CSEP 590A
Computational Biology
Summer 2006

Lecture 7

Gene Prediction

Some References
(more on schedule page)

An extensive online bib

http://www.nslij-genetics.org/gene/

A good intro survey

JM Claverie (1997) "Computational methods for the identification of genes in vertebrate genomic sequences” Human Molecular Genetics, 6(10)(review issue): 1735-1744.

A gene finding bake-off

M Burset, R Guigo (1996), "Evaluation of gene structure prediction programs", Genomics, 34(3): 353-367.

Motivation

Sequence data flooding into Genbank

What does it mean?

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

Protein Coding Nuclear DNA

Focus of this lecture

Goal: Automated annotation of new sequence data

State of the Art:

- predictions ~ 60% similar to real proteins
- ~80% if database similarity used
- lab verification still needed, still expensive

Biological Basics

Central Dogma:

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

Codons: 3 bases code one amino acid

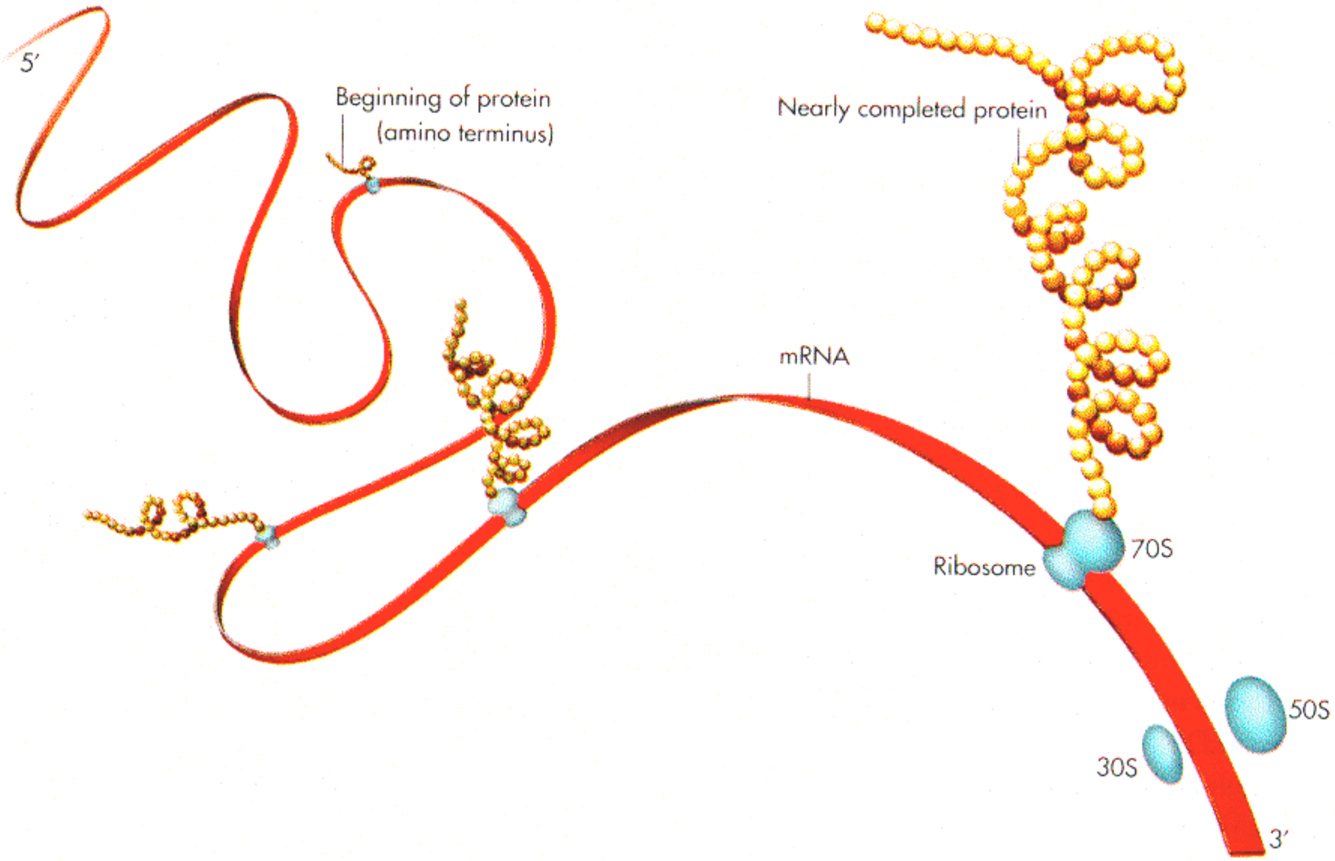
- Start codon
- Stop codons
- 3’, 5’ Untranslated Regions (UTR’s)

Codons & The Genetic Code

		Second Base					
		U	C	A	G		
First Base	U	Phe	Ser	Tyr	Cys	U	Third Base
		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	

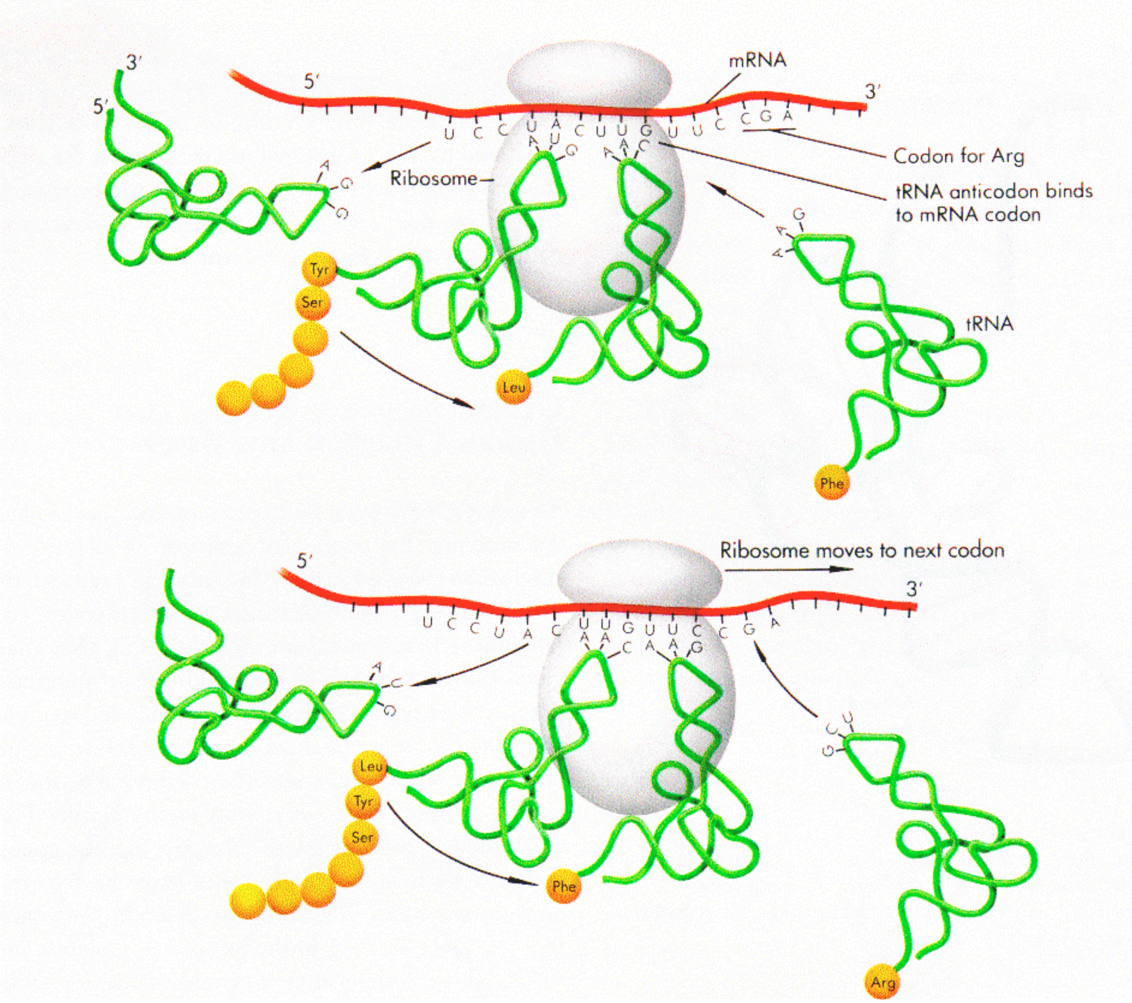
Ala	: Alanine
Arg	: Arginine
Asn	: Asparagine
Asp	: Aspartic acid
Cys	: Cysteine
Gln	: Glutamine
Glu	: Glutamic acid
Gly	: Glycine
His	: Histidine
Ile	: Isoleucine
Leu	: Leucine
Lys	: Lysine
Met	: Methionine
Phe	: Phenylalanine
Pro	: Proline
Ser	: Serine
Thr	: Threonine
Trp	: Tryptophane
Tyr	: Tyrosine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

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

Ribosomes


3′
5′
5′
mRNA
3′
U C C U A C U U G U U C C G A
A U G
A A C
A G G
Ribosome
Codon for Arg
tRNA anticodon binds to mRNA codon
A A G
tRNA
Tyr
Ser
Leu
Phe
Ribosome moves to next codon
5′
3′
U C C U A C U U G U U C C G A
A A C
A A G
A C G
G C U
Leu
Tyr
Ser
Phe
Arg

Watson, Gilman, Witkowski, & Zoller, 1992

Idea #1: Find Long ORF’s

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

Open Reading Frame: No stop codons

In random DNA

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

But average protein ~ 1000bp

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. histone, enhancer, splice interactions

Recognizing Codon Bias

Assume

- Codon usage i.i.d.; abc with freq. f(abc)
- $a_1a_2a_3a_4\ldots a_{3n+2}$ is coding, unknown frame

Calculate

- $p_1 = f(a_1a_2a_3)f(a_4a_5a_6)\ldots f(a_{3n-2}a_{3n-1}a_{3n})$
- $p_2 = f(a_2a_3a_4)f(a_5a_6a_7)\ldots f(a_{3n-1}a_{3n}\,a_{3n+1})$
- $p_3 = f(a_3a_4a_5)f(a_6a_7a_8)\ldots f(a_{3n}\,a_{3n+1}a_{3n+2})$
- $P_i = p_i\,/\,(p_1+p_1+p_3)$

