

# Bioinformatics for Biologists

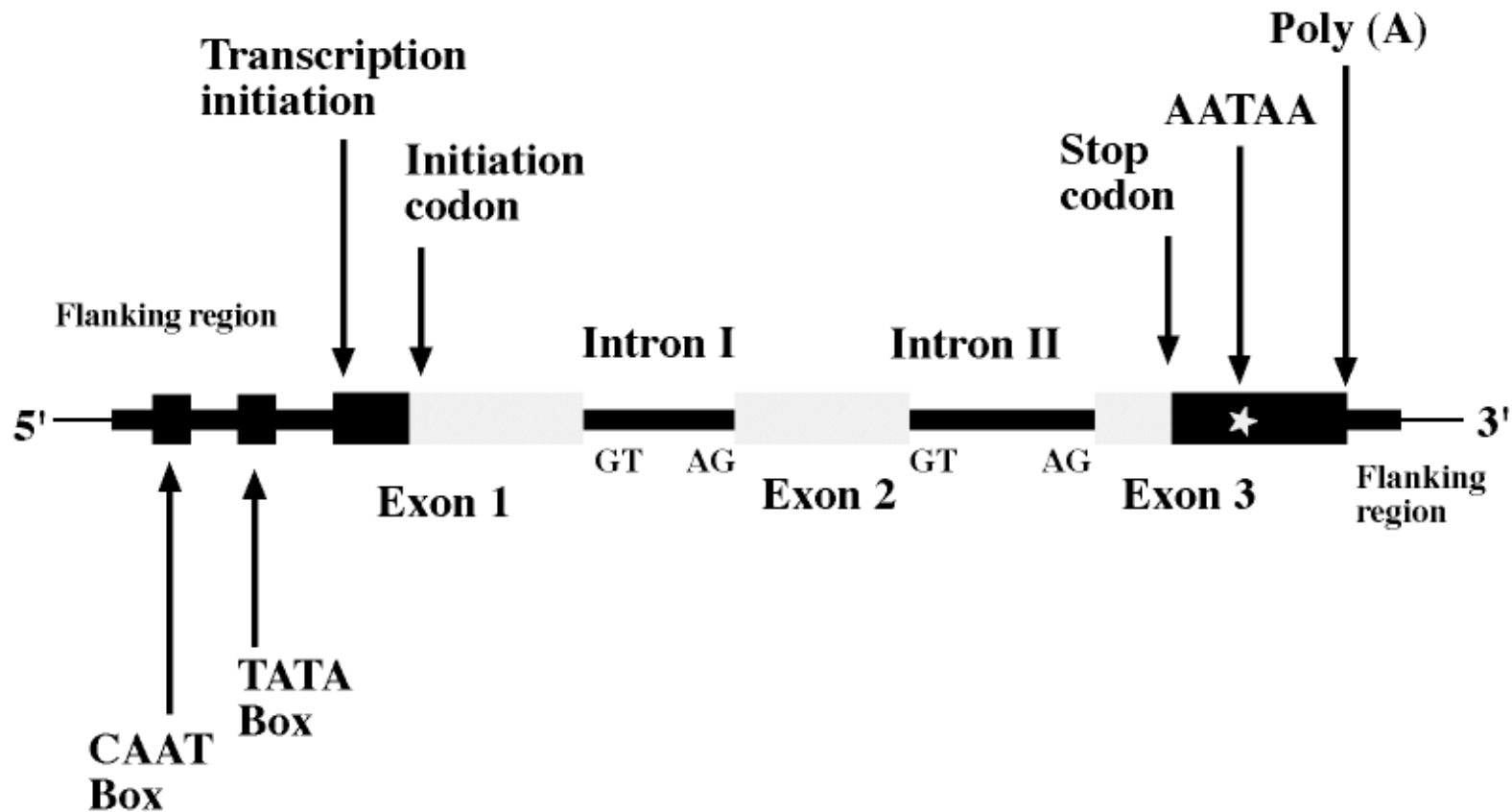
## Sequence Analysis: Part III. Gene Finding

Fran Lewitter, Ph.D.  
Head, Biocomputing  
Whitehead Institute



# Problem to Solve

---



# Types of Signals to Detect

---

- Transcriptional
  - TSS
  - TATA box
  - PolyA
- Translational
  - Kozak (CC A/G CCAUGG)
  - Termination codon (UAA, UAG, UGA)
- Splicing
  - Introns - GT.....AG

