# Bioinformatics

## CS300

## Genome annotation: Advanced (eukaryotic) gene prediction Continued...

**Spring 2021**

**Oliver BONHAM-CARTER**

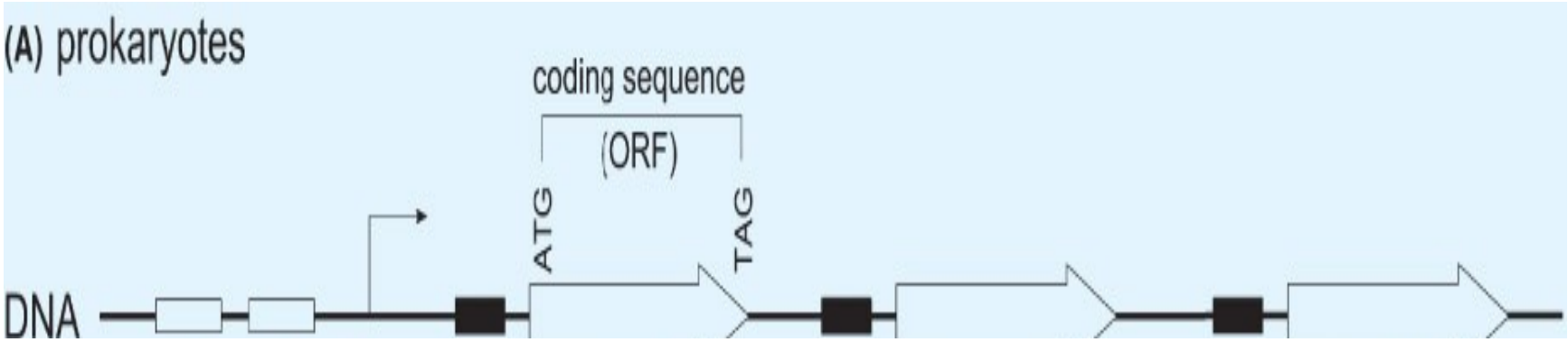
# Prediction Algorithms

- **Alignment-based:** Sequence similarity to previously identified gene in another organism (BLAST)
- **Sequence-based:** Search for specific sequences - e.g. ORF–finder – searches start and stop codons
- **Probabilistic:** Combination of sequence- and content-based plus probability that sequence is part of a gene
- **Content-based:** Search for patterns such as nucleotide or codon frequencies

# Prokaryotic Gene Structure

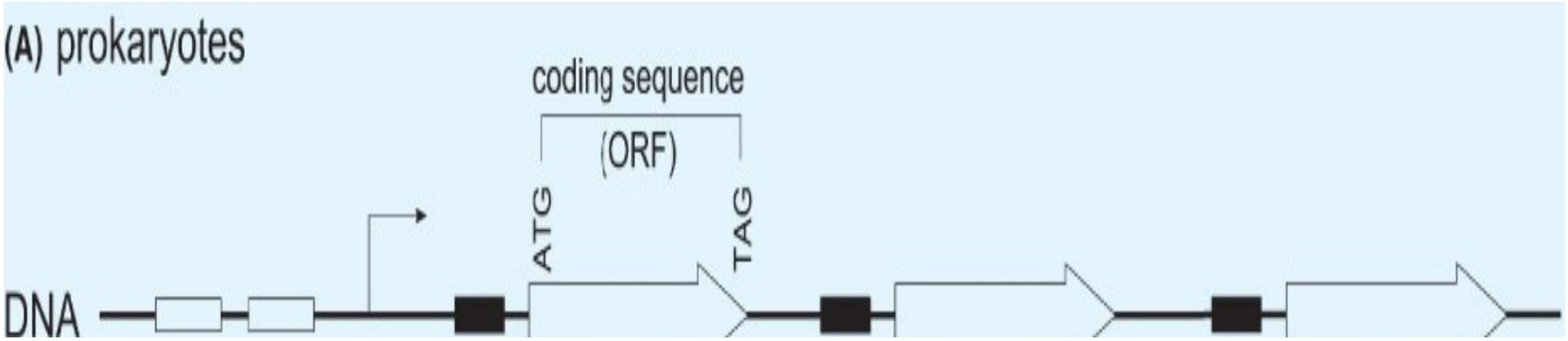
- Highly conserved sequences 10 and 35 bases upstream (before) start codon
- Highly conserved Shine-Dalgarno sequence immediately before start codon

| Sequence | Consensus (5′ → 3′) | Function |
|---|---|---|
| Prokaryotes | | |
| –10 sequence | TATAAT | RNA polymerase binds to start transcription |
| –35 sequence | TTGACA 17±2 from –10 | RNA polymerase binds to start transcription |
| Shine-Dalgarno | AGGAGG 5±2 from ATG | Ribosome binds to find start codon |



