Symmetry, homology, and phrasing in the recognition of helical regulatory sequences in DNA

Michael C. O'Neill

Department of Biological Sciences, University of Maryland Baltimore County,
5401 Wilkens Avenue, Baltimore, MD 21228, USA

Received 14 November 1977

ABSTRACT

Regulatory regions in DNA which have been sequenced have generally been found to contain one or more axes of two-fold rotational symmetry. If this symmetry is to be maintained in the helical sequence, the axis of rotation must be aligned with one of the two dyad axes of the helix. This is equivalent to saying that the rotational symmetry of the sequence can only be seen from certain viewing points in a circuit about the helix. More surprising is the fact that new symmetrical sequence arrangements can be seen at $+36^\circ$, $+72^\circ$, $+108^\circ$, and $+144^\circ$ relative to the point at which the rotational symmetry is seen. This "amplification" of symmetry suggests a three-dimensional approach to sequence analysis. A specific reading frame, suggested by the geometry of the helix, is examined with regard to its elucidation of intra- and inter-sequence homologies. Two sequences are thus identified as being recurrent in a number of different regulatory sequences.

A number of regulatory regions in DNA have now been sequenced. The underlying strategy was the hope of deducing "rules" of DNA-protein recognition from comparisons of sequences known to be recognized by the same protein (e.g. the multiple promoters for *E. coli* RNA polymerase) or competing proteins (e.g. promoter-operators recognized by both a repressor and RNA polymerase). Although certain sequences must contain recognition elements for various proteins, such as RNA polymerase, repressors, CAP protein, etc., in no case does the linear sequence clearly indicate which elements of the sequence are important for recognition. This, in turn, suggests that this approach may be too simple. What sources of information for specifying DNA-protein recognition are neglected in analyzing only the linear sequence? There are two additional factors. One is the positional information conferred on the sequence by its three-dimensional context within the helix. The other is the stabilizing effect of cooperative interaction with adjacent protein(s). This paper is concerned with the first of these, the effect the double helical format on the symmetry of regulatory sequences and on the utilization of that symmetry in recognition. This analysis develops the following three points:

- 1.) Recognition sequences which show rotational symmetry in the linear sequence present new symmetrical arrangements when viewed in their helical form; these symmetries are not obvious in the linear form.
- 2.) It is known that many regulatory sequences interact with two or more proteins. The geometry of the double helix provides, implicitly, a means of interlacing or "multiplexing" recognition sites for several proteins within a single short stretch of DNA.
- 3.) If it is assumed that helical sequences are read in a regular pattern, dictated by the regular repeats in the helix geometry, regions of homology are seen between the lac and lambda operators and, indeed, in the majority of known regulatory sequences.

A few general remarks about representations of double helical sequences of DNA are in order at the outset of this discussion. If one were to walk around a model of the double helix, one would see that the base pairs reproduce the original viewing register at every 36° , a consequence of the fact that there are 10 base pairs per complete turn of the helix. For example, the base pair lying perpendicular to the line of view centered in a major groove at a starting point would differ from the base pair in this same position seen at 36° in either direction from the first view, and this would again be true at each additional 36° increment. In this sense the helical sequence presents ten views or faces, spaced at 36° , in which the base pairs present the same positional register.

Let us compare a linear sequence with helical views of the same sequence. Figure 1 shows such a representation of a consensus sequence drawn from all of the rightward and leftward repressor binding sites in lambda where the rightward and leftward operators each contain three repressor binding sites.¹ (Each position in this sequence is assigned on the basis of the base pair occurring most often at that position in a comparison of the six sites.) The central helix is designated as the 0-face (it will be shown below that this is not an arbitrary choice); viewing points offset 36° and 72° in either direction produce the +1, +2 and -1, -2 faces. Three base pairs are exposed in each groove, a major and a minor groove per complete turn; two unseen base pairs lie behind the backbone at each half turn, almost parallel ($\pm 18^\circ$) to the line of view. The sequence reading "down" the left side of a major groove has 5' $\rightarrow$ 3' orientation. The sequence exposed in the grooves is different at each register of the five faces of the helix shown. (For considerations of symmetry, it is not necessary to examine all ten faces because any symmetrical arrangement appearing on the "rearward" faces can also be recognized on the "forward" faces.) One can obtain an appreciation of the changing perspective on different sides of the helix by focusing on a

Figure 1. Linear and Helical Views of a Consensus Sequence of the Lambda Operators. The linear sequence shown at the left was constructed by taking, at each position, a consensus of the rightward and leftward repressor binding sites, O_R1 , O_R2 , O_R3 , O_L1 , O_L2 and half of O_L3 . The five helices show those parts of this sequence which appear in the grooves as the viewing position shifts in 36° increments either to the right (+1, +2 faces) or to the left (-1, -2 faces) of the central 0-face view. The face showing rotational symmetry was chosen as the 0 face. The two base pairs most nearly parallel to the line of view, hidden by the sugar-phosphate backbone at each half turn, are not shown.

single base pair, for instance the A-T pair centered in the major groove at the lower right of Figure 1, and observing it as the view point shifts to the left around the helix. On the +1 face the same pair is seen at the top of the major groove. Continuing, one loses sight of this pair behind the backbone on the 0 and -1 faces until it reappears as a T-A pair at the bottom of a minor groove on the -2 face. Another 36° would bring you half circle; the A-T pair, originally centered in a major groove, would be seen from the opposite side as an T-A pair centered in a minor groove. Likewise, the effect of moving to the right rather than the left can be seen in tracing a base pair from the -2 face to the +2 face. The second operation one must consider is a rotation, not about the center of the

helix, but about an axis perpendicular to the axis of the helix. If, instead of moving about the helix, one turns the model upside down while maintaining the same viewpoint, a sequence which reads A-T, G-C, G-C in a major groove will now read C-G, C-G, T-A in a major groove. With these two operations in hand, we can proceed to a consideration of the symmetry of helical sequences.

