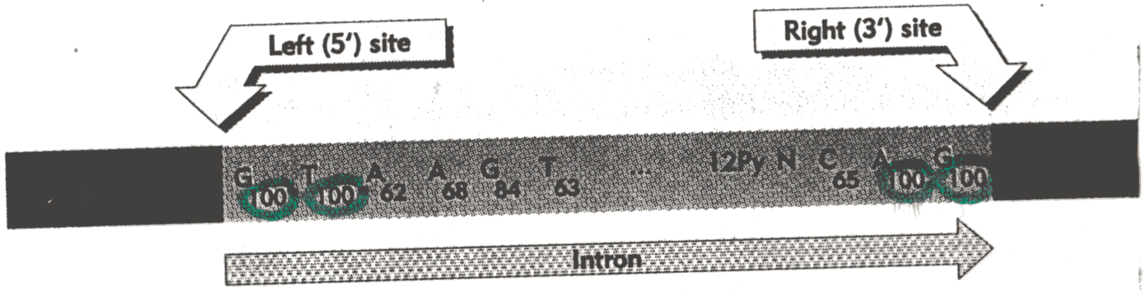
```
AGCGGATCCCCCGGTGGCCTCATGTCGCGCAGTGGAACCGGATCCTCAGCAACGCCCAGGCAGGCGTCAGAGG
000000000000000000000CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

CGGACGCCGGCAGCAGCAACCTTCCGGGCAAACGGTAACTGCACCGCGGCAGGGACTCGCTGGGGCGCGGA
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

GCCGAGCCCTCCCCTTCCTTAGGAAGCTTTCGTCCCCTCCGAAGGTTGGAACGCTCATCCCGAGCCAGA
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

CCGACAAGGCGTACAGTCTGCAGGCCTCTACGAGCAGCAGGCCAATTGGCGCTGGGAAAGTCCAATCCTG
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

GGCCTCTAGCTCCTGAGCGGGACAGGGCCGAGAGGGCGCTCCCGAGCTTGGGCCTGCTGGTGGGTGAGAC
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

CCAGGAGAGAGGGAGCTAGAGGGGGAGCTCTGAGGACTGATCTTGACTGTCTGCCCCCAGACCATCAGC
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIICCCCCCCCC

ATATCCGCTACAACCCGCTGCAGGATGAGTGGGTGCTGGTGTCAGCTCACCGCATGAAGCGGCCCTGGTA
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

GGGTCAAGTGGAGCCCCAGCTTCTGAAGACAGTGCCCCGCCATGACCCTCTCAACCCTCTGTGTCCTGGG
C000000000000000000000000000000000000000000000000000000000000000000000

GCCATCCGAGCCAACGGAGAGGTAAGCCTGTAGAGCCCTGCATCTGCAGGCTGGGCCACGG
0000000000000000000000000000000000000000000000000000000000000
```

Figure 1: A DNA sequence containing one gene. For each nucleotide its label is written below. The coding regions are labeled ‘C’, the introns ‘I’, and the intergenic regions ‘0’. The shaded areas are the coding regions. (Genbank sequence HUMGALT54X, Homo sapiens galactose-1-phosphate uridyl transferase (GALT) mutant).

Figure 31.2

The ends of the intron are defined by the GT-AG rule.



