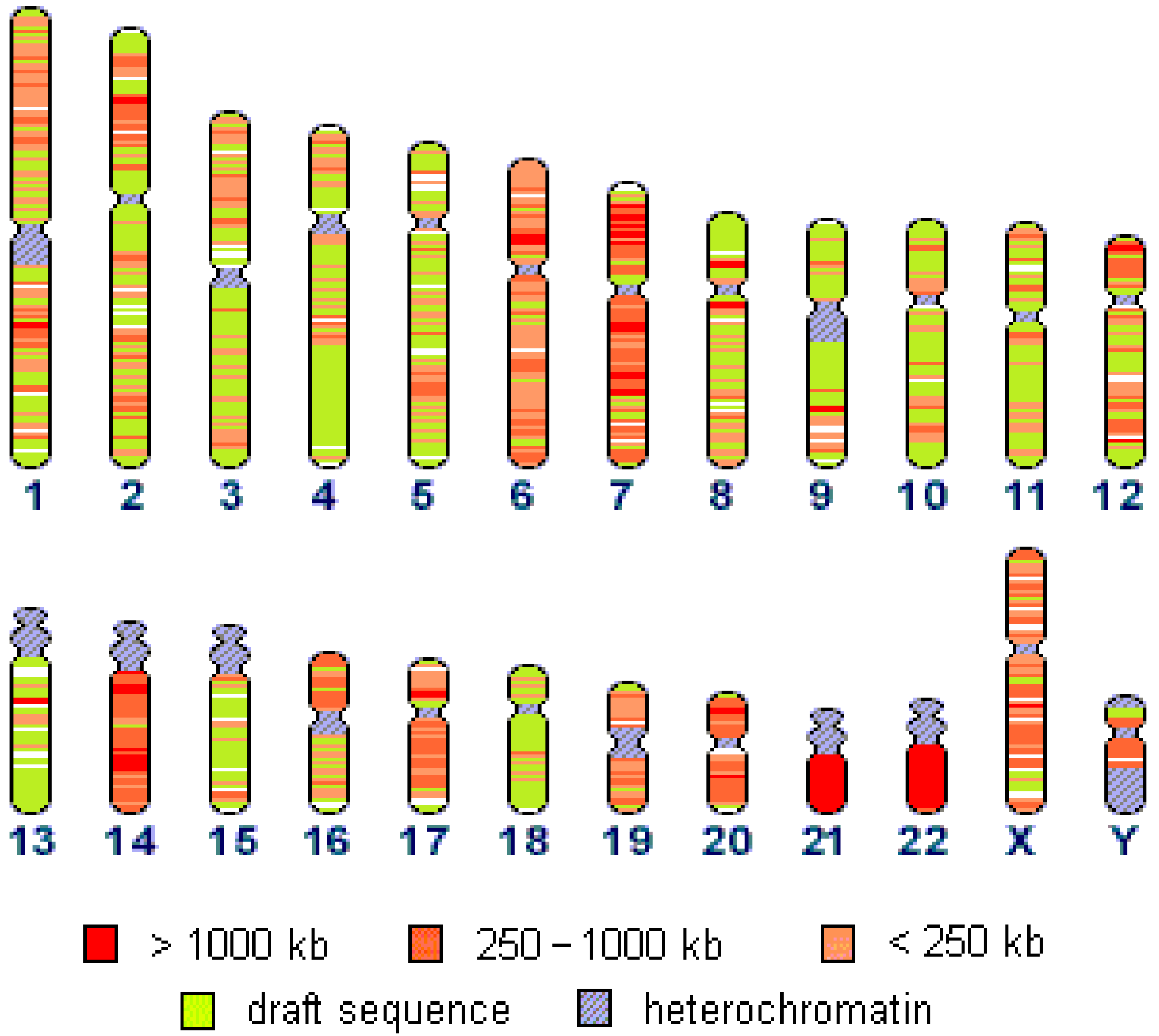
Human Genome Project: sequencing

Dec 12, 2000

Draft ⬜🟩

Finished 🟥



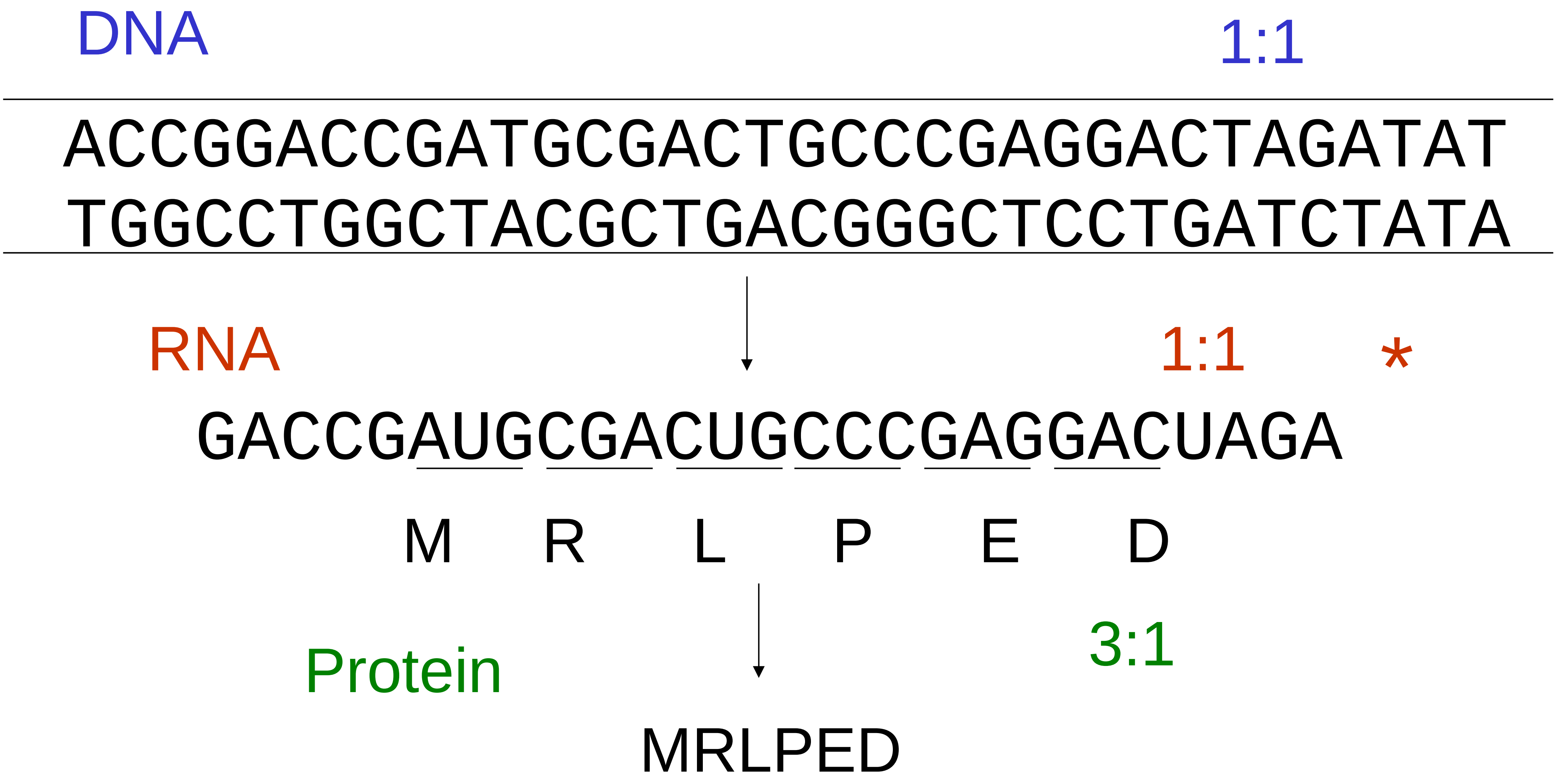
Structure of a Human Gene (PSA)

```
GGTGTCTTAGGCACACTGGTCTTGGAGTGCAAAGGATCTAGGCACGTGAGGCTTTGTATGAAGAATCGGGGATCGTACCCACCCCCTGTTTCTGTTTCATCCTGGGCATGTCTCCTCTGCCTTTGTCCCC
TAGATGAAGTCTCCATGAGCTACAAGGGCCTGGTGCATCCAGGGTGATCTAGTAATTGCAGAACAGCAAGTGCTAGCTCTCCCTCCCCTTCCACAGCTCTGGGTGTGGGAGGGGGTTGTCCAGCCTCCAG
CAGCATGGGGAGGGCCTTGGTCAGCCTCTGGGTGCCAGCAGGGCAGGGGCGGAGTCCTGGGGAATGAAGGTTTTATAGGGCTCCTGGGGGAGGCTCCCCAGCCCCAAGCTTACCACCTGCACCCGGAGAG
CTGTGTCACCATGTGGGTCCCGGTTGTCTTCCTCACCCTGTCCGTGACGTGGATTGGTGAGAGGGGCCATGGTTGGGGGGATGCAGGAGAGGGAGCCAGCCCTGACTGTCAAGCTGAGGCTCTTTCCCCC
CCAACCCAGCACCCCAGCCCAGACAGGGAGCTGGGCTCTTTTCTGTCTCTCCCAGCCCCACTTCAAGCCCATACCCCCAGTCCCCTCCATATTGCAACAGTCCTCACTCCCACACCAGGTCCCCGCTCCC
TCCCACTTACCCCAGAACTTTCTTCCCATTTGCCCAGCCAGCTCCCTGCTCCCAGCTGCTTTACTAAAGGGGAAGTTCCTGGGCATCTCCGTGTTTCTCTTTGTGGGGCTCAAAACCTCCAAGGACCTCT
CTCAATGCCATTGGTTCCTTGGACCGTATCACTGGTCCATCTCCTGAGCCCCTCAATCCTATCACAGTCTACTGACTTTTCCCATTCAGCTGTGAGTGTCCAACCCTATCCCAGAGACCTTGATGCTTGG
CCTCCCAATCTTGCCCTAGGATACCCAGATGCCAACCAGACACCTCCTTCTTTCCTAGCCAGGCTATCTGGCCTGAGACAACAAATGGGTCCCTCAGTCTGGCAATGGGACTCTGAGAACTCCTCATTCC
CTGACTCTTAGCCCCAGACTCTTCATTCAGTGGCCCACATTTTCCTTAGGAAAAACATGAGCATCCCCAGCCACAACTGCCAGCTCTCTGAGTCCCCAAATCTGCATCCTTTTCAAAACCTAAAAACAAA
AAGAAAAACAAATAAAACAAAACCAACTCAGACCAGAACTGTTTTCTCAACCTGGGACTTCCTAAACTTTCCAAAACCTTCCTCTTCCAGCAACTGAACCTCGCCATAAGGCACTTATCCCTGGTTCCTA
GCACCCCTTATCCCCTCAGAATCCACAACTTGTACCAAGTTTCCCTTCTCCCAGTCCAAGACCCCAAATCACCACAAAGGACCCAATCCCCAGACTCAAGATATGGTCTGGGCGCTGTCTTGTGTCTCCT
ACCCTGATCCCTGGGTTCAACTCTGCTCCCAGAGCATGAAGCCTCTCCACCAGCACCAGCCACCAACCTGCAAACCTAGGGAAGATTGACAGAATTCCCAGCCTTTCCCAGCTCCCCCTGCCCATGTCCC
AGGACTCCCAGCCTTGGTTCTCTGCCCCCGTGTCTTTTCAAACCCACATCCTAAATCCATCTCCTATCCGAGTCCCCCAGTTCCCCCTGTCAACCCTGATTCCCCTGATCTAGCACCCCCTCTGCAGGCG
CTGCGCCCCTCATCCTGTCTCGGATTGTGGGAGGCTGGGAGTGCGAGAAGCATTCCCAACCCTGGCAGGTGCTTGTGGCCTCTCGTGGCAGGGCAGTCTGCGGCGGTGTTCTGGTGCACCCCCAGTGGGT
CCTCACAGCTGCCCACTGCATCAGGAAGTGAGTAGGGGCCTGGGGTCTGGGGAGCAGGTGTCTGTGTCCCAGAGGAATAACAGCTGGGCATTTTCCCCAGGATAACCTCTAAGGCCAGCCTTGGGACTGG
GGGAGAGAGGGAAAGTTCTGGTTCAGGTCACATGGGGAGGCAGGGTTGGGGCTGGACCACCCTCCCCATGGCTGCCTGGGTCTCCATCTGTGTCCCTCTATGTCTCTTTGTGTCGCTTTCATTATGTCTC
TTGGTAACTGGCTTCGGTTGTGTCTCTCCGTGTGACTATTTTGTTCTCTCTCTCCCTCTCTTCTCTGTCTTCAGTCTCCATATCTCCCCCTCTCTCTGTCCTTCTCTGGTCCCTCTCTAGCCAGTGTGTC
TCACCCTGTATCTCTCTGCCAGGCTCTGTCTCTCGGTCTCTGTCTCACCTGTGCCTTCTCCCTACTGAACACACGCACGGGATGGGCCTGGGGGACCCTGAGAAAAGGAAGGGCTTTGGCTGGGCGCGGT
GGCTCACACCTGTAATCCCAGCACTTTGGGAGGCCAAGGCAGGTAGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACTGGTGAAACCCCATCTCTACTAAAAATACAAAAAATTAGCCAGGC
GTGGTGGCGCATGCCTGTAGTCCCAGCTACTCAGGAGCTGAGGGAGGAGAATTGCATTGAACCTGGAGGTTGAGGTTGCAGTGAGCCGAGACCGTGCCACTGCACTCCAGCCTGGGTGACAGAGTGAGAC
TCCGCCTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAAAAGAAAAGAAAAGAAAAGGAAGTGTTTTATCCCTGATGTGTGTGGGTATGAGGGTATGAGAGGGCCCCTCTCACTCCATTCCTTCTCCA
GGACATCCCTCCACTCTTGGAGACACAGAGAAGGGCTGGTTCCAGCTGGAGCTGGGAGGGGCAATTGAGGGAGGAGGAAGGAGAAGGGGGAAGGAAAACAGGGTATGGGGGAAAGGACCCTGGGGAGCG
AAGTGGAGGATACAACCTTGGGCCTGCAGGCAGGCTACCTACCCACTTGGAAACCCACGCCAAAGCCGCATCTACAGCTGAGCCACTCTGAGGCCTCCCCTCCCCGGCGGTCCCCACTCAGCTCCAAAGT
CTCTCTCCCTTTTCTCTCCCACACTTTATCATCCCCCGGATTCCTCTCTACTTGGTTCTCATTCTTCCTTTGACTTCCTGCTTCCCTTTCTCATTCATCTGTTTCTCACTTTCTGCCTGGTTTTGTTCTT
CTCTCTCTCTTTCTCTGGCCCATGTCTGTTTCTCTATGTTTCTGTCTTTTCTTTCTCATCCTGTGTATTTTCGGCTCACCTTGTTTGTCACTGTTCTCCCCTCTGCCCTTTCATTCTCTCTGCCCTTTTA
CCCTCTTCCTTTTCCCTTGGTTCTCTCAGTTCTGTATCTGCCCTTCACCCTCTCACACTGCTGTTTCCCAACTCGTTGTCTGTATTTTGGCCTGAACTGTGTCTTCCCAACCCTGTGTTTTCTCACTGTT
TCTTTTTCTCTTTTGGAGCCTCCTCCTTGCTCCTCTGTCCCTTCTCTCTTTCCTTATCATCCTCGCTCCTCATTCCTGCGTCTGCTTCCTCCCCAGCAAAAGCGTGATCTTGCTGGGTCGGCACAGCCTG
TTTCATCCTGAAGACACAGGCCAGGTATTTCAGGTCAGCCACAGCTTCCCACACCCGCTCTACGATATGAGCCTCCTGAAGAATCGATTCCTCAGGCCAGGTGATGACTCCAGCCACGACCTCATGCTGC
TCCGCCTGTCAGAGCCTGCCGAGCTCACGGATGCTGTGAAGGTCATGGACCTGCCCACCCAGGAGCCAGCACTGGGGACCACCTGCTACGCCTCAGGCTGGGGCAGCATTGAACCAGAGGAGTGTACGCT
GGGCCAGATGGTGCAGCCGGGAGCCCAGATGCCTGGGTCTGAGGGAGGAGGGGACAGGACTCCTGGGTCTGAGGGAGGAGGGCCAAGGAACCAGGTGGGGTCCAGCCCACAACAGTGTTTTTGCCTGGCC
CGTAGTCTTGACCCCAAAGAAACTTCAGTGTGTGGACCTCCATGTTATTTCCAATGACGTGTGTGCGCAAGTTCACCCTCAGAAGGTGACCAAGTTCATGCTGTGTGCTGGACGCTGGACAGGGGGCAAA
AGCACCTGCTCGGTGAGTCATCCCTACTCCCAAGATCTTGAGGGAAAGGTGAGTGGGACCTTAATTCTGGGCTGGGGTCTAGAAGCCAACAAGGCGTCTGCCTCCCCTGCTCCCCAGCTGTAGCCATGCC
ACCTCCCCGTGTCTCATCTCATTCCCTCCTTCCCTCTTCTTTGACTCCCTCAAGGCAATAGGTTATTCTTACAGCACAACTCATCTGTTCCTGCGTTCAGCACACGGTTACTAGGCACCTGCTATGCACC
CAGCACTGCCCTAGAGCCTGGGACATAGCAGTGAACAGACAGAGAGCAGCCCCTCCCTTCTGTAGCCCCCAAGCCAGTGAGGGGCACAGGCAGGAACAGGGACCACAACACAGAAAAGCTGGAGGGTGTC
AGGAGGTGATCAGGCTCTCGGGGAGGGAGAAGGGGTGGGGAGTGTGACTGGAGGAGACATCCTGCAGAAGGTGGGAGTGAGCAAACACCTGCGCAGGGGAGGGGAGGGCCTGCGGCACCTGGGGGAGCA
GAGGGAACAGCATCTGGCCAGGCCTGGGAGGAGGGGCCTAGAGGGCGTCAGGAGCAGAGAGGAGGTTGCCTGGCTGGAGTGAAGGATCGGGGCAGGGTGCGAGAGGGAACAAAGGACCCCTCCTGCAGGG
CCTCACCTGGGCCACAGGAGGACACTGCTTTTCCTCTGAGGAGTCAGGAACTGTGGATGGTGCTGGACAGAAGCAGGACAGGGCCTGGCTCAGGTGTCCAGAGGCTGCGCTGGCCTCCTATGGGATCAGA
CTGCAGGGAGGGAGGGCAGCAGGGATGTGGAGGGAGTGATGATGGGGCTGACCTGGGGGTGGCTCCAGGCATTGTCCCCACCTGGGCCCTTACCCAGCCTCCCTCACAGGCTCCTGGCCCTCAGTCTCTC
CCCTCCACTCCATTCTCCACCTACCCACAGTGGGTCATTCTGATCACCGAACTGACCATGCCAGCCCTGCCGATGGTCCTCCATGGCTCCCTAGTGCCCTGGAGAGGAGGTGTCTAGTCAGAGAGTAGTC
CTGGAAGGTGGCCTCTGTGAGGAGCCACGGGGACAGCATCCTGCAGATGGTCCTGGCCCTTGTCCCACCGACCTGTCTACAAGGACTGTCCTCGTGGACCCTCCCCTCTGCACAGGAGCTGGACCCTGAA
GTCCCTTCCTACCGGCCAGGACTGGAGCCCCTACCCCTCTGTTGGAATCCCTGCCCACCTTCTTCTGGAAGTCGGCTCTGGAGACATTTCTCTCTTCTTCCAAAGCTGGGAACTGCTATCTGTTATCTGC
CTGTCCAGGTCTGAAAGATAGGATTGCCCAGGCAGAAACTGGGACTGACCTATCTCACTCTCTCCCTGCTTTTACCCTTAGGGTGATTCTGGGGGCCCACTTGTCTGTAATGGTGTGCTTCAAGGTATCA
CGTCATGGGGCAGTGAACCATGTGCCCTGCCCGAAAGGCCTTCCCTGTACACCAAGGTGGTGCATTACCGGAAGTGGATCAAGGACACCATCGTGGCCAACCCCTGAGCACCCCTATCAAGTCCCTATTG
TAGTAAACTTGGAACCTTGGAAATGACCAGGCCAAGACTCAAGCCTCCCCAGTTCTACTGACCTTTGTCCTTAGGTGTGAGGTCCAGGGTTGCTAGGAAAAGAAATCAGCAGACACAGGTGTAGACCAGG
TGTTTCTTAAATGGTGTAATTTTGTCCTCTCTGTGTCCTGGGGAATACTGGCCATGCCTGGAGACATATCACTCAATTTCTCTGAGGACACAGTTAGGATGGGGTGTCTGTGTTATTTGTGGGATACAGA
GATGAAAGAGGGGTGGGATCCACACTGAGAGAGTGGAGAGTGACATGTGCTGGACACTGTCCATGAAGCACTGAGCAGAAGCTGGAGGCACAACGCACCAGACACTCACAGCAAGGATGGAGCTGAAAAC
ATAACCCACTCTGTCCTGGAGGCACTGGGAAGCCTAGAGAAGGCTGTGAGCCAAGGAGGGAGGGTCTTCCTTTGGCATGGGATGGGGATGAAGTAAGGAGAGGGACTGGACCCCCTGGAAGCTGATTCAC
TATGGGGGGAGGTGTATTGAAGTCCTCCAGACAACCCTCAGATTTGATGATTTCCTAGTAGAACTCACAGAAATAAAGAGCTCTTATACTGTGGTTTATTCTGGTTTGTTACATTGACAGGAGACACACT
GAAATCAGCAAAGGAAACAGGCATCTAAGTGGGGATGTGAAGAAAACAGGGAAAATCTTTCAGTTGTTTTCTCCCAGTGGGGTGTTGTGGACAGCACTTAAATCACACAGAAGTGATGTGTGACCTTGTG
TATGAAGTATTTCCAACTAAGGAAGCTCACCTGAGCCTTAGTGTCCAGAGTTCTTATTGGGGGTCTGTAGGATAGGCATGGGGTACTGGAATAGCTGACCTTAACTTCTCAGACCTGAGGTTCCCAAGAG
TTCAAGCAGATACAGCATGGCCTAGAGCCTCAGATGTACAAAAACAGGCATTCATCATGAATCGCACTGTTAGCATGAATCATCTGGCACGGCCCAAGGCCCCAGGTATACCAAGGCACTTGGGCCGAAT
GTTCCAAGGGATTAAATGTCATCTCCCAGGAGTTATTCAAGGGTGAGCCCTGTACTTGGAACGTTCAGGCTTTGAGCAGTGCAGGGCTGCTGAGTCAACCTTTTACTGTACAGGGGGGTGAGGGAAAGGG
```

