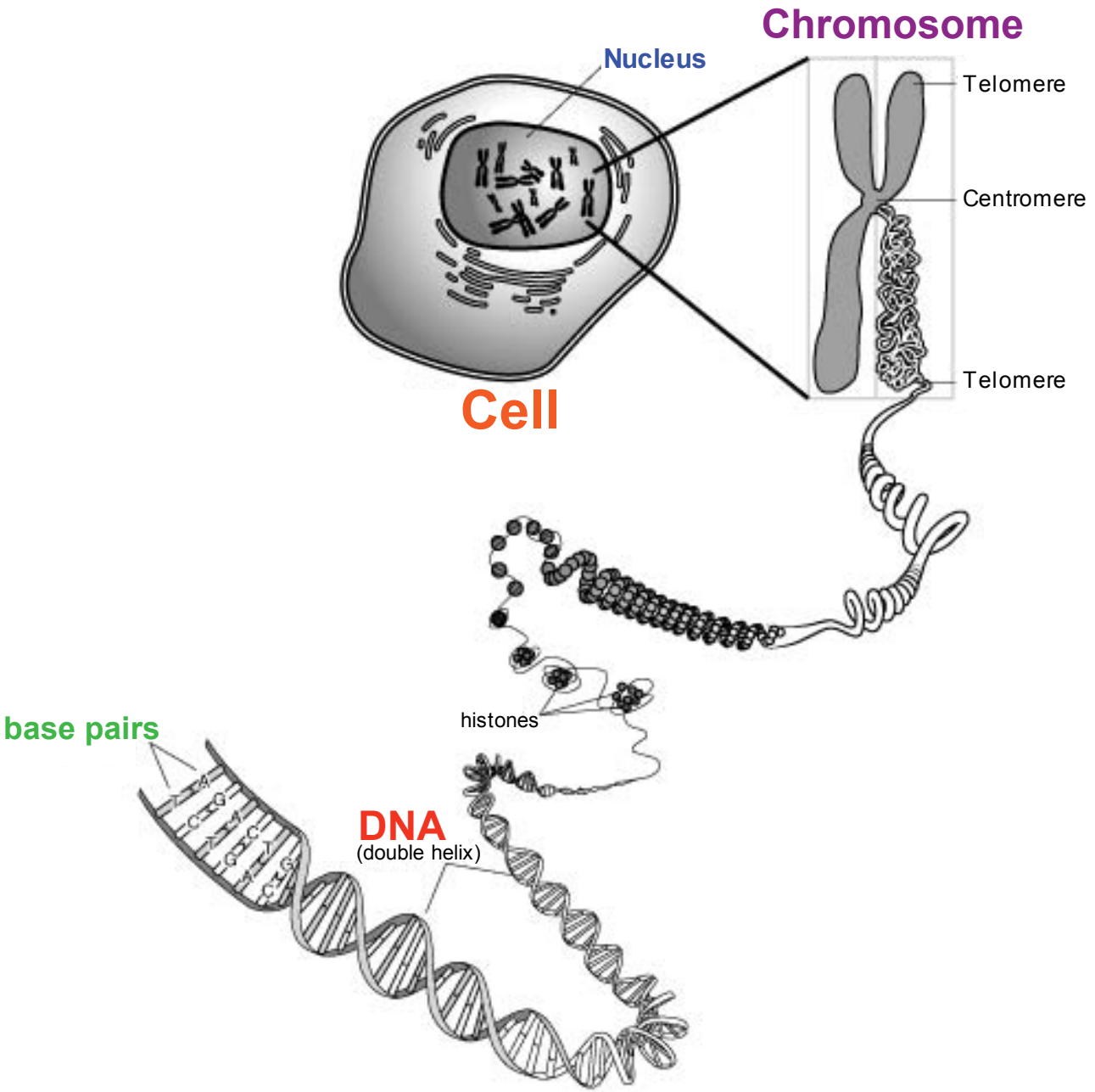
Overview of Eukaryotic Gene Prediction

CBB 231 / COMPSCI 261

W.H. Majoros

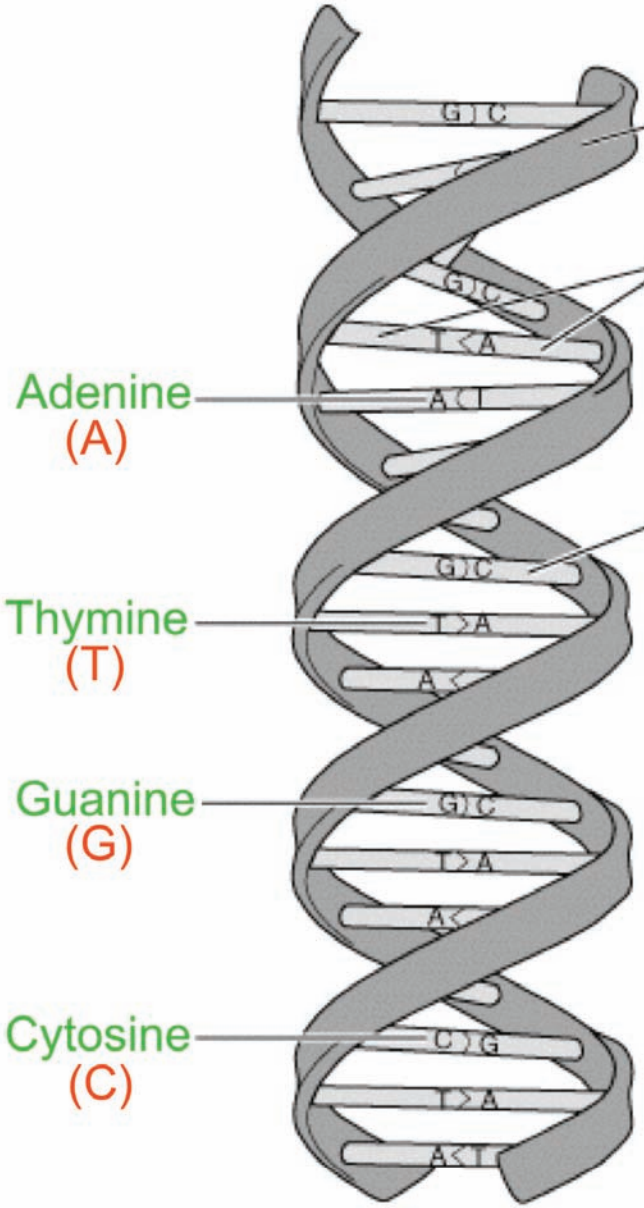
Duke
UNIVERSITY

What is DNA?



Duke UNIVERSITY

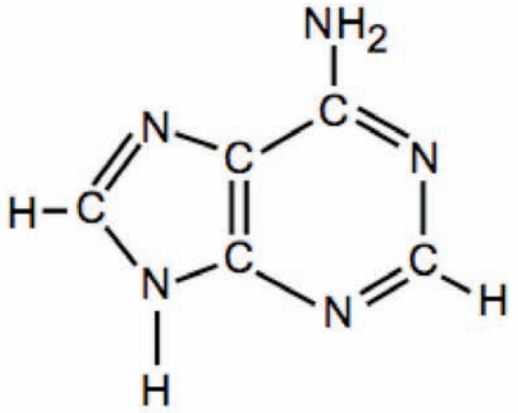
DNA is a Double Helix



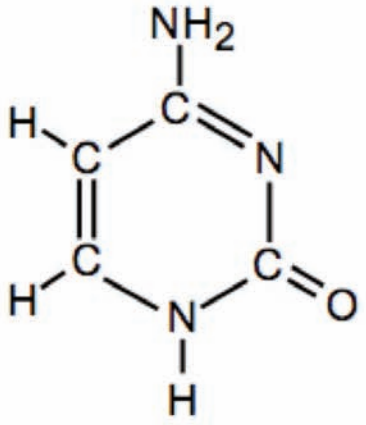
What is DNA?

- DNA is the main repository of hereditary information
- Every cell contains a copy of the genome encoded in DNA
- Each chromosome is a single DNA molecule
- A DNA molecule may consist of an arbitrary sequence of *nucleotides*
- The discrete nature of DNA allows us to treat it as a sequence of A's, C's, G's, and T's
- DNA is replicated during cell division
- Only mutations on the *germ line* may lead to evolutionary changes

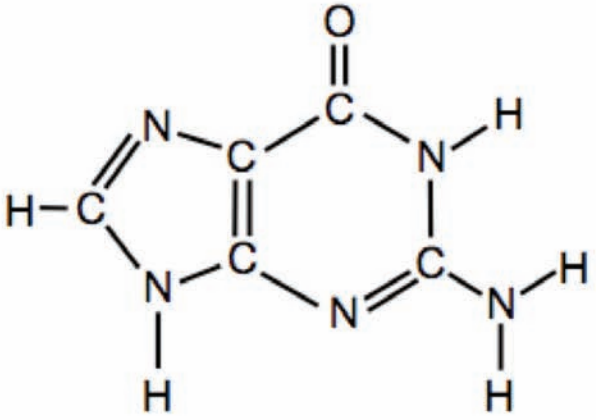
Molecular Structure of Nucleotides



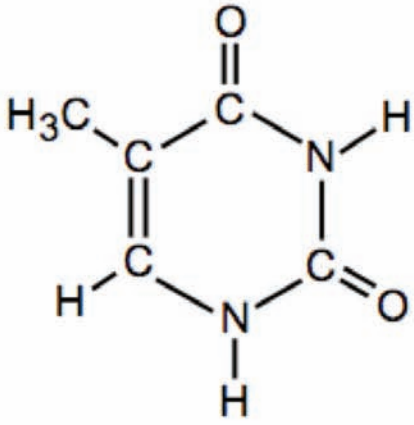
adenine



cytosine



guanine



thymine

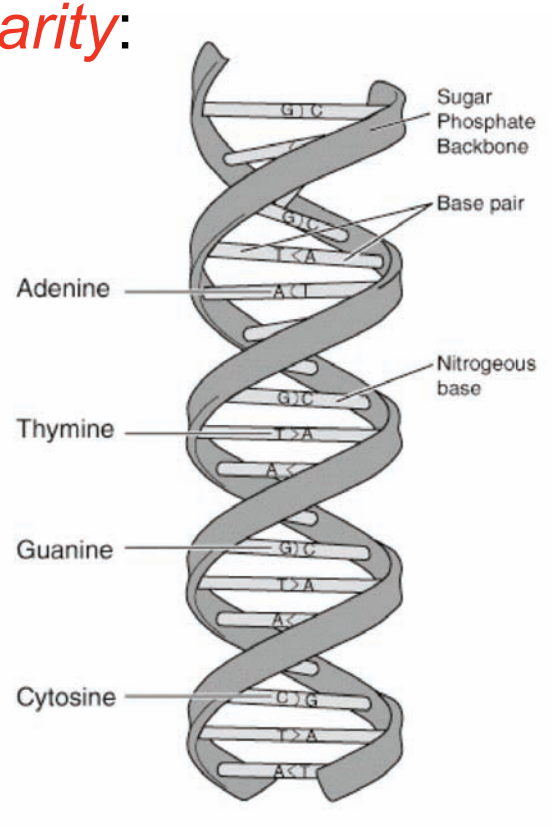
Duke UNIVERSITY

Base Complementarity

Nucleotides on opposite strands of the double helix pair off in a strict pattern called *Watson-Crick complementarity*:

A only pairs with T

C only pairs with G



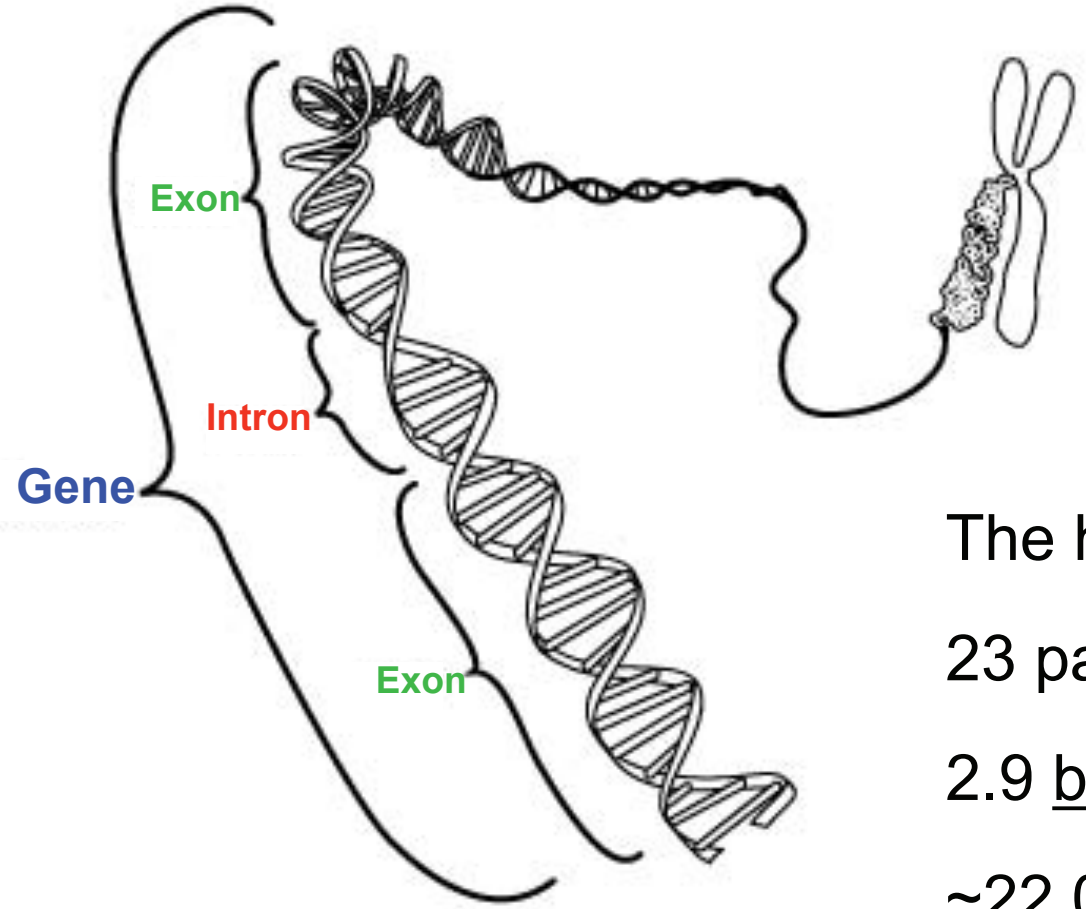
The A-T pairing involves *two* hydrogen bonds, whereas the G-C pairing involves *three* hydrogen bonds.

In RNA one can sometimes find G-T (actually, G-U) pairings, which involve only *one* H-bond.

Note that the bonds forming the “rungs” of the DNA “ladder” are the *hydrogen bonds*, whereas the bonds connecting successive nucleotides along each helix are *phosphodiester bonds*.

Duke UNIVERSITY

Exons, Introns, and Genes



The human genome:

23 pairs of chromosomes

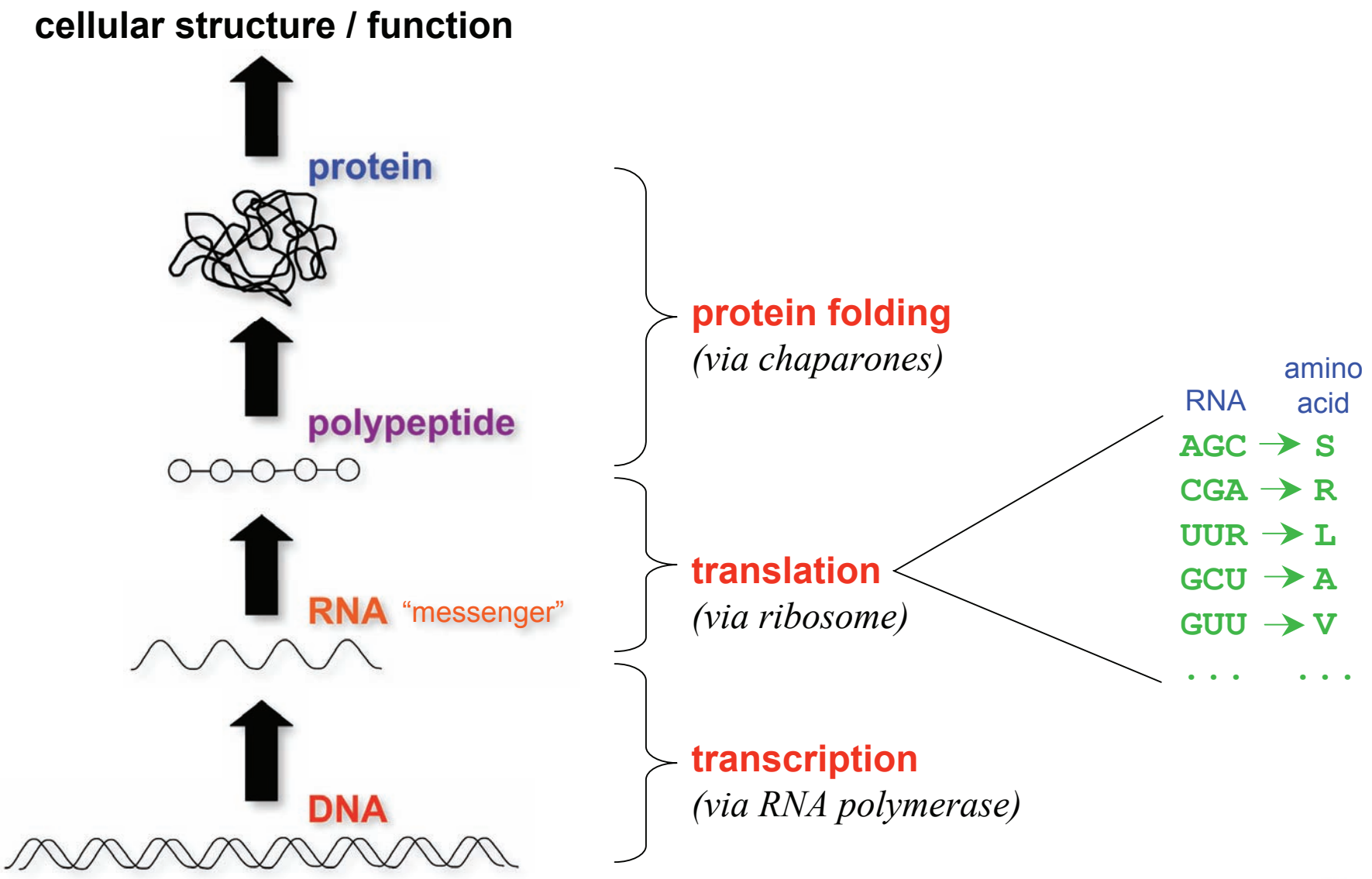
2.9 billion A's, T's, C's, G's

~22,000 genes (?)

~1.4% of genome is coding

Duke UNIVERSITY

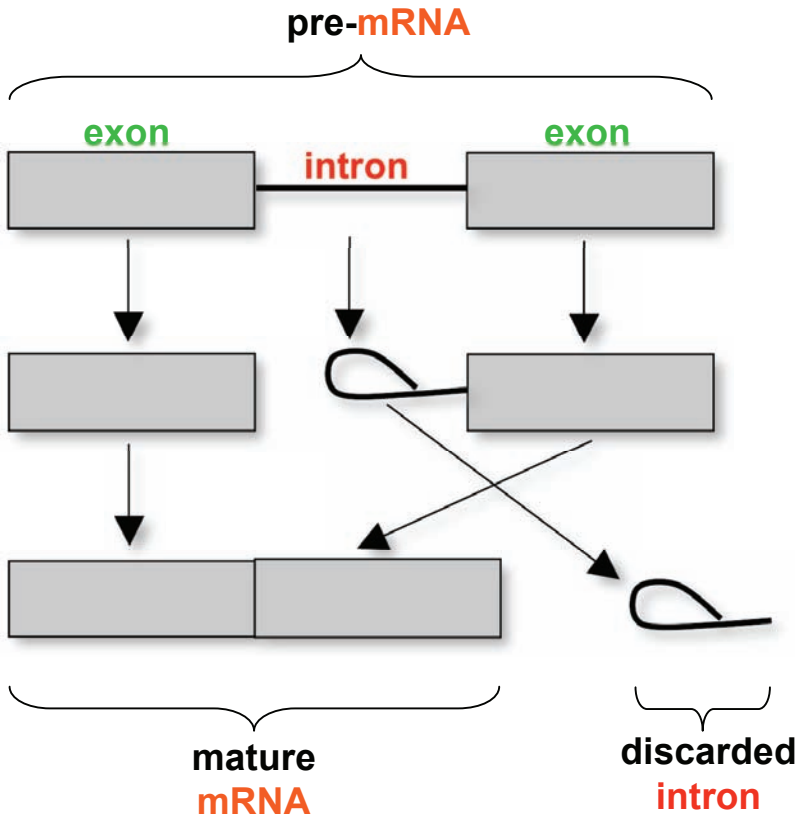
The Central Dogma



Duke UNIVERSITY

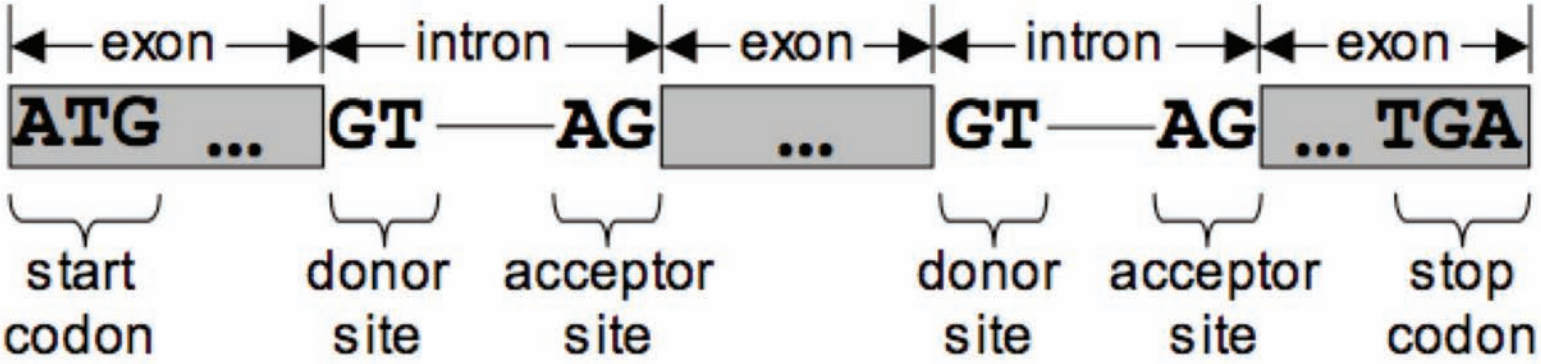
Splicing of Eukaryotic mRNA's

After transcription by the *polymerase*, eukaryotic pre-mRNA's are subject to splicing by the *spliceosome*, which removes introns:



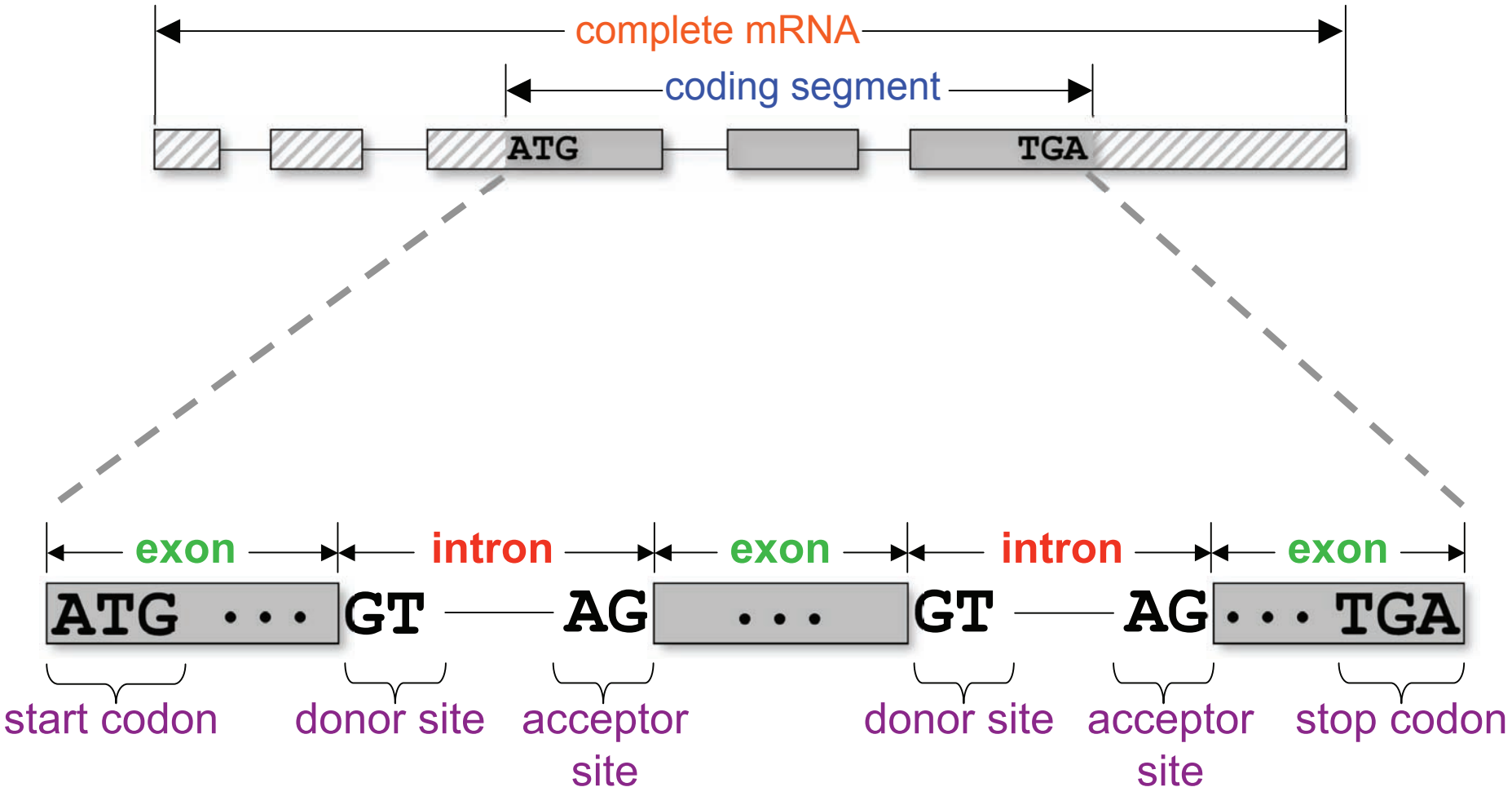
Signals Delimit Gene Features

Coding segments (*CDS*'s) of genes are delimited by four types of signals: *start codons* (ATG in eukaryotes), *stop codons* (usually TAG, TGA, or TAA), *donor sites* (usually GT), and *acceptor sites* (AG):



For initial and final exons, only the coding portion of the exon is generally considered in most of the gene-finding literature; thus, we redefine the word "exon" to include only the coding portions of exons, for convenience.

Eukaryotic Gene Syntax



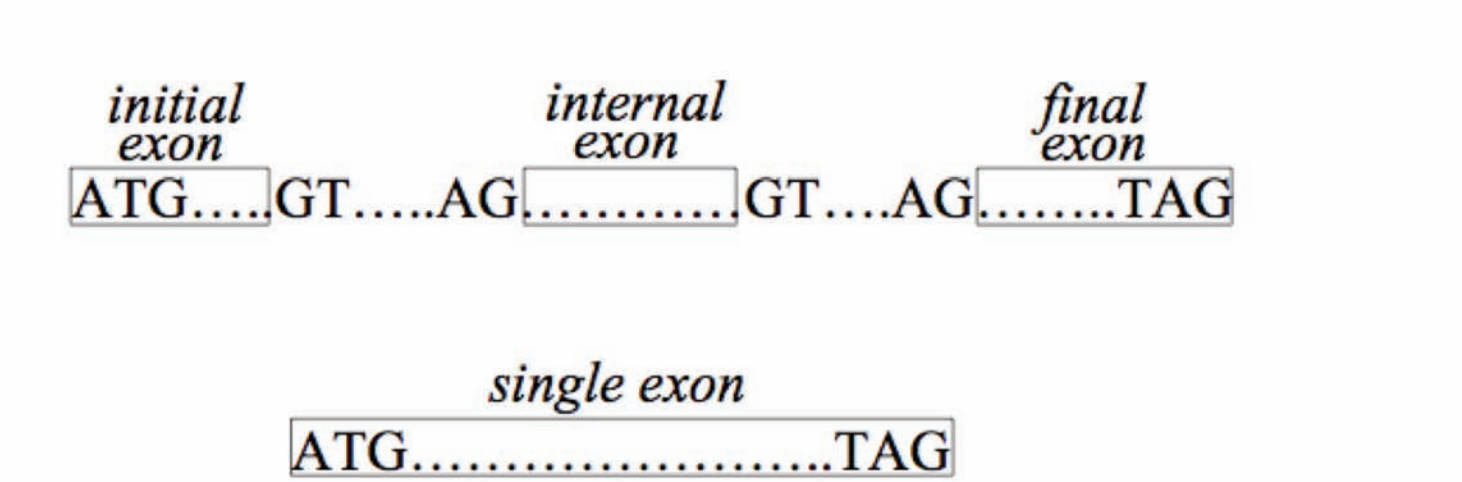
Regions of the gene outside of the CDS are called *UTR*'s (*untranslated regions*), and are mostly ignored by gene finders, though they are important for regulatory functions.

Duke UNIVERSITY

Types of Exons

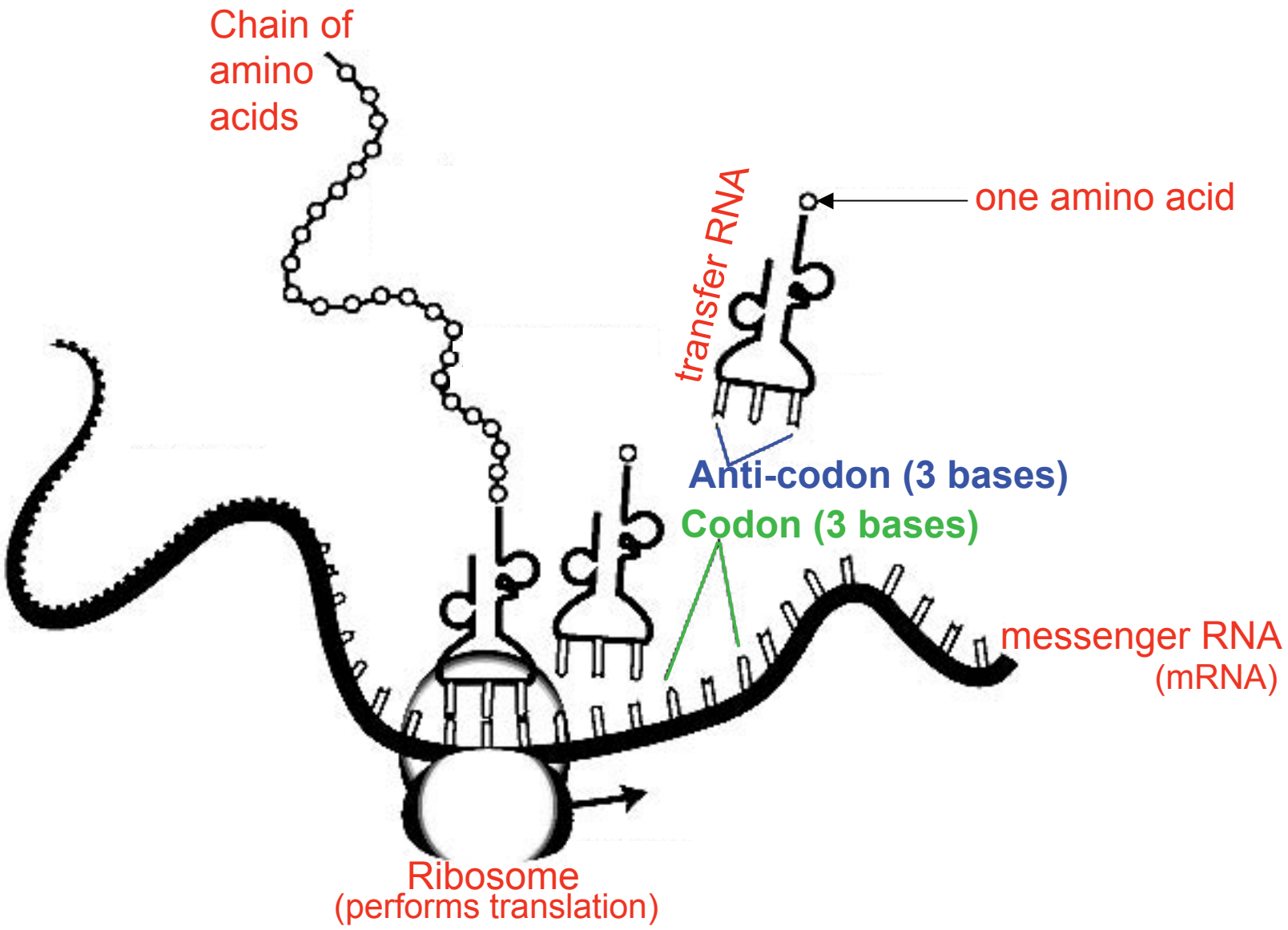
Three types of exons are defined, for convenience:

- *initial exons* extend from a start codon to the first donor site;
- *internal exons* extend from one acceptor site to the next donor site;
- *final exons* extend from the last acceptor site to the stop codon;
- *single exons* (which occur only in *intronless genes*) extend from the start codon to the stop codon:



Duke UNIVERSITY

Translation



Degenerate Nature of the Genetic Code

acid	*codons*	*acid*	*codons*	*acid*	*codons*	*acid*	*codons*
A	GCA GCC GCG GCT	G	GGA GGC GGG GGT	M	ATG	S	AGC AGT TCA TCC TCG TCT
C	TGC TGT	H	CAC CAT	N	AAC AAT	T	ACA ACC ACG ACT
D	GAC GAT	I	ATA ATC ATT	P	CCA CCC CCG CCT	V	GTA GTC GTG GTT
E	GAA GAG TTC TTT	K	AAA AAG	Q	CAA CAG	W	TGG
F	TTC TTT	L	CTA CTC CTG CTT TTA	R	AGA AGG CGA CGC CGG CGT	Y	TAC TAT

Each amino acid is encoded by one or more *codons*.

