CSEP 527
Computational Biology

Genes and Gene Prediction

Gene Finding: Motivation

Sequence data flooding in

What does it mean?

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

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

Protein Coding Nuclear DNA

Focus of these slides

Goal: Automated annotation of new seq data

State of the Art:

- In Eukaryotes:
  - predictions ~ 60% similar to real proteins
  - ~80% if database similarity used
- Prokaryotes
  - better, but still imperfect
- Lab verification still needed, still expensive
- Largely done for Human; unlikely for most others
  - and *non*-coding is poorly understood even in human

Biological Basics

Central Dogma:

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

Codons: 3 bases code one amino acid

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

RNA Transcription

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

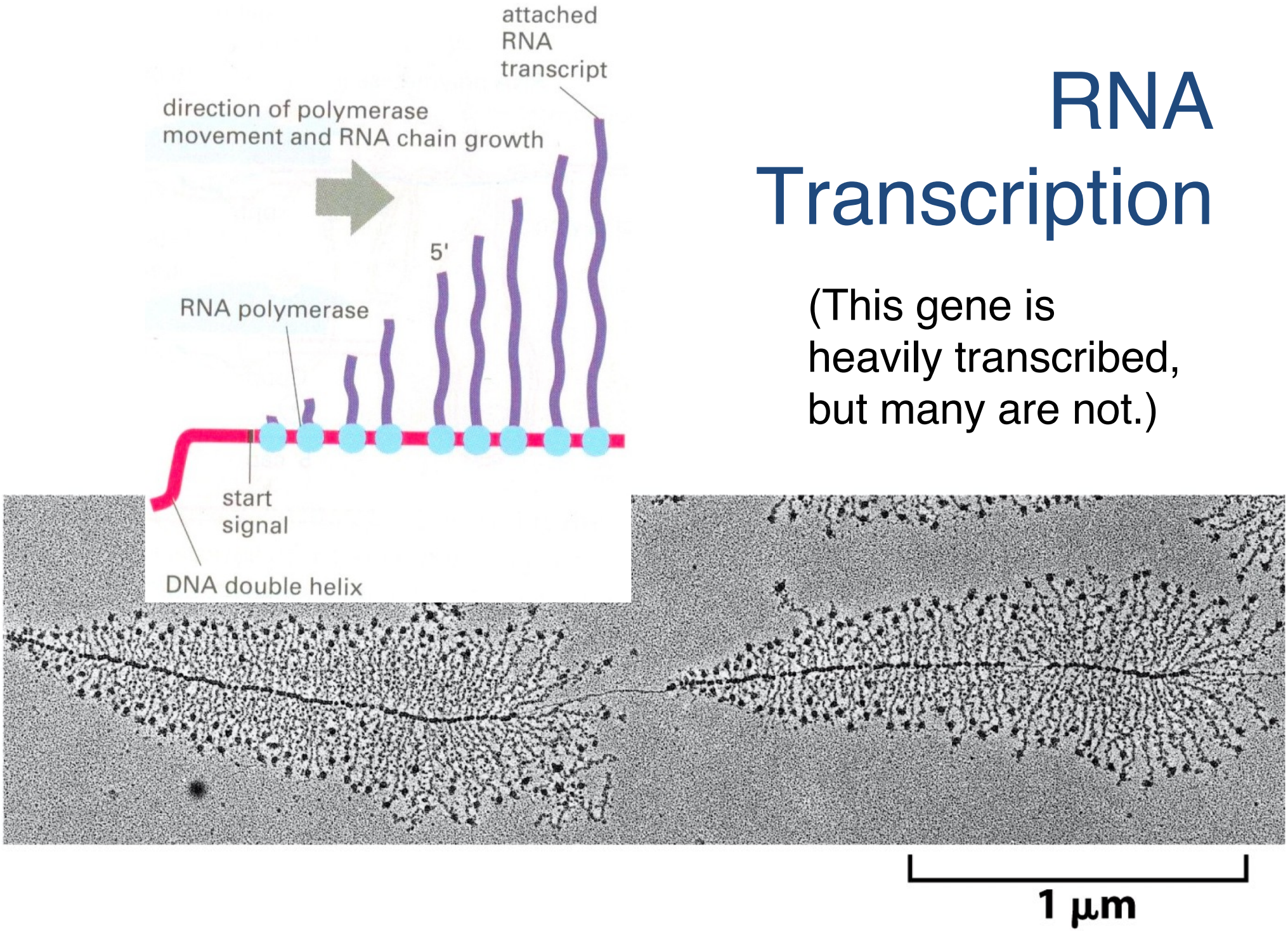
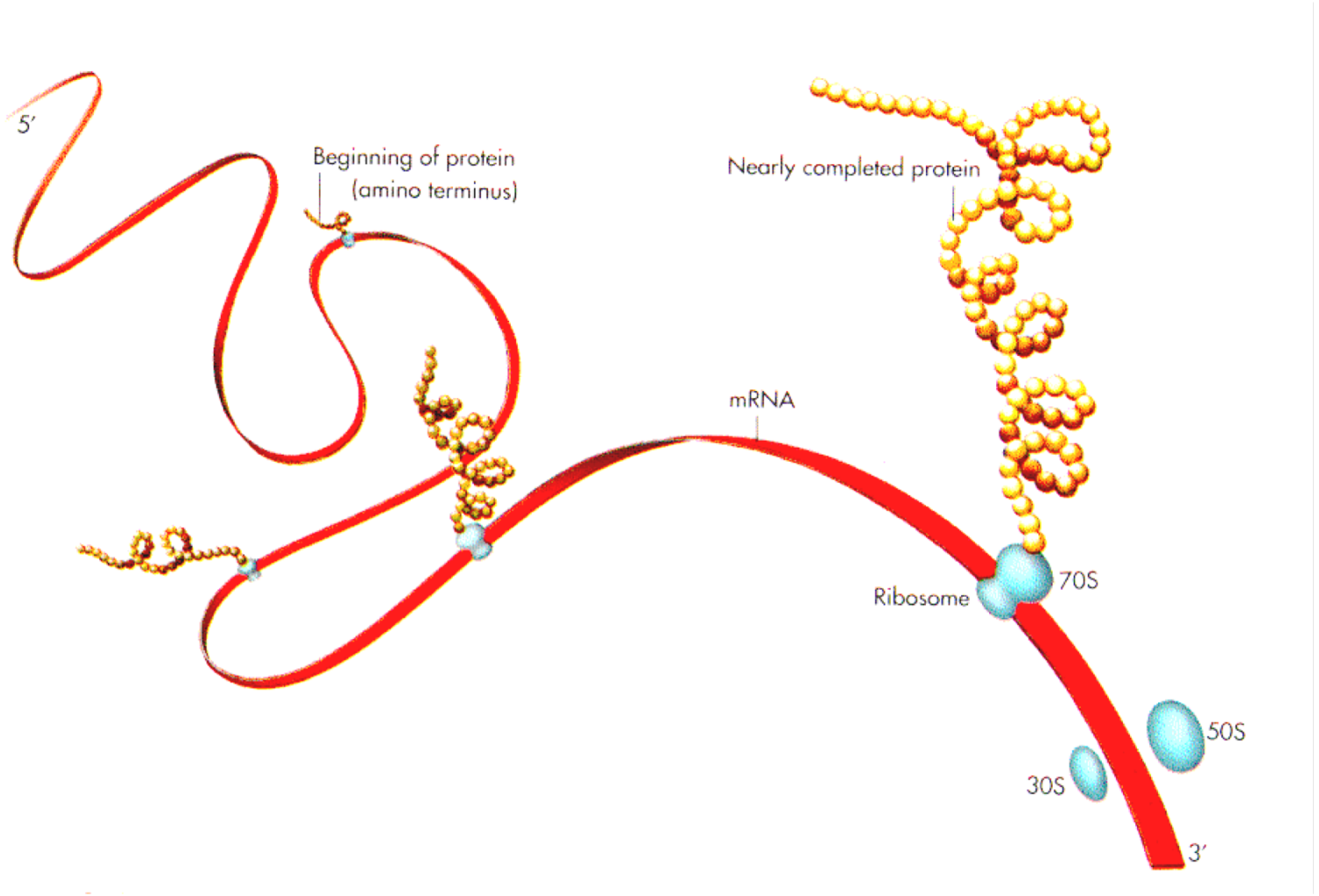


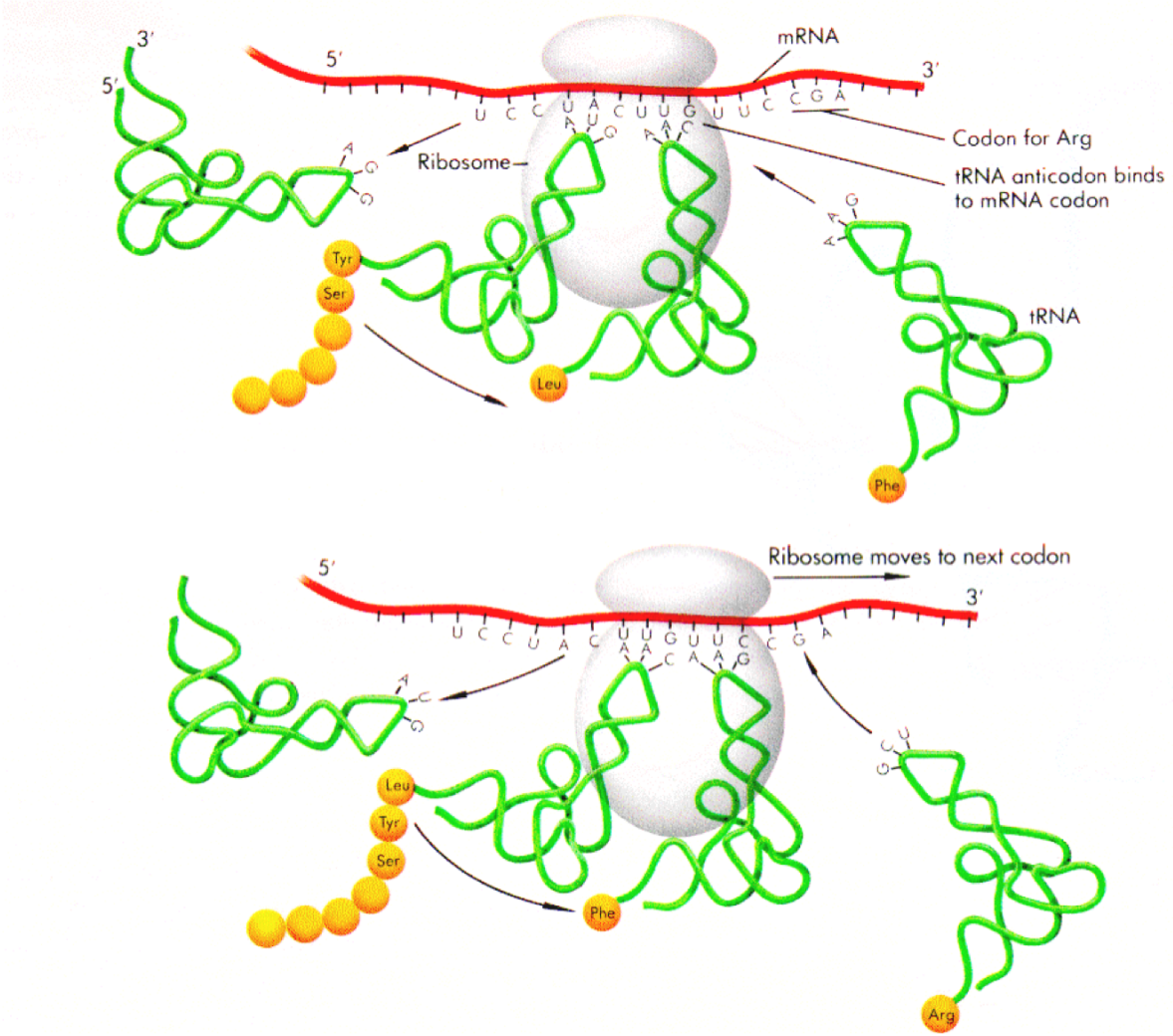
Figure 6-9 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

Ribosomes



Watson, Gilman, Witkowski, & Zoller, 1992

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

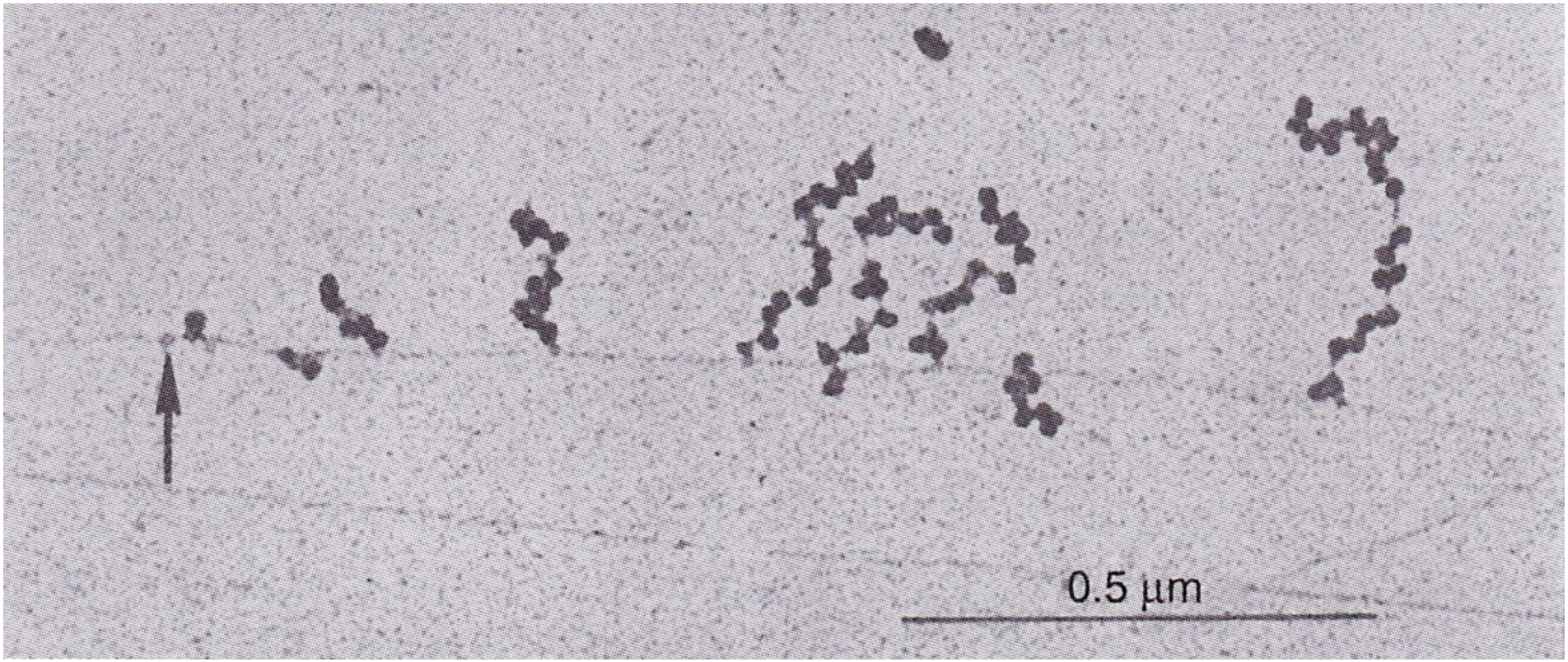


Figure 3-7. Coupled transcription/translation in bacteria is visualized. Oscar Miller and colleagues lysed *E. coli* cells and immediately collected the cell contents on electron microscope grids. They saw threads of mRNA still associated with DNA (thin lines), and ribosomes—several at a time—were already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

