Functional Genomics 3: Genes and HMMs

Michael Schatz

March 6, 2019

Lecture 12: Applied Comparative Genomics

Assignment 5: Due Monday March 11

Assignment 5: RNA-seq and differential expression

Assignment Date: Monday, March 4, 2019
Due Date: Monday, March 11, 2019 @ 11:59pm

Assignment Overview

In this assignment, you will analyze gene expression data and learn how to make several kinds of plots in the environment of your choice. (We suggest Python or R.) **Make sure to show your work/code in your writeup!** As before, any questions about the assignment should be posted to Piazza.

Question 1. Gene Annotation Preliminaries [10 pts]

Download the annotation of build 38 of the human genome from here: ftp://ftp.ensembl.org/pub/release-87/gtf/homo_sapiens/Homo_sapiens.GRCh38.87.gtf.gz

- Question 1a. How many annotated protein coding genes are on each autosome of the human genome? [Hint: Protein coding genes will have "gene" in the 3rd column, and contain the following text: gene_biotype "protein_coding"]
- Question 1b. What is the maximum, minimum, mean, and standard deviation of the span of protein coding genes? [Hint: use the genes identified in 1b]
- Question 1c. What is the maximum, minimum, mean, and standard deviation in the number of exons for protein coding genes? [Hint: you should separately consider each isoform for each protein coding gene]

Question 2. Time Series [10 pts]

This file contains pre-normalized expression values for 100 genes over 10 time points. Most genes have a stable background expression level, but some special genes show increased expression over the timecourse and some show decreased expression.

a. Cluster the genes using an algorithm of your choice. Which genes show increasing expression and which genes show decreasing expression, and how did you determine this? What is the background expression level (numerical value) and how did you determine this? [Hint: K-means and hierarchical clustering are common clustering algorithms you could try.]

b. Calculate the first two principal components of the expression matrix. Show the plot and color the points based on their cluster from part (a). Does the PC1 axis, PC2 axis, neither, or both correspond to the clustering?

c. Create a heatmap of the expression matrix. Order the genes by cluster, but keep the time points in numerical order.

Question 3. Sampling Simulation [10 pts]

A typical human cell has ~250,000 transcripts, and a typical bulk RNA-seq experiment may involve millions of cells. Consequently in an RNAseq experiment you may start with trillions of RNA molecules, although your sequencer will only give a few million to billions of reads. Therefore your RNAseq experiment will be a small sampling of the full composition. We hope the sequences will be a representative sample of the total population, but if your sample is very unlucky or biased it may not represent the true distribution. We will explore this concept by sampling a small subset of transcripts (1000 to 5000) out of a much larger set (1M) so that you can evaluate this bias.

In data1.txt with 1,000,000 lines we provide an abstraction of RNA-seq data where normalization has been performed and the number of times a gene name occurs corresponds to the number of transcripts sequenced.

a. Randomly sample 1000 rows. Do this simulation 10 times and record the relative abundance of each of the 15 genes. Plot the mean vs. variance.

b. Do the same sampling experiment but sample 5000 rows each time. Again plot the mean vs. variance.

c. Is the variance greater in (a) or (b)?, and explain why. What is the relationship between abundance and variance?

Proposal: Due Friday March 15

Project Proposal

Assignment Date: Wednesday March 6, 2019
Due Date: Friday, March 15, 2019 @ 11:59pm

Review the Project Ideas page

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

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)

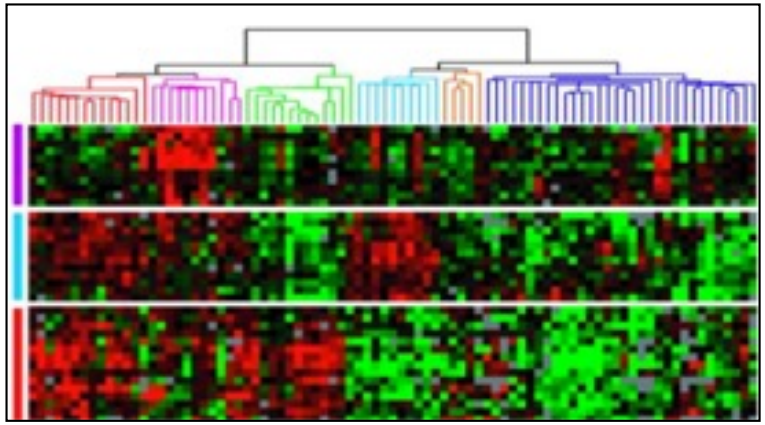
Submit the proposal as a single page PDF on GradeScope (each team member should submit the same PDF). After submitting your proposal, we will schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner 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

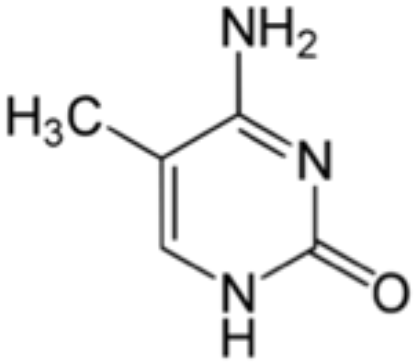
Please use Piazza to coordinate proposal plans!

*-seq in 4 short vignettes

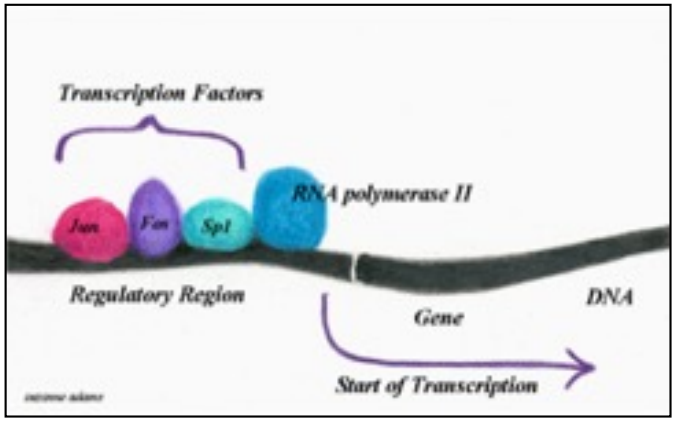
RNA-seq



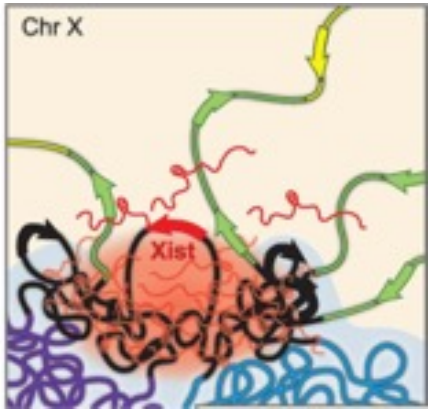
Methyl-seq



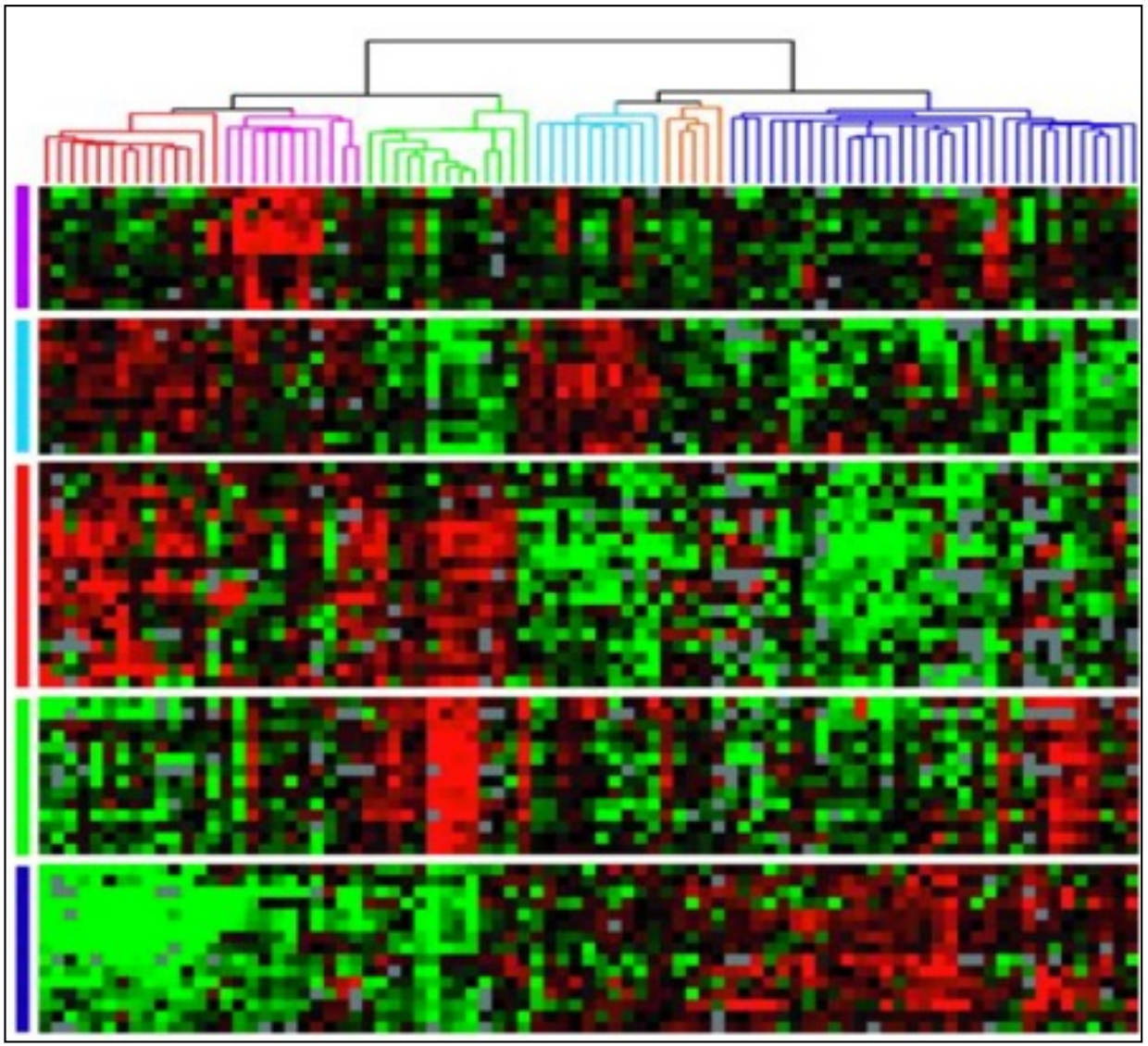
ChIP-seq



Hi-C

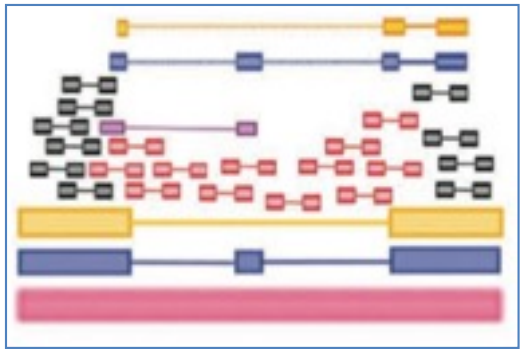


RNA-seq



Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.
Sørlie et al (2001) *PNAS*. 98(19):10869-74.

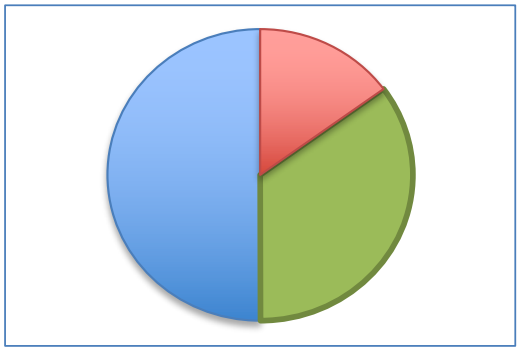
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

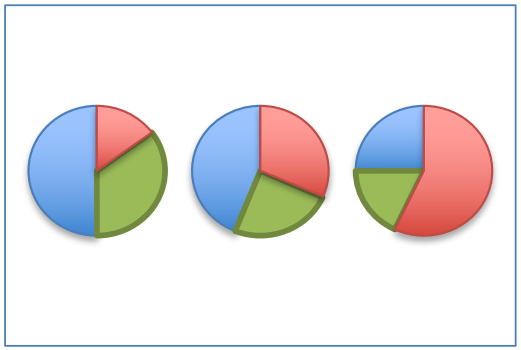
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. TPM)

Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

RNA-seq differential expression studies: more sequence or more replication?
Liu et al (2013) *Bioinformatics.* doi:10.1093/bioinformatics/btt688

Goal: Genome Annotations

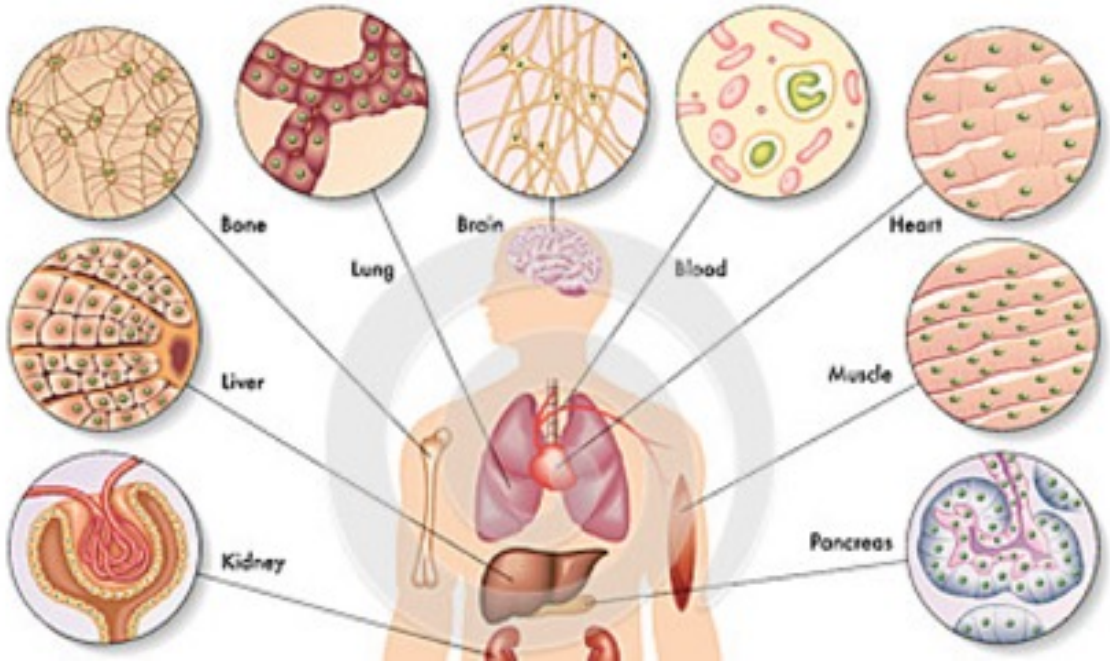
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaat
gcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatgct
aatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg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

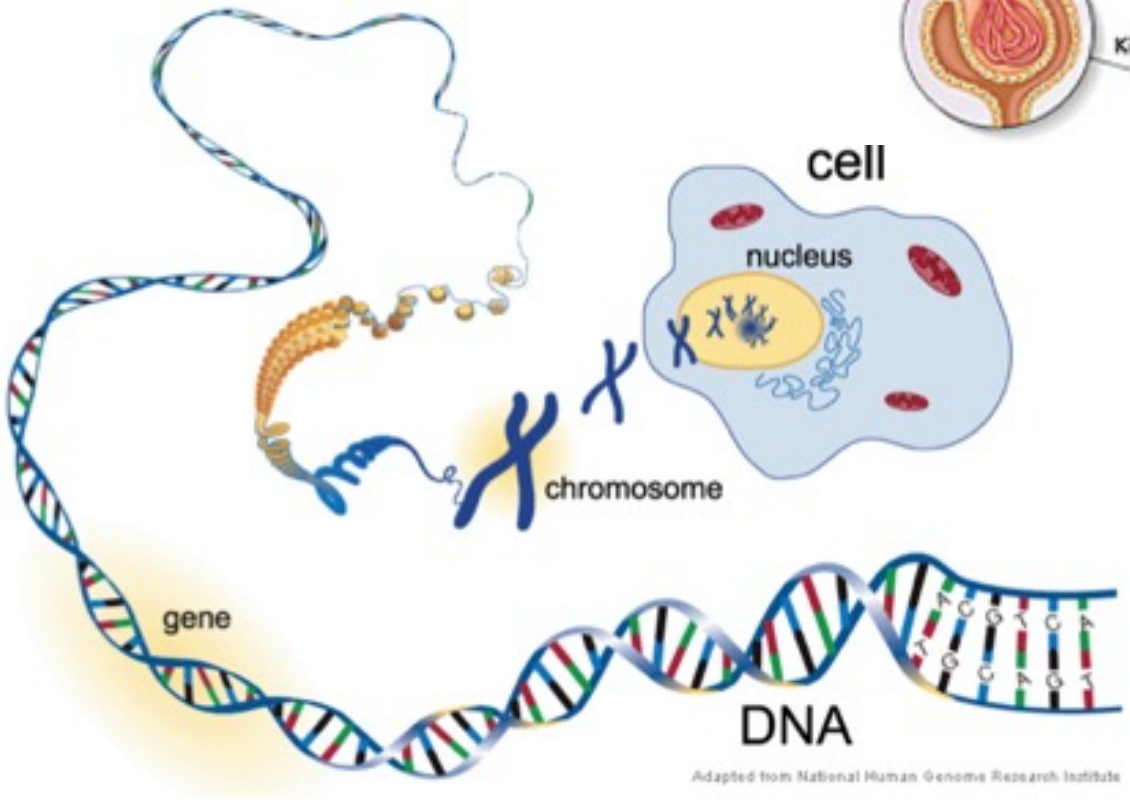
Goal: Genome Annotations

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

Why Genes?

