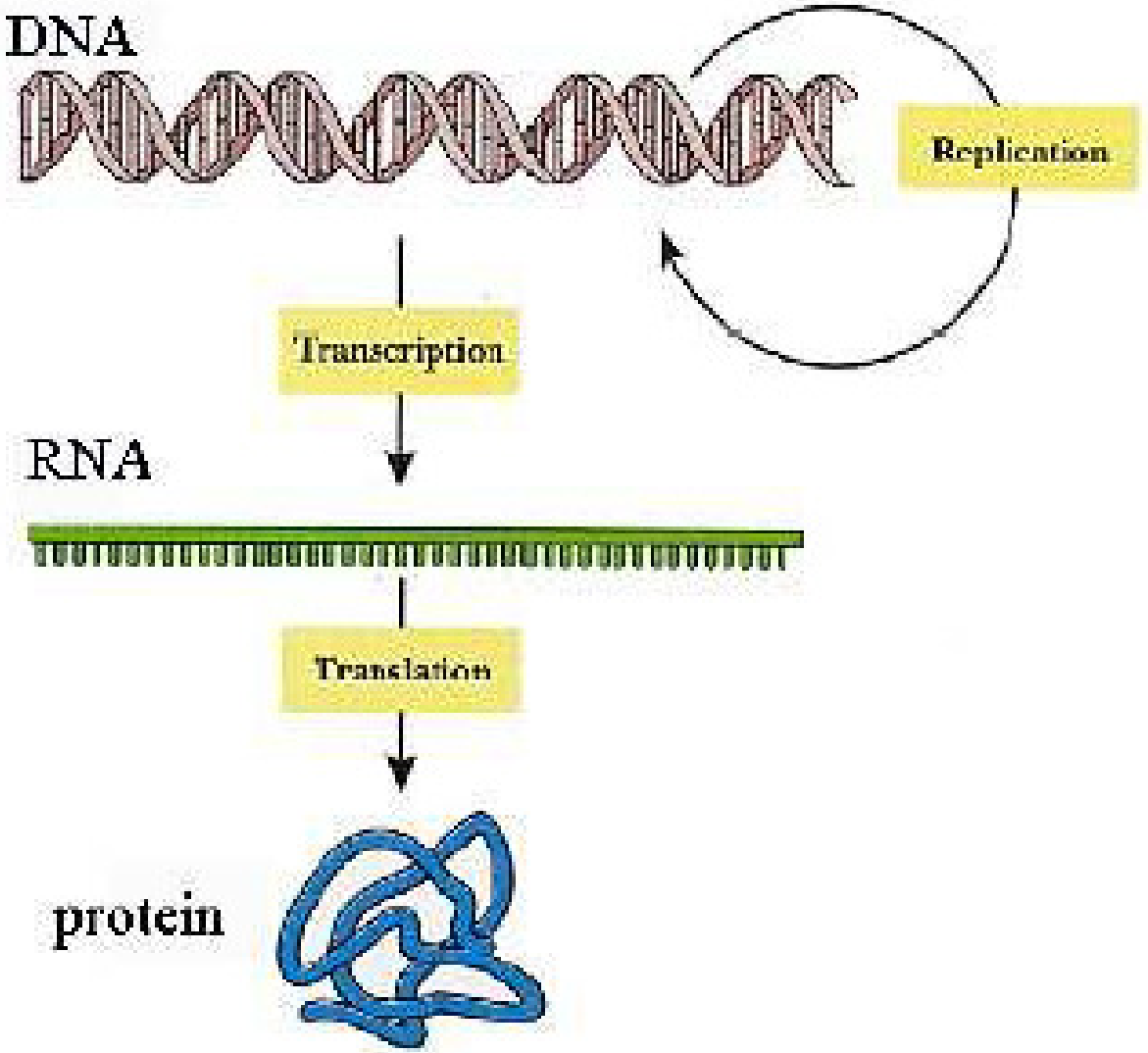
Lecture 2
Gene discovery

The Central Dogma



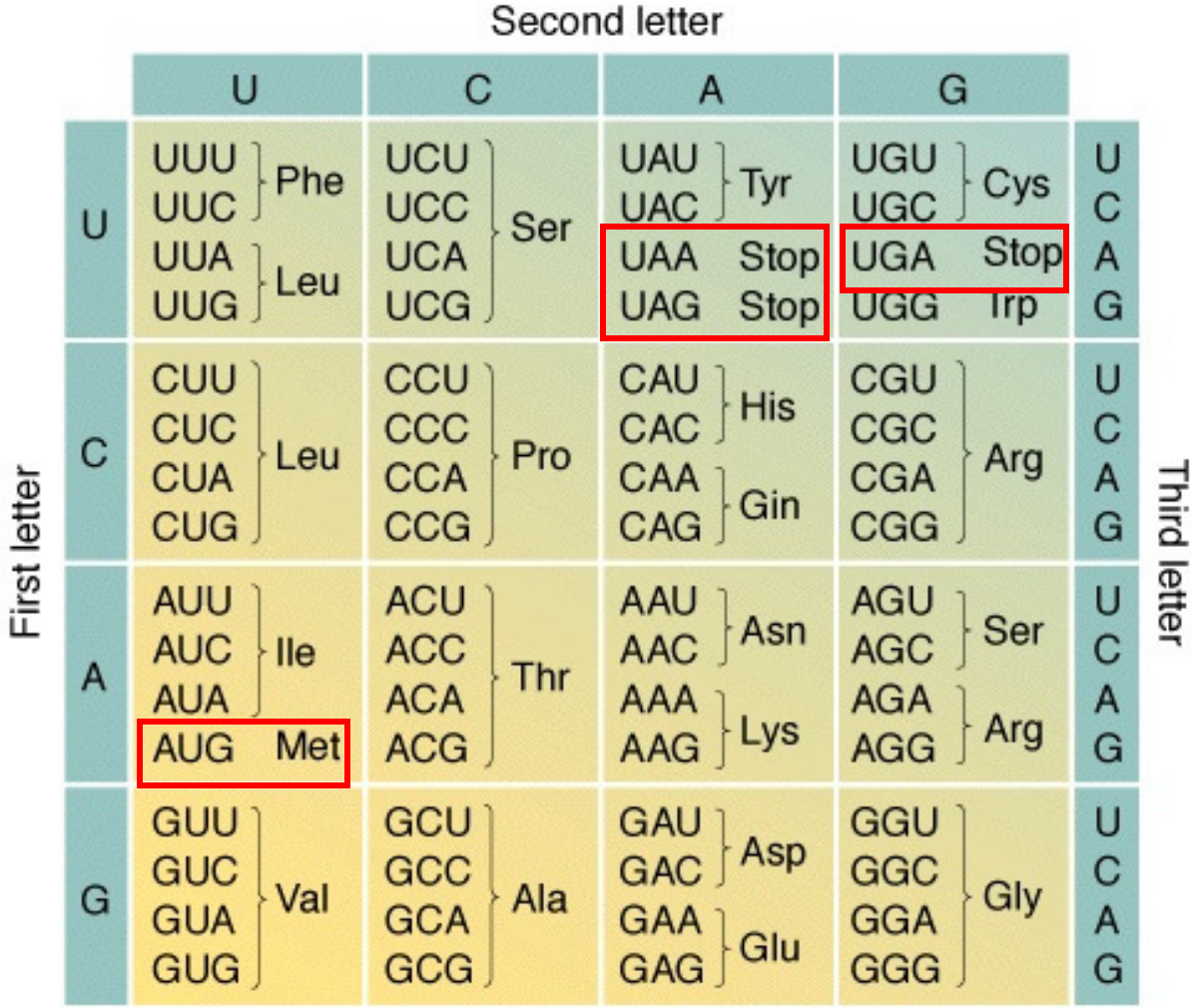
Transcription

- *RNA polymerase* is the enzyme that builds an RNA strand from a gene
- RNA that is transcribed from a gene is called *messenger RNA (mRNA)*