Outline

- Exon-intron structure of genes
- Models of gene grammar
  - Example: Genscan
- Models of exon-intron sequence
- Integrating intrinsic, extrinsic information
  - Example: GenomeScan
- The RNA splicing code

Central Dogma



DNA

1:1

ACCGGACCGATGCGACTGCCCGAGGACTAGATAT
TGGCCTGGCTACGCTGACGGGCTCCTGATCTATA

RNA

1:1

*

GACCGAUGCGACUGCCCGAGGACUAGA

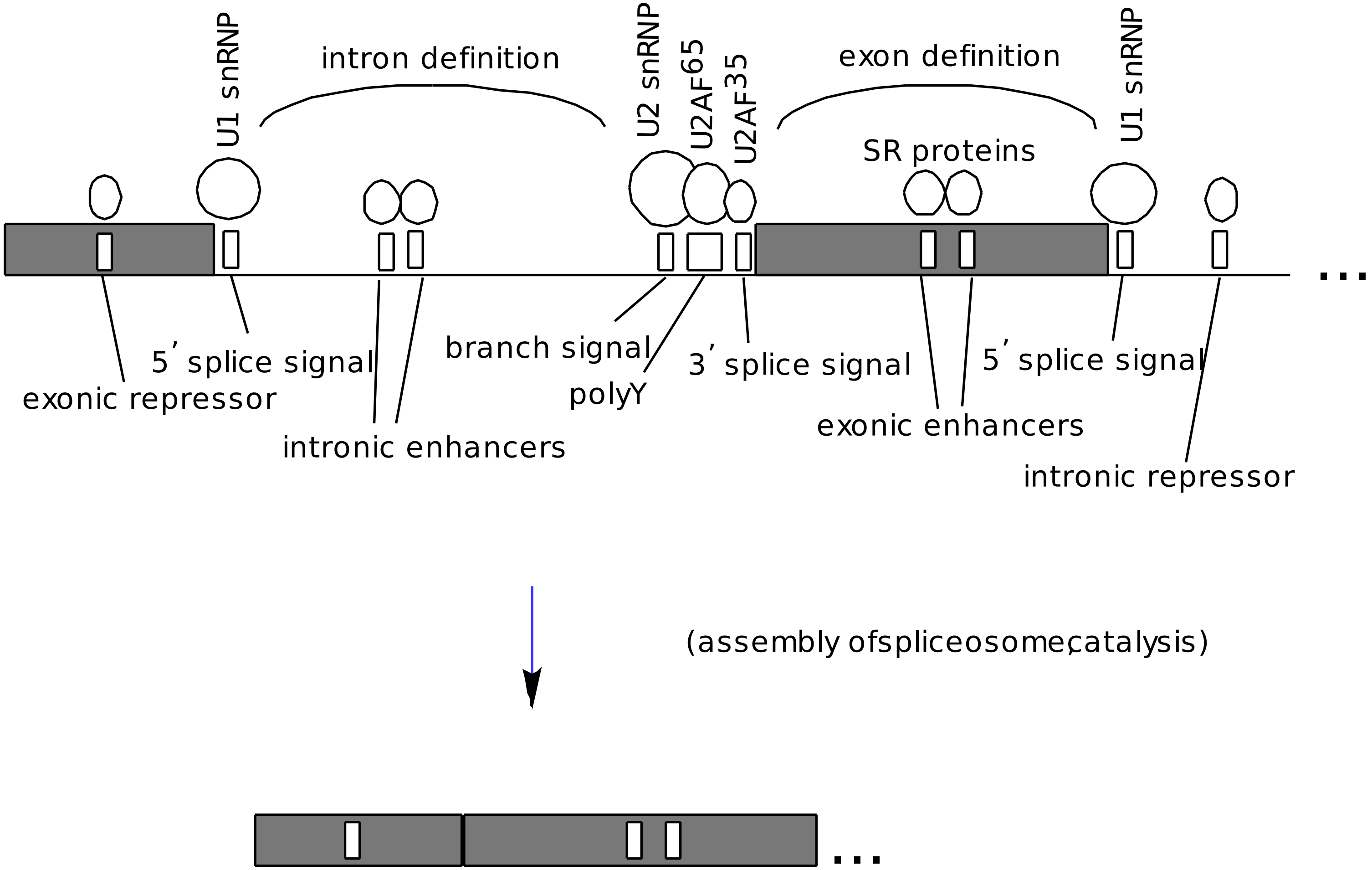
M R L P E D

Protein

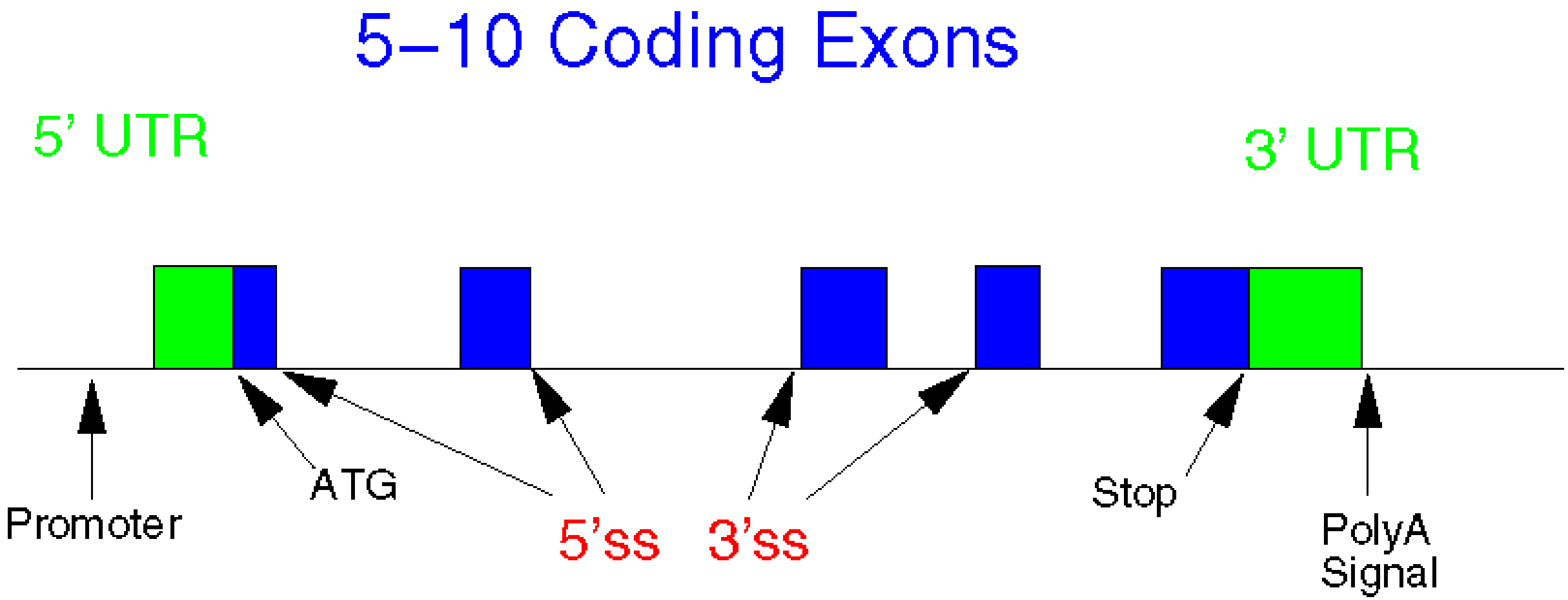
3:1

MRLPED

Pre-mRNA Splicing

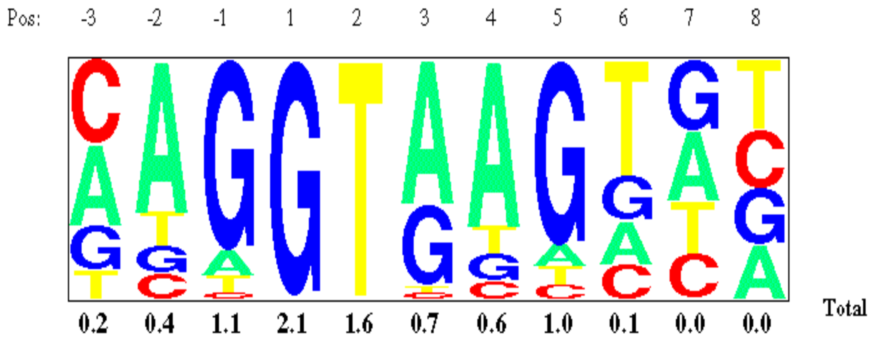


Structure of a Typical Human Gene

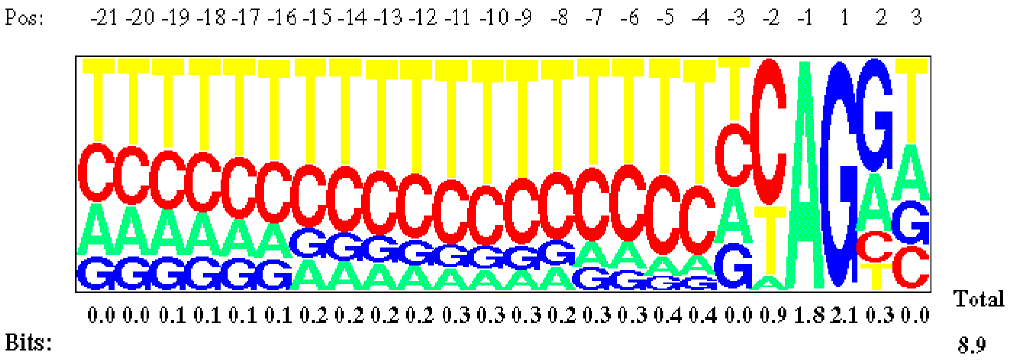


Human Splice Signal Motifs

5' splice signal



3' splice signal



Molecular Codes

Genetic Code

mRNA → Protein

Gene Finding Code

Genomic DNA → Genes

RNA Splicing Code

pre–mRNA → mRNA

GENSCAN – Basic Idea

Model of what a human gene "looks like" in terms of:

exon–intron structure

sequence composition

In principle, given a sequence, assign a probability to every possible gene structure compatible w/ sequence

In practice, use a Dynamic Programming algorithm to determine the most probable gene structure(s).

