DNA SEQUENCE ANALYSIS OF THE REPEAT AND ADJOINING UNIQUE REGION OF THE LONG SEGMENT OF HERPES SIMPLEX VIRUS TYPE 1

by

Lise Jane Perry

A Thesis Presented for the Degree of
Doctor of Philosophy

in

The Faculty of Science
at the University of Glasgow

Institute of Virology
Church Street
Glasgow
G11 5JR

June 1986

TABLE OF CONTENTS

MATERIALS AND METHODS

MATERIALS

METHODS

RESULTS

DISCUSSION

ACKNOWLEDGEMENTS

I am grateful to Professor J.H. Subak-Sharpe for providing me with the opportunity of working at the Institute of Virology, and for his interest throughout the course of this study. I thank Dr. Duncan McGeoch for his supervision and continued interest for the duration of the work, and his help in the preparation of this thesis.

I thank Drs. Duncan McGeoch, Frazer Rixon, Andrew Davison, Margaret Frame and Roger Everett for helpful discussion and for access to their data prior to publication. I am grateful to Prof. G. Darai for unpublished data on the deletion in the HSV-1 strain HFEM. I thank Drs. Duncan McGeoch, Andrew Davison and Philip Taylor for assistance with the computing. I wish to thank Drs. Frazer Rixon, Val Preston and Chris Preston for plasmids. I also thank those I have worked with for their time and help.

My thanks go to Fiona Conway for printing this thesis. I thank my family and friends for their support and encouragement.

During the course of this study, the author was supported by a grant from The Shell Trust for Higher Education. Unless otherwise stated, all results reported in this thesis were by the author's own work.

SUMMARY

The DNA sequences of the BamHI b fragment and of a SmaI-BamHI subfragment of BamHI e from the herpes simplex virus type 1 (HSV-1) genome have been determined. These restriction fragments are located at each end of the U_L sequence and span the R_L/U_L junctions. The BamHI b fragment contains over 6 kb of the R_L element. The DNA sequence of the two restriction fragments was determined by chain terminator sequencing reactions of recombinant M13 clones generated from sonicated DNA or restriction enzyme digested DNA.

The precise locations of the junctions between the R_L and U_L sequences have been determined. This was achieved by aligning the homologous R_L sequences of both fragments and noting where they diverged. The R_L/U_L junction was found to be located adjacent to a set of tandem reiterations, which are contained in the R_L sequences. The copy number of this reiteration set was found to vary in the BamHI b sequence.

The natures of the DNA sequences of the two fragments were investigated. The R_L sequence was found to contain several tandem reiteration sets and numerous minor repeated elements. R_L was further characterised by the high G+C content of the DNA sequence. The base composition of the R_L sequence determined is 70.6% G+C. The U_L sequences investigated contained no reiteration sets, and had lower G+C contents, of 65.6% and 61.8% for the BamHI b and BamHI e fragments respectively.

Two genes had previously been mapped to the BamHI b fragment. These are the immediate early genes, IE gene 1 and IE gene 2. IE gene 1, encoding IE110, is entirely contained within the R_L element. The gene is

3.6 kb in length, and extends beyond BamHI b. IE gene 1 contains two introns of 764 bp and 135 bp in length. The introns are composed of repetitive DNA sequences. All three of the exons are believed to be polypeptide coding. The 775 amino acid sequence of IE110, predicted from the DNA sequence, has a M_r 78,452. IE110 has a cysteine rich region which may be involved in functional binding to DNA.

IE gene 2, encoding IE63, has been previously mapped to the U_L component of the BamHI b fragment, and had been shown to contain no introns. The predicted amino acid sequence of IE63 is 512 residues in length, and has a M_r 55,376.

The DNA sequences determined have been analysed in a search for further genes. This analysis indicated clearly that two genes, encoding proteins of M_r 20,491 and 21,182, are located in the U_L sequence of BamHI b, downstream of IE gene 2. Three genes are believed to be contained in the U_L sequence of BamHI e. These genes have been named UL1, UL2 and UL3, and encode proteins with M_r 24,932, 27,327, and 24,446. None of these predicted genes are thought to contain introns. No function has been assigned to any of the predicted proteins.

Examination of the 3 kb sequence in R_L between the R_L/U_L junction and the 3' end of IE gene 1 suggests that it may not be polypeptide coding. The open reading frames in this region have been analysed, but do not resemble known HSV-1 polypeptide coding sequences.

The arrangement of the genes in the R_L and U_L sequences of the BamHI b and e restriction fragments of HSV-1 has been examined. The genes in the U_L sequences determined show a compact arrangement. None of the genes are thought to contain introns, and the distances

between adjacent genes is generally short. Two of the genes, UL1 and UL2, are believed to be transcribed into 3' coterminal mRNAs. The organisation of the genes adjacent to the R_L/U_L junctions was studied. Each junction is located very close to the promoter sequences of a gene transcribed in a direction away from R_L.

Both IE gene 1 and IE gene 2 have homologues in the genome of the alphaherpesvirus, varicella-zoster virus (VZV). Four of the predicted genes also have VZV homologues. The relative arrangement of the genes encoding the homologous proteins was found to be similar in the two viruses. However, comparisons between the arrangement of the genes relative to the R_L elements and the R_L/U_L junctions in HSV-1 and VZV indicated a change in sequence organisation. These differences have implications for the evolution of the DNA sequence in the regions characterised.

Two of the proteins, IE63 and the M_r 27,327 protein encoded by UL2, also have homologues in the genome of the gammaherpesvirus, Epstein-Barr virus (EBV). An extensive search for EBV homologues was carried out. However, no detectable level of homology was seen between any EBV polypeptide and IE110 or any of the remaining proteins predicted from the DNA sequence. The homologous proteins show considerable amino acid sequence conservation. The level of homology between the proteins was comparable with that observed for these proteins in VZV. In addition, the segments of the amino acid sequences showing greatest conservation was the same for both VZV and EBV.

The organisation of the two homologous genes in EBV differs greatly from the arrangement of the corresponding genes in HSV-1. As no EBV genes neighbouring the conserved genes have HSV-1 homologues, it is not possible to reconstruct in any detail the

evolutionary events which have resulted in this extensive relocation of genetic material.

NON-STANDARD ABBREVIATIONS

A	deoxyadenylic acid residue in DNA
BCIG	5-chloro-4-bromo-3-indolyl-β-D-galactoside
BSA	bovine serum albumin
bp	base pair(s)
C	deoxycytidylic acid residue in DNA
C terminus	carboxy terminus of a protein
CAT	chloramphenicol acetlytransferase
CCV	Channel Catfish Virus
dATP	2'deoxyadenosine 5'-triphosphate
dCTP	2'deoxycytidine 5'-triphosphate
dGTP	2'deoxyguanosine 5'-triphosphate
dTTP	2'deoxythymidine 5'-triphosphate
dITP	2'deoxyinosine 5'-triphosphate
ddATP	2',3'-dideoxyadenosine 5'-triphosphate
ddCTP	2',3'-dideoxycytidine 5'-triphosphate
ddGTP	2',3'-dideoxyguanosine 5'-triphosphate
ddTTP	2',3'-dideoxythymidine 5'-triphosphate
DNase	deoxyribonuclease
DTT	dithiothreitol
EBV	Epstein-Barr virus
EDTA	ethylene diaminetetra acetic acid
G	deoxyguanylic acid residue in DNA
HCMV	human cytomegalovirus
HSV-1	herpes simplex virus type 1
HSV-2	herpes simplex virus type 2
HVS	herpesvirus saimiri
IPTG	isopropyl-β-D-thiogalactoside
kb	kilobase(s)
MDV	Marek's disease virus
M_r	molecular weight
N terminus	amino terminus of a protein
ORF	open reading frame
PEG	polyethylene glycol
pi	post infection

PRV	psuedorabies virus
R	purine nucleotide
SDS	sodium dodecylsulphate
T	deoxythymidylic acid residue in DNA
TK	thymidine kinase
TEMED	N,N,N',N''tetramethyl ethylenediamine
tris	tris(hydroxymethyl)aminomethane
ts	temperature sensitive
UV	ultraviolet radiation
V_{mw}	viral protein
VZV	varicella-zoster virus
Y	pyrimidine nucleotide

Amino acid symbols

amino acid	three letter symbol	one letter symbol
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V
any amino acid	-	X

INTRODUCTION

INTRODUCTION

1.1 OBJECTIVES

The aim of the work described in this thesis was to investigate the nature of the long repeat region and adjacent long unique sequences of the genome of herpes simplex virus type 1 (HSV-1). With this objective, the DNA sequences of two regions of the HSV-1 genome have been determined. The data have been interpreted in relation to the sequence organisation of the genome. The arrangement of the genes in the sequence has been deduced, and the nature of their encoded proteins examined. The evolutionary implications of the results, in relation to other herpesviruses, have been explored.

This introduction aims to provide a general background on the herpesviruses and on the biology of herpes simplex virus. A description of topics more directly related to the experimental work then follows. These include the structure and organisation of HSV-1 genes, the regulation of HSV-1 gene expression, and the relationships between HSV-1 and other herpesviruses.

1.2 THE HERPESVIRIDAE AND THEIR CLASSIFICATION

Herpes simplex virus (HSV) is a member of the family Herpesviridae (Matthews, 1982). Members of the Herpesviridae characteristically have a large virus particle, 120-200 nm in diameter, containing a single, linear double-stranded DNA molecule with a M_r ranging between 80 and 150 X 10^6 (Honess and Watson, 1977). The genomes of herpesviruses contain regions of unique

and repeat (or reiterated) nucleotide sequences. The genomes of several herpesviruses exist in different isomeric forms, with respect to the orientations of the unique sequences (Honess and Watson, 1977). Figure 1.1 shows the structure of the six genome arrangements known, with arrows showing the relative orientations of the unique sequences in the various isomers.

The herpesviruses have a wide host range, infecting fish (for example, Channel catfish virus, CCV), birds (for example, Marek's disease virus of chickens, MDV), and mammals. Humans are known to be the natural host of five herpesviruses: herpes simplex virus types 1 and 2, varicella-zoster virus (VZV), Epstein-Barr virus (EBV), and human cytomegalovirus (HCMV) (Honess and Watson, 1977).

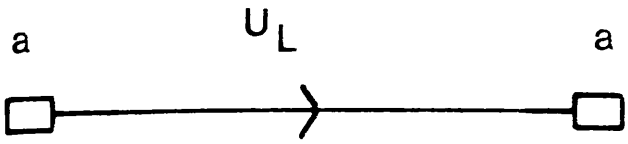
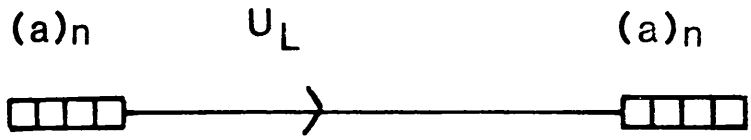
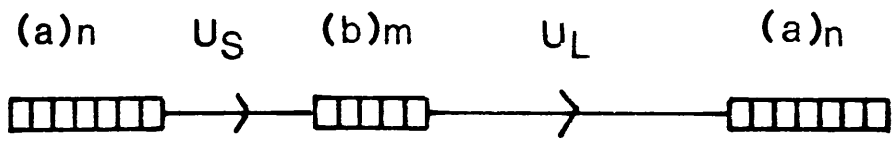
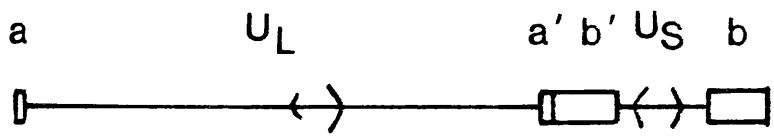
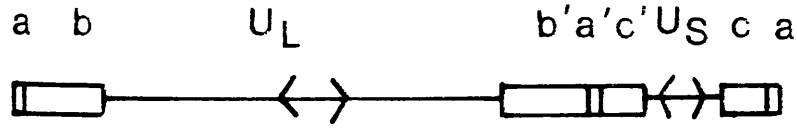
Herpesviruses of vertebrates differ in the cell types infected and in the pattern of latency in the host. This has been used as a basis for classification (Matthews, 1982), as described below.

1.2.1 Alphaherpesvirinae

Members of this sub-family have a short replicative cycle, less than 24 h, which can be observed as a rapid spread of infection in cell culture. Primary infection generally gives rise to skin or respiratory tract infections. The viruses often persist subclinically as latent infections of the central nervous system and may reactivate to give periodic recurrences of disease. Members of the Alphaherpesvirinae include herpes simplex virus types 1 and 2 and varicella-zoster virus.

Figure 1.1 Genomic structure of the herpesviruses

The structures of herpesvirus genomes are illustrated. Repeat sequences are represented as open boxes. U_L and U_S indicate long and short unique sequences, a, b, etc., indicate repeat sequences, with b', a', their complement. Arrows indicate the relative orientations of unique sequences. The number of isomeric forms of each structure is given. CCV, Channel catfish virus; HVS, Herpesvirus saimiri; EBV, Epstein-Barr virus; PRV, Pseudorabies virus; VZV, varicella-zoster virus; HSV-1, Herpes simplex virus type 1.

	Genome Structure	Number of Isomers
;CV		1
IVS		1
EBV		1
RV		2
/ZV		2 major 2 minor
HSV-1		4

1.2.2 Betaherpesvirinae

Members of this sub-group have a narrow host range. Infection is usually subclinical, although it can result in serious disease in neonates or immunocompromised hosts. They have a long replicative cycle, greater than 24 h, and replicate best in fibroblasts. The virus can become latent, although the site of persistence is presently unclear. Human cytomegalovirus is a member of the Betaherpesvirinae.

1.2.3 Gammaherpesvirinae

Members of this group have a narrow host range. They all replicate in lymphoblastoid cells, although the length of the replicative cycle varies between viruses. The Gammaherpesvirinae frequently give rise to latent infections, giving transformed lymphoblastoid cells. EBV is a member of the Gammaherpesvirinae.

1.3 PATHOLOGY AND BIOLOGY OF HERPES SIMPLEX VIRUS

Generally, HSV-1 infections give rise to oral and facial lesions, whereas infection with HSV-2 gives rise to genital lesions. However, a proportion of genital HSV infections are caused by HSV-1, and HSV-2 has likewise been associated with non-genital disease (Whitley, 1985). Individuals are usually subjected to primary infection by HSV-1 as children, and most adults have antibodies against HSV-1. There is a great variability in symptoms of primary infection, ranging from being totally asymptomatic to a combination of fever, sore throat, ulcerative and vesicular lesions

and gingivostomatitis. Following primary infection, HSV can remain latent in nervous tissue. Reactivation of latent virus involves transmission from the ganglion sites through the sensory nervous tissue to the skin. Recurrent infection is usually characterised by lip lesions (herpes labialis), as reviewed by Whitley (1985).

Although less common, there are a number of serious diseases also attributable to herpes simplex virus. These include herpes keratoconjunctivitis, eczema herpeticum and herpes simplex encephalitis (Whitley, 1985). In addition, serious disease can arise in the immunocompromised host and through infection of the newborn (Rawls, 1973). Further, genital HSV-2 infection has been associated with cervical carcinoma. This followed the observation that women with genital HSV infection had an increased incidence of cervical carcinoma (Whitley, 1985). Both HSV-1 and HSV-2 can induce malignant transformation of cells *in vitro*, but only after the virus has been inactivated to prevent productive infection and cell death. This has increased speculation that HSV-2 may provide at least a contributory factor in cervical cancer (Spear and Roizman, 1980).

1.4 STRUCTURE AND REPLICATION OF THE VIRION

The virion consists of four structural components, namely, the core, the capsid, the tegument and the envelope (Spear and Roizman, 1980). The core of the virion contains the DNA in association with a proteinaceous structure (Furlong *et al.*, 1972). Surrounding the core is the capsid, which consists of 162 capsomeres arranged in an icosahedron,

approximately 100 nm in diameter (Spear and Roizman, 1980). The tegument, a variable layer of amorphous material, surrounds the capsid (Roizman and Furlong, 1974). These structures are enclosed within a lipid bilayer, the envelope. The envelope originates from the nuclear membrane and contains a number of virally encoded structural glycoproteins (Roizman and Furlong, 1974).

The virus gains entry to the host cell by adsorption of envelope glycoproteins to receptors on the plasma membrane of the host cell, followed by fusion of the envelope with the membrane (Spear, 1985). The capsid is released into the cytoplasm of the cell (Spear and Roizman, 1980) and the DNA in a protein complex is translocated to the nucleus. Inside the nucleus, the DNA is transcribed into mRNA, which is translated in the cytoplasm (Wagner, 1985). The DNA is replicated in the nucleus as described below, and is incorporated into newly formed nucleocapsids (Stow, 1985). The nucleocapsids bud through the inner nuclear membrane, thereby aquiring the outer envelope (Spear and Roizman, 1980). The virus particles are released by transport to the surface of the cell through the endoplasmic reticulum (Spear, 1985).

Following primary infection, HSV will often establish a latent infection (Wildy *et al.*, 1982). HSV DNA persists in the sensory ganglia, and can be reactivated from explanted ganglia cultured *in vitro* (Al-Saadi *et al.*, 1983; Tullo *et al.*, 1983). Work with *ts* mutants has shown that HSV genes are required to induce latency (Lofgren *et al.*, 1977). It has been claimed that viral transcripts and proteins, including IE175, can be detected in latently infected neurones (Green *et al.*, 1981). The physical state of the HSV

DNA is uncertain, although it is believed to exist in "endless" structures, such as circles or concatemers (Rock and Frazer, 1983).

1.5 GENOME STRUCTURE OF HSV

The genome of HSV is a linear, double-stranded DNA molecule of approximately 98 X 10^6 molecular weight, as determined by the summation of the molecular weights of restriction endonuclease fragments of genomic DNA (Clements et al., 1976; Cortini and Wilkie, 1978). HSV-1 DNA has an average base composition of 67% G+C, and HSV-2 DNA, 69% G+C (Kieff et al., 1971). HSV DNA is unmethylated (Low et al., 1969). There appear to be single strand gaps in the DNA, revealed by alkaline denaturation and by digestion of HSV DNA from internal sites by lambda exonuclease (Kieff et al., 1971; Wilkie 1973; Wadsworth et al., 1976). The HSV genome is believed to have a single unpaired nucleotide at each 3' terminus (Mocarski and Roizman, 1982).

HSV DNA can be regarded as consisting of two segments, designated long (L, 121 kbp) and short (S, 26 kbp) (Delius and Clements, 1976; Roizman, 1979; McGeoch et al., 1985 and 1986a). Each segment comprises a unique region, the long unique (U_L) and short unique (U_S) sequences, and these are bounded by inverted repetitions, the terminal and internal long repeats (TR_L and IR_L) and the terminal and internal short repeats (TR_S and IR_S) (Sheldrick and Berthelot, 1974). The base composition differs between the repeat and unique sequences. The R_S sequence of HSV-1 is composed of 79.5% G+C DNA, and U_S 64.3% G+C (McGeoch et al., 1985 and 1986a).
The inverted repeat sequences are considered to contain

two sequence elements, as described below.

A terminal redundancy of approximately 400 bp (dependent on HSV strain), known as the a sequence, is present as direct repeats at both termini and in inverted orientation at the joint between the L and S segments (Davison and Wilkie, 1981). The a sequence is present as a single copy at the S terminus and in variable copy numbers at the L-S joint and at the L terminus (Wagner and Summers, 1978; Locker and Frenkel, 1979). The sequence in TR_L and IR_L excluding the a sequence, is termed the b sequence. Likewise, the TR_S and IR_S elements, are composed of an a and a c sequence (Roizman, 1979). The structure of the HSV genome is illustrated in figure 1.1.

In addition to these major repeat elements, the HSV genome contains a number of short tandemly reiterated DNA sequences (reviewed by Rixon et al., 1984). Examining eight tandem reiteration sets, Rixon et al. (1984) noted that although the individual sequences differ, they all have high G+C contents and show a marked asymmetry in the distribution of purines and pyrimidines on the two DNA strands. The reiterated sequences studied are generally short (between 5 and 35 bp) and their copy number can vary between individual virus genomes (Rixon et al., 1984). The majority of the reiterations presently known are located in the major inverted repeats, R_L and R_S (Davison and Wilkie, 1981; McGeoch et al., 1986a; work described in this thesis). However, two reiterated sequences lie within U_S (McGeoch et al., 1985). At least three tandem reiteration sets are believed to occur within polypeptide coding regions. These include both the sets in U_S, which are located in genes US7 and US11 (Rixon and McGeoch, 1984; McGeoch et al., 1985).

The third reiteration set is located within a gene in R_L near the a sequence (Chou and Roizman, 1986).

1.5.1 Isomerisation of the HSV genome

Sheldrick and Berthelot (1974) suggested that as the terminal regions are repeated internally, recombination between these sequences might occur readily. This would result in the inversion of the L or S component, thereby producing four isomeric forms. Restriction analysis of HSV DNA confirmed that isomeric forms do indeed exist (Hayward et al., 1975; Clements et al., 1976). The presence of four genome arrangements has also been demonstrated by partial denaturation mapping of heteroduplex DNA (Delius and Clements, 1976).

DNA from wild-type virus consists of four equimolar populations. These are termed P (Prototype), I_L (L inverted), I_S (S inverted) and I_{SL} (S and L inverted) (Roizman et al., 1979). The four isomers are illustrated in figure 1.2.

Studying intertypic recombinants of HSV-1 and HSV-2, Davison and Wilkie (1983a), demonstrated that segment inversion specifically depends upon the a sequence. Homology between a sequences alone at the L terminus and L-S joint was sufficient for inversion of U_L. Deletion of the L-S joint sequences prevents inversion, and insertion of additional copies of the a sequence promotes inversion (Poffenberger et al., 1983; Mocarski et al., 1980).

Figure 1.2 The structure and four isomeric forms of the HSV-1 genome

U_L and U_S indicate the long and short unique sequences; TR_L, IR_L, IR_S and TR_S, the major repeat sequences. Arrows indicate the relative orientation of the unique sequences. The four isomeric forms of HSV-1 are labelled as follows.

P = Prototype arrangement
I_L = Inversion of the long region
I_S = Inversion of the short region
I_{SL} = Inversion of both the long and short regions

1.5.2 Replication of HSV DNA

Replication of the HSV genome occurs in the nucleus of the infected cell. The precise mode of replication has not been determined. The genome appears to circularise, perhaps as a result of direct ligation of the termini (Poffenberger and Roizman, 1985), possibly involving the complementary unpaired nucleotides present at both 3' termini of the genomic DNA (Mocarski and Roizman, 1982). Replication is initiated at one or more origins of replication and continues in a rolling circle, to yield head to tail concatemers. Late in infection newly synthesised HSV DNA appears as tandem repeats of the viral genome (Jacob *et al*., 1979; Stow *et al*., 1983; Stow, 1985). Cleavage of unit lengths of virus DNA is believed to occur by site specific events and may be directed by signals within the *a* sequence (Davison and Wilkie, 1981; Varmuza *et al*., 1985). Cleavage and encapsidation of the DNA is believed to be tightly coupled (Deiss and Frenkel, 1986). The cis-acting signals necessary for cleavage and packaging of the HSV genome are all believed to be contained within the *a* sequence (Stow *et al*., 1986).

The HSV genome contains three origins of replication. Two lie in identical sequences in IR_S and TR_S and the third is near the centre of U_L (Friedmann *et al*., 1977; Kaerner *et al*., 1979; Stow, 1982). The origin of replication in U_L (Ori_L) contains a 144 bp perfect palindromic sequence (Quinn and McGeoch, 1985; Weller *et al*., 1985). Ori_S contains a sequence similar to Ori_L and has a near perfect palindromic sequence 45 bp in length (Murchie and McGeoch, 1982). Of the 144 bp palindromic sequence in Ori_L, 90 bp are conserved in Ori_S. Greatest homology occurs at and near to the

centre of the palindrome (Weller et al., 1985).

1.6 EXPRESSION, STRUCTURE AND ARRANGEMENT OF HSV GENES

HSV DNA is transcribed in the nucleus of the infected cell by host cell RNA polymerase II (Costanzo et al., 1977; Lowe, 1978). As with host mRNA, HSV mRNA is polyadenylated at the 3' end (Bachenheimer and Roizman, 1972), 5' capped, and is internally methylated (Moss et al., 1977).

HSV promoters have features similar to those of cell genomes and will function when introduced by transfection into cells (McKnight and Gavis, 1980). A number of HSV genes studied have recognisable TATA boxes in their 5' upstream sequences, positioned 25 to 30 nucleotides upstream of the transcriptional start site (Corden et al., 1980; McGeoch et al., 1985). Further transcriptional control sequences upstream of the cap site operate in HSV genes, and will be discussed later. Non-translated leader sequences range from about 15 to over 300 bp in length (Murchie and McGeoch, 1982; McGeoch et al., 1985).

HSV mRNA differs from cellular mRNA in its low level of splicing. Apart from work which will be described in this thesis, the only genes which are known to be spliced are IE gene 4 and IE gene 5, and a late gene (Rixon and Clements, 1982; Costa et al., 1985). The splice donor and acceptor sites of the introns studied all comply with recognised consensus sequences (Mount, 1982). The introns of IE genes 4 and 5 share identical sequences in the IR_S and TR_S regions of the S segment, respectively. The introns are short and contain tandemly reiterated non-coding sequences

(Murchie and McGeoch, 1982). The late gene described by Costa et al. (1985) has a 4 kb intron which contains a nested set of three 3' coterminal genes, located on the opposite DNA strand.

The sequence AATAAA or its variant ATTAAA have been found near the 3' termini of all HSV mRNAs mapped (for example, McGeoch et al., 1985). As with cellular genes, the sequence is associated with termination of transcription and polyadenylation of the 3' end of the mRNA (Proudfoot and Brownlee, 1976; Wickens and Stephenson, 1984).

A sequence associated with efficient termination of mRNA transcription has been identified (McLauchlan et al., 1985). This sequence element with the consensus YGTGTTYY usually lies approximately 30 bp downstream of the polyadenylation site. The sequence has been found to be present in many HSV genes (McLauchlan et al., 1985).

The S region of HSV-1 has been completely sequenced, and the mRNAs mapped, thereby enabling an analysis of gene organisation in HSV-1 (Rixon et al., 1982; McGeoch et al., 1985 and 1986a; Rixon and McGeoch, 1985). The arrangement of genes in U_S is compact, utilizing both strands of the DNA, with 94% of the sequence transcribed and 74% of the sequence believed to be polypeptide coding. The sequence contains the polypeptide coding regions of twelve genes. All of the genes entirely contained in U_S have been shown to be unspliced.

Of the twelve genes, ten are organised into four 3' coterminal families. This results in a wide range in length of the 3' untranslated sequence, from

approximately 50 nucleotides to over 3 kb. Although this appears to be a common feature in HSV, adjacent genes on the same DNA strand do not always possess 3' coterminal mRNAs. Each of the genes in U_S has a distinct promoter. Several may lie in an adjacent gene, within either untranslated or polypeptide coding sequences. mRNAs on opposite strands also overlap. In addition, the polypeptide coding regions of two genes in U_S overlap (Rixon and McGeoch, 1984).

Both strands of HSV DNA encode genes. Examples of head to head and tail to tail genes occur in U_S (Rixon and McGeoch, 1985). The genes for the major DNA binding protein and the DNA polymerase have been mapped and sequenced (Quinn and McGeoch, 1985). These genes are located on either side of an origin of replication (Ori_L) and are transcribed divergently.

1.7 PROTEINS OF HSV

HSV codes for an estimated 80 polypeptides (Honess and Watson, 1977; McGeoch and Davison, 1986). A large number of HSV induced proteins can be detected experimentally. For example, Haar and Marsden (1981) detected approximately 230 virus induced polypeptides by two dimensional gel analysis. However, these certainly included many processed species, and the presence of multiple forms obscures experimental determination of numbers of separate proteins. The functions of a number of HSV-1 proteins were first determined using *ts* or drug resistant mutants. Physical mapping of *ts* mutants by marker rescue has been useful in determining their genomic location (for example, Stow *et al.*, 1978). The functions of a number of HSV-1 proteins have been characterised, but for the majority of proteins the function is unknown. Proteins encoded by the HSV genome can

be classified as structural proteins of the virion, proteins involved in DNA replication, and proteins responsible for the regulation of gene expression.

1.7.1 Structural proteins

The virion contains 15-33 polypeptides (Spear and Roizman 1980). These include the proteins of the nucleocapsid, the tegument and the glycosylated proteins associated with the virion envelope.

The capsid is composed of at least six polypeptides. The major capsid protein is a large constituent of the capsid, with an estimated 850-1000 copies present in each virion (Spear and Roizman, 1980). In addition to determining the regular structure of the virion, the proteins of the nucleocapsid may play a role in the encapsidation of the viral DNA in the core.

The tegument is a variable layer, containing an undefined number of proteins. Most of the proteins have yet to be functionally characterised, but are believed to be important in the envelopment of the nucleocapsid through interactions with the proteins of the nucleocapsid and envelope (Spear and Roizman, 1980). A major constituent of the tegument is Vmw65, a regulatory protein (Campbell *et al.*, 1984) which will be discussed later.

The glycosylated proteins are the best characterised group of the structural proteins. HSV-1 is known to encode six glycoproteins, namely, gB, gC, gD, gE, gG and gH (Spear 1985; Frame *et al.*, 1986; Buckmaster *et al.*, 1984; McGeoch *et al.*, unpublished). No HSV-2 counterpart to gH has yet been characterised, but all other known HSV-1 glycoproteins have HSV-2 equivalents (Frame *et al.*, 1986; Spear, 1985).

Glycoproteins are important determinants of viral pathogenicity (Spear, 1985). The glycoproteins are intergrated in the virion envelope and form the spikes projecting from the particle. They are also present on the surface of infected cells. The glycoproteins are believed to be involved in the adsorption and penetration of the virion into the host cell, and to mediate cell to cell spread of infection. They may also play a role in budding and envelopment (Spear, 1985; Spear and Roizman, 1980). Antibodies directed against some of the glycoproteins have been shown to have neutralising activity (Powell et al., 1974).

1.7.2 HSV proteins involved in DNA metabolism and replication

HSV codes for a number of proteins involved in DNA replication. These include enzymes participating in the metabolism of DNA precursors, such as thymidine kinase (TK), ribonucleotide reductase and dUTPase (Kit and Dubbs, 1963; Dutia, 1983; Frame et al., 1985; Cohen, 1972; Wohlrab et al., 1982). The genes for TK and both subunits of the ribonucleotide reductase have been sequenced (McKnight, 1980; McLauchlan and Clements, 1983; Nikas, personal communication). HSV also encodes proteins which are directly involved in DNA synthesis. These proteins include the DNA polymerase, DNA binding protein, alkaline exonuclease and possibly topoisomerase (Keir and Gold, 1963; Keir et al., 1966; Littler et al., 1983; Morrison and Keir, 1968; Muller et al., 1985).

Since the activity of a virally encoded DNA polymerase was first detected (Keir and Gold, 1963), the protein has been extensively characterised. The gene for the DNA polymerase has been sequenced (Quinn and McGeoch, 1985). In

addition to DNA polymerisation, the enzyme has a 3' to 5' exonuclease activity, believed to serve as a proof reading function. The virally encoded alkaline exonuclease is also required for HSV DNA replication (Morrison and Keir, 1968; Moss, 1986). The enzyme is believed to function together with the DNA polymerase and major DNA binding protein (Littler *et al.*, 1983). The function of the major DNA binding protein has not been clarified although it is known that *in vitro* it can separate the strands of DNA of a double helix (Powell *et al.*, 1981). The gene for this protein has also been sequenced (Quinn and McGeoch, 1985). A topoisomerase activity is associated with the HSV virion (Muller *et al.*, 1985). The enzyme is believed to be a component of the envelope or tegument of the virion.

1.7.3 HSV regulatory proteins

The regulatory proteins of HSV control viral expression. A number of virally encoded proteins including the IE proteins and Vmw65 regulate expression of viral genes at the transcriptional level. The control of gene expression at the transcriptional level by the IE proteins and Vmw 65 will be discussed fully later.

A number of virally encoded proteins undergo post-translational modifications, including phosphorylation, glycosylation and sulphation. Modification may affect the protein by altering its activity or location in the infected cell. Similarly, modification of host proteins may affect their activity. An HSV-1 gene encoding a protein showing homology to the amino acid sequence of the protein kinase of other organisms has been sequenced (McGeoch and Davison, 1986). It is postulated that this protein may play a role in the modulation of cellular processes during lytic infection and possibly also in latent infection (McGeoch and

Davison, 1986).

1.8 CONTROL OF GENE EXPRESSION OF HSV

HSV gene expression is coordinately regulated and sequentially ordered in a cascade fashion. The first genes to be expressed are the five immediate early (IE or α) genes (Watson et al., 1979). The IE genes can be expressed without de novo protein synthesis in the newly-infected cell nucleus (Honess and Roizman, 1974; Harris-Hamilton and Bachenheimer, 1985). The five IE genes are located at or near the inverted repeats, as shown in figure 1.3 (Clements et al., 1979; Watson et al., 1979; Anderson et al., 1980; Watson et al., 1981).

Following translation ot the IE proteins, the second set of genes are transcribed (Wagner et al., 1972). The early (E or β) genes can be subdivided into two subgroups (β_1 and β_2) according to their time of expression (Mavromara-Nazos et al., 1986).

At late times of infection, with the onset of viral DNA replication, the late (L or γ) genes are transcribed (Wagner, 1972). The late genes can also be subdivided into two groups. In the absence of viral DNA replication, the leaky-late (γ_1) mRNAs become detectable in the cytoplasm. Transcription of the true late genes (γ_2) appears to have an absolute requirement for viral DNA replication (Holland et al., 1980).

The regulation of viral gene expression in the infected cell is considered to be primarily at the transcription level, through both cis-acting promoter sequences and trans-acting viral products. The factors involved in the control of each stage of transcription will

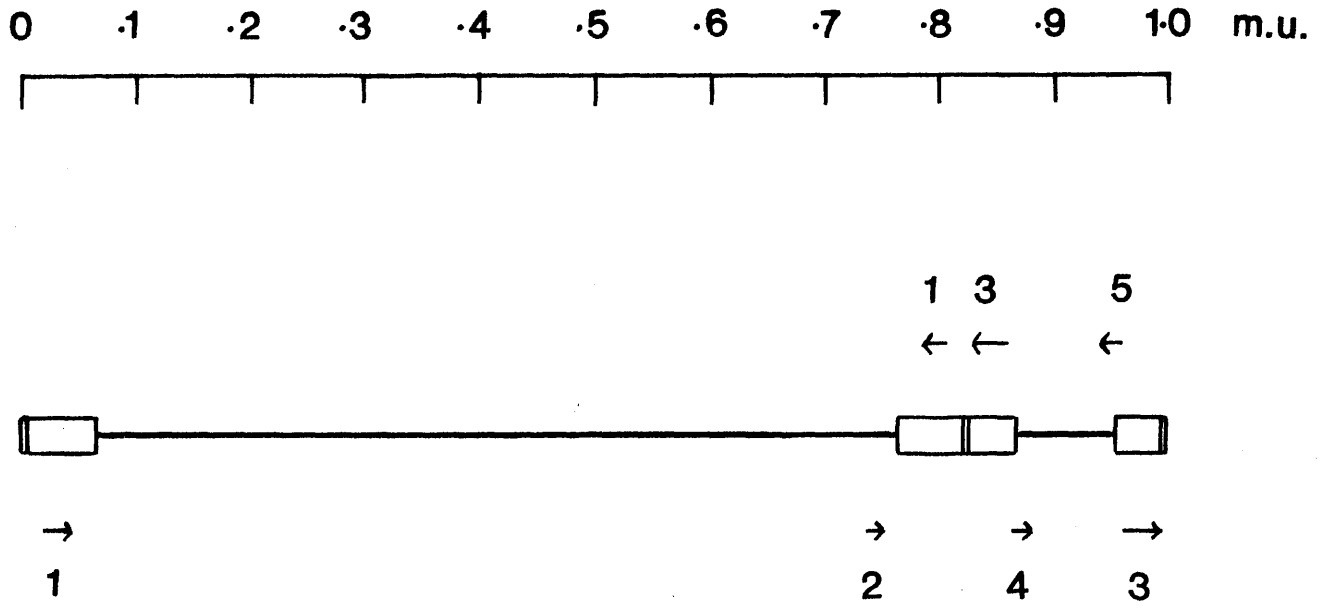

0 ·1 ·2 ·3 ·4 ·5 ·6 ·7 ·8 ·9 1·0 m.u.
1 3 5
1 2 4 3

Figure 1.3 Location and orientation of the IE genes

The locations and orientations of the five IE genes of HSV-1 are shown. The scale represents map units. (Clements et al., 1979; Watson et al., 1979; Anderson et al., 1980).

be considered below.

1.8.1 Regulation of IE gene expression

Even the IE genes, transcribed immediately on infection of the cell, are subject to modulation by the virion (Batterson and Roizman, 1983). The late polypeptide $V_{mw}65$, a major species of the virus tegument, will stimulate IE gene expression from basal levels (Campbell et al., 1984). The DNA sequence of the gene encoding $V_{mw}65$ has now been sequenced (Dalrymple et al., 1985). Stimulation of expression by $V_{mw}65$ is specific to IE gene promoters (O'Hare and Hayward, 1985b).

IE175, itself an IE protein encoded by IE gene 3, down-regulates IE gene transcription (Preston, 1979). Conversely, IE110, also an IE protein, appears to stimulate expression from IE promoters (O'Hare and Hayward, 1985b). The control of IE gene expression provided by IE175 and IE110 may maintain a balance required during normal infection (O'Hare and Hayward, 1985b).

$V_{mw}65$, a structural component of the virion, induces expression of the IE genes (Cordingley et al., 1983). The significance of sequences upstream of the mRNA cap site in IE gene transcription regulation has been recognised (Post et al., 1981; Mackem and Roizman, 1982a). Sequence data from the upstream regions of all five IE genes of HSV-1 have been analysed (Murchie and McGeoch, 1982; Mackem and Roizman, 1982a and b; Whitton et al., 1983). A sequence with the consensus 5' TAATGARATTC 3' (R = purine) is present in the upstream regions of all IE genes (Mackem and Roizman, 1982b; Preston et al., 1984). The highly conserved sequence is present in one or more copies. Further, the upstream regions have palindromic sequences (Mackem and Roizman,

1982b).

On comparing the regulatory region of the HSV-1 IE gene 2 with the HSV-2 equivalent, Whitton et al. (1983 and 1984) found that sequences believed to be involved in the induction of IE gene expression were highly conserved. The TAATGARATTC sequence element is conserved in the HSV-2 IE gene 2. In addition, palindromic sequences are present at similar positions in the upstream regions of both the HSV-1 and HSV-2 IE gene 2.

The presence of these sequence characteristics upstream of all HSV-1 IE genes and their conservation between the two HSV serotypes suggest they may be of importance in the coordinate induction and regulation of IE gene expression. The requirements for particular upstream sequences in stimulation of transcription by $V_{mw}65$ have been investigated (Cordingley et al., 1983; Preston et al., 1984; Campbell et al., 1984; Bzik and Preston, 1986). Two distinct regions have been identified.

For efficient transcription only 69 bp of the upstream sequence from the promoter region of IE genes 4 and 5, was required. This region includes the TATA sequence element (Corden et al., 1980). A second region, far upstream from the IE gene cap site, was demonstrated (Cordingley et al., 1983). This region was shown to be important for stimulation by $V_{mw}65$. This far upstream region contains the TAATGARATTC sequence element at positions -338 to -328 relative to the transcriptional start site . This sequence, crucial for stimulation of IE transcription by $V_{mw}65$, appears to be unique to the IE genes (Preston et al., 1984).

The sequences flanking the TAATGARATTC element are also conserved between the IE genes, and have some effect on the degree of response to $V_{mw}65$ (Preston et al., 1984). The

longer consensus, including these flanking nucleotides is 5' GYATGNTAATGARATTCYTTGNGGG 3' (Y= pyrimidine)(Mackem and Roizman, 1982b; Preston et al., 1984). In the upstream regions of IE genes 4 and 5, these flanking regions have not been conserved. This would suggest that they are not crucial for the induction of the IE genes (Preston et al., 1984).

Further upstream and downstream of the conserved element are G+C and G+A rich sequences. In the case of IE genes 4 and 5, the extent of stimulation by $V_{mw}65$ declines as the upstream G+C rich regions are removed (Preston et al., 1984). The G+A rich element was found to be required for full activation of IE gene 3 by $V_{mw}65$ (Bzik and Preston, 1986). Deletion of the G+C rich sequences downstream of the TAATGARATTC element also results in reduced stimulation. This effect however, could be due to altering the distance between the regulatory sequences and the promoter, or changing the relative positions of sequences within the far upstream region. These two G+C rich sequences contain inverted and direct repeats. Hairpin loop formation may modulate the response to $V_{mw}65$ (Mackem and Roizman, 1982a,b).

Low levels of IE175 have been found to give a small stimulatory effect on IE gene expression in transient assays (DeLuca and Schaffer, 1985). Plasmids containing the promoter region of an IE gene linked to the coding regions of the chloramphenicol acetyltransferase (CAT) gene have been used as expression vectors to measure this response. The stimulatory effect is diminished or abolished, however, with increasing levels of IE175 (DeLuca and Schaffer, 1985).

In normal infection, IE175 inhibits expression of IE genes. ts mutants with a defect in IE gene 3 do not shut down IE gene expression (Preston, 1979). Similar results

have been reported for mutants with deletions in IE gene 3 (DeLuca et al., 1985). Further evidence for the negative autoregulatory role of IE175 is provided by O'Hare and Hayward (1985b), who found that IE175 inhibited both basal levels and $V_{mw}65$- and IE110-activated levels of expression from its own promoter in a chimaeric IE-CAT construct in short-term cotransfection assays.

The stimulatory and inhibitory activities of IE175 on IE gene transcription are probably separate functions, possibly involving different domains of the protein. At non-permissive temperatures, a ts mutant of IE175 stimulated expression from IE-CAT constructs in transient assays. This effect required expression of the ts IE175 and the presence of the cis-acting regulatory sequences in the CAT construct (DeLuca and Schaffer, 1985).

The sequences involved in IE transcription regulation by IE175 have not been identified. IE175 is a DNA binding protein (Hay and Hay, 1980), and may be acting directly, binding to specific sequences in the upstream regions of IE gene promoters. Freeman and Powell (1982) have reported, however, that on purification, IE175 loses its ability to bind DNA. The association between IE175 and DNA may therefore be indirect, through forming a complex with another protein. O'Hare and Hayward (1985b) speculate that the IE175 negative autoregulation may be due to binding of IE175 to its own promoter. This would explain IE175 abolishing IE110-stimulation of transcription from the IE175 promoter. Similar conclusions have been drawn by DeLuca and Schaffer (1985), who suggest that both the stimulatory and inhibitory effects of IE175 on IE gene expression could be due to IE175 either alone or in a complex, binding at multiple sites on IE gene promoters. The ability of mutant forms of IE175 to bind may be altered, only allowing the interaction responsible for the stimulation of IE gene

expression. The ability of a ts mutant of IE175 to inhibit the activity of IE175 also supports the suggestion that specific binding or interactions involving the IE175 protein are essential for IE transcriptional regulation (DeLuca et al., 1985).

In short term transfection assays IE110 has been shown to stimulate expression from IE-CAT constructs (O'Hare and Hayward, 1985b). No specific sequence elements have been identified in the promoter or far upstream region of IE genes which could be responsible for the effect.

The specificity of transcriptional stimulation by IE110 is uncertain. In addition to up-regulation of IE gene expression, IE110 transactivates later classes of HSV genes (Everett, 1984a). IE110 is also a DNA binding protein (Hay and Hay, 1980). However, little is known of how transcriptional stimulation by IE110 is achieved. There is presently no evidence that IE110 and IE175 interact directly.

A protein produced at late times of infection may be involved in the down-regulation of IE gene expression, possibly in conjunction with IE175 (DeLuca et al., 1984). Two ts mutants of IE175, permissive for E gene expression, do not reinitiate IE gene expression on raising the temperature to non-permissive levels at late times (after 6 h postinfection). This differs from other ts IE175 mutants, which will recommence IE transcription at any stage of infection, on raising the temperature to non-permissive levels. This may suggest that a late function, acting in conjunction with or in addition to IE175, contributes in the negative regulation of IE gene expression (DeLuca et al., 1984).

1.8.2 Regulation of E gene expression

Other than its role in switching off IE genes, IE175 is required to activate E gene expression (Watson and Clements, 1980). E gene transcription is also stimulated by IE110 (Everett, 1984a). A further IE protein, IE12, has also been implicated in E gene transcription regulation (O'Hare and Hayward, 1985a).

IE175 by itself can induce E gene expression in short term transfection assays (Everett, 1984b). The induction of E genes by IE175 appears to be specific, involving cis-acting sequences in the upstream regulatory region. The presence of the TK promoter sequences alone was insufficient for transactivation of the TK gene by IE175 (O'Hare and Hayward, 1985a). In addition, some cis-acting regulatory domain, present further upstream of the cap site, was required. However, Everett (1983) reported that far upstream sequences were not required for the transactivation of the E glycoprotein D gene, but that the required sequences were all contained within 83 bp of the mRNA cap site.

In short term transfection assays, IE110 can also independently stimulate expression from E genes (O'Hare and Hayward, 1985a). The reported level of transactivation by IE110 relative to levels achieved by IE175 is variable (Gelman and Silverstein, 1985; O'Hare and Hayward, 1985a). Not all E genes may be induced by IE110 alone. Where transactivation does occur it appears to be less specific than $V_{mw}65$ or IE175, stimulating HSV genes of different classes (Everett, 1984a).

Stimulation of transcription of E genes is greatly increased in assays where genes for both IE175 and IE110 are cotransfected with the target gene (Everett, 1984b; Quinlan

and Knipe, 1985). This may be partly explained by higher levels of IE175 present due to IE110 stimulation of expression of IE175 (O'Hare and Hayward, 1985b), although this is thought unlikely (Everett, 1984b).

The conditional lethal mutant tsK fails to express E genes at the non-permissive temperature although it has functional IE110 genes (Gelman and Silverstein, 1985). This may suggest that IE175 and IE110 interact, possibly forming a transcriptional complex which is inactivated by the ts lesion. Alternatively, the ts IE175 may interact with or bind to the regulatory sequences of the E genes, thereby making them unavailable to IE110 (Gelman and Silverstein, 1985). Either suggestion is supported by the observation that transactivation by IE110 of a TK-CAT (CAT coding sequences linked to the promoter regions of the TK gene) construct is inhibited by cotransfection of a ts IE175 gene at non-permissive temperature (DeLuca and Schaffer, 1985).

It has been reported that IE12 may augment activation of E promoters by IE175 or IE110 or both together (O'Hare and Hayward, 1985a), although on its own IE12 shows no inducing activities (Everett, 1984b).

IE63 and IE68 have been found not to stimulate transcription from E promoters either independently or when transfected with other IE genes (Everett, 1984b; O'Hare and Hayward, 1985a). Interestingly, IE68 appears to slightly decrease IE175 and IE110 stimulation of E gene transcription in transient assays using E-CAT constructs (O'Hare and Hayward, 1985a). A mutant with a deletion in the IE68 gene has recently been described (Sears et al., 1985). The mutant exhibits a delayed shut-off of E gene expression, perhaps suggesting that IE68 may play a role in E gene expression regulation. The function of IE68 in HSV infection is unknown, but it is not essential for growth in

human and primate cell culture (Post and Roizman, 1981; Jacquemont et al., 1984). Growth of the IE68 mutant is restricted, however, in rodent cell lines (Sears et al., 1985).

1.8.3 Regulation of L gene expression

L promoters, although structurally similar to E promoters, have additional requirements for transcription (Wagner, 1985). The two subgroups of late genes, γ_1 and γ_2, differ in their time of expression and in the stringency of requirement for viral DNA synthesis (Conley et al., 1981). Only IE175 and IE110 have been shown to regulate L gene expression in transfection assays, although further IE proteins may be involved. Viral DNA replication appears to play a significant role in the activation of L genes (Hall et al., 1982; Johnson et al., 1986).

IE175 is required for L gene expression (Watson and Clements, 1980). DeLuca et al. (1984) have described a ts mutant of IE175 that expresses IE and E genes, induces viral DNA synthesis, but produces greatly reduced levels of L proteins at the non-permissive temperature. In transient assays, IE175 can induce transcription from L promoters (Silver and Roizman, 1985; Mavromara-Nazos et al., 1986). However, higher levels of IE175 and additional trans-acting HSV gene products may be necessary for induction of some (particularly γ_2) L genes (DeLuca and Schaffer, 1985). Generally, cotransfections with constructs containing IE175 together with the other IE genes increased the expression from L promoters (DeLuca and Schaffer, 1985).

The domains in IE175 responsible for induction of late promoters may be separate from those stimulating E gene transcription. ts mutations in the IE175 gene affecting

both E and L, or only L gene expression, have been reported (DeLuca et al., 1984).

IE110 by itself will stimulate expression from L promoters to levels comparable to E genes, in short term transfection assays (Mavromara-Nazos et al., 1986). The stimulation of L gene expression by IE110 may be responsible for the increased transcription of L genes by a plasmid containing all five IE genes, relative to the levels achieved by IE175 alone (DeLuca and Schaffer, 1985).

Recent studies have indicated that true late (γ_2) genes are differentially regulated depending on their genomic environment. Late genes introduced into cells in plasmid constructs or integrated in the cellular genome appear to be regulated in a manner similar to E genes, and are transactivated by IE110 and IE175, but not by IE63, IE68 or IE12 (Mavromara-Nazos et al., 1986). L genes resident in the viral genome, however, are not expressed until late times of infection (Silver and Roizman, 1985). Further evidence of this differential regulation comes from the characterisation of several ts mutants of IE63 (Sacks et al., 1985). Although infection at non-permissive temperature advanced to the stage of DNA replication, the mutants produced drastically reduced levels of several late proteins, implying a role for IE63 in L gene expression.

Two explanations have been given for the apparent requirement for additional factors. It has been suggested that host transcriptional factors required for late gene expression are sequestered during HSV infection and only become available through an effect of the viral DNA polymerase. The second hypothesis predicts that late gene transcription is blocked by specific viral proteins bound to the viral genome which are removed during DNA replication. Either suggestion could begin to explain the observed

differences in regulation of L genes isolated from or integrated in the viral genome (Mavromara-Nazos et al., 1986).

1.9 RELATIONS BETWEEN HSV AND OTHER HERPESVIRUSES

The diversity of the herpesviruses, demonstrated by their differing biological properties, has been discussed. At the genomic level, this diversity can be seen as large differences in size, structure and base composition.

The base composition of herpesvirus DNAs ranges widely, from 32% G+C for canine herpesvirus, to 75% for B-virus (Honess and Watson, 1977). Herpesviruses infecting a single host can also show a wide range in base composition. For example, VZV and HSV-2, both human herpesviruses, have G+C contents of 46% and 69% respectively (Ludwig et al., 1972; Kieff et al., 1971). The exception is the cytomegaloviruses, which all have G+C contents between 50% and 60% (Honess and Watson, 1977). The uneven distribution of high G+C sequences within the genome is a further feature of many herpesviruses. An extreme example is HVS, with 70% G+C repeat sequences and a 35% G+C unique region (Honess and Watson, 1977).

The range in base composition of the herpesviruses approaches the limits of protein coding capacity (Woese, 1967). There is evidence of HSV-1 making maximal use of degenerate codons and tolerable amino acid replacements which increase the use of codons with high G+C contents. To illustrate this, the coding region of the HSV-1 gene for the protein IE175, with a base composition of 81.5% G+C, has G+C contents in the three codon positions of 81.9%, 66.2% and 96.3% respectively. Although there is a strong bias towards

a G or C residue in the third codon position, the four most common amino acids are Ala, Pro, Gly and Arg, all of which have codons containing only G+C residues (McGeoch et al., 1986a).

Despite great differences in base composition, a closer relationship between HSV and other alphaherpesviruses than with members of the beta- or gammaherpesviruses has been demonstrated. First detected as common antigens between viruses, similarities have been demonstrated as homology at the DNA sequence level and in the colinearity of gene arrangement. Homology between amino acid sequences of related proteins coded by the viruses has further emphasised the closer relationship between viruses within the subgroup. The relationship between HSV and viruses of the alpha-, beta- and gammaherpesviruses will be discussed below.

1.9.1 Alphaherpesviruses

HSV-1 and HSV-2 share multiple antigens and will cross-neutralise (Honess et al., 1974). The very close relationship between HSV-1 and HSV-2 has been demonstrated by their ability to recombine (Timbury and Subak-Sharpe, 1973). Limited cross-reactivity between HSV and other alphaherpesviruses has demonstrated a relationship between the viruses (Honess, 1984).

a) DNA cross hybridization

Extensive DNA sequence homology between the genomes of HSV-1 and HSV-2 has been demonstrated (Kieff et al., 1972). Southern blot analyses of restriction enzyme digests of genomic DNA have been used to identify the regions of conserved sequences. Almost all regions of the HSV-1 and

HSV-2 genomes show some homology. Least conserved sequences include the repeat sequences and the majority of the short segment. The analysis confirmed the belief that the two genomes are colinear (Davison and Wilkie, 1983b).

Heteroduplex analysis of cloned fragments of HSV-1 and HSV-2 has shown that this homology is not uniform throughout the length of the genome (Kudler et al., 1983). Three regions of the genome were investigated. The region between 0.3 and 0.4 map units showed extensive sequence homology between HSV-1 and HSV-2. This region contains the gene for gB, which shows cross-reactivity between the two serotypes (Spear and Roizman, 1980). The region between 0.2 and 0.3 map units, also in U_L, showed a high level of DNA sequence homology. Heteroduplex analysis of fragments spanning most of the U_S region however, suggested that this sequence is less highly conserved between HSV-1 and HSV-2 (Kudler et al., 1983).

Homology between the DNA sequences of HSV and other alphaherpesviruses has also been investigated. Rand and Ben-Porat (1980) detected a low level of sequence homology between PRV and HSV by DNA hybridisation. Homologous sequences were located in the U_L region and to a lesser extent in U_S. Only low levels of homology were detected in the repeat sequences. PRV appears to contain sequences in U_S showing greater homology to HSV-2 than to HSV-1. Davison and Wilkie (1983b) also reported finding additional regions of homology between PRV and HSV-2, and located these homologous sequences to the R_S region of HSV-2. They reported finding no homology in the U_S region. The PRV genome was found to be generally colinear with the I_L or I_{SL} arrangement of HSV, although a region within the U_L sequence of PRV was inverted relative to HSV.

Davison and Wilkie (1983b) also demonstrated sequence

homology between HSV and VZV. A detectable level of hybridisation was observed in most regions of the HSV genomes. Regions showing least homology included all of U_S and sequences within U_L between approximately 0.1 and 0.3 map units and between 0.45 and 0.55 map units. The L segment of VZV genome was found to be inverted relative to the P arrangement of HSV (Davison and Wilkie, 1983b).

The EHV-1 genome also showed some homology to HSV by DNA hybridisation. No homology between the DNAs of EHV-1 and HSV in the U_S sequences was found. The extent of homology was sufficient to conclude that the EHV-1 genome is colinear with the I_L or I_{SL} arrangement of HSV (Davison and Wilkie, 1983b).

b) Gene homology and arrangement

The determination of the complete DNA sequence of the short segment of the HSV-1 and VZV genomes has enabled a comparison of the predicted amino acid sequences of the encoded genes (Murchie and McGeoch, 1982; McGeoch et al., 1985; McGeoch et al., 1986a; Davison, 1983; Davison and Scott, 1985). The S segment of the HSV-1 genome contains 14 genes (including two copies of IE gene 3). The HSV-1 R_S sequence contains one entire gene, and the 5' non-coding regions of two further genes. The U_S sequence contains ten genes plus the coding regions of the two genes located across the R_S/U_S junction (McGeoch et al., 1985 and 1986a). In VZV the S segment encodes only seven genes. The U_S sequence contains two entire genes, and the majority of two further genes. Each copy of R_S contains three entire genes.

All seven of the VZV proteins have counterparts in the HSV-1 S segment. Although only low levels of DNA sequence homology in S were detected by DNA hybridization,

significant similarities between related proteins are visible (Davison and McGeoch, 1986).

Similarities between the corresponding proteins varied. Some proteins showed extensive amino acid sequence homology, such as the HSV-1 IE175 protein and the VZV counterpart. Other proteins showed only low levels of sequence conservation. Homologous amino acid sequences were not generally distributed evenly throughout the proteins. Least conservation of amino acid sequences was usually found at the amino terminal portion of the proteins. Differences in sizes between pairs of homologous proteins was also usually due to heterogeneity at the amino terminus.

With the identification of related proteins in the two viruses, the arrangement of the genes for the homologous proteins was examined (Davison and McGeoch, 1986). The genetic organisation of the S segment of the two virus differs, as shown in figure 1.4. The genes coding for corresponding proteins in each virus are indicated.

The changes in gene arrangement between the two viruses occurring during their evolution from a common ancestor are believed to have arisen through a number of recombinational events. This could have resulted in the expansion and contraction of the repeat sequences in S. The encroachment of R_S into U_S would result in the inclusion of genes in R_S, originating from the U_S sequence, and the loss of other genes (Davison and McGeoch, 1986).

1.9.2 Betaherpesviruses and gammaherpesviruses

The characterization of several of these viruses has suggested they show some similarity to HSV. At the DNA sequence level, HCMV (a betaherpesvirus), EBV and MDV

Figure 1.4 Relationship between genes in the S segments of HSV-1 and VZV

The location and orientation of predicted polypeptide coding regions in the S segment of the two viruses are indicated. The R_S regions are represented as open boxes and the U_L sequences as solid lines. Homologous genes are indicated by arrows. For clarity, relationships are indicated for only one copy of genes in the inverted repeats. From Davison and McGeoch (1986).

(gammaherpesviruses) show no detectable sequence homology by cross hybridisation to HSV DNA (Honess and Watson, 1977). Evidence for an evolutionary relationship between HSV and other herpesviruses has often been based on the identification of apparently equivalent virally encoded proteins in infected cells. A DNA polymerase activity distinguishable from that of the host, has been identified in cells infected with HSV-1, HSV-2, HCMV, EBV and other herpesviruses (Honess and Waton, 1977).

The complete DNA sequence determination of EBV (Baer et al., 1984) has enabled comparisons of the predicted amino acid sequences of the encoded genes with data from HSV. Clear homology has been found between the amino acid sequences of the major DNA binding protein, the DNA polymerase, glycoprotein B and the two ribonucleotide reductase subunits of HSV-1 and EBV (Quinn and McGeoch, 1985; Gibson et al., 1984). The relative arrangement and location in the genome of the DNA polymerase, DNA binding protein and gB genes has not been entirely conserved between the two viruses (Quinn and McGeoch, 1985).

A 7.8 kbp DNA sequence from near the left end of the U_L region of HSV-1 has also been determined (McGeoch et al., 1986b) and the transcripts mapped (Costa et al., 1983). Of the seven polypeptides believed to be coded in this region, five (including the alkaline exonuclease) show homology to translated EBV ORFs. In addition to the amino acid sequence homology between the predicted polypeptide sequences, the genes are similarly arranged although their location within the genome differs (McGeoch et al., 1986b). The organisation of homologous genes found to date are illustrated in figure 1.5. (McLauchlan and Clements, 1983; Gibson et al., 1984; Quinn and McGeoch, 1985; McGeoch et al., 1986b).

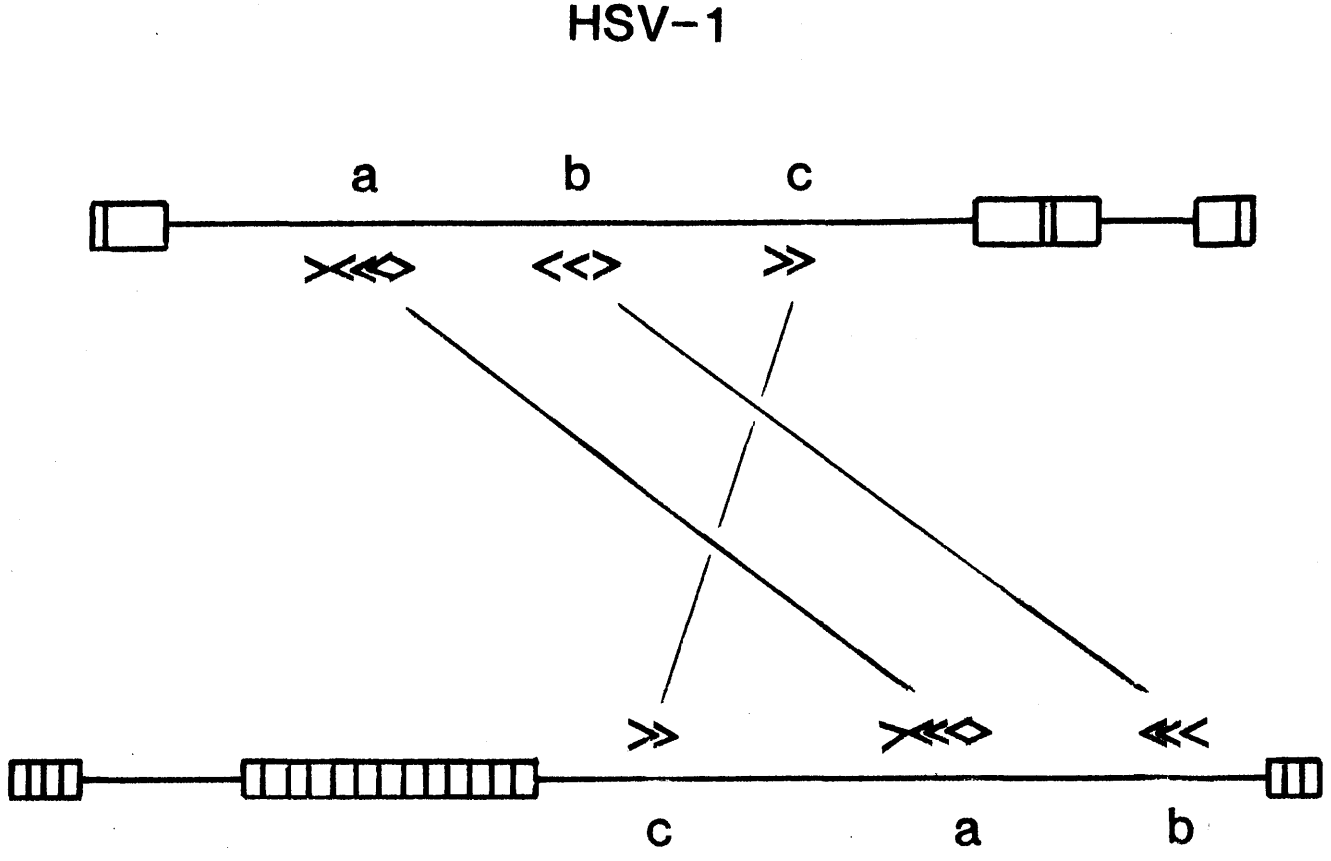

HSV-1
a
b
c
c
a
b
EBV

Figure 1.5 Location of homologous genes in HSV-1 and EBV

The HSV-1 prototype genome (above) and EBV genome are shown. Repeat elements are shown as open boxes, unique sequences as solid lines. All genes encoding detectable homologues reported to date are illustrated. From left to right, in the HSV-1 genome, homologous genes labelled are as follows. a) exonuclease and neighbouring genes; b) glycoprotein B, major DNA binding protein and polymerase genes; c) genes for both of the ribonucleotide reductase subunits (McLauchlan and Clements, 1983; Gibson et al., 1984; Quinn and McGeoch, 1985; McGeoch et al., 1986b).

Therefore, despite the identification of homologous genes in HSV-1 and EBV, and a similar local arrangement of genes, large scale rearrangements of DNA sequences must have occurred during the evolution of these viruses. All the homologous genes identified to date are located within the U_L region of the viruses. No homology has been detected between any EBV ORF and the genes located in the short region of HSV-1 (McGeoch and Davison, 1986). The structure of the EBV genome in relation to the organisation of unique and repeat sequences is markedly different from that of HSV-1, as described in section 1.2. Further, no DNA sequence homology between the two viruses has been detected by cross hybridisation experiments (Honess and Watson, 1977).

MATERIALS AND METHODS

MATERIALS

2.1 Enzymes

Restriction endonucleases, T4 DNA polymerase and T4 polynucleotide kinase were obtained from Bethesda Research Laboratories.
The large fragment of Escherichia coli polymerase I (Klenow), was obtained from Boehringer Corporation Limited, or provided by A.J. Davison.
Calf intestinal phosphatase was obtained from Boehringer Corporation Limited.
E.coli DNA polymerase I and T4 DNA ligase were obtained from New England Biolabs.
Lysozyme and proteinase K were purchased from Sigma Chemical Company Limited.

2.2 Chemicals

5-chloro-4-bromo-3-indolyl-β-D-galactosidase (BCIG) and isopropyl-D-thiogalactoside (IPTG) were obtained from Sigma Chemical Company Limited.
3'-deoxyribonucleoside 5'-triphosphates and 2',3'-dideoxyribonucleoside triphosphates were supplied by Pharmacia P-L Biochemicals.
Synthetic oligonucleotide primers were obtained from Celltech Limited.
Other laboratory chemicals were obtained from BDH and Sigma.

2.3 Radiochemicals

All radiochemicals were obtained from the Radiochemical centre, Amersham:
[α^{32}P]dNTPs at 3000 Ci/mmol, 10 mCi/ml.
[γ^{32}P]ATP at 5000 Ci/mmol, 10 mCi/ml.

2.4 Miscellaneous Materials

3MM chromatography paper was purchased from Whatman Limited. Nitrocellulose 82 mm discs (BA85, 0.45 um pore) and nitrocellulose paper (BA85, 0.45 um pore) were obtained from Schleicher and Schuell. Wacker silicone was provided by Wacker-Chemie GmbH, Munich.
Repelcote (2% solution of dimethyldichlorosilane in 1,1,1-trichloroethane) was supplied by Hopkin and Williams.

2.5 Buffers and Solutions

TBE	90 mM Tris-base, 90 mM boric acid, 2.5 mM EDTA. pH 8.3.
TE	10 mM Tris.HCl, 0.1 mM EDTA. pH 8.0.
NTE	10 mM Tris.HCl, 10 mM NaCl, 1mM EDTA. pH7.4
SSC	0.15M NaCl, 15 mM Na citrate. pH 7.5.
Denhardts	0.02% (w/v) each Ficoll, polyvinylpyrrolidone, BSA (bovine serum albumin) in 0.06 X SSC.

2.6 Agar and Bacterial Growth Media

2YT	85 mM NaCl, 1.0% (w/v) bactopeptone, 1.0% (w/v) yeast extract.
L Broth	177mM NaCl, 1.0% (w/v) bacto-peptone, 0.5% (w/v) yeast extract.
L Broth Agar	1.5% (w/v) agar in L Broth.
Top Agar	1.0% (w/v) agar in water.

