The EMBO Journal vol.3 no.4 pp.879–885, 1984

Complete nucleotide sequence of a gene encoding a functional human class I histocompatibility antigen (HLA-CW3)

Régis Sodoyer, Michèle Damotte, Terry L. Delovitch, Jeannine Trucy, Bertrand R.Jordan and Tom Strachan*

Centre d'Immunologie INSERM-CNRS de Marseille-Luminy, Case 906, 13288 Marseille Cedex 9, France

*To whom reprint requests should be sent

Communicated by B.R.Jordan

The HLA-CW3 gene contained in a cosmid clone identified by transfection expression experiments has been completely sequenced. This provides, for the first time, data on the structure of HLA-C locus products and constitutes, together with that of the gene coding for HLA-A3, the first complete nucleotide sequences of genes coding for serologically defined class I HLA molecules. In contrast to the organisation of the two class I HLA pseudogenes whose sequences have previously been determined, the sequence of the HLA-CW3 gene reveals an additional cytoplasmic encoding domain, making the organisation of this gene very similar to that of known H-2 class I genes and also the HLA-A3 gene. The deduced amino acid sequences of HLA-CW3 and HLA-A3 now allow a systematic comparison of such sequences of HLA class I molecules from the three classical transplantation antigen loci A, B, C. The compared sequences include the previously determined partial amino acid sequences of HLA-B7, HLA-B40, HLA-A2 and HLA-A28. The comparisons confirm the extreme polymorphism of HLA classical class I molecules, and permit a study of the level of diversity and the location of sequence differences. The distribution of differences is not uniform, most of them being located in the first and second extracellular domains, the third extracellular domain is extremely conserved, and the cytoplasmic domain is also a variable region. Although it is difficult to determine locus-specific regions, we have identified several candidate positions which may be C locus-specific.

Key words: HLA-C locus gene/nucleotide sequence/class I HLA gene organisation/amino acid comparisons

Introduction

The major histocompatibility complex is a cluster of genes playing an important role in the immune response (for recent reviews, see Steinmetz and Hood, 1983; Hood *et al.*, 1983; Coligan and Kindt, 1984). It contains genes coding for class I antigens, including the classical transplantation antigens and also certain differentiation antigens known as Qa, Tla (see below); class II antigens involved in regulation of the immune response; and class III molecules, complement factors (Klein, 1975; Ploegh *et al.*, 1981). We have focused on human class I molecules known to be involved in graft rejection, and also in CTL recognition of virus-infected cells. The class I antigens consist of a highly polymorphic heavy chain (45 000 daltons) transmembrane glycoprotein encoded on chromosome 6 in man, non-covalently bound to a non-polymorphic 12 000--dalton light chain (β2-microglobulin) encoded on chromosome 15 (Goodfellow *et al.*, 1975; Michaelson *et al.*, 1977; Grey *et al.*, 1973; Coligan *et al.*, 1981). Although the heavy chains of the classical transplantation antigens are highly polymorphic, those of the closely similar Qa and Tla antigens, which have been identified in mouse, are much less so (Steinmetz *et al.*, 1981). Classical class I antigens are found at the surface of almost all somatic cells but the cellular distribution of Qa and Tla molecules is more limited. The number of genes coding for class I molecules (in the case of the analogous H-2 complex in mouse) is ~3–5 for classical class I molecules and ~20–30 for Qa and Tla products (Winoto *et al.*, 1983).

In the case of H-2 and HLA complexes, genes coding for classical class I molecules have been isolated and used in transfection expression experiments, resulting in the expression of some of them (Barbosa *et al.*, 1982; Evans *et al.*, 1982; Le Bouteiller *et al.*, 1983; Lemonnier *et al.*, 1983a, 1983b). Several H-2 class I active genes have been sequenced and in all cases the gene consists of eight exons separated by seven introns (Steinmetz *et al.*, 1981; Evans *et al.*, 1982; Moore *et al.*, 1982; Weiss *et al.*, 1983; Kvist *et al.*, 1983; Schulze *et al.*, 1983). For those genes the first exon encodes a signal peptide, the three extracellular domains of the molecule are encoded by exons 2, 3 and 4, exon 5 corresponds to the transmembrane peptide, and the cytoplasmic region is encoded by three small exons. So far, only two human class I gene sequences have been published (Malissen *et al.*, 1982a; Biro *et al.*, 1983). These derive from two genomic clones pHLA 12.4 and LN-11A which represent pseudogenes and apparently contain only seven encoding domains.

In the HLA complex, genes encoding classical class I molecules are split between three loci B, C and A. No information is yet available on the presumptive human Qa and Tla-like genes although class I-like HLA genes have also been shown to map telomeric to the A locus (Orr and DeMars, 1983). Amino acid sequences of the three extracellular domains of some HLA-A and HLA-B molecules have been published (Lopez de Castro *et al.*, 1982, 1983). Here, we present the complete nucleotide sequence of a gene which has been shown by transfection experiments (Lemonnier *et al.*, 1982) to encode an HLA-CW3 molecule. This report and the accompanying paper which documents the complete nucleotide sequence of the gene HLA-A3 (Strachan *et al.*, accompanying paper) provides, for the first time, complete nucleotide sequences of active HLA class I genes. In addition, the present work furnishes the first sequence information about HLA-C locus products whose amino acid sequences are extremely difficult to obtain, because of the very low expression level of those molecules on the human cell surface. We discuss the exon-intron organisation of HLA-CW3 gene and we compare its amino acid sequence with the data available for several HLA class I antigens.

