

Gene Finding

Slides by Carl Kingsford



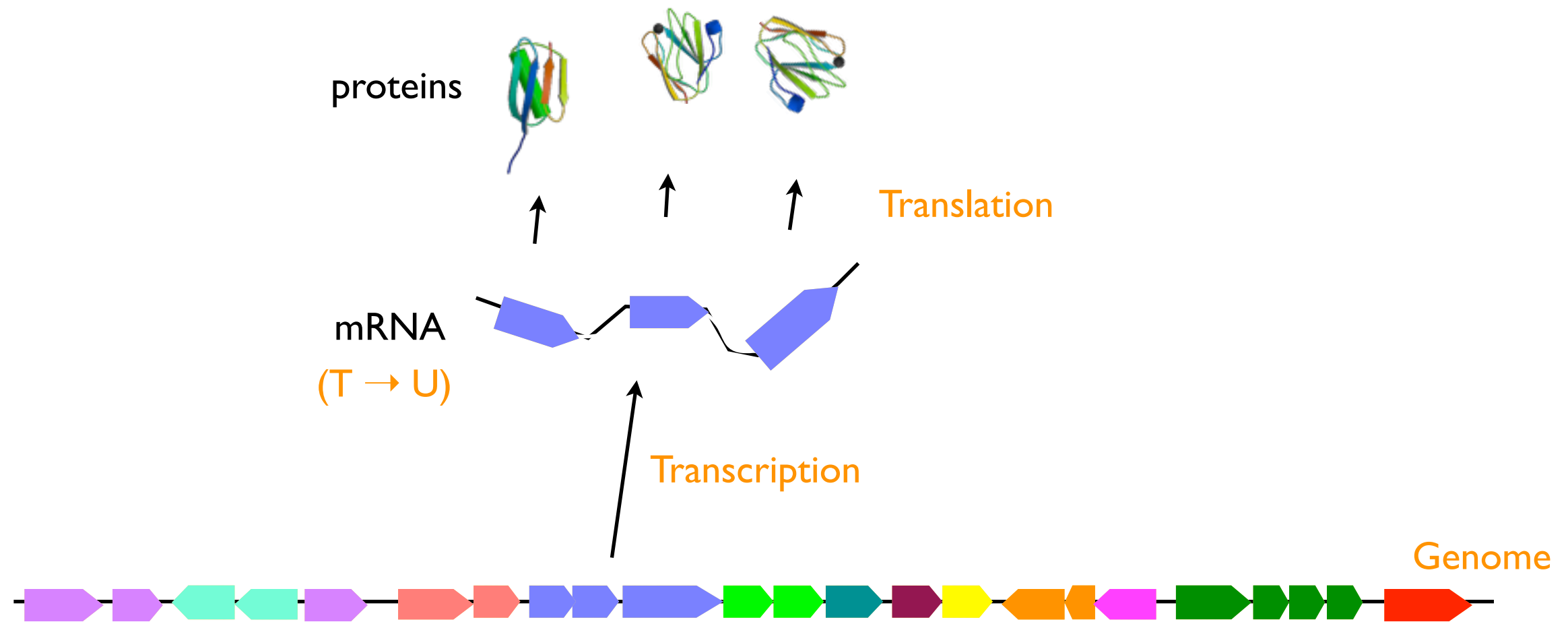
Genome of the Cow

a sequence of 2.86 billion letters


enough letters to fill a million pages of a typical book.

```
TATGGAGCCAGGTGCCTGGGGCAACAAGACTGTGGTCACTGAATTCATCCTTCTTGGTCTAACAGAGAACATAG  
AACTGCAATCCATCCTTTTTTGCCATCTTCCTCTTTGCCTATGTGATCACAGTCGGGGGCAACTTGAGTATCCTG  
GCCGCCATCTTTGTGGAGCCCAAACCTCCACACCCCCATGTACTACTTCCTGGGGAACCTTTCTCTGCTGGACAT  
TGGGTGCATCACTGTCACCAATTCCTCCCATGCTGGCCTGTCTCCTGACCCACCAATGCCGGGTTCCTATGCAG  
CCTGCATCTCACAGCTCTTCTTTTTCCACCTCCTGGCTGGAGTGGACTGTCACCTCCTGACAGCCATGGCCTAC  
GACCGCTACCTGGCCATTTGCCAGCCCCTCACCTATAGCATCCGCATGAGCCGTGACGTCCAGGGAGCCCTGGT  
GGCCGTCTGCTGCTCCATCTCCTTCATCAATGCTCTGACCCACACAGTGGCTGTGTCTGTGCTGGACTTCTGCG  
GCCCTAACGTGGTCAACCACTTCTACTGTGACCTCCCGCCCCCTTTTCCAGCTCTCCTGCTCCAGCATCCACCTC  
AACGGGCAGCTACTTTTCGTGGGGGCCACCTTCATGGGGGTGGTCCCCATGGTCTTCATCTCGGTATCCTATGC  
CCACGTGGCAGCCGCAGTCCTGCGGATCCGCTCGGCAGAGGGCAGGAAGAAAGCCTTCTCCACGTGTGGCTCCC  
ACCTCACCGTGGTCTGCATCTTTTATGGAACCGGCTTCTTCAGCTACATGCGCCTGGGCTCCGTGTCCGCCTCA  
GACAAGGACAAGGGCATTGGCATCCTCAACACTGTCATCAGCCCCATGCTGAACCCACTCATCTACAGCCTCCG  
GAACCCTGATGTGCAGGGCGCCCTGAAGAGGTTGCTGACAGGGAAGCGGCCCCCGGAGTG ...
```

"Central Dogma" of Biology



DNA =

- double-stranded, linear molecule
- each strand is string over {A,C,G,T}
- strands are complements of each other ($A \leftrightarrow T$; $C \leftrightarrow G$)
- substrings encode for genes  most of which encode for proteins



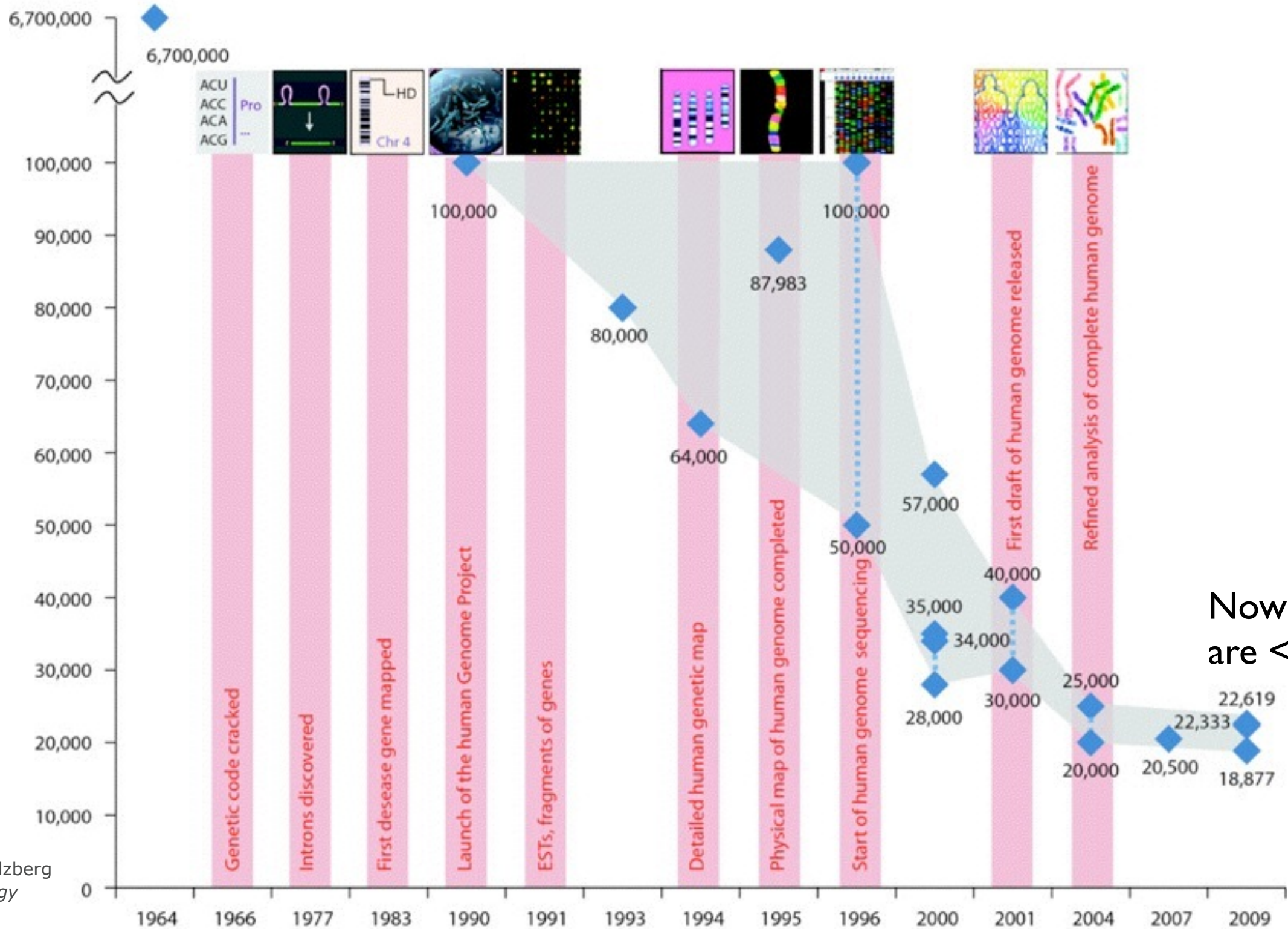
The Genetic Code

- There are 20 different amino acids & 64 different codons.
- Lots of different ways to encode for each amino acid.
- The 3rd base is typically less important for determining the amino acid
- Three different “stop” codons that signal the end of the gene
- Start codons differ depending on the organisms, but AUG is often used.

