Genome Characteristics and Annotation

COMP 571 - Spring 2015

Luay Nakhleh, Rice University

Outline

* Gene prediction in prokaryotic genomes
* Features used in eukaryotic gene detection
* Predicting eukaryotic gene signals
* Complete eukaryotic gene models
* Genome annotation

Gene Prediction in Prokaryotic Genomes

A simple gene structure

* Although introns do exist in prokaryotes, they are extremely rare and often ignored by gene prediction tools.

* The relative simplicity of bacterial gene structure has led to some very successful gene prediction techniques that use functional signals, such as the ribosome-binding site, the stop codon that signals the end of translation, and other well-defined features.

Illustration

* One can easily enumerate all potential open reading frames (ORFs) present in the genome.

* The longer the potential ORF, the more likely it is to really be a gene.

* A key problem then is to distinguish the true and false genes in the set of short potential ORFs of, say, 150 bases or fewer.

* A consequence of this situation is that by accepting some false positives, a gene detection method can achieve very high rates of detection!

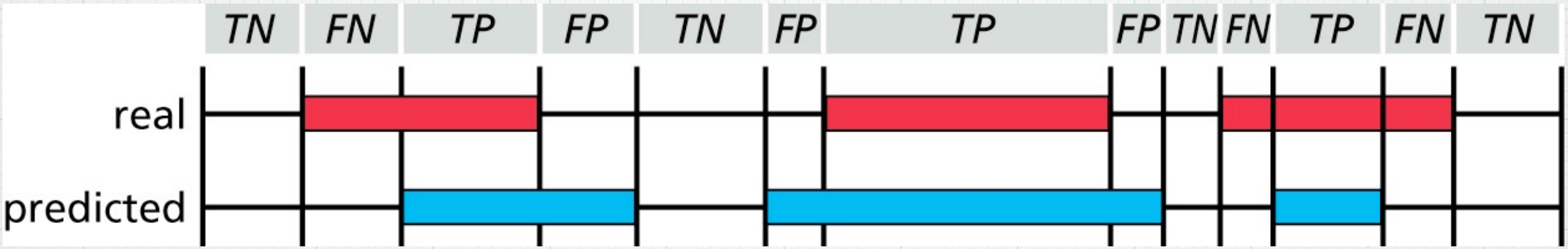
* Put another way, these methods should really be detecting the false ORFs.

* One must be wary of some of the high success rates quoted (even over 98%), and false positive rates would be more informative, but are often not quoted.

Measures of Gene Prediction Accuracy

* In the field of gene prediction, accuracy can be measured at three different levels:
  * Coding nucleotides: The base level
  * Exon structure: The exon level
  * Protein product: The protein level

Measures of Gene Prediction Accuracy at the Nucleotide Level

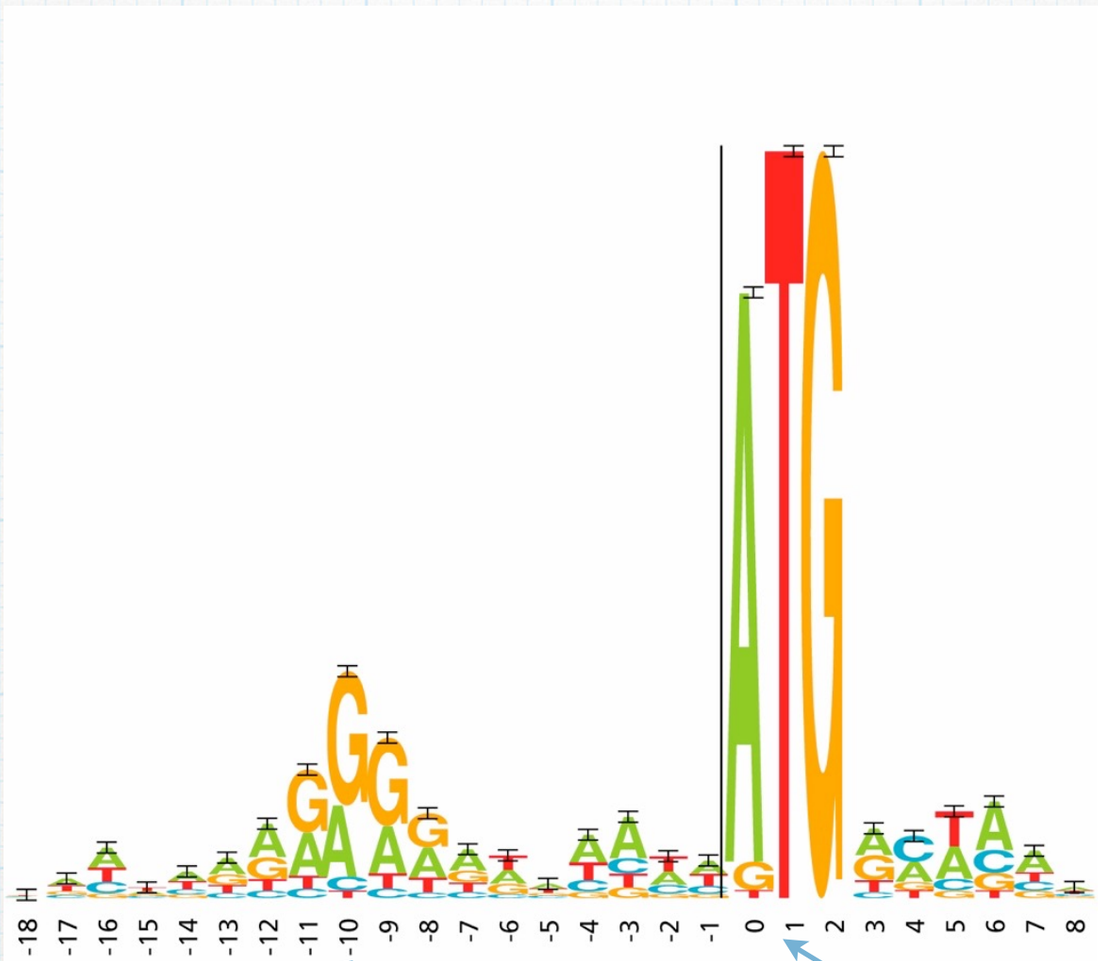


$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TP}{TP + FP}$$

* The most basic characteristic of a gene is that it must contain an open reading frame (ORF) that begins with a start codon (ATG) and ends with a stop codon (TAA, TAG, or TGA).

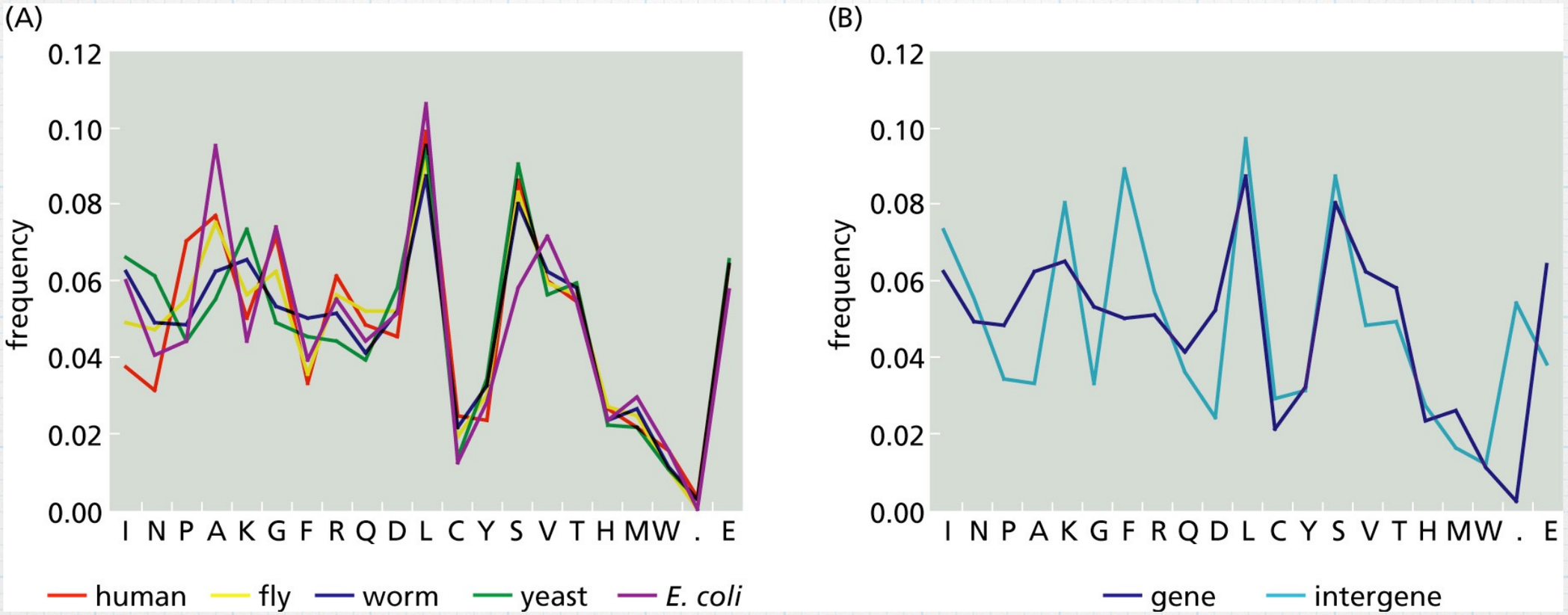
* There are some exceptions (for example, E. coli uses GTG for 9% and TTG for 0.5% of start codons).



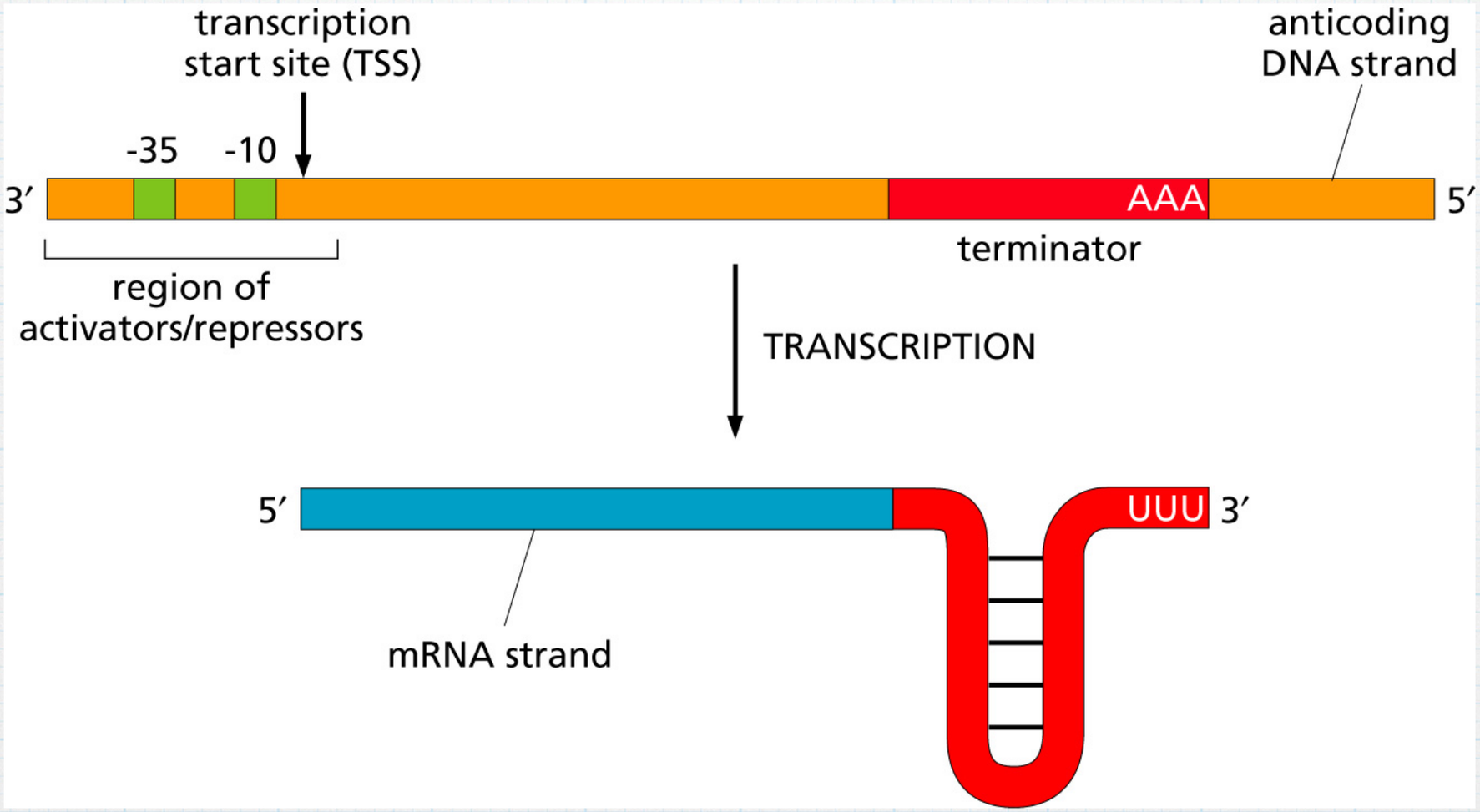
Ribosome-binding site

Start codon in E. coli genes

* Another characteristic that can be used to detect genes is the relative frequency of codon occurrences.


