Bioinformatics

CS300

Genome annotation: Advanced (eukaryotic) gene prediction

Fall 2017

Oliver Bonham-Carter

Prediction Algorithms

- **Alignment-based** – sequence similarity to previously identified gene in another organism (BLAST)

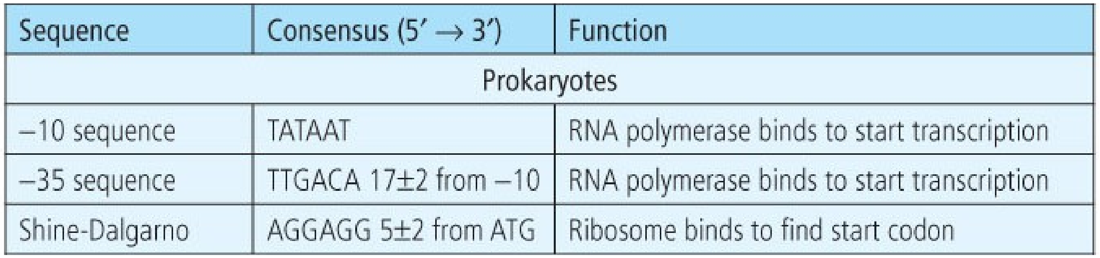
- **Sequence-based** – search for specific sequences - e.g. ORF–finder – searches start and stop codons

- **Content-based** – search for patterns – e.g. nucleotide or codon frequency

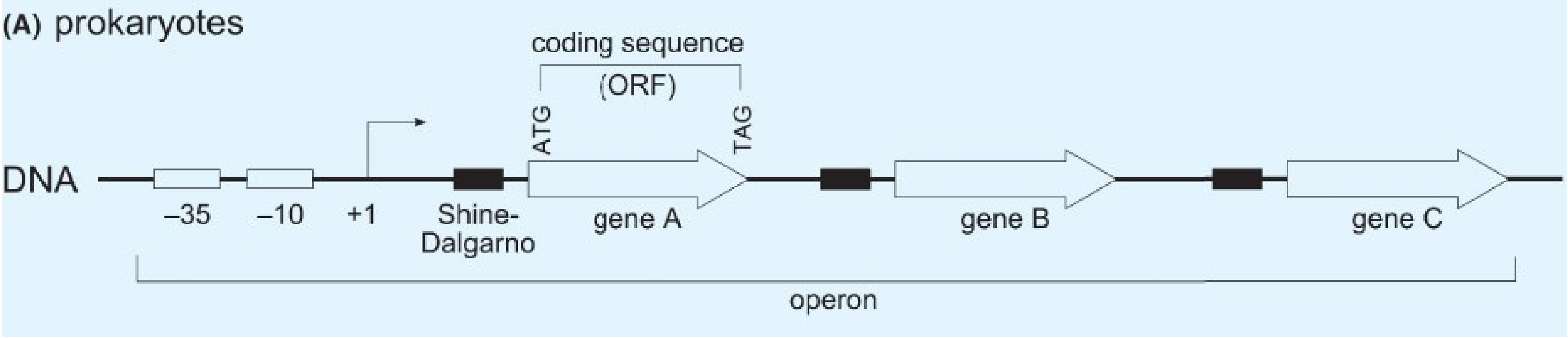
- **Probabilistic** – combination of sequence- and content-based plus probability that sequence is part of a gene

Prokaryotic Gene Structure

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

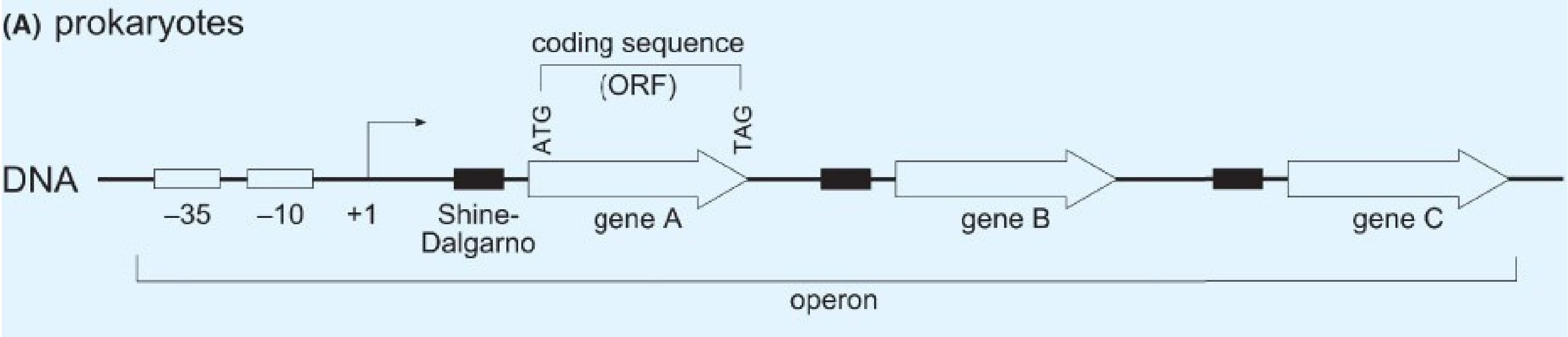


Sequence	Consensus (5′ → 3′)	Function
Prokaryotes		
–10 sequence	TATAAT	RNA polymerase binds to start transcription
–35 sequence	TTGACA 17±2 from –10	RNA polymerase binds to start transcription
Shine-Dalgarno	AGGAGG 5±2 from ATG	Ribosome binds to find start codon



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)

