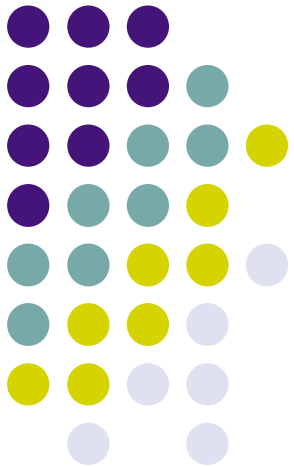
Graduate Computational Genomics

02-710 / 10-810 & MSCBIO2070

Computational gene prediction

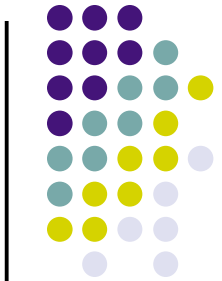


Takis Benos

Lecture #9, February 13, 2007

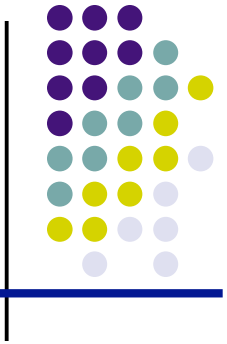
Reading: hand-outs & papers

Outline



- The problem
- Predicting the coding parts of the genome
- Predicting splice sites
- Predicting exons
- Predicting whole gene structures
  - *GenScan*
  - *TwinScan*
  - *N-SCAN*

The problem



Given a genomic DNA piece, predict the regions that are more likely to be part of a (protein coding) gene.

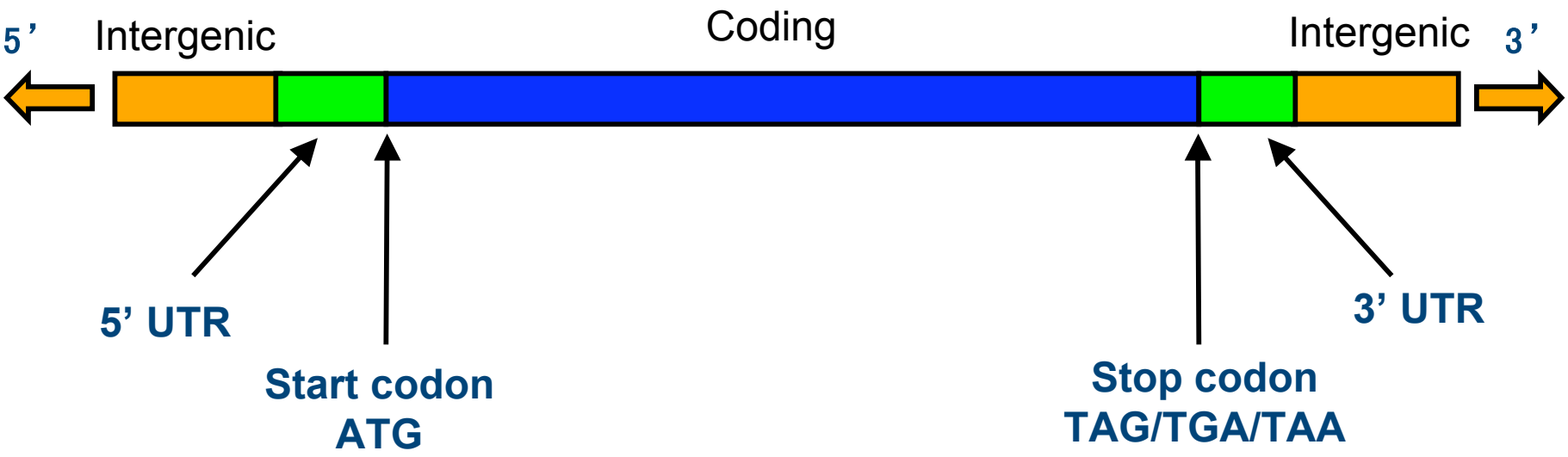
Protein coding gene parts:

- (promoter region)
- 5' Untranslated Region (*5' UTR*)
- Open Reading Frame (*ORF*)
- 3' Untranslated Region (*3' UTR*)

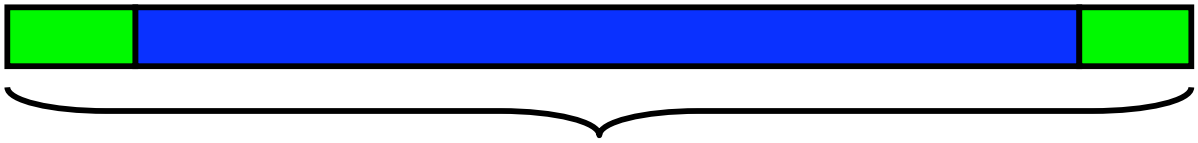
→ *Eukaryotes*

- Introns
- Exons

Finding Genes in Yeast

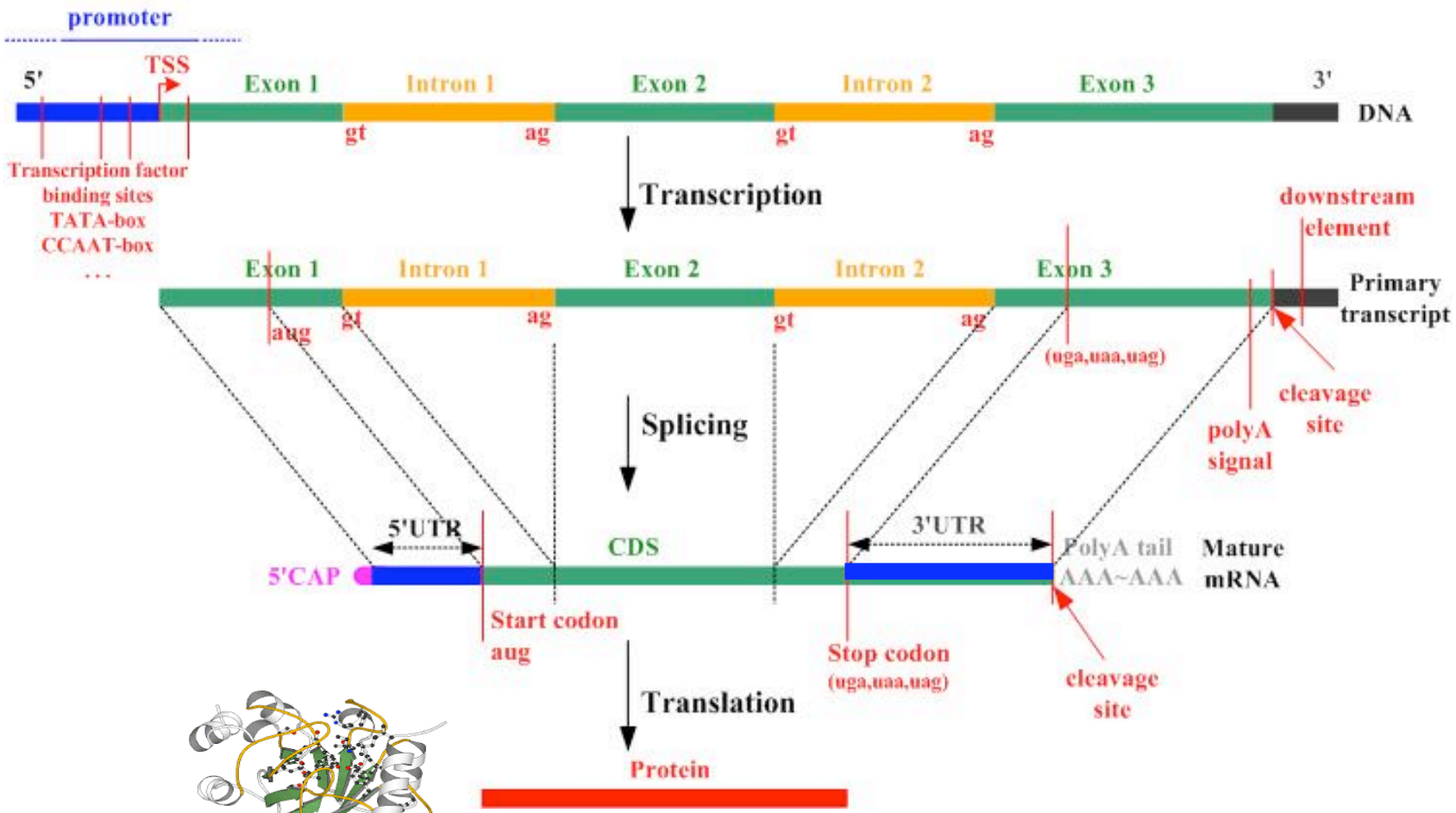


Mean coding length about 1500bp (500 codons)

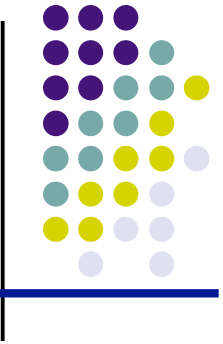


Transcript

Gene structure (eukaryotes)



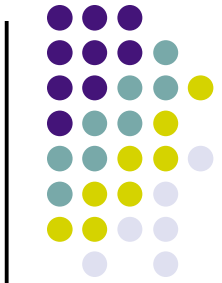
Codon bias



First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	
U	UUU UUC Phenylalanine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine	UGU UGC Cysteine	U C A G
	UUA UUG Leucine		UAA UAG Stop codon Stop codon	UGA Stop codon; UGG Tryptophan	
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine	CGU CGC CGA CGG Arginine	U C A G
			CAA CAG Glutamine		
A	AUU AUC AUA Isoleucine	ACU ACC ACA ACG Threonine	AAU AAC Asparagine	AGU AGC Serine	U C A G
	AUG Methionine; initiation codon		AAA AAG Lysine	AGA AGG Arginine	
G	GUU GUC GUA GUG Valine	GCU GCC GCA GCG Alanine	GAU GAC Aspartic acid	GGU GGC GGA GGG Glycine	U C A G
			GAA GAG Glutamic acid		

E. coli (per 1,000)		*yeast* (per 1,000)	
Phe	UUU 22.3	Phe	UUU 26.1
Phe	UUC 16.6	Phe	UUC 18.2
Leu	UUA 13.9	Leu	UUA 26.4
Leu	UUG 13.7	Leu	UUG 27.1
Leu	CUU 11.0	Leu	CUU 12.2
Leu	CUC 11.1	Leu	CUC 5.4
Leu	CUA 3.9	Leu	CUA 13.4
Leu	CUG 52.6	Leu	CUG 10.4
Ile	AUU 30.3	Ile	AUU 30.2
Ile	AUC 25.1	Ile	AUC 17.1
Ile	AUA 4.4	Ile	AUA 17.8
Met	AUG 27.9	Met	AUG 20.9
Val	GUU 18.3	Val	GUU 22.0
Val	GUC 15.3	Val	GUC 11.6
Val	GUA 10.9	Val	GUA 11.8
Val	GUG 26.4	Val	GUG 10.7

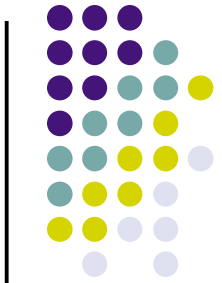
Coding exon finding



Measures of “coding potentials” related to codon usage:

- Codon usage measure
- Hexamer measure
- Amino acid usage measure
- Di-amino acid usage measure

Codon usage measure



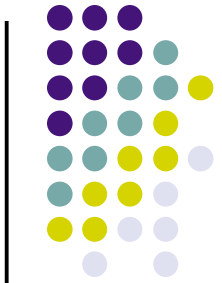
- Staden & McLachlan (1982)
  - "Coding" probability of a window is the product of the codon probabilities calculated from a reference set (mRNA):

$$P(w_j) = \prod_{i=1}^{L/3} P(c_i)$$

- Gribskov, Devereux & Burgess (1984)
  - "Coding" probability of a window is the product of log-likelihood of the in-frame *vs*. out-of-frame codon frequencies:

$$S(w_j) = \prod_{i=1}^{L/3} \frac{P(c_i \mid coding)}{P(c_i \mid non-coding)}$$

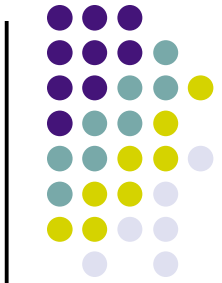
Codon usage measure (cntd)



- Hinds & Blake (1985)
  - Similar to Staden & McLachlan, but the P(c_i) is calculated as the frequency of the in-frame codon, c_i, over all occurrences (both in-frame and out-of-frame).

