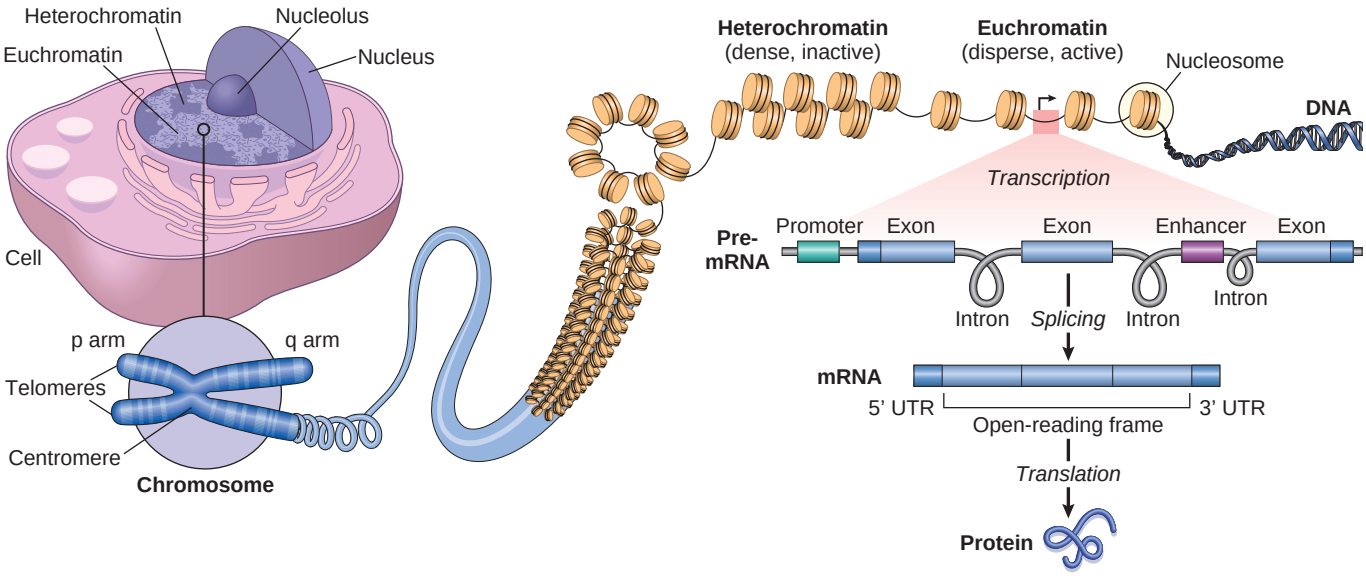

Figure 1-1 The organization of nuclear DNA. At the light microscopic level, the nuclear genetic material is organized into dispersed, transcriptionally active *euchromatin* or densely packed, transcriptionally inactive *heterochromatin*; chromatin can also be mechanically connected with the nuclear membrane, and nuclear membrane perturbation can thus influence transcription. Chromosomes (as shown) can only be visualized by light microscopy during cell division. During mitosis, they are organized into paired chromatids connected at *centromeres*; the centromeres act as the locus for the formation of a *kinetochore* protein complex that regulates chromosome segregation at metaphase. The *telomeres* are repetitive nucleotide sequences that cap the termini of chromatids and permit repeated chromosomal replication without loss of DNA at the chromosome ends. The chromatids are organized into short "P" ("*petite*") and long "Q" ("*next letter in the alphabet*") arms. The characteristic banding pattern of chromatids has been attributed to relative GC content (less GC content in bands relative to interbands), with genes tending to localize to interband regions. Individual chromatin fibers are comprised of a string of nucleosomes—DNA wound around octameric histone cores—with the nucleosomes connected via DNA linkers. Promoters are noncoding regions of DNA that initiate gene transcription; they are on the same strand and upstream of their associated gene. Enhancers are regulatory elements that can modulate gene expression over distances of 100 kB or more by looping back onto promoters and recruiting additional factors that are needed to drive the expression of pre-mRNA species. The intronic sequences are subsequently spliced out of the pre-mRNA to produce the definitive message that is translated into protein—without the 3′- and 5′-untranslated regions (UTR). In addition to the enhancer, promoter, and UTR sequences, noncoding elements are found throughout the genome; these include short repeats, regulatory factor binding regions, noncoding regulatory RNAs, and transposons.

lies in the 98.5% of the human genome that does not encode proteins. It has been known for some time that protein-coding genes in higher organisms are separated by long stretches of DNA whose function has been obscure for many years—sometime denoted as "dark matter" of the genome. That viewpoint has subsequently been modified, driven by the multinational ENCODE (Encyclopedia of DNA Elements) project that set out in 2007 to identify all regions of the genome that could be ascribed some function. The striking conclusion is that as much as **80% of the human genome either binds proteins, implying it is involved in regulating gene expression, or can be assigned some functional activity, mostly related to the regulation of gene expression, often in a cell-type specific fashion.** It follows that while proteins provide the building blocks and machinery required for assembling cells, tissues and organisms, it is the noncoding regions of the genome that provide the critical "architectural planning." Practically stated, the difference between worms and humans apparently lies more in the genomic "blueprints" than in the construction materials.

The major classes of functional non–protein-coding sequences found in the human genome are the following (Fig. 1-1):

- *Promoter* and *enhancer* regions that provide binding sites for transcription factors
- Binding sites for factors that organize and maintain higher order *chromatin structures*
- *Noncoding regulatory RNAs.* More than 60% of the genome is transcribed into RNAs that are never translated into protein, but which nevertheless can regulate gene expression through a variety of mechanisms. The two best-studied varieties—micro-RNAs and long noncoding RNAs—are described later.
- *Mobile genetic elements* (e.g., *transposons*). Remarkably, more than one third of the human genome is composed of these elements, popularly denoted as "jumping genes." These segments can move around the genome, exhibiting wide variation in number and positioning even amongst closely related species (i.e., humans and other primates). They are implicated in gene regulation and chromatin organization, but their function is still not well established.
- Special structural regions of DNA, in particular *telomeres* (chromosome ends) and *centromeres* (chromosome "tethers")

One of the reasons these findings have generated so much interest is that many, and perhaps most, of the genetic variations (*polymorphisms*) associated with diseases are located in non–protein-coding regions of the genome. Thus, variation in gene regulation may prove to be more important in disease causation than structural changes in specific proteins.