METHODS

2.7 Glycerol stocks of bacteria

Bacterial stocks were prepared from 20 ml standing cultures grown overnight in 2YT, at 37°C. After centrifugation at 5000 rpm for 5 min, 4°C, the bacterial cell pellets were resuspended in 2 ml 2% Difco bactopeptone (w/v) and 2 ml 80% glycerol (v/v) was added. The stocks were stored at -70°C.

2.8 HSV-1 recombinant plasmids

The bacterial hosts used were E.coli K12 strains HB101 (Boyer and Roulland-Dussoix, 1969) and DH1 (Hanahan, 1983). The following recombinant plasmids carrying restriction fragments of HSV-1 cloned into pAT153 (Twigg and Sherratt, 1980) were used in this work:

plasmid	HSV-1 fragment(s)	cloning site	source
pGX48	BamHI b	BamHI	F.J. Rixon
pGX23	BamHI e	BamHI	V.G. Preston
pGX190	HpaI s and v	EcoRI/BamHI	C.M. Preston
pGX53	SalI-BamHI subfragment of BamHI b	BamHI/SalI	F.J. Rixon

2.9 Preparation of plasmid DNA

A standing culture of 100 ml 2YT broth with 100 µg/ml ampicillin inoculated with 50 µl from a glycerol

stock was incubated overnight at 37°C. This was added to 1.6 l L broth and incubated at 37°C in an orbital shaker for 3-4 h until it reached an optical density at 550 nm of approximately 0.7 ($O.D._{550}$ 0.7). The plasmid was amplified by the addition of chloramphenicol to 25 µg/ml (Clewell, 1972) and the culture shaken overnight at 37°C.

The isolation of plasmid was by soft lysis as described by Katz et al. (1973). The bacteria were pelleted by centrifugation at 5000 rpm for 5 min and the cell pellet was resuspended in 10 ml 25% sucrose (w/v) in 50 mM Tris.HCl (pH 8.0). All subsequent procedures were carried out on ice. Lysozyme was added to a concentration of 2 mg/ml to disrupt bacterial cell walls. After 5 min, 4 ml 250 mM EDTA (pH 8.0) was added, and incubation on ice continued for a further 5 min. Lysis of the bacterial cells was achieved by adding 16 ml of 2% (v/v) Triton X-100 in 50 mM Tris.HCl (pH 8.0), 2.5 mM EDTA and standing 20 min on ice. The lysate was centrifuged for 20 min at 15000 rpm at 4°C to pellet cell debris and the majority of the chromosomal DNA. The supernatant was mixed vigorously with an equal volume of NTE-equilibrated phenol and centrifuged at 5000 rpm for 5 min. The upper aqueous layer was phenol extracted again and the DNA precipitated with three volumes of ethanol and a tenth volume 4 M NaOAc at -20°C.

The plasmid DNA was purified by isopycnic banding on caesium chloride gradients to remove residual host cell DNA and RNA. To prepare the gradient, the DNA was recovered from the ethanol precipitation by centrifugation at 8000 rpm for 30 min, rinsed in 70% ethanol, and dried. The pellet was resuspended in 0.75 g/ml CsCl containing 0.5 mg/ml ethidium bromide.

Centrifugation was carried out at 40000 rpm for 18 h, at 15°C in a TV865 or TV865B rotor. The lower fluorescent red band, visualised under long wave (365 nm) UV irradiation contained the supercoiled plasmid DNA and was removed using a needle and syringe. The ethidium bromide was removed from the sample by extraction with CsCl-saturated isopropanol. To precipitate the DNA, three volumes of water and nine volumes of ethanol were added, and the sample held at -20°C.

2.10 Transformation of bacterial cells with plasmid DNA

Host bacterial cells were grown to an $O.D._{630}$ 0.3 and pelleted by centrifugation at 5000 rpm for 5 min at 4°C. The pellet was resuspended in a half volume of ice-cold 50 mM $CaCl_2$ and incubated on ice for 20 min. The cells were centrifuged again and resuspended in a tenth volume of ice-cold $CaCl_2$. Plasmid DNA (20-200 ng) was added to 0.2 ml aliquots of calcium shocked cells and incubated on ice for 1 h. The samples were heated to 42°C for 2 min and transferred to 1 ml 2YT and grown as shaking cultures at 37°C. After 1 h, 0.2 ml aliquots of each sample were spread on L broth agar plates containing the appropriate selective antibiotic and incubated overnight at 37°C. Colonies picked from the plates were streaked on duplicate plates containing antibiotics; one antibiotic to which it was resistant and one to which it was sensitive. For example bacteria containing the plasmid pGX48 are ampicillin resistant and tetracyclin sensitive. Colonies were restreaked onto L broth agar plates and grown overnight. Individual colonies were picked and glycerol stocks prepared, as described in section 2.7.

2.11 Isolation and preparation of DNA prior to cloning

DNA fragments to be sequenced were digested with the appropriate restriction enzyme and purified from the vector by gel electrophoresis. Generally, fragments greater than 2000 base pairs in length were purified by elution from low melting point agarose and smaller fragments from polyacrylamide gels.

a) Elution from low melting point agarose

The digested DNA was electrophoresed through a 1% agarose minigel run in 1 X TBE with 1 µg/ml ethidium bromide at 30 mA. Under long wave UV illumination, a slice containing the DNA fragment was cut from the agarose. The agarose was heated to 65°C in two volumes of TE for 10 min. The sample was extracted twice with an equal volume of phenol equilibrated with TE, and ethanol precipitated.

b) Elution from polyacrylamide gel

The digested DNA was electrophoresed through a 4% polyacrylamide gel run in TBE at 40 mA for 6 h. The gel was soaked in 5 µg/ml ethidium bromide for 5 min and rinsed. Under long wave UV illumination, a slice containing the DNA fragment was cut from the gel. The gel slice was extruded through a hole pierced in a 0.5 ml Eppendorf tube by centrifugation. The acrylamide was mixed vigorously with 0.5 ml elution buffer [0.5 M NH_4OAc, 1 mM EDTA, 0.1% SDS (sodium dodecyl sulphate)] and incubated overnight at 55°C. The sample was filtered through sterile glass wool, extracted with TE saturated phenol and ethanol precipitated.

The DNA was prepared for insertion into M13 by two

methods. a) The DNA was digested with a restriction enzyme yielding flush ends, such as AluI, RsaI and SmaI, for ligation into the SmaI site of the vector. Alternatively, the DNA was digested with Sau3AI and ligated into the BamHI site of the vector.

b) A large proportion of the recombinant M13 clones were prepared using sonicated DNA. After the DNA was isolated from the plasmid, the ends of the fragment were ligated together using 2 units of T4 DNA ligase in ligation buffer [10 mM Tris.HCl (pH 7.5), 10 mM $MgCl_2$, 10 mM DTT (dithiothreitol), 1 mM ATP], to a final volume of 10 µl, at 15°C for 48 h. Random DNA fragments were generated by sonication as described by Deininger (1983). The DNA (5-20 µg) was suspended in 0.5 ml sonication buffer [500 mM NaCl, 100 mM Tris.HCl (pH 8.00), 1 mM EDTA]. This was sonicated in 5 X 5 s bursts, with 40 s intervals, using a Daves soniprobe set at minimum power.

The ends were made flush by incubation with 4 units of T4 DNA polymerase at 37°C for 1 h in 67 mM Tris.HCl (pH 8.0), 7 mM $MgCl_2$, 0.025mM each of dCTP, dGTP, dATP, dTTP and 5 mM DTT. The reaction was stopped by the addition of an equal volume of 10 mM EDTA followed by extraction with phenol equilibrated with TE, and ethanol precipitation.

The products of the sonication were size fractionated prior to cloning. The DNA was run on a 1.5% agarose, 1 X TBE, 1 µg/ml ethidium bromide, minigel at 50 mA alongside DNA markers of appropriate size. pAT153 DNA digested with HinfI, giving fragment sizes of 1651, 517, 396, 298, 221, 220, 154, 145 and 75 bp, was generally used. Under long wave UV illumination, a trough was cut in the sample track just ahead of the

220/221 bp marker band, and this was filled with buffer. The gel run was continued and the buffer in the trough collected and replaced at 30 s intervals. Fractions of buffer were collected until, upon UV visualisation, the size of the fragments in the trough was estimated to be greater than 500 bp. The DNA was extracted with TE saturated phenol and ethanol precipitated. The DNA was pelleted and resuspended in TE buffer to a final concentration of 50-500 ng/µl.

2.12 Construction of recombinant M13 clones

The double stranded replicative form (RF) of M13mp8 (either prepared as for plasmids, Katz *et al.*, 1973, or obtained from Amersham) was used for all cloning experiments. The vector was linearised with SmaI and 5' phosphate groups removed with calf intestinal phosphatase, to abolish recircularisation of the vector during ligation. Vector DNA (10 ng) was incubated with 10-500 ng DNA insert with 2 units of T4 DNA ligase in ligation buffer, as described above in section 2.11, for 48 h at 15°C (Sanger *et al.*, 1980).

2.13 Transfection of bacterial cells with M13

E.coli JM101 were grown to an $O.D._{630}$ 0.3 and calcium shocked as described in section 2.10. The ligation mixture was added to 0.2 ml aliquots of the calcium shocked cells and incubated on ice for 40 min. The samples were heated to 42°C for 2 min. 3 ml of melted top agar at 42°C containing 20 ul IPTG (100 mM) and 25 ul of 2% BCIG in dimethylformamide was added to the sample and the mixture poured on to 90 mm L broth agar plates and incubated at 37°C overnight.

2.14 Growth and Extraction of recombinant M13 clones

An overnight standing culture of JM101 was used to inoculate 2YT broth (1:100). This was dispensed in 1.5 ml aliquots into 25 ml Universal bottles. The recombinant (colourless) plaques were toothpicked from the plates into the broth and incubated at 37°C for 6-7 h with shaking.

The clones were centrifuged at 10000 rpm for 5 min to pellet the bacterial cells. The supernatant (0.8 ml) was transferred to a 1.5 ml Eppendorf tube and the phage precipitated by the addition of 0.2 ml 2.5 M NaCl, 20% polyethylene glycol (PEG) 6000 for 30 min at room temperature or 4°C overnight. After centrifugation at 10000 rpm for 5 min, the supernatant was removed. After a brief centrifugation, any remaining supernatant was removed with a glass capillary tube. The pellet was resuspended in 100 ul TE, extracted with 50 µl phenol equilibrated with TE, and ethanol precipitated. After centrifugation for 5 min at 10000 rpm, the DNA pellet was resuspended in 50 µl TE and stored at -20°C.

2.15 Screening recombinant clones

The DNA fragment was not always isolated from the plasmid prior to sonication. In this case a proportion of the recombinant M13 clones would contain pAT153 DNA inserts. Dot blot hybridisation of these clones was used to identify useful clones. Generally two probes were used on duplicate filters, namely, one probe of the DNA fragment to be sequenced and the second probe of pAT DNA.

2.16 Preparation of nick translated DNA

The DNA (0.5 µg) to be used as a probe was labelled with 2 µCi [α^{32}P]dATP using 1 unit of DNA polymerase I and 1 µl DNase (10^{-4} µg/ul) in a reaction mix containing 0.05% BSA, 0.1 mM each of dTTP, dCTP and dGTP in nick translation buffer [50 mM Tris.HCl (pH 7.8), 5 mM $MgCl_2$, 1 mM DTT] in a final volume of 20 µl. The reaction was carried out at 15°C for 90 min. The DNA was isopropanol precipitated for 30 min with a half volume of isopropanol and a tenth volume 4 M NaOAc on dry ice. After centrifugaton at 10000 rpm for 5 min, the pellet was resuspended in TE and isopropanol precipitated again.

2.17 Probing recombinant M13 clones

A sample of DNA from each clone processed (0.5 µl) was spotted on an 82 mm nitrocellulose grid. The nitrocellulose was air dried and baked under vacuum at 80°C for 2 h. The nitrocellulose was prehybridised at 68°C, shaking gently, in hybridisation buffer [20 mM Tris.HCl (pH 7.5), 6 X SSC, 10 X Denhardt's buffer, 1 mM EDTA, 0.5% SDS, 50 µg denatured calf thymus DNA]. The nick translated probe was pelleted as before, taken up in 80% formamide and heated at 100°C for 5 min. The nitrocellulose was then incubated in fresh hybridisation buffer containing the nick translated probe. Hybridisation was continued overnight, with shaking, at 68°C. The nitrocellulose was washed three times for 30 min in 2 X SSC, 0.2% SDS at 60°C, with shaking. The filters were air dried and autoradiographed.

Table 2.1 Nucleotide concentrations used in DNA sequencing reactions

component	sequencing solution			
(concentration)	T	C	G	A
dNTPs (0.5mM)				
dTTP	25	500	500	500
dCTP	500	25	500	500
dGTP	500	500	25	500
DDNTPS (5MM)				
DDTTP	30	---	---	---
ddCTP	---	2	---	---
ddGTP	---	---	5	---
ddATP	---	---	---	3
TE buffer	945	973	970	497

2.18 Sequence analysis of recombinant M13 clones

Single stranded DNA template (2 µl) was annealed to 2.5 ng of a commercial oligonucleotide primer (purchased from Celltech) in 10 mM Tris.HCl (pH 8.5), 10 mM NaCl in a volume of 10 µl at 37°C for 30 min. The annealed DNA was aliquoted in 2 µl fractions into four 0.5 ml Eppendorf tubes, corresponding to the specific T, C, G and A reactions of each clone (Sanger et al., 1977). An equal volume of nucleotide mix containing dNTPs and specific ddNTPs (see Table 2.1), 1 µM dATP, [α^{32}P]dATP (0.25 µCi/clone), and polymerase I (Klenow fragment) was added to each tube. The tube contents were mixed and the reaction allowed to proceed for 15 min at room temperature. The reactions were chased by addition of dATP to a final concentration of 0.1 mM for 30 min at room temperature. The reaction was stopped by the addition of formamide dye mix [0.1% bromophenol blue (w/v), 0.1% xylene cyanol (w/v) in formamide]. The samples were heated to 100°C for 1 min and electrophoresed through polyacrylamide gels.

2.19 Electrophoresis and autoradiography of sequence gels

Electrophoresis was carried out through vertical gels 40 X 20 X 0.03 cm in size. Spacers and gel combs were cut from Plasticard. The notched plate was treated with Repelcote. The plain plate was treated with 0.5% Wacker silicone in 0.3% acetic acid and ethanol (Garoff and Ansorge, 1981). This bonds the acrylamide to the plate allowing the gel to be dried down after electrophoresis.

2.20 Gradient gels

Generally, electrophoresis was carried out through TBE gradient gels (Biggin et al., 1983). In this system a buffer gradient is produced in the gel, with 0.5 X TBE in the top and 2.5 X TBE in the bottom of the gel. This was achieved using two acrylamide gel mixes. The top mix was of 0.5 X TBE, 6% acrylamide and 9 M urea. The bottom mix was composed of 2.5 X TBE, 6% acrylamide, 9 M urea, 5% sucrose (w/v) and 0.1% bromophenol blue (w/v), to visualise the gradient. When preparing a 40 X 20 X 0.03 cm gel, ammonium persulphate and TEMED were added to both the top (0.016% and 0.16%, respectively) and bottom (0.02% and 0.2%) gel mixes. 4 ml of top mix was drawn into a 25 ml syringe followed by 6 ml of bottom mix. A few bubbles were passed through to allow some mixing. The syringe contents were slowly expelled between the plates to the bottom. The remainder of the top gel mix was used to make up the rest of the gel. The gel comb was inserted into the top and the gel rested in a near horizontal position until polymerisation was complete. The tape was removed from the bottom of the gel, and the plates set up with 0.5 X TBE in the top and 1 X TBE in the bottom tank of a gel kit. After removing the comb, the wells were flushed with 0.5 X TBE and the DNA samples loaded. The gels were run at a constant power of 40 W for 2 h 30 min.

2.21 Alternative sequencing systems

In some instances, the sequence of the DNA insert in the template will form secondary structures which prevent the polymerase reading through the region. This results in "pile-ups". Pile-ups are observed on the

autoradiograph of the gel as a concentration of label in all four sequencing tracks at the same position. Usually, the bands above this region are of a lower intensity. This effect can be alleviated by running the sequencing reactions at 37°C, rather than at room temperature.

The sequence of the newly synthesized DNA strand may lead to the formation of secondary structures which result in an anomalous gel migration pattern, called compressions. To resolve this, the gels can be maintained at a higher temperature using a jacket of circulating water at 80°C. In this system, the sequencing reactions were carried out as described and run on a 0.5 X TBE, 6% acrylamide, 9 M urea gel. Prior to loading the samples, the gel was heated, using the hot water jacket, for one hour and prerun at 40 W for 30 min. Once the samples were loaded the gel was run at 40 W for 3 h.

Formamide gels containing 30% deionised formamide, 6% acrylamide, 8 M urea and 1 X TBE were also used to resolve ambiguous DNA migration patterns. Gels were prerun for 30 min, and the run continued at 40 W, after loading the samples.

The analogue of dGTP, deoxyinosine triphosphate (dITP), was also used to reduce ambiguities in migration pattern. Sequencing reactions were carried out as described using a different set of sequencing solutions. The solutions were made up as described in Table 2.1, but substituting the 0.5 mM dGTP with 2.0 mM dITP, and the 5.0 mM ddGTP with 50 mM ddITP. The reactions were chased with 0.2 mM dGTP and 0.2 mM dATP.

2.22 Autoradiography

After electrophoresis, the plates were dismantled and the gel, bonded to the plain plate, immersed for 30 min in 10% acetic acid to fix the DNA and remove the urea from the gel. The gels were dried down on the plate in an oven at 125°C for 1-2 h. An overnight exposure against Kodak X-Omat film at room temperature generally gave the required intensity.

2.23 Accumulation, storage and handling of sequence data

DNA sequence data were handled and interpreted using the Institute of Virology's PDP 11/44 computer operating under RSX11M. The gel readings were read and typed into the account using BATIN (Batch input) which stores data from gel readings under a chosen file name (Staden, 1977). Additionally, gel readings were entered using the digitising pad using the program DPAD (Digitising pad, written by P. Taylor), which stores the data in the same manner.

The compilation of the individual gel readings entered into the database was achieved using a group of programs designated BATCH, based on the DBAUTO system of Staden (1982). Firstly, a search is made to detect any restriction site chosen by the operator in the gel readings. This helps to avoid entering gel readings which extend beyond the cloning site into vector sequences, or to read across religated termini of the starting DNA fragment. The gel readings which pass this test are compared against the cloning vector sequences. M13 clones which do not have an insert, or have inserts containing pAT153 sequences are rejected. The data are

then compared against a consensus of the gel readings currently held. Gel readings which show a level of homology to the latest consensus sequence are entered into the database in the appropriate region of the sequence. Where necessary, the program will insert padding characters into either the newly entered gel reading or into the consensus sequence, to enable the sequences to be correctly aligned. Gel readings which show no homology to any sequence held in the database, are also entered, but are held individually as separate "contigs".

The sequence held in the database could be manipulated with DBUTIL (database utility) (Staden, 1980). This allows individual gel readings to be entered directly. In addition it gives access to all the data in the database, including lists of the entered gel readings with details of the location, length and orientation of the clone.

During the analysis of the determined sequences, several programs were used. These are now listed together with a brief description.

BASES This counts the number of each base in a nucleotide sequence at a specified position (for example every third base), or in the whole sequence (P.Taylor, unpublished).

CINTHOM The program searches for homology between two amino acid or nucleotide sequences and displays the results in a matrix plot. Homologous sequences are displayed as letters. Greatest homology is scored with capital letters. The parameters such as window length are set by the operator (Pustell and Kafatos, 1982). This program has been adapted by P.Taylor (unpublished)

to score for conservative amino acid changes, as evaluated by Dayhoff (1983).

CHOP This provides a rapid editing facility which will delete strings of nucleotide or amino acid sequences from a file (P. Taylor, unpublished).

DIAG Results from CINTHOM can be converted to a dot plot using this program (P.Taylor, unpublished).

DSPLY This gives a graphic display of open reading frames (ORFs) in a DNA sequence. Only ORFs greater than a chosen length, or ORFs starting with an initiation codon can be specified (Cold Spring Harbor Laboratory).

FRMSCN This is used to evaluate the codon usage of all three reading frames in a DNA sequence. The coding region of a similar gene is used as a reference (Staden and McLachlan, 1982).

HOMOL This program will optimally align two amino acid or DNA sequences and indicate identical residues. The parameters are set by the operator (Taylor, 1984).

MOLGEL Using marker tracks as reference, this program will calculate the size of an unknown DNA fragment on a gel (P. Taylor, submitted).

PROFILE This will give a graphic display of hydropathicity plots of amino acid sequences (Based on Kyte and Doolittle, 1982), or G+C contents of DNA sequences (P. Taylor, unpublished).

PTRANS This will give an amino acid translation of a specified region of a DNA sequence. It will also provide amino acid content and codon usage tables (P.

Taylor, 1986)

SEARCH This will search for specified strings of characters in a nucleotide or amino acid sequence (R. Staden, unpublished).

SEQLST This was used to provide listings of DNA sequence files (P.Taylor, unpublished).

TURN This program will reverse and complement a nucleotide sequence (P.Taylor, unpublished).

WORDSEARCH The National Biomedical Research Foundation protein sequence database (version 6) was searched using the wordsearch program of the "Wisconsin package" (Devereux et al., 1984) as implemented on a VAX at the Edinburgh Regional computing centre.

2.24 Calculation of DNA length by electrophoresis of labelled fragments

To determine the size of a particular restriction fragment which includes a tandem reiteration, 0.5 µg pGX48 was digested with SmaI, and the 5' phosphate groups removed with calf intestinal phosphatase. After complete digestion, the DNA was extracted once with TE equilibrated phenol and ethanol precipitated. The pelleted DNA was labelled with [γ ^{32}P]ATP using 5 units T4 polynucleotide kinase in 40 mM Tris.HCl (pH 7.5), 10 mM $MgCl_2$, 5 mM DTT for 30 min. at room temperature. Radioactive size markers were prepared by digesting 1 µg pAT153 DNA with 5 units HaeII in 50 mM Tris.HCl (pH 8.0), 10 mM $MgCl_2$, 50 mM $CaCl_2$. TBE dyes [0.1% bromophenol blue (w/v), 0.1% xylene cyanol (w/v) in 5 X TBE] were added to the samples. The DNA was loaded

on a 1 X TBE, 3% polyacrylamide gel and run at 40 v, alongside the pAT153/HaeII markers. The dyes were used as size markers to estimate the length of the run. After drying down the gel, an overnight exposure on Kodak X-Omat film was taken. Using the computer program MOLGEL (Taylor, unpublished), with the pAT153/HaeII fragments as reference, the size of the DNA fragments was calculated.

RESULTS

RESULTS

3.1 SEQUENCE DETERMINATION

The DNA sequences of the HSV-1 restriction fragment BamHI b and part of BamHI e were determined. The plasmids used for sequencing are shown in figure 3.1. The entire BamHI b fragment and the SmaI - BamHI subfragment containing the U_L component of BamHI e were sequenced using dideoxy chain termination reactions of recombinant M13 clones containing random DNA fragments. Examples of three sequenced clones are shown in figure 3.2. The recombinant M13 clones generated data covering the whole sequence investigated. The data were compiled and analysed using the computer programs described in the previous chapter.

The databases containing the compiled sequences are shown in figures 3.3 and 3.4. Listings of the M13 clones used to sequence the two fragments are given in figures 3.5 and 3.6. The sequence determined for the BamHI b fragment, shown in figure 3.7, is 10041 bp in length and has a base composition of 68.8% G+C. The sequence was determined from 116,136 characters from 611 sequence readings, so that residues were sequenced an average of 11.6 times. 99.5% of the sequence was obtained for both strands. In BamHI e, the sequence determined for the SmaI-BamHI subfragment, shown in figure 3.8, is 3160 bp in length and has a G+C content of 64.2%. The sequence was determined from 109 gel readings giving 22037 characters, or 7.0 characters per base. 89.3% of the sequence was obtained for both strands. 98.6% of the sequence of BamHI e in U_L was determined from both strands.

Difficulties in obtaining the sequence due to

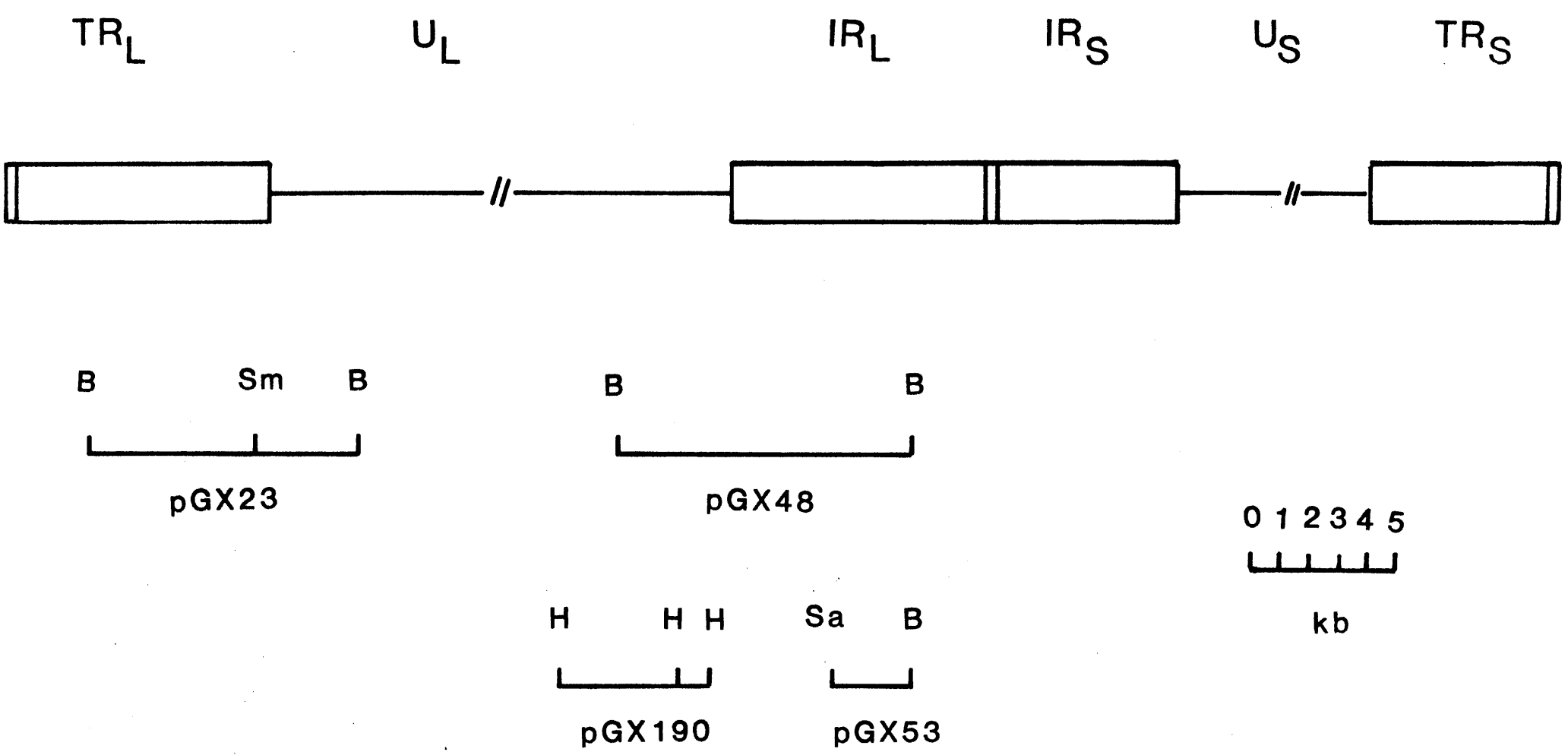

TRL
UL
IRL
IRS
US
TRS
B
Sm
B
pGX23
B
B
pGX48
0 1 2 3 4 5
kb
H
H H
Sa
B
pGX190
pGX53

Figure 3.1 Plasmids used in DNA sequencing

The upper part of the figure shows the prototype arrangement of the HSV-1 genome. The location of restriction fragments used for DNA sequence analysis are shown. Plasmids pGX23, BamHI e; pGX48, BamHI b; pGX190, HpaI s and v; pGX53, SalI-BamHI subfragment of BamHI b. B = BamHI, H = HpaI, Sa = SalI, Sm = SmaI.

1
2
3

Figure 3.2 Sequencing gels

The figure shows three clones sequenced as described in the text and run on 6% polyacrylamide, 9M urea, TBE gradient gels. The sequencing reactions in the four tracks are from left to right, T,C,G,A.

1. The first gel shows a sequence in which both the beginning and end of the DNA insert in the M13 clone are clearly visible. These are indicated by the arrowheads.
2. The sequence of the second M13 clone contains a reiteration set, which gives a characteristic ladder pattern of fragments.
3. The insert of the third M13 clone contains the DNA sequence spanning the TR_L/U_L junction. The arrowhead indicates the location of the junction.

Figure 3.3 BamHI b database

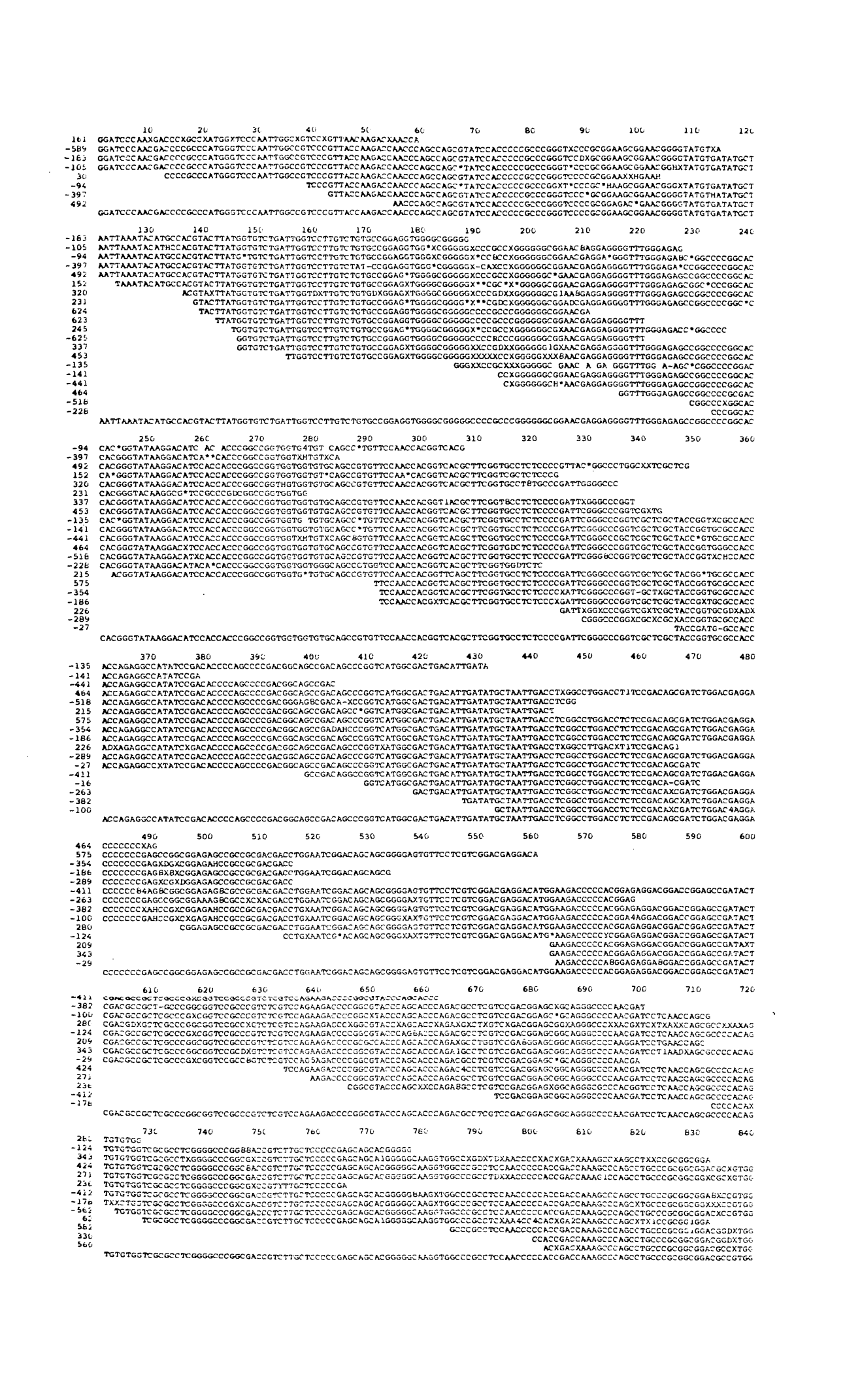
The compiled gel readings of recombinant M13 clones covering the BamHI b restriction fragment held in the database are given in the following pages. The sequence is orientated as in the prototype genome, and numbering starts at the left BamHI site.

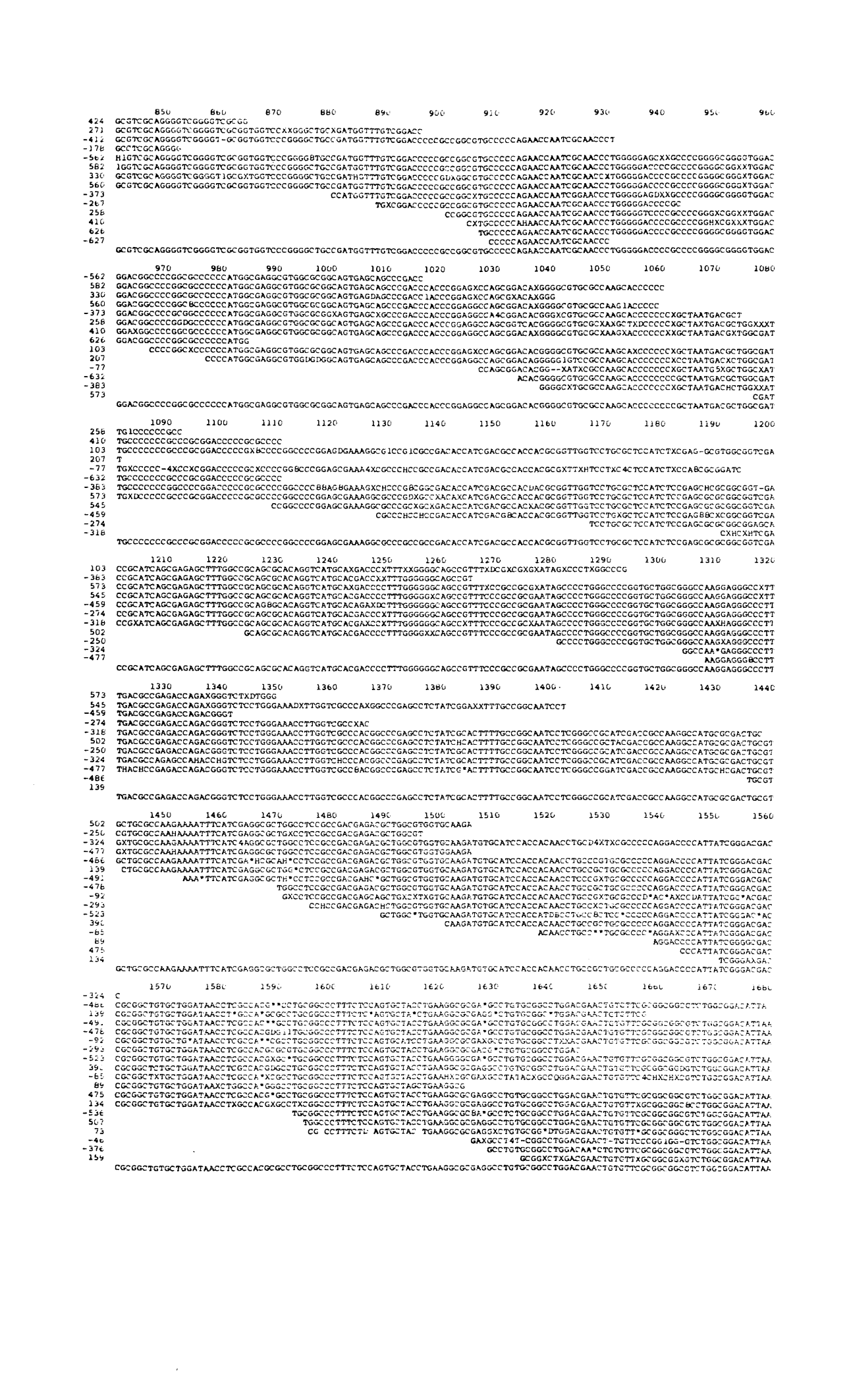
The columns of numbers on the left side of the sequence are the clone identity numbers (see figure 3.5). Clones with negative identity numbers contain sequences from the opposite DNA strand, and their complementary sequence is displayed. During the compilation of the data, padding characters are used to optimally align the sequences. These are seen as X, *, or as a blank. The consensus sequence derived from the amassed data is shown in the bottom line of the display.

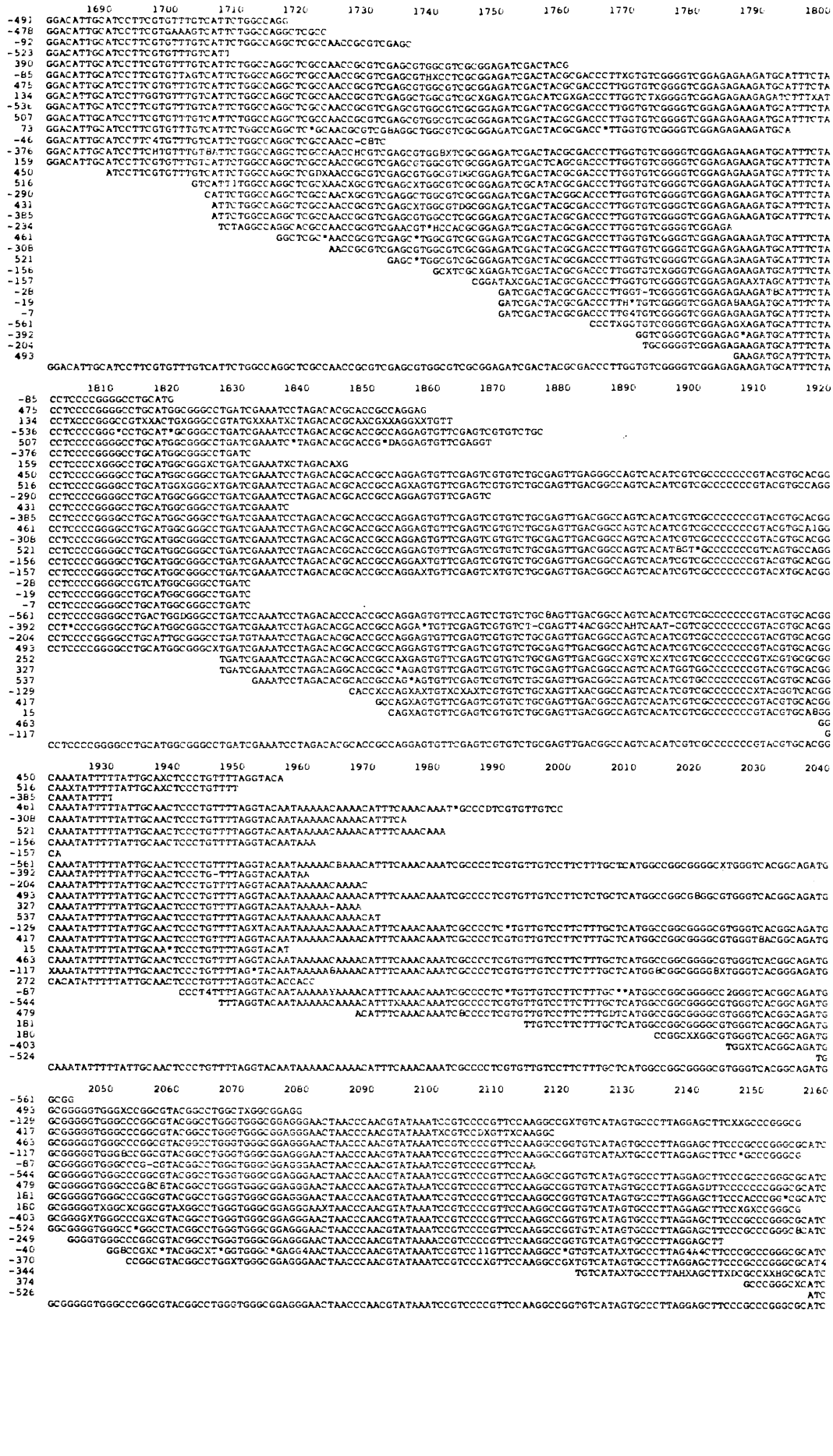
In the database the following "uncertainty" codes are used:

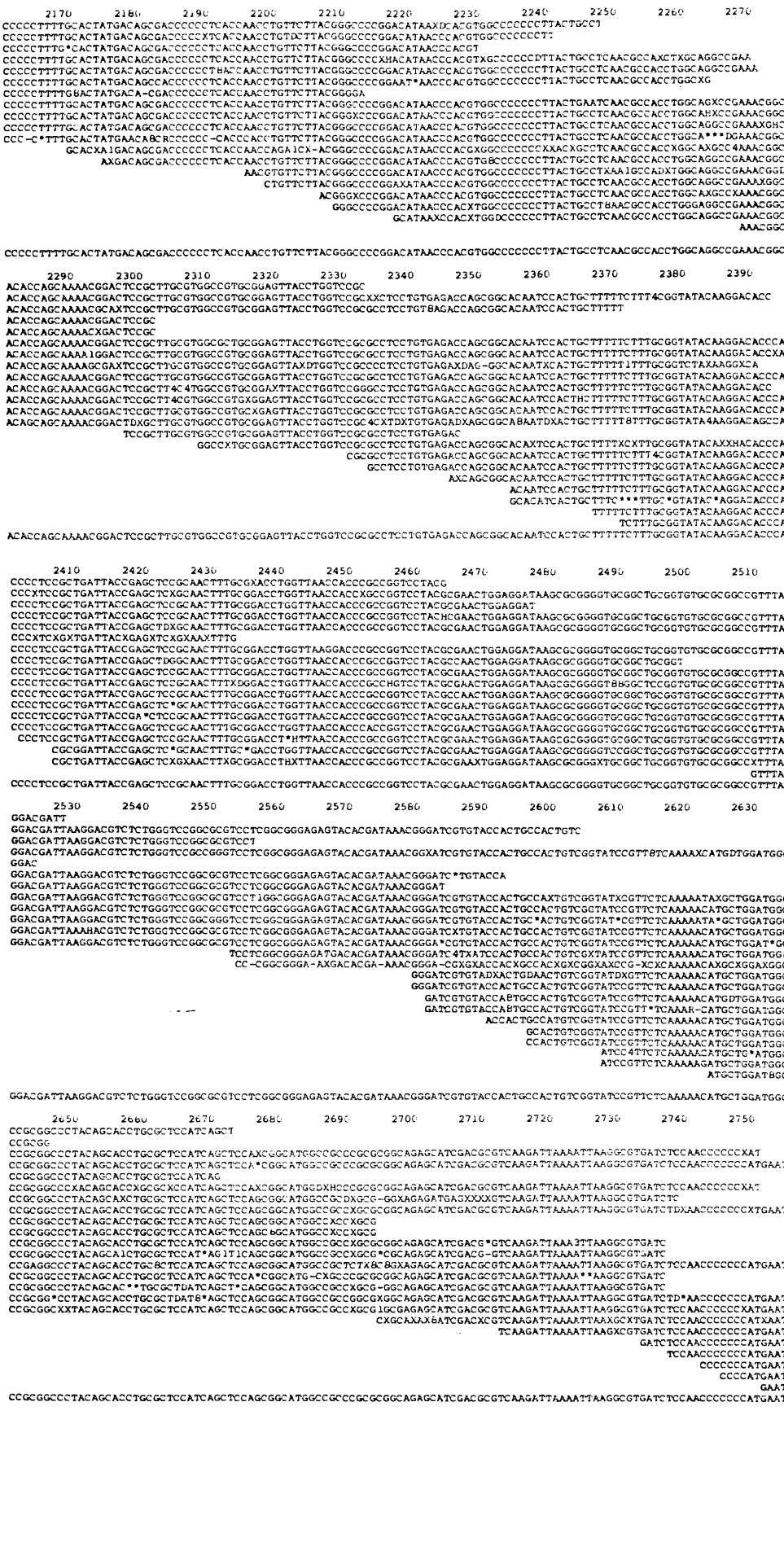
Symbol	Meaning
1	probably C
2	" T
3	" A
4	" G
D	Definitely C, possibly CC
V	" T, " TT
B	" A, " AA
H	" G, " GG
R	A or G
Y	C or T
5	A or C
6	G or T
7	A or T
8	G or C
-	A, C, G or T

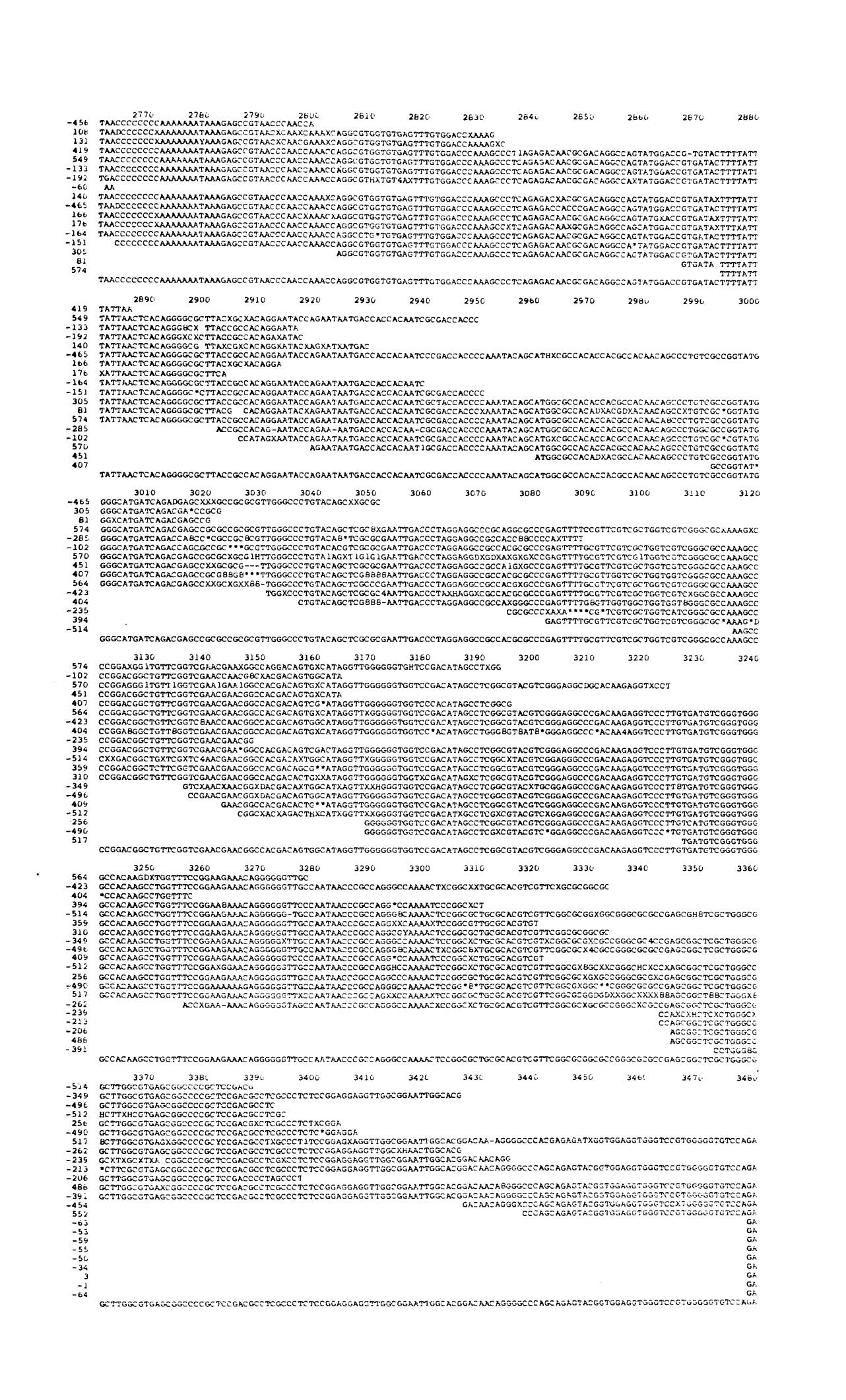
(Staden, 1980)

