Volume 13 Number 5 1985 Nucleic Acids Research

Complete nucleotide sequence of a functional HLA-DPβ gene and the region between the DPβ1 and DPα1 genes: comparison of the 5' ends of HLA class II genes

Adrian Kelly and John Trowsdale

Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London, WC2A 3PX, UK

Received 4 January 1985; Accepted 29 January 1985

ABSTRACT

The complete nucleotide sequence of an HLA-DPβ1 gene and part of the adjacent DPα1 gene, up to and including the signal sequence exon, were determined. The sequence of the DPβ1 gene identified it as the DPw4 allele. The six exons of the DPβ1 gene spanned over 11,000 bp of sequence. The arrangement of the gene was broadly analogous to genes of other class II β chains. The β1 exon was flanked by introns of over 4 kb. Comparisons with published sequences of cDNA clones indicated that an alternative splice junction, at the 3' end of the gene, is used in at least one allele. Variation in choice of splice junction indicates an additional mechanism for allelic variation in class II genes. The sequence also indicated that the DPβ1 and DPα1 genes are separated by only 2 kb at their 5' ends. Comparison of the 5' ends of the DPα1 and β1 genes with other class II sequences, including the DZα gene, showed conservation of several blocks of sequences thought to be involved in control of expression. Some areas of the introns were partially conserved in the DQβ gene, and several other intron sequences were homologous to sequences found in other unrelated genes.

INTRODUCTION

The class II HLA antigens are heterodimeric, cell surface glycoproteins, consisting of α and β chains, of approximately 34,000 daltons and 28,000 daltons, respectively (1,2). A detailed understanding of the HLA-D region, containing the class II genes, is emerging from analysis of the glycoproteins and, more recently, their genes and cDNA clones of their transcripts (3,4). Both α and β chains are organised into two extracellular domains. The membrane-proximal domain of each chain shows homology to immunoglobulin constant domains. Both chains have a transmembrane domain of hydrophobic amino acids and a short cytoplasmic tail of charged or hydrophilic residues (3,4).

An important feature of class II antigens is their extensive polymorphism, which is located on the β chains of DR, and both α and β chains of DQ, predominantly in the amino-terminal domains (5-9). At least six HLA-D region α chain genes have been reported, and there are over seven β chain

genes (10-14). The three most clearly established regions containing these genes are called DP, DQ and DR (15).

The HLA-DP region has been analysed in considerable detail, after its original description by primed lymphocyte typing (16). There are two DPα and two DPβ genes, arranged in the order: DPβ2, DPα2, DPβ1, DPα1 (17). The two pairs of α and β genes have their promoter ends adjacent: DPβ1 with DPα1; and, though separated by a larger distance, DPβ2 and DPα2. The DPβ sequences are more closely related to each other than to genes from the other loci, although only an incomplete sequence is published for the DPβ2 gene (17,18). The DPα2 and DPβ2 genes are probably non-functional pseudogenes (18), but DPα1 and β1 have been shown, in transfection experiments, to encode DP antigens which, after expression in mouse L cells, could function to present antigen to appropriate DP-restricted T cell clones (19).

In order to facilitate manipulation of the DPβ1 gene in studies of the regulation of its expression we determined the nucleotide sequence of 15 kb of DNA encompassing the functional DPβ1 gene up to, and including, the signal sequence of the adjoining DPα1 gene, covering the promoter regions of both genes. In this paper the promoter regions of both of genes are compared with those from published class II gene sequences, including the DZα gene.

MATERIALS AND METHODS

Sources of Materials

DH1 bacteria were obtained from Dr. D. Hanahan. RNAase was from Sigma. Merck proteinase K was supplied by British Drug Houses (Poole, England). The Klenow fragment of DNA polymerase, for sequencing, and T4 DNA polymerase, were supplied by Bethesda Research Laboratories. Cosmid LC11 was derived from DNA taken from a lung carcinoma (9,17).

Nucleic Acid Technique

Procedures for preparing plasmids and cosmids, DNA isolation, Southern blot hybridization and ^{32}P-labelled probes have been described in recent papers from this laboratory (9,17,19,20). Any modifications to these procedures are noted in the Text.

Sequencing

Subcloned DNA fragments, as shown in Fig. 1, were prepared from recombinant plasmids grown in the DH1 strain of *E. coli*. After self ligation about 10 μg of each DNA fragment was sonicated four times for 4 secs. in an MSE sonicator, end repaired using T4 DNA polymerase, and ligated into SmaI-cut MP8 vector (21). Random clones of about 300 bp were then sequenced using the chain-termination method (22,23). Multiple overlapping sequences were aligned to provide accurate consensus sequences using computer programs designed by Staden (24). The DBUTIL program was modified by Dr. P. Stockwell (unpublished), to provide a screen-editing system, called VTUTIL. The strategy for DNA sequencing is outlined in Fig. 1.

RESULTS AND DISCUSSION

Overlapping cosmid clones covering the HLA-DP genes were described in a

recent paper from this laboratory (17). One of the clones, LC11, containing the DPα2, DPβ1 and DPα1 genes was used in this work. The nucleotide sequence for the whole of the DPβ1 gene and part of the adjoining DPα1 gene was determined by the chain termination method, using the strategy outlined in Figure 1. This sequence is presented in Figure 2. From comparisons to published sequences for HLA-DPα and β chains it was possible to determine the locations of the transcripts from the genes, as depicted in Figure 2. Detailed analyses of the gene sequences are presented below.

Exon-intron organisation

The exon organisation of the DPβ1 gene and the adjacent DPα1 gene first exon, derived by alignment with cDNA clones, is shown in Figure 2. The DPα1 gene first exon lies 3' to 5' in Figure 2 and has a splice junction at bp 440 and the initiating methionine codon of the signal sequence at bp 540. The exact 5' limit of this exon has not been mapped, however, comparison of DPβ1 with a DPα cDNA clone suggests a 5' untranslated leader sequence of 79 bp and a signal sequence of 31 mainly hydrophobic amino acid residues (Figure 2). Promoter sequences upstream of the methionine codon conform to a pattern conserved between all of the class II genes consistent with this being the start of the DPα gene sequence (see below). The DPα1 and DPβ1 genes lie just over 2 kb apart and are arranged 5' to 5' in respect to direction of transcription of the two genes.

