Transcription Factor Binding Site motifs

Wstęp do biologii obliczeniowej

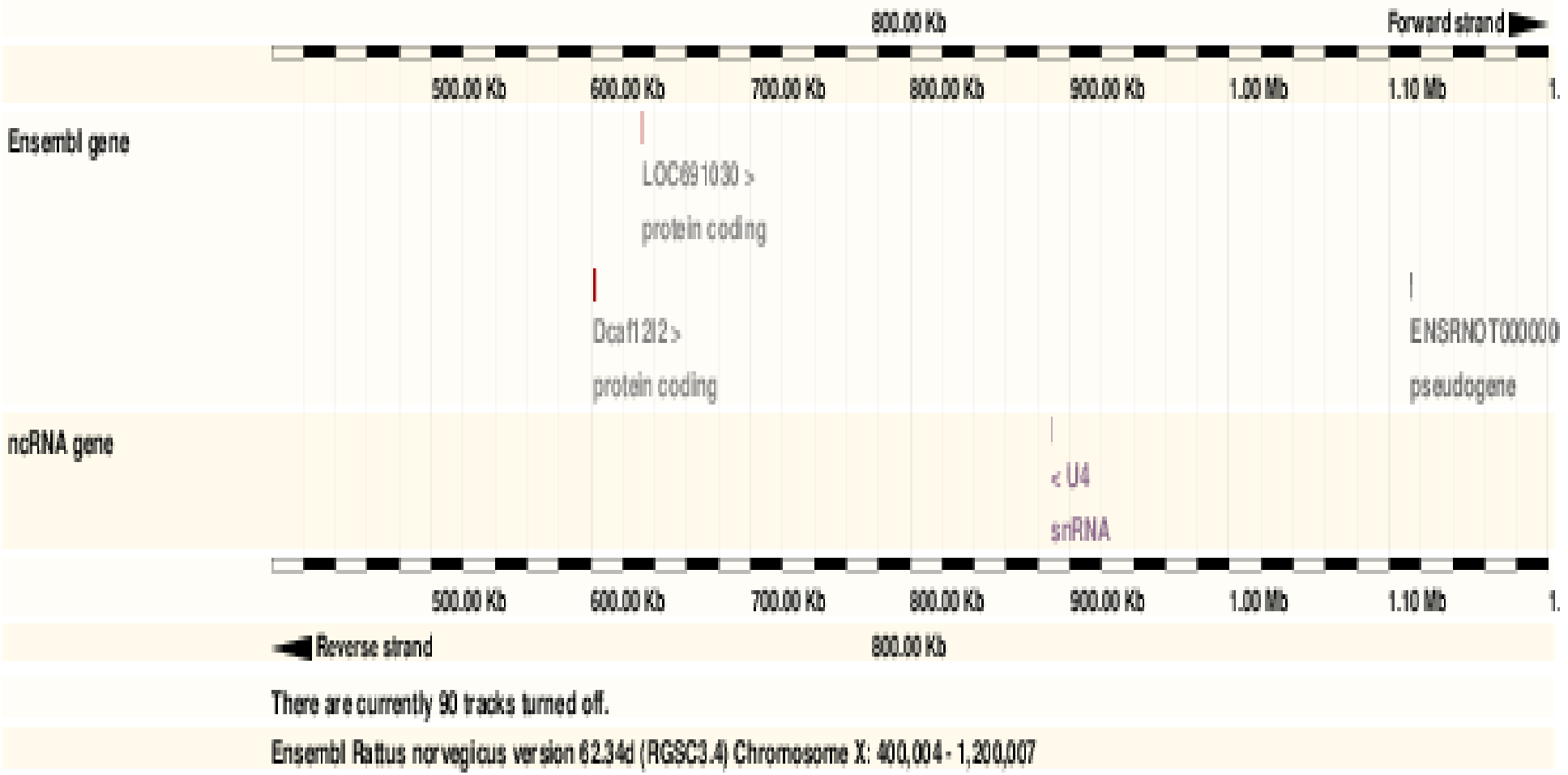
Lecture 11

May 12th 2020

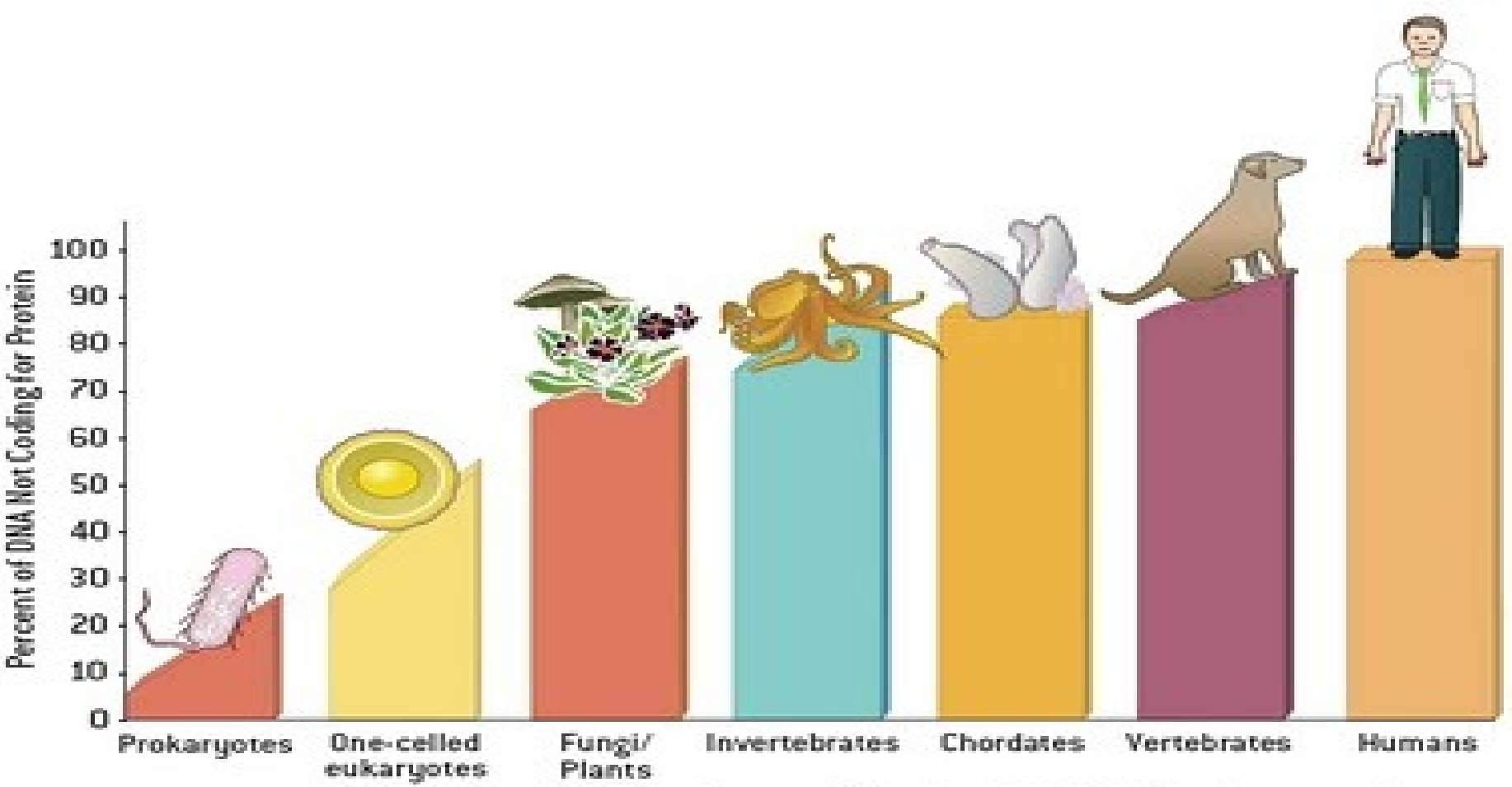
We know about genes, what's with the rest of the genome?

- We know that genomes have “genes” which are the “important” bits of inherited information
- We know that protein coding sequences are “genes”, as they are very important for cellular function
- Non-coding sequences include introns and intergenic sequences, and comprise large proportions of the genome

Random genomic locus from Human genome...