What happens to the two-fold rotational symmetry commonly observed in the linear sequence of regulatory regions when the sequence is viewed in its helical configuration? (Although supercoiling may be involved with some recognition sites opening up, several cases that have been studied to date, RNA polymerase and the lac and lambda repressors,^{2,3,4} show little or no unwinding in the initial recognition event.) First note that rotational symmetries fall into two classes: one contains sequences which are symmetric about an axis running through a base, e.g., $\begin{matrix} \text{AAAAGTTT} \\ \text{TTTCAAAA} \end{matrix}$; the others are symmetric about an axis falling between two base pairs, $\begin{matrix} \text{GAAAATTTT} \\ \text{CTTTTAAAAG} \end{matrix}$. The lambda operator sequences provide an example of the first type of symmetry. Although the rotational symmetry of many of these sites is weak, the symmetry of the consensus sequence is very good, $\begin{matrix} 5' \text{TAAATTATCACCGCCAGTGATAATTAA} \\ 3' \text{AATTAATAGTGGCGGTCACTATTAATT} \end{matrix}$, how might a repressor which exists as a dimer or higher order oligomer^{5,6} recognize and bind this specific sequence? By considering only the linear sequence, one might imagine that each half of the sequence must accommodate a subunit in a tail to tail arrangement. However, a different perspective comes from viewing binding of repressor to the sequence in its helical form. To a precision of one degree, there are 360 possible views (36 sets of 10 faces each) of any helical sequence. Assuming the rotational symmetry of these sequences to be important, one can ask if symmetry considerations provide any clues as to which faces are involved in recognition. The answer is yes in that the rotational symmetry of the sequence is not apparent in most views of the helical sequence. Only those (2) views which find the symmetry axis of the sequence exactly centered in a groove of the helix will present the rotational symmetry originally evident in the linear sequence.

In Figure 1, the 0-face view shows the same rotational symmetry present in the linear sequence; this is, in fact, a defining characteristic in that the 0 and 5 faces are the two views in which the sequence axis is superimposed on the helix axis. Yet the symmetry of the linear sequence is not only reproduced but amplified. The sequence on the +1 face is closely duplicated by turning the -1 face sequence upside down; the sequence on the +2 face is similarly a rotated version of that on the -2 face. Here each sequence in the symmetric couple is twice the size of the rotational element (half sequence) in the linear sequence. (The same symmetrical arrangements appear again on the rearward faces of the

helix but contain different sequence information, since AT in a major groove will appear as TA in a minor groove on the opposite side of the helix.) Is this apparent amplification of the symmetry seen in the linear sequence real? The same base pairs are appearing on these different faces. Can they be read simultaneously on different faces, for instance, by subunits of a dimer? In the case of the $+2$ faces and the $+3$ faces, the paired faces lie at 144° relative to one another, with the result that base pairs being read on, say, the $+2$ face are being read from the opposite side on the -2 face with no necessary steric hindrance. In the case of the $+1$ faces, a subunit might read the upper half of the -1 face and another read the lower half of the $+1$ face, again without necessary steric hindrance. Since there are 4 paired faces, plus the 0 and 5 faces each with rotational symmetry, and each symmetric couple can potentially be read two ways - tail to tail or head to head - it is easy to model specific sites for 20 or more distinct dimers on the sequence of Figure 1, with as many as 3 of these sites occupied simultaneously!

It is a rather simple exercise to show that the face symmetry arrangements described here are a general property of linear sequences with rotational symmetry. For instance the base pairs third from the top and third from the bottom on the 0 face in Figure 1 are related to each other by a 180° rotation; as the view moves to the left 72° (-2 face), the pair originally third from the top is now seen at the top of the groove; in the same way, at 72° to the right ($+2$ face) of the 0 face, the pair originally third from the bottom is now seen at the bottom position. Thus, the top base pair on the -2 face is still related to the bottom pair on the $+2$ face by a 180° rotation. Since the same is true of any symmetric pair of base pairs on the 0-face, the whole sequence on the -2 face is a rotated version of that of the $+2$ face to the extent that the rotational symmetry of the linear sequence is good.

It is interesting to consider the extreme symmetry case. Figure 2 shows a hypothetical sequence which shows near perfect rotational symmetry on all ten faces. In the lambda example, there were a number of different arrangements which could accommodate a dimer; but here a hexamer of identical subunits could be accommodated using the $+2$ faces or some combination of the 0, $+2$, and $+4$ faces, all of which contain repeats of the same sequence element.

Let us apply this type of analysis to the interaction between the lac repressor and its operator⁷⁻¹⁰. In this system, the repressor is known to be a tetramer of identical subunits¹¹, and the linear operator sequence shows good symmetry of the first type (e.g. $\begin{smallmatrix} \text{AAGTT} \\ \text{TTCAA} \end{smallmatrix}$). Figure 3 shows the linear sequence along with its helical representations. Again, besides the rotational symmetry of the 0-face, the $+1$ faces and $+2$ faces, as well as their rearward counterparts,

Figure 2. A Sequence with Maximum Helical Symmetry. A hypothetical sequence is shown which presents rotational symmetry on all ten faces around the helix. The five "forward" faces, relative to the 0-face, are depicted. Here the major and minor grooves have been artificially aligned for the purpose of comparison. The two base pairs which are considered as hidden by the backbone at each half turn are shown in smaller type. The arrow indicates the axis of rotation.

show additional symmetries. If one assumes that all four subunits are involved in recognition and that each must find a region within the operator presenting approximately the same sequence, the problem becomes tightly constrained. This is especially true when one considers the placement of four subunits whose individual dimensions are $30\text{\AA} \times 45\text{\AA} \times 60\text{\AA}$.⁸ If one reads only the major grooves on the +2 faces in Figure 3, one finds

where the spaces represent seven intervening base pairs and the arrows indicate rotational axes. As seen in Figure 4, this would provide each subunit of the tetrameric repressor with a recognition sequence of size six, of which an average of 4 base pairs repeat in each of the 4 subunit sites.

What evidence is there which supports such a model? The +2 and -2 faces are 144° apart, yielding an offset of 4 base pairs between the ends of the recognition sequences on the two faces; therefore the minimum overall length of the repressor binding sequence would be $126\text{\AA}$ in good agreement with the size ($130\text{--}140\text{\AA}$) determined by Steitz et al. for the long axis dimension of the repressor molecule.⁸ Because of the offset, the sites on the two faces overlap over a distance of only approximately $99\text{\AA}$ or 29 base pairs compared to the 27 base pair repressor-protected fragment found by Gilbert and Maxam.⁷ The first base pair of the sequence they determined lies 1 base inside the region where the two face sites overlap, placing the protected fragment symmetrically within the 29 base pair overlap region. This would in turn suggest that the DNA is protected from nuclease attack only within the region covered on both sides by repressor subunits. Moreover, with only one exception, the locations containing 0° mutations,⁹ base pairs conserved in the operator and pseudo operators¹⁰ (Gilbert, personal communication), and bases whose methylation is blocked by repressor¹⁰ fall either in the proposed recognition sites (19/25) or within 2 base pairs of the "seam" where two subunits meet on one face (5/25). The arrangement, with subunits at the corners of a parallelogram, is consistent with D_2 symmetry.⁸ If each subunit contains two DNA binding sites as Steitz et al. suggest,⁸ the electron microscope observations of Hirsch and Schleif,¹² which seem to show the DNA running along one edge of the repressor molecule, as opposed to a more central location, could also be accommodated. None of the 0° mutations improves the symmetry of the proposed recognition groups if the sequence unit is $\begin{smallmatrix} 5' \text{GTG} \\ 3' \text{CAC} \end{smallmatrix}$ as opposed to $\begin{smallmatrix} \text{TTG} \\ \text{AAC} \end{smallmatrix}$. To date, no 0° mutations have been located in the four outer triplets shown in Figure 4. However, an examination of Figure 3 will show that there is no sequence, even of size 3, which appears four times without overlap within the central 17 base pair segment which contains all the 0° mutations and which has recently been determined to be the minimum sequence required for repressor binding.¹³

