CSE 427
Computational Biology

Genes and Gene Prediction

Gene Finding: Motivation

Sequence data flooding in

What does it mean?

protein genes, RNA genes, mitochondria, chloroplast, regulation, replication, structure, repeats, transposons, unknown stuff, …

More generally, how do you: learn from complex data in an unknown language, leverage what's known to help discover what's not

Protein Coding Nuclear DNA

Focus of these slides

Goal: Automated annotation of new seq data

State of the Art:

- In Eukaryotes:
  - predictions ~ 60% similar to real proteins
  - ~80% if database similarity used
- Prokaryotes
  - better, but still imperfect
- Lab verification still needed, still expensive
- Largely done for Human; unlikely for most others

Biological Basics

Central Dogma:

DNA $\xrightarrow{\text{transcription}}$ RNA $\xrightarrow{\text{translation}}$ Protein

Codons: 3 bases code one amino acid

- Start codon
- Stop codons
- 3', 5' Untranslated Regions (UTR's)

RNA Transcription

(This gene is heavily transcribed, but many are not.)

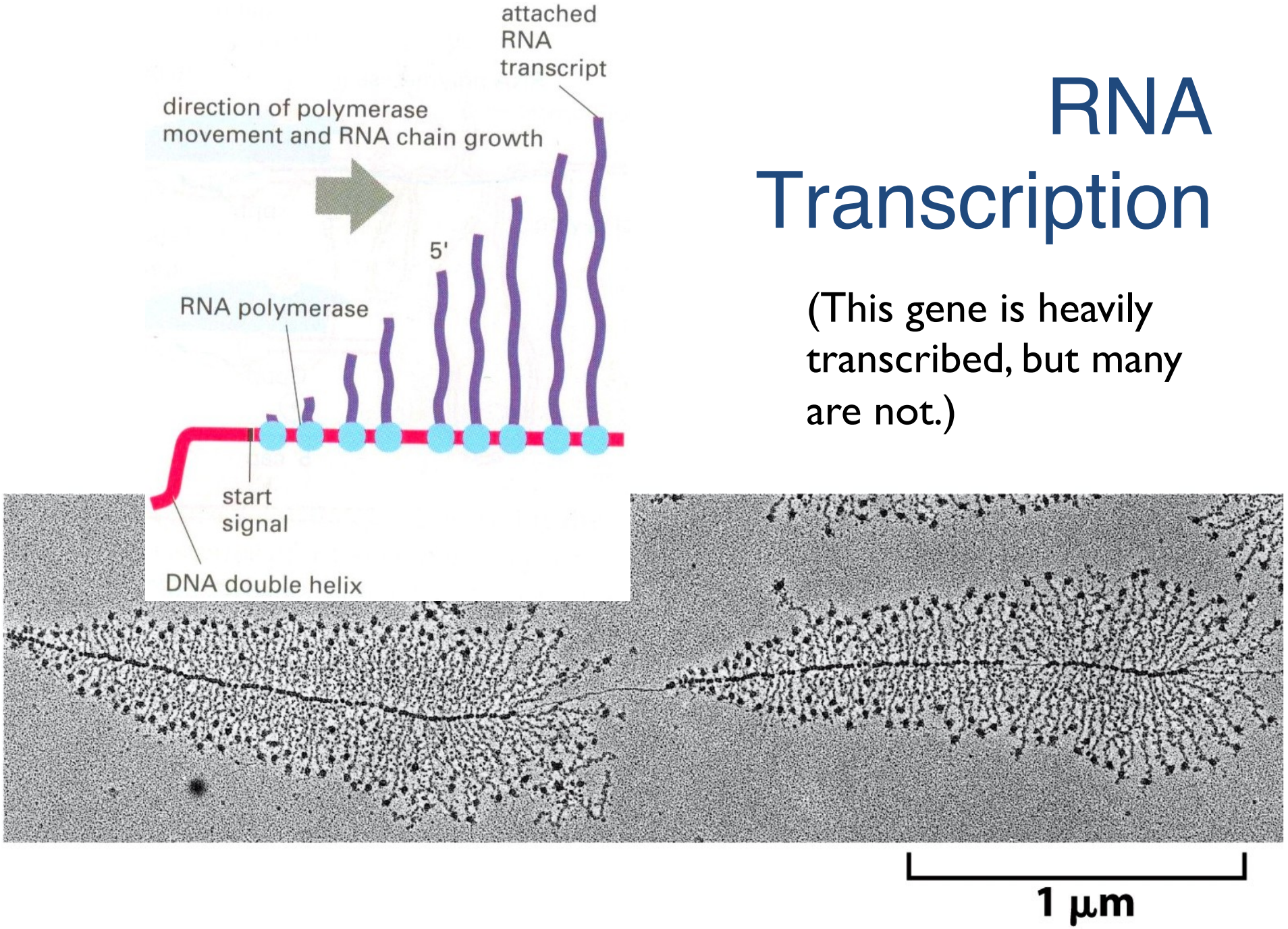
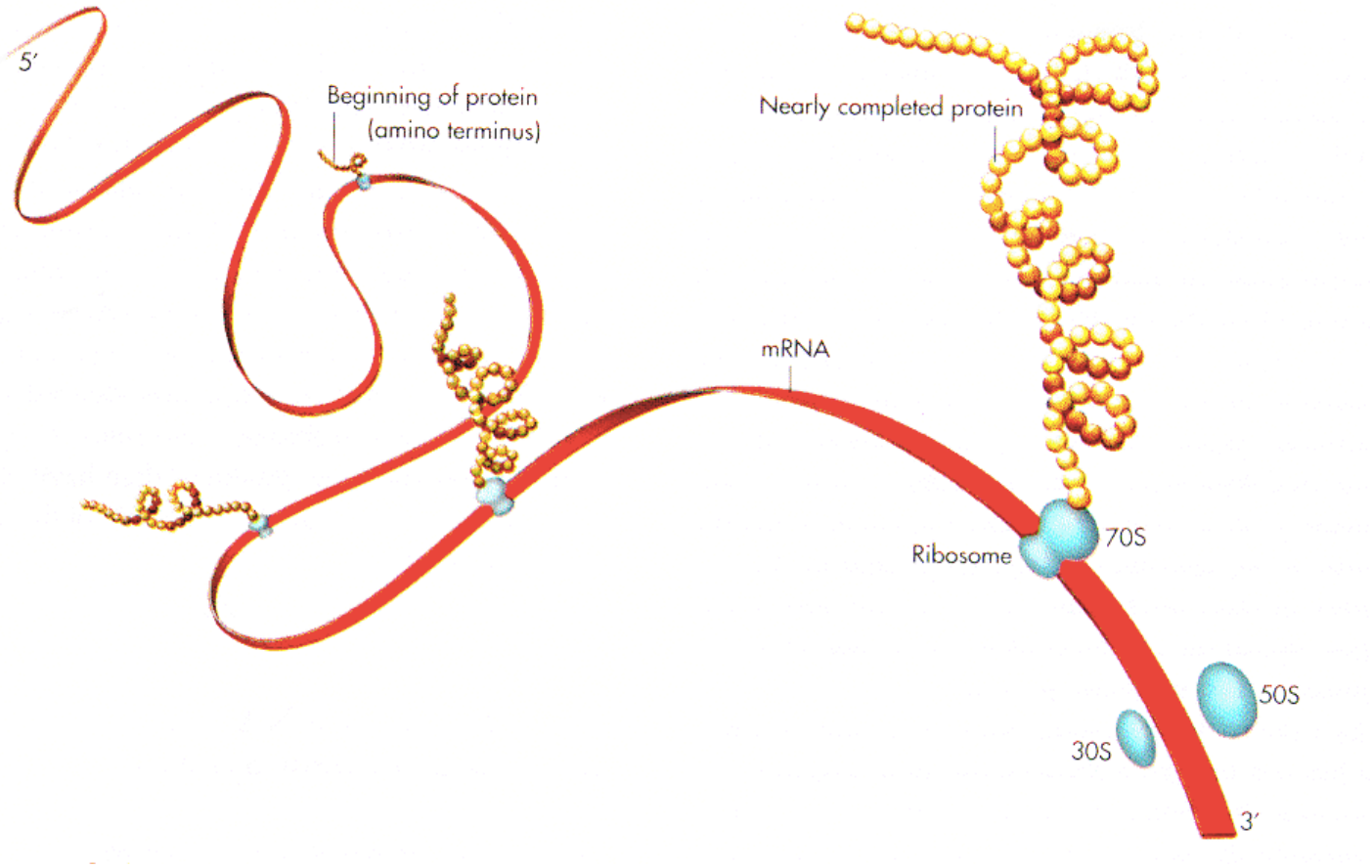


