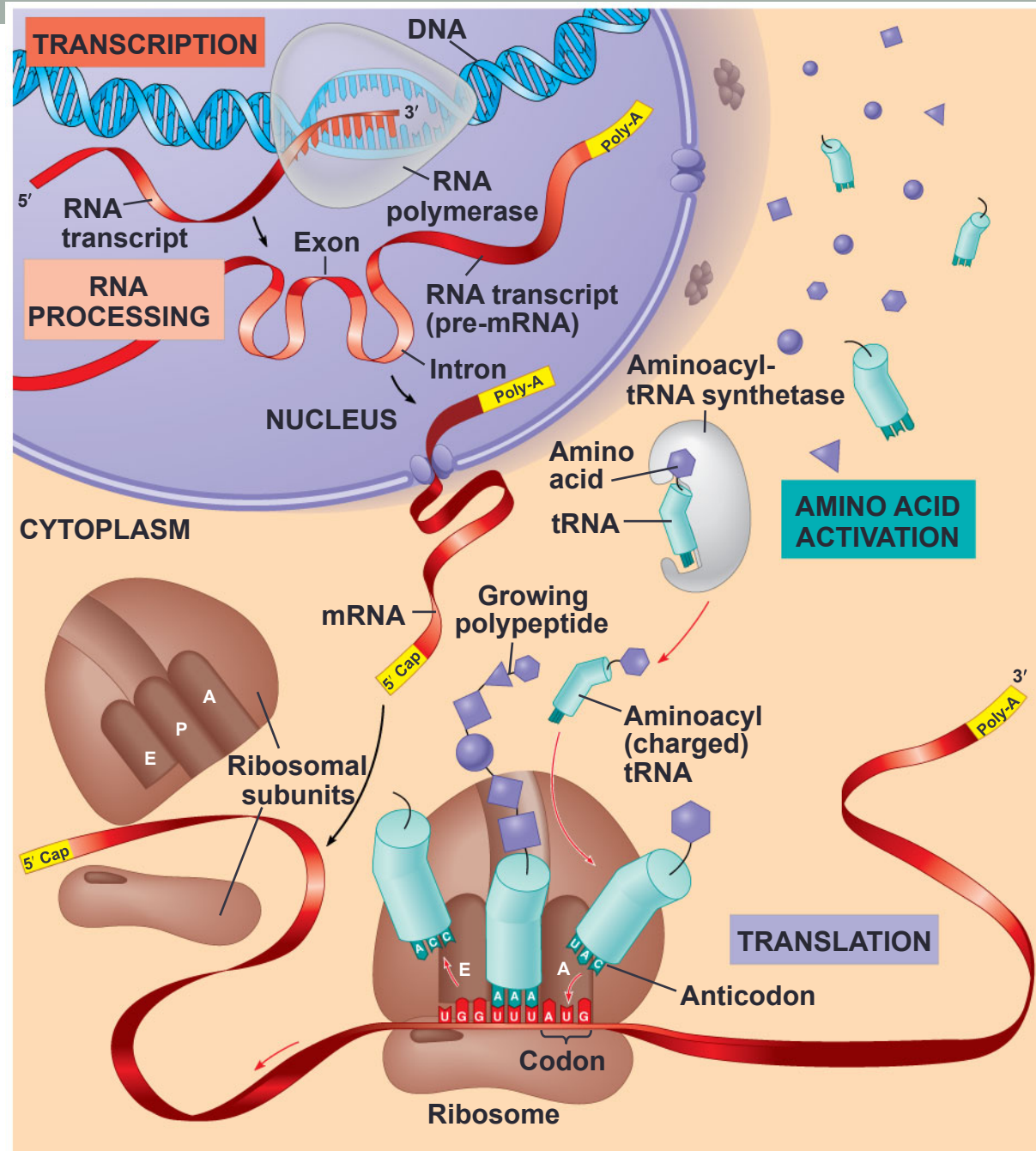