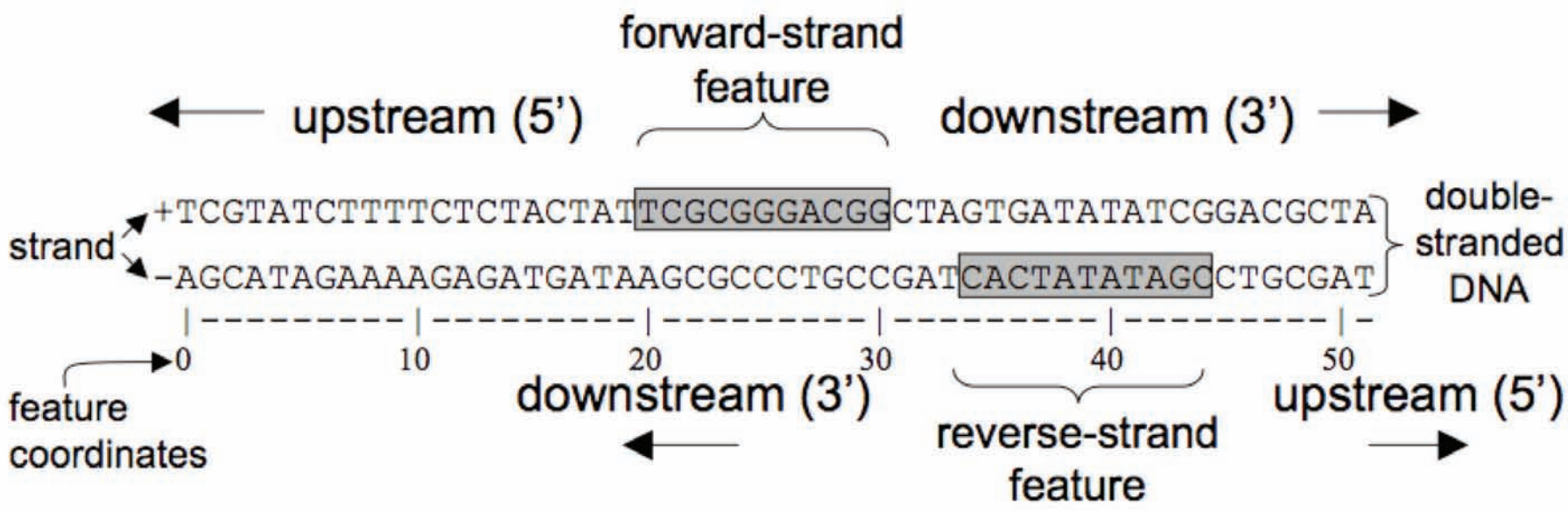
Each codon encodes a single *amino acid*.

The *third position* of the codon is the most likely to vary, for a given amino acid.

Duke UNIVERSITY

Orientation

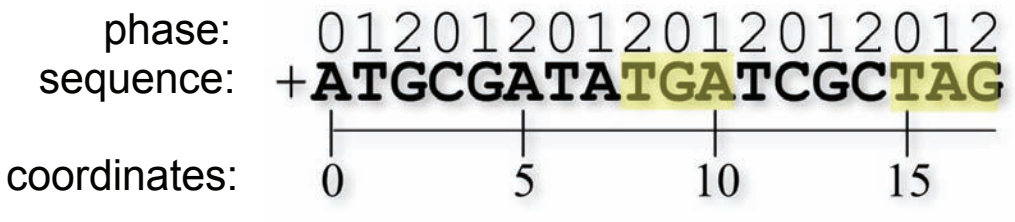
DNA replication occurs in the *5'-to-3'* direction; this gives us a natural frame of reference for defining orientation and direction relative to a DNA sequence:



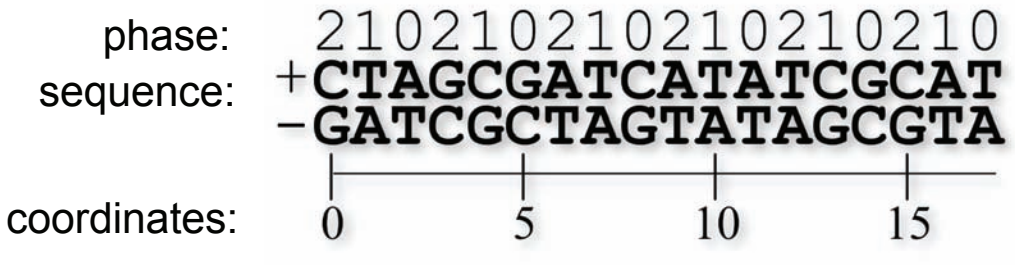
The input sequence to a gene finder is always assumed to be the forward strand. Note that genes can occur on either strand, but we can model them relative to the forward-strand sequence.

