E.coli promoter spacer regions contain nonrandom sequences which correlate to spacer length

Bruce A.Beutel[1] and M.Thomas Record Jr.[1,2]*
[1]Program in Molecular Biology and [2]Departments of Chemistry and Biochemistry, University of Wisconsin, Madison, WI 53706, USA

Received February 12, 1990; Revised and Accepted May 10, 1990

ABSTRACT

The −10 and −35 regions of *E. coli* promoter sequences are separated by a spacer region which has a consensus length of 17 base-pairs. This region is thought to contribute to promoter function by correctly positioning the two conserved regions. We have performed a statistical evaluation of 224 spacer sequences and found that spacers which deviate from the 17 base-pair consensus length have nonrandom sequences in their upstream ends. Spacer regions which are shorter than 17 base-pairs in length have a significantly higher than expected frequency of purine-purine and pyrimidine-pyrimidine homo-dinucleotides at the six upstream positions. Spacer regions which are longer than 17 base-pairs in length have a significantly higher than expected frequency of purine-pyrimidine and pyrimidine-purine hetero-dinucleotides at these positions. This suggests that the nature of the purine-pyrimidine sequence at the upstream end of spacer regions affect promoter function in a manner which is related to the spacer length. We examine the spacer sequences as a function of spacer length and discuss some possible explanations for the observed relationship between sequence and length.

INTRODUCTION

Our present understanding of promoter sites for *E. coli* Eσ70 RNA polymerase (RNAP) and of polymerase-promoter interactions is based on a large number of genetic, biochemical, and biophysical experiments, as well as on statistical analyses of promoter sequences. Promoters for RNAP are characterized by two regions of conserved sequences (hexamers), designated the −35 and −10 regions to indicate their respective distances from the transcriptional start site. Although no naturally occurring promoter has been found to have the consensus sequence, most promoters have only a few deviations from consensus.[1,2] The vast majority of point mutations which have been found to alter rates of transcription initiation *in vivo* are located within these two hexameric regions, and *in vitro* studies have upheld the general conclusion that the extent to which a promoter is homologous to the consensus sequence in these two regions is the major determinant of the rate of transcription initiation. These findings have recently been reviewed.[3,4,5]

The spacer region which separates the −35 and −10 regions of a promoter is usually 17 base-pairs in length. A recent compilation of 231 naturally occurring promoter sequences contains promoters with spacer lengths which range from 15 to 21 base-pairs, with nearly all being between 16 and 18.[2] Recent experiments have shown that the substitution of large central portions of the spacer region with homopolymeric and heteropolymeric sequences suspected to have altered helical conformations gives rise to measurable effects on promoter function *in vivo* and *in vitro*.[6,7] Certain groups of promoters (such as ribosomal RNA promoters) have conserved sequences in the spacer region, but most spacer regions appear to consist of relatively random sequences.[1,2] Recent work on a set of promoter sequences has shown that the average spacer sequence is different for promoters of differing spacer length, but previous studies of point mutations indicate that the spacer sequence of naturally occurring promoters is not a significant determinant of promoter function.[8,9,10] However, the length of the spacer region is an important determinant of promoter function. Constructs in which the spacer length is either increased or decreased from the consensus 17 base-pair length exhibit a reduced promoter strength *in vitro* and *in vivo*.[11–14] In addition, studies of promoters with single-stranded gaps in spacer regions of various lengths are consistent with the hypothesis that the structural and physical properties (e.g. length, torsional rigidity, etc.) of the spacer region, rather than its exact sequence, are important in promoter function.[15]

We have performed a statistical evaluation of the sequences of 224 nonhomologous spacer regions and found that the order of purines and pyrimidines in these sequences correlates to the length of the spacer. Spacers longer than 17 base-pairs tend to have nonrandom sequences biased to favor alternating purine-pyrimidine (RY) and pyrimidine-purine (YR) base steps; spacers shorter than 17 base-pairs tend to have nonrandom sequences biased to favor purine-purine (RR) and pyrimidine-pyrimidine (YY) base steps. Sequences of 17 base-pair spacers are random. In spacers which have lengths differing from 17 base-pairs, the nonrandom sequences are located near the upstream end of the spacer (adjacent to the −35 region), rather than throughout the

* To whom correspondence should be addressed at the Department of Chemistry, 1101 University Avenue, University of Wisconsin, Madison, WI 53706, USA

entire spacer sequence. We propose that these nonrandom sequences serve to compensate (in part) for the effects of the nonconsensus spacer length on promoter strength and promoter function. Some possible mechanisms for this compensation are described.

