

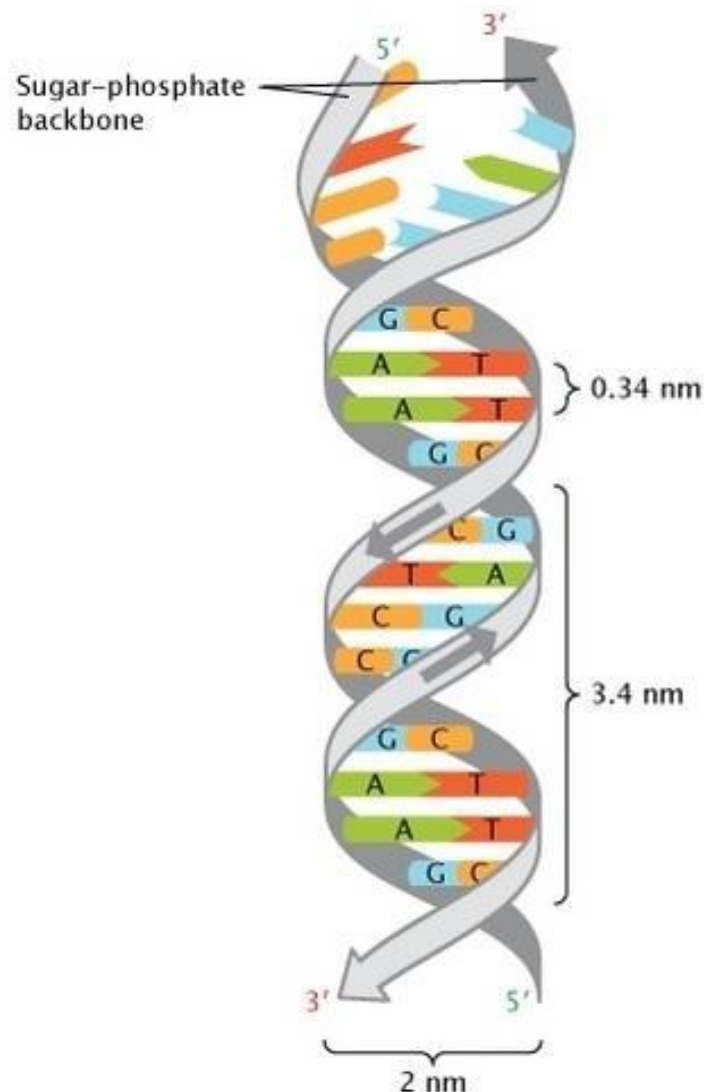
- Genomic DNA is a long string of the characters {T, C, A, G}.
- For example, the 4th chromosome of the yeast genome consists of about 1.5 million such characters.

```
chrIV.fa
1 chrIV
2 ACACCACACCCACACCACCCACACACCACCCACACACCACCC
3 ACACCACACACCCACACCACCCACACACCACCCACACCACCC
4 ACACCACACACCCACACCACCCACACACCACCCACCCACACCA
5 CACACTACCCCTAACACTACCCCTATTCTAACCCCTGATTTTACCTGTCTCC
6 AAACCTACCCCTCACATTACCCCTACCTCCCACTCGTTACCCCTGCCCACT
7 CAACCATCCACTCCCAACCACCATCCATCTCTACTTACCACTAACCCAC
8 CGTCCACCATAACCGTTACCCCTCCAATTACCCATATCCAACCTCCACTACC
9 ACTTACCCCTACTATTACCCCTACCATCCACCATGTCCTACTCACTGTACTG
10 TTGTTCTACCCCTCCATATTGAAACGTTAACAAATGATCGTAAATAATACA
11 TACATACTTACCCCTACCACTCTATACCACCACTACCAACACCGCCACTTG
12 CCACACTCACCTTCACCTTCTACTGATATGTCATACGCACCGGATGCTAC
13 AGTATATACCCTCTCAAACTTACCCCTACTTTTATATTCCACACCATGGCC
14 CCATTCTCACTAAATCAGTACCAATGCACTCACATCATTATTCACGGCA
15 CTTGCTCAGCGGTCTATACCCCTGTGCCATTACGCATAACTCCCACGAT
16 TATCCACATTTTATTACCTATATCTCATTGCGGGGCCCAAAATATTGTAT
17 AACTGCTCTTAATACATACGTTATACCACTTTTACACCGTATACACCA
18 CTCAATTTATATACACTTATGCCAATATTACAAAAAATCACCCTAAAA
19 TCACCTAAACATAAAAAATTTCTATTCTTCAACCATATATACATAACACA
20 CTTAATTGCGTCTTAATACTATCATGGTATCATTAACTTAAAGTTCCCTA

:

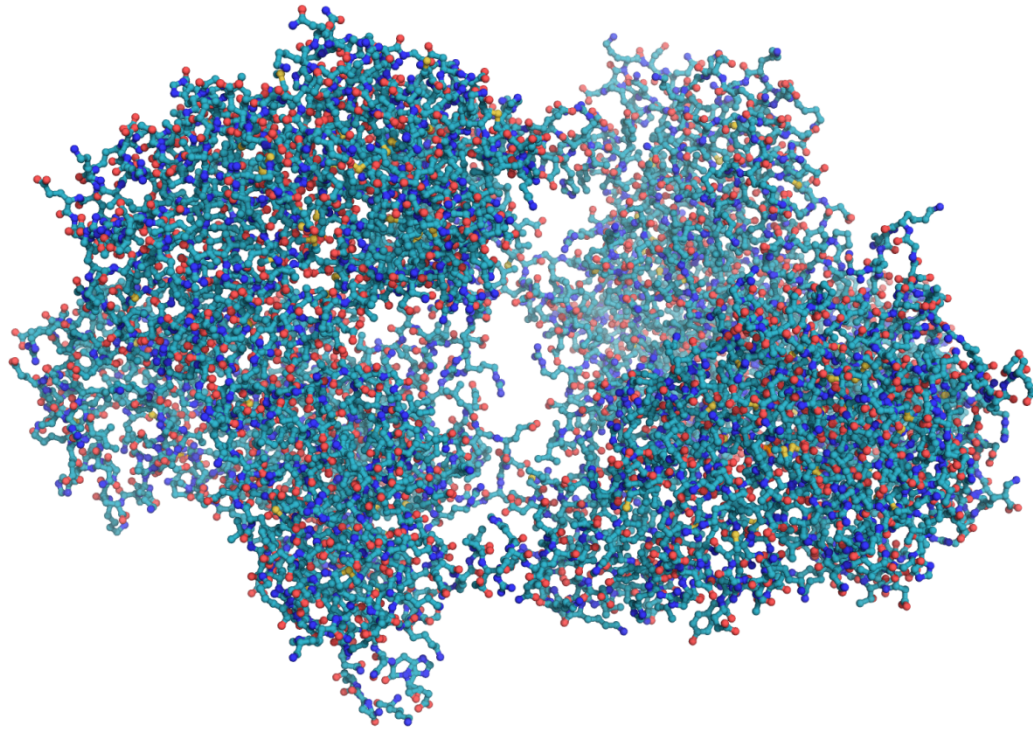