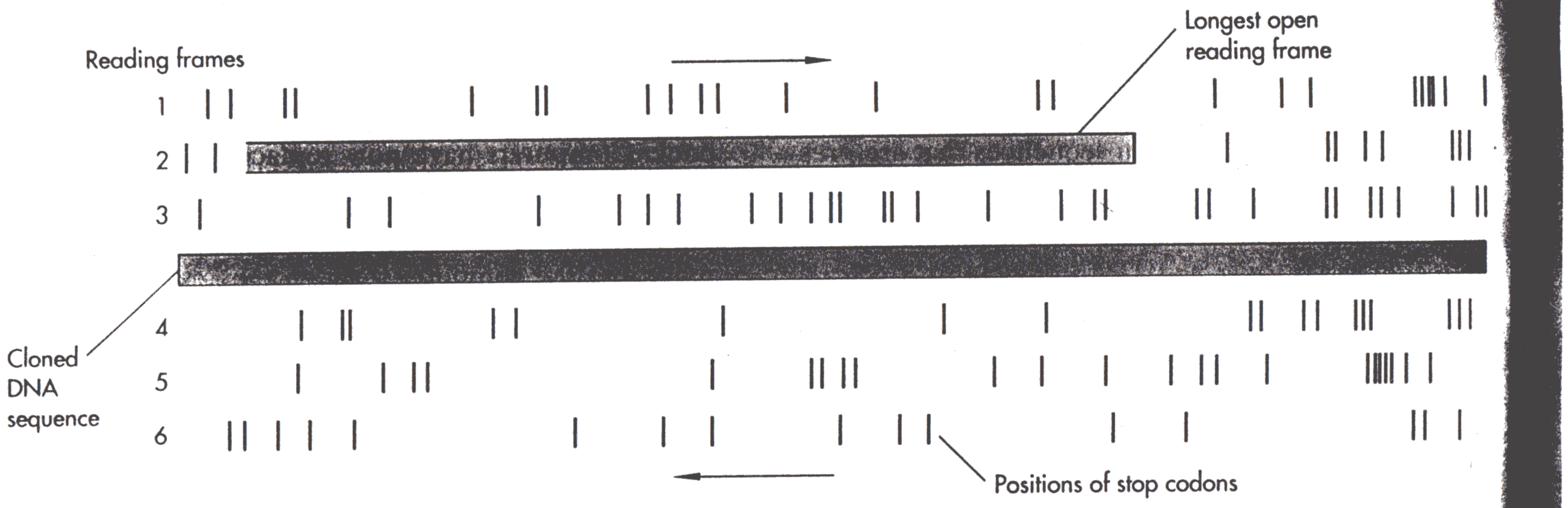


FIGURE 7-15
Finding an open reading frame. A simple method to spot an open reading frame with the potential to encode a protein is to ask a computer to locate the translation termination codons in all reading frames in both orientations (six frames in all). For example, above and below the bar representing the cDNA sequence of the human c-*FOS* protooncogene (see Chapter 18) are lines that represent the termination codons in each reading frame in the leftward and rightward orientation, respectively. In five of the frames, termination codons are scattered throughout the sequence, so that it is clear these frames cannot be translated into a sizeable protein. In the second frame above the cDNA, however, there is a long gap in which there is no termination codon. It would be expected that the contiguous open reading frame encodes the protein, and in fact it does encode the c-*FOS* protein.