Darnell, p120

Codons & The Genetic Code

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

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

Idea #1: Find Long ORF's

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

Open Reading Frame: No internal stop codons

In random DNA

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

But average protein ~ 1000bp

A Simple ORF finder

start at left end

scan triplet-by-non-overlapping triplet for AUG

then continue scan for STOP

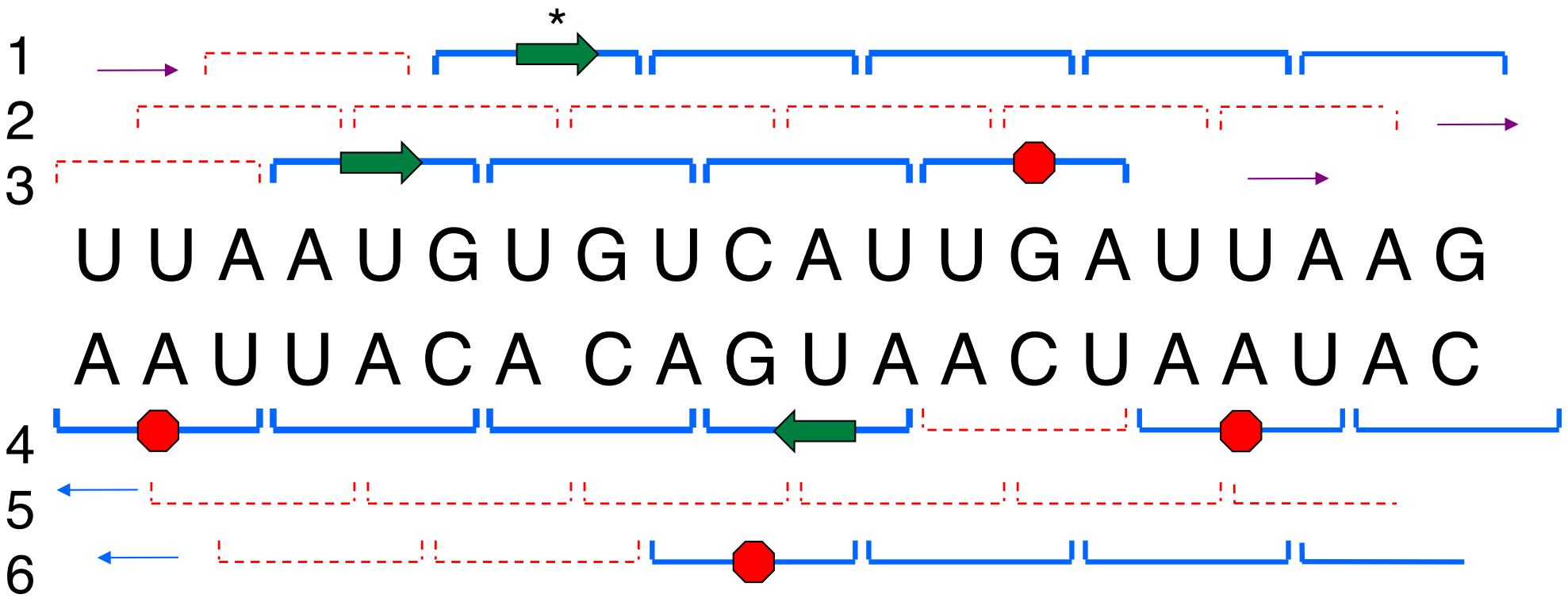
repeat until right end

repeat all starting at offset 1

repeat all starting at offset 2

then do it again on the other strand

Scanning for ORFs



* In bacteria, GUG is sometimes a start codon…

Idea #2: Codon Frequency

In random DNA

Leucine : Alanine : Tryptophan = 6 : 4 : 1

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

So, coding DNA is not random

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

examples known with 90% AT 3rd base

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

Idea #3: Non-Independence

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

How to model this? Markov models

Markov Chains

A sequence $x_1, x_2, \ldots$ of random variables is a *k-th order Markov chain* if, for all *i*, *i*th value is independent of all but the previous *k* values:

$$P(x_i \mid x_1, x_2, \ldots, x_{i-1}) = P(x_i \mid x_{i-k}, x_{i-k+1}, \ldots, x_{i-1})$$

i-1

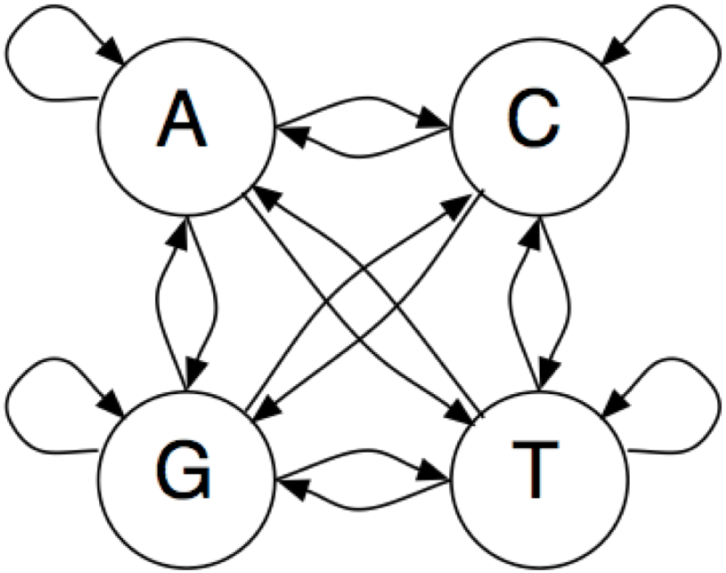
k typically ≪ i-1

Example 1: Uniform random ACGT } 0th order

Example 2: Weight matrix model } 0th order

Example 3: ACGT, but ↓ Pr(G following C) } 1st order

A Markov Model (1st order)



States: A,C,G,T

Emissions: corresponding letter

Transitions: $a_{st} = P(x_i = t \mid x_{i-1} = s)$ ⟵ 1st order

Pr of emitting sequence *x*

$$
\begin{aligned}
x &= x_1\ x_2\ \ldots\ x_n \\
P(x) &= P(x_1, x_2, \ldots, x_n) \\
&= P(x_1) \cdot P(x_2 \mid x_1) \cdots P(x_n \mid x_{n-1}, \ldots, x_1) \\
&= P(x_1) \cdot P(x_2 \mid x_1) \cdots P(x_n \mid x_{n-1}) \\
&= P(x_1) \prod_{i=1}^{n-1} a_{x_i, x_{i+1}} \\
&= \prod_{i=0}^{n-1} a_{x_i, x_{i+1}} \quad \text{(with Begin state)}
\end{aligned}
$$

a law of probability ("the chain rule")

if 1st order MC

Discrimination/Classification

Log likelihood ratio of CpG model vs background model

$$S(x) = \log \frac{P(x|\text{model}+)}{P(x|\text{model}-)} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}} = \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

β	A	C	G	T
A	−0.740	0.419	0.580	−0.803
C	−0.913	0.302	1.812	−0.685
G	−0.624	0.461	0.331	−0.730
T	−1.169	0.573	0.393	−0.679

From DEKM

CpG Island Scores

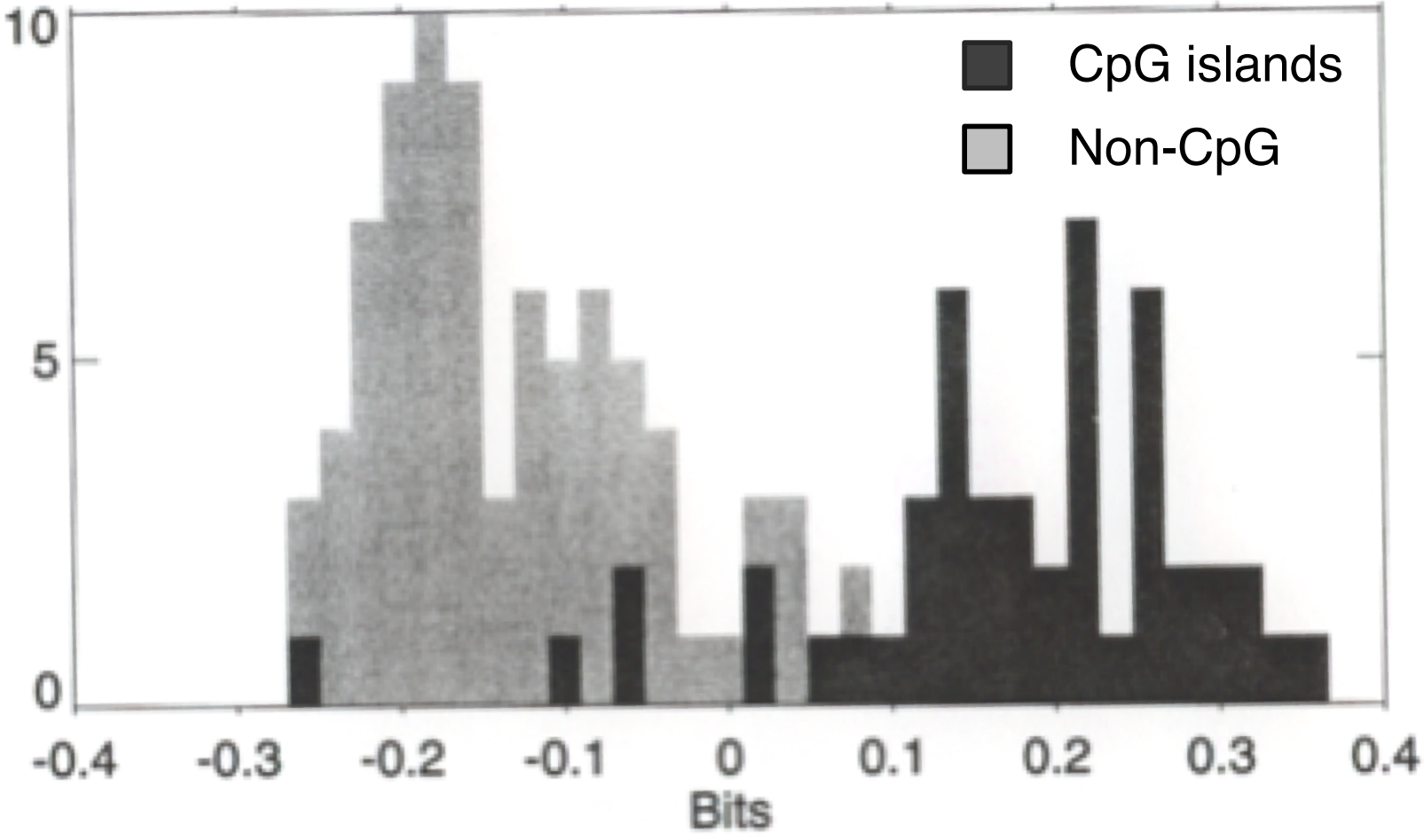


Figure 3.2 *Histogram of length-normalized scores.*

From DEKM

A Gene Finding Example

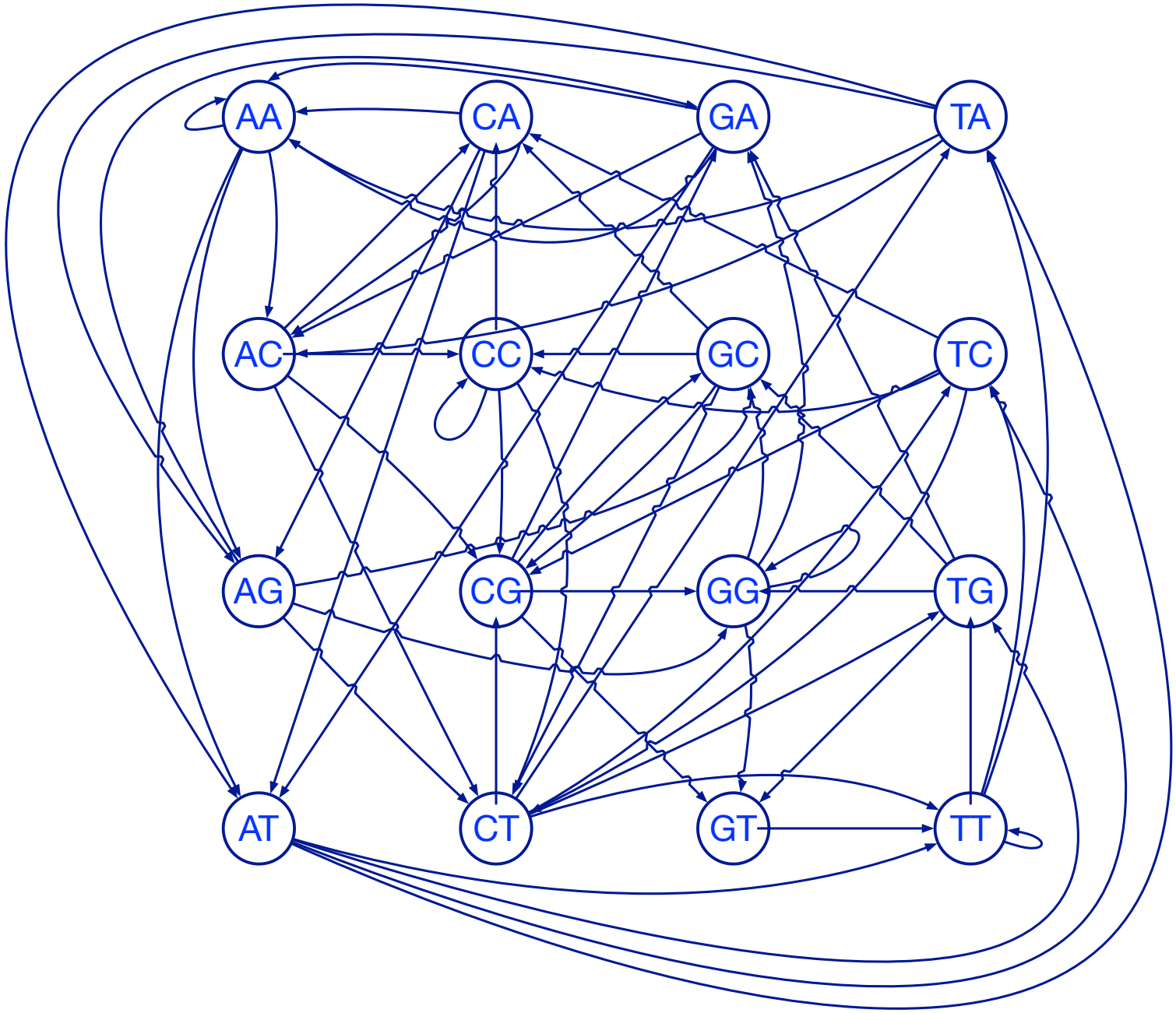
All +-strand ORFs in a prokaryote

"Truth" based on stop codon matching a Genbank-annotated protein coding gene

Built ROC curve for classification by:

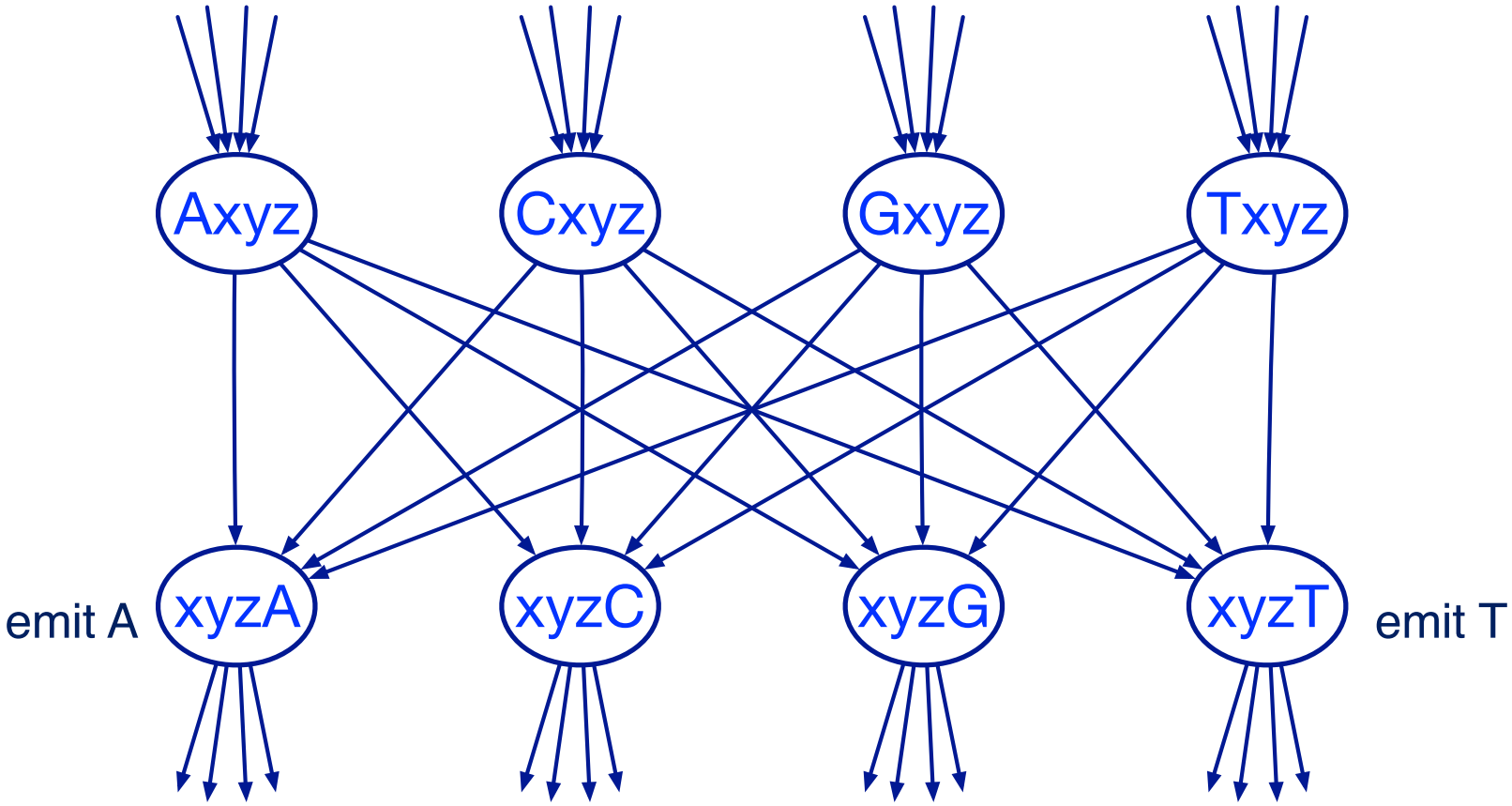
- length
- 6-th order Markov model
- both