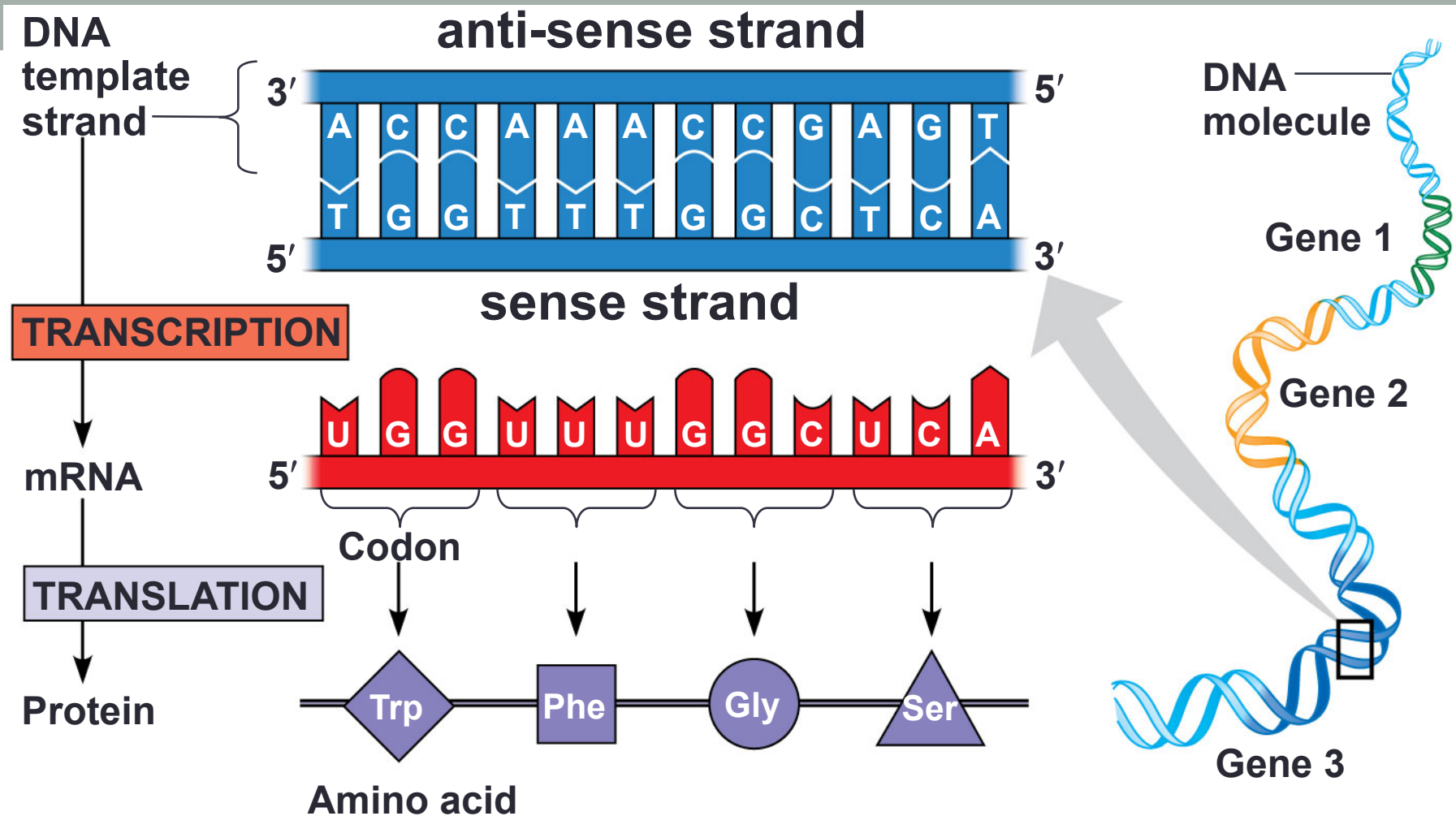# CHAPTER 9: GENE PREDICTION