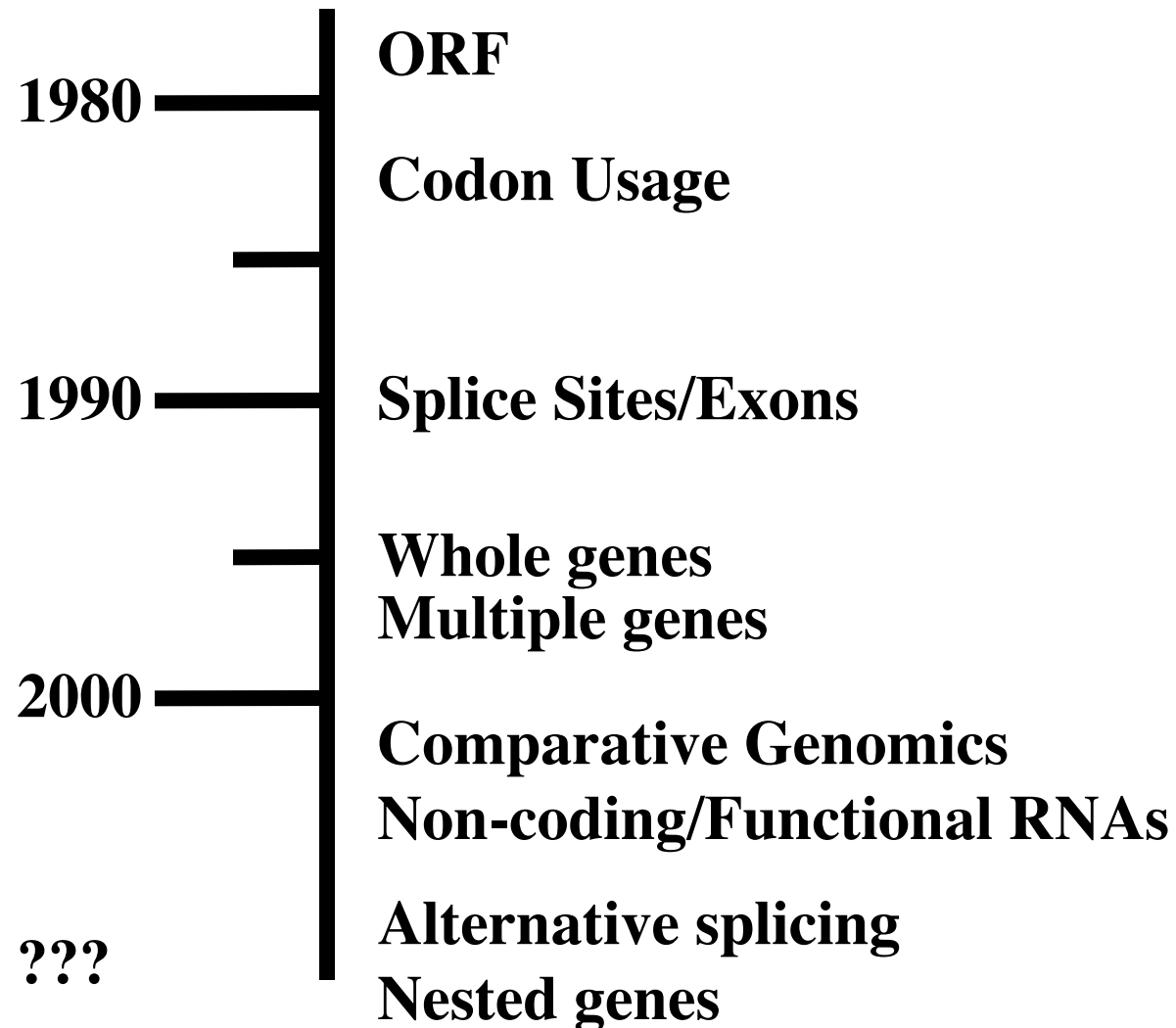
# Gene Finding Strategies

---

- Content-based methods
  - codon usage, compositional complexity
- Site-based methods
  - presence or absence of specific pattern or sequence
- Comparative methods
  - determination based on homology

# Evolution of Gene Finding Programs

---



# RepeatMasker

RepeatMasker Server

RepeatMasker is a program that screens DNA sequences for low complexity DNA sequences and interspersed repeats. The masked out sequence can be used to for BLAST search.  
*Please refer to: Smit, AFA & Green, P "RepeatMasker" at <http://repeatmasker.genome.washington.edu>*

[Home](#) || [Help](#) || [Check Queue](#) || [Your Suggestion](#) || [References](#) || [RepBase Update](#)

Enter your sequence ( *sequence in fasta format* )

( OR ) Upload the file

DNA Source is from

Primates

Rodents

Other Mammals

Other Vertebrates

Arabidopsis

Grasses

Drosophila

**Running options**

Fast (*quick search, 3-4 times faster*)

Slow (*slow search, 2.5 times slower*)

**Repeat Options**

Do not mask low\_complexity DNA or simple repeats

only masks Alus (and 7SLRNA, SVA and LTR5)(only for primate DNA)

only masks low complex/simple repeats (no interspersed repeats)

