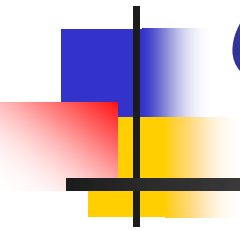


Computational gene finding

Devika Subramanian

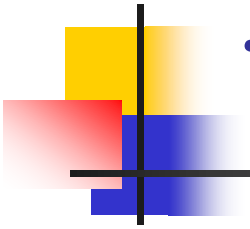
Comp 470

Outline (3 lectures)

Lec 1
- The biological context
- Markov models and Hidden Markov models

Lec 2
- Ab-initio methods for gene finding
- Comparative methods for gene finding

Lec 3
- Evaluating gene finding programs

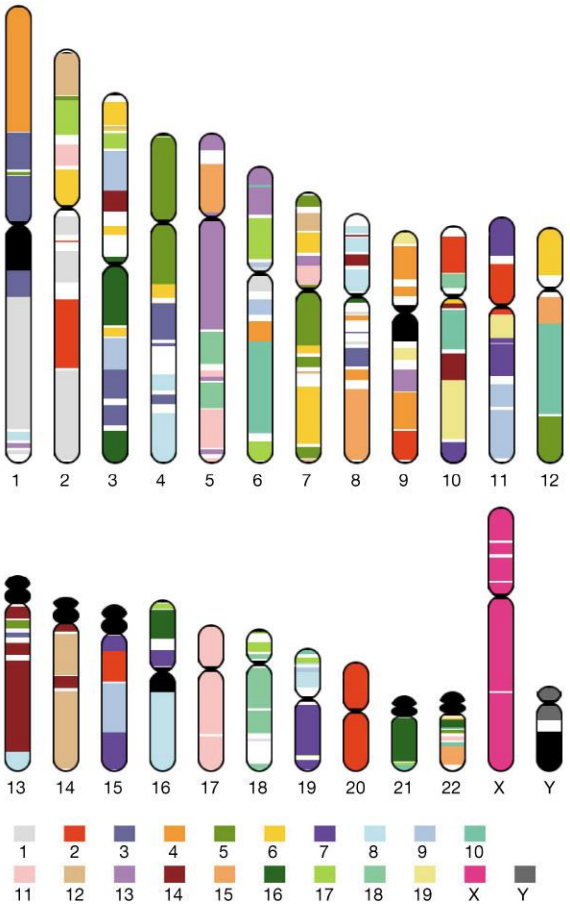


The biological context

- Introduction to the human genome and genes
- The central dogma: transcription and translation

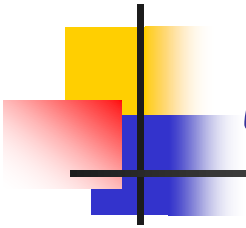
Facts about the human genome

- The human genome contains 3 billion chemical nucleotide bases (A, C, T, and G).
- About 20,500 genes are estimated to be in the human genome. Chromosome 1 (the largest human chromosome) has the most genes (2968), and the Y chromosome has the fewest (231).



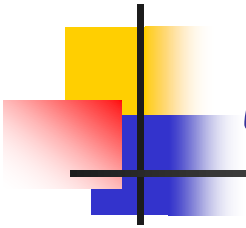
http://www.pnas.org/content/104/49/19428.abstract

~17,000 genes confirmed experimentally/Ensembl2009



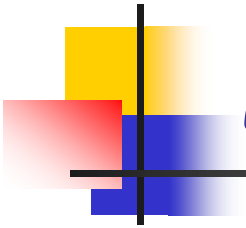
More facts

- The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.



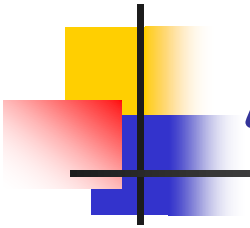
More facts

- Genes appear to be concentrated in random areas along the genome, with vast expanses of non-coding DNA between.
- About 2% of the genome encodes instructions for the synthesis of proteins.
- We do not know the function of more than 50% of the discovered genes.



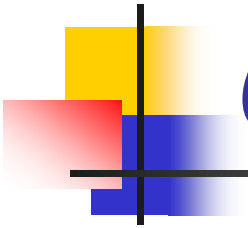
More facts

- The human genome sequence is almost (99.9%) exactly the same in all people. There are about 3 million locations where single-base DNA differences occur in humans (Single Nucleotide Polymorphisms or SNPs).
- Over 40% of the predicted human proteins share similarity with fruit-fly or worm proteins.



A great site to learn more

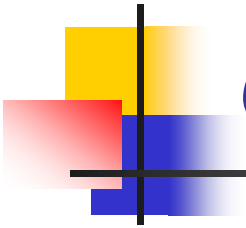
http://www.dnai.org/index.htm



Genome sizes

Organism	Genome Size (Bases)	Estimated Genes
Human (*Homo sapiens*)	3 billion	20,000
Laboratory mouse (*M. musculus*)	2.6 billion	20,000
Mustard weed (*A. thaliana*)	100 million	20,000
Roundworm (*C. elegans*)	97 million	19,000
Fruit fly (*D. melanogaster*)	137 million	13,000
Yeast (*S. cerevisiae*)	12.1 million	6,000
Bacterium (*E. coli*)	4.6 million	3,200
Human immunodeficiency virus (HIV)	9700	9

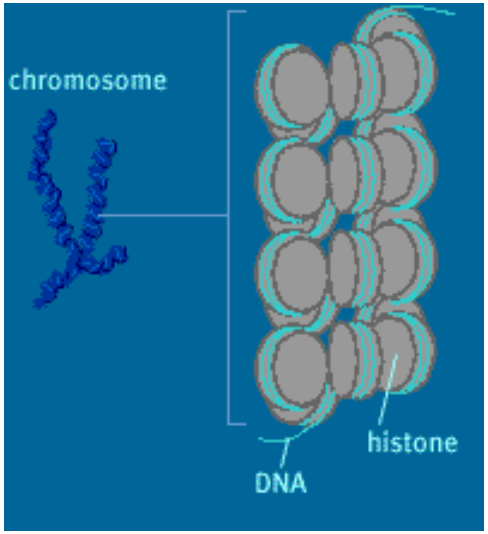
No correlation between amount of DNA in a species and number of genes