The DPβ1 gene, oriented 5'-3' on Figure 2 (bp 2943-13736) is contained within 6 exons, encompassing approximately 11 kb of DNA. As with other HLA genes, the exons correspond with the envisaged structural domains of the mature protein, i.e. exon 1 comprises the 5' untranslated leader sequence, a signal sequence of 29 predominantly hydrophobic amino acids, and the first five amino acids of the β1 exon. A second possible signal sequence, previously identified, is present at position 6043 to 6136, however, it is not known if this sequence is used (17). It is not directly preceded by promoter sequences characteristic of other class II genes, described in a later section, and has an uncharacteristically high proline content.

The second (AAs 6-93) and third (AAs 94-187) exons encode the two extracellular domains β1 and β2 respectively. Four cysteine residues are available at amino acid positions, 15, 77, 115 and 171 for intradomain disulphide bond formation and a potential carbohydrate attachment site [ASN, GLY, THR] is found at amino acid position 19. This sequence is common to all of the human and mouse class II β chains described so far. Another potential site, [ASN, VAL, SER] is located at amino acid 98, in the β2 domain.

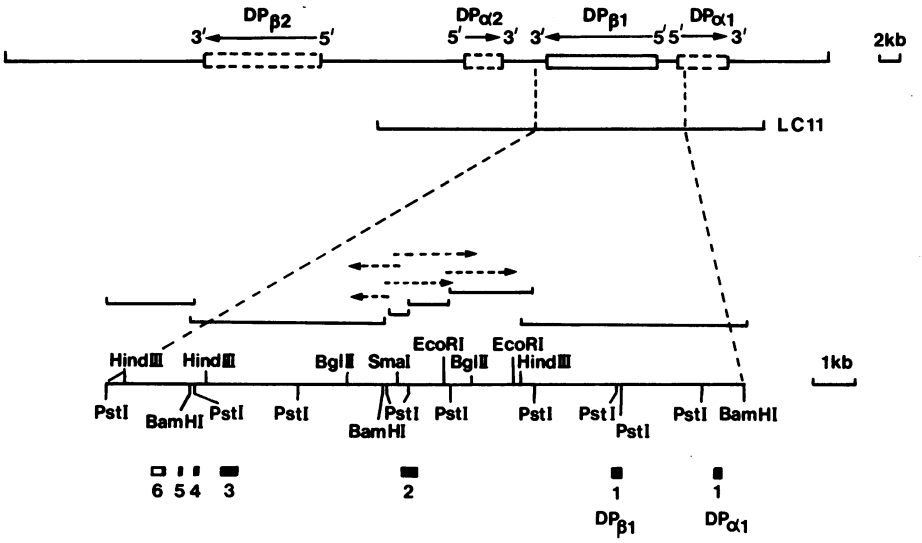


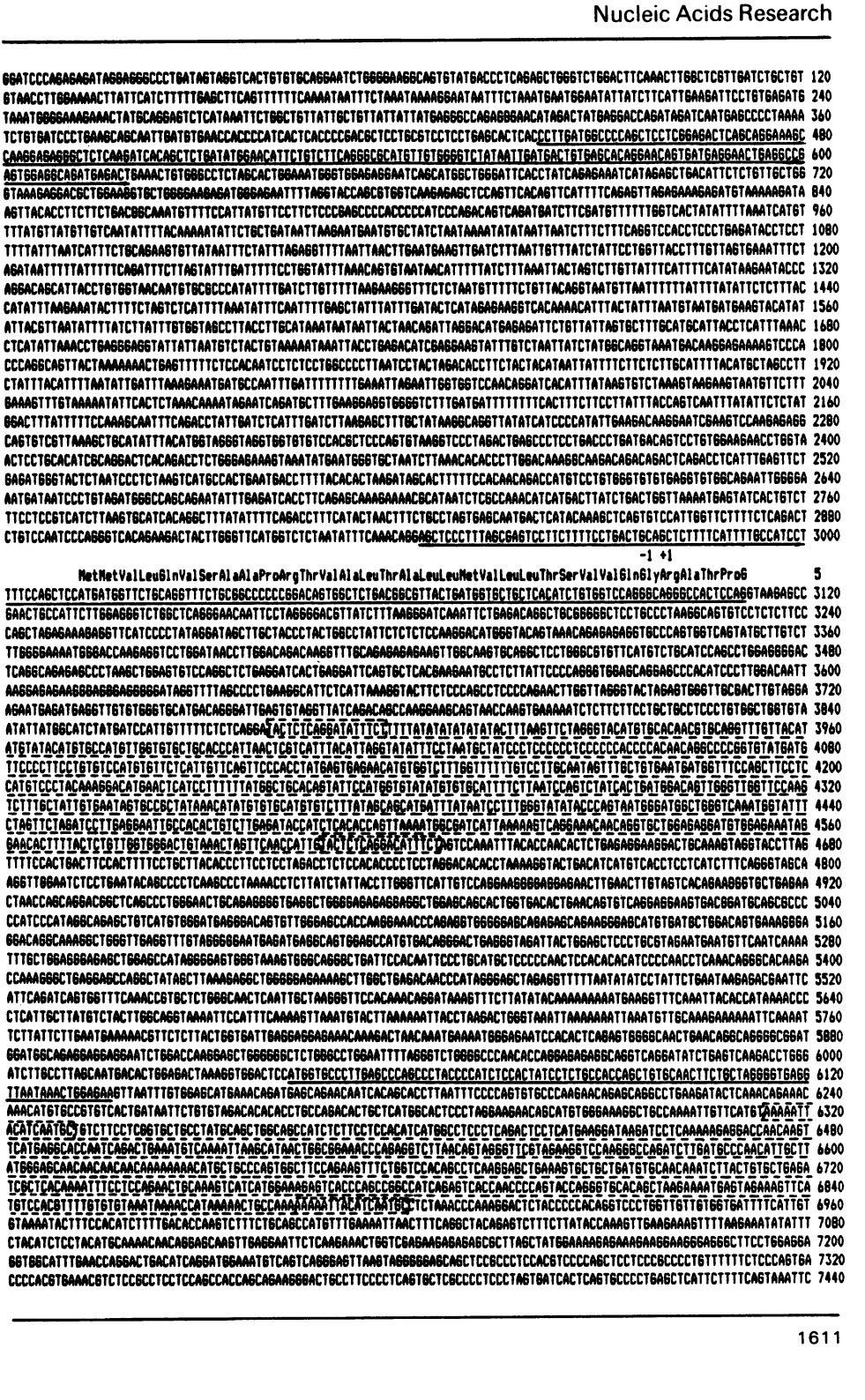
Figure 1

Molecular map of the HLA-DP region and sequencing strategy for the DPβ1 gene and its adjacent DPα1 gene first exon. Dashed boxes show the approximate positioning of DP genes as determined by restriction endonuclease mapping (17). Genes covered by DNA sequencing are shown by boxes with solid lines. The region spanned by cosmid LC11 is indicated and an expanded view of the sequenced insert, with cutting sites of the major restriction endonucleases utilised during sequencing is shown.