Results

Organisation of the HLA-CW3 gene

The gene is similar in size to the two pseudogenes pHLA 12.4

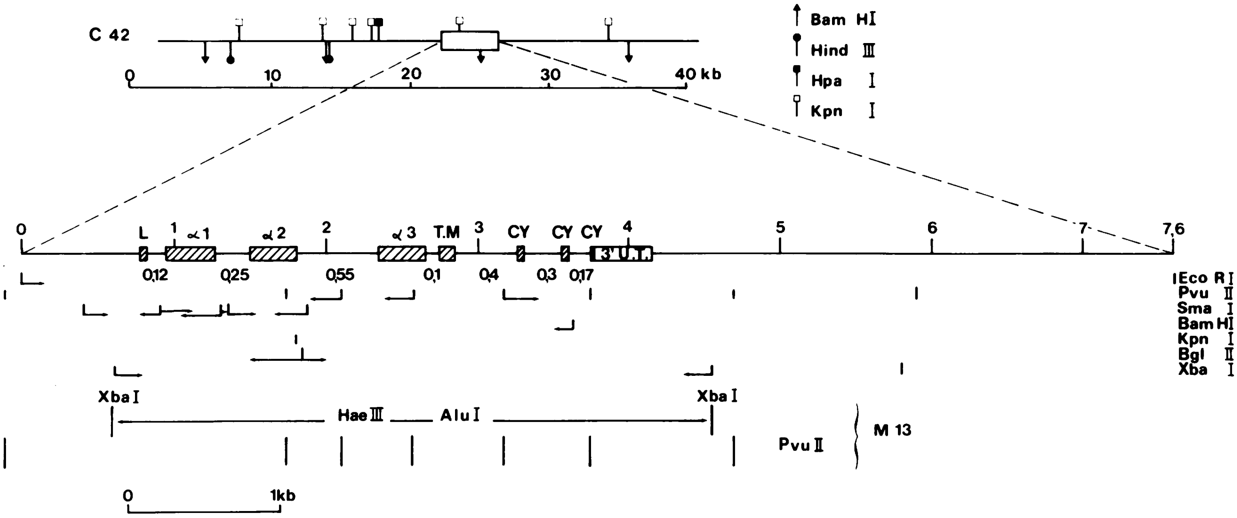

Fig. 1. Partial restriction map of cosmid c42. The 7.6-kb *Eco*RI fragment was subcloned into plasmid pBR328 and studied by restriction mapping, Southern blotting and DNA sequencing. Organisation of exons and introns in HLA-CW3 is: ▨, position of exons; ■, the supplementary encoding domain; □, 3′-untranslated region. Arrows indicate sequence obtained by the Maxam-Gilbert procedure. The 3.9-kb *Xba*I fragment containing the gene was used as a source of DNA for M13 dideoxy sequencing.

and LN-11A (Malissen *et al.*, 1982a; Biro *et al.*, 1983) but its organisation is slightly different. The HLA-CW3 gene consists of at least eight encoding domains and seven introns spanning ~3500 nucleotides (Figure 1). All of the exon-intron junctions are characterized by the splicing signals described by Breathnach and Chambon (1981).

The first exon encodes a typical signal sequence polypeptide (Ploegh *et al.*, 1979). In the H-2 complex those signal sequences of classical class I genes which have been analyzed have been shown to consist of 21 amino acids. For HLA-CW3 the length of the signal sequence seems to be 24 amino acids if it is assumed that the translation begins at the first ATG triplet (Figure 2) (for discussion see Strachan *et al.*, accompanying paper). In addition, the first exon codes for a stretch of hydrophobic amino acids consistent with an α helix structure normally required for this kind of polypeptide (Emr and Silhavy, 1983). The three external domains of the molecule are encoded by exons 2, 3 and 4. Exon 5 encodes a 40 amino acid long transmembrane segment containing an additional amino acid when compared with the nucleotide sequences of the gene HLA-A3 (Strachan *et al.*, accompanying paper) the pseudogene pHLA 12.4 (Malissen *et al.*, 1982a) and a cDNA clone representing HLA-B7 (Biro *et al.*, 1983) and also the HLA-A2 protein sequence (Lopez de Castro *et al.*, 1982). A notable fact is that the supplementary residue results from a single deletion of three nucleotides and two insertions of three nucleotides each (Strachan *et al.*, accompanying paper). A somewhat similar situation is found in the class I genes of the mouse where transmembrane exons code for 39 or 40 amino acids, the two alternatives being related by single insertion/deletion events of three nucleotides. The first part of the transmembrane peptide (Glu, Pro, Ser, Ser) is hydrophilic (considered as a part of the third extracellular domain), its middle part, a stretch of 29 amino acids is devoid of charged residues, while the remaining segment of seven amino acids is also hydrophilic. Consequently, the cytoplasmic segment is really split between four exons (including the last part of the exon encoding the transmembrane segment) instead of three found in the previously determined HLA class I sequences (Malisen *et al.*, 1982a; Biro *et al.*, 1983) and this remarkable feature which indicates for this 'expressed gene' an intron-exon organisation similar to mouse H-2 class I genes is discussed below. The last exon contains the complete 3′-untranslated region terminated by the AATAAA polyadenylation site (Breathnach and Chambon, 1981).

Evidence for an additional encoding domain

Although the cytoplasmic region of mouse class I antigens is known to be encoded by four exons (the last part of the transmembrane exon and three cytoplasmic exons) only three exons appear to encode the equivalent region in the human class I pseudogenes pHLA 12.4 and LN-11A, the second cytoplasmic exon encodes a 14 amino acid peptide and ends with a TGA termination codon. However, in the case of both HLA-CW3 and HLA-A3 (Strachan *et al.*, accompanying paper) this termination codon has been replaced by TGT (Figure 3). Two possible explanations can be considered: the first possibility is that the second cytoplasmic exon is the last exon and is longer in the HLA-CW3 and HLA-A3 genes; in this case a termination codon should be found in the same reading frame. We think this possibility is very unlikely; in the case of HLA-CW3 this would mean an additional eight amino acids and in the case of HLA-A3 (Strachan *et al.*, accompanying paper) an extra 32 amino acids which is highly improbable given the available amino acid sequence data established for class I antigens (Orr *et al.*, 1979).