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

TRANSCRIPTION

DNA

RNA polymerase

RNA transcript

Exon

RNA PROCESSING

RNA transcript (pre-mRNA)

Intron

NUCLEUS

CYTOPLASM

Amino acid

tRNA

Aminoacyl-tRNA synthetase

AMINO ACID ACTIVATION

mRNA

Growing polypeptide

Aminoacyl (charged) tRNA

Ribosomal subunits

TRANSLATION

Anticodon

Codon

Ribosome

Poly-A

5' Cap

- 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 <u>protein</u> database using a <u>translated nucleotide query</u>
  - tblastx – search a <u>translated nucleotide</u> database using a <u>translated nucleotide</u> 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 proteins that are produced?

  - 5' – GATGGATGACGCGATGA – 3'

- Let's look at the Expasy Translate tool:
  - http://web.expasy.org/translate/

- In both prokaryotes and eukaryotes, genes can occur in *any* reading frame and on either strand (always in the 5' to 3' direction)

- Also, not all genes are coding.

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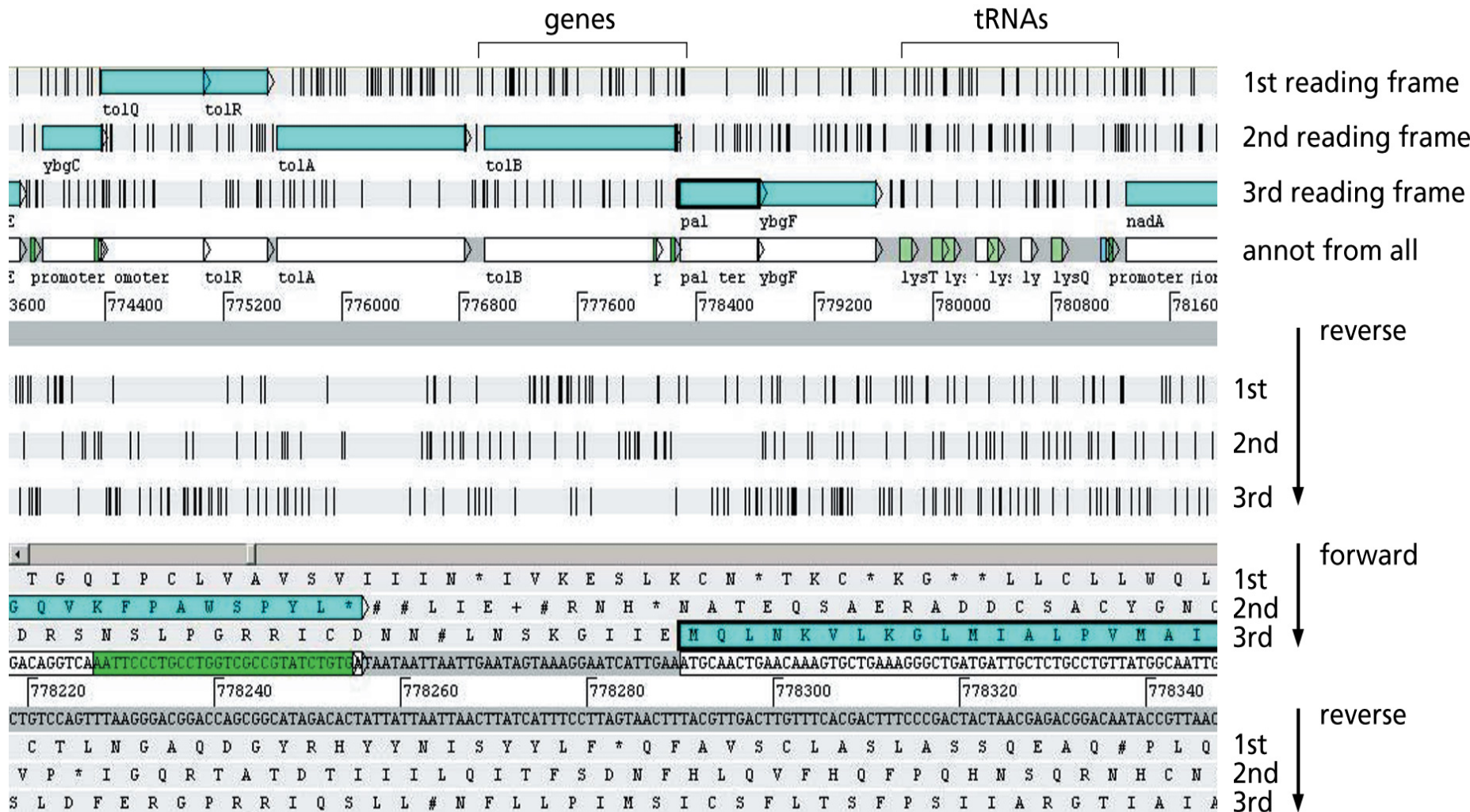
Observations

- 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 begins with the start codon ATG and ends with a stop codon
- The actual genes correspond to regions of DNA with large open reading frames
- Simple algorithm:
  - 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

