The EMBO Journal Vol.2 No.12 pp.2203 – 2209, 1983

DNA sequence of the U_S component of the varicella-zoster virus genome

A.J.Davison

MRC Virology Unit, Institute of Virology, Church Street, Glasgow G11 5JR, UK

Communicated by J.H.Subak-Sharpe
Received on 29 July 1983; revised on 25 August 1983

The linear duplex DNA molecule of varicella-zoster virus is 120 000 bp in size and has the sequence arrangement U_L-IR_S-U_S-TR_S, where U_L and U_S are unique sequences and IR_S and TR_S are inverted repeats flanking U_S. The primary structure of the cloned *Sst*I g DNA fragment containing U_S (5232 bp) and adjacent portions of IR_S and TR_S (426 bp of each) was determined, and the following model for genetic expression was derived from an analysis of the sequence. The region specifies four mRNAs encoding primary translation products with mol. wts. of 11, 44, 39 and either 74 or 70 kd. The 39-and 70-kd proteins have primary structures characteristic of membrane proteins. The mRNAs encoding the 11- and 74/70-kd proteins extend from opposite sides of U_S into IR_S/TR_S, thus sharing a common 3′ terminus. These proteins do not share a common carboxy terminus because the coding region for the 11-kd protein terminates at the junction between U_S and IR_S, whereas that for the 74/70-kd protein extends into TR_S. The analysis affirms the hypothesis that the extent of inverted repeats in herpesvirus genomes is primarily a result of constraints imposed by adjacent protein coding sequences.

Key words: varicella-zoster virus/DNA sequence/inverted repeats

Introduction

Man is host to five herpesviruses: herpes simplex virus types 1 and 2, varicella-zoster virus (VZV), human cytomegalovirus and Epstein-Barr virus. VZV causes two common clinical conditions, chickenpox (varicella) and shingles (herpes zoster), but the molecular biology of this virus has lagged behind that of the other human herpesviruses and, indeed, behind that of some animal herpesviruses. This was probably owing to difficulties associated with growth of VZV *in vitro*. However, recent technical advances in this area have allowed several groups to begin to characterize the genetic material of VZV.

The linear double-stranded DNA molecule of VZV has a guanine plus cytosine (G + C) content of 46% (Ludwig *et al.*, 1972), and the genome structure has been elucidated by electron microscopy and restriction endonuclease analysis (Dumas *et al.*, 1980 and 1981; Straus *et al.*, 1981 and 1982; Ecker and Hyman, 1982; Gilden *et al.*, 1982). The VZV genome is represented structurally as U_L-IR_S-U_S-TR_S, where U_L (~100 000 bp) and U_S (~5000 bp) are unique sequences and IR_S and TR_S (each ~7500 bp) are inverted repeats flanking U_S. Virion DNA contains two isomeric genome arrangements in equimolar amounts owing to inversion of U_S (or IR_S-U_S-TR_S) relative to U_L, and these are defined arbitrarily as the P (prototype) and I_S arrangements. Davison and Scott (1983) have excluded the presence of a precise terminal redundancy greater in size than 20 bp. Straus *et al.* (1981) reported the presence of superhelical circular DNA in a proportion (5 – 10% or more) of VZV nucleocapsids from all of the virus strains they examined, but the significance of this observation awaits explanation. Physical maps of VZV DNA for eleven restriction endonucleases and cloned DNA fragments are available (Dumas *et al.*, 1981; Ecker and Hyman, 1982; Straus *et al.*, 1982 and 1983; Davison and Scott, 1983).

No data are available at present on the genetic organization and expression of the VZV genome. Therefore, the aim of this study was to determine the DNA sequence of U_S and to derive a model for transcription and translation of this region.