2nd order Markov Model

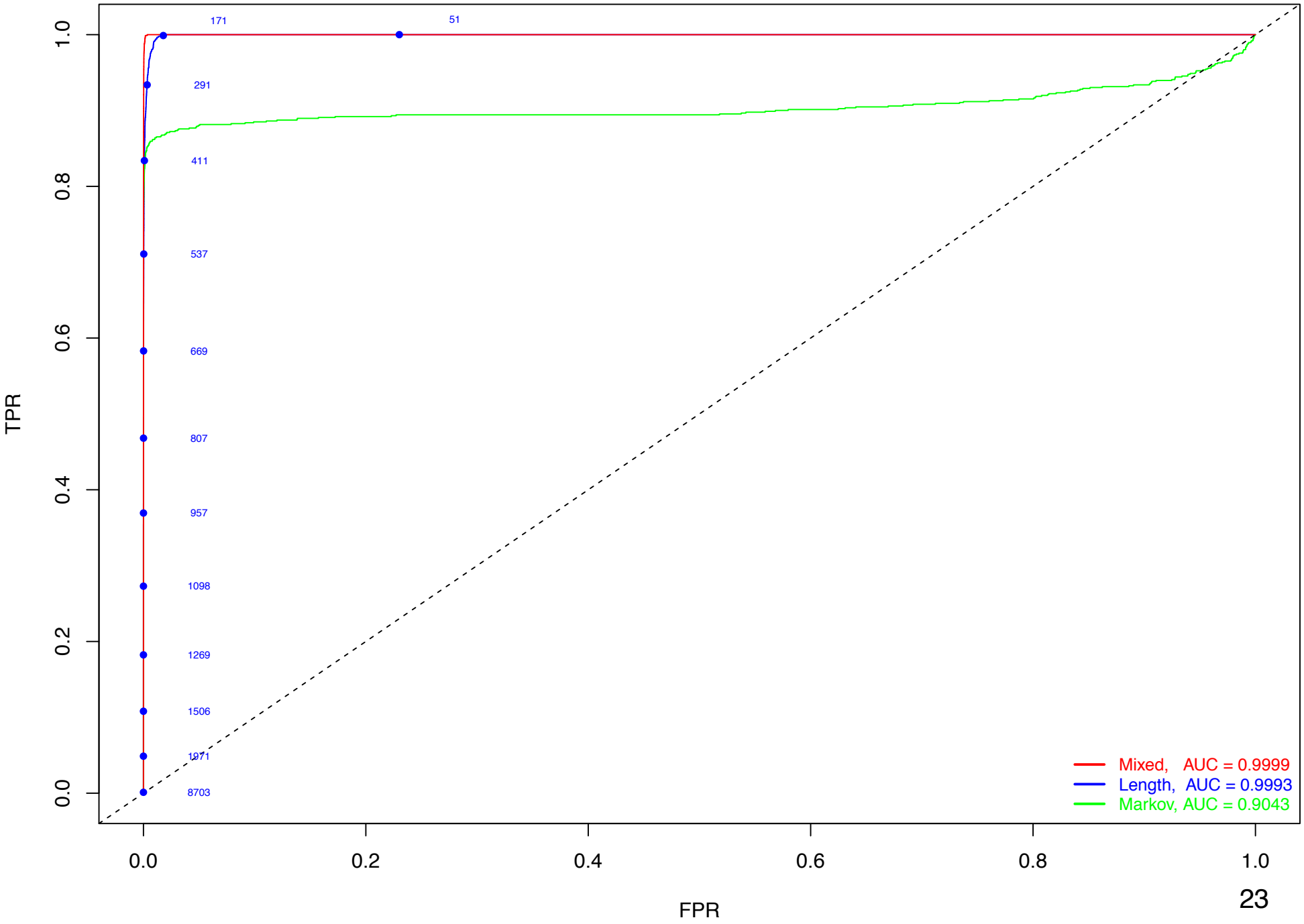


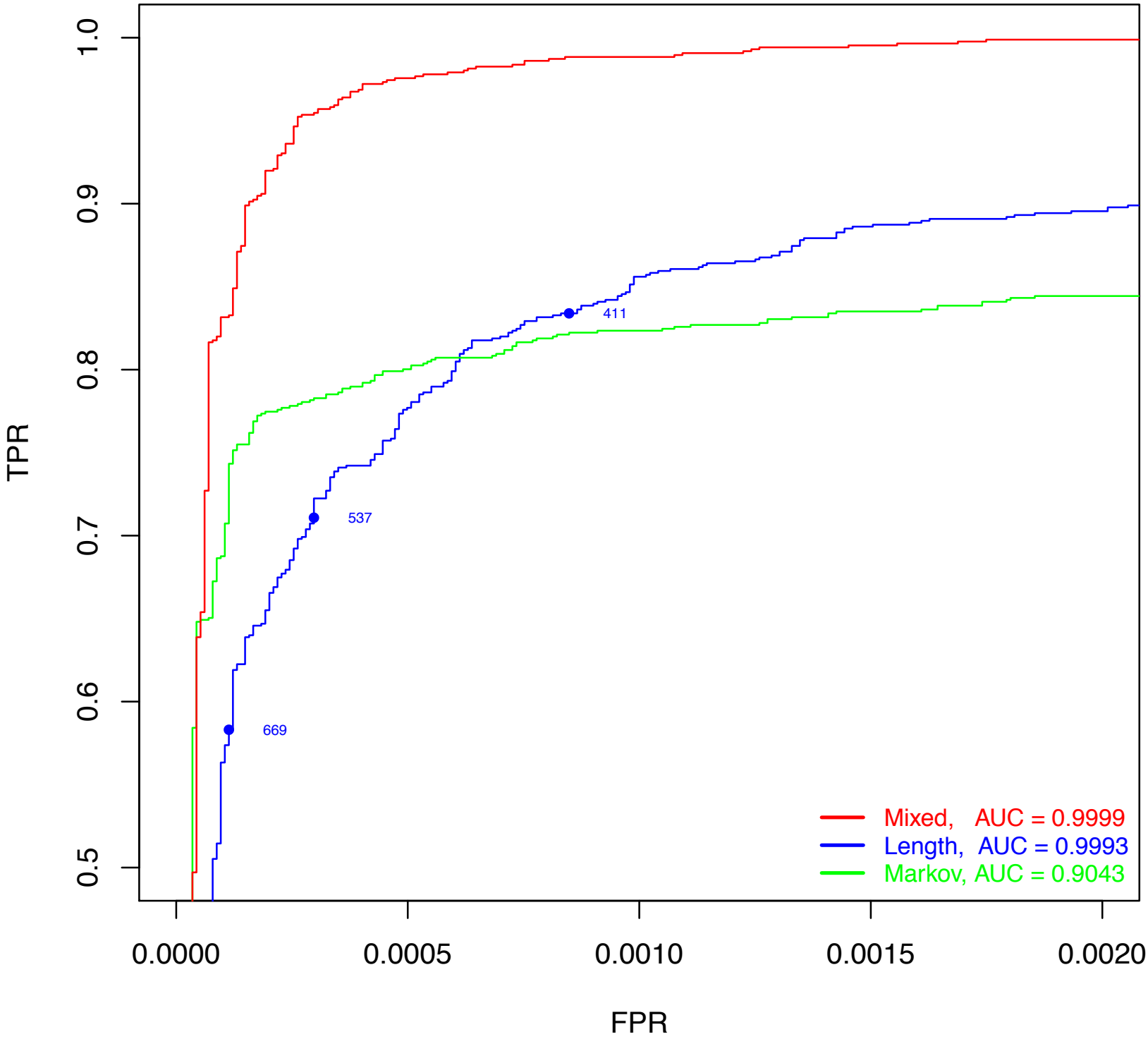
k^{th} Order Markov Model

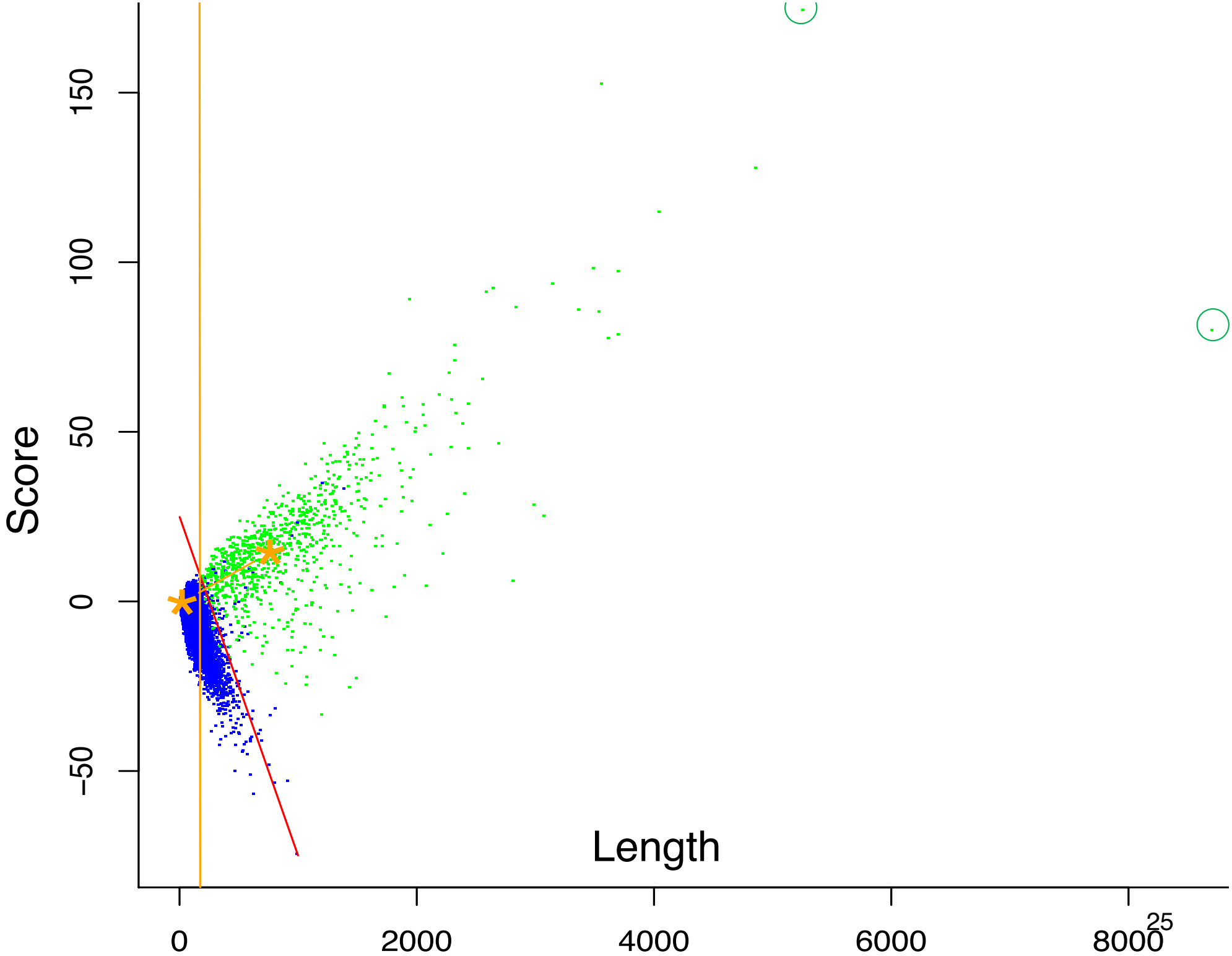
4^k states, each in/out-degree 4, joined as follows:



(where xyz is a length k-1 string over {A,C,G,T}, and, e.g., Axyz and xyzA are the *same* state when $xyz=A^{k-1}$)


171
51
291
411
537
669
807
957
1098
1269
1506
1971
8703
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
Mixed, AUC = 0.9999
Length, AUC = 0.9993
Markov, AUC = 0.9043


TPR
1.0
0.9
0.8
0.7
0.6
0.5
0.0000
0.0005
0.0010
0.0015
0.0020
FPR
669
537
411
Mixed, AUC = 0.9999
Length, AUC = 0.9993
Markov, AUC = 0.9043


Score
150
100
50
0
−50
Length
0
2000
4000
6000
8000

Summary

In prokaryotes, most DNA is coding

E.g. ~ 70% in *H. influenzae*

Long ORFs + codon/nucleotide stats do well

Can improve by modeling associated features (TATA boxes, promoters, etc.)

e.g. via WMM or higher-order Markov models

But obviously won't be perfect

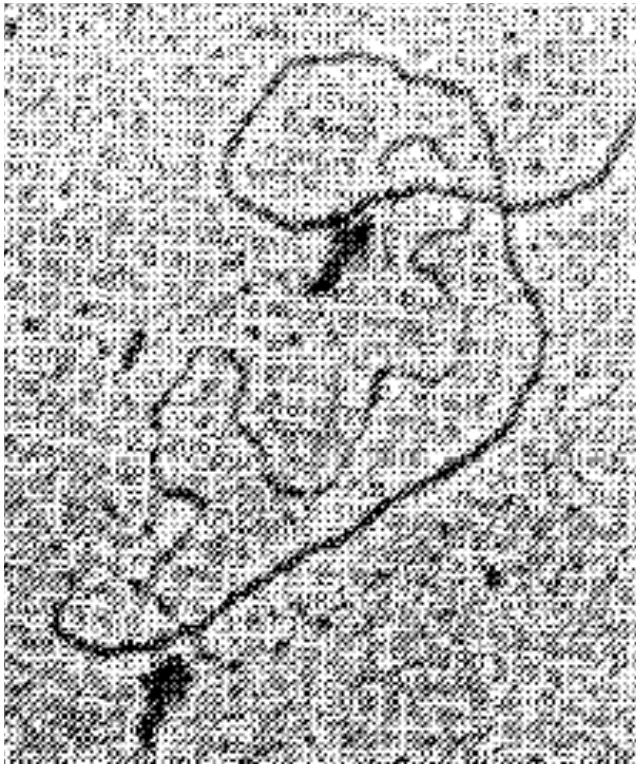
short genes, frame shifts, 5′ & 3′ UTR's, …

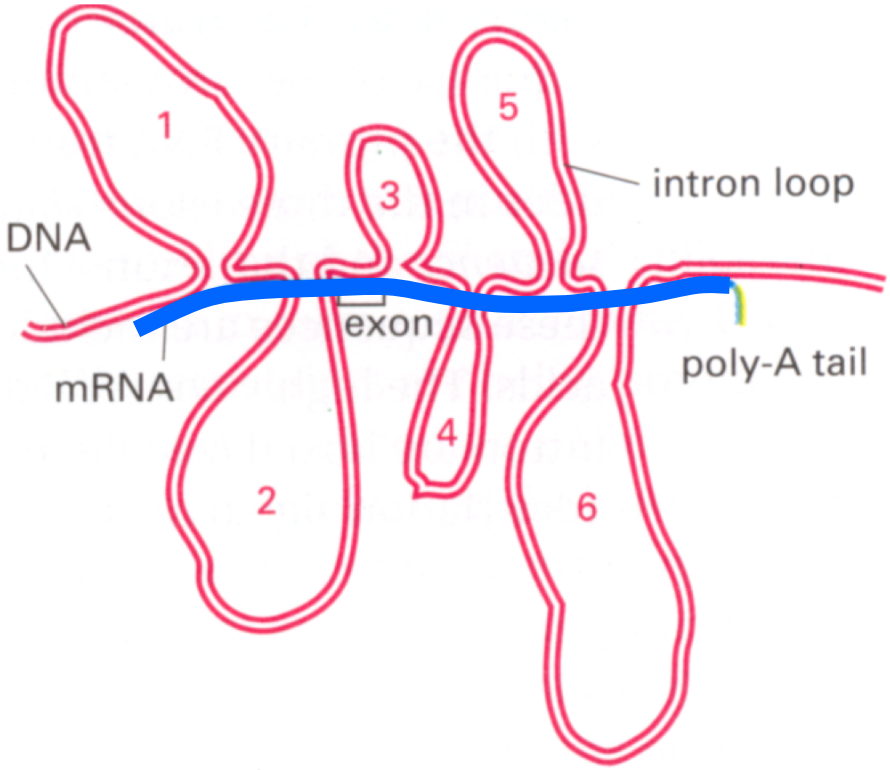
Eukaryotes

As in prokaryotes (but maybe more variable)

- promoters
- start/stop transcription
- start/stop translation

And then…





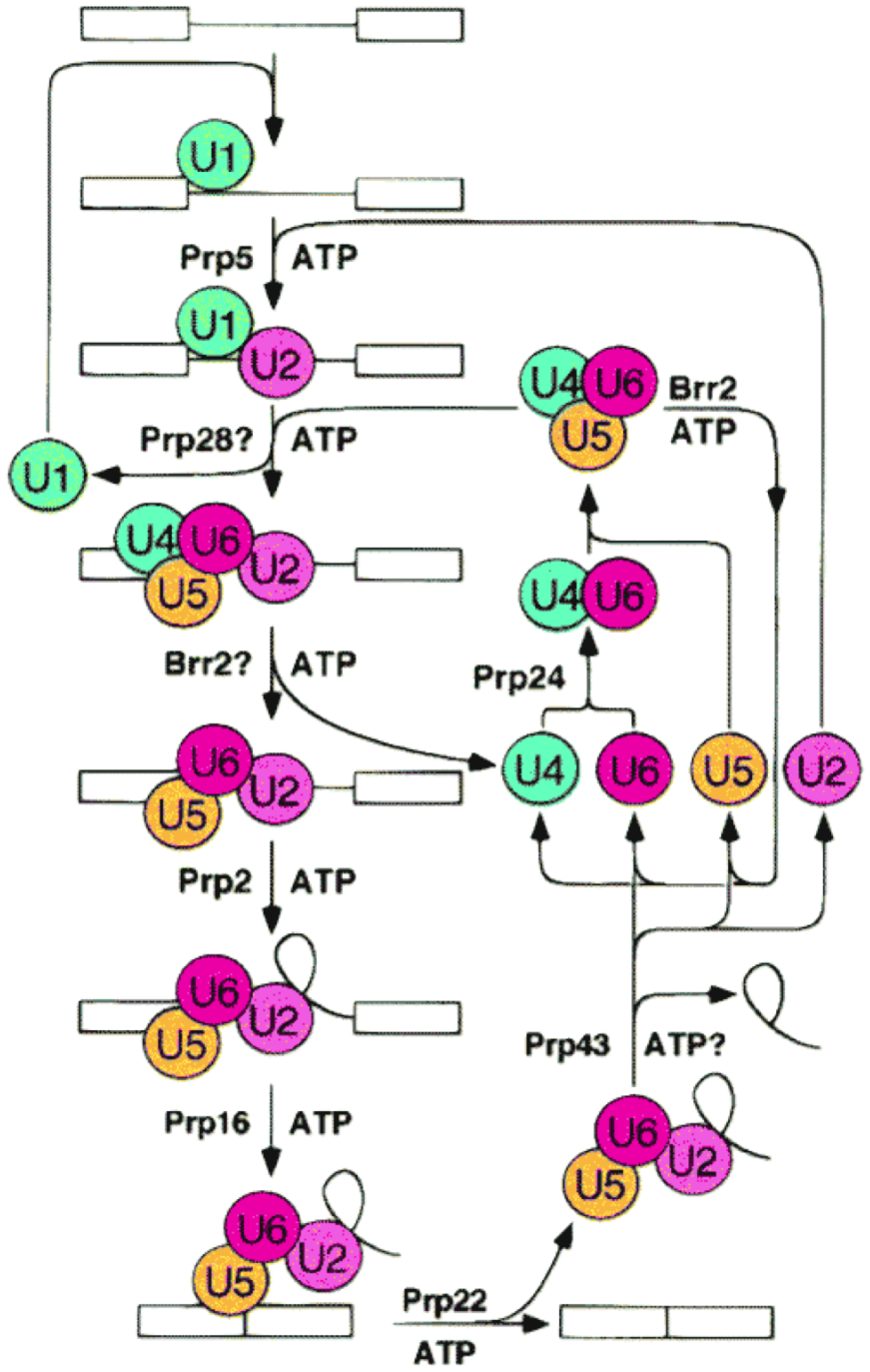
Nobel Prize of the week: P. Sharp, 1993, Splicing

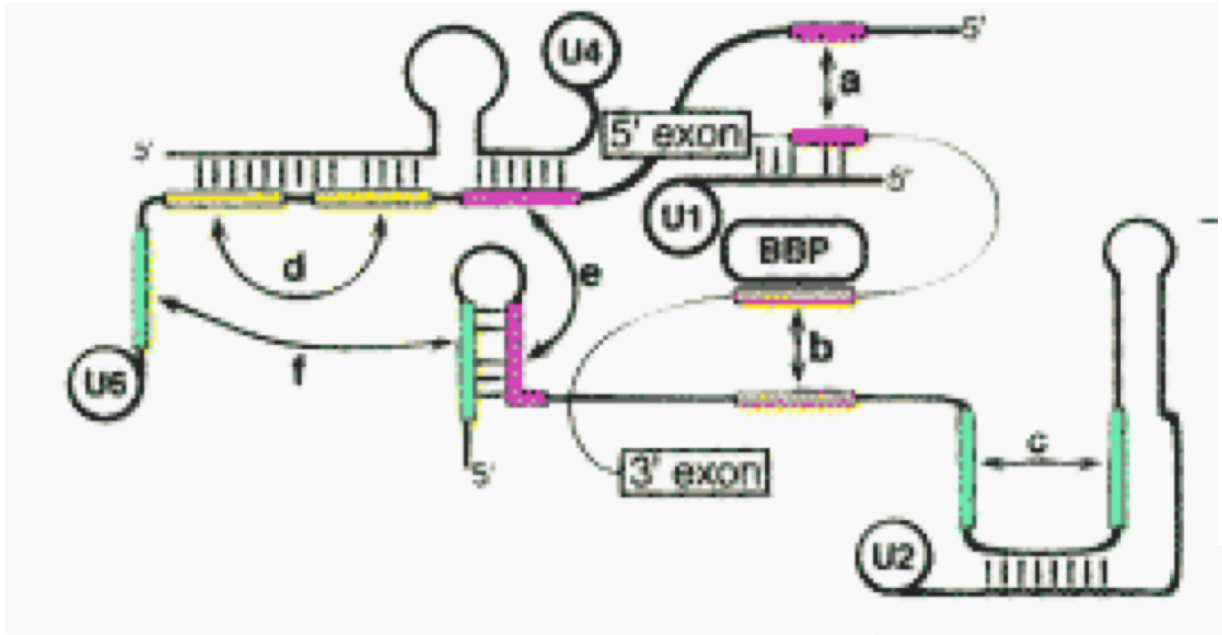
Mechanical Devices of the Spliceosome: Motors, Clocks, Springs, and Things