RNA polymerase

CTG**TTGACA**CTTTATGCTTCCGGCTCG**TATAAT**GTGTGGA**A**T
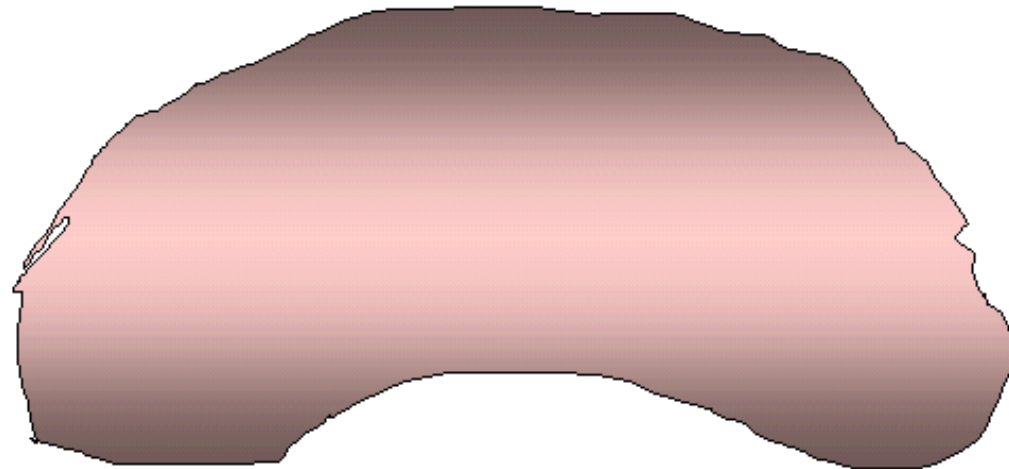
-35   -10   +1

Transcription start site

TRANSCRIPTION

Upstream region   Downstream region

- 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

Consensus logo of -10 sequence

The -10 region of 350 E. coli promoters

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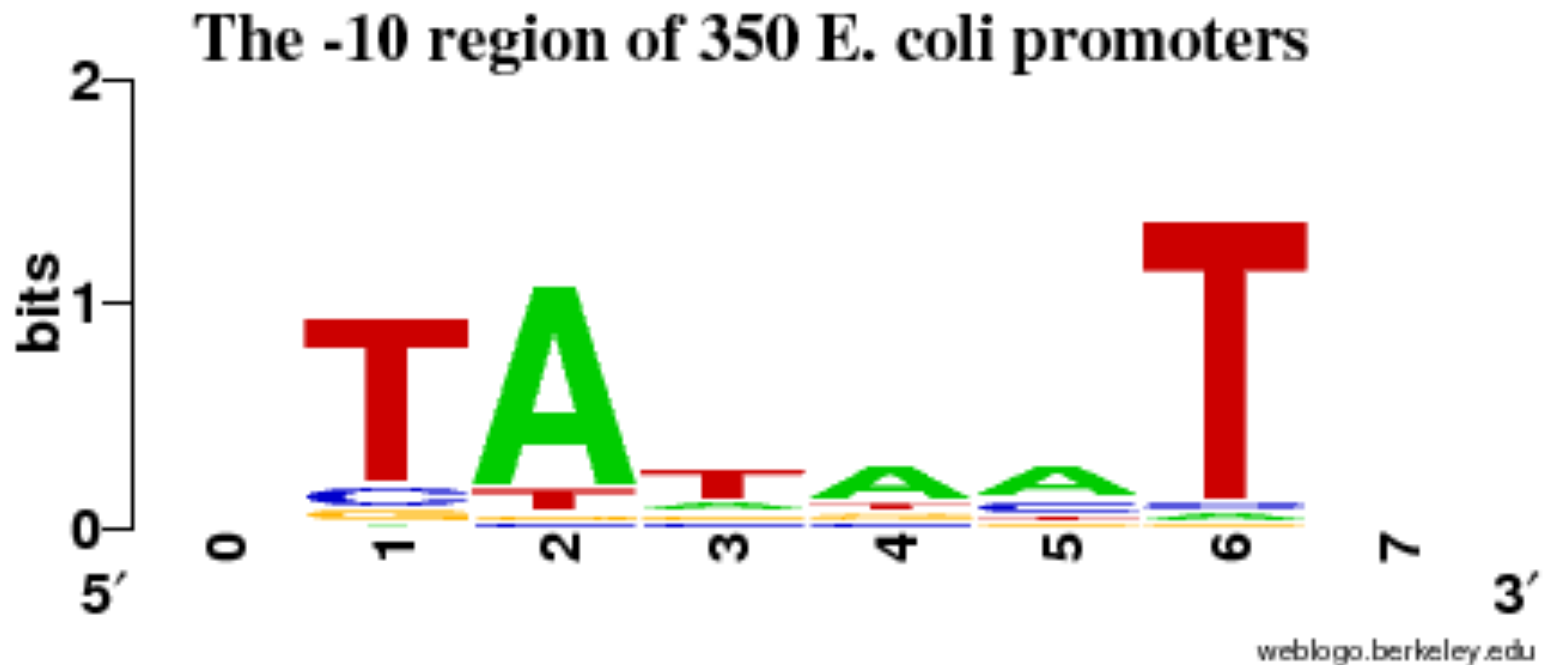
- The height of a *position* corresponds to how conserved the position is
- At each position, the height of each character is proportional to its frequency