RNA

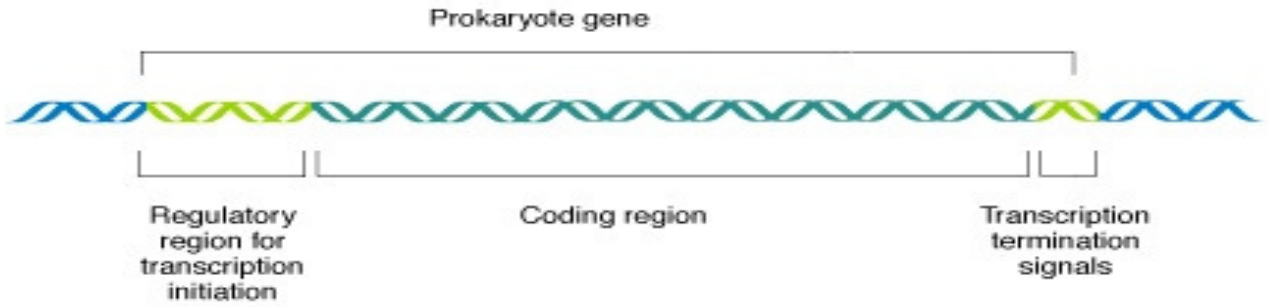
- RNA is like DNA except:
  - backbone is a little different
  - usually single stranded
  - the base uracil (U) is used in place of thymine (T)
- A strand of RNA can be thought of as a string composed of the four letters: A, C, G, U

The Genetic Code



64 combinations: 20 amino acids + stop codon

Genes include both coding regions as well as control regions



Fasta format

```
>YAH1 sacCer1.chr16:73363-73881
ATGCTGAAAATTGTTACTCGGGCTGGACACACAGCTAGAATATCGAACAT
CGCAGCACATCTTTTACGCACCTCTCCATCTCTGCTCACACGCACCACCA
CAACCACAAGATTTCTGCCCTTCTCTACGTCTTCGTTCTTAAACCATGGC
CATTTGAAAAAACCGAAACCAGGCGAAGAACTGAAGATAACTTTTATTCT
GAAGGATGGCTCCCAGAAGACGTACGAAGTCTGTGAGGGCGAAACCATCC
TGGACATCGCTCAAGGTCACAACCTGGACATGGAGGGCGCATGCGGCGGT
TCTTGTGCCTGCTCCACCTGTCACGTCATCGTTGATCCAGACTACTACGA
TGCCCTGCCGGAACCTGAAGATGATGAAAACGATATGCTCGATCTTGCTT
ACGGGCTAACAGAGACAAGCAGGCTTGGGTGCCAGATTAAGATGTCAAAA
GATATCGATGGGATTAGAGTCGCTCTGCCCCAGATGACAAGAAACGTTAA
TAACAACGATTTTAGTTAA
>GAL4 sacCer1.chr16:79711-82356
ATGAAGCTACTGTCTTCTATCGAACAAGCATGCGATATTTGCCGACTTAA
AAAGCTCAAGTGCTCCAAAGAAAAACCGAAGTGCGCCAAGTGTCTGAAGA
ACAACTGGGAGTGTCGCTACTCTCCCAAAACCAAAAGGTCTCCGCTGACT
AGGGCACATCTGACAGAAGTGGAATCAAGGCTAGAAAGACTGGAACAGCT
ATTTCTACTGATTTTTCCTCGAGAAGACCTTGACATGATTTTGAAAATGG
ATTCTTTACAGGATATAAAAGCATTGTTAACAGGATTATTTGTACAAGAT
AATGTGAATAAAGATGCCGTCACAGATAGATTGGCTTCAGTGGAGACTGA
TATGCCTCTAACATTGAGACAGCATAGAATAAGTGCGACATCATCATCGG
AAGAGAGTAGTAACAAAGGTCAAAGACAGTTGACTGTATCGATTGACTCG
GCAGCTCATCATGATAACTCCACAATTCCGTTGGATTTTATGCCCAGGGA
TGCTCTTCATGGATTTGATTGGTCTGAAGAGGATGACATGTCGGATGGCT
TGCCCTTCCTGAAAACGGACCCCAACAATAATGGGTTCTTTGGCGACGGT
TCTCTCTTATGTATTCTTCGATCTATTGGCTTTAAACCGGAAAATTACAC
```