Figure 6-9 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

DNA (thin lines), RNA Pol (Arrow), mRNA with attached Ribosomes (dark circles)

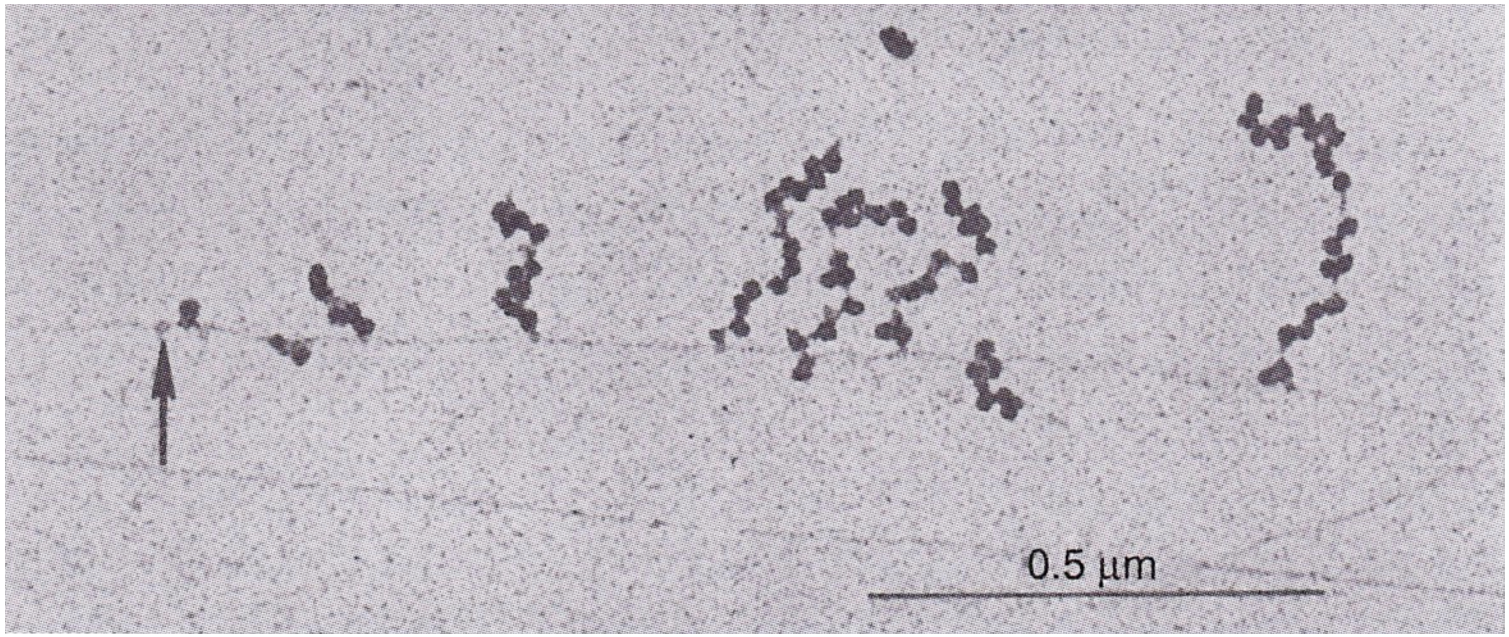
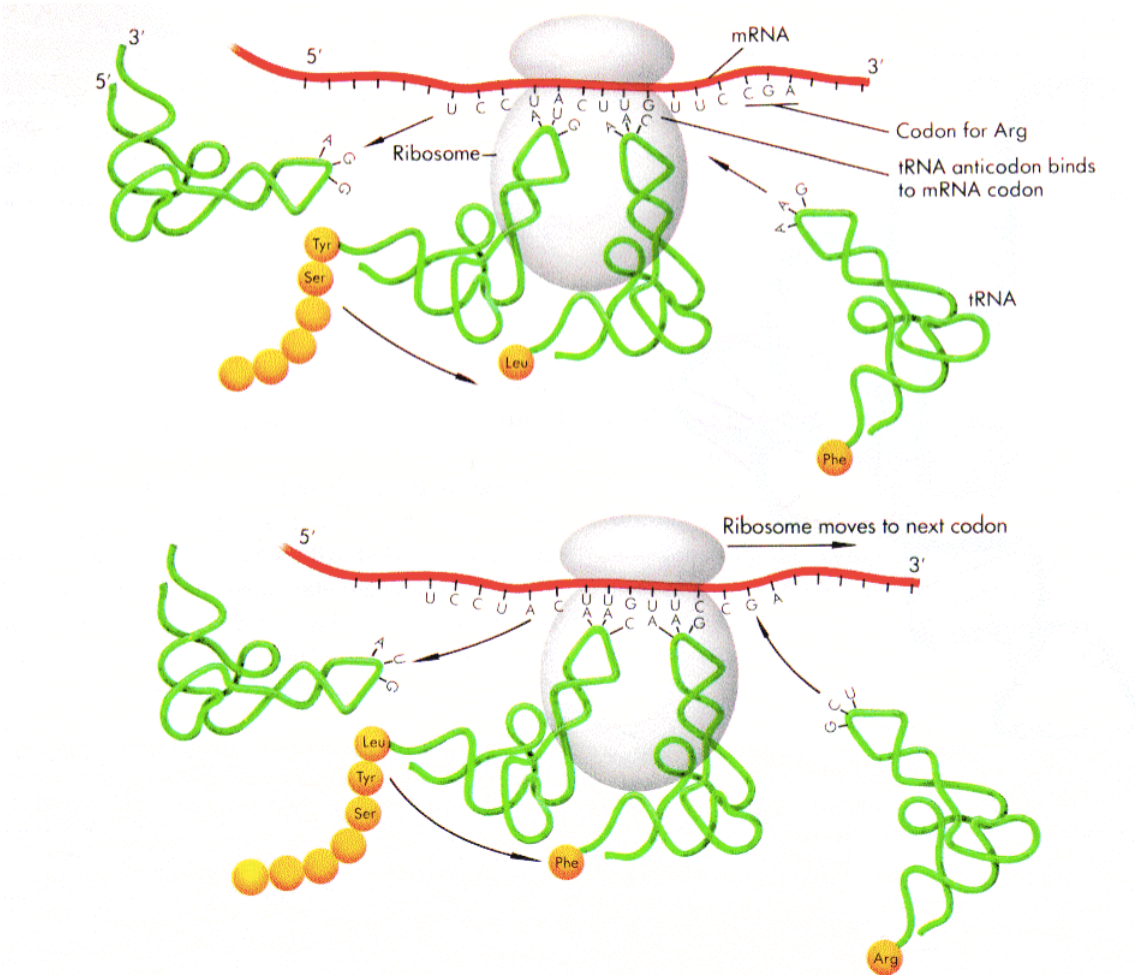