**Output Options**

Show Alignments

Mask with X's to distinguish masked regions from Ns already in query

Produce an annotation table with fixed width columns

html validate this page

# RepeatMasker

---

## Repeat sequence:

SW score	perc div.	perc del.	perc ins.	query sequence	position in query begin end (left)	matching repeat repeat	repeat class/family	position in repeat begin end (left)	repeat ID
2117	3.9	0.0	1.2	myseq	7 263 (8487)	+	AluY SINE/Alu	58 311 (0)	1
7658	20.9	6.8	5.7	myseq	2193 2517 (6233)	C	L1MDa LINE/L1	(1460) 5080 4760	2
2516	4.9	2.0	0.0	myseq	2518 2822 (5928)	+	AluY SINE/Alu	1 311 (0)	3
7658	20.9	6.8	5.7	myseq	2823 4135 (4615)	C	L1MDa LINE/L1	(1780) 4760 3462	2
5685	6.2	0.8	0.0	myseq	4136 4864 (3886)	+	L1PA10 LINE/L1	5417 6151 (17)	4
7658	20.9	6.8	5.7	myseq	4865 5181 (3569)	C	L1MDa LINE/L1	(3078) 3462 3150	2
2130	11.2	3.0	0.7	myseq	5182 5303 (3447)	+	AluSq SINE/Alu	1 119 (194)	5
351	0.0	0.0	0.0	myseq	5304 5342 (3408)	+	(TAAA)n Simple_repeat	2 40 (0)	6
2130	11.2	3.0	0.7	myseq	5343 5514 (3236)	+	AluSq SINE/Alu	119 302 (11)	5
2593	6.6	2.8	0.0	myseq	5525 5886 (2864)	+	L1PA13 LINE/L1	5792 6163 (0)	7
7658	20.9	6.8	5.7	myseq	5887 6390 (2360)	C	L1MDa LINE/L1	(3398) 3142 2645	2
2092	9.6	1.0	3.6	myseq	6391 6693 (2057)	C	AluSc SINE/Alu	(13) 296 2	9
7658	20.9	6.8	5.7	myseq	6694 8738 (12)	C	L1MDa LINE/L1	(3895) 2645 491	2



# RepeatMasker

---

## Summary:

Total length: 8750 bp

GC level: 35.61%

Bases masked: 6803 bp (77.75%)

	number of elements*	length occupied	percentage of sequence
SINES:	4	1159 bp	13.25 %
ALUs	4	1159 bp	13.25 %
MIRs	0	0 bp	0.00 %
LINEs:	3	5605 bp	64.06 %
LINE1	3	5605 bp	64.06 %
LINE2	0	0 bp	0.00 %
L3/CR1	0	0 bp	0.00 %
LTR elements:	0	0 bp	0.00 %
MaLRs	0	0 bp	0.00 %
ERV_L	0	0 bp	0.00 %
ERV_classI	0	0 bp	0.00 %
ERV_classII	0	0 bp	0.00 %
DNA elements:	0	0 bp	0.00 %
MER1_type	0	0 bp	0.00 %
MER2_type	0	0 bp	0.00 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats:		6764 bp	77.30 %
Small RNA:	0	0 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	1	39 bp	0.45 %
Low complexity:	0	0 bp	0.00 %

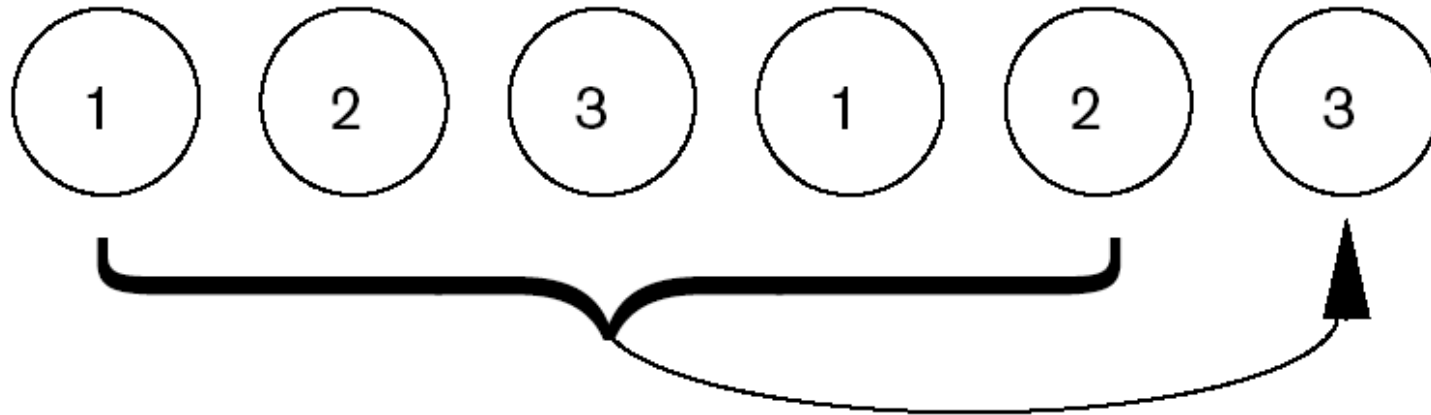
# Coding Measures

---

- Look at frequencies of codons (e.g. redundancy of genetic code; Leucine = UUA, UUG, CUU, CUC, CUA, CUG)
- 6-tuple or hexamer approach  
ACCTCG TACTCG GCCCTC  
Thr Ser Tyr Ser Ala Leu