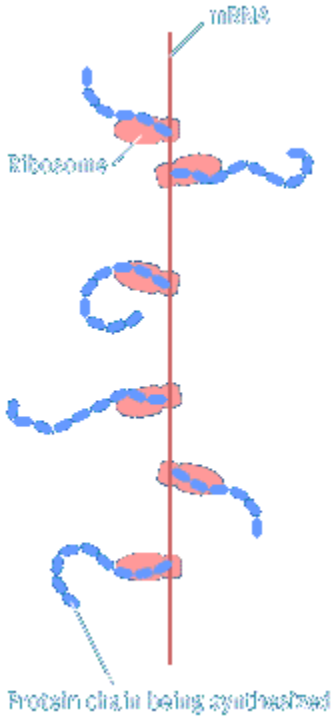


Codons

- 3 consecutive DNA bases code for an amino acid. There are 64 possible codons, but only 20 amino acids (some amino acids have multiple codon representations).
- Special codons: start codon (ATG) [and two others] and three stop codons (TAG, TGA, TAA). They indicate the start and end of translation regions.

The central dogma



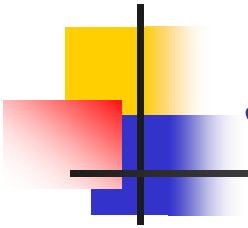


mRNA produced by transciption

DNA → mRNA → proteins

Transcription

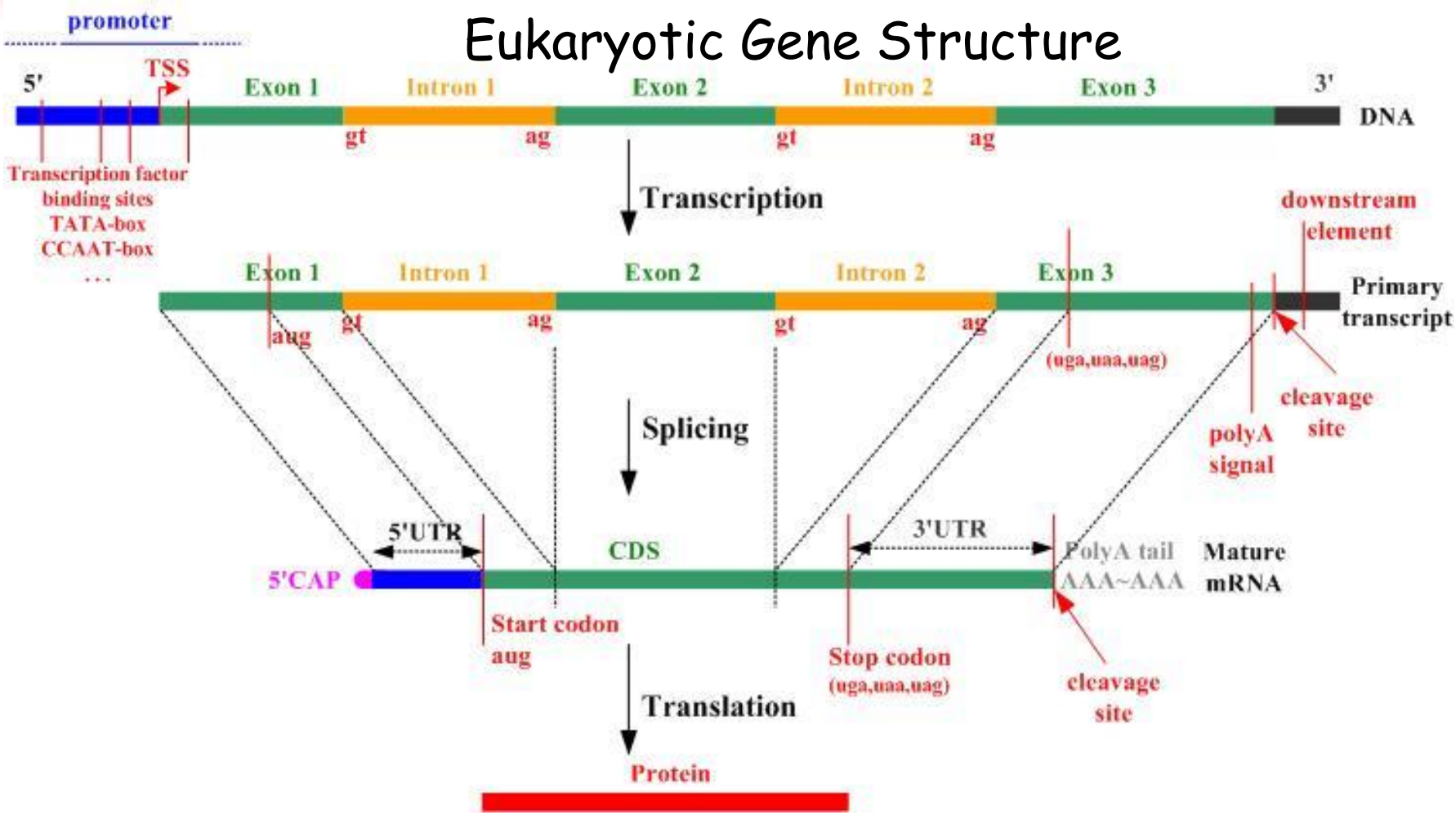
- When a gene is "expressed" the sequence of nucleotides in the DNA is used to determine the sequence of amino acids in a protein in a two step process.
- First, the enzyme RNA polymerase uses one strand of the DNA as a template to synthesize a complementary strand of messenger RNA (mRNA) in a process called transcription. RNA is identical to DNA except that in RNA T is replaced with U (for uracil). Also, unlike DNA, RNA usually exists as a single stranded molecule.



