Comparative Genomics

2. Gene Prediction

Zvelebil Chapters 9, 10

Outline

- Introduction genes
- Gene prediction
  - ORFs
  - Prokaryote prediction
  - Eukaryote prediction
  - Experimental support
  - Database searches
- Promoter prediction
  - Prokaryotes
  - Eukaryotes

Introduction
why?

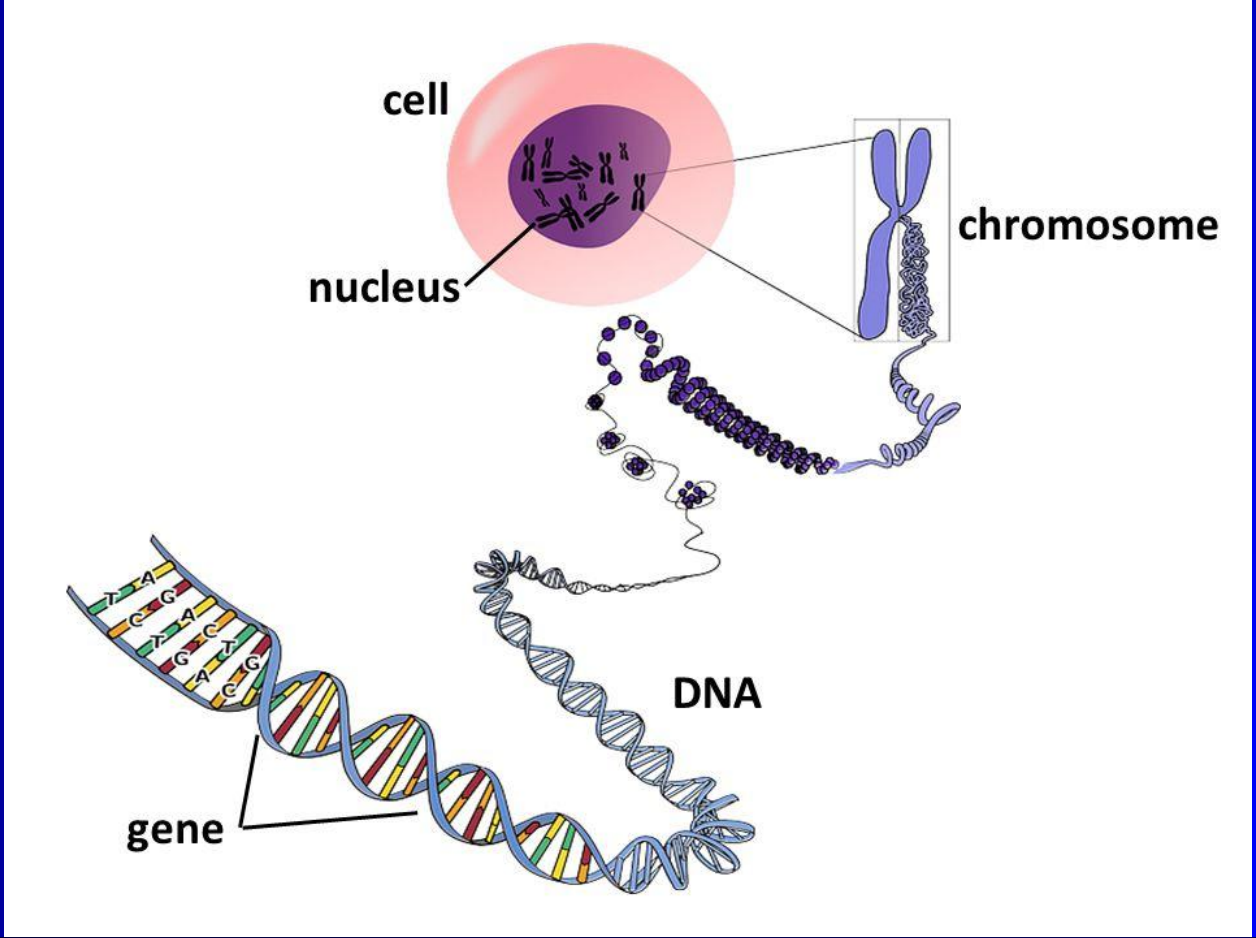
- What can be said based solely upon raw sequence?
- Why do we need special methods to do this?
- ASLDITALSKDJMIGHTPOWEIURBELERNBLK
  JDHARDLASKDJJDETOASKDGETJWOVNBRG
  JINFORMATIONASNREFROMAJKRACNERAW
  ALKEJFSEQUENCELKHFENLKNAMGIRASDF

Introduction
why?

- What can be said based solely upon raw sequence?
- Why do we need special methods to do this?
- ASLDITALSKDJMIGHTPOWEIURBELERNBLK
  JDHARDLASKDJJDETOASKDGETJWOVNBRG
  JINFORMATIONASNREFROMAJKRACNERAW
  ALKEJFSEQUENCELKHFENLKNAMGIRASDF

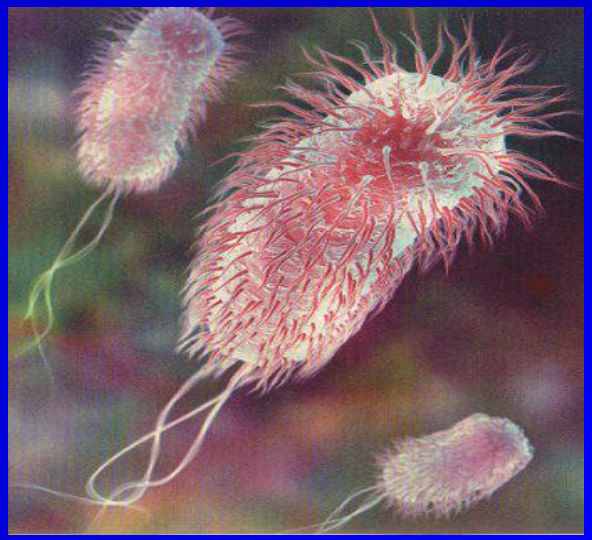
Introduction
Genes

- What is a gene ?

