CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

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

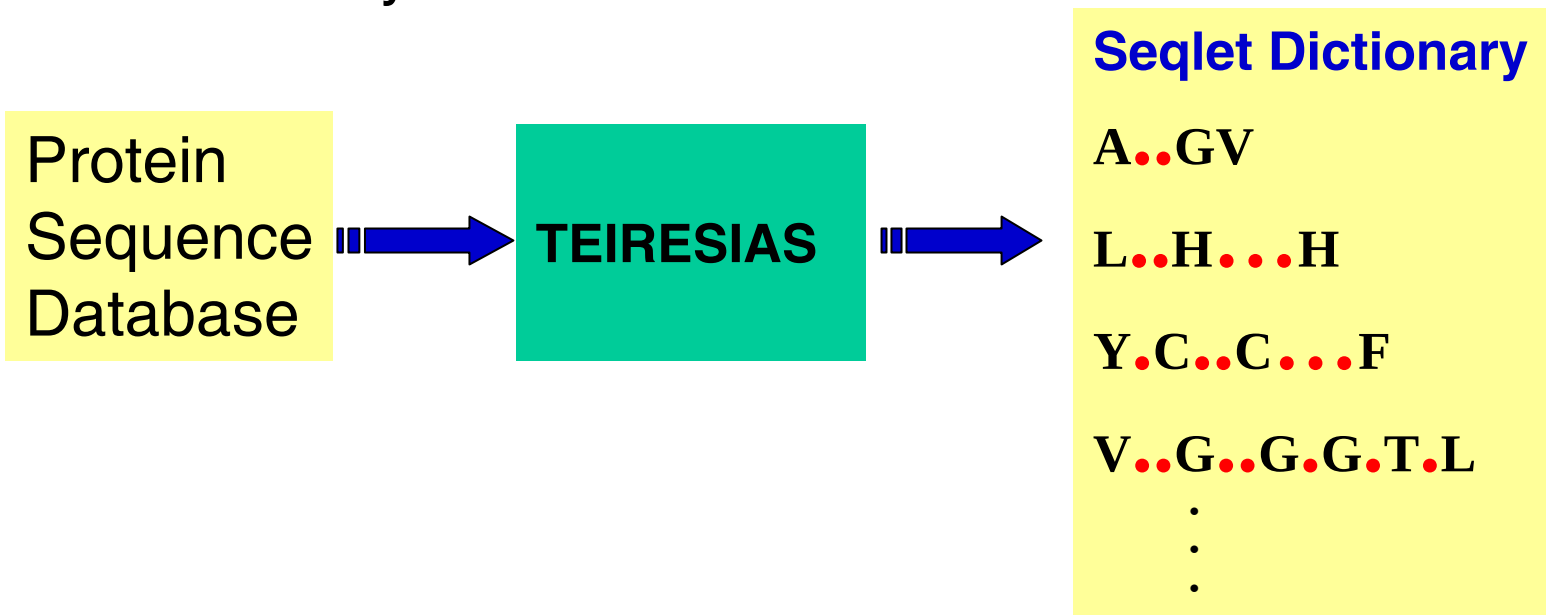
Pattern Discovery

What we have discussed so far

- Why Pattern discovery?
- Types of patterns
- How to represent and store patterns?
- Types of pattern discovery
  - Supervised pattern discovery
    - Motif Detection
  - Unsupervised pattern discovery
- Evaluation of pattern discovery

Unaligned Pattern Discovery

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



Rigoutsos & Floratos, Bioinformatics, '98

TEIRESIAS: Key Features

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

  A..GV.S ⊕ V.S.GR = A..GV.S.GR

- **Order of Evaluation** carefully chosen so that long patterns get longer first
- Finds **all maximal patterns.**
- **Combinatorial explosion** avoided by generating only relevant maximal patterns.