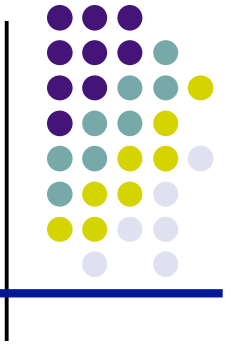
- Claverie & Bougueleret (1986 & 1990)
  - p(w_j) and q(w_j) are the frequencies of the hexamer w_j in exons and introns, resp.
  - Two hexamer measures were used:
    1. $d_1(w_j) = p(w_j) / (p(w_j) + q(w_j))$
    2. $d_2(w_j) = p(w_j) - q(w_j)$

Codon usage measure (cntd)

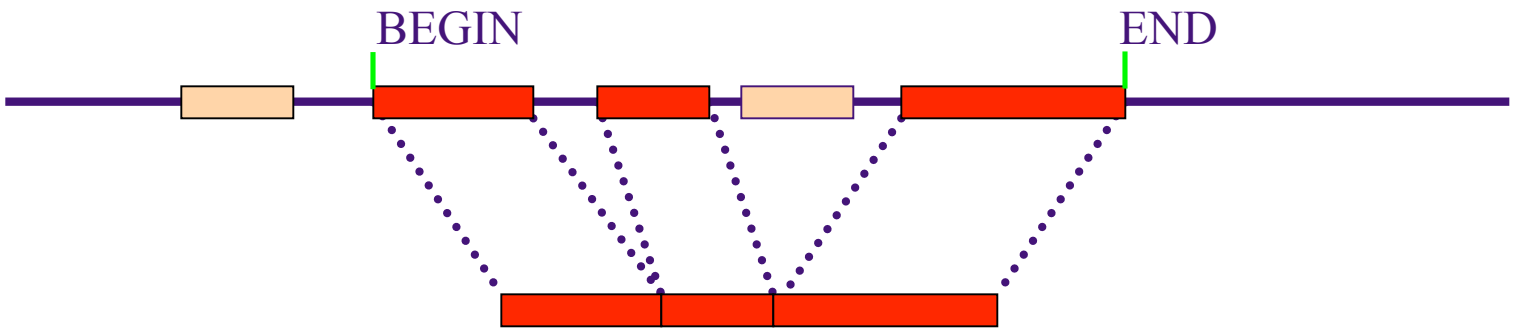


- Borodovsky *et al.* (1986)
  - Construct four Markov Models from reference set: (1-3) each of the three frames of the coding regions, (4) non-coding regions
  - For each window, calculate the (four) window probabilities given a particular model
  - Calculate the probabilities of each model given the window (using Bayes)
  - Markov models of order 1 through 5 were used.
  - Best results were obtained with MM of order 5 (hexamer frequencies)

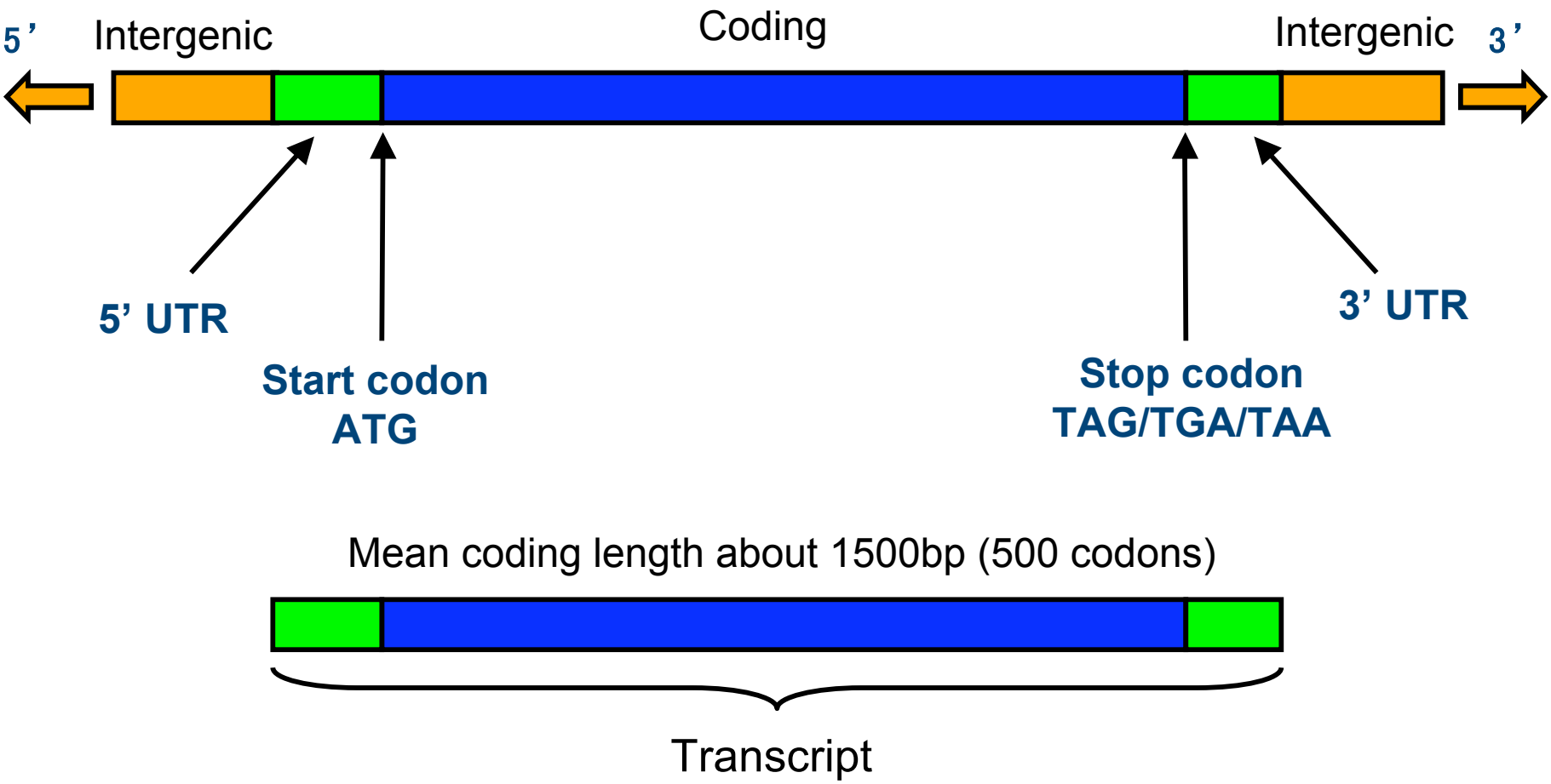
Predicting genes



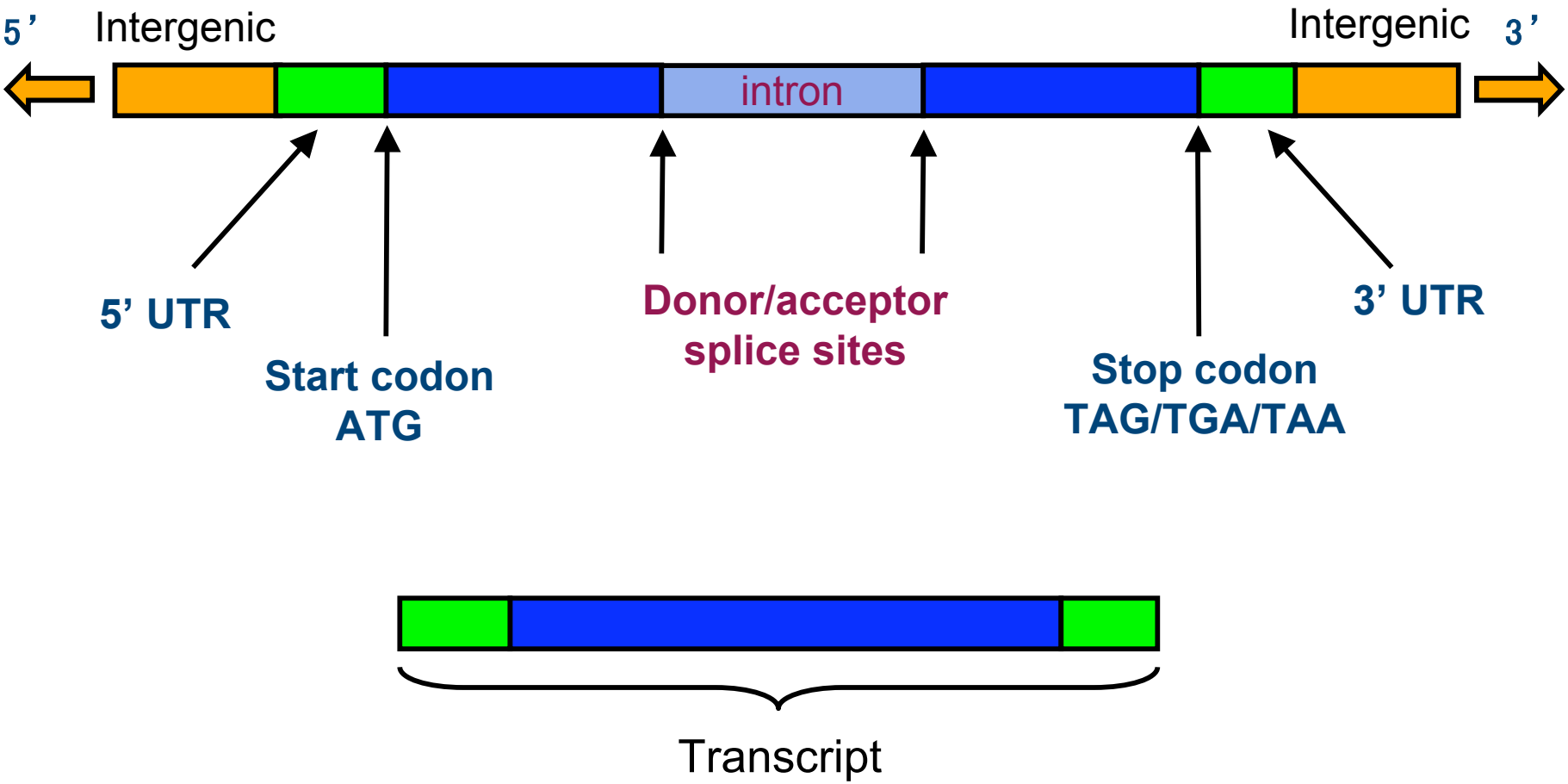
- ✓Prediction of exon(s)
- Prediction of exact splice sites
- Prediction of begin/end of the gene
- Selection of exons