		2nd base							
		U	C	A	G				
1st base	U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine
		UUC	(Phe/F) Phenylalanine	UCC	(Ser/S) Serine	UAC	(Tyr/Y) Tyrosine	UGC	(Cys/C) Cysteine
		UUA	(Leu/L) Leucine	UCA	(Ser/S) Serine	UAA	Ochre Stop	UGA	Opal Stop
		UUG	(Leu/L) Leucine	UCG	(Ser/S) Serine	UAG	Amber Stop	UGG	(Trp/W) Tryptophan
	C	CUU	(Leu/L) Leucine	CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine
		CUC	(Leu/L) Leucine	CCC	(Pro/P) Proline	CAC	(His/H) Histidine	CGC	(Arg/R) Arginine
		CUA	(Leu/L) Leucine	CCA	(Pro/P) Proline	CAA	(Gln/Q) Glutamine	CGA	(Arg/R) Arginine
		CUG	(Leu/L) Leucine	CCG	(Pro/P) Proline	CAG	(Gln/Q) Glutamine	CGG	(Arg/R) Arginine
	A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine
		AUC	(Ile/I) Isoleucine	ACC	(Thr/T) Threonine	AAC	(Asn/N) Asparagine	AGC	(Ser/S) Serine
		AUA	(Ile/I) Isoleucine	ACA	(Thr/T) Threonine	AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
		AUG ^[A]	(Met/M) Methionine	ACG	(Thr/T) Threonine	AAG	(Lys/K) Lysine	AGG	(Arg/R) Arginine
	G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine
		GUC	(Val/V) Valine	GCC	(Ala/A) Alanine	GAC	(Asp/D) Aspartic acid	GGC	(Gly/G) Glycine
		GUA	(Val/V) Valine	GCA	(Ala/A) Alanine	GAA	(Glu/E) Glutamic acid	GGA	(Gly/G) Glycine
		GUG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

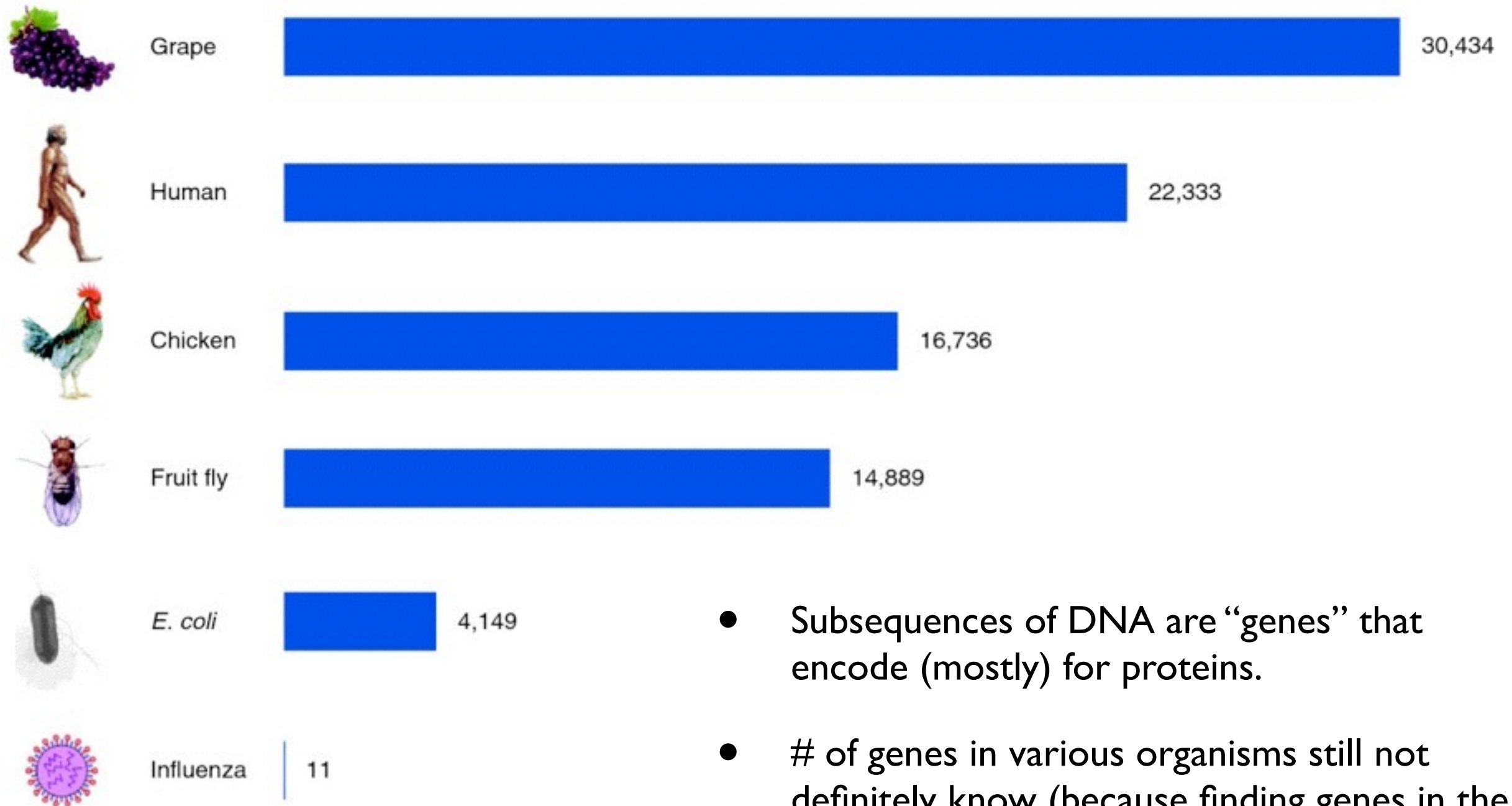
Estimates for the # of Human Genes

Before human genome sequence was available, many (but not all) estimates for # of genes were high (> 80,000).



Now estimates are < 23,000.

of Genes in Various Organisms



- Subsequences of DNA are “genes” that encode (mostly) for proteins.
- # of genes in various organisms still not definitely know (because finding genes in the sequence is a hard problem that we will talk about).
- But there are reasonably good estimates.

The Prokaryotic Gene Finding Problem

- Genes are subsequences of DNA that tell the cell how to make specific proteins.
- How can we find which subsequences of DNA are genes?

Start Codon: ATG

Stop Codons: TGA, TAG, TAA

—————→
ATAGAGGGT**AT**GGGGGACCCGGACACG**AT**GGCAGAT**TG**ACGATGACGATGACGATGACGGGT**TGA**AGTGAGTCAACACATGAC

Challenges:

- The start codon can occur in the middle of a gene (where it encodes for the amino acid methionine)
- The stop codon can occur in nonsense DNA between genes.
- The stop codon can occur “out of frame” inside a gene.
- Don’t know what “phase” the gene starts in.

A Simple Gene Finder

1. Find all stop codons in genome
2. For each stop codon, find the in-frame start codon farthest upstream of the stop codon, without crossing another in-frame stop codon.

GGC **TAG** **ATG** AGG GCT CTA ACT **ATG** GGC GCG **TAA**

Each substring between the start and stop codons is called an ORF
“open reading frame”

3. Return the “long” ORF as predicted genes.

3 out of the 64 possible codons are stop codons \Rightarrow in random DNA,
every 22nd codon is expected to be a stop.

Gene Finding as a Machine Learning Problem

- Given training examples of some known genes, can we distinguish ORFs that are genes from those that are not?
- **Idea:** can use distribution of codons to find genes.
 - every codon should be about equally likely in non-gene DNA.
 - every organism has a slightly different bias about how often certain codons are preferred.
 - could also use frequencies of longer strings (k-mers).

Bacillus anthracis (anthrax) codon usage

UUU	F	0.76	UCU	S	0.27	UAU	Y	0.77	UGU	C	0.73
UUC	F	0.24	UCC	S	0.08	UAC	Y	0.23	UGC	C	0.27
UUA	L	0.49	UCA	S	0.23	UAA	*	0.66	UGA	*	0.14
UUG	L	0.13	UCG	S	0.06	UAG	*	0.20	UGG	W	1.00
CUU	L	0.16	CCU	P	0.28	CAU	H	0.79	CGU	R	0.26
CUC	L	0.04	CCC	P	0.07	CAC	H	0.21	CGC	R	0.06
CUA	L	0.14	CCA	P	0.49	CAA	Q	0.78	CGA	R	0.16
CUG	L	0.05	CCG	P	0.16	CAG	Q	0.22	CGG	R	0.05
AUU	I	0.57	ACU	T	0.36	AAU	N	0.76	AGU	S	0.28
AUC	I	0.15	ACC	T	0.08	AAC	N	0.24	AGC	S	0.08
AUA	I	0.28	ACA	T	0.42	AAA	K	0.74	AGA	R	0.36
AUG	M	1.00	ACG	T	0.15	AAG	K	0.26	AGG	R	0.11
GUU	V	0.32	GCU	A	0.34	GAU	D	0.81	GGU	G	0.30
GUC	V	0.07	GCC	A	0.07	GAC	D	0.19	GGC	G	0.09
GUA	V	0.43	GCA	A	0.44	GAA	E	0.75	GGA	G	0.41
GUG	V	0.18	GCG	A	0.15	GAG	E	0.25	GGG	G	0.20

An Improved Simple Gene Finder

- Score each ORF using the product of the probability of each codon:

$$\text{GFScore}(g) = \text{Pr}(\text{codon}_1) \times \text{Pr}(\text{codon}_2) \times \text{Pr}(\text{codon}_3) \times \dots \times \text{Pr}(\text{codon}_n)$$

But: as genes get longer, $\text{GFScore}(g)$ will decrease.

So: we should calculate $\text{GFScore}(g[i\dots i+k])$ for some window size k .

The final $\text{GFSCORE}(g)$ is the average of the Scores of the windows in it.

