

DNA SEQUENCE ANALYSIS OF BACILLUS PHAGE Ø29 RIGHT EARLY REGION  
AND LATE GENES 14, 15 AND 16

by

Kevin James Garvey

---

A Dissertation Submitted to the Faculty of the  
DEPARTMENT OF MOLECULAR AND CELLULAR BIOLOGY

In Partial Fulfillment of the Requirements  
For the Degree of

DOCTOR OF PHILOSOPHY  
WITH A MAJOR IN MOLECULAR BIOLOGY

In the Graduate College  
THE UNIVERSITY OF ARIZONA

1 9 8 6

## INFORMATION TO USERS

This reproduction was made from a copy of a manuscript sent to us for publication and microfilming. While the most advanced technology has been used to photograph and reproduce this manuscript, the quality of the reproduction is heavily dependent upon the quality of the material submitted. Pages in any manuscript may have indistinct print. In all cases the best available copy has been filmed.

The following explanation of techniques is provided to help clarify notations which may appear on this reproduction.

1. Manuscripts may not always be complete. When it is not possible to obtain missing pages, a note appears to indicate this.
2. When copyrighted materials are removed from the manuscript, a note appears to indicate this.
3. Oversize materials (maps, drawings, and charts) are photographed by sectioning the original, beginning at the upper left hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is also filmed as one exposure and is available, for an additional charge, as a standard 35mm slide or in black and white paper format.\*
4. Most photographs reproduce acceptably on positive microfilm or microfiche but lack clarity on xerographic copies made from the microfilm. For an additional charge, all photographs are available in black and white standard 35mm slide format.\*

**\*For more information about black and white slides or enlarged paper reproductions, please contact the Dissertations Customer Services Department.**

**U·M·I** Dissertation  
Information Service

University Microfilms International  
A Bell & Howell Information Company  
300 N. Zeeb Road, Ann Arbor, Michigan 48106



8623827

**Garvey, Kevin James**

DNA SEQUENCE ANALYSIS OF BACILLUS PHAGE PHI29 RIGHT EARLY  
REGION AND LATE GENES 14, 15 AND 16

*The University of Arizona*

PH.D. 1986

University  
Microfilms  
International 300 N. Zeeb Road, Ann Arbor, MI 48106



DNA SEQUENCE ANALYSIS OF BACILLUS PHAGE Ø29 RIGHT EARLY REGION  
AND LATE GENES 14, 15 AND 16

by

Kevin James Garvey

---

A Dissertation Submitted to the Faculty of the  
DEPARTMENT OF MOLECULAR AND CELLULAR BIOLOGY

In Partial Fulfillment of the Requirements  
For the Degree of

DOCTOR OF PHILOSOPHY  
WITH A MAJOR IN MOLECULAR BIOLOGY

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 8 6

THE UNIVERSITY OF ARIZONA  
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read  
the dissertation prepared by Kevin James Garvey

entitled DNA Sequence Analysis of Bacillus Phage Ø29

Right Early Region and Late Genes 14, 15 and 16

and recommend that it be accepted as fulfilling the dissertation requirement  
for the Degree of Doctor of Philosophy.

John Spiggin

5/30/86  
Date

Colin W. Little

5/30/86  
Date

Harris Bernstein

5/30/86  
Date

David Mount

5/30/86  
Date

Janet Ito

May 30, 1986  
Date

Final approval and acceptance of this dissertation is contingent upon the  
candidate's submission of the final copy of the dissertation to the Graduate  
College.

I hereby certify that I have read this dissertation prepared under my  
direction and recommend that it be accepted as fulfilling the dissertation  
requirement.

Janet Ito  
Dissertation Director

May 30, 1986  
Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Signed: \_\_\_\_\_

*Karin J. J.*

## DEDICATION

This work and the resulting dissertation is dedicated to my wife,  
Toni, and to my son Brendan.

## ACKNOWLEDGMENTS

I wish to thank Dr. Junetsu Ito for his advice, support and guidance in the completion of this project. I am also indebted to my committee members; Drs. John Spizizen, Harris Bernstein, John Little and David Mount for their advice and critical reviews of this manuscript. David Mount is to be thanked for the use of his computer programs. My numerous friends and colleagues are also to be acknowledged for their friendship and moral support. I would particularly like to thank Mohammad Saedi, with whom I have had a number of productive collaborations.

This research was supported in part by a National Institute of Health Grant, GM28013, an American Cancer Society Grant, MV-229, and by Public Health Service Grant, T32 CA09213 from the Department of Health and Human Services.

## TABLE OF CONTENTS

	Page
LIST OF ILLUSTRATIONS . . . . .	vii
LIST OF TABLES . . . . .	viii
ABSTRACT . . . . .	ix
1. INTRODUCTION . . . . .	1
Overview . . . . .	1
Genome . . . . .	2
$\phi$ 29 Development . . . . .	2
Transcription . . . . .	6
$\phi$ 29 Morphogenesis . . . . .	8
$\phi$ 29 DNA Replication . . . . .	11
Project . . . . .	16
2. MATERIALS AND METHODS . . . . .	18
Bacterial Strains, Plasmids and Culture Conditions . . . . .	18
Enzymes and Chemicals . . . . .	19
DNA Preparation . . . . .	19
DNA Sequence Determination . . . . .	19
DNA Sequence Analysis . . . . .	19
In Vitro Protein Synthesis . . . . .	20
Complementation of Lysis Defective T4 Phage . . . . .	20
3. RESULTS . . . . .	21
DNA Sequence . . . . .	21
Coding Capacity . . . . .	25
Translation Initiation Regions . . . . .	26
In Vitro Protein Synthesis . . . . .	28
Transcription Signals . . . . .	32
Computer Analysis . . . . .	35
Complementation of Lysis Defective T4 Phages . . . . .	43
4. DISCUSSION . . . . .	47
Overview . . . . .	47
Genetic Organization . . . . .	47

TABLE OF CONTENTS--Continued

	Page
Translation Initiation Sites . . . . .	48
Transcription Signals . . . . .	49
Early Gene Functions . . . . .	53
Late Gene Functions . . . . .	57
Conclusions . . . . .	61
LIST OF REFERENCES . . . . .	62

## LIST OF ILLUSTRATIONS

Figure	Page
1. Genetic and Transcription Maps . . . . .	3
2. Schematic Diagram of $\phi$ 29 . . . . .	9
3. 5' Gap Created after RNA Priming of DNA Synthesis of a Linear Genome . . . . .	12
4. Protein Priming Mechanism of $\phi$ 29 DNA Synthesis . . . . .	14
5. Sequencing Strategy and Genetic Organization . . . . .	22
6. The Nucleotide Sequence of the Right Early Region . . . . .	23
7. The Nucleotide Sequence of the Late Region . . . . .	24
8. Autoradiographs of L-[ <sup>35</sup> S]Methionine-Labeled Proteins in the E. coli Coupled Transcription-Translation System . . . . .	31
9. Early Promoters of $\phi$ 29 . . . . .	33
10. Features of the $\phi$ 29 Early/Late Terminator Region . . . . .	36
11. Fastp Alignments of $\phi$ 29 gp15 and P22 gp19 . . . . .	39
12. Fastp Alignments of $\phi$ 29 gp15 and T4 gpe . . . . .	40
13. Fastp Alignments of P22 gp19 and T4 gpe . . . . .	41
14. Comparison of $\phi$ 29 gp15 Carboxy Terminal Duplications . . . . .	42
15. Construction of pMS2 . . . . .	45

LIST OF TABLES

	Page
Table	
1. ø29 Genes, Proteins and Functions . . . . .	4
2. ø29 Translation Initiation Sites . . . . .	27
3. Physicochemical Properties of the Putative Right Early Region Gene Products and Genes 14, 15 and 16 . . . . .	30
4. Complementation of T4 Mutant Infection . . . . .	46

## ABSTRACT

The sequence of the rightmost 4,626 bp of the Bacillus phage  $\phi$ 29 genome is presented and analyzed. Nine large open reading frames (ORF's) have been found. Three of these ORF's are correlated with the late genes 14, 15 and 16. The remaining six ORF's are in the right early region. One of these early ORF's has been identified as gene 17 (g17), the only early gene have been genetically mapped in this region. The remaining ORF's (16.5, 16.6, 16.7, 16.8 and 16.9) were previously unknown. The biological efficacies of some of these putative early ORF's were demonstrated using an in vitro E. coli transcription-translation system. The primary amino acid sequences, molecular weights, translational initiation sequences and genetic organization of these nine genes are presented and discussed. Gene product 15 (gp15) was found to have a strong homology with Salmonella phage P22 gp19, a lysozyme. gp15 also has a lesser but possibly significant homology with T4 gene product e (gpe), also a lysozyme. Using a clone containing  $\phi$ 29 g15 it was shown that gp15 can complement T4 gene e (ge) mutant infections, leading to the conclusion that  $\phi$ 29 g15 encodes a lysozyme. Three transcriptional initiation sites ( $P_{E3}$ ,  $P_{(EC)3}$  and B2) were previously mapped in this region. The sequences of the putative  $P_{(EC)3}$  and B2 promoter sites are presented and shown to have homology with the Bacillus  $\sigma^{55}$  consensus sequence. Sequences having homology to a minor Bacillus sigma factor recognition site,  $\sigma^{32}$ , are also presented and discussed. The region between the last late gene (g16)

and the last early gene (ORF-16.5) consists of only 30 bp. Analysis of potential secondary structures of transcripts across this region suggests that the same sequences may be involved in the termination of both late and early transcription.

## CHAPTER 1

## INTRODUCTION

Overview

Bacillus phage  $\phi$ 29 has played an increasingly useful and important role in the elucidation of molecular processes in Bacillus (Geiduschek and Ito, 1982). The genome (18 kb) is one of the smallest of all the Bacillus phages and is genetically and transcriptionally well characterized (Mellado et al., 1976). The relative genetic and transcriptional simplicity of  $\phi$ 29 has been exploited by a number of researchers to examine transcriptional and translational requirements in Bacillus (Sogo et al., 1979<sub>a</sub>; Davison, Murray, and Rabinowitz, 1980; Hwang and Doi, 1980; Murray and Rabinowitz, 1982; McLaughlin, Murray, and Rabinowitz, 1981; Hager and Rabinowitz, 1985). The  $\phi$ 29 virion is composed of only seven proteins, yet is morphologically complex (Anderson, Hickman, and Reilly, 1966; Anderson and Reilly, 1976; Mendez et al., 1971; Reilly and Spizizen, 1965). This has made  $\phi$ 29 an attractive model for phage morphogenesis; the study of which, has resulted in the development of a well defined, highly efficient in vitro packaging system (Bjornsti, Reilly, and Anderson, 1981; Guo, Grimes and Anderson, 1986). Finally,  $\phi$ 29 DNA replication occurs in a novel manner; de novo DNA synthesis initiates via a protein priming mechanism (for a review see Salas, 1983).

### Genome

The genome of  $\phi 29$  is an 18 kb linear double-stranded DNA molecule, the 5'-termini of which are covalently bound to the terminal protein, which initiates replication (Watabe, Shih, and Ito, 1983). Eighteen genes have been mapped as diagrammed in Fig. 1 (Mellado et al., 1976; Reilly, Nelson, and Anderson, 1977). The late genes (7-16) are clustered in the center of the genome, flanked by early genes (1-6 and 17) at either end of the genome. The early genes, therefore, are divided into two distinct transcriptional units. Approximately twenty-five phage specific proteins have been identified in infected cells and protein products have been assigned to all the genes except genes 13 and 14. However, there are several phage specific proteins that have not been given gene assignments (Hawley et al., 1973; Hagen et al., 1976; Anderson and Reilly, 1976).

The functions of most known  $\phi 29$  genes have been established as outlined in Table 1. The late genes are involved in morphogenesis (as structural or morphogenic proteins) or in cell lysis. The early genes are particularly interesting since all but one ( $g_4$ ) has been implicated in the replication process (Carrascosa et al., 1976).

### $\phi 29$ Development

There are several features of  $\phi 29$  development that warrant a review. The phage exhibits a rather limited host range infecting only selected strains of *B. subtilus*, *B. licheniformis*, *B. pumilus* and *B. amyloliquifaciens* (Reilly, 1976). This may be due, in part, to the inability of the phage to be adsorbed on the cells, a reaction shown

Fig. 1. Genetic and Transcription Maps.

A: The genetic map was adapted from those of Mellado et al. (1976), Reilly et al. (1977) Garvey K.J., Yoshikawa, H. and Ito J. (1985<sub>b</sub>). Genes 1-6 and 16.5-17 are early genes and 7-16 are late genes. The arrows represent the direction and extent of the early and late transcripts.

B: Early transcription map based on the results of Davison et al. (1980).

C: A schematic diagram of the transcriptional signals. The letters A,B,C,D and E refer to the products of an EcoRI digest of the  $\phi$ 29 genome.

D: Transcription map based on the results of Sogo et al. (1979<sub>a</sub>). The vertical bars represent the in vitro mapped B. subtilis RNA polymerase binding sites. The arrows designate the direction and extent transcription, determined in vivo.

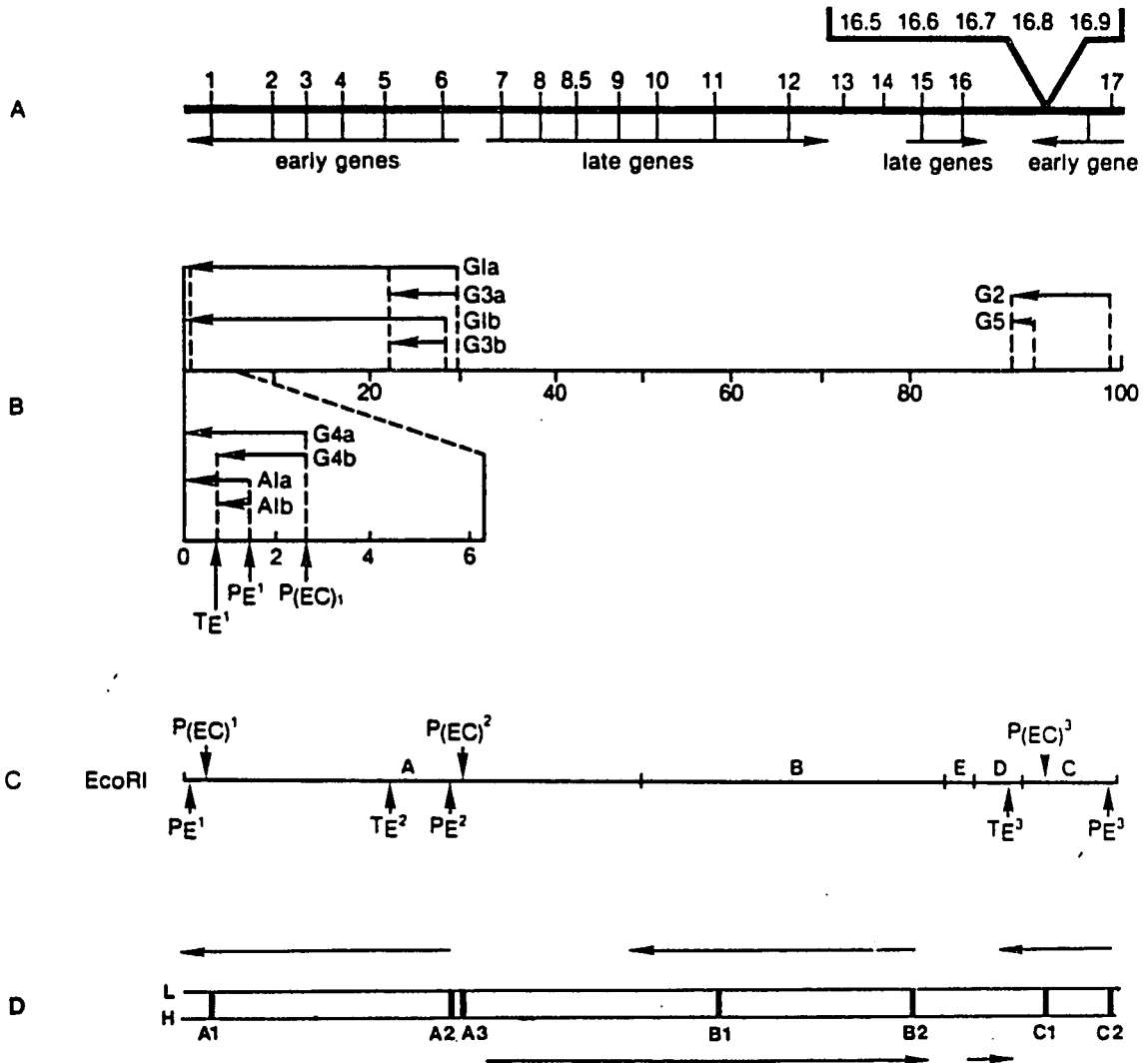


Fig. 1. Genetic and Transcription Maps.