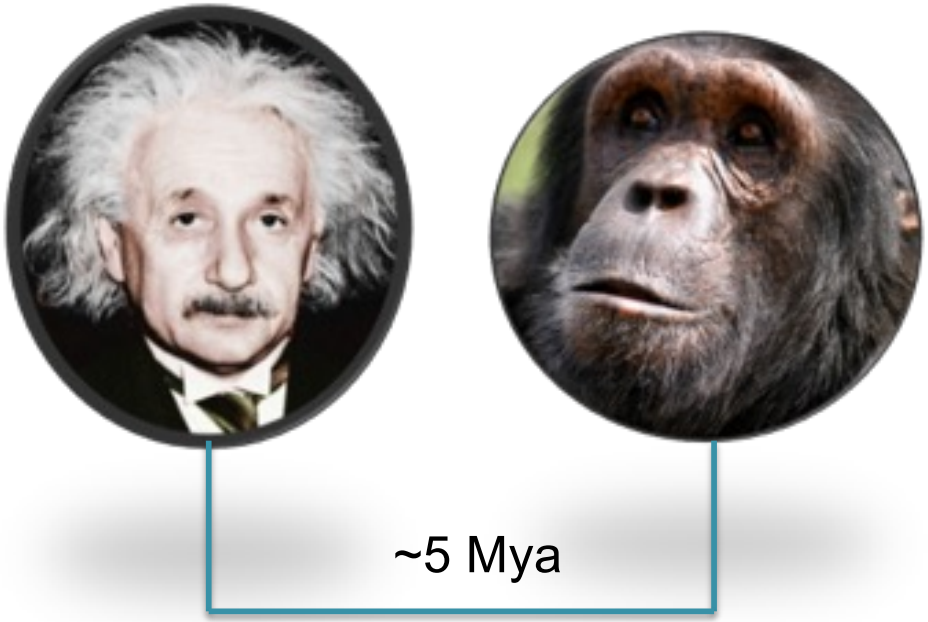
Each cell of your body contains an exact copy of your 3 billion base pair genome.





Your body has a few hundred (thousands?) major cell types, largely defined by the gene expression patterns

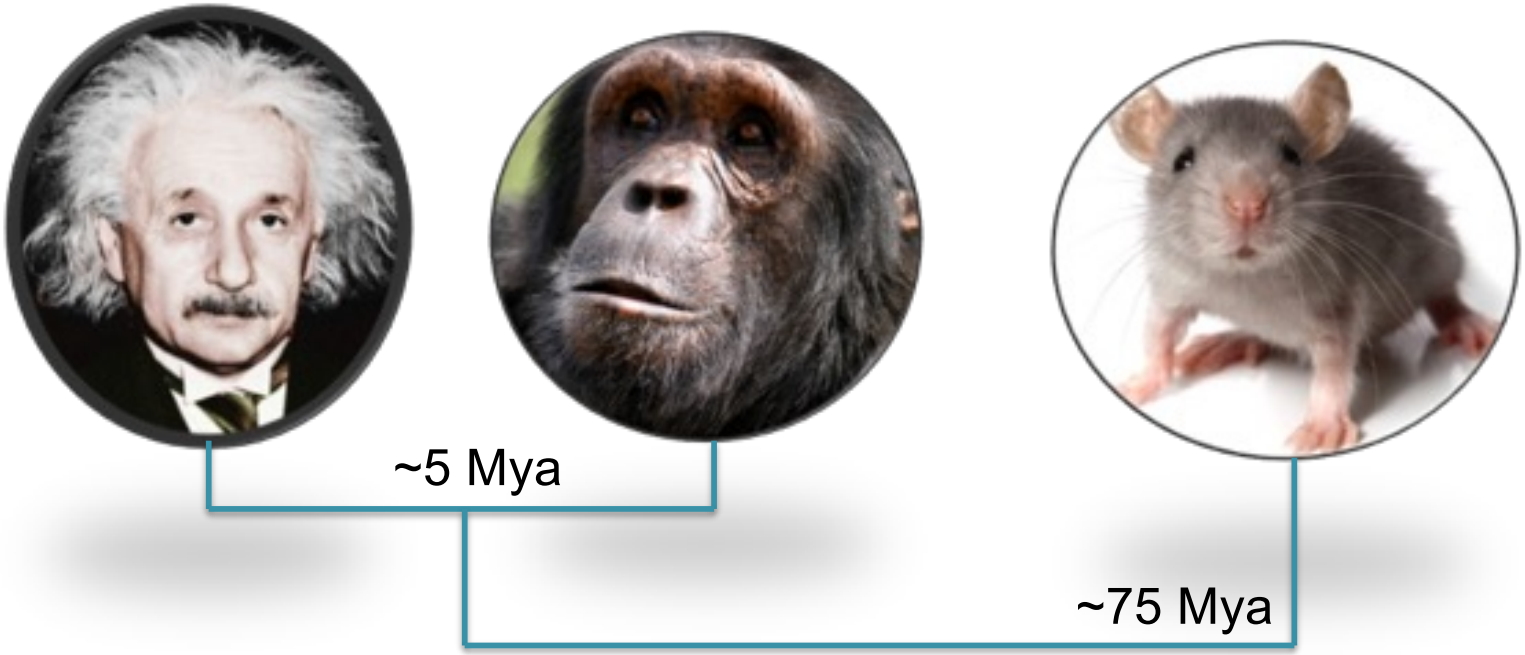
Human Evolution



- Humans and chimpanzees shared a common ancestor ~5-7 million years ago (Mya)
- Single-nucleotide substitutions occur at a mean rate of 1.23% but ~4% overall rate of mutation: comprising ~35 million single nucleotide differences and ~90 Mb of insertions and deletions
- Orthologous proteins in human and chimpanzee are extremely similar, with ~29% being identical and the typical orthologue differing by only two amino acids, one per lineage

Initial sequence of the chimpanzee genome and comparison with the human genome (2005) *Nature 437, 69-87 doi:10.1038/nature04072*

Human Evolution

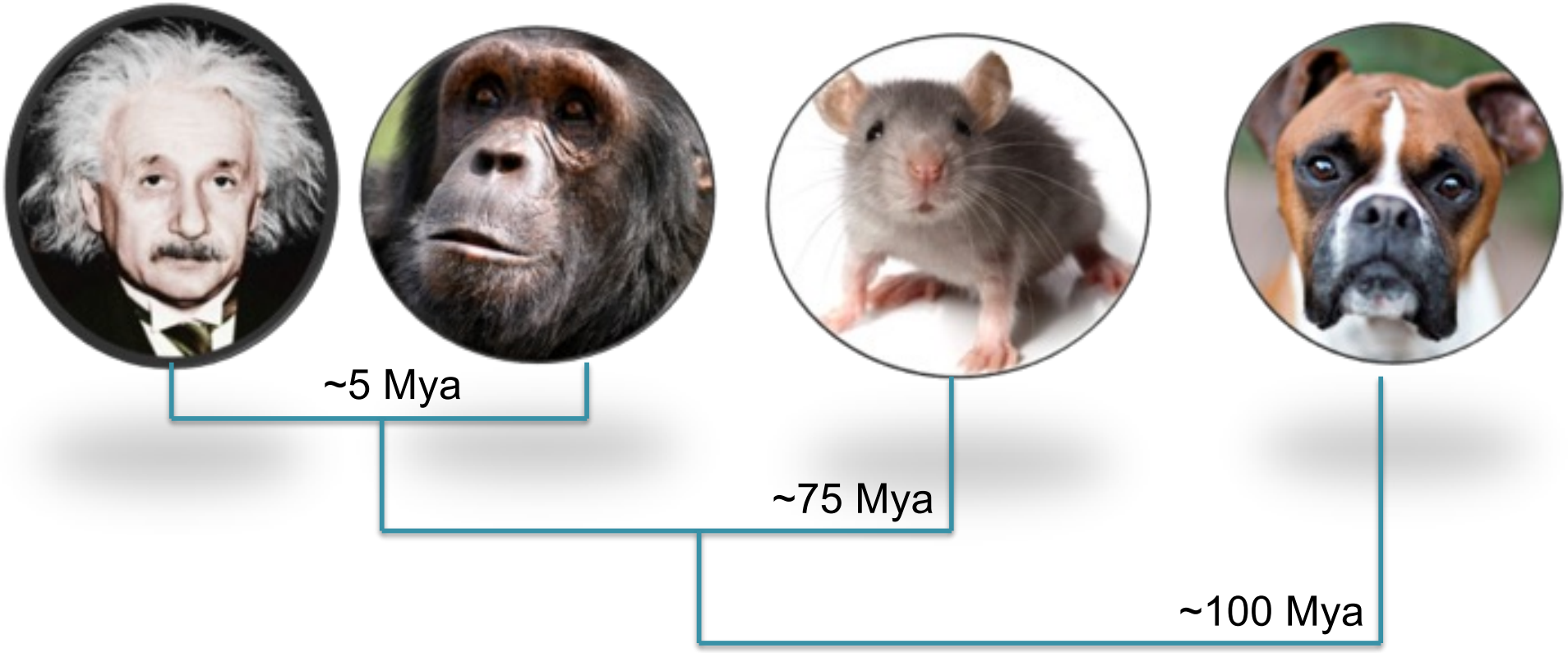


"In the roughly 75 million years since the divergence of the human and mouse lineages, the process of evolution has altered their genome sequences and caused them to diverge by ***nearly one substitution for every two nucleotides***"

"The mouse and human genomes each seem to contain about 30,000 protein-coding genes. These refined estimates have been derived from both new evidence-based analyses that produce larger and more complete sets of gene predictions, and new de novo gene predictions that do not rely on previous evidence of transcription or homology. The proportion of mouse genes with a single identifiable orthologue in the human genome seems to be approximately 80%. ***The proportion of mouse genes without any homologue currently detectable in the human genome (and vice versa) seems to be less than 1%.***"

Initial sequencing and comparative analysis of the mouse genome
Chinwalla et al (2002) *Nature.* 420, 520-562 doi:10.1038/nature01262

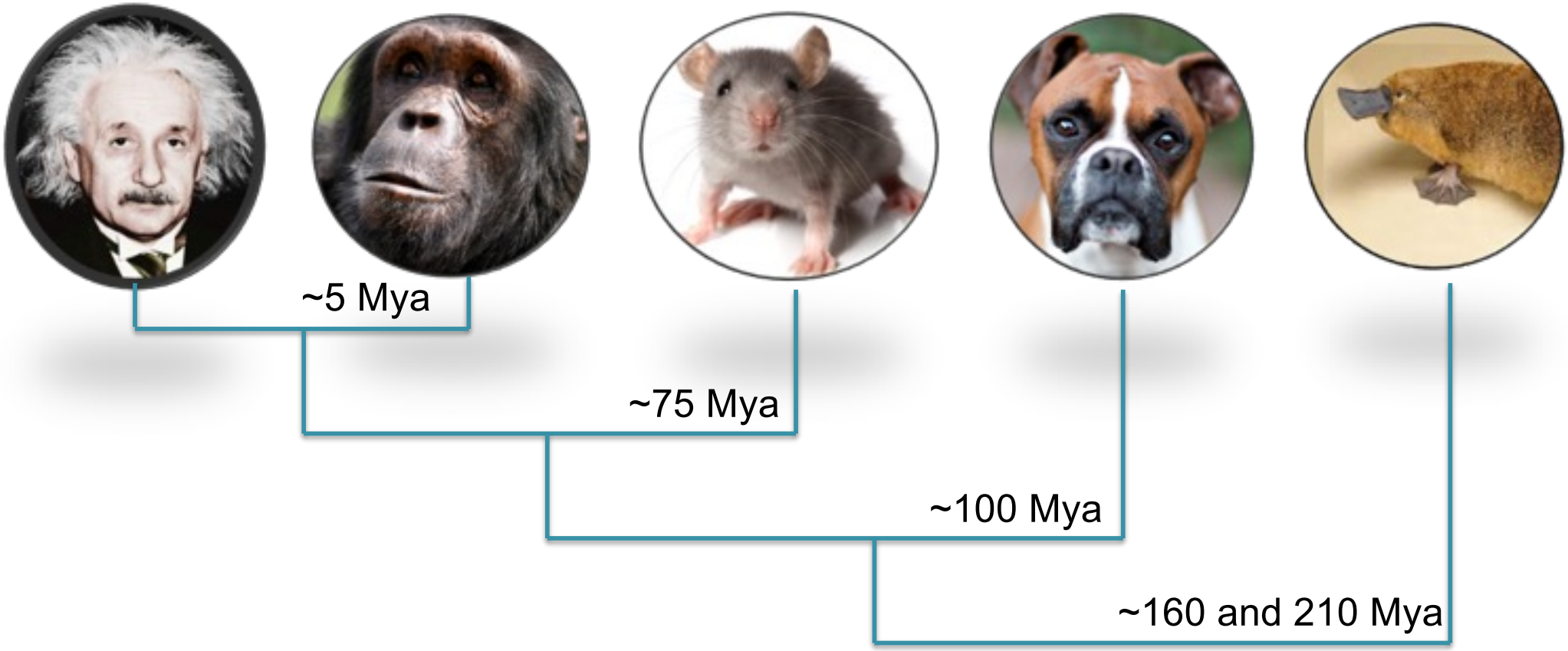
Human Evolution



"We generated gene predictions for the dog genome using an evidence-based method (see Supplementary Information). The resulting collection contains ***19,300 dog gene predictions, with nearly all being clear homologues of known human genes***. The dog gene count is substantially lower than the ~22,000-gene models in the current human gene catalogue (EnsEMBL build 26). For many predicted human genes, we find no convincing evidence of a corresponding dog gene. Much of the excess in the human gene count is attributable ***to spurious gene predictions in the human genome***"

Genome sequence, comparative analysis and haplotype structure of the domestic dog
Lindblad-Toh et al (2005) *Nature. 438, 803-819 doi:10.1038/nature04338*