Solid lines (———) indicate subcloned fragments used for the generation of random M13 inserts and dashed lines (- - ->) depict specific M13 constructs, that were made using restriction enzyme sites.

Exon 4 encodes amino acids 188 to 224, which comprise the so-called connecting peptide, a transmembrane domain of 22 predominantly hydrophobic residues and 5 amino acids of the cytoplasmic tail. The remaining 6 amino acids forming the carboxyl terminal of the DPβ1 glycoprotein are contained on exon 5, along with the TAA stop codon and 4 bp of 3' untranslated sequence. The organisation of this area of the gene is analogous to that of other class II β chain genes, both human and mouse, except for DQβ, which has a shorter cytoplasmic tail and apparently lacks the splicing signals to bring into play the 5th exon (11,27).

Exon 6 comprises an estimated 220 bp of 3' untranslated sequence, however transcripts from different alleles show some polymorphic variation. All splice junctions within the DPβ1 gene conform to the GT/AG rule, and noticably four out of the five 5' splice junctions comprise the sequence AGGT. Alignment of cDNA clones pHAβ (25) and p11-β-7 (26) with the DPβ1 gene



```
TCTCTCTGCGTGGTGAGAAAACAGGCCTGGAGAGGCTCTGCGACCCGCTTAGGACCACAGAACTCGGTACTAGGAAAACTCCTATTTTAAAATCCAGCCCTGGGTGGGAAGATTTGGGAA 7560
                                                                                       luAsnThrLeuPheGlnGlyArgGlnGluCysTyrAl 17
GAATCGTTAATATTGAGAGAGAGAGGGAGAAAGAGGATTAGATGAGAGTGGCGCCTCCGCTCATGTCCGCCCCCTCCCCGCAGAGAATTACCTTTTCCAGGGACGGCAGGAATGCTACGC 7680
aPheAsnGlyThrGlnArgPheLeuGluArgTyrIleTyrAsnArgGluGluPheAlaArgPheAspSerAspValGlyGluPheArgAlaValThrGluLeuGlyArgProAlaAlaGl 57
GTTTAATGGGACACAGCGCTTCCTGGAGAGATACATCTACAACCGGGAGGAGTTCGCGCGCTTCGACAGCGACGTGGGGGAGTTCCGGGCGGTGACGGAGCTGGGGCGGCCTGCTGCGGA 7800
uTyrTrpAsnSerGlnLysAspIleLeuGluGluLysArgAlaValProAspArgMetCysArgHisAsnTyrGluLeuGlyGlyProMetThrLeuGlnArgArgV            93
GTACTGGAACAGCCAGAAGGACATCCTGGAGGAGAAGCGGGCAGTGCCGGACAGGATGTGCAGACACAACTACGAGCTGGGCGGGCCCATGACCCTGCAGCGCCGAGGTGAGTGAGGGCT 7920

TTGGGCCGGCGGTCCCAGGGCAGCCCCGCGGGCCCGTGCCCAGGGCGCAGGAGCAGCCGGGTTGGCCTAAGGGACCTTAGTGCCGGGCGGAAAGGGGACTTTGGGTTGGGGATTCATGGG 8040
GGGAGCCCATCTGGAGCTTGTCAGGGGAGCGAGCGCGGGGACCTGGACTGGGCTGAGCATGGAGTGAGGAGGACGAGAGCAGAGAGACCCCCGGGAGCTTCATCAGGCCTGGCAGCTGAC 8160
TGCATGTGGGGTGAAAAAAGGAAGCCACAGGACAGCGCACAAGGGTATGGTGTGGAGATGGAGGTGGAGATGGCACAGCAGGCCACACAGAGAAGAAACCTACAGGGAGGTAGCTGGGTT 8280
TGAGGTGCTTGAGGGGCAGATGGGTGGTCTGATGGGCAGGTAGACAGAAGGGTCTGCAGCCGGGGAGGAGACTGAGATACATGAGACCATCCAGGGAGAGGGGACCCAGGGGGGAAGAGCA 8400
AAGGACCGGATCCTGGGAACTGGACAGTTGTGATTTGGCCAAGACAGAAAAGCCTGTGAAAGAGACCAAAAAAACCCAAGTGCAGTGTGAGGAGAGGCCCGCAGAGAAGAGTCTTGGAAG 8520
CTGAGGGGAGGTGACCTCAGCAGCACAGTGGACAGCGGTGCCAGTGACTTGGGAAGGTCAGAAAACAGAAGATGGAAAGTGGGTTTGGAAACCAGGGAGACCTGGGGAGAGCAGGTTGGC 8640
CGCAGCGGCAGGAGCTGGAATGGGAGGGGGTGCATGAGGCTGAGTGTGGCGCATCCTCCTCGGGGCTGAGATGGATTTTACTTGTCTTGGGTTCCCCACGGCTGTCACAGGGCAGTGTCT 8760
CAGTTCATTCGTCTTTTTCCTTCAGGAAGTCTGGGTGTAAAGGGATGGAGAGAGGTGAGGTGTGTGCAGTAAGAGGATTTCTCAAGGATGGGACAGGAAGGCCTTGGAGCTTTGGCTTCC 8880
TCCTGTGAACTTGTGGGGTGGGGAGCCTGGTGCACCAACCTGAGGGACTTGAGGGAGTAGTATCAGGATGTGGGATTGAGCCCTGGACCTTTTTTCTAGAAAGAGGAAAAAAATGAAGGG 9000
AGGAGGAGGAGGAAGCTGGGGAGATCACACCTTTGATTTTCTTGTTCCTGGAAAGTGAAAGGAAGTTCACCTGCTATGAGTGAGAAGGTGGACACACTGGGTGGGGATGAGGTGAGTGAC 9120
ATGAGCTTAGGAAAGTTGCTGAGGTAATTGGTTGAGAGAGGTGTTCAAATAAAAATAACGCAATTGGCAAAAACTGTTACTAAGACTTTGTAGAGGCACCAATCAGTGACATGGCAGCAT 9240
TTTCTTTCACAGTAATCAACTGCCAGATTGCAGACAGCCCTGATGCCAGCCTAAGGAGTGTGGGTTTCTCCTCCAGGCCCGCAGGTCCCCAACCTCACTCCTCTGAAGACTCTTCTGGAG 9360
ATCCTCTGTGATGCACAGATCTCCAGACTCAGTGCCCCCAGACTCAGATTCCCTGGGTGGGGAGGTCTGGGGATCTCTGCTTGTAATCAGCTCCCTAGAGGTTCCCATGTAGCCAGATAA 9480
GTATTGTCAGAACACTGAAGATTTTTGAAAAATGAAAAAGAGAAGGTTGGAGATGTGTCTTCAGAAGACTACTAAGGGTGCTGGCTAGAGGAGGGACCAGAGGCAGGGAGATGAGGTAGG 9600
AAACTGCTATTATTTGTCAGGGAAATTGCAATCAAGGCATGAGTTAGAACAGGGAAAACACAGAGGCAAGGGAGAGGTGGAAGGGGGGAGGAAAGAAGTAGTGACAATTCCAGGGTGGATG 9720
TCCACCCAAATCTAGAAGTAATTGAGCAAATGTTTTCTGGGCATTAGAGAAGGCAACTAGAACAAACAGGAATCCTTGCCTTGGTGAAATGTATTTGAACTGGGTCAGAAATGAGGCCAT 9840
TGGGTATCAGGCCTTAACTCCAGCGCACCCTGGAGGTCACTGATGTGGCTCCAGGCTGACCTGCTCCTGTCAAAGAATATTGAGCAAGATGCCTCTCGTGGAATGTTCTGGGACCTTAAA 9960
ACAGATACCCAAGTATTCCCCCTGATTTCATGGTTCCCAGAAGCTCTATGGGGAAGAAATTGTAGGTAATTCACAACTGAGATTTAGACATAAGTTGAATAGTGTAATGGACATTGAGTT 10080
AACCGAGGTAATGAAGTAGTGAGACACAGGTGCCCCTGAAATAAACTCACATTGAGGGAAGAGGCTGACAATGTGGATCAGTCTGAAAACAAGGCAAAAATACAATAGGGAGTAAGGGTT 10200
GTGTGTCAGTTCAAGACTGTACTTTTACCTGGCCCAGCGCCATGTTAGGGTATTTGTGTTCTCCAGGAAGTAGAAAGGAAAGAACTGAGTGATTAGGGACCTAGAAGACTAATTTGAGAC 10320
