Nucleic Acids Research, 1997, Vol. 25, No. 23 **4703–4709**

Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by *Escherichia coli* RNA polymerase

O. N. Ozoline*, A. A. Deev[1] and M. V. Arkhipova[2]

Institute of Cell Biophysics, Russian Academy of Sciences (RAS) and [1]Institute of Theoretical and Experimental Biophysics, RAS, Pushchino, 142292 Moscow region, Russia and [2]Department of Biology, Moscow State University, Leninskie gory, 119899 Moscow, Russia

Received June 16, 1997; Revised and Accepted October 14, 1997

ABSTRACT

Nucleotide sequences of 441 promoters recognized by *Escherichia coli* RNA polymerase were subjected to a site-specific cluster analysis based on the hierarchical method of classification. Five regions permitting promoter subgrouping were identified. They are located at –54 ± 4, –44 ± 3, –35 ± 3 (–35 element), –29 ± 2 and –11 ± 4 (–10 element). Promoters were independently subgrouped on the basis of their sequence homology in each of these regions and typical sequence elements were determined. The putative functional significance of the revealed elements is discussed on the basis of available biochemical data. Those promoters that have a high degree of homology with the revealed sequence elements were selected as representatives of corresponding promoter groups and the presence of other sequence motifs in their structure was examined. Both positive and negative correlations in the presence of particular sequence motifs were observed; however, the degree of these interdependencies was not high in all cases, probably indicating that different combinations of the signal elements may create a promoter. The list of promoter sequences with the presence of different sequence elements is available on request by Email: ozoline@venus.iteb.serpukhov.su.

INTRODUCTION

A statistical analysis of promoter sequences recognized by *Escherichia coli* RNA polymerase (1–3) has made it possible to identify two consensus hexamers, TTGACA and TATAAT, located ~35 and 10 bp upstream from the transcription start point, respectively. Each base pair of these elements is most often found at a specific position, nevertheless an average conservation in both regions is 7.9 nt per promoter (3) and ~10% of the efficient promoters match the consensus in only five or six positions (3). This variability causes problems in generating an efficient promoter-site-searching algorithm by means of statistical weighting of canonical base pairs (4–7), and several self-learning pattern-recognition softwares employing neural networks were suggested for this purpose (8–12). In spite of some limitations based on the bias selection of the promoters for training and on uncertainty in cut-off rules when a sequence is passed through a series of networks which give contradictory results and polling has to be done, these approaches demonstrate the highest efficiency in promoter recognition (up to 98%). Preliminary classification of promoters according to their spacer length (6,7,10,11) or functional specificity (13) increases the predictability of promoter-search algorithms and those which take into account sequences flanking consensus hexamers gave better results (6,7,9,12). Non-random base pair distribution, specific for particular promoter groups, has been observed in non-canonical promoter regions (13–15), suggesting that additional signal elements may contribute to promoter specificity. Recent biochemical studies support this possibility: TG dinucleotide located in some promoters 1 bp upstream from the –10 element has an established functional significance (16,17) and a group of promoters has been identified for which transcription efficiency depends upon the interaction of the C-terminal domain of the RNA polymerase α-subunit with the sequences located one to two helix turns upstream from the –35 element (16–19). Both topology of the complex formed by RNA polymerase with different promoters (reviewed in 20) and molecular mechanism of complex formation (21) display a pronounced variability. It is clear that this variability could not be conditioned by the elements common for all promoters, and approaches revealing non-canonical elements may be favorable for further progress.

An analytical method for finding unknown patterns that occur imperfectly in a set of promoter sequences has been developed in this study. Specific elements were searched without any prior assumption as to their nature and quantity within one and the same promoter region. Two main questions were addressed: (i) what kind of sequence motifs besides canonical hexamers may be found in the promoter structure and (ii) whether the presence of a particular signal element in the promoter group correlates with the presence of any other sequence motif in the same group?

A set of 441 promoter sequences recognized by *E.coli* RNA polymerase and ranging from –70 to +10 was compiled. Promoters were aligned according to their +1 position(s) and

*To whom correspondence should be addressed. Tel: +7 095 9237467; Fax: +7 0967 790509; Email: ozoline@venus.iteb.serpukhov.su

subjected to classification by software CLUSTER based on the hierarchical method of clustering (22). Promoters were compared by their subsequences of varying length (2–10 bp) at all positions from –70 to –1. Upstream and downstream shifting up to 2 bp was allowed for better fitting and the highest match number was used as a measure of promoter resemblance. At every position promoters were subgrouped on the basis of their similarity and the distribution of group sizes along the promoter was analyzed. The strongest clustering was observed when the 5′ end of compared subsequences was located at –54 ± 4, –44 ± 3, –35 ± 3, –29 ± 2 and –11 ± 4. The typical base pair motifs were determined in any of these regions as a consensus sequence for the largest groups. The presence of these sequences in the promoter compilation was examined and those motifs which display a maximum in distribution in the region of clustering were suggested as a putative non-canonical signal elements. Those promoters which have high degree of similarity with revealed sequences were selected as representatives of the corresponding promoter group and the presence of other sequence elements in the promoter structure of the group was examined. Both positive and negative correlations in the presence of a particular sequence elements were observed; the degree of these correlations was not high in all cases, however, indicating that different sets of signal elements can create an efficient promoter. Functional significance of the revealed sequence motifs is discussed on the basis of footprinting data.

