Volume 2 number 3 March 1975 Nucleic Acids Research

On the statistical significance of primary structural features found in DNA-protein interaction sites

Gene Dykes, Robert Bambara, Kenneth Marians and Ray Wu

Biochemistry, Molecular and Cell Biology, Wing Hall, Cornell University, Ithaca, New York 14853, USA

Received 16 December 1974

ABSTRACT

Probabilities of occurrence for a number of the symmetries and other sequence regularities found in DNA-protein interaction site sequences have been calculated for segments of random DNA sequence. Results show that many of the symmetrical and repetitive features seen in these interaction sites are likely to have occurred by chance. Other features are so unlikely to have occurred by chance that they are probably involved in the DNA-protein interaction processes.

INTRODUCTION

Jacob and Monod[1] predicted that regulation of transcription would occur at the molecular level, brought about by interaction between a regulator species and certain defined genetic loci on the chromosome. They postulated that negative control elements which inhibit transcription, termed repressors, would be small protein or RNA molecules. It was shown by the isolation of the repressor function by Gilbert and Müller-Hill[2] for the *E. coli lac* system and by Ptashne[3] for the λ phage system that repressors were, in fact, proteins. The operator was defined genetically as the locus of repressor action[1]. Both Gilbert and Müller-Hill[4] and Ptashne[5] demonstrated that the binding of their repressors to DNA was sensitive to O^c mutations, which rendered the operator region insensitive to repressor action, thus identifying the operator as the DNA binding site of the repressor. Since then, much attention has been focused on the nature of the repressor-operator interaction as a model for DNA-regulator protein interaction. Other attention has also been given to the sites of interaction of RNA polymerase with specific DNA segments known as promoters, and other sites of regulatory protein interaction.

It is clear that a prerequisite to the understanding of the nature of DNA-protein interaction is the determination of the primary structure of a number of interaction sites on defined segments of DNA. Gilbert and Maxam[6] and Gilbert, Gralla and Maxam[7] have obtained a 27 base pair sequence for the

region of DNA protected from pancreatic DNase digestion by the *lac* repressor protein. Since then, Maniatis *et al*[8] have determined the nucleotide sequence of the primary repressor binding site of the left-hand operator of phage λ. Very recently Dickson *et al*[9] have determined a sequence of 122 nucleotides from the *i* gene to the *z* gene of the *lac* operon. This sequence presumably contains the catabolic activator protein (CAP) binding site, promoter and operator. Their operator sequence agrees with that obtained in Gilbert's laboratory[6,7]. Schaller[10] has obtained the sequence of a promoter from the bacteriophage fd. Zain *et al.*[11] have determined the sequence of the region preceding the start of the E-strand transcript on SV40 DNA, while Sekiya and Khorana[12] have determined the sequence preceding the tyr t-RNA gene.

Inspection of these sequences reveals the following common features. There is the presence of some form of symmetry in all of the sequences. These symmetries have been discussed as being of possible signficance in DNA-protein interaction by the workers who reported the sequences. Various types of symmetry which are found at these sites are illustrated in figure 1. There is true 2-fold molecular symmetry in which atomic centers in all symmetrical nucleotides are symmetrical about a 2-fold axis (figure 1a). Such symmetry appears, upon examination of a single strand, as a complementary palindrome,

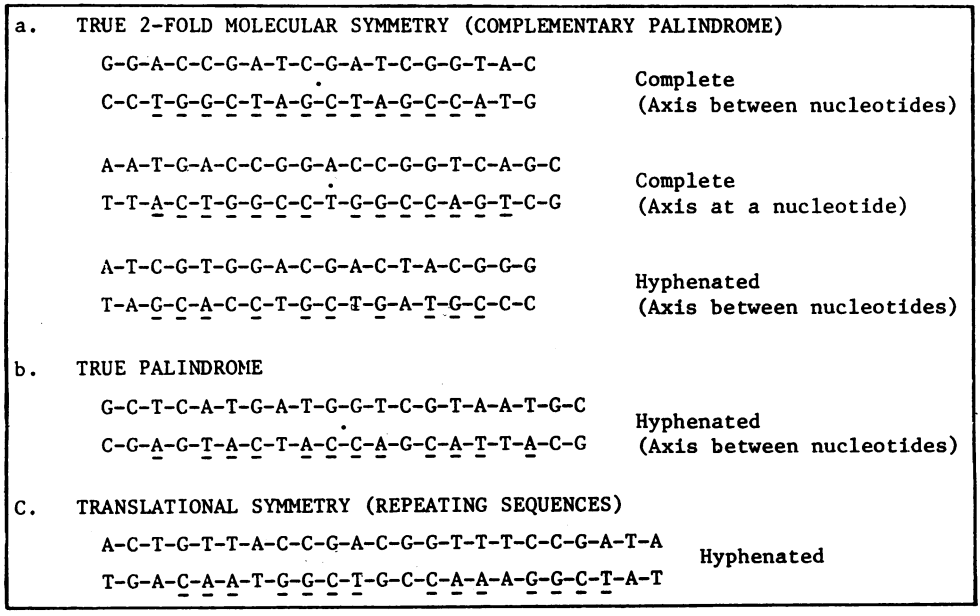

FIGURE 1--Examples of types of symmetry which can be found in nucleic acids. Underlined base pairs are participating in the symmetry. Other information about these types of symmetry is given in the text.

in which bases are symmetrical with complementary bases across a central axis (e.g. A-C-C-G-A-T-C·G-A-T-C-G-G-T). Sequences are also found where each strand contains a true palindrome, in which bases in a single strand are symmetrical with the same bases across a central axis (e.g. T-A-C·C-A-T) (figure 1b). Double-stranded true palindrome structures are not two-fold symmetrical with respect to atomic centers. Other sequences are found in which extensive regions of symmetrical purine-pyrimidine arrangements occur (e.g. Pu-Py-Pu·Pu-Py-Pu).

So far, all of the symmetries found in operator and promoter sequences are hyphenated, i.e., they have symmetrical nucleotides interspersed with non-symmetrical nucleotides, as shown in the example in figure 1a. An interesting feature of the complementary palindrome sequences is that they can be drawn so that the complementary sequences are matched, forming two double-stranded loops projecting from the original double-stranded linear DNA. True palindrome symmetries and sections of linearly repeated sequence (translational symmetry) also have been found (figure 1c). Besides symmetries, adjacent alternating G-C rich and A-T rich regions of varying lengths have been found in each of these sites. Many of these features are indicated in the DNA-protein interaction site sequences shown in figure 2.