Table 1.  $\phi$ 29 Genes, Proteins and Functions

Gene	Time of expression	Protein or phenotype
1	Early	DNA synthesis <sup>a</sup>
2	Early	DNA polymerase
3	Early	Terminal protein
4	Early	Control of late transcription
5	Early	DNA synthesis
6	Early	DNA binding protein
7	Late	Scaffolding protein
8	Late	Major head protein
8.5	Late	Head fiber protein
9	Late	Tail protein
10	Late	Upper collar protein
11	Late	Lower collar protein
12	Late	Neck appendage protein precursor
13	Late	Morphogenesis
14	Late	Lysis
15	Late	Lysis (morphogenesis)
16	Late	Encapsidation protein
17	Early	DNA synthesis <sup>a</sup>

<sup>a</sup> Some sus mutants in genes 1 and 17 synthesize  $\phi$ 29 DNA, and others do not (Hagen et al., 1976; Harding and Ito, 1976).

to require glucosylated teichoic acid (Young, 1967). This is supported by the observation that competent cells of strains unable to adsorb  $\phi 29$  DNA, can be transfected with  $\phi 29$  DNA (Reilly and Spizizen, 1965).  $\phi 29$  can also transfect *B. subtilis* protoplasts (Stahly and Ito, 1981).

Subsequent to  $\phi 29$  adsorption, the genome is injected into the cell with the right end apparently entering the cell first (Bjornsti, Reilly, and Anderson, 1983; Krawiec et al., 1981). Early protein synthesis begins almost immediately, but not all early proteins appear or disappear at the same times (Hawley et al., 1973). DNA synthesis begins at about 10 minutes after infection and continues until lysis. The late proteins appear about this time with one exception; the major head protein is synthesized early in infection (4-6 minutes). Synthesis of many early proteins continues until very late in infection. Lysis occurs 55-60 minutes after infection at 37 C and burst size is significantly affected by growth conditions (Kawamura and Ito, 1975).

Host cell functions are not greatly affected by  $\phi 29$  infection; DNA, RNA, and protein synthetic processes appear to function normally until very late in infection (Schachtele et al., 1972). When  $\phi 29$  infects sporulating cells phage DNA synthesis is inhibited and  $\phi 29$  DNA is incorporated into the spores in a heat stable form (Kawamura and Ito, 1974, Kawamura and Ito, 1975). This occurs at about the same point at which host DNA synthesis is also terminated (Ito, J. unpublished data). The incorporated phage genome is expressed only after

germination and spore outgrowth. Suggesting that  $\phi 29$  may serve as a probe for investigating host sporulation and germination processes.

### Transcription

Transcription maps of the  $\phi 29$  genome have been published by several laboratories (Fig. 1). Early mRNA's are transcribed from right to left on from the L-strand of  $\phi 29$  DNA and are synthesized throughout infection (Loskutoff, Pene, and Andrews, 1973; Schachtele, DeSain, and Anderson, 1973). Early mRNA's are transcribed from two regions, the right early region at the right terminus and the left early region at the left terminus (Kawamura and Ito, 1977). Synthesis of late mRNA is correlated with, but not dependent upon, initiation of  $\phi 29$  DNA synthesis and is transcribed left to right from the H-strand (Loskutoff et al., 1973; Schachtele et al., 1973). Gene 4 product is required for late mRNA synthesis (Anderson and Reilly, 1974; Sogo et al., 1979<sub>a</sub>) and is thought to function as a sigma type factor (Salas et al., 1984).

Transcription of both late and early regions appears to be quite complex. Loskutoff et al. (1973) and Loskutoff and Pene (1973) demonstrated that late mRNA consisted of at least three species (MW,  $1.75 \times 10^6$ ,  $0.93 \times 10^6$ , and  $0.07 \times 10^6$ ), and early mRNA was comprised of at least six species (MW, 0.04 to  $0.75 \times 10^6$ ). Subsequent investigations by Kawamura and Ito (1977) found at least 13 early transcripts (MW, 0.09 to  $1.0 \times 10^6$ ). It was proposed, based on transcript size and coding capacity, that some of the mRNA's were redundant. This could

be due to RNA processing, degradation or regulation of initiation and termination sites.

Davison et al. (1980) have mapped a number of in vitro early transcripts synthesized using E. coli and B. subtilis RNA polymerases (Fig. 1B). These results indicate that the early regions contain at least four B. subtilis and six E. coli polymerase promoter sites and three termination sites. These results suggest that some of the transcripts are redundant as proposed by Kawamura and Ito (1977).

Sogo et al. (1979<sub>a</sub>) also examined  $\phi$ 29 transcription and the results are diagrammed in Fig. 1D. The authors visualized B. subtilis RNA polymerase binding to  $\phi$ 29 DNA using electron microscopy and determined the positions of the binding sites. The extent of transcription was determined by DNA:RNA hybridization. Comparisons of E. coli and B. subtilis polymerase binding sites indicated that the E. coli enzyme bound to the same sites as the B. subtilis enzyme, as well as two additional sites. These results are consistent with those of Davison et al. (1980) with two important exceptions. Sogo et al. (1979<sub>a</sub>) observed that the H (late) strand was transcribed in vitro, initiated at A3, but to a lesser extent than the in vivo results. Also it was observed that anti-late mRNA synthesis occurred in the EcoRI-B region both in vivo and in vitro. This region contains two B. subtilis RNA polymerase binding sites. The nature of this symmetric transcription remains unknown.

The entire left and right early regions have been sequenced and the transcription signals identified (Yoshikawa and Ito, 1981;

Garvey, Yoshikawa, and Ito, 1985<sub>b</sub>; Dobinson and Spiegelman, 1985). The sequences of the promoter regions are compiled in Fig. 9 and they all show good homology with the consensus sequence. These must be considered putative, however, except P<sub>(EC)3</sub> which was identified by S1 mapping (Dobinson and Spiegelman, 1985). The RNA polymerase binding sites are known to a high degree of precision ( $\pm 54$  to 216 bp), increasing the probability that the sites selected are correct. Putative terminators have also been identified. The T<sub>E</sub>2 and T<sub>E</sub>3 regions have sequences capable of forming stable hairpin structures (Yoshikawa and Ito, 1981; Garvey, Yoshikawa, and Ito, 1985<sub>b</sub>) similar to factor independent terminator sequences in *E. coli* (Platt and Bear, 1983; Holmes, Platt, and Rosenberg, 1983). No such sequences were found in the T<sub>E</sub>1 region which may be similar to the rho factor dependent terminators of *E. coli*. Consistent with this proposal, it has been observed that in vitro transcription of  $\phi$ 29 DNA was affected by purified *B. subtilis* rho factor (Hwang and Doi, 1980). Unfortunately, the specific sites affected were not determined.

#### $\phi$ 29 Morphogenesis

The  $\phi$ 29 virion is a morphologically intricate construct despite the small genome size. The virion, as diagrammed in Fig. 2, features an anisometric head, a collar region and a short non-contractile tail (Anderson et al., 1966). Attached to the head are numerous fibers (gp8.5) (Tosi and Anderson, 1973; Reilly, et al., 1977). The neck is composed of an upper and lower collar. The upper collar is attached to the phage head. The lower collar is the attachment point



for the twelve appendage proteins, which are thought to be involved in the adsorption process (Tosi and Anderson, 1973). The tail is attached to the neck region and is larger at the bottom than the top (Anderson et al., 1966).

Morphogenesis of the  $\phi$ 29 virion occurs via a single pathway, for which a number of provisional pathways have been proposed (Nelson Reilly, and Anderson, 1976; Jimenez et al., 1977; Murialdo and Becker, 1978; Carrascosa et al., 1981). In brief, proheads, composed of the major head protein (gp8), the head fiber protein (gp8.5), the neck collar protein (gp10) and the scaffolding protein (gp7), are filled with  $\phi$ 29 DNA-gp3 complex in the presence of the encapsidation protein (gp16). None of the structural proteins, except gp12\*, appear to be products of proteolytic cleavage (Tosi, Reilly, and Anderson, 1975). The products of gp9 and gp10 may, in addition to their structural functions, play roles in morphogenesis (i.e. DNA packaging) (Nelson et al., 1976).

An in vitro packaging system has recently been developed that has confirmed the provisional assembly pathway and, because of its high efficiency, has facilitated analysis of the  $\phi$ 29 DNA-gp3 packaging mechanism (Bjornsti et al., 1981). The reaction has an absolute requirement for functional  $\phi$ 29 DNA-gp3 complex. Deproteinized DNA is not packaged (Bjornsti, Reilly, and Anderson, 1982). Bjornsti, Reilly, and Anderson (1985) have also demonstrated that gp2 and other undetermined early proteins may be involved in, but not required for, morphogenesis. Another essential morphogenic gene product is gp16,

which catalyzes the encapsidation reaction. This gene has recently been cloned and gp16 purified (Guo et al, 1986). It is the only protein required for the in vitro prohead- $\phi$ 29 DNA-gp3 encapsidation reaction (Bjornsti, Reilly, and Anderson, 1984).

This is the only in vitro packaging system available in Bacillus and is about 20x more efficient than the lambda system (Bjornsti et al., 1984). This system has recently been used to package restriction digested Bacillus DNA ligated to similarly digested  $\phi$ 29 DNA terminal fragments (Ganesan and Hoch, 1984). Transduction of auxotrophs to prototrophy was demonstrated, indicating that  $\phi$ 29 has potential as a cloning vector.

#### $\phi$ 29 DNA Replication

The fact that no known DNA polymerase can initiate de novo DNA synthesis places certain constraints on the replication of linear genomes such as  $\phi$ 29. If initiation were to occur by an RNA priming mechanism, subsequent removal of the primer would result in a single-stranded 3'-OH tail (Fig.3). A number of models have been proposed to circumvent this problem; requiring the existence of terminal structures such as cohesive ends (Wu and Taylor, 1982), terminal redundancies (Watson, 1972), palindromic termini (Cavalier-Smith, 1974), and crosslinked ends (Estaban, Flores, and Holowczak, 1977). Sequence analysis has shown that  $\phi$ 29 termini have only a six basepair inverted repeat at each end (Yoshikawa, Friedman, and Ito, 1981). No other structures suitable for the models described above were found.

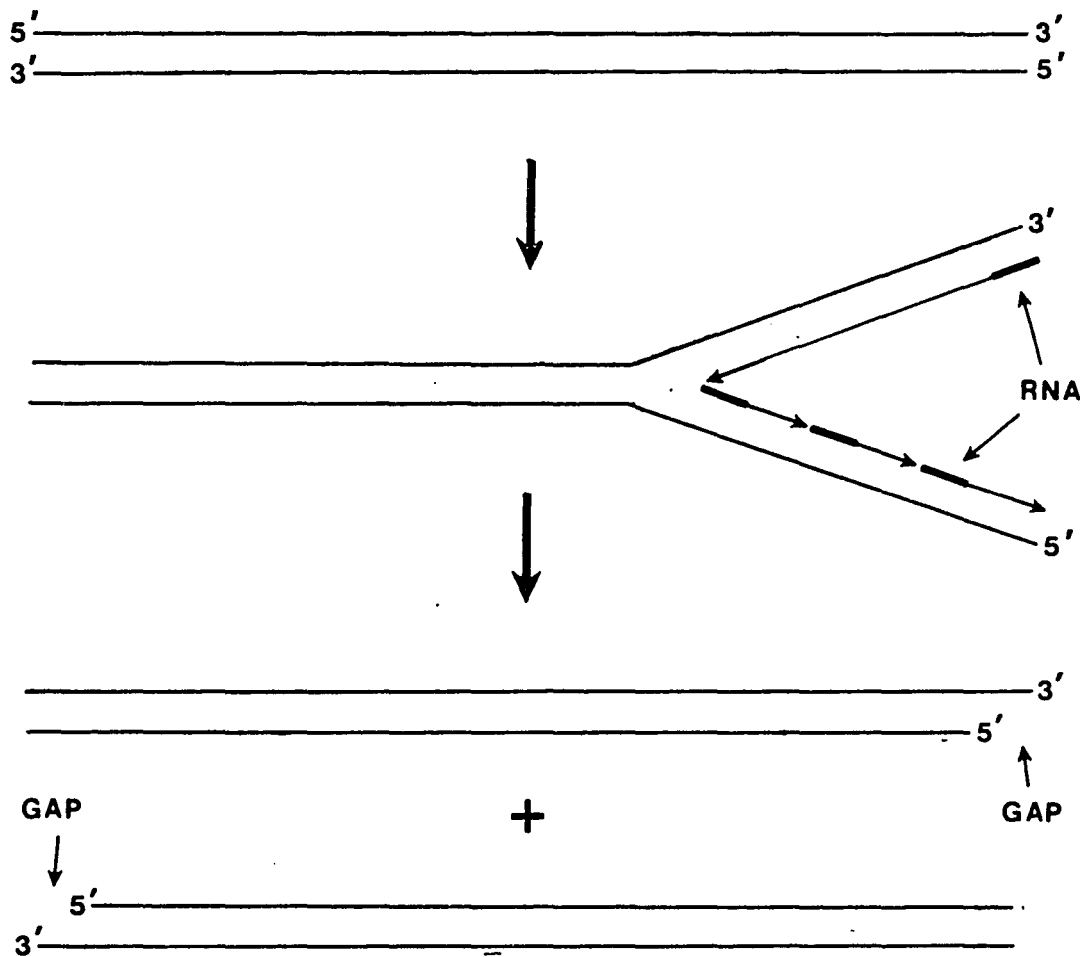


Figure 3. 5' Gap Created after RNA Priming of DNA Synthesis of a Linear Molecule.

Adenoviridae are eukaryotic transforming viruses that have proteins linked to the 5' ends of their duplex linear DNA genome (Rekosh et al., 1977). As with  $\phi 29$ , the protein is bound to the 5' nucleotide (dCMP) via an O-5'-(nucleotydy)-L-serine bond (Desiderio and Kelly, 1981). Rekosh et al., (1977) proposed that the pre-terminal protein (pTP), which is later processed to yield the terminal protein found in the virion, serves as a primer for adenovirus replication (Fig. 3). This hypothesis has since been extended to  $\phi 29$  DNA replication (Harding and Ito, 1980; Mellado et al., 1980). According to this model (Fig. 4), newly-synthesized gp3 reacts with the appropriate deoxyribonucleotide triphosphate, forming a gp3-dNMP complex, and thereby providing the 3'-OH necessary for elongation by DNA polymerase. Replication would occur via strand displacement, and as a consequence would not require lagging strand synthesis.

An in vitro DNA replication system has been developed for  $\phi 29$  (Watabe, Shih, and Ito, 1982; Penalva and Salas, 1982). Synthesis is dependent upon exogenous  $\phi 29$  DNA-gp3 complex and does not occur if deproteinized  $\phi 29$  DNA is used as a template. Elongation occurred via asymmetric strand displacement. This system also catalyzes the reaction of gp3 with  $\alpha$ - $^{32}\text{P}$ -dATP to form a gp3-dAMP complex (Shih, Watabe, and Ito, 1982). This system has since been further refined and DNA synthesis was shown to require only two phage encoded activities; DNA polymerase (gp2) and terminal protein (gp3) (Watabe, Leusch, and Ito, 1984; Blanco and Salas, 1985). In this defined system gp3-dAMP complex is formed which subsequently serves as a primer for chain

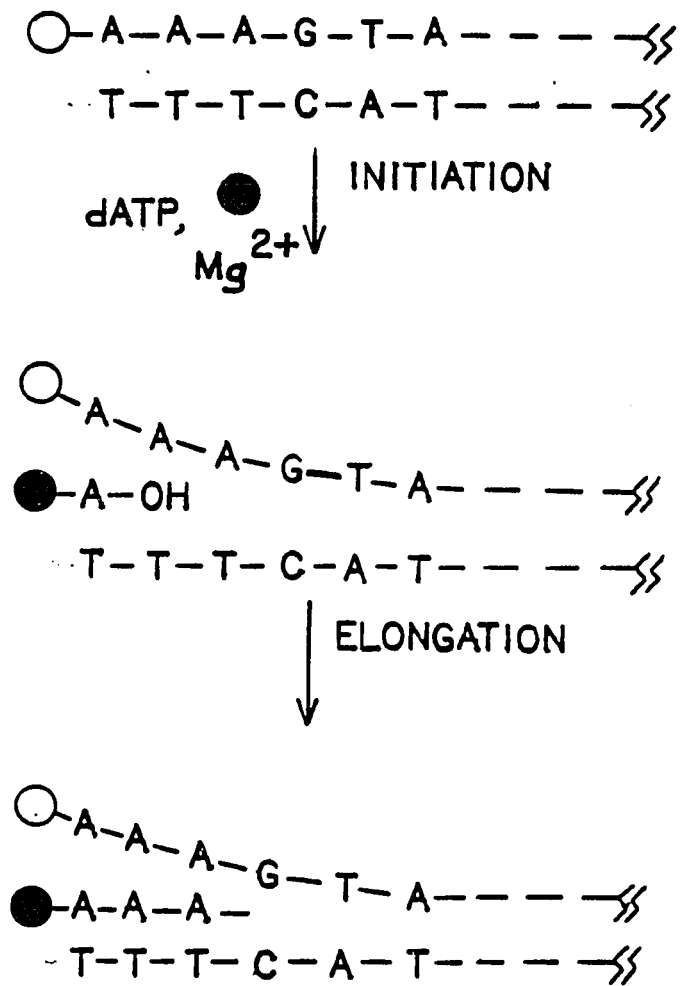


Figure 4. Protein Priming Mechanism of  $\phi$ 29 DNA Synthesis.

elongation as in the crude system. Preliminary evidence suggests that initiation reaction was stimulated by host cell extracts, indicating that host proteins may be involved. In any case, elongation was inefficient, suggesting that additional host or viral encoded proteins are required.