An alternative explanation is the presence of a supplementary intron. In accordance with this suggestion a potential splicing signal GGT is found only three nucleotides downstream from the TGT triplet for both HLA-CW3 and HLA-A3 (Figure 3). The distance between the termination codon TGA in the second cytoplasmic exon and the polyadenylation site AATAAA in the B7 cDNA clone (Biro *et al.*, 1983) is ~430 nucleotides. The equivalent distance in the case of the gene sequences of HLA-CW3, HLA-A3, pHLA 12.4, HLA LN-11A is ~600 nucleotides. These observations are consistent with the presence of an ≅170-bp intron. In the sequence of HLA-CW3 there are several possible supplementary cytoplasmic exons, one of which is ~170 bp after the potential GGT splicing signal. More convincingly in the case of HLA-A3, the first possibility is found ~170 bp after the GGT splic-

```
AATCTGCGTCGGGTCCTTCTTCCTGAATGACTCATGACGCGTCCCCAATTCCCACTCCCATTGGGTGTCGGACCNNTCTAGAAGGCCGGTCAGCGTCTCC   100

                                             Exon 1 (Signal)    M   R   V   M   A   P   R   T   L   I
GCAGTCCCGGTTCTGAAGTCCCCAGTCACCCACCCGGACTCAGATTCTCCCCAGACGCCGAG ATG CGG GTC ATG GCG CCC CGG ACC CTC ATC  192

 L   L   L   S   G   A   L   A   L   T   E   T   W   A
CTG CTG CTC TCG GGA GCC CTG GCC CTG ACC GAG ACC TGG GCC G GTGAGTGCGGGGTTGGGAGGAATCGGCCTCTTGCGGAGAGG   277

                                                                         Exon 2 (α 1)     G  S   H
AGCGAGGGGCCCGCCCGGCGGAGGGCGCAGGACCCGGGGAGCCGCGCAGGGAGGAGGGTCGGGCGGTCTCAGCCCCTCCTCGCCCCAG GC TCC CAC   374

 S   M   R   Y   F   C   T   A   V   S   R   P   G   R    G   E    P   H   F   I    A   V   G   Y   V
TCC ATG AGG TAT TTC TGC ACC GCT GTG TCC CGG CCC GGA CGC  GGG GAG  CCC CAC TTC ATC  GCC GTG GGC TAC GTG  449

 D   D   T   Q   F   V   R   F   D   S   D   D   E   S    P   R    G   E   P   R    A   P   W   V   E
GAC GAC ACG CAG TTC GTG CGG TTC GAC AGC GAC GAC GAG AGT  CCG AGA  GGG GAG CCG AGG  GCG CCG TGG GTG GAG  524

 R   K   G   P   E   Y   W   D   R   E   T   Q   K   Y    K   P    Q   A   Q   T    D   R   V   S   L
CGG AAG GGG CCG GAG TAT TGG GAC CGG GAG ACA CAG AAG TAC  AAG CCC  CAG GCA CAG ACT  GAC CGA GTG AGC CTG  599

 R   N   L   R   G   Y   Y   N   Q   S   E   A
CGG AAC CTG CGC GGC TAC TAC AAC CAG AGC GAG GCC G GTGAGTGGACCCCGGCCCGGGGCGCAGGTCACGACCCCTCCTCATCCCCC   686
ACGGACGGCCCGGGTCGCCCCAAGTCTCCCGGTCTGAGATCCACCCCGAGGCTGCGGAACCCGAGACCCTCGACCGGAGAGAGCCCCAGTCACCTTTACC   786

                                                                           Exon 3 (α 2)     G  S
CGGTTTCATTTTCAGTTTAGGCCAAAATCCCCGCGGGTTGGTCGGGGCGGGGCGGGGCTCGGGGACGGGGCTGACCGCGGGGCGGGCCAG GG TCT   883

 H   I   I   Q   R   M   Y   G   C   D   V   G   P   D   G   R   L   L   R   G   Y   D   Q   H   A
CAC ATC ATC CAG AGG ATG TAT GGC TGC GAC GTG GGG CCC GAC GGG CGC CTC CTC CGC GGG TAT GAC CAG CAC GCC   958

 Y   D   G   K   D   Y   I   A   L   N   E   D   L   R   S   W   T   A   A   N   T   A   A   Q   I
TAC GAC GGC AAG GAT TAC ATC GCC CTG AAC GAG GAT CTG CGC TCC TGG ACC GCC GCG AAC ACG GCG GCT CAG ATC   1033

 T   Q   R   K   W   E   A   A   R   E   A   E   Q   L   R   A   Y   L   E   G   L   C   V   E   W
ACC CAG CGC AAG TGG GAG GCG GCC CGT GAG GCG GAG CAG CTG AGA GCC TAC TTG GAG GGC CTG TGC GTG GAG TGG   1108

 L   R   R   Y   L   K   N   G   K   E   T   L   Q   G   A
CTC CGC AGA TAC CTG AAG AAT GGG AAG GAG ACG CTG CAG GGC GCG G GTACCAGGGGCAGTGGAGCGTTCCCCATCTCCTGTAG   1192
ATCTCCCGGGATGGCCTCCCACGAGGAGGGGAGGAAAATGGGATCAGCGCTAGAATATCGCCCCTCCCTTGAATGGAGAATGGGATGAGTTTTCCTGAGTT   1292
TCCTCTGAGGGCCCCCTCTGCTCTCTGAGGACAATTAAGGGATGAAGTCCTTGAAGAAATGGAGGGGAAGACAGTCCCTAGAATACTCATCAGGGGTCCC   1392
CTTTGACCACTTTGACCACTGCAGCAGCTGTGGTCAGGCTGCTGACCTTTCTCTCAGGCCTTGTTCTCTGCCTCACGCTCAATGTGTTTGAAGGTTTGAT   1492
TCCAGCTTTTCTGAGTCCTTCGGCCTCCACTCAGGTCAGGACCAGAAGTCGCTGTTCCTCCCTCAGAGACTAGAACTTTCCAATGAATAGGAGATTATCC   1592
CAGGTGCCTGTGTCCAGGCTGGCGTCTGGGTTCTGTGCCCCCTTCCCCACCCCAGGTGTCCTGTCCGTTCTCAGGATGGTCACATGGGCGCTGTTGGAGT   1692

                                   Exon 4 (α3)     E  H   P   K   T   H   V   T   H   H   P   V   S
GTCGCAAGAGAGATACAAAGTGTCTGAATTTTCTGACTCTTCCCGTCAG AA CAC CCA AAG ACA CAC GTG ACC CAC CAT CCC GTC TCT   1779

 D   H   E   A   T   L   R   C   W   A   L   G   F   Y   P   A   E   I   T   L   T   W   Q   W   D
GAC CAT GAG GCC ACC CTG AGG TGC TGG GCC CTG GGC TTC TAC CCT GCG GAG ATC ACA CTG ACC TGG CAG TGG GAT   1854

 G   E   D   Q   T   Q   D   T   E   L   V   E   T   R   P   A   G   D   G   T   F   Q   K   W   A
GGG GAG GAC CAA ACT CAG GAC ACT GAG CTT GTG GAG ACC AGG CCA GCA GGA GAT GGA ACC TTC CAG AAG TGG GCA   1929

 A   V   V   V   P   S   G   E   E   Q   R   Y   T   C   H   V   Q   H   E   G   L   P   E   P   L
GCT GTG GTG GTG CCT TCT GGA GAA GAG CAG AGA TAC ACG TGC CAT GTG CAG CAC GAG GGG CTG CCG GAG CCC CTC   2004

 T   L   R   W
ACC CTG AGA TGG G GTAAGGAGGGGGATGAGGGGTGATGTGTCTTCTCAGGGAAAGCAGAAGTCCTGGAGCCCTTCAGCCAGGTCAGGGCTGAGG   2098

                         Exon 5 (TM)     E  P   S   S   Q   P   T   I   P   I   V   G   I   V   A
CTTGGGGTCAGGGCCCCTCACCTTCCCCTCCTTTCCCAG AG CCG TCT TCC CAG CCC ACC ATC CCC ATC GTG GGC ATC GTT GCT   2182

 G   L   A   V   L   A   V   L   A   V   L   G   A   V   V   A   V   V   M   C   R   R   K   S   S
GGC CTG GCT GTC CTG GCT GTC CTA GCT GTC CTA GGA GCT GTG GTG GCT GTT GTG ATG TGT AGG AGG AAG AGC TCA G  2258
GTAGGGAAGGGGTGAGGAGTGGGGTCTGGGTTTTCTTGTTCCACTGGGAGTTTCAAGCCCCAGGTAGAAGTGTGCCCCACCTCGTTACTGGAAGCACCAT   2358
CCACACATGGGCCCCATCCCAGCCTGGGACCCTATGTGCCAGCACTTACTCTGTTGTGAAGCACATGACAATGAAGGACAGATGTATCACCTTGATGATT   2458
ATGGTGTTGGGGTCCTTGATTCCAGCATTCATGAGTCAGGGGAAGGTCCCTGCTAAGGACAGACCTTAGGAGGGCAGTTGCTCCAACAACCACAGCTGCT   2558
TTCCCCGTGTTTCCTGATCCTGCCCTGGGTCTGCAGTCATAGTTCTGGAAACTTCTCTTGGGTCCAAGACTAGGAGGTTCCCCTAAGATCGCATGGCCCT   2658

                           Exon 6 (C1)     G  G   K   G   G   S   C   S   Q   A   A
GACTCCTCCCTGTCCCCTCACAGGGCATTTTCTTCCCACAG GT GGA AAA GGA GGG AGC TGC TCT CAG GCT GCG T GTAAGTGATGGCG   2745
GTGGGCGTGTGGAGGAGCTGCTCTCAGGCTGCGTGTAAGTGATGGCGGTGGGCGTGTGGAGGAGCTCACCCACCCCATAATTCCTCTTGTCCCACATCTC   2845

                           Exon 7 (C2)     S  S   N   S   A   Q   G   S   D   E   S   L   I   A   C
CTGCGGGCTCTGACCAGGTCTTTTTTTTTGTTCTACCCCAG CC AGC AAC AGT GCC CAG GGC TCT GAT GAG TCT CTC ATC GCT TGT   2930

 K
AAA G GTGAGATTCTGGGGAGCTGAAGTGGTCGGGGGTGGGGCAGAGGGAAAAGGCCTAGGTAATGGGGATCCTTTGATTGGGACGTTTCGAATGTGTG   3028

                                                  Exon 8 (C3+3'UT)      A
GTGAGCTGTTCAGAGTGTGATCACTTACCATGACTGACCTGAATTTGTTCATGACTATTGTGTTCTGTAG CC TGAGACAGCTGCCTGTGTGGGACTGA   3126
GATGCAGGATTTCTTCACACCTCTCCTTTGTGACTTCAAGAGCCTCTGGCATCTCTTTCTGCAAAGGCATCTGAATGTGTCTGCGTTCCTGTTAGCATAA   3226
TGTGAGGAGGTGGAGAGACAGCCCACCCCCTGTCCACCGTGACCCCTGTCCCCACACTGACCTGTGTTCCCTCCCCGATCATCTTTCCTGTTCCAGAGAA   3326
GTGGGCTGGATGTCTCCATCTCTGTCTCAACTTCATGGTGCGCTGAGCTGCAACTTCTTACTTCCCTAATGAAGTTAGGAACCTGAATATAAATTTGTTT   3426
TCTCAAATATTTGCTATGAAGGGTTGATGGATTAATTAAATAAGTCAATTCCTGGAAGTTGAGAGAGCAAATAAAGACCTGAGAAGCTTTCCAGAATCCG   3526
CATGTTCTCTGTGGCTGAGTCTGTTGCAGGTGGGGGTGGGGAAGGCTGTGAGGAGCCGAGTGTGGACGGGGGCCTGTGCCTAGTTGCTGTTCAGTTCTTC   3626
ATGGGCTTTATGTAGTCAGTCCTTAGCTGGGTCACCTTCACTGCTCCATTGTCCTTGTCCCTTCAGTGGAAACTTGTCCAGCAGGAGC               3714
```