Before determination of protein interaction site sequences, it was proposed that 2-fold symmetry would be found to play a role in protein-DNA interactions[13-17]. Bernardi[13] first proposed that true 2-fold molecular symmetry would be involved in the DNA recognition site of certain nucleases. Proteins which interact with DNA often exist as groups of identical subunits which have true molecular symmetry. Such proteins might be expected to bind regions of nucleic acid which display similar symmetry. Gierer[14] proposed that regulatory proteins would bind to and stabilize loop structures in DNA, so that repressors would act as a physical block to the passage of RNA polymerase. Sobell[15] expanded this hypothesis and proposed that the recognition sequence would be tandemly duplicated so that formation of a loop structure would allow a protein with 2 or 4 subunits to interact, enabling each identical subunit to bind an identical sequence. Crick[16] also proposed that regulatory proteins would bind to denatured regions of DNA.

Steitz *et al.*[18] have recently determined the approximate shape of the *lac* repressor, and based on this information have proposed that the *lac* repressor interacts with native helical DNA recognizing the true two-fold molecular symmetry (complementary palindrome) in the operator.

Since symmetries and other sequence regularities have been found in

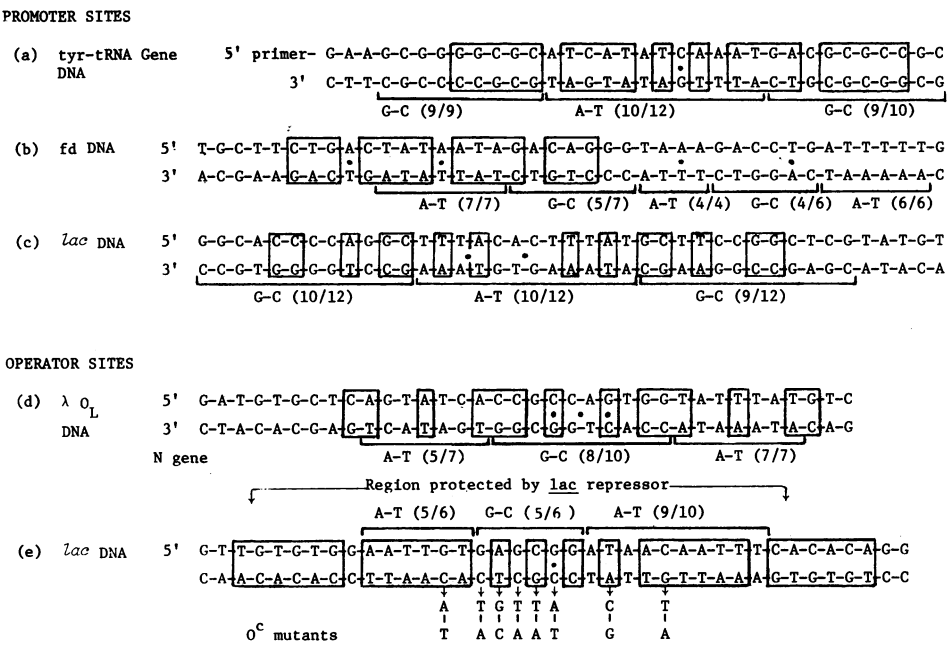

FIGURE 2--The nucleotide sequences of various DNA-Protein interaction sites. Regions of high G-C or A-T content are indicated. A single complementary palindrome sequence is indicated by boxed in areas in each sequence. Other complementary palindrome sequences are indicated only by the dots at their symmetry axes. True palindromes are not shown.

transcriptional control regions, the question of their exact role in the interaction process must be considered. Do some of these sequence regularities participate mostly as an initial recognition signal between the regulatory protein and DNA; do they assure the tight binding of the protein to DNA, do they represent subtle means of binding control, in fact, do they function in protein binding at all? It is reasonable to assume that in any segment of DNA, including segments known to interact with proteins, there are likely to be noticeable symmetrical orrepetitive sequences which have nothing to do with the biological function of the DNA, but instead are occurring merely by chance. It is possible to use statistical analysis to point out which of these features are likely to have been chance occurrences, to be assigned no particular importance unless later structural studies prove otherwise, and which are such rare occurrences that they are probably required for the recognition and binding processes.

RESULTS AND DISCUSSION

We have used the probability of occurrence of different symmetries and

other sequence regularities found in random DNA segments to predict the probability of occurrence of such features in DNA segments known or suspected to bind protein specifically. This information has allowed us to examine the statistical significance of the presence of sequence regularities in these DNA segments. Using this information we have been able to specify which sequence regularities are likely to be implicated in protein recognition.

In order to predict the occurrence of symmetry in random segments of DNA, we have used the following formula[19], which is a standard expression of the binomial distribution function,

$$P(N,B,X) = \frac{N!}{B!(N-B)!} X^{B}(1-X)^{(N-B)}$$

which yields the probability, P, of an event happening B times in N trials. In this formula P is the probability of occurrence of symmetry to the extent specified by the formula, N is the total number of nucleotides examined on one side of the symmetry axis, B is the number of these nucleotides which are symmetrical, (N-B) is the number of these nucleotides which are not symmetrical, X is the probability of a given nucleotide being symmetrical, and (1-X) is the probability that it is not symmetrical. In this analysis, the event for which a probability is calculated is the matching of bases in symmetrical positions with respect to axes drawn perpendicular to a polynucleotide sequence. The number of trials is determined by the number of bases on one side of the axis. Each base on one side of the axis is compared with its corresponding base on the other side. The term $X^{B}(1-X)^{N-B}$ may be considered as the probability that the nucleotides to the right of the symmetry axis are properly matched by a particular sequence of symmetric and non-symmetric elements to the left of the axis. The term $\frac{N!}{B!(N-B!)}$ gives the number of ways of arranging the symmetric and non-symmetric elements into distinguishable sequences. The probability of a match, X, is based on the assumption that for completely random sequences in a DNA the four bases are present in equal amounts. Thus, the quantitative results as presented here are only rigorougly true for DNA with A = T = G = C = 25%, but the conclusions will still be valid for DNA molecules not differing markedly from this percentage. This assumption, then, yields a value of 0.25 for X when attempting to match complementary bases and 0.5 when matching purines to purines and pyrimidines to pyrimidines. As a sample case, consider a sequence of 12 nucleotides in which 4 nucleotides on one side of a central axis match their complementary nucleotides in symmetrical positions on the other side of the axis. In this case, N = 6 and B = 4. Calculation yields the result, P = 0.033. In other words, there is a 3.3% chance that

out of 12 nucleotides in a random sequence, 8 will be involved in a particular type of symmetry around a central axis, producing, for example, a complementary palindrome. This formula works equally well for either true or complementary palindromes, or repeating sequences predicting an equivalent chance of random occurrence for either of these three types of symmetry.