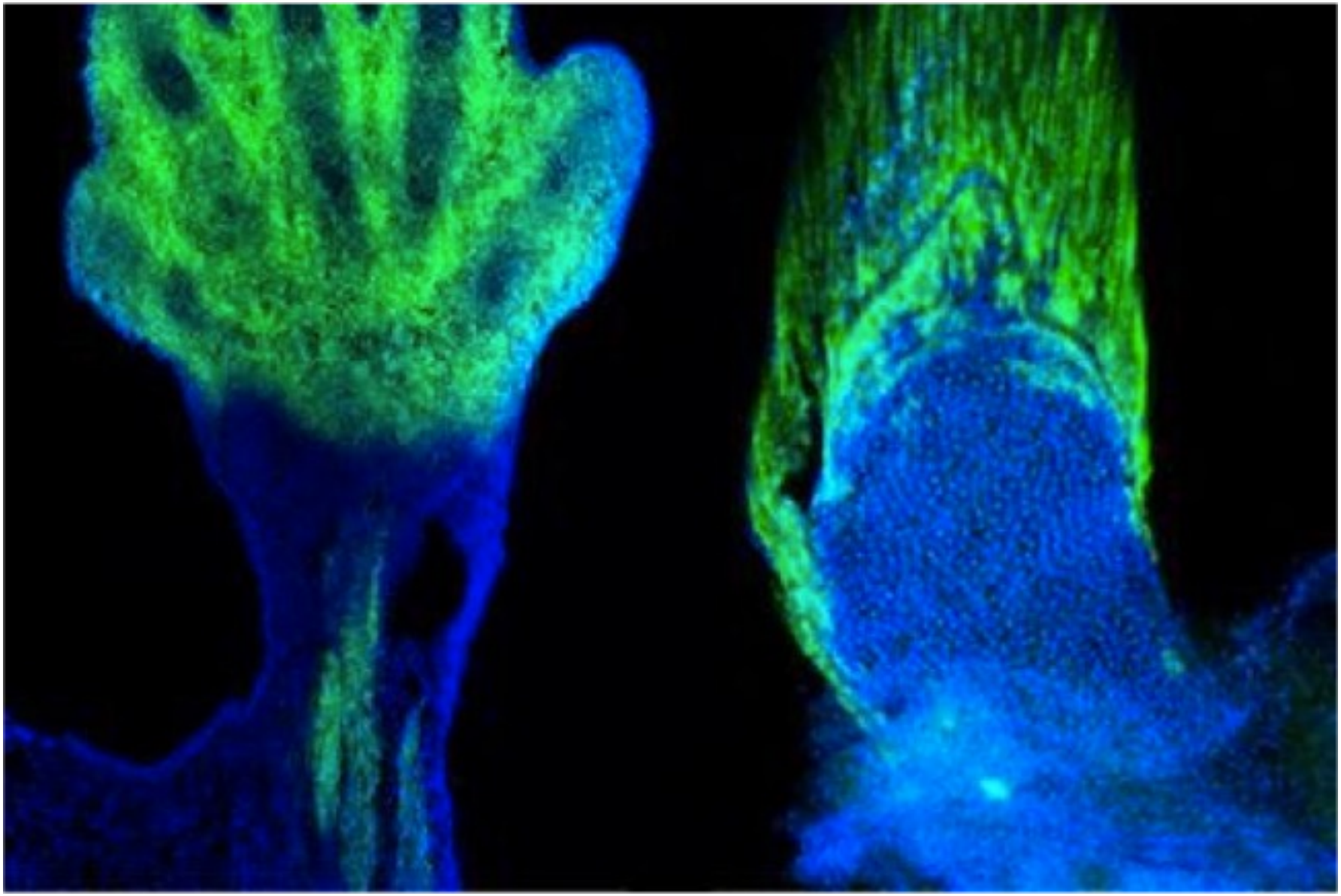
Human Evolution



As expected, the majority of platypus genes (82%; 15,312 out of 18,596) have orthologues in these five other amniotes (Supplementary Table 5). The remaining 'orphan' genes are expected to primarily reflect rapidly evolving genes, for which no other homologues are discernible, erroneous predictions, and true lineage-specific genes that have been lost in each of the other five species under consideration.

Genome analysis of the platypus reveals unique signatures of evolution
(2008) *Nature.* 453, 175-183 doi:10.1038/nature06936

Human Evolution

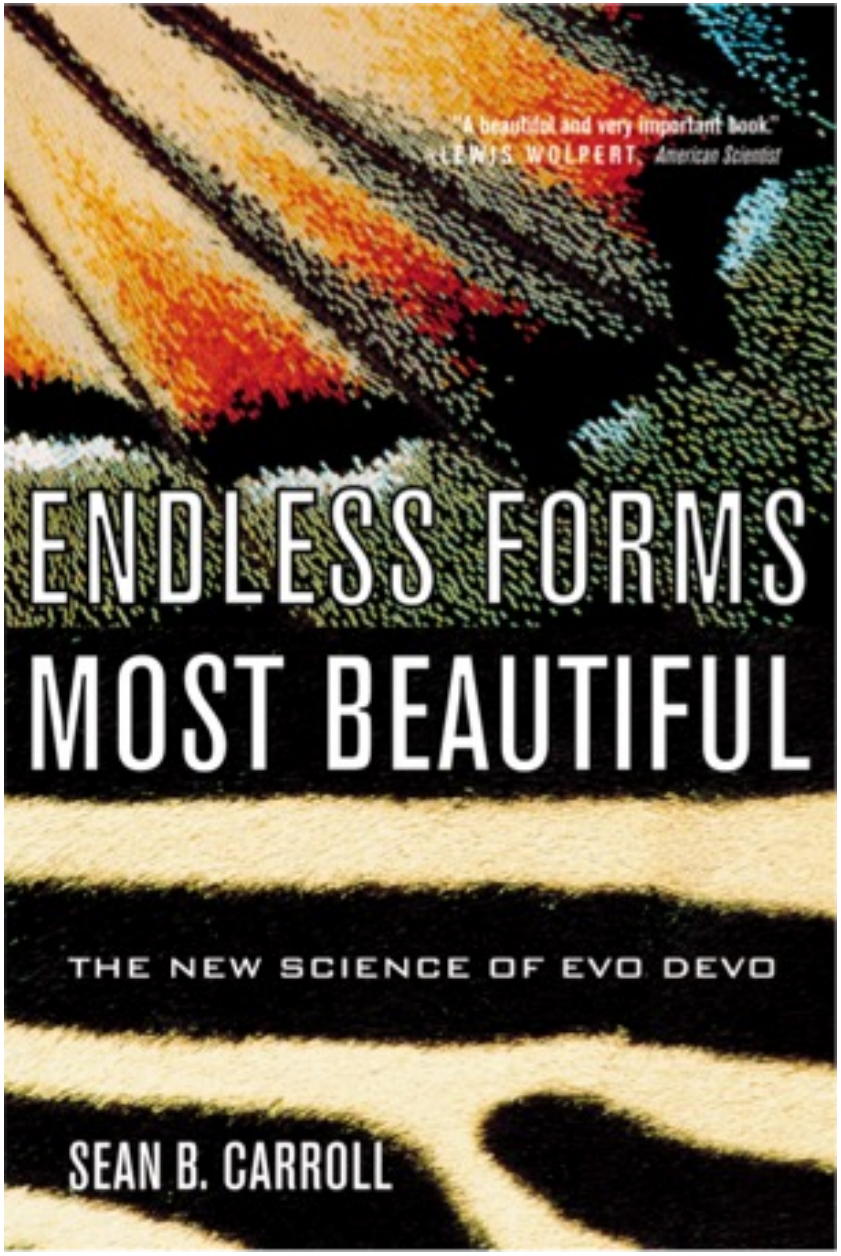


Digits and fin rays share common developmental histories
Nakamura et al (2016) *Nature.* 537, 225–228. doi:10.1038/nature19322

More Information

“Anything found to be true of E. coli must also be true of elephants”

-Jacques Monod



Outline

1. Experimental: RNAseq
2. Homology: Alignment to other genomes
3. Prediction: "Gene Finding"

Outline

1. Experimental: RNAseq
   - ☺ Direct evidence for expression!
     - Including novel genes within a species
   - ☹ Typical tissues only express 25% to 50% of genes
     - Many genes are restricted to very particular cell types, developmental stages, or stress conditions
     - Our knowledge of alternative splicing is very incomplete
   - ☹ Can resolve gene structure, but nothing about gene function
     - Co-expression is sometimes a clue, but often incomplete

Outline

2. Homology: Alignment to other genomes

- :-/ Indirect evidence for expression
  - Works well for familiar species, but more limited for unexplored clades
  - Relatively few false positives, but many false negatives
- ☺ Universal across tissues (and species)
  - Proteins often have highly conserved domains, whereas genome/transcript may have many mutations (especially "wobble" base)
- :-/ Transfer gene function across species
  - Reciprocal best blast hit a widely used heuristic
  - Often works, but examples where single base change leads to opposite function

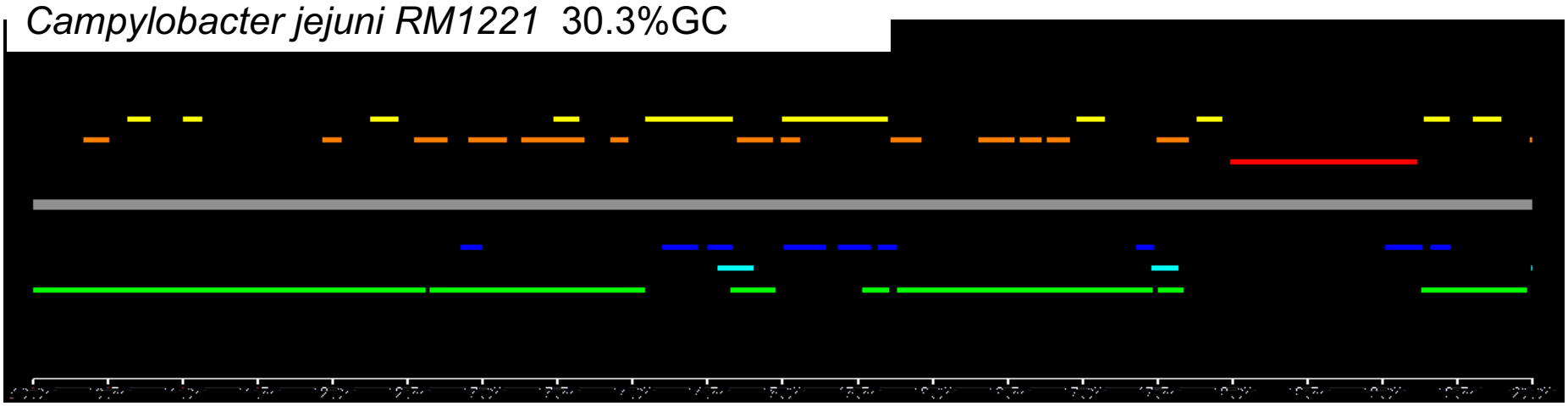
Outline

1. Experimental: RNAseq
2. Homology: Alignment to other genomes
3. Prediction: "Gene Finding"

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 School of Medicine

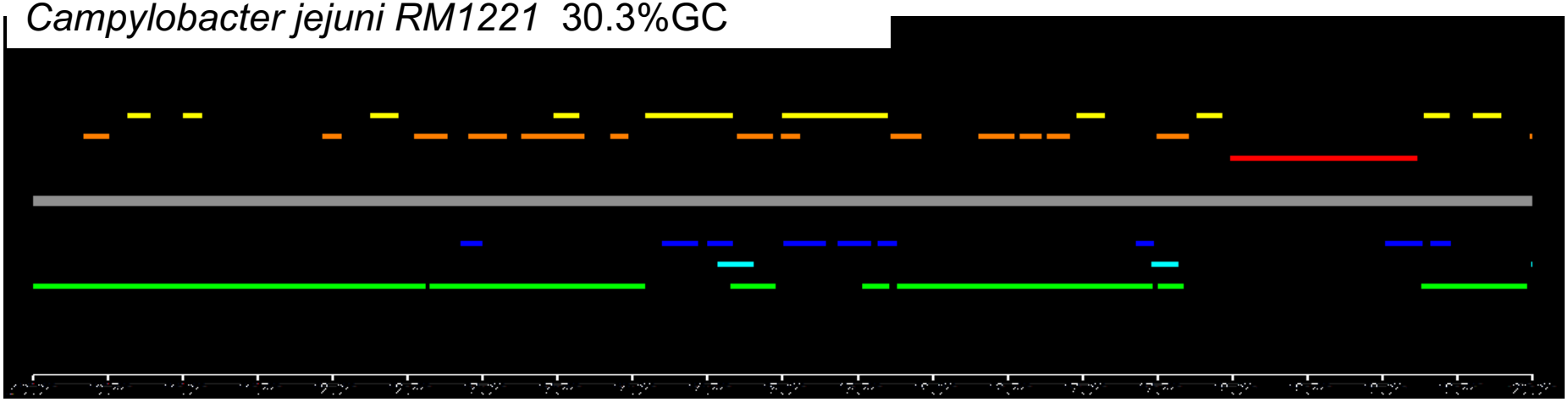


All ORFs longer than 100bp on both strands shown
- color indicates reading frame

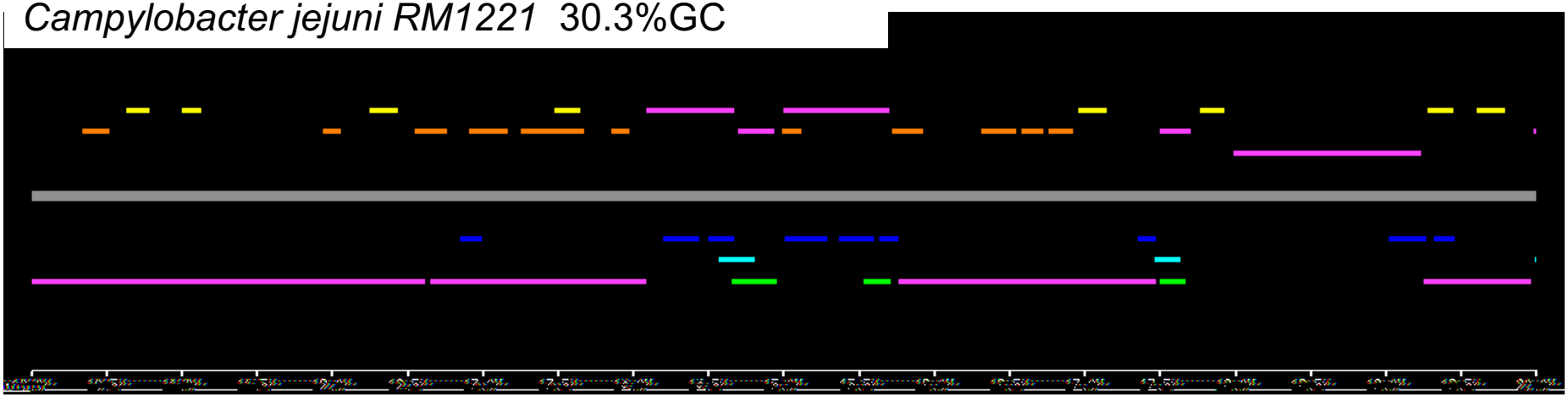
Note the low GC content
- many A+T -> many stop codons (TAA/TAG/TGA)

All genes are ORFs but not all ORFs are genes
- Longest ORFs likely to be protein-coding genes