Fig. 2. DNA sequence of a 3.7-kb fragment of p42 containing a gene coding for HLA-CW3. Splicing signals, the termination codon and the polyadenylation site are underlined. Amino acid sequences encoded by the exons are shown above the DNA sequence using the one letter code for amino acids (Dayhoff, 1978; see also legend to Figure 4).

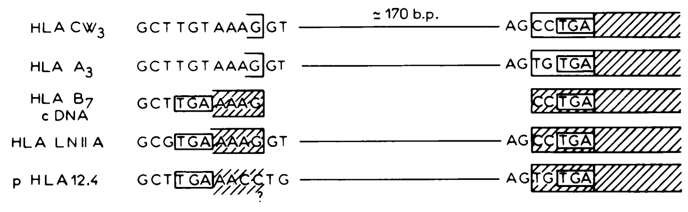

Fig. 3. Determination of the supplementary encoding domain position by comparison of the HLA-CW3, HLA-A3, HLA LN-11A, and pHLA 12.4 gene sequences and the HLA-B7 cDNA sequence. GT and AG dinucleotides indicate the exon-intron junctions. ▨ 3′-untranslated region. ? indicates no evidence for a GT splicing signal in the case of pHLA 12.4.

ing signal. In both cases the candidate supplementary exon is preceded by the requisite AG splicing signal and encodes a single amino acid, either alanine (HLA-CW3) or valine (HLA-A3), a situation strikingly reminiscent of mouse class I genes, of which six are known to code for a single amino acid in the third cytoplasmic exon, either alanine (in four cases) or valine (the other two cases).

This result provokes an alternative interpretation of the HLA-B7 DNA sequence, i.e., there also exists a supplementary exon in the HLA-B7 sequence and in the case of the genes HLA-A3, HLA-B7 and HLA-CW3, and also the pseudogene LN-11A, the position of this exon is conserved (Figure 3). The only exception appears to be the pHLA 12.4 pseudogene which lacks the splicing signal GT although in all cases the requisite AG splicing signal precedes the position of the supplementary exon. Although the HLA-B7 gene and the pseudogene LN-11A resemble the genes HLA-A3 and HLA-CW3 in containing at least eight exons, they differ in having the coding sequence terminated in the penultimate exon.

