Bacterial Gene Finding

CMSC 423

Finding Signals in DNA

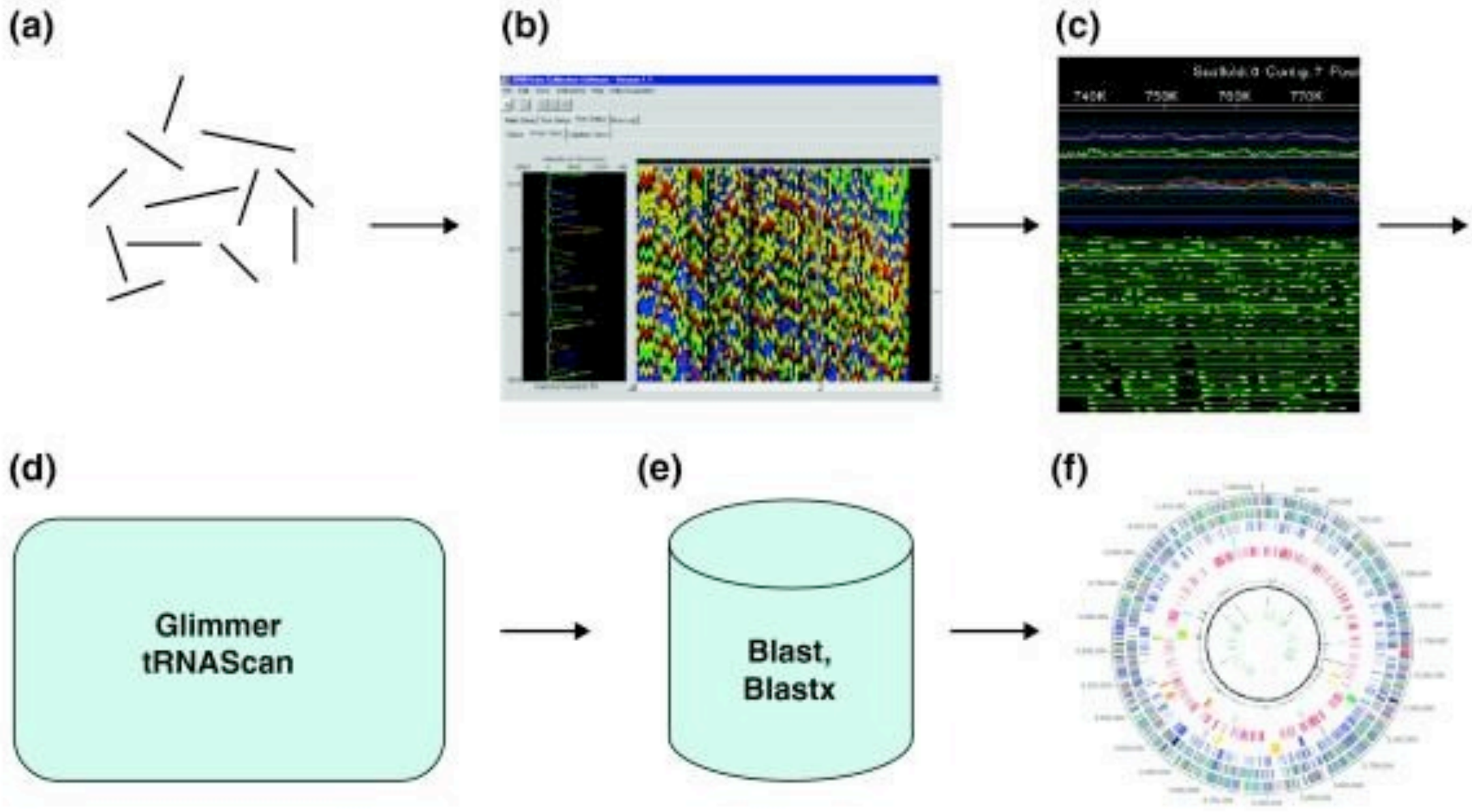
- We just have a long string of A, C, G, Ts. How can we find the "signals" encoded in it?

- Suppose you encountered a language you didn't know. How would you decipher it?

- **Idea #1:** Based on some external information, build a model (like an HMM) for how particular features are encoded.