Glimmer

Salzberg et al., NAR, 1998

- Score ORFs using 6 HMMs:
 - 1 model for each reading frame (3 forward, 3 reverse)
- ORFs for which the correct reading frame is the highest score are saved as candidates.
- Use “Interpolated Markov models” to adapt to data availability
- Handle overlapping ORFs

Interpolated HMMs

Sequence

Model

Length of the sequence

String ending at position x

$$P(S|M) = \sum_{x=1}^n \text{IMM}_8(S_x)$$

IMM score is a linear combination of 8th, 7th, ..., 0th order models:

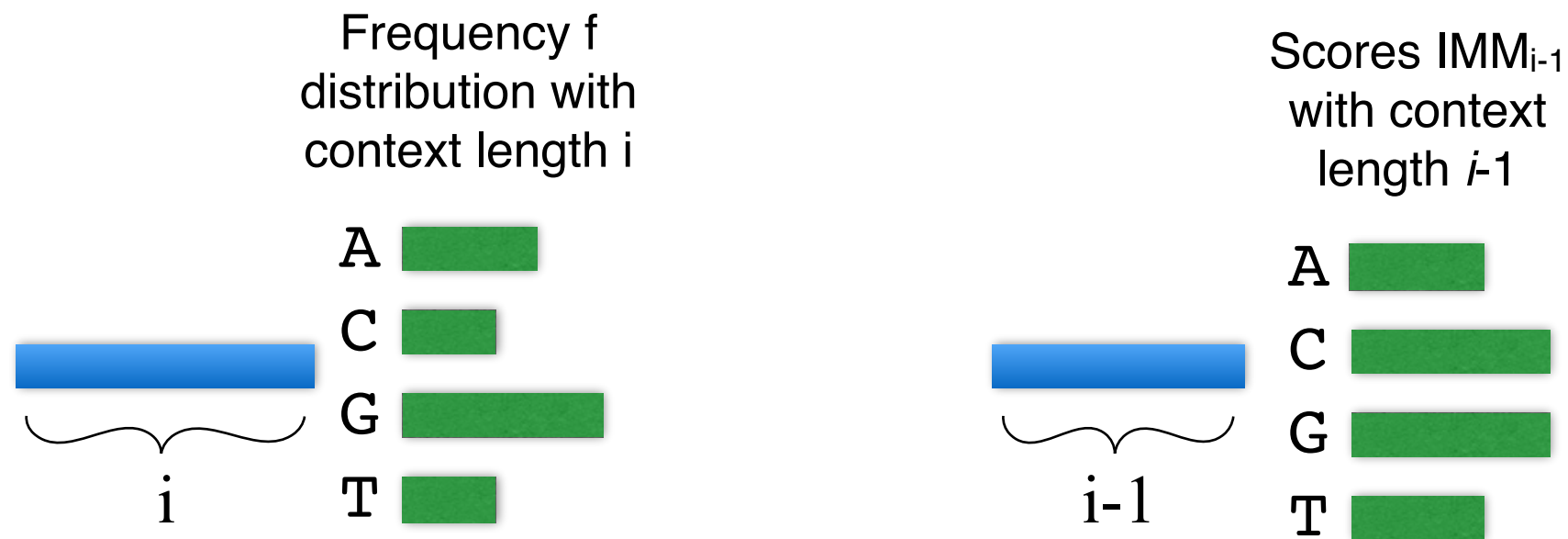
Weight of the k -mer ending at position $x-1$

$$\text{IMM}_k(S_x) = \lambda_k(S_{x-1}) \cdot P_k(S_x) + [1 - \lambda_k(S_{x-1})] \cdot \text{IMM}_{k-1}(S_x)$$

Probability of letter at position x from a k th-order model

Setting Parameters

- If # of occurrences of context k-mer ≥ 400 , $\lambda = 1$
- Otherwise compare the following with a χ^2 statistic:

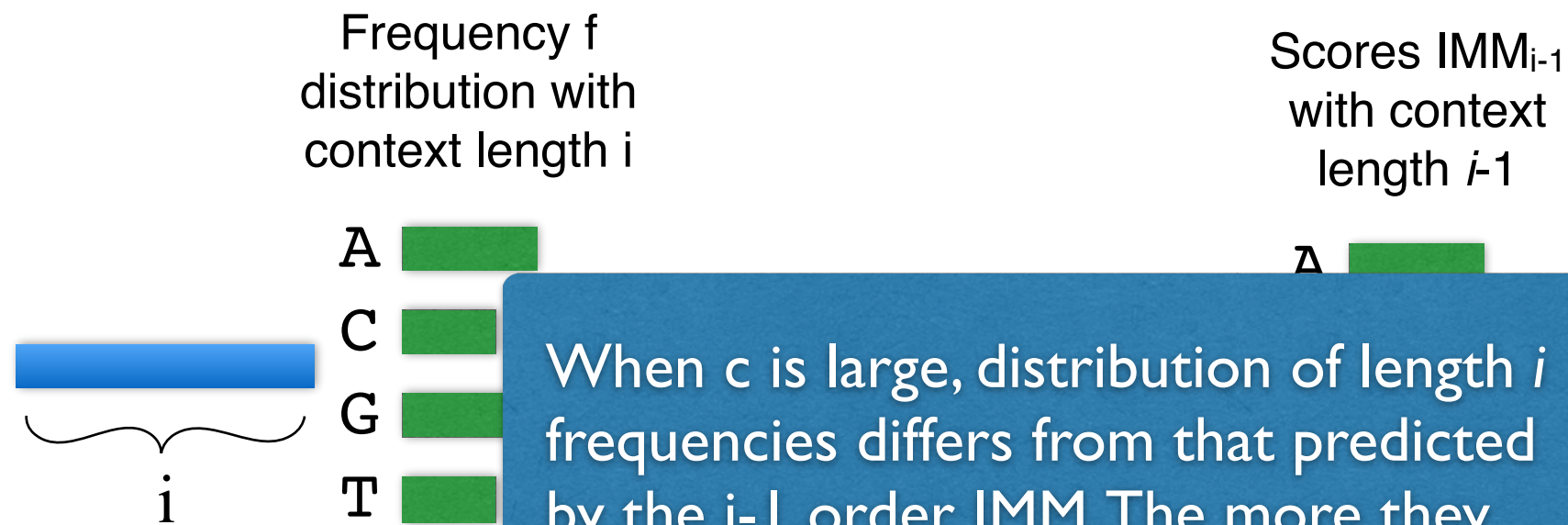


- Set λ as follows, where c is the χ^2 statistic that the frequencies did not come from the IMM distribution:

$$\lambda_i(S_{x-1}) = \begin{cases} 0.0 & \text{if } c < 0.50 \\ \frac{c}{400} \sum_{b \in \{acgt\}} f(s_1 s_2 \dots s_i b) & \text{if } c \geq 0.50 \end{cases}$$

Setting Parameters

- If # of occurrences of context k-mer ≥ 400 , $\lambda = 1$
- Otherwise compare the following with a χ^2 statistic:



When c is large, distribution of length i frequencies differs from that predicted by the $i-1$ order IMM. The more they differ, the more we weight them.

- Set λ as follows, where c is the number of occurrences of context $i-1$ and f come from the IMM distribution:

$$\lambda_i(S_{x-1}) = \begin{cases} 0.0 & \text{if } c < 0.50 \\ \frac{c}{400} \sum_{b \in \{acgt\}} f(s_1 s_2 \dots s_i b) & \text{if } c \geq 0.50 \end{cases}$$

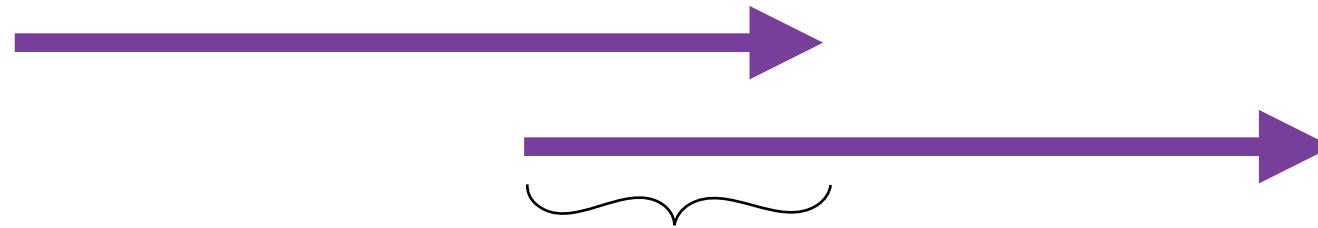
IMM vs. 5th Order HMM

Model	Genes found	Genes missed	Additional genes
GLIMMER IMM	1680 (97.8%)	37	209
5 th -Order Markov	1574 (91.7%)	143	104

The first column indicates how many of the 1717 annotated genes in *H.influenzae* were found by each algorithm. The 'additional genes' column shows how many extra genes, not included in the 1717 annotated entries, were called genes by each method.

Salzberg et al., NAR, 1998

Overlaps



Scored separately with the
two IMMs for the reading
frames for the two genes

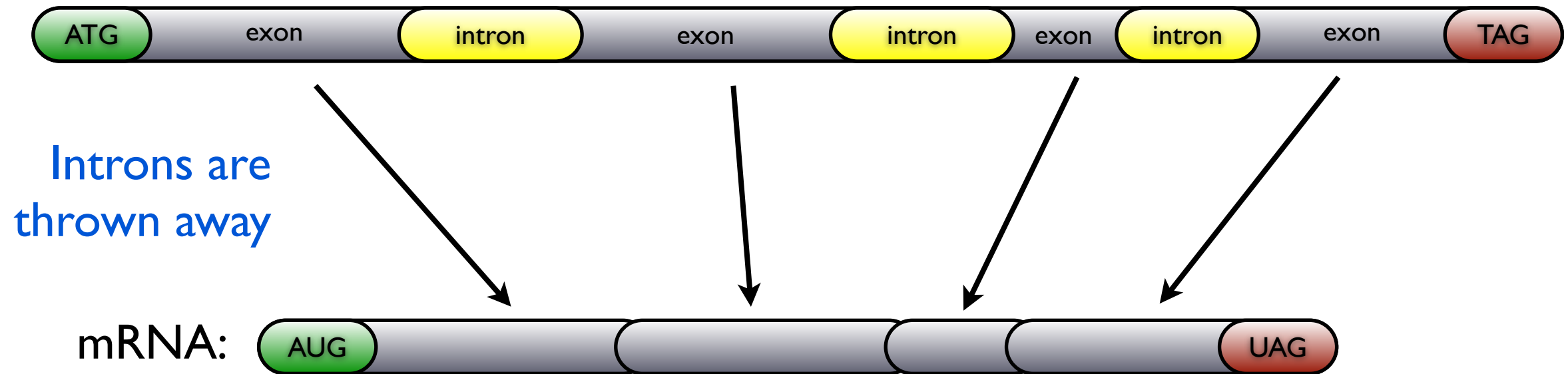
Discard the shorter gene if the longer gene's reading frame scores higher

Eukaryotic Genes & Exon Splicing

Prokaryotic (bacterial) genes look like this:



Eukaryotic genes usually look like this:

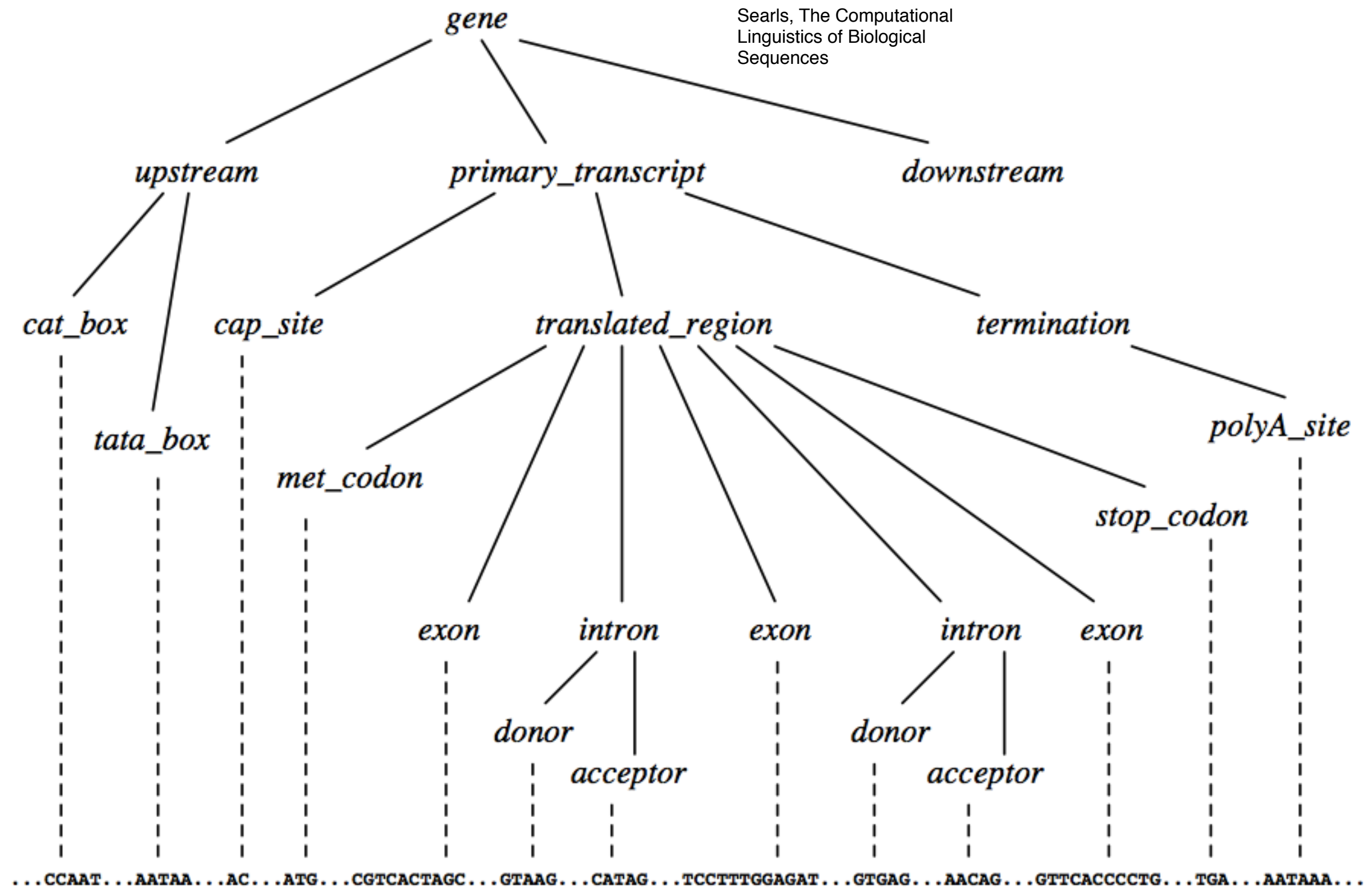


Exons are concatenated together

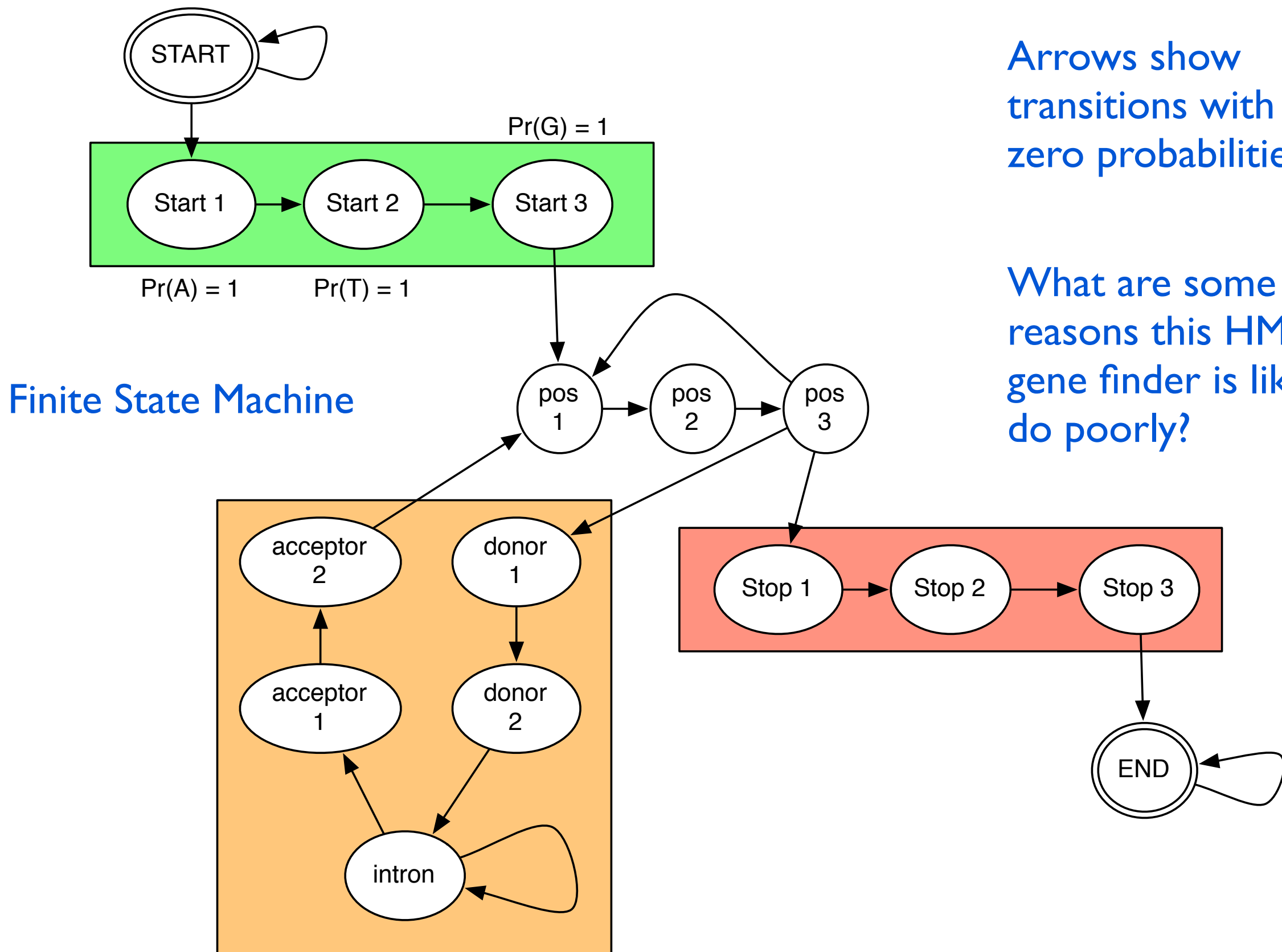
This spliced RNA is what is translated into a protein.

Hypothetical Eukaryotic Gene Parse Tree

Searls, The Computational Linguistics of Biological Sequences



A (Bad) Eukaryotic Gene Finder



Arrows show transitions with non-zero probabilities

What are some reasons this HMM gene finder is likely to do poorly?

Finite State Machine

Bad Eukaryotic Gene Finder

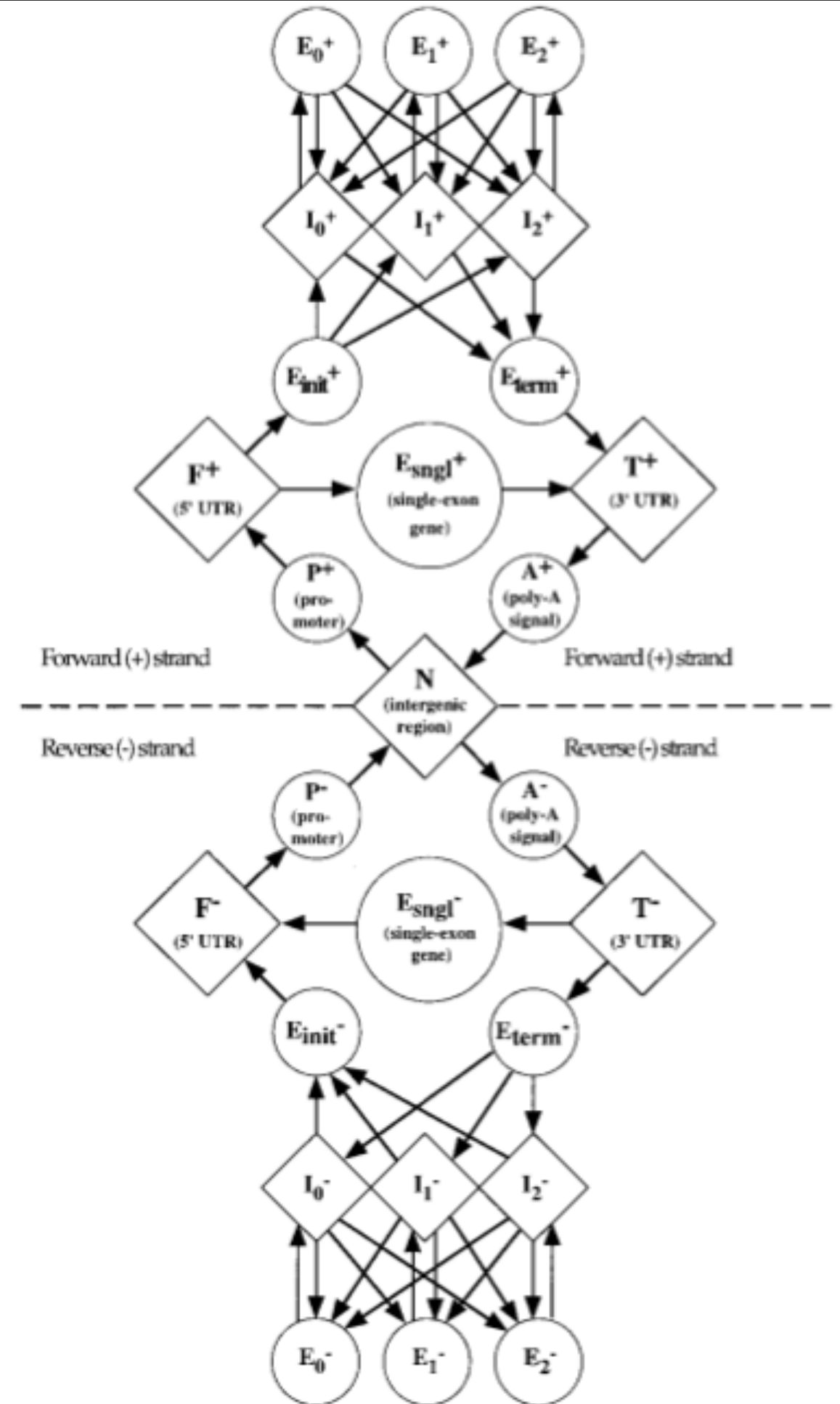
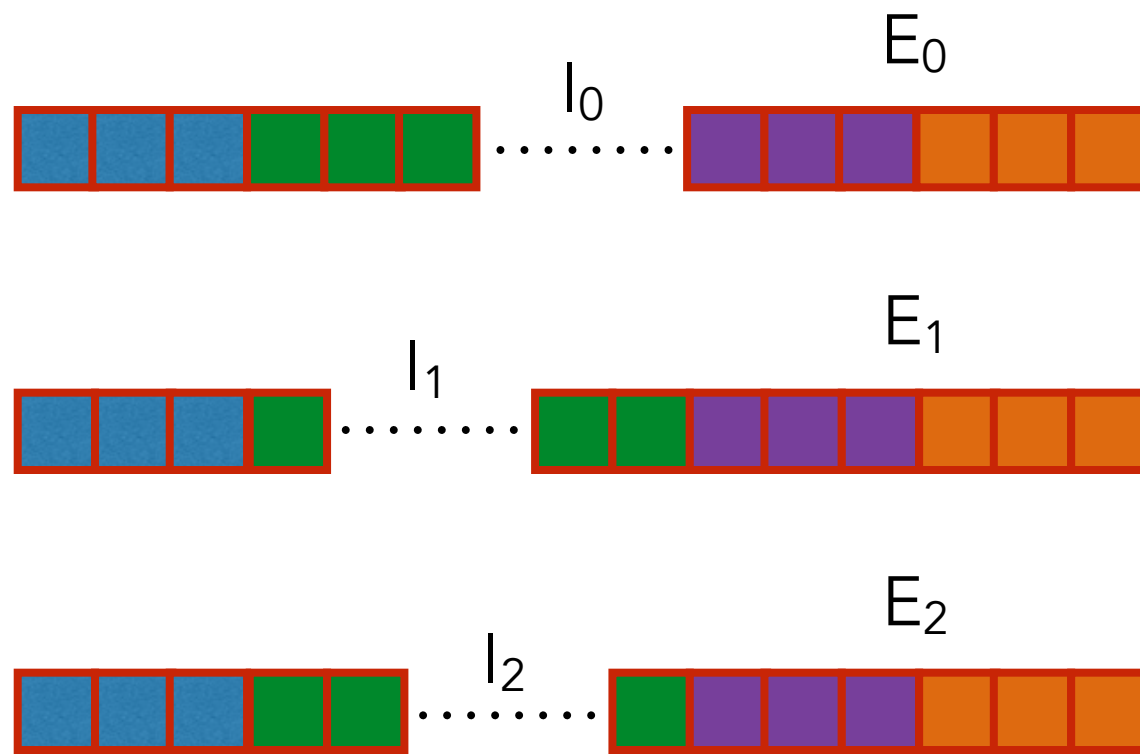
Why is it so bad?