Jonathan P. Staley and Christine Guthrie

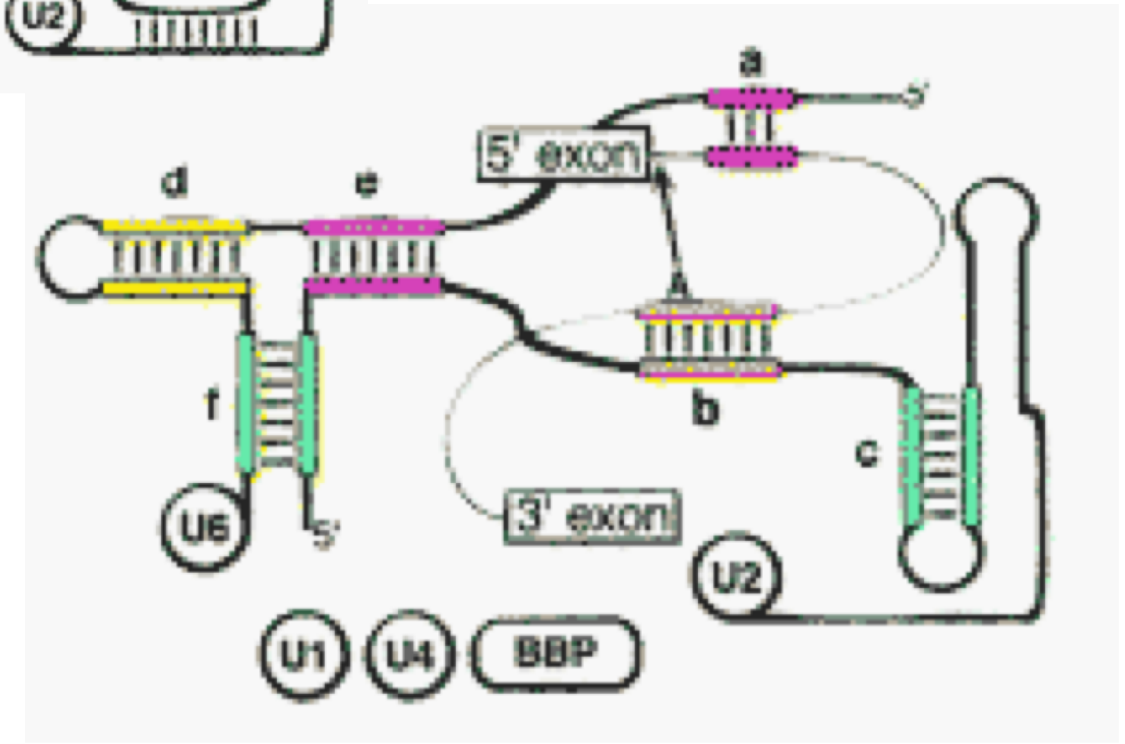
CELL Volume 92, Issue 3 , 6 February 1998, Pages 315-326

Figure 2. Spliceosome Assembly, Rearrangement, and Disassembly Requires ATP, Numerous DExD/H box Proteins, and Prp24. The snRNPs are depicted as circles. The pathway for *S. cerevisiae* is shown.

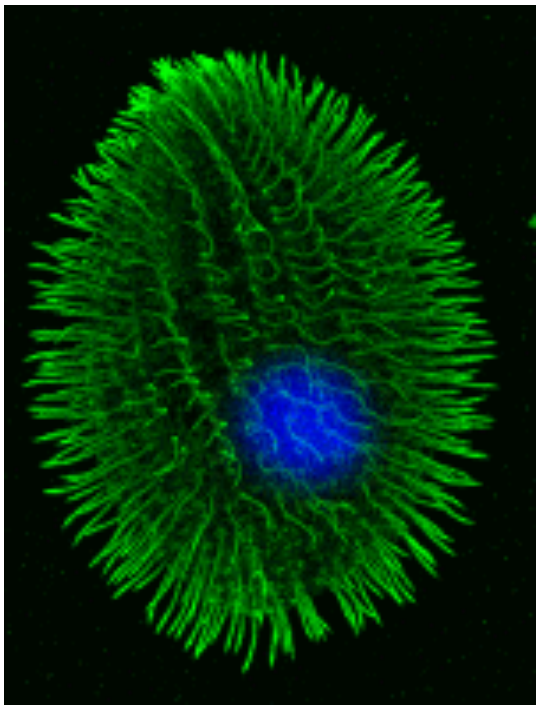




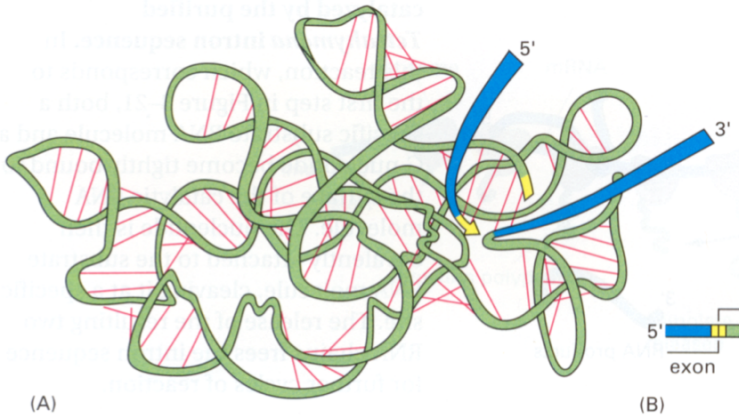
E.g.:
exchange of
U1 for U6

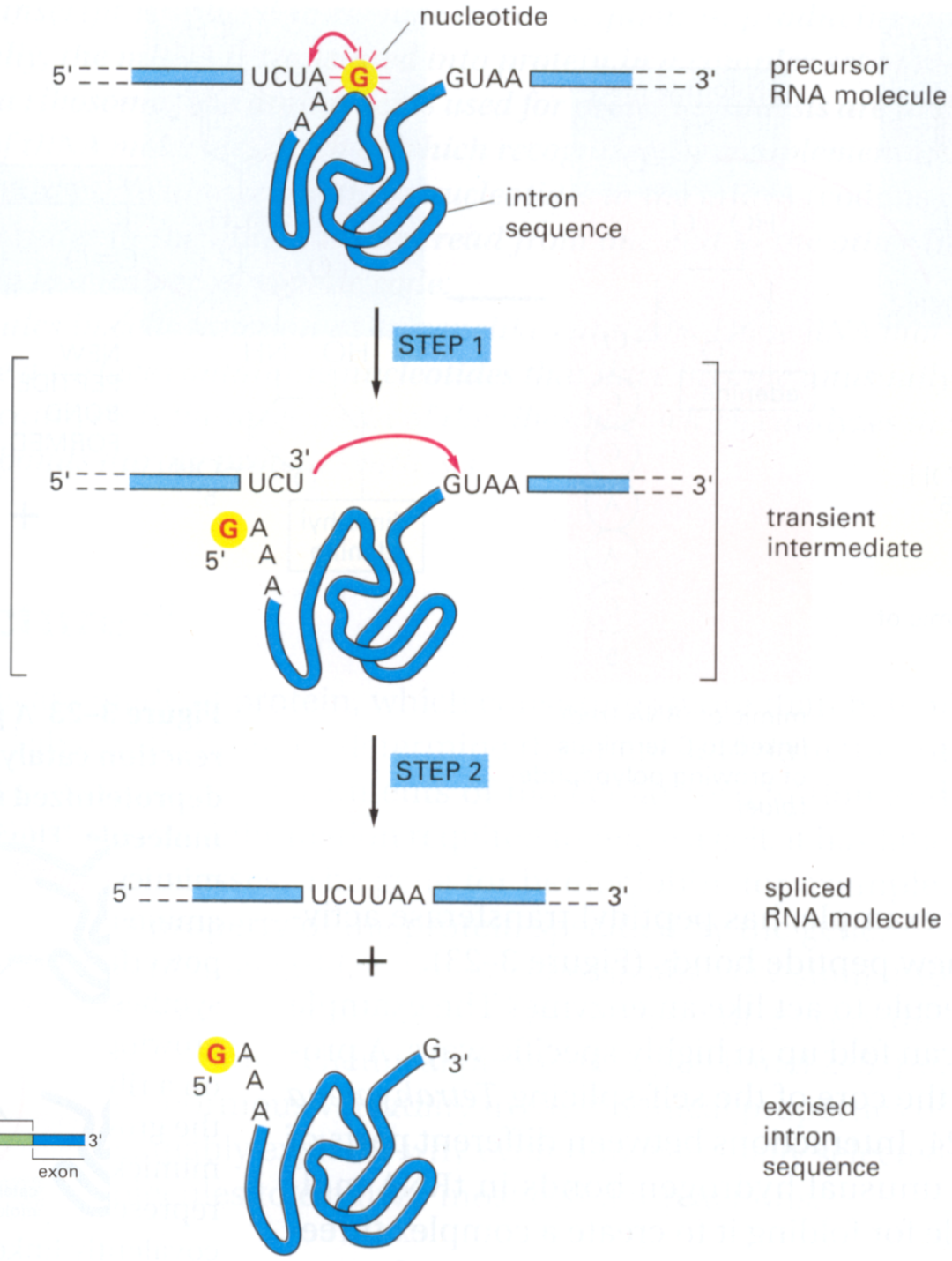


Hints to Origins?



Tetrahymena thermophila





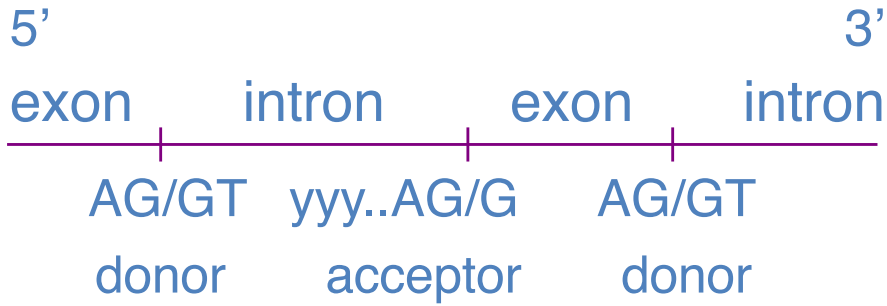
Genes in Eukaryotes

As in prokaryotes (but maybe more variable)

- promoters
- start/stop transcription
- start/stop translation

New Features:

- introns, exons, splicing
- branch point signal
- alternative splicing
- polyA site/tail



Characteristics of human genes

(Nature, 2/2001, Table 21)

	Median	Mean	Sample (size)
Internal exon	122 bp	145 bp	RefSeq alignments to draft genome sequence, with confirmed intron boundaries (43,317 exons)
Exon number	7	8.8	RefSeq alignments to finished seq (3,501 genes)
Introns	1,023 bp	3,365 bp	RefSeq alignments to finished seq (27,238 introns)
3′ UTR	400 bp	770 bp	Confirmed by mRNA or EST on chromo 22 (689)
5′ UTR	240 bp	300 bp	Confirmed by mRNA or EST on chromo 22 (463)
Coding seq	1,100 bp	1340 bp	Selected RefSeq entries (1,804)*
(CDS)	367 aa	447 aa	
Genomic span	14 kb	27 kb	Selected RefSeq entries (1,804)*

* 1,804 selected RefSeq entries were those with full-length unambiguous alignment to finished sequence

Big Genes

Many genes are over 100 kb long,

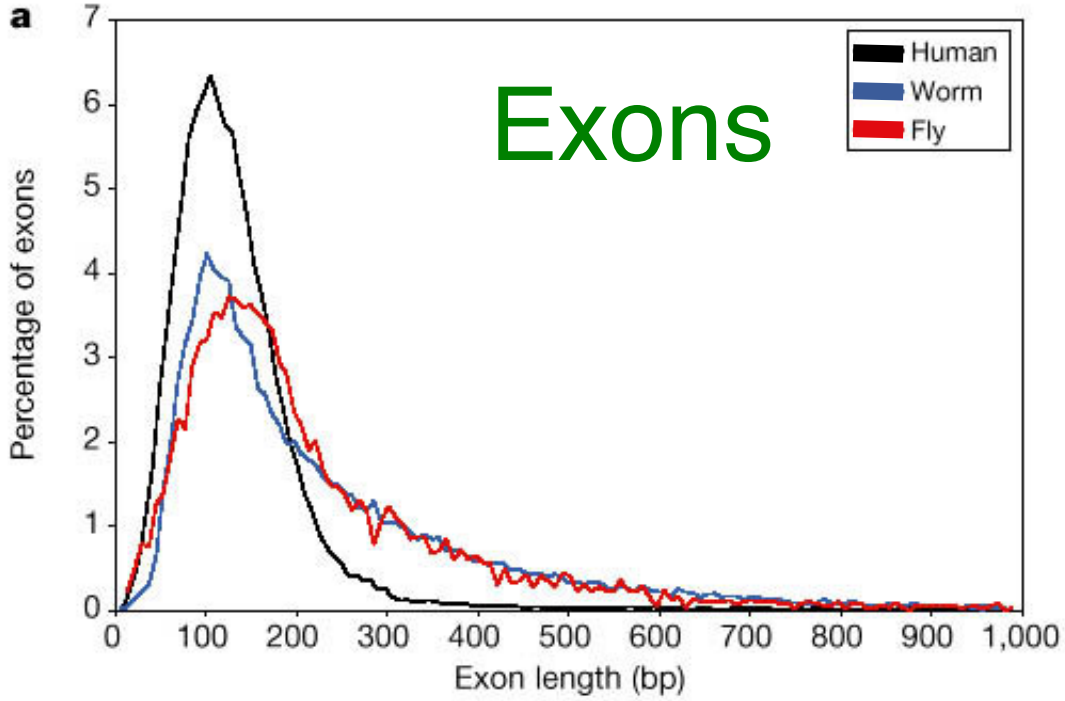
Max known: dystrophin gene (DMD), 2.4 Mb.

The variation in the size distribution of coding sequences and exons is less extreme, although there are remarkable outliers.