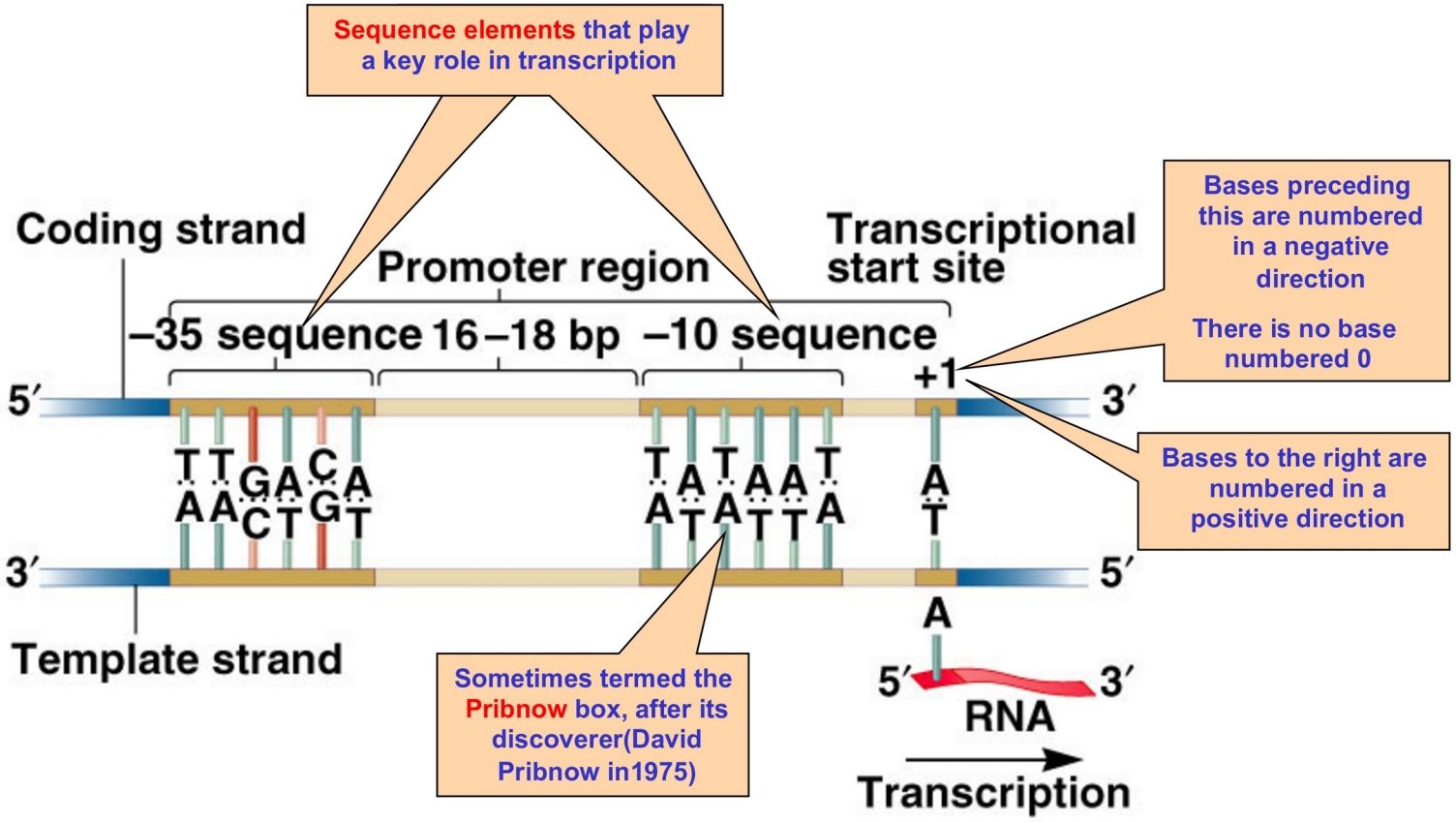
# Prokaryotic Gene Structure

- Highly conserved sequences 10 and 35 bases upstream (before) start codon
- Highly conserved Shine-Dalgarno sequence immediately before start codon
- Genes rarely contain introns
  - Present as ORFs (start codon through stop codon – all protein-coding sequence)



# Prokaryotic Gene Structure

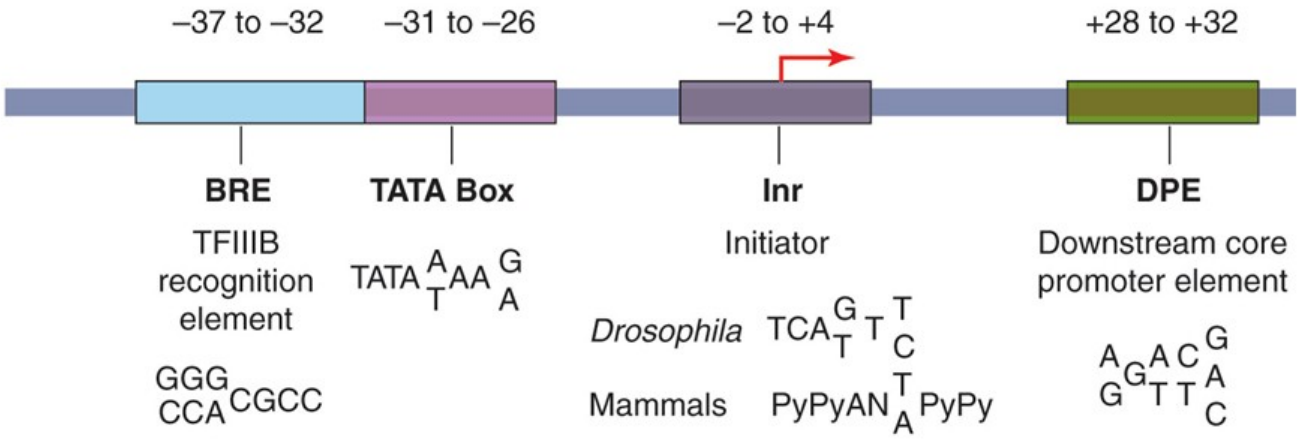
- **The bases in a promoter sequence are numbered in relation to the transcription start site.**




**The conventional numbering system of promoters**

# Eukaryotic Gene Structure

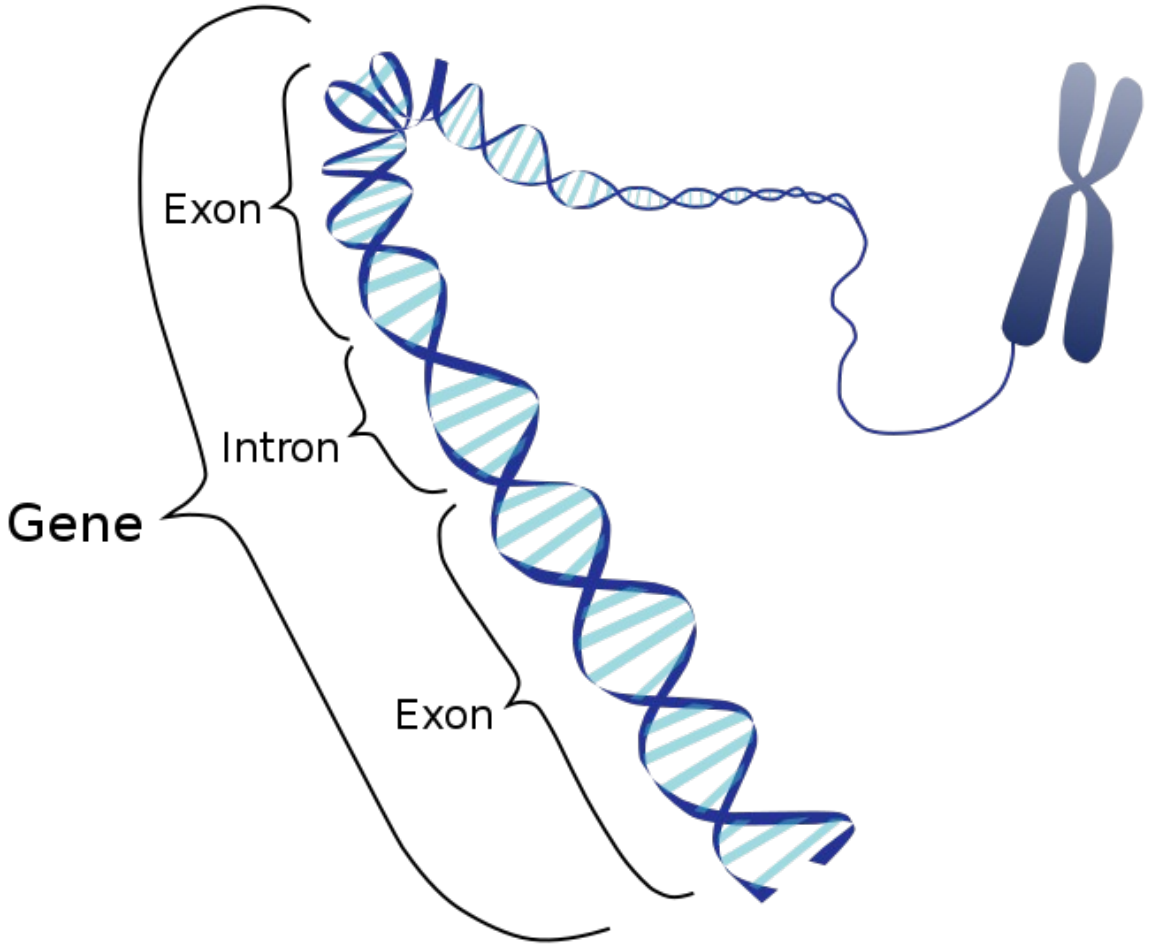
- Variable promoter structure
  - Not all promoter elements present in all genes
  - Promoter element sequence can vary between genes
  - No conserved Shine-Delgarno-like sequence