More generally: k-th order Markov model

- k=5 or 6 is typical

Codon Usage in Φx174

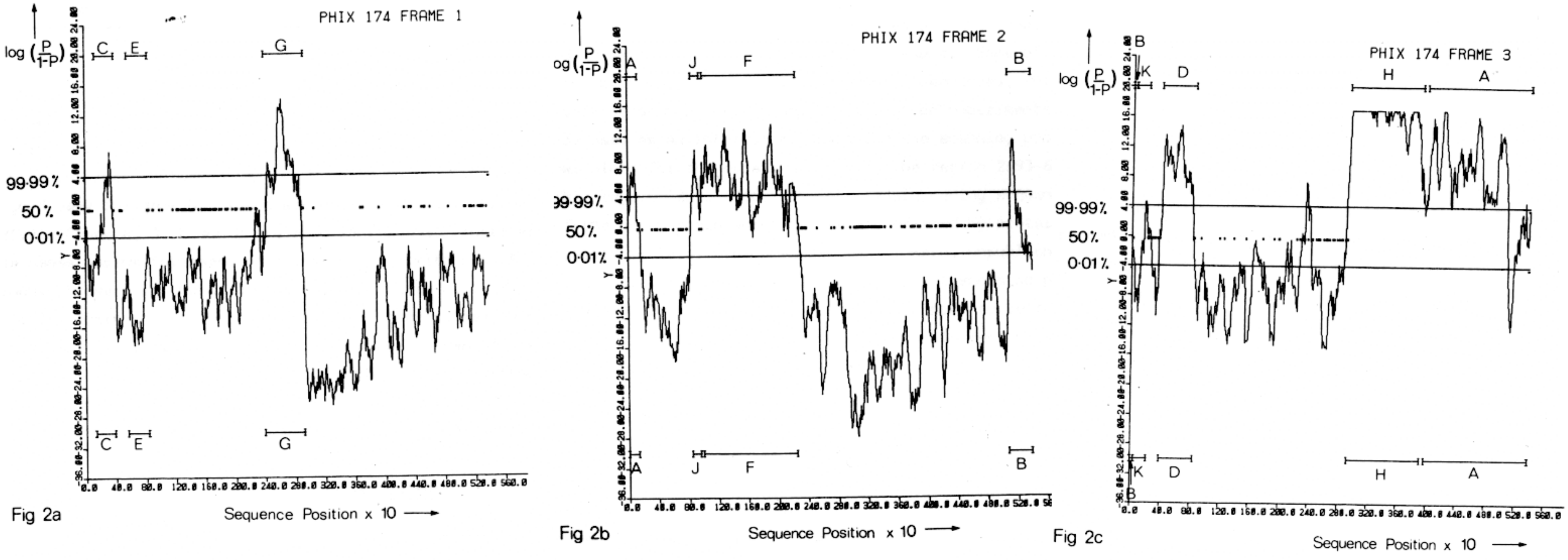


Fig 2a

Fig 2b

Fig 2c

Staden & McLachlan, NAR 10, 1 1982, 141-156

Promoters, etc.

- In prokaryotes, most DNA coding
    - E.g. ~ 70% in *H. influenzae*
- Long ORFs + codon stats do well
- But obviously won’t be perfect
    - short genes
    - 5’ & 3’ UTR’s
- Can improve by modeling promoters & other signals
    - e.g. via WMM or higher-order Markov models

Eukaryotes

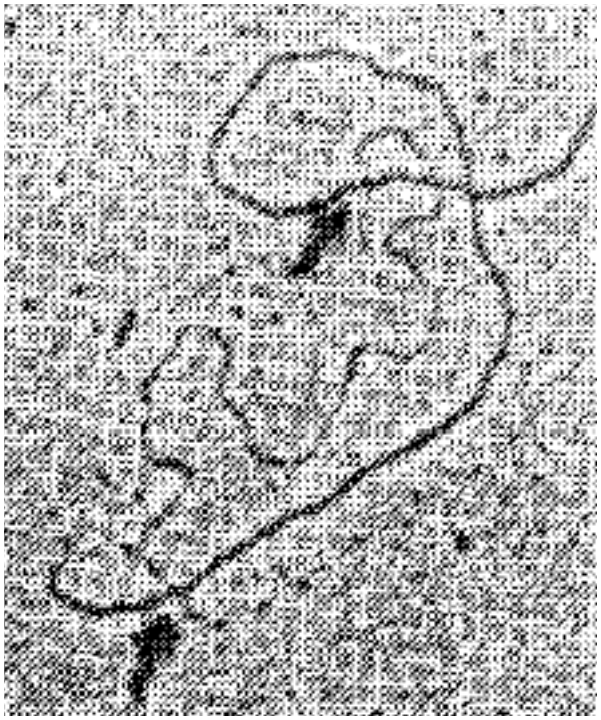
As in prokaryotes (but maybe more variable)

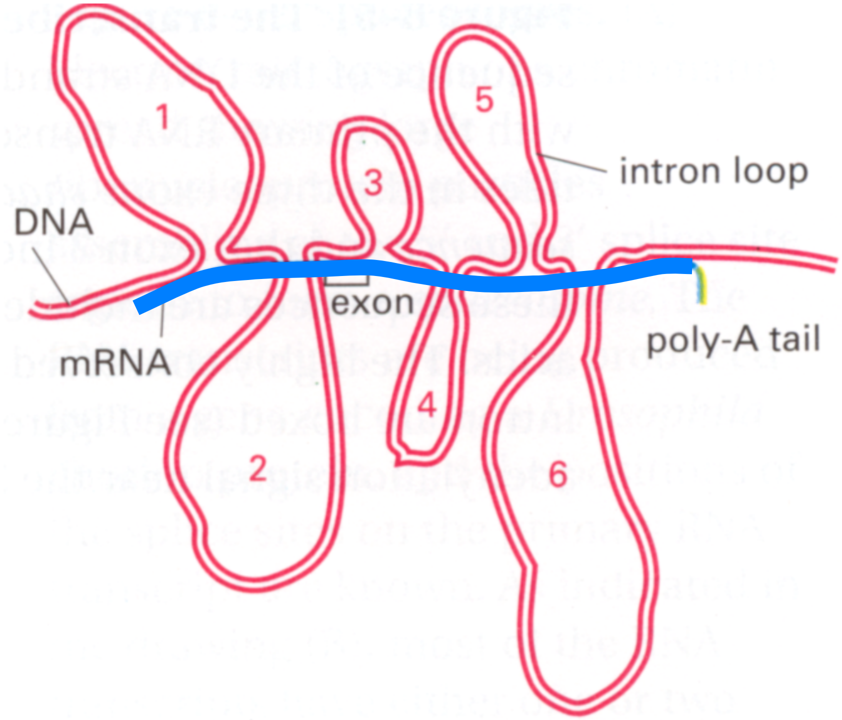
promoters

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

doi:10.1016/S0092-8674(00)80925-3

Figure 2. Spliceosome Assembly, Rearrangement, and Disassembly Requires ATP, Numerous DExD/H box Proteins, and Prp24The snRNPs are depicted as circles. The pathway for *S. cerevisiae* is shown.

