

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 SstI 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 SstI 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 BamHI 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

10 20 30 40 50 60 70 80 90 100 110 120
 GAGCTCAGG AGGGCGTAT ATTGTGCGA TATTGGCGT GTTGGCGAC CGGGGACTT CGACGATTCG GATTGCGCC CGATRAFAA TCAACGCC CGAAATTAA TGACGCCCG
 CTGAGCGC TGCGCGATA TAATCGCTG ATAATACCA CAAACGCTG GGGCGCTGA GGCAGTAAAC TCAACAGCG GCGTAAATTG AGTTAGGGG GCGTTAAAT TGCTGGGGGG
 130 140 150 160 170 180 190 200 210 220 230 240
 GGGCGAAATTG GGTAGAGCT TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 CGCTTAAATG AGCTCGCG AATACCGCG AGGGCGCTG TCAATAGCGG ATTCAGCGT TCGGCGTCTG TGGAGCGA CGTTTAAATG CGAAATTAA TGCTGGGGGG
 250 260 270 280 290 300 310 320 330 340 350 360
 AACACAGCGG CGATTAATG GAGCGCGT TTAATTAATG TCAACGCC CGGGGACTT CGACGATTCG GATTGCGCC CGATRAFAA TCAACGCC CGAAATTAA TGACGCCCG
 TCTGAGCGC TGCGCGATA TAATCGCTG ATAATACCA CAAACGCTG GGGCGCTGA GGCAGTAAAC TCAACAGCG GCGTAAATTG AGTTAGGGG GCGTTAAAT TGCTGGGGGG
 370 380 390 400 410 420 430 440 450 460 470 480
 CGCTTAAATG TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 GGGCGAAATTG GGTAGAGCT TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 490 500 510 520 530 540 550 560 570 580 590 600
 TAAACGCC CGAAATTAA TGCTGGGGGG CGCTTAAATG TTAATTAATG TCAACGCC CGGGGACTT CGACGATTCG GATTGCGCC CGATRAFAA TCAACGCC CGAAATTAA TGACGCCCG
 AACACAGCGG CGATTAATG GAGCGCGT TTAATTAATG TCAACGCC CGGGGACTT CGACGATTCG GATTGCGCC CGATRAFAA TCAACGCC CGAAATTAA TGACGCCCG
 630 640 650 660 670 680 690 700 710 720 730 740
 CGCTTAAATG TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 GGGCGAAATTG GGTAGAGCT TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 750 760 770 780 790 800 810 820 830 840 850 860
 CGCTTAAATG TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 GGGCGAAATTG GGTAGAGCT TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 11 →
 890 900 910 920 930 940 950 960 970 980 990 1000
 CGCTTAAATG TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 GGGCGAAATTG GGTAGAGCT TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 11 →
 1030 1040 1050 1060 1070 1080 1090 1090 1100 1110 1120 1130
 CGCTTAAATG TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 GGGCGAAATTG GGTAGAGCT TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 11 →
 1140 1150 1160 1170 1180 1190 1200 1210 1220 1230 1240 1250
 CGCTTAAATG TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 GGGCGAAATTG GGTAGAGCT TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 1260 1270 1280 1290 1300 1310 1320 1330 1340 1350 1360 1370
 CGCTTAAATG TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 GGGCGAAATTG GGTAGAGCT TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 1390 1400 1410 1420 1430 1440 1450 1460 1470 1480 1490 1500
 CGCTTAAATG TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 GGGCGAAATTG GGTAGAGCT TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 1530 1540 1550 1560 1570 1580 1590 1600 1610 1620 1630 1640
 CGCTTAAATG TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 GGGCGAAATTG GGTAGAGCT TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 1660 1670 1680 1690 1700 1710 1720 1730 1740 1750 1760 1770 1780 1790 1800
 CGCTTAAATG TTGGAGCGG TCGGCGTCTG TGGAGCGA AGAAACGAA AGAGAGCTG ATAGCGCG CGTTTAAATG GAAAAAAAATAAATTTAA