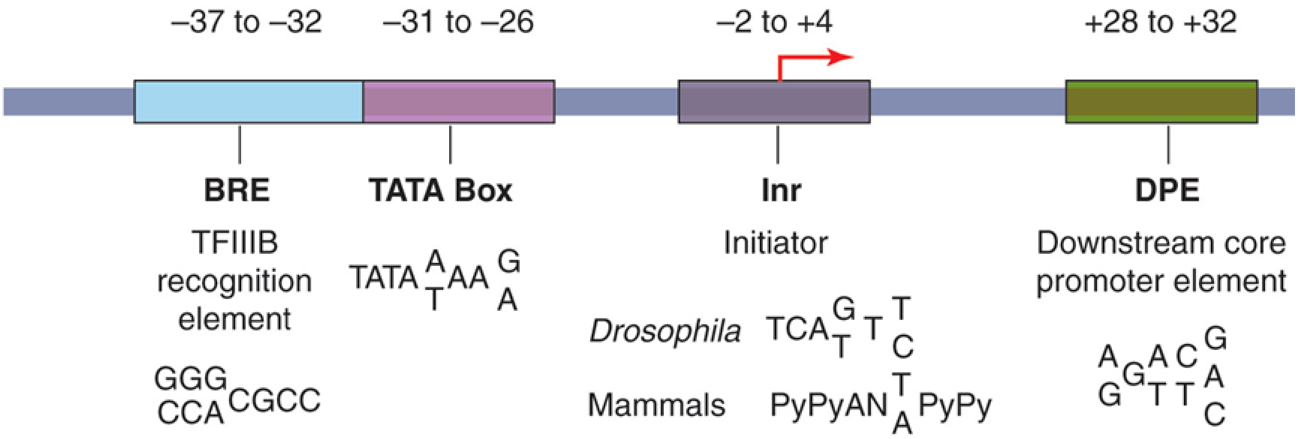


Prokaryotic Gene Structure

- Conserved -10 and -35 promoter sequences
- Conserved Shine-Dalgarno sequence marks the start codon
- Few interrupted ORFs (introns are rare)

Eukaryotic Gene Structure

- Variable promoter structure
  - not all promoter elements present in all gene
  - promoter element sequence can vary between genes
  - no conserved Shine-Delgarno-like sequence



–37 to –32

–31 to –26

–2 to +4

+28 to +32

BRE

TFIIIB
recognition
element

GGG
CCA CGCC

TATA Box

TATA A/T AA G/A

Inr

Initiator

Drosophila TCA G/T T T/C

Mammals PyPyAN T/A PyPy

DPE

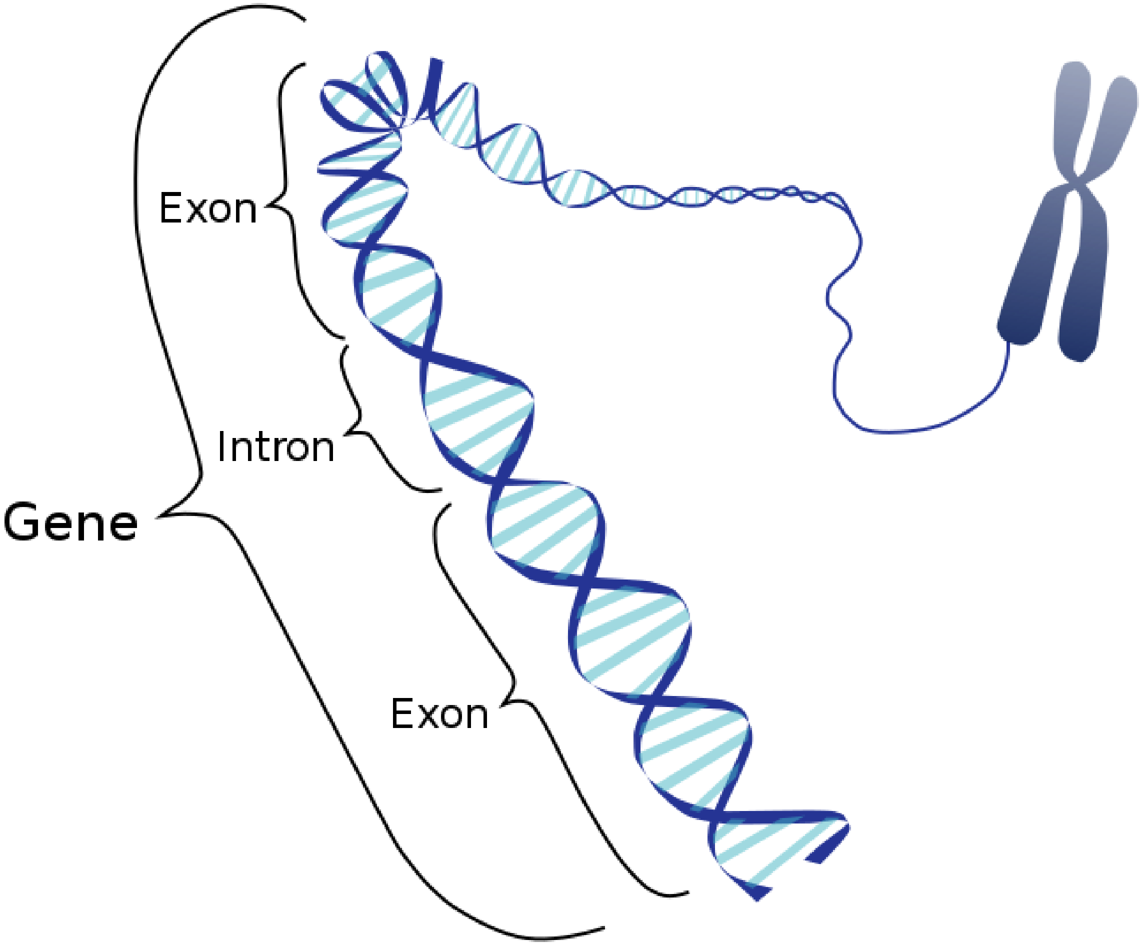
Downstream core
promoter element

A/G G A/T C/T G/A/C

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.

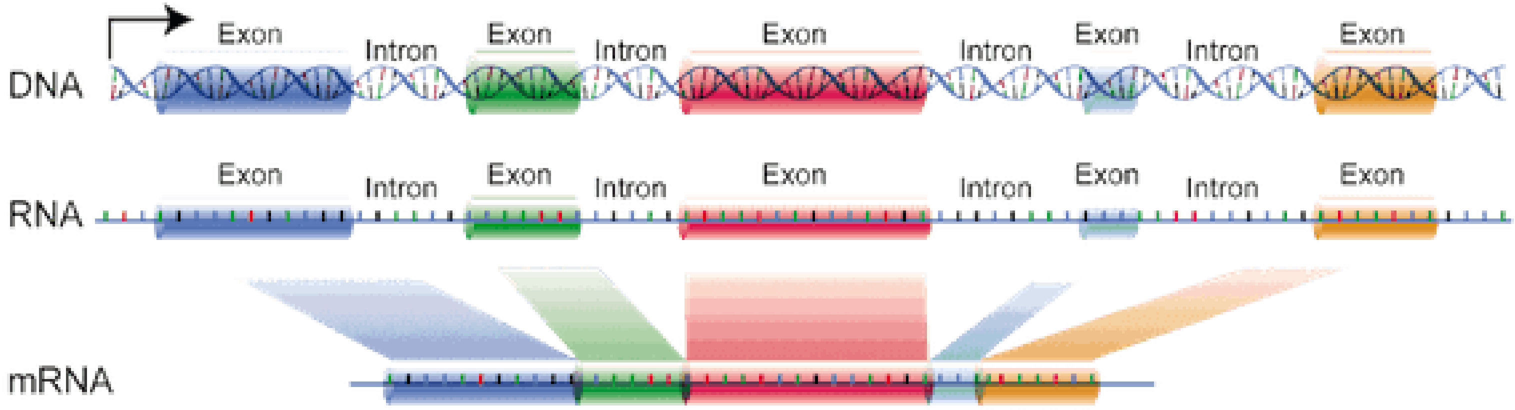
- An **exon** (expressed region) is any part of a gene that will encode a part of the final mature RNA produced by that gene after introns have been removed by RNA splicing.



