DNA sequence of the *int-xis-P_I* region of the bacteriophage lambda; overlap of the *int* and *xis* genes

R.Wayne Davies

Department of Biology, University of Essex, Colchester, CO4 3SQ, UK

Received 18 February 1980

ABSTRACT

The DNA sequence of the *int* and *xis* genes of lambda bacteriophage, as well as that of the *P_I* promoter and a region of unknown function beyond this, was determined by the chain termination procedure. The Pribnow box sequence of the *P_I* promoter lies just within the *xis* gene, and both possible sites of mRNA initiation from *P_I* are within the *xis* gene. The end of the *xis* gene in its turn overlaps the start of the *int* gene by 23 base pairs, in a different reading frame. This overlap may play a role in ensuring efficient excision of the prophage in response to natural induction signals.

INTRODUCTION

Lambda is a temperate bacteriophage that can establish stable lysogeny by turning off its lytic functions and integrating its DNA into the host chromosome. Integration occurs by a single reciprocal site-specific recombination event between unique sites on the bacteriophage and bacterial chromosomes known as attachment (*att*) sites (1, 2). Integration requires the *int* gene product (3,4,5) (integrase, *Int*) while excision requires both integrase and the *xis* gene product (excisionase, *Xis* 6,7). Both these genes map next to the attachment site (Fig. 1), and belong to the major leftward operon transcribed from *P_L*. The *int* gene, but not the *xis* gene, is also expressed from another promoter called *P_I* at a low constitutive level even when transcription from *P_L* is stopped by the λ repressor (8,9). Expression of the *int* gene from *P_I* is stimulated by the products of the *cII* and *cIII* genes (10,11) which also stimulate *cI* (repressor) gene transcription from *P_{RE}* during lysogenisation. This control system coordinates the high level synthesis of the gene products, repressor and integrase, needed to achieve stable lysogeny. Mutations which increase the expression of integrase by the prophage (*int* *c* mutations, 9), map at the right end of *xis* (12) and are deficient in *xis* gene function. Some of these have

Fig. 1a. Physical map of the lambda bacteriophage DNA molecule showing the position of some important genes and DNA sites, and the positions of the recognition sites for restriction endonucleases Eco RI (19), HindIII (43) and BamHI (44).

b. The positions of genetic elements in the Eco R1C fragment of lambda, based on ref. 45, personal communication from W. Szyhalski and this present paper.

c. Partial restriction map of the lambda att-int-xis-P_I region corresponding to the DNA sequence shown in Fig. 3. Recognition sites shown are as follows: HinfI ∇ , TaqI ∇ , AluI ∇ , Sau3A ∇ , MboII ∇ . The positions of other sites can be determined from Fig. 3. The primer fragments that were used in the sequence determination are shown below the line.

been shown to be due to the addition of insertion sequences (IS elements) carrying a promoter, but others are clearly point mutations. These *int* c point mutations have been interpreted as improvements in a preexisting promoter-like sequence, usually supposed to be P_I, or as defects in a

sequence signalling termination of transcription from P_I in the absence of cII and cIII proteins (called t_I , 13).

The nucleotide sequence of the int and xis genes and the region to the right of the xis gene was determined in order to increase our understanding of the regulation of int gene expression by the cII and cIII gene products, and to find out the protein sequences of integrase and excisionase. The DNA sequence not only contains the expected information, but also suggests strongly that the int and xis genes overlap. That genes can overlap was convincingly demonstrated by the determination of the complete nucleotide sequence of $\phi\chi 174$ (14,15,16). Since then there has been considerable speculation about the reasons for genes to overlap and how widespread this phenomenon might be. In the case of int and xis, our knowledge of the physiology of lambda allows one to propose that these genes most probably overlap because this structure is part of the system ensuring that the correct relative levels of the int and xis gene products are produced in response to natural induction signals, so that excision of the prophage is efficiently coupled to induction.

MATERIALS AND METHODS

Phage and plasmids

Large amounts of purified lambda phage were produced by heat induction of C600 (λ cI₈₅₇ S am₇), followed by one CsCl block gradient and two CsCl equilibrium density gradient steps. The DNA of the plasmids pMG1409 (17) and pac1 29 (G. Cesarini and S. Brenner, manuscript in preparation) was purified by the method of Clewell and Helinski (18) with slight modifications; pMG1409 contains the whole of the lambda Eco.R1 C fragment (19, Fig 1) carrying the attachment site and the int and xis genes, while pac1 29 contains the HindIII - Bam HI fragment covering the attachment site and the C-terminal coding end of the int gene.