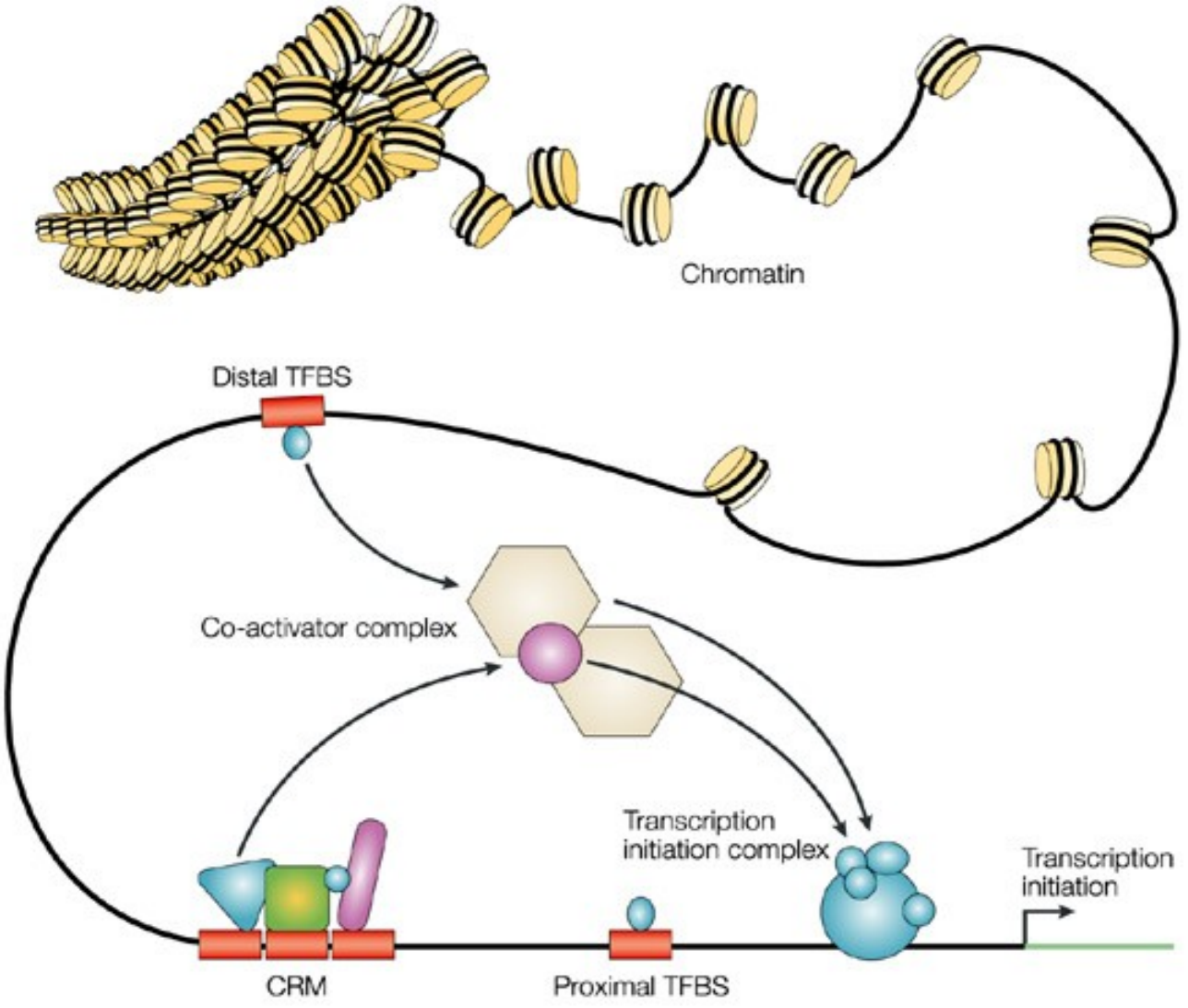


Is this “junk” DNA?

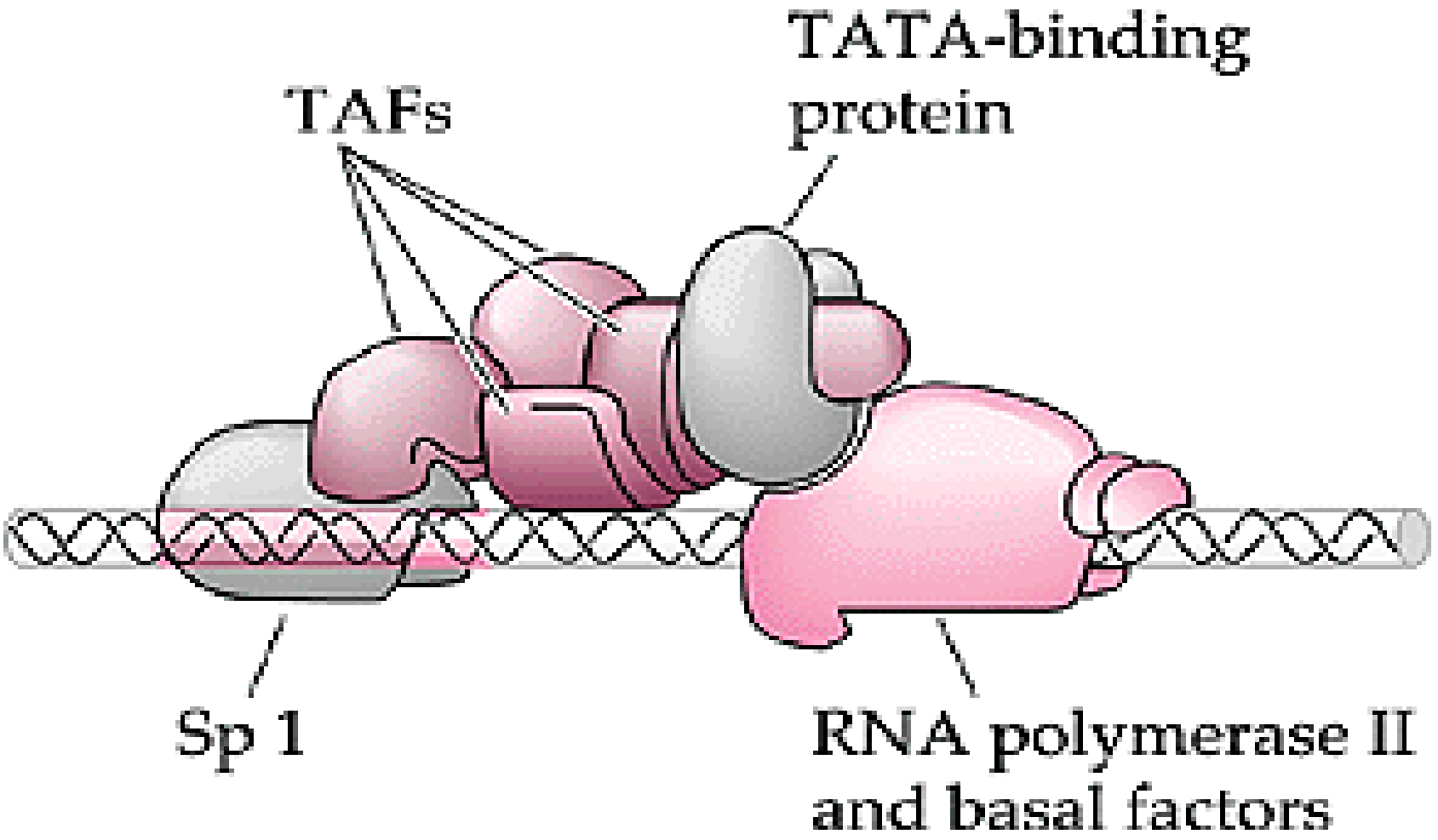


NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to explain organisms' complexity.

