J. Mol. Biol. (1972) **71**, 383–405

The Topography of Lambda DNA: Polyriboguanylic Acid Binding Sites and Base Composition

JAMES J. CHAMPOUX† AND DAVID S. HOGNESS

Department of Biochemistry
Stanford University School of Medicine
Stanford, Calif. 94305, U.S.A.

(*Received 16 March 1972*)

The positions of the ten poly(rG) binding sites on the two strands of λ DNA have been determined with the aid of the families of overlapping DNA fragments described in the preceding paper and the DNA of deletion mutants. Nine sites are on the *r*-strand‡: two are located in the right half, $0{\cdot}84 \pm 0{\cdot}03$ and $0{\cdot}91 \pm 0{\cdot}02$ molecular length unit from the left end; and seven are evenly distributed within 0·4 unit from the left end. The tenth site is on the *l*-strand, $0{\cdot}60 \pm 0{\cdot}03$ unit from the left end. Another site was mapped on the *r*-strand of the *bio* segment of *Escherichia coli* DNA found in λ *bio* variants; the *l*-strand of this segment contains no sites.

All λ sites are alike in binding the same amount of poly(rG) and are considered to be derived from a set of repeated sequences with two orientations: *gr* sequences generate sites on *r*-strands; *gl* sequences, on *l*-strands. Without exception, this orientation corresponds to the orientation of transcription at the location of the sequence, transcription being from the *r*-strand at *gr* sites and from the *l*-strand at the *gl* site. The location of these *g* sequences indicates that they are not equivalent to operators, promoters or rho-dependent termination sites. However, *g* sequences do exhibit a positive topographic correlation with the multiple mRNA molecules derived from single transcriptons in λ. This correlation suggests that *g* sequences determine where the mRNA from a single transcripton is to be divided. The selective advantages that could result from this division and the likely mechanisms for effecting it are discussed.

The base composition of λ DNA as a function of molecular length has been determined from the buoyant density of the overlapping fragments, complementing and extending previous maps.

1. Introduction

The right and left families of λ DNA fragments defined in the previous article (Egan & Hogness, 1972) can be considered as ordered sets of overlapping deletions to which the principles of deletion mapping can be applied. Since two members of the same family have in common all sequences of the smaller member, chemical differences between the two members result from and characterize the region of non-overlap in the larger member. In the present article we have used this analysis of differences (a) to identify the position of the sequences responsible for the binding of poly(rG) and (b) to

† Present address: Department of Microbiology, University of Washington Medical School, Seattle, Wash. 98105, U.S.A.

‡ Terminology. The right and left of linear λ DNA correspond to the right and left of the linkage map for vegetative λ (Hogness & Simmons, 1964). The *r*-strand of λ DNA is transcribed rightward, having its 5′ end to the right. The *l*-strand is its complement (see Szybalski, 1970).

amplify previous maps of the base composition as a function of molecular length (Skalka, Burgi & Hershey, 1968). Our primary emphasis will be on the mapping of the poly(rG) binding sites.

Poly(rG) exhibits a differential affinity for the two strands of λ DNA, which results in a differential increase in their buoyant densities which can easily be measured (Kubinski, Opara-Kubinska & Szybalski, 1966). This differential affinity varies with the region of the λ DNA examined and is correlated with the strand that is transcribed in the region, at least at the level of half-molecules (Hradecna & Szybalski, 1967). Transcription in the left arm of λ DNA is from the *r*-strand (Fig. 6) and poly(rG) is preferentially bound to that strand in this region. The right arm can be divided into two regions of approximately equal length such that transcription is predominantly from the *r*-strand in the quarter containing the right end, but is from the *l*-strand in the quarter immediately to its left (Fig. 6). Hradecna & Szybalski (1967) have interpreted their results to indicate (a) that poly(rG) is bound equally to the *r*- and *l*-strands in the right half, and (b) that the binding sites on one strand are in a different region of the right half from those on the complementary strand. The location and size of these regions were not defined, though their size was presumably in the order of quarter molecules.

This rough positive correlation between transcription and binding of poly(rG) in λ DNA and in other systems (e.g. at the level of whole molecules of phage T3 and T7 DNA (Summers & Szybalski, 1968*a*)) led to the suggestion that the sequences responsible for binding may act in duplex DNA as signals for the initiation and/or termination of transcription (Szybalski, Kubinski & Sheldrick, 1966; Szybalski *et al.*, 1969). Our purpose is to test these and other proposals of function by increasing the resolution for the mapping of the binding sequences.

2. Experimental Procedure

(a) *Materials*

(i) *Bacteria and phage*

The λ*b*2 and λ*b*221 DNA's referred to in the text were obtained from phages λ*b*2*c* (Kellenberger, Zichichi & Weigle, 1960,1961) and λ*b*221*c*26 (Parkinson & Huskey, 1971). The λ*bioP* and λ*bioy* DNA's originate from the λ*bio* transducing variants M38–5 and M58–2, respectively (Kayajanian, 1968,1970). The deletions and substitutions in all four DNA's are indicated in Table 2.

W602 is a *bio*$^-$ derivative of *E. coli* K12 (Rothman, 1965). Double lysogens of the constitution W602 (λ*bio*, λ*ind*$^-$*cI*857) were constructed by transduction with the above λ*bio*'s, using λ*ind*$^-$*cI*857 (Sussman & Jacob, 1962) as helper. The λdv plasmid DNA was obtained from the *E. coli* derivative KM314 (Matsubara & Kaiser, 1968). All other bacteria and phage have been previously described (Egan & Hogness, 1972).

(ii) *Media and reagents*

Media and all but the following reagents have been described (Egan & Hogness, 1972). Poly(rG) was obtained from Biopolymers, Inc. and ^{3}H-labeled poly(rG) from Miles Laboratories.

(b) *Methods*

(i) *DNA and phage preparations*

The isolation of DNA from phage, and of λ*ind*$^-$*cI*857 (used as the source of λ DNA) and 5-bromouracil-labeled λ*ind*$^-$*cI*857 phages are described in the preceding article (Egan & Hogness, 1972). The λ*bio* phages were prepared by thermal induction of W602 (λ*bio*, λ*ind*$^-$*cI*857) and purified according to the above procedures except that centrifugation in CsCl was repeated until the less dense λ*bio* contained less than one λ*ind*$^-$*cI*857 contaminant

per 500 λ*bio*'s. λ*b2c* and λ*b221c26* were prepared by infection of C600 and subsequent purification in CsCl gradients.

Members of the right and left families of λ DNA fragments are those described in the preceding paper (Egan & Hogness, 1972). The λdv DNA was obtained from G. Hobom (Universität Freiburg).

(ii) *Buoyant densities and molecular weights*

Buoyant densities of the DNA's were determined in the Beckman-Spinco model E ultracentrifuge as described by Doerfler & Hogness (1968), except that in some cases the monochromator-scanner system ($\lambda = 265$ nm) was used instead of photography to determine the ultraviolet absorbance pattern. The density of alternating poly[d(A-T)] was taken as 1·6790 g/ml. (Schildkraut, Marmur & Doty, 1962) and β as $1{\cdot}19 \times 10^9$ cm^5 g^{-1} sec^{-2} for the computation of densities (Ifft, Voet & Vinograd, 1961).

The molecular weights of DNA's from deletion and substitution variants relative to that of λ DNA were determined by the mixed-oligomer method of Baldwin, Barrand, Fritsch, Goldthwait & Jacob (1966), as recently modified by Baldwin (1971).

(iii) *Analysis of poly(rG) binding*

Standard assay. The DNA to be analyzed (2 μg) was diluted to 0·44 ml. in 0·001 M-EDTA, 0·01 M-Tris·HCl buffer (pH 7·1) and the solution made alkaline by the addition of 0·02 ml. of 1·0 M-NaOH. After 5 min at room temperature the solution was chilled to 0°C, neutralized with 0·03 ml. of 1·0 M-Tris·HCl buffer (pH 7·1) plus 0·02 ml. of 1·0 M-HCl, and the poly(rG) added immediately (0·02 ml. of the hydrolyzed stock solution described below). After addition of poly[d(A-T)] (5 μl.; $A_{260} = 3{\cdot}1$), solid CsCl was added to a density of 1·712 g/ml. and the solution (final vol. 0·70 ml.) centrifuged at 44,000 rev./min and 25°C until the position of the peaks representing the two strands stabilized (approx. 20 hr). The same procedure, except for omission of the poly(rG), was followed to determine the density of the strands alone, and thereby, the increase in density due to poly(rG) ($\Delta\rho$). The density of the reference poly[d(A-T)] is not affected by poly(rG). When the scanner was used instead of photography, the DNA concentration was reduced 2-fold. Typical absorbance patterns obtained by both techniques are given in Fig. 1. The general rise in absorbance near the bottom of the cell that is apparent in all the photographic tracings is due to free poly(rG) and can be cancelled out with the scanner if the solvent sector of the 2-sector cell contains the same amount of free poly(rG) as the sector containing the DNA (see Fig. 1, λ*b221* DNA).

Alkaline hydrolysis of poly(rG). The poly(rG) (0·50 mg/ml.) was hydrolyzed in 0·33 M-NaOH at 37°C for varying times and tested in the above procedure with λ DNA to find the optimum time of hydrolysis (5 min) as determined by the maximum ratio of the interband distance to average band width of the two strands. Shorter hydrolysis times yield greater values for $\Delta\rho$ and therefore greater interband distances, but increase the band width even more; longer times result in less separation with less than a proportionate decrease in band width. The treatment rendered 8% of the material acid soluble (as measured by A_{260}) and zone sedimentation of ^{3}H-labeled poly(rG) treated identically yielded a broad band with $S_{20,w} = 3{\cdot}5$ S. The poly(rG) stock solution was neutralized after hydrolysis and contained 0·33 mg poly(rG)/ml. after dialysis against 0·001 M-EDTA, 0·01 M-Tris·HCl buffer (pH 7·1). The same solution was used for all analyses and exhibited no change after storage at 5°C for more than 1 year. The molar absorbancy of this solution was taken as $8{\cdot}0 \times 10^3$ M^{-1} cm^{-1}, this being the average of two values: $8{\cdot}2 \times 10^3$ (Pochon & Michelson, 1965) and $7{\cdot}8 \times 10^3$ (determined directly using the orcinol reaction with GMP as a standard; Mejbaum, 1939).