Table 1 shows the probability of occurrence of different degrees of hyphenated symmetries, including the degrees of symmetries found in DNA segments

Probabilities of Occurrence of Symmetries at a Single Position in a Random DNA Segment

Symmetry Type	P (%)	Symmetry Type	P (%)	Symmetry Type	P (%)
2/3	14.5	8/8	0.00152	10/13	0.0126
2/4	26.2	5/9	4.89	7/14	3.83
4/4	0.391	6/9	0.999	10/14	0.0342
3/5	10.3	8/9	0.0101	12/14	3.21×10^{-4}
4/5	1.56	5/10	7.81	5/15	31.3
5/5	0.0976	[d]6/10	1.97	8/15	1.73
3/6	16.9	[f]8/10	0.0416	10/15	0.0794
4/6	3.76	9/10	0.00296	12/15	0.00124
5/6	0.464	6/11	3.43	14/15	4.28×10^{-6}
6/6	0.0244	8/11	0.119	5/16	37.0
4/7	7.01	9/11	0.0126	9/16	0.747
[b]5/7	1.29	6/12	5.44	[d]11/16	0.0285
7/7	0.00610	9/12	0.0392	15/16	1.14×10^{-6}
4/8	11.4	[a]10/12	0.00376	9/17	1.24
[f]5/8	2.73	11/12	2.21×10^{-4}	12/17	0.0100
6/8	0.423	5/13	20.6	[e]14/17	1.14×10^{-4}
[b]7/8	0.0381	[b,d]7/13	2.43	16/17	3.03×10^{-7}

a Present in sequence a, figure 2
b " " " b, " "
c " " " c, " "
d Present in sequence d, figure 2
e " " " e, " "
f " " " shown in Fig. 6

TABLE 1--Probabilities were calculated on a Data General Nova Minicomputer using the formula given in the text. The extent of symmetry is expressed as B/N where B represents the number out of N nucleotides on one side of a two-fold axis which are symmetrical with B out of N nucleotides on the other side of the axis. For example, in the sequence, AATTG·CAACT, containing a complementary palindrome for which the dot indicates the symmetry axis and the underlined nucleotides are symmetrical, B = 4, and N = 5. P is expressed as the probability of occurrence of (B or greater number of symmetrical nucleotides)/N. In this case P is the probability, in a 10 nucleotide long region, of matching 4 or 5 on the opposite side of a central symmetry axis.

which specifically bind protein. Degrees of symmetry are expressed in the form of B/N where B and N are as defined above and in the legend for table 1. It should be noted that probabilities given in table 1 are for a B or greater number of symmetrical nucleotides out of N. Thus the probability of 4/6 is given as (P of 4/6) + (P of 5/6) + (P of 6/6). This calculation requires three separate uses of the formula and summation of the results. In all results presented here, probability of occurrence of symmetry has been calculated in this way.

Probabilities given in table 1 are only for the chance occurrence of symmetry at one point in a random DNA segment. In order to determine the probability of occurrence of a particular degree of symmetry in a large DNA segment of defined length, the probability of occurrence of the symmetry must be considered around a large number of symmetry axes along the DNA segment. This is the only way to determine the probability of expected occurrence of degrees of symmetry found in sequences of DNA segments known to contain protein interaction sites, because these DNA segments are usually much larger than the symmetrical regions they contain. Another set of calculations was then performed, designed to yield the probability of finding a specific degree of symmetry in a given large segment of nucleotides. First the formula was used to determine the probability of occurrence of the degree of symmetry in question around a single axis. Next, we can again use the binomial formula, but this time to test the occurrence of a different event - the occurrence of a particular type of symmetry in a certain number of trials. To distinguish between the two applications of the binomial formula, N, X and B have been redefined as N', X', and B'. N', the number of trails, is the total number of possible axes which can be searched for symmetry. Inspection reveals that $N' = 2(n_1 - 2n_2) + 1$ where n_1 is the number of nucleotides in the DNA segment, and n_2 is the length of the symmetrical region on one side of the axis. $2n_2$ is subtracted from n_1 because there is a region of n_2 nucleotides on each end of the DNA segment in which a symmetry axis cannot be allowed to fall, since it would not be surrounded by the necessary number of nucleotides and still remain within the defined segment. The factor of 2 accounts for the fact that axes can fall between nucleotides or bisect nucleotides. One is added because there is always one more possible axis between nucleotides than the sum of axes bisecting nucleotides. X', the probability of occurrence in one trail, is the probability of occurrence of the degree of symmetry around each axis. For the example which has been used where 8 or more out of 12 nucleotides are symmetrical (4/6 in B/N notation), X' = 0.0376. For a 50 nucleotide long segment, N' would then be 2(50-12) + 1 or 77. B' equals the number of

occurrences of the symmetry, for which the probability is to be calculated. We would like to know the probability of one or more occurrences of the symmetry. If B' is set to zero then the probability of zero occurrences of the symmetry can be calculated. 100% - (the probability of zero occurrences) = (the probability of one or more occurrences). In our example, the probability of zero occurrences of 4/6 symmetry in a 50 nucleotide long segment is 5.2%, so the probability of one or more occurrences is 94.8%. Such calculations have been done for various symmetries found in various sized DNA segments. Table 2 shows a compilation of results from calculations of this type.

To test the correct application of the formula and the accuracy of the calculated probabilities, our computer was used to generate over 14,000 random sequences and search them for symmetrical sequences. Results of these calculations are given in table 3. In all cases, with numbers large enough for statistical accuracy, the observed occurrence of symmetries agreed very well with the occurrence predicted by two successive applications of the formula.

As shown in tables 1 and 2, the expected frequency of occurrence of many palindrome type sequences is quite high. These results are in general agreement with the report of Gralla and DeLisi[20] that a high degree of possible base-pairing is found in random sequences of single-stranded RNA.