Discussion

Comparison of the protein sequence of HLA-CW3 with several HLA class I sequences

We have compared the sequence of the HLA-CW3 protein with the protein sequences of HLA-B7 (Orr *et al.*, 1979), HLA-B40 (Lopez de Castro *et al.*, 1983), HLA-A2 and HLA-A28 (Lopez de Castro *et al.*, 1982) and HLA-A3 (Strachan *et al.*, accompanying paper) and also with the protein sequence that would be encoded by the pseudogene pHLA 12.4. Such a comparison was done to locate clusters of variability and also hypervariable positions, which are putative sites for alloantigenic determinants. The distribution of differences is not uniform, most of them being located in the first and second extracellular domains. In the first domain (positions 9–12, 40–45, 52–55, 63–83, conserved segments alternate with variable clusters. In the second domain, differences are located throughout the total sequence which makes the identification of high variability clusters somewhat hazardous. The third extracellular domain is, as expected, extremely conserved (see below). The transmembrane segment displays a large number of differences, but the variable amino acids are alanine, valine, leucine or isoleucine, thereby conserving the hydrophobic character of the central part of this segment. Finally the cytoplasmic domain is also a variable region, as has also been shown in the case of mouse class I genes (Kvist *et al.*, 1983; Weiss *et al.*, 1983). The variability of the cytoplasmic region may reflect a role in its interaction with certain viral antigens (Signas *et al.*, 1982). The repertoire of such interactions is expected to be considerably increased by the possibility of alternative RNA splicing, which has been shown to generate different carboxy termini in the case of the mouse class I H-2K gene (Kress *et al.*, 1983).

HLA-C locus versus HLA-A and HLA-B

The amino acid sequence comparison between HLA-CW3 and the other HLA class I sequences shows that, in general, those positions where HLA-CW3 differs from other HLA proteins correspond to previously determined variable regions. Some additional differences are however apparent: whether they are 'C-locus specific' must await sequencing of other C-locus alleles. Among these additional differences some are positions where CW3 uniquely differs from all the other proteins. These 'C-locus specific' differences are found at residues 21, 40, 52, 54, 55, 173, 181, 183, 219, 268, 308, 321, 338 if the comparison is made with the active genes (see Figure 4), but these differences must be interpreted with caution because only a few amino acid sequences are being considered. In fact, four out of these 13 differences (positions 40, 52, 54, 268) disappear if the comparative analysis includes the pseudogene pHLA 12.4. Locus-specific areas are difficult to find, the only exception being the cluster at position 177–184 which has already been suggested to be a putative locus-specific structural marker of HLA class I molecules (Lopez de Castro *et al.*, 1983).

In addition to the variable regions, there are some conserved or extremely conserved areas. For example, the third extracellular domain is highly conserved, which is consistent with its central importance in the interaction with β2-microglobulin (Peterson *et al.*, 1972; Orr *et al.*, 1979; Yokoyama and Nathenson, 1983). A more remarkable constant region is the hexapeptide Val-Arg-Phe-Asp-Ser-Asp found at positions 34–39. This sequence is conserved in all class I heavy chains and many class II β chains antigens in mouse and man. In addition, those exceptions which have been reported for class II sequences show single conservative substitutions of the type Phe→Tyr, Val→Leu or Val→Ala (Choi *et al.*,1983; Malissen *et al.*, 1983; Saito *et al.*, 1983; Larhammar *et al.*, 1983). This degree of sequence conservation suggests that it plays an important structural role.

Nucleotide sequence analysis of HLA-CW3

The complete nucleotide sequence of the gene coding for the HLA-CW3 molecule illustrates the strong homology between HLA-C locus products and the other classical HLA class I molecules (Table I). Comparisons between nucleotide sequences allow homology studies of introns as well as exons, which may be important for a critical test of hypotheses on the evolution of these genes (Strachan *et al.*, accompanying paper).

The organisation of the genes encoding HLA-CW3 and also HLA-A3 reveals an additional 'encoding domain' at the 3′ end by comparison with known HLA class I genes, and shows for these two genes an exon-intron organisation closely similar to that of mouse class I genes. However, we have not been able to detect a classical promoter site in the available 160 or so nucleotides upstream from the signal exon. This therefore implies either the existence of a non-classical promoter site or an additional exon upstream from the signal exon. In the case of class I genes from mouse which code for the H-2K^{b} and H-2K^{d} alleles (Weiss *et al.*, 1983; Kvist *et al.*, 1983) the transcriptionally important sequences CCAAT and TATAAA are located at positions ~80 bp and 55 bp, respectively upstream from the start of the signal exon. A very similar situation occurs in the case of the two human pseudogenes pHLA 12.4 (Malissen *et al.*, 1982a) and LN-11A (Biro

```
Signal peptide exon
HLA CW3         MRVMAPRTLILLLSGALALTETWA

HLA A3     MARGDQA       L          Q

HLA 12.4        MVL      L          Q

First domain exon
                   10        20        30        40        50        60        70        80        90
HLA CW3    GSHSMRYFCTAVSRPGRGEPHFIAVGYVDDTQFVRFDSDDESPRGEPRAPWVERKGPEYWDRETQKYKPQAQTDRVSLRNLRGYYNQSEA

HLA A3             F S         R                  AA Q M      I QE      Q  RNV A S     D GT
HLA A2             F S         R                  AA Q M      I QE       --- V AH H V  D GT        Z
HLA A28            Y S         R                  AA Q M      I QE       N RNV A S     D GT

HLA B7             Y S         R  S               AA   E      I QE       N  I  A      E
HLA B40            H  M        R  T       L       AT   K      I QE          IS TNT  Y E

HLA 12.4           Y TM    A   R  S                A   E      M  E  K    N  IC A    E EN  IALR      G

Second domain exon
                   100       110       120       130       140       150       160       170       180
HLA CW3    GSHIIQRMYGCDVGPDGRLLRGYDQHAYDGKDYIALNEDLRSWTAANTAAQITQRKWEAAREAEQLRAYLEGLCVEWLRRYLKNGKETLQGA

HLA A3        T  I       S   F    R D             ↓      DM     K      H         D T         E       RT
HLA A2        TLZ        S W-F    - Y          K  /--    -M   T KH     HV          T     --   E---    RT
HLA A28       T --       S   F    R D          K  L      -M   T KH     HV   -      T     - - E       RT

HLA B7        TL S               H  Y                    D                  R      E         E   DK ER
HLA B40       TL                 HN Y                    D     -   L    V          E         E   DK ER

HLA 12.4      TM V          PF    E                      DM     K       R   R V    EF        E       R

Third domain exon
                 190       200       210       220       230       240       250       260       270
HLA CW3    EHPKTHVTHHPVSDHEATLRCWALGFYPAEITLTWQWDGEDQTQDTELVETRPAGDGTFQKWAAVVVPSGEEQRYTCHVQHEGLPEPLTLRW

HLA A3     DP    M    I                        R                                                K
HLA A2     DA    H   A      -                  R                      E          Q              K   ---
HLA A28    DA  ---   A                         -                    ---- V       Q              K  ----

HLA B7     DP         I                        R                   R  E                         K   ---
HLA B40    DP         I                        R                   R                            K   ---

HLA 12.4   DP    M    I                        R

Transmembrane and cytoplasmic exons
               280       290         300       310            320             330       340
HLA CW3    EPSSQPTIPIVGIVAGL(A)VLAVLAVLGAVVAVVMCRRKSS     GGKGGSCSQAA     SSNSAQGSDESLIACK      A

HLA A3      L           I        LGAVIT     A  W          DR    YT          D      V  T         V
HLA A2     ------------------------------------V          DR    Y           D  Z   V  T

HLA B7          S V -  -  -    AV-AVLV-     F                   Y         C D      V  T

HLA 12.4          V              L AV T     A  W K              Y                  V  T
```

