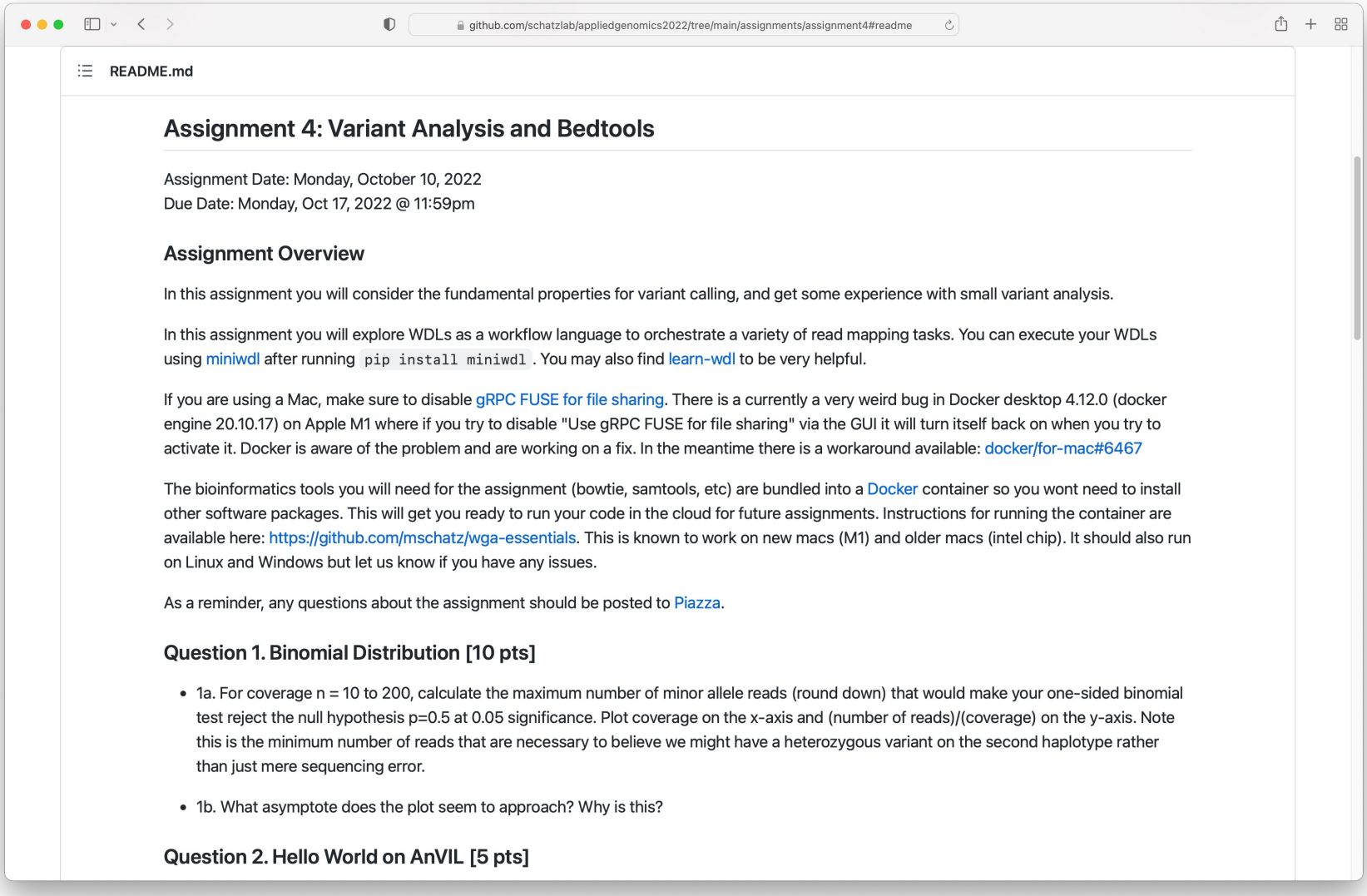
Gene Finding

Michael Schatz

October 17, 2022

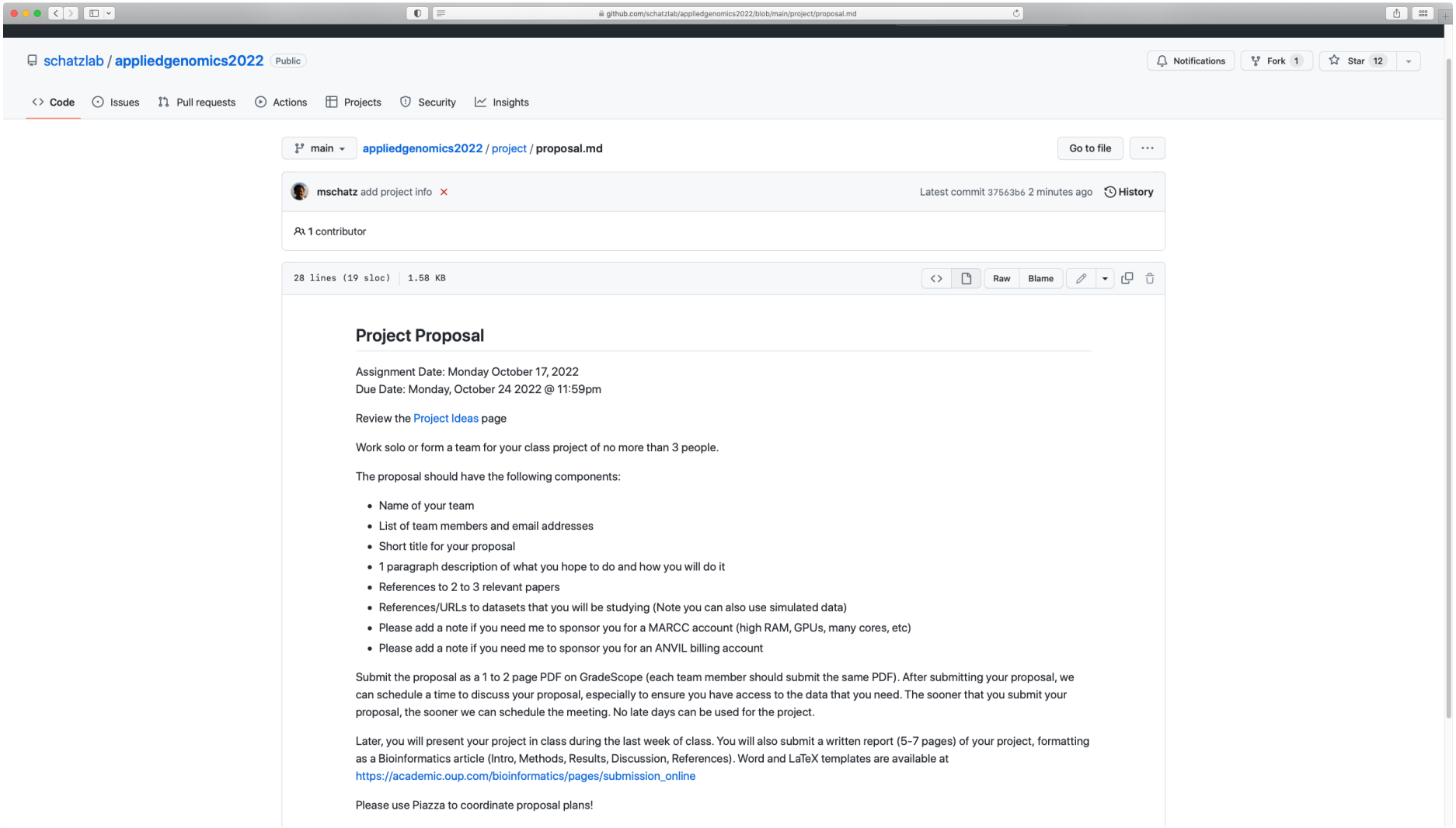
Lecture 14. Applied Comparative Genomics

Assignment 4: Variant Analysis and bedtools
Due Monday Oct 17 by 11:59pm



https://github.com/schatzlab/appliedgenomics2022/tree/main/assignments/assignment4
Check Piazza for questions!

Project Proposal
Due Monday Oct 24 by 11:59pm



schatzlab / appliedgenomics2022 Public

Notifications Fork 1 Star 12

Code Issues Pull requests Actions Projects Security Insights

main appliedgenomics2022 / project / proposal.md Go to file

mschatz add project info Latest commit 37563b6 2 minutes ago History

1 contributor

28 lines (19 sloc) 1.58 KB Raw Blame

Project Proposal

Assignment Date: Monday October 17, 2022
Due Date: Monday, October 24 2022 @ 11:59pm

Review the Project Ideas page

Work solo or form a team for your class project of no more than 3 people.

The proposal should have the following components:

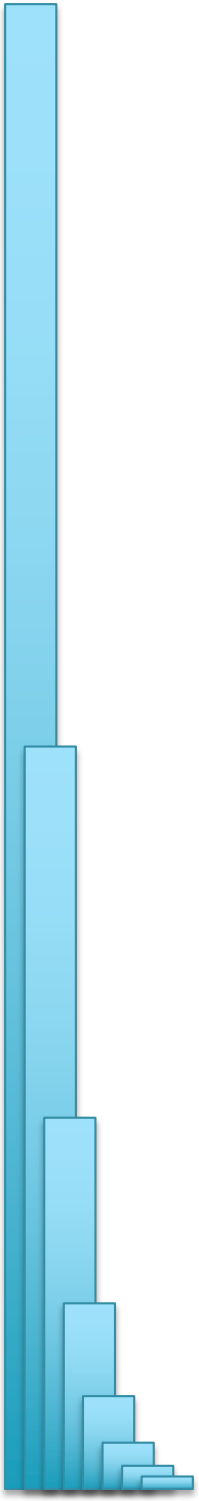
- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to 2 to 3 relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)
- Please add a note if you need me to sponsor you for a MARCC account (high RAM, GPUs, many cores, etc)
- Please add a note if you need me to sponsor you for an ANVIL billing account

Submit the proposal as a 1 to 2 page PDF on GradeScope (each team member should submit the same PDF). After submitting your proposal, we can schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for the project.

Later, you will present your project in class during the last week of class. You will also submit a written report (5-7 pages) of your project, formatting as a Bioinformatics article (Intro, Methods, Results, Discussion, References). Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online

Please use Piazza to coordinate proposal plans!

https://github.com/schatzlab/appliedgenomics2022/tree/main/project/proposal.md
Check Piazza for questions!



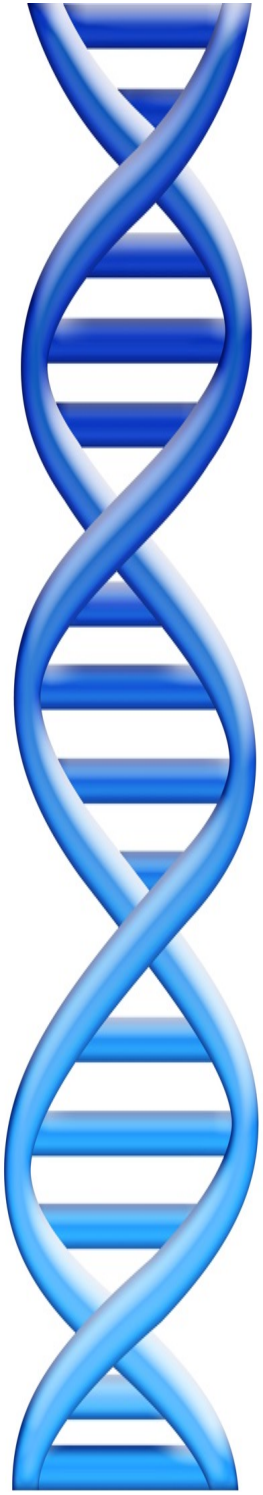
Annotation

Goal: Genome Annotations

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaat
gcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatgct
aatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcg
gctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcct
gcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcg
gctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctgggaat
gcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctg
ggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgg
gatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatc
cgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcg
gctatgctaatgcatgcggctatgctaagctcatgcgg

Goal: Genome Annotations

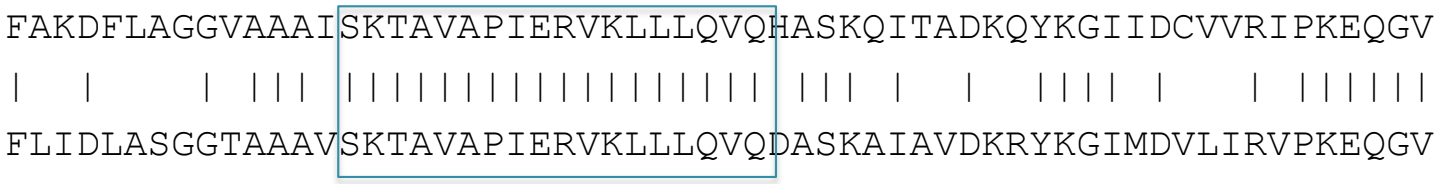
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaat
gcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatgct
aatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcg
gctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcct
gcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatt[Gene!]gctatgctaagctgggaatgcatgcg
gctatgctaagctgggatccgat[redacted]atgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctgggaat
gcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctg
ggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgg
gatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatc
cgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcg
gctatgctaatgcatgcggctatgctaagctcatgcgg



Outline

1. Alignment to other genomes
2. Prediction aka “Gene Finding”
3. Experimental & Functional Assays

Seed and Extend



- Homologous sequences are likely to contain a short high scoring word pair, a seed.
  - Smaller seed sizes make the sense more sensitive, but also (much) slower
  - Typically do a fast search for prototypes, but then most sensitive for final result

- BLAST then tries to extend high scoring word pairs to compute high scoring segment pairs (HSPs).
  - Significance of the alignment reported via an e-value

BLAST E-values

E-value = the number of HSPs having alignment score S (or higher) expected to occur by chance.