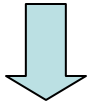
Translation

```
>YAH1 sacCer1.chr16:73363-73881
ATGCTGAAAATTGTTACTCGGGCTGGACACACAGCTAGAATATCGAACATCGCAGCA
CATCTTTTACGCACCTCTCCATCTCTGCTCACACGCACCACCACAACCACAAGATTT
CTGCCCTTCTCTACGTCTTCGTTCTTAAACCATGGCCATTTGAAAAAACCGAAACCA
GGCGAAGAACTGAAGATAACTTTTATTCTGAAGGATGGCTCCCAGAAGACGTACGAA
GTCTGTGAGGGCGAAACCATCCTGGACATCGCTCAAGGTCACAACCTGGACATGGAG
GGCGCATGCGGCGGTTCTTGTGCCTGCTCCACCTGTCACGTCATCGTTGATCCAGAC
TACTACGATGCCCTGCCGGAACCTGAAGATGATGAAAACGATATGCTCGATCTTGCT
TACGGGCTAACAGAGACAAGCAGGCTTGGGTGCCAGATTAAGATGTCAAAAGATATC
GATGGGATTAGAGTCGCTCTGCCCCAGATGACAAGAAACGTTAATAACAACGATTTT
AGTTAA
```

Codon: triplet of nucleotides
Start codon: ATG
Stop codon: TAA

Translation

>YAH1 sacCer1.chr16:73363-73881
ATGCTGAAAATTGTTACTCGGGCTGGACACACAGCTAGAATATCGAACATCGCAGCA
CATCTTTTACGCACCTCTCCATCTCTGCTCACACGCACCACCACAACCACAAGATTT
CTGCCCTTCTCTACGTCTTCGTTCTTAAACCATGGCCATTTGAAAAAACCGAAACCA
GGCGAAGAACTGAAGATAACTTTTATTCTGAAGGATGGCTCCCAGAAGACGTACGAA
GTCTGTGAGGGCGAAACCATCCTGGACATCGCTCAAGGTCACAACCTGGACATGGAG
GGCGCATGCGGCGGTTCTTGTGCCTGCTCCACCTGTCACGTCATCGTTGATCCAGAC
TACTACGATGCCCTGCCGGAACCTGAAGATGATGAAAACGATATGCTCGATCTTGCT
TACGGGCTAACAGAGACAAGCAGGCTTGGGTGCCAGATTAAGATGTCAAAAGATATC
GATGGGATTAGAGTCGCTCTGCCCCAGATGACAAGAAACGTTAATAACAACGATTTT
AGTTAA



M--L--K--I--V--T--R--A--G--H--T--A--R--I--S--N--I--A--A--
H--L--L--R--T--S--P--S--L--L--T--R--T--T--T--T--T--R--F--
L--P--F--S--T--S--S--F--L--N--H--G--H--L--K--K--P--K--P--
G--E--E--L--K--I--T--F--I--L--K--D--G--S--Q--K--T--Y--E--
V--C--E--G--E--T--I--L--D--I--A--Q--G--H--N--L--D--M--E--
G--A--C--G--G--S--C--A--C--S--T--C--H--V--I--V--D--P--D--
Y--Y--D--A--L--P--E--P--E--D--D--E--N--D--M--L--D--L--A--
Y--G--L--T--E--T--S--R--L--G--C--Q--I--K--M--S--K--D--I--
D--G--I--R--V--A--L--P--Q--M--T--R--N--V--N--N--N--D--F--
S--*

MLKIVTRAGHTARISNIAAHLLRTSPSLLTRTTTTTRFLPFSTSSFLNHGHLKKPKPG
EELKITFILKDGSQKTYEVCEGETILDIAQGHNLDMEGACGGSCACSTCHVIVDPDYY
DALPEPEDDENDMLDLAYGLTETSRLGCQIKMSKDIDGIRVALPQMTRNVNNNDFS*

If reading frame is unknown

TCTCTACGATGCTGAAAATTGTTACTCGGGCTGGACACACAGCTAGAATATCGAACA
TCGCAGCACATCTTTTACGCACCTCTCCATCTCTGCTCACACGCACCACCACAACCA
CAAGATTTCTGCCCTTCTCTACGTCTTCGTTCTTAAACCATGGCCATTTGAAAAAAC
CGAAACCAGGCGAAGAACTGAAGATAACTTTTATTCTGAAGGATGGCTCCCAGAAGA
CGTACGAAGTCTGTGAGGGCGAAACCATCCTGGACATCGCTCAAGGTCACAACCTGG
ACATGGAGGGCGCATGCGGCGGTTCTTGTGCCTGCTCCACCTGTCACGTCATCGTTG
ATCCAGACTACTACGATGCCCTGCCGGAACCTGAAGATGATGAAAACGATATGCTCG
ATCTTGCTTACGGGCTAACAGAGACAAGCAGGCTTGGGTGCCAGATTAAGATGTCAA
AAGATATCGATGGGATTAGAGTCGCTCTGCCCCAGATGACAAGAAACGTTAATAACA
ACGATTTTAGTTAATGCCCTGC