(A)
0.12
0.10
0.08
0.06
0.04
0.02
0.00
frequency
I N P A K G F R Q D L C Y S V T H M W . E
human
fly
worm
yeast
E. coli
(B)
0.12
0.10
0.08
0.06
0.04
0.02
0.00
frequency
I N P A K G F R Q D L C Y S V T H M W . E
gene
intergene

Gene Structure in Prokaryotes



* Bacterial promoters typically occur immediately before the position of the **transcription start site** (TSS), and contain two characteristic short sequences, or motifs, that are almost the same in the promoters for different genes.

* The termination of transcription is controlled by the **terminator signal** which in bacteria differs from the promoter is that it is active when transcribed to form the end of the mRNA strand (forms a loop structure that prevents the transcription apparatus from continuing).

* Single type of RNA polymerase transcribes all genes.

Algorithms for Gene Detection in Prokaryotes

* GeneMark
* GeneMark.hmm
* GLIMMER
* ORPHEUS
* ...

GeneMark

* GeneMark uses a fifth-order Markov chain model to represent the statistics of coding and noncoding reading frames.
* The method uses the dicodon statistics to identify coding regions.

GeneMark

$$\mathbf{P}(a|b_1b_2b_3b_4b_5) = \frac{n_{b_1b_2b_3b_4b_5a}}{\sum_{\alpha\in\{A,C,T,G\}} n_{b_1b_2b_3b_4b_5\alpha}}$$

The number of times $b_1b_2b_3b_4b_5\alpha$ occurs in the training data

GeneMark assumes each reading frame has unique dicodon statistics, and thus has its own model probabilities P1, P2, P3, P4, P5, P6.

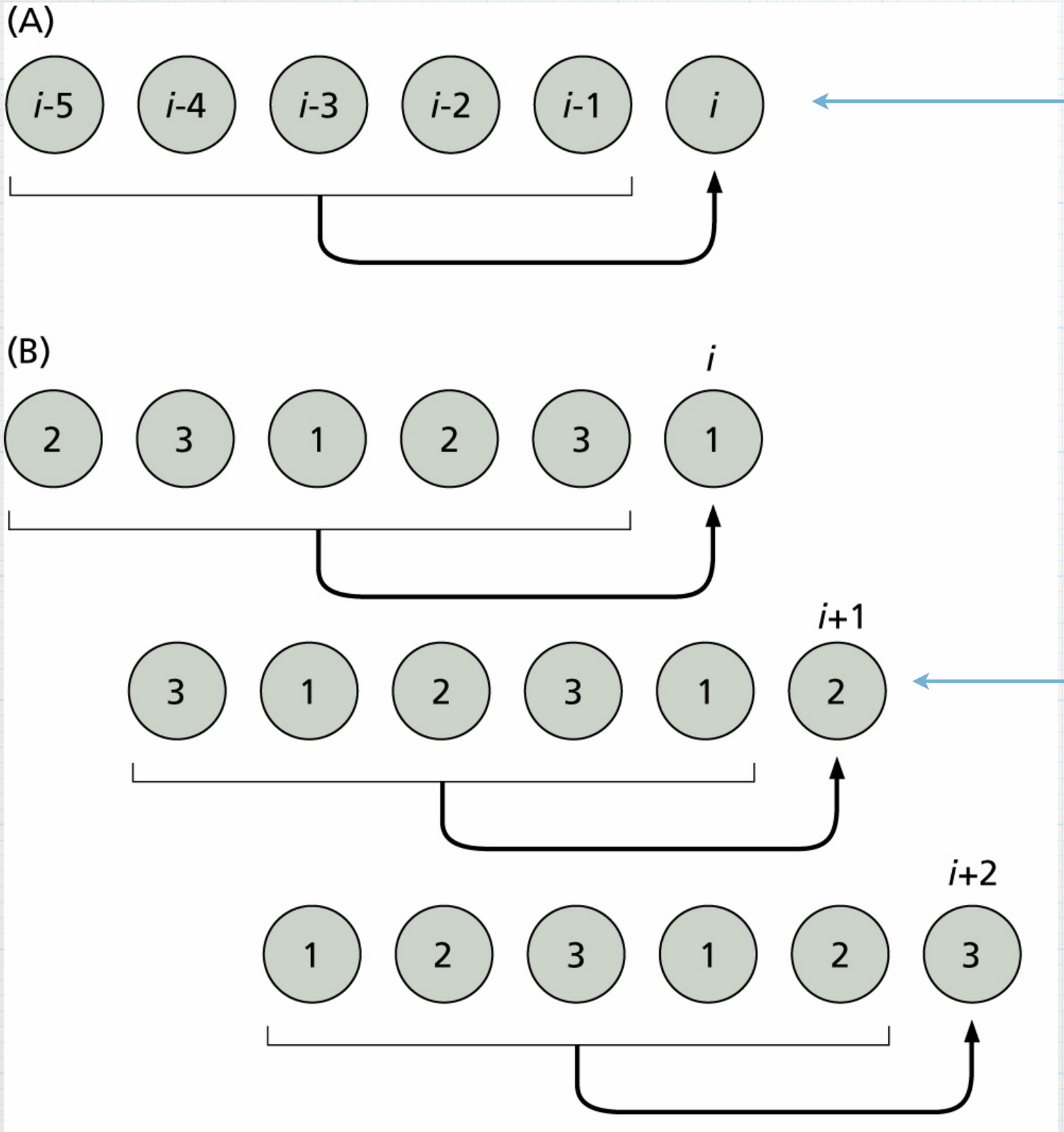
For noncoding regions, there is $P_{nc}(a|b_1b_2b_3b_4b_5)$.

GeneMark

* For example, the probability of obtaining a sequence $x=x_1x_2...x_9$ if $x_1x_2x_3$ is a translated codon (that is, x9 is in the third position of a translated codon) is given by

$$\begin{aligned}\mathbf{P}(x|3) &= \mathbf{P_2}(x_1x_2x_3x_4x_5)\mathbf{P_2}(x_6|x_1x_2x_3x_4x_5)\mathbf{P_3}(x_7|x_2x_3x_4x_5x_6) \\ &\quad \times\mathbf{P_1}(x_8|x_3x_4x_5x_6x_7)\mathbf{P_2}(x_9|x_4x_5x_6x_7x_8)\end{aligned}$$

This is called a periodic, phased, or inhomogeneous Markov model.



homogeneous

inhomogeneous

GeneMark

* We want P(3|x), which can be derived using Bayes rule as

$$\mathbf{P}(3|x) = \frac{\mathbf{P}(x|3)\mathbf{P}(3)}{\mathbf{P}(x|nc)\mathbf{P}(nc) + \sum_{i=1}^{6} \mathbf{P}(x|i)\mathbf{P}(i)}$$

* Similar formulas can be derived for P(i|x) for any value of i.

GeneMark

* In GeneMark, P(nc) was assumed to be 1/2, and P(1)-P(6) were assumed to all be 1/12.
* Sliding windows of 96 nucleotides were scored in steps of 12 nucleotides.
* If P(i|x) exceeds a certain threshold, the window is predicted to be in coding reading frame i.

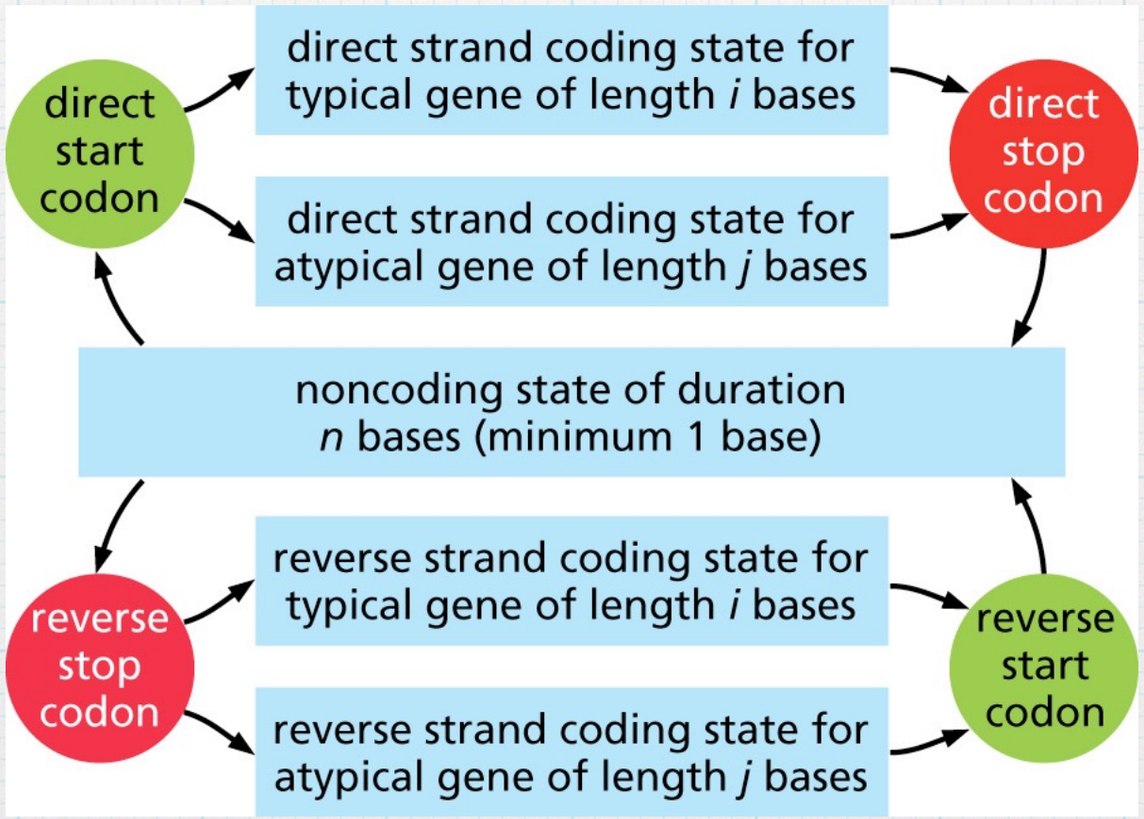
GeneMark

* The final predicted gene boundaries are defined by start and stop codons in that reading frame.