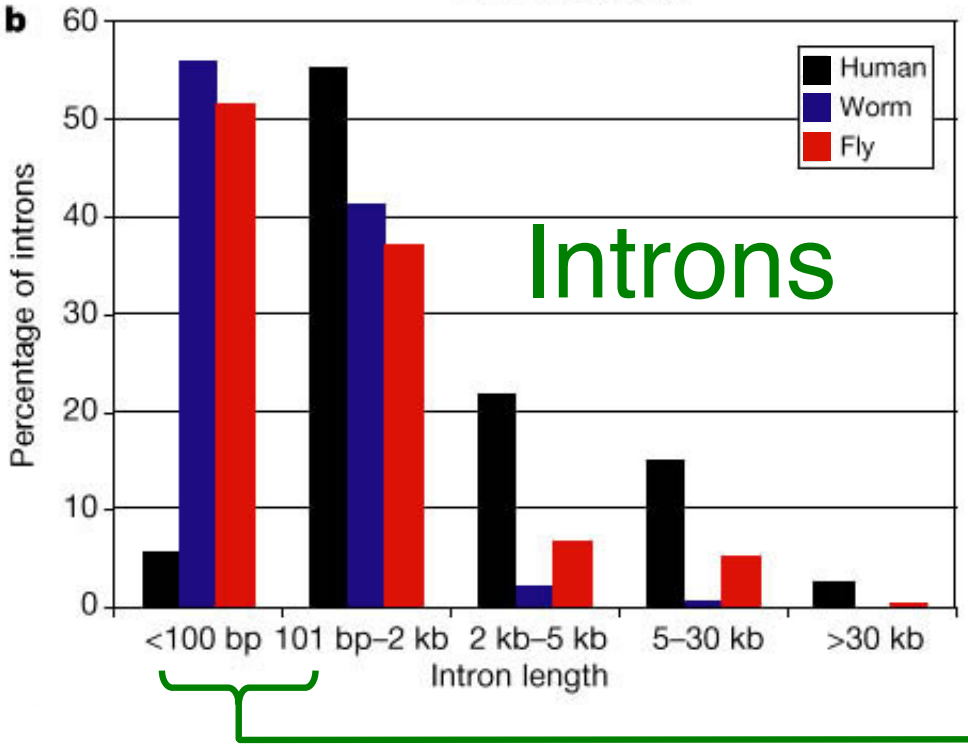
- The titin gene has the longest currently known coding sequence at 80,780 bp; it also has the largest number of exons (178) and longest single exon (17,106 bp).

RNApol rate: 1.2-2.5 kb/min ≥16 hours to transcribe DMD

Nature 2/2001



Introns



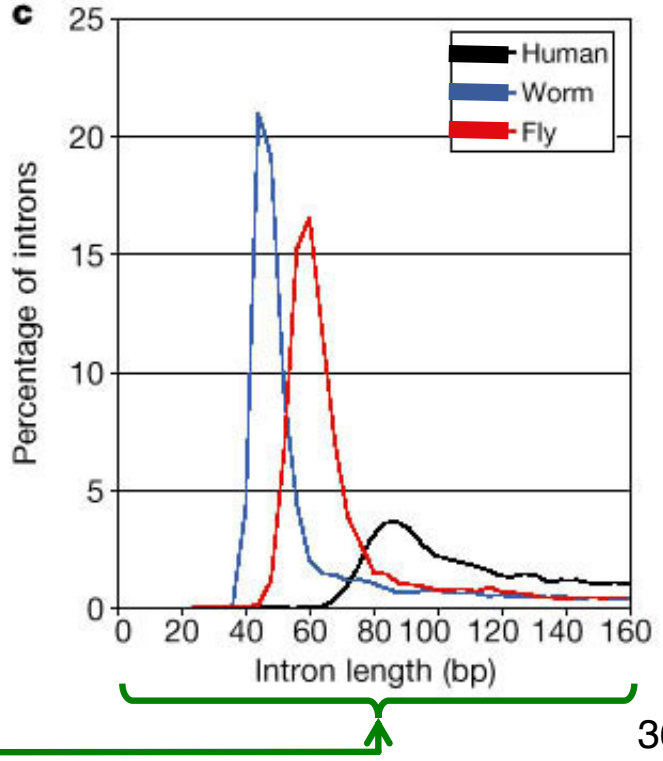
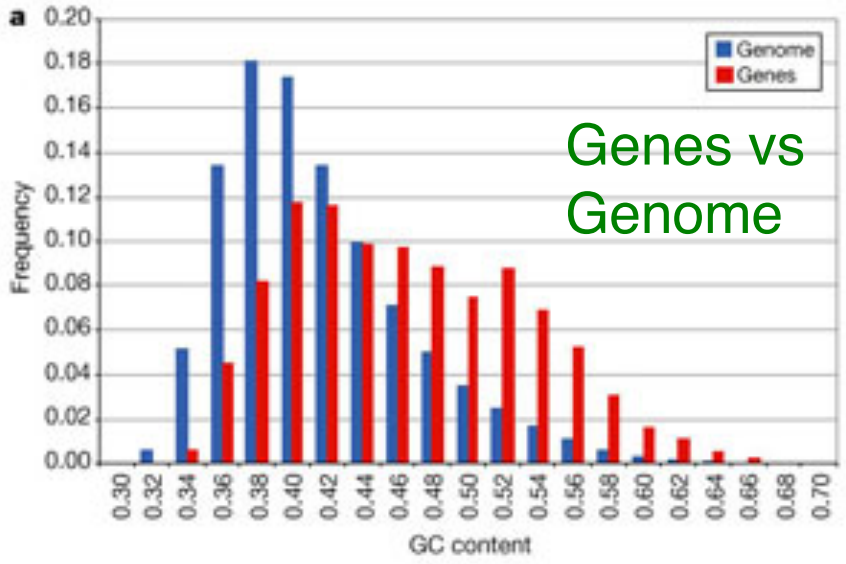
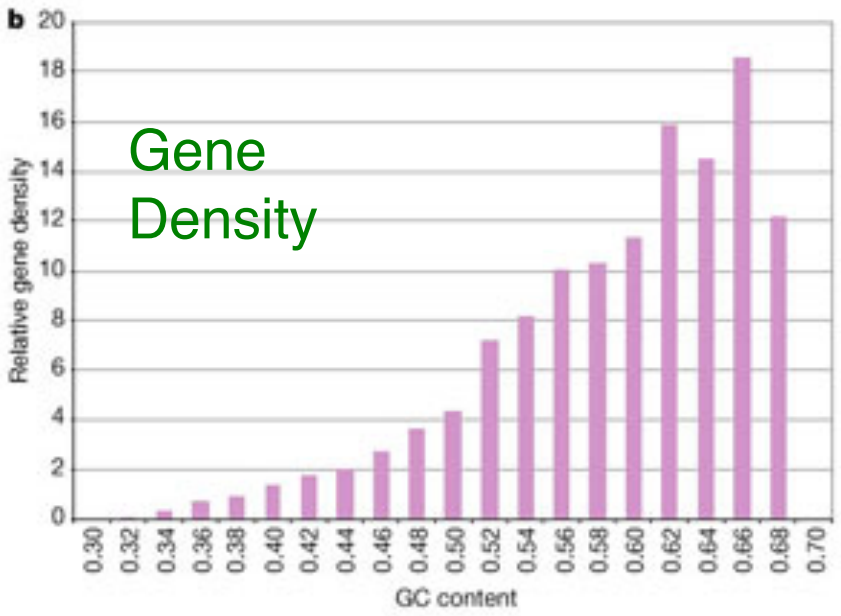


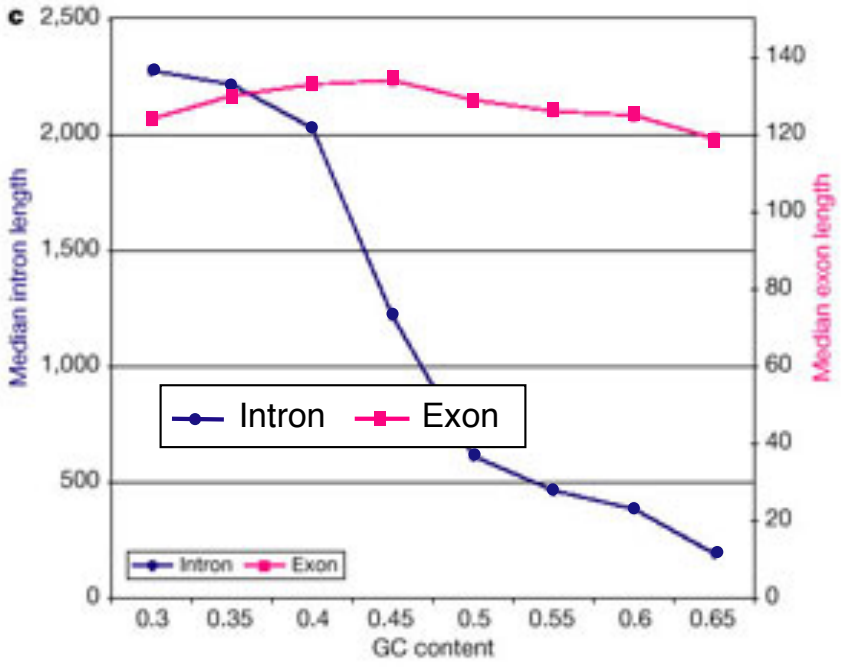
Figure 36 GC content

Nature 2/2001





a: Distribution of GC content in genes and in the genome. For 9,315 known genes mapped to the draft genome sequence, the local GC content was calculated in a window covering either the whole alignment or 20,000 bp centered on midpoint of the alignment, whichever was larger. Ns in the sequence were not counted. GC content for the genome was calculated for adjacent nonoverlapping 20,000-bp windows across the sequence. Both distributions normalized to sum to one.



b: Gene density as a function of GC content (= ratios of data in a. Less accurate at high GC because the denominator is small)

c: Dependence of mean exon and intron lengths on GC content. The local GC content, based on alignments to finished sequence only, calculated from windows covering the larger of feature size or 10,000 bp centered on it

Other Relevant Features

PolyA Tails

- 100-300 A's typically added to the 3′ end of the mRNA after transcription–*not* templated by DNA

Processed pseudogenes

- Sometimes mRNA (*after* splicing + polyA) is reverse-transcribed into DNA and re-integrated into genome
- ~14,000 in human genome

Alternative Splicing

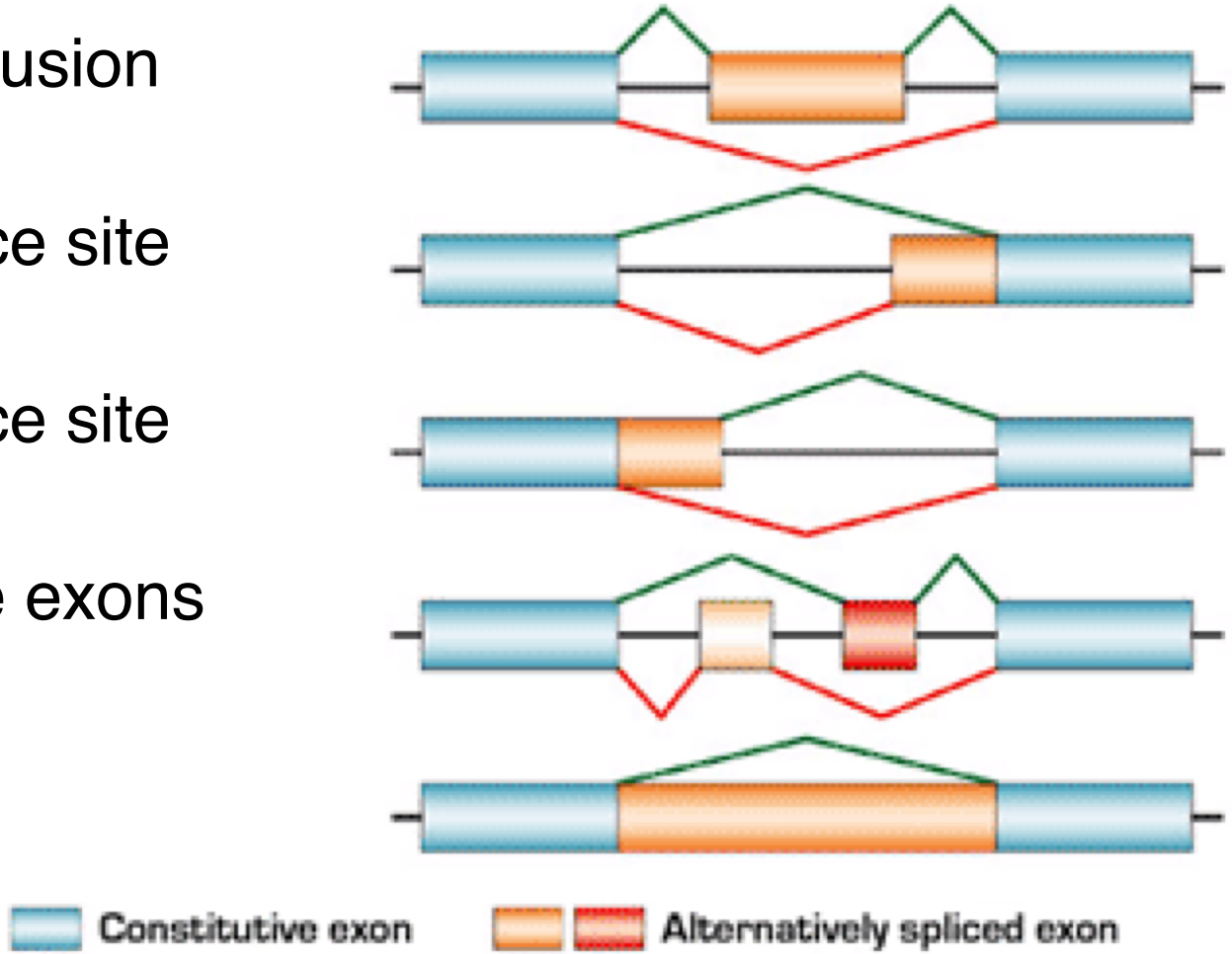
Exon skipping/inclusion

Alternative 3′ splice site

Alternative 5′ splice site

Mutually exclusive exons

Intron retention



These are *regulated,* not just errors

Other Features (cont)

Alternative start sites (5′ ends)

Alternative PolyA sites (near 3′ ends)

Alternative splicing

Collectively, these affect an estimated 95% of genes,
with ~5 (a wild guess) isoforms per gene
(but can be huge; fly Dscam: 38,016, potentially)

Trans-splicing and gene fusions

(rare in humans but important in some tumors)

Computational Gene Finding?