→ Smaller E-value, more significant in statistics

→ Bigger E-value, less significant

→ Over 1 means expect this totally by chance

(not significant at all!)

The expected number of HSPs with the score at least S is :

$$E = K*n*m*e^{-\lambda S}$$

K, λ are constant depending on model

n, m are the length of query and sequence

E-values quickly drop off for better alignment bits scores

Very Similar Sequences

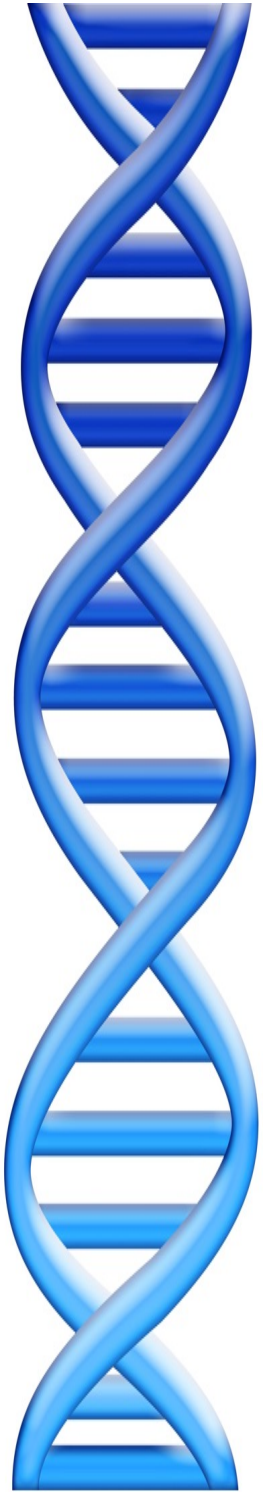
```
Query: HBA_HUMAN Hemoglobin alpha subunit
Sbjct: HBB_HUMAN Hemoglobin beta subunit

Score =  114 bits (285),  Expect = 1e-26
Identities = 61/145 (42%), Positives = 86/145 (59%), Gaps = 8/145 (5%)

Query  2   LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV 55
           L+P +K+ V A WGKV  +  E G EAL R+ + +P T+ +F  F      D   G+ +V
Sbjct  3   LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV 60

Query  56  KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA 115
           K HGKKV  A ++ +AH+D++    + LS+LH  KL VDP NF+LL + L+  LA H
Sbjct  61  KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK 120

Query  116 EFTPAVHASLDKFLASVSTVLTSKY 140
           EFTP V A+  K +A V+  L  KY
Sbjct  121 EFTPPVQAAYQKVVAGVANALAHKY 145
```



Outline

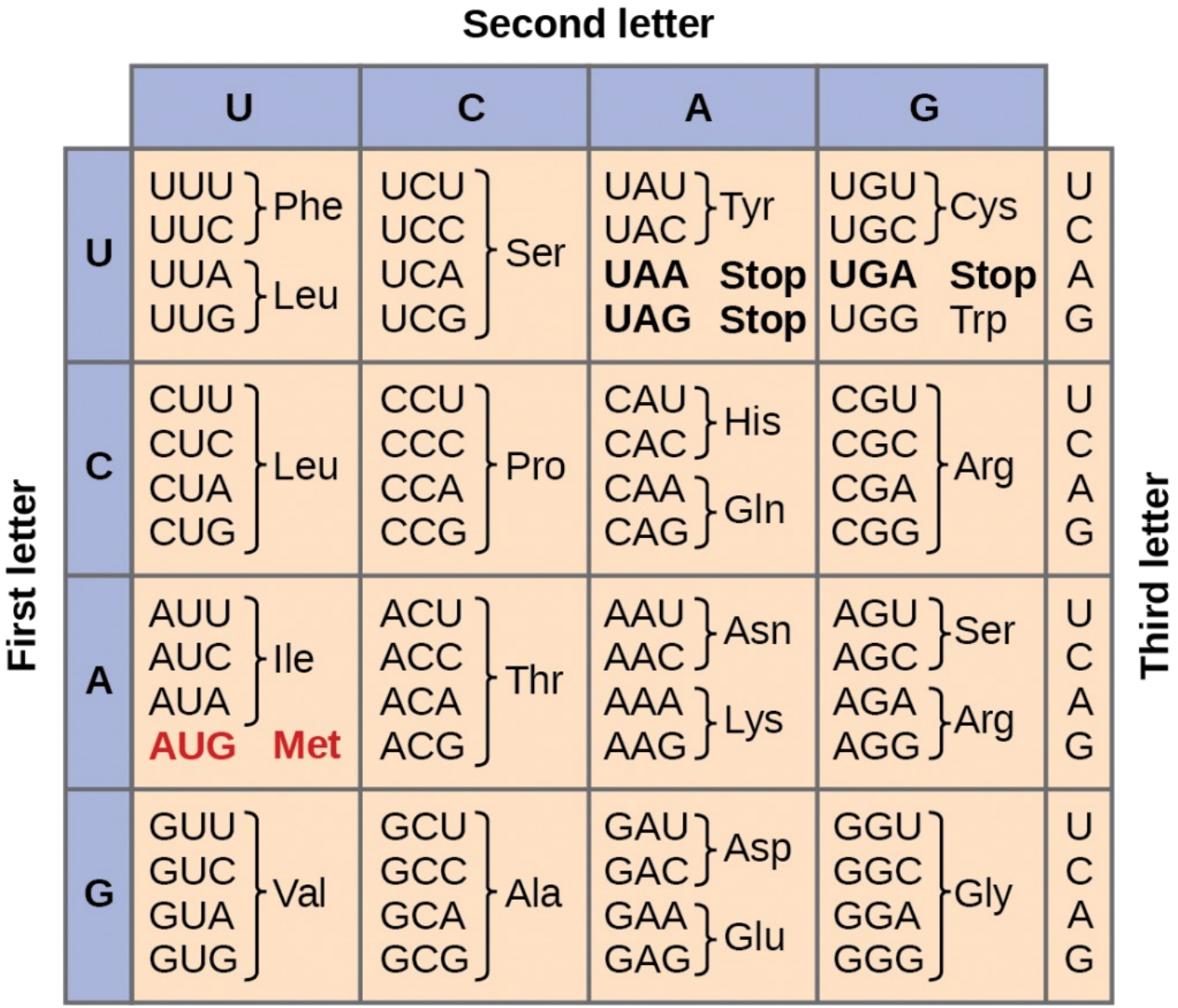
1. Alignment to other genomes
2. Prediction aka "Gene Finding"
3. Experimental & Functional Assays

Bacterial Gene Finding and Glimmer

(also Archaeal and viral gene finding)

Arthur L. Delcher and Steven Salzberg
Center for Bioinformatics and Computational Biology
Johns Hopkins University

Genetic Code



Start:
- AUG

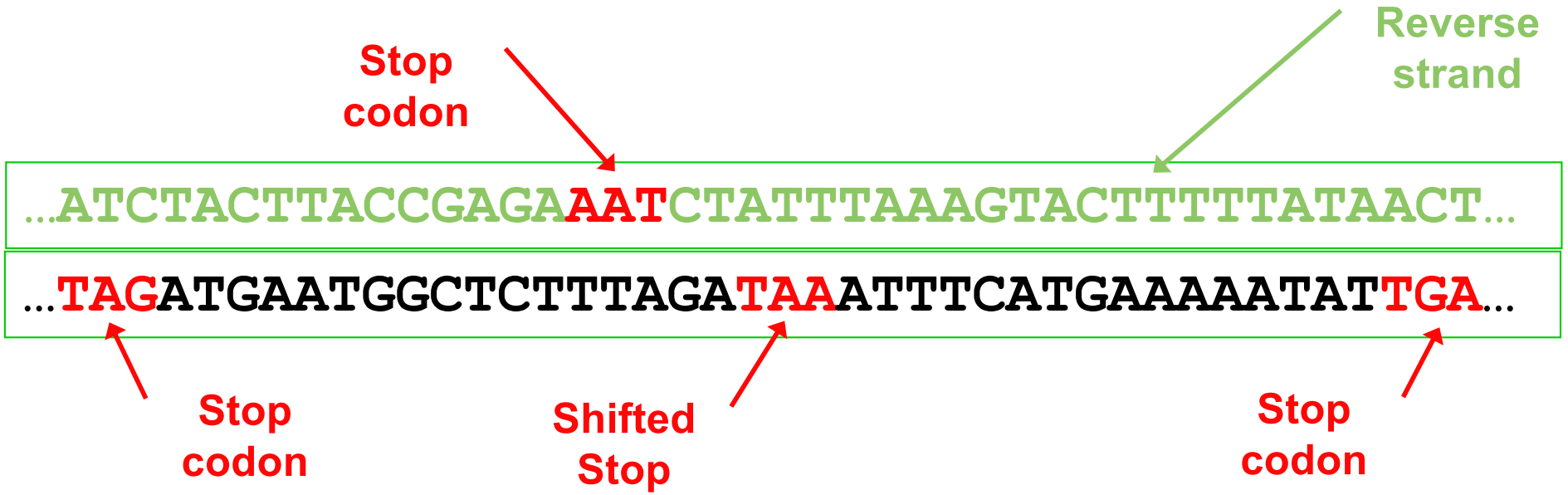
Stop:
- UAA
- UAG
- UGA

Step One

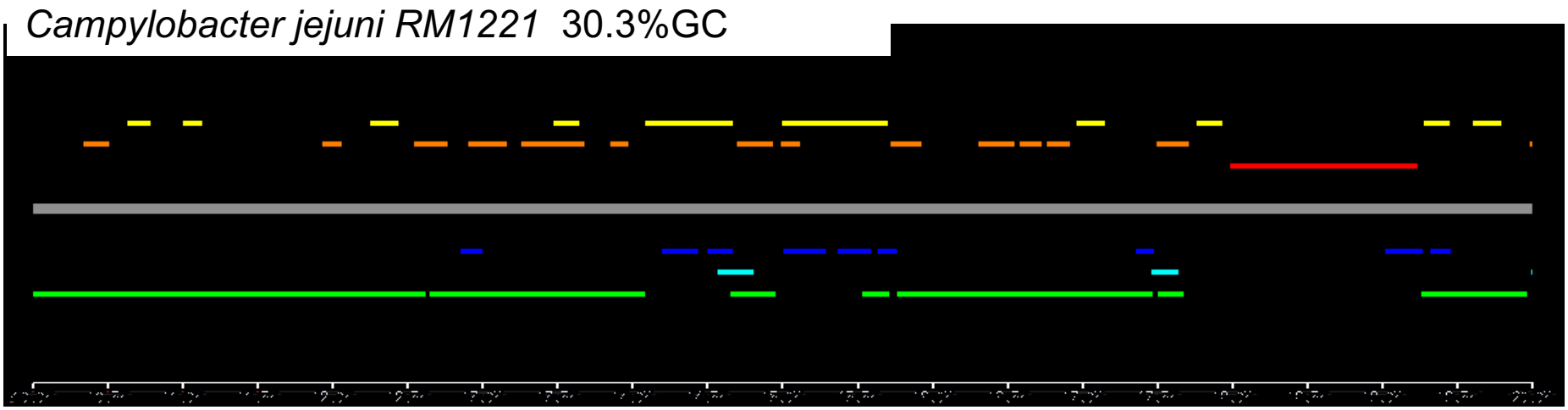
- Find open reading frames (ORFs).

Step One

- Find open reading frames (ORFs).



