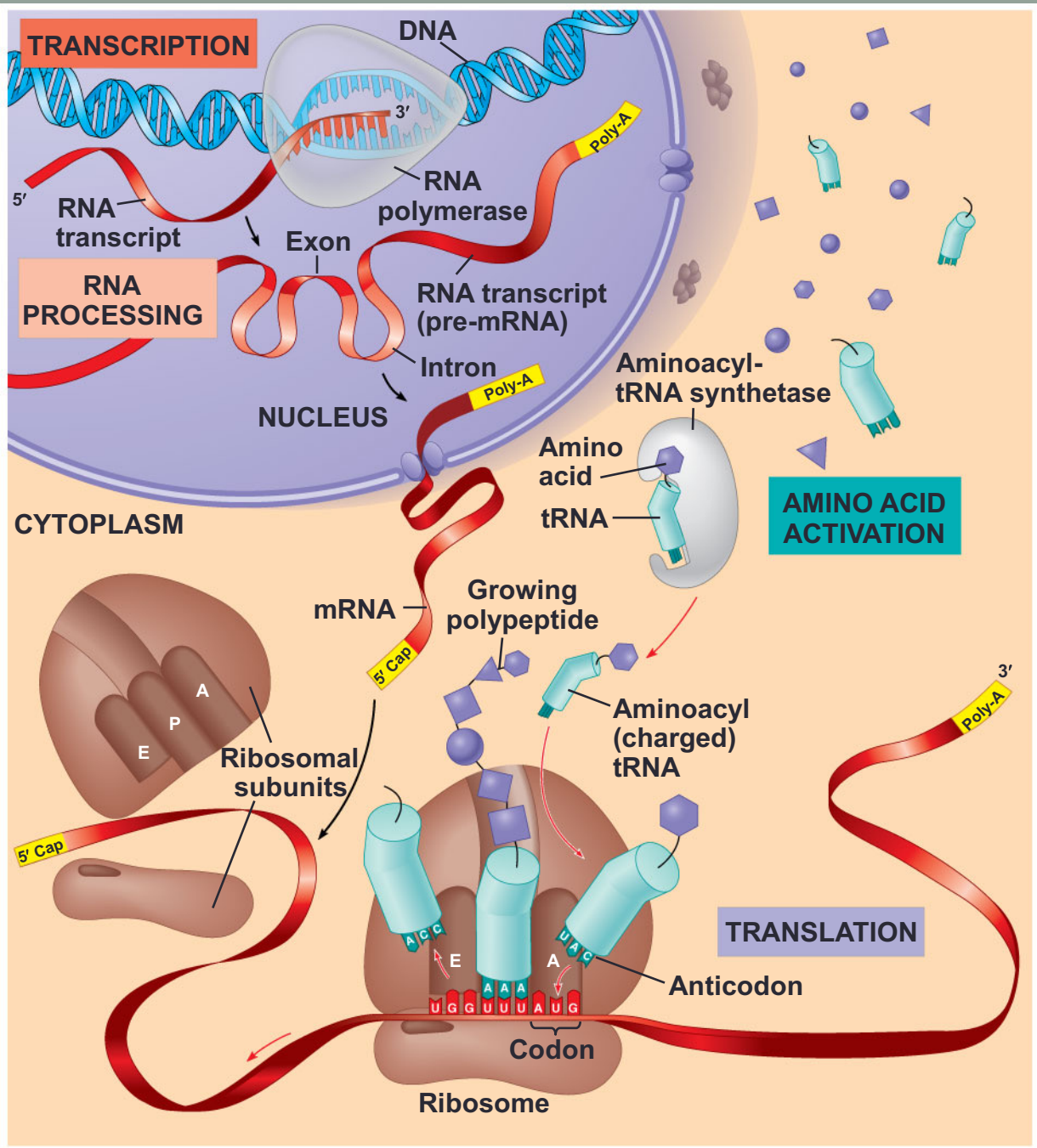
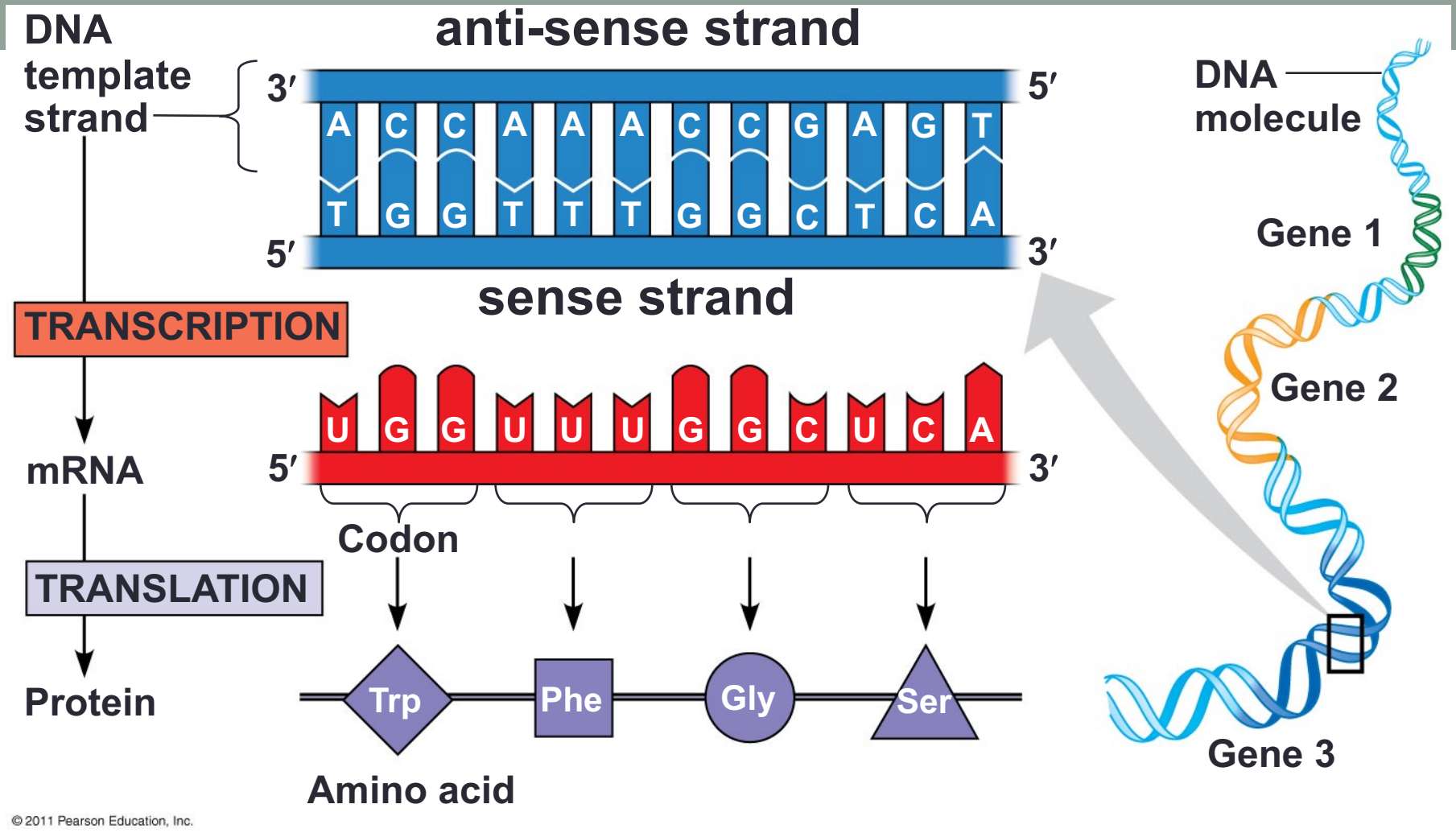
CHAPTER 9: GENE PREDICTION