How do we algorithmically account for all this complexity…

A Case Study – Genscan

C Burge, S Karlin (1997), “Prediction of complete gene structures in human genomic DNA”, Journal of Molecular Biology, 268: 78-94.

Training Data

238 multi-exon genes

142 single-exon genes

total of 1492 exons

total of 1254 introns

total of 2.5 Mb

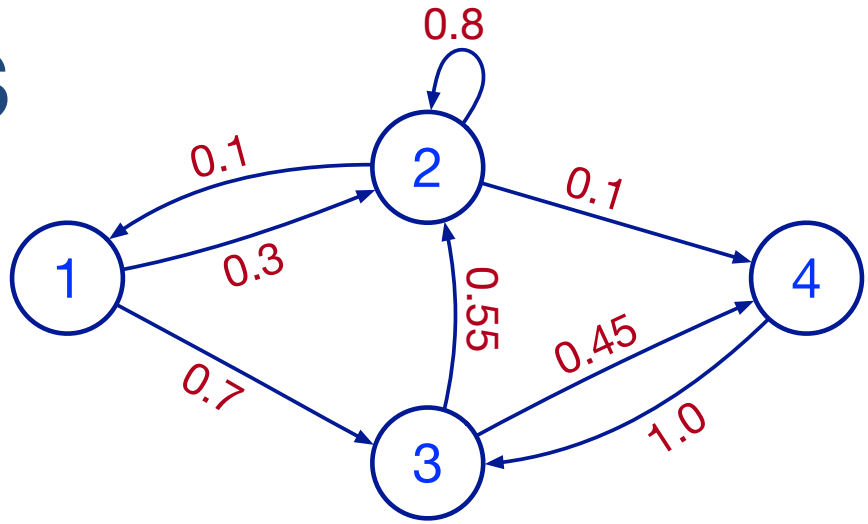
NO alternate splicing, none > 30kb, ...

Performance Comparison

	Accuracy						
	per nuc.		per exon				
Program	**Sn**	**Sp**	**Sn**	**Sp**	**Avg.**	**ME**	**WE**
GENSCAN	0.93	0.93	0.78	0.81	0.80	0.09	0.05
FGENEH	0.77	0.88	0.61	0.64	0.64	0.15	0.12
GeneID	0.63	0.81	0.44	0.46	0.45	0.28	0.24
Genie	0.76	0.77	0.55	0.48	0.51	0.17	0.33
GenLang	0.72	0.79	0.51	0.52	0.52	0.21	0.22
GeneParser2	0.66	0.79	0.35	0.40	0.37	0.34	0.17
GRAIL2	0.72	0.87	0.36	0.43	0.40	0.25	0.11
SORFIND	0.71	0.85	0.42	0.47	0.45	0.24	0.14
Xpound	0.61	0.87	0.15	0.18	0.17	0.33	0.13
GeneID‡	0.91	0.91	0.73	0.70	0.71	0.07	0.13
GeneParser3	0.86	0.91	0.56	0.58	0.57	0.14	0.09

After Burge&Karlin, Table 1. Sensitivity, Sn = TP/AP; Specificity, Sp = TP/PP

Generalized Hidden Markov Models



States: *1, 2, …*

π: Initial state distribution

a_{ij}: Transition probabilities

One submodel per state

Outputs are *strings* generated by submodel

Given length *L*

- Pick start state q_1 *(~π)*
- While $\sum d_i < L$
  - Pick d_i & string s_i of length d_i ~ submodel for q_i
  - Pick next state q_{i+1} *(~a_{ij})*
- Output $s_1 s_2 \ldots$

Decoding

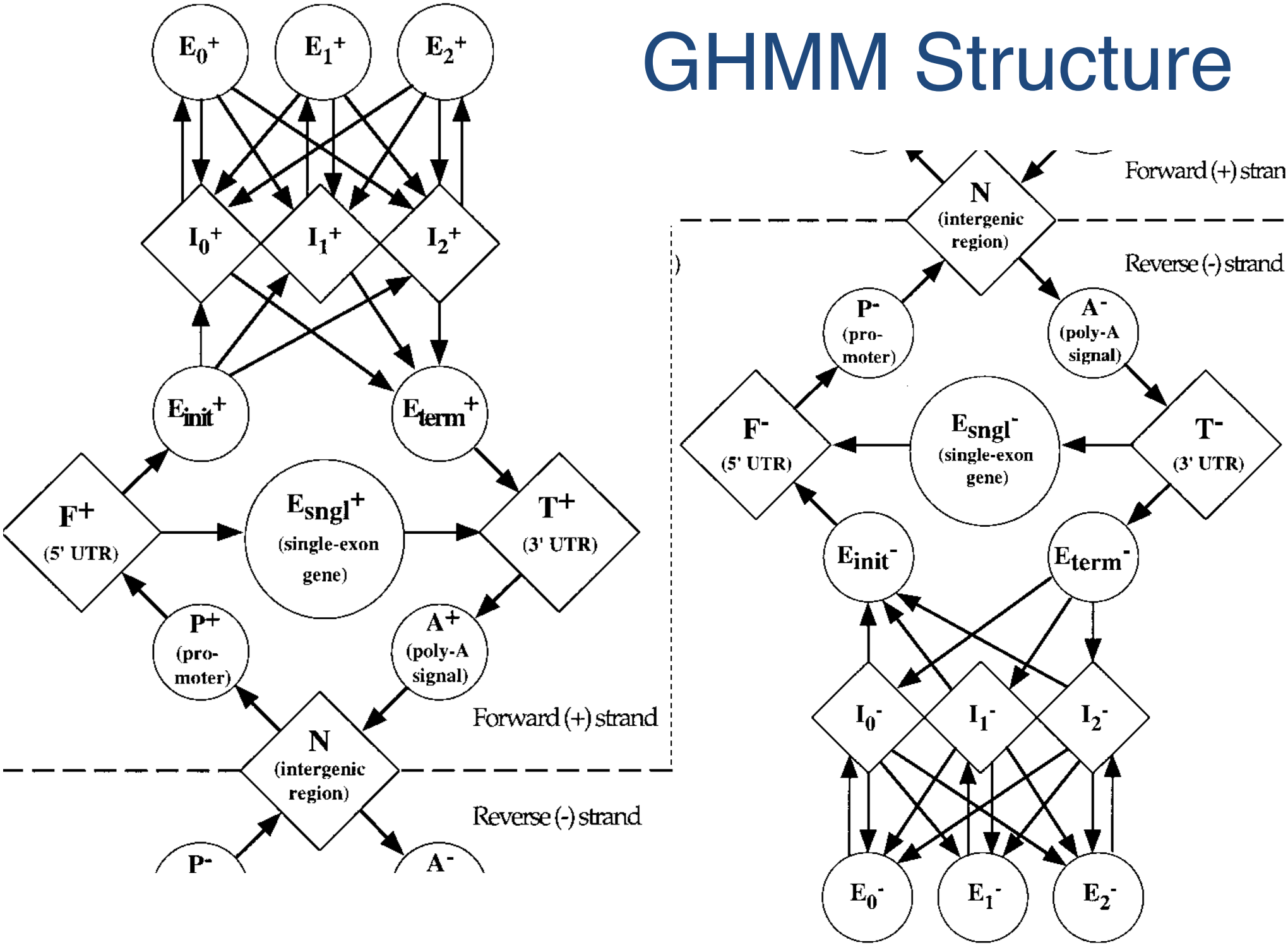
A "parse" ϕ of seq $s = s_1 s_2 \ldots s_L$ is a pair $d = d_1 d_2 \ldots d_k$, $q = q_1 q_2 \ldots q_k$ with $\sum d_i \geq L$

A Viterbi-like alg calculates prob of most probable path emitting s; traceback gives ϕ

Similarly, forward/backward-like algs can find, e.g.
Pr(emit $s_1 s_2 \ldots s_i$ & end in state j)

(summing over possible predecessor states, possible d_k, *etc.)*

GHMM Structure



Length Distributions

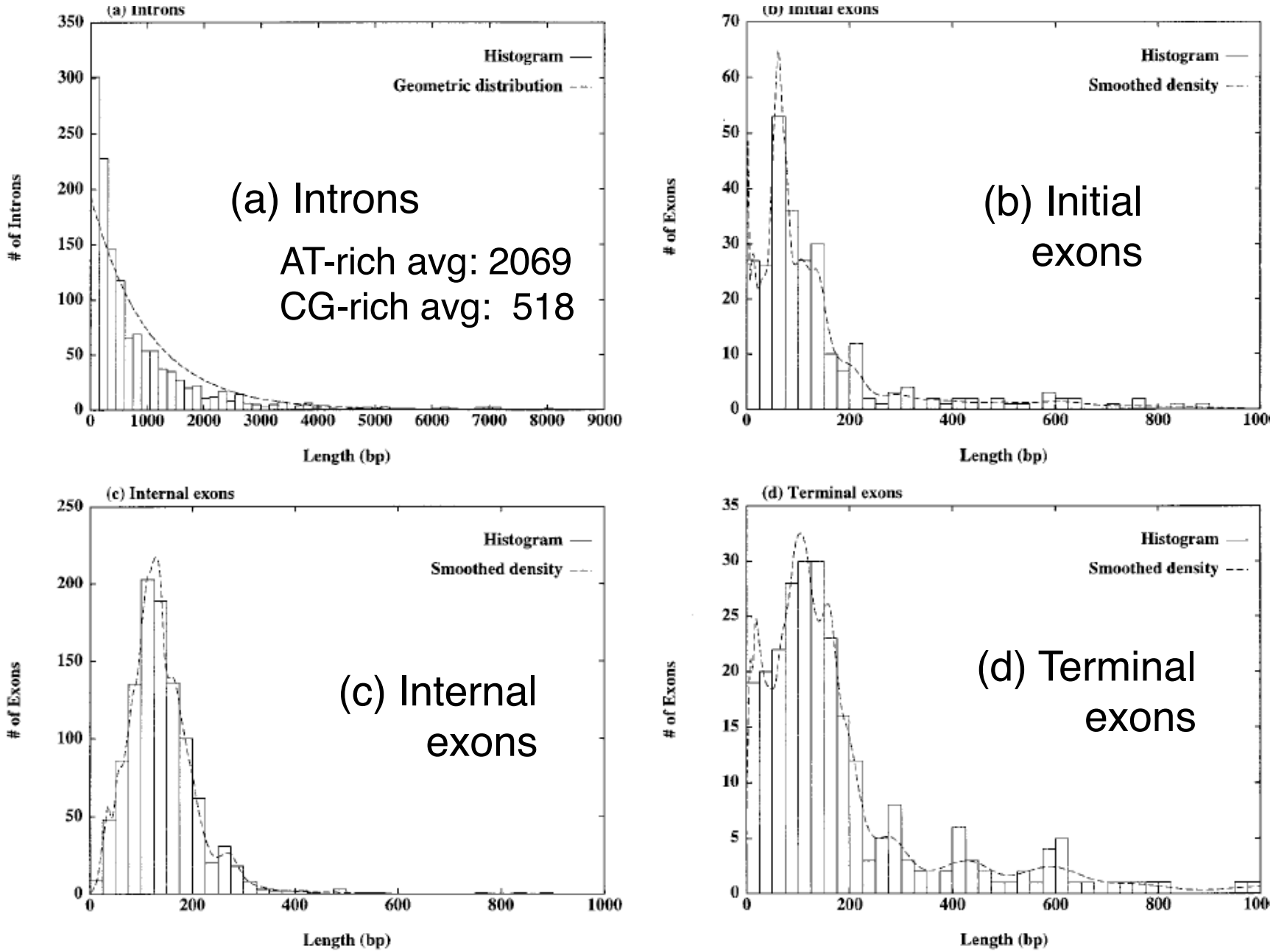


Figure 4. Length distributions are shown for (a) 1254 introns; (b) 238 initial exons; (c) 1151 internal exons; and (d) 238 terminal exons from the 238 multi-exon genes of the learning set $\mathscr{L}$. Histograms (continuous lines) were derived with a bin size of 300 bp in (a), and 25 bp in (b), (c), (d). The broken line in (a) shows a geometric (exponential) distribution with parameters derived from the mean of the intron lengths; broken lines in (b), (c) and (d) are the smoothed empirical distributions of exon lengths used by GENSCAN (details given by Burge, 1997). Note different horizontal and vertical scales are used in (a), (b), (c), (d) and that multimodality in (b) and (d) may, in part, reflect relatively

Effect of G+C Content

Group	**I**	**II**	**III**	**IV**
C + G% range	<43	43-51	51-57	>57
Number of genes	65	115	99	101
Est. proportion single-exon genes	0.16	0.19	0.23	0.16
Codelen: single-exon genes (bp)	1130	1251	1304	1137
Codelen: multi-exon genes (bp)	902	908	1118	1165
Introns per multi-exon gene	5.1	4.9	5.5	5.6
Mean intron length (bp)	2069	1086	801	518
Est. mean transcript length (bp)	10866	6504	5781	4833
Isochore	L1+L2	H1+H2	H3	H3
DNA amount in genome (Mb)	2074	1054	102	68
Estimated gene number	22100	24700	9100	9100
Est. mean intergenic length	83000	36000	5400	2600
Initial probabilities:				
Intergenic (N)	0.892	0.867	0.54	0.418
Intron (I+, I-)	0.095	0.103	0.338	0.388
5’ Untranslated region (F+, F-)	0.008	0.018	0.077	0.122
3’ Untranslated region (T+, T-)	0.005	0.011	0.045	0.072

Submodels

5′ UTR

L ~ geometric(769 bp), s ~ MM(5)

3′ UTR

L ~ geometric(457 bp), s ~ MM(5)

Intergenic

L ~ geometric(GC-dependent), s ~ MM(5)

Introns

L ~ geometric(GC-dependent), s ~ MM(5)

Submodel: Exons

Inhomogenious 3-periodic 5th order Markov models

Separate models for low GC (<43%), high GC

Track “phase” of exons, i.e. reading frame.