Finding Genes in Yeast

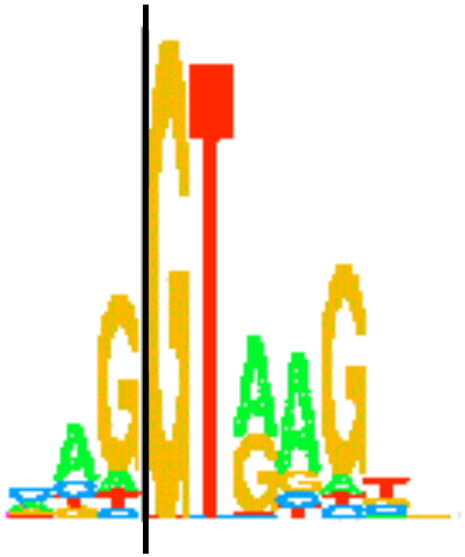


Splice sites



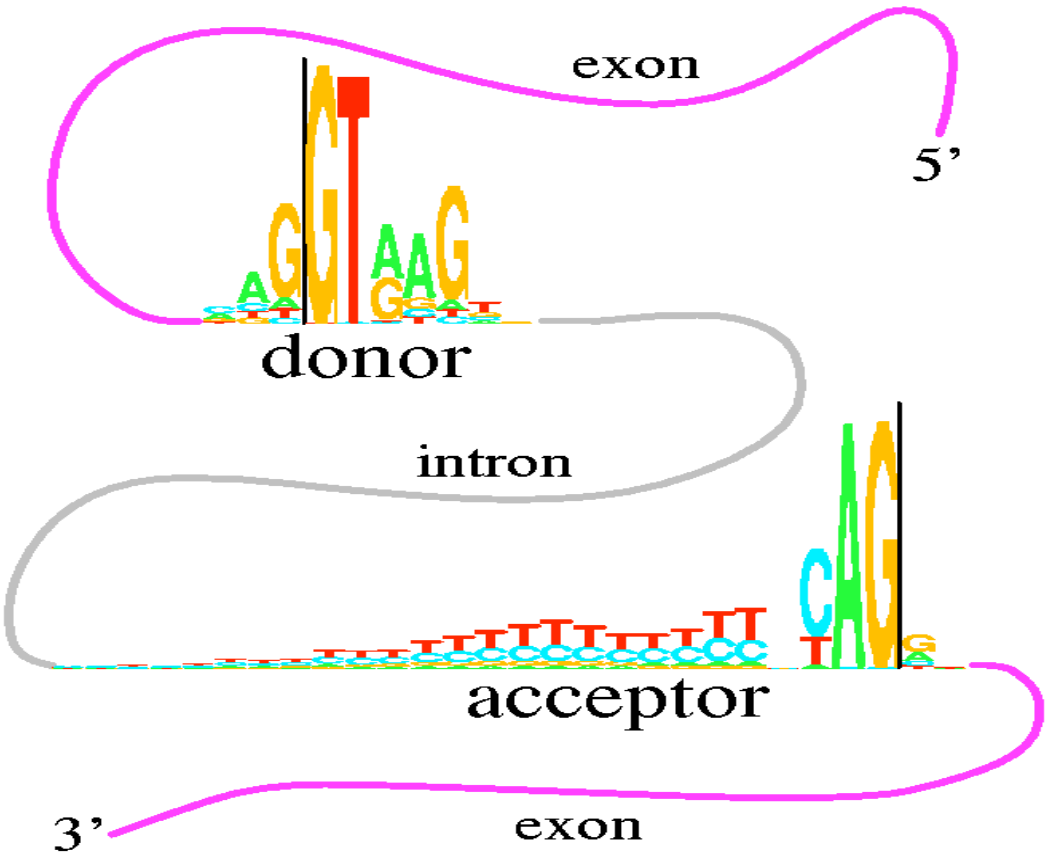
Splice sites (cntd)

%	-8	…	-2	-1	0	1	2	…	17
A	26	…	60	9	0	0	54	…	21
C	26	…	15	5	0	1	2	…	27
G	25	…	12	78	100	0	41	…	27
T	23	…	13	8	0	99	3	…	25

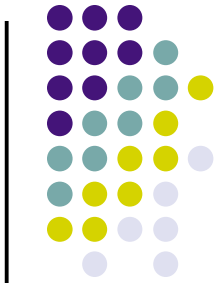


Splice sites (cntd)

This figure shows two "sequence logos" which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern "CAG|GT", which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992)



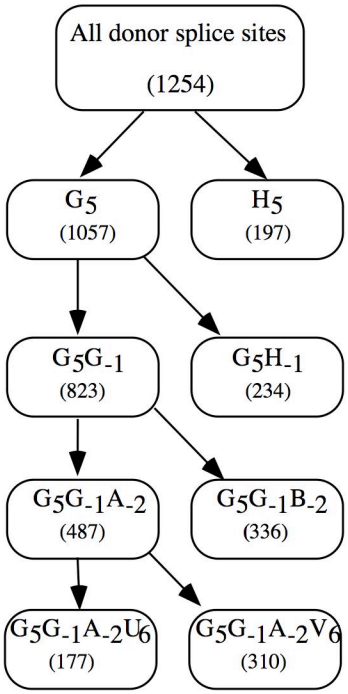
Modeling splice sites



- **WMM:** weight matrix model = PSSM (Staden 1984)
- **WAM:** weight array model = 1st order Markov (Zhang & Marr 1993)
- **MDD:** maximal dependence decomposition (Burge & Karlin 1997)
  - Decision-tree algorithm to take pairwise dependencies into account
  - Train separate WMM models for each subset

Maximal Dependence Decomposition

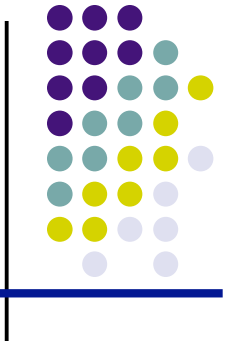
Pos	A%	C%	G%	U%
-3	33	36	19	13
-2	56	15	15	15
-1	9	4	78	9
+3	44	3	51	3
+4	75	4	13	9
+6	14	18	19	49
-3	34	37	18	11
-2	59	10	15	16
+3	40	4	53	3
+4	70	4	16	10
+6	17	21	21	42
-3	37	42	18	3
+3	39	5	51	5
+4	62	5	22	11
+6	19	20	25	36
-3	32	40	23	5
+3	27	4	59	10
+4	51	5	25	19



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'**

Prediction of gene structure



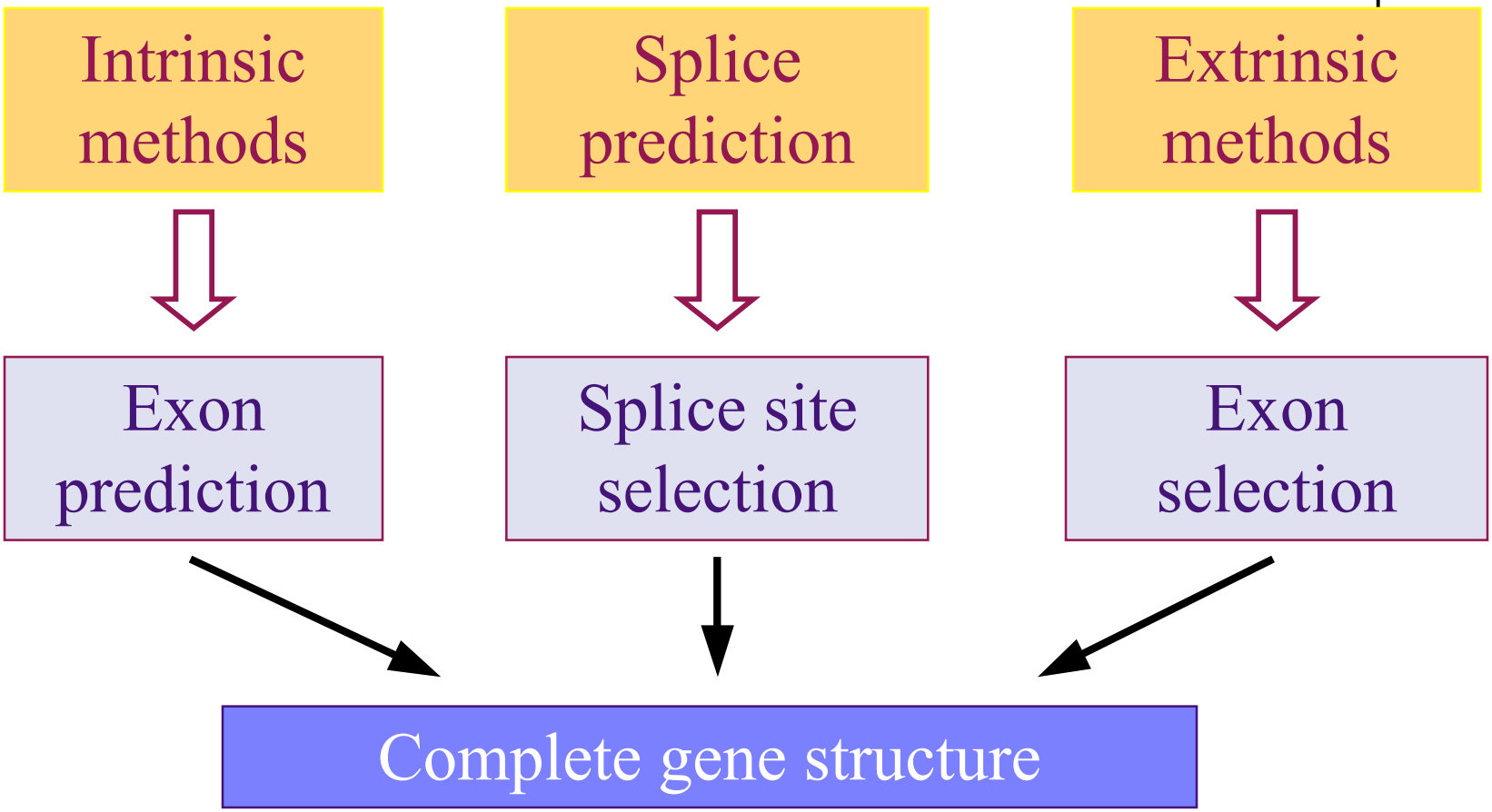
Extrinsic information:

- Alignment with known protein sequences
- Alignment with ESTs or cDNAs
- Alignment with genomic DNA of related species