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

Eukaryotic Gene Structure

- Most genes contain introns

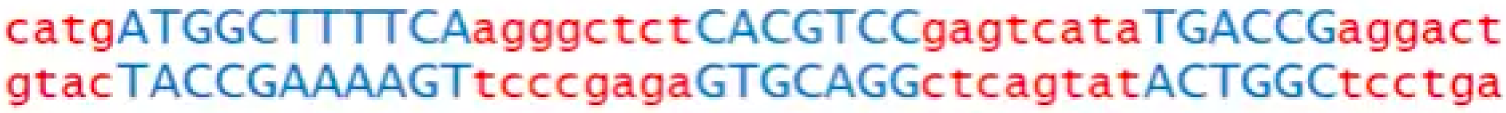


- Only interested in coding region
  - Exons only
    - Sequence with codons

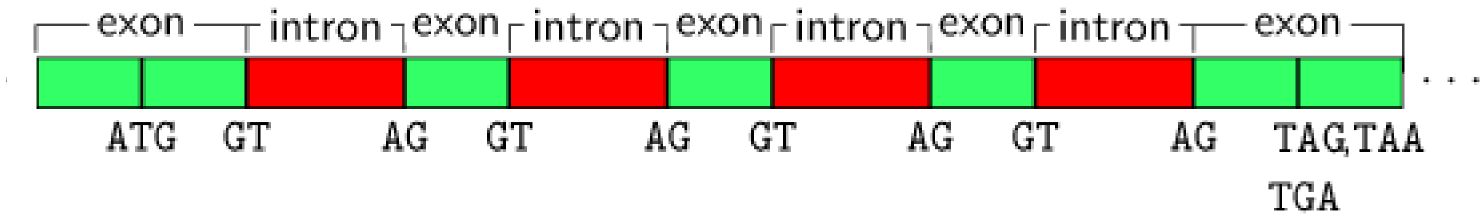
Eukaryotic Gene Structure

- 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.

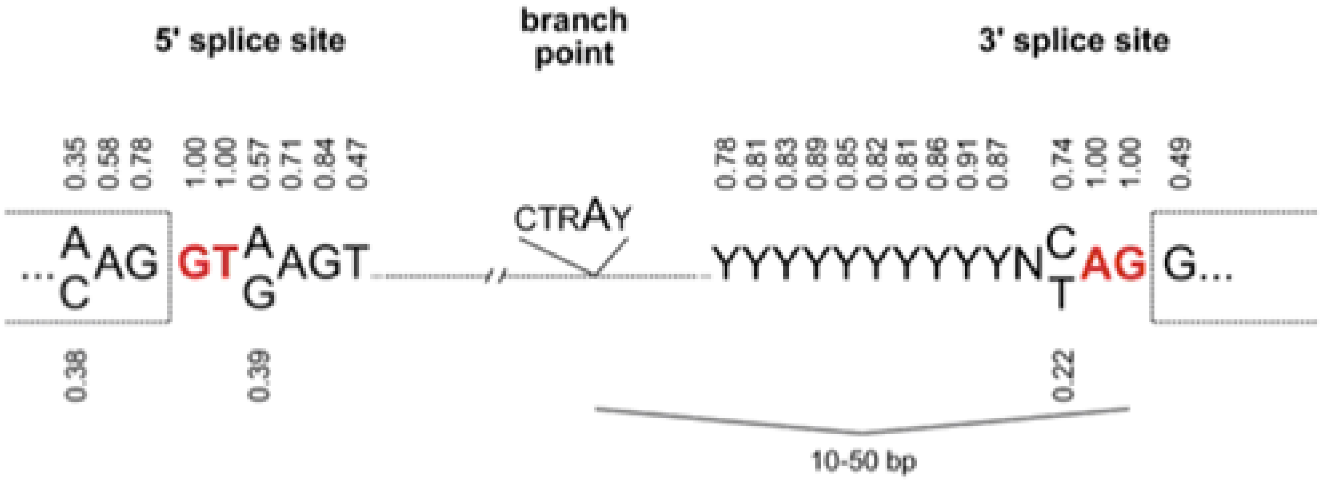
- **Exon**/**Intron** boundaries not highly conserved



catgATGGCTTTTCAagggctctCACGTCCgagtcataTGACCGaggact
gtacTACCGAAAAGTtcccgagaGTGCAGGctcagtatACTGGCtcctga