Preparation of templates

Purified λ single strands were prepared from purified phage by the method of Hradecna and Szybalski (20). Lambda 1-strand template was also prepared from pMG1409 DNA using the exonuclease III method of Smith (21), utilising the fact that pMG1409 has a single site for the restriction endonuclease HindIII within the λ fragment just 350 base pairs away from the end of the int gene (22). 100 μ g of pure supercoiled DNA were digested with HindIII in 7mM Tris HCl pH 7.4, 7 mM MgCl₂, 50 mM NaCl, 1 mM DTT. The reaction was stopped by heating at 70°C for 10 minutes, and the DNA was.

ethanol precipitated, washed in 95% ethanol, dried and taken up in 50 μ l of sterile water. 25 μ l of this solution was treated with exonuclease III by adding 3 μ l of a solution containing 700 mM Tris HCl pH 8, 10 mM $MgCl_2$ and 100 mM DTT and 0.5 units of exonuclease III (New England Biolabs) per μ l total volume. After 7 hours digestion at 20°C the DNA was treated twice with freshly distilled phenol and excess phenol removed by six extractions with water saturated ether. After a further ethanol precipitation and wash with 95% ethanol, the dried precipitate was taken up in 25 μ l of sterile water, and 1 μ l of this was used per sequencing reaction.

Preparation of DNA fragments

The starting material for the preparation of primer fragments was either the HindIII - Bam HI fragment from p α cl 29, the λ Eco RI C fragment from pMG1409, or a HincII fragment derived from this and containing the attachment site and almost all of the int gene. The HincII cut at the right end of this fragment is shown in Fig 1c. These fragments were purified by agarose gel electrophoresis. Pieces of gel containing the DNA fragment in question were crushed by passing them through a 1 ml syringe with no needle, sometimes followed by passage through a large-bore needle (e.g Becton-Dickinson G22). After incubation at 37°C for more than 12 hours in a solution containing 0.15 M NaCl, 10 mM Tris HCl pH 8 and 1 mM EDTA, solid agarose was removed by centrifugation. The supernatant was passed through a small (0.25 ml) DEAE column that had been equilibrated with the extraction buffer. After washing the column with 5 mls or more of extraction buffer, the DNA was eluted by washing with 2 mls of 1 M NaCl and ethanol precipitated using the ultracentrifuge.

DNA fragments to be used as primers in the DNA sequencing reactions were prepared from these large fragments by further restriction endonuclease digestion using enzymes that cut frequently; Mbo_{II}, HinfI, AluI, Sau3A and TaqI were used. All restriction endonucleases were purchased from New England Biolabs or Bethesda Research Laboratories, except for Eco.RI which was bought from Boehringer. Digestions were always carried out in 7 mM Tris HCl pH 7.4, 7 mM $MgCl_2$, 50 mM NaCl, 1 mM DTT at 37°C, or 65°C for TaqI. The products of these reactions were separated from one another in 8% polyacrylamide thin gels. Usually a small portion (1/40) of the preparation was radioactively labelled by filling in the ends with the Klenow fragment of DNA polymerase I (23) (Boehringer), or by nick-translation for blunt-ended fragments. The labelled and unlabelled fragments were mixed and run together on the polyacrylamide gel. This allows the position of each

fragment to be identified by autoradiography, and also allows the subsequent extraction and concentration steps to be monitored. The segment of gel was placed whole in 1 ml of a solution containing 0.15 M NaCl, 10 mM Tris HCl pH 8 and 1 mM EDTA. After 24 hours at 37°C, the supernatant was taken and ethanol precipitated in the ultracentrifuge.

DNA sequence determination

DNA sequencing was carried out using the chain termination method (24). The amount of primer fragment used was that derived from 0.5 µg of the HindIII - Bam HI fragment or from 5 µg of the λ EcoRI C fragment. This was mixed with 2.5 µg of purified λ single-strand, or 1 µl of exonuclease-III treated pMG1409 prepared as described above, and 0.5 µl of H x 10 buffer (25) and the volume adjusted to 5 µl with sterile distilled water. The solution was sealed in a capillary tube, boiled for 3 minutes and incubated at 65°C for 30 minutes. The sequencing reaction and subsequent gel electrophoresis and autoradiography were carried out as described previously according to Sanger et al (24,26,27). On several occasions sequences within a primer fragment were determined by treating the primer with exonuclease III before annealing to the template, a method developed by A. J. H. Smith (personal communication).

RESULTS AND DISCUSSION

The int and xis genes have long been known from genetic mapping data to be located very close together near their site of action, the phage attachment site. The attachment site was therefore taken as a primary reference point, since its sequence has been determined (22,28) and the position of the cross-over point is known to be within or at the ends of a 15 base pair region common to the phage and bacterial attachment sites (the common core). Previously, an open reading frame at the right end of the 317 base pair HinfI fragment containing att was tentatively identified as coding for the C-terminal end of the int gene.

