# The complete genome sequence of the *Streptomyces* temperate phage ϕC31: evolutionary relationships to other viruses

Margaret C. M. Smith*, R. Neil Burns, Stuart E. Wilson and Matthew A. Gregory

Institute of Genetics, University of Nottingham, Queen's Medical Centre, Nottingham NG7 2UH, UK

Received January 22, 1999; Revised and Accepted March 30, 1999 DDBJ/EMBL/GenBank accession no. AJ006589

## ABSTRACT

**The completed genome sequence of the temperate *Streptomyces* phage ϕC31 is reported. ϕC31 contains genes that are related by sequence similarities to several other dsDNA phages infecting many diverse bacterial hosts, including *Escherichia*, *Arthrobacter*, *Mycobacterium*, *Rhodobacter*, *Staphylococcus, Bacillus, Streptococcus, Lactobacillus* and *Lactococcus*. These observations provide further evidence that dsDNA phages from diverse bacterial hosts are related and have had access to a common genetic pool. Analysis of the late genes was particularly informative. The sequences of the head assembly proteins (portal, head protease and major capsid) were conserved between ϕC31, coliphage HK97, staphylococcal phage ϕPVL, two *Rhodobacter capsulatus* prophages and two *Mycobacterium tuberculosis* prophages. These phages and prophages (where non-defective) from evolutionarily diverse hosts are, therefore, likely to share a common head assembly mechanism i.e. that of HK97. The organisation of the tail genes in ϕC31 is highly reminiscent of tail regions from other phage genomes. The unusual organisation of the putative lysis genes in ϕC31 is discussed, and speculations are made as to the roles of some inessential early gene products. Similarities between certain phage gene products and eukaryotic dsDNA virus proteins were noted, in particular, the primase/helicases and the terminases (large subunits). Furthermore, the complete sequence clarifies the overall transcription map of the phage during lytic growth and the positions of elements involved in the maintenance of lysogeny.**

## INTRODUCTION

The Gram-positive bacteria in the genus *Streptomyces* are prolific producers of complex secondary metabolites, many of which have antibiotic or other biological activities. A major aim in the study of *Streptomyces* species has been to understand the biosynthesis of antibiotics and to exploit genetic recombination so that new compounds with novel activities can be created. Tools for genetic engineering of *Streptomyces* have been developed from ϕC31, most notably phage cloning vectors (1,2), site-specific integration vectors (3,4) and a cosmid cloning vector (5). About half of the DNA sequence of ϕC31 has been determined, giving information on the repressor locus (6), the essential early genes (7), the integrase region and *attP* site (8,9) and part of the late region (10). The completed ϕC31 sequence should help to develop new molecular tools for *Streptomyces* research and improve existing ones.

Another reason for completing the sequence of ϕC31 is to compare its sequence to other phage genomes. There are now approximately 30 completed phage genome sequences in the databases, a tiny fraction of the estimated total numbers of phages in the world (~$10^{30}$; 11,12; R.Hendrix, personal communication). Most of the completed genome sequences are for phages that infect Gram-negative eubacteria or the A+T-rich Gram-positive bacteria such as the *Lactococci*, *Streptococci* or *Bacilli*. There are, however, two completed mycobacteriophage genomes, L5 (13) and D29 (14), which, as phages that infect close relatives of the streptomycetes, might have features similar to ϕC31. Indeed L5, D29 and ϕC31 are unusual amongst temperate phages in that they encode genes for DNA polymerases (7,13,14). Sequencing bacterial genomes has also been extremely productive for obtaining the sequences of prophages from diverse bacteria (15–18). The analyses of phage genomes by sequencing (originally using heteroduplex analysis) have shown that phages are exceptionally diverse. A large part of this diversity is due to mosaicism arising by homologous and illegitimate recombination between members of a phage family such as the lambdoid family (19,20) or amongst the closely related streptococcal phages (21,22). Recently, however, sequence conservation between individual genes within genomes of phages that infect evolutionarily very diverged hosts has been observed (23,24). These data strongly suggest that all dsDNA phages share common ancestry and are in genetic contact with each other by horizontal exchanges from a common genetic pool. However, the degree of access to the global phage gene pool is not thought to be uniform (23); there are clearly areas of freer exchange, such as within the lambdoid or streptococcal phages (19,21,22). Between phages that infect more diverse hosts there are barriers (e.g. host range) which reduce the frequency of exchange. By analysing more

*To whom correspondence should be addressed. Tel: +44 115 9709778; Fax: +44 115 9709906; Email: maggie.smith@nottingham.ac.uk phage genomes from diverse hosts it should be possible to deduce the nature of, and mechanisms for overcoming, these barriers.

Whilst there are numerous completed phage genomes, the transcriptional circuitries of rather fewer phages have been studied. In the case of ϕC31 the control of the lytic and lysogenic cycles involves some novel mechanisms. A global transcription map of ϕC31 showed that the early and late genes were clustered, with the late genes mainly on the left arm and the early genes on the right arm (25). Early transcripts arise from multiple phage specific promoters, and inefficient termination results in overlapping mRNAs (25–28). Late transcription is thought to occur via a single unstable transcript arising from a promoter located just downstream of the integrase gene at the extreme right hand end of the genome (10,25). Both early and late promoters contain a highly conserved 21 bp sequence (27,29). These promoters, which are completely inactive in uninfected cells or in uninduced lysogens, are activated during phage growth. We presume that the phage encodes an activator of these promoters, which must be repressed during lysogenic growth. The mechanism of temporal control of early and late lytic promoters is not understood. The control of lysogeny occurs via the action of the products of the *c* gene, which expresses three N-terminally different inframe isoforms of 74, 54 and 42 kDa (30). The 54 and 42 kDa isoforms have been shown to bind to a 17 bp conserved inverted repeat (CIR) sequence located in multiple copies all along the phage genome (31–33). Characterisation of the genetic lesion in a virulent mutant of ϕC31 and DNA binding studies suggested that one CIR site, CIR6, was important for controlling the lytic–lysogenic switch in ϕC31 (32). If CIR6 controls the expression of an activator of lytic promoters then a typical control circuit can be envisaged. Why there are so many repressor binding sites is not clear.

In this paper we present an analysis of the complete genome sequence of ϕC31 with the aim of providing information concerning the evolution of phage genomes and the specific adaptations acquired by ϕC31 for growth in *Streptomyces* spp. We also aim to provide a global view of the repressor binding sites, phage specific promoters and terminators. The sequence of ϕC31 is the first *Streptomyces* phage genome to be completed.

## MATERIALS AND METHODS

Sequence of the late and inessential regions of ϕC31 were determined using ABI Prism, Dye Terminator, cycle sequencing kits and an ABI373 sequencer (34). The reactions were primed from universal or customised primers and purified plasmid DNA was used as templates. The sequencing strategy employed plasmids containing restriction fragments from *Eco*RI, *Sph*I, *Hin*dIII or *Kpn*I digests isolated from wild type ϕC31 Norwich stock. The sequence of each restriction fragment was determined by designing primers to extend the sequences from the ends of the inserts, ultimately to obtain the DNA sequence from both strands. The sequences across the cloning sites were determined using overlapping restriction fragments. Ms Damji and Dr Leskiw kindly donated ~3 kb of sequence overlapping SphI-F; this region was re-sequenced using customised primers to confirm accuracy. Plasmids were prepared for sequencing using standard techniques (35). ϕC31 DNA was prepared as described previously (36).

Sequence analysis was performed using the University of Wisconsin Genetics Computer Group programs (37), the BLAST and FASTA searches at the Sanger Centre (http://www.sanger.ac.uk/ ) and BLAST2 and PSI_BLAST searches at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/ ).

N-terminal sequence analysis of the phage coat proteins was performed as described previously (38). Approximately 100 µg of phage coat proteins were loaded onto a 9% SDS–polyacrylamide gel, blotted onto polyvinylidene difluoride (PVDF) paper and subjected to Edman degradation on an Applied Biosystems 473A protein sequencing machine.

