Sequencing the Genes

enetic 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

MT I Bgi II Hpa II Ava II CAGATCTEGAA ATG GAC CCC AAC TGC TCC TGC TCC ACC GGC AGC ACC TGC ACC TGC TCC AGT TCC TGT GGC met asp pro asn cvs ser cvs ser thr giv ser thr cvs thr cvs ser ser ser cvs giv Alu I BgI I TOC AAA GAC TGC AAG TGC ACC TCC TGC AAG AAG AGC TGC TGC TGC TGC TGC CCA GTG GGC TGC TCC AAG cvs ivs aso cvs ivs cvs in ser cvs ivs ivs ser cvs cvs ser cvs sys oro val giv cvs ser TGT GCC CAG GGC TGC GTC TGC AAA GGG GCA TCG GAC AAG TGC ACG TGC TGT GCC TAATGGGAGGACGATGCCGC cys ala gin gly cys val cys lys gly ala ser asp lys cys thr cys cys dia xxx Dde l Hinf I Hpa II, Alu I CTCCCACGTGTAAATAGTGCCCGGAGCTCTACCCGTTTTACTAAGTCCCCTTTTCTACGAAATATGTGAATAAAAAACCCAATGTGATTCT(19A) MT II Bam HI Hpa II Ava II DEGTEGTETTEACTOGEC ATG GAC CCC AAC TGC TEC TGT GET ACA GAT GGA TEC TGC TCC CARCOGGTOTOGA nel esp pro asn cvs ser cys ale thr asp giv ser cys ser Hha I Alu I TEC GCT GGG TCT TGC AAA TGC AAA GAG TGC AAA TGC ACC ACC TGC AAG AAA AGC TGC TGC TGC TGC cys ala gly ser cys lys cys lys glu cys lys cys thr thr cys lys ser cys cys Ser cys Hpa II, Boi I Alu I THE CCO OTO BOC TOT DCO AND TOC TCC CAS BOC TOC OTC TOC ANA GAS BOT TOG GAC AND TGC AGC TGC cys, pro val giv cys ala lys cys ser gin giv cys val cys lys giu ala eer asp lys cys eer cys Hha I Hinf | Dde I, Alu I Hpa II, Hae III cys ala xxx

CCACACCCCTTTCTATAAAGCATGTAATTGATAATAAAAGGGGTGTGGCGAAT

5' Untranslated Region
Coding and Start, Stop Codes
3' Untranslated Region

be used to define potential restrictionenzyme 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. ■