Gene Features Modeled by Genscan

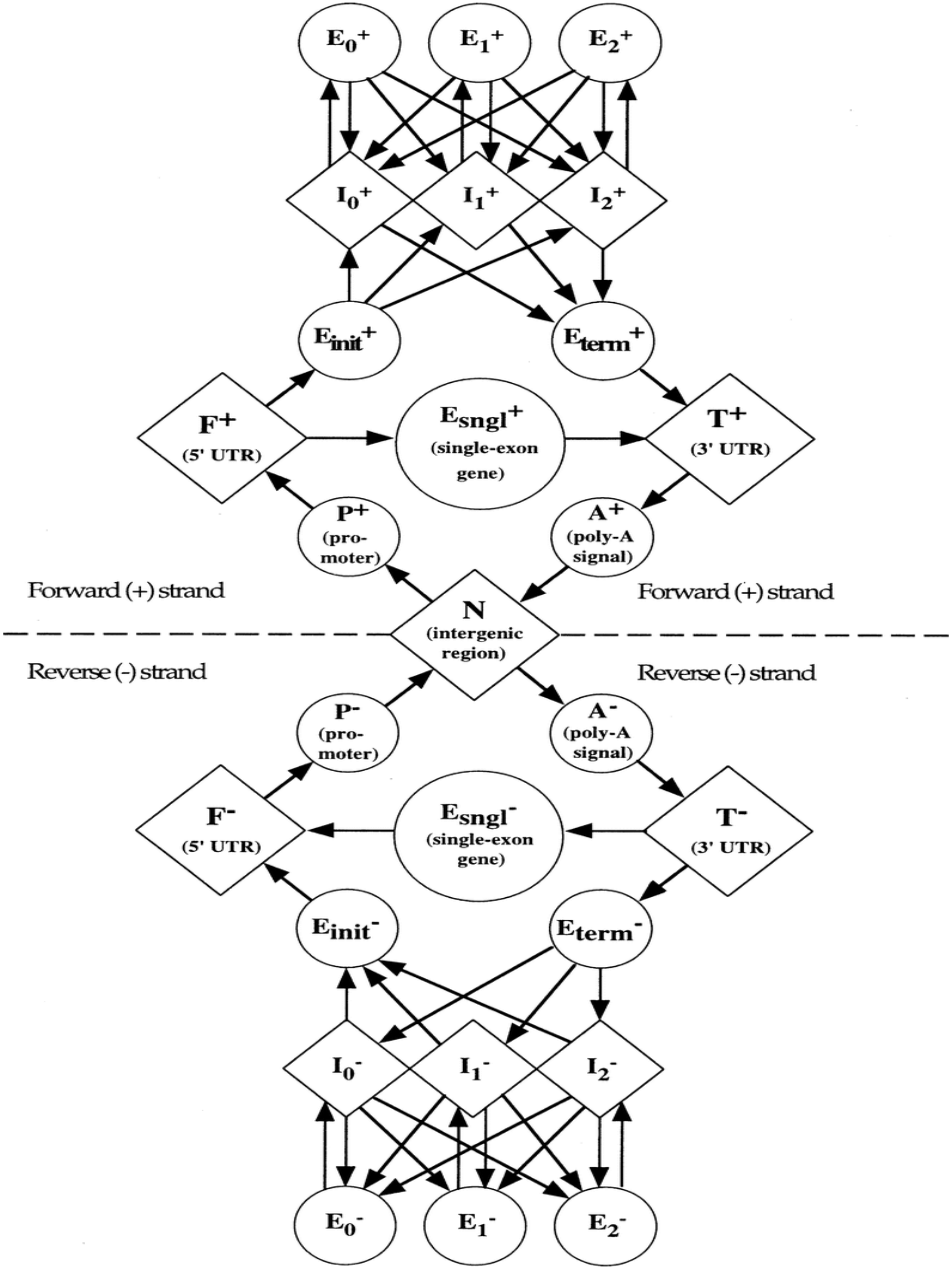
Semi–Markov HMM Model of human gene structure
and composition

Features modeled:

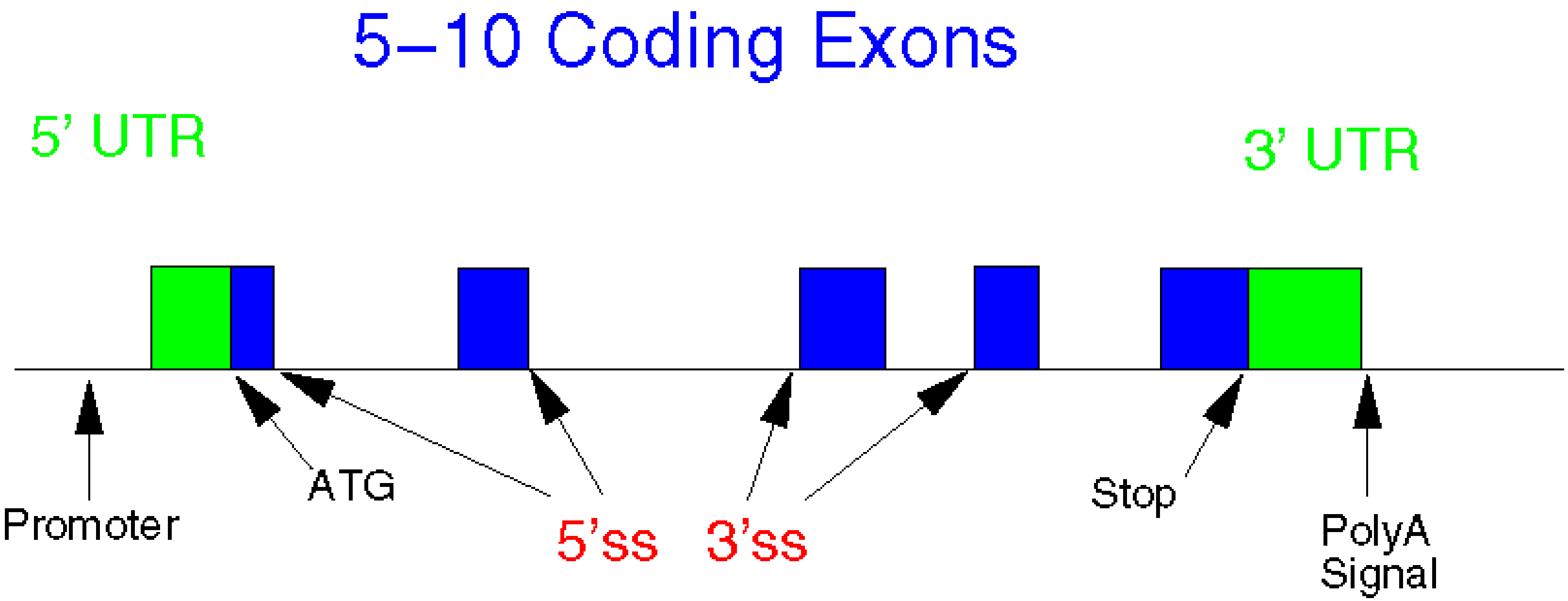
- Hexamer composition of exons/introns
- Extended 5′ and 3′ splice signals
- Reading frame consistency of exons
- Exon/intron length distributions
- Promoter and polyA signals
- Isochore differences

C. Burge & S. Karlin, 1997, 1998

Genscan HSMM

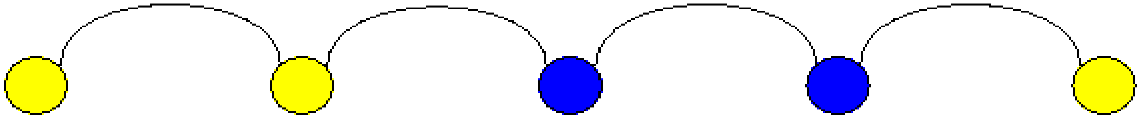


Structure of a Typical Human Gene



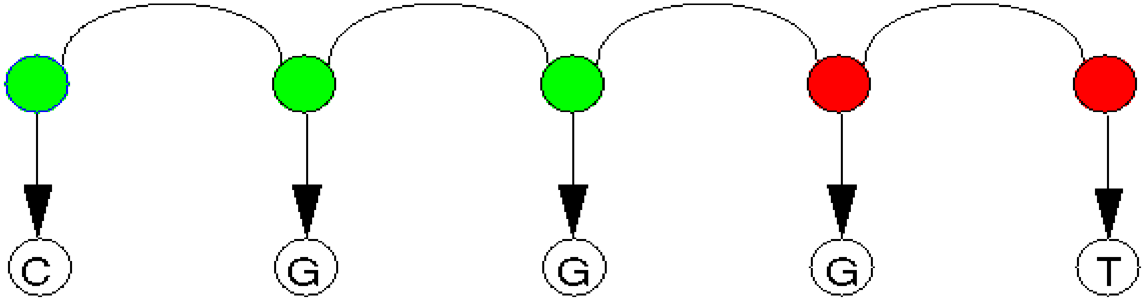
Markov and Hidden Markov Models

Markov



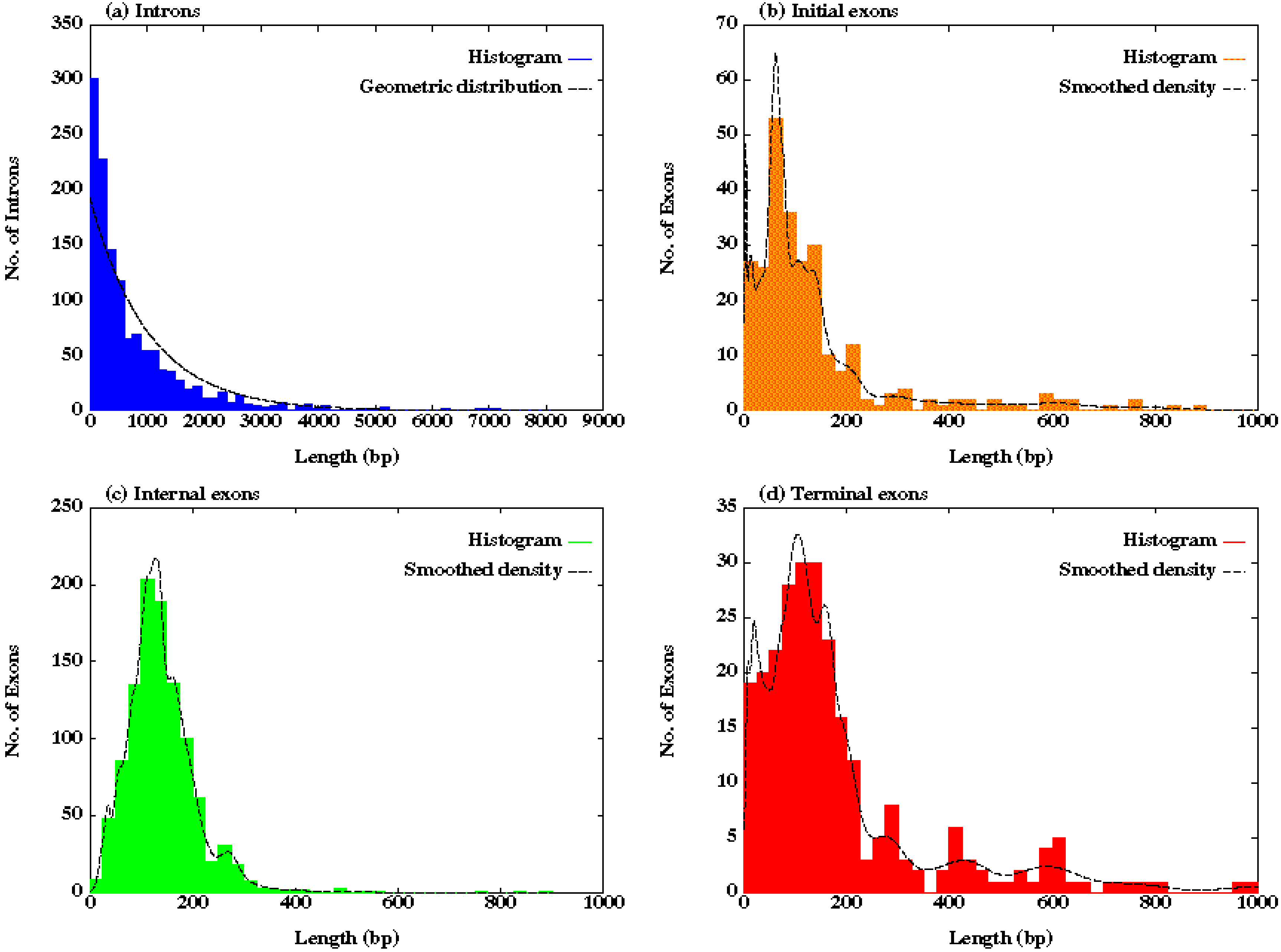
State **n+1** depends on state **n**, but not on previous states

Hidden Markov



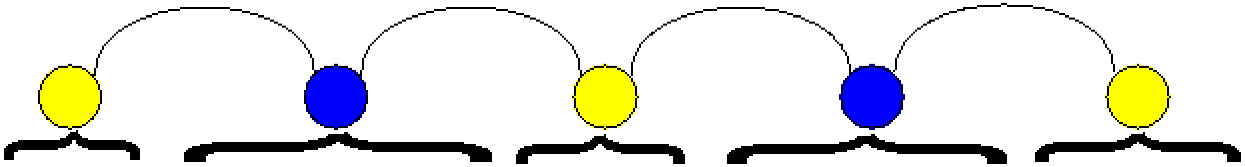
Hidden states have Markov dependence; observable states generated from hidden

Length distributions of human introns and initial, internal and terminal exons



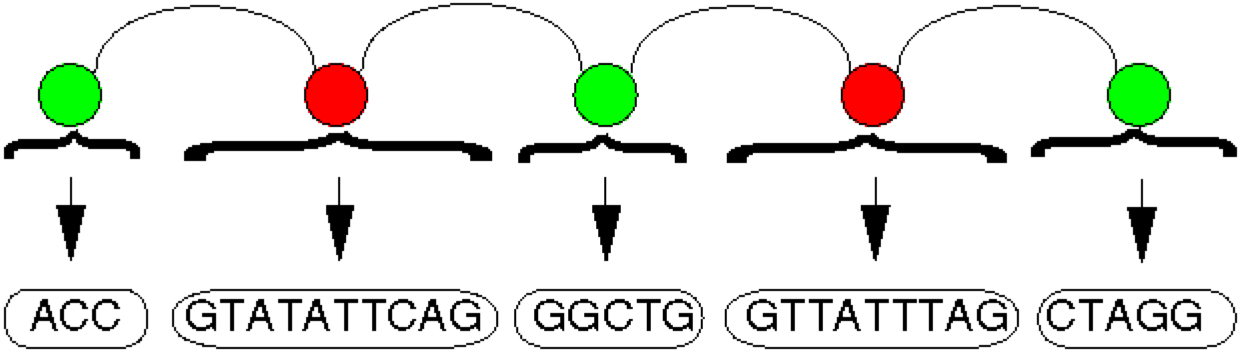
Semi–Markov and Hidden Semi–Markov

Semi–Markov



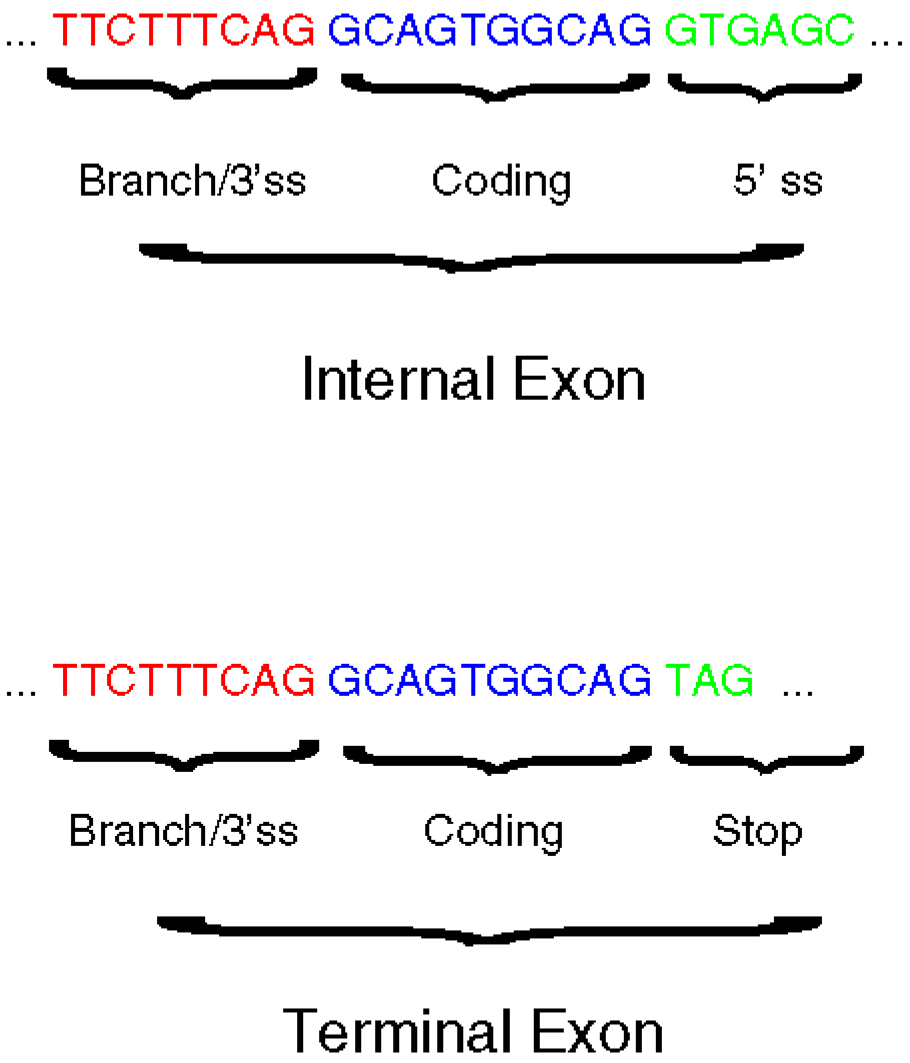
States have Markov dependence;
each state has an associated length

Hidden Semi–Markov



Hidden states semi–Markov;
observable generated from hidden

Sample Exon Models



... TTCTTTCAG GCAGTGGCAG GTGAGC ...