30618 TCGAGCTCTGTGCTGTTGGGCTGGCCAGAAAGTATCGTATCGGAGGATG
30619 GCTTGGGAAGCACTAGCTGTGGAGAGAATGCTGCGAAATGACGAGGAATA
30620 CAAAGAATACTTGAAGACATCGAGCCATATCATGGGACCCCTGTAGGAT
30621 ATTTGAAATATTTTAGCGTAAAAAGGGGAGAGATCTACTCTCAGATACAG
30622 AGAAATTATGCTTGGTACCTGGCCATTACTAGAAGAAGAGAAACAATTAG
30623 TGTATTGGATTTCGACAAGAGGCAAGCAAGGGAGCCAAAGTTTCCGCATGT
30624 CTGGAAGGCAGATCAAAGAGTTGTATTATAAAGTATGGAGCAACTTGCGT
30625 GAATCGAAGACAGAGGTGCTGCAGTACTTTTGAACCTGGGACGAGAAAAA
30626 GTGCCGGGAAGAATGGGAGGCAAAAGACGATACGGTCTTTGTGGAAGCGC
30627 TCGAGAAAGTTGGAGTTTTCAGCGTTTTCGCTTCCATGACGAGCGCTGGA
30628 CTGAGGGTCCGACGTACGTCAAGCTGCAGTTTAGCAGGCATCATCGAC
30629 GTTGAGGAGCAGATATGAATTAAGTCTAGGAATGCACTTGCAGATCAGC
30630 TTGCGCTGGGAGTTACCCCATCTAAAGTGCCGATTGGACGGCATTCCTG
30631 TCGATGCTGATAGGGCTGTTCTACAATAAAACATTTTCGGCAGAAACTGGA
30632 ATATCTTTTGGAGCAGATTTTCGGAGGTGTGGTTGTTACCACTTGGCTTG
30633 ATTTGGCAAACGTTGAAGTTCTCGCTGCAGATAACACGAGGGTACCGCTG
30634 TACATGCTGATGGTAGCGGTTACAAAGAGCTGGATAGCGATGATGTTCC
30635 AGACGGTAGATTTGATATAATATTACTATGTAGAGATTTCGAGCAGAGAAG
30636 TTGGAGAGTGAAGGAAATTGTTGTACGAAAGTCAGTGATTATGATTGT
30637 GTAGTATAGTATATTGTAAGAAATTTTTTTTCTAGGGAATATGCGTTTT
30638 GATGTAGTAGTATTTCACTGTTTTGATTTAGTGTGTTGTCACGGCAGTA
30639 GCGAGAGACAAGTGGGAAAGAGTAGGATAAAAAAGACAATCTATAAAAGT
30640 AAACATAAAATAAAGGTAGTAAGTAGCTTTTGG
```