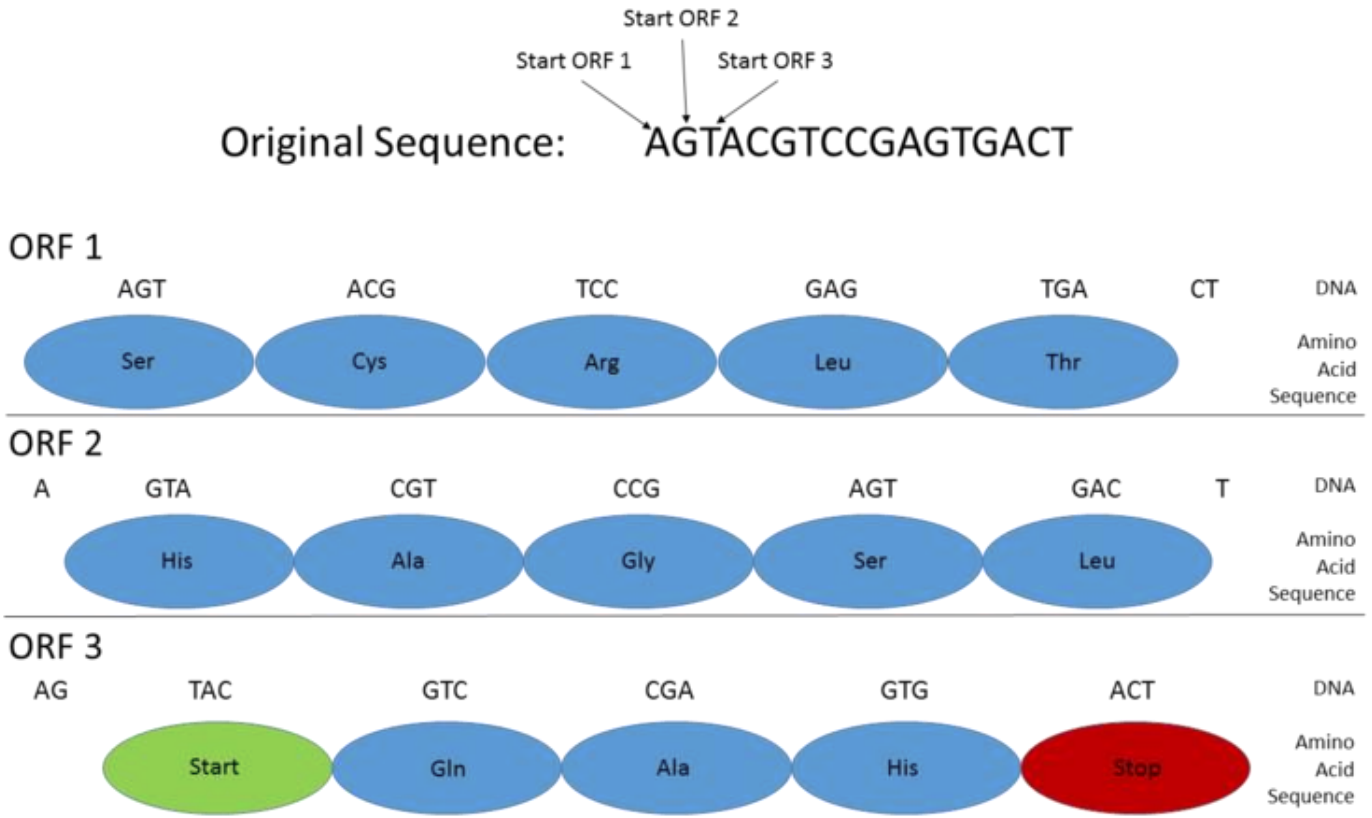


Introduction
Eukaryotic vs prokaryotic genes

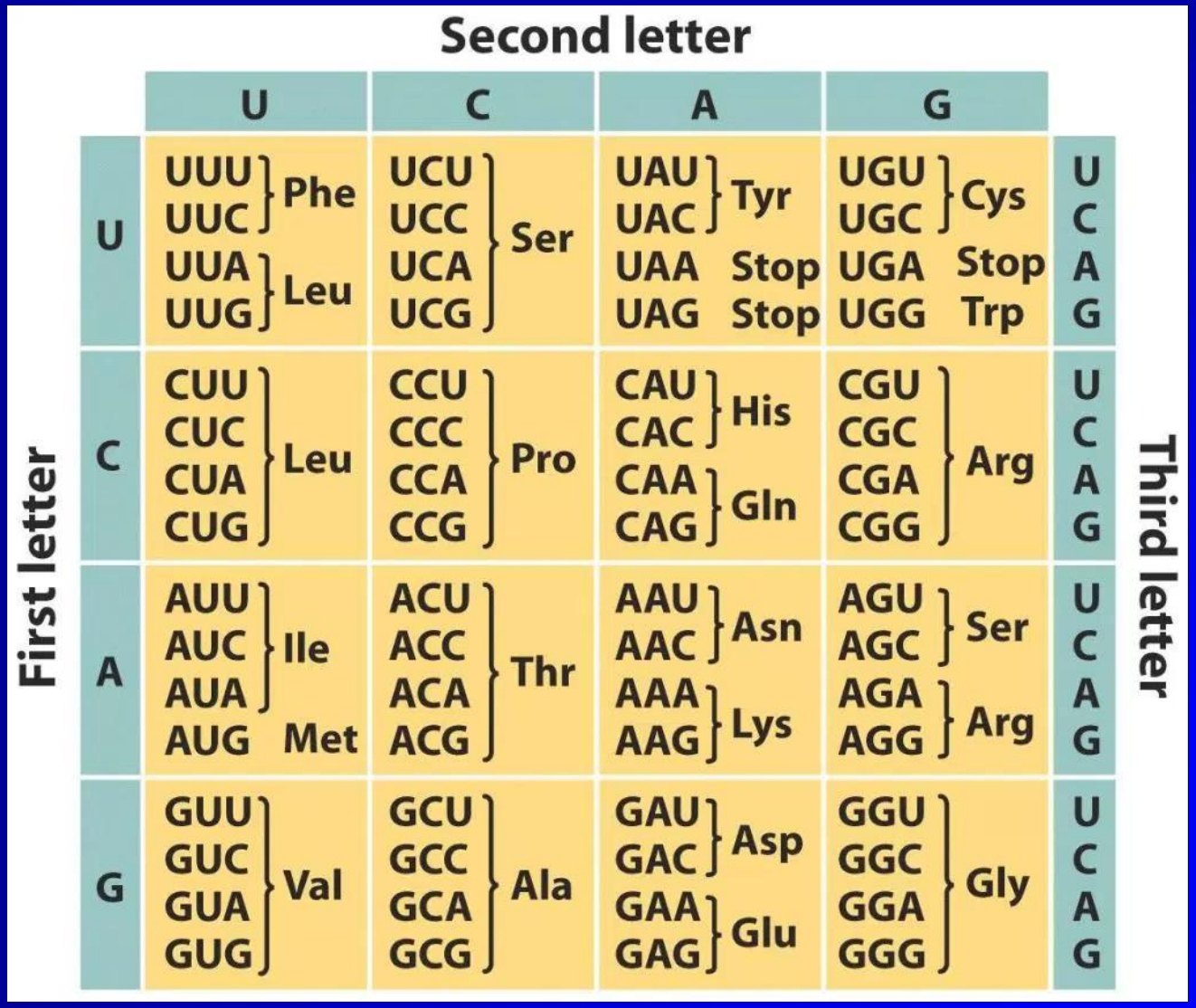
- What do they have in common?
- What is different?