MATERIALS AND METHODS

The Harley and Reynolds compilation of 231 naturally occurring promoters was used as our data set. The single promoter which has a spacer length of 20 was deleted, as were 6 promoters (colE1-P2, M1rna, rrnG-P1, rrnG-P2, Tn2660bla-P3, and tyrT/6) which are completely homologous to others in the set. (We thank a referee for pointing these out, though including them did not significantly alter the results.) The remaining 224 promoters were used in our analysis. The sequences of the spacer regions were analyzed using a dinucleotide frequencies test and a run test. All calculations were performed on a Digital VAX 8650 with programs written in Fortran by one of the authors (B. Beutel).

The dinucleotide frequencies test used to analyze the spacer sequences was performed as follows. The number of RR, YY, RY, and YR dinucleotide base-steps and the number of R and Y were counted for each group of spacer sequences. The expected number of RR, YY, RY, and YR were determined from the product of the frequency of the occurrence of R and Y and the total number of dinucleotides observed. For example, the expected number of RR equals $[R/(R+Y)]^2(RR+YY+RY+YR)$. The observed and expected numbers of homo-dinucleotides (RR+YY) and hetero-dinucleotides (RY+YR) were compared using a two-class chi-squared test (discussed below). Those groups with greater than 95% significance were considered to have a significantly nonrandom arrangement of purines and pyrimidines.

The run test used to analyze the spacer sequences was performed as follows. In a sequence of purines (R) and pyrimidines (Y), for example RYRRYYYR, a run is defined as a stretch of either R or Y *of any length* which is flanked by either the other character or an end. In this example, there are 5 runs: R, Y, RR, YYY, and R. Given the number of R and Y in a sequence (4 R and 4 Y in this case) the number of runs expected if the arrangement were random is given by a noncontinuous distribution described and tabulated by Owen.[16] Note that if the sequence contains at least one R and at least one Y, then there must be at least two runs. The maximum possible number of runs for a sequence of length n is n, in the case of a completely alternating sequence with an equal number of R and Y, such as RYRYRYRY. The spacer sequences were grouped by spacer length, and the sequences within each group were categorized further by their observed number of runs. This categorization results in a distribution of the observed number of spacer sequences which have each possible number of runs. This discrete distribution was compared to that expected for a group of randomly arranged sequences each with the same composition of R and Y as the actual group. As an example, the sequences of the 6 upstream positions of the four spacers of length 15bp are the following:

1) TTT AAA, 2) A CTT AA, 3) CCT GA C, 4) TC GAGA

Each sequence is shown with spaces separating the runs of purines and pyrimidines. For the particular composition of R and Y of each of these four sequences, the probabilities of having each possible number of runs (from Owen) are compared with the observed number of runs in Table 1. Note that in this case, if the sequences were random, at least half of them (1.6 + 0.8 + 0.2 = 2.6 out of 4) would have 4 or more runs. However, all four sequences have 3 or less runs. Of course, in such a small dataset this deviation is not very significant. A chi-squared test was used to calculate the significance of the differences between the observed and expected distributions. If the chi-squared test resulted in a greater than 95% probability that the observed and expected distributions were not the same, then the group of spacers was considered to have a statistically significant number of spacer sequences which have either fewer or greater runs than expected.

Note that the dinucleotide and run tests address slightly different questions. The dinucleotide test measures the extent to which purines and pyrimidines are randomly arranged within each sequence *and* across the sequences of the group. This is because the hypothesis tested in the dinucleotide test is that the purines and pyrimidines in the *group* of promoters are randomly arranged. The run test measures only the extent to which the purines and pyrimidines are randomly arranged within each sequence, but not across the sequences in the group. Unlike the dinucleotide test, the run test takes into account the individual composition of each sequence in the group.

Chi-squared tests were performed as described by Snedecor and Cochran for single classification with two or more classes.[17] Briefly, if there are n classes, such as n=4 for classes of dinucleotides (RR, YY, RY, and YR) then chi-squared = $[\Sigma\ (f_i - F_i)^2/F_i$, with n−1 degrees of freedom, where f_i = observed number in class i and F_i = expected number in class i.

TABLE 1. Sample Calculation Using Run Test (15bp Spacers)

Sequence[1]	Composition	Runs Obs.	Theoretical Probability of n Runs (Owen)				
			n=2	3	4	5	6
YYYRRR	3R, 3Y	2	0.10	0.20	0.40	0.20	0.10
RYYYRR	3R, 3Y	3	0.10	0.20	0.40	0.20	0.10
YYYRRY	2R, 4Y	3	0.13	0.27	0.40	0.20	0
YYRRRR	4R, 2Y	2	0.13	0.27	0.40	0.20	0
Expected distribution of runs:			0.46	0.94	1.60	0.80	0.20
Observed distribution of runs:			2	2	0	0	0

[1] Purine/pyrimidine sequence for 6 positions at upstream end of 15bp spacers. See Methods for details.