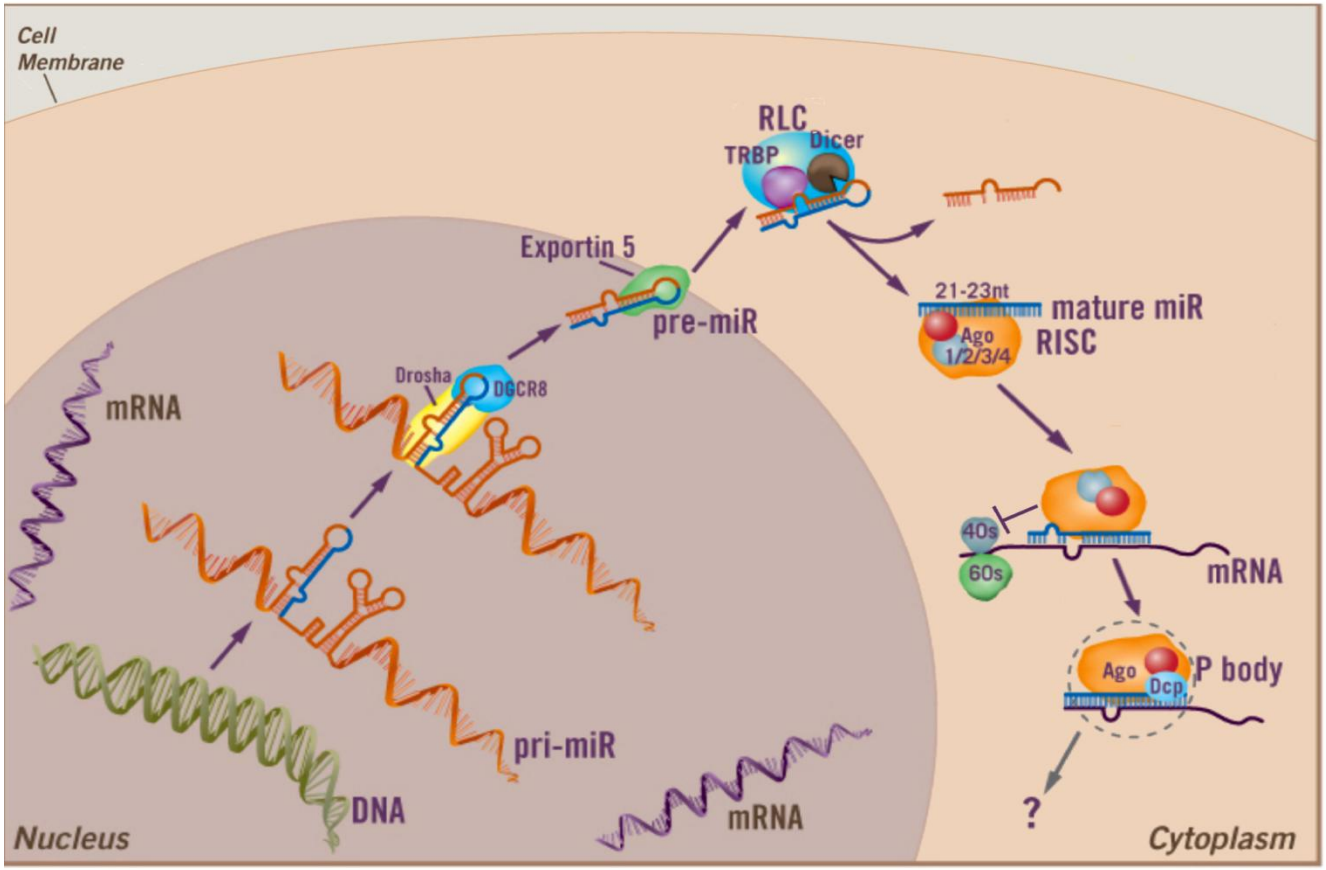
Splicing and Translation

- In eukaryotes, after a gene is transcribed the introns are removed from the mRNA and the adjacent exons are spliced together in the nucleus prior to translation outside the nucleus. (alternative splicing)
- After the mRNA for a particular gene is made it is used as a template with which ribosomes synthesize the protein in a process called translation.

The Biological Model

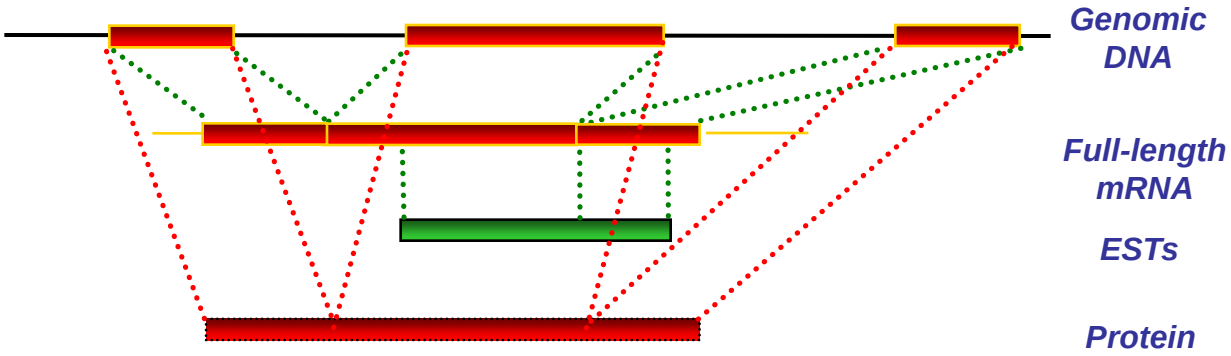


The current view



Philip A. Sharp, MIT, 2009

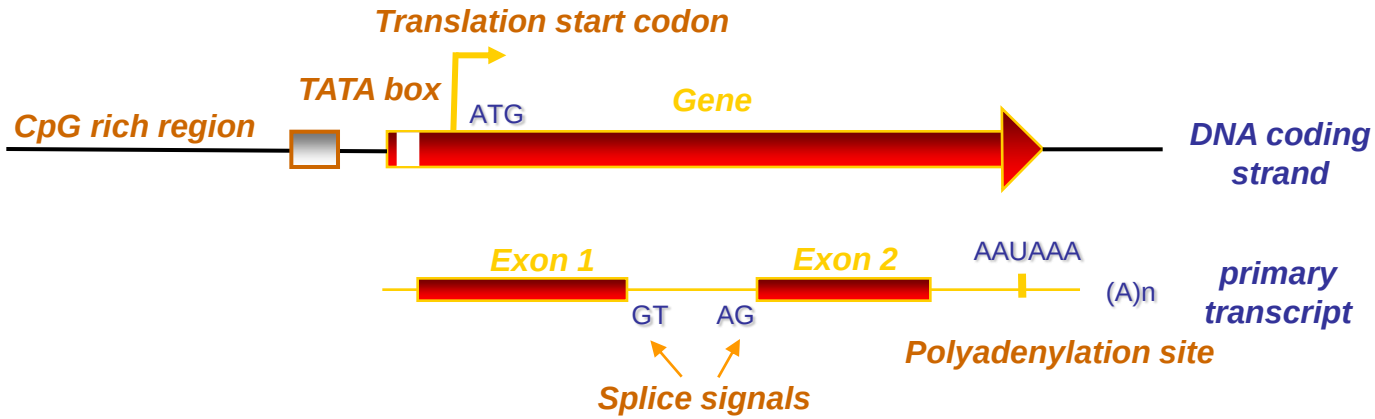
How genes are validated



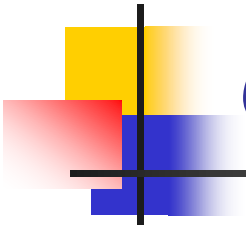
- cDNA - single-stranded DNA complementary to an RNA, synthesized from it by reverse transcription