One can look at the probability of occurrence for palindrome symmetry in two ways. It is tempting to assume that once the sequence of a known control region has been determined, any sequence regularity found within it has biological significance. By this assumption the discovery of 7/13 symmetry, which might allow formation of a looped structure, out of a protein protected segment 50 nucleotides long, implies that this symmetry plays a significant role in the interaction process. Our results cast doubt on such an interpretation. We feel that it is more valid to reason that because of the high frequency of occurrence of some palindromes (70% chance in this example), the finding of symmetry within control regions is not necessarily significant.

Clearly, if a symmetrical sequence at one position in a DNA segment were known to be involved in binding a specific protein with each symmetrical nucleotide contributing to the recognition process, the probability of random occurrence of the binding sequence would be very small. If the 7/13 symmetry is considered as 14 symmetrical nucleotides out of 26, each of the 14 nucleotides a specified base, it would have only a $1/4^{14}$ or 0.000000372% chance of random occurrence at any one position on the DNA. However, in order to see if the symmetry is likely to have occurred by chance, the probability of just this degree of symmetry 7/13, and not of any specific sequence should be con-

Occurrence of Various Degrees of Symmetry in Regions of DNA Containing Different Numbers of Nucleotides

Degree (B/N)	Prob. of one or more occurrences at any one position %	Probabilities of occurrence in DNA of sizes:						
		30	50	100	1000	50,000	1.0×10^6	3.0×10^6
4/6	3.76	75.8	94.8	>99.9	-	-	-	-
5/6	0.464	15.8	30.1	56.1	99.9	>99.9		
6/6	0.0244	0.9	1.9	4.2	38.3	>99.9	-	-
6/8	0.423	11.6	25.3	51.1	>99.9	-	-	-
7/8	0.0381	1.1	2.6	6.2	52.8	>99.9	-	-
8/8	0.00152	0.04	0.11	0.26	3.0	78.2	>99.9	-
6/10	1.97	34.2	70.3	96.0	>99.9	-	-	-
8/10	0.0416	0.87	2.51	6.48	55.8	>99.9	-	-
7/13	2.43	19.9	70.0	94.4	>99.9	-	-	-
9/13	0.099	0.89	4.73	13.7	85.5	>99.9	-	-
11/13	0.0011	0.01	0.05	0.16	2.13	66.9	>99.9	-
10/12	3.76×10^{-3}	0.0489	0.199	0.574	7.08	97.7	>99.9	-
14/17	1.14×10^{-4}		0.00377	0.0152	0.221	10.8	89.8	99.9
16/17	3.03×10^{-7}	-	0.00001	0.00004	0.000[illegible]	0.03	0.60	1.80

Table 2 - Values of P were calculated using the formula and procedure given in text. Values are probabilities of one or more occurrences in DNA of the given lengths. It may be mentioned that if we calculate the probabilities of only a single occurrence, the value will reach a maximum and then decrease as the size of the DNA increases, instead of approaching 100% asymptotically as does the probability of one or more occurrences.

sidered. An important sequence, whether it is part of the symmetry or not, cannot be picked out just by inspection. As calculated by the formula, a symmetry of 7/13, which may contain any of many possible nucleotide sequences has a 2.43% chance of occurrence in any one position in a random DNA segment 26 nucleotides long, and a 70% chance of occurrence in a random DNA segment 50 nucleotides in length (table 2). Therefore, this degree of symmetry is actually expected to occur. The appearance of such a symmetrical sequence in a segment of DNA known to bind protein should be assigned little significance unless more direct information is available to confirm its importance.

The formula given above is also applicable for calculating the probability of the occurrence of regions of high A-T or G-C content. In this case, X = (1 - X) = 0.5, for a completely random sequence. Calculations have been done for various sized regions of A-T and G-C content with varying degrees of G-C or A-T concentrations. Table 4 provides a compilation of data of this type. Again, probabilities are given for the specific base concentration or greater. Protein interaction site sequences shown in Figure 2 have been analyzed for the probability of occurrence of A-T and G-C rich regions as well as symmetries, and the results are given in table 5. Probabilities of the adjacent alternating G-C and A-T rich regions found in DNA-protein interaction sites are presented as double the product of the probabilities of each of the adjacent groups as given in table 4. The product is doubled because the central con-

Accuracy Test for Predicted Probabilities of Occurrence of Degrees of Symmetry Found in Random DNA Segments

(1) (B)	(2) P(%) of segments with (B or greater)/17 sym.	(3) Actual no. of segments with (B or greater)/17 sym.	(4) % of segments with (B or greater)/17 sym.	(5) % error
6	99.9	14807	100	0
7	99.9	14807	100	0
8	99.6	14731	99.5	0.10
9	80.9	12023	81.0	0.12
10	33.8	5083	34.3	1.5
11	8.00	1193	8.06	0.75
12	1.32	200	1.35	2.2
13	0.164	26	0.176	6.8
14	0.0152	4	0.0270	44
15	9.87×10^{-4}	0	0	--
16	4.03×10^{-5}	0	0	--

Table 3--14807 random segments of DNA were generated, each 100 nucleotides in length. Each segment was searched by the computer for the maximum value of B found among all possible complementary palindrome symmetries with N = 17. For each B in column 1, the actual number of segments found to contain this B or greater/17 as the maximum extent of symmetry are given in column 3. The percentage of the segments with the value of B given in column 1 or greater, are given in column 4. The predicted values of this percentage, obtained by two successive uses of tne binomial formula, as described in the text, are given in column 2. The percent error, comparing the predicted (column 2) and observed (column 4) values, is given in column 5.