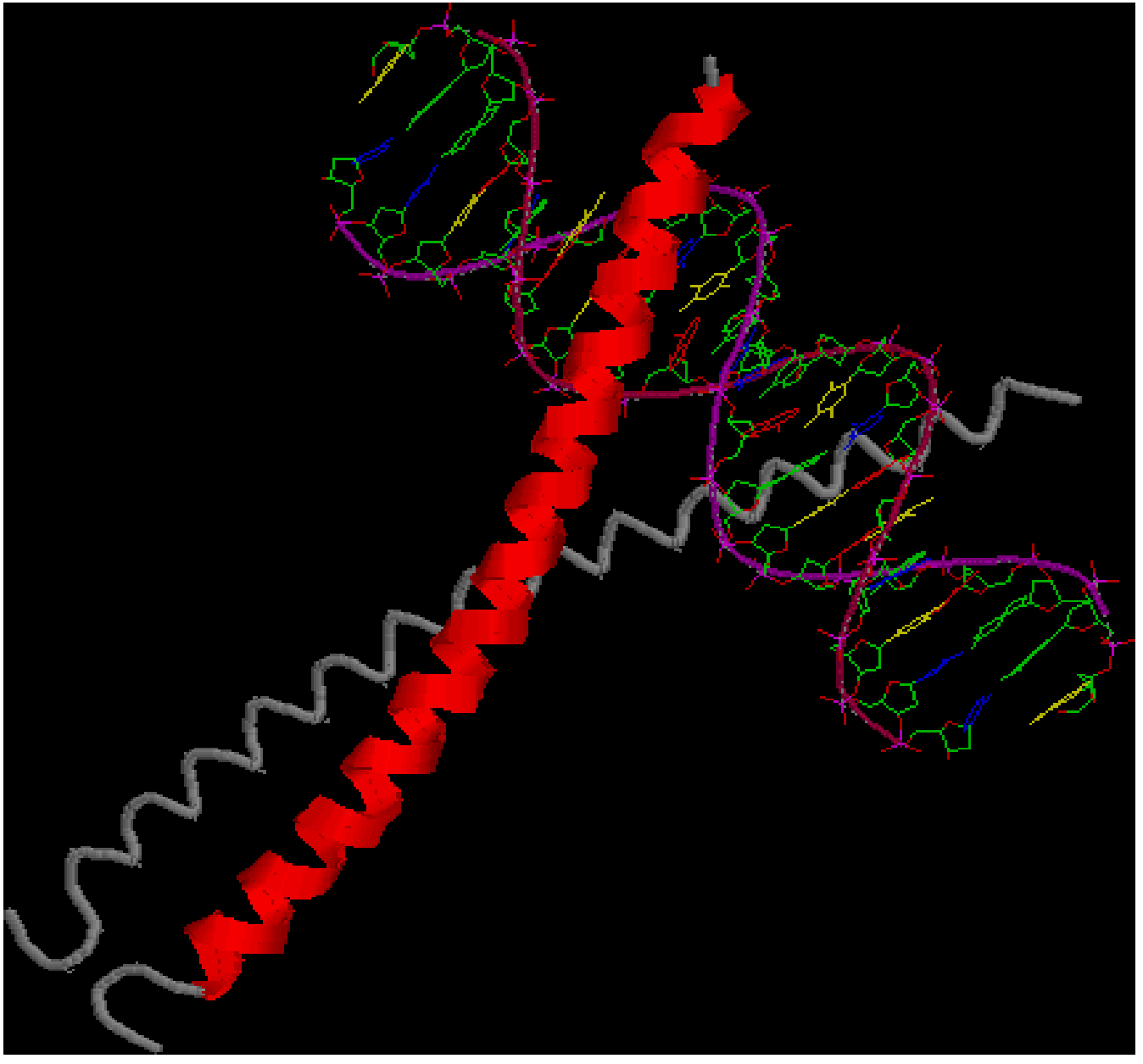
Let us look at transcription initiation



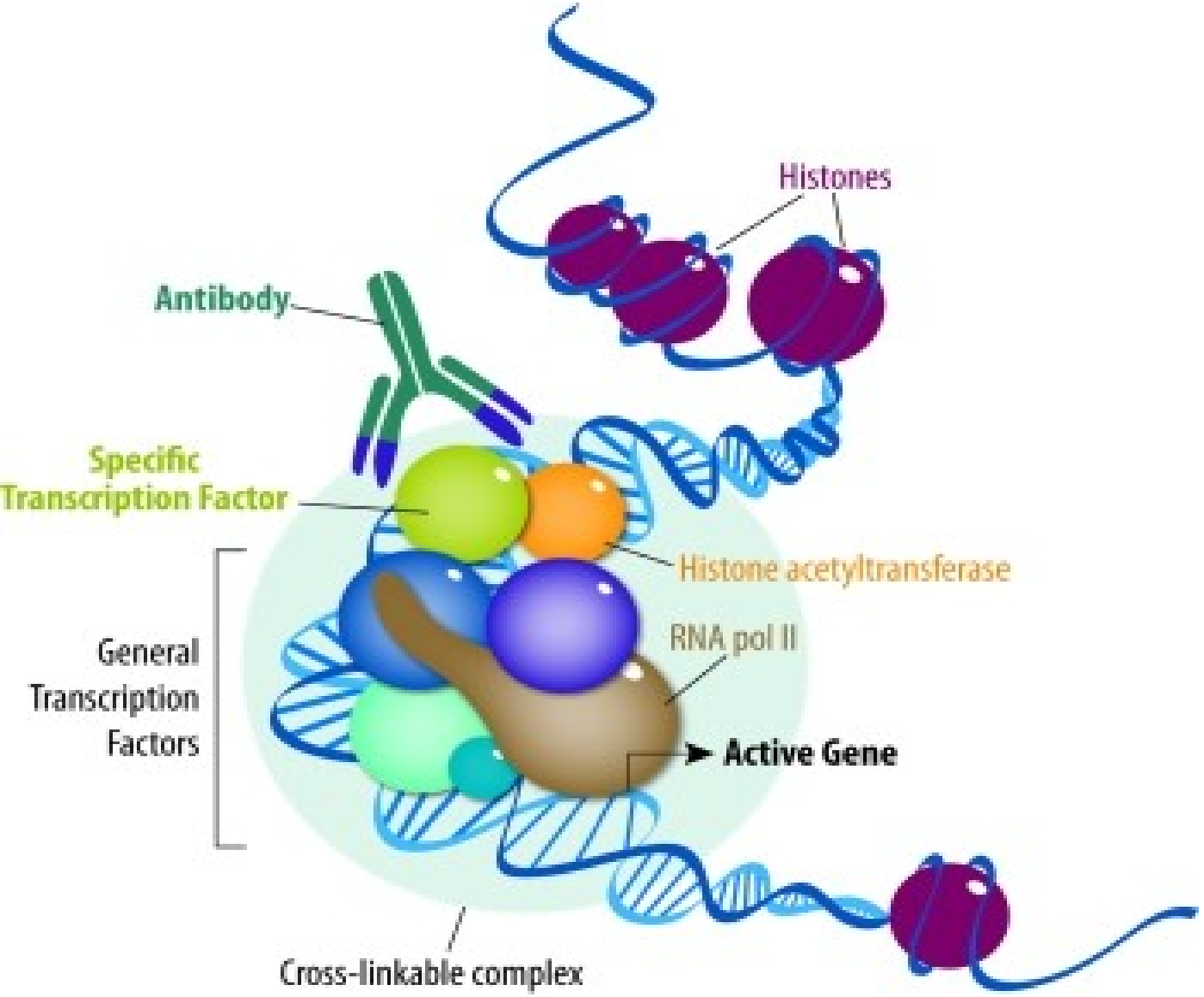
How does a promoter look like



Transcription factor binding



Discovering binding sites



TFBS representation

- Consensus words
- Regular expressions (IUPAC code)
- Position specific frequency Matrices (probability)
- Position specific Weight Matrices (log-odds)
- Higher Order Markov Models
- Bayesian Networks

Position specific matrices

- We can represent a sequence alignment by position specific matrix, which retains only the counts of symbols and not the correlations