- full-length mRNAs (GenBank *RefSeq,* ~16000 human sequences)/Ensembl 2009
- ESTs – Expressed Sequence Tags
  - relatively short, 500 bp long on average
  - span one or more exons
  - large data sets required (GenBank *dbEST* – 4.3 M human sequences)

Signals



- Upstream regulatory signals (TATA boxes)
- Translation start codon (ATG)
- Translation stop codon (*e.g.*, TAA)
- Polyadenylation signal (~AATAAA)
- Splice recognition signals (*e.g.*, GT-AG, branch point)

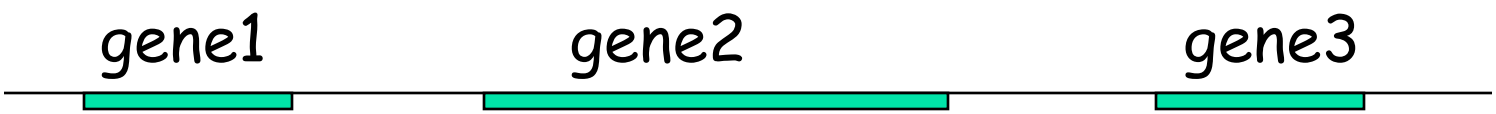


Computational gene finding

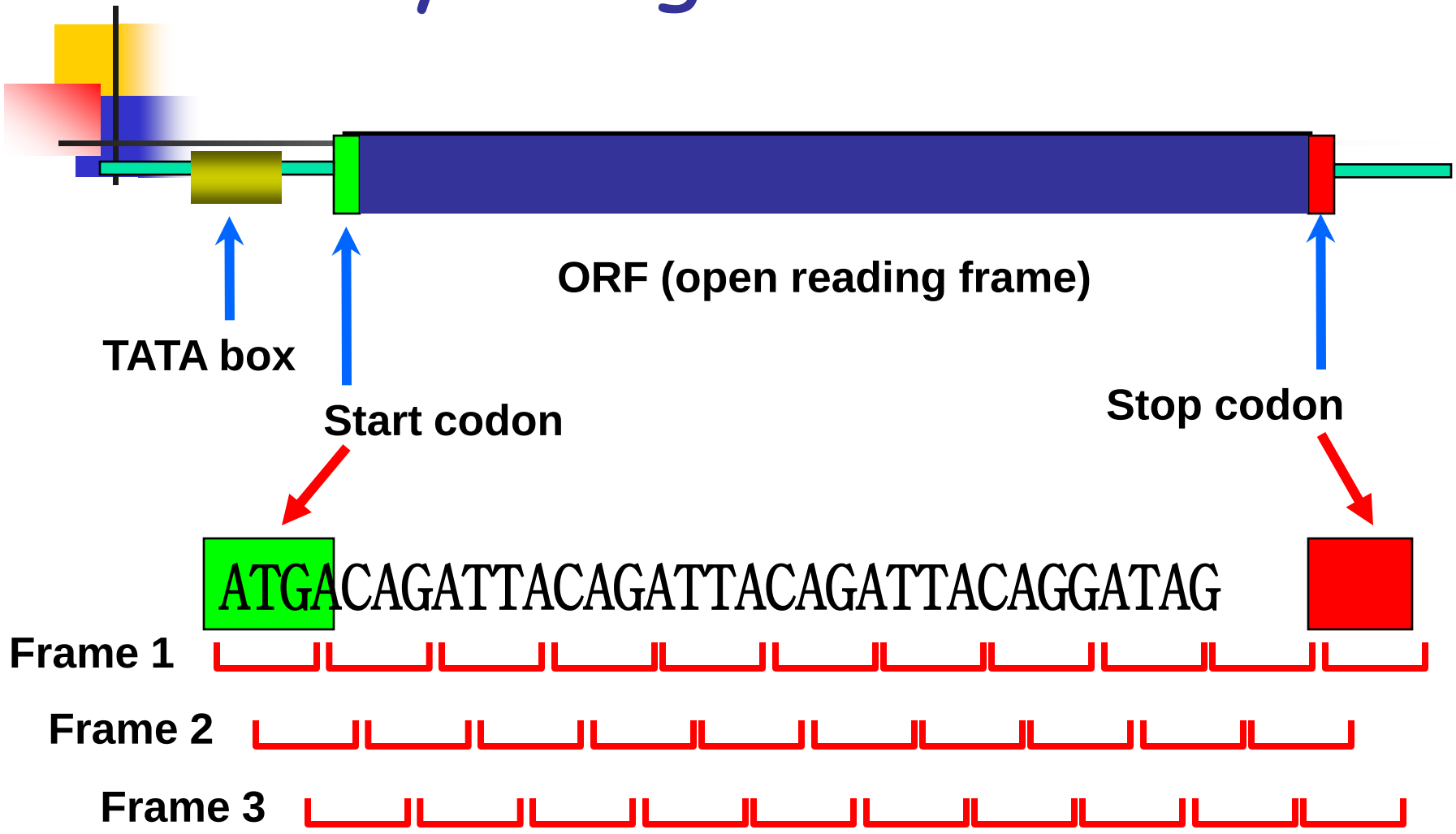
- Gene finding in prokaryotes
- Gene finding in eukaryotes