Saturation conditions. The change in $\Delta\rho$ for the strands of λ DNA as a function of poly(rG) concentration is given in Table 1. The concentration of poly(rG) used in the standard assay (9·4 μg/ml.) is as close to saturation as is practical for the photographic method because of the interfering absorbance caused by small poly(rG) molecules. Effective saturation at this poly(rG) concentration is also indicated by the fact that the $\Delta\rho$ for the *r*-strand is increased by only 3% when the DNA concentration was lowered 10-fold; decreasing the DNA concentration 2-fold, as is done when the scanner is used, does not produce an observable change in $\Delta\rho$.

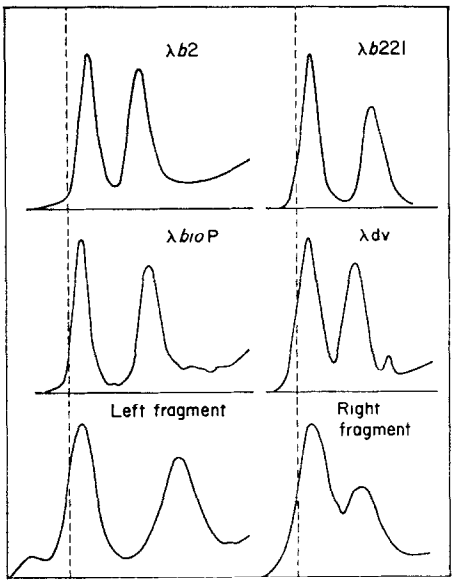


FIG. 1. Effect of poly(rG) on the distribution of strands in CsCl gradients. Density increases toward the right, and the peak of greatest density represents the *r*-strand of each DNA. The three distributions in the left column derive from densitometer tracings of absorption photographs; those in the right column are from the scanner. The distributions in each column are aligned so that the vertical line represents the density of the strands of each DNA in the absence of poly(rG), all distributions being normalized to constant peak amplitude for the *l*-strands (ordinate) and to proportionality between interpeak radial distance and the difference in density between the two peaks (abscissa). The left and right fragments have lengths of $f_L = 0{\cdot}214$ and $f_R = 0{\cdot}202$, respectively. The fraction of absorbance due to the band of *r*-strands is 0·49, 0·47, 0·49, 0·46, 0·47 and 0·42 for λb2, λb221, λ*bioP*, λdv, left and right fragments, respectively, after correction for the absorbance of free poly(rG) (see text), minor amounts of renatured DNA (e.g. trace of left fragments) and the geometry of the cell sectors.

TABLE 1

Increase in density of the strands of λ DNA as a function of poly(rG) concentration

Poly(rG) (μg/ml.)	0·47	1·4	2·4	4·7	9·4	19·0
$\Delta\rho$ (g/ml.) *l*-strand	0·0025	0·0041	0·0045	0·0050	0·0059	0·0064
r-strand	0·0088	0·0138	0·0150	0·0167	0·0187	0·0207

Except for the poly(rG) concentration, the conditions of the standard assay (2·9 μg DNA/ml.) were used throughout. $\Delta\rho$ is the difference between the buoyant density in the presence and absence of poly(rG). Values for $\Delta\rho$ at 9·4 μg poly(rG)/ml. are the means from 12 determinations, the other values resulting from one or two determinations.

3. Results

(a) *The map of base composition*

(i) *Experimental plan*

Consider an overlapping pair of DNA molecules where the longer member contains all sequences in the shorter member; e.g. two members of the right family of fragments

or λ DNA and DNA from a simple deletion mutant of λ. Designating the fraction of λ DNA in the shorter member by f, and that in the longer member by $f + \Delta f$, the G + C content of the region of non-overlap represented by Δf is related to the G + C contents of the two members by:

$$(\Delta f)\,(G + C)_\Delta = (f + \Delta f)\,[(G + C)_f + \Delta(G + C)] - f(G + C)_f$$

or

$$(G + C)_\Delta = [(G + C)_f + \Delta(G + C)] + f[\Delta(G + C)/\Delta f], \quad (1)$$

where $(G + C)_\Delta$ is the G + C content of the Δf region, and $(G + C)_f$ and $[(G + C)_f + \Delta(G + C)]$ represent the G + C contents of the shorter and longer members, respectively.

Since λ DNA contains very few modified bases (0·05 to 0·55 mole %, depending upon the host (Arber, 1968)), the base compositions of duplex DNA molecules derived from λ DNA can be reliably calculated from differences in their buoyant densities in CsCl solutions relative to that of λ DNA according to:

$$(G + C) = (G + C)_\lambda + 10{\cdot}2\,(\rho - \rho_\lambda) \quad (2)$$

(Schildkraut *et al.*, 1962). Skalka *et al.* (1968) obtained a value of 0·505 for the G + C content of λ DNA by analysis of the nucleotides released after enzymic hydrolysis, while Mushynski & Spencer (1970*b*), using the same method, report a value of 0·488, yielding a mean of 0·496. This mean agrees well with values obtained by two other techniques: 0·496 obtained by analysis of the nucleosides released after formic acid hydrolysis (Mushynski & Spencer, 1970*b*), and 0·497 obtained by spectral analysis (Falkow & Cowie, 1968). Since we obtain a value of 1·7085 g/cm^3 for the buoyant density of λ DNA under our standard conditions (see section (b) Experimental Procedure), the preceding equation becomes:

$$G + C = 0{\cdot}496 + 10{\cdot}2\,(\rho - 1{\cdot}7085), \quad (3)$$

where ρ is the density of any λ-derived duplex DNA determined under these conditions and G + C is its G + C content.

Use of equations (1) and (3) allows the mapping of G + C content *versus* molecular length from the densities of overlapping DNA molecules derived from λ DNA.

(ii) *The right arm*

A plot of the densities of 19 fragments from the right family as a function of their lengths measured from the right end of λ DNA is given in Figure 2. An ordinate transform according to equation (3) allows this same curve to serve as a plot of G + C content of the fragments *versus* their length. The G + C content of small segments of DNA $(G + C)_\Delta$ as a function of their distance from the right end (f_R, equivalent to f in equation (1)) can be calculated from this curve and equation (1), which for very small segments (Δf approaches zero) becomes:

$$(G + C)_\Delta = (G + C)_f + f_R\,(G + C)', \quad (4)$$

where $(G + C)'$ is the slope at f_R of the curve of the G + C content of fragments, $(G + C)_f$, *versus* their length, f_R.

The resulting curve of $(G + C)_\Delta$ *versus* f_R is given by the heavy line in Figure 2. The DNA of the λdv plasmid is contained within this map ($f_R = 0{\cdot}121$ to 0·266; M. Fiandt & W. Szybalski as cited by Davidson & Szybalski, 1971). Its density is 1·7063 g/ml., corresponding to a G + C content of 0·474. This is in good agreement with the G + C value of 0·471 calculated from the map in Figure 2.

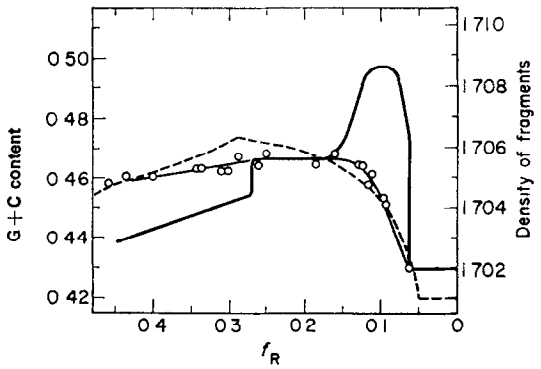


FIG. 2. Map of the base composition of the right arm. The circles represent the buoyant density in CsCl (right-hand ordinate) or the G+C content (left-hand ordinate) of fragments in the right family plotted against their lengths, f_R, as determined in the preceding paper (Egan & Hogness, 1972). The thin line drawn through these points consists of three straight line segments (least squares) connected by a curve for the region $f_R = 0{\cdot}09$ to $0{\cdot}16$ such that all points are within experimental error ($\pm 0{\cdot}0003$ g/cm^3, right ordinate). The heavy line indicates the G+C content of λ DNA at f_R computed from the preceding curve and equation (4). The dashed line indicates the expected densities of the fragment, were the map of Skalka *et al.* (1968) valid (see Fig. 3 and Discussion).

(iii) *The left arm and center*

The densities of seven fragments from the left family with lengths evenly distributed over the range $f_L = 0{\cdot}18$ to $0{\cdot}44$ have been measured (where f_L is the fraction of the molecular length of λ DNA measured from the left end). The densities of the two fragments closest to the left end ($f_L = 0{\cdot}181$ and $f_L = 0{\cdot}214$) do not differ significantly and correspond to a mean G + C content and range of $0{\cdot}560 \pm 0{\cdot}001$. The densities of the five fragments closest to the center ($f_L = 0{\cdot}252$ to $0{\cdot}44$) form a second homogeneous group the corresponding G + C content of which is $0{\cdot}548 \pm 0{\cdot}003$, significantly below that of the first group. Applying equation (1) to the decrease in the G + C content of fragments as their length increases from $0{\cdot}214$ to $0{\cdot}252$ yields a value of $0{\cdot}48$ for this interval. Since the G + C content of the fragments remains constant for lengths greater than $f_L = 0{\cdot}252$, segments in the interval $f_L = 0{\cdot}252$ to $0{\cdot}44$ must have G + C contents equal to that of the longer fragments, i.e. $0{\cdot}548$ (Fig. 3). The depth of the calculated depression at $f_L = 0{\cdot}23$ is inversely proportional to its estimated width. The width could be smaller and the depth greater, since the boundaries were determined from a limited population of fragments. Errors of about 6% in the relative lengths of fragments (Egan & Hogness, 1972) also allow a greater width which could raise the G + C content in the depression as high as $0{\cdot}51$.