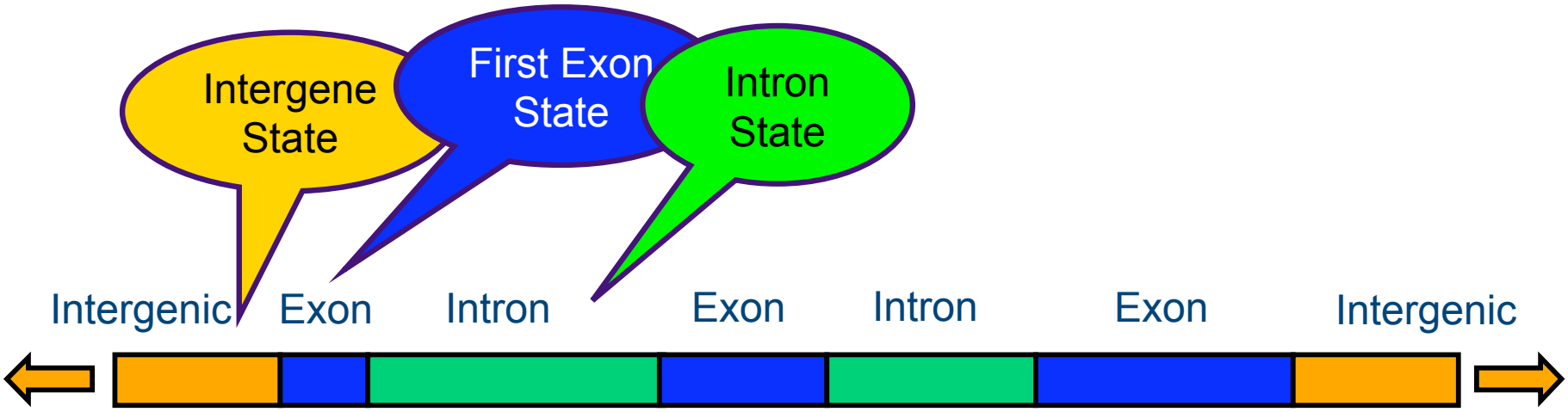
Potential sources of problems:

- Genes with no similarity in the database
- Poor database quality
- Genes expressed in specific conditions or at low levels
- Alternatively spliced genes

All together now…

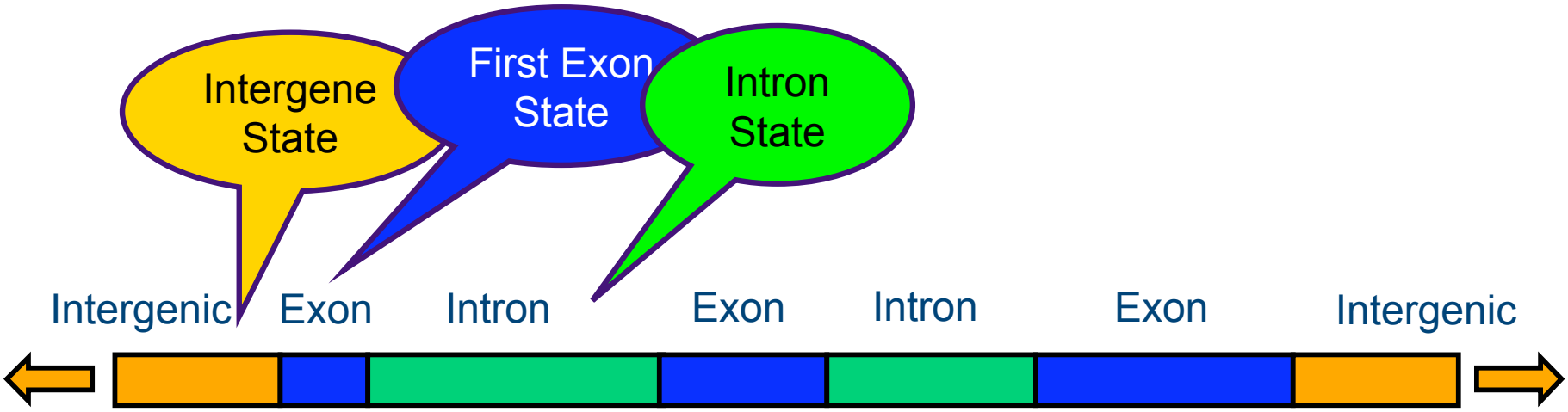


HMMs for Gene Finding



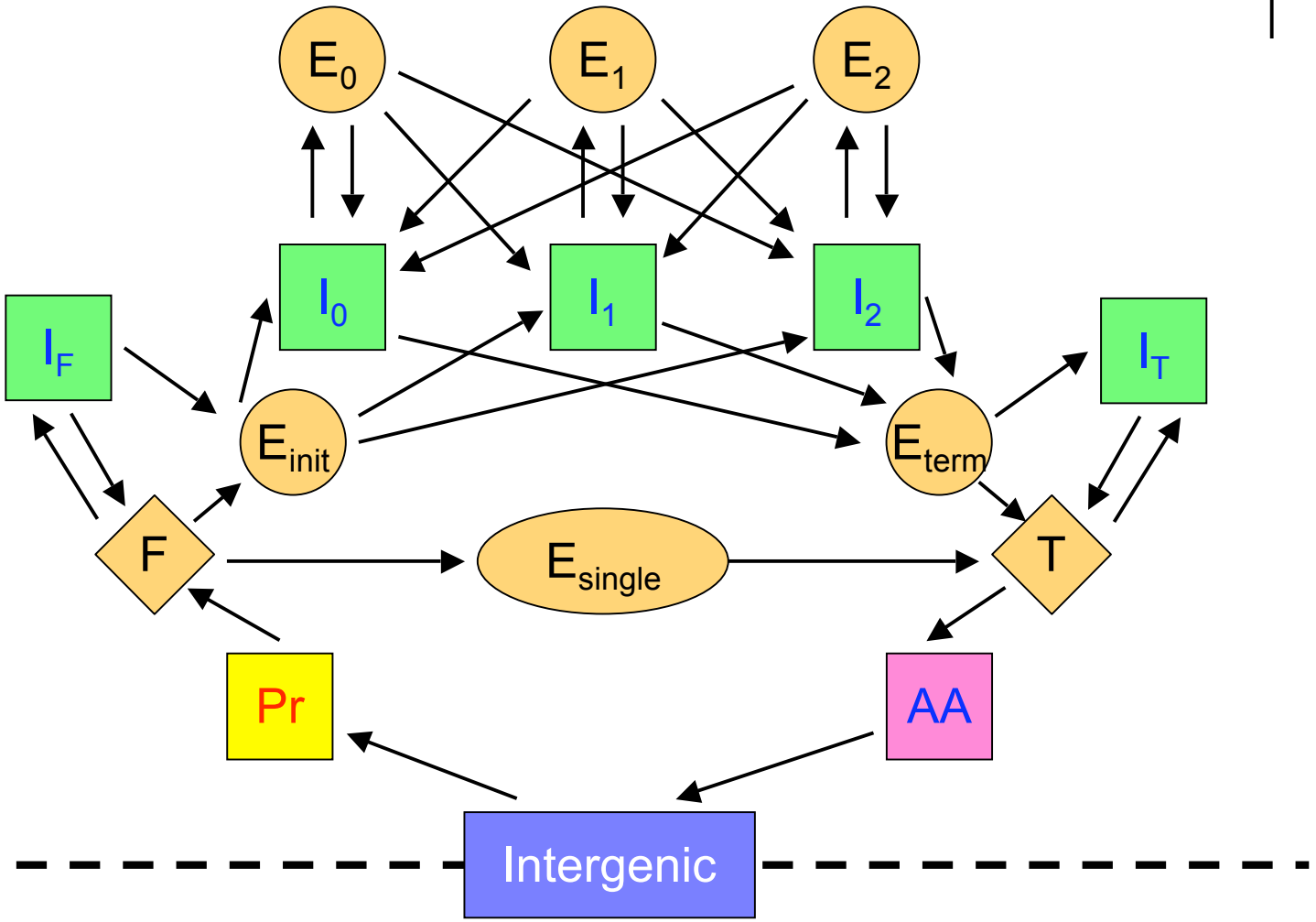
GTCAGATGAGCAAAGTAGACACTCCAGTAACGCGGTGAGTACATTAA

HMMs for Gene Finding

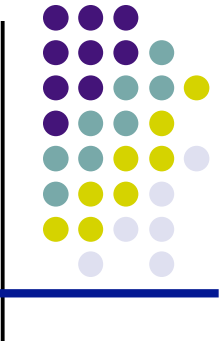


GTCAGATGAGCAAAGTAGACACTCCAGTAACGCGGTGAGTACATTAA

GENSCAN

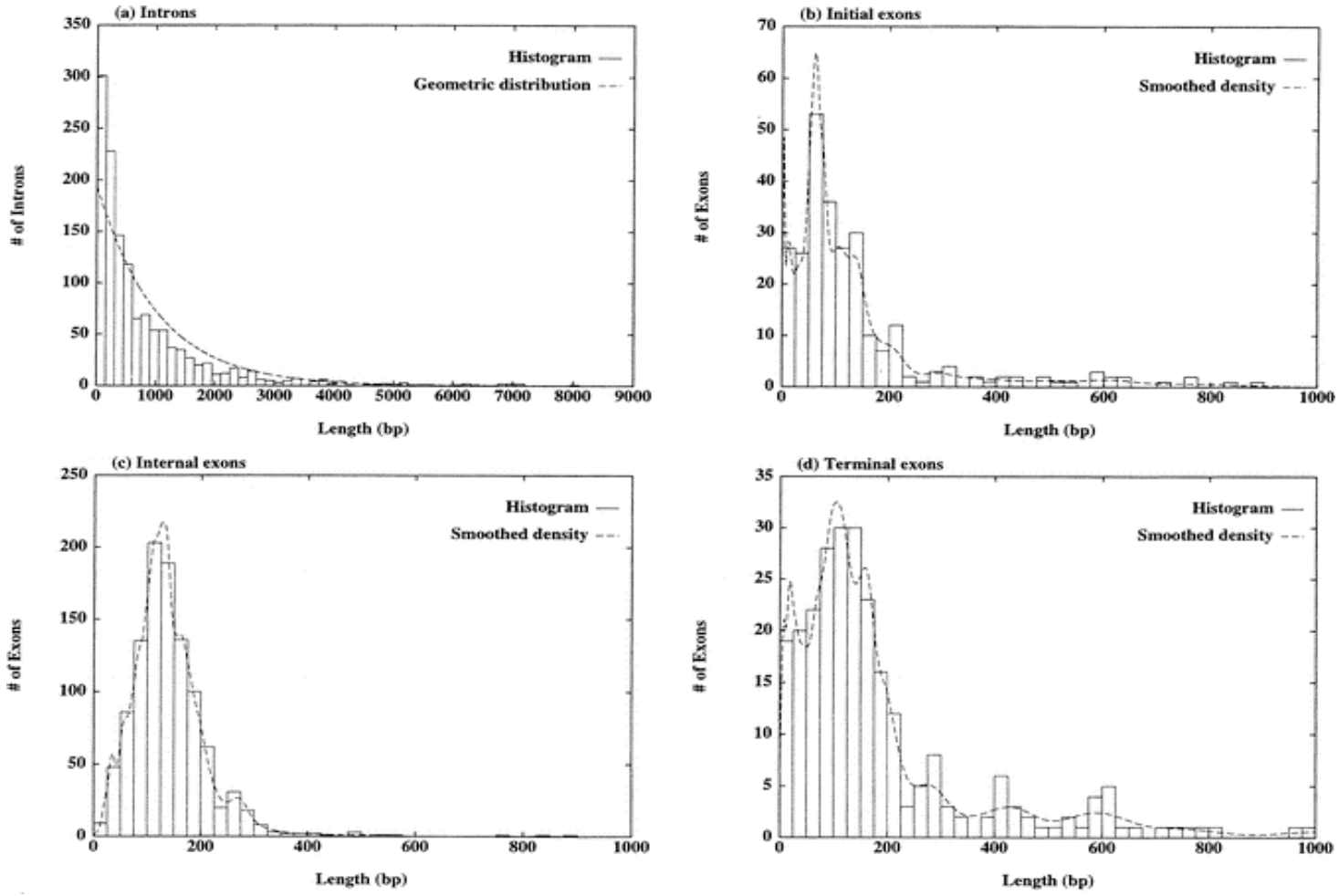


GENSCAN Characteristics

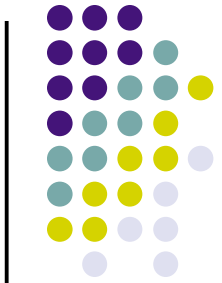


- Burge and Karlin, Stanford, 1997
- Explicit state duration HMM (with tricks)
  - Intergenic and intronic regions have geometric length distribution
  - Exons are only possible when correct flanking sequences are present
- Output probabilities for NC and CDS are 5th-order Markov
- Each CDS frame has its own model
- WAM models for start/stop codons and acceptor sites
- MDD model for donor sites
- Separate parameters for regions of different GC content