Finding genes in prokaryotes

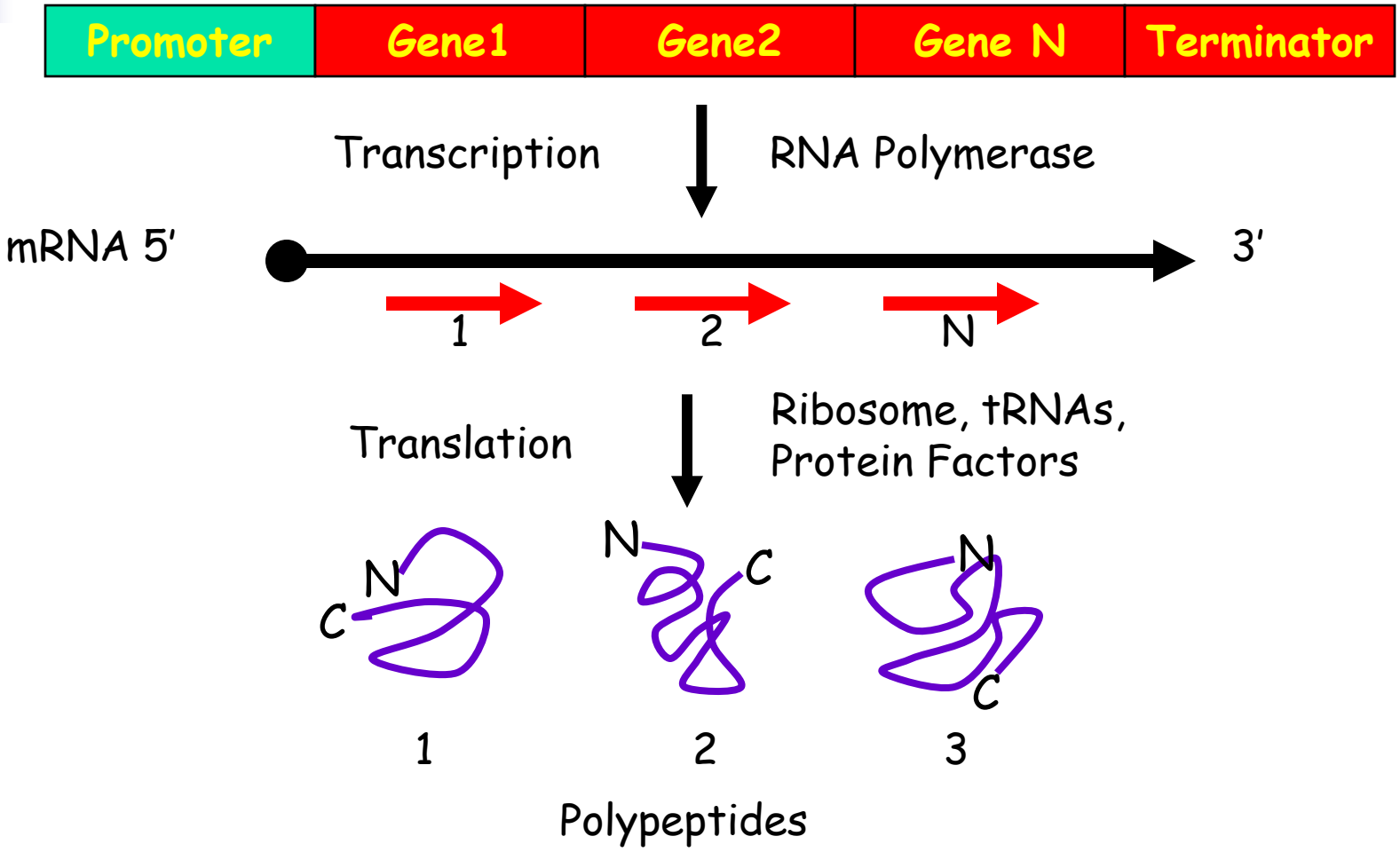
- Prokaryotes are single-celled organisms without a nucleus (e.g., bacteria).
- Few introns in prokayotic cells. Over 70% of H. influenzae genome codes for proteins.
- No introns in coding region.

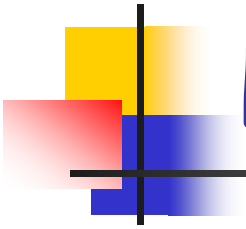


Prokaryotic gene structure



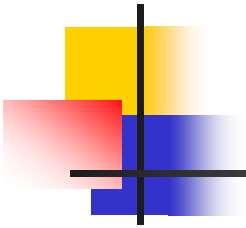
Prokaryotes stack multiple genes together for expression ("operons")





Prokaryote gene finding

- Advantages
  - Simple gene structure
  - Small genomes (0.5 to 10 million bp)
  - No introns
  - High coding density (>90%)
- Disadvantages
  - Some genes overlap (nested)
  - Some genes are quite short (<60 bp)



Finding genes in prokaryotes

- Main idea: if bases were drawn uniformly at random, then a stop codon is expected once every 64/3 (about 21) bases. Since coding regions are terminated by stop codons, a simple technique to find genes is to look for long stretches of bases without a stop codon. Once a stop codon is found, we work backward to find the start codon corresponding to the gene.
- Main problems: misses short genes, overlapping ORFs.

Gene finding based on start/stop codons

- Look for putative start codon (ATG)
- Staying in same frame, scan in groups of three until a stop codon is found
- If # of codons >=50, assume it's a gene
- If # of codons <50, go back to last start codon, increment by 1 & start again
- At end of chromosome, repeat process for reverse complement