ATTCCTCTTGATGAGCTGTTCTCTAGGGTAGTCCTCTGAAAGAGCTGTTCTCTAGTGGATCTCCCTGAATGAACTGTTCTCTAGGAGCACTTGACCTTTTCTGTGTTTGTTTTTTGTTT 10440
TGTGTTTGTGTTTGTTTTTGAGACAGGTTCTCACTTTGTCTCCCAGGCTGGAGTGCTGTGGCACCATCATGGCTCACTGCAGCCTCAACCTCCTGGGCTCAAGTGATCCTCCTGCCTCAG 10560
CCTCCCATGTAGCTAGAACTACAGATACACGTACCACCATGTCTGGCTAATTTATTTTTCTTTTTAGAGATGGGTTCTCACTATGTTGCCCAGGCCGGTCTCAAAACCCTGGGCTCAAGT 10680
GATCCTCATGCCTCAACCTCCCAAAGTGCTAAGATTATAGGCATGACCACCATGCCTGGCCTTTTCTGCTTTCTGAGGAGGAAAAAGGTACTGGTGGCAGAGATCCAAAAGAAAAGTTGC 10800
CAGTGGCAGTGTGGAAATTCACCTGAGAACAACAGGACAAGCTGGGGCACAAATGCAAAGATGCAGAGGGAGGCAACACCTGGTCATCTGTGAGACCTTCATGGGACCTGAAGACGCAGC 10920
ACAGAGGAGGAACTTGAAAAAGGACGGGATTTCTACTACTCAAGCATGTAGGAGCTCAGGATATTCTGTAAATATGAAGATTTTGAGTTTTTGTAGGTGAGGTAAAAAAATACATAGGTT 11040
TTTTACAGAATAAGACATGTAAAGCTCTCTTCATTTTCTTTGTATTTTCATGAAGTTATTAGATTCACAGGCCACCATAATGCCATTGTCTGTATATCTTAATTTCAAGATATTATTTGA 11160
GTAAATTTTGCTTCCTTTGTATCAAGATAGAACTTTGAAAAGGTAGGTAATTTCACAGTTGATCAAATATTCTTTGCCCAAATTACTTTTGGTTAAAATTTCTCCTAAATGTGCTACAGA 11280
GTGCAAACTCTGTCTCCCTGCCATTCCGCTATATACTTACTAACTATTATTTTATTCAAGATCATGCATGCTCTACTTGAAGGTCTATTTCTATCTTTTCAATGCTACCCTTACCCACTA 11400
GCCTAATCACATTATTCCTATTTTCAACATCTAGGAATCAATTACATAGTGAACATGCCTAAGAAATAATAATCTGGGCAGATGCAGTGGCTCAGGCCCGTAATCCCAGCCCTTTGAGAG 11520
GCCGAGCGGGTGGATCACTTGAGGTCAGGCGTTGGTCAAGTGCTCCTAGAGAACCAGGCTGACCAACATGGAGAAACCTTGTCTCTACTAATAATACAAAAATTAGCCAGGTGAAGTGGC 11640
AGGCACCTATAATCCCAGCTATTCGGGAGGCTGAGGAAGGAGAATTGGTTGAAGCCCAGAGGTGGAGGTTGCAGTGAGCCAATATTGCGCCACTGCATTCCAGACTTGGCAACAGAGTGA 11760
CACTCCATCTCAACAAAAAGAAAGAATGAAAGAAAGAAAGAGCGAGATTATGTCTCAAAAAAAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGACAATC 11880
                                       alGlnProArgValAsnValSerProSerLysLysGlyProLeuGlnHisHisAsnLeuLeuValCysHisValThrAspPh 120
TCAAATTCTATTTCATTATTTTTCTTCCACGCTCCTAGTCCAGCCTAGGGTGAATGTTTCCCCCTCCAAGAAGGGGCCCTTGCAGCACCACAACCTGCTTGTCTGCCACGTGACGGATTT 12000
eTyrProGlySerIleGlnValArgTrpPheLeuAsnGlyGlnGluGluThrAlaGlyValValSerThrAsnLeuIleArgAsnGlyAspTrpThrPheGlnIleLeuValMetLeuGl 160
CTACCCAGGCAGCATTCAAGTCCGATGGTTCCTGAATGGACAGGAGGAAACAGCTGGGGTCGTGTCCACCAACCTGATCCGTAATGGAGACTGGACCTTCCAGATCCTGGTGATGCTGGA 12120
uMetThrProGlnGlnGlyAspValTyrThrCysGlnValGluHisThrSerLeuAspSerProValThrValGluTrpL                                        187
AATGACCCCCCAGCAGGGAGATGTCTACACCTGCCAAGTGGAGCACACCAGCCTGGATAGTCCTGTCACCGTGGAGTGGAGTGAGTCTCTGATGACCCTCTAGACCCCACCTCTGAAGAG 12240
CAGGGGACTCTCTGGCTCTGGGGTCCACTCATCTTATCTTCTGCATCTATACCCTGGGGCCATGTCCAAACCCCATCTTTCTTCTATACCAGCTCCTGAGCATAGTTTGAAGCCAGGGAA 12360
ATGGAGACTTCCTGACCTTGGCTTAGGGGTTCCTGAAGATTCATAGTTCTCCCCCTTGTCAGAGAATCTAGGGACACTGACTGGTCTCGAAACCCTCACACTTAGGAACTGACCTCACAC 12480
ATAGGAACAGTTCTCTTCCTTCAGCATTTTAGCCTCTTCTCAGGCATTTTGAGAGGCAACTTCCAGAATCAGCATTTGCCACCTTGTTGAGGTCACACCCCTGTTCCAGATATGAGGGTG 12600
GCTCTTTCTGAATTTCCTCTTAGCAAGCTTTTTCCGCTGCACTGTCCTCATCCCGATATGCTGCATCAGGCTCCAGAATCTCAGACAGGACATGAGTAGGGATGCAGCTGGTGGAGGTGA 12720
                              ysAlaGlnSerAspSerAlaArgSerLysThrLeuThrGlyAlaGlyGlyPheValLeuGlyLeuIleIleCysGlyValGlyIlePheMetH 218
CACTAAACCTGGGTCTGTCCTTCCCAGAGGCACAGTCTGATTCTGCCCGGAGTAAGACATTGACGGGAGCTGGGGGCTTCGTGCTGGGGCTCATCATCTGTGGAGTGGGCATCTTCATGC 12840
                 [GlyGluLysAlaCysArg]                                                                                   [229]
isArgArgSerLysLysV                                                                                                      224
ACAGGAGGAGCAAGAAAGGTGAGAAAGCCTGCAGGGTGAGCGGGACTTACCTTCCCCTGGCATATTCACACTTATTCCACGATGAGGGGTTTGACAGAAAAGAAATGTCAGAAAGCTCTA 12960
GAGGCCACTGATATCAGATAATCGGGGAACAAACATGACCTATAGCGAGAGAGGGATCCCAGGCTGGGATCTTAATGCAGCCAGATGCATGAGGTCCCAAGTACTCAGGCTCCTGCGGAG 13080
                                                                                                          [PheAsnGluAspL [234]
                                                                                                           alGlnArgGlySe 228
CGTCCATTGAGTGATGGGCAATGGAATTTGGTGGGATGGAAATGTTTCTCTAATTATCTGAGGTGGTTTCAATGGCTGATTATATAACCTTTCGTCTTTCATTTCAGTTCAACGAGGATC 13200
euHisLysGln]                                                                                                            [237]
rAla***                                                                                                                 229
TGCATAAACAGGTAATATTCCTGCTTTGATTTCCTTGTGGGGTGGGTTGCAGGAGGATATGAGTCCTTTCTGTGCATTGTAACACTGAGGCTCCTCCAGGAAGGGAATCTCAGGCATGAA 13320
CCCCTCTTTCAATGTCAGCCTTCAGGCAAGTGGGGAAAGAGCATTGCTTGGCTCCATTGCTGAAGGAAGCAGAGATCAACTCTGTTATTTATCAGCCTGAGACGCATCCTCTCACCATAA 13440
```