Dr. Garrett Dancik

What is a gene?

- a region of DNA that can be expressed to produce a final functional product, either
  - a polypeptide or
  - an RNA molecule





- The genetic code is a triplet code where a 3-nucleotide DNA word codes for a 3-nucleotide mRNA word (a **codon**) which codes for an amino acid

Gene Prediction by Homology

- New DNA sequences can be searched (e.g., BLASTED) against various databases
  - blastx – search a protein database using a translated nucleotide query
  - tblastx – search a translated nucleotide database using a translated nucleotide query
- Generally, >50% of prokaryotic genes can be identified by homology
- Gene prediction in this manner is more difficult for eukaryotic organisms
  - Why?

Sequence Translation Revisited

- Suppose you have a sequence of DNA that includes a gene (you don't know exactly where the gene is). What are the possible polypeptide sequences that could (theoretically) be produced?
  - 5' – GATGGATGACGCGATGATCC – 3'
- Let's look at the Expasy Translate tool:
  - http://web.expasy.org/translate/

Sequence Translation Revisited

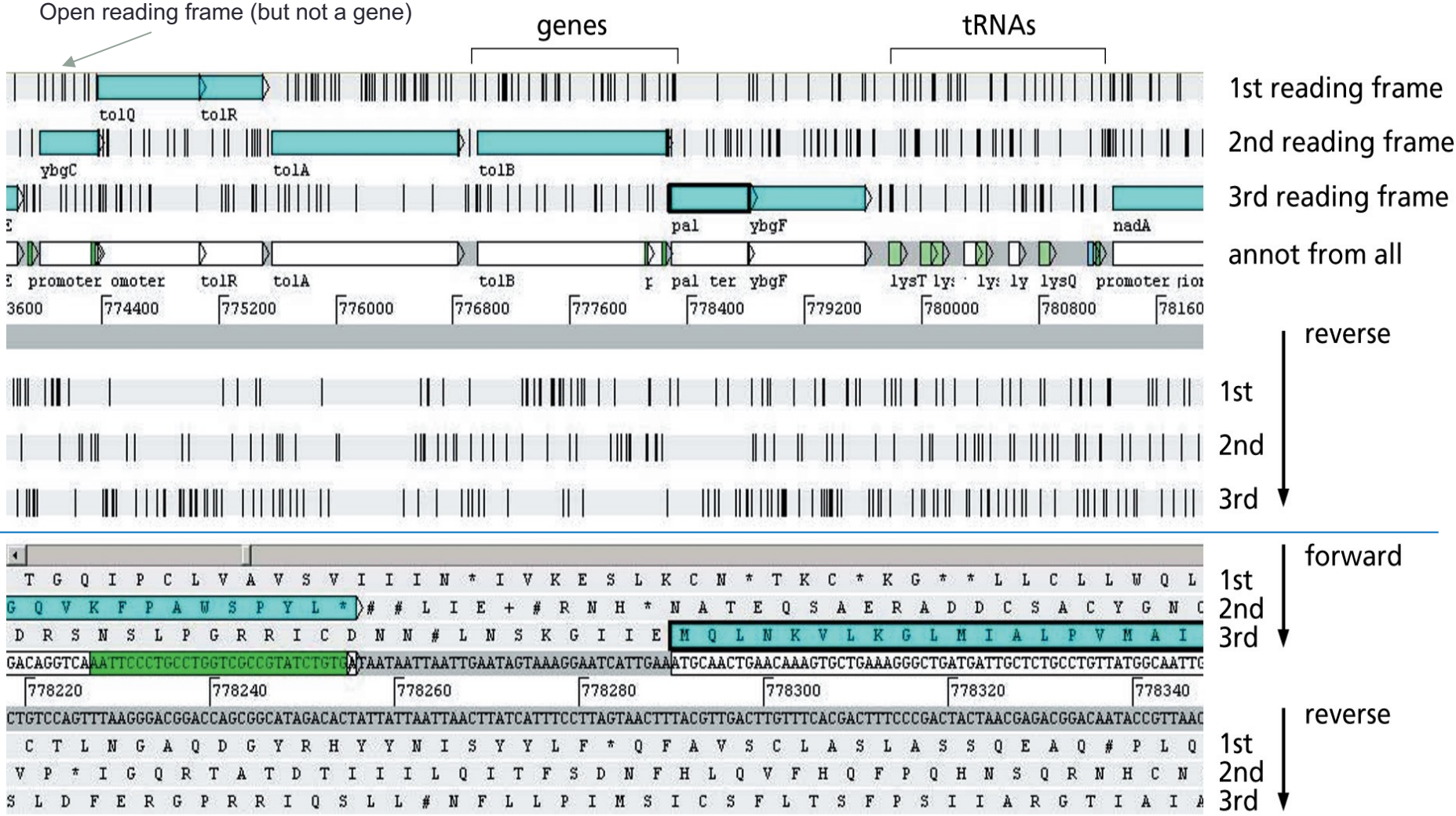
- We don't know where the first codon (in the sequence) begins.
  - 5' – GATGGATGACGCGATGA – 3'

Reading frame 1	GAT	GGA	TGA	CGC	GAT	GA
Reading frame 2	ATG	GAT	GAC	GCG	ATG	A
Reading frame 2	TGG	ATG	ACG	CGA	TGA	

- We don't know which *strand* is the *sense* strand (need to consider the *reverse complement*)
  - 5' – TCATCGCGTCATCCATC – 3'

Reading frame 4	TCA	TCG	CGT	CAT	CCA	TC
Reading frame 5	CAT	CGC	GTC	ATC	CAT	C
Reading frame 6	ATC	GCG	TCA	TCC	ATC	

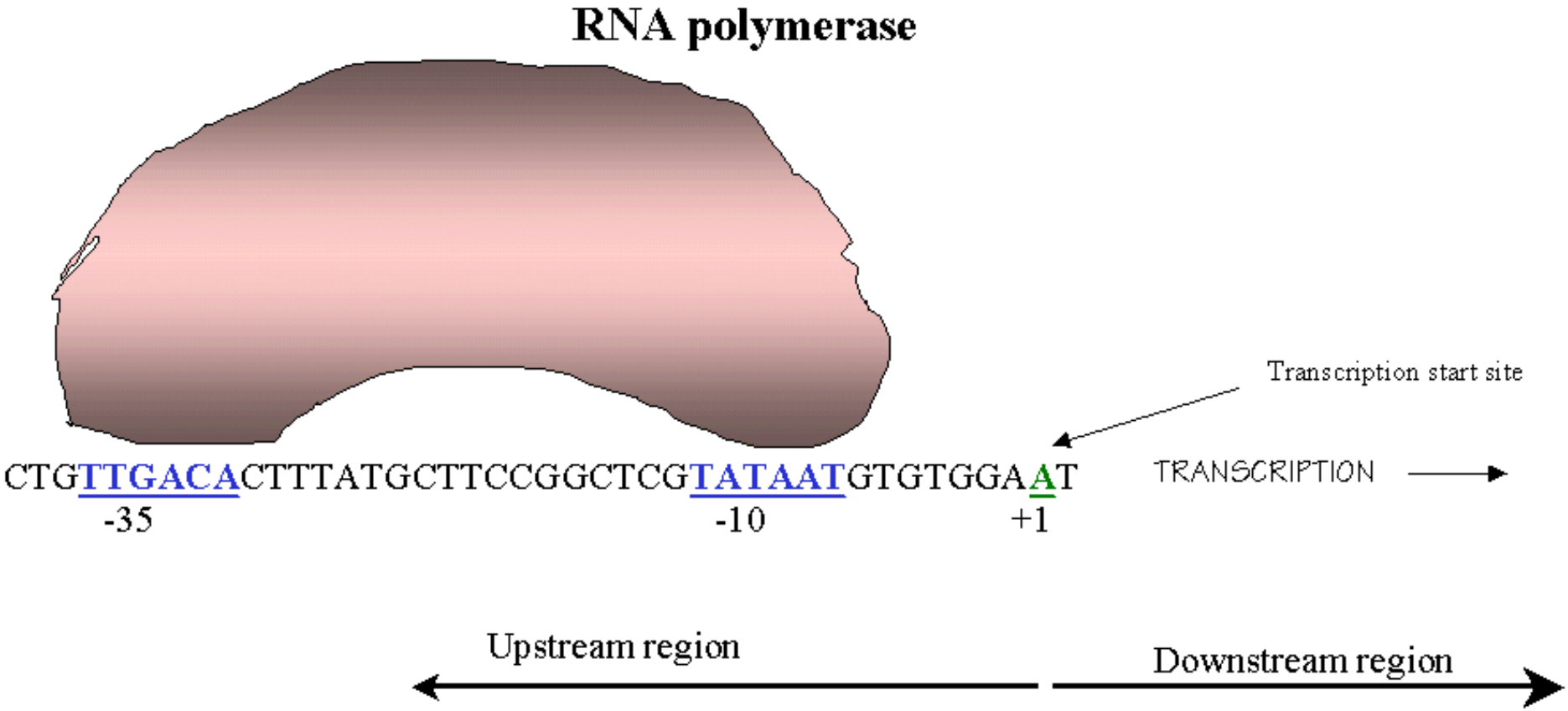
Annotation of a segment of the *E. coli* genome