Duke UNIVERSITY

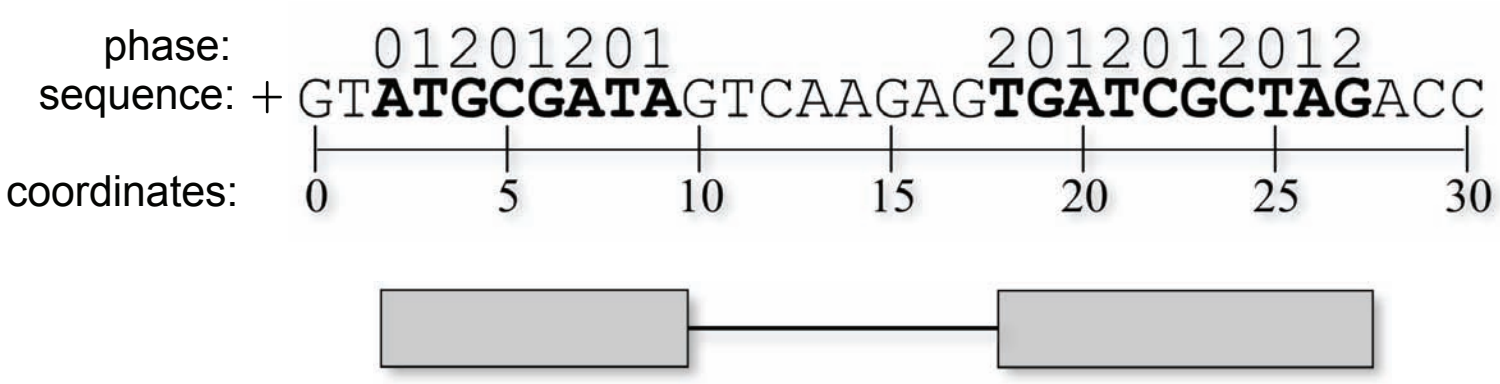
The Notion of Phase



forward strand



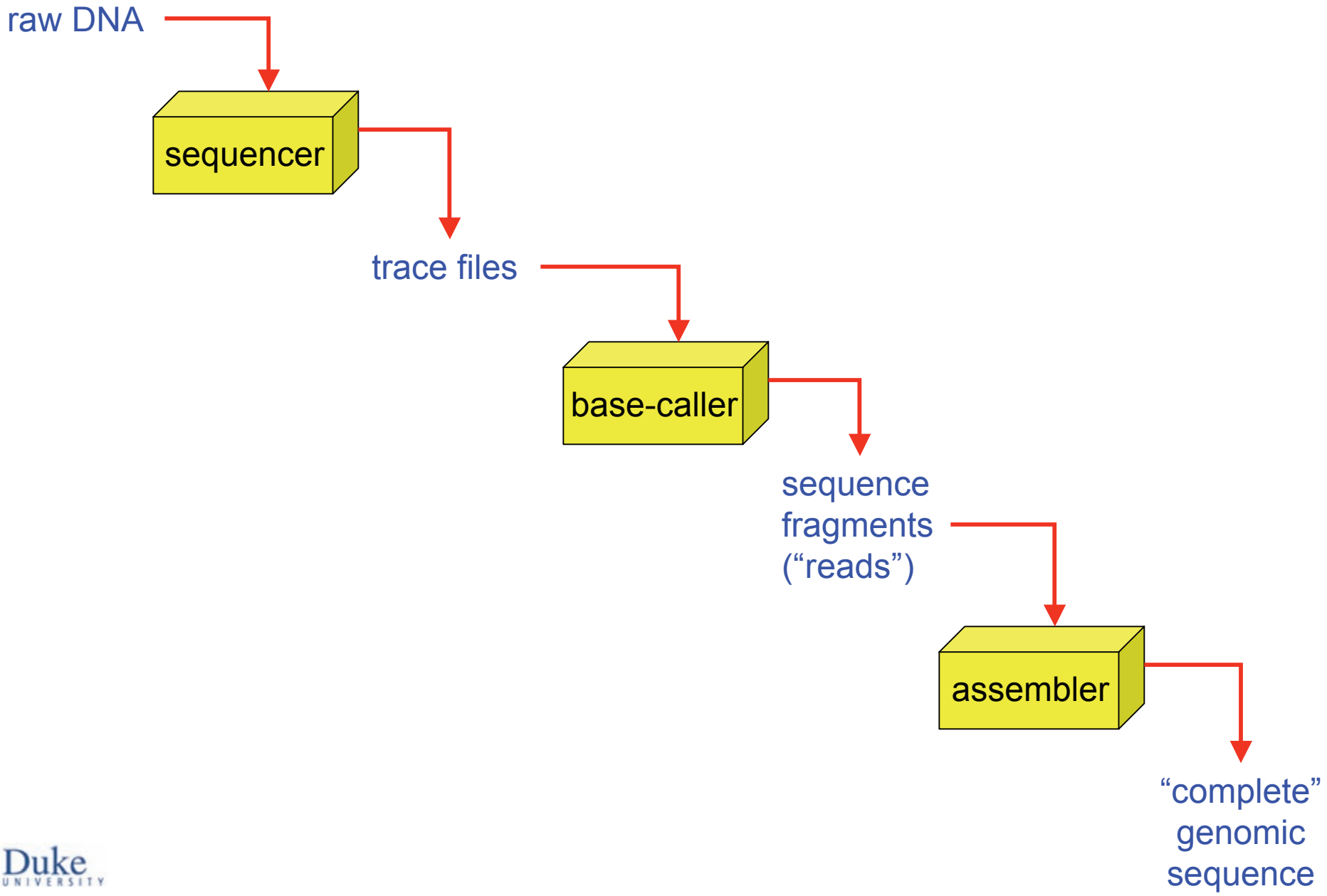
reverse strand



forward strand, spliced

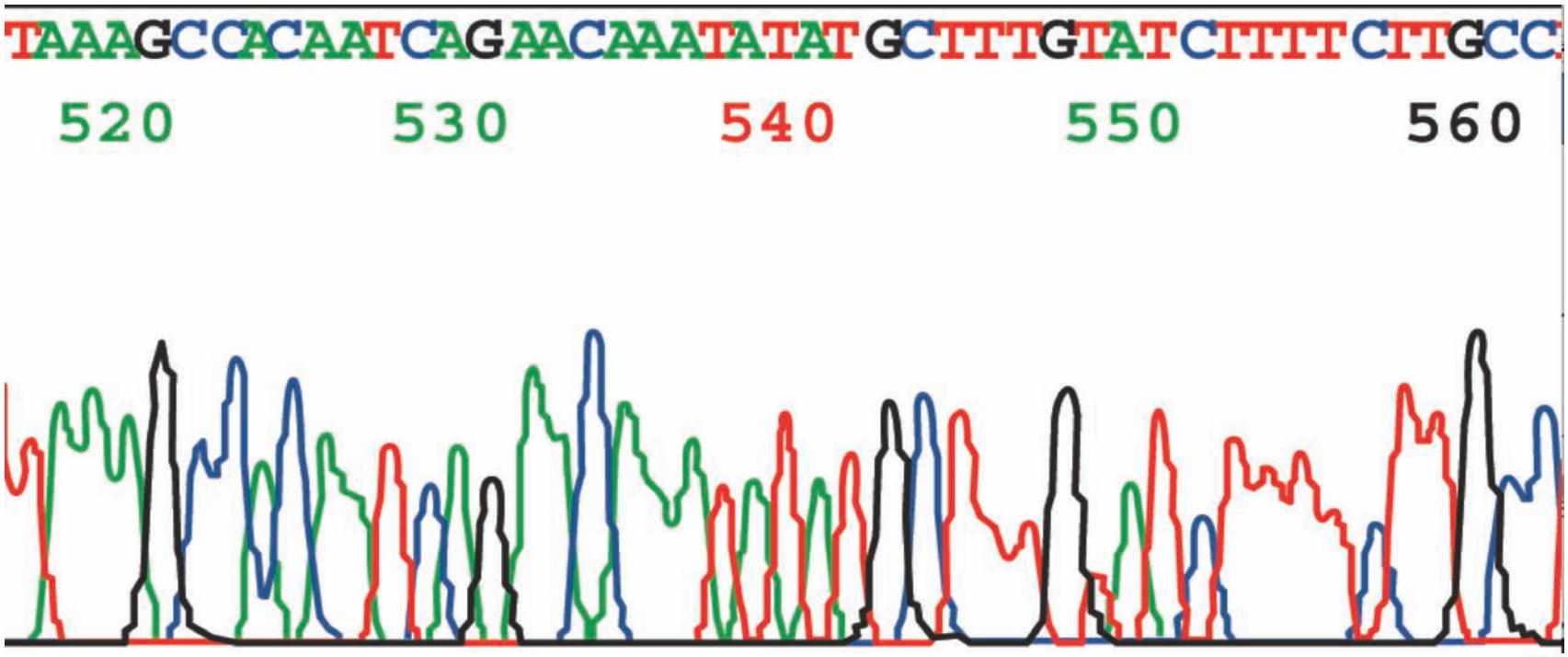
Duke UNIVERSITY

Sequencing and Assembly



DNA Sequencing

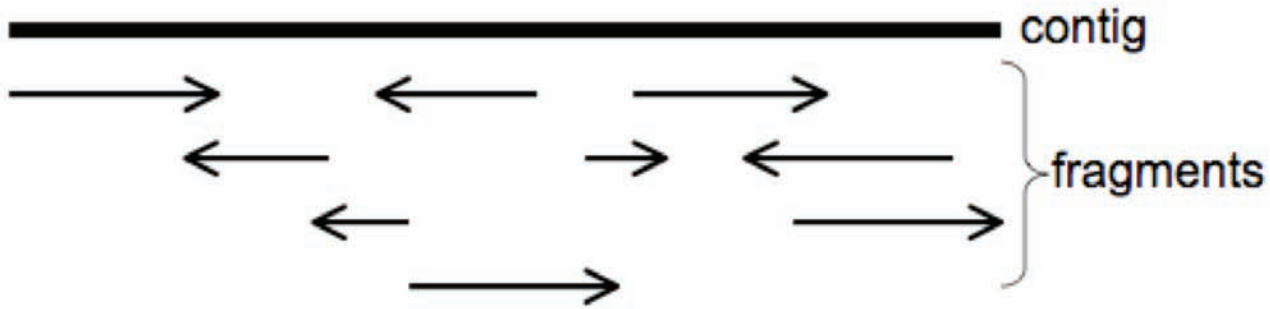
Nucleotides are induced to fluoresce in one of four colors when struck by a laser beam in the sequencer. A sensor in the sequencing machine records the levels of fluoresence onto a *trace diagram* (shown below). A program called a *base caller* infers the most likely nucleotide at each position, based on the peaks in the trace diagram:



Duke UNIVERSITY

Genome Assembly

Fragments emitted by the sequencer are assembled into *contigs* by a program called an *assembler*:

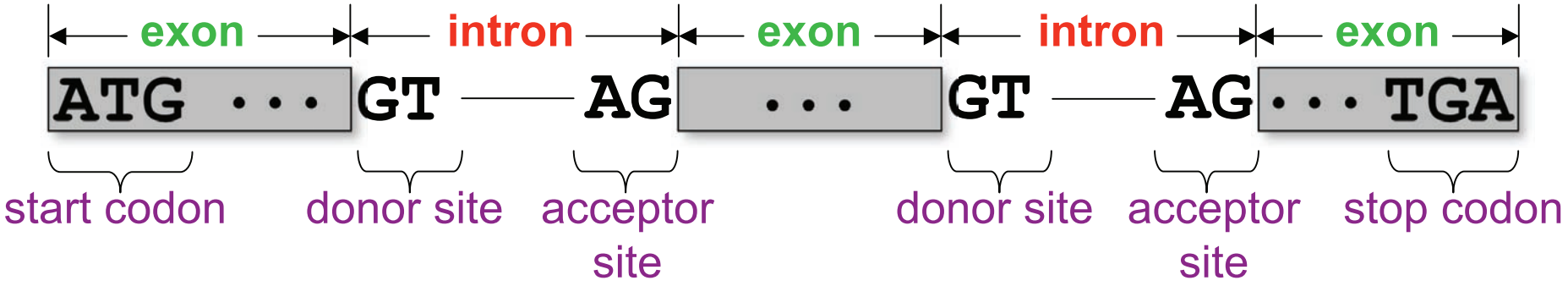


Each fragment has a *clear range* (not shown) in which the sequence is assumed of highest quality. Contigs can be ordered and oriented by *mate-pairs*:

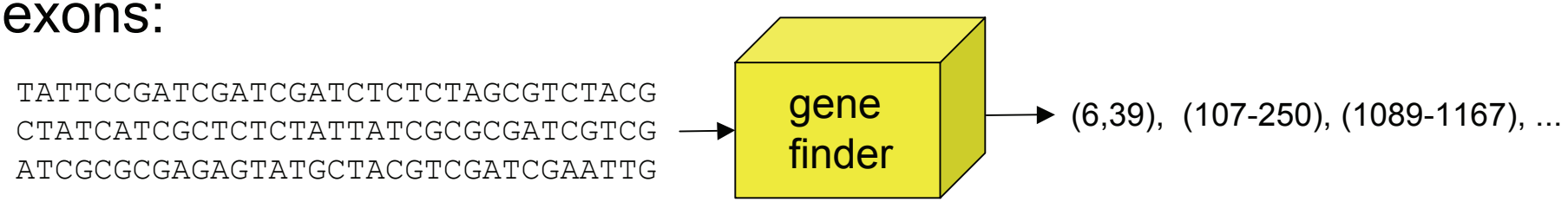
Mate-pairs occur because the sequencer reads from both ends of each fragment. The part of the fragment which is actually sequenced is called the *read*.

Gene Prediction as Parsing

The problem of eukaryotic gene prediction entails the identification of putative exons in unannotated DNA sequence:



This can be formalized as a process of identifying intervals in an input sequence, where the intervals represent putative coding exons:

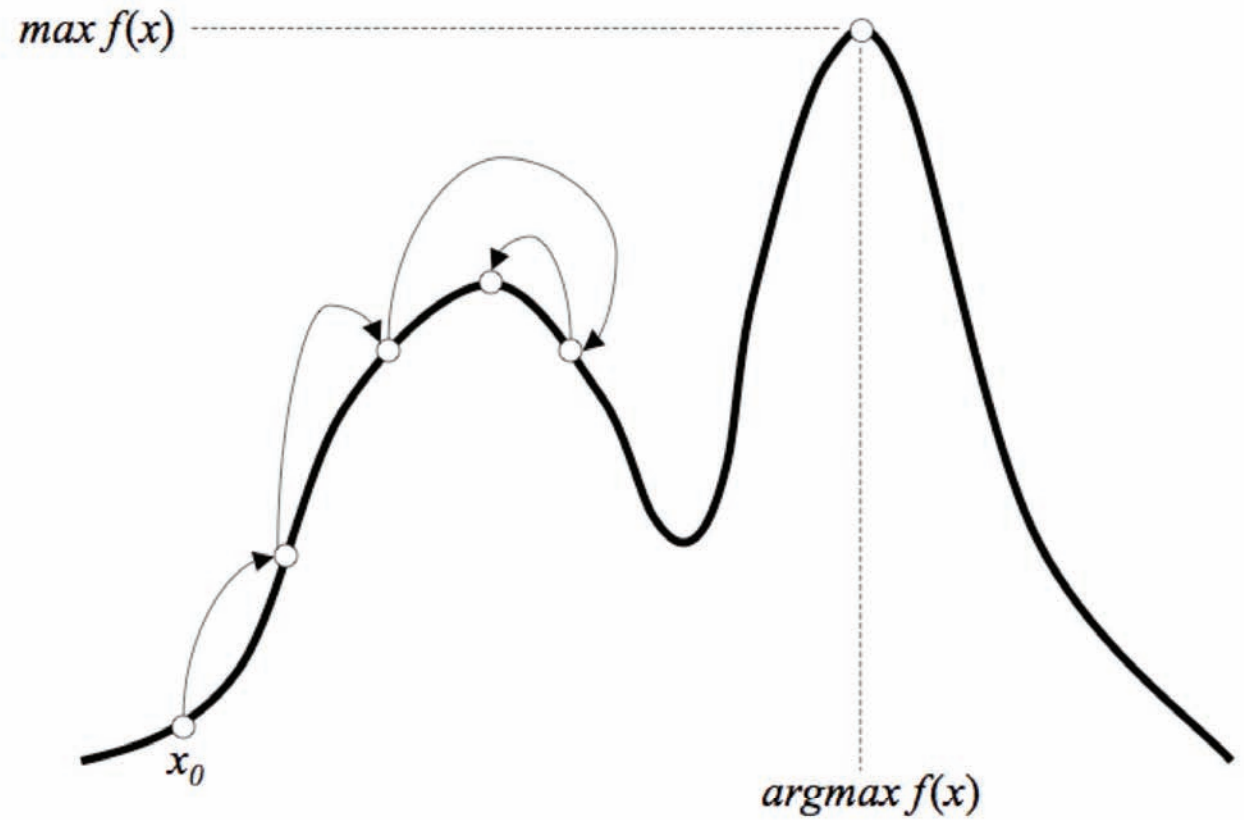


These putative exons will generally have associated scores.

Duke UNIVERSITY

The Notion of an Optimal Gene Structure

If we could enumerate all putative gene structures along the x-axis and graph their scores according to some function f(x), then the highest-scoring parse would be denoted argmax f(x), and its score would be denoted max f(x). A gene finder will often find the *local maximum* rather than the *global maximum*.



Eukaryotic Gene Syntax Rules

The syntax of eukaryotic genes can be represented via series of *signals* (ATG=start codon; TAG=any of the three stop codons; GT=donor splice site; AG=acceptor splice site). Gene *syntax rules* (for forward-strand genes) can then be stated very compactly:

$$ATG \rightarrow TAG$$
$$ATG \rightarrow GT$$
$$GT \rightarrow AG$$
$$AG \rightarrow GT$$
$$AG \rightarrow TAG$$
$$TAG \rightarrow ATG$$

For example, a feature beginning with a start codon (denoted ATG) may end with either a TAG (any of the three stop codons) or a GT (donor site), denoting either a single exon or an initial exon.