There is also a suggestion that the +2 faces are utilized in the lambda operator (Figure 1). Two facts support an increased conservation specifically of

Figure 3. Linear and Helical Representations of the Lac Operator Sequence. Positions -8 to +29 of the lac operator sequence are shown both in linear and in five "forward" face helical representations. The symmetry elements of the linear sequence are most apparent on the 0-face helical sequence. Again, as in Figure 2, the major and minor grooves have been aligned. The bars indicate the three distinct symmetrical aspects of the sequence presented on different faces of the helix. The arrow indicates the axis of rotation, located here in the center of a narrow groove.

those base pairs exposed in the grooves on the +2 faces: 1) In a comparison of O_L^2 , O_L^3 , O_R^1 , O_R^2 , and O_R^3 with O_L^1 , agreement is lowest on the 0-face (59% of the exposed pairs the same as O_L^1) and highest on the +2 faces (77% agreement); 2) five of the six O^C mutations¹ fall behind the backbone on the 0-face, whereas all appear in groove positions on the +2 face with five of the six exposed on both faces.

Figure 4. A Model for Lac Repressor Binding. This is a model of the tetrameric lac repressor interaction site within the lac operator sequence. The symmetrical elements from the +2 and -2 faces in Figure 3 are shown here, appearing in the wide grooves. The faces shown lie almost opposite one another on the helix (144°). Therefore the repressor subunits, indicated here schematically as side by side would actually sandwich the DNA. The positions of O^C mutations are indicated. The base pairs conserved in the operator and pseudo-operators are indicated by C. The purines whose methylation is blocked by repressor are indicated by arrowheads.

Now let us consider a sequence with rotational symmetry of the second type, for instance $AAGCGCGCGCGCGCTT$. As seen in the second column of Figure 5, such a sequence can present a mirror symmetry when viewed in its helical form. At first, this seems of little biological importance since identical subunits in a dimer cannot utilize sequences which, as mirror images, are stereochemically distinct. However, it is a general property of this type of symmetry that the face sequence which shows mirror symmetry will be repeated on the exact opposite side of the helix, but running in the opposite direction (Figure 5). Such an arrangement could accommodate a dimer, or it could increase the number of available binding sites for a regulatory protein which can function in either orientation. An example of this type can be found in a portion of the attenuator

TRP ATTENUATOR

	-144°	-72°	0°	+36°	100°	+180°
AT	GC*	GC*	AT	AT	AT	AT
AT	CG*	AT	TA*	AT	AT	AT
GC	GC	GC	GC	AT*	AT*	GC
	cq	cq	at	at	at	cq
	gc	cq	cq	ta	cq	gc
GC	GC	GC	GC	TA*	GC	GC
CG	CG	GC*	CG	CG	GC*	CG
GC	GC	GC	GC	GC	GC	GC
	cq	cq	gc	qc	cq	cq
	qc	cq	cq	cq	cq	gc
GC	GC	GC	GC	GC	GC	GC
CG	CG	GC*	CG	CG	GC*	CG
GC	TA*	GC	GC	GC	GC	GC
	ta	cq	gc	qc	cq	cq
	at	at	cq	cq	cq	at
GC	AT*	AT*	GC	GC	GC	GC
AT	AT	AT	AT	CG*	AT	TA*
AT	AT	AT	AT	GC*	GC*	AT

Figure 6. A Multifaceted Sequence. The sequence shown is a portion of the trp leader of Bertrand et al.¹⁴ in the attenuator region. A consensus of 6 faces is shown at the left; the asterisks mark divergence from the consensus. All six faces show approximately the same sequence. Note that the positive faces here are not generated as in the previous figures; they are the opposite sides of the -144°, -72° and 0° faces, all shown inverted.

the line of view versus three out of five if a base pair is centered in the groove as in the previous figures.

With the exception of the model for the repressor site in the lac operator, there has been little in the discussion to this point that is conjectural. The symmetry properties described are a direct product of combining sequence symmetry with the symmetry of the double helix. It is clear that the first is amplified by the second. How these symmetries are used, however, is quite another question, and one for which there are only speculative answers at present. How does a protein read a sequence? Does it manage to sense all the bases along a continuous stretch, or are some bases at unfavorable angles or unrecognizable because of steric hindrance from the sugar-phosphate backbone? In the latter event it is possible that the regular geometry of the double helix dictates a reading frame in the form of a regular pattern base pairs exposed to a globular protein sitting

-2	-1	0	+1	+2
TA				
TA	TA			
CG	TA	TA		
C	G	A	A	
GC	CG	GC	AT	AT
CG	GC	CG	GC	AT
GC	CG	GC	CG	GC
CG	GC	CG	GC	CG
C	G	C	G	C
GC	CG	GC	CG	GC
CG	GC	CG	GC	CG
GC	CG	GC	CG	GC
CG	GC	CG	GC	CG
C	G	C	G	C
GC	CG	GC	CG	GC
CG	GC	CG	GC	CG
TA	CG	GC	CG	GC
TA	TA	CG	GC	CG
	A	A	G	C
		AT	AT	GC
			AT	AT
				AT

Figure 7. A Helical View Preserving Rotational Symmetry of the Second Type. The 0 face shows rotational symmetry about its center. The +1 and +2 faces are rotated versions of one another.

on a particular side of the helix. This is, of course, a large assumption, considering the myriad of other possibilities. It has the advantage that it suggests a specific approach to sequence analysis which can be quickly invalidated if it produces no interesting results. In the previous discussion, two alignments of the bases were considered, one which had a base pair centered in each groove and one which had the space between two base pairs centered in each groove. These alignments correspond to the two dyad axes of the double helix and, as such, are the only two of all possible alignments which allow the rotational symmetry of the linear sequences to appear again on the helix. As noted above the former of these permits greater exposure of base pairs in the grooves. In this case, a space-filled model suggests a pattern of three exposed base pairs, two hidden behind the backbone, three exposed, two hidden, and so on. Several reasons suggested an examination of this pattern as a possible "reading frame": 1.) If proteins read the continuous sequence, one would expect more homology to be obvious in, say, the E. coli promoters, since the repeated occurrence of a

block of bases should be easily identified.