## RESULTS AND DISCUSSION

### General features

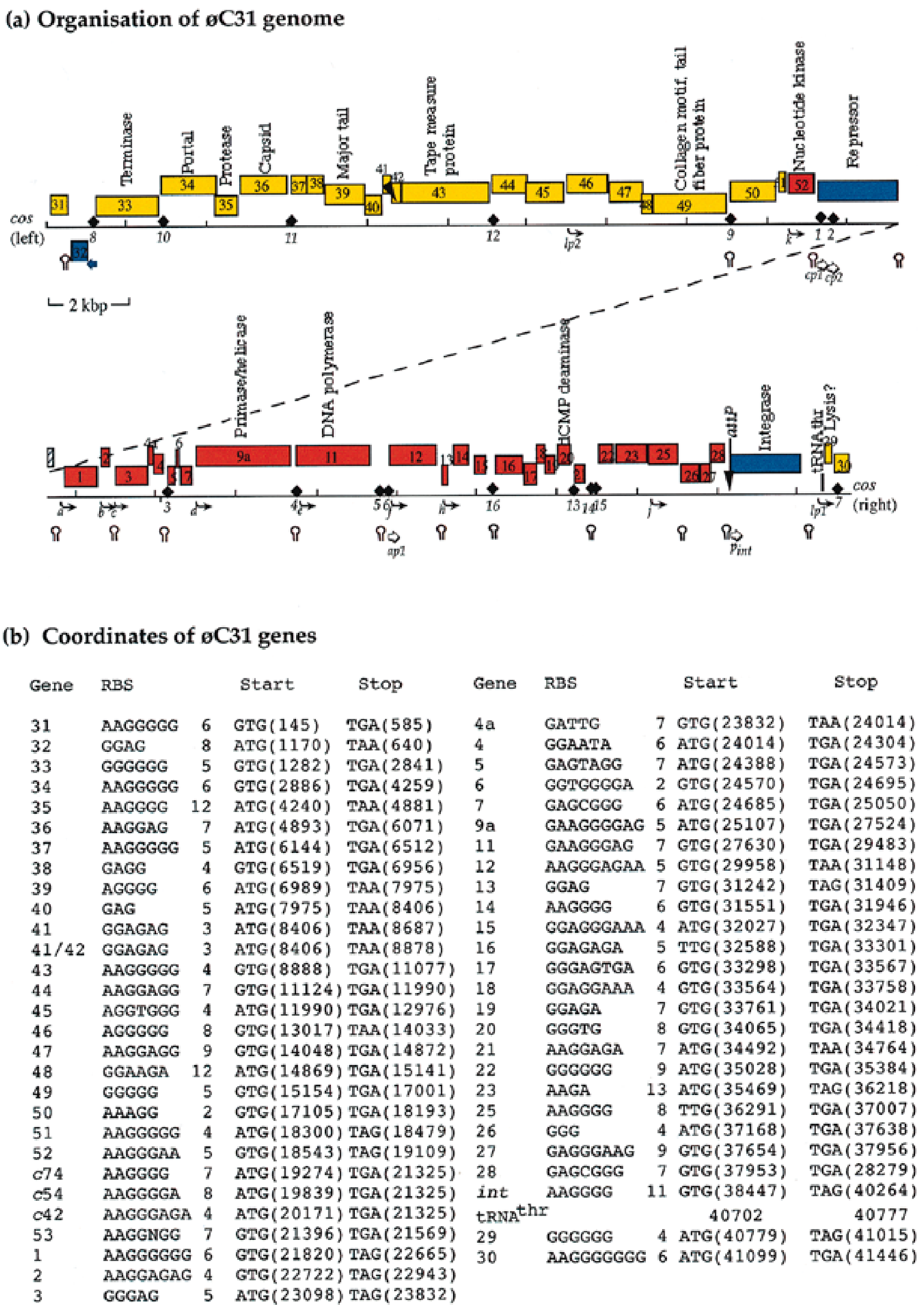
The sequences of the late and the inessential early regions of ϕC31 were determined. These were incorporated into a contiguous sequence with the previously published essential early region (7), integrase gene and phage attachment site (8,9), the repressor region (6) and part of the late region including the *cos* ends (10). The completed sequence was 41 491 bp in length. The average G+C content of the late region (coordinates 41320–18512) was 63.1%, the inessential early region (coordinates 31952–38340) was 64.9% and overall was 63.6%, which is in good agreement with the previously published sequence (63.8%).

### Analysis of putative gene products

Analysis of the sequence with CODONPREFERENCE using a CODONFREQUENCY table calculated from the published genes from ϕC31 revealed a total of 54 genes in the ϕC31 genome (Fig. 1). The individual gene G+C compositions varied from 59.6% (gene 37) to 69% (gene 15). The G+C content of the third position of the codons of individual genes varied from 66.4% (gene 30) to 85.1% (gene 2) and was, on average, 75.1%. Overall, there was no evidence for significant differences in G+C usage in the early versus late regions. During the sequence analysis the endpoint of gp9 and the start of gp10 of the sequence of Hartley *et al*. (7) was found to fall in the middle of a putative helicase domain, suggestive of a sequence error. Using custom primers the endpoints between genes 8, 9 and 10 were re-sequenced; additional Gs were detected at positions 6840 and 8376 of the sequence of Hartley *et al*. (7). The final sequence predicts a single gene, which we call gene 9a, encompassing genes 8, 9 and 10 from Hartley *et al*. (7). The relationships between gp9a and other similar sequences in the database are discussed below.

Confirming the findings of previous workers (6–10), none of the ϕC31 genes contain the rare UUA leucine codon, which along with its cognate $\text{tRNA}_{\text{UUA}}$, control the expression of genes involved in sporulation or secondary metabolism (39). As ϕC31 reproduces best in young, rapidly growing mycelia before the onset of antibiotic or aerial mycelia production, the absence of any UUA codons is as expected (40). The genes are preceded in each case by a ribosome binding site (with the exception of $\text{tRNA}^{\text{Thr}}$) and initiate with ATG or GTG with approximately equal frequency (Fig. 1). Two early genes in the inessential region initiate with TTG (Fig. 1).

All of the genes are expressed from left to right with the exception of gene 32. Previous work demonstrated that gene 32 is expressed in lysogens at a low level via a transcript of ~600 nt (10). An inverted repeat is located immediately downstream of gene 32 and it is likely that this acts as a terminator of transcription (10). This arrangement, i.e. a gene expressed in lysogens but located within a lytic operon and transcribed in the opposite



(a) Organisation of øC31 genome

(b) Coordinates of øC31 genes

| Gene | RBS | | Start | Stop | Gene | RBS | | Start | Stop |
|---|---|---|---|---|---|---|---|---|---|
| 31 | AAGGGGG | 6 | GTG(145) | TGA(585) | 4a | GATTG | 7 | GTG(23832) | TAA(24014) |
| 32 | GGAG | 8 | ATG(1170) | TAA(640) | 4 | GGAATA | 6 | ATG(24014) | TGA(24304) |
| 33 | GGGGGG | 5 | GTG(1282) | TGA(2841) | 5 | GAGTAGG | 7 | ATG(24388) | TGA(24573) |
| 34 | AAGGGGG | 6 | GTG(2886) | TGA(4259) | 6 | GGTGGGGA | 2 | GTG(24570) | TGA(24695) |
| 35 | AAGGGG | 12 | ATG(4240) | TAA(4881) | 7 | GAGCGGG | 6 | ATG(24685) | TGA(25050) |
| 36 | AAGGAG | 7 | ATG(4893) | TGA(6071) | 9a | GAAGGGGAG | 5 | ATG(25107) | TGA(27524) |
| 37 | AAGGGGG | 5 | ATG(6144) | TGA(6512) | 11 | GAAGGGAG | 7 | GTG(27630) | TGA(29483) |
| 38 | GAGG | 4 | GTG(6519) | TGA(6956) | 12 | AAGGGAGAA | 5 | GTG(29958) | TAA(31148) |
| 39 | AGGGG | 6 | ATG(6989) | TAA(7975) | 13 | GGAG | 7 | GTG(31242) | TAG(31409) |
| 40 | GAG | 5 | ATG(7975) | TAA(8406) | 14 | AAGGGG | 6 | GTG(31551) | TGA(31946) |
| 41 | GGAGAG | 3 | ATG(8406) | TAA(8687) | 15 | GGAGGGAAA | 4 | ATG(32027) | TGA(32347) |
| 41/42 | GGAGAG | 3 | ATG(8406) | TAA(8878) | 16 | GGAGAGA | 5 | TTG(32588) | TGA(33301) |
| 43 | AAGGGGG | 4 | GTG(8888) | TGA(11077) | 17 | GGGAGTGA | 6 | GTG(33298) | TGA(33567) |
| 44 | AAGGAGG | 7 | GTG(11124) | TGA(11990) | 18 | GGAGGAAA | 4 | GTG(33564) | TGA(33758) |
| 45 | AGGTGGG | 4 | ATG(11990) | TGA(12976) | 19 | GGAGA | 7 | GTG(33761) | TGA(34021) |
| 46 | AGGGGG | 8 | GTG(13017) | TAA(14033) | 20 | GGGTG | 8 | GTG(34065) | TGA(34418) |
| 47 | AAGGAGG | 9 | GTG(14048) | TGA(14872) | 21 | AAGGAGA | 7 | ATG(34492) | TAA(34764) |
| 48 | GGAAGA | 12 | ATG(14869) | TGA(15141) | 22 | GGGGGG | 9 | ATG(35028) | TGA(35384) |
| 49 | GGGGG | 5 | GTG(15154) | TGA(17001) | 23 | AAGA | 13 | ATG(35469) | TAG(36218) |
| 50 | AAAGG | 2 | GTG(17105) | TGA(18193) | 25 | AAGGGG | 8 | TTG(36291) | TGA(37007) |
| 51 | AAGGGGG | 4 | ATG(18300) | TAG(18479) | 26 | GGG | 4 | ATG(37168) | TGA(37638) |
| 52 | AAGGGAA | 5 | GTG(18543) | TAG(19109) | 27 | GAGGGAAG | 9 | GTG(37654) | TGA(37956) |
| *c*74 | AAGGGG | 7 | ATG(19274) | TGA(21325) | 28 | GAGCGGG | 7 | GTG(37953) | TGA(28279) |
| *c*54 | AAGGGGA | 8 | ATG(19839) | TGA(21325) | *int* | AAGGGG | 11 | GTG(38447) | TAG(40264) |
| *c*42 | AAGGGAGA | 4 | ATG(20171) | TGA(21325) | tRNA$^{thr}$ | | | 40702 | 40777 |
| 53 | AAGGNGG | 7 | GTG(21396) | TGA(21569) | 29 | GGGGGG | 4 | ATG(40779) | TAG(41015) |
| 1 | AAGGGGGG | 6 | GTG(21820) | TAG(22665) | 30 | AAGGGGGGG | 6 | ATG(41099) | TGA(41446) |
| 2 | AAGGAGAG | 4 | GTG(22722) | TAG(22943) | | | | | |
| 3 | GGGAG | 5 | ATG(23098) | TAG(23832) | | | | | |