Figure 2

The nucleotide sequence of the DNA fragments of cosmid LC11 encompassing the complete DPβ1 gene and the first exon of the DPα1 gene. Transcribed regions, as determined by comparison with cDNA clones, for DPβ (18,25,26) and DPα (pSBα-318; H. Ehrlich, personal communication), are underlined. (The putative signal sequence (bp 6043-6136) previously identified is also shown; 17). The most probable transcription product of the DPβ1 gene is indicated and an alternative transcript produced by a differential splicing at the 3' end of the fourth exon, as seen in cDNA clone p11-β-7 (26) is shown in brackets. Termination codons (xxx) and polyadenylation signals AATAAT and AATAAA are shown. Alu and other repeat sequences are underlined (_____) and their terminal repeats boxed. For further details, see Text.

from cosmid LC11 shows that the variation in both length and composition of the cytoplasmic domains of these clones appears to be a result of an alternative splicing event at the 3' end of exon 4. The splice junction (AGGT) present at position 12859 in the DPβ1 genes appears to have been used in the processing of the pHAβ transcript, whereas, a second splice site (GGGT), located 17 bp downstream was used for the p11-β-7 transcript. There is a mutation in p11-β-7 which has altered the second splice juntion from GGGT observed in the DPβ1 gene, to AGGT. This interesting observation indicates that there is polymorphism for choice of splice junction in the DPβ genes. There are no indications that either of the variations prohibits the function of the resulting glycoprotein chain, but the substantial differences between the 3' ends of the transcripts from different alleles may result in subtle effects that remain undetected.

Variability in the 3' end of the DPβ1 gene is further complicated by the observation that the atypical polyadenylation signal AATAAT, present in the DPβ1 gene at bp 13723, and in the corresponding position in the transcripts for both p11-β-7 and pHAβ, appears to have directed polyA addition to two different sites, approximately 30 bp apart (bp 13736 for pHAβ transcripts and bp 13759 in p11-β-7). A second polyadenylation signal, apparently not represented in any of the published cDNA clones, is present

101 bp downstream of the first, at position 13824. Additional and, in some cases, alternative polyA+ addition sites are a common feature of several eukaryotic genes, for example DRα (20) and dihydrofolate reductase (28).