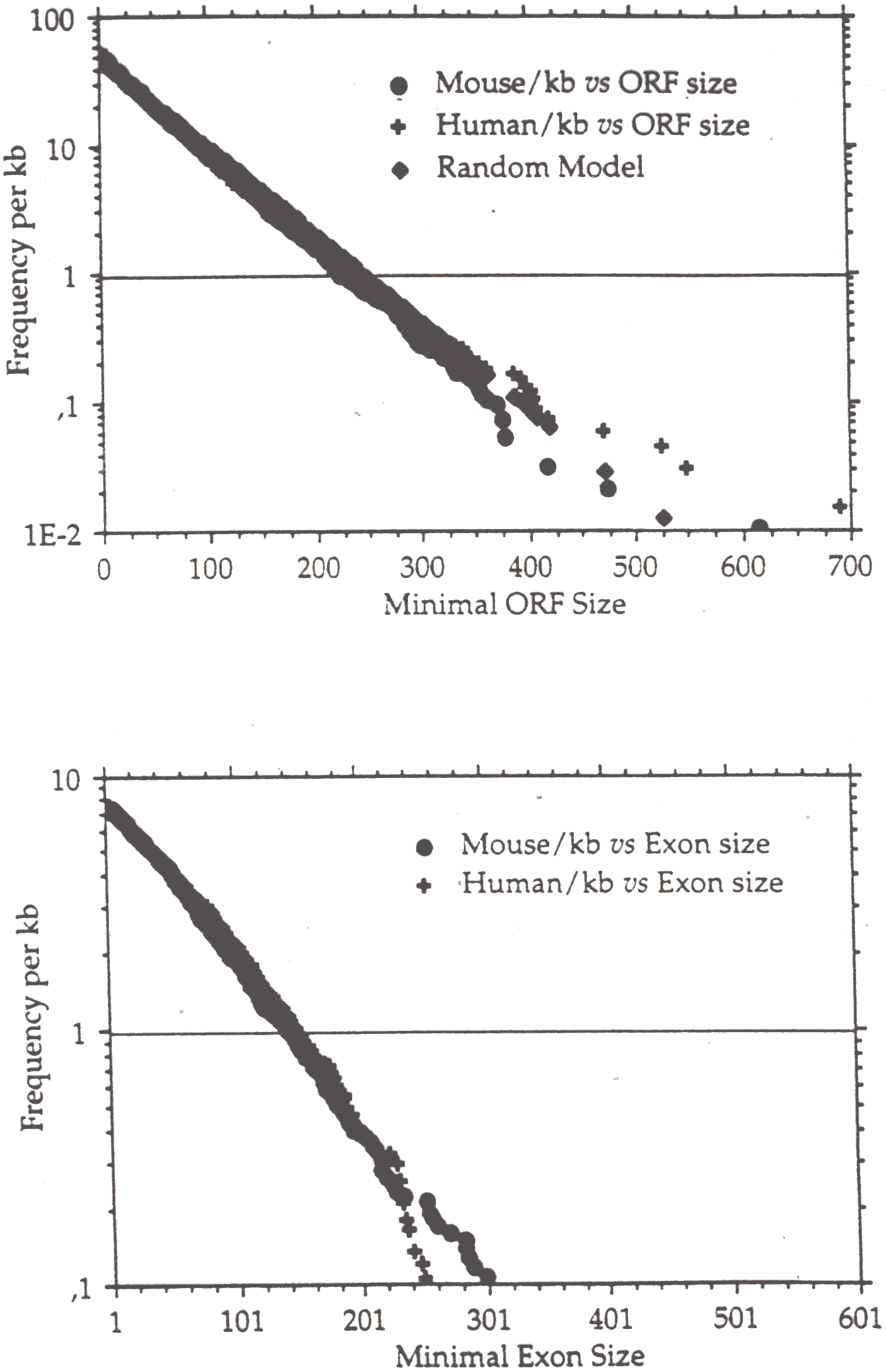


Fig. 1. Open reading frames (ORF) and candidate exons in actual sequences. Top: open reading frame (ORF) size distribution per kilobase (kb) in two unrelated long vertebrate genomic sequences. The mouse sequence is from a 94 kb contig from the X chromosome. The human sequence is from a 67 kb contig from the Xp22.3 region. Nucleotide compositions are {A: 30%, T: 27.5%, C: 21%, G:21.5%} for the murine sequence and {A: 30%, T: 31%, C: 19.5%, G:19.5%} for the human sequence. The two distributions are very close, and fit a simple theoretical model (see text) for length up to 350 nt. The random occurrence of at least one ORF of size ≥ 250 nt is expected for each kb. Actual protein coding ORFs are thus indistinguishable from random noise. Bottom: candidate exon size distribution per kilobase in the same sequences. Reasonable splice acceptor and donor consensus (Fig. 2) are now required to flank the 5′ and 3′ extremities of the ORF. The murine and human sequences exhibit very similar length distributions. An average of one candidate exon of size ≥ 150 nt is found for each kilobase.