**Figure 1.** Genes and regulatory elements encoded by ϕC31. (**a**) Organisation of the ϕC31 genome: the ϕC31 genome is represented as a line broken just after the repressor gene for convenience. The ends of the genome are labelled as *cos* (left) and *cos* (right). The genes are numbered and represented by coloured boxes. The early genes are coloured orange, the late genes are yellow and those known to be expressed in lysogens are blue. The single tRNA gene is shown as a black bar and is transcribed late. All the genes above the line are transcribed left to right and gene 32, located below the line is read right to left. The tailless arrow between genes 41 and 42 represents the proposed translational frameshift to express a gene 41/42 fusion protein. The regulatory signals in ϕC31 are shown. The black diamonds represent the repressor binding sites or CIRs (see text) and are numbered 1–16. The phage-specific lytic promoters are represented by curved arrows and are labelled *a*–*k* for the early promoters and lp1 and lp2 for the late promoters. The immediate early promoters (recognised by host RNA polymerase) are shown by open arrows and are labelled, except for the proposed immediately-early promoter upstream from gene 32 which is coloured blue. The transcription terminators are shown by the stem–loop icons, where those that are known to be functional (10,26–28,30) are coloured pink and those that are proposed terminators are black. The functions of some of the proteins are shown. (**b**) Coordinates of ϕC31 genes: the sequence coordinates of the ϕC31 genes, the start and stop codons, and the sequences of the ribosome binding sites are shown. The accession number for the ϕC31 genome sequence is AJ006589.

**Table 1.** Table of similarities

| Gene | Amino acids | Mol. Wt. (kDa) | Similar proteins | E value (BLAST2) and % identity | Nature of protein/probablefunction |
|---|---|---|---|---|---|
| **Early genes** | | | | | |
| Gp52 | 189 | 20.9 | Gp76 HSVI1, unknown | 5e-06(30% in 133aa) | Deoxynucleotide monophosphate kinase. |
| | | | Gp77 HSVI1, unknown | 6e-06 (31% in 106aa) | |
| | | | Coliphage T4 dexynucleotide kinase | 4e-05 (32% in 102aa) | |
| Gp1 | 282 | 31.6 | ORF1, *Arthrobacter* phage øAAU | 2.4e-44 (43% in 239aa) | |
| Gp2 | 74 | 7.7 | | | |
| Gp3 | 245 | 27.4 | | | |
| Gp4a | 60 | 6.7 | | | |
| Gp4 | 97 | 10.6 | | | |
| Gp5 | 62 | 6.6 | | | |
| Gp6 | 42 | 4.2 | | | |
| Gp7 | 122 | 13.9 | *M. tuberculosis*, Rv2469c, unknown | 5e-09 (43% in 69aa) | Large family of zinc finger proteins, possibly with endonuclease activity (61) |
| | | | *Synechocystis*, unknown | 5e-08 (30% in 107aa) | |
| | | | *M. tuberculosis*, Rv3074, unknown | 3e-06 (50% in 42aa) | |
| | | | ORF4, *Lactococcus* phage ϕ31 | 9e-05 (33% in 102aa) | |
| Gp9a | 805 | 87.1 | ORF11, *Bacillus* phage ϕ105 | 2e-72(33%in 456aa) | Primase/helicase. Contains N-terminal zinc finger domain. Proposed alternative translation start at amino acid Met86. |
| | | | *M. tuberculosis*, ϕRv1 | 2e-50(29% in 411aa) | |
| | | | *S. coelicolor*, ϕSc2E1 | 2e-48 (30% in 388aa) | |
| | | | Mycobacteriophage TM4 gp70 | 3e-41 (29% in 441aa) | |
| | | | 103R, Chilo iridescent virus | 1e-16 (24% in 338aa) | |
| | | | Primase/helicase, coliphage P4 | 5e-11 (22% in 532aa) | |
| | | | Phage R73 | 6e-11 (22% in 532aa) | |
| | | | ORF382, *Streptococcus thermophilus* phage, Sfi21. | 4e-08 (20% in 295aa) | |
| | | | ORF904, *Sulfolobus islandicus* pRN1. | 1e-06 (23% in 239aa) | |
| | | | MC094R, Molluscum contagiosum virus | 1e-06 (27% in 323aa | |
| | | | ORF2, *Strep. thermophilus* phage, Sfi18 | 2e-06 (21% in 232aa) | |
| | | | pC9262R, Africa swine fever virus | 4e-05 (24% in 223aa) | |
| | | | C5, Rabbit fibroma virus | 2e-04 (25% in 240aa) | |
| Gp11 | 618 | 68.5 | DNA polymerase, mycobacteriophage D29 | e-108 (38% in 624aa) | DNA polymerase |
| | | | Plus 31 others (<2e-04), all to DNA polymerases. | | |
| Gp12 | 397 | 42.7 | | | |
| Gp13 | 56 | 5.9 | | | |
| Gp14 | 132 | 14.6 | | | |
| Gp15 | 107 | 11.7 | | | |
| Gp16 | 238 | 26.0 | Gp48, mycobacteriophage L5. | 8e-53 (48% in 245aa) | Highly conserved protein |
| | | | *M. tuberculosis*, Rv2754c, unknown | 5e-39 (48% in 198aa) | |
| | | | *Corynebacterium glutamicum*, DAPB-DAP intergenic region. | 2e-34 (41% in 218aa) | |
| | | | *Pyrococcus horikoshii,* unknown. | 3e-14 (32% in 172aa) | |
| | | | *Helicobacter pylori,* unknown. | 5e-05 (28% in 184aa) | |
| | | | *Paramecium bursaria* Chlorella virus (PBCV1) unknown. | 2e-04 (27% in 154aa) | |
| Gp17 | 90 | 10.1 | | | |
| Gp18 | 65 | 6.8 | | | Possible secreted protein. |
| Gp19 | 89 | 9.5 | | | |
| Gp20 | 118 | 12.0 | Gp36.1 mycobacteriophage D29 | 6e-24 (54% in 117aa) | dCMP deaminase |
| | | | Plus 31 others (<4e-05) all to dCMP deaminases. | | |
| Gp21 | 91 | 10.0 | | | |
| Gp22 | 119 | 12.8 | | | |
| Gp23 | 250 | 25.5 | | | Proline rich, possible secreted protein. |
| Gp25 | 239 | 25.7 | SpdB2 from pJV1,*Streptomyces phaeochromogenes*. | 7e-18 (34% in 119aa) | Proline rich C-terminal domain, possible membrane protein. |
| Gp26 | 157 | 16.9 | *M. tuberculosis*, Rv0494, unknown | 2e-05 (42% in 69aa) | |
| | | | FadRV, *Vibrio alginolyticus* | 4e-05 (39% in 66aa) | Member of the GntR family of repressors. |
| | | | Plus 7 others (<8e-04), all transcriptional regulators. | | |
| Gp27 | 101 | 11.5 | øC31 Gp28 | 38% identical | |
| | | | *M. tuberculosis* FadE7 | 0.16 (27% in 88aa) | |
| Gp28 | 109 | 12.7 | øC31Gp27 | 38% identical | |
| **Late genes** | | | | | |
| $tRNA^{Thr}$ | - | - | *Methanococcus vannielii* $tRNA^{Thr}$ | 8e-04 (88% in 43 bp) | $tRNA^{Thr}$ |
| Gp29 | 8.7 | | | | Putative holin (10) |
| Gp30 | 116 | 12.8 | Intron contained hypothetical protein Calothrix sp. | 4e-04 (43% in 48aa) | Large family of Zinc finger proteins, possibly with endonuclease activity (61). |
| | | | Gp7 øC31 | 0.5 (30% in 62aa) | |
| | | | *Rhodobacter capsulatus* øRcP1 protein | 2.5 (26% in 108aa) | |
| Gp31 | 147 | 16.0 | *M. tuberculosis,* øRv2 | 0.008 (23% in 116aa) | |
| Gp33 | 520 | 57.7 | Gp2, terminase large subunit coliphage HK97. | 28% identical | Terminase, large subunit. |
| | | | *Rhodobacter capsulatus* øRcP1 protein | 7e-32 (26% in 492aa) | |
| | | | ORF5, *Lactobacillus casei* phage A2 | 2e-22 (26% in 500aa) | |
| | | | Gp13 mycobacteriophage D29 | 2e-15 (32% in 213aa) | |
| | | | Gp13 mycobacteriophage L5 | 4e-15 (35% in 177aa) | |
| | | | Plus 7 others, all putative terminases. | <0.77 | |
| Gp34 | 458 | 49.8 | Gp3, portal protein, coliphage HK97. | 3e-36 (29% in 375aa) | Portal protein |
| | | | *Staphylococcus* phage øPVL putative portal | 2e-14 (23% in 364aa) | |
| | | | *Rhodobacter capsulatus* øRcP1 protein | 6e-10 (22% in 359aa) | |