Figure 3-7. Coupled transcription/translation in bacteria is visualized. Oscar Miller and colleagues lysed *E. coli* cells and immediately collected the cell contents on electron microscope grids. They saw threads of mRNA still associated with DNA (thin lines), and ribosomes—several at a time—were already translating protein along the mRNA. Thus, in bacterial cells, the picture of information recovery and use, at least in broad outline, was complete: mRNA was made on demand; ribosomes recognized the 5′ end of the mRNA, bound, and began protein synthesis even before the mRNA had been completely synthesized. (In this photo, the arrow indicates a presumptive RNA polymerase [the faint disk to the left of the first ribosome]. The DNA thread at the top is being copied into mRNA, but the one at the bottom is not. Both are presumably double stranded.) (Reprinted, with permission, from Miller et al. 1970 [©AAAS].)

Darnell, p120

Ribosomes



Watson, Gilman, Witkowski, & Zoller, 1992

Codons & The Genetic Code

First Base	Second Base: U	Second Base: C	Second Base: A	Second Base: G	Third Base
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met/Start	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Ala : Alanine
Arg : Arginine
Asn : Asparagine
Asp : Aspartic acid
Cys : Cysteine
Gln : Glutamine
Glu : Glutamic acid
Gly : Glycine
His : Histidine
Ile : Isoleucine
Leu : Leucine
Lys : Lysine
Met : Methionine
Phe : Phenylalanine
Pro : Proline
Ser : Serine
Thr : Threonine
Trp : Tryptophane
Tyr : Tyrosine
Val : Valine

Idea #1: Find Long ORF's

Reading frame: which of the 3 possible sequences of triples does the ribosome read?