  - Find the END but identifying the START of genes is more difficult
  - :-(

# Eukaryotic Gene Structure

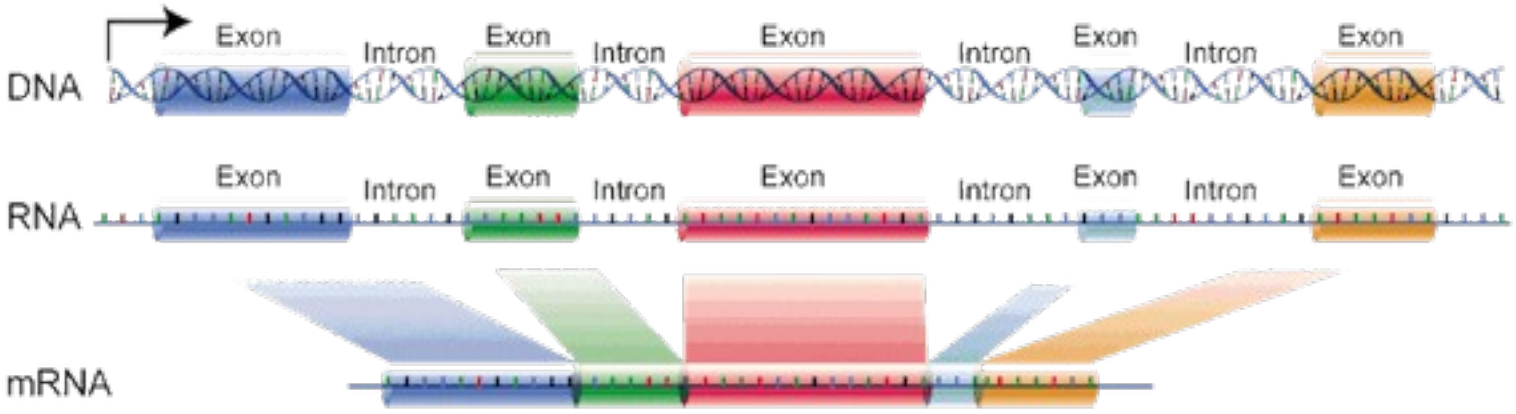
- An **Intron** (intragenic region) is any nucleotide sequence within a gene that is removed by RNA splicing during maturation of the final RNA product.
- An **exon** (expressed region) is any part of a gene that will encode a part of the final mature RNA produced by that gene after introns have been removed by RNA splicing.




- https://www.youtube.com/watch?v=YtKoTOCJGt4 (1 min)
- https://www.youtube.com/watch?v=_asGjfCTLNE (6.5 mins)

# Boundaries in RNA

- Most genes contain introns that do not code for protein



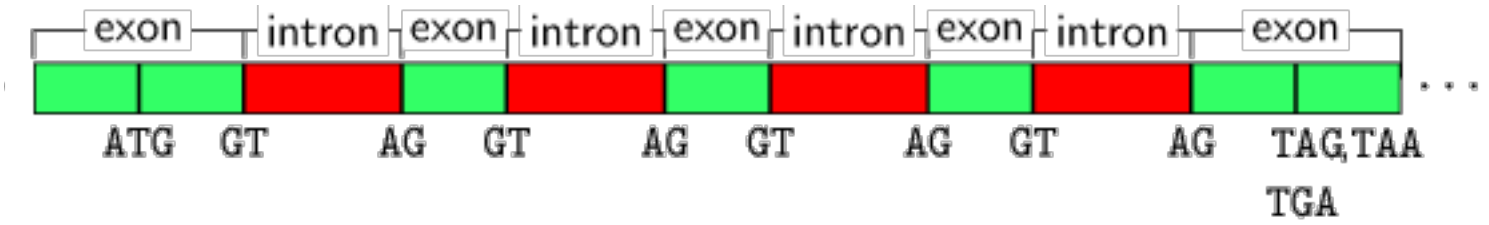

- There is commonly a G-C boundary between EXON's )coding regions) and INTRONs (non-coding regions)
  - Separations between regions help in removing editing out non-protein building code

# Little Conservation of Boundaries

- Nuclear pre-mRNA introns (spliceosomal introns) are characterized by specific intron sequences located at the boundaries between introns and exons.
- These sequences are recognized by spliceosomal RNA molecules when the splicing reactions are initiated.

- **Exon**/**Intron** boundaries not highly conserved (hard to identify boundaries between genes)

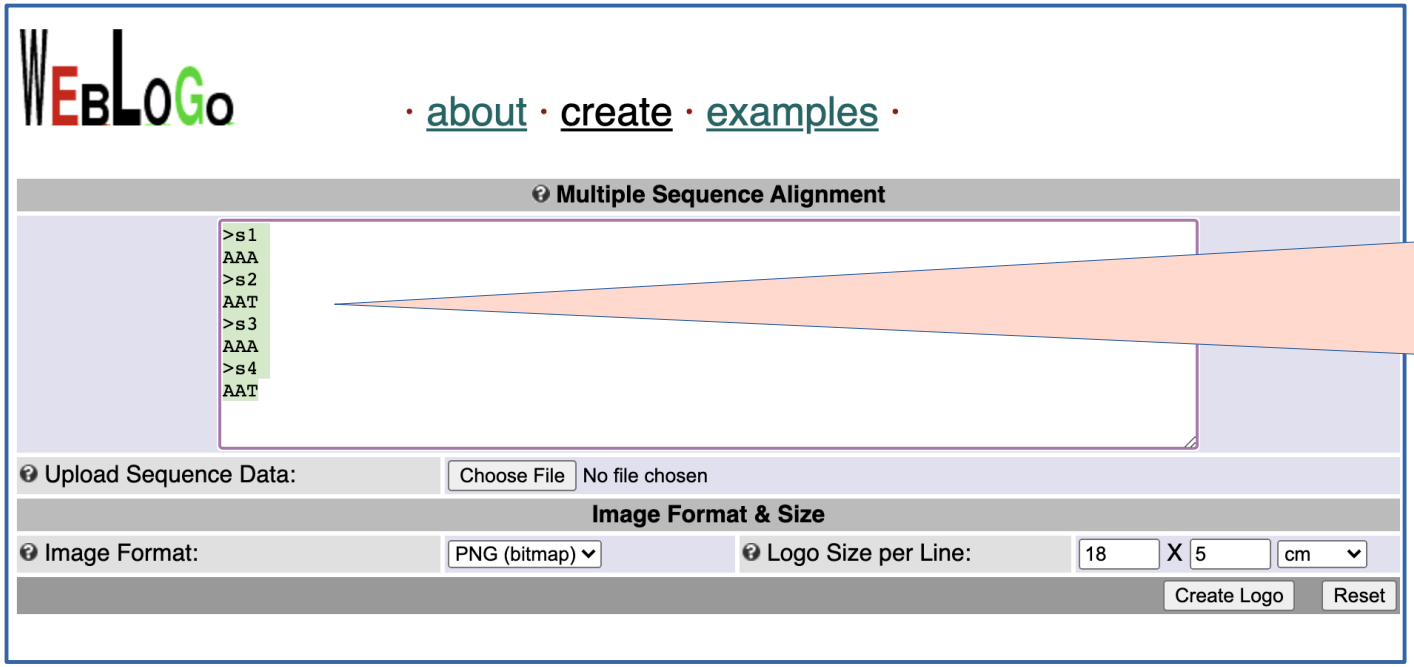
```
catgATGGCTTTTCAagggctctCACGTCCgagtcataTGACCGaggact
gtacTACCGAAAAGTtcccgagaGTGCAGGctcagtatACTGGCtcctga
```