The available evidence indicates that DPα1 and DPβ1 are functional genes. It is known, for instance, that transcripts identified as DPα and β correspond closely to the DPβ1 and DPα1 sequences and to available protein sequence data (see legend to Figure 2 and ref. 29). Moreover, when transfected into mouse L cells, cosmids containing the DPβ1 gene (in conjunction with DPα genes 1 and 2) gave rise to expressed HLA-DP glycoprotein on the cell surface. The antigen so produced was shown to be capable of presenting antigen to an appropriate DP-restricted T cell clone (19). From sequence data DPα2 and DPβ2 are pseudogenes so it seems fair to assume that all of the published transcripts originated from the DPβ1 gene (18).

Southern blots were used to demonstrate that there are only the two DPβ genes per haploid genome. If, therefore, only one pair is functional, one has to find other explanations for published evidence that more than one DP locus is expressed. This was suggested by primed lymphocyte tests (30), and by the fact that the monoclonal antibody ILR1 binds to B cell mutants that lack one complete haplotype, and only have the DPβ2 gene in the other (31). Unless the DPβ2 gene is functional in some haplotypes (and this has not been completely ruled out), it seems possible that there are additional class II genes centromeric to DPβ2.

Partial sequences of several different DPβ1 alleles have been determined (18). Comparison of the sequence of the DPβ1 gene on Figure 2 with DPw2, w3 and w4 sequences revealed an exact match with the coding regions of DPw4, and only 4 nucleotide differences, in intron regions. On the other hand, there were numerous differences between our sequence and those of the other two alleles. In addition, comparison of 3 kb of our sequence with that of another, independently derived DPw4 allele, identified only one nucleotide difference, in an intron (K. Gustafsson and D. Larhammar, personal communication). The gene that we have sequenced, from an untyped lung carcinoma, is therefore identified as DPw4. Indeed, as pointed out by Kappes et al., there is remarkably little sequence difference between DPw4 alleles (18). Even in the introns, from the sequences we have compared, the number of substitutions is less than 0.2 percent.

Intron sequences

A computer search of the current nucleotide sequence databanks (EMBO -

version 4, GENBANK - version 25), with the sequences shown on Figure 2, using the Wilbur and Lipman algorithm for rapid sequence comparisons, revealed some interesting sharing of sequences with those from other genes (32). The presence of a processed pseudogene, flanked by a 17 bp direct repeat sequence, about 700 bp upstream of the β1 exon, has already been described in detail (17). This has now been identified as a pseudogene for ribosomal protein L32, details of which will be published elsewhere (J. Young and J. Trowsdale, manuscript in preparation). Some other interesting matches are outlined on Figure 2. They include two Alu repeat units present in the intron separating exons two and three. The first (bp 10416-10748) matches 241/314 bp of an Alu repetitive sequence present in the 5'-flanking intergenic region of a pseudo alpha-globin gene (33). It possesses a 9 bp direct repeat at each end (CCTTTTCTG), is composed of two homologous units approximately 150 bp long, and shares 76% homology with a 114 bp section of the second Alu unit (bp 11475-11775). This latter repetitive sequence, also composed of two roughly equal length (150 bp) units, apparently lacks terminal repeat sequences but is flanked at its 3' end by a 102 bp highly deoxyadenosine rich region (57%) which includes the sequence (GGAA)11. All these characteristics are common to repetitive DNA sequences (34).

A fourth region of repetitive DNA (bp 3880-4625) flanked by an 18 bp terminal repeat, with a single mismatch (TACTCTCAGGACATTTCT), is present in the first intron of the DPβ1 gene. This sequence, containing several shorter internal repeats, appears to comprise a central Bam5 element approximately 250 bp long (67% homologous to 195 bp of a Bam5 determinant described by Wilson and Storb (35)) flanked on one side by a sequence 93% homologous to a 113 bp region of repetititve DNA associated with the leukocyte interferon gene cluster (36), and on the other side by DNA, presumably also of the middle repetitive class, which is over 79% homologous with a 138 bp sequence present in the first intron of the mouse kallikrein gene (37). It seems likely therefore that this sequence has evolved through multiple insertion of one repetitive sequence into another.

The function, if any, of such sequences is unknown, however Singer et al. have noted a specific association of classes of repetitive DNA with HLA genes and it is possible therefore that they have some function associated with the regulation of expression, or the polymorphism of these genes (38). Recently, the functional insertion of an Alu type 2 sequence into a murine class I gene has been reported (39).


CONSENSUS A
CONSENSUS B
Alpha consensus
I-E alpha
DR alpha
DZ alpha
DP alpha
JOINT CONSENSUS
DP beta
DQ beta
I-A beta
I-E beta
Beta consensus
CONSENSUS A
CONSENSUS B
CAT CONSENSUS
TRANSCRIPTION START
INITIATION CODON
<--15/16 bp-->
<--13-21 bp-->
CCAATCC
<--34-63 bp-->
<--47-78 bp-->
ATG
<---35 bp---->
<---37 bp---->
TGCTGATTGGC
GCTGATTGG
GCTGATTTGCAT
CTCATTTGCAT
<----------162 bp----------(E.coli gsh-11-------->
<----------156 bp----------(H2A histone)--------->
<----------~90 bp----------(Vλ and Vκ Ig light chain)->
<----------~125 bp----------(H2B histone)--------->

Promoter regions of class II genes

As previously mentioned, the precise limits of the 5' ends of both the DPβ1 and DPα1 genes have not been determined although if it is assumed that cDNA clones pDD2 and pSBα 318 (see Figure 2), are full length, then untranslated leader sequences of 79 and 69 bp, respectively, could be predicted. This compares favourably with an estimate of about 63 bp obtained by comparison of class II sequences in general (Fig. 3), where cDNA clone analysis, primer extension and S1 mapping studies have been used to locate probable sites of transcriptional initiation. Examination of sequences 5' to the initiation MET codon of class II sequences from both man and mouse (Fig. 3) reveals several regions of strong sequence conservation. In particular two regions (A and B, Fig. 3) present in DPα1 and to a lesser extent DPβ1, bear striking homology to upstream promoter region sequences, previously reported as conserved between I-Eα, DRα and I-Eβ genes (42-45). Although the homology of DPα1 with these sequences was immediately apparent, a more extensive comparison of class II genes was required to reveal the corresponding blocks in the DPβ1 gene. It is evident from Figure 3 that whilst there is a high degree of inter group (α and β chain) sequence conservation, in the above mentioned locations, there is an even greater level of intra group (α or β chain) homology. This is highlighted firstly by the exact sequence conservation observed between the DZα and DPα genes in the larger conserved region and secondly by the observation that the distance separating the two conserved regions is kept constant at precisely 16 bp in all the α chains and 15 bp in all the β chains.