Fig. 4. Comparison of the amino acid sequences of HLA-CW3 (this paper), HLA-A3 (Strachan *et al.*, accompanying paper), HLA-B7 (Orr *et al.*, 1979), HLA-B40 (Lopez de Castro *et al.*, 1983), HLA-A28, HLA-A2 (Lopez de Castro *et al.*, 1982) and pHLA 12.4 (Malissen *et al.*, 1982a). The one letter code for amino acids is used: A, Ala; B, Asx; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Z, Glx. Dashes correspond to non-assigned residues.

et al., 1983) and the HLA-A3 gene (see Strachan *et al.*, 1984) where the sequence CCAAT has been conserved at the same position but the TATAAA sequence has been replaced by TCTAAA. However, sequence alignments involving the HLA-CW3 gene reveal substantial differences: the CCAAT sequence has been replaced by CCGGT and the TATAAA sequence by TCTGAA. A classical CCAAT sequence can be detected at an alternative position ~120 bp upstream from the signal exon, while the nearest equivalent to a TATA box is the sequence TAGA which is found ~80 bp before the start of the signal exon.

Finally, establishment of the complete nucleotide sequences of HLA class I genes belonging to different loci is highly desirable for identifying locus-specific sequences or even allele-specific sequences, which facilitates preparation of probes of great specificity. In the case of HLA-CW3 and HLA-A3 for example, the nucleotide sequence between the seventh and eighth exons is variable enough to permit the preparation of specific probes.

HLA class I genes in serological recognition studies

Comparisons of the amino acid sequence of several HLA class I molecules confirm the point that the structural polymorphism of these molecules is extremely pronounced in the first and second extracellular domains (Figure 5). To understand the polymorphism of HLA class I molecules it is possible to modify *in vitro*, in a predetermined way, HLA class I genes which have been previously isolated, serologically characterized and sequenced. Recent results from 'exon shuffling' experiments with HLA class I antigens (Jordan *et al.*, 1983) and with murine class I molecules (Evans *et al.*, 1983) provide information about the parts of the molecule involved in recognition by monoclonal antibodies. Comparison of several HLA class I protein sequences, in the context of secondary structure predictions (Vega *et al.*, 1984) might be helpful in suggesting more precise experiments than simple exon shuffling. We have now initiated mutagenesis experiments in which limited exchanges, involving only a part of one domain or a single amino acid change in a given gene, are effected both by the use of restriction sites (already existing or created by site-directed mutagenesis) or by site-directed mutagenesis using synthetic oligonucleotides.

Table I. Amino acid sequence homology between HLA class I molecules (expressed as percentages)

Exon	2	3	4	5	6	7
	(α1)	(α2)	(α3)	(T.M.)	(C1)	(C2)
Allele						
CW3/A3	78.9	83.7	94.6	72.5	63.6	87.5
CW3/A2	78.1	76.8	91.3	–	72.7	66.7
CW3/A28	78.8	80.0	92.7	–	–	–
C/A	78.2	80.2	92.9	72.5	68.2	77.1
CW3/B7	84.4	85.9	92.4	70.0	90.9	66.7
CW3/B40	80.0	84.8	93.5	–	–	–
C/B	82.2	85.4	93.0	70.0	90.9	66.7
A3/A2	92.0	86.6	92.4	–	90.9	68.8
A3/A28	96.7	94.1	92.7	–	–	–
A2/A28	92.0	95.1	98.8	–	–	–
A/A	93.6	91.9	94.6	70.0	90.9	68.8
A3/B7	84.4	82.2	95.7	70.0	72.7	68.8
A3/B40	78.8	80.2	96.7	–	–	–
A2/B7	83.9	76.8	94.6	–	81.8	85.7
A2/B40	78.1	79.0	93.5	–	–	–
A28/B7	87.7	76.5	91.5	–	–	–
A28/B40	77.7	77.4	91.5	–	–	–
A/B	81.4	78.7	93.9	70.0	77.3	77.3
B7/B40	85.5	94.5	98.9	–	–	–
P12/CW3	74.4	82.6	94.6	77.5	72.7	66.7
P12/A3	71.1	84.8	98.9	77.5	90.9	81.3
P12/A2	72.4	79.3	91.3	–	100	85.7
P12/A28	74.4	82.1	92.7	–	–	–
P12/A	72.6	82.1	94.3	77.5	95.5	83.5
P12/B7	83.3	83.7	95.7	–	81.8	85.7
P12/B40	74.4	80.2	96.7	–	–	–
P12/B	78.8	82.0	96.2	–	81.8	85.7

The regions compared are: the three external domains in all cases and the transmembrane and cytoplasmic region when data are available. C/A, A/A, A/B, p12/A, p12/B correspond to mean values obtained from the appropriate interlocus and intralocus comparisons.

Materials and methods

Reagents and enzymes

Restriction enzymes were obtained from Bethesda Research Laboratories (BRL) and from Boehringer (Mannheim, FRG) and used according to the manufacturer's specifications. The Klenow fragment of *Escherichia coli* DNA polymerase I was purchased from Boehringer. [α-^{32}P]dNTPs (10 mCi/ml, 3000 Ci/mmol) and deoxyadenosine 5'-[α-^{35}S]thiotriphosphate, triethyl ammonium salt (7.7 mCi/ml, 650 Ci/mmol) were from Amersham, UK. M13 primers were purchased from Collaborative Research and Biolabs.