**Table 1.** *Continued*

| | | | | | |
|---|---|---|---|---|---|
| Gp35 | 214 | 23.5 | Gp4, protease coliphage HK97.<br>ORFs5/6 *Staphylococcus* phage øPVL | 28% identical<br>33% identical | Protease |
| Gp36 | 393 | 41.7 | *R. capsulatus*, ϕRcM1 RRC01383<br>*M. tuberculosis*, øRv2<br>*M. tuberculosis*, øRv1<br>Gp5, coliphage HK97 major capsid | 9e-20 (29% in 326aa)<br>0.016 (23% in 190aa)<br>0.081 (23% in 219aa)<br>19.5% identical | Major capsid protein |
| Gp37 | 123 | 13.8 | | | |
| Gp38 | 146 | 15.4 | | | |
| Gp39 | 329 | 34.8 | | | Major tail protein |
| Gp40 | 144 | 16.0 | | | |
| Gp41 | 94 | 10.3 | | | |
| Gp41/2 | 158 | 17.7 | | | Putative frameshift event between genes 41 and 42. |
| Gp43 | 730 | 75.4 | Hypothetical protein in *B. subtilis* SpoIIIC-CwlA intergenic region; *skin* prophage.<br>ORF15 *Staphylococcus* phage øPVL.<br>*Strep. thermophilus* phage Sfi19<br>Gp14, tail tape measure protein, coliphage HK97.<br>Plus 4 others, all putative minor tail proteins. | 9e-18 (28% in 241aa)<br>5e-07 (23% in 350aa)<br>5e-06 (24% in 280aa)<br>23% identity.<br><0.100 | Tail tape measure protein. |
| Gp44 | 289 | 31.4 | | | |
| Gp45 | 329 | 34.3 | | | |
| Gp46 | 339 | 35.9 | | | |
| Gp47 | 275 | 29.9 | | | |
| Gp48 | 91 | 10.0 | | | |
| Gp49 | 616 | 63.0 | *Ephydatia muelleri* (sponge) short chain collagen.<br>Plus >100 other, all containing collagen repeats. | 1.3e-13 (32% in 237aa)<br><9e-09 | Tail fibre protein. |
| Gp50 | 363 | 38.3 | *Bacillus licheniformis* N-acetylmuramoyl-L-alanine amidase (autolysin). | 0.31 (29% in 168aa) | Cell wall hydrolase |
| Gp51 | 60 | 5.9 | | | |
| **Others** | | | | | |
| Gp32 | 177 | 20.0 | | | Expressed in lysogens. |
| Rep | 683 | 74.0 | | | Repressor isoforms. |
| Gp53 | 57 | 5.7 | | | |
| Int | 606 | 67.0 | *B. subtilis* bacteriophage SPBc2 | 2e-05 (24% in 193aa) | Site-specific recombinase of the resolvase family. |

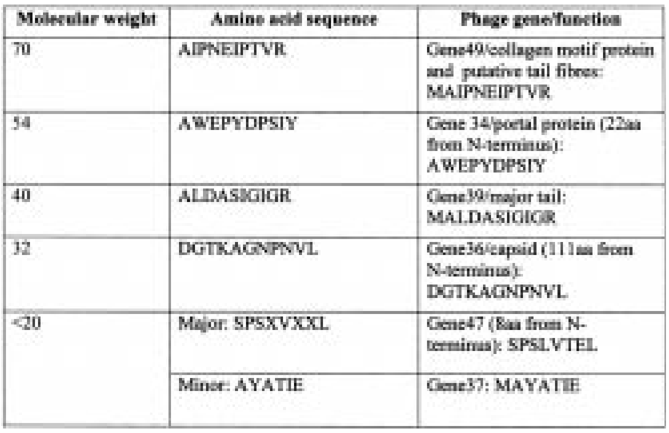
direction, is highly unusual in phage genomes, potentially disrupting expression of, in this case, the late proteins. It seems most likely that gene 32, its promoter and the terminator have been inserted by an illegitimate recombination event forming a completely self-contained, mono-cistronic operon. Comparisons of genomic sequences of closely related phages such as the lambdoid phages, the mycobacteriophages or the streptococcal phages have shown that insertions or deletions involving complete genes occurs frequently (14,19,21,22). What is remarkable is the precision of the insertion or deletion and any honing to include the minimum amount of flanking DNA sequence. Why is such an insertion in the late operon of ϕC31 tolerated? Presumably its expression during lysogeny may be of such a selective advantage that its persistence in this unusual position is permitted. As the function of gp32 is not known and has no clear homologues in the databases, the selective advantage incurred can only be speculated upon. Some possibilities could include phage exclusion or blockage of ϕC31 receptors so as to avoid inactivation of progeny phages after induction.

BLAST2, FASTA and PSI-BLAST searches of the protein databases with the predicted amino acid sequences from the phage genes (omitting the tRNA$^{Thr}$ gene) revealed that whilst most (33/54) of the gene products do not have any homologues in the database, 21 do, frequently to proteins encoded by phages that infect evolutionarily diverse bacteria including *Escherichia*, *Arthrobacter*, *Mycobacterium*, *Rhodobacter*, *Staphylococcus*, *Bacillus*, *Streptococcus* and *Lactobacillus* (Table 1). The genetic similarities between the diverse phages and other relationships, notably to the dsDNA viruses and other cellular genes, are discussed in the following sections.

## Diverse phages grouped together by a common module for capsid assembly

N-terminal sequence analysis of the most abundant ϕC31 structural protein, most likely the major capsid protein, indicated that it was encoded by gene 36 (41) (Table 2). Comparison of the predicted amino acid sequence of gp36 and the mature protein indicated that the former has an additional 111 amino acids at the N-terminus, suggesting that the primary translation product was processed. This cleavage is reminiscent of a similar event during the assembly process in coliphage HK97, which has been studied in some detail (42). HK97 proheads are assembled from pentamers and hexamers of gp5 (capsid). HK97 is unusual amongst phages as it does not require a separate scaffold protein for assembly and it is speculated that the 102 N-terminal amino acids that are later cleaved off from gp5 substitutes as a scaffold (42,43). Upstream of gene 5 is a gene (gene 4) encoding a protease, which is responsible for the processing (44). Also important in the formation of HK97 proheads is the portal protein (encoded by gene 3), which forms an aperture through which the DNA passes during the packaging process (42). The overall assembly process follows an ordered pathway of covalent and conformational changes, largely determined by the activities of the major capsid protein, eventually to form mature phage heads (42).

**Table 2.** N-terminal sequences of ϕC31 structural proteins



| Molecular weight | Amino acid sequence | Phage gene/function |
|---|---|---|
| 70 | AIPNEIPTVR | Gene49/collagen motif protein and putative tail fibres: MAIPNEIPTVR |
| 54 | AWEPYDPSIY | Gene 34/portal protein (22aa from N-terminus): AWEPYDPSIY |
| 40 | ALDASIGIGR | Gene39/major tail: MALDASIGIGR |
| 32 | DGTKAGNPNVL | Gene36/capsid (111aa from N-terminus): DGTKAGNPNVL |
| <20 | Major: SPSXVXXL | Gene47 (8aa from N-terminus): SPSLVTEL |
|  | Minor: AYATIE | Gene37: MAYATIE |