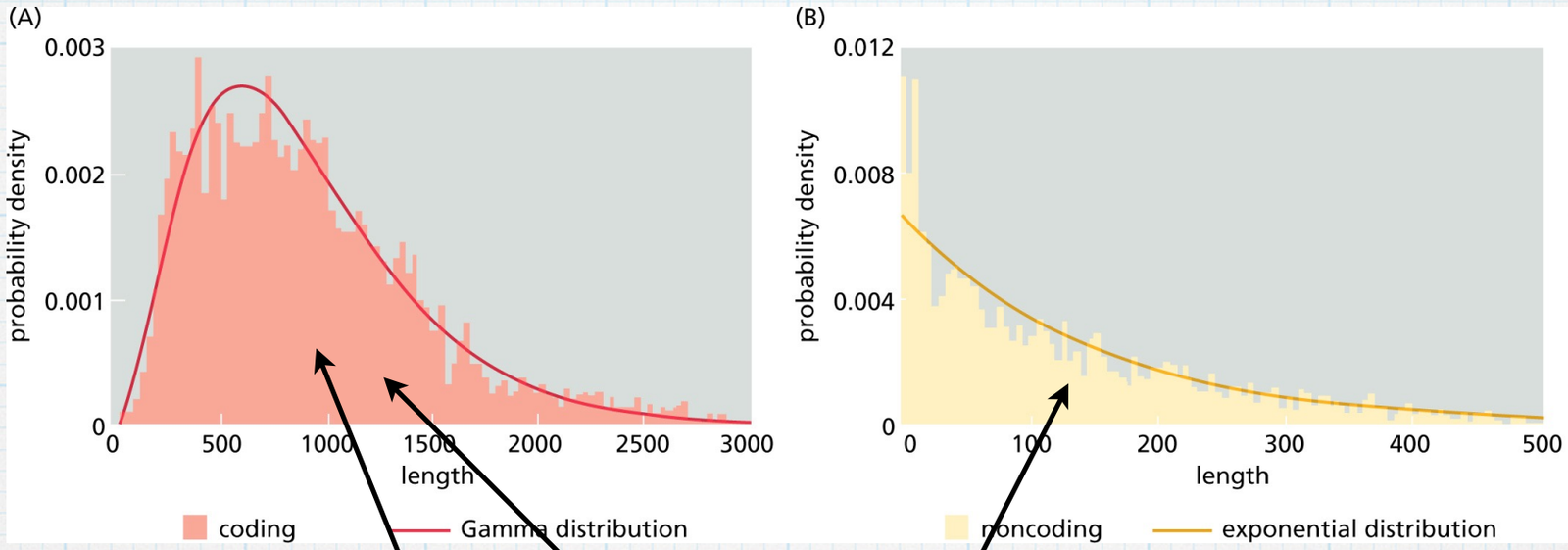
GeneMark.hmm

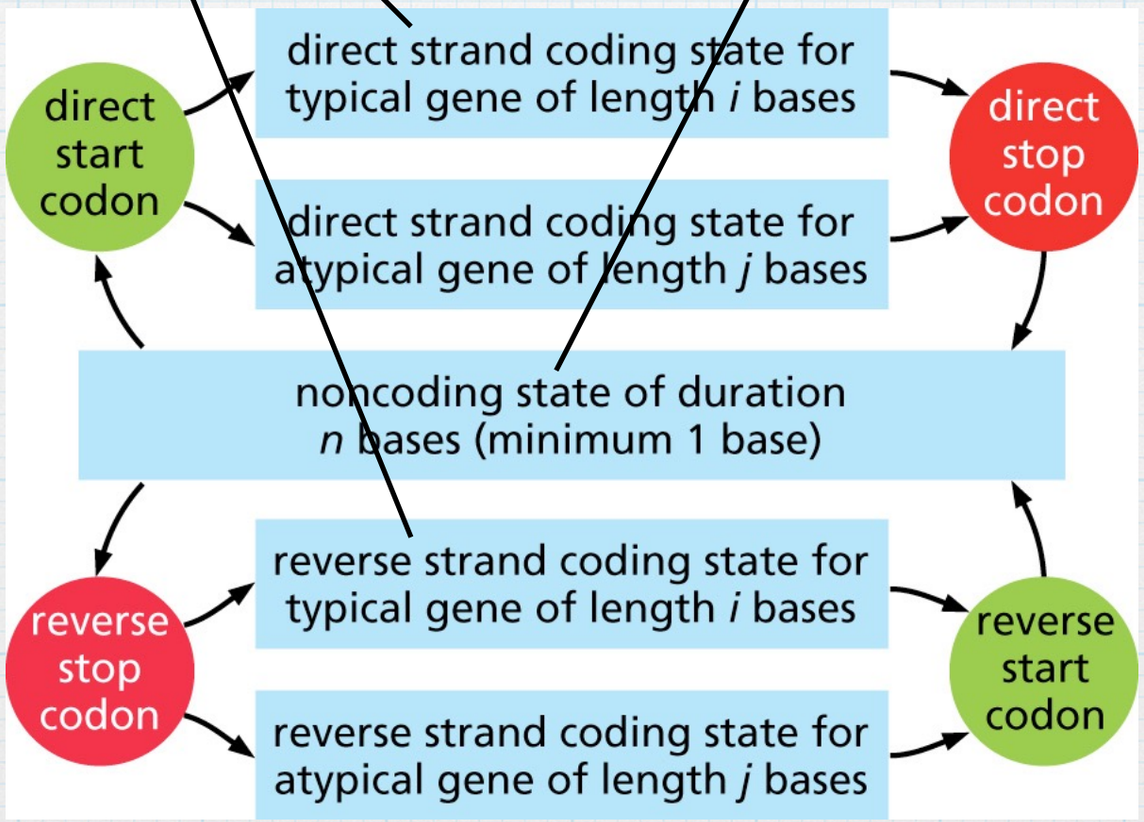
* GeneMark uses a sliding window, and doesn't do a good job at defining the gene boundaries.
* GeneMark.hmm is an extension to ameliorate these issues.

GeneMark.hmm



GeneMark.hmm





Features Used in Eukaryotic Gene Detection

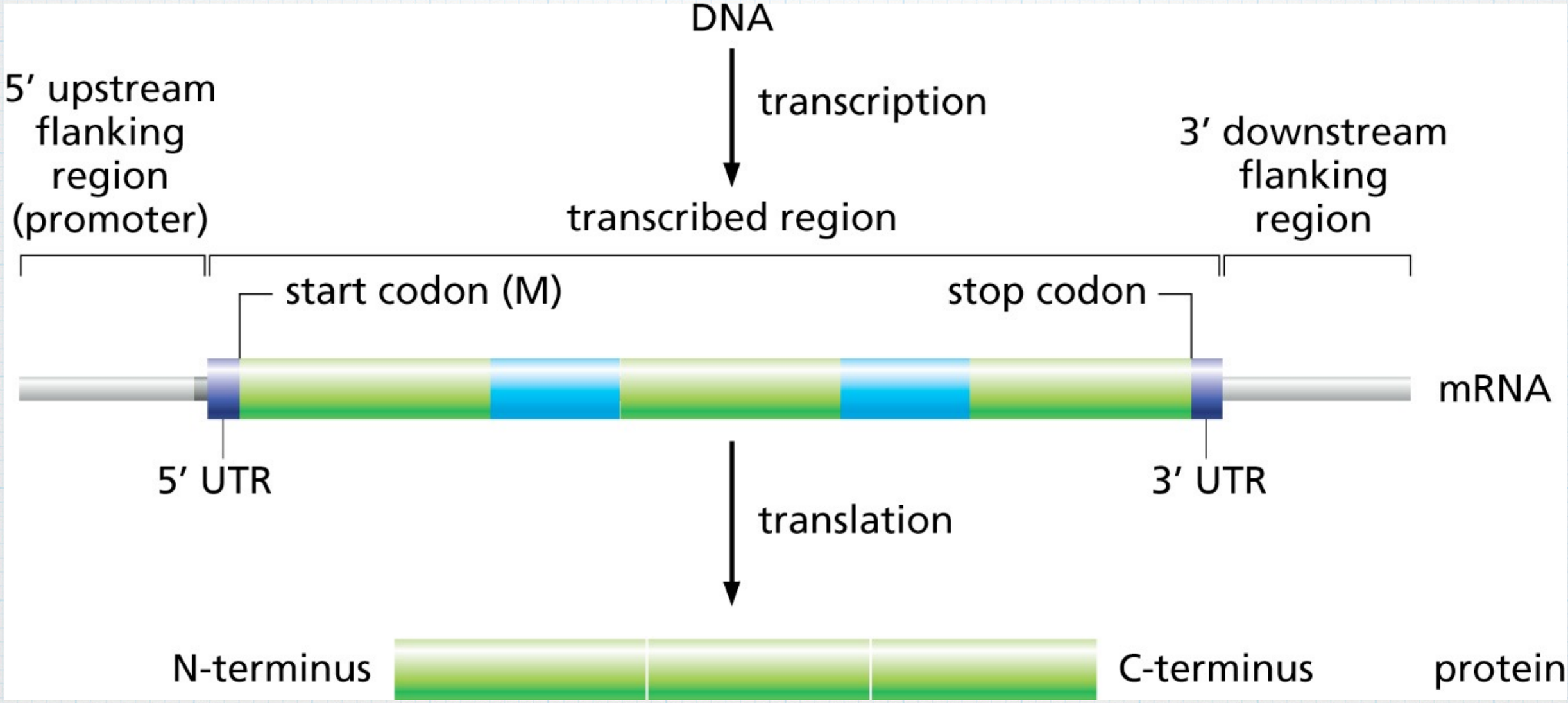
* Many of the principles that apply to the detection of genes in prokaryotes also apply to gene finding in eukaryotes.

* For example, the coding regions of eukaryotic genomes have distinct base statistics similar to those found in prokaryotes.

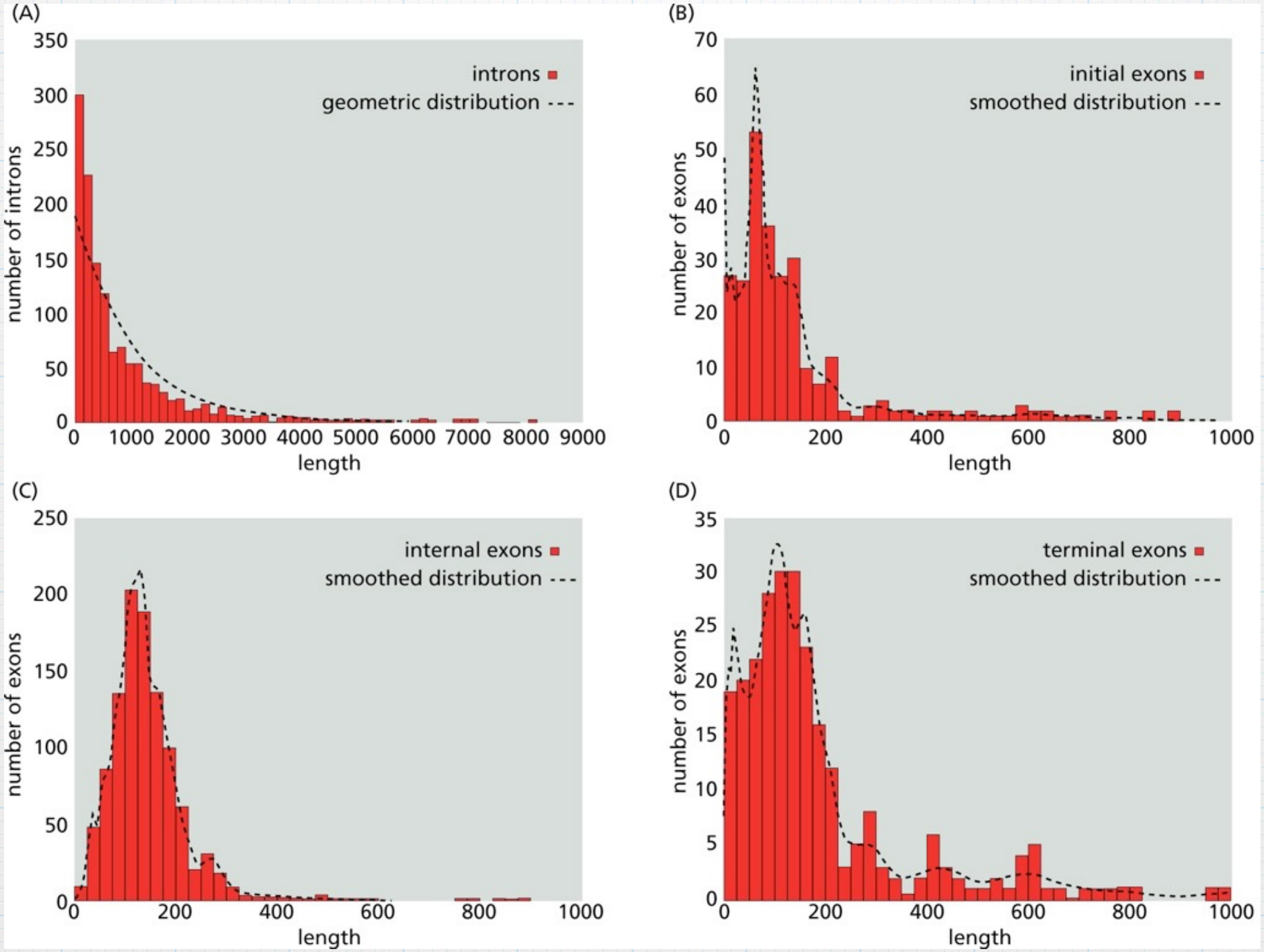
* In addition, although the signals differ, there are equivalent transcription and translation start and stop signals.