Observations

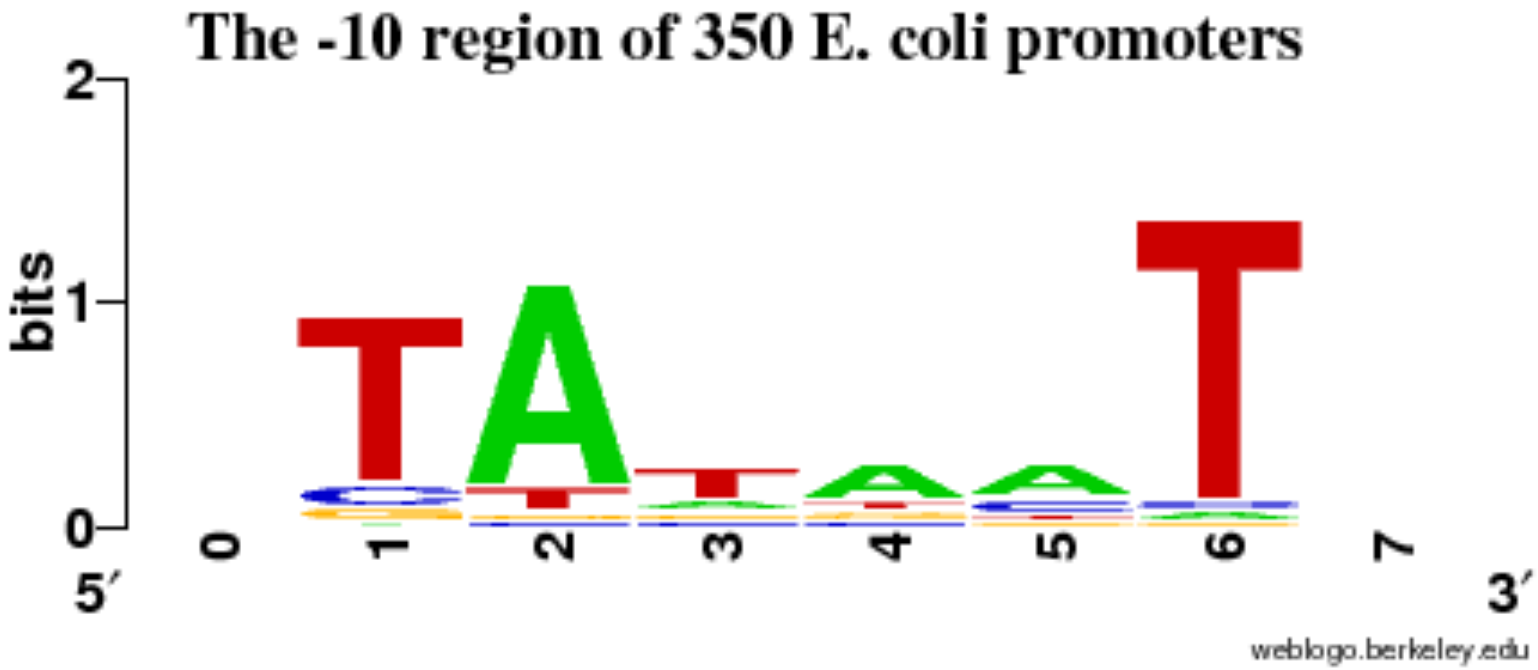
- Non-coding genes such as tRNAs do not have corresponding proteins
  - These have conserved structures that aid in their identification
- Definition: an *open reading frame* is a DNA sequence that contains no stop codons
- Actual protein-coding genes correspond to regions of DNA with large open reading frames, that begin with a start codon and end with a stop codon.
- Simple algorithm (for prokaryotes):
  - Search for a start codon. If not found, then there are no protein coding genes in this sequence
  - Search for a stop codon in the same reading frame as the start codon. Discard the ORF if its length is less than a threshold (e.g., 100 amino acids)
  - Repeat until all candidate genes are found