Rigoutsos & Floratos, Bioinformatics, '98

SPLASH

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

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

Precomputed Sequence Patterns

- PROSITE
- BLOCKS and PRINTS
- *e*MOTIF
- SPAT
- PRODOM
- Pfam

Motif Detection Tools

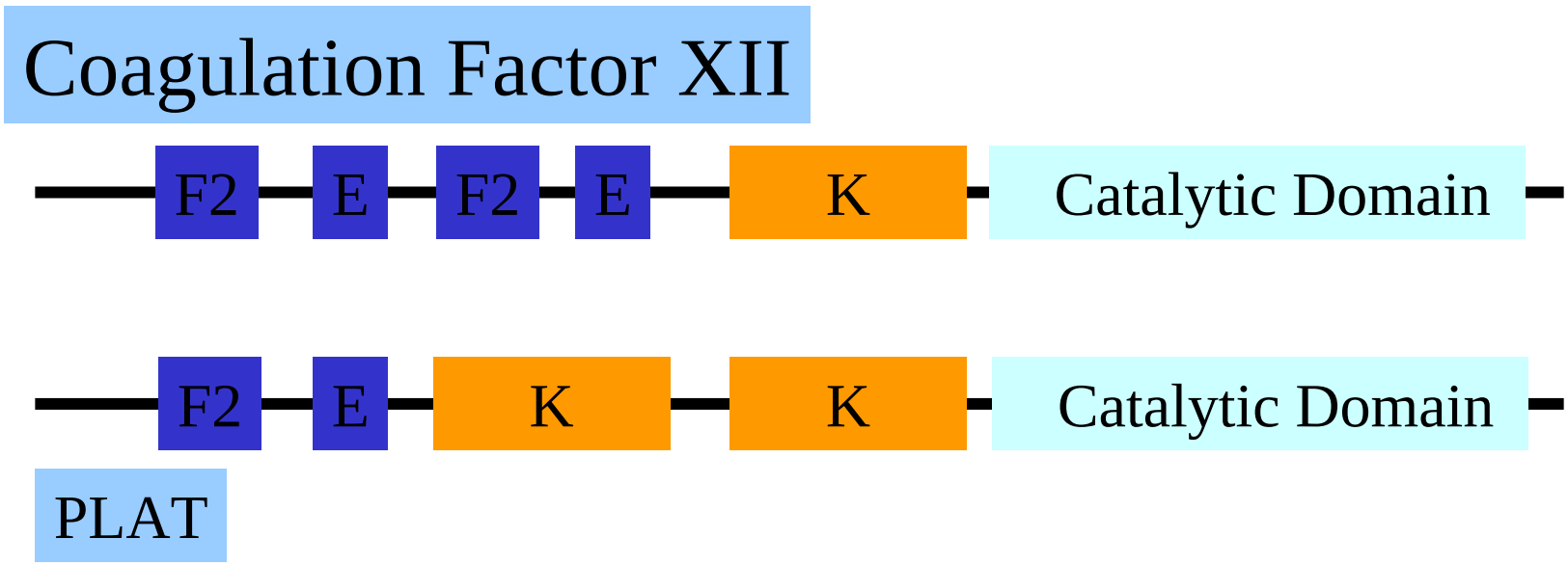
- PROSITE (Database of protein families & domains)
  - Try PDOC00040. Also Try PS00041
- PRINTS Sample Output
- BLOCKS (multiply aligned ungapped segments for highly conserved regions of proteins; automatically created) Sample Output
- Pfam (Protein families database of alignments & HMMs)
  - Multiple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

Protein Information Sites

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

Modular Nature of Proteins

- Proteins are collections of "modular" domains. For example,



Domain Architecture Tools

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

Predicting Specialized Structures

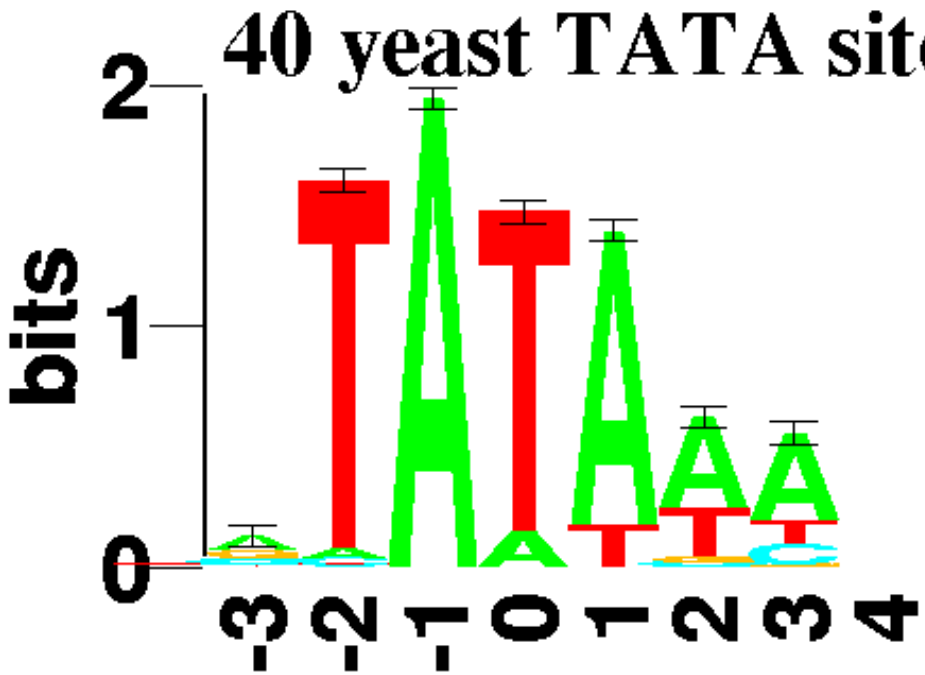
- COILS – Predicts coiled coil motifs
- TMPred – predicts transmembrane regions
- SignalP – predicts signal peptides
- SEG – predicts nonglobular regions