RESULTS

The spacer sequences of 224 promoters were grouped by spacer length. Each group was analyzed for the occurrence of homo- and hetero-dinucleotides (see methods). The results of this analysis are shown in figure 1. The ratio of observed to expected homo-dinucleotides appears to decrease with increasing spacer length. Spacers of length 17 (the most common length) have a ratio of 1.0, which would be expected for completely random sequences.

In order to investigate the source of this trend, we repeated the calculations for three different segments of the spacer regions (figure 2): the 6 positions at the upstream end (counting from the −35 region), 6 central positions downstream to the first six (7−12), and the 6 positions at the downstream end (adjacent to the −10 region). Note that these segments overlap for spacers shorter than 18 base-pairs, and that some positions are untested for spacers longer than 18 base-pairs. The origin of the trend in figure 1 is clearly confined to the six positions at the upstream end of the spacer region. The upstream segments in the 16 and 18 base-pair spacer groups are significantly nonrandom (> 95% confidence level). The groups of spacers were also analyzed to examine whether the trend which is observed in terms of purines and pyrimidines also exists in terms of frequencies of particular dinucleotides, such as AA, AC, TA, etc. Of the sixteen possible dinucleotides, AA and GC occur at higher than expected frequencies in the 17 base-pair long spacer sequences. *E. coli* DNA sequences have previously been reported to have a significant preference for AA and GC dinucleotides.[18] The sequences of those spacers which are longer or shorter than 17 base-pairs do not show a significant preference for any dinucleotide. Therefore, the trend observed in terms of purine and pyrimidine homo- and hetero-dinucleotides is not apparent in terms of individual dinucleotide sequences (data not shown).

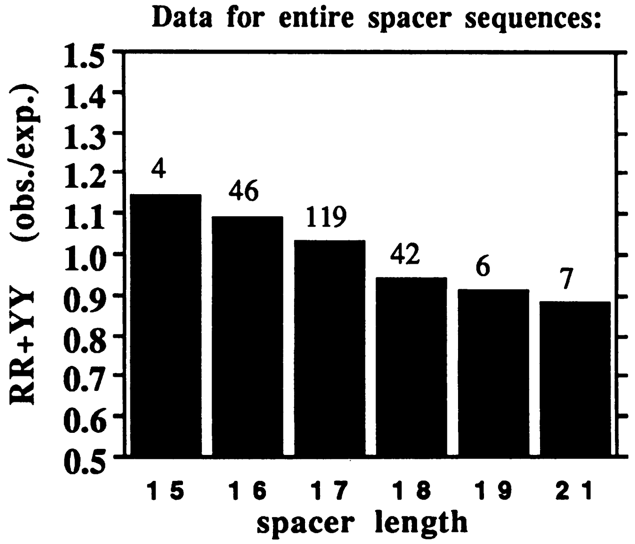

Figure 1. Dinucleotide distribution as a function of spacer length. The ratio of observed/expected homo-dinucleotides is shown for the spacer sequences of each length. The number of promoter sequences with each spacer length is shown above each bar. The expected number of purine-purine (RR) and pyrimidine-pyrimidine (YY) dinucleotides in each group of spacer sequences was calculated from the frequency of purines and pyrimidines as described in the Methods section. The distributions of (RR+YY) vs. (RY+YR) in the 16 base-pair spacer sequences are significantly nonrandom as determined by the chi-squared test (see Methods).