Example ORF


```
 5'                                                    3'
  atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa

1 atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *
2  tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat
    C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y
3   gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata
     A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

Segment of Influenza Virus

- http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=CY018024
- 1151 bp segment of Influenza B Virus.
- Has two genes: 4 to 750 and 750 to 1079

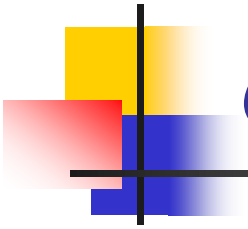
The sequence

1 aaa**atg**tcgc tgtttggaga cacaattgcc tacctgcttt cattgacaga agatggagaa
61 ggcaaagcag aactagcaga aaaattacac tgttggttcg gtgggaaaga atttgaccta
121 gactctgcct tggaatggat aaaaaacaaa agatgcttaa ctgatataca aaaagcacta
181 attggtgcct ctatctgctt tttaaaaccc aaagaccagg aaaggaaaag aagattcatc
241 acagagcctc tatcaggaat gggaacaaca gcaacaaaaa agaaaggcct gattctagct
301 gagagaaaaa tgagaagatg tgtgagcttt catgaagcat ttgaaatagc agaaggccat
361 gaaagctcag cgctactata ttgtctcatg gtcatgtacc tgaatcctgg aaattattca
421 atgcaagtaa aactaggaac gctctgtgct ttgtgcgaga aacaagcatc acattcacac
481 agggctcata gcagagcagc gagatcttca gtgcccggag tgagacgaga aatgcagatg
541 gtctcagcta tgaacacagc aaaaacaatg aatggaatgg gaaaaggaga agacgtccaa
601 aagctggcag aagagctgca aagcaacatt ggagtattga gatctcttgg agcaagtcaa
661 aagaatgggg aaggaattgc aaaggatgta atggaagtgc taaagcagag ctctatggga

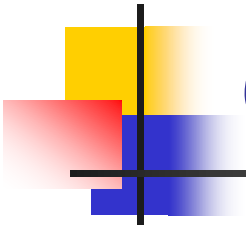
721 aattcagctc ttgtgaagaa atatcta**taa tg**ctcgaacc atttcagatt ctttcaattt
781 gttcttttat cttatcagct ctccatttca tggcttggac aatagggcat ttgaatcaaa
841 taaaaagagg agtaaacatg aaaatacgaa taaaaggtcc aaacaaagag acaataaaca
901 gagaggtatc aattttgaga cacagttacc aaaaagaaat ccaggccaaa gaaacaatga
961 aggaagtact ctctgacaac atggaggtat tgagtgacca catagtgatt gaggggcttt

1021 ctgccgaaga gataataaaa atgggtgaaa cagttttgga gatagaagaa ttgcat**taa**a
1081 ttcaattttt tactgtattt cttattatgc atttaagcaa attgtaatca atgtcagcaa
1141 ataaactgga a

Problems with start/stop codon based approaches