Patterns in DNA Sequences

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

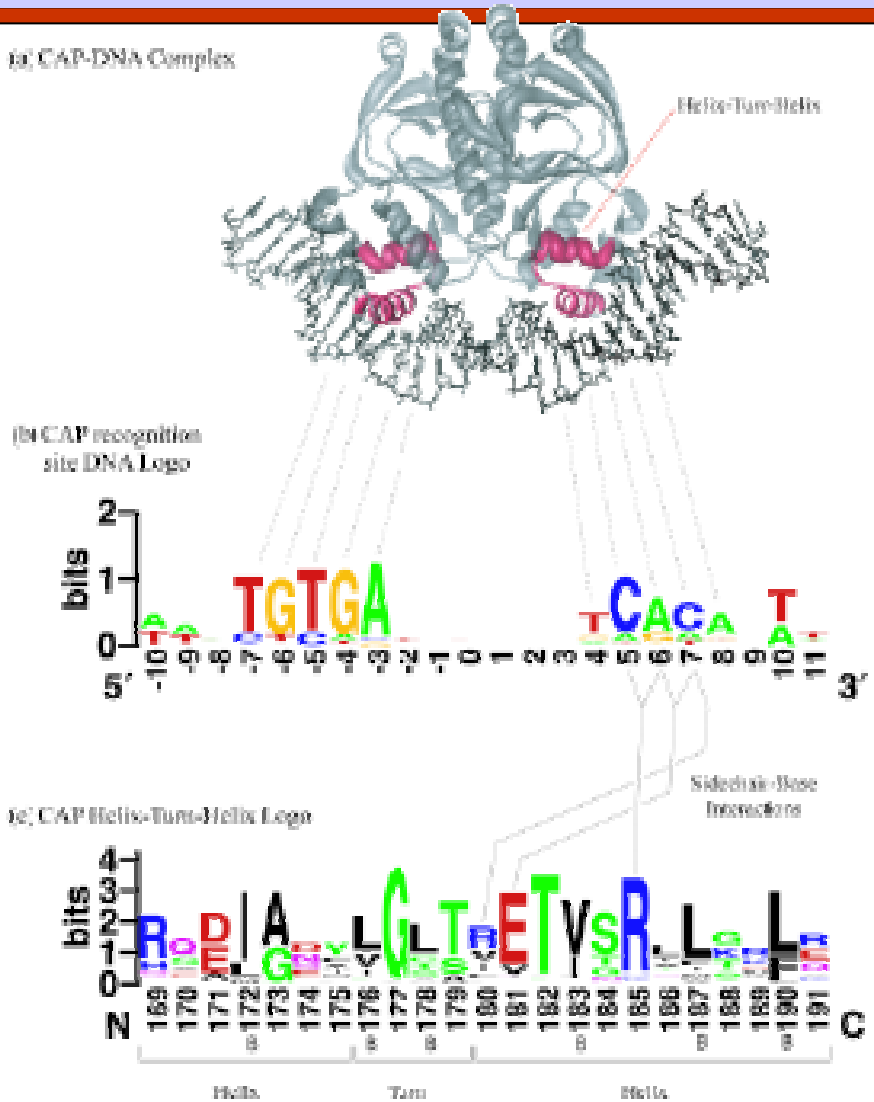
Motifs in DNA Sequences

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



