Genes & Gene Finding

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Department of Computer Science

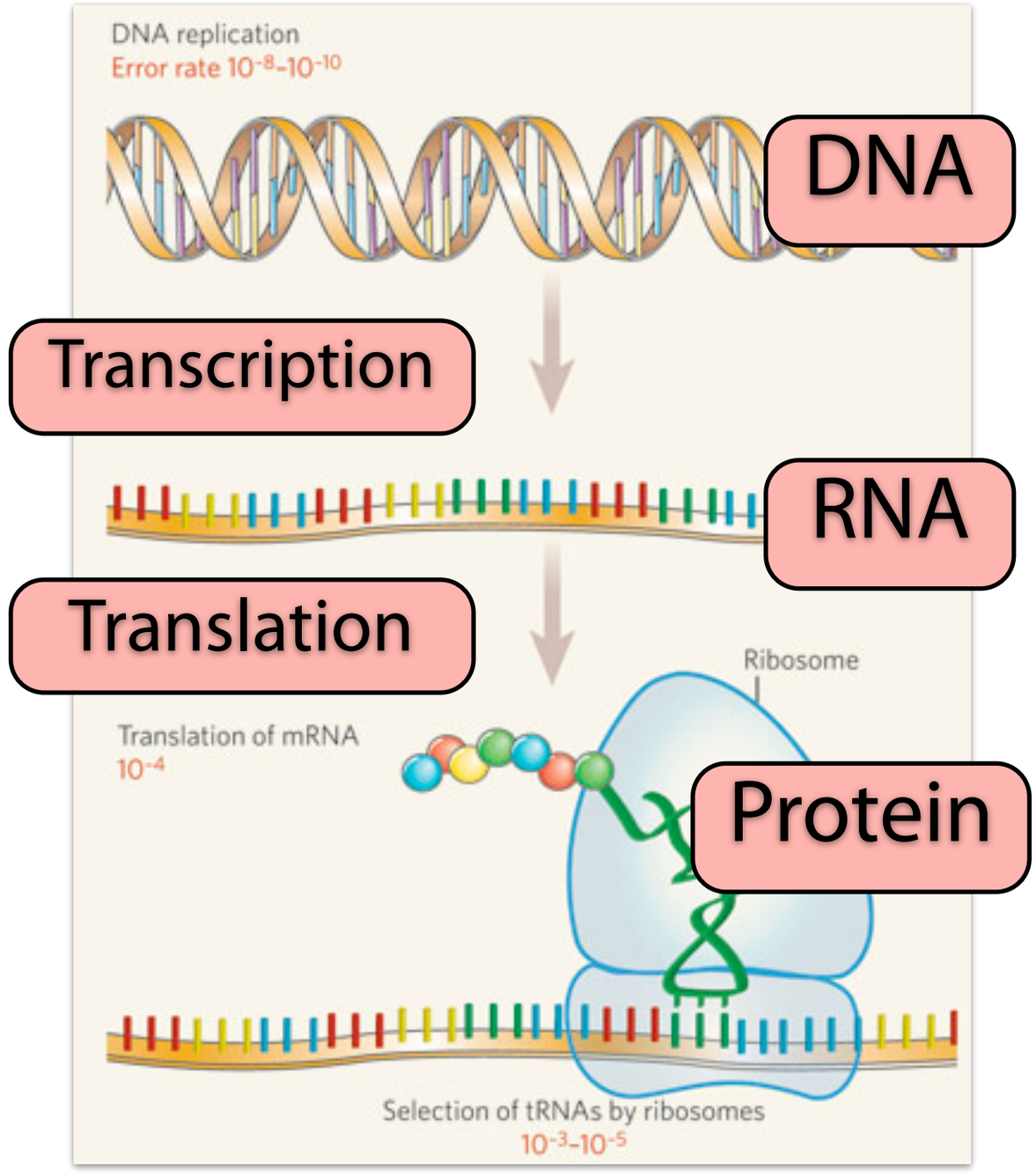
Please sign guestbook (www.langmead-lab.org/teaching-materials) to tell me briefly how you are using the slides. For original Keynote files, email me (ben.langmead@gmail.com).

Gene finding

Recall the "Central Dogma" and the centrality of genes

DNA molecules contain information about how to create proteins; this is *transcribed* into RNA molecules, which, in turn, direct chemical machinery to *translate* the message into a protein.

Hunter, Lawrence. "Life and its molecules: A brief introduction." *AI Magazine* 25.1 (2004): 9.



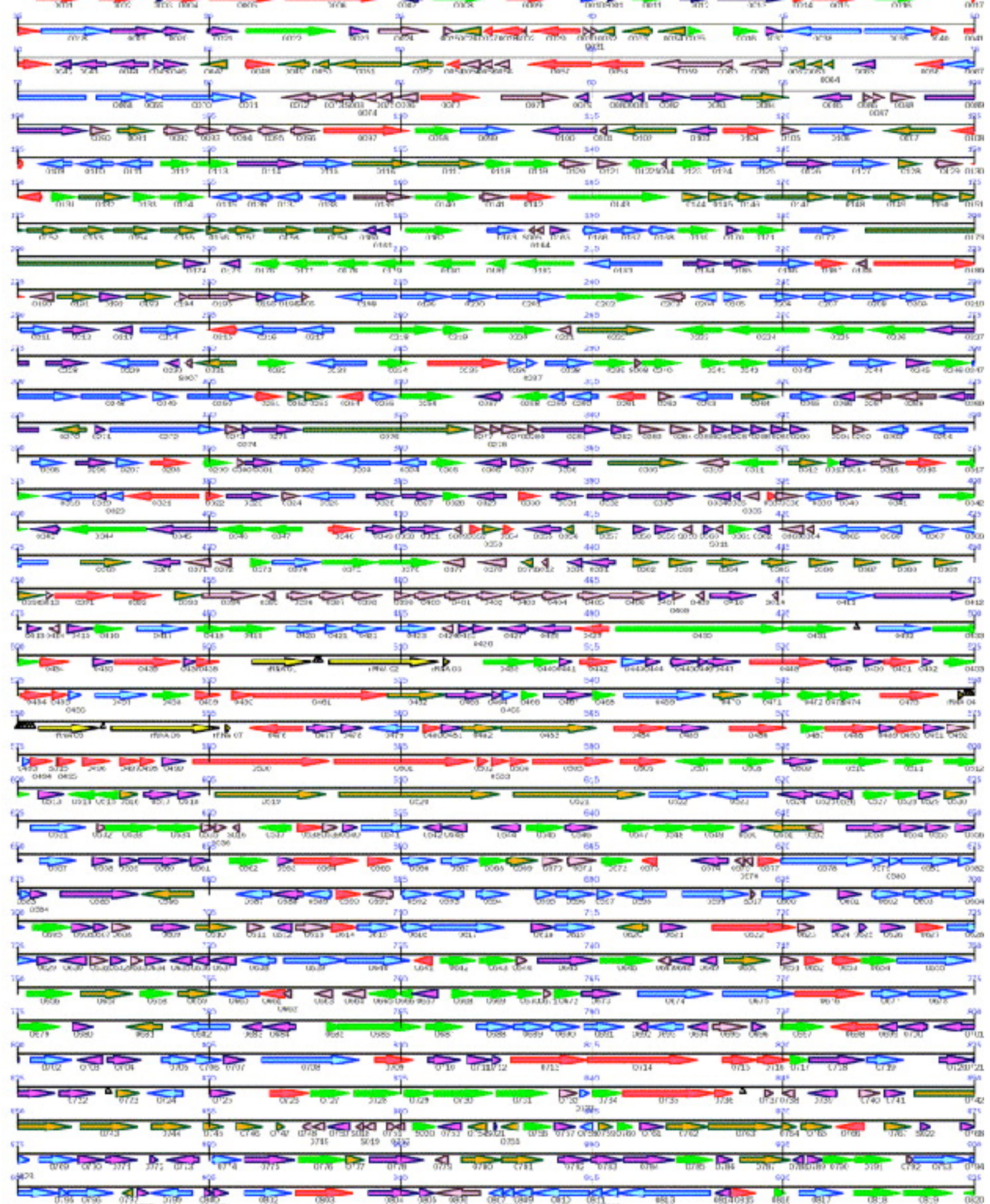
Picture from: Roy H, Ibba M. Molecular biology: sticky end in protein synthesis. Nature. 2006 Sep 7;443(7107):41-2.