Promoter identification



- A promoter is a region of DNA where RNA polymerase binds.
- Prokaryotic gene promoters have two conserved sequences
  - -10 sequence: TATAAT approximately 10 bp upstream of the transcription start site
  - -35 sequence TTGACA approximately 35 bp upstream of transcription start site
  - The two above sequences may not be exact

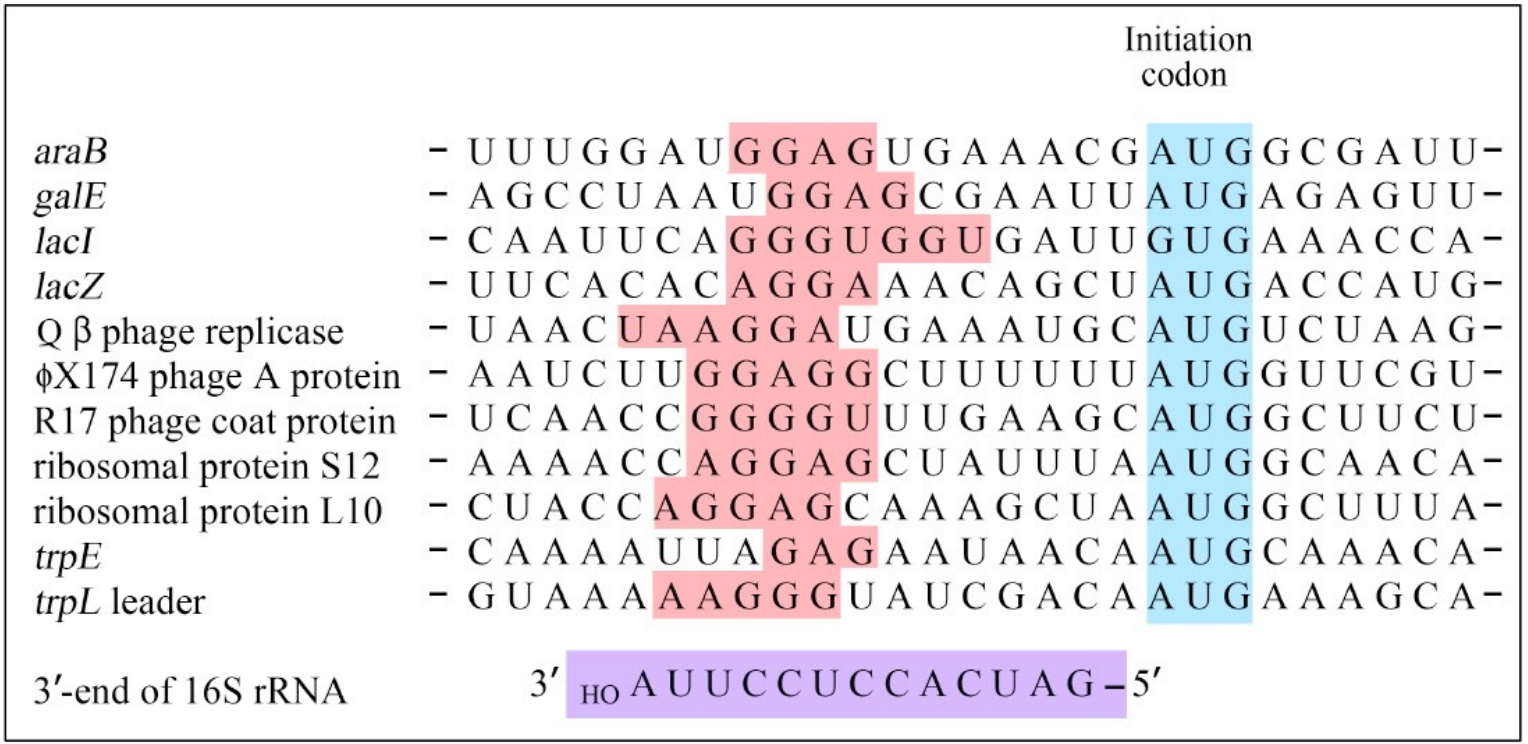
Sequence logo of -10 sequence



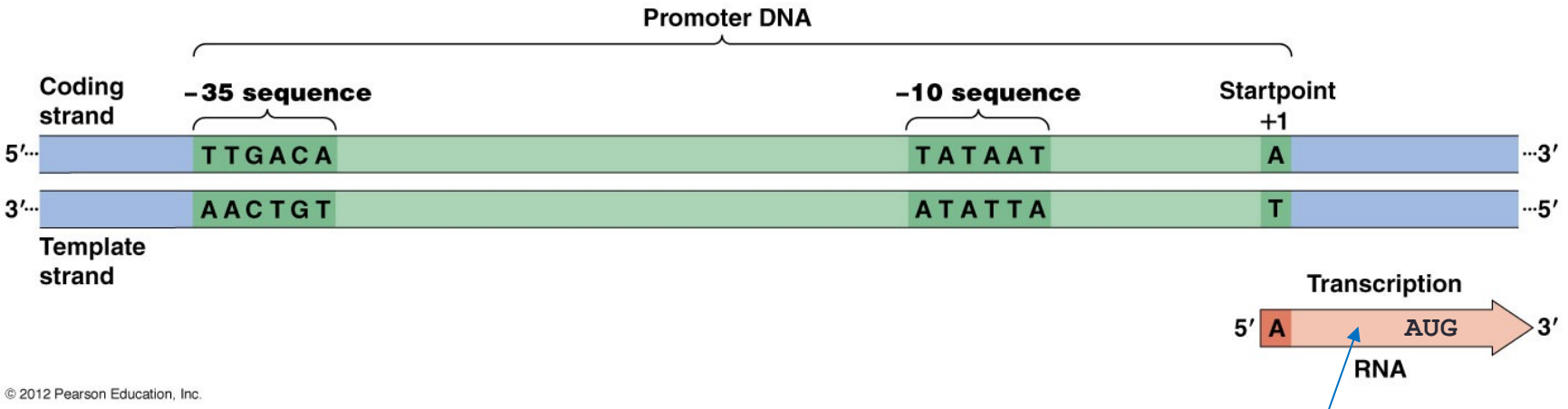