Probabilities of Occurrence of A-T or G-C Rich Regions at a Single Position in a Random DNA Segment

Concentration	P (%)	Concentration	P (%)	Concentration	P (%)
[b]4/4	6.25	6/8	14.5	10/10	0.0976
5/5	3.13	7/8	3.52	8/11	11.3
[b]4/6	34.4	8/8	0.391	9/11	3.27
[e]5/6	10.9	7/9	8.98	10/11	0.586
[b]6/6	1.56	8/9	1.95	[c]9/12	7.30
[b,d]5/7	22.6	[a]9/9	0.195	[a,c]10/12	1.93
6/7	6.25	[d]8/10	5.47	11/12	0.317
[b,d]7/7	0.781	[a,e]9/10	1.07	[f]15/17	0.12

a Present in sequence a, figure 2
b " " " b, " "
c " " " c, " "
d Present in sequence d, figure 2
e " " " e, " "
f " " " " shown in figure 6

TABLE 4--P was calculated as in Table 1. In this case B/N represents the number of base pairs that are either A-T or G-C out of the total region considered, e.g.,

5/7 = -A-A-A-G-G-A-A-
-T-T-T-C-C-T-T-

Again P is given as the probability of (B or greater)/N identical base pairs in a group. It should be noted that probabilities given here are for occurrence of a base pair concentration rich in one defined base pair. If the region may be either A-T or G-C rich the probabilities given here should be doubled.

centration may be of A-T pairs or G-C pairs. The central concentration defines which base pair concentrations are in the adjacent groups. These data are presented as the probabilities of occurrence at a single position on a random segment of DNA. Probabilities are also given for one or more occurrences in random DNA segments the size of the given search area. A second application of the formula, as described for symmetries, is used to obtain probabilities in the search area. In these cases N', the total number of positions where the adjacent alternating concentrations can occur, is equal

Probabilities of Occurrence of Complementary Palindrome Symmetries And Alternating Nucleotide Concentrations in Known DNA-Protein Interaction Sites

Segment	Symmetry	P(%) (In S.A.)	Search Area	Alternating Nucleotide Concentration	P(%) (single)	P(%) (In S.A.)
tyr t-RNA	10/12	0.079	34	9/9,10/12,9/10	0.00008	0.00032
fd	7/8	1.927	41	7/7,5/7,4/4,6/6	0.00034	0.0062
	5/7	50.98	"			
	7/13	53.34	"			
lac promotor	5/6 repeating	23.99	41	10/12,10/12,9/12	0.0054	0.032
	7/14	65.13	41			
λ operator$_L$	7/13	40.33	36	5/7,8/10,7/7	0.0192	0.249
	6/10	48.19	"			
	8/12	6.73	"			
lac operator	14/17	1.14×10^{-4}	35	5/6,5/6,9/10	0.024	0.34
		10.18	50,000			
tet function recognition site[27,20]	5/8	2.7 (one position)	16	15/17	0.24	0.24

TABLE 5--Values of P for complementary palindrome symmetry were calculated as in the text for the given search area. Values of P for alternating sequences were obtained by doubling the product of the probability of occurrence of the individual concentrations. B/N is defined as in table 1 for symmetries, and as in table 4 for concentrations. Values in the column headed "Single" are the probabilities of occurrence at a single position on a random DNA segment. Values in the columns headed "In S.A." are the probabilities of occurrence in a random segment the size of the search area.

to $(m_1-m_2) + 1$, where m_1 = the number of nucleotides in the search region and m_2 = the number of nucleotides in the adjacent alternating concentrations. X' = the probability of occurrence at a single position, and B' = the number of occurrences for which the probability is to be calculated.

It is clear from the results shown in table 5 that many of the adjacent alternating regions of high G-C and A-T concentrations (three or more) found in DNA-protein interaction sites, when taken together, are as rare or rarer events than palindromes. This suggests that more attention should be given to their possible role in the protein binding mechanisms.

If each of the sequences presented in table 5 is examined individually, additional conclusions can be drawn:

The symmetry present in *lac* operator is striking in its extremely low probability of expected random occurrence. Here, with 28 of 34 nucleotides (14/17) involved in true molecular symmetry (complementary palindrome symmetry), the probability of occurrence of such an event at any one position on a DNA segment is 0.000114%. Symmetry this rare suggests that the symmetrical nucleotides play a crucial role in the DNA-protein interaction at the *lac* operator site. Possibly this extensive true two-fold symmetry matches true two-fold symmetry in protein contact points on the repressor molecule.

The mechanism of DNA-protein interaction and the exact role of the symmetry is still a matter of debate. One of the more popular proposed mechanisms for protein binding of sequences containing complementary palindromes has been the formation of a loop structure at the binding site[14]. A large loop structure would be a prominent feature on the DNA recognized because of its size and shape as well as sequence. The symmetry found in the *lac* operator is such a unique and unlikely occurrence, that one may propose that the degree of symmetry alone, producing a particular loop size, may be recognized by the *lac* repressor with only minor effects from the sequence. Our statistical data, shown in table 2, suggest that most of the degrees of symmetry observed in DNA-protein interaction sites are expected to occur in random segments much smaller than *E. coli* DNA, and so should occur hundreds or thousands of times on a random segment the size of E. coli DNA. A specific symmetrical site could only be recognized on the basis of its specific sequence. The *lac* symmetry should not occur in random DNA segments smaller than about 200,000-500,000 base pairs, and so should be present a few times on a random segment the size of *E. coli* DNA. In real *E. coli* DNA, most of these 14/17 symmetries could be lost by selection, or necessary constraints for loop formation such as hyphen positions, or G-C content could make the specific *lac*

symmetry unique. However, a model for recognition on the basis of symmetry alone is not in agreement with the results of Gilbert, Gralla and Maxam[7] who have shown that O^c mutations, which inhibit *lac* repressor binding, sometimes raise, sometimes lower and sometimes *do not affect* the extent of symmetry in the operator (see figure 2e). The loop structure model itself has not stood up well against recent evidence. The work of Wang *et al.*[21] studying changes in superhelicity of DNA containing the *lac* operator upon repressor binding, suggests that binding does not cause loop formation. Preliminary results from this laboratory on the action of S_1 nuclease on isolated restriction enzyme fragments containing the *lac* operator indicate that in the absence of repressor the operator does not exist, even part of the time, as a single-stranded loop structure[22]. In view of this information, the mechanism of Steitz *et al.*[18], in which the repressor is a molecule with two-fold symmetry interacting with the true two-fold symmetry in the native helical form of the operator seems more likely.