COMPILATION PROCEDURE

The classification procedure applied for analysis requires as large a collection of objects as possible. For the greater part (292 promoters) our compilation (see table 1 in the supplementary material) was composed from *E.coli* mRNA promoters (published in ref. 3) while 94 phage and plasmid promoters recognized by *E.coli* RNA polymerase were added from previous compilations (1,2) and 54 from the original papers (23–51). The nucleotide sequences of the upstream region of phage and plasmid promoters were taken from the papers cited in references 1 and 2. Only those promoter sequences were included in our list if upstream sequences at least up to –50 were available. Promoter alignment is a crucial point in any statistical method. The generally used alignment, according to regions –35 and –10, requires two preliminary steps: (i) the canonical regions should be identified in every promoter and (ii) blank positions should be inserted in the spacer and between –10 region and start point. The former procedure often gives alternative possibilities, while the later one makes impossible the cluster analysis in the regions with blank positions. To overcome these problems, we aligned promoters according to their +1 position, which is the strongest experimentally determined characteristic of a promoter DNA. In cases where transcription initiation is possible from numerous sites, all were independently tested and the one providing the best fit with comparable promoter in the analyzed region was selected.

CLUSTERING PROCEDURE

Analysis was performed using our own software CLUSTER (available by Email: deev@venus.iteb.serpukhov.su) based on the hierarchical clustering method (22). Promoters were compared for the local positions by their subsequences at the distance **X** from the transcription start point, varied in the range $-70 \le \mathbf{X} \le -1$. The value **R** of resemblance for two promoters Pi and Pj was determined as a number of coinciding nucleotides in the selected promoter subsequences, the length **L** of which was varied in the range $2 \le \mathbf{L} \le 10$. If there was more than one +1 position in the Pi and (or) Pj, **R** was calculated as the maximum number of coinciding nucleotides for optimal selection of +1 point in Pi and Pj. Taking into account the fact that the length of the spacer and the distance between the –10 region and the start point of transcription vary in individual promoters, we allowed them to shift along each other for a value $\mathbf{V} \le 2$ to provide a more obvious comparison. In this case Pi and Pj were initially aligned according to +1. Then the Pi promoter was consecutively shifted along Pj for a value **s** in the range $-\mathbf{V} \le \mathbf{s} \le +\mathbf{V}$ to get the maximum number of matching nucleotides in the positions corresponding to

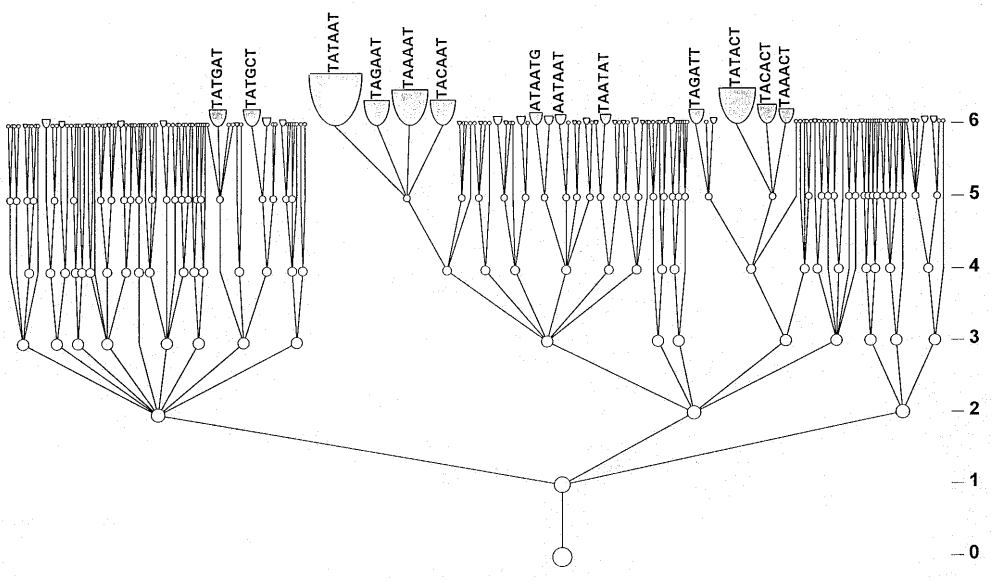


Figure 1. An example of dendrogram obtained for **X** = –12, **L** = 6, **V** = 2. Ciphers on the right indicate the value of **H**. The difference in the group size at the level **H** = 6 is schematically represented by the size of half-ellipses. Consensus sequences of the largest groups are indicated.

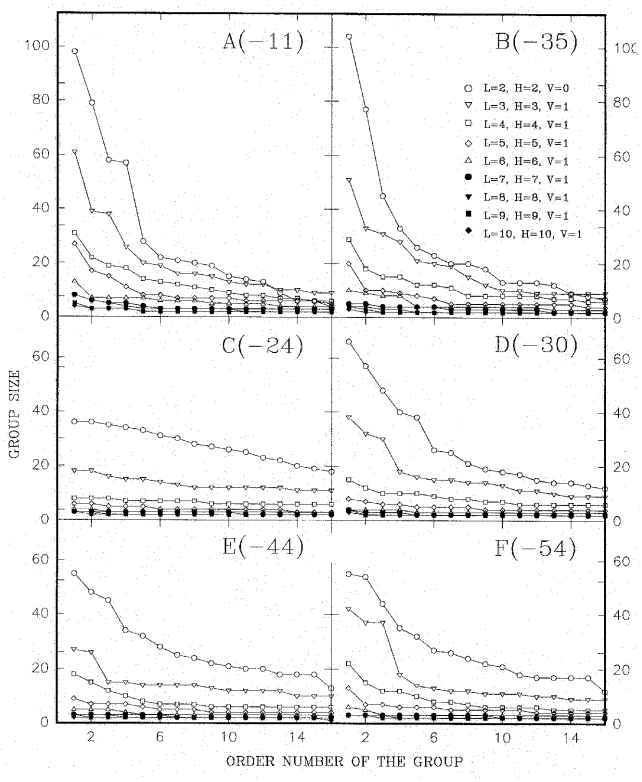


Figure 2. Number of promoters in the 16 largest groups obtained in some critical promoter positions. Parameters used for clustering are indicated in the panel B.