# Studying Gene Structures

## http://weblogo.berkeley.edu/logo.cgi

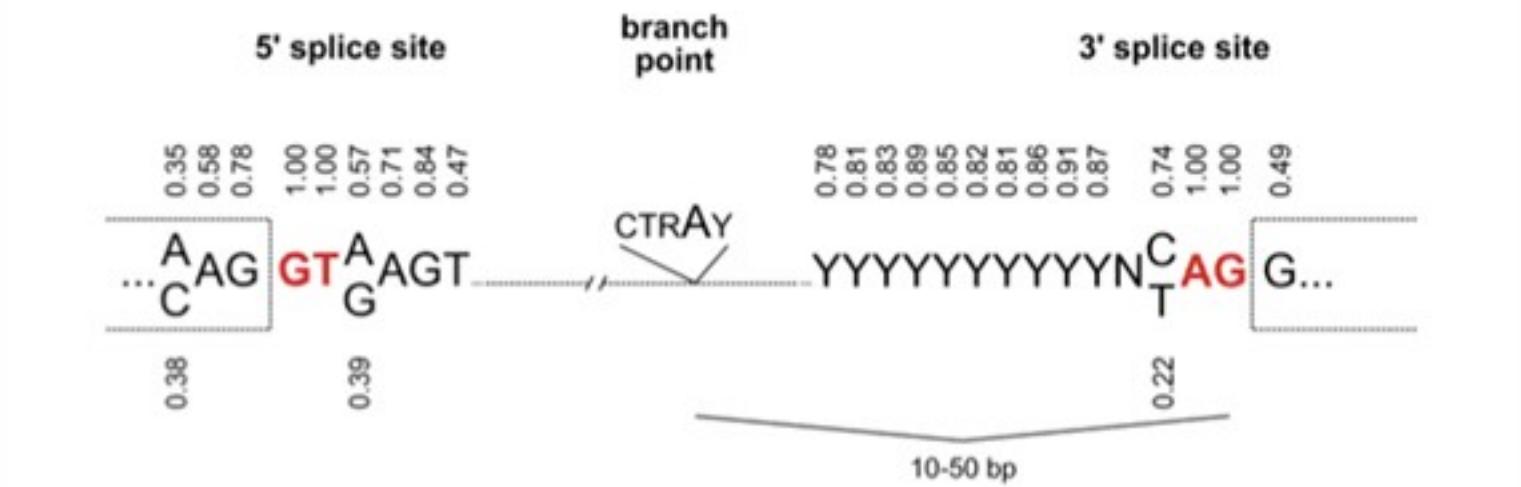
- How to quickly compare bases in sequences (of similar length)?




>s1
AAA
>s2
AAT
>s3
AAA
>s4
AAT

All sequences have two A's, then differ by an A or T

# Eukaryotic Gene Boundaries

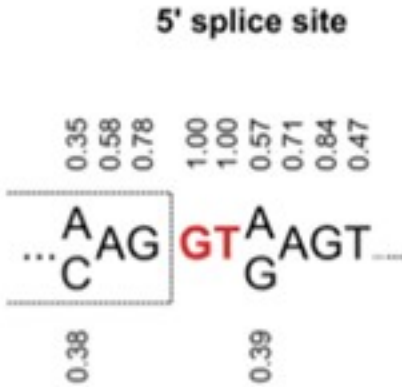


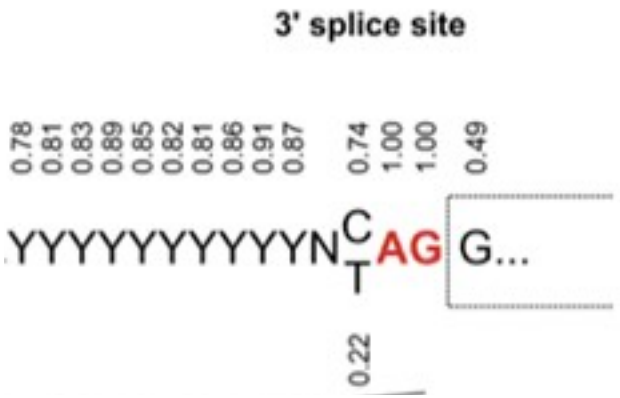

- Exon/Intron boundaries not highly conserved: different genes, different boundary signatures ...
- In general, codon usage is the same within one organism.
- Observed changes in codon usage in parts of code suggest intron /exon boundaries exist
- **Energy signals**: DNA adopts a unique structural and energy state at the boundary junctions of protein coding genes (in human DNA).
- (Article)**: Intron exon boundary junctions in human genome have in-built unique structural and energetic signals**:
  - Link; https://academic.oup.com/nar/article/49/5/2674/6148175