Branch/3'ss Coding 5' ss

Internal Exon

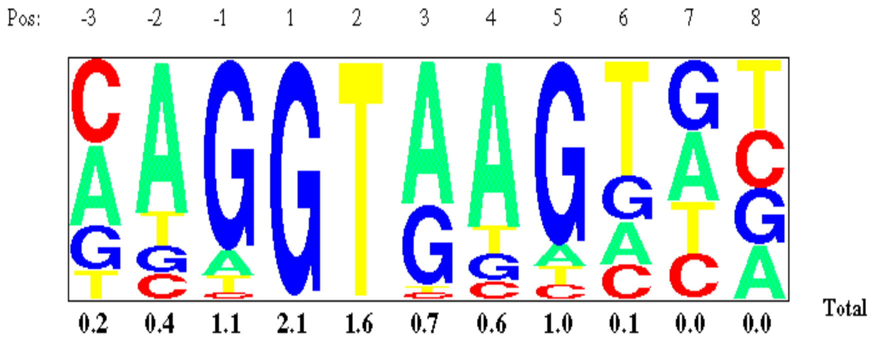
... TTCTTTCAG GCAGTGGCAG TAG ...

Branch/3'ss Coding Stop

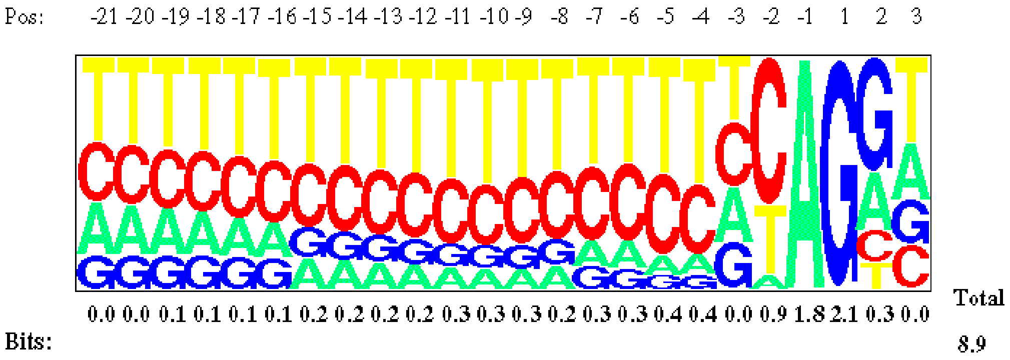
Terminal Exon

Human Splice Signal Motifs

5' splice signal

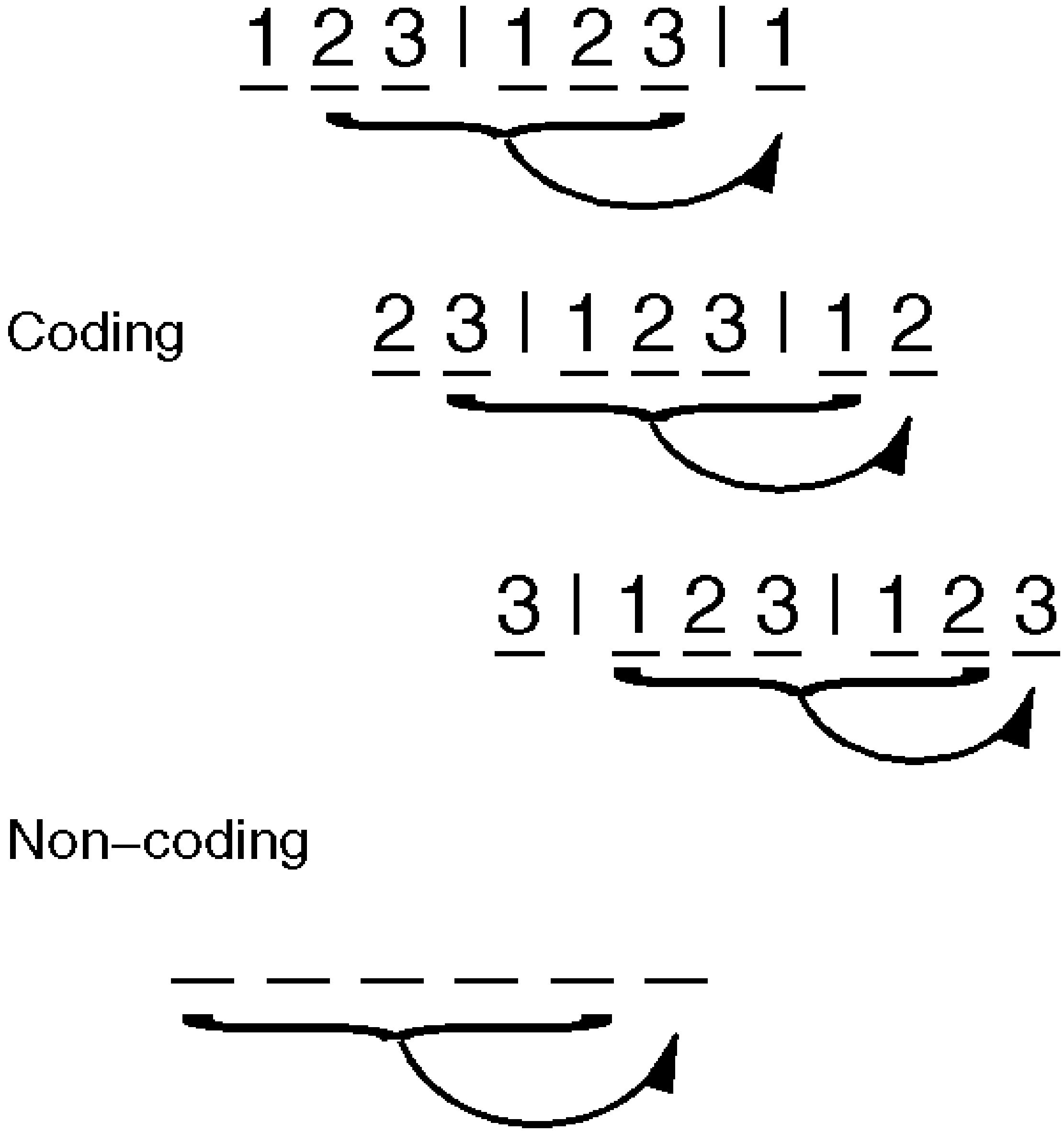


3' splice signal



http://genes.mit.edu/pictogram.html

Models of Coding and Non–Coding DNA



Coding

Non–coding

Viterbi Algorithm – Basic Idea

Goal: Maximize $P(\phi_i,S)$

(Find optimal ’parse’ of sequence)

Approach:

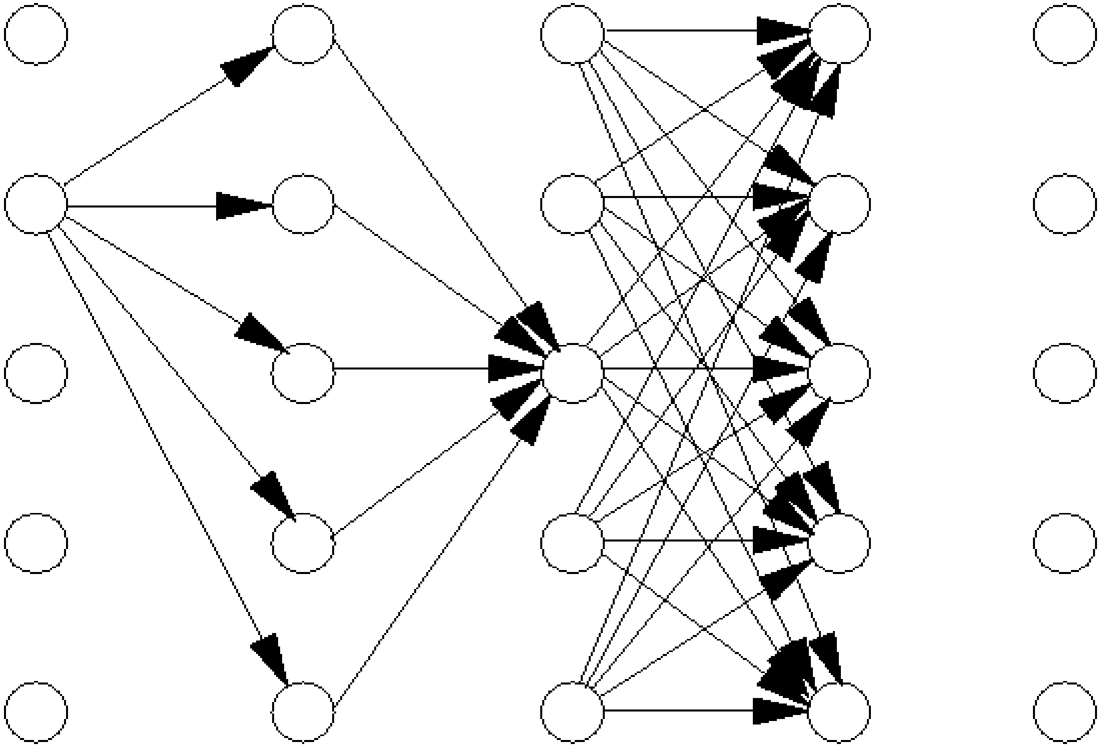
Define variables which store Pr of optimal parse of subsequence up to pos. j ending in each possible state

Solve recursively

Forward/backward algorithms are similar but calculate *sum* of Pr of all parses

Viterbi, A J (1967), Forney, G D (1973), Rabiner, R (1989).

Viterbi Algorithm in HMM Case

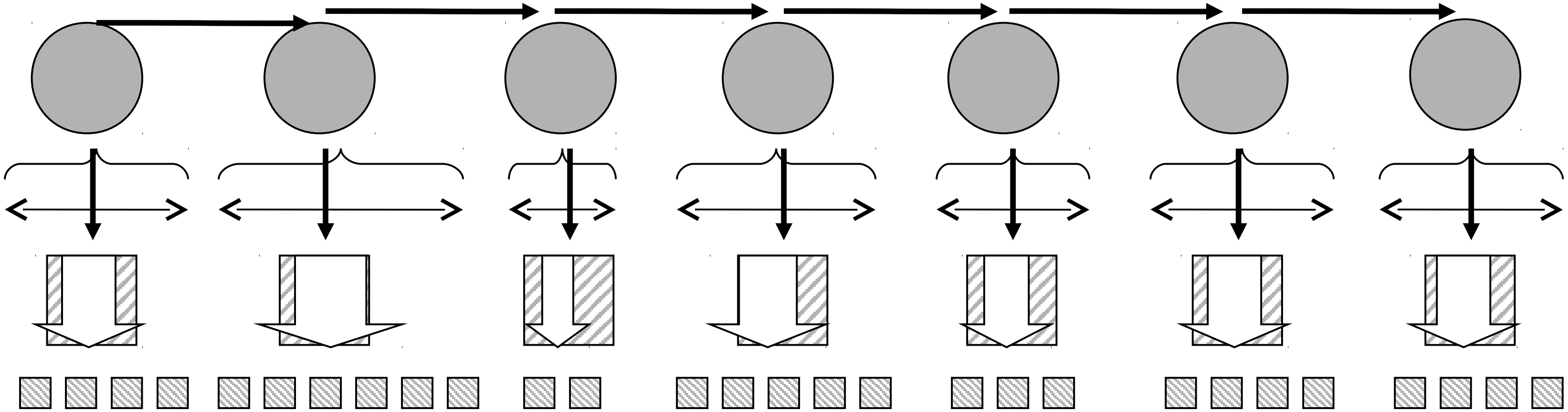


Optimal paths derivable from "single step" recursion.

For N state model, seq length L : $O(N^2L)$

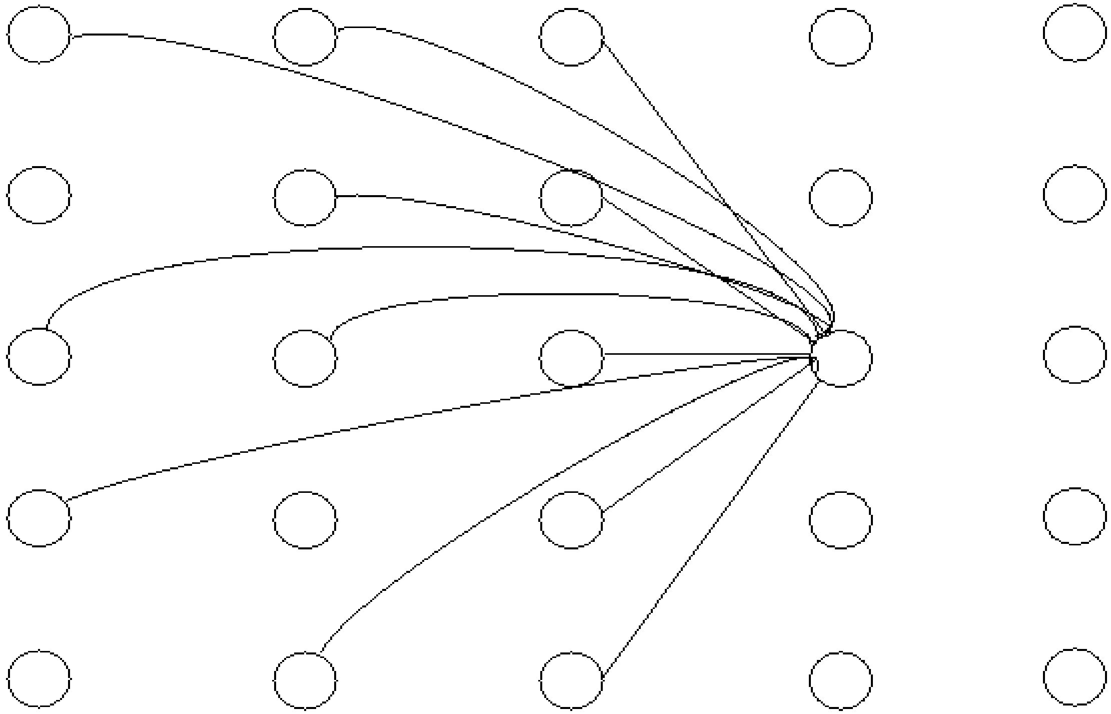
Viterbi, A J (1967), Forney, G D (1973), Rabiner, R (1989).

Semi-Markov HMM Model



Viterbi Algorithm for

(Hidden) "Semi–Markov" Model



Paths involve "jumps" as well as "steps":

For **N** state model, seq length **L** : $O(N^2L^3)$

Howard, RA (1971) "Dynamic Probabilistic Systems Volume II:

Semi–Markov and Decision Processes." See also Rabiner (1989).

How Well did Genscan Work on Chromosome 22?

Annotated genes:

94% predicted at least partially: ~6% of genes missed

Annotated exons:

84% predicted at least partially: ~16% of exons missed

Predicted exons:

Approx 30% more than annotated

How many of these are real?