The G + C content of the central region can also be determined from density differences to complete a map constructed by a single technique. The deletion mutant $\lambda b2$ lacks $12{\cdot}6 \pm 0{\cdot}6\%$ of λ DNA located between $f_L = 0{\cdot}448 \pm 0{\cdot}005$ and $f_L = 0{\cdot}573$ to $0{\cdot}574$ (Westmoreland, Szybalski & Ris, 1969; Davis & Parkinson, 1971). The difference in density between $\lambda b2$ and λ DNA is $0{\cdot}0014 \pm 0{\cdot}0003$ g/ml. ($0{\cdot}0016$—Miyazawa & Thomas, 1965; $0{\cdot}0011$—Hirschman, Gellert, Falkow & Felsenfeld, 1967; $0{\cdot}0017$—Hradecna & Szybalski, 1967; $0{\cdot}0013$—Falkow & Cowie, 1968; $0{\cdot}0013$—our data), indicating that the G + C content of $\lambda b2$ DNA is $0{\cdot}014 \pm 0{\cdot}003$ greater than that of

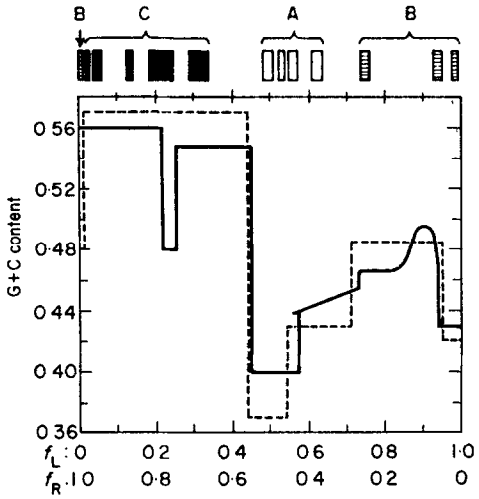


FIG. 3. Maps of the base composition of λ DNA. The continuous line indicates the map determined from density differences as described in the text. The dashed line is the map of Skalka *et al.* (1968). The segments of partial denaturation mapped by Inman (1967) and Inman & Schnös (1970) are given above the maps, and the class (A, B or C) to which they belong indicated. The classes are defined according to the temperature or time at alkaline pH which causes denaturation within the segment, A segments exhibiting denaturation before B, and B before C.

λ DNA. The G + C content of the deleted region is therefore 0·40 ± 0·03 (equation (1)).

Integration over our entire map (Fig. 3) yields a G + C value of 0·492. This is in good agreement with the reference value of 0·496 used in equation (3) and indicates that our analysis is internally consistent.

(b) *The map of poly(rG) binding sites*

(i) *Experimental plan*

If poly(rG) binds to n specific sites on a DNA strand, then the buoyant density of the complex is given by:

$$\rho = \frac{m_D + nm_G}{(m_D/\rho_D) + (nm_G/\rho_G)}, \tag{5}$$

where m_D is the mass of the DNA strand, m_G is the average mass of the poly(rG) bound per site, ρ_D is the buoyant density of the DNA strand without poly(rG) bound to any of the n sites, ρ_G is the buoyant density of poly(rG), and it is assumed that these densities are unchanged in the complex. We are interested in determining the number of nucleotide residues of poly(rG) bound to the n sites in a strand of known size from the resulting increase in buoyant density ($\Delta\rho_n = \rho - \rho_D$). From equation (5) we obtain:

$$\begin{aligned} nm_G &= m_D[\rho_G/\rho_D]\,[\Delta\rho_n/(\rho_G - \rho_D - \Delta\rho_n)], \text{ or} \\ nN_G &= N_D\,[M_D\rho_G/M_G\rho_D]\,[\Delta\rho_n/(\rho_G - \rho_D - \Delta\rho_n)], \end{aligned} \tag{6}$$

where N_D and N_G are the number of nucleotide residues in the DNA strand and in the average poly(rG) bound per site, and M_G and M_D are the molecular weights (as cesium salts) per nucleotide residue of poly(rG) and λ DNA, respectively.

Since $N_D = fN_\lambda$ (where $N_\lambda = 46{,}500$, the number of nucleotide residues per strand of whole λ DNA; Davidson & Szybalski, 1971), N_D is calculable for any strand of known

f. The buoyant density of poly(rG) in CsCl at 25°C (ρ_G) approaches that of saturated CsCl and has not been directly determined. We have adopted an estimated value of 1·96 g/cm^3 from the density of poly(rC) (1·88—Schildkraut, Marmur, Fresco & Doty, 1961; 1·886—Bruner & Vinograd, 1965) and the difference in density between poly(dG) and poly(dC) of 0·078 (Wells, Larson, Grant & Sweet, quoted by Szybalski & Szybalski, 1970). This value is in good agreement with the value of 1·95 determined indirectly by Summers & Szybalski (1968*b*). While the buoyant density of the complex (ρ) is that measured under the standard conditions (see Experimental Procedure), the determination of ρ_D requires special comment because poly(rG) causes a small non-specific increment in the buoyant densities of all DNA strands that we have investigated.

The increases in buoyant density resulting from poly(rG) ($\Delta\rho$) for both *r*- and *l*-strands of some fragments of the left family of DNA from two *λbio* substitution mutants and from two simple deletion mutants, and of λ DNA are given in Table 2. The increase in density observed for the *l*-strand in all these molecules, except those of λ and the smaller of the two deletions, *λb2*, is within experimental error of their mean, 0·0039 g/cm^3. We consider this increase to be non-specific, i.e. to be independent of specific regions, or sites. This non-specific increment has also been observed for the *l*-strand in fragments of the right family, and for the *r*-strand in the smallest of these fragments (Fig. 4(a)). We have also observed this increment for the *l*-strand of DNA in phage T7 which is generally regarded not to bind poly(rG) (Summers & Szybalski, 1968*a*,*b*). In sum, under our conditions all DNA strands that we have tested exhibit at least this small non-specific increment in density.

The buoyant densities, ρ and ρ_D, are so defined that their difference, $\Delta\rho_n$, equals the increment caused by binding of poly(rG) only to the n specific sites. As ρ necessarily includes the non-specific increment, so must ρ_D; hence ρ_D equals the buoyant density of the strand measured in the absence of poly(rG) plus the non-specific increment. Using the value of 0·0039 g/cm^3 for the non-specific increment, 1·7228 g/cm^3 for the buoyant density of denatured λ DNA in the absence of poly(rG) (the mean of 12 independent determinations under our conditions), and the appropriate molecular weights in equation (6), one obtains

$$nN_G = (4{\cdot}87 \times 10^4)[f\Delta\rho_n/(0{\cdot}233 - \Delta\rho_n)]. \quad (7)$$

Since $\Delta\rho_n = \Delta\rho$ — (non-specific increment), nN_G can be computed for any strand of known length from the observed $\Delta\rho$†.

(ii) *Specific binding to the* l*-strand*

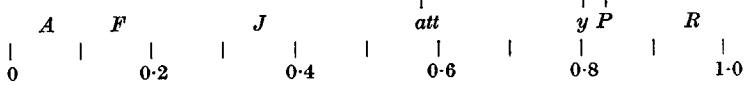
The *l*-strands of DNA from λ and *λb2* exhibit increases in density due to poly(rG) which are significantly higher than the non-specific increment and correspond to nN_G values of 420 and 310, respectively (Table 2). As will become apparent, these values are consistent with a single binding site on the *l*-strand ($n = 1$) with N_G equal to 350 nucleotide residues.

† The buoyant densities of strands from the same duplex do not differ significantly in the absence of poly(rG). However, the densities of different duplexes derived from λ DNA vary sufficiently relative to $\Delta\rho$ that the densities of the strand and of the complex must be determined in each case to obtain an accurate value of $\Delta\rho$ (see Experimental Procedure, section (b)). On the other hand, the simplification in equation (7) of using a constant value for ρ_D where it is not part of $\Delta\rho n$ results in an error of no more than 3% in nN_G.

TABLE 2

Binding of poly(rG) to the strands of λ, substitution and deletion mutants, and left fragments

Type of DNA	f	$\Delta\rho$ (g/ml.)		nN_G		n	
		l	*r*	*l*	*r*	*l*	*r*
(a) Fragments—left family							
	0·181	0·0039	0·0306	0	1140	0	*3·3*
	0·214	0·0037	0·0309	(−10)	1370	0	*3·9*
	0·290	0·0038	0·0315	(−10)	1900	0	*5·4*
	0·382	0·0037	0·0309	(−20)	2440	0	*7·0*
(b) Substitution mutants							
λbio y	0·863	0·0036	0·0227	(−50)	3690	0	*10·5*
λbio P	0·866	0·0040	0·0227	20	3700	0·1	*10·6*
(c) Deletion mutants							
λb221	0·777	0·0043	0·0215	60	3090	0·2	*8·8*
λb2	0·874	0·0056	0·0201	310	3180	0·9	*9·1*
(d) λ	1·000	0·0059	0·0187	420	3300	1·2	*9·4*



Left fragments were obtained and their lengths (f) determined as described by Egan & Hogness (1972). The molecular weights of the *λbio* DNA's relative to that of λ DNA (f) were determined by the mixed-oligomer method (see Experimental Procedure). The f values for *λb221* and *λb2* DNA's and the deletion boundaries result from electron microscopy (see Davidson & Szybalski, 1971); these f values are in good agreement with values determined by the mixed-oligomer method: *λb221* = 0·771 ± 0·007; *λb2* = 0·874 ± 0·004 (mean and range of values determined here and by Baldwin (1971)).