The sequencing strategy was twofold. First to use primer fragments from the right end of the HindIII - Bam HI fragment to obtain DNA sequence as far into the int gene from its C-terminal coding end as possible. The DNA sequence determined in this way could now be scanned for restriction endonuclease recognition sites. Endonucleases whose recognition sites fell within the new sequence were then used to digest purified λ EcoRI C fragment (or the HincII fragment derived from it), and since the size of the new primer fragment was known it was usually simple to recognise it in

polyacrylamide gels and purify it. In this way, given a large number of restriction enzymes and a reasonably random sequence, it is possible to extend the region of known sequence along a DNA molecule. However, because of the multiple steps involved, this approach actually yielded information more slowly than the second more direct strategy. This was simply to take the λ EcoRI C fragment or its HincII subfragment, to digest it with a restriction endonuclease that cuts frequently and purify all resulting fragments that were less than 200 base pairs long. These fragments were then used as primers in both directions, the resulting sequence data collated and overlaps found by inspection or by using computer programmes devised by Staden (29,30). Single-enzyme digestions with HinfI, TaqI, AluI and Sau3A were used, but it would in general be more efficient to use double or multiple digests to obtain small primer fragments from every region. Once sufficient sequence data had been amassed, any remaining gaps were filled by making particular primers using restriction cuts known from the sequence to be near the gap. In fact only one region, in the middle of the int gene, proved difficult, and this was sequenced using primers obtained by an AluI - Sau3A double digest. Whenever a sequence was not completely clear in a set of primings in one direction, the complementary sequence on the other strand was determined. Both l and r-strand sequences of all of the xis-P_I region were determined; only in the middle of the int gene was the sequence determined on one strand only, and this region preserves the reading frame and has no nonsense codons. Initially, only small fragments were used as primers, but larger fragments can be used as primers after a pretreatment with exonuclease III (A. J. H. Smith, personal communication). This enables one to determine sequences within primer fragments with ease. The range of primer fragments that gave the clearest information is shown in Fig. 1c; many smaller pieces of corroborative sequence were obtained using other fragments.

Autoradiographs of two typical sequencing gels are shown in Fig. 2. One is of an experiment in which pure λ r-strand was used as a template, the other of an experiment in which exonuclease III treated HindIII cut pMG1409 was used as a template. The latter shows the sequences that can be identified as P_I and the beginning of the xis gene. The DNA sequence that was finally obtained is given in Fig. 3, in the form of the r-strand sequence only, since this is equivalent to the mRNA sequence. The amino acid sequences of the proteins are also given in the single letter code described in reference 31.

The int gene is easy to identify. It has been shown by genetic

Fig. 2. DNA sequence analysis of the int-xis- P_I region of lambda. Autoradiograms of two representative 7M urea 8% polyacrylamide thin gels (26) displaying the products of sequencing experiments using the chain termination method (24).

- a. The Mbo_{II} - Bam_{HI} fragment that lies just to the right of att (Fig. 1) was used as a primer on λ r strands purified as in ref. 20. This is 1-strand sequence, complementary to the sequence in Fig. 3. The first band visible (arrowed) corresponds to G¹⁴⁷⁰ in Fig. 3. and longer runs of the same material gave unambiguous data up to G¹²⁷¹, shorter runs down to A¹⁵³¹.
- b. The 78 bp $TaqI$ fragment (cuts between bp294 and 295, and bp373 and 374 in Fig. 3) was used as a primer on λ 1-strand template. The 1-strand template in this case was prepared by treating $HindIII$ -cut pMG1409 (17), which contains the λ $EcoRI$ C fragment (19) and only has one $HindIII$ site (that left of att) with exonuclease III as described in Materials and Methods (21). This autoradiogram shows r-strand sequence from bp380 to bp535 (Fig. 3), which is the putative cII-cIII- site and P_I region plus the beginning of the xis gene.

Y K Y S M K V Y K S N E D K S P F C N V R E M
CATACAAATATTCCATGAAGGTTTATAAGTCTAATGAAGACAAATCCCCATTCTGCAATGTGCGAGAAAT
10 20 30 40 50 60 70

T D T V A K L L S R V R G N D T C P L C T K H
GACTGATACCGTGGCAAAATTATTATCACGAGTACGTGGAAACGATACTTGCCCTCTCTGTACAAAACAT
80 90 100 110 120 130 140

I D D *
ATAGATGATTAACCCAATATTACATAACAATCCTCGCACTCGCGGGGATTATTTTATCTGAACTCGCT
150 160 170 180 190 200 210

M H F R V T G E W N G E P
M I N A L P S H R R M E W R A I
ACGGCGGGTTTTGTTTTATGGAGATGATAAATGCACTTCCGAGTCACAGGAGAATGGAATGGAGAGCCAT
220 230 240 250 260 270 280