Bacterial genes

Genes (colored arrows) packed tightly into the *Staphylococcus aureus* genome

In bacteria, one gene corresponds to one continuous interval on the genome

We have good methods for predicting where these genes are



Bacterial gene finding

Table 3. Glimmer3 prediction accuracy with `long-orfs` training

Genome			Glimmer3 Predictions					versus Glimmer2.13		
Organism	GC%	# Genes	3′ Matches		5′ & 3′ Matches		Extra	3′ Match	5′ & 3′	Extra
A.fulgidus	49	1165	1161	99.7%	873	74.9%	1332	−2	−34	−64
B.anthracis	35	3132	3125	99.8%	2751	87.8%	2419	−1	+752	−144
B.subtilis	44	1576	1562	99.1%	1391	88.3%	3020	+3	+421	−724
C.tepidum	57	1292	1289	99.8%	934	72.3%	835	+3	+26	−400
C.perfringens	29	1504	1501	99.8%	1383	92.0%	1192	−1	+267	−20
E.coli	51	3603	3534	98.1%	3112	86.4%	1002	+11	+784	−843
G.sulfurreducens	61	2351	2337	99.4%	1933	82.2%	1165	+7	+575	−734
H.pylori	39	915	910	99.5%	795	86.9%	788	+2	+57	−103
P.fluorescens	63	4535	4510	99.4%	3598	79.3%	1953	+35	+895	−2359
R.solanacearum	67	2512	2485	98.9%	2028	80.7%	1183	+341	+1044	−2184
S.epidermidis	32	1650	1646	99.8%	1514	91.8%	791	+8	+358	−32
T.pallidum	53	575	567	98.6%	391	68.0%	567	−2	+50	−281
U.parvum	26	327	324	99.1%	295	90.2%	297	−1	+21	−11
Averages:				99.3%		83.1%		+31	+401	−608

Genomes and columns are as in the preceding table. Glimmer3 was run by using the output of its `long-orfs` program to train an IMM. The output of an initial run of Glimmer3 was used to set start codon frequencies and to find a ribosome-binding-site motif. A second run of Glimmer3 using those values generated the above predictions. Glimmer2 was trained on the output of its version of the `long-orfs` program.

Approaches can identify exact bacterial genes with as high as 92% accuracy; can identity gene ends with ≥98% accuracy

Delcher, Arthur L., et al. "Identifying bacterial genes and endosymbiont DNA with Glimmer." *Bioinformatics* 23.6 (2007): 673-679.

Eukaryotic genes

Eukaryotic genes are more complex than prokaryotic (bacterial) genes for several reasons, as we'll see

Likewise, *finding* eukaryotic genes computationally is harder

Gene finding

During the Human Genome Project, there was public debate about how many genes were in the human genome

A range of predictions were made: ~40K to ~100K

Answer turned out to be ~20K, and the number of protein-coding genes has slowly but steadily decreased since then

	chromosomes --diploid	base pairs	genome size (#genes)	Reference
fruit fly	8	$1.65x10^8$	13,600	ref
Budding yeast	16	12,462,637	6,275	ref
human	46	3.3×10^9	~21,000	ref
human mitochondria		16,569	13	ref
rice	24	4.66×10^8	46,022 -55,615	ref
dog	78	2.4×10^9	~25,000	ref
mouse	40	3.4×10^9	~23,000	ref

https://www.edinformatics.com/math_science/human_genome.htm

But how did they find the genes given the genome sequence?

A human gene

chr11:5246500-5248500 (reverse strand):

ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACTCCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTC
ATCACTTAGACCTCACCCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGG
GCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAA
CAGACACC**ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTT**
GGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGA
CAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAG**GCTGCTGGTG**
GTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGT
GAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA
CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGGTGAGTCTATGGGACGCTTGATGTTTT
CTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAG
ACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTT
GTTTAATTCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAA
AGGAAATATCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAATAT
ATGTGTGCTTATTTGCATA[illegible]TACATAATCATTATACATAT
TTATGGGTTAAAGTGTAA[illegible]TTGCATTTGTAATTTTAAAA
AATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAA
TGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCT
CTGCATATAAATATTTCTGCATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTA
CCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCA
TACCTCTTATCTTCCTCCCACAG**CTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACC**
CCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAAGCTCGCTT
TCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTT
GAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAATGATGTATTTAAATTATTTCTGAATATTTTACTA
AAAAGGGAATGTGGGAGGTCAGTGCATTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATACACTATA
TCTTAAACTCCATGAAAGAAGGTGAGGCTGCAAACAGCTAATGCACATTGGCAACAGCCCCTGATGCATATGCCTTATTC

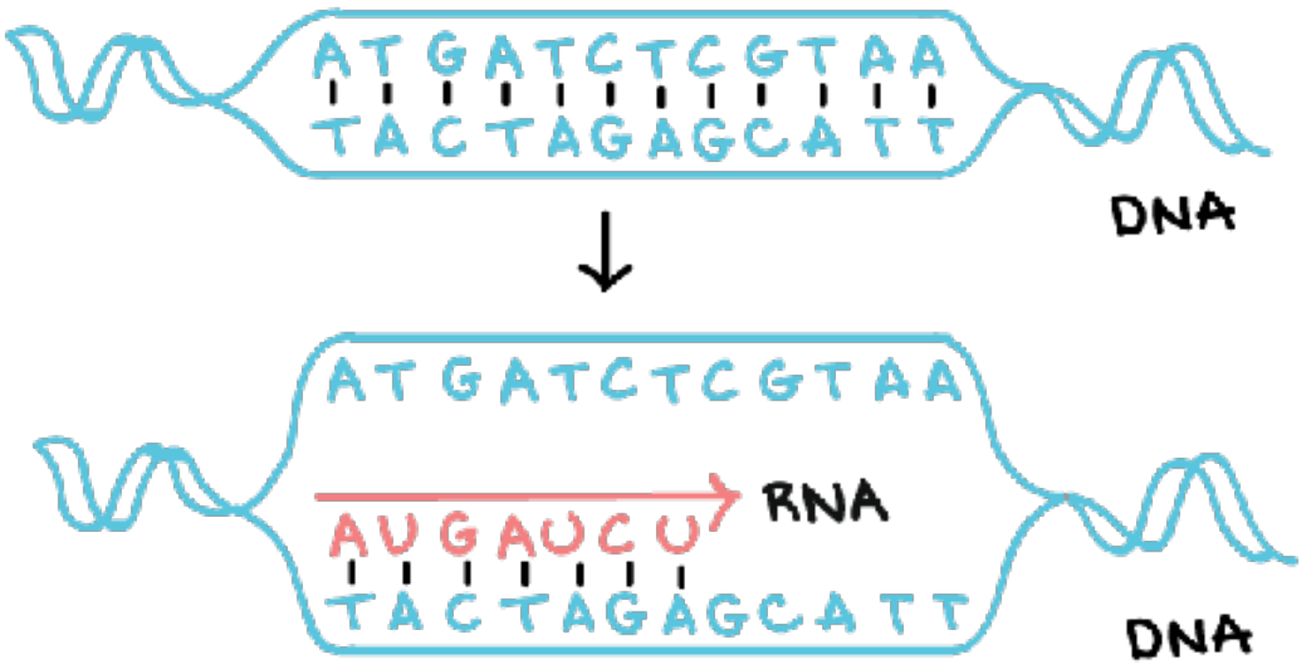
Homo sapiens hemoglobin, beta (HBB)

Genes

Sequence models will allow us to "see" gene sequences in the ocean of the genome

Recall a few things we've learned about genes and transcription

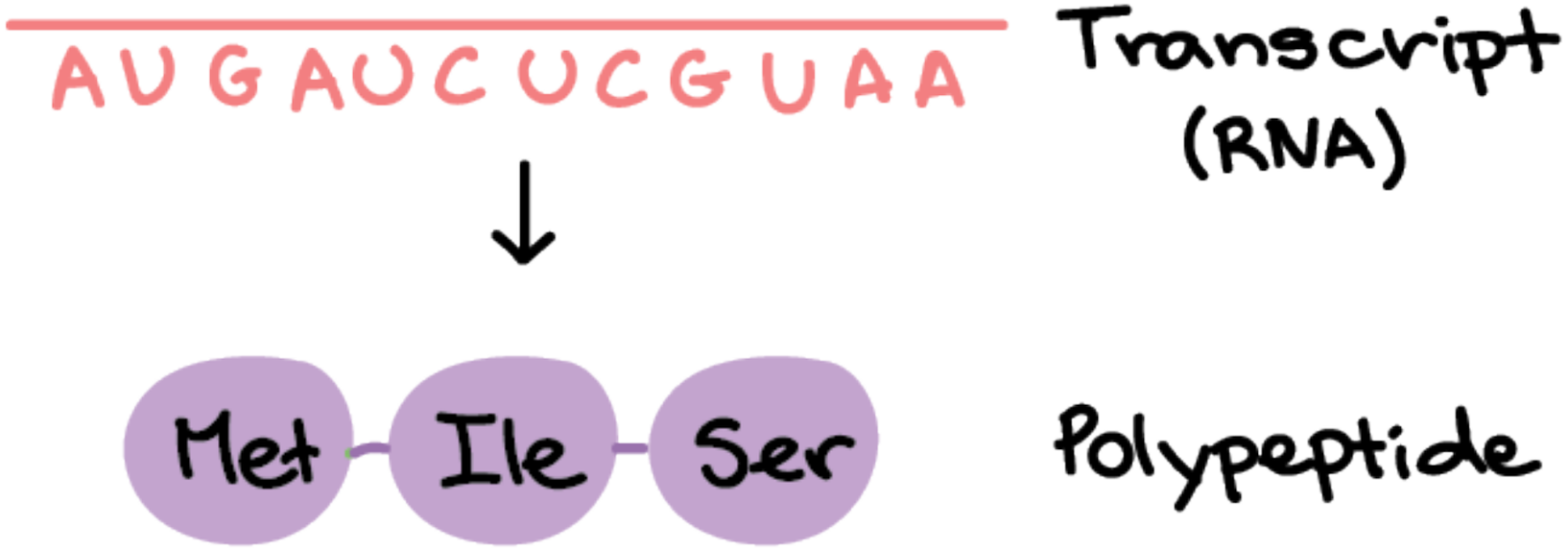
Transcription produces an RNA copy of a stretch of DNA but with Ts (thymine) replaced by Us (uracil)