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

Motifs



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

Motifs in DNA Sequences

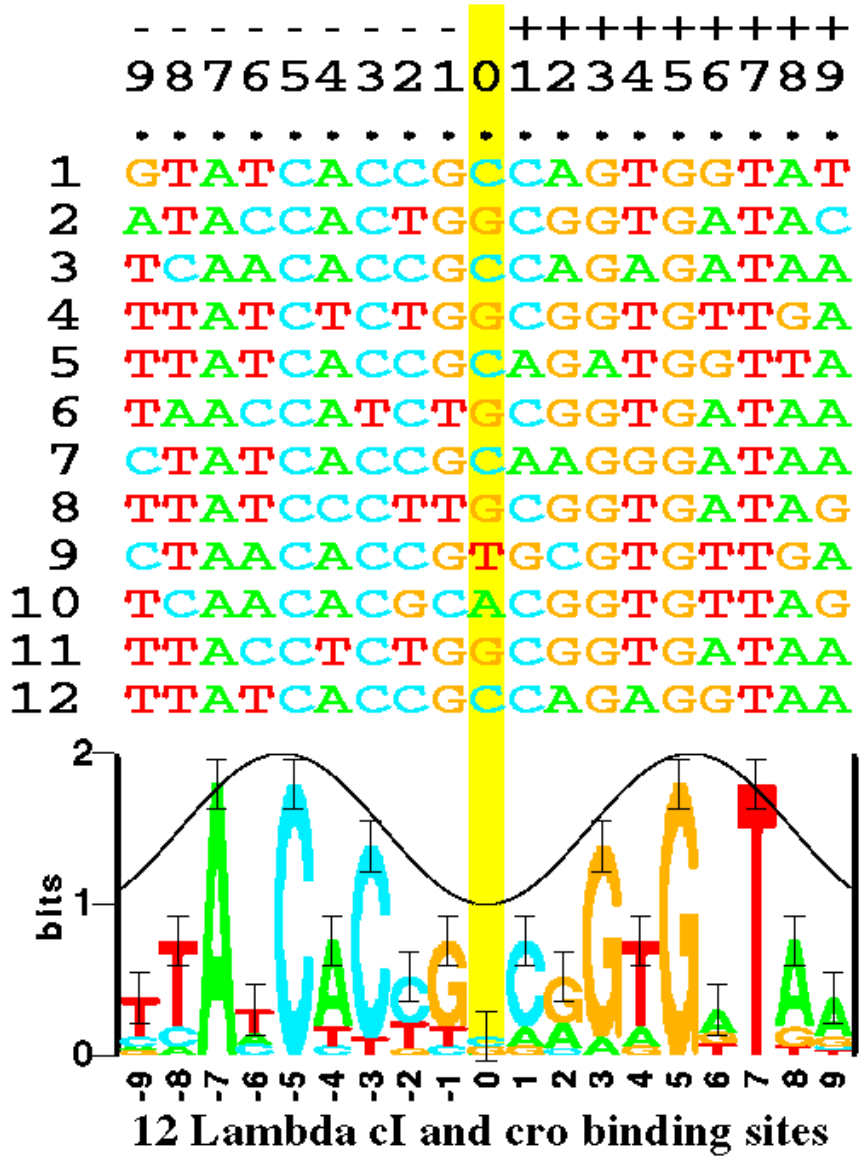
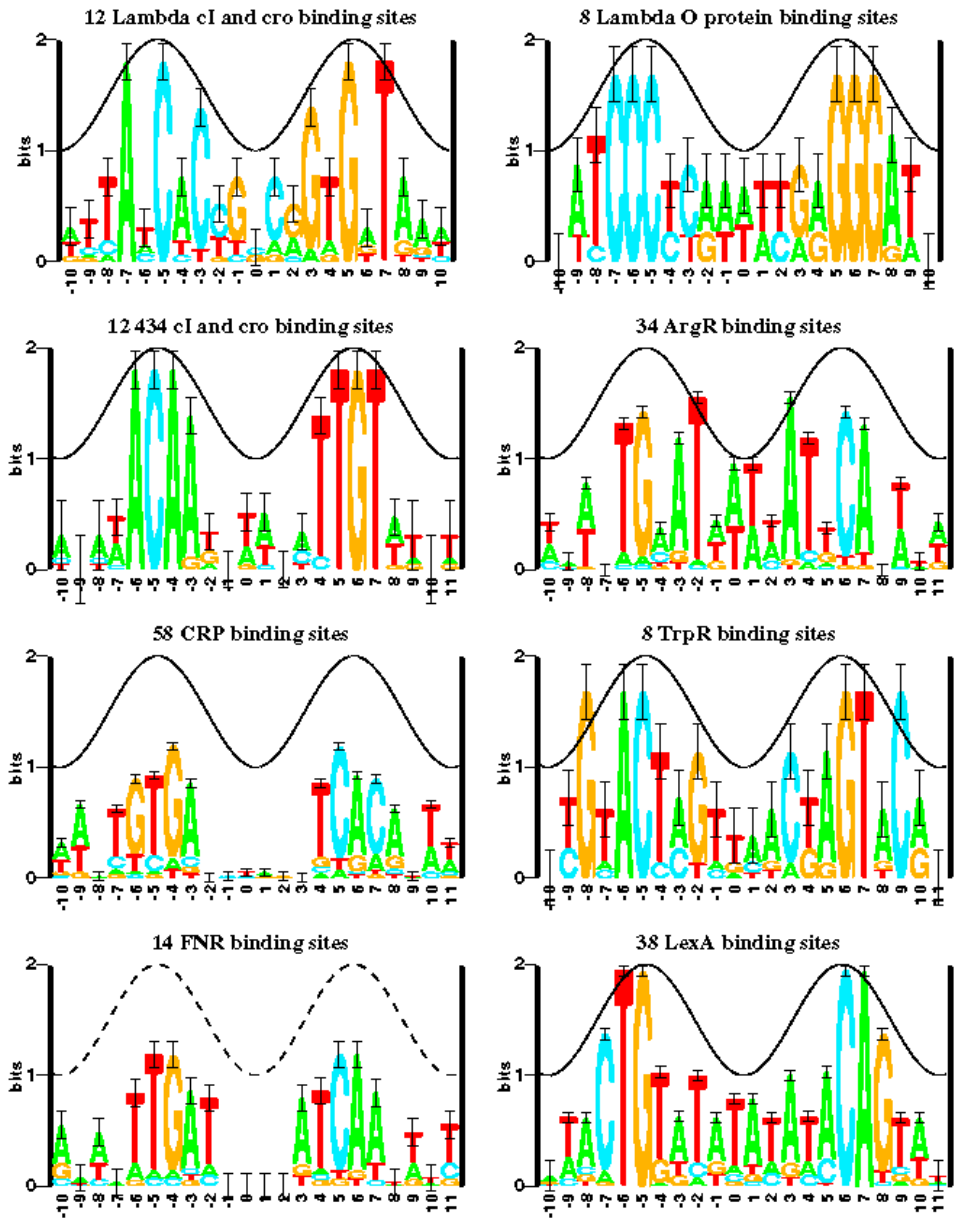


Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered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

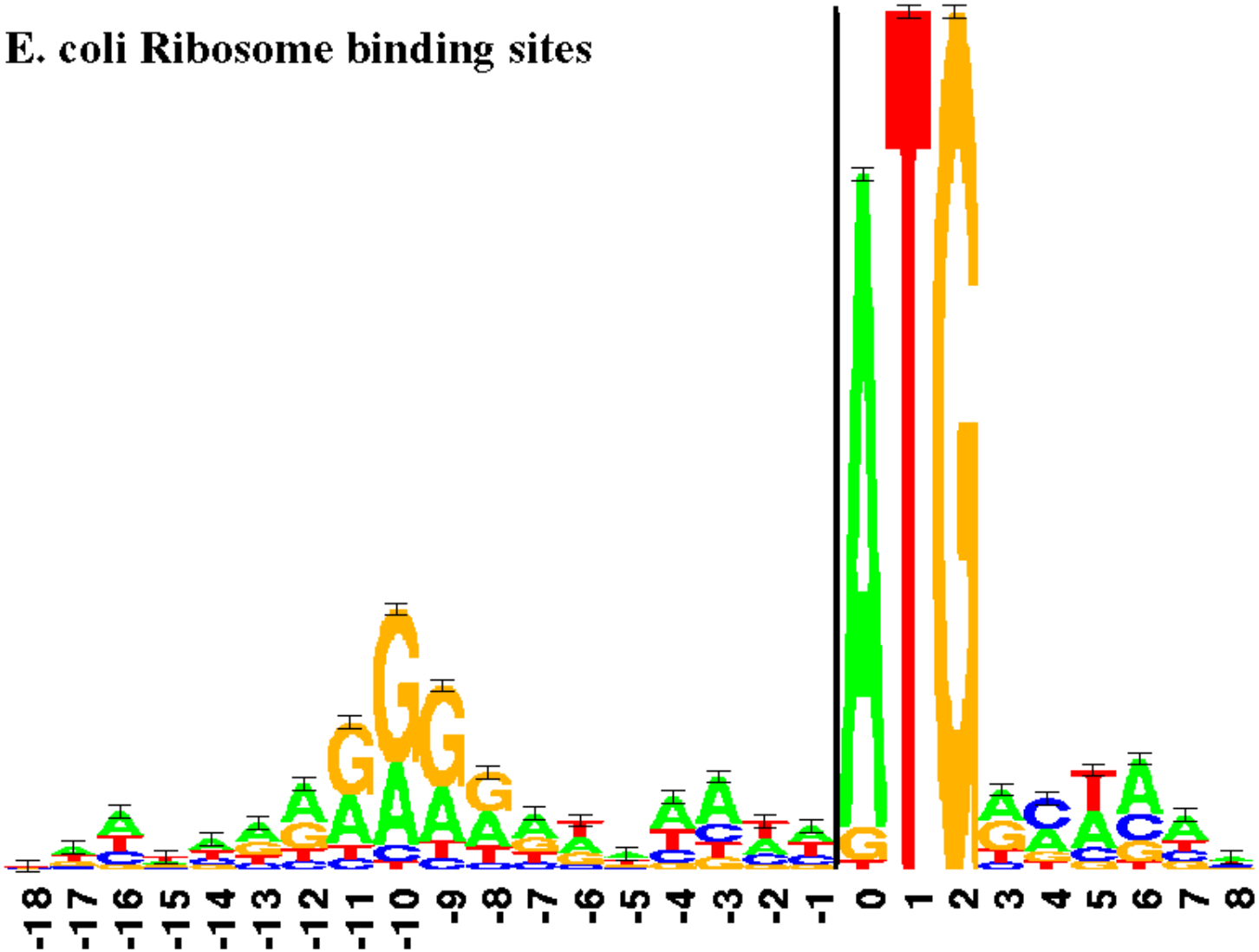
http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