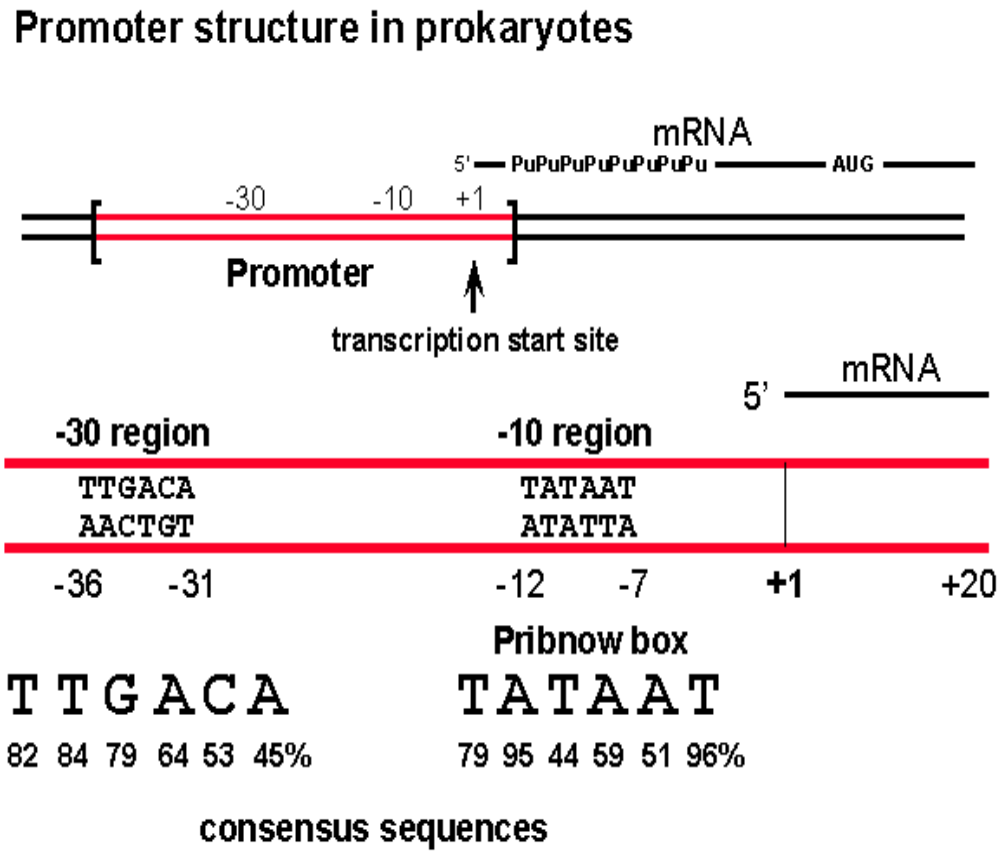
- Advantages
  - Simple and fairly sensitive (>50%)
- Disadvantages
  - Prokaryotic genes are not always so simple to find
  - ATG is not the only possible start site (e.g. CTG, TTG – class I alternates)
  - Small genes tend to be overlooked and long ones over-predicted
- Solution? Use additional information to increase confidence in predictions

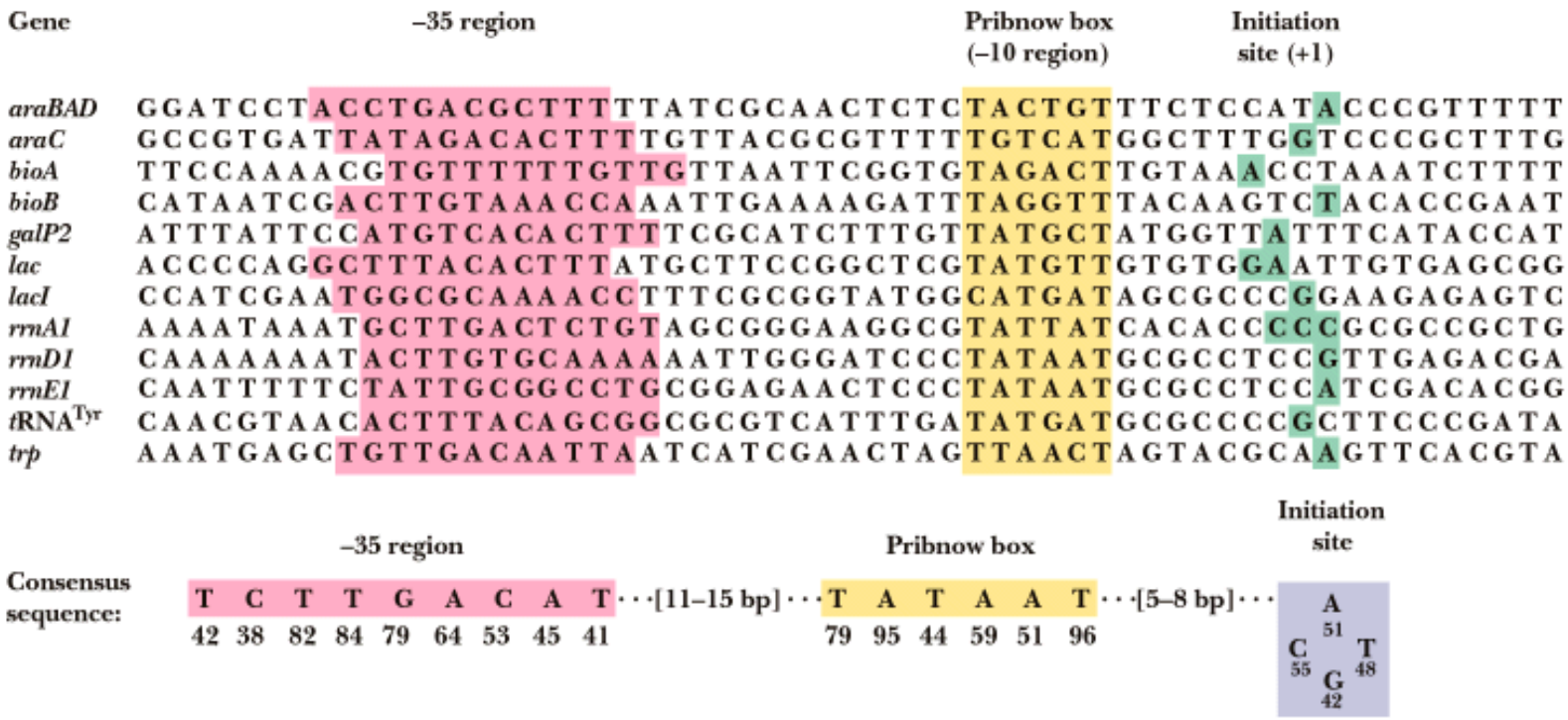


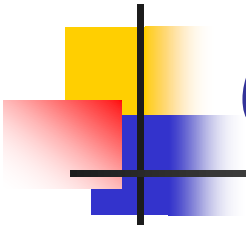
Other content features

- RNA polymerase promoter site (-10, -30 site or TATA box)
- Shine-Dalgarno sequence (+10, Ribosome Binding Site) to initiate protein translation
- Codon biases
- High GC content
- Search for nucleotide sequences and coded proteins in related organisms

Promotor structure




Gene
–35 region
Pribnow box
(–10 region)
Initiation
site (+1)
araBAD GGATCCTACCTGACGCTTTTTATCGCAACTCTCTACTGTTTCTCCATACCCGTTTTT
araC GCCGTGATTATAGACACTTTTGTTACGCGTTTTTGTCATGGCTTTGGTCCCGCTTTG
bioA TTCCAAAACGTGTTTTTTTGTTGTTAATTCGGTGTAGACTTGTAAACCTAAATCTTTT
bioB CATAATCGACTTGTAAACCAAATTGAAAAGATTTAGGTTTACAAGTCTACACCGAAT
galP2 ATTTATTCCATGTCACACTTTTCGCATCTTTGTTATGCTATGGTTATTTCATACCAT
lac ACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGG
lacI CCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTC
rrnA1 AAAATAAATGCTTGACTCTGTAGCGGGAAGGCGTATTATCACACCCCCGCGCCGCTG
rrnD1 CAAAAAAATACTTGTGCAAAAAATTGGGATCCCTATAATGCGCCTCCGTTGAGACGA
rrnE1 CAATTTTTCTATTGCGGCCTGCGGAGAACTCCCTATAATGCGCCTCCATCGACACGG
tRNA^Tyr CAACGTAACACTTTACAGCGGCGCGTCATTTGATATGATGCGCCCCGCTTCCCGATA
trp AAATGAGCTGTTGACAATTAATCATCGAACTAGTTAACTAGTACGCAAGTTCACGTA
Initiation
site
–35 region
Pribnow box
Consensus
sequence:
T C T T G A C A T ···[11–15 bp]··· T A T A A T ···[5–8 bp]···
42 38 82 84 79 64 53 45 41
79 95 44 59 51 96
A 51
C 55
G 42
T 48

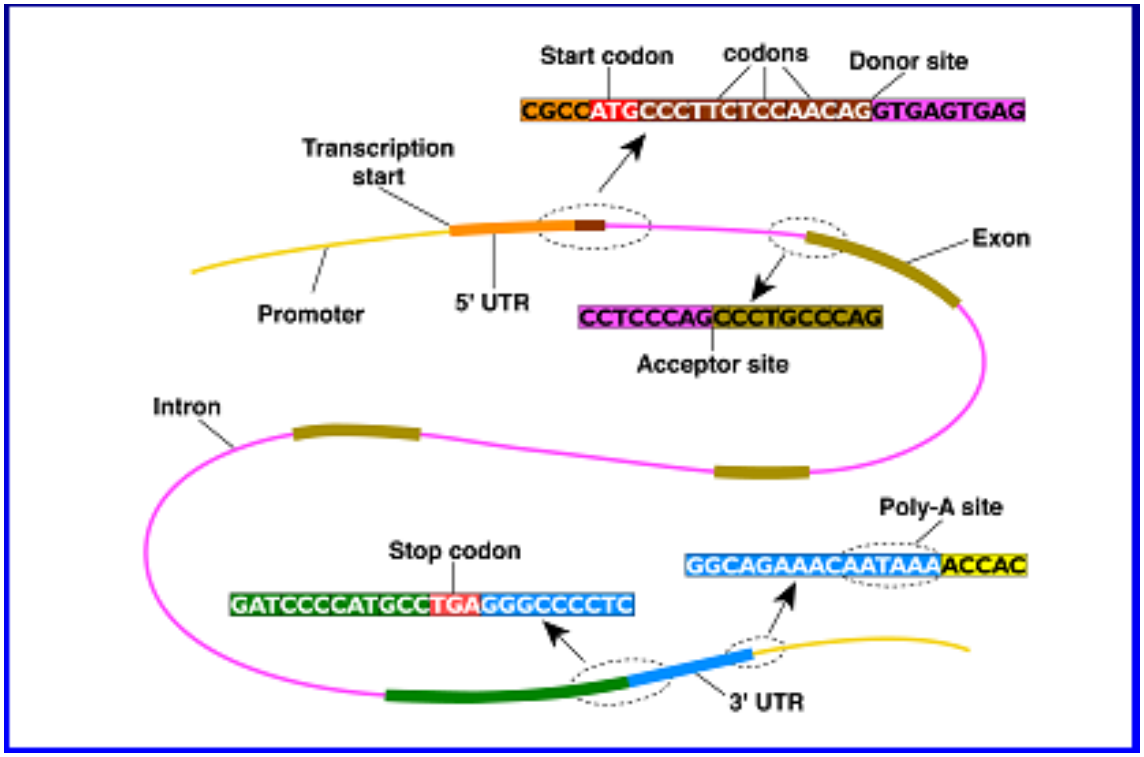


GLIMMER

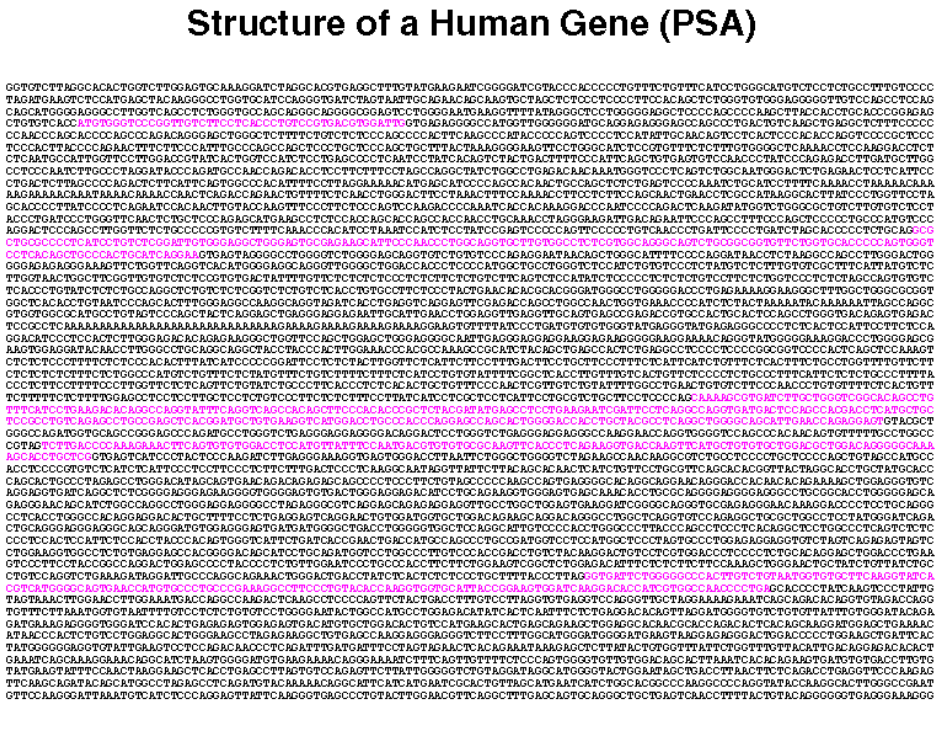
- State of the art prokaryotic gene finder. Based on interpolated Markov models.
- Available at http://cbcb.umd.edu/software/glimmer
- 98% accuracy in identifying viral and microbial genes. 2007 paper in Bioinformatics that shows latest version of tool.