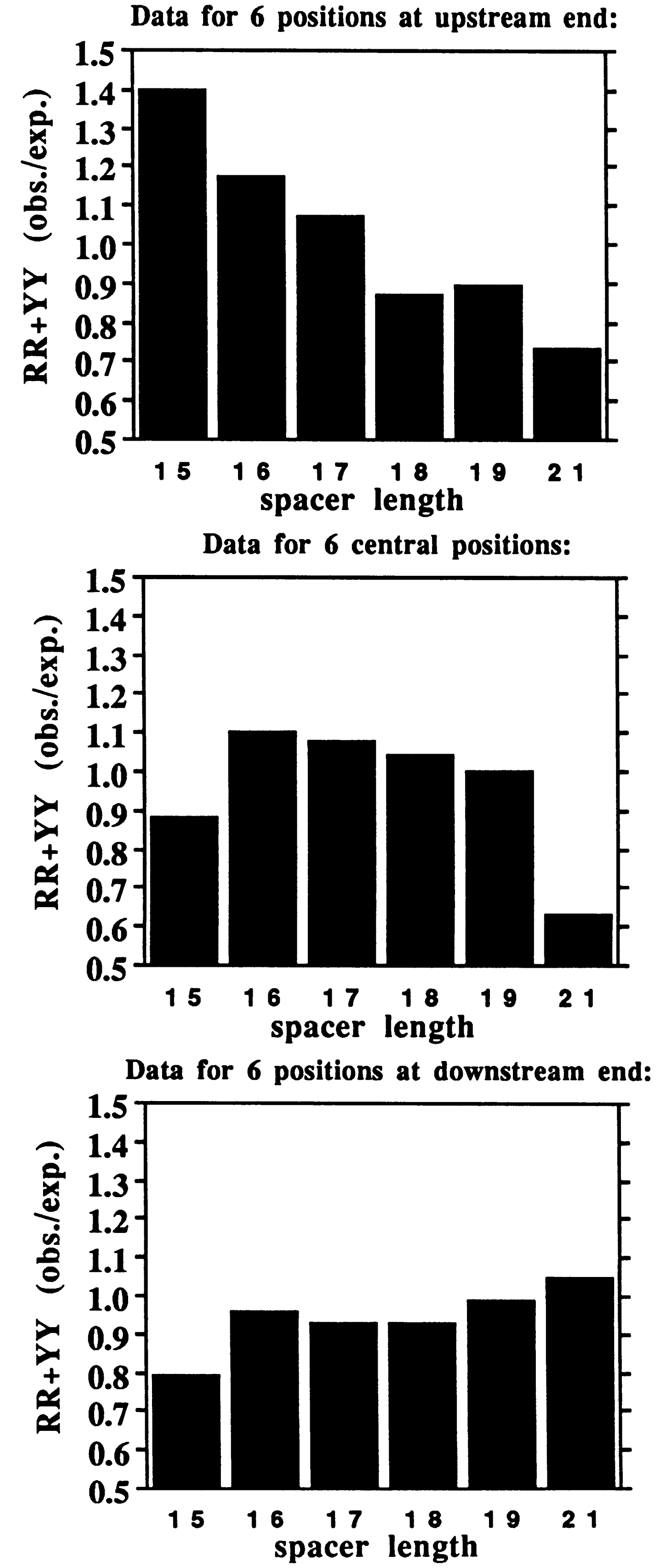

Figure 2. Dinucleotide distribution for three segments of spacer region. Three different segments of the spacer region—the 6 positions (1−6) at the upstream end, 6 central positions (7−12), and the 6 positions at the downstream end—were analyzed as described in figure 1. The distributions of (RR+YY) vs. (RY+YR) in the upstream end of the 16 and 18 base-pair spacer sequences are significantly nonrandom as determined by the chi-squared test (see Methods).

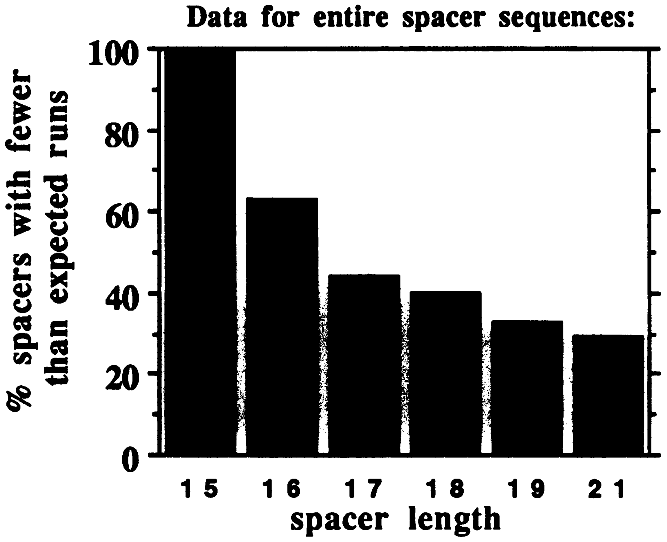

Figure 3. Percentage of spacer sequences which are more homopolymeric than expected. The number of runs were counted for each individual spacer sequence in each group of sequences and compared with the calculated average number of runs expected for each group (see Methods). The percentage of spacers which have fewer runs than expected is shown for each spacer length. A sequence which has fewer runs than expected has more homo-dinucleotides than expected. The distribution of spacer sequences (in terms of number of runs) for the 16 base-pair spacer sequences is significantly nonrandom as determined by a two class chi-squared test (see Methods).