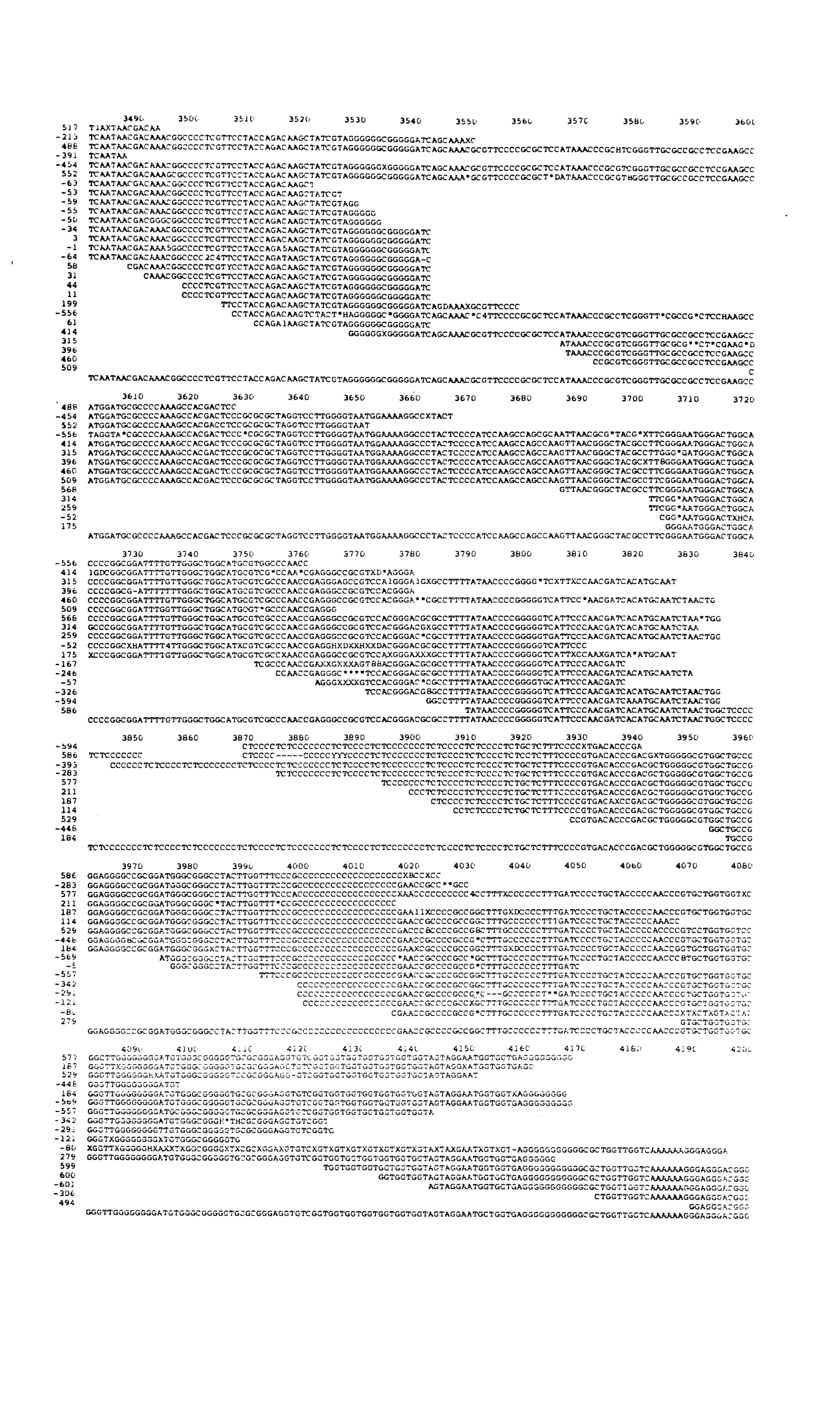


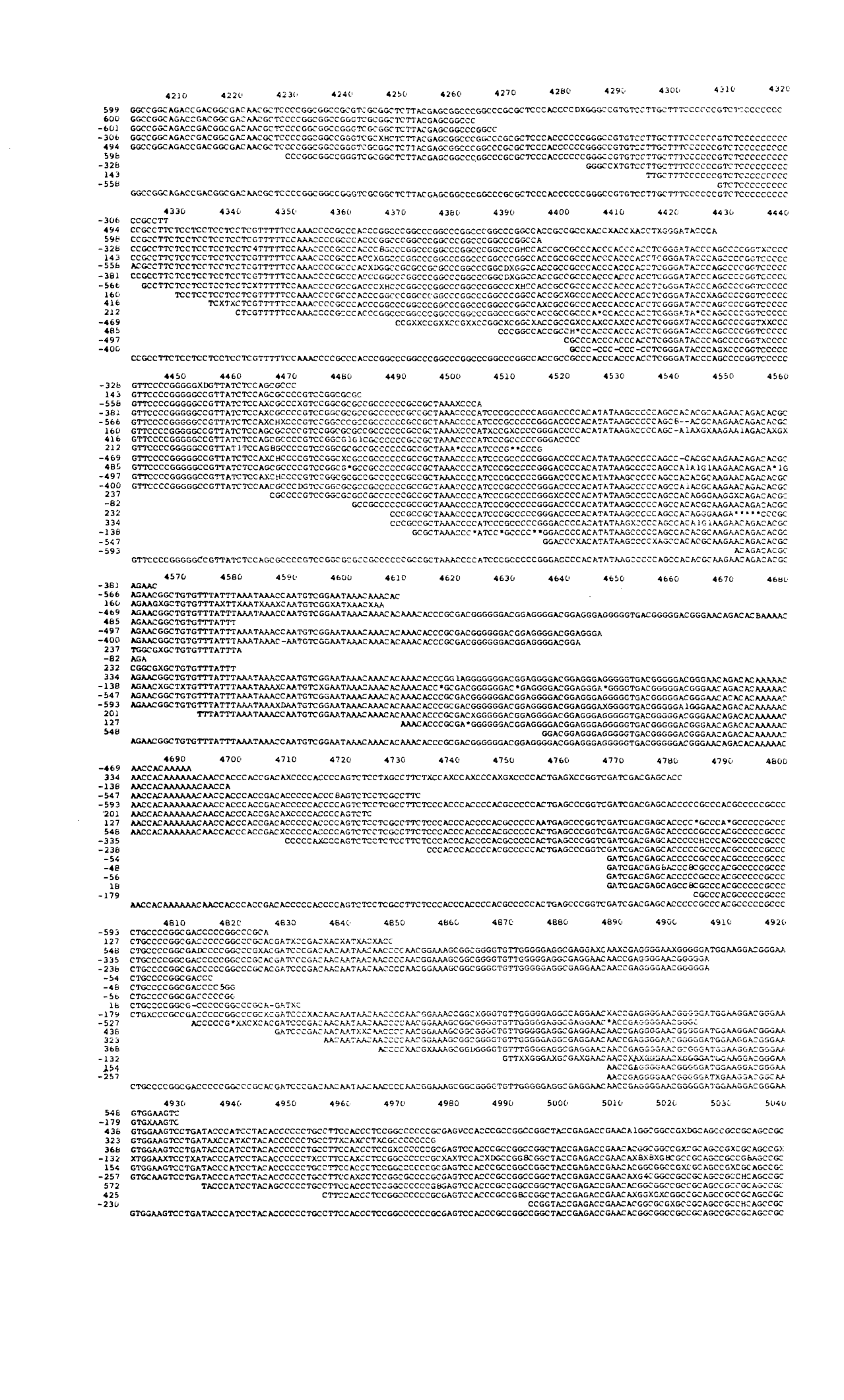


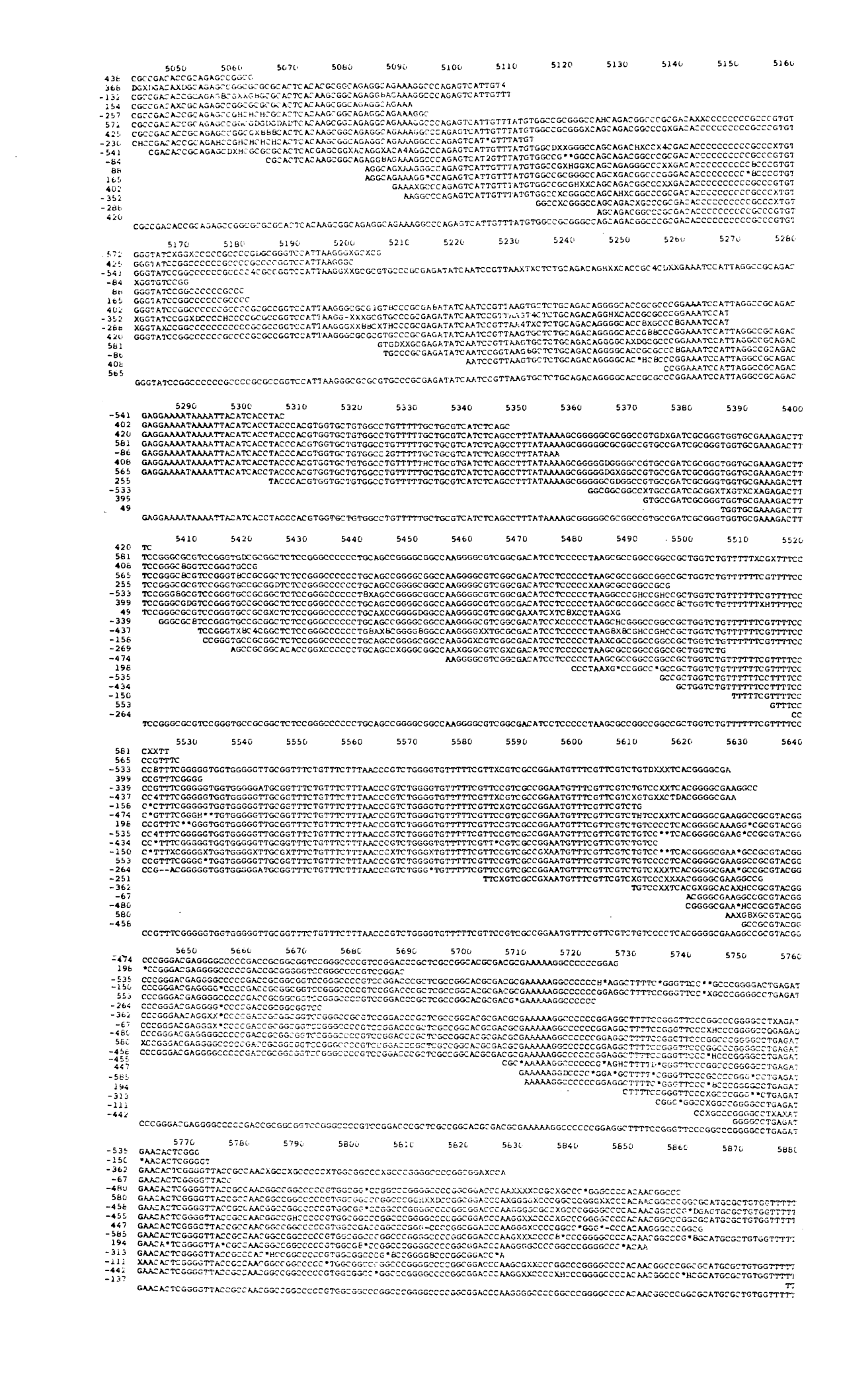


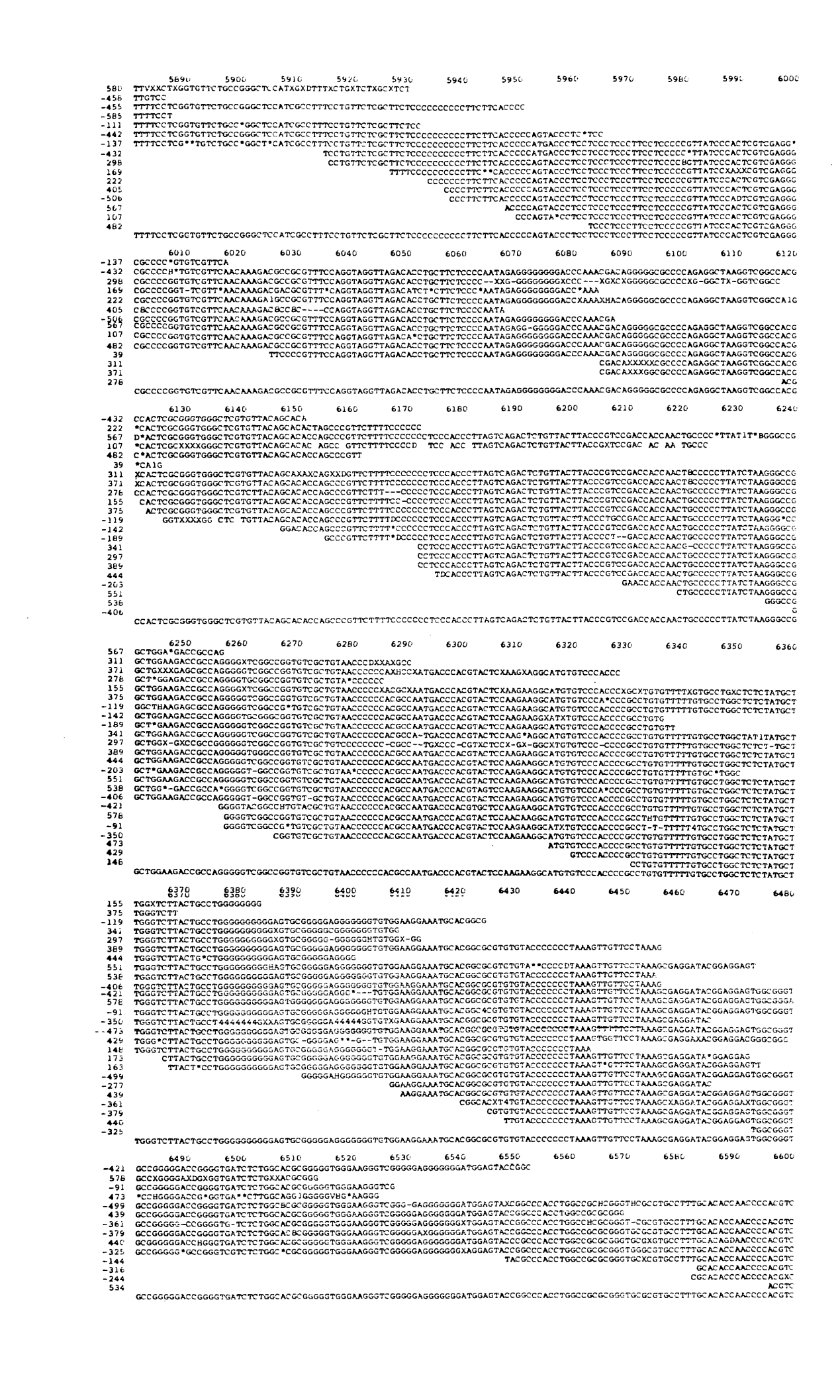


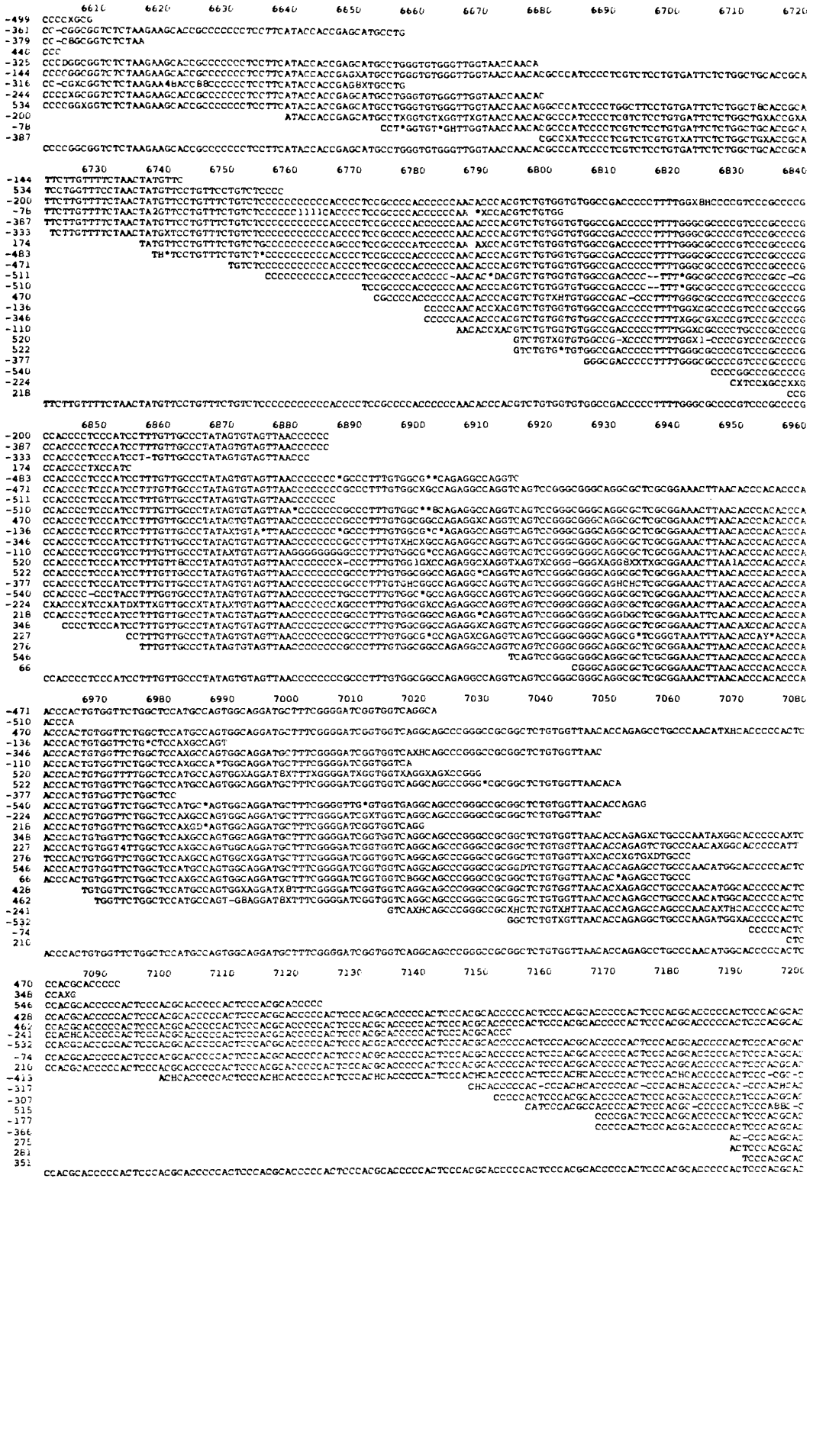


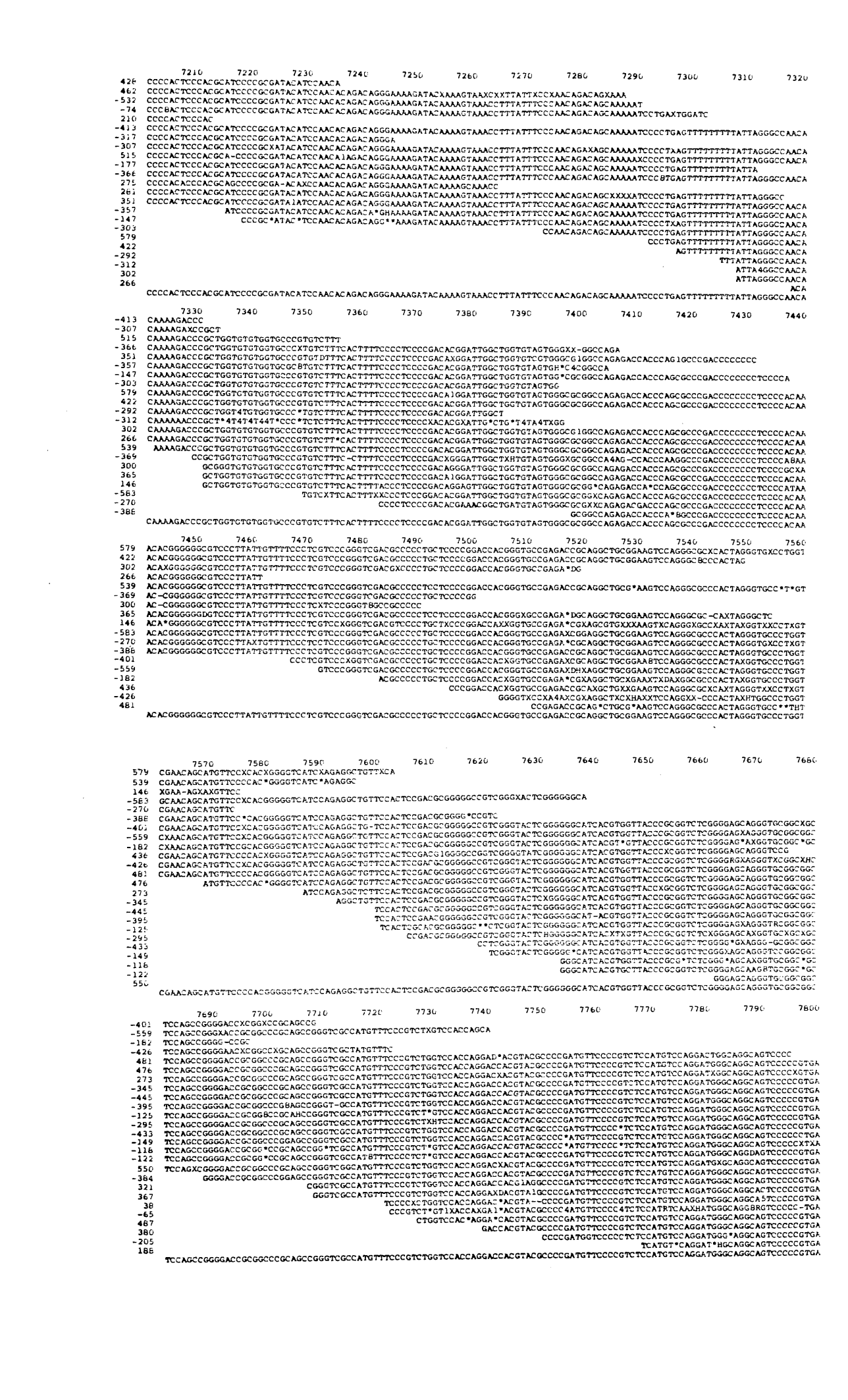


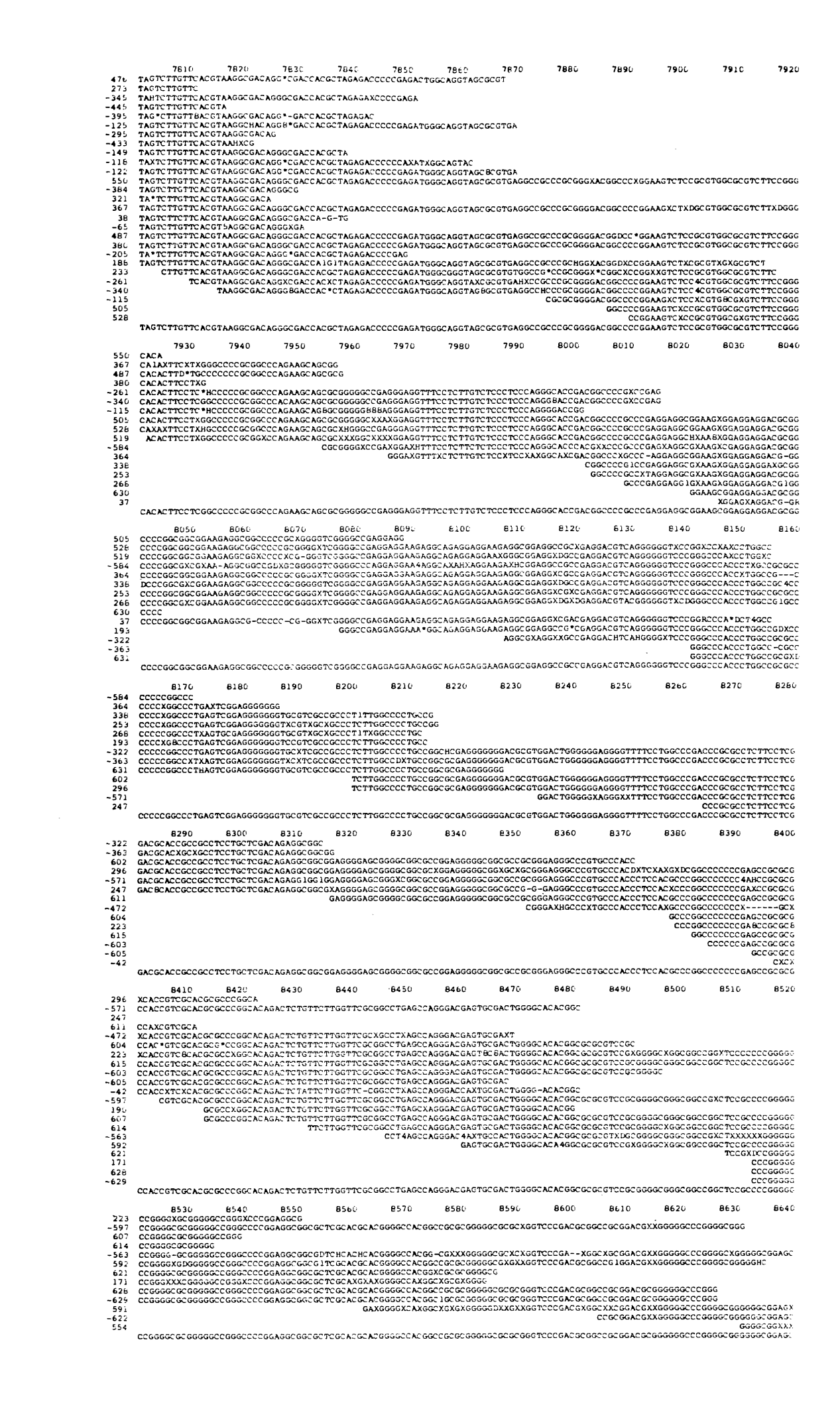


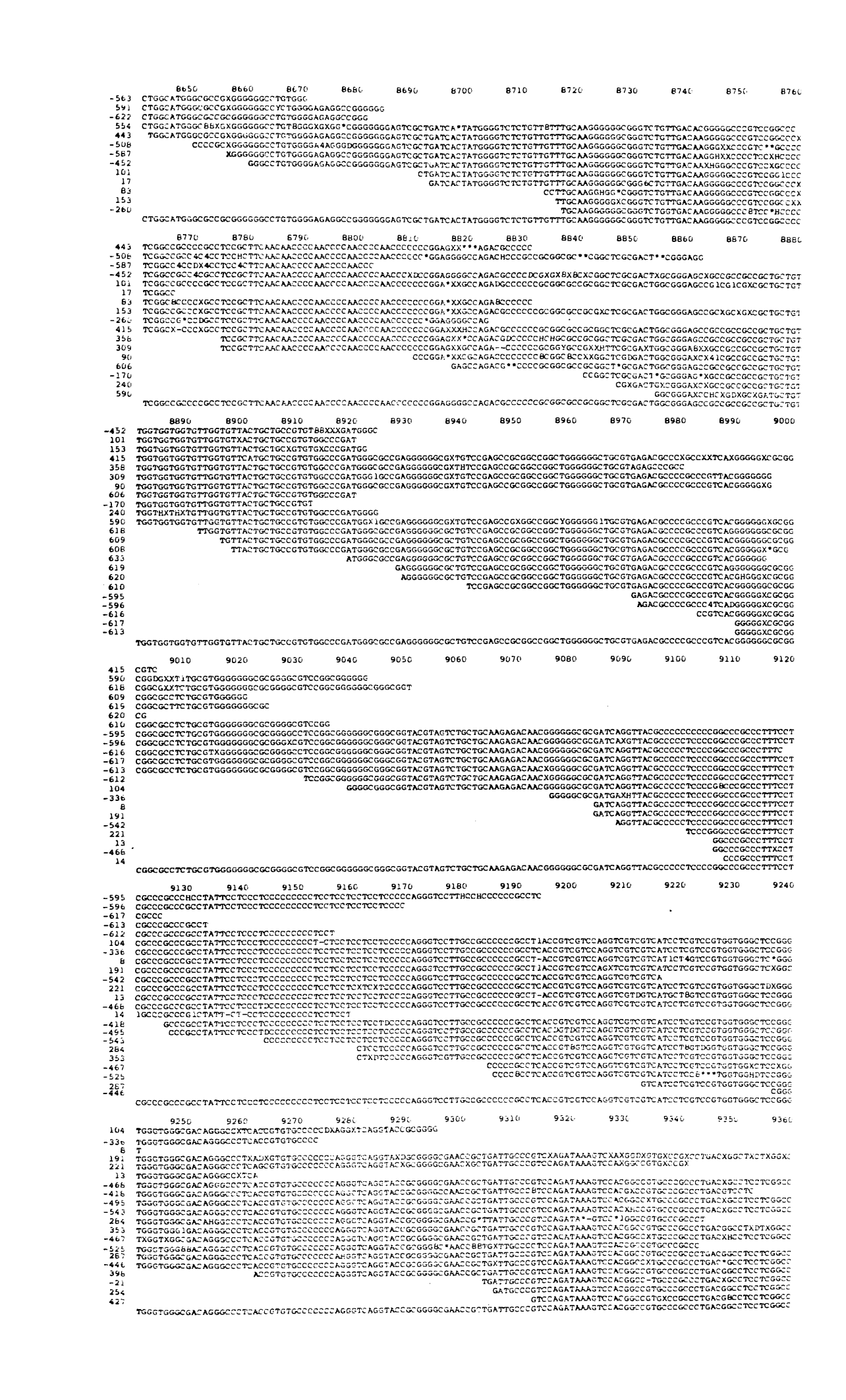












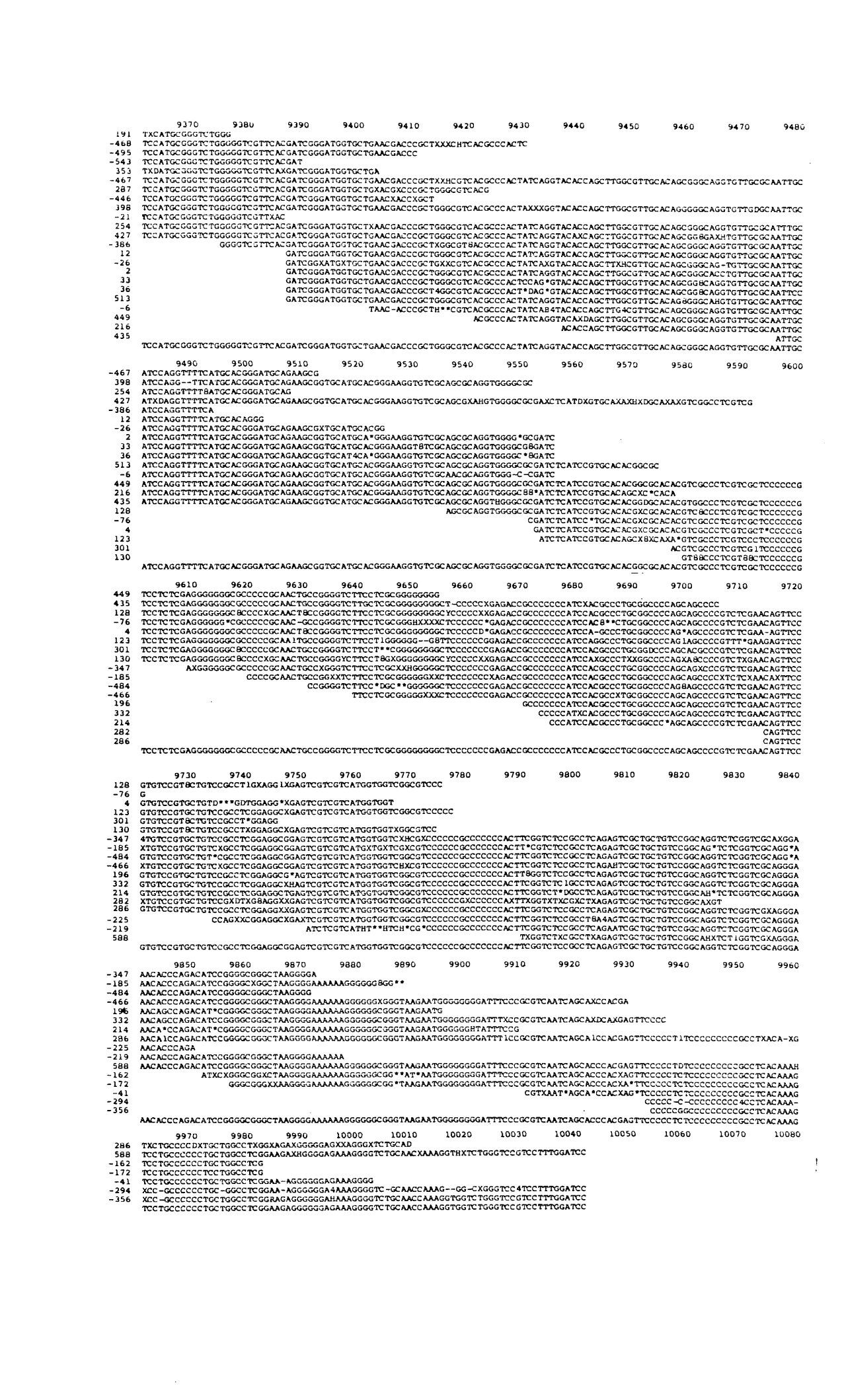


Figure 3.4 BamHI e database

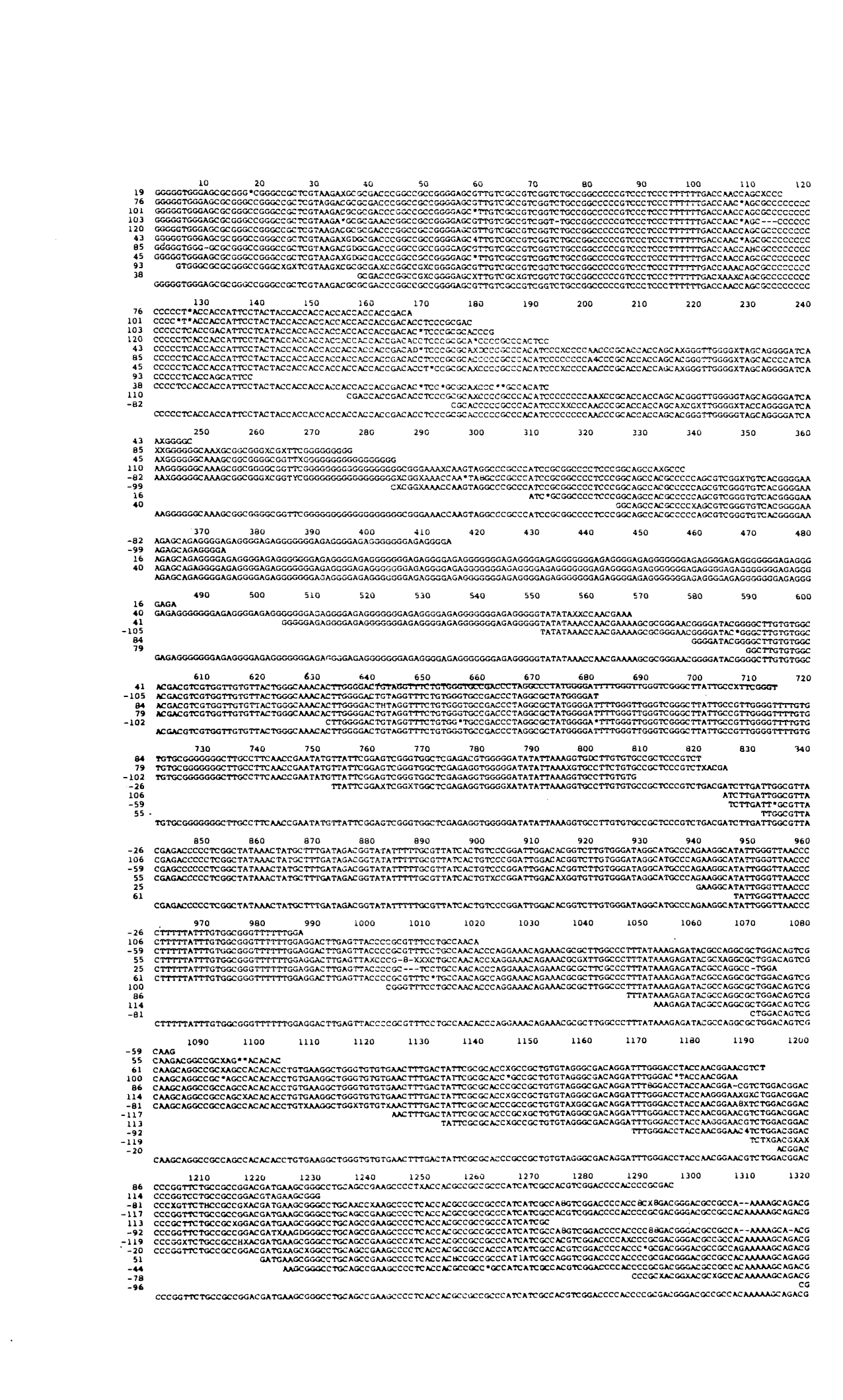
The compiled gel readings of recombinant clones covering the SmaI-BamHI subfragment of BamHI e held in the database are listed in the following pages. The sequence is orientated as in the prototype genome, and numbering starts at the SmaI site.

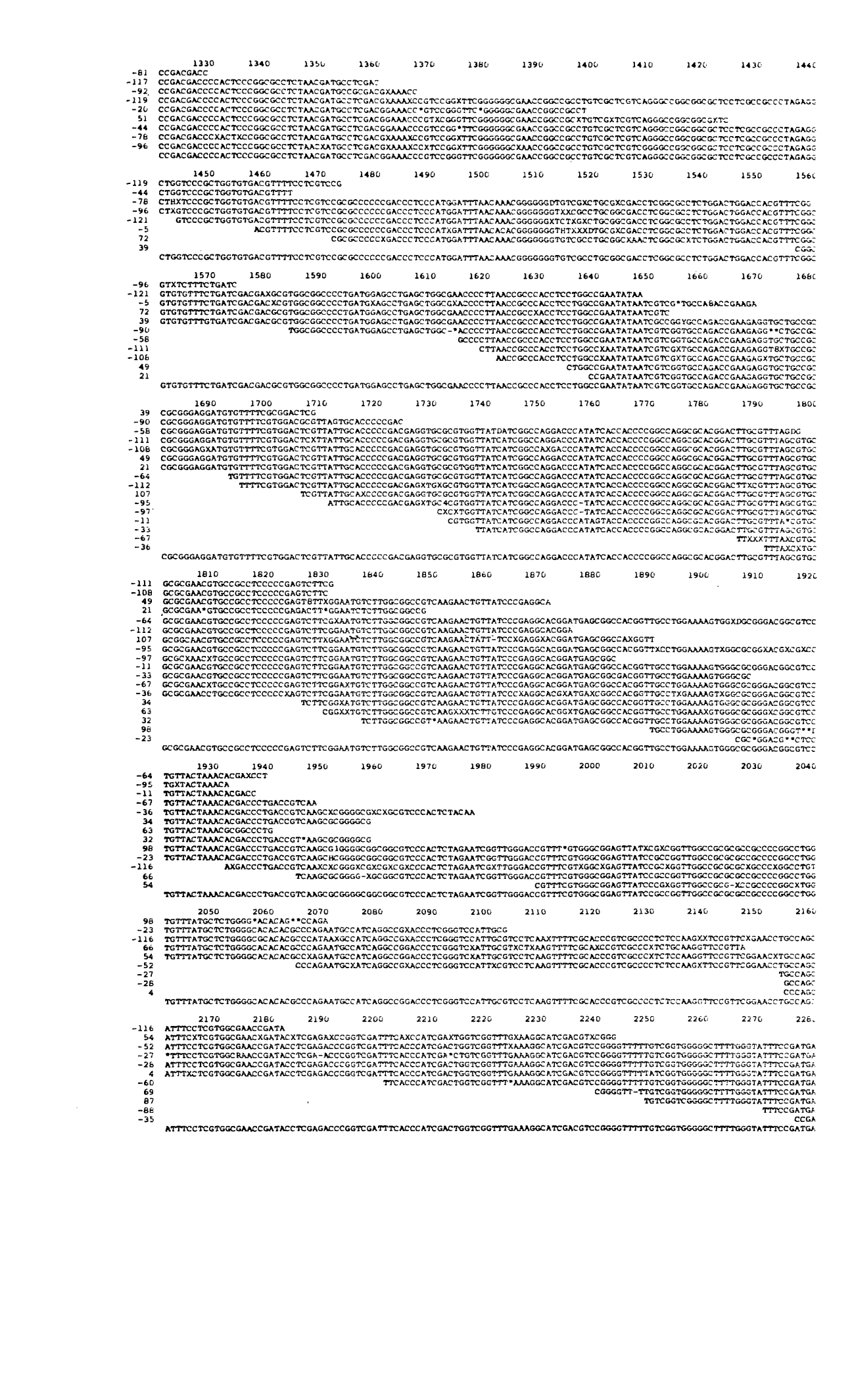
The numbers in the column to the left of the sequences, are the clone identity numbers (see figure 3.6). Clones with negative identity numbers contain sequences from the complementary DNA strand. The padding characters *, X or a blank, were used to align the gels. The bottom line in the display shows the consensus sequence derived.

In the database the following "uncertainty" codes are used:

Symbol	Meaning
1	probably C
2	" T
3	" A
4	" G
D	Definitely C, possibly CC
V	" T, " TT
B	" A, " AA
H	" G, " GG
R	A or G
Y	C or T
5	A or C
6	G or T
7	A or T
8	G or C
-	A, C, G or T

(Staden, 1980)





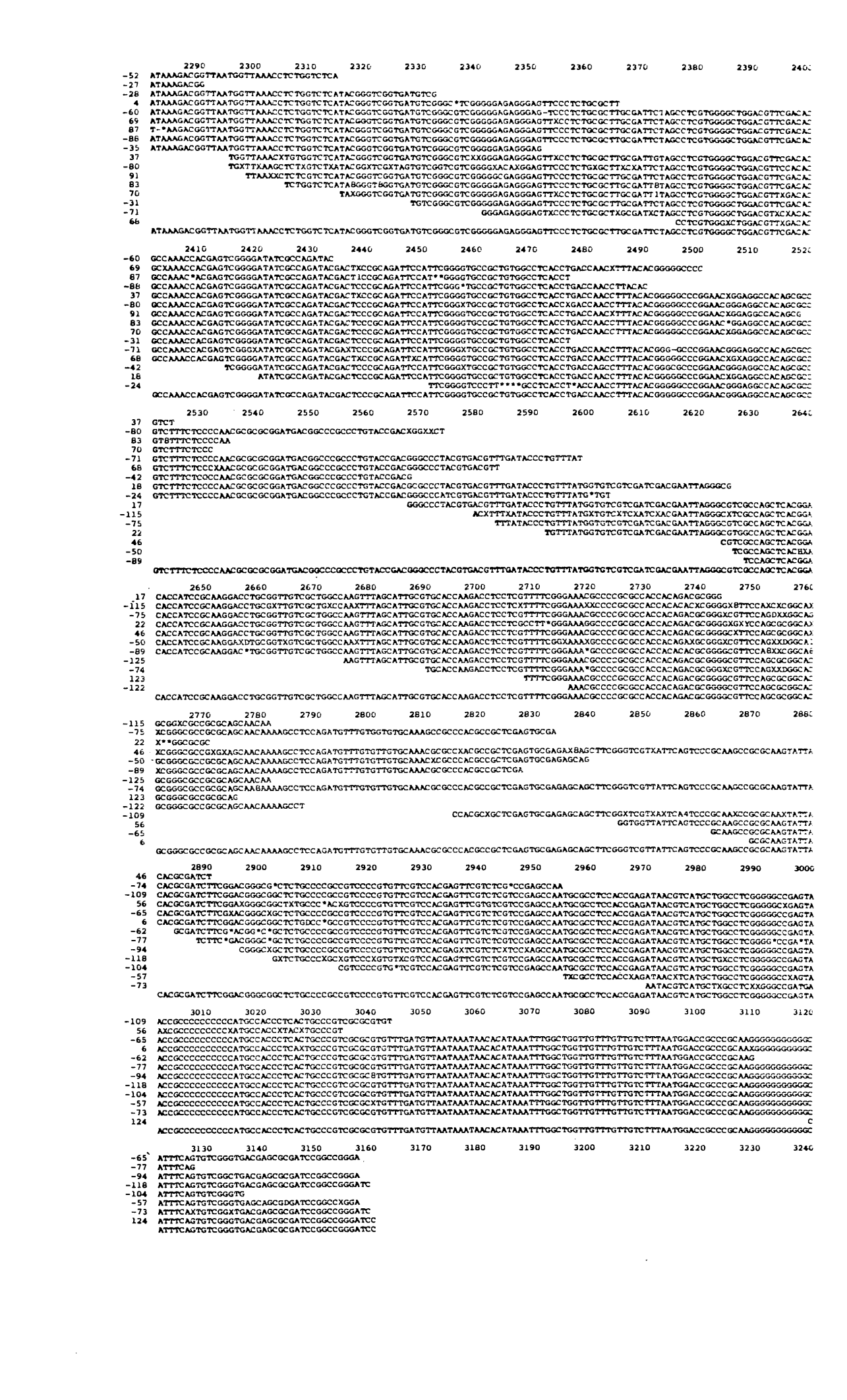


Figure 3.5 Recombinant M13 clones used to sequence BamHI b

A sorted list of recombinant M13 clones used to sequence BamHI b is given in the following pages. Alongside the clone name (given by the experimenter) is the identity number, allocated to individual sequences as they are entered into the database, as used in the previous figures. This listing also gives the position in the database where the clone starts, the orientation of the sequence, the length of the reading and the identity numbers of the clones sequenced to the left and to the right of this clone in the database.

BR655A.24	161	1	59	0	589
BR730.115	589	1	-114	161	183
BR730.31	183	1	-188	589	105
BR566.12	105	1	-227	183	30
BS170.A	30	13	91	105	94
BR556.11	94	39	-269	30	397
BR1316.84	397	43	-242	94	492
BR1428.101	492	55	294	397	152
BR675.25	152	124	201	492	320
BR1247.71	320	136	200	152	231
BR899.46	231	138	141	320	624
BR675.18A	624	139	72	231	623
BR1247.18A	623	142	79	624	245
BR940.48	245	145	91	623	625
BR566.18A	625	146	-75	245	337
BR940.75	337	146	193	625	453
BR1378.97	453	155	189	337	135
BR617.20	135	186	-246	453	141
BR631.22	141	194	-185	135	441
BR1367.96	441	195	-209	141	464
BR1397.98	464	216	275	441	518
BR1459.105	518	229	-219	464	228
BR892.46	228	233	-89	518	215
BR986.41	215	243	201	228	575
BR1523.112	575	291	270	215	354
BR1144.77	354	292	-224	575	186
BR731.31	186	292	-242	354	226
BR1025.45	226	325	142	186	289
BR1144.61	289	329	-187	226	27
BS131.A	27	346	-125	289	411
BR1318.85	411	398	-265	27	16
BS111.A	16	409	-62	411	263
BR1083.55	263	418	-161	16	382
BR1288.81	382	427	-274	263	100
BR5302.10	100	433	-280	382	280
BR1066.58	280	495	233	100	124
BR592.18	124	514	-264	280	209
BR950.38	209	563	149	124	343
BR950.76	343	563	270	209	29
BS169.A	29	564	-136	343	424
BR1332.89	424	634	231	29	271
BR1052.52	271	639	258	424	236
BR897.46	236	646	120	271	412
BR1319.85	412	673	-259	236	178
BR707.29	178	713	-140	412	562
BR1512.111	562	723	-294	178	62
BS209.A	62	728	105	562	582
BR1534.113	582	785	276	62	330
BR1277.74	330	800	241	582	560
BR1505.110	560	802	258	330	373
BR1281.81	373	880	-195	560	267
BR1102.57	267	888	-56	373	258
BR1058.53	258	901	222	267	410
BR1317.85	410	905	206	258	626
BR1058.19A	626	907	78	410	627
BR1319.19A	627	909	-22	626	103
BR564.12	103	967	327	627	207
BR952.38	207	977	105	103	77
BR510.1	77	1026	-169	207	632
CP26.47	632	1034	-77	77	383
BR1289.81	383	1038	-228	632	573
BR1521.112	573	1077	273	383	545
BR1491.108	545	1109	294	573	459
BR1401.98	459	1128	-214	545	274
BR1092.56	274	1167	-200	459	318
BR1241.71	318	1191	-248	274	502
BR1438.102	502	1224	282	318	250
BR1033.50	250	1281	-216	502	324
BR1272.74	324	1304	-258	250	477
BR1418.100	477	1308	-198	324	486
BR1443.103	486	1436	-244	477	139
BR625.21	139	1439	219	486	491
BR1425.101	491	1453	-265	139	478
BR1419.100	478	1470	-254	491	92
BR541.11	92	1471	-266	478	293
BR1183.65	293	1476	-170	92	523
BR1476.107	523	1489	-220	293	390
BR1305.83	390	1501	260	523	85
BR523.3	85	1518	-302	390	89
BR853.8	89	1539	104	85	475
BR1416.100	475	1544	315	89	134
BR613.20	134	1551	312	475	536
BR1485.108	536	1593	-284	134	507
BR1449.104	507	1595	274	536	73
BR502.1	73	1596	199	507	46
BS187.A	46	1626	-107	73	376
BR1290.82	376	1628	-204	46	159
BR667.25	159	1635	212	376	450
BR1368.96	450	1690	267	159	516
BR1454.104	516	1703	247	450	290
BR1056B.52	290	1705	-164	516	431
BR1343.94	431	1706	132	290	385
BR1293.82	385	1706	-225	431	234
BR919.47	234	1707	-79	385	461
BR1409.99	461	1716	284	234	306
BR1213.68	306	1724	-252	461	521
BR1467.106	521	1733	249	306	156
BR658.24	156	1740	-222	521	157
BR638.24	157	1746	-177	156	28
BS168.A	28	1750	-84	157	19
BS144B.A	19	1750	-84	28	7
BS164.A	7	1750	-84	19	561
BR1509.111	561	1764	-281	7	392
BR1308.84	392	1771	-190	561	204
BR955.38	204	1772	-198	392	493
BR1412.100	493	1786	294	204	252
BR941.48	252	1827	94	493	327
BR941.75	327	1827	142	252	537
BR1487.108	537	1832	140	327	129
BR587.18	129	1847	-310	537	417
BR1337.87	417	1851	266	129	15
BS105.A	15	1853	105	417	463
BR1396.98	463	1919	330	15	117
BR587.15	117	1920	-237	463	272
BR1054.52	272	1921	41	117	87
BR544.9	87	1941	-175	272	544
BR1489.108	544	1947	-295	87	479
BR1434.102	479	1968	263	544	181
BR711.30	181	1994	277	479	180
BR710.29	180	2014	143	181	403
BR1312.84	403	2024	-248	180	524
BR1478.107	524	2039	-227	403	249
BR930.49	249	2044	-101	524	40
PBBR119.A	40	2050	-164	249	370
BR1271.79	370	2053	-281	40	344
BR953.76	344	2122	-272	370	374
BR1283.81	374	2148	224	344	526
BR1480.107	526	2158	-145	374	319
BR1242.71	319	2170	-133	526	501
BR1447.103	501	2175	290	319	576
BR1524.112	576	2196	333	501	329
BR1276.74	329	2199	193	576	500
BR1446.103	500	2207	-271	329	208
BR953.38	208	2209	-185	500	331
BR860.74	331	2218	-302	208	498
BR1440.103	498	2269	-248	331	68
BR114.A	68	2277	157	498	530
BR1469.106	530	2298	-50	68	503
BR1426.101	503	2309	-296	530	197
BR866.35	197	2331	-169	503	372
BR1282.81	372	2334	223	197	531
BR1470.106	531	2346	290	372	195
BR860.34	195	2355	-170	531	489
BR1421.101	489	2355	-239	195	378
BR1295.82	378	2367	-218	489	112
BR572.13	112	2371	266	378	457
BR1399.98	457	2402	-272	112	145
BR627.21	145	2407	240	457	360
BR1184.78	360	2407	-345	145	456
BR1398.98	456	2510	-290	360	72
BA107.2	72	2553	-119	456	299
BR1184.65	299	2554	-198	72	71
BR124.A	71	2576	165	299	108
BR571.13	108	2580	254	71	24
BS151.A	24	2582	113	108	25
BS154.A	25	2582	115	24	32
BS173.A	32	2591	147	25	47
BS188.A	47	2597	141	32	131
BR601.19	131	2597	258	47	10
BS167.A	10	2608	-130	131	43
BS184.A	43	2608	130	10	419
BR1327.89	419	2624	263	43	549
BR1498.109	549	2639	311	419	133
BR609.19	133	2694	-223	549	192
BR750.33	192	2713	-205	133	60
BS207.A	60	2734	-30	192	140
BR624.21	140	2738	223	60	465
BR1402.98	465	2743	-309	140	166
BR700.27	166	2746	169	465	176
BR697.28	176	2752	152	166	164
BR703.27	164	2761	-179	176	151
BR673.25	151	2764	-187	164	305
BR1218.69	305	2804	224	151	81
BR538.4	81	2867	154	305	574
BR1522.112	574	2873	321	81	285
BR1140.61	285	2902	-187	574	102
BR543.11	102	2904	-261	285	570
BR1518.112	570	2919	306	102	451
BR1369.96	451	2960	205	570	407
BR1324.85	407	2992	204	451	564
BR1515.111	564	3001	278	407	423
BR1330.89	423	3031	-303	564	404
BR1320.85	404	3037	221	423	235
BR920.47	235	3075	-74	404	394
BR1311.84	394	3075	236	235	514
BR1451.104	514	3116	-273	394	359
BR1224.79	359	3121	200	514	310
BR1224.69	310	3121	213	359	349
BR1070.76	349	3136	-291	310	496
BR1435.102	496	3137	-256	349	409
BR1328.86	409	3143	179	496	512
BR1423.104	512	3146	-249	409	256
BR1038.50	256	3169	236	512	490
BR1422.101	490	3169	-239	256	517
BR1456.105	517	3227	267	490	262
BR1070.54	262	3256	-171	517	239
BR916.47	239	3343	-93	262	213
BR982.41	213	3343	-208	239	206
BR959.39	206	3345	-53	213	488
BR1445.103	488	3345	283	206	391
BR1307.83	391	3353	-135	488	454
BR1380.97	454	3427	-240	391	552
BR1503.110	552	3438	214	454	63
BS212.A	63	3479	-43	552	53
BS196.A	53	3479	-48	63	59
BS203.A	59	3479	-51	53	55
BS200.A	55	3479	-54	59	50
BS193.A	50	3479	-55	55	34
BS175.A	34	3479	-64	50	3
BS157.A	3	3479	64	34	1
BS129.A	1	3479	-64	3	64
BS112A.A	64	3479	-64	1	58
BS202.A	58	3488	55	64	31
BS172.A	31	3491	52	58	44
BS185A.A	44	3498	45	31	11
BS101A.A	11	3498	45	44	199
BR863.34	199	3505	54	11	556
BR1508.110	556	3507	-254	199	61
BS208.A	61	3511	32	556	414
BR1338.87	414	3528	252	61	315
BR1244.71	315	3566	261	414	396
BR1315.84	396	3567	213	315	460
BR1407.99	460	3572	262	396	509
BR1458.105	509	3600	166	460	568
BR1529.113	568	3686	150	509	314
BR1246.71	314	3701	130	568	259
BR1061	259	3701	135	314	52
BS195.A	52	3704	-107	259	175
BR683.28	175	3704	123	52	167
BR031.27	167	3751	-67	175	246
BR946.48	246	3755	-75	167	57

BS201B.A	57	3762	-56	246	326
BR946.75	326	3771	-70	57	594
BR1333.114	594	3782	-159	326	586
BR1061.115	586	3789	236	594	393
BR1310.84	393	3845	-116	586	283
BR1133.60	283	3875	-155	393	577
BR1527.113	577	3894	275	283	211
BR973.40	211	3899	118	577	187
BR732.31	187	3903	259	211	114
BR580.14	114	3907	161	187	529
BR1468.106	529	3928	224	114	448
BR1381.97	448	3953	-145	529	184
BR714.30	184	3956	212	448	569
BR1532.113	569	3974	-195	184	5
BS160.A	5	3976	-74	569	557
BR1510.111	557	3992	-152	5	342
BR1159.77	342	3999	-126	557	291
BR1159.63	291	3999	-127	342	121
BR552.6	121	4000	-127	291	80
BR505.3	80	4016	-181	121	279
BR1119.59	279	4068	98	80	599
BR1119.134	599	4124	196	279	600
BR1119.137	600	4134	130	599	601
BR1220.137	601	4143	-125	600	306
BR1220.69	306	4173	-155	601	494
BR1415.100	494	4190	238	306	598
BR1415.134	598	4229	167	494	328
BR1273.74	328	4284	-187	598	143
BR635.22	143	4295	188	328	558
BR1500.110	558	4308	-197	143	381
BR1287.81	381	4321	-245	558	566
BR1517.111	566	4323	-287	381	160
BR674.25	160	4329	278	566	416
BR1336.87	416	4335	188	160	212
BR976.40	212	4340	176	416	469
BR1406.99	469	4369	-323	212	485
BR1439.103	485	4388	192	469	497
BR1431.102	497	4400	-247	485	400
BR1304.83	400	4401	-242	497	237
BR901.46	237	4466	115	400	82
BR547.6	82	4481	-83	237	232
BR905.46	232	4488	92	82	334
BR905.75	334	4488	294	232	138
BR622.21	138	4492	-207	334	547
BR1493.108	547	4516	-218	138	593
BR915.114	593	4551	-275	547	201
BR873.35	201	4573	152	593	127
BR611.19	127	4610	239	201	548
BR1496.109	548	4636	294	127	335
BR915.75	335	4709	-198	548	238
BR915.47	238	4735	-172	335	54
BS197.A	54	4768	-48	238	48
BS190.A	48	4768	-52	54	56
BS201A.A	56	4768	-52	48	18
BS134.A	18	4768	64	56	179
BR709.29	179	4784	-146	18	527
BR1481.107	527	4812	-93	179	438
BR1352.95	438	4827	239	527	323
BR1265.73	323	4836	141	438	368
BR1265.79	368	4846	263	323	132
BR607.19	132	4868	-242	368	154
BR641.23	154	4888	204	132	257
BR1039.50	257	4888	-207	154	572
BR1520.112	572	4934	272	257	425
BR1335.89	425	4956	245	572	230
BR895.46	230	4994	-120	425	541
BR1464.108	541	5044	-263	230	84
BR551.6	84	5065	-107	541	88
BR852.8	88	5083	98	84	165
BR852.27	165	5083	99	88	402
BR1309.84	402	5088	260	165	352
BR1135.77	352	5090	-179	402	288
BR1135.61	288	5114	-155	352	420
BR1329.89	420	5125	278	288	581
BR1533.113	581	5205	321	420	86
BR542.8	86	5206	-151	581	408
BR1325.85	408	5221	201	86	565
BR1516.111	565	5257	271	408	255
BR936.50	255	5304	199	565	533
BR1474.107	533	5362	-266	255	399
BR1303.83	399	5372	160	533	49
BS192.A	49	5386	106	399	339
BR937.75	339	5404	-229	49	437
BR1348.95	437	5411	-218	339	158
BR666.25	158	5412	-200	437	269
BR932A.52	269	5417	-90	158	474
BR1414.100	474	5456	-272	269	198
BR867.35	198	5479	210	474	535
BR1484.108	535	5495	-277	198	434
BR1354.95	434	5498	-118	535	150
BR664.25	150	5508	-266	434	553
BR1504.110	553	5515	209	150	264
BR932.52	264	5519	-155	553	251
BR937.48	251	5583	-51	264	362
BR1236.78	362	5610	-218	251	67
BR102.A	67	5620	-158	362	480
BR1436.102	480	5620	-240	67	580
BR1531.113	580	5627	304	480	458
BR1400.98	458	5630	-257	580	455
BR1395.98	455	5707	-245	458	447
BR1379.97	447	5710	154	455	585
BR1371.115	585	5711	-177	447	194
BR858.34	194	5729	125	585	313
BR1236.70	313	5735	-93	194	111
BR508.10	111	5742	-191	313	442
BR1370.96	442	5749	-217	111	137
BR621.21	137	5879	-139	442	432
BR1345.94	432	5915	-237	137	298
BR1172.64	298	5916	202	432	169
BR679.28	169	5927	158	298	222
BR971.40	222	5934	239	169	405
BR1322.85	405	5937	131	222	506
BR1448.104	506	5938	-150	405	567
BR1525.112	567	5948	109	506	107
BR570.13	107	5950	275	567	482
BR1430.102	482	5963	199	107	39
PBBR117.A	39	6025	101	482	311
BR1225.69	311	6085	206	39	371
BR1225.79	371	6085	244	311	278
BR1087.58	278	6118	168	371	155
BR644.23	155	6122	262	278	375
BR1286.81	375	6123	246	155	119
BR579.14	119	6126	-299	375	142
BR634.22	142	6147	-190	119	189
BR745.33	189	6155	-184	142	341
BR1155.77	341	6172	237	189	297
BR1155.62	297	6172	241	341	389
BR1301.83	389	6172	285	297	444
BR1375.97	444	6175	226	389	203
BR964.39	203	6209	-142	444	551
BR1501.110	551	6219	255	203	538
BR1490.108	538	6235	221	551	406
BR1223.85	406	6240	-217	538	421
BR1331.89	421	6256	-297	406	578
BR1528.113	578	6257	257	421	91
BR5301.10	91	6257	-271	578	350
BR1076.77	350	6266	-200	91	473
BR1413.100	473	6316	209	350	429
BR1340.94	429	6320	160	473	148
BR653.24	148	6331	113	429	173
BR700.29	173	6366	107	148	163
BR702.27	163	6367	108	173	499
BR1442.103	499	6391	-218	163	277
BR1076.58	277	6407	-59	499	439
BR1364.96	439	6409	161	277	361
BR1233.78	361	6420	-238	439	379
BR1284.81	379	6425	-192	361	440
BR1365.96	440	6428	176	379	325
BR939.75	325	6473	-206	440	144
BR636.22	144	6548	-195	325	316
BR1233.70	316	6561	-97	144	244
BR939.48	244	6582	-98	316	534
BR1477.107	534	6596	163	244	200
BR865.35	200	6639	-247	534	78
BR514.1	78	6654	-151	200	387
BR1296.82	387	6679	-207	78	333
BR865.74	333	6722	-161	387	174
BR705.29	174	6736	119	333	483
BR1432.102	483	6738	-178	174	471
BR1410.99	471	6750	-273	483	511
BR1462.105	511	6756	-131	471	510
BR1461.105	510	6771	-195	511	470
BR1408.99	470	6773	320	510	136
BR619.21	136	6781	-209	470	346
BR994.76	346	6781	-268	136	110
BR575.13	110	6786	-233	346	520
BR1463.105	520	6795	235	110	522
BR1475.107	522	6795	257	520	377
BR1292.82	377	6806	-176	522	540
BR1495.108	540	6826	-230	377	224
BR994.42	224	6829	-220	540	218
BR999.42	218	6838	187	224	348
BR1017.76	348	6844	242	218	227
BR1017.44	227	6854	226	348	276
BR1105.57	276	6856	207	227	546
BR1492.108	546	6914	211	276	66
BR101	66	6924	139	546	428
BR1384.93	428	6967	269	66	462
BR1394.98	462	6969	319	428	241
BR906.46	241	7015	-140	462	532
BR1471.106	532	7034	-257	241	74
BR504.1	74	7072	-231	532	210
BR967.39	210	7078	135	74	413
BR1321.85	413	7099	-232	210	317
BR1235.70	317	7148	-98	413	307
BR1174.68	307	7152	-182	317	515
BR1452.104	515	7157	199	307	177
BR703.29	177	7168	-144	515	366
BR1229.78	366	7168	-239	177	275
BR1101.57	275	7189	76	366	281
BR1101.58	281	7189	128	275	351
BR1001.77	351	7191	241	281	357
BR1178.78	357	7215	-190	351	147
BR642.23	147	7218	-220	357	303
BR1178.68	303	7273	-129	147	579
BR1530.113	579	7292	312	303	422
BR1333.89	422	7297	253	579	292
BR1174.64	292	7305	-81	422	312
BR1229.70	312	7308	-95	292	302
BR1195.66	302	7308	212	312	266
BR1096.56	266	7318	144	302	539
BR1494.108	539	7322	275	266	369
BR1270.79	369	7329	-172	539	300
BR1187.65	300	7331	160	369	365
BR1187.78	365	7331	224	300	146
BR639.23	146	7331	248	365	583
BR1535.113	583	7349	-290	146	270
BR935.52	270	7363	-212	583	388
BR1299.83	388	7398	-225	270	401
BR1306.83	401	7467	-242	388	559
BR1502.110	559	7472	-269	401	182
BR716.30	182	7483	-214	559	436
BR1346.94	436	7496	180	182	426
BR1355.92	426	7505	-218	436	481
BR1437.102	481	7511	285	426	476
BR1417.100	476	7569	299	481	273
BR1091.56	273	7587	225	476	345
BR958.76	345	7593	-259	273	445
BR1376.97	445	7600	-217	345	395
BR1313.84	395	7600	-244	445	125
BR594.18	125	7601	-270	395	295
BR1149.62	295	7606	-220	125	433
BR1353.95	433	7619	-203	295	149
BR663.25	149	7619	-220	433	118
BR594.15	118	7634	-228	149	122
BR594.16	122	7634	-237	118	550
BR1499.109	550	7662	263	122	384
BR1291.82	384	7688	-142	550	321
BR1264.73	321	7707	126	384	367
BR1264.79	367	7708	249	321	38
PBBR109.A	38	7721	118	367	65
PBBR116.A	65	7722	-109	38	487

BR1444.103	487	7727	231	65	380
BR1285.81	380	7739	194	487	205
BR958.38	205	7750	-101	380	188
BR733.32	188	7768	147	205	233
BR917.47	233	7805	116	188	261
BR1069.54	261	7810	-206	233	340
BR1069.76	340	7815	-201	261	115
BR584.14	115	7875	-127	340	505
BR1372.96A	505	7886	203	115	528
BR1465.106	528	7890	266	505	519
BR1460.105	519	7922	234	528	584
BR1526.112	584	7954	-217	519	364
BR1263.79	364	7966	221	584	338
BR931.75	338	8002	213	364	253
BR922.49	253	8003	213	338	268
BR931.52	268	8009	204	253	630
BR1922.120	630	8021	24	268	37
PBBR103.A	37	8026	130	630	193
BR857.34	193	8077	137	37	322
BR1255.72	322	8107	-208	193	363
BR1255.78	363	8141	-176	322	631
BSMA8.39	631	8141	87	363	602
BR1153.138	602	8200	172	631	296
BR1153.62	296	8200	223	602	571
BR1519.112	571	8234	-248	296	247
BR923.49	247	8264	138	571	611
BR923.140	611	8316	96	247	472
BR1411.99	472	8352	-118	611	604
BR981.139	604	8378	114	472	223
BR981.41	223	8379	172	604	615
BR981.143	615	8382	139	223	603
BR1519.138	603	8385	-112	615	605
BR1411.139	605	8393	-76	603	42
PBBR125.A	42	8397	-85	605	597
BF981.F	597	8405	-227	42	190
BR749.33	190	8413	68	597	607
BR749.138	607	8413	127	190	614
BR981.141	614	8432	103	607	563
BR1513.111	563	8446	-225	614	592
BR981.114	592	8460	176	563	621
BR690.143	621	8508	79	592	171
BR690.28	171	8513	72	621	628
BRSMA1.40	628	8513	114	171	629
BRSMA4.40	629	8513	-114	628	591
BR690.114	591	8562	123	629	622
BR1513.143	622	8605	-77	591	554
BR1506.110	554	8631	129	622	443
BR1374.97	443	8642	190	554	508
BR1455.105	508	8650	-213	443	587
BR1455.115	587	8656	-147	508	452
BR1371.96	452	8660	-265	587	101
BR540.10	101	8691	230	452	17
BS133.A	17	8693	74	101	83
BR548.6	83	8714	118	17	153
BR640.23	153	8716	207	83	260
BR1077.55	260	8717	-107	153	415
BR1335.87	415	8761	244	260	358
BR1217.78	358	8775	206	415	309
BR1217.69	309	8775	222	358	90
BR518.10	90	8810	187	309	606
BR1217.139	606	8817	104	90	170
BR681.28	170	8841	-72	606	240
BR1013.44	240	8847	79	170	590
BR1217.115	590	8854	189	240	618
BR1217.142	618	8892	159	590	609
BR1217.140	609	8896	125	618	608
BR1335.140	608	8898	102	609	633
BR1013.HOT	633	8919	77	608	619
BR1013.142	619	8926	97	633	620
BR1335.142	620	8929	74	619	610
BR1013.140	610	8941	95	620	595
BR938.114	595	8971	-224	610	596
BR1483.114	596	8972	-198	595	616
BR938.142	616	8983	-136	596	617
BR1483.142	617	8990	-136	616	613
BR1483.140	613	8990	-144	617	612
BR938.140	612	9031	-126	613	104
BR565.12	104	9039	257	612	336
BR938.75	336	9076	-198	104	8
BS165.A	8	9084	158	336	191
BR736.32	191	9084	293	8	542
BR1483.108	542	9088	-129	191	221
BR968.40	221	9101	241	542	13
BS103.A	13	9106	157	221	468
BR1405.99	468	9106	-325	13	14
BS104.A	14	9108	52	468	418
BR1339.87	418	9126	-228	14	495
BR1420.100	495	9127	-284	418	543
BR1488.108	543	9144	-246	495	284
BR1137.61	284	9161	184	543	353
BR1137.77	353	9161	243	284	467
BR1404.99	467	9185	-328	353	525
BR1479.107	525	9186	-158	467	287
BR1127.60	287	9213	211	525	446
BR1377.97	446	9237	-184	287	398
BR1302.83	398	9262	290	446	21
BS147.A	21	9304	-83	398	254
BR925.49	254	9306	202	21	427
BR1383.92	427	9313	279	254	386
BR1294.82	386	9375	-119	427	12
BS101B.A	12	9387	116	386	26
BS106.A	26	9387	-138	12	2
BS156.A	2	9387	169	26	33
BS174.A	33	9387	169	2	36
BS177.A	36	9387	169	33	513
BR1450.104	513	9387	188	36	6
BS162.A	6	9402	-154	513	449
BR1366.96	449	9421	234	6	216
BR987.41	216	9437	141	449	435
BR1326.94	435	9476	230	216	128
BR595.16	128	9536	240	435	76
BR507.1	76	9551	-177	128	4
BS158.A	4	9552	215	76	123
BR589.18	123	9553	225	4	301
BR1192.66	301	9577	169	123	130
BR598.19	130	9579	196	301	347
BR1005.76	347	9609	-264	130	185
BR718.30	185	9620	-269	347	484
BR1433.102	484	9631	-241	185	466
BR1403.99	466	9639	-291	484	196
BR861.34	196	9670	226	466	332
BR861.74	332	9673	264	196	214
BR983.41	214	9675	235	332	282
BR1115.58	282	9705	122	214	286
BR1115.60	286	9714	296	282	225
BR1005.43	225	9734	-117	286	219
BR1003.43	219	9751	-127	225	588
BR1115.115	588	9790	252	219	162
BR697.27	162	9852	-131	588	172
BR696.28	172	9857	-126	162	41
PBBR121.A	41	9911	-95	172	294
BR1161.63	294	9933	-109	41	356
BR1161.77	356	9934	-108	294	0

Figure 3.6 Recombinant M13 clones used to sequence BamHI e

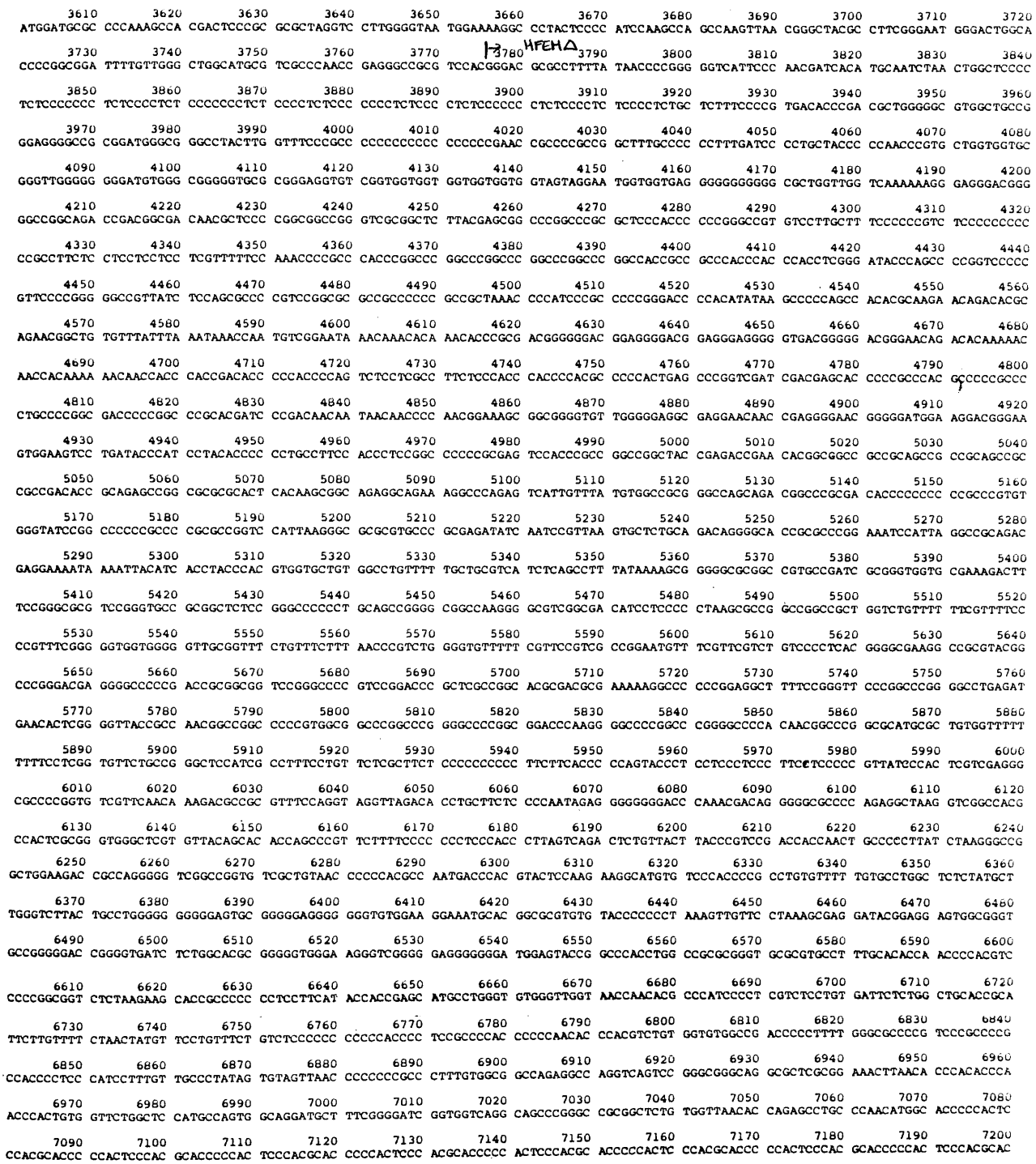
A sorted list of recombinant M13 clones used to sequence the SmaI-BamHI subfragment of BamHI e is given. Alongside the clone name is the identity number of the sequence reading, the position and orientation in the database, the length of the reading and the identity numbers of the clones to the left and right in the database.