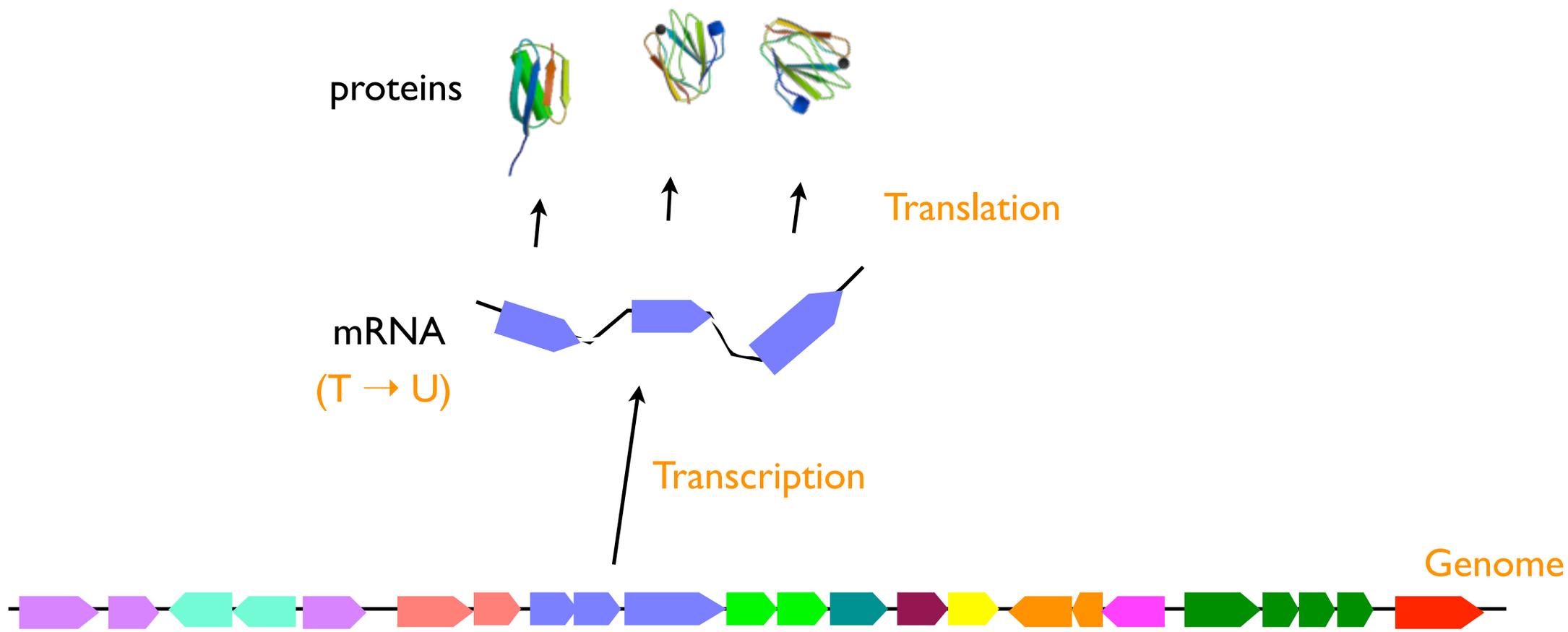
- **Idea #2:** Find patterns that appear more often than you expect by chance. ("the" occurs a lot in English, so it may be a word.)

- Gibbs sampling was an example of how to implement Idea #2.
  We will soon see how to implement idea #1.



Salzberg *Genome Biology* 2007 **8**:102

“Central Dogma” of Biology



DNA =

- double-stranded, linear molecule
- each strand is string over {A,C,G,T}
- strands are complements of each other (A ↔ T; C ↔ G)
- substrings encode for genes most of which encode for proteins