 TGGACAGT TAAAGGAAACG TTACGCTATA ACATAATGAC ATGTCGCTTA TTACGACGTT ACCGACAGAA TTATACCTGC TARCTAGCTG CAAAGCGAA CCTCCCGATA TGTGACATT
 AACATGCGA ATTTCCTGCA AAATGCGAT PGTTCGCTG TTTGTCGACT TGTACACG TAAATGCGT CGCTTAAATG AAATATGCG AGATAGCGC GMPTCGCGT GGAGGGTAT ACACGTTAA
 1810 1820 1830 1840 1850 1860 1870 1880 1890 1900 1910 1920
 ACTTTGGAGC AGCGTGTTC CGCTGGATA TTAACTGCA CAGGTATTAT ATCAAATCTT GTCGACGAT CGAACAGCGC AGCAACGCG AGCGTAAAC CACCCCGCG
 TGGACACCG CGCACAAGA GGGCACCTA ATTAAGCGT GTGTCGCTG TCTGTGCTG AGCAAAACAC TTGCTGCGTGG TGGCGCGTGG CGTACGCTG CGTACGCTG
 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020 2030 2040
 ACATATGGAG TGCGGGATT GATTTGGATT AAATGGCTAC AGGACAGAAC TCGTTATGG AACGAGAGCG TTAGATGGC AATTTGACA GTGAGGCGCA AAATAAACCT ATTATACCG
 TGATACCGC ACGGCGCTAA CATAATACAC TTTCACGCTG TCTGTGCTG AGCAAAACAC TTGCTGCGTGG TGGCGCGTGG CGTACGCTG CGTACGCTG
 2050 2060 2070 2080 2090 2100 2110 2120 2130 2140 2150 2160
 GATCTGACG TCAATCCAA GAATTTCCA TTAACTGCA ATCAAATCTT CGTGACGAT ACATGGCTT CGACGAGCG TGGACGCGC ATGAGCGAG CGGGCATTT CGTACGCTG
 CTGACGCTG AGTAGGGGTG CTTAAAGGGT AATTTGGATG TAGTTGGAGA CGAGCTGTTA TGTAACCAA CGTGTGCTG TGGAGCGAG CGAGAGCGT TGGCGCTAG
 2170 2180 2190 2200 2210 2220 2230 2240 2250 2260 2270 2280
 TATATGAGT CGCAATGTT TTGGAGCTG TGATATGGAA GTATGGPACG TTTGGACGC ACAGCGAGCG GTGATCGGCC ATGAGCGAG CGTGTGCTG CGTACGCTG
 ATATACCA CGGGTAACTA CAACTCTATA ATCATACATG CTACAATAGG AAACCTGCGT CGTAGCGCG TAGTGTGCG TGGAGCGAG CGAGAGCGT TGGCGCTAG
 2290 2300 2310 2320 2330 2340 2350 2360 2370 2380 2390 2400
 CATATCCAA TCAATGCAA GTGGAGCTG AAATTCATI AAGCGTGTG ATAAATATG GATAAAATG TTGTTAAATG GTATAACCG TAAAGGCAA TGGGTAAAC
 GTATAGCTT AGCTTACCTT CAACCTCTAA TTAAAGTAA TTOGGGACAT TATTTTATA CATATTTAC ACAAAATATG CATATGGCG AATTCGTTT ATCCCGTGTG
 2410 2420 2430 2440 2450 2460 2470 2480 2490 2500 2510 2520
 TTTGAAATA CTAATGAA AAACATACAC AAAPAGAACG TAAATACAGC TCAGGCCCA TTAACTGCA CGTGTGCTG TGGAGCGAG CGGGCATTT CGTACGCTG
 AAACATCTG TATTTGAGT TTGGTGGTGG ATTATGGCTT ATGTCGCTG AGCGCGGGT AATTTGCGT CTATGGCTG CGCTTAAATG AAATGGGTGA
 2530 2540 2550 2560 2570 2580 2590 2600 2610 2620 2630 2640
 CGCTTAAATG GGTCAACAGA CGCTGCTTC CGTGACGTT GGTATACGG GGTATACGG GATAACGGG CGTTTGGTT CGGGTGTGTA CGAATGAAAC
 GGGGAGGGGG CGAGGTGCTG CGAGGAGAG CGACATCGAC CGTATGACG CGCGGGAGTA AAATGGCGT ACAAAATTA GGTACAAAC ACATGGCGG
 ACCAACTGTT CGATCTGCTA GGGGACCAA GTGACGCTGC AGGTTAACG CAGTTCACG TGTATCTGTA TCCCGATCGA AAATGATAA TATAACGAGA GTCTGGTT
 TGGTGGCGA ACTAGAACGT CGCGCTGCTG CACTGACG TTCAATGGT CGTACGCG AGATAGGGT TTACATGTTA ATGTCGCTG ATTTGGCTG
 2770 2780 2790 2800 2810 2820 2830 2840 2850 2860 2870 2880
 ATTGGAGAGC AACTACCTAC CGGGACAAAC TATAGGGAA CACTGGACT GTTATACGG GATAACGGG CGTTTGGTT CGGGTGTGTA CGAATGAAAC
 TAACCTCTG TTGATGGATG CGCTGTTTG ATATCGCTT GTGACGCTG CAATATGCGC CTATGGCACC GCAAAACAAAGA GGCAGCTAC GMPTCGCGT
 2890 2900 2910 2920 2930 2940 2950 2960 2970 2980 2990 3000
 ATTAGACGA GGGCTTTTAT TTGGTGGTGG TACAAACATT CGTGGCATTA TGTTAACTCA ACGGATCGGA TATCAACAGA CGCGGATGCT GGTGATATGT
 TAATCTGCTG CGCGGAAATA AAGCACACAT ATGTTGAA CGACCGTAA ACCATGGATG TGGCTAGCTT ATGTTGCTG CGGCTACGA CGACATACAA
 3010 3020 3030 3040 3050 3060 3070 3080 3090 3100 3110 3120
 ATAATGATG CTGGTGTGTA TGCTACTCTT GTTCCGGTAG ACCATGAGC ATTCACCGAT GTTTCGCTT TTGGCTGATA TGTATATACA CGGGGCTGC
 TTTACTGCTG CGCGGAAATA AAGCACACAT ATGTTGAA CGACCGTAA ACCATGGATG TGGCTAGCTT ATGTTGCTG CGGCTACGA CGACATACAA
 3130 3140 3150 3160 3170 3180 3190 3200 3210 3220 3230 3240