- The height of a *position* corresponds to how conserved the position is
- At each position, the height of each character is proportional to its frequency

Shine-Delgarno sequence (prokaryotes)

- The Shine-Dalgarno sequence (or ribosome binding site) precedes the start codon by a few bases and is where the ribosome binds to the corresponding mRNA.
- Consensus sequence is AGGAGG



Putting it together…



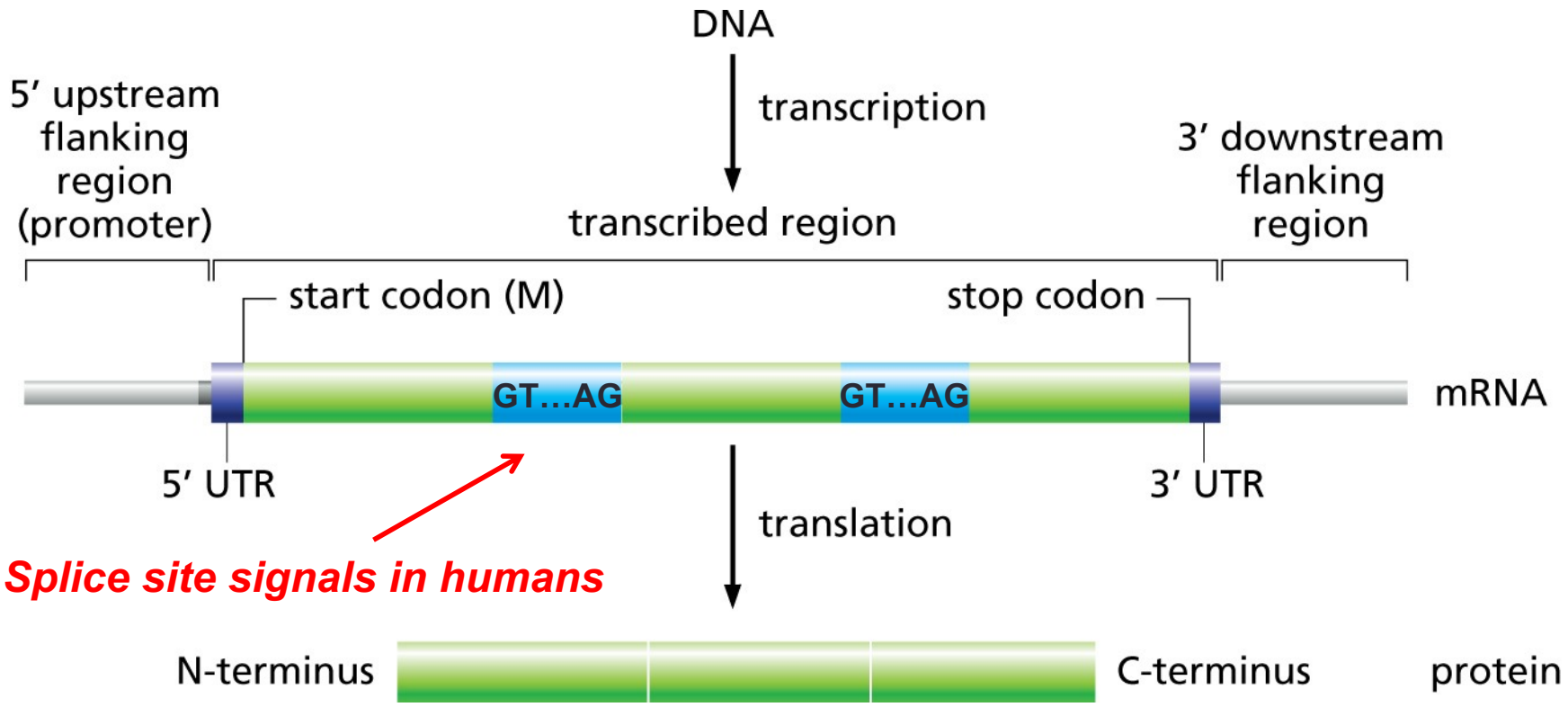
Shine-Dalgarno Sequence, AGGAGG, ~ 5bp upstream of start codon