Campylobacter jejuni RM1221 30.3%GC



Campylobacter jejuni RM1221 30.3%GC



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

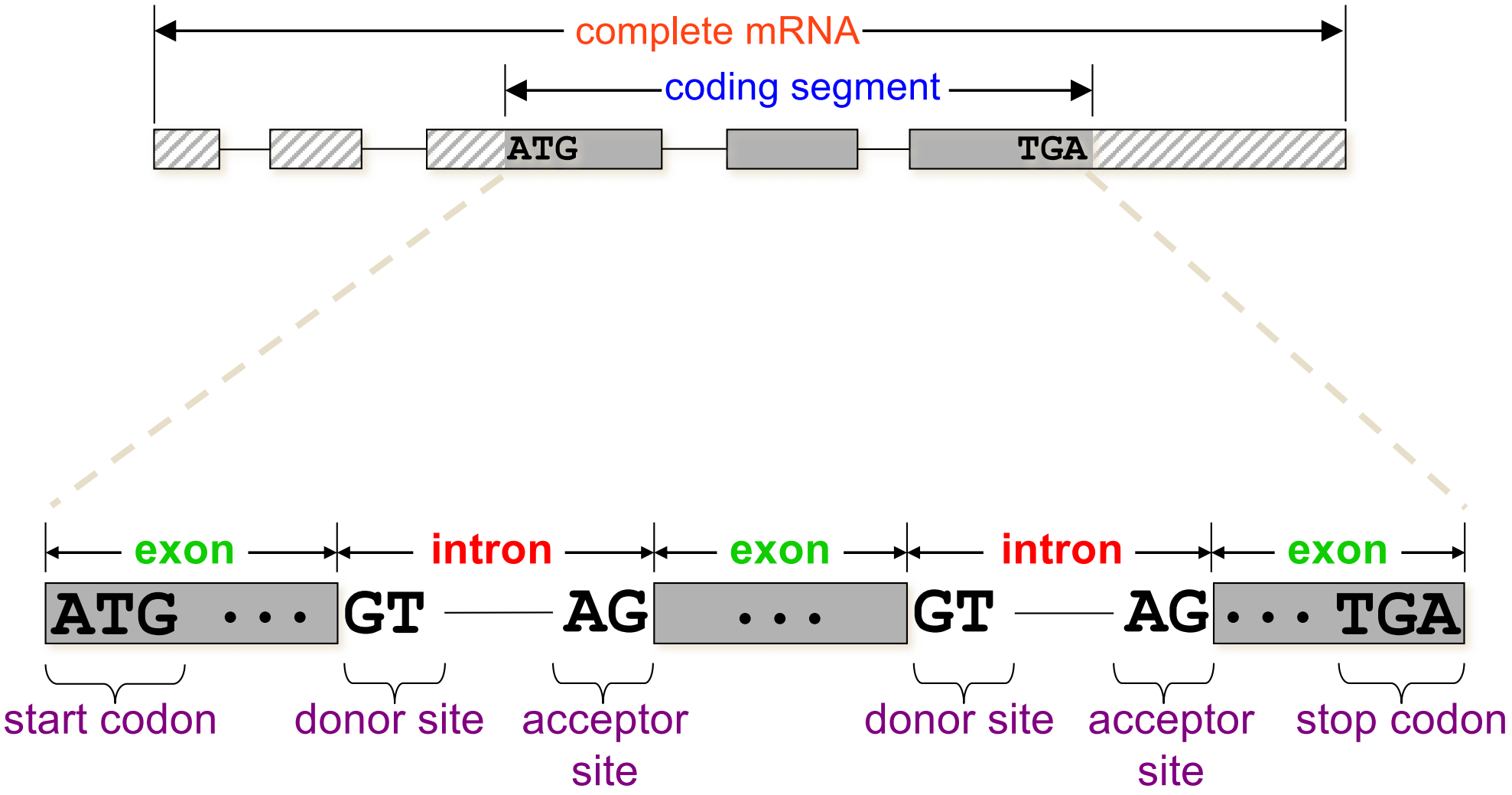
Overview of Eukaryotic Gene Prediction

CBB 231 / COMPSCI 261

W.H. Majoros

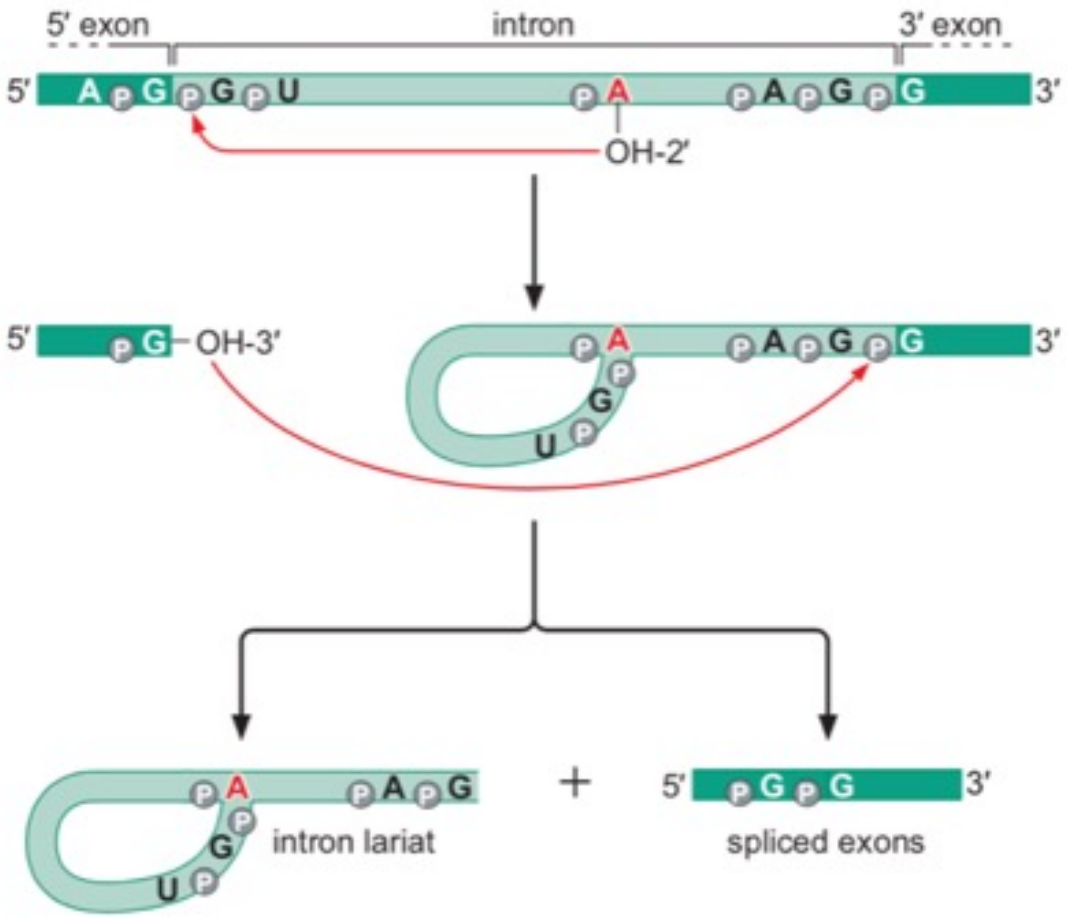
Duke
UNIVERSITY

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.

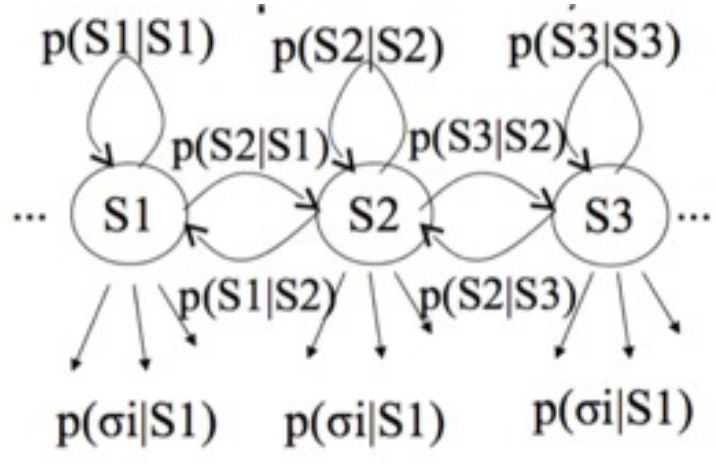
Eukaryotic Gene Splicing



https://www.youtube.com/watch?v=FVuAwBGw_pQ

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



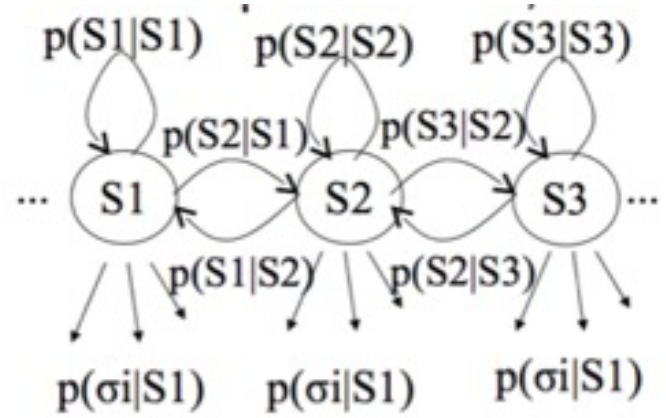
  - A set of emission characters
    - E={H,T} for coin tossing
    - E={1,2,3,4,5,6} for dice tossing
    - E={A,C,G,T} for DNA

  - 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
    - $P(s_i=Exon \mid s_{i-1}=Intergenic)$, …

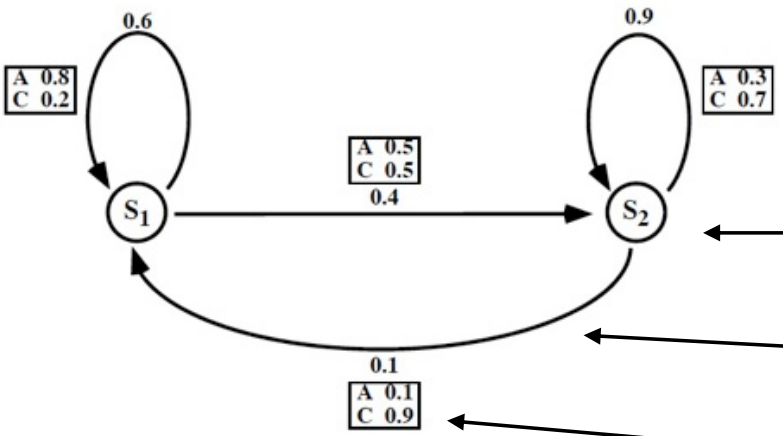
Why Hidden?

- 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

HMM Example – Two State DNA Model



Symbol emission prob

States

State transition probs.

Observation Symbols

Observation Sequence

AAACCACACCCACCCACCCAAAACCACA

State Sequence

11122122222221111122211112222

Motivation: Given a sequence of As and Ts, can you tell which states are being used?

(Note the State Sequence must be inferred!)

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?

Three classic HMM problems

1. **Evaluation**: given a model and an output sequence, what is the probability that the model generated that output?

- To answer this, we consider all possible paths through the model

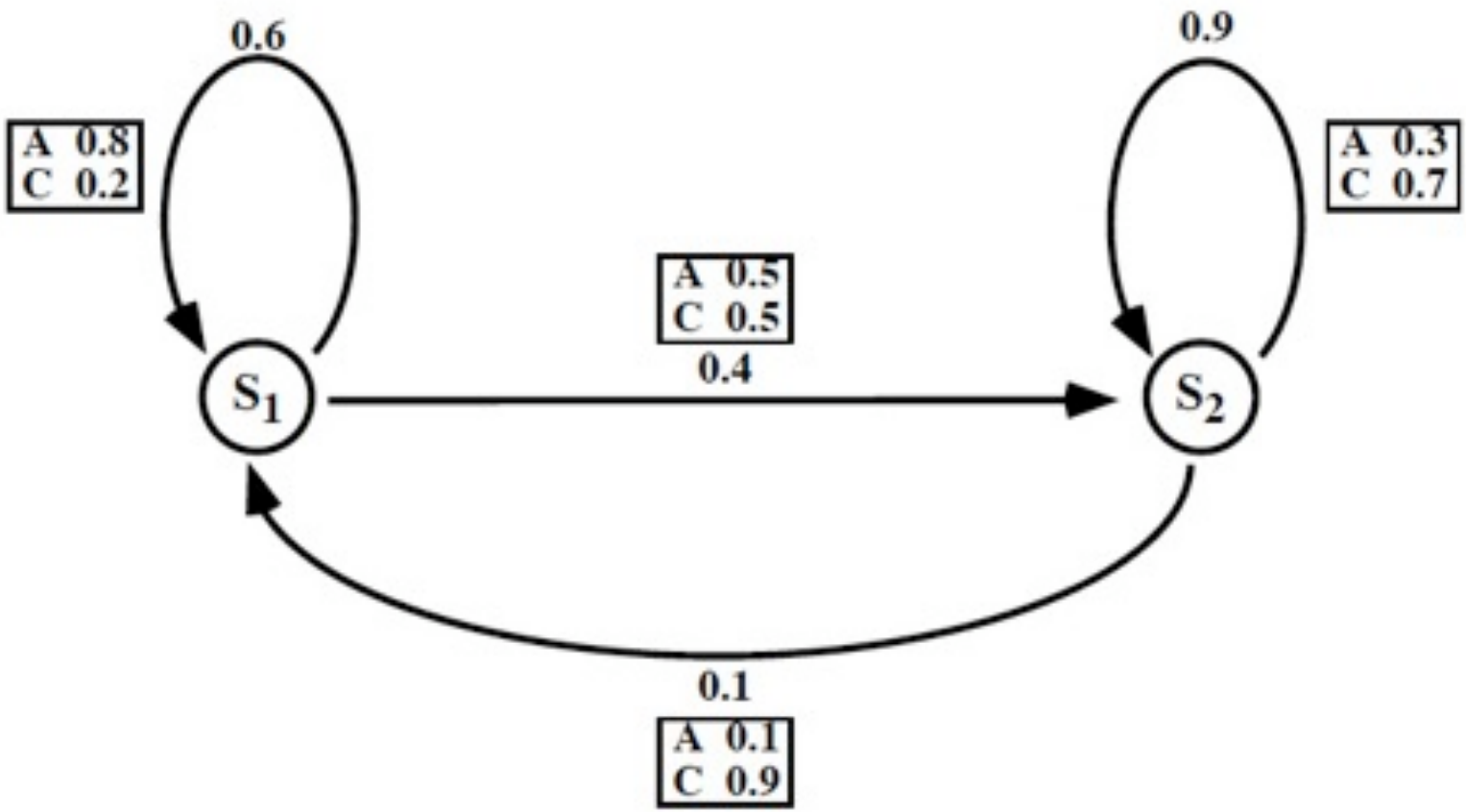
- Example: we might have a set of HMMs representing protein families -> pick the model with the best score

Solving the Evaluation problem: The Forward algorithm

- To solve the Evaluation problem (probability that the model generated the sequence), we use the HMM and the data to build a *trellis*

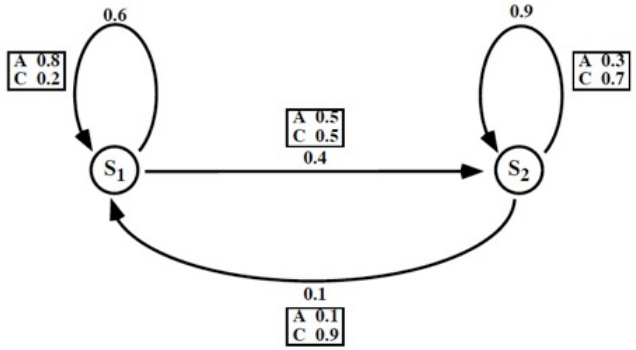
- Filling in the trellis will give tell us the probability that the HMM generated the data by finding all possible paths that could do it
  - Especially useful to evaluate from which models, a given sequence is most likely to have originated