The right-hand boundaries of the *λbio P* deletion intersects the *P* gene of λ (Kayajanian, 1968) and is therefore located at f_L = 0·814 (Egan & Hogness, 1972; see also Davidson & Szybalski, 1971). Heteroduplex mapping places the right-hand boundary of the *λbio y* deletion at, or just to the right of, the right-hand boundary of *imm*^{434} (M. Fiandt & W. Szybalski, personal communication), indicating it cuts the *y* gene, or is very close to the left end of *y* (f_L = 0·791; Davidson & Szybalski, 1971). The left-hand boundary of both *λbio* deletions is taken as the *att* position (f_L = 0·573 to 0·574; Davidson & Szybalski, 1971). The length of the inserted *bio* segment of *E. coli* DNA, indicated by the horizontal dashed line, is the difference between the f value and the remaining λ DNA, being 0·08 (or 3700 base pairs) for *λbio y*, and 0·11 (or 5100 base pairs) for *λbio P*. These segments contain the *bio* gene sequences *att-bio A-bio B* and *att-bio A-bio B-bio F-bio C*, respectively (A. Campbell, personal communication).

The $\Delta\rho$ value is the difference between the density in the presence and absence of poly(rG); the values for λ DNA are the means of 12 determinations, and those for the mutant DNA's of at least 3 determinations. Values of nN_G were computed from equation (7) (taking $\Delta\rho_n = \Delta\rho - 0{\cdot}0039$), and n was determined for N_G = 350.

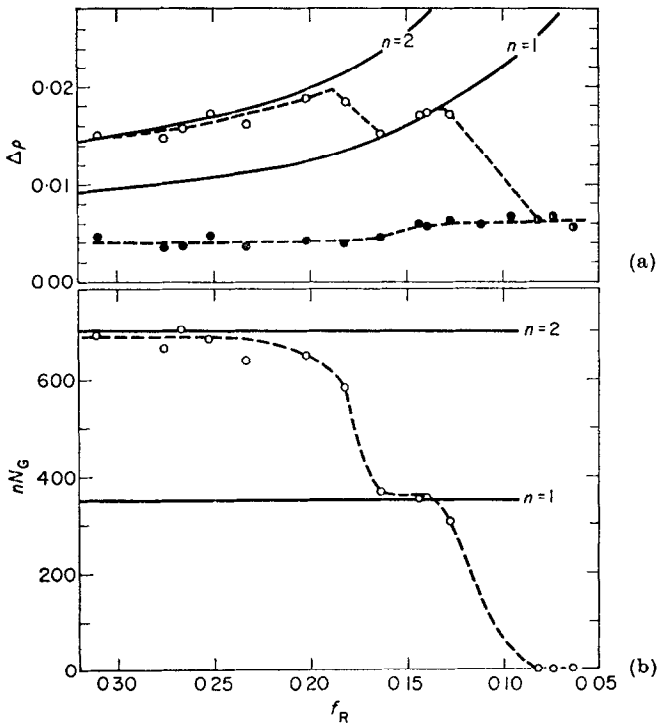


FIG. 4. Poly(rG) binding to the strands of fragments in the right family.

(a) (○) $\Delta\rho$ for the *r*-strand; (●) the *l*-strand; (◑) the common $\Delta\rho$ for both strands. Fractions with $f_R \geq 0{\cdot}128$ exhibit two peaks in the CsCl gradients (Fig. 1) which yield the different $\Delta\rho$ values for each strand. Fractions with $f_R \leq 0{\cdot}082$ yield unimodal density distributions, indicating both strands exhibit about the same $\Delta\rho$. The density distributions for the two intermediate fractions ($f_R = 0{\cdot}096$ and $0{\cdot}112$) consist of a main peak (*l*-strand) with a shoulder at higher densities (*r*-strand); only the $\Delta\rho$ for the *l*-strand is given, as that for the *r*-strand could not be accurately determined. The densities can be assigned to the proper strand by correlation with the results given in Table 2. This assignment was confirmed in separate experiments by allowing the fragment strands to anneal with an excess of purified *l*- or *r*-strands of bromouracil-labeled λ DNA before centrifugation with poly(rG) (method B, Champoux, 1969). This procedure removes the fragment strand that is complementary to the added bromouracil-labeled strand from the density region of interest, and allows the observed $\Delta\rho$ to be assigned to the remaining fragment strand. The solid lines are theoretical curves for strands containing one or two sites capable of binding 350 poly(rG) residues per site calculated from equation (7) (see text).

(b) The values of nN_G for the fragment *r*-strands were calculated from equation (7) as described in the text. The solid horizontal lines indicate the values expected for $N_G = 350$ and $n = 1$ or 2.

The density increment for the *l*-strand of the DNA from the larger deletion, λ*b*221, indicates that this binding site has been deleted and hence is located in the region of non-overlap between λ*b*221 and λ*b*2. The absence of the site in the two λ*bio* substitution mutants localizes it to the region between the right ends of the λ*b*2 ($f_L = 0{\cdot}574$) and λ*b*221 ($f_L = 0{\cdot}629$) deletions, a region containing 2560 bases (see Fig. 6).

(iii) *Specific binding to the* r*-strand*

Right arm. The effect of poly(rG) on the densities of the strands in the fragments of the right family is given in Figure 4(a), the lower curve representing the *l*-strand, and the upper dashed curve, the *r*-strand. The solid lines are theoretical curves of $\Delta\rho$ *versus* f_R for fragment strands containing one or two sites, each capable of binding

350 nucleotide residues of poly(rG) and located to the right of f_R. The values of $\Delta\rho$ observed for the *l*-strands are, as expected, incompatible with a binding site on this strand in the interval shown. By contrast, the correspondence between the theoretical and observed curves for the *r*-strand indicates the existence of two separate binding sites to the right of $f_R = 0{\cdot}18$.

The amount of poly(rG) specifically bound to the *r*-strand of these fragments (nN_G) was calculated from equation (7) and the results plotted in Figure 4(b). The two-step curve indicates that the first site (i.e. the site closest to the right end) binds 360 nucleotide residues of poly(rG) and the second, 330. These values are not significantly different from each other, nor from the mean of 365 nucleotide residues calculated for the single site on the *l*-strand (Table 2). The average value for the three sites is 350 nucleotide residues, which we take to be the amount of poly(rG) bound to each, and which was therefore used to compute the theoretical curves shown in Figure 4(a).

The strands of right fragments which are longer than those included in Figure 4 exhibit density distributions in the presence of poly(rG) which are not sufficiently bimodal to assign accurately $\Delta\rho$ values to either strand. However, the mean and width of these distributions are compatible with the expected single site on the *l*-strand and the two sites on the *r*-strand.

Where are these two binding sites located on the *r*-strand? The position of each step in the curves shown in Figure 4 will be displaced to the left of the location of each binding site because of the length heterogeneity of the fragments in each fraction (Egan & Hogness, 1972) and our limited analysis of the density distribution of the poly(rG)-strand complexes†. However, since this displacement will be approximately the same for each step, the difference between the two step positions ($\Delta f = 0{\cdot}07 \pm 0{\cdot}01$) should equal the molecular distance between the two sites. Hence a determination of the position of the first binding site is sufficient to determine the location of both sites.

We have determined the position of the first binding site on the *r*-strand by the same method used in the preceding paper (Egan & Hogness, 1972) to determine gene positions. A population of fragments from the right family with lengths both shorter and longer than required to include the site was subject to zone sedimentation and successive fractions analyzed for the relative number of molecules containing the site. As the lengths of the fragments in successive fractions become shorter, the number containing the site will decrease to indicate the "cut-off" size, i.e. the length of the shortest fragment containing the site. The results of this analysis are given in Figure 5. We used a population for which the sedimentation distributions of the *P* and *Q* genes had been determined so that the cut-off and hence the position of the first binding site could be determined relative to that for these genes. The cut-offs for the first binding site and for the *Q* gene are seen to be indistinguishable. The first binding site is therefore located at $f_R = 0{\cdot}09 \pm 0{\cdot}02$, the range resulting from errors inherent in the mapping of the *Q* gene (Egan & Hogness, 1972) and from those associated with the

† Given a fraction with f_R equal to the position of the first site and an estimated length heterogeneity such that the duplex fragments with length equal to or greater than this f_R comprise $\leq 50\%$ of the DNA in the fraction (Egan & Hogness, 1972), then no more than 25% of the denatured DNA will contain the site and exhibit the higher $\Delta\rho$. This percentage is just below the threshold that our experience and criteria dictate is necessary for scoring the higher $\Delta\rho$ in Figure 4(a) (the actual percentage should be even smaller, since the high-density component generally comprises somewhat less than the theoretical proportion (Fig. 1)). Such a fraction would be recorded as were the two fractions with $f_R = 0{\cdot}096$ and $0{\cdot}112$ (see legend to Fig. 4(a)), indicating that the first site is near these end-points, and to the right of the first step.

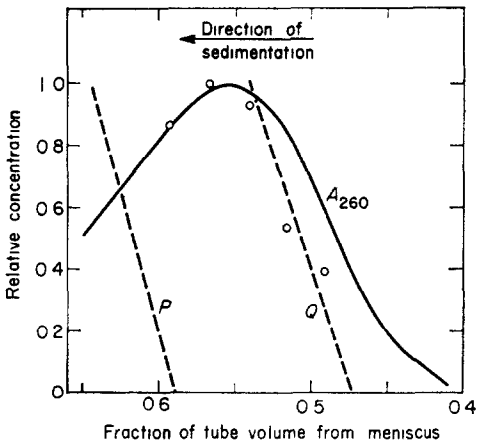


FIG. 5. Location of a binding site on the *r*-strand near the *Q* gene. The zone sedimentation of the R6 population of fragments from the right family in a sucrose gradient is described in Figure 6 of Egan & Hogness (1972) from which the curves for genes *P* and *Q* and for A_{260} are taken. The gene distributions represent the relative molar concentration of fragments containing the given gene in different fractions from the gradient, only that portion of the curve exhibiting decreasing concentration with decreasing fragment length being shown. To determine the relative molar concentration of fragments containing the first site on the *r*-strand, the DNA in fractions containing fragments with average $f_R \leq 0{\cdot}16$ were subject to the standard assay for poly(rG) binding and the percentage of the absorbance due to *r*-strands exhibiting the higher $\Delta\rho$ was determined as described in Figure 1. Multiplying this percentage by the A_{260} for the fraction and dividing the product by the average f_R yields a value proportional to the desired molar concentration. These values after normalizing to the maximum value, are given by the open circles.

assay of fragments containing the binding site. The second binding site must then be located at $f_R = 0{\cdot}16 \pm 0{\cdot}03$, the position of both sites being indicated in Figure 6.