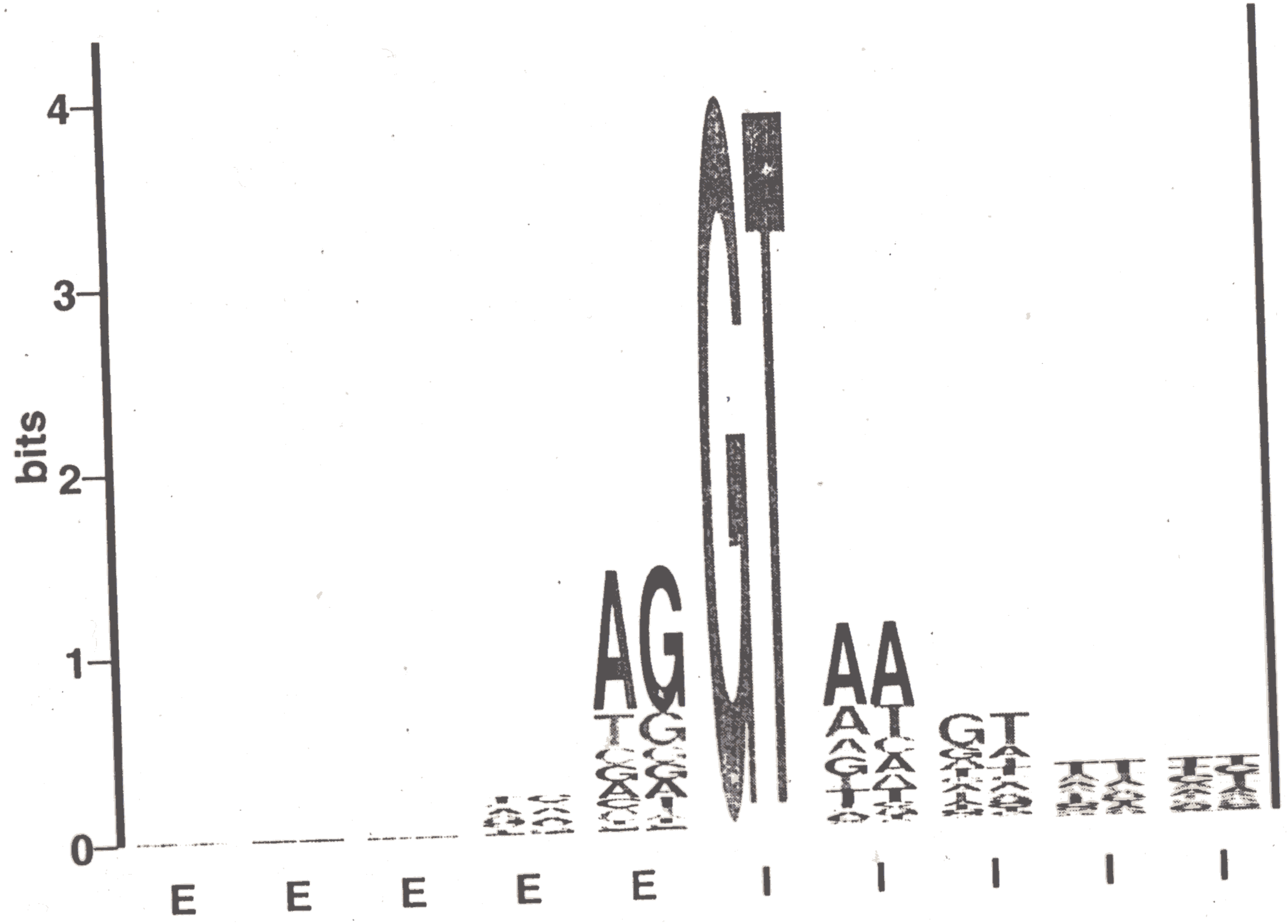


Figure 1.8: A Logo of Donor Splice Sites from the Dicot Plant *A. thaliana* (cress). The logo is based on frequencies of nonoverlapping dinucleotides in the exon/intron transition region, using the standard Shannon information measure entering equation (1.8) with the alphabet size $|A| = 16$. The logo was prepared on a nonredundant data set of sequences extracted from GenBank [246]. See Plate IV for illustration in color.

Splice Signal Position Weight Matrices

Splice acceptor signal

Position

	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0	0	0	0	0	0	0	10	0
C	1	1	1	1	1	1	1	1	0	2	0	0
G	0	0	0	0	0	0	0	0	0	0	0	10
T	1	1	1	1	1	1	1	1	0	1	0	0

Minimal score: 26

Splice donnor signal

Position

	1	2	3	4	5	6
A	0	0	0	1	1	0
C	0	0	0	0	0	0
G	1	10	0	1	1	1
T	0	0	10	0	0	0