Exon/Intron Lengths

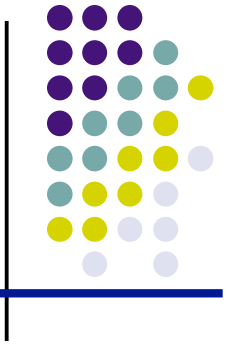


GENSCAN Performance



- Performed well on realistic sequences (complementary to *Genefinder*, Phil Green, University of Washington, *unpublished*)
  - *Genscan*: long, multiple genes in both orientations
  - *Genefinder*: short, multiple genes in both orientations
- Pretty good sensitivity, poor specificity
  - Exon: Sn = 69%; Sp = 33%
  - Exact gene: Sn = 10%; Sp = 4%
- Was the best gene predictor for about 4 years

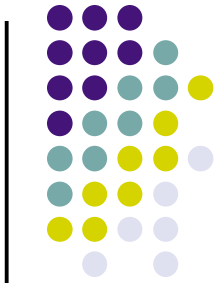
TWINSCAN



- Korf, Flicek, Duan, Brent, Washington University in St. Louis, 2001

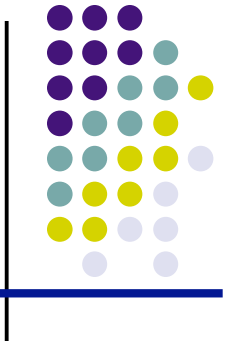
- Uses an informant sequence to help predict genes
  - Informant sequence consists of three states (match, mismatch, unaligned)
  - Informant sequence assumed independent of target sequence

The TWINSCAN Model



- Extends GENSCAN by adding a model for conservation sequence
- 5^{th}-order models for CDS and NC, 2^{nd}-order models for start and stop codons and splice sites
  - One CDS model for all frames
- Mouse was used as informant (the best of those tried)

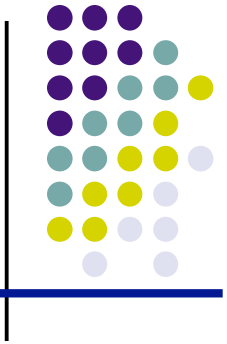
TWINSCAN Performance



- Slightly more sensitive than GENSCAN, much more specific
  - Exon: Sn = 71%; Sp = 61%
  - Exact gene: Sn = 25%; Sp = 15%

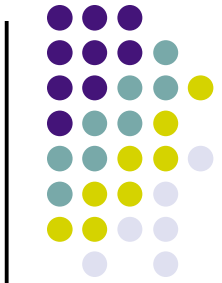
- Was the best gene predictor for about 4 years

N-SCAN



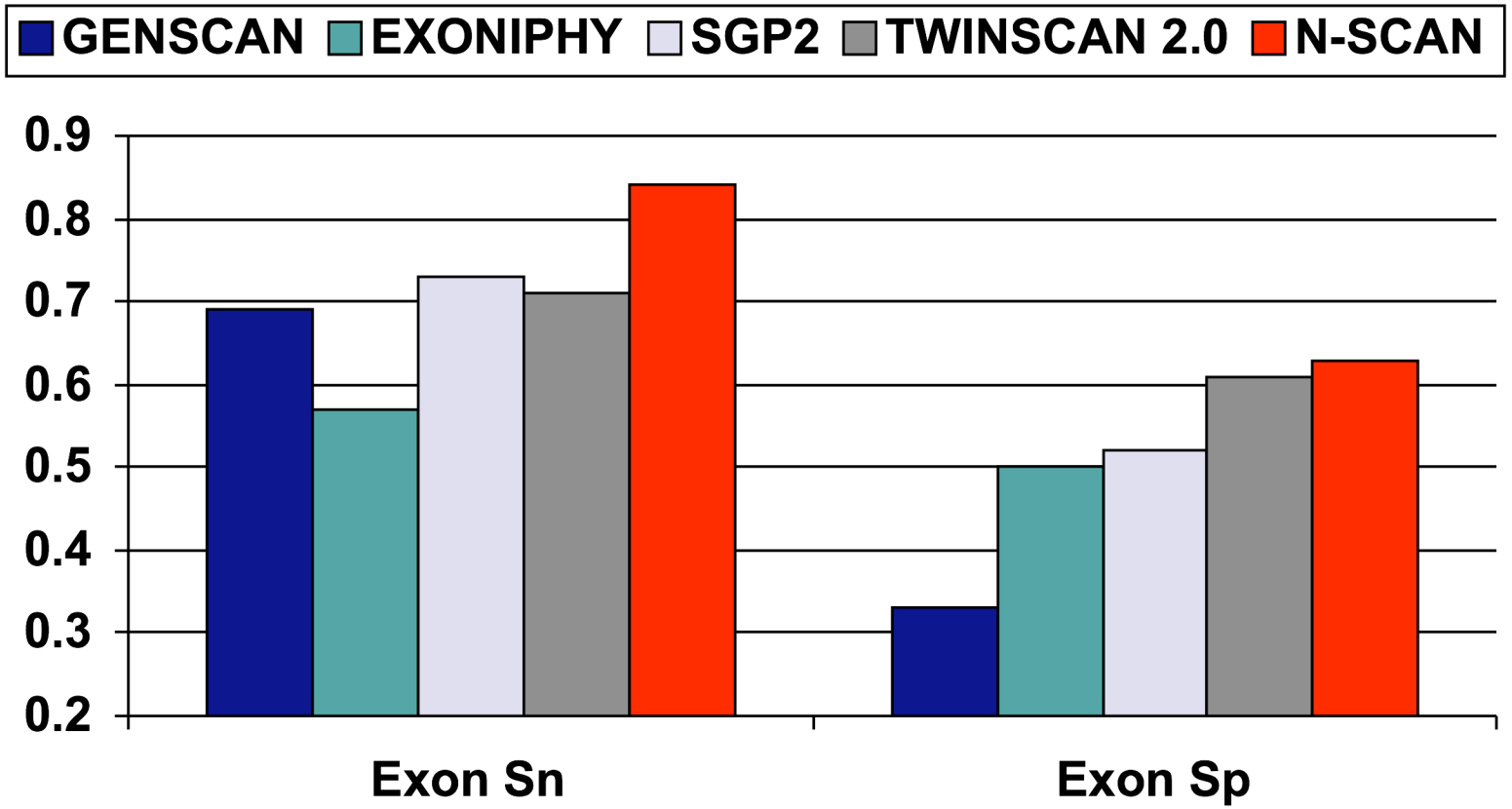
- Brown, Gross and Brent, Washington University in St. Louis, 2005
- Multiple informants
- Richer models of sequence evolution
- Frame-specific CDS conservation model
- Conserved noncoding sequence model
- 5’ UTR structure model

Evaluating Performance

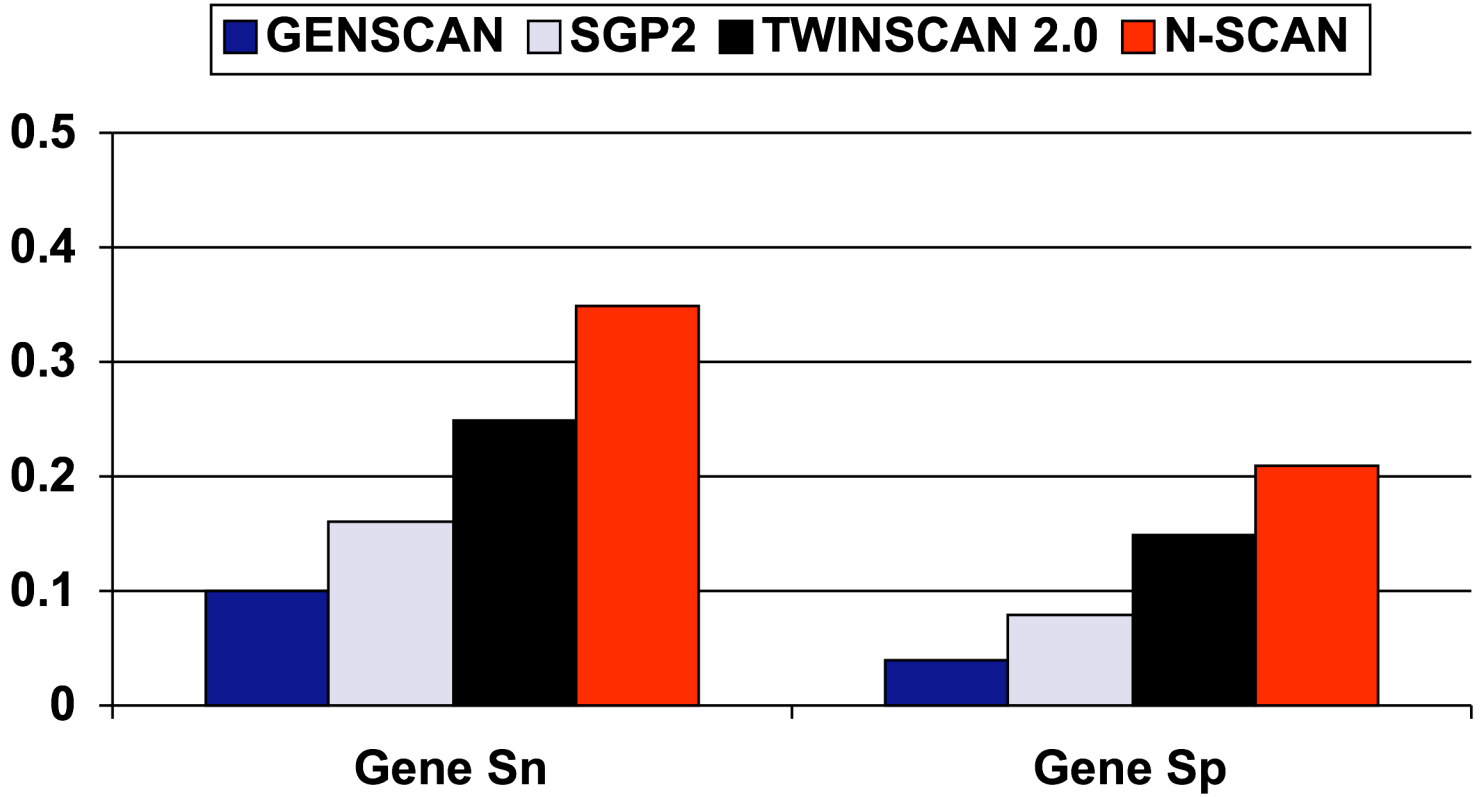


- Three levels of performance: gene, exon, nucleotide
- Two measures of performance:
  - Sensitivity = TP / (TP + FP)
  - Specificity = TP / (TP + FP)
- Testing standard is whole-genome prediction
  - Predicting on single-gene sequences is easier and less interesting