Eukaryotic Gene Structure

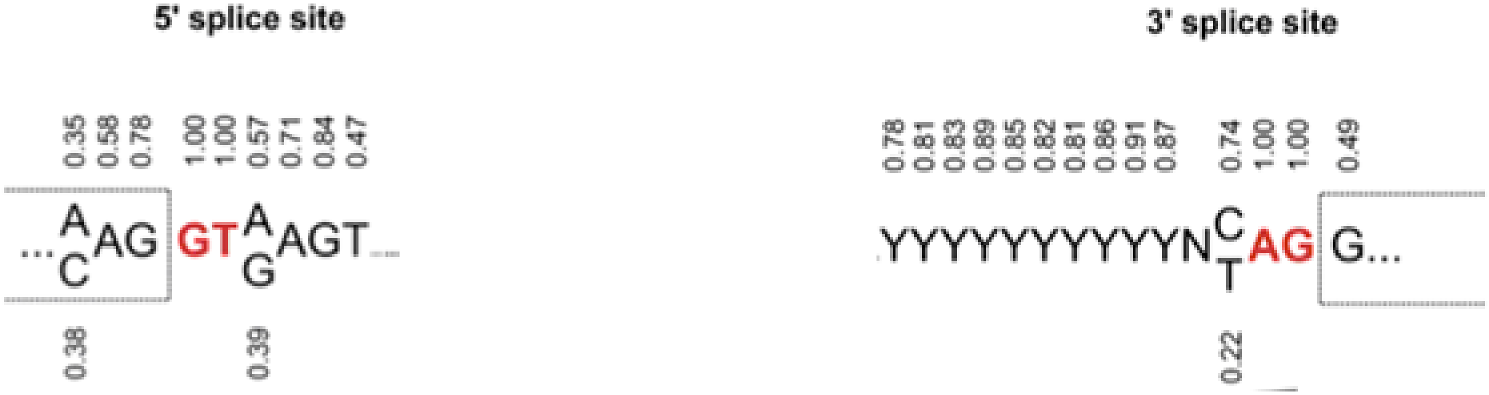


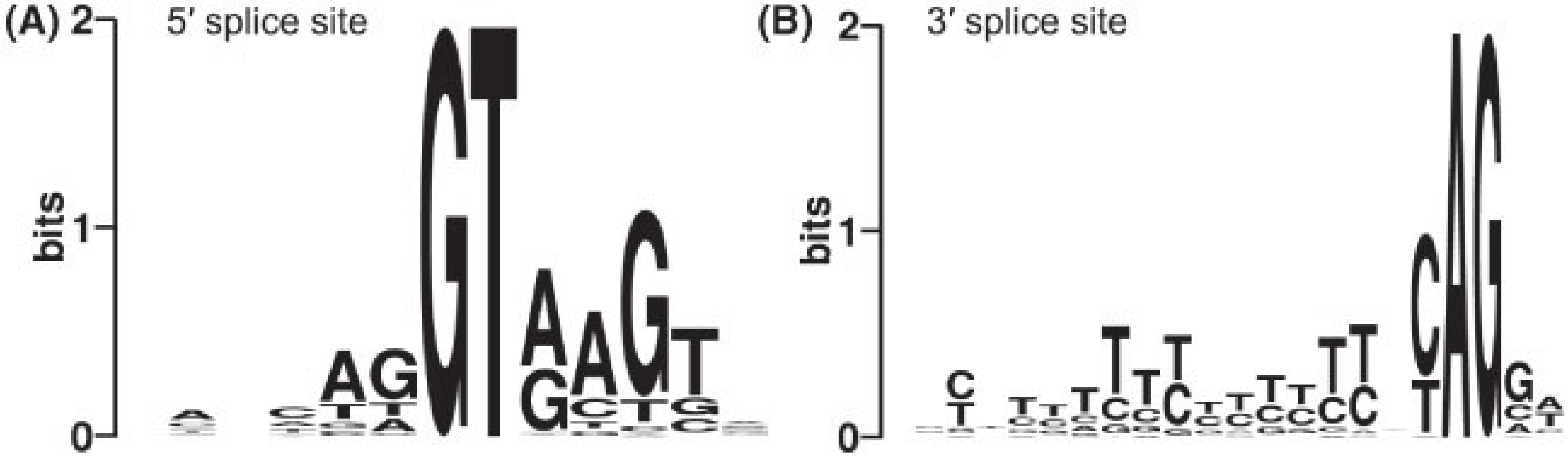
- Exon/Intron boundaries not highly conserved
- In general, codon usage is the same within one org.
- However, when the usage fluctuates noticably from the norm, we suspect an intron /exon boundry has been crossed.

Eukaryotic Gene Structure

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

- Exon/Intron boundaries not highly conserved





Prokaryotic Gene Structure

- -10 and -35 promoter sequences
- Shine-Dalgarno sequence marks the start codon
- Few interrupted ORFs (introns are rare)

Eukaryotic Gene Structure

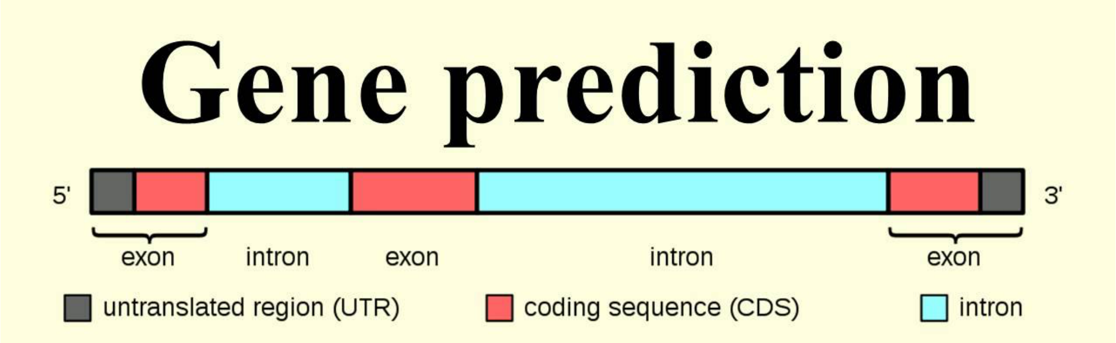
- Promoter sequences vary in number and sequence
- No Shine-Dalgarno – unambiguous identification of transcriptional start site is difficult
- Nearly all genes contain introns
- Intron/exon boundaries are hard to discern

Bioinformatics Solution?

- Content- and Probability-Based Gene Prediction

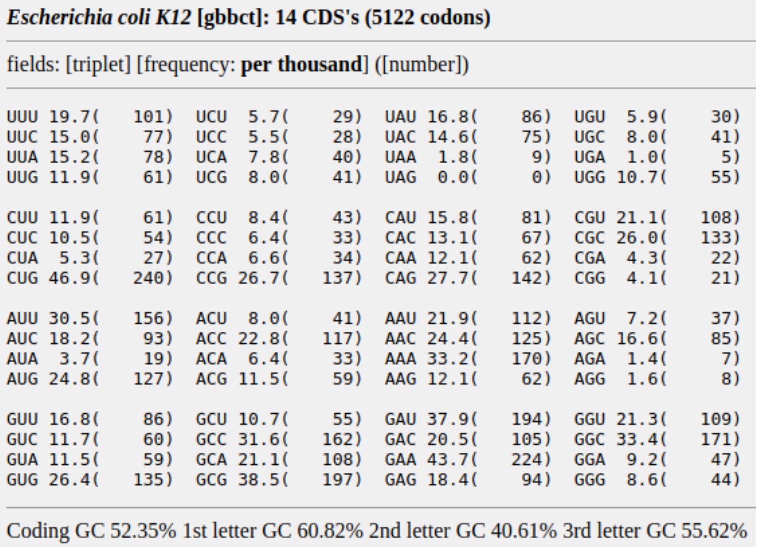
- Content-based gene prediction
  - Alone not very precise
  - Better result when used in combination with other algorithms