The functions of the remaining early gene thought to be involved in  $\phi$ 29 replication are still unknown. Recently, g6 has been cloned and the product purified. The purified protein may stimulate the in vitro replication system (Pastrana et al., 1985), but other results are contradictory (Hodges, 1986). No specific activity, such as topoisomerase activity (although the protein did associate with E. coli topoisomerase I), helicase or exonuclease activities (Hodges, 1986). The protein did bind to double and single stranded DNA, albeit weakly (Hodges, 1986).

Carrascosa et al. (1976) have reported that conditionally lethal mutants of genes 1, 2, 3, 5, 6 and 17 were unable to synthesize DNA under non-permissive conditions. Hagen et al. (1976) also assayed for DNA synthesis using conditional lethal mutants and confirmed that DNA synthesis did not take place in mutants of genes 2, 3 and 5. However, mutants of genes 1 and 6 supported limited DNA synthesis, and a gene 17 mutant exhibited wild type incorporation levels. The mutants used (in genes 1, 6 and 17) were the same as those used by Carrascosa et al. (1976). Another gene 17 mutant examined by Hagen et al. (1976) could support only limited DNA synthesis. These discrepancies may be due to the different host strains used.

Mellado et al. (1980) performed shift up experiments using temperature sensitive (ts) mutants of genes 2, 3, 5 and 6. It was determined that ts mutants of genes 2 and 3 formed full-length DNA after the temperature shift, indicating that genes 2 and 3 are involved in initiation. Mutants of genes 5 and 6 produced no full-length DNA or production was delayed after shift up. It was proposed that these genes are, therefore, involved in the maturation process or in elongation. These results were supported by *in vitro* experiments reported by Blanco et al. (1983). The researchers assayed gp3-dAMP complex formation using extracts from  $su^-$  cells infected with  $\phi 29$  suppressor sensitive mutants in genes 2, 3, 5, 6 and 17. Extracts from sus2 and sus3 infected cells could not catalyze this reaction. Extracts from sus5, sus6 and sus17 infected cells could catalyze this reaction. It appears, then, that genes 2 and 3 are involved in initiation, and genes 5, 6, and 17 are involved in the elongation or maturation process.

#### Project

DNA sequence analysis has greatly aided the study of phages such as T7 and  $\lambda$  by providing a concise map of their genetic organizations and a description of the overall patterns of gene expression during infection at the nucleotide level. With benefits in mind, an effort to completely sequence the genome of Bacillus phage  $\phi 29$  has been under taken (Yoshikawa and Ito, 1981; Garvey et al., 1985<sub>a</sub>; Garvey et al., 1985<sub>b</sub>). Presented here is the sequence of 4,626 bp of the right end of the genome. The primary amino acid sequences of both

late and early proteins encoded in this region are provided and examined. A total of nine genes, of which only four were previously known, were found. One late gene, function previously unknown, is shown to be structurally and functionally similar to other phage lysozymes. This analysis also discusses the structural significance of the transcriptional and translational control regions.

## CHAPTER 2

## MATERIALS AND METHODS

Bacterial Strains, Plasmids and Culture Conditions

Bacteriophage  $\phi$ 29 was grown using Bacillus amyloliquefaciens H strain as the host.  $\phi$ 29 strains sus14(1241)sus16(300) and sus8(22)-sus10(302)sus14(1241) were grown using Bacillus subtilis MO-101-P spoA<sup>-</sup> {met<sup>-</sup>}<sup>+</sup> thr<sup>-</sup> su<sup>+44</sup> as the host (Bjornsti et al., 1981). The phage were purified as described previously (Kawamura and Ito, 1977).

Escherichia coli strain SCR 1271, C600 rk<sup>-</sup> mk<sup>-</sup> recBC<sup>-</sup> containing the  $\phi$ 29 Hind III-C fragment cloned into pBR313 (Ito and Roberts, 1980) was grown as described (Marko, Chipperfield, and Birnboim, 1982).

Bacteriophage T4 was obtained from John Obringer. The T4 gene e mutants, eG79 (a gene e deletion mutant), am882 and amH26x5 were obtained from Pat Tedesco (University of Colorado). E. coli strains S/6 and CR63 (S, supD60, lamB63), also obtained from John Obringer, were used as hosts for the wildtype and amber T4 strains.

E. coli strain K12 (sm<sup>r</sup>, lacZam, bio-uvrB, trpEA2 (Nam53, cI857, H1)), designated ER69, was the host for plasmids pMS2, pMS6 and pPLc245. The parent strain and the plasmid containing derivatives were obtained from Mohammad Saedi.

### Enzymes and Chemicals

Restriction endonucleases were purchased from Bethesda Research Laboratories or New England Biolabs. Calf intestinal alkaline phosphatase was obtained from Boehringer Mannheim. Polynucleotide kinase was purchased from P-L Biochemicals. [ $\gamma$ - $^{32}$ P]ATP (>7000 Ci/mmol) was purchased from ICN. N,N'-methylenebisacrylamide, acrylamide and urea were obtained from Bio Rad Laboratories.

### DNA Preparation

Phage  $\phi$ 29 DNA was prepared as described previously (Ito, 1978). Plasmid DNA was isolated as described by Marko et al., (1982).

### DNA Sequence Determination

The DNA was digested with an appropriate restriction enzyme; the digestion products were treated with calf intestinal alkaline phosphatase (Boehringer Mannheim) and labeled with [ $\gamma$ - $^{32}$ P] ATP ( $\geq$  7000 Ci/mmol; ICN) using T $_4$ -polynucleotide kinase (P-L Biochemicals). The DNA was sequenced by the method of Maxam and Gilbert (1980), as modified by Smith and Calvo (1980). The chemical cleavage reactions used were the G, A+G, C+T and C specific reactions. Cleavage products were separated on 6%, and 20% polyacrylamide gels according to the method of Sanger and Coulson (1978). The gels were usually dried and autoradiographed at -70 $^{\circ}$  with or without intensifying screens (Kodak).

### DNA Sequence Analysis

Determination of open reading frames, translation product sequences and restriction sites were analyzed using the microcomputer

programs of Mount and Conrad (1984). The National Biomedical Research Foundation (NBRF) protein library was searched on an IBM-PC using the programs of Lipman and Pearson (1985). The library and programs were generously provided by David Mount.

#### In Vitro Protein Synthesis

The in vitro transcription-translation kit was purchased from Worthington and used accordingly to the protocol provided by the manufacturer. The kit consisted of an S-30 extract of E. coli (recB<sup>-</sup> recC<sup>-</sup>) (Nirenberg and Matthau, 1961; Chen and Zubay, 1983) and the components necessary for RNA and protein synthesis. This system will transcribe and translate exogenous linear or supercoiled DNA if suitable signals are present on the template. The newly synthesized proteins were labeled by adding [<sup>35</sup>S] methionine (>1000 Ci/mmol, New England Nuclear) to the reaction. The reaction products were separated using SDS-polyacrylamide gel electrophoresis (Laemmli, 1970). The gels were dried and exposed to film (Kodak XRP-6), without chemical enhancers, for 1-8 days.

#### Complementation of Lysis Defective T4 Phage

E. coli strain K12 H1 trp bearing the plasmids to be tested was grown to  $2 \times 10^8$  cells/ml and plated in Hersheys soft agar (Steinberg and Edgar, 1962) on LB plates (Lennox, 1955) containing 100 ug/ml ampicillin when appropriate. Dilutions of the T4 gene e mutants to be tested were spotted and plaque formation was noted.

## CHAPTER 3

## RESULTS

DNA Sequence

The sequence of the  $\phi 29$  rightmost region (4,626 bp) is presented in Figs. 6 and 7. The rightmost 274 bp (Fig. 6) of sequence was published previously (Yoshikawa et al., 1981) and is included here for continuity. The DNA sequencing strategy is diagrammed in Fig. 5 along with a detailed restriction map and genetic map. The DNA sequence was determined from both strands for approximately 90% of the DNA. Sequencing was done across all sequencing start sites to prevent errors due to closely spaced restriction sites.

The sequence is relatively A-T rich, having a G-C content of only 38%. This is identical to the value obtained from the sequence of the left early region of  $\phi 29$  (Yoshikawa and Ito, 1981). It is also consistent with the estimate derived from chemical hydrolysis data (Rubio et al., 1974). The G-C content of this sequence was examined using a 50 bp window (data not shown). The G-C% ranged from 28% to 56% with a relatively A-T rich region occurring between 50-400bp (average G-C% = 32%). This agrees with previous partial denaturation results, which suggested that the right terminus of  $\phi 29$  is A-T rich (Sogo et al, 1979<sub>b</sub>).

**Figure 5. Sequencing Strategy and Genetic Organization.**

(i) A detailed restriction map. Key: AccI=A, EcoRI=E, HpaII=H, HindIII=Hd, HinfI=Hf, HincPI=Hn, HpaI=Hp, MboI=M, NcoI=N, TaqI=T.

(ii) Sequencing strategy. The arrows represent the direction and extent of sequence readings for each of the labeled sites.

(iii) The arrows represent the direction and extent of the putative ORF's in this region.  $P_E3$  and  $P_{(EC)3}$  designates the early promoters;  $T_E3$  the early terminator.

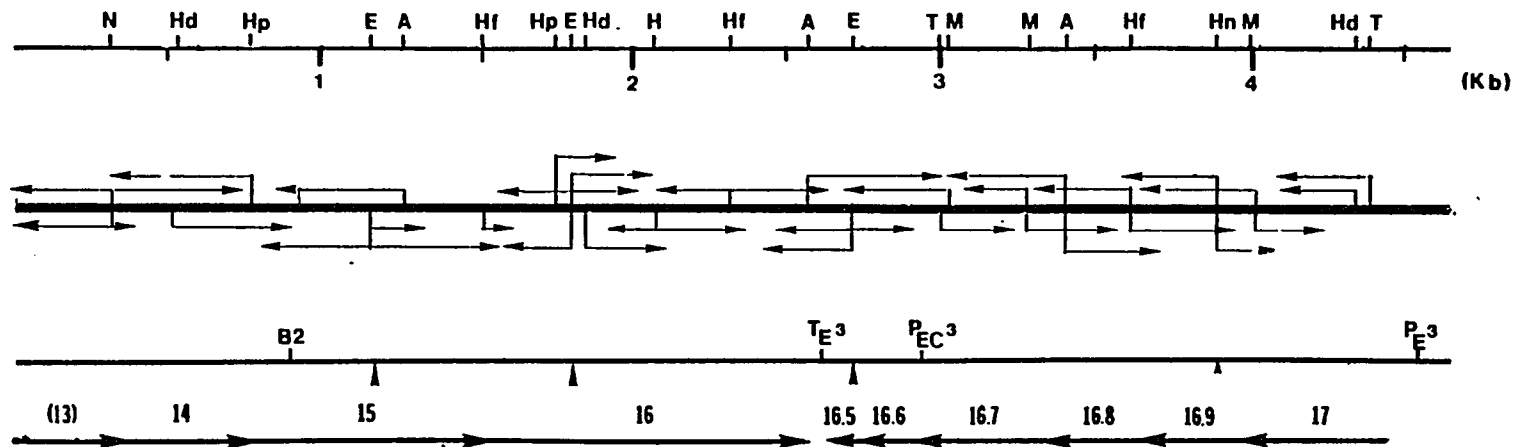


Figure 5. Sequencing Strategy and Genetic Organization.

Figure 6. The Nucleotide Sequence of the Right Early Region.

The inferred amino acid sequences of the putative open reading frames are also shown. The \* indicates a stop codon. The presumptive "-35" and "-10" regions of the  $P_{(EC)3}$  and  $P_{E3}$  are as indicated. The putative termination site  $T_{E3}$  is as indicated.

100  
 AAAGTAGGGTACAGCGACAACATACACCATTTCGCCATTGACCGACTACTCTCGACAAGAACTCTAACAACCTAAATCAGGACTATATACCTATACTATT  
 -35 F<sub>3</sub> -10  
 200  
 TTATCATCAATTTGTCGAAAAGGGTAGACAAACTATCGTTTAACTGTTATACTATAATAGAAAGTAAGGTAATAAGACAACCAATCATAGGAGGAATTAC  
 GENE-17  
 300  
 ACATGAAATAACTATCAATTAACACTCAATGAGGTAATCGACATCAACACAAAACACAGAGATCAATAGCTTGTAGCCAAGAAAAGAAAATCTATTCC  
 MetAsnAsnTyrGluLeuThrIleAsnGluValIleAspIleIleAsnThrAsnThrGluIleAsnLysLeuValAlaLysLysGluAsnLeuPhePro  
 400  
 TACAGACTTATACGACTTAGACAAACAAGAACTTATCGCTATCATCCTTAACAGTGACTTTGCCTATCTAGCATTAAAGAGTGCTACTAGAAGTAACT  
 ThrAsnLeuTyrAspLeuAspLysGlnGluLeuIleAlaIleIleLeuAsnSerAspPheAlaLeuSerSerIleLysArgValLeuLeuGluValThr  
 500  
 GTTGAAGAGTTAGGCACACAAAGACAACGACGAAGATGATGAGTTAGAAGATTAGACGGTGAATAGATAGAGTTGACTATATTGATAAAGACGGTATCA  
 ValGluGluLeuGlyThrGlnAspAsnAspGluAspAspGluLeuGluAspLeuAspGlyGluIleAspArgValAspTyrIleAspLysAspGlyIleArg  
 600  
 GATTGATGTTCCACGTGAAACATCACCACAGTAGATAAGTCAATCGTTACATTCATGATGAGCTTCTGTGTAAGCGAATAAGATCGCCAAGTCAAT  
 PheAspValProArgGluThrSerProHisValAspLysSerIleValThrPheAsnAspGluLeuLeuAspGluAlaAsnLysIleAlaLysSerIle  
 700  
 ACAAGAACATGACTTTAATGACAAAGCTATAGAAGAAGCAGAAGCTTAAATATTCAAGAACCACTTACCATCTATCTACAGCATGAAGAAGGAGAACAAG  
 GlnGluHisAspPheAsnAspLysAlaIleGluGluAlaGluLeuLysIlePheLysAsnHisLeuProSerIleTyrSerMetLysLysGluAsnLys  
 ORF-16.9 Hin PI 800  
 TAACATGAGCGTAGCAACTTAATGCATTCACTTTCATCTTAGAGCGCGAGGTGGGAATGGTATGCTACGAACAACATAACAACAATGGCACAGCATC  
 \* MetSerValGlnLeuAspAlaPheThrPheIleLeuGluArgArgGlyTrpArgMetValCysTyrGluGlnGluThrThrAspGlyThrArgIle  
 900  
 TTACATTCTATCTCAAGGACAACCCACCGTCTTCTGCTACATACTCATCGCAATTCITATCTGATACTAAAATGATAAGACGCTTTGCATCATGGAGCG  
 LeuHisPheTyrLeuLysAspAsnProThrPhePheAlaThrTyrSerSerGlnPheGluSerAspThrLysMetIleArgArgPheAlaSerTrpSerGly  
 1000  
 GTCAGTTACTAGAAGGCTCAAACTCTGTGTTTTGGACAAACATCACACCATTGAAACCAATGATGAGGAAACAGCAGAAGACATCAAGAACTTGATAA  
 GlnLeuLeuGlnGlySerAsnSerValPheTrpThrAsnIleThrProPheGluProIleAspGluGluThrAlaGluAspIleLysAsnLeuAspLys  
 ORF-16.8 1100  
 AGTCGTTGAGGGGATGAATTTACATTATGATAGACATCATGTAAGAGGACAAGCGGCTAATCACTGTTCAAACACAGAGGGAGACGAAGTATTTT  
 ValValGluGlyMetAsnPheThrLeu \*IleAspIleIleValLysGluAspLysArgLeuIleThrValGlnThrProGluGlyAspGluValPheTyr  
 Met 1200  
 ACACCTTGTCTTTCACAGCGGTACAAAGATACTGAAACGTTCAAGTGCAGGACTAAGAAACAACATTTATGCAATTTGGTGTGGCTAACATCAGATGGAT  
 ThrLeuSerPheSerAspGlyHisLysIleLeuLysArgSerSerAlaArgLeuArgAsnAsnIleTyrAlaIleGlyValAlaAsnIleArgTrpMet  
 1300  
 GCTAGTAGACATGGATAACATGATACTTACTAGTACATACATCATGTTGATATTTTAAAGACATCGACCGTAAAATGAGGAAATGGGTTATATAGTT  
 LeuValAspMetAspAsnMetIleLeuSerLeuTyrIleHisHisValAspIleLeuLysAspIleAspArgLysMetArgGluMetGlyTyrIleVal  
 ORF-16.7 1400  
 ATTTGAGAAATGGCAACAGCCAAATAAAAAGGGGACAAGGAGATAACATGGAAGCTATTTAATGATCGGGTACTTGCATTGTGCGTTATATCCTCTGA  
 IleSerGluTrpGluHisAlaAsnLysLysGlyThrArgArg \* MetGluAlaIleLeuMetIleGlyValLeuAlaLeuCysValIlePheLeuLeu  
 1500  
 TCAGGACGAAACAACAAAAGAAACAGGAAGCAAGGGAGCTAGAAGATTATCTTGAAGACCTCAACAAACGAGTTGTTCAACGAACACAAATACTCAGCG  
 SerGlyArgAsnAsnLysLysLysGlnGluAlaArgGluLeuGluAspTyrLeuGluAspLeuAsnLysArgValValGlnArgThrGlnIleLeuSerGlu  
 1600  
 AGCTTAAACGAAGTTATCTCAACAGAAGCATTGACAAAACAGTCAACCTGTGAGTGTGAAAGTCCCGTCTGTGATCTGTATGAGCAGTCAAATATCCG  
 LeuAsnGluValIleSerAsnArgSerIleAspLysThrValAsnLeuSerAlaCysGluValAlaValLeuAspLeuTyrGluGlnSerAsnIleArg  
 -35 F<sub>3</sub> -10  
 1700  
 CATTCTAGTGACATCATCGAAGATTGGTTAACTCAACGTTTACAAAGTGAACAGGAAGTGTAAACTATAAGACACAGCGGACATCTGGAAATTG  
 IleProSerAspIleIleGluAspLeuValAsnGlnArgLeuGlnSerGluGlnGluValLeuAsnTyrIleGluThrGlnArgThrTyrTrpLysLeu  
 ORF-16.6 1800  
 GAGAATCAGAAAAAATAATCGGGGTCATTGAAATGAAGTTGCTTACACAGCTTTGCTACTACTGTAGCTTCTCATTCTTCACTCGAAAGTTGATGT  
 GluAsnGlnLysLysLeuTyrArgLysSerLeuLys \*LysLeuLeuThrHisValCysHisTyrCysSerPheSerPhePheThrArgLysPheAspVal  
 Met ORF-  
 1900  
 GTTTGGTCAATAACCAAGAAAGATACTCCTGTTGCTCTGTCACCTGTGGAATCAATCTCTTTTCAGTATCACATATCGAGGAGGAAATCAGATGA  
 PheGlyAlaIleThrLysLysAspThrProValValPheCysProThrCysGlyAsnGlnSerLeuSerValSerHisIleGluGluGluIleArg \*Asn  
 16.5 Met  
 2000  
 ATCAGAAAGAATTCCAAGCCGTTTTAGACTGGATGCTGTCAACCCACTATTATACAATCCATGAATACAATATATGCTACAAAAGAGCCTACCGTTCT  
 GlnLysGluPheGlnAlaValLeuAspTrpMetLeuSerProThrIleIleGlnPheHisGluTyrAsnTyrMetLeuGluLysSerLeuProPheLeu  
 T<sub>p</sub>3 2100  
 CAGACGATAGGCTTTTCTTTGATTTGTTCCACGTGAAACATTACTGTATACGCATCTTACGAAACAACCTATATGCGATATTTCTGATAACCTGATTGTC  
 ArgArg \* 2200  
 AAACCTCAGATAACCAATTCATGAACGCACTAGCTAACTACGTAAATGATAATTATTCTTATAGTTGGTTATCAACATCATGTTTTCATTAAGATCATCT  
 2216  
 GTTGCAATGTGTTATA