- But ORFs generally overlap …



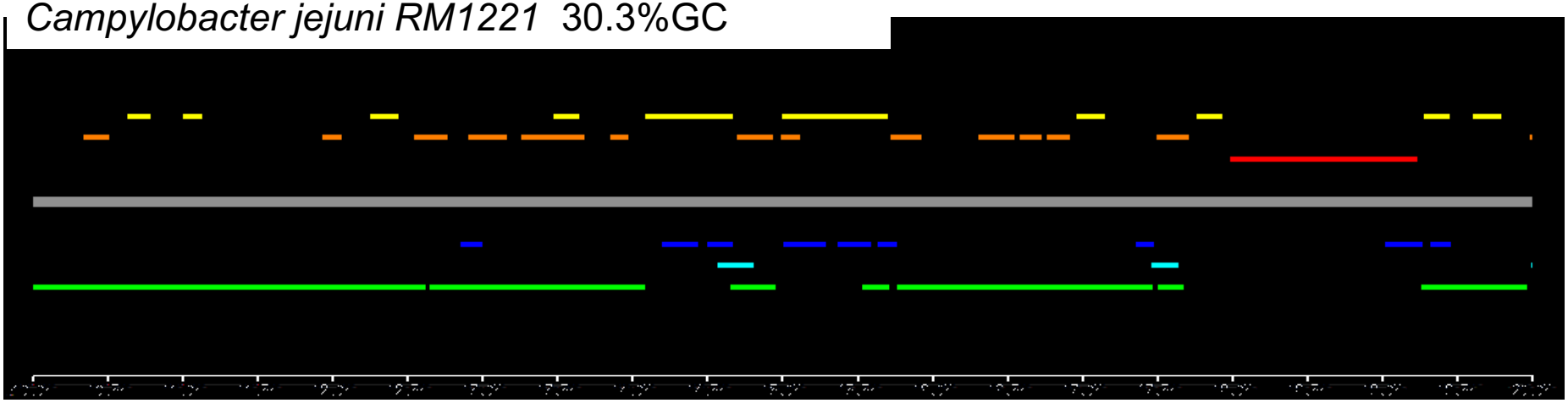
All ORFs longer than 100bp on both strands shown
- color indicates reading frame

Longest ORFs likely to be protein-coding genes

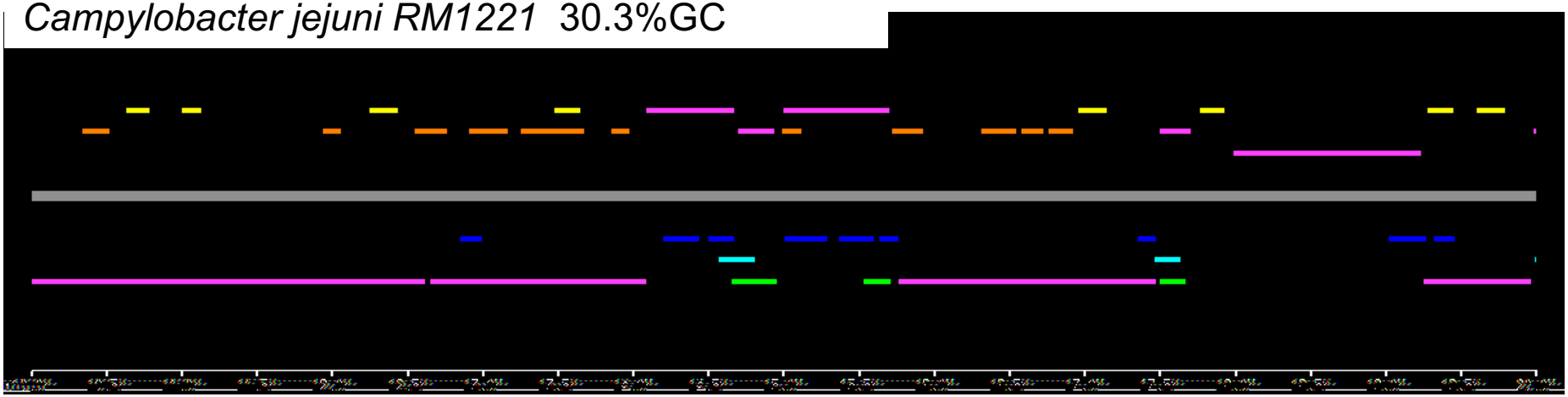
Note the low GC content

All genes are ORFs but not all ORFs are genes

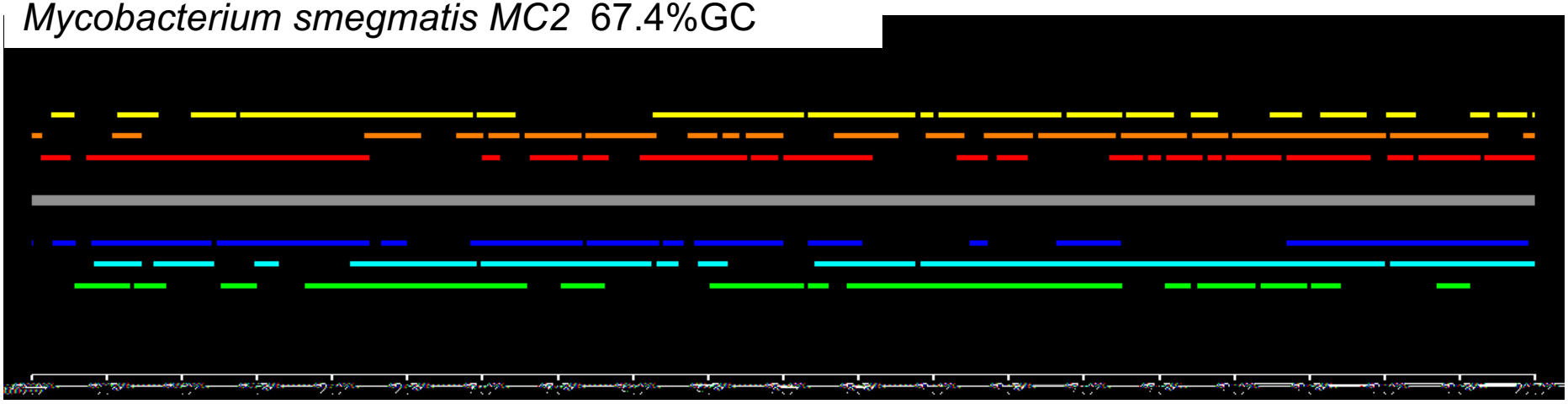
Campylobacter jejuni RM1221 30.3%GC



Campylobacter jejuni RM1221 30.3%GC

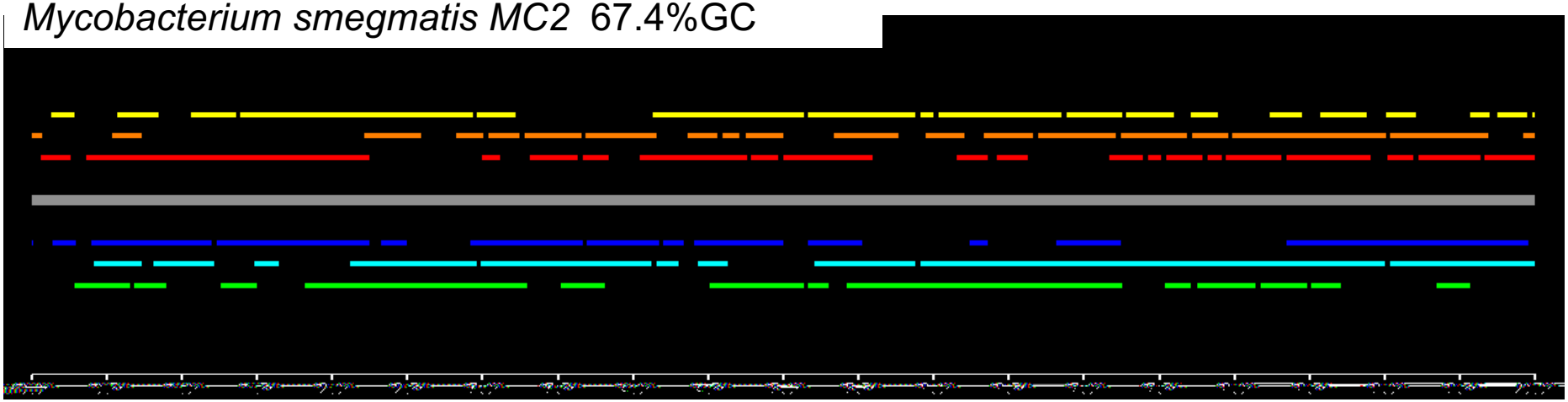


Mycobacterium smegmatis MC2 67.4%GC

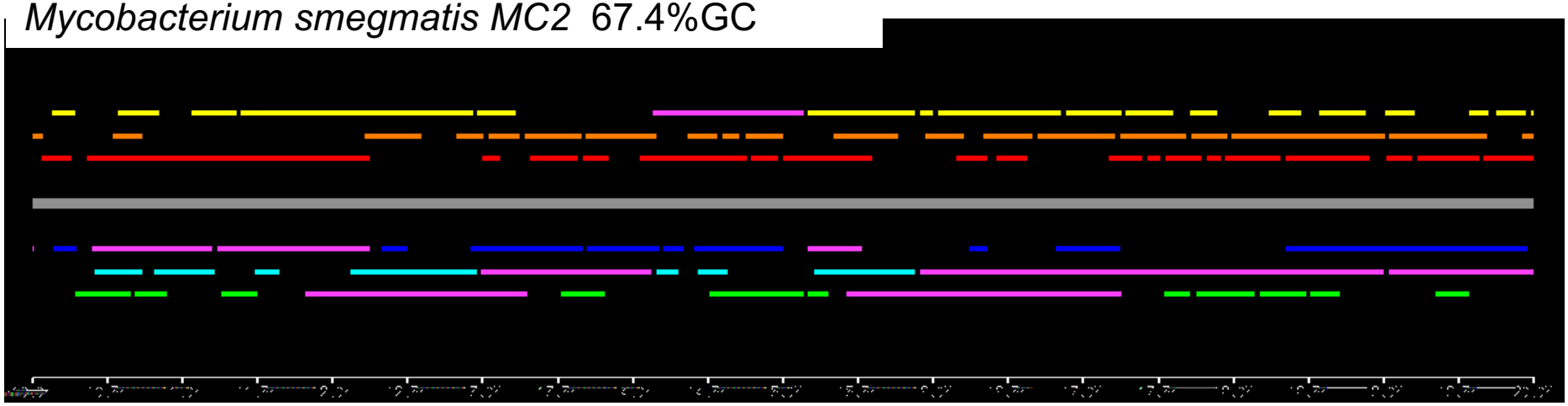


Note what happens in a high-GC genome

Mycobacterium smegmatis MC2 67.4%GC



Mycobacterium smegmatis MC2 67.4%GC



Flipping a Biased Coin

P(heads) = 61/64 (95.4%)   P(tails) = 3/64 (4.6%)