https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription

Genes

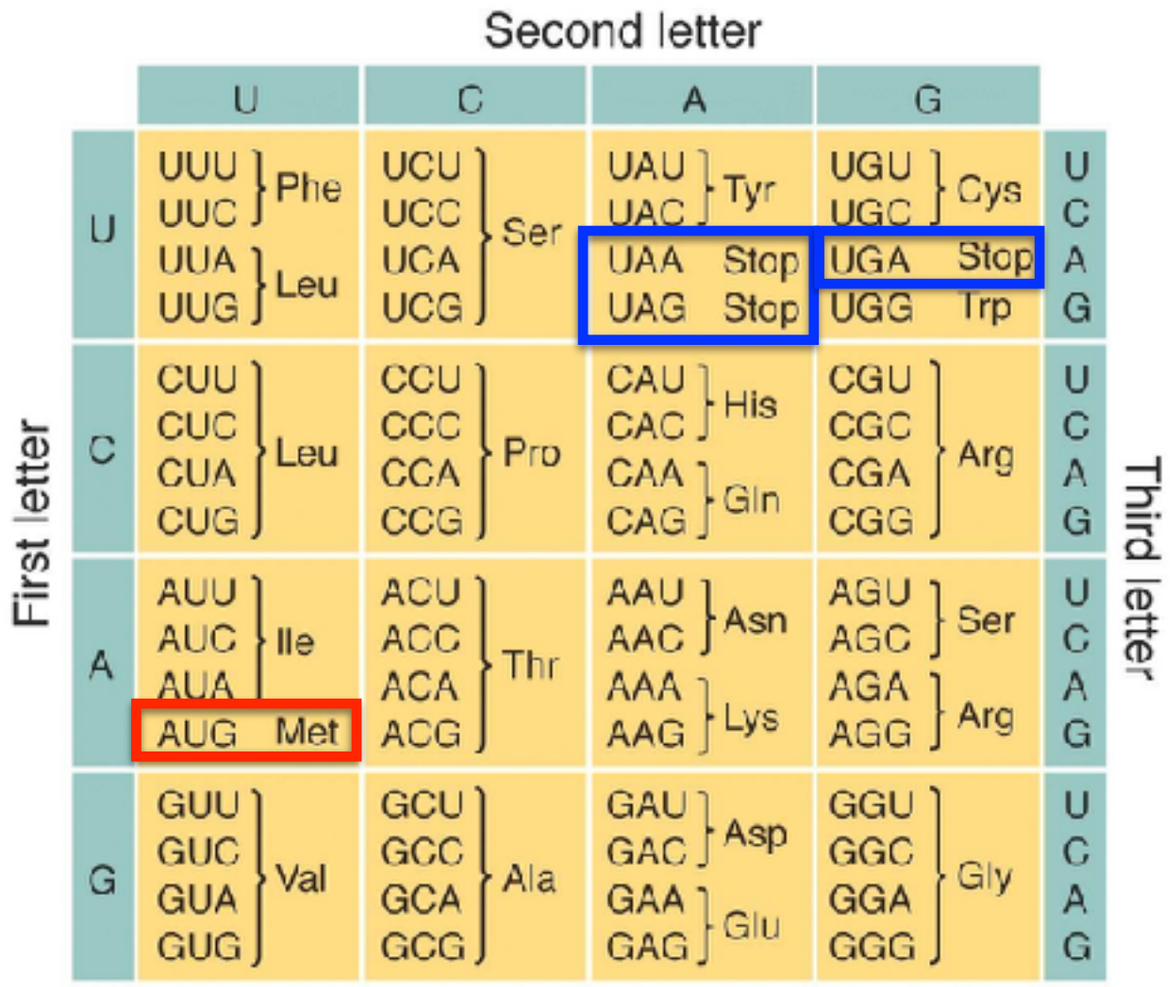
Triples of nucleotides ("codons") are translated into amino acids via the genetic code



https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription

Genes

Some codons are special ("stop codons"), signaling that translation of the protein should stop



"Start codon" (AUG) signals translation should begin; also codes for amino acid Methionine

Genes

Splicing is a process by which some portions of the mRNA are cut out prior to translation

ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACTCCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTCATCACTTAGACCTCACCC
TGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGC
TTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT
GGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCA
TGTGGAGACAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTGGTCTACCCTT
GGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAA
CTTCAGGGTGAGTCTATGGGACGCTTGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGT
TTAGAATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTTGT
TTAATTCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATATCTCTGAGATAC
ATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAATATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTT
ATTTTCTTTTATTTTTAATTGATACATAATCATTATACATATTTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGT
AATTTTGCATTTGTAATTTTAAAAAATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGG
CAATAATGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCTCTGCATATAAAT
ATTTCTGCATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTACCATTCTGCTTTTATTTTATGGTTGGGATA
AGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTCTTATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTG
TGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTAT
CACTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGC
ATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAATGATGTATTTAAATTATTTCTGAATATTTTACTAAAAAGGGAATGTGGGAGGTCAG
TGCATTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATACACTATATCTTAAACTCCATGAAAGAAGGTGAGGCTGCAAACAGCTA
ATGCACATTGGCAACAGCCCCTGATGCATATGCCTTATTC

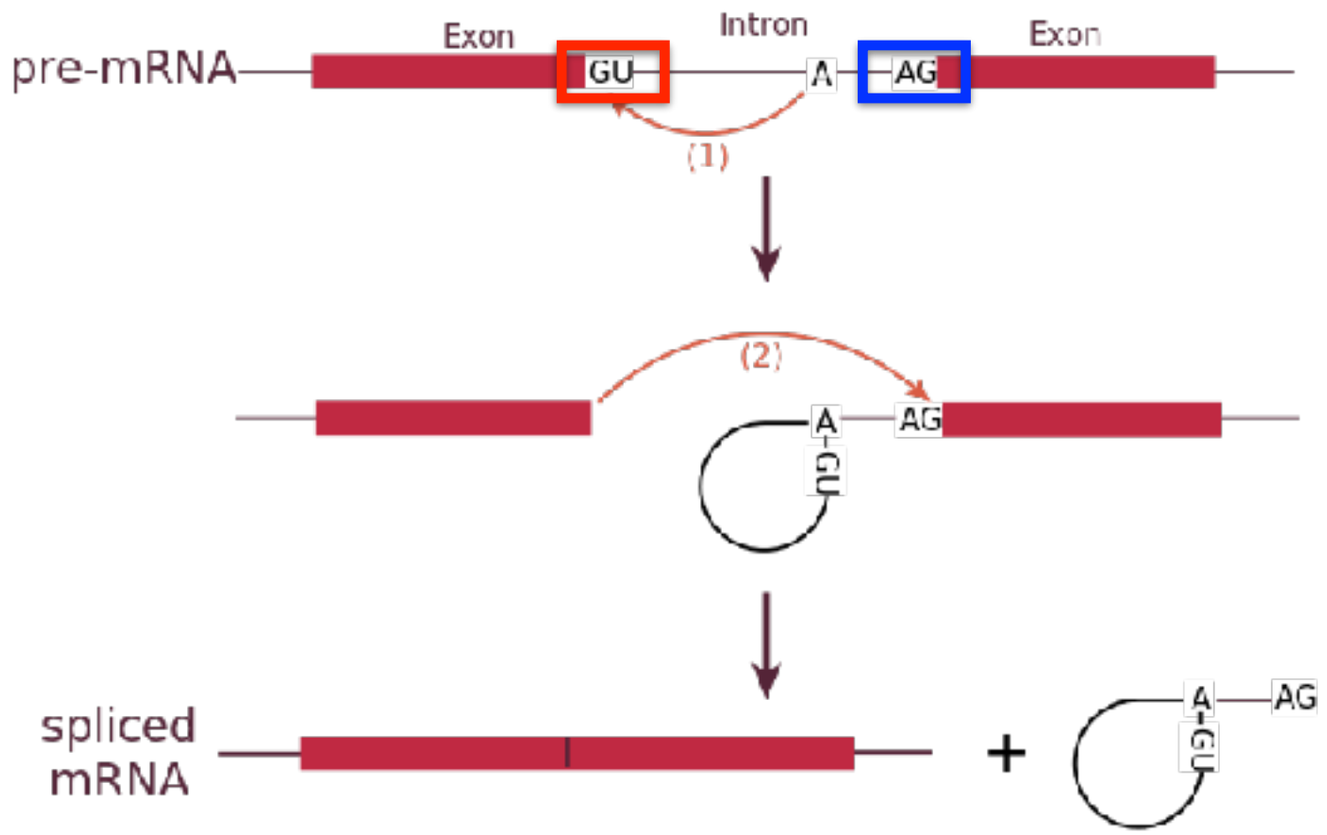
Genes

Splicing is a process by which some portions of the mRNA are cut out prior to translation

ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACTCCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTCATCACTTAGACCTCACCC
TGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGC
TTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT
GGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCA
TGTGGAGACAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTGGTCTACCCTT
GGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAA
CTTCAGGGTGAGTCTATGGGACGCTTGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGT
TTAGAATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTTGT
TTAATTCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATATCTCTGAGATAC
ATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAATATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTT
ATTTTCTTTTATTTTTAATTGATACATAATCATTATACATATTTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGT
AATTTTGCATTTGTAATTTTAAAAAATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGG
CAATAATGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCTCTGCATATAAAT
ATTTCTGCATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTACCATTCTGCTTTTATTTTATGGTTGGGATA
AGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTCTTATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTG
TGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTAT
CACTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGC
ATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAATGATGTATTTAAATTATTTCTGAATATTTTACTAAAAAGGGAATGTGGGAGGTCAG
TGCATTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATACACTATATCTTAAACTCCATGAAAGAAGGTGAGGCTGCAAACAGCTA
ATGCACATTGGCAACAGCCCCTGATGCATATGCCTTATTC

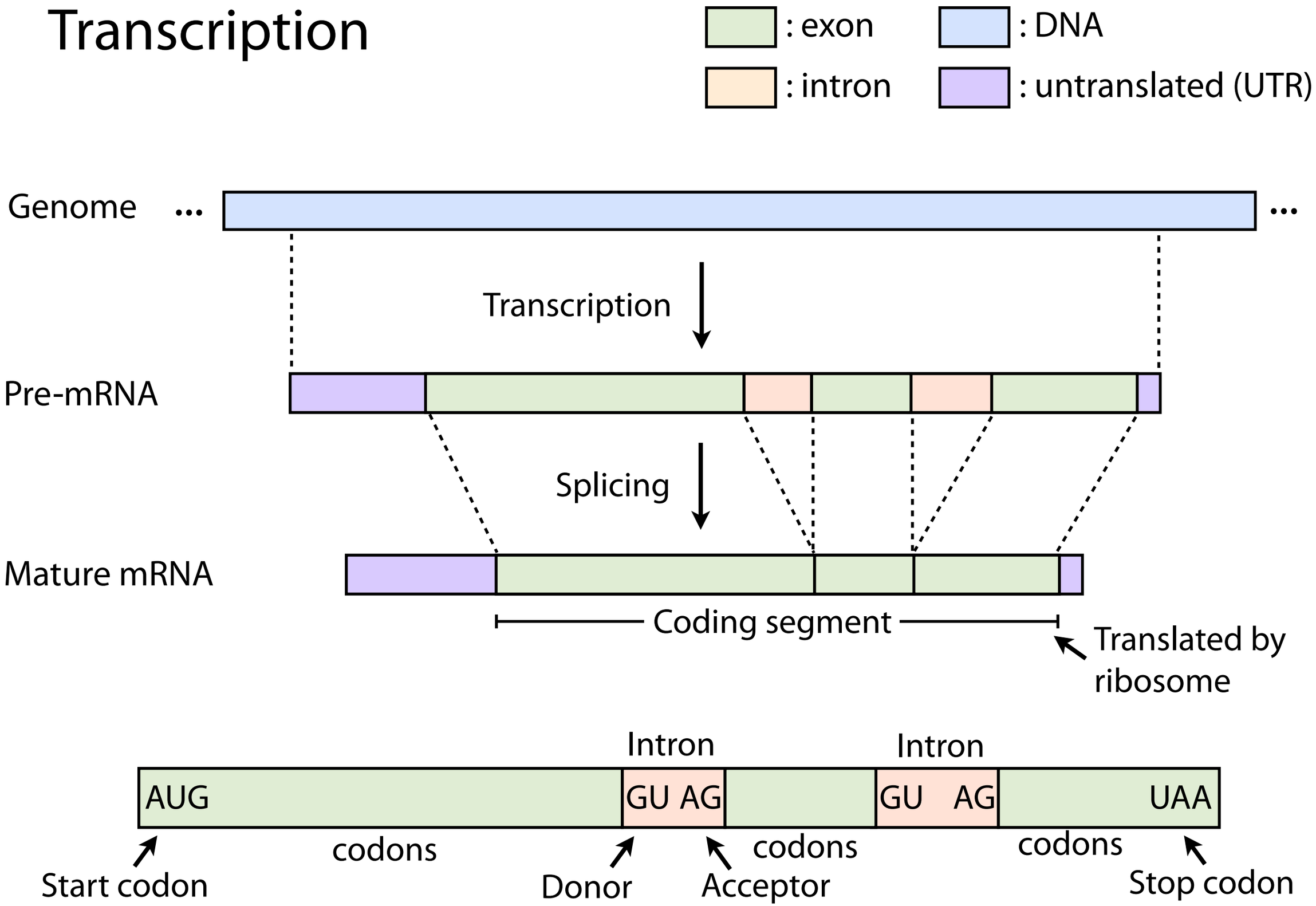
Genes

Splicing happens at certain nucleotide patterns: GU and AG



Animation of splicing: https://evolutionnews.org/2013/09/the_spliceosome_1/

Transcription



A human gene

chr11:5246500-5248500 (reverse strand):

ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACTCCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTC
ATCACTTAGAC[CACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGG
GCTGGGCATAA[TCTATTGCTTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAA
CAGACAC**ATG**GTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT[GTGGATGAAGTT
GGTGGTGAGGCCCTGGGCAG**GT**TGGTATCAAGGTTACAAGACAGGTTTAAGGAG[GGCATGTGGAGA
CAGAGAAGACTCTTGGGTTTCTC[CTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAG**GCTGCTGGTG
GTCTACCCTTGGACCCAGAGGT[TTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGT
GAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA
CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGG**GT**GAGTCTATGGGACGCTTGATGTTTT
CTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAG
ACGAATGATTGCATCAGTGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTT
GTTTAATTCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAA
AGGAAATATCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAATAT
ATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTTAATTGATACATAATCATTATACATAT
TTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAATTTTGCATTTGTAATTTTAAAA
AATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATAA
TGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCT
CTGCATATAAATATTTCTGCATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTA
CCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTGGATTATTCTGAGTCCA[AATCATGTTCA
TACCTCTTATCTTCCTCCCAC**AG**CTCCTGGGCAACGTGCTGGTCTGTGTGCT[AAGAATTCACC
CCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCAC**TAA**GCTCGCTT
TCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTT
GAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAATGATGTATTTAAATTATTTCTGAATATTTTACTA
AAAAGGGAATGTGGGAGGTCAGTGCATTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATACACTATA
TCTTAAACTCCATGAAAGAAGGTGAGGCTGCAAACAGCTAATGCACATTGGCAACAGCCCCTGATGCATATGCCTTATTC