DNA library and screening

HLA class I genes were isolated (Malissen *et al.*, 1982b) from a human genomic library constructed by F.G.Grosveld and R.A.Flavell using the vector pOPFI (Grosveld *et al.*, 1982). The probe used to screen the library was the 5.6-kb *Hin*dIII fragment from plasmid pHLA 12.4 (Malissen *et al.*, 1982a) containing a complete HLA class I gene and devoid of repetitive sequences. The HLA class I positive clones were screened for the presence of active genes by transfection of mouse LMTK$^-$ cells using the calcium phosphate-mediated technique with purified HLA class I gene DNA (Lemonnier *et al.*, 1983b). The corresponding class I molecules expressed on TK$^+$ cell surface were identified using monomorphic monoclonal antibodies, alloantisera and specific monoclonal antibodies (Lemonnier *et al.*, 1983b). L cells transfected with the cosmid clone c42 expressed a human class I molecule which was serologically characterized as HLA-CW3 (Lemonnier *et al.*, 1983b).

DNA subcloning and sequencing

The 7.6-kb *Eco*RI fragment from cosmid c42 (Figure 1), which contained the complete gene as shown by transfection experiments, was subcloned into the corresponding site of pBR328 generating the plasmid p42. The 7.6-kb inserted DNA was separated from the vector by agarose electrophoresis and electroelution and used as a source of DNA for sequencing. DNA sequencing was conducted by both Maxam-Gilbert (1980) and M13 dideoxy techniques (Sanger *et al.*, 1977, 1980; Sanger and Coulson, 1978). In the case of Maxam-

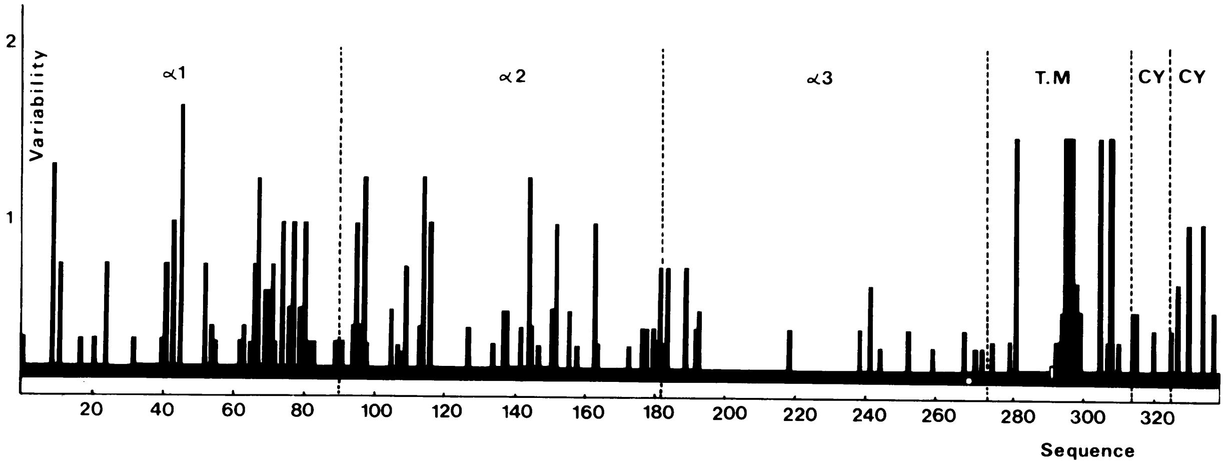

Fig. 5. Variability of the amino acid sequence calculated from data of HLA-CW3, HLA-A3, HLA-A2, HLA-A28, HLA-B7, HLA-B40 and pHLA 12.4 in the three external domains, and from data of HLA-CW3, HLA-A3, HLA-A2, HLA-B7 and pHLA 12.4 in the transmembrane and cytoplasmic regions. The supplementary residue in the transmembrane domain of HLA-CW3 is located at position 292.

Gilbert sequencing, selected fragments were labeled at their 3′ ends using the Klenow fragment of DNA polymerase I. Uniquely labeled fragments were obtained by secondary restriction enzyme digestion or strand separation on 5%, 7% or 10% neutral acrylamide gels. In the case of dideoxy sequencing the p42 DNA was digested with *Xba*I and a 3.9-kb fragment spanning the gene region was isolated, digested with either *Alu*I, *Hae*III or *Pvu*II and subcloned into the *Sma*I site of M13 mp8. Positive clones were selected by hybridization to nick-translated p42 DNA (or region-specific probes derived from p42) prior to isolation of single-stranded templates and dideoxy sequencing using [α-^{32}P]dATP, or more recently [^{35}S]dATP (Biggin *et al.*, 1983). Gels were the thin 0.3 mm. Acrylamide/urea type using an acrylamide percentage of 4%, 6%, 8% and 20% in the case of Maxam-Gilbert sequencing and 6% in the case of the M13 dideoxy sequencing. 60% of the gene sequence was established by the Maxam-Gilbert technique and 95% by the M13 dideoxy technique; in the latter case all the sequence was determined independently on both strands. The sequence homology program of Staden (1982) was used to align overlapping sequences by comparison with the reference sequence of the pseudogene pHLA 12.4 (Malissen *et al.*, 1982a).

Acknowledgements

We thank Dr. F.A.Lemonnier for conducting the transfection expression experiments, and Drs. Haser and Frey (Centre de Recherches sur les Mécanismes de Croissance Cristalline, Marseille Luminy) for conformational calculations. We also thank Drs. M.A.Vega, J.A.Lopez de Castro, D.Larhammar and P.A.Peterson for communication of their results prior to publication. Finally we are grateful to Véronique Bernay for typing this manuscript. This work was supported by grants from INSERM (PRC 134026) and from the Fondation pour la Recherche Médicale. T.S. was the recipient of an EMBO long-term fellowship.

References

Barbosa,J.A., Kamarck,M.E., Biro,P.A., Weissman,S.M. and Ruddle,F.H. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 6322-6331.

Biggin,M.D., Gibson,T.J. and Hong,G.F. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3963-3965.

Biro,P.A., Pan,J., Sood,A.K., Kole,R., Reddy,V.B. and Weissman,S.M. (1983) *Cold Spring Harbor Symp. Quant. Biol.*, **47**, 1082-1086.

Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349-383.

Choi,E., McIntyre,K., Germain,R.N. and Seidman,J.G. (1983) *Science (Wash.)*, **221**, 283-286.