Statistics from I. Dunham et al. Nature 402, 489–95, 1999

Genes on Human Chromosome 22

Class	No. of Genes	
Known	247	545
Related	150	
EST–supported	148	
Pseudo	134	
Predicted novel	325	100

Estimated ~45K genes in genome

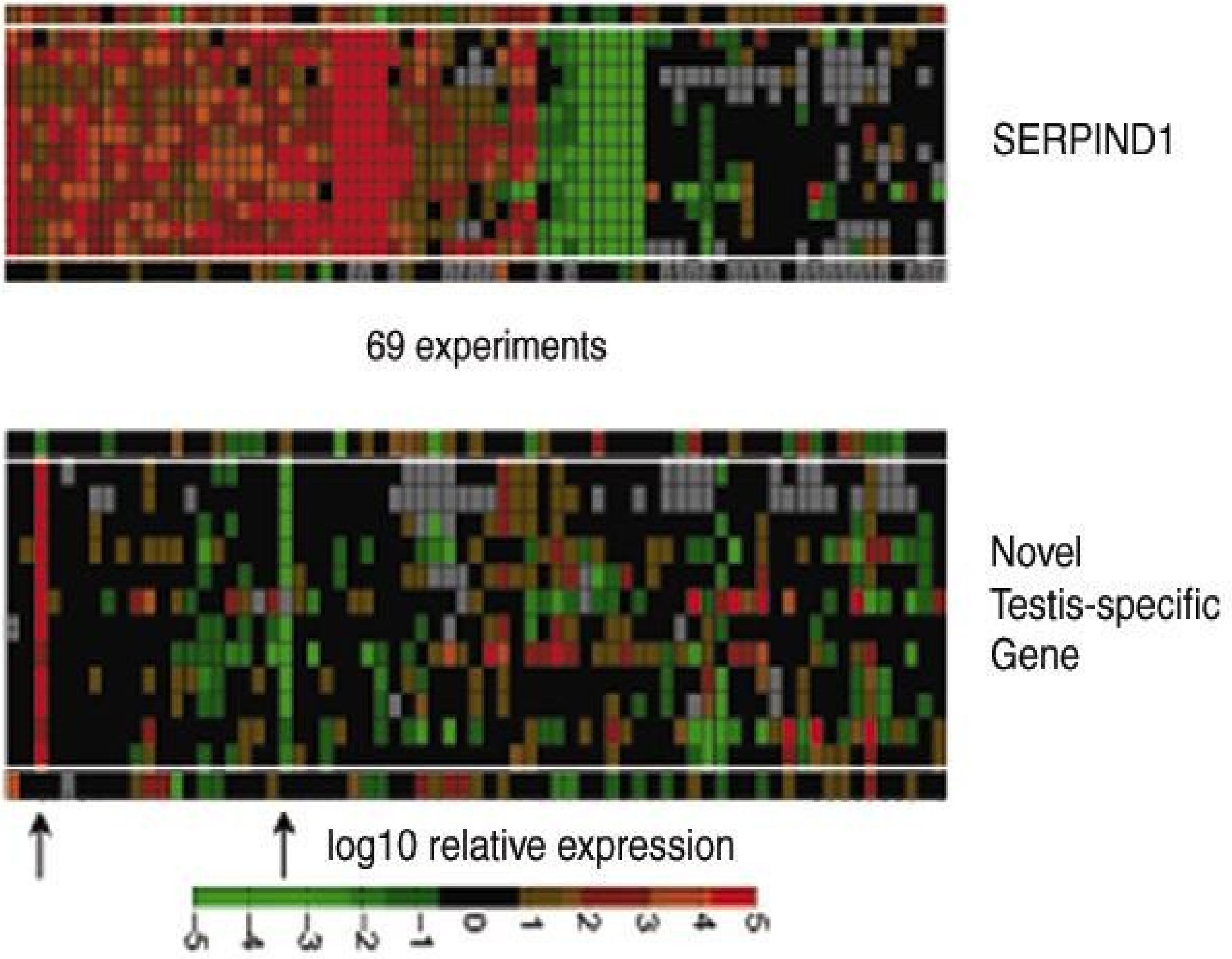
Dunham, I. et al. Nature 1999

Genome Scale Gene Finding Strategies

Strategy	Based on	Examples
Ab initio prediction	Models of gene structure/comp	Genscan, GRAIL GenLang, hmmgene
Microarray	Hybridization	Exon-scanning array
Gene inference	Homology	GenomeScan
Genomic:genomic alignment	Homology	ExoFish GLASS/Rosetta
DNA:protein alignment	Homology	GeneWise
cDNA sequencing	Sequencing	RIKEN

C. Burge Nature Genet. 27, 5-7, 2001

EXON-SCANNING ARRAYS



D. Shoemaker et al. Nature 409, 922-7, 2001

(C. Burge Nature Genet. 27, 5-7, 2001)

ExoFish

Homo sapiens

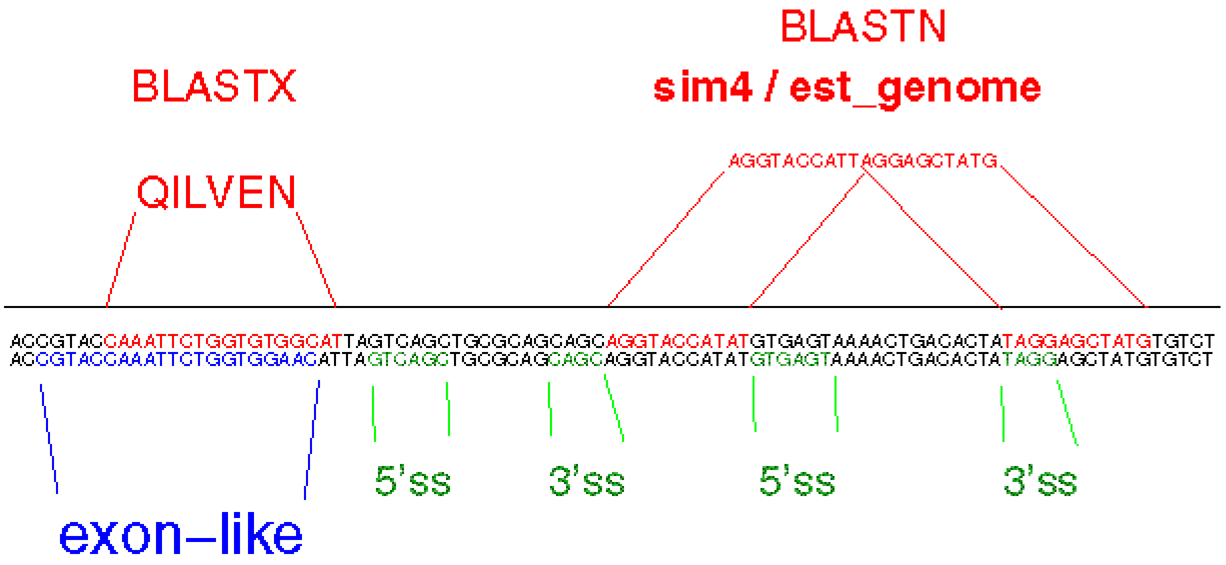
⇔

Tetraodon nigroviridis

Roest Crollius et al., Nature Genet., 2000

Extrinsic & Intrinsic Information about Gene Locations

Extrinisic



BLASTN

sim4 / est_genome

BLASTX

QILVEN

AGGTACCATTAGGAGCTATG

ACGGTACGCAAATTCTGGTGTGGCATTAGTCAGCTGCGCAGCAGCAGGTACCATATGTGAGTAAAACTGACACTATAGGAGCTATGTGTCT

ACGGTACCAAATTCTGGTGGAACATTAGTCAGCTGCGCAGCAGCAGGTACCATATGTGAGTAAAACTGACACTATAGGAGCTATGTGTCT

5'ss 3'ss 5'ss 3'ss

exon–like

Intrinsic

GenomeScan Objectives

- Combine probabilistic ‘extrinsic’ information (BLAST hits) with a probabilistic model of gene structure/composition

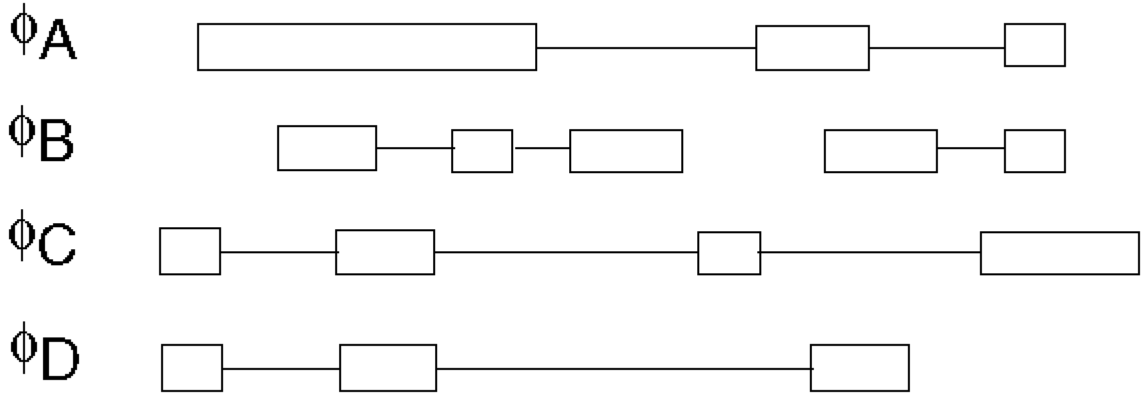
- Make method efficient and reliable enough to run on an entire vertebrate genome without human supervision

- Focus on ‘typical case’ when homologous but not identical proteins are available.

Genscan:

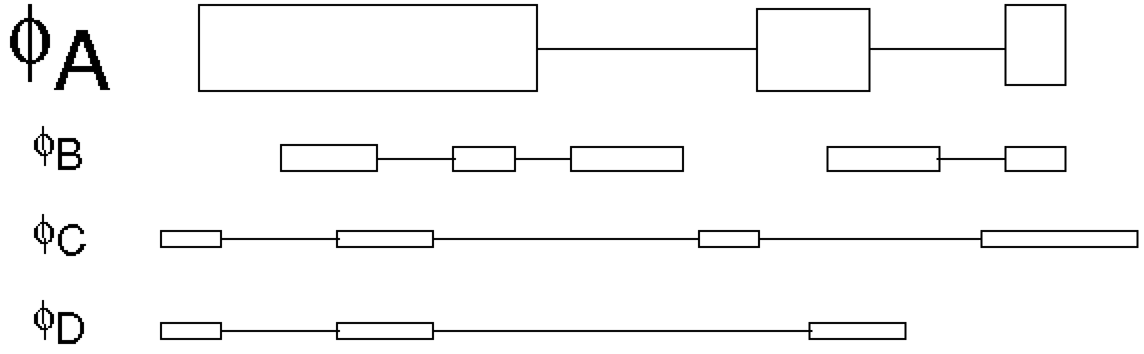
use likelihood to choose among possible gene structures

Prior: $P(\phi_A) \cong P(\phi_B) \cong P(\phi_C) \cong P(\phi_D)$



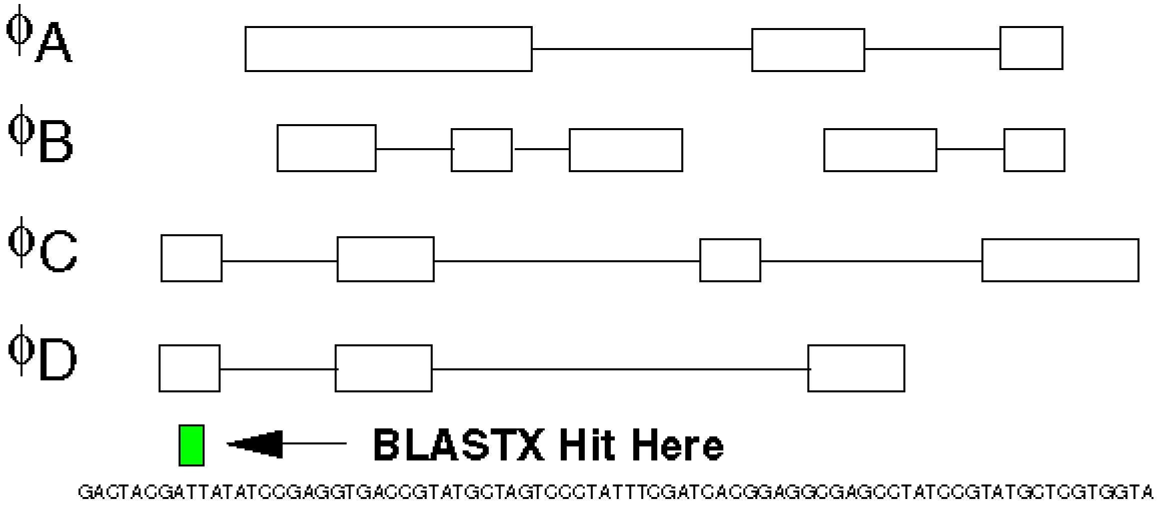
GACTACGATTATATCCGAGGTGACCGTATGCTAGTCCCTATTTCGATCACGGAGGCGAGCCTATCCGTATGCTCGTGGTA

Joint: $P(\phi_A,S) = P(\phi_A) \times P(S|\phi_A)$ (Prior x Likelihood)

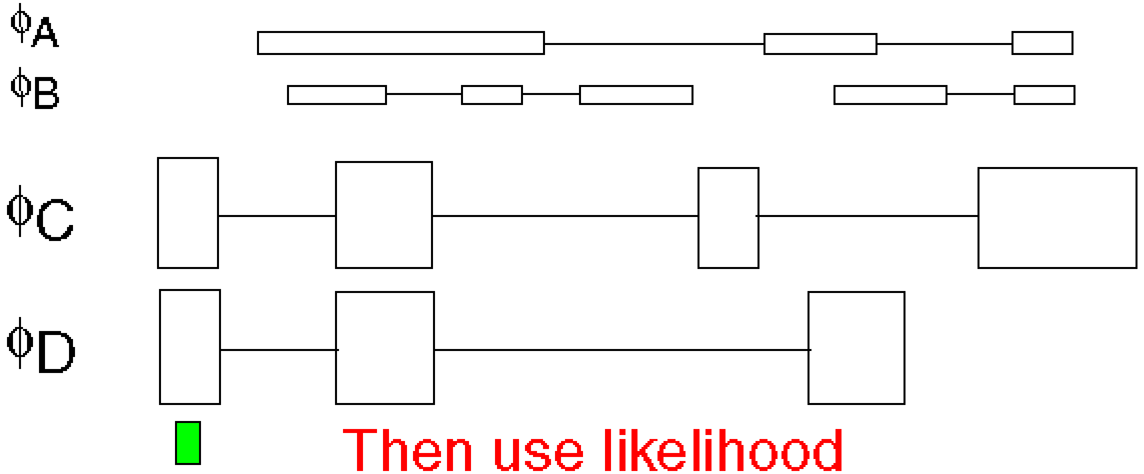


Using similarity information in GenomeScan

Prior: $P(\phi_A) \cong P(\phi_B) \cong P(\phi_C) \cong P(\phi_D)$


ϕA
ϕB
ϕC
ϕD
BLASTX Hit Here
GACTACGGATTATATCCGAGGTGACCGTATGCTAGTCCCTATTTCGATCACGGAGGCGAGCCTATCCGTATGCTCGTGGTA

New Prior: $P(\phi_A) \cong P(\phi_B) << P(\phi_C) \cong P(\phi_D)$



Then use likelihood

When Good Alignments Go Bad...