Figure 3

Alignment of sequences upstream of the ATG initiating methionine codon of human and mouse class II genes. Areas of sequence highly conserved between: I-Eα and DRα, DQβ and I-Aβ, and DZα and DPα, are shown aligned (dashed boxes). Blocks of sequence strongly conserved in both alpha and beta chains are shown boxed (solid lines). The alpha/beta chain consensuses define positions with 100% nucleotide conservation whilst the joint consensus defines positions where a single base occurs in >75% of cases or two bases occur in 100% of cases. The ATG initiating MET codons are underlined. Putative transcriptional start positions (^), as defined by cDNA clones, primer extension studies or S1 mapping are shown, and potential CAT/TATA sequences are also underlined. Sequence information was from the following sources. DQ β (11), I-Eβ (42), I-Aβ (43), DRα (44), DZα - Trowsdale and Kelly (manuscript in preparation). I-Eα (45).

The joint consensus is drawn again, underneath the main figure, but including some frequent alternative nucleotides and additional positions. The upstream sequences from the following genes are also given, for comparison: E. coli glutathione synthetase gsh-II (49); sea urchin histone H2A (48); human and mouse V_k and V_λ Immunoglobulin light chains (47); H2B histone from a variety of species (50). For details, see Text.

Interestingly, similar upstream determinants have been observed in immunoglobulin genes (40,46,47). Conserved decanucleotide (dc) and pentadecanucleotide (pd) elements, shown to be essential for correct gene transcription, are located upstream of all sequenced human and mouse immunoglobulin K genes. Furthermore, a consensus sequence (CGTGATTTGC) spanning almost the entire dc element matches 9 out of 10 consecutive bases in the smaller class II conserved element. It therefore seems plausable to speculate that these elements are fulfilling some similar, as yet undetermined function. There is evidence for similar sequences in a number of different genes from various organisms (48). We have aligned some conserved upstream blocks of sequence from immunoglobulin and histone genes with the concensus sequences from class II genes on Figure 3. Also shown, are sequences upstream of the E. coli glutathione synthetase gsh-II gene, which exhibit remarkable similarity to the conserved blocks A and B (49).

The degree of intersequence homology outside the above mentioned elements, excluding alleles of the same gene is generally low. Comparison of the I-Eα and DRα genes, where one might expect considerable conservation, shows a 90 bp highly homologous (85%) region spanning regions A and B on Figure 3, a 75 bp region with some detectable homology encompassing the putative transcriptional initiation sites and a 12 bp region of exact homology directly preceding the methionine (initiation) codon. The degree of homology throughout the I-Aα and DRα signal sequences then remains high (83%) over an 82 bp section. A similar relationship exists between I-Aβ and DQβ, in that aproximately 60 bp of homologous DNA with 79% homology spans regions A and B, but further downstream little detectable homology exists through the 5' untranslated leader sequence and homology between signal sequences is lower (64%) than that observed in the case of I-Eα and DRα genes. Upstream of blocks A and B no detectable sequence conservation is observed. Throughout all the other sequences little homology exists outside the A and B units, with the exception of a possible CAT consensus sequence CCAATCC which lies approximately 18 bp 3' to B in all the β chains. A similar but less easily recognisable sequence is present in the DP and DZ α genes in a similar position, but in neither I-E nor DR α genes. There appears to be no strong requirement for highly conserved "TATA" like sequences; and where marked, possible candidates lie uncharacteristically close to the 5' end of the genes (Fig. 3). From these data a concensus for the 5' region of class II genes was derived (Fig. 3).

The existence of regulatory factors 5' to those described is not

excluded. Examination of the DPα1 and β1 intergenic regions reveals two approximately 70% AT rich 400-500 bp regions respectively 400 and 800 bp upstream of these genes. Such regions would offer situations of reduced interstrand basepairing, possibly facilitating strand separation at these points. Several small repeats and inverted repeats are also present in this region, however the significance of any of these sequences is uncertain.

Comparison of the DPβ1 gene to the other human class II β gene that has been sequenced, DQβ, is consistent with the derivation of the genes by duplication of a common precursor sequence followed by rapid sequence diversification in the introns. Some regions of the introns do show a residual level of homology, bp 13209-13240 in Fig. 2, for example, with the analogous intron in the DQβ gene (11), but there are few strikingly conserved regions that might indicate sequences of functional importance in the introns. Nevertheless, the length of the DPβ1 gene is remarkable in comparison to class II α genes. In particular, the introns flanking the β1 domain are over 4 kb in length.

The separation of the DPβ1 and α1 genes by only 2 kb at their 5' ends is a feature which provides an ideal opportunity to study the influence of the promoter regions of both genes upon their regulation. The insertion of these sequences into appropriate expression vectors should enable us to identify which, if any, of the conserved sequences are important for transcription of the genes under appropriate conditions.

Finally, the proximity of the promoter regions of the two genes suggests that they may be controlled co-ordinately by an enhancer, or other sequence, in or near the intergenic region. This possibility is under investigation.

ACKNOWLEDGEMENTS

We should like to thank Dr. W.F. Bodmer for continual advice and encouragement, as well as our colleagues in the laboratory: P. Austin, S. Carson, H. Meunier, J. Young and A. So. Frank Grosveld kindly allowed us to screen his genomic libraries, from which clone LC11 was derived. Dr. D Kappes gave us unpublished information on the signal sequence of the DPβ1 gene and Dr. H. Ehrlich similar information on DPα1.

REFERENCES

1. Snary D., Barnstable C., Bodmer W.F., Goodfellow P. and Crumpton M.J. (1976) Cold Spring Harbour Symp. Quant. Biol., 41, 379-386.
2. Springer T.A., Kaufman J.F., Terhorst C. and Strominger J.L. (1977) Nature, 268, 213-218.
3. Shackelford D.A., Kaufman J.F., Korman A.J. and Strominger J.L. (1982) Immunol. Rev., 66, 133-187.