* A crucial difference in gene structure causes eukaryotic gene detection to be far harder: there are numerous introns present in many genes.

From Eukaryotic DNA to Protein



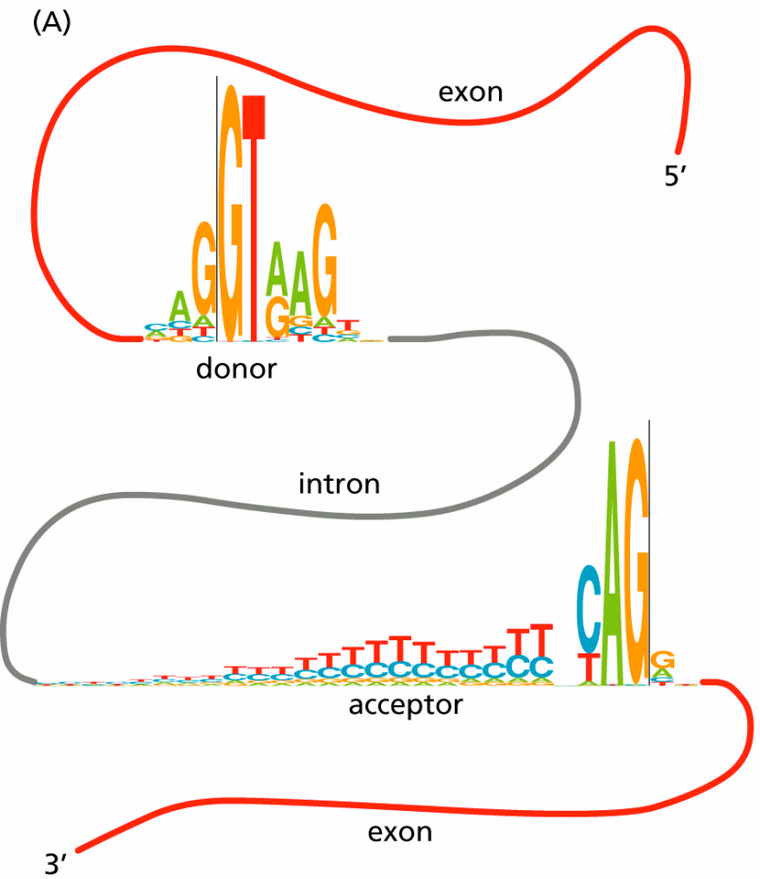
* The length of the protein-coding segments (exons) is on average smaller in eukaryotes than in prokaryotes, resulting in poorer base statistics, and making their detection more difficult.

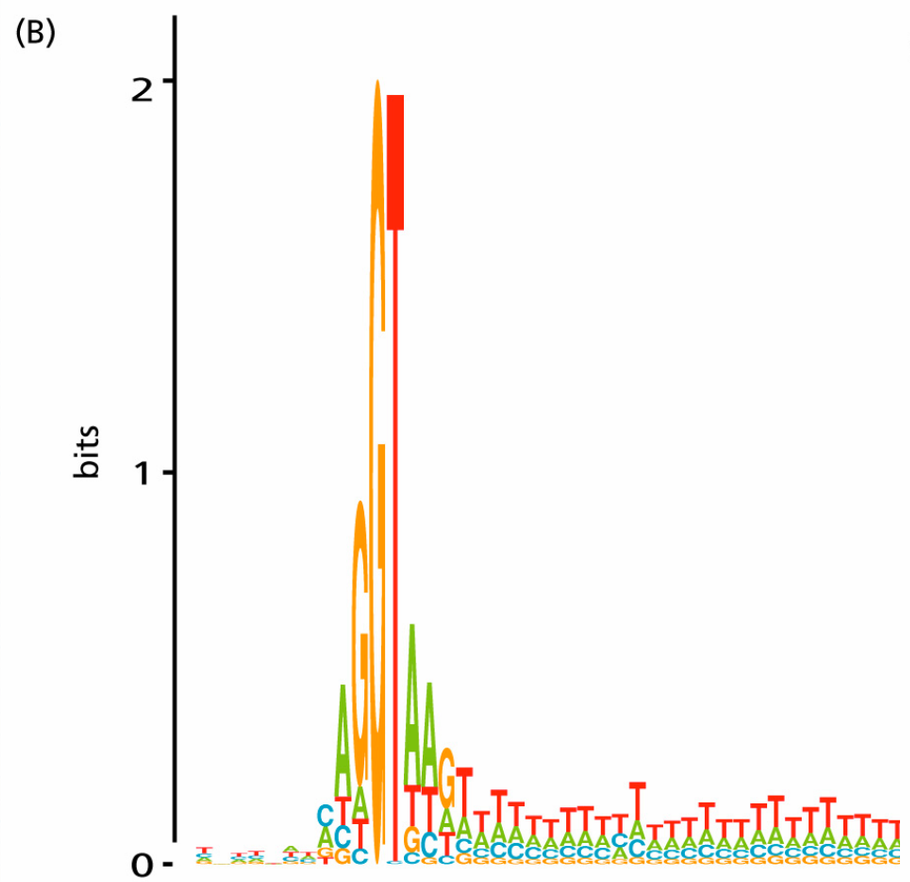


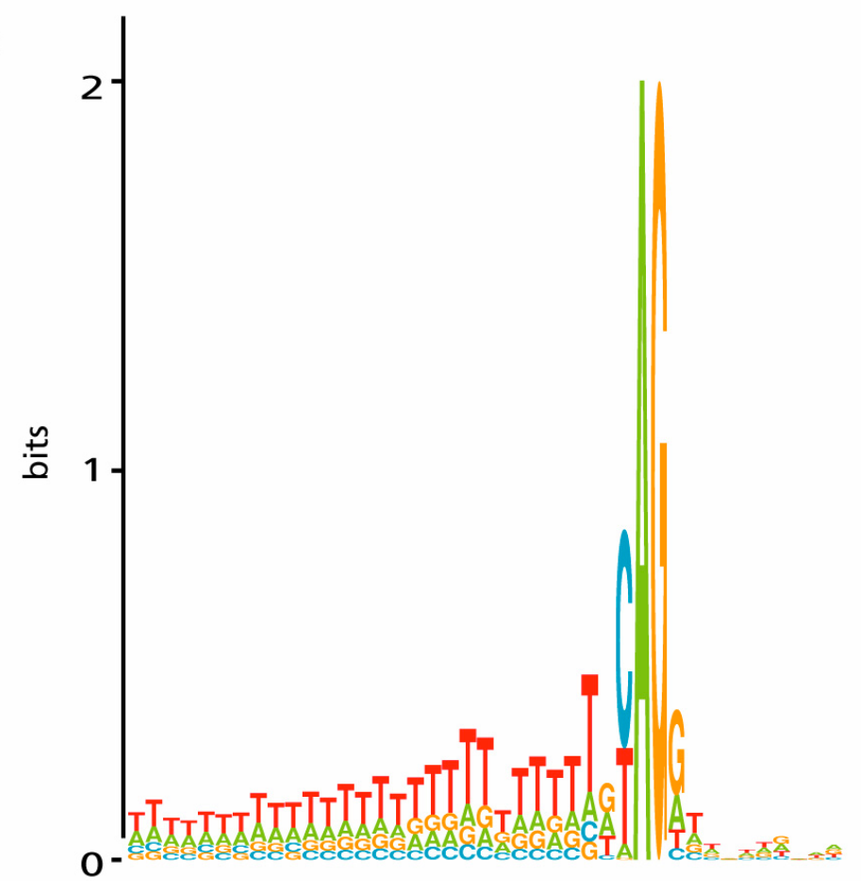
Distributions in the human genome

* An additional difference that can also cause difficulties is that the density of genes in most segments of eukaryotic genomes is significantly less than in prokaryotes.

* The splice signals at intron-exon boundaries are quite variable, making them hard to locate accurately.







human donor/acceptor sites

donor sites

acceptor sites

in Arabidopsis

Alternative Splicing

* A particularly difficult problem can arise in eukaryotic genomes when moving from gene detection to protein prediction, a trivial step in prokaryotes.
* The splicing of introns in the RNA is not always identical for a given gene (the phenomenon of alternative splicing).

Alternative Splicing

* Alternative splicing can give rise to the production of two or more different proteins from the same gene, and these are often known as splice variants.

Promoter Sequences and Binding Sites for Transcription Factors

* A further difference between prokaryotic and eukaryotic gene structures is that the sequence signals in the upstream regions are much more variable in eukaryotes, both in composition and position.

Promoter Sequences and Binding Sites for Transcription Factors

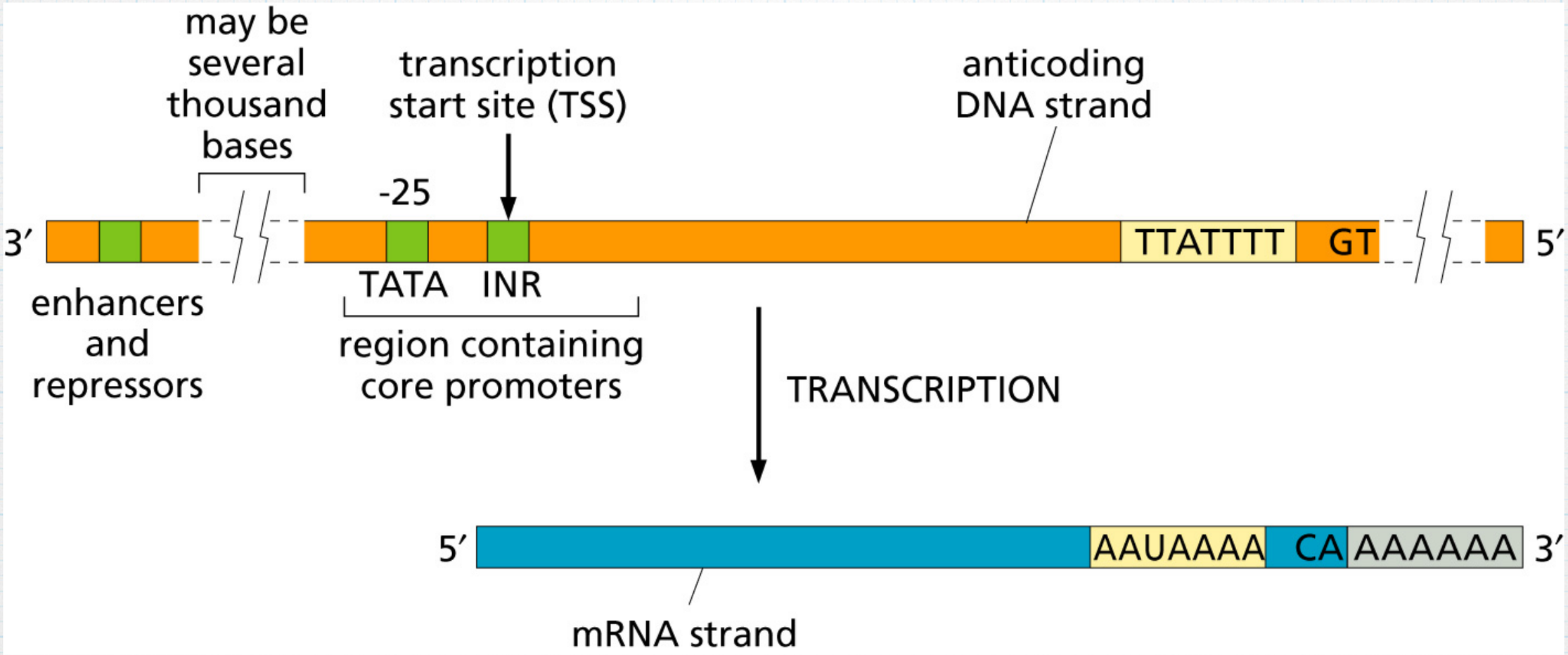
* The control of gene expression is more complex in eukaryotes than prokaryotes, and can be affected by many molecules binding the DNA in the region of the gene.

Promoter Sequences and Binding Sites for Transcription Factors

* This leads to many more potential promoter binding signals spread over a much larger region (possibly several thousand bases) in the vicinity of the transcription start site.