Figure 6. The Nucleotide Sequence of the Right Early Region.

**Figure 7. The Nucleotide Sequence of the Late Region.**

The inferred amino acid sequences of the putative open reading frames are also shown. The \* indicates a stop codon. The sus14(1241) and sus16(300) mutations (C to T transitions) are at 425 and 1798 bp respectively. The location of the putative B2 promoter -35 and -10 regions are as indicated.

(ORF-13)

TCGCTATATTACATGGGTAACCGTTCATGAAAGTCCCTTTGGCCCTTTGATGTGGTAAGAAGCTCAAAAAAGGCGATCTGATGGGACACACAGGTATCGGA 100  
 ArgTyrIleThrTrpValAsnValHisGluSerProLeuProPheAspValGlyLysLysLeuLysLysGlyAspLeuMETGlyHisThrGlyIleGly  
 200  
 GGAAACGTAACAGGCGACCAATTGGCACTTCAATGTTATTGACGGTAAGGAGTACCAAGGATGGACAAAGAAACCTGATTCGTGTTTAGCAGGGACAGAGT  
 GlyAsnValThrGlyAspHisTrpHisPheAsnValIleAspGlyLysGluTyrGlnGlyTrpThrLysLysProAspSerCysLeuAlaGlyThrGluLeu  
 300  
 TACACATATATGATGTTTTCGCTGTCAACAACGTTGGAGATAATCAACGGAAACGGCTACGACTGGAAAACCTAGTGATTGGCAAGACGGGACGGTGGGA  
 HisIleTyrAspValPheAlaValAsnAsnValGluIleIleAsnGlyAsnGlyTyrAspTrpLysThrSerAspTrpGlnAspGlyAspGlyGlyAsp  
 400  
 TGGCGACGACGACAACGATACAATAAAAAAGATTTAATAGCCCTTTACTATCTGACGCCCTCCATGGTTGGAAAAGCATAGAAAAGGAGAATGGGA  
 GlyAspAspAsnAspAsnAsnLysThrLysAspLeuIleAlaLeuLeuLeuSerAspAlaLeuHisGlyTrpLysAla \*  
 500  
 Gene-14  
 GATATGAAATGATAGCGTGGATGCAACACTTTTGTAGAGACAGACGAAACAAGCTTATTACTGGTTAACATTCCCTTATGGTTTGTATGGTTGTTGATA  
 METLysMETIleAlaTrpMETGlnHisPheLeuGluThrAspGluThrLysLeuIleTyrTrpLeuThrPheLeuMETValCysMETValValAspThr  
 600  
 CAGTTTGGGGTGTATTGCAAAAGCTTAAACCAACATTAATTTTTCATCATTTAAATCAAAAACAGGGGTGTGATTAAAGTCAGTAAATGATTCT  
 ValLeuGlyValLeuPheAlaLysLeuAsnProAsnIleLysPheSerSerPheLysIleLysThrGlyValLeuIleLysValSerGluMETIleLeu  
 700  
 AGCGTTATTGGCTATCCCTTCGCTGTCCCTTCCCTGCGGGTTTACCCTTATTATACACGGTTTATACGGCTTGTGTGTATCGGAAATATATTTCTATT  
 AlaLeuLeuAlaIleProPheAlaValProPheProAlaGlyLeuProLeuLeuTyrThrValTyrThrAlaLeuCysValSerGluIleTyrSerIle  
 800  
 TTCGGACATCTGAGATTAGTAGATGATAAAAGTATTCTTGAATACTTGAACACTCTTTAAGCGCACATCCGGTAAAAATAAGGAGGAAAAATAAC  
 PheGlyHisLeuArgLeuValAspAspLysSerAspPheLeuGluIleLeuLeuLysPhePheLysArgThrSerGlyLysAsnLysGluGluLys \*  
 900  
 Gene-15  
 ATGCAATTCACAAGCGGTATCAACTTAATTAAGAGCTTTAGGGTTTACAACCTGAAAGCATATAAAGCTGTTCCGACTGAGAAGCATTACACCATG  
 METGlnIleSerGlnAlaGlyIleAsnLeuIleLysSerPheGluGlyLeuGlnLeuLysAlaTyrLysAlaValProThrGluLysHisTyrThrIleGly  
 1000  
 GTTACCGTCAATTACGGTCCCGATGTTTACCTAGGCAGGTTATCACTGCTAAACAGGCTGAAGACATGTTGCGTGATGATGTCAGGCTTTTGTGGATGG  
 TyrGlyHisTyrGlySerAspValSerProArgGlnValIleThrAlaLysGlnAlaGluAspMETLeuArgAspAspValGlnAlaPheValAspGly  
 -18 -35  
 TAATTT B2 TTAGTT 1100  
 TGTAATAAAGCATTAAAGTATCTGTCCACCAAAATCAATTTGATGCACITGTCTCATTCGCTTACAACGTTGGGTTAGGGGCTTTCAGGTCTTCTTCT  
 ValAsnLysAlaLeuLysValSerValThrGlnAsnGlnPheAspAlaLeuValSerPheAlaTyrAsnValGlyLeuGlyAlaPheArgSerSerSer  
 1200  
 CTACTGGAATCTTGAATGAAGGAAGAACAGCTCTAGCGCGCGTGAATCCCTAAATGGAATAAGTCAGCGGTAAGTTTATCAAGGGTTGATTAACC  
 LeuLeuGluTyrLeuAsnGluGlyArgThrAlaLeuAlaAlaIleGluPheProLysTrpAsnLysSerGlyGlyLysValTyrGlnGlyLeuIleAsnArg  
 1300  
 GTAGAGCACAGGAGCAAGCCTTGTAAATAGTGAACACCTAAAATGTTTACCGTGAACATCGCTCTACTAAAACGACACCTAAGTAAAGGTGAAGAG  
 ArgAlaGlnGluGlnAlaLeuPheAsnSerGlyThrProLysAsnValSerArgGlyThrSerSerThrLysThrThrProLysTyrLysValLysSer  
 1400  
 TGGTGACAACCTTACTAAAATCGCTAAAAGCATAATAACAACGGTGTACTTTGTTGAAGTTGAATCCGAGTATCAAAGACCCGAACATGATTAGAGTT  
 GlyAspAsnLeuThrLysIleAlaLysLysHisAsnThrThrValAlaThrLeuLeuLysLeuAsnProSerIleLysAspProAsnMETIleArgVal  
 1500  
 GGCAACAATAAATGTTACAGGTAGCGGGGCAAAACACATAAAGTGAAGTGGTGACACACTCAGTAAAATGCGGTTGATAACAAAACGACTGTGA  
 GlyGlnThrIleAsnValThrGlySerGlyGlyLysThrHisLysValLysSerGlyAspThrLeuSerLysIleAlaValAspAsnLysThrThrValSer  
 Gene-16 1600  
 GTAGATTGATGAGTAAACCTGAAATACGAATCCAAATCATATAAAGTACGGTCAAACAATTAGATTAAAGTGAAGGTGTAATCATCGACAAGAGTT  
 ArgLeuMETSerLeuAsnProGluIleThrAsnProAsnIleLysValGlyGlnThrIleArgLeuSer \* METAspLysSerLeu  
 1700  
 TATTTTATAATCCACAGAAAATGTFATCATACGATCGCATACTGAACCTTGTATCGGTGCTCGTGGTATCGGTAATCATATGCAATGAAGGTGATACC  
 PheTyrAsnProGlnLysMETLeuSerTyrAspArgIleLeuAsnPheValIleGlyAlaArgGlyIleGlyLysSerTyrAlaMETLysValTyrPro  
 T  
 TATTAATCGCTTTATTAAGTACGGAGAACAATTCATATATGTTTCGACGATACAAAACCGGAGCTTGCAGAGGTCCTCAACTATTTTAAATGATGTAGCTCAA  
 IleAsnArgPheIleLysTyrGlyGluGlnPheIleTyrValArgArgTyrLysProGluLeuAlaLysValSerAsnTyrPheAsnAspValAlaGln  
 1900  
 GAATTCCTGACCATGAGTTGGTTGTGAAGGGTGAAGGTTTACATTGATGGTAAAGCTTGCAGGGTGGGCTATTCCTCTGAGTGTGTGGCAGAGTGAA  
 GluPheProAspHisGluLeuValValLysGlyArgArgPheTyrIleAspGlyLysLeuAlaGlyTrpAlaIleProLeuSerValTrpGlnSerGluLys  
 2000  
 AATCTAATGCATATCTCAACGTAAGCACAATAGTATTTGATGAGTTTATCAGGAGAAAAGACAATAGCAACTATATTCCTAATGAGGTTTCAGCTTACT  
 SerAsnAlaTyrProAsnValSerThrIleValPheAspGluPheIleArgGluLysAspAsnSerAsnTyrIleProAsnGluValSerAlaLeuLeu  
 2100  
 AAACCTTATGGACACCGTATTCGTAACCGTGAAGCGGTGTCAGATGCATTGTTTAAAGTAAATGCTGFATCCGTTGTTAACCCCTATTTTCTGTCTTCAAC  
 AsnLeuMETAspThrValPheArgAsnArgGluArgValArgCysIleCysLeuSerAsnAlaValSerValValAsnProTyrPheLeuPhePheAsn  
 2200  
 CTTGTCCTGATGTCAACAACCGTTCATGTATATGACGATGCTTAAATGAAATACCTGATAGTCTTGACTTCTCATCTGAAAGCGTAAAACAAGGT  
 LeuValProAspValAsnLysArgPheAsnValTyrAspAspAlaLeuIleGluIleProAspSerLeuAspPheSerSerGluArgArgLysThrArgPhe  
 2300  
 TTGGGCGGCTAATTGATGGAACCGAGTACGGTGAAGTGGTATTAGATAACCAAGTTTATCGGAGATAGTCAGGTGTTTATAGAAAAGCGCAGTAAGGATAG  
 GlyArgLeuIleAspGlyThrGluTyrGlyGluMETSerLeuAspAsnGlnPheIleGlyAspSerGlnValPheIleGluLysArgSerLysAspSer  
 2400  
 TAAGTTGATFCTCCATCGTCTAATGGATTCACTCTTGGTGTGGGTTGATGTTAATCAAGGCTTATGTACATTGATACAGCAGATGACCCGTC  
 LysPheValPheSerIleValTyrAsnGlyPheThrLeuGlyValTrpValAspValAsnGlnGlyLeuMETTyrIleAspThrAlaHisAspProSer  
 2500  
 ACTAAGAATGTATACACATTGACAACAGATGATCTTAATGAAAACATGATGTTGATAACCAACTATAAGAATAAATTCATTTACGTAAGTTAGCTAGTG  
 ThrLysAsnValTyrThrLeuThrThrAspAspLeuAsnGluAsnMETMETLeuIleThrAsnTyrLysAsnAsnTyrHisLeuArgLysLeuAlaSerAla  
 2586  
 CGTTCAATGAATGTTATCTGAGGTTTGACAATCAGGTTATCAGAAATATCGCATATGAGTGTGTTTCGTAAGATGCGTATACAGTAA  
 PheMETAsnGlyTyrLeuArgPheAspAsnGlnValIleArgAsnIleAlaTyrGluLeuPheArgLysMETArgIleGln \*

Figure 7. The Nucleotide Sequence of the Late Region.

### Coding Capacity

In Figs. 6 and 7 ten open reading frames that may code for viral gene products have been designated. These conclusions are based on the following criteria:

- 1) the presence of an ORF beginning with a GTG, TTG or ATG codon.
- 2) the presence of a Shine-Dalgarno sequence appropriately spaced upstream from the start codon.
- 3) previous biochemical and genetic data relating to genes 14, 15, 16 and 17. (Mellado et al., 1976; Murray and Rabinowitz, 1982; Yoshikawa et al., 1981.)
- 4) identification of ORF's by sequencing  $\phi$ 29 sus mutants.
- 5) the results of an E. coli in vitro transcription-translation study presented here.

The ORF's were initially chosen based upon the first three criteria and are designated as gene-17 (g17), ORF-16.9, ORF-16.8, ORF-16.7, ORF-16.6, ORF-16.5, g16, g15, and g14. Genes 14, 15, and 16 are late genes and the remaining genes (ORF's) are early genes. The right-most early translational initiation site has been previously identified as g17 (Yoshikawa and Ito, 1981; Murray and Rabinowitz, 1982) which is the only early gene to be genetically mapped in this region (Mellado et al., 1976). The ORF's designated g14 and g16 were identified by sequencing  $\phi$ 29 sus14(1241) and sus16(300) mutants (the sus16(300) mutation was sequenced by M. Saedi). The mutant sequences were distinguished by a C to T transition in a CAA codon, resulting in the

creation of ochre mutations (TAA). The positions of these transitions are indicated in Figs. 6 and 7.

The six putative early ORF's are organized in tandem, with at most a single nucleotide between the stop codon of one ORF and the start codon of the next (Fig. 6; Table 2). In fact at three junctions, the ORF-16.9:16.8, the ORF-16.7:16.6 and the 16.6:16.5 junctions, the termination codon of the preceding ORF overlaps with the start codon of the succeeding ORF (Table 2). In all three cases the structure is the same, ATGA. The late genes are also organized in tandem (Fig. 7), genes 14 and 15 are separated by a single nucleotide but genes 15 and 16 are separated by 10 nucleotides and ORF-13 and g14 are separated by 18 nucleotides.