The symmetry present in tyrosine t-RNA promoter region is also highly unlikely to have occurred by chance (P at one position = 0.00376%). Since the promoter region sequence was not determined by isolation of the segment of DNA

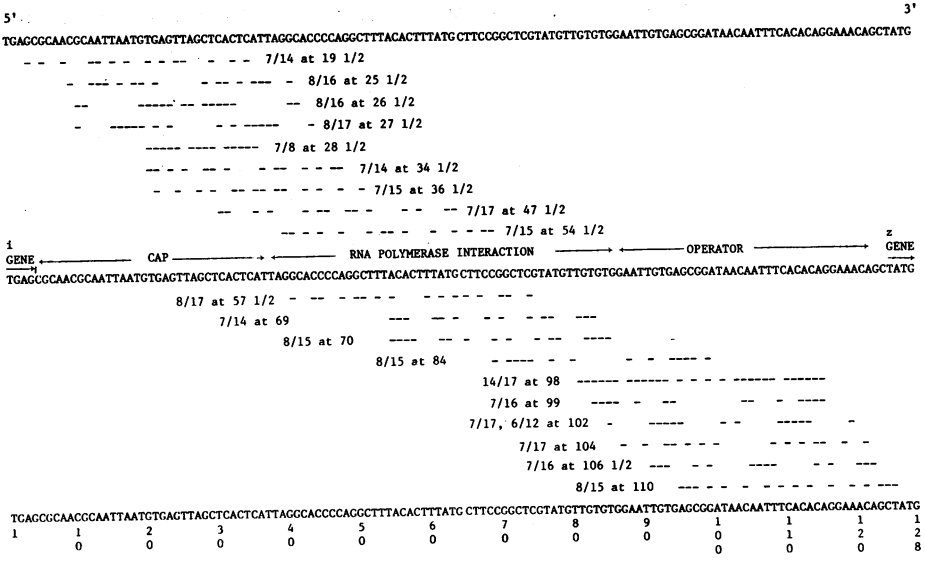

FIGURE 3 Complementary palindromes in the sequence of the *lac* control region as reported by Dickson, Abelson, Barnes and Reznikoff[9].

The sequence was searched by computer for symmetries of >6/17. The symmetrical regions are indicated by the positions of the dashes. The position of each symmetry axis is given as the number of nucleotides in the sequence to the left of the symmetry axis. A position number in the form M 1/2 refers to a symmetry axis lying between nucleotide M and M + 1.

which binds RNA polymerase, it is not clear exactly where the RNA polymerase-DNA contact site is. The presence of symmetry alone is not sufficient evidence to assume binding centered on the symmetrical region. It is necessary that there be more direct evidence available, such as isolation of a protein protected fragment, or discovery of sequence changes in the symmetrical region in promoter mutants. However, our calculations show that the degree of symmetry in the tyr t-RNA suspected promoter region is a very rare event. Such a degree of symmetry, as shown in table 2, is not expected to occur in random DNA segments smaller than several thousand nucleotides in length. Discovery of this degree of symmetry suggests that it could play a signficant role in the biological function of the DNA segment.

Although there are a number of symmetrical sequences found in the segment of SV40 DNA shown in Figure 5, the most outstanding symmetry is the 17-nucleotide long true palindrome[11]. Since the probability of occurrence of such an event (8/8) is very low (0.15% in a 65-long segment, and 15.4% in the 5500-long SV40 DNA), this symmetrical sequence may play a special role in the SV40 DNA-RNA polymerase interaction.

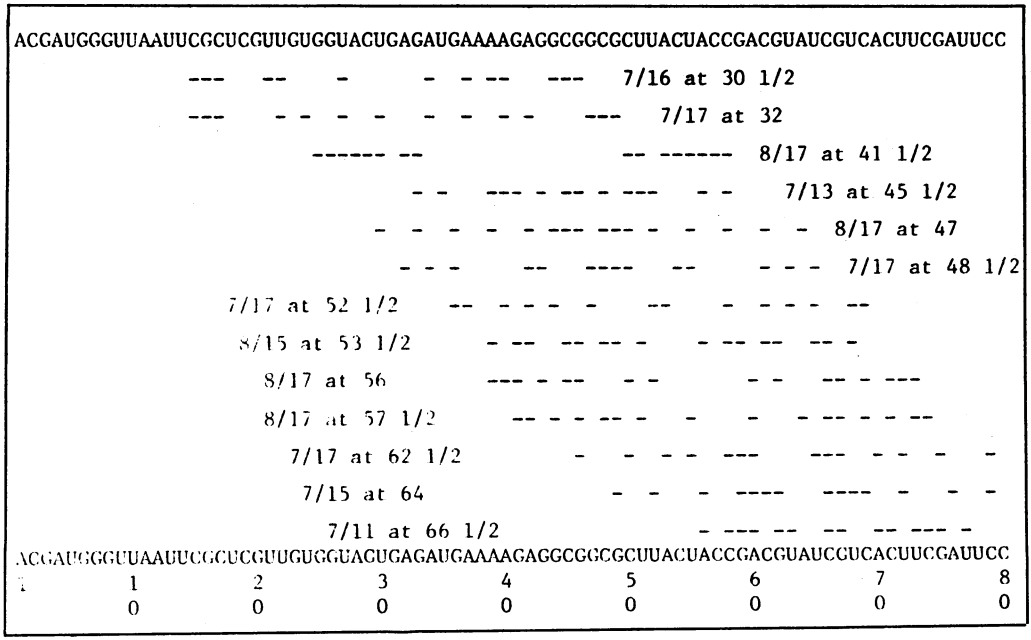

FIGURE 4--Complementary palindromes in the first 80 nucleotides of the sequence of the 6S RNA from *in vitro* transcription of λ DNA[23]. The sequence was searched by computer for symmetries of 6/17. The symmetrical regions are indicated by the positions of the dashes. The position of each symmetry axis is given as explained in the legend for figure 3. The positional arrangment of nucleotides 52-54 is ambiguous.

The alternating adjacent A-T rich and C-G rich nucleotide groups which pervade the *lac* and tyr t-RNA symmetrical regions also have a very low probability of chance occurrence. Such a low probability suggests that the arrangement of these nucleotides may play a role in the DNA-protein recognition process. The probable significance of such alternating adjacent A-T rich and G-C rich regions in the *lac* promoter area has been pointed out by Dickson *et al.*[9].

As shown in table 5, many of the other segments of DNA known to bind control proteins have degrees of symmetry which should be quite common on any DNA segment. In order to demonstrate that any segment of DNA will be covered with symmetries of this degree, we have located all complementary palindromes where B/N is (7 or greater)/15 in the 122 nucleotide long segment of the *lac* control region DNA[9]. As shown in figure 3, the DNA is covered with a maze of "background noise" symmetries, against which the *lac* operator symmetry stands out strikingly. True and complementary palindromes occur with about equivalent frequency. Similarly, smaller segments of DNA, such as the λ operator (left-hand) and the *lac* and fd promoter sites apparently also contain background symmetries. The 6/10 and 7/13 symmetries in the λ operator, the 5/7 and 7/13 symmetries in the fd promoter region, and the 7/14 and 5/6 symmetries in *lac* promoter region, all with very high probabilities of chance occurrence, are most likely to be such background symmetries. The 7/8 symmetry with P = 1.927% in the fd promoter region does seem considerably more likely to have a biological function. The 8/12 symmetry in the λ operator, with P = 6.73% is difficult to assign.