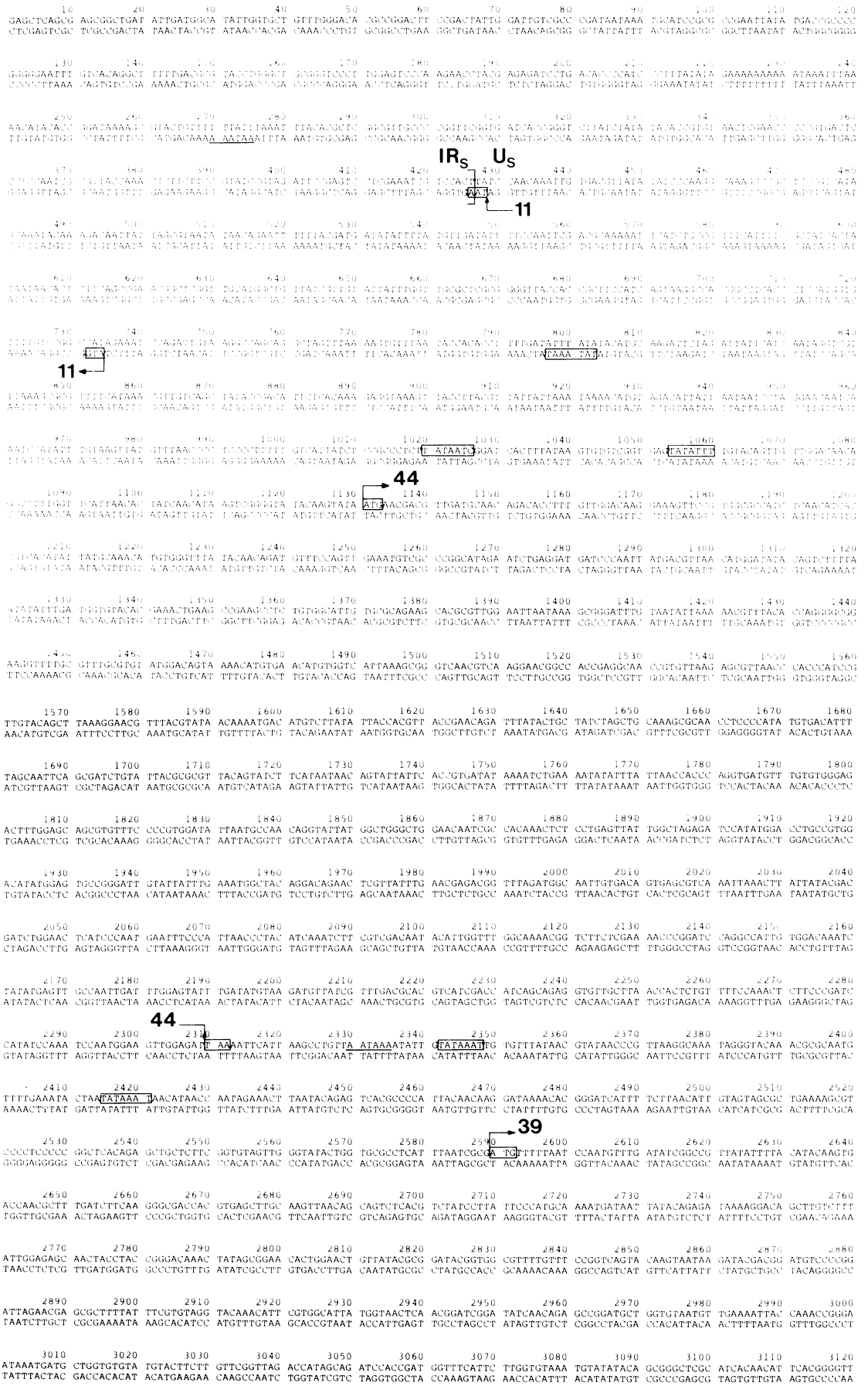
Results and Discussion

The DNA sequence of VZV *Sst*I g is shown in Figure 1. The fragment is 6084 bp in size and consists of U_S (5232 bp) flanked on each side by 426 bp of IR_S/TR_S. The G + C content of U_S is 42.8%, somewhat lower than that of the complete VZV genome (46%, Ludwig *et al.*, 1972), whereas the G + C content of the portion of IR_S/TR_S is slightly higher at 48.8%. No direct or inverted repeats of significant length were found in U_S.

The size of U_S is consistent with that estimated previously (Dumas *et al.*, 1981; Ecker and Hyman, 1982; Davison and Scott, 1983). The significantly larger size of U_S (8700 bp) measured by Straus *et al.* (1982) using electron microscopy is probably incorrect, since comparison of restriction profiles of their strain with those of the strain used in this paper shows little or no size difference in this region (Straus *et al.*, 1981; Dumas *et al.*, 1981; Davison and Scott, 1983). It is likely that the even larger size of U_S (14 700 bp) determined by Gilden *et al.* (1982) using similar techniques is also in error, since they detected no differences between the *Bam*HI restriction endonuclease profiles of their strain and that used by Straus *et al.* (1981). It is significant that Straus *et al.* (1983) did not detect size variability in U_S in their analysis of the restriction endonuclease profiles of several VZV strains.