#### Translation Initiation Regions

In Table 2, the translation initiation regions of the genes in this region are compared. The results of previously sequenced  $\phi 29$  genes are also included for comparison (Yoshikawa and Ito, 1981; Escaramis & Salas, 1982; Murray & Rabinowitz, 1982). All of the previously sequenced  $\phi 29$  genes (3, 4, 6 and 17) have low free energies of mRNA-rRNA interaction in the Shine-Dalgarno region (Shine and Dalgarno, 1974), as do the putative ORF's presented here. ORF-16.7 presents a slight problem because there are two regions of Shine-Dalgarno complementarity. The first, AAAGGGG, has a  $\Delta G$  of interaction of -16.6 kcal, but the spacing (13 b) is larger than those found previously for either Gram-positive or *E. coli* translation initiation regions (McLaughlin et al., 1981; Moran et al., 1982; Gold et al.,

Table 2.  $\phi$ 29 Translation Initiation Sites

Gene (ORF)	16S rRNA <sup>a</sup> UCUUUCCUCCACUAG	$\Delta G^b$ (kcal/mol)	Spacer (bp)
Gene-3	UGGUUGA <u>AAGGAG</u> AUAA <u>CGCAACA</u> <u>AUGGCGA</u> <sup>+1</sup>	-16.2	10
Gene-4	AUAAAC <u>AGGAGG</u> UAAAAU <u>UAGA</u> <u>AUGCCUA</u>	-18.8	9
Gene-6	AAAU <u>AGAAA</u> AGUGGGACGAAGAA <u>AUGGCAA</u>	-18.0	8
Gene-17	CCAAUCA <u>UAGGAGG</u> AAU <u>UACACA</u> <u>AUGAAUA</u>	-16.6	8
ORF-16.9	CUGA <u>AGAAGG</u> GAGAACA <u>AGUAACA</u> <u>AUGAGCG</u>	-14.2	10
ORF-16.8	UUG <u>AGGGGA</u> UGAAU <u>UUCACA</u> U <u>AUG</u> AUAG	-17.8	10
ORF-16.7a	UAAAAGGGGACA <u>AGGAG</u> AU <u>AAACA</u> <u>AUGGAAG</u>	-13.2	5
ORF-16.7b	UAAAAGGGGACA <u>AGGAG</u> AU <u>AAACA</u> <u>AUGGAAG</u>	-16.6	13
ORF-16.6	AACUAU <u>AUCGGGG</u> GUC <u>AUUGAAA</u> <u>AUG</u> AAGU	-16.2	7
ORF-16.5	CACAU <u>AUCGAGG</u> AGGAAA <u>UACAGA</u> <u>AUG</u> AAUC	-16.6	6
Gene-16	AGAU <u>UAAGU</u> UG <u>AGGUG</u> UAAA <u>UCA</u> <u>AUGGACA</u>	-18.2	6
Gene-15	AAAAU <u>AAGGAGG</u> AAAA <u>UAAACA</u> <u>AUG</u> CAAA	-16.6	9
Gene-14a	<u>AUAGAAA</u> AGGAGAA <u>UUGGG</u> GAGAU <u>AUG</u> AAAA	-15.4	10
Gene-14b	AAGGAGAA <u>UUGGG</u> GAGAU <u>AUG</u> AAAA <u>AUG</u> AUAG	-13.0	8

The sequences are aligned according their respective start codons; the first nucleotide of which is designated by a +1. The regions of complementarity to 16s rRNA are underlined as are the start codons. The stop codons of the preceding genes are overlined.

a. C. Woese, cited in McLaughlin et al., (1981).

b. Calculated by the rules of Tinoco et al. (1973).

1981). The other site, GACAAGGAG, has a slightly higher  $\Delta G$  of interaction (-13.2 kcal) but the spacing (5 b) is within previously determined values. Therefore, the latter would be the more likely ribosome binding site. Gene 14 has three ATG start codons and two potential Shine-Dalgarno sequences in the putative initiation region. The one designated Gene-14a in Table 2 was chosen as the most likely start region since the free energy value was lower than the alternative. However, either one or both may be able to serve as a translational start site. The results from the other putative ORF's are within previously determined free energy values (-11.6 to -21.0 kcal) and spacings (4-10 bp) found for Gram-positive Shine-Dalgarno regions (McLaughlin et al., 1981; Moran et al., 1982; see Stephens et al., 1984 for a correction of the translational start site of the 0.3 kb gene in Moran et al., 1982). Therefore, the putative ORF's in this region appear to satisfy the minimum, but not necessarily sufficient, structural requirements of a protein coding region.

#### In Vitro Protein Synthesis

In order to examine whether or not these ORF's specify proteins, in vitro transcription-translation was carried out using an E. coli system and  $\phi 29$  EcoRI-C fragment DNA as the template. The EcoRI-C fragment contains all the right early genes except ORF-16.5. Carrascosa et al. (1975) previously demonstrated that  $\phi 29$  DNA could be transcribed and translated by an E. coli in vitro system. The results of SDS-polyacrylamide electrophoresis of the in vitro synthesized products are shown in Fig. 8. In lane c, three major protein bands

whose synthesis was directed by  $\phi$ 29 EcoRI-C fragment DNA are seen: corresponding bands were found when whole  $\phi$ 29 DNA was used as a template (lane d). The estimated molecular weights of 21, 14 and 13 kilodaltons correlate well with the predicted molecular weights of 19.1, 15.2 and 12.6 (or 12.4) kilodaltons (Table 3). If it is assumed that the predicted 12.6 kd and 12.4 kd proteins would probably not be resolved by this gel system then possibly four of the five predicted proteins are synthesized in vitro. The only other putative protein missing is the 6.2 kd band. However, the marker proteins with molecular weights of 6.2 kd and 3 kd are also not seen; indicating that the gel system was unable to resolve very low molecular weight proteins. Alternative protein gel methods have been examined in order to resolve such small proteins but without success.

Murray and Rabinowitz (1982) used a B. subtilis derived in vitro transcription-translation system to study  $\phi$ 29 gene expression. In contrast to our results, the authors concluded that only one major protein was synthesized when  $\phi$ 29 EcoRI-C fragment was used to direct the reaction. This may be due to the differences of the respective systems used. The estimated molecular weight of the protein was 22.4 kd and was shown to be g17. Furthermore, Murray and Rabinowitz (1982) showed that this protein was encoded at the right end of the EcoRI-C fragment by demonstrating that, when EcoRI-C fragment was pre-digested with HaeII, which cuts just past the first ORF (Fig. 6), the 22.4 kd protein was still synthesized in vitro. This experiment was repeated using HinPI, which also cuts the HaeII site, and the E. coli in vitro

Table 3. Physicochemical Properties of the Putative Gene Products of the ø29 Right Early Region and Late Genes 14, 15 and 16.

	total residues	M <sub>r</sub> (kd)	% acidic residues	% basic residues	% hydrophobic residues <sup>a</sup>
Gene-17	166	19,231	24.1	13.2	33.6
ORF-16.9	108	12,634	11.1	10.2	37.9
ORF-16.8	106	12,389	14.1	19.8	39.6
ORF-16.7	130	15,168	15.4	13.8	37.7
ORF-16.6	54	6,192	9.3	16.7	35.2
ORF-16.5	37	4,612	8.1	13.5	45.9
Gene-16	331	38,920	11.2	12.9	40.6
Gene-15	258	28,022	7.0	16.0	29.0
Gene-14	131	15,014	9.9	12.2	38.9

a. Includes tyrosine, valine, leucine, isoleucine, phenylalanine, tryptophane and methionine.

Figure 8. Autoradiographs of L-[<sup>35</sup>S]Methionine-Labeled Proteins Synthesized in the *E. coli* Coupled Transcription-Translation System.

Conditions were as described in Materials and Methods.

Lane a. Protein standards (BRL, low  $M_r$  standards: insulin, 3 kd; bovine trypsin inhibitor, 6.2 kd; cytochrome C, 12.3 kd; lysozyme, 14.3 kd; lactoglobulin, 18.4 kd; chymotrypsinogen, 25.7 kd; ovalalbumin, 43.0 kd).

Lane b.  $\phi$ 29 EcoRI-C fragment digested with HinPI (0.5 ug/reaction). Lane c,  $\phi$ 29 EcoRI-C fragment (0.5 ug/reaction).

Lane d.  $\phi$ 29 DNA (1 ug/reaction).

Lane e. Control (without DNA template).

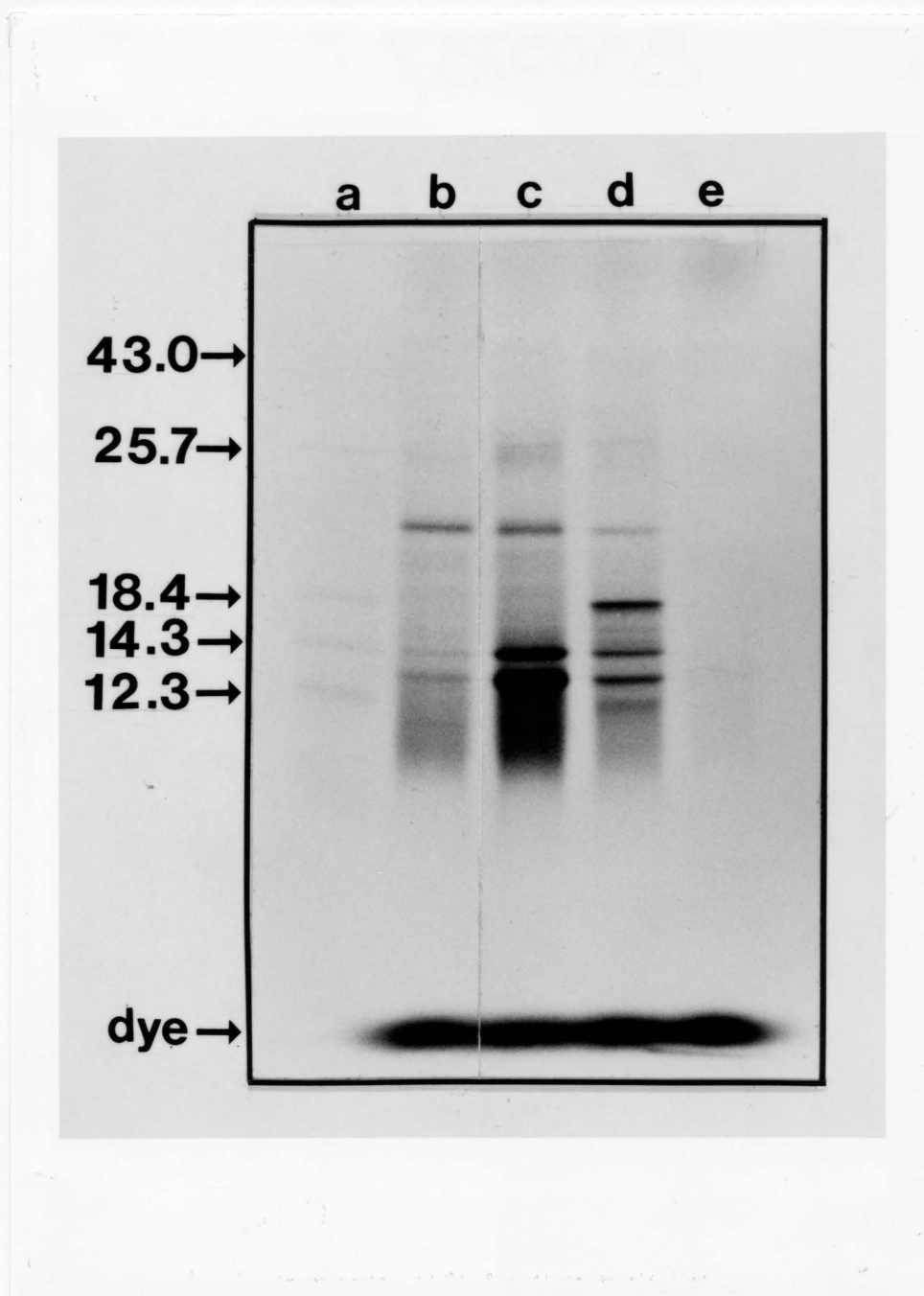


Figure 8. Autoradiographs of L-[<sup>35</sup>S]Methionine-Labeled Proteins Synthesized in the *E. coli* Coupled Transcription-Translation System.

protein synthesis system. In lane b, it can be seen that only one major band is present, at 21 kd, confirming the results of Murray and Rabinowitz (1982) and indicating that the other proteins are located further downstream from gene 17, as the sequence results predict.

There are faint bands at the 14 kd and 13 kd positions (lane b). Such residual bands were found in all the gels and could be due to undigested EcoRI-C fragment, although all the restriction digests were checked by agarose gel electrophoresis. Another possibility could be terminal transcription initiation which can occur during transcription of linear DNA molecules in vitro (Chen and Zubay, 1983). It is also noted that a faint band appears at 26 kd in all the lanes. The origin of this is not clear, but upon over exposure of the gel a similar band appears in the control lane lacking DNA and therefore could be due to an endogenous factor.

#### Transcription Signals

In the right early region (Fig.1), two B. subtilis RNA polymerase binding sites have been identified (Sogo et al., 1979<sub>a</sub>; Sogo, Lozano, and Salas, 1984). Their location was estimated to be at 91.3 ( $\pm 0.6$ ) (C1) and 99.3 ( $\pm 0.4$ ) (C2) map units (designated here as P<sub>(EC)3</sub> and P<sub>E3</sub> respectively); expressed as the percentage of the distance from the left end of the genome.

The P<sub>(EC)3</sub> promoter has been tentatively identified between 1640 bp and 1668 bp (Fig. 6). The putative -35 and -10 regions show homology with the E. coli and Bacillus consensus sequence (Fig. 9). The

## Left Early Promoters

$P_{E1}$       ATTAATGTTTTGACAACTATTACAGAGTATGCTATAATGGTAGTAT  
 $P_{E2}$       GAAAAGTGTTGAAAATTGTCGAACAGGGTGATATAATAAAAGAGT  
 $P_{(EC)1}$     AACTATATTTGACATCTGATAAGGAGGGTTATATCATGAAGCGTG  
 $P_{(EC)2}$     GGTTTTAATGGCATATGTTTCACCTCTTCTATAATCTATTAGTA

## Right Early Promoters

$P_{E3}$       GAAAAGGGTAGACAAACTATCGTTTAAACATGTTATACTATAATAG  
 $P_{(EC)3}$     AATCAACGTTTACAAAGTGAACAGGAAGTGTTAAACTATATAGAG  
 B2        GCATCAAATTGATTTTGGGTGACAGATACTTTTAATGCTTTATTT

Concensus <sup>a</sup>	TTGACA	TATAAT
	-35	-10

## Putative Early Sporulation Promoters

$P_R$       TAACAACTAAATCACGACTATATACCTATACTATTTATTATCATC  
 $P_L$       CTTTTATTAAAACCTTCTAAACTTTGTCGAACTTTTTTATAGAAA

Concensus <sup>b</sup>	AAATC	TANTGNTTNTA
	-35	-10

Figure 9. Early Promoters of  $\phi 29$ .  
<sup>a</sup> Moran et al. (1982). <sup>b</sup> Johnson et al. (1983).  $P_R$  and  $P_L$  refer to sequences at the right and left early regions respectively.

(16-18 bp) for E. coli and Bacillus promoters (Rosenberg and Court, 1979; Hawley and McClure, 1983; Murray and Rabinowitz, 1982; Moran et al., 1982). A search was made for further potential RNA polymerase binding sites but this was the only region that exhibited good -35 and -10 homology with the proper spacing.

Gene regulation in Bacillus is modulated by proteins that modify the promoter specificity of RNA polymerase (Losick and Pero, 1981). The right early region sequence was searched for homologies to such promoter consensus sequences. A region homologous to the Bacillus  $\sigma^{32}$  consensus sequence (Johnson, Moran, and Losick, 1983) was found just upstream from the  $P_{E3}$  promoter. A search of the regions upstream from the  $P_{E1}$  and  $P_{E2}$  revealed a similar sequence upstream from the  $P_{E2}$  promoter sequence. The results are compiled in Fig. 9.

In addition to the transcription of the  $\phi 29$  early coding regions, Sogo et al. (1979<sub>a</sub>) reported symmetrical transcription over most of the EcoRI-B fragment. The in vitro B. subtilis RNA polymerase binding sites were mapped, as were the in vivo transcripts. Both methods agree that the start site is at approximately 79 map units. This would be at about 900 bp in Fig. 7. A search of this region (+200 bp) revealed a sequence at 1013-1040 bp with a putative -35 and -10 regions that conform to the Bacillus  $\sigma^{55}$  consensus sequence (Fig. 9). A search was made for coding sequences both up and downstream from this putative promoter site. Two large ORF's were found upstream which could encode proteins >10 kd, but no ribosome

binding sites were found in either case. No large ORF's were found downstream.

Previous evidence suggests that both the right early transcripts and the late transcript(s) terminate in the EcoRI-D fragment (Sogo et al., 1979<sub>a</sub>; Davison et al., 1980; Sogo et al., 1984). In Fig. 10 the region between gene 16 and ORF-16.5, which appears to be the last right early region gene, has been diagrammed. These genes are separated by a 30 bp region containing two relatively large dyad symmetries. The one designated as T<sub>E3</sub> conforms to the consensus structure for an E. coli factor independent terminator (Holmes et al., 1983; Platt and Bear, 1983): that is, a potential hairpin loop followed by a poly-U sequence.

The potential secondary structures of both the early and late transcripts across this region are shown in Fig. 10. The primary differences between the structures are the increased stability of the late transcript secondary structure and the more extensive poly-U sequence of the early transcript. For either transcript the extent of the poly-U region can be increased but at the expense of hairpin stability. The late and early transcripts also differ in the position of the second dyad symmetry, shown in Fig. 10. This region would probably not form a stable hairpin structure but its position relative to T<sub>E3</sub> may affect termination.

