in a group of related proteins. (*A*) A section of an msa. The illustration shows the columns generated in an msa. Each column may include matches and mismatches (*red* positions), insertions (*green* positions), and deletions (*purple* positions). (*B*) The HMM. Each column in the model represents the possibility of a match, insert, or delete in each column of the alignment in *A*. The HMM is a probabilistic representation of a section of the msa. Sequences can be generated from the HMM by starting at the beginning state labeled BEG and then by following any one of many pathways from one type of sequence variation to another (states) along the state transition arrows and terminating in the ending state labeled END. Any sequence can be generated by the model and each pathway has a probability associated with it. Each square match state stores an amino acid distribution such that the probability of finding an amino acid depends on

# G-Protein Couple Receptors

- Transmembrane proteins with 7 $\alpha$-helices and 6 loops; many subfamilies
- Highly variable: 200-1200 aa in length, some have only 20% identity.
- [Baldi & Chauvin, '94] HMM for GPCRs
- HMM constructed with 430 match states (avg length of sequences) ; Training: with 142 sequences, 12 iterations

# GPCR - Analysis

- Compute main state entropy values

$$H_i = -\sum_a e_{ia} \log e_{ia}$$

- For every sequence from test set (142) & random set (1600) & all SWISS-PROT proteins
  - Compute the negative log of probability of the most probable path $\pi$

$$Score(S) = -\log\left(P(\pi \mid S, M)\right)$$

# Entropy

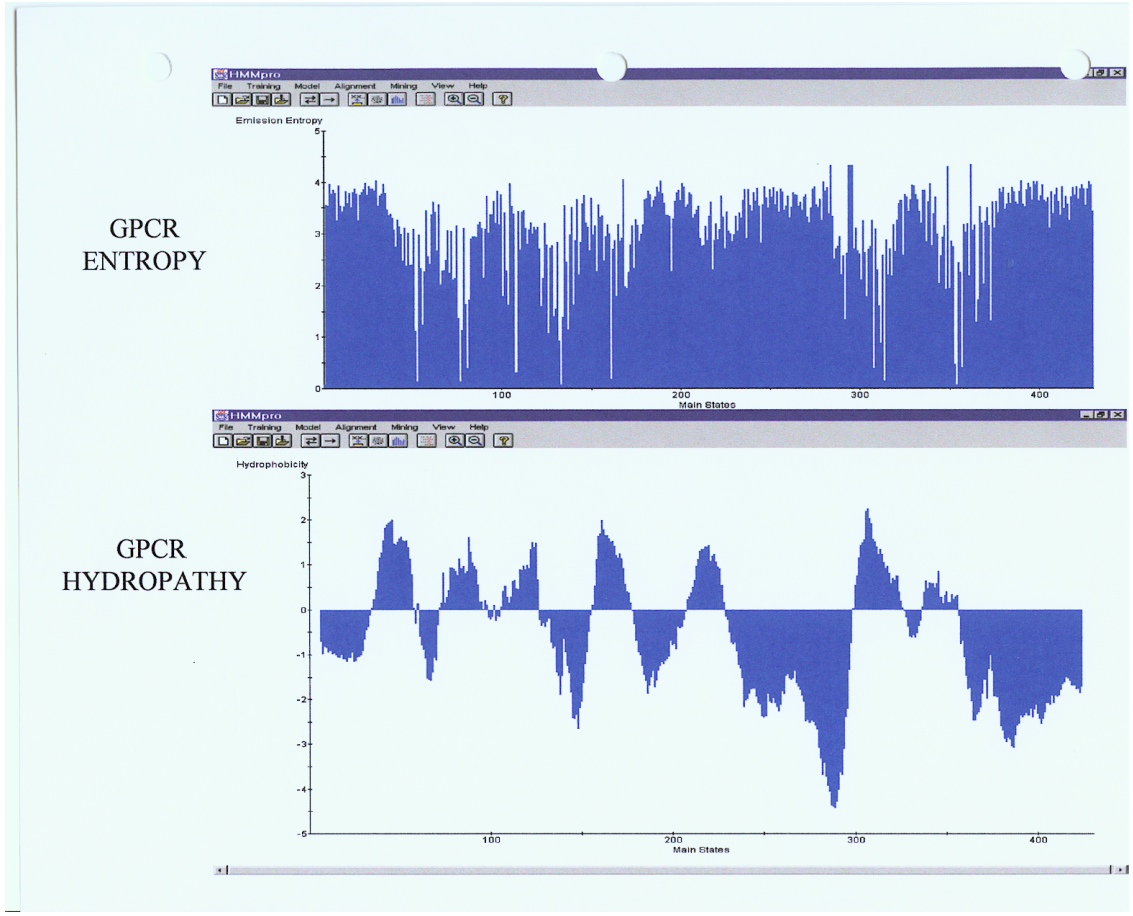
- Entropy measures the variability observed in given data.

$$E = -\sum_{c} p_c \log p_c$$

- Entropy is useful in multiple alignments & profiles.

- Entropy is max when uncertainty is max.