2.) Though other regular reading frames are, of course, possible -- the second type alignment might be compatible with a read four, skip one, read four, or with a read two, skip three, read two -- the three-two pattern is the most appealing compromise in terms of geometry and information content, i.e., the former would require the reading of steeply angled base pairs (36°) and the latter would provide only 12 bases for information in $102\text{\AA}$ along the helix.

3.) Although not a sine qua non for multiplexing, a discontinuous reading frame is appealing in this regard since base pairs behind the backbone would be free for specification on a different face, allowing greater flexibility in inter-lacing separate regulatory sites.

4.) Such a reading frame would allow globular peptide subunits to occupy linear sites on the DNA aligned with the axis of the helix.

If such a reading pattern is used in DNA-protein recognition, reading sequences with a mask which duplicates the pattern might be expected to show homologies that were otherwise not apparent. In a previous paper¹⁵ I noted that regions of the lac and lambda operators are related by mirror symmetry when read in this way. Further consideration of the properties of mirror symmetry suggested that these sites might show a direct homology on other faces. The top half sequence from the -1 face of the lac operator in Figure 3 and the bottom half sequence from the 0-face of the lambda (consensus) site of Figure 1 read

Remembering that each of these sites has a rotational axis and applying the symmetry arguments above, one realizes that the other "halves" of these sites must also show distinct similarities. In the case of lambda the half sequence appears again in the top half of the 0-face, and in the case of lac the bottom half of the +1 face should approximately repeat the top half of the -1 face. One finds

This unlikely coincidence suggested an evolutionary relationship between the two operators with both sequences evolving from a common half sequence. If a base pair were deleted or added to this half sequence, in one of the two cases, prior to the duplication events producing the inverted repeats, the pattern observed in the two sites would result. Figure 8 shows such a scheme, indicating the steps

Figure 8. A Hypothetical Precursor to the Lac and Lambda Operators. The central sequence shown is a hypothetical half-sequence from which operator sequences of either the *lac* type or λ type could have evolved by the steps indicated.

required had the two sites evolved from a common "precursor" sequence. In each case, one or two mutations in the hypothetical precursor, followed by the duplication of a 16 base pair region which begins with the sequence (T) GTGT produce sites of the *lac* or lambda type. Examination of the half-sequences of the *lac* and lambda repressor binding sites in Figure 9 (lambda sequences are shifted down one position in Figure 9 relative to Figure 8) will show that the sequences obtained in Figure 8 are truly representative of both types though not identical to any one of the sequences in either group. The deletion has the effect of offsetting the rotational (and derivative) symmetry in the two cases. The other

mutations effect maximum diversion in the lac tetramer recognition group suggested above. The one base pair which is conserved in all the half-sequences of the three lac repressor binding sites and the two pairs conserved in all of the lambda half-sequences are located precisely at the two (non-deletion) mutation sites suggested here, with the result that heterologous repressor interaction would be further discouraged. Another protein could still recognize the homologous regions, noted earlier, to accomplish a regulatory function shared by the two operators.

A computer program was written to utilize this reading frame in a search of additional regulatory sequences. Figure 9 shows partial sequences from the lambda operators, the lac operator and pseudo-operators, the lac CAP protein binding site,^{16,17} the trp attenuator region, the trp promoter-operator,¹⁸ the phage fd gene 3 promoter,¹⁹ the ara BAD promoter (Kaplan, Maxam, and Schlieff, personal communication), the gal operator and promoter,²⁰ the SV40 promoter,²¹ the SV40 replication origin,²⁷ and the T7A2 and T7A3 promoters.²³ In many of these sequences, a gross pattern can be discerned in which the central minor groove (16-18) is occupied by G-C pairs, most often C-G with the next minor groove on either side (6-8 and 26-28) tending toward all A-T pairs. In accordance with the scheme of Figure 10 the major groove below the central groove tends toward GC (λ) or AT (lac) albeit with some variation. The major groove above center either repeats this pattern (lac, ara, CAP, gal) or favors TA (λ, fd, trp atten, trp P-02, T7A3). A more quantitative assessment of the degree of homology between these sequences is provided in the lower panel of Figure 9. Here the bars show the location in the two indicated sequences of a match in which at least 8 of 12 base pairs appearing in 4 consecutive grooves are identical. The matches do not necessarily fall on the particular face shown in Figure 9. For instance, O_L1 and the T7A2 promoter do not appear very similar in Figure 9, however, if the +2 faces are compared beginning at position 4, one finds

In all, 138 matches are found, exclusive of matches within the lac operator and pseudo-operators and within the lambda operators. It should be noted that the criterion adopted here for a "match" is deliberately liberal. That is, the probability of a single match of 8 out of 12 in comparing two sequences of the length under consideration here is high. For two random sequences with 24 base