AR131.4	19	1	115	0	76
AR217.15	76	1	167	19	101
AR384.22	101	1	178	76	103
AR391.23	103	1	182	101	120
AR404.27	120	1	192	103	43
AR165.8	43	1	247	120	85
AR148.19	85	1	276	43	45
AR169.8	45	1	284	85	93
AR365.20	93	5	134	45	38
AR154.6	38	38	155	93	110
AR385.22	110	156	182	38	82
AR216.15	82	175	-240	110	99
AR375.22	99	284	-90	82	16
AR128.4	16	309	176	99	40
AR159.7	40	325	243	16	41
AR158.7	41	504	211	40	105
AR401.24	105	551	-131	41	84
AR221.15	84	579	242	105	79
AR210.14	79	589	237	84	102
AR387.23	102	632	-177	79	26
AR140.5	26	753	-235	102	106
AR402.24	106	825	195	26	59
AR194.11	59	826	-259	106	55
AR177.9	55	832	272	59	25
AR139.5	25	940	135	55	61
AR201.12	61	947	246	25	100
AR380.22	100	1003	185	61	86
AR215.19	86	1047	249	100	114
AR393.23	114	1052	179	86	81
AR213.14	81	1070	-260	114	117
AR397.24	117	1124	-237	81	113
AR390.23	113	1133	140	117	92
AR363.20	92	1168	-200	113	119
AR403.24	119	1190	-284	92	20
AR132.4	20	1195	-204	119	51
AR166.8	51	1220	205	20	44
AR167.8	44	1224	-241	51	78
AR202.14	78	1288	-272	44	96
AR373.21	96	1319	-256	78	121
AR405.27	121	1443	-206	96	5
AR125.1	5	1458	-212	121	72
AR208.12	72	1472	182	5	39
AR155.7	39	1557	152	72	90
AR376.21	90	1584	-141	39	58
AR182.10	58	1615	-183	90	111
AR386.23	111	1619	-213	58	108
AR381.22	108	1622	-209	111	49
AR179.9	49	1635	237	108	21
AR133.4	21	1639	210	49	64
AR192.11	64	1693	-247	21	112
AR388.23	112	1695	-181	64	107
AR379.22	107	1706	185	112	95
AR367.20	95	1711	-222	107	97
AR378.21	97	1731	-153	95	11
AR116.3	11	1733	-205	97	33
AR156.7	33	1738	-171	11	67
AR198.11	67	1786	-163	33	36
AR145.6	36	1791	-187	67	34
AR157.7	34	1826	133	36	63
AR191.11	63	1830	111	34	32
AR146.6	32	1837	122	63	98
AR372.22	98	1890	181	32	23
AR135.5	23	1906	-198	98	116
AR395.24	116	1932	-251	23	66
AR195.11	66	1945	203	116	54
AR172.8	54	1989	255	66	52
AR168.8	52	2065	-250	54	27
AR141.5	27	2154	-137	52	28
AR142.5	28	2154	-179	27	4
AR124.1	4	2155	211	28	60
AR199.11	60	2201	-233	4	69
AR204.12	69	2240	261	60	87
AR364.20	87	2249	228	69	88
AR369.20	88	2271	-220	87	35
AR113.3	35	2277	-75	88	37
AR147.6	37	2295	230	35	80
AR214.14	80	2295	-279	37	91
AR361.20	91	2298	221	80	83
AR220.15	83	2305	230	91	70
AR203.12	70	2315	217	83	31
AR143.6	31	2328	-149	70	71
AR205.12	71	2341	-258	31	68
AR200.12	68	2376	208	71	42
AR161.7	42	2414	-154	68	18
AR130.4	18	2420	206	42	24
AR138.5	24	2451	-153	18	17
AR129.4	17	2567	177	24	115
AR394.23	115	2579	-203	17	75
AR212.14	75	2583	-251	115	22
AR134.5	22	2592	179	75	46
AR170.8	46	2624	267	22	50
AR164.8	50	2626	-214	46	89
AR371.21	89	2628	-200	50	125
394.H	125	2675	-107	89	74
AR163.14	74	2690	-265	125	123
AR170.34	123	2708	68	74	122
AR212.34	122	2716	-72	123	109
AR383.22	109	2815	-229	122	56
AR181.9	56	2845	190	109	65
AR193.11	65	2862	-296	56	6
AR126.1	6	2869	252	65	62
AR206.12	62	2884	-226	6	77
AR218.15	77	2888	-240	62	94
AR366.20	94	2896	-262	77	118
AR400.24	118	2902	-258	94	104
AR398.24	104	2914	-223	118	57
AR178.9	57	2955	-203	104	73
AR209.13	73	2970	-190	57	124
AR126.33	124	3120	41	73	0

Figure 3.7 Sequence of BamHI b

The entire sequence of the BamHI b restriction fragment determined using the recombinant M13 clones listed in figure 3.5, as compilled in figure 3.3 is shown in the following pages. Numbering starts at the left BamHI site. The sequence is given for the 5' to 3' strand of the fragment, left to right, as orientated in the P genome arrangement.

```
        10         20         30         40         50         60         70         80         90        100        110        120
GGATCCCAAC GACCCCGCCC ATGGGTCCCA ATTGGCCGTC CCGTTACCAA GACCAACCCA GCCAGCGTAT CCACCCCCGC CCGGGTCCCC GCGGAAGCGG AACGGGGTAT GTGATATGCT

       130        140        150        160        170        180        190        200        210        220        230        240
AATTAAATAC ATGCCACGTA CTTATGGTGT CTGATTGGTC CTTGTCTGTG CCGGAGGTGG GGCGGGGGCC CCGCCCGGGG GGCGGAACGA GGAGGGGTTT GGGAGAGCCG GCCCCGGCAC

       250        260        270        280        290        300        310        320        330        340        350        360
CACGGGTATA AGGACATCCA CCACCCGGCC GGTGGTGGTG TGCAGCCGTG TTCCAACCAC GGTCACGCTT CGGTGCCTCT CCCCGATTCG GGCCCGGTCG CTCGCTACCG GTGCGCCACC

       370        380        390        400        410        420        430        440        450        460        470        480
ACCAGAGGCC ATATCCGACA CCCCAGCCCC GACGGCAGCC GACAGCCCGG TCATGGCGAC TGACATTGAT ATGCTAATTG ACCTCGGCCT GGACCTCTCC GACAGCGATC TGGACGAGGA

       490        500        510        520        530        540        550        560        570        580        590        600
CCCCCCCGAG CCGGCGGAGA GCCGCCGCGA CGACCTGGAA TCGGACAGCA GCGGGGAGTG TTCCTCGTCG GACGAGGACA TGGAAGACCC CCACGGAGAG GACGGACCGG AGCCGATACT

       610        620        630        640        650        660        670        680        690        700        710        720
CGACGCCGCT CGCCCGGCGG TCCGCCCGTC TCGTCCAGAA GACCCCGGCG TACCCAGCAC CCAGACGCCT CGTCCGACGG AGCGGCAGGG CCCCAACGAT CCTCAACCAG CGCCCCACAG

       730        740        750        760        770        780        790        800        810        820        830        840
TGTGTGGTCG CGCCTCGGGG CCCGGCGACC GTCTTGCTCC CCCGAGCAGC ACGGGGGCAA GGTGGCCCGC CTCCAACCCC CACCGACCAA AGCCCAGCCT GCCCGCGGCG GACGCCGTGG

       850        860        870        880        890        900        910        920        930        940        950        960
GCGTCGCAGG GGTCGGGGTC GCGGTGGTCC CGGGGCTGCC GATGGTTTGT CGGACCCCCG CCGGCGTGCC CCCAGAACCA ATCGCAACCC TGGGGGACCC CGCCCCGGGG CGGGGTGGAC

       970        980        990       1000       1010       1020       1030       1040       1050       1060       1070       1080
GGACGGCCCC GGCGCCCCCC ATGGCGAGGC GTGGCGCGGC AGTGAGCAGC CCGACCCACC CGGAGGCCAG CGGACACGGG GCGTGCGCCA AGCACCCCCC CCGCTAATGA CGCTGGCGAT

      1090       1100       1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
TGCCCCCCCG CCCGCGGACC CCCGCGCCCC GGCCCCGGAG CGAAAGGCGC CCGCCGCCGA CACCATCGAC GCCACCACGC GGTTGGTCCT GCGCTCCATC TCCGAGCGCG CGGCGGTCGA

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300       1310       1320
CCGCATCAGC GAGAGCTTTG GCCGCAGCGC ACAGGTCATG CACGACCCCT TTGGGGGGCA GCCGTTTCCC GCCGCGAATA GCCCCTGGGC CCCGGTGCTG GCGGGCCAAG GAGGGCCCTT

      1330       1340       1350       1360       1370       1380       1390       1400       1410       1420       1430       1440
TGACGCCGAG ACCAGACGGG TCTCCTGGGA AACCTTGGTC GCCCACGGCC CGAGCCTCTA TCGCACTTTT GCCGGCAATC CTCGGGCCGC ATCGACCGCC AAGGCCATGC GCGACTGCGT

      1450       1460       1470       1480       1490       1500       1510       1520       1530       1540       1550       1560
GCTGCGCCAA GAAAATTTCA TCGAGGCGCT GGCCTCCGCC GACGAGACGC TGGCGTGGTG CAAGATGTGC ATCCACCACA ACCTGCCGCT GCGCCCCCAG GACCCCATTA TCGGGACGAC

      1570       1580       1590       1600       1610       1620       1630       1640       1650       1660       1670       1680
CGCGGCTGTG CTGGATAACC TCGCCACGCG CCTGCGGCCC TTTCTCCAGT GCTACCTGAA GGCGCGAGGC CTGTGCGGCC TGGACGAACT GTGTTCGCGG CGGCGTCTGG CGGACATTAA

      1690       1700       1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
GGACATTGCA TCCTTCGTGT TTGTCATTCT GGCCAGGCTC GCCAACCGCG TCGAGCGTGG CGTCGCGGAG ATCGACTACG CGACCCTTGG TGTCGGGGTC GGAGAGAAGA TGCATTTCTA

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900       1910       1920
CCTCCCCGGG GCCTGCATGG CGGGCCTGAT CGAAATCCTA GACACGCACC GCCAGGAGTG TTCGAGTCGT GTCTGCGAGT TGACGGCCAG TCACATCGTC GCCCCCCCGT ACGTGCACGG

      1930       1940       1950       1960       1970       1980       1990       2000       2010       2020       2030       2040
CAAATATTTT TATTGCAACT CCCTGTTTTA GGTACAATAA AAACAAAACA TTTCAAACAA ATCGCCCCTC GTGTTGTCCT TCTTTGCTCA TGGCCGGCGG GGCGTGGGTC ACGGCAGATG

      2050       2060       2070       2080       2090       2100       2110       2120       2130       2140       2150       2160
GCGGGGGTGG GCCCGGCGTA CGGCCTGGGT GGGCGGAGGG AACTAACCCA ACGTATAAAT CCGTCCCCGT TCCAAGGCCG GTGTCATAGT GCCCTTAGGA GCTTCCCGCC CGGGCGCATC

      2170       2180       2190       2200       2210       2220       2230       2240       2250       2260       2270       2280
CCCCCTTTTG CACTATGACA GCGACCCCCC TCACCAACCT GTTCTTACGG GCCCCGGACA TAACCCACGT GGCCCCCCCT TACTGCCTCA ACGCCACCTG GCAGGCCGAA ACGGCCATGC

      2290       2300       2310       2320       2330       2340       2350       2360       2370       2380       2390       2400
ACACCAGCAA AACGGACTCC GCTTGCGTGG CCGTGCGGAG TTACCTGGTC CGCGCCTCCT GTGAGACCAG CGGCACAATC CACTGCTTTT TCTTTGCGGT ATACAAGGAC ACCCACCACA

      2410       2420       2430       2440       2450       2460       2470       2480       2490       2500       2510       2520
CCCCTCCGCT GATTACCGAG CTCCGCAACT TTGCGGACCT GGTTAACCAC CCGCCGGTCC TACGCGAACT GGAGGATAAG CGCGGGGTGC GGCTGCGGTG TGCGCGGCCG TTTAGCGTCG

      2530       2540       2550       2560       2570       2580       2590       2600       2610       2620       2630       2640
GGACGATTAA GGACGTCTCT GGGTCCGGCG CGTCCTCGGC GGGAGAGTAC ACGATAAACG GGATCGTGTA CCACTGCCAC TGTCGGTATC CGTTCTCAAA AACATGCTGG ATGGGGGCCT

      2650       2660       2670       2680       2690       2700       2710       2720       2730       2740       2750       2760
CCGCGGCCCT ACAGCACCTG CGCTCCATCA GCTCCAGCGG CATGGCCGCC CGCGCGGCAG AGCATCGACG CGTCAAGATT AAAATTAAGG CGTGATCTCC AACCCCCCCA TGAATGTGTG

      2770       2780       2790       2800       2810       2820       2830       2840       2850       2860       2870       2880
TAACCCCCCC CAAAAAAATA AAGAGCCGTA ACCCAACCAA ACCAGGCGTG GTGTGAGTTT GTGGACCCAA AGCCCTCAGA GACAACGCGA CAGGCCAGTA TGGACCGTGA TACTTTTATT

      2890       2900       2910       2920       2930       2940       2950       2960       2970       2980       2990       3000
TATTAACTCA CAGGGGCGCT TACCGCCACA GGAATACCAG AATAATGACC ACCACAATCG CGACCACCCC AAATACAGCA TGGCGCCACA CCACGCCACA ACAGCCCTGT CGCCGGTATG

      3010       3020       3030       3040       3050       3060       3070       3080       3090       3100       3110       3120
GGGCATGATC AGACGAGCCG CGCCGCGCGT TGGGCCCTGT ACAGCTCGCG CGAATTGACC CTAGGAGGCC GCCACGCGCC CGAGTTTTGC GTTCGTCGCT GGTCGTCGGG CGCCAAAGCC

      3130       3140       3150       3160       3170       3180       3190       3200       3210       3220       3230       3240
CCGGACGGCT GTTCGGTCGA ACGAACGGCC ACGACAGTGG CATAGGTTGG GGGGTGGTCC GACATAGCCT CGGCGTACGT CGGGAGGCCC GACAAGAGGT CCCTTGTGAT GTCGGGTGGG

      3250       3260       3270       3280       3290       3300       3310       3320       3330       3340       3350       3360
GCCACAAGCC TGGTTTCCGG AAGAAACAGG GGGGTTGCCA ATAACCCGCC AGGGCCAAAA CTCCGGCGCT GCGCACGTCG TTCGGCGCGG CGCCGGGCGC GCCGAGCGGC TCGCTGGGCG

      3370       3380       3390       3400       3410       3420       3430       3440       3450       3460       3470       3480
GCTTGGCGTG AGCGGCCCCG CTCCGACGCC TCGCCCTCTC CGGAGGAGGT TGGCGGAATT GGCACGGACA ACAGGGGCCC AGCAGAGTAC GGTGGAGGTG GGTCCGTGGG GGTGTCCAGA

      3490       3500       3510       3520       3530       3540       3550       3560       3570       3580       3590       3600
TCAATAACGA CAAACGGCCC CTCGTTCCTA CCAGACAAGC TATCGTAGGG GGGCGGGGGA TCAGCAAACG CGTTCCCCGC GCTCCATAAA CCCGCGTCGG GTTGCGCCGC CTCCGAAGCC
```


HFEHΔ

HFEM A ←

```
      7210       7220       7230       7240       7250       7260       7270       7280       7290       7300       7310       7320
CCCCACTCCC ACGCATCCCC GCGATACATC CAACACAGAC AGGGAAAAGA TACAAAAGTA AACCTTTATT TCCCAACAGA CAGCAAAAAT CCCCTGAGTT TTTTTTTATT AGGGCCAACA

      7330       7340       7350       7360       7370       7380       7390       7400       7410       7420       7430       7440
CAAAAGACCC GCTGGTGTGT GGTGCCCGTG TCTTTCACTT TTCCCCTCCC CGACACGGAT TGGCTGGTGT AGTGGGCGCG GCCAGAGACC ACCCAGCGCC CGACCCCCCC CTCCCCACAA

      7450       7460       7470       7480       7490       7500       7510       7520       7530       7540       7550       7560
ACACGGGGGG CGTCCCTTAT TGTTTTCCCT CGTCCCGGGT CGACGCCCCC TGCTCCCCGG ACCACGGGTG CCGAGACCGC AGGCTGCGGA AGTCCAGGGC GCCCACTAGG GTGCCCTGGT

      7570       7580       7590       7600       7610       7620       7630       7640       7650       7660       7670       7680
CGAACAGCAT GTTCCCCACG GGGGTCATCC AGAGGCTGTT CCACTCCGAC GCGGGGGCCG TCGGGTACTC GGGGGGCATC ACGTGGTTAC CCGCGGTCTC GGGGAGCAGG GTGCGGCGGC

      7690       7700       7710       7720       7730       7740       7750       7760       7770       7780       7790       7800
TCCAGCCGGG GACCGCGGCC CGCAGCCGGG TCGCCATGTT TCCCGTCTGG TCCACCAGGA CCACGTACGC CCCGATGTTC CCCGTCTCCA TGTCCAGGAT GGGCAGGCAG TCCCCCGTGA

      7810       7820       7830       7840       7850       7860       7870       7880       7890       7900       7910       7920
TAGTCTTGTT CACGTAAGGC GACAGGGCGA CCACGCTAGA GACCCCCGAG ATGGGCAGGT AGCGCGTGAG GCCGCCCGCG GGGACGGCCC CGGAAGTCTC CGCGTGGCGC GTCTTCCGGG

      7930       7940       7950       7960       7970       7980       7990       8000       8010       8020       8030       8040
CACACTTCCT CGGCCCCCGC GGCCCAGAAG CAGCGCGGGG GCCGAGGGAG GTTTCCTCTT GTCTCCCTCC CAGGGCACCG ACGGCCCCGC CCGAGGAGGC GGAAGCGGAG GAGGACGCGG

      8050       8060       8070       8080       8090       8100       8110       8120       8130       8140       8150       8160
CCCCGGCGGC GGAAGAGGCG GCCCCCGCGG GGGTCGGGGC CGAGGAGGAA GAGGCAGAGG AGGAAGAGGC GGAGGCCGCC GAGGACGTCA GGGGGGTCCC GGGCCCACCC TGGCCGCGCC

      8170       8180       8190       8200       8210       8220       8230       8240       8250       8260       8270       8280
CCCCCGGCCC TGAGTCGGAG GGGGGGTGCG TCGCCGCCCT CTTGGCCCCT GCCGGCGCGA GGGGGGGACG CGTGGACTGG GGGGAGGGGT TTTCCTGGCC CGACCCGCGC CTCTTCCTCG

      8290       8300       8310       8320       8330       8340       8350       8360       8370       8380       8390       8400
GACGCACCGC CGCCTCCTGC TCGACAGAGG CGGCGGAGGG GAGCGGGGCG GCGCCGGAGG GGGCGGCGCC GCGGGAGGGC CCGTGCCCAC CCTCCACGCC CGGCCCCCCC GAGCCGCGCG

      8410       8420       8430       8440       8450       8460       8470       8480       8490       8500       8510       8520
CCACCGTCGC ACGCGCCCGG CACAGACTCT GTTCTTGGTT CGCGGCCTGA GCCAGGGACG AGTGCGACTG GGGCACACGG CGCGCGTCCG CGGGGCGGGC GGCCGGCTCC GCCCCGGGGG

      8530       8540       8550       8560       8570       8580       8590       8600       8610       8620       8630       8640
CCGGGGCGCG GGGGCCGGGC CCCGGAGGCG GCGCTCGCAC GCACGGGGCC ACGGCCGCGC GGGGGCGCGC GGGTCCCGAC GCGGCCGCGG ACGCGGGGGG CCCGGGGCGG GGGGCGGAGC

      8650       8660       8670       8680       8690       8700       8710       8720       8730       8740       8750       8760
CTGGCATGGG CGCCGCGGGG GGCCTGTGGG GAGAGGCCGG GGGGGAGTCG CTGATCACTA TGGGGTCTCT GTTGTTTGCA AGGGGGGCGG GTCTGTTGAC AAGGGGGCCC GTCCGGCCCC

      8770       8780       8790       8800       8810       8820       8830       8840       8850       8860       8870       8880
TCGGCCGCCC CGCCTCCGCT TCAACAACCC CAACCCCAAC CCCAACCCCC CCGGAGGGGC CAGACGCCCC CCGCGGCGCC GCGGCTCGCG ACTGGCGGGA GCCGCCGCCG CCGCTGCTGT

      8890       8900       8910       8920       8930       8940       8950       8960       8970       8980       8990       9000
TGGTGGTGGT GTTGGTGTTA CTGCTGCCGT GTGGCCCGAT GGGCGCCGAG GGGGGCGCTG TCCGAGCCGC GGCCGGCTGG GGGGCTGCGT GAGACGCCCC GCCCGTCACG GGGGGCGCGG

      9010       9020       9030       9040       9050       9060       9070       9080       9090       9100       9110       9120
CGGCGCCTCT GCGTGGGGGG GCGCGGGGCG TCCGGCGGGG GGCGGGCGGT ACGTAGTCTG CTGCAAGAGA CAACGGGGGG CGCGATCAGG TTACGCCCCC TCCCCGGCCC GCCCTTTCCT

      9130       9140       9150       9160       9170       9180       9190       9200       9210       9220       9230       9240
CGCCCGCCCG CCTATTCCTC CCTCCCCCCC CCTCCTCCTC CTCCTCCCCC AGGGTCCTTG CCGCCCCCCG CCTCACCGTC GTCCAGGTCG TCGTCATCCT CGTCCGTGGT GGGCTCCGGG

      9250       9260       9270       9280       9290       9300       9310       9320       9330       9340       9350       9360
TGGGTGGGCG ACAGGGCCCT CACCGTGTGC CCCCCCAGGG TCAGGTACCG CGGGGCGAAC CGCTGATTGC CCGTCCAGAT AAAGTCCACG GCCGTGCCCG CCCTGACGGC CTCCTCGGCC

      9370       9380       9390       9400       9410       9420       9430       9440       9450       9460       9470       9480
TCCATGCGGG TCTGGGGGTC GTTCACGATC GGGATGGTGC TGAACGACCC GCTGGGCGTC ACGCCCACTA TCAGGTACAC CAGCTTGGCG TTGCACAGCG GGCAGGTGTT GCGCAATTGC

      9490       9500       9510       9520       9530       9540       9550       9560       9570       9580       9590       9600
ATCCAGGTTT TCATGCACGG GATGCAGAAG CGGTGCATGC ACGGGAAGGT GTCGCAGCGC AGGTGGGGCG CGATCTCATC CGTGCACACG GCGCACACGT CGCCCTCGTC GCTCCCCCCG

      9610       9620       9630       9640       9650       9660       9670       9680       9690       9700       9710       9720
TCCTCTCGAG GGGGGGCGCC CCCGCAACTG CCGGGGTCTT CCTCGCGGGG GGGGCTCCCC CCCGAGACCG CCCCCCCATC CACGCCCTGC GGCCCCAGCA GCCCCGTCTC GAACAGTTCC

      9730       9740       9750       9760       9770       9780       9790       9800       9810       9820       9830       9840
GTGTCCGTGC TGTCCGCCTC GGAGGCGGAG TCGTCGTCAT GGTGGTCGGC GTCCCCCCGC CCCCCCACTT CGGTCTCCGC CTCAGAGTCG CTGCTGTCCG GCAGGTCTCG GTCGCAGGGA

      9850       9860       9870       9880       9890       9900       9910       9920       9930       9940       9950       9960
AACACCCAGA CATCCGGGGC GGGCTAAGGG GAAAAAAGGG GGGCGGGTAA GAATGGGGGG GGATTTCCCG CGTCAATCAG CACCCACGAG TTCCCCCTCT CCCCCCCCCG CCTCACAAAG

      9970       9980       9990      10000      10010      10020      10030      10040      10041
TCCTGCCCCC CTGCTGGCCT CGGAAGAGGG GGGAGAAAGG GGTCTGCAAC CAAAGGTGGT CTGGGTCCGT CCTTTGGATC C
```

Figure 3.8 Sequence of the SmaI-BamHI subfragment of BamHI e

The DNA sequence of the SmaI-BamHI subfragment of BamHI e determined using the recombinant M13 clones listed in Figure 3.6, and compiled in figure 3.4, is listed opposite. Numbering starts at the SmaI site. The sequence is given for the 5' to 3' strand of the fragment, left to right, as orientated in the P arrangement of the HSV-1 genome.

"compressions" arising from aberrant gel mobility of DNA fragments, were resolved by increasing the denaturing qualities of the gel system. This was achieved by including 30% formamide in the gel mix, or, more succesfully, by running the gel at an increased temperature. Some experiments also used dITP as a substitute for dGTP. Figure 3.9 shows an example of a compression in a sequence, alongside the same clone run on a gel maintained at a higher temperature with a hot water jacket. Although running gels at an increased temperature minimises compressions, there is an associated loss of resolution, so that the sequence cannot be read as far. This system is therefore not used for general sequencing.

Although a large number of difficult regions could be resolved using these methods, occasionally they proved insufficient. Regions of anomaly, usually with particularly high G+C contents, were subcloned as follows. The sequence surrounding the troublesome area was first examined for restriction sites. Short target sequences were then isolated by restriction endonuclease digestion from the parental plasmid clone or plasmid subclone (for example pGX53 containing the SalI-BamHI fragment from the right end of BamHI b), and cloned into M13. Sequencing reactions using M13 clones starting near the troublesome sequence, run on hot gels, usually provided good data and clarified short regions of difficulty.

On analysis of the sequence for IE gene 2, it was found that there was no continuous open reading frame. The sequences of M13 clones spanning the gene were examined closely, but no error could be found. Another plasmid, pGX190, containing the HpaI s and v fragments of HSV-1 was then examined. This plasmid was isolated

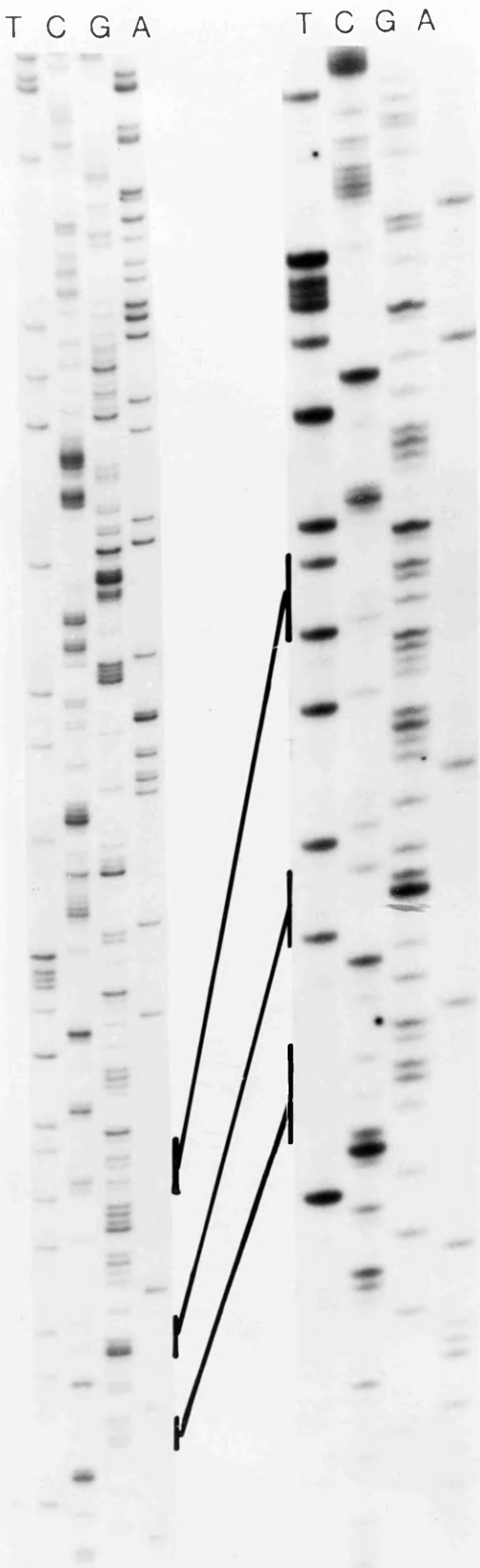

T C G A
T C G A

Figure 3.9 Resolving a compression with a hot gel

The figure shows an M13 clone sequenced as a normal run, alongside the same clone run on a hot gel. The resolved compressions are indicated. The sequence is from the BamHI b fragment, approximately between position 825 and 870. The gradient gel is on the left, the hot gel on the right.

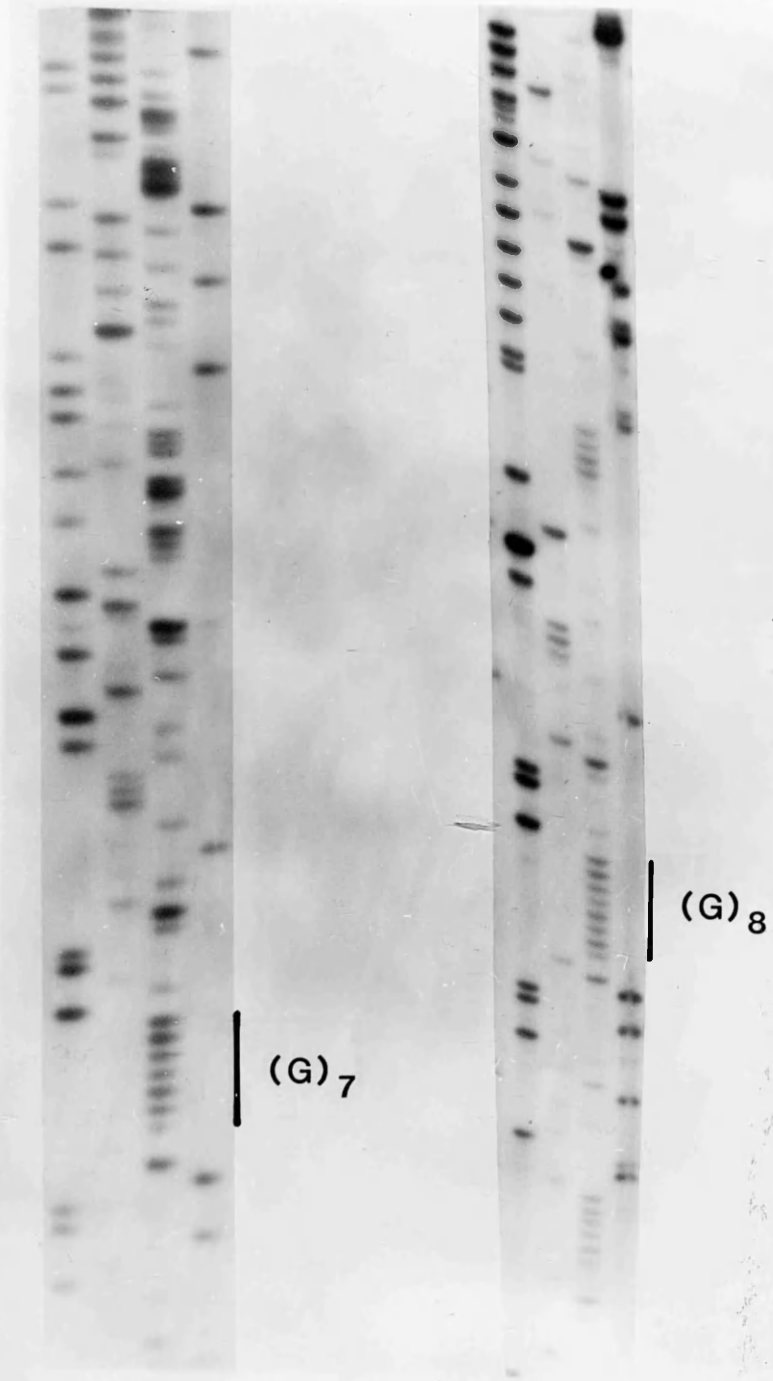

(G)7
(G)8

Figure 3.10 Sequence difference between recombinant M13 clones generated from two distinct plasmids

Two clones spanning the same sequence, showing a difference in number in a tract of G residues at position 1062 in the BamHI b fragment, are shown. Clone BR1281.36 on the left, was generated from plasmid pGX48, and CP16 on the right, from pGX190.

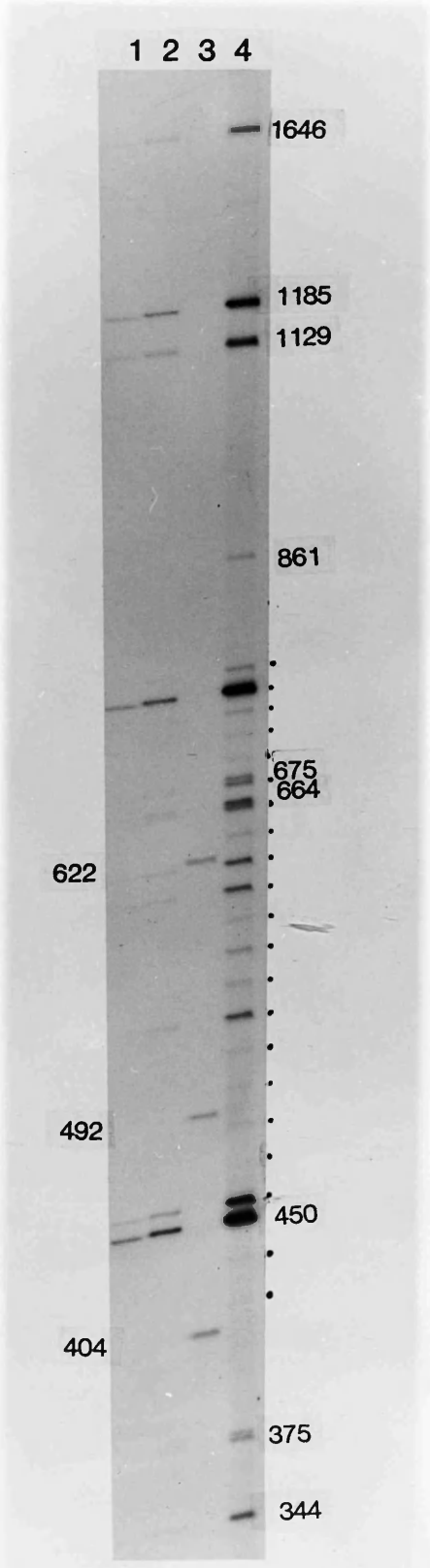

1 2 3 4
1646
1185
1129
861
675
664
622
492
450
404
375
344

Figure 3.11 Determination of the copy numbers of reiteration 1 in BamHI b

The copy numbers of reiteration 1 in plasmid cloned BamHI b were examined by digesting isolated BamHI b with SmaI, 5' end labelling with [γ^{32}P]ATP, and fractionating the DNA as single strands on a 3% polyacrylamide gel, tracks 1, 2 and 4, alongside pAT/HaeII size markers, track 3. The sizes (nucleotides) of the pAT/HaeII marker bands are shown on the left of the figure, known fragment sizes of BamHI b/SmaI digest are shown on the right. The ladder of fragments containing the variable copy numbers of the reiteration is indicated by dots to the right of track 4.

fragments (figure 3.11). The ladder represents the fragments containing the reiteration. Other bands in the track were predicted from the sequence data. Taking the known bands as size markers, the computer program MOLGEL was used to calculate the sizes of the variable fragment. With reference to the sequence data it was possible to calculate the number of copies of the reiteration present in each band. The copy number was found to range from one copy to at least 21. Fragments with copy numbers of three and 20 were the most abundant. Four copies of the reiteration are present in the BamHI b database. The copy number of this reiteration set was not seen to vary in BamHI e. Only one M13 clone, containing ten copies of the reiteration, spans this region of BamHI e.

3.2 LOCATION OF THE R_L/U_L JUNCTION AND COMPARISON OF THE R_L SEQUENCES

In addition to the unique sequences, the SmaI-BamHI fragment of BamHI e contains 549 bp of TR_L. The junction between the long unique and repeat sequences was identified by aligning the homologous R_L sequences of BamHI b and BamHI e and noting where they diverged. The junction is located directly adjacent to the set of tandem reiterations with a variable copy number, described above. The reiterations are part of the repeat structure. The R_L/U_L junction in BamHI e is at positions 549/550 in the sequence, as orientated in the database. The U_L/R_L junction in BamHI b is at positions 3836/3837 in the sequence as listed in the database.

The R_L sequences adjacent to the R_L/U_L junction have been aligned, and are listed in figure 3.12. (The

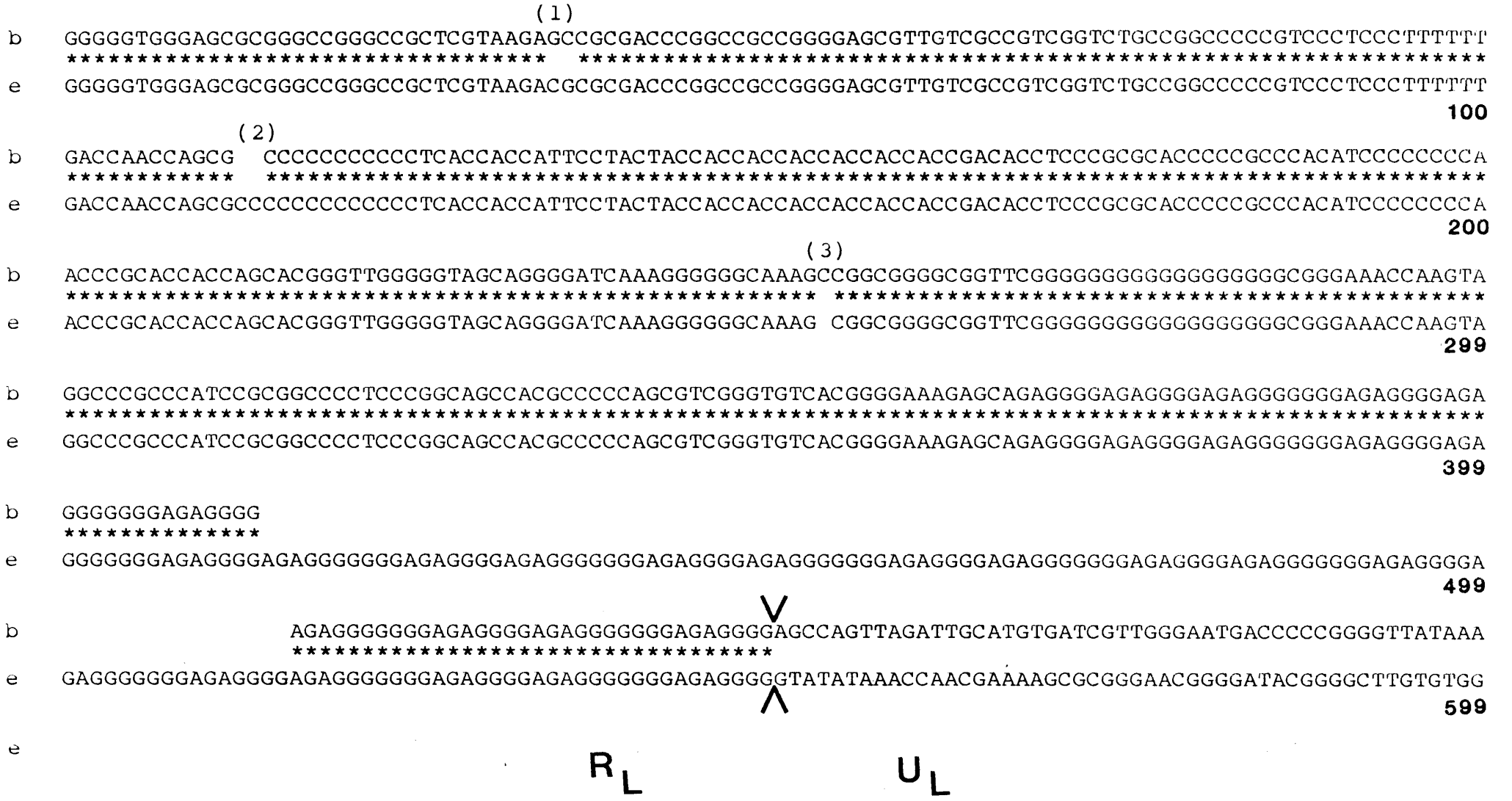

(1)
b GGGGGTGGGAGCGCGGGCCGGGCCGCTCGTAAGAGCCGCGACCCGGCCGCCGGGGAGCGTTGTCGCCGTCGGTCTGCCGGCCCCCGTCCCTCCCTTTTTT
e GGGGGTGGGAGCGCGGGCCGGGCCGCTCGTAAGACGCGCGACCCGGCCGCCGGGGAGCGTTGTCGCCGTCGGTCTGCCGGCCCCCGTCCCTCCCTTTTTT
100
(2)
b GACCAACCAGCG CCCCCCCCCCCTCACCACCATTCCTACTACCACCACCACCACCACCACCGACACCTCCCGCGCACCCCCGCCCACATCCCCCCCCA
e GACCAACCAGCGCCCCCCCCCCCCCCTCACCACCATTCCTACTACCACCACCACCACCACCACCGACACCTCCCGCGCACCCCCGCCCACATCCCCCCCCA
200
(3)
b ACCCGCACCACCAGCACGGGTTGGGGGTAGCAGGGGATCAAAGGGGGGCAAAGCCGGCGGGGCGGTTCGGGGGGGGGGGGGGGGGGCGGGAAACCAAGTA
e ACCCGCACCACCAGCACGGGTTGGGGGTAGCAGGGGATCAAAGGGGGGCAAAG CGGCGGGGCGGTTCGGGGGGGGGGGGGGGGGGCGGGAAACCAAGTA
299
b GGCCCGCCCATCCGCGGCCCCTCCCGGCAGCCACGCCCCCAGCGTCGGGTGTCACGGGGAAAGAGCAGAGGGGAGAGGGGAGAGGGGGGGAGAGGGGAGA
e GGCCCGCCCATCCGCGGCCCCTCCCGGCAGCCACGCCCCCAGCGTCGGGTGTCACGGGGAAAGAGCAGAGGGGAGAGGGGAGAGGGGGGGAGAGGGGAGA
399
b GGGGGGGAGAGGGG
e GGGGGGGAGAGGGGAGAGGGGGGGAGAGGGGAGAGGGGGGGAGAGGGGAGAGGGGGGGAGAGGGGAGAGGGGGGGAGAGGGGAGAGGGGGGGAGAGGGGA
499
V
b AGAGGGGGGGAGAGGGGAGAGGGGGGGAGAGGGGAGCCAGTTAGATTGCATGTGATCGTTGGGAATGACCCCCGGGGTTATAAA
e GAGGGGGGGAGAGGGGAGAGGGGGGGAGAGGGGAGAGGGGGGGAGAGGGGGTATATAAACCAACGAAAAGCGCGGGAACGGGGATACGGGGCTTGTGTGG
599
^
e
R_L
U_L

Figure 3.12 Sequence across the R_L/U_L junctions

The sequence across the TR_L/U_L junction is listed below the IR_L/U_L junction sequence. The sequences start from the SmaI site, numbering is as in the BamHI e database. Homologous R_L sequences are indicated with an *. Three differences between the R_L sequences of the two fragments are numbered. The BamHI b sequence shown contains four copies of the reiteration. These are aligned with the first two and last two copies in the BamHI e sequence, which contains ten copies. The sequence divergence at the R_L/U_L junctions is indicated.

BamHI b sequence is in reverse orientation, relative to the database). The figure shows that the two sequences are not completely identical. As stated above, the copy number of the reiteration set at the R_L/U_L junction in the bamHI b and BamHI e databases differs. There are three further differences between the two R_L sequences.

At positions 35 and 36 in the sequence presented for BamHI e, there is an inversion of two nucleotides relative to BamHI b. Referring to the BamHI b database, this region is well represented by M13 clones on both strands. In the BamHI e database, all the clones spanning this region are in one orientation. There is also evidence of a compression in the sequence of several clones. Therefore it is possible that this difference does not reflect the situation in the plasmid. The second difference is at pos ition 113 in the BamHI e sequence, where there is an insertion of two nucleotides, relative to BamHI b. Examination of the sequences of the M13 clones in these regions suggests this difference is real and not due to a compression. This situation resembles the mutations described in the previous section as it involves a homopolymer tract (13 bp in BamHI e and 11 bp in BamHI b). Finally, there is a deletion of a single residue at position 253 in the BamHI e sequence, relative to BamHI b. There is no evidence of compressions, and the region is well represented by M13 clones covering both strands in both BamHI b and e. This probably also represents the real situation in the plasmid.

It is not possible to conclude from this data whether the differences observed in the R_L sequences of the two restriction fragments originated in the virus, or occurred at some point during the cloning and growth

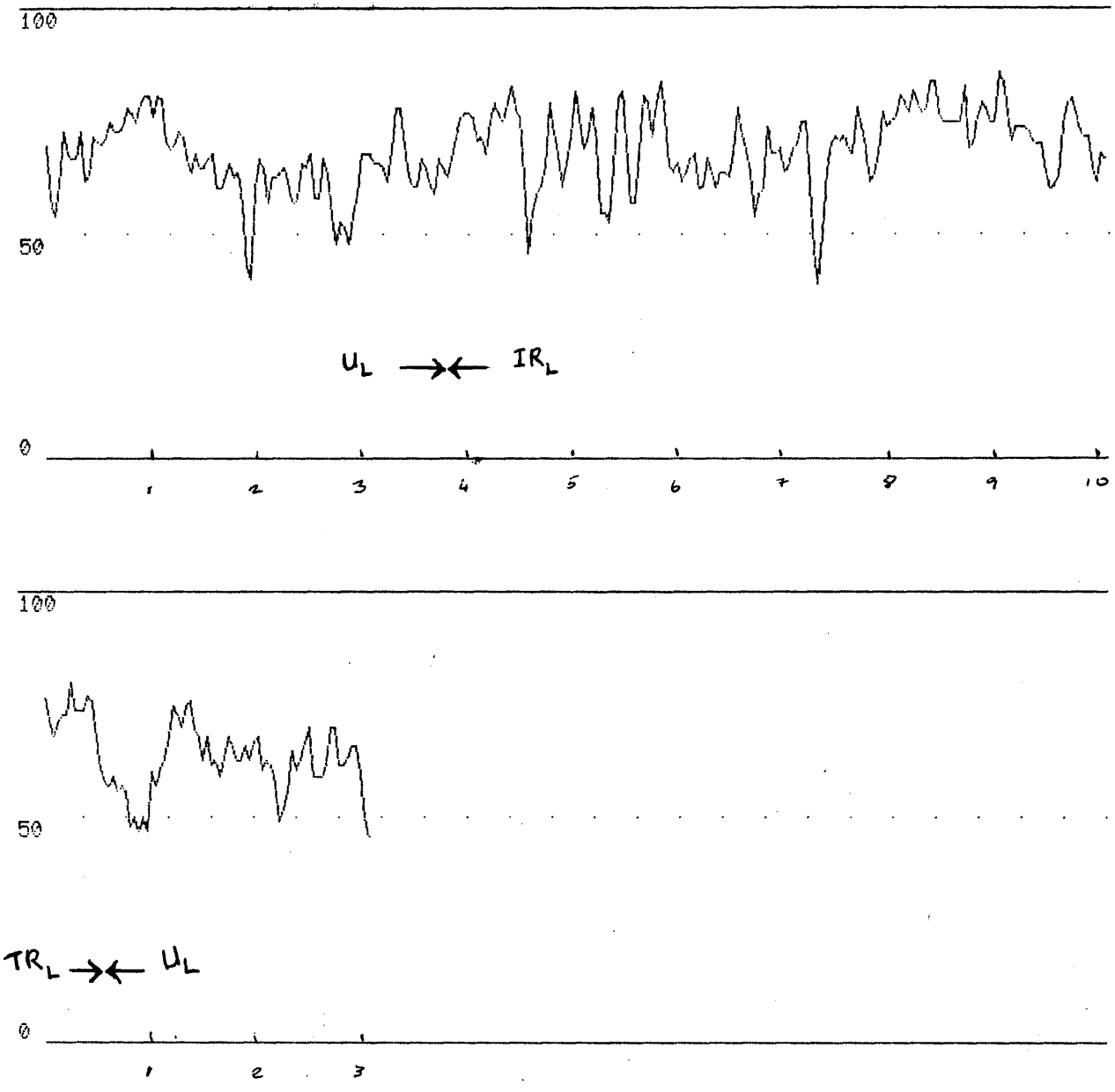

100
50
U_L → ← IR_L
0
1
2
3
4
5
6
7
8
9
10
100
50
TR_L → ← U_L
0
1
2
3

Figure 3.13 Base compositions of the BamHI b and SmaI/BamHI subfragment of BamHI e

The base compositions along the two DNA sequences determined are shown. The upper part of the figure illustrates the G+C content of BamHI b, as orientated in the database. The G+C content of the BamHI e sequence determined is shown below, also as orientated in the database. High scores represent regions with high G+C contents. A window length of 100 was used, and moved along the sequence in steps of 40.

The boundaries between the repeat and unique sequences are indicated in the figures. The y axis represents 0 to 100% G+C, the x axis represents kilobase pairs.

of the plasmids.

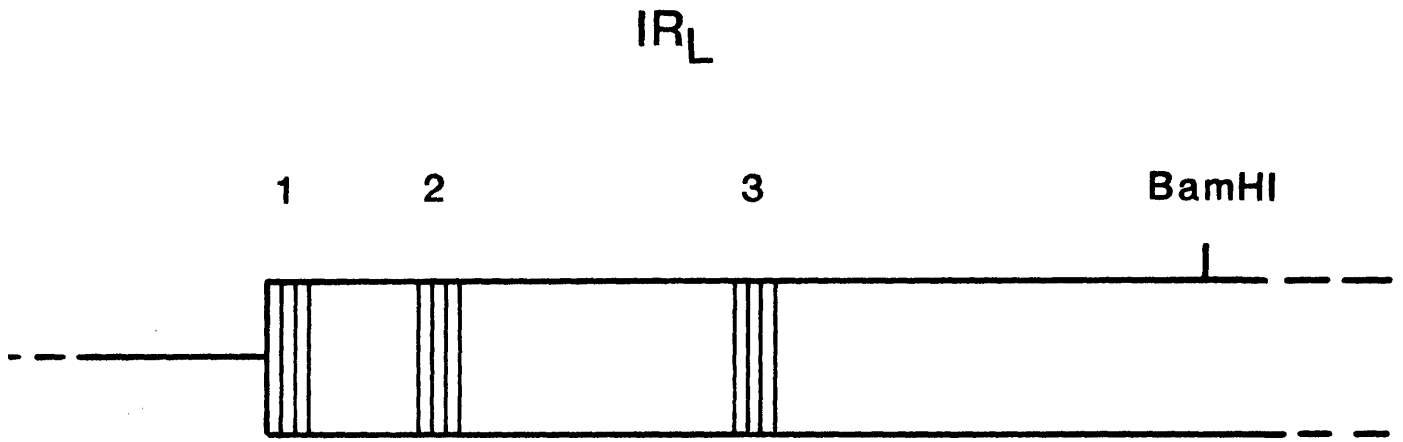
3.3 CHARACTERISTICS OF THE DNA SEQUENCE OF BAMHI b AND e

As previously stated, the sequence analysed has a high G+C content. The local base composition varies within the region studied, most notably between the repeat and unique sequences. The 6204 bp of R_L contained in BamHI b has a G+C content of 70.6%. The U_L components of BamHI b and e are 65.6% and 61.8% G+C respectively. Figure 3.13 shows the base composition along the two sequences.

A characteristic of the sequence is the presence of tandem reiterations (several identical tandemly repeated copies of a sequence). Reiterated sequences are particularly numerous in R_L. Figure 3.14 lists the sequence and locations of reiterations in the R_L sequence contained in BamHI b. Reiteration set (1) is present in variable copy numbers in pGX48 DNA. No other reiteration was seen to vary in copy number. In addition to these sets of reiterations, R_L contains a large number of smaller, imperfect repeat elements. There are eight copies of the triplet CCG directly repeated between positions 4444 and 4515, and 19 copies of the same triplet on the opposite strand, between positions 5653 and 5862. Nine copies of another G+C rich sequence, CCCGG are directly repeated between positions 5741 and 5861.

3.4 ANALYSIS OF THE GENETIC CONTENT OF THE SEQUENCE

The determined DNA sequences of BamHI b and BamHI



(1) (CCCCTCTCCCCCCCTCT)n

(2) (CGGCC)6

(3) (GCACCCCCACTCCCAC)9

Figure 3.14 Location and sequence of tandem reiterationsin the IR_L sequence of BamHI b

The upper part of the figure represents the BamHI b fragment of HSV-1. IR_L is illustrated as an open box. Reiteration sets are indicated as vertical lines. The sequences of the reiteration sets are listed below. The sequences are given for the 5' to 3' strand of the DNA, right to left as orientated in the figure above.

e were examined to identify previously mapped genes, and to predict further genes. Initially, the locations of genes known to lie in the sequences were identified. The only two proteins known to be coded in these sequences are IE110 and IE63 (Clements et al., 1977 and 1979). The positions of the 5' and 3' termini of the transcripts of IE gene 1 and IE gene 2 in the sequence were identified from previous studies (Mackem and Roizman, 1982b; Whitton et al., 1983; Rixon et al., 1984).

Evidence for further genes in these regions from studies with ts or deletion mutants is minimal. The strain of HSV-1 HFEM has a large deletion across the IR_L/U_L junction (Rosen and Darai, 1985). Referring to the work described in this thesis, the deletion has been exactly mapped, and is known to remove sequences between positions 3776 and 7227 (fig 3.7) (Darai, personnal communication). As the strain HFEM has an altered phenotype, exhibiting reduced pathogenicity, it was concluded that the deletion must affect one or more genes. Therefore one would expect to find at least a part of one gene within this sequence. Jenkins et al., 1985 have described a method for targeting large inserts into the HSV-1 genome. The insertion site in the two mutants described has not been exactly identified, however, refering to the sequence presented in this thesis, both appear to lie within IR_L. As genes in R_L are diploid, the mutation of any gene entirely contained within R_L would not be expected to be deleterious.

The next stage in this analysis was to locate ORFs in the sequence. Due to the high G+C content of the DNA, there are few out of frame nonsense codons. To illustrate this Figure 3.15 shows all the ORFs

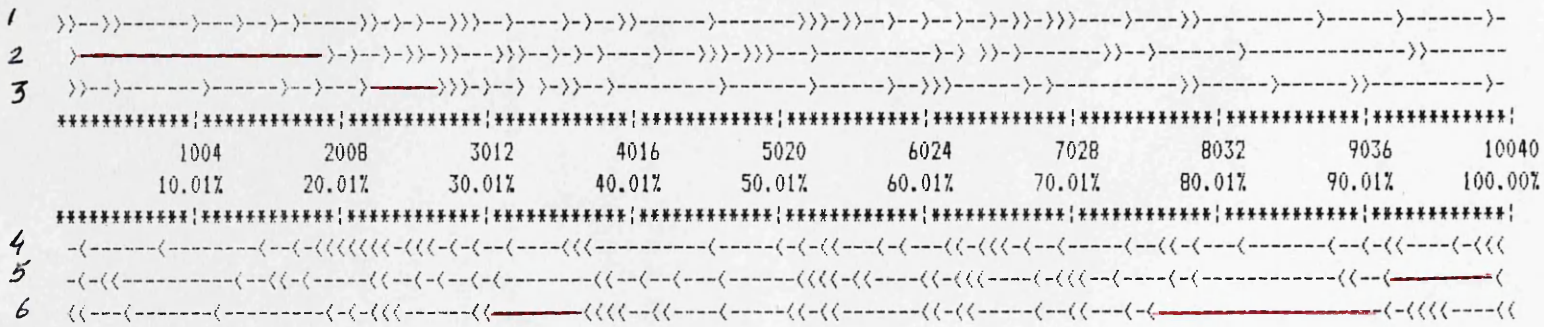

1
2
3
4
5
6
1004
10.01%
2008
20.01%
3012
30.01%
4016
40.01%
5020
50.01%
6024
60.01%
7028
70.01%
8032
80.01%
9036
90.01%
10040
100.00%

Figure 3.15 Open reading frames in the BamHI b sequence

All open reading frames on both strands of the sequence are shown. The three rightward reading frames are shown above, and the three leftward reading frames below. Open reading frames are represented by ----->. Angled brackets, > and <, indicate in-frame stop codons.

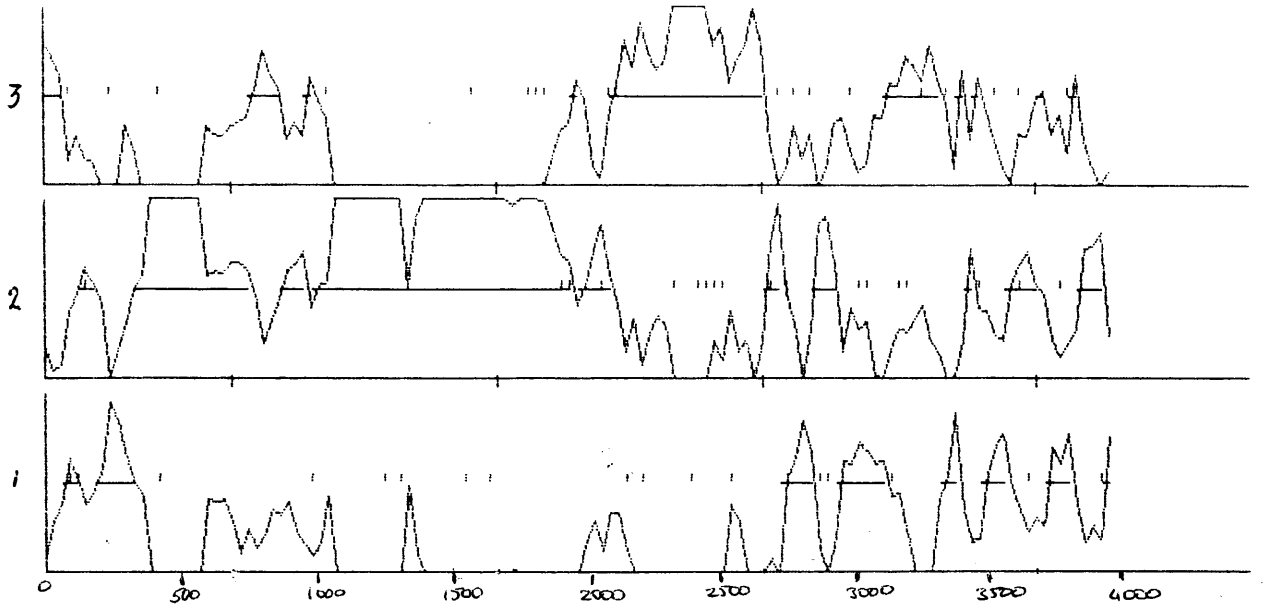
The reading frames are numbered on the left. Open reading frames proposed in the text to encode IE110, IE63 and two predicted polypeptides of 20.5K and 21.2K are shown in red.

present in the BamHI b sequence determined. This analysis must therefore be supplemented by further criteria. Initially, only ORFs beginning with an ATG were considered. This is valid as the majority of HSV-1 genes characterised are known to be unspliced. Where appropriate, this was supplemented with the identification of all ORFs, usually of a minimum size. This allows for spliced genes, and was used in regions where no obvious ORF was identified.

The sequences flanking the ORFs were examined for transcriptional control signals. TATA box homologies (Corden et al., 1980) would be expected 25 to 30 bp upstream of the initiation codon of an ORF. A search was also made for polyadenylation signals, downstream of the ORFs (Fitzgerald and Shenk, 1980). HSV-1 genes are often arranged in 3' coterminal families (for example, McGeoch et al., 1985). Consequently, polyadenylation signals are not invariably found short distances downstream of the polypeptide coding sequences, and transcripts will often have extensive untranslated trailer sequences.

A transcription termination signal 25 to 30 bp downstream of the polyadenylation signal of many genes (including HSV-1 genes) has been identified (McLauchlan et al., 1985). The sequences downstream of all the polyadenylation signals were examined for homologues to this sequence. However, it is recognised that not all genes have this sequence.

The polypeptide coding regions of potential genes identified using these parameters were analysed. Using the polypeptide coding sequences of a known gene, the codon usage of the ORFs was evaluated (Staden and McLachlan, 1982). An appropriate reference gene was


3
2
1
0
500
1000
1500
2000
2500
3000
3500
4000

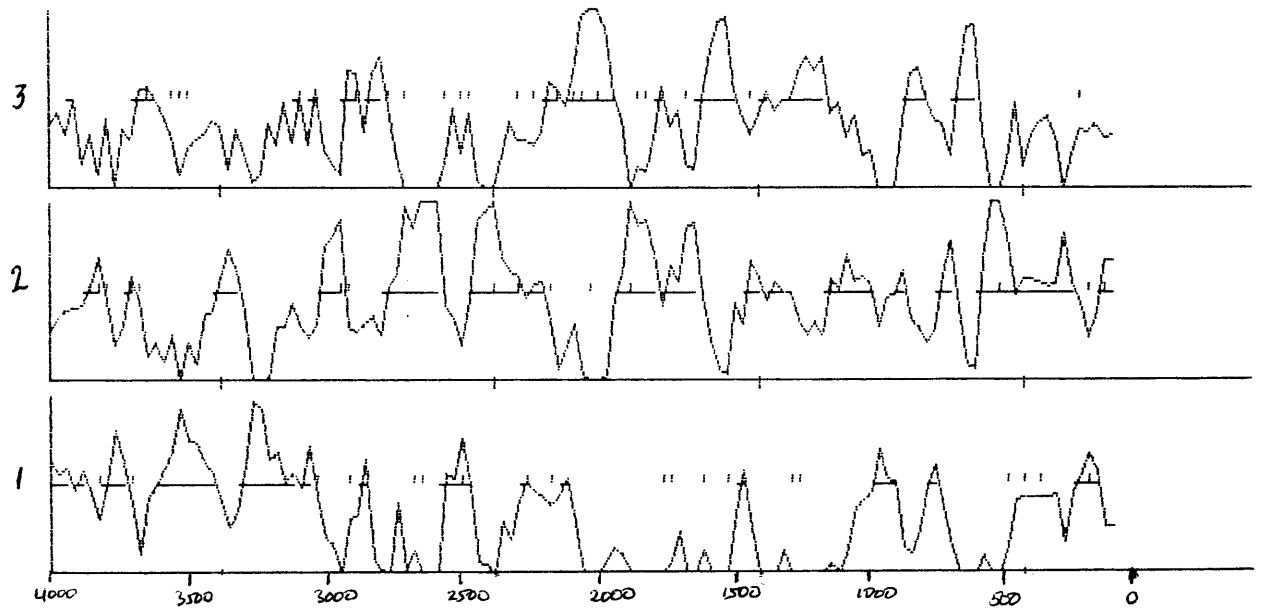
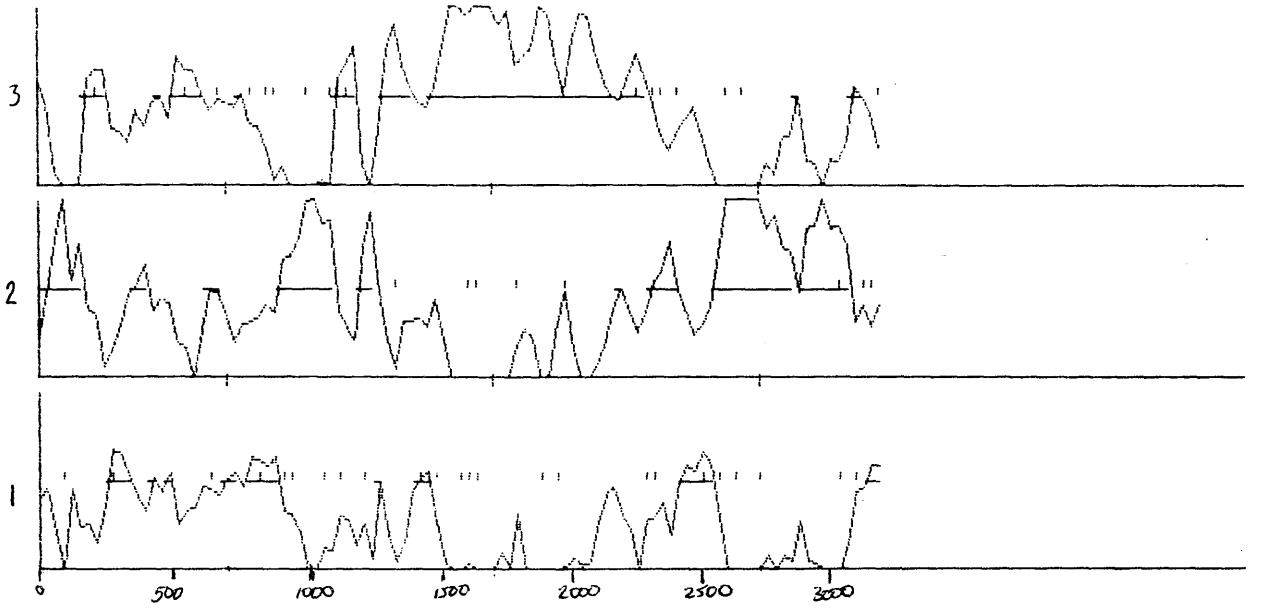

3
2
1
4000
3500
3000
2500
2000
1500
1000
500
0

Figure 3.16 Codon usage evaluation for all reading frames in the U_L region of BamHI b

The codon usage of the three reading frames in both strands of BamHI b, from positions 1 to 4000, was evaluated using the program FRMSCN (Staden and McLachlan, 1982), using the coding region of the gene US3 as reference. In frame stop codons are shown as vertical lines on the central axis. Reading frames with similar codon usage score highest.

The top panel represents the left to right DNA strand, reading frames 1, 2 and 3, are marked. The lower panel represents reading frames on the right to left strand, reading frames 1, 2 and 3 are marked.

The x axis represents the numbered position in the DNA sequence, as in the standard orientation. The y axis represents a log probability function, with an arbitrary scale (Staden and McLachlan, 1982).


3
2
1
0
500
1000
1500
2000
2500
3000

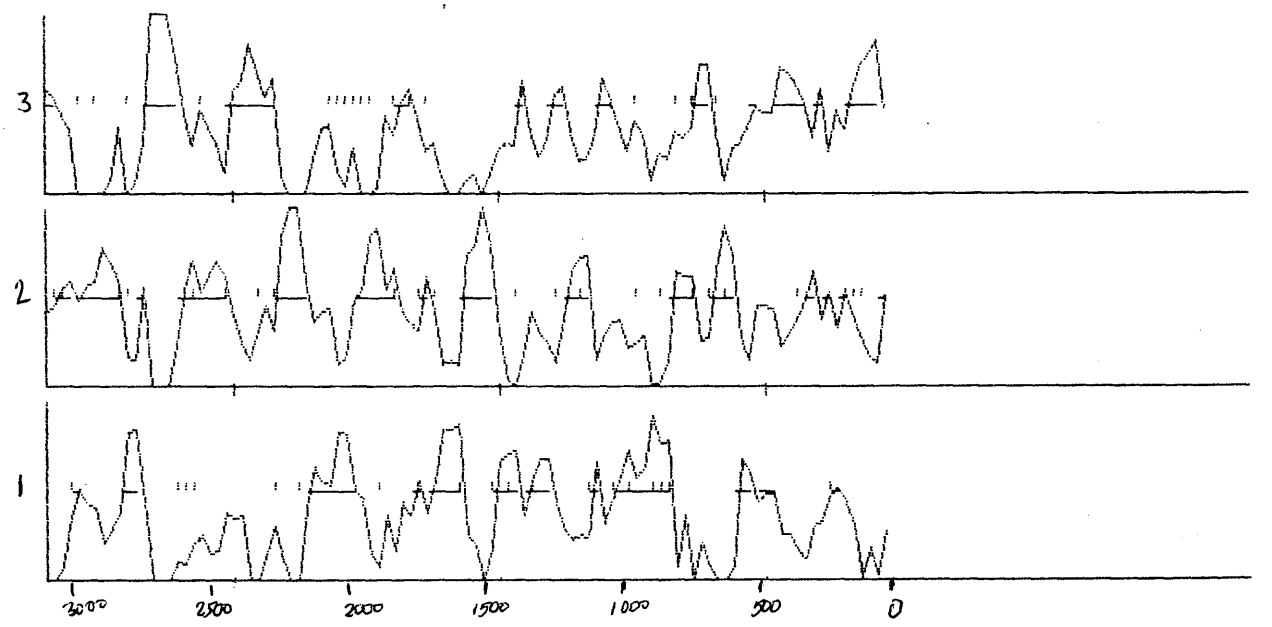

3
2
1
3000
2500
2000
1500
1000
500
0

Figure 3.17 Codon usage evaluation of all reading frames in the SmaI/BamHI subfragment of BamHI e

The codon usage of the three reading frames in both strands of BamHI e, from position 1 to the end of the determined sequence, was evaluated using the program FRMSCN (Staden and McLachlan, 1982), using the coding regions of IE gene 2 as reference. In frame stop codons are shown as vertical lines on the central axis. Reading frames with similar codon usage score highest.

The top panel represents the left to right DNA strand, reading frames 1, 2 and 3, are marked. The lower panel represents reading frames on the right to left strand, reading frames 1, 2 and 3 are marked.

The x axis represents the numbered position in the DNA sequence, as in the standard orientation. The y axis represents a log probability function, with an arbitrary scale (Staden and McLachlan, 1982).

chosen for each evaluation. For the evaluation of genes in the U_L regions, for example, the coding region of IE gene 2 or of the U_S genes were generally used. Figures 3.16 and 3.17 show the codon evaluation of all three reaing frames from both strands, in the U_L sequences of BamHI b and BamHI e respectively. To evaluate the ORFs in R_L, IE gene 1 or IE gene 3, both coded in the major repeat elements which are composed of G+C rich DNA, were used as reference genes. As previously mentioned, there is a significant difference in base compositions between the R_L (70.6% G+C) and U_L (65.6% and 61.8% G+C) sequences determined. The base composition of the DNA will affect the codon usage of the gene.

There are a number of characteristics of HSV-1 genes associated with the base composition of the DNA. Both termini of HSV-1 genes generally have a lower G+C content. Consequently, on a graphic representation of the base composition of the sequence, as in figure 3.13, it may be possible to identify genes. To counteract the high G+C content of the DNA and to maximise coding potential, there is a tendency towards a G or C residue in the third, redundant position. This can be illustrated graphically, by plotting the G+C contents of the first, second and third codon positions of an ORF. Figure 3.18 shows the G+C contents of the three codon positions of the polypeptide coding sequences of IE gene 1. Generally, the lowest G+C content, where there is least flexibility, is in the second codon position.

Subsequent transcript mapping of the proposed genes will substantiate the analysis. This has been used to confirm the proposed structure of IE gene 1, and will be described later. Further supportive

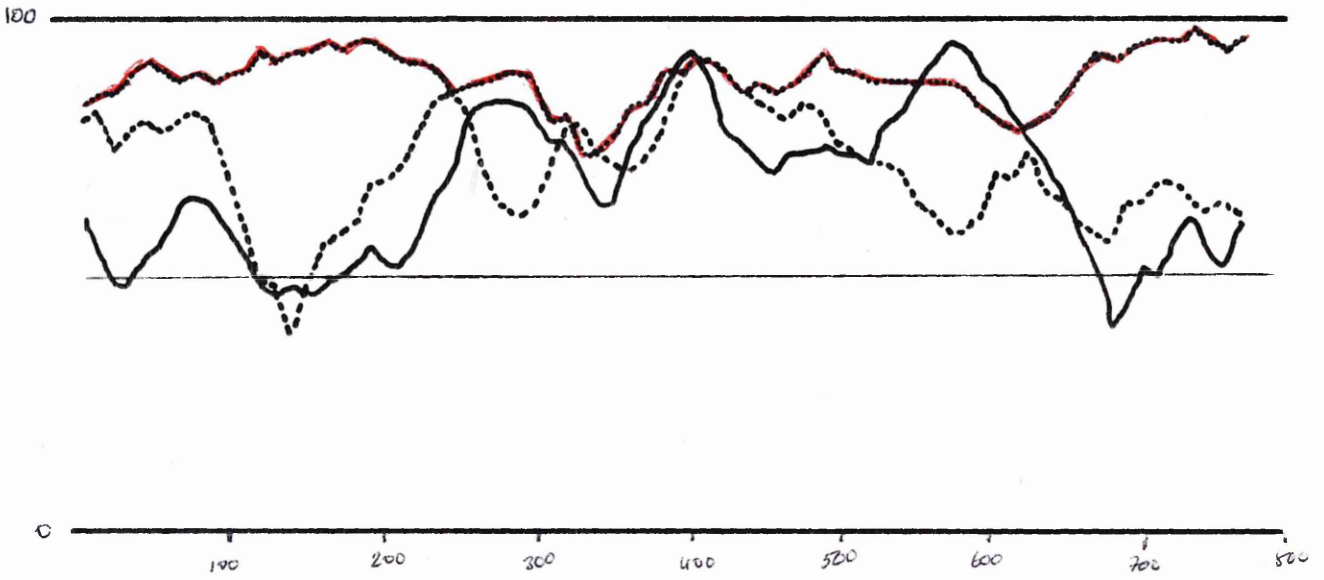

100
0
100
200
300
400
500
600
700
800

Figure 3.18 Base composition of the three codon positions of IE gene 1

The G+C contents of the coding regions of IE gene 1 are illustrated. The program PROFILE (P. Taylor, unpublished) was used. A window length of 50 and a shift of 25 were the parameters used. The three codon positions are represented as follows. -------, 1st position; ———, 2nd position; and -·-·-·-, third position.

The x axis represents nucleotide position in the DNA sequence of the non-coding strand of IE gene 1. Numbering is as in the standard sequence. The y axis represents 0 to 100% G+C.

evidence for the interpretation can come from mutational analysis of the gene, also based on the data. The predicted amino acid sequence can be used to produce an oligopeptide and subsequently an antiserum directed against the protein. This can be used in infected cells to precipitate the native protein. This will validate the chosen ORF in this region of the protein.

The translated amino acid sequences of the proposed polypeptide coding regions can be compared with closely related proteins using the CINTHOM plot program (Pustell and Kafatos, 1982), described in section 2.23. During this work, data from VZV and EBV were extensively used. Both viruses have been entirely sequenced (Davison and Scott, 1986; Baer et al., 1984), and an extensive comparison with all ORFs was possible. These yielded several positive results which will be illustrated later with the coverage of each protein. Where appropriate and available, comparisons were made with the amino acid sequences of other HSV-1 and HCMV proteins. The amino acid sequence of IE110 was compared with the amino acid sequences of both HSV-1 IE175 and with the major IE protein of HCMV, but neither displayed any homology. In addition, a comprehensive search for homologous proteins was made using the protein sequence database of the National Biomedical Research Foundation. This was unsuccessful in finding any homologues.

Several features are only usually present in non-coding regions. Although reiterated DNA sequences have been reported in the polypeptide coding regions of several HSV-1 genes (McGeoch et al., 1985; Chou and Roizman, 1986), the sequences are always multiples of three, and therefore encode repeated amino acid

sequences. There are no reports of reiterated sequences of other lengths present within the polypeptide coding regions of HSV-1 genes. Therefore ORFs spanning reiterated sequences are unlikely to be polypeptide coding, unless it codes for a repeated amino acid sequence.

Where data from more than one source is available, for example where two overlapping restriction fragments have been sequenced, it is possible to look for variability in the DNA sequence. In this work several instances were identified. In the BamHI b/XhoI c overlap, a deletion/insertion of one nucleotide was found; in IE gene 2, the insertion/deletion of a single residue; in the R_L sequences determined of BamHI b and e, three instances of an insertion/deletion of one or two residues, and a reiteration with a highly variable copy number. Although one of these instances was found to lie within the coding region of a gene, these occurrences may reflect the situation in the viral population. If the heterogeneity does originate from the virus, one may expect a higher level, and tolerance, of mutations outside of the coding regions and transcriptional control signals of the genes. Tentative use of this information may be helpful in some instances.