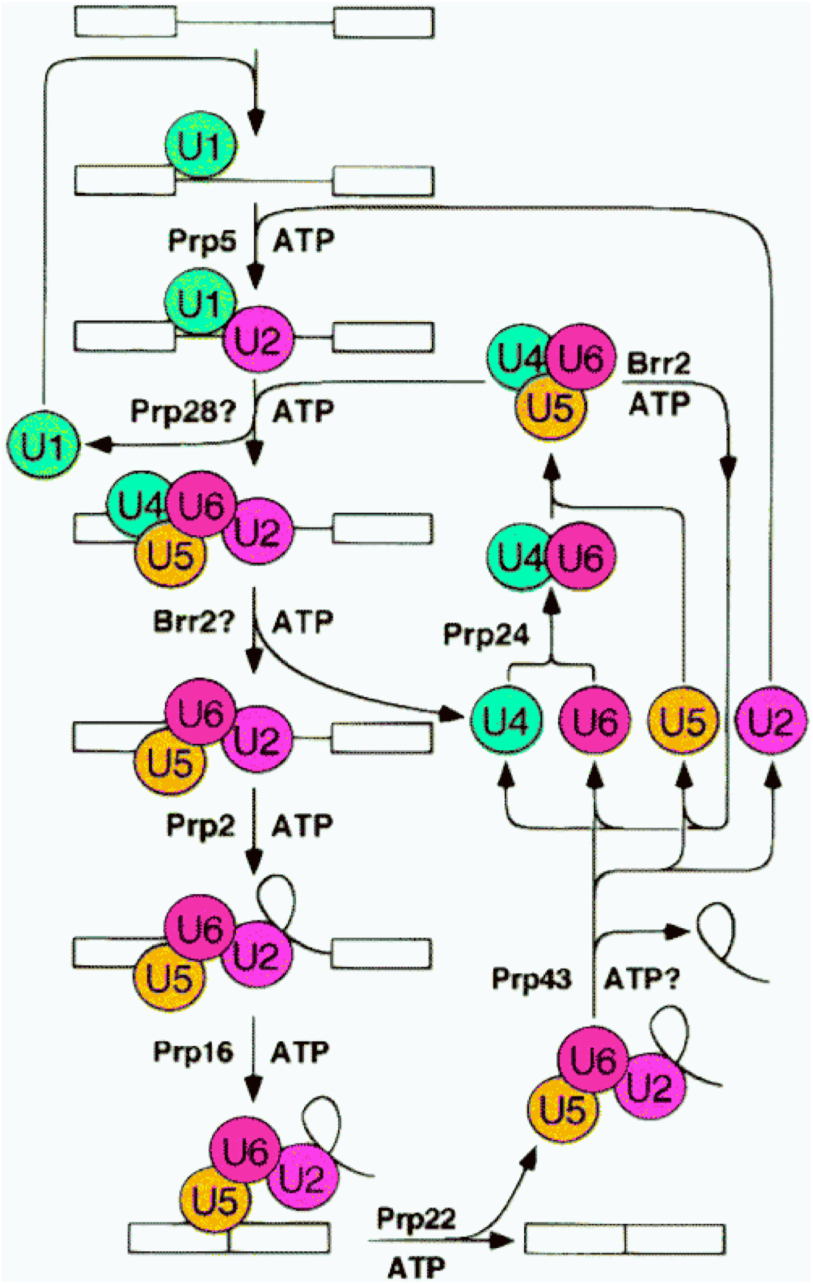


Figure 3. Splicing Requires Numerous Rearrangements

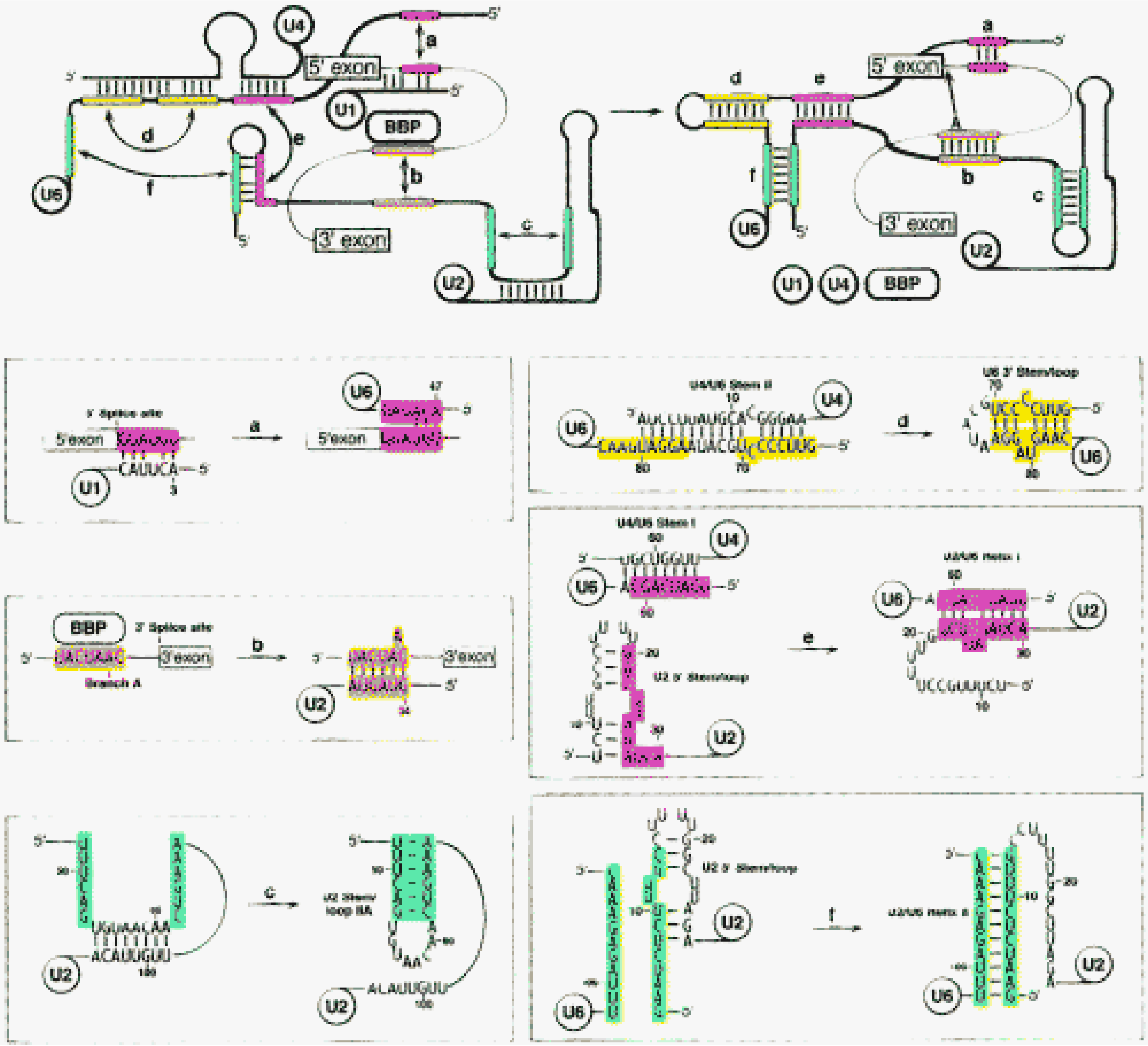
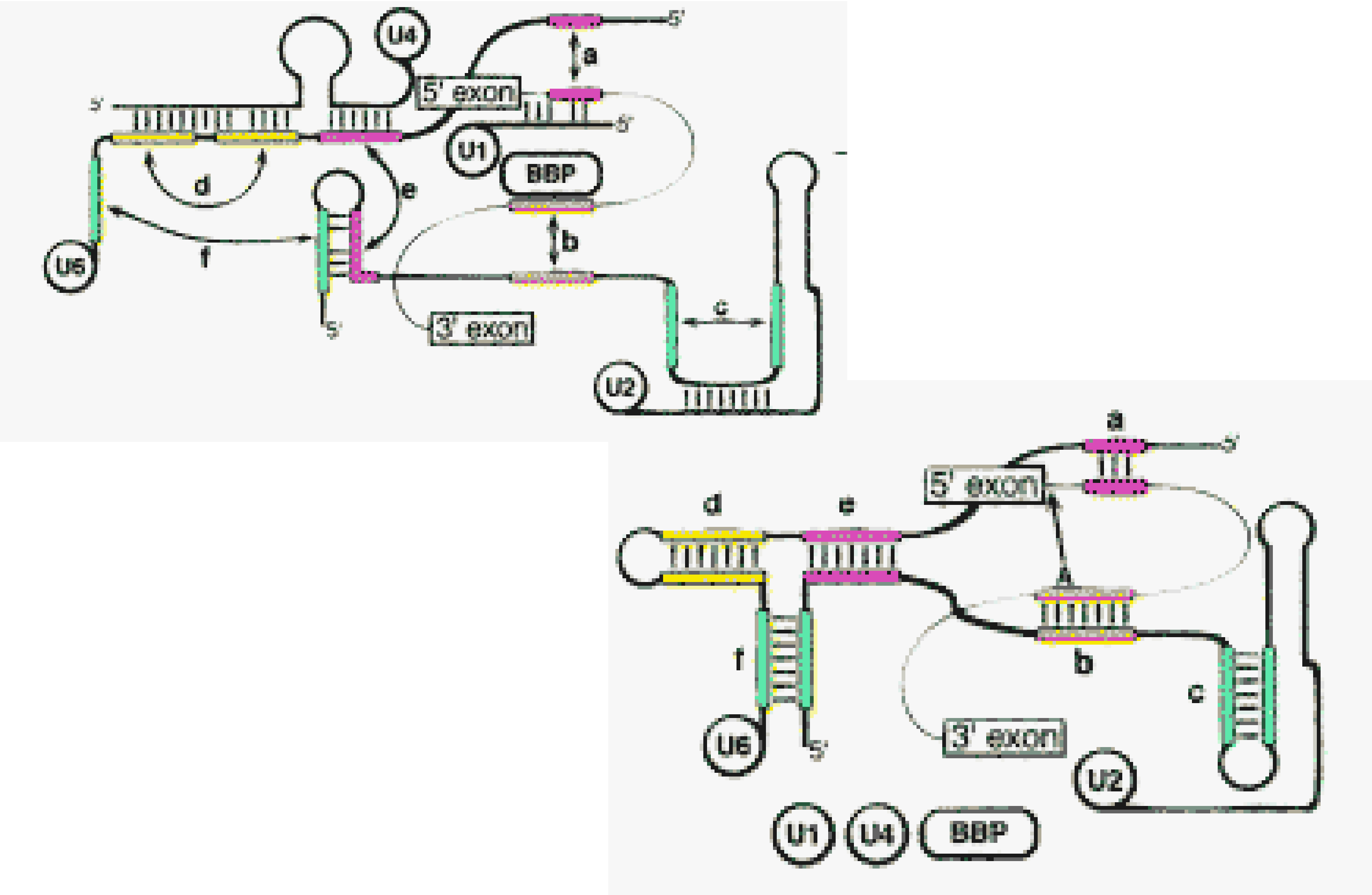


Figure 3. Splicing Requires Numerous Rearrangements



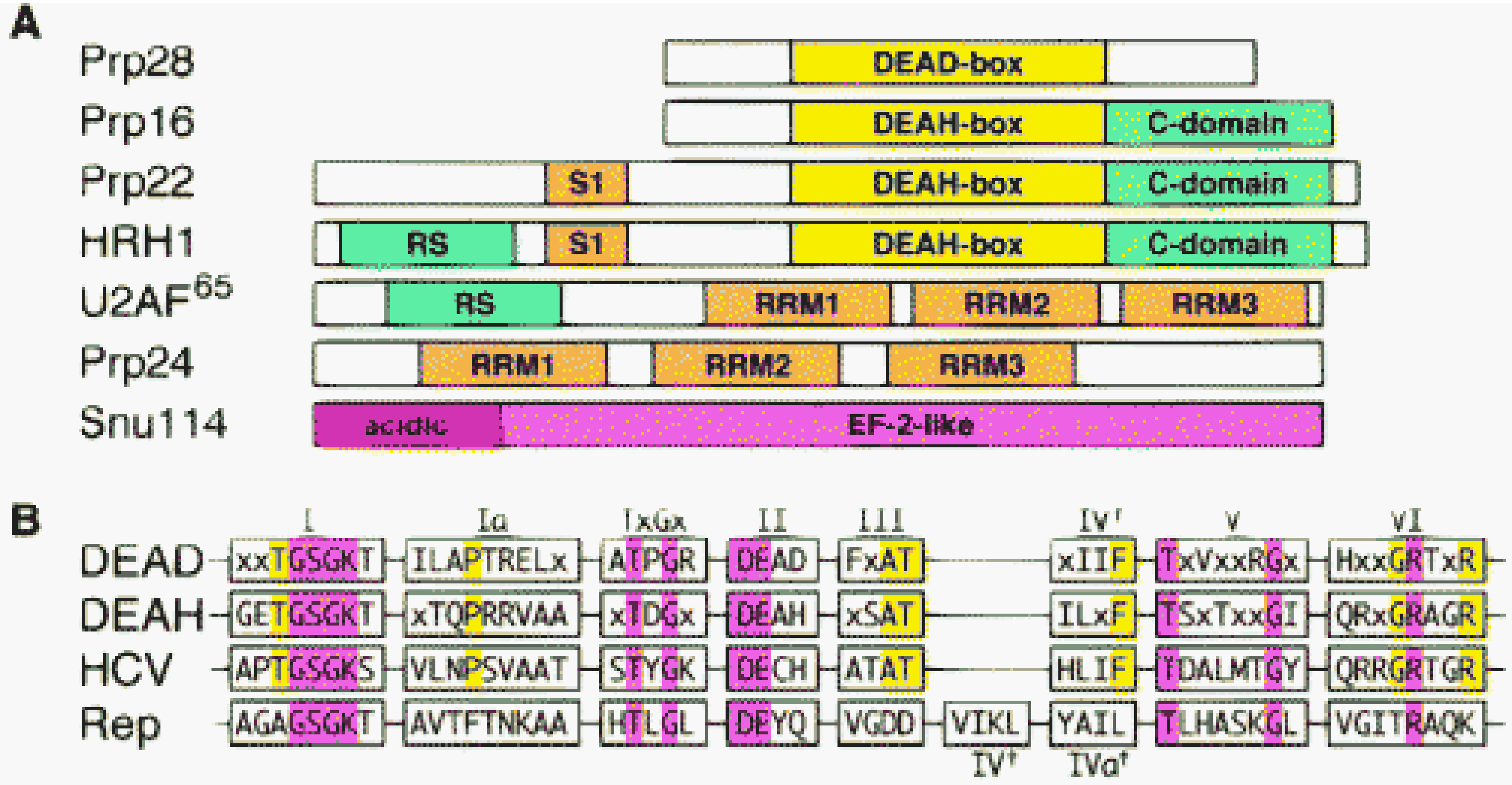

A
Prp28
DEAD-box
Prp16
DEAH-box
C-domain
Prp22
S1
DEAH-box
C-domain
HRH1
RS
S1
DEAH-box
C-domain
U2AF65
RS
RRM1
RRM2
RRM3
Prp24
RRM1
RRM2
RRM3
Snu114
acidic
EF-2-like
B
I Ia IxGx II III IV' V VI
DEAD xxTGSGKT ILAPTRELx AIPGR DEAD FxAT xIIF TxVxxRGx HxxGRTxR
DEAH GETGSGKT xTQPRRVAA xIDGx DEAH xSAT ILxF TSxTxxGI QRxGRAGR
HCV APTGSGKS VLNPSVAAT STYGK DECH ATAT HLIF TDALMTGY QRRGRTGR
Rep AGAGSGKT AVTFTNKAA HTLGL DEYQ VGDD VIKL YAIL TLHASKGL VGITRAQK
IV† IVa†