- DNA is double-stranded – a sequence is paired with its reverse complement (A → T, G → C).



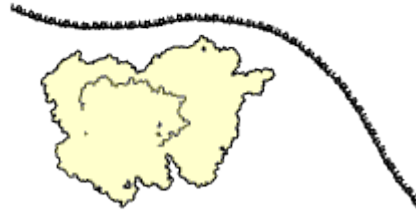
(from nature.com)

- Some (but not all) parts of this string encode instructions for making complicated molecular machines called proteins out of building blocks called amino acids.



(from wikipedia)

- The process of making a protein from these instructions is called translation.



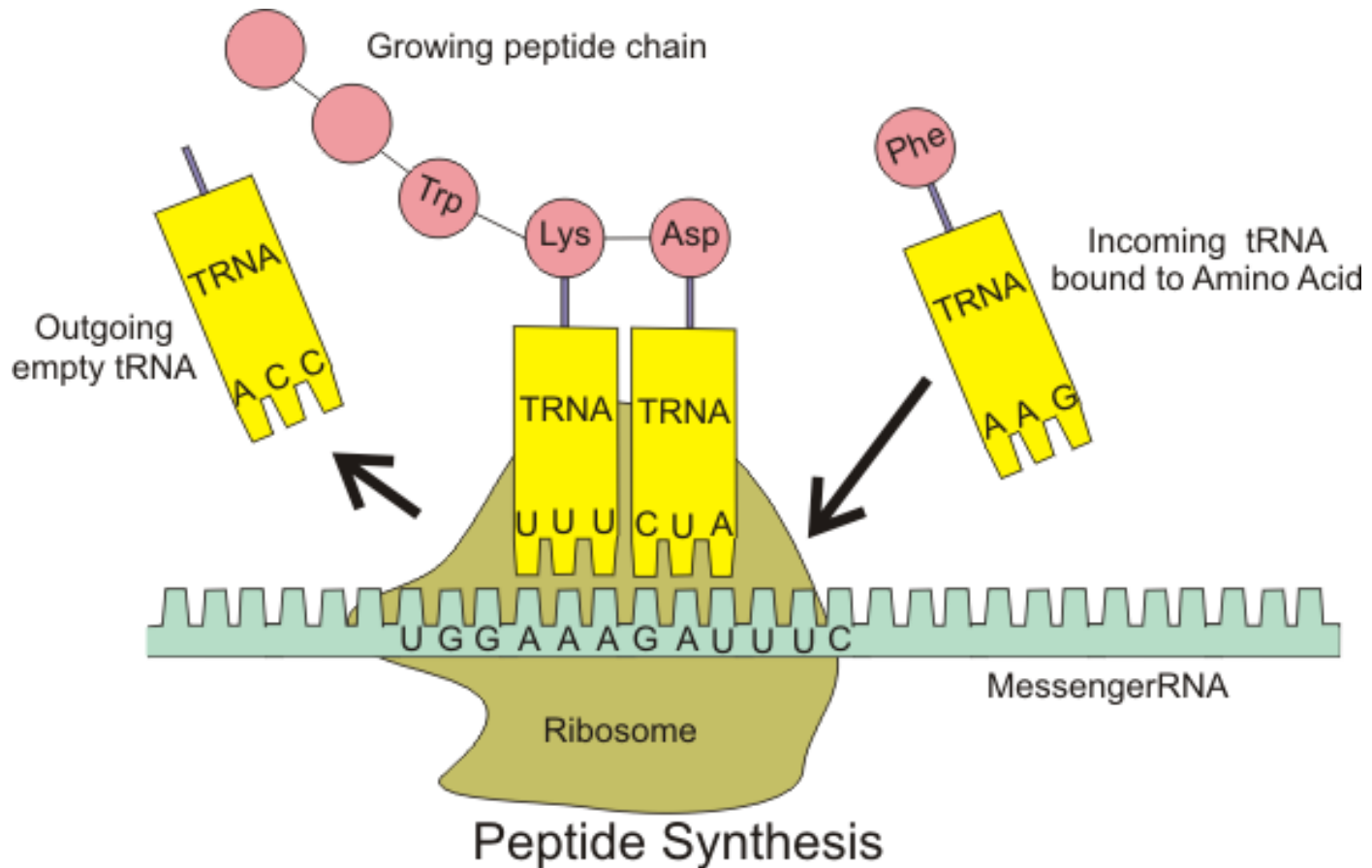
(from wikipedia)

- Each triplet of nucleotides (called a *codon*) encodes the identity of an amino acid to be used to build the protein.