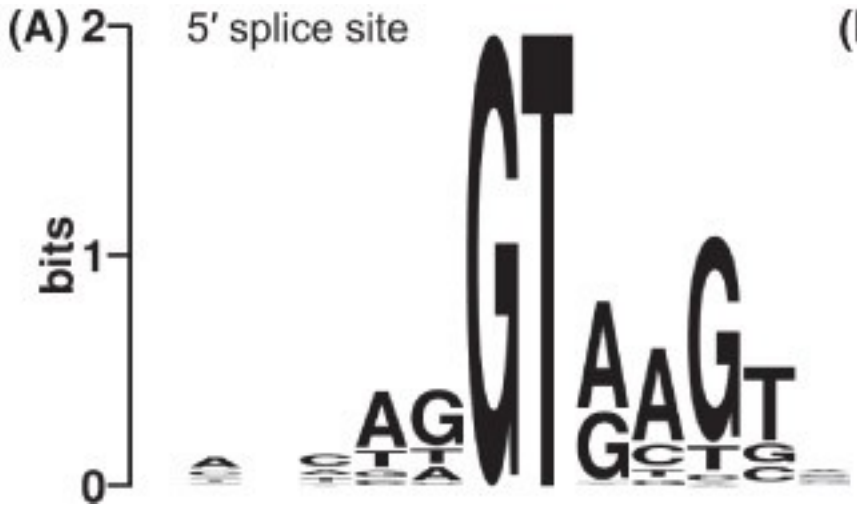
# Eukaryotic Gene Structure

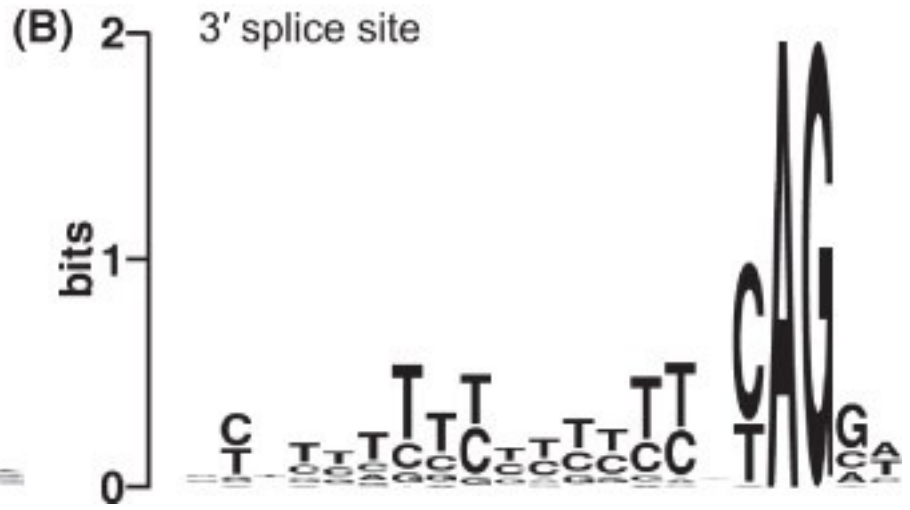
http://weblogo.berkeley.edu/logo.cgi

- Exon/Intron boundaries not highly conserved; samples have differing bases



# Differences

| Prokaryotic Gene Structure | Eukaryotic Gene Structure |
| --- | --- |
| -10 and -35 promoter sequences | Promoter sequences vary in number and sequence |
| Shine-Dalgarno sequence marks the start codon | No Shine-Dalgarno – unambiguous identification of transcriptional start site is difficult |
| Few interrupted ORFs (introns are rare) | Nearly all genes contain introns |
| | Intron/exon boundaries are hard to discern |

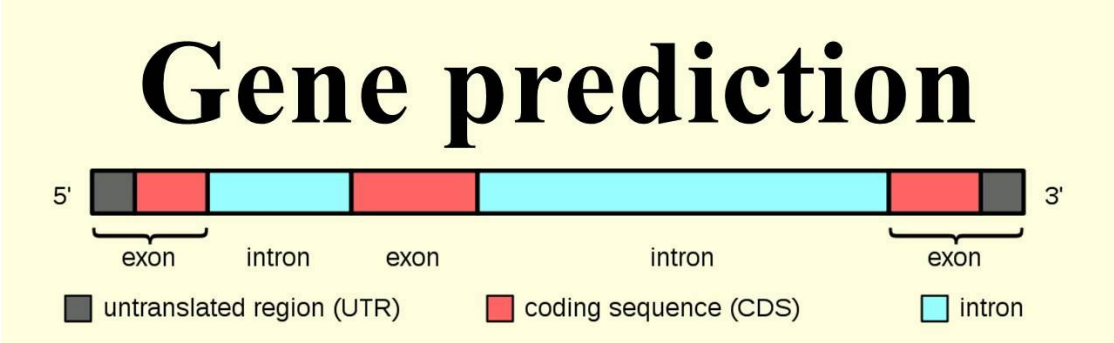
ALLEGHENY COLLEGE

# Bioinformatics Solution?

- Content and Probability-Based Gene Prediction

- Content-based gene prediction
  - Alone, not very precise
  - Better results when used in combination with other algorithms

- Content
  - Codon usage
  - CpG Islands



# Codon Usage and Frequency

Synonymous codons for same amino acids but not used with equal frequencies

***Escherichia coli K12*** **[gbbct]: 14 CDS's (5122 codons)**

fields: [triplet] [frequency: **per thousand**] ([number])

```
UUU 19.7(   101)  UCU  5.7(    29)  UAU 16.8(    86)  UGU  5.9(    30)
UUC 15.0(    77)  UCC  5.5(    28)  UAC 14.6(    75)  UGC  8.0(    41)
UUA 15.2(    78)  UCA  7.8(    40)  UAA  1.8(     9)  UGA  1.0(     5)
UUG 11.9(    61)  UCG  8.0(    41)  UAG  0.0(     0)  UGG 10.7(    55)

CUU 11.9(    61)  CCU  8.4(    43)  CAU 15.8(    81)  CGU 21.1(   108)
CUC 10.5(    54)  CCC  6.4(    33)  CAC 13.1(    67)  CGC 26.0(   133)
CUA  5.3(    27)  CCA  6.6(    34)  CAA 12.1(    62)  CGA  4.3(    22)
CUG 46.9(   240)  CCG 26.7(   137)  CAG 27.7(   142)  CGG  4.1(    21)

AUU 30.5(   156)  ACU  8.0(    41)  AAU 21.9(   112)  AGU  7.2(    37)
AUC 18.2(    93)  ACC 22.8(   117)  AAC 24.4(   125)  AGC 16.6(    85)
AUA  3.7(    19)  ACA  6.4(    33)  AAA 33.2(   170)  AGA  1.4(     7)
AUG 24.8(   127)  ACG 11.5(    59)  AAG 12.1(    62)  AGG  1.6(     8)

GUU 16.8(    86)  GCU 10.7(    55)  GAU 37.9(   194)  GGU 21.3(   109)
GUC 11.7(    60)  GCC 31.6(   162)  GAC 20.5(   105)  GGC 33.4(   171)
GUA 11.5(    59)  GCA 21.1(   108)  GAA 43.7(   224)  GGA  9.2(    47)
GUG 26.4(   135)  GCG 38.5(   197)  GAG 18.4(    94)  GGG  8.6(    44)
```

Coding GC 52.35% 1st letter GC 60.82% 2nd letter GC 40.61% 3rd letter GC 55.62%

***Homo sapiens*** **[gbpri]: 93487 CDS's (40662582 codons)**

fields: [triplet] [frequency: **per thousand**] ([number])

```
UUU 17.6(714298)  UCU 15.2(618711)  UAU 12.2(495699)  UGU 10.6(430311)
UUC 20.3(824692)  UCC 17.7(718892)  UAC 15.3(622407)  UGC 12.6(513028)
UUA  7.7(311881)  UCA 12.2(496448)  UAA  1.0( 40285)  UGA  1.6( 63237)
UUG 12.9(525688)  UCG  4.4(179419)  UAG  0.8( 32109)  UGG 13.2(535595)

CUU 13.2(536515)  CCU 17.5(713233)  CAU 10.9(441711)  CGU  4.5(184609)
CUC 19.6(796638)  CCC 19.8(804620)  CAC 15.1(613713)  CGC 10.4(423516)
CUA  7.2(290751)  CCA 16.9(688038)  CAA 12.3(501911)  CGA  6.2(250760)
CUG 39.6(1611801)  CCG  6.9(281570)  CAG 34.2(1391973)  CGG 11.4(464485)

AUU 16.0(650473)  ACU 13.1(533609)  AAU 17.0(689701)  AGU 12.1(493429)
AUC 20.8(846466)  ACC 18.9(768147)  AAC 19.1(776603)  AGC 19.5(791383)
AUA  7.5(304565)  ACA 15.1(614523)  AAA 24.4(993621)  AGA 12.2(494682)
AUG 22.0(896005)  ACG  6.1(246105)  AAG 31.9(1295568)  AGG 12.0(486463)

GUU 11.0(448607)  GCU 18.4(750096)  GAU 21.8(885429)  GGU 10.8(437126)
GUC 14.5(588138)  GCC 27.7(1127679)  GAC 25.1(1020595)  GGC 22.2(903565)
GUA  7.1(287712)  GCA 15.8(643471)  GAA 29.0(1177632)  GGA 16.5(669873)
GUG 28.1(1143534)  GCG  7.4(299495)  GAG 39.6(1609975)  GGG 16.5(669768)
```