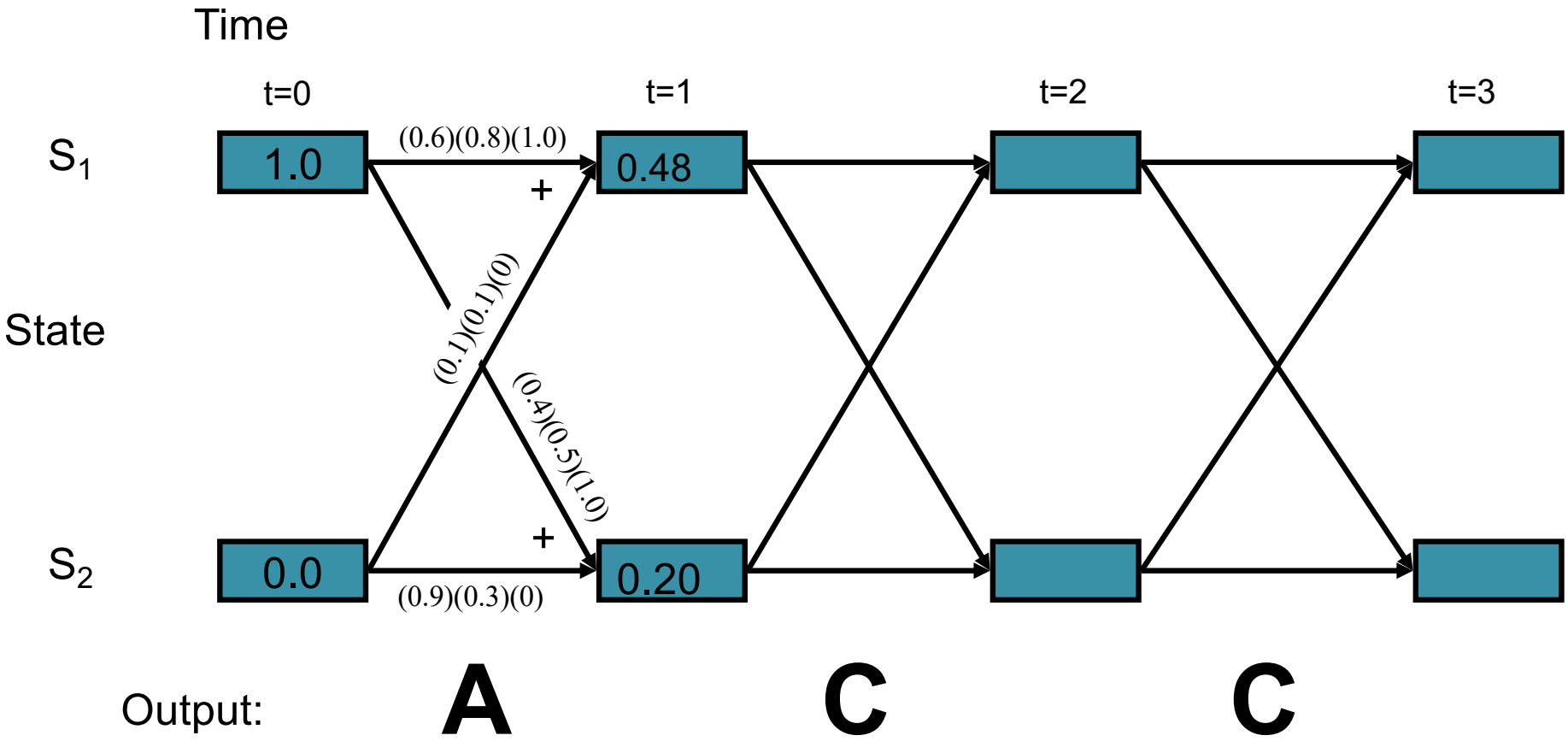
Our sample HMM



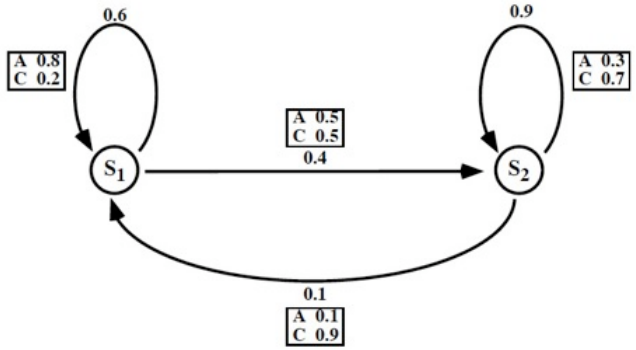
Let S_1 be initial state, S_2 be final state

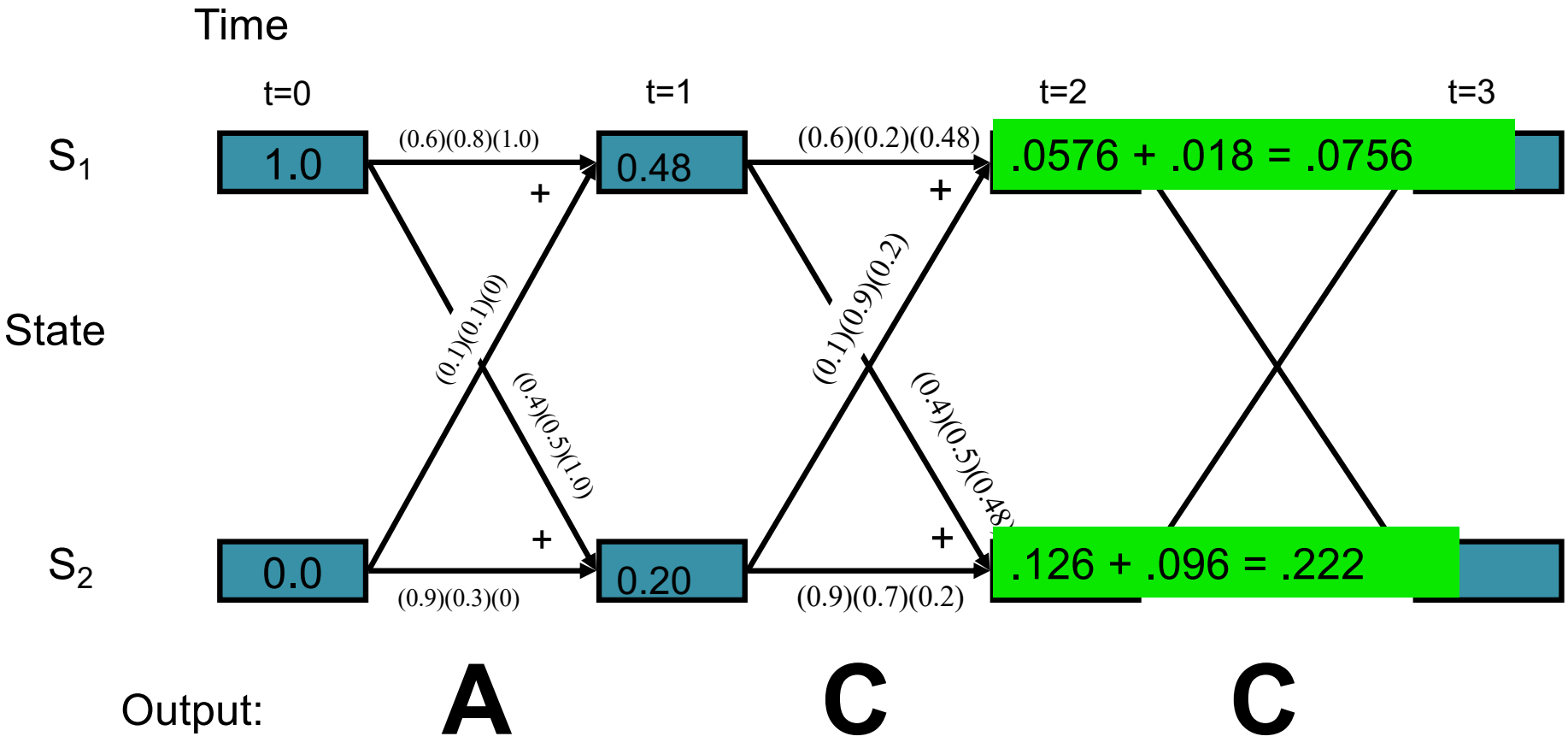
A trellis for the Forward Algorithm



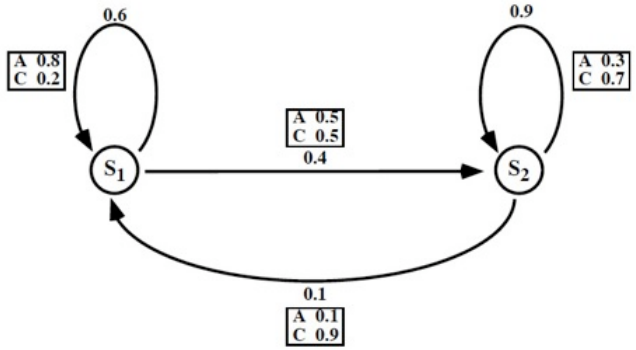


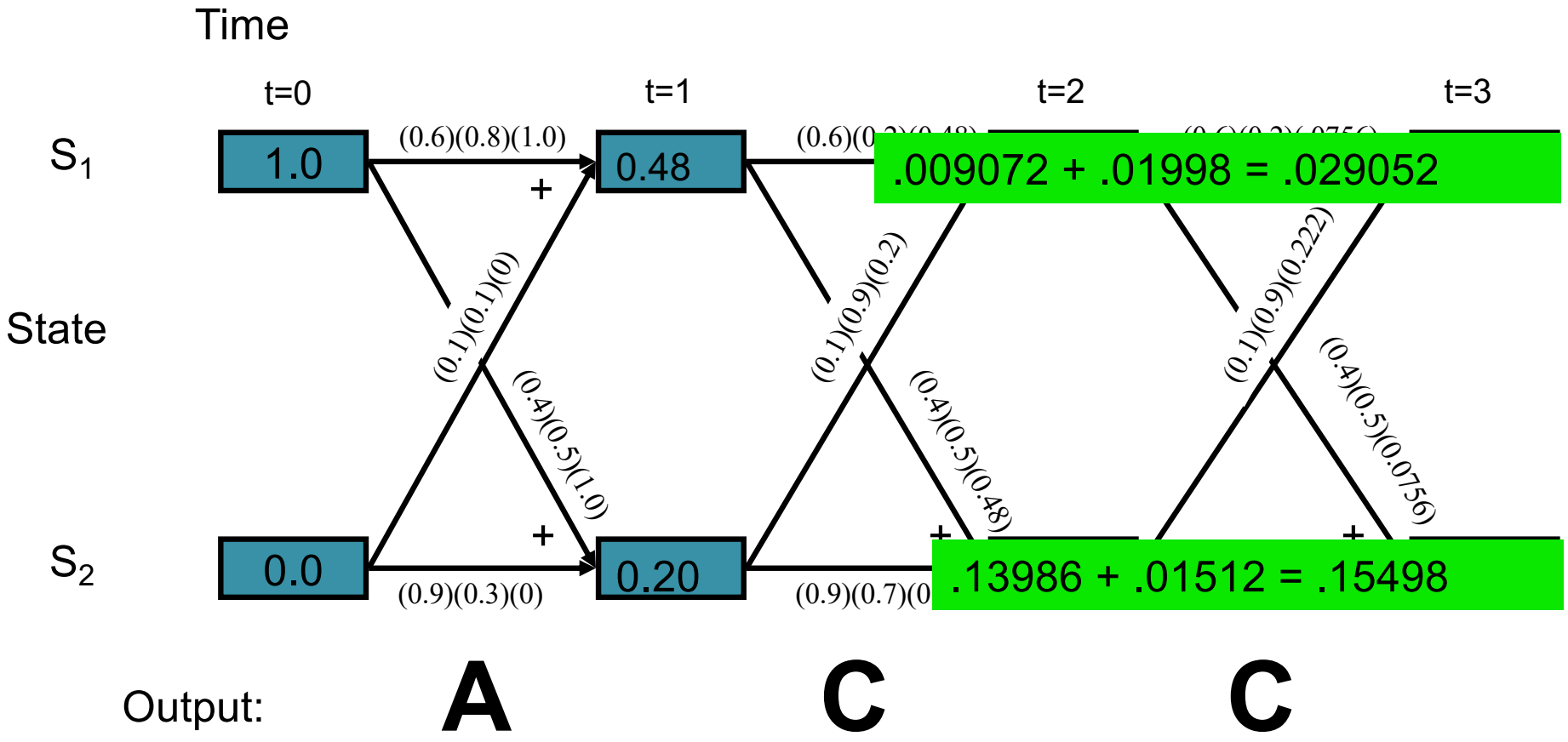
A trellis for the Forward Algorithm



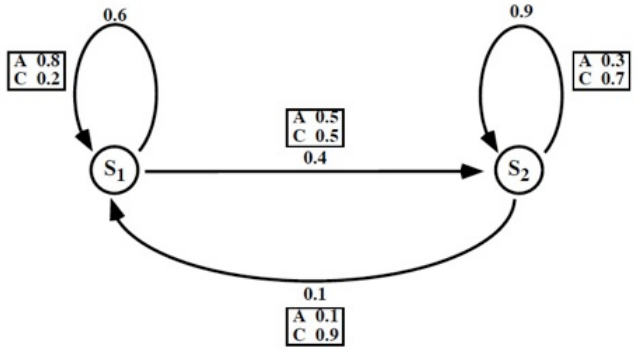


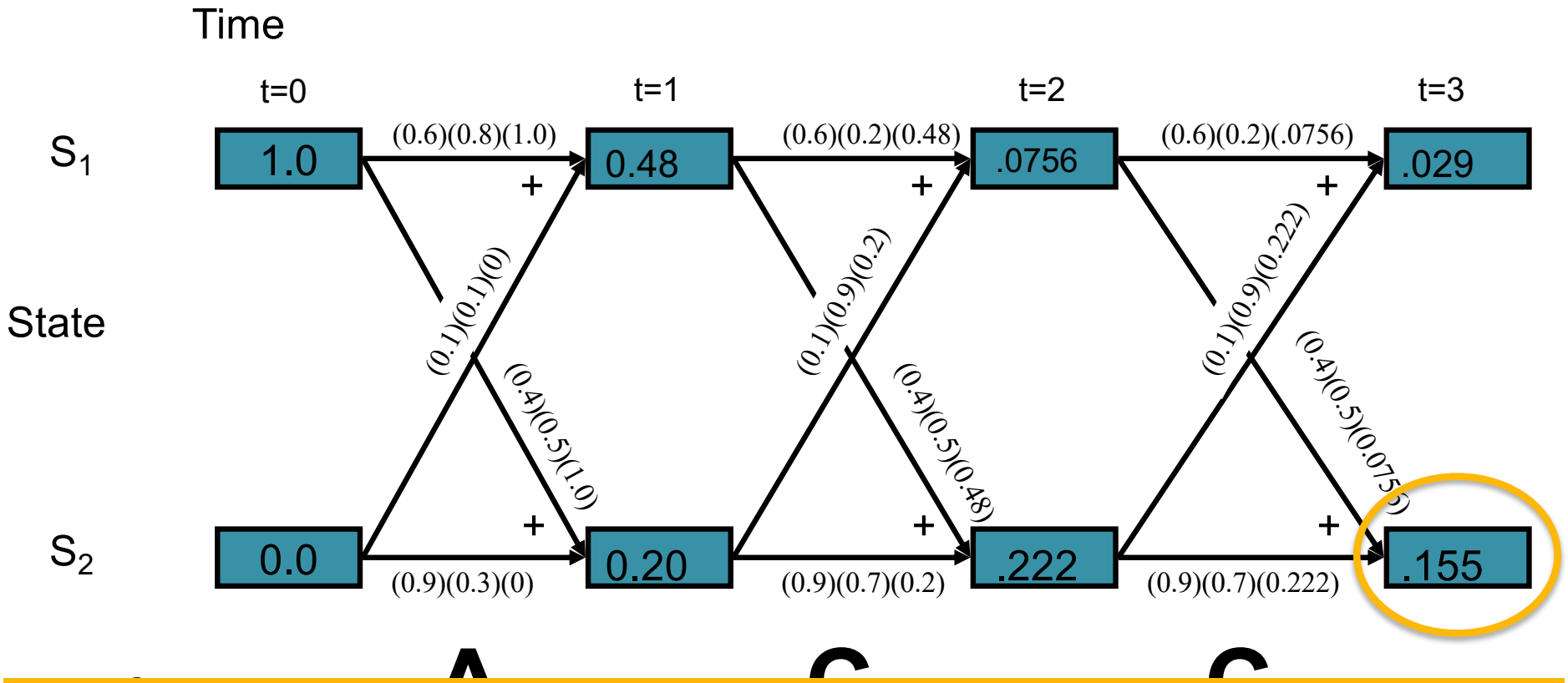
A trellis for the Forward Algorithm





A trellis for the Forward Algorithm





S2 is final state➔ 15.5% probability of this sequence **given** this model was used

Probability of the model

- The Forward algorithm computes *P(y|M)*

- If we are comparing two or more models, we want the likelihood that each model generated the data: *P(M|y)*

  – Use Bayes' law: $$P(M \mid y) = \frac{P(y \mid M)P(M)}{P(y)}$$

  – Since P(y) is constant for a given input, we just need to maximize *P(y|M)P(M)*

Three classic HMM problems

2. **Decoding**: given a model and an output sequence, what is the most likely state sequence through the model that generated the output?

- A solution to this problem gives us a way to match up an observed sequence and the states in the model.