- The positions in the codons are treated independently: the probability of emitting a base can't depend on which previous base was emitted.
- Only one strand of the DNA is considered at once.
- Length distributions of introns and exons are not considered.

Genscan

Burge & Karlin. J. Mol. Biol.
(1997) 268, 78±94

- Explicitly double stranded
- One of the first to handle sequences with ≥ 1 gene in them



Generalized HMMs

- Each state can emit a *sequence* of symbols.
- In the diagram on the previous slide, each state emitted a *complete gene feature* (e.g. an entire exon):

$$\max_{\pi} \prod_{i=1}^n \underbrace{\Pr(x_i \dots x_{i+d_i} \mid \pi_i, d_i)}_{\text{Probability of emitting the string of length } d_i} \underbrace{\Pr(d_i \mid \pi_i)}_{\text{Probability that the state will emit } d_i \text{ symbols.}} \underbrace{\Pr(\pi_i \rightarrow \pi_{i+1})}_{\text{Probability of transitioning to the next state}}$$

Generalized HMMs

- Each state can emit a *sequence* of symbols.
- In the diagram on the previous slide, each state emitted a *complete gene feature* (e.g. an entire exon):

$$\max_{\pi} \prod_{i=1}^n \underbrace{\Pr(x_i \dots x_{i+d_i} \mid \pi_i, d_i)}_{\text{Probability of emitting the string of length } d_i} \underbrace{\Pr(d_i \mid \pi_i)}_{\text{Probability that the state will emit } d_i \text{ symbols.}} \underbrace{\Pr(\pi_i \rightarrow \pi_{i+1})}_{\text{Probability of transitioning to the next state}}$$

This probability could itself be computed by an HMM or a Markov chain, etc.

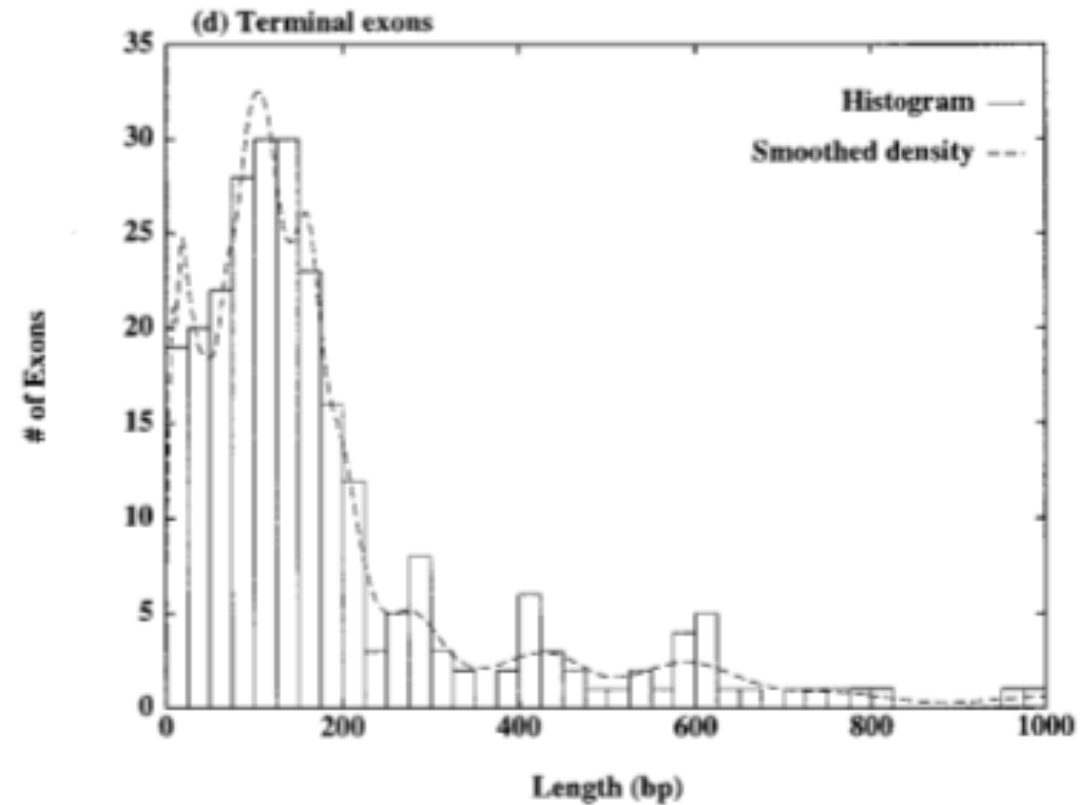
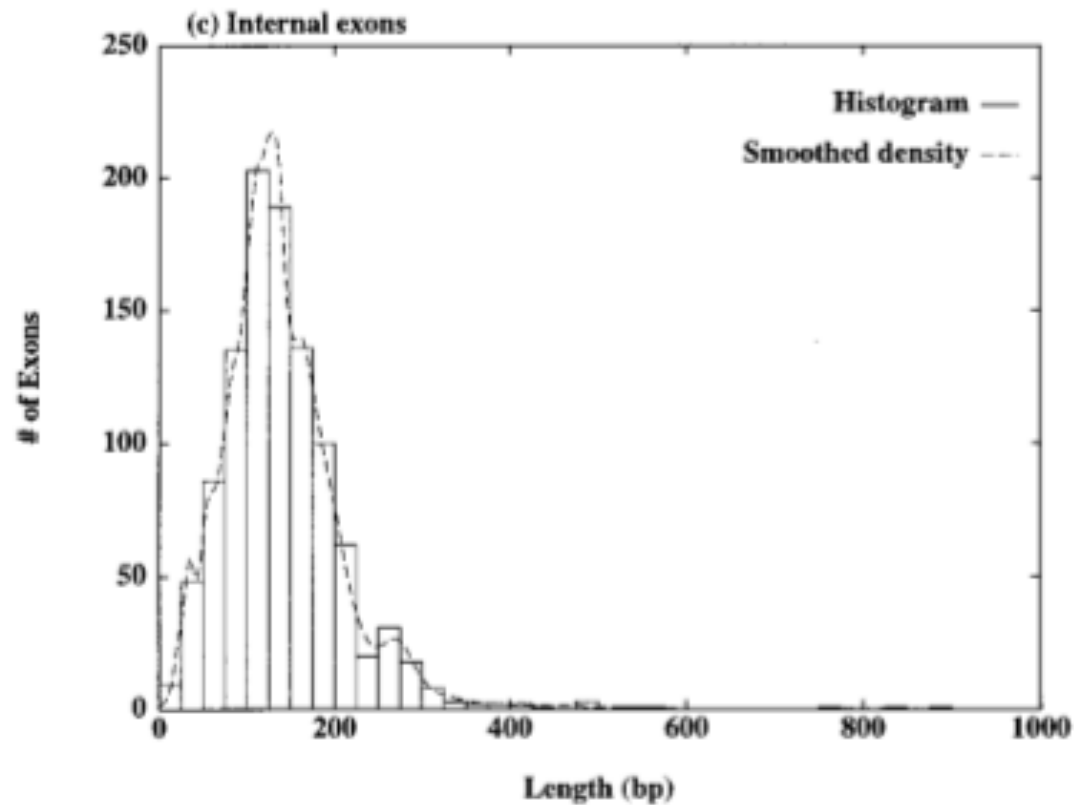
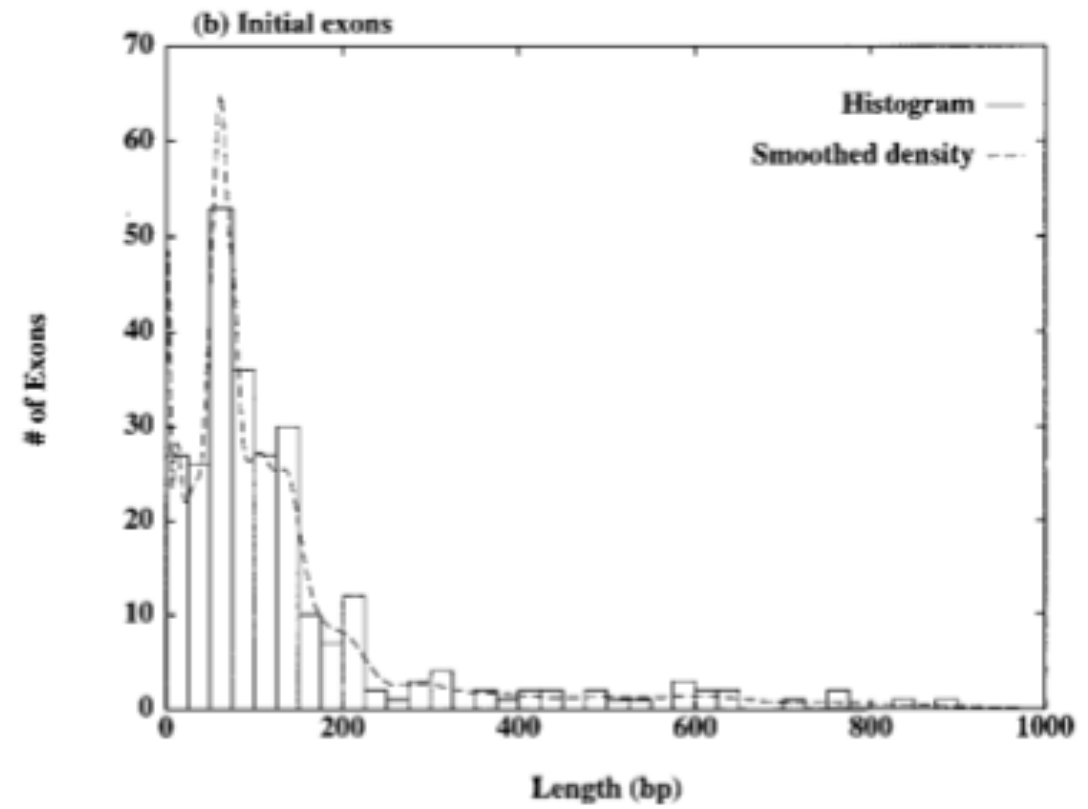
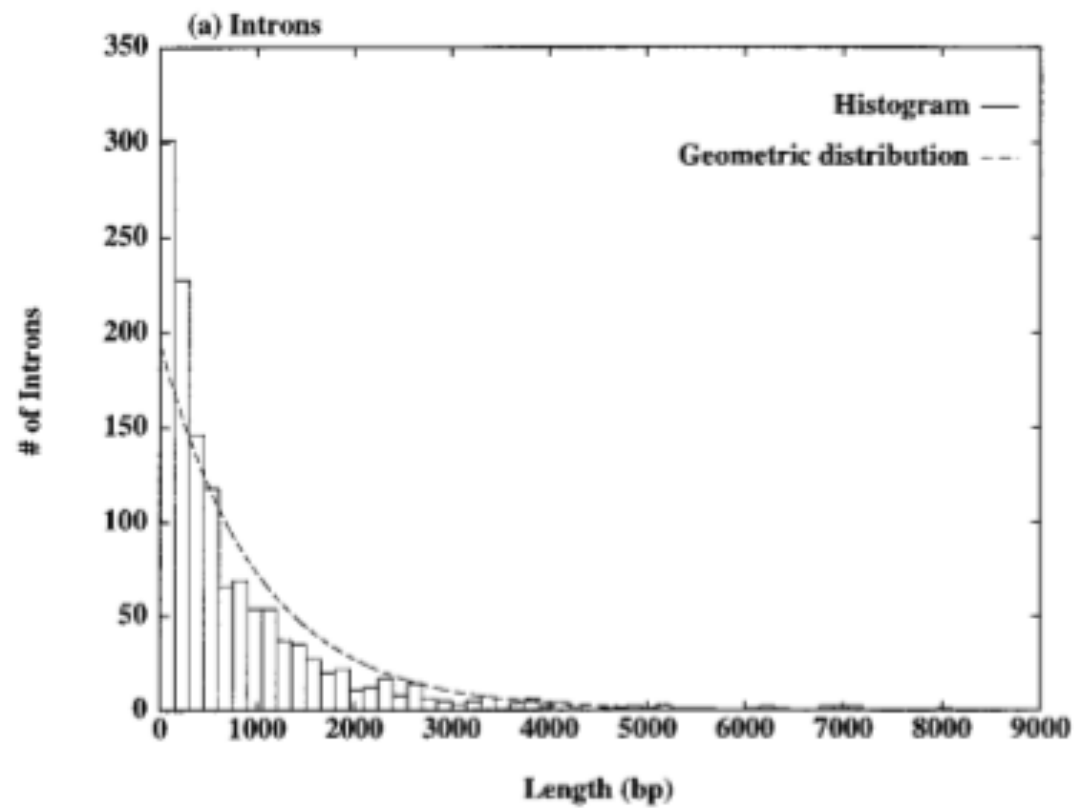
Components Needed