More Motifs in *E. Coli* DNA Sequences



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

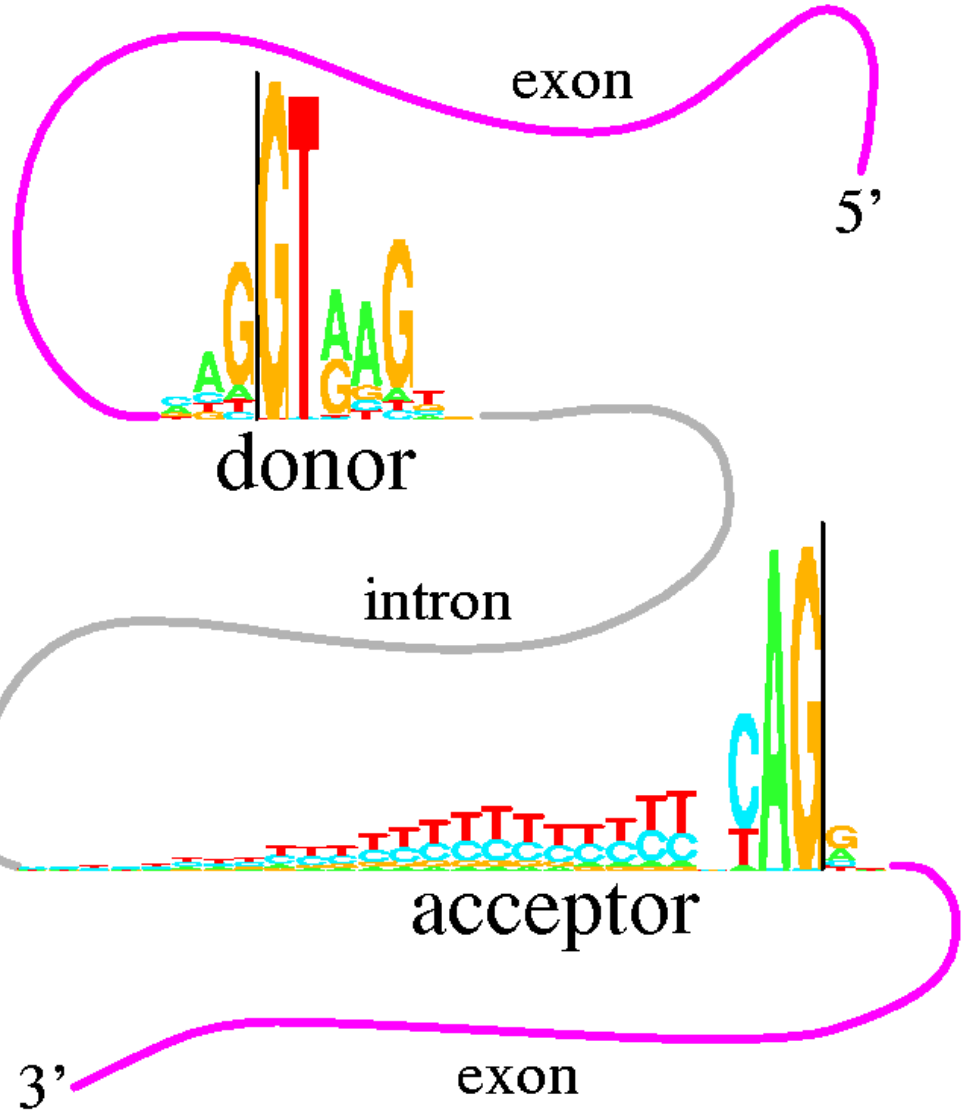
E. coli Ribosome binding sites