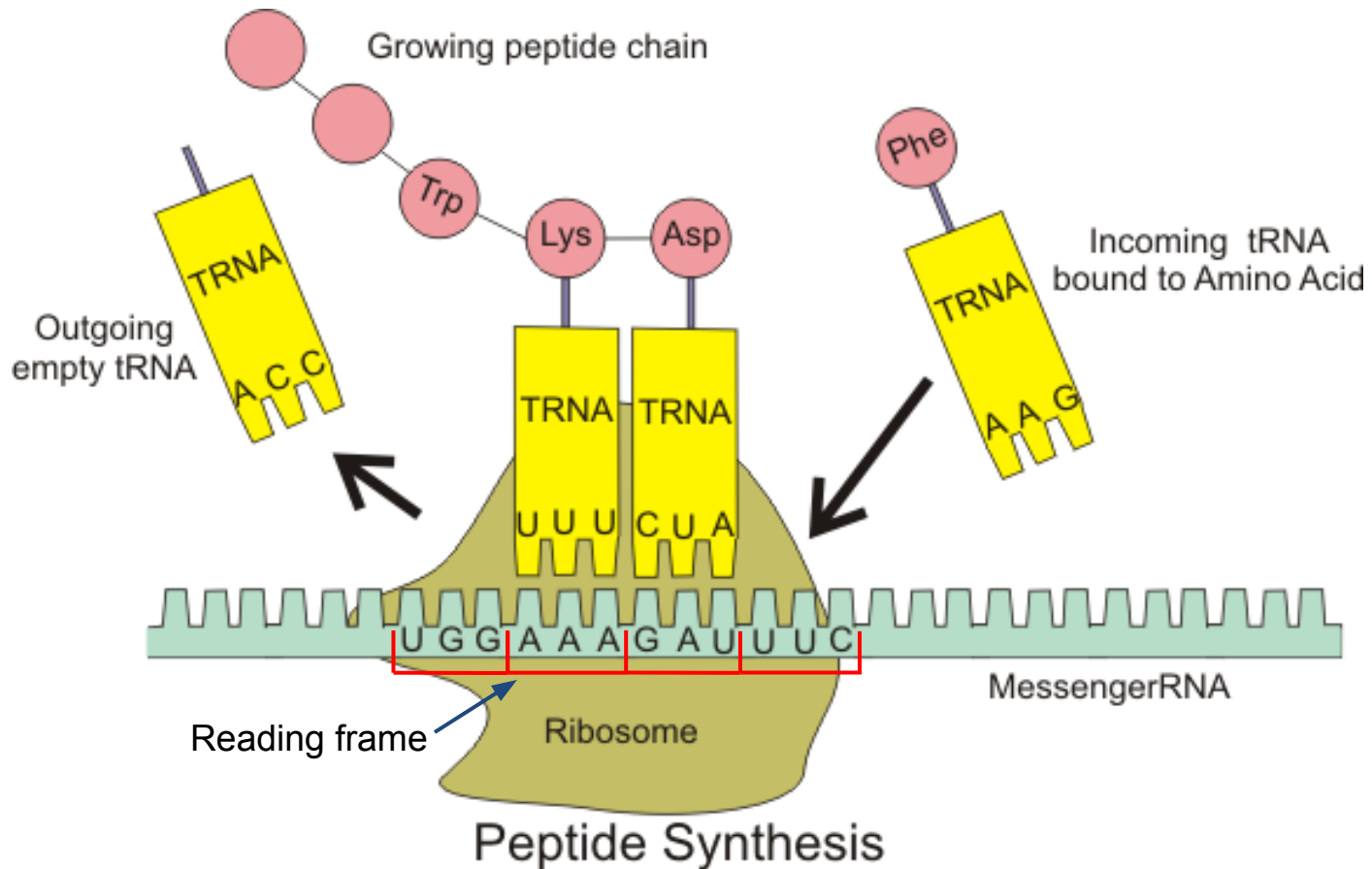
		Second Letter					
		T	C	A	G		
First Letter	T	TTT } Phe TTC } TTA } Leu TTG }	TCT } TCC } Ser TCA } TCG }	TAT } Tyr TAC } TAA Stop TAG Stop	TGT } Cys TGC } TGA Stop TGG Trp	T	C
	C	CTT } CTC } Leu CTA } CTG }	CCT } CCC } Pro CCA } CCG }	CAT } His CAC } CAA Gln CAG }	CGT } CGC } Arg CGA } CGG }	T	C
	A	ATT } ATC } Ile ATA } ATG Met	ACT } ACC } Thr ACA } ACG }	AAT } Asn AAC } AAA Lys AAG }	AGT } Ser AGC } AGA Arg AGG }	T	C
	G	GTT } GTC } Val GTA } GTG }	GCT } GCC } Ala GCA } GCG }	GAT } Asp GAC } GAA Glu GAG }	GGT } GGC } Gly GGA } GGG }	T	C
						Third Letter	

(from <http://plato.stanford.edu/entries/information-biological/>)

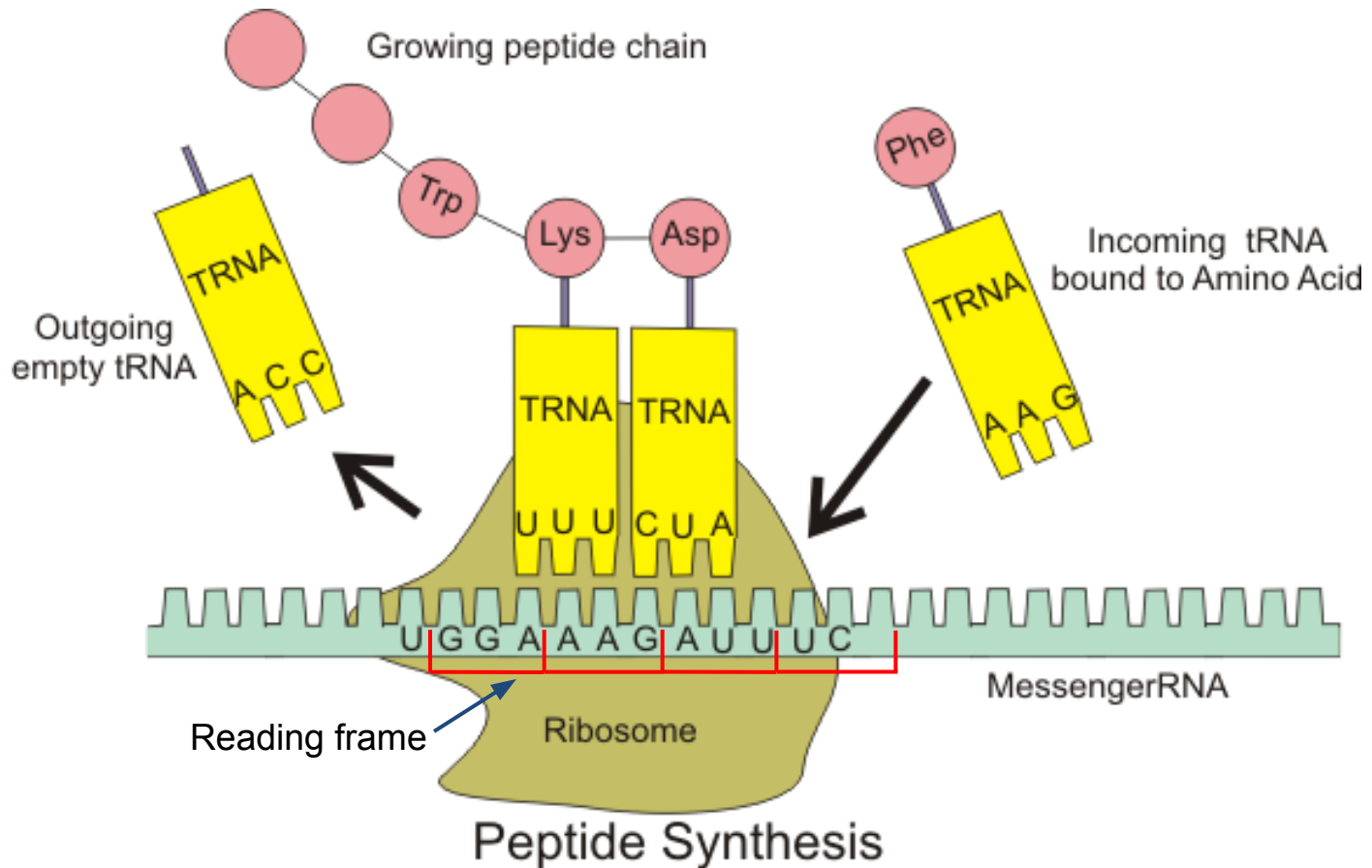
- Translation converts a sequence of non-overlapping codons to a sequence of amino acids.