The results of the dinucleotide tests indicate that spacer sequences, when grouped according to their lengths, have a nonrandom frequency of occurrence of dinucleotides in terms of purines and pyrimidines. The dinucleotide results alone do not give information about *individual* spacer sequences, since a subset of sequences in each group, if very skewed, could make the entire group appear to be nonrandom. The run test was used to examine the composition of each group of spacer sequences so that the generality of the dinucleotide trend could be assessed (see methods). The results for whole spacer sequences, grouped according to length, are shown in figure 3. The fraction of individual spacer sequences within each group which have fewer runs than expected appears to decrease with increasing spacer length. An examination of three segments of the spacer region (figure 4) shows the same trend is apparent only in the upstream end of the spacer region. These results are consistent with the dinucleotide frequency results in figures 1 and 2. This indicates that as length increases, individual spacer sequences have a tendency to contain fewer homo-dinucleotides (i.e., more runs). For example, the araBAD promoter has a 16 base-pair long spacer with 6 pyrimidines in a row adjacent to the −35 region, while the cloDFrnaI promoter has an 18 base-pair long spacer with 7 bases of alternating purines and pyrimidines adjacent to the −35 region. These are extreme examples, but they are representative of the tendencies we observe in the majority of promoter sequences. A summary of our results for the entire spacer region and for the six positions at the upstream end are shown in Table 2.

We attempted to determine the exact location in the spacer regions which exhibits the trends we observe. Starting with the first dinucleotide downstream of the −35 region, each group of spacer sequences was scanned, and the percentage of sequences in the group which have a homo-dinucleotide was calculated for each position. The results are shown in figure 5. Although there