ORFs - open reading frames

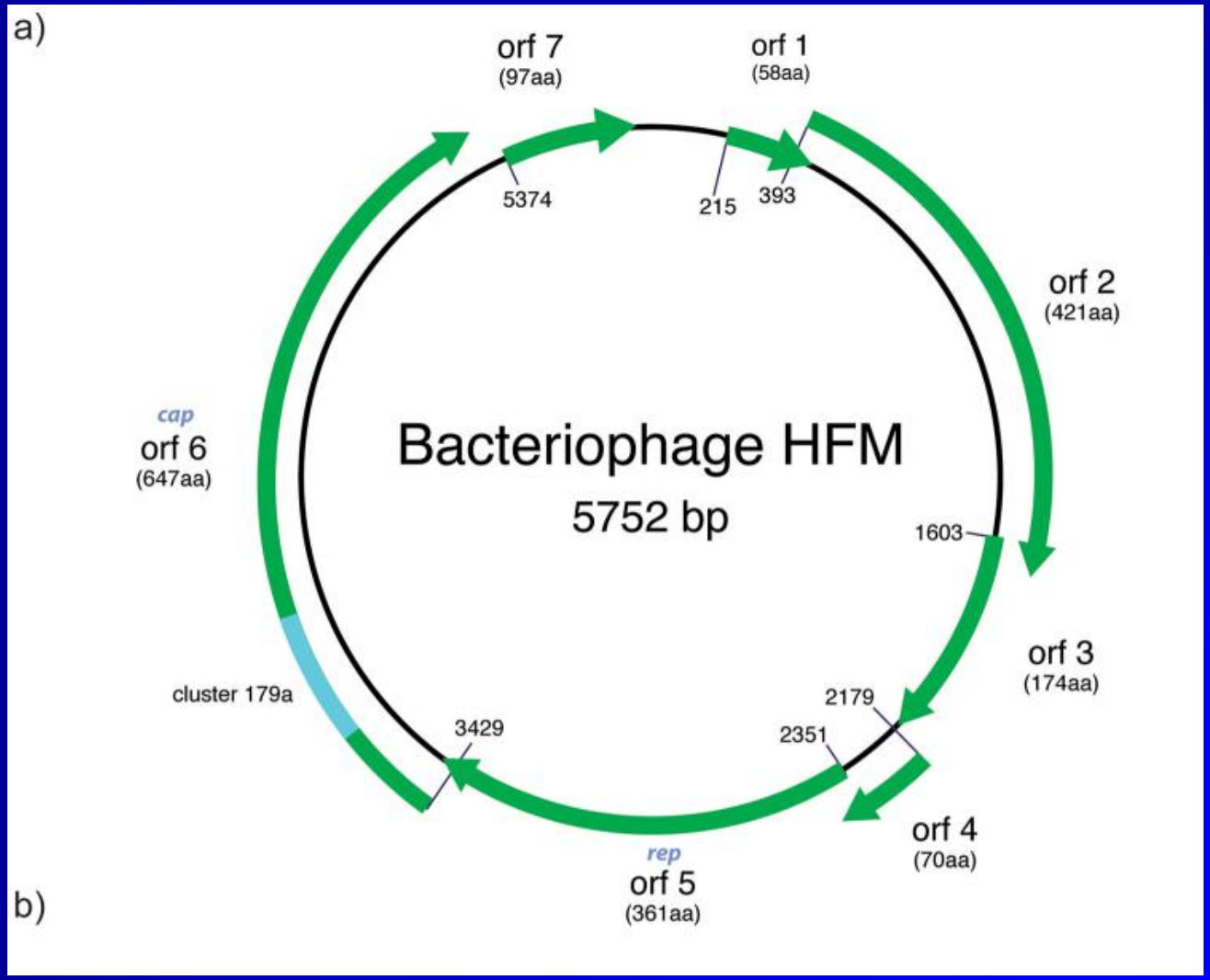


The genetic code



E. coli in Artemis (Zvelebil Figure 9.3)

ORFs in a virus genome



ORFs
Summary

- A stretch without stop codons is an Open Reading Frame.
- The ORF sequence is a list of codons from start to stop.
- Each species has characteristic pattern of use of synonymous codons, “codon bias”
- Different syn. codons often used in strongly versus weakly expressed genes.
- Organisms with high GC content have a bias towards G and C in the third codon position

Gene Prediction
Protein coding considerations

- Is the universal genetic code used?
- Is the mRNA edited?
- Is the mRNA chemically modified?

Gene Prediction
testing ORF quality

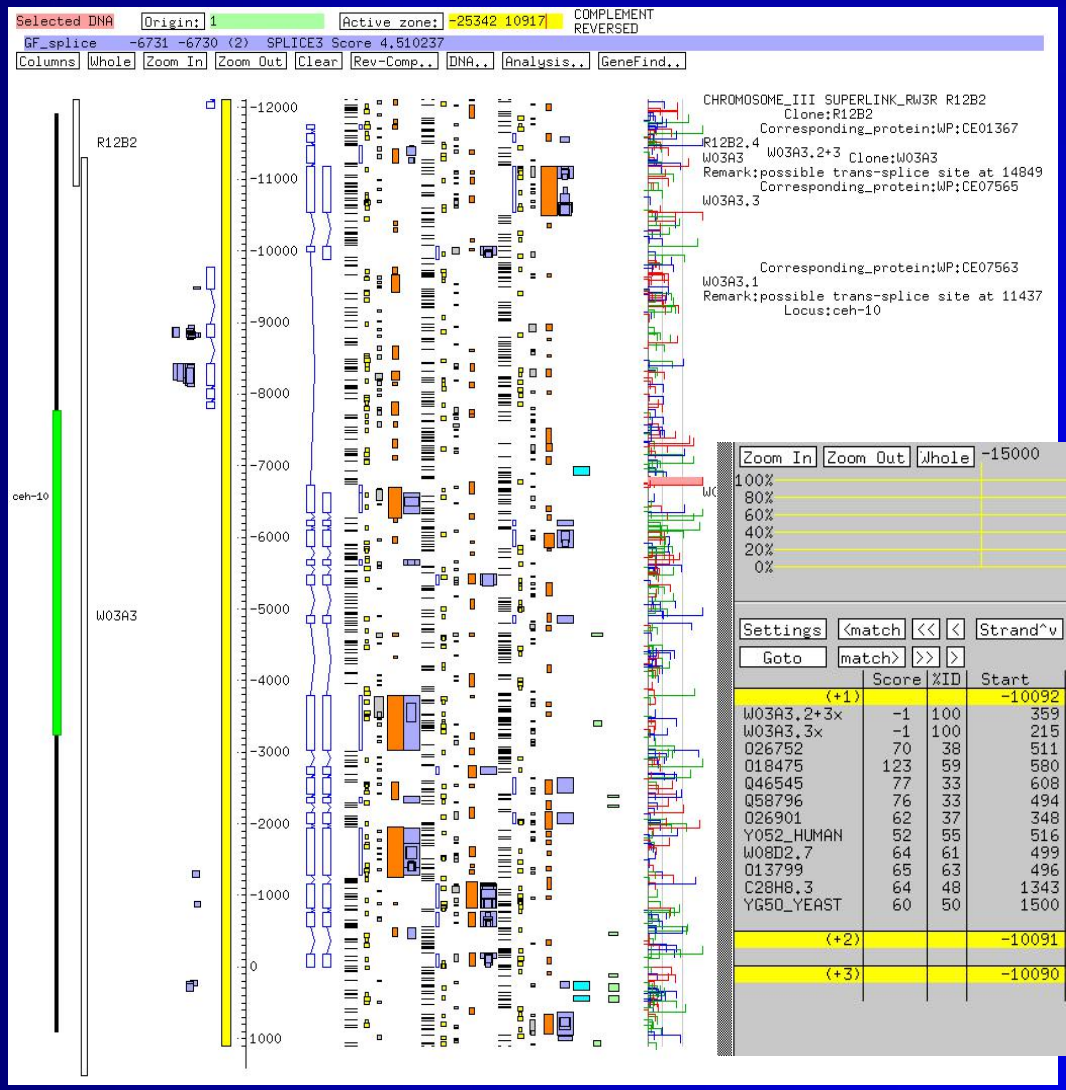
- Period three compositional bias
  - TESTCODE
- Compare codon usage in gene with average codon usage for organism
  - CODONFREQUENCY
- ORF translated into aa seq and compared with database of known proteins.