The location of the binding sites in the right arm is confirmed by the results obtained with the DNA from the λdv plasmid, which, as has been indicated, contains the segment of λ DNA, $f_R = 0{\cdot}121$ to $0{\cdot}266$. The *r*-strand of this DNA should contain the second but not the first site and the *l*-strand should be devoid of sites (Fig. 6). Nicked circles of λdv DNA (formed either by introducing a single break with pancreatic DNase or by radioactive decay of ^{3}H-labeled supercoils provided by G. Hobom) were isolated and analyzed in the standard assay for the binding of poly(rG). The $\Delta\rho$ observed for the *l*- and *r*-strands is 0·0032 and 0·0166 g/cm^3, respectively. The small non-specific density increment exhibited by the *l*-strand is that expected from the data in Figure 4(a). The $\Delta\rho$ for the *r*-strand, when corrected for this non-specific increment, yields a value for nN_G of 420 nucleotide residues, equivalent to 1·2 binding sites.

Left arm, center and the bio *segment.* The data for the binding of poly(rG) to the *r*-strand in fragments of the left family is given in Table 2. The *r*-strand in the largest of these fragments ($f_L = 0{\cdot}382$) binds 2440 nucleotide residues of poly(rG), which is equivalent to seven sites, each capable of binding 350 residues. The *r*-strand in the shortest fragment ($f_L = 0{\cdot}181$) binds an amount of poly(rG) equivalent to about three binding sites (3·3, Table 2), indicating that four of the presumed sites are located in the interval, $f_L = 0{\cdot}18$ to $0{\cdot}38$ since a plot of nN_G *versus* f_L (similar to that in Fig. 4(b)) for the *r*-strands of the four fragments ending within this interval is perfectly linear, with a slope equivalent to one site every $0{\cdot}055\,f$ unit, or 2560 bases, there is no evi-

dence for a concentration of binding sites within the interval. Rather, this result is consistent with a uniform distribution of sites, since a curve consisting of four equal steps ($\Delta n N_G = 350$) which are uniformly spaced ($\Delta f_L = 0{\cdot}055$) fits the data equally well. We have no information regarding the distribution of poly(rG) bound to the *r*-strand in the interval $f_L = 0{\cdot}0$ to $0{\cdot}18$; hence the three binding sites that we presume are in this region could be distributed in any manner.

Since the entire *r*-strand of λ DNA binds an amount of poly(rG) equivalent to nine sites (9·4, Table 2), the seven sites in the left arm plus the two in the right arm account for all of the λ sites. The conclusion that there are no binding sites on the *r*-strand in the central region is sustained by the observation that the *r*-strand of the deletion mutants, λ*b*2 and λ*b*221, also bind an amount of poly(rG) equivalent to nine sites (9·1 and 8·8, Table 2).

The *r*-strand in each of the two substitution mutants, λ*bio y* and λ*bio P*, binds an amount of poly(rG) significantly greater than the *r*-strand of λ, λ*b*2, or λ*b*221 DNA's (Table 2). This amount is equivalent to one or two additional binding sites which must be assigned to the *r*-strand of the *bio* segment of *E. coli* DNA inserted into these mutants (see legend to Table 2). Since λ*bio y* contains only the *A* and *B* cistrons of the biotin operon, we assume the site(s) to be restricted to the *r*-strand of the *bioA-bioB* segment. The *l*-strand of even the longer *bioA-bioB-bioF-bioC* segment found in λ*bioP* does not contain a binding site.

4. Discussion

(a) *Maps of base composition*

There is a general qualitative agreement between the maps obtained by Skalka *et al.* (1968) and by us (Fig. 3). There is, however, a significant difference between the maps of the right arm for the region, $f_R = 0{\cdot}06$ to $0{\cdot}40$. A review of the data of Skalka *et al.* (1968) does not indicate any clear discrepancy between their results and ours. Rather, the difference in the two maps of the right arm depends upon two simplifying assumptions that they employ: namely, that the right arm contains DNA of only two G + C classes (0·42 to 0·43 and 0·485), and that the number of boundaries between segments of each class be minimized.

We have three reasons to think that these assumptions, appropriate for their analysis of a small number of terminal fragments, represent an oversimplification with respect to our more extensive data. First, consider the dashed line in Figure 2 which represents the curve of densities expected were the map of Skalka *et al.* (1968) to obtain. The difference between our data and the expectation from their map is significant in two regions centering about $f_R = 0{\cdot}13$ and $0{\cdot}29$. It is this steeper rise to a plateau that is both lower and closer to the right end than predicted from their map that forces us to break up two of their three segments of constant G + C content (Figs 2 and 3). Second, the G + C content of λdv DNA (0·474, see Results, section (a)(ii)) confirms our map and negates theirs for one segment ($f_R = 0{\cdot}121$ to $0{\cdot}266$) of the disputed region. The G + C values computed from our map and from theirs for this segment are 0·471 and 0·485, respectively. Third, the G + C content of a messenger RNA which has been localized to the remainder of the region in dispute ($f_R = 0{\cdot}29$ to $0{\cdot}40$) is 0·45 (Marcaud, Portier, Kourilsky, Cohen & Gros, 1971). This is in better agreement with the value of 0·45 computed from our map than with the value of 0·43 computed from theirs.

The differences between the maps in other regions are more simply explained. We did not examine small enough fragments from the left family to detect the depression in G + C content at the left terminus, and Skalka *et al.* (1968) did not examine fragments that would have facilitated the detection of the depression at f_L = approx. 0·23. Hence we suppose that both exist. Finally, the difference in the central region near the *b2* deletion is within the error of the two methods.

The zones of partial denaturation mapped by electron microscopy after exposure of λ DNA to high temperature or pH in the presence of formaldehyde (Inman, 1967; Inman & Schnös, 1970) are also given in Figure 3. Skalka *et al.* (1968) have noted the rough qualitative correlation between the zones which denature first (class A) and second (class B) with the region of lowest G + C content at the center and with the depressions in G + C content at the two termini; the changes we have made in the map do not affect these. The largest zone in class C, the third in order of denaturation, corresponds to the depression in G + C content that we detect at $f_L = 0{\cdot}23$.

The lack of a quantitative correspondence between zones and the maps of base composition is most easily attributed to the higher resolution obtainable with the electron microscope. For example, a third zone of partial denaturation in class B which is centered about $f_L = 0{\cdot}75$ does not correspond to an obvious trough in either map (Fig. 3). This zone is in the *rex-cI* gene region ($f_L = 0{\cdot}74$ to 0·78; Davidson & Szybalski, 1971) and messenger RNA from this region appears to have a G + C content of 0·43 (Bear & Skalka, 1969). We should not expect to detect such a small trough of G + C = 0·43 in this region; e.g. the density of a right fragment extending to the right-hand boundary of the trough ($f_R = 0{\cdot}22$) would exceed the density of a fragment extending to left-hand trough boundary ($f_R = 0{\cdot}26$) by only 0·0005 g/cm^3. This difference is at the borderline of reliability and would be even smaller for pairs of fragments with ends not so precisely aligned with the boundaries of the trough. It should also be noted that the resolution of our method decreases with increasing distance from an end of λ DNA. A change of 0·05 in G + C content will be detectable only if the length of the segment is at least one-tenth the average length of that pair of fragments the broken ends of which coincide with the boundaries of the segment, the magnitude of the detectable change in G + C content being inversely proportional to this required length ratio. The resolution obtainable by observation of partial denaturation in the electron microscope is higher than this, particularly for regions well removed from the ends of the molecule. Unfortunately, the G + C content of the denatured segments cannot be calculated, since the extent of the denaturation depends upon the time of reaction with formaldehyde and is not an equilibrium process.

(b) *Nature of the poly(rG) binding sites*

(i) *Definition*

We have determined the location of three discrete regions of specific poly(rG) binding in the right half of λ DNA (Fig. 6). Each binds approximately the same amount of poly(rG) (350 residues) and there are no discrete regions binding a lesser amount. We define such regions as single sites.

(ii) *Size*

The amount of poly(rG) bound per site under saturating conditions increases above 350 nucleotides if the mean molecular weight of the poly(rG) is greater than that we normally use, and decreases if it is less, the change in molecular weight being effected

by alkaline hydrolysis (see Experimental Procedure). We, therefore, infer that a significant fraction of the 350 poly(rG) residues associated with each site does not directly interact with bases in the DNA.

Using resistance to T_1 RNase as a criterion, Summers & Szybalski (1968*b*) observed 1150 directly interacting poly(rG) residues per *r*-strand of T7 DNA, which contains all the binding loci. From the estimated size of the resistant poly(rG) molecules, they concluded that the interacting residues were distributed at an average of 20 residues per each of 58 binding loci on the strand.

Our definition of a binding *site* is clearly different from that for the T7 binding *locus*. Using equation (6) and the fact that a given sample of poly(rG) (or poly(rI,G)) interacts with the *r*-strands of λ and T7 to produce the same density increment in each ($\pm$ 25%; Kubinski *et al.*, 1966; Hradecna & Szybalski, 1967; Summers & Szybalski, 1968*a*), we compute that one λ site is capable of binding 8($\pm$ 2)-fold more poly (rG) than a T7 locus. We therefore infer that a λ binding site directly interacts with 160 ($\pm$ 40) poly(rG) residues. The simplest model combining the results from both phages under the supposition that the same functional units are being measured is that T7 DNA contains a total of seven binding sites equivalent to those in λ DNA, and that the T_1 RNase produces internal breaks in the bound poly(rG) molecules in regions of infrequent base pairing. A corollary supposition is that such sites are divisible into several (8 $\pm$ 2) binding loci wherein the frequency of base-pairing is sufficiently high to make the poly(rG) resistant to T_1 RNase.

(iii) *Composition*

It has generally been assumed that poly(rG) binding sites result from a localized concentration of dC residues in the DNA. This assumption is consistent with the following correlation between poly(rG) binding and dC content for the strands of both λ and T7 DNA.