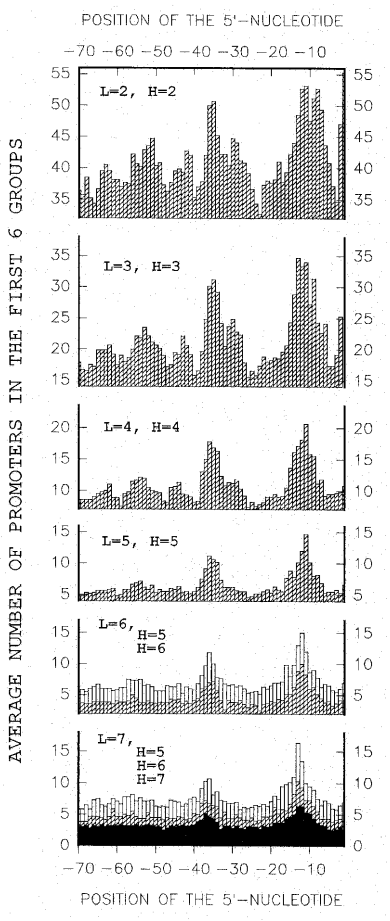


Figure 3. Position dependence for the average number of promoters in the six largest groups obtained for different values of **L** and **H**.

the Pi subsequence. This procedure was repeated by shifting the Pj promoter along Pi and the resulting value of **R** was determined as the maximum number of coinciding nucleotides for an optimal position of shifting. If identical values were obtained for several alignments, the one which corresponded to the smaller shift was chosen. Based on this definition for promoter resemblance **R**, hierarchical dendrograms (Fig. 1) were generated by sequential grouping of the promoters according to their resemblance in the range $1 \leq \mathbf{R} \leq \mathbf{L}$. On the first step the clustering procedure collects the promoters which have $\mathbf{R} = \mathbf{L}$. On the second and subsequent steps the groups are combined to get the highest value of overall similarity. For further description we have designated the groups of the level **H** those groups in which $\mathbf{R} \geq \mathbf{H}$ for each pair of promoters.

RESULTS

Site-specific promoter subgrouping

The hierarchical dendrograms were sequentially generated by the clustering procedure and analyzed at all promoter positions in the range from $-70 \leq \mathbf{X} \leq -1$ for the length **L** of compared subsequences varied from 2 to 10 and **V** varied from 0 to 2. The number of promoter groups varied from 16 (**L** = 2 at all positions) up to 428 (**L** = 10, **H** = 10, **X** = –24) whereas their size varied from 1 to 119 (TA dinucleotide at position –12). Figure 2 represents the real number of promoters in the largest groups obtained at some critical positions. An average size of promoter groups and their number characterize an efficiency of promoter subgrouping and, hence, a degree of promoter homology at every position according to the transcription start site. Both these parameters exhibit a pronounced dependence on **X**, revealing the specific regions with non-random distribution of base pairs. Figure 3 demonstrates the site-dependence for an average number of promoters in the largest six groups (a point where the curves of the group sizes approach a plateau, Fig. 2). An average number of promoters in the groups increases when the value of **H** is taken smaller than **L** (lower panels, Fig. 3) and decreases when averaging is performed for all groups with a size >1 (data not shown). In all cases, the character of the histograms remains the same: when the 5′ end of the compared subsequences was located at –54 ± 4, –44 ± 3, –35 ± 3, –29 ± 2 and –11 ± 4 (regions I–V, respectively) an average size of promoter groups significantly increases (Fig. 3). The number of promoters in the largest groups obtained in these critical positions exhibits specific distribution, different from that observed, for example, at position –24, which shows the lowest level of clustering and can probably be used as a convenient intrinsic reference point. Some degree of subgrouping is also observed around the position –65, showing a level of clustering lower than other promoter regions. Since the experimentally observed border of the contact area with RNA polymerase usually did not reach this position (reviewed in 20) we omitted the –65 region from further analysis. Efficiency of clustering in all regions dramatically decreases when the value of $\mathbf{L} > 6$, confining a

reasonable extent of subsequences for further analysis (histograms for $8 \leq \mathbf{L} \leq 10$ are not presented to avoid overloading the figure).

Thus, three additional regions with non-random distribution of base pairs were identified in promoters. Two of them are located in the upstream region, where specific base pair content is widely discussed in connection with promoter function. It was unexpected, however, to observe two separate patterns of clustering, which appeared to be phased with the –35 signal element. The third new region with non-random distribution of base pairs is located in the spacer just downstream from the –35 signal element.

Sequence motifs typical for the revealed promoter regions

The consensus sequences of all the promoter groups were automatically determined as output parameters of any clustering circle (Fig. 1): however, due to some fundamental features of the method, their contexts could not be directly accepted as a typical sequence motif of the region. For example, if three promoters with local sequences TATAATG, ATATAAT and ATAATGC are compared by their subsequences of **L** = 6, the first promoter, due to the shifting procedure, will be occasionally combined either with the second or with the third, giving different consensus sequences. Thus the size of the group with a specific consensus, generally determined by the frequency in the presence of a corresponding subsequence, is also dependent on the value of **V** and on the order of promoters in the compilation. To reveal the dominate sequence motifs, typical for different promoter regions, we used two-step analysis.