Output of BLASTX:

Score = 129 bits (321), Expect = 2e−29

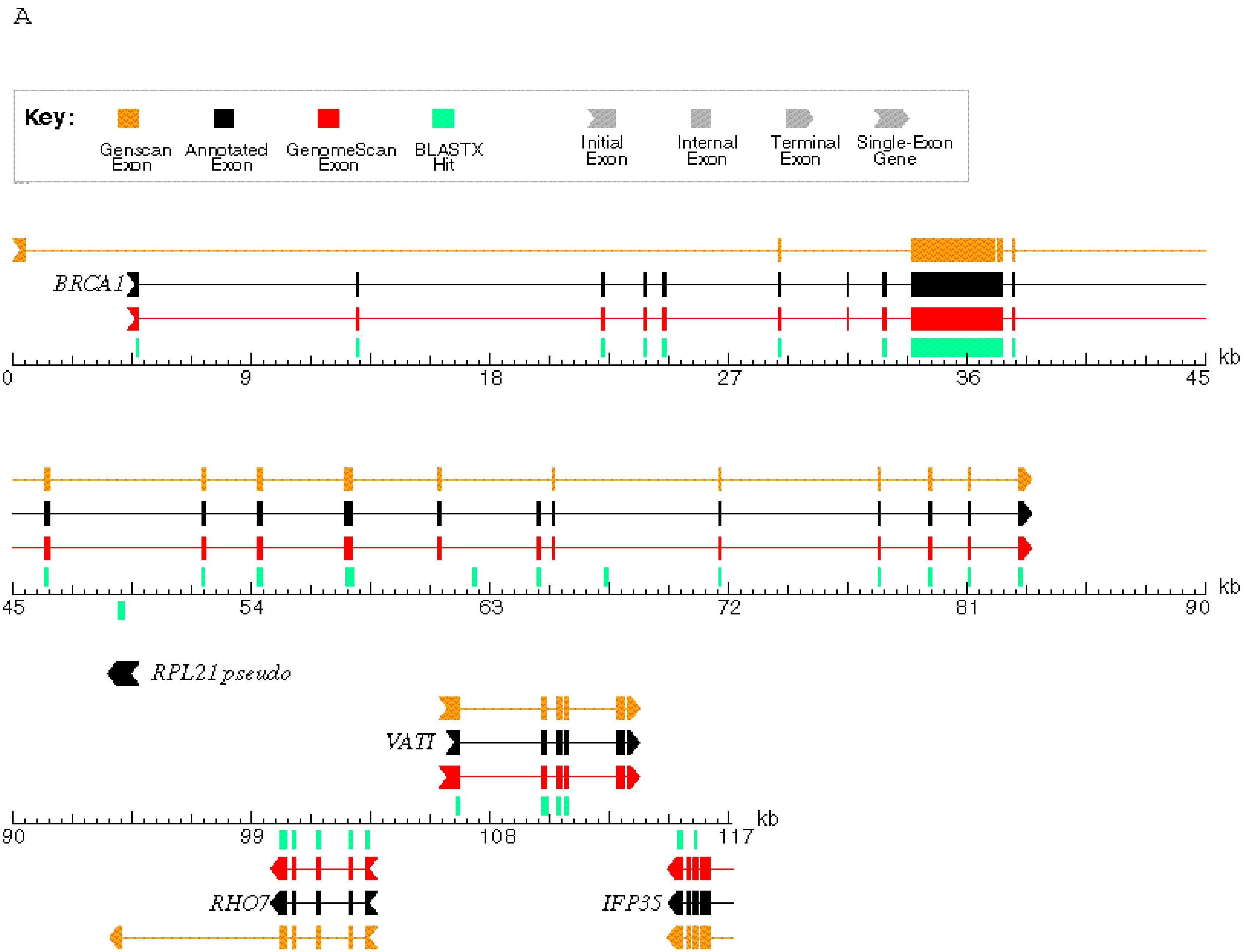
```
DGGWGWIVLFGCFVITGFSYAFPKAVSVYFKELMKDFHVGYSDTA
DGGWGW VLFGCF+ITGFSYAFPKAVSV+FKELM +F +GYSDTA
DGGWGWAVLFGCFIITGFSYAFPKAVSVFFKELMHEFGIGYSDTA


WISSIMLAMLYGTGDAWIYFPLPNPCLPCPARVPNRVPVGMLNGL
WISSI+LAMLYGTG      PL   C  C  R   R PV ++ GL
WISSILLAMLYGTG------PL---CSMCVNRFGCR-PVMLVGGL
```

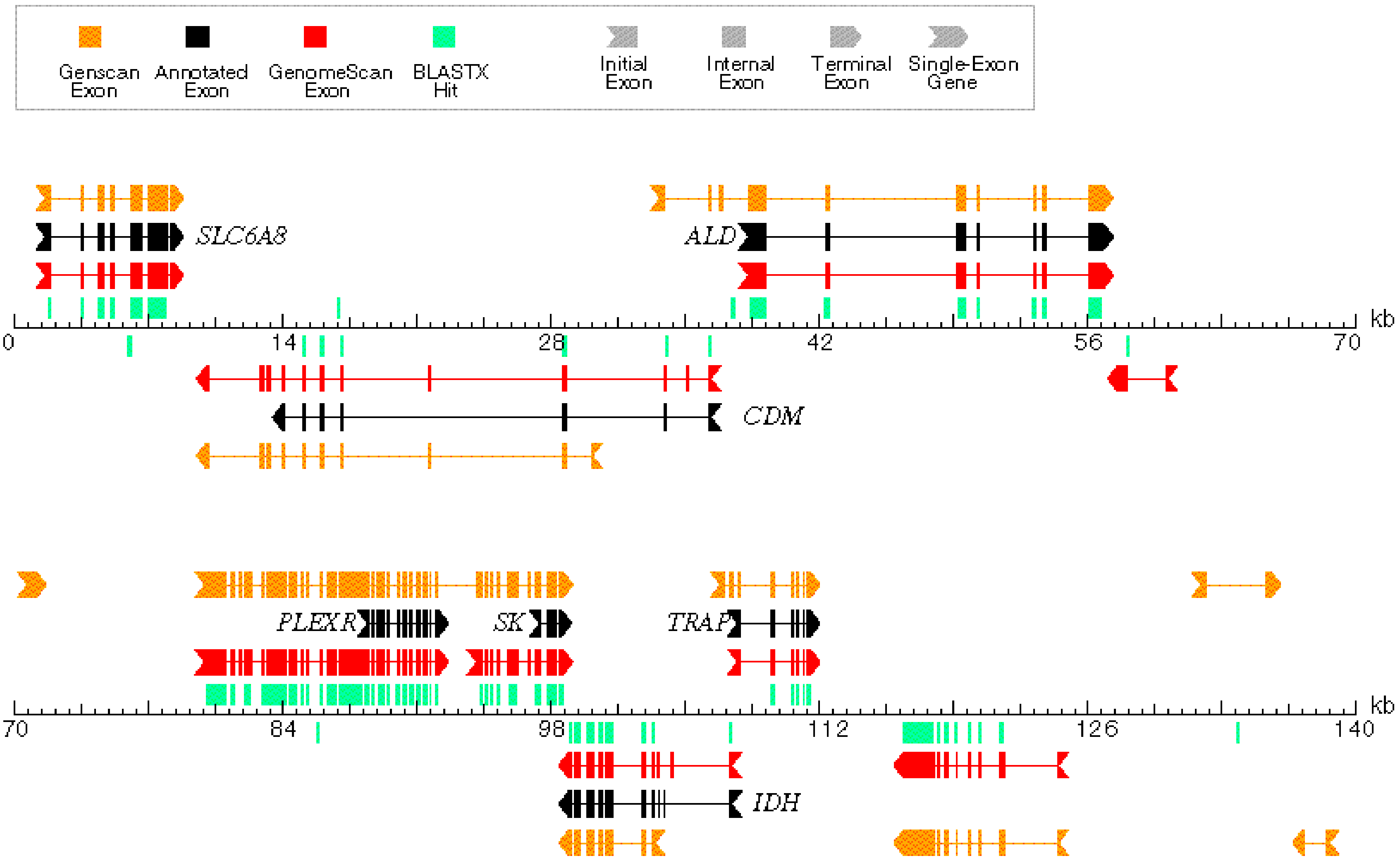
Questionable Region

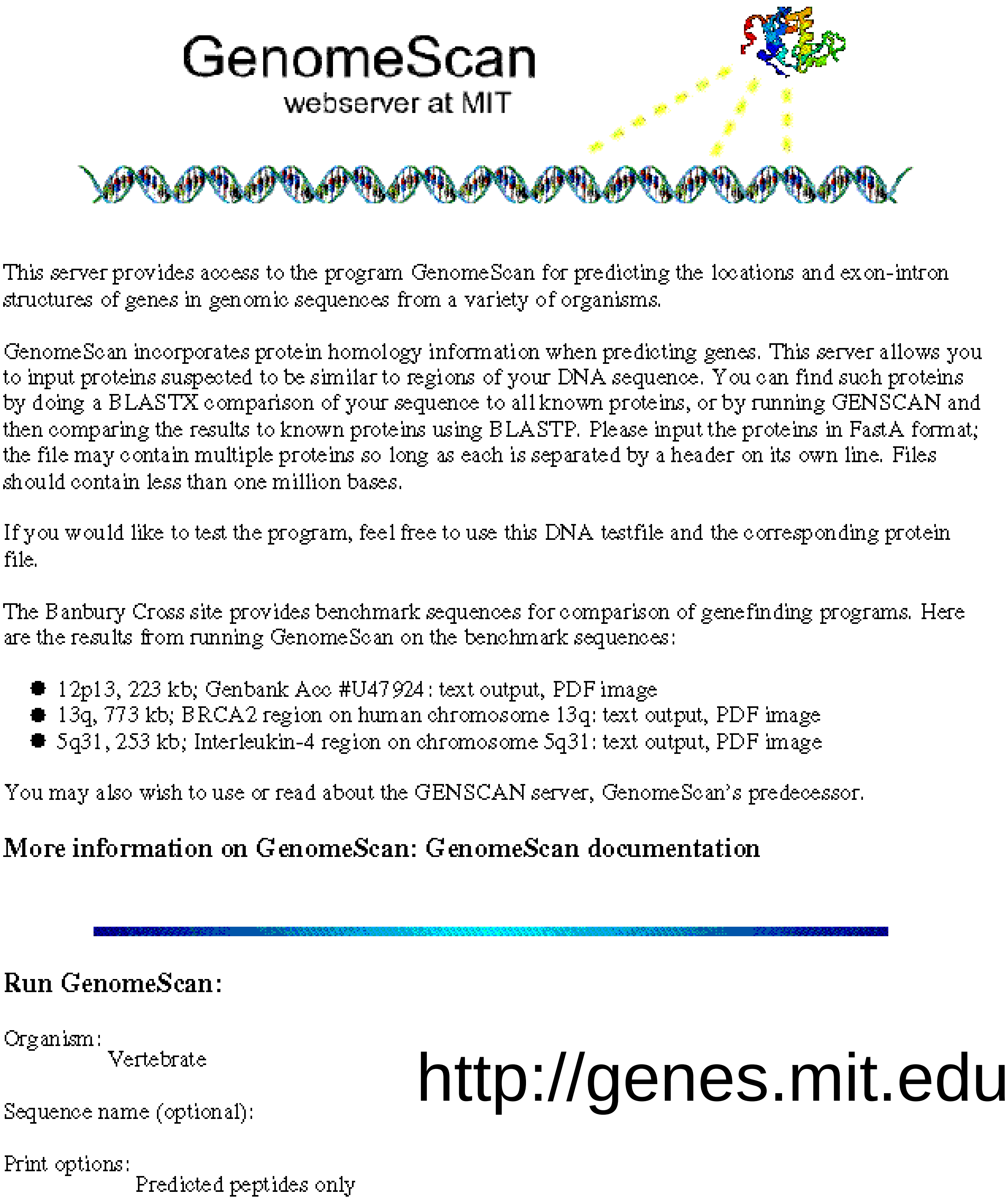
Solution

Post−process BLAST hits w/ steepest slope heuristic


A
Key:
Genscan Exon
Annotated Exon
GenomeScan Exon
BLASTX Hit
Initial Exon
Internal Exon
Terminal Exon
Single-Exon Gene
BRCA1
0
9
18
27
36
45
kb
45
54
63
72
81
90
kb
RPL21 pseudo
VATI
90
99
108
117
kb
RHO7
IFP35

B


Genscan Exon
Annotated Exon
GenomeScan Exon
BLASTX Hit
Initial Exon
Internal Exon
Terminal Exon
Single-Exon Gene
SLC6A8
ALD
CDM
0
14
28
42
56
70
kb
PLEXR
SK
TRAP
IDH
70
84
98
112
126
140
kb



GenomeScan
webserver at MIT

This server provides access to the program GenomeScan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

GenomeScan incorporates protein homology information when predicting genes. This server allows you to input proteins suspected to be similar to regions of your DNA sequence. You can find such proteins by doing a BLASTX comparison of your sequence to all known proteins, or by running GENSCAN and then comparing the results to known proteins using BLASTP. Please input the proteins in FastA format; the file may contain multiple proteins so long as each is separated by a header on its own line. Files should contain less than one million bases.

If you would like to test the program, feel free to use this DNA testfile and the corresponding protein file.

The Banbury Cross site provides benchmark sequences for comparison of genefinding programs. Here are the results from running GenomeScan on the benchmark sequences:

- 12p13, 223 kb; Genbank Acc #U47924: text output, PDF image
- 13q, 773 kb; BRCA2 region on human chromosome 13q: text output, PDF image
- 5q31, 253 kb; Interleukin-4 region on chromosome 5q31: text output, PDF image

You may also wish to use or read about the GENSCAN server, GenomeScan's predecessor.

More information on GenomeScan: GenomeScan documentation

Run GenomeScan:

Organism:
Vertebrate

Sequence name (optional):

Print options:
Predicted peptides only

http://genes.mit.edu/genomescan

Current Human Gene Annotation Efforts

- Ensembl [http://www.ensembl.org]

  Genscan (ab initio) + BLAST (homology) + GeneWise (protein:DNA alignment)

- NCBI [http://ncbi.nlm.nih.org]

  acembly (cDNA,EST alignments)

- Burge lab [http://genes.mit.edu/genomescan]

  GenomeScan (ab initio + protein sequence homology)

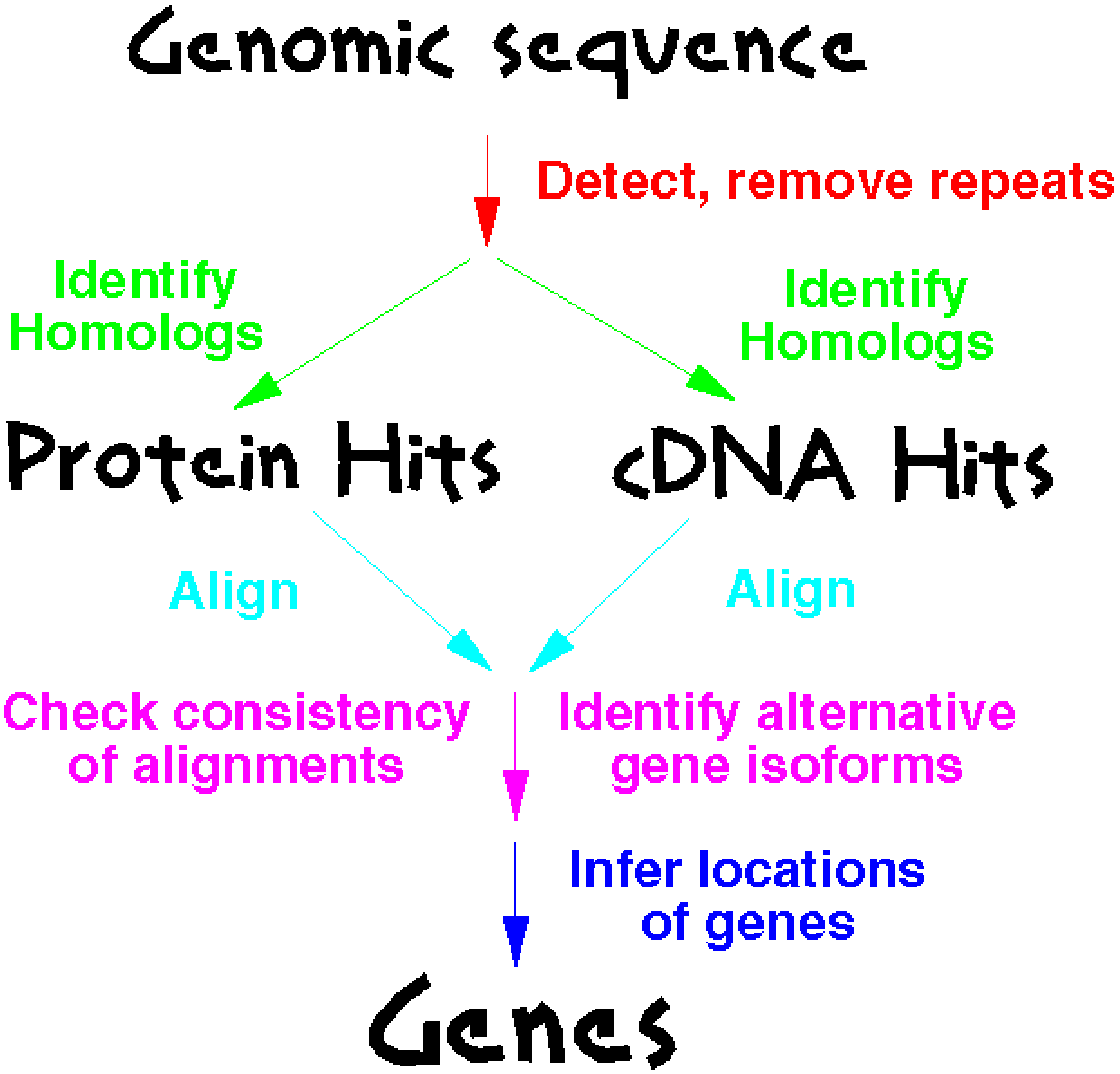
- Neomorphic/Affymetrix

  Genie (ab initio + EST)

- Celera

  Otto (???)