These analyses were systematically used to establish the arrangement of genes in the DNA sequences determined. The organisation of genes within the U_L sequences of both BamHI b and BamHI e appears to have been resolved using this system. The proposed arrangement of genes in the U_L sequences of BamHI b and BamHI e are illustrated in Figures 3.19 and 3.20. In addition to IE gene 2, two potential genes have been identified in the U_L region of BamHI b, and three

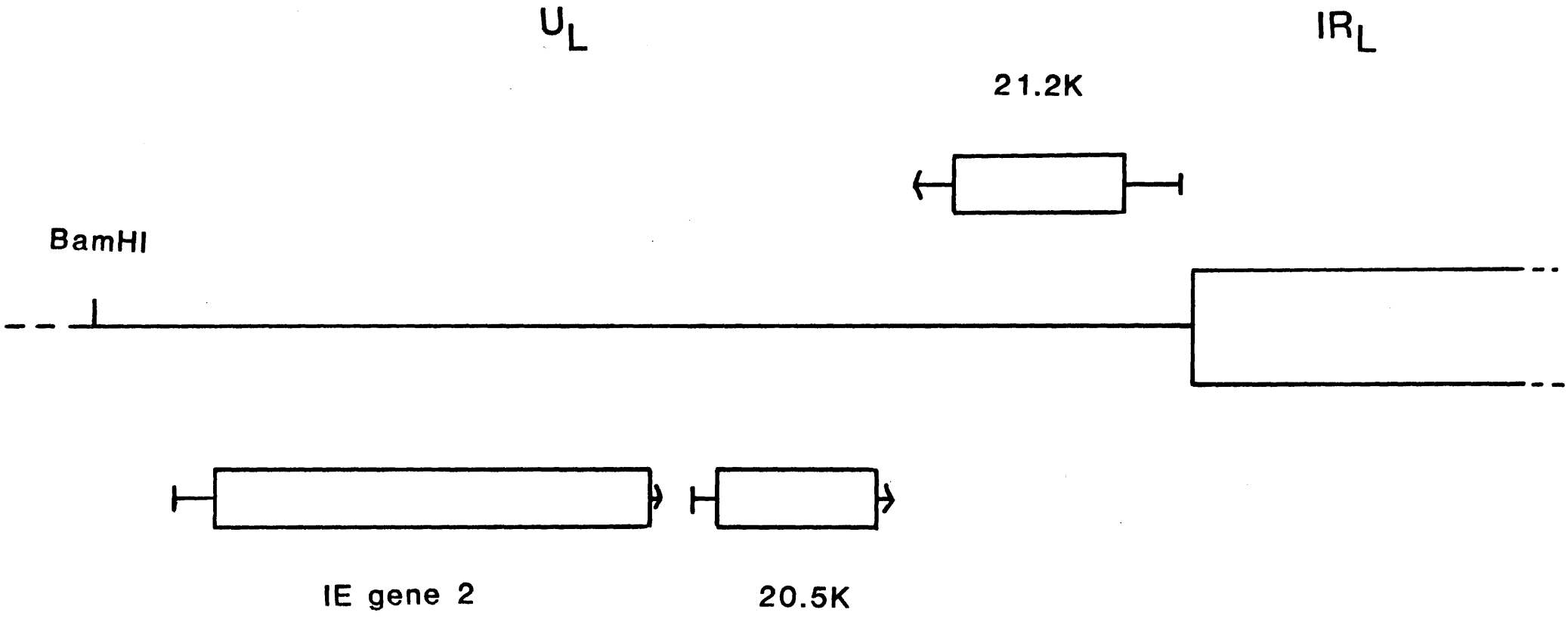

U_L
IR_L
21.2K
BamHI
IE gene 2
20.5K

Figure 3.19 Proposed location and orientation of genes in the U_L portion of BamHI b

The position of IE gene 2 and two predicted genes in the U_L sequence of BamHI b is illustrated. The location and orientation of genes is shown with a representation of the U_L sequence (solid line) adjacent to IR_L (open box). ORFs are shown as open boxes. Distances between the proposed coding region and the mapped 5' and 3' ends of IE gene 2 mRNA are shown as solid lines. Distances between the ORFs and proposed TATA box and polyadenylation signal sequences of the two predicted genes, representing proposed 5' and 3' non-coding regions, are shown as solid lines.

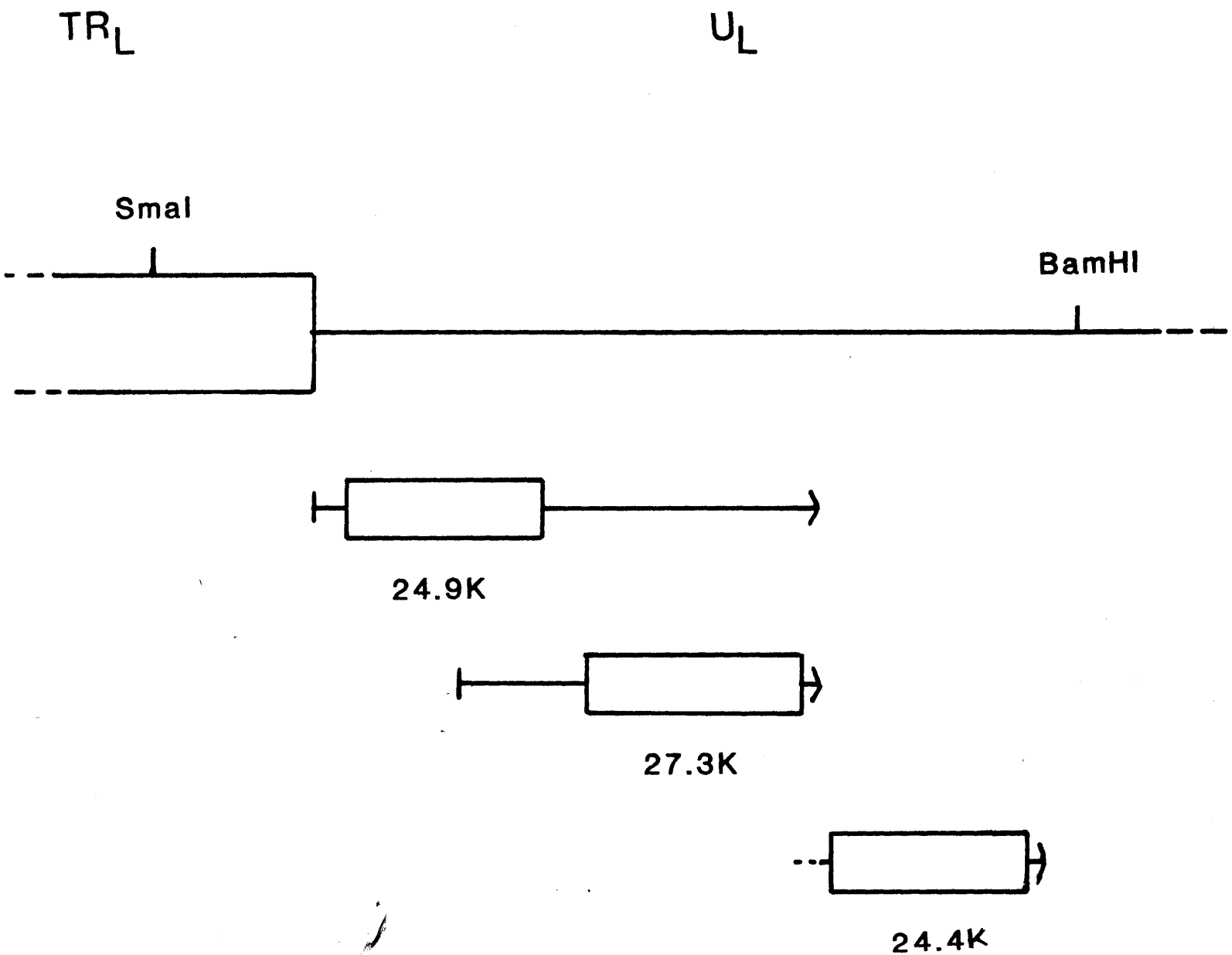

TR_L
U_L
SmaI
BamHI
24.9K
27.3K
24.4K

Figure 3.20 Proposed location and orientation of genes in the U_L portion of BamHI e

The position of three predicted genes in the U_L sequence of BamHI e is illustrated. The location and orientation of genes is shown with a representation of the U_L sequence (solid line) adjacent to TR_L (open box). ORFs are shown as open boxes. Distances between the ORFs and the proposed TATA box and polyadenylation signal sequences of the predicted genes, representing proposed 5' and 3' non-coding regions, are shown as solid lines.

potential genes in the U_L component of BamHI e. The location and structure of IE gene 1, in R_L has been established, and the structure of the gene will be described below. Each of these genes and their encoded proteins will be dealt with individually. In R_L, the situation is complex. No gene in addition to IE gene 1 was identified. An analysis of this region will be given later, in detail.

3.5 IE GENE 1

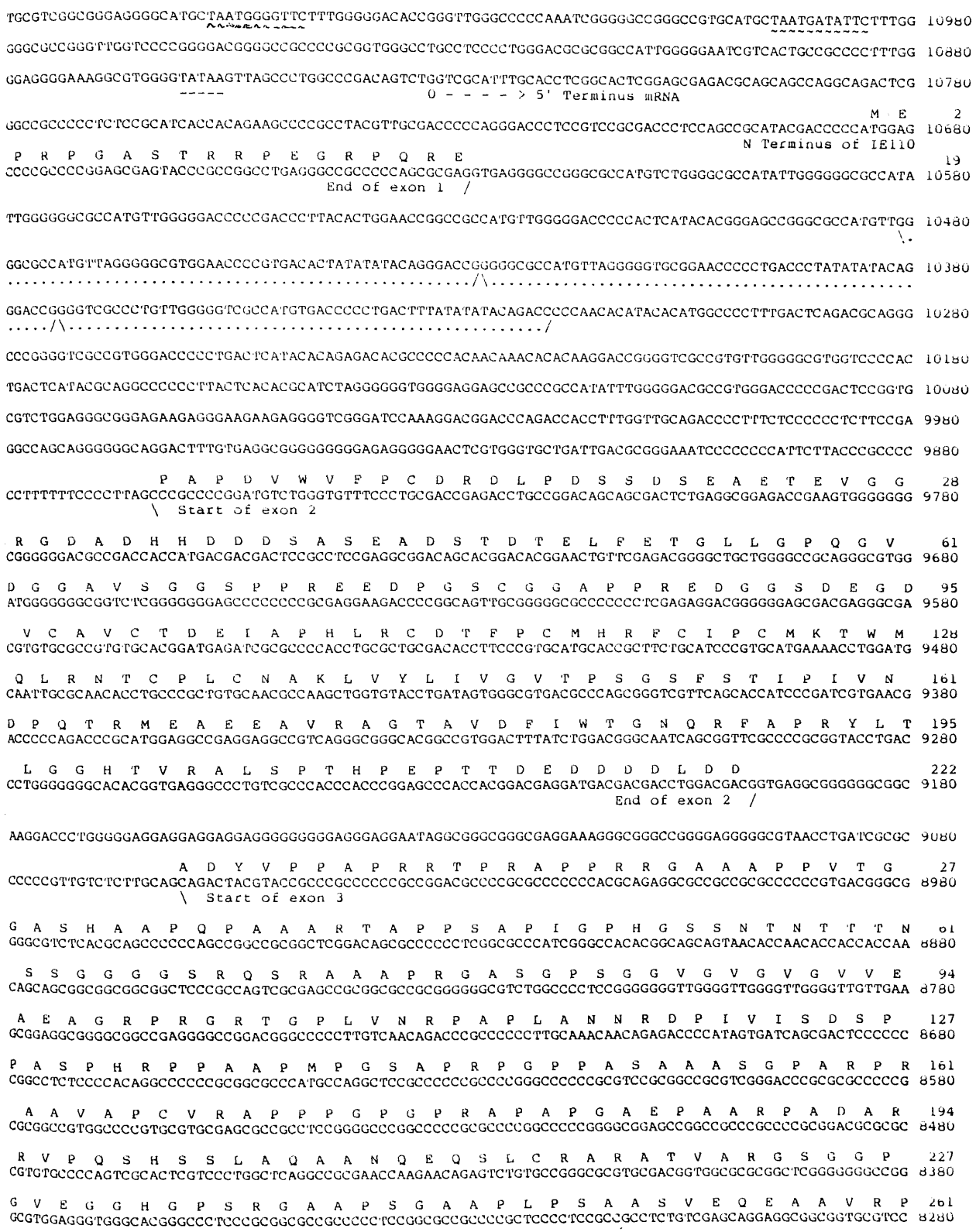
3.5.1 DNA sequence of IE gene 1

IE gene 1 encodes the protein IE110. The DNA and amino acid sequences are listed in figure 3.21. Numbering of the DNA sequence starts at the left BamHI site of the BamHI b fragment, as in the database. This numbering will be used for all genes in BamHI b. As the gene extends beyond the right BamHI site, data from this region, made available by D.J. McGeoch, have been included.

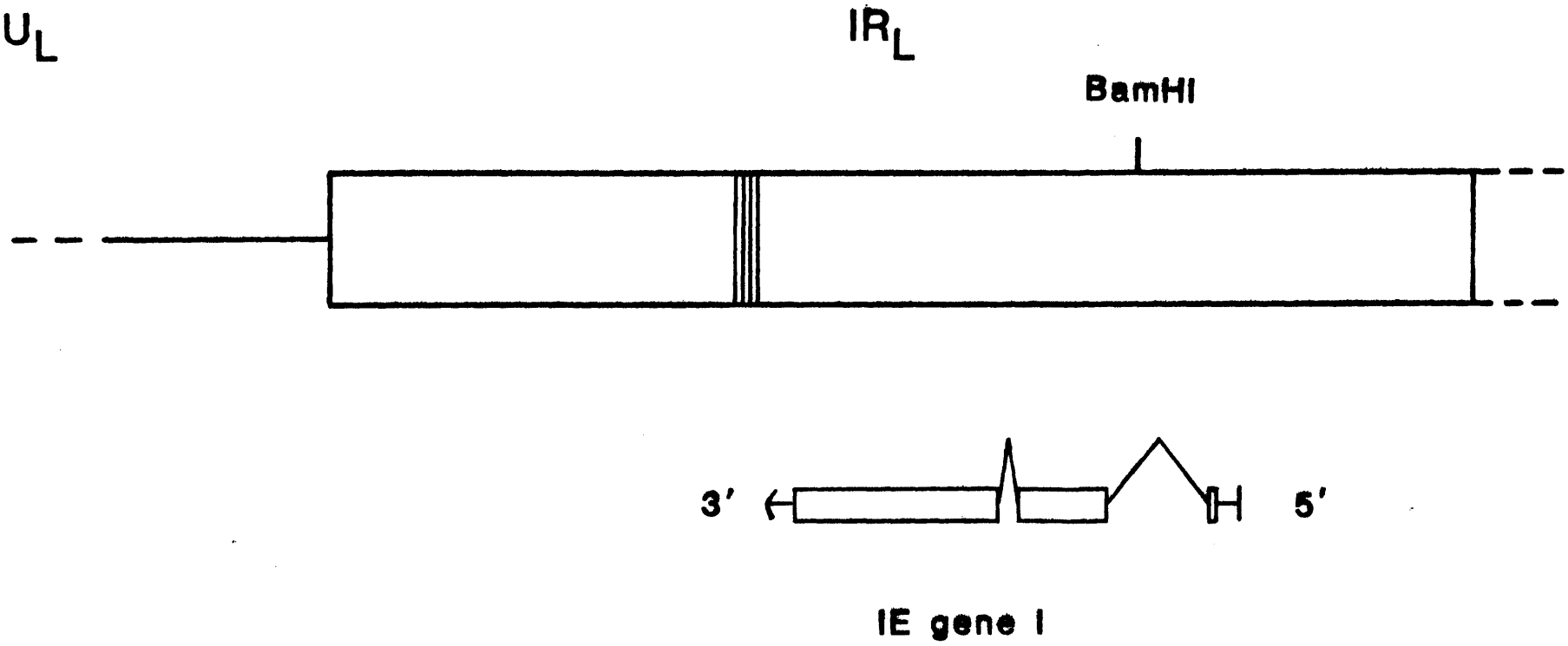
The 5' and 3' termini of the mRNA have been mapped to positions 10833 and 7248 respectively (Mackem and Roizman, 1982b; Rixon et al., 1984; Perry et al., submitted). A TATA box can be identified 27 bp upstream of the 5' terminus. The far upstream regulatory elements involved in coordinate induction of IE gene expression are located further upstream (Mackem and Roizman, 1982b,) These include palindromic G+C rich regions and the sequence TAATGATATTC at position 10995, which is an example of the TAATGARATTC sequence element, discussed in the introduction. Downstream of the coding region of the gene, at positions 7313 and 7273, are two copies of the sequence AATAAA, associated with polyadenylation (Fitzgerald and Shenk, 1980). At

Figure 3.21 DNA sequence of IE gene 1

The DNA and sequence of IE gene 1 and the proposed amino acid sequence of its encoded protein IE110 is given. The sequence is inverted, or complementary to the consensus sequence in figure 3.7, and is listed 5' to 3'. Numbering is from the left BamHI site, as in the database. The predicted amino acid sequence of IE110 is given as a single letter code. Proposed TATA box and polyadenylation signal sequences are underlined. The homologues to the TAATGARATTC sequence are indicated. Sets of tandem reiterations are marked as \...../.


0 - - - - > 5' Terminus mRNA
N Terminus of IE110
End of exon 1 /
End of exon 2 /
Start of exon 2
Start of exon 3

```
  R  K  R  R  G  S  G  Q  E  N  P  S  P  Q  S  T  R  P  P  L  A  P  A  G  A  K  R  A  A  T  H  P  P   294
GAGGAAGAGGCGCGGGTCGGGCCAGGAAAACCCCTCCCCCCAGTCCACGCGTCCCCCCCTCGCGCCGGCAGGGGCCAAGAGGGCGGCGACGCACCCCCCC 8180

  S  D  S  G  P  G  G  R  G  Q  G  G  P  G  T  P  L  T  S  S  A  A  S  A  S  S  S  S  A  S  S  S  S   327
TCCGACTCAGGGCCGGGGGGGCGCGGCCAGGGTGGGCCCGGGACCCCCCTGACGTCCTCGGCGGCCTCCGCCTCTTCCTCCTCTGCCTCTTCCTCCTCGG 8080

  A  P  T  P  A  G  A  A  S  S  A  A  G  A  A  S  S  S  A  S  A  S  S  G  G  A  V  G  A  L  G  G  R  Q   361
CCCCGACCCCCGCGGGGGCCGCCTCTTCCGCCGCCGGGGCCGCGTCCTCCTCCGCTTCCGCCTCCTCGGGCGGGGCCGTCGGTGCCCTGGGAGGGAGACA 7980

  E  E  T  S  L  G  P  R  A  A  S  G  P  R  G  P  R  K  C  A  R  K  T  R  H  A  E  T  S  G  A  V  P   394
AGAGGAAACCTCCCTCGGCCCCCGCGCTGCTTCTGGGCCGCGGGGGCCGAGGAAGTGTGCCCGGAAGACGCGCCACGCGGAGACTTCCGGGGCCGTCCCC 7880

  A  G  G  L  T  R  Y  L  P  I  S  G  V  S  S  V  V  A  L  S  P  Y  V  N  K  T  I  T  G  D  C  L  P   427
GCGGGCGGCCTCACGCGCTACCTGCCCATCTCGGGGGTCTCTAGCGTGGTCGCCCTGTCGCCTTACGTGAACAAGACTATCACGGGGGACTGCCTGCCCA 7780

  I  L  D  M  E  T  G  N  I  G  A  Y  V  V  L  V  D  Q  T  G  N  M  A  T  R  L  R  A  A  V  P  G  W  S   461
TCCTGGACATGGAGACGGGGAACATCGGGGCGTACGTGGTCCTGGTGGACCAGACGGGAAACATGGCGACCCGGCTGCGGGCCGCGGTCCCCGGCTGGAG 7680

  R  R  T  L  L  P  E  T  A  G  N  H  V  M  P  P  E  Y  P  T  A  P  A  S  E  W  N  S  L  W  M  T  P   494
CCGCCGCACCCTGCTCCCCGAGACCGCGGGTAACCACGTGATGCCCCCCGAGTACCCGACGGCCCCCGCGTCGGAGTGGAACAGCCTCTGGATGACCCCC 7580

  V  G  N  M  L  F  D  Q  G  T  L  V  G  A  L  D  F  R  S  L  R  S  R  H  P  W  S  G  E  Q  G  A  S   527
GTGGGGAACATGCTGTTCGACCAGGGCACCCTAGTGGGCGCCCTGGACTTCCGCAGCCTGCGGTCTCGGCACCCGTGGTCCGGGGAGCAGGGGGCGTCGA 7480

  T  R  D  E  G  K  Q  -                                                                               534
CCCGGGACGAGGGAAAACAATAAGGGACGCCCCCCGTGTTTGTGGGGAGGGGGGGGTCGGGCGCTGGGTGGTCTCTGGCCGCGCCCACTACACCAGCCAA 7380
 C terminus of IE110

TCCGTGTCGGGGAGGGGAAAAGTGAAAGACACGGGCACCACACACCAGCGGGTCTTTTGTGTTGGCCCTAATAAAAAAAAAACTCAGGGGATTTTTGCTGT 7280
                                                                                    ------
CTGTTGGGAAATAAAGGTTTACTTTTGTATCTTTTCCCTGTCTGTGTTGGATGTATCGCGGGGATGCGTGGGAGTGGGGGTGCGTGGGAGTGGGGGTGCG 7180
         ------      - - - - :          ........              ../\............../\............../\
               3' Terminus  mRNA

TGGGAGTGGGGGTGCGTGGGAGTGGGGGTGCGTGGGAGTGGGGGTGCGTGGGAGTGGGGGTGCGTGGGAGTGGGGGTGCGTGGGAGTGGGGGTGCGTGGG 7080
.............../\............../\............../\............../\............../\............../\....

AGTGGGGGTGCCATGTTGGGCAGGCTCTGGTGTTAACCACAGAGCCGCGGCCCGGGCTGCCTGACCACCGATCCCCGAAAGCATCCTGCCACTGGCATGG 6980
.........../
```



IE gene 1

Figure 3.22 Structure and location of IE gene 1

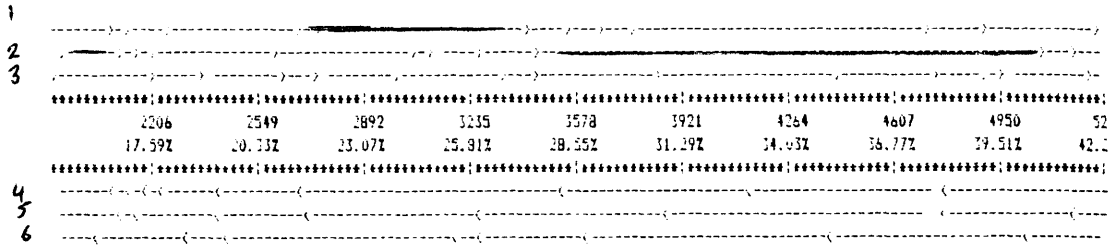
The location and orientation of IE gene 1 in IR_L is shown. The mapped 5' and 3' termini of the transcript are indicted. Proposed coding regions are shown as open boxes, non-coding regions as solid lines. The position of reiteration 3 and the BamHI site are indicated.

position 7237, 11 bp downstream of the 3' terminus of the mRNA, is the sequence TGTGTTGG, resembling the consensus YGTGTTYY required for efficient transcription termination (McLauchlan et al., 1985). Downstream of the gene are nine copies of a 16 bp tandemly reiterated sequence.

The 5' and 3' termini of the mRNA were mapped at a distance of 3585 bp apart, whereas the mRNA had a size estimated from its gel electrophoretic mobility of approximately 3 kb, including poly(A) tract (Watson et al., 1979). Examination of the DNA sequence between the mapped 3' and 5' termini did not show an ORF appropriate for IE110. Figure 3.23 shows all the ORFs in this sequence between the mapped 5' and 3' ends of the mRNA. These data suggested that the gene was spliced. Subsequent S1 nuclease and exonuclease analyses of the mRNA by F.J. Rixon confirmed that the gene is spliced and contains two introns (Perry et al., submitted). Figure 3.22 shows the structure and location of IE gene 1, in IR_L.

There are consensus splice donor and acceptor sequences at the mapped boundaries of both introns (Mount, 1982). From the DNA sequence, the first intron is calculated to be 764 bp in length. It contains three imperfect tandem copies of a 54 bp sequence, and a number of other imperfect repeat elements. The homopolymer tract containing a variable residue number, discussed previously, lies within this intron. The second intron is 135 bp long and is extremely purine rich.

The polypeptide coding sequences of IE110 were located by examination of the codon usage of the ORFs. Figure 3.24


1
2
3
2206 2549 2892 3235 3578 3921 4264 4607 4950 5293
17.59% 20.33% 23.07% 25.81% 28.55% 31.29% 34.03% 36.77% 39.51% 42.25%
4
5
6

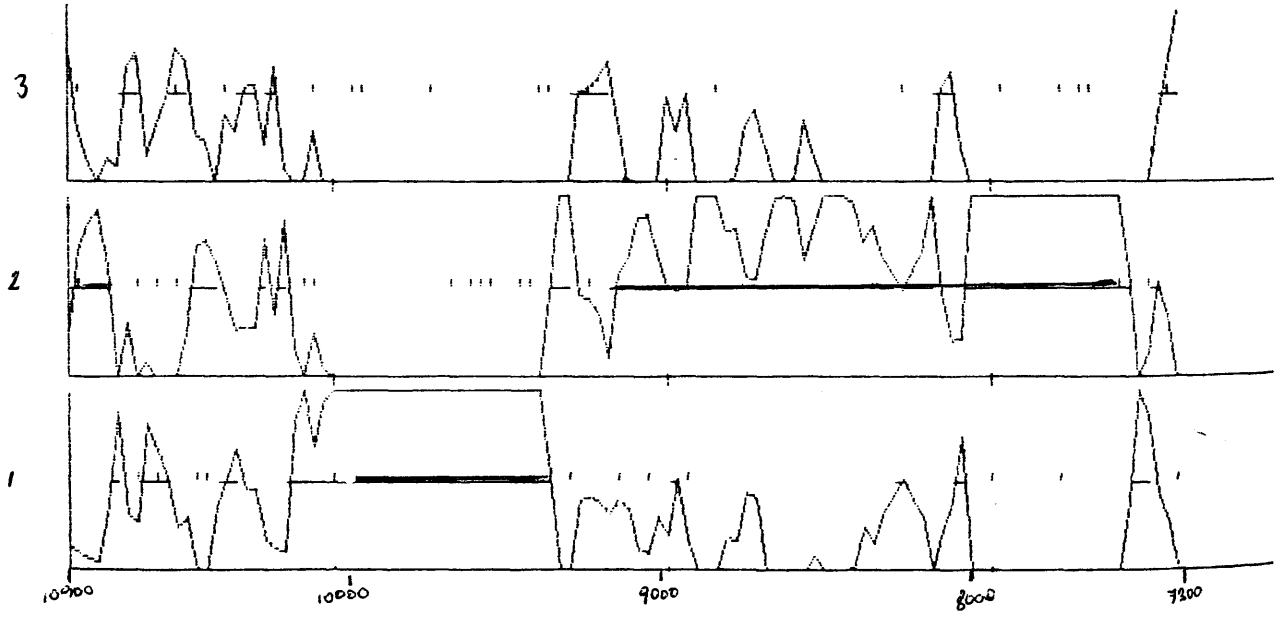

3
2
1
10900
10000
9000
8000
7200

Figure 3.23 Open reading frames in IE gene 1

All open reading frames in the sequence of IE gene 1 are shown. The sequence is represented in the 5' to 3' direction, from position 10901 in the standard sequence, and is in inverted orientation. The reading frames are numbered. Proposed polypeptide coding sequences are shown in red.

Figure 3.24 Evaluation of the codon usage of the reading frames in IE gene 1

The codon usage of the three reading frames on the rightward 5' to 3' strand of IE gene 1 where compared with the polypeptide coding region of IE gene 3. High scores represent regions with a similar codon usage. Short vertical lines on the central axis represent in-frame stop codons. Proposed polypeptide coding regions are shown in red. The x axis represents the numbered position in the DNA sequence, as in the standard orientation. The y axis represents a log probability function, with an arbitrary scale (Staden and McLachlan, 1982).

shows an evaluation of the codon usage of each reading frame in the gene. The coding regions of the gene for IE175 (McGeoch et al., 1986a) was used as the reference gene as it is also coded in a major repeat structure with a high G+C content. The reading frame of the coding sequences can be deduced from this analysis. A reading frame with a codon usage similar to that of IE175 achieves a high score as is illustrated in the figure. The introns appear as regions with none of the reading frames scoring well. Using this analysis and referring to the sequence, it was concluded that translation begins at the first ATG after the 5' terminus of the mRNA, at position 10685. This gives a 5' non-coding region of 148 residues. This ATG conforms to the initiation codon consensus CCRCCAUG of Kozak, 1984). All three exons are thus protein coding. Translation terminates at the TAA codon at position 7459, giving a protein 775 amino acids in length. The 3' untranslated region is 209 residues in length. The polypeptide coding region has a base composition of 75.4% G+C. This is a very high value, even for an HSV-1 gene. The effect of the base composition on coding potential is minimised to some extent by the high level of G and C residues in the redundant third position of the codons (see figure 3.18).

3.5.2 Amino acid sequence of IE110

The 775 amino acid polypeptide predicted from the DNA sequence has a M_r 78,452. This is considerably lower than the previous estimate of 110,000 obtained from gel electrophoretic mobility of the protein (Marsden et al., 1976). Apparent molecular weights estimated from mobility through polyacrylamide gels can be fairly inaccurate. The high Pro content of many HSV

Table 3.25 Predicted Amino Acid Composition of IE110

Res	No.	%	Res	No.	%
Ala	110	14.2	Leu	36	4.6
Arg	66	8.5	Lys	8	1.0
Asn	18	2.3	Met	11	1.4
Asp	39	5.0	Phe	9	1.2
Cys	14	1.8	Pro	103	13.3
Gln	20	2.6	Ser	78	10.1
Glu	36	4.6	Thr	48	6.2
Gly	94	12.1	Trp	7	0.9
His	15	1.9	Tyr	7	0.9
Ile	13	1.7	Val	43	5.5

Table 3.26 Codon usage catalogue of the IE110

TTT Phe	2	TCT Ser	12	TAT Tyr	0	TGT Cys	1
TTC Phe	7	TCC Ser	29	TAC Tyr	7	TGC Cys	13
TTA Leu	0	TCA Ser	1	TAA ---	1	TGA ---	0
TTG Leu	1	TCG Ser	17	TAG ---	0	TGG Trp	7
CTT Leu	2	CCT Pro	4	CAT His	1	CGT Arg	4
CTC Leu	6	CCC Pro	67	CAC His	14	CGC Arg	29
CTA Leu	1	CCA Pro	3	CAA Gln	4	CGA Arg	4
CTG Leu	26	CCG Pro	29	CAG Gln	16	CGG Arg	17
ATT Ile	0	ACT Thr	2	AAT Asn	1	AGT Ser	4
ATC Ile	11	ACC Thr	23	AAC Asn	17	AGC Ser	15
ATA Ile	2	ACA Thr	1	AAA Lys	2	AGA Arg	4
ATG Met	11	ACG Thr	22	AAG Lys	6	AGG Arg	8
GTT Val	5	GCT Ala	5	GAT Asp	4	GGT Gly	4
GTC Val	11	GCC Ala	54	GAC Asp	35	GGC Gly	33
GTA Val	1	GCA Ala	4	GAA Glu	7	GGA Gly	5
GTG Val	26	GCG Ala	47	GAG Glu	29	GGG Gly	52

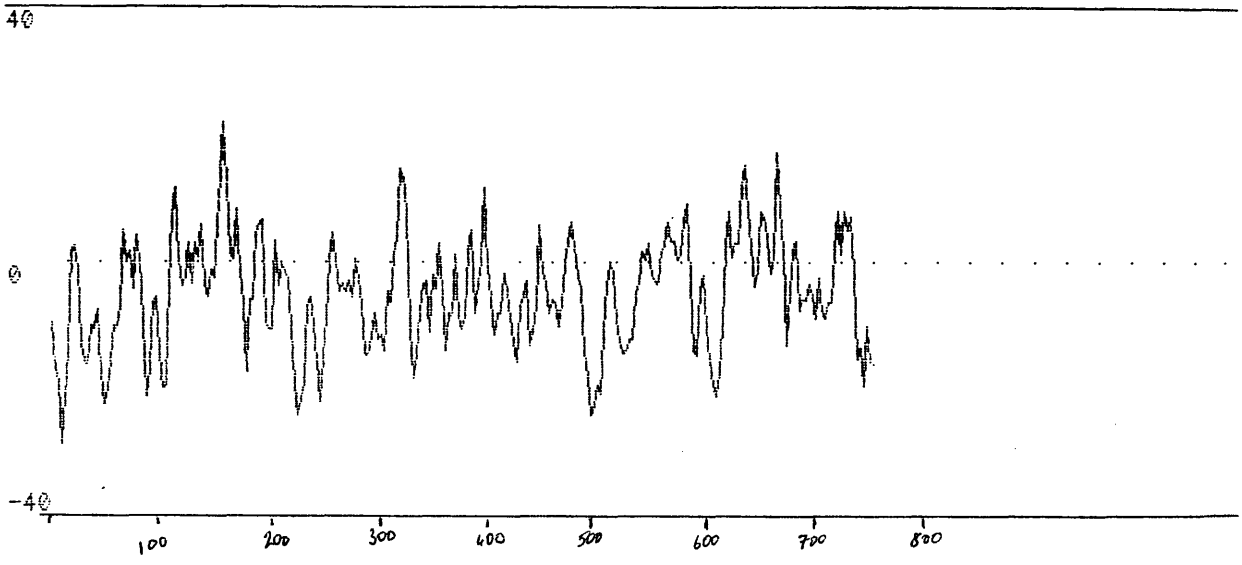

40
0
-40
100
200
300
400
500
600
700
800

Figure 3.27 Hydropathicity of the IE110 protein

Hydropathicity of the proposed amino acid sequence of IE110 is illustrated opposite. The predicted nature of the amino acid sequence is based on the parameters of Kyte and Doolittle (1982). The amino terminus is at the left of the figure, the carboxy terminus at the right. Higher scores represent hydrophobic regions.

The y axis represents hydropathicity, -40 to +40, summed over nine residues. The scale along the x axis is in 100 amino acids.

proteins, as well as post-translational modifications such as phosphorylation, may give an anomalous mobility to polypeptides.

The IE110 protein contains a small excess of basic over acidic residues (table 3.25). The five most common amino acids are Ala, Pro, Gly, Ser and Arg. Ala, Pro, Gly and Arg are amino acids translated from codons containing only G and C residues (table 3.26). Their prominence correlates with the extreme base composition of the gene. The distribution of these residues is not uniform: they are most abundant in the region beginning with the start of exon 3 (amino acid residue 242) and ending with residue 560. A further notable feature of this protein is the high Ser and Ala content between amino acid residues 554 and 594. The hydropathicity of the protein is illustrated in figure 3.27.

There is a Cys rich region between residues 99 and 156. Nine of the 14 Cys residues in IE110 are located in this 58 amino acid sequence. Many DNA binding proteins contain Cys rich regions (Miller et al., 1985; Berg, 1986). A common sequence element within these regions is a pair of Cys residues separated by two other amino acids. IE110 has three of these sequences, as shown in the sequence listing, figure 3.21.

Using the predicted amino acid sequence of IE110, an antiserum was prepared by M.C. Frame against a 12-mer oligopeptide with the amino acid sequence of the predicted carboxy terminus linked to a Tyr residue. The antiserum precipitated native IE110 (Perry et al., submitted), so validating the predicted amino acid sequence in this locality.

VZV
HSV-1
N
N
C
C

Figure 3.28 Relationship between HSV-1 IE110 and the 51K VZV protein encoded by gene 61

The relationship between the IE110 protein of HSV-1 and the 51K protein encoded by VZV gene 61, is shown opposite. Homology was calculated by the CINTHOM program (Pustell and Kafatos, 1982), adapted to allow for conservative amino acid changes, and displayed as a dot matrix using DIAG (P. Taylor, unpublished). The parameters used were as follows, range, 20; minimum value, 20; others, as default. The x axis represents the 51K VZV protein, the y axis the HSV-1 IE110 protein. The diagonal line represents amino acid sequences conserved between the two proteins.

```
                                    ----                ----          ----
          -                         -  -          -     -  -          -  -
IE110   99 CGGAPPREDGGSDEGD     VCAVCTDEIAPHLRCDTFPCMHRFCIPCMKTWMQLRNTCPLCNAKLVYLIVGVTPSG
                     ***    *     *  *        *    * ** * **  *    *      ****     *  *      *
VZV 61   1     MDTILAGGSGTSDASDNTCTICM STVSDLG KTMPCLHDFCFVCIRAWTSTSVQCPLC RCPVQSILHKIVSD
                                    -  -             -    -  -          -  -  -
                                    ----                ----          ----
```

Figure 3.29 Alignment of the conserved amino acid residues in HSV-1 IE110 and VZV 51K encoded by gene 61

The amino acid sequences in the conserved region were optimally aligned using the program HOMOL (P. Taylor, 1984). Default parameters were used. Identical residues are indicated with *. The Cys residues in both sequences are under/over-lined. The three sequences C X X C conserved between the two proteins are indicated.

A search was made to find a homologue to IE110 in proteins encoded by the alphaherpesvirus VZV (Davison and Scott, 1986). A predicted protein of VZV encoded by gene 61, with a M_r 50913, (hereafter referred to as 51K), showed some homology to IE110. A matrix plot showing the level of homology between the two proteins is shown in figure 3.28. The most highly conserved amino acid sequences have been aligned using the HOMOL program, and are shown in figure 3.29. The region conserved between the two proteins includes the Cys rich sequence. The majority of the Cys residues appear to have been conserved between the two proteins, suggesting they may play an essential role in function.

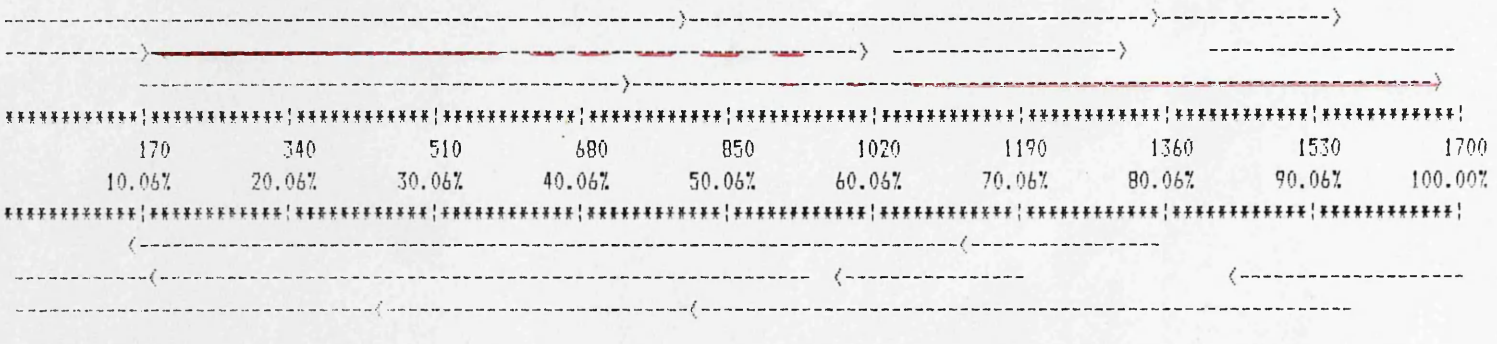
The amino acid sequence of IE110 was also compared to the sequence of the major IE protein of the betaherpesvirus HCMV, (Stenberg et al., 1984), but they did not show any homology. IE110 also did not show any homology to any polypeptides predicted from the complete genome sequence of the gammaherpesvirus EBV (Baer et al., 1984).

GENES IN THE U_L REGION OF BAMHI B

3.6 IE GENE 2

3.6.1 DNA sequence of IE gene 2

IE gene 2 encodes the protein IE63. This is the only gene previously mapped to U_L in the DNA sequence presented here. During the analysis of the sequence for IE gene 2, the reading frame believed to code for IE63 was seen to be blocked. Fine examination of the sequence of M13 clones did not resolve the problem. As


170 340 510 680 850 1020 1190 1360 1530 1700
10.06% 20.06% 30.06% 40.06% 50.06% 60.06% 70.06% 80.06% 90.06% 100.00%

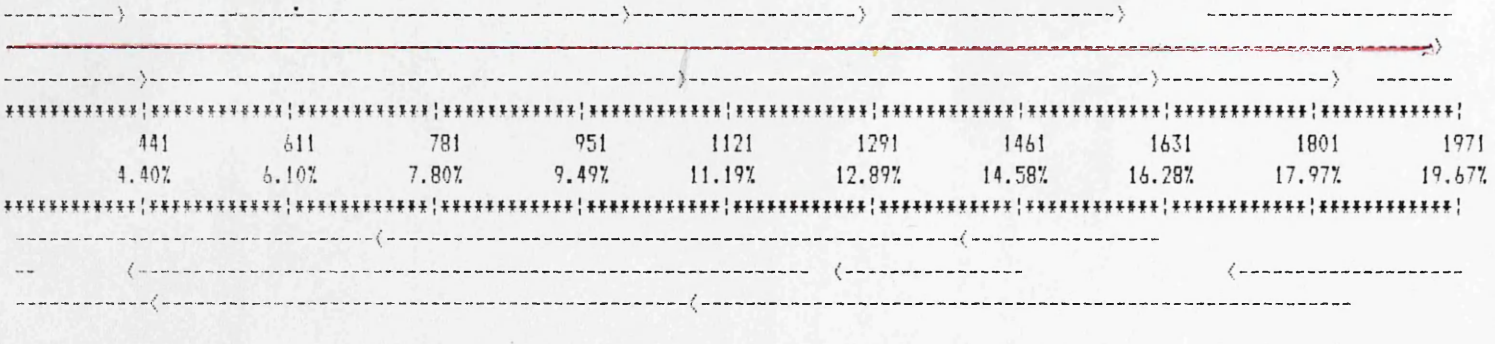
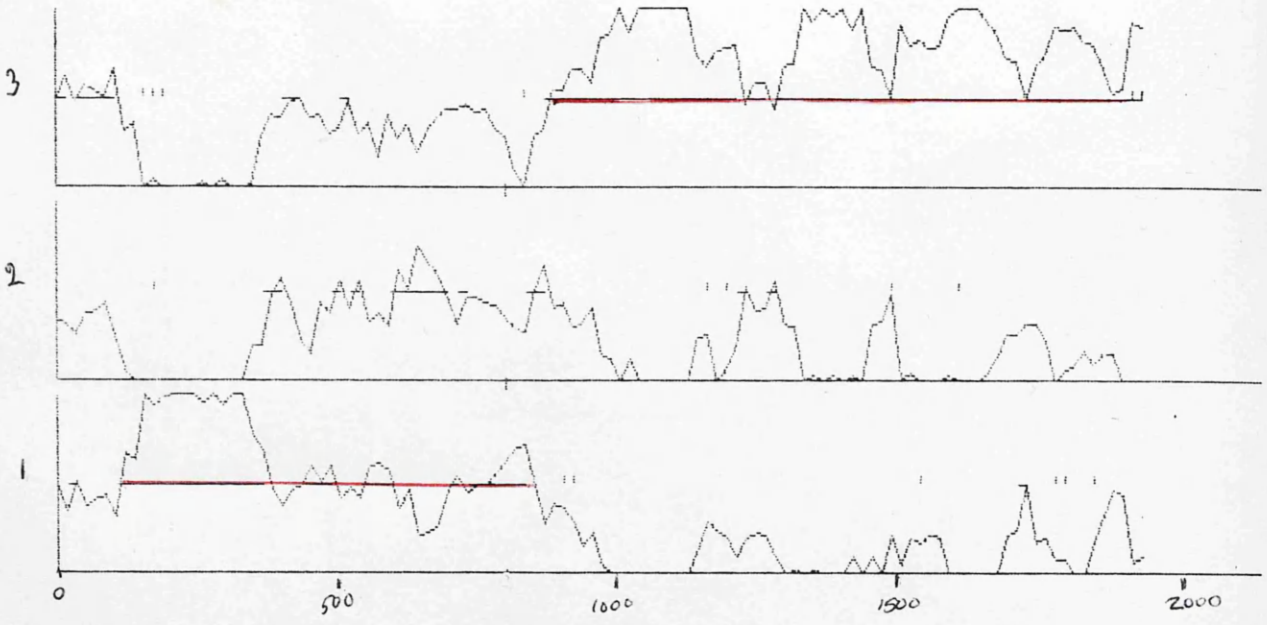

441 611 781 951 1121 1291 1461 1631 1801 1971
4.40% 6.10% 7.80% 9.49% 11.19% 12.89% 14.58% 16.28% 17.97% 19.67%

Figure 3.30 Open reading frames in IE gene 2

All open reading frames in the sequences of IE gene 2 determined from plasmids pGX48, shown above, and pGX190. The sequence is represented rightwards, in the 5' to 3' direction. The region displayed opposite lies between the mapped positions of the 5' and 3' termini of the mRNA.

The continuous red line through pGX190 frame 2 represents the predicted IE63 polypeptide coding sequence. This is marked with a dotted red line in pGX48.


3
2
1
0
500
1000
1500
2000

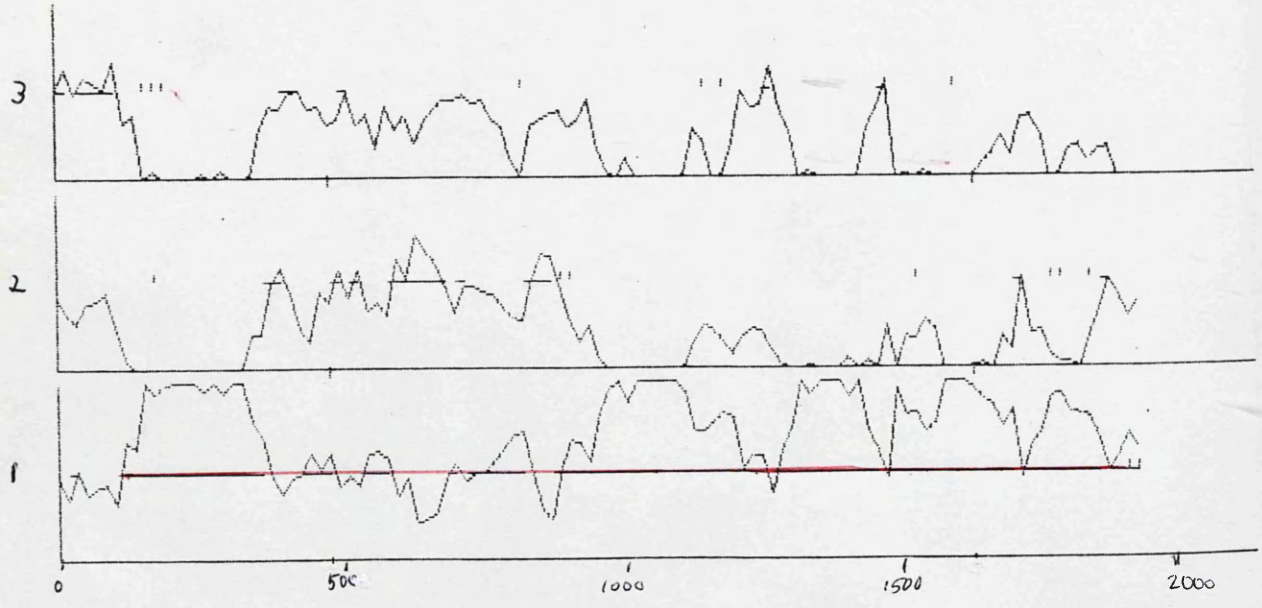

3
2
1
0
500
1000
1500
2000

Figure 3.31 Codon usage of IE gene 2

The codon usage evaluation of the three rightward reading frames in IE gene 2 are shown (Staden and McLachlan, 1982). The upper three reading frames are from the sequence of the plasmid pGX48, with a deletion in IE gene 2. The lower three reading frames are from the sequence of plasmid pGX190. The coding region of gene US3 was used as reference. Short vertical lines on the central axis represent in-frame stop codons.

The x axis represents the numbered position in the DNA sequence in the standard orientation. The y axis represents a log probability function, with an arbitrary scale (Staden and McLachlan, 1982). Proposed polypeptide coding regions are shown in red.

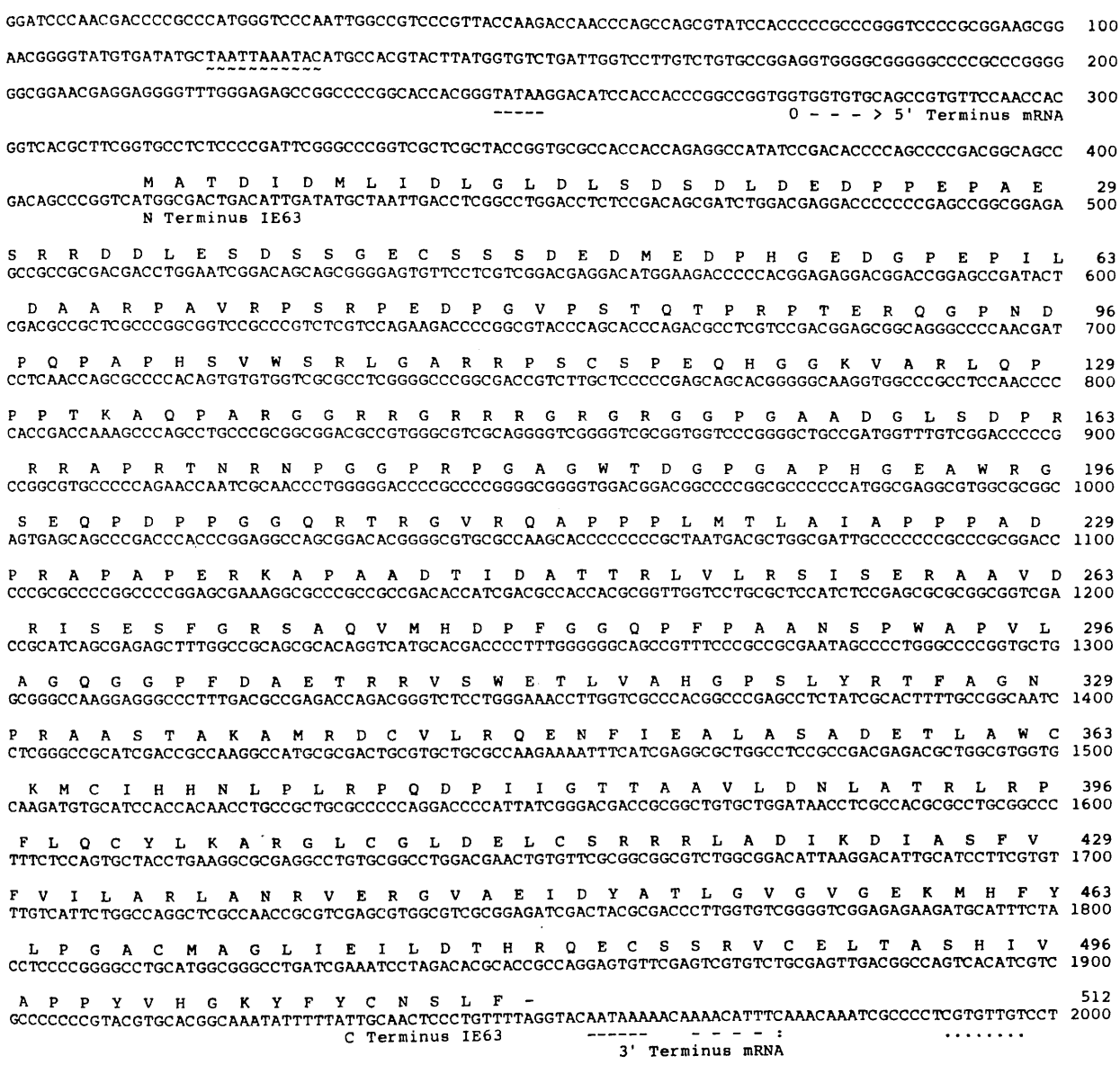
described above, the region was resequenced using another plasmid (pGX190) and a single residue deletion was found within the coding region of the gene in the plasmid pGX48, at position 1062. Figure 3.30 shows all the ORFs in the DNA sequence between the mapped 5' and 3' ends of the mRNA. The additional nucleotide in plasmid pGX190 opened a reading frame of appropriate size, which is proposed to code for IE63. The codon usage evaluation of IE gene 2 shown in figure 3.31 supports this interpretation of the sequence.

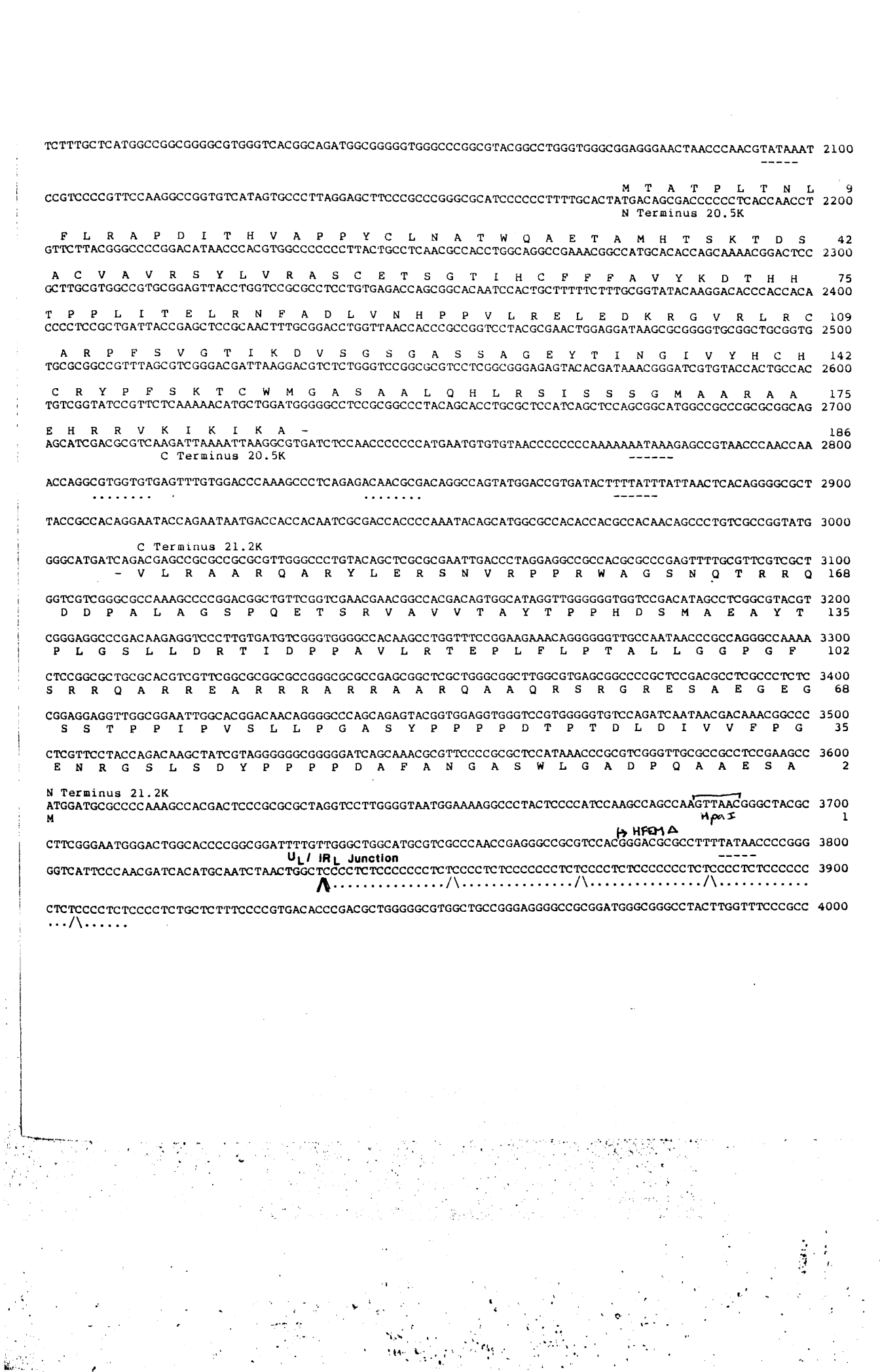
The DNA sequence of IE gene 2 is listed in figure 3.32. The amino acid sequence is given as the single letter code The 5' and 3' termini of the mRNA have been finely mapped to positions 275 and 1974 respectively (Whitton et al., 1983). A TATA box is located upstream at position 247. As with IE gene 1, the far upstream regulatory elements involved in transcription regulation have been identified (Mackem and Roizman, 1982; Preston et al., 1984). The sequence TAATAAATAC, representing the TAATGARATTC sequence element, is located between positions 120 and 130. Near the 3' terminus of the mRNA is the polyadenylation signal sequence AATAAA, at position 1956. Further downstream, at position 1990, is the sequence CGTGTTGT, showing homology to the consensus transcription termination sequence (McLauchlan et al., 1985). By mapping the mRNA, Whitton et al. (1983) concluded that IE gene 2 is unspliced.

The mRNA is believed to be translated from the first ATG at position 413, giving a 5' non-coding region of 138 nucleotides. This ATG conforms to Kozak's rules and initiates an ORF 515 amino acids in length. Translation terminates at TAG, at position 1949, leaving a 3' non-coding region 23 nucleotides in

Figure 3.32 The amino acid sequences of proteins encoded by genes in the U_L portion of the BamHI b fragment

The DNA sequence of BamHI b between positions 1 and 4000 is shown as the rightward 5' to 3' strand only, with numbering starting at the left BamHI site, as in the database. The predicted amino acid sequences of the proteins are given as a single letter code, rightward encoded proteins are shown above, and the leftward encoded protein below the corresponding DNA sequence. The mapped termini of the IE gene 2 mRNA are marked (0---->, 5' terminus; ----:, 3' terminus). Proposed TATA box and polyadenylation signal sequences are underlined. Homologues to the TAATGARATTC sequence and transcription termination (McLauchlan) sequence are also indicated. The IR_L/U_L junction is labelled. Copies of reiteration set 1 are labelled as \...../.


M A T D I D M L I D L G L D L S D S D L D E D P P E P A E
N Terminus IE63
0 - - - > 5' Terminus mRNA
C Terminus IE63
3' Terminus mRNA


N Terminus 20.5K
C Terminus 20.5K
C Terminus 21.2K
N Terminus 21.2K
HpaI
HFQM Δ
UL/ IRL Junction

length. The base composition of the coding region varies considerably along its length. The 5' terminus of the coding region is particularly G+C rich. The coding region of the amino terminal 300 residues is 73.7% G+C, and the remainder of the coding region 63.0% G+C.

3.6.2 Amino acid sequence of IE63

The predicted amino acid sequence of the protein has a M_r 55,376. This is lower than that estimated from its electrophoretic mobility (Marsden et al., 1976). Overall, the protein is slightly basic (table 3.33). There is some grouping of acidic residues near the amino terminus. The protein is quite hydrophilic, although it has a strongly hydrophobic region between residues 143 and 160 (figure 3.33). Again, amino acids translated from codons containing only G and C residues, such as Pro, Arg and Ala, are over-represented (table 3.35). These residues are distributed unevenly along the protein, with the lowest number near the carboxy terminus.

A search was made for a homologue to IE63 in VZV. The VZV protein with an apparent M_r 51540, encoded by gene 4 (Davison and Scott, 1986) hereafter referred to as 52 K, showed good homology (figure 3.36). This homology was most pronounced at the carboxy terminus, and negligible in the amino half of the protein. All three reading frames in this region, in each protein, were compared to ensure that the loss of homology was not due to an error resulting in a frameshift in either of the sequences. The aligned amino acid sequences of the two proteins in the homologous region are shown in figure 3.37. The considerable homology in this region

Table 3.33 Predicted Amino Acid Composition of IE63

Res.	No.	%	Res	No.	%
Ala	64	12.4	Leu	42	8.2
Arg	53	10.3	Lys	10	1.9
Asn	11	2.1	Met	9	1.7
Asp	32	6.2	Phe	12	2.3
Cys	14	2.7	Pro	55	10.7
Gln	13	2.5	Ser	36	7.0
Glu	23	4.5	Thr	25	4.9
Gly	41	8.0	Trp	5	1.0
His	13	2.5	Tyr	9	1.7
Ile	18	3.5	Val	30	5.8

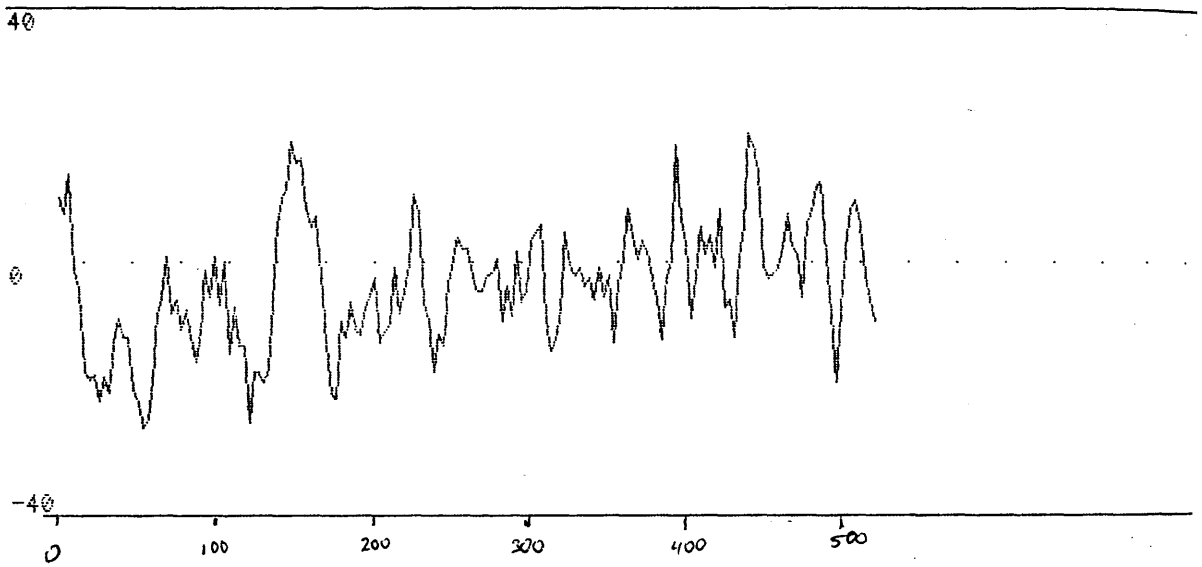

40
0
-40
0
100
200
300
400
500

Figure 3.34 Hydropathicity of IE63

Hydropathicity of the proposed amino acid sequence of IE63 is illustrated graphically. The nature of the amino acid sequences is based on the parameters of Kyte and Doolittle (1982). The amino terminus of the protein is at the left, carboxy terminus at the right side of the figure. Higher scores represent hydrophobic regions.

The y axis represents hydropathicity, -40 to +40, summed over nine residues. The scale along the x axis is in 100 amino acids.

Table 3.35 Codon usage catalogue of the IE63

TTT Phe	9	TCT Ser	0	TAT Tyr	3	TGT Cys	5
TTC Phe	3	TCC Ser	10	TAC Tyr	6	TGC Cys	9
TTA Leu	0	TCA Ser	0	TAA ---	0	TGA ---	0
TTG Leu	3	TCG Ser	8	TAG ---	1	TGG Trp	5
CTT Leu	2	CCT Pro	2	CAT His	1	CGT Arg	5
CTC Leu	11	CCC Pro	32	CAC His	12	CGC Arg	20
CTA Leu	3	CCA Pro	4	CAA Gln	3	CGA Arg	7
CTG Leu	23	CCG Pro	17	CAG Gln	10	CGG Arg	14
ATT Ile	7	ACT Thr	2	AAT Asn	4	AGT Ser	3
ATC Ile	11	ACC Thr	11	AAC Asn	7	AGC Ser	15
ATA Ile	0	ACA Thr	1	AAA Lys	1	AGA Arg	4
ATG Met	9	ACG Thr	11	AAG Lys	9	AGG Arg	3
GTT Val	1	GCT Ala	2	GAT Asp	3	GGT Gly	3
GTC Val	19	GCC Ala	33	GAC Asp	29	GGC Gly	17
GTA Val	0	GCA Ala	5	GAA Glu	5	GGA Gly	5
GTG Val	10	GCG Ala	24	GAG Glu	18	GGG Gly	16

VZV
N
C
HSV-1
C

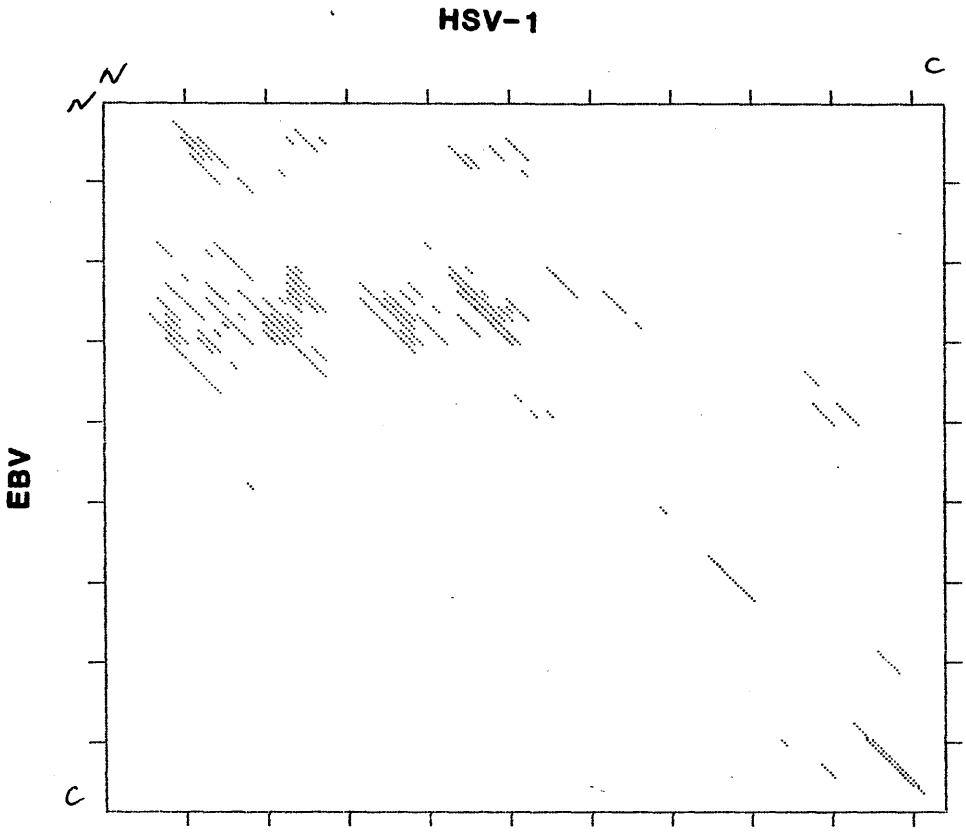
Figure 3.36 Relationship between HSV-1 IE63 and VZV gene 4 protein

The relationship between the IE63 protein of HSV-1 and the protein encoded by VZV gene 4, calculated by the CINTHOM program (Pustell and Kafatos, 1982), adapted to allow for conservative amino acid changes, and displayed as a dot matrix using DIAG (P. Taylor, unpublished) is shown opposite. The parameters used were as follows, range, 12; minimum value, 20; others, as default. The x axis represents the VZV protein encoded by gene 4, the y axis the HSV-1 IE63K protein. The diagonal line represents amino acid sequences conserved between the two proteins.

```
300 AGQGGPFDAETRRVSWETLVAHGPSLYRTFAGNPRAASTAKAMRDCVLRQENFIEALASADETLAWCKMCIHHN
      ** *       **  *   *  *****          * * *  * * *  **** **** * * **    *
232 WASGGCFPGIKQNTSWPELMLYGHELYRTFESYKMDSRIARALRERVIRGESLIEALESADELLTWIKMLAAKN

374 LPLRPQDPIIGTTAAVLDNLATRLRPFLQCYLKAR    GLCGLDELCSRRRLADIKDIASFVFVILARLANRV
    **     **  *    * **   * **  * *  *    *   * **    *  **  *    **  ** ** *
306 LPIYTNNPIVATSKSLLENLKLKLGPFVRCLLLNRDNDLGSRTLPELLRQQRFSDITCITTYMFVMIARIANIV

444 ERGVAEIDYATLGVGVGEKMHFYLPGACMAGLIEILDTHRQECSSRVCEL TASHIVAPPYVHGKYFYC NSLF
     **    *      *      * ** * **  * * **  **    * * *           **** *
380 VRGSKFVEYDDISCNV QVLQEYTPGSCLAGVLEALITHQRECGRVECTLSTWAGHLSDARPYGKYFKCSTFNC
```

Figure 3.37 Homology between the carboxy amino acid sequences of the HSV-1 IE63 and 52K gene 4 proteins

The amino acid sequences of the carboxy regions of the proteins are optimally aligned using the program HOMOL (Taylor, 1984). Default parameters were used. Identical residues are marked with an *. Blanks represent gaps introduced into the sequence to optimise alignment. The amino acid sequence of IE63 is listed above the 52K sequence.




```
401 LQCYLKARGLCGLDELCSRRRLADIKDIASFVFVILARLANRVERGVAEIDYAT
     *        ** *                  *    *  *        *   *
350 KQ       PLCLL                  AAYAAVAPAYINANCRRRHDE

455 LGVGVGEKMHFYLPGACMAGLIEILDTHRQECSSRVCELTASHIVAP PYVHGK
        *     * **      * *   **   * * *       * *     *
378  VEFLGHYIKNYNPGTLSSLLTEAVETHTRDCRSASCSRLVRAILSPGTGSLGL

508 YFYCNSLF
     *
431 FFVPGLNQ
```

Figure 3.38a Relationship between HSV-1 IE63 and EBV protein encoded by BMLF1

The relationship between the IE63 protein of HSV-1 and the protein encoded by EBV BMLF1, calculated by the CINTHOM program (Pustell and Kafatos, 1982), adapted to allow for conservative amino acid changes, and displayed as a dot matrix using DIAG (P. Taylor, unpublished) is shown opposite. The parameters used were as follows, range, 20; minimum value, 20; others, as default. The x axis represents the HSV-1 IE63K protein, the y axis the EBV protein encoded by BMLF1. The diagonal line represents amino acid sequences conserved between the two proteins.