Signal Models I: WMM's

Polyadenylation

- 6 bp, consensus AATAAA

Translation Start

- 12 bp, starting 6 bp before start codon

Translation stop

- A stop codon, then 3 bp WMM

Signal Models II: more WMM's

- Promoter
  - 70% TATA
    - 15 bp TATA WMM
    - s ~ null, L ~ Unif(14-20)
    - 8 bp cap signal WMM
  - 30% TATA-less
    - 40 bp null

Signal Models III: W/WAM's

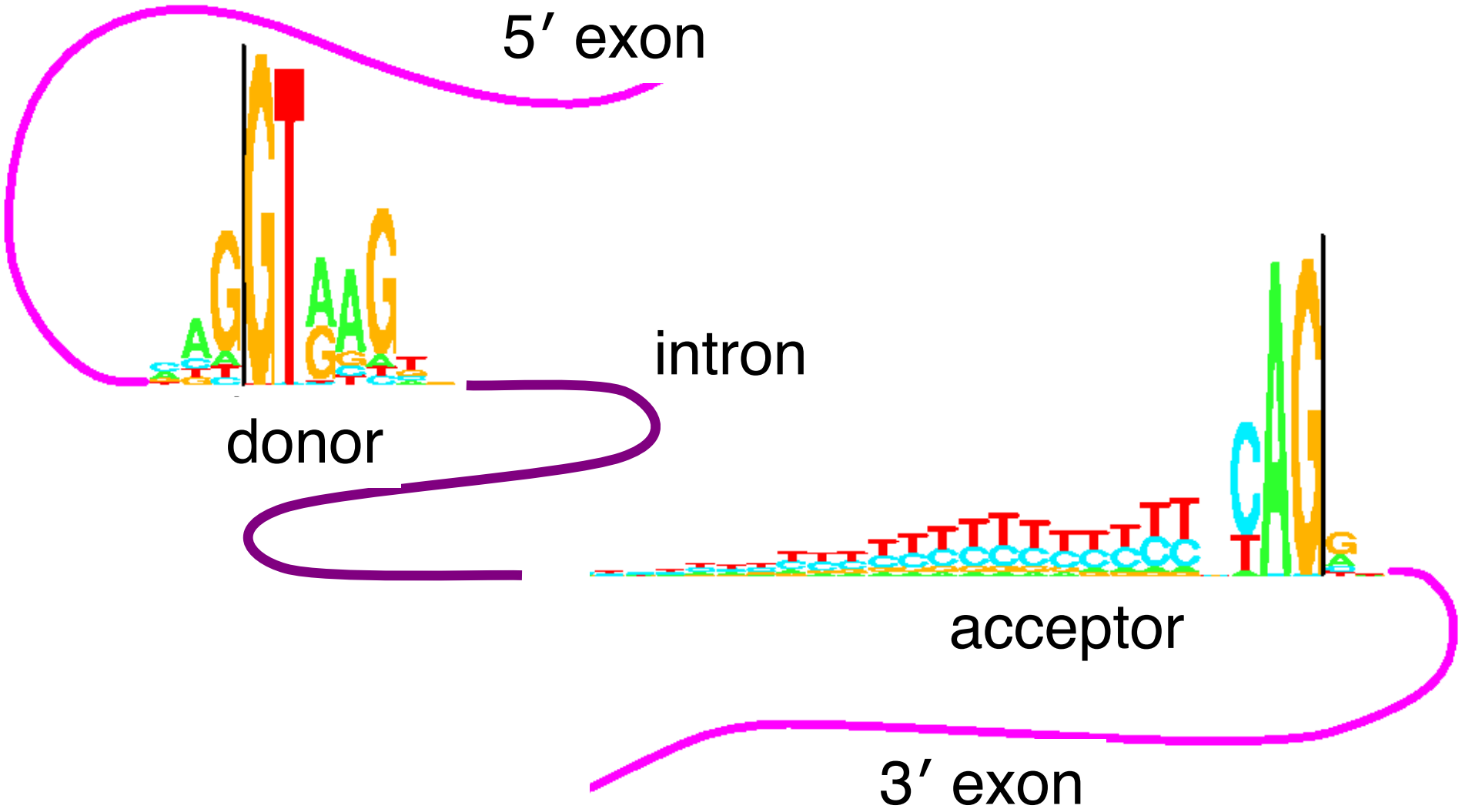
Acceptor Splice Site (3' end of intron)

- [-20..+3] relative to splice site modeled by "1st order weight array model"

Branch point & polypyrimidine tract

- Hard. Even weak consensus like YYRAY found in [-40..-21] in only 30% of training
- "Windowed WAM": 2nd order WAM, but averaged over 5 preceding positions
  "captures weak but detectable tendency toward YYY triplets and certain branch point related triplets like TGA, TAA, …"

What do splice sites look like?



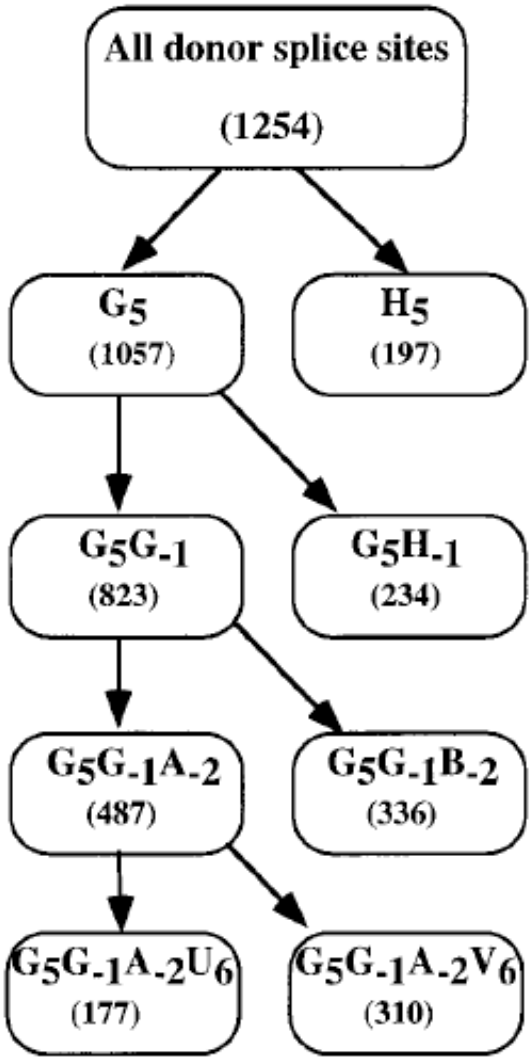
Signal Models IV: Maximum Dependence Decomposition

Donor splice sites (5′ end of intron) show dependencies between non-adjacent positions, e.g. poor match at one end compensated by strong match at other end, 6 bp away

Model is basically a decision tree

Uses χ^2 test to quantitate dependence

Pos	A%	C%	G%	U%
-3	33	36	19	13
-2	56	15	15	15
-1	9	4	78	9
+3	44	3	51	3
+4	75	4	13	9
+6	14	18	19	49
-3	34	37	18	11
-2	59	10	15	16
+3	40	4	53	3
+4	70	4	16	10
+6	17	21	21	42
-3	37	42	18	3
+3	39	5	51	5
+4	62	5	22	11
+6	19	20	25	36
-3	32	40	23	5
+3	27	4	59	10
+4	51	5	25	19



Pos	A%	C%	G%	U%
-3	35	44	16	6
-2	85	4	7	5
-1	2	1	97	0
+3	81	3	15	2
+4	51	28	9	12
+6	22	20	30	28
-3	29	31	21	18
-2	43	30	17	11
+3	56	0	43	0
+4	93	2	3	3
+6	5	10	10	76
-3	29	30	18	23
+3	42	1	56	1
+4	80	4	8	8
+6	14	21	16	49
-3	39	43	15	2
+3	46	6	46	3
+4	69	5	20	7

All sites:	Position								
Base	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	**33**	**60**	8	0	0	**49**	**71**	6	15
C%	**37**	13	4	0	0	3	7	5	19
G%	18	14	**81**	**100**	0	**45**	12	**84**	20
U%	12	13	7	0	**100**	3	9	5	**46**

U1 snRNA: 3' G U C C A U U C A 5'

Many dependencies, such as 5′/3′ compensation, e.g. G_{-1} vs G_5/H_5

χ^2 test : Are events A & B independent ?

	B	not B	
A	8	4	12
not A	2	6	8
	10	10	20

← Event counts plus marginals

$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

"Expected" means expected assuming independence, e.g. expect B 10/20; A 12/20; both 120/400*20 = 6, etc.

Significance: table look up (or approximate as normal)

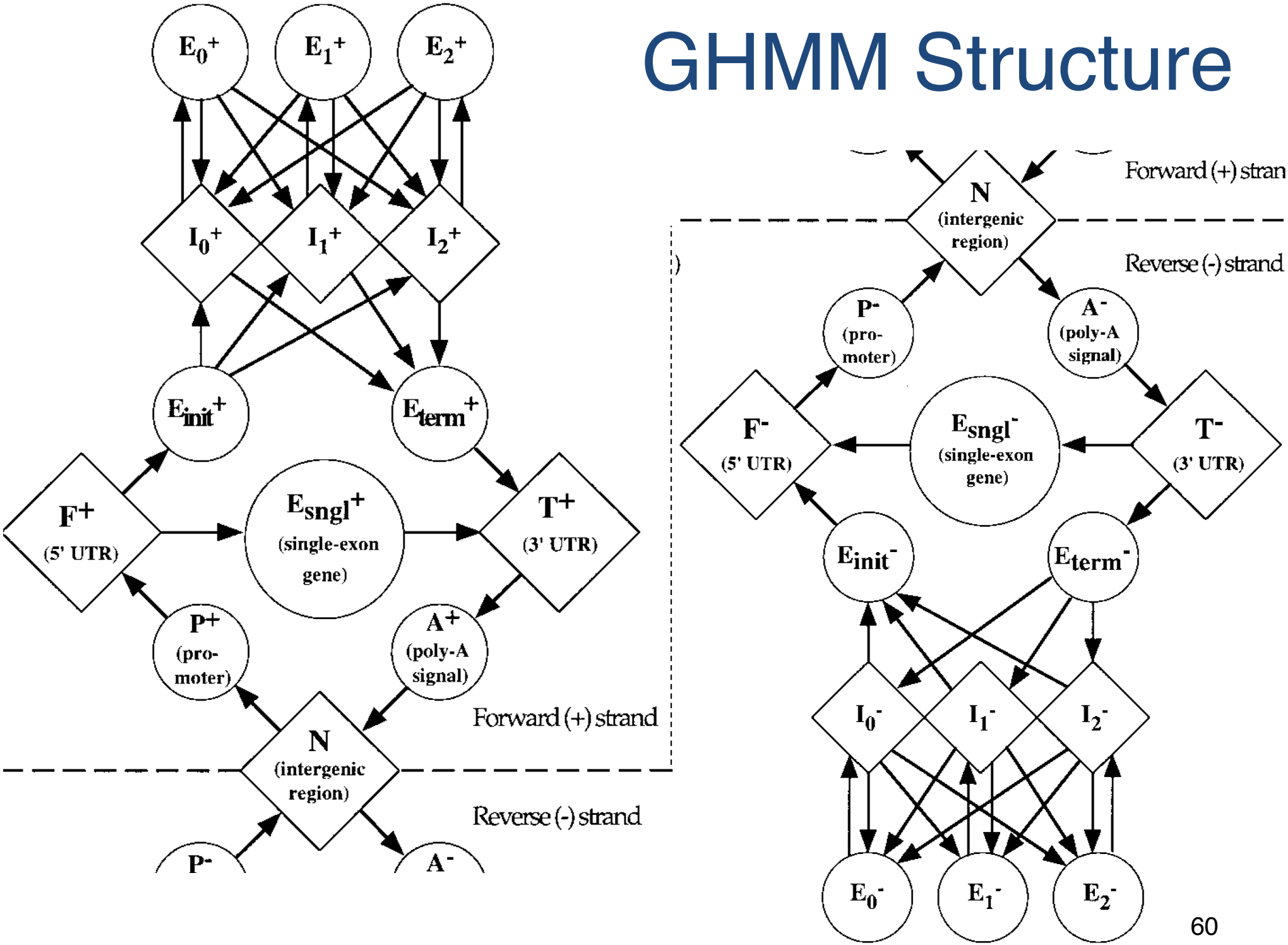
χ^2 test for independence of nucleotides in donor sites

i	Con	j: -3	-2	-1	+3	+4	+5	+6	Sum
-3	c/a	---	61.8*	14.9	5.8	20.2*	11.2	18.0*	131.8*
-2	A	115.6*	---	40.5*	20.3*	57.5*	59.7*	42.9*	336.5*
-1	G	15.4	82.8*	---	13.0	61.5*	41.4*	96.6*	310.8*
+3	a/g	8.6	17.5*	13.1	---	19.3*	1.8	0.1	60.5*
+4	A	21.8*	56.0*	62.1*	64.1*	---	56.8*	0.2	260.9*
+5	G	11.6	60.1*	41.9*	93.6*	146.6*	---	33.6*	387.3*
+6	t	22.2*	40.7*	103.8*	26.5*	17.8*	32.6*	---	243.6*

*** means chi-squared p-value < .001**

Technically – build a 2 x 4 table for each (i,j) pair:
Pos i does/does not match consensus vs pos j is A, C, G, T
calculate χ^2 as on previous slide, e.g. χ^2 for +6 vs -1 = 103.8
If independent, you'd expect $\chi^2 \leq 16.3$ all but one in a 1000 times.

GHMM Structure



Summary of Burge & Karlin

- Coding DNA & control signals are nonrandom
  - Weight matrices, WAMs, etc. for controls
  - Codon frequency, etc. for coding
- GHMM nice for overall architecture
- Careful attention to small details pays

Problems with BK training set

1 gene per sequence

Annotation errors

Single exon genes over-represented?

Highly expressed genes over-represented?

Moderate sized genes over-represented? (none > 30 kb) …

Similar problems with other training sets, too

(Of course we can now do better for human, mouse, etc., but what about cockatoos or cows or endangered frogs, or …)

Problems with all methods

Pseudo genes (~ 14,000 in human)

Short ORFs

Sequencing errors

Non-coding RNA genes & spliced UTR's

Overlapping genes

Alternative TSS/polyadenylation/splicing

Hard to find novel stuff – not in training

Species-specific weirdness – spliced leaders, polycistronic transcripts, RNA editing…

Other important ideas

Database search - does gene you're predicting look anything like a known protein? If that protein is an important player in some pathway, are related genes also present?

Comparative genomics - what does this region look like in related organisms?