Open reading frame (ORF)

- One can represent a genome of length n as a sequence of n/3 codons
- The three ‘stop’ codons (TAA,TAG, and TGA) break this sequence into segments, one between every two consecutive stop codons
- The subsegments of these that start from a start codon (ATG) are ORFs

Six reading frames

```
TCTCTACGATGCTGAAAATTGTTACTCGGGCTGGACACACAGCTAGAATATCGTGAA


reading frame 1
TCTCTACGATGCTGAAAATTGTTACTCGGGCTGGACACACAGCTAGAATATCGTGAA
S--L--R--C--*--K--L--L--L--G--L--D--T--Q--L--E--Y--R--E--

reading frame 2
 CTCTACGATGCTGAAAATTGTTACTCGGGCTGGACACACAGCTAGAATATCGTGAA
 L--Y--D--A--E--N--C--Y--S--G--W--T--H--S--*--N--I--V--

reading frame 3
  TCTACGATGCTGAAAATTGTTACTCGGGCTGGACACACAGCTAGAATATCGTGAA
  S--T--M--L--K--I--V--T--R--A--G--H--T--A--R--I--S--*

reading frame 4 (reverse complement frame 1)
TTCACGATATTCTAGCTGTGTGTCCAGCCCGAGTAACAATTTTCAGCATCGTAGAGA
F--T--I--F--*--L--C--V--Q--P--E--*--Q--F--S--A--S--*--R

reading frame 5 (reverse complement frame 2)
 TCACGATATTCTAGCTGTGTGTCCAGCCCGAGTAACAATTTTCAGCATCGTAGAGA
 S--R--Y--S--S--C--V--S--S--P--S--N--N--F--Q--H--R--R

reading frame 6 (reverse complement frame 3)
  CACGATATTCTAGCTGTGTGTCCAGCCCGAGTAACAATTTTCAGCATCGTAGAGA
  H--D--I--L--A--V--C--P--A--R--V--T--I--F--S--I--V--E
```

Size of ORF

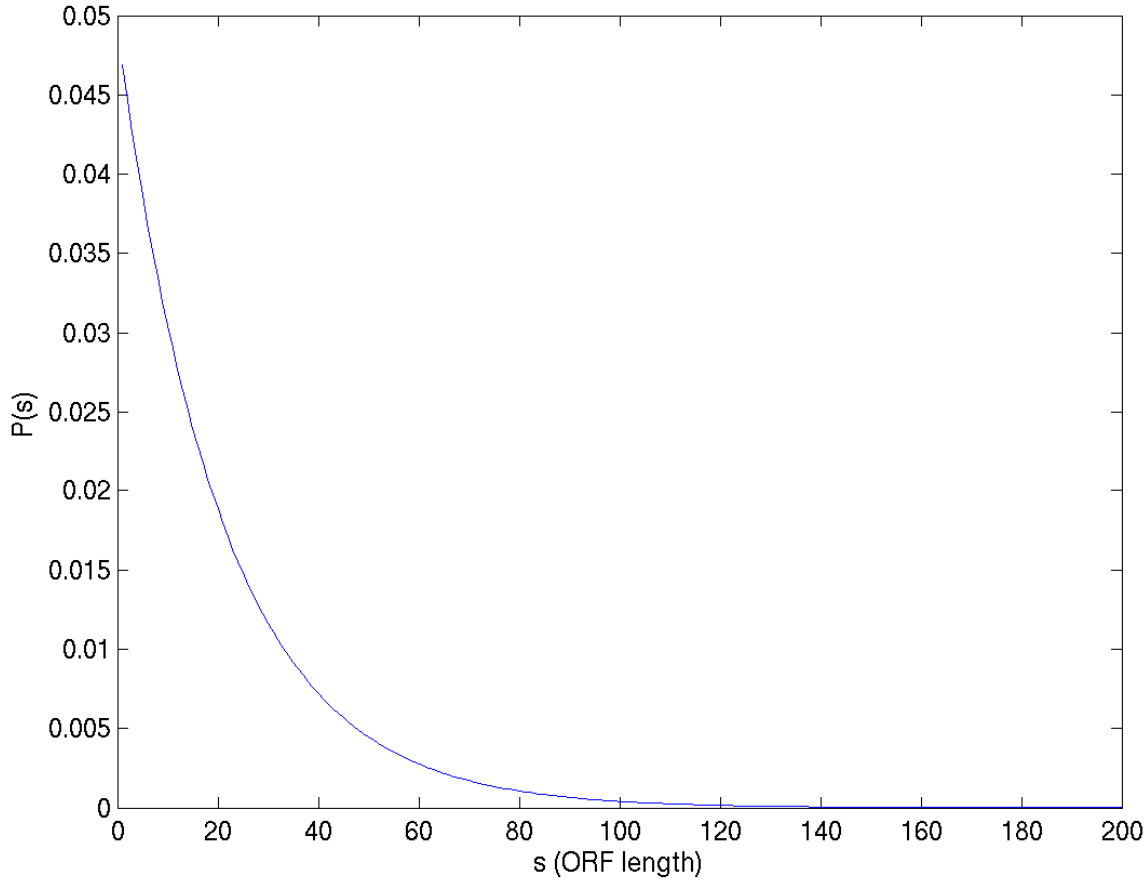
- Total number of codons: $4^3 = 64$
- Assuming random occurrences of A,C,G,Ts with equal probability:
- The probability of a codon being start codon is: 1/64
- The probability of a codon being stop codon is: 3/64