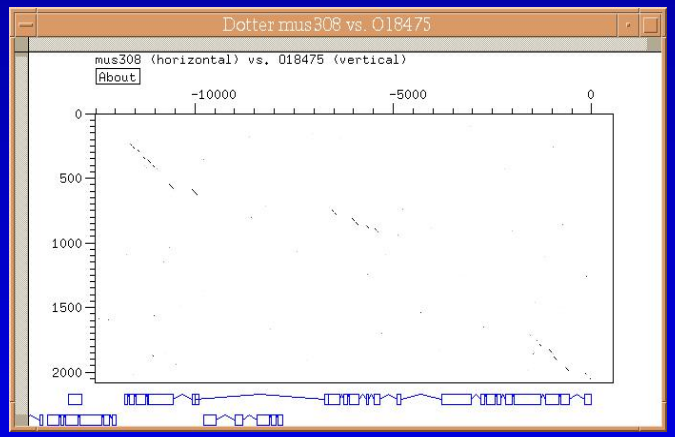
Gene Prediction
2 main approaches

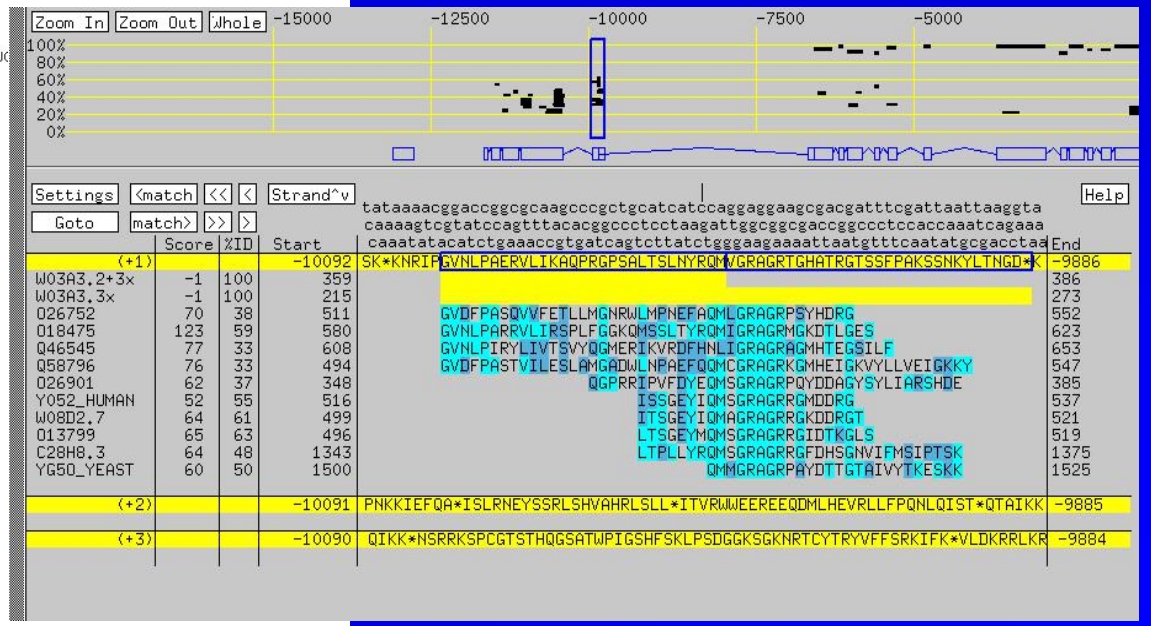
- Ab initio methods
  - Searching by sequence motifs, e.g Open Reading Frames, Promoters, Splice motifs, Breakpoints
  - Searching by content, e.g. different composition coding/non-coding regions

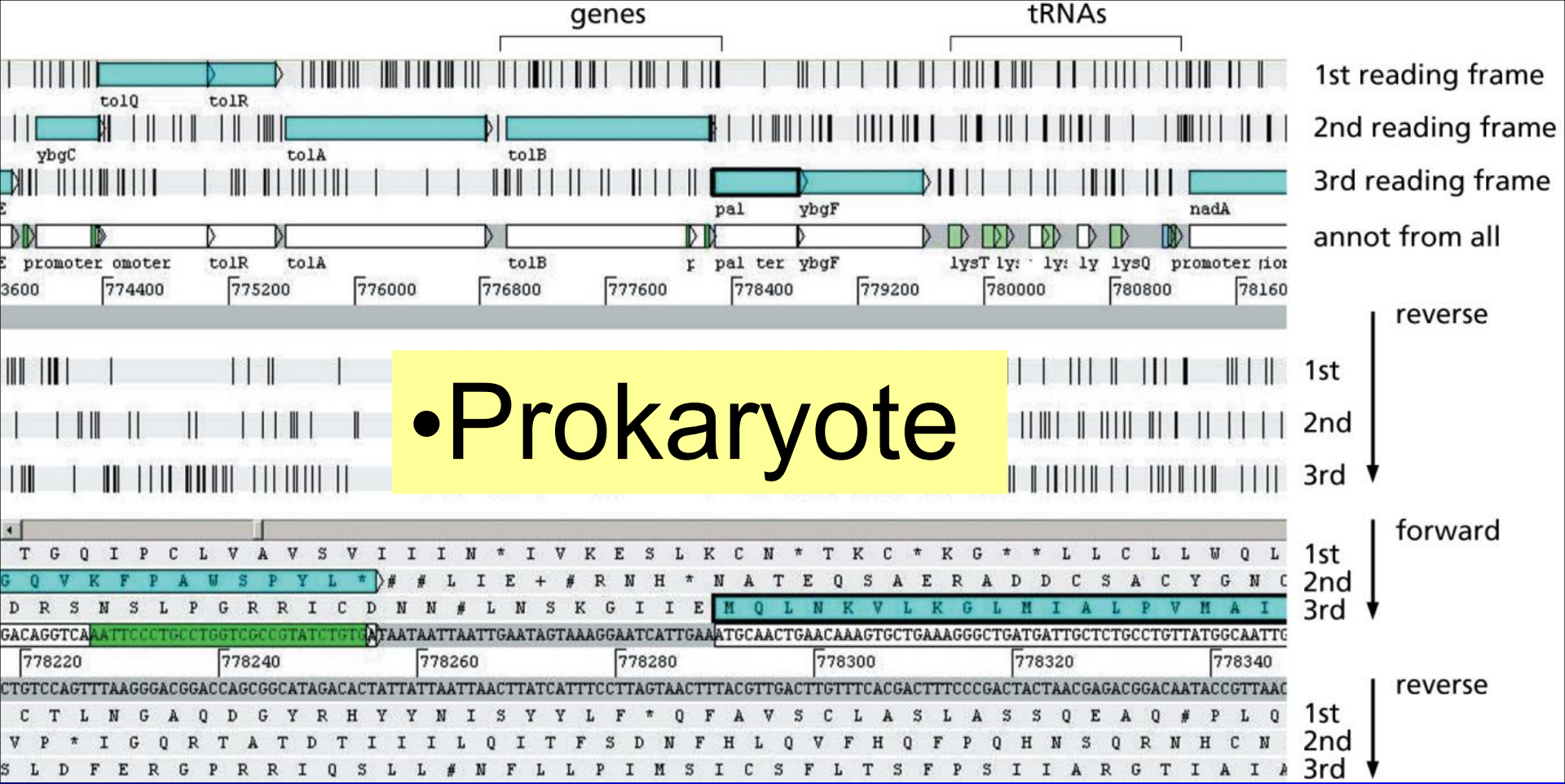
- Extrinsic evidence methods
  - E.g. sequence match to known gene, EST, or protein
  - E.g. similarity to aligned closely related genomes