- Content
  - Codon usage
  - CpG Islands



Codon Usage and Frequency

Synonymous codons for same amino acid not used with equal frequency



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

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

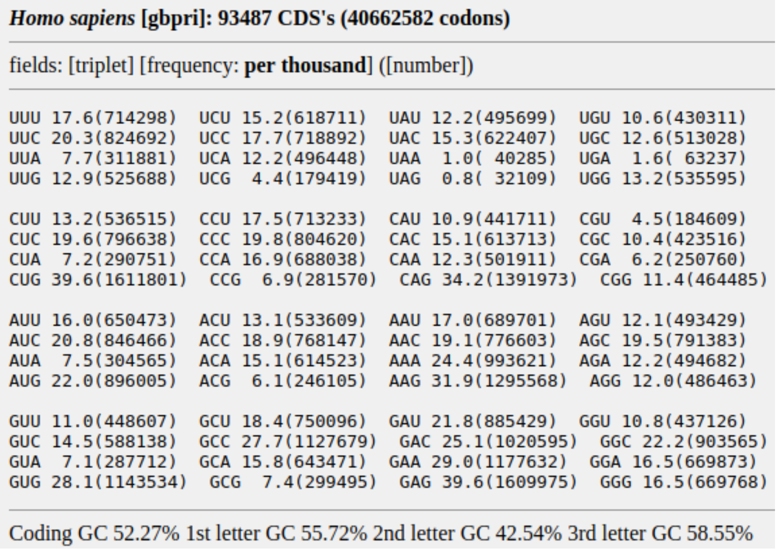
```
UUU 19.7(   101)  UCU  5.7(    29)  UAU 16.8(    86)  UGU  5.9(    30)
UUC 15.0(    77)  UCC  5.5(    28)  UAC 14.6(    75)  UGC  8.0(    41)
UUA 15.2(    78)  UCA  7.8(    40)  UAA  1.8(     9)  UGA  1.0(     5)
UUG 11.9(    61)  UCG  8.0(    41)  UAG  0.0(     0)  UGG 10.7(    55)

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%

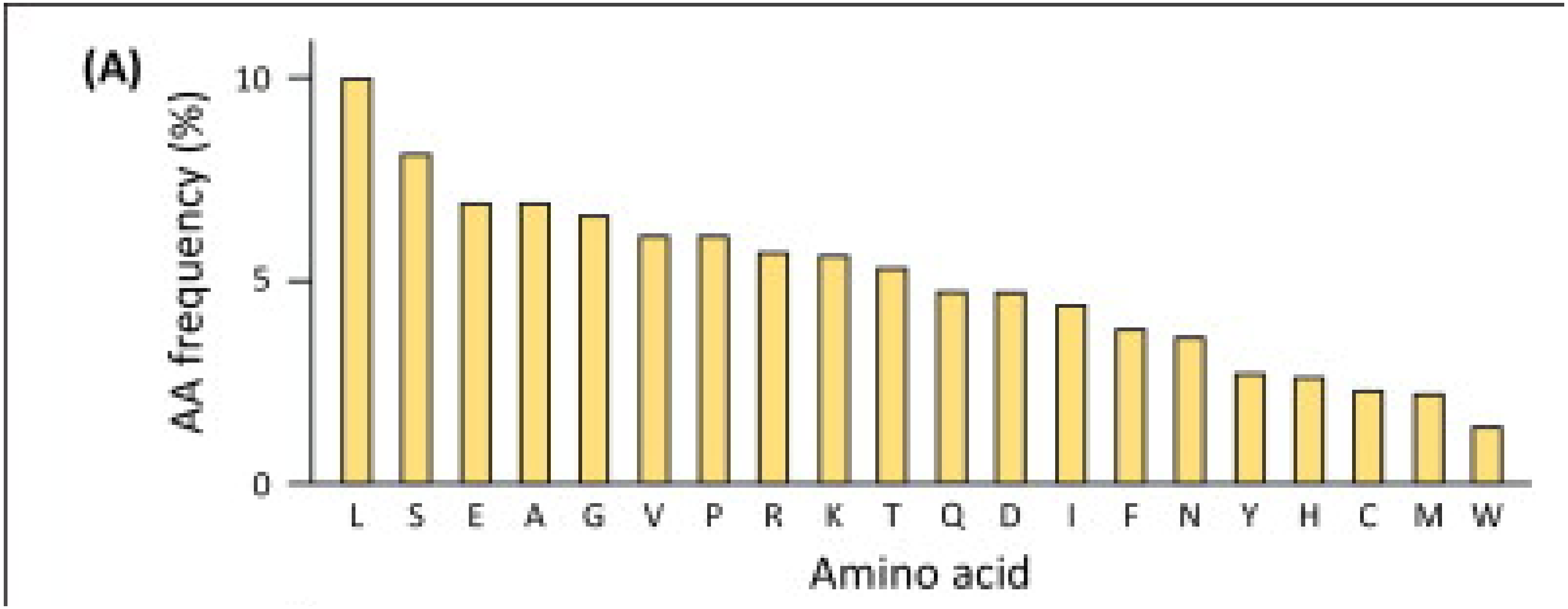
E. coli