On the first step, promoters were occasionally mixed three times and subgrouping was performed for **V** = 0 and 2 within the ranges $-58 \leq \mathbf{X} \leq -50$; $-48 \leq \mathbf{X} \leq -40$; $-39 \leq \mathbf{X} \leq -32$; $-31 \leq \mathbf{X} \leq -27$ and $-18 \leq \mathbf{X} \leq -4$. **L** was varied in the range from 3 to 6 and groups of the level **H** = **L** and **H** = **L** – 1 were examined. Consensus sequences were determined for the six largest groups and those which take place for the all mixed sets of promoters, were subjected to the second step. At this step we analyzed the histograms of appearance for revealed sequence motifs through the promoter compilation and those elements which have a pronounced maximum in the region of clustering are summarized in Table 1. Only the longest sequence elements identified in different promoter regions are indicated for simplicity. Some promoter groups have a consensus differing in only one or two positions or a consensus shifted several steps. We combine them to obtain a larger groups. For example, six sequence motifs revealed in the –10 region are combined into two groups: one contains canonical TATAAT and motifs bearing all possible substitutions in the place of central T, while the other is composed of sequence elements displaying a conserved C 1 bp upstream of the 3′ end (see also Fig. 1). Some additional considerations were taken into account at this stage. For example, the ACAAGC motif (group **5**) looks like a prolongation of CACAA (group **4**), nevertheless they are placed into separate subsets because the corresponding groups contain only five common promoters (intersection of line **5** and column **4** in Table 2) and the group with consensus CACAAG was not identified by cluster analysis.

Three types of sequences are found upstream of the –10 element. One is a TG dinucleotide which is usually separated from the –10 sequence by A, C or T, but not G. Two types of sequence motifs are identified downstream from the –10 element. The GCGC is typical for the stringently controlled promoters and is responsible for their inhibition by ppGpp. Promoters may be

Table 1. Sequence motifs typical for the regions with non-random distribution of base pairs

-54±4		Upstream of the -35 element		Upstream of the -10 element	
1.	AAAAAT (–58/–53) AAAAAA (–59/–53) AAAGAT (–54/–51)	a).	TCTTGA (–40/–34) TTCTTG (–41/–33)	d).	CGTATA (–16/–11) GCGTAA (–17/–12) CCGTAT (–16/–12) AAGCGT (–20/–17)
2.	ATTTTT (–54/–50) TTTTTT (–57/–50)	b).	GTTGAC (–39/–35) GTTGCA (–38/–34)	e).	TGATA (–17/–12) TGTTA (–17/–12) TGCTA (–16/–12)
3.	TGATTT (–58/–53) CTGATT (–58/–56) TGTTTA (–57/–55) GTTTT (–57/–52)	c).	CATTGA (–38/–35) GCATTG (–39/–36)	f).	CCTATA (–16/–14) CCCTAT (–17/–14) TCCCTA (–18/–16)
4.	ATTCAC (–58/–55) AATCA (–60/–51) TCACA (–59/–51) CACAA (–59/–51)	-35 element		-10 element	
5.	AAGCTT (–51/–50) ACAAGC (–59/–52) CAAGCT (–52/–50)	14.	TTGACA (–38/–32) TTGAAA (–35/–32) TTTACA (–38/–32)	22.	TATAAT (–14/–9) TAAAAT (–15/–9) TACAAT (–14/–9) TAGAAT (–14/–12)
6.	CTTTAC (–52/–50)	15.	TTGCAA (–37/–32) TGCAAA (–36/–32)	23.	TATACT (–14/–10) TACACT (–15/–9)
-44±4		16.	TTGACT (–38/–33) TGACTC (–37/–33)	Downstream from the -10 element	
7.	TGAAAT (–49/–45) GTGTAA (–49/–47)	-29±2		g).	TGCGCC (–9/–6) TGCGCG (–10/–5) CGCCGC (–8/–4) CGCCTC (–7) GCGCGG (–9/–4)
8.	AATTCT (–49/–43) ATTCTC (–48/–42) AGTTCT (–43/–41)	17.	TGAATA (–32/–27) GATAAA (–32/–30)	h).	ACTGAA (–9) CTGAAA (–8) TGAAAG (–7)
9.	TTTTTC (–49/–40) TGTTTC (–48/–42)	18.	TTTACC (–30/–25) ATTTAC (–31/–26) TTACCC (–31/–25)		
10.	AAAGAA (–45/–44) AAAAA (–49/–41)	19.	AAAATA (–32/–28) AAAAA (–31/–27)		
11.	ATCAAA (–49/–42)	20.	AAACTG (–32/–26)		
12.	TTGCAT (–48/–41)	21.	CCTTTT (–29/–28)		
13.	CATTGA (–49/–43) ACATT (–48/–42)				

Typical sequence elements revealed in the promoter structure and combined on the basis of their similarity. Sequence elements obtained in the peak positions of regions I–V are numbered. Elements characterizing sequences flanking regions III and V are designated by lower case letters. The numbers in the brackets indicate promoter regions containing the peak fraction of the corresponding sequence elements with **L** = **R**.

