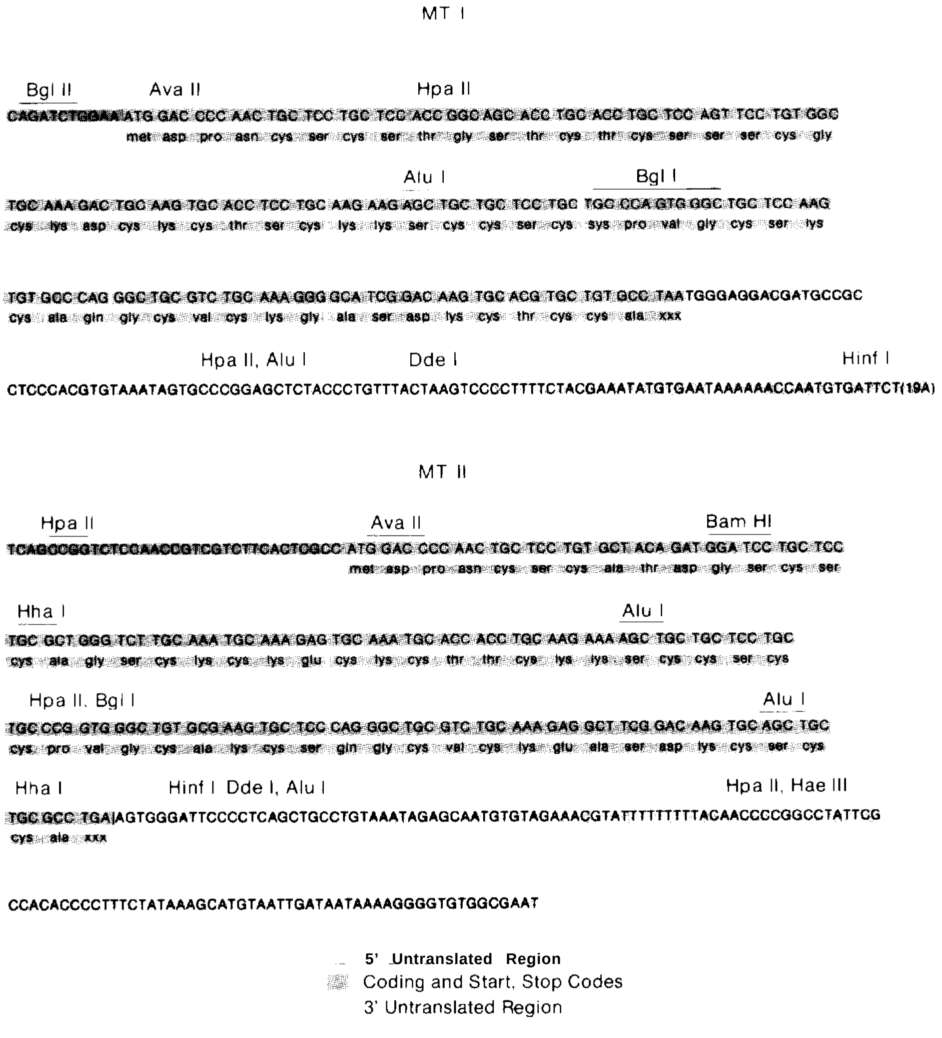
Sequencing the Genes



Genetic information is stored as particular sequences of nucleotides or bases along a strand of DNA. What are the base sequences encoding the metallothioneins? To answer this question we isolated cDNA probes I and II, which encode the two major metallothioneins (MT I and II) of Chinese hamster cells. We then cloned in plasmids the double-stranded form of each of these probes, generating sufficient material to determine base composition and sequence by biochemical cleavage with restriction enzymes. The nucleotide sequence determined for each probe included both the complete protein-coding region and portions of the adjacent 3' and 5' untranslated regions present in processed mRNA.

The accompanying figure gives the base sequence for each of the probe molecules, with C standing for cytosine, A for adenine, G for guanine, and T for thymine. Under each base triplet, or codon, in the coding regions (blue) is the decoded amino acid (abbreviated), These amino acids make up the protein molecules. The first triplet in each coding region (ATG) also codes for the start of protein synthesis; the last triplet (TAA or TGA) codes for the stop of synthesis. The abbreviations above the base sequences represent restriction enzymes that recognize and cleave at the sequence indicated by the line.

The metallothionein sequences can be compared by computer-assisted analyses with each other and with sequences that encode the metallothioneins of other species. Once sequence data are stored, information can be retrieved and homologies calculated using programs such as those developed at Los Alamos for the Genetic Sequence Data Bank. Such homology searches were essential to the development of sequence-specific hybridization experiments, as described in the text. Additionally, computer analysis can be used to define potential restriction-enzyme cleavage sites that can be used in a variety of experiments. Clearly, the establishment of a DNA sequence data base and associated software for sequence analysis is essential to the problem of information handling in molecular biology—especially when one notes the complexity of the sequence data for the metallothioneins, which are relatively short proteins. ■