Model for expression

Figure 2 shows that the distribution of the 572 stop codons present in both strands clearly defines four open reading frames (ORFs) capable of encoding proteins containing >80 amino acids. Table I summarizes an evaluation of probable sites for initiation and termination of transcription and translation for each ORF. The following criteria were used in the analysis. (Sequences are given with respect to the non-coding DNA strand, which is equivalent to the mRNA strand.) (i) Sequences immediately adjacent to first and subsequent in-frame ATG codons were analysed, since Kozak (1981) concluded from a detailed study of functional initiator codons that the most common sequence is (A/G)--ATGG, and therefore postulated this sequence as most favourable for initiation of translation. (ii) Computer-aided analysis of codon usage was used to evaluate coding probability in the region of the first and subsequent in-frame ATG codons (Staden and McLachlan, 1982). (iii) The sequence at the beginning of each ORF was examined for the presence of a promoter element


IRS
US
11
11
44
44
39

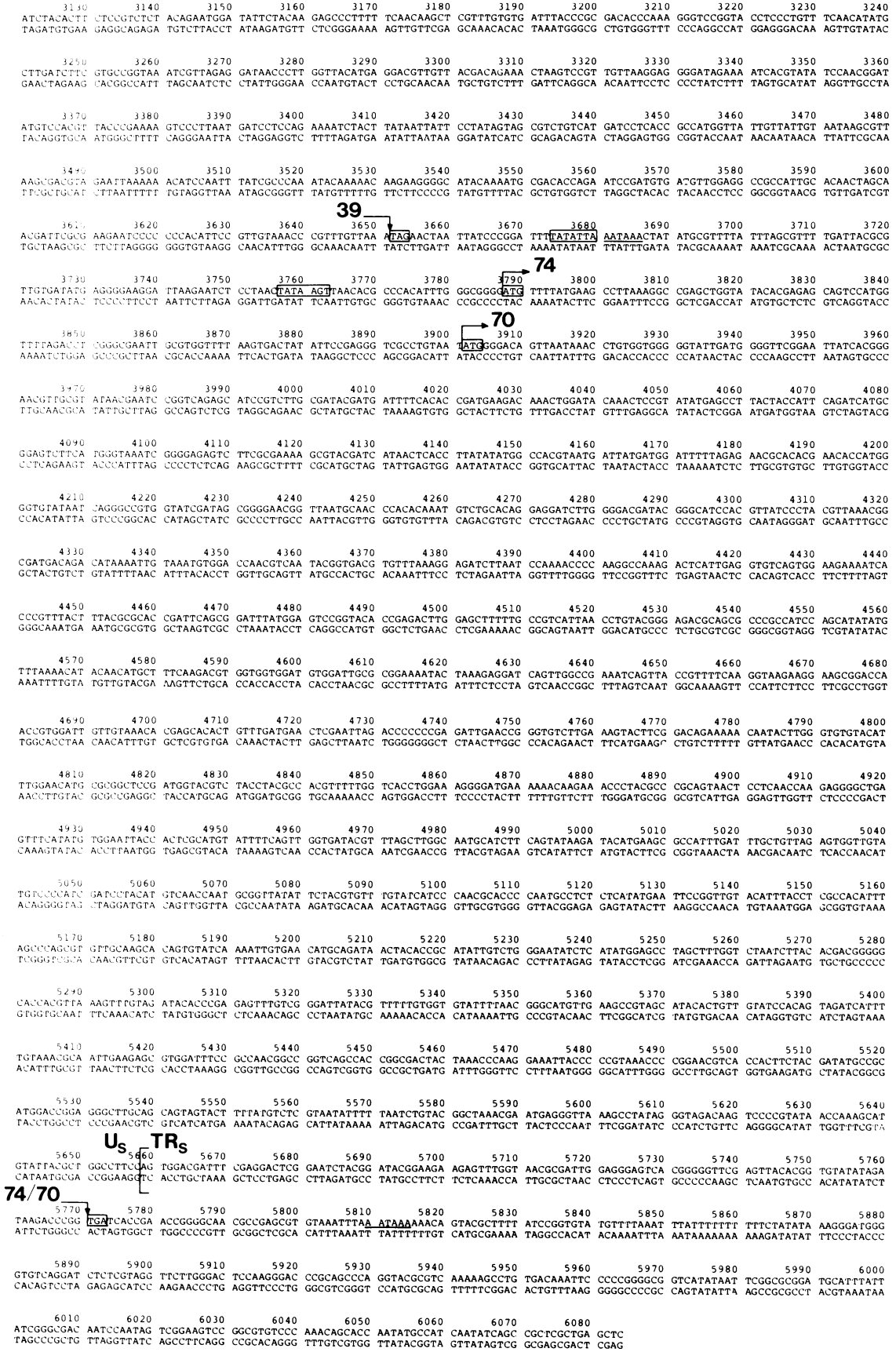


Fig. 1. DNA sequence of VZV *Sst*I g displayed with respect to the I_S genome arrangement. The junctions of U_S with IR_S and TR_S are denoted by square brackets. The following data from Table I are included: transcriptional promoter (boxed) and terminator elements (underlined) and initiation and stop codons (boxed) defining the 11-, 44-, 39- and 74/70-kd proteins.