Coligan,J.E. and Kindt,T.J. (1984) *Handbook of Experimental Immunology*, published by Blackwell Scientific Publications, in press.

Coligan,J.E., Kindt,T.J., Uehara,H., Marlinko,J. and Nathenson,S.G. (1981) *Nature*, **291**, 35-39.

Dayhoff,M.O. (1978) *Atlas of Protein Sequence and Structure*, vol. **5**, suppl. 3, published by National Biomedical Research Foundation, MD, p. 197.

Emr,S.D. and Silhavy,T.J. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 4599-4603.

Evans,G.A., Margulies,D.M., Camerini-Otero,R.D., Ozato,K. and Seidman, J.G. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 1994-1998.

Evans,G.A., Margulies,D.H., Shykind,B., Seidman,J.G. and Ozato,K. (1983) *Nature*, **300**, 755-757.

Goodfellow,P.N., Jones,E.A., van Heyningen,V., Solomon,E., Bobrow,M., Miggiano,V. and Bodmer,W.F. (1975) *Nature*, **254**, 267-269.

Grey,M.M., Kubo,R.T., Colon,S.M., Poulik,M.D., Creswell,P., Spinger, T.A., Turner,M. and Strominger,J.L. (1973) *J. Exp. Med.*, **138**, 1608-1612.

Grosveld,F.G., Lund,T., Murray,E., Mellor,A., Dahl,A. and Flavell,R.A. (1982) *Nucleic Acids Res.*, **10**, 6715-6732.

Hood,L., Steinmetz,M. and Malissen,B. (1983) *Annu. Rev. Immunol.*, **1**, 529-568.

Jordan,B.R., Lemonnier,F.A., Caillol,D.H. and Trucy,J. (1983) *Immunogenetics*, **18**, 165-171.

Klein,J. (1975) *Biology of the Mouse Histocompatibility Complex*, Springer Verlag, Berlin.

Kress,M., Glaros,D., Khoury,G. and Jay,G. (1983) *Nature*, **306**, 602-604.

Kvist,S., Roberts,L. and Dobberstein,B. (1983) *EMBO J.*, **2**, 249-254.

Larhammar,D., Andersson,G., Rask,R. and Peterson,P.A. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 7313-7317.

Le Bouteiller,P.P., Mishal,Z., Lemonnier,F.A. and Kourilsky,F.M. (1983) *J. Immunol. Methods*, **61**, 301-316.

Lemonnier,F.A., Malissen,M., Golstein,P., Le Bouteiller,P., Rebaï,N., Damotte,M., Birnbaum,D., Caillol,D., Trucy,J. and Jordan,B.R. (1982) *Immunogenetics*, **16**, 355-361.

Lemmonier,F.A., Le Bouteiller,P.P., Malissen,B., Golstein,P., Malissen,M., Mishal,Z., Caillol,D.H., Jordan,B.R. and Kourilsky,F.M. (1983a) *J. Immunol.*, **130**, 1432-1437.

Lemonnier,F.A., Dubreuil,P.C., Layet,C., Malissen,M., Bourel,D., Mercier, P., Jakobsen,B.K., Caillol,D.H., Svejgaard,A., Kourilsky,F.M. and Jordan,B.R. (1983b) *Immunogenetics*, **18**, 65-78.

Lopez de Castro,J.A., Strominger,J.L., Strong,D.M. and Orr,H.T. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 3813-3817.

Lopez de Castro,J.A., Bragado,R., Strong,D.H. and Strominger,J.L. (1983) *Biochemistry (Wash.)*, **22**, 3961-3969.

Malissen,M., Malissen,B. and Jordan,B.R. (1982a) *Proc. Natl. Acad. Sci. USA*, **79**, 893-897.

Malissen,M., Damotte,M., Birnbaum,D., Trucy,J and Jordan,B.R. (1982b) *Gene*, **20**, 485-489.

Malissen,M., Hunkapiller,T. and Hood,L. (1983) *Science (Wash.)*, **221**, 750-754.

Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499-560.

Michaelson,J., Flaherty,L., Vitetta,E. and Poulik,M.D. (1977) *J. Exp. Med.*, **145**, 1066-1070.

Moore,K.W., Sher,B.T., Sun,H.Y., Eakle,K.A. and Hood,L. (1982) *Science (Wash.)*, **215**, 679-682.

Orr,H.T. and DeMars,R. (1983) *Nature*, **302**, 534-536.

Orr,H.T., Lopez de Castro,J.A., Lancet,D. and Strominger,J.L. (1979) *Biochemistry (Wash.)*, **18**, 5711-5720.

Peterson,P.A., Cunningham,B.A., Berggard,I. and Edelman,G.M. (1972) *Proc. Natl. Acad. Sci. USA*, **69**, 1697-1701.

Ploegh,H.L., Cannon,L.E. and Strominger,J.L. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 2273-2277.

Ploegh,H.L., Orr,H.T. and Strominger,J.L. (1981) *Cell*, **24**, 287-299.

Saito,M., Maki,R.A., Clayton,L.D. and Tonegawa,S. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 5520-5524.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.

Sanger,F. and Coulson,A.R. (1978) *FEBS Lett.*, **87**, 107-110.

Sanger,F., Coulson,A.R., Barrell,B.G., Smith,A.J.H. and Roe,B.A. (1980) *J. Mol. Biol.*, **143**, 161-178.

Schulze,D.H., Pease,L.R., Geier,S.S., Reyes,A.A., Sarmiento,L.A., Wallace R.B. and Nathenson,S.G. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 2007-2011.

Signas,C., Katze,M.G., Persson,M. and Philipson,L. (1982) *Nature*, **299**, 175-178.

Staden,R. (1982) *Nucleic Acid Res.*, **10**, 2951-2961.

Steinmetz,M., Moore,K.W., Frelinger,J.G., Sher,B.T., Sher,F.W., Boyse, E.A. and Hood,L. (1981) *Cell*, **25**, 683-692.

Steinmetz,M. and Hood,L. (1983) *Science (Wash.)*, **222**, 727-733.

Vega,M.A., Bragado,R., Ezquerra,A. and Lopez de Castro,J.A. (1984) *Biochemistry (Wash.)*, in press.

Weiss,E., Golden,L., Zakut,R., Mellor,A., Fahrner,K., Kvist,S. and Flavell, R.A. (1983) *EMBO J.*, **2**, 453-462.

Winoto,A., Steinmetz,M. and Hood,L. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3425-3429.

Yokoyama,K. and Nathenson,S.G. (1983) *J. Immunol.*, **130**, 1419-1425.

Received on 30 January 1984