#### Computer Analysis

A search was made of the latest version of the NBRF protein sequence library using the FASTP program of Lipman and Pearson (1985).

Figure 10. Features of the  $\phi$ 29 Early/Late Terminator Region.

A. The arrows designate the regions of dyad symmetry. T<sub>E3</sub> is the putative early terminator. Gene 16 and the last ORF of the right early region, ORF-16.5 are as designated.

B. Putative secondary structure of the early and late transcripts across the terminator region. The G's of interaction were calculated according to the rules of Tinoco et al. (1973). These structures do not necessarily represent those with the minimum free energy values.



This algorithm rapidly searches the data base for amino acid homology to an input sequence. The output provides a histogram of the alignment scores, the mean of these scores and the alignments if desired. Such a search can provide valuable information as to the function of new protein sequences, if homologies can be demonstrated. Unfortunately no such homologies could be found for any of the sequences presented here. Searches were done for Ktup 1 and 2.

A subsequent search of the literature revealed that a number of prokaryotic sequences were not contained in the NBRF library. These sequences were organized into a library. A search of this library revealed an extensive homology between ø29 gp15 and Salmonella phage P22 gp19, a lysozyme. The RDF program of Lipman and Pearson (1985) was used to evaluate the significance of this homology using a Monty Carlo analysis. The output displays a histogram of the shuffled scores, the initial and optimized scores and the z-value statistics where  $z = (\text{similarity score} - \text{mean of random score}) / (\text{standard deviation of random scores})$ . In other words, it calculates the number of standard deviations the score is from the randomized mean alignment scores. The guidelines for the significance of the z values are as follows:

- $z \geq 3$  possibly significant
- $z \geq 6$  probably significant
- $z \geq 10$  significant

The results of this analysis showed a very significant degree of homology between these two proteins. The optimized score was 218,

which was 28.21 standard deviations above the mean for the shuffled comparisons. Rennell and Poteete (1985) demonstrated that P22 lysozyme (gp19) exhibits a significant but limited homology with *E. coli* phage T4 lysozyme (gene product e; gpe). This analysis concurs with the above; the aligned score was 76 with a z-value of 6.64. A similar analysis of ø29 gp15 and T4 gpe suggests little homology between these proteins since the aligned score was 59 and the z-value was 3.72.

The FASTP program was used to determine amino acid alignments between these three proteins. ø29 gp15 and P22 gp19 share a 37.5% identity in a 144 residue overlap (Fig. 11). If conserved amino acid substitutions are considered, then the degree of homology very extensive. T4 gpe and P22 gp19 show a less extensive homology having only 18.2% identity in a 137 residue overlap (Fig. 12). ø29 gp15 and T4 gpe show the least homology; 26.8% identity in an 82 residue overlap (Fig. 13).

ø29 gp15 (28.0 kd) is considerably larger than either T4 gpe (18.7 kd) or P22 gp19 (16.1 kd) (Owen et al., 1983; Rennell and Poteete, 1985) and in the alignments only the amino end of gp15 is homologous; the carboxy end is not. A dot matrix analysis for direct repeats was done on the gp15 sequence. The results suggested the presence of direct repeats in the carboxy end of the molecule. In order to analyze the significance of this, the carboxy end was divided in two, and the sequences subjected to RDF analysis; FASTP was used to align the sequences. The FASTP alignment confirms that an extensive

37.5% identity in 144 aa overlap

```

      10      20      30      40      50
gp15  MQISQAGINLIKSFEGQLKAYKAVPTEKHYTIGY-GHY-GSDVSPRQVITAKQAEDMLR
      :::  .::  .:  ::  :::::  ..  .::  .:  :...  :::::
gp19  MQISSNGITRLKREEGERLKAYSDSRGIPTIGVGHTGKVDGNSVASGMTITAEKSSELLK
      10      20      30      40      50      60

      60      70      80      90      100     110
gp15  DDVQAEVDGVNKALKVSVTONQFDALVSFAYNVGLGAFRSSLLEYLNEGRTALAAAEFP
      .::  :...  :::::  :::::  .:  .::  .:  .:::  ::  .  :::::
gp19  EDLQWVEDAISSLVRVPLNQNYDALCSLIFNIGKSAFAGSTVLRQLNLKNYQAAADAF
      70      80      90      100     110     120

      120     130     140     150     160     170
gp15  KWNKSGGKVYQGLINRRAEQALFNSGTPKNVSRGTSSTKTPKYKVKSGDNLTKIAKKH
      .:  .:  .  .  ::  .:::  :
gp19  LWKK-AGKDPDILLPRRRRERALFLS
      130     140

      180     190     200     210     220     230
gp15  NTTVATLLKLNPSIKDPNMIRVGQTINVTGSGGKTHKVKSGDTLSKIAVDNKTTVSRLMS

      240     250
gp15  LNPEITNPNHIKVGQTIRLS

```

Figure 11. Fastp Alignments of ø29 gp15 and P22 gp19.  
 The ":" and "." indicate identical matches and conserved  
 substitutions respectively (Lipman and Pearson, 1985).

26.8% identity in 82 aa overlap

```

                                10      20      30      40      50
gp15      MQISQAGINLIKSFEGQLKAYKAVPTEKHHTIGYGHYGSQVSPRQVITA
                                :::
gpe      MNIFEMLRIDEGLRLKIYKDTTEGYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITK
                                10      20      30      40      50      60

                                60      70      80      90      100
gp15      KQAEMLRDDVQAFVDGVNKALKVSVTQNFDF-----ALVSFAYNVGLGAFRS--SSLLEY
..:.....::: : : . . :. . . . : : :.....: . . . . :
gpe      DEAEKLEFNQDVDAAVRGILRNAKLPVYDSLDAVRRCALINMVFQMGETGVAGFTNSLRM
                                70      80      90      100      110      120

                                110      120      130      140      150      160
gp15      LNEGRALAAAEPK--WNKSGGKVYQGLINRRAQEQALFNSGTPKNVSRGTSSTKTPK
:. . : . : : : : : :
gpe      LQQKRWDEAAVNLAKSRYNQTPNRAKRVITTFRTGTWDAYKNL
                                130      140      150      160

                                170      180      190      200      210      220
gp15      YKVKSGDNLTAKIAKKHNTTVATLLKLNPSIKDPNMIRVGQTIINVITGSGGKTHKVKSGDTL

                                230      240      250
gp15      SKIAVDNKTTVSRLMSLNPEITNPNIKVGQTIRLS

```

Figure 12. Fastp Alignments of ø29 gp15 and T4 gpe. The ":" and "." indicate identical matches and conserved substitutions respectively (Lipman and Pearson, 1985).

P22 gp19 vs T4 gpe

18.2% identity in 137 aa overlap

```

      10      20      30      40      50
gp19  MQISSNGITRLKREEGERLKAYSDSRGIPTIGVGH---TGKVDGNSVASGMTITAEKSS
      : ..... :. :. :. : : :. :. :. :. :. :. :. :. :. :. :. :. :.
gpe   MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNG
      10      20      30      40      50

      60      70      80      90      100     110
gp19  ELLKEDLQWV-EDAISSLVRVPLNQYDALCSLIENIGKSAFAGSTVLRQLNLKNYQAA
      . :. . . :. :. :. :. :. :. :. :. :. :. :. :. :. :. :. :. :.
gpe   VITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRCAL--INMVFQMGETGVAGF
      60      70      80      90      100     110

      120     130     140
gp19  ADAF-LLWKKAGKDPDILLPRRRRERALFLS
      .... :. :. :. :. :. :. :. :. :. :. :. :. :. :. :. :. :.
gpe   TNSLRMLQQKRWDEAAVNLAKSRYNQTPNRAKRVITTFRTGTWDAYKNL
      120     130     140     150     160

```

Figure 13. Fastp Alignments of P22 gp19 and T4 gpe.  
 The ":" and "." indicate identical matches and conserved  
 substitutions respectively (Lipman and Pearson, 1985).

60.5% identity in 43 aa overlap

```

      *           *           *           *           *
156-STKTPPKYKVKSGDNLTAKKHNTTVATLLKLNPSIKDPNMIRVGQTINV-206
      :           :           :           :           :
207-TGSGGKTHKVKSGDTLSKIAVDNKTTVSRMLSLNPEITNPNHIKVGQTIRLS-258
      *           *           *           *           *

```

Figure 14. Comparison of ø29 gp15 Carboxy Terminal Duplications. The ":" and "." indicate identical matches and conserved substitutions respectively (Lipman and Pearson, 1985).

direct repeat does exist (Fig. 14) and the RDF analysis ( $z=16.1$ ) suggests that the alignment is probably significant. The RDF value must be taken with caution since to examine just segments that have homology may bias the results.

#### Complementation of Lysis Defective T4 Phages

Rennell and Poteete (1985) demonstrated that cloned T4 gene e could complement P22 am19 mutant infections. This experiment was repeated here using  $\phi$ 29 g15 cloned into the  $P_L$  promoter plasmid pPLc45 (Fig. 15). Plasmid pMS2 contained the gene cloned in the proper orientation for expression by the  $P_L$  promoter. Plasmid pMS6 contained the same DNA fragment but cloned in the opposite direction. pMS2 when induced overproduces a protein with an approximate molecular weight of 26,000 which is the approximate predicted size of gene 15 (Table 3); pMS6 does not (M. Saedi, unpublished results).

Complementation was determined using spot tests of the defective T4 phage. In Table 4 it can be seen that the T4 amber and deletion mutants of gene e were complemented by clones pMS2 and pMS6 but not by the parent plasmid pPLc45. These studies were conducted at 30 C because pMS2 did not grow well at 42 C. Since both clones complemented the T4 infections and complementation occurred under non-inducing conditions, it appears that transcription of g15 occurs independent of the  $P_L$  promoter. This initiation of transcription is probably not plasmid mediated since complementation occurs in both orientations therefore the promoter is probably on the cloned fragment. However, an analysis of the sequence upstream of g15 did

not reveal any obvious candidates for a promoter site. The plasmids themselves have not been sequenced to detect promoter sequences arising from sequence changes.

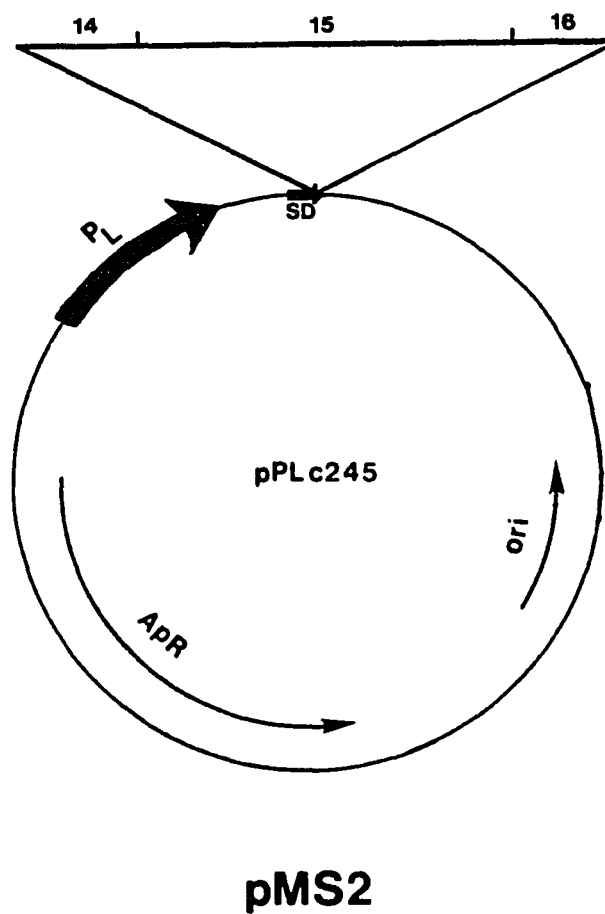


Figure 15. Construction of pMS2 (M. Saedi, unpublished results). Plasmid pMS6 has the insert in the opposite direction. pPLc245 was constructed by Remaut, Stanssens and Fiers (1983).

Table 4. Complementation of T4 Mutant Infection<sup>a</sup>

Strain (Plasmid)	Phage			
	am882	amH26x5	eG79	WT
ER69 (none)	-	-	-	+
ER69 (pPLc45)	-	-	-	+
ER69 (pMS2)	+	+	+	+
ER69 (pMS6)	+	+	+	+

a. Plasmid containing strains were spotted with the indicated phage and lysis was noted. Plating efficiencies were determined and in all cases were approximately 1.0 relative to CR63 (supD60) host cells.

## CHAPTER 4

## DISCUSSION

Overview

The sequence of the rightmost 4,626 bp of the  $\phi$ 29 genome has been presented. The identities of three of the late genes and evidence for an additional upstream ORF (ORF-13) has been established. The right early region appears to be composed of two overlapping transcriptional units composed of 6 and 2 ORF's each. However, only one of these ORF's (g17) has been identified as authentic. The transcriptional and translational signals present in this sequence are consistent with established signal sequences from E. coli and Bacillus (Rosenberg and Court, 1979; Hawley and McClure, 1983; Murray and Rabinowitz, 1982; Moran et al., 1982). Finally, data have been presented that strongly suggest gp15 functions as a lysozyme and can complement a lysozyme defective T4 phage infection.

Genetic Organization

The ten genes (ORF's) presented here are separated into two temporal classes, late and early. As predicted from earlier studies these genes are transcribed from opposite DNA strands and are oriented such that the transcripts converge (Loskutoff et al., 1973; Schachtele et al., 1973; Kawamura and Ito, 1977; Sogo et al., 1979<sub>a</sub>). The two gene classes do not overlap and are separated by a 30 bp

intercistronic region (Fig. 10). The reading frames are compactly organized in tandem. In most cases, the stop codon of the preceding ORF overlaps the start codon of the succeeding ORF or at most a single codon separates the start and stop codons (Table 2). This economic mode of genetic organization has been observed previously in phages  $\lambda$ , T7, G4 and  $\phi$ X174 (Sanger et al., 1982; Dunn and Studier, 1983; Godson et al., 1978; Sanger et al., 1978). The reasons for the evolution of such a concise genetic organization may be many fold; including constraints placed on the size of the genome by the size of the virion; it may serve to maximize information capacity while enhancing replication; or finally, as has been demonstrated so clearly in the E. coli trp operon, it may serve to translationally couple the termination of synthesis of one protein with the initiation of another (Oppenheim and Yanofski, 1980). Translational coupling has also been observed in Bacillus using artificial constructs (Sprengel, Reiss, and Schaller, 1985). In fact, the ribosome binding site, thought to be essential in Bacillus (Band and Henner, 1984; McLaughlin et al. 1981), of the downstream gene could be completely eliminated.

#### Translation Initiation Sites

Translation initiation sites, in general, feature a region of complementary to 16S rRNA (the Shine-Dalgarno sequence) spaced 4-10 bp upstream from the start codon (for reviews see Gold et al. 1981 and Kozak, 1983). This applies to initiation regions from Gram-positive organisms as well as E. coli (McLaughlin et al., 1981; Moran et al., 1982). There is however accumulating evidence that efficient

translation by Gram-positive organisms requires a more extensive Shine-Dalgarno complementarity (Band and Henner, 1984; McLaughlin et al., 1981; Moran et al., 1982). Table 2 clearly demonstrates that the sequences of  $\phi$ 29 translational initiation regions are consistent with this hypothesis.

The requirement by Bacillus for a strong Shine-Dalgarno region has been proposed to be one of the barriers to efficient translation of E. coli sequences in Bacillus. However, recent results indicate that this requirement is necessary but not sufficient for efficient translation of T7 mRNA, as compared to  $\phi$ 29 mRNA (Hager and Rabinowitz, 1985). Therefore, additional sequences are required for efficient translation in Bacillus and the translational barrier to heterologous gene expression in Bacillus is more complex than originally proposed. In more practical terms, this would limit the utility of Bacillus as a general cloning vector. Perhaps continued examination of  $\phi$ 29 translational initiation sites will provide additional insights.

#### Transcription Signals

In the right early region, Davison et al. (1980) found that the  $P_{E3}$  promoter (G2 transcript, 2000 b; the G represents the initiating nucleotide) was used efficiently by B. subtilis RNA polymerase in vitro but the  $P_{(EC)3}$  promoter (G5 transcript; 330 b) was only used at high RNA polymerase/DNA ratios. However, the  $P_{(EC)3}$  promoter was used efficiently by E. coli RNA polymerase resulting in the synthesis of one large transcript (G2), initiated at  $P_{E3}$ , and two small transcripts,  $G5_a$  (200 b) and  $G5_b$  (180 b) initiated at  $P_{(EC)3}$ , when  $\phi$ 29

EcoRI-C fragment was used as a template. By contrast, Sogo et al. (1984) observed that the P<sub>(EC)3</sub> promoter is efficiently utilized by B. subtilis RNA polymerase in vitro; and in fact, concluded that it was a stronger promoter than the P<sub>E3</sub> promoter.