F N R V I E A E N I N D C Y D H W M I W A Q I A
Q Q S Y R S G E H Q R L L R P L D D M G A D S
TCAACAGAGTTATCGAAGCGGAGAACATCAACGACTGCTACGACCACTGGATGATATGGGCGCAGATAGC
290 300 310 320 330 340 350

H A D V T N I R I E E L K E H Q A A *
T C R R N Q Y S N *
ACATGCAGACGTAACCAATATTGCAATTGAAGAACTGAAAGAACACCAAGCCGCCTGATGGCGGTTTTTT
360 370 380 390 400 410 420

P R S L E T V R R W V R E C R I F P P P V K D G
CAAGAAGCCTTGAAACAGTTTCGTGATGGGTTCCGGAATGCAGGATATCCACCTCCGGTTAAGGATGG
500 510 520 530 540 550 560

R E Y L F H E S A V K V D L N R P V T G G L L
 AAGAGAGTATCTGTTCACGAATCAGCGGTAAGGTTGACTTAAATCGACCAGTAACAGGTGGCCTTTTG
 570 580 590 600 610 620 630

INT INT XIS
 INT RBS M G R R R S H E R R D L P P N L Y I R
 K R I R N G K K A K S *
 AAGAGGATCAGAAATGGGAAGAAGCGGAAGTCATGAGCGCGGGGATTACCCCTAACCTTTATATAAGA
 640 650 660 670 680 690 700
 START STOP

N N G Y Y C Y
 R D P R T G K E F G L G R D R R
 AACAAATGGATATTACTGCTACAGGGACCCAAGGACGGGTAAAGAGTTTGGATTAGGCAGAGACAGCGGAA
 710 720 730 740 750 760 770

I A I T E A I Q A N I E L F S G H K H K P L T A
 TCGCAATCACTGAAGCTATACAGGCCAACATTGAGTTATTTTCAGGACACAAACACAAGCCTCTGACAGC
 780 790 800 810 820 830 840

R I N S D N S V T L H S W L D R Y E K I L A S
 GAGAAATCAACAGTGATAATTCCGTTACGTTACATTCATGGCTTGATCGCTACGAAAAATCCTGGCCAGC
 850 860 870 880 890 900 910

R G I K Q K T L I N Y M S K I K A I R R G L P
 AGAGGAATCAAGCAGAAGACACTCATAAATTACATGAGCAAAATTAAGCAATAAGGAGGGGTCTGCCTG
 920 930 940 950 960 970 980

D A P L E V I T T K E I A A M L N G Y I D E G K
 ATGCTCCACTTGAAGTCATCACCACAAAAGAAATTGCGGCAATGCTCAATGGATACATAGACGAGGGCAA
 990 1000 1010 1020 1030 1040 1050

A A S A K L I R S T L S D A F R E A I A E G H
 GCGGCGTCAGCCAAGTTAATCAGATCAACACTGAGCGATGCATTCCGAGAGGCAATAGCTGAAGGCCAT
 1060 1070 1080 1090 1100 1110 1120

I T T N H V A A T R A A K S E V R R S R L T A
 ATAACAACAAACCATGTCGCTGCCACTCGCGCAGCAAAATCAGAGGTAAGGAGATCAAGACTTACGGCTG
 1130 1140 1150 1160 1170 1180 1190

D E Y L K I Y Q A A E S S P C W L R L A M E L A
 ACGAATACCTGAAAATTTATCAAGCAGCAGAATCATCACCATGTTGGCTCAGACTTGCAATGGAAGTGGC
 1200 1210 1220 1230 1240 1250 1260

V V T G Q R V G D L C E M K W S D I V D G Y L
 TGTGTGTACCGGGCAACGAGTTGGTGATTTATGCGAAATGAAGTGGTCTGATATCGTAGATGGATATCTT
 1270 1280 1290 1300 1310 1320 1330

Y V E Q S K T G V K I A I P T A L H I D A L G
 TATGTCGAGCAAAAGCAAACAGCGTAAAAATTGCCATCCCAACAGCATTGCATATTGATGCTCTCGGAA
 1340 1350 1360 1370 1380 1390 1400

I S M K E T L D K C K E I L G G E T I I A S T R
 TATCAATGAAGGAAACACTTGATAAATGCAAAGAGATTCTTGGCGGAGAAACCATAATTGCATCTACTCG
 1410 1420 1430 1440 1450 1460 1470

R E P L S S G T V S R Y F M R A R K A S G L S
 TCGCGAACCGCTTTTCATCCGGCACAGTATCAAGGTATTTTATGCGCGCAGCAAAAGCATCAGGTCTTTCC
 1480 1490 1500 1510 1520 1530 1540