# Fifth Order Markov Models

---

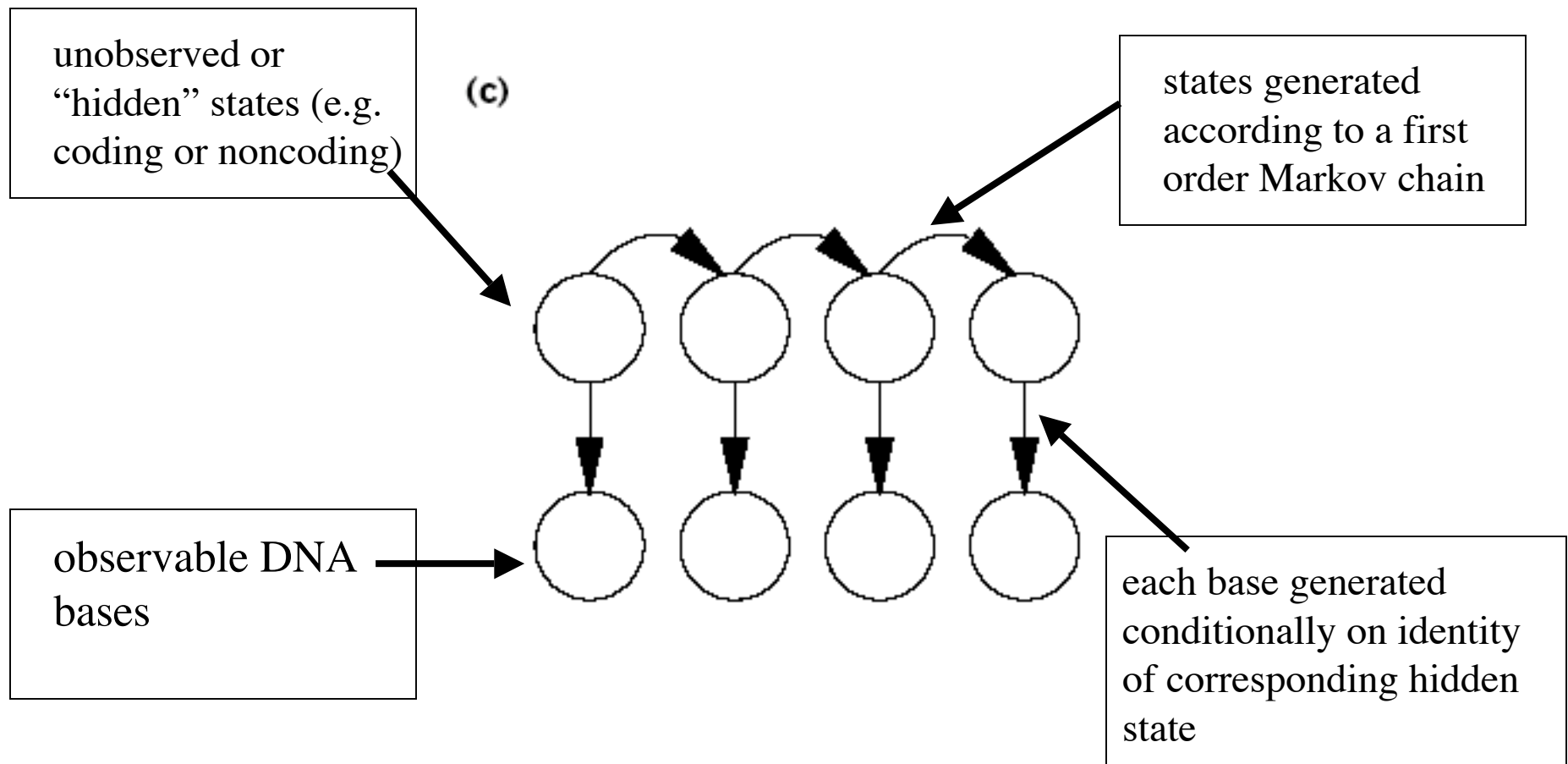


Periodic fifth order Markov model. Circles represent consecutive DNA bases, numbers indicate codon position, and the arrows indicate that the next base is generated conditionally on the previous five and on the codon position.

Burge and Karlin, *Current Opinions in Structural Biology* 1998, 8:346-354.

# Hidden Markov Models

---



Burge and Karlin, *Current Opinions in Structural Biology* 1998, 8:346-354.