# GPCR Analysis



# Entropy

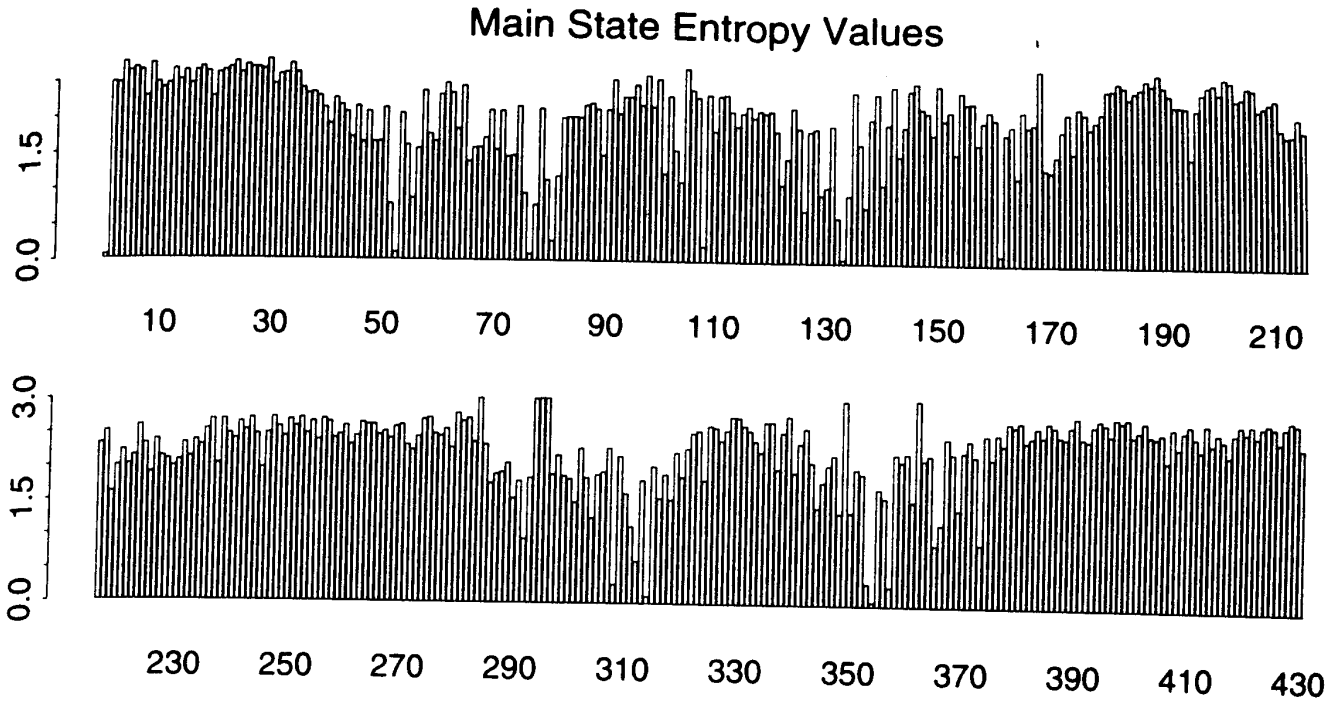



Figure 8.1: Entropy Profile of the Emission Probability Distributions Associated with the Main States of the HMM After 12 Cycles of Training.

# GPCR Analysis (Cont'd)

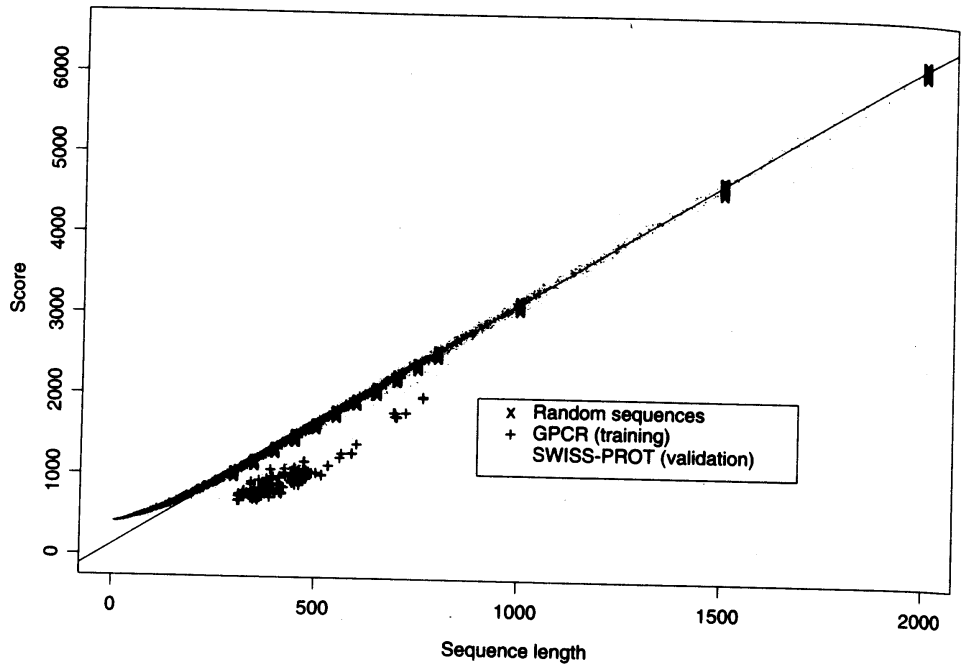



Figure 8.2: Scores (Negative Log-likelihoods of Optimal Viterbi Paths). Represented sequences consist of 142 GPCR training sequences, all sequences from the SWISS-PROT database of length less than or equal to 2000, and 220 randomly generated sequences with same average composition as the GPCRs of length 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800 (20 at each length). The regression line was obtained from the 220 random sequences. The horizontal distances in the histogram correspond to normalized scores (6).

# Applications of HMM for GPCR

- Bacteriorhodopsin
  - Transmembrane protein with 7 domains
  - But it is not a GPCR
  - Compute score and discover that it is close to the regression line. Hence not a GPCR.
- Thyrotropin receptor precursors
  - All have long initial loop on INSERT STATE 20.
  - Also clustering possible based on distance to regression line.

# HMMs – Advantages

- Sound statistical foundations
- Efficient learning algorithms
- Consistent treatment for insert/delete penalties for alignments in the form of locally learnable probabilities
- Capable of handling inputs of variable length
- Can be built in a modular & hierarchical fashion; can be combined into libraries.
- Wide variety of applications: Multiple Alignment, Data mining & classification, Structural Analysis, Pattern discovery, Gene prediction.