F E G D P P T F H E L R S L S A R L Y E K Q I
 TTCGAAGGGGATCCGCCTACCTTTACGAGTTGCGCAGTTTGTCTGCAAGACTCTATGAGAAGCAGATAA
 1550 1560 1570 1580 1590 1600 1610

S D K F A Q H L L G H K S D T M A S Q Y R D D R
 GCGATAAGTTTGCTCAACATCTTCTCGGGCATAAGTCCGACACCATGGGCATCACAGTATCGTGATGACAG
 1620 1630 1640 1650 1660 1670 1680

Fig. 3 Nucleotide sequence of the att-int-xis-P_L region; protein sequences of the int and xis gene products. Only r-strand sequence is given, numbered running from right to left on the standard lambda map (Fig. 1). The r-strand is the DNA equivalent of the mRNA sequence in this region of lambda. Base-pairs 1590-1911 were sequenced previously by Davies et al. (22) and the att sequence beyond bp 1800 is given in that publication. The numbering has no biological significance, starting simply with the rightmost base in the sequence that was obtained. Referring to the eighth base in the common core as 0 (28), base pair 1795 in this Figure is 0, base pair 1695 is +100, 1595 is +200, etc. Features of interest mentioned in the text are boxed and identified. Ribosome binding sites are labelled RBS and underlined. The Int binding site is that defined by protection against dimethyl sulphate (42) (Minter, S. J., Kotewicz, M. L., Davies, R. W., Schreier, P. H. and Echols, H., manuscript in preparation). The single letter amino acid code is given in ref. 31.

mapping to be situated immediately adjacent to att and to be about 1250 base pairs long (12). The gene product, integrase, has been purified, and is a protein of M.W. 40000 (32). In this part of lambda the l-strand is the template for transcription from P_L (and P_I), so that the r-strand is the DNA equivalent of the mRNA sequence. In Fig 3 a 1068 bp long open reading frame can be seen, beginning with A⁶⁴⁴TG and ending with T¹⁷¹²AATGA, a double stop that is 83 bp from the middle of the common core and immediately adjacent to a tight binding site for integrase that must form part of att (33, Minter, S. J., Kotewicz, M. L., Davies, R. W., Schreier, P. H. and Echols, H., manuscript in preparation). This is the same reading frame as that identified during the sequencing of the attachment site (22). There is no other open reading frame of comparable length in this region, so that this must be the coding sequence of the int gene. A ribosome binding site should have at

least three bases complementary to the 5' GAUCACCUCCUUA_{OH} 3' sequence at the 3' end of 16S ribosomal RNA which has been shown to be involved in the initiation of protein synthesis, and be within 15 bases of the initiation codon (34,35). The A⁶⁴⁴TG int gene initiation codon is preceded by a 5' G⁶³³AGG 3' sequence complementary to 3' CUCC 5', which is a good candidate for a ribosome binding site. Thus the int gene codes for a protein with 356 amino acid residues, which agrees well with the known M.W. of integrase. The protein sequence is given in Fig. 3 in single letter code (31).

A second feature of the sequence that can be unambiguously identified is the P_I promoter, since a sequence upstream from the int gene bears a very close resemblance to λ P_{RE}, the other lambda promoter that, like P_I, is dependent upon the function of the cII and cIII gene products before it can be used efficiently. Comparison of 46 promoters recognised by E. coli RNA polymerase (36) shows homologies between them around positions -10 and -35 upstream from the mRNA startpoint. The -10 region of the λ P_{RE} promoter has been identified by the DNA sequence of a mutant (37); the -35 region associated with this has no homology with other promoters, consistent with this promoter being absolutely dependent upon positive regulation by the cII and cIII gene products. Both genetic and structural evidence indicate that the region defined by the five -35 λ P_{RE} mutations is the site of action of the cII and cIII gene products (38,39). Since the cII and cIII gene products also exert positive control over transcription from P_I, the occurrence of a sequence close to and upstream from the int gene ribosome binding site, and homologous to the λ P_{RE} -35 site, identifies and defines the position of P_I. The specialised -35 site in P_{RE} is TTGCGTTGTGTTGC; the sequence T⁴²²TTGCGTGAATTGC in Fig 3 is identical to this in 11 out of 14 places. No other sequence in the 643 base pair sequence upstream of the start of the int gene resembles P_{RE} at all closely. Fourteen bases downstream from these sequences occur the sequences AAGTAT (P_{RE}) and TGTACT (P_I), both poor versions of -10 sequences. The structure of these promoters are thus virtually identical. This would place the initiation of transcription from P_I at either A⁴⁵⁹ or A⁴⁶⁴, well outside the int gene. Since there is no sequence between P_I and the start of the int gene that is an obvious candidate for a messenger RNA termination site, the DNA sequence data seem to support a promoter stimulation rather than antitermination model for the action of the cII gene product, but this will only be definitively shown by the sequences of int_c point mutants that are being determined in the laboratories of A. Landy and H. Echols (personal communication), and by

further work on transcription in this region. There is in fact a possible terminator sequence upstream of the P_I sequence from C^{401} to G^{414} , but no good promoter can be seen in the 400 bases before it; this may be one of the series of termination sites in the N antitermination control system.