4. Larhammar D., Andersson G., Andersson M., Bill P., Böhme J., Claesson L., Denaro M., Emmoth E., Gustafsson K., Hammarling U., Heldin E., Hyldig-Nielsen J.J., Lind P., Schenning L., Servenius B., Widmark E., Rask L. and Peterson P.A. (1983) Hum. Immunol., 8, 95-103.
5. Kaufman J.F. and Strominger J.L. (1982) Nature, 297, 694-697.
6. Chang H.C., Moriuchi T. and Silver J. (1983) Nature, 305, 813-815.
7. Schenning L., Larhammar D., Bill P., Wiman K., Jonsson A., Rask L. and Peterson P.A. (1984) EMBO J., 3, 447-452.
8. Auffray C., Ben-Nun A., Roux-Dosseto M., Germain R.N., Seidman J.G. and Strominger J.L. (1983) EMBO J., 2, 121-124.
9. Trowsdale J., Lee J., Carey J., Grosveld F., Bodmer J. and Bodmer W.F. (1983) Proc. Natl. Acad. Sci., USA, 80, 1972-1976.
10. Böhme J., Owerbach D., Denaro M., Lernmark A., Peterson P.A. and Rask L. (1983) Nature, 301, 82-84.
11. Larhammar D., Hyldig-Nielson J.J., Servenius B., Andersson G., Rask L. and Peterson P.A. (1983) Proc. Natl. Acad. Sci., USA, 80, 7313-7317.
12. Kratzin H., Yang C., Götz H., Pauly E., Kölbel S., Egert G., Thinnes F., Wernet P., Altevogt P. and Hilschmann N. (1981) Hoppe-Seyler's Z. Physiol. Chem., 362, S, 1665-1669.
13. Spielman R.S., Lee J.S., Bodmer W.F., Bodmer J.G. and Trowsdale J. (1984) Proc. Natl. Acad. Sci., USA, 81, 3461-3465.
14. Wake C.T., Long E.O. and Mach B. (1982) Nature, 300, 372-374.
15. Bodmer J. and Bodmer W.F. (1984) Immunology Today, 5, 251-254.
16. Shaw S., Kavathas P., Pollack M.S., Charmot D. and Mawas C. (1981) Nature, 293, 745-747.
17. Trowsdale J., Kelly A., Lee J., Carson S., Austin P. and Travers P. (1984) Cell, 38, 241-249.
18. Kappes D.J., Arnol D., Okada K. and Strominger J.L. (1984) EMBO J., 3, 2985-2993.
19. Austin P., Trowsdale J., Rudd C., Bodmer W.F., Feldman M. and Lamb J. (1984) Nature, in press.
20. Lee J.S., Trowsdale J. and Bodmer W.F. (1982) Proc. Natl. Acad. Sci., USA, 79, 545-549.
21. Vieira J. and Messing J. (1982) Gene, 19, 259-268.
22. Sanger F., Coulson A.R., Barrell B.G., Smith A.J.H. and Roe B.A. (1980) J. Mol. Biol., 143, 161-178.
23. Bankier A.T. and Barrell B.G. (1983) In, Techniques in Life Sciences, Nucleic Acid Biochemistry, p 1-34, Elsevier Scientific Publishers, Ireland.
24. Staden R. (1982) Nucl. Acids Res., 10, 4731-4751.
25. Roux-Dosseto M., Auffray C., Lillie J., Boss J., Cohen D., DeMars R., Mawas C., Seidman J. and Strominger J. (1983) Proc. Natl. Acad. Sci., USA, 80, 6036-6040.
26. Gustafsson K., Emmoth E., Widmark E., Bohme J., Peterson P.A. and Rask L. (1984) Nature, 309, 76-78.
27. Boss J.M. and Strominger J.L. (1984) Proc. Natl. Acad. Sci., USA, 81, 5199-5203.
28. Frayne E.G., Leys, E.J., Crouse G.F., Hook A.G. and Kellems R.E. (1984) Mol. Cell Biol., 4, 2921-2924.
29. Hurley C.K., Shaw S., Nadler L., Schlossman S. and Capra J.D. (1982) Exp. Med., 156, 1557-1562.
30. Pawelec G., Shaw S. and Wernel P. (1982) Immunogenetics, 15, 187-198.
31. DeMars R., Chang C.C., Shaw S., Reitnauer P.J. and Sondel P.M. (1984) Human Immunol., 11, 77-97.
32. Wilbur W.J. and Lipman D.J. (1983) Proc. Natl. Acad. Sci., USA, 80, 726-730.

33. Sawada I., Beal M.P., Shen, Chapman B., Wilson A.C. and Schmid C. (1983) Nucl. Acids Res., 11, 8087-8101.
34. Rogers J. (1984) Int. Rev. Cytol. Suppl., 17, in press.
35. Wilson R. and Storb U. (1983) Nucl. Acids Res., 11, 1803-1817.
36. Ullrich A., Gray A., Goeddel D.V. and Dull T.J. (1982) J. Mol. Biol., 156, 467-486.
37. Mason A.J., Evans B.A., Cox D.R., Shine J. and Richards R.I. Nature, 303, 300-307.
38. Singer D.S., Lifshitz R., Abelson L., Nyirjesy P. and Rudikoff S. (1983) Mol. Cell. Biol., 3, 903-913.
39. Kress M., Barra Y., Seidman J.G., Khoury G. and Jay G. (1984) Science, 226, 974-977.
40. Saito H., Matri R.A., Clayton L.K. and Tonegawa S. (1983) Proc. Natl. Acad. Sci., USA, 80, 5520-5524.
41. Mathis D., Benoist C., Williams V., Kanter M. and McDevitt H. (1983) Cell, 32, 745-754.
42. Gillies S.D., Folsom V. and Tonegawa S. (1984) Nature, 310, 594-597.
43. Malissen M., Hunkapiller T. and Hood L. (1983) Science, 221, 750-754.
44. Das H.K., Lawrance S.K. and Weissman S.M. (1983) Proc. Natl. Acad. Sci., USA, 80, 3543-3547.
45. Hyldig-Nielsen J.J., Schenning L., Hammerling U., Widmark E., Heldin E., Lind P., Servenius B., Lund T., Flavell R., Lee J.S., Trowsdale J., Scheier P.H., Zablitsky F., Larhammar D., Peterson P.A. and Rask L. (1983) Nucl. Acids Res., 11, 5055-5071.
46. Parslow T.G., Blair D.L., Murphy W.J., and Granner D.K. (1984) Proc. Natl. Acad. Sci., USA, 81, 2650-2654.
47. Falkner F.G. and Zachau H.G. (1984) Nature, 310, 71-74.
48. Grosschedl R. and Birnstiel M. (1980) Proc. Natl. Acad. Sci., USA, 77, 7102-7106.
49. Gushima, H., Yasuda, S., Soeda, E., Yokota, M., Kondo, M. and Kimura, A. (1984) Nucl. Acids Res., 12, 9299-9307.
50. Harvey, R.P., Robins, A.J. and Wells, J.R.E. (1982) Nucl. Acids Res., 10, 7851-7863.