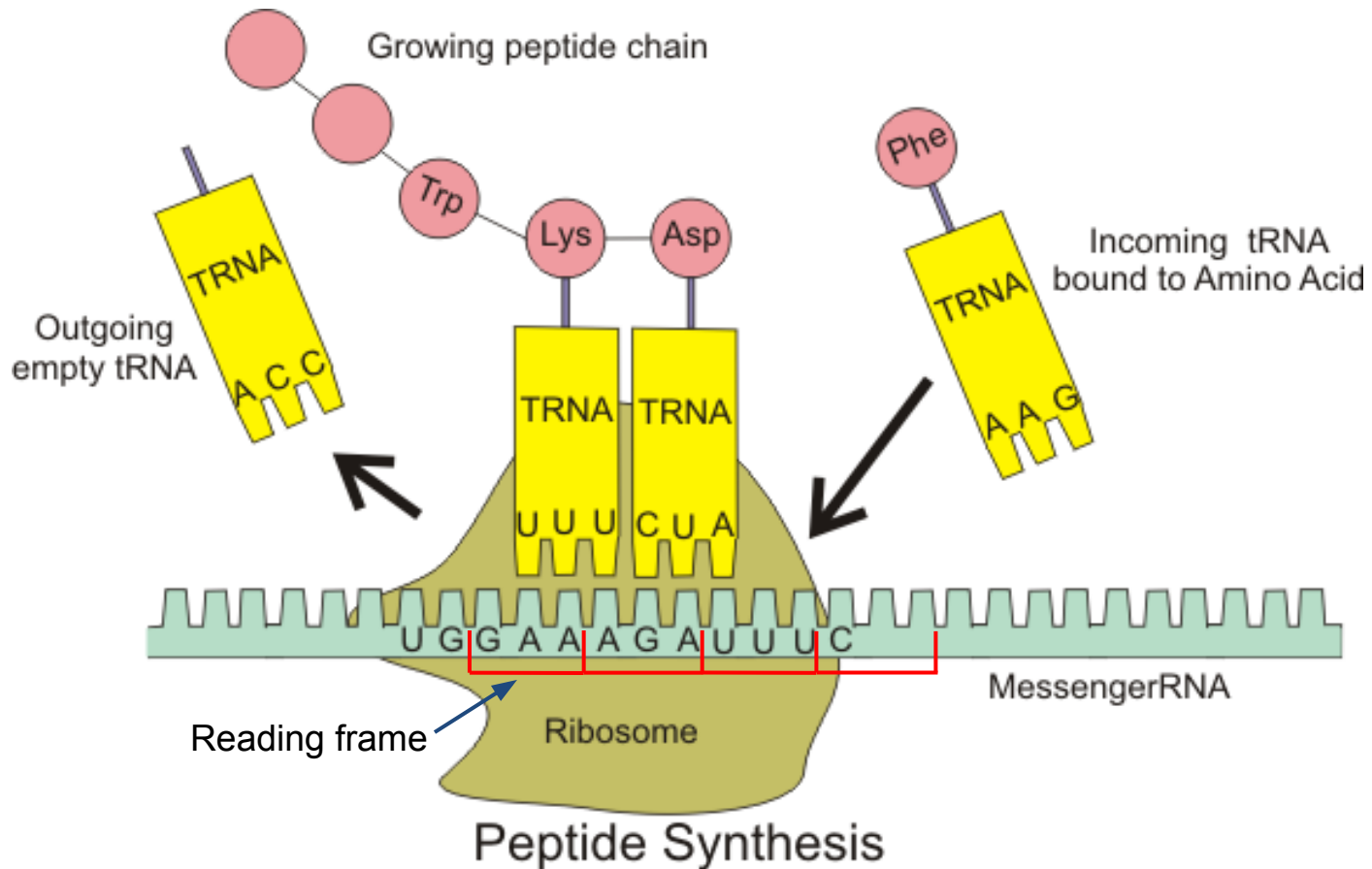
- A nucleotide sequence could potentially be decoded into 3 different amino sequences depending on how the nucleotides are grouped into codons.