H. sapiens

http://www.kazusa.or.jp/codon/

Codon Usage and Frequency

Some amino acids are much more common in proteins than others

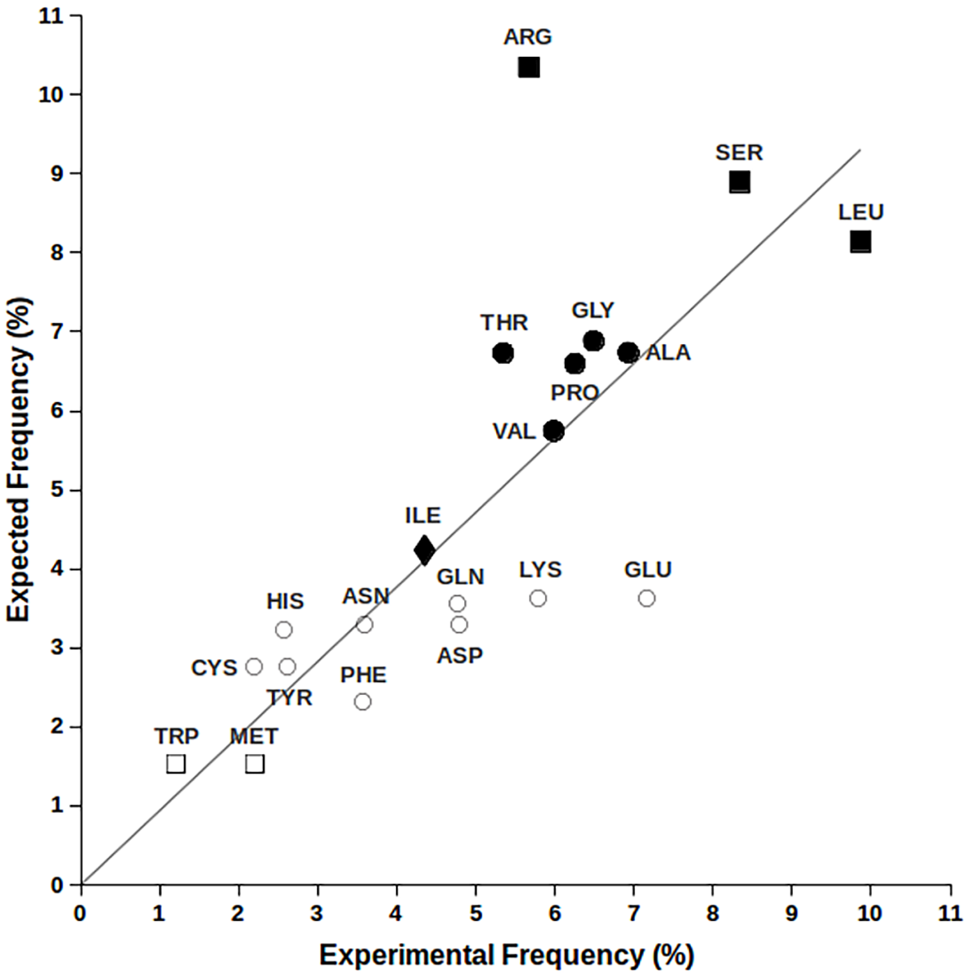


Codon Usage and Number

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 a R2 = 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.

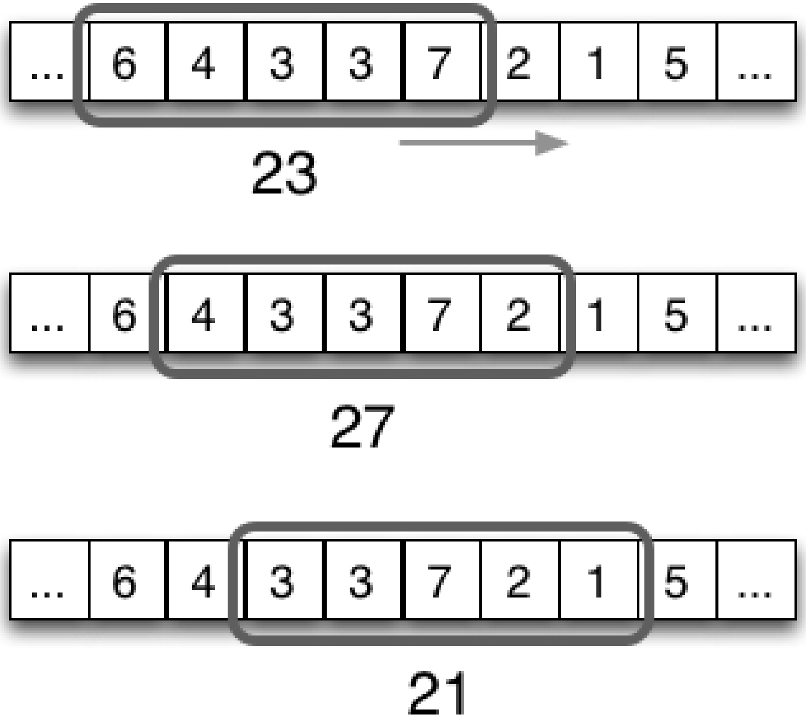
Using Codon Frequencies to Find Exon/Intron Boundaries

- Expectations:
  - Exon – codon frequency closely matches expected frequency for a gene
  - Intron – “codon” frequency poorly matches expected frequency for a gene (because not really codons!!)
  - Boundary – point where frequencies shift