Figure 5. Sequence Characteristics of the Spliceosome's Mechanical Gadgets(A) Examples of domain structure. DEAD and DEAH, helicase-like domains; C-domain, conserved in the DEAH proteins; S1, a ribosomal motif implicated in RNA binding; RS, rich in serine/arginine dipeptides; RRM, RNA recognition motif; EF-2, elongation factor 2. All factors are from *S. cerevisiae* except for the mammalian factors U2AF^{65} and HRH1, the human ortholog of Prp22.(B) Sequence motifs of the DExD/H box domains. DEAD, residues identical between Prp5, Prp28, and U5⋚100 kDa (Table 1). DEAH, amino acid residues identical between Prp2, Prp16, Prp22, Prp43, hPRP16, and HRH1 (Table 1). x, any amino acid. The specific

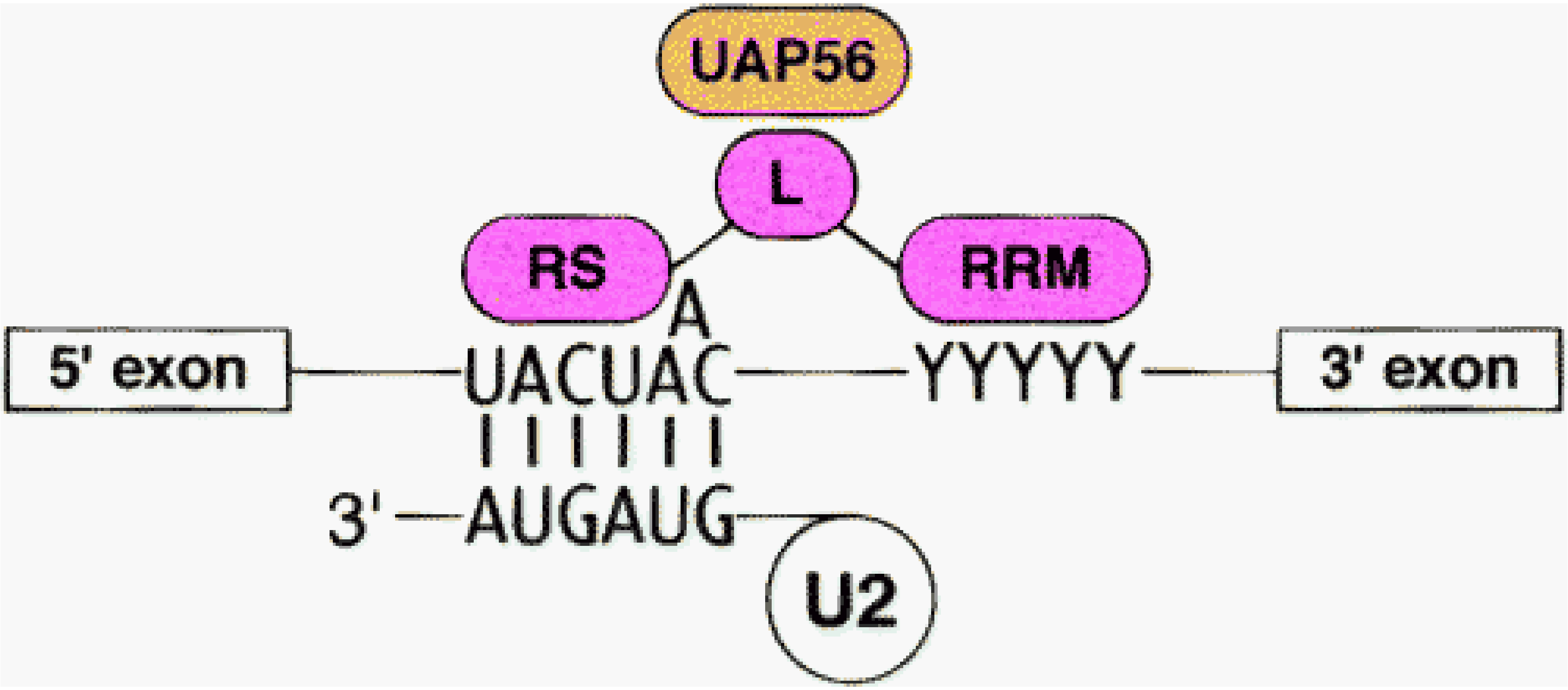


Figure 6. A Paradigm for Unwindase Specificity and Timing?The DExD/H box protein UAP56 (orange) binds U2AF^{65} (pink) through its linker region (L). U2 binds the branch point. Y's indicate the polypyrimidine stretch; RS, RRM as in Figure 5A. Sequences are from mammals.

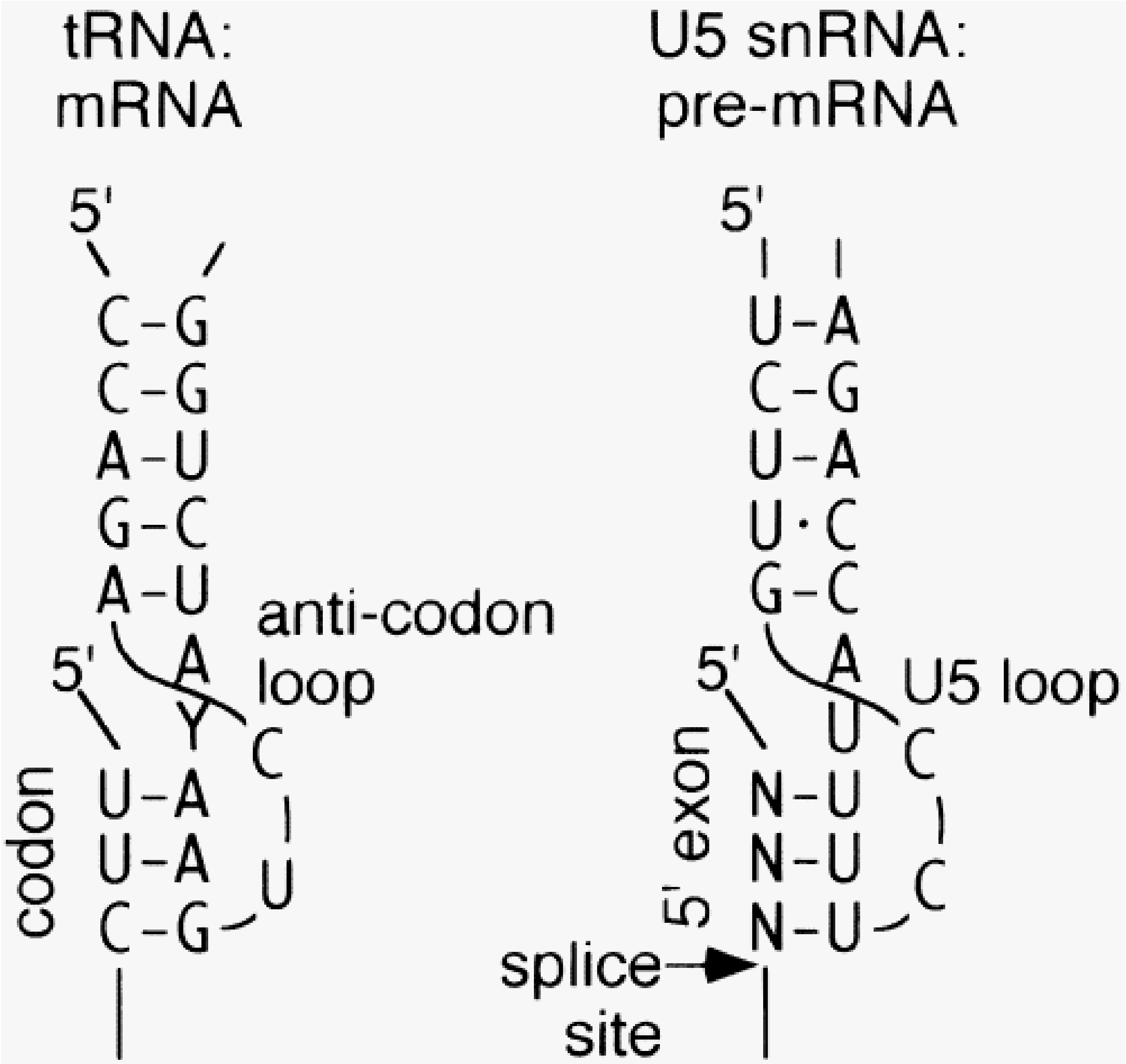
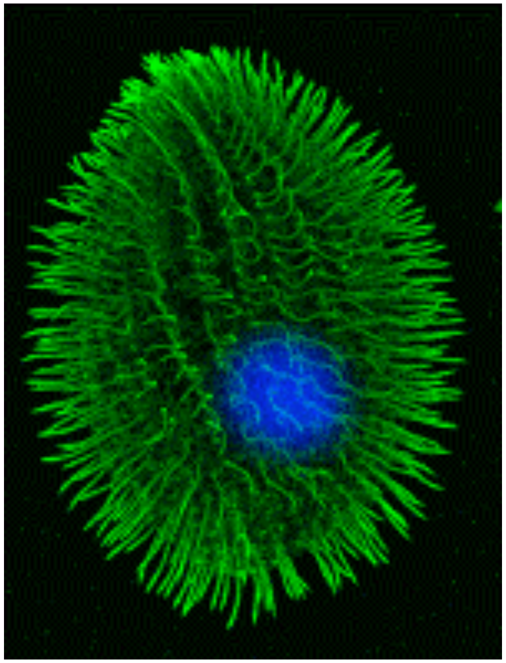
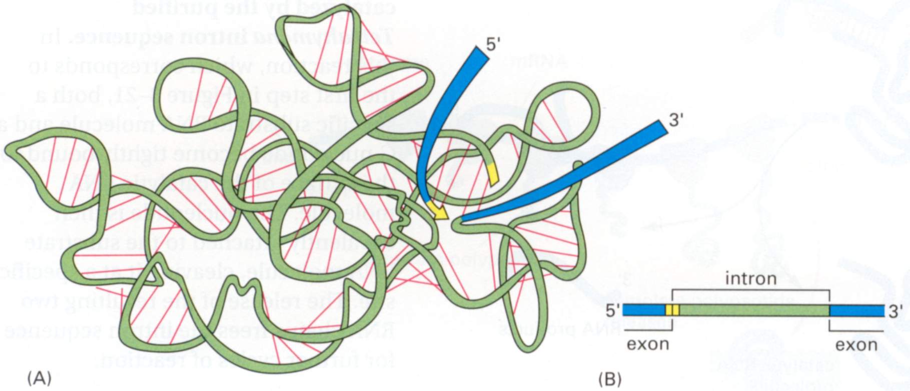
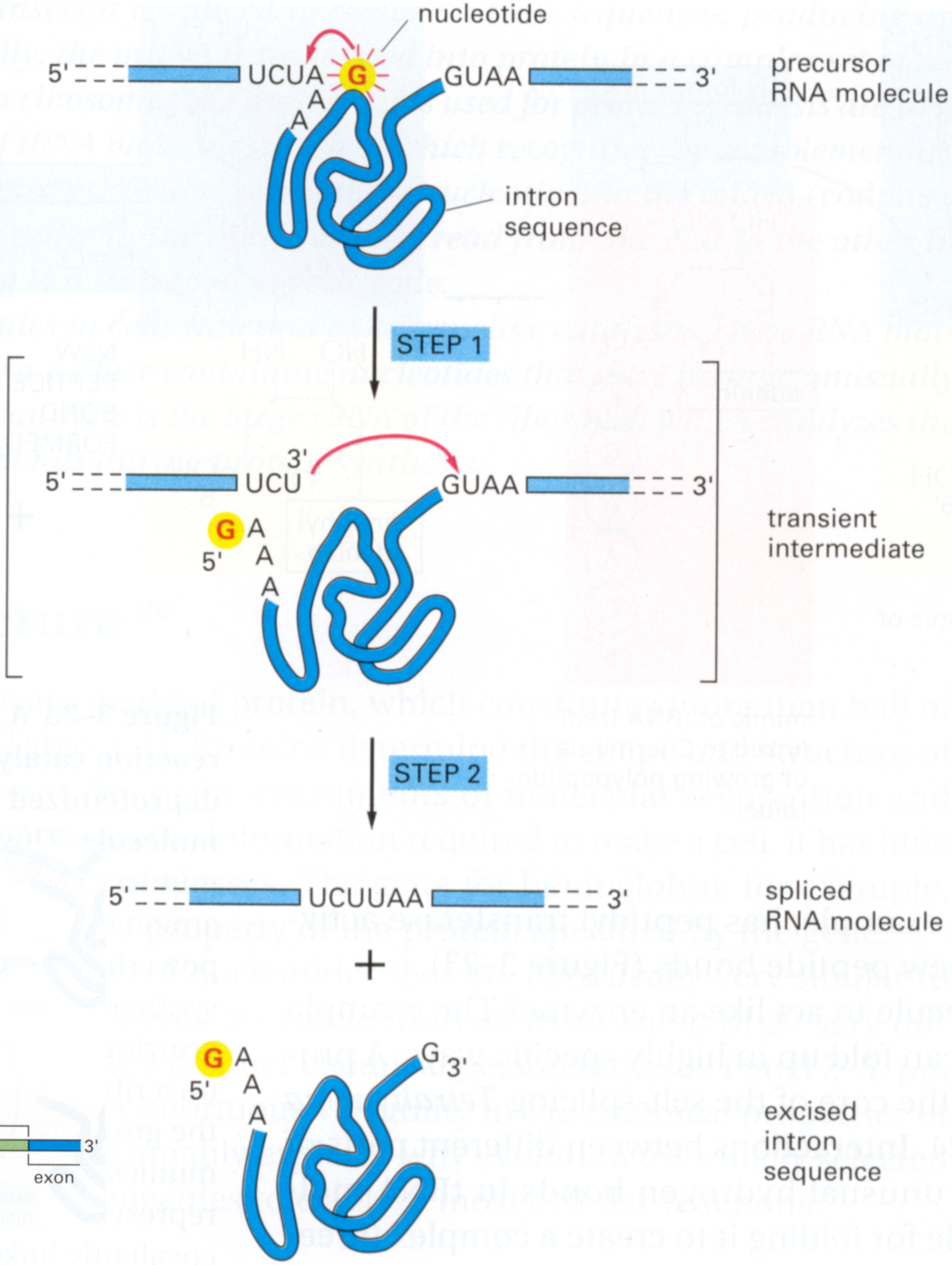