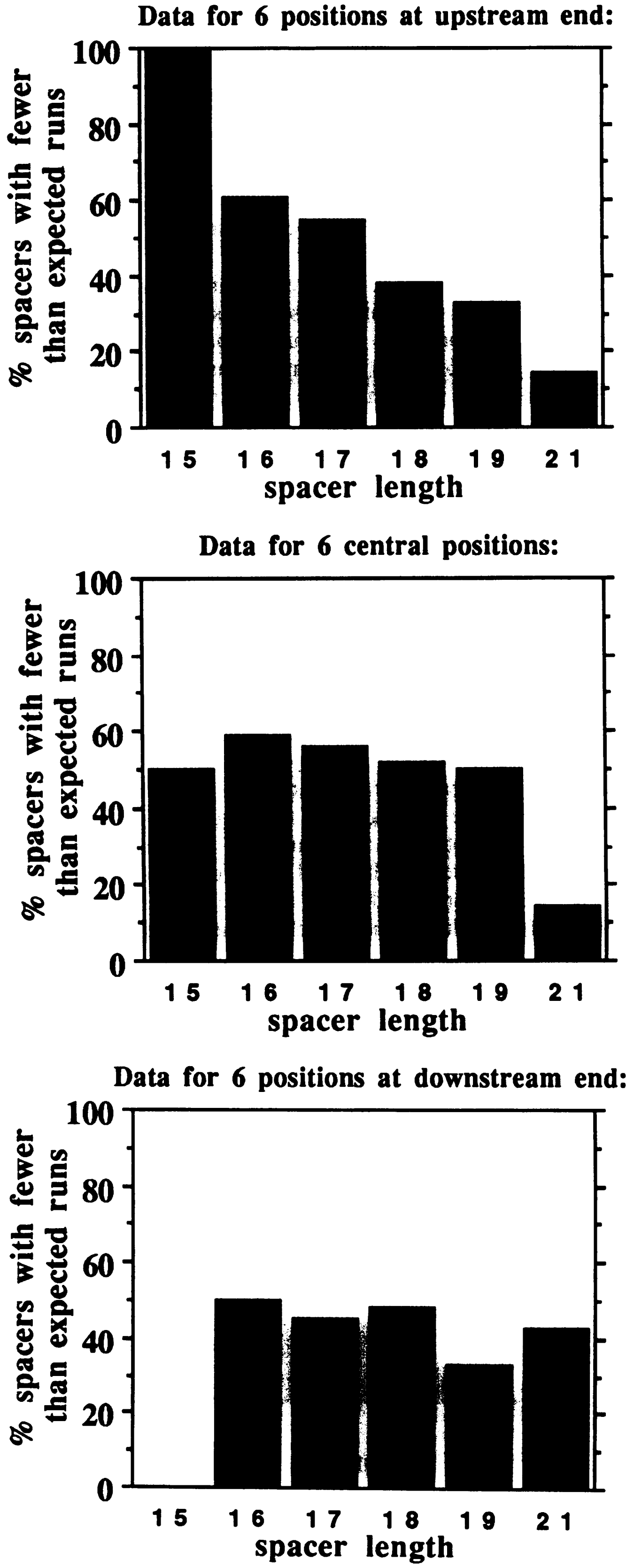

Figure 4. Percentage of homopolymeric sequences for three segments of spacer region. Three segments of the spacer region, as described in figure 2 and text, were analyzed as described in figure 3. The distributions of spacer sequences (in terms of number of runs) at the upstream end for the 16 and 18 base-pair spacer sequences are significantly nonrandom as determined by a two class chi-squared test (see Methods).

TABLE 2. Summary of Statistical Analysis for Entire Spacer Region and Six Upstream Positions

a. Entire Spacer Region				Dinucleotide Test		Run Test	
Length	Set Size	%R	%Y	O/E[1]	CL[2]	%<E[3]	CL[2]
15	4	51.7	48.3	1.14	–	100	–
16	46	50.0	50.0	1.09	97%	63	94%
17	119	48.3	51.7	1.03	–	44	–
18	42	49.5	50.5	0.94	–	40	–
19	6	49.1	50.9	0.91	–	33	–
21	7	57.8	42.2	0.88	–	29	–

b. Six Positions at Upstream End				Dinucleotide Test		Run Test	
Length	Set Size	%R	%Y	O/E[1]	CL[2]	%<E[3]	CL[2]
15	4	50.0	50.0	1.40	–	100	–
16	46	47.8	52.2	1.17	99%	61	95%
17	119	47.9	52.1	1.07	–	55	–
18	42	48.8	51.2	0.87	94.5%	38	96%
19	6	38.9	61.1	0.89	–	33	–
21	7	57.1	42.9	0.73	–	14	–

[1] Ratio of observed/expected homodinucleotides.

[2] Confidence level that result is nonrandom.

[3] The percentage of sequences in group which have fewer than the expected number of runs.