3140 3140 3150 3160 3170 3180 3190 3200 3210 3220 3230 3240
ATCTGACATT CTCCGTCCTC ACAGAATGGA TATTCCTACAA GAGCCCTTT TCAACAAAGT CGTTTGTG TG ATTACCCGC GACACCCAA GGGTCGGTA CCTCCCGT TT CAACATATG
TAGATGTGAA GAGGAGAGA TGCTTACTC ATAAGATGTT CTCGGAAA AGTTGTGCA GCAAACACAC TAATGGCG CTGTTGGTT CCCAGGCC ATGGGGACAA AGTTGTATAC

3250 3260 3270 3280 3290 3300 3310 3320 3330 3340 3350 3360
CTTGATCTTC GTGCCGGTAA ATCGTTGAG GATAACCCCT GGTTACATGA GGACGTTGTT ACAGCACAAA CTAAGCTCGT GTTTAAGGAG GGGATAGAAA ATCACGTT CTCACCGGAT
GAACTAGAG CACGCCATT TAGCAATCTC CTATTGGAA CCAATGACT CCTGCAACAA TGCTGTTTT GATTCAAGGCA ACAATTCCC CCCTATCTT TAGTGATAT AGGTTGGCTA

3370 3380 3390 3400 3410 3420 3430 3440 3450 3460 3470 3480
ATGTTGACGT TACCGGAAA GTCCCCTAA GATCCCTCCAG AAAATCTACT TATAAATTATT CCTATAGTG CGTCTGTCAT GATCTTCACCC GGCATAGGTTA TTGTTATTGTT AATAAGCGTT
TAAGGTTGAA ATGGGCTTT CAGGGAAATA CTAGGGAGTC TTTAGATGAG ATTTAATAA GGATATCATC CGACGACAGA CTAGGAGTGG CGGTACCAAT AACAAATAACA TTATTGCGA

3490 3500 3510 3520 3530 3540 3550 3560 3570 3580 3590 3600
AAUGGAGGAA GAATAAAA ACATCCAATT TATCGGAAATC ATACAAAAAC AAGAAAGGGGAC ATACAGAACG CGACACCCAG ATCCGAGTGC ATGTTGGAGG CGCGGATTGAC ACAACTAGCA
PPGCTGAACT TTAAATTTTT TGAGGTTAA ATAGCGGGTT ATATGTTTTG TTCTTCCCCG TATGTTTAC GCTGTTGCT TAGGTCACAC TACAACCTCC GGCGTAAACG TGTTGATCGT

39
3610 3620 3630 3640 3650 3660 3670 3680 3690 3700 3710 3720
ACGATTCGGG AAGAAATCCC CCCACATTCC GTTGTAACC CGTTTGTAA **AAGAACTAA** TTATCCCGA **TATATA** AATAAATCT ATGCGTTTTA TTAGCGTT TGATTAGCG
TGCTAAGGCG TTCTTGGGG GGGTGTAAAGG CAACATTTGG GCAACAAATT TATCTTGTATT AAATAGGCCCT **AATAATAT** TTATTTGATA TACGCAAAAT AAATCGCAA ACTAATGCGC