Using Codon Frequencies to Find Exon/Intron Boundaries

- Sliding-Window Approach

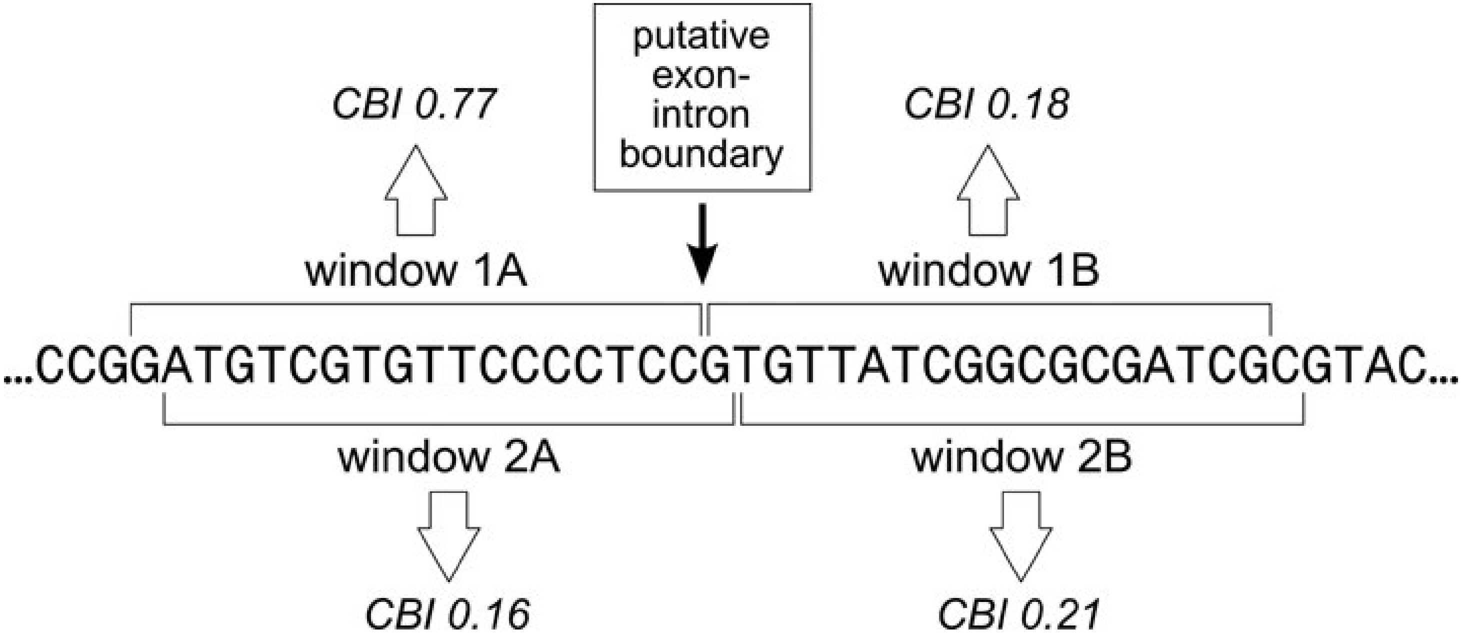
Slide a window along the sequence to read the frequency scores of the codons.



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.

Using Codon Frequencies to Find Exon/Intron Boundaries



putative exon-intron boundary

CBI 0.77

window 1A

CBI 0.18

window 1B

...CCGGATGTCGTGTTCCCCTCCGTGTTATCGGCGCGATCGCGTAC...

window 2A

CBI 0.16

window 2B

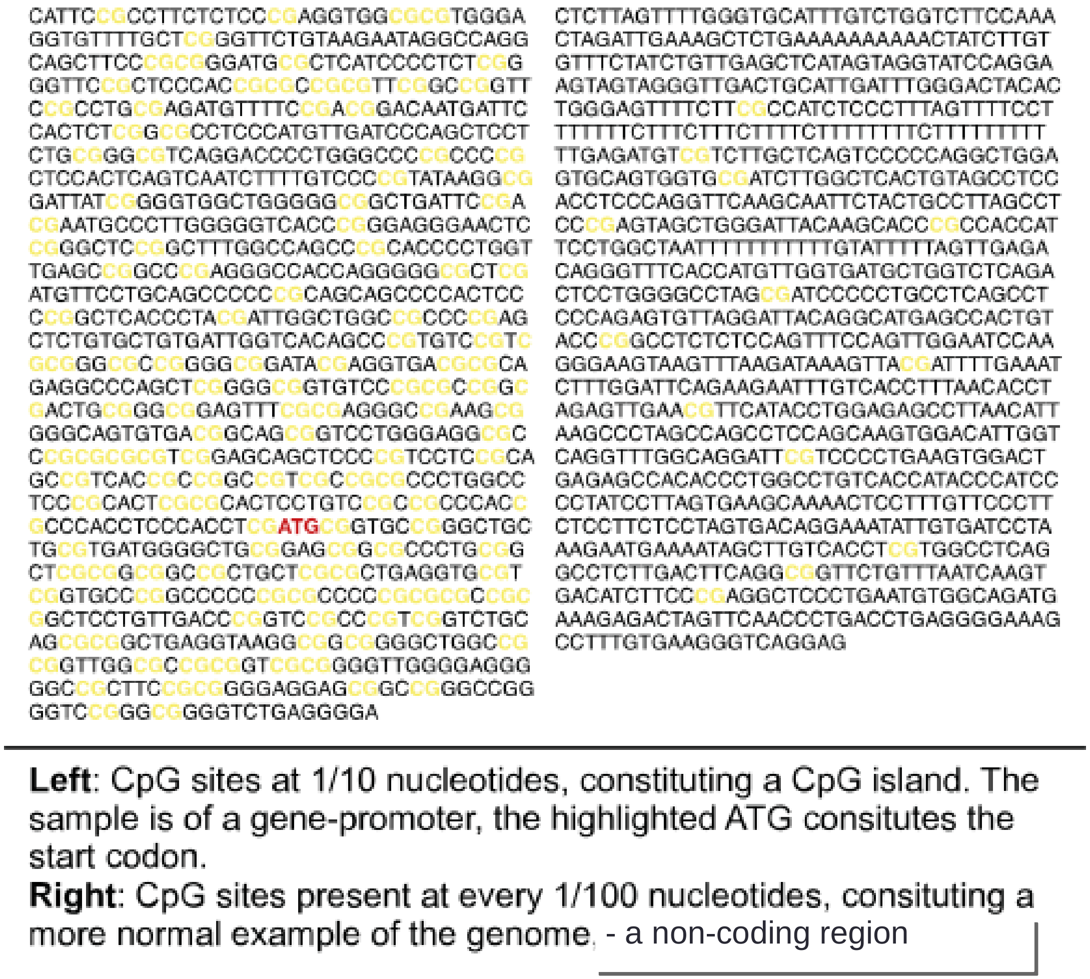
CBI 0.21

- **CBI** = **codon bias index**
- 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

- 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

Finding Promoters Using CpG Islands