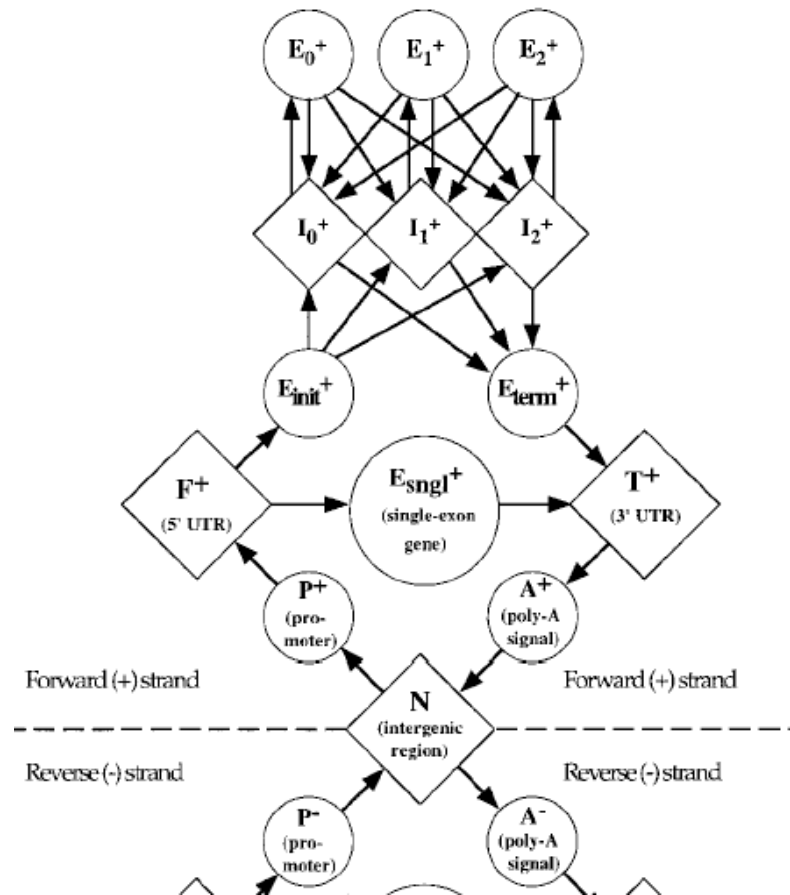
# GENSCAN

---

- MM - prob for a given nuc to occur at position  $p$  depends on nuc occupying previous  $k$  positions
- Generalized Hidden Markov Model (GHMM)
- Optimize module performing signal recognition
- Incorporates influence of C+G content
- Considers gene models on both strands
- Can identify multiple genes

Burge and Karlin, JMB:268:78-94, 1997

# GENSCAN



Burge and Karlin,  
JMB:268:78-94, 1997

# Gene Finding Programs

---

- FGENESH - Softberry
- GeneID - Barcelona
- GeneMark HMM - Georgia Tech
- Genie - UCSC & LBNL
- Genscan - Stanford and MIT
- GenomeScan - MIT
- MZEF/First exon - Cold Spring Harbor
- Twinscan - WU

# HMR195 Test Set

---

- 103 human, 82 mouse, 10 rat sequences
- Sequence new since August, 1997
- Genomic sequences containing exactly one gene
- No mRNA sequences, pseudogenes or alternatively spliced genes
- The mean length of sequences is 7,096 bp

*Rogic, Mackworth and Ouellette, Genome Research 11:817-832, 2001*



# HMR195 Test Set (con't)

---

- 43 single-exon genes; 152 multi-exon genes
- Average number of exons per gene is 4.86
- Mean exon length = 208 bp,  
mean intron length = 678 bp,  
mean coding length per gene = 1,015 bp  
(~330 aa)
- Coding sequence 14%, intronic sequence 46% and intergenic DNA 40%.

*Rogic, Mackworth and Ouellette, Genome Research 11:817-832, 2001*

# Definitions

---

- ***Sensitivity***: the proportion of true sites (e.g., exons or donor splice sites) which are correctly predicted =  $TP / (TP + FN)$
- ***Specificity***: the proportion of predicted sites which are correct =  $TP / (TP + FP)$

# Program Comparisons Results

- Genscan and HMMgene had reliable scores for exons
- Nucleotide Sn = .95 for Genscan and .93 for HMMgene.
- Sp = .90 and .93, respectively
- Accuracy dependent on G+C content

*Rogic, Mackworth and Ouellette, Genome Research 11:817-832, 2001*

**Table 2.** Accuracy versus Signal Type

Programs	Signal type			
	start codon (195)	acceptor site (753)	donor site (753)	stop codon (195)
FGENES	0.67 (0.63)	0.80 (0.77)	0.85 (0.82)	0.75 (0.72)
GeneMark.hmm	0.46 (0.60)	0.81 (0.75)	0.82 (0.78)	0.57 (0.64)
Genie	0.56 (0.57)	0.77 (0.82)	0.78 (0.83)	0.72 (0.73)
Genscan	0.61 (0.78)	0.87 (0.80)	0.90 (0.84)	0.76 (0.86)
HMMgene	0.75 (0.78)	0.81 (0.85)	0.83 (0.87)	0.78 (0.81)
Morgan	0.43 (0.43)	0.66 (0.57)	0.65 (0.56)	0.39 (0.39)
MZEF	—	0.59 (0.65)	0.66 (0.73)	—

For each program, the proportion of actual signals identified correctly (the upper number) and the proportion of predicted signals that are correct (the lower number) are averaged over all signals belonging to a particular type. The number in parenthesis in the header of each column represents the number of signals of each type in the HMR195 dataset.