74
3730 3740 3750 3760 3770 3780 3790 3800 3810 3820 3830 3840
TTTGTGATAG AGGGGAAGG TAAAGAATCT CTCAACT**AATAAAGTACACG** CCCACATTGG GGGCGGAG **TATATA** TTATAGAG CTTAAAGGCG CGAGCTGGTA TACACGAGAG CAGTCATGG
AAACATATACTC TCCCTTCTT AATTCTTAGA GGATTGATAT TCAATTGTC GGGGTAAAC CGCCGCTTAC AAAAATCTC GGAATTCCG GCTGACCAT ATGTCGTC TGTAGGAC

70
3850 3860 3870 3880 3890 3900 3910 3920 3930 3940 3950 3960
TTTTGAGAT CGGGGAAGT GCGTGGTTT AAGTGAATCT ATTCCGAGGG TCAGCTGAA **TATATA** GTTAATAAC CTGTTGGTTG GGTATTGAG GGGTTGGAA TTATCACGGG
AAATCTGGA GGGCGCTTAA CGACCAAACTA CTACTGTATA TAAGGCTCCC AGCGGACATT ATACCCCTGT CAATTATTG GACACCCAC CCATAACTAC CCCAACGCTT AATAGTGCCT

3970 3980 3990 4000 4010 4020 4030 4040 4050 4060 4070 4080
AACGTTGGGT AATAACGAA CGGTGAGAGC ATCCGTCCTG CGATACGATG ATTTCACAC CGATGAGAC AAACCTGGT ATATGAGCCT TACTACATT CAGATCATGC
TTGCAACGCA TATGCTTACG GCCAGCTCTG TAGGCAGAAC GCTATGTC TAAAGTGTG GCTACTCTG TTTGACCTAT GTTGAGGCA TATACTCGGA ATGATGGTAA GTCTAGTACG

4090 4100 4110 4120 4130 4140 4150 4160 4170 4180 4190 4200
GGAGCTTCA PGGGAAATC GGGGAGGAGC TTCCGAAAGG CGGTGAGTC ATTAACTCAGC TTATATATGG CCACGTAATG ATTATGAGG AAGGACACAG AACACCATGG
CTGAGAAGT ACCCTTACG CCCCCCTCAG AAGGCTTTT CGCATGCTAG TATTGAGTAC AATATACCC GGTGCAATTAC TAATACTACCC TAAAAATCTC TTGCGTGTGC TTGTTGGTAC

4210 4220 4230 4240 4250 4260 4270 4280 4290 4300 4310 4320
GGTGATAAT JAGGGCCCGT GTATGATAG CGGGGAAACGG TTAATGCAAC CCACACAAAT GTCGTCACAG GAGGATCTTG GGGACGATAC GGGCATCCAC GTTATCCCTA CGTTAACCGG
CCACATATAA GTCCCGGCAC CATAGCTATC GCCCCTTGCC AATTACCTTG GGTGTTTTA CAGACGTC GTCCTAGAAC CCCTGCTATG CCCGTAGGTG CAATAGGGAT GCAATTGCG

4330 4340 4350 4360 4370 4380 4390 4400 4410 4420 4430 4440
CGATGACAG CATAAAATG TAAATGTTG GCAACGTC TACGGTACG TGTTTAAAGG AGATCTTAA CTAAACCTCC TCTAGAATTG GGGTTGGG TTCCGGTT TGAGTAACCTC CACAGTCACC TTCTTTAGT

4450 4460 4470 4480 4490 4500 4510 4520 4530 4540 4550 4560
CCCCTTACT TTACGCCAC CGATTCAAGG GATTATGGA GTCCCGTACA CGGAGACTTG GAGCTTTTG CGCGTCAATT CCTGTACGGG AGACGAGCGG CCCGCCATCC AGCATATATG
GGGCAATGA ATGCGCTG GCTAACTACCT CAGGGCATGT GGCTCTGAAC CTGCAAAACAC GGCAGTAAATT GGACATGCC TCTGCGTCC GGGCGTAGG TCGTATATAC

4570 4580 4590 4600 4610 4620 4630 4640 4650 4660 4670 4680
TTAAACACAT ACAACATGCT TTCAAGACGT GGTGGTGGAT GTGGATTGGC CGGAAATAC TAAAGAGGAT CAGTTGGCCG AAATCAGTTA CCGTTTTCAA GTTAAGAAGG AAGCGGACCA
AAATTGTA TGTGTAACG CACACTTGTG GCTCGTGTGA CAAACTACTT GAGCTTAAATC TGGGGGGGT CTAACATTGGC CCACAGAACT TTCATGAGC 4780 4790 4800