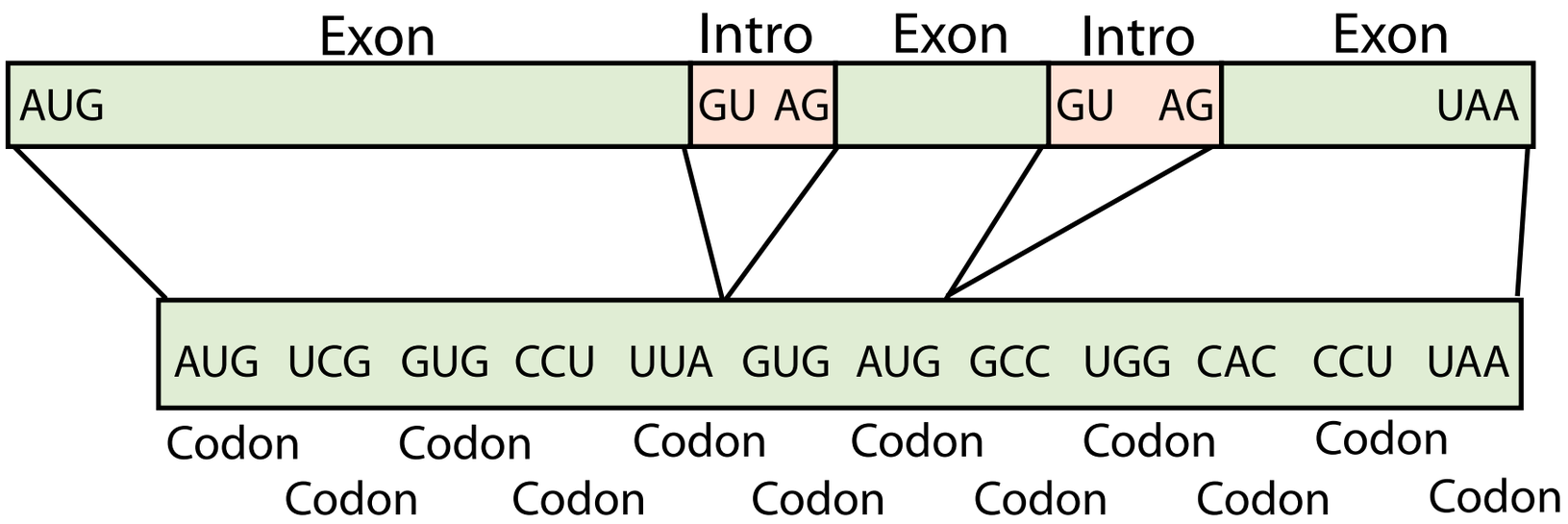
Start codon

Donor

Acceptor

Stop codon

Codons and the genetic code

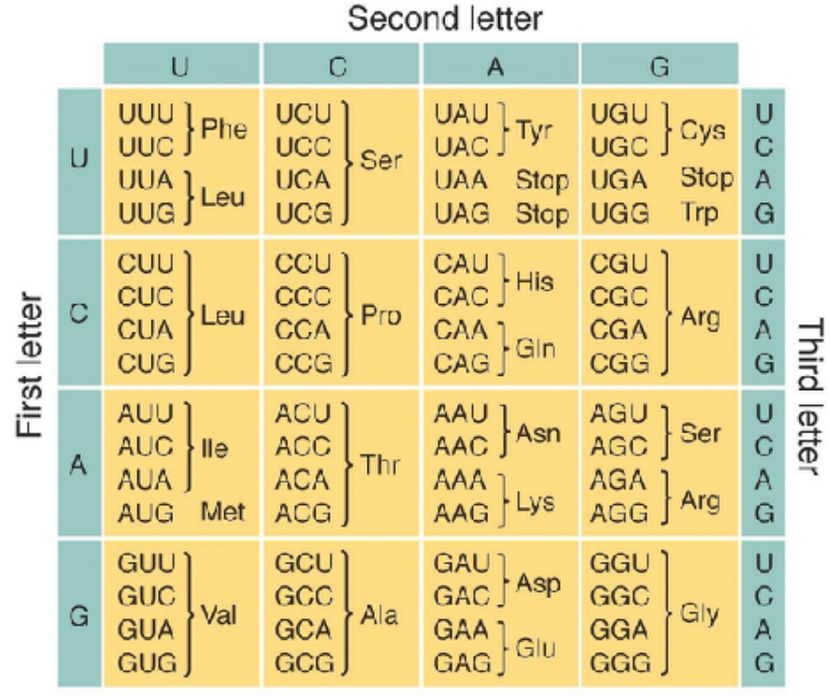


These are the signals we want to capture with a sequence model

Probabilistic model is appropriate since the signals are "fuzzy" to various degrees

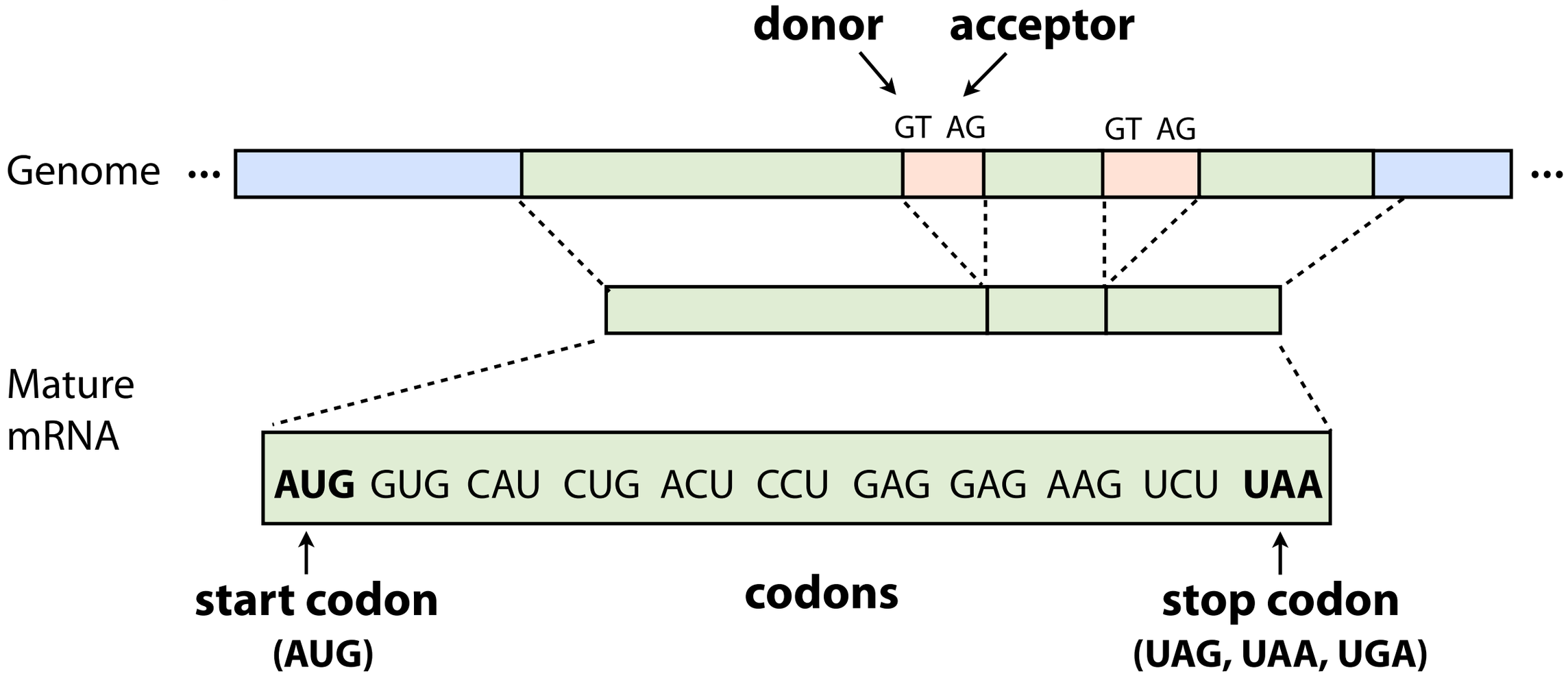
There are a lot of start codons, stop codons, donors and acceptors in the genome that have nothing to do with translation or splicing

Only some donor or acceptors *in a gene* are involved in splicing



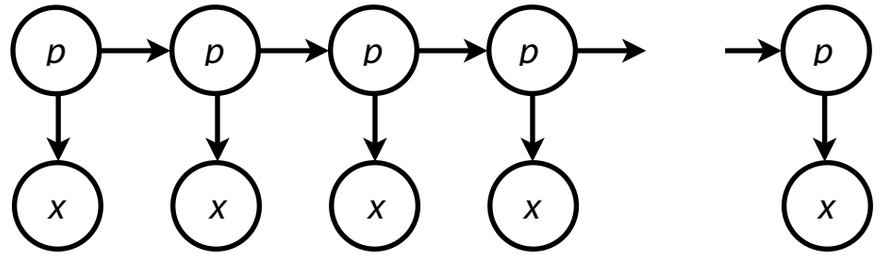
Exception: stop codon in a gene always stops translation

Transcription



Can HMM help us find eukaryotic genes and their constituent parts?

What will be the states? Emissions?

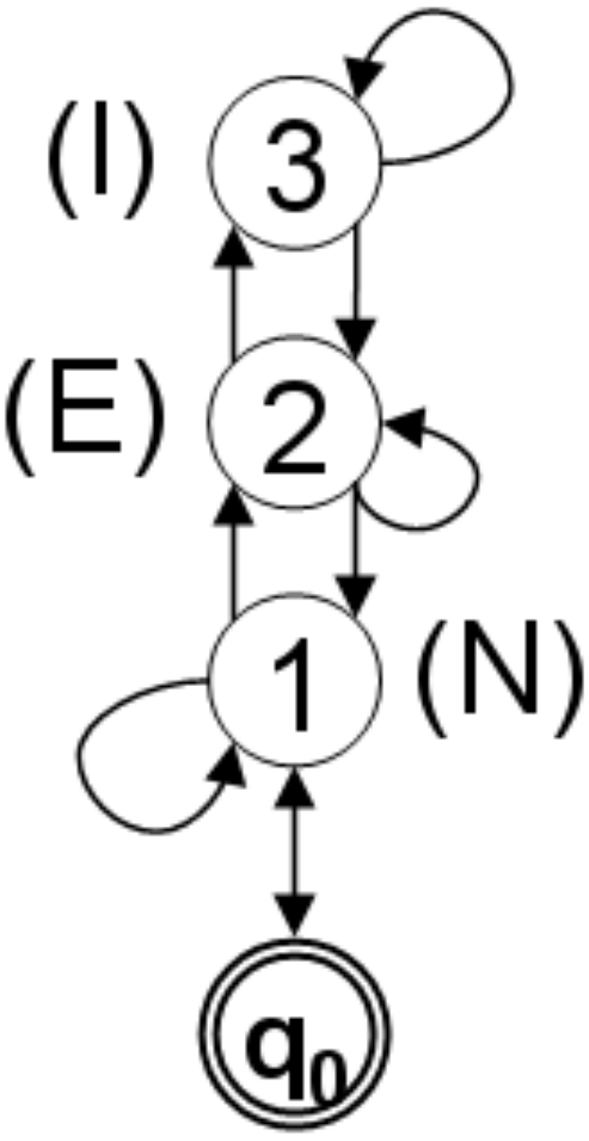


Eukaryotic gene finding

Emissions are nucleotides

I = intron
E = exon
N = intergenic
(between genes)

q0 is a *start state*; guarantees we start in the N (intergenic) state



Model captures:

Exons and introns and space between genes

Does not capture:

Acceptors & donors, start & stop codons, other codons

Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.