# GenomeScan

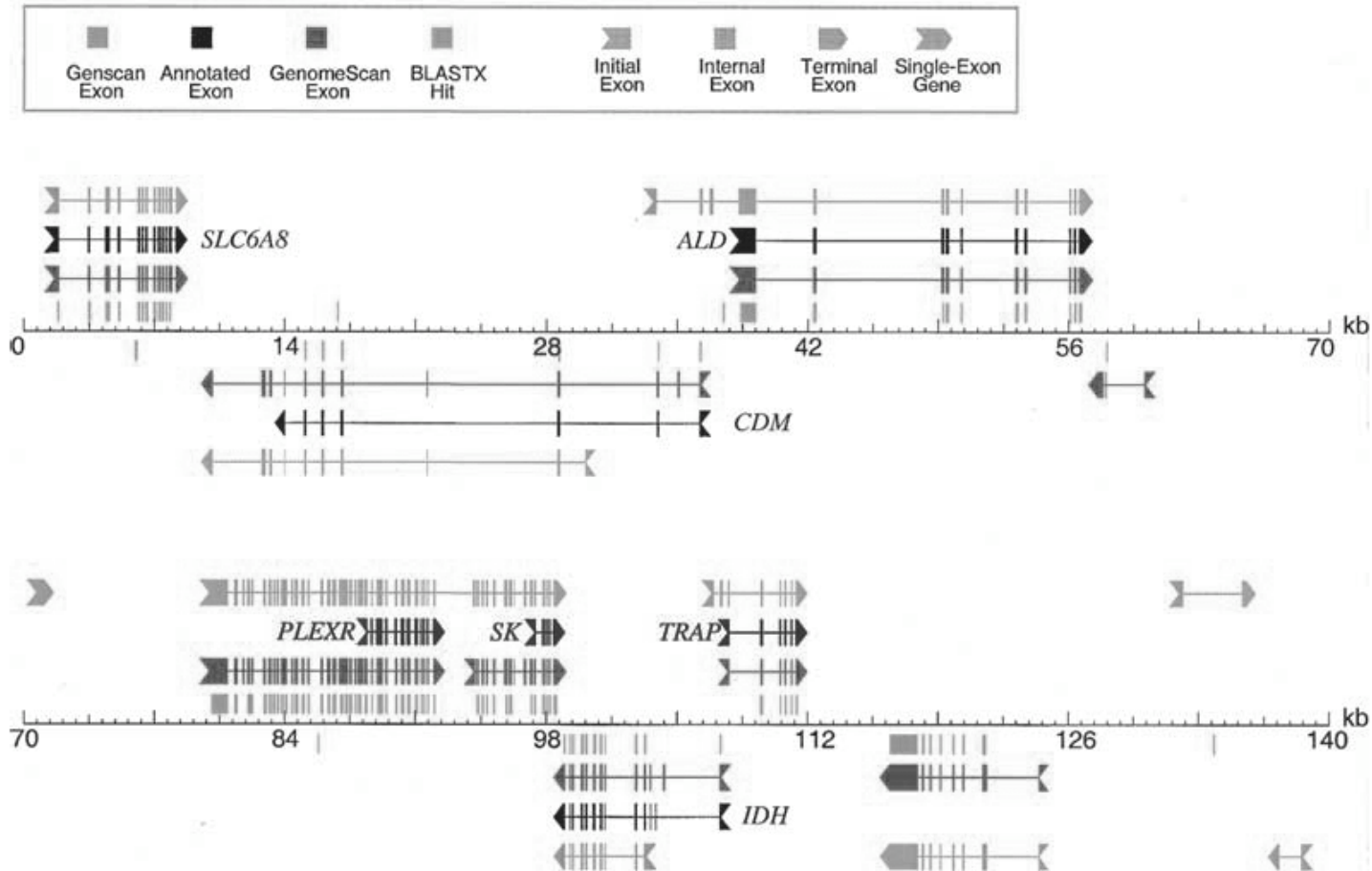
---

- Combines exon-intron and splice signal models with similarity to known proteins
- Used to identify genes in human draft sequence
- Uses GENSCAN and BLASTX

Yeh, Lim, and Burge, Genome Research 11:803-816, 2001.

# GenomeScan

B



Yeh, Lim, and Burge, *Genome Research* 11:803-816, 2001.