4690 4700 4710 4720 4730 4740 4750 4760 4770 4780 4790 4800
ACCGTGGATT GTTGAACACA CGACCAACT GTTTGTGATG CTCGAAATTG ACCCCCCCGA GATTGAACCC GGTGCTTGC AAGTACTCTG GACAGAAAAA CAATACTTGG GTTGTGACAT
TGGCCTCAA CAACATTGT GCTCGTGTGA CAAACTACTT GAGCTTAAATC TGGGGGGGT CTAACATTGGC CCACAGAACT TTCATGAGC 4780 4790 4800

4810 4820 4830 4840 4850 4860 4870 4880 4890 4900 4910 4920
TTGGACATG CGGGCTCG ATGGTACCGC TACCTACGGC AGCTTTGG TGACCTGGAA GGGGAAAGG AAAACAGAA ACCCTACGGC CGCAGTAACCT CCTCAACCC GAGGGGCTAC
AACCTGAGC TGGGCGAGC TACCATGAG ATGGATGGG CGTACCAACTT TGCCAAACCG AGTGGACATT TTCCCTACTT TTGTTCTT TGGGATCGG GCCTCATTTGA GGAGTTGGTT CTCCCCGACT

4930 4940 4950 4960 4970 4980 4990 5000 5010 5020 5030 5040
GTTTCAATG TGGAATTACG ACCTGAGATG ATTTCAGTT GGTGATAGCTT TTAGCTTGGC ATAGCATTT CAATGATAAGA TACATGAGC GGCATTGAT TTGCTGTTAG ATGTTGTTGA
CAAATGATA ACCTTAATGG TGACCTACA TAAAAGTCAA CCACTATGCA AATGCAACCC TTACCTGAGA ATGTTGAGTAC TACATGAGC AACGACAATC TCACCAACAT

5050 5060 5070 5080 5090 5100 5110 5120 5130 5140 5150 5160
TGTCGATC GATGCTATC GTCACCACT GGGGTATAT TCTACGTTGTT TGATCATCC CAACGCCACCC CAATGCTCTT CTATGATGAA TTCCGGTTGT ACATTTCACCT CGCCACATT
ACAGGGGTAG ATAGGATGTA JAGTGTTTA CGGCAATAAT AGATGCAACCA ATAGTACTG GTTGGCTGGG TTACGGAGA GAGTACTATC AAGGCCAACA TGTAATGGA GCGGTGTA

5170 5180 5190 5200 5210 5220 5230 5240 5250 5260 5270 5280
AGCCAGGGT GTPGCAAGCA CAGTGATACCA AAATTGTTGAA CATGCAGATA ACTACACGGC ATATTGCTG GGAATATCTC ATATGGAGCC TAGCTTTGGT CTAACTCTAC ACAGCGGGG
TGGGCGAGA CAAAGTTGCT GTACACATAG TTTAACACCTT TGACGCTTAT GATGTTGGC TATAACAGAC CCTTATAGAG TATACTCCG ATCGAACCA GATTAGAAATG TGCTGCCCG

5290 5300 5310 5320 5330 5340 5350 5360 5370 5380 5390 5400
CACCGGTTA AGGTTGTTG ATACACGGC GAGTTGTCG GGATTATAGC TTGTTGTTGTT GTATTTAAC GGGCATCTG AAGGCGTAGC ATACACTGTT GATATCAGC TAGTCAATT
GTTGGCGAAC TTCAACACAT TGTGGGGCT CTCAAACAGC CCTTAATATGC AAAAACACCA CATAAAATTG CCCGTACAC TATGTCGACAA CATAGGTGTC ATCTAGTAA

5410 5420 5430 5440 5450 5460 5470 5480 5490 5500 5510 5520
TGPAACGCC AATGGAAGAGC GTGTTTACCC GCAACGGCC GGTGACGGC CGGCGACTAC TAAACCAAG GAAATTACCC CGCTAACCC CGGAACGTCA CCACCTCTAC GATATCCGC
ACATTGGGT TAATCTCTG CACCTAAAGG CGGTTGGCCG CGAGTCGGTG ATTTGGGTC TTAAATGGG GGCATTGGG GCTTGTGAGT GTGAGAATG CTATACGGC