Previously sequenced  $\phi 29$  early promoter regions have exhibited strong homologies with the E. coli consensus sequence (Yoshikawa and Ito, 1981; Yoshikawa et al., 1981; Murray and Rabinowitz, 1982). Thus suggesting that,  $\phi 29$  early transcription is initiated by the B. subtilis  $\sigma^{55}$ -RNA polymerase complex (Moran et al., 1982). The P<sub>E3</sub> promoter region has been published previously, (Yoshikawa and Ito, 1981) but the sequence of the P<sub>(EC)3</sub> promoter region was unknown until now. If the P<sub>(EC)3</sub> transcript were initiated at the first G residue after the designated -10 region, then the run-off RNA from a  $\phi 29$  EcoRI-C fragment template would be 240 bp long. Which is reasonably close to the estimated 200 (180) b transcript observed by Davison et al. (1980). The lack of further potential RNA polymerase binding sites supports the contention of Davison et al. (1980) that the size difference of the G5<sub>a</sub> and G5<sub>b</sub> transcripts was due to nonspecific binding of the RNA polymerase to the DNA terminus, causing premature termination of transcription. The importance of this promoter in vivo is unknown, but if utilized, it would allow transcription of ORF-16.6 and ORF-16.5 (Fig. 6). Subsequent to completion of this sequence, Dobinson and Spiegelman (1985) confirmed the position of P<sub>(EC)3</sub> by S1 mapping and provided evidence that it is used in vivo.

The results presented here suggest that  $\phi 29$  transcription may be regulated by  $\sigma^{32}$  as well as by the  $\sigma^{55}$  RNA polymerase complexes. The  $\sigma^{32}$  factor is found in late log to stationary phase vegetative cells and regulates expression of two sporulation genes of unknown function (Johnson et al., 1983). There is evidence that  $\phi 29$  transcription is turned off in sporulating cells (Kawamura and Ito, 1975) but, to my knowledge, no studies have been done to investigate  $\phi 29$  transcription in early stationary to stationary vegetatively growing cells. Therefore, the possibility of regulation of  $\phi 29$  gene expression by  $\sigma^{32}$  is very tentative but certainly worthy of further investigation.

The genetic map of  $\phi 29$  clearly defines the early and late coding regions (Fig. 1), however Sogo et al. (1979<sub>a</sub>) observed that early transcripts from the late region were found in vivo. Two Bacillus promoter sites (B1 and B2; Fig 1) were mapped in this region and were active in vitro (Sogo et al., 1979<sub>a</sub>; Sogo et al., 1984). The putative sequence of the B2 site is provided here and is consistent with previously sequenced  $\phi 29$  promoter regions (Fig. 9). Early transcripts from this region do not appear to be synthesized late in infection, unlike other early transcripts (Sogo et al., 1979<sub>a</sub>). The existence of the two Bacillus promoters suggests that anti-late RNA synthesis is not due to initiation errors or to read-through from the right early region (Sogo et al., 1979<sub>a</sub>).

The biological significance of symmetric transcription in  $\phi 29$  is not clear, particularly since no known early genes map in this

region and none were found by sequence analysis. However, anti-sense RNA has been shown to serve a regulatory role in ColE1 plasmid replication (Tomizawa, 1984) and in regulating translation of ompF mRNA in *E. coli* (Coleman, Green, and Inouye, 1984) and Tn10 transposase mRNA (Simons and Kleckner, 1983). Perhaps  $\phi$ 29 anti-sense RNA serves a similar role since the late proteins are differentially expressed and at least one, gp8, is expressed early (4 min.) in infection (Hawley et al. 1973).

Previous evidence suggests that both the right early transcript and the late transcript terminate in the EcoRI-D fragment (Sogo et al., 1979<sub>a</sub>; Davison et al., 1980; and Sogo et al., 1984). The early terminator ( $T_{E3}$ ) appears to be relatively inefficient, both in vivo and in vitro (Sogo et al., 1979<sub>a</sub>; Davison et al., 1980). The late terminator appears to be very efficient since no late mRNA hybridizes with the downstream EcoRI-C fragment (Kawamura and Ito, 1977).

Based upon the sequence and potential secondary structure of transcripts from the gene-16:ORF16.5 intercistronic region it is possible that the same sequences may act to terminate both the late and early transcripts (Fig. 10). This is not unprecedented, bidirectional terminators have been found in *E. coli* (tonB-p14 junction) (Postle and Good, 1985) and the *E. coli* transposon, Tn10, (tetA-orfL junction) (Schollmeir, Gartner, and Hillen, 1985). In both cases the genes are separated by a short intercistronic region, exhibiting potential secondary structures. An alternative hypothesis is that either one or both of these transcripts may terminate via a factor dependent

mechanism. The observation that termination of  $\phi 29$  transcripts in vitro is affected by a purified Bacillus "rho" factor supports this proposal (Hwang and Doi, 1980). However, either of these proposals are based upon sequence analysis and therefore are predictive only and require biological verification.

The above analysis is limited by the lack of knowledge of Bacillus terminators. Several researchers have identified putative Bacillus terminators (based on correlations with E. coli) but few have been identified as such, but, to my knowledge, not a single terminator has been studied in molecular detail. It is within this context that  $\phi 29$  assumes its importance; particularly since the transcription and genetic maps have been so well established.

#### Early Gene Functions

The right early region appears to be organized into two transcriptional units (Sogo et al., 1979<sub>a</sub>; Davison et al., 1980; Dobinson and Spiegelman, 1985; Garvey, et al. 1985<sub>b</sub>) (Fig. 1). The largest encompasses gene 17 through ORF 16.5; the smallest would transcribe only ORF 16.6 and ORF 16.5. The ORF encoding g17 has been identified (McLaughlin et al., 1981; Yoshikawa and Ito, 1981; Garvey et al., 1985<sub>b</sub>) but the other ORF's are putative. The biological efficacy of at least two of these ORF's was supported by in vitro protein synthesis experiments (Fig. 8). Hagen et al. (1976) and Anderson and Reilly (1976) discovered at least 4 early phage proteins, LM4 (15.2 kd), LM5 (14.7 kd), LM6 (13.0 kd) and LM6B (8.5-13 kd) and one small protein

LM10 (<4.5 kd) that have not been given genetic assignments. Therefore some of these proteins could be encoded by the right early region.

The functions of the early genes (ORF's) are unknown but gp17 is of particular interest, since it has been implicated in the DNA replication process, although the results are not unequivocal (Harding and Ito, 1976; Carrascosa et al., 1976; Hagen et al., 1976; Jimenez et al., 1977). Hawley et al. (1973) have reported the loss of two proteins, A1 (22.4 kd) and A2 (21.8 kd) from sus17 infected cells. Revertants regained expression of both proteins, suggesting that the observed electrophoretic differences were due to post-translational modification. This is supported by the existence of only one ORF in this region encoding a protein greater than 19 kd (Table 2). The proteins are produced in large amounts early during infection (2 min.) and in reduced quantities after the onset of DNA replication (Hawley et al., 1973; Hagen et al., 1976; Jimenez et al., 1977).

If gp17 is involved in DNA replication, it could have any number of functions.  $\phi$ 29 DNA replication in vitro requires only terminal protein (gp3), DNA polymerase (gp2) and  $\phi$ 29 DNA-gp3 complex (Watabe, et al., 1983; Watabe, et al., 1984; Prieto, et al., 1984; Blanco and Salas, 1984; Blanco and Salas, 1985). This system has proved inefficient and probably requires additional enzymes or accessory proteins. Two possible activities are suggested by the Adenovirus in vitro DNA synthesis system which requires a single-strand DNA binding protein and topoisomerase (Kaplan et al., 1979; Stillman, White, and Grodzicker, 1984; Nagata, Guggenheimer, and Hurwitz, 1983).

Several other possibilities are: helicase, double stranded DNA binding protein, exonuclease activity, an accessory role in replication or a variety of other functions that may regulate expression of genes required for replication.

Mutants in the remaining right early ORF's have never been genetically mapped (Mellado et al., 1976; Reilly et al., 1977) but the existence of  $P_E3$  and  $P_{(EC)3}$  and the size of right early transcripts (Sogo et al., 1979<sub>a</sub>; Sogo et al., 1984; Davison et al., 1980; Kawamura and Ito, 1977; Dobinson and Speigelman, 1985) suggest that they are expressed. Perhaps these genes are non-essential under the conditions used for mutant isolation: supported by the observation that, mutants unable to synthesize certain  $\phi 29$  encoded low molecular weight proteins, can exhibit plaque formation (Anderson and Reilly, 1976). As to the function of these proteins, several lines of evidence are suggestive. Involvement in the replication, as discussed above, is one possibility. Another, may be involvement in RNA processing. Kawamura and Ito (1977) found at least 13 major early transcripts (three in the right early region) and observed that small transcripts were more stable than larger ones. They proposed that this may be due to processing of the RNA. Anti-late mRNA synthesis also appears to be regulated, since its transcription is different from the other early mRNA's (Sogo et al., 1979<sub>a</sub>). Finally, there is evidence that genes 3 17 and 12 are post-translationally modified and that other early proteins are differentially expressed during infection; although early mRNA synthesis continues through-out infection (Hawley et al., 1973;

Hagen et al., 1976; Anderson and Reilly, 1976). Roles in protein processing and regulation of translation are also to be considered.

Several predictions can be made based on the amino acid content (Table 3) concerning the properties of these early proteins. The most notable feature of gp17 (19,231 d) is the high acidic residue content (24% of the total residues) as opposed to the number of basic residues (13.2%). Most of these acidic residues are found in the central portion of the molecule. Within this region, from residue 64 to residue 99, 14 of the 25 residues are glutamate or aspartate, no basic residues are found in this sequence. Therefore gp17 is probably a very acidic protein, lacking cysteine and tryptophane.

Although gp16.9 and gp16.8 (12,634 and 12,389 d respectively) are very similar in size they can be differentiated by the lack of cysteine in gp16.8 and the high phenylalanine content of gp16.9 (9.3%). In addition, gp16.8 may be a basic protein (20.6% basic vs 14.7% acidic residues) whereas, gp16.9 has approximately equal numbers of basic and acid amino acids. A comparison of the amino acid sequences of these two proteins found no extensive homologies.

gp16.7 (15,168 d) contains a relatively large number of hydrophobic residues (37.6%) and approximately equal numbers of acid and basic residues. The amino terminus of this protein is very hydrophobic, as are the amino termini of gp16.9 and gp16.6; reminiscent of signal peptides (Michaelis and Beckwith, 1982).

gp16.6 (6,192 d) should be a basic polypeptide lacking tryptophane, but having a large number of phenylalanine residues (11%).

Most of these phenylalanine residues are found between residues 12 and 24; 5 of the 11 amino acids are phenylalanine, the significance of this is not known.

gp16.5 (4,612 d) is expected to be the smallest of all these proteins. It will be lacking in glycine but having a relatively large proportion of methionine residues. This protein may also be the most hydrophobic of these proteins since 45.9% of its amino acids have hydrophobic side groups.

These predictions must, of course, be considered tentative, particularly the predictions concerning the basic natures of gp 16.7 and gp16.6 since the tertiary structure of the protein determines which amino acids are exposed to the medium. These predictions may however be useful for further identification and characterization of these proteins.

#### Late Gene Functions

ø29 g16 encodes a late non-structural protein (36 kd) shown to catalyze the genome encapsidation reaction (Bjornsti et al., 1984). The gene has been recently cloned (Bjornsti et al., 1984; Guo et al., 1986) and localized to the EcoRI-D and -E fragments and is presumably the last gene in the late region (Carrascosa et al., 1976). The results provided here are consistent with all these data (Fig. 7; Table 3). The encapsidation reaction in vitro is highly efficient and requires only proheads, DNA-gp3 and gp16 (Bjornsti et al., 1984; Guo et al., 1986). Functional DNA-gp3 is required and packaging of the DNA-gp3 occurs on discrete lengths, with the left end of genome

entering the prohead first (Bjornsti et al., 1982; Bjornsti et al., 1984). In conjunction with the head filling process, the scaffolding protein (gp7) exits the prohead and a conversion of the rounded prohead to the angular morphology of the mature phage occurs (Bjornsti et al., 1983). This entire cascade of events, resulting in the condensation of the DNA into the phage head is catalyzed by just one protein, gp16 (Bjornsti et al., 1984). Now that this gene has been cloned and purified (Guo et al., 1986) and sequenced (Garvey et al., 1985<sub>a</sub>), its functions may be studied in greater detail.

Gene 15 encodes a late non-structural protein with an estimated molecular weight of about 26 kd and is synthesized in large amounts (Carrascosa et al., 1976; Hawley et al., 1973; Hagen et al., 1976; Jimenez et al., 1977). The gene product is not present in mature phage but has been associated, in small amounts, with DNA free proheads from cells infected with mutants in genes 9, 11, 12, and 13 (Hagen et al., 1976; Jimenez et al., 1977). The phenotype of sus15 or sus14sus15 double mutants is that of delayed lysis. Mature and proheads are produced as in a sus14 infection, indicating that morphogenesis is complete. It has also been observed that the necessity of gp15 may depend on the growth conditions (Jimenez et al., 1977). Based on these data it was concluded that gp15 plays a role in morphogenesis.

The results presented here (Fig. 7; Table 3) are consistent with these data in relation to the size (28 kd) and location of gp15 and with the delayed lysis phenotype, but are not consistent with the

morphogenic factor hypothesis.  $\phi$ 29 gp15 exhibits a highly significant homology with the lambdoid Salmonella phage protein, P22 gp19 (Fig. 11) and considerably less homology with E. coli phage protein T4 gpe (Fig. 12). These two proteins, T4 gpe and P22 gp19, encode lysozymes and, as confirmed here, share significant homologies themselves (Owen et al., 1983; Weaver et al., 1984; Rennell and Poteete, 1985) (Fig. 13).  $\phi$ 29 gp15 has been cloned into an expression vector and cell lysing activity was found under inducing conditions (M. Saedi, unpublished results). This clone (Fig. 15) was used here (Table 4) to demonstrate that it could complement T4 infections mutant in gene e, again consistent with the predicted lysozyme function of  $\phi$ 29 gp15. These results do not, however eliminate the possibility of a dual function for gp15.

This is the first lysozyme from a Gram-positive system to be sequenced. Only two other phage lysozymes, T4 ge and P22 g19, have been sequenced (Owen et al., 1983; Rennell and Poteete, 1985), ignoring T2 lysozyme which differs from T4 ge by 3 residues (Tsugita, 1971).  $\phi$ 29 gp15 is by far the largest of the three lysozymes and the entire extent of its homology with the other two proteins is within the amino portion of gp15. The non-homologous carboxy terminus is composed of two direct repeats (Fig. 14). Structural comparisons of T4 lysozyme with hen egg white lysozyme indicated that, despite a lack of sequence homology, the two proteins share a significant structural homology and probably diverged from a common precursor (Mathews et al., 1981). Similar analysis may provide insights as to structural

relatedness of  $\phi$ 29 gp15, P22 gp19 and T4 gene e product, particularly since they share limited homology but  $\phi$ 29 gp15 is so much larger than T4 gene e at the carboxy terminus.

The three phage lysozymes genes are obviously interrelated but how this arose, considering they infect different bacterial genera, is another matter. This can be considered in the context of viral origins in general. Viruses can be viewed as arising from degenerate cellular parasites or as having evolved from genetic components of the host (Luria et al., 1978). The latter hypothesis is preferred to explain the origins of lambdoid phages (Campbell and Botstein, 1983). According to this hypothesis, various functions such as replication, recombination, assembly, and lysis were derived, in chimeric fashion, from analogous cellular components. Therefore, the genomes of  $\phi$ 29, T4 and P22 would be composed of evolutionarily and functionally distinct modules. Accordingly, these phages could have independently acquired the "lysozyme gene" from their respective hosts (genes from Bacillus and E. coli have demonstrated considerable homology). Therefore, the relatedness of the genes may reflect the evolutionary interrelationships of the hosts and not necessarily those of the phages.

The phages T4 (t,e),  $\lambda$  (S,R-RZ) and P22 (13,19) all have at least two lysis functions. The first (t,S,13) are thought to disrupt the inner membrane (Reader and Siminovitch 1971; Josslin, 1970; Rennell and Poteete, 1985) allowing access of a second protein (e,R-RZ,19) to the cell wall which is degraded (Tsugita et al., 1968; Rao and Burma, 1971; Bienkowska-Szewczyk and Taylor, 1980).  $\phi$ 29 also

encodes at least two lysis functions, g14 and g15. Gene 15 has been demonstrated to be a probable lysozyme and, by analogy, g14 would encode a protein that disrupts the cell membrane. However, no homology could be found between ø29 gp14 and P22 gp13 or gpS. There may also be a third function involved in ø29 lysis since sus14sus15 mutants do exhibit lysis (Jimenez et al., 1977) and a spontaneous deletion mutant of ø29, shown to be deleted in g14 and g15, forms small plaques (M. Saedi, unpublished results). This proposed third function may be either phage or host encoded and requires further elucidation.