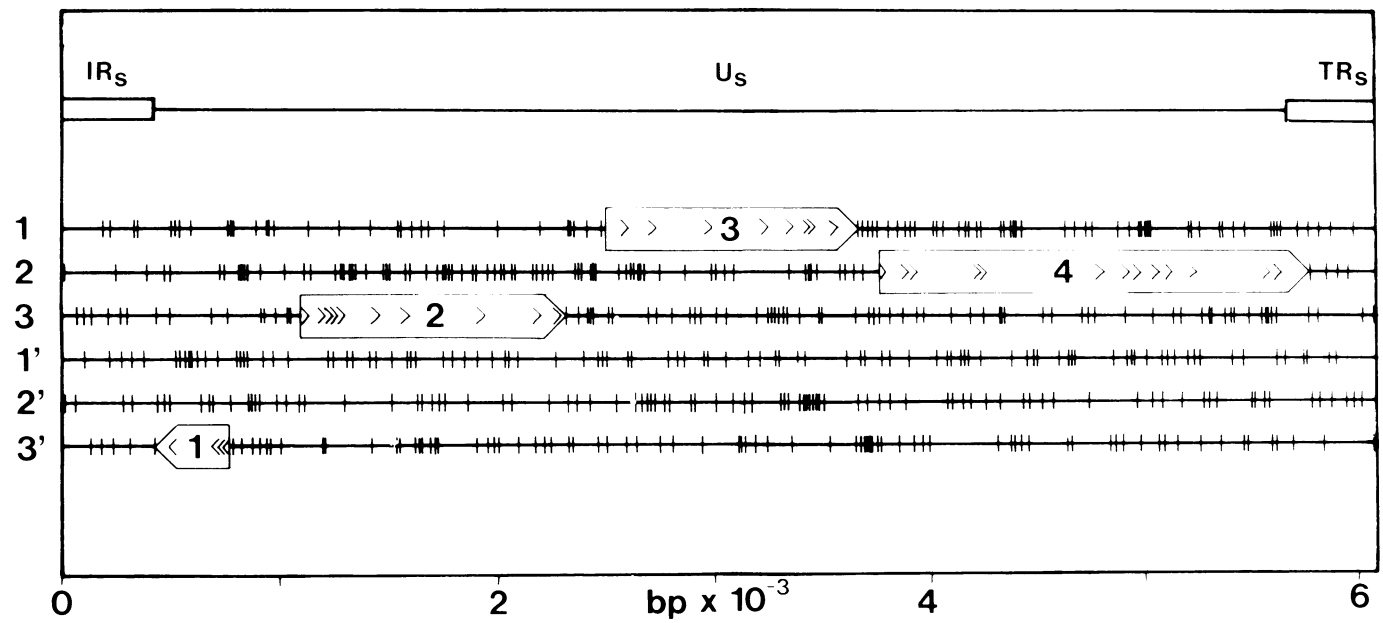


Fig. 2. Distribution of stop codons (vertical lines) in the six reading frames of VZV *Sst*I g. ORFs are numbered and internal in-frame ATG codons are shown as arrowheads pointing in the direction of translation.

Table I. Evaluation of probable sites[a] for initiation and termination of transcription and translation

ORF	Location of ORF	Location of best initiator ATG[b]		Location of best transcriptional promoter	Location of AATAAA	Approximate location of mRNA	Location of protein coding region	Size of primary translation product (kd)
		Local sequence[c]	Codon usage[d]					
1	734′ – 429′	716′	–[e]	803′	275′	770′ – 250′	734′ – 429′	11
2	1131 – 2309	1131	1131	1020 or 1054	2330	1050 – 2350 or 1080 – 2350	1131 – 2309	44
3	2590 – 3651	2590 or 2977	–[e]	2342 or 2415	3681	2380 – 3700 or 2440 – 3700	2590 – 3651	39
4	3788 – 5770	3902	3788	3674 or 3757	5810	3700 – 5830 or 3790 – 5830	3788 – 5770 or 3902 – 5770	74 or 70

[a]Nucleotide locations from the upper strand in Figure 1. Numbers with a prime indicate nucleotides on the lower strand.
[b]First and 3 subsequent in-frame ATG codons evaluated.
[c]Evaluated according to the findings of Kozak (1981).
[d]Evaluated from codon usage in the 50 (ORF1), 200 (ORF2, ORF3) or 400 (ORF4) codons immediately prior to the appropriate stop codon.
[e]Possible initiator codons were not clearly differentiated.

similar to the consensus TATA(A/T)A(T/A), which is usually found 20 – 30 nucleotides upstream from transcriptional initiation sites (Corden *et al.*, 1980). (iv) The sequence close downstream from the end of each ORF was examined for the presence of the element AATAAA, which is usually found 11 – 30 nucleotides upstream from transcriptional termination sites (Proudfoot and Brownlee, 1976).