Gene finding in eukaryotes

- Gene finding in eukaryotic DNA



Structure of a human gene



Exon-intron structure


> HSCKIIBE, Human gene for casein kinase II subunit beta (EC 2.7.1.37).
ggggctgagatgtaaattagaggagctggagaggagtgcttcagagtttgggttgctttaagaaagggt
ggttccgaattctcccgtggttggagggccgaatgtgggaggag
gaacttgagctttactgacactgttctttttctagctgacgtga
ctggatttcctggttctgtgggctccgtggcaatgaattcttct
ctacttgccagcttcacatatcttcccaccagacgttccttcacatattccacttctacactgttctct
aaagcttttatgggagagagtgtaggtgaactagggagagacacaagtacttctgctgagttgggagtg
agaaacaagcacaacagatgcagttgtgttgatgataaggcatc
agatgaggattttgatatgggttccctcttggcttccatgtcct
ggacaaatt
ggacctgga
atctctatt
tgctttattttgatcctccacctgactcttgtctagttttgtgacgtatatcacttgttctcatgtttt
tgcaagggtcagaagcccaggtttctgggtcccatgcccagatgttggatggggtaaggcccaaaagta
ggtgctaggcaaactgaatagcccgca
agtagttgcatttggccgggctgtggt
attg
gccc
gcaa
ccctgccctaatttggaagaaaggcaacacagaagtttgagagcccatctagtccagagaaggggcct
ctggacagagttggaaggagtgccgacagagttg
ttacatctacctgccaacccctccattgtattc
gttactgtcctcgtgtgtactgtgagaaccagcc
ggaaagcaccgtgtggcagtcttatgggaaggagttggggctcaa
ggaggttaggtaggaataggggatacctggcctgctgagtctggctgt
cccaggtgaagccatggtgaagctctactgccccaagtgcatggatgtg
acaccatcacacggatggcgcctacttcggcactgg
gtaccggcccaagagacctgccaaccagtttgtgcc
aaggaaaggcccaagatcccccagagaggggaggac
cagaatcagggcatctccctgctgagtgactgtgggaaagttatttgattatctgtgcttgagttacct
tattgtagaatgttcttgagctgagaagttgggaaccacgaggctttagctctgagcaggtccatagag
gagctcaggtggggaggtgggaatgcaggtgactggcagggcctggatggggctcatgctgctgcctct
ctgacctctgccctggcctaggctctacggtttcaaga
agccgccagcaacttcaagagcccagtcaagacgattcg
tttgtcttttccttctttttgccacccttcaggaac
gtcgttatcgtggtgggaatatgaaataaagtagaagaaaaggccatgagctagtctgctggtgcttgc
ggaaggggtggagcgtggccatggaaatcgggctccacggcccagggatgg
atg
caggtg
ggtgag
cagatg
ggtgag
cagttg
ggtgag
caggcc
ggtgag
tga

Ab initio methods

- Use information embedded in the genomic sequence *exclusively* to predict the gene structure.
- Find structure G representing gene boundaries + internal gene structure which maximizes the probability P(G|genomic sequence).
- Hidden Markov models are the predominant generative method for modeling the problem.