WIBR Bioinformatics Course, © Whitehead Institute, October 2003

# GenomeScan

**Table 4.** Summary of GenomeScan-predicted Genes and Partial Genes in the Human Genome

Similarity category	Type of predicted gene						
	Complete genes (>2 exons)			Partial genes		All genes (partial + complete)	
	No. of genes	No. of exons/gene	No. of aa/gene	No. of genes	No. of exons/gene	No. of genes	% of all predicted genes
Known (cDNA)	5698	9.6	496	8901	4.9	16040	41.5
Protein + EST	4502	8.8	510	6537	5.5	12546	32.5
Proteins only	2767	5.2	303	4600	3.1	10061	26.0
All	12967	8.4	460	20038	4.7	38647	100.0

Genes were predicted in the September 2000 GoldenPath human genome sequence as described in Methods. Predicted coding sequences (CDS) were first compared to cDNAs in the RefSeq cDNA database (September 2000) using BLASTN; those which had a hit at least 100 bp long with at least 98% identity are listed as "known". The remaining predicted coding sequences were searched against dbEST (September 2000 release) using BLASTN; those which had a hit at least 100 bp long with at least 97% identity are listed as "Protein + EST". All other predicted genes are categorized as "Protein only" because all GenomeScan-predicted genes have at least modest similarity to a known protein. Statistics are listed separately for predicted partial genes and predicted complete genes with at least three exons; the category "all genes" includes these two groups as well as predicted 1- and 2-exon genes.

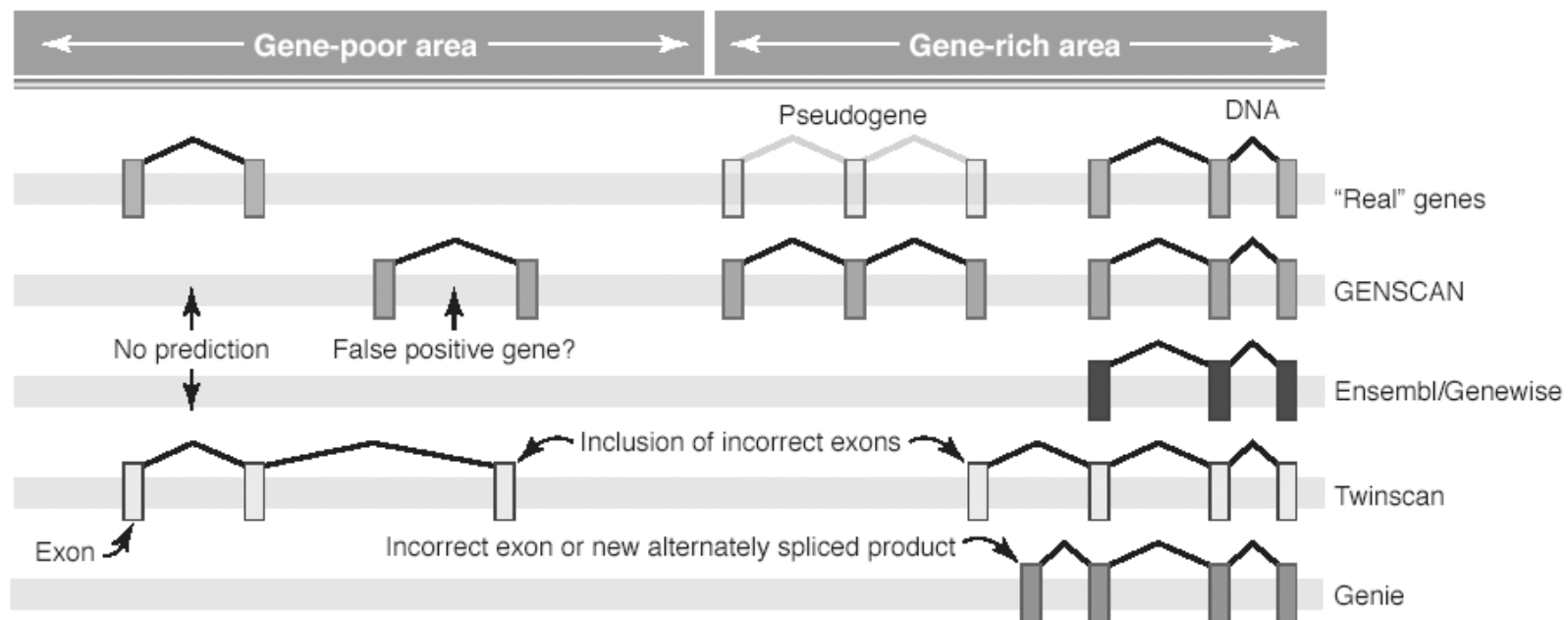
Yeh, Lim, and Burge, Genome Research 11:803-816, 2001.