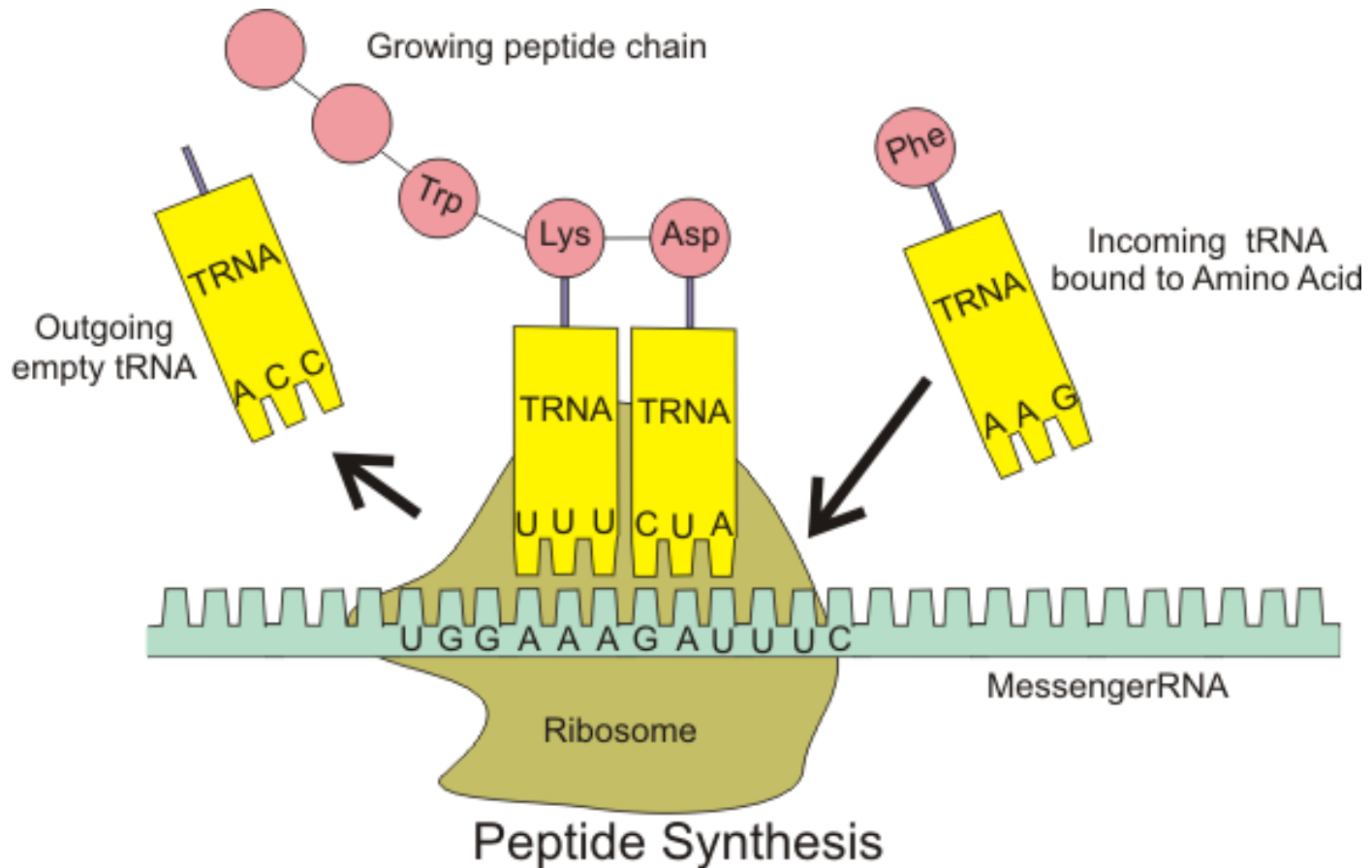
- A nucleotide sequence could potentially be decoded into 3 different amino sequences depending on how the nucleotides are grouped into codons.



- A nucleotide sequence could potentially be decoded into 3 different amino sequences depending on how the nucleotides are grouped into codons.



- 3 special codons called 'stop codons' signal the end of a coding sequence – TAA, TAG, and TGA.



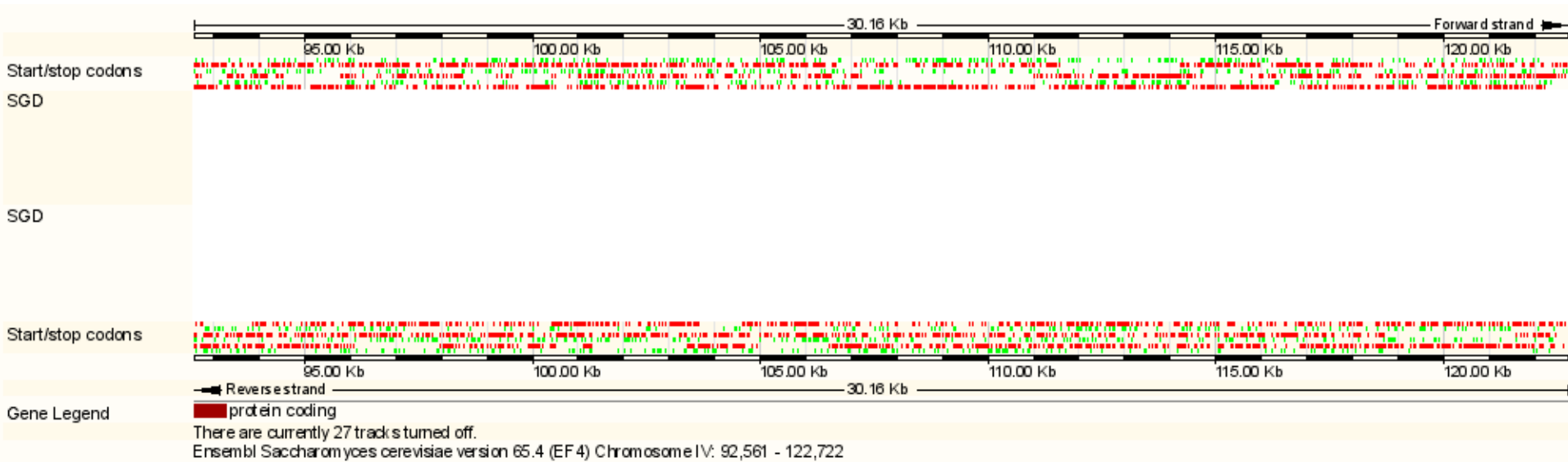
- An open reading frame (ORF) is a stretch of sequence that does not contain any in-frame stop codons.