Gene prediction in ACEDB

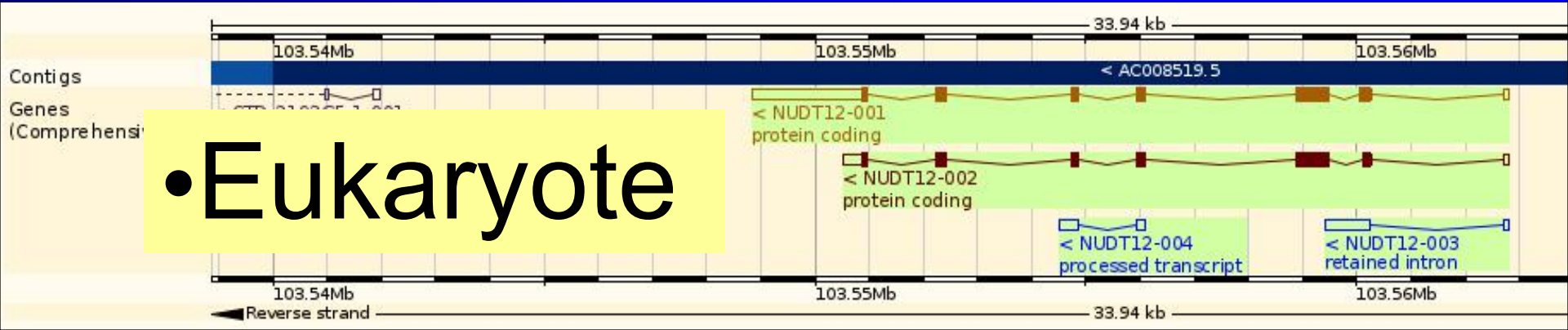








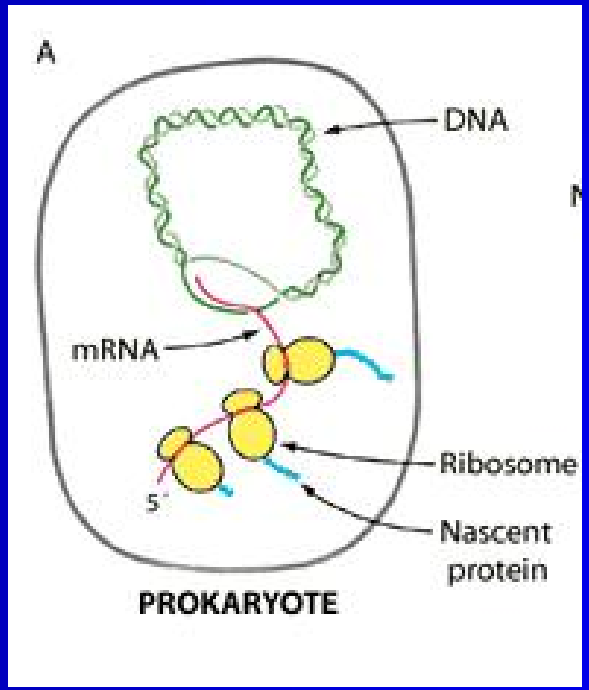
•Prokaryote



•Eukaryote

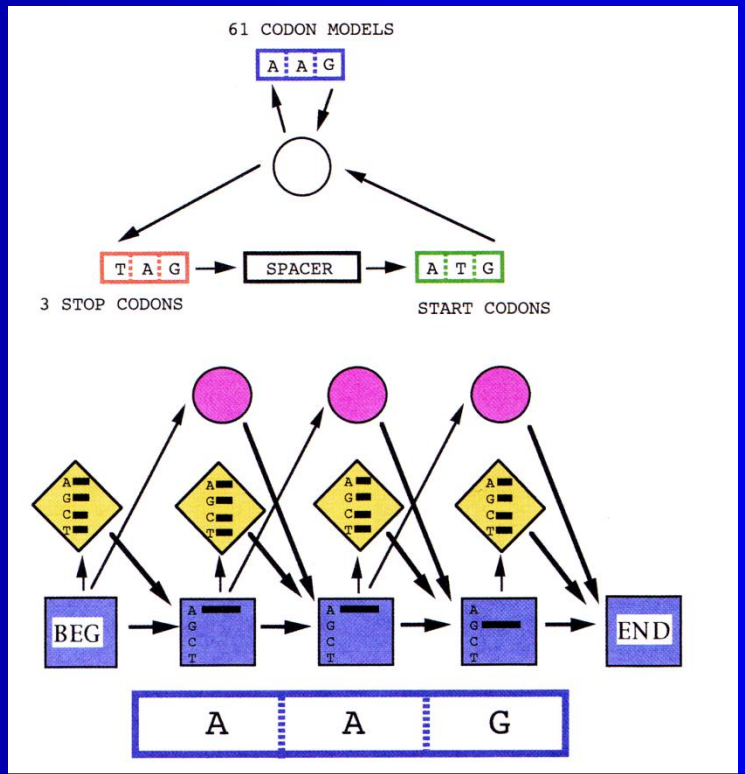
Gene Prediction
Prokaryotes

- Usually no sequence modification from DNA -> mRNA -> Protein
- Simply find the longest ORF from start- to stop codon



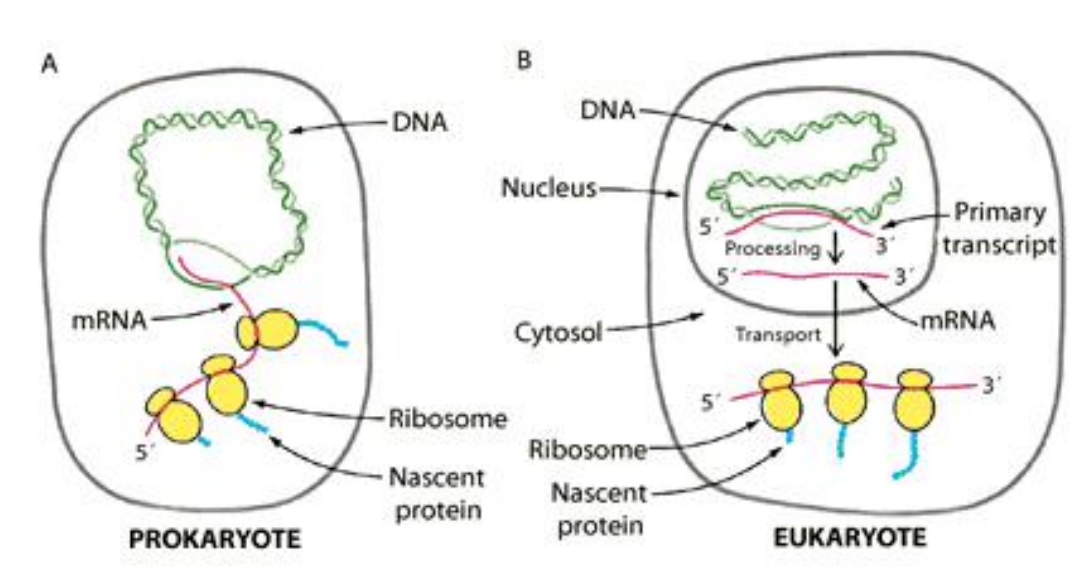
Gene Prediction
Prokaryote prediction tools

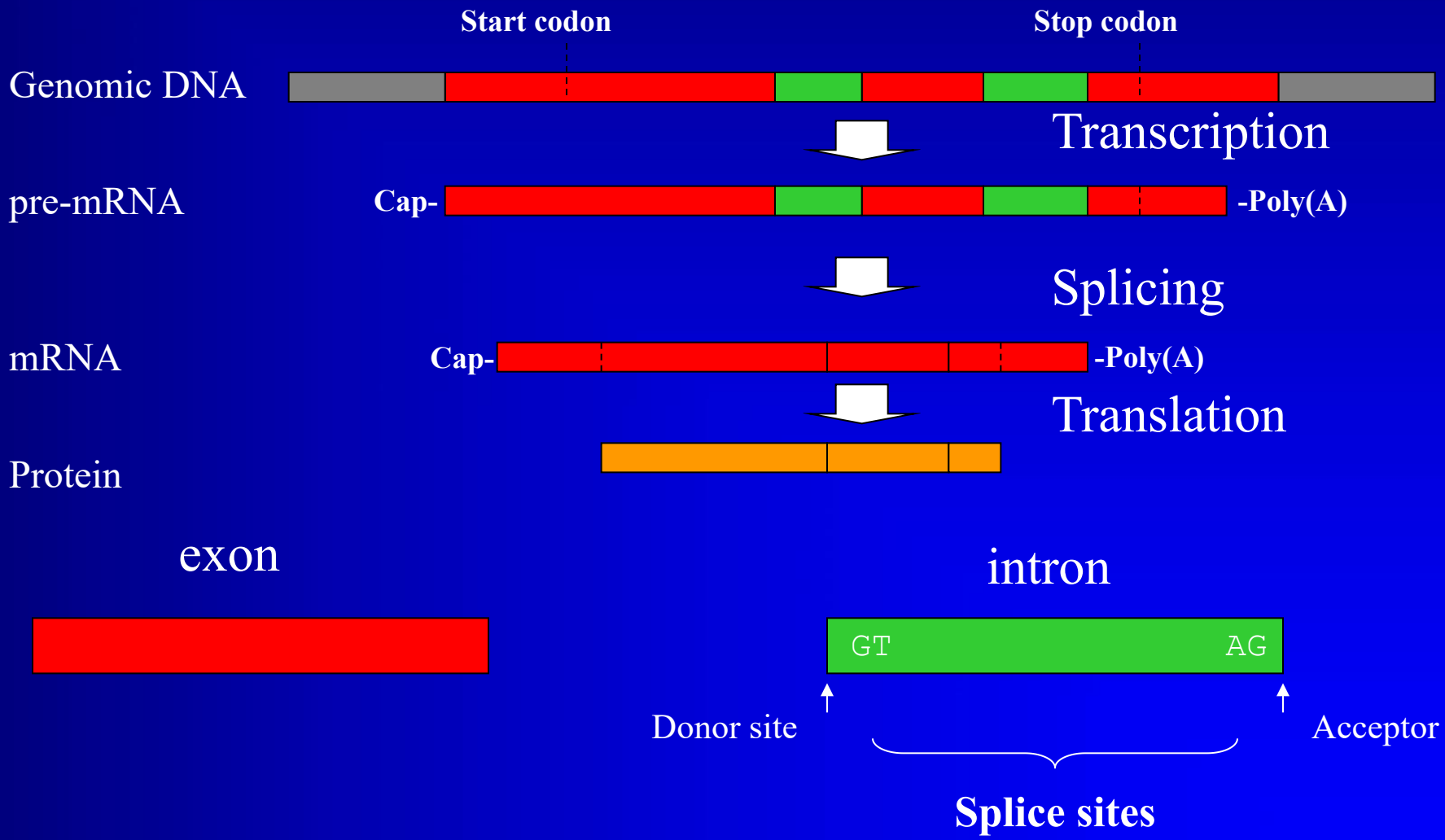
- Use ORF, coding potential, promoter signals
- Hidden markov models perform best
  - GeneMark.HMM, Glimmer (Interpolated Markov model)