Prokaryotic Gene Prediction Algorithm

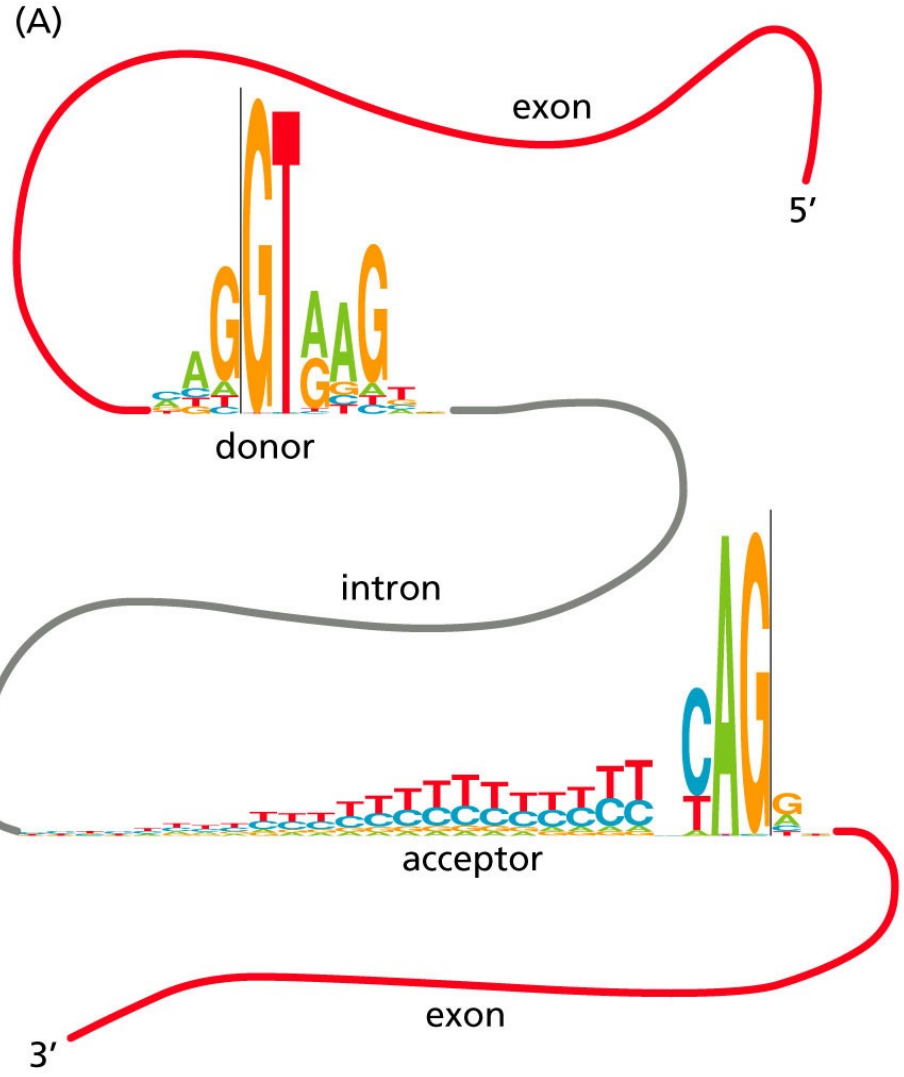
Sequences that include an ORF of a minimum length, a Shine-Dalgarno sequence, and conserved promoter elements are candidate genes:

1. Search for the next start codon. If no start codon is found, end.
2. Search for a stop codon in the same reading frame as the start codon. Continue only if the ORF length is greater than a threshold (e.g., 100 amino acids). Otherwise start over.
3. Search for a Shine-Dalgarno sequence 3-7 bases upstream of the start codon. The sequence should pass a matching threshold (e.g., 5/6 identity). If not found, start over.
4. Search 500 nucleotides upstream of the Shine-Dalgarno sequence for a promoter. The TTGACA promoter should be located 15-19 nucleotides upstream of TATAAT. Allow for one mismatch in each sequence (use of other consensus sequences is possible)

Gene expression in eukaryotes
(introns are spliced out)