Using the density difference between strands of whole and half molecules of λ DNA in alkaline CsCl, Doerfler & Hogness (1968) concluded that the *r*-strand contained 1440 $\pm$ 230 more dC + dA residues than the *l*-strand (calculated for 46,500 bases/strand) and that 87% of these excess residues are localized in the left half. Mushynski & Spencer's (1970*b*) direct analysis of the base composition of the two whole strands indicates that the *r*-strand contains 1300 $\pm$ 200 more dC residues than the *l*-strand and that there is no significant bias between the strands for dA. Combining these consistent results and assuming that the halves, like the whole molecule, exhibit no bias for dA between strands, the *r*-strand should then contain an excess of 1130 dC residues in the left half and 170 in the right half. This distribution correlates well with an excess of seven binding sites on the *r*-strand in the left half and only one in the right half (Fig. 6), the ratio of excess dC residues to excess sites being 166 $\pm$ 4 for either half. This value is in remarkably good agreement with the estimate made in the preceding section that 160 nucleotide residues of poly(rG) directly interact with each λ binding site. The implication of this correlation is that the directly interacting rG residues form dC·rG pairs. A similar correlation results if one compares Mushynski & Spencer's (1970*a*) finding that the *r*-strand of T7 contains an excess of 1050 dC residues with Summers & Szybalski's (1968*b*) finding that 1150 nucleotide residues of poly(rG) directly interact with this strand.

The simplest model for a binding locus, namely that it consists of long runs of about 20 dC residues, is excluded by Mushynski & Spencer's (1970*a*,*b*) analyses of pyrimidine

tracts ($dC_n\ dT_m$) in T7 and λ DNA which yield upper limits of $n = 8$ and $n + m = 13$ for both DNA's. These workers proposed that the 18 largest tracts ($n + m = 11$ to 13, $n = 4$ to 8) represent the poly(rG) binding loci in T7 as they are restricted to the *r*-strand. We think this unlikely as they contain a total of only 102 dC residues, and the 18 tracts of the same composition in λ are distributed 11 : 7 between *r*- and *l*-strands. One has to descend to smaller tracts ($n + m \leqslant 9$, $n \leqslant 5$) in λ to find one which appears nine or more times in the *r*-strand and at least once in the *l*-strand, the comparison of interest if all λ sites are the same (see section (c) below). We conclude that the small binding loci consist of multiple small runs of dC separated not only by dT but also by purine residues.

(c) *Topographic correlations with the poly(rG) binding sites*

(i) *Positive correlation with the orientation of transcription*

The sequences of base pairs which generate poly(rG) binding sites (*g* sequences) can be divided into two classes according to their orientation: *gr* sequences generate sites on the *r*-strand; and *gl* sequences generate sites on the *l*-strand. The topographic comparisons made in Figure 6 indicate that without exception *gr* sequences are in regions of rightward transcription (i.e. from the *r*-strand), and *gl* sequences are in regions of leftward transcription.

(ii) *Negative correlations*

At the level of resolution indicated in Figure 6, positive topographic correlations can suggest but not identify the functions of the *g* sequences; negative correlations on the other hand can eliminate functions from consideration. We first clear the air of some previously proposed functions (Szybalski *et al.*, 1969) by the elimination process.

Promoters and operators. Inspection of Figure 6 reveals that all three of the known promoters for leftward transcription and one of two known promoters for rightward transcription are located in a region ($f_R = 0{\cdot}18$ to $0{\cdot}37$) which is devoid of *g* sequences. Hence, these promoters do not contain *g* sequences. Since the two operators, o_L and o_R, are adjacent to or interdigitated with two of these promoters, p_L and p_R, respectively (Ordal, 1971; Ptashne, 1971), they also do not contain *g* sequences.

The remaining known promoter, p'_R, cannot be similarly eliminated since it appears to lie within the *gr2* region (Fig. 6). However, we tentatively reject the suggestion that the *gr2* sequence is the p'_R promoter, since acceptance would demand that *g* sequences at different loci have different functions. This is so because the p'_R promoter requires the protein product of the *Q* gene for activation, and transcription to the left of *gl1* or to the right of *gr1* (Fig. 6) is not *Q*-dependent (Nijkamp, Bøvre & Szybalski, 1970; Hendrix, 1971). Because the different *g* sequences yield sites with the same binding capacities for poly(rG), we wish to maintain the unifying assumption that *g* sequences are repeating sequences, yielding a set of ten repeated functions within the λ genome.

The negative correlation between promoters and *g* sequences can be extended to the *bio* segment of *E. coli* DNA. The results of Guha, Saturen & Szybalski (1971) indicate that a promoter for transcription from the *l*-strand of this segment is incorporated into λ *bio y* and λ*bio P* DNA's, which we have shown contain no *gl* sequences (Table 2).

Terminators. Roberts (1969,1970) has identified three sites at which the transcription of λ DNA is terminated *in vitro* in the presence but not in the absence of the rho factor of the host. Two of these rho-dependent terminators (t_L and t_{R1}) are directly eliminated by their positions (Fig. 6). We tentatively eliminate the third (t_{R2}) on the

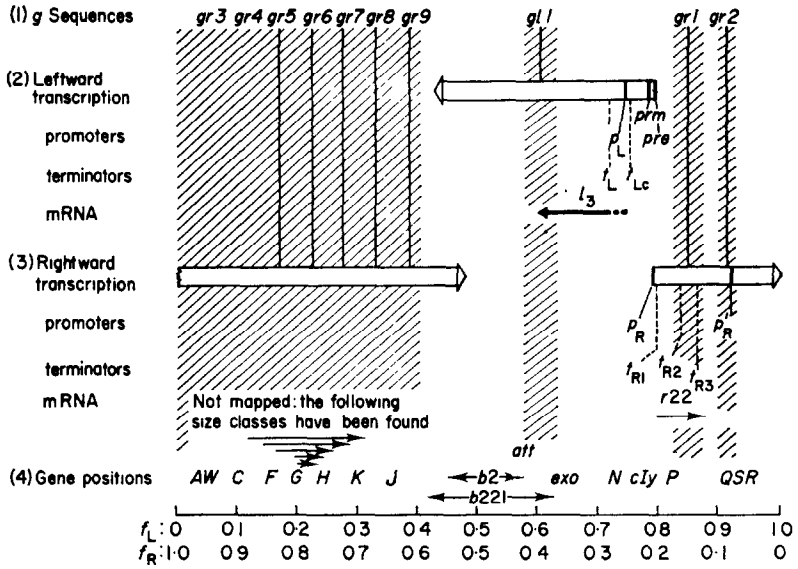


FIG. 6. Topographic correlations between *g* sequences and other functions of λ DNA. The *gr* sequences (*gr1–gr9*) generate poly(rG) binding sites on the *r*-strand; the *gl1* sequence forms the single site on the *l*-strand. The orientation of transcription for the different intervals is taken from Szybalski *et al.* (1969) and Champoux (1970), the small amount of leftward transcription in the *y-P* interval detected by Champoux not being indicated. The position of the promoters p_L and p_R is from Davidson & Szybalski (1971); of *prm* and *pre*, the promoters for *cI*, from Reichardt & Kaiser (1971); and of p'_R, the *Q*-dependent promoter of late transcription lying between *Q* and *S* (Herskowitz & Signer, 1970*a*), from the *Q* and *R* gene positions (Egan & Hogness, 1972). The positions of the ρ-dependent terminators, t_L, t_{R1} and t_{R2} (see Szybalski, 1970, for terminology), were calculated from the 12 s, 7 s and 16 s RNA's initiated at p_L, p_R and c_{17} (a mutant promoter in *y*), while the position of the ρ-independent terminator, t_{R3}, was calculated from the 22 s RNA, also initiated at c17 (Roberts, 1969,1970). The t_{Lc} terminator position was calculated from the size and orientation of the mRNA made in lysogenic cells (Gros, Kourilsky & Marcaud, 1969) which is initiated at *prm* (Reichardt & Kaiser, 1971). The *l3* mRNA is that described by Kourilsky *et al.* (1969,1970) and Marcaud *et al.* (1971); the *r22* RNA is the above 22 s RNA found by Roberts (1970); and the size classes of mRNA from the left arm are from Oda *et al.* (1969). The gene positions in the lower row are taken from the maps of Davidson & Szybalski (1971) and Egan & Hogness (1972).

grounds that termination sequences recognized by the rho-factor would be expected to be the same regardless of their location.

A fourth terminator, t_{Lc}, which may or may not be rho-dependent, can be eliminated directly from its position (Fig. 6). This terminator is active in lysogenic cells to produce the messenger RNA of the *cI* gene.

A fifth terminator, labeled t_{R3} in Figure 6, was characterized from transcription of λ DNA *in vitro* as being rho-*in*dependent and located to the right of t_{R2} (Roberts, 1970). It is located within the *gr1* region (Fig. 6 and legend). We adopt this positive correlation as part of a model for *g* function which is given in the last section.

N-controlled loci. Three loci in λ DNA have the following properties: (a) they connect two DNA segments transcribed in the same direction; (b) transcription of both segments depends upon a single promoter located upstream; (c) efficient transcription downstream from the locus depends upon the presence of the protein product of the *N* gene (Court & Sato, 1969; Heinemann & Spiegelman, 1970; Kumar, Calef & Szybalski, 1970; Nijkamp *et al.*, 1970; Butler & Echols, 1970; Franklin, 1971; Fiandt, Hradecna, Lozeron & Szybalski, 1971). As the three *N*-controlled loci are close to, if

not coincident with, the three rho-dependent terminators, it has been suggested that the *N* protein acts to prevent rho termination (Roberts, 1969,1970), or that rho termination may expose a quasi-promoter which requires the *N* protein for activation. Whatever the mechanism, such loci are eliminated by the same arguments used above to eliminate the rho-dependent terminators.