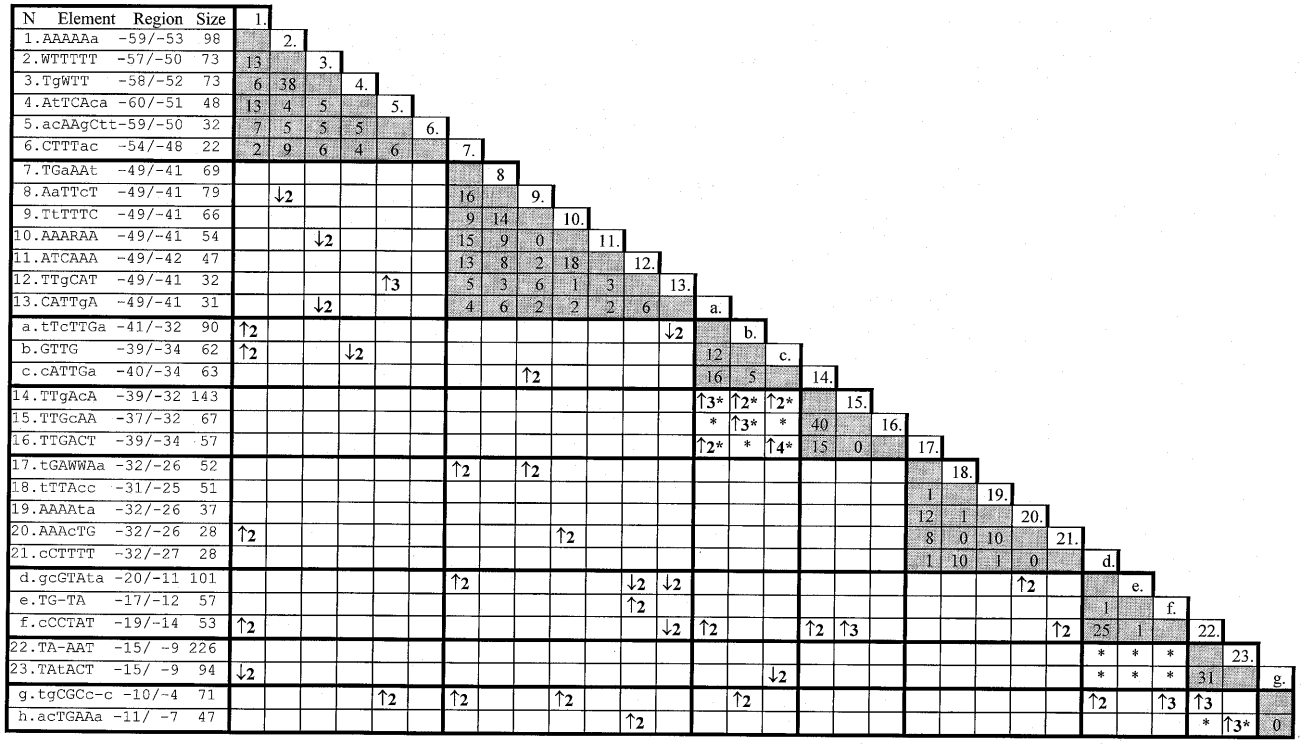
subgrouped on the basis of their sequence similarity upstream of the –35 element, where they often possess TC, CA or G. The presence of C in this region was already discussed (52). Five dominant sequence motifs are identified in the –30 region. Three are enriched in homopurine (homopyrimidine) dinucleotides, previously reported for promoters with long spacer length (15). It was surprising to find that many promoters have sequences resembling the –35 element one helix turn upstream from its normal location (groups 12 and 13).

Thus, numerous sequence motifs are identified in the promoter structure. Since the cluster analysis was performed for segments of different length and different levels of coincidence, the set presented in Table 1 shall probably be considered as the maximum list of putative promoter-specific elements encoded in the nucleotide sequence.

Correlation in the presence of different sequence elements in the one promoter group

The last part of this study was performed with the aim of revealing a possible correlation in the presence of different sequence elements in the promoter structure. A rationale for this analysis was governed by the fact that promoters bearing TG dinucleotides upstream of the –10 element form tight contacts with RNA polymerase in the upstream region and permit the complete elimination of –35 elements (17,19). This suggests a possibility

Table 2. Distribution of promoter-specific sequence elements between different promoter groups



N	Element	Region	Size
1.	AAAAAa	−59/−53	98
2.	WTTTTT	−57/−50	73
3.	TgWTT	−58/−52	73
4.	AtTCAca	−60/−51	48
5.	acAAgCtt	−59/−50	32
6.	CTTTac	−54/−48	22
7.	TGaAAt	−49/−41	69
8.	AaTTcT	−49/−41	79
9.	TtTTTC	−49/−41	66
10.	AAARAA	−49/−41	54
11.	ATCAAA	−49/−42	47
12.	TTgCAT	−49/−41	32
13.	CATTgA	−49/−41	31
a.	tTcTTGa	−41/−32	90
b.	GTTG	−39/−34	62
c.	cATTGa	−40/−34	63
14.	TTgAcA	−39/−32	143
15.	TTGcAA	−37/−32	67
16.	TTGACT	−39/−34	57
17.	tGAWWAa	−32/−26	52
18.	tTTAcc	−31/−25	51
19.	AAAAta	−32/−26	37
20.	AAAcTG	−32/−26	28
21.	cCTTTT	−32/−27	28
d.	gcGTAta	−20/−11	101
e.	TG-TA	−17/−12	57
f.	cCCTAT	−19/−14	53
22.	TA-AAT	−15/ −9	226
23.	TAtACT	−15/ −9	94
g.	tgCGCc-c	−10/−4	71
h.	acTGAAa	−11/ −7	47

N and symbols upstream from the columns designate the name of the promoter group according to the Table 1. Numbers in the shaded cells indicate a quantity of the common promoters in the groups named on the top of the column and on the left from the crossing line. Arrows and ciphers in the white cells indicate the sign and the value of **B** in deviation of **n** from the expected value ('2' means $1.96 \le \mathbf{B} < 3$; '3' means $3 \le \mathbf{B} < 4$ and so on). To analyze the correlation with maximum efficiency, promoters bearing sequences exactly corresponding to the typical elements of the groups **14**, **16** and **22**, **23** were ascribed to only one group. Otherwise the size of these groups is: 163 (**14**); 75 (**16**); 253 (**22**); 132 (**23**) and they contain a large number of intersections. Group **b** contains promoters which have at least five matches with indicated sequences and also possess G on the 5′-end. Borders of some regions were slightly changed comparing to Table 1 in accordance with the specific distribution exhibited by the motif when **R** = 5 was taken lower than **L** = 6. Asterisks designate those cases when overlapping in sequence motifs was not taken into account, otherwise overlapping was allowed for not more than one position. W, A or T; R, A or G.

that specific pathways requiring particular non-canonical elements can be used by RNA polymerase for complex formation with some promoters. To test this possibility in principle, promoters which have at least five matches with every revealed sequence motif were extracted as representatives of the corresponding promoter group and their consensus sequences in the examined regions were determined. The value 5 for **R** was chosen to obtain a statistically reliable pool of promoters in every group and, on the other hand, to avoid a large number of intersections between different groups for the same promoter region. In those cases where different sequence elements of the group have shifted 5′ ends (see for example groups **3**, **4** and **5**), flanking sequences were taken into account so as to align them completely before weighting. Different bases of the sequence elements show some degree of variation. Those which are present in a specific position in >75% of promoters in the group are designated by capital letters and those which were found in 50–75% of promoters are designated by lower case letters (Table 2). The presence of alternative bases in one and the same position was indicated only if both of them are present in 40–50% of selected promoters.