Gene Prediction
Eukaryotes

- More complex gene structure
- Not just a matter of finding longest ORF
- Complex exon/intron structure

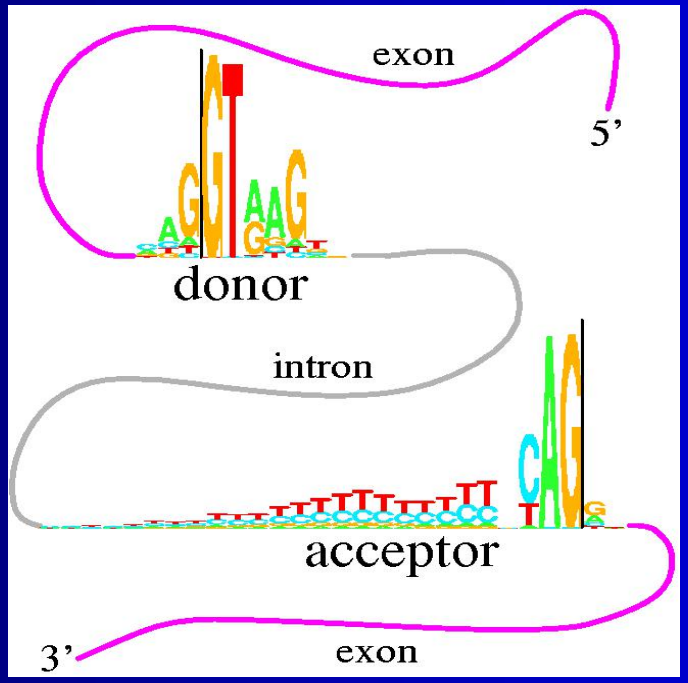


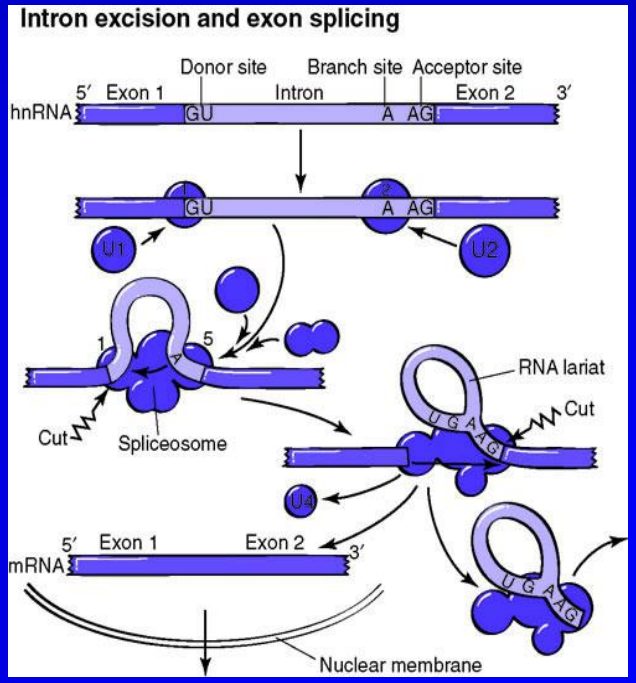


(Brendel 2005)