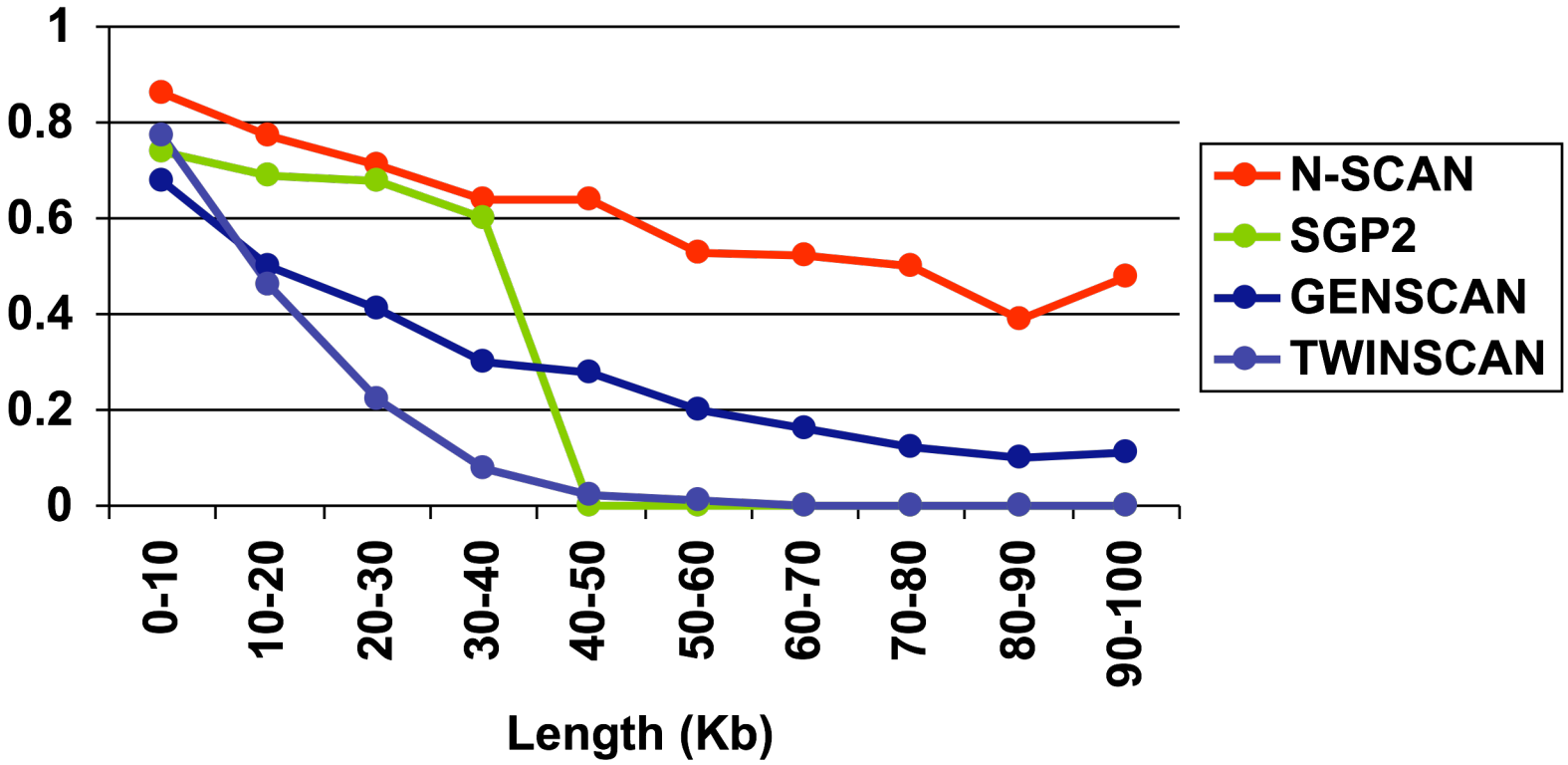
Exact Exon Accuracy



Exact Gene Accuracy

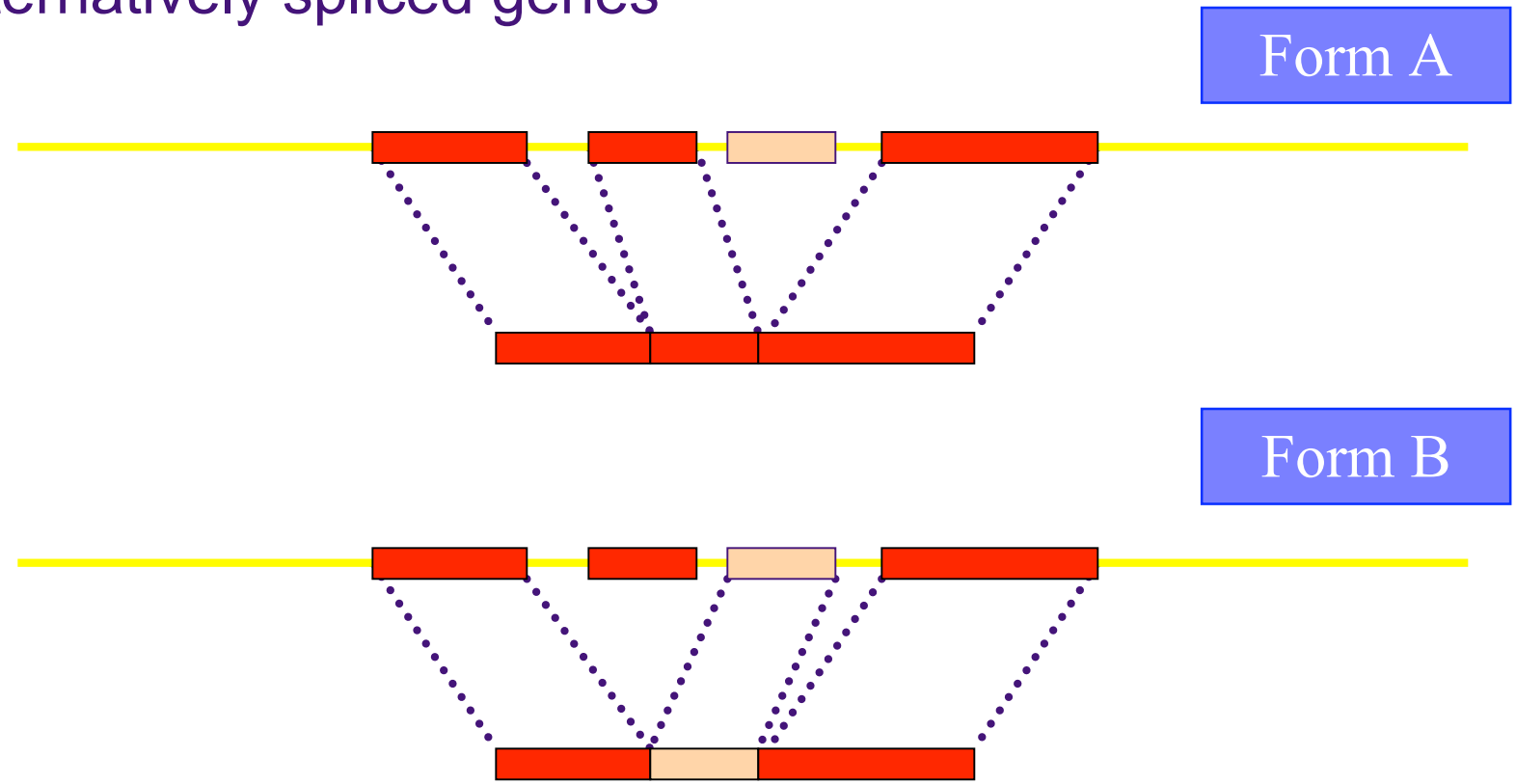


Intron Sensitivity By Length



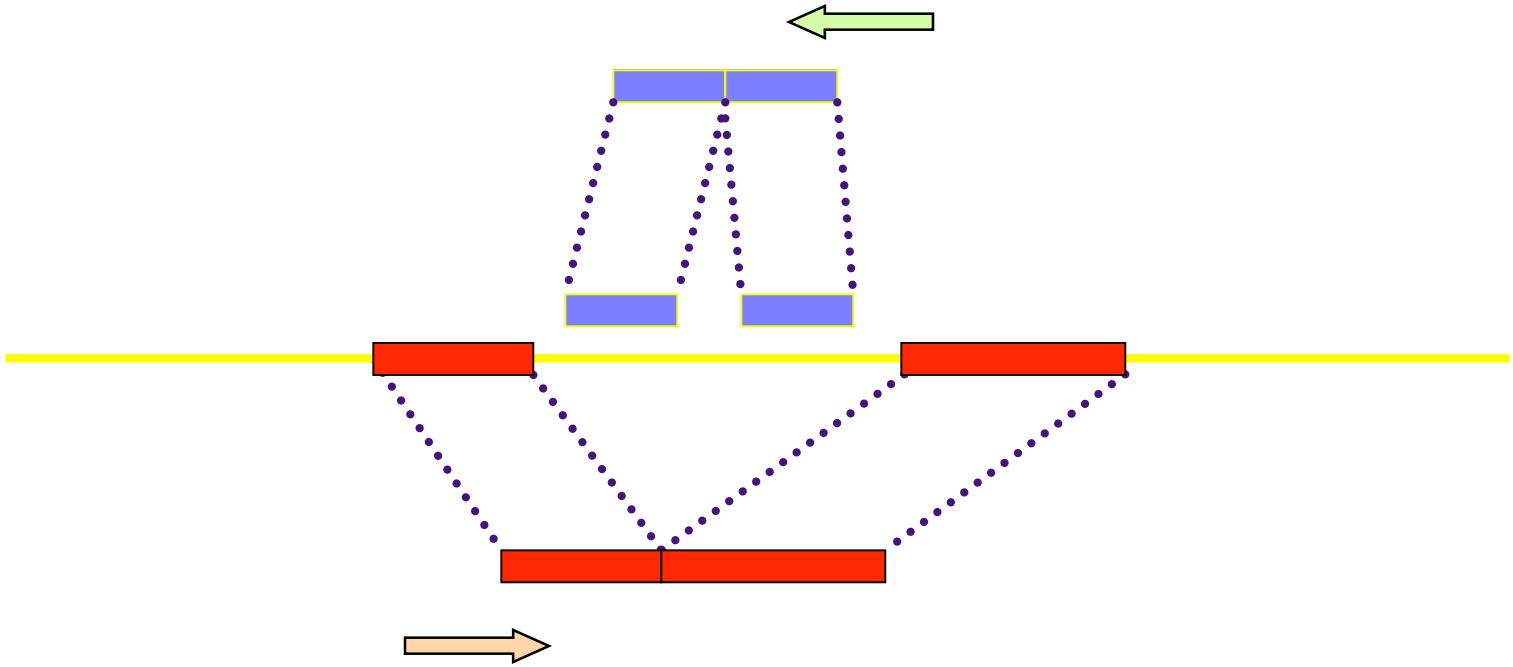
Missing parts…

- Alternatively spliced genes

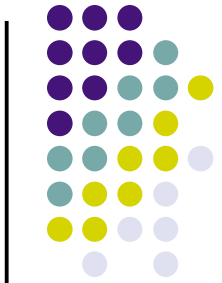


Missing parts…(cntd)