Eukaryotic gene finding

What if we wanted to model the three codon positions separately?

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe, UUC Phe, UUA Leu, UUG Leu	UCU Ser, UCC Ser, UCA Ser, UCG Ser	UAU Tyr, UAC Tyr, UAA Stop, UAG Stop	UGU Cys, UGC Cys, UGA Stop, UGG Trp	U C A G
C	CUU Leu, CUC Leu, CUA Leu, CUG Leu	CCU Pro, CCC Pro, CCA Pro, CCG Pro	CAU His, CAC His, CAA Gln, CAG Gln	CGU Arg, CGC Arg, CGA Arg, CGG Arg	U C A G
A	AUU Ile, AUC Ile, AUA Ile, AUG Met	ACU Thr, ACC Thr, ACA Thr, ACG Thr	AAU Asn, AAC Asn, AAA Lys, AAG Lys	AGU Ser, AGC Ser, AGA Arg, AGG Arg	U C A G
G	GUU Val, GUC Val, GUA Val, GUG Val	GCU Ala, GCC Ala, GCA Ala, GCG Ala	GAU Asp, GAC Asp, GAA Glu, GAG Glu	GGU Gly, GGC Gly, GGA Gly, GGG Gly	U C A G

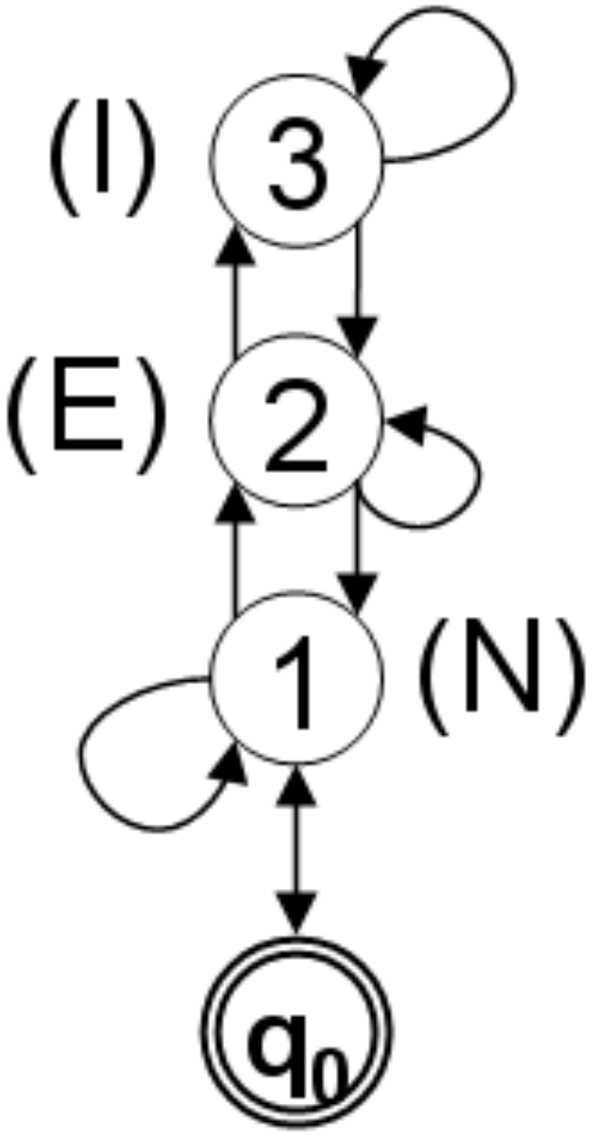


Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.

Eukaryotic gene finding

As before:

I = intron
E = exon
N = intergenic
(between genes)

Can we additionally model start & stop codons and donors & acceptors?

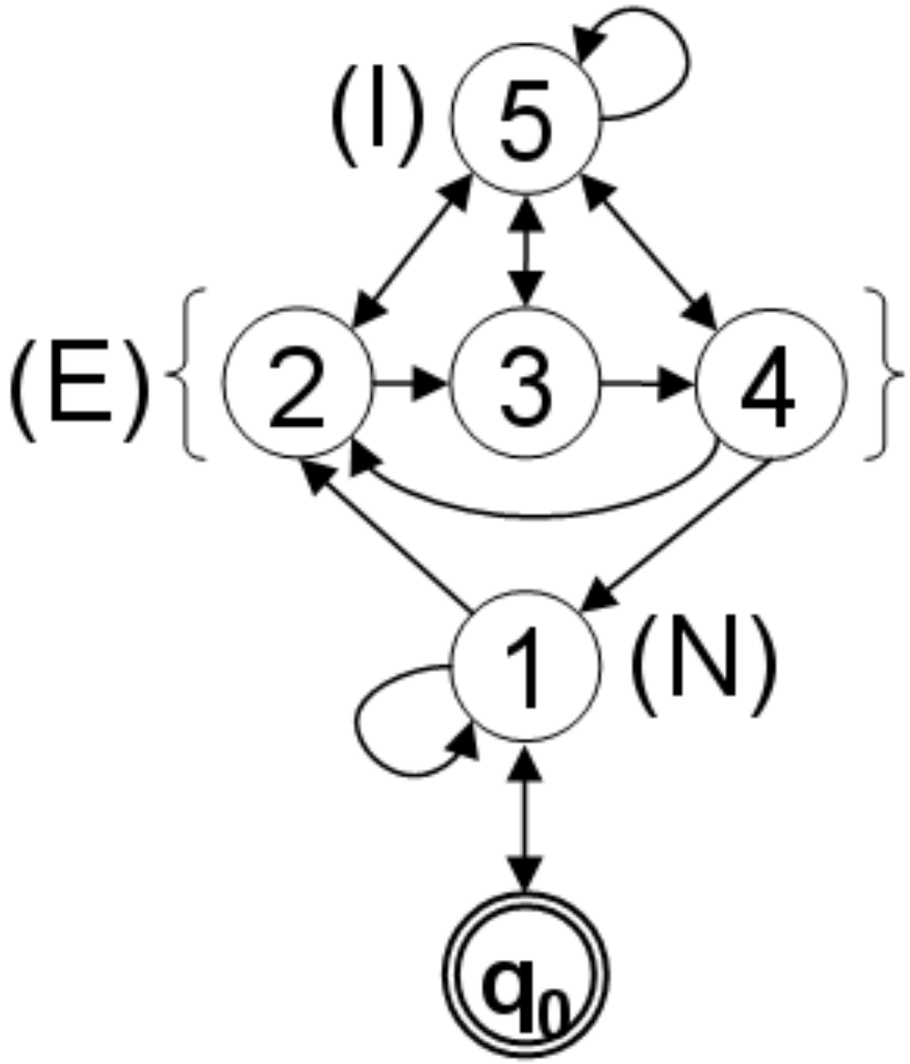


Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.