Since the int gene and the P_I promoter can be so clearly identified, we can use this positional information and the known facts about the genetics of this region to identify the xis gene unambiguously. Enquist and Weisberg (12) mapped 29 xis⁻ and 56 int⁻ mutants against a set of 25 deletions and substitutions in the int-xis region. With one exception (int 2266, see below) all xis⁻ mutants were to the right of all int⁻ mutants, and the two groups were essentially contiguous. The maximum recombination distance within the int gene was 3 cM (4% λ), within xis 0.2 cM (0.3% λ), and this together with heteroduplex mapping of the deletions and substitutions used for genetic mapping lead them to propose sizes for the int and xis genes of 1100-1300 (1240) and 100-250 (110) base pairs. As according to the DNA sequence int codes for 356 amino acids, the proportion of xis⁻ and int⁻ mutants among excision defective mutants (12) gives an estimate of 60 amino acids for xis. Clearly xis is a very small gene very close to the beginning of int. A search for ATG initiation codons with associated ribosome binding sites in the region upstream from int yields only three promising candidates, which are shown in Fig. 3. They are $A^{234}TG$ to $T^{378}GA$ (48 amino acids), $A^{241}TG$ to $T^{406}GA$ (55 amino acids) and $A^{448}TG$ to $T^{664}GA$ (72 amino acids). This last 72 amino acid coding sequence, which overlaps the beginning of the int gene by 23 bp including the stop signal, is the best candidate for the xis gene. The ribosome binding site is 5' $G^{436}GAG$, which pairs with 3' CCUC.

Three lines of argument support this conclusion. Transcription from P_I leads to the production of an mRNA containing the information for the synthesis of Int protein but not Xis protein (8,9). The int_c mutations that make the expression of the int gene constitutive and independent of the cII and cIII gene products all make the phage xis⁻ simultaneously, and some of them are clearly point mutants (9,12). The point mutants can either be interpreted as changes in P_I that make it resemble the normal E. coli general promoter sequence more closely, in which case xis must at least overlap with P_I , or as defects in a termination sequence that must lie between P_I and the int gene, in which case xis must also be situated between P_I and the int gene. The only open reading frame that fits is the $A^{448}TG$ to $T^{664}GA$ sequence, since the T of the putative initiation codon $A^{448}TG$ is the first base in the most probable -10 sequence of

the P_I promoter, so that changes in this sequence could well also be xis^- mutations. Moreover, the proposed P_I dependent transcription starts at A^{459} or A^{464} are both within this probable xis gene sequence, so that the mRNA would lack the xis gene ribosome binding site and initiation codon. Two other known facts provide further support of this interpretation; the proximity of int^- and xis^- mutants in the genetic map, and the phenotype of the int 2266 mutant (12). Thus the mapping of int^- and xis^- mutants shows a linear array with no break between the genes. While the number of deletion classes used was not large (fifteen), it is very unlikely that a gap one quarter the size of the int gene would not have been detected, as must be the case if one of the alternative open reading frames were really the xis gene. Int 2266 is a recessive mutant that is both int^- and xis^- , maps as the rightmost int^- and the leftmost xis^- mutant and is not a physically detectable deletion. The phenotype and map position of int 2266

Fig. 4. The decision whether to integrate or excise. P_L and P_I are the leftward and int -promoters of lambda. n is an unknown value above which integration occurs more frequently than excision. The model is not all-inclusive; protein stabilities (46) also play a role, for instance, and although Int - Xis is written as a complex, this need not be the case.

is neatly explained if int and xis overlap, and if xis is separated from int by 240 bp or more it can only be explained by the complex assumption that it is an int⁻ mutation which causes the Int protein to interact negatively with the Xis protein. Thus all those arguments taken together leave no room for doubt that xis is a gene coding for 72 amino acids that overlaps the first part of the int gene.