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	
	O _L ¹	O _L ²	O _L ³	O _R ¹	O _R ²	O _R ³	f _{dg3}	trp	T7A3	trp	PO-1	trp	SV40P	SV40P	SV40	T7A2	lac O	lac p-O ₂	lac p-O ₁	ara 1	ara 2	gal 1	gal 2	lac CAP
	ATTEN																							
1	TA			TA						AT		TA			AT	GC	GC	CG	CG	AT				
2	GC			GC						CG		TA			TA	TA	TA	CG	CG	AT	CG			
3	TA			CG						TA		TA			TA	CG	GC	GC	CG	TA				
4	GC			AT					AT	GC	CG	TA			GC	TA	TA	AT	AT	CG				
5	GC			TA					GC	CG	CG	TA			GC	CG	CG	CG	AT	CG				
6	AT			GC					AT	CG	AT	TA			AT	CG	TA	CG	AT	CG	TA	GC	TA	
7	GC			GC					CG	CG	AT	TA			TA	TA	TA	CG	TA	TA	TA	TA	AT	TA
8	TA			TA					TA	CG	TA	AT			CG	TA	TA	CG	TA	TA	TA	TA	AT	TA
9	CG			AT					CG	AT	CG	TA			CG	AT	TA	AT	CG	TA	TA	CG	CG	CG
10	TA			TA					TA	CG	AT	TA			TA	TA	TA	TA	TA	TA	TA	TA	TA	TA
11	AT			AT					AT	CG	TA	AT			AT	GC	CG	CG	CG	CG	CG	CG	CG	CG
12	TA			TA					TA	TA	AT	TA			AT	TA	TA	TA	TA	TA	TA	TA	TA	TA
13	CG			CG					CG	CG	CG	CG			CG	CG	TA	TA	CG	TA	CG	CG	AT	CG
14	AT			AT					AT	CG	TA	AT			AT	AT	AT	AT	AT	AT	AT	AT	AT	AT
15	GC			GC					GC	CG	CG	CG			GC	CG	AT	CG	CG	CG	CG	CG	CG	CG
16	GC			GC					GC	CG	CG	CG			CG	TA	CG	CG	CG	CG	CG	CG	CG	CG
17	CG			CG					CG	CG	CG	CG			CG	TA	CG	CG	CG	CG	CG	CG	CG	CG
18	GC			GC					GC	CG	CG	CG			CG	TA	CG	CG	CG	CG	CG	CG	CG	CG
19	GC			GC					GC	CG	CG	CG			CG	TA	CG	CG	CG	CG	CG	CG	CG	CG
20	AT			AT					AT	CG	TA	TA			AT	TA	CG	CG	CG	CG	CG	CG	CG	CG
21	GC			GC					GC	CG	CG	CG			AT	TA	CG	CG	CG	CG	CG	CG	CG	CG
22	TA			TA					TA	CG	TA	TA			TA	TA	CG	CG	CG	CG	CG	CG	CG	CG
23	GC			GC					GC	CG	CG	CG			CG	CG	CG	CG	CG	CG	CG	CG	CG	CG
24	GC			GC					GC	CG	CG	CG			CG	CG	CG	CG	CG	CG	CG	CG	CG	CG
25	AT			AT					AT	CG	AT	TA			TA	TA	CG	CG	CG	CG	CG	CG	CG	CG
26	TA			TA					TA	CG	TA	CG			TA	TA	CG	CG	CG	CG	CG	CG	CG	CG
27	TA			TA					TA	CG	CG	TA			TA	TA	CG	CG	CG	CG	CG	CG	CG	CG
28	AT			AT					AT	CG	CG	TA			AT	TA	CG	CG	CG	CG	CG	CG	CG	CG
29	AT			AT					AT	CG	AT	TA			AT	TA	CG	CG	CG	CG	CG	CG	CG	CG
30	AT			AT					AT	CG	AT	TA			AT	TA	CG	CG	CG	CG	CG	CG	CG	CG
31	TA			TA					TA	TA	AT	TA			TA	TA	CG	CG	CG	CG	CG	CG	CG	CG
32	GC			GC					GC	TA	AT	TA			GC	CG	CG	CG	CG	CG	CG	CG	CG	CG
33	TA			TA					TA	TA	AT	TA			TA	TA	CG	CG	CG	CG	CG	CG	CG	CG

Figure 9A. Partial Homologies in Regulatory Sequences. Partial sequences from a number of operators and promoters are shown here, aligned to emphasize partial homologies. These are, from left to right, the λ operators, the phage fd gene 3 promoter (-39 to -63), part of the trp leader sequence in the attenuator region (-50 to -23), the T7A3 promoter (+17 to -8), two partially overlapping regions of the trp promoter-sequence (-18 to +15 and -31 to -9) two regions of the SV40 replication origin, the T7A2 promoter (-3 to +18 the lac operator, the Z gene lac pseudo-operator; the I gene lac pseudo operator, two regions of the ara operon (-18 to +10 and -79 to -49), and the CAP binding sequence in the lac operon. The lac operator and pseudo operators are shown as half-sequences, split at the axis of rotation. In this format, the left base in position 1 is a 5' end with the 5' $\rightarrow$ 3' strand appearing on alternate sides in alternate grooves. The two backbone base pairs at each half turn are set off by lines, beginning with positions 4 and 5.

Figure 9B. The bars in this figure indicate regions of partial homology. A bar in the O_L 1 sector represents a sequence match between O_L 1 and a second sequence identified by the number (from Figure 9A) at the top of the bar. The location of the bar shows the position of the match in O_L 1; the starting position of the match in the second sequence is given by the number at the bottom of the bar. In some cases the matches are found in opposite orientation, i.e., running top to bottom in one sequence bottom to top in the second; these are indicated by an R beside the sequence identification number. The search employed here required that 8 out of 12 base pairs appearing in 4 consecutive grooves be identical on some face of the helical sequences. The numerous matches with the λ operators and within the lac operator and pseudo operators have been ignored. A given match is represented only once so that a match between O_L 1 and fdG3 will appear under O_L 1 but will not be shown again under fdG3. This lack of redundancy results fewer matches appearing in the right hand side of the figure.

pairs each, one or more such matches would occur with a probability of about one in four; for 30 base pairs, the probability becomes three in five. However, multiple matches and/or more perfect matches (9, 10, 11 out of 12) become rapidly more improbable, e.g., two or more matches of size 9 out of 12 between two sequences of 24 base pairs (average length in Figure 9) would have a random probability of ~ 0.007 . Even at the 8 of 12 level, meaningful homologies may be missed; there is no match of this size between the two lac pseudo-operators. At the more stringent 9 of 12 level, one begins to miss clear functional homologs, with no match being found between O_R2 and either O_R1 or O_L2 . Figure 9 suggests a grouping of the λ , fdG3, trp atten, trp P- O_2 , T7A2, T7A3 and SV40 sequences on the one hand and on the other lac, ara 1, CAP, and to a lesser extent ara 2, gal 1, and gal 2. Table 1 lists more stringent matches.

In an attempt to run an empirical control on the significance of these matches, the O_L1 , trp attenuator, T7A3 promoter, the trp P- O_2 , and the lac operator sequences were compared to a group of 24 "sequences" arbitrarily chosen from the total sequence of the MS2 A-protein gene²⁴ which, as a structural gene in an RNA phage, was assumed to be unrelated to regulatory sequences. In this comparison, U in the RNA sequence was read as T. The test sequences were drawn from positions 129 to 848 of the A gene sequence in lengths of either 24 or 30 base pairs each, in two separate comparisons. The frequencies predicted for random matches were generally born out in comparing the regulatory sequences and the test series, with one rather glaring exception. The third test sequence (189 to 219) gave multiple matches (8 size 8, 4 size 9) with O_L1 . A closer examination showed that positions 213 to 197 of the A gene matched 13 of the 17 base pairs in the top half sequence of O_L1 as it is shown in Figure 9. This is a more extensive match than is found between the two half sequences of O_L1 . The random occurrence of such a match in this test series has a very low probability. At this time, it cannot be decided whether this is in fact a low probability coincidence or presages, possibly along with the pseudo-operators in the I and Z genes of lac, the existence of regulatory regions within structural genes. One still has the operational criterion that a stricter definition of a match than that employed in Figure 9 results in the failure to identify known homologous sequences.