ATCCTGTGTGATCTTATCGACGGTTGA

The diagram shows a DNA sequence: ATCCTGTGTGATCTTATCGACGGTTGA. Below the sequence, blue vertical lines mark the boundaries of codons. A red bracket above the sequence spans from the start of the third codon (GTG) to the end of the eighth codon (GTT), indicating an open reading frame. The sequence is divided into 8 codons: ATC, CTG, TGT, GAT, CTT, ATC, GAC, and GTT. The final 'GA' is a partial codon.

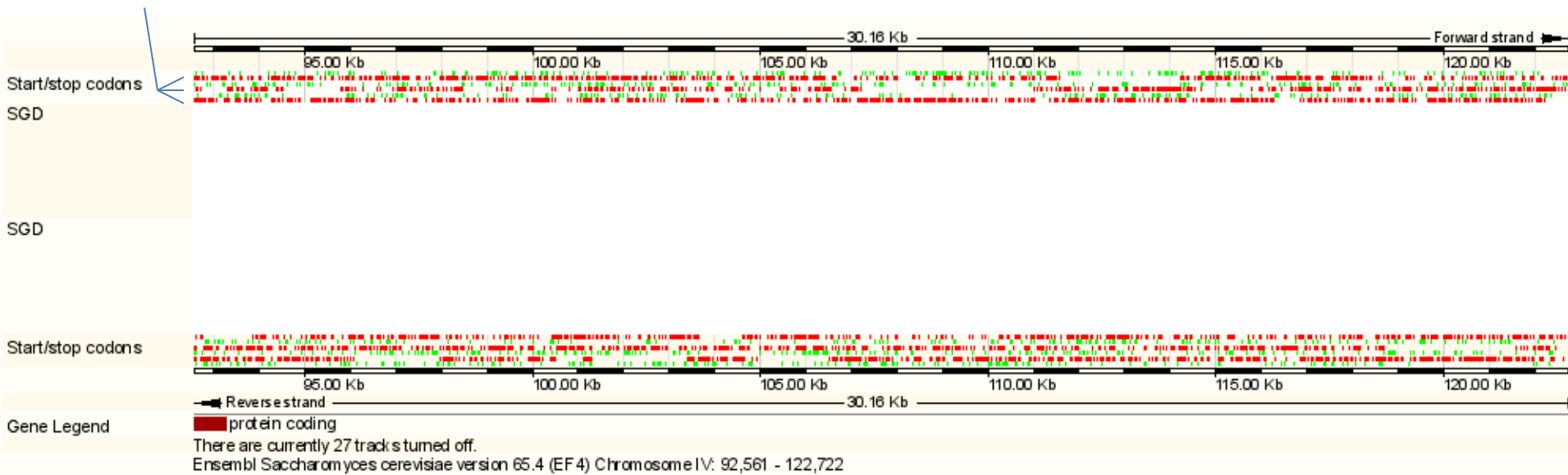
- Long ORFs are likely to be regions that encode a protein.
 - Sequence for a protein needs to avoid in-frame stop codons.
 - Non-translated regions have no such constraint and so will occasionally contain stop codons in all 3 frames by chance.