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

Motifs in DNA Sequences

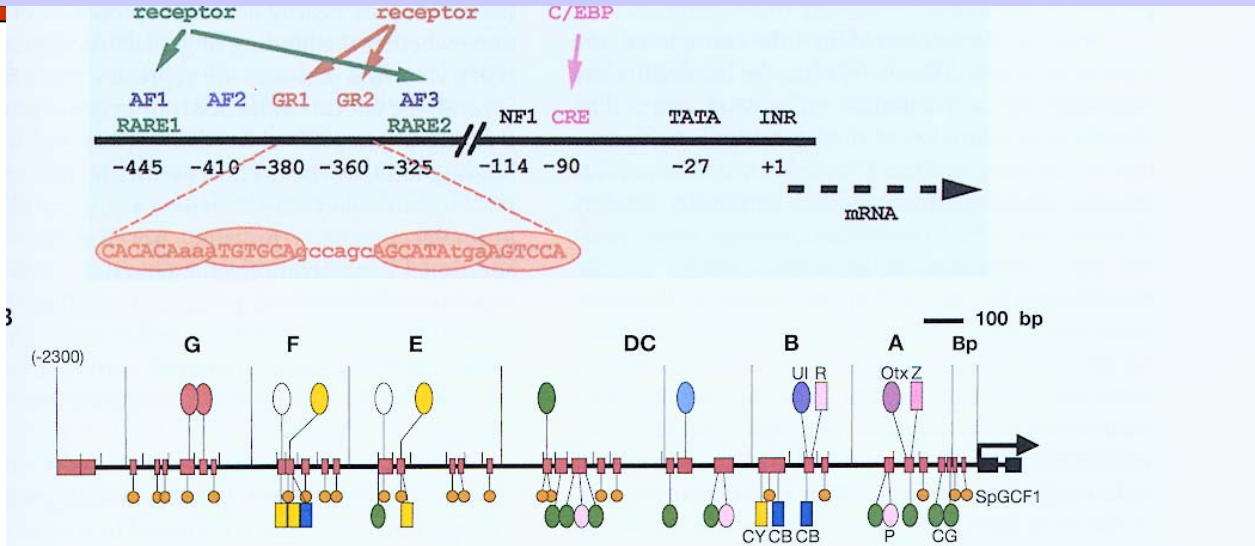
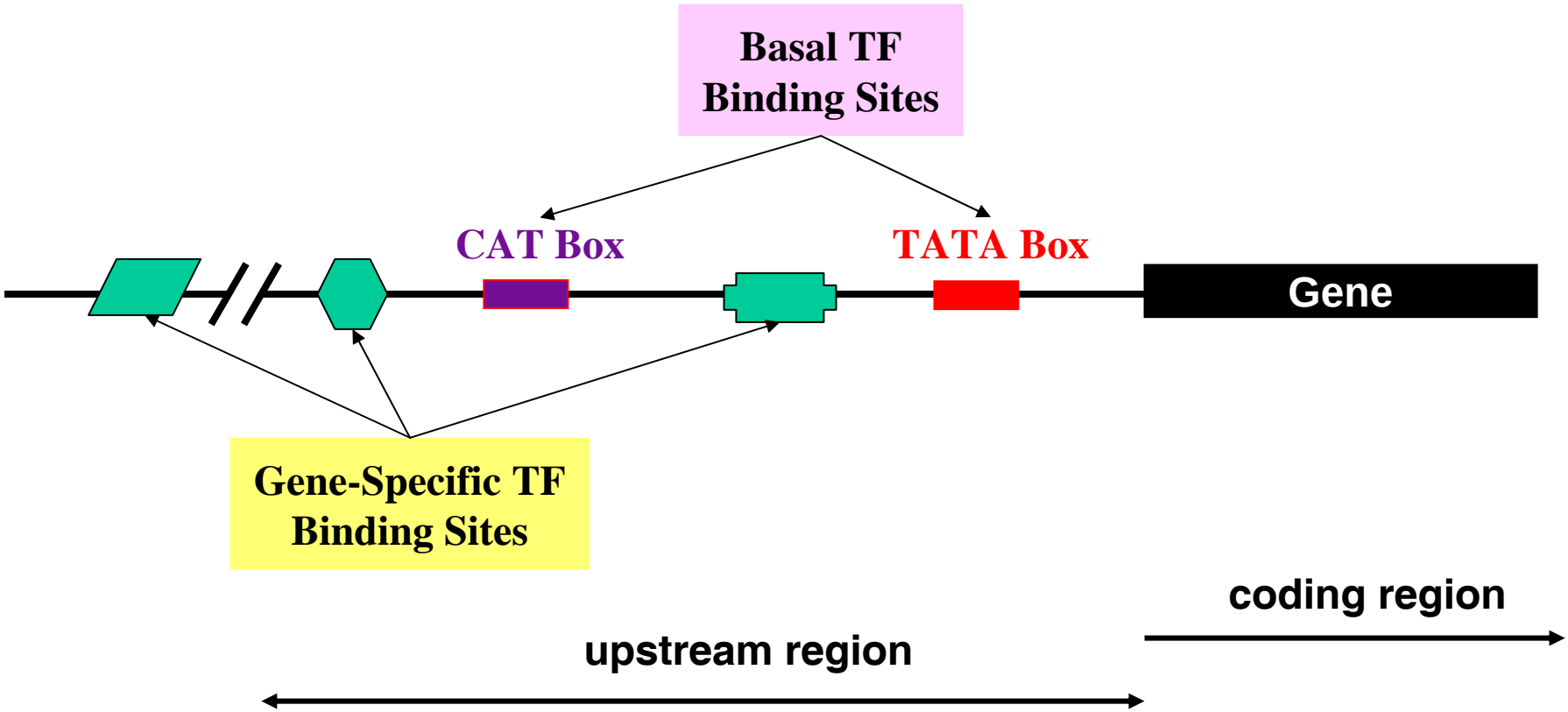
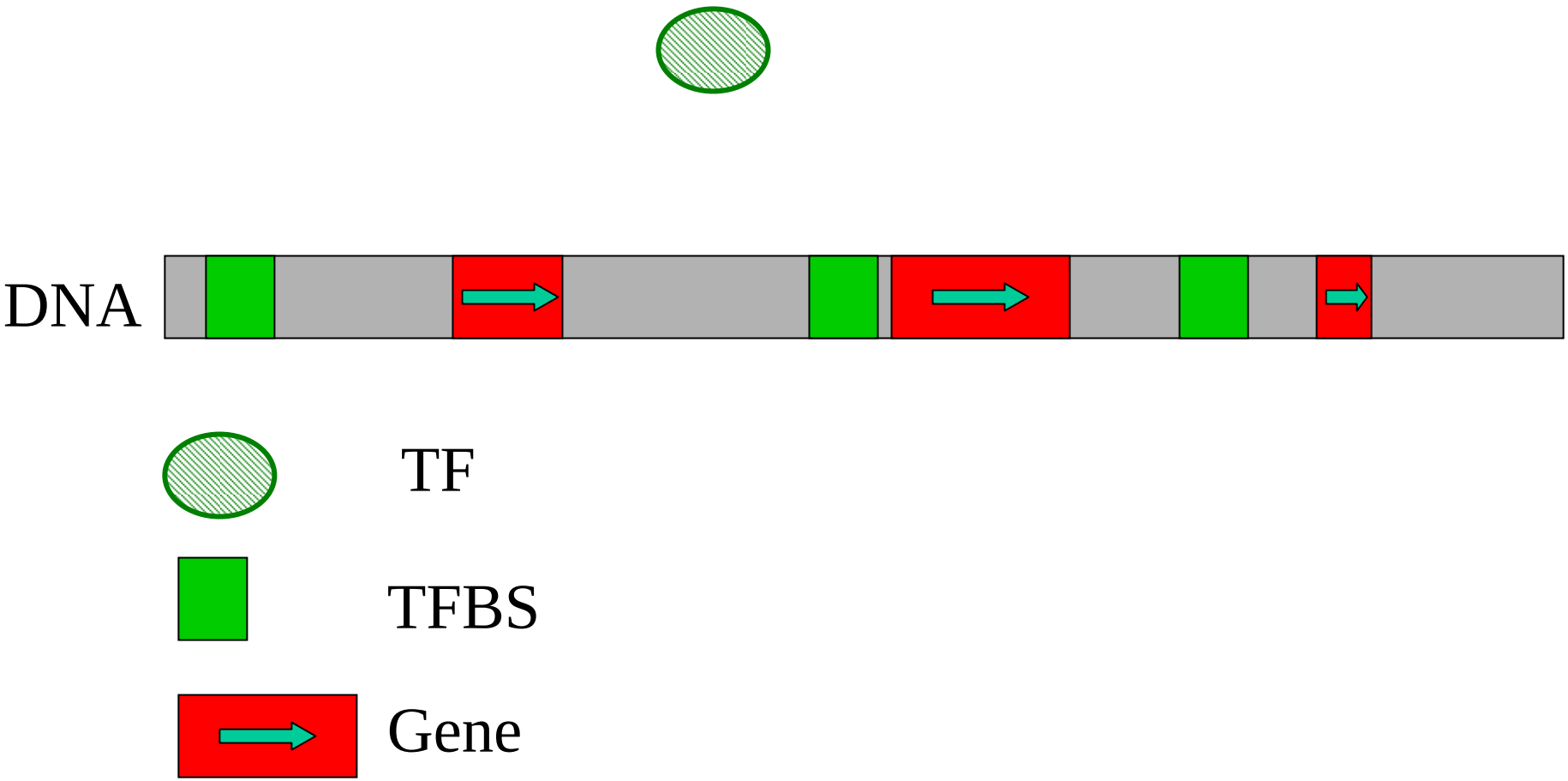