A curious fact is evident upon examination of the background symmetries of the *lac* control region. All complementary palindrome symmetries of (7 or greater)/17 found in the CAP interaction and RNA polymerase binding site regions are centered between nucleotides, whereas all complementary symmetries but one in the operator region are centered on a nucleotide. The effect can be contrasted to the essentially random distribution of symmetry axes of the background symmetries found in the first 80 nucleotides of an RNA molecule transcribed *in vitro* from λ phage DNA by Lebowitz, Weissman and Radding[23], shown in figure 4. The possible significance of this non-random grouping of symmetry axes in the *lac* control region remains to be elucidated.

Another interesting aspect of the sequence in figure 3 is that the symmetries seem to be clustered in particular areas. Inspection reveals a cluster of symmetries centered in the regions assigned as the CAP, RNA polymerase and repressor binding sites. Perhaps clusters of symmetries may be involved in defining functional regions of the chromosome.

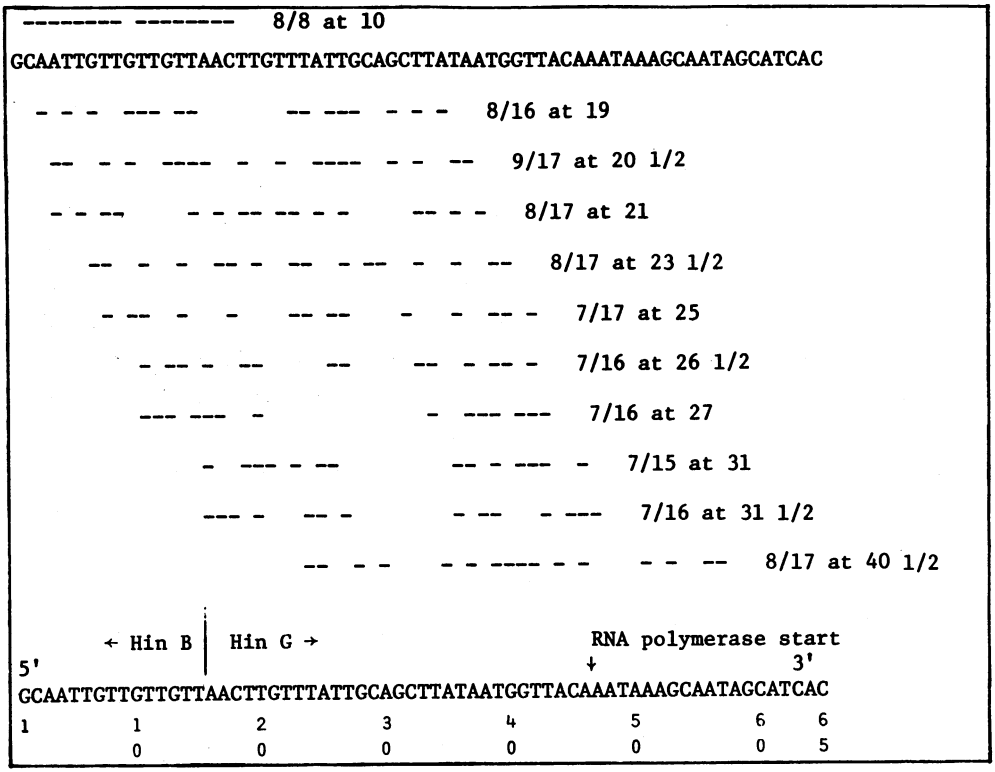

Figure 5 True palindromes in the sequence of the L-strand of SV40 DNA which precedes the preferred initiation site of E. coli RNA polymerase. Sequence of nucleotides 16-57 is derived from the L-strand transcript by using HinG DNA fragment as template, and nucleotides 1-15 is derived by using Ad2+ND3A DNA as template.[11] Sequence of nucleotides 47-65 is derived from the E-strand transcript by using intact SV40 DNA as template, and a sequence of 160 nucleotides beyond nucleotide 65 has been reported.[32] The RNA polymerase binding site, which represents the preferred initiation site for the E-strand transcript, is presumably within nucleotides 1-46. This sequence was also searched for complementary palindromes. The most significant complementary palindrome in this region is a 6/9 match at position 43. P = 1.0% at any one position.

Both the λ operator, and the fd and *lac* promoters have unusually low probability for the presence by chance of the adjacent alternating A-T and G-C rich regions. It would seem reasonable to predict that the alternating A-T and G-C rich regions rather than the symmetrical sequences present will play a role in the protein DNA interaction. The alternating A-T and G-C rich regions may assume unusual helical structures which could form the basis of recognition by specific protein molecules. The length, base composition and the pattern of these alternating regions could provide the specificity for

the interaction. This interpretation is consistent with the analyses of the helical structure of poly(dA-dT).poly(dA-dT) and poly(dG-dC).poly(dG.dC) by Wells *et al.*[24] and Arnott *et al.*[25].

It is interesting to note that regions of alternating high concentrations of G-C and A-T can, by their very non-random nature, give rise to symmetry more easily than a completely random sequence. This fact is illustrated in figure 3 where a considerable number of the symmetries which have been found are centered over regions of G-C and A-T base concentrations. The existence of a concentration of A-T and G-C pairs on a DNA segment significantly raises the probability of occurrence of symmetry at the center of the concentration. This situation is especially evident in the regions of circular λ and Ø80 DNA which are cleaved by the *ter* function endonuclease to produce the cohesive ends The sequences of the cohesive end and 3'-terminal regions of λ DNA[26,27] and Ø80 DNA[28-30] have been determined and are shown in figure 6. Weigel *et al.*[27] have proposed that the recognition site of the *ter* endonuclease is a complementary palindrome centered in the middle of the cohesive ends directly between the cleavage sites (all shown in figure 6). For this one position in a random sequence, the probability of occurrence of the 5/8 symmetry which was observed is 2.7%. However, since this region is very G-C rich, having 15/17 G-C base pairs, the expected probability is considerably higher. For a random region of 17/17 G-C pairs the probability of occurrence of 5/8 symmetry at one position is 21.9%. Therefore, the symmetry is likely to have occurred by chance, while the run of 15/17 G-C pairs (P = 0.24%) is improbably enough to have an important role in either cohesive end recognition or endonuclease function.