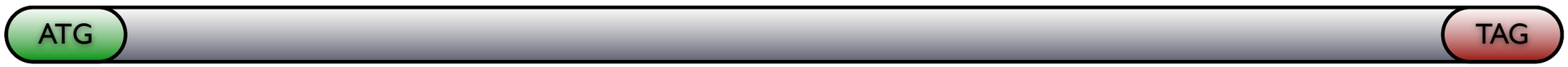
The Genetic Code

1st base	2nd base: U		2nd base: C		2nd base: A		2nd base: G	
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine
U	UUC	(Phe/F) Phenylalanine	UCC	(Ser/S) Serine	UAC	(Tyr/Y) Tyrosine	UGC	(Cys/C) Cysteine
U	UUA	(Leu/L) Leucine	UCA	(Ser/S) Serine	UAA	Ochre *Stop*	UGA	Opal *Stop*
U	UUG	(Leu/L) Leucine	UCG	(Ser/S) Serine	UAG	Amber *Stop*	UGG	(Trp/W) Tryptophan
C	CUU	(Leu/L) Leucine	CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine
C	CUC	(Leu/L) Leucine	CCC	(Pro/P) Proline	CAC	(His/H) Histidine	CGC	(Arg/R) Arginine
C	CUA	(Leu/L) Leucine	CCA	(Pro/P) Proline	CAA	(Gln/Q) Glutamine	CGA	(Arg/R) Arginine
C	CUG	(Leu/L) Leucine	CCG	(Pro/P) Proline	CAG	(Gln/Q) Glutamine	CGG	(Arg/R) Arginine
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine
A	AUC	(Ile/I) Isoleucine	ACC	(Thr/T) Threonine	AAC	(Asn/N) Asparagine	AGC	(Ser/S) Serine
A	AUA	(Ile/I) Isoleucine	ACA	(Thr/T) Threonine	AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
A	AUG [A]	(Met/M) Methionine	ACG	(Thr/T) Threonine	AAG	(Lys/K) Lysine	AGG	(Arg/R) Arginine
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine
G	GUC	(Val/V) Valine	GCC	(Ala/A) Alanine	GAC	(Asp/D) Aspartic acid	GGC	(Gly/G) Glycine
G	GUA	(Val/V) Valine	GCA	(Ala/A) Alanine	GAA	(Glu/E) Glutamic acid	GGA	(Gly/G) Glycine
G	GUG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

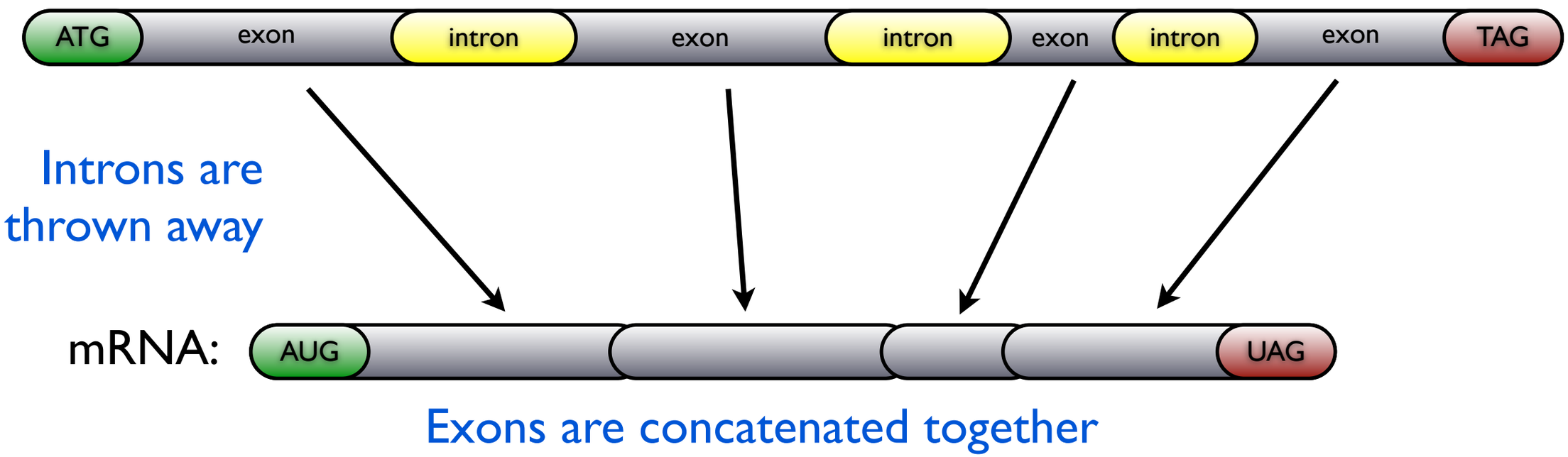
- There are 20 different amino acids & 64 different codons.
- Lots of different ways to encode for each amino acid.
- The 3rd base is typically less important for determining the amino acid
- Three different "stop" codons that signal the end of the gene
- Start codons differ depending on the organisms, but AUG is often used.