- Probability distribution of initial state
 - = the fraction of known genome corresponding to each state, divided into groups by GC content.
- State transition probabilities
 - = the probability X follows Y in known genes
- Length distributions for each state
 - For exons: = estimated from empirically observed distribution (next slide)
 - For introns: = geometric distribution with parameter q_{GC} , where is the best fit parameter for regions with a given GC content.
- Sequence models for each state/length

for states with
strong motifs:

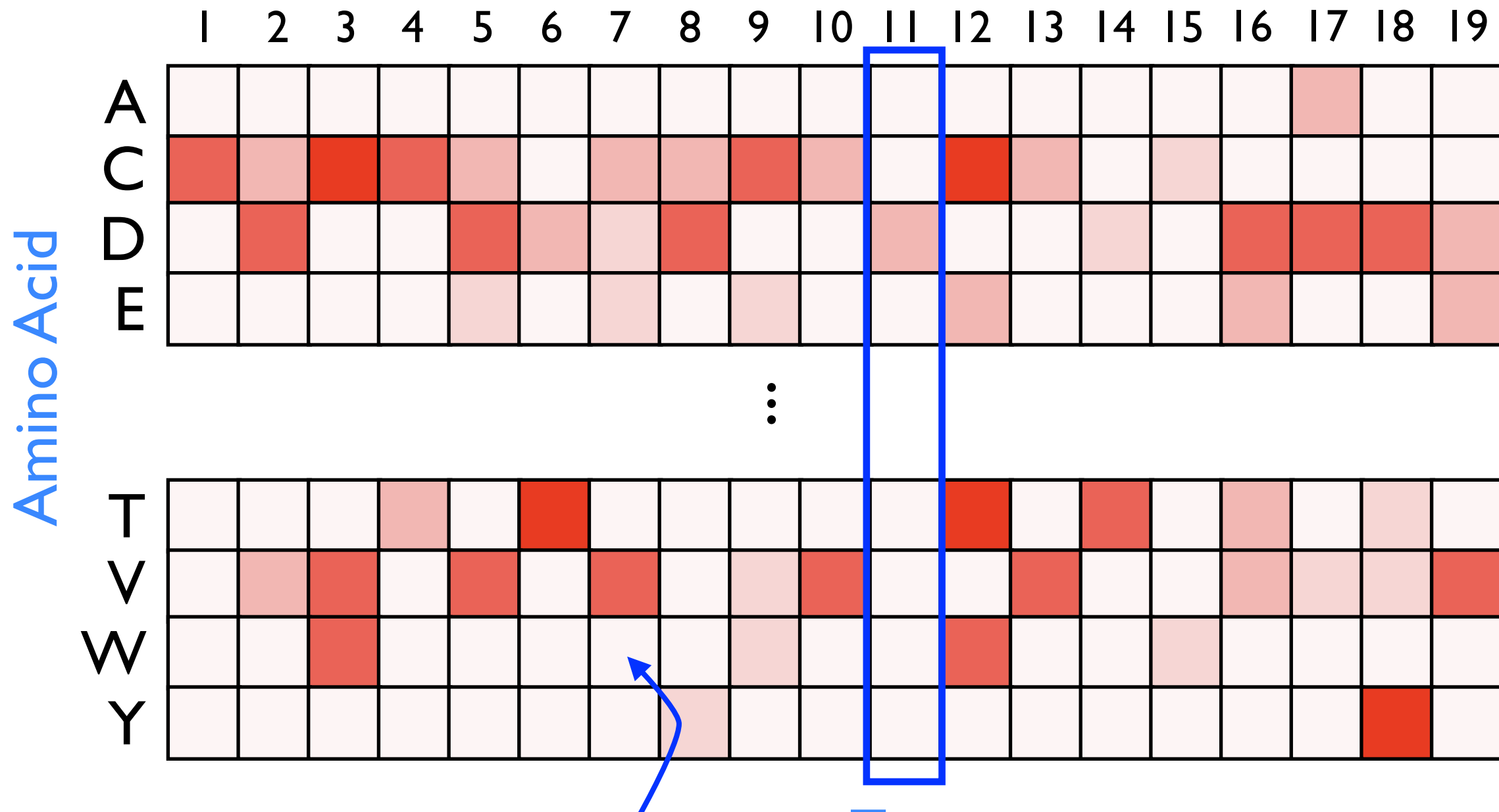


Feature Length Distributions



Sequence Profiles (PSSM)

Motif Position



Color \approx Probability that the i^{th} position has the given amino acid = $e_i(x)$.

$$\sum = 1$$

Sequence Generators

Exons: 3 different 5th-order Markov models:

- 1 model for each base of a codon
- Sequence generated by repeatedly applying model 1, then 2, then 3, and so on.
- Separate models for regions with GC content $< 43\%$