Sites of initiation of transcription and translation are more difficult to predict than those of termination for two reasons. First, the transcriptional promoter is less well-defined in sequence than the termination element AATAAA; thus, only the elements close to the beginning of ORFs and most similar to the promoter consensus are listed in Table I. Second, translation is not necessarily initiated at the first in-frame ATG codon in an ORF, whereas the stop codon defining the end of the ORF is a clear site for termination of translation. Kozak (1981) has noted that the first transcribed ATG codon functions as the translational initiator in the majority of mRNAs, and in those cases where the first ATG codon is not used, it is located in an unfavourable sequence environment and is almost always followed closely by a stop codon. However, she has given two examples of mRNAs which encode two proteins owing to initiation at the first and second ATG codons, in one of which both proteins are coded in the same frame (Preston and McGeoch, 1981).

It was concluded from the analysis shown in Table I that translation in each of ORF1, 2 and 3 is initiated at the first in-frame ATG codon, giving rise to primary translation products with approximate mol. wts. of 11, 44 and 39 kd, respectively. The predicted mRNAs contain no ATG codons in any frame upstream of the first in-frame ATG codon. It is possible that initiation of translation in ORF1 could also occur at the second in-frame ATG codon, which is more favourable as a potential initiation codon (Kozak, 1981), giving rise to a slightly smaller protein.

The analysis of ORF4 is more difficult to interpret. It is possible that the predicted mRNA does not contain the first in-frame ATG, since one of the two transcriptional promoters of best fit is located only ~30 nucleotides upstream. However, although considerations of local sequence also

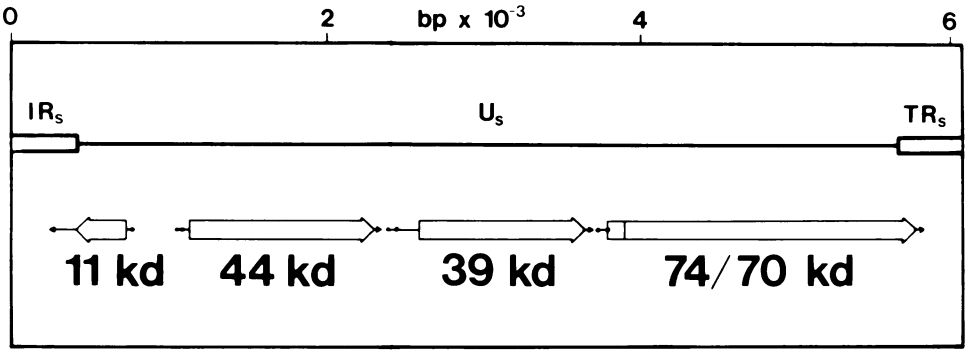


Fig. 3. Model for the expression of VZV *Sst*I g. The approximate positions of non-coding regions of predicted mRNAs are shown as horizontal lines extending from open arrows, which indicate the locations and orientations of protein coding regions.

11 kd

MAGQNTMEGEAVALLMEAVVTPRAQPNNTTITAIQPSRSAEKCYYSDSEN
ETADEFLRRIGKYQHKIYHRKKFCYITLIIVFVFAMTGAAFALGYITSQF
VG

44 kd

MNDVDATDTFVGQGKFRGAISTSPSHIMQTCGFIQQMFPVEMSPGIESED
DPNYDVNMDIQSFNIFDGVHETEAEASVALCAEARVGINKAGFVILKTFT
PGAEGFAFACMDSKTCEHVVIKAGQRQGTATEATVLRALTHPSVVQLKGT
FTYNKMTCLILPRYRTDLYCYLAAKRNLPICDILAIQRSVLRALQYLHNN
SIIHRDIKSENIFINHPGDVCVGDFGAACFPVDINANRYYGWAGTIATNS
PELLARDPYGPAVDIWSAGIVLFEMATGQNSLFERDGLDGNCDSERQIKL
IIRRSGTHPNEFPINPTSNLRRQYIGLAKRSSRKPGSRPLWTNLYELPID
LEYLICKMLSFDARHRPSAEVLLNHSVFQTLPDPYPNPMEVGD

39 kd

MFLIQCLISAVIFYIQVTNALIFKGDHVSLQVNSSLTSILIPMQNDNYTE
IKGQLVFIGEQLPTGTNYSGTLELLYADTVAFCFRSVQVIRYDGCPRIRT
SAFISCRYKHSWHYGNSTDRISTEPDAGVMLKITKPGINDAGVYVLLVRL
DHSRSTDGFILGVNVYTAGSHHNIHGVIYTSPSLQNGYSTRALFQQARLC
DLPATPKGSGTSLFQHMLDLRAGKSLEDNPWLHEDVVTTETKSVVKEGIE
NHVYPTDMSTLPEKSLNDPPENLLIIIPIVASVMILTAMVIVIVISVKRR
RIKKHPIYRPNTKTRRGIQNATPESDVMLEAAIAQLATIREESPPHSVVN
PFVK