Predicting Eukaryotic Gene Signals

Gene Structure in Eukaryotes



* Regulatory elements in eukaryotes are more complex.

* Three types of RNA polymerase transcribe genes: RNA polymerase II transcribes all protein coding genes, whereas other RNA polymerase tpyes transcribe genes for tRNAs, rRNAs and other types of RNA

* In 1990 P. Bucher derived weight matrices to identify four separate RNA polymerase II promoter elements: the TATA box, the cap signal (INR), the CCATT box, and the GC box.

* Using more than 500 aligned eukaryotic sequences, the weights of different bases a at position u is a signal sequence were obtained from the general equation

$$w_u(a) = \ln\left(\frac{n_u(a)}{e_u(a)} + \frac{c}{100}\right) + c_u$$

number of occurrences of base a at position u

expected number of bases a at position u

a small number (often 2)

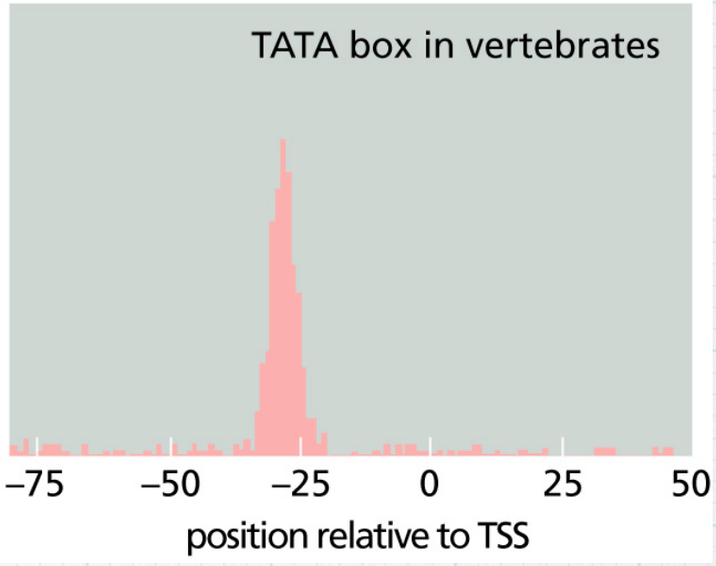
adjusted to make the greatest $w_u(a)$ zero

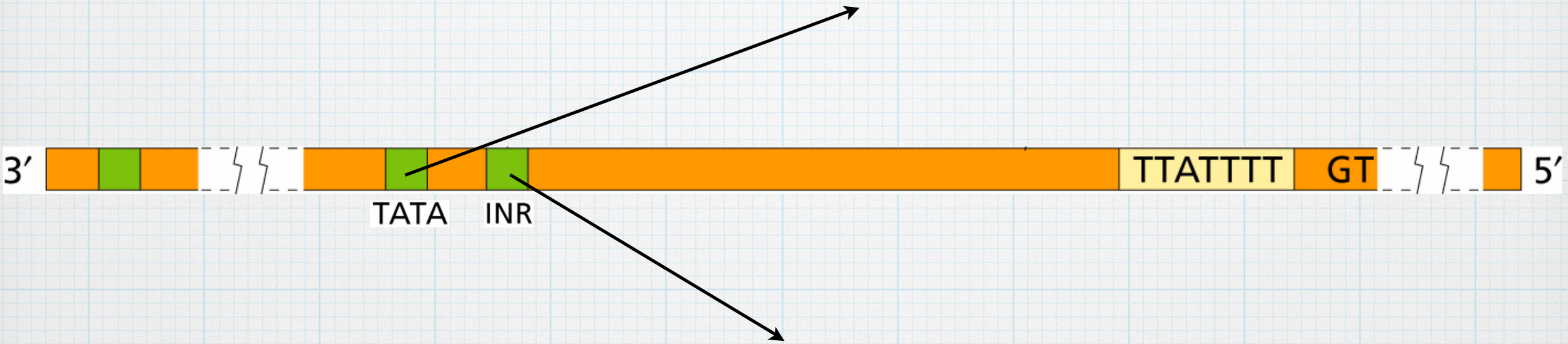
(A)

sequence position

		-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
observed bases	A	61	16	352	3	354	268	360	222	155	56	83	82	82	68	77
	C	145	46	0	10	0	0	3	2	44	135	147	127	118	107	101
	G	152	18	2	2	5	0	20	44	157	150	128	128	128	139	140
	T	31	309	35	374	30	121	6	121	33	48	31	52	61	75	71
weight matrix	A	-1.02	-3.05	0.00	-4.61	0.00	0.00	0.00	0.00	-0.01	-0.94	-0.54	-0.48	-0.48	-0.74	-0.62
	C	-0.28	-2.06	-5.22	-3.49	-5.17	-4.63	-4.12	-3.74	-1.13	-0.05	0.00	-0.05	-0.11	-0.28	-0.40
	G	0.00	-2.74	-4.38	-4.61	-3.77	-4.73	-2.65	-1.50	0.00	0.00	-0.09	0.00	0.00	0.00	0.00
	T	-1.68	0.00	-2.28	0.00	-2.34	-0.52	-3.65	-0.37	-1.40	-0.97	-1.40	-0.82	-0.66	-0.54	-0.61
consensus		G/C	T	A	T	A	A/T	A	A/T	G/A	G/C	C/G	G/C	G/C	G/C	G/C

(B)



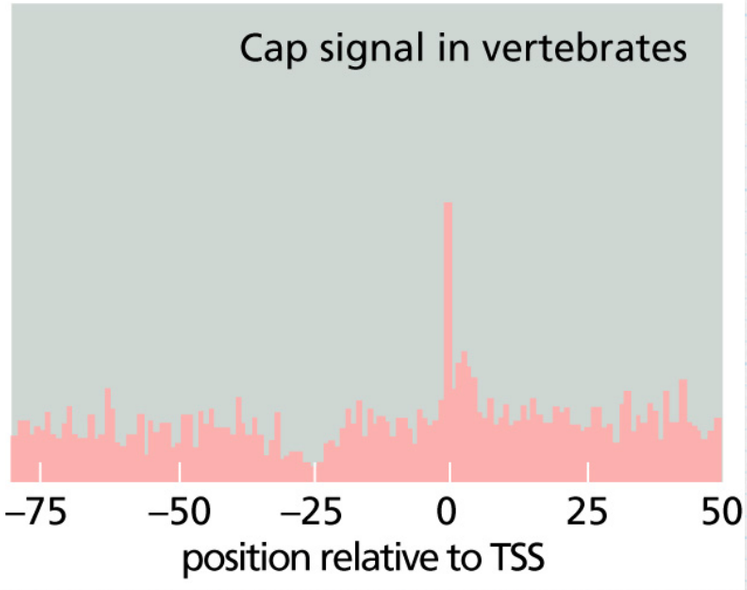


(C)

sequence position

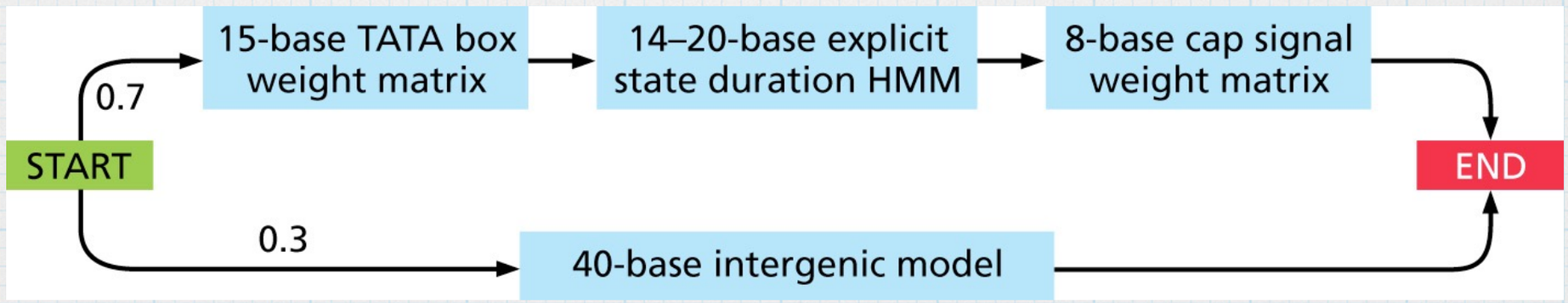
		-2	-1	0	1	2	3	4	5
observed bases	A	49	0	288	26	77	67	45	50
	C	48	303	0	81	95	118	85	96
	G	69	0	0	116	0	46	73	56
	T	137	0	15	80	131	72	100	101
weight matrix	A	-1.14	-5.26	0.00	-1.51	-0.65	-0.55	-0.91	-0.82
	C	-1.16	0.00	-5.21	-0.41	-0.45	0.00	-0.29	-0.18
	G	-0.75	-5.26	-5.21	0.00	-4.56	-0.86	-0.38	-0.65
	T	0.00	-5.26	-2.74	-0.29	0.00	-0.36	0.00	0.00
consensus		T	C	A					

(D)



* In GenScan (a popular gene detection method; more later), the promoter detection component uses Bucher's TATA-box and cap-signal models.

* To avoid missing genes that lack a TATA-box, the model allows for both possibilities.



Predicting Exons and Introns

* All internal introns and exons in a eukaryotic gene are delimited by the splice sites at which introns are cut out of the RNA transcript and the exon sequences joined together.

Predicting Exons and Introns

* The splice sites have distinct sequence signals.
* There are programs that predict introns and exons without reference to splice sites, and other programs that predict splice sites without information about introns and exons.

Predicting Exons and Introns

* For example, GenScan identifies eukaryotic coding regions by dicodon statistics, as in the prokaryotic example given earlier, but it uses an explicit state duration HMM based on the observed length distribution of real exons.
* The length of the potential exon is generated from this distribution, and its sequence generated with probabilities based on the dicodon statistics.

Predicting Exons and Introns

* Measures of gene prediction accuracy at the exon level make use of:
  * AE: the number of actual (real) exons in the data
  * PE: The number of predicted exons
  * CE: the number of exons predicted exactly
  * ME: the number of real exons that do not overlap any of the predicted ones
  * WE: the number of predicted exons that have no overlap with the actual ones

Predicting Exons and Introns

* We then define:
  * Sn_1=CE/AE
  * Sp_1=CE/PE
  * Sn_2=ME/AE
  * Sp_2=WE/PE

Predicting Exons and Introns

* Many of these ab initio prediction programs have been modified to take account of homology to experimental gene sequences.
* For example, GenScan allows BLAST hits to the genome sequence to be used as an extra guide.

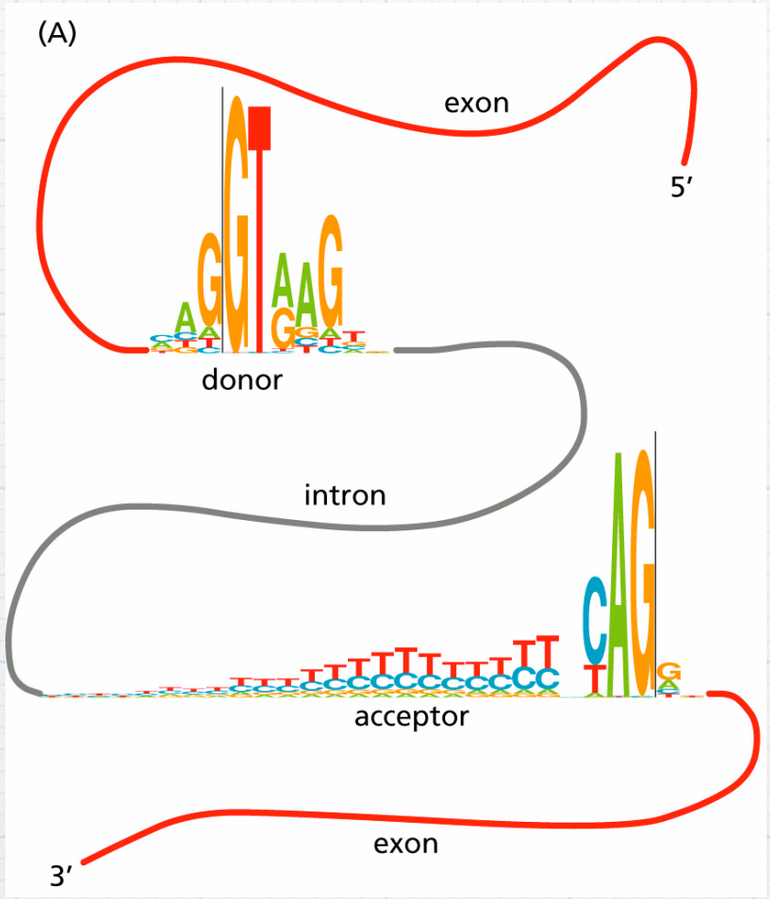
Splice Site Prediction

* Most introns start with (i.e., have a 5' end) a GU dinucleotide in the RNA (GT in the DNA) at what is referred to as the donor splice site.
* The 3' end of introns (acceptor splice site) is mostly AG dinucleotide.