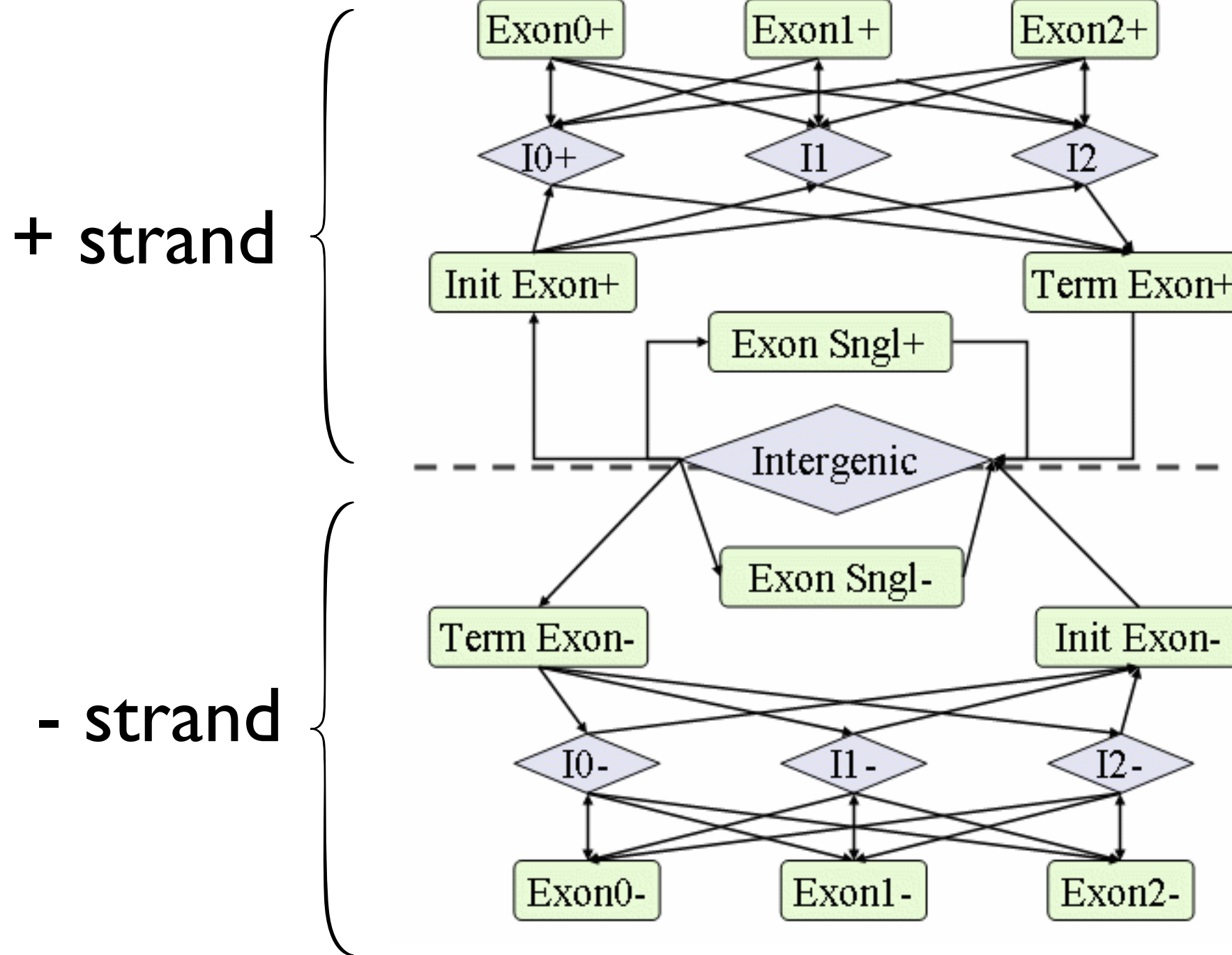
Non-coding states: (F, T, I_i)

- 5th-order Markov model
- Separate model for regions with GC content $< 43\%$

Acceptor / donor sites: a more complicated model that accounts for dependencies between positions.

GlimmerHMM

Majoros et al, 2004



Differences:

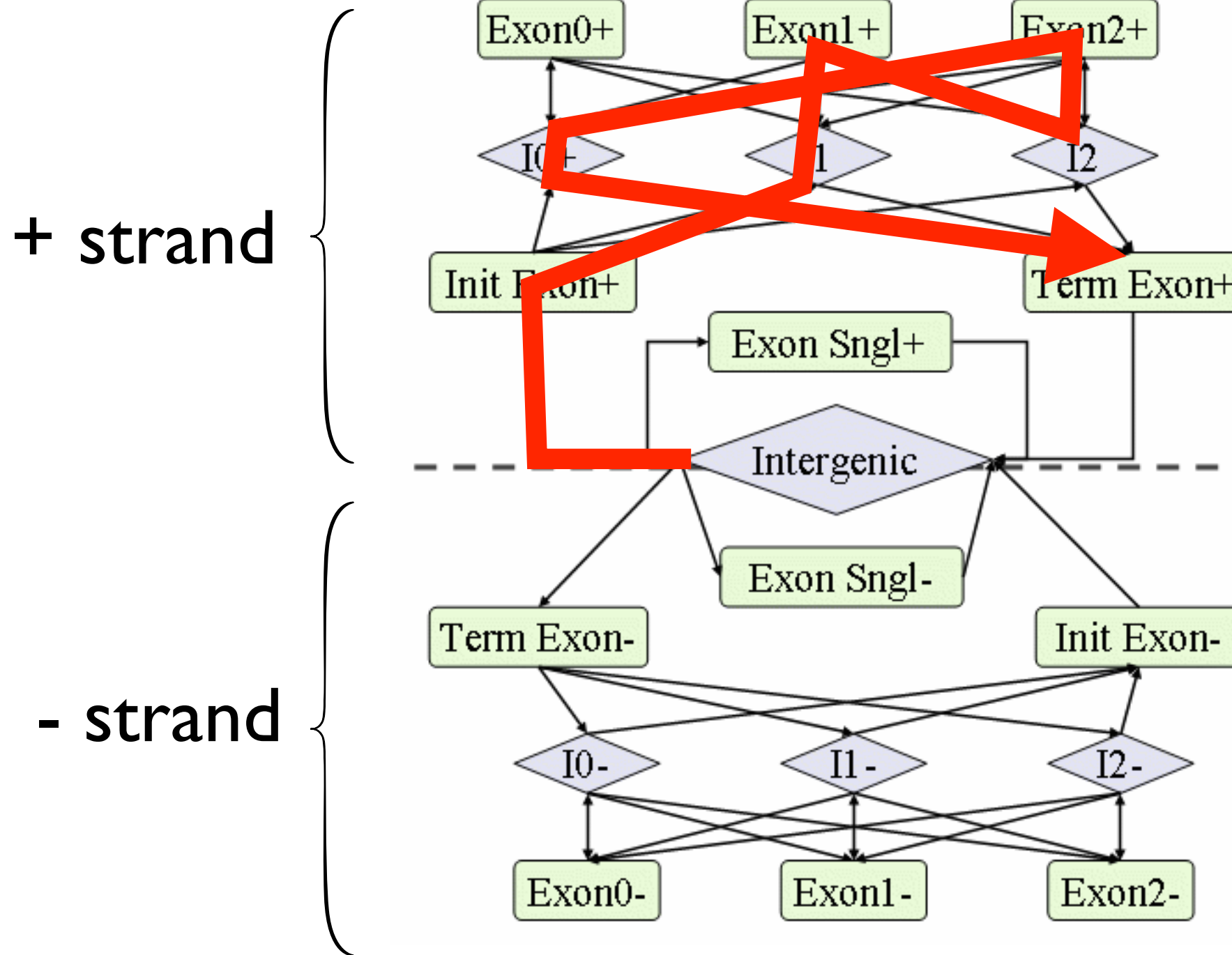
Interpolated HMM for coding sequences

New splicing model

GlimmerHMM model

GlimmerHMM

Majoros et al, 2004



Differences:

Interpolated HMM for coding sequences

New splicing model

GlimmerHMM model

GlimmerHMM Performance

% of predicted in-gene nucleotides that are correct

% of predicted exons that are true exons.

	<i>Nuc Sens</i>	<i>Nuc Prec</i>	<i>Nuc Accur</i>	<i>Exon Sens</i>	<i>Exon Prec</i>	<i>Exact Genes</i>	<i>Size of test set</i>
<i>D.rerio</i>	93%	78%	86%	77%	69%	24%	549 genes
<i>C.elegans</i>	96%	95%	96%	82%	81%	42%	1886 genes
<i>Arabidopsis</i>	97%	99%	98%	84%	89%	60%	809 genes
<i>Cryptococcus</i>	96%	99%	98%	86%	88%	53%	350 genes
<i>Coccidioides</i>	99%	99%	99%	84%	86%	60%	503 genes
<i>Brugia</i>	93%	98%	95%	78%	83%	25%	477 genes

% of true gene nucleotides that GlimmerHMM predicts as part of genes.

% of true exons that GlimmerHMM found.

% of genes perfectly found

Compare with GENSCAN

- On 963 human genes:

	<i>Nuc Sens</i>	<i>Nuc Prec</i>	<i>Nuc Acc</i>	<i>Exon Sens</i>	<i>Exon Prec</i>	<i>Exon Acc</i>	<i>Exact Genes</i>
<i>GlimmerHMM</i>	86%	72%	79%	72%	62%	67%	17%
<i>Genscan</i>	86%	68%	77%	69%	60%	65%	13%

- Note that overall accuracy is pretty low.

Generalized Pair HMMs

Use: find genes simultaneously in 2 genomes
increased signal b/c the structure of homologous genes is often very similar.

- Pair: Each state emits two symbols, one for each sequence
- Generalized Pair: a pair of lengths d, e is drawn from a joint probability distribution and a pair of sequences X, Y of length d, e , respectively, are generated at each state.