Shine-Delgarno sequence

- The Shine-Delgarno 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

	Initiation codon
araB	– U U U G G A U G G A G U G A A A C G A U G G C G A U U –
galE	– A G C C U A A U G G A G C G A A U U A U G A G A G U U –
lacI	– C A A U U C A G G G U G G U G A U U G U G A A A C C A –
lacZ	– U U C A C A C A G G A A A C A G C U A U G A C C A U G –
Q β phage replicase	– U A A C U A A G G A U G A A A U G C A U G U C U A A G –
φX174 phage A protein	– A A U C U U G G A G G C U U U U U U U A U G G U U C G U –
R17 phage coat protein	– U C A A C C G G G G U U U G A A G C A U G G C U U C U –
ribosomal protein S12	– A A A A C C A G G A G C U A U U U U A A U G G C A A C A –
ribosomal protein L10	– C U A C C A G G A G C A A A G C U A A U G G C U U U U A –
trpE	– C A A A A U U A G A G A A U A A C A A U G C A A A C A –
trpL leader	– G U A A A A A G G G U A U C G A C A A U G A A A G C A –

3′-end of 16S rRNA   3′ HO A U U C C U C C A C U A G – 5′

Putting it together…


© 2012 Pearson Education, Inc.
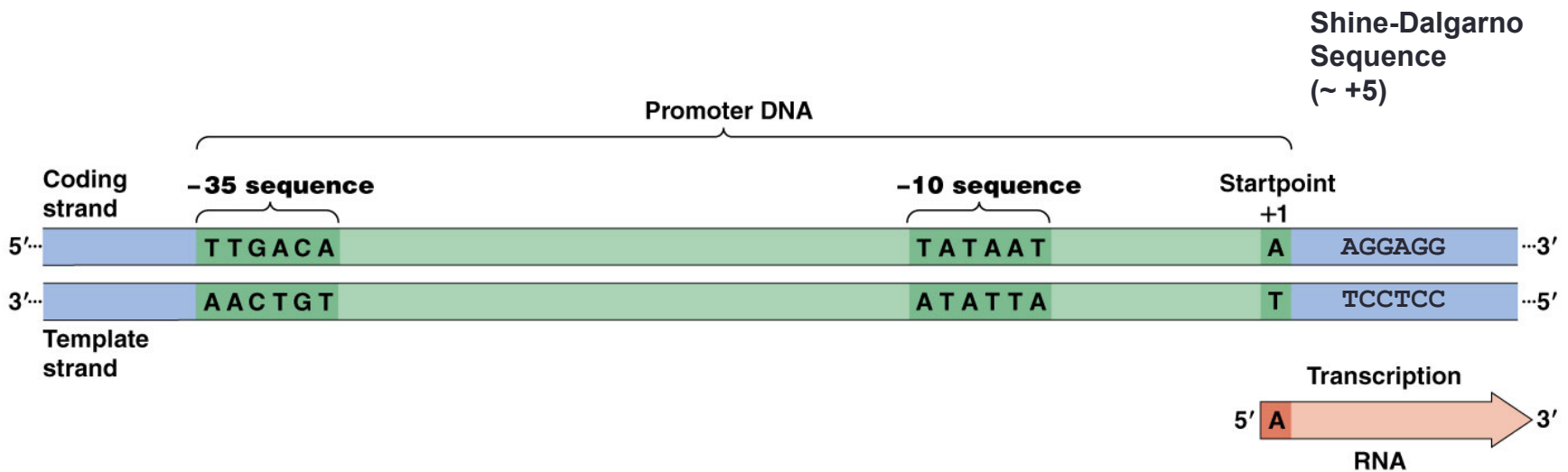
Prokaryotic Gene Prediction Algorithm