IGI (International Gene Index) / IPI (EBI)



Genomic sequence

Detect, remove repeats

Identify Homologs

Identify Homologs

Protein Hits

cDNA Hits

Align

Align

Check consistency of alignments

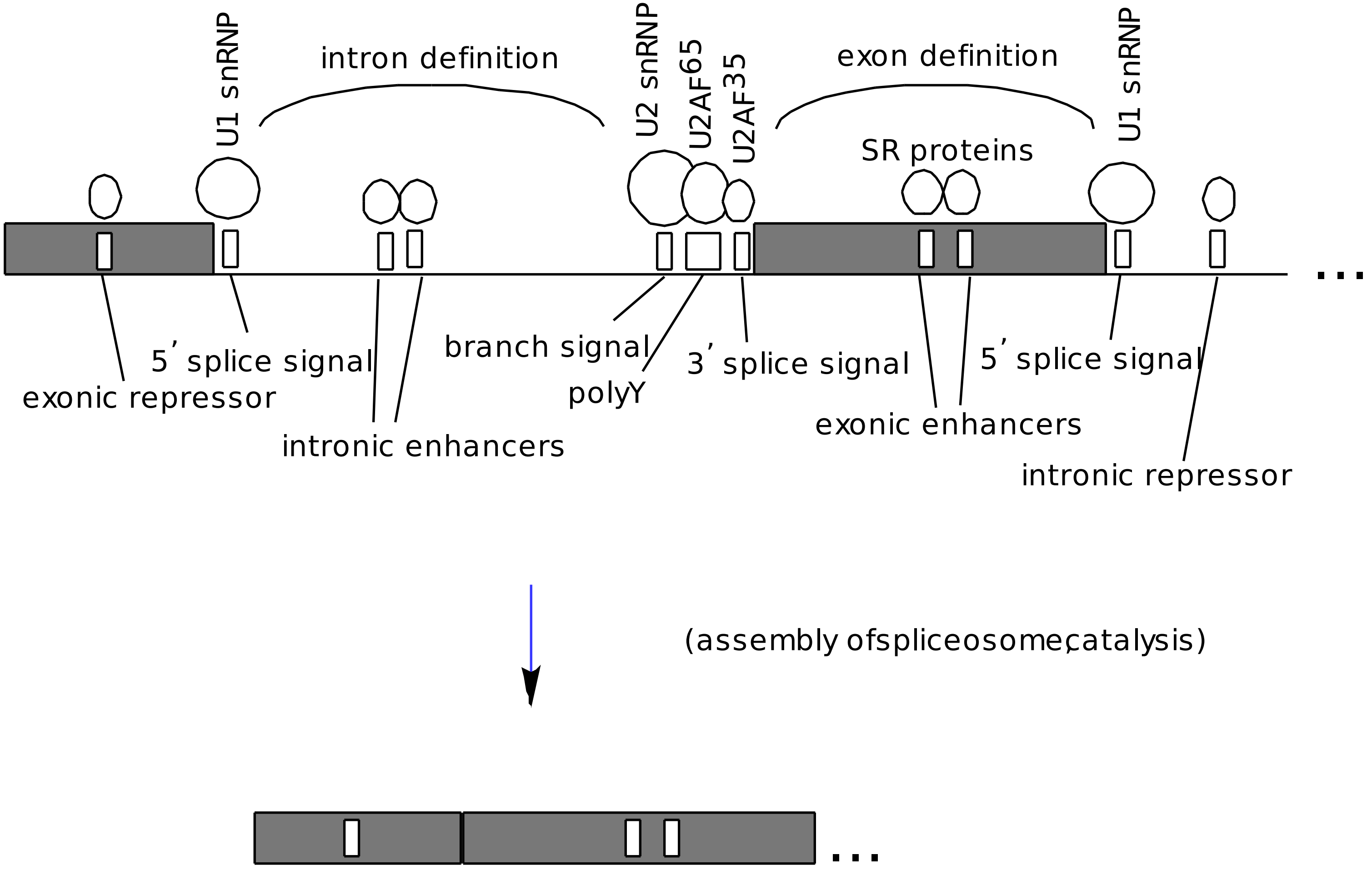
Identify alternative gene isoforms

Infer locations of genes

Genes

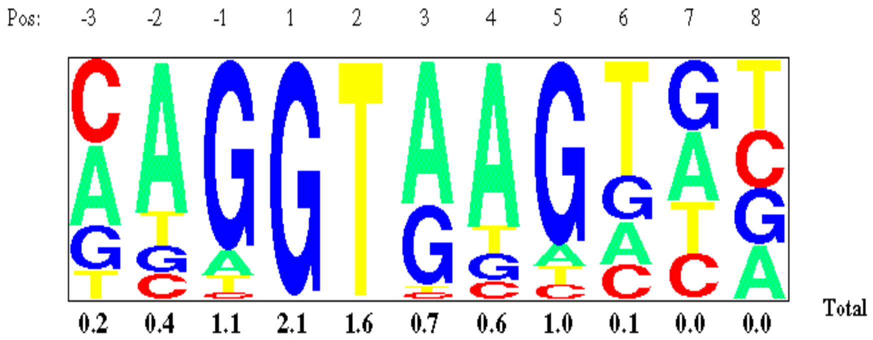
E. Birney & C. Burge

Pre-mRNA Splicing

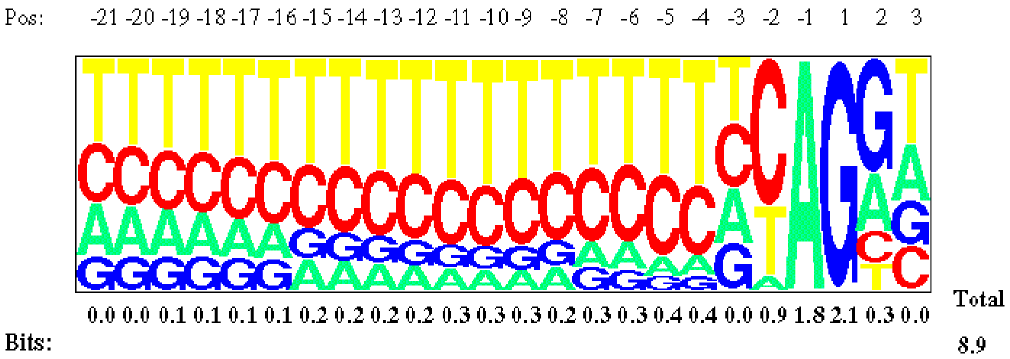


Human Splice Signal Motifs

5' splice signal



3' splice signal



Splice Signal Models I

Sequence $S = S_1 S_2 S_3 \ldots S_n$

Weight Matrix Model (WMM)

$P(S|+) = P_1(S_1)P_2(S_2)P_3(S_3) \ldots P_n(S_n)$

Assumes independence between positions

Weight Array (Markov) Model (WAM)

$$P(S|+) = P_1(S_1)P_2(S_2|S_1)P_3(S_3|S_2) \ldots P_n(S_n|S_{n-1})$$

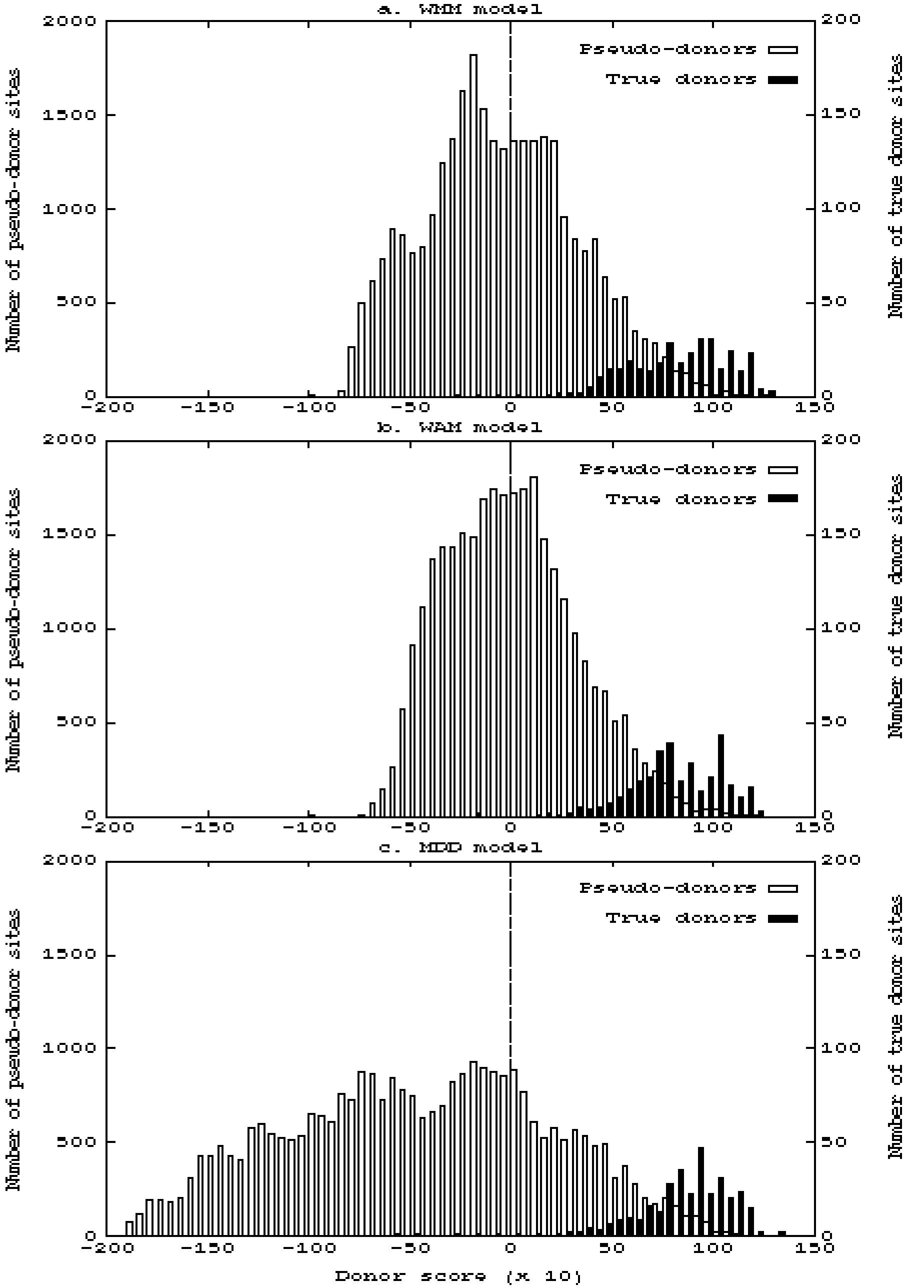
Allows for nearest−neighbor dependence

In either case, discriminate based on score:

$$s(S) = \log_2(P(S|+)/P(S|-))$$

5’ Splice Signal Scores

Fig. 6. Comparison of donor splice signal models



Comparison of Human

5′ Splice Signal Models

Sensitivity =

% of true sites above score cutoff

Specificity =

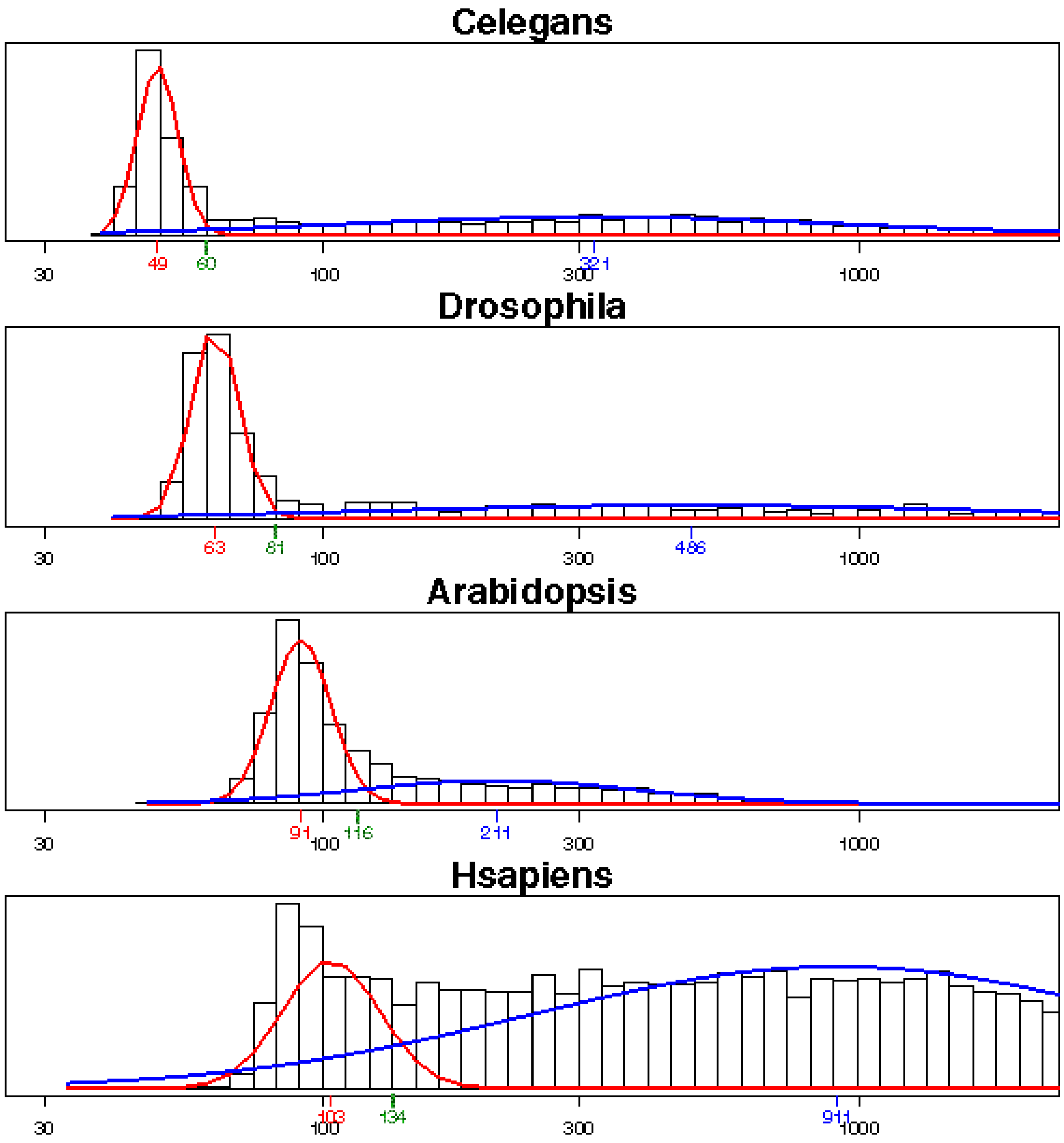
% of sites above cutoff which are true

Sensitivity Level

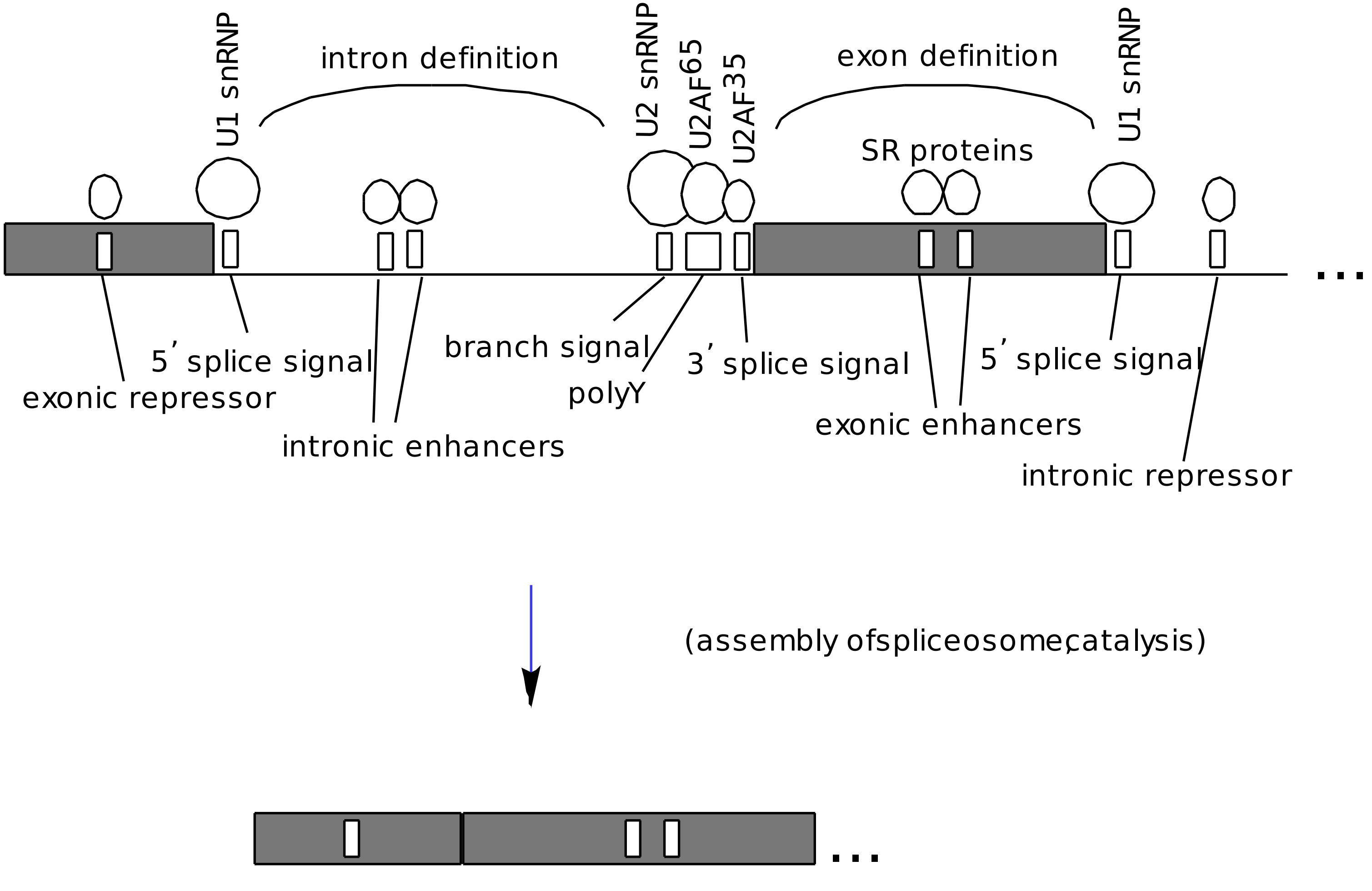
Model	20%	50%	90%
WMM	50%	32%	7%
WAM	50%	33%	7%
MDD	54%	36%	9%