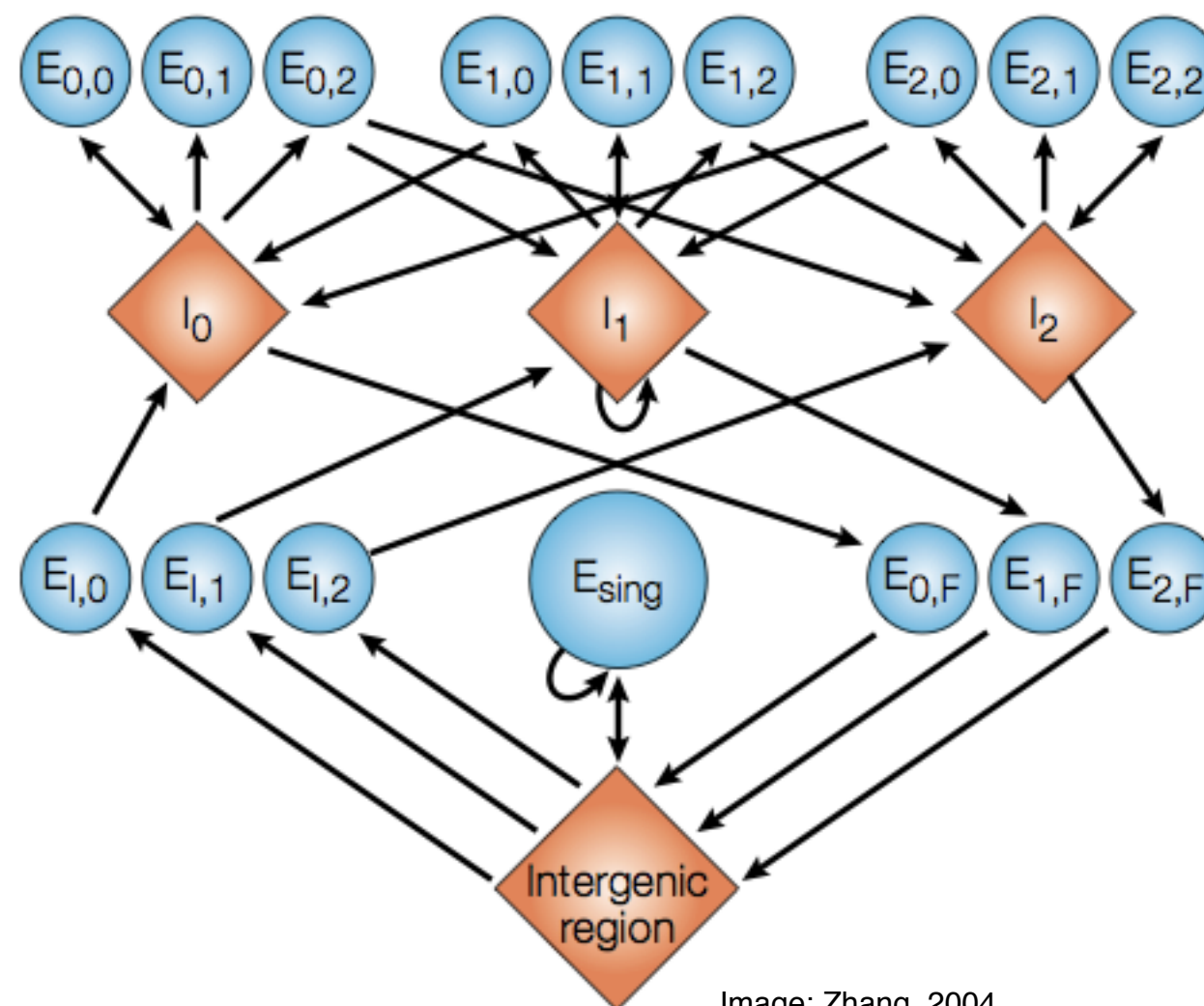


Image: Zhang, 2004

Pachter et al. J Comp Biol, 9(2), 2002

Reverse strand: mirror reflection of above

Generalized Pair HMMs

Use: find genes simultaneously in 2 genomes
increased signal b/c the structure of homologous genes is often very similar.

- Pair: Each state emits two symbols, one for each sequence
- Generalized Pair: a pair of lengths d, e is drawn from a joint probability distribution and a pair of sequences X, Y of length d, e , respectively, are generated at each state.

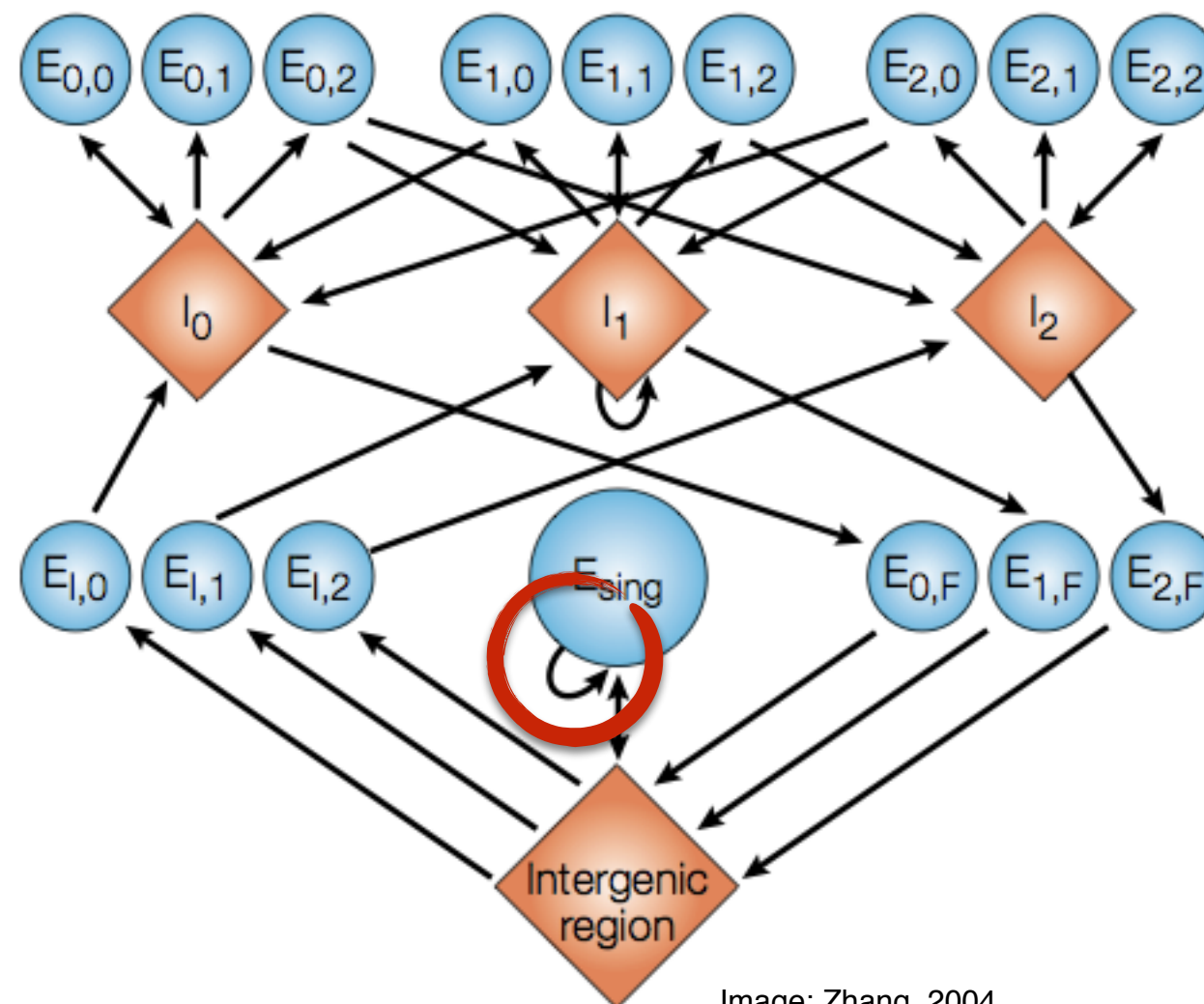
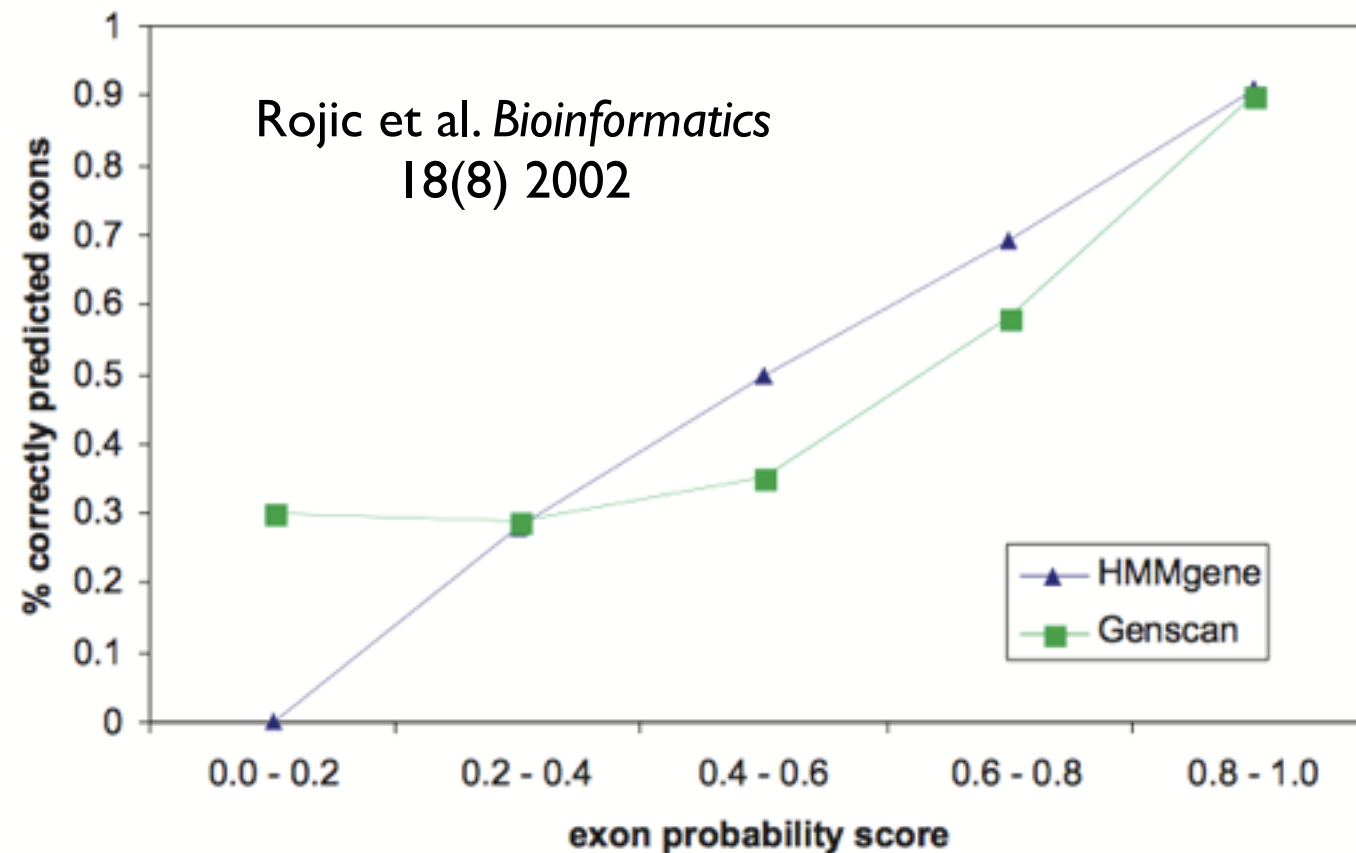


Image: Zhang, 2004

Pachter et al. J Comp Biol, 9(2), 2002

Reverse strand: mirror reflection of above

Combining Several Predictors



- Use each programs exon probability scores (probability that exon is included in the parse).
- Example: keep disagreeing exons only if score is above a threshold.

Recap

- Simple gene finding approaches use codon bias and long ORFs to identify genes.
- Many top gene finding programs for Eukaryotes are based on generalizations of Hidden Markov Models because multiple types of signals are present in a gene (intron, exon, etc.)
- Basic HMMs must be generalized to emit variable sized strings.