Eukaryotic gene finding

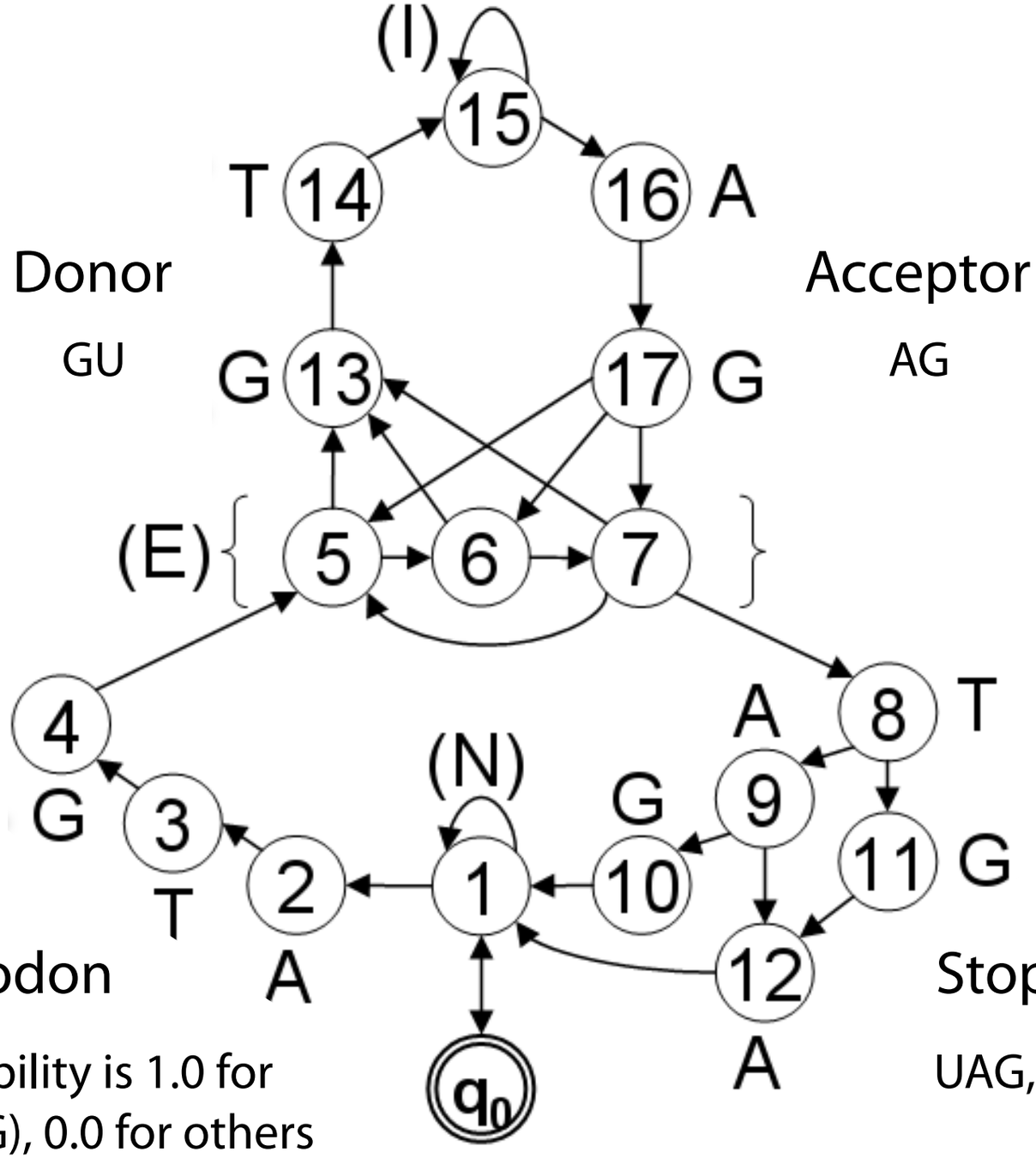


Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.

Eukaryotic gene finding

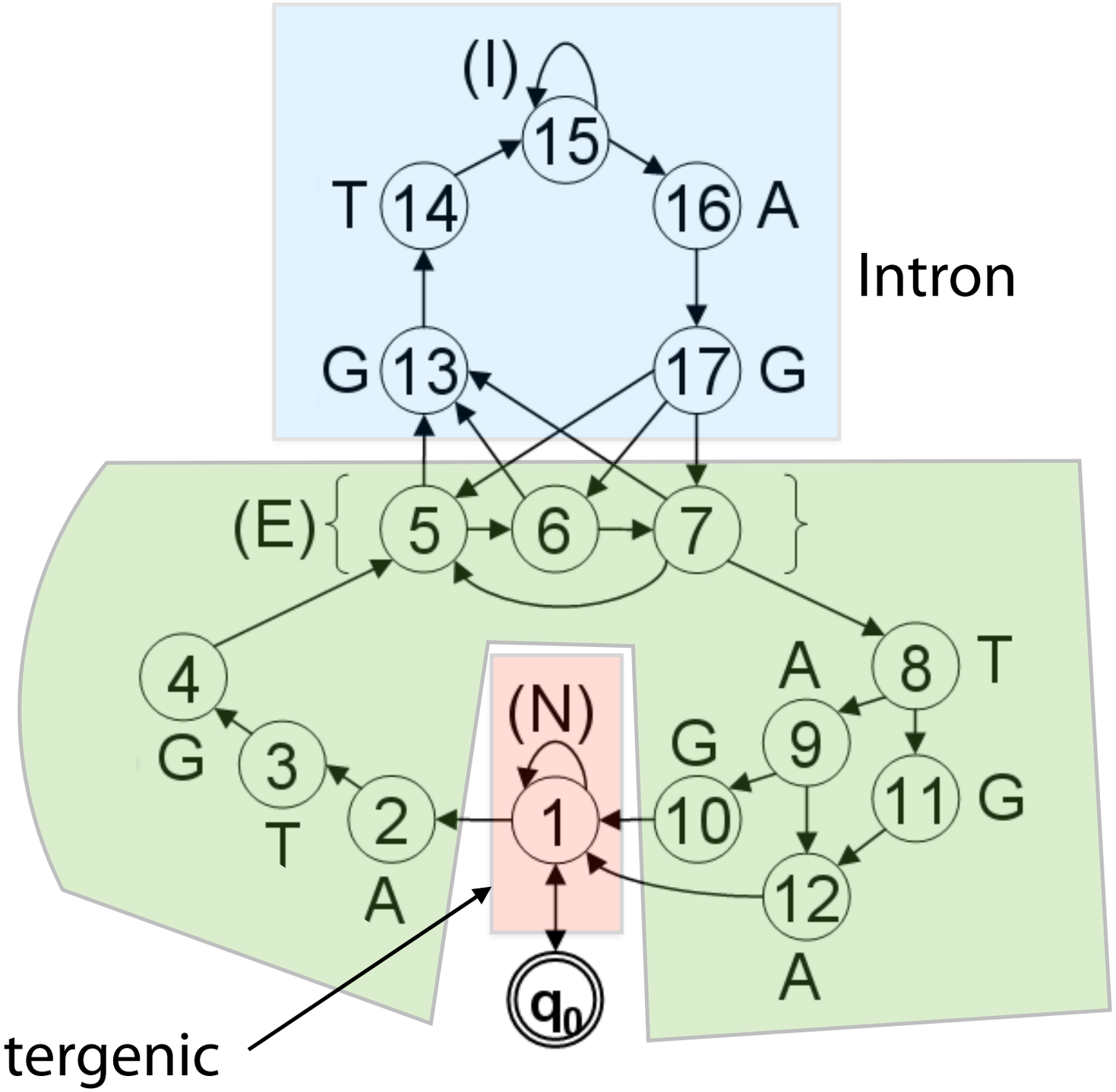


Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.

Eukaryotic gene finding

Which nodes have non-trivial emission probabilities that we must learn?