The Stochastic Nature of Signal Motifs

(start codons)

A T G

(stop codons)

T G A
T A A
T A G

(donor splice sites)

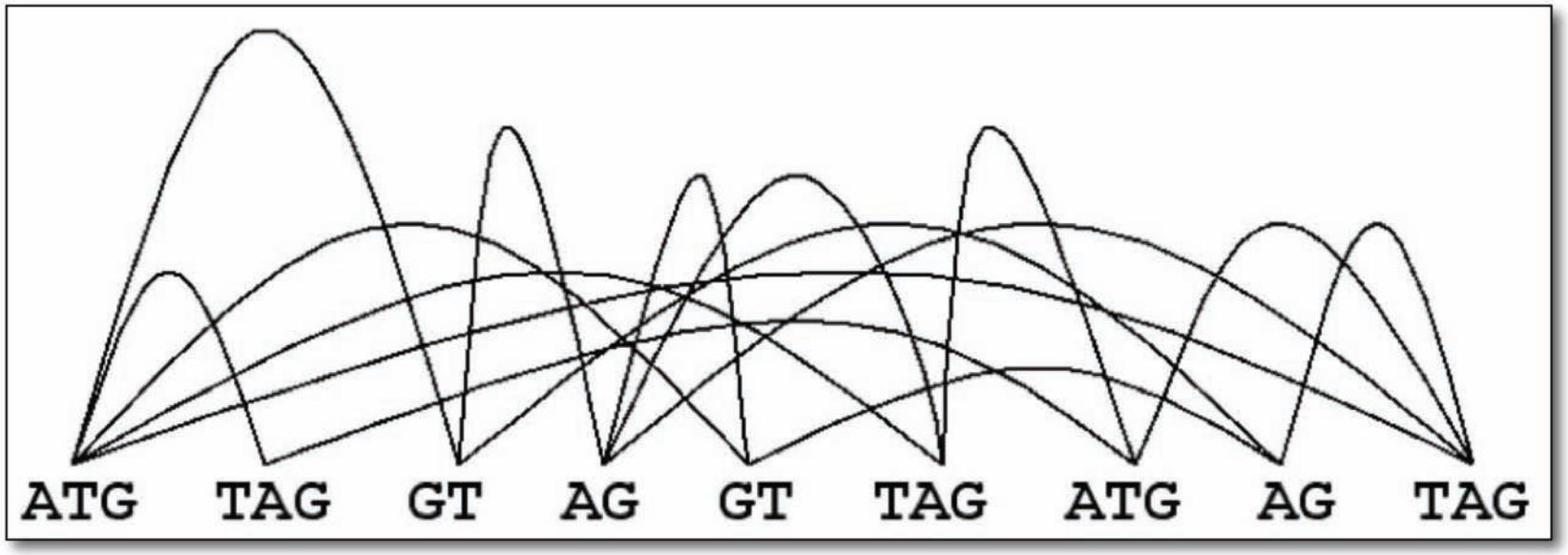
G T

(acceptor splice sites)

A G

Representing Gene Syntax with ORF Graphs

After identifying the most promising (i.e., highest-scoring) signals in an input sequence, we can apply the gene syntax rules to connect these into an *ORF graph*:

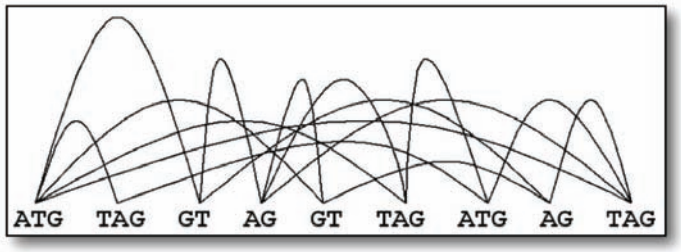


An ORF graph represents all possible *gene parses* (and their scores) for a given set of putative signals. A *path* through the graph represents a single gene parse.

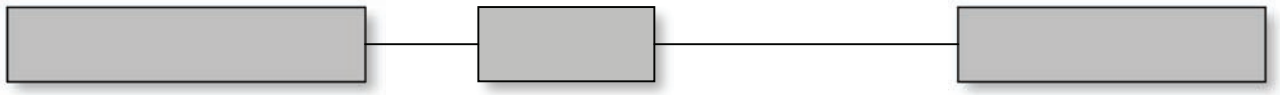
Conceptual Gene-finding Framework

```
TATTCCGATCGATCGATCTCTCTAGCGTCTACG
CTATCATCGCTCTCTATTATCGCGCGATCGTCG
ATCGCGCGAGAGTATGCTACGTCGATCGAATTG
```

identify most promising signals, score signals and content regions between them; induce an ORF graph on the signals

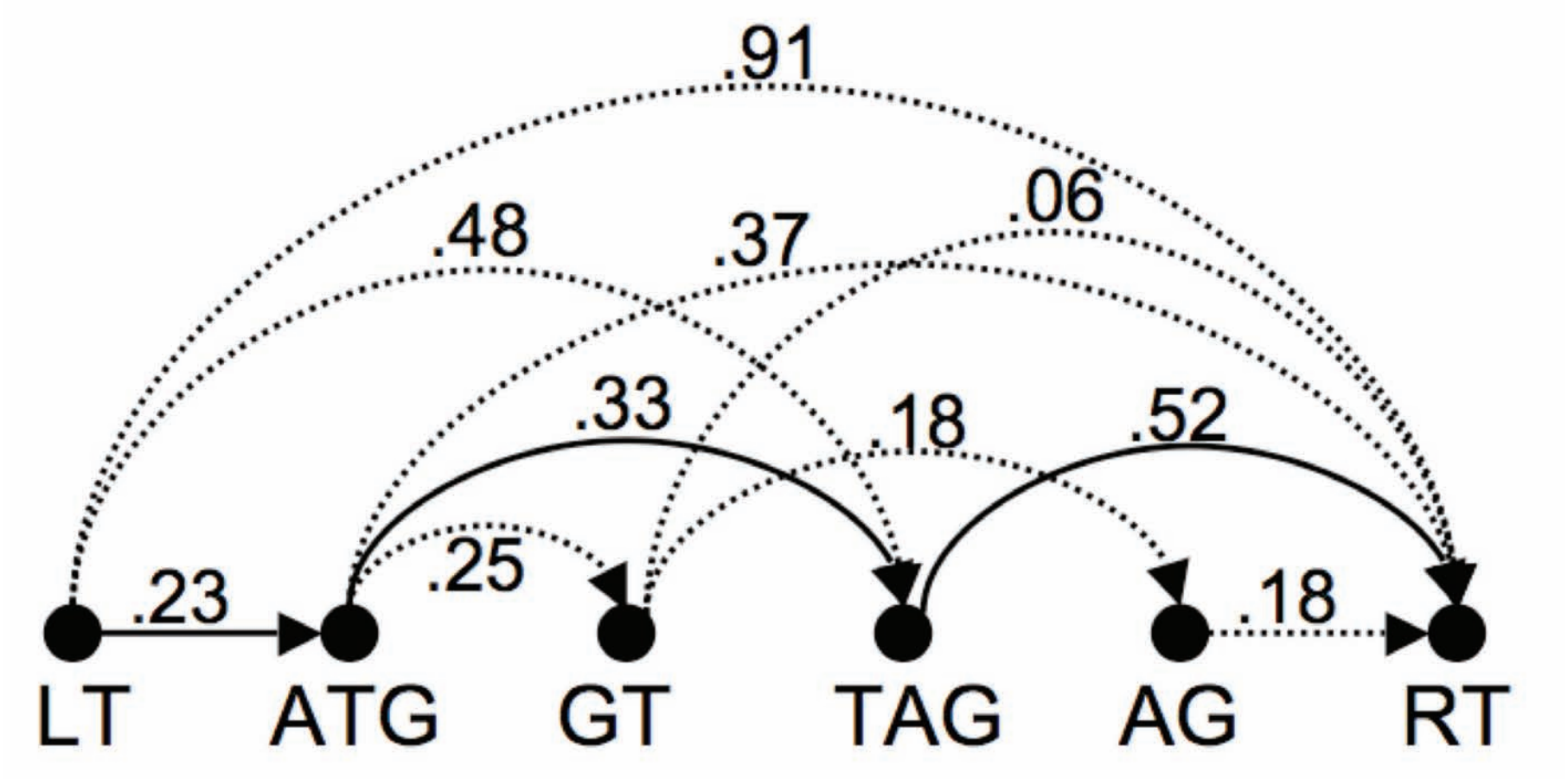


find highest-scoring path through ORF graph; interpret path as a gene parse = gene structure



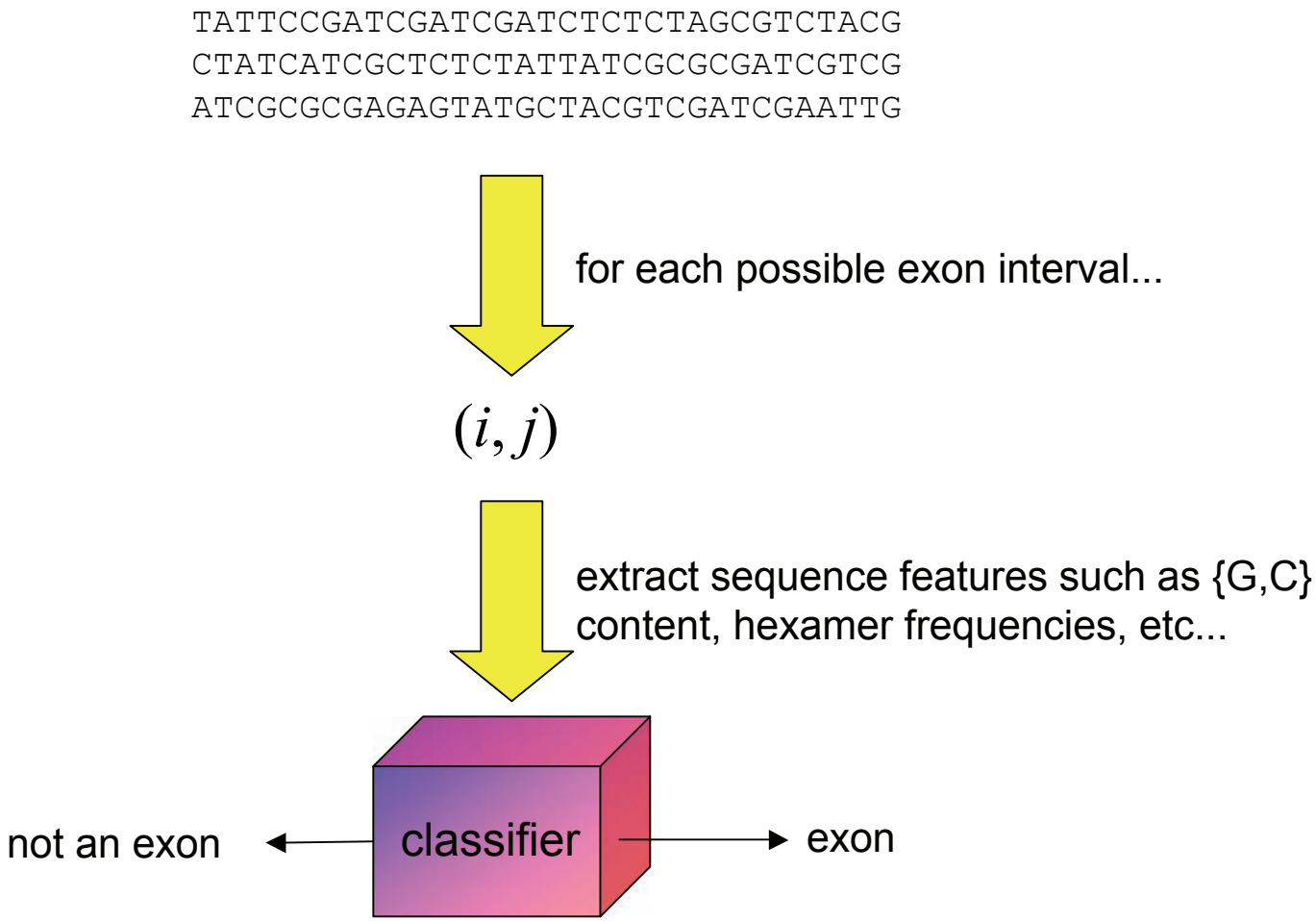
ORF Graphs and the Shortest Path

A standard *shortest-path algorithm* can be trivially adapted to find the highest-scoring parse in an ORF graph:



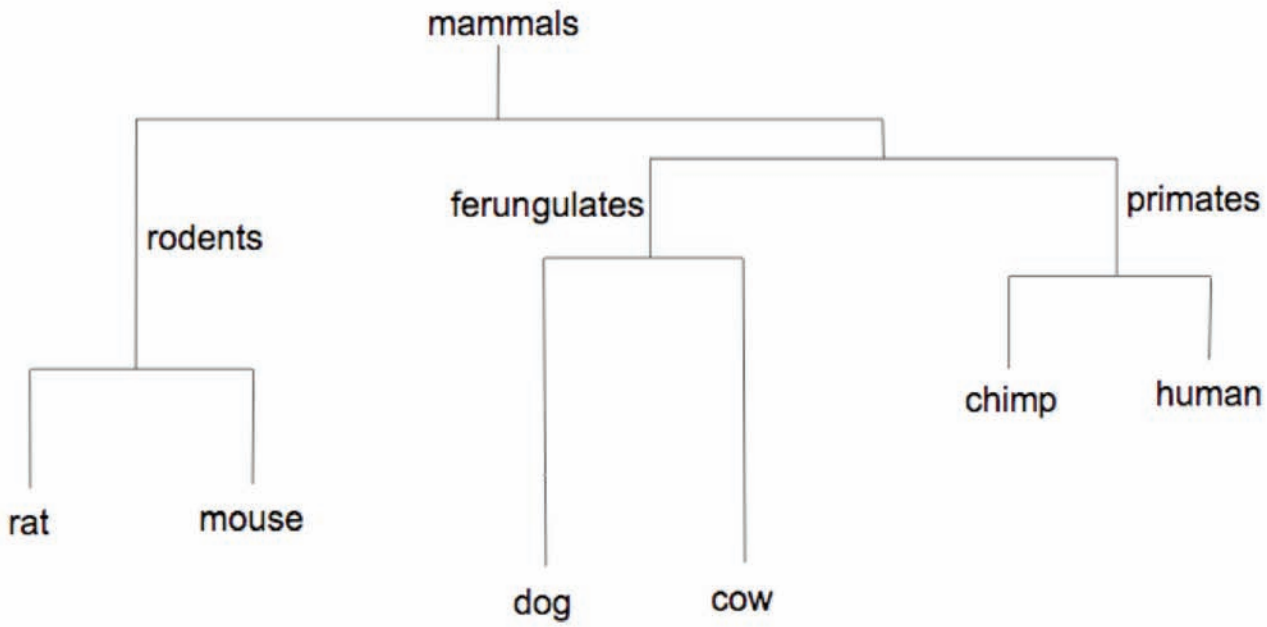
Gene Prediction as Classification

An alternate formulation of the gene prediction process is as one of *classification* rather than *parsing*:



Evolution

The evolutionary relationships (i.e., *common ancestry*) among sequenced genomes can be used to inform the gene-finding process, by observing that *natural selection* operates more strongly (or at different *levels of organization*) within some genomic features than others (i.e., coding versus noncoding regions). Observing these patterns during gene prediction is known as *comparative gene prediction*.



GFF - General Feature Format

GFF (and more recently, GTF) is a standard format for specifying features in a sequence:

```
52  fumigatus  initial-exon  42155  42915  .  +  0  transgrp=10
52  fumigatus  internal-exon 42980  43210  .  +  2  transgrp=10
52  fumigatus  internal-exon 44004  44214  .  +  2  transgrp=10
52  fumigatus  internal-exon 44278  44525  .  +  0  transgrp=10
52  fumigatus  final-exon    44593  44758  .  +  2  transgrp=10
52  fumigatus  initial-exon  59987  60513  .  -  0  transgrp=16
52  fumigatus  final-exon    59549  59930  .  -  2  transgrp=16
52  fumigatus  single-exon   73702  74544  .  +  0  transgrp=24
```

Columns are, left-to-right: (1) contig ID, (2) organism, (3) feature type, (4) begin coordinate, (5) end coordinate, (6) score or dot if absent, (7) strand, (8) phase, (9) extra fields for grouping features into transcripts and the like.

What is a FASTA file?

Sequences are generally stored in FASTA files. Each sequence in the file has its own *defline*. A defline begins with a '>' followed by a *sequence ID* and then any free-form textual information describing the sequence.

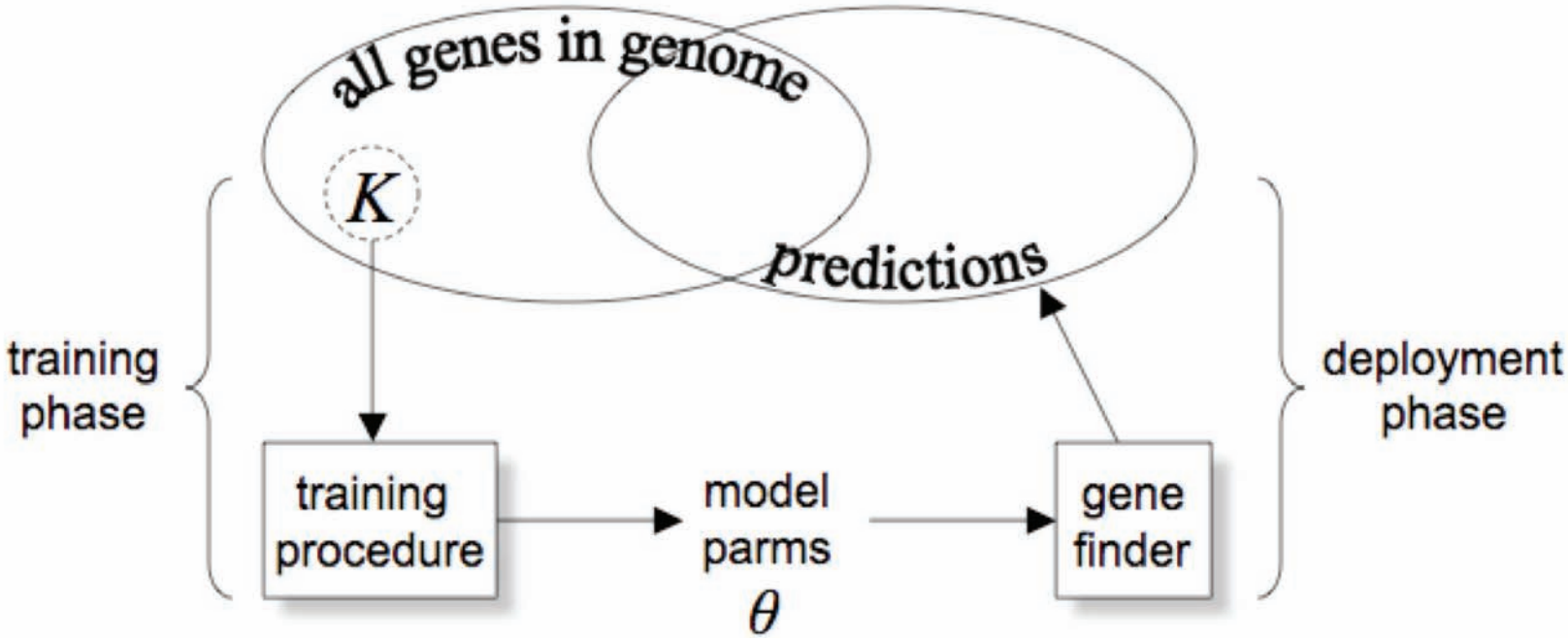
```
>7832 contig assembled on Nov-19-04 by chASM v3.2
ATCGATCGATCGGCGATGCTAGCTACTAGCTGATTCTCTCTAGAGAGCTAGCTGAC
GGCGTAGCTAGCTAGCTGCGATTCAGCGTACGTAGCTAGCTATCTACTTCGATCGT
AGCTATTCGATCTAGCTAGTCGATGTCAGCGCGCGATTATATCGTGCTATCGTGCG
TATCATATATATAGCGCGCGATCGTCGGCGCATGCGAGAGAGTCGTAGTAGTCGTA
GCGCTAGCTGATGCTGTCGTAGCTATCTTTCAGTAG
```

Sequence lines can be formatted to arbitrary length. Deflines are sometimes formatted into a set of *attribute-value pairs* or according to some other convention, but no standard syntax has been universally accepted.

Duke UNIVERSITY

Training Data vs. The Real World

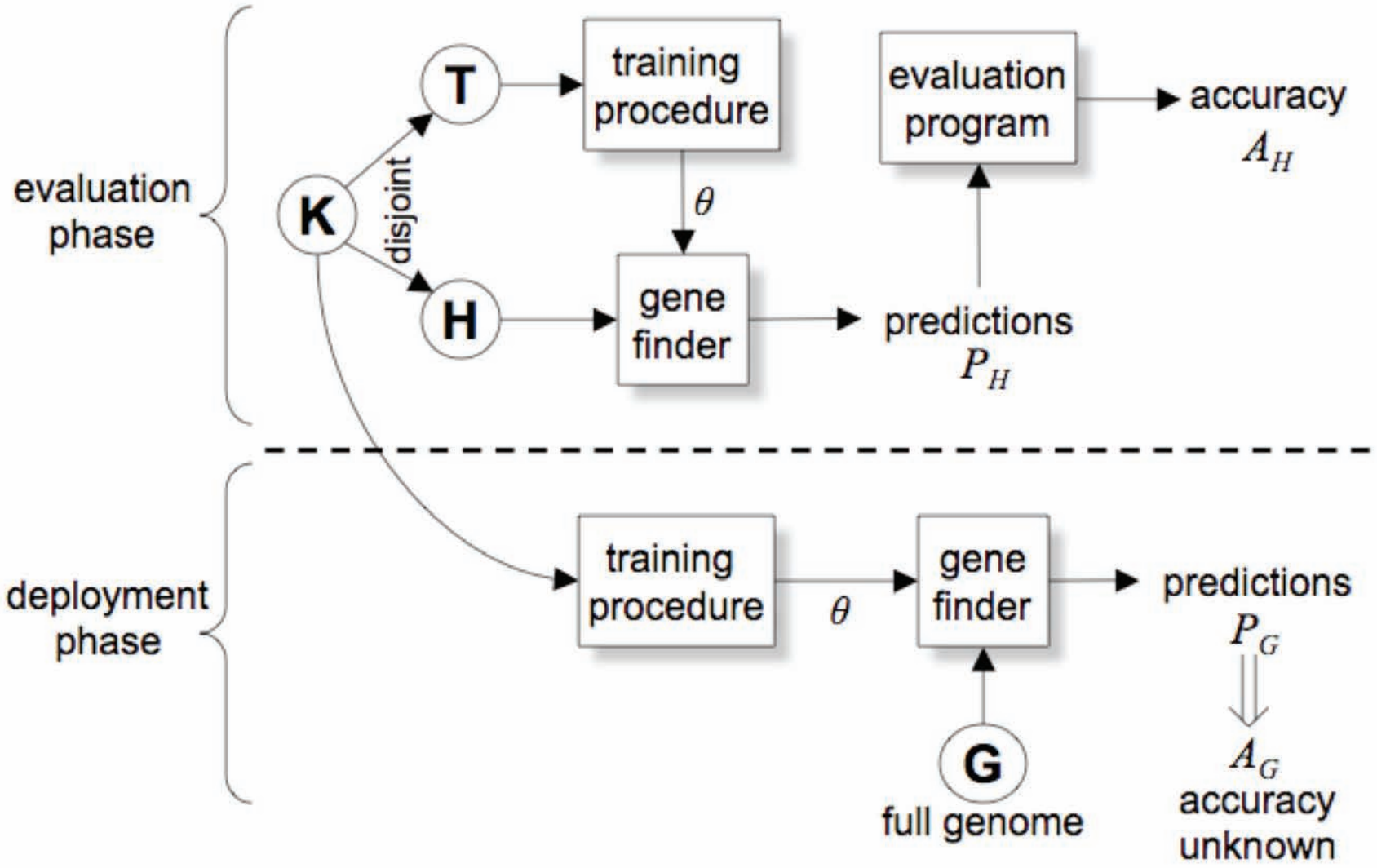
During *training* of a gene finder, only a subset *K* of an organism's gene set will be available for training:



The gene finder will later be *deployed* for use in predicting the rest of the organism's genes. The way in which the *model parameters* are inferred during training can significantly affect the accuracy of the deployed program.