Splice Site Prediction

* Thus, locating occurrences of AG and GT would identify all possible splice sites with these sequence properties, but in addition there would be about 30 to 100 false predicted sites for every true one.
* Properties of the surrounding sequence is used to reduce the false-negative prediction rate to a manageable level.

Splice Site Prediction



Splice Site Prediction

* GenScan uses a sophisticated variant of weighted matrices to predict donor splice sites.

Splice Site Prediction

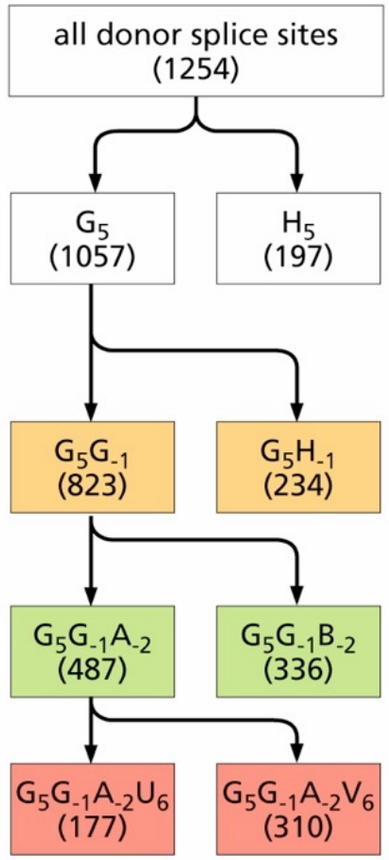
(A)

	sequence position								
	-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
consensus	A/C	A	G	G	U	A/G	A	G	U

observed bases

(B)

Position	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

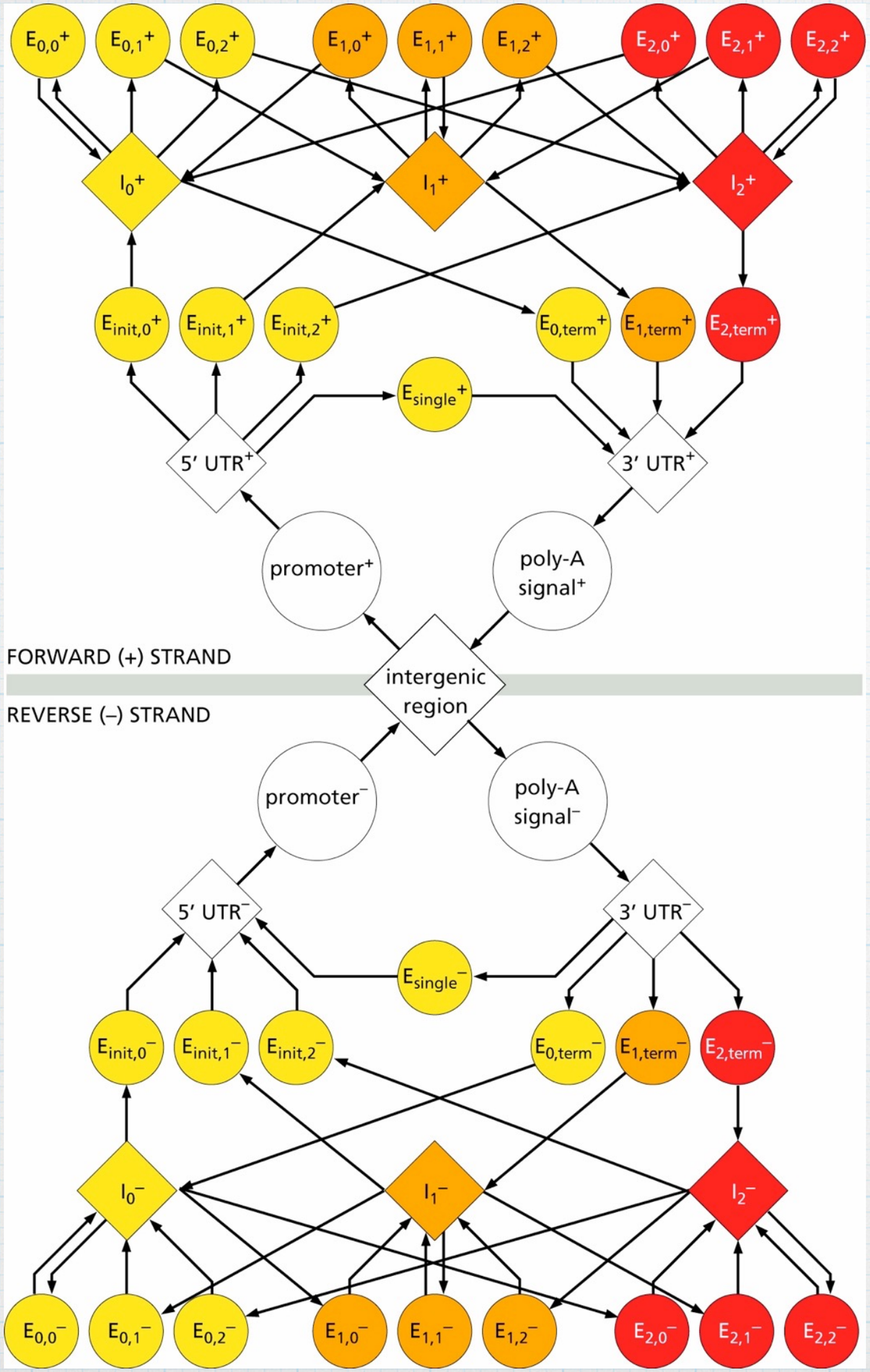


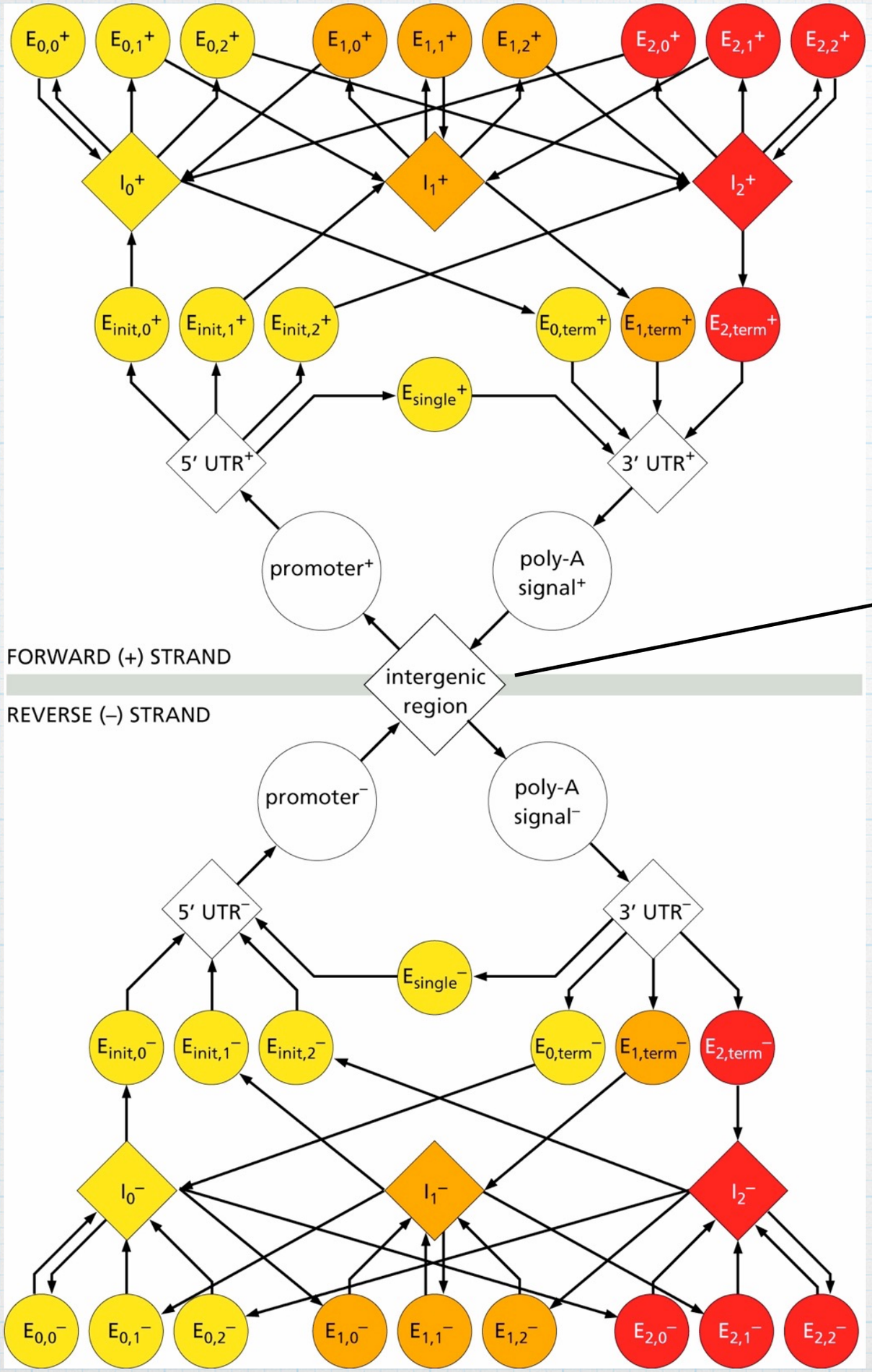
Position	A%	C%	G%	U%
-3	35	44	16	6
-2	85	4	7	5
-1	2	1	97	0
+3	81	3	15	2
+4	51	28	9	12
+6	22	20	30	28
-3	29	31	21	18
-2	43	30	17	11
+3	56	0	43	0
+4	93	2	3	3
+6	5	10	10	76
-3	29	30	18	23
+3	42	1	56	1
+4	80	4	8	8
+6	14	21	16	49
-3	39	43	15	2
+3	46	6	46	3
+4	69	5	20	7

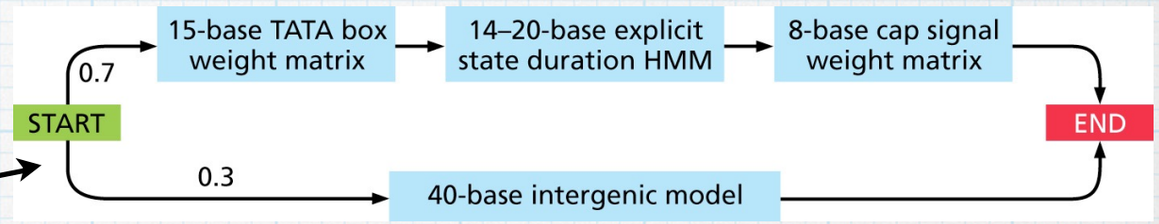
Complete Eukaryotic Gene Models

* Once all the separate components of a gene have been predicted, it is possible to put them all together to predict a complete gene structure.

* GenScan uses an HMM that considers both the forward and backward strands of DNA simultaneously.