Eukaryotic Genes & Exon Splicing

Prokaryotic (bacterial) genes look like this:



Eukaryotic genes usually look like this:



This spliced RNA is what is translated into a protein.

The Prokaryotic Gene Finding Problem

- Genes are subsequences of DNA that (generally) tell the cell how to make specific proteins.

- How can we find which subsequences of DNA are genes?

```
Start Codon: ATG
Stop Codons: TGA, TAG, TAA
```

ATAGAGGGT**ATG**GGGGACCCGGACACG**ATG**GCAGA**TGA**CGATGACGATGACGATGACGGG**TGA**AGTGAGTCAACACATGAC

Challenges:

- The start codon can occur in the middle of a gene (where it encodes for the amino acid methionine)
- The stop codon can occur in nonsense DNA between genes.
- The stop codon can occur "out of frame" inside a gene.
- Don't know what "phase" the gene starts in.

A Simple Gene Finder

1. Find all stop codons in genome

2. For each stop codon, find the in-frame start codon farthest upstream of the stop codon, without crossing another in-frame stop codon.

GGC **TAG** **ATG** AGG GCT CTA ACT **ATG** GGC GCG **TAA**

Each substring between the start and stop codons is called an ORF
"open reading frame"

3. Return the "long" ORF as predicted genes.

3 out of the 64 possible codons are stop codons ⇒ in random DNA, every 22nd codon is expected to be a stop.

Gene Finding as a Machine Learning Problem

- Given training examples of some known genes, can we distinguish ORFs that are genes from those that are not?

- **Idea**: can use distribution of codons to find genes.
  - every codon should be about equally likely in non-gene DNA.
  - every organism has a slightly different bias about how often certain codons are preferred.
  - could also use frequencies of longer strings (k-mers).

Bacillus anthracis (anthrax) codon usage

UUU	F	0.76	UCU	S	0.27	UAU	Y	0.77	UGU	C	0.73
UUC	F	0.24	UCC	S	0.08	UAC	Y	0.23	UGC	C	0.27
UUA	L	0.49	UCA	S	0.23	UAA	*	0.66	UGA	*	0.14
UUG	L	0.13	UCG	S	0.06	UAG	*	0.20	UGG	W	1.00
CUU	L	0.16	CCU	P	0.28	CAU	H	0.79	CGU	R	0.26
CUC	L	0.04	CCC	P	0.07	CAC	H	0.21	CGC	R	0.06
CUA	L	0.14	CCA	P	0.49	CAA	Q	0.78	CGA	R	0.16
CUG	L	0.05	CCG	P	0.16	CAG	Q	0.22	CGG	R	0.05
AUU	I	0.57	ACU	T	0.36	AAU	N	0.76	AGU	S	0.28
AUC	I	0.15	ACC	T	0.08	AAC	N	0.24	AGC	S	0.08
AUA	I	0.28	ACA	T	0.42	AAA	K	0.74	AGA	R	0.36
AUG	M	1.00	ACG	T	0.15	AAG	K	0.26	AGG	R	0.11
GUU	V	0.32	GCU	A	0.34	GAU	D	0.81	GGU	G	0.30
GUC	V	0.07	GCC	A	0.07	GAC	D	0.19	GGC	G	0.09
GUA	V	0.43	GCA	A	0.44	GAA	E	0.75	GGA	G	0.41
GUG	V	0.18	GCG	A	0.15	GAG	E	0.25	GGG	G	0.20

An Improved Simple Gene Finder

- Score each ORF using the product of the probability of each codon:

$$GFScore(g) = Pr(codon_1) x Pr(codon_2) x Pr(codon_3) x ... x Pr(codon_n)$$

But: as genes get longer, GFScore(g) will decrease.

So: we should calculate GFScore(g[i...i+k]) for some window size k.

The final GFSCORE(g) is the average of the Scores of the windows in it.

Recap

- Simple gene finding approaches use codon bias and long ORFs to identify genes.

- Many top gene finding programs for Eukaryotes are based on generalizations of Hidden Markov Models because multiple types of signals (many "authors") are present in a gene (intron, exon, etc.)

- Basic HMMs must be generalized to emit variable sized strings.