Database searches using genes from the ϕC31 late cluster showed that the head gene organisation and predicted amino acid sequences were similar to the head assembly genes of coliphage HK97, staphylococcal phage ϕPVL and two apparent prophages from the *Rhodobacter capsulatus* genome sequence (ϕRcM1 and ϕRcP1; Fig. 2). (The nomenclature for the *Rhodobacter* prophages identifies the contigs on which they are encoded, i.e. contigs M1 and P1.) ϕC31 gp34 is most similar to RRC01381 from ϕRcM1 and to gp3 from HK97, which encodes the portal protein. Gp34 was shown by N-terminal sequence analysis to be present in ϕC31 particles and is, like the major capsid protein, processed (Table 2). It is worth noting here that the portal protein of phage λ is also processed 22 amino acids from the N-terminus (45). The C-terminal domain of ϕC31 gp34 differs from that of the other portal homologues described here (Fig. 2) as it is unusually rich in proline and acidic residues; the function of this domain is not known. ϕC31 gp35 was aligned in a BLAST2 search to both open reading frames (ORFs) 5 and 6 from ϕPVL. If a single base change is introduced in the TAA termination codon of gene 5 of ϕPVL, then a single ORF can be generated which is 33% identical to the whole length of ϕC31 gp35 (Fig. 2). The HK97 homologue of ϕC31 gp35 (HK97 gene 4), is the protease that processes the HK97 major capsid protein. We therefore believe that gp35 and ORF5/6 are both proteases that process the capsid proteins from ϕC31 and ϕPVL, respectively. Similarly, protease homologues were observed in ϕRcM1, and in two *Mycobacterium tuberculosis* prophages, ϕRv2 and ϕRv1 (15,23). In ϕRcM1 the protease domain is fused to a long C-terminal sequence that has 29% identity over 326 amino acids with the ϕC31 gp36 (28% identity overall). Generally the major capsid proteins have less conservation than the proteases and portal proteins (Fig. 2). Taken together these observations strongly suggest that the portal, protease and capsid proteins encoded by ϕC31 and ϕPVL, and the prophages (where non-defective) assemble to form phage heads using the same mechanism as that for HK97. Whilst similar organisation of late genes has been observed between diverse phages before (13,24,46,47), the sequence similarities between the capsid assembly proteins described here are indicative of a conserved functional module derived from a common ancestor. The question arises of how these HK97-like capsid assembly modules from diverse phages relate in an evolutionary sense to non-homologous capsid assembly genes from other phages. Could capsid assembly have arisen more than once during evolution or has selection operated differently in phages that have been dealt a different combination of assembly genes, or has the rate of horizontal exchange of the HK97-like module just been particularly rapid? Sequencing of more phage genomes will surely provide answers to some of these questions.

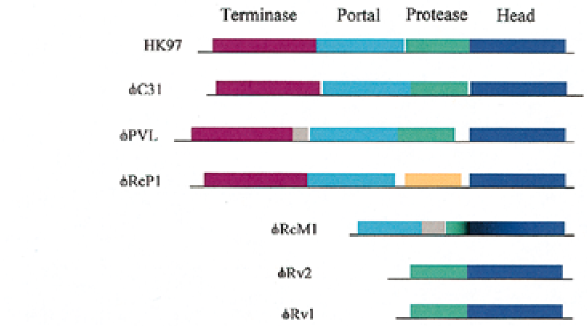
Even within the head assembly modules described here, there is also evidence for shuffling of the individual genes between phages by horizontal exchange. Whilst the two mycobacterial prophages' protease/head proteins are extremely similar (>88% identical), the portal/capsid proteins from the two *Rhodobacter* prophages are no more similar to each other than to the other homologues shown in Figure 2. Indeed, the most similar portal protein, protease and capsid proteins to those of the *Rhodobacter* prophage ϕRcM1 are from the *Streptomyces* phage ϕC31, coliphage HK97 and ϕC31, respectively. A similar patchwork of relatedness is evident from the presence in ϕRcP1 of a λ/N15-family protease rather than the HK97-like protease. This inability to discern which phage genome is most closely related to which is highly suggestive of horizontal exchange of genetic material between phages. It is also of interest that the prophages in the *M.tuberculosis* genome have capsid and protease homologues of those of the HK97 group but no portal or terminase homologues are located next to them. It would seem likely that these mycobacterial prophage genomes, whilst apparently incomplete, can still play a part in phage evolution as a resource for horizontal exchange. Indeed it is possible that some defective prophage genomes may actually be accidentally taking up interlopers, originally derived from phages that normally infect a different genus from that in which they now reside. Exchange between viable phages and these 'foreign' prophage genomes could provide one mechanism for horizontal transfer of DNA between phages that infect different genera.

Another unusual feature of HK97 capsid assembly is the formation of crosslinks within the capsid protein to link subunits together like a kind of chain-mail (48). The amino acid residues in the HK97 gp5 protein that form the crosslinks are K169 and N356 and the end result of the covalent modification can be observed by polyacrylamide gel electrophoresis as the appearance of very high molecular weight protein bands (48). Using this assay, chain-mail has not been observed in ϕC31 heads (41). Furthermore none of the capsid protein sequences related to HK97 gp5 have candidate lysine or asparagine residues at or near the equivalent positions to those in the HK97 sequence.

## Tail assembly in ϕC31

N-terminal sequence analysis of the second most abundant protein in ϕC31 particles corresponded to the product of gene 39 (41) (Table 2). Also present in phage particles, running as a single broad band with a mobility of <20 kDa, were two late proteins corresponding to gp37 (13.8 kDa) and gp47 (30 kDa). Clearly, the expected and observed molecular weights of these proteins were not in agreement, suggesting that gp47 is processed and gp37 may have an unusual mobility. We expect both of these proteins, however, to be tail or tail-fibre proteins. Searches of the databases using BLAST2 and PSI-BLAST with genes flanking gene 39 revealed that gene 43 was similar to many minor tail proteins. The position of gene 43 in the ϕC31 genome compared with the organisation and known functions of other phage tail genes, and the size of the predicted product strongly suggested that gp43 is a tail length determination protein with a role analogous to that of

(a) Organisation of head assembly genes.



(b) Sequence identities between portal proteins

| | ϕC31 | ϕRcM1 | HK97 | ϕRcP1 | ϕPVL |
|---|---|---|---|---|---|
| ϕC31 | 100 | | | | |
| ϕRcM1 | 31 | 100 | | | |
| HK97 | 29 | 29 | 100 | | |
| ϕRcP1 | 24 | 26 | 23 | 100 | |
| ϕPVL | 23 | 22 | 21 | 19 | 100 |

(c) Sequence identities between head proteases

| | ϕC31 | ϕPVL | ϕRv2 | ϕRv1 | ϕRcM1 | HK97 |
|---|---|---|---|---|---|---|
| ϕC31 | 100 | | | | | |
| ϕPVL | 33 | 100 | | | | |
| ϕRv2 | 28 | 20 | 100 | | | |
| ϕRv1 | 31 | 22 | 89 | 100 | | |
| ϕRcM1 | 28 | 21 | 30 | 31 | 100 | |
| HK97 | 28 | 25 | 29 | 25 | 38 | 100 |

(d) Sequence identities between capsid proteins

| | ϕC31 | ϕRv2 | ϕRv1 | ϕRcM1 | ϕRcP1 | HK97 | ϕPVL |
|---|---|---|---|---|---|---|---|
| ϕC31 | 100 | | | | | | |
| ϕRv2 | 23 | 100 | | | | | |
| ϕRv1 | 25 | 88 | 100 | | | | |
| ϕRcM1 | 28 | 23 | 24 | 100 | | | |
| ϕRcP1 | 25 | 20 | 22 | 23 | 100 | | |
| HK97 | 20 | 23 | 21 | 21 | 19 | 100 | |
| ϕPVL | 19 | 17 | 19 | 21 | 16 | 19 | 100 |

**Figure 2.** Comparisons of head assembly proteins from diverse phages and prophages. (**a**) Organisation of head assembly genes in coliphage HK97, *Streptomyces* phage ϕC31, the staphylococcal phage ϕPVL, the *Rhodobacter* prophages, ϕRcP1 and ϕRcM1, and the *M.tuberculosis* phages ϕRv1 and ϕRv2. Except for the grey blocks, which are proteins of unknown function and lack relatives in the databases, the coloured blocks represent related proteins. The major head proteins are shown in dark blue, the head proteases in green, the portal proteins in light blue and the terminases in pink. Shown in yellow in ϕRcP1 is a protease that has similarity to the λ/N15 family of proteases and is not related to the other proteases shown here. ϕRcM1 head and proteases are fused to form a single protein represented by the shadowing between the green (protease) and the blue (head) blocks. The accession numbers for the sequences aligned here and in (b) to (d) below are AF068845 (HK97), AB009866 (ϕPVL), Z80225 (ϕRv2) and Z95586 (ϕRv1). The *R.capsulatus* prophages ϕRcP1 and ϕRcM1 are located on contigs P1 and M1, respectively, and can be accessed at http://capsulapedia.uchicago.edu . ϕRcP1 can also be accessed by accession no. AF010496. (**b**) Sequence identities between the portal protein homologues. % identities between the portal homologues are shown, calculated using BESTFIT from the GCG package (37). The proteins can be identified as gp3 (HK97), gp34 (ϕC31), ORF4 (ϕPVL), RRC00515 (ϕRcP1) and RRC01381 (ϕRcM1). (**c**) Sequence identities between the head protease homologues. % identities between the protease homologues are shown, calculated using BESTFIT. The proteins can be identified as gp4 (HK97), gp35 (ϕC31), ORFs5/6 (ϕPVL; see text), RRC01383 (ϕRcM1), Rv2651c (ϕRv2) and Rv1577c (ϕRv1). (**d**) Sequence identities between the major capsid homologues. % identities between the capsid homologues are shown, calculated using BESTFIT. The proteins can be identified as gp5 (HK97), gp36 (ϕC31), ORF7 (ϕPVL; see text), RRC00514 (ϕRcP1), RRC01383 (ϕRcM1), Rv2650c (ϕRv2) and Rv1576c (ϕRv1).

the phage λ H protein (49). The products of the two small ORFs just upstream of gene 43, genes 41 and 42, lack homology with any proteins in the database. However, mainly on the basis of the positions of genes 41 and 42, we believe that they are the ϕC31 versions of tail assembly proteins from several other tailed dsDNA phages including the products of genes G and T in phage λ and genes 11 and 12 in HK97 (50; R.Hendrix, personal communication). The unusual feature of G and T in λ is that a translational frameshift can occur to generate a G-T fusion protein (49). The translationally slippery sequence in λ is GGGAAAG, and similar sequences have been observed in many other putative G-T homologues (e.g. HK97, D29 and L5, P2; R.Hendrix, G.Hatfull and G.Christie, personal communication). In ϕC31 gene 42 does not have a good start codon or ribosome binding site, consistent with it being expressed via a translational frameshift within gene 41. Indeed within gene 41 there is a sequence, GAAGGGGAAGG, which could be analogous to frameshifting sequences from other phages. Overall the organisation of putative tail genes in ϕC31 resembles that in other phages.