```
CATTCCGCCTTCTCTCCCGAGGTGGCGCGTGGGA
GGTGTTTTGCTCGGGTTCTGTAAGAATAGGCCAGG
CAGCTTCCCGCGGGATGCGCTCATCCCCTCTCGG
GGTTCCGCTCCCACCGCGCCGCGTTCGGCCGGTT
CCGCCTGCGAGATGTTTTCCGACGGACAATGATTC
CACTCTCGGCGCCTCCCATGTTGATCCCAGCTCCT
CTGCGGGCGTCAGGACCCCTGGGCCCCGCCCCG
CTCCACTCAGTCAATCTTTTGTCCCCGTATAAGGCG
GATTATCGGGGTGGCTGGGGGCGGCTGATTCCGA
CGAATGCCCTTGGGGGTCACCCGGGAGGGAACTC
CGGGCTCCGGCTTTGGCCAGCCCGCACCCCTGGT
TGAGCCGGCCCGAGGGCCACCAGGGGGCGCTCG
ATGTTCCTGCAGCCCCCCGCAGCAGCCCCACTCC
CCGGCTCACCCTACGATTGGCTGGCCGCCCCGAG
CTCTGTGCTGTGATTGGTCACAGCCCGTGTCCGTC
GCGGGCGCCGGGGCGGATACGAGGTGACGCGCA
GAGGCCCAGCTCGGGGCGGTGTCCCGCGCCGGC
GACTGCGGGCGGAGTTTCGCGAGGGCCGAAGCG
GGGCAGTGTGACGGCAGCGGTCCTGGGAGGCGC
CCGCGCGCGTCGGAGCAGCTCCCCGTCCTCCGCA
GCCGTCACCGCCGGCCGTCGCCGCGCCCTGGCC
TCCCGCACTCGCGCACTCCTGTCCGCCGCCCACC
GCCCACCTCCCACCTCGATGCGGTGCCGGGCTGC
TGCGTGATGGGGCTGCGGAGCGGCGCCCTGCGG
CTCGCGGCGGCCGCTGCTCGCGCTGAGGTGCGT
CGGTGCCCGGCCCCCCGCGCCCCCGCGCGCCGC
GGCTCCTGTTGACCCGGTCCGCCCGTCGGTCTGC
AGCGCGGCTGAGGTAAGGCGGCGGGGCTGGCCG
CGGTTGGCGCCGCGGTCGCGGGGTTGGGGAGGG
GGCCGCTTCCGCGGGGAGGAGCGGCCGGGCCGG
GGTCCGGGCGGGGTCTGAGGGGA
```

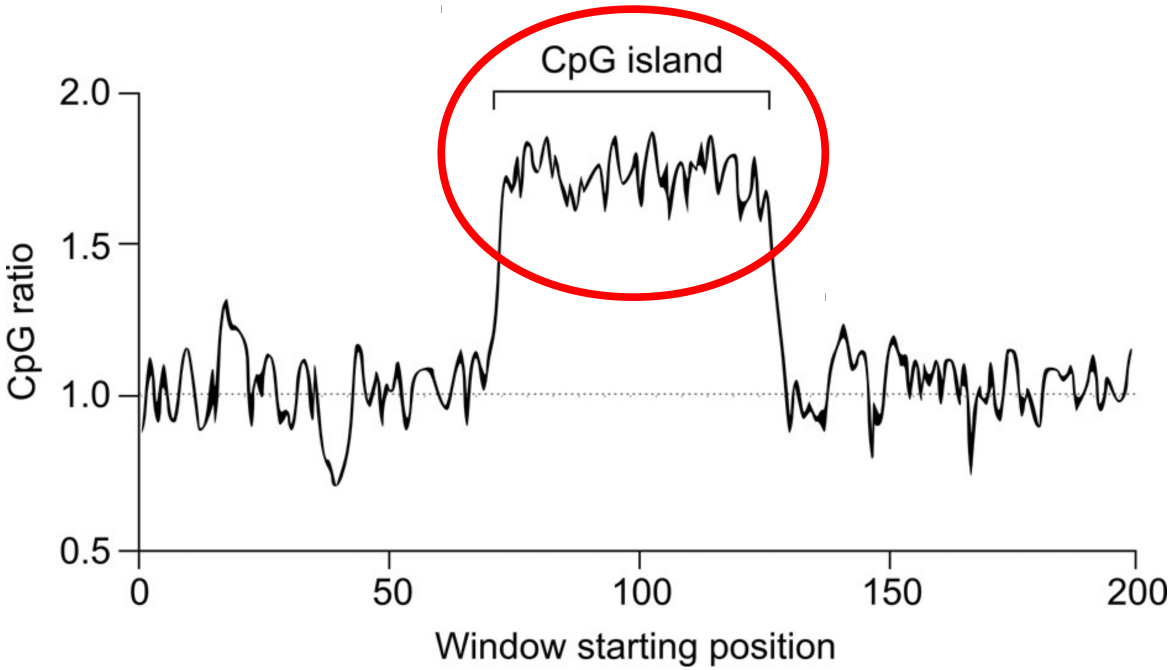
```
CTCTTAGTTTTGGGTGCATTTGTCTGGTCTTCCAAA
CTAGATTGAAAGCTCTGAAAAAAAAAAACTATCTTGT
GTTTCTATCTGTTGAGCTCATAGTAGGTATCCAGGA
AGTAGTAGGGTTGACTGCATTGATTTGGGACTACAC
TGGGAGTTTTCTTCGCCATCTCCCTTTAGTTTTCCT
TTTTTTCTTTCTTTCTTTTCTTTTTTTTTCTTTTTTTTTT
TTGAGATGTCGTCTTGCTCAGTCCCCCAGGCTGGA
GTGCAGTGGTGCGATCTTGGCTCACTGTAGCCTCC
ACCTCCCAGGTTCAAGCAATTCTACTGCCTTAGCCT
CCCGAGTAGCTGGGATTACAAGCACCCGCCACCAT
TCCTGGCTAATTTTTTTTTTTTGTATTTTTAGTTGAGA
CAGGGTTTCACCATGTTGGTGATGCTGGTCTCAGA
CTCCTGGGGGCCTAGCGATCCCCCTGCCTCAGCCT
CCCAGAGTGTTAGGATTACAGGCATGAGCCACTGT
ACCCGGCCTCTCTCCAGTTTCCAGTTGGAATCCAA
GGGAAGTAAGTTTAAGATAAAGTTACGATTTTGAAAT
CTTTGGATTCAGAAGAATTTGTCACCTTTAACACCT
AGAGTTGAACGTTCATACCTGGAGAGCCTTAACATT
AAGCCCTAGCCAGCCTCCAGCAAGTGGACATTGGT
CAGGTTTGGCAGGATTCGTCCCCTGAAGTGGACT
GAGAGCCACACCCTGGCCTGTCACCATACCCATCC
CCTATCCTTAGTGAAGCAAAACTCCTTTGTTCCCTT
CTCCTTCTCCTAGTGACAGGAAATATTGTGATCCTA
AAGAATGAAAATAGCTTGTCACCTCGTGGCCTCAG
GCCTCTTGACTTCAGGCGGTTCTGTTTAATCAAGT
GACATCTTCCCGAGGCTCCCTGAATGTGGCAGATG
AAAGAGACTAGTTCAACCCTGACCTGAGGGGAAAG
CCTTTGTGAAGGGTCAGGAG
```

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

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



$$\frac{\textit{observed CG pairs} \quad .}{\textit{C nucleotides x G nucleotides/total nucleotides}}$$