Coding GC 52.27% 1st letter GC 55.72% 2nd letter GC 42.54% 3rd letter GC 58.55%

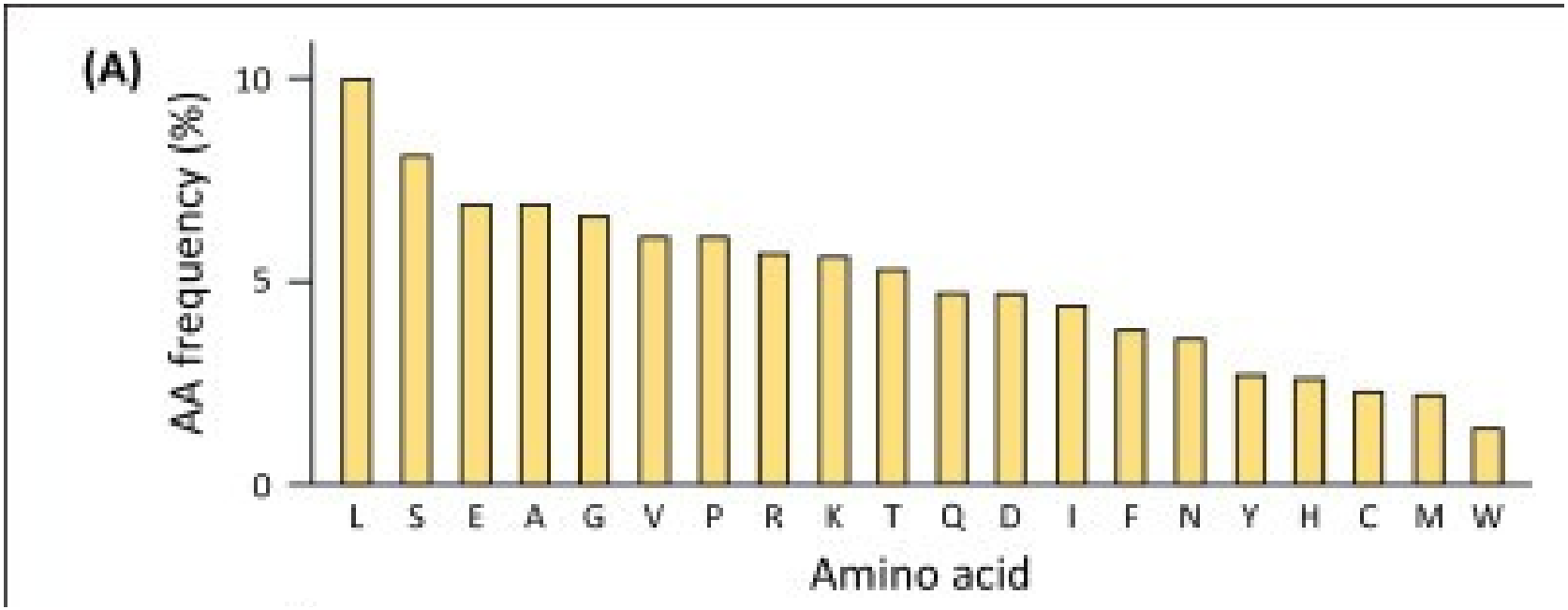
*E. coli*

*H. sapiens*

Codon Usage Database:
http://www.kazusa.or.jp/codon/

# Amino Acids and Frequency

**Some amino acids are much more common in proteins than others**

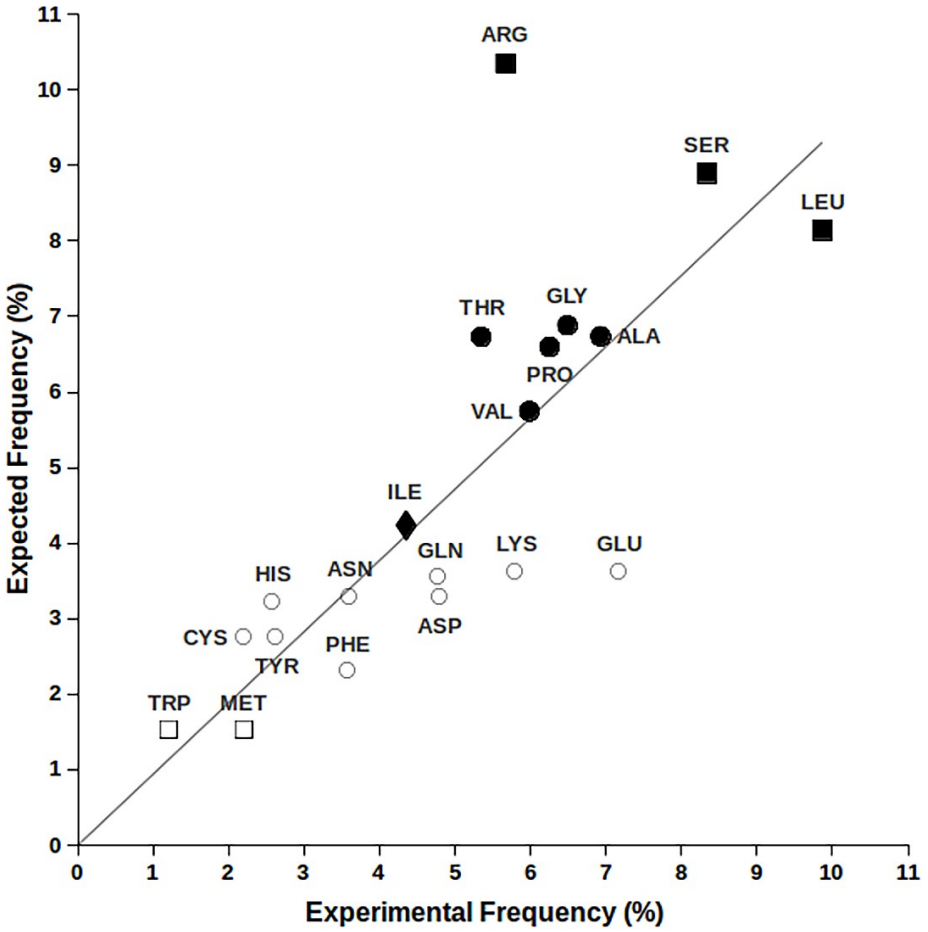


# Codon Usage: expected and observed

The sum of the expectation values for each codon of natural amino acids, as resulting from the product of their nucleotide occurrence, is compared to the amino acid frequency found in human protein sequences.

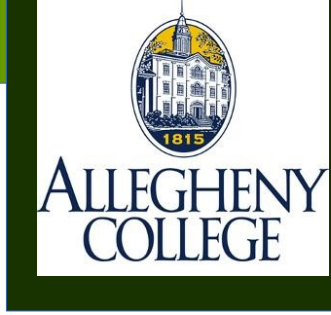
It is possible to note that expected and observed amino acid frequencies exhibit a good correlation with R squared = 0.91

Amino acids with 6, 4, 3, 2 and 1 codons are labelled respectively with “■”, “●”, “♦”, “○” and “□”.




Gardini, Simone, et al. "On nature’s strategy for assigning genetic code multiplicity." PloS one 11.2 (2016): e0148174.