(iii) *Further positive correlations leading to a model for g function*

All ten *g* sequences are located within the three large units of transcription (transcriptons) defined by the promoters, p_L, p_R and p'_R (Fig. 6). Leftward transcription dependent upon p_L extends more than 12,000 base pairs to the left of p_L to include *gl1*; rightward transcription dependent upon p_R extends at least to p'_R (approx. 7000 base pairs) to include *gr1* and possibly *gr2*; and rightward transcription dependent upon p'_R extends some 26,000 base pairs from this promoter around circular λ DNA (formed by joining left and right ends) to include *gr3-gr9* (see Szybalski *et al.*, 1970 and Herskowitz & Signer, 1970*b* for review).

The clear advantage of such large transcriptons is the small number of control sites required to turn on and off a large number of genes at a given time. A clear problem inherent to very large transcriptons is that of providing a mechanism to regulate the relative rate of synthesis of the many different protein products. The evidence cited in the following paragraphs indicates that the transcript of a given transcripton does not appear as a single long mRNA, but rather takes the form of a set of discrete size classes of mRNA, each representing a fraction of the entire transcripton. This phenomenon suggests a solution to the regulatory problem: namely, that the genes within a transcripton are divided into smaller linkage groups defined by members of the mRNA set, and that translation frequencies are regulated by the concentration and structure of these different mRNA's. We suggest that the *g* sequences function in some manner to indicate the loci at which such division occurs, adopting the term "divider" to indicate this function. We present the topographic correlation between *g* sequences and divider loci which lead to this supposition before a consideration of the possible mechanisms of division and regulation.

p'_R *Transcripton.* A single mRNA containing the entire base sequence of the p'_R transcripton would exhibit a sedimentation coefficient of 60 to 70 S. As indicated in Figure 6, the observed distribution of mRNA from this transcripton includes five discrete size-classes transcribed from the left half which exhibit sedimentation coefficients between 9 and 35 S (Oda, Sakakibara & Tomizawa, 1969). Hence the number of mRNA species formed from a region which contains seven *gr* sequences (*gr3-gr9*) is at least five.

p_R *Transcripton.* Under normal conditions, where the *N*-protein is present to relieve the blocks caused by the rho-dependent terminators (t_{R1} and t_{R2}), two size classes of *N*-dependent mRNA with sedimentation coefficients of 22 S and 28 S are formed from the p_R transcripton in the region containing the two *g*-sequences, *gr1* and *gr2* (Oda *et al.*, 1969). Hybridization of each of these classes† to DNA from deletion mutants led

† Under the conditions employed by Kourilsky, Marcaud, Sheldrick, Luzzati & Gros (1968) and Kourilsky, Marcaud, Portier, Zamansky & Gros (1969), the smaller or *r*1 class of mRNA is formed at significant rates in the N^- state, while Oda *et al.* (1969) indicate this does not occur under their conditions. We attribute this difference to a difference in the degree of leakage past the rho-dependent terminator(s) in the absence of *N* protein, a leakage documented by indirect data; see Herskowitz & Signer (1970*b*) for review. Both groups agree that the larger, *r*2, class is *N*-dependent.

Kourilsky *et al.* (1969) and Kourilsky, Bourguignon, Bouquet & Gros (1970) to place them in tandem order, with the smaller mRNA (*r1*) extending rightward from p_R, and the larger mRNA (*r2*) occupying the adjacent segment which ends at a point to the right of *Q*.

This placement, and the observation that *r1* exhibits the same sedimentation coefficient as the RNA (given as *r22* in Fig. 6) produced *in vitro* to identify the rho-independent terminator, t_{R3} (Roberts, 1970), indicate that a divider is located at or near t_{R3} within the *gr1* region. The errors in estimating the size of these two RNA's and the apparent breadth of their size distributions (Kourilsky *et al.*, 1968; Oda *et al.*, 1969; Roberts, 1970) limit the resolution for the localization of this divider and of t_{R3} to $f_R = 0{\cdot}16 \pm 0{\cdot}02$; however, this range is contained within that for the *gr1* sequence (Fig. 6). If the *r2* mRNA covers the adjacent segment, its estimated length ($f = 0{\cdot}11 \pm 0{\cdot}03$; calculated from data of Oda *et al.*, 1969, and of Kourilsky *et al.*, 1968) places a second divider at $f_R = 0{\cdot}05 \pm 0{\cdot}05$, a region which overlaps that assigned to *gr2* ($f_R = 0{\cdot}09 \pm 0{\cdot}02$).

We do not wish to emphasize the above placement of *r2*, since the hybridization data indicate only that many or most of its sequences are located to the right of gene *P*, and do not exclude sequences to the left of *P* (Kourilsky *et al.*, 1969; Marcaud *et al.*, 1971). Hence it is possible that the *r2* class contains mRNA which originates at p_R ($f_R =$ approx. 0·22), escapes termination at t_{R3}, and ends at a divider located in the region $f_R = 0{\cdot}11 \pm 0{\cdot}03$ which also overlaps with *gr2*. The sedimentation or electrophoretic distributions of *r2* are sufficiently broad to include two subclasses of mRNA, one initiated at p_R, the other at *gr1*, and both ending at a common divider in *gr2*. Indeed, the most recent data on the mapping of *r2* is consistent with this interpretation (Marcaud *et al.*, 1971).

p_L *Transcripton.* The p_L transcripton produces a major *N*-dependent mRNA with a length equivalent to $f = 0{\cdot}11 \pm 0{\cdot}03$ (Kourilsky *et al.*, 1968; Oda *et al.*, 1969). This mRNA is a transcript of the interval represented by the *l3* mRNA shown in Figure 6 (Kourilsky *et al.*, 1968,1969,1970; Marcaud *et al.*, 1971), indicating that a divider is located within the *gl1* region. This location of *l3* is confirmed by the correspondence between its G + C content of 0·45 (Marcaud *et al.*, 1971) and the value of 0·45 calculated for the equivalent interval in our map of base composition (Fig. 2).

The length distribution of the mRNA from the region to the left of *gl1* has not been determined. We predict this mRNA will exhibit considerable length heterogeneity, since the region contains no *gl* sequences and there is evidence that it lacks specific termination sites; i.e. the p_L and p'_R transcriptons overlap in this region, suggesting that termination may result from random "collision" of opposing RNA polymerase molecules (Bøvre & Szybalski, 1969; Nijkamp, Bøvre & Szybalski, 1971).

Mechanisms. Although no single correlation between *g* sequence and divider is convincing, the sum of these correlations is impressive and provides the best formal model for the function of *g* sequences that we have. If we now consider the mechanisms by which this division of the mRNA from a single transcripton may occur, three basic models come to mind.

We can imagine that *g* sequences act to release the mRNA from the transcribing complex without release of the RNA polymerase, which continues down the DNA synthesizing the next mRNA segment—"RNA release" model (see Davis & Hyman, 1970, for a discussion of this mechanism in T7 DNA). According to the simplest version of this model, all segments would exhibit equal transcription frequencies per RNA

polymerase molecule completing a passage along the entire transcripton. However, different frequencies can be included in the model if it is assumed that reinitiation of transcription at a given *g* sequence occurs with a probability less than one, and that the polymerase can slip downstream without transcription to the next *g* sequence where reinitiation may occur. This would lead to decreasing frequencies of transcription as the distance from the promoter increased, unless reinitiation sites coupled to the *g* sequences exhibited different efficiencies.

By contrast, we can imagine that *g* sequences act to increase the probability that the RNA polymerase will dissociate from the transcribing complex, but to a value less than one—"polymerase release" model. By itself, this mechanism would lead to a decrease in the frequency of transcription as the polymerase moved downstream. To permit an increase, some *g* sequences would have to be coupled with quasi-promoters which act as receptor sites for additional polymerase molecules, the activity of such quasi-promoters being dependent upon prior interaction between the *g* sequence and the original polymerase (see Herskowitz & Signer, 1970*a*,*b*, for a discussion of such models).

Finally, one can suppose that the transcript of the *g* sequence in the RNA acts as a site where the RNA is cut (e.g. by a specific endonuclease), transcription itself being uninterrupted—"RNA cut" model.

The topographic correlation between t_{R3} and the divider at *gl1* favors the RNA and polymerase release models where transcription is interrupted, over the RNA cut model where it is not. Similarly, the evidence for weak promoters in the p'_R transcription (between genes *A* and *W*, *C* and *F*, *F* and *G*, and *H* and *K*—see Fig. 6; Toussaint, 1969; Herskowitz & Signer, 1970*a*; Sharp, Donta & Freifelder, 1971) may be indicative of the quasi-promoters in the extension of the polymerase release model, or conceivably, of reinitiation sites in the RNA release model, but cannot be simply related to the RNA cut model.

The fact that the net rate of mRNA formation per DNA base pair in the region to the left of *gl1* is considerably less than that in the region to its right (Kumar, Bøvre, Guha, Hradecna, Maher & Szybalski., 1969; Nijkamp *et al.*, 1971; and unpublished data of Nijkamp, Saturen & Szybalski, as cited in the preceding reference) supports the concept that *g* sequences act as dividers and that such division can result in differences in concentration for different mRNA from the same transcripton. Unfortunately this observation does not differentiate among the basic models, since even the RNA cut model could yield this result if the different mRNA's exhibited different sensitivities to degradation.

We have used the neutral term, divider, not only because the mechanism for division of the transcript is undetermined, but also to emphasize the conceptual difference between dividers and the terminators such as t_L, t_{R1} and t_{R2}, which act to turn off transcription further downstream, unless counteracted by regulatory elements such as the *N* protein. The purpose of this latter class appears to be an enforcement of a desired *off* state of the large transcriptons (e.g. in lysogenic cells) in the event of leakage through the repressor–operator block, and/or to effect a chronologic regulation of transcription. The postulated purpose of dividers, on the other hand, is to divide the very long message from a single transcripton into subunits to facilitate the differential regulation of the translation frequencies of individual genes or groups of genes.

Division of the mRNA can facilitate such differential regulation in two ways. Division can lead directly to different sequences within a transcripton being repre-

sented by different concentrations of the corresponding mRNA's, as indicated in the descriptions of the RNA and polymerase release models. Alternatively, we can imagine that division permits each of the resulting mRNA molecules to adopt secondary structures which either could not be formed, or would have a lower probability of formation if the mRNA were not divided. Not only could such structures control the half-lives of the different mRNA's, but they could also serve to regulate the relative frequencies of translation of different genes within a given mRNA, as evidently occurs in certain RNA phages (Kozak & Nathans, 1972).