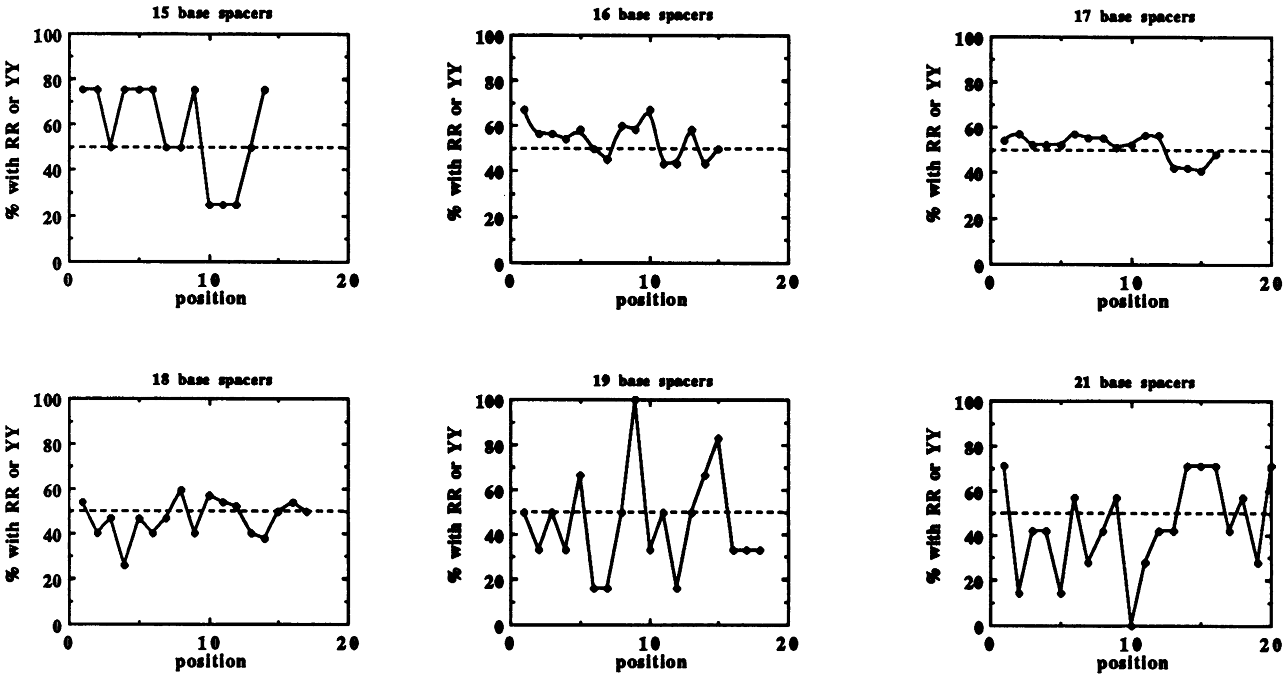

Figure 5. Dinucleotide frequencies at each position of the spacer region. The spacer sequences of each group were scanned from upstream to downstream. The dinucleotides at each position were counted, and are shown here as the percentage of sequences which have homo-dinucleotide (RR or YY) at each position. See text for discussion of results.

does not appear to be a distinct boundary, the results are consistent with the previous observation that the nonrandom sequences are located in the six positions at the upstream end of the spacer region. In this segment of the spacer region, short spacers (lengths 15 and 16) have a higher percentage (> 50%) of homo-dinucleotides, and the long spacers (lengths > 17) have a lower percentage (< 50%) of homo-dinucleotides. It is clear that no subset of the upstream segment is responsible for the trends we observe.

DISCUSSION

DNA sequences of the spacer region separating the −35 and −10 regions in *E. coli* promoters have generally been considered to be random. We observe a trend in which spacer length correlates to an over- or under-representation of homo-dinucleotides, in terms of purines and pyrimidines. Those spacer sequences which are 17 base-pairs long (consensus) appear to be random at this level, and nonrandom only in their preference for the dinucleotides AA and GC. *E. coli* sequences (unrelated to promoter sequences) have been shown to have a significant preference for these two dinucleotides. Therefore, this preference is not likely to be related to spacer (or promoter) function. This observation is consistent with the prevalent model of spacer function in which the spacer region serves only to separate the −35 and −10 regions in a manner which is independent of DNA sequence.

The 17 base-pair long spacer sequences have approximately the same representation of homo-dinucleotides, in terms of purines and pyrimidines, as that expected for random sequences. Those spacers shorter than 17 base-pairs in length exhibit a higher than expected frequency of occurrence of homo-dinucleotides. Those spacers longer than 17 base-pairs in length have a higher than expected frequency of occurrence of hetero-dinucleotides. These statistically significant preferences result from the tendency for individual spacer sequences in each group to have the preference observed for the group. These preferences are entirely analogous to having a consensus sequence for a group of sequences. The spacer sequences which are shorter than 17 in length have *conservation* of the tendency to have homo-dinucleotides—i.e., a significantly large fraction of these sequences have more homo-dinucleotides than expected. The spacer sequences greater than 17 in length have just the opposite tendency, and those of length 17 show neither tendency. Although very few promoters in our data set have spacer lengths other than 16, 17, or 18, it appears that these tendencies vary monotonically with spacer length over the range from 15 to 21 base-pairs (figures 3,4).

The trends we observe for whole spacer sequences appear to be a result of trends in the upstream end of the sequences. The downstream regions of the spacers do not exhibit the trends we describe above. This localization may be important in eventually understanding the reasons for these trends. Spacer lengths less than or greater than 17 result in a decrease in promoter strength.[11−14] It is possible that *E. coli* spacer sequences have evolved to modulate or compensate for this loss of strength. Spacer sequences would be used to 'fine tune' promoter strength. Of course, not every promoter would have the same kind of spacer sequence, but it is likely that if spacer sequences did play such a role, the regions involved would contain nonrandom sequences. (This is analogous to the observation of consensus sequences for the −35 and −10 regions, even though no natural promoters have this exact sequence.) A regulatory role for the spacer region is consistent with the observations of other investigators that substitutions of sequences in the center of spacer regions result in measurable effects on promoter function.[6,7] Our results suggest that these effects might have been even larger had the substitutions been in the upstream end of the spacer region.