- We can also compute Position specific probability matrix, normalizing all columns to 1

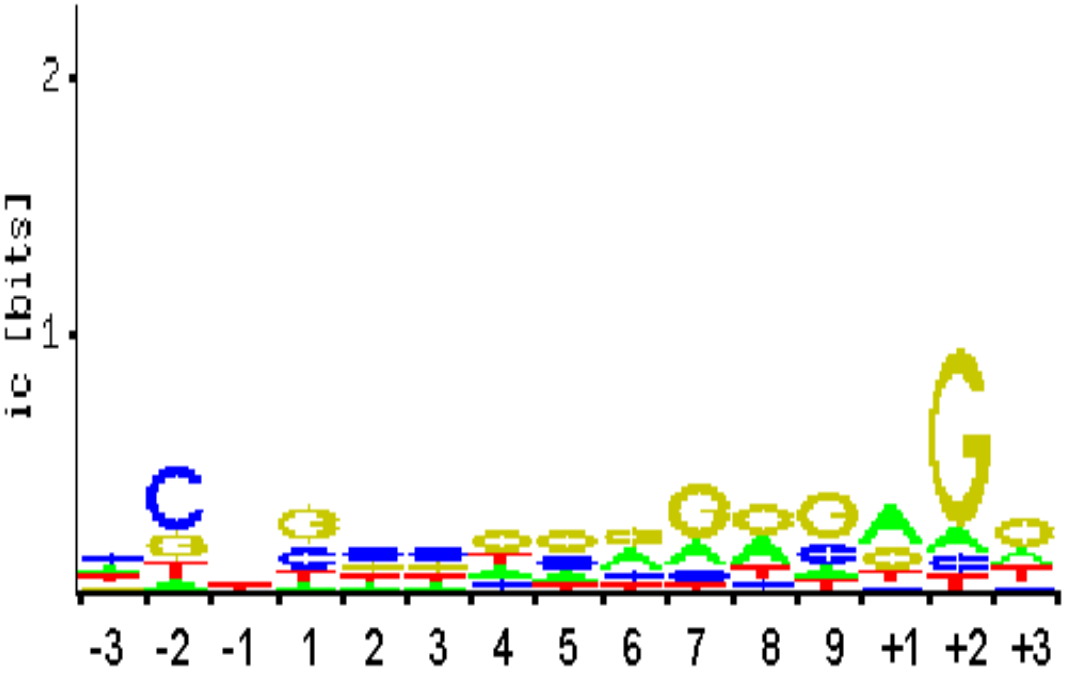
```
cacagtCGCGTGt
actatCGCGTGtt
actgttCGCGTGc
tctcatCGCGTG
aggaatCGCGTGc
gagtgtCGCGTG
aggggtCGCGTG
  ggatCGCGTGtcc
   atgTGCGTGaagg
 ttaggTGCGTG
      TGCGTGccacctc
  gagtTGCGTGc
 aaggtTGCGTGc
 aggtgTGCGTGc
 tagcgTGCGTGc
 tgattTGCGTG
 atggaTGCGTGc
   ggcTGCGTGacc
 cttatTGCGTGc
 tgggttAGCGTGc
 ttacttAGCGTGc
 gatccgGGCGTGa
tgggttaGTCGTG
 gtgtgaAGTTTGc
```