74/70 kd

MFYEALKAELVYTRAVHGFRPRANCVVLSDYIPRVACNMGTVNKPVVGVL
MGFGIITGTLRITNPVRASVLRYDDFHTDEDKLDTNSVYEPYYHSDHAES
SWVNRGESSRKAYDHNSPYIWPRNDYDGFLENAHEHHGVYNQGRGIDSGE
RLMQPTQMSAQEDLGDDTGIHVIPTLNGDDRHKIVNVDQRQYGDVFKGDL
NPKPQGQRLIEVSVEENHPFTLRAPIQRIYGVRYTETWSFLPSLTCTGDA
APAIQHICLKHTTCFQDVVVDVDCAENTKEDQLAEISYRFQGKKEADQPW
IVVNTSTLFDELELDPPEIEPGVLKVLRTEKQYLGVYIWNMRGSDGTSTY
ATFLVTWKGDEKTRNPTPAVTPQPRGAEFHMWNYHSHVFSVGDTFSLAMH
LQYKIHEAPFDLLLEWLYVPIDPTCQPMRLYSTCLYHPNAPQCLSHMNSG
CTFTSPHLAQRVASTVYQNCEHADNYTAYCLGISHMEPSFGLILHDGGTT
LKFVDTPESLSGLYVFVVYFNGHVEAVAYTVVSTVDHFVNAIEERGFPPT
AGQPPATTKPKEITPVNPGTSPLLRYAAWTGGLAAVVLLCLVIFLICTAK
RMRVKAYRVDKSPYNQSMYYAGLPVDDFEDSESTDTEEEFGNAIGGSHGG
SSYTVYIDKTR

Fig. 4. One-letter amino acid sequences of predicted primary translation products of VZV *Sst*I g. The amino terminus of the 70-kd protein is indicated by an arrowhead, and hydrophobic regions close to the termini of this and the 39-kd protein are underlined.

favour the use of the second in-frame ATG codon, those of codon usage favour the first. Moreover, two relatively unfavourable out-of-frame ATG codons which are closely followed by stop codons are present between the first and second in-frame ATG codons, and this would be an unusual situation among mRNAs if initiation of translation occurs at the second ATG. These observations make it unwise at present to predict whether the primary translation product of ORF4 has a mol. wt. of 74 or 70 kd.

A model for the expression of *Sst*I g is presented in Figure 3. More complex models have not been ruled out, but it is not necessary at present to invoke RNA splicing since potential transcriptional control elements are present near the beginning and end of each ORF. Sequences at the termini of predicted mRNAs are relatively low in G + C content, and the AATAAA element close to the end of each ORF is located < 35 bp upstream from the complementary sequence TTTATT (ORF1, 3 and 4) or the closely related sequence TTTATA (ORF2). Six of the remaining nine AATAAA sequences in *Sst*I g are present within the ORFs in either strand. The ORFs are in close proximity, ~85% of U_S potentially coding for polypeptide, and conservation of codon usage within and between the ORFs is strong evidence that they do indeed code for proteins.

An interesting comparison may be made between the locations of the junctions between U_S and IR_S/TR_S in the genomes of herpes simplex virus types 1 and 2 (HSV-1 and HSV-2) and VZV. In both HSV-1 and HSV-2, the two mRNAs spanning the junctions share a common 5′ terminus, but the encoded proteins do not share a common amino terminus because the initiator ATG codons are located 8 and 40 nucleotides (HSV-1) or 1 and 33 nucleotides (HSV-2) inside U_S (Watson *et al.*, 1981; Murchie and McGeoch, 1982; Rixon and Clements, 1982; J.L.Whitton and J.B.Clements, in preparation). In VZV, the predicted mRNAs containing ORF1 and ORF4 share a common 3′ terminus, but the encoded proteins do not share a common carboxy terminus because the stop codon for ORF1 spans the junction between U_S and IR_S, whereas that for ORF4 is located inside TR_S. The VZV data add weight to the interpretation of the above HSV-1 and HSV-2 data proposed by J.L.Whitton and J.B.Clements (in preparation) that the extent of IR_S/TR_S is primarily a result of constraints imposed by the locations of adjacent protein coding sequences. Therefore, the extent of IR_S/TR_S could be limited when the two proteins coded by the mRNAs spanning the IR_S-U_S and TR_S-U_S junctions cannot possess a region of common primary structure at either the amino or the carboxy terminus and retain their function, and when the same sequence in IR_S/TR_S cannot encode dissimilar amino or carboxy termini in two reading frames.