Other ϕC31 structural proteins identified by N-terminal sequence analysis included the product of gene 49. Gp49 is remarkable as it contains GXY repeats that are typical of eukaryotic collagen. A growing number of phages have been found to contain tail fibre proteins that have collagen-like repeats (51,52). Gp49 is therefore most likely to be a tail fibre protein.

## Lysis proteins

Most dsDNA phages encode at least two proteins required for cell lysis; the holin protein, which forms holes in the cell membrane through which a cell wall hydrolase passes after a defined time interval and breaks down the cell wall, causing lysis (53). The temporal control on lysis is critical: too soon and there are reduced numbers of mature phages; too late and the phage infection may be outrun by competing phages. Previously we identified a putative holin protein, gp29, based on the presence of two putative membrane spanning domains (10). This holin was not typical of many of the holin-like proteins, which have a 'dual start' motif (i.e. one of two start codons are used, located usually two or three codons apart, in the same frame resulting in the synthesis of two N-terminally different isoforms from the same ORF). The two isoforms are thought to regulate the formation of the pore (53). The regulation of other holins, which do not contain the dual start motif, is still unclear. In the lambdoid phages the holin and the glycosidase are located adjacent to each other (53,19). Analysis of the ϕC31 late region gene products did not unambiguously reveal a cell wall degrading enzyme, although gp50 was weakly similar to a cell wall hydrolase from *Bacillus licheniformis* on a BLAST2 search (Table 1). Gene 50 is unlikely to be expressed as an early gene as there is no indication of a promoter located after gene 49 (a putative tail fibre gene and therefore probably a late expressed gene) and before gene 50 (Fig. 1 and see below). If gene 50 does encode a lysis protein then the organisation of the lysis genes in ϕC31, in which the proposed holin and hydrolase genes are located at opposite ends of the late operon, is significantly different from other phages in which the lysis genes tend to be grouped together. Possibly in ϕC31 the lysis event is controlled by delaying the expression of the hydrolase. It is notable that gene 50 lies downstream of the only recognisable terminator at the end of the late operon (Fig. 1). If this terminator is, like those of the early region, inefficient, gp50 will be expressed at a very low level, perhaps building up slowly to an effective level during the lytic cycle. Alternatively this terminator may protect gene 50 from any rogue transcription during lysogeny.

### Inessential genes: homologues to mycobacteriophage genes and adaptations to growth in *Streptomyces*?

The inessential region of ϕC31 was defined as that region that can be deleted to allow capacity for insertion of foreign DNA in the ϕC31 derived vectors, without impairing the lytic growth of the phage. Database searches with the putative gene products from this region reveal two genes with very high similarities to mycobacteriophage genes. ϕC31 gp16 is 49% identical with L5 gp48 (Table 1) (13). The function of this protein is not clear but homologues are also present in the chromosomes of several bacteria including *Corynebacteria glutamicum*, *M.tuberculosis*, *Pyrococcus horikoshii*, *Helicobacter pylori* and in *Paramecium bursaria Chlorella* virus (PBCV1) (Table 1). The locations of the homologues of gene 16 in the bacterial chromosomes are different, giving no real clues as to the function of these proteins. In *M.tuberculosis* and *C.glutamicum* the gene 16 homologues are located adjacent to *dapA*, a gene required for lysine biosynthesis.

ϕC31 gp20 is 54% identical to D29 gp36.1 and other proteins in the databases that have been identified as dCMP deaminases. This enzyme converts dCMP to dUMP, a precursor of dTTP (Table 1) (14). D29 and ϕC31 may encode these enzymes to boost the flux of metabolites to DNA synthesis, thereby ensuring the maximum burst size. Possibly the flux of DNA precursors provided by the host enzymes is not ideally suitable for the rapid DNA synthesis required during phage infection or, in the case of ϕC31, to meet the different G+C compositions of the phage DNA (63%) compared to its host (74%). Indeed ϕC31 carries another (inessential) gene whose product is, putatively, involved in nucleotide metabolism. This is gene 52, which encodes a putative nucleotide kinase. (It is also notable in view of the discussion below that gp52 has greatest resemblance to two HSVI1 enzymes.)

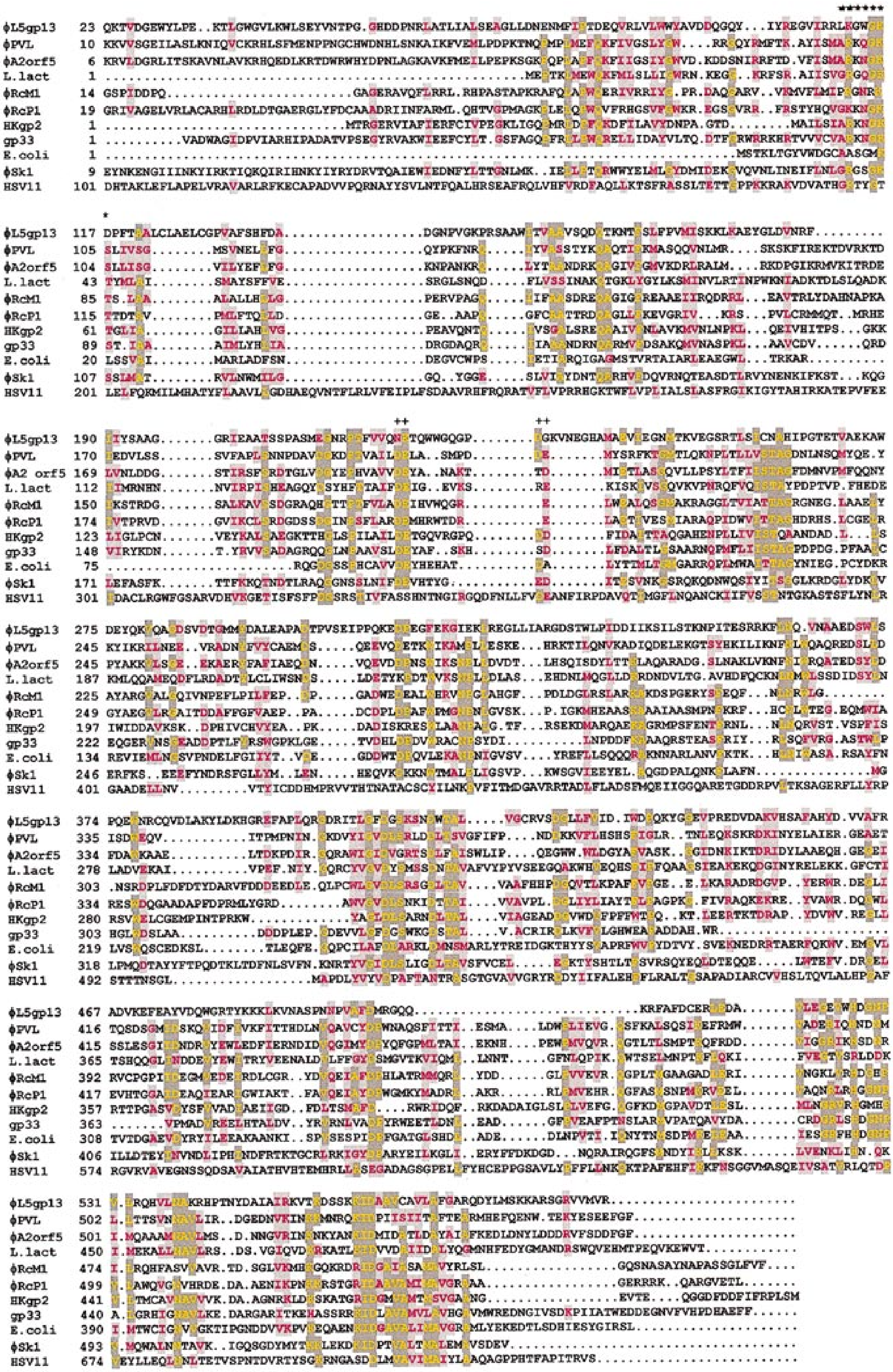
ϕC31 gp25 is most similar to a so-called ‘spread’ gene from a transferable *Streptomyces* plasmid pJV1 (Table 1) (54); the *spd* genes were identified as being required for the spread of plasmids after the initial transfer into a new host strain. Whilst the phenomenon of plasmid spread is poorly understood, it is intriguing that a homologue of one of these proteins is present in the ϕC31 genome. The presence of a lytic promoter upstream of gp25 strongly suggests that gp25 is expressed during the lytic cycle. One possibility is that gp25 allows the phage genome to travel between mycelial compartments during lytic growth. A Kyte–Doolittle hydropathy prediction of gp25 is consistent with the N-terminal half of this protein being embedded in the membrane. Furthermore, two other inessential proteins, gp18 and gp23, could be secreted as both are predicted to contain signal sequences.