A search to determine whether any single sequence occurred repeatedly in the collection of Figure 9 suggested that variations on the linear sequence 5'TTatCTTaaACTatTATa3' recurred frequently. Variants of this sequence lie within the promoter regions discussed by Pribnow, Pirrotta, and others,^{19,22,25} with the Pribnow sequence often partially or wholly overlapping one or the other end of the sequence. Table 2 shows the occurrence and location of this sequence

TABLE I
HIGHER ORDER HOMOLOGIES

1st Sequence	2nd Sequence	# Matches	Size of Match
O _L 1	T7A2	1	9
		1	11
O _L 1	trp atten	2	9
		1	10
O _L 2		1	9
O _R 1		1	9
O _L 1	fdG3	1	10
O _L 2		1	9
O _R 1		1	9
O _R 1	trp P-O ₂	1	9
O _R 3		1	9
O _R 1	SV40P ₁	1	10
O _R 2	ara ₁	1	9
O _R 3	T7A3	1	9
lac p-O _Z	fdG3	2	9
lac O		1	9
fdG3	ara ₁ (also see above)	2	9
CAP	trp P-O ₁	3	9
	trp P-O ₂	1	9
	trp atten	1	9
	trp P-O ₁	4	9
	(also see above)	1	10
gal ₁	ara ₂	3	9
ara ₁	gal ₂	1	9

within the collection of Figure 9, also included are the lac promoter,¹⁶ and the hypothetical λ P_{RM} promoter;²⁵ sequences adjacent to the transcription and translation start points of the genes of ϕ X174 are shown in Table 3.²⁶ (The E gene is particularly interesting since the hypothetical promoter (565-582) is itself preceded by a sequence (523-542) in which 15 of 20 base pairs are identical to the trp attenuator (positions 8 to 27 in Figure 9), suggesting that the overlapping genes may be transcriptionally controlled with a D gene attenuator followed by a

Table 2. Recurrent Homology within Regulatory Sequences

SEQUENCE SOURCE	LINEAR SEQUENCE (5' → 3')	POSITION RELATIVE * TO TRANSCRIPTION START
λ O _R 1	TTTAcCTCtgGCGgtGATa	-27 → -9
λ O _L 1	TATcaCCGccAGTggTATt	-9 → -27 (I)
λ P _{RM} (?)	TTTatCCCTtGCGgtGATa	~ -27 → -9
λ P _O	TTTtgCTGtaTTTgtCATa	-28 → -10
fdG2P	TTCctCTTaaTCTttTTGa	-54 → -36
"	TTTgcTTCtgACTatAAa	-25 → -7
fdG3P	TGTatCATcgCCTgaTAAa	-41 → -59 (I)
"	TTTgaCCCaGCGatTATA	+9 → -10 (I)
SV40	GTTgtTGTtaACTtgTTTa	-39 → -21
"	TTTatTTGtaACCatTATA	+7 → -12 (I)
trp P-O	TTTtaCGTgaACTtgCGTa	+14 → -5 (I)
lac P	TATgcTTCcgGCTcgTATg	-27 → -9
ara BAD P	GTTtcCCCatACCcgTTTt	-8 → +11
"	TTTccCCAaTaCCcgTTTt	-7 → +12
"	TTCccCATacCCgtTTTt	-6 → +13
gal P	TCTttGTTatGCTatGGTt	-24 → -6
	TTAggCTTatGGTatGAAa	+16 → -3 (I)
T7A2 P	AGTaaCATgcAGTaaGATc	-25 → -7
T7A3 P	TGTcgTTTcaTGTggTACa	+12 → -7 (I)
CONSENSUS (19 sequences)	TTT--CTTtaACTgtTATA # 13 16 13 9 9 7 8 7 14 12 7 11 11 12 13 13	
% pyrimidine	84 74 95 42 58 95 74 79 63 37 32 79 32 58 68 26 84 26	

* The position numbering is given with reference to the 5' TTT to TAT 3' strand; (I) indicates that this sequence read in the 5' → 3' direction is running opposite the direction of transcription. The complement to the bracketed region of the consensus strand approximates the Pribnow sequence.²¹ The criterion for a match here is ≥ 7 of 12 bases identical with those of the consensus sequence.

separate promoter for the E gene). The sequence appears in both orientations with respect to the direction of transcription, often in tandem or in overlapping clusters. The latter arrangement would result in a repetition of the sequence on several faces of the helix where the cluster appears. These homologous regions may provide a common recognition site for a single protein. RNA polymerase, which is involved with all of these sites in some way, is a possible candidate.

Table 3. Recurrent Sequence Associated

with Start Points in Genes of ϕ X174.