AAAGCATGCATTTAACGAGAGCACAAGGGCTCTAATGCCG

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

Three classic HMM problems

2. **Decoding**: given a model and an output sequence, what is the most likely state sequence through the model that generated the output?

- A solution to this problem gives us a way to match up an observed sequence and the states in the model.

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

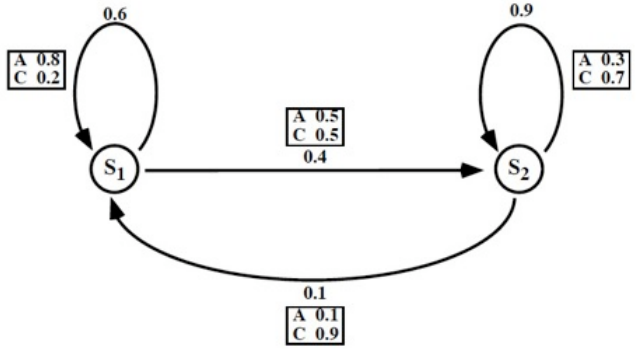
Solving the Decoding Problem: The Viterbi algorithm

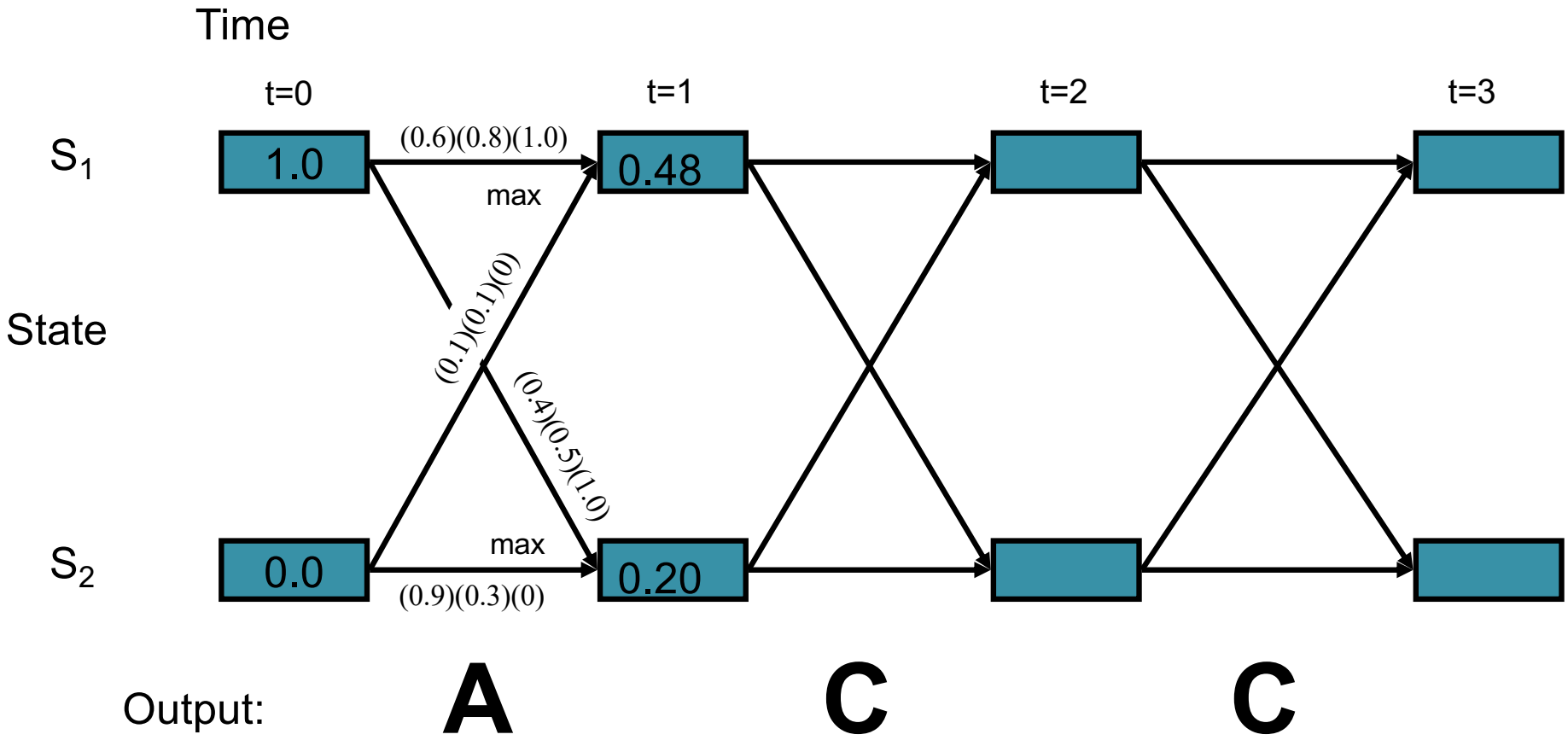
- To solve the decoding problem (find the most likely sequence of states), we evaluate the Viterbi algorithm

$$V_i(t) = \begin{cases} 0 & : \quad t = 0 \wedge i \neq S_I \\ 1 & : \quad t = 0 \wedge i = S_I \\ \max V_j(t-1) a_{ji} b_{ji}(y) & : \quad t > 0 \end{cases}$$

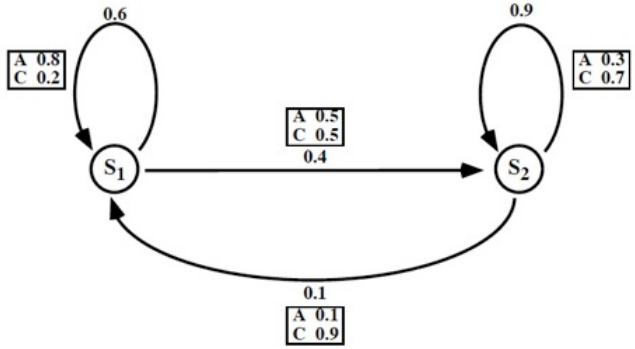
Where $V_i(t)$ is the probability that the HMM is in state i after generating the sequence $y_1, y_2, \ldots, y_t$, following the *most probable path* in the HMM