# HMMs – Disadvantages

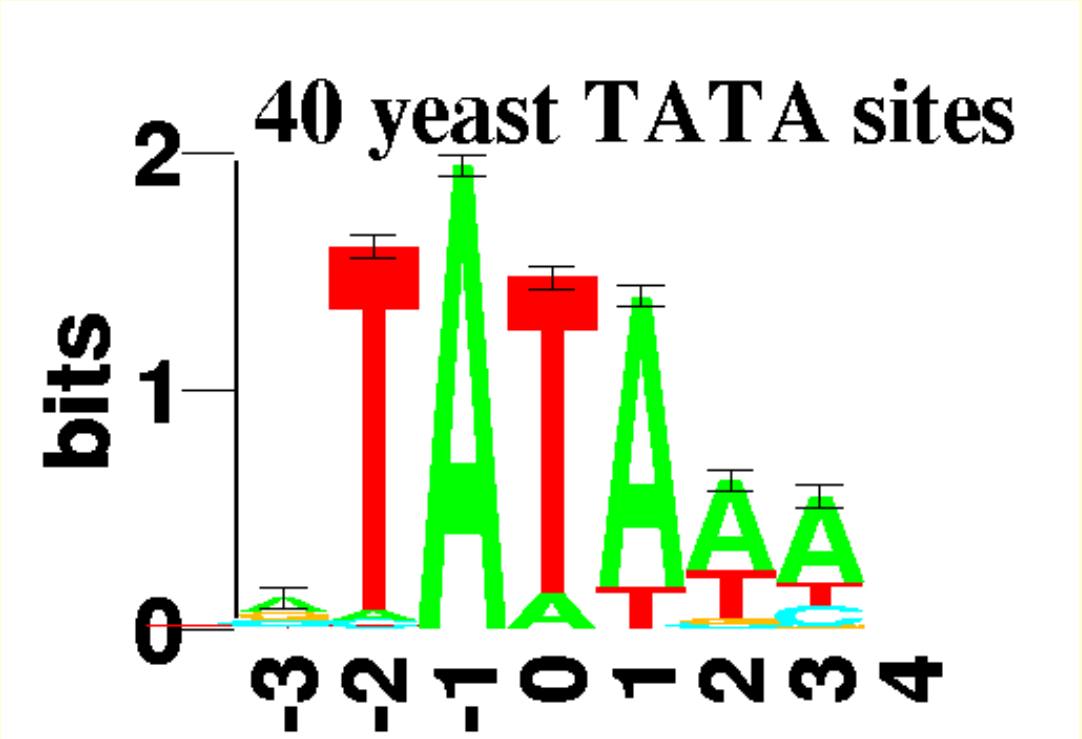
- Large # of parameters.
- Cannot express dependencies & correlations between hidden states.

# 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

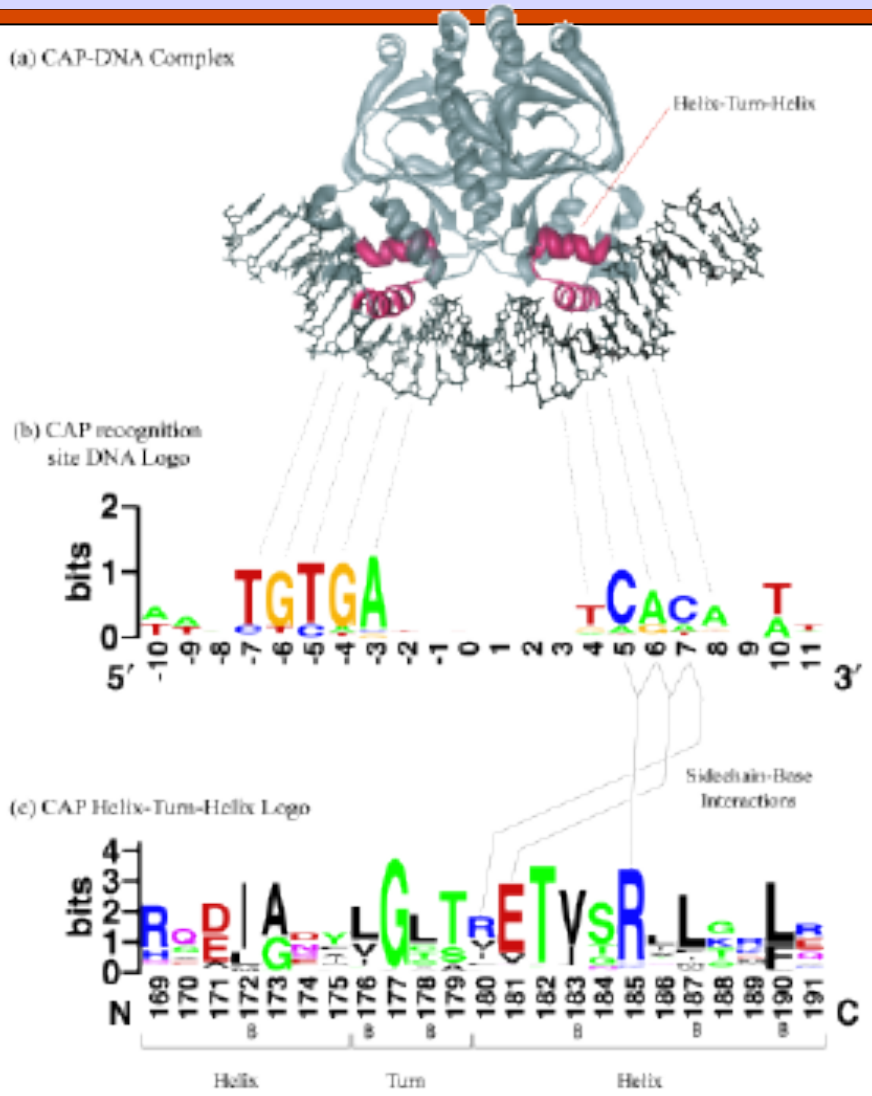
- Given a collection of DNA sequences of promoter regions, locate the transcription factor binding sites (also called regulatory elements)
  - Example:






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

# Motifs






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

# Motifs in DNA Sequences

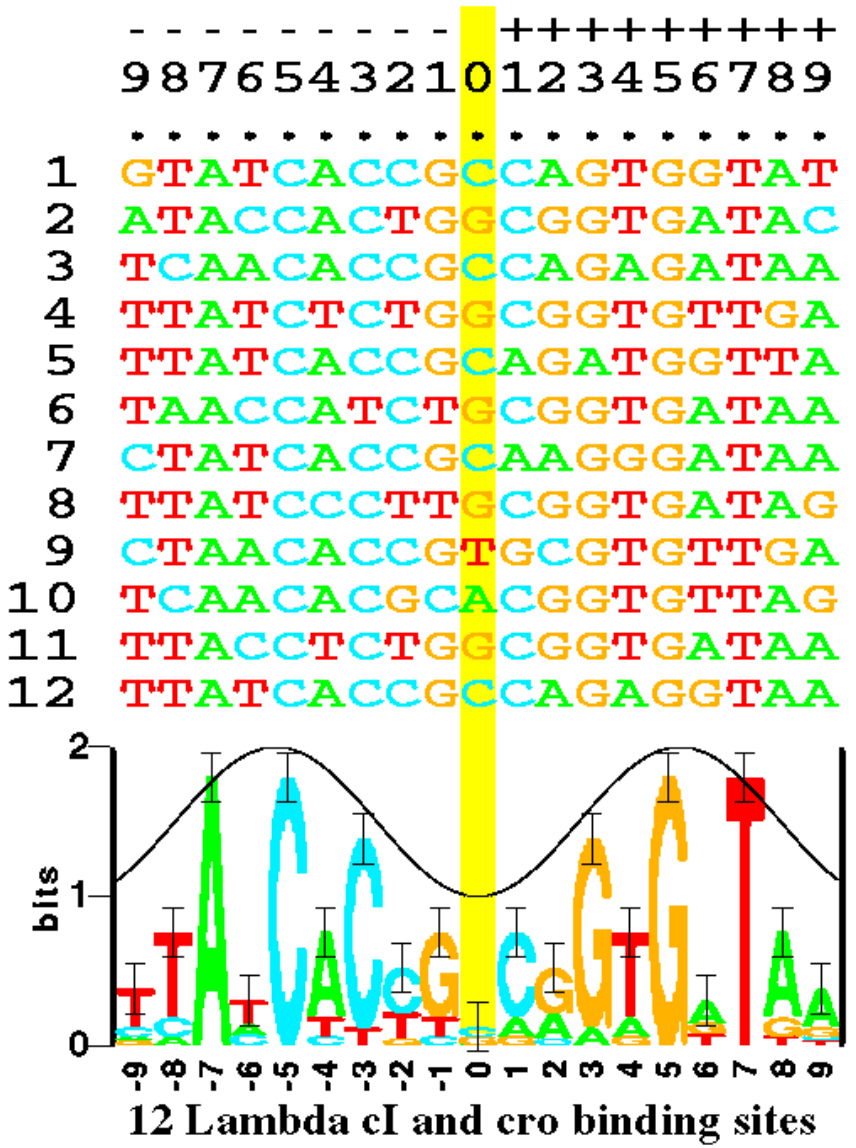
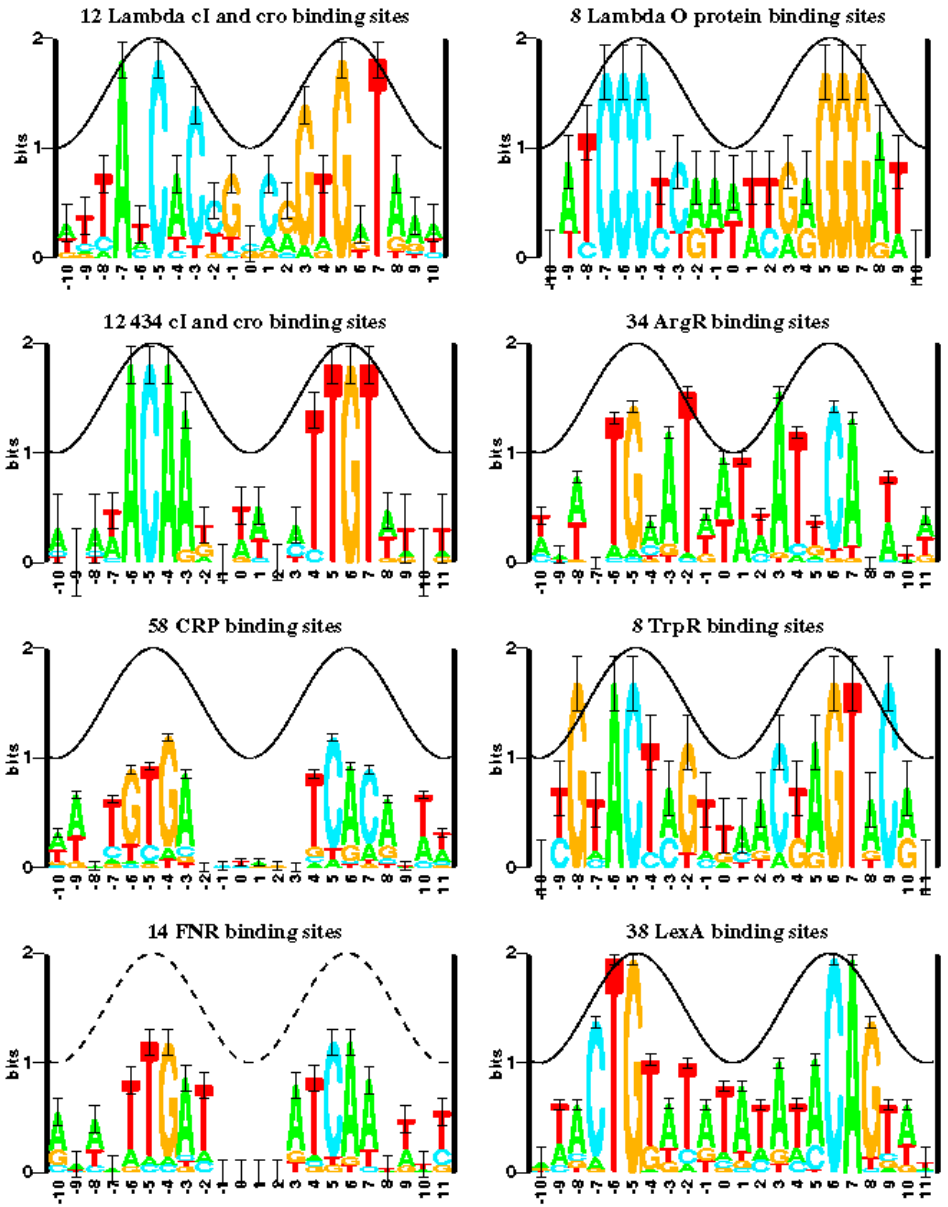



Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the $P_L$ and $P_R$ control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered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