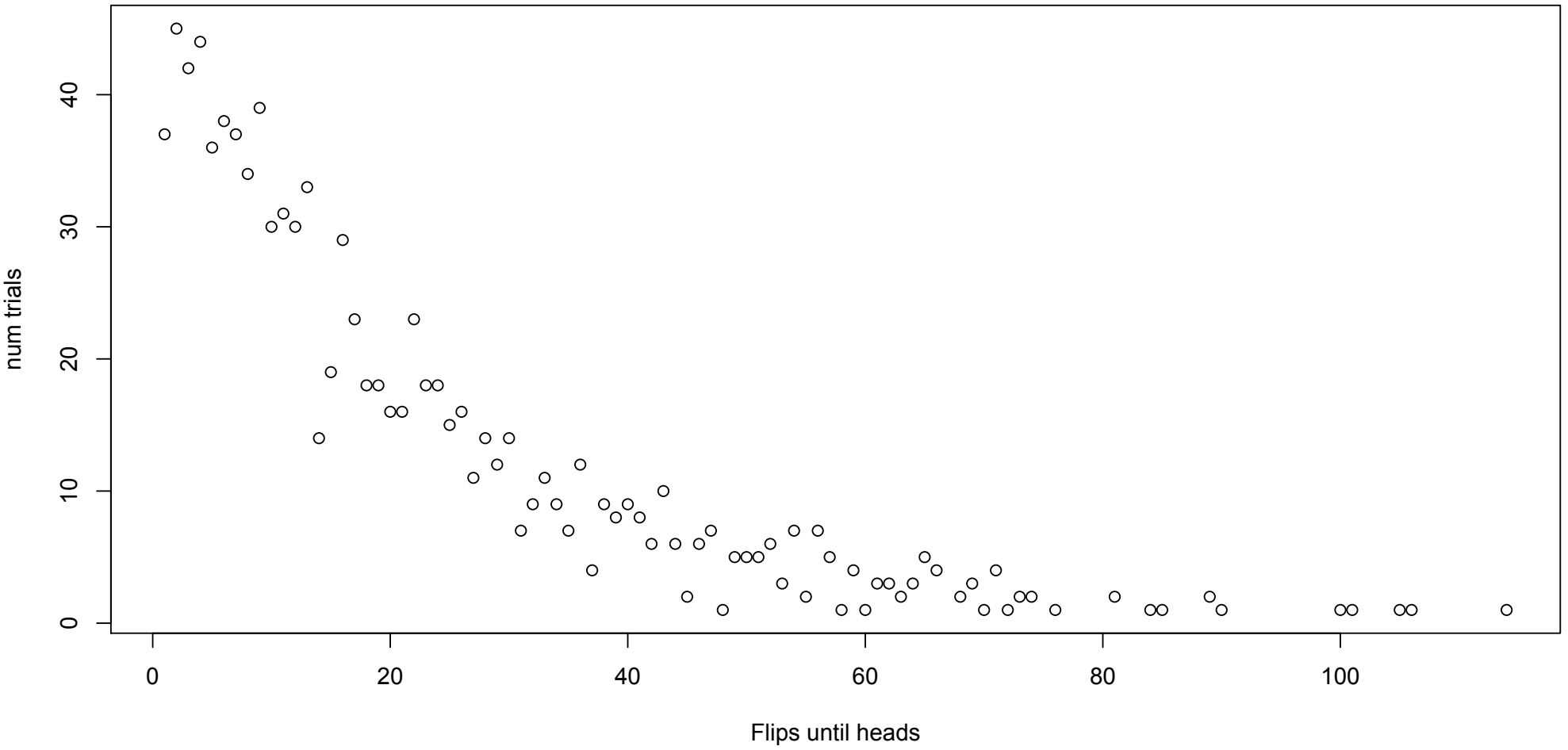
How many flips until my first tail?

```
$ ./coinflip.pl 0.046875 1000
0:   HHHHHHHHHHHHHHT   15
1:   HHHHHHT 7
2:   HHHHHHHHHHHT   12
3:   HHHHHHHHHHHHHHHHHHHHHHHT  24
4:   HT  2
5:   HHHHHHHHHHHHHT 14
6:   HHHHHHHHHT  10
7:   HHHHHHHHHHHHHT 14
8:   HHHHHT  6
9:   HHHHHHHHHHT    11
10:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
11:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT   40
12:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT  45
13:  HHHT    4
14:  HHHHHHHHHHHHHHT   15
15:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT 39
16:  HHHHHT  6
17:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT  38
18:  HHHHHHHHHHHHHHHHHHHHHHHHHT   26
19:  HHHHHHHHHHHT   12
```

Flipping a Biased Coin

P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

How many flips until my first tail?

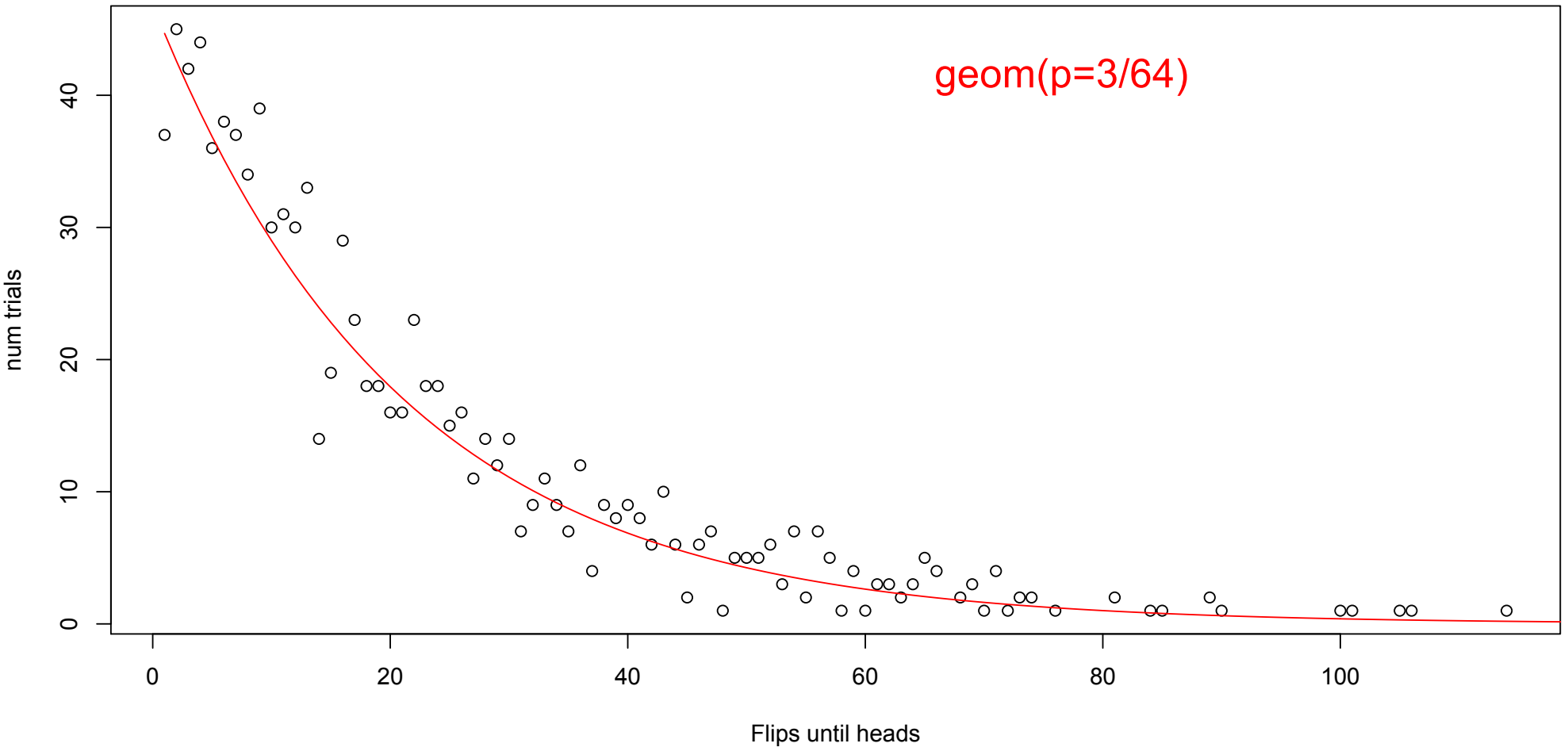


Flipping a Biased Coin

P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

How many flips until my first tail?

Geometric Distribution: $P(X=x) = p_{heads}^{x-1} p_{tails}$

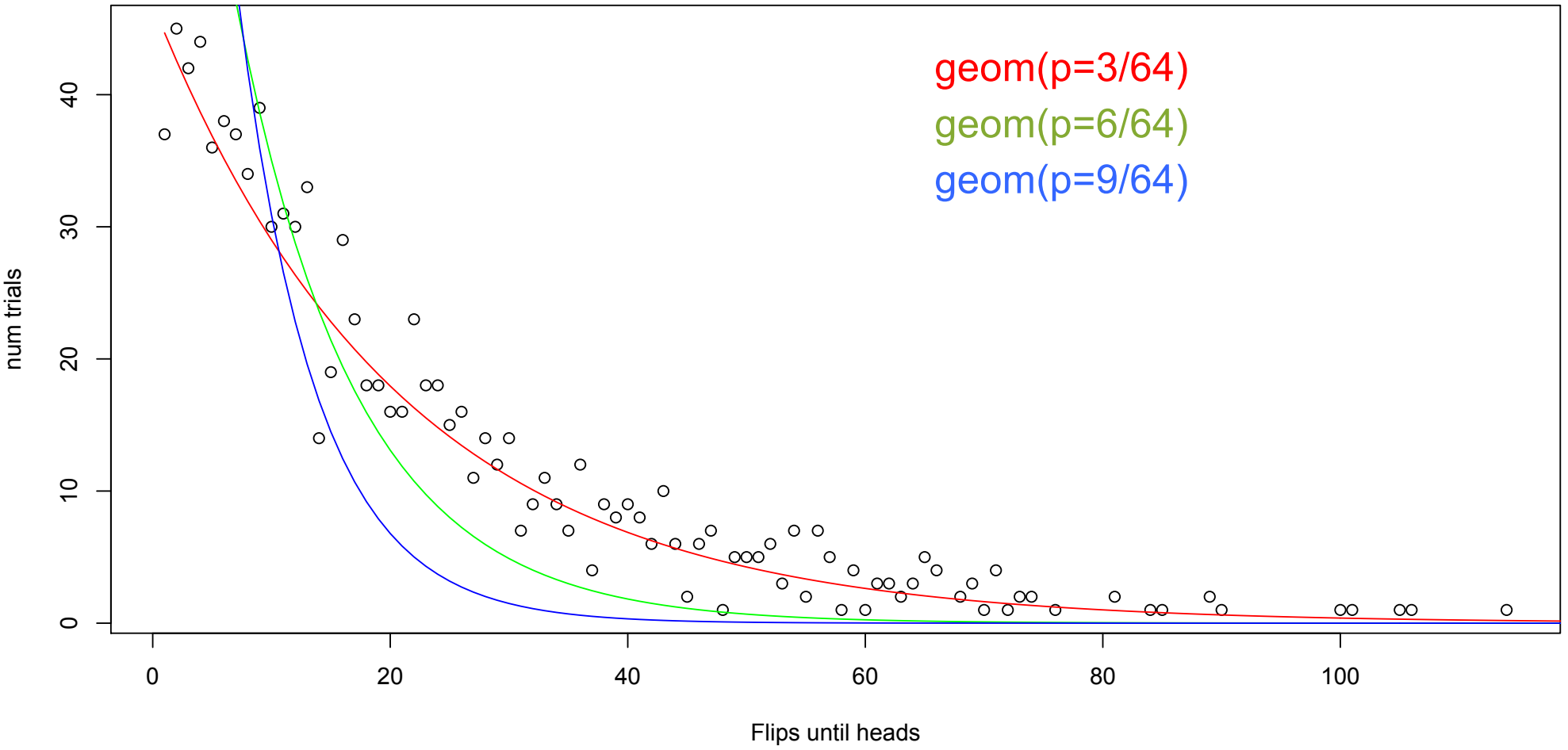


Flipping a Biased Coin

P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

How many flips until my first tail?