Features of predicted primary translation products

Figure 4 shows the amino acid sequences of the four predicted primary translation products and includes features derived from the hydrophobicity plots shown in Figure 5. Data for both the 74- and 70-kd proteins are included because it is not possible to predict which of the two possible amino

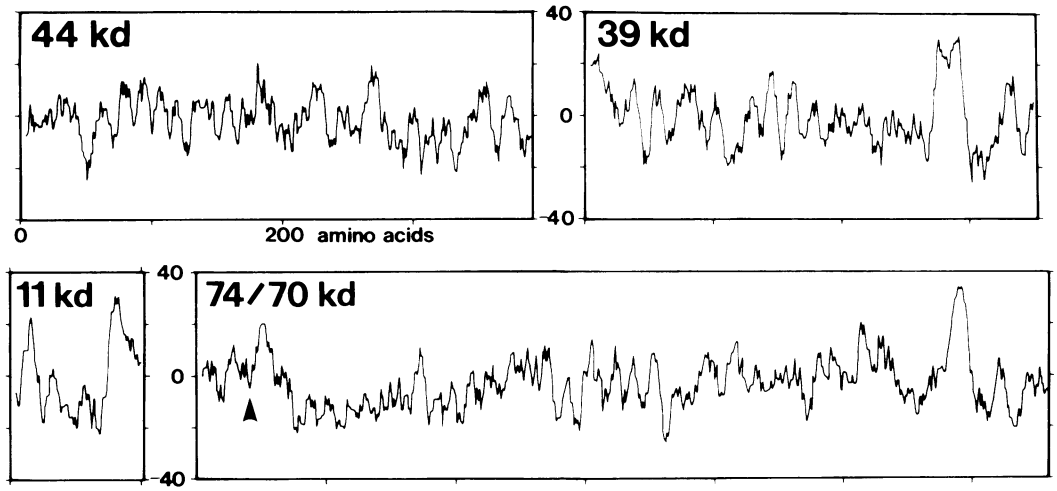


Fig. 5. Hydrophobicity plots of predicted primary translation products of VZV *Sst*I g. The plot was computed by a program using the 'hydropathicity' parameters of Kyte and Doolittle (1982), and involved moving a window spanning nine amino acids along the sequence one amino acid at a time. Peaks indicate hydrophobic regions and the amino terminus of the 70-kd protein is indicated by an arrowhead.

termini of ORF4 is correct. The 39- and 74/70-kd proteins each show a marked hydrophobic region close to the carboxy terminus, followed immediately by a region containing several basic amino acids. This is a characteristic feature of membrane proteins, the hydrophobic portion corresponding to the transmembrane region (Tomita and Marchesi, 1975). Moreover, the 39- and 70-kd proteins each have a hydrophobic region very close to the amino terminus which perhaps corresponds to a signal peptide for membrane insertion (Emr *et al.*, 1980). This may indicate that ORF4 encodes the 70-kd rather than the 74-kd protein. The 11-kd protein also possesses a hydrophobic region close to each terminus, but the lack of basic residues following that at the carboxy terminus makes the properties of this protein difficult to predict. The 44-kd protein lacks the characteristics of membrane proteins.

In conclusion, the likelihood that two of the four genes encode membrane proteins will have a major influence upon future investigation of the expression of *Sst*I g, especially in view of the fact that VZV induces at least four membrane-associated glycoproteins with apparent mol. wts. of 62, 88, 98 and 118 kd (Grose, 1980; Grose and Friedrichs, 1982).

Materials and methods

Cloned DNA fragment

A recombinant plasmid containing *Sst*I g (Davison and Scott, 1983) was transferred from the original host (*Escherichia coli* strain HB101) to a modification-plus host [*E. coli* K12 strain DH1 (Hanahan, 1983)]. *Sst*I g fragment was isolated by agarose gel electrophoresis from plasmid DNA purified as described previously (Davison and Wilkie, 1981).

DNA sequencing

The DNA sequence of *Sst*I g was determined from ~60 000 nucleotides of data derived using the M13-dideoxynucleotide technology (Sanger *et al.*, 1977 and 1980). 95% of the fragment was sequenced on both strands.

Restriction endonuclease fragments or random fragments (400–1000 bp) generated by sonication were inserted into the *Sma*I site of vector M13 mp8 (Messing and Vieira, 1982). Recombinant phage DNA was prepared under conditions of good microbiological practice from infected *E. coli* K12 strain JM101 (Messing *et al.*, 1979) and sequenced using pentadecamer primer (New England Biolabs), large fragment DNA polymerase I (Bethesda Research Laboratories) and [α-^{32}P]dATP (PB 10204; Amersham International). Products were separated in thin 6% polyacrylamide-urea gels (Sanger and Coulson, 1978). Each gel was bonded to one glass plate prior to electrophoresis and then dried prior to autoradiography (Garoff and Ansorge, 1981).