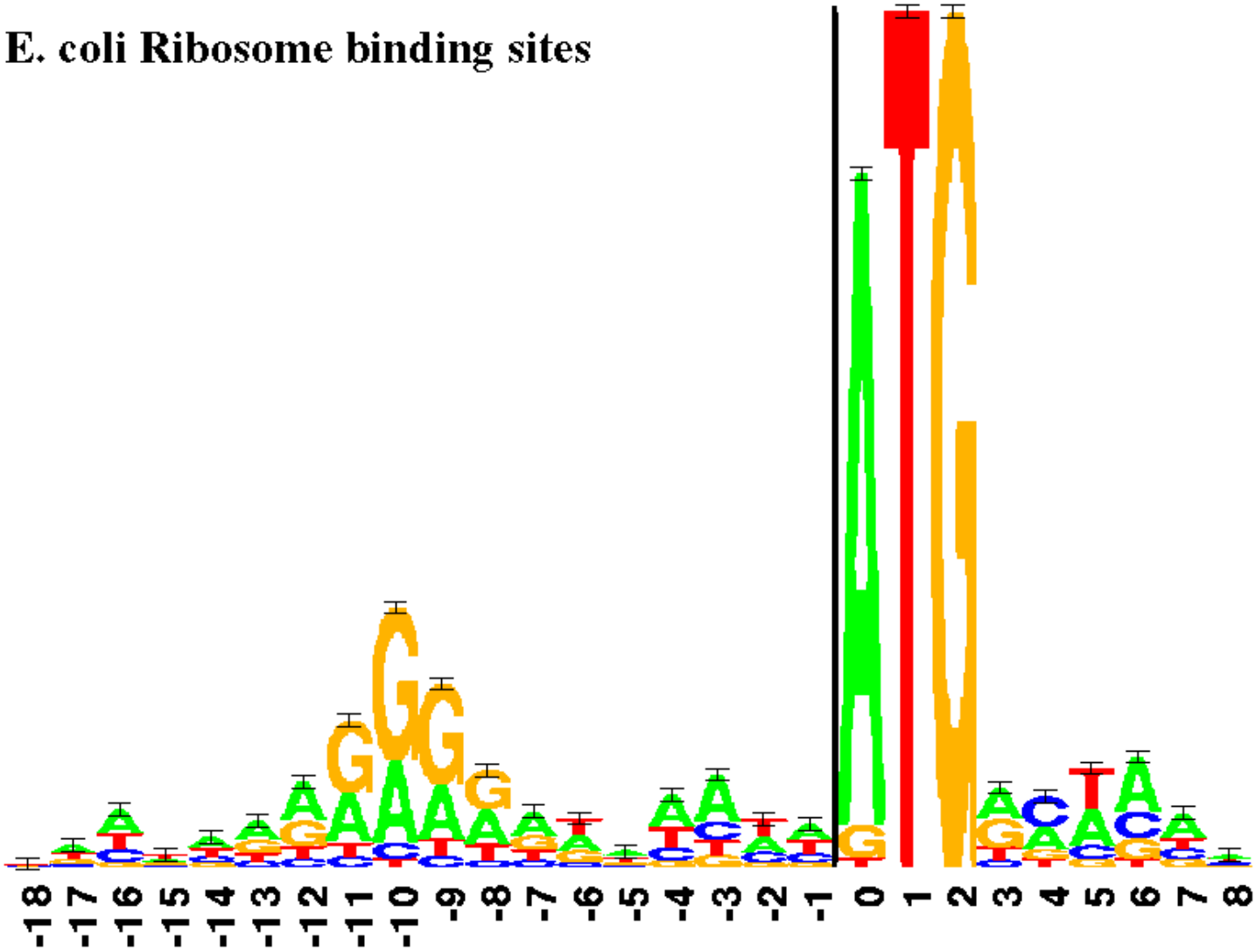
# More Motifs in *E. Coli* DNA Sequences