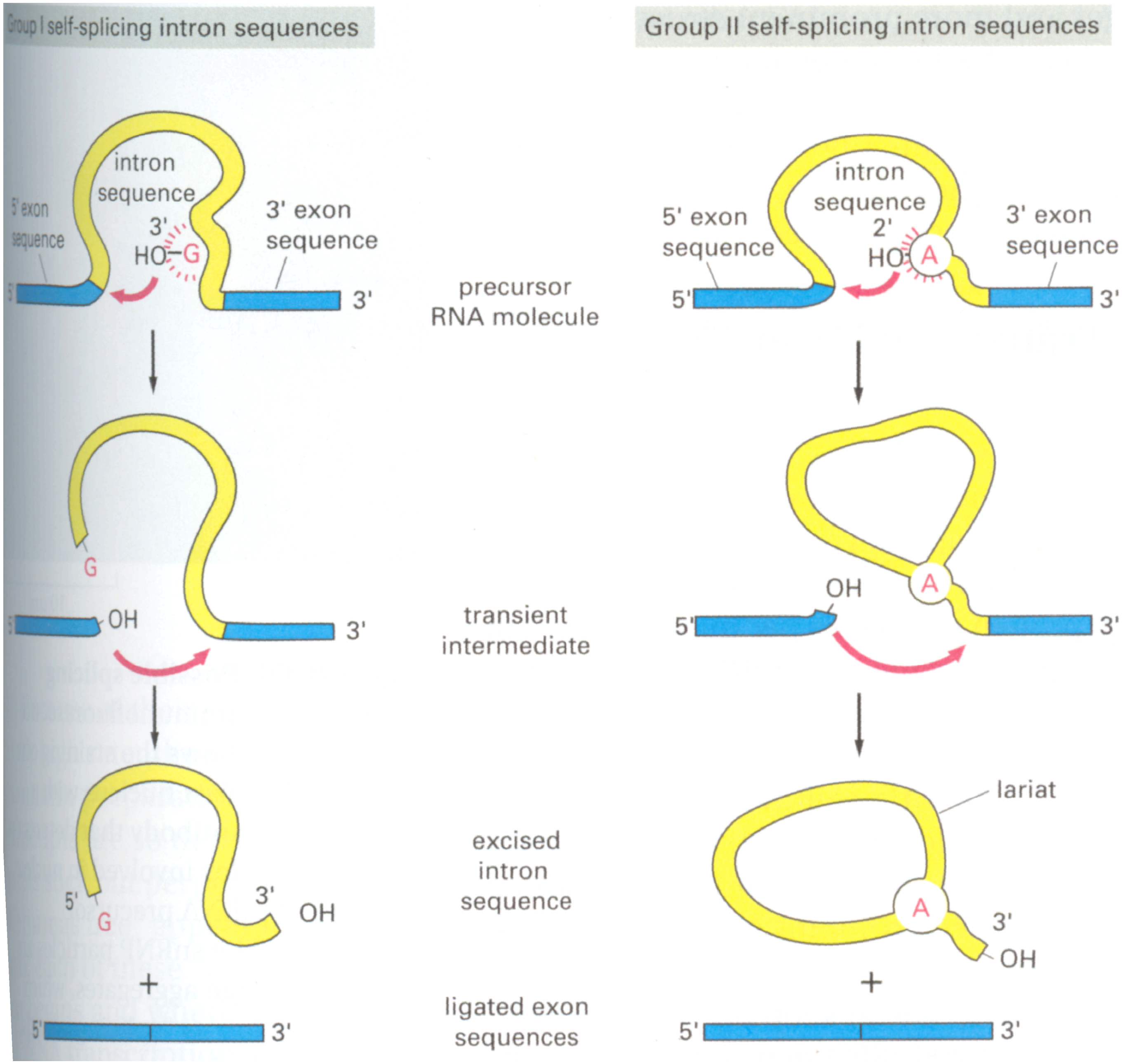
Figure 7. A Parallel between the Spliceosome and the Ribosome?The binding of a yeast Phe codon by the anticodon loop of the cognate tRNA is compared with the binding of a 5′ exon by the yeast U5 loop in a hypothetical, yet provocative, configuration. N, any nucleotide.



Tetrahymena thermophila






Group I self-splicing intron sequences

Group II self-splicing intron sequences

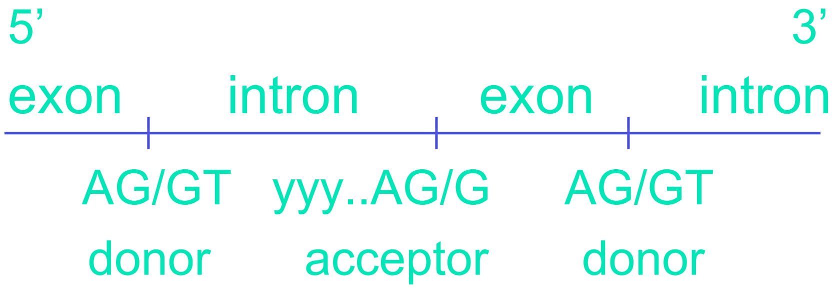
Eukaryotes

As in prokaryotes (but maybe more variable)

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

New Features:

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



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	1340bp	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,

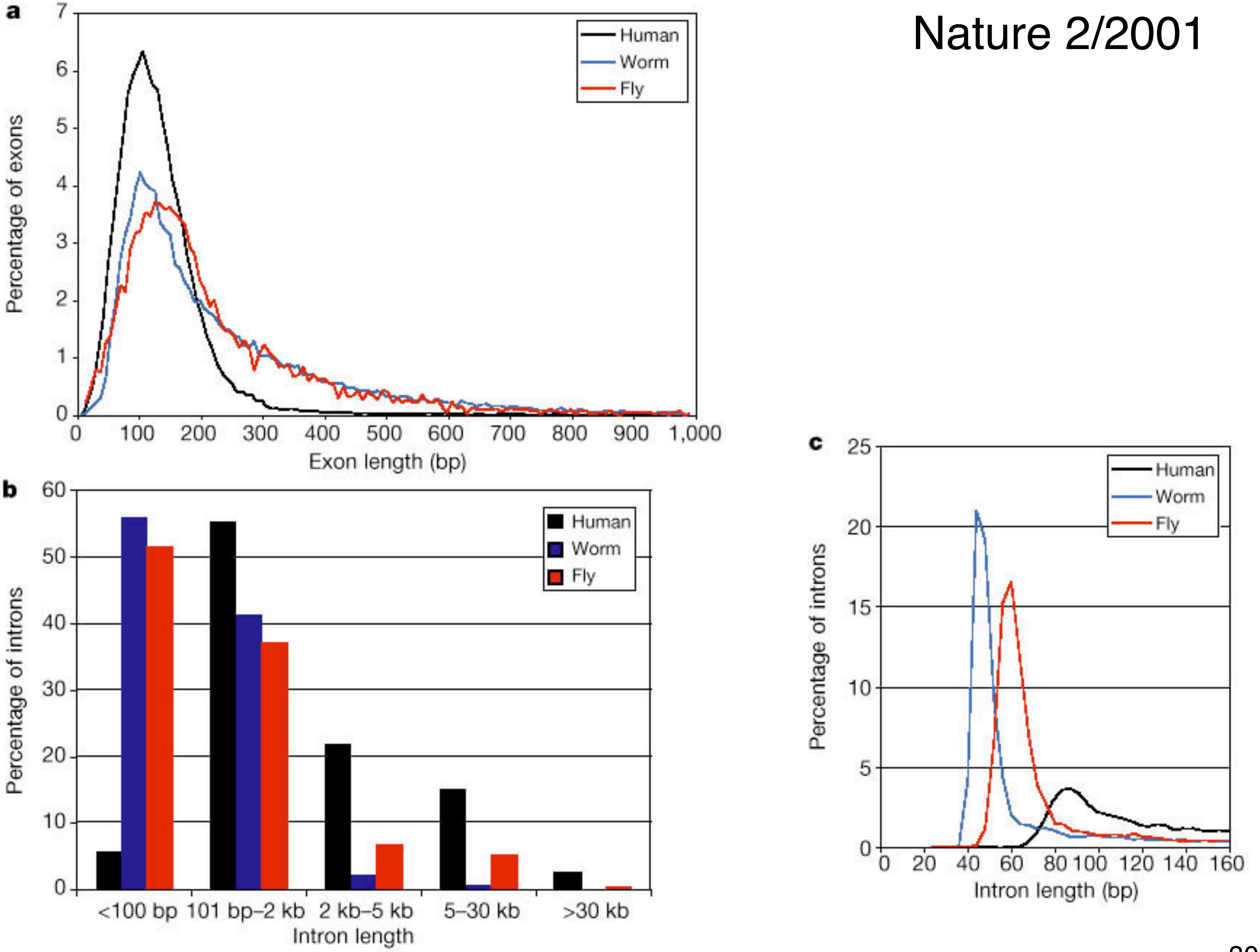
Max known: dystrophin gene (DMD), 2.4 Mb.