5530 5540 5550 5560 5570 5580 5590 5600 5610 5620 5630 5640
ATGGACCGG GGGCTTGCAG CAGTAGACTT TTGATGCTC GTAATTTTT TAATCTGTC GGCTAACCGA ATGAGGTTA AAGCCTATAG GGTAGACAG TCCCCGTATA ACCAAAGCAT
TACCTGGCT CCCGACGTC GTCATCATGAA AAATACAGC CATTAAATTG ATTAGACATG CGGATTTGGCT TACTCCCTAT TTGCGATATC CCATCTGTC AGGGGATAT TGGTTGCGA

U_S **TR_S**
5650 5660 5670 5680 5690 5700 5710 5720 5730 5740 5750 5760
GTATTAAGCT GGGCTTGCAG TGGACGATT CGAGGACTCG GAATCTACGG ATACGGAAAGA AGATGTTGTT ACGCGTATG CTTGCGTAACT CTCCTCTAG GCCCCAACAG TCAATGTC
CATAATGGCA CGGAGGAGTC ACCTGCTAA GCTCTCTAGG CTAGATGTC TTAGCTTCT TCTCAACCA TTGCGTAACT CTCCTCTAG GCCCCAACAG TCAATGTC ACATATATCT

74/70
5770 5780 5790 5800 5810 5820 5830 5840 5850 5860 5870 5880
TAAGCCGGT **GTGTCACGCA** ACAGGGGAA CGCCGGCGT GAAATTTAA **ATAA** AAC GTCACCTTT ATCCGGTTA TGTTTTAAAT TTATTTTTTTT TTCTATATAA AAGGGATGGG
ATTCTGGGGC ACTAGTGGCT TGGGGCGCTT GGGCTCGCA CATTAAATT TATTGTTGTT CATGGCAAA TAGGCCACAT ACAAATTTA AAAAAAAA AAAGATATAT TTCCCTACCC

5890 5900 5910 5920 5930 5940 5950 5960 5970 5980 5990 6000
GTGTCAGGAT CTCTCGTAGG TTCTGGGAC TCCAAGGGAC CGCGACGCC GGTACGGCTC AAAAGGCCG TGACAAATTG CCGCGGGGG GTCATATAAT TCGGCGGGA TGCAATTATT
CACAGCTCA GAGAGCATC AAGAACCCCTG AGGTTCCCTG CGGCTCGGGT CCATGCGAG TTTTCCGGAC ACTGTTTAAG GGGGCCCGC CAGTATATAA ACAGCCGCCT ACGTAAATAA

6010 6020 6030 6040 6050 6060 6070 6080
ATCGGGGAGC AATCCAAATG TCGGAAGTC GCGCTGTCAC AACACGCC AATATGCCAT CAATATCAGG CGCTCGCTGA GCTC
TAGCCCGCTG TTAGGTTATC AGCCTCTAGG CGGCACAGGG TTGTCGTTG TTATACGCTA GTTATAGTCG CGGAGGCACT CGAG

Fig. 1. DNA sequence of VZV SstI 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.