Define ***S*** to be the length of an ORF (the number of codons, excluding the stop-codon)

Question: what is the probability distribution of ***S*** ?

Distribution of randomly occurred ORF length

- $P(S=s) = (1-p)^{s-1} p$ where $p = 3/64$, $s>0$



Significance measure

Suppose you discovered an ORF with length s.
How surprised is this, if assuming A,C,G,Ts are randomly distributed?

Statistics:

- **Null model:** A,C,G,Ts are randomly distributed with equal probability -> $P(s)=(1-p)^{s-1}p$

- **P-value:** The probability of observing an ORF with **S≥s** under the null model.

P-value = $P(S \geq s) = \sum_{x=s}^{\infty} (1-p)^{x-1}p = 1 - \sum_{x=1}^{s} (1-p)^{x-1}p$

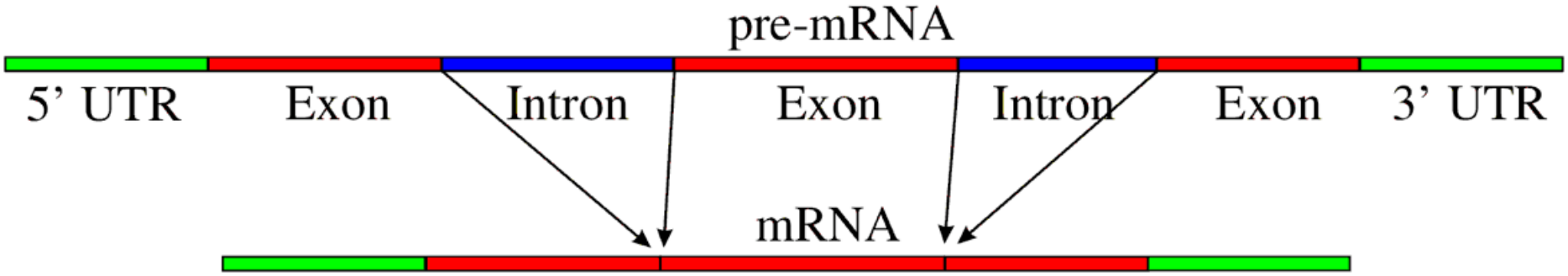
Gene discovery in higher order organisms

More complicated than ORF discovery due to more complex gene structure: multiple exons separated by introns.

Methods:

1. Statistical models of codon usage
2. Markov models of gene structure
3. Comparing across different species

RNA Splicing: pre mRNA --> mRNA



Genes include both coding regions as well as control regions