- Genes within genes



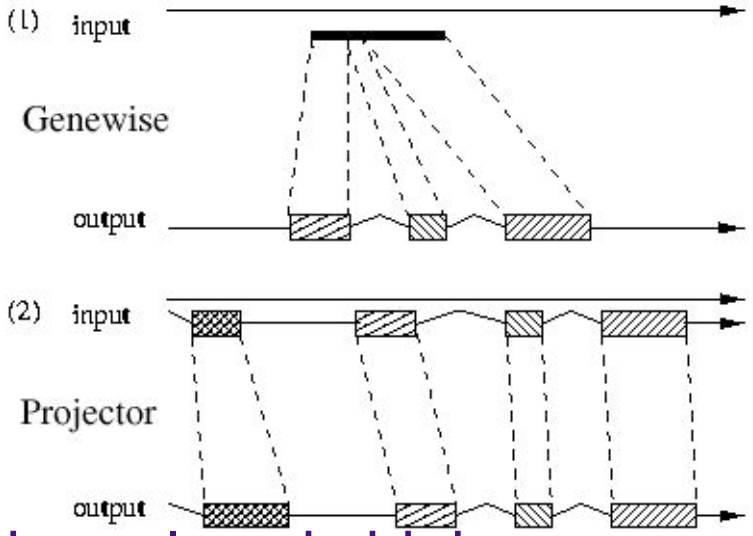
Homology-Based Gene Prediction



- Idea: Try to predict a gene in one organism using a known orthologous gene or protein from another organism

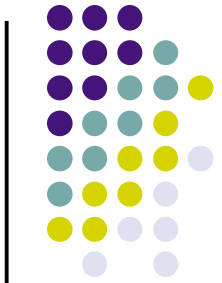
- Genewise
  - Protein homology

- Projector
  - Gene structure homology



- Very accurate if (and only if??) homology is high

For more reading



- GENSCAN: Burge & Karlin, "*Prediction of complete gene structures in human genomic DNA*" (1997) *J Mol Biol* **268:** 78-94
- TWINSCAN: Korf, Flicek, Duan, Brent, "*Integrating genomic homology into gene structure prediction*" (2001) *Bioinformatics* **17** Suppl 1: S140-S148
- N-SCAN: Gross & Brent ,"*Using multiple alignments to improve gene prediction*" (2006) *J Comput Biol* ***13:*** *379-393*

- Fickett & Tung, "*Assessment of protein coding measures*" (1992) *Nuc Acids Res* **20:** 6441-6450
- Rogic *et al.* "Evaluation of Gene-Finding Programs on Mammalian Sequences" (2001) *Genome Res* **11:** 817-832
- Mathe *et al.* "Current methods of gene prediction, their strengths and weaknesses" (2002) *Nuc Acids Res* **30:** 4103-4117

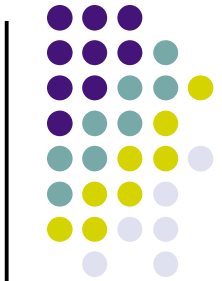
Acknowledgements

Some of the slides used in this lecture are adapted or modified slides from lectures of:

- Serafim Batzoglou, Stanford University

Theory and examples from the following:

- C. Burge, PhD thesis, 1997, Stanford University
- http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html



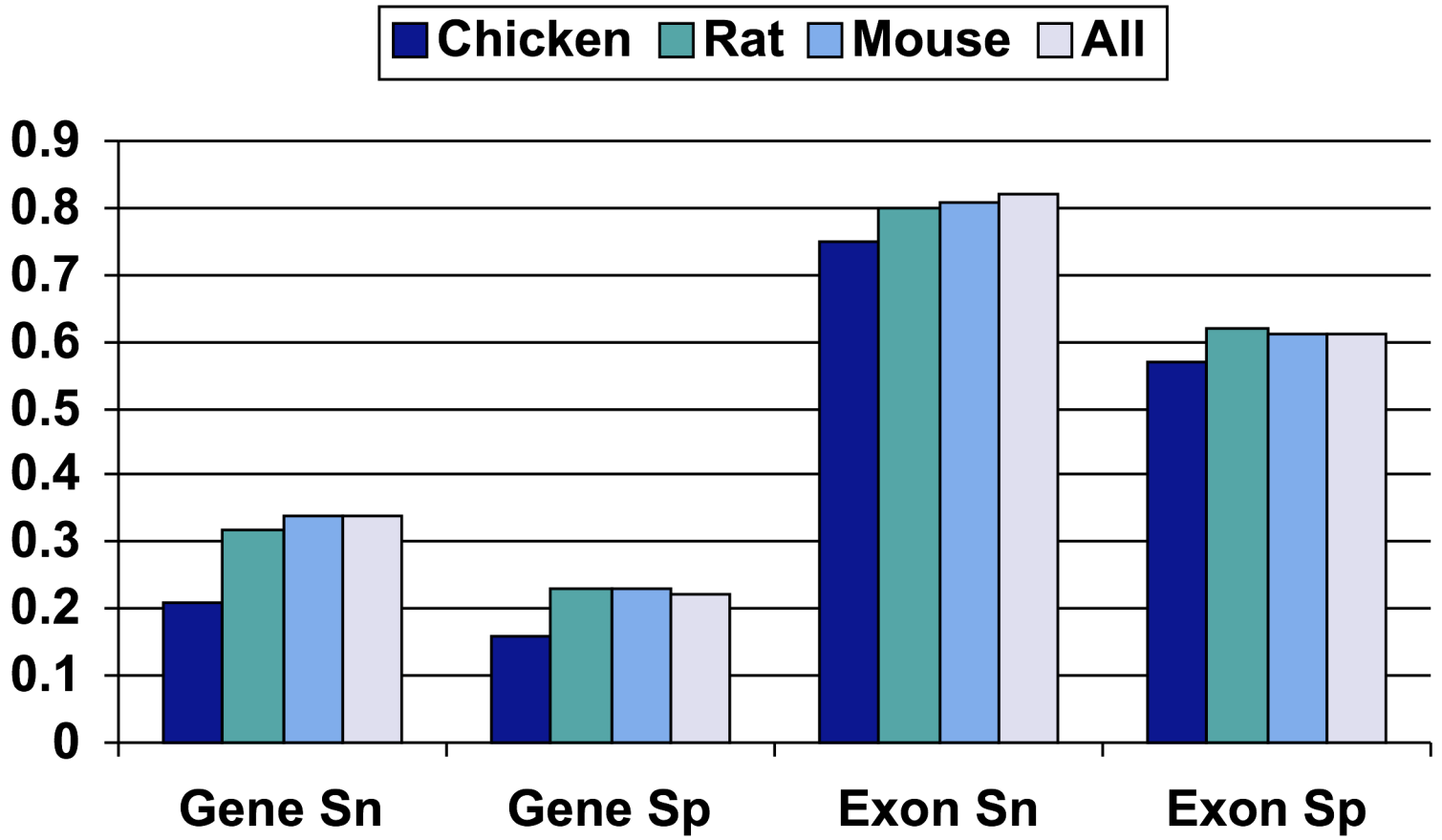
GENSCAN Performance

Programs	No. of sequences	Nucleotide accuracy			
		Sn	Sp	AC	CC
FGENES	195 (5)	0.86	0.88	0.84 ± 0.19	0.83
GeneMark.hmm	195 (0)	0.87	0.89	0.84 ± 0.18	0.83
Genie	195 (15)	0.91	0.90	0.89 ± 0.16	0.88
Genscan	195 (3)	0.95	0.90	0.91 ± 0.12	0.91
HMMgene	195 (5)	0.93	0.93	0.91 ± 0.13	0.91
Morgan	127 (0)	0.75	0.74	0.70 ± 0.21	0.69
MZEF	119 (8)	0.70	0.73	0.68 ± 0.21	0.66

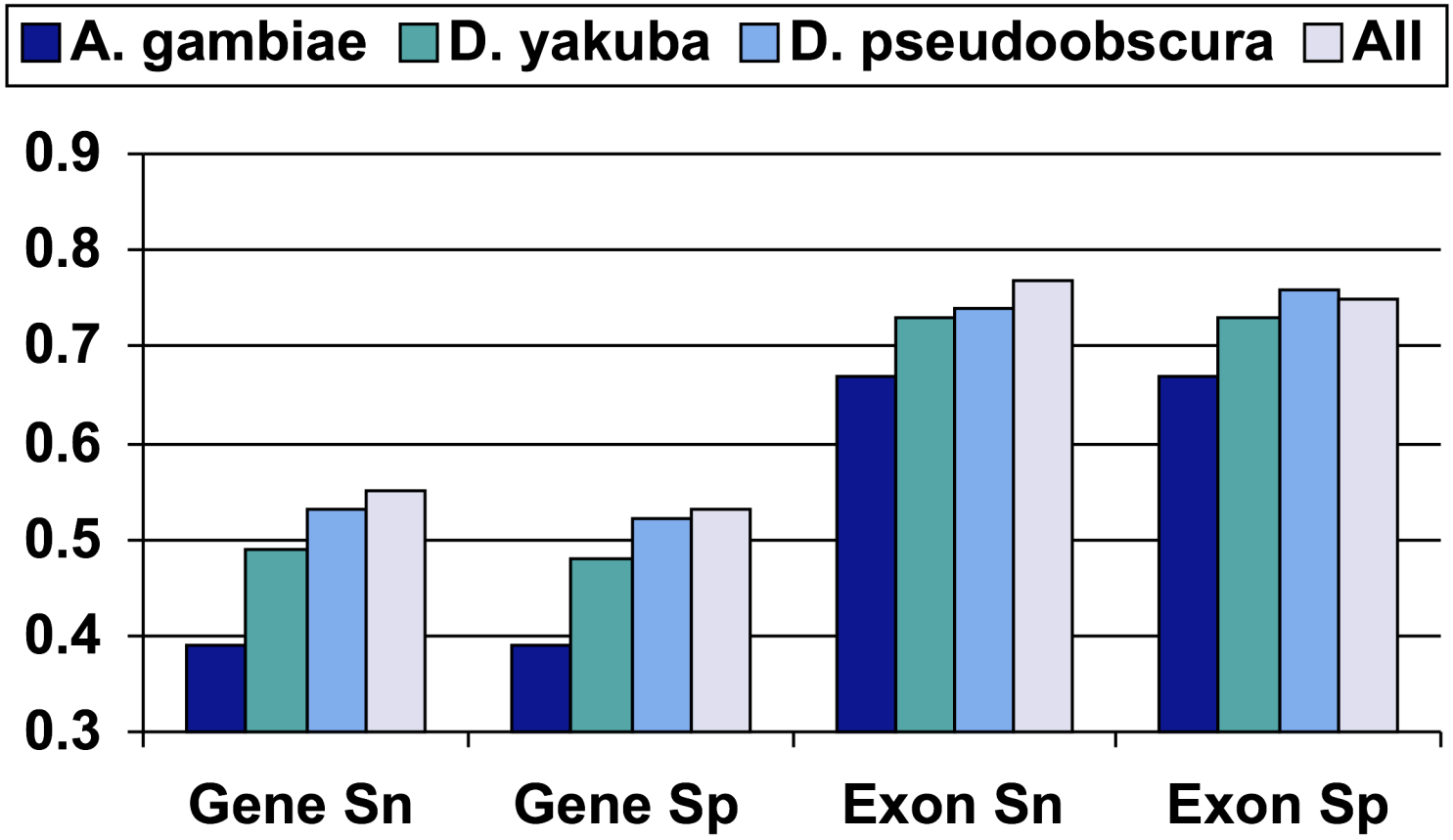
Programs	No. of sequences	Exon accuracy							
		ESn	ESp	(ESn+ESp)/2	ME	WE	PCa	PCp	OL
FGENES	195 (5)	0.67	0.67	0.67 ± 0.32	0.12	0.09	0.20	0.17	0.02
GeneMark.hmm	195 (0)	0.53	0.54	0.54 ± 0.36	0.13	0.11	0.29	0.27	0.09
Genie	195 (15)	0.71	0.70	0.71 ± 0.30	0.19	0.11	0.15	0.15	0.02
Genscan	195 (3)	0.70	0.70	0.70 ± 0.32	0.08	0.09	0.21	0.19	0.02
HMMgene	195 (5)	0.76	0.77	0.76 ± 0.30	0.12	0.07	0.14	0.14	0.02
Morgan	127 (0)	0.46	0.41	0.43 ± 0.26	0.20	0.28	0.28	0.25	0.07
MZEF	119 (8)	0.58	0.59	0.59 ± 0.28	0.32	0.23	0.08	0.16	0.01