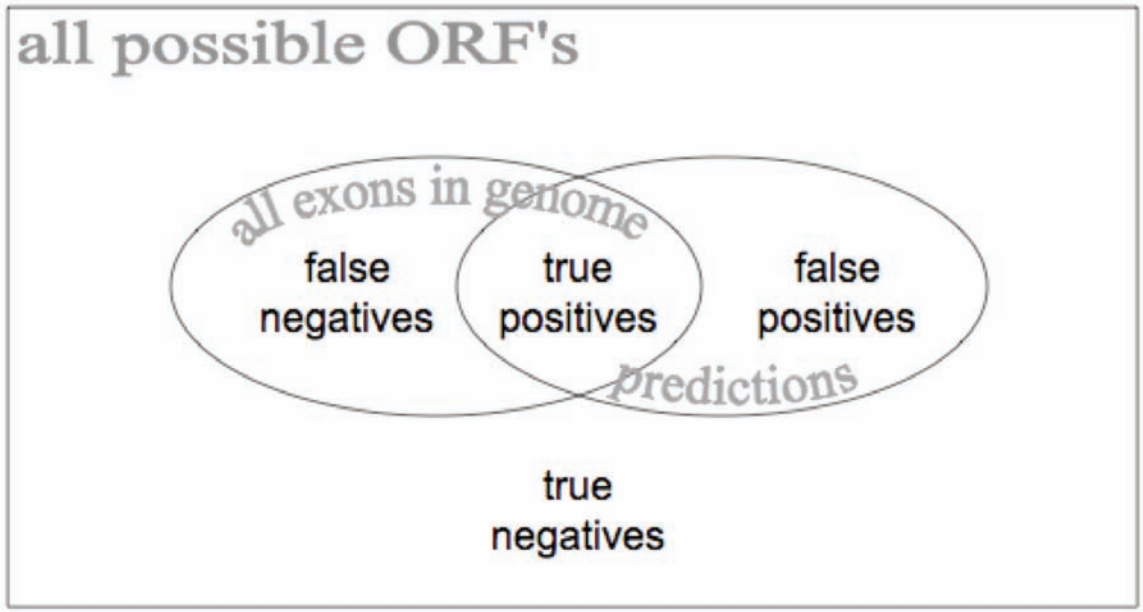
Duke UNIVERSITY

Estimating the Expected Accuracy



TP, FP, TN, and FN

Gene predictions can be evaluated in terms of *true positives* (predicted features that are real), *true negatives* (non-predicted features that are not real), *false positives* (predicted features that are not real), and *false negatives* (real features that were not predicted:



These definitions can be applied at the *whole-gene*, *whole-exon*, or *individual nucleotide* level to arrive at three sets of statistics.

Evaluation Metrics for Prediction Programs

$$Sn = \frac{TP}{TP + FN} \qquad Sp = \frac{TP}{TP + FP} \qquad F = \frac{2 \times Sn \times Sp}{Sn + Sp}$$

$$SMC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}.$$

$$ACP = \frac{1}{n}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right),$$

$$AC = 2(ACP - 0.5).$$

Duke UNIVERSITY

A Baseline for Prediction Accuracy

$$Sn_{rand} = \frac{TP}{TP + FN} = \frac{dp}{dp + d(1 - p)} = p,$$

$$Sp_{rand} = \frac{TP}{TP + FP} = \frac{dp}{dp + (1 - d)p} = d,$$

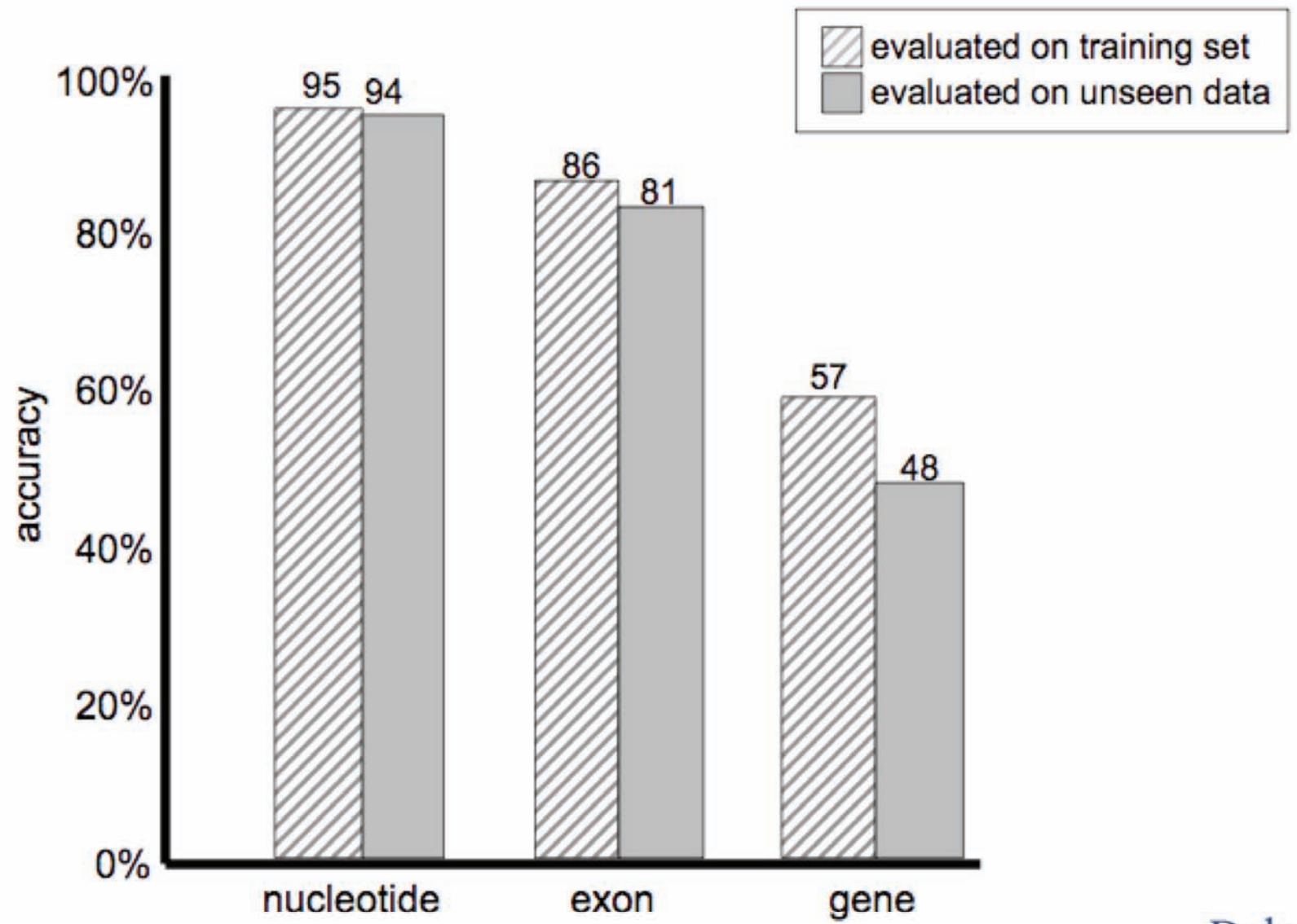
$$F_{rand} = \frac{2 \times Sn \times Sp}{Sn + Sp} = \frac{2dp}{d + p}.$$

$$F_{rand} = \frac{2d}{d + 1}.$$

$$SMC_{rand} = \max(d, 1 - d)$$

Duke UNIVERSITY

Never Test on the Training Set!

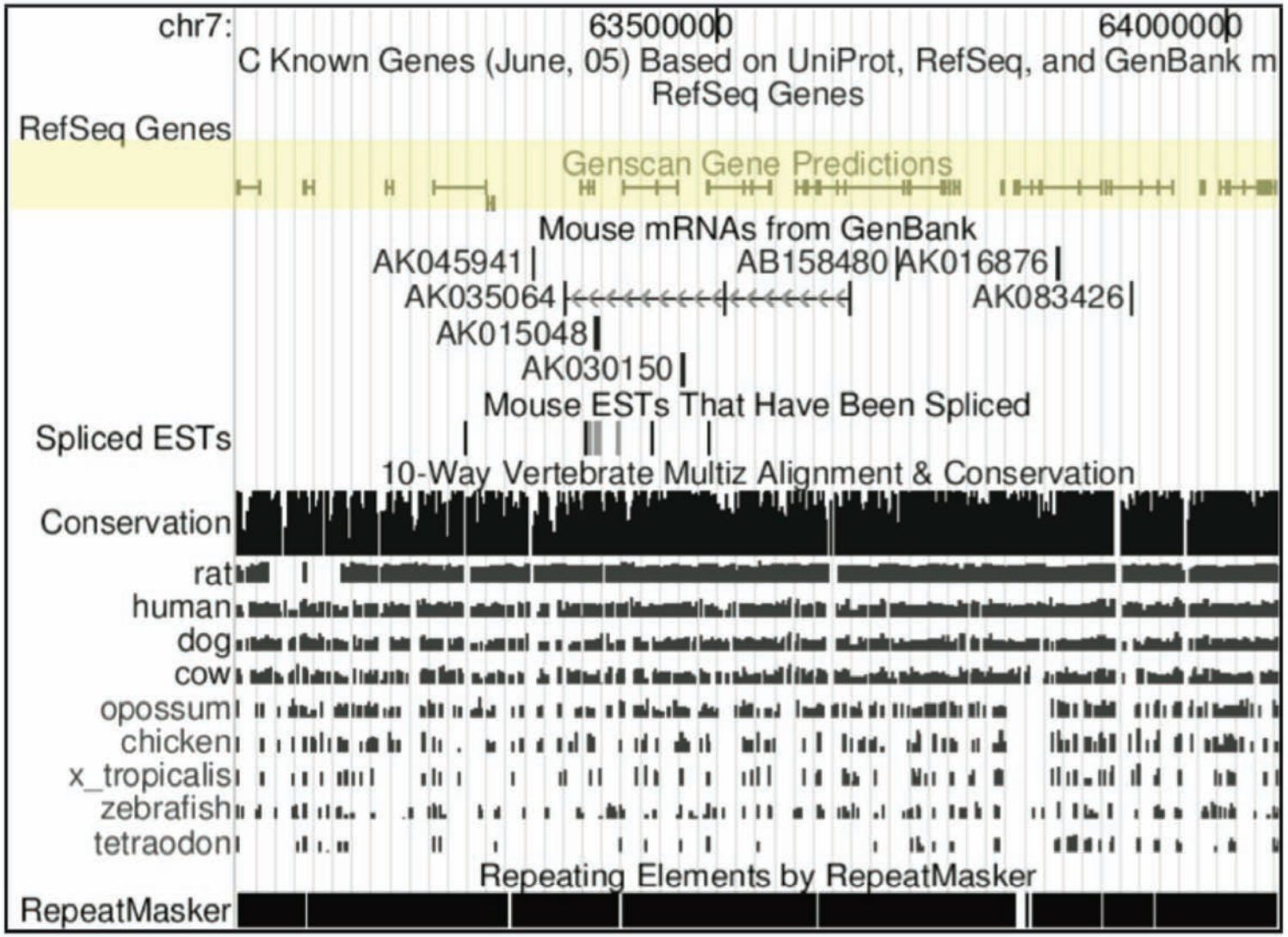


Common Assumptions in Gene Finding

- No overlapping genes
- No nested genes
- No frame shifts or sequencing errors
- Optimal parse only
- No split start codons (ATGT...AGG)
- No split stop codons (TGT...AGAG)
- No alternative splicing
- No selenocysteine codons (TGA)
- No ambiguity codes (Y,R,N, etc.)

Genome Browsers

Manual curation is performed using a graphical browser in which many forms of evidence can be viewed simultaneously. Gene predictions are typically considered the *least reliable* form of evidence by human annotators.



Duke UNIVERSITY