Geometric Distribution: $P(X=x) = p_{heads}^{x-1} p_{tails}$

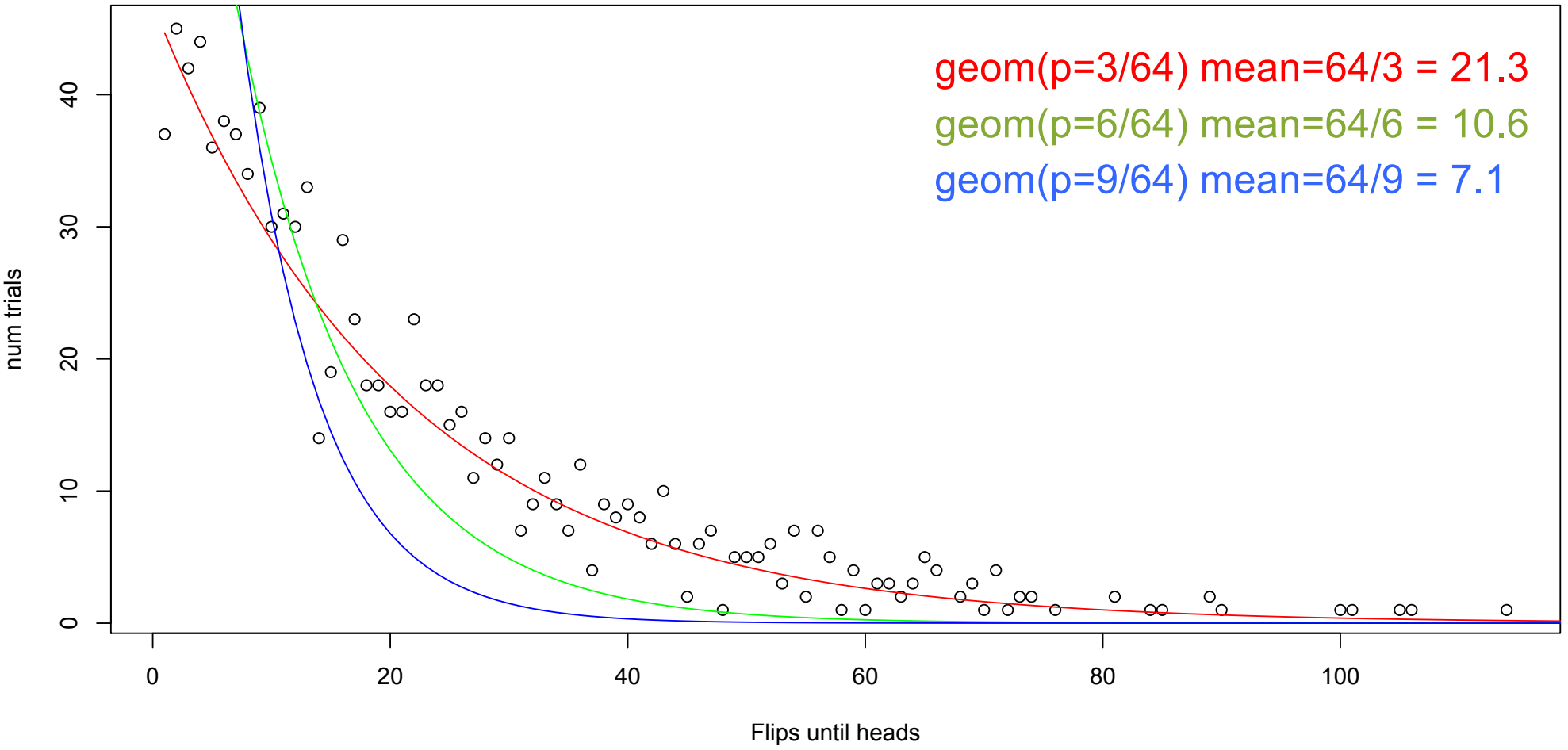


Flipping a Biased Coin

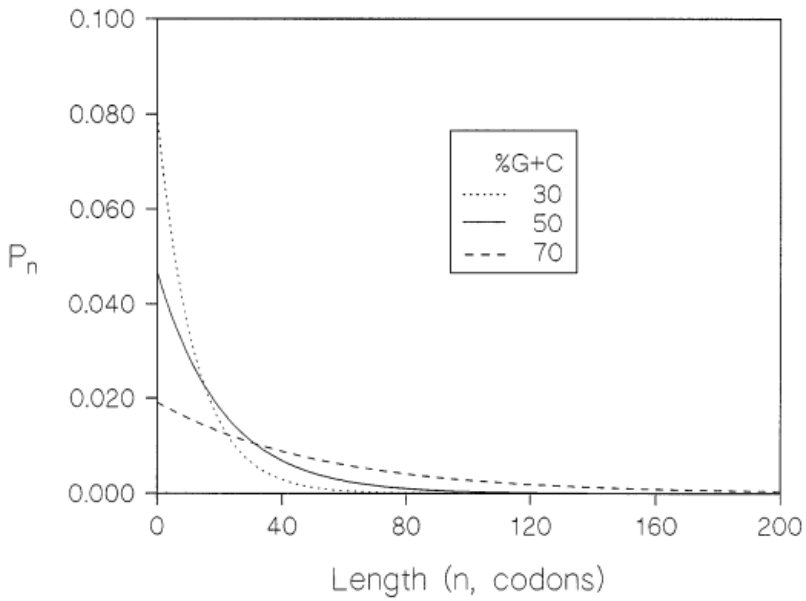
P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

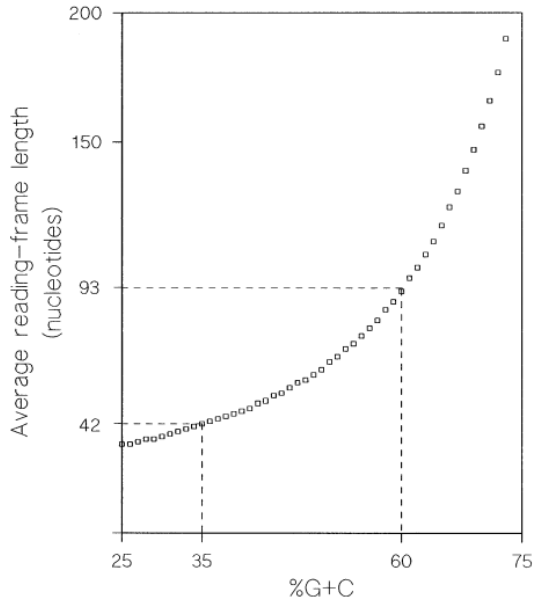
How many flips until my first tail?

Geometric Distribution: $P(X=x) = p_{heads}^{x-1} p_{tails}$



Stop Codon Frequencies



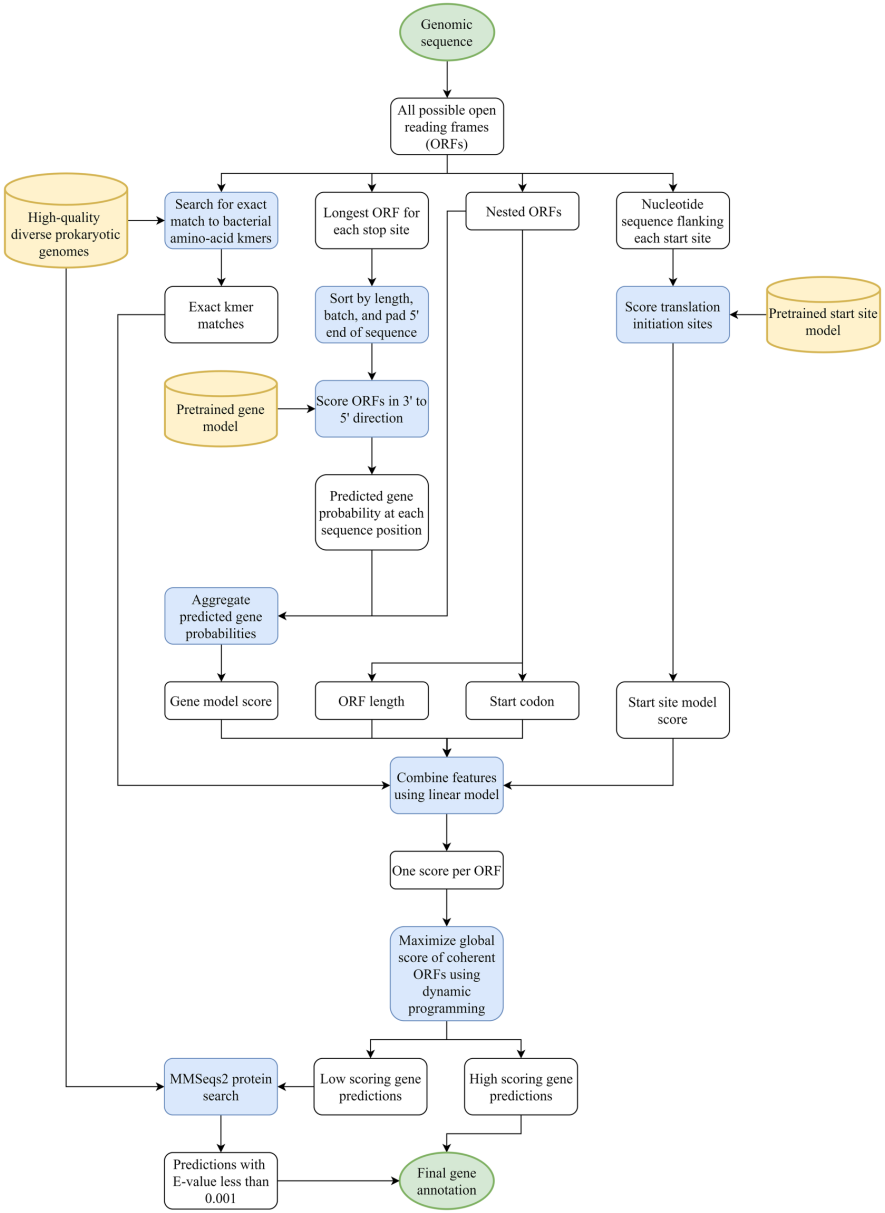


If the sequence is mostly A+T, then likely to form stop codons by chance!

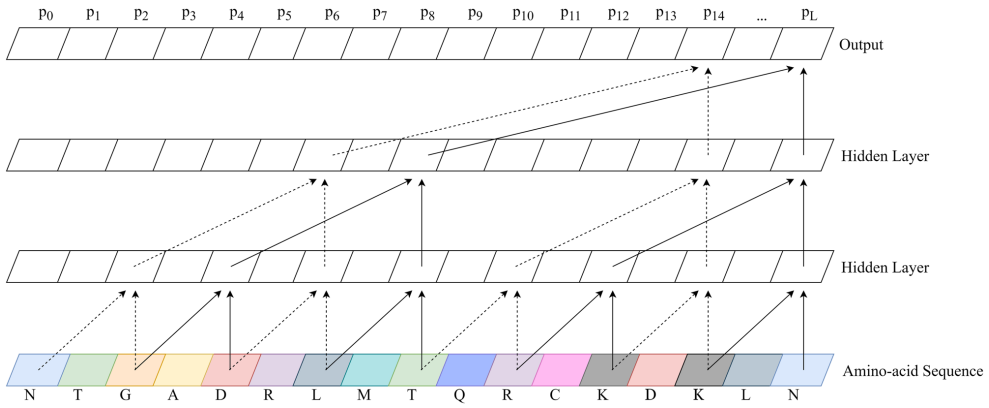
In High A+T (Low G+C):
Frequent stop codons; Short Random ORFs; long ORFs likely to be true genes

In High G+C (Low A+T):
Rare stop codons; Long Random ORFs; harder to identify true genes

A relationship between GC content and coding-sequence length.
Oliver & Marín (1996) J Mol Evol. 43(3):216-23.



Temporal Convolutional Network



Balrog: A universal protein model for prokaryotic gene prediction

Sommer, MJ, Salzberg, SL (2021) PLOS Comp. Bio. doi: 10.1371/journal.pcbi.1008727

Probabilistic Methods

- Create models that have a probability of generating any given sequence.
  - Evaluate gene/non-genome models against a sequence
- Train the models using examples of the types of sequences to generate.
  - Use RNA sequencing, homology, or "obvious" genes
- The "score" of an orf is the probability of the model generating it.
  - Most basic technique is to count how kmers occur in known genes versus intergenic sequences
  - More sophisticated methods consider variable length contexts, "wobble" bases, other statistical clues