All non-motif nodes

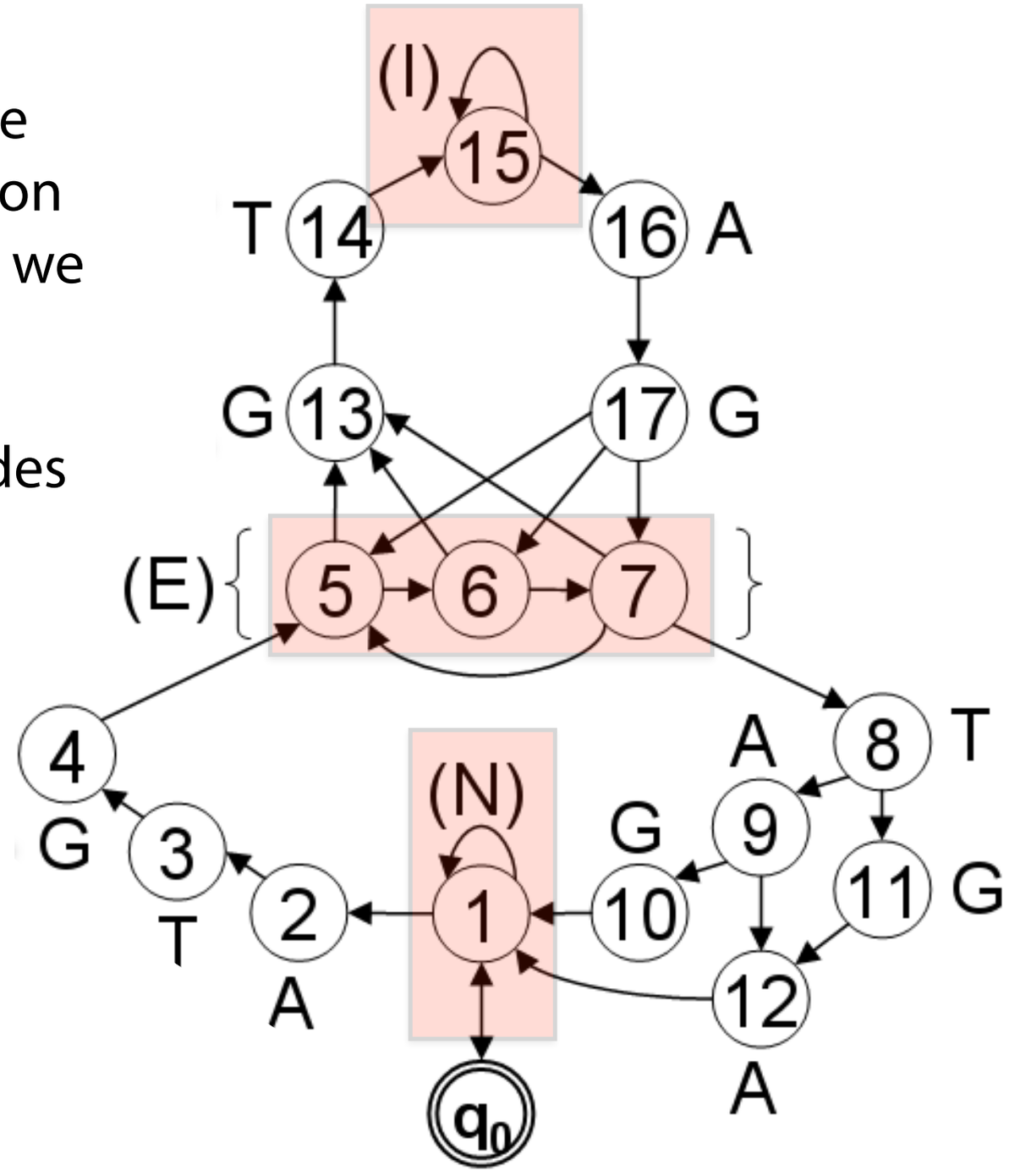
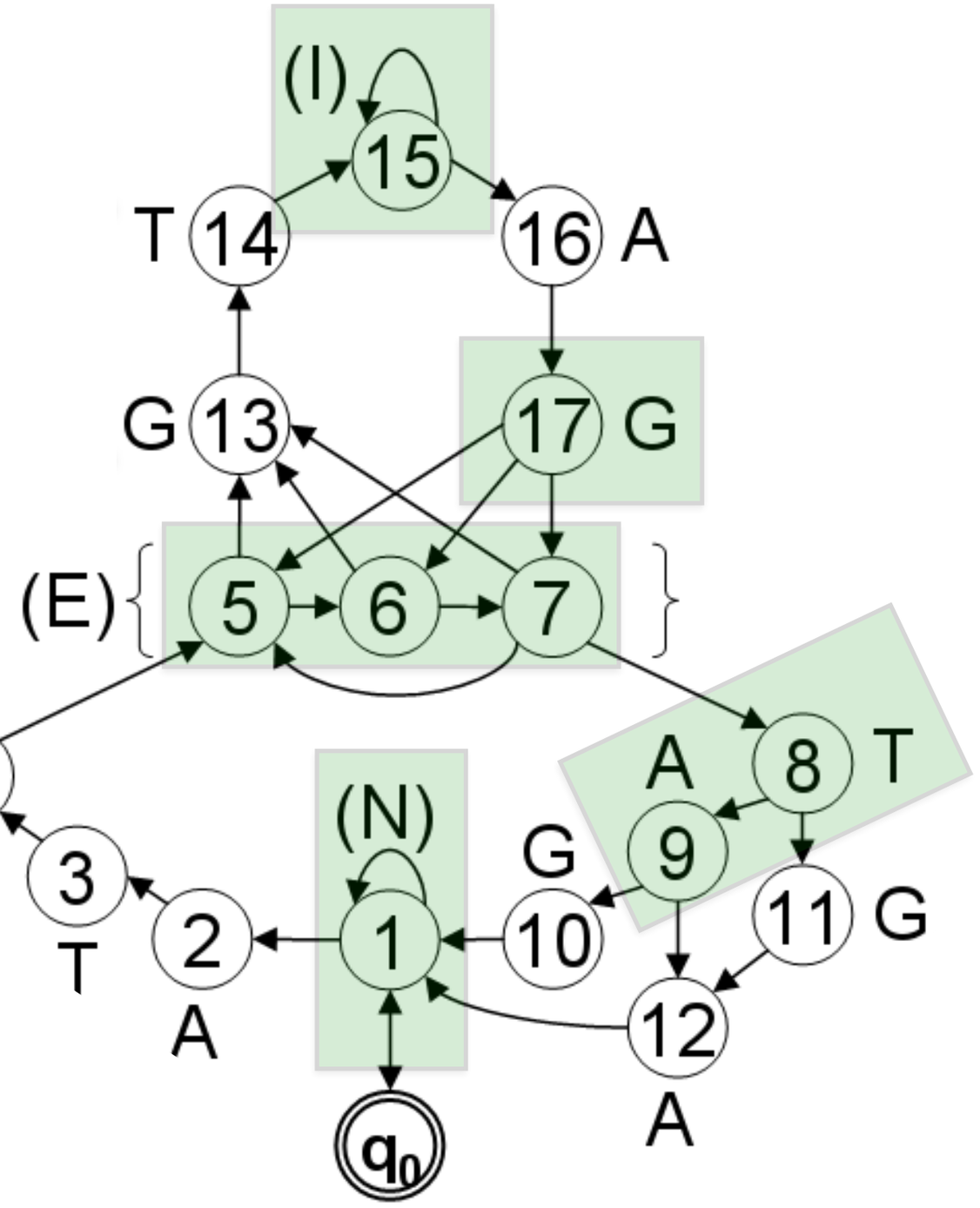


Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.

Eukaryotic gene finding

Which edges have non-trivial transition probabilities that we must learn?

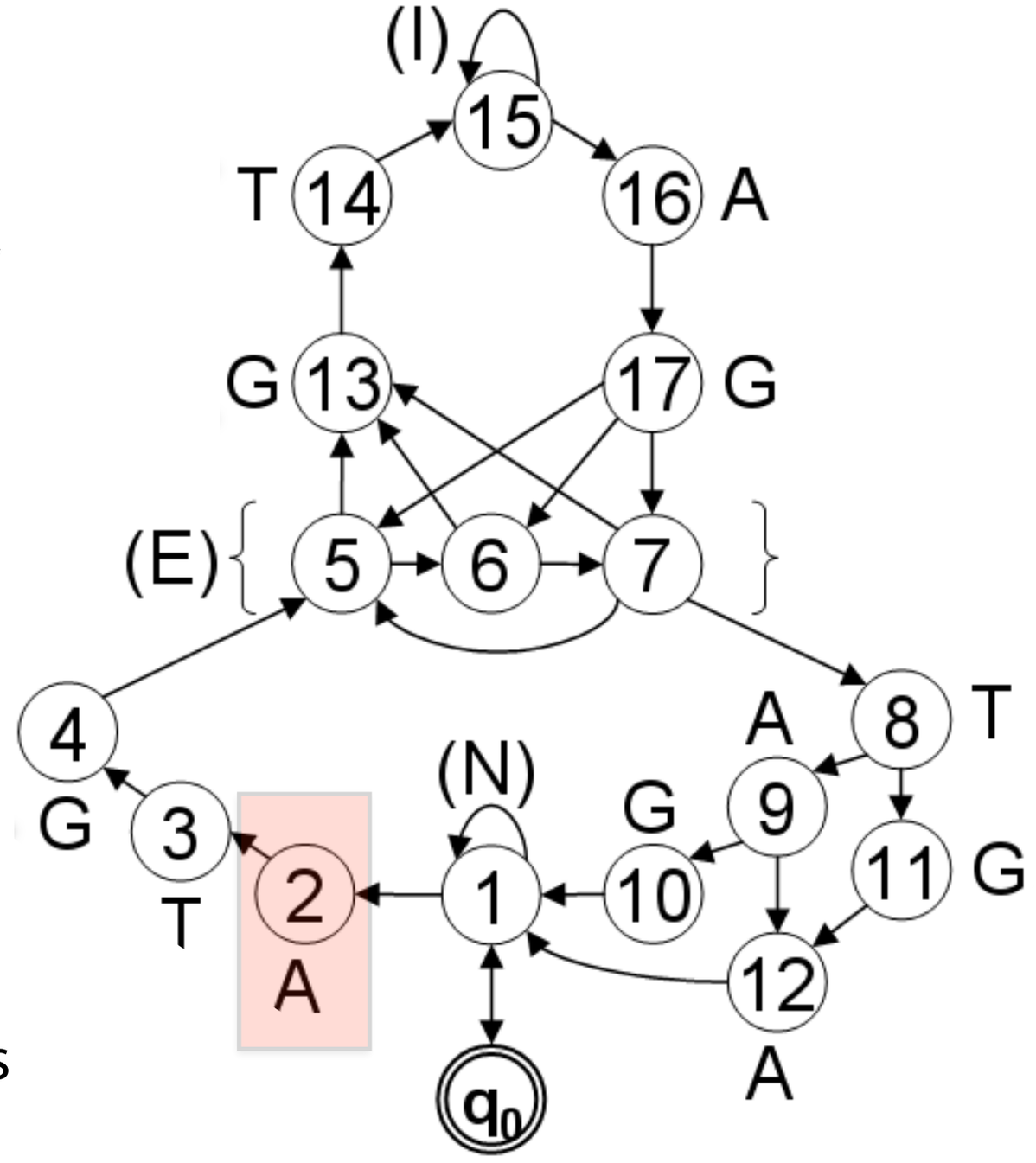


Edges outgoing from nodes with multiple outgoing edges.

Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.

Eukaryotic gene finding

Given a trained model and emission string, is it possible for a backtrace (not necessarily the Viterbi backtrace) to have probability 0?



Yes: e.g. any backtrace that puts us in state 2 at a step where the emission string does not have "A"

Figure: Majoros, William H. *Methods for computational gene prediction*. Vol. 1. Cambridge: Cambridge University Press, 2007.