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

## E. coli Ribosome binding sites

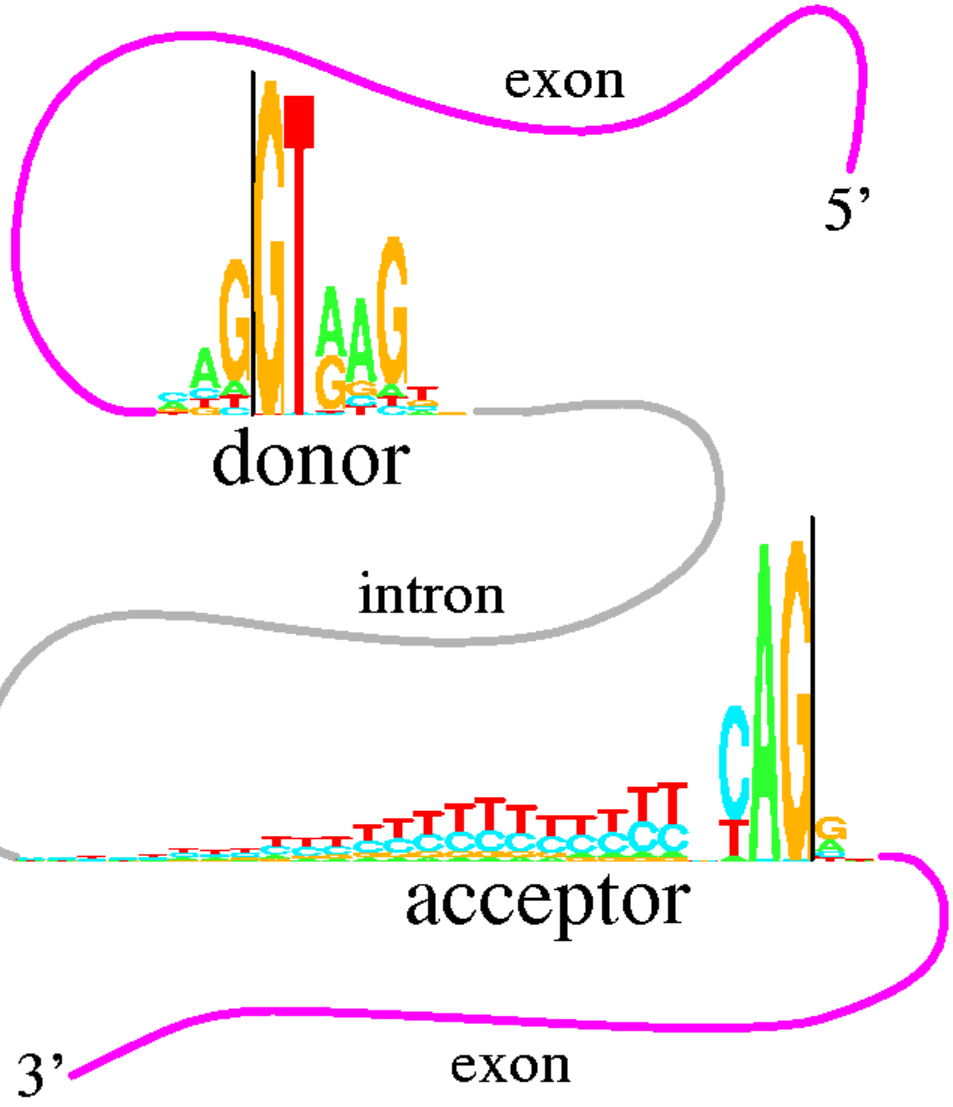





http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

# Other Motifs in DNA Sequences: Human Splice Junctions

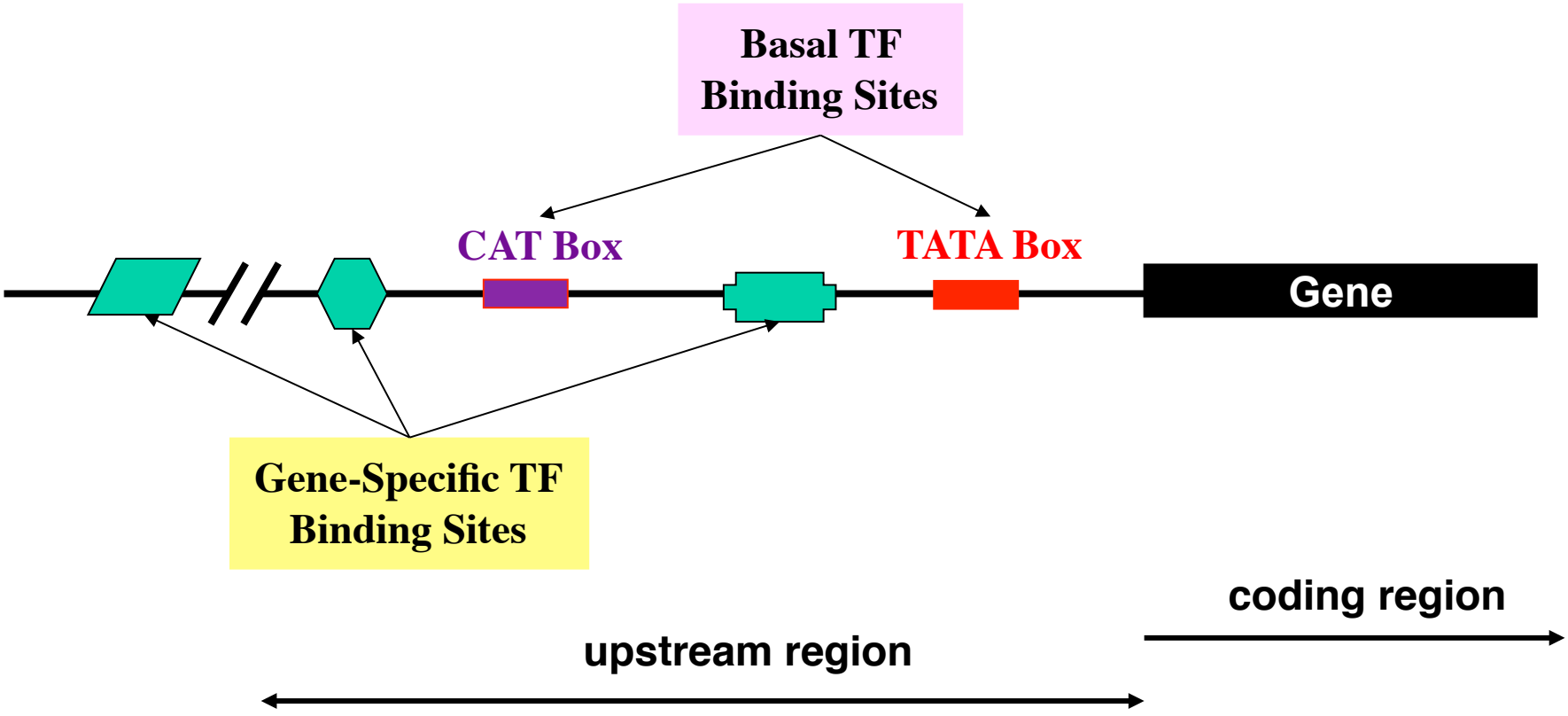
This figure shows two "sequence logos" which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides 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

# Transcription Regulation



# 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 | ACCGTTTAGCCCCTCAAAAACAGGGCAATAAAGTCGAAC | |
| 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.

# Motifs in DNA Sequences

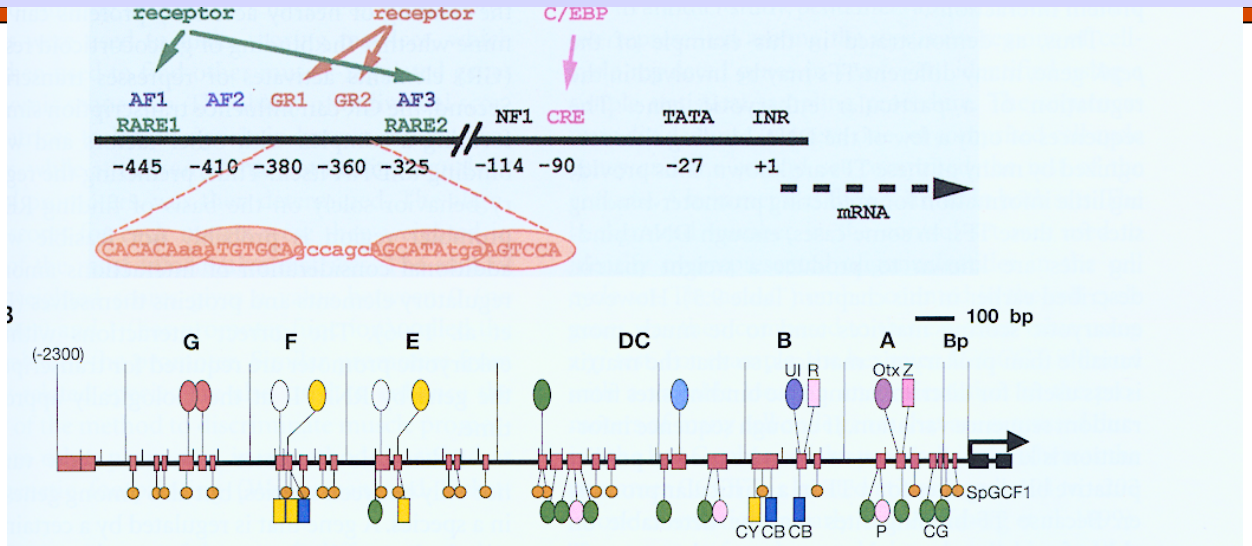
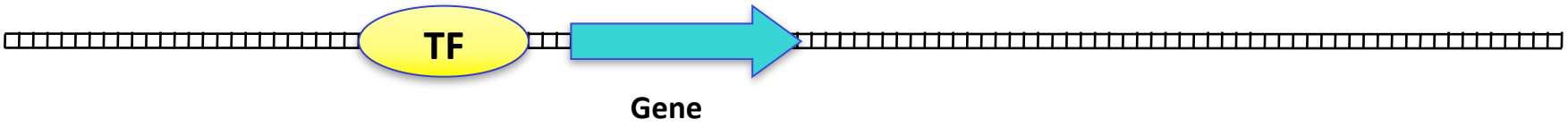