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: 2.5 kb/min

Nature 2/2001



Nature 2/2001

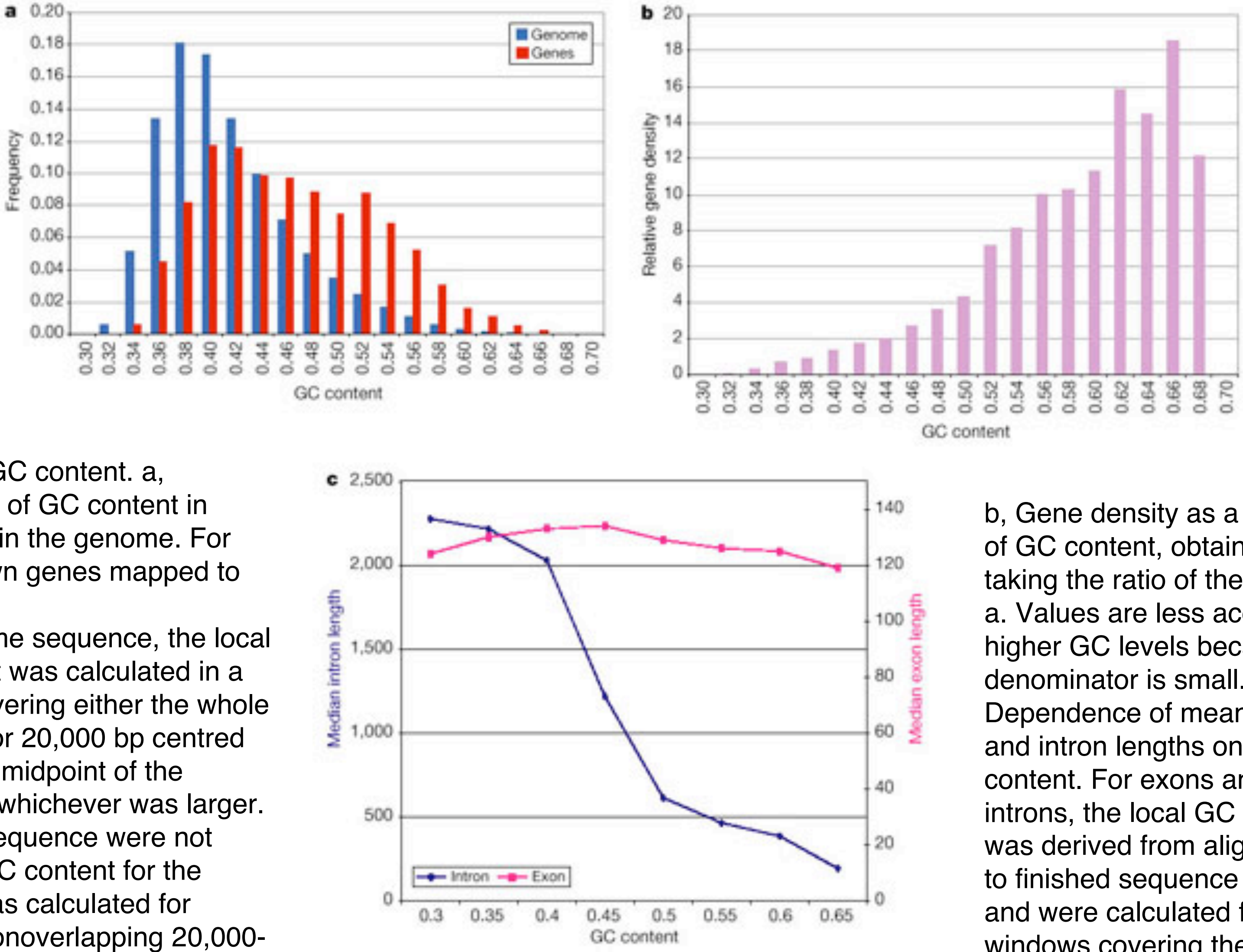


Figure 36 GC content. a, Distribution of GC content in genes and in the genome. For 9,315 known genes mapped to the
draft genome sequence, the local GC content was calculated in a window covering either the whole alignment or 20,000 bp centred around the midpoint of the alignment, whichever was larger. 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 the gene and genome distributions have been normalized to sum to one.

b, Gene density as a function of GC content, obtained by taking the ratio of the data in a. Values are less accurate at higher GC levels because the denominator is small. c, Dependence of mean exon and intron lengths on GC content. For exons and introns, the local GC content was derived from alignments to finished sequence only, and were calculated from windows covering the feature or 10,000 bp centred on the feature, whichever was larger.

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

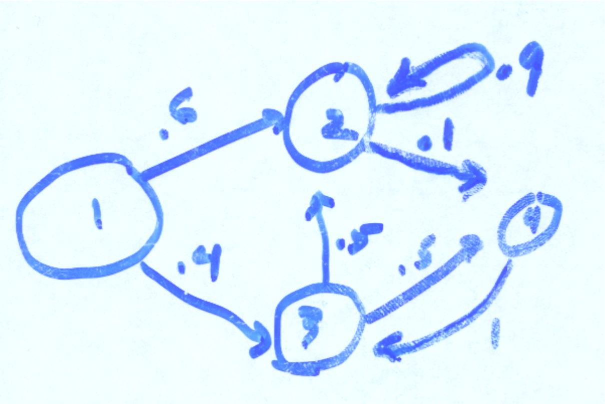
NO alternate splicing, none > 30kb, ...

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



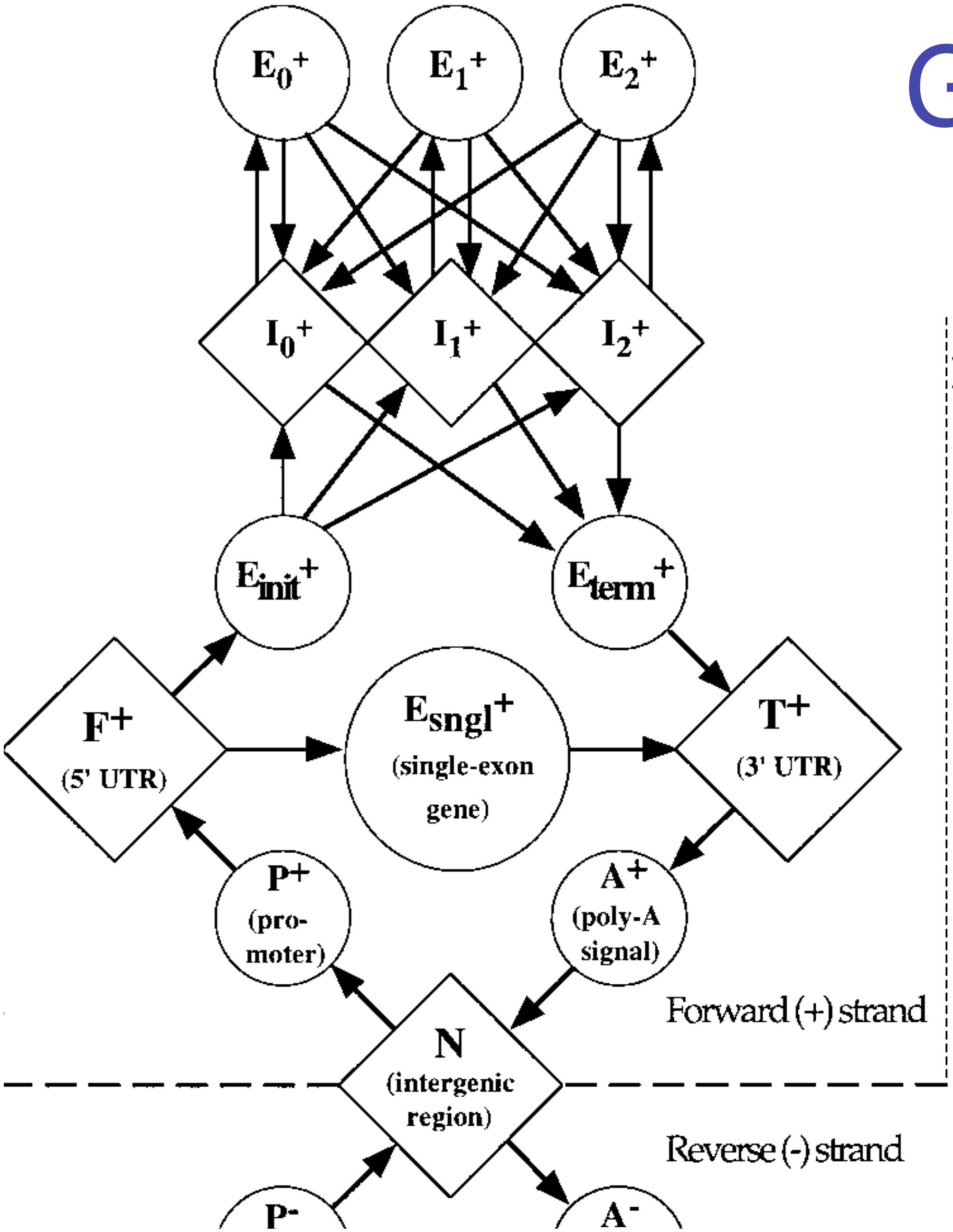
- π: Initial state distribution
- a_{ij}: Transition probabilities
- One submodel per state
- Outputs are *strings* gen'ed by submodel
- Given length L
    - Pick start state q_1 (~π)
    - While $\sum d_i < L$
        - Pick d_i
        - Pick string s_i of length $d_i = |s_i|$ ~ submodel for q_i
        - Pick next state q_{i+1} (~a_{ij})
    - Output $s_1 s_2$…