Gene sequences including an ORF of a certain length, a Shine-Dalgarno sequence, and promoter, 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

Gene expression in eukaryotes


Splice site signals in humans
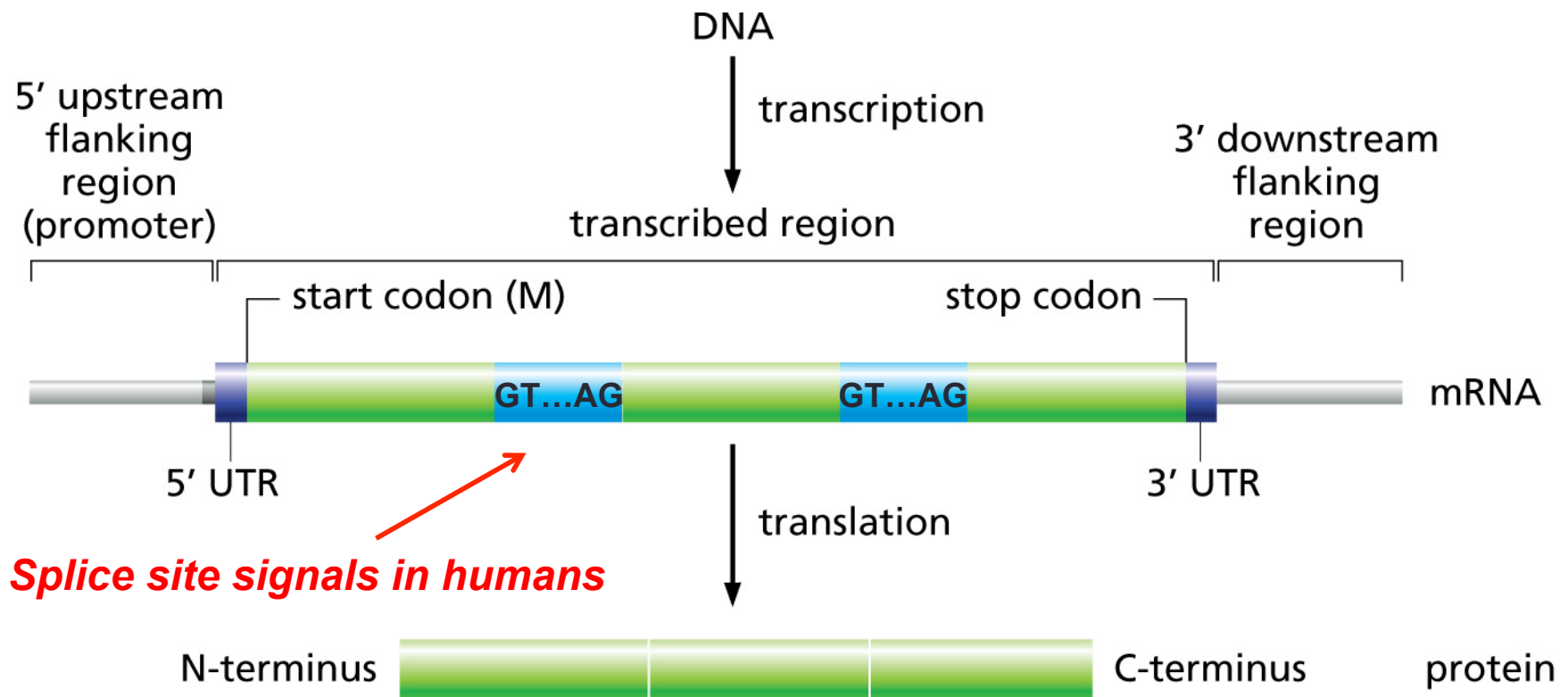
Sequence conservation of splice sites in humans
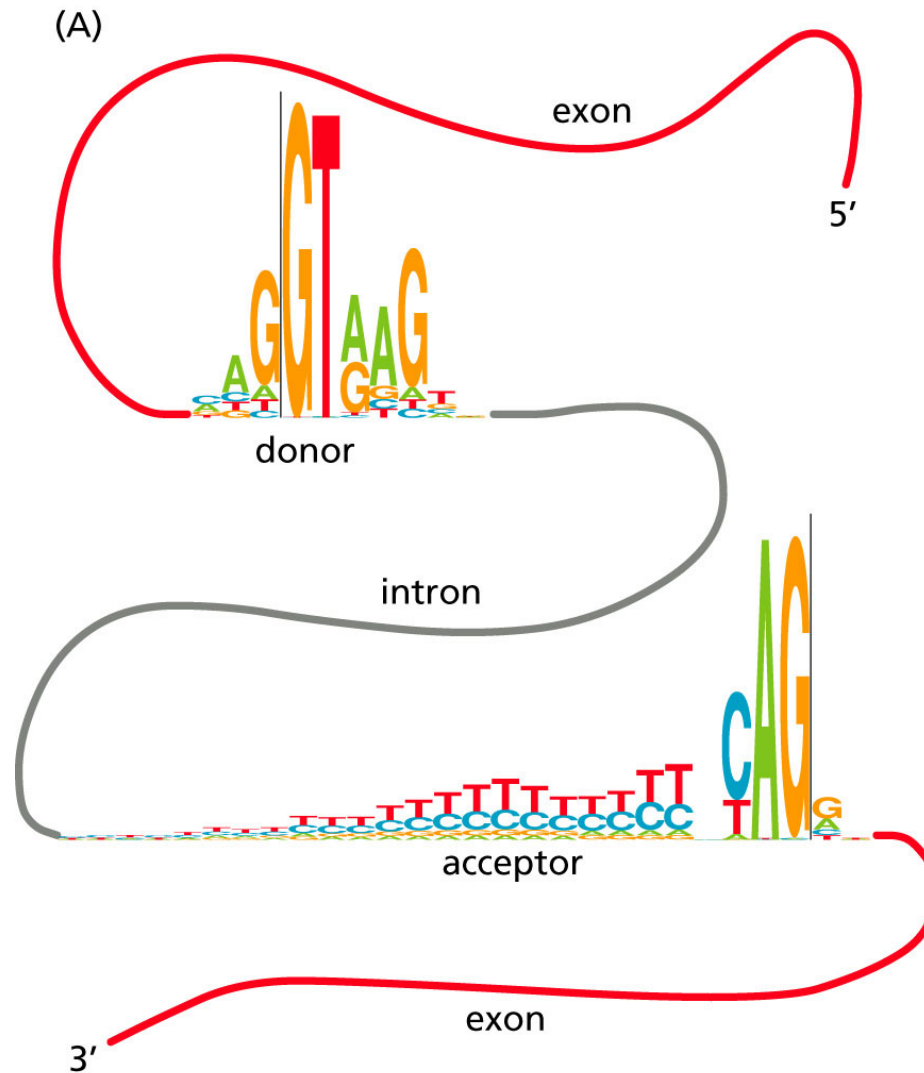
# Gene Prediction in Eukaryotes

- 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
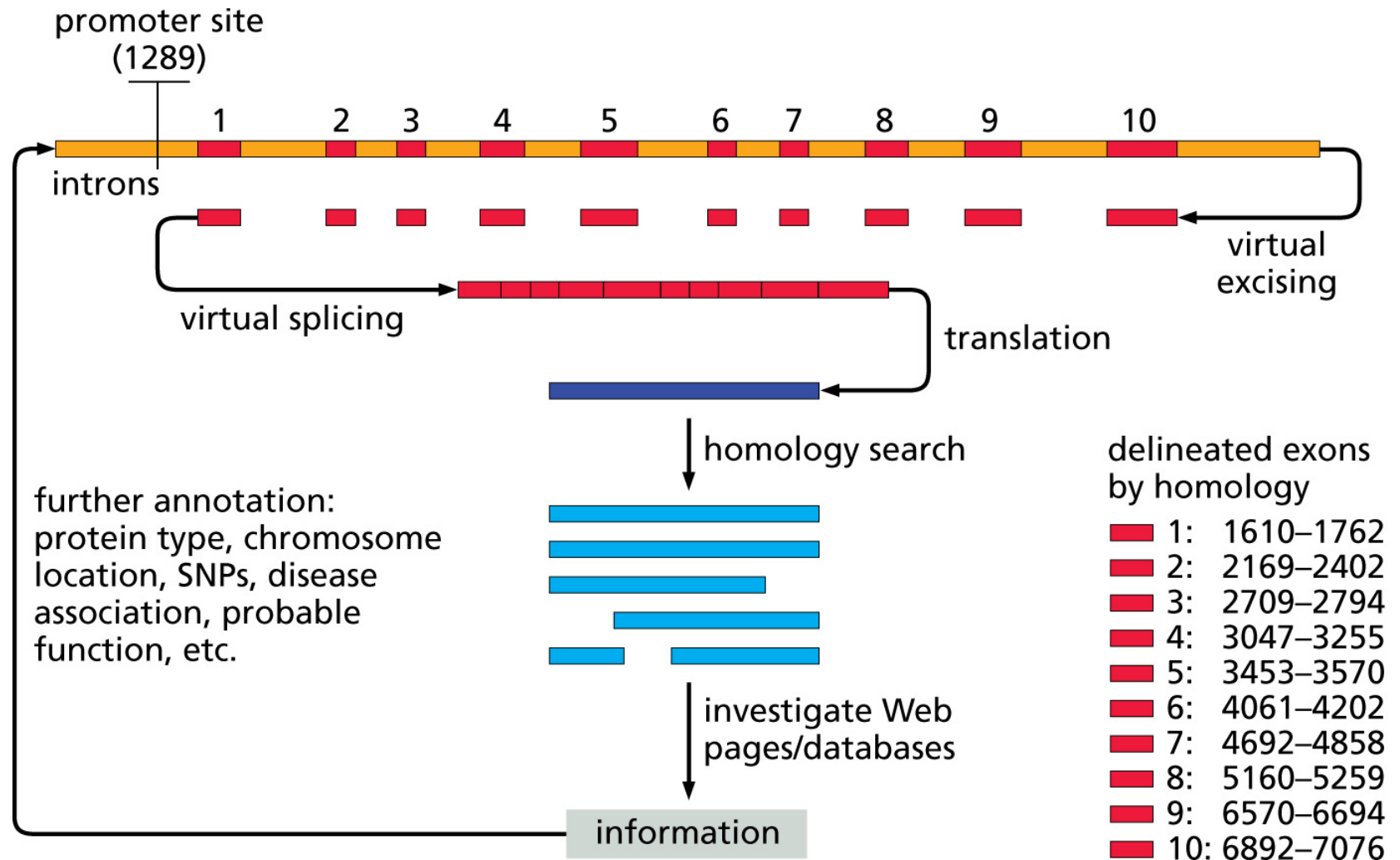
- http://bioinf.uni-greifswald.de/augustus/
- Uses a Hidden Markov Model (HMM)
- Probabilitistic intron length model
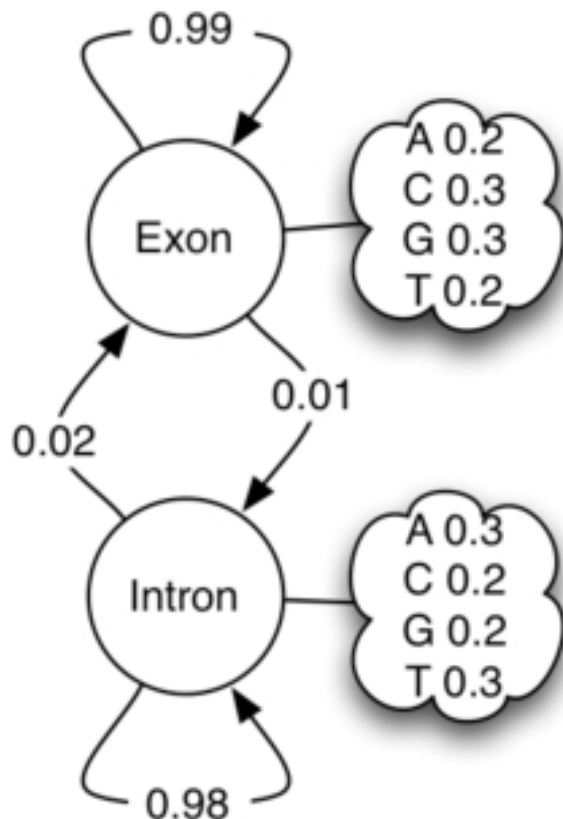
A very simple HMM for gene structure

- 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)?