Overlapping genes are now a well-accepted phenomenon in the biology of small DNA bacteriophages and small DNA animal viruses. For all the well-established cases, space-saving has been the most popular explanation. Barrell (41) suggested for one case in $\phi\lambda 174$ that the relative position of two overlapping genes might be important in determining the relative rates of production of the two gene products. The case of the int and xis genes of lambda supports this kind of interpretation strongly. The argument is as follows. The integrase protein causes insertion of λ DNA into the host chromosome. Excisionase is not needed for integration, but is essential for excision of viral DNA from the chromosomes of lysogenic cells. The existence of different catalytic requirements for integration and excision allows the direction of the reaction to be controlled so that integration is favoured at some stages of the life cycle and excision at others. An easy way to visualise it (40) is to propose that Int and Xis form a protein complex, and the decision whether to integrate or excise depends upon the relative levels of free integrase and Int-Xis complex. This need not be the way it actually happens. If after infection the life cycle is switched into the lysogenic pathway, expression of the int and xis genes from P_L is swamped by cII and cIII gene product stimulated int mRNA synthesis from P_I , excess free integrase is synthesised and integrative recombination predominates. In the lytic pathway Int and Xis are unimportant. In the lysogenic state P_L is turned off by repressor, and low level cII and cIII gene product independent expression of the int gene from P_I combined with the absence of Xis keeps the prophage locked in place. Induction of the prophage under natural conditions is likely to be usually weak and transient, but the phage has to respond to these signals that the host has become a poor environment by excising itself from the chromosome and making progeny. Thus the natural induction process will involve intermittent bursts of synthesis from P_L (and P_R), leading to the synthesis of Int and Xis from P_L -initiated mRNA alone, since some time is required before concentrations of the cII and cIII gene products reach levels at which the cII gene product significantly stimulates P_I . In this situation the relative levels of integrase and excisionase

depend solely upon the structure of the leftward operon and on the relative translation efficiencies for the two genes. To ensure efficient excision *Xis* should be in excess, or the ratio of free integrase to Int-*Xis* complex has to be below a critical value (n in Fig. 4); precisely the opposite situation from lysogenisation. This could be achieved by the genes being in tandem and the ribosome binding site for *int* being less efficient, but this mechanism is disadvantageous during lysogenisation. The way it seems to be achieved is to overlap the end of *xis* with the ribosome binding site and initiation codon of the *int* gene in of course a different reading frame. If translation of *xis* is efficient, ribosomes reading *xis* will block free ribosomes from starting to read *int*. In this way the *int* gene can maintain a good ribosome binding site, but *xis* expression can nevertheless predominate during induction. By the time enough *cII* has been made, replication is underway and the lytic pathway has been initiated. This is summarised diagrammatically in Fig. 4.

Thus overlapping two genes puts expression of the "internal" gene under control of the efficiency of translation of the "external" gene. If the products of the two genes are needed in particular relative concentrations in this cell, having the genes overlap combined with fine tuning through ribosome binding site efficiency and messenger structure is a good way to achieve the correct ratio. This, as well as space-saving, may be a general reason why genes are found to overlap.

Acknowledgements

I thank Dr. F. Sanger for encouragement and the use of laboratory facilities, Dr. P. H. Schreier and D. E. Büchel for their important contributions to the sequence near *att* (22), Drs. B. G. Barrell and P. H. Schreier for discussion and A. J. H. Smith and Dr. H. Echols for communicating results prior to publication. This work was supported by an MRC project grant.

REFERENCES

1. Nash, H. A. (1977) *Curr. Topics Microbiol. Immunol.* 78, 171-199.
2. Weisberg, R. A., Gottesman, S. and Gottesman, M. E. (1977) *Comprehensive Virology* 8 (Fraenkel-Conrat, H. and Wagner, R. eds.), 197-288, New York, Plenum Press.
3. Zissler, J. (1967) *Virology* 31, 189.
4. Gingery, R. and Echols, H. (1967) *Proc. Nat. Acad. Sci. U.S.A.* 58, 1507-1514.
5. Gottesman, M. E. and Yarmolinsky, M. B. (1968) *J. Mol. Biol.* 31, 487-505.
6. Kaiser, A. D. and Masuda, T. (1970) *J. Mol. Biol.* 47, 557-564.
7. Guarneros, G. and Echols, H. (1970) *J. Mol. Biol.* 47, 565-574.