# Other Approaches

---

- Use microarrays to identify expressed genes based on the coexpression of sets of adjacent exons as predicted by GENSCAN (Shoemaker, et al, Nature 409:922-927, 2001)
- RT-PCR with radio-labeled primers targeted to pairs of adjacent predicted exons, followed by sequencing of the amplified product (Das, Burge et al, Genomics. 77:71-8, 2001)

# E. Pennisi, Science 301:1040, 2003



**Never perfect.** No program calls all genes correctly. Some see genes (shown here as coding regions, or exons, connected by bent lines) where there are none; some miss a gene altogether; and some don't put all the gene's parts in the right places.



# Comparative Genomics

---

Cliften P, et al. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*. 301:71-6, 2003.

Kellis M, et al Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241-54, 2003.

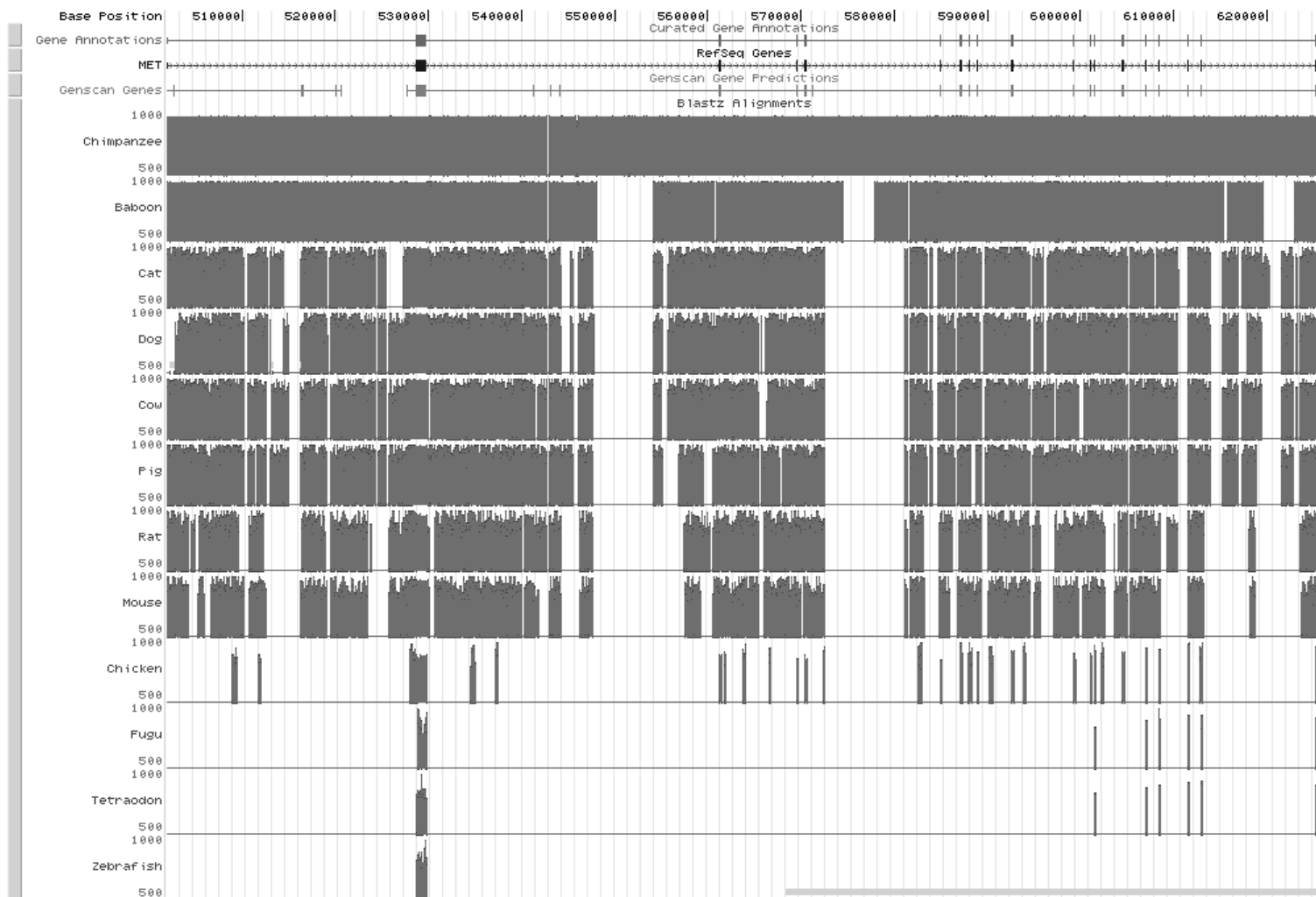
Thomas JW, et al. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424:788-93, 2003

# Some other useful Tools

---

- GeneWise - compares protein sequence to genomic DNA sequence, allowing for introns and frameshifting errors
- Pipmaker/Multipipmaker - long alignments of two or multiple genomic regions (local)
- VISTA - whole genome alignments - human, mouse, rat (global)
- UCSC Genome Browser

# UCSC Genome Browser



# Future Challenges

---

- Alternative Splicing
- Gene products functioning at RNA level
- Nested genes
- 5' end of genes
- Other unusual characteristics