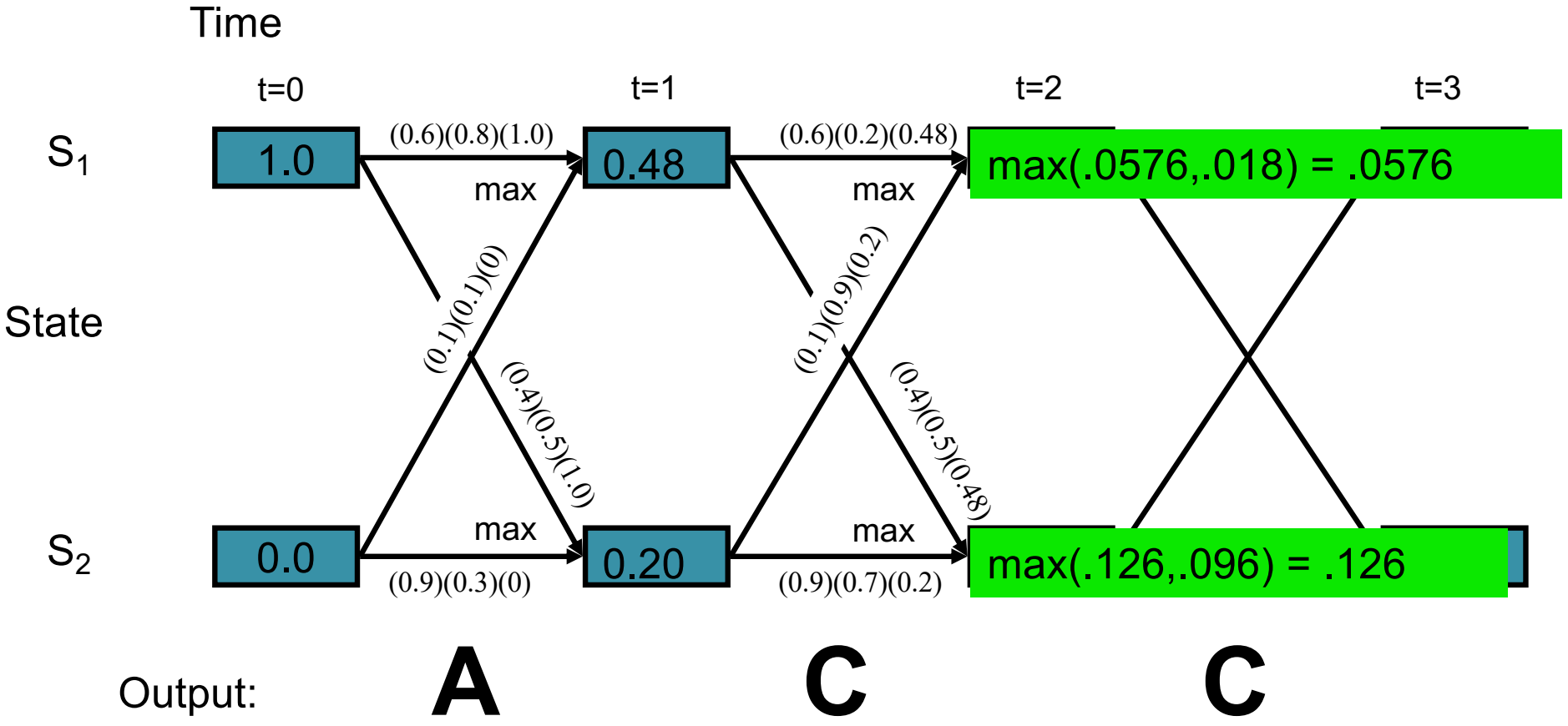
A trellis for the Viterbi Algorithm



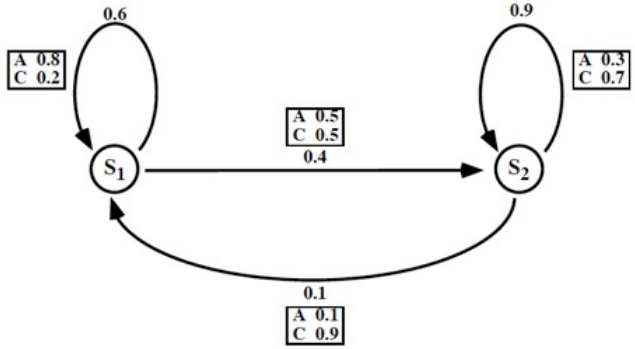


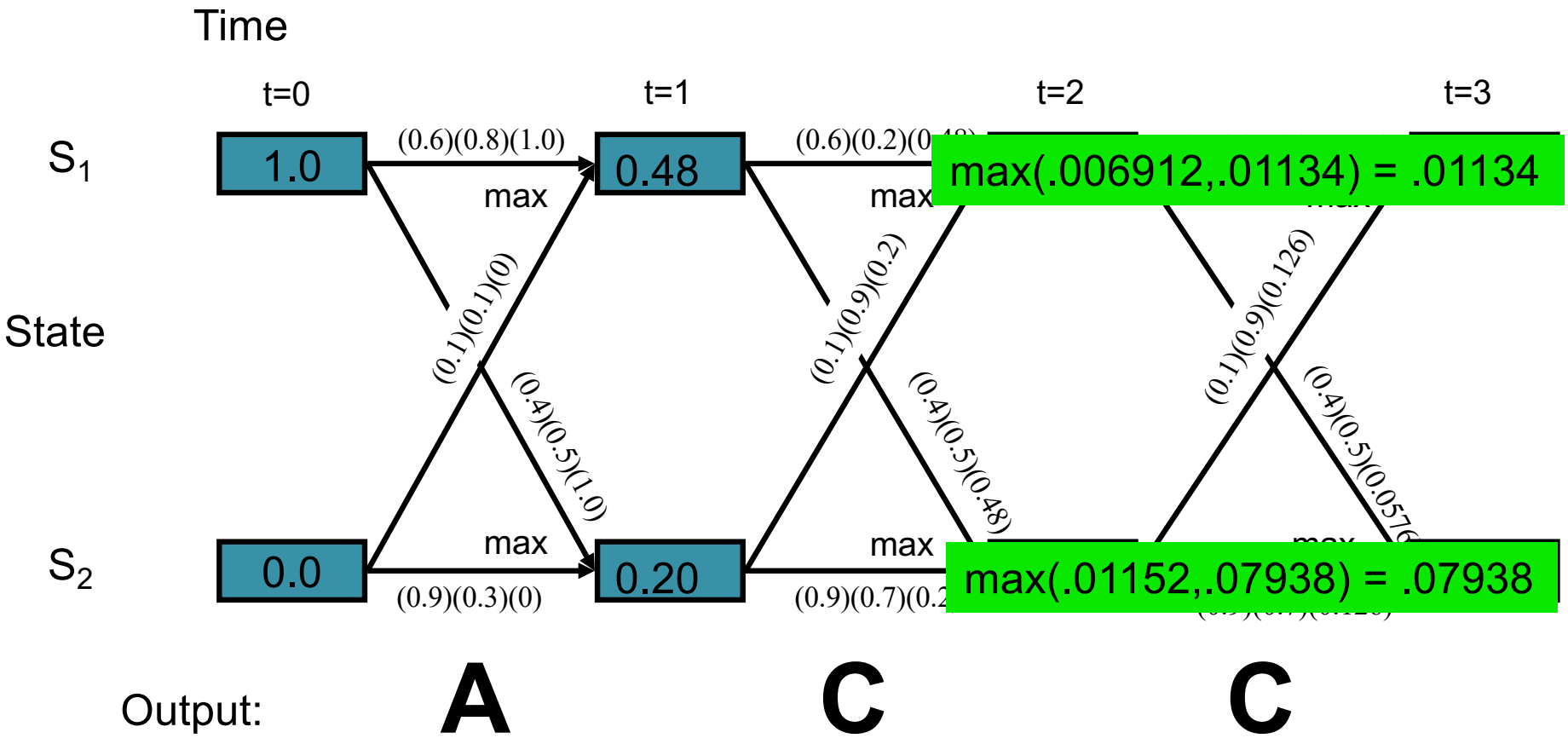
A trellis for the Viterbi Algorithm



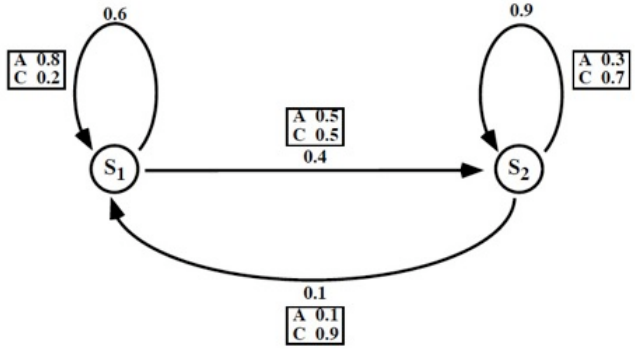


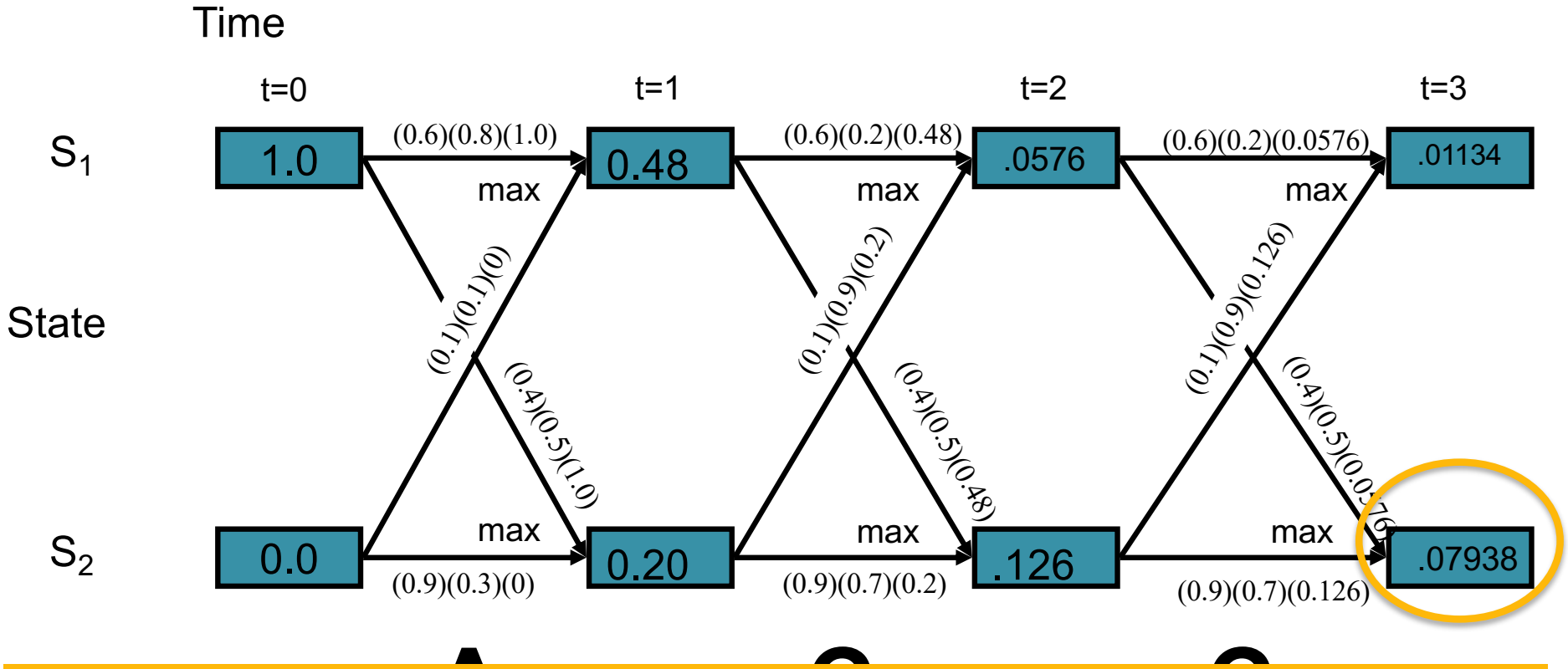
A trellis for the Viterbi Algorithm





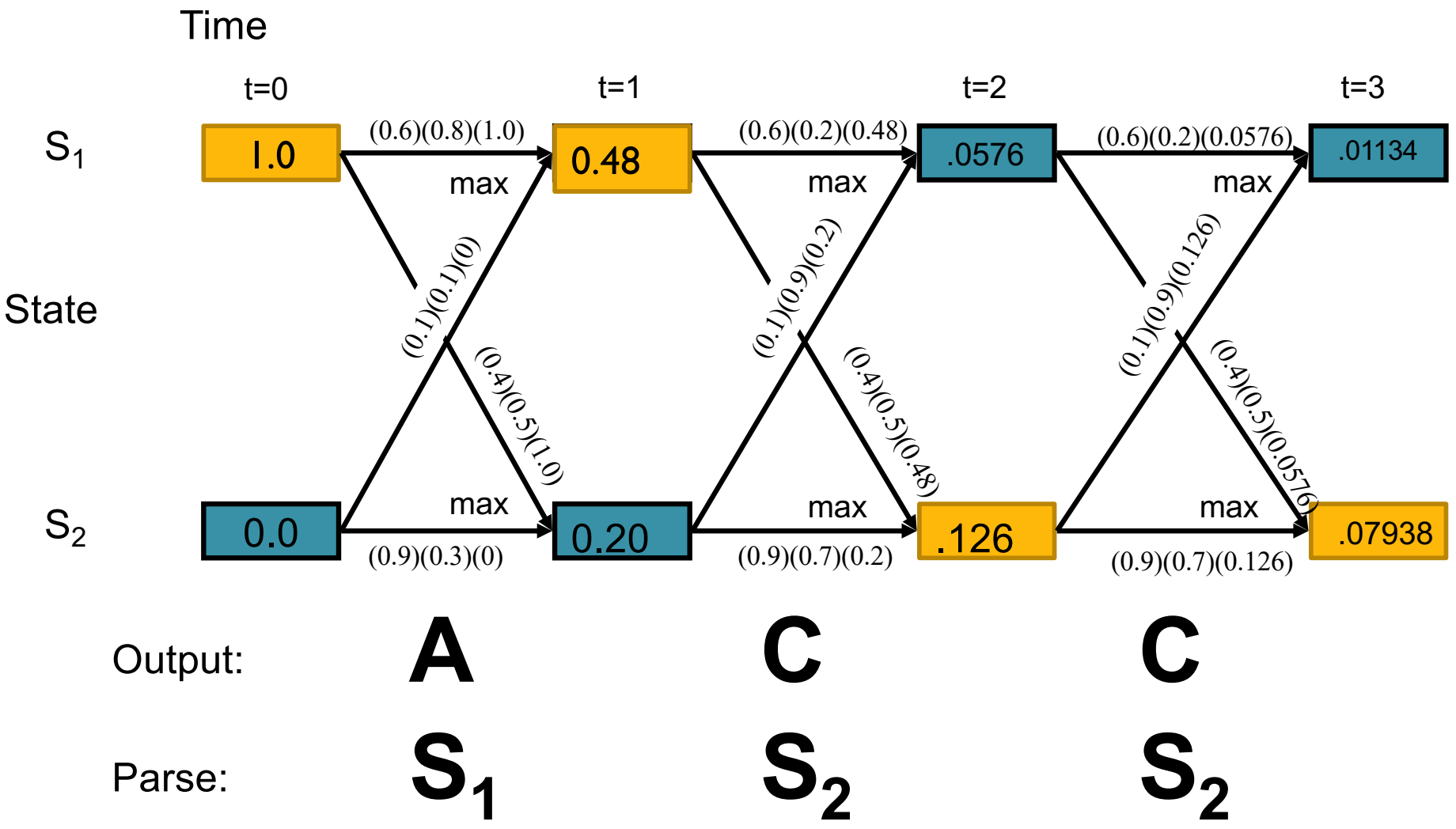
A trellis for the Viterbi Algorithm





S2 is final state➔ the most probable sequence of states has a 7.9% probability

A trellis for the Viterbi Algorithm



Three classic HMM problems

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?

- This is perhaps the most important, and most difficult problem.

- A solution to this problem allows us to determine all the probabilities in an HMMs by using an ensemble of training data

Learning in HMMs:

- The learning algorithm uses Expectation-Maximization (E-M)
  - Also called the Forward-Backward algorithm
  - Also called the Baum-Welch algorithm

- In order to learn the parameters in an "empty" HMM, we need:
  - The topology of the HMM
  - Data - the more the better
  - Start with a random (or naïve) probability, repeat until converges

Eukaryotic Gene Finding with GlimmerHMM

Mihaela Pertea

Assistant Professor

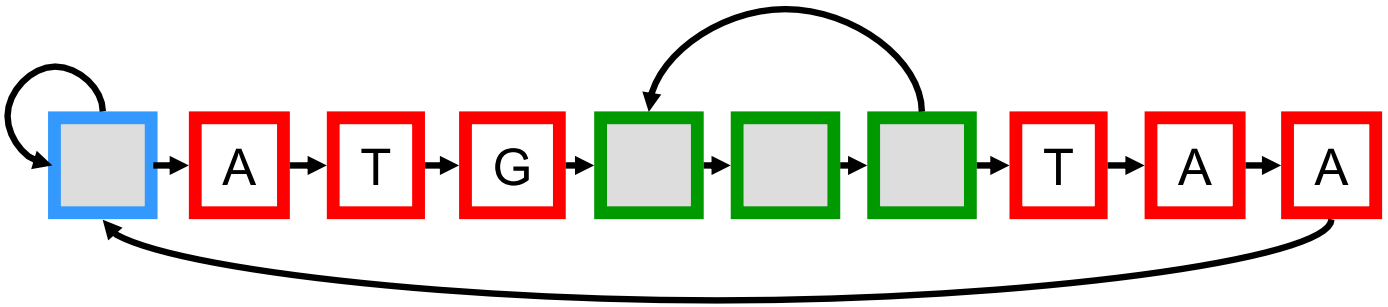
JHU

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:



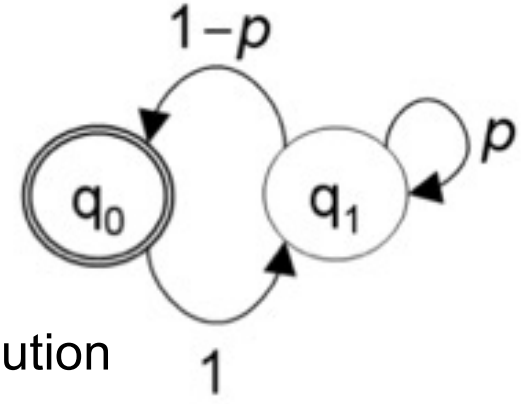
  AAAGC ATG CAT TTA ACG AGA GCA CAA GGG CTC TAA TGCCG

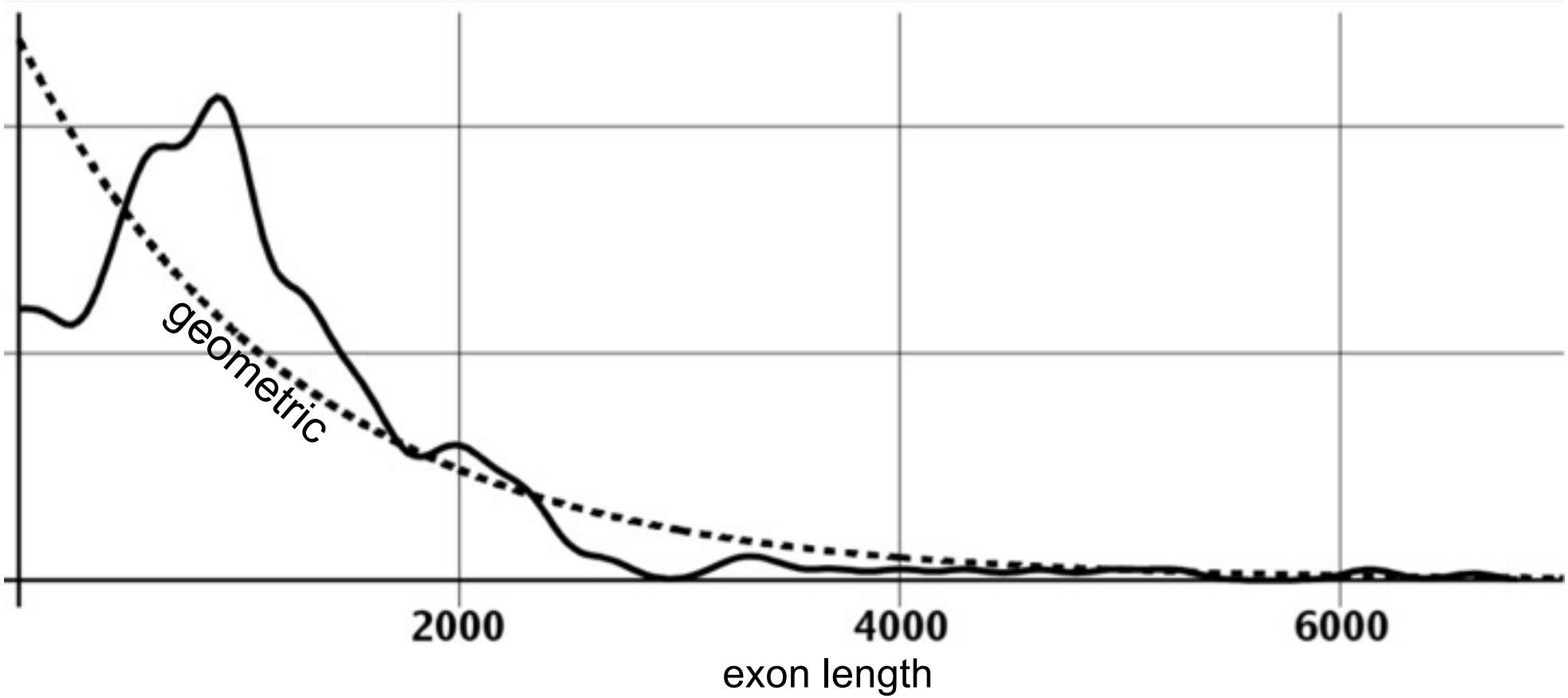
- The sequence of states is an annotation of the generated string – each nucleotide is generated in intergenic, start/stop, coding state

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

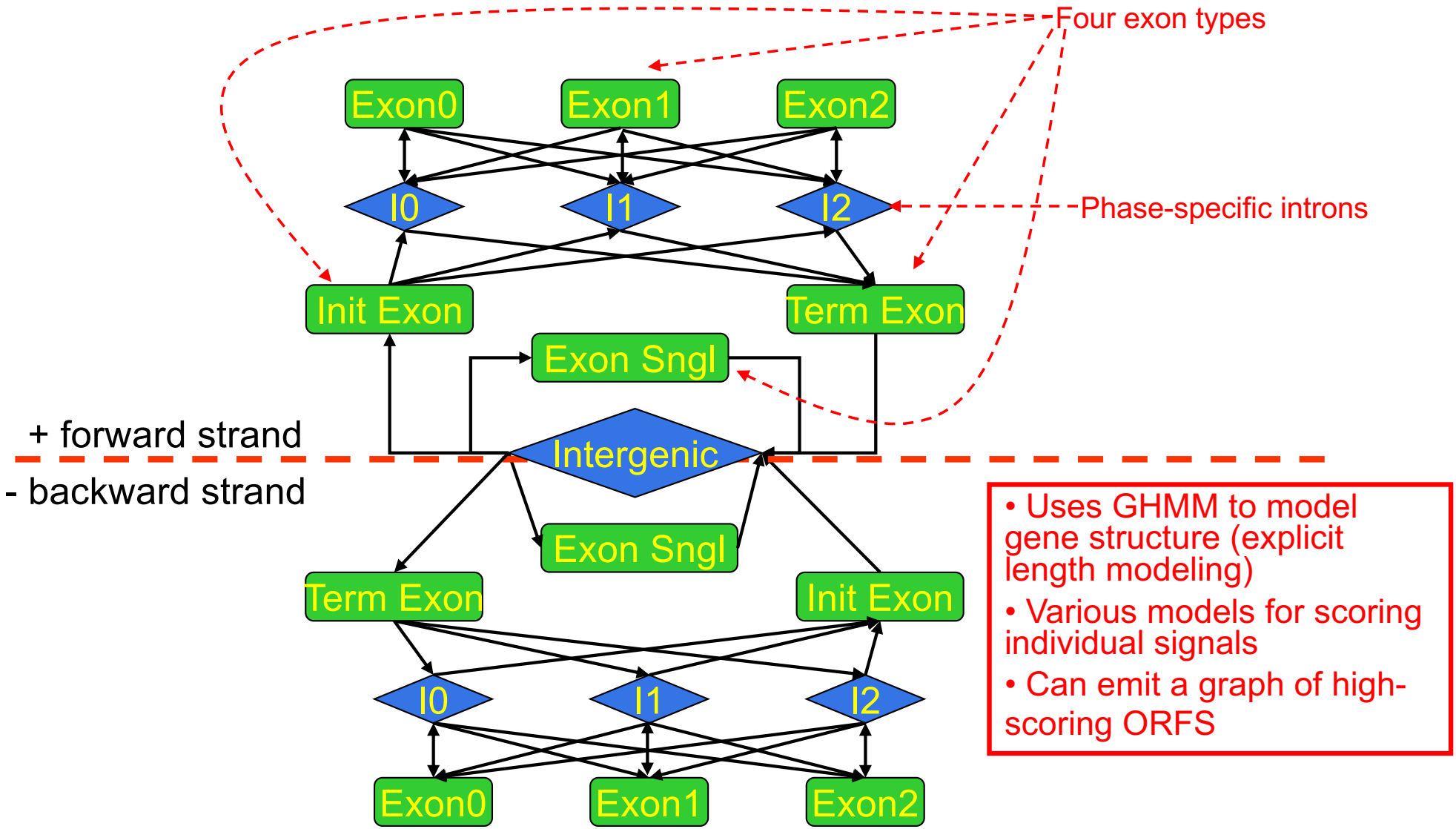
- GHMMs generalize HMMs by allowing each state to emit a subsequence rather than just a single symbol