Figure 3.38b Homology between the carboxy amino acid sequences of the HSV-1 IE63 and EBV BMLF1 protein

The amino acid sequences of the carboxy regions of the proteins are optimally aligned using the program HOMOL (Taylor, 1984). Default parameters were used. Identical residues are marked with an *. Blanks represent gaps introduced into the sequence to optimise alignment. The amino acid sequence of IE63 is listed above the EBV BMLF1 amino acid sequence.

Table 3.39 Codon usage catalogue of 20.5K

TTT Phe	4	TCT Ser	1	TAT Tyr	1	TGT Cys	3
TTC Phe	3	TCC Ser	7	TAC Tyr	5	TGC Cys	5
TTA Leu	1	TCA Ser	1	TAA ---	0	TGA ---	1
TTG Leu	0	TCG Ser	1	TAG ---	0	TGG Trp	2
CTT Leu	0	CCT Pro	2	CAT His	1	CGT Arg	0
CTC Leu	3	CCC Pro	2	CAC His	9	CGC Arg	7
CTA Leu	2	CCA Pro	0	CAA Gln	0	CGA Arg	1
CTG Leu	7	CCG Pro	6	CAG Gln	2	CGG Arg	6
ATT Ile	4	ACT Thr	0	AAT Asn	0	AGT Ser	1
ATC Ile	3	ACC Thr	9	AAC Asn	5	AGC Ser	5
ATA Ile	2	ACA Thr	3	AAA Lys	3	AGA Arg	0
ATG Met	4	ACG Thr	4	AAG Lys	5	AGG Arg	0
GTT Val	1	GCT Ala	1	GAT Asp	1	GGT Gly	0
GTC Val	5	GCC Ala	11	GAC Asp	5	GGC Gly	3
GTA Val	1	GCA Ala	1	GAA Glu	2	GGA Gly	1
GTG Val	5	GCG Ala	9	GAG Glu	5	GGG Gly	5

(from residue 300 to the carboxy terminus of IE63), is reflected in that 37% of amino acids are identical between the amino acid sequences of the two proteins.

A protein of M_r 51,347, coded by the gene BMLF1, as predicted from the DNA sequence of EBV (Baer et al., 1984), also shows some homology to IE63 (figure 3.38a). Again most conservation of amino acid sequences was observed at the carboxy terminus of the proteins. The level of homology between the two amino acid sequences in the conserved region is not very high, at 20%, as shown in figure 3.38b.

3.7 U_L GENE ENCODING 20.5K

This gene is proposed to lie on the same strand as and immediately downstream of IE gene 2. The amino acid sequence of the 20.5K protein is shown with the DNA sequence in figure 3.32. A good TATA box is located at position 2094. A polyadenylation signal is located at position 2777. Downstream of the polyadenylation signal is a G+T rich sequence resembling the McLauchlan consensus (1985).

Translation is considered to start at the first ATG codon, position 2175. Termination of translation of the ORF is at TGA, position 2733. The coding region is 62.4% G+C, three percentage points lower than the average of the region. As in the previous genes discussed, the bias towards a G or C residue in the third redundant position of the codon, diminishes the effect of the base composition on the coding potential. Other than the high Ala content (11.8%), the amino acids translated from G+C only codons are not over-represented (table 3.39). The 186 amino acid

Table 3.40 Predicted Amino Acid Composition of 20.5K

Res	No.	%	Res	No.	%
Ala	22	11.8	Leu	13	7.0
Arg	14	7.5	Lys	8	4.3
Asn	5	2.7	Met	4	2.2
Asp	6	3.2	Phe	7	3.8
Cys	8	4.3	Pro	10	5.4
Gln	2	1.1	Ser	16	8.6
Glu	7	3.8	Thr	16	8.6
Gly	9	4.8	Trp	2	1.1
His	10	5.4	Tyr	6	3.2
Ile	9	4.8	Val	12	6.5

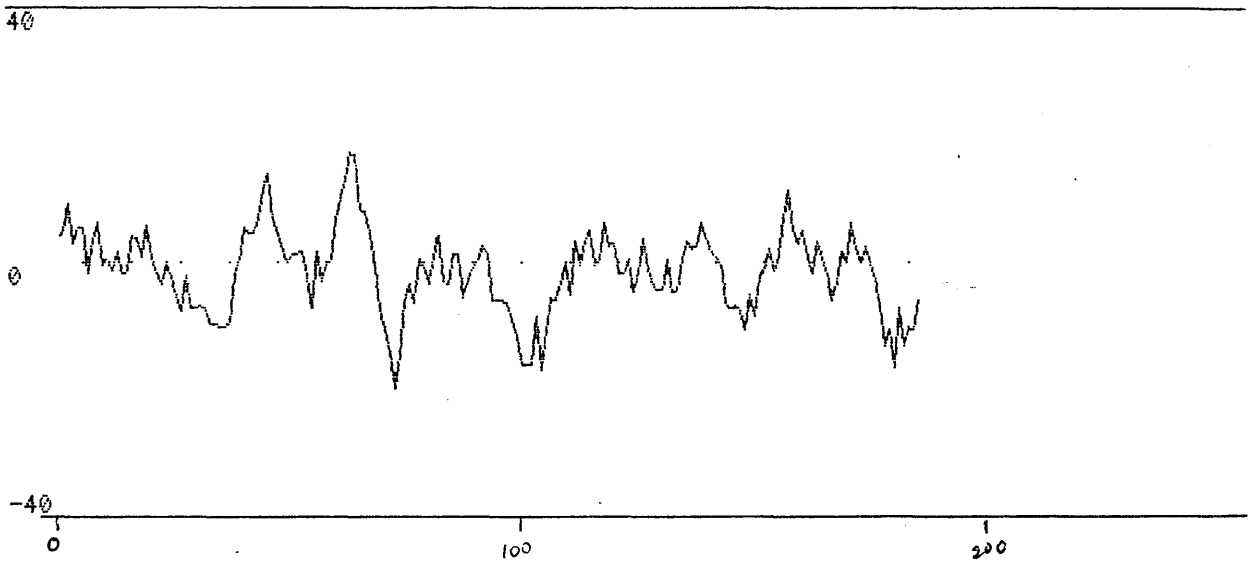
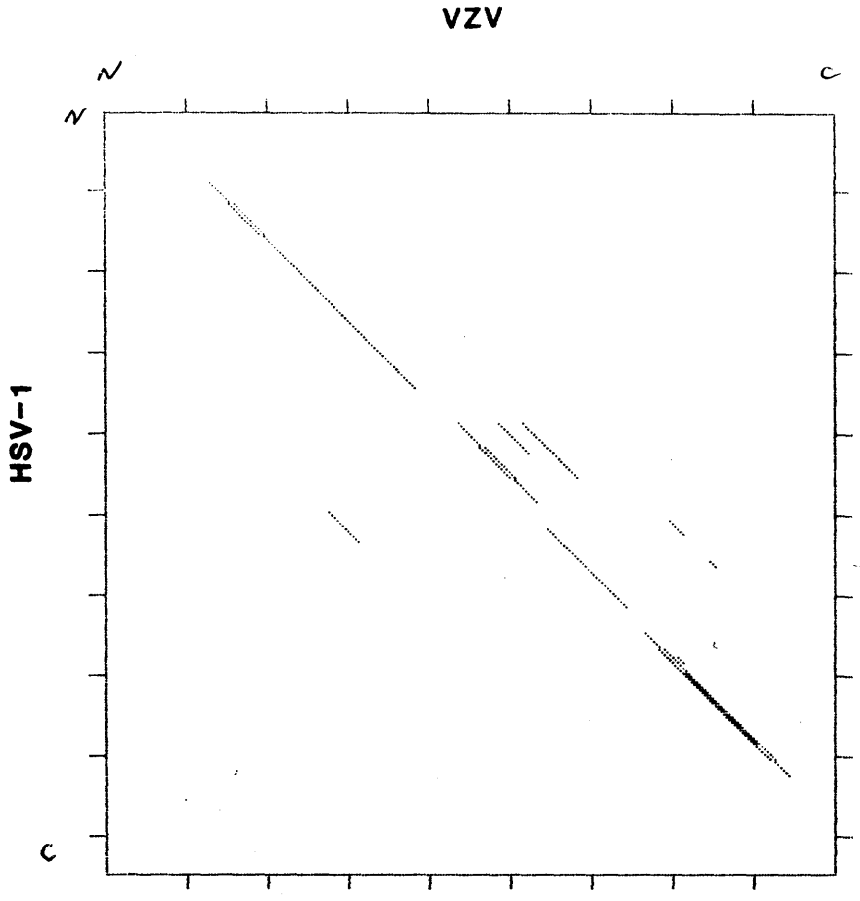

40
0
-40
0
100
200

Figure 3.41 Hydropathicity of the 20.5K protein

Hydropathicity of the amino acid sequence of the predicted 20.5K protein is illustrated graphically. The nature of the protein is predicted by the parameters of Kyte and Doolittle (1982). The amino terminus is at the left, carboxy terminus at the right side of the figure. Higher scores represent hydrophobic regions.

The y axis represents hydropathicity, -40 to +40, summed over nine residues. The scale along the x axis is in 100 amino acids.




```
  1                MTATPLTNLFLRAPDITHVAPPYCLNATWQAETAMHTSKTD
                                     * * **  *  **       *
  1 MDTTGASESSQPIRVNLKPDPLASFTQVIPPLALETTWTCPANSHAPTPS

 42 SACVAVRSYLVRASCETSGTIHCF FFAVYKDTHHTPPLITELRNFADLV
          *    ** *       * *         *     *    *  *  *
 51 PLYGVKRLCALRATCGRADDLHAFLIGLGRRDKPSESPMYVDLQPFCSLL

 91 NHPPVLRELEDKRGVRLRCARPFSVGTIKDVSGSGASSAGEYTINGIVYH
    *     * *           *  *** *       **           * **
101 N   SQRLLPEMANYNTLCDAPFSAATQQMMLESG   QLGVHLAAIGYH

141 CHCRYPFSKTCWMGASAALQHLRSISSSGMAARAAEHRRVKIKIKA
    ***  ***  ** *** *  *        *  ****
145 CHCKSPFSAECWTGASEAYDH    VVCGGKARAA       VGGL
```

Figure 3.42 Relationship between the HSV-1 20.5K and VZV 19.5K protein encoded by gene 3

The relationship between the 20.5K protein of HSV-1 and the protein encoded by VZV gene 3, calculated by the CINTHOM program (Pustell and Kafatos, 1982), adapted to allow for conservative amino acid changes, and displayed as a dot matrix using DIAG (P. Taylor, unpublished) is shown opposite. The parameters used were as follows, range, 12; minimum value, 20; others, as default. The x axis represents the 19.5K VZV protein, the y axis the HSV-1 20.5K protein. The diagonal line represents amino acid sequences conserved between the two proteins.

Figure 3.43 Homology between the amino acid sequences of the HSV-1 20.5K protein and the 19K protein encoded by VZV gene 3

The amino acid sequences of the HSV-1 20.5K and the VZV 19K were optimally aligned using the program HOMOL (Taylor, 1984). Default parameters were used. Identical residues are marked with an *. Blanks represent gaps introduced into the sequence to optimise alignment. The amino acid sequence of the HSV-1 20.5K protein is listed above the VZV 19K amino acid sequence.

Table 3.44 Predicted Amino Acid Composition of 21.2K

Res	No.	%	Res	No.	%
Ala	28	14.2	Leu	16	8.1
Arg	25	12.7	Lys	0	0.0
Asn	4	2.0	Met	2	1.0
Asp	11	5.6	Phe	4	2.0
Cys	0	0.0	Pro	27	13.7
Gln	8	4.1	Ser	17	8.6
Glu	10	5.1	Thr	11	5.6
Gly	14	7.1	Trp	2	1.0
His	1	0.5	Tyr	5	2.5
Ile	3	1.5	Val	9	4.6

protein has a M_r 20,491. The protein has twice the number of basic as acidic residues (table 3.40). The hydropathicity of the protein is shown in figure 3.41

A homologue to the 20.5K protein has been found in VZV. The M_r 19,149 VZV protein (19.1K) coded by gene 3 (Davison and Scott, 1986) is of similar size and shows homology throughout its length (figure 3.42). The alignment of the amino acid sequences of the two proteins in figure 3.43 illustrates the level of conservation, which is greatest near to the carboxy terminus. 25% of amino acids are conserved between the two proteins. No ORF encoding a homologue to the 20.5K protein was found in the EBV genome. No data exist on the function of the protein.

3.8 U_L GENE ENCODING 21.2K

This gene is proposed to lie between IR_L and the gene for 20.5K, described above, which lies on the opposite strand. The DNA and amino acid sequences are listed in figure 3.32. A good TATA box is situated aproximately 40 bp from the U_L/IR_L junction at position 3793. The polyadenylation signal for the mRNA is located at position 2880. A sequence resembling the transcription termination sequence (McLauchlan et al., 1985) lies at position 2849.

The ATG considered to initiate translation is at position 3602. Translation terminates at TGA, position 3011. The protein is 197 amino acids in length, and has a M_r 21,182. The amino acid content and codon usage reflect the 67.8% G+C content of the coding region of the gene (tables 3.44 and 3.45). For example the Ala, Pro and Arg levels are high at 14.2%, 13.7%

Table 3.45 Codon usage catalogue of 21.2K

TTT Phe	4	TCT Ser	2	TAT Tyr	1	TGT Cys	0
TTC Phe	0	TCC Ser	4	TAC Tyr	4	TGC Cys	0
TTA Leu	2	TCA Ser	1	TAA ---	0	TGA ---	1
TTG Leu	5	TCG Ser	7	TAG ---	0	TGG Trp	2
CTT Leu	2	CCT Pro	4	CAT His	0	CGT Arg	2
CTC Leu	3	CCC Pro	10	CAC His	1	CGC Arg	8
CTA Leu	0	CCA Pro	6	CAA Gln	4	CGA Arg	4
CTG Leu	4	CCG Pro	7	CAG Gln	4	CGG Arg	6
ATT Ile	2	ACT Thr	1	AAT Asn	1	AGT Ser	1
ATC Ile	1	ACC Thr	6	AAC Asn	3	AGC Ser	2
ATA Ile	0	ACA Thr	1	AAA Lys	0	AGA Arg	0
ATG Met	2	ACG Thr	3	AAG Lys	0	AGG Arg	5
GTT Val	2	GCT Ala	6	GAT Asp	3	GGT Gly	2
GTC Val	4	GCC Ala	10	GAC Asp	8	GGC Gly	6
GTA Val	0	GCA Ala	1	GAA Glu	3	GGA Gly	1
GTG Val	3	GCG Ala	11	GAG Glu	7	GGG Gly	5

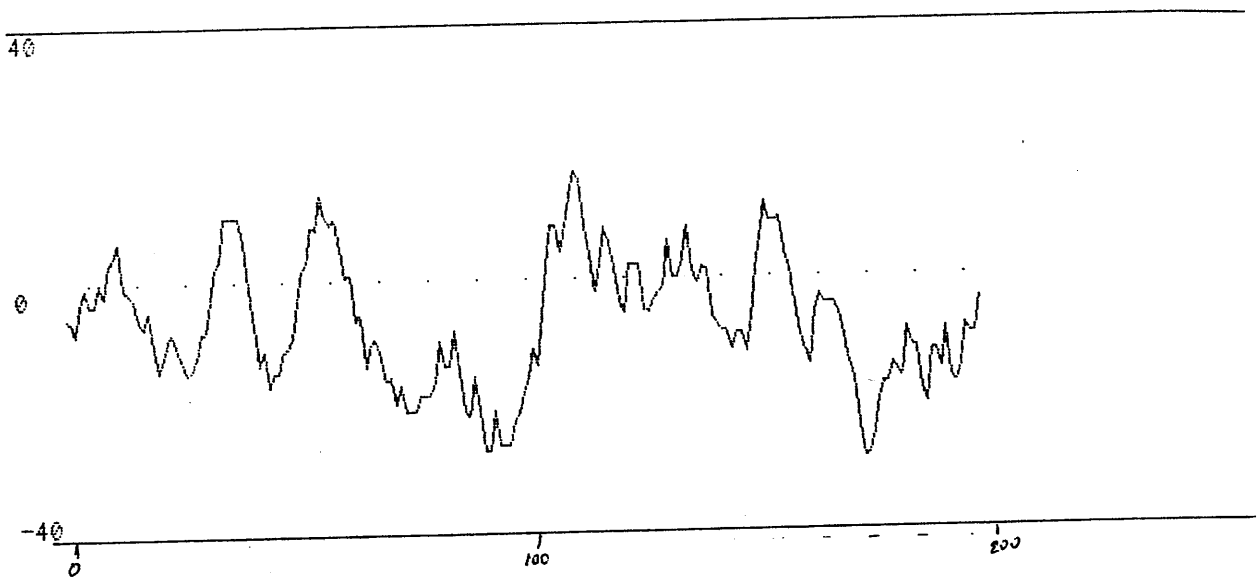

40
0
-40
0
100
200

Figure 3.46 Hydropathicity of the 21.2K protein

Hydropathicity of the amino acid sequence of the predicted 21.2K protein is illustrated graphically. The predicted nature of the amino acid sequence is based on the paramenters of Kyte and Doolittle (1982). The amino terminus of the protein is at the left, carboxy terminus at the right side of the figure. Higher scores represent hydrophobic regions. A * residue window length was used.

The y axis represents hydropathicity, -40 to +40, summed over nine residues. The scale along the x axis is in 100 amino acids.

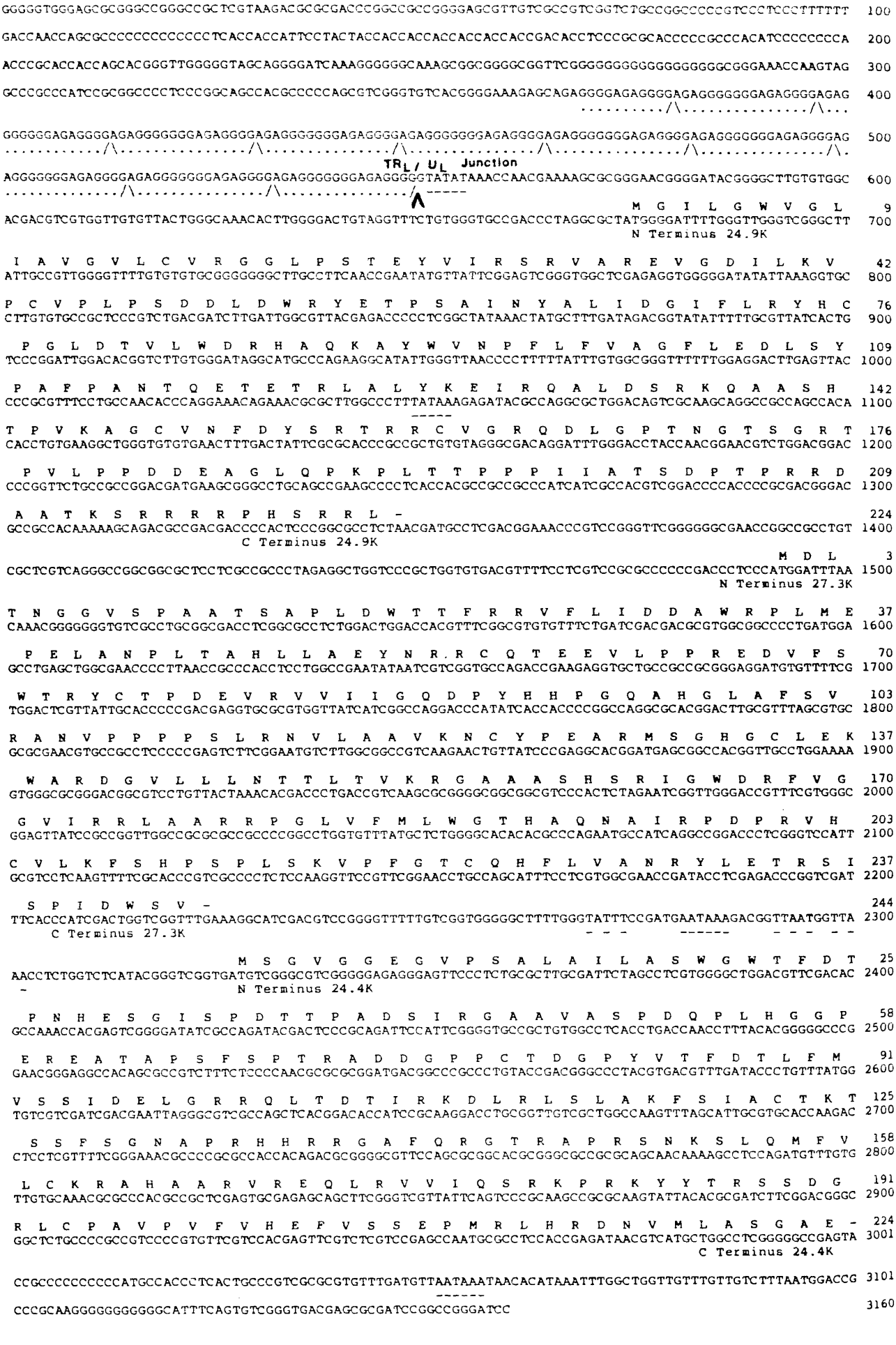

TRL / UL Junction
C Terminus 24.9K
N Terminus 27.3K
C Terminus 27.3K
N Terminus 24.4K
C Terminus 24.4K

Figure 3.47 The amino acid sequences of proteins encoded by genes in the U_L portion of BamHI e

The DNA sequence is shown as the rightward 5' to 3' strand only, with numbering starting at the SmaI site, as in the database. The predicted amino acid sequences of the proteins are given as a single letter code, above the corresponding DNA sequence. Proposed TATA box and polyadenylation signal sequences are underlined. The TR_L/U_L junction is labelled. Copies of reiteration set 1 are labelled as \...../.

Table 3.48 Codon usage catalogue of 24.9K

TTT Phe	6	TCT Ser	2	TAT Tyr	6	TGT Cys	5
TTC Phe	0	TCC Ser	1	TAC Tyr	2	TGC Cys	0
TTA Leu	2	TCA Ser	1	TAA ---	1	TGA ---	0
TTG Leu	11	TCG Ser	3	TAG ---	0	TGG Trp	4
CTT Leu	3	CCT Pro	5	CAT His	1	CGT Arg	2
CTC Leu	3	CCC Pro	8	CAC His	3	CGC Arg	8
CTA Leu	0	CCA Pro	0	CAA Gln	0	CGA Arg	5
CTG Leu	3	CCG Pro	9	CAG Gln	6	CGG Arg	6
ATT Ile	3	ACT Thr	0	AAT Asn	0	AGT Ser	3
ATC Ile	2	ACC Thr	8	AAC Asn	5	AGC Ser	2
ATA Ile	5	ACA Thr	3	AAA Lys	2	AGA Arg	1
ATG Met	1	ACG Thr	5	AAG Lys	5	AGG Arg	1
GTT Val	5	GCT Ala	4	GAT Asp	6	GGT Gly	3
GTC Val	2	GCC Ala	9	GAC Asp	9	GGC Gly	2
GTA Val	1	GCA Ala	1	GAA Glu	4	GGA Gly	4
GTG Val	8	GCG Ala	4	GAG Glu	4	GGG Gly	7

and 12.7% respectively. There is a slight excess of basic over acidic residues. Near the middle of the protein is a strongly hydrophilic region (figure 3.46).

A search was made for homologues to the 21.2K protein in VZV and EBV. None was found.

GENES IN THE U_L REGION OF BAMHI E

3.9 GENE UL1 ENCODING 24.9K

The numbering of the DNA sequence for the following three genes is from the SmaI site in BamHI e, as in the database. The DNA sequence of the gene encoding the 24.9K protein is listed in figure 3.47. The TATA box of this gene lies directly adjacent to the TR_L/U_L junction, at position 551. Any upstream promoter region must therefore lie within TR_L. The mRNA has a long 3' non-coding region, the polyadenylation signal is located at position 2280. The mRNA of this gene is therefore proposed to be 3' coterminal with the adjacent, downstream gene encoding the 27.3K protein (figure 3.20). There is apparently no transcription termination sequence (McLauchlan et al., 1985) downstream of this polyadenylation signal.

The ATG at position 674 is considered to initiate translation of a 224 codon ORF, terminating at TAA, position 1348. The protein has a M_r 24,932. The G+C content of the coding region is low for HSV-1, at 58.1%. As a result, the biased codon usage and amino acid content, described for previous genes, is less apparent (table 3.48). The protein has a slight excess of basic over acidic residues (Table 3.49). The carboxy terminus of the protein is particularly basic

Table 3.49 Predicted Amino Acid Composition of 24.9K

Res	No.	%	Res	No.	%
Ala	18	8.0	Leu	22	9.8
Arg	23	10.3	Lys	7	3.1
Asn	5	2.2	Met	1	0.4
Asp	15	6.7	Phe	6	2.7
Cys	5	2.2	Pro	22	9.8
Gln	6	2.7	Ser	12	5.4
Glu	8	3.6	Thr	16	7.1
Gly	16	7.1	Trp	4	1.8
His	4	1.8	Tyr	8	3.6
Ile	10	4.5	Val	16	7.1

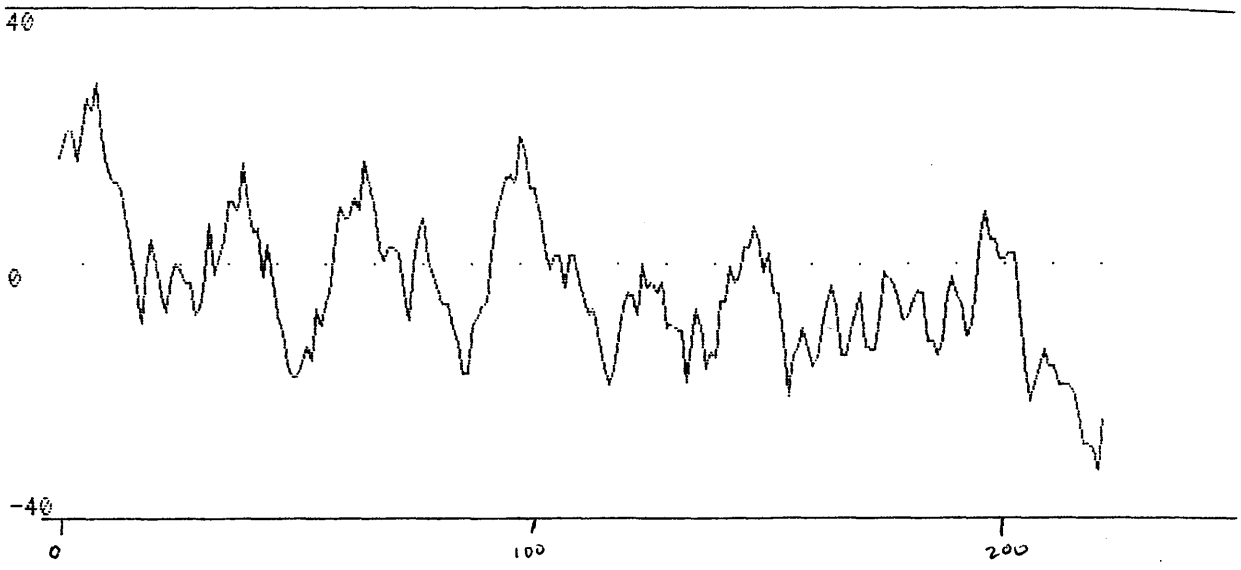
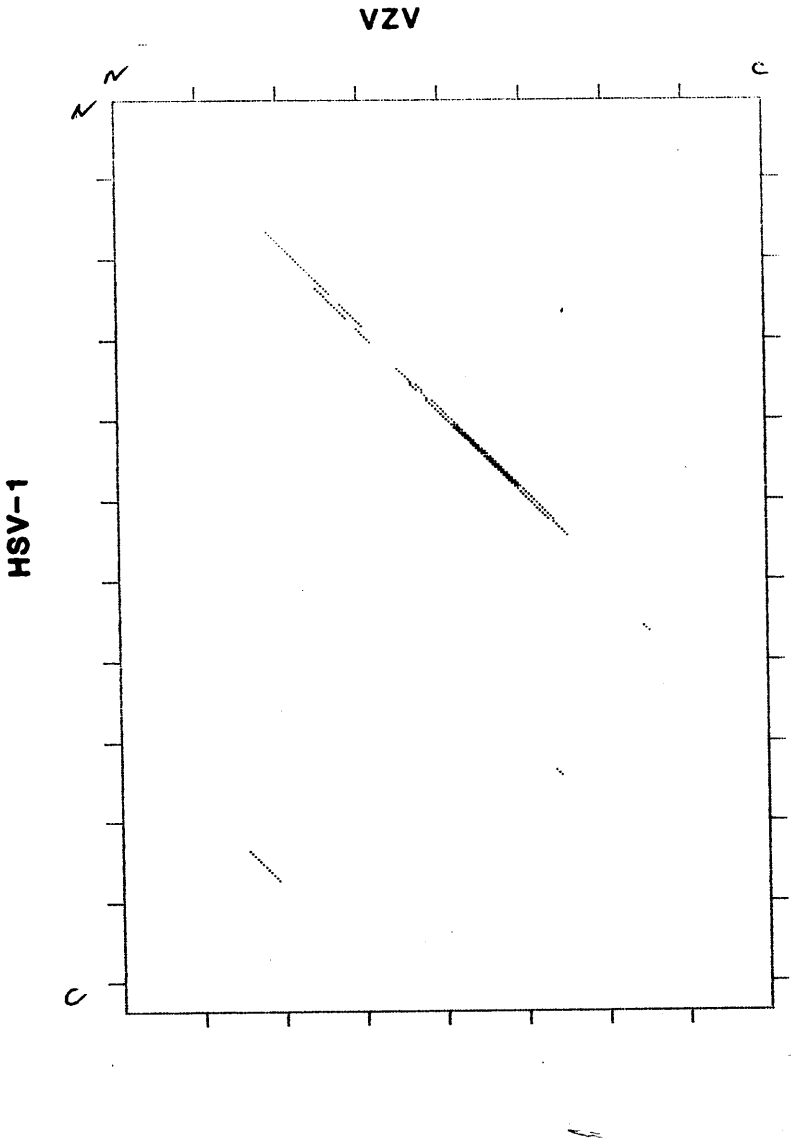

40
0
-40
0
100
200

Figure 3.50 Hydropathicity of the 24.9K protein

Hydropathicity of the amino acid sequence of the predicted 24.9K protein is illustrated graphically. The predicted nature of the amino acid sequence is based on the paramenters of Kyte and Doolittle (1982). The amino terminus of the protein is at the left, carboxy terminus at the right side of the figure. Higher scores represent hydrophobic regions.

The y axis represents hydropathicity, -40 to +40, summed over nine residues. The scale along the x axis is in 100 amino acids.




```
 1      MGILGWVGLIAVGVLCVRGGLPSTEYVIRSRVAREVGDILKVPCVPLPSD
           *  * *              *             *  *   ***
 1 MASHKWLLQIVFLKTITIAYCLHLQDDTPLFFGAKPLSDVSLIITEPCVSSVYE

51 DLDWRYETPSAINYALIDGIFLRYHCPGLDTVLWDRHAQKAYWVNPFLFVAGFL
    *       *    **  **      **      **    * *** **      * *
55 AWDYAAPPVSNLSEAL SGIVVKTKCPVPEVILWFKDKQMAYWTNPYVTLKG L
```

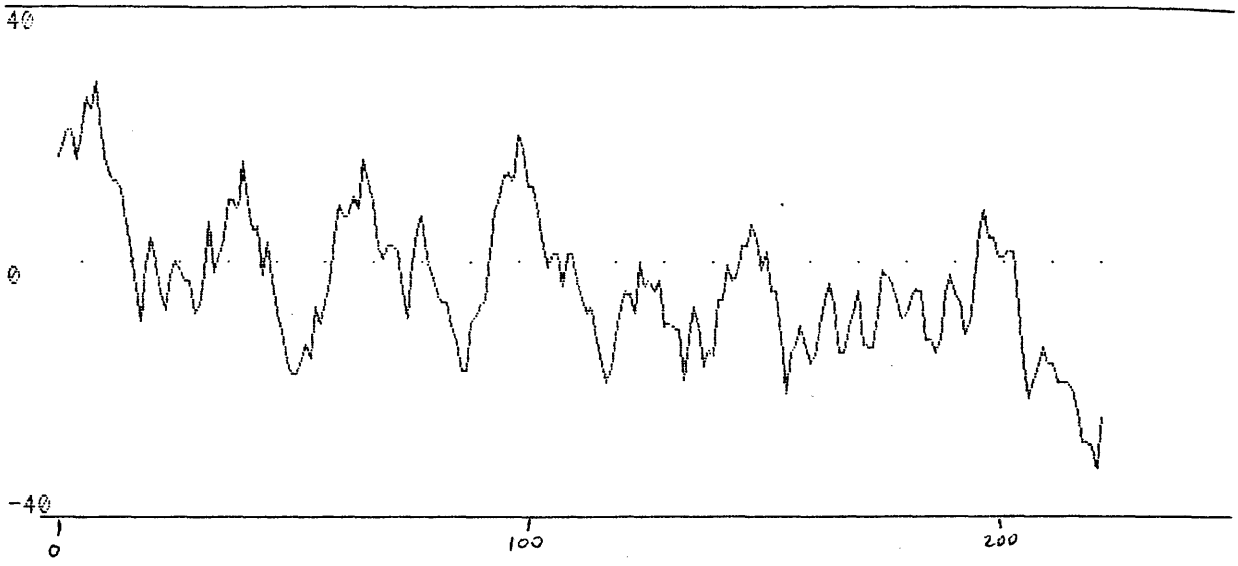
Figure 3.51 Relationship between the HSV-1 24.9K protein encoded by UL1 and the VZV 18K protein encoded by gene 60

The relationship between the 24.9K protein encoded by gene UL1 of HSV-1 and the protein encoded by VZV gene 60, is shown. Homology was calculated by the CINTHOM program (Pustell and Kafatos, 1982), adapted to allow for conservative amino acid changes, and displayed as a dot matrix using DIAG (P. Taylor, unpublished). The parameters used were as follows, range, 12; minimum value, 20; others, as default. The x axis represents the VZV 18K protein, the y axis the HSV-1 18K protein. The diagonal line represents amino acid sequences conserved between the two proteins.

Figure 3.52 Homology between the amino acid sequences in the conserved regions of the HSV-1 24.9K protein encoded by gene UL1 and the VZV 18K protein encoded by gene 60

The amino acid sequences of the two proteins, including the most conserved regions, near to the middle of the proteins, have been optimally aligned using the program HOMOL (Taylor, 1984). Default parameters were used. Identical residues are marked with an *. Blanks represent gaps introduced into the sequence to optimise alignment. The amino acid sequence of the HSV-1 24.9K protein is listed above the sequence of VZV 18K.

HSV-1



VZV

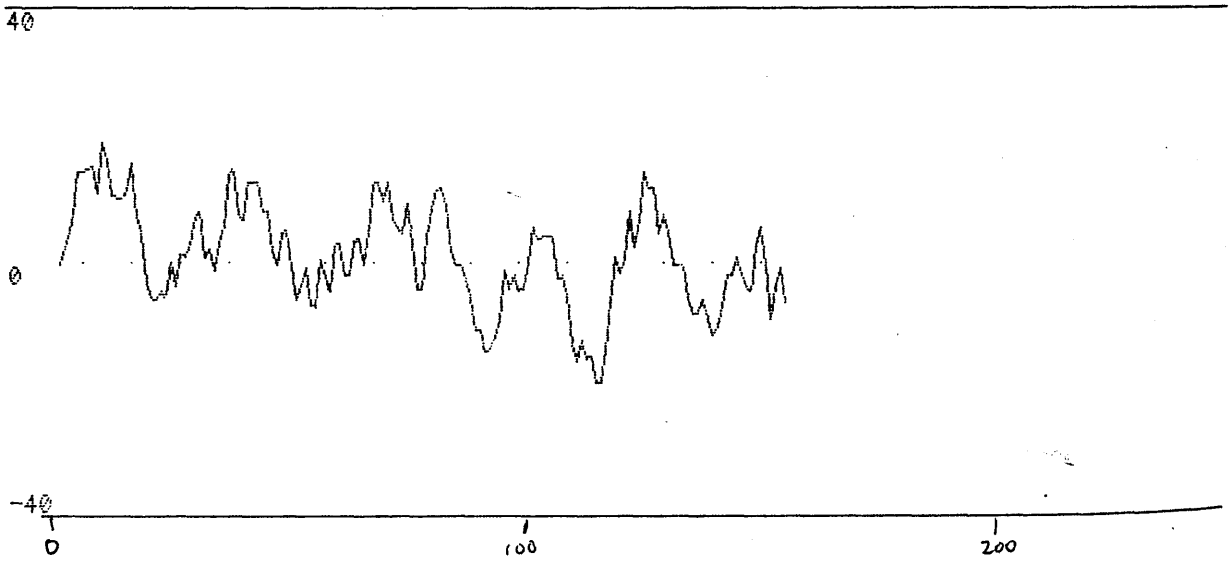


Figure 3.53 Hydropathicity profiles of the 24.9K HSV-1 and 18K VZV proteins

The hydropathicity profiles, using the same parameters as before, of the two homologous proteins is shown. The amino termini of the proteins are both strongly hydrophobic although their amino acid sequences have not been conserved.

The y axis represents hydropathicity, -40 to +40, summed over nine residues. The scale along the x axis is in 100 amino acids.

Table 3.54 Codon usage catalogue of 27.3K

TTT Phe	6	TCT Ser	1	TAT Tyr	4	TGT Cys	1
TTC Phe	3	TCC Ser	2	TAC Tyr	1	TGC Cys	5
TTA Leu	3	TCA Ser	1	TAA ---	0	TGA ---	1
TTG Leu	2	TCG Ser	7	TAG ---	0	TGG Trp	7
CTT Leu	2	CCT Pro	5	CAT His	2	CGT Arg	4
CTC Leu	6	CCC Pro	9	CAC His	8	CGC Arg	6
CTA Leu	1	CCA Pro	1	CAA Gln	0	CGA Arg	1
CTG Leu	10	CCG Pro	7	CAG Gln	5	CGG Arg	10
ATT Ile	1	ACT Thr	1	AAT Asn	3	AGT Ser	1
ATC Ile	7	ACC Thr	9	AAC Asn	6	AGC Ser	2
ATA Ile	0	ACA Thr	2	AAA Lys	0	AGA Arg	1
ATG Met	4	ACG Thr	2	AAG Lys	5	AGG Arg	1
GTT Val	4	GCT Ala	0	GAT Asp	2	GGT Gly	2
GTC Val	6	GCC Ala	6	GAC Asp	9	GGC Gly	7
GTA Val	0	GCA Ala	1	GAA Glu	3	GGA Gly	3
GTG Val	11	GCG Ala	15	GAG Glu	7	GGG Gly	3

Table 3.55 Predicted Amino Acid Composition of 27.3K

Res	No.	%	Res	No.	%
Ala	22	9.0	Leu	24	9.8
Arg	23	9.4	Lys	5	2.0
Asn	9	3.7	Met	4	1.6
Asp	11	4.5	Phe	9	3.7
Cys	6	2.5	Pro	22	9.0
Gln	5	2.0	Ser	14	5.7
Glu	10	4.1	Thr	14	5.7
Gly	15	6.1	Trp	7	2.9
His	10	4.1	Tyr	5	2.0
Ile	8	3.3	Val	21	8.6

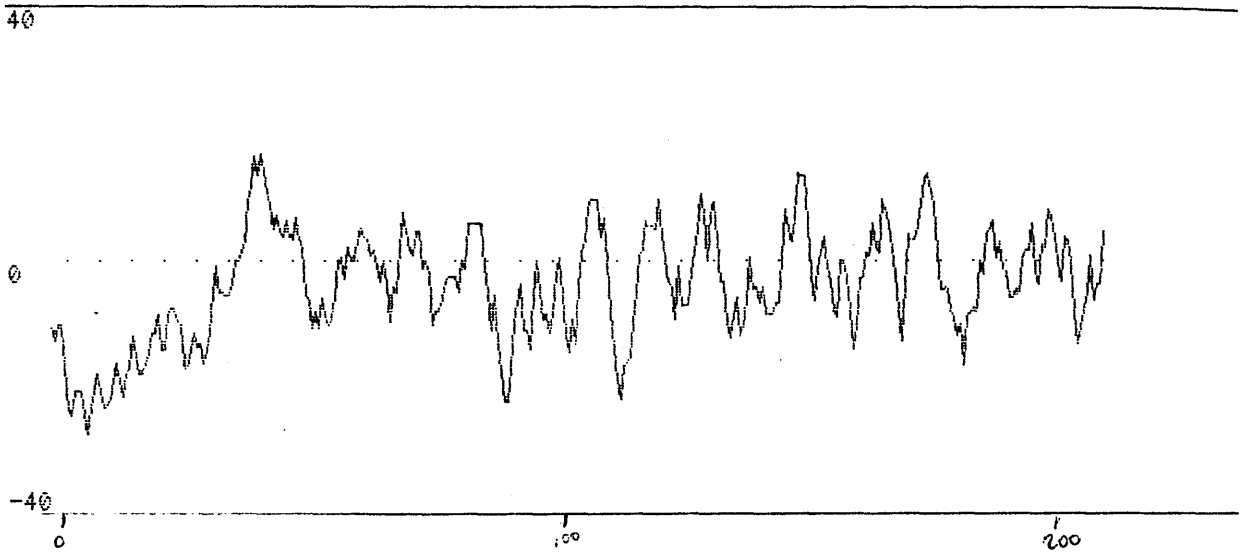

40
0
-40
0
100
200

Figure 3.56 Hydropathicity of the 27.3K protein

Hydropathicity of the amino acid sequence of the predicted 27.3K protein encoded by UL2 is illustrated graphically. The nature of the amino acid sequences is based on the parameters of Kyte and Doolittle (1982). The amino terminus of the protein is at the left, carboxy terminus at the right side of the figure. Higher scores represent hydrophobic regions.

The y axis represents hydropathicity, -40 to +40, summed over nine residues. The scale along the x axis is in 100 amino acids.

and hydrophilic. The amino terminus of the protein is strongly hydrophobic (figure 3.50), suggesting it may be membrane translated. As there are no other long regions of hydrophobic residues, it is probably not a membrane inserted protein.

A homologue to the 24.9K protein has been found in VZV (figure 3.51). The VZV protein has a M_r 17,616 and is coded by gene 60 (Davison and Scott, 1986). The VZV protein, here designated 17.6K, at only 159 amino acids in length, is considerably shorter than its HSV-1 homologue. However, the two proteins are clearly related. Greatest similarity in amino acid content can be seen near the middle of the two proteins. The carboxy terminal regions show least, if any, homology. The amino terminal 100 residues aligned in Figure 3.52, show the level of amino acid conservation near the centre of the proteins. Although there is a loss of homology near to the amino termini of the proteins, the hydrophobic nature of this region of the protein has been conserved in VZV, as shown in Figure 3.53. Indeed, the two proteins have similar hydropathicity profiles. The hydrophobic nature of the amino terminus of the protein may be required for its function. No EBV homologue to the 24.9K protein has been found.

3.10 Gene UL2 encoding 27.3K

There is a good TATA box sequence at position 1049. The mRNA for the 27.3K protein is coterminal with the mRNA for the 24.9K protein , as described above.

There are two potential translation initiation sites at positions 1491 and 1596. Although both

conform to Kozak's rules, the first ATG resembles the consensus initiation codon more closely. The codon evaluation of this region suggests the codon usage in the sequences preceding the second ATG are not typical of HSV-1 genes (figure 3.17). However it is not possible to rule out translation from the first initiation codon. For the purpose of this analysis, translation of the ORF will be taken from this first ATG. Translation terminates at TGA, position 2223. The 244 amino acid protein has a M_r 27,327. The effect of the 63.7% G+C content of the coding region is minimised by the codon usage, as described previously (table 3.54). There is a slight excess of basic over acidic residues (table 3.55). The amino terminus of the protein is hydrophilic, as shown in figure 3.56.

The protein has a close homologue in VZV. The M_r 34,375 protein is coded by gene 59 (Davison and Scott, 1986), and will be termed 34.4K. Conservation of amino acid sequences is shown throughout the protein, except at the amino terminus, as illustrated in figure 3.57. The aligned sequences of the HSV-1 and VZV proteins shows the extent of this homology (figure 3.58). In the region included in figure 3.58, 49% of residues are conserved between the two proteins. Both figures 3.57 and 3.58 suggest that translation of the HSV-1 protein may start at the second ATG. However, amino acid sequence conservation between the amino terminus of the VZV protein and the second ATG of the HSV-1 27.3K protein is minimal, and no definitive conclusions can be drawn. The possibility of two translation initiation sites functioning in this gene must be considered. Examples of multiple translation-initiation sites in HSV-1 have been reported (Marsden et al., 1983).

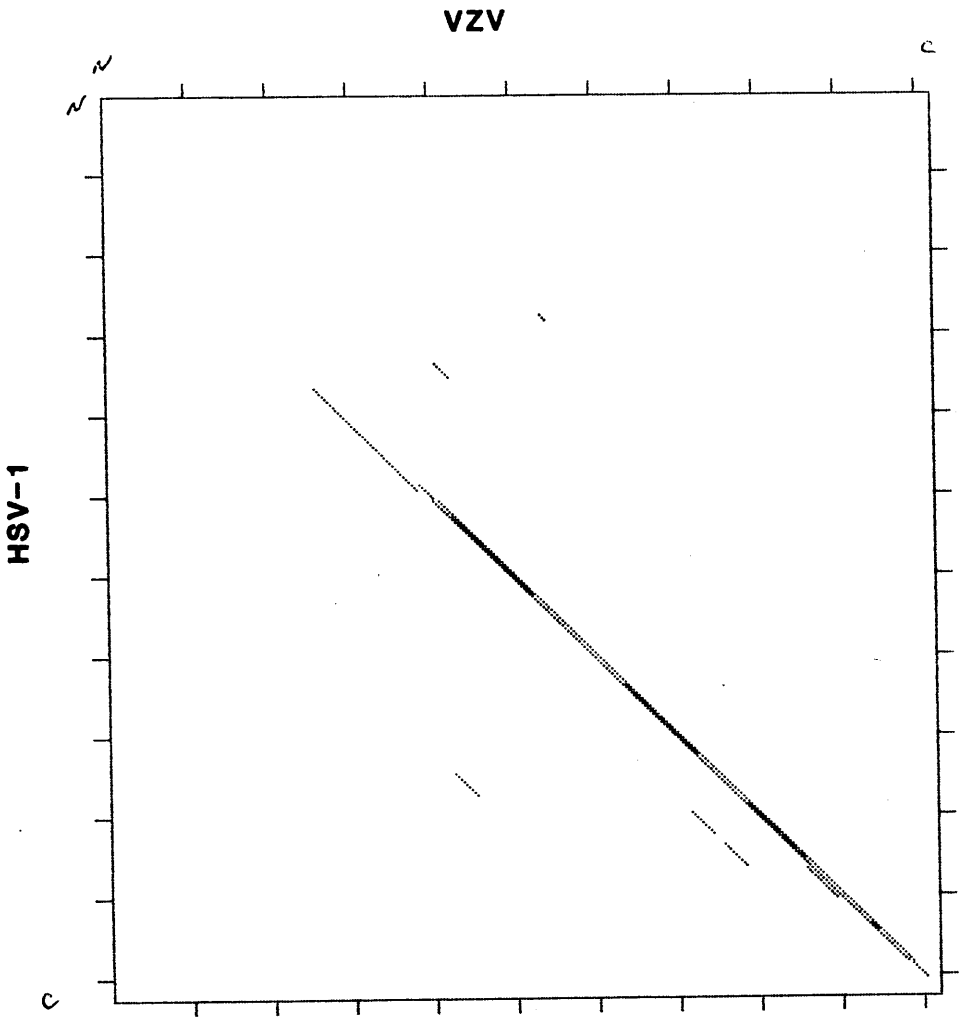

VZV
HSV-1
N
N
C
C

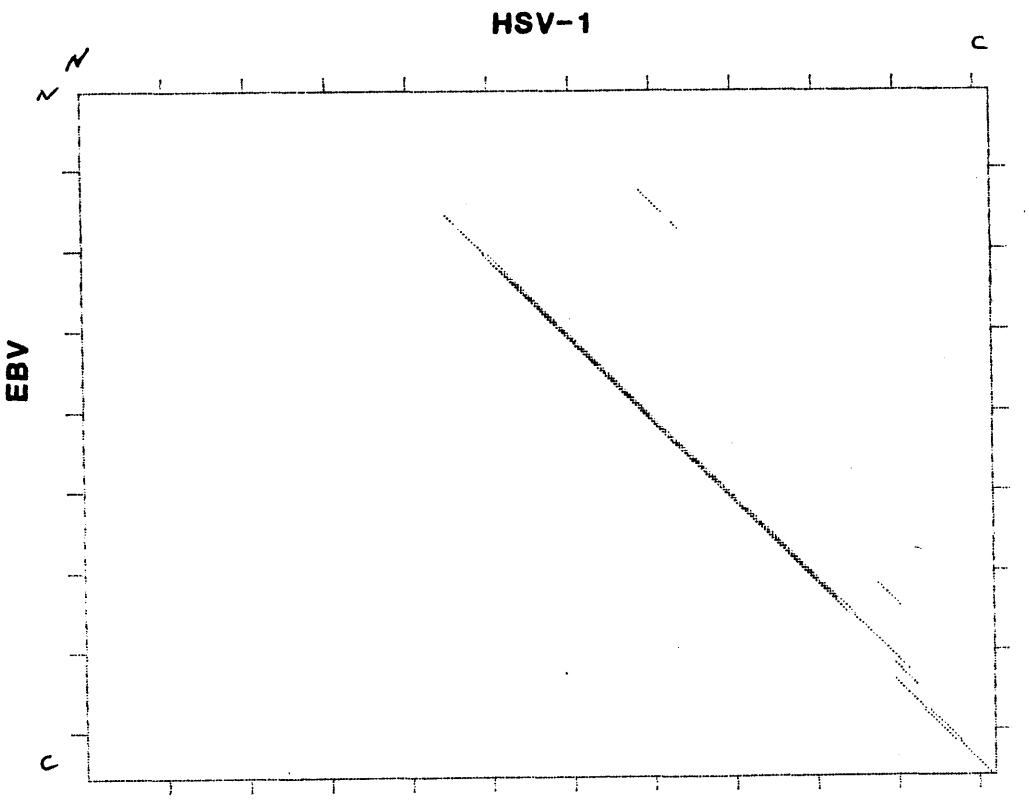
Figure 3.57 Relationship between the HSV-1 27.3K protein encoded by gene UL2 and VZV 34K protein encoded by gene 59

The relationship between the 27.3K protein of HSV-1 and the 34K protein encoded by VZV gene 59 is shown opposite. Homology was calculated by the CINTHOM program (Pustell and Kafatos, 1982), adapted to allow for conservative amino acid changes, and displayed as a dot matrix using DIAG (P. Taylor, unpublished). The parameters used were as follows, range, 12; minimum value, 20; others, as default. The x axis represents the 34K VZV protein, the y axis the HSV-1 27.3K protein. The diagonal line represents amino acid sequences conserved between the two proteins.

```
 91 MDLTNGGVSPAATSAPLDWTTFRRVFLIDDAWRPLMEPELANPLTAHLLA
    *              *  **        * *   *  *   **  *
 60 MSHHYDTETFTPVSSQLDSVEVFSKFNISPEWYDLLSDELKEPYAKGIFL

141 EYNR RCQTEEVLPPREDVFSWTRYCTPDEVRVVIIGQDPYHHPGQAHGL
    ****     ** **   * * *** * *   **********   * ****
110 EYNRLLNSGEEILPSTGDIFAWTRFCGPQSIRVVIIGQDPYPTAGHAHGL

190 AFSVRANVPPPPSLRNVLAAVKNCYPEARMSGHGCLEKWARDGVLLLNTT
    ****     ** ** *  **     **       ***** *** ********
160 AFSVKRGITPPSSLKNIFAALMESYPNMTPPTHGCLESWARQGVLLLNTT

240 LTVKRGAAASHSRIGWDRFVGGVIRRLAARRPGLVFMLWGTHAQNAIRPD
    *** **   **   ** * *  *  **   * ******** ***    *
210 LTVRRGTPGSHVYLGWGRLVQRVLQRLCENRTGLVFMLWGAHAQKTTQPN

290 PRVHCVLKFSHPSPLSKVPFGTCQHFLVANRYLETRSISPIDWSV
     * * **   ****** ***  * **  ** *        *****
260 SRCHLVLTHAHPSPLSRVPFRNCRHFVQANEYFTRKGEPEIDWSVI
```

Figure 3.58 Homology between the carboxy region of the HSV-1 27.3K encoded by UL2 and VZV 34K from gene 59

The amino acid sequences of the carboxy regions of the proteins were optimally aligned using the program HOMOL (Taylor, 1984). Default parameters were used. Identical residues are marked with an *. Blanks represent gaps introduced into the sequence to optimise alignment. The amino acid sequence of the HSV-1 27.3K protein is listed above the VZV 34K sequence.




```
134 LTAHLLAEYNRRCQTEEVLPPREDVFSWTRYCTPDEVRVVIIGQDPYHHPGQAH
      *   *   *       * *       * * * *    *** ***** * ***
 48  LAAVIACVRRLRTQATVYPEEDMCMAWARFCDPSDIKVVILGQDPY HGGQAN

188 GLAFSVRANVPPPPSLRNVLAAVKNCYPEARMSGHGCLEKWARDGVLLLNTTLT
    ******    * ******  *       **      ****  **  ******* **
100 GLAFSVAYGFPVPPSLRNIYAELHRSLPEFSPPDHGCLDAWASQGVLLLNTILT

242 VKRGAAASHSRIGWDRFVGGVIRRLAARRPGLVFMLWGTHA QNAIRPDPRVHC
    *  *   **  ***  *   **  *  *    ******  *   *        *
154 VQKGKPGSHADIGWAWFTDHVISLLSERLKACVFMLWGAKAGDKASLINSKKHL

295 VLKFSHPSPL          SKVPFGTCQHFLVANRYLETRSISPIDWSV
    **   *****             *    **  **  *       ***
208 VLTSQHPSPLAQNSTRKSAQQKFLGNNHFVLANNFLREKGLGEIDWRL
```

Figure 3.59 Relationship between the 27.3K protein encoded by HSV-1 gene UL2 and an EBV protein encoded by BKRF3

The relationships between the predicted 27.3K protein encoded by UL2 of HSV-1 and the EBV protein encoded by BKRF3, is shown. Homology was calculated by the CINTHOM program (Pustell and Kafatos, 1982), adapted to allow for conservative amino acid changes, and displayed as a dot matrix using DIAG (P. Taylor, Unpublished). The parameters used were as follows, range 20; minimum value, 20; others as default. The x axisrepresents the HSV-1 protein, the y axis the EBV protein. The diagonal line represents amino acid sequences conserved between the two proteins.

Figure 3.60 Homology between the aligned amino acid sequences of the HSV-1 27.3K protein and the protein encoded by BKRF3 of EBV

The amino acid sequences of the carboxy regions of the proteins have been optimally aligned using the program HOMOL (Taylor, 1984). Default parameters were used. Identical residues are marked with an *. Blanks represent gaps introduced into the sequence to optimise alignment.

Table 3.61 Codon usage catalogue of 24.4K

TTT	Phe	5	TCT Ser	3	TAT Tyr	1	TGT Cys	1
TTC	Phe	5	TCC Ser	5	TAC Tyr	2	TGC Cys	3
TTA	Leu	2	TCA Ser	1	TAA ---	1	TGA ---	0
TTG	Leu	2	TCG Ser	12	TAG ---	0	TGG Trp	2
CTT	Leu	2	CCT Pro	2	CAT His	0	CGT Arg	1
CTC	Leu	4	CCC Pro	6	CAC His	7	CGC Arg	11
CTA	Leu	1	CCA Pro	4	CAA Gln	1	CGA Arg	4
CTG	Leu	4	CCG Pro	6	CAG Gln	5	CGG Arg	6
ATT	Ile	4	ACT Thr	1	AAT Asn	0	AGT Ser	0
ATC	Ile	2	ACC Thr	5	AAC Asn	4	AGC Ser	3
ATA	Ile	1	ACA Thr	1	AAA Lys	2	AGA Arg	1
ATG	Met	5	ACG Thr	8	AAG Lys	5	AGG Arg	0
GTT	Val	2	GCT Ala	2	GAT Asp	5	GGT Gly	1
GTC	Val	6	GCC Ala	11	GAC Asp	8	GGC Gly	5
GTA	Val	0	GCA Ala	1	GAA Glu	2	GGA Gly	3
GTG	Val	6	GCG Ala	7	GAG Glu	7	GGG Gly	8

Table 3.62 Predicted Amino Acid Composition of 24.4K

Res	No.	%	Res	No.	%
Ala	21	9.4	Leu	15	6.7
Arg	23	10.3	Lys	7	3.1
Asn	4	1.8	Met	5	2.2
Asp	13	5.8	Phe	10	4.5
Cys	4	1.8	Pro	18	8.0
Gln	6	2.7	Ser	24	10.7
Glu	9	4.0	Thr	15	6.7
Gly	17	7.6	Trp	2	0.9
His	7	3.1	Tyr	3	1.3
Ile	7	3.1	Val	14	6.2

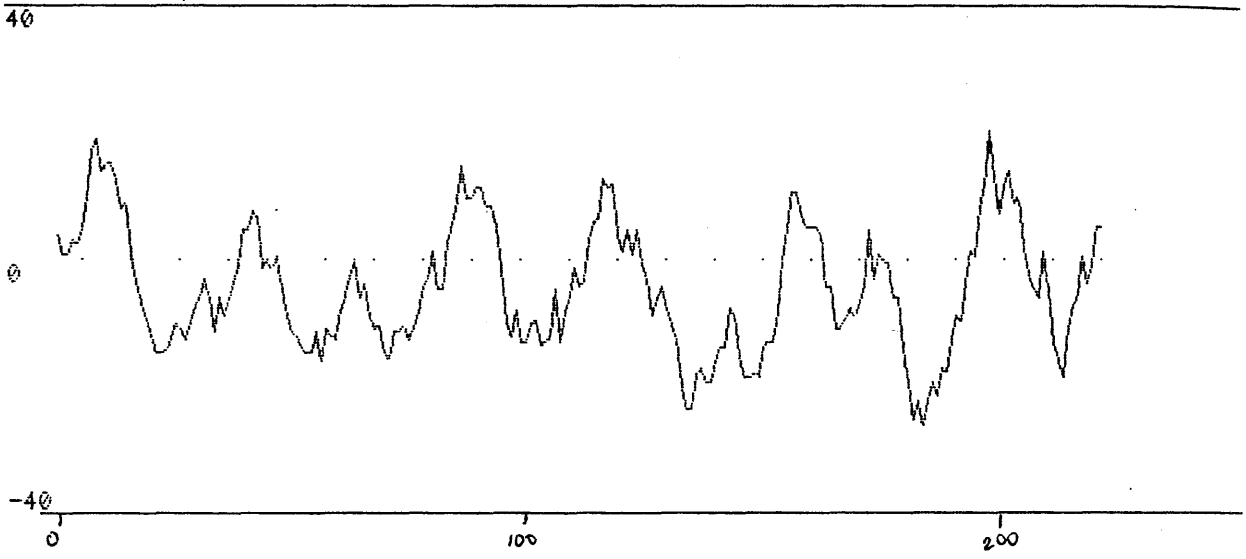

40
0
-40
0
100
200

Figure 3.63 Hydropathicity of the 24.4K protein

Hydropathicity of the amino acid sequence of the predicted 24.4K protein encoded by HSV-1 UL3 is illustrated graphically. The nature of the amino acid sequences is based on the parame ters of Kyte and Doolittle (1982). The amino terminus of the protein is at the left, carboxy terminus at the right side of the figure. Higher scores represent hydrophobic regions.

The y axis represents hydropathicity, -40 to +40, summed over nine residues. The scale along the x axis is in 100 amino acids.

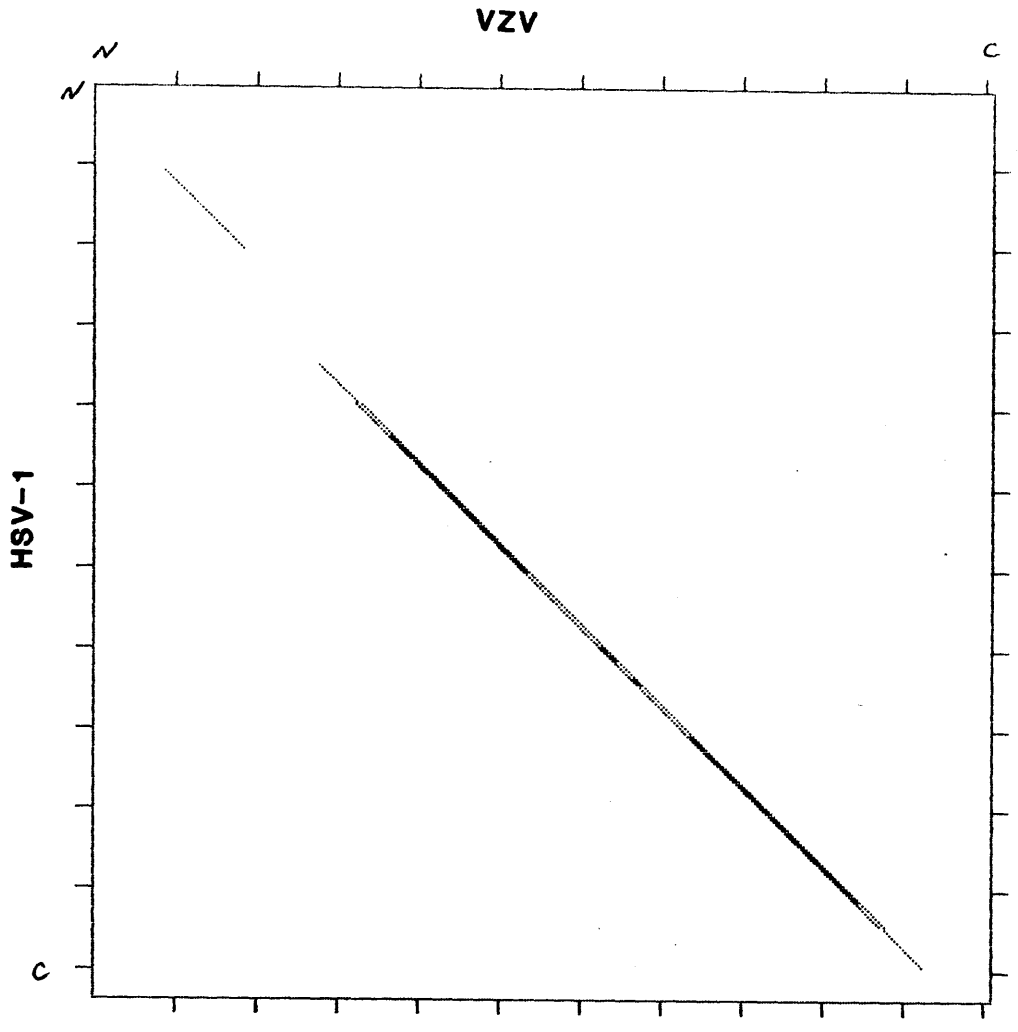
A homologue to the 27.3K protein has also been found in EBV. The protein is coded by gene BKRF3 and has a M_r 31,606 (Baer et al., 1984). The 31.6K EBV protein shows considerable amino acid sequence conservation, as illustrated in figure 3.59. The optimally aligned amino acid sequences of the two homologous proteins are shown in figure 3.60. In this region of the proteins, 42% of residues are conserved.

3.11 GENE UL3 ENCODING 24.4K

This gene is located on the same strand as the two genes described above. Although there is no good TATA box homologue upstream of the coding region, there are several A+T rich sequences, including the polyadenylation signal of the mRNAs for the 24.9K and 27.3K proteins. The polyadenylation signal of the mRNA for the gene encoding the 24.4K protein is located at position 3052.

Translation of the mRNA is considered to begin at the ATG at position 2327, and terminate at TAA, position 2999. The coding region is 64.0% G+C. The amino acid content of the protein and the codon usage of the gene are given in tables 3.61 and 3.62. The protein is particularly rich in Ser and Arg residues, which are distributed uniformly throughout its length. The protein is slightly basic. The predicted protein has a M_r 24,446. The protein contains both hydrophobic and hydrophilic regions (figure 3.63).