8. Shimada, K., Weisberg, R. A. and Gottesman, M. E. (1973) *J. Mol. Biol.* 80, 279-314.
9. Shimada, K. and Campbell, A. (1974) *Proc. Nat. Acad. Sci. U.S.A.* 71, 237-241.
10. Katzir, N., Oppenheim, A., Belfort, M. and Oppenheim, A. B. (1976) *Virology* 74, 324-331.
11. Court, D., Adhya, S., Nash, H. and Enquist, L. (1977) *DNA Insertion Elements, Plasmids and Episomes* (Bukhari, A. I., Shapiro, J. A. and Adhya, S. L. eds.) pp 389-394. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
12. Enquist, L. and Weisberg, R. A. (1977) *J. Mol. Biol.* 111, 97-120.
13. Pilacinski, W. et al. (1977) *Gene* 2, 61-74.
14. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A. III., Slocombe, P. M. and Smith, M. (1977) *Nature* 265, 687-695.
15. Smith, M., Brown, N. L., Air, G. M., Barrell, B. S., Coulson, A. R., Hutchison, C. A. III and Sanger, F. (1977) *Nature* 265, 702-705.
16. Barrell, B. G., Air, G. M. and Hutchison, C. A., III (1976) *Nature* 264, 34-41.
17. Schreier, P. H., Davies, R. W., Büchel, D. E., Gronenborn, B., Fanning, T. G., von Wilcken, B. and Messing, J. (1977) *Nature* 267, 555-557.
18. Clewell, D. B. and Helinski, D. R. (1959) *Proc. Nat. Acad. Sci. U.S.A.* 62, 1159-1166.
19. Thomas, M. and Davis, R. W. (1975) *J. Mol. Biol.* 91, 315-328.
20. Hradecna, Z. and Szybalski, W. (1967) *Virology* 32, 633-643.
21. Smith, A. J. H. (1979) *Nucleic Acids Res.* 6, 831-848.
22. Davies, R. W., Schreier, P. H. and Büchel, D. E. (1977) *Nature* 270, 757-760.
23. Davies, R. W., Schreier, P. H., Kotewicz, M. L. and Echols, H. (1979) *Nucleic Acids Res.* 7, 2255-2273.
24. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) *Proc. Nat. Acad. Sci. U.S.A.* 74, 5463-5467.
25. Air, G. M., Sanger, F. and Coulson, A. R. (1976) *J. Mol. Biol.* 108, 519-533.
26. Sanger, F. and Coulson, A. R. (1978) *FEBS Lett.* 87, 107.
27. Davies, R. W., Schreier, P. H. and Büchel, D. E. (1978) *Nucleic Acids Res.* 5, 3209-3218.
28. Landy, A. and Ross, W. (1977) *Science* 197, 1147-1160.
29. Staden, R. (1977) *Nucleic Acids Res.* 4, 4037-4051.
30. Staden, R. (1978) *Nucleic Acids Res.* 5, 1013-1015.
31. *Biochem. J.* (1969) 113, 1-4.
32. Kotewicz, M. L. et al., *J. Biol. Chem.* in press.
33. Ross, W., Landy, A., Kikuchi, Y., and Nash, H. A. (1979) *Cell* 18, 297-308.
34. Shine, J. and Dalgarno, L. (1975) *Eur. J. Biochem.* 57, 221-230.
35. Steitz, J. A. and Jakes, K. (1978) *Proc. Nat. Acad. Sci. U.S.A.* 72, 4734-4738.
36. Rosenberg, M. and Court, D. (1979) *Ann. Rev. Genet.* in press.
37. Rosenberg, M., Court, D., Shimatake, H., Brady, C. and Wulff, D. L. (1978) *Nature* 272, 414-423.
38. Rosenberg, M., Court, D., Shimatake, H., Brady, C. and Wulff, D. L. (1978) in *The Operon* (Miller, J. H. and Reznikoff, W. S., eds) pp 345-371, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
39. Wulff, D. L., Behr, M., Izumi, S., Beck, J., Mahoney, M., Shimatake, H., Brady, C., Court, D. and Rosenberg, M. (1980) *J. Mol. Biol.* in press.

40. Echols, H. (1970) *J. Mol. Biol.* 47, 575-583.
41. Barrell, B. G. (1978) *International Review of Biochemistry, Biochemistry of Nucleic Acids* 11, 17 (B. F. C. Clark, ed.), pp.125-179.
42. Gilbert, W., Maxam, A. M. and Mirzabekov, A. D. (1971) in *Control of Ribosome Synthesis, Alfred Benson Symposium IX* (Kjeldgaard, N. and Maaloe, U. eds.) pp 139-148, Munksgaard, Copenhagen.
43. Murray, K. and Murray, N. E. (1975) *J. Mol. Biol.* 98, 551-564.
44. Haggerty, D. M. and Schleif, R. F. (1976) *J. Virol.* 18, 659-663.
45. Szybalski, W. (1976) in *Handbook of Biochemistry and Molecular Biology* (ed. Fasman, G. D.) 3rd Ed., Nucleic Acids Vol. II pp 677-685, CRC Press, Cleveland.
46. Weisberg, R. A. and Gottesman, M. E. (1971) in *The Bacteriophage Lambda* (Hershey, A. D., ed.) pp 489-500, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.