All pairs of promoter groups were analyzed (Table 2) and in most cases the quantity of common promoters in two different groups corresponded to the value theoretically expected for the size of these groups. Nevertheless, in some cases, statistically relevant deviations were found. The significance **B** of this phenomena was estimated using a non-parametric statistical method (53):

$$\mathbf{B} = (\mathbf{n} - \mathbf{Nq})/(\mathbf{Nq}(1 - \mathbf{q}))^{1/2}$$

where **n** is the number of common promoters for two analyzed groups; **N** is the size of one group; **q** is the percentage of another group in the overall promoter compilation.

Within the 95% confidence interval, the correlation in the presence of two different sequence elements in the structure of one and the same promoter group was considered as insignificant if $|\mathbf{B}| < 1.96$ (empty white cells in the Table 2). Only 41 out of 408 pairs of signal elements show certain positive ($\mathbf{n} > \mathbf{Nq}$) or negative ($\mathbf{n} < \mathbf{Nq}$) correlations in their simultaneous presence. The level of these interdependencies was not high if compared with a **B** value of 95% confidence interval, probably indicating that different sets of the signal elements may create a functional promoter.

DISCUSSION

Our present knowledge on the structure of promoter sites for *E.coli* RNA polymerase is based on a large number of experimental data, as well as on statistic analysis of promoter sequences. The priority

of the later approaches was especially evident in revealing consensus hexamers. Their significance for promoter functioning was confirmed by later genetic, biochemical and biophysical data. Experimental data in turn attract attention to the non-canonical promoter regions and were the challenge of searching for additional promoter-specific patterns which occur imperfectly in many sequences. The cluster analysis performed in this paper was governed by this problem.

Five regions permitting promoter subgrouping were revealed by cluster analysis (Fig. 3). Canonical –35 and –10 regions show the highest clustering efficiency, displaying TTGACA and TATAAT as dominant elements. The occurrence of particular base pairs in these regions, however, shows specific variations allowing a possibility for promoter subgrouping into several subsets which show slightly different correlation with the presence of other sequence motifs. Putative promoter-specific elements are also identified within –54 ± 4, –44 ± 4 and –29 ± 2 regions (Table 1). The situation here is however quite different: sequence motifs typical for any of these regions show no mutual sequence similarity, implying a difference in their functional manifestation.

Approximately 99% of promoters from our compilation possess a high degree of sequence homology with at least one of the revealed sequence motifs (table 1 in the supplementary material). The average size of the largest groups determined for regions I–V decreases in order: V > III > (IV) ≅ I ≥ II (Fig. 2). The number of promoters bearing at least one of the sequence motifs in these regions displays slightly different dependence: V > III ≅ II ≥ I > IV (table 1 in the supplementary material), indicating a statistical significance of the elements found in the upstream region.

At least three different mechanisms affecting transcription initiation are functioning upstream from the –35 element. First of all, this region frequently bears target sites for a group of regulatory proteins (54,55). Second, the upstream regions of promoters are often A/T-rich. A/T tracts positioned in phase with helix turn are capable of inducing DNA bending which sometimes shows transcription enhancing activity (reviewed in 56). Third, this region in some promoters is recognized by the C-terminal domain of α-subunit (18).

Cluster analysis revealed a group of promoters with TCACA as a consensus sequence upstream of –50 (group **4**). This element is a part of cAMP-CRP binding site and 15 promoters from this group are regulated by different regulatory proteins: only four of them, however, *mtl*, *ompB*, *rbs*, *tnaA*, are regulated by CRP through interaction with TCACA. The presence of this element in other promoters probably contributes to promoter function in some other way. The fact that eight of 20 promoters which complex with RNA polymerase display a hyperreactivity to DNase I at ~–47 have ATTCA or TCAC motifs upstream from the hyperreactive site suggests that this element can form direct contacts with RNA polymerase (20).

Prolonged A/T tracts are found in the upstream region of many promoters (groups **1**, **2**, **6**, **9** and **10**, Table 1 and 2). It is natural to suggest that these elements create a specific three-dimensional structure, and the possibility of DNA bending in these regions is widely discussed as an efficient transcription-enhancing factor. Several repeats of A/T tracts positioned in phase with helix turn are necessary to induce a pronounced DNA bend and many promoters really possess these repeats. Nevertheless, the presence of A/T stretches in the –54 region show no positive correlation with the presence of the same or complementary stretches around the –44 region, implying that some other promoter-specific signals may be created by these elements. The AAAT motif was found near the hyperreactive sites in polymerase-promoter complexes and it is duplicated in the UP element of *rrnBP1*, which forms contacts with the RNA polymerase α-subunit. Thus, this particular element may be suggested as a possible target for interaction with the C-terminal domain of α-subunit.

All elements revealed in the –45 region have similarity with sequences found near the sites hyperreactive to DNAse I at position –47 or –37 (20), suggesting their participation in complex formation with RNA polymerase. Most of them bear elements resembling motifs revealed in –55 and –29 regions. Two sequence motifs (groups **12** and **13**) have similarity with –35 elements (groups **15** and **c**, respectively), indicating that these elements may also be duplicated in some promoters. This observation is probably of large importance since it allows the possibility that promoter signals similar in structure but different in location may be employed by RNA polymerase. Therefore the topology of the enzyme surfaces involved in this interaction becomes of interest.