SEQUENCE SOURCE	LINEAR SEQUENCE (5' $\rightarrow$ 3')	POSITION RELATIVE*	
		TO TRANSCRIPTION START	TO TRANSLATION START
ϕ X174 A	ATTggCTCtaATTtgTCTa	-71 $\rightarrow$ -53	-90 $\rightarrow$ -72
	TATttCCTagACAaatTAG	-46 $\rightarrow$ -64 (I)	-65 $\rightarrow$ -83 (I)
	ACaccCTCccAATtgTATg	-32 $\rightarrow$ -14	-51 $\rightarrow$ -33
	TTTatGGTtcGTTctTATt	+17 $\rightarrow$ +35	-3 $\rightarrow$ +16
B	TTTttGTTaaCGTatTTAg		-43 $\rightarrow$ -61 (I)
	CTTttTGTTaACGtaTTTa		-42 $\rightarrow$ -60 (I)
	GTTaaCAAAaAGTcaGATa		-53 $\rightarrow$ -35
	TTTtagCAGcaAGGtcCATa		-18 $\rightarrow$ -36 (I)
	CCTtgCTGctAAAggTCTa		-30 $\rightarrow$ -12
	TAGctCCTagACcctTAGc		-5 $\rightarrow$ -23 (I)
	CCAttCTTtaGCTccTAGa		+4 $\rightarrow$ -15 (I)
	CTTggTTTttAGTgaGTTg		+25 $\rightarrow$ +7 (I)
proposed C ²⁵	AATtaCTActGCTtgTTTa		-51 $\rightarrow$ -33
	CTTgtTTAcgAATtaAATc		-40 $\rightarrow$ -22
	CTTcgATTtaATTcgTAAa		-18 $\rightarrow$ -36 (I)
	GTCcaCTCcgATTtaATTc		-13 $\rightarrow$ -31 (I)
	TTTctCATttTCCgcCAGc		+8 $\rightarrow$ -11 (I)
D	ATTctCTTgtTGAcattTt	-47 $\rightarrow$ -31	-79 $\rightarrow$ -61
	AAGagCGTggATTactATc	-26 $\rightarrow$ -8	-58 $\rightarrow$ -40
	TTActATCtgAGTccGATg	-15 $\rightarrow$ +4	-47 $\rightarrow$ -29
	ATGctCTTcaACCactAAt	+2 $\rightarrow$ +20	-31 $\rightarrow$ -13
	CTGttCAaccACTaaTAGg	+5 $\rightarrow$ +23	-28 $\rightarrow$ -10
	CTTactATtAGTggTTGa	+27 $\rightarrow$ +9 (I)	-6 $\rightarrow$ -24 (I)
	TTGttCAGtaACTgACTc	+53 $\rightarrow$ +35 (I)	+21 $\rightarrow$ +3 (I)
E	TTTatGGTtacGCTggACTt		-3 $\rightarrow$ +16
J	GTTttCTTctGCGtcAGTa		-36 $\rightarrow$ -54 (I)
	TTTactCTTtaGACatTAc		+15 $\rightarrow$ -4 (I)
	TTTtaCCTttAGAcattAc		+16 $\rightarrow$ -3 (I)
F	TTTtaATTgcAGGggCTTc		-42 $\rightarrow$ -24
	TTTatCCTcaAGTaaGGGg		-3 $\rightarrow$ -21 (I)
	ATAaaTTAtgTCTaatATt		-7 $\rightarrow$ +12
	TTTgaATAttAGAcataAt		+16 $\rightarrow$ -3 (I)
	TATgtCTAatATTcaAAct		-1 $\rightarrow$ +18
G	ACAtgATTaaACTccTAAg		+4 $\rightarrow$ -15 (I)
	TTAggAGTttAATcaTGtt		-14 $\rightarrow$ +5
	CATgtTTCagACTttTATt		-1 $\rightarrow$ +18
	TCAGaCTTttATTtcTCGc		+6 $\rightarrow$ +24
H	TTTcgCTGaatCAGgTTAT		-61 $\rightarrow$ -43
	TTTgtCTCcaCCactTAA		-31 $\rightarrow$ -13
	ATTtaTGTTtGTTgcTATt		-4 $\rightarrow$ +15
	TATtgCTTctGCTctTGct		+24 $\rightarrow$ +42
CONSENSUS (41)	TTT-tCTTt-Act--TAT-		
	20 25 19 23 24 16 27 17 28 29 20 22		

* The position numbering is given with reference to the 5'TTT to TAT3' strand;
(I) indicates that this sequence read in the 5' $\rightarrow$ 3' direction is running opposite
the direction of transcription. The criterion for a match here is ≥ 7 of 12 bases
identical with those of the consensus sequence.

However, it should be noted that the DNA corresponding to the ribosome binding sites in three of the four sites recently determined in ϕ X174²⁷ reads:

site I	5'CCA--CTT--GCT--TAG3'
and in a different frame	5'TTT--CTC--AGA--TTT3'
site II	5'TTT--CTT--GAC--TAG3'
site III	5'ATA--TTA--TCT--TAT3'

This and the clustering of this sequence at translation as well as transcription start points leave it unclear whether the sequence is important for ribosome recognition, RNA polymerase recognition, or both.

A second sequence, similar to the first, was also found to yield matches in a large number of regulatory sequences: 5' TTT--AAG--T^G_CA--TTT3' These matches are shown in Table 4. This sequence is rather interesting because,

Table 4. Recurrent Matches (> 8/12) with the Sequence TTT--AAG--TG/CA--TTT.

SEQUENCE SOURCE	LINEAR SEQUENCE 3' → 3'	POSITION RELATIVE	
		TO TRANSCRIPTION START	TO TRANSLATION START
ara BAD P	TTTcaCGGcgTCacaCTT	-97 → -80	
	ATaaaAAGcgTCaggTAG	-23 → -40 (I)	
fd G2 P	AATtaAAAacgCGAtaTTT	-86 → -69	
	CTTcaAATatCGGctTTT	-64 → -82 (I)	
	GGTaaAGAccTGAttTTT	= 1 → +17	
fd G3 P	-CTaaAACacTCAttTTT	+24 → + 7 (I)	
gal P	ATTcaATGtcACActTTT	-46 → -29	-72 → -55
lac psO ₂	TGTtaCTGcTCacaTTT	NA	
λ P _O	TTGggCAGagGTAacTTT	-38 → -55 (I)	
tyr tRNA P	TTTctCAAcgTAAcaCTT	-51 → -34	
	TCTcaACGtaACActTTA	-49 → -32	
	CTGtaAAGtgTTAcgTTG	-29 → -46 (I)	
trp PO	TTAacTAGttCGAtgATT	-10 → -27 (I)	
MS2 coat protein	GUUugAAGcgUGGcuUCU		- 9 → + 9
Xenopus 5S Gene Spacer	TTTtcAAGgtTTGatTTT	-308 → -291	
	TTTcgAGGttTGAttTTT	-307 → -290	
	TTTaaAGTttTCAaaGTT	-291 → -274	
	TTTtcAAAggtTTGagTTT	-284 → -267	
	TTTcaAAGgtTCAGtTTT	-283 → -266	
	TTTcaAAGgtTCAGtTTT	-268 → -251	
	TTTcaAAGgtTAAttTTT	-253 → -236	
	TTTcaAAGgtTTCaaATT	-239 → -213	
	TTTtcAAAttTGAAttTTC	-222 → -205	
	TTTcaAAGgtTTCaaATT	-208 → -191	
	TTTtcAAAttTAAAttTTC	-199 → -182	
	TTTtcAAGgtTCAagTTG	-185 → -168	
	TTTcaACGgtTCAagTTG	-146 → -129	
	TTTcaACGgtTCAagTTT	-136 → -109	
	TTTcaACGgtTCAaaGTT	-106 → - 89	
	TTTcaAAGgtTCAAttTCA	- 71 → - 54	
Green Monkey (175 base pair reiterated seq.)	TTTctGAGaaACTgcTGT	NA	
	TCTcaCAGagTTAcaTGT		
	TCAcgGACgtTACAtTTT		
	TCTgaAACggAAAgcTGT		
	TGGctAAGgcTGTctTTG		
	ATTggCAAgCGAtaTTT		
	CTTccAAAtaTCCctTTC		
	TTTgaACGgaACAtaTTT		

QX 174 22 occurrences associated with the transcription or translation starts of nine genes, 1-5 occurrences per gene.

consensus (60 sequences)	TTT	AAG	T ^G _C A	TTT
	44 41 50	45 37 58	45 22 43	43 48 44

Figure 10. Reading Frame Analysis of Gal Regulatory Region. A reading frame of read-three-skip-two suggests that this regulatory region may be composed of overlaps of a single sequence.

in addition to numerous promoter matches, it bridges the two highly reiterated sequences now known, the spacer sequence for *Xenopus* 5S DNA (N. Federoff, personal communication) and the repeated sequence which comprises 10% of the total DNA in green monkey.²⁸ It remains to be seen if these sequences serve a regulatory function.