Gene Prediction
splicing

- Intron signals
  - Donor
  - Acceptor

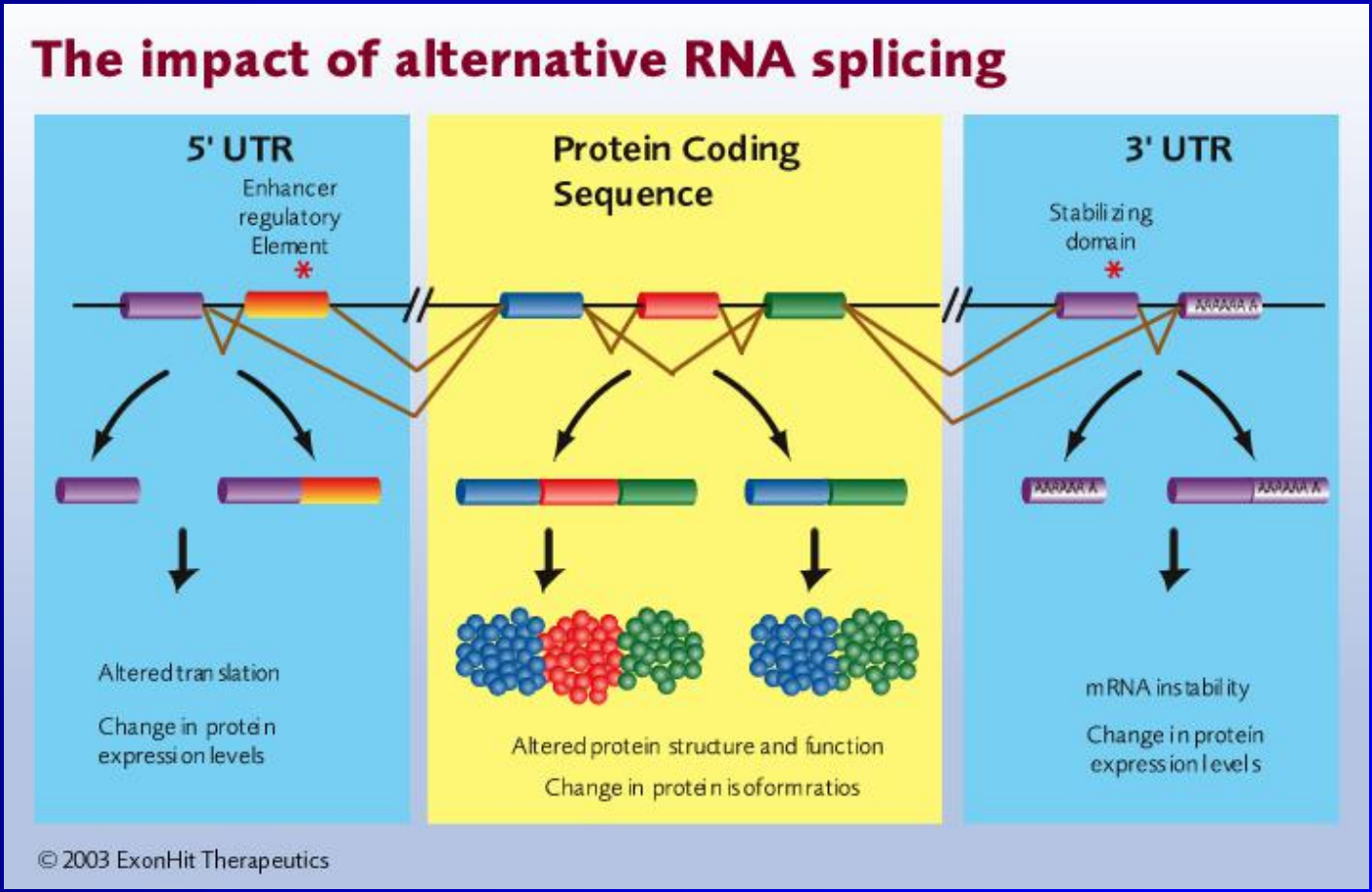


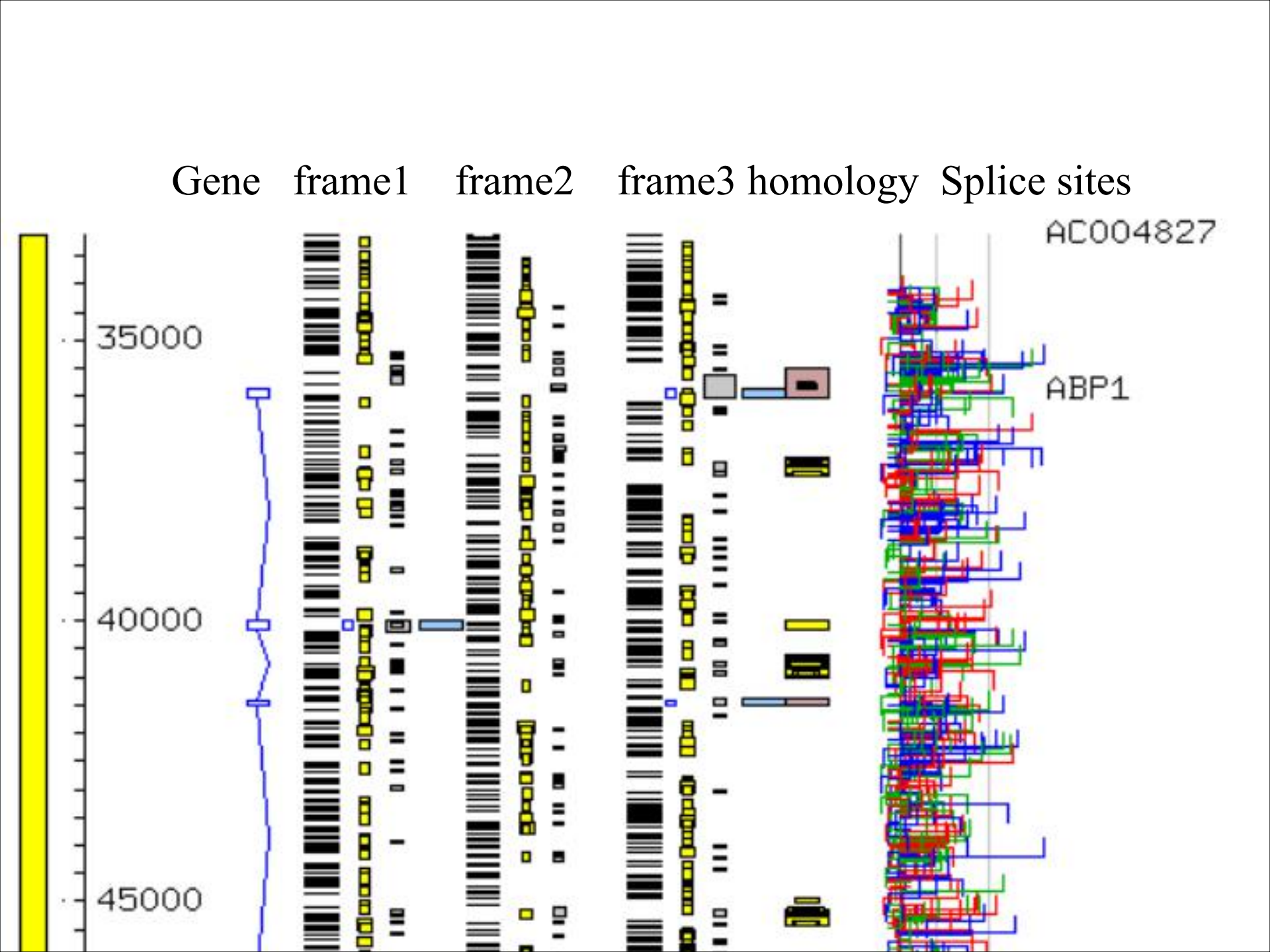


Introns and alternative splicing

- Complex eukaryotes have their genes split up into a series of exons (sometimes hundreds) separated by intron sequences.
- After transcription, a special machinery (splicing) removes the introns.
- This machinery may also remove exons selectively, so the same gene can give rise to different final proteins: ***alternative splicing.***
- Alternative splicing and more complex regulatory systems, enabling more fine-grained reaction patterns, may explain why complexity grows faster than gene number.

Alternative splicing




Gene
frame1
frame2
frame3
homology
Splice sites
AC004827
ABP1
35000
40000
45000

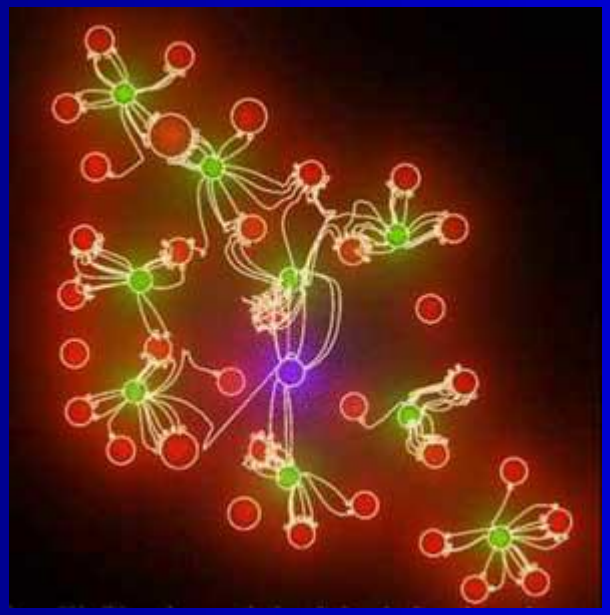
Gene Prediction
Eukaryote predictions

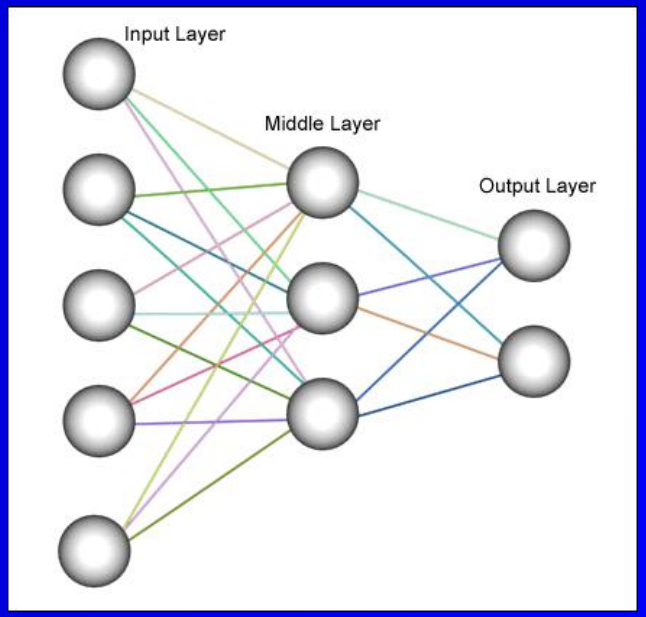
- Find the difference between coding/noncoding regions
  - Compositional bias
  - Uneven use of synonymous/nonsynonymous codons