# Using Codon Frequencies to Find Exon/Intron Boundaries

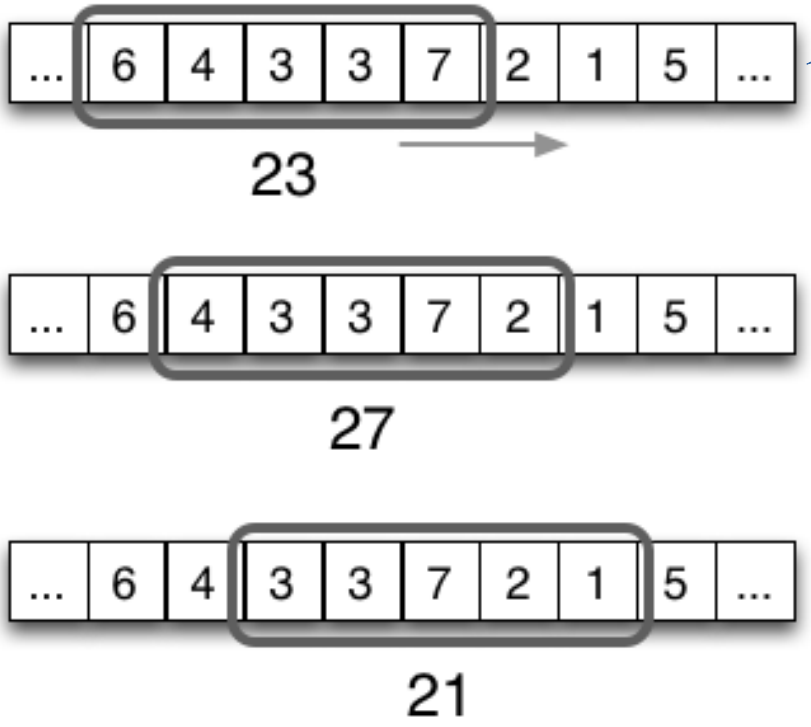



- Expectations:
  - Exon: codon frequency closely matches expected frequency for a gene
  - Intron: “codon” frequency poorly matches expected frequency for a gene (because not really a functional codon that codes for amino acids ...)
  - Boundary: locations in genes where base frequencies shift (observable differences)

# Using Codon Frequencies to Find Exon/Intron Boundaries

- Sliding-Window Approach

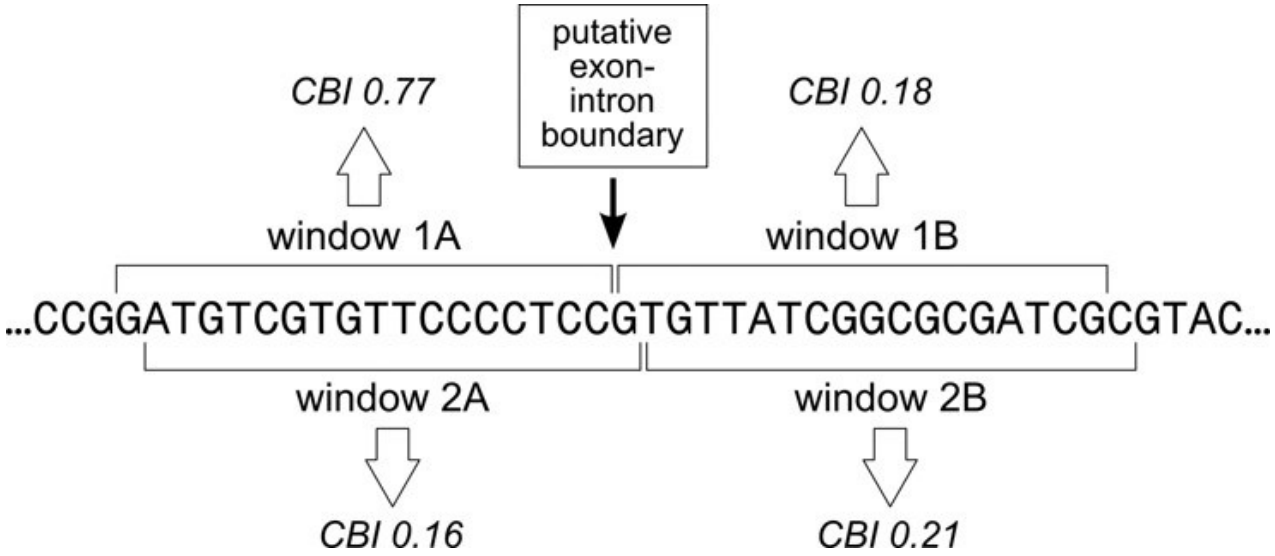
Slide a window along the sequence to read the frequency scores of the codons.




*As the sliding window advances, the slice of its input data changes. Here the algorithm uses the current sliding window data to compute the sum of the window's elements.*

The CpG island is a short stretch of DNA in which the frequency of the CG sequence is higher than other regions.

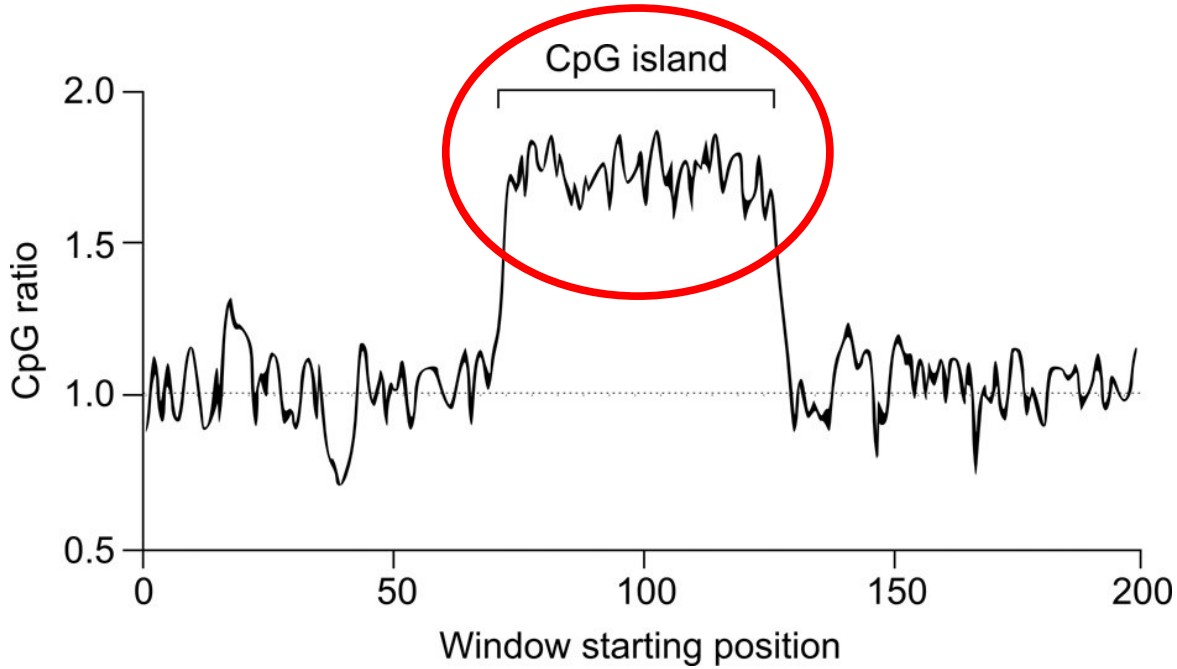
# Codon Biases:
# Find Exon/Intron Boundaries




- **CBI** = **codon bias index**
- Codon bias index is a measure of directional codon bias. It is a measurement of the extent to which a gene uses a subset of optimal codons. CBI is similar to another measurement called the “Frequency of Optimal Codon” with an additional expected usage calculation used as a scaling factor.
  - Compares usage of common codons to the random occurrence of the same codons
  - The algorithm is in the Exploring Bioinformatics textbook, page 198.