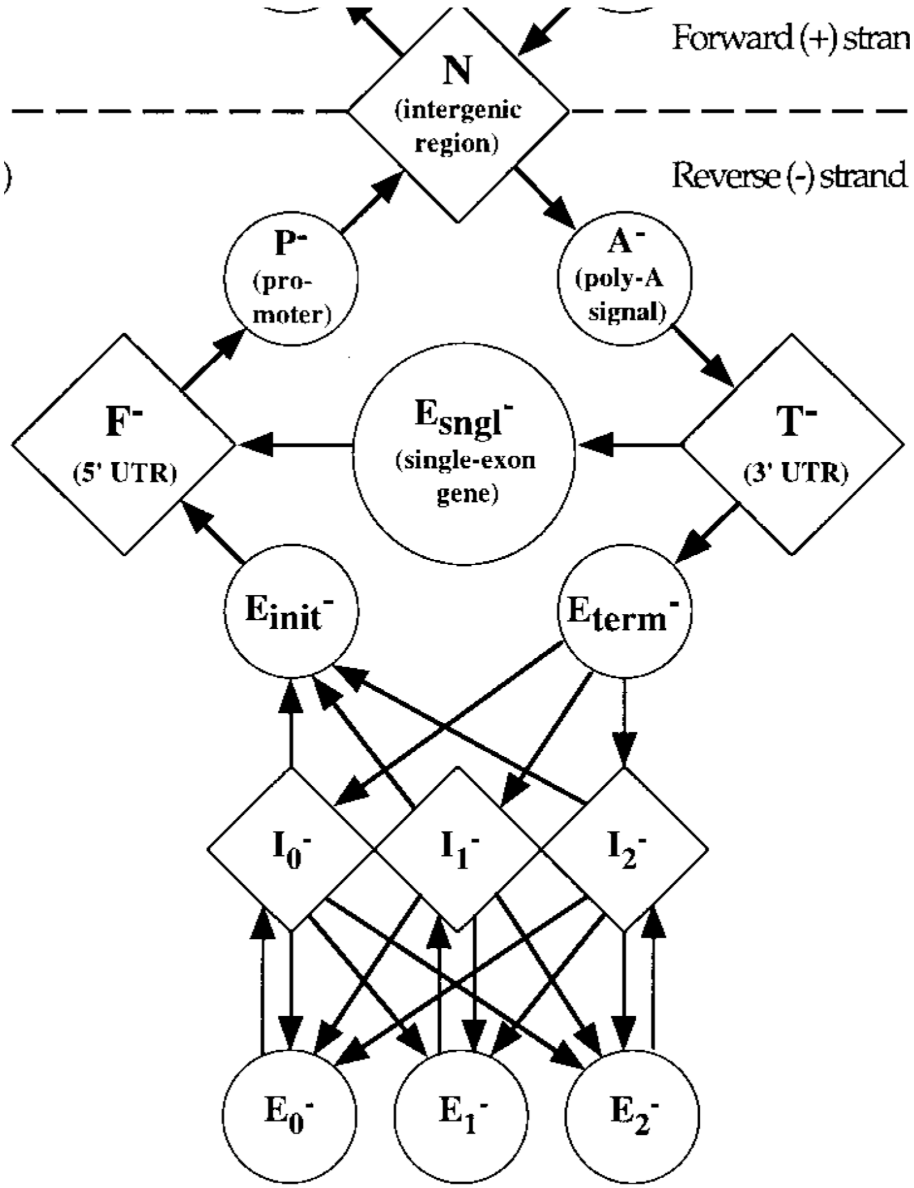
Decoding

- A “parse” ϕ of $s = s_1 s_2 \ldots s_L$ is a pair
  $d = d_1 d_2 \ldots d_k$ $q = q_1 q_2 \ldots q_k$ with $\sum d_i = L$
- Now use something like the forward/backward algorithms to calculate probabilities like “P(seq up to position i generated ending in state q_k)”, which involves summing over possible predecessor states q_{k-1} and possible d_k

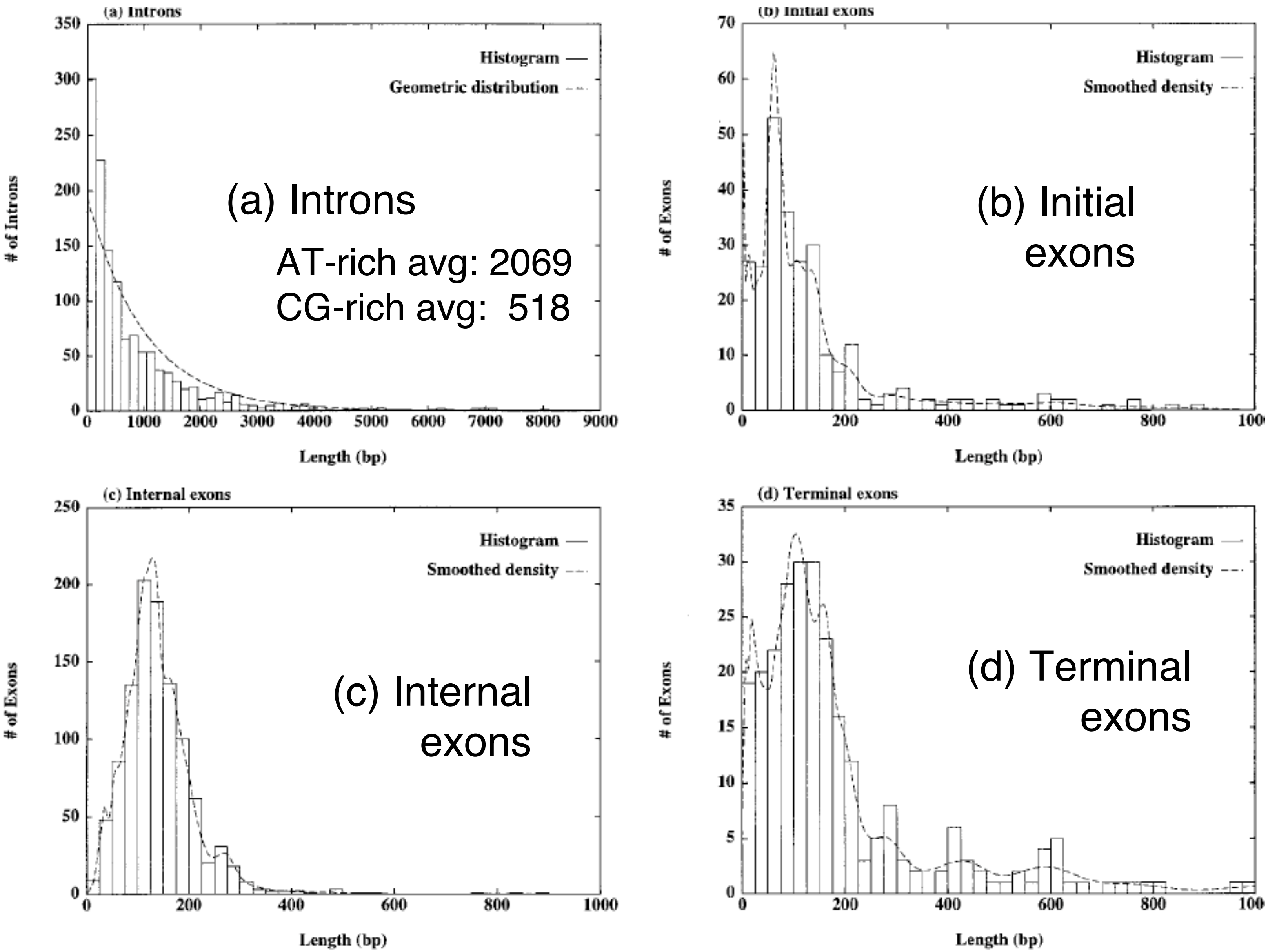
$$Pr(\phi \mid s) = \frac{P_r(\phi \wedge s)}{P_r(s)} \quad \ldots$$

GHMM Structure





Length Distributions



(a) Introns

AT-rich avg: 2069
CG-rich avg: 518

(b) Initial exons

(c) Internal exons

(d) Terminal exons

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

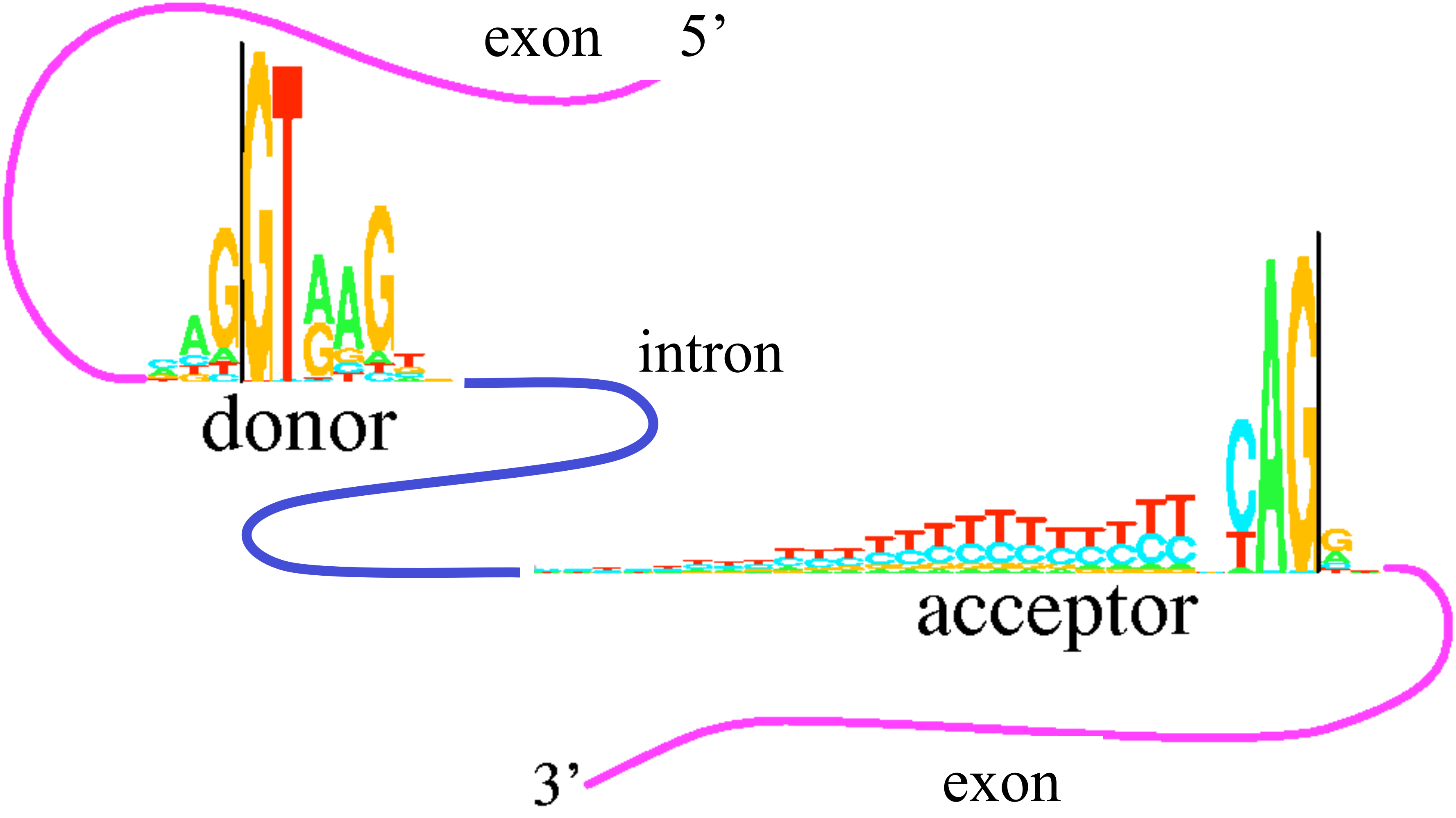
Acceptor Splice Site (3' end of intron)

- [-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's in the Primary Sequence?