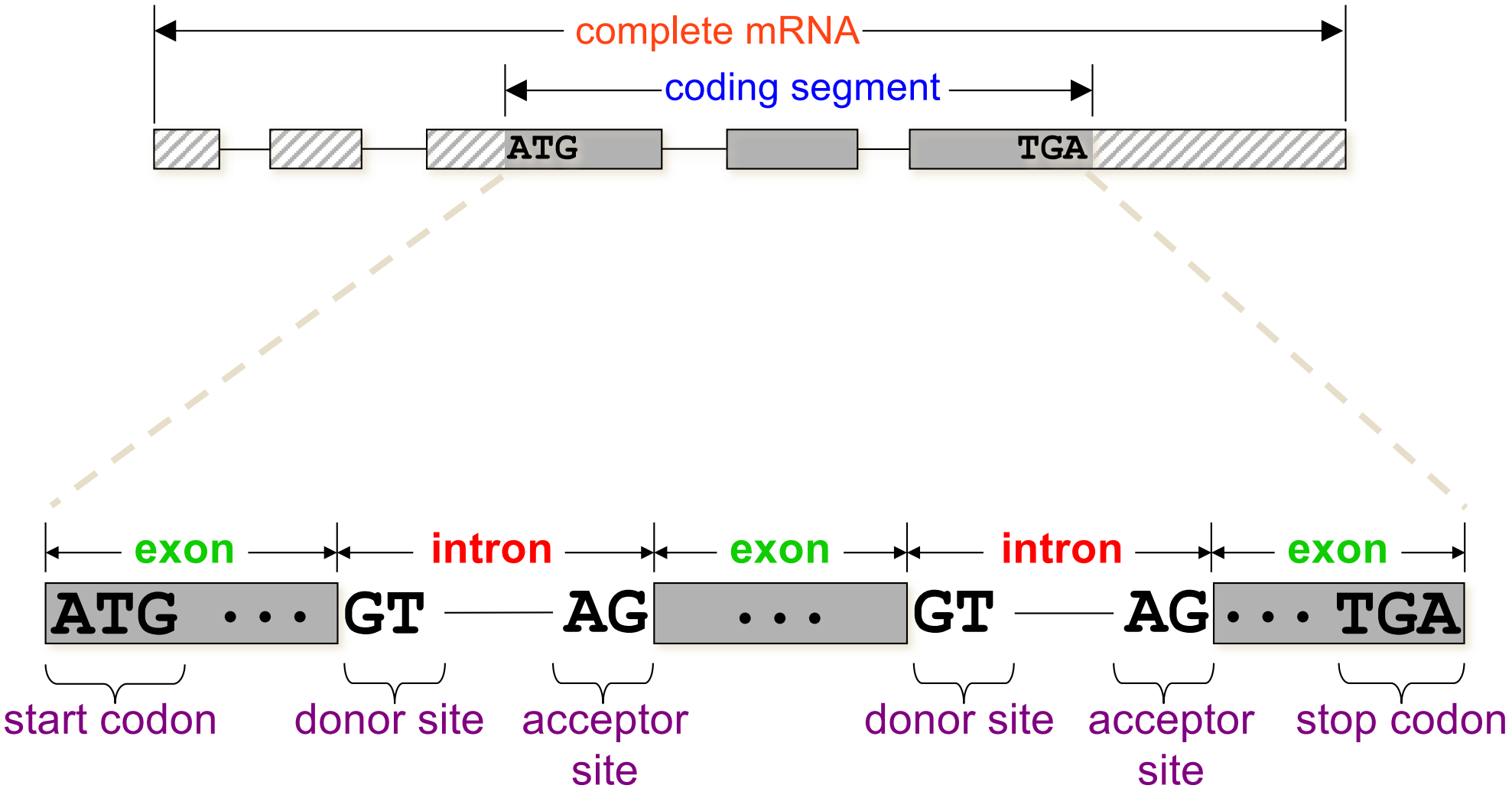
Overview of Eukaryotic Gene Prediction

CBB 231 / COMPSCI 261

W.H. Majoros

Duke UNIVERSITY

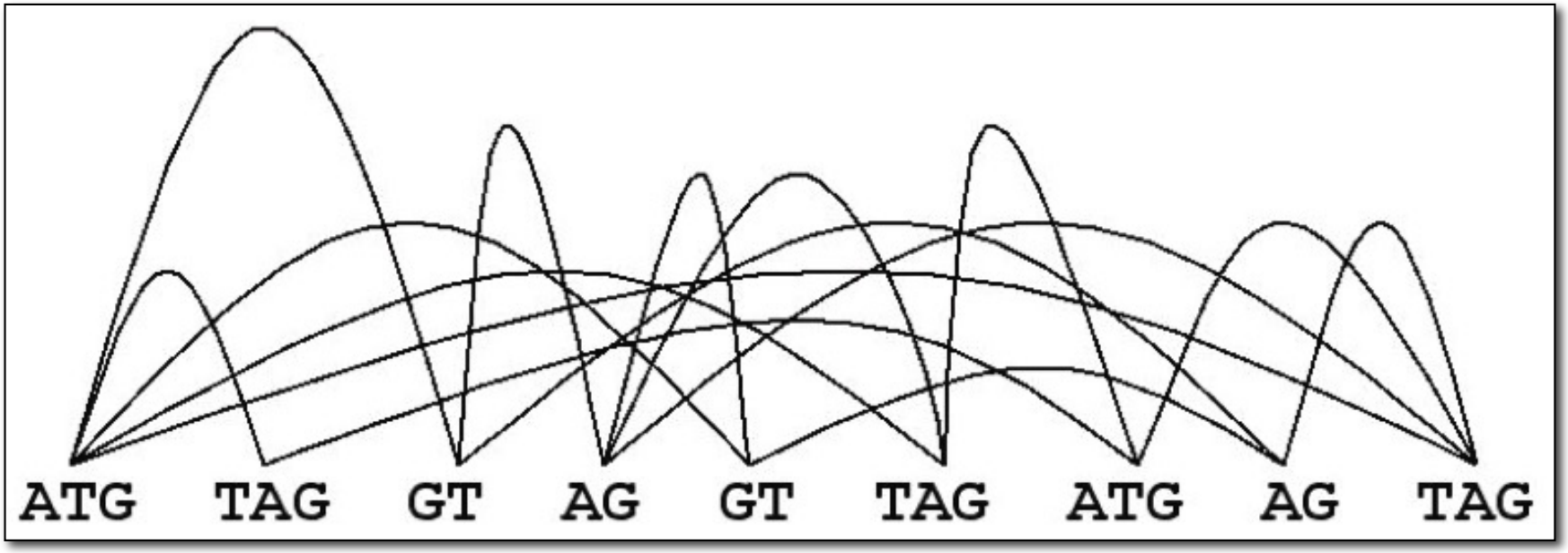
Eukaryotic Gene Syntax



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.

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*:

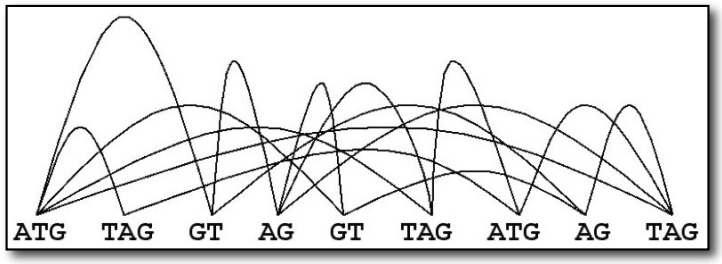


An ORF graph represents all possible *gene parses* (and their scores) for a given set of putative signals. A *path* through the graph represents a single gene parse.

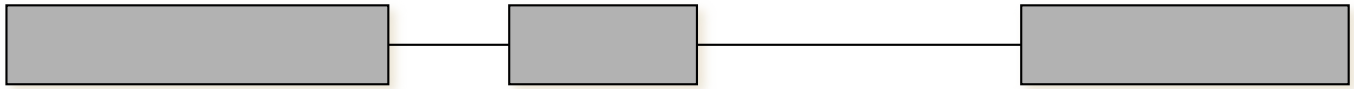
Conceptual Gene-finding Framework

```
TATTCCGATCGATCGATCTCTCTAGCGTCTACG
CTATCATCGCTCTCTATTATCGCGCGATCGTCG
ATCGCGCGAGAGTATGCTACGTCGATCGAATTG
```

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

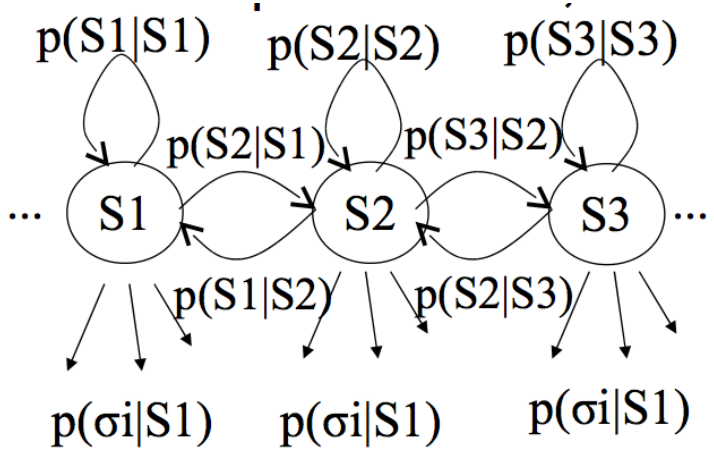


find highest-scoring path through ORF graph; interpret path as a gene parse = gene structure



What is an HMM?

- Dynamic Bayesian Network
  - A set of states
    - {Fair, Biased} for coin tossing
    - {Gene, Not Gene} for Bacterial Gene
    - {Intergenic, Exon, Intron} for Eukaryotic Gene
    - {Modern, Neanderthal} for Ancestry



  - A set of emission characters
    - E={H,T} for coin tossing
    - E={1,2,3,4,5,6} for dice tossing
    - E={A,C,G,T} for DNA

  - State-specific emission probabilities
    - P(H | Fair) = .5, P(T | Fair) = .5, P(H | Biased) = .9, P(T | Biased) = .1
    - P(A | Gene) = .9, P(A | Not Gene) = .1 …

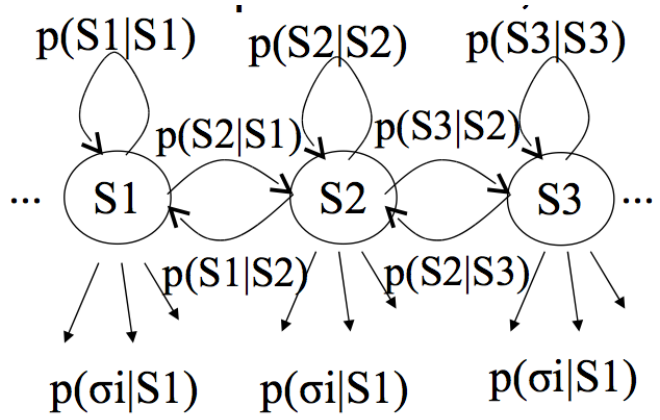
  - A probability of taking a transition
    - $P(s_i=Fair|s_{i-1}=Fair)$ = .9, $P(s_i=Bias|s_{i-1}$ = Fair) .1
    - $P(s_i=Exon \mid s_{i-1}=Intergenic)$, …

Three classic HMM problems

1. **Evaluation**: given a model and an output sequence, what is the probability that the model generated that output?
2. **Decoding**: given a model and an output sequence, what is the most likely state sequence through the model that generated the output?
3. **Learning**: given a model and a set of observed sequences, how do we set the model's parameters so that it has a high probability of generating those sequences?

Why Hidden?

- Similar to Markov models used for prokaryotic gene finding, but system may transition between multiple models called states (gene/non-gene, intergenic/exon/intron)
- Observers can see the emitted symbols of an HMM (i.e., nucleotides) but have no ability to know which state the HMM is currently in.
  - But we can ***infer*** the most likely hidden states of an HMM based on the given sequence of emitted symbols.



AAAGCATGCATTTAACGTGAGCACAATAGATTACA

Eukaryotic Gene Finding with GlimmerHMM

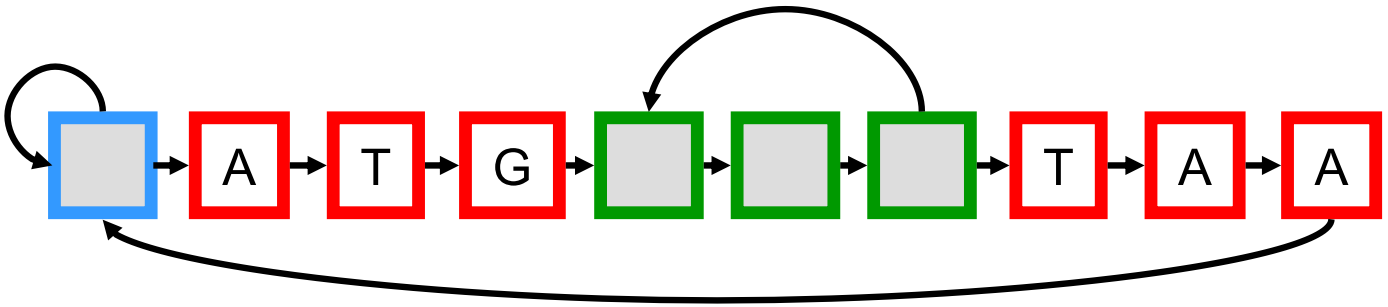
Mihaela Pertea

JHU

HMMs and Gene Structure

- Nucleotides {*A,C,G,T*} are the observables
- Different states generate nucleotides at different frequencies

A simple HMM for unspliced genes:



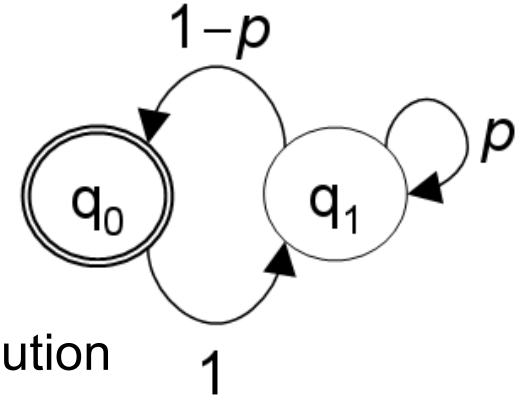
AAAGC ATG CAT TTA ACG AGA GCA CAA GGG CTC TAA TGCCG