We propose a model to account for the nonrandom nature of the spacer sequences. This model leads to direct predictions of the effect of sequence mutations in the upstream end of spacer regions. It is possible that the sequence in the spacer region provides a sensitive means of adjusting the orientation of the −35 and −10 regions with respect to each other. Crystallographic and NMR studies demonstrate that the purine-pyrimidine sequence of DNA is an important determinant of the local DNA structure.[19−21] Homo-dinucleotides have been shown to have greater helical twist than hetero-dinucleotides.[19,20] Consider the case of a promoter with a 17 base-pair long spacer which has the −10 region optimally aligned with respect to the −35 region. Changing the spacer length to 16 would alter this alignment by approximately 34 degrees. Variations in helical twist per base-step on the order of 5−10 degrees have been observed.[19] Therefore, an increase in the number of homo-dinucleotides in the spacer region could at least partially correct for this 34 degree displacement. Only those regions of the spacer which are inert (i.e. not contacted by RNA polymerase) would be able to serve such a function. DNA modification and protection studies have shown that the upstream end of the spacer exhibits very few close contacts with RNA polymerase during open complex formation *in vitro*.[10,22] These same studies demonstrate that the downstream end of the spacer shows a large number of such contacts. If the downstream end of the spacer region binds, even nonspecifically, to the RNA polymerase, then it is not inert. Since the upstream end of the spacer region is only minimally involved in contacts with RNA polymerase, it may be used to orient the −35 region with respect to the downstream end of the spacer and the −10 region. This is consistent with our observation that only the upstream end of the spacer sequences is nonrandom and exhibits the trends we report here.

It is also possible that sequence effects on DNA flexibility, unwinding, and other physical properties thought to be involved in promoter function could be the underlying reason for the nonrandom upstream ends of spacer sequences. Experiments will be required to further understand the function of the spacer sequence as a part of overall promoter function. We suggest that such experiments focus on the upstream end of the spacer region, because the sequences at this end exhibit the kind of nonrandomness which is usually associated with functional DNA sites.

ACKNOWLEDGEMENTS

This work was supported by NIH grant GM23467. We would like to thank Pieter deHaseth, Bill Reznikoff, Rick Gourse, and Jeff Livingstone for their helpful discussions and comments on the manuscript, and a special thanks to the referees for their thoughtful comments and suggestions.

REFERENCES

1. Hawley, D. K. and McClure, W. R. (1983) *Nucleic Acids Res.* **11**, 2237−2255.
2. Harley, C. B. and Reynolds, R. P. (1987) *Nucleic Acids Rec.* **15**, 2343−2361.
3. von Hippel, P. H., Bear, D. G., Morgan, W. D. and McSwiggen, J. A. (1984) *Annu. Rev. Biochem.* **53**, 389−446.
4. McClure, W. R. (1985) *Annu. Rev. Biochem.* **54**, 171−204.
5. Travers, A. A. (1987) *Crit. Rev. Biochem.* **22**, 181−219.
6. Auble, D. T., Allen, T. L. and deHaseth, P. L. (1986) *J. Biol. Chem.* **261**, 11202−11206.
7. Lozinski, T., Markiewicz, W. T., Wyrzykiewicz, T. K. and Wierzchowski, K. C. (1989) *Nucleic Acids Res.* **17**, 3855−3863.
8. O'Neill, M. C. (1989) *J. Biol. Chem.* **264**, 5522−5530.
9. Youderian, P., Bouvier, S. and Susskind, M. M. (1982) *Cell* **30**, 843−853.
10. Siebenlist, U., Simpson, R. B. and Gilbert, W. (1980) *Cell* **20**, 270−281.
11. Mandecki, W. and Reznikoff, W. S. (1982) *Nucleic Acids Res.* **10**, 903−912.
12. Brosius, J. Erfle, M. and Storella, J. (1985) *J. Biol. Chem.* **260**, 3539−3541.
13. Mulligan, M. E., Brosius, J. and McClure, W. R. (1985) *J. Biol. Chem.* **260**, 3529−3538.
14. Stefano, J. E. and Gralla, J. D. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 1069−1072.
15. Ayers, D. G., Auble, D. T. and deHaseth, P. L. (1989) *J. Mol. Biol.* **207**, 749−756.

16. Owen, D. B. (1962) *Handbook of Statistical Tables*, Addison-Wesley, NY, New York.
17. Snedecor, G. W. and Cochran, W. G. (1980) *Statistical Methods*, The Iowa State University Press, Ames, Iowa.
18. Nussinov, R. (1984) *Nucleic Acids Res.* **12**, 1749–1763.
19. Dickerson, R. E. and Drew, H. R. (1981) *J. Mol. Biol.* **149**, 761–786.
20. Dickerson, R. E. (1983) *J. Mol. Biol.* **166**, 419–441.
21. Schroeder, S. A., Roongta, V., Fu, J. M., Jones, C. R. and Gorenstein, D. G. (1989) *Biochemistry* **28**, 8292–8303.
22. Duval-Valentin, G. and Ehrlich, R. (1986) *Nucleic Acids Res.* **14**, 1967–1983.