Data handling and analysis

DNA sequence data were manipulated and analysed using the programs described by Staden (1977, 1978, 1979, 1980), Staden and McLachlan (1982), Pustell and Kafatos (1982a, 1982b) and Kyte and Doolittle (1982) in a DEC PDP-11/44 computer operating under the RSX11M system.

Acknowledgements

I am indebted to my colleague Dr. Duncan McGeoch for help and discussion at all stages of this work. My thanks are due to Prof. John Subak-Sharpe for critical reading of the manuscript, to Mr. John Quinn for aid with the sequencing technology, to Dr. Philip Taylor for help with computer programs, and to Mr. Jim Scott for expert technical assistance.

References

Corden,J., Wasylyk,B., Buchwalder,A., Sassone-Corsi,P., Kedinger,C. and Chambon,P. (1980) *Science*, **209**, 1406-1414.
Davison,A.J. and Wilkie,N.M. (1981) *J. Gen. Virol.*, **55**, 315-331.
Davison,A.J. and Scott,J.E. (1983) *J. Gen. Virol.*, in press.
Dumas,A.M., Geelen,J.L.M.C., Maris,W. and van der Noordaa,J. (1980) *J. Gen. Virol.*, **47**, 233-235.
Dumas,A.M., Geelen,J.L.M.C., Weststrate,M.W., Wertheim,P. and van der Noordaa,J. (1981) *J. Virol.*, **39**, 390-400.
Ecker,J.R. and Hyman,R.W. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 156-160.
Emr,S.D., Hedgpeth,J., Clement,J.-M., Silhavy,T.J. and Hofnung,M. (1980) *Nature*, **285**, 82-85.
Garoff,H. and Ansorge,W. (1981) *Anal. Biochem.*, **115**, 450-457.
Gilden,D.H., Shtram,Y., Friedmann,A., Wellis,M., Devlin,M., Fraser,N. and Becker,Y. (1982) *J. Gen. Virol.*, **60**, 371-374.
Grose,C. (1980) *Virology*, **101**, 1-9.
Grose,C. and Friedrichs,W.E. (1982) *Virology*, **118**, 86-95.
Hanahan,D. (1983) *J. Mol. Biol.*, **166**, 557-580.
Kozak,M. (1981) *Nucleic Acids Res.*, **9**, 5233-5252.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105-132.
Ludwig,H., Haines,H.G., Biswal,N. and Benyesh-Melnick,M. (1972) *J. Gen. Virol.*, **14**, 111-114.
Messing,J. (1979) *Recombinant DNA Tech. Bull.*, **2**, 43-48.
Messing,J. and Vieira,J. (1982) *Gene*, **19**, 269-276.
Murchie,M.-J. and McGeoch,D.J. (1982) *J. Gen. Virol.*, **62**, 1-15.
Preston,C.M. and McGeoch,D.J. (1981) *J. Virol.*, **38**, 593-605.
Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature*, **263**, 211-214.
Pustell,J. and Kafatos,F.C. (1982a) *Nucleic Acids Res.*, **10**, 51-59.
Pustell,J. and Kafatos,F.C. (1982b) *Nucleic Acids Res.*, **10**, 4765-4782.
Rixon,F.J. and Clements,J.B. (1982) *Nucleic Acids Res.*, **10**, 2241-2256.
Sanger,F. and Coulson,A.R. (1978) *FEBS Lett.*, **87**, 107-110.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.
Sanger,F., Coulson,A.R., Barrell,B.G., Smith,A.J.H. and Roe,B.A. (1980) *J. Mol. Biol.*, **143**, 161-178.

Staden,R. (1977) *Nucleic Acids Res.*, **4**, 4037-4051.
Staden,R. (1978) *Nucleic Acids Res.*, **5**, 1013-1015.
Staden,R. (1979) *Nucleic Acids Res.*, **6**, 2601-2610.
Staden,R. (1980) *Nucleic Acids Res.*, **8**, 3673-3694.
Staden,R. and McLachlan,A.D. (1982) *Nucleic Acids Res.*, **10**, 141-156.
Straus,S.E., Aulakh,H.S., Ruyechan,W.T., Hay,J., Casey,T.A., Vande Woude,G.F., Owens,J. and Smith,H.A. (1981) *J. Virol.*, **40**, 516-525.
Straus,S.E., Owens,J., Ruyechan,W.T., Takiff,H.E., Casey,T.A., Vande Woude,G.F. and Hay,J. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 993-997.
Straus,S.E., Hay,J., Smith,H. and Owens,J. (1983) *J. Gen. Virol.*, **64**, 1031-1041.
Tomita,M. and Marchesi,V.T. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 2964-2968.
Twigg,A.J. and Sherratt,D.J. (1980) *Nature*, **283**, 216-218.
Watson,R.J., Sullivan,M. and Vande Woude,G.F. (1981) *J. Virol.* **37**, 431-444.