Those promoters which have a high degree of homology with canonical –35 element often possess TC, G or CA as a flanking nucleotide at the 5′ end while the promoters with TTGACT sequence in the –35 region prefer CA, and with TTGCAA, G in this position. The presence of these particular nucleotides upstream of the –35 element probably indicates that 'extended –35' elements may sometimes be recognized by RNA polymerase. It should be taken into account, however, that the consensus elements of groups **a–c** possess TTG common with the consensus elements of the groups **14–16**. That means that the selection of the promoters in groups **a–c** and **14–16** was not independent and absolute values of **B** in these cases are overestimated.

The non-random distribution of purine–purine homo-dinucleotides and purine–pyrimidine hetero-dinucleotides in the six upper positions of the spacer was previously described for promoters with a spacer length different from 17 bp (15). Some sequence motifs found in this region are really enriched with homodinucleotides. However only two groups have a percentage of promoters with non-optimal spacer length essentially different from average (46%): the subset **20** bearing AAAcTG motif near the –30 have 61% of promoters with non-optimal spacer length, whereas the subset **17**, characterized by sequence motif tGAWWAa, only 26%. This, in principle, is in line with the suggestion that some specific sequence motifs may be functioning in the spacer region to compensate variation in its length (15) and indicates that the presence of other motifs requires a stronger correspondence of this distance to the optimal value.

Two types of elements suggested for the –10 region show some difference in correlation with the presence of sequence motifs **1**, **c**, **h** and **g**. Three typical dinucleotides were found in the region flanking the –10 element. The presence of TG 1 bp upstream from the –10 element is important for transcription activity of *galP1* and *cysG* (17,19), forming a so-called 'extended –10' element. These promoters show an extended footprint in the upstream region and are active in the absence of the –35 element. Nine promoters (instead of an expected four), which have TG upstream of the –10 element have a high degree of homology with TTgCAT near position –45. TTgCAT is well expressed in *galP1*. *CysG* has only three matches with this sequence bearing four coincidences with TTTACA, another motif typical for –35 region. Both *galP1* and *cysG* have weak homology with –35 elements in their normal location. Thus it could not be excluded that the specific topology

of the complexes formed by these two promoters with RNA polymerase is conditioned by unusual interaction of σ (or some other subunit) with the –35-like element one helix turn upstream from its normal position. This statement however could not be attributed to all promoters with TG dinucleotide upstream from –10 element or promoters with TTGCAT motif near –45 site, since neither of these groups showed a pronounced decrease in the presence of typical –35 elements.

Two types of sequence motifs were revealed downstream from the –10 element. One of them (group **g**) corresponds to the well known element responsible for stringent regulation (57). The other has not been discussed as such; nevertheless, it was reported that promoters which have A/T-rich sequences between the –10 element and the start point of transcription are activated in the presence of ppGpp (58), implying a possibility that two different types of elements located between –10 region and start point of transcription are employed to provide transcription regulation by metabolic agents.

The presence of a particular sequence element in the promoter group shows no strong correlation with the presence of any other sequence motif. This observation indicates that practically all combinations of signal elements (with or without certain preferences) may be found in the promoter structure and is in line with the model of alternative pathways proposed for promoter activation on the basis of experimental data (16,17,20,21). This model suggests that at every step of complex formation, the promoter recognition center of RNA polymerase can identify a set of promoter specific elements that can be different in nature and position. Specific conformational transitions accompany interaction with a particular combination of signal elements and create prerequisites for the next stage, at which some alternative possibilities for the enzyme to form further contacts can also exist. An absence of strong correlations in the presence of particular signal elements in the promoter structure implies that the pathway of complex formation is not strongly determined by initial interaction of the enzyme with any of them. Those cases which show slight positive correlation in the presence of certain sequence elements may probably give some information on the nature of combinations preferentially used by enzyme at the same or subsequent stages of complex formation. Special attention should probably be paid to the elements which show negative correlation in their simultaneous presence since they may mark out evolutionary unfavorable combinations.

ACKNOWLEDGEMENTS

This work was supported by the Foundation for Basic Research of Russian Government (grants 97-04-49418 and 97-04-48404).

See supplementary material available in NAR Online.