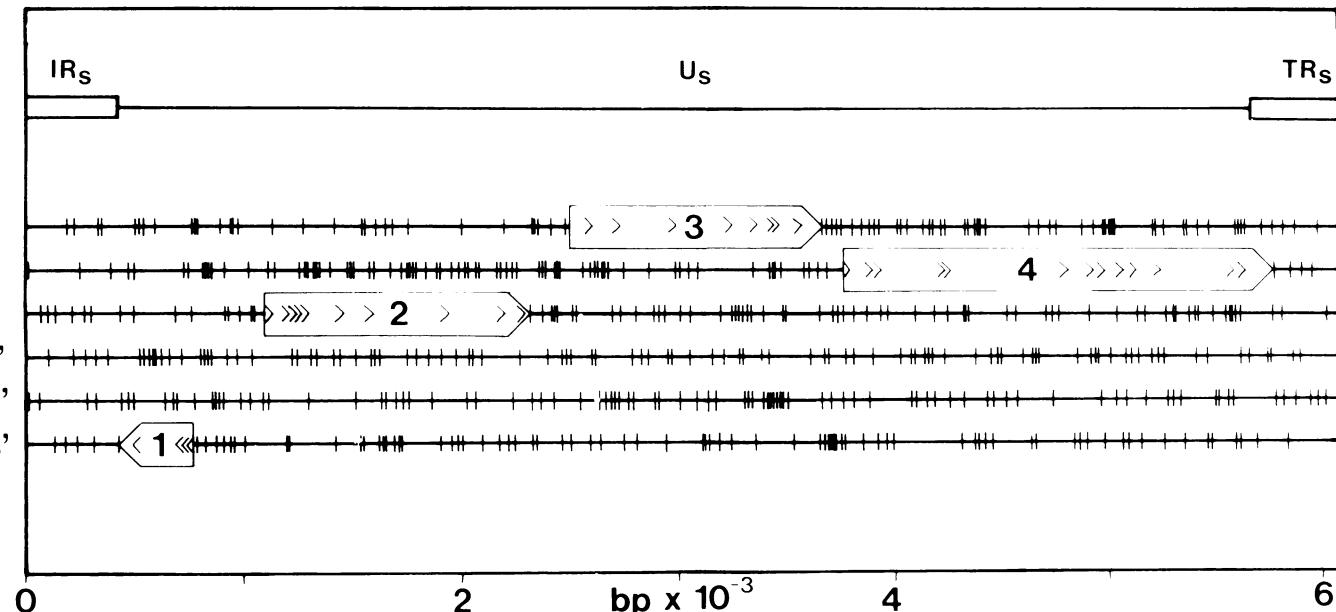


Fig. 2. Distribution of stop codons (vertical lines) in the six reading frames of VZV SstI g. ORFs are numbered and internal in-frame ATG codons are shown as arrowheads pointing in the direction of translation.

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 |

^aNucleotide locations from the upper strand in Figure 1. Numbers with a prime indicate nucleotides on the lower strand.

^bFirst and 3 subsequent in-frame ATG codons evaluated.

^cEvaluated according to the findings of Kozak (1981).

^dEvaluated from codon usage in the 50 (ORF1), 200 (ORF2, ORF3) or 400 (ORF4) codons immediately prior to the appropriate stop codon.

^ePossible 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

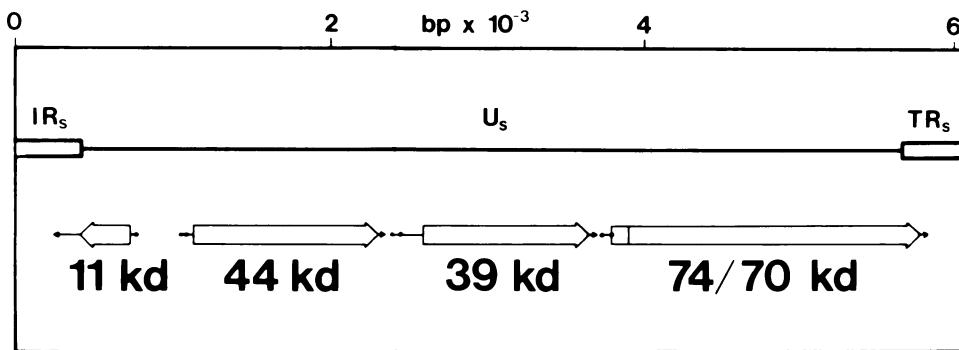


Fig. 3. Model for the expression of VZV *SsI g*. The approximate positions of non-coding regions of predicted mRNAs are shown as horizontal lines extending from open arrows, which indicate the locations and orientations of protein coding regions.

11 kd

MAGQNTMEGEAVALLMEAVVTPRAQPNNTTITAIQPSRSAEKCYYSDFSEN
ETADEFLRRIGKYQHKIYHRKKFCYITLIIIVFVFAMTGAFAFLGYITSQF
VG

44 kd

MNDVADATDTFVGQGKFRGAISTSPSHIMQTCGFQQMFPVEMSPGIESED
DPNYDVNMDIQSFNIFDGVHETEAEASVALCAEARVGINKAGFVILKFT
PGAEQAFACMDSKTCHEVVAKAGQRQCTATEATVLRALTHPSVQVLKGT
FTYNKMTCLIPRYTDLCYLAALKRNLPCIDILAIQRQSVRLAQYLHNN
SIIHRDIKSENFIINHPGDVCVGDFGAFCFVDINANRYWGAWTIATNS
PELLARDPYGPADWIWSAGIVLFEMATGQNSLFERDGLDGNCDSERQIKL
IIRRSGTHPNEFPINPTSNLRRQYIGLAKRSRKPGSRPLWTLNLYELPID
LEYLIKMLSFDARHRSPEVLLNHHSVQTLPPDPYPNPMPMEVGD