Data from Burge, 1998 "Comp. Methods in Mol. Biology"

Intron Length Distributions



Pre-mRNA Splicing



Characterizing the sources of information used for splicing

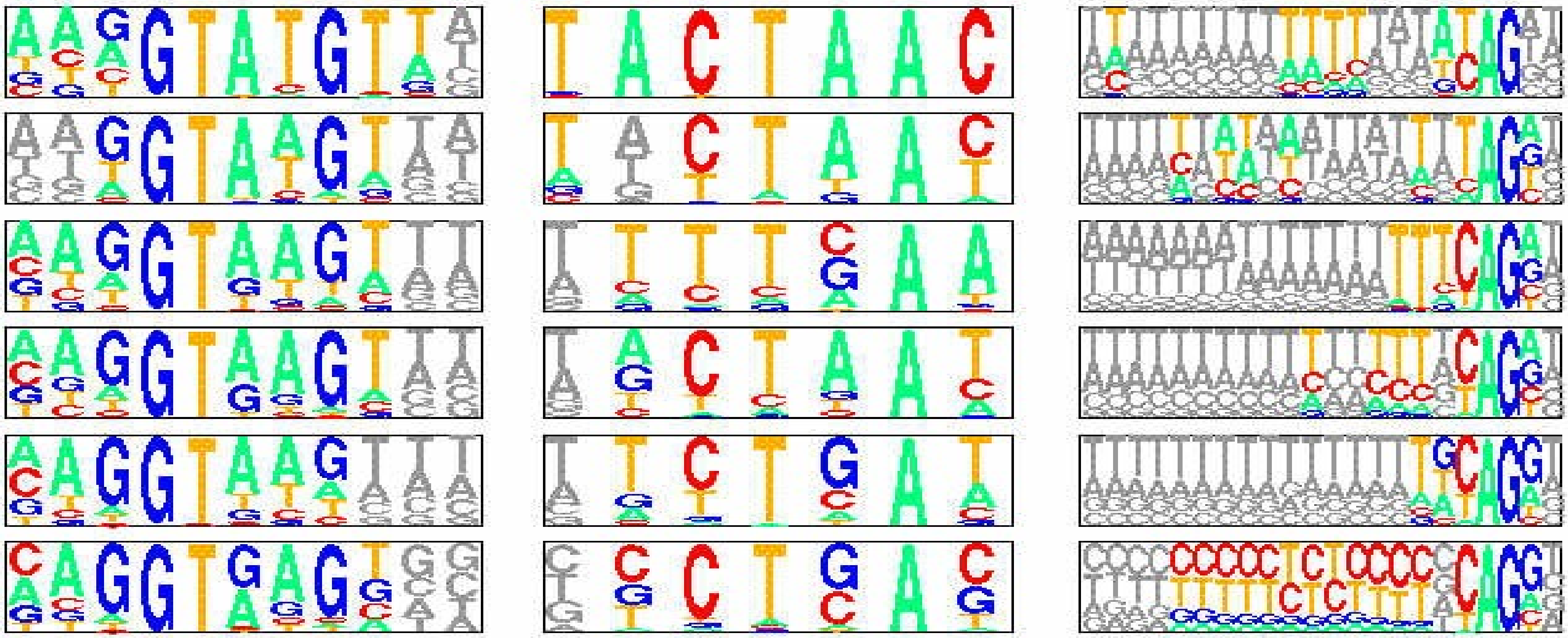
- 5ʼ splice signal (.AG/GTRAGt)
- 3ʼ splice signal (…YYYYYY.YAG/)
- Branch signal (…CTGAC..)
- Intron length preference
- Intron composition

Splicing-verified Transcripts

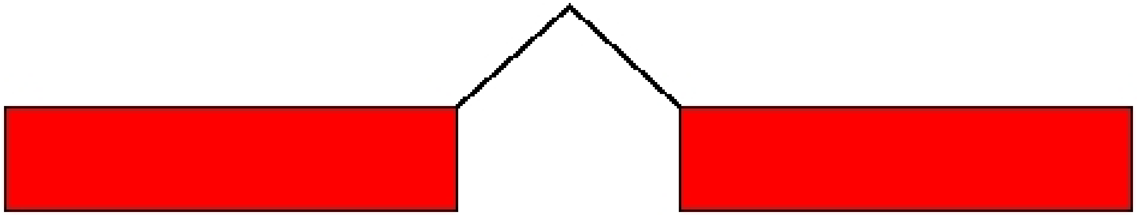
Org	**MBp**	**i-Tx**	**Introns**	**Int/iTx**	**%Short**
Yeast	12	152	152	~ 1	~50
Worm	100	691	3,577	~ 7	46
Fly	140	1,310	3,737	~ 4	54
Arab	125	1,121	5,265	~ 5	63
Human	3,000+	8,165	33,666	~ 9	10

Data from Sep, 2000 GenBank release

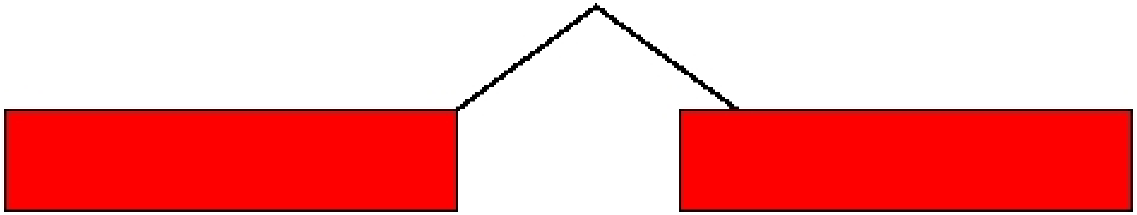
Splice Signal Sequences



"exact prediction"



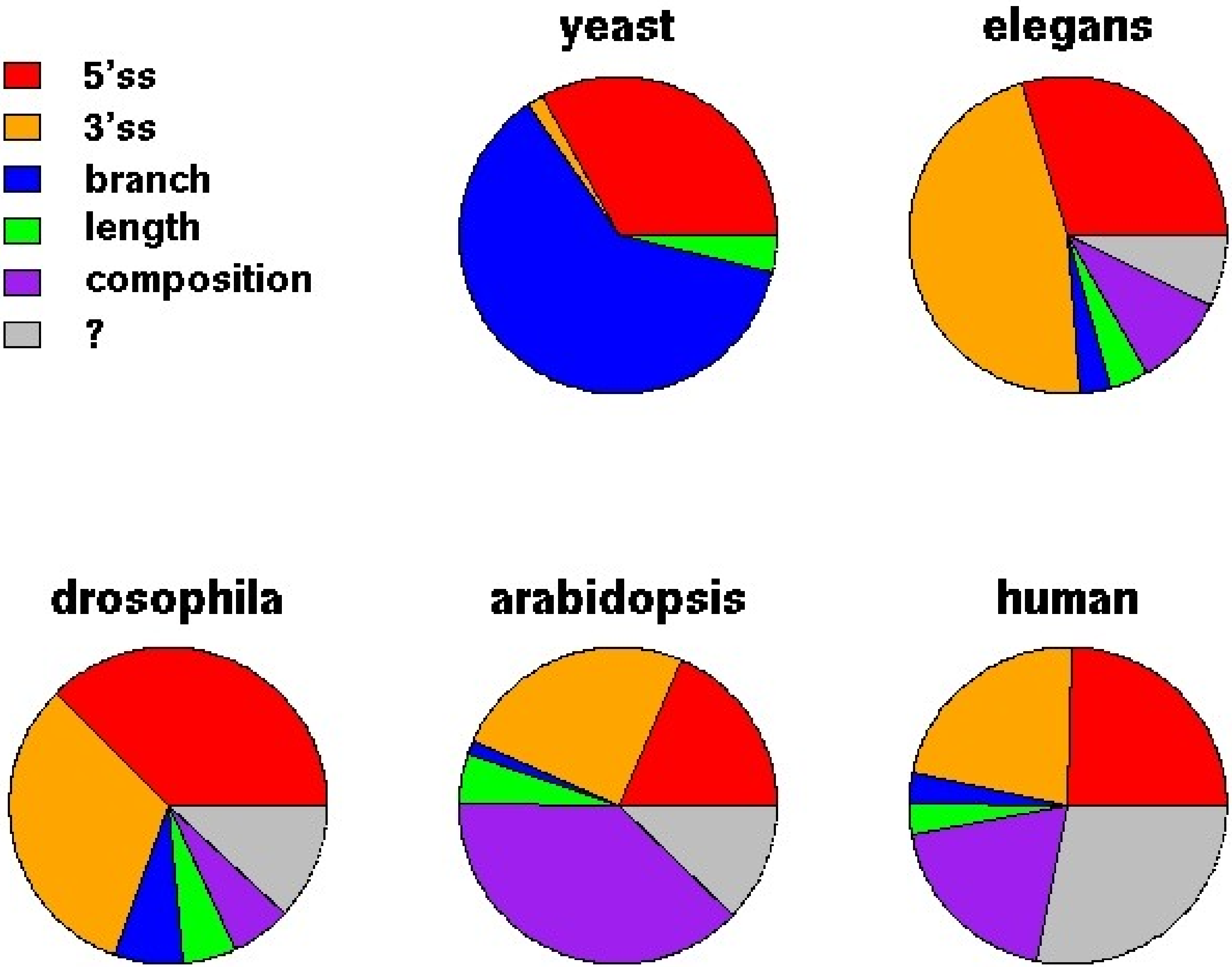
"detection"

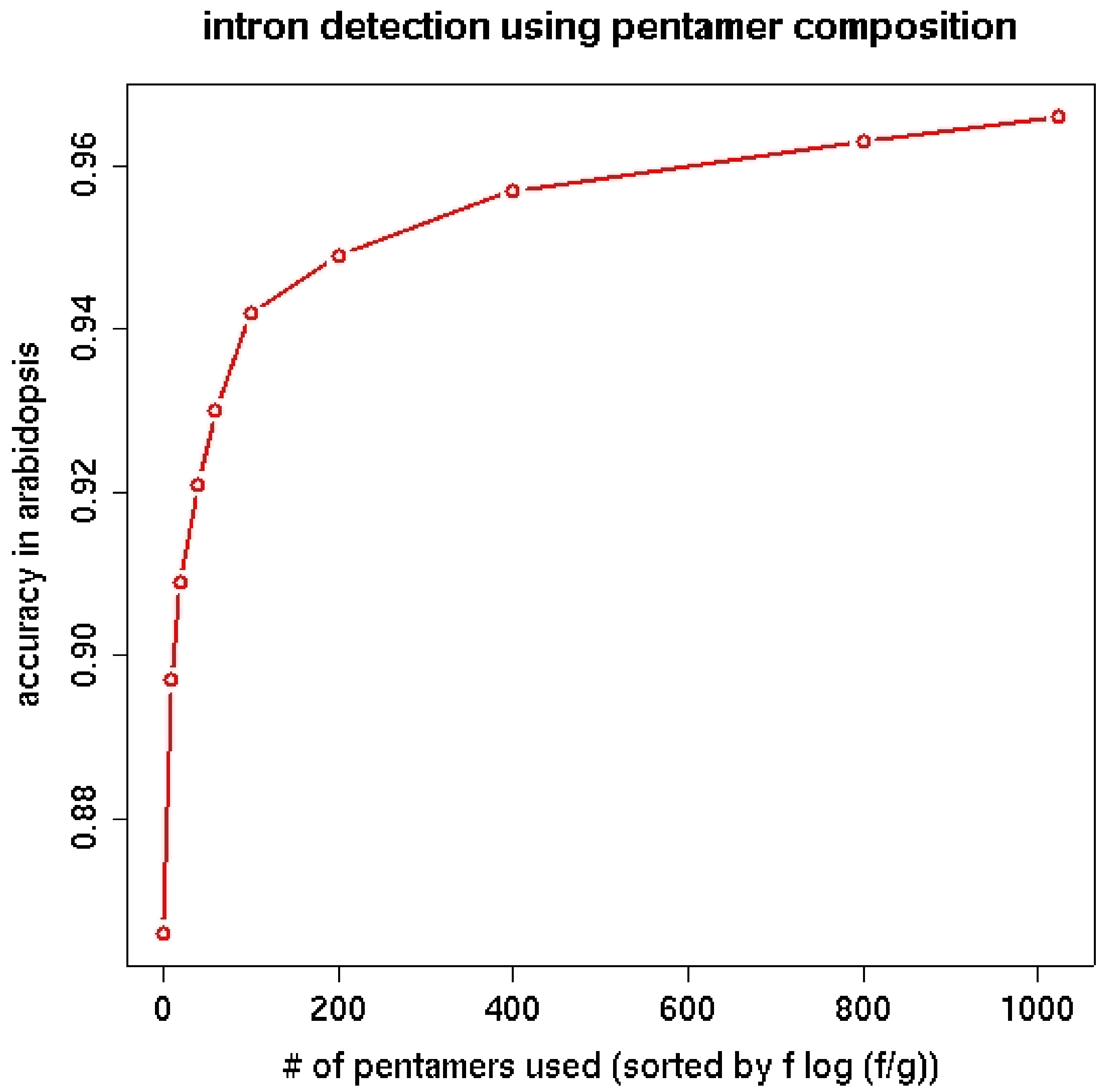


IntronScan Accuracy

	5’ss and 3’ss only		Complete model	
Organism	Detect	Exact	Detect	Exact
Yeast	90	43	98	86
Elegans	95	92	97	95
Fly	92	88	96	94
Arabidopsis	82	68	96	92
Human	76	65	88	85

Fivefold cross-validated


5'ss
3'ss
branch
length
composition
?
yeast
elegans
drosophila
arabidopsis
human


intron detection using pentamer composition
accuracy in arabidopsis
0.88
0.90
0.92
0.94
0.96
0
200
400
600
800
1000
# of pentamers used (sorted by f log (f/g))

Top Ten Intronic Pentamers

Arabidopsis	Drosophila	Human
TCTCT	**ATATA**	**GTGGG**
TTTTT	**AAATA**	**CTGGG**
TTTGT	**TATAT**	**GAGGG**
TCTTT	**TGATT**	**CAGGG**
TGTTT	**ACTTA**	**TGGGG**
TCTGT	**ACATA**	**GCAGG**
TTCTT	**TTTGT**	**GGTGG**
TGTGT	**CATTT**	**GGAGG**
CTTTT	**TTAAA**	**GCGGG**
TTTCT	**TCATT**	**GCTGG**

Top Ten Exonic Pentamers

Arabidopsis	Drosophila	Human
TGAAG	GGCGG	GATGA
CAAAG	CGAGG	CAGAA
AGAAG	CGCTG	GAAGA
TGCTG	AGGAG	CAGCA
TCTGA	TGGCC	CACCA
TGCAG	AGCTG	CTGAA
TGGAG	TGCTG	GTGGA
GGAAG	AGCAG	CAGGA
CGAAG	AGAAG	GAGGA
GAAGG	TGCAG	CTGGA

Summary

- Genes have a grammatical structure

  probabilistic models of this structure are interesting and useful
- Computational methods interact with experimental methods in modern biology
- Introns also have a grammatical structure

  sequence analysis may help us to deduce aspects of this structure
- There are many interesting related problems:
  - Finding RNA genes, identifying regulatory elements,
  - Understanding transcription, regulatory networks, etc.