$E_{0,0}^+$
$E_{0,1}^+$
$E_{0,2}^+$
$E_{1,0}^+$
$E_{1,1}^+$
$E_{1,2}^+$
$E_{2,0}^+$
$E_{2,1}^+$
$E_{2,2}^+$
I_0^+
I_1^+
I_2^+
$E_{init,0}^+$
$E_{init,1}^+$
$E_{init,2}^+$
$E_{0,term}^+$
$E_{1,term}^+$
$E_{2,term}^+$
E_{single}^+
5′ UTR$^+$
3′ UTR$^+$
promoter$^+$
poly-A signal$^+$
intergenic region
FORWARD (+) STRAND
REVERSE (–) STRAND
promoter$^-$
poly-A signal$^-$
5′ UTR$^-$
3′ UTR$^-$
E_{single}^-
$E_{init,0}^-$
$E_{init,1}^-$
$E_{init,2}^-$
$E_{0,term}^-$
$E_{1,term}^-$
$E_{2,term}^-$
I_0^-
I_1^-
I_2^-
$E_{0,0}^-$
$E_{0,1}^-$
$E_{0,2}^-$
$E_{1,0}^-$
$E_{1,1}^-$
$E_{1,2}^-$
$E_{2,0}^-$
$E_{2,1}^-$
$E_{2,2}^-$


$E_{0,0}^+$
$E_{0,1}^+$
$E_{0,2}^+$
$E_{1,0}^+$
$E_{1,1}^+$
$E_{1,2}^+$
$E_{2,0}^+$
$E_{2,1}^+$
$E_{2,2}^+$
I_0^+
I_1^+
I_2^+
$E_{init,0}^+$
$E_{init,1}^+$
$E_{init,2}^+$
$E_{0,term}^+$
$E_{1,term}^+$
$E_{2,term}^+$
E_{single}^+
5′ UTR$^+$
3′ UTR$^+$
promoter$^+$
poly-A signal$^+$
intergenic region
FORWARD (+) STRAND
REVERSE (–) STRAND
promoter$^-$
poly-A signal$^-$
5′ UTR$^-$
3′ UTR$^-$
E_{single}^-
$E_{init,0}^-$
$E_{init,1}^-$
$E_{init,2}^-$
$E_{0,term}^-$
$E_{1,term}^-$
$E_{2,term}^-$
I_0^-
I_1^-
I_2^-
$E_{0,0}^-$
$E_{0,1}^-$
$E_{0,2}^-$
$E_{1,0}^-$
$E_{1,1}^-$
$E_{1,2}^-$
$E_{2,0}^-$
$E_{2,1}^-$
$E_{2,2}^-$


15-base TATA box weight matrix
14–20-base explicit state duration HMM
8-base cap signal weight matrix
0.7
START
END
0.3
40-base intergenic model

Genome Annotation

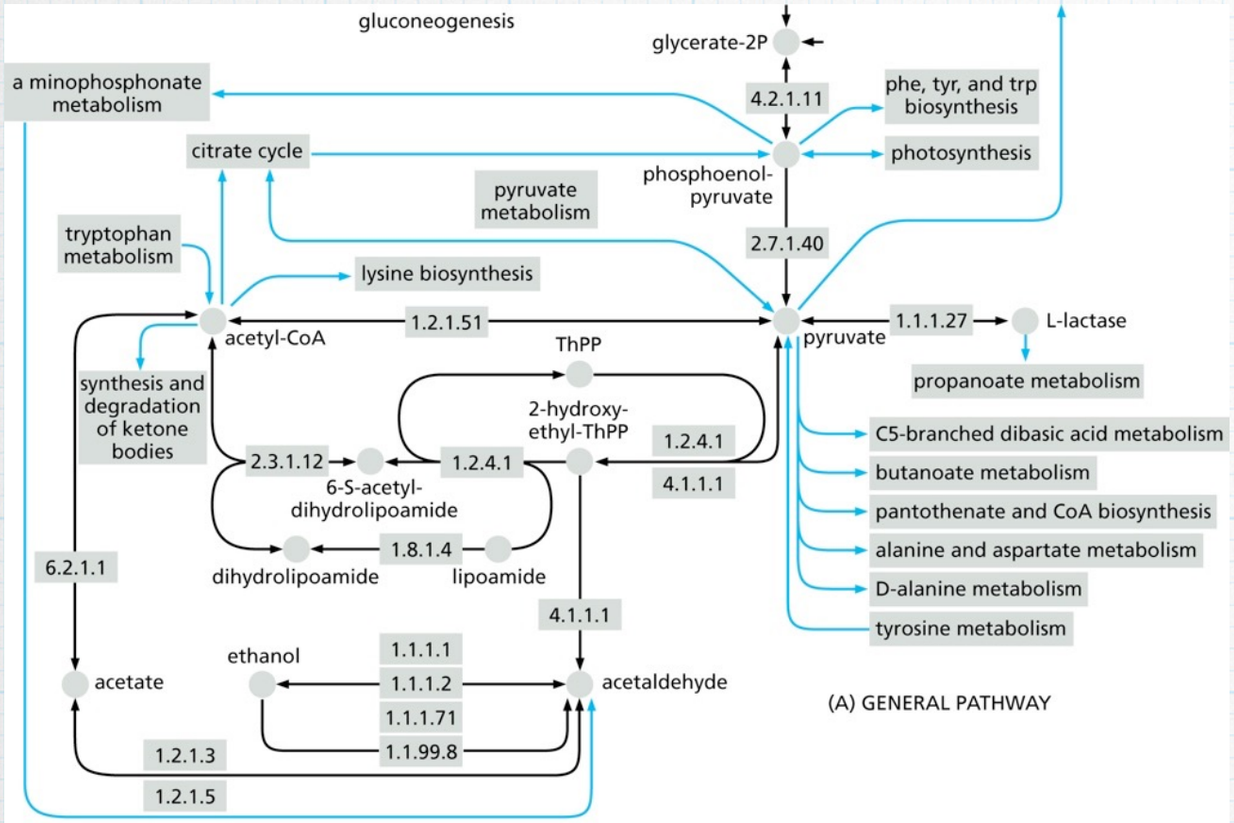
* Once all the genes have been predicted, it remains to be determined what function the encoded proteins might play.

* The obvious way to start to determine gene function is by sequence analysis.

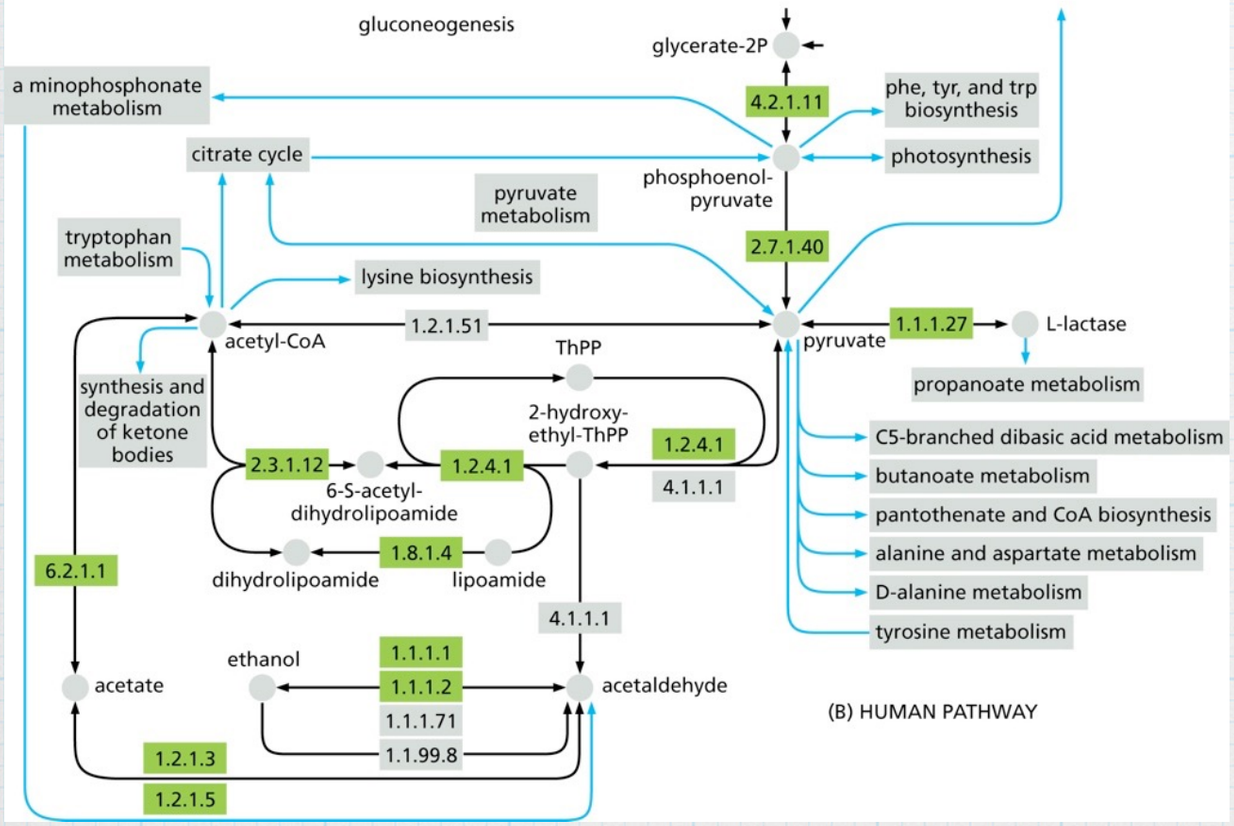
* If the encoded protein has one or more significant matches against sequence and pattern databases, the function and other properties can be predicted with considerable confidence to be similar to those of the matches.

Pathway Information

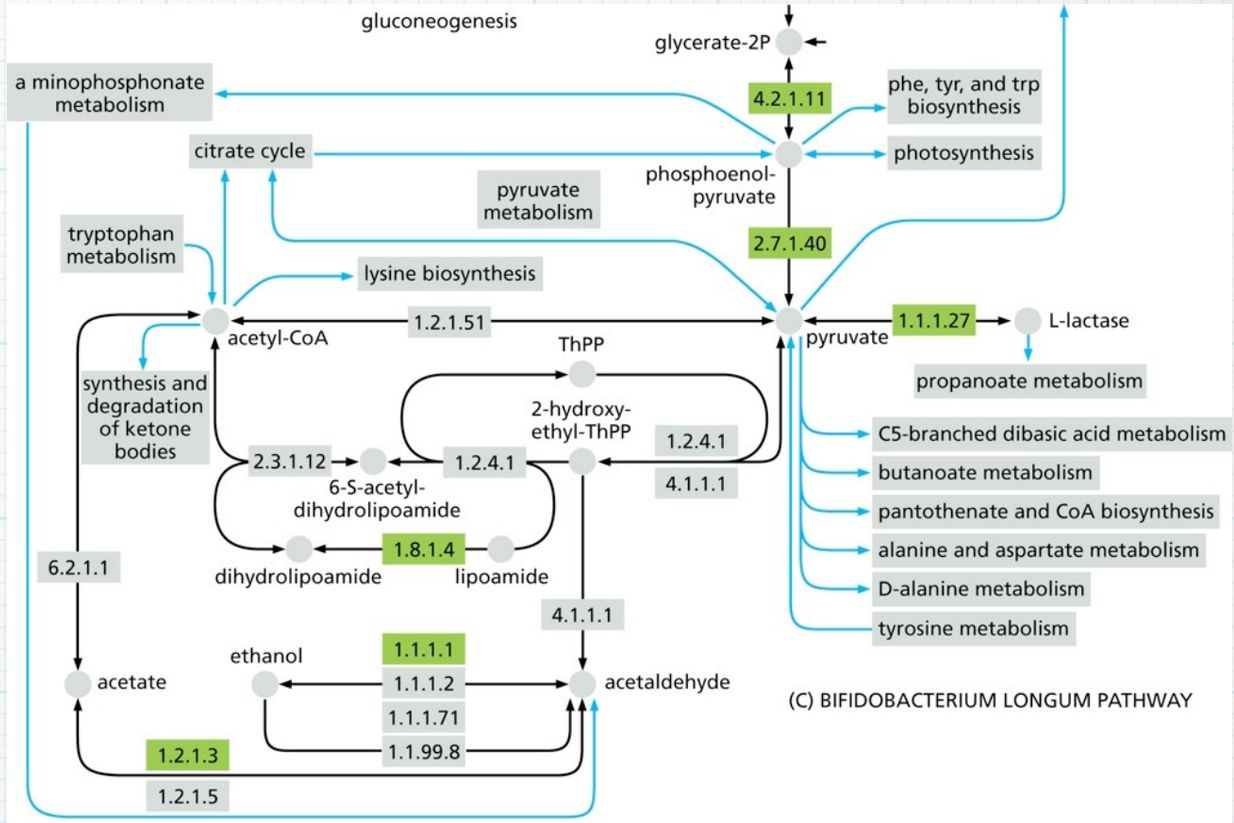
* The use of pathway information can aid gene and genome annotation.
* Comparing a new genome to a well-annotated and functionally defined one can aid the analysis of specific pathways and may identify missing components or blanks in the pathway.