39 kd

MFLIQCLISAVIFYIQVTNALIFKGDHVSQVNSSLTSIILPMQNDNYTE
IKGQLVFIGEQLPTGPNYSGTLELLYADTVAFCRSVQVIRYDGCPRI
SAFISCRYKHSWHYGNSTDRISTEPDAVGMLKIKTPGINADGVVLLVR
DHSRSTDGFILGVNVYTAGSHHNHGVYITSPSLQNGYSTRALFQQARLC
DLPATPKGSQTSLFQHMDDLRAGKSLEDNPWLHEDVVTETKSVVKEGIE
NHVYPTDMSTLPEKSLNDPPENLLIIPIVAVSAMVILAMIVIVSKRR
RIKKHPIYRPNTKTRRGIQNATPESDMVLEAAIAQLATIREESPSPHSV
PFVK

74/70 kd

MFYEALKAELVYTRAVHGFRPRANCVVLSDYIPRVACNMGTVNKPVVGVL
MGFGIITGLRITINPVRASVLRYDDFTEDKDLDNTNSVYEPYHSDHAES
SWVNRECESSRKAHDNSPYIWPRLYDGFLENAEHGGVYQNQGRGIDS
RLMQPTQMSAQEDLGDDTGIHVITLNGDDDRHKIVNVQDRQYGDVFKGDL
NPKPQGQRLIEVSVEENHPFTLRAPIQRIYGVRVETWSFLPSLTCTGDA
APAIQHICLKHTTCFQDVVVDVDAENTKEDQLAEISYRFQGKKEADQPW
IVVNTSTLDELELDPPEIEPGVLKVLRTKEQYLGVIWNMRGSDGTSTY
ATFLVTKGDEKTRNPTPAVTQPGRAEFHMWNYHSHVFSVGDFTSLAMH
LQYKIHEAPFDLLWEWLVPIDPTCQPMRLYSTCLYHPNAPQCLSHMN
CTFTSPHLAQRVASTVYQNCEADNYTAYCLGISHMEPSFGLILHDGGTT
LKFVDTDPESLSGYVVFVVFNGHVEARTVVSTDHVFNNAIEERGFPT
AGQPPATTKPKETPPVNPGBTSPLLRYAAWTGGLAAVLLCLVIFLICKA
RMRVKAYRVDKSPYNQSMYYAGLPVDDFEDSESTDTEEEFGNAIGGSHGG
SSYTYYIDKTR

Fig. 4. One-letter amino acid sequences of predicted primary translation products of VZV *SsI 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 *SsI 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 *SsI 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

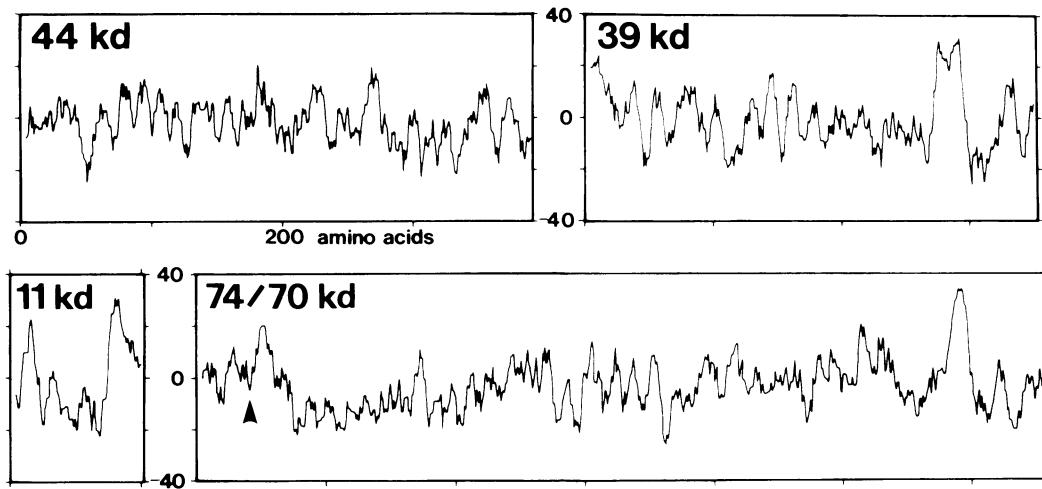


Fig. 5. Hydrophobicity plots of predicted primary translation products of VZV SstI g. The plot was computed by a program using the 'hydropathicity' parameters of Kyte and Doolittle (1982), and involved moving a window spanning nine amino acids along the sequence one amino acid at a time. Peaks indicate 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 SstI 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 SstI 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)]. SstI 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 SstI 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 -³²P]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., Wasylkyk,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.