### Phage proteins with homologues in eukaryotic viruses

ϕC31 gp9a aligns in a BLAST2 search with proteins encoded by the dsDNA viruses Chilo iridescent virus, Molluscum contagiosum virus, African swine fever virus and Rabbit fibroma virus (Table 1). Other hits on this search are all either phage or prophage-encoded proteins or encoded by archaeal plasmids (e.g. *Sulfolobus* pRN1). One of the proteins with similarity to ϕC31 gp9a is from the *S.coelicolor* sequence, cosmid 2E1 and, as the region encoding the gp9a homologue has other features consistent with it being within a prophage (including a homologue to the ϕC31 integrase gene), we propose to name it ϕSc2E1. The function of gp9a can be deduced from its similarity with a coliphage P4 protein, known to have primase and helicase activities (Table 1) (55) and because of the conserved residues (motifs A, B and C) that have previously been noted as being present in a virally-encoded superfamily of helicases, family SF3 (56). Like the P4 primase/helicase, ϕC31 gp9a also has a zinc finger domain at the N-terminus. The presence of a ribosome binding site (AAGGAGAAG) located 7 bp from an internal, in frame ATG (encoding residue 86) is suggestive of expression of a truncated gp9a which would lack the zinc finger protein. Several of the proteins aligned by the BLAST2 search are much shorter than gp9a and consist only of the helicase domain and flanking conserved sequences. Originally the helicase motifs identified by Gorbalenya *et al.* (56) were limited to the small DNA and RNA viruses and phage P4 primase/helicase was the only bacteriophage-encoded member of the family. They suggested that the SF3 helicase family have evolved from a common ancestor. The analysis shown here confirms that the SF3 helicases are frequently encoded by eukaryotic and prokaryotic viruses and by archaeal plasmids. Chromosomally-encoded helicases usually belong to one of the remaining four superfamilies (57).

Another phage-encoded protein that is related to proteins from eukaryotic viruses is gp33 encoding the putative large subunit of the terminase (Fig. 3). After six iterations of a PSI-BLAST search using the ϕC31 gp33, 16 phage or prophage terminases and seven homologues from bacterial genomes (also possibly from prophage sequences) were above the E-value threshold (0.001). In addition, four animal virus packaging proteins (terminases) above the E-value threshold were identified, and out of the first 20 hits below the threshold, 13 were putative terminases from eukaryotic viruses. Figure 3 is an alignment of nine putative phage or prophage-encoded terminases and a probable DNA packaging protein from HSVI1. Many of the highly conserved residues are also present in the HSVI1 homologue (Fig. 3). Terminase is a protein performing an exclusively viral function, namely transporting DNA into the proheads prior to attachment of the tail (45,58). Another exclusively viral protein, and one which interacts with the terminase, is the portal protein (45,59). However, none of the ϕC31, HK97 or ϕPVL portals pick out any putative portal protein homologues in the eukaryotic viruses even

**Figure 3.** (Opposite) Alignment of putative terminases from phages, prophages and a dsDNA virus. The alignment was performed using the PILEUP programme from the GCG package. The accession nos for the proteins shown in this alignment are X97563 (*Lactobacillus casei* phage A2; orf5), AF010496 (*R.capsulatus* prophage ϕRcP1; 3128374), Q05219 (mycobacteriophage L5; gp13), P75978 (*intE-pin* region in *Escherichia coli*), AJ223961 (*Lactococcus lactis* genome; e1323764), AB009866 (*Staphylococcus* phage ϕPVL; orf2), AF011378 (*L.lactis* phage ϕSk1; terminase large subunit), AF068845 (*E.coli* phage HK97;gp3) and P04295 (HSV11). The *R.capsulatus* prophage ϕRcM1 is located on contig M1 and can be accessed at http://capsulapedia.uchicago.edu and the putative terminase is encoded by RRC02334. The putative ϕC31 terminase is encoded by gp33. Marked by * or + above the alignment are putative nucleotide binding, Walker A and B motifs, respectively (62).


ϕL5gp13
ϕPVL
ϕA2orf5
L.lact
ϕRcM1
ϕRcP1
HKgp2
gp33
E.coli
ϕSk1
HSV11

after six iterations of PSI-BLAST. The portal protein from phage λ has even fewer homologues in the database i.e. N15 and P21 portal proteins. The terminase proteins and the gp9a-like primase/helicases (discussed above) have been conserved in viruses of both prokaryotes and eukaryotes (and in the case of the helicases, also in the archaeal plasmids). Ancestral proteins from both families could have been present before the split between the kingdoms, although it is also possible that horizontal transfer has produced the relationships noted here.

## Global regulation of ϕC31 genes

*Promoters and terminators*. Previous work analysing the transcription of ϕC31 during lytic growth showed that the right arm is mainly expressed early (25,26). Transcripts arising from a multiplicity of phage-specific promoters yield overlapping transcripts due to the inefficient nature of the intrinsic terminators (27,28). The left arm is mainly expressed late (25); one late promoter has been characterised on the extreme right end of the genome and transcription proceeds over the annealed *cos* ends (10,29). The phage specific promoters contain a highly conserved 21 bp sequence and analysis of the newly determined sequence revealed additional early and late promoters (Epj and lp2, respectively; Fig. 1). Lp2 has previously been identified as a fragment of DNA with promoter activity using a ϕC31-derived promoter probe vector (2; C.Bruton and K.F.Chater, personal communication). The sequence was also analysed for the presence of terminators and, again, additional terminators were found in each of the early and late regions (Fig. 1). The paucity of late promoters and the presence of only one potential terminator (upstream of gene 50) are in agreement with a previous proposal that the late region is expressed largely as a single operon during lytic growth (25). [The terminator downstream of gene 32 was shown to be non-functional in stopping transcription in the left to right direction during lytic transcription (10).] Processing of late transcripts has been proposed, in particular close to CIR8 (10,25). Lp2 could be required to boost the expression of the relatively abundant tail fibre proteins. The possible significance of the terminator in the late region located just upstream of gene 50 was discussed above.

## Repressor binding sites

The DNA sequence contains a total of 16 potential repressor binding sites (Fig. 1). These sites were located by searching for the conserved 17 bp sequence identified previously as the binding site for the 42 and 54 kDa repressor isoforms (31–33). The 17 bp sequence contains a conserved inverted repeat and the sites are therefore referred to as CIR sites (Fig. 1). Thirteen CIRs contain an identical 16 bp core sequence, the remaining three only differing from the core by 1 or 2 bp. Except for CIR2 which lies in the coding sequence of the repressor gene, the CIRs are located in intergenic regions. The locations of the CIR sites with respect to other transcription signals do not provide a strong indication of the precise roles of the CIR sites in the maintenance of lysogeny. Four CIRs lie near known or proposed immediate-early or early phage-specific promoters, six others lie near or overlap with factor independent terminators, and six are within intergenic sequences that appear to lack any transcriptional signals. The genomes of mycobacteriophages L5 and D29 also have multiple repressor binding sites and it is thought that, in the absence of any factor independent terminators, repressor bound to these sites prevents any rogue transcription during lysogeny (60). In ϕC31 there are many factor independent terminators which can presumably fulfil this role (Fig. 1).

The locations of the CIR sites in the late region are suggestively non-random. CIRs 10 and 11 appear to neatly flank the head assembly module of portal, protease and capsid genes whose relatives are found in diverse phage and prophages (see above). It is tempting to speculate that CIRs 11 and 12 flank the tail cluster and CIRs 12 and 9 flank the tail fibre cluster. As short conserved sequences such as terminators have previously been proposed to be sites of recombination between phages of the lambdoid family (19,61), the arrangement of CIRs in the late region could permit exchange of whole assembly modules between phages of a putative 'ϕC31 family'. Perhaps in the early region the multiple phage specific promoters serve a similar function. Additionally, as noted previously (27), the mechanisms of evolution of the regulatory elements in ϕC31 are not likely to involve genetic exchange of a complete regulatory module as occurs in the lambdoid family of phages.