#### Conclusions

The sequence analysis presented in this paper provides a firm structural bases for further analysis of the ø29 right early region genes and three late genes, their organization and their transcriptional and translational regulatory sequences. More specifically, this analysis reveals the primary sequences for gp17, a protein that may be required for ø29 DNA replication, and for five previously undetected genes in the early region. In addition, the primary sequences of gp16, which catalyzes the genomes encapsidation reaction, gp14, a lysis gene, and gp15, which is a lysozyme, are presented. Although the transcriptional and translational regulatory sequences must be considered putative at this point, the data are consistent with previous analyses of Bacillus sequences and facilitate a more complete examination of these regions.

## LIST OF REFERENCES

- Anderson, D.L., Hickman, D.D. and Reilly, B.E. (1966). *J. Bacteriol.* 91, 2081-2089.
- Anderson, D.L. and Reilly, B.E. (1974). *J. Virol.* 13, 211-221.
- Anderson, D.L. and Reilly, B.E. (1976). In "Microbiology, 1976" (Schlessinger, D., ed.). pp. 254-274. American Society for Microbiology, Washington, D.C.
- Band, L. and Henner, D.J. (1984). *DNA* 3, 17-21.
- Bienkowska-Szewczyk, K. and Taylor, A. (1980). *Bioc. Biop. Acta* 615, 489-496.
- Bjornsti, M.A., Reilly, B.E. and Anderson, D.L. (1981). *Proc. Natl. Acad. Sci. USA* 78, 5861-5865.
- Bjornsti, M.A., Reilly, B.E. and Anderson, D.L. (1982). *J. Virol.* 41, 508-517.
- Bjornsti, M.A., Reilly, B.E. and Anderson, D.L. (1983). *J. Virol.* 45, 383-396.
- Bjornsti, M.A., Reilly, B.E. and Anderson, D.L. (1984). *J. Virol.* 50, 766-772.
- Bjornsti, M.A., Reilly, B.E. and Anderson, D.L. (1985). *J. Virol.* 53, 858-861.
- Blanco, L., Garcia, J.A., Penalva, M.A. and Salas, M. (1983). *Nuc. Acids Res.* 11, 1309-1323.
- Blanco, L. and Salas, M. (1984). *Proc. Natl. Acad. Sci. USA* 81, 5325-5329.
- Blanco, L. and Salas, M. (1985). *Proc. Natl. Acad. Sci. USA* 82, 6404-6408.
- Campbell, A. and Botstein, D. (1983). In "Lambda II" (Hendrix, R.W., Roberts, J.W., Stahl, F.W. and Weisberg, R.A., eds.). pp. 365-380. Cold Spring Harbor Laboratory, New York.

- Carrascosa, J.L., Camacho, A., Moreno, F., Jimenez, F., Mellado, R.P., Vinuela, E. and Salas, M. (1976). *Eur. J. Biochem* 66, 229-241.
- Carrascosa, J.L., Jimenez, F., Vinuela, E. and Salas, M. (1975). *Eur. J. Biochem.* 51, 587-591
- Carrascosa, J.L., Mendez, E., Corral, J., Rubio, V., Ramirez, G., Salas, M. and Vinuela, E. (1981). *Virology* 111, 401-413.
- Cavalier-Smith, T. (1974). *Nature* 250, 467-470.
- Chen, H. and Zubay, G. (1983). In "Methods in Enzymology" (Wu, R., Grossman, L. and Moldave, K., eds.). pp. 674-960. Academic Press, New York.
- Coleman, J., Green, P.J. and Inouye, M. (1984). *Cell* 37, 429-436.
- Davison, B.L., Murray, C.L. and Rabinowitz, J.C. (1980). *J. Biol. Chem.* 255, 8819-8830.
- Desiderio, S.V. and Kelly, T.J. (1981). *J. Mol. Biol.* 145, 319-337.
- Dobinson, K.F. and Spiegelman, G.B. (1985). *J. Biol. Chem.* 260, 5950-5955.
- Dunn, J.J. and Studier, F.W. (1983). *J. Mol. Biol.* 166, 477-535.
- Escarmis, C. and Salas, M. (1982). *Nuc. Acids Res.* 10, 5785-5789.
- Estaban, M., Flores, L. and Holowczak, J.A. (1977). *Virology* 83, 467-73.
- Ganesan, A.T. and Hoch, J.A. (1984). "Genetics and Biotechnology of Bacilli" pp. 403-405. Academic Press, New York.
- Garvey, K.J., Saedi, M.S. and Ito, J. (1985<sub>a</sub>). *Gene* 40, 311-316.
- Garvey, K.J., Yoshikawa, H. and Ito, J. (1985<sub>b</sub>). *Gene* 40, 301-309.
- Geiduschek, E.P. and Ito, J. (1982). In "Molecular Biology of Bacilli" (Dubnau, D.A., ed.). pp. 203-245. Academic Press, New York.
- Godson, G.N., Barrel, B.G., Staden, R. and Fiddes, F.C. (1978). *Nature* 176, 236-247.
- Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B.S. and Stormo, G. (1981). *Ann. Rev. Microbiol.* 35, 365-403.
- Guo, P., Grimes, S. and Anderson, D.L. (1986). *Proc. Nat. Acad. Sci. USA* 83, 3505-3509.

- Hagen, E.W., Reilly, B.E., Tosi, M.E. and Anderson, D.L. (1976). *J. Virol.* 19, 501-517.
- Hager, P.W. and Rabinowitz, J.C. (1985). *J. Biol. Chem.* 260, 15163-15167.
- Harding, N.E. and Ito, I. (1976). *Virology* 73, 389-401.
- Harding, N.E. and Ito, I. (1980). *Virology* 104, 323-338.
- Hawley, D.K. and McClure, W.R. (1983). *Nuc. Acids Res.* 11, 2237-2255.
- Hawley, L.A., Reilly, B.E., Hagen, E.W. and Anderson, D.L. (1973). *J. Virol.* 12, 1149-1159.
- Hodges, Y.K. (1986). Doctoral Dissertation, University of Arizona, Tucson, AZ.
- Holmes, M.W., Platt, T. and Rosenberg, M. (1983). *Cell* 32, 1029-1032.
- Hwang, J.Y. and Doi, R.H. (1980). *Eur. J. Biochem.* 104, 313-320.
- Ito, J., Department of Microbiology and Immunology, University of Arizona, Tucson, AZ 85724
- Ito, J. (1978). *J. Virol.* 28, 895-905.
- Ito, J. and Roberts, R.J. (1979). *Gene* 5, 1-7.
- Jimenez, F. Camacho, A., De La Torre, J., Vinuela, E. and Salas, M. (1977). *Eur. J. Biochem.* 73, 57-72.
- Johnson, W.C., Moran, C.P. and Losick, R. (1983). *Nature* 302, 800-804.
- Josslin, R. (1970). *Virology* 40, 719-726.
- Kaplan, L.M., Ariga, H. Hurwitz, J. and Horwitz, M.S. (1979). *Proc. Natl. Acad. Sci. USA* 76, 5534-5538,
- Kawamura, F. and Ito, J. (1974). *Virology* 62, 414-425.
- Kawamura, F. and Ito, J. (1975). In "Spores VI" (Gerhardt, P., Costilow, R.N. and Sadoff, H.L., eds.), pp. 231-240. American Society for Microbiology, Washington, D.C.
- Kawamura, F. and Ito, J. (1977). *J. Virol.* 23, 562-577.
- Kozak, M. (1983). *Microbiol. Rev.* 47, 1-45.

- Krawiec, S., Jimenez, F., Garcia, J.A., Villanueva, N., Sogo, J.M. and Salas, M. (1981). *Virology* 11, 440-454.
- Leamli, U.K.: (1970). *Nature* 227, 680-685.
- Lennox, E.S.: (1955). *Virology* 1, 190-206.
- Lipman, D.J. and Pearson, W.R. (1985). *Science* 22, 1435-1440.
- Losick, R. and Pero, J. (1981). *Cell* 25, 582-584.
- Loskutoff, D.J. and Pene, J.J. (1973). *J. Virology* 11, 87-97.
- Loskutoff, D.J., Pene, J.J. and Andrews, D.P. (1973). *J. Virology* 11, 78-86.
- Luria, S.E., Darnell, J.E., Baltimore, D. and Campbell, A. (1978). "General Virology" pp.481-490. John Wiley and Sons, New York.
- Marko, M.A., Chipperfield, R. and Birnboim, H.C. (1982). *Analyt. Biochem.* 121, 382-387.
- Mathews, B.W., Remington, S.J., Grutter, M.G. and Anderson, W.F. (1981). *J. Mol. Biol.* 147, 545-558.
- Maxam, A.M., and Gilbert, W. (1980). In "Methods in Enzymology" (Grossman, L. and Moldave, K., eds.). Vol. 65, pp. 590-650. Academic Press, New York.
- McLaughlin, J.R., Murray, C.L. and Rabinowitz, J.C.: (1981). *Proc. Natl. Acad. Sci. USA* 78 4912-4916.
- Mellado, R.P., Moreno, F., Vinuela, E., Salas, M., Reilly, B.E. and Anderson, D.L. (1976). *J. Virology* 19, 495-500.
- Mellado, R.P., Penalva, M.A. Incarte, M.R. and Salas, M. (1980). *Virology* 104, 84-96.
- Mendez, E., Ramirez, G., Salas, M. and Vinuela, E. (1971). *Virology* 45, 567-576.
- Michaelis, S. and Beckwith, J. (1982). *Ann. Rev. Microbiol.* 36, 435-465.
- Moran, C.P., Lang, N., LeGrice, S.F.J., Lee, G., Stephens, M., Sonenshein, A.L., Pero, J. and Losick, R. (1982). *Mol. Gen. Genet.* 186, 339-346.
- Mount, D.W. and Conrad, B. (1984). *Nuc. Acids Res.* 12 (1984) 811-817.

- Murialdo, H. and Becker, A. (1978). *Microbiol. Rev.* 42, 529-576.
- Murray, C.L. and Rabinowitz, J.C. (1982). *J. Biol. Chem.* 257, 1053-1062.
- Nagata, K. Guggenheimer, R.A. and Hurwitz, J. (1983). *Proc. Natl. Acad. Sci. USA* 80, 6177- 6181.
- Nelson, R.A., Reilly, B.E. and Anderson D.L. (1976). *J. Virol.* 19, 518-532.
- Nirenberg, M.W. and Mathau, J.H. (1961). *Proc. Natl. Acad. Sci. USA* 47, 1588-1602.
- Oppenheim, D.S., and Yanofsky, C. (1980). *Genetics* 95, 785-795.
- Owen, J.E., Schultz, D.W., Taylor, A. and Smith, G.R. (1983). *J. Mol. Biol.* 165, 229-248.
- Pastrana, R., Lazaro, J.M., Blanco, L. Garcia, J.A., Mendez, E. and Salas, M. (1985). *Nuc. Acids Res.* 13, 3083-3100.
- Penalva, M.A. and Salas, M. (1982). *Proc. Natl. Acad. Sci. USA* 79, 5522-5526.
- Platt, T. and Bear, D.G. (1983). In "Gene Function in Prokaryotes" (Beckwith, J., Davies, J. and Gallant, J.A., eds.). pp. 123-161. Cold Spring Harbor Laboratories, New York.
- Postle, K. and Good, R.F. (1985). *Cell* 41, 577-585.
- Prieto, I., Lazaro, J.M., Garcia, J.A., Hermoso, J.M. and Salas, M. (1984). *Proc. Natl. Acad. Sci. USA* 81, 1639-1643.
- Rao, G.R.K. and Burma, D.P. (1971). *J. Biol. Chem.* 146, 6474-6479.
- Reader, R.W. and Siminovitch, L. (1971). *Virol.* 43, 607-622.
- Reilly, B.E. (1976). In "Microbiology, 1976" (Schlessinger, D., ed.), pp. 228-237. American Society for Microbiology, Washington, D.C.
- Reilly, B.E., Nelson, R.A. and Anderson, D.L. (1977). *J. Virol.* 24, 363-377.
- Reilly, B.E. and Spizizen, J. (1965). *J. Bact.* 89, 782-790.
- Rekosh, D.M.K., Russell, W.C., Bellett, A.J.D. and Robinson, A.J. (1977). *Cell* 11, 283-295.

- Remaut, D. Stanssens, P. and Fiers, W. (1983) *Nuc. Acids Res.* 11, 4677-4688.
- Rennell, D. and Poteete, A.R. (1985). 143, 280-289.
- Rosenberg, M. and Court, D. (1979). *Ann. Rev. Genet.* 13, 319-353.
- Rubio, V., Salas, M., Vinuela, E., Usobiaga, P., Saiz, J.L. and Lloplis, J.F. (1974). *Virology* 57, 112-121.
- Saedi, M., Department of Microbiology and Immunology, University of Arizona, Tucson AZ 85724.
- Salas, M. (1983). *Current Topics in Microbiology and Immunology.* 109, 89-106.
- Salas, M., Mellado, R.P., Lazaro, J.M. and Sogo, J.M. (1984). In "Genetics and Biotechnology of Bacilli". (Ganesan, A.T. and Hoch, J.A., eds.), pp. 195-208. Academic Press, New York.
- Sanger, F., Air, G.M. Barrel, B.G., Brown, N.L., Coulsen, A.R., Fiddes, J.C., Hutchison, C.A. III, Slocombe, P.M. and Smith, M. (1978). *Nature* 265, 687-695.
- Sanger, F. and Coulsen, A.R. (1978). *FEBS Lett.* 87, 107-110.
- Sanger, F., Coulsen, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982). *J. Mol. Biol.* 162, 729-723.
- Schachtele, C.F., DeSain, S.V. and Anderson, D.L. (1973). *J. Virol.* 11, 9-16.
- Schachtele, C.F., Desain, S.V., Hawley, L.H. and Anderson, D.L. (1972). *J. Virol.* 10, 1170-1178.
- Schollmeir, K., Gartner, D. and Hillen, W. (1985). *Nuc. Acids Res.* 13, 4227-4237.
- Shih, M., Watabe, K. and Ito, J. (1982). *Bioc. Biop. Res. Comm.* 105, 1031-1036.
- Shine, J. and Dalgarno, L. (1974). *Proc. Nat. Acad. Sci. USA* 71, 1342-1346.
- Simons, R.W. and Kleckner, N. (1983). *Cell*, 34, 683-693.
- Smith, D.R. and Calvo, J.M. (1980). *Nuc. Acids Res.* 8, 2255-2274.
- Sogo, J.M., Inciarte, M.R., Corral, J., Vinuela, E. and Salas, M. (1979<sub>a</sub>). *J. Mol. Biol.* 127, 411-436.

- Sogo, J.M., Lozano, M. and Salas, M. (1984). *Nuc. Acids Res.* 12, 512-522.
- Sogo, J.M., Rodeno, P., Koller, T.H., Vinuela, E. and Salas, M. (1979<sub>b</sub>). *Nuc. Acids Res.* 7, 107-120.
- Sprengel, R. Reiss, B. and Schaller, H. (1985). *Nuc. Acids Res.* 13, 893-909.
- Stahly, D.P. and Ito, J. (1981). *Mol. Gen. Genet.* 182, 180-182.
- Steinberg, C.M. and Edgar, R.S. (1982). *Genetics* 47, 187-208.
- Stephens, M.A., Lang, N., Sandman, K. and Losick, R. (1984). *J. Mol. Biol.* 176, 333-348.
- Stillman, B.W. White, E. and Grodzicker, T. (1984). *Virology* 50, 598-605.
- Tinoco, J., Borer, P.N., Dengler, B., Levine, M.D., Ulenbeck, O.C., Crother, D.M. and Gralla, J. (1973). *Nature (New Biology)* 246, 40-41.
- Tomizawa, J. (1984). *Cell* 38, 861-870.
- Tosi, M. and Anderson, D.L. (1973). *J. Virol.* 1548-1559.
- Tosi, M., Reilly, B.E. and Anderson, D.L. (1975). *J. Virol.* 16, 1282-1295.
- Tsugita, A. (1971). In "The Enzymes". (Boyer, P.D. ed.), Vol. 5. pp. 343-411, Academic Press, New York.
- Tsugita, A., Inouye, M., Terzaghi, E. and Streisinger, G. (1968). *J. Biol. Chem.* 243, 391-397.
- Watabe, K., Leusch, M. and Ito, J. (1984). *Proc. Natl. Acad. Sci. USA* 81, 5374-5378.
- Watabe, K., Shih, M. and Ito, J. (1982). *Proc. Natl. Acad. Sci. USA* 79, 5245-5248.
- Watabe, K., Shih, M. and Ito, J. (1983). *Proc. Natl. Acad. Sci. USA* 80, 4248-4252.
- Watson, J.D. (1972). *Nature (New Biology)* 239, 197-201.
- Weaver, L.H., Rennell, D., Poteete, A.R. and Mathews, B.W. (1985). *J. Mol. Biol.* 184, 739-741.

- Wu, R. and Taylor, E. (1971). *J. Mol. Biol.* 57, 491-511.
- Yoshikawa, H., Friedmann, T. and Ito, J. (1981). *Proc. Natl. Acad. Sci. USA* 78, 1336-1340.
- Yoshikawa, H. and Ito, J. (1981). *Gene* 17, 323-335.
- Young, F.E. (1967). *Proc. Natl. Acad. Sci. USA* 58, 2377-2384.