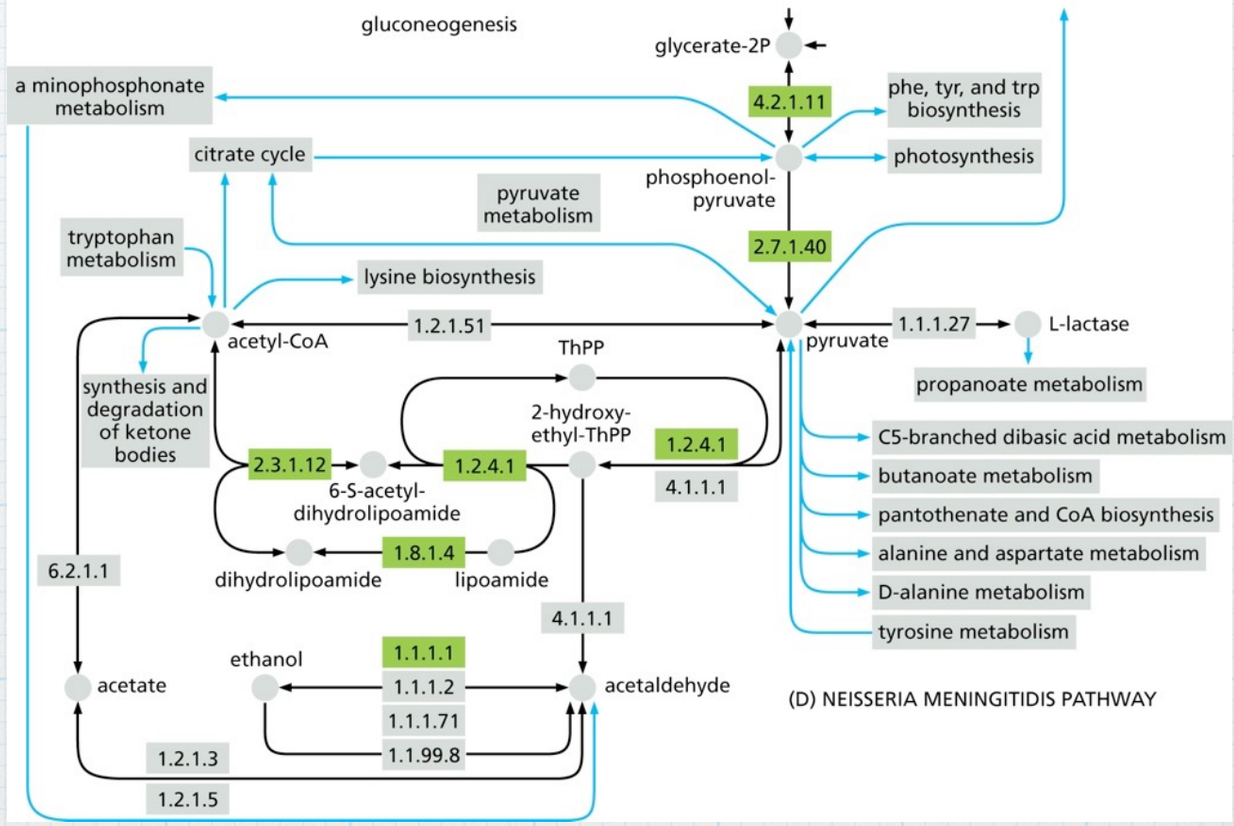
(A) GENERAL PATHWAY



(B) HUMAN PATHWAY



(C) BIFIDOBACTERIUM LONGUM PATHWAY



(D) NEISSERIA MENINGITIDIS PATHWAY

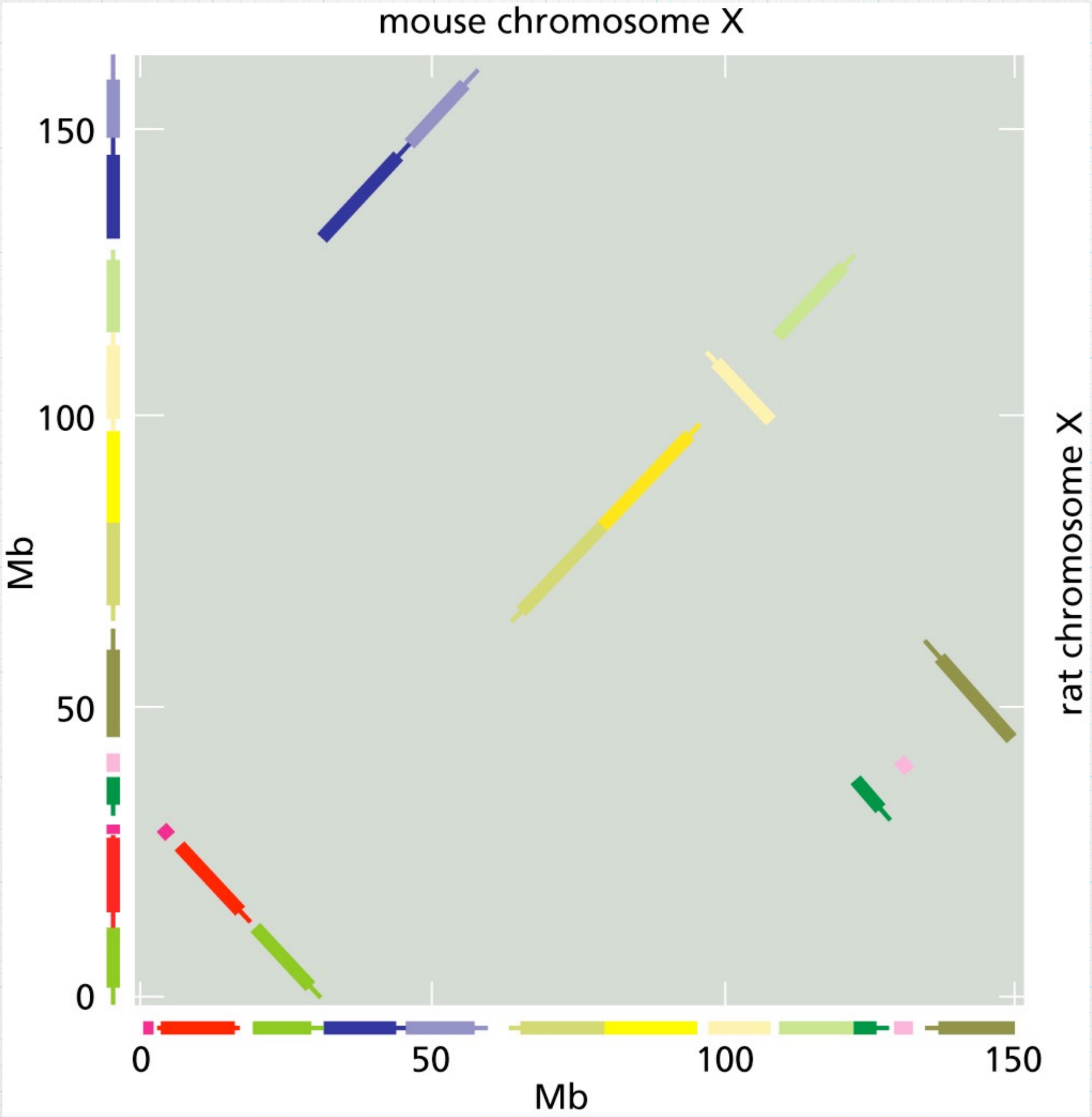
Gene Ontology

* One of the important aspects of genome annotation has been the recognition of the importance of gene ontology.

* Gene ontology is a set of standardized and accepted terms that encompass the range of possible functions and can be found on the Gene Ontology Consortium's website (geneontology.org)

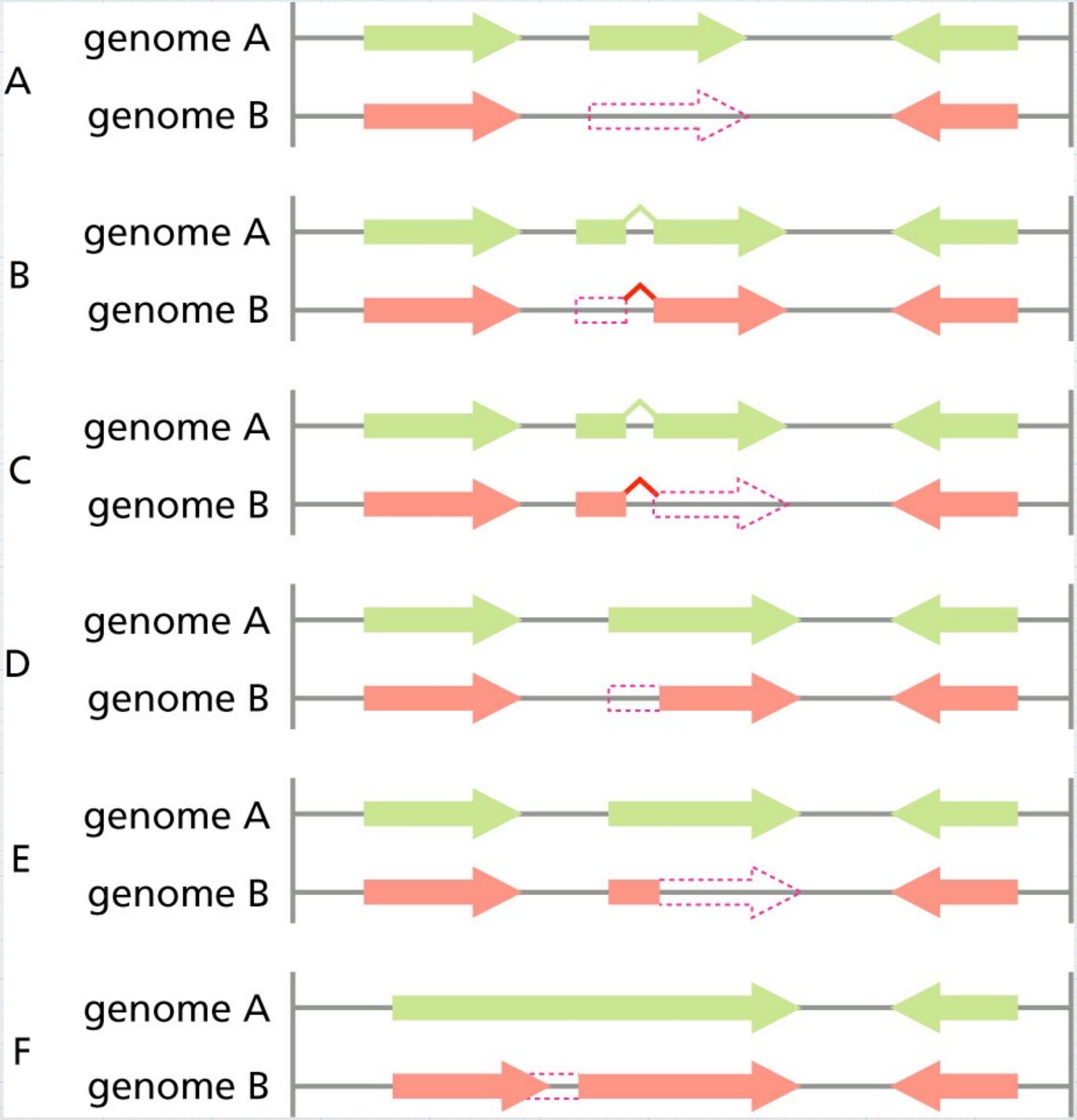
Genome Comparison

* Comparison of two genomes can be a very powerful tool in determining the status of uncertain gene predictions.
* Aligning two genomes is not easy, as large-scale rearrangements are common, but it should be possible to find regions of synteny where the gene structure is sufficiently similar as to make their common evolutionary ancestry apparent.


mouse chromosome X
150
100
50
0
0
50
100
150
Mb
rat chromosome X

Genome Comparison

* Comparing syntenic regions can reveal errors in one or other genome annotation.


A
genome A
genome B
B
genome A
genome B
C
genome A
genome B
D
genome A
genome B
E
genome A
genome B
F
genome A
genome B

Acknowledgments

* "Understanding Bioinformatics" by Zvelebil and Baum.