A [	3	0	0	0	0	0	]
C [	8	0	23	0	0	0	]
G [	2	23	0	23	0	24	]
T [	11	1	1	1	24	0	]

Log-odds and PWMs

- PSPMs can be improved by including background information and converting to an additive (log) scale, this is usually called a position weight matrix (PWM)
- We can score new sequences against a PWM, by summing up scores, it corresponds to calculating log-odds of observing a sequence from PWM against random model
- It is also important to correct for missing values, by adding pseudo-counts

We can find important positions



A:	21	7	21	9	13	12	18	15	34	23	35	12	46	10	20
C:	39	58	21	32	40	37	14	27	18	14	10	23	7	7	10
G:	19	22	22	46	24	30	47	44	37	55	42	53	31	77	49
T:	20	12	33	12	21	19	19	12	9	6	10	9	14	5	19

- Information content is defined as the entropy of the probability distribution
- Information content of a position in a PWM is the relative entropy of the motif compared to the background

Predicting Binding sites in genomes

- Using PWMs, we can score new sequences
- If we want to predict binding sites, we need to decide on the threshold on the score to find predicted sites (above the threshold sequences)
- This is in fact a problem in space of all words of certain length L
- We can estimate false positive/false negative rates of any threshold
- But the complexity grows exponentially with L

Motif databases

- There are two major eucaryotic TFBS motif databases: JASPAR and TRANSFAC
- TRANSFAC is commercial (BIOBASE GMBH)
- JASPAR is an open alternative
- They collect binding sites and matrices
- Provide tools for scanning sequences and motif comparisons