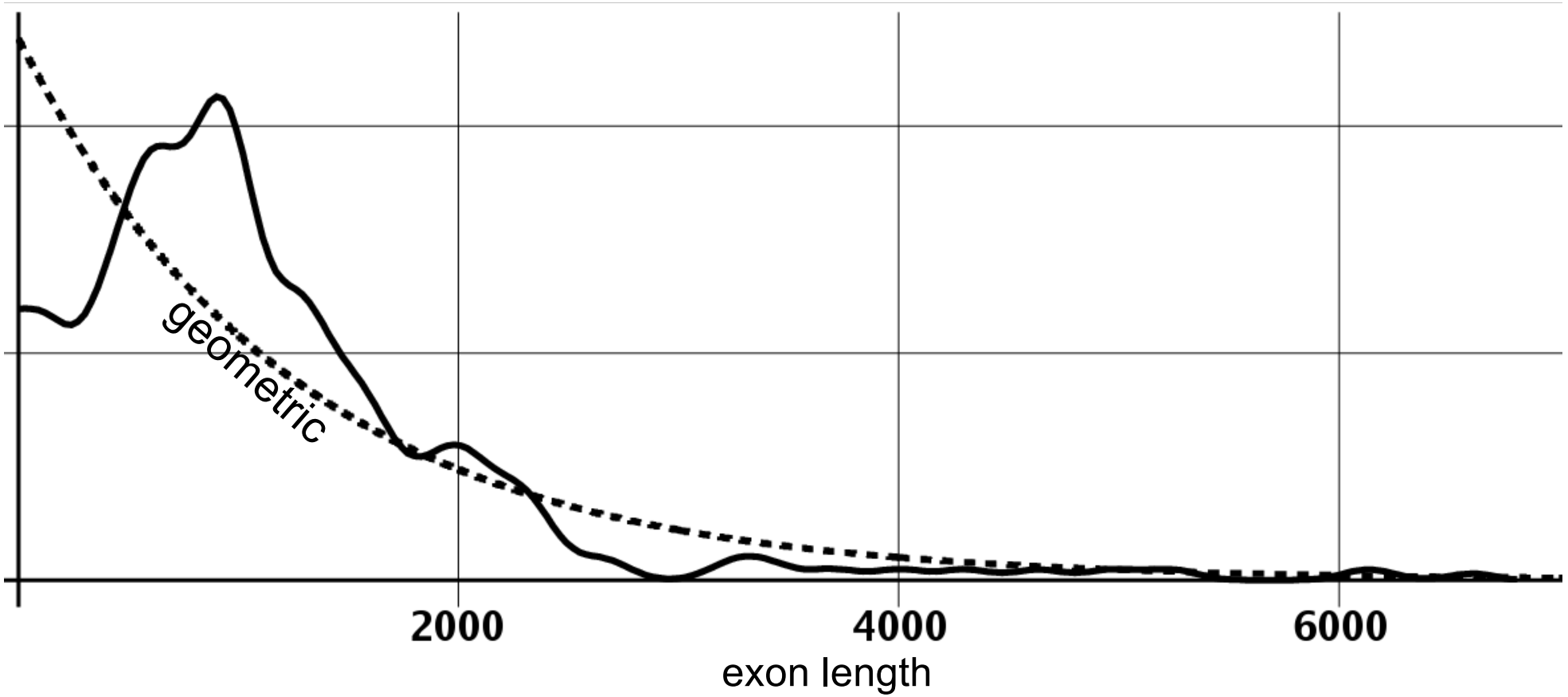
- The sequence of states is an annotation of the generated string – each nucleotide is generated in intergenic, start/stop, coding state

HMMs & Geometric Feature Lengths

$$P(x_0 \ldots x_{d-1} \mid \theta) = \left(\prod_{i=0}^{d-1} P_e(x_i \mid \theta) \right) p^{d-1}(1-p)$$

geometric distribution





Generalized HMMs Summary

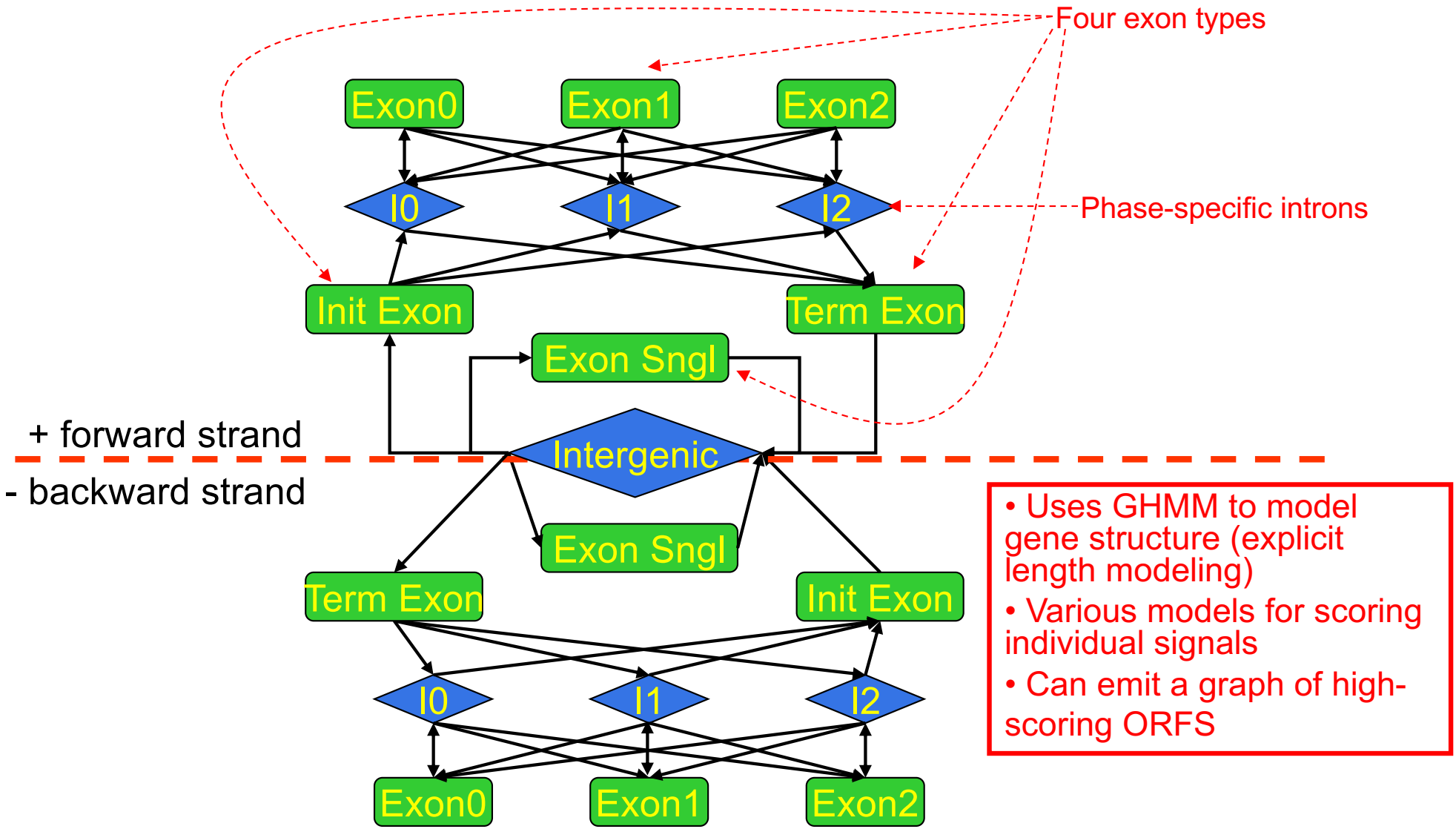
- GHMMs generalize HMMs by allowing each state to emit a subsequence rather than just a single symbol

- Whereas HMMs model all feature lengths using a geometric distribution, coding features can be modeled using an arbitrary length distribution in a GHMM

- Emission models within a GHMM can be any arbitrary probabilistic model ("submodel abstraction"), such as a neural network or decision tree

- GHMMs tend to have many fewer states => simplicity & modularity

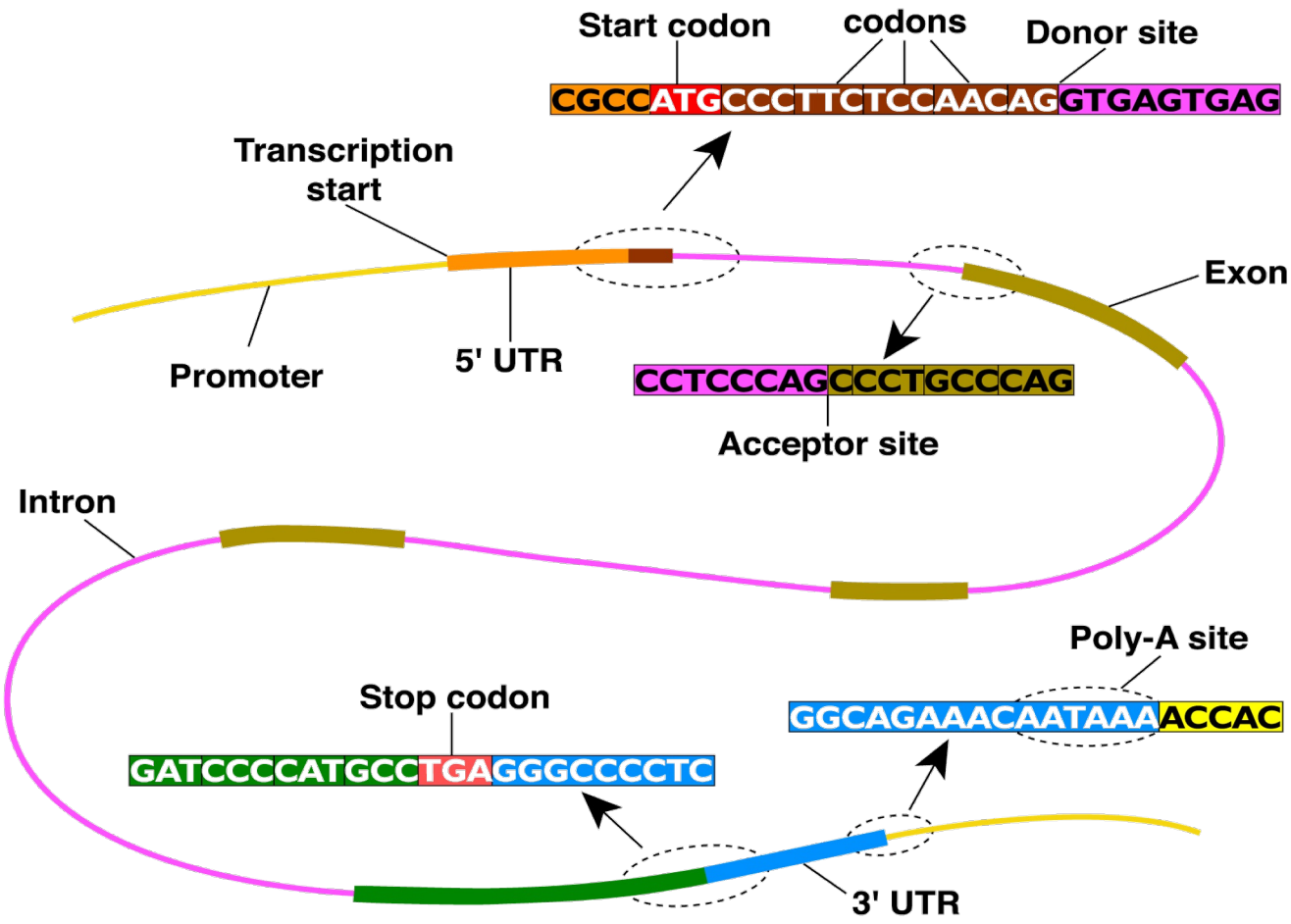
GlimmerHMM architecture



- Uses GHMM to model gene structure (explicit length modeling)
- Various models for scoring individual signals
- Can emit a graph of high-scoring ORFS

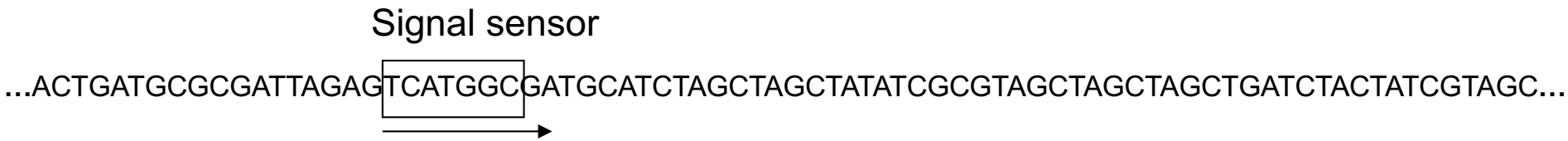
Signal Sensors

Signals – short sequence patterns in the genomic DNA that are recognized by the cellular machinery.