ensEMBL genome browser view of a 30 kb stretch of chr4



ensEMBL genome browser view of a 30 kb stretch of chr4

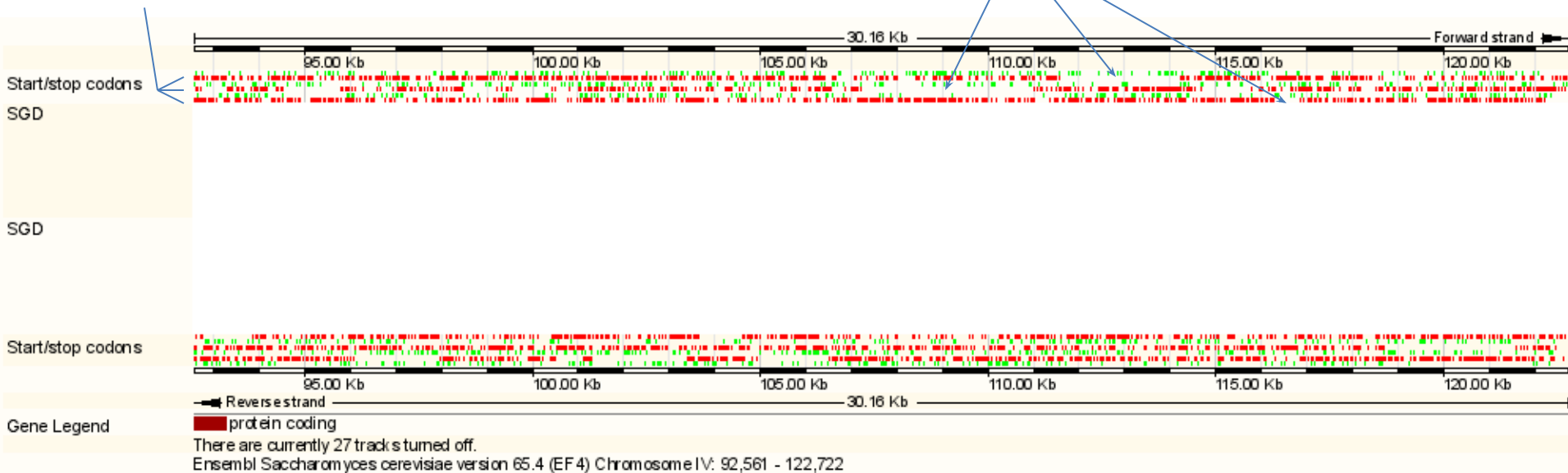
Occurrences of stop codons in each
of the 3 reading frames on the
forward strand



ensEMBL genome browser view of a 30 kb stretch of chr4

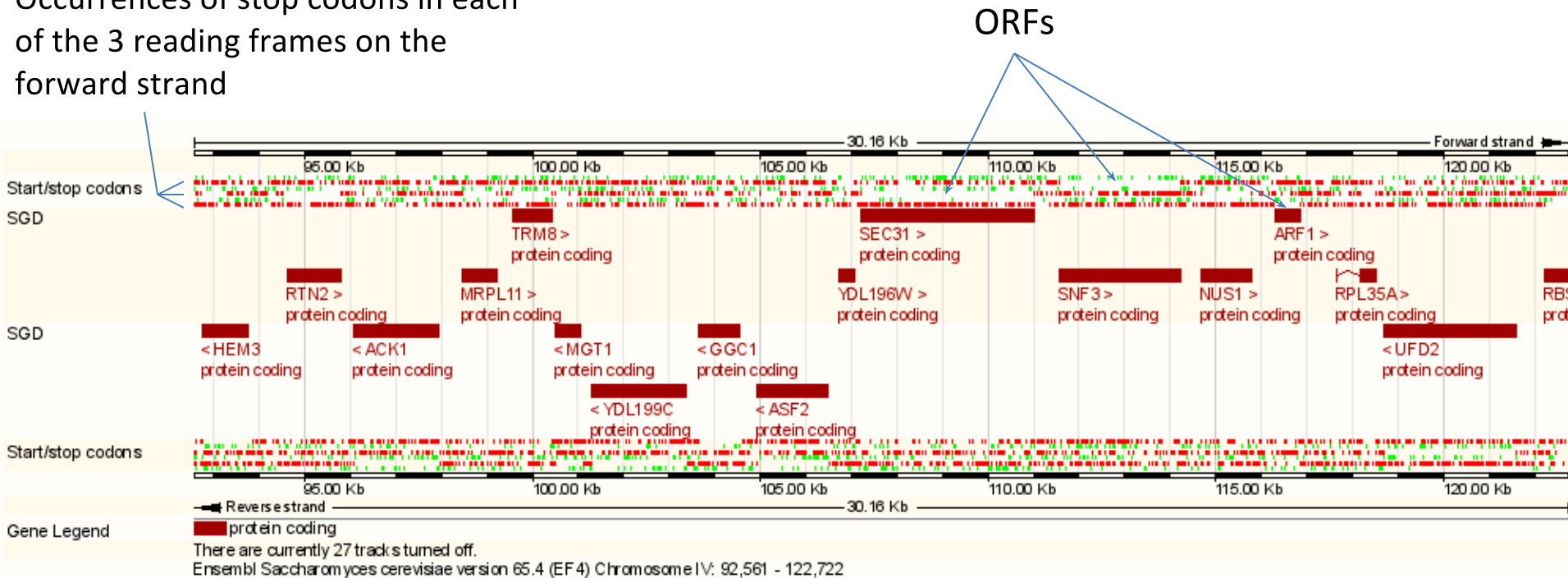
Occurrences of stop codons in each
of the 3 reading frames on the
forward strand

ORFs



ensEMBL genome browser view of a 30 kb stretch of chr4

Occurrences of stop codons in each of the 3 reading frames on the forward strand



Experimental evidence can then confirm that these ORFs really are translated into proteins.

- chrIV.txt, a text file of the sequence of chromosome 4, is available from today's segment page or on the class server.
- Your job is to find the regions of chromosome 4 that code for proteins by identifying ORFs that are longer than you would expect to see by chance.
- First you need to define exactly what this means. Invent a statistical model for ORF lengths.
 - In other words, decide how to answer the question “what is the probability that a random sequence of length $3n$ is an ORF?”

- Measure the length of the first ORF in the 0-offset reading frame and assign a p-value to this length according to your (null) model. Is this is a statistically significantly long ORF?

- Measure the lengths of all ORFs in the 0-offset reading frame. Make a histogram of the p-values you assign to all of the ORFs. Which ORFs are statistically significantly long?
- Time permitting:
 - Include all ORFs in the 1- and 2-offset reading frames.
 - Include all ORFs on the opposite strand.