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 Res.*, **15**, 2343–2361.
3 Lisser,S. and Margalit,H. (1993) *Nucleic Acids Res.*, **21**, 1507–1516.
4 Mulligan,M.E., Hawley,D.K., Entriken,R. and McClure,W.R. (1984) *Nucleic Acids Res.*, **12**, 789–800.
5 Mulligan,M. and McClure,W.R. (1986) *Nucleic Acids Res.*, **14**, 109–126.
6 O'Neill,M.C. (1989) *J. Biol. Chem.*, **264**, 5522–5530.
7 O'Neill,M.C. and Chiafari,F. (1989) *J. Biol. Chem.*, **264**, 5531–5534.
8 Lukashin,A.V., Anshelevich,V.V., Amirikyan,B.R., Gragerov,A.I. and Frank-Kamenetskii,M.D. (1989) *J. Biomol. Struc. Dynam.*, **6**, 1123–1133.
9 Demeler,B. and Zhou,G. (1991) *Nucleic Acids Res.*, **19**, 1593–1599.
10 O'Neill,M.C. (1991) *Nucleic Acids Res.*, **19**, 313–318.
11 O'Neill,M.C. (1992) *Nucleic Acids Res.*, **20**, 3471–3477.
12 Mahadevan,I. and Ghosh,I. (1994) *Nucleic Acids Res.*, **22**, 2158–2165.
13 Rozkot,F., Sazelova,P. and Pivec,L. (1989) *Nucleic Acids Res.*, **17**, 4799–4815.
14 Cardon,L.R. and Stormo,G.D. (1992) *J. Mol. Biol.*, **223**, 159–170.
15 Beutel,B.A. and Record,M.T.,Jr (1990) *Nucleic Acids Res.*, **18**, 3597–3603.
16 Chan,B. and Busby,S. (1989) *Gene*, **84**, 227–236.
17 Belyaeva,T., Griffiths,L., Minchin,S., Cole,J. and Busby,S. (1993) *Biochem. J.*, **296**, 851–857.
18 Ross,W., Gosink,K.K., Salomon,K., Igarashi,K., Zou,C., Ishihama,A., Severinov,K. and Gourse,R.L. (1993) *Science*, **262**, 1407–1413.
19 Belyaeva,T.A., Bown,J.A., Fujita,N., Ishihama,A. and Busby,S. (1996) *Nucleic Acids Res.*, **24**, 2243–2251.
20 Ozoline,O.N. and Tsyganov,M.A. (1995) *Nucleic Acids Res.*, **22**, 4533–4541.
21 Ozoline,O.N., Uteshev,T.A., Masulis,I.S. and Kamzolova,S.G. (1993) *Biochim. Biophys. Acta*, **1172**, 251–261.
22 Duda,R. and Hart,P. (1973) Pattern recognition and scene analysis, NY.
23 Sakena,P. and Walker,J.R. (1992) *J. Bacteriol.*, **174**, 1956–1964.
24 Schnetz,K. and Bak,B. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 1244–124.
25 Arnqvist,A., Olsen,A. and Normark,S. (1994) *Mol. Microbiol.*, **13**, 1021–1032.
26 Aiba,H. (1985) *J. Biol. Chem.*, **260**, 3063–3070.
27 Kukolj,G. and DuBow,M.S. (1992) *J. Biol. Chem.*, **267**, 17827–17835.
28 Lewis,L.K. and Mount,D.W. (1992) *J. Bacteriol.*, **174**, 5110–5116.
29 Hofer,B., Muller,D. and Koster,H. (1985) *Nucleic Acids Res.*, **13**, 5995–6013.
30 Weickert,M. and Adhya,S. (1993) *J. Bacteriol.*, **175**, 251–258.
31 Brun,J.V., Sanfacon,H., Breton,R. and Lapointe,J. (1990) *J. Mol. Biol.*, **214**, 845–864.
32 Kohno,K., Wada,M., Kano,Y. and Imamoto,F. (1990) *J. Mol. Biol.*, **213**, 27–36.
33 Mulvey,M.R. and Loewen,P.C. (1989) *Nucleic Acids Res.*, **17**, 9979–9993.
34 Tartaglia,L.A., Storz,G. and Ames,B.N. (1989) *J. Mol. Biol.*, **210**, 709–719.
35 Peterson,M.L. and Reznikoff,W.S. (1985) *J. Mol. Biol.*, **185**, 525–533.
36 Bell,A.J., Cole,J.A. and Busby,S.J.W. (1990) *Mol. Microbiol.*, **4**, 1753–1763.
37 Womble,D.D., Sampathkumar,P., Easton,A.M., Luckow,V. and Rownd,R.H. (1985) *J. Mol. Biol.*, **181**, 395–410.
38 Durwin,A., Hussain,H., Griffiths,L., Grove,J., Sambongi,Y., Busby,S. and Cole,J. (1993) *Mol. Microbiol.*, **9**, 1255–1265.
39 Huang,L., Tsui,P. and Freundlich,M. (1992) *J. Bacteriol.*, **174**, 664–670.
40 Kubota,M., Yamazaki,Y. and Ishihama,A. (1991) *Jpn. J. Genet.*, **66**, 399–409.
41 Kukochi,Y., Yoda,K., Yamasaki,M. and Tamura,G. (1981) *Nucleic Acids Res.*, **9**, 5671–5678.
42 Surin,B.P., Jans,D.A., Fimmel,A.L., Shaw,D.C., Cox,G.B. and Rosenberg,H. (1984) *J. Bacteriol.*, **157**, 772–778.
43 Mikuni,O., Ito,K., Maffat,J., Matsumura,K., McCanghan,K., Nogukuni,T., Tate,W. and Nakamura,Y. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 5798–5802.
44 Makino,K., Shinagawa,H., Muemura,M., Kimura,S., Nakata,A. and Ishihama,A. (1988) *J. Mol. Biol.*, **203**, 85–95.
45 Slany,R.K. and Kersten,H. (1992) *Nucleic Acids Res.*, **20**, 4193–4198.
46 Yamagishi,M., Matsushima,H., Wada,A., Sakagami,M., Fujita,N. and Ishihama,A. (1993) *EMBO J.*, **12**, 625–630.
47 Burton,Z.F., Gross,C.A., Watanabe,K.K. and Burgess,R.R. (1983) *Cell*, **32**, 335–349.
48 Young,R.A. and Steitz,J.A. (1979) *Cell*, **17**, 225–234.
49 Dunn,J. and Studier,F.W. (1983) *J. Mol. Biol.*, **166**, 477–535.
50 Prosen,D.E. and Cesh,C.L. (1986) *Biochemistry*, **25**, 5378–5388.
51 Kanazawa,H., Mabuchi,K. and Futai,M. (1982) *Biochem. Biophys. Res. Commun.*, **107**, 568–575.
52 Galas,D.J., Eggert,M. and Waterman,M.S. (1985) *J. Mol. Biol.*, **186**, 117–128.
53 Hollander,M. and Wolfe,D.A. (1973) *Nonparametric Statistical Methods*. J.Willey and Sons, New York–London–Sydney–Toronto, pp. 15–26.
54 Busby,S. and Kolb,A. (1995) in Lin,E.C.C. and Lynch,A.S. (eds), *Regulation of Gene Expression in Escherichia coli*. R.G.Landes Company, pp. 255–279.
55 Collado-Vides,J., Magasanik,B. and Gralla,J.D. (1991) *Microbiol. Rev.*, **55**, 371–394.
56 Travers,A. (1987) *CRC Crit. Rev. Biochem.*, **22**, 181–219.
57 Travers,A. (1984) *Nucleic Acids Res.*, **12**, 2605–2618.
58 Travers,A.A. (1980) *J. Mol. Biol.*, **141**, 91–97.