In addition to showing inter-sequence homologies, analysis with a reading frame can point out rather sophisticated internal repeat patterns within a sequence. Figure 10 shows an example of this, using the gal regulatory region.²⁰ Here a stretch of 70 base pairs from -57 to +13 shows no obvious repeating pattern in the linear sequence; however, in a read three, skip two frame, the same sequence shows 7 repeats approximating 5'ATG--^AAA--TAT--CAT--CA3' interwoven in both tandem and overlapping arrays on both strands. Patterns of this sort frequently occur in regulatory regions with the repeated sequence being one of the recurrent sequences mentioned above (see Tables 2-4).

Whether or not the homologies discussed evolved in common, then to be moved about, or were rather the result of convergence, perhaps dictated by physical constraints on regulatory regions, cannot yet be decided. Again, whether or not a single protein does in fact act in common on a sequence recurrent in different locations cannot be deduced from considerations of homology. But it is clear that any approach to these or related questions should take into account the

helical appearance of the sequence(s), with particular regard to the amplification of symmetry induced by the geometry of the helix. The correlative possibility of a regular reading frame, again determined by the helix, offers a level of complexity for both coding and multiplexing information that is implied by the apparent sophistication of promoter recognition on the one hand and by regulatory regions supporting multiple specific interactions (e.g. lambda rightward operator) on the other. Even if the accumulation of further biochemical and genetic data on protein-DNA interactions proves this reading frame incorrect, the homologies exist, as defined, spanning regulatory and suspected regulatory sequences from bacteria to primates.

References

1. Maniatis, T., Ptashne, M., Backman, K., Kleid, D., Flashman, S., Jeffrey, A., Maurer, R. 1975. *Cell* 5, 109-113.
2. Chamberlin, M.J. 1974. *Ann. Rev. Biochem.* 43, 721-775.
3. Wang, J.C., Barkley, M.D., Bourgeois, S. 1975. *Nature* 251, 247-249.
4. Maniatis, T. and Ptashne, M. 1973. *Proc. Natl. Acad. Sci. U.S.A.* 70, 1531-1535.
5. Pirrotta, V., Chadwick, P., and Ptashne, M. 1970. *Nature* 227, 41-44.
6. Chadwick, P., Pirrotta, V., Steinberg, R., Hopkins, N., and Ptashne, M. 1970. *Cold Spring Harbor Symp. Quant. Biol.* 35, 283-294.
7. Gilbert, W. and Maxam, A. 1973. *Proc. Natl. Acad. Sci. U.S.A.* 70, 3581-3584.
8. Steitz, T.A., Richmond, T.J., Wise, D., Engelman, D. 1974. *Proc. Natl. Acad. Sci. U.S.A.* 71, 593-597.
9. Gilbert, W., Gralla, J., Majors, J., Maxam, A. 1975. in *Protein Ligand Interactions*, H. Sund and G. Blauer, eds. (Berlin: Walter de Greyter and Co.) pp. 193-210.
10. Gilbert, W., Maxam, A., Mirzabekov, A. 1976. in *Control of Ribosome Synthesis*, N.C. Kjeldgaard and O. Maaloe, eds. (Munksgaard/New York: Academic Press) pp. 139-148.
11. Miller-Hill, B., Beyreuther, K., and Gilbert, W., *Methods in Enzymology*, 21, part D, 483 (ed. Grossman, L. and Moldave, K., 1971).
12. Hirsch, J. and Schleif, R. 1976. *J. Mol. Biol.* 108, 471-490.
13. Bahl, C.P., Wu, R., Stavinsky, J. and Narang, S.A. 1977. *Proc. Natl. Acad. Sci. U.S.A.* 74, 966-970.
14. Bertrand, K., Korn, L., Lee, F., Platt, T., Squires, C.S., Yanofsky, C. 1975. *Science* 189, 22-26.
15. O'Neill, M.C. 1976. *Nature* 260, 550-554.
16. Dickson, R.C., Abelson, J., Barnes, W.M., Reznikoff, W.S. 1975. *Science* 187, 27-35.
17. Sobell, H.M. 1976. *Ann. Rev. Biophys. Bioeng.* 5, 307-335.
18. Bennett, G.N., Schweingruber, M.E., Brown, K.D., Squires, C., Yanofsky, C. 1976. *Proc. Natl. Acad. Sci. U.S.A.* 73, 2351-2355.
19. Takanami, M., Sugimoto, K., Sugisaki, H., Okamoto, T. 1976. *Nature* 260, 297-302.
20. Musso, R., DiLauro, R., Rosenberg, M., deCrombrughe, B. 1977. *Proc. Natl. Acad. Sci. U.S.A.* 74, 106-110.
21. Dhar, R., Weissman, S.M., Zain, B.S., Pan, J., Lewis, A.M., Jr. 1974. *Nucl. Acid Res.* 1, 595-613.

22. Jay, E., Roychoudhury, R., Wu, R. 1976. *Biochem. Biophys. Res. Comm.* 69, 678-686.
23. Pribnow, D. 1975. *J. Mol. Biol.* 99, 419-443.
24. Fiers, W., Contreras, R., Duerinck, F., Haeghebaert, G., Merregaert, J., Min Jou, W., Raeymakers, A., Volchaert, G., Ysebaert, M., Van deKerckhove, J., Noll, F., Van Montagu, M. 1975. *Nature* 256, 273-278.
25. Pirrotta, V. 1976. *Current Topics in Microbiology and Immunology* 74, 21-50.
26. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchinson, C.A., III, Slocumbe, P.M., Smith, M. 1977. *Nature* 265, 687-695.
27. Ravetch, J.V., Model, P., Robertson, H.D. 1977. *Nature* 265, 698-702.
28. Rosenberg, M., Segal, S., Kuff, E.L., Singer, M.F. 1977. *Cell* 11, 845-858.
29. The author is supported by a grant from the Muscular Dystrophy Society of America. The computer time was provided by a grant from the University of Maryland Computer Science Center. The assistance of Wayne N. Bushell, who wrote the search program used in much of this analysis is gratefully acknowledged.