- 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

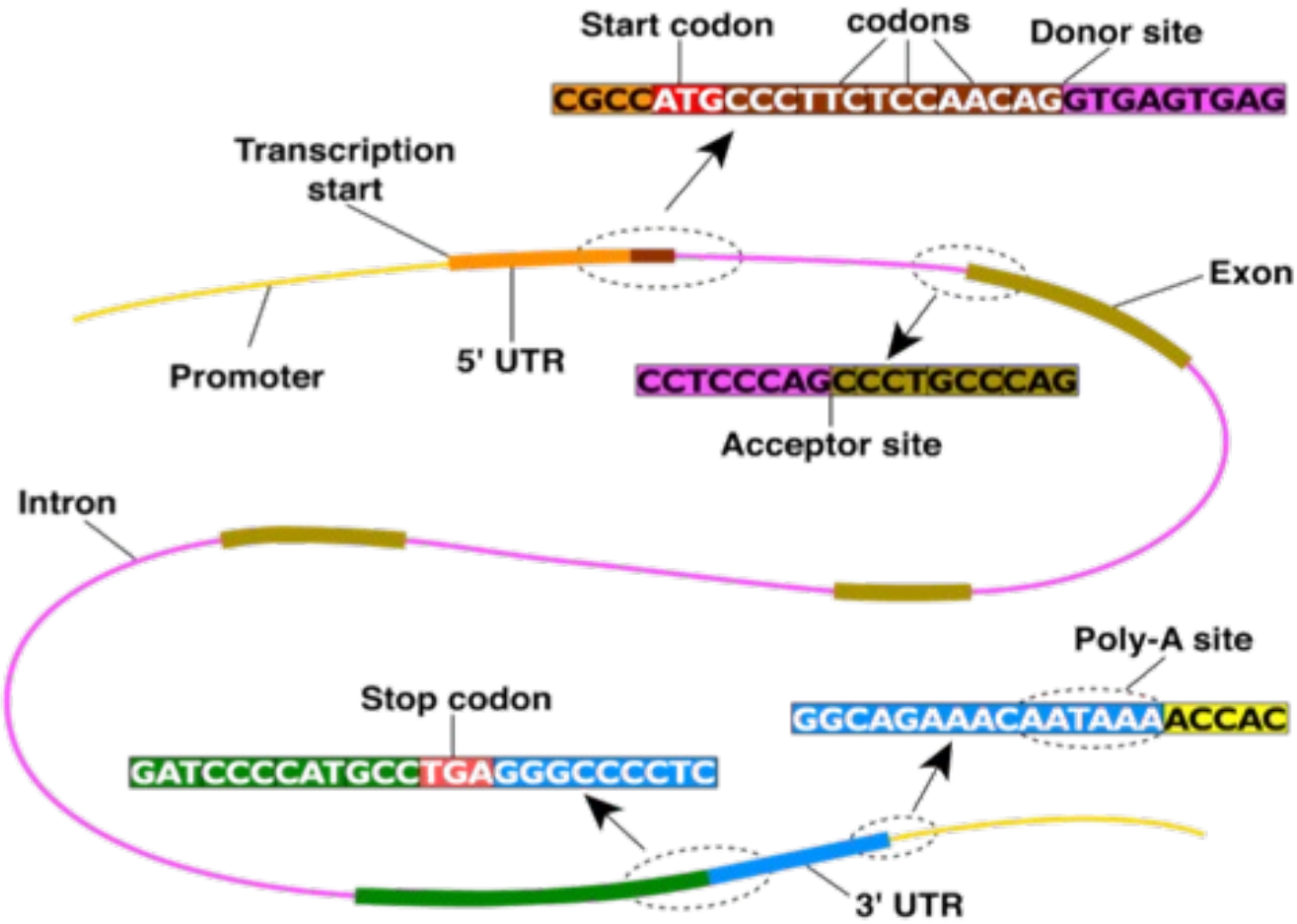
GlimmerHMM architecture



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

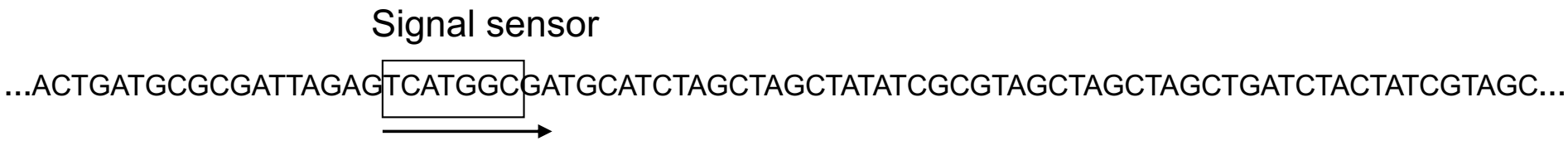
Signal Sensors

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.

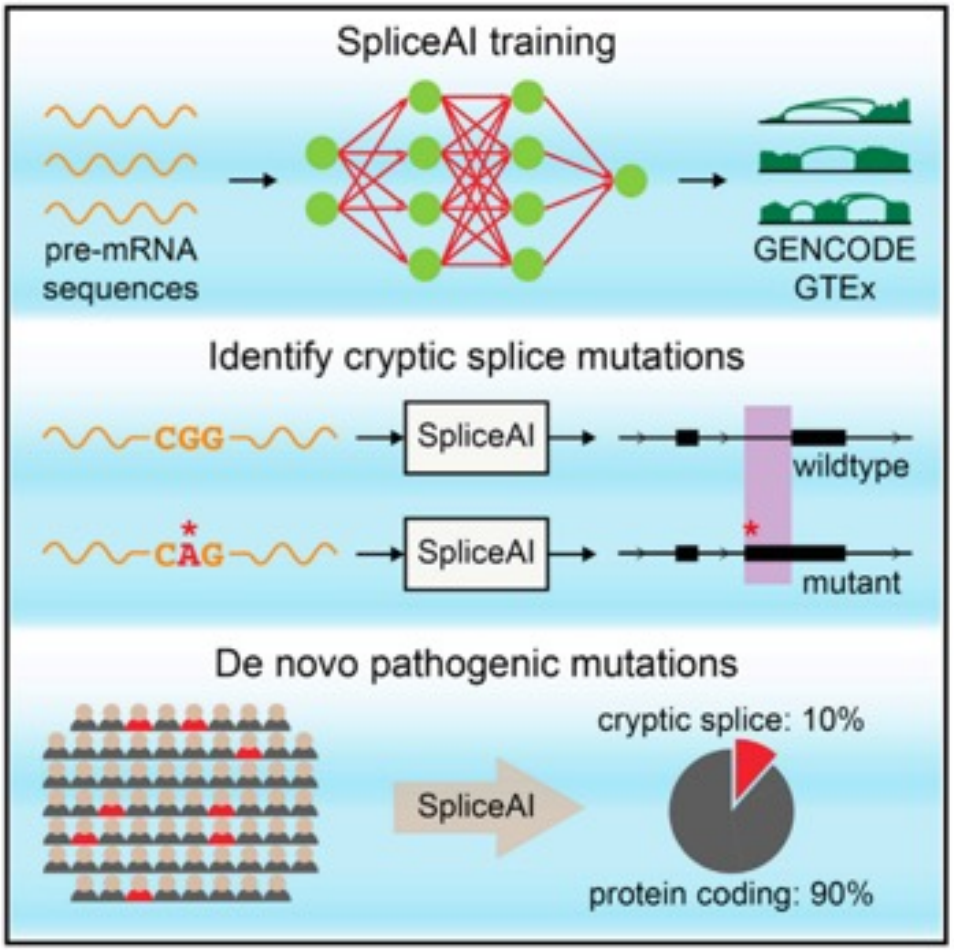
The most common signal sensor is the Position Weight Matrix:

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%

Cell

Predicting Splicing from Primary Sequence with Deep Learning

Graphical Abstract



Authors

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

Correspondence

kfarh@illumina.com

In Brief

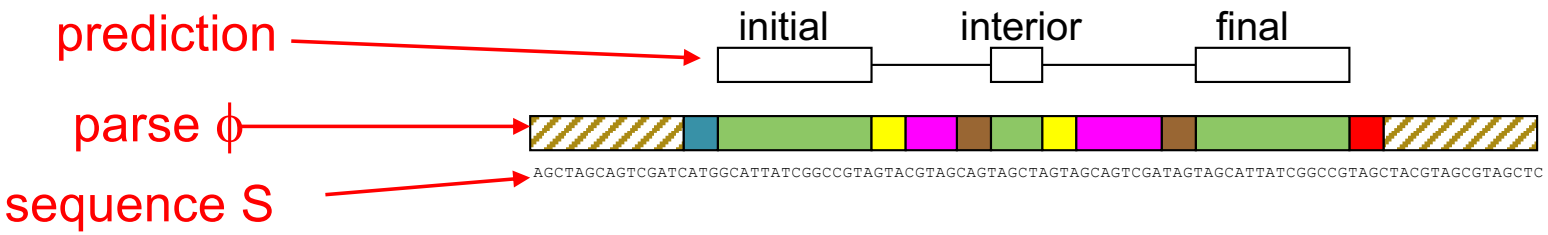
A deep neural network precisely models mRNA splicing from a genomic sequence and accurately predicts noncoding cryptic splice mutations in patients with rare genetic diseases.

Highlights

- SpliceAI, a 32-layer deep neural network, predicts splicing from a pre-mRNA sequence

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:

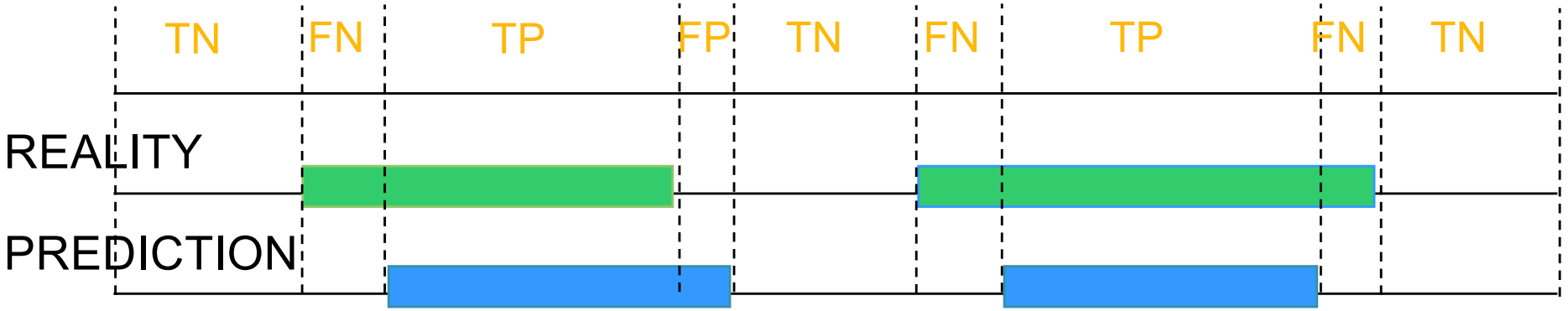


The parse ϕ consists of the coordinates of the predicted exons, and corresponds to the precise sequence of states during the operation of the GHMM (and their duration, which equals the number of symbols each state emits).

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

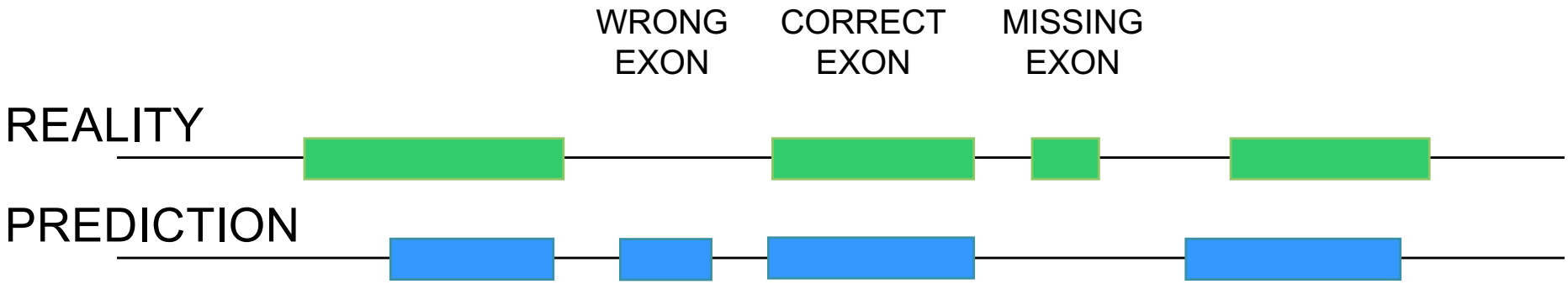


Sensitivity: $$Sn = \frac{TP}{TP + FN}$$ What fraction of reality did you predict?

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%

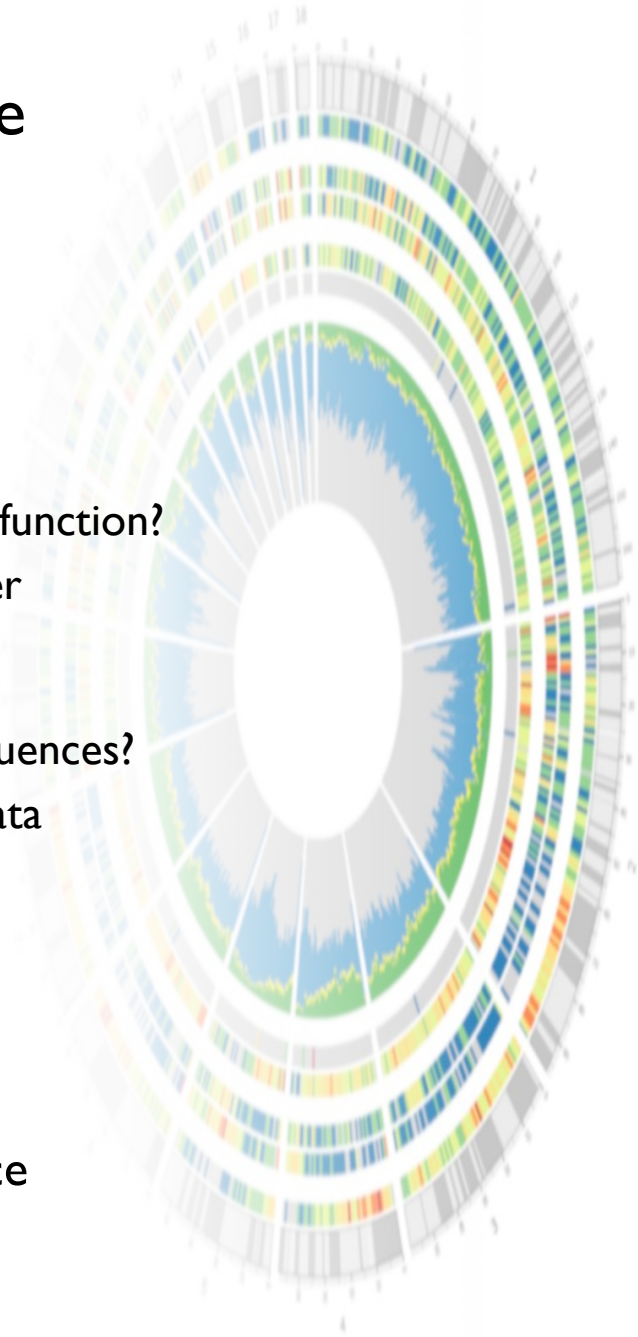
GlimmerHMM's performace compared to Genscan on 963 human RefSeq genes selected randomly from all 24 chromosomes, non-overlapping with the training set. The test set contains 1000 bp of untranslated sequence on either side (5' or 3') of the coding portion of each gene.

Gene Prediction 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

Annotation Summary

- Three major approaches to annotate a genome
  1. Experimental:
     - Lets test to see if it is transcribed/methylated/bound/etc
     - Strongest but expensive and context dependent
  2. Alignment:
     - Does this sequence align to any other sequences of known function?
     - Great for projecting knowledge from one species to another
  3. Prediction:
     - Does this sequence statistically resemble other known sequences?
     - Potentially most flexible but dependent on good training data



- Many great resources available
  - Learn to love the literature and the databases
  - Standard formats let you rapidly query and cross reference
  - Google is your number one resource ☺