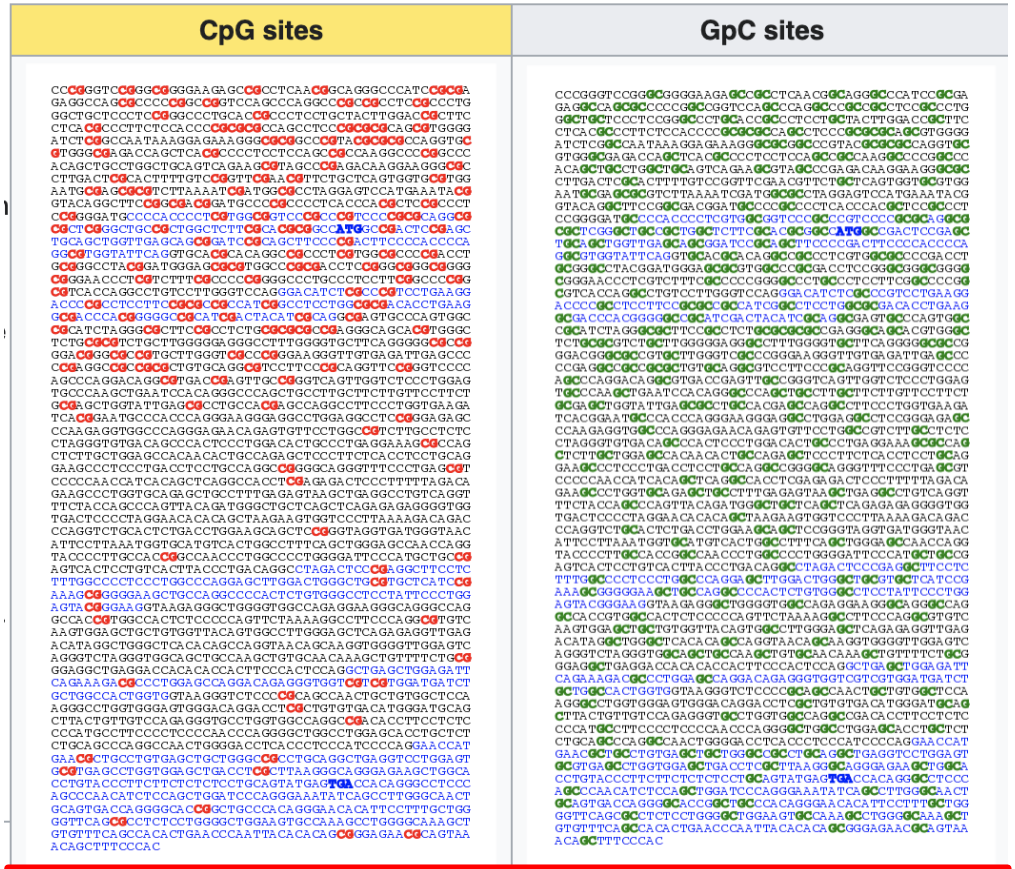
# Finding Promoters Using CpG Islands

- Sliding-window approach + pattern matching algorithm
  - Just one window
- CpG ratio = 1 for no difference between random and naturally occurring codons.




$$\text{CG Freq} = \frac{\textit{observed CG pairs}}{\textit{C nucleotides x G nucleotides/total nucleotides}}$$

# Genetic Monuments (CG and GC islands)




Distribution of CpG sites (left: in red) and GpC sites (right: in green) in the human APRT gene. CpG are more abundant in the upstream region of the gene, where they form a CpG island, whereas GpC are more evenly distributed. The 5 exons of the APRT gene are indicated (blue), and the start (ATG) and stop (TGA) codons are emphasized (bold blue).

https://en.wikipedia.org/wiki/CpG_site

# Finding Promoters Using CpG Islands (CG islands)

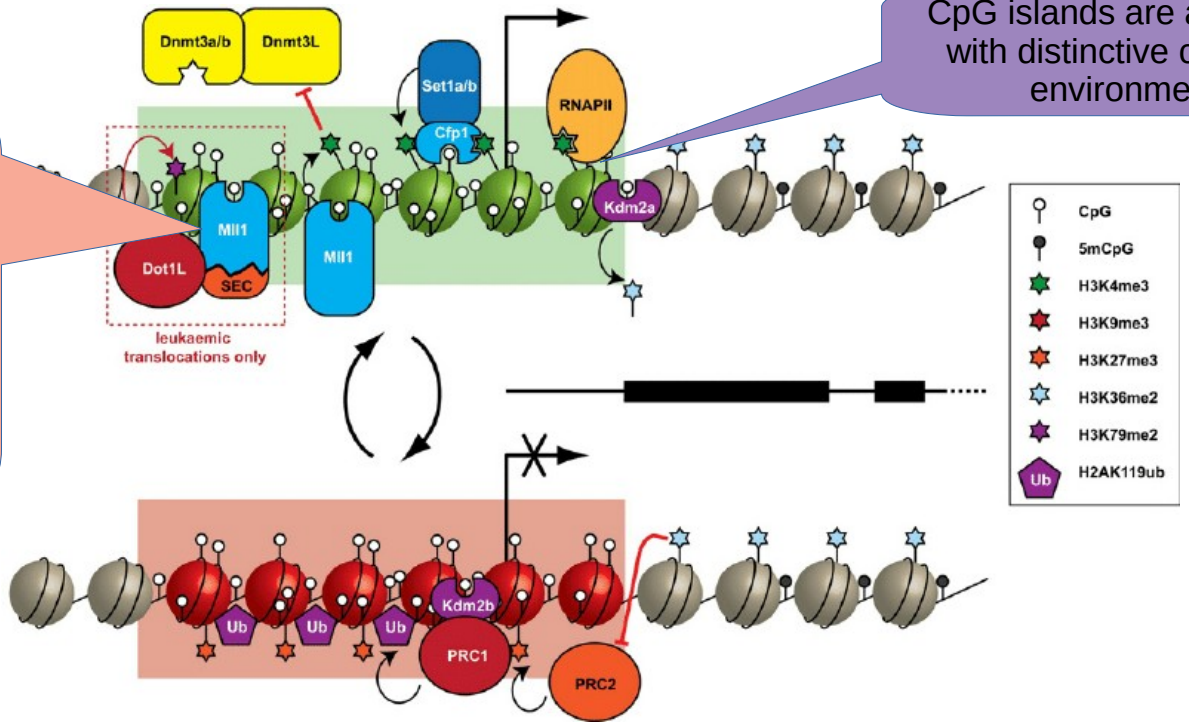

**Left**: CpG sites at 1/10 nucleotides, constituting a CpG island. The sample is of a gene-promoter, the highlighted ATG consitutes the start codon.

**Right**: CpG sites present at every 1/100 nucleotides, consituting a more normal example of the genome - a non-coding region

# Finding Promoters Using CpG Islands

CpG islands are associated with distinctive chromatin environments.

CpG islands associated with actively transcribed genes recruit H3K4 Methyltransferases (Mll1/2, Set1a/b)




- Promoter regions tend to have a higher frequency of C and G nucleotides relative to A and T nucleotides
- The CG dinucleotide occurs in promoter regions more frequently than would be expected by chance
- CpG targets for methylation and epigenetic regulation of gene expression