exon 5’

intron

donor

acceptor

3’

exon

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

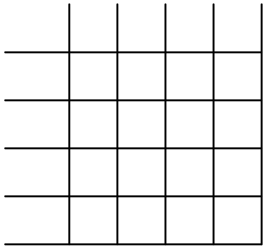
Uses χ^2 test to quantitate dependence

χ^2 test for independence

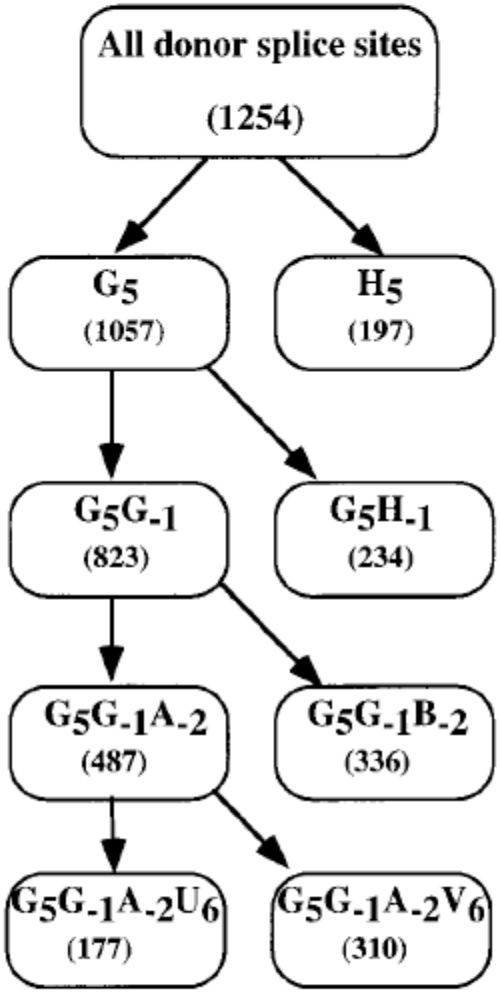
i	Con	j: -3	-2	-1	+3	+4	+5	+6	Sum
-3	c/a	---	61.8*	14.9	5.8	20.2*	11.2	18.0*	131.8*
-2	A	115.6*	---	40.5*	20.3*	57.5*	59.7*	42.9*	336.5*
-1	G	15.4	82.8*	---	13.0	61.5*	41.4*	96.6*	310.8*
+3	a/g	8.6	17.5*	13.1	---	19.3*	1.8	0.1	60.5*
+4	A	21.8*	56.0*	62.1*	64.1*	---	56.8*	0.2	260.9*
+5	G	11.6	60.1*	41.9*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

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



“expected” means expected assuming independence



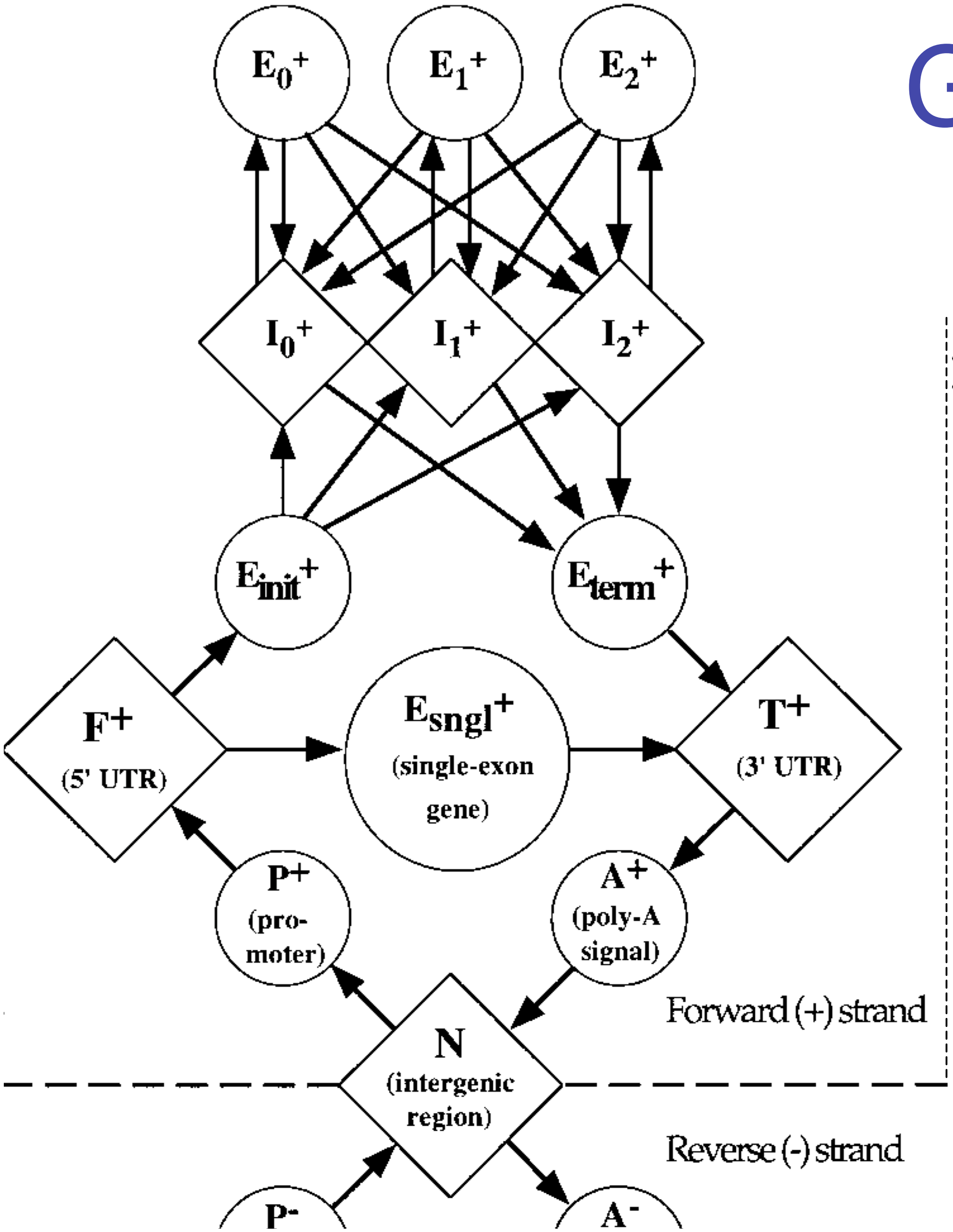
Pos	A%	C%	G%	U%
-3	33	36	19	13
-2	56	15	15	15
-1	9	4	78	9
+3	44	3	51	3
+4	75	4	13	9
+6	14	18	19	49
-3	34	37	18	11
-2	59	10	15	16
+3	40	4	53	3
+4	70	4	16	10
+6	17	21	21	42
-3	37	42	18	3
+3	39	5	51	5
+4	62	5	22	11
+6	19	20	25	36
-3	32	40	23	5
+3	27	4	59	10
+4	51	5	25	19

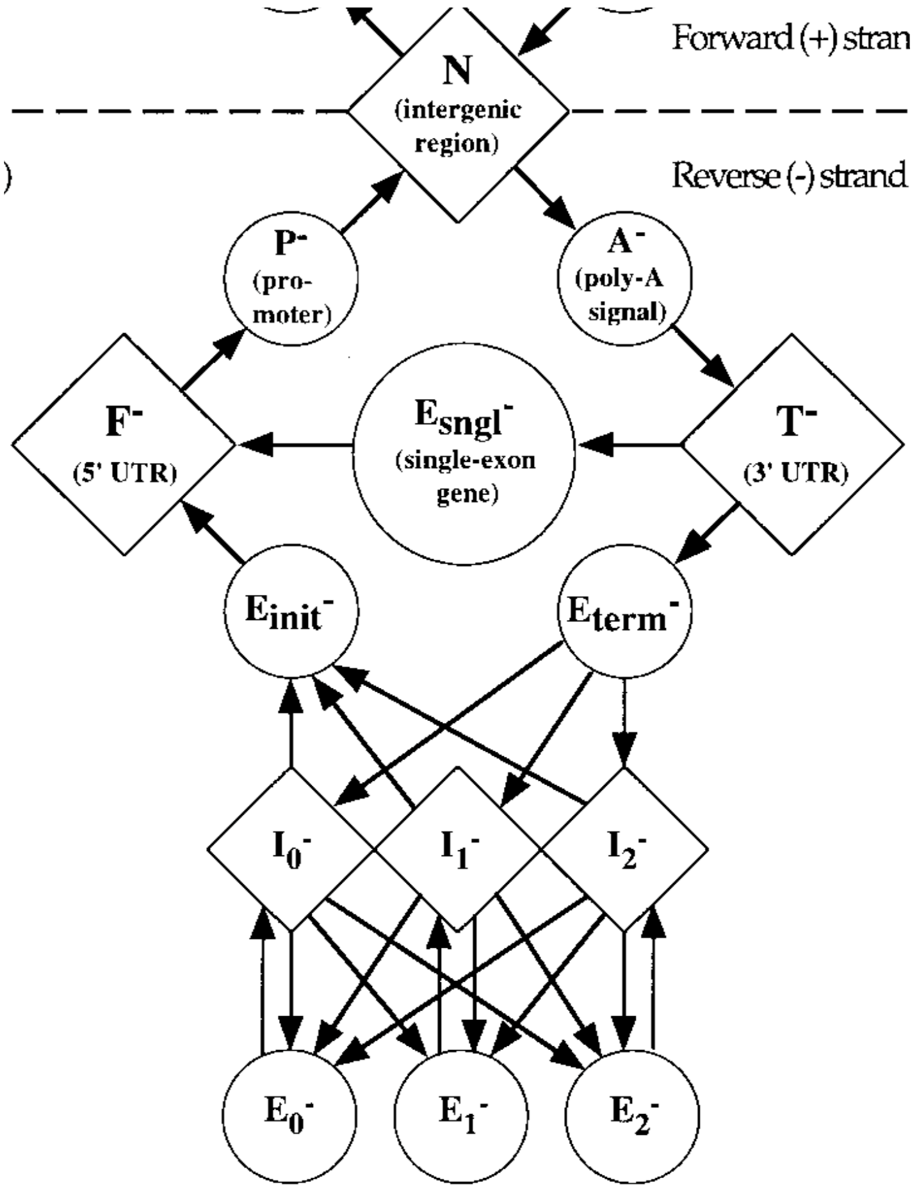
Pos	A%	C%	G%	U%
-3	35	44	16	6
-2	85	4	7	5
-1	2	1	97	0
+3	81	3	15	2
+4	51	28	9	12
+6	22	20	30	28
-3	29	31	21	18
-2	43	30	17	11
+3	56	0	43	0
+4	93	2	3	3
+6	5	10	10	76
-3	29	30	18	23
+3	42	1	56	1
+4	80	4	8	8
+6	14	21	16	49
-3	39	43	15	2
+3	46	6	46	3
+4	69	5	20	7

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'

GHMM Structure





Summary of Burge & Karlin

Coding DNA & control signals 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

Problems with all methods

Pseudo genes

Short ORFs

Sequencing errors

Non-coding RNA genes & spliced UTR’s

Overlapping genes

Alternative splicing/polyadenylation

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?

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