## ACKNOWLEDGEMENTS

We would like to thank Ms Farzana Damji and Dr Brenda Leskiw for donation of part of the ϕC31 sequence from the late region and to Professor Keith Chater and Ms Celia Bruton for release of data prior to publication. We would also like to thank Professors Roger Hendrix and Graham Hatfull for stimulating discussions on phage evolution and for letting us see other phage sequences prior to publication. This work was supported by the BBSRC (Project grant G05986) and the Royal Society.

## REFERENCES

1 Chater,K.F. (1986) In Queener,S.E. and Day,L.E. (eds), *The Bacteria*, Vol. 9. Academic Press, Orlando, FL, pp. 119–158.
2 Bruton,C.J., Guthrie,E.P. and Chater,K.F. (1991) *BioTechnology*, **9**, 652–656.
3 Kuhstoss,S., Richardson,M.A. and Rao,R.N. (1991) *Gene*, **97**, 143–146.
4 Bierman,M., Logan,R., O'Brien,K., Seno,E.T., Rao,S.N. and Schoner,B.E., (1992) *Gene*, **116**, 43–49.
5 Kobler,L., Schwertfirm,G., Schmieger,H., Bolotin,A. and Sladkova,I. (1991) *FEMS Microbiol. Lett*., **78**, 347–354.
6 Sinclair,R.B. and Bibb,M.J. (1988) *Mol. Gen. Genet*., **213**, 269–277.
7 Hartley,N.M., Murphy,G.J.P., Bruton,C.J. and Chater,K.F. (1994) *Gene*, **147**, 29–40.
8 Kuhstoss,S. and Rao,R.N. (1991) *J. Mol. Biol*., **222**, 897–908.
9 Rausch,H. and Lehmann,M. (1991) *Nucleic Acids Res*., **19**, 5187–5189.
10 Howe,C.W. and Smith,M.C.M. (1996) *Microbiology*, **142**, 1357–1367.
11 Bergh,O., Borsheim,K.Y., Bratbak,G. and Heldal,M. (1989) *Nature*, **340**, 467–468.
12 Whitman,W.B., Coleman,D.C. and Wiebe,W.J. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6578–6583.
13 Hatfull,G.F. and Sarkis,G.J. (1993) *Mol. Microbiol*., **7**, 395–405.
14 Ford,M.E., Sarkis,G.J., Belanger,A.E., Hendrix,R.W. and Hatfull,G.F. (1998*) J. Mol. Biol*., **279**, 143–164.
15 Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S. and Barry,C.E.,III *et al*. (1998) *Nature*, **393**, 537–544.
16 Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al*. (1997) *Science*, **277**, 1453–1462 .
17 Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.-F., Dougherty,B.A. and Merrick,J.M. (1995) *Science*, **269**, 496–512.
18 Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertoro,M.G., Bessieres,P., Bolotin,A., Borchert,S. *et al*. (1997) *Nature*, **390**, 249–256.

19 Casjens,S., Hatfull,G. and Hendrix,R. (1992) *Semin. Virol.*, **3**, 383–397.
20 Sandmeier,H. (1994) *Mol. Microbiol.*, **12**, 343–350.
21 Neve,H., Zenz,K.I., Desiere,F., Koch,A., Heller,K.J. and Brussow,H. (1998) *Virology*, **241**, 61–72.
22 Desiere,F., Lucchini,S. and Brussow,H. (1998) *Virology*, **241**, 345–356.
23 Hendrix,R.W., Smith,M.C.M., Burns,R.N., Ford,M.E. and Hatfull,G.F. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 2192–2197.
24 Lucchini,S., Desiere,F. and Brussow,H. (1998) *Virology*, **246**, 63–73.
25 Suarez,J.E., Clayton,T.M., Rodriguez,A., Bibb,M.J. and Chater,K.F. (1992*) J. Gen. Microbiol.*, **138**, 2145–2157.
26 Ingham,C.J. and Smith,M.C.M. (1992) *Gene*, **122**, 77–84.
27 Ingham,C.J., Crombie,H.J., Bruton,C.J., Chater,K.F., Hartley,N.M., Murphy,G.J.P. and Smith,M.C.M. (1993) *Mol. Microbiol.*, **9**, 1267–1274.
28 Ingham,C.J., Hunter,I.S. and Smith,M.C.M. (1995) *Nucleic Acids Res.*, **23**, 370–376.
29 Howe,C.W. and Smith,M.C.M. (1996) *J. Bacteriol.*, **178**, 2127–2130.
30 Smith,M.C.M. and Owen,C. (1991) *Mol. Microbiol.*, **5**, 2833–2844.
31 Ingham,C.J., Owen,C.E., Wilson,S.E., Hunter,I.S. and Smith,M.C.M. (1994) *Nucleic Acids Res.*, **22**, 821–827.
32 Wilson,S.E., Ingham,C.J., Hunter,I.S. and Smith,M.C.M. (1995) *Mol. Microbiol.*, **16**, 131–143.
33 Wilson,S.E. and Smith,M.C.M. (1998) *Nucleic Acids Res.*, **26**, 2457–2463.
34 Connell,C., Fung,S., Heiner,C., Bridgham,J., Chakerian,V., Heron,E., Jones,B., Menchen,S., Mordan,W., Raff,M. *et al.* (1987) *Biotechniques*, **5**, 342.
35 Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
36 Hopwood,D.A., Bibb,M.J., Chater,K.F., Kieser,T., Bruton,C.J., Kieser,H.M., Lydiate,D.J., Smith,C.P., Ward,J.M. and Schrempf,H. (1985) *Genetic Manipulation of Streptomyces: A Laboratory Manual.* The John Innes Institute, Norwich, UK.
37 Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
38 Matsudaira,P.T. (1989) *A practical guide to protein and peptide purification for microsequencing*. Academic Press, San Diego, CA.
39 Chater,K.F. (1998) *Microbiology*, **144**, 1465–1478.
40 Rodriguez,A., Caso,J.L., Hardisson,C. and Suarez,J.E. (1986) *J. Gen. Microbiol.*, **132**, 1695–1701.
41 Suarez,J.E., Caso,J.L., Rodriguez,A. and Hardisson,C (1984) *FEMS Microbiol. Letts.*, **22**, 113–117.
42 Hendrix,R.W. and Duda,R.L. (1998) *Adv. Virus Res.*, **50**, 235–288.
43 Duda,R.L., Martincic,K. and Hendrix,R.W. (1995) *J. Mol. Biol.*, **247**, 636–647.
44 Duda,R.L., Hempel,J., Michel,H., Shabanowitz,J., Hunt,D. and Hendrix,R.W. (1995) *J. Mol. Biol.*, **247**, 618–635.
45 Catalano,C.E., Cue,D. and Feiss,M. (1995) *Mol. Microbiol.*, 1075–1086.
46 Stanley,E., Fitzgerald,G.F., Le Marrec,C., Fayard,B. and van Sinderen,D. (1997) *Microbiology*, **143**, 3417–3429.
47 Esposito,D., Fitzmaurice,W.P., Benjamin,R.C., Goodman,S.D., Waldman,A.S. and Scocca,J.J. (1996) *Nucleic Acids Res.*, **24**, 2360–2368.
48 Duda,R.L. (1998) *Cell*, **94**, 55–60.
49 Casjens,S. and Hendrix,R.W. (1988) In Calender,R. (ed.), *The Bacteriophages*, Vol. 1. Plenum Press, New York and London, pp. 15–91.
50 Levin,M.E., Hendrix,R.W. and Casjens,S.R. (1993) *J. Mol. Biol.*, **234**, 124–139.
51 Smith,M.C.M., Burns,N., Sayers,J.R., Sorrell,J.A., Casjens,S.R. and Hendrix,R.W. (1998) *Science*, **279**, 1834.
52 Engel,J. and Bachinger,H.P. (1999) *Indian Acad. Sci.*, in press.
53 Young,R. (1992) *Microbiol. Rev.*, **56**, 430–481.
54 Servin-Gonzales,L., Sampieri,A.I., Cabello,J., Galvan,L., Juarez,V. and Castro,C. (1995) *Microbiology*, **141**, 2499–2510.
55 Ziegelin,G., Scherzinger,E., Lurz,R. and Lanka,E. (1993) *EMBO J.*, **12**, 3703–3708.
56 Gorbalenya,A.E., Koonin,E.V. and Wolf,Y.I. (1990) *FEBS Lett.*, **262**, 145–148.
57 Gorbalenya,A.E. and Koonin,E.V. (1993) *Curr. Opin. Struct. Biol.*, **3**, 419–429.
58 Black,L.W. (1988) In Calender,R. (ed.), *The Bacteriophages*, Vol. 2. Plenum Press, New York and London, pp. 321–373.
59 Yeo,A. and Feiss,M. (1995*) J. Mol. Biol.*, **245**, 141–150.
60 Brown,K.L., Sarkis,G.J., Wadsworth,C. and Hatfull,G.F. (1997) *EMBO J.*, **16**, 5914–5921.
61 Baker,J., Limberger,R., Schneider,S. and Campbell,A. (1991) *New Biol.*, **3**, 297–308.
62 Walker,J.E., Saraste,M., Runswick,M.J. and Gay,N.J. (1982) *EMBO J.*, **1**, 945–951.