A close homologue to the 24.4K protein has been found in VZV (figure 3.64). The VZV M_r 25,093 protein coded by gene 58 (Davison and Scott, 1986) is of similar size and is here termed 25.1K. The VZV protein




```
 75 GPPCTDGPYVTFDTLFMVSSIDELGRRQLTDTIRKDLRLSLAKFSIACTK
       *    ** ************************  *          ****
 61 CIEKTKDDYVPFDTLFMVSSIDELGRRQLTDTIRRSLVMNACEITVACTK

125 TSSFSGNAPRHHRRGAFQRGTRAPRSNKSLQMFVLCKRAHAARVREQLRV
    *  ***                    * * *********    ** ***  *
111 TAAFSGRGVSRQKHVTLSKNKFNPSSHKSLQMFVLCQKTHAPRVRNLLYE

175 VIQSRKPRKYYTRSSDGRLCPAVPVFVHEFVSSEPMRLHRDN        V
     *  * ** ***** **   * ***** **        **  *
161 SIRARRPRRYYTRSTDGKSRPLVPVFVYEFTALDRVLLHKENTLTDQPIN

218 MLASG    AE
       **
211 TENSGHGRTRT
```

Figure 3.64 Relationship between the 24.4K protein encoded by HSV-1 gene UL3 and VZV 25K protein encoded by gene 59

The relationship between the predicted 24.4K protein encoded by gene UL3 of HSV-1 and the 25K protein encoded by VZV gene 58 is shown. Homology was calculated by the CINTHOM program (Pustell and Kafatos, 1982), adapted to allow for conservative amino acid changes, and displayed as a dot matrix using DIAG (P. Taylor, unpublished). The parameters used were as follows, range, 20; minimum value, 20; others, as default. The x axis represents the 25K VZV protein, the y axis the HSV-1 24.4K protein. The diagonal line represents amino acid sequences conserved between the two proteins.

Figure 3.65 Homology between the carboxy region of the 24.4K protein encoded by HSV-1 gene UL3 and the 25K VZV protein encoded by gene 58

The amino acid sequences of the carboxy regions of the HSV-1 24.4K and VZV 25K proteins were optimally aligned using the program HOMOL (Taylor, 1984). Default parameters were used. Identical residues are marked with an *. Blanks represent gaps introduced into the sequence to optimise alignment. The amino acid sequence of the HSV-1 24.4K protein is listed above the VZV 25K sequence.

shows homology to its HSV-1 counterpart along most of its sequence (figure 3.64). Amino acid sequence conservation is less pronounced at the amino terminus. In figure 3.65, 47% of the residues are shown to be conserved in the most homologous region of the proteins. No EBV homologue to the 24.4K protein has been detected.

3.12 SEQUENCE IN R_L OUTSIDE OF IE GENE 1

No gene has yet been assigned to the region in R_L downstream of IE gene 1. The sequence was systematically analysed for the presence of genes. Much of the sequence is composed of ORFs, as shown in figure 3.15.

Potential transcriptional control sequences have been identified, and are indicated on figure 3.66, together with the location and orientations of ORFs in the region. There are four potential polyadenylation signals (Wickens and Stephenson, 1984). Three lie on the rightward 5' to 3' strand, at positions 4581, 4597 and 5287. The fourth lies on the opposite strand at position 4578. The ORFs upstream of each polyadenylation signal has been examined. Upstream of the first polyadenylation signal is an ORF 387 bp in length, designated ORF 1. This would encode a polypeptide of M_r 13.8K. The closest TATA box homologue lies in U_L, at position 3789. This sequence is believed to function as the TATA box for the 21.2K protein in U_L, described above. There is no homologue to the sequence associated with efficient transcriptional termination downstream of this polyadenylation site.

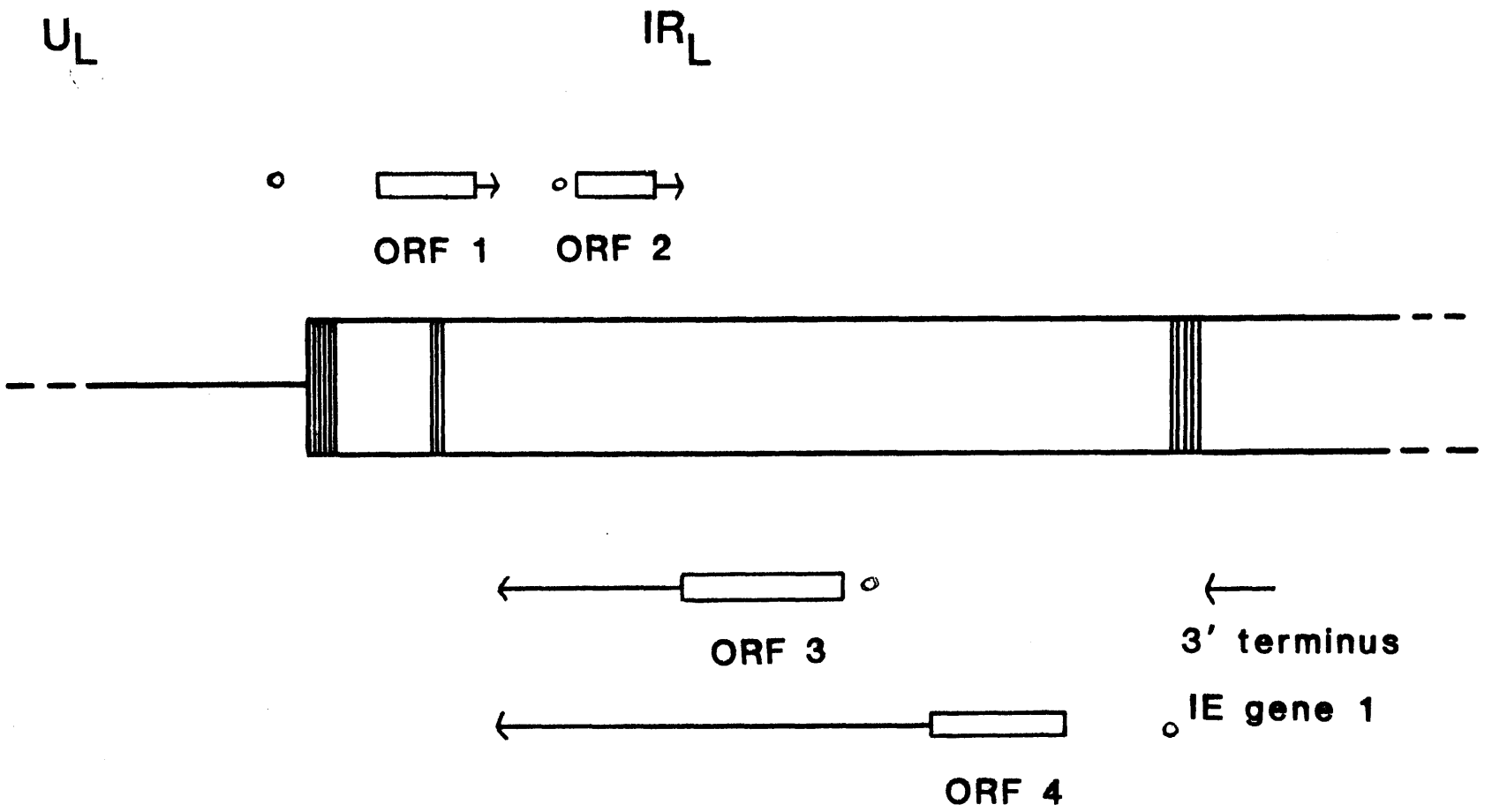

U_L
IR_L
ORF 1
ORF 2
ORF 3
3' terminus
IE gene 1
ORF 4

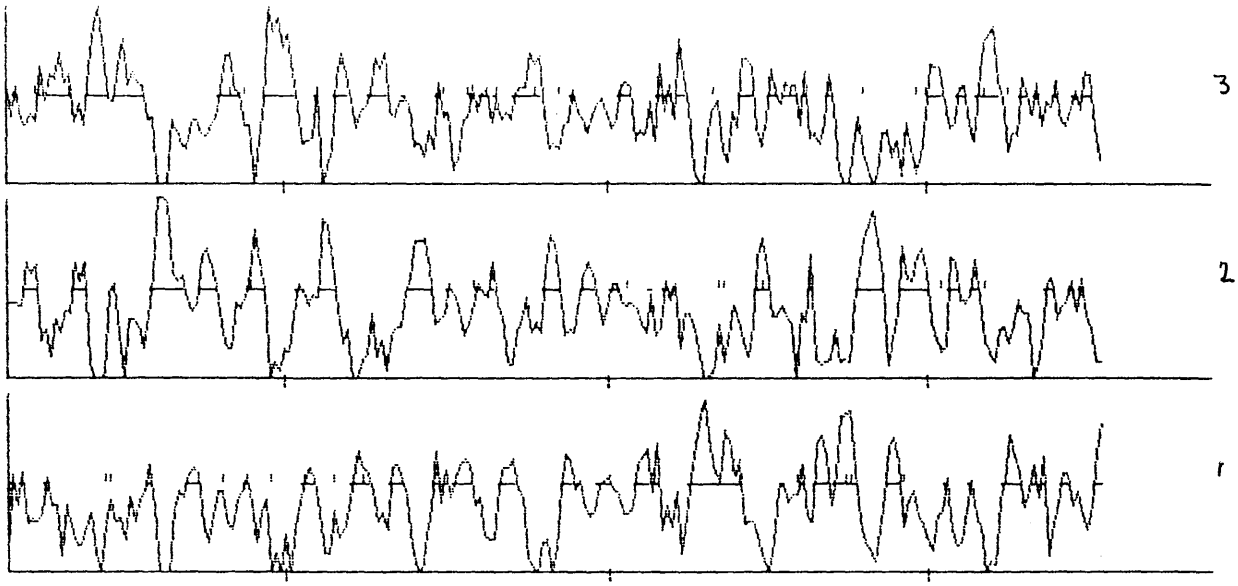
Figure 3.66 Major open reading frames in IR_L, downstream of IE gene 1

Possible polypeptide coding regions of the ORFs in IR_L downstream of IE gene 1 are shown as open boxes. The orientation of the ORFs is indicated. Solid lines represent the distance to the closest polyadenylation signal. The locations of tandem reiterations and the position of the mapped 3' terminus of IE gene 1 are shown. Potential TATA box sequences are indicated as circles.

ORF2, initiated at position 4906, is 288 bp in length and would encode a polypeptide of M_r 10.5K. ORF2 has sequences resembling a TATA box, a polyadenylation signal and a transcription termination sequence. On the opposite strand, ORF3, initiated at position 5908 is 603 bp in length, and would encode a polypeptide of M_r 22.1K. There are several A+T rich sequences upstream which may function as TATA boxes. The closest polyadenylation signal is at position 4573. Downstream of the polyadenylation signal is a homologue to the transcription termination consensus sequence (McLauchlan et al., 1985). ORF4, upstream of ORF3, initiated at position 6721 is 492 bp in length, and would encode a polypeptide of M_r 18.0K. The closest TATA box homologue is over 400 bp upstream, and the closest polyadenylation signal over 1.5 kb downstream of the termination codon.

The codon usages of the four ORFs were evaluated using IE gene 1 and IE gene 3, both coded in the major repeat elements by G+C rich DNA. All four of the ORFs showed atypical codon usage. To determine whether the genes as a set employed an alternative codon usage pattern, the ORFs were compared against each other. Again, the results were negative. Examples of the codon usage evaluations for the whole of this region are illustrated in figure 3.67. In this figure both strands of the DNA have been analysed for potential coding regions. Similarly negative results were achieved using genes from U_L as reference.

The repetitive nature of the R_L sequence has been described. The sequences of each of the four ORFs was examined to locate any reiterations. ORF1 contains a number of repeated sequences, including six copies of the sequence [CCCGG]. In addition, it contains long


3
2
1

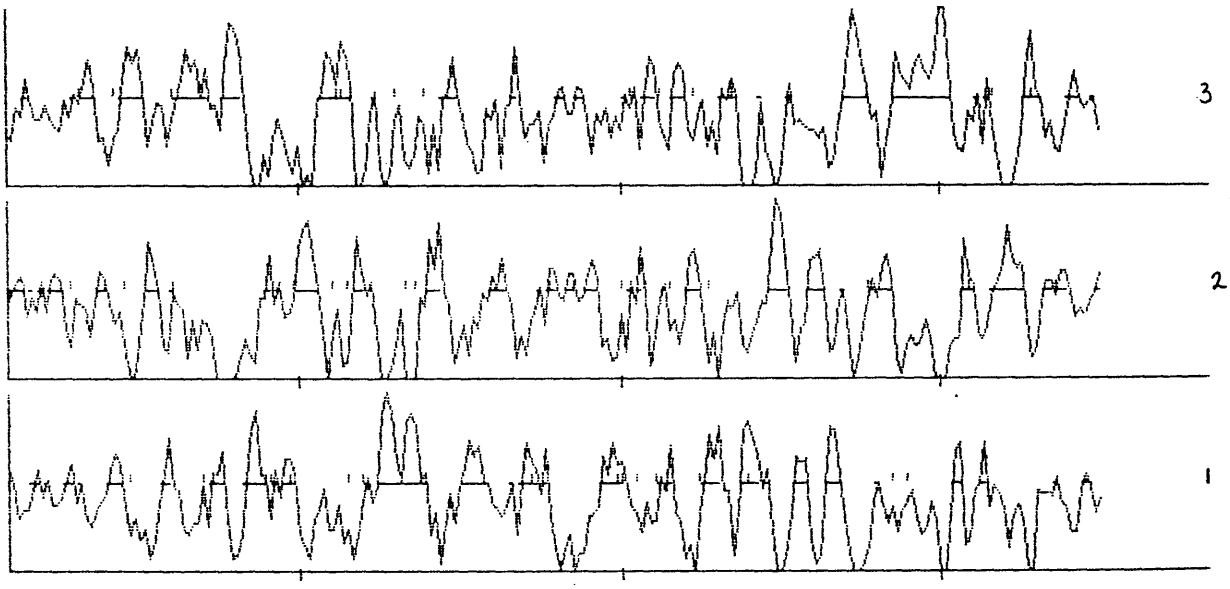

3
2
1

Figure 3.67 Codon usage evaluations of the DNA sequence in R_L, downstream of IE gene 1

The sequence in IR_L downstream of the position of the mapped terminus of the mRNA of IE gene 1 was examined for potential polypeptide coding regions. An analysis of the codon usage of this sequence is shown in this figure. The two sets of results are for the two DNA strands. In this instance the proposed coding regions of IE gene 1 were used as the reference. As before, high scores represent possible coding regions, short vertical lines on the central axis represent in-frame stop codons. A window of 15 residues, in steps of 5 were used in this analysis. The reading frames are numbered at the right of the figure.

strings of single bases, namely $[G]_{11}$ and $[C]_{11}$ at positions 4160 and 4312. ORF4 also contains a number of long strings of C residues. These characteristics may suggest that the ORFs are not polypeptide coding, unless the mRNA is spliced to exclude the repeated sequences. Two of the differences observed in the overlapping sequences in R_L of BamHI b and the SmaI/BamHI subfragment of BamHI e lie within ORF1. This may also suggest that the sequence does not encode a polypeptide. There was no overlap of sequence data in the other three reading frames.

Five of the six genes identified in the U_L sequences, and IE gene 1 in R_L, have homologues in VZV. The sequence of R_L was analysed to see whether an amino acid sequence encoded within this region showed homology to any VZV protein. All three reading frames from both strands were translated. An extensive search was made using the CINTHOM matrix program developed by P.Taylor from Pustell and Kafatos (1982). No homology was observed between the translated amino acid sequences and any VZV protein.

To conclude, the region is almost entirely composed of ORFs, and there are several sequences which closely resemble transcriptional control signals. However, the potential codon usages of the sequences are atypical of HSV-1 genes, and the region contains a high level of repeated sequences. It is not possible to arrive at a clear conclusion concerning possible genes encoded in this region. Northern blotting data, supplimented with fine mapping of transcripts from the region, are now required to identify any genes and to interpret the DNA sequence data in this region.

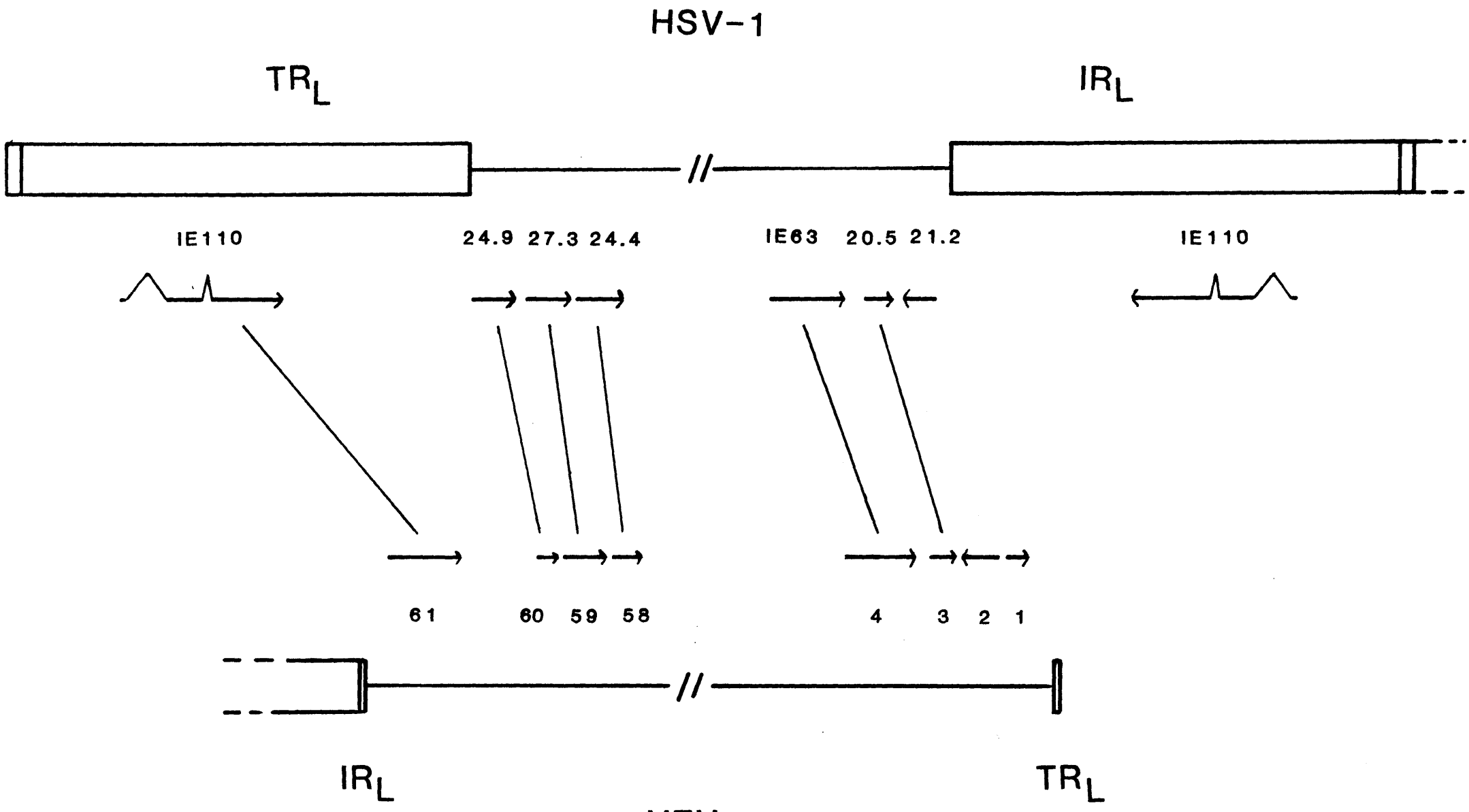

HSV-1
TRL
IRL
IE110
24.9 27.3 24.4
IE63 20.5 21.2
IE110
61
60 59 58
4 3 2 1
IRL
TRL
VZV

Figure 3.68 Relationship between the genes located near the termini of the U_L sequence of HSV-1 and VZV

The HSV-1 genome is shown in the prototype arrangement at the top of the figure. The VZV genome is drawn in the I_L arrangement. Open boxes represent the inverted repeat elements (R_L). Potential polypeptide coding regions showing homology between the two viruses are indicated. The relation between the gene corresponding to IE110 in HSV-1 is drawn only once, for clarity. The predicted genes in HSV-1 are labelled with their M_r. The VZV ORFs are numbered according to Davison and Scott (1986).

The alternative naming system for three of the genes used in this thesis is as follows. UL1, 24.9K; UL2, 27.3K; and UL3, 24.4K.

3.13 RELATIVE ARRANGEMENT OF HOMOLOGOUS GENES IN HSV-1 AND VZV

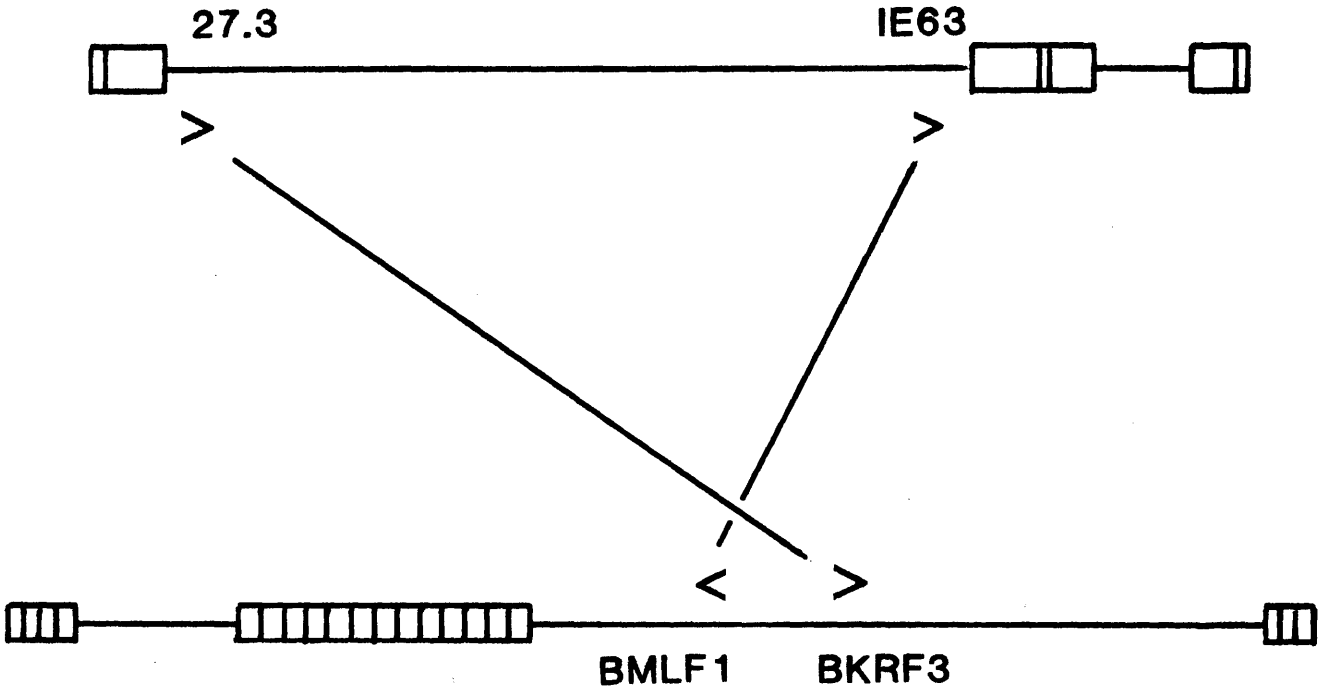
IE gene 1 and five of the genes located in the U_L sequences described, have homologues in VZV. Only one gene in U_L, encoding the 21.2K protein, does not show homology to any VZV ORF.

Figure 3.68 shows the arrangement of genes in the HSV-1 genome, described in this section. The VZV genome is shown underneath. The pairs of corresponding genes are indicated. For the purpose of this comparison the orientation of the VZV genome has been inverted, relative to the prototype, so that the short region is at the left. The figure illustrates the similar arrangement of homologous genes near the termini of U_L in HSV-1 and VZV. The L segment of VZV does not have large inverted repeat sequences. The VZV gene encoding the homologue to IE110 is located at the terminus of U_L near to the junction with IR_L. Although VZV has two ORFs at the other end of U_L downstream of the homologue to the 20.5K protein, neither shows homology to the HSV-1 21.2K protein. No equivalent to the HSV-1 21.2K gene was found elsewhere in the VZV genome, suggesting it had not been relocated through non-homologous recombination. As mentioned previously, no homology was found between the amino acid sequences translated from all six reading frames from R_L with any VZV ORF.

3.14 RELATIVE ARRANGEMENT OF HOMOLOGOUS GENES IN HSV-1 AND EBV

Only two of the proteins described above appear to have homologues coded in the EBV genome. These are

HSV-1



EBV

Figure 3.69 Relationship between HSV-1 and EBV in the region investigated

The HSV-1 genome is shown at the top of the figure, in the prototype arrangement. The EBV genome is shown below. Open boxes represent repeat elements. Solid lines, the unique sequences. Genes are drawn as arrowheads, indicating the direction of transcription. The location of the two pairs of genes encoding homologous proteins is shown. The two HSV-1 genes are IE gene 2, encoding IE63, and UL2, encoding a 27.3K protein.

IE63 and the 27.3K protein encoded by UL2. The locations of the corresponding genes are illustrated in figure 3.69. The figure shows extensive rearrangement of the two genes relative to HSV-1. Close examination of the ORFs neighbouring the EBV homologues did not indicate that they might encode proteins related to any of the HSV-1 proteins described here.

DISCUSSION

DISCUSSION

4.1 EVALUATION OF THE DNA SEQUENCE ANALYSIS

4.1.1 Problems associated with base composition

HSV DNA has a high G+C content which presents difficulties in sequencing, resulting in gel artefacts such as compressions and pile-ups. A compression is a term used to describe the situation when the newly synthesised strand of DNA forms secondary structures, often as a result of binding of short palindromic G+C rich sequences. Pile-ups are caused by the template DNA forming secondary structures which the polymerase does not read through efficiently. To minimise compressions, a denaturing electrophoresis system is used, which includes 9 M urea in the gel. Pile-ups were reduced by conducting polymerase reactions at 37^{o}C. Despite this, regions with extreme G+C contents still prove difficult to sequence. Most of the sequence of both DNA fragments was obtained from clones generating data for both strands. This helps to resolve artefacts and can show up previously undetected compressions, as these will often occur at different positions on the two strands.

Resolving compressions was most successfully achieved by running hot gels. Raising the temperature of the gel using a jacket of hot circulating water reduces the formation of secondary structures in the DNA. Where the running of hot gels proved insufficient, the troublesome sequence was recloned, so that the compression was positioned near the start of the clone. This was then sequenced and run on a hot gel.

4.1.2 Mutations within genes

A second source of ambiguity which arose during the sequencing of the DNA was the insertion or deletion of a single residue in a homopolymer tract. Both detected instances of this became apparent when the DNA sequences of two independently cloned fragments were compared. The first was observed in the overlapping sequences of the pGX48 BamHI b and the pGX58 XhoI c fragments, and lies within the first intron of IE gene 1. Detailed examination of the sequence autoradiographs showed that this difference was real and not a gel artefact. The origin of this insertion/deletion is unknown. As the residue lies within an intron, it may be of no consequence to the virus, and could have been present in the viral DNA when it was cloned.

In the second case, the mutation occurred within the coding region of IE gene 2. Plasmid pGX190, which overlaps this region, was sequenced. pGX190 contains an additional nucleotide and opens a reading frame of appropriate size for IE63, which has a codon usage similar to other HSV-1 genes. In functional assays of IE63, using both pGX48 and pGX190 as the source of IE gene 2, only pGX190 gave IE63 activity (C.M. Preston, personal communication), which further supports the interpretation. The deletion in pGX48 may have originated from a defective viral genome, but it could not have survived as an independently infectious species if it resulted in the loss of an essential function.

Length variability of homopolymer tracts has been previously reported. In the U_S sequence of HSV-1,

McGeoch et al. (1985) observed variability within a C tract. Gingeras et al. (1982) observed a similar length variability within an A tract in the genome of adenovirus type 2. This heterogeneity may result from slippage during DNA replication or from unequal recombination. From the sequence determined to date, the HSV-1 genome has an abundance of long C:G tracts. It is uncertain how many of these vary in length. As the gene for IE63 had been mapped (Whitton et al., 1983), a mutation causing a frameshift, thereby blocking the ORF, could be detected in the sequence. However, mutations lying within non-coding regions cannot be similarly detected. Of more serious consequence, a mutation lying within the coding region of a previously uncharacterised gene may remain unnoticed.

4.1.3 Heterogeneity in R_L sequences

The R_L sequences determined in each fragment were compared, and several differences were observed. One of these differences was an inversion of two nucleotides, and may be an error in the BamHI e sequence resulting from aberrant gel migration of the sequenced DNA. The second instance involved the insertion/deletion of two nucleotides in a homopolymer tract. This resembles the situation described in section 4.1.2, above. A further insertion/deletion event of a single nucleotide in the R_L sequence was observed. The source of these differences was not determined. It is possible that they originated in the viral genome. If this region of the genome does not encode polypeptide, it may tolerate these mutations. It is possible that an HSV-1 population may show a level of heterogeneity in regions of the genome

not under strong selective pressure.

4.1.4 Variation in copy number of a reiterated sequence

The R_L sequences adjacent to the R_L/U_L junctions were found to contain a tandemly reiterated 17 bp sequence. The copy number of this sequence in M13 clones of BamHI b appeared to vary between 2.5 and at least 18. Only four copies of the sequence were entered into the database. M13 clones with greater numbers of this reiteration were sequenced, but the gel migration patterns of the samples were not sufficiently clear to register possible imperfect copies of the sequence. In BamHI e, only one M13 clone spanned the reiteration set. This clone contained ten perfect copies of the sequence.

To discover whether the variation observed in BamHI b arose during the generation of the M13 clones, the copy number of the reiteration in the plasmid pGX48 was examined. Copy numbers ranging from 1 to at least 21 could be detected in the plasmid DNA. Copy numbers of 3 and 20 were most abundant. It thus appears that this sequence is unstable in the plasmid. Copy numbers of the reiteration in the BamHI e plasmid DNA were not determined.

Deletions of specific sequences within plasmids have been reported. Usually the sequences deleted contain inverted repeats. For instance, the 144 bp palindromic sequence of HSV-1 Ori_L, when contained in a plasmid, has been found to delete at a very high frequency (Quinn and McGeoch, 1985). However, the sequence described in this present case is a tandem

reiteration, not a palindromic repeat sequence.

The DNA sequence at or near the reiterations may represent a recombinational hot spot. There is a Chi sequence (GCTGGTGG) in R_L, at position 4070 in BamHI b, and at position 208 in the BamHI e sequences presented. Chi sites enhance homologous recombination by the RecBC pathway of E.coli (Ponticelli et al., 1985). The presence of the Chi recognition site may well increase recombination in the plasmid near to the sequence. The reiteration showing the variable copy number is less than 200 bp from the Chi sequence. Unequal recombination between short homologous sequences within the reiteration, possibly influenced by the Chi sequence, may have been responsible for this heterogeneity. There are no other Chi sites in the BamHI b and BamHI e sequences.

Variation in copy number of reiteration sets in the HSV genome has been previously described. Size heterogeneity of the a sequence has been shown to be due to variable copy numbers of several short reiterated sequences, and is presumed to result from unequal recombination (Davison and Wilkie, 1981).

Post et al. (1980) reported that in digests of HSV-1 DNA, the BamHI b fragment does not form a discrete band after gel electrophoresis. The wide, diffuse band was believed to be due to a variation in size of BamHI b fragments. This probably results from a variation in copy number of the reiterated sequences in the viral genome. However, the authors did not observe a similar variation in size of the BamHI e fragment and concluded that the variable region must lie within the U_L sequences contained in BamHI b.

4.1.5 Limitations of large scale sequence analyses

The difficulties encountered during this work indicate the limitations of the system. The high G+C content of the DNA often resulted in ambiguous gel migration patterns, although these could usually be resolved using alternative sequencing systems.

The potentially serious problem of frame-shifting mutations within genes was also encountered. Both cases were seen when overlapping sequences of independently cloned genomic DNA, were compared. In the case of IE gene 2, the deletion blocked the reading frame of IE63. As the gene had been mapped, the problem was recognised, and could be resolved. Sequencing uncharacterised regions of the genome must be seen as carrying a risk of misinterpretation. To overcome the problems resulting from mutations within the DNA sequences, it may, in principle, be necessary to sequence two or more fragments generated from separate cloning experiments, although this would significantly increase the work involved. It may be preferable to rely on computer analyses to detect mutations within the sequence of a single clone: for example, the codon usage evaluation program (Staden and McLachlan, 1982) which predicts polypeptide coding reading frames. Abrupt changes in reading frame, where there is no evidence of splicing, could be due to the insertion or deletion of a residue.

These findings emphasise the difficulties in performing large scale DNA sequence analysis independently of other supportive evidence, such as mRNA mapping or previous genetic analyses. However, evaluation of the codon usage of ORFs, taken together

with other recognised features of HSV-1 gene arrangements, can give a good indication of the organisation of genes. With the complete DNA sequence determination and transcript mapping of the entire S segment of HSV-1 (Rixon et al., 1982; McGeoch et al., 1985 and 1986a; Rixon and McGeoch, 1985), and the subsequent increased understanding of the coding strategy of the virus, greater confidence can be placed in these predictions. Transcript mapping subsequent to the sequence determination can be used to confirm the organisation of genes deduced from the analysis of the DNA sequence.

4.2 INTERPRETATION OF THE SEQUENCE

The most striking characteristic of the DNA sequence investigated was the base composition. The G+C content of the DNA however, is not constant throughout the sequence, or within the genes. There is generally no observable difference in the base composition between the coding and untranslated regions, other than the drop in G+C content at both termini of the genes. Within the coding region of the genes there is a periodicity in the G+C content. This is caused by the bias towards a G or C residue in the third, redundant position in the codon. In IE gene 1, the G+C content of the third position is 88.3%, compared with 70.8% and 67.0% for the first two positions. The G+C content of the coding regions may affect the coding capacity of a gene, although conservative amino acid changes can compensate for the extreme base composition, at least to some extent. In addition, more drastic amino acid changes might be tolerated in less crucial regions of the protein.

4.2.1 Structure of IE gene 1

IE gene 1, encoding the protein IE110, is entirely contained within the R_L sequences. The gene is orientated from right to left in IR_L, and the 5' end of the gene extends beyond the right boundary of the BamHI b fragment. The 5' and 3' termini of the transcript have been previously mapped (Mackem and Roizman, 1982b; Rixon et al., 1984) but did not correlate with the estimated size of the transcript (Watson et al., 1979). The gene has subsequently been shown to be spliced (Perry et al., 1986). The intron boundaries have been finely mapped by F.J. Rixon, and the results agree with the sequence analysis. Mutational analysis of the structure of the gene by R.D. Everett also correlates well.

Both introns contain repetitive DNA sequences. Downstream of the 3' end of the gene is a further set of reiterated sequences. These repeat sequences possibly play no role in the regulation of transcription, but may define the functional limits of the gene. Rixon et al. (1984) have reviewed several examples of sets of reiterated sequences occurring within the introns and downstream of the 3' terminus of genes. Analysis of the DNA sequences suggested that all three exons are polypeptide coding. An antiserum prepared by M.C. Frame against an oligopeptide representing the predicted sequence validated the reading frame assignment at the carboxy terminus.

4.2.2 Gene organisation in the U_L sequences of BamHI b and BamHI e

The 5' and 3' termini of IE gene 2 have been mapped, and the gene has been shown to be unspliced (Mackem and Roizman, 1982b; Whitton *et al.*, 1983). The five additional predicted proteins were identified by ORF analysis. Confidence in the identification of polypeptide coding sequences is based on the similarity of codon usage of the reading frames to other HSV-1 genes.

The gene arrangement in these regions of U_L appears to be fairly compact, and this is also the impression from other sequence studies within U_L (see for instance, McGeoch *et al.*, 1986b). The proposed 5' and more particularly 3' untranslated sequences of the genes in the U_L regions of BamHI *b* and *e* are shorter than those found with IE gene 1. Further, none of the genes in U_L analysed here are believed to be spliced. An example of a spliced late gene mapping in U_L has been reported (Costa *et al.*,1985). The 4 kb intron lying within the coding region of this late gene contains a 3' coterminal family of genes, further demonstrating the compact genetic arrangement.

The U_L sequences of BamHI *b*, adjacent to IR_L, contain three entire genes. The genes all have distinct promoter and polyadenylation signals. None of the genes overlap. IE gene 2, lying in a left to right organisation, is located near to the BamHI site. Just downstream of IE gene 2, in the same orientation, lies the gene for the 20.5K protein. Between this gene and the IR_L/U_L junction lies the third gene. This is located on the opposite DNA strand. The TATA box sequence of this gene lies close (44 bp) to the IR_L/U_L

junction.

The presence of the gene for the 21.2K protein is supported by the phenotype of a mutated strain of HSV-1, HFEM. HFEM contains a large deletion between positions 3776 and 7227, which is associated with loss of virulence (Rosen and Darai, 1985; Darai, personal communication). Pathogenicity can be rescued with the BamHI b fragment (Rosen et al., 1985). The mutation must therefore lie within the BamHI b sequences and is probably due to the deletion, although this has not been formally shown. The deletion removes the TATA box and upstream sequences of the gene for the 21.2K protein. Any gene entirely contained within IR_L and affected by the deletion, should not alter the phenotype of the virus, unless the copy in TR_L was also non-functional.

This interpretation is supported by the structure of three further strains. R19, another apathogenic strain, contains a 0.4 kb deletion which starts at the same position in U_L, but does not extend as far into IR_L (Becker et al., 1986). HSV-1 strains SP11 and LP contain mutations which result in the loss of the HpaI site at position 3686, 90 bp to the left of the HFEM deletion, but retain their virulence. This led to the suggestion that the apathogenic phenotype of HFEM is not due to the loss of sequences at the left border of the deletion (Becker et al., 1986). Referring to the sequence interpretation, this HpaI site lies within the 5' non-coding region of the 21.2K gene. It could be predicted that a small sequence change at this position would have no effect on 21.2K expression. This is consistent with the sequence interpretation.

The U_L sequence in BamHI e is believed to contain

three genes. All are located on the same DNA strand and are in a left to right orientation. The TATA box of the gene UL1, encoding the 24.9K protein, is immediately adjacent to the TR_L/U_L junction. This suggests that upstream regulatory sequences may lie within R_L. It is thought that the mRNA for the 24.9K protein must be 3' coterminal with the mRNA for the next gene, gene UL2, encoding the 27.3K protein. The coding regions for the two proteins do not overlap. This situation is not unusual for HSV-1, as discussed by McGeoch et al. (1985) and Rixon and McGeoch (1985). There is no good TATA box homology upstream of the third gene (UL3) encoding the 24.4K protein. However, there are several A+T rich sequences upstream of the coding region, including the polyadenylation signal of the two upstream genes.

4.2.3 R_L sequences downstream of IE gene 1

Examination of the sequence data in this region of R_L does not allow any firm conclusions on the presence of additional genes. Several features suggesting the presence of further genes can be found, including transcriptional control sequences. In particular, there are four potential polyadenylation sites in addition to those of IE gene 1. Further A+T rich sequences closely resembling TATA boxes are present, and are very apparent against the G+C rich DNA in this region.

Evaluation of the codon usage of the ORFs in this region showed that none contained sequences typical of HSV-1 coding regions. Both the IE110 and IE175 genes were used as references, as their genes lie within the two major repeat elements and also have very high G+C

contents. A further analysis, using genes from U_L as reference, also gave similarly negative results. To investigate whether a distinct sub-set of genes existed in R_L, the codon usage of the individual ORFs were compared against each other. It could be envisaged that these potential genes, with their close proximity to each other, may be more closely related. However, none of the ORFs appeared to share a similar coding strategy.

As described in the results, much of the sequence in this region is composed of minor tandemly reiterated sequences. Reiterated sequences do occasionally occur in polypeptide coding sequences in HSV-1 (Rixon and McGeoch, 1984; McGeoch et al., 1985; Chou and Roizman, 1986). Usually however, the presence of reiterated sequences is believed to reflect regions which are not under selective pressure (Smith, 1976).

It is therefore possible that these sequences are non-coding. This would reflect the situation in VZV. Approximately 1.4 kb of sequence downstream of gene 61, the VZV homologue of IE gene 1, is believed to be non coding (Davison and Scott, 1986). This VZV sequence includes an 88 bp direct repeat of unknown function. It was suggested that this sequence may contain promoter elements of the downstream gene, gene 60. An HSV-1 homologue to the VZV gene 60 exists, present in an equivalent position and orientation, and will be discussed later.

4.3 NATURE OF THE R_L/U_L JUNCTIONS

4.3.1 Gene arrangement around the R_L/U_L junctions

From the sequence interpretation, it is suggested that there are no transcripts spanning the R_L/U_L junctions. The proposed TATA box sequences of two genes in U_L lie next to the two R_L/U_L junctions. Further upstream regulatory sequence elements may thus be present in the R_L sequence. Both genes are orientated so that they are transcribed in the direction away from R_L. The location of the junctions may define the functional limit of the two genes. It is possible that these two genes are differentially regulated by sequences upstream of the TATA boxes. For this reason, the R_L sequences, which are thought to be dynamic, mobile elements, may not be able to expand to include the 5' ends of the genes.

4.3.2 Comparison of the R_L/U_L and R_S/U_S junctions

TR_S and IR_S contain the upstream sequences and 5' non coding regions of genes which are transcribed into U_S (Watson *et al.*, 1981; Rixon and Clements, 1982; Murchie and McGeoch, 1982). The arrangement around the R_L/U_L junctions is quite similar, although the repeats probably do not extend into the transcribed regions of the genes. It is quite possible, as mentioned above, that the upstream regulatory sequences of these genes are located in R_L. If the two genes are necessarily differentially regulated, then perhaps R_L cannot expand further into the genes, to give an arrangement directly comparable to R_S.

R_L contains a 17 bp reiteration directly adjacent

to U_L. R_S also has a reiteration of a 22 bp sequence, although this is located 91 bp from the R_S/U_S junctions. This reiteration lies within IE genes 4 and 5, the two genes mentioned above (Murchie and McGeoch, 1982).

4.4 PROTEINS ENCODED IN THE BAMHI B AND BAMHI E SEQUENCES

4.4.1 IE110

Although its function in viral growth and replication is unknown, IE110 has been shown to regulate the expression of all classes of HSV-1 genes. The transactivational activities of IE110 were described in the introduction.

The coding region of the gene has a particularly extreme base composition, of 75.4% G+C. The only sequenced gene to exceed this value is another HSV-1 gene, IE gene 3, which is contained within R_S (McGeoch, et al., 1986a). Within the coding region of IE110 are regions with particularly high G+C contents. Most notable is a 1 kb region starting at or near the beginning of the third exon, which has a G+C content of over 80%. This compares with the base compositions of the coding regions upstream and downstream of this sequence, of 70-72% G+C. This feature may indicate a distinction between essential and non-critical domains of the protein.

The functional domains of the IE110 protein have been investigated. R.D. Everett has constructed a series of IE gene 1 mutants containing frameshifts or deletions within the coding sequences of the gene.

IE110 activity was totally abolished by the deletion taking out a large part of the second and third exon. However, a frameshift mutation near to the carboxy terminus of IE110 retained approximately 50% of IE110 activity. This suggests that the carboxy terminal amino acids are not essential for IE110 activity. Mutations within the second and third exons reduced transactivational activities by varying amounts.

A variety of regulatory and nucleic acid binding proteins contain regions with repetitive amino acid sequences containing high levels of Cys and/or His residues, which are thought to be involved in the coordination of metal ions. The most characteristic sequence element is a pair of Cys residues separated by two other amino acids (Miller et al., 1985). IE110 appears to belong to this class of proteins. Of the 14 Cys residues in IE110, nine are located between residues 99 and 156. Six of these take the form C-X-X-C. This suggests that the proposed interaction between IE110 and DNA during transcription regulation may be through direct binding of IE110 to the DNA.

HSV-2 IE118 is a functionally equivalent protein to the HSV-1 IE110. HSV-1 X HSV-2 intertypic recombinants containing only one functional IE gene 1 from either serotype are viable (Davison et al., 1981). The level of conservation of amino acid sequences between the two proteins is presently unknown and awaits the DNA sequence determination of this region of HSV-2. IE gene 1 lies within R_L, which has been shown to have only a low level of DNA sequence homology between the two serotypes (Davison and Wilkie, 1981). However this may be due to DNA sequences outside of the polypeptide coding regions.

In cotransfection assays, the effect of IE110 is similar to IE175. The amino acid sequences of the two proteins were compared. IE110 shows no homology to IE175, other than a Ser rich region between amino acids 554 to 594 of IE110 and between 177 and 199 in IE175. In addition, no homology has been found between the amino acid sequences of IE110 and the other HSV-1 IE proteins (IE175, IE68 and IE12 sequences from Murchie and McGeoch, 1982; McGeoch et al., 1985 and 1986a).

A homologue to IE110 was detected in VZV. The two proteins differ in size, and the level of homology is not uniform throughout the proteins. The most highly conserved region corresponds to the Cys rich locality of IE110. All the C-X-X-C sequences have been conserved between the two proteins. In addition a further Cys residue and a His residue have been conserved. The conservation of this region may suggest that the ability to bind nucleic acid is an essential requirement of the proteins.

4.4.2 IE63

IE63 is an essential IE protein which plays a role in the regulation of HSV-1 gene expression (Sacks et al., 1985). The activities of IE63 have been described in the introduction.

The gene for IE63, IE gene 2, is located near to the right end of U_L (Clements et al., 1977). The complete DNA sequence of the gene and the amino acid sequence of the protein were determined during this work. The 5' and 3' ends of the transcript have been finely mapped, and the gene has been shown to be unspliced (Whitton et al., 1983).

The polypeptide coding region of the gene has a G+C content of 69.4%. This base composition is tolerated in a similar manner to IE gene 1, with a biased codon usage. The base composition varies considerably within the polypeptide coding regions of the gene. The amino 50% portion of the protein is coded by sequences with 73.9% G+C content, whilst the carboxy 50% of the protein is coded by 64.7% G+C DNA.

A VZV protein, encoded by gene 4, shows considerable homology to IE63. The amino portion of the protein shows least conservation of amino acid residues. The pattern of amino acid sequence conservation between IE63 and the EBV homologue (BKRF3) is similar to that observed with VZV. This may reflect a non-critical domain, or possibly an HSV-specific function, at the amino terminus of the protein. This would support the suggestion that within a single gene, polypeptide coding sequences with higher G+C contents may represent non-crucial regions of HSV-1 proteins (McGeoch et al., 1986a).

4.4.3 Predicted proteins encoded in BamHI b and BamHI e

In addition to IE110 and IE63 genes, the genes for five further proteins were sequenced. None of these proteins have previously been characterised, and no function has been assigned to them.

Four of the predicted proteins, the 20.5K in BamHI b, and the UL1, UL2 and UL3 products in BamHI e, have homologues in VZV. A homologue to UL2 was also detected in EBV. The presence of homologues supports

the interpretation given to the sequence data. The final predicted protein, 21.2K, has no detectable homologue. This may suggest that its function is specific to HSV. The 21.2K protein is encoded by the gene affected by the HFEM deletion, and may be involved, directly or indirectly, in determining virulence.

The UL1 protein has a strongly hydrophobic amino terminus, suggesting it is membrane translated. Although the VZV homologue of IE63 is considerably shorter, and shows only a low level of homology at the amino terminus, this feature has been retained.

4.5 RELATION OF HSV-1 WITH OTHER HERPESVIRUSES

4.5.1 Relation between HSV-1 and VZV

VZV and HSV-1 are distantly related alphaherpesviruses. A large number of homologous proteins have been found in the two viruses (see for example, Davison and McGeoch, 1986). The DNA sequence determination and interpretation of the entire S segments of both HSV-1 and VZV has enabled a complete analysis of homologous genes in the region (Davison and McGeoch, 1986). The structures and genetic arrangements of the S segments of the two viruses and the location of corresponding genes has been described in the introduction. The comparison of the two sequences has suggested that the repeats are dynamic entities, which have undergone recombination during their evolution which has resulted in the rearrangement of sequences within the S segment, and the change in length of R_S and U_S.

VZV has no major repeat structures equivalent to the HSV-1 IR_L and TR_L b sequence, although it does have a short 88.5 bp repeat sequence flanking U_L (Davison, 1984). This sequence shows no homology to the HSV-1 sequences presented here. Possibly due to the absence of the b and also the a sequence, in only a small proportion of genomes, the U_L sequence of VZV is present in an inverted orientation. The 5% of isomers observed with an inverted U_L sequence may be generated by aberrant cleavage of unit lengths at the incorrect L-S joint in the nascent concatemeric DNA during replication (Davison, 1984). The predominant arrangement of U_L in VZV is inverted relative to the arbitrary prototype arrangement of HSV-1. (The prototype VZV genome is equivalent to the I_L arrangement of HSV-1.)

The size of the U_L sequence of the HSV-1 genome, estimated from restriction enzyme digests, is approximately 106 kbp (Clements et al., 1976). The U_L sequence of VZV has been entirely sequenced and is 105 kbp in length (Davison and Scott, 1986). The base composition of the two sequences varies markedly. VZV has a G+C content of 46.0%. Regions of HSV-1 U_L that have been sequenced have base compositions of around 62-67% G+C (McGeoch et al., 1986; Quinn and McGeoch, 1985; Dalrymple et al., 1985) The R_L sequences of VZV is only 88.5 bp in length and shows no similarity to the R_L sequences of HSV-1. The absence of conserved sequences in this region was first demonstrated by DNA hybridisation experiments (Davison and Wilkie, 1983).

VZV has far fewer reiterated sequence elements than HSV-1. This became apparent in the comparison between the S sequences of HSV-1 and VZV (Davison and McGeoch, 1986). The reiterations VZV does contain are

not related in either sequence or location in the genome (Davison and Scott, 1986). All the reiteration sets in the HSV-1 sequence presented here lie within R_L. The 88.5 bp R_L sequence of VZV contains no reiterations. None of the sequences show homology to any of the reiteration sets present in the HSV-1 sequence presented here. If reiterations represent "parasitic" sequences which are not under selective pressure (Davison and Scott, 1986), homology between these sequences in distantly related viruses would not be expected.

Of the seven genes sequenced, five have detectable homologues in VZV. The three genes in the BamHI e U_L sequence all have homologues in VZV. The orientation and relative arrangement of the three genes is the same in both viruses. In VZV there is a gene located between this set of homologous genes and the U_L boundary. This gene encodes the protein corresponding to IE110. The relative arrangement of the four genes has been conserved between the two viruses, although the location of the R_L/U_L junction has altered. Chou and Roizman (1986) have mapped a gene to the R_L sequences upstream of IE110. This shows no homology to any VZV ORF in this region.

The genes for IE63 and the 20.5K protein in the U_L sequence of BamHI b have homologues in VZV. The location and orientation of the two genes is similar. However, HSV-1 has a further gene between the U_L boundary and these two genes. The gene is located on the opposite strand of DNA and encodes the 21.2K protein. There is no corresponding gene elsewhere in the VZV genome. In VZV, this region of the genome contains two ORFs. Neither shows homology to any HSV-1 gene sequenced. These two VZV genes may have

homologues elsewhere in the HSV-1 genome, or alternatively, may encode VZV specific functions.

The level of conservation of amino acid sequences between pairs of homologous proteins varies. In the majority of cases however, the amino terminal portions are least conserved, both in sequence and in length. This feature has also been observed in the proteins coded by genes in the S segment (McGeoch et al., 1986; Davison and McGeoch, 1986).

The organisation of the corresponding genes in the two genomes gives an indication of the recombinational events which must have occurred to generate the rearrangements. Although the relative organisation of the homologous genes has been maintained, the position of the R_L/U_L junction is altered. HSV-1 R_L contains at least two genes, whereas the R_L sequence of VZV contains no genes. The expansion/contraction of the R_L sequences appears to have resulted in the loss of genes in both viruses. This can be observed by identifying the equivalent genes in the U_L sequences adjacent to TR_L in VZV, with those in the corresponding region in HSV-1, contained in the BamHI b fragment. This region of the VZV genome contains two genes, neither of which show any homology to the HSV-1 gene located in the corresponding position.

The only region sequenced to which no genes could be assigned lies within R_L downstream of IE gene 1. All the ORFs in this region were compared to the predicted ORFs of the entire VZV genome. No homology was found between any ORFs. Although this is not conclusive evidence, it is consistent with the DNA sequence interpretation. It is of interest to note

that in the equivalent position in the VZV genome, there is a 1.4 kb region to which no gene has been asigned. This region contains an 88 bp direct repeat. This region may prove to have some role other than encoding proteins, which has not yet been recognised.

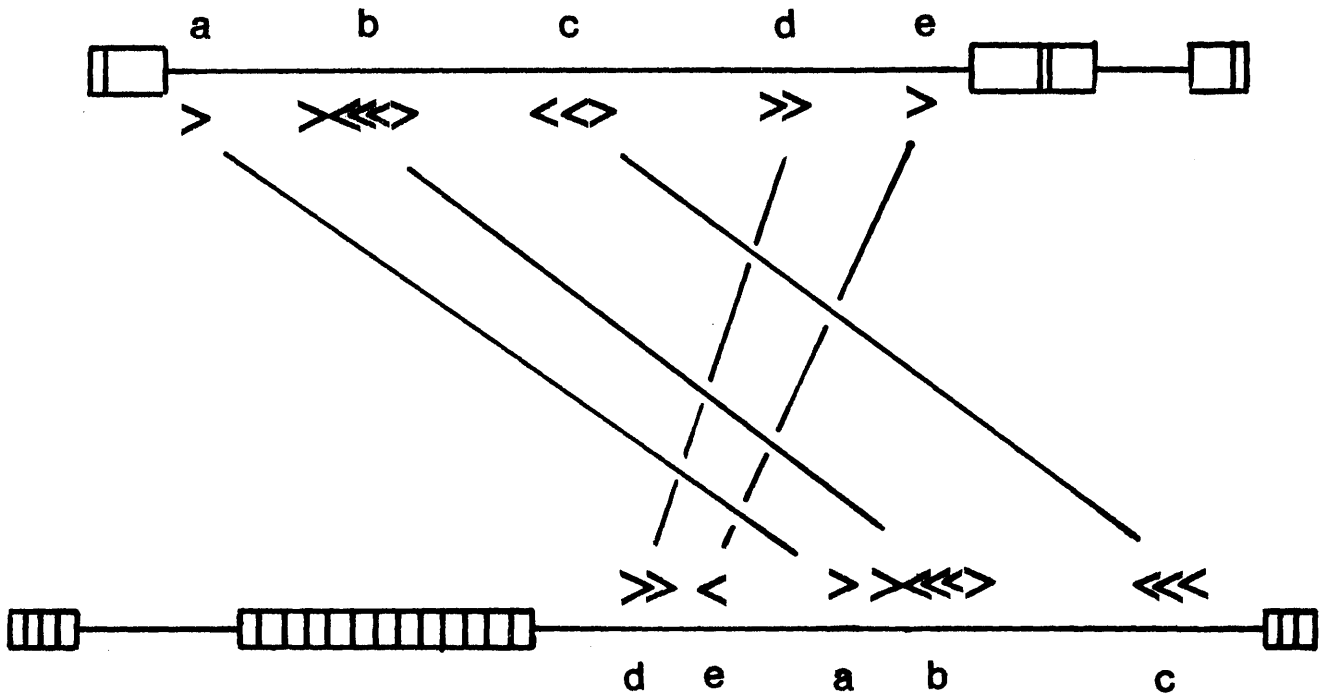
4.5.2 Relation between HSV-1 and EBV

The structures of the genomes of HSV-1 and EBV are very different, as described in the introduction. EBV has no equivalent to the R_L sequence. Consequently there are difficulties in predicting where corresponding genes may be located in the two genomes. A general search was made of all EBV ORFs for homologues to the HSV-1 proteins described here. This was followed by a closer examination of ORFs near to the homologues found. Only two proteins from the EBV ORFs showed homology to the HSV-1 proteins. Homology between HSV-1 UL2 and EBV BKRF3 was very high.

IE gene 2 is the only IE gene which has a detectable homologue in EBV. The protein encoded by the homologous gene, BMLF1, is expressed early in infection (Cho et al., 1985). Although IE63 is expressed at IE times, it appears to be required later in infection, when it plays an essential role in the induction of late gene expression (Sacks et al., 1985). It is thus plausible that the two genes encode equivalent functions.

The location of the two pairs of corresponding genes in the two genomes differed. This suggests that extensive rearrangement of genetic material has occurred during the evolution of the two viruses. Several homologous EBV genes have been detected. These

HSV-1



EBV

Figure 4.1 Relationship between the HSV-1 and EBV genomes

The location and orientation of all homologous proteins detected to date, including the work described here, are illustrated. The HSV-1 genome is in the P orientation at the top of the figure. From left to right, in the HSV-1 genome, homologous genes indicated are as follows. a) the UL2 gene, b) the exonuclease and neighbouring genes, c) glycoprotein B, major DNA binding protein and polymerase genes, d) genes for both the ribonucleotide reductase subunits, e) and IE gene 2 (McLauchlan and Clements, 1983; Gibson et al., 1984; Quinn and McGeoch, 1985; McGeoch et al., 1986b; and the work described in this thesis).

include the genes equivalent to the HSV-1 glycoprotein B, major DNA binding protein, DNA polymerase, both ribonuclease reductase subunits, exonuclease, and four genes neighbouring the exonuclease (Gibson et al., 1984; Quinn and McGeoch, 1985; McGeoch et al, 1986b). Figure 4.1 shows the arrangement of these genes together with the location of the genes equivalent to IE63 and UL2. This figure illustrates the large scale rearrangements of genes between the two viruses, but also shows that the relative arrangements of several genes has been maintained. This is most clearly observed in the region containing the exonuclease gene. The location of the two EBV homologues detected in the BamHI b and e sequences is consistent with the location of other homologous genes identified.

4.6 EVOLUTION OF R_L

The comparative study of the genetic arrangement of the S segment of HSV-1 and VZV by Davison and McGeoch (1986) was described in the introduction. A variable level of amino acid sequence conservation between pairs of homologous proteins was observed together with an extensive relocation of genes. To explain the differences in gene layout, the report described a process by which R_S could expand and contract, thereby including or expelling genes from its sequence.

An analogous comparison of the HSV-1 R_L genes has been attempted. The VZV genome does not have a major repeat sequence flanking U_L, though the HSV-1 IE gene 1 contained in R_L does have a VZV homologue in the U_L sequence of the VZV genome. The homologous genes have retained their same relative positions in VZV and

HSV-1. There are additional genes present at one terminus of U_L in both viruses, but their origin and evolution cannot be deduced from the present data.

Contraction and expansion of the R_L sequence of HSV-1 could be envisaged. For example, this could result in the inclusion/exclusion of IE gene 1 in the R_L sequence. According to the DNA sequence interpretation, there is approximately 3 kb of non-coding sequence in R_L downstream of IE gene 1. Loss of these sequences may not be strongly detrimental to the virus, providing transcriptional regulatory sequences w ere retained and the overall length of the genome remained within the size limits required for packaging into nucleocapsids.

Further expansion of R_L into U_L, involving loss of sequences from one terminus of U_L, would be limited by the genetic content of the region. The two genes located at each end of U_L are both orientated in a direction away from R_L. The functions of their encoded proteins have not been established. Exchange of the upstream sequences could only be tolerated if the subsequent regulation of expression of the modified gene did not adversely affect viral growth and replication. The effect of more drastic changes cannot be predicted without knowledge of the functions of the proteins encoded by the genes which would be modified or deleted.

The origin of the R_L sequence cannot be deduced from existing knowledge. Features such as the high level of reiterations suggest that the apparently redundant sequences within R_L may have been generated through aberrant recombinational events within sequences not under selective pressure. The

transcriptional regulatory sequences present within this redundant sequence may represent remnants of genes previously located in this region of the genome.

REFERENCES

REFERENCES

Al-Saadi, S.A., Clements, G.B. and Subak-Sharpe, J.H. (1983). Viral genes modify herpes simplex virus latency in mouse footpad and sensory ganglia. J. Gen. Virol. 64, 1175-1179.

Anderson, K.P., Costa, R.H., Holland, L.E. and Wagner, E.K. (1980). Characterisation of herpes simplex virus type 1 RNA present in the absence of de novo protein synthesis. J. Virol. 34, 9-27.

Bachenheimer, S.L. and Roizman, B. (1972) Ribonucleic acid sequences in cells infected with herpes simplex virus: VI Polyadenylic acid sequences in viral messenger ribonucleic acid. J.Virol. 10, 875-879.

Baer, R., Bankier, A.T., Biggin, M.D., Deininger, P.L., Farrell, P.J., Gibson, T.J., Hatfull, G., Hudson, G.S., Satchwell, S.C., Seguin, C., Tuffnell, P.S. and Barrell, B.G. (1984). DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310, 207-211.

Batterson, W and Roizman, B. (1983). Characterisation of the herpes simplex virion-associated factor responsible for the induction of genes. J. Virol. 46, 371-377.

Becker, Y., Hadar, J., Tabor, E., Ben-Hur, T., Raibstein, I., Rosen, A. and Daza, G. (1986). A sequence in Hpal-P fragment of herpes simplex virus-1 DNA determines intraperitoneal virulence in mice. Virology 149, 255-259.

Berg, J.M. (1986). Potential metal-binding domains in nucleic acid binding proteins. Science 232, 485-487.

Biggin, M.D., Gibson, T.J. and Huang, G.F. (1983). Buffer gradient gels and ^{35}S label as an aid to rapid sequence determination. Proc. Natl. Acad. Sci., U.S.A. 80, 3963-3965.

Boyer, H.W. and Roulland-Dussoix, D. (1969). A complementation analysis of the restriction modification of DNA in Escherichia coli. J. Mol. Biol. 41, 459-472.

Buckmaster, E.A., Gompels, U. and Minson, A. (1984). Characterisation and physical mapping of an HSV-1 glycoprotein of approximately 115 X 10^3 molecular weight. Virology 139, 408-413.

Bzik, D.J. and Preston, C.M. (1986). Analysis of DNA sequences which regulate the transcription of herpes simplex virus immediate early gene 3: DNA sequences required for enhancer-like activity and response to trans-activation by a virion polypeptide. Nucleic Acids Res. 14, 929-943.

Campbell, M.E., Palfreyman, J.W. and Preston, C.M. (1984). Identification of herpes simplex virus DNA sequences which encode a transacting polypeptide responsible for stimulation of immediate-early transcription. J. Mol. Biol. 180, 1-19.

Cho, M.-S., Jeang, K.-T. and Hayward, S.D. (1985). Localization of the coding region for an Epstein-Barr virus early antigen and inducible expression of this 60-kilodalton nuclear protein

in transfected fibroblast cell lines. J. Virol. 56, 852-859.

Chou, J. and Roizman, B. (1986). The terminal a sequence of the herpes simplex virus genome contains the promoter of a gene located in the repeat sequences of the L component. J. Virol. 57, 629-637.

Clements, J.B., Cortini, R. and Wilkie, N.M. (1976). Analysis of herpesvirus DNA substructure by means of restriction endonucleases. J. Gen. Virol. 30, 243-256.

Clements, J.B., McLauchlan, J. and McGeoch, D.J. (1979). Orientation of herpes simplex virus type 1 immediate-early mRNAs. Nucleic Acids Res. 7, 77-91.

Clements, J.B., Watson, R.J. and Wilkie, N.M. (1977). Temporal regulation of herpes simplex virus type 1 transcription: location of transcripts on the viral genome. Cell 12, 275-285.

Clewell, D.B. (1972). Nature of Col El plasmid replication in Escherichia coli in the presence of chloramphenicol. J. Bact. 110, 667-676.

Cohen, G.H. (1972). Ribonucleotide reductase activity of synchronised KB cells infected with herpes simplex virus. J. Virol. 9, 408-418.

Conley, A.J., Knipe, D.M., Jones, P.C. and Roizman, B. (1981). Molecular genetics of herpes simplex virus. VII. Characterisation of a temerature-sensitive mutant produced by in vitro

mutagenesis and defective in DNA synthesis and accumulation in γ -polypeptides. J. Virol. 37, 191-206.

Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C. and Chambon, P. (1980). Promoter sequences of eukaryotic protein-coding genes. Science 209, 1406-1414.

Cordingley, M.G., Campbell, M.E. and Preston, C.M. (1983). Functional analysis of a herpes simplex type 1 promoter: identification of far-upstream regulatory sequences. Nucleic Acids Res. 11, 2347-2365

Cortini, R. and Wilkie, N.M. (1978). Physical maps for HSV type 2 DNA with five restriction endonucleases. J. Gen. Virol. 39, 259-280.

Costa, R.H., Draper, K.G., Banks, L., Powell, K.L., Cohen, G., Eisenberg, R. and Wagner, E.K. (1983). High resolution mapping of herpes simpex virus type 1 transcripts encoding alkaline exonuclease and a 50,000-dalton protein tentatively identified as a capsid protein. J, Virol. 48, 591-603.

Costa, R.H., Draper, K.G., Kelly, T.J. and Wagner, E.K. (1985). An unusual spliced herpes simplex virus type 1 transcript with sequence homology to Epstein-Barr virus DNA. J. Virol. 54, 317-328.

Costanzo, F., Campadelli-Fiume, G., Foa-Tomasi, L. and Cassai, E. (1977). Evidence that herpes simplex virus DNA is transcribed by cellular RNA polymerase B. J. Virol. 21, 996-1001.

Dalrymple, M.A., McGeoch, D.J., Davison, A.J. and Preston, C.M. (1985). DNA sequence of the herpes simplex virus type 1 gene whose product is responsible for transcriptional activation of immediate early promoters. Nucleic Acids Res. 13, 7865-7879.

Davison, A.J. (1983). DNA sequence of the U_S component of the varicella-zoster virus genome. EMBO J. 2, 2203-2209.

Davison, A.J. (1984). Structure of the genome termini of varicella-zoster virus. J. Gen. Virol. 65, 1969-1977.

Davison, A.J., Marsden, H.S. and Wilkie, N.M. (1981). One functional copy of the long terminal repeat gene specifying the immediate-early polypeptide IE 110 suffices for a productive infection of human feotal lung cells by herpes simplex virus. J. gen. Virol. 55, 179-191.

Davison, A.J. and McGeoch, D.J. (1986). Evolutionary comparisons of the S segments in the genomes of herpes simplex virus type 1 and varicella-zoster virus. J. gen. Virol. 67, 597-611.

Davison, A.J. and Scott, J.E. (1985). DNA sequence of the major inverted repeat in the varicella-zoster virus genome. J. gen. Virol. 66, 207-220.

Davison, A.J. and Scott, J.E. (1986). The complete DNA sequence of the varicella-zoster virus. J. Gen. Virol. In press.

Davison, A.J. and Wilkie, N.M. (1981). Nucleotide

sequences of the joint between the L and S segments of herpes simplex virus types 1 and 2. J. gen. Virol. 55, 315-331.

