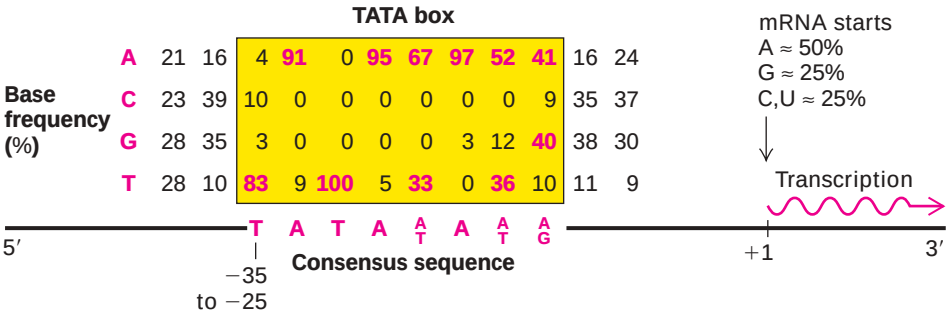

▲ FIGURE 11-9 **Determination of consensus TATA box sequence.** The nucleotide sequences upstream of the start site in 900 different eukaryotic protein-coding genes were aligned to maximize homology in the region from −35 to −26. The tabulated numbers are the percentage frequency of each base at each position. Maximum homology occurs over an eight-base region, referred to as the *TATA box*, whose consensus sequence is shown at the bottom. The initial base in mRNAs encoded by genes containing a TATA box most frequently is an A. [See P. Bucher, 1990, *J. Mol. Biol.* **212**:563, and http://www.epd.isb-sib.ck/promoter_elements.]

where A^{+1} is the base at which transcription starts, Y is a pyrimidine (C or T), N is any of the four bases, and T/A is T or A at position +3.

Transcription of genes with promoters containing a TATA box or initiator element begins at a well-defined initiation site. However, transcription of many protein-coding genes has been shown to begin at any one of multiple possible sites over an extended region, often 20–200 base pairs in length. As a result, such genes give rise to mRNAs with multiple alternative 5′ ends. These genes, which generally are transcribed at low rates (e.g., genes encoding the enzymes of intermediary metabolism, often called "housekeeping genes"), do not contain a TATA box or an initiator. Most genes of this type contain a CG-rich stretch of 20–50 nucleotides within ≈100 base pairs upstream of the start-site region. The dinucleotide CG is statistically underrepresented in vertebrate DNAs, and the presence of a CG-rich region, or *CpG island*, just upstream from a start site is a distinctly nonrandom distribution. For this reason, the presence of a CpG island in genomic DNA suggests that it may contain a transcription-initiation region.

Promoter-Proximal Elements Help Regulate Eukaryotic Genes

Recombinant DNA techniques have been used to systematically mutate the nucleotide sequences upstream of the start sites of various eukaryotic genes in order to identify transcription-control regions. By now, hundreds of eukaryotic genes have been analyzed, and scores of transcription-control regions have been identified. These control elements, together with the TATA-box or initiator, often are referred to as the *promoter* of the gene they regulate. However, we prefer to reserve the term *promoter* for the TATA-box or initiator sequences that determine the initiation site in the template. We use the term **promoter-proximal elements** for control regions lying within 100–200 base pairs upstream of the start site. In some cases, promoter-proximal elements are cell-type-specific; that is, they function only in specific differentiated cell types.

One approach frequently taken to determine the upstream border of a transcription-control region for a mammalian gene involves constructing a set of 5′ deletions as discussed earlier (see Figure 11-3). Once the 5′ border of a transcription-control region is determined, analysis of *linker scanning mutations* can pinpoint the sequences with regulatory functions that lie between the border and the transcription start site. In this approach, a set of constructs with contiguous overlapping mutations are assayed for their effect on expression of a reporter gene or production of a specific mRNA (Figure 11-10a). One of the first uses of this type of analysis identified promoter-proximal elements of the thymidine kinase (*tk*) gene from herpes simplex virus (HSV). The results demonstrated that the DNA region upstream of the HSV *tk* gene contains three separate transcription-control sequences: a TATA box in the interval from −32 to −16, and two other control elements farther upstream (Figure 11-10b).

To test the spacing constraints on control elements in the HSV *tk* promoter region identified by analysis of linker scanning mutations, researchers prepared and assayed constructs containing small deletions and insertions between the elements. Changes in spacing between the promoter and promoter-proximal control elements of 20 nucleotides or fewer had little effect. However, insertions of 30 to 50 base pairs between a promoter-proximal element and the TATA box was equivalent to deleting the element. Similar analyses of other eukaryotic promoters have also indicated that considerable flexibility in the spacing between promoter-proximal elements is generally tolerated, but separations of several tens of base pairs may decrease transcription.