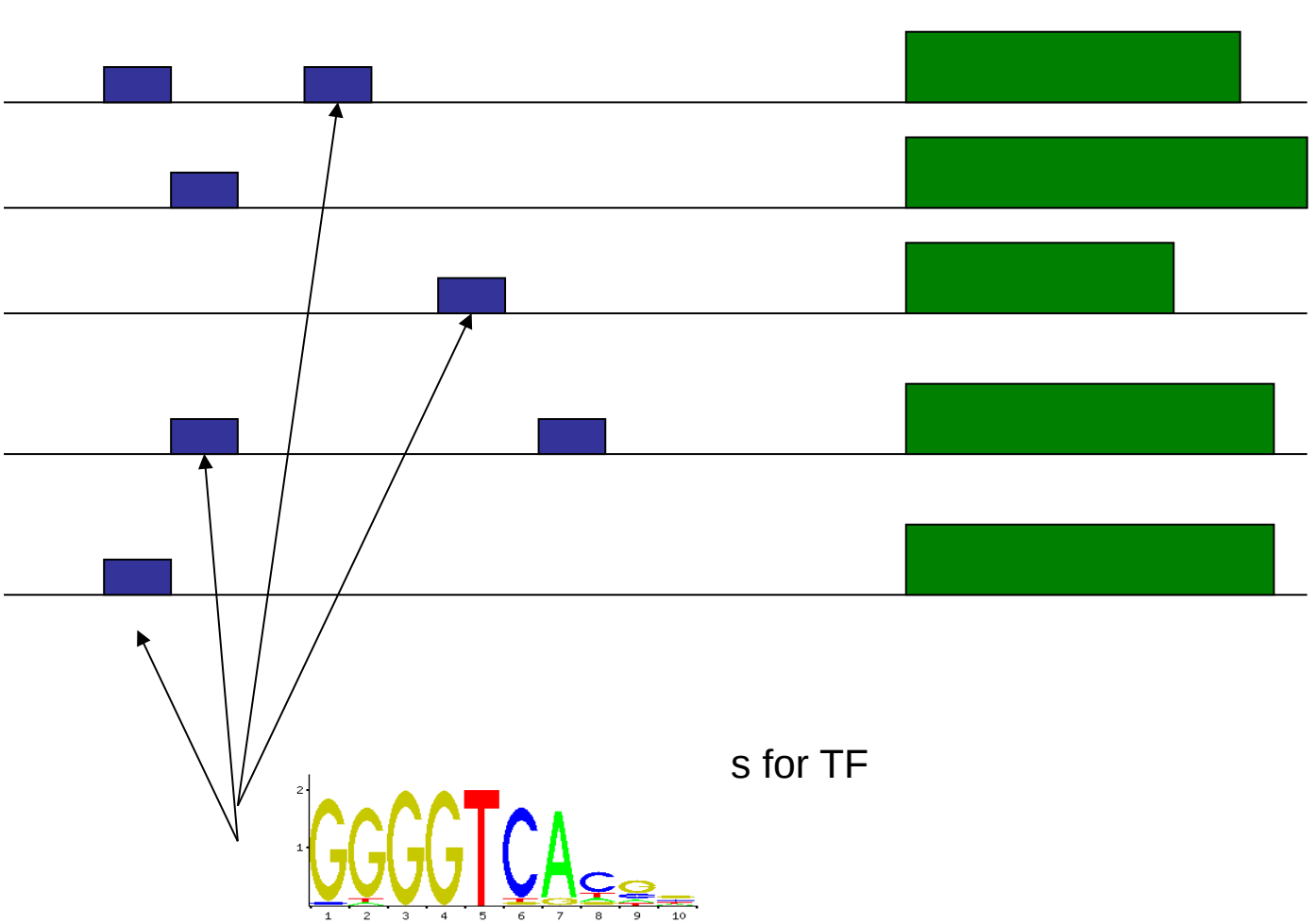
Comparing motifs

- It is important to be able to compare different motifs
- There are several approaches:
  - alignment based methods (like STAMP: gap-less alignment of PSSMs)
  - probability based (like MOSTA: probability of co-occurence of the motifs in random sequences
  - Distribution based: comparing motif score distributions across some sequences of interest (MEMOFinder)

Motif finding – problem statement