Davison, A.J and Wilkie, N.M. (1983a). Inversion of the two segments of the herpes simplex virus genome in intertypic recombinants. J. Gen. Virol. 64, 1-18.

Davison, A.J. and Wilkie, N.M. (1983b). Location and orientation of homologous sequences in the genomes of five herpesviruses. J. gen. Virol. 64, 1927-1942.

Dayhoff, M.O., Barker, W.C. and Hunt, L.T. (1983). Establishing homologies in protein sequences. Methods Enzymology 91, 524-545.

Deininger, P.L. (1983). Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. Analytical Biochem. 129, 216-223.

Deiss, L. and Frenkel, N. (1986). Herpes simplex virus amplicon: cleavage of concatemeric DNA is linked to packaging and involves amplification of the terminally reiterated a sequence. J. Virol. 57, 933-941.

Delius, H. and Clements, J.B. (1976). A partial denaturation map of herpes simplex virus type 1 DNA: Evidence for inversions of the unique DNA regions. J. Gen. Virol. 33, 125-133.

DeLuca, N.A. and Schaffer, P.A. (1985). Activation of immediate-early, early and late promoters by temperature-sensitive and wild-type forms of herpes simplex virus type 1 protein ICP4. Mol.

Cell. Biol. 5, 1997-2008.

DeLuca, N.A., Courtney, M.A. and Schaffer, P.A. (1984). Temperature-sensitive mutants in herpes simplex virus type 1 ICP4 permissive for early gene expression. J. Virol. 52, 767-776.

DeLuca, N.A., McCarthy, A.M. and Schaffer, P.A. (1985). Isolation and characterisation of deletion mutants of herpes simplex virus type 1 in the gene encoding immediate-early regulatory protein ICP4. J. Virol. 56, 558-570.

Devereux, J., Haeberli, P. and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12, 387-395.

Dutia, B.M. (1983). Ribonucleotide reductase induced by herpes simplex virus has a virus-specified constituent. J. Gen. Virol. 64, 513-521.

Everett, R.D. (1983). DNA sequence elements required for regulated expression of the HSV-1 glycoprotein D gene lie within 83 bp of the RNA capsites. Nucleic Acids Res. 11, 6647-6666.

Everett, R.D. (1984a). A detailed analysis of an HSV-1 early promoter: sequences involved in trans-activation by viral immediate-early gene products are not early-gene specific. Nucleic Acids Res. 12, 3037-3056.

Everett, R.D. (1984b). Transactivation of transcription by herpes virus products: requirement for two HSV-1 immediate-early polypeptides for maximum activity. EMBO J. 3, 3135-3141.

Fitzgerald, M. and Shenk, T. (1981). The sequence 5'-AAUAAA-3' forms part of the recognition site for polyadenylation of late SV40 mRNAs. Cell 24, 251-260.

Frame, M.C., Marsden, H.S. and Dutia, B.M. (1985). The ribonucleotide reductase induced by herpes simplex virus type 1 involves minimally a complex of two polypeptides (136K and 38K). J. Gen. Virol. 66, 1581-1587.

Frame, M.C., Marsden, H.S. and McGeoch, D.J. (1986). Novel herpes simplex virus type 1 glycoprotein identified by antiserum against a synthetic oligopeptide from the predicted product of US4. J. Gen Virol. 67, 745-751.

Freeman, M.J. and Powell, K.L. (1982). DNA-binding properties of a herpes simplex virus immediate early protein. J.Virol. 44, 1084-1087.

Friedman, A., Shlomai, J. and Becker, Y. (1977). Electron microscopy of herpes simplex virus DNA molecules isolated from infected cells by centrifugation in CsCl. J. Gen. Virol. 34, 507-522.

Furlong, D., Swift, H. and Roizman, B. (1972). Arrangement of herpesvirus deoxyribonucleic acid in the core. J. Virol. 10, 1071-1074.

Garoff, H. and Ansorge, W. (1981). Improvements of DNA sequencing gels. Analytical Biochem. 115, 450-457.

Gelman, I.H. and Silverstein, S. (1985). Identification of immediate early genes from herpes simplex virus that transactivate the virus thymidine kinase. Proc. Natl. Acad. Sci., U.S.A. 82, 5265-5269.

Gibson, T, Stockwell, P., Ginsburg, M. and Barrell, B. (1984). Homology between two EBV early genes and HSV ribonucleotide reductase and 38K genes. Nucleic Acids Res. 12, 5087-5099.

Gingeras, T.R., Sciaky, D, Gelinas, R.E, Bing-Dong, J., Yen, C.E., Kelly, M.M., Bullock, P.A., Parsons, B.L., O'Neill, K.E. and Roberts, R.J. (1982). Nucleotide sequences from the adenovirus-2 genome. J. Biol. Chem. 257, 13475-13491.

Green, M.T., Courtney, R.J. and Dunkel, E.C. (1981). Detection of an immediate early herpes simplex virus type 1 polypeptide in trigeminal ganglia from newly infected animals. Infect. Immunol. 34, 987-992.

Haar, L. and Marsden, H.S. (1981). Two dimensional gel analysis of HSV type 1 induced polypeptides and glycoprotein processing. J. Gen. Virol. 52, 77-92.

Hall, L.M., Draper, K.G., Frink, R.J., Costa, R.H. and Wagner, E.K. (1982). Herpes simplex virus mRNA species mapping in EcoRI fragment I. J. Virol. 43, 594-607.

Hanahan, D. (1983). Studies on transformation of Escherichia coli with plasmids. J. Mol. Biol. 166, 557-580.

Harris-Hamilton, E. and Bachenheimer, S.L. (1985).

Accumulation of herpes simplex virus type 1 RNAs of different kinetic classes in the cytoplasm of infected cells. J. Virol. 53, 144-151.

Hay, R.T. and Hay, J. (1980). Properties of herpesvirus-induced "immediate-early" polypeptides. Virology 104, 230-234.

Hayward, G.S., Jacob, R.J., Wadsworth, S.C. and Roizman, B. (1975). Anatomy of herpes simplex virus DNA: evidence for four populations of molecules that differ in the relative orientations of their long and short components. Proc. Natl. Acad. Sci. 72, 4243-4247.

Holland, L.E., Anderson, K.P., Shipman, C. and Wagner, E.K. (1980). Viral DNA synthesis is required for the efficient expression of specific herpes simplex virus type 1 mRNA species. Virology 101, 10-24.

Honess, R.W. (1984). Herpes simplex and "the herpes complex": diverse observations and a unifying hypothesis. J. Gen. Virol. 65, 2077-2107.

Honess, R.W., Powell, K.L., Robinson, D.J., Sim, C. and Watson, D.H. (1974). Type specific and type common antigens in cells infected with herpes simpex virus type 1 and on the surface of naked and enveloped particles of the virus. J. Gen. Virol. 22, 159-169.

Honess, R.W. and Roizman, B. (1974). Regulation of herpesvirus macromolecular synthesis. I. Cascade regulation of the synthesis of three groups of viral proteins. J. Virol. 14, 8-19.

Honess, R.W. and Roizman, B. (1975). Regulation of herpesvirus macromolecular synthesis: sequential transition of polypeptide synthesis requires functional viral polypeptides. Proc. Natl. Acad. Sci. U.S.A. 72, 1276-1280.

Honess, R.W. and Watson, D.H. (1977). Unity and diversity in the herpesviruses. J. gen. Virol. 37, 15-37.

Jacob, R.J., Morse, L.S. and Roizman, B. (1979). Anatomy of herpes simplex virus DNA. XII. Accumulation of head-to-tail concatamers in nuclei of infected cells and their role in the generation of the four isomeric arrangements of viral DNA. J. Virol. 29, 448-457.

Jacquemont, B., Verrier, B., Epstein, A.L. and Machuca, I. (1984). Expression of immediate-early genes in herpes simplex virus type 1-infected XC cells: lack of ICP22(68K) polypeptide. J. gen. Virol. 65, 1331-1340

Jenkins, F.J., Casadaban, M.J. and Roizman, B. (1985). Application of the mini-Mu-phage for target-sequence- specific insertional mutagenesis of the herpes simplex virus genome. Proc. Natl. Acad. Sci. 82, 4773-4777.

Johnson, P.A., MacLean, C., Marsden, H.S., Dalziel, R.G. and Everett, R.D. (1986). The product of gene US11 of herpes simplex virus type 1 is expressed as a true late gene. J. Gen. Virol. 67, 871-883.

Kaerner, H.C., Maichle, I.B., Ott, A. and Schroder, C.H. (1979). Origin of two different classes of defective HSV-1 Angelotti DNA. Nucleic Acids Res. 6, 1467-1478.

Katz, L., Kingsbury,D.T. and Helinski, D.R. (1973). Stimulation by cyclic adenosine monophosphate of plasmid deoxyribonucleic acid replication and catabolic repression of the plasmid deoxyribonucleic acid-protein relaxation complex. J. Bact. 114, 577-591.

Keir, H.M. and Gold, E. (1963). Deoxyribonucleic acid nucleotidyltransferase and deoxribonuclease from cultured cells infected with herpes simplex virus. Biochim. Biophys. Acta 72, 263-276.

Keir, H.M., Subak-Sharpe, H., Shedden, W.I.H., Watson, D.H. and Wildy, P. (1966). Immunological evidence for a specific DNA polymerase produced after infection by herpes simplex virus. Virology 30, 154-157.

Kieff, E.D., Bachenheimer, S.L. and Roizman, B. (1971). Size, composition and structure of the deoxyribonucleic acid of herpes simplex virus subtypes 1 and 2. J. Virol. 8, 125-132.

Kieff, E.D., Hoyer, B., Bachenheimer, S.C. and Roizman, B. (1972). Genetic relatedness of type 1 and type 2 herpes simplex viruses. J. virol. 9, 738-745.

Kit, S. and Dubbs, D.R. (1963). Acquisition of thymidine kinase activity by herpes simplex virus infected mouse fibroblast cells. Biochem.

Biophys. Acta 11, 55-59.

Kozak, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Res. 12, 857-872.

Kudler, C., Jones, T.R., Russell, R.J. and Hyman, R.W. (1983). Heteroduplex analysis of cloned fragments of herpes simplex virus DNAs. Virology 124, 86-99.

Kyte, J. and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105-132.

Littler, E., Purifoy, D.J.M., Minson, A.C. and Powell, K.L. (1983). Herpes simplex non-structural proteins. III. Function of the major DNA binding protein. J. Gen. Virol. 64, 983-995.

Locker, H. and Frenkel, N. (1979). Structure and origins of defective genomes contained in serially passaged herpes simplex virus type 1 (Justin). J. Virol. 29, 1065-1077.

Lofgren, K.W., Stevens, J.G., Marsden, H.S. and Subak-Sharpe, J.H. (1977). Temperature sensitive mutants of herpes simplex virus differ in their capacity to establish latent infections in mice. Virology 76, 440-443.

Low, M., Hay, J. and Keir, H.M. (1969). DNA of herpes simplex virus is not a substrate for methylation in vivo. J. Mol. Biol. 46, 205-207.

Lowe, P.A. (1978). Levels of DNA-dependent RNA

polymerases in herpes simplex virus-infected BHK 21 C13 cells. Virology 86, 577-580.

Ludwig, H.O., Biswal, N. and Benyesh-Melnick, M. (1972). Studies on the relatedness of herpesviruses through DNA-DNA hybridization. Virology 49, 95-101.

Mackem, S. and Roizman, B. (1982a). Differentiation between α-promoter and regulator regions of herpes simplex virus 1: the functional domains and sequence of a movable regulator. Proc. Natl. Acad. Sci. U.S.A. 79, 4917-4921.

Mackem, S. and Roizman, B. (1982b). Structural features of the herpes simplex virus α gene 4, 0, and 27 promoter-regulatory sequences which confer α regulation on chimeric thymidine kinase genes. J. Virol. 44, 939-949.

Marsden, H.S., Crombie, I.K. and Subak-Sharpe, J.H. (1976). Control of protein synthesis in herpes virus-infected cells: Analysis of the polypeptides induced by wild type and sixteen temperature-sensitive mutants. J. Gen. Virol. 31, 347-372.

Marsden, H.S., Haarr, L. and Preston, C.M. (1983). Processing of herpes simplex virus proteins and evidence that translation of thymidine kinase mRNA is initiated at three separate AUG codons. J. Virol. 46, 434-445.

Matthews, R.E.F. (1982). Classification and nomenclature of viruses. Intervirology 17, 47-51.

Mavromara-Nazos, P., Silver, S., Hubenthal-Voss, J., McKnight, J.C.L. and Roizman, B. (1986). Regulation of herpes simplex virus 1 genes: α gene sequence requirements for transient induction of indicator genes regulated by β or late (γ_2) promoters. Virology 149, 152-164.

McGeoch, D.J. and Davison, A.J. (1986). Alphaherpesviruses possess a gene homologous to the protein kinase gene family of eukaryotes and retroviruses. Nucleic Acids Res. 14, 1765-1777.

McGeoch, D.J., Dolan, A., Donald, S. and Brauer, D.H.K. (1986a). Complete DNA sequence of the short repeat region in the genome of herpes simplex virus type 1. Nucleic Acids Res. 14, 1727-1745.

McGeoch, D.J., Dolan, A., Donald, S. and Rixon, F.J. (1985). Sequence determination and genetic content of the short unique region in the genome of herpes simplex virus type 1. J. Mol. Biol. 181, 1-13.

McGeoch, D.J., Dolan, A. and Frame, M.C. (1986b). DNA sequence of the region of the genome of herpes simplex virus type 1 containing the exonuclease gene and neighbouring genes. Nucleic Acids Res. 14, 3435-3447.

McKnight, S.L. (1980). The nucleotide sequence and transcript map of the herpes simplex virus thymidine kinase gene. Nucleic Acids Res. 8, 5949-5964.

McKnight, S.L. and Gavis, E.R. (1980). Expression of the herpes thymidine kinase gene in Xenopus laevis oocytes: an assay for the study of deletion

mutants constructed *in vitro*. Nucleic Acids Res. *8*, 5931-5948.

McLauchlan, J. and Clements, J.B. (1983). DNA sequence homology between two colinear loci on the HSV genome which have different transforming abilities. EMBO J. *2*, 1953-1961.

McLauchlan, J., Gaffney, D., Whitton, J.L. and Clements, J.B. (1985). The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. Nucleic Acids Res. *13*, 1347-1368.

Miller, J., McLachlan, A.D. and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. EMBO J. *4*, 1609-1614.

Mocarski, E.S. and Roizman, B. (1982). Structure and role of the herpes simplex virus DNA termini in inversion, circularisation and generation of virion DNA. Cell *31*, 89-97.

Mocarski, E.S., Post, L.E. and Roizman, B. (1980). Molecular engineering of the herpes simplex virus genome: insertion of a second L-S junction into the genome causes additional genome inversions. Cell *22*, 243-255.

Morrison, J.M. and Keir, H.M. (1968). A new DNA-exonuclease in cells infected with herpesvirus: partial purification and properties of the enzyme. J. Gen. Virol. *3*, 337-347.

Moss, B., Gershowitz, A., Stringer Jr., Holland, L.E.

and Wagner, E.K. (1977). 5' terminal and internal methylated nucleotides in herpes simplex virus type 1 mRNA. J. Virol. 23, 234-239.

Moss, H. (1986) The herpes simplex virus type 2 alkaline DNase activity is essential for replication and growth. J. Gen. Virol. 67, in press.

Mount, S.M. (1982). A catalogue of splice junction sequences. Nucleic Acids Res. 10, 459-472.

Muller, M.T., Bolles, C.S. and Parris, D.S. (1985). Association of type 1 DNA topoisomerase with herpes simplex virus. J. Gen. Virol. 66, 1565-1574.

Murchie, M.-J. and McGeoch, D.J. (1982). DNA sequence analysis of an immediate-early gene region of the herpes simplex virus type 1 genome (map coordinates 0.950 to 0.978). J. gen. Virol. 62, 1-15.

O'Hare, P. and Hayward, G.S (1985a). Evidence for a direct role for both the 175,000- and 110,000-molecular weight immediate-early proteins of herpes simplex virus in the trans activation of delayed-early promoters. J. Virol. 53, 751-760.

O'Hare, P. and Hayward, G.S (1985b). Three trans-acting regulatory proteins of herpes simplex virus modulate immediate-early gene expression in a pathway involving positive and negative feedback regulation. J. Virol. 56, 723-733.

Perry, L.J., Rixon, F.J., Everett, R.D., Frame, M.C.

and McGeoch, D.J. (1986). The IE110 gene of herpes simplex virus type 1: characterisation by mRNA mapping, DNA sequence, oligopeptide antiserum and mutational analysis. Submitted to J. Gen. Virol., May 1986.

Poffenberger, K.L. and Roizman, B. (1985). A noninverting genome of a viable herpes simplex virus 1: Presence of head-to-tail linkages in packaged genomes and requirements for circularization after infection. J. Virol. 53, 587-595.

Poffenberger, K.L., Tabares, E. and Roizman, B. (1983). Characterisation of a viable, non-inverting herpes simplex virus 1 genome derived by insertion of sequences at the L-S component junction. Proc. Natl. Acad. Sci., U.S.A. 80, 2690-2694.

Ponticelli, A.S., Schultz, D.W., Taylor, A.F. and Smith, G.R. (1985). Chi-dependent DNA strand cleavage by RecBC enzyme. Cell 41, 145-151.

Post, L.E., Conley, A.J., Mocarski, E.S. and Roizman, B. (1980). Cloning of reiterated and nonreiterated herpes simplex virus 1 sequences as BamHI fragments. Proc. Natl. Acad. Sci. 77, 4201-4205.

Post, L.E., Mackem, S. and Roizman, B. (1981). Regulation of α genes of herpes simplex virus: expression of chimeric genes produced by fusion of thymidine kinase with α gene promoters. Cell 24, 555-565

Post, L.E. and Roizman, B. (1981). A generalised technique for deletion of specific genes in large

genomes: α gene 22 of herpes simplex virus 1 is not essential for growth. Cell 25, 227-232.

Powell, K.L., Buchan, A., Sim, C. and Watson, D.H. (1974). Type specific protein in herpes simplex virus envelopes reacts with neutralising antibody. Nature 249, 360-361.

Powell, K.L., Littler, E. and Purifoy, D.J.M. (1981). Non-structural proteins of herpes simplex virus. II. Major virus-specific DNA binding proteins. J. Virol 39, 894-902.

Preston, C.M. (1979). Control of herpes simplex type 1 mRNA synthesis in cells infected with wild-type virus or the temperature-sensitive mutant tsK. J. Virol. 29, 275-284.

Preston, C.M., Cordingley, M.G. and Stow, N.D. (1984). Analysis of DNA sequences which regulate the transcription of a herpes simplex virus immediate early gene. Virology 50, 708-716.

Proudfoot, N.J. and Brownlee, G.G. (1976). 3' non-coding region sequences in eukaryotic mRNA. Nature 263, 211-214.

Pustell, J. and Kafatos, F.C. (1982). A high speed, high capacity homology matrix: zooming through SV40 and polyoma. Nucleic Acids Res. 10, 4765-4782.

Quinlan, M.P. and Knipe, D.M. (1985). Stimulation of expression of a herpes simplex virus DNA-binding protein by two viral functions. Mol. Cell. Biol. 5, 957-963.

Quinn, J.P. and McGeoch, D.J. (1985). DNA sequence of the region in the genome of herpes simplex virus type 1 containing the genes for DNA polymerase and the major DNA binding protein. Nucleic Acids Res. 13, 8143-8163.

Rand, T.H. and Ben-Porat, T. (1980). Distribution of sequences homologous to the DNA of herpes simplex virus types 1 and 2, in the genome of pseudorabies virus. Intervirology 13, 48-53.

Rawls, W.E. (1973). Herpes simplex virus. In "The herpesviruses" pp291-323. A.S. Kaplan (Ed.). Academic Press, New York.

Rixon, F.J., Campbell, M.E. and Clements, J.B. (1982). The immediate-early mRNA that encodes the regulatory polypeptide $V_{mw}175$ of herpes simplex virus type 1 is unspliced. EMBO J. 1, 1273-1277.

Rixon, F.J., Campbell, M.E. and Clements, J.B. (1984). A tandemly reiterated DNA sequence in the long repeat region of herpes simplex virus type 1 found in close proximity to immediate-early mRNA 1. J. Virol. 52, 715-718.

Rixon, F.J. and Clements, J.B. (1982). Detailed structural analysis of two spliced immediate early mRNA. Nucleic Acids Res. 10, 2241-2256.

Rixon, F.J. and McGeoch, D.J. (1984). A 3' co-terminal family of mRNAs from the herpes simplex virus type 1 short region: Two overlapping reading frames encode unrelated polypeptides one of which has a highly reiterated amino acid sequence. Nucleic

Acids Res. 12, 2473-2487.

Rixon, F.J. and McGeoch, D.J. (1985). Detailed analysis of the mRNAs mapping in the short unique region of herpes simplex virus type 1. Nucleic Acids Res. 13, 953-973.

Rock, D.L. and Frazer, N.W. (1983). Detection of HSV-1 genome in central nervous system of latently infected mice. Nature 302, 523-525.

Roizman, B. (1979). The structure and isomerization of herpes simplex virus genomes. Cell 16, 481-494.

Roizman, B. and Furlong, D. (1974). The replication of herpesviruses. In "Comprehensive Virology" vol. 3, pp229-403. Ed. H. Fraenkel-Conrat and R.R. Wagner. Plenum Press, N.Y.

Roizman, b., Jacob, R.J., Knipe, D., Morse, C.S. and Ruyechan, W.T. (1979). On the structure, functional equivalence and replication of the four arrangements of herpes simplex virus DNA. Cold Spring Harbor Symp. Quant. Biol. U.S.A. 43, 809-826.

Rosen, A. and Darai, G. (1985). Mapping of the deletion in the genome of HSV-1 strain HFEM responsible for its avirulent phenotype. Med. Microbiol. Immunol. 173, 329-343.

Rosen, A., Gelderblom, H. and Darai, G. (1985). Transduction of virulence in herpes simplex virus type 1 from a pathogenic to an apathogenic strain by a cloned viral DNA fragment. Med. Microbiol. Immunol. 173, 257-278.

Sacks, W.R., Greene, C.C., Aschman, D.P. and Schaffer, P.A. (1985). Herpes simplex virus type 1 ICP27 is an essential regulatory protein. J. Virol. 55, 796-805.

Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. (1980). Cloning in single stranded bacteriophage as an aid to rapid DNA sequencing. J. Mol. Biol. 143, 161-178.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977). DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci., U.S.A. 74, 5463-5467.

Sears, A.E., Halliburton, I.W., Meignier, B., Silver, S. and Roizman, B. (1985). Herpes simplex virus 1 mutant deleted in the α22 gene: Growth and gene expression in permissive and restrictive cells and establishment of latency in mice. J. Virol. 55, 338-346.

Sheldrick, P. and Berthelot, N. (1974). Inverted repetitions in the chromosome of herpes simplex virus. Cold Spring Harbor Symp. Quant. Biol. U.S.A. 39, 667-678.

Silver, s. and Roizman,B. (1985). γ_2-thymidine kinase chimeras are identically transcribed but regulated as γ_2 genes in herpes simpex virus genome s and as β genes in cell genomes. Mol. Cell Biol. 5, 518-528.

Smith, G.P. (1976). Evolution of repeated DNA sequences by unequal crossover. Science 191, 528-535.

Spear, P.G. (1985). Glycoproteins specified by herpes simplex viruses. In "The Herpesviruses", Vol. 3, pp315-356. B. Roizman (Ed.). Plenum Press

Spear, P.G. and Roizman, B. (1980). Herpes simplex viruses. In "Molecular Biology of Tumour viruses", part 2. DNA tumour viruses. pp615-745. J. Tooze (Ed.). Cold Spring Harbor Monograph Series.

Staden, R. (1977). Sequence handling by computer. Nucleic Acids Res. 4, 4037-4051.

Staden, R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. Nucleic Acids Res. 8, 3673-3694.

Staden, R. (1982). Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res. 10, 4731-4751.

Staden, R. and McLachlan, A.D. (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucleic Acids Res. 10, 141-156.

Stenberg, R.M., Thomsen, D.R. and Stinski, M.F. (1984). Structural analysis of the major immediate early gene of human cytomegalovirus. J. Virol. 49, 190-199.

Stow, N.D. (1982). Localization of an origin of DNA replication within the TR_S/IR_S repeated region of the herpes simplex virus type 1 genome. EMBO J. 1,

863-867.

Stow, N.D. (1985). Mutagenesis of a herpes simplex virus origin of DNA replication and its effect on viral interference. J. gen. Virol. 66, 31-42.

Stow, N.D., McMonagle, E.C. and Davison, A.J. (1983). Fragments from both termini of the herpes simplex virus type 1 genome contain signals required for the encapsidation of viral DNA. Nucleic Acids Res. 11, 8205-8220.

Stow, N.D., Murray, M.D. and Stow, E.C. (1986). Cis-acting signals involved in the replication and packaging of herpes simplex virus type 1 DNA. Cold Spring Harbor Symp. Quant. Biol. In Press.

Stow, N.S., Subak-Sharpe, J.H. and Wilkie, N.M. (1978)Physical mapping of herpes simplex virus type 1 mutations by marker rescue. J. Virol. 28, 182-192.

Taylor, P. (1984). A fast homology program for aligning biological sequences. Nucleic Acids Res. 12, 447-455.

Taylor, P. (1986). A computer program for translating DNA sequences into protein. Nucleic Acids Res. 14, 437-441.

Timbury, M.C. and Subak-Sharpe, J.H. (1973). Genetic interaction between temperature sensitive mutants of types 1 and 2 herpes simpex virus. J. Gen. Virol. 18, 347-357.

Tullo, A.B., Shimeld, C., Easty, D.L. and Darville,

J.M. (1983). Distribution of latent herpes simplex virus infection in human trigeminal ganglion. Lancet 1, 353.

Twigg, A.J. and Sherratt, D. (1980). Trans-complementable copy-number mutants of plasmid ColE1. Nature 283, 216-218.

Varmuza, S.L. and Smiley, J.R. (1985). Signals for site-specific cleavage of HSV DNA: Maturation involves two separate cleavage events at sites distal to the recognition sequences. Cell 41, 793-802.

Wadsworth, J., Hayward, G.S. and Roizman, B. (1976). Anatomy of herpes simplex virus DNA. V. Terminally repetitive sequences. J. Virol. 17, 503-512.

Wagner, E.K. (1985). Individual HSV transcripts: Characterisation of specific genes., in "The Herpesviruses", vol.3, pp45-104., B. Roizman (Ed.), Plenum Press, New York and London.

Wagner, E.K., Swanstrom, R.I. and Stafford, M.G. (1972). Transcription of the herpes simplex virus genome in human cells. J. Virol. 10, 675-682.

Wagner, M.J. and Summers, W.C. (1978). Structure of the joint region and the termini of the DNA of herpes simplex virus. J. Virol. 27, 347-387.

Watson, R.J. and Clements, J.B. (1980). Identification of a herpes simplex virus type 1 function continuously required for synthesis of early and late virus RNAs. Nature 285, 329-330.

Watson, R.J., Preston, C.M. and Clements, J.B. (1979). Separation and characterization of herpes simplex virus type 1 immediate-early mRNAs. J. Virol. 31, 42-52.

Watson, R.J., Sullivan, M. and Vande Woude, G.F. (1981). Structures of two herpes simpex virus type 1 immediate-early mRNAs which map at the junctions of the unique and reiterated regions of the virus DNA S component. J Virol. 37, 431-444.

Weller, S.K., Spadaro, A., Schaffer, J.E., Murray, A.W., Maxam, A.M. and Schaffer, P.A. (1985). Cloning, sequencing, and functional analysis of ori_L, a herpes simplex virus type 1 origin of DNA synthesis. Mol. Cell. Biol. 5, 930-942.

Whitley, R.J. (1985). Epidemiology of herpes simplex viruses., in "The Herpesviruses", vol.3, ppl-44., B. Roizman (Ed.), Plenum Press.

Whitton, J.L. and Clements, J.B. (1984). Replication origins and a sequence involved in coordinate induction of the immediate-early gene family are conserved in an intergenic region of herpes simplex virus. Nucleic Acids Res. 12, 2061-2079.

Whitton, J.L., Rixon, F.J., Easton, A.J. and Clements, J.B. (1983). Immediate-early mRNA2 of herpes simplex viruses types 1 and 2 is unspliced: conserved sequences around the 5' and 3' termini correspond to transcription regulatory signals. Nucleic Acids Res. 11, 6271-6287.

Wickens, M. and Stephenson, P. (1984). Role of the conserved AAUAAA sequence: four AAUAAA point mutations prevent messenger RNA 3' end formation. Science 226,

1045-1051.

Wildy, P., Field, H.J. and Nash, A.A. (1982). Classical herpes latency revisited. In "Virus persistence" ppl33-167. B.W.J. Mahy, A.C. Minson and G.K. Darby (Eds.). Cambridge University Press.

Wilkie, N.M. (1973) The synthesis and substructure of herpesvirus DNA: the distribution of alkali labile single strand interruptions in HSV-1 DNA. J. Gen. Virol. 21, 453-467.

Woese, C.R. (1967). The genetic code. The molecular basis for genetic expression. Harper and Row.

Wohlrab, F., Garrett, B.K. and Franke, B. (1982). Control of expression of herpes simplex virus-induced deoxypyrimidine triphosphatase in cells infected with mutants of herpes simplex virus types 1 and 2 and intertypic recombinants. J. Virol. 43,

The IE110 gene of herpes simplex virus type 1: Characterization by mRNA mapping, DNA sequence, oligopeptide antiserum and mutational analysis.

Lise J. Perry, Frazer J. Rixon[+], Roger D. Everett[+], Margaret C. Frame[+] and Duncan J. McGeoch[*+]

Institute of Virology,
University of Glasgow,
Church Street,
Glasgow, G11 5JR,
U.K.

[*] to whom reprint requests should be sent

[+] members of the MRC Virology Unit

Short title: Structure of the HSV-1 IE110 gene

Keywords: DNA sequencing/ herpes simplex virus/ mRNA mapping/ trans activating factor/ oligopeptide antiserum

Journal of General Virology, in press

ABSTRACT

We have determined the DNA sequence of the herpes simplex virus type 1 gene encoding the immediate early protein IE110, which is involved in transcriptional activation of later virus genes. The locations of the 5' and 3' termini of IE110 mRNA, together with the positions of two introns, were identified. Examination of the DNA sequence suggested that translation starts at the first ATG after the 5' terminus of the mRNA, and that both introns occur in protein coding sequence. The predicted IE110 polypeptide contains 775 amino acids, and has a molecular weight of 78452. It contains a cysteine rich region resembling regions found in several proteins which interact functionally with DNA. An antiserum was raised to the predicted C terminal amino acid sequence of the IE110 polypeptide and shown to immunoprecipitate the native protein from HSV-1 infected cell extracts. Evaluations of functional importance of regions of the protein were made by construction of frameshift and deletion mutants of a plasmid borne IE110 gene. The mutants were tested for IE110 function by short-term transfection assays, and the results were correlated with the DNA sequence and RNA mapping studies.

INTRODUCTION

After infection of tissue culture cells with herpes simplex virus type 1 (HSV-1), the first virus genes to be expressed are the five immediate early (IE) genes (Watson et al., 1979). It has long been recognized that the product of IE gene 3, IE175 (also known as ICP4), has a major role in activation of transcription of early and late genes (Preston, 1979; Watson & Clements, 1980; Dixon and Schaffer, 1980). The function of IE175 was demonstrated by the study of HSV-1 mutants ts in IE gene 3, but until recently lack of similar mutants in the other IE genes has hindered studies on their function. In the last two years experimental schemes have been developed for assay of IE gene function using plasmid cloned IE genes (Everett, 1983; Everett & Dunlop, 1984; Gelman & Silverstein, 1985; O'Hare & Hayward, 1985). These can be introduced into culture cells to give expression of IE proteins, whose effect on transcription from HSV-1 promoters, introduced in a separate plasmid, can then be measured. Experiments of this type have shown that another IE protein, IE110 (also called ICP 0), encoded by IE gene 1, can act as a transcriptional activator, either independently or in cooperation with IE175 (Everett, 1984; O'Hare & Hayward, 1985).

For both IE175 and IE110 the mechanisms of trans activation remain obscure. It is not clear whether either protein acts directly (that is, by binding to DNA in the locality of the target transcription initiation site) or indirectly (for instance, by interaction with some other component, which ultimately affects transcription). Nor are relations between the actions of the two proteins clear. For instance, they could act at the same point of the activation process, or sequentially at different levels of the one process, or in parallel in different activation processes.

However, the existence of at least two HSV specified trans activators of virus gene expression certainly supports the view that control of HSV transcription may be far from simple.

Information on the IE gene 1 and its encoded protein is limited. The gene is located entirely within the long repeat element (R_L: See Figure 1) of the HSV-1 genome, and is thus diploid (Preston et al., 1978). However, the existence of deletion variants of HSV-1 shows that only one copy suffices for virus replication (Davison et al., 1981). IE110 is a phosphoprotein which accumulates in the nuclei of infected cells (Pereira et al., 1977). The protein has an estimated molecular weight of 110000 (Marsden et al., 1976) and has been shown to bind DNA (Hay & Hay, 1980).

In this paper we report the DNA sequence of IE gene 1 and describe mRNA mapping experiments which show that the gene contains two introns, both of which are considered to be in protein coding DNA. We give the deduced amino acid sequence of IE110, and show that an antipeptide antiserum to the predicted C terminus of IE110 binds to the whole protein. Finally, we have constructed deletion and frameshift mutations within the gene and have evaluated the functionality of the resulting mutated IE110 derivatives.

MATERIALS AND METHODS

Plasmids

Recombinant plasmids carrying HSV-1 strain 17 restriction fragments were used as follows. DNA sequence analysis used BamHI [illegible]oned into the BamHI site of pAT153 (this plasmid was called [illegible]HI k, also in pAT153 (from A.J. Davison), and XhoI c in

the XhoI site of pMK16 (pGX58, from N.D. Stow). In addition to pGX48 and pGX58, mRNA mapping experiments also used pGX53, which consists of the SalI-BamHI subfragment from the right end of the BamHI b fragment cloned into pAT153, and plasmid pJR3 consisting of the SstI-PstI fragment from IR_L (including the complete IE110 gene) cloned into pUC9, obtained from J. Russell and C.M. Preston. Functional assays of the IE110 gene used pJR3 described above, and p110 consisting of the HindIII-SstI fragment of pJR3 (containing all the HSV sequences of pJR3) cloned into the HindIII-SstI sites of pBRSst (where pBRSst is a derivative of pBR322 in which the PvuII site has been converted to an SstI site). pJR3 and p110 are similar except that p110 contains only a single BamHI site, while pJR3 has an additional BamHI site in the vector sequences. p175, used in later experiments, was derived from pGX58 and contained the coding sequences of the IE175 gene under the control of the SV40 early promoter and enhancer, cloned into the EcoRI-PvuII fragment of pBR322, so that upstream parts of IE genes 1 and 4 present in pGX58 were excluded. Finally, plasmid pgDCAT has the promoter the glycoprotein gD gene of HSV-1 linked to the chloramphenicol acetyltransferase (CAT) gene (McLauchlan et al., 1985). The CAT coding region is linked to an HSV-2 polyadenylation signal and the construct cloned into pUC9. Additional derivatives of pJR3 and p110 were constructed as described below.

DNA sequence analysis

The DNA sequence was determined by M13 chain termination reactions (Sanger et al., 1980), using M13 clones generated by ligation of sonicated DNA into the SmaI site of M13mp8 (Messing & Vieira, 1982) . The products of the sequencing reactions, labelled with [α-^{32}P]dATP, were electrophoresed through 9M urea,

6% polyacrylamide, buffer gradient gels (Biggin et al., 1983). Where necessary, to resolve compressions arising from the high G+C composition of the DNA, 6% polyacrylamide gels were maintained at a high temperature during the run using a hot water jacket (80-85°C).

Computing

Computing was performed with a DEC PDP 11/44 under RSX11M. Sequence analysis used the database system of Staden (1982). Codon usage evaluations used the program of Staden & McLachlan (1982). Sequence homologies were evaluated with a matrix comparison program (Pustell & Kafatos, 1982) and an alignment optimizing program (Taylor, 1984).

mRNA mapping

Cytoplasmic RNA was prepared by the method of Kumar and Lindberg (1972). For production of HSV-1 IE mRNA, cell monolayers were infected at a multiplicity of infection of 50 p.f.u./cell and the cell monolayers were pretreated and maintained in medium containing cycloheximide (Clements et al., 1977).

Structural analysis was performed as described previously (Rixon & Clements 1982). Briefly, either 3' or 5' labelled DNA was co-precipitated with cytoplasmic RNA from infected or mock-infected cells. The DNA/RNA mixture was denatured at 100°C for 3 min and incubated for 16 h at 57.5°C in 20 ul of 90% (v/v) deionized formamide, 0.4 M NaCl, 40 mM PIPES, pH 6.8, 1 mM EDTA. Nuclease S1 and exonuclease VII digestion were performed at 37°C.

Production of anti-oligopeptide serum

The peptide NH_2-Tyr-Glu-Gly-Ala-Ser-Thr-Arg-Asp-Glu-Gly- (synthesized by Cambridge Research Biochemicals,

Cambridge, England) corresponds to the C terminal eleven amino acids of the predicted IE110 polypeptide linked to a Tyr residue. After coupling the peptide to bovine serum albumin (BSA), antisera were raised in rabbits as previously reported (Frame et al., 1986). Antibodies raised against BSA were removed by passing the immune-serum through a BSA-Sepharose column.

Immunoprecipitation

Confluent monolayers of BHK cells were infected with 20 p.f.u./cell of HSV-1 strain 17, and the proteins labelled with 100uCi/ml [^{35}S]-methionine. Conditions for increased production of immediate-early polypeptides using cycloheximide block have previously been described (Preston, 1979). Immunoprecipitations were carried out as described (Frame et al., 1986) and analysed on 5-12·5% gradient SDS-polyacrylamide gels (Marsden et al., 1976).

Construction of plasmids carrying mutated versions of IE gene 1

Derivatives of the plasmid pJR3 were constructed as follows. p110del4 was made by SalI and XhoI digestion of pJR3 followed by ligation. p110del6 and del8 were made by digestion of pJR3 with KpnI and XhoI respectively, followed by S1 nuclease treatment and ligation. p110del10 and del11 were made by digestion of pJR3 with SalI and p110 with BamHI followed by treatment with DNA polymerase I large fragment in the presence of all four dNTPs (filling in) and ligation. p110del7 was made by digestion of pJR3 with HindIII and HpaI, followed by filling in and ligation. p110del1 was isolated from the ligation that produced p110del8. p110del14 was derived from a partial SmaI digest of p110del7. p110E11F was isolated after (i) elimination of the EcoRI site in p110 to give p111; (ii) insertion of a 12 bp oligonucleotide, containing an EcoRI recognition sequence, into partially HaeIII cut p111; (iii)

screening of the resulting plasmids to isolate pl10El1, containing an EcoRI linker in the HaeIII site at position 3440; and (iv) elimination of the EcoRI site by filling in with DNA polymerase I large fragment, and ligation. Structures of these plasmids are shown in Figure 7. The sequences around manipulated sites in del1, del6, del8, del10, del11, and in pl10El1 and pl10El1F, were confirmed by DNA sequence analysis (Maxam & Gilbert, 1980), and all plasmids were subjected to extensive restriction enzyme analysis.

Transfection and CAT assays

HeLa cell monolayers were transfected with a total of 12 ug of plasmid DNA (4 ug per plasmid) by calcium phosphate precipitation (Corsalo & Pearson, 1981). pBR322 was used to make up the amount of DNA to 12 ug where necessary. After 24 h the monolayers were washed and fresh medium added. After a further 24 h, extracts were prepared and CAT activities assayed as described by Gorman et al. (1982). In all experiments, a positive control (pgDCAT + pJR3 + pGX58) was included. The radioactivity in the substrate and in monoacetylated product spots was measured. The protein concentration of each extract was determined and the percentage conversion per mg protein calculated. This value was then expressed as a percentage of that of the positive control.

RESULTS

Mapping of IE gene 1 mRNA

As shown in Figure 1, IE gene 1 is located in the R_L element HSV-1 genome. We have determined the complete sequence of ith adjacent regions of U_L: the whole sequence will

be presented elsewhere (Perry & McGeoch, in preparation). Residue numbering in this paper is based on the DNA sequence of R_L, starting with the residue in IR_L adjacent to the a' sequence (see legend to Figure 1). Previous work had mapped the 5' terminus of mRNA-1 to residue 1714 (Mackem & Roizman, 1982) and the 3' terminus to residue 5301 (Rixon et al., 1984) on the leftward 5' to 3' strand of IR_L, giving a span of 3587 residues. Since the size of the mRNA, including the poly(A) tail, was estimated from agarose gel electrophoresis to be 3 kb (Watson et al., 1979) , it seemed probable that the mRNA was spliced.

Figure 2 gives a summary of the nuclease digestion mapping data described in the text and shows the structure of IE mRNA-1 as determined by these analyses. In the following descriptions, single positions are given for the mRNA 5' and 3' ends and for all splice sites. Where detailed analysis of small nuclease-resistant products revealed a number of closely spaced bands (probably due to imprecise nuclease cleavage at the hybrid ends), the size given is that which aligns the RNA with the appropriate splice junction recognition site on the DNA. As described below, the DNA sequence encoding IE mRNA-1 has a high G+C composition. Furthermore, the distribution of G+C is not uniform throughout the gene, with certain regions having a particularly extreme G+C content. This causes difficulties in the analysis of RNA structure by nuclease digestion, since localised melting of double-stranded molecules in regions of lower G+C will occur and complex annealing of DNA and RNA involving more than two molecules may take place. In several of the examples described below more than one band was present following nuclease digestion. In each of these cases only one of the bands coincided with a recognizable RNA processing signal. The other bands generally correspond to short A+T rich stretches in an otherwise high G+C sequence. These bands were assumed to

result from localised melting of DNA/DNA or DNA/RNA hybrids. However, it is possible that some of these bands may represent minor or aberrant splicing events.

Initial attempts to determine the structure of IE mRNA-1 by nuclease digestion procedures were made using DNA fragments which had been labelled at the BamHI site at position 2508. No nuclease-resistant material was detected with either 5' labelled or 3' labelled DNA. In view of the discrepancy between the size of the mRNA and of the DNA sequence encoding it, mentioned above, it seemed probable that the BamHI site lay within an intron.

The 5' end of IE mRNA-1 had previously been identified by Mackem and Roizman (1982). This was confirmed using pJR3 DNA which had been 5' labelled at the unique NcoI site at position 1866. This generated nuclease-resistant bands of around 151 bases length, placing the 5' end of IE mRNA-1 at position 1716. The 5' boundary of the putative intron was identified using a 3' labelled, 491 bp HinfI sub-fragment of pJR3, (positions 1767-2256). This generated nuclease S1-resistant and exonuclease VII-resistant bands of around 154 bases (Figure 3A) placing the splice donor site at position 1920. To locate the 3' boundary of this intron, the 435 bp XhoI/BamHI sub-fragment of pGX48 was 5' labelled at the XhoI site at position 2943. This generated two nuclease S1 resistant species, a major band of around 258 bases and a minor band of around 280 bases, and a single exonuclease VII-resistant band of around 258 bases (Figure 3B). Examination of the DNA sequence encoding this region revealed that only the 258 base band coincides with a recognizable splice acceptor sequence at position 2686. The 280 base band coincides with the start of a relatively A+T rich sequence which does not resemble a [illegible]ice acceptor site. The data presented here do not exclude the [illegible]that small exons may exist between the mapped splice

donor and acceptor sites. However, since the exonuclease VII and nuclease S1-resistant bands are of similar sizes, the introns must extend beyond the limits of the DNA fragments used for mapping (to within positions 2256-2508). Conclusive proof of the absence of more complicated splicing patterns will probably require cDNA cloning and sequence determination.

To determine whether the remaining 3' portion of IE mRNA-1 gene was unspliced, the 2561 bp XhoI/HpaI sub-fragment of pGX48 (positions 2943-5503) was 3' labelled at the XhoI site (position 2943). This generated three nuclease S1-resistant bands of 409, 427 and 470 bases (Figure 3C). Examination of the DNA sequence of this region revealed that the 409 base band coincides with a potential splice donor sequence at position 3352. The two other nuclease S1-resistant products coincide with short AT rich regions which do not resemble any known splice donor signals. The position of this splice site was confirmed using a 3' labelled, 1063 bp HinfI sub-fragment (positions 2800-3861) of pGX53. This generated three nuclease S1-resistant bands of 552, 570 and 613 bases (data not shown), which correspond to those observed with the XhoI/HpaI fragment. It is clear from these results that the gene encoding IE mRNA-1 contains a second intron. Exonuclease VII digestion generated only full length probe DNA, indicating that the intron lay entirely within this HinfI fragment. Therefore, the 3' boundary of this second intron was identified using the same 1063 bp HinfI sub-fragment of pGX53, which had been 5' labelled. This hybridization was performed at both 57·5°C and at 60°C and gave a complicated pattern of bands at both temperatures (Figure 3D). However, there were differences in the abundance of the bands at the two temperatures, with a considerable increase in the relative intensity of a 375 base band at the higher temperature suggesting that it represents a more stable hybrid.

This band coincides with a possible splice acceptor signal at position 3488, which we propose represents the intron boundary. A similar pattern of bands was obtained (data not shown) with a 5' labelled, 549 bp HincII/KpnI fragment (positions 3262-3811). In this case the increase in relative intensity at the higher temperature was exhibited by a 322 base band which corresponds to the 375 base band produced with the HinfI fragment. We emphasize that the difficulties with these analyses are considered to result from the extreme base composition, and that finally definitive results will probably require cDNA cloning and sequence analysis.

The remainder of the 3' portion of IE mRNA-1 was analysed using 3' labelled 1470 bp HinfI (positions 4376-5843) and 1685 bp Sau3AI (positions 3855-5540) sub-fragments of pJR3. These generated nuclease resistant bands of 925 bases (HinfI) and 1446 bases (Sau3AI) respectively (data not shown), placing the 3' end at the previously mapped location, position 5301 (Rixon et al., 1984).

From the above data the structure of the gene for IE mRNA-1 is shown to consist of three exons of 205, 667 and 1812 bases, separated by two introns of 767 and 136 bases.

Sequence of IE gene 1

Figure 4 presents the DNA sequence of residues 1201 to 5520 of IR_L. This was determined by the M13/chain terminator method, using plasmid cloned copies of restriction fragments BamHI b, BamHI k and XhoI c, as shown in Figure 1. The positions of the 5' and 3' termini of mRNA-1 are marked at residues 1716 and 5301 respectively. There is a potential TATA sequence at 1689-1693. The promoter and activator regions 5' to the mRNA's 5' terminus [illegible] been examined by Mackem & Roizman (1982). At the downstream [illegible]ene, adjacent to the 3' terminus of the mRNA, there

are two copies of the polyadenylation associated sequence AATAAA, at residues 5239 and 5279. Downstream of the 3' terminus lies a set of 9 tandem copies of a 16 bp sequence (Rixon et al., 1984). Also present downstream of the 3' terminus of mRNA-1 is the sequence TGTGTTGG, at 5312 to 5319, which resembles the consensus YGTGTTYY identified by McLauchlan et al. (1985) as being required for efficient formation of mRNA 3' termini.

There are consensus splice donor and acceptor sequences close to the mapped boundaries of the two introns (Mount, 1982). We thus consider that intron 1 runs from 1921 to 2685 and intron 2 from 3353 to 3488. The sequence of intron 1 contains 3 imperfect tandem copies of a 54 bp sequence and also a number of other imperfect repeat elements. Starting at residue 2647, in intron 1, Figure 4 shows a run of eight C residues, as found in the BamHI b clone analysed, while the XhoI c clone contains 9 residues here. The sequence of intron 2 is particularly purine rich.

We have located the polypeptide coding sequence for IE110 by examination of open reading frames and use of the codon usage evaluation program of Staden & McLachlan (1982). We conclude that translation begins with the first ATG after the mRNA's 5' terminus, at 1864, giving a 5' noncoding region of 148 residues. This ATG, in exon 1, conforms to the initiation codon consensus of Kozak (1984), and initiates an open reading frame (allowing for introns) of 775 codons, terminating with TAA at 5090. This leaves a 3' noncoding region of 209 residues. The polypeptide coding region of IE110 has a base composition of 75.4% G+C. This is a very high value, even for an HSV gene (see Discussion).

Amino acid sequence of IE110

The 775 amino acid polypeptide predicted by the above reading frame has a molecular weight of 78452. This is

considerably lower than the previous estimate of 110000 obtained from gel electrophoretic mobility (Marsden et al., 1976) . However, such revisions of molecular weights have occurred frequently in HSV sequence analysis (see, for example, McGeoch et al., 1985). The IE110 sequence contains a small excess of basic over acidic residues (Table 2). The five most common amino acid types are Ala, Pro, Gly, Ser and Arg. Ala, Pro, Gly and Arg are the amino acids which possess codons containing only G and C residues (see Table 1), so that their prominence correlates with the extreme base composition of the gene. The distribution of these residues is not uniform: they are most abundant in a region beginning with the start of exon 3 (amino acid residue 242) and ending with residue 560. In addition, amino acid residues 554 to 594 consist mostly of Ser and Ala residues.

Recently, it has been found that many DNA binding proteins contain a Cys rich locality, which is thought to be involved in coordination of metal ions (Miller et al., 1985; Berg, 1986). The most characteristic sequence element within these regions comprises two Cys residues separated by two other amino acids. We have now found that IE110 also belongs to this class: as shown in Figure 5, 9 of the 14 Cys residues in IE110 are located in a 58 residue region between amino acids 99 and 156, and these include three C--C pairs.

We investigated the possibility that the IE110 protein might be related to the IE175 protein, which is also involved in trans activation of later virus genes, but could not detect any homology, global or local, between the IE110 and IE175 amino acid sequences, beyond noting that both contain a very Ser rich region (amino acids 554 to 594 of IE110, and 177 to 199 of IE175; McGeoch [illegible], 1986). Conceivably, these residues could be involved in [illegible]lation of the proteins. In addition, no homology of

IE110 sequences was detected with other HSV IE proteins (sequences from Murchie & McGeoch, 1982; McGeoch et al., 1985; Perry & McGeoch, unpublished data), or with the major IE protein of the betaherpesvirus, human cytomegalovirus, (Stenberg et al., 1984) or with any polypeptides predicted from the complete genome sequence of the gammaherpesvirus, Epstein-Barr virus (Baer et al., 1984). A probable homologue to IE gene 1 was detected in the genome of the related alphaherpesvirus, varicella-zoster virus (VZV). This is gene 61 of VZV, which is proposed to lack introns and to encode a 467 amino acid polypeptide (Davison & Scott, 1986). Similarity between the two amino acid sequences is extremely limited (so that the correspondence was not detected in tests carried out by Davison & Scott (1986)) and is essentially limited to the Cys rich region, where all three C--C pairs are conserved (Figure 5). We consider that this limited similarity, taken together with the corresponding genomic locations of the two genes, constitutes a reasonable case for them being related and having similar functions. Confirmation of this assignment, however, must await experimental demonstration of transcriptional activation by the VZV gene 61 product.

Detection of IE110 using anti-oligopeptide serum

An antiserum raised to the C terminal eleven amino acids of the predicted sequence of IE110 immunoprecipitates the native protein from extracts of cells infected with HSV-1 and labelled with [^{35}S]-methionine under immediate-early conditions (Figure 6, lane 2). Specificity was shown by inhibition of the immunoprecipitation by inclusion of an excess of the synthetic oligopeptide in the reaction mix (Figure 6, lane 6). This confirms our reading frame assignment at the C terminus. In addition, no reaction with extracts of mock-infected cells was

observed. Also specifically immunoprecipitated is a lower molecular weight species (apparent molecular weight of 40000) which may be a breakdown product of IE110 or a species associated with it in the infected cell.

Mutational analysis of the IE110 gene

Recent work from several laboratories has shown that IE110 can activate transcription from HSV promoters in transient transfection assays (Everett, 1984; O'Hare & Hayward, 1985; Quinlan & Knipe, 1985; Gelman & Silverstein, 1985). The promoter of the HSV-1 glycoprotein D gene was found to be slightly activated by co-transfection of plasmids encoding IE175, while inclusion of plasmids expressing both IE175 and IE110 resulted in very substantial activation (Everett, 1984). This observation provides the basis for an assay of the function of IE110 in this experimental system. The assay was used to determine if the effect of frameshift and deletion mutants within IE gene 1 could be correlated with the mRNA mapping and DNA sequence data presented above. In the experiments reported here, plasmids pJR3 and p110, which contain HSV-1 R_L sequences between the SstI site at 895 and the PstI site at 7308, were used as a source of functional IE110 gene, and a set of mutated derivatives was constructed. The ability of these mutant plasmids to activate gene expression was measured by co-transfection with a plasmid carrying a CAT gene linked to the gD promoter, and with pGX58 or p175, which express IE175. The efficiency of gD promoter activation was then assayed by measurement of CAT enzyme activity in extracts of transfected HeLa cells.

The IE110 gene mutations made are shown in Figure 7 and the [illegible] obtained are summarized in Table 3. Deletion of sequences [illegible] the HpaI site at 5501 (plasmid p110del7) resulted in

an increase in trans activation. While the reason for the increase is not resolved, the important conclusion for this plasmid is that the functionally complete IE110 gene lies upstream of residue 5501. On the other hand, deletion between the XhoI site at 2939 and the SalI site at 5065 (plasmid p110del4) eliminated IE110 activity. Thus, these trans activation data are consistent with the DNA sequence interpretation. Frameshift mutations, p110del6 and p110del8, both mapping in exon 2, gave very low levels of activation. We do not know whether the apparent, small effects with these plasmids are actually due to a residual activity of the N terminal fragments of IE110 expressed. Plasmid p110del10, containing a frameshift mutation at the SalI site at 5065, near the assigned translational termination site at 5095, gave an intermediate activation. We interpret this as indicating that the immediate C terminal residues are not crucial for IE110 activity. We also constructed two in-frame deletions within predicted coding sequence. Plasmid p110del1 has lost 354 bp (118 codons) from exon 2 (including almost all of the Cys rich region), and p110del14 has lost 483 bp (181 codons) from exon 3. Rather surprisingly, both retain low levels of trans activation activity. Evidently, the IE110 protein can still function to some extent despite substantial structural alterations in two different regions. Finally, frameshift mutations were constructed at the BamHI site, residue 2508 (p110del11), and at the HaeIII site at residue 3440 (p110E11F). These plasmids gave similar activation levels to those obtained with their parent. These results are consistent with the mRNA mapping data, which placed the BamHI site at position 2508 and the HaeIII site at position 3440 within introns 1 and 2, respectively.

DISCUSSION

This paper describes the DNA sequence, mRNA mapping and preliminary mutational analysis of IE gene 1, encoding IE110. In addition we have demonstrated a specific interaction of anti-C terminal oligopeptide serum with the native protein extracted from HSV-1 infected cells. The HSV genome contains five separate IE genes (Watson et al., 1979). Of these, IE genes 2 and 3 possess no introns (Whitton at al., 1983; Rixon et al., 1982), while genes 4 and 5 each contain a single intron in the 5' noncoding region (Watson et al., 1981; Rixon & Clements, 1982; Murchie & McGeoch, 1982). IE gene 1 is the only IE gene with multiple introns, and the only one containing introns in polypeptide coding sequence. Costa et al. (1985) have described a late gene of HSV-1 containing an intron, which lies within polypeptide coding sequences. (Fontichiaro et al. (1985) have claimed to detect a second intron in IE genes 4 and 5, but we consider that their interpretation is erroneous and results from a secondary structure artifact in high G+C DNA (Rixon & Clements, 1982)).

HSV-1 DNA has an estimated overall base composition of 67% G+C (Kieff et al., 1971) and most genes so far sequenced have coding regions around 65% G+C (see McGeoch et al., 1985). The coding region of IE gene 1 is strikingly higher, at 75.4% G+C. This is the second highest value in any determined protein coding sequence, so far as we are aware, and is exceeded only by that of another HSV-1 gene, the IE gene 3, which is situated in the R_S region of the genome and has a coding region composition of 81.5% (McGeoch et al., 1986). Inspection of the IE gene 1 codon usage catalogue (Table 1) shows that the base composition is achieved [illegible]ly by encoding high levels of amino acids with G+C rich [illegible]partly by a heavy bias towards G and C in the

redundant third positions. The phenomenon has been discussed for IE gene 3 (McGeoch et al., 1986), and the arguments adduced then appear also to apply in the present case. Thus, it is clear in general that the base composition is not driven to extremity by requirements of amino acid sequence functionality. The nature of evolutionary forces underlying the effect remains obscure, although they are evidently potent. For both IE gene 1 and gene 3, it is a major repeat element of the HSV genome that is involved.

In both genes there are, within the coding sequence, regions of particularly high G+C content. In the present study, the central portion of the IE110 coding sequence (residues 3489 to 4564) has a base composition of 80.8%, while the upstream and downstream portions (allowing for introns) are 71.2% and 70.4% respectively. It is interesting that the upstream boundary of this high G+C region is at or near the start of exon 3: this could suggest a possible distinction in evolutionary history or in function of encoded polypeptide between exons 1 plus 2, on the one hand, and exon 3. In the case of IE gene 3, McGeoch et al. (1986), suggested that particularly extreme base composition might correlate with sequences encoding relatively non-crucial regions of protein. The plasmid pl10del14 (which has lost 483 bp from the region of high G+C) retained some transactivational activity, which would support this suggestion; however plasmid pl10del1 contains a deletion outside of the high G+C region and also retains some transactivational activity.

There is at present no firm model for IE110's mode of action. The existence of a Cys rich locality in IE110, similar to that found in many proteins which interact directly with nucleic acids (Berg, 1986) does suggest a rather direct mechanism of transcriptional activation, involving binding of IE110 to DNA.

The finding that a residual activity remains after deletion of this region we regard as probably indicative of the complexity of functional interaction between IE110 and IE175, which is also present in our assay.

In conclusion, the mRNA mapping, DNA sequencing and sequence interpretation give a precise description of IE gene 1. Our studies on modification of the gene correlate well with the sequence analysis, and represent a start in molecular genetic analysis of IE110 function.

ACKNOWLEDGEMENTS

We wish to thank J.H. Subak-Sharpe for advice and encouragement during this work. Our thanks are also due to N.D. Stow, A.J. Davison, C.M. Preston and J. Russell for plasmids, and to A. Dolan, M. Wilkie and S. McCallum for technical assistance. L.J.P. was in receipt of a grant from The Shell Trust for Higher Education.

REFERENCES

BAER, R., BANKIER, A.T., BIGGIN, M.D., DEININGER, P.L., FARRELL, P.J., GIBSON, T.J., HATFULL, G., HUDSON, G.S., SATCHWELL, S.C., SEGUIN, C., TUFFNELL P.S. & BARRELL, B.G. (1984). DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310, 207-211.

BERG, J.M. (1986). Potential metal-binding domains in nucleic acid binding proteins. Science 232, 485-487.

BIGGIN, M.D., GIBSON, T.J. & HONG, G.F. (1983). Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proceedings of the National Academy of Sciences, U.S.A. 80, 3963-3965.

CLEMENTS, J.B., WATSON, R.J. & WILKIE, N.M. (1977). Temporal regulation of herpes simplex virus type 1 transcription; location of transcripts on the viral genome. Cell 12, 275-285.

CORSALO, C.M. & PEARSON, M.L. (1981). Enhancing the efficiency of DNA-mediated gene transfer in mammalian cells. Somatic Cell Genetics 7, 603-616.

COSTA, R.H., DRAPER, K.G., KELLY, T.J. & WAGNER, E.K. (1985). An unusual spliced herpes simplex virus type 1 transcript with sequence homology to Epstein-Barr virus DNA. Journal of Virology 54,317-328.

DAVISON, A.J. & SCOTT, J.E. (1986). The complete DNA sequence of varicella-zoster virus. Journal of General Virology (in press).

DAVISON, A.J. & WILKIE, N.M. (1981). Nucleotide sequences of the joint between the L and S segments of herpes simplex virus types 1 and 2. Journal of General Virology 55, 315-331.

DAVISON, A.J., MARSDEN, H.S. & WILKIE, N.M. (1981). Nucleotide

sequences of the joint between the L and S segments of herpes simplex virus types 1 and 2. Journal of General Virology 55, 315-331.

DIXON, R.A.F. & SCHAFFER, P.A. (1980). Fine structure mapping and functional analysis of temperature-sensitive mutants in the gene encoding the herpes simplex virus type 1 immediate early protein VP175. Journal of Virology 36, 189-203.

EVERETT, R.D. (1983). DNA sequence elements required for regulated expression of the HSV-1 glycoprotein D gene lie within 83 bp of the RNA capsites. Nucleic Acids Research 11, 6647-6666.

EVERETT, R.D. (1984). Transactivation of transcription by herpes virus products; requirements for two HSV-1 immediate early polypeptides for maximum activity. EMBO Journal 3, 3135-3141.

EVERETT, R.D. & DUNLOP, M. (1984). Trans activation of plasmid-borne promoters by adenovirus and several herpes group viruses. Nucleic Acids Research 12, 5969-5978.

FONTICHIARO, K.L.E., BECK, T.W. & MILLETTE, R.L. (1985). In vitro transcription of herpes simplex virus genes: Identification of a new initiation site and second intervening sequence in the immediate-early RNA-5 gene. Journal of Virology 53, 235-242.

FRAME, M.C., MCGEOCH, D.J., RIXON, F.J., ORR, A.C. & MARSDEN, H.S. (1986). The 10K virion phosphoprotein encoded by U_S gene 9 from herpes simplex virus type 1. Virology, in press.

GELMAN, I.H. & SILVERSTEIN, S. (1985). Identification of immediate early genes from herpes simplex virus that transactivate the virus thymidine kinase gene. Proceedings of the National Academy of Sciences, U.S.A. 82, 5265-5269.

GORMAN, C.M., MOFFATT, L.F. & HOWARD, B.H. (1982). Recombinant [illegible]omes which express chloramphenicol acetyl transferase in [illegible]ells. Molecular and Cellular Biology 2, 1044-1051.

HAY, R.T. & HAY, J. (1980). Properties of herpesvirus-induced "immediate early" polypeptides. Virology 104, 230-234.

KIEFF, E.D., BACHENHEIMER, S.L. & ROIZMAN, B. (1971). Size, composition, and structure of the deoxyribonucleic acid of herpes simplex virus subtypes 1 and 2. Journal of Virology 8, 125-132.

KOZAK, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Research 12, 857-872.

KUMAR, A. & LINDBERG, U. (1972). Characterization of messenger ribonucleoprotein and messenger RNA from KB cells. Proceedings of the National Academy of Sciences, U.S.A. 69, 681-685.

MACKEM, S. & ROIZMAN, B. (1982). Structural features of the herpes simplex virus α gene 4, 0 and 27 promoter-regulatory sequences which confer regulation on chimeric thymidine kinase genes. Journal of Virology 44, 939-949.

MARSDEN, H.S., CROMBIE, I.K. & SUBAK-SHARPE, J.H. (1976). Control of protein synthesis in herpesvirus-infected cells: analysis of the polypeptides induced by wild type and sixteen temperature sensitive mutants of HSV strain 17. Journal of General Virology 31, 347-372.

MAXAM, A.M. GILBERT, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. Methods in Enzymology 65, 499-560.

McGEOCH, D.J., DOLAN, A., DONALD, S. & BRAUER, H.K. (1986). Complete DNA sequence of the short repeat region in the genome of herpes simplex virus type 1. Nucleic Acids Research 14, 1727-1745.

McGEOCH, D.J., DOLAN, A. DONALD, S. & RIXON, F.J. (1985). Sequence determination and genetic content of the short unique region in the genome of herpes simplex virus type 1. J.

...iology 181, 1-13.

...CHLAN, J., GAFFNEY, D. WHITTON, J.L. & CLEMENTS, J.B. (1985). The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. Nucleic Acids Research 13, 1347-1368.

MESSING, J. & VIEIRA, J. (1982). A new pair of M13 vectors for selecting either strand of double-digest restriction fragments. Gene 19, 269-276.

MILLER, J., McLACHLAN, A.D. & KLUG, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. The EMBO Journal 4, 1609-1614.

MOUNT, S.M. (1982). A catalogue of splice junction sequences. Nucleic Acids Research 10, 459-472.

MURCHIE, M.-J. & McGEOCH, D.J. (1982). DNA sequence analysis of an immediate-early gene region of the herpes simplex virus type 1 genome (map coordinates 0.950 to 0.978). Journal of General Virology 62, 1-15.

O'HARE, P. & HAYWARD, G.S. (1985). Evidence for a direct role for both the 175,000- and 110,000-molecular-weight immediate-early proteins of herpes simplex virus in the transactivation of delayed-early promoters. Journal of Virology 53, 751-760.

PEREIRA, L., WOLFF, M.H., FENWICK, M. & ROIZMAN, B. (1977). Regulation of herpesvirus macromolecular synthesis. Virology 77, 733-749.

PRESTON, C.M. (1979). Control of herpes simplex virus type 1 mRNA synthesis in cells infected with wild-type virus or the temperature-sensitive mutant tsK. Journal of Virology 29, 275-284.

PRESTON, V.G., DAVISON, A.J., MARSDEN, H.S., TIMBURY, M.C., SUBAK-SHARPE, J.H. & WILKIE, N.M. (1978). Recombinants between herpes simplex virus types 1 and 2. Analysis of genome

structure and expression of immediate early polypeptides. Journal of Virology 28, 499-517.

PUSTELL, J. & KAFATOS, F.C. (1982). A high speed, high capacity homology matrix: zooming through SV40 and polyoma. Nucleic Acids Research 10, 4765-4782.

QUINLAN, M.P. & KNIPE, D.M. (1985). Stimulation of expression of a herpes simplex virus DNA binding protein by two viral proteins. Molecular & Cellular Biology 5, 957-963.

RIXON, F.J., CAMPBELL, M.E. & CLEMENTS, J.B. (1982). The immediate-early mRNA that encodes the regulatory polypeptide $V_{mw}175$ of herpes simplex virus type 1 is unspliced. EMBO Journal 1, 1273-1277.

RIXON, F.J., CAMPBELL, M.E. & CLEMENTS, J.B. (1984). A tandemly reiterated DNA sequence in the long repeat region of herpes simplex virus type 1 found in close proximity to immediate early mRNA-1. Journal of Virology 52, 715-718.

RIXON, F.J. & CLEMENTS, J.B. (1982). Detailed structural analysis of two spliced HSV-1 immediate-early mRNAs. Nucleic Acids Research 10, 2241-2256.

SANGER, F., COULSON, A.R., BARRELL, B.J., SMITH, A.J.H. & ROE, B.A. (1980). Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. Journal of Molecular Biology 143, 161-178.

STADEN, R. (1982). Automation of computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Research 10, 4731-4751.

STADEN, R. & MCLACHLAN, A.D. (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucleic Acids Research 10, 141-156.

STENBERG, R.M., THOMSEN, D.R. & STINSKI, M.F. (1984). Structural analysis of the major immediate early gene of human