FIGURE 9.13. Regulatory elements of two promoters. (*A*) The rat *pepCK* gene. The relative positions of the TF-binding sites are illustrated (Yamada et al. 1999). The glucocorticoid response unit (GRU) includes three accessory factor–binding sites (AF1, AF2, and AF3), two glucocorticoid response elements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

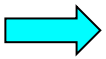
# Single Gene Activation





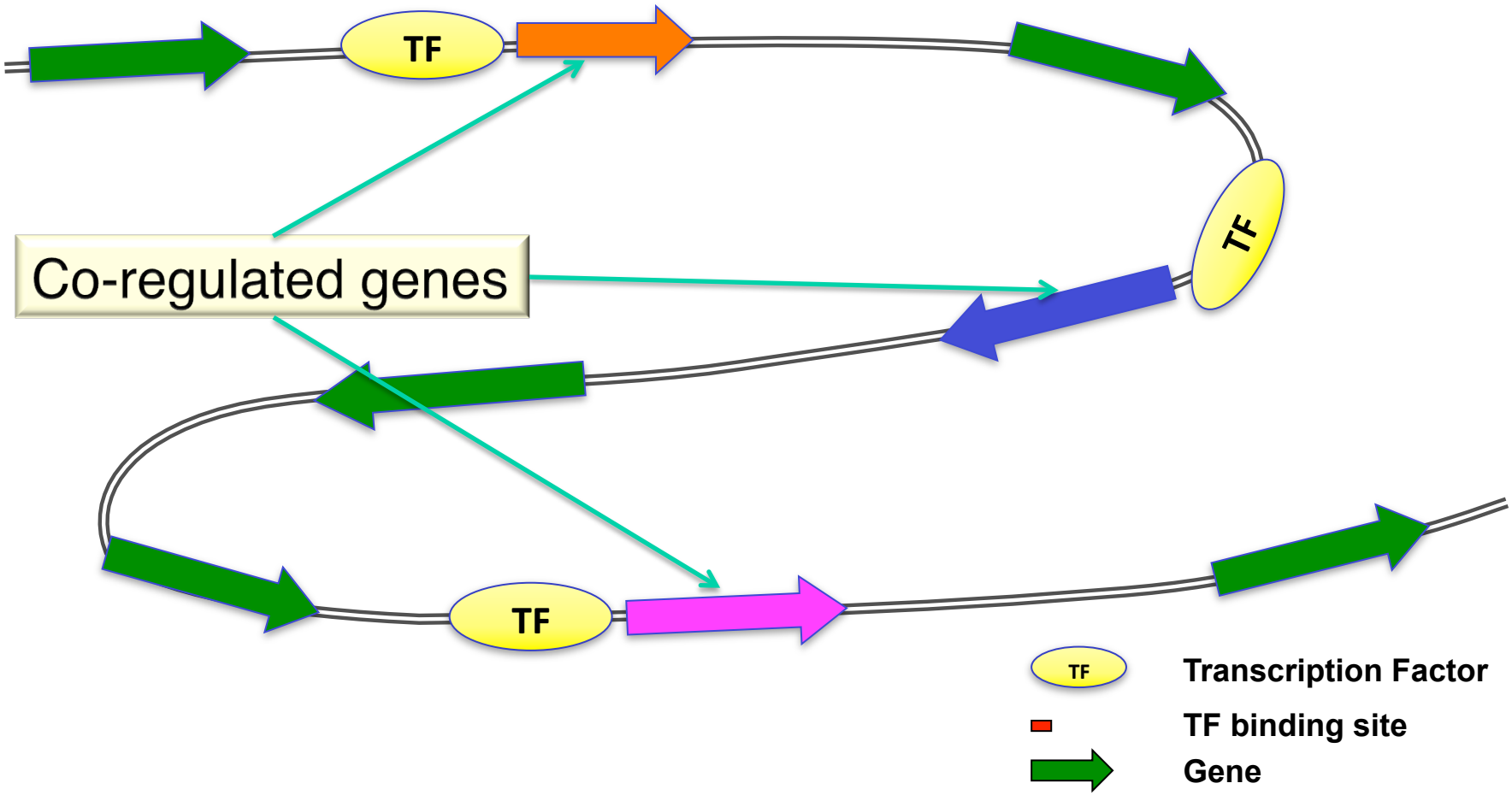
Transcription Factor

TF binding site

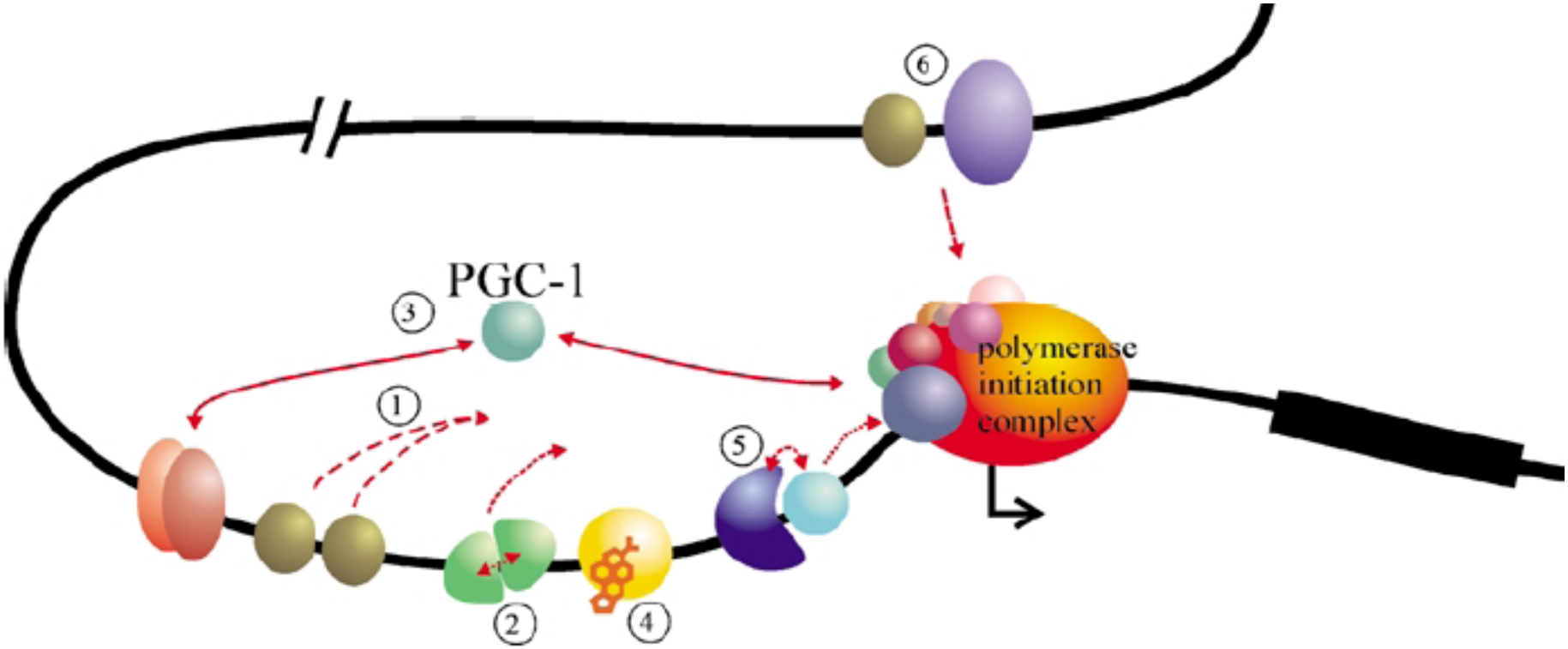


Gene

# Multiple Gene Activation



# Transcription Regulation

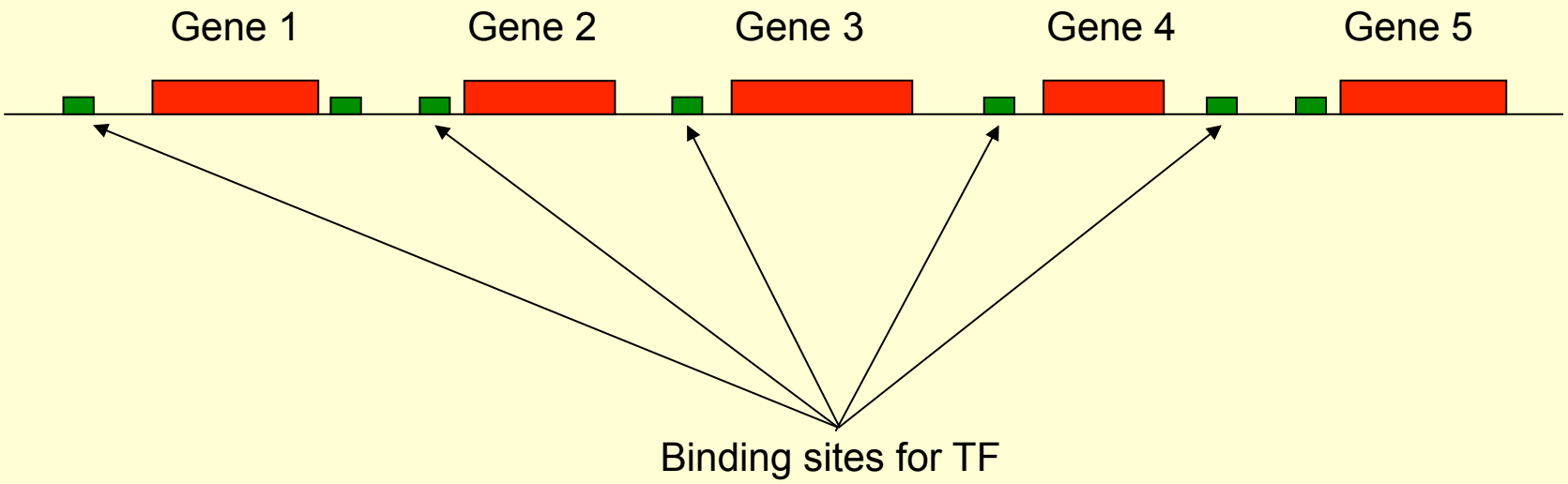



[ Goffart *et al. Exp. Physiology* (2003) ]

# Motif-prediction: Whole genome

Problem: Given the upstream regions of all genes in the genome, find all over-represented sequence signatures.

Basic Principle: If a TF co-regulates many genes, then all these genes should have at least 1 binding site for it in their upstream region.



# 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

# EM Algorithm

**Goal**: Find θ, Z that maximize Pr (X, Z | θ)

**Initialize**: random profile

**E-step**: Using profile, compute a likelihood value $z_{ij}$ for each *m*-window at position *i* in input sequence *j*.

**M-step**: Build a new profile by using every *m*-window, but weighting each one with value $z_{ij}$.

**Stop** if converged

MEME [Bailey, Elkan 1994]

# Gibbs Sampling for Motif Detection