- We are presented with a number of experimentally defined regions (several hundred bases in size) bound by the factor of interest
- What is the most likely sequence motif bound by this factor? (motif of a given length 5-20bp)
- Hint: what are the positions of the supposed binding sites of that motif in input sequences?

Motif finding - illustration



Terminology

- Information content of the motif

$$\sum_i \sum_j p_{ij} log_2 \frac{p_{ij}}{b_j}$$

- log-odds of a motif occurrence

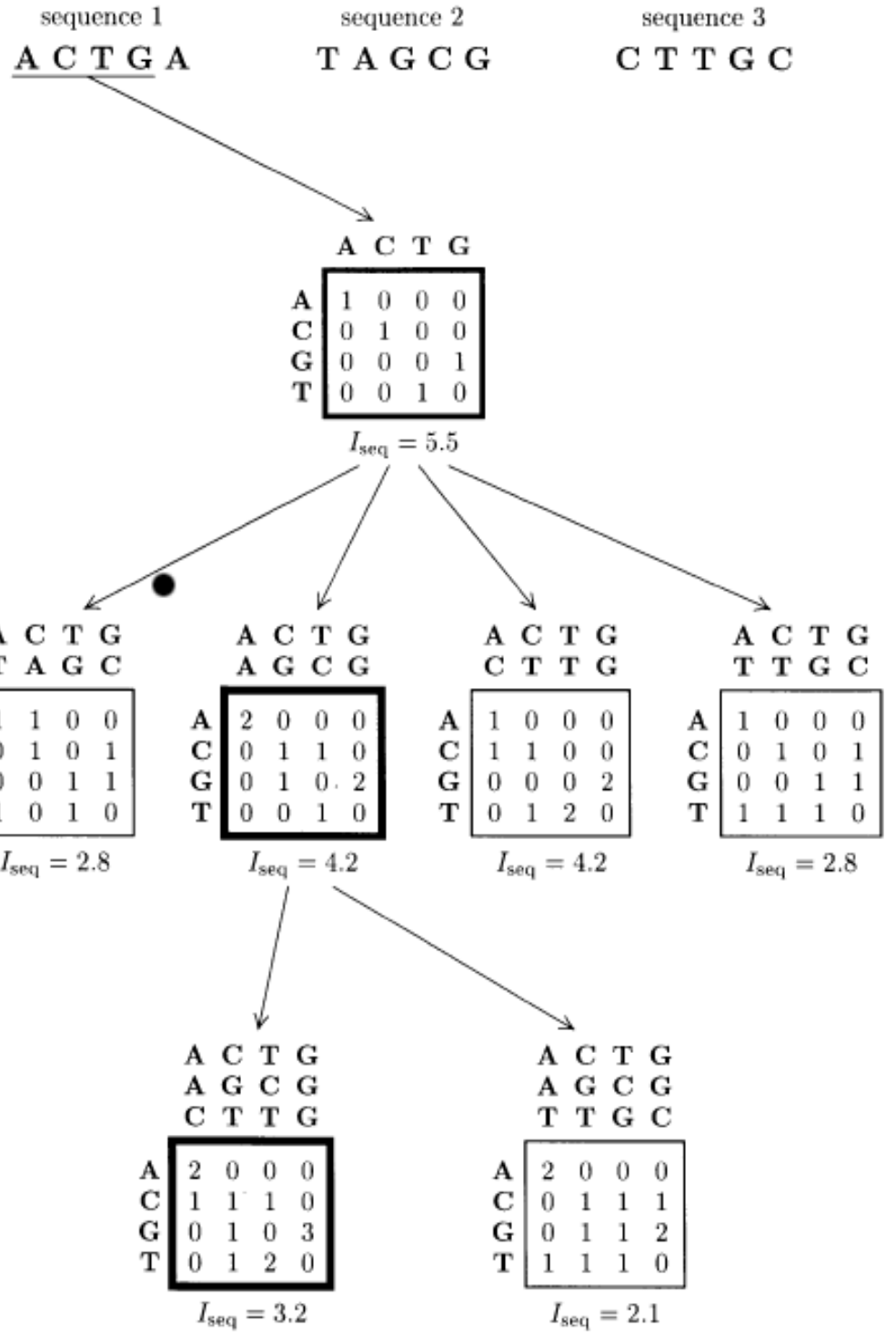
$$PWM(w) = \sum_j log_2 \frac{p_{w[j]}}{b_j}$$