In conclusion, we note another possible selective advantage resulting from division which was suggested to us by W. F. Dove. Namely, division could be used to restrict the proteins synthesized on a single polysome to a given class, as for example those proteins which must form a complex to exhibit a required activity. Clearly this advantage and that arising from the differential regulation of translation frequencies are not mutually exclusive.

We thank J. Barry Egan for providing the DNA fragments used in this work and for many helpful suggestions. This research was supported by grants from the National Institutes of Health and the National Science Foundation.

REFERENCES

Arber, W. (1968). *Symp. Soc. Gen. Microbiol.* **18**, 295.
Baldwin, R. L. (1971). In *Procedures in Nucleic Acid Research*, ed. by G. L. Cantoni & D. R. Davies, vol. 2, p. 355. New York and London: Harper & Row.
Baldwin, R. L., Barrand, P., Fritsch, A., Goldthwait, D. A. & Jacob, F. (1966). *J. Mol. Biol.* **17**, 343.
Bear, P. D. & Skalka, A. (1969). *Proc. Nat. Acad. Sci., Wash.* **62**, 385.
Bøvre, K. & Szybalski, W. (1969). *Virology*, **38**, 614.
Bruner, R. & Vinograd, J. (1965). *Biochim. biophys. Acta*, **108**, 18.
Butler, B. & Echols, H. (1970). *Virology*, **40**, 212.
Champoux, J. J. (1969). Dissertation, Stanford University.
Champoux, J. J. (1970). *Cold Spr. Harb. Symp. Quant. Biol.* **35**, 319.
Court, D. & Sato, K. (1969). *Virology*, **39**, 348.
Davidson, N. & Szybalski, W. (1971). In *The Bacteriophage Lambda*, ed. by A. D. Hershey, p. 45. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.
Davis, R. W. & Hyman, R. W. (1970). *Cold Spr. Harb. Symp. Quant. Biol.* **35**, 269.
Davis, R. W. & Parkinson, J. S. (1971). *J. Mol. Biol.* **56**, 403.
Doerfler, W. & Hogness, D. S. (1968). *J. Mol. Biol.* **33**, 635.
Egan, J. B. & Hogness, D. S. (1972). *J. Mol. Biol.* **71**, 363.
Falkow, S. & Cowie, D. B. (1968). *J. Bact.* **96**, 777.
Fiandt, M., Hradecna, Z., Lozeron, H. A. & Szybalski, W. (1971). In *The Bacteriophage Lambda*, ed. by A. D. Hershey, p. 329. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.
Franklin, N. C. (1971). In *The Bacteriophage Lambda*, ed. by A. D. Hershey, p. 621. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.
Gros, F., Kourilsky, P. & Marcaud, L. (1969). *CIBA Foundation Symposium on Homeostatic Regulation*, p. 107.
Guha, A., Saturen, Y. & Szybalski, W. (1971). *J. Mol. Biol.* **56**, 53.
Heinemann, S. F. & Spiegelman, W. G. (1970). *Cold Spr. Harb. Symp. Quant. Biol.* **35**, 315.
Hendrix, R. W. (1971). In *The Bacteriophage Lambda*, ed. by A. D. Hershey, p. 355. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.
Herskowitz, I. & Signer, E. R. (1970*a*). *J. Mol. Biol.* **47**, 545.
Herskowitz, I. & Signer, E. R. (1970*b*). *Cold Spr. Harb. Symp. Quant. Biol.* **35**, 355.
Hirschman, S. Z., Gellert, M., Falkow, S. & Felsenfeld, G. (1967). *J. Mol. Biol.* **28**, 469.

Hogness, D. S. & Simmons, J. R. (1964). *J. Mol. Biol.* **9**, 411.
Hradecna, Z. & Szybalski, W. (1967). *Virology*, **32**, 633.
Ifft, J. B., Voet, D. H. & Vinograd, J. R. (1961). *J. Phys. Chem.* **65**, 1138.
Inman, R. B. (1967). *J. Mol. Biol.* **28**, 103.
Inman, R. B. & Schnös, M. (1970). *J. Mol. Biol.* **49**, 93.
Jovin, T. M., Englund, P. T. & Bertsch, L. L. (1969). *J. Biol. Chem.* **244**, 2996.
Kayajanian, G. (1968). *Virology*, **36**, 30.
Kayajanian, G. (1970). *Virology*, **41**, 170.
Kellenberger, G., Zichichi, M. L. & Weigle, J. L. (1960). *Nature*, **187**, 161.
Kellenberger, G., Zichichi, M. L. & Weigle, J. L. (1961). *J. Mol. Biol.*, **3**, 399.
Kourilsky, P., Bourguignon, M.-F., Bouquet, M. & Gros, F. (1970). *Cold Spr. Harb. Symp. Quant. Biol.* **35**, 305.
Kourilsky, P., Marcaud, L., Portier, M.-M., Zamansky, M.-H. & Gros, F. (1969). *Bull. Soc. Chim. biol.* **51**, 1429.
Kourilsky, P., Marcaud, L., Sheldrick, P., Luzzati, D. & Gros, F. (1968). *Proc. Nat. Acad. Sci., Wash.* **61**, 1013.
Kozak, M. & Nathans, D. (1972). *Bact. Rev.* **36**, 109.
Kubinski, H., Opara-Kubinska, Z. & Szybalski, W. (1966). *J. Mol. Biol.* **20**, 313.
Kumar, S., Bøvre, K., Guha, A., Hradecna, Z., Maher, V. M. & Szybalski, W. (1969). *Nature*, **221**, 823.
Kumar, S., Calef, E. & Szybalski, W. (1970). *Cold Spr. Harb. Symp. Quant. Biol.* **35**, 331.
Marcaud, L., Portier, M.-M., Kourilsky, P., Cohen, A. & Gros, F. (1971). *Molec. Gen. Genetics*, **113**, 355.
Matsubara, K. & Kaiser, A. D. (1968). *Cold Spr. Harb. Symp. Quant. Biol.* **33**, 769.
Mejbaum, W. (1939). *Z. physiol. Chem.* **258**, 117.
Miyazawa, Y. & Thomas, C. A. (1965). *J. Mol. Biol.* **11**, 223.
Mushynski, W. E. & Spencer, J. H. (1970*a*). *J. Mol. Biol.* **52**, 91.
Mushynski, W. E. & Spencer, J. H. (1970*b*). *J. Mol. Biol.* **52**, 107.
Nijkamp, H. J. J., Bøvre, K. & Szybalski, W. (1970). *J. Mol. Biol.* **54**, 599.
Nijkamp, H. J. J., Bøvre, K. & Szybalski, W. (1971). *Molec. Gen. Genetics*, **111**, 22.
Oda, K., Sakakibara, Y. & Tomizawa, J. (1969). *Virology*, **39**, 901.
Ordal, G. W. (1971). In *The Bacteriophage Lambda*, ed. by A. D. Hershey, p. 565. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.
Parkinson, J. S. & Huskey, R. J. (1971). *J. Mol. Biol.* **56**, 569.
Pochon, F. & Michelson, A. M. (1965). *Proc. Nat. Acad. Sci., Wash.* **53**, 1425.
Ptashne, M. (1971). In *The Bacteriophage Lambda*, ed. by A. D. Hershey, p. 221. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.
Reichardt, L. & Kaiser, A. D. (1971). *Proc. Nat. Acad. Sci., Wash.* **68**, 2185.
Roberts, J. W. (1969). *Nature*, **224**, 1168.
Roberts, J. W. (1970). *Cold Spr. Harb. Symp. Quant. Biol.* **35**, 121.
Rothman, J. L. (1965). *J. Mol. Biol.* **12**, 892.
Schildkraut, C. L., Marmur, J. & Doty, P. (1962). *J. Mol. Biol.* **4**, 430.
Schildkraut, C. L., Marmur, J., Fresco, J. & Doty, P. (1961). *J. Biol. Chem.* **236**, PC3.
Sharp, J. D., Donta, S. & Freifelder, D. (1971). *Virology*, **43**, 176.
Skalka, A., Burgi, E. & Hershey, A. D. (1968). *J. Mol. Biol.* **34**, 1.
Summers, W. C. & Szybalski, W. (1968*a*). *Virology*, **34**, 9.
Summers, W. C. & Szybalski, W. (1968*b*). *Biochim. biophys. Acta*, **166**, 371.
Sussman, R. & Jacob, F. (1962). *C.R. Acad. Sci. Paris*, **254**, 1517.
Szybalski, E. H. & Szybalski, W. (1970). In *Handbook of Biochemistry, Selected Data for Molecular Biology*, 2nd Edition, p. H-16. Cleveland: The Chemical Rubber Co.
Szybalski, W. (1970). In *Handbook of Biochemistry, Selected Data for Molecular Biology*, 2nd Edition, p. I-35. Cleveland: The Chemical Rubber Co.
Szybalski, W., Bøvre, K., Fiandt, M., Guha, A., Hradecna, Z., Kumar, S., Lozeron, H. A., Maher, V. M., Sr., Nijkamp, H. J. J., Summers, W. C. & Taylor, K. (1969). *J. Cell Physiol.* **74** (suppl. 1), 33.
Szybalski, W., Bøvre, K., Fiandt, M., Hayes, S., Hradecna, Z., Kumar, S., Lozeron, H. A., Nijkamp, H. J. J. & Stevens, W. F. (1970). *Cold Spr. Harb. Symp. Quant. Biol.* **35**, 341.

Szybalski, W., Kubinski, H. & Sheldrick, P. (1966). *Cold Spr. Harb. Symp. Quant. Biol.* **31**, 123.
Toussaint, A. (1969). *Molec. Gen. Genetics*, **106**, 89.
Westmoreland, B. C., Szybalski, W. & Ris, H. (1969). *Science*, **163**, 1343.