- HMM
  - E.g. Genscan
- Neural networks
  - E.g. GRAIL, GeneParser
- Pattern discriminating methods
  - E.g. HEXON, FGENEH, MZEF

Gene Prediction
Neural networks

- Nodes read the sequence, perform evaluation and pass on information to lower layers
- Final node makes a prediction
- Requires training on known genes

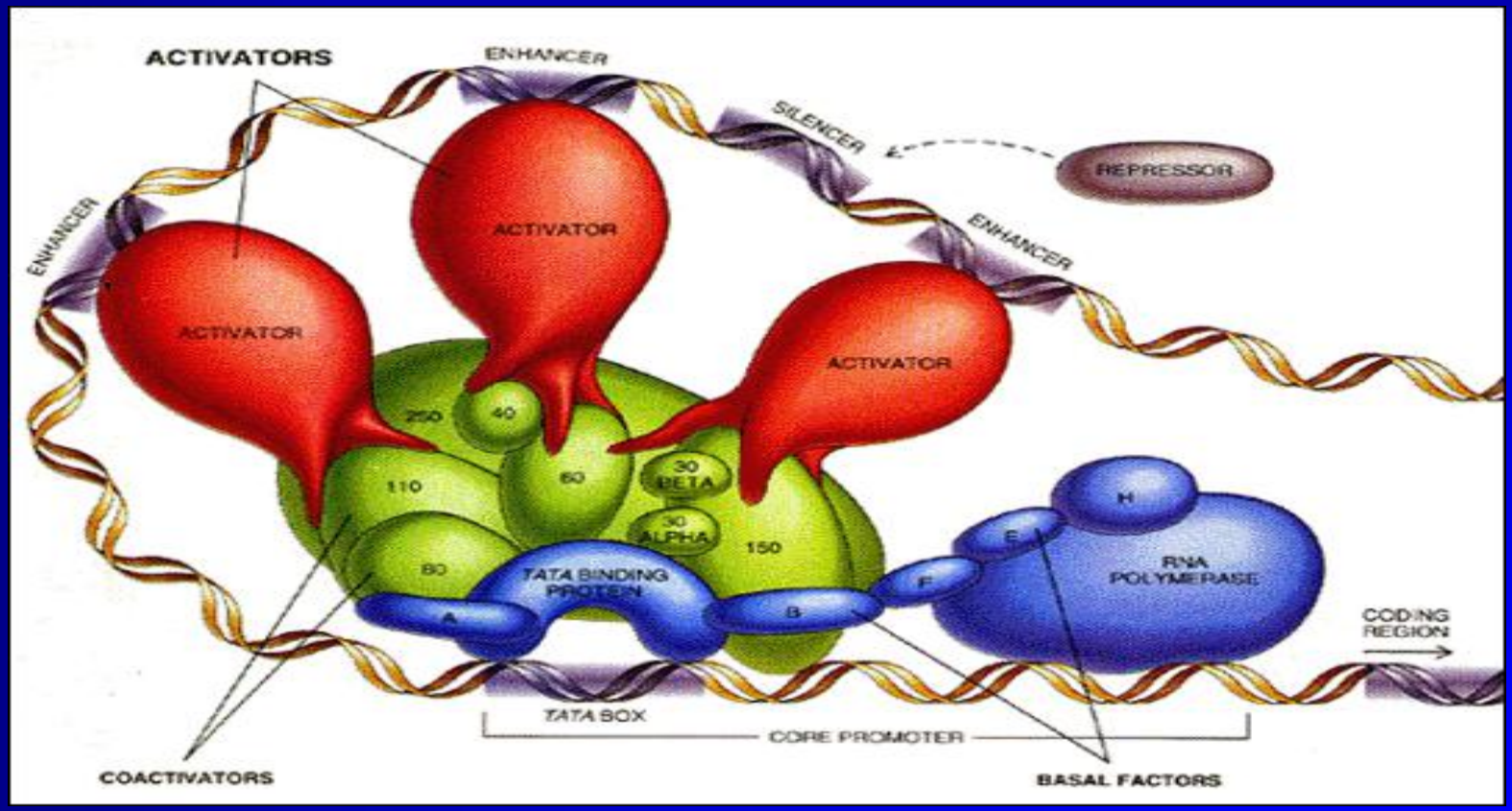




Gene Prediction
Sequence database searches

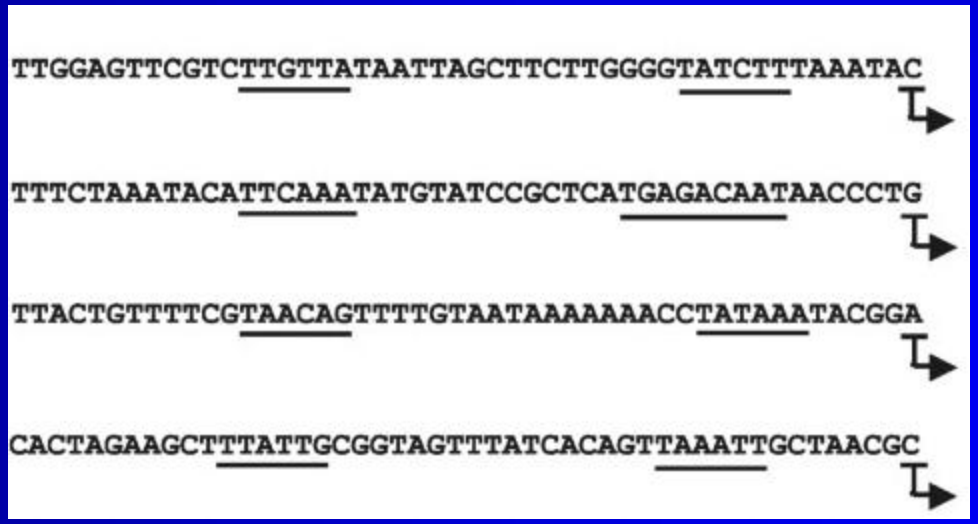
- I: Find proteins/genes in a DNA sequence
  - Translate sequence in all reading frames
  - Search protein db using e.g. blastx
- II: Scan DNA sequence for a particular protein
  - Search for a protein against DNA sequence (db) translated in all reading frames with e.g. tblastn

Promoter prediction



Promoter prediction
Prokaryotes

- Characteristic promoters from *E.coli*
  - consensus TATAAT at position -10
  - consensus TTGACA at position -35
  - AT rich region before the -35 region



Promoter prediction
Methods

- Conserved patterns
  - Align promotor regions
  - Check for conserved patterns
  - Check query sequence for these patterns
- Scoring matrix
  - Align promotor regions
  - Count base frequencies for each column and convert to log odds scores in a matrix
  - ”Slide” matrix over query sequence and calculate score
- Neural Networks
- Works well when the regions are well conserved

Promoter prediction
Statistical methods

- Expectation maximization
  - Initial scoring matrix from guessed alignment
  - Scan sequences with matrix
  - Update scoring matrix with sequence pattern found at each position weighted by probability of match to position

- HMMs

Promoter prediction
Eukaryotes

- Predicting promotors not as easy in eukaryotes

- NN trained on TATA and Inr sites/NN-genetic alg. To identify conserved patterns and spacings in RNAPII promoters
- Recognition of TATA with weight matrix and an analysis of the density of TF sites
- Linear discriminant model using features of promoter sequences
- Weight matrixes from different organism against query sequence
- Evaluation of query sequence for presence of clustered groups or modules of TF-binding sites that are characteristic of a given pattern

Conclusions

- Gene prediction in prokaryotes is relatively easy
  - Simple methods are often sufficient; more advanced methods give a modest improvement
- Introns in eukaryotes make gene prediction more difficult
  - For lower organism this is still doable with more advanced methods
  - Higher organisms have more and longer introns making prediction even harder
- The need for experimental support/validation increases with genome complexity.