FIGURE 9.13. Regulatory elements of two promoters. (*A*) The rat *pepCK* gene. The relative positions of the TF-binding sites are illustrated (Yamada et al. 1999). The glucocorticoid response unit (GRU) includes three accessory factor–binding sites (AF1, AF2, and AF3), two glucocorticoid response elements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

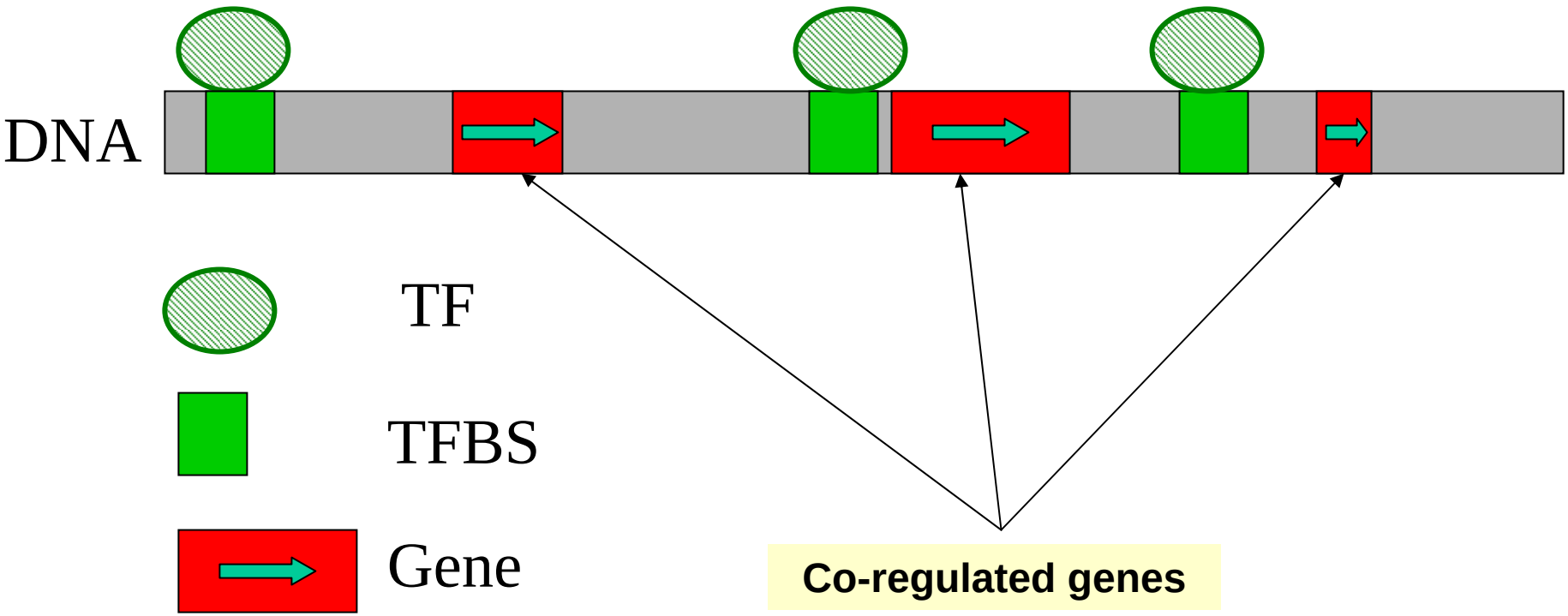
Transcription Regulation



Transcription Regulation



Transcription Regulation

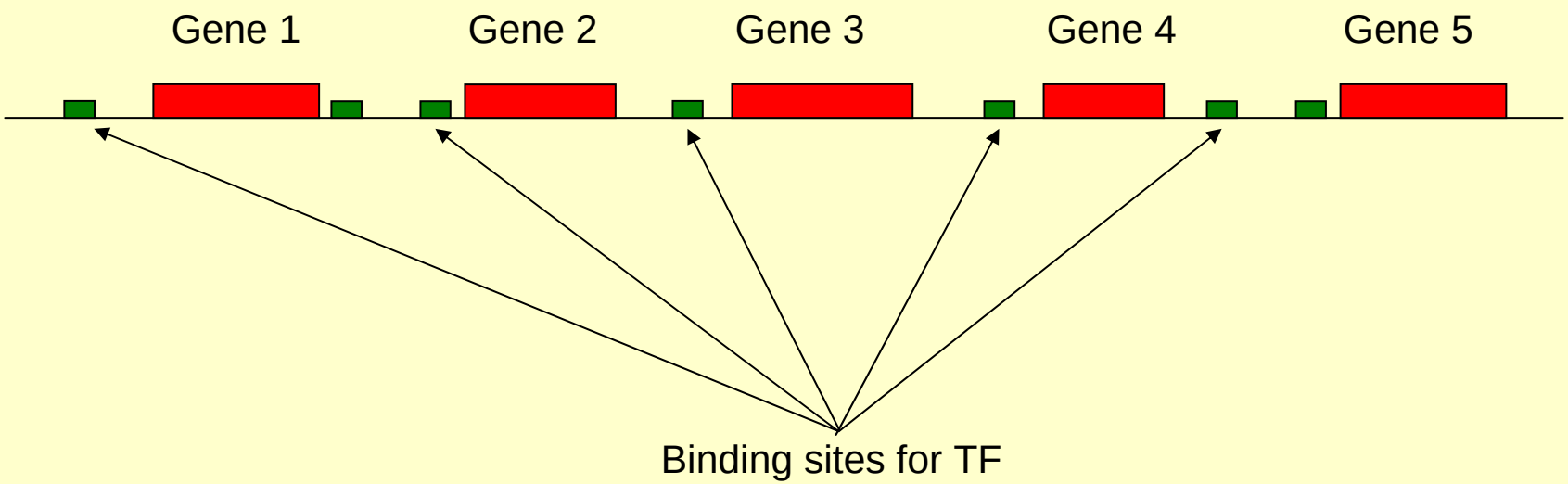


Motif-prediction: Whole genome

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

Motif-prediction: Whole genome

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

Predicting Motifs in Whole Genome

- **MEME:** **EM algorithm** [Bailey *et al.*, 1994]
- **AlignACE:** **Gibbs Sampling Approach** [Hughes *et al.*, 2000]
- **Consensus:** **Greedy Algorithm Based** [Hertz *et al.*, 1990]
- **ANN-Spec:** **Artificial Neural Network and a Gibbs sampling method** [Workman *et al.*, 2000]
- **YMF:** **Enumerative search** [Sinha *et al.*, 2003]
- **...**

EM Method: Model Parameters

- Input: upstream sequences
  - $X = \{X_1, X_2, \ldots, X_n\}$,
- Motif profile: 4×k matrix $\theta = (\theta_{rp})$,
  - $r \in \{A,C,G,T\}$
  - $1 \le p \le k$
  - θ_{rp} = Pr(residue r in position p of motif)
- Background distribution:
  - θ_{r0} = Pr(residue r in background)

EM Method: Hidden Information

- $Z = \{Z_{ij}\}$, where

$$Z_{ij} = \begin{cases} 1, & \text{if motif instance starts at position } i \text{ of } X_j \\ 0, & \text{otherwise} \end{cases}$$

- Iterate over probabilistic models that could generate X and Z, trying to converge on this solution

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