Identifying Signals In DNA

We slide a fixed-length model or "window" along the DNA and evaluate `score(signal)` at each point:

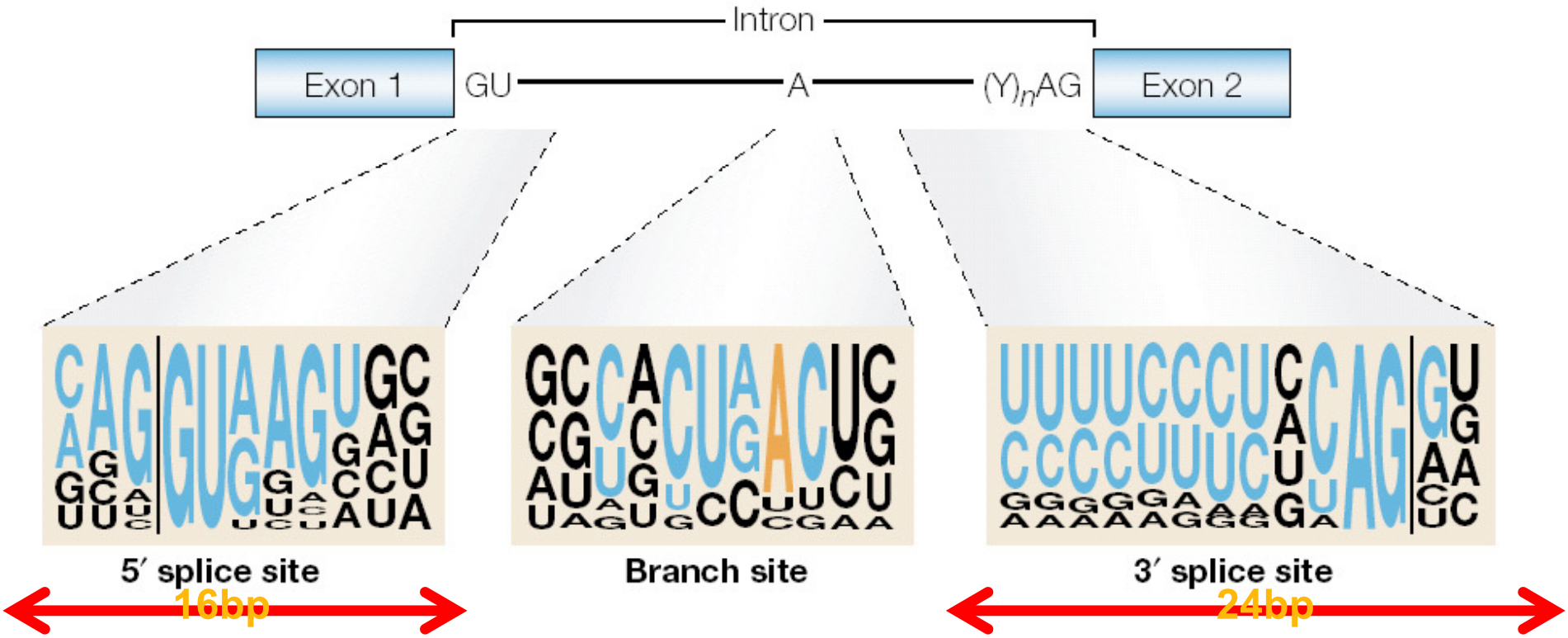


When the `score` is greater than some threshold (determined empirically to result in a desired sensitivity), we remember this position as being the potential site of a signal.

The most common signal sensor is the Position Weight Matrix:

A = 31% T = 28% C = 21% G = 20%	A = 18% T = 32% C = 24% G = 26%	**A** 100%	**T** 100%	**G** 100%	A = 19% T = 20% C = 29% G = 32%	A = 24% T = 18% C = 26% G = 32%

Splice site prediction

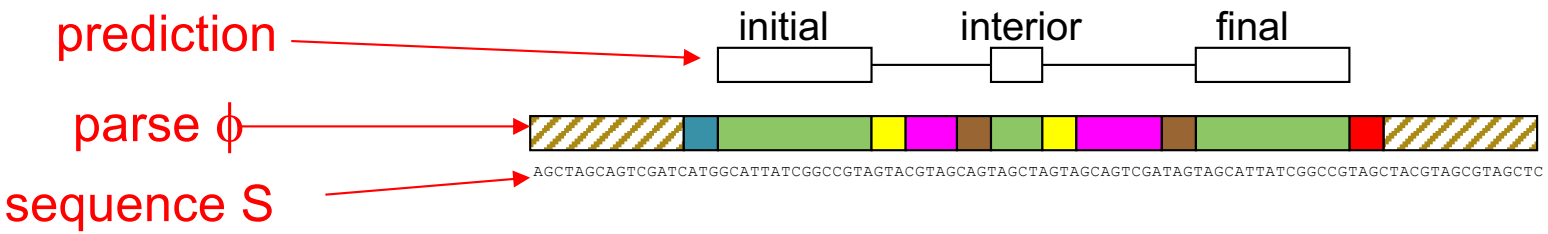


The splice site score is a combination of:

- first or second order inhomogeneous Markov models on windows around the acceptor and donor sites
- Maximal dependence decomposition (MDD) decision trees
- longer Markov models to capture difference between coding and non-coding on opposite sides of site (optional)
- maximal splice site score within 60 bp (optional)

Gene Prediction with a GHMM

Given a sequence S, we would like to determine the parse ϕ of that sequence which segments the DNA into the most likely exon/intron structure:

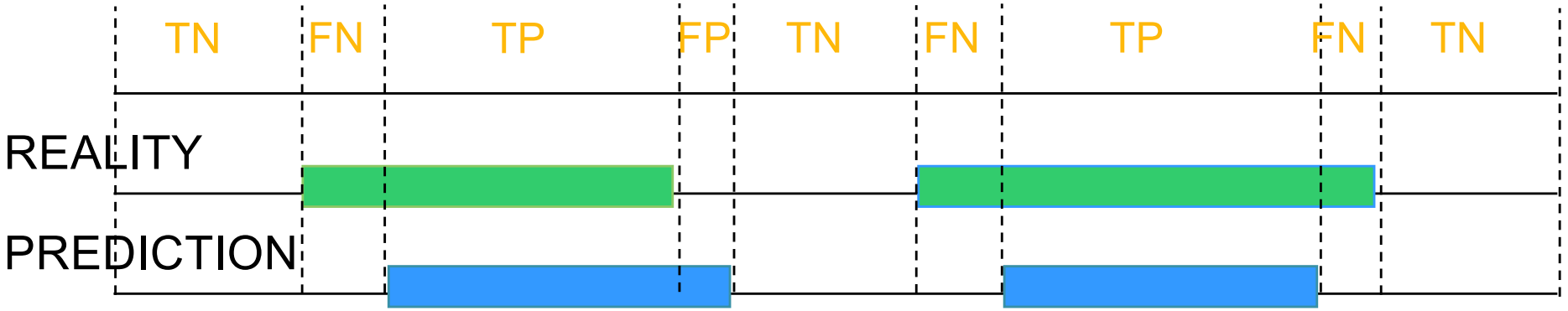


The parse ϕ consists of the coordinates of the predicted exons, and corresponds to the precise sequence of states during the operation of the GHMM (and their duration, which equals the number of symbols each state emits).

This is the same as in an HMM except that in the HMM each state emits bases with fixed probability, whereas in the GHMM each state emits an entire feature such as an exon or intron.

Evaluation of Gene Finding Programs

Nucleotide level accuracy

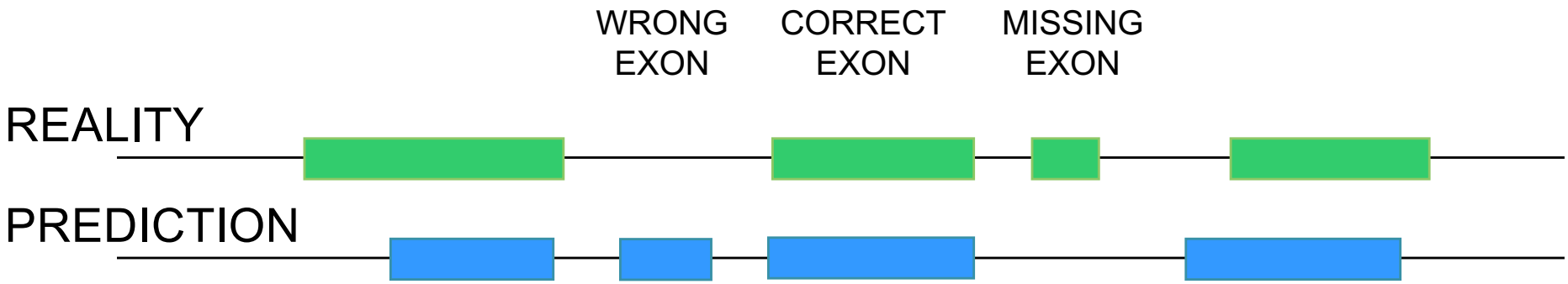


Sensitivity: $Sn = \frac{TP}{TP + FN}$ What fraction of reality did you predict?

Specificity: $Sp = \frac{TP}{TP + FP}$ What fraction of your predictions are real?

More Measures of Prediction Accuracy

Exon level accuracy



$$ExonSn = \frac{TE}{AE} = \frac{\text{number of correct exons}}{\text{number of actual exons}}$$

$$ExonSp = \frac{TE}{PE} = \frac{\text{number of correct exons}}{\text{number of predicted exons}}$$

GlimmerHMM is a high-performance ab initio gene finder

Arabidopsis thaliana test results

	Nucleotide			Exon			Gene		
	Sn	Sp	**Acc**	Sn	Sp	**Acc**	Sn	Sp	**Acc**
GlimmerHMM	97	99	**98**	84	89	**86.5**	60	61	**60.5**
SNAP	96	99	**97.5**	83	85	**84**	60	57	**58.5**
Genscan+	93	99	**96**	74	81	**77.5**	35	35	**35**

- All three programs were tested on a test data set of 809 genes, which did not overlap with the training data set of GlimmerHMM.
- All genes were confirmed by full-length Arabidopsis cDNAs and carefully inspected to remove homologues.

GlimmerHMM on human data

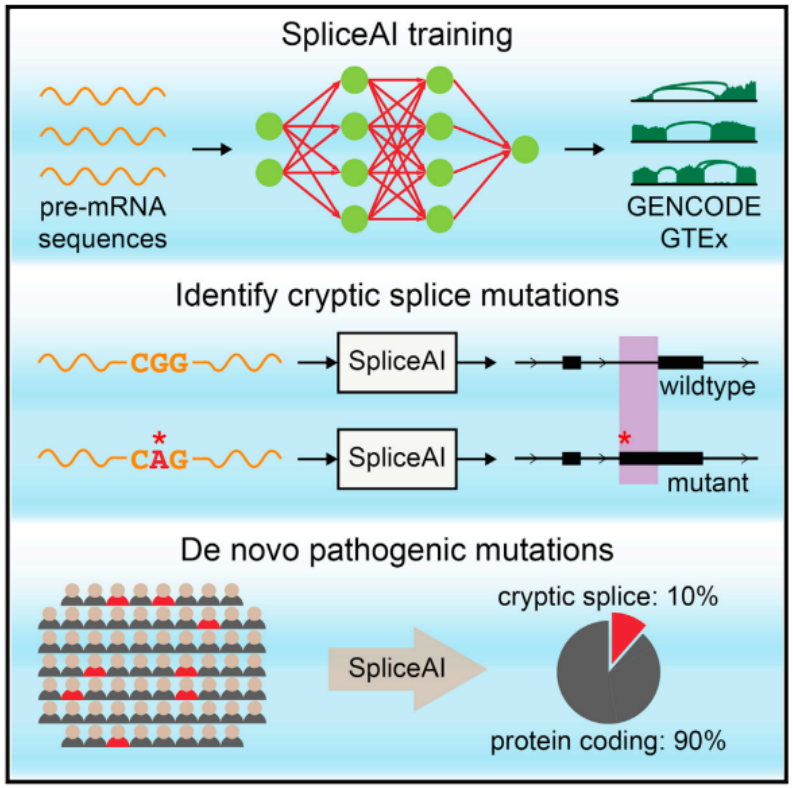
	Nuc Sens	*Nuc Spec*	*Nuc Acc*	*Exon Sens*	*Exon Spec*	*Exon Acc*	*Exact Genes*
GlimmerHMM	86%	72%	79%	72%	62%	67%	17%
Genscan	86%	68%	77%	69%	60%	65%	13%

GlimmerHMM's performace compared to Genscan on 963 human RefSeq genes selected randomly from all 24 chromosomes, non-overlapping with the training set. The test set contains 1000 bp of untranslated sequence on either side (5' or 3') of the coding portion of each gene.

Cell

Predicting Splicing from Primary Sequence with Deep Learning

Graphical Abstract



Highlights

- SpliceAI, a 32-layer deep neural network, predicts splicing from a pre-mRNA sequence
- 75% of predicted cryptic splice variants validate on RNA-seq
- Cryptic splicing may yield ~10% of pathogenic variants in neurodevelopmental disorders
- Cryptic splice variants frequently give rise to alternative splicing

Authors

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, ..., Serafim Batzoglou, Stephan J. Sanders, Kyle Kai-How Farh

Correspondence

kfarh@illumina.com

In Brief

A deep neural network precisely models mRNA splicing from a genomic sequence and accurately predicts noncoding cryptic splice mutations in patients with rare genetic diseases.

Gene Finding Overview

- Prokaryotic gene finding distinguishes real genes and random ORFs
  - Prokaryotic genes have simple structure and are largely homogenous, making it relatively easy to recognize their sequence composition
- Eukaryotic gene finding identifies the genome-wide most probable gene models (set of exons)
  - "Probabilistic Graphical Model" to enforce overall gene structure, separate models to score splicing/transcription signals
  - Accuracy depends to a large extent on the quality of the training data