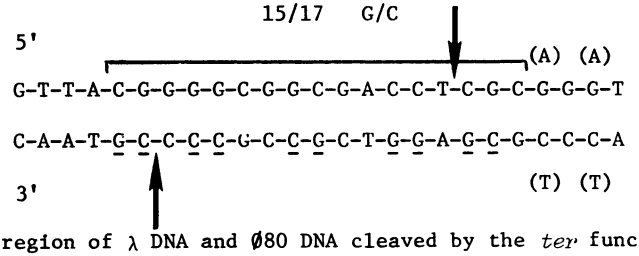

FIGURE 6--The region of λ DNA and Ø80 DNA cleaved by the *ter* function endonuclease to generate the cohesive ends.

The arrows point to the sites of cleavage. The 5/8 complementary palindrome symmetry is outlined by underlining the nucleotides involved. In Ø80 DNA the two A-T pairs in parenthesis replace the G-C pairs present in the λ sequence.

It is not possible to present any theory on the mode of interaction between regulator proteins and DNA from the results presented here. However,

we have been able to show that many of the symmetrical sequences present in DNA-protein interaction sites are expected to occur by chance, and, therefore, without further evidence should not be assumed to be involved in the interaction process. Some symmetrical sequences found at DNA-protein interaction sites are highly unlikely to have occurred there by chance and so are more likely to be present to perform some specific role in the interaction process. As it has been pointed out by Wu, Bambara and Jay[31], adjacent alternating G-C and A-T rich regions are present in all of the sequences shown in figure 2, and are very unlikely to have occurred by chance in each case. These nucleotide concentrations should also be expected to be involved in the biological f tion of the regions. Additional conclusions concerning these sequences await the detailed structural studies of the DNA-protein complexes at each of these sites.

REFERENCES

1 Jacob, F. and Monod, J. (1961) *J. Mol. Biol. 3,* 318.
2 Gilbert, W. and Müller-Hill, B. (1966) *Proc. Nat. Acad. Sci. USA 56,* 1891.
3 Ptashne, M. (1967) *Proc. Nat. Acad. Sci. USA 57,* 306.
4 Gilbert, W. and Müller-Hill, B. (1967) *Proc. Nat. Acad. Sci. USA 58,* 2415.
5 Ptashne, M. (1967) *Nature 214,* 232.
6 Gilbert, W. and Maxam, A. (1973) *Proc. Nat. Acad. Sci. USA 70,* 3581.
7 Gilbert, W., Gralla, J. and Maxam, A. *Personal communication.*
8 Maniatis, T., Ptashne, M., Barrell, B. and Donelson, J. (1974) *Nature 250,* 394.
9 Dickson, R., Abelson, J., Barnes, W. and Reznikoff, B. *Personal communication.*
10 Schaller, H. *Personal communication.*
11 Zain, B.S., Weissman, S.M., Dhar, R. and Pan, J. (1974) *Nucleic Acid Research 1,* 577.
12 Sekiya, T., and Khorana, H. G. (1974) *Proc. Nat. Acad. Sci. USA 71,* 2978.
13 Bernardi, G. (1968) *Advan. Enzymol. 31,* 1.
14 Gierer, A. (1966) *Nature 212,* 1480.
15 Sobell, H. (1972) *Proc. Nat. Acad. Sci. USA 69,* 2483.
16 Crick, F. (1971) *Nature 234,* 25.
17 Adler, K., Beyreuther, K., Fanning, E., Geisler, N., Gronenborn, B., Klemm, A., Müller-Hill, B., Pfahl, M. and Schmitz, A. (1972) *Nature 237,* 322.
18 Steitz, T. A., Richmond, T. J., Wise, E. and Engelman, D. (1974) *Proc. Nat. Acad. Sci. USA 71,* 593.
19 Darlington, R. B. (1974) *Radicals and Squares,* Logan Hill Press, 477.
20 Gralla, J. and DeLisi, C. (1974) *Nature 248,* 330.
21 Wang, J. C., Barkley, M. D. and Bourgeois, S. (1974) *Nature 249,* 247.
22 Marians, K. J., Bahl, C. P. and Wu, R. *Unpublished results.*
23 Lebowitz, P., Weissman, M. and Radding C. M. (1971) *J. Biol. Chem. 246,* 5120.
24 Wells, R.D., Larson, J.E., Grant, R.C., Shortle, B.E. and Cantor, C.R. (1970) *J. Mol. Biol. 54,* 465.
25 Arnott, S., Chandrasekaran, R., Hukins, D., Smith, P. J., and Watts, L. (1974) *J. Mol. Biol. 88,* 523.
26 Wu, R. and Taylor, E. (1971) *J. Mol. Biol. 57,* 491.
27 Weigel, P. H., Englund, P. T., Murray, K. and Old, R. W. (1973) *Proc. Nat. Acad. Sci. USA 70,* 1151.
28 Bambara, R., Padmanabhan, R. and Wu, R. (1973) *J. Mol. Biol. 75,* 741.
29 Murray, K. and Murray, N. E. (1973) *Nature New Biol. (Lond.) 243,* 134.

30 Ghangas, G. S., Jay E., Bambara, R. and Wu, R. (1973) *Biochem. Biophys. Res. Commun. 54,* 998.
31 Wu, R., Bambara, R. and Jay, E. (1974) *CRC Critical Reviews in Biochemistry, in press.*

The authors thank Dr. Keith Moffat for many helpful comments and suggestions, Kay Koffel for sharing her knowledge of statistical theory, and Larry Brenner for doing some of the computations. They thank Drs. Abelson, Reznikoff, Gilbert, Schaller and Khorana for making their results available prior to publication. They also thank Drs. Elizabeth Keller, Jeffrey Roberts and Joseph Calvo for critical reading of the manuscript. This is paper XXII in a series of NUCLEOTIDE SEQUENCE ANALYSIS OF DNA. Paper XXI is by Bambara, R., and Wu, R., J. Biol. Chem.1975. This work was supported by Research Grants GM-18887 and CA-14939 from the National Institutes of Health, and NIH Training Grant GM-00824.