Background models

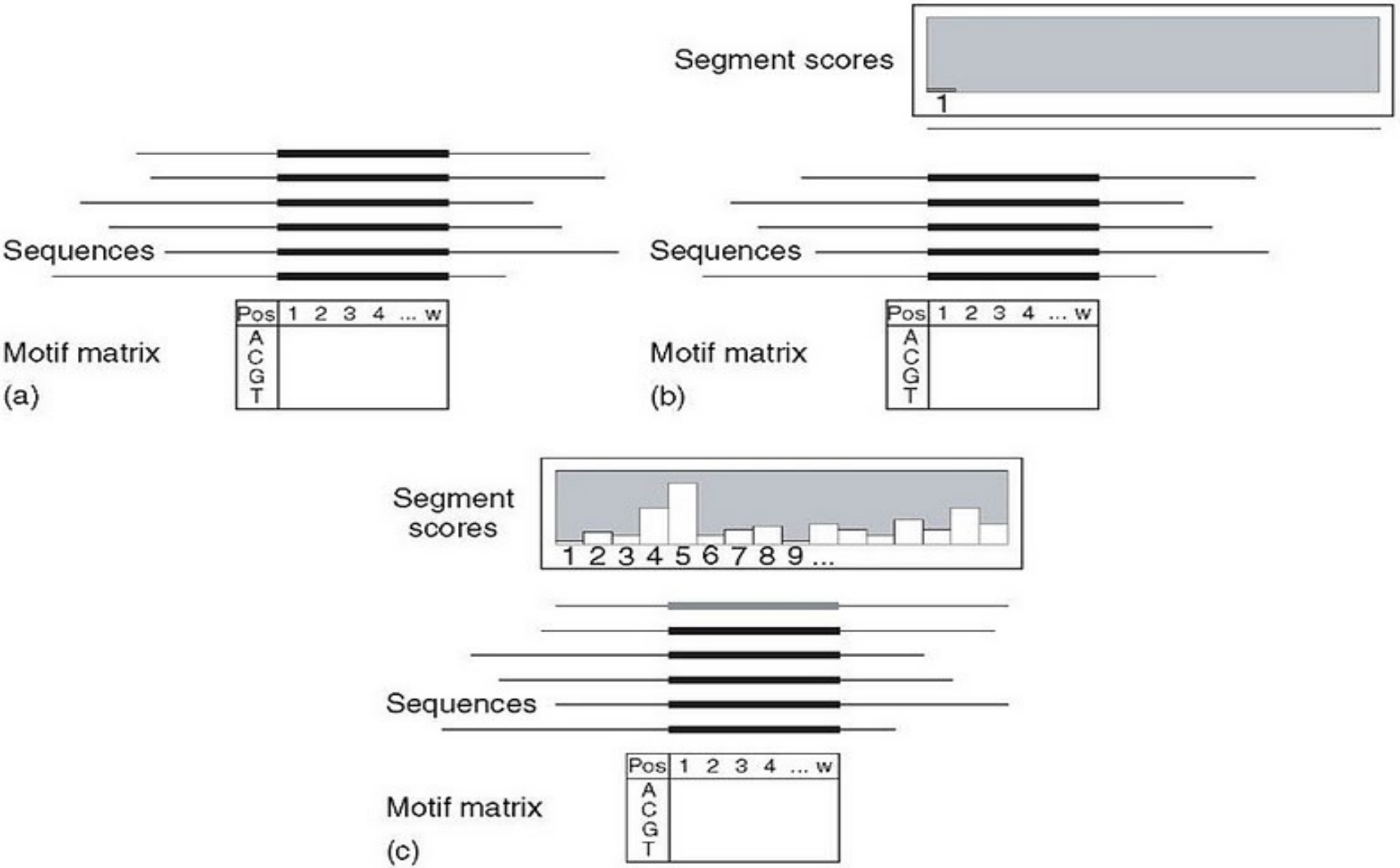
- Different models
  - Uniform nucleotide distributions (weak)
  - GC-content (most commonly used)
  - Higher order Markov models (strongest, but complex and dependent on surce sequences)
- Different types of sources
  - Whole genome
  - Non-coding sequences
  - Promoter-related
  - Input sequences

Consensus method

- Greedy algorithm
- Dependant on sequence ordering
- Take only a few “most informant alignments



Gibbs sampling



Motif Expectation Maximization (MEME)

S	unaligned set of sequences (training data) S_1, S_2, …, S_i, …, S_n each of length L
W	width of motif
Z	matrix of probabilities that the motif starts in position j in S_i
ρ	matrix representing the probability of character c in column k (the character c will be A, C, G, or T for DNA sequences or one of the 20 protein characters)
ε	epsilon value

```
1.  EM (S, W) {
2.      choose starting point and initial value for ρ
3.      do {
4.          re-estimate Z from ρ  //the estimation step
5.          re-estimate ρ from Z  //the maximization step
6.      } until (change in ρ < ε)
7.      return ρ , Z
8.  }
```