Rogic et al. (2001) Genome Res ***11:****817-832*

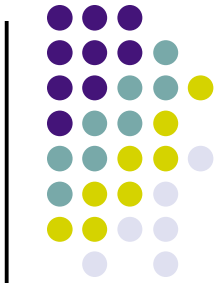
Human Informant Effectiveness



Drosophila Informant Effectiveness

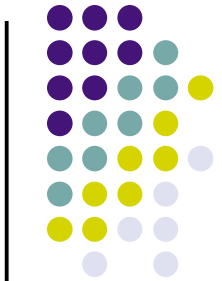


Codon usage measure (cntd)



- Lapedes *et al.* (1990 & 1992)
  - A neural network approach.
  - Uses one hidden layer to summarize codon usage data from a window input.
  - Initially using a 64-space vector, later expanded to include dicodons.

Locating Genes

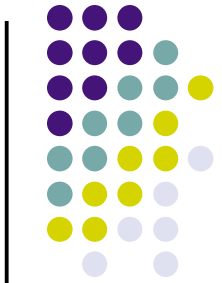


- We have a genome sequence, maybe with related genomes aligned to it…where are the genes?

- Yeast genome is about 70% protein coding
  - About 6000 genes

- Human genome is about 1.5% protein coding
  - About 22,000 genes

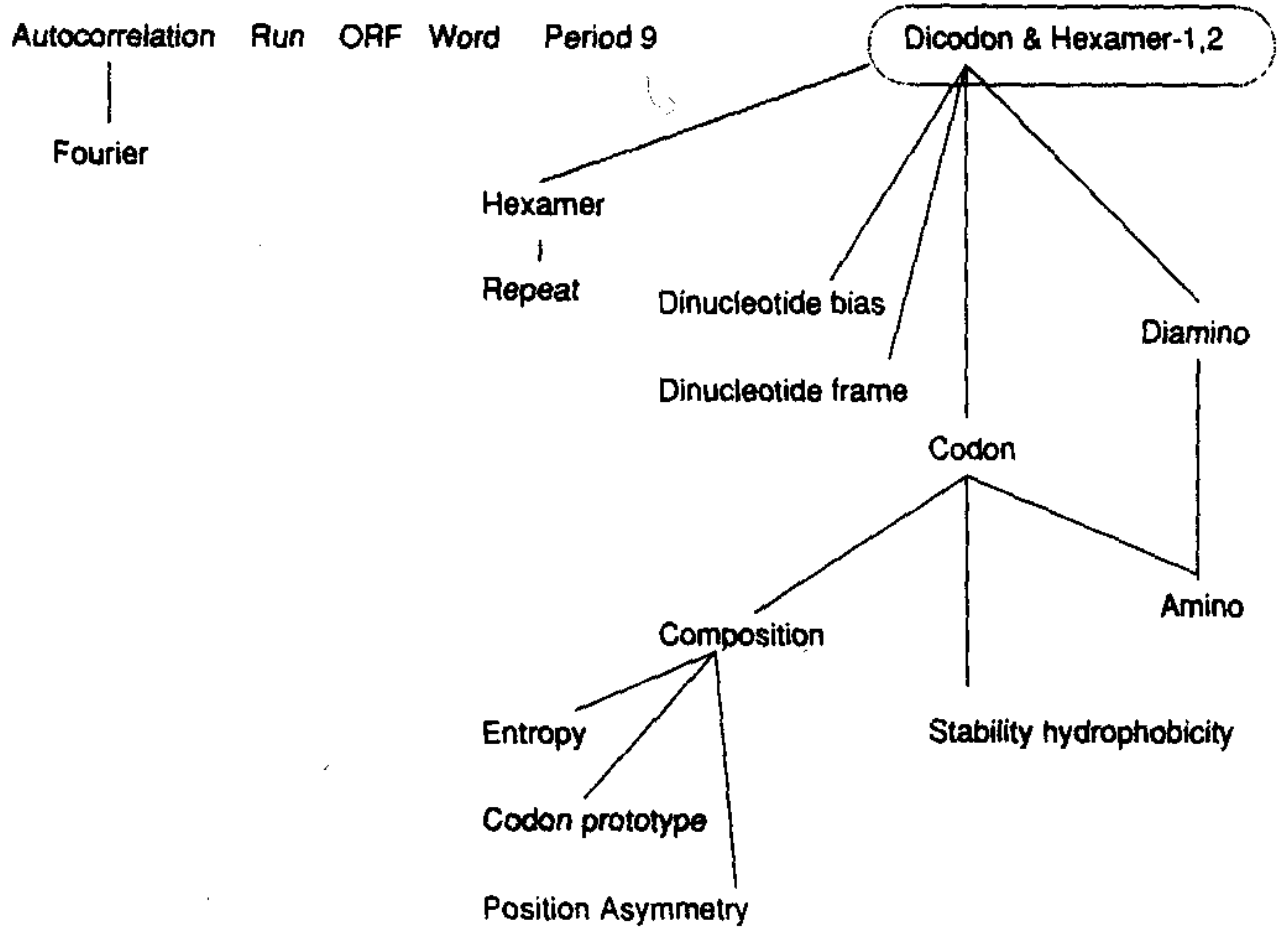
Introns: The Bane of ORF Scanning



- Drosophila:
  - 3.4 introns per gene on average
  - mean intron length 475, mean exon length 397
- Human:
  - 8.8 introns per gene on average
  - mean intron length 4400, mean exon length 165

- ORF scanning is defeated 

Comparison of exon prediction methods



Fickett & Tung (1992) Nuc Acids Res **20**:6441-6450

Comparison of exon prediction methods (cntd)

Table 2. Percentage accuracy (average of specificity and sensitivity) of the coding measures in predicting region coding

Measure	Human 54 Penrose	Human 108 Penrose	Human 162 Penrose	*E.coli* 54 Penrose	Human 54 Classical
Hexamer	70.5	73.1	74.2	67.5	–
Position Asymmetry	70.2	76.6	80.6	61.6	70.3
Dicodon Usage	70.2	72.9	73.9	67.5	–
Fourier	69.9	76.5	80.8	61.3	69.9
Hexamer-1	69.9	72.6	73.8	66.8	–
Hexamer-2	69.9	72.6	73.8	66.7	–
Run	66.6	70.3	71.3	63.6	67.9
Codon Usage	65.2	68.0	69.5	64.1	66.
Repeat	65.1	69.9	73.1	62.4	–
Autocorrelation	64.9	71.1	77.0	58.2	64.9
Dinucleotide Bias	62.9	55.5	50.7	55.9	62.7
Diamino Acid Usage	62.8	66.3	67.8	61.3	–
Composition	61.7	64.1	65.9	61.7	61.3
Amino Acid Usage	60.6	63.4	64.7	59.7	61.3
Word	59.5	66.4	71.4	56.6	61.0
Entropy	58.4	63.1	66.2	55.0	58.4
Dinucleotide Frame	58.0	62.9	66.6	54.6	58.0
Open Reading Frame	57.8	59.2	60.7	57.4	57.8
Stability Hydrophobicity	55.5	57.5	58.7	55.5	55.5
Codon Prototype	54.7	56.1	56.4	54.7	54.7
Period 9	52.5	53.0	52.8	51.8	52.4

Data from five benchmark situations are shown, with varying data set (Human or *E.coli*), window length (54, 108, or 162) and decision method (Penrose discriminant or Classical linear discriminant).

Fickett & Tung (1992) Nuc Acids Res **20**:6441-6450

Comparison of exon prediction methods (cntd)

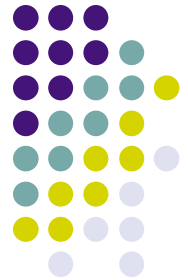


Table 3. Percentage accuracy (average of specificity and sensitivity) of the coding measures in predicting phase-specific coding

Measure	Human 54 Penrose	Human 108 Penrose	Human 162 Penrose	*E.coli* 54 Penrose	Human 54 Classical
Dicodon Usage	80.7	84.3	85.4	88.7	–
Hexamer-2	79.5	82.8	84.2	87.2	–
Hexamer-1	78.6	82.0	83.3	87.1	–
Codon Usage	78.0	81.9	82.1	86.9	81.7
Diamino Acid Usage	77.2	84.9	87.7	84.2	–
Amino Acid Usage	75.3	81.1	83.6	83.3	76.2
Codon Prototype	74.3	78.2	80.5	78.8	74.3
Open Reading Frame	72.9	83.3	88.0	75.6	72.9
Composition	72.2	74.7	75.9	78.8	75.0
Hexamer	71.7	74.3	75.4	70.5	–
Position Asymmetry	68.1	74.7	77.5	59.7	68.3
Fourier	67.8	74.8	77.6	54.7	67.5
Run	66.1	69.6	71.1	62.5	67.0
Repeat	65.5	70.4	73.8	63.0	–
Autocorrelation	64.5	71.4	76.3	58.6	64.6
Dinucleotide Bias	61.9	56.4	55.5	50.3	61.4
Entropy	61.1	64.7	69.2	56.2	61.2
Stability Hydrophobicity	59.8	62.5	63.8	60.4	59.8
Word	58.4	65.6	72.9	57.6	60.7
Dinucleotide Frame	58.4	62.6	65.7	52.1	56.5
Period 9	55.0	58.4	58.9	53.9	55.0

Data from five benchmark situations are shown, with varying data set (Human or *E.coli*), window length (54, 108, or 162) and decision method (Penrose discriminant or Classical linear discriminant).

Fickett & Tung (1992) Nuc Acids Res **20**:6441-6450