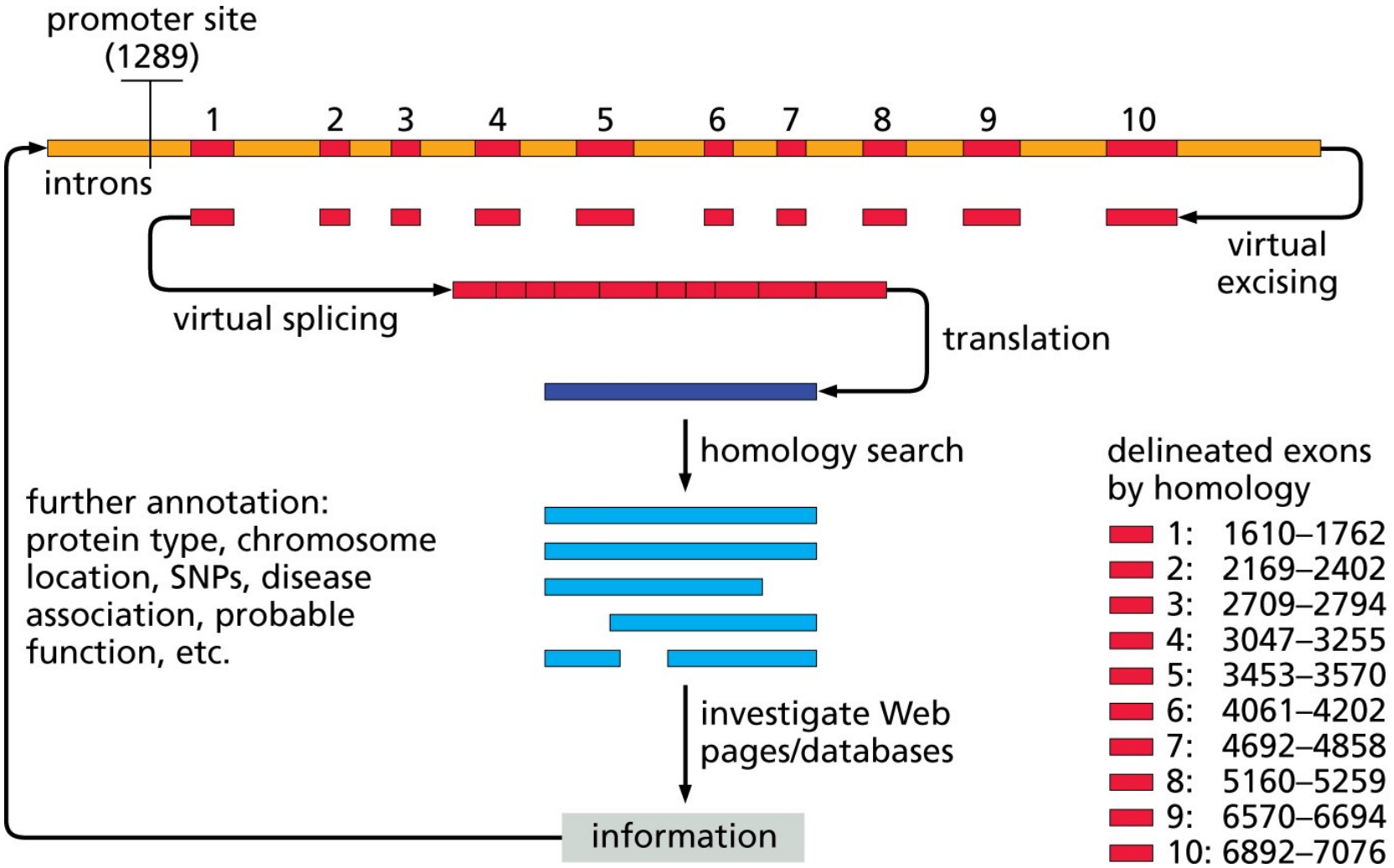
Sequence conservation of splice sites in humans



Gene Prediction in Eukaryotes

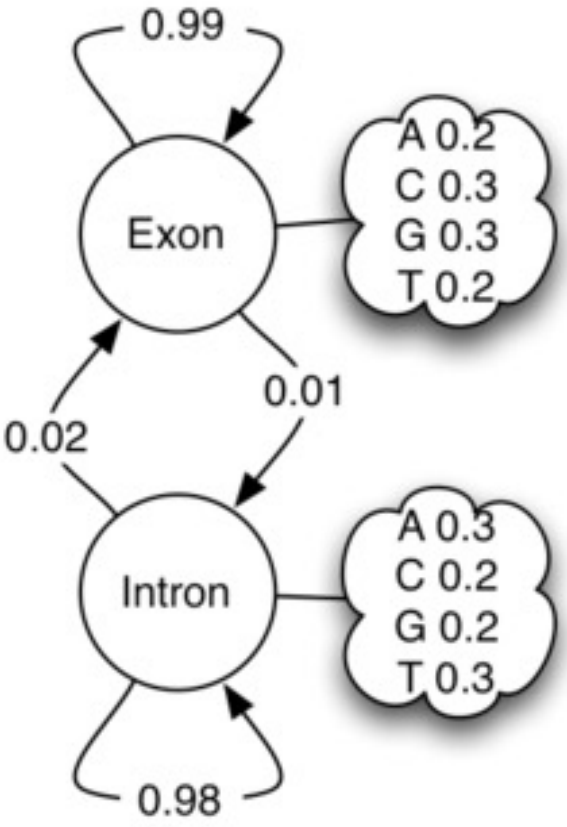
- Involves prediction of exons and introns
  - Based on statistical gene models and query sequence
  - Based on statistical gene models, sequence similarity, and a query sequence
- Must preserve the correct reading frame
- Involves prediction of the promoter

Eukaryotic Gene Prediction and Gene Annotation



Augustus

- http://bioinf.uni-greifswald.de/augustus/
- Uses a Hidden Markov Model (HMM)
- Probabilistic intron length model



A very simple HMM for gene structure

- Hidden states: exon and intron
- Transition probabilities
  - exon → exon: 0.99 intron → intron: 0.98
  - exon → intron: 0.01 intron → exon: 0.02
- Emission probabilities for observed values
  - Exon: A,C,G,T (0.2, 0.3, 0.3, 0.2)
  - Intron: A,C,G,T (0.3, 0.2, 0.2, 0.3)
- Objective: identify the most likely states (gene structure) given the observed values (the sequence)?