Open Reading Frame: No internal stop codons

In random DNA

- average ORF ~ 64/3 = 21 triplets
- 300bp ORF once per 36kbp per strand

But average protein ~ 1000bp

A Simple ORF finder

start at left end

scan triplet-by-non-overlapping triplet for AUG

then continue scan for STOP

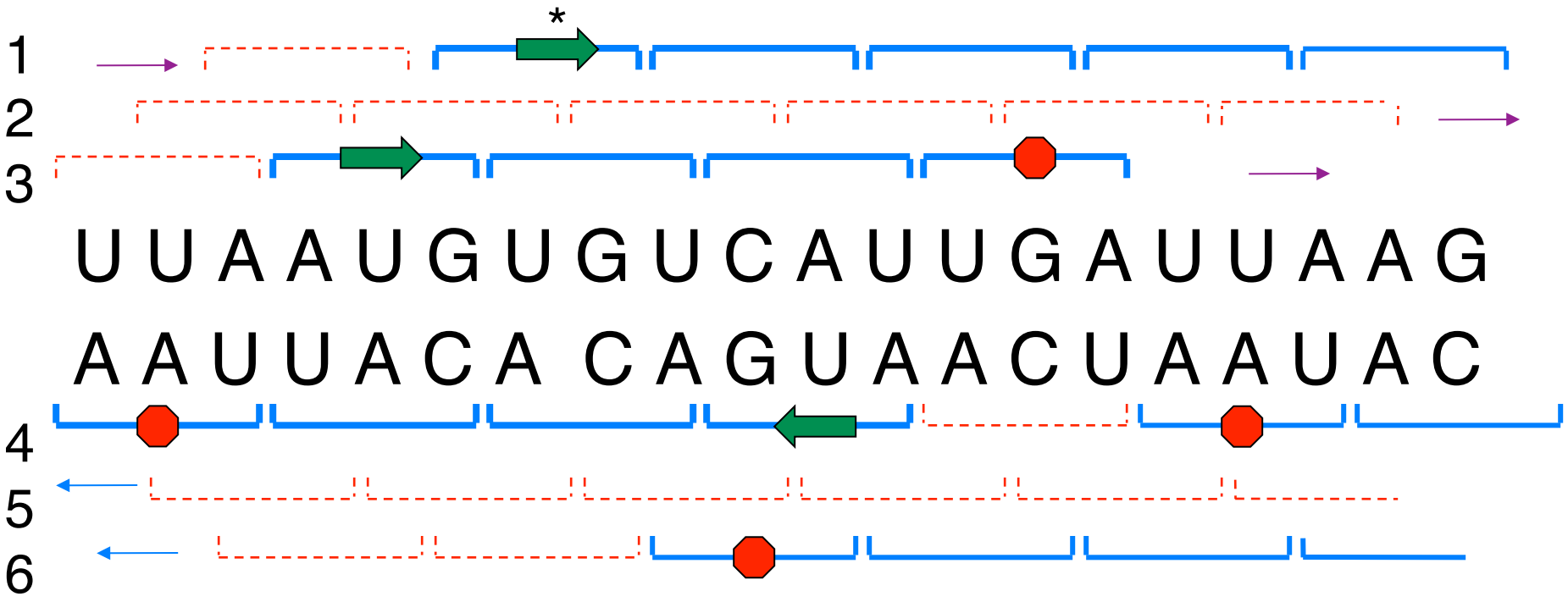
repeat until right end

repeat all starting at offset 1

repeat all starting at offset 2

then do it again on the other strand

Scanning for ORFs



* In bacteria, GUG is sometimes a start codon…

Idea #2: Codon Frequency

In random DNA

Leucine : Alanine : Tryptophan = 6 : 4 : 1

But in real protein, ratios ~ 6.9 : 6.5 : 1

So, coding DNA is not random

Even more: synonym usage is biased (in a species dependant way)

examples known with 90% AT 3rd base

Why? E.g. efficiency, histone, enhancer, splice interactions

Idea #3: Non-Independence

Not only